MATH 170S Introduction to Probability and Statistics II
Lecture Note

Hanbaek Lyu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095

Email address: hlyu@math.ucla.edu

WWW.HANAEKLYU.COM

Contents

CHAPTER 1

Review of basic probability theory

1. Probability measure and probability space

Many things in life are uncertain. Can we 'measure' and compare such uncertainty so that it helps us to make more informed decision? Probability theory provides a systematic way of doing so.

We begin with idealizing our situation. Let Ω be a finite set, called *sample space*. This is the collection of all possible outcomes that we can observe (think of six sides of a die). We are going to perform some experiment on Ω, and the outcome could be any subset E of Ω, which we call an *event*. Let us denote the collection of all events $E \subseteq \Omega$ by 2^{Ω}. A *probability measure* on Ω is a function $\mathbb{P} : 2^{\Omega} \rightarrow [0,1]$ such that for each event $E \subseteq \Omega$, it assigns a number $\mathbb{P}(E) \in [0,1]$ and satisfies the following properties:

(i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(ii) If two events $E_1, E_2 \subseteq \Omega$ are disjoint, then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_1)$.

In words, $\mathbb{P}(E)$ is our quantization of how likely it is that the event E occurs out of our experiment.

Exercise 1.1.1. Let $\mathbb{P}$ be a probability measure on sample space Ω. Show the following.

(i) Let $E = \{x_1, x_2, \cdots x_k\} \subseteq \Omega$ be an event. Then $\mathbb{P}(E) = \sum_{i=1}^{k} \mathbb{P}(\{x_i\})$.

(ii) $\sum_{x \in \Omega} \mathbb{P}(\{x\}) = 1$.

If $\mathbb{P}$ is a probability measure on sample space Ω, we call the pair $(\Omega, \mathbb{P})$ a *probability space*. This is our idealized world where we can precisely measure uncertainty of all possible events. Of course, there could be many (in fact, infinitely many) different probability measures on the same sample space.

Exercise 1.1.2 (coin flip)**.** Let $\Omega = \{H, T\}$ be a sample space. Fix a parameter $p \in [0,1]$, and define a function $\mathbb{P}_p : 2^{\Omega} \rightarrow [0,1]$ by $\mathbb{P}_p(\emptyset) = 0$, $\mathbb{P}_p(\{H\}) = p$, $\mathbb{P}_p(\{T\}) = 1 - p$, $\mathbb{P}_p(\{H, T\}) = 1$. Verify that $\mathbb{P}_p$ is a probability measure on Ω for each value of p.

A typical way of constructing a probability measure is to specify how likely it is to see each individual element in Ω. Namely, let $f : \Omega \rightarrow [0,1]$ be a function that sums up to 1, i.e., $\sum_{x \in \Omega} f(x) = 1$. Define a function $\mathbb{P} : 2^{\Omega} \rightarrow [0,1]$ by

$$\mathbb{P}(E) = \sum_{\omega \in E} f(x). \tag{1}$$

Then this is a probability measure on Ω, and f is called a *probability distribution* on Ω. For instance, the probability distribution on $\{H, T\}$ we used to define $\mathbb{P}_p$ in Exercise 1.1.2 is $f(H) = p$ and $f(T) = 1 - p$.

Example 1.1.3 (Uniform probability measure)**.** Let $\Omega = \{1, 2, \cdots, m\}$ be a sample space and let $\mathbb{P}$ be the *uniform probability measure* on Ω, that is,

$$\mathbb{P}(\{x\}) = 1/m \qquad \forall x \in \Omega. \tag{2}$$

Then for the event $A = \{1, 2, 3\}$, we have

$$\begin{align} \mathbb{P}(A) &= \mathbb{P}(\{1\} \cup \{2\} \cup \{3\}) \tag{3} \\ &= \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{3\}) \tag{4} \\ &= \frac{1}{m} + \frac{1}{m} + \frac{1}{m} = \frac{3}{m} \tag{5} \end{align}$$

Likewise, if $A \subseteq \Omega$ is any event and if we let $|A|$ denote the size (number of elements) of A, then

$$\mathbb{P}(A) = \frac{|A|}{m}. \tag{6}$$

For example, let $\Omega = \{1,2,3,4,5,6\}^2$ be the sample space of a roll of two fair dice. Let A be the event that the sum of two dice is 5. Then

$$A = \{(1,4),(2,3),(3,2),(4,1)\}, \tag{7}$$

so $|A| = 4$. Hence $\mathbb{P}(A) = 4/36 = 1/9$. ▲

2. Random variables

Given a finite probability space $(\Omega, \mathbb{P})$, a (discrete) *random variable* (RV) is any real-valued function $X : \Omega \to \mathbb{R}$. We can think of it as the outcome of some experiment on Ω (e.g., height of a randomly selected friend). We often forget the original probability space and specify a RV by is *probability mass function* (PMF) $f_X : \mathbb{R} \to [0,1]$,

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}). \tag{8}$$

Namely, $\mathbb{P}(X = x)$ is the likelihood that the RV X takes value x.

Example 1.2.1. Say you win \$1 if a fair coin lands heads and lose \$1 if lands tails. We can set up our probability space $(\Omega, \mathbb{P})$ by $\Omega = \{H, T\}$ and $\mathbb{P}$ = uniform probability measure on Ω. The RV $X : \Omega \to \mathbb{R}$ for this game is $X(H) = 1$ and $X(T) = -1$. The PMF of X is given by $f_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 1/2$ and likewise $f_X(-1) = 1/2$.

Exercise 1.2.2. Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a RV. Let f_X be the PMF of X, that is, $f_X(x) = \mathbb{P}(X = x)$ for all x. Show that f_X adds up to 1, that is,

$$\sum_x f_X(x) = 1, \tag{9}$$

where the summation runs over all numerical values x that X can take.

There are two useful statistics of a RV to summarize its two most important properties: Its average and uncertainty. First, if one has to guess the value of a RV X, what would be the best choice? It is the *expectation* (or mean) of X, defined as below:

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x). \tag{10}$$

Exercise 1.2.3. For any RV X and real numbers $a, b \in \mathbb{R}$, show that

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{11}$$

On the other hand, say you play two different games where in the first game, you win or lose \$1 depending on a fair coin flip, and in the second game, you win or lose \$10. In both games, your expected winning is 0. But the two games are different in how much the outcome fluctuates around the mean. This notion if fluctuation is captured by the following quantity called *variance*:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{12}$$

Namely, it is the expected squared difference between X and its expectation $\mathbb{E}(X)$.

Exercise 1.2.4. Let X be a RV. Show that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Here are some of the simplest and yet most important RVs.

Exercise 1.2.5 (Bernoulli RV)**.** A RV X is a *Bernoulli* variable with (success) probability $p \in [0,1]$ if it takes value 1 with probability p and 0 with probability $1-p$. In this case we write $X \sim \mathrm{Bernoulli}(p)$. Show that $\mathbb{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.

Exercise 1.2.6 (Indicator variables)**.** Let $(\Omega, \mathbb{P})$ be a probability space and let $E \subseteq \Omega$ be an event. The *indicator variable* of the event E, which is denoted by $\mathbf{1}_E$, is the RV such that $\mathbf{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbf{1}_E(\omega) = 0$ if $\omega \in E^c$. Show that $\mathbf{1}_E$ is a Bernoulli variable with success probability $p = \mathbb{P}(E)$.

Example 1.2.7 (Uniform RV)**.** Let $\Omega = \{x_1, x_2, \cdots x_n\}$ be a finite sample space of real numbers. A RV X is a *uniform* variable on Ω, denoted as $X \sim \text{Uniform}(\Omega)$, if it takes each of the element in Ω with equal probability. That is,

$$\mathbb{P}(X = x_i) = \frac{1}{n} = \frac{1}{|\Omega|} \qquad \text{for all } 1 \le i \le n. \tag{13}$$

Let $X \sim \text{Uniform}(\Omega)$ as above. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{14}$$

Also,

$$\text{Var}(X) = \sum_{i=1}^{n} (x_i - \mathbb{E}[X])^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathbb{E}[X])^2. \tag{15}$$

The following exercise ties the expectation and the variance of a RV into a problem of finding a point estimator that minimizes the mean squared error.

Exercise 1.2.8 (Variance as minimum MSE)**.** Let X be a RV. Let $\hat{x} \in \mathbb{R}$ be a number, which we consider as a 'guess' (or 'estimator' in Statistics) of X. Let $\mathbb{E}[(X - \hat{x})^2]$ be the *mean squared error* (MSE) of this estimation.

(i) Show that

$$\begin{aligned} \mathbb{E}[(X - \hat{x})^2] &= \mathbb{E}[X^2] - 2\hat{x}\mathbb{E}[X] + \hat{x}^2 \qquad (16) \\ &= (\hat{x} - \mathbb{E}[X])^2 + \mathbb{E}[X^2] - \mathbb{E}[X]^2 \qquad (17) \\ &= (\hat{x} - \mathbb{E}[X])^2 + \text{Var}(X). \qquad (18) \end{aligned}$$

(ii) Conclude that the MSE is minimized when $\hat{x} = \mathbb{E}[X]$ and the global minimum is $\text{Var}(X)$. In this sense, $\mathbb{E}[X]$ is the 'best guess' for X and $\text{Var}(X)$ is the corresponding MSE.

To define a discrete RV, it was enough to specify its PMF. For a continuous RV, *probability distribution function* (PDF) plays an analogous role of PMF. We also need to replace summation $\sum$ with an integral $\int dx$. Namely, X is a *continuous RV* if there is a function $f_X : \mathbb{R} \to [0, \infty)$ such that for any interval $[a, b]$, the probability that X takes a value from an interval $(a, b]$ is given by integrating f_X over the interval $(a, b]$:

$$\mathbb{P}(X \in (a, b]) = \int_a^b f_X(x)\, dx. \tag{19}$$

The *cumulative distribution function* (CDF) of a RV X (either discrete or continuous), denoted by F_X, is defined by

$$F_X(x) = \mathbb{P}(X \le x). \tag{20}$$

By definition of PDF, we get

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt. \tag{21}$$

Conversely, PDFs can be obtained by differentiating corresponding CDFs.

Exercise 1.2.9. Let X be a continuous RV with PDF f_X. Let a be a continuity point of f_X, that is, f_X is continuous at a. Show that $F_X(x)$ is differentiable at $x = a$ and

$$\left.\frac{dF_X}{dx}\right|_{x=a} = f_X(a). \tag{22}$$

The expectation of a continuous RV X with pdf f_X is defined by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx, \tag{23}$$

and its variance $\mathrm{Var}(X)$ is defined by the same formula (12).

CHAPTER 2

Descriptive Statistics

1. Basic sample statistics

One of the core problem in statistics is making inference on an unknown random variable X using its sample values $x_1, x_2, \cdots, x_n$ (not necessarily distinct). From these sample values of X, how can we 'infer' the true distribution of X? The basic idea is to approximate the distribution of X by the *empirical distribution* X.

Definition 2.1.1. Let X be a RV, and $x_1, \cdots, x_n$ be n sample values of X. Then the *empirical distribution* X is a probability distribution on the set $S = \{x_1, \cdots, x_n\}$ with

$$f(x_i) = \frac{1}{n}(\text{\# of times that the value } x_i \text{ appeared}) \tag{24}$$

Remark 2.1.2. If the sample values $x_1, \cdots, x_n$ are all distinct, then the empirical distribution becomes the uniform distribution on the set $S = \{x_1, \cdots, x_n\}$.

Example 2.1.3. Say we have a coin with unknown probability p of coming up heads. We want to figure out the true value of p. A natural way to do this would be simply flipping the coin many times, and see how many heads come up versus tails. For example, suppose we flip this coin 10 times and the result was

$$(x_1, x_2, \cdots, x_{10}) = (H, H, T, H, H, T, T, H, H, T). \tag{25}$$

Then the corresponding empirical distribution is given by

$$f(H) = \frac{6}{10}, \qquad f(T) = \frac{4}{10}. \tag{26}$$

Can we conclude that $p \approx 6/10$? ▲

In many cases, computing the exact distribution of an unknown RV maybe costly or difficult. Instead, computing its mean and variance gives a rough idea on how large and 'spread out' it is, respectively. Since we do not have the full distribution of X, but only its sample values $x_1, \cdots, x_n$, we can only approximately compute the true mean $\mathbb{E}[X]$ and variance $\mathrm{Var}(X)$.

Definition 2.1.4. Let X be a RV and $x_1, \cdots, x_n$ be its sample values. Define the following:

$$\text{(sample mean)} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{27}$$

$$\text{(variance of the empirical dist.)} \qquad v = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{28}$$

$$\text{(sample variance)} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n}{n-1} v, \tag{29}$$

$$\text{(sample standard deviation)} \qquad s = \sqrt{s^2} \tag{30}$$

Remark 2.1.5. Why do we have $n-1$ instead of n in the definition of sample variance? In later sections, we will see that s^2 is in some sense a 'better estimator' than v for the population variance $\mathrm{Var}(X)$. See Exercise 3.2.4.

Exercise 2.1.6. Let $X \sim \text{Bernoulli}(p)$ for unknown parameter p. Say after some experiment, we have its sample values

$$(x_1, x_2, \cdots, x_{10}) = (1, 1, 0, 1, 1, 0, 0, 1, 1, 0). \tag{31}$$

Compute the corresponding sample mean, variance of the empirical distribution, sample variance, and sample standard deviation.

Exercise 2.1.7. Let $X \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$, which can also be thought as the outcome of rolling a fair die. Suppose we have sample values

$$(x_1, x_2, \cdots, x_{10}) = (2, 4, 2, 5, 6, 1, 3, 3, 2, 6). \tag{32}$$

Compute the corresponding sample mean, variance of the empirical distribution, sample variance, and sample standard deviation.

Example 2.1.8 (Excerpted from [HTZ77])**.** Let X be the weight of a candy bar produced in a factory A. Due to the uncertainty in the production process, it is reasonable to think of X as a random variable. In order to get to knowledge on X, we may collect actual weights of n candy bars from this factory. This will give us *sample values* of X, which are shown in table below:

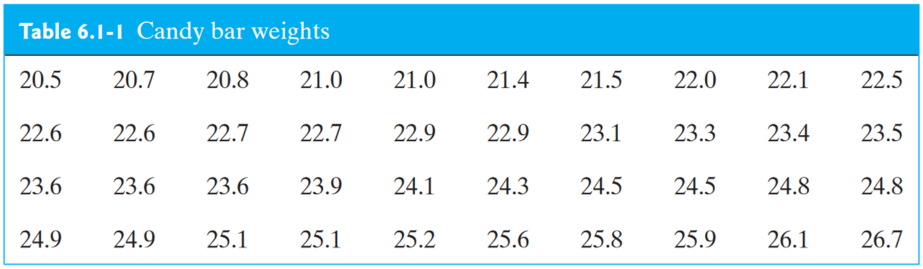


Table 6.1-1 Candy bar weights

20.5	20.7	20.8	21.0	21.0	21.4	21.5	22.0	22.1	22.5
22.6	22.6	22.7	22.7	22.9	22.9	23.1	23.3	23.4	23.5
23.6	23.6	23.6	23.9	24.1	24.3	24.5	24.5	24.8	24.8
24.9	24.9	25.1	25.1	25.2	25.6	25.8	25.9	26.1	26.7

FIGURE 1. Sample values of candy bar weights.

We can of course directly think of the empirical distribution corresponding to the above sample values of X. But in some cases (especially for hugh sample size), it is often convenient to group the sample values and consider the resulting 'coarse-grained' empirical distribution. We will illustrate this through this example.

We start from observing that the minimum and maximum sample values are 20.5 and 26.7, respectively. The interval $[20.5, 26.7]$ has length 6.2, so it can be covered with $k = 7$ sub-intervals of length 0.9. So we will divide the interval $[20.45, 26.75]$ in $k = 7$ sub-intervals of equal length 0.9. For each sub-interval, its midpoint is called its *class mark*. We then count how many sample values fall in each of these seven intervals and make histogram, where the ith sub-interval I_i has height equals to its f_i.

A more probabilistic way of making such histogram is called the *relative frequency histogram* or the *density histogram*. Namely, for each sub-interval I,

$$\text{Area of density histrogram over } I_i = \frac{f_i}{n} \approx \mathbb{P}(X \in I). \tag{33}$$

The corresponding density histogram of our running example is shown below.

▲

Exercise 2.1.9. Compute the sample mean and sample variance of the raw sample values given in Example 2.1.8. Also compute the sample mean and sample variance of the grouped data, using the class marks and with their respective frequencies. How do they compare?

Table 6.1-2 Frequency table of candy bar weights

Class Interval	Class Limits	Tabulation	Frequency (f_i)	$h(x)$	Class Marks
(20.45, 21.35)	20.5–21.3	𝍸	5	5/36	20.9
(21.35, 22.25)	21.4–22.2	||||	4	4/36	21.8
(22.25, 23.15)	22.3–23.1	𝍸 |||	8	8/36	22.7
(23.15, 24.05)	23.2–24.0	𝍸 ||	7	7/36	23.6
(24.05, 24.95)	24.1–24.9	𝍸 |||	8	8/36	24.5
(24.95, 25.85)	25.0–25.8	𝍸	5	5/36	25.4
(25.85, 26.75)	25.9–26.7	|||	3	3/36	26.3

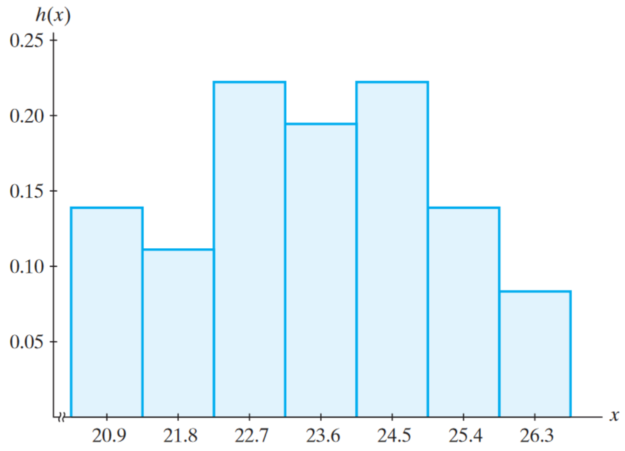


Figure 6.1-1 Relative frequency histogram of weights of candy bars

2. Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods, first promoted by John Tukey. Since modern datasets are usually too large to be able to comprehended by naked human eyes, advanced EDA techniques such as dimensionality reduction, clustering, and data visualization are becoming very important. This could lead to a better understanding of the data set, yielding better modeling or formulating more insightful hypothesis.

In this section, we will learn elementary methods of EDA, such as *stem-and-leaf display*, *order statistics*, *sample percentiles*, and *box-plot*.

Example 2.2.1 (Ordered Stem-and-leaf display, Excerpted from [HTZ77])**.** Suppose the scores in Midterm 1 of Math 170S turned out to be as below:

Drawing histogram of the sample data would be one way to visually express the data set. Tukey suggested the following more informative 'version' of histogram, which does not 'collapse' sample values into a class as in histogram. For instance, we think of the sample value 93 as a stem of 9 with leaf of 3. So 'stems' would be the first digits of the exam scores, and 'leaves' would be the second digits. We can also order the leaves from the smallest to the largest for more clean representation. Finally, we also record

93	77	67	72	52	83	66	84	59	63
75	97	84	73	81	42	61	51	91	87
34	54	71	47	79	70	65	57	90	83
58	69	82	76	71	60	38	81	74	69
68	76	85	58	45	73	75	42	93	65

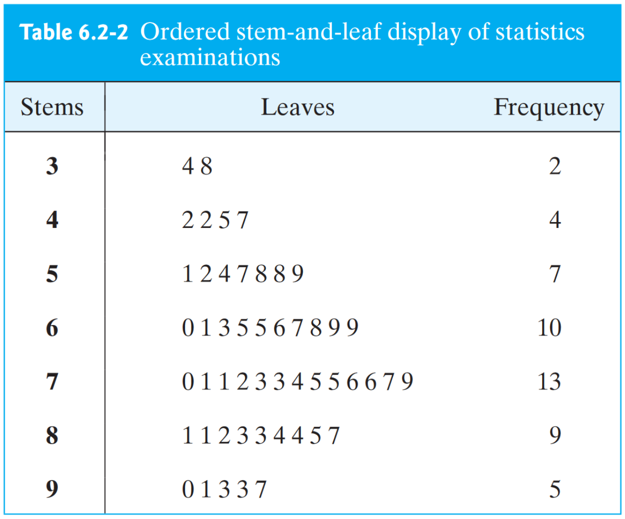


Table 6.2-2 Ordered stem-and-leaf display of statistics examinations

Stems	Leaves	Frequency
3	4 8	2
4	2 2 5 7	4
5	1 2 4 7 8 8 9	7
6	0 1 3 5 5 6 7 8 9 9	10
7	0 1 1 2 3 3 4 5 5 6 6 7 9	13
8	1 1 2 3 3 4 4 5 7	9
9	0 1 3 3 7	5

the frequency of each stem, which equals to the number of its leaves. The resulting diagram is as above, which is called an (ordered) *stem-and-leaf display*.

▲

Given a set of sample values $x_1, x_2, \cdots, x_n$, we can rearrange these values so that they are ordered from the smallest to the largest. Namely, let y_1 be the smallest value in the sample; let y_2 be the second smallest value in the sample, and so on. So we can rearrange the sample as $y_1 \leq y_2 \leq \cdots \leq y_n$. We call y_k as the kth order statistics of the sample.

The *median* of the sample, roughly speaking, is the value m so that half of the sample values are $< m$ and the other half are $> m$; so m is the 'middle' value. One should not confuse the median with the sample mean. For example, if the sample values for $1,3,8$, then the median is 3, but the mean is $(1+3+8)/3 = 4$. But what if there is no middle value? Say, the sample values are $1,3,5,8$. When the median should be somewhere 'between' 3 and 5. In this case, we take a balanced average $(3+5)/2 = 4$ to be the median.

In general, fix $p \in (0,1)$. The $(100p)$*th sample percentile* (or *sample percentile of order* p) of the sample, which we denote by $\tilde{\pi}_p$, is the value such that (approximately) np of the sample values are $< \tilde{\pi}_p$, and $n(1-p)$ of them are $> \tilde{\pi}_p$. In particular, if $p = 1/2$, then the 50th sample percentile is exactly the median. If $(n+1)p$ is an integer r, according to the Exercise 2.2.2, we choose rth order statistic y_r as the $(100p)$th sample percentile $\tilde{\pi}_p$. What if $(n+1)p$ is not integer, but an integer r plus a proper rational $\delta \in (0,1)$? We then take the weighted average of y_r and y_{r+1} with weights δ and $1-\delta$. In general, if $(n+1)p = r + \delta$ for a unique integer r and $\delta \in [0,1)$, then

$$\tilde{\pi}_p = (1-\delta) y_r + \delta y_{r+1}. \tag{34}$$

The reasoning behind this formula is provided in the following exercise.

Exercise 2.2.2. Suppose we have a sample of values $x_1, x_2, \cdots, x_n$. Let y_k denote the kth order statistics of this sample. Fix $p \in (0,1)$. Write $(n+1)p = r + \delta$ for a unique integer r and $\delta \in [0,1)$. Show the following:

(i) (# of sample values $\leq y_r$) $= r = np + (p - \delta)$.

(ii) (# of sample values $> y_r$) $= n - r = n(1-p) - p + \delta$.

Example 2.2.3 (Order statistics and sample percentiles, Excerpted from [HTZ77])**.** The following table shows order statistics of the $n = 50$ exam scores in Example 2.2.1. We will compute some sample percentiles. First, let $p = 1/2$. The since, $(n+1)p = 51 * (1/2) = 25.5$, the 50th sample percentile is

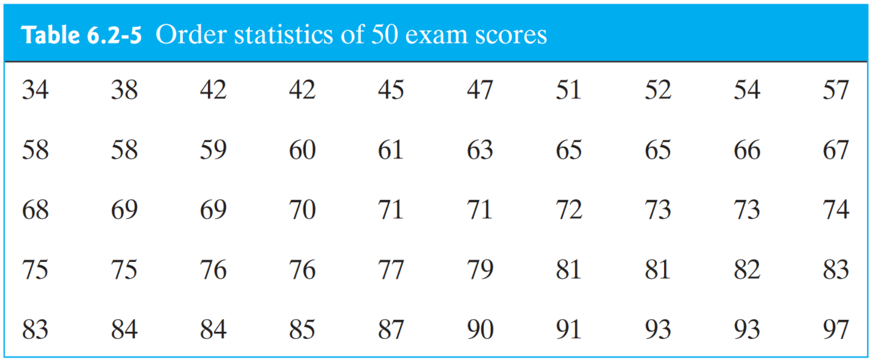


Table 6.2-5 Order statistics of 50 exam scores

34	38	42	42	45	47	51	52	54	57
58	58	59	60	61	63	65	65	66	67
68	69	69	70	71	71	72	73	73	74
75	75	76	76	77	79	81	81	82	83
83	84	84	85	87	90	91	93	93	97

$$\tilde{\pi}_{0.5} = (0.5)y_{25} + (0.5)y_{26} = (71+71)/2 = 71. \tag{35}$$

For $p = 1/4$, we have $(n+1)p = 51 * (1/4) = 12.75$, so the 25th sample percentile is

$$\tilde{\pi}_{0.25} = (0.25)y_{12} + (0.75)y_{13} = (0.25)(58) + (0.25)(59) = 58.75. \tag{36}$$

Lastly, for $p = 3/4$, we have $(n+1)p = 51 * (3/4) = 38.25$, so the 75th sample percentile is

$$\tilde{\pi}_{0.75} = (0.75)y_{38} + (0.25)y_{39} = (0.25)(81) + (0.25)(82) = 81.25. \tag{37}$$

▲

Special names are given to certain sample percentiles. We have noted that the 50th percentile is the *median* of the sample. The 25th, 50th, and 75th percentiles are also called the *first, second, and third quartiles* of the sample, respectively. We also use special notation for them:

$$\tilde{q}_1 = \tilde{\pi}_{0.25}, \qquad \tilde{q}_2 = \tilde{m} = \tilde{\pi}_{0.5}, \qquad \tilde{q}_3 = \tilde{\pi}_{0.75}. \tag{38}$$

The *five-number summary* of a sample consists of its three quantiles as well as their minimum y_1 and maximum y_n. The *box-plot* is a diagramatic representation of this five-number summary. See the following example.

Example 2.2.4 (Five-number summary and box plot)**.** We use the sample of 50 exam scores in Example 2.2.1. According to the computation in Example 2.2.3, the five-number summary of the sample is

$$y_1 = 34, \qquad \tilde{q}_1 = 58.75, \qquad \tilde{q}_2 = \tilde{m} = 71, \qquad \tilde{q}_3 = 81.25, \qquad y_{50} = 97. \tag{39}$$

The corresponding box plot is shown below. The middle line corresponds to the median, whereas the rectangular box are bounded by the first and third quantiles. The vertial line spans from the minimum y_1 to the maximum y_{50}.

▲

Exercise 2.2.5. Suppose we have the following sample of Google's stock price for the past 50 weeks (unit in \$ per stock).

320 326 325 318 322 320 329 317 316 331 320 320 317 329 316 308 321
319 322 335 318 313 327 314 329 323 327 323 324 314 308 305 328 330
322 310 324 314 312 318 313 320 324 311 317 325 328 319 310 324

(i) Compute the sample mean $\bar{x}$ and sample standard deviation s.

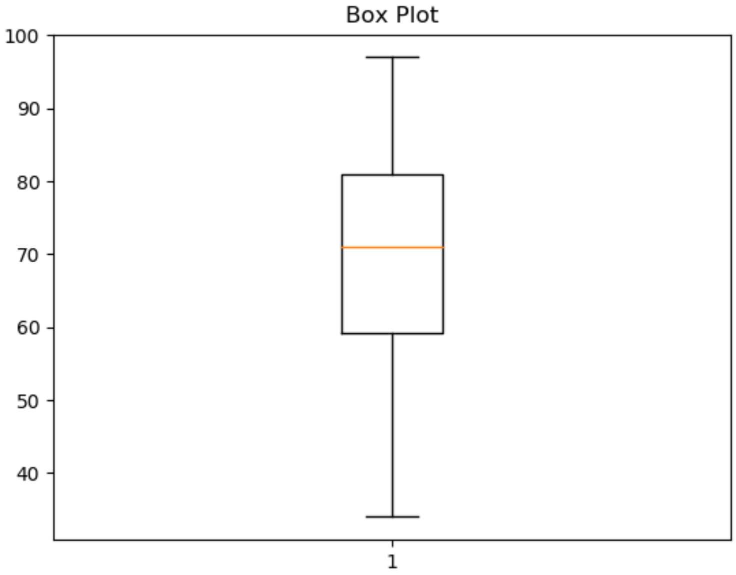


TABLE 1. Box plot of the 50 exam scores data in Example 2.2.1.

(ii) Draw the ordered stem-and-leaf display. How many sample values are between $\bar{x} \pm s$, and $\bar{x} \pm 2s$?
(iii) Give the five-number summary of the sample. Draw the corresponding box plot.

3. Order statistics

3.1. Distribution of Order Statistics. *Order statistics* are the observations of the random sample, rearranged in magnitude from the smallest to the largest. They are extremely important in statistics for determining basic sample statistics such as sample median, sample quantile, sample range, and the empirical CDF. In recent years, they are also used in nonparametric inference and robust procedures.

In this section, we assume that we have n independent and identically distributed (i.i.d) continuous random variables $X_1, X_2, \cdots, X_n$. According to the following exercise,

Exercise 2.3.1. Let $X_1, \cdots, X_n$ be i.i.d. continuous RVs. Show that
(i) $\mathbb{P}(X_i = x) = 0$ for all $1 \le i \le n$ and $x \in \mathbb{R}$.
(ii) Use conditioning and (i) to show that

$$\mathbb{P}(X_i = X_j) = \mathbb{E}[\mathbb{P}(X_i = X_j | X_j)] = 0. \tag{40}$$

(iii) From (ii), deduce that

$$\mathbb{P}(\text{there is a tie in } X_1, \cdots, X_n) = \mathbb{P}\left(\bigcup_{1 \le i < j \le n} \{X_i = X_j\} \right) \le \sum_{1 \le i < j < n} \mathbb{P}(X_i = X_j) = 0 \tag{41}$$

Now we define the order statistics for the random sample $X_1, X_2, \cdots, X_n$. Let

$$Y_1 = \text{smallest of } X_1, X_2, \cdots, X_n \tag{42}$$
$$Y_2 = \text{second smallest of } X_1, X_2, \cdots, X_n \tag{43}$$
$$\vdots \tag{44}$$
$$Y_n = \text{largest of } X_1, X_2, \cdots, X_n. \tag{45}$$

We call Y_k the *kth order statistic* of the sample $X_1, \cdots, X_n$.

Below, we will derive the CDF of order statistics. To begin, we start with the largest one Y_n. Note that for each $y \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(Y_n \le y) &= \mathbb{P}(X_1 \le y, X_2 \le y, \cdots, X_n \le y). && (46)\\
&= \mathbb{P}(X_1 \le y)\mathbb{P}(X_2 \le y)\mathbb{P}(X_n \le y) && (47)\\
&= \mathbb{P}(X_1 \le y)^n, && (48)
\end{aligned}$$

where we have used the fact that X_i's are independent for the second equality, and that they have identical distribution for the second. Next, consider the minimum Y_1.

$$\begin{aligned}
\mathbb{P}(Y_1 \le y) &= 1 - \mathbb{P}(Y_1 > y) && (49)\\
&= 1 - \mathbb{P}(X_1 > y, X_2 > y, \cdots, X_n > y) && (50)\\
&= 1 - \mathbb{P}(X_1 > y)\mathbb{P}(X_2 > y)\cdots\mathbb{P}(X_n > y) && (51)\\
&= 1 - \mathbb{P}(X_1 > y)^n. && (52)
\end{aligned}$$

So we can reduce the CDF of Y_n and Y_1 to some functions of that of X_1. We generalize this observation in Exercise 2.3.2.

There is also a quite reasonable heuristic argument that gives the PDF of Y_k. Consider the following infinitesimal probability

$$\begin{aligned}
\mathbb{P}(y \le Y_k < y+h) &= \mathbb{P}(\text{smallest } k \text{ of } X_i\text{'s are} < y+h,\ k\text{th smallest one lies in } [y, y+h)) && (53)\\
&= \mathbb{P}(\text{exactly } k-1\ X_i\text{'s are} \le y, \text{ one is in } [y,y+h), \text{ the others are} > y+h) && (54)\\
&\quad + O\big(\mathbb{P}(\text{more than one } X_i\text{'s are in } [y,y+h))\big). && (55)
\end{aligned}$$

Since X_i's are i.i.d., we can write

$$\begin{aligned}
\mathbb{P}(y \le Y_k < y+h) &= n\binom{n-1}{k-1}\mathbb{P}(X_1 \le y)^{k-1}\mathbb{P}(y \le X_1 < y+h)\mathbb{P}(X_1 > y+h)^{n-k} && (56)\\
&\quad + O\big(\mathbb{P}(\text{more than one } X_i\text{'s are in } [y,y+h))\big). && (57)
\end{aligned}$$

Note that the coefficients above is the number of ways to choose one X_i to put inside $[y, y+h)$ and $k-1$ X_i's in the remaining $n-1$ to put inside $(-\infty, y)$; the rest goes to $[y+h, \infty)$. Also note that $\mathbb{P}(y \le X_1 < y+h) \approx f_{X_1}(y) \cdot h$ for h small[1], and the probability that more than two X_i's fall in $[y, y+h)$ is small in the order of $O(h^2)$. A bit more precisely,

$$\mathbb{P}(y \le X_1, X_2 < y+h) = \mathbb{P}(y \le X_1 < y+h)^2 = (f_{X_1}(y) \cdot h)^2 = O(h^2), \tag{58}$$

so this yields

$$\begin{aligned}
\mathbb{P}(\text{more than one } X_i\text{'s are in } [y,y+h)) &= \mathbb{P}(\text{some two } X_i\text{'s are in } [y,y+h)) && (59)\\
&= \binom{n}{2}\mathbb{P}(y \le X_1, X_2 < y+h) = O(h^2). && (60)
\end{aligned}$$

Combining with previous estimates, we get

$$f_{Y_k}(y) = \lim_{h \to 0} \frac{\mathbb{P}(y \le Y_k < y+h)}{h} = \frac{n!}{(k-1)!(n-k)!}\mathbb{P}(X_1 \le y)^{k-1}\mathbb{P}(X_1 > y)^{n-k} f_{X_1}(y). \tag{61}$$

A more standard derivation of the PDF of Y_k is also given in Exercise 2.3.2.

Exercise 2.3.2. Let $Y_1, \cdots, Y_n$ be the order statistics of the sample $X_1, \cdots, X_n$.

[1] Use Taylor expansion of the PDF of Y_k or mean value theorem for integrals

(i) Show that

$$\mathbb{P}(X_i > y \text{ for exatly one } 1 \le i \le n) = \mathbb{P}\left(\bigcup_{i=1}^{n} \{X_i > y \text{ and } X_j \le y \text{ for all other } j\text{'s}\,\}\right) \tag{62}$$

$$= \sum_{i=1}^{n} \mathbb{P}\left(X_i > y \text{ and } X_j \le y \text{ for all } j \ne i\right) \tag{63}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_i > y)\mathbb{P}\left(X_j \le y \text{ for all } j \ne i\right) \tag{64}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_1 > y)\mathbb{P}(X_1 \le y)^{n-1} \tag{65}$$

$$= n\mathbb{P}(X_1 > y)\mathbb{P}(X_1 \le y)^{n-1}. \tag{66}$$

(ii) Show that

$$\mathbb{P}(X_i > y \text{ for exactly two } i\text{'s}\,) = \mathbb{P}\left(\bigcup_{1\le i<j\le n} \{X_i, X_j > y \text{ and } X_k \le y \text{ for all other } k\text{'s}\,\}\right) \tag{67}$$

$$= \sum_{1\le i<j\le n} \mathbb{P}\left(X_i, X_j > y\right)\mathbb{P}\left(X_j \le y \text{ for all } k \notin \{i,j\}\right) \tag{68}$$

$$= \sum_{1\le i<j\le n} \mathbb{P}(X_1 > y)^2\mathbb{P}(X_1 \le y)^{n-2} \tag{69}$$

$$= \binom{n}{2}\mathbb{P}(X_1 > y)^2\mathbb{P}(X_1 \le y)^{n-2}. \tag{70}$$

(iii) Using a similar argument as before, show that

$$\mathbb{P}(X_i > y \text{ for exactly } k \ i\text{'s}\,) = \binom{n}{k}\mathbb{P}(X_1 > y)^k\mathbb{P}(X_1 \le y)^{n-k}. \tag{71}$$

(iv) Use (iii) to conclude that

$$\mathbb{P}(Y_k \le y) = \mathbb{P}\left(X_i \le y \text{ for at least } k \ i\text{'s}\,\right) \tag{72}$$

$$= \sum_{\ell=k}^{n} \mathbb{P}\left(X_i \le y \text{ for exacly } \ell \ i\text{'s}\,\right) \tag{73}$$

$$= \sum_{\ell=k}^{n} \mathbb{P}\left(X_i > y \text{ for exacly } n-\ell \ i\text{'s}\,\right) \tag{74}$$

$$= \sum_{\ell=k}^{n} \binom{n}{\ell}\mathbb{P}(X_1 > y)^{n-\ell}\mathbb{P}(X_1 \le y)^{\ell}. \tag{75}$$

(v) Differentiate the CDF of Y_k in (iv) to get is pdf f_{Y_k}:

$$f_{Y_k}(y) = \frac{n}{(k-1)!(n-k)!}\mathbb{P}(X_1 \le y)^{k-1}\mathbb{P}(X_1 > y)^{n-k} f_{X_1}(y). \tag{76}$$

Exercise 2.3.3. Let $X_1, X_2, \cdots, X_5$ be i.i.d. Exp(λ) RVs. Recall that the PDF of Exp(λ) RVs is given by

$$f_X(t) = \lambda e^{-\lambda t}\mathbf{1}(t \ge 0). \tag{77}$$

Let $Y_1 < Y_2 < \cdots < Y_5$ denote the order statistics of $X_1, \cdots, X_5$.

(i) Compute the CDFs of $Y_1 < \cdots < Y_5$.
(ii) Compute the PDFs of $Y_1 < \cdots < Y_5$.
(iii) Plot the PDFs of $Y_1 < \cdots < Y_5$ assuming $\lambda = 1$.

4. Order statistics and the q-q plot

Suppose for an unknown RV X, we have a sample $x_1,\cdots,x_n$ with sample mean $\bar{x}=3.4$ and sample variance $s^2=1.3$. Probably it is reasonable to claim that $\mathbb{E}[X]\approx 3.4$ and $\mathrm{Var}(X)\approx 1.3$ (we will learn in what precise sense we could do this). Furthermore, if we would like to claim that X is in fact a normal RV with mean 3.4 and variance 1.3, how do we test this claim and quantitatively see if it is reasonable? Moreover, if the sample size n is small, then sample mean and variance may not be good approximations for there population counterpart. In this case, can we still test if our sample is arising from a normal distribution?

The core problem here is to compare the empirical distribution of X and a normal distribution with unknown mean and variance. A widely used technique in EDA for such distribution comparison is called the *quantile-quantile plot* (or *q-q plot*). The basic idea is compare the order statistics of the sample with the quantiles of the hypothetical distribution. Namely, let $y_1<\cdots<y_n$ denote the order statistics of the sample $x_1,\cdots,x_n$. For each integer $1\le r\le n$, let $p_r=r/(n+1)$. Then according to (34), we have

$$y_r=\widetilde{\pi}_{p_r}. \tag{78}$$

Suppose X has distribution close to $N(\mu,\sigma^2)$. If we denote by π_p the quantile of order p for this hypothetical distribution, then we should have

$$y_r=\widetilde{\pi}_{p_r}\approx\pi_{p_r}\qquad\text{for all } 1\le r\le n. \tag{79}$$

In this case, the scatter plot of the points (y_r,π_r) will lie on the line $y=x$.

But what if we do not know μ and σ^2? The following exercise show that we can simply use $N(0,1)$ instead of $N(\mu,\sigma^2)$, and see if the resulting points of percentiles (y_r,π_r) lie on a straight line.

Exercise 2.4.1. Let $Z\sim N(0,1)$ and $X\sim N(\mu,\sigma^2)$. Show the following.

(i) $\sigma Z+\mu\sim N(\mu,\sigma^2)$.

(ii) $\mathbb{P}(X\le t)=\mathbb{P}(Z\le(t-\mu)/\sigma)$.

(iii) Let q_p be the quantile of X of order p, i.e., $\mathbb{P}(X\le q_p)=p$. Recall that for $Z\sim N(0,1)$, z_α is defined so that $\mathbb{P}(Z>z_\alpha)=\alpha$. Then use (ii) to derive that

$$q_p=\mu+\sigma z_{1-p}. \tag{80}$$

That is, the points (z_{1-p},q_p) lie on the line $y=\mu+\sigma x$.

Example 2.4.2 (Excerpted from [HTZ77])**.** Suppose for an unknown RV X, we have the following sample of size $n=30$:

1.24	1.36	1.28	1.31	1.35	1.20	1.39	1.35	1.41	1.31	1.28	1.26	1.37	1.49	1.32
1.40	1.33	1.28	1.25	1.39	1.38	1.34	1.40	1.27	1.33	1.36	1.43	1.33	1.29	1.34

We can compute $\bar{x}=1.33$ and $s^2=0.0040$. From this, can we conclude that X approximately follows $N(1.33,0.0040)$? To help answer this question, we shall construct a q–q plot of the standard normal percentiles that correspond to $p=1/31,2/31,\cdots,30/31$ versus the ordered observations.

The figure above shows three q-q plot for the sample in the example against three hypothetical distributions: t-distribution with 20 degrees of freedom, $N(0,1)$, and $N(3,5)$. According to the q-q plots, can we conclude that the X is approximately normal? ▲

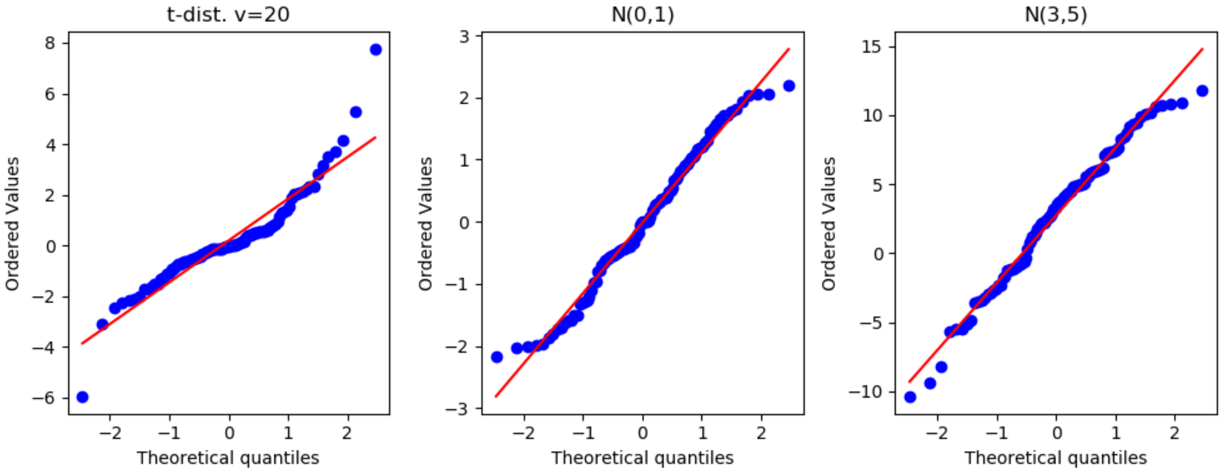


FIGURE 2. q-q plot of the sample in this example against three hypothetical distributions.

CHAPTER 3

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a classical notion in statitics, which has a prominent use in the context of modern machine learning. Arguably it is one of the most important concept that we will learn in 170S.

1. Definition and Examples of MLE

We have mentioned that, the core problem in statistics, is to infer an unknown RV X from its sample values $x_1, \cdots, x_n$. When we had absolutely no knowledge on X, in the previous sections we have seen some methods for EDA on the sample, using sample mean, variance, and quantiles, etc. However, in many cases, we can narrow down our inference problem by imposing a probabilistic (statistical) model for X. Namely, say we know X is a Bernoulli RV with unknown success probability p. We then only need to estimate the parameter p using our sample, not the entire distribution of X. MLE gives a systematic approach to this parameter estimation problem. Below we give the pipeline of MLE.

Pipeline of MLE.
Input: An unknown RV X with distribution $f_{X;\theta}$, parameterized by θ in a parameter space Ω. Also have sample values $x_1, x_2, \cdots, x_n$.
Objective: Obtain an estimation $\hat{\theta}$ of θ using the sample.
Method: Choose $\hat{\theta} \in \Omega$ so that it maximizes the following *likelihood function*

$$L(x_1, \cdots, x_n\,; \theta) = \prod_{i=1}^{n} f_{X,\theta}(x_i). \tag{81}$$

Here the RV $\hat{\theta}(X_1, \cdots, X_n)$ is called the *maximum likelihood estimator* of θ, and its observed value $\hat{\theta}(x_1, \cdots, x_n)$ is called the *maximum likelihood estimate* of θ.

What is the reasoning behind maximizing the likelihood function as above? The underlying assumption is that, *you are seeing the sample values $x_1, \cdots, x_n$ because it was the most likely to see them!* Namely, suppose we have designed the sampling procedure well-enough so that we get i.i.d. samples $X_1, \cdots, X_n$ each with distribution $f_{X;\theta}$. Then we have

$$L(x_1, \cdots, x_n\,; \theta) = \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n\,; \theta). \tag{82}$$

Namely, the likelihood function on the LHS is the probability of observing a sequence of specific sample values $x_1, \cdots, x_n$ under the i.i.d. assumption and assuming the parameter value θ. Hence, MLE chooses $\hat{\theta}$ to be the value of the parameter in Ω under which the probability of obtaining the current sample is maximized.

Example 3.1.1 (MLE for Bernoulli RVs)**.** Suppose $X \sim \text{Bernoulli}(p)$ for some unknown $p \in [0,1]$. Suppose we have sample values $x_1, \cdots, x_n$ obtained from an i.i.d. sampling for X. Denote $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. We will show that the MLE for p is $\bar{X}$, that is,

$$\hat{p} = \bar{X}. \tag{83}$$

Let $X_1, \cdots, X_n$ be i.i.d. with distribution Bernoulli(p). Note that

$$\mathbb{P}(X_i = x_i; p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases} \tag{84}$$

$$= p^{x_i}(1-p)^{1-x_i}. \tag{85}$$

Hence the likelihood function is given by

$$\begin{aligned} L(x_1,\cdots,x_n;p) &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i;p) && (86)\\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} && (87)\\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. && (88) \end{aligned}$$

In order to maximize the likelihood function, it suffices to maximize its logarithm [1]. The *log likelihood function* is given by

$$l(x_1,\cdots,x_n;p) := \log L(x_1,\cdots,x_n;p) = n\bar{x}\log p + (n-n\bar{x})\log(1-p). \tag{89}$$

To maximize the log likelihood function, we take partial derivative in p:

$$\frac{\partial l(x_1,\cdots,x_n;p)}{\partial p} = \frac{n\bar{x}}{p} - \frac{n-n\bar{x}}{1-p}. \tag{90}$$

Setting this equal to zero, we obtain

$$\frac{\bar{x}}{p} = \frac{1-\bar{x}}{1-p}. \tag{91}$$

Rearranging, we obtain $p = \bar{x}$. Hence we have (83) as desired. ▲

Exercise 3.1.2 (MLE for Binomial RV)**.** Let $X \sim \text{Binomial}(n,p)$ where n is known but p is not.

(i) Write $x = x_1 + \cdots + x_m$. Show that the log likelihood function is given by

$$l(x_1,\cdots,x_m;p) = \left(\sum_{i=1}^{m} \log\binom{n}{x_i}\right) + x\log p + (n-x)\log(1-p). \tag{92}$$

(ii) Show that the MLE for p is $\bar{X}$.

Exercise 3.1.3 (MLE for Geometric RV)**.** Let $X \sim \text{Geom}(p)$, which is a discrete RV with PMF $\mathbb{P}(X = k) = (1-p)^{k-1}p$, $k = 1,2,3,\cdots$.

(i) Show that the log likelihood function is given by

$$l(x_1,\cdots,x_n;p) = n\log p + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right)\log(1-p). \tag{93}$$

(ii) Show that the MLE for p is $1/\bar{X}$.[2]

Exercise 3.1.4 (MLE for Poisson RV)**.** Let $X \sim \text{Poisson}(\lambda)$, which is a discrete RV with PMF $\mathbb{P}(X = k) = \lambda^k e^{-\lambda}/k!$, $k = 0,1,2,\cdots$.

(i) Show that the log likelihood function is given by

$$l(x_1,\cdots,x_n;\lambda) = \log\lambda \sum_{i=1}^{n} x_i - n\lambda - \log(x_1!x_2!\cdots x_n!). \tag{94}$$

(ii) Show that the MLE for λ is $\bar{X}$.[3]

[1] since log is a strictly increasing function.

[2] Recall that $\mathbb{E}[\text{Geom}(p)] = 1/p$, so we are saying the MLE for the population mean is $\bar{X}$.

[3] Recall that $\mathbb{E}[\text{Poisson}(\lambda)] = \lambda$, so we are saying the MLE for the population mean is $\bar{X}$.

Example 3.1.5 (MLE for Exponential RVs)**.** Suppose $X \sim \text{Exp}(\lambda)$ for some unknown $\theta \in [0,1]$. We have sample values $x_1, \cdots, x_n$ obtained from an i.i.d. sampling for X. Denote $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. We will show that the MLE for λ is $1/\bar{X}$, that is,

$$1/\hat{\lambda} = \bar{X}. \tag{95}$$

Let $X_1, \cdots, X_n$ be i.i.d. with distribution Exp(λ), which have the following PDF

$$f_X(x;\lambda) = \lambda e^{-\lambda x}\mathbf{1}(x \geq 0). \tag{96}$$

Noting that $x_1, \cdots, x_n \geq 0$, it follows that the likelihood function is given by

$$L(x_1, \cdots, x_n; \lambda) = \prod_{i=1}^n f_X(x_i;\lambda) \tag{97}$$

$$= \lambda^n \prod_{i=1}^n e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n \bar{x}} = \lambda^n e^{-\lambda n \bar{x}}. \tag{98}$$

So the log likelihood function is given by

$$l(x_1, \cdots, x_n; \lambda) = n \log \lambda - \lambda n \bar{x}. \tag{99}$$

To maximize the log likelihood function, we take partial derivative in λ:

$$\frac{\partial l(x_1, \cdots, x_n; \lambda)}{\partial \lambda} = \frac{n}{\lambda} - n\bar{x}. \tag{100}$$

Setting this equal to zero, we obtain $1/\lambda = \bar{x}$. Hence $1/\hat{\lambda} = \bar{X}$, as desired. ▲

Example 3.1.6 (MLE for Normal RVs)**.** Suppose $X \sim N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R}$ and $\sigma \geq 0$. Let $X_1, \cdots, X_n$ be i.i.d. with distribution $N(\mu, \sigma^2)$. Denote the sample mean and variance of the empirical distribution by

$$\bar{X} = n^{-1}\sum_{i=1}^n X_i, \qquad V = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2. \tag{101}$$

We will show that the MLE for μ and σ^2 are given by sample mean and variance of the empirical distribution:

$$\hat{\mu} = \bar{X}, \qquad \widehat{(\sigma^2)} = V. \tag{102}$$

Recall that $N(\mu, \sigma^2)$ has the following PDF

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{103}$$

It follows that the likelihood function is given by

$$L(x_1, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^n f_X(x_i;\mu,\sigma^2) \tag{104}$$

$$= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^n \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \tag{105}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right). \tag{106}$$

So the log likelihood function is given by

$$l(x_1, \cdots, x_n; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2. \tag{107}$$

To maximize the log likelihood function, we take partial derivatives in μ and σ^2:

$$\frac{\partial l(x_1,\cdots,x_n;\mu,\sigma^2)}{\partial\mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = -\frac{n}{\sigma^2}\left(\bar{x}-\mu\right), \tag{108}$$

$$\frac{\partial l(x_1,\cdots,x_n;\mu,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{109}$$

Setting these equations to be zero, we obtain

$$\begin{cases}\bar{x}=\mu \\ -n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=0.\end{cases} \tag{110}$$

Using the first equation for the second, we obtain

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{111}$$

This shows (102). ▲

Exercise 3.1.7. Consider a RV X with the following PDF

$$f_{X;\theta}(x) = \left(\frac{\theta}{1-\theta}\right)x^{(2\theta-1)/(1-\theta)}\mathbf{1}(0<x\le 1). \tag{112}$$

(i) Show that $\mathbb{E}[X]=\theta$.
(ii) If $X_1,\cdots,X_n$ are i.i.d. from the same distribution $f_{X;\theta}$ above, find the MLE $\hat{\theta}$ for θ.
(iii) Conclude that the MLE of the population mean is not always the sample mean.

2. Unbiased estimators

Suppose we have a RV X whose distribution has an unknown parameter θ. Let $X_1,\cdots,X_n$ be i.i.d. copies of X. I the previous section, we have seen various examples where the method of MLE gives a function u such that $u(X_1,\cdots,X_n)$ is an estimator of the parameter θ (e.g., $u(X_1,\cdots,X_n)=\bar{X}$). We say this estimator is *unbiased* if

$$\mathbb{E}[u(X_1,\cdots,X_n)]=\theta. \tag{113}$$

Exercise 3.2.1. Let $X_1,\cdots,X_n$ be i.i.d. copies of a RV X of unknown expectation $\mathbb{E}[X]=\mu$. Show that

$$u(X_1,\cdots,X_n)=\bar{X} \tag{114}$$

is an unbiased estimator of μ.

Example 3.2.2 (MLE for Uniform RV)**.** Let $X_1,\cdots,X_n$ be i.i.d. samples from Uniform$([0,\theta])$ distribution, and let $Y_1<\cdots<Y_n$ be their order statistics. Recall that their PDF is given by

$$f_X(x)=\frac{1}{\theta}\mathbf{1}(x\in[0,\theta]). \tag{115}$$

Hence the likelihood function is given by

$$L(x_1,\cdots,x_n;\theta)=\theta^{-n}\mathbf{1}(x_1,\cdots,x_n\in[0,\theta]). \tag{116}$$

In order to maximize this, we need to choose θ as small as possible. But since x_i's have to be between 0 and θ, we need to have $\theta\ge\max(x_1,\cdots,x_n)$. Thus the MLE $\hat{\theta}$ for θ is

$$\hat{\theta}=\max(X_1,\cdots,X_n)=Y_n. \tag{117}$$

Is this an unbiased estimator for θ? Recall from the previous section that

$$\mathbb{P}(Y_n\le y)=\mathbb{P}(X_1,\cdots,X_n\le y)=\mathbb{P}(X_1\le y)^n=(y/\theta)^n\mathbf{1}(y\in[0,\theta]). \tag{118}$$

Hence the PDF of Y_n is

$$f_{Y_n}(y) = \frac{d}{dy}\mathbb{P}(Y_n \le y) = n\theta^{-n}y^{n-1}\mathbf{1}(y \in [0,\theta]). \tag{119}$$

It follows that

$$\mathbb{E}[Y_n] = \int_0^\theta n\theta^{-n}y^n\,dy = n\theta^{-n}\left[\frac{y^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1}\theta. \tag{120}$$

Hence Y_n is *not* an unbiased estimator of θ. However, $\frac{n+1}{n}Y_n$ is an unbiased estimator of θ, since

$$\mathbb{E}\left[\frac{n+1}{n}Y_n\right] = \frac{n+1}{n}\mathbb{E}[Y_n] = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta. \tag{121}$$

In particular, MLE does *not* always give unbiased estimator. ▲

Exercise 3.2.3 (Pythagorian theorem for variances)**.** Let $X_1,\cdots,X_n$ be i.i.d. samples from some distribution with mean μ and finite variance. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i-\bar{X})^2$.

(i) Show that

$$\begin{aligned}\sum_{i=1}^n (X_i-\mu)^2 &= \sum_{i=1}^n (X_i-\bar{X}+(\bar{X}-\mu))^2 &(122)\\ &= \sum_{i=1}^n (X_i-\bar{X})^2 + 2\sum_{i=1}^n (X_i-\bar{X})(\bar{X}-\mu) + n(\bar{X}-\mu)^2 &(123)\\ &= \left[\sum_{i=1}^n (X_i-\bar{X})^2\right] + n(\bar{X}-\mu)^2. &(124)\end{aligned}$$

(ii) From (i), deduce that

$$(n-1)S^2 = \left[\sum_{i=1}^n (X_i-\mu)^2\right] - n(\bar{X}-\mu)^2. \tag{125}$$

Exercise 3.2.4 (Sample variance is unbiased)**.** Let $X_1,\cdots,X_n$ be i.i.d. samples from some distribution with mean μ and finite variance σ^2. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i-\bar{X})^2$. We will show that the sample variance S^2 is an unbiased estimator of the population variance σ^2.

(i) Show that

$$\mathbb{E}\left[\sum_{i=1}^n (X_i-\bar{X})(\bar{X}-\mu)\right] = 0. \tag{126}$$

(ii) Show that

$$\mathbb{E}[(\bar{X}-\mu)^2] = \mathbb{E}\left[n^{-2}\left(\sum_{i=1}^n (X_i-\mu)\right)^2\right] = n^{-1}\sigma^2. \tag{127}$$

(iii) Use Exercise 3.2.3 and show that

$$\mathbb{E}\left[\sum_{i=1}^n (X_i-\mu)^2\right] = \mathbb{E}\left[\sum_{i=1}^n (X_i-\bar{X})^2\right] + n\mathbb{E}\left[(\bar{X}-\mu)^2\right]. \tag{128}$$

(iv) From (ii) and (iii), deduce that

$$\mathbb{E}\left[\sum_{i=1}^n (X_i-\bar{X})^2\right] = (n-1)\sigma^2. \tag{129}$$

(v) From (iv), show that

$$\mathbb{E}\left[S^2\right] = \sigma^2. \tag{130}$$

Exercise 3.2.5 (Sample variance is consistent)**.** Let $X_1, \cdots, X_n$ be i.i.d. samples from some distribution with mean μ and finite variance σ^2. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. We will show that the sample variance S^2 is a *consistent estimator* of σ^2, that is,

$$\lim_{n\to\infty} S^2 = \sigma^2 \qquad \text{in probability.} \tag{131}$$

(This justifies the heuristic of approximating unknown σ^2 with observed value s^2 of S^2.)

(i) Use Strong Laws of Large Numbers to deduce that

$$\mathbb{P}\left(\lim_{n\to\infty} n^{-1}\sum_{i=1}^n (X_i-\mu)^2 = \sigma^2\right) = 1, \quad and \quad \mathbb{P}\left(\lim_{n\to\infty} (\bar{X}-\mu)^2 = 0\right) = 1. \tag{132}$$

(ii) Use Exercise 3.2.3 to write

$$S^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\right] - \frac{n}{n-1}(\bar{X}-\mu)^2. \tag{133}$$

Use (i) to deduce that

$$\mathbb{P}\left(\lim_{n\to\infty} S^2 = \sigma^2\right) = 1. \tag{134}$$

(You may use the fact that if $X_n \to x$ and $Y_n \to y$ almost surely as $n\to\infty$, then $X_n + Y_n \to x + y$ and $X_n Y_n \to xy$ almost surely as $n\to\infty$.)

(iii) From (ii), deduce (131). (You may use the fact that almost sure convergence implies convergence in probability.)

Example 3.2.6 (MLE for normal RV revisited)**.** From Example 3.1.6, we have seen that the MLEs for the mean μ and variance σ^2 of $N(\mu,\sigma^2)$ RVs are given by

$$\hat{\mu} = \bar{X}, \qquad \widehat{(\sigma^2)} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2. \tag{135}$$

From Exercise 3.2.1, we know that $\bar{X}$ is an unbiased estimator of μ. However, what about the variance? According to Exercise 3.2.4 (iv), we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right] = \frac{n-1}{n}\sigma^2. \tag{136}$$

Hence the MLE for σ^2 is a biased estimator. If we had $n-1$ instead of n, we would have

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2. \tag{137}$$

Hence the sample variance S^2 is an unbiased estimator of σ^2, as we have also seen in Exercise 3.2.4. ▲

3. Method of moments

Before the method of Maximum Likelihood Estimation was popularized by R.A. Fischer during 1912 and 1922, a different method for parameter estimation, called the *method of moments*, was widely used. This still carries some advantages over MLE nowadays, especially in its computational efficiency. Recall that the kth moment of a RV X is $\mathbb{E}[X^k]$. Basically, the method of moments is to let

$$k \text{ th sample moment} = k \text{ th population moment.} \tag{138}$$

Example 3.3.1 (Gamma distribution)**.** A continuous random variable X is a *Gamma* RV with parameters α,β if it has the following PDF

$$f_{X;\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\mathbf{1}(x>0), \tag{139}$$

where $\Gamma(x)$ is the *Gamma function* defined by

$$\Gamma(x) = \int_0^\infty t^x e^{-t}\,dt. \tag{140}$$

In this case we write $X \sim \mathrm{Gamma}(\alpha,\beta)$.

If $X_1, \cdots, X_n$ are i.i.d. samples of $X \sim \mathrm{Gamma}(\alpha,\beta)$, then the likelihood function is

$$L(x_1,\cdots,x_n;\alpha,\beta) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n (x_1 x_2,\cdots,x_n)^{\alpha-1} \exp\left(-\beta\sum_{i=1}^n x_i\right). \tag{141}$$

Now maximizing the above function is computationally difficult, especially due to the presence of the Gamma function in the parameters.

Instead, using the fact that $\Gamma(\alpha,\beta)$ has mean α/β and variance α/β^2, method of moments gives

$$\frac{\alpha}{\beta} = \bar{X}, \qquad \frac{\alpha}{\beta^2} = V, \tag{142}$$

where $V = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is the variance of the empirical distribution. This gives the following method of moments estimators

$$\widetilde{\alpha} = \frac{\bar{X}^2}{V}, \qquad \widetilde{\beta} = \frac{V}{\bar{X}}. \tag{143}$$

▲

Exercise 3.3.2. Show that Maximum Likelihood Estimation and Method of Moments give the same estimators for $N(\mu,\sigma^2)$.

CHAPTER 4

Linear regression

Suppose we have a data set consisting of the pairs of numbers $(x_1, y_1), \cdots, (x_n, y_n)$. Think of y_k is the price of stock you are holding and x_k is NASDAQ composite at time k. You want to obtain a simple functional explaining the *dependent variable* y_k in terms of the *independent variable* x_k. If we think of the data as a sample for the pair of random variables (X, Y), then what you want to compute is the 'best guess' on Y given X. This is well-known to be nothing but the conditional expectation $\mathbb{E}[Y|X]$. Below we give a brief recap of conditional expectation and variance.

1. Conditional expectation and variance

Let X, Y be discrete RVs. Recall that the expectation $\mathbb{E}[Y]$ is the 'best guess' on the value of Y when we do not have any prior knowledge on Y. But suppose we have observed that some possibly related RV X takes value x. What should be our best guess on Y, leveraging this added information? This is called the *conditional expectation of Y given $X = x$*, which is defined by

$$\mathbb{E}[Y|X = x] = \sum_y y\,\mathbb{P}(Y = y|X = x). \tag{144}$$

This best guess on Y given $X = x$, of course, depends on x. So it is a function in x. Now if we do not know what value X might take, then we omit x and $\mathbb{E}[Y|X]$ becomes a RV, which is called the *conditional expectation of Y given X*.

As we have defined conditional expectation, we could define the variance of a RV Y given that anther RV X takes a particular value. Recall that the (unconditioned) variance of Y is defined by

$$\text{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2]. \tag{145}$$

Note that there are two places where we take expectation. Given X, we should improve both expectations so the *conditional variance of Y given X is defined by*

$$\text{Var}(Y\,|\,X) = \mathbb{E}[(Y - \mathbb{E}[Y\,|\,X])^2\,|\,X]. \tag{146}$$

Exercise 4.1.1. Let X and Y be RVs. Then show that

$$\text{Var}(Y\,|\,X) = \mathbb{E}[Y^2\,|\,X] - \mathbb{E}[Y\,|\,X]^2. \tag{147}$$

The following exercise explains in what sense the conditional expectation $\mathbb{E}[Y\,|\,X]$ is the best guess on Y given X, and that the minimum possible mean squared error is exactly the conditional variance $\text{Var}(Y\,|\,X)$.

Exercise 4.1.2. Let X, Y be RVs. For any function $g : \mathbb{R} \to \mathbb{R}$, consider $g(X)$ as an estimator of Y. Let $\mathbb{E}[(Y - g(X))^2\,|\,X]$ be the *mean squared error.*

(i) Show that

$$\mathbb{E}[(Y - g(X))^2\,|\,X] = \mathbb{E}[Y^2\,|\,X] - 2g(X)\mathbb{E}[Y\,|\,X] + g(X)^2 \tag{148}$$

$$= (g(X) - \mathbb{E}[Y\,|\,X])^2 + \mathbb{E}[Y^2\,|\,X] - \mathbb{E}[Y\,|\,X]^2 \tag{149}$$

$$= (g(X) - \mathbb{E}[Y\,|\,X])^2 + \text{Var}(Y\,|\,X). \tag{150}$$

(ii) Conclude that the mean squared error is minimized when $g(X) = \mathbb{E}[Y\,|\,X]$ and the global minimum is $\text{Var}(Y\,|\,X)$.

2. A simple linear regression problem

Suppose we have two unknown *dependent* RVs X and Y, for which we want to figure out the relationship between them. In order to test this, we have i.i.d. random samples $(X_1, Y_1), \cdots, (X_n, Y_n)$, each with the same distribution as (X, Y). When we have to guess a functional relationship between two RVs Y and X, our first choice should be a linear function. We will have our probabilistic model of Y in terms of X as follows:

$$Y_i \,|\, X_i = \alpha + \beta X_i + \varepsilon_i, \tag{151}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ denotes independent Gaussian 'noise'. Namely, given that we observe the value of X_i, we think of Y_i as a linear transform $\alpha + \beta X_i$ plus some Gaussian noise ε_i occurring whenever we make a measurement. This probabilistic model has three parameters, α, β, and σ^2. We will use MLE to obtain estimators for each of them.

Proposition 4.2.1. *The MLEs for the paramters α, β, and σ^2 in the linear model in (151) are given by*

$$\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \tag{152}$$

$$\widehat{\beta} = \frac{\left(\sum_{i=1}^n X_i Y_i\right) - n^{-1}\left(\sum_{i=1}^n Y_i\right)\left(\sum_{i=1}^n X_i\right)}{\left(\sum_{i=1}^n X_i^2\right) - n^{-1}\left(\sum_{i=1}^n X_i\right)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{153}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\alpha} - \widehat{\beta} X_i)^2. \tag{154}$$

PROOF. First, observe that $Y_i | X_i = x_i$'s are independent with distribution $N(\alpha + \beta x_i, \sigma^2)$. Thus the likelihood function is

$$L(x_1, y_1, \cdots, x_n, y_n; \alpha, \beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right) \tag{155}$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right). \tag{156}$$

So the log likelihood function is

$$l(x_1, y_1, \cdots, x_n, y_n; \alpha, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2. \tag{157}$$

In order to maximize this, we differentiate it with respect to α, β, and σ^2. This gives

$$\frac{\partial l}{\partial \alpha} = \sigma^{-2}\sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \tag{158}$$

$$\frac{\partial l}{\partial \beta} = \sigma^{-2}\sum_{i=1}^n x_i(y_i - \alpha - \beta x_i) = 0 \tag{159}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sigma^{-4}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = 0. \tag{160}$$

These equations yield

$$\begin{cases} \left(\sum_{i=1}^n y_i\right) - n\widehat{\alpha} - \widehat{\beta}\left(\sum_{i=1}^n x_i\right) & = 0 \\ \left(\sum_{i=1}^n x_i y_i\right) - \widehat{\alpha}\left(\sum_{i=1}^n x_i\right) - \widehat{\beta}\left(\sum_{i=1}^n x_i^2\right) & = 0 \\ n\widehat{\sigma^2} - \sum_{i=1}^n (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 & = 0. \end{cases} \tag{161}$$

Combining the first two, we get

$$\left(\sum_{i=1}^n x_i y_i\right) - \left[n^{-1}\left(\sum_{i=1}^n y_i\right) - \widehat{\beta} n^{-1}\left(\sum_{i=1}^n x_i\right)\right]\left(\sum_{i=1}^n x_i\right) - \widehat{\beta}\left(\sum_{i=1}^n x_i^2\right) = 0, \tag{162}$$

which yields

$$\widehat{\beta}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n^{-1}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \left(\sum_{i=1}^{n} x_i y_i\right) - n^{-1}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right). \tag{163}$$

This gives the first desired expression for $\widehat{\beta}$. The second expression for $\widehat{\beta}$ is easy to verify from the first by simply distributing the product and the sum. Then using the first and last equations in (161), we get the MLEs for α and σ^2, as desired. □

Exercise 4.2.2 (Method of Least Squares)**.** Suppose we have data samples $(x_1, y_1), \cdots, (x_n, y_n)$. Let $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $\mathbf{x} = (x_1, \cdots, x_n)^T$. We would like to find the best approximation of the vector $\mathbf{y}$ by a linear transform of $\mathbf{x}$. Namely, we want to find α and β such that

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} \alpha \\ \vdots \\ \alpha \end{bmatrix} + \begin{bmatrix} \beta x_1 \\ \vdots \\ \beta x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{164}$$

Write the above matrix equation as $\mathbf{Y} \approx \mathbf{XB}$. The *Least Squares Estimator* $\widehat{\mathbf{B}}$ for this problem is obtained by solving the following optimization problem

$$\begin{aligned} &\text{minimize} && \|\mathbf{Y} - \mathbf{XB}\|_2^2 && (165) \\ &\text{subject to} && \mathbf{B} \in \mathbb{R}^{2\times 1}, && (166) \end{aligned}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. (d.f. For matrix algebra and calculus, see The Matrix Cookbook.)

(i) Show that (165) is equivalent to following optimization problem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 && (167) \\ &\text{subject to} && \alpha, \beta \in \mathbb{R}. && (168) \end{aligned}$$

(ii) Show that

$$\begin{aligned} \|\mathbf{Y} - \mathbf{XB}\|_2^2 &= (\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB}) && (169) \\ &= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{XB} + \mathbf{B}^T\mathbf{X}^T\mathbf{XB}. && (170) \end{aligned}$$

(iii) Show that

$$\frac{\partial}{\partial \mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_2^2 = 2\mathbf{X}^T\mathbf{XB} - 2\mathbf{X}^T\mathbf{Y}. \tag{171}$$

Conclude that the solution of the optimization problem in this exercise is given by

$$\widehat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{172}$$

Verify that this gives

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{173}$$

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n^{-1}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - n^{-1}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{174}$$

Exercise 4.2.3. Let $\widehat{\alpha}, \widehat{\beta}$, and $\widehat{\sigma^2}$ be the MLEs for the parameters α, β, and σ^2 in the simple linear regression model

$$Y | X = \alpha + \beta X + \varepsilon, \tag{175}$$

where $\varepsilon \in N(0, \sigma^2)$ is an independent Gaussian noise. (See Proposition 4.2.1.) Suppose we are given a sample $X_1 = x_1, \cdots, X_n = x_n$.

(i) Using the fact that $Y_i|X_i = x_i$ is a $N(\alpha + \beta x_i, \sigma^2)$ RV, show that

$$\mathbb{E}[\widehat{\beta}(x_1, \cdots, x_n)] = \beta, \qquad \mathrm{Var}(\widehat{\beta}(x_1, \cdots, x_n)) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{176}$$

(ii) Use Proposition 4.2.1 and (i) to show that

$$\mathbb{E}[\widehat{\alpha}(x_1, \cdots, x_n)] = \alpha, \qquad \mathrm{Var}(\widehat{\alpha}(x_1, \cdots, x_n)) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right). \tag{177}$$

Exercise 4.2.4 (Excerpted from [HTZ77])**.** The midterm and final exam scores of 10 students in a statistics course are tabulated as shown.

Midterm:	70	74	80	84	80	67	70	64	74	82
Final:	87	79	88	98	96	73	83	79	91	94

(178)

(i) Calculate the least squares regression line for these data. (See Proposition 4.2.1 and Exercise 4.2.2.)

(ii) Plot the points and the least squares regression line onthe same graph.

(iii) Find the value of $\widehat{\sigma^2}$ (see Proposition 4.2.1).

CHAPTER 5

Sufficient Statistics

1. Definition and motivating example

Let $X_1, \cdots, X_n$ be i.i.d. copies of a RV X, whose distribution has an unknown parameter θ. We may construct an estimator $Y = u(X_1, \cdots, X_n)$ for θ. So far we have seen two ways to do so – Maximum Likelihood Estimation and Method of Moments – but there are many others. To be the most effective, we would like to extract as much information as possible from the given sample $X_1, \cdot, X_n$. For example, if our estimator is made of the first three samples X_1, X_2, X_3, then probably it is not the most informative one and we should be making use of all the other samples.

Roughly speaking, an estimator (or statistics) $Y = u(X_1, \cdots, X_n)$ for a parameter θ is called a *sufficient statistics* of θ if there is no other statistic $v(X_1, \cdots, X_n)$ that provides any additional information about θ.

Definition 5.1.1. Let $X_1, \cdots, X_n$ be i.i.d. RVs whose distribution has a parameter θ. A statistic $T = T(X_1, \cdots, X_n)$ for θ is *sufficient* if the following conditional probability

$$\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = n \,|\, T = t) \tag{179}$$

does not depend on θ.

Note that the conditional probability in (179) is typically a function of both t and θ. In the following motivating example, we will see that in some cases, knowing the value of T gets rid of the dependence on θ.

Example 5.1.2 (Sufficient statistic for Bernoulli RV)**.** Let $X_1, \cdots, X_n$ be i.i.d. RVs with distribution Bernoulli(p). Recall that we can write down the full joint distribution of the n samples:

$$f(x_1, \cdots, x_n; p) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i; p) \tag{180}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{181}$$

$$= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \tag{182}$$

From Example 3.1.1, we have seen that the sample mean $\bar{X}$ is a MLE for p.

Let $Y := X_1 + \cdots + X_n = n\bar{X}$. Suppose we know that $Y = y$ (and hence $\bar{X} = y/n$). Is it possible to use the values of the sample $X_1, \cdots, X_n$ in some clever way to extract further information for p? The answer is negative. To see this, let us compute the conditional joint distribution of $(X_1, \cdots, X_n)$ on event $Y = y$: For any $x_1, \cdots, x_n$ such that $\sum_{i=1}^n x_i = y$,

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n \,|\, n\bar{X} = y) = \frac{\mathbb{P}(X_1 = x, \cdots, X_n = x_n)}{\mathbb{P}(n\bar{X} = y)} \tag{183}$$

$$= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} \tag{184}$$

$$= \frac{p^y(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} \tag{185}$$

$$= \frac{1}{\binom{n}{y}}. \tag{186}$$

Indeed, the conditional joint distribution does not depend on the parameter p, so it does not contain any additional information on p. Hence Y (and hence $\bar{X}$) is a sufficient statistic for p. ▲

2. Sufficient statistics by factorization theorem

The following factorization theorem due to Neyman and Fisher provides a convenient way to compute sufficient statistics, which we will illustrate through a number of examples.

Theorem 5.2.1 (Neyman-Fisher)**.** *Let $X_1, \cdots, X_n$ denote RVs with joint distribution $f(x_1, \cdots, x_n; \theta)$, which depends on the parameter θ. A statistic $T = T(X_1, \cdots, X_n)$ is sufficient for θ if and only if there exists functions ϕ and g such that*

$$f(x_1, \cdots, x_n; \theta) = \phi\big(T(x_1, \cdots, x_n); \theta\big)\, g(x_1, \cdots, x_n), \tag{187}$$

where g does not depend on θ.

PROOF. (Optional*) We prove the assertion when X_i's are discrete RVs, but the argument for continuous case is the same. First suppose $T = T(X_1, \cdots, X_n)$ is a sufficient statistic for θ. Then we can write

$$\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}_\theta(T = t)\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n \,|\, T = t). \tag{188}$$

Note that the conditional probability on the right hand side does not depend on θ, since T is sufficient for θ. Thus the above give the desired factorization in (187).

Conversely, suppose (187) holds. Then

$$\mathbb{P}_\theta(T = t) = \sum_{\substack{x_1, \cdots, x_n \\ T(x_1, \cdots, x_n) = t}} \mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n) \tag{189}$$

$$= \sum_{\substack{x_1, \cdots, x_n \\ T(x_1, \cdots, x_n) = t}} \phi\big(t; \theta\big)\, g(x_1, \cdots, x_n) \tag{190}$$

$$= \phi\big(t; \theta\big) \sum_{\substack{x_1, \cdots, x_n \\ T(x_1, \cdots, x_n) = t}} g(x_1, \cdots, x_n). \tag{191}$$

Thus we have

$$\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n \,|\, T = t) = \frac{\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n, T = t)}{\mathbb{P}_\theta(T = t)} \tag{192}$$

$$= \frac{\mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n)}{\mathbb{P}_\theta(T = t)} \tag{193}$$

$$= \frac{\phi\big(Y; \theta\big)\, g(x_1, \cdots, x_n)}{\phi\big(t; \theta\big) \sum_{T(x_1, \cdots, x_n) = t} g(x_1, \cdots, x_n)} \tag{194}$$

$$= \frac{g(x_1, \cdots, x_n)}{\sum_{T(x_1, \cdots, x_n) = t} g(x_1, \cdots, x_n)}. \tag{195}$$

Hence $T = T(X_1, \cdots, X_n)$ is a sufficient statistic for θ, as desired. □

Example 5.2.2 (Sufficient statistic for Poisson RV)**.** Let $X_1, \cdots, X_n$ be i.i.d. with distribution Poisson(λ). In order to compute a sufficient statistics for λ, we write down the full joint distribution of the n samples:

$$f(x_1, \cdots, x_n; \lambda) = \prod_{i=1}^{n} f_{X_i}(x_i; \lambda) \tag{196}$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \tag{197}$$

$$= (\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}) \left(\frac{1}{x_1!x_2!\cdots x_n!} \right). \tag{198}$$

Hence any constant multiple of $\sum_{i=1}^{n} X_i$ is a sufficient statistic for λ. Recall that $\bar{X}$ is also an unbiased estimator of $\mathbb{E}[\text{Poisson}(\lambda)] = \lambda$. Hence it is natural to choose $u(X_1, \cdot, X_n) = \bar{X}$. ▲

Example 5.2.3 (Sufficient statistic for Normal RV)**.** Let $X_1, \cdots, X_n$ be i.i.d. with distribution $N(\mu, 1)$. In order to compute a sufficient statistics for μ, we write down the full joint distribution of the n samples:

$$f(x_1, \cdots, x_n; \mu) = \prod_{i=1}^{n} f_{X_i}(x_i; \mu) \tag{199}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2} \right) \tag{200}$$

$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right). \tag{201}$$

To separate the dependence of each x_i on μ, we note that

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 \tag{202}$$

$$= \sum_{i=1}^{n} \left[(x_i - \bar{x}) + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2 \right] \tag{203}$$

$$= \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right) + n(\bar{x} - \mu)^2, \tag{204}$$

where for the last equality, we have used $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$. So we can decompose the joint PDF as

$$f(x_1, \cdots, x_n; \lambda) = \left[(2\pi)^{-n/2} \exp\left(n(\bar{x} - \mu)^2 \right) \right] \left[\exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \right]. \tag{205}$$

Hence $\bar{X}$ is a sufficient statistic for μ. ▲

Exercise 5.2.4. Let $X_1, \cdots, X_n$ be i.i.d. RVs with the following PDF of exponential form

$$f_{X;\theta}(x) = \exp\left(K(x)p(\theta) + S(x) + q(\theta) \right) \mathbf{1}(x \in A), \tag{206}$$

where the support A does not depend on the parameter θ.

(i) Show that the joint likelihood function is given by

$$f(x_1, \cdots, x_n; \theta) = \exp\left(p(\theta) \left(\sum_{i=1}^{n} K(x_i) \right) + nq(\theta) \right) \exp\left(\sum_{i=1}^{n} S(x_i) \right) \mathbf{1}(x_1, \cdots, x_n \in A). \tag{207}$$

Deduce that $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for θ.

(ii) Derive sufficient statistics for the parameters of Bernoulli(p), Exp(λ), and Poisson(λ) using (i).

3. Rao-Blackwell Theorem

Theorem 5.3.1 (Rao-Blackwell)**.** *Let $X_1, \cdots, X_n$ be i.i.d. RVs of distribution $f_{X;\theta}$ with paramter θ. Let $T_1 = T_1(X_1, \cdots, X_n)$ and $T_2 = T_2(X_1, \cdots, X_n)$ be two statistics for θ. Let $T_3 = \mathbb{E}[T_2 \,|\, T_1]$.*

(i) *One can improve T_2 by conditioning on T_1. That is, for the new statistic T_3,*

$$\text{Var}(T_3) \le \text{Var}(T_2). \tag{208}$$

(ii) *If T_2 is unbiased, then T_3 is also an unbiased estimator for θ.*

(iii) *If T_1 is sufficient for θ, then T_3 depends only on the sample values $X_1, \cdots, X_n$ and not on θ.*

PROOF. Recall the law of total variance:

$$\mathrm{Var}(T_2) = \mathbb{E}[\mathrm{Var}(T_2 \,|\, T_1)] + \mathrm{Var}(\mathbb{E}[T_2 \,|\, T_1]). \tag{209}$$

Since $\mathrm{Var}(T_2 \,|\, T_1) \ge 0$,

$$\mathrm{Var}(T_3) = \mathrm{Var}(\mathbb{E}[T_2 \,|\, T_1]) \le \mathbb{E}[\mathrm{Var}(T_2 \,|\, T_1)] + \mathrm{Var}(\mathbb{E}[T_2 \,|\, T_1]) = \mathrm{Var}(T_2). \tag{210}$$

This shows (i). To show (ii), suppose $\mathbb{E}[T_2] = \theta$. Then

$$\theta = \mathbb{E}[T_2] = \mathbb{E}[\mathbb{E}[T_2 \,|\, T_1]] = \mathbb{E}[T_3]. \tag{211}$$

Hence T_3 is an unbiased estimator for θ.

Now we show (iii). By definition of conditional expectation, write

$$T_3 = \mathbb{E}[T_2 \,|\, T_1] = \sum_t t\,\mathbb{P}_\theta(T_2(X_1, \cdots, X_n) = t \,|\, T_1). \tag{212}$$

Since T_1 is sufficient for θ, the conditional probabilities in the last expression do not depend on θ. This shows that T_3 does not depend on θ. □

Remark 5.3.2. Notice that parts (i) and (ii) in the Rao-Blackwell theorem above do not require sufficiency of T_1. What is the meaning of part (iii)? Notice that, the conditional expectation $T_3 = \mathbb{E}[T_2 \,|\, T_1]$ defining T_3 may depend on θ. If it is this case, then we won't be able to *compute* the value of T_3 by only looking at the sample values $X_1, \cdots, X_n$ (e.g., consider $T_3 = \bar{X} + \theta$). This is problematic, since a 'statistic' should be computable just from the sample values to give insight into θ.

Exercise 5.3.3. Let $X_1, \cdots, X_n$ be i.i.d. Poisson(λ) RVs.

(i) Show that X_1 is an unbiased estimator for λ.

(ii) Show that $T = X_1 + \cdots + X_n$ is a sufficient statistic for λ.

(iii) (Rao-Blackwellization) Show that

$$\mathbb{E}[X_1 \,|\, T] = T/n = \bar{X}. \tag{213}$$

Deduce that $\bar{X}$ is an unbiased estimator for θ with $\mathrm{Var}(\bar{X}) \le \lambda$ from Rao-Blackwell theorem. (In fact, $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_1)/n^2 = \lambda/n^2$.)

CHAPTER 6

Bayesian Estimation

In this section, we discuss another widely used method of parameter estimation, called the *Bayesian estimation*. Just like in MLE, we would like to estimate an unkonwn parameter θ in a distribution $f_{X;\theta}$ from analyzing an i.i.d. sample $X_1, \cdots, X_n$ drawn from that distribution. One of the key aspect in Bayesian estimation is to quantify our prior knowledge on the unknown parameter θ as yet another probabilistic model, which we call as *belief*. After observing a new *data* $\{X_1 = x_1, \cdots, X_n = x_n\}$, our current best belief on θ (*prior distribution*) will be updated to a new belief (posterior distribution) according to Bayes' Theorem. Since we are making some additional modeling assumption on θ, we will be able to draw more quantitative inference on θ using Bayesian estimation.

One of the early bottleneck in Bayesian estimation was the computational overhaed in computing posterior distributions. Due to the recent explosion in our computational capacity, coupled with advanced sampling techniques such as MCMC, Bayesian methods have become one of the indispensible tool in modern statistics and machine learning.

1. Bayes' Theorem and Inference – Discrete setting

If we have a model for bitcoin price, then we can tune the parameters accordingly and attempt to predict its future prices. But how do we tune the parameters? Say the bitcoin price suddenly dropped by 20% overnight (with no surprise). Which factor is the most likely to have caused this drop? In general, Bayes' Theorem can be used to infer likely factors when we are given the effect or outcome. This is all based on conditional probability and partitioning.

We begin by the following trivial but important observation:

$$\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B). \tag{214}$$

And this is all we need.

Theorem 6.1.1 (Bayes' Theorem)**.** *Let $(\Omega, \mathbb{P})$ be a probability space, and let $A_1, \cdots, A_k \subseteq \Omega$ be events of positive probability that form a partition of Ω. Then*

$$\mathbb{P}(A_1 \,|\, B) = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1) + \mathbb{P}(B \,|\, A_2)\mathbb{P}(A_2) + \cdots + \mathbb{P}(B \,|\, A_k)\mathbb{P}(A_k)}. \tag{215}$$

PROOF. Note that (224) yields the first equality. Then partitioning and rewrite $\mathbb{P}(B)$ gives the second equality. □

What's so important about the Bayes' theorem is its interpretation as a means of *inference*, which is one of the fundamental tool in modern machine learning.

Example 6.1.2. Suppose Bob has a coin with unknown probability of heads, which we denote by Θ. This is called the *parameter*. Suppose Alice knows that Bob has one of the three kinds of coins A, B, and C, with probability of heads being $p_A = 0.2$, $p_B = 0.5$, and $p_C = 0.8$, respectively. This piece of information is called the *model*. Since Alice has no information, she initially assumes that Bob has one of the three coins equally likely. Namely, she assumes the uniform distribution over the sample space $\Omega = \{0.2, 0.5, 0.8\}$. This knowledge is called *prior*.

Now Bob flips his coin 10 times and got 7 heads, and reports this information, which we call `Data`, to Alice. Now that Alice has more information, she needs to update her *prior* to *posterior*, which is the probability distribution on Ω that best explains the `Data`. Namely, it is the conditional probability distribution $\mathbb{P}(\Theta=\theta\,|\,\mathtt{Data})$.

First, using our prior, we compute $\mathbb{P}(\mathtt{Data})$, the probability of seeing this particular data at hand. This can be done by partitioning:

$$\begin{aligned}\mathbb{P}(\mathtt{Data}) &= \mathbb{P}(\mathtt{Data}\,|\,\Theta=0.2)\mathbb{P}(\Theta=0.2)+\mathbb{P}(\mathtt{Data}\,|\,\Theta=0.5)\mathbb{P}(\Theta=0.5) && (216)\\ &\qquad +\mathbb{P}(\mathtt{Data}\,|\,\Theta=0.8)\mathbb{P}(\Theta=0.8) && (217)\\ &= \mathbb{P}(7/10\text{ heads}\,|\,\Theta=0.2)\frac{1}{3}+\mathbb{P}(7/10\text{ heads}\,|\,\Theta=0.5)\frac{1}{3}+\mathbb{P}(7/10\text{ heads}\,|\,\Theta=0.8)\frac{1}{3} && (218)\\ &= \frac{1}{3}\left(\binom{10}{7}(0.2)^7(0.8)^3+\binom{10}{7}(0.5)^7(0.5)^3+\binom{10}{7}(0.8)^7(0.2)^3\right)\approx 0.1064, && (219)\end{aligned}$$

where $\binom{10}{7}=120$ is the number of ways to choose 7 out of 10 objects.

Second, we reformulate the first equality of Theorem 6.2.1 as

$$\mathbb{P}(\Theta\,|\,\mathtt{Data})=\frac{\mathbb{P}(\mathtt{Data}\,|\,\Theta)\mathbb{P}(\Theta)}{\mathbb{P}(\mathtt{Data})}. \tag{220}$$

Hence we can compute the posterior distribution by

$$\mathbb{P}(\Theta=0.2\,|\,\mathtt{Data})=\frac{\mathbb{P}(\mathtt{Data}\,|\,\Theta=0.2)\mathbb{P}(\Theta=0.2)}{\mathbb{P}(\mathtt{Data})}=\frac{\binom{10}{7}(0.2)^7(0.8)^3\frac{1}{3}}{0.1064}\approx 0.0025 \tag{221}$$

$$\mathbb{P}(\Theta=0.5\,|\,\mathtt{Data})=\frac{\mathbb{P}(\mathtt{Data}\,|\,\Theta=0.5)\mathbb{P}(\Theta=0.5)}{\mathbb{P}(\mathtt{Data})}=\frac{\binom{10}{7}(0.5)^7(0.5)^3\frac{1}{3}}{0.1064}\approx 0.3670 \tag{222}$$

$$\mathbb{P}(\Theta=0.8\,|\,\mathtt{Data})=\frac{\mathbb{P}(\mathtt{Data}\,|\,\Theta=0.8)\mathbb{P}(\Theta=0.8)}{\mathbb{P}(\mathtt{Data})}=\frac{\binom{10}{7}(0.8)^7(0.2)^3\frac{1}{3}}{0.1064}\approx 0.6305. \tag{223}$$

Note that according to the posterior distribution, $\Theta=0.8$ is the most likely value, which is natural given that we have 7 heads out of 10 flips. However, our knowledge is always incomplete so our posterior knowledge is still a probability distribution on the sample space.

What if Bob flips his coin another 10 times and reports only 3 heads to Alice? Then she will have to use her current prior $\pi=[0.0025,0.3060,0.6305]$ (which was obtained as the posterior in the previous round) to compute yet another posterior using the new data. This will be likely to give higher weight to $\Theta=0.2$. ▲

Exercise 6.1.3. Suppose we have a prior distribution $\pi=[0.0025,0.3680,0.6305]$ on the sample space $\Omega=\{0.2,0.5,0.8\}$ for the inference problem of unknown parameter Θ. Suppose we are given the data that ten independent flips of probability Θ coin comes up heads twice. Compute the posterior distribution using this data and Bayesian inference.

Exercise 6.1.4. A test for pancreatic cancer is assumed to be correct %95 of the time: if a person has the cancer, the test results in positive with probability 0.95, and if the person does not have the cancer, then the test results in negative with probability 0.95. From a recent medical research, it is known that only %0.05 of the population have pancreatic cancer. Given that the person just tested positive, what is the probability of having the cancer?

2. Bayes' Theorem and Inference – Continuous setting

We begin by recalling the following basic observation, from which Bayes' theorem can be easily derived:

$$\mathbb{P}(B|A)\mathbb{P}(A)=\mathbb{P}(A\cap B)=\mathbb{P}(A|B)\mathbb{P}(B). \tag{224}$$

In case of the events involving continuous random variables taking a certain value, we interpret the above probabilities as probability densities. The following is a random variable version of Bayes' Theorem discussed in the previous subsection.

Theorem 6.2.1 (Bayes' Theorem)**.** *Let X and Θ be random variables. Then*

$$\mathbb{P}(\Theta=\theta\,|\,X=x)=\frac{\mathbb{P}(X=x\,|\,\Theta=\theta)\mathbb{P}(\Theta=\theta)}{\mathbb{P}(X=x)}. \tag{225}$$

PROOF. Follows from (224). □

Remark 6.2.2. Depending on whether X and Θ are discrete or continuous, we interpret the probabilities in the above theorem as PDF or PMF, accordingly.

Bayesian inference is usually carried out in the following procedure.

Bayesian inference:

(i) To explain an observable $\mathbf{x}$, we choose a *probabilistic model* $p(x\,|\,\theta)$, which is a probability distribution on the possible values x of $\mathbf{x}$, depending on a parameter θ.

(ii) Choose a probability distribution $\pi(\theta)$, called the *prior distribution*, that expresses our beliefs about a parameter θ to the best of our current knowledge.

(iii) After observing data $\mathcal{D}=\{x_1,\cdots,x_n\}$, we update our beliefs and compute the *posterior distribution* $p(\theta\,|\,\mathcal{D})$ according to Bayes' Theorem.

When we generate data $\mathcal{D}=\{x_1,\cdots,x_n\}$, we assume that each x_i is obtained by an independent sample X_i of the observable $\mathbf{x}$ from the true distribution of $\mathbf{x}$. According to Bayes' Theorem, we can compute the posterior distribution as

$$\mathbb{P}(\Theta=\theta\,|\,X_1=x_1,\cdots,X_n=x_n)=\frac{\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta=\theta)\mathbb{P}(\Theta=\theta)}{\mathbb{P}(X_1=x_1,\cdots,X_n=x_n)}, \tag{226}$$

or in a more compact form,

$$p(\theta\,|\,x_1,\cdots,x_n)=\frac{p(x_1,\cdots,x_n\,|\,\theta)\pi(\theta)}{p(x_1,\cdots,x_n)}. \tag{227}$$

By a slight abuse of notation, we use lowercase p to denote either PMF or PDF, depending on the context.

The conditional probability $p(x_1,\cdots,x_n\,|\,\theta)$ is called the *likelyhood function*, which is the probability of obtaining independent data sample $\mathcal{D}=\{x_1,\cdots,x_n\}$ according to our probability model $p(x\,|\,\theta)$ assuming model parameter $\Theta=\theta$. By the independence between samples, the likelihood function factors into the product of marginal likelihood of each sample:

$$\begin{aligned} p(x_1,\cdots,x_n\,|\,\theta)&=\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta=\theta) && (228)\\ &=\prod_{i=1}^{n}\mathbb{P}(X_i=x_i\,|\,\Theta=\theta)=\prod_{i=1}^{n}p(x_i\,|\,\theta). && (229)\end{aligned}$$

On the other hand, the joint probability $\mathbb{P}(X_1=x_1,\cdots,X_n=x_n)$ of obtaining the data sample $\mathcal{D}=\{x_1,\cdots,x_n\}$ from our probability model can be computed by conditioning on the values of model parameter Θ:

$$\begin{aligned} \mathbb{P}(X_1=x_1,\cdots,X_n=x_n)&=\mathbb{E}\left[\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta)\,\middle|\,\Theta\right] && (230)\\ &=\int_{-\infty}^{\infty}p(x_1,\cdots,x_n\,|\,\theta)\pi(\theta)\,d\theta. && (231)\end{aligned}$$

Example 6.2.3 (Signal detection)**.** A binary signal $X\in\{-1,1\}$ is transmitted, and we are given that

$$\mathbb{P}(X=1)=p,\qquad \mathbb{P}(X=-1)=1-p \tag{232}$$

for some $p \in [0,1]$, as the prior distribution of X. There is a white noise in the channel so we receive a perturbed signal

$$Y = X + Z, \tag{233}$$

where $Z \sim N(0,1)$ is a standard normal variable.

Conditional on $Y = y$, what is the probability that the actual signal X is 1? In order to use Bayes theorem, first observe that $Y|X = x \sim N(x,1)$. Hence

$$f_{Y|X=x}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}}. \tag{234}$$

So by Bayes Theorem,

$$\mathbb{P}(X = 1 \,|\, Y = y) = \frac{f_{Y|X=1}(y)\mathbb{P}(X=1)}{f_Y(y)} \tag{235}$$

$$= \frac{\frac{p}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}}{\frac{p}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} + \frac{1-p}{\sqrt{2\pi}} e^{-\frac{(y+1)^2}{2}}} \tag{236}$$

Thus we get

$$\mathbb{P}(X = 1 \,|\, Y = y) = \frac{pe^y}{pe^y + (1-p)e^{-y}}, \tag{237}$$

and similarly,

$$\mathbb{P}(X = -1 \,|\, Y = y) = \frac{(1-p)e^{-y}}{pe^y + (1-p)e^{-y}}, \tag{238}$$

This is our posterior distribution of X after observing $Y = y$. ▲

Example 6.2.4 (Bernoulli model and uniform prior)**.** Bob has a coin with unknown probability Θ of heads. Alice has no information whatsoever, so her prior distribution π for Θ is the uniform distribution Uniform$([0,1])$. Bob flips his coin independently n times, and let $X_1, \cdots, X_n$ be the outcome, X_i's are i.i.d. Bernoulli(Θ) variables. Let x_i be the observed value of X_i, and let $s_n = x_1 + \cdots + x_n$ be the number of heads in the n flips. Given data $\mathscr{D} = \{x_1, \cdots, x_n\}$, Alice wants to compute her posterior distribution on Θ.

The the likelyhood function is given by

$$p(x_1, \cdots, x_n \,|\, \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \tag{239}$$

$$= \theta^{s_n} (1-\theta)^{n-s_n}, \tag{240}$$

where we denote $s_n = x_1 + \cdots + x_n$. Since $\pi \equiv 1$, the posterior distribution is given by

$$p(\theta \,|\, x_1, \cdots, x_n) = \frac{\theta^{s_n}(1-\theta)^{n-s_n}}{p(x_1, \cdots, x_n)}, \tag{241}$$

Note that

$$p(x_1, \cdots, x_n) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\, d\theta = \frac{1}{\binom{n}{s_n}(n+1)}, \tag{242}$$

where the last equality follows from Exercise 6.2.6. Hence

$$p(\theta \,|\, x_1, \cdots, x_n) = \binom{n}{s_n}(n+1)\theta^{s_n}(1-\theta)^{n-s_n} \tag{243}$$

$$= \frac{(n+1)!}{s_n!(n-s_n)!} \theta^{(s_n+1)-1} (1-\theta)^{(n-s_n+1)-1}. \tag{244}$$

Hence, according to Exercise 6.2.7, we can write

$$\Theta \,|\, x_1, \cdots, x_n \sim \text{Beta}(s_n + 1, n - s_n + 1). \tag{245}$$

▲

Exercise 6.2.5. Bob has a coin with unknown probability Θ of heads. Alice has the following Beta prior (See Exercise 6.2.7 for the definition of Beta distribution):

$$\pi = \text{Beta}(\alpha, \beta). \tag{246}$$

Suppose that Bob gives Alice the data $\mathscr{D}_n = \{x_1, \cdots, x_n\}$, which is the outcome of n independent coin flips. Denote $s_n = x_1 + \cdots + x_n$. Show that Alice's posterior distribution is $\text{Beta}(\alpha + s_n, \beta + n - s_n)$. Namely,

$$\Theta \,|\, \mathscr{D}_n \sim \text{Beta}(\alpha + s_n, \beta + n - s_n). \tag{247}$$

Exercise 6.2.6. Let $Y \sim \text{Uniform}([0,1])$ and $X \sim \text{Binomial}(n, Y)$ be independent RVs.

(i) Use iterated expectation for probability to write

$$\mathbb{P}(X = k) = \binom{n}{k} \int_0^1 y^k (1-y)^{n-k}\, dy. \tag{248}$$

(ii) Write $A_{n,k} = \int_0^1 y^k (1-y)^{n-k}\, dy$. Use integration by parts and show that

$$A_{n,k} = \frac{k}{n-k+1} A_{n,k-1}. \tag{249}$$

for all $1 \le k \le n$. Conclude that for all $0 \le k \le n$,

$$A_{n,k} = \frac{1}{\binom{n}{k}} \frac{1}{n+1}. \tag{250}$$

(iii) Conclude that $X \sim \text{Uniform}(\{0, 1, \cdots, n\})$.

Exercise 6.2.7 (Beta distribution)**.** A random varible X taking values from $[0,1]$ has Beta distribution of parameters α and β, which we denote by $\text{Beta}(\alpha, \beta)$, if it has PDF

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \tag{251}$$

where $\Gamma(z)$ is the Euler Gamma function defined by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx. \tag{252}$$

(i) Use integration by parts to show the following recursion

$$\Gamma(z+1) = z\Gamma(z). \tag{253}$$

Deduce that $\Gamma(n) = (n-1)!$ for all integers $n \ge 1$.

(ii) Let $X \sim \text{Beta}(k+1, n-k+1)$. Use (i) to show that

$$f_X(x) = \frac{n!(n+1)}{k!(n-k)!} x^k (1-x)^{n-k} = \frac{x^k (1-x)^{n-k}}{1/\binom{n}{k}(n+1)}. \tag{254}$$

Use Exercise 6.2.6 to verify that the above function is indeed a PDF (i.e., it integrates to 1).

(iii)* Show that if $X \sim \text{Beta}(\alpha, \beta)$, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{255}$$

3. More Bayesian concepts

Example 6.3.1 (Bayesian estimator)**.** Recall Exercise 6.2.5, where we want to infer an unknown parameter θ in Bernoulli(θ), starting from the beta prior $\pi = \text{Beta}(\alpha, \beta)$. After observing a data $\mathscr{D} = (x_1, \cdots, x_n)$ consisting of $s_n = \sum_{i=1}^n x_i$ successes, our posterior distribution is Beta distribution with parameters $s_n + 1$ and $\beta + n - s_n$. In other words,

$$\Theta | \mathscr{D} \sim \text{Beta}(\alpha + s_n, \beta + n - s_n). \tag{256}$$

Now suppose we want to get a point estimator $\hat{\theta}$ for the unknown parameter from our posterior distribution. Any choice of estimator will result in some kind of error, so we would like our estimator to be minimizing an error function of choice. A standard choice is the mean squared error (MSE), which in our case would be under conditioning on the observed data. That is, we want our estimator $\hat{\theta}$ to be such that

$$\hat{\theta} = \operatorname{argmin} \mathbb{E}[(\Theta - \hat{\theta})^2 \,|\, \mathscr{D}]. \tag{257}$$

According to Exercise 4.1.2, the above MSE will be minimized when $\hat{\theta} = \mathbb{E}[\Theta \,|\, \mathscr{D}]$, and the minimum MSE is the conditional variance $\text{Var}(\Theta \,|\, \mathscr{D})$. Noting the mean of beta distribution in Exercise 6.2.7, we obtain the following Bayesian estimator

$$\hat{\theta} = \mathbb{E}[\Theta \,|\, \mathscr{D}] = \mathbb{E}[\text{Beta}(\alpha + s_n, \beta + n - s_n)] = \frac{\alpha + s_n}{\alpha + \beta + n}. \tag{258}$$

▲

Example 6.3.2 (Excerpted from [HTZ77], Berry 1996)**.** We are concerned with breakage of glass panels in high-rise buildings. One such case involved 39 panels, and of the 39 panels that broke, it was known that 3 broke due to nickel sulfide (NiS) stones found in them. Loss of evidence prevented the causes of breakage ofthe other 36 panels from being known. So the court wanted to know whether the manufacturer of the panels or the builder was at fault for the breakage of these 36 panels. In other words, we would like to know the probability θ of a glass panel being broken due to NiS stones.

We are going to use Bayesian estimation using beta prior $\pi = \text{Beta}(\alpha, \beta)$ with parameters α and β that we are going to determine using some external knowledge. From expert testimony, it was thought that usually about 5% breakage is caused by NiS stones. Hence we should have

$$0.05 = \mathbb{E}_\pi[\theta] = \mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}. \tag{259}$$

Moreover, the expert thought that if two panels from the same lot break and one breakage was caused by NiS stones, then, due to the pervasive nature of the manufacturing process, the probability of the second panel breaking due to NiS stones increases to about 95%. Hence after observing one data sample of broken glass panel, our posterior distribution is $\text{Beta}(\alpha + 1, \beta)$. Hence 95% should equal to the mean of this posterior distribution:

$$0.95 = \mathbb{E}[\Theta \,|\, \text{1st panel broken due to NiS}] = \mathbb{E}[\text{Beta}(\alpha + 1, \beta)] = \frac{\alpha + 1}{\alpha + \beta + 1}. \tag{260}$$

Solving these two equations, we find

$$\alpha = \frac{1}{360}, \qquad \beta = \frac{19}{360}. \tag{261}$$

Recall that we know that three glass panels were broken due to NiS. Then the posterior estimate $\hat{\theta}$ that the fourth glass panel is also broken due to NiS, given the the first three were so, is given by

$$\mathbb{E}[\Theta \,|\, \text{first three panels broken due to NiS}] = \mathbb{E}[\text{Beta}(\alpha + 3, \beta)] = \frac{\alpha + 3}{\alpha + \beta + 3}. \tag{262}$$

A similar calculation holds for the fourth panel, and so on. Hence by multiplying out the Bayesian estimators of conditional probabilities (predictive probabilities),

$$\mathbb{P}\left(\begin{matrix}\text{remaining 36 panels are}\\ \text{broken due to NiS}\end{matrix}\middle|\,\text{first 3 broken due to NiS}\right) \tag{263}$$

$$= \prod_{i=3}^{38} \mathbb{P}\left(\begin{matrix} i+1\text{th panels is}\\ \text{broken due to NiS}\end{matrix}\middle|\,\text{first i broken due to NiS}\right) \tag{264}$$

$$\approx \left(\frac{\alpha+3}{\alpha+\beta+3}\right)\left(\frac{\alpha+4}{\alpha+\beta+4}\right)\cdots\left(\frac{\alpha+38}{\alpha+\beta+38}\right) = 0.8664. \tag{265}$$

Hence according to our Bayesian estimation, the probability of all remaining 36 panels were broken due to NiS stones given the first three were so is about 87%, which is the needed value in the court's decision. ▲

CHAPTER 7

Interval estimation

1. Estimating mean of normal distribution with known variance

Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with finite mean μ and variance σ^2. We have seen many times that the sample mean $\bar{X}$ arises as a natural estimator for the population mean μ (from MLE and MoM). Recall that the sample mean is an unbiased estimator of μ:

$$\mathbb{E}[\bar{X}] = \mu. \tag{266}$$

Suppose after some experiment, we obtained a data of samples $\mathscr{D} = (x_1, \cdots, x_n)$. From this we will estimate $\mu \approx \bar{x}$. But what do we really mean by this approximation here? Can we say more quantitative about the location of μ? Namely, we want to make statements like

$$\text{“}\mu \text{ is contained in the interval } [\bar{x} - \varepsilon, \bar{x} + \varepsilon] \text{ with confidence at least } \alpha\text{”}. \tag{267}$$

More precisely, this may be rephrased as

$$\mathbb{P}\left(|\mu - \bar{x}| \leq \varepsilon\right) \geq \alpha. \tag{268}$$

However, what is really random in the above probability? μ is an unknown parameter, and $\bar{x}$ is an observed sample mean. So what we really mean here is the following:

$$\mathbb{P}\left(|\mu - \bar{X}| \leq \varepsilon\right) \geq \alpha. \tag{269}$$

Namely, if we randomly sample the RVs $X_1, \cdots, X_n$ and take their sample mean $\bar{X}$ (which is still random), then this random estimator $\bar{X}$ should be close to the unknown parameter μ with some large probability.

In order to compute the above probability in the left hand side, we need to know the distribution of the random sample mena $\bar{X}$. In this subsection, we first consider a special case where we know the distribution of $\bar{X}$ exactly. Our derivation of what is called ‘confidence interval’ in this section relies on the following two assumptions:

A1. We have i.i.d. samples $X_1, \cdots, X_n$ from normal distribution $N(\mu, \sigma^2)$.
A2. We know the population variance $\mathrm{Var}(X) = \sigma^2$.

Example 7.1.1 (Confidence interval for the mean of normal distribution)**.** Let $X_1, \cdots, X_n$ be i.i.d. RVs with distribution $N(\mu, \sigma^2)$. Suppose we know the variance σ^2, but not the mean μ. First note that

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mathbb{E}[X_1] = \mu \tag{270}$$

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\mathrm{Var}(X_1) = n\sigma^2, \tag{271}$$

where the first uses linearity of expectation and the second uses independence between X_i’s. Moreover, recalling that normal distribution is additive, we have the precise distribution of the sum:

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2). \tag{272}$$

Since $\mathbb{E}[\bar{X}] = \mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(\bar{X}) = n^{-2}\mathrm{Var}(X_1 + \cdots + X_n) = \sigma^2/n$, it follows that

$$\bar{X} \sim N(\mu, \sigma^2/n). \tag{273}$$

From this we deduce

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{274}$$

From this we can construct a *confidence interval* for μ. Namely, recall that the 'z-score' for 'confidence' $\alpha \in [0,1]$, denoted by z_α, is defined so that $\mathbb{P}(N(0,1) > z_\alpha) = \alpha$. Hence we have

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha. \tag{275}$$

Rewriting, this gives

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{276}$$

That is, the probability that the random interval $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ containing the population mean μ is $1-\alpha$. In this sense, this random interval is called the $100(1-\alpha)\%$ *confidence interval* for μ. For instance, noting that $z_{0.05/2} = 1.96$, $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is a 95% confidence interval for μ. ▲

Exercise 7.1.2 (Excerpted from [HTZ77])**.** Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{x} = 1478$ hours, what is the corresponding 95% confidence interval for μ?

2. Central limit theorem

In this subsection, we handle constructing confidence interval for population mean μ for the case of unknown population distribution but known population variance σ^2. The main tool we use is the celebrated and one of the most important result in probability and statistics – the Central Limit Theorem.

As before, let $(X_t)_{t\ge 0}$ be a sequence of i.i.d. RVs with finite mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for $n \ge 1$. We have calculated the mean and variance of the sample mean S_n/n:

$$\mathbb{E}[S_n/n] = \mu, \quad \mathrm{Var}(S_n/n) = \sigma^2/n. \tag{277}$$

Since $\mathrm{Var}(S_n/n) \to 0$ as $n \to \infty$, we expect the sequence of RVs S_n/n to converge its mean μ in probability. This is the famous laws of large numbers.

Central limit theorem is a limit theorem for the sample mean with different regime, namely, it describes the 'fluctuation' of the sample mean around its expectation, as $n \to \infty$. For this purpose, we need to standardize the sample mean so that the mean is zero and variance is unit. Namely, let

$$Z_n = \frac{S_n/n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \tag{278}$$

so that

$$\mathbb{E}[Z_n] = 0, \quad \mathrm{Var}(Z_n) = 1. \tag{279}$$

Since the variance is kept at 1, we should not expect the sequence of RVs $(Z_n)_{n\ge 0}$ converge to some constant in probability, as in the law of large number situation. Instead, Z_n should converge to some other RV, if it ever converges in some sense. Central limit theorem states that Z_n becomes more and more likely to be a standard normal RV $Z \sim N(0,1)$.

Let us state the central limit theorem.

Theorem 7.2.1 (CLT)**.** *Let $(X_k)_{k\ge 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \ge 1$. Suppose $\mathbb{E}[X_1] < \infty$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}. \tag{280}$$

Then Z_n converges to Z as $n \to \infty$ in distribution, namely,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \mathbb{P}(Z \le z). \tag{281}$$

As a typical application of CLT, we can approximate Binomial(n,p) variables by normal RVs.

Exercise 7.2.2. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - np)/\sqrt{np(1-p)}$. Use CLT to deduce that, as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim$ Binomial(n,p), then

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \Rightarrow Z \sim N(0,1). \tag{282}$$

(iii) From (ii), deduce that have the following approximation

$$\mathbb{P}(Y_n \leq x) \approx \mathbb{P}\left(Z \leq \frac{x - np}{\sqrt{np(1-p)}}\right), \tag{283}$$

which becomes more accurate as $n \to \infty$.

Example 7.2.3 (Confidence interval for the mean with known variance)**.** Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with unknown mean μ but known variance σ^2. Here we do not know if X_i's are drawn from a normal distribution. By CLT, we have

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Longrightarrow Z \sim N(0,1), \tag{284}$$

as the sample size $n \to \infty$. In other words, Z_n becomes more and more likely to be a standard normal RV, so for any $z \in \mathbb{R}$

$$\mathbb{P}(Z_n \leq z) \approx \mathbb{P}(Z \leq z) \tag{285}$$

when n is large enough.

From this we can construct a *confidence interval* for μ similarly as in Example 7.4.2. Namely,

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha. \tag{286}$$

Rewriting, this gives

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) \approx 1 - \alpha. \tag{287}$$

That is, the probability that the random interval $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ containing the population mean μ is approximately $1-\alpha$. In this sense, this random interval is called the $100(1-\alpha)\%$ *confidence interval* for μ. For instance, noting that $z_{0.05/2} = 1.96$, $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is a 95% confidence interval for μ. ▲

Exercise 7.2.4. Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of $\text{Var}(X) = 1296$. If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{x} = 1478$ hours, what is the corresponding 95% confidence interval for μ?

3. Confidence interval for the mean with unknown variance

In this subsection, we handle constructing confidence interval for population mean μ for the case of normal population distribution with unknown population variance σ^2.

Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with normal distribution $N(\mu, \sigma^2)$ with unknown finite mean μ and variance σ^2. Using our argument in Example 7.4.2, we can construct the following $\%100(1-\alpha)$ confidence interval

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha. \tag{288}$$

However, since we do not know σ, this does not really give us a valid confidence interval for μ.

Recall that the underlying observation for the above equation was that the standardized sample mean from normal distribution follows standard normal:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{289}$$

But since σ is unknown, a reasonable move here is to replace σ by some of its estimator. A natural choice would be the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{290}$$

which we have shown to be an unbiased estimator of σ^2 in Exercise 3.2.4. So what happens if we replace σ by S in (289)? It turns out that this substitution changes the distribution from the standard normal $N(0,1)$ to something that is known as the 't-distribution with $n-1$ degrees of freedom'. This is based on chi-square distribution.

Definition 7.3.1 (chi-square distribution)**.** Let $Z_1,\cdots,Z_k$ be i.i.d. standard normal RVs. Let $V = Z_1+\cdots+Z_k$. Then the distribution of X is called the *chi-square distribution* of k degrees of freedom, which we denote by $V \sim \chi^2(k)$.

Definition 7.3.2 (t-distribution)**.** Let $Z \sim N(0,1)$ and $V \sim \chi^2(k)$ be independent. Let $T = Z/\sqrt{V/k}$. Then the distribution of T is called the *t-distribution* of k degrees of freedom, which we denote by $T \sim t(k)$.

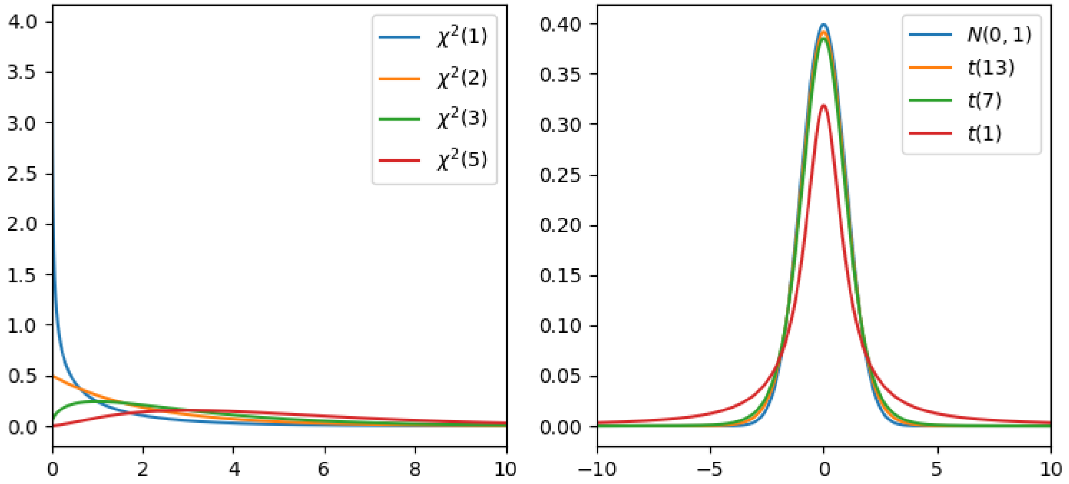


FIGURE 1. PDF of χ^2- and t-distributions for some values of degrees of freedom.

Exercise 7.3.3. **(i)** Let $Z \sim N(0,1)$. Show that $Z^2 \sim \chi^2(1)$ has the following MGF

$$\mathbb{E}[e^{tZ^2}] = \int_0^\infty e^{tx^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx = \int_0^\infty \frac{1}{\sqrt{2\pi}}e^{-(1-2t)x^2/2}\,dx = \frac{1}{\sqrt{1-2t}}. \tag{291}$$

(ii) Let $V \sim \xi^2(k)$. Use (i) to deduce that

$$\mathbb{E}[e^{tV}] = (1-2t)^{-k/2}. \tag{292}$$

(iii) Let $V \sim \chi^2(k)$. Show that V has the following PDF

$$f_V(x) = \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(k/2)}\mathbf{1}(x \ge 0), \tag{293}$$

where $\Gamma(\cot)$ is the Gamma function. (Hint: Compute the MGF of the above PDF, and show that it equals to the MGF of $\xi^2(k)$.)

Exercise 7.3.4. (Optional) Let $T \sim t(k)$. Show that T has the following PDF

$$f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\Gamma\left(\frac{k}{2}\right)} \left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \tag{294}$$

where $\Gamma(\cdot)$ is the Gamma function.

The following is the theoretical background in this subsection, whose proof (optional) is provided at the end of this subsection.

Proposition 7.3.5. *Let $X_1, \cdots, X_n$ be i.i.d samples from $N(\mu, \sigma^2)$. Then*

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \tag{295}$$

Using this result, we can now construct confidence interval for the case of normal distribution with unknown variance:

Example 7.3.6 (Confidence interval for the mean of normal with unknown variance)**.** Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with distribution $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. By Proposition 7.3.5, we have

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \tag{296}$$

From this we can construct a *confidence interval* for μ similarly as in Example 7.4.2. Namely, denote that the 't-score' $t_\alpha(k)$ by $\mathbb{P}(T \geq t_\alpha(k)) = \alpha$, where $T \sim t(k)$. Then

$$\mathbb{P}\left(-t_{\alpha/2}(n-1) \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right) = 1-\alpha. \tag{297}$$

Rewriting, this gives

$$\mathbb{P}\left(\mu \in \left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{298}$$

That is, the probability that the random interval $\left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$ containing the population mean μ is $1-\alpha$. In this sense, this random interval is called the $100(1-\alpha)\%$ *confidence interval* for μ. For instance, noting that $t_{0.05}(19) = 1.729$, a 90% confidence interval for μ would be $\left[\bar{x} - 1.729\frac{s}{\sqrt{n}}, \bar{x} + 1.729\frac{s}{\sqrt{n}}\right]$. ▲

Exercise 7.3.7 (Excerpted from [HTZ77])**.** Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of X is $N(\mu, \sigma^2)$. To estimate μ, a farmer measured the butter fat production for $n = 20$ cows and obtained the following data:

481	537	513	583	453	510	570	500	457	555
618	327	350	643	499	421	505	637	599	392

(i) Give a point estimate of μ.
(ii) Give a 90% confidence interval for μ.

Exercise 7.3.8 (Excerpted from [HTZ77])**.** An automotive supplier of interior parts places several electrical wires in a harness. A pull test measures the force required to pull spliced wires apart. A customer requires that each wire spliced into the harness must with-stand a pull force of 20 pounds. Let X equal the pull force required to pull 20 gauge wires apart. Assume that $X \sim N(\mu, \sigma^2)$. The following data give 20 observations of X:

28.8	24.4	30.1	25.6	26.4	23.9	22.1	22.5	27.6	28.1	(299)

$$20.8 \quad 27.7 \quad 24.4 \quad 25.1 \quad 24.6 \quad 26.3 \quad 28.2 \quad 22.2 \quad 26.3 \quad 24.4 \tag{300}$$

(i) Find point estimates for μ and σ.

(ii) Find a 99% one-sided confidence interval for μ that provides a lower bound for μ. That is, first find the smallest constant c such that

$$\mathbb{P}\left(\bar{X} - c\frac{S}{\sqrt{n}} \le \mu\right) \ge 0.99. \tag{301}$$

Then obtain the desired one-sided confidence interval for μ by computing $\bar{x}$ and s from the given sample.

Proof of Proposition 7.3.5. (Optional*) We first write

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)\left(\frac{1}{S/\sigma}\right). \tag{302}$$

Since X_i's are i.i.d. with distribution $N(\mu,\sigma^2)$, we know that the standardized sample mean $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ has distribution $N(0,1)$. Hence by definition, it suffices to show that $(n-1)(S/\sigma)^2 \sim \chi^2(n-1)$.

Now we show that $(n-1)(S/\sigma) \sim \chi^2(n-1)$. First by using Exercise 3.2.4 (iii), we write

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) + n(\bar{X}-\mu)^2 \tag{303}$$

$$= (n-1)S^2 + n(\bar{X}-\mu)^2. \tag{304}$$

Divide both sides by σ^2 to get

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = (n-1)(S/\sigma)^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2. \tag{305}$$

Notice that the left hand side has distribution $\xi^2(n)$, and the last term in the right hand side has distribution $\xi^2(1)$. We claim that the two RVs in the right hand side are independent. The assertion then follows by taking MGFs, since

$$\text{MGF}(\xi^2(n)) = \text{MGF}((n-1)(S/\sigma)^2)\text{MGF}(\xi^2(1)) \tag{306}$$

so that

$$\text{MGF}((n-1)(S/\sigma)^2) = \frac{\text{MGF}(\xi^2(n))}{\text{MGF}(\xi^2(1))} = \frac{[\text{MGF}(\xi^2(1))]^n}{\text{MGF}(\xi^2(1))} = [\text{MGF}(\xi^2(1))]^{n-1} = \text{MGF}(\xi^2(n-1)). \tag{307}$$

It remains to show that $\sum_{i=1}^{n}(X_i-\bar{X})^2$ and $n(\bar{X}-\mu)^2$ are independent. For this, it is enough to show that the RV $\bar{X}$ is independent of the random vector $\mathbf{X} := (X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X})$. We will do this by using joint MGF and factorization theorem. Note that

$$\text{MGF}_{\bar{X},\mathbf{X}}(s,t_1,\cdots,t_n) = \mathbb{E}\left[\exp\left(s\bar{X} + t_1(X_1-\bar{X}) + \cdots + t_n(X_n-\bar{X})\right)\right] \tag{308}$$

$$= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} t_iX_i + \left(s - \sum_{i=1}^{n} t_i\right)\bar{X}\right)\right] \tag{309}$$

$$= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} t_iX_i + \left((s/n) - \bar{t}\right)\sum_{i=1}^{n} X_i\right)\right] \tag{310}$$

$$= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n}(t_i - \bar{t} + (s/n))X_i\right)\right], \tag{311}$$

where we have denoted $\bar{t} = (t_1+\cdots+t_n)/n$. Now since X_i's are i.i.d. with $N(\mu,\sigma^2)$ distribution, we have

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{n}(t_i-\bar{t}+(s/n))X_i\right)\right] = \prod_{i=1}^{n}\text{MGF}_{X_i}(t_i-\bar{t}+(s/n)) \tag{312}$$

$$= \prod_{i=1}^{n} \exp\left(\mu(t_i - \bar{t} + (s/n)) + \frac{\sigma^2}{2}(t_i - \bar{t} + (s/n))^2\right) \tag{313}$$

$$= \exp\left(\sum_{i=1}^{n} \mu(t_i - \bar{t} + (s/n)) + \frac{\sigma^2}{2}(t_i - \bar{t} + (s/n))^2\right) \tag{314}$$

$$= \exp\left(s\mu + \frac{\sigma^2}{2}\left(\sum_{i=1}^{n}(t_i - \bar{t})^2 + \frac{s^2}{n}\right)\right) \tag{315}$$

$$= \exp\left(s\mu + \frac{\sigma^2 s^2}{2n}\right)\exp\left(\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2\right). \tag{316}$$

This shows that the joint MGF of $\bar{X}$ and $\mathbf{X}$ factorizes, which implies their independence. This shows the assertion. □

4. Confidence intervals of difference of two means

In this section, we study how to statistically compare two unknown RVs X and Y by sampling and making confidence intervals. We start with the ideal situation when both X and Y are normal with known variances.

Example 7.4.1 (Confidence interval for the difference of two means I)**.** Let $X \sim N(\mu_X, \sigma_Y^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, where both σ_X and σ_Y are known, but the means are unknown. We want to know whether $\mu_X \neq \mu_Y$. To test this, draw i.i.d. samples $X_1, \cdots, X_n$ for X and $Y_1, \cdots, Y_m$ for Y. Recall that

$$\bar{X} \sim N(\mu_X, \sigma_X^2/n), \qquad \bar{Y} \sim N(\mu_X, \sigma_Y^2/m). \tag{317}$$

If it were the case that $\mu_X \neq \mu_Y$, then it should be that $\bar{X} \neq \bar{Y}$. To make this quantitative, we consider the RV $\bar{X} - \bar{Y}$. Note that

$$\mathbb{E}[\bar{X} - \bar{Y}] = \mu_X - \mu_Y, \qquad \mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}. \tag{318}$$

We make a further assumption that the X_i's and Y_j's are also independent. Then $\bar{X}$ and $\bar{Y}$ are also independent, so

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right). \tag{319}$$

Thus if we standardize $\bar{X} - \bar{Y}$ and make a statistic Z, we get

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1). \tag{320}$$

From this we can construct a confidence interval for $\mu_X - \mu_Y$. Namely, (441) implies

$$\mathbb{P}\left(-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq (\bar{X} - \bar{Y}) - (\mu_X - \mu_Y) \leq z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right) = 1 - \alpha. \tag{321}$$

Rewriting, this gives

$$\mathbb{P}\left(\mu_X - \mu_Y \in \left[(\bar{X} - \bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{X} - \bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right]\right) = 1 - \alpha. \tag{322}$$

That is, the probability that the random interval in the above probablity containing the difference $\mu_X - \mu_Y$ of the population means is $1 - \alpha$. In this sense, this random interval is called the $100(1-\alpha)\%$ *confidence interval* for $\mu_X - \mu_Y$.

For instance, let $n=15$,$m=8$, $\bar{x}=70.1$,$\bar{y}=75.3$, $\sigma_X^2=60$, $\sigma_Y^2=40$, and $1-\alpha$=0.90. Note that $1-\alpha/2=0.95=\mathbb{P}(Z\le 1.645)$ for $Z\sim N(0,1)$. Hence $z_{\alpha/2}=1.645$, so

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}=1.645\sqrt{\frac{60}{15}+\frac{40}{8}}=1.645\cdot 3=4.935. \tag{323}$$

Noting that $\bar{x}-\bar{y}=-5.2$, we conclude that $[-5.2-4.935,-5.2+4.935]=[-10.135,-0.265]$ is a 90% confidence interval for $\mu_X-\mu_Y$. In particular, we can claim that, with at least 90% confidence, μ_Y is larger than μ_X by at least 0.265. ▲

Example 7.4.2 (Confidence interval for the difference of two means II)**.** Let X and Y be independent RVs such that $\mathbb{E}[X]=\mu_X$, $\mathrm{Var}(X)=\sigma_X^2$, $\mathbb{E}[Y]=\mu_Y$, and $\mathrm{Var}(Y)=\sigma_Y^2$. Suppose we do not know if X and Y have normal distribution but we know σ_X and σ_Y. Then by using CLT, all our confidence interval estimates in the previous example hold asymptotically. Namely, by CLT,

$$\bar{X}\Rightarrow N(\mu_X,\sigma_X^2/n), \qquad \bar{Y}\Rightarrow N(\mu_X,\sigma_Y^2/m), \tag{324}$$

as $n,m\to\infty$. As all samples are independent, this yields that, as $n,m\to\infty$,

$$\bar{X}-\bar{Y}\Rightarrow N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right). \tag{325}$$

Thus when n and m are large, $\bar{X}-\bar{Y}$ nearly follows normal distribution above. This implies the following $100(1-\alpha)$% (approximate) confidence interval for $\mu_X-\mu_Y$:

$$\mathbb{P}\left(\mu_X-\mu_Y\in\left[(\bar{X}-\bar{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},\,(\bar{X}-\bar{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right]\right)\approx 1-\alpha. \tag{326}$$

For instance, let $n=60$,$m=40$, $\bar{x}=70.1$,$\bar{y}=75.3$, $\sigma_X^2=60$, $\sigma_Y^2=40$, and $1-\alpha=0.90$. Note that $1-\alpha/2=0.95=\mathbb{P}(Z\le 1.645)$ for $Z\sim N(0,1)$. Hence $z_{\alpha/2}=1.645$, so

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}=1.645\sqrt{\frac{60}{60}+\frac{40}{40}}=1.645\cdot\sqrt{2}=2.326. \tag{327}$$

Noting that $\bar{x}-\bar{y}=-5.2$, we conclude that $[-5.2-2.326,-5.2+2.326]=[-7.526,-2.87]$ is a 90% confidence interval for $\mu_X-\mu_Y$. In particular, we can claim that, with at least 90% confidence, μ_Y is larger than μ_X by at least 2.87. ▲

Example 7.4.3 (Confidence interval for the difference of two means III)**.** Let $X_1,\cdots,X_n$ i.i.d. from $N(\mu_X,\sigma_Y^2)$ and $Y_1,\cdots,Y_m$ i.i.d. from $N(\mu_Y,\sigma_Y^2)$. Also assume that all such samples are independent of each other. In this discussion, we suppose that σ_X and σ_Y are both unknown.

As in the previous section for estimating a mean, we may try to replace population variances with sample variances. This would result in the following random interval

$$\left[(\bar{X}-\bar{Y})-z_{\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}},\,(\bar{X}-\bar{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right] \tag{328}$$

This may heuristically serve as $100(1-\alpha)$% (approximate) confidence interval for $\mu_X-\mu_Y$ at least when $n,m\gg 1$. However, this is not rigorously justified (as opposed to Example 7.3.6).

In fact, we can use t-distribution to handle this situation, at least when we further assume that $\sigma_X=\sigma_Y=\sigma$ (that is, unknown but equal variances). Namely, recall that

$$\frac{(n-1)S_X^2}{\sigma^2}\sim\chi^2(n-1), \qquad \frac{(m-1)S_Y^2}{\sigma^2}\sim\chi^2(m-1). \tag{329}$$

As all samples are independent, the two RVs above are also independent. It follows that their sum is a χ^2 RV with $(n-1)+(m-1)=n+m-2$ degrees of freedom:

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2). \tag{330}$$

Moreover, at the end of the proof of Proposition 7.3.5, we have shown that the sample mean and sample variance are independent. Hence if we define Z to be the standardization of the difference of sample means (see (441)), then U and Z are indpendent. Thus by definition of T distribution (see (7.3.2)), we have

$$T := \frac{Z}{\sqrt{U/(n+m-2)}} = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \frac{1}{\sqrt{\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}}\Big/\sqrt{n+m-2}} \tag{331}$$

$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}} \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2). \tag{332}$$

Denoting the *pulled estimator*

$$S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}, \tag{333}$$

this implies the following $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$:

$$\mathbb{P}\left(\mu_X-\mu_Y \in \left[(\bar{X}-\bar{Y})-t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}},\ (\bar{X}-\bar{Y})+t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]\right) = 1-\alpha. \tag{334}$$

For instance, suppose the problem of comparing the means of $N(\mu_X,\sigma^2)$ and $N(\mu_Y,\sigma^2)$ for unknown σ. Suppose we have $n=9$ samples from the first distribution with $\bar{x}=81.31$, $s_X^2=60.76$, and $m=15$ samples from the second distribution with $\bar{y}=78.61$, $s_Y^2=48.24$. The positive square root of the pulled estimator is given by

$$s_p = \sqrt{\frac{(9-1)(60.76)+(15-1)(48.24)}{9+15-2}} = 7.2658 \tag{335}$$

Noting that $t_{0.025}(22)=2.074$, we compute the endpoints of the corresponding 95% confidenceinterval for $\mu_X-\mu_Y$ as

$$(81.31-78.61) \pm (2.074)(8.4923)\sqrt{\frac{1}{9}+\frac{1}{15}} = -3.6538, 9.0538. \tag{336}$$

Since this interval contains 0, we cannot conclude that $\mu_X \neq \mu_Y$ with high confidence as before. ▲

Remark 7.4.4. What if we do not know population variances and also that they are equal? See the discussion at the end of [HTZ77, Sec. 7.2]. We give an alternative solution to this in Example 7.4.6.

Example 7.4.5 (Confidence interval for the difference of two means IV)**.** Suppose we are interested in two measurements (e.g., height and weight) for each subject in some experiment. We will model this as a pair (X,Y) of RVs. We are interested in whether $\mathbb{E}[X] \neq \mathbb{E}[Y]$. In order to test this, we made i.i.d. observations $(X_1,Y_1),\cdots,(X_n,Y_n)$. Notice that the pairs are independent as random vectors, but each X_i and Y_i should be dependent (height and weight of the same ith subject). To build confidence interval for $\mathbb{E}[X]-\mathbb{E}[Y]$, we start from the observation that $X_1-Y_1,\cdots,X_n-Y_n$ are i.i.d. RVs. Denote their sample mean by $\bar{D}$. First note that

$$\mu_D := \mathbb{E}[\bar{D}] = \mathbb{E}[\bar{X}]-\mathbb{E}[\bar{Y}] = \mathbb{E}[X]-\mathbb{E}[Y], \tag{337}$$

$$\sigma_D^2 := \mathrm{Var}(\bar{D}) = \frac{\mathrm{Var}(X_1-Y_1)}{n} = \frac{\mathrm{Var}(X-Y)}{n}. \tag{338}$$

(Note here that $\mathrm{Var}(X-Y) \neq \mathrm{Var}(X)+\mathrm{Var}(Y)$ since X and Y are dependent.) Note that being the sample mean of n i.i.d. RVs, $\bar{D}$ converges to a normal distribution by CLT:

$$\bar{D} \Rightarrow N\left(\mu_D, \sigma_D^2\right). \tag{339}$$

Hence if we know σ_D, this gives the following (approximate) $100(1-\alpha)\%$ confidence interval for μ_D :

$$\mathbb{P}\left(\mu_D \in \left[(\bar{X}-\bar{Y}) - z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}, (\bar{X}-\bar{Y}) + z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}\right]\right) \approx 1-\alpha. \tag{340}$$

In case σ_D is unknown, but if $X_i - Y_i$'s are normal, then we may use the sample variance S_D^2 of the differences $X_i - Y_i$ to replace σ_D^2 and use the t-distribution. Namely, we have

$$\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \sim t(n-1). \tag{341}$$

This gives

$$\mathbb{P}\left(\mu_D \in \left[\bar{D} - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{342}$$

Recalling that $\bar{D} = \bar{X} - \bar{Y}$, the following random interval

$$\left[(\bar{X}-\bar{Y}) - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, (\bar{X}-\bar{Y}) + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right] \tag{343}$$

contains $\mathbb{E}[X] - \mathbb{E}[Y]$ with probability $1-\alpha$.

Example 7.4.6 (Confidence interval for the difference of two means V)**.** We revisit the problem in Example 7.4.3. Let $X_1, \cdots, X_n$ i.i.d. from $N(\mu_X, \sigma_Y^2)$ and $Y_1, \cdots, Y_m$ i.i.d. from $N(\mu_Y, \sigma_Y^2)$, with $n < m$. Also assume that all such samples are independent of each other and that σ_X and σ_Y are unknown. In case when $\sigma_X = \sigma_Y$, we can use pooled estimator and make use of the full samples, as discussed in Example 7.4.3. In this discussion, suppose we do not know if $\sigma_X = \sigma_Y$. Since all we are interested in is the difference $\mu_X - \mu_Y$, we will only use first n samples from the larger sample $Y_1, \cdots, Y_m$, and form the following n i.i.d. samples of differences:

$$X_1 - Y_1, X_2 - Y_2, \cdots, X_n - Y_n. \tag{344}$$

These are i.i.d. RVs with mean $\mu_X - \mu_Y$ and unknown variance. We can use the sample variance of the above differences to replace this unknown variance. This reduces our problem to the exact situation as in the previous example. Namely, we use the sample variance S_D^2 of the differences $X_i - Y_i$ and use the t-distribution so that

$$\frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{S_D/\sqrt{n}} \sim t(n-1). \tag{345}$$

Thus the following random interval

$$\left[(\bar{X}-\bar{Y}) - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, (\bar{X}-\bar{Y}) + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right] \tag{346}$$

contains $\mu_X - \mu_Y$ with probability $1-\alpha$.

Example 7.4.7 (Excerpted and modified from [HTZ77])**.** An experiment was conducted to compare people's reaction times to a red light versus a green light. When signaled with either the red or the green light, the subject was asked to hit a switch to turn off the light. When the switch was hit, a clock was turned off and the reaction time in seconds was recorded. The following results give the reaction times for eight subjects with some missing data:

Suppose we do know that these samples are from some normal distribution, but not the value of their variances and if they equal to each other. Since each X and Y in the same row are from the same

Subject	Red (x)	Green (y)	Difference $(x-y)$
1	0.30	0.43	-0.13
2	0.23	0.32	-0.09
3	0.41	0.58	-0.17
4	0.53	0.46	0.07
5	0.24	0.27	-0.03
6	0.36	0.41	-0.05
7	0.38	0.38	0.00
8	0.51	0.61	-0.10
9	?	0.83	?
10	0.49	?	?

subject, they are likely to be dependent of each other. By disregarding the last two incomplete data, we may use the approach in Example 7.4.5. Namely, using only the first 8 rows, we compute

$$\bar{d} = -0.0625, \quad s_d = 0.0765. \tag{347}$$

Noting that $t_{0.025}(7) = 2.365$, we compute the following 95% confidence interval for $\mu_X - \mu_Y$:

$$-0.0625 \pm 2.365 \frac{0.0765}{\sqrt{7}}, \quad \text{or} \quad [-0.1265, 0.0015]. \tag{348}$$

Since this interval contains 0, we cannot conclude $\mu_X - \mu_Y$ with 95% confidence.

Next, consider the samples X and Y even from the same subject are independent. Also assume that X and Y have the same variance. Then we can use the approach in Example 7.4.3 using pooled estimator. This approach will make use of all 10 rows in the table (You may carry out this computation). However, even if we do not have this equal variance assumption, we may proceed in the same way as before and obtain the same conclusion as in the previous paragraph. This illustrates our discussion in Example 7.4.6. ▲

5. Confidence interval for proportions

Example 7.5.1. Let E_A be the event that a randomly select voter supports candidate A. Using a poll, we would like to estimate $p = \mathbb{P}(E_A)$, which can be understood as the proportion of supporters of candidate A. As before, we observe a sequence of i.i.d. indicator variables $X_k = \mathbf{1}(E_A)$. Let $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$. be the empirical proportion of supporters of A out of n samples. We know by WLLN that $\hat{p}_n$ converges to p in probability. But if we want to guarantee a certain significance level α for an error bound ε, how many samples should be take?

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. According to the CLT, it is immediate to deduce the following convergence in distribution

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \Rightarrow Z \sim N(0,1). \tag{349}$$

Hence for any $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(|\hat{p}_n - p| \leq \varepsilon\right) &= \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \leq \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) && (350)\\
&\geq \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \leq 2\varepsilon\sqrt{n}\right) && (351)\\
&\approx \mathbb{P}\left(|Z| \leq 2\varepsilon\sqrt{n}\right) = 2\mathbb{P}(0 \leq Z \leq 2\varepsilon\sqrt{n}), && (352)
\end{aligned}$$

where for the inequality we have used the fact that $p(1-p) \le 1/4$ for all $0 \le p \le 1$. The last expression is at least $1-\alpha$ if and only if

$$\mathbb{P}(0 \le Z \le 2\varepsilon\sqrt{n}) \ge 0.5 - (\alpha/2). \tag{353}$$

This shows the following implication (for n large)

$$n \ge \left(\frac{z_{\alpha/2}}{2\varepsilon}\right)^2 \quad \Longrightarrow \quad \mathbb{P}\left(|\hat{p}_n - p| \le \varepsilon\right) \ge 1-\alpha. \tag{354}$$

See Exercise 7.6.3 for more details.

For instance, from the table of standard normal distribution, we know that $\mathbb{P}(0 \le Z \le 1.96) = 0.475$. Hence

$$n \ge \left(\frac{0.98}{\varepsilon}\right)^2 \tag{355}$$

implies that p is within $\hat{p}$ with at least probability 0.95. For instance, $\varepsilon = 0.01$ gives $n \ge 9604$. ▲

In the discussion above, we made use of the simple fact $p(1-p) \le 1/4$ for all $p \in [0,1]$ to get rid of the dependence on the error term on the unknown parameter p. In the exercise below, we will present a slightly more precise analysis.

Exercise 7.5.2. Let $X_1, \cdots, X_n$ be i.i.d. samples from Bernoulli(p) and denote $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$.

(i) Show that

$$\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le z \quad \Longleftrightarrow \quad n(\hat{p}_n - p)^2 \le z^2 p(1-p) \tag{356}$$

$$\Longleftrightarrow \quad \left(1 + (z^2/n)\right)p^2 - 2\left(\hat{p}_n + (z^2/(2n))\right)p + \hat{p}_n^2 \le 0. \tag{357}$$

(ii) Solving the quadratic inequality in (i) for p and using CLT, show that an approximate $100(1-\alpha)\%$ confidence interval for p is given by the end points

$$\frac{\hat{p}_n + (z^2/(2n))}{1 + (z^2/n)} \pm \frac{z\sqrt{\hat{p}_n(1-\hat{p}_n)/n + (z/(2n))^2}}{1 + (z^2/n)} \tag{358}$$

for $z = z_{\alpha/2}$.

(iii) For $n \gg 1$, show that an approximate $100(1-\alpha)\%$ confidence interval is given by the endpoints

$$\hat{p}_n \pm z\sqrt{\hat{p}_n(1-\hat{p}_n)/n}. \tag{359}$$

Example 7.5.3 (Excerpted from [HTZ77]). In a certain political campaign, one candidate has a poll taken at random among the voting population. The results are that 185 votes out of $n = 351$ voters favor this candidate. Even though $\hat{p} = 185/351 = 0.527$, should the candidate feel very confident of winning? From Exercise 7.5.2, an approximate 95% confidence interval for the fraction p of the voting population who favor the candidate is

$$0.527 \pm 1.96\sqrt{\frac{(0.527)(0.473)}{351}}, \tag{360}$$

or, equivalently, $[0.475, 0.579]$. Thus, there is a good possibility that p is less than 50%, and the candidate should certainly take this possibility into account in campaigning. ▲

6. Sample size

So far, our focus was to obtain a confidence interval for the population mean for a confidence $1-\alpha$ of choice, given that we have i.i.d. samples $X_1, \cdots, X_n$ from an unknown distribution $f_{X;\theta}$. In other words, we tried to get the most out of a given sample. In this section, we consider a reverse question: If we want to get confidence interval for the population mean of a certain confidence and error bound, how large should the sample size n has to be? We have already seen an example of this discussion in Example 7.5.1.

In fact, using the confidence intervals we have derived, we can estimate the needed sample size for an easy computation.

Exercise 7.6.1 (Sample size for estimating the mean)**.** Let $X_1,\cdots,X_n$ be i.i.d. samples from some unknown distribution of mean μ. Let $\bar{X}$ and S^2 denote the sample mean and sample variance. Fix $\alpha \in (0,1)$ and $\varepsilon > 0$.

(i) Suppose the population distribution is $N(\mu,\sigma^2)$ for known $\sigma^2 > 0$. Recall that we have the following $100(1-\alpha)\%$ confidence interval for μ:

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{361}$$

Deduce that

$$n \ge (2z_{\alpha/2}\sigma/\varepsilon)^2 \iff z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \frac{\varepsilon}{2} \iff \mathbb{P}\left(\mu \in \left[\bar{X} - \frac{\varepsilon}{2}, \bar{X} + \frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{362}$$

(ii) Suppose the population distribution is not necessarily normal but has known variance $\sigma^2 > 0$. Recall that CLT gives us the following approximate $100(1-\alpha)\%$ confidence interval for μ:

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) \approx 1-\alpha. \tag{363}$$

Deduce that, approximately for large $n \ge 1$,

$$n \ge (2z_{\alpha/2}\sigma/\varepsilon)^2 \iff z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \frac{\varepsilon}{2} \iff \mathbb{P}\left(\mu \in \left[\bar{X} - \frac{\varepsilon}{2}, \bar{X} + \frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{364}$$

(iii) Suppose the population distribution is normal but unknown variance σ^2. Recall that we have the following $100(1-\alpha)\%$ confidence interval for μ:

$$\mathbb{P}\left(\mu \in \left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{365}$$

Deduce that

$$n \ge (2t_{\alpha/2}(n-1)S/\varepsilon)^2 \iff t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le \frac{\varepsilon}{2} \iff \mathbb{P}\left(\mu \in \left[\bar{X} - \frac{\varepsilon}{2}, \bar{X} + \frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{366}$$

(iv) In each cases of (i)-(ii), suppose $\sigma^2 = 5$. If we would like to get a 95% confidence interval for μ of length ≤ 0.01, what is the smallest sample size n that guarantees this?

(v) In the cases of (iii), suppose $s^2 = 5$ and $n = 30$. If we would like to get a confidence interval of length ≤ 0.01, what is the largest confidence we can guarantee?

Exercise 7.6.2 (Sample size for estimating the difference of two means)**.** Let $X_1,\cdots,X_n$ and $Y_1,\cdots,Y_m$ be i.i.d. samples from some unknown distributions of mean μ_X and μ_Y, respectively. Fix $\alpha \in (0,1)$ and $\varepsilon > 0$.

(i) Suppose the population distribution for X_i's and Y_i's are normal with known variances σ_X^2 and σ_Y^2, respectively. Recall that we have the following $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$:

$$\mathbb{P}\left(\mu_X - \mu_Y \in \left[(\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right]\right) = 1-\alpha. \tag{367}$$

Deduce that

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \le \frac{\varepsilon}{2} \iff \mathbb{P}\left(\mu \in \left[\bar{X} - \frac{\varepsilon}{2}, \bar{X} + \frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{368}$$

(ii) Suppose the population distribution for X_i's and Y_i's are not necessarily normal with known variances σ_X^2 and σ_Y^2, respectively. Recall that CLT gives us the following $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$:

$$\mathbb{P}\left(\mu_X - \mu_Y \in \left[(\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},\, (\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right]\right) \approx 1-\alpha. \tag{369}$$

Deduce that (approximately for large n)

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}} \le \frac{\varepsilon}{2} \quad\Longleftrightarrow\quad \mathbb{P}\left(\mu \in \left[\bar{X}-\frac{\varepsilon}{2}, \bar{X}+\frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{370}$$

(iii) Suppose the population distribution are normal but with unknown but equal variances. Recall that we have the following $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$:

$$\mathbb{P}\left(\mu_X - \mu_Y \in \left[(\bar{X}-\bar{Y}) - t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}},\, (\bar{X}-\bar{Y}) + t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]\right) = 1-\alpha. \tag{371}$$

where $S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$. Deduce that

$$t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}} \le \frac{\varepsilon}{2} \quad\Longleftrightarrow\quad \mathbb{P}\left(\mu \in \left[\bar{X}-\frac{\varepsilon}{2}, \bar{X}+\frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{372}$$

(iv) Let $(X_1,Y_1),\cdots,(X_n,Y_n)$ be i.i.d. copies of a random vector (X,Y) with unknown mean and variance. Suppose $X_i - Y_i$'s have normal distribution. Let S_D^2 denote the sample variance for the differences $X_1-Y_1,\cdots,X_n-Y_n$. Recall that we have the following

$$\mathbb{P}\left((\mathbb{E}[X]-\mathbb{E}[Y]) \in \left[(\bar{X}-\bar{Y}) - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}, (\bar{X}-\bar{Y}) + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right]\right) = 1-\alpha. \tag{373}$$

Deduce that

$$z_{\alpha/2}\frac{S_D}{\sqrt{n}} \le \frac{\varepsilon}{2} \quad\Longleftrightarrow\quad \mathbb{P}\left((\mathbb{E}[X]-\mathbb{E}[Y]) \in \left[(\bar{X}-\bar{Y})-\frac{\varepsilon}{2}, (\bar{X}-\bar{Y})+\frac{\varepsilon}{2}\right]\right) \ge 1-\alpha. \tag{374}$$

(v) In each cases of (i)-(ii), suppose $\sigma_X^2 = 5$ and $\sigma_Y^2 = 3$. If we would like to get a 95% confidence interval for $\mathbb{E}[X]-\mathbb{E}[Y]$ of length ≤ 0.01, what is the choice of sample sizes n and m such that $n+m$ is minimized?

(vi) In the cases (iii), suppose $s_X^2 = 5$, $s_Y^2 = 3$, $n = 20$, and $m = 30$, depending on the assumption. If we would like to get a confidence interval of length ≤ 0.01, what is the largest confidence we can guarantee?

(v) In the cases of (iv), suppose $n = 20$, $\sum_{i=1}^n x_i^2 = 100$, $\sum_{i=1}^n y_i^2 = 60$, and $\sum_{i=1}^n x_i y_i = 60$. If we would like to get a confidence interval of length ≤ 0.01, what is the largest confidence we can guarantee?

Exercise 7.6.3 (Sample size for estimating proportion)**.** Let $X_1,\cdots,X_n$ be i.i.d. samples from Bernoulli(p) for unknown p. Let $\bar{X}$ and S^2 denote the sample mean and sample variance. Fix $\alpha \in (0,1)$ and $\varepsilon > 0$. Let $\hat{p}_n = n^{-1}(X_1+\cdots+X_n)$.

(i) Use CLT to show that we have the following convergence in distribution as $n \to \infty$:

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \Rightarrow Z \sim N(0,1). \tag{375}$$

(ii) Show that $t(1-t) \le 1/4$ for all $t \in [0,1]$. Use this and (i) to show that

$$\mathbb{P}\left(|\hat{p}_n - p| \le \varepsilon\right) = \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \tag{376}$$

$$\geq \mathbb{P}\left(\left| \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \right| \leq 2\varepsilon\sqrt{n} \right) \approx \mathbb{P}\left(|Z| \leq 2\varepsilon\sqrt{n} \right) = 2\mathbb{P}(0 \leq Z \leq 2\varepsilon\sqrt{n}), \tag{377}$$

(iii) Suppose $2\varepsilon\sqrt{n} \geq z_{\alpha/2}$. Show that

$$2\mathbb{P}(0 \leq Z \leq 2\varepsilon\sqrt{n}) \geq 2\mathbb{P}(0 \leq Z \leq z_{\alpha/2}) = 1-\alpha. \tag{378}$$

(iv) From (ii) and (iii), deduce the following implication (for n large)

$$n \geq \left(\frac{z_{\alpha/2}}{2\varepsilon} \right)^2 \quad \Longrightarrow \quad \mathbb{P}\left(|\hat{p}_n - p| \leq \varepsilon \right) \geq 1-\alpha. \tag{379}$$

(v) Suppose we want to estimate p with confidence at least 99% with error at most 0.01. How large should our sample size n be?

7. Distribution-free confidence intervals for percentiles

The main question we will address in this section is *how we can estimate population percentiles using sample percentiles*. Our main tool are order statistics, which we have discussed in Section 2. The methods presented here does not make any assumption on the population distribution (except that it is continuous), hence the name "distribution-free confidence intervals".

Given a random variable X with distribution f_X, its *median* is the number m such that $\mathbb{P}(X \leq m) = 1/2$. In case there are several such values of m, we may take the smallest of such:

$$m = m(X) := \inf\{x \in \mathbb{R} \,|\, \mathbb{P}(X \leq x) = 1/2\}. \tag{380}$$

For instance, if $X \sim N(3,4)$, then its median is $m = 3$. In general, for given $p \in [0,1]$, the percentile of order p for X (or for its distribution f_X) is defined by

$$\pi_p = \pi_p(X) := \inf\{x \in \mathbb{R} \,|\, \mathbb{P}(X \leq x) = p\}. \tag{381}$$

Recall the definition of sample percentile of order p in Section 2.

We begin our discussion with the following motivating example.

Example 7.7.1. Suppose we have i.i.d. samples $X_1, X_2, \cdots, X_5$ of an unknown continuous RV X, for which we have no idea what its distribution looks like. We are interested in estimating its median $m = m(X)$. For this, we order the samples to obtain the following order statistics

$$Y_1 < Y_2 < Y_3 < Y_4 < Y_5. \tag{382}$$

From this sample, it is tempting that the 'middle value' Y_3 might be close to the population median m. However, since we have only five samples, we will go for a more conservative estimate and claim that m is between Y_1 and Y_5. In what confidence can we make this claim? Namely, we would like to compute the following probability

$$\mathbb{P}\big(m(X) \in (Y_1, Y_5)\big) = \mathbb{P}\left(Y_1 < m(X) < Y_5\right). \tag{383}$$

In order to compute the above probability, observe that by definition of the median $m = m(X)$ and since X is continuous, $\mathbb{P}(X_i < m) = 1/2$ for all $i = 1, 2, \cdots, 5$. For $Y_1 < m(X)$, we need at least one X_i such that $X_i < m$. For $m < Y_5$, we need at least one X_j such that $X_j > m$. By viewing the event $X_i < i$ a 'success' at ith trial, we want a Binomial$(5, 1/2)$ RV to take values from 1 up to 4. Namely,

$$\mathbb{P}\left(Y_1 < m(X) < Y_5\right) = \mathbb{P}\left(k\ X_i\text{'s are} < m \text{ and the rest are} > m \text{ for } k = 1,2,3,4\right) \tag{384}$$

$$= \sum_{k=1}^{4} \mathbb{P}\left(k \text{ sucesses out of 5 trials}\right) \tag{385}$$

$$= \mathbb{P}\left(1 \leq \text{ Binomial } (5,1/2) \leq 4\right). \tag{386}$$

This gives

$$\mathbb{P}\left(Y_1 < m(X) < Y_5\right) = 1 - \mathbb{P}\left(\text{ Binomial } (5,1/2) = 0\right) - \mathbb{P}\left(\text{ Binomial } (5,1/2) = 5\right) \tag{387}$$

$$= 1 - \left(\frac{1}{2}\right)^5 - \left(\frac{1}{2}\right)^5 = \frac{15}{16} \approx 0.94. \tag{388}$$

This means that the population median m is in the interval (Y_1, Y_5) with confidence 94%. ▲

Example 7.7.2 (Excerpted from [HTZ77])**.** The lengths in centimeters of $n = 9$ fish of a particular species captured off the New England coast were 32.5, 27.6, 29.3, 30.1, 15.5, 21.7, 22.8, 21.2, and 19.0. Thus, the observed order statistics are

$$15.5 < 19.0 < 21.2 < 21.7 < 22.8 < 27.6 < 29.3 < 30.1 < 32.5. \tag{389}$$

Before the sample is drawn, we know that

$$\begin{aligned} \mathbb{P}\,(Y_2 < m < Y_8) &= \mathbb{P}\,(2 \le \text{ Binomial } (5,1/2) < 8) \qquad (390)\\ &= \mathbb{P}\,(\text{ Binomial } (5,1/2) \le 7) - \mathbb{P}\,(\text{ Binomial } (5,1/2) \le 1) \qquad (391)\\ &\approx 0.9805 - 0.0195 = 0.9610, \qquad (392)\end{aligned}$$

where for the last computation, we have used the binomial table in [HTZ77, Table II, Appendix B]. From the samples drawn above, we have $y_2 = 19.0$ and $y_8 = 30.1$. Thus the interval $(19.0, 30.1)$ is a 96% confidence interval for the the population median m. ▲

When we have large sample size, we can use CLT for a normal approximation for binomial distribution (see Exercise 7.2.2).

Exercise 7.7.3. Let $X_1, \cdots, X_n$ be i.i.d. samples of a continuous RV X, and let $Y_1 < \cdots < Y_n$ be their order statistics. Fix $p \in [0,1]$ and let $\pi_p = \pi_p(X)$ denote the percentile of X of order p.

(i) For each X_i, show that $\mathbb{P}(X_i < \pi_p) = p$. We will call the event $\{X_i < \pi_p\}$ a 'success at the ith trial'.

(ii) Show that

$$\begin{aligned} \mathbb{P}\left(Y_i < \pi_p < Y_j\right) &= \mathbb{P}(i \le \#\text{ of successes in } n \text{ trials} < j) \qquad (393)\\ &= \mathbb{P}\left(i \le \text{Binomial}(n,p) < j\right) \qquad (394)\\ &= \mathbb{P}\left(i - 0.5 \le \text{Binomial}(n,p) \le j - 0.5\right). \qquad (395)\end{aligned}$$

(iii) Use CLT (or Exercise 7.2.2) and (ii) to deduce that

$$\mathbb{P}\left(Y_i < \pi_p < Y_j\right) \approx \mathbb{P}\left(\frac{i - 0.5 - np}{\sqrt{np(1-p)}} \le N(0,1) < \frac{j - 0.5 - np}{\sqrt{np(1-p)}}\right). \tag{396}$$

Example 7.7.4. Let $Y_1 < \cdots < Y_{16}$ be order statistics of $n = 16$ i.i.d. samples $X_1, \cdots, X_n$ of a continuous RV X. According to Exercise 7.7.3 (ii) and the binomial table in [HTZ77, Table II, Appendix B],

$$\begin{aligned} \mathbb{P}\,(Y_5 < m < Y_{12}) &= \mathbb{P}(5 \le \text{ Binomial } (16,1/2) \le 11) \qquad (397)\\ &= \mathbb{P}(\text{ Binomial } (16,1/2) \le 11) - \mathbb{P}(\text{ Binomial } (16,1/2) \le 4) \qquad (398)\\ &= 0.9616 - 0.0384 = 0.9232. \qquad (399)\end{aligned}$$

On the other hand, using the normal approximation in Exercise 7.7.3 (iii),

$$\begin{aligned} \mathbb{P}\,(Y_5 < m < Y_{12}) &\approx \mathbb{P}\left(\frac{5 - 0.5 - 8}{\sqrt{4}} \le N(0,1) \le \frac{12 - 0.5 - 8}{\sqrt{4}}\right) \qquad (400)\\ &= \mathbb{P}\,(-1.75 \le N(0,1) \le 1.75) = 0.9198, \qquad (401)\end{aligned}$$

where the last numerical value can be computed from the standard normal table given at the end of this note. Thus the normal approximation here gives quite accurate result in computing the confidence. ▲

Exercise 7.7.5 (Excerpted from [HTZ77])**.** Let the following numbers represent the order statistics of the $n = 27$ observations obtained in a random sample from a certain population of incomes (measured in hundreds of dollars):

$$261 \quad 269 \quad 271 \quad 274 \quad 279 \quad 280 \quad 283 \quad 284 \quad 286 \quad 287 \quad 292 \quad 293 \quad 296 \quad 300 \tag{402}$$

$$304 \quad 305 \quad 313 \quad 321 \quad 322 \quad 329 \quad 341 \quad 343 \quad 356 \quad 364 \quad 391 \quad 417 \quad 476 \tag{403}$$

We are interested in estimating the 25th percentile, $\pi_{0.25}$, of the population.

Since $(n+1)p = 28(1/4) = 7$, the seventh order statistic, namely, $y_7 = 283$, would be a point estimate of $\pi_{0.25}$. For interval estimate, we would like to use the interval (y_4, y_{10}) for estimating $\pi_{0.25}$. Compute the confidence of this interval estimate.

CHAPTER 8

Statistical testing

1. Tests about one mean

Suppose we have an unknown RV X. According to some external information, we know that its distribution is either $N(1,5)$ or $N(3,5)$. How can we test which one is correct? We will propose statistical testing procedures for this type of problems, which are based on analysis of sample statistics and confidence intervals that we have developed before.

Example 8.1.1 (Simple alternative, Excerpted from [HTZ77])**.** Let X equal the breaking strength of a steel bar. If the bar is manufactured by process I, we know that $X \sim N(50,36)$. It is hoped that if process II (a new process) is used, then it will be that $X \sim N(55,36)$. Given a large number of steel bars manufactured by process II, how could we test whether the five-unit increase in the mean breaking strength was realized?

There are two hypothesis on the mean μ of $X \sim N(\mu,36)$ that we want to test:

1. Null hypothesis $H_0 : \mu = 50$
2. Alternative hypothesis $H_1 : \mu = 55$.

Both of the above hypotheses are *simple hypotheses*, since they completely describe the distribution of X. Otherwise, they would be called a *composite hypothesis*. We will be testing which of the above hypothesis is more likely to be true, using our given samples:

3. Sampled values $x_1, x_2, \cdots, x_n$ of X manufactured by process II

Since we are testing about the unknown mean of X, we will use the usual test statistic:

4. Test statistic: $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$.

Next, we set up a decision rule based on the observed sample mean $\bar{x}$. Since large values of $\bar{x}$ would indicate H_1 is more likely than H_0, our rule may be of the form "Reject H_0 and accept H_1 if $\bar{x} \geq 53$".

5. Critical region: $C = \{(x_1, \cdots, x_n) \,|\, \bar{x} \geq 53\}$. Reject H_0 and accept on C, otherwise test is inconclusive.

The decision rule above could lead to the wrong conclusion, in each cases when H_0 or H_1 is true. Hence we will compute two types of errors as below:

6. Two types of error probabilities:

$$\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}(\text{Reject } H_0 \text{ when } H_0 \text{ is true}) \tag{404}$$

$$\beta = \mathbb{P}(\text{Type II error}) = \mathbb{P}(\text{Not reject } H_0 \text{ when } H_1 \text{ is true}) \tag{405}$$

In our case, we can compute these error probabilities using the known distribution of $\bar{X}$ under each hypothesis:

$$\bar{X}; H_0 \sim N(50, 36/n), \qquad \bar{X}; H_1 \sim N(55, 36/n). \tag{406}$$

Hence we can compute

$$\alpha = \mathbb{P}(\bar{X} > 53; H_0) = \mathbb{P}\left(\frac{\bar{X}-50}{6/\sqrt{n}} > \frac{53-50}{6/\sqrt{n}}; H_0\right) = \mathbb{P}\left(N(0,1) > \frac{53-50}{6/\sqrt{n}}\right) \tag{407}$$

$$\beta = \mathbb{P}(\bar{X} < 53; H_1) = \mathbb{P}\left(\frac{\bar{X}-55}{6/\sqrt{n}} < \frac{53-55}{6/\sqrt{n}}; H_1\right) = \mathbb{P}\left(N(0,1) < \frac{53-55}{6/\sqrt{n}}\right). \tag{408}$$

Assuming $n = 16$, we get the following values

$$\alpha = 0.0228, \qquad \beta = 0.0913. \tag{409}$$

Hence the probability of making both typo of errors are quite small. See Figure 1.

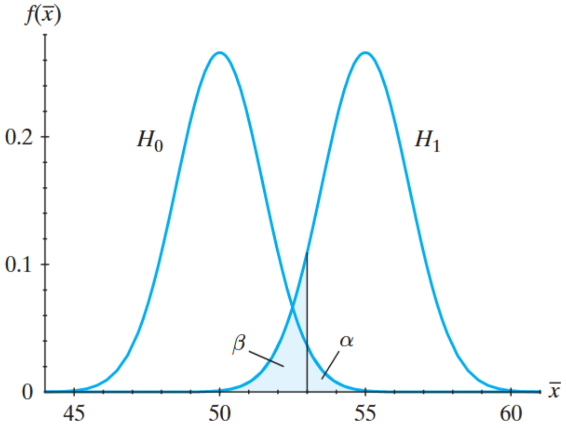


FIGURE 1. PDF of $\bar{X}$ under H_0 and H_1 with type I and II errors.

▲

Example 8.1.2 (Composite alternative, Excerpted and modified from [HTZ77])**.** Assume that the underlying distribution is normal with unknown mean μ but known variance $\sigma^2 = 100$. Say we are testing the simple null hypothesis $H_0 : \mu = 60$ against the composite alternative hypothesis $H_1 : \mu > 60$ with a sample mean $\bar{X}$ based on $n = 52$ observations. Suppose that we obtain the observed sample mean of $\bar{x} = 62.75$, which is a bit larger than μ and biased toward H_1. How should we make our conclusion from this data?

As in the previous example, our decision rule will be based on whether the observed sample mean $\bar{x}$ is biased toward H_1 or not. In this example, the critical interval will be of the form

$$C = \{(x_1, \cdots, x_n) \,|\, \bar{x} > \theta\}, \tag{410}$$

where θ is a certain threshold that we will set. Then the type I error will be

$$\text{type I error} = \mathbb{P}(\bar{X} > \theta; H_0) = \mathbb{P}\left(\frac{\bar{X} - 60}{10/\sqrt{52}} > \frac{\theta - 60}{10/\sqrt{52}}; H_0\right) = \mathbb{P}\left(N(0,1) > \frac{\theta - 60}{10/\sqrt{52}}\right). \tag{411}$$

Note that we can tolerate type I error up to probability 0.05 (also called the *significance level*). We can make type I error arbitrary small by making the rejection threshold θ large, but then the resulting test would not be very useful. So we would like to find smallest value of θ which guarantees type I error to occur with probability at most 0.05. This is when $(\theta - 60)/(10/\sqrt{52}) = z_{0.05}$, so our optimal test for significance level $\alpha = 0.05$ would be

$$\text{Reject } H_0 \quad \text{if } \bar{x} > 60 + z_{0.05}\frac{10}{\sqrt{52}} = 60 + (1.745)\frac{10}{\sqrt{52}} = 62.281. \tag{412}$$

In our case, $\bar{x} = 62.75$ so we will reject H_0 according to the criterion above.

There is an alternative way to form a significance level $\alpha = 0.05$ test. Note that

$$\mathbb{P}(\bar{X} \geq \bar{x}; H_0) < 0.05 \iff \mathbb{P}\left(\frac{\bar{X} - 60}{10/\sqrt{52}} \geq \frac{\bar{x} - 60}{10/\sqrt{52}}; H_0\right) < 0.05 \tag{413}$$

$$\iff \mathbb{P}\left(N(0,1) \geq \frac{\bar{x} - 60}{10/\sqrt{52}}; H_0\right) < 0.05 \tag{414}$$

$$\iff \frac{\bar{x} - 60}{10/\sqrt{52}} > z_{0.05} \tag{415}$$

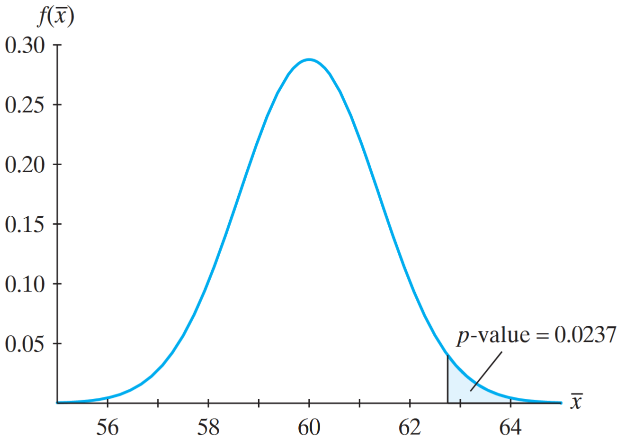


FIGURE 2. Illustration of p-value for $H_0 : \mu = 60$ and $H_1 : \mu > 60$.

$$\iff \quad \bar{x} > 60 + z_{0.05}\frac{10}{\sqrt{52}}. \tag{416}$$

Hence we may reformulate the test in (477) as

$$\text{Reject } H_0 \quad \text{if } \mathbb{P}(\bar{X} \geq \bar{x}; H_0) < 0.05. \tag{417}$$

Here the probability $\mathbb{P}(\bar{X} \geq \bar{x}; H_0)$ is called the *p-value* associated with observed sample mean $\bar{x}$ for this example. ▲

In the previous example, we have introduced the p-value associated with sample mean $\bar{x}$. Roughly speaking, it is the probability of obtaining another random sample mean that is more extreme than the observed value $\bar{x}$ under H_0. As in the MLE, our grounding belief is that we are observing a data because it was likely to occur. Hence if such a p-value is small, this means our observed data is extremely unlikely under H_0, which goes against our belief, so we better reject H_0. Below we formally introduce the p-value depending on the composite alternate hypotheses.

$$p\text{-value associated with } \bar{x} = \mathbb{P}\begin{pmatrix}\text{getting a sample } X_1, \cdots, X_n \text{ uner } H_0 \text{ s.t. } \bar{X} \text{ is biased} \\ \text{toward } H_0 \text{ more than the observed sample mean } \bar{x}\end{pmatrix} \tag{418}$$

$$= \begin{cases} \mathbb{P}(\bar{X} \geq \bar{x}; H_0) & \text{if } H_1 : \mu > \mu_0 \\ \mathbb{P}(\bar{X} \geq \bar{x}; H_0) & \text{if } H_1 : \mu < \mu_0 \\ \mathbb{P}(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|; H_0) & \text{if } H_1 : \mu \neq \mu_0 \end{cases} \tag{419}$$

By solving the equation (p-value) $< \alpha$ in the case when the RV X of interest is known to follow normal distribution with variance σ^2, we can summarize our test about one mean of significance level α.

Remark 8.1.3. If we do not know X has normal distribution but know the variance $\text{Var}(X) = \sigma^2$, we can use CLT to get the above critical regions approximately for large n. Hence the same hypothesis testing holds approximately for large n.

When we do not know population variance but still know that it has normal distribution, then we can use t-scores instead of z-scores for testing one mean.

Example 8.1.4 (Excerpted from [HTZ77]). Let X (in millimeters) equal the growth in 15 days of a tumor induced in a mouse. Assume that the distribution of X is $N(\mu, \sigma^2)$. We shall test the null hypothesis $H_0 : \mu = \mu_0 = 4.0$ against the two-sided alternative hypothesis $H_1 : \mu \neq 4.0$. If we use $n = 9$ observations and a significance level of $\alpha = 0.10$, the critical region is given by

$$\left|\frac{\bar{x} - 4}{s/\sqrt{9}}\right| \geq t_{\alpha/2}(8) = 1.860. \tag{422}$$

Algorithm 1 Hypothesis testing for one mean, known variance σ^2

```
Input:
    Null hypothesis $H_0 : X \sim N(\mu_0, \sigma^2)$ (or only $\mathrm{Var}(X) = \sigma^2$ without normal distribution)
    Alternative hypothesis $H_1$ : one of $\mu < \mu_0$ or $\mu > \mu_0$ or $\mu \neq \mu_0$, where $\mu = \mathbb{E}[X]$
    Observed sample: $x_1, x_2, \cdots, x_n$.
        Sample mean $\bar{x} = n^{-1}(x_1 + \cdots + x_n)$
    significance level: $\alpha \in (0,1)$
    Test statistic: $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$
Do: Compute the critical region associated with $\bar{x}$:
    $$\begin{cases} \bar{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} & \text{if } H_1 : \mu > \mu_0 \\ \bar{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} & \text{if } H_1 : \mu < \mu_0 \\ |\bar{x} - \mu_0| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}} & \text{if } H_1 : \mu \neq \mu_0 \end{cases} \tag{420}$$
Do:
    If   Inside critical region, Reject $H_0$
    Else   Test inconclusive and cannot reject $H_0$
```

Algorithm 2 Hypothesis testing for one mean, unknown variance

```
Input:
    Null hypothesis $H_0 : X \sim N(\mu_0, ?)$
    Alternative hypothesis $H_1$ : one of $\mu < \mu_0$ or $\mu > \mu_0$ or $\mu \neq \mu_0$, where $\mu = \mathbb{E}[X]$
    Observed sample: $x_1, x_2, \cdots, x_n$.
        Sample mean $\bar{x} = n^{-1}(x_1 + \cdots + x_n)$ and sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
    significance level: $\alpha \in (0,1)$
    Test statistic: $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$
Do: Compute the critical region associated with $\bar{x}$:
    $$\begin{cases} \bar{x} > \mu_0 + t_\alpha \frac{s}{\sqrt{n}} & \text{if } H_1 : \mu > \mu_0 \\ \bar{x} < \mu_0 - t_\alpha \frac{s}{\sqrt{n}} & \text{if } H_1 : \mu < \mu_0 \\ |\bar{x} - \mu_0| > t_{\alpha/2} \frac{s}{\sqrt{n}} & \text{if } H_1 : \mu \neq \mu_0 \end{cases} \tag{421}$$
Do:
    If   Inside critical region, Reject $H_0$
    Else   Test inconclusive and cannot reject $H_0$
```

Specifically, if we are given the data $n = 9$, $\bar{x} = 4.3$, and $s = 1.2$, then

$$\frac{4.3 - 4}{1.2/\sqrt{9}} = 0.75 < 1.860. \tag{423}$$

Thus in this case cannot reject $H_0 : \mu = 4.0$ at significance level $\alpha = 0.1$.

In terms of the p-value, let $T \sim t(8)$. Then

$$p\text{-value} = \mathbb{P}(|T| > 0.075) = 2\mathbb{P}(T > 0.075) = 2(0.2374) = 0.4748. \tag{424}$$

Since this p-value is much larger than the significance level $\alpha = 0.1$, we cannot reject $H_0 : \mu = 4.0$ at significance level $\alpha = 0.1$.

▲

Exercise 8.1.5 (Excerpted and modified from [HTZ77])**.** Twenty-four girls in the 9th and 10th grades were put on an ultra heavy rope-jumping program. Someone thought that such a program would increase

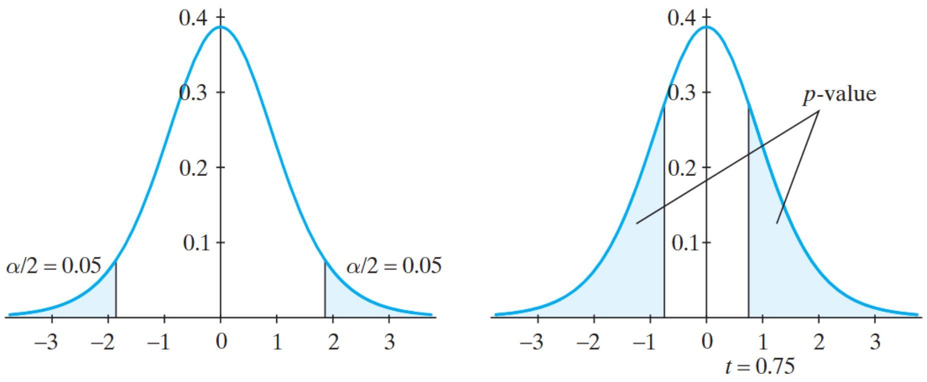


FIGURE 3. Test about mean of tumor growth

their speed in the 40-yard dash. Let W equal the difference in time to run the 40-yard dash—the "before-program time" minus the "after-program time." Assume that the distribution of W is approximately $N(\mu_W, \sigma_W^2)$. We shall test the null hypothesis $H_0 : \mu_W = 0$ against the alternative hypothesis $H_1 : \mu_W > 0$.

(i) Show that the test statistic and the critical region that has an $\alpha = 0.05$ significance level are given by

$$t = \frac{\bar{w} - 0}{s_W / \sqrt{24}} > t_{0.05}(23) = 1.714, \tag{425}$$

where $\bar{w}$ is the observed value of the sample mean for the random variable W.

(ii) The following data give the difference in time that it took each girl to run the 40-yard dash, with positive numbers indicating a faster time after the program:

$$0.28 \quad 0.01 \quad 0.13 \quad 0.33 \quad -0.03 \quad 0.07 \quad -0.18 \quad -0.14 \tag{426}$$

$$-0.33 \quad 0.01 \quad 0.22 \quad 0.29 \quad -0.08 \quad 0.23 \quad 0.08 \quad 0.04 \tag{427}$$

$$-0.30 \quad -0.08 \quad 0.09 \quad 0.70 \quad 0.33 \quad -0.34 \quad 0.50 \quad 0.06 \tag{428}$$

Compute the observed value of the test statistic t. Can we reject the null hypothesis $H_0 : \mu_W = 0$ at significance level $\alpha = 0.05$? Also, compute the p-value.

(iii) Find the smallest value of significance level α under which we can reject the null hypothesis $H_0 : \mu_W = 0$ using the data given in (ii). Does it equal to the p-value found in (ii)? Explain why.

2. Test about two means

In this section, we study how to statistically compare the means of two unknown RVs X and Y. Our statistical testing we will develop in this section is based on the confidence interval for difference of two means, which are summarized in Exercise 7.6.2.

We start with the ideal situation when both X and Y are normal with known variances.

Example 8.2.1 (Normal with known variances)**.** Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, where their variances are known. Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. Suppose we have i.i.d. samples for $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ for X and Y, respectively. Our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ will be of the form

$$C = \{(x_1, \cdots, x_n, y_1, \cdots, y_m) \,|\, |\bar{x} - \bar{y}| \geq \theta\}, \tag{429}$$

for some rejection threshold θ. To compute the corresponding type I error, note that

$$\text{type I error} = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \,|\, H_0) \tag{430}$$

$$= \mathbb{P}(|\bar{X} - \bar{Y}| \geq \theta \,|\, H_0). \tag{431}$$

Recall from Example 7.4.2 that

$$Z := \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1). \tag{432}$$

Since $\mu_X = \mu_Y$ under H_0, we have

$$\text{type I error} = \mathbb{P}(|\bar{X}-\bar{X}| \geq \theta \,|\, H_0) = \mathbb{P}\left(|Z| \geq \frac{|\theta-(\mu_X-\mu_Y)|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \,|\, H_0 \right) = \mathbb{P}\left(|Z| \geq \frac{|\theta|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \right). \tag{433}$$

Hence it follows that

$$\text{type I error} \leq \alpha \quad \Longleftrightarrow \quad \frac{|\theta|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \geq z_{\alpha/2}. \tag{434}$$

Hence this gives the following critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ at significance level α:

$$|\bar{x}-\bar{y}| \geq z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}. \tag{435}$$

▲

Example 8.2.2 (Known variances, not necessarily normal)**.** Suppose we have two RVs X and Y of interest, and their variances are known as σ_X^2 and σ_Y^2, respectively, but not that they follow normal distribution. Using CLT, all our discussion in the 8.5.4 holds approximately for large sample sizes n and m.

Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. Suppose we have i.i.d. samples for $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ for X and Y, respectively. Our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ will be of the form

$$C = \{(x_1, \cdots, x_n, y_1, \cdots, y_m) \,|\, |\bar{x}-\bar{y}| \geq \theta\}, \tag{436}$$

for some rejection threshold θ. By CLT, we have

$$Z := \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \Rightarrow N(0,1) \tag{437}$$

as both $n, m \to \infty$. Hence we can approximately compute the type I error:

$$\begin{align} \text{type I error} &= \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \,|\, H_0) \tag{438} \\ &= \mathbb{P}(|\bar{X}-\bar{Y}| \geq \theta \,|\, H_0) \tag{439} \\ &= \mathbb{P}\left(|Z| \geq \frac{|\theta|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \right) \approx \mathbb{P}\left(|N(0,1)| \geq \frac{|\theta|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \right) \tag{440} \end{align}$$

Recall from Example 7.4.2 that

$$Z := \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1). \tag{441}$$

Hence it follows that, approximately for large n, m,

$$\text{type I error} \leq \alpha \quad \Longleftrightarrow \quad \frac{|\theta|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \geq z_{\alpha/2}. \tag{442}$$

Hence this gives the following approximate critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ at significance level α:

$$|\bar{x} - \bar{y}| \geq z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}. \tag{443}$$

▲

Example 8.2.3 (Normal with unknown, equal variances)**.** Let $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$, where their variances are known. Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. Suppose we have i.i.d. samples for $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ for X and Y, respectively. We first recall from Example 7.4.3 that

$$T := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2), \tag{444}$$

where S_p^2 denotes the pulled estimator

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}. \tag{445}$$

Hence in this case, we will form our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ as

$$C = \left\{ (x_1, \cdots, x_n, y_1, \cdots, y_m) : \left| \frac{\bar{x} - \bar{y}}{s_p\sqrt{(1/n) + (1/m)}} \right| \geq \theta \right\}, \tag{446}$$

for some rejection threshold θ. To compute the corresponding type I error,

Hence we can (exactly) compute the type I error in this case:

$$\text{type I error} = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \mid H_0) \tag{447}$$

$$= \mathbb{P}\left(\left| \frac{\bar{X} - \bar{Y}}{S_p\sqrt{(1/n) + (1/m)}} \right| \geq \theta \mid H_0 \right) = \mathbb{P}\left(|T| \geq \theta \mid H_0\right) = \mathbb{P}\left(|t(n+m-2)| \geq \theta\right). \tag{448}$$

Hence it follows that

$$\text{type I error} \leq \alpha \quad \Longleftrightarrow \quad \theta \geq t_{\alpha/2}(n+m-2). \tag{449}$$

Hence this gives the following critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ at significance level α:

$$\frac{|\bar{x} - \bar{y}|}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \geq t_{\alpha/2}(n+m-2). \tag{450}$$

▲

Exercise 8.2.4 (Paired test, known variance of the difference)**.** Let X, Y be RVs. Denote $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_Y$. Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. Suppose we have i.i.d. pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the joint distribution of (X, Y). Further assume that we know the value of $\sigma^2 = \mathrm{Var}(X - Y)$.

(i) Noting that $X_1 - Y_1, \cdots, X_n - Y_n$ are i.i.d. with finite variance σ^2, show using CLT that, as $n \to \infty$,

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma/\sqrt{n}} \Longrightarrow N(0,1). \tag{451}$$

(ii) Our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ will be of the form

$$C = \{(x_1, \cdots, x_n, y_1, \cdots, y_m) \mid |\bar{x} - \bar{y}| \geq \theta\}, \tag{452}$$

for some rejection threshold $\theta > 0$. Show that

$$\text{type I error} = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \mid H_0) \tag{453}$$

$$= \mathbb{P}(|\bar{X}-\bar{Y}| \geq \theta \,|\, H_0) \tag{454}$$

$$= \mathbb{P}\left(|Z| \geq \frac{\theta}{\sigma/\sqrt{n}}\,;\, H_0\right) \approx \mathbb{P}\left(|N(0,1)| \geq \frac{\theta}{\sigma/\sqrt{n}}\right). \tag{455}$$

(iii) Conclude that (approximately for large n)

$$\text{type I error} \leq \alpha \quad \Longleftrightarrow \quad \frac{\theta}{\sigma/\sqrt{n}} \geq z_{\alpha/2} \quad \Longleftrightarrow \quad \theta \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{456}$$

(iv) Show that the critical region of rejecting $H_0 : \mu_X = \mu_Y$ in favor of $H_1 : \mu_Y \neq \mu_Y$ with significance level α is given by

$$|\bar{x}-\bar{y}| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{457}$$

Exercise 8.2.5 (Paired test, known normality of the difference)**.** Let X, Y be RVs. Denote $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_Y$. Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. Suppose we have i.i.d. pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the joint distribution of (X, Y). Further assume that we know the $X - Y$ follows a normal distribution.

(i) Noting that $X_1 - Y_1, \cdots, X_n - Y_n$ are i.i.d. with normal distribution, show that (exactly)

$$T := \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{S/\sqrt{n}} \sim t(n-1), \tag{458}$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}((X_i - Y_i) - (\bar{X}-\bar{Y}))^2$ denotes the sample variance.

(ii) Our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ will be of the form

$$C = \left\{(x_1, \cdots, x_n, y_1, \cdots, y_m) \,\middle|\, \frac{|\bar{x}-\bar{y}|}{s/\sqrt{n}} \geq \theta\right\}, \tag{459}$$

for some rejection threshold $\theta > 0$. Show that

$$\text{type I error} = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \,;\, H_0) \tag{460}$$

$$= \mathbb{P}\left(\frac{|\bar{X}-\bar{Y}|}{S/\sqrt{n}} \geq \theta \,;\, H_0\right) = \mathbb{P}\left(|t(n-1)| \geq \theta\right). \tag{461}$$

(iii) Conclude that (exactly for all n)

$$\text{type I error} \leq \alpha \quad \Longleftrightarrow \quad \theta \geq t_{\alpha/2}(n-1). \tag{462}$$

(iv) Show that the critical region of rejecting $H_0 : \mu_X = \mu_Y$ in favor of $H_1 : \mu_X \neq \mu_Y$ with significance level α is given by

$$\frac{|\bar{x}-\bar{y}|}{s/\sqrt{n}} \geq t_{\alpha/2}(n-1). \tag{463}$$

Example 8.2.6 (Excerpted from [HTZ77])**.** A product is packaged by a machine with 24 filler heads numbered 1 to 24, with the odd-numbered heads on one side of the machine and the even on the other side. Let X and Y equal the fill weights in grams when a package is filled by an odd-numbered head and an even-numbered head, respectively. Assume that the distributions of X and Y are $N(\mu_X, \sigma_2)$ and $N(\mu_Y, \sigma^2)$, respectively, and that X and $Y$are independent. We would like to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_x \neq \mu_Y$. To perform the test, after the machine has been setup and is running, we shall select one package at random from each filler head and weigh it. The test statistic is that given by Equation 8.2-1 with $n = m = 12$. At an $\alpha = 0.10$ significance level, the critical region is given by

$$\frac{|\bar{x}-\bar{x}|}{x_p\sqrt{\frac{1}{12}+\frac{1}{12}}} \geq t_{0.05}(22) = 1.717. \tag{464}$$

Suppose we have the following 12 observations each for X and Y:

$$X: \quad 1071 \quad 1076 \quad 1070 \quad 1083 \quad 1082 \quad 1067 \quad 1078 \quad 1080 \quad 1075 \quad 1084 \quad 1075 \quad 1080 \tag{465}$$

$$Y: \quad 1074 \quad 1069 \quad 1075 \quad 1067 \quad 1068 \quad 1079 \quad 1082 \quad 1064 \quad 1070 \quad 1073 \quad 1072 \quad 1075 \tag{466}$$

We can compute $\bar{x} = 1076.75$, $s_X^2 = 29.30$, $\bar{y} = 1072.33$, and $s_Y^2 = 26.24$. Hence the observed value of the test statistic is

$$t = \frac{|1076.75 - 1072.33|}{\sqrt{\frac{(11)(29.30)+(11)(26.24)}{22}\frac{1}{12} + \frac{1}{12}}} = 2.05. \tag{467}$$

Since $t = 2.05 > 1.717 = t_{0.05}(22)$, we reject the null hypothesis $H_0 : \mu_X = \mu_Y$ at significance level $\alpha = 0.1$. ▲

Exercise 8.2.7 (Excerpted from [HTZ77])**.** Plants convert carbon dioxide (CO_2) in the atmo-sphere, along with water and energy from sunlight, in to the energy they need for growth and reproduction. Experiments were performed under normal atmospheric air conditions and in air with enriched CO_2 concentrations to determine the effect on plant growth. The plants were given the same amount of water and light for a four-week period. The following table gives the plant growths in grams:

$$\text{Normal Air:} \quad 4.67 \quad 4.21 \quad 2.18 \quad 3.91 \quad 4.09 \quad 5.24 \quad 2.94 \quad 4.71 \quad 4.04 \quad 5.79 \quad 3.80 \quad 4.38 \tag{468}$$

$$\text{Enriched Air:} \quad 5.04 \quad 4.52 \quad 6.18 \quad 7.01 \quad 4.36 \quad 1.81 \quad 6.22 \quad 5.70 \tag{469}$$

On the basis of these data, determine whether CO_2-enriched atmosphere increases plant growth.

3. Tests about proportions

Suppose we have a Bernoulli RV X of unknown success probability p. The goal of this section is to develop statistical testing procedures that infers the value of p from a given set of i.i.d. samples $X_1, X_2, \cdots, X_n$.

Example 8.3.1. Suppose $X \sim \text{Bernoulli}(p)$ with unknown p. Say we are testing the simple null hypothesis $H_0 : p = p_0$ against the composite alternative hypothesis $H_1 : p > p_0$. Let $X_1, \cdots, X_n$ be i.i.d. samples of X and let $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$ denote the sample frequency. We know $\hat{p}_n$ is an unbiased estimator for p (e.g., $\mathbb{E}[\hat{p}_n] = p$) and strong law of large numbers tells us $\hat{p}_n \to p$ almost surely as $n \to \infty$. But we only have a finite number of samples (say, $n = 30$) and we need to make a decision whether $p = p_0$ or $p > p_0$.

As before, our decision rule will be based on whether the observed sample frequency $\hat{p}_n$ is biased toward H_1 or not. In this example, the critical interval will be of the form

$$C = \{(x_1, \cdots, x_n) \,|\, \bar{p}_n > \theta\}, \tag{470}$$

where θ is a certain threshold that we will set. Note that we can tolerate type I error up to probability α (say 0.05) (also called the *significance level*). We can make type I error arbitrary small by making the rejection threshold θ large, but then the resulting test would not be very useful. So we would like to find smallest value of θ which guarantees type I error to occur with probability at most α.

The key observation here is that

$$n\hat{p}_n = (X_1 + \cdots + X_n) \sim \text{Binomial}(n, p). \tag{471}$$

So the type I error will be

$$\text{type I error} \le \alpha = \mathbb{P}(\bar{p}_n > \theta; H_0) = \mathbb{P}\left(n\hat{p}_n > n\theta; H_0\right) = \mathbb{P}(\text{Binomial}(n, p_0) > n\theta). \tag{472}$$

If we use the binomial table, the above equation will give us the exact type I error. But when n is sufficiently large, we can proceed by using CLT:

$$\mathbb{P}(\bar{p}_n > \theta; H_0) = \mathbb{P}\left(\hat{p}_n > \theta; H_0\right) = \mathbb{P}\left(\frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} > \frac{\theta - p_0}{\sqrt{p_0(1-p_0)/n}}\,; H_0\right) \tag{473}$$

$$= \mathbb{P}\left(\frac{n\hat{p}_n - np_0}{\sqrt{np_0(1-p_0)}} > \frac{n\theta - np_0}{\sqrt{np_0(1-p_0)}} ; H_0 \right) \tag{474}$$

$$\approx \mathbb{P}\left(N(0,1) > \frac{n\theta - np_0}{\sqrt{np_0(1-p_0)}} \right). \tag{475}$$

Thus we have (approximately for large n)

$$\text{type I error} \le \alpha \quad \iff \quad \frac{n\theta - np_0}{\sqrt{np_0(1-p_0)}} \ge z_\alpha \quad \iff \quad \theta \ge p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}. \tag{476}$$

Hence our optimal test for significance level α would be

$$\text{Reject } H_0 \quad \text{if } \hat{p}_n > p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}. \tag{477}$$

▲

A similar argument as above gives hypothesis testing about the null $H_0 : p = p_0$ against various alternatives $H_1 : p > p_0$, $H_1 : p < p_0$, and $H_1 : p \neq p_0$ population proportion against various alternative hypothesis.

Algorithm 3 Hypothesis testing for proportion

```
Input:
    Null hypothesis $H_0 : X \sim \text{Bernoulli}(p_0)$
    Alternative hypothesis $H_1$ : one of $p > p_0$ or $p < p_0$ or $p \neq p_0$, where $p = \mathbb{E}[X]$
    Observed sample: $x_1, x_2, \cdots, x_n$.
        Sample mean $\bar{x} = n^{-1}(x_1 + \cdots + x_n)$ and sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
    significance level: $\alpha \in (0,1)$
    Test statistic: $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$
Do: Compute the critical region associated with $\bar{x}$:
        $$\begin{cases} \hat{p}_n > p_0 + z_\alpha\sqrt{p_0(1-p_0)/n} & \text{if } H_1 : p > p_0 \\ \hat{p}_n < p_0 - z_\alpha\sqrt{p_0(1-p_0)/n} & \text{if } H_1 : p < p_0 \\ |\hat{p}_n - p_0| > z_{\alpha/2}\sqrt{p_0(1-p_0)/n} & \text{if } H_1 : p \neq p_0 \end{cases} \tag{478}$$
Do:
    If   Inside critical region, Reject $H_0$
    Else   Test inconclusive and cannot reject $H_0$
```

Example 8.3.2 (Excerpted from [HTZ77])**.** It was claimed that many commercially manufactured dice are not fair because the "spots" are really indentations, so that, for example, the 6-side is lighter than the1-side. Let p equal the probability of rolling a 6 with one of these dice. To test $H_0 : p = 1/6$ against the alternative hypothesis $H_1 : p > 1/6$, several such dice will be rolled to yield a total of $n = 8000$ observations. Let Y equal the number of times that 6 resulted in the 8000 trials. The test statistic is

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} = \frac{(Y/8000) - (1-6)}{\sqrt{(1/6)(5/6)/8000}}. \tag{479}$$

If we use a significance level of $\alpha = 0.05$, the critical region is

$$z > z_{0.05} = 1.645. \tag{480}$$

Suppose the results of the experiment yielded $y = 1389$ success. Then the observed value of the test statistic is

$$z = \frac{(1389/8000) - (1-6)}{\sqrt{(1/6)(5/6)/8000}} = 1.67 > 1.645. \tag{481}$$

Hence the null hypothesis is rejected at significance level $\alpha = 0.05$, and the experimental results indicate that these dice favor a 6 more than a fair die would. ▲

4. The Wilcoxon Tests

When we cannot verify normality of the unknown random variable subject to our test, and if we do not know its variance, then our hypothesis testing on its mean, which is based on either normality (so that we can compute the distribution of sample mean) or known variance (so that we can apply CLT), do not apply. In this case we may use 'non-parameteric' testing methods, which are essentially testing percentiles rather than mean or variance. The simplest of such is called the *sign test*, which is explained below.

Example 8.4.1 (Sign test)**.** Let X be a continuous RV of unknown distribution with median $m = m(X)$. We would like to test the null hypothesis $H_0 : m = m_0$ against the alternative hypothesis $H_1 : m > m_0$. Let $X_1, \cdots, X_n$ be i.i.d. samples of X. We will use the following statistic

$$W = \sum_{i=1}^{n} \mathbf{1}(X_i > m_0) = (\#X_i\text{'s s.t. } X_i > m_0). \tag{482}$$

If H_1 is true, then it is more likely to exceed m_0 than even. Hence it is reasonable to use the critical region of the form

$$C = \{(x_1, \cdots, x_n) \,|\, w > c\} \tag{483}$$

to reject H_0 in favor of H_1. Under the null hypothesis $H_0 : m = m_0$, each X_i exceeds its (hypothetical) median m_0 with probability $1/2$. Hence $W; H_0$ follows Binomial$(n, 1/2)$ distribution. This gives

$$\begin{aligned} \text{type I error} &= \mathbb{P}(\text{reject } H_0 : m = m_0\,;\, H_0) && (484)\\ &= \mathbb{P}(W > c\,;\, H_0) && (485)\\ &= \mathbb{P}\left(\text{Binomial}(n, 1/2) > c\right). && (486) \end{aligned}$$

While one can use a binomial table to compute the above probability for any given c, it is more convenient to use normal approximation by CLT when n is large. This gives

$$\begin{aligned} \text{type I error} &= \mathbb{P}\left(\text{Binomial}(n, 1/2) > c\right) && (487)\\ &\approx \mathbb{P}\left(N(0,1) > \frac{c - n/2}{\sqrt{n/4}}\right). && (488) \end{aligned}$$

Thus we have (approximately for large n)

$$\text{type I error} \le \alpha \quad \Longleftrightarrow \quad \frac{c - n/2}{\sqrt{n/4}} \ge z_\alpha \quad \Longleftrightarrow \quad c \ge \frac{n}{2} + z_\alpha \sqrt{n/4}. \tag{489}$$

Hence our optimal test for significance level α would be

$$\text{Reject } H_0 : m = m_0 \quad \text{if } w > \frac{n}{2} + z_\alpha \sqrt{n/4}. \tag{490}$$

For instance, suppose a random sample of size 20 yielded the following data

$$6.5 \quad 5.7 \quad 6.9 \quad 5.3 \quad 4.1 \quad 9.8 \quad 1.7 \quad 7.0 \quad 2.1 \quad 19.0 \tag{491}$$

$$18.9 \quad 16.9 \quad 10.4 \quad 44.1 \quad 2.9 \quad 2.4 \quad 4.8 \quad 18.9 \quad 4.6 \quad 7.9 \tag{492}$$

Consider $H_0 : m = 6.2$ and $H_1 : m > 6.2$. Then the (approximate) critical region for rejecting H_0 in favor of H_1 with significance level $\alpha = 0.05$ is given by

$$w > 10 + z_{0.05}\sqrt{20/4} = 10 + 1.645\sqrt{5} = 13.67. \tag{493}$$

Since there are only $w = 9$ samples that exceed the claimed median $m = 6.2$, we cannot reject H_0 in favor of H_1 at significance level 5%. ▲

One of the objection to the above sign test is that it does not leverage the information of magnitude of sample data, but only the sign of $X_i - m_0$. Below we present the *Wilcoxon signed rank test*, which combines the sign test with extra information about the magnitudes of the differences $X_i - m_0$ between data and claimed median.

Example 8.4.2 (Wilkcoxon's test)**.** Let X be a continuous RV of unknown distribution with median $m = m(X)$. We would like to test the null hypothesis $H_0 : m = m_0$ against the alternative hypothesis $H_1 : m > m_0$. Let $X_1, \cdots, X_n$ be i.i.d. samples of X. We will use the following *signed rank statistic*

$$W = \sum_{i=1}^{n} \operatorname{sgn}(X_i - m_0) \cdot \operatorname{Rank}(|X_i - m_0|), \tag{494}$$

where $\operatorname{sgn}(a) = 1$ if $a > 0$ and -1 if $a < 0$, and $\operatorname{Rank}(|X_i - m_0|)$ denotes the rank (from the smallest) of the magnitude $|X_i - m_0|$ among the n values $|X_1 - m_0|, |X_2 - m_0|, \cdots, |X_n - m_0|$.

For instance, suppose we have the following $n = 10$ sample values of X:

$$5.0 \quad 3.9 \quad 5.2 \quad 5.5 \quad 2.8 \quad 6.1 \quad 6.4 \quad 2.6 \quad 2.0 \quad 4.3 \tag{495}$$

Suppose we want to test $H_0 : m = 3.7$ against $H_1 : H_1 > 3.7$. Then

$$\begin{array}{rcccccccccc} x_i - m_0 : & 1.3 & 0.2 & 1.5 & 1.8 & -0.9 & 2.4 & 2.7 & -1.1 & -1.7 & 0.6 \\ |x_i - m_0| : & 1.3 & 0.2 & 1.5 & 1.8 & 0.9 & 2.4 & 2.7 & 1.1 & 1.7 & 0.6 \\ \text{Ranks} : & 5 & 1 & 6 & 8 & 3 & 9 & 10 & 4 & 7 & 2 \\ \text{Signed Ranks} : & 5 & 1 & 6 & 8 & -3 & 9 & 10 & -4 & -7 & 2 \end{array} \tag{496}$$

Then the observed value w of the signed rank statistic W is

$$w = 5 + 1 + 6 + 8 + (-3) + 9 + 10 + (-4) + (-7) + 2 = 27. \tag{497}$$

Notice that large values of W supports $H_1 : m > m_0$, since if the actual median m is larger than claimed median m_0, then each $X_i - m_0$ is more likely to be positive than negative. Hence, as in the case of the sign test, we use the critical region of the form

$$C = \{(x_1, \cdots, x_n) \,|\, w > c\} \tag{498}$$

to reject H_0 in favor of H_1.

In order to compute type I error of this test, we need to know the distribution of the signed rank statistic under the null hypothesis $H_0 : m = m_0$. Observe that, under H_0, each $X_i - m_0$ is either positive or negative with equal probability. Also, since we are adding up the signed ranks and order of summation does not matter, we can rewrite

$$W = \sum_{k=1}^{n} \xi_k k, \tag{499}$$

where $\xi_1, \cdots, \xi_k$ are i.i.d. RVs with $\mathbb{P}(\xi_1 = 1) = \mathbb{P}(\xi_1 = -1) = 1/2$. It is easy to see that $\mathbb{E}[W] = 0$ and $\operatorname{Var}(W) = \sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$. So if we have a normal approximation of W, we should have

$$\frac{W_n - \mathbb{E}[W_n]}{\sqrt{W_n}} = \frac{W_n}{\sqrt{n(n+1)(2n+1)/6}} \Longrightarrow N(0,1) \tag{500}$$

as $n \to \infty$. However, note that the increments $\xi_k k$ are independent, not but identically distributed. So standard CLT does not apply here. Nonetheless, a more general version of CLT (called Lyapunov's CLT) can be used to obtain a normal approximation of W (see Exercise 8.4.3) as above.

Now we can proceed in a standard way:

$$\begin{aligned} \text{type I error} &= \mathbb{P}(\text{reject } H_0 : m = m_0\,; H_0) && (501) \\ &= \mathbb{P}(W > c\,; H_0) && (502) \\ &= \mathbb{P}\left(\frac{W}{\sqrt{n(n+1)(2n+1)/6}} > \frac{c}{\sqrt{n(n+1)(2n+1)/6}} ; H_0 \right) && (503) \end{aligned}$$

$$\approx \mathbb{P}\left(N(0,1) > \frac{c}{\sqrt{n(n+1)(2n+1)/6}}\right). \tag{504}$$

Thus we have (approximately for large n)

$$\text{type I error} \le \alpha \quad \Longleftrightarrow \quad \frac{c}{\sqrt{n(n+1)(2n+1)/6}} \ge z_\alpha \Longleftrightarrow \quad c \ge z_\alpha \sqrt{n(n+1)(2n+1)/6}. \tag{505}$$

Hence our optimal test for significance level α would be

$$\text{Reject } H_0 : m = m_0 \quad \text{if } w \ge z_\alpha \sqrt{n(n+1)(2n+1)/6}. \tag{506}$$

For the example we considered above, we had $n = 10$ and $w = 25$. Noting that $z_{0.1} = 1.282$, the critical region for rejecting $H_0 : m = 3.7$ in favor of $H_1 : m > 3.7$ at significance level $\alpha = 0.1$ is given by

$$w > (1.282)\sqrt{10(11)(21)/6} = 25.147 \tag{507}$$

As our observed valued of signed rank statistic is $w = 27$, we reject $H_0 : m = 3.7$ in favor of $H_1 : m > 3.7$ at significance level $\alpha = 0.1$. ▲

Exercise 8.4.3 (Normal approximation of signed rank statistic)**.** Let $V_1, \cdots, V_n$ be independent RVs where for each $1 \le k \le n$,

$$\mathbb{P}(V_k = k) = \mathbb{P}(V_k = -k) = 1/2. \tag{508}$$

Define a RV $W_n = \sum_{k=1}^n V_k$.

(i) Show that $\mathbb{E}[W_n] = 0$ and $\mathrm{Var}(W_n) = \sum_{k=1}^n k^2 = n(n+1)(2n+1)/6$.

(ii) Show that for any $\delta > 0$,

$$\frac{\sum_{i=1}^n \mathbb{E}[|X_k - \mathbb{E}[X_k]|^{2+\delta}]}{\sqrt{\sum_{k=1}^n \mathrm{Var}(X_k)}^{2+\delta}} = \frac{\sum_{i=1}^n k^{2+\delta}}{\sqrt{n(n+1)(2n+1)/6}^{2+\delta}} \le \frac{n^{3+\delta}}{(n/3)^{3+3\delta/2}} \to 0 \quad \text{as } n \to \infty. \tag{509}$$

Hence W_n verifies the following *Lyapunov's condition*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}[|X_k - \mathbb{E}[X_k]|^{2+\delta}]}{\sqrt{\sum_{k=1}^n \mathrm{Var}(X_k)}^{2+\delta}} = 0 \quad \text{for some } \delta > 0. \tag{510}$$

Then Lyapunov's central limit theorem tells that as $n \to \infty$,

$$\frac{W_n - \mathbb{E}[W_n]}{\sqrt{\mathrm{Var}(W_n)}} = \frac{W_n}{\sqrt{n(n+1)(2n+1)/6}} \Longrightarrow N(0,1). \tag{511}$$

(Note that V_k's are independent but *not* identically distributed, so standard CLT does not apply to W_n.)

Exercise 8.4.4. The weights of the contents of $n_1 = 8$ and $n_2 = 8$ tins of cinnamon packaged by companies A and B, respectively, selected at random, yielded the following observations of X and Y:

$$\begin{array}{rcccccccc} x: & 117.1 & 121.3 & 127.8 & 121.9 & 117.4 & 124.5 & 119.5 & 115.1 \\ y: & 123.5 & 125.3 & 126.5 & 127.9 & 122.1 & 125.6 & 129.8 & 117.2 \end{array} \tag{512}$$

Let $T = X - Y$. As a means of comparing X and Y, we want to test $H_0 : m(T) = 0$ against $H_1 : m(T) < 0$.

(i) Compute the sample values of differences $x - y$. This gives a $n = 8$ sample for T.

(ii) Compute the signed rank statstic w for the sample of T in (i) for the null hypothesis $H_0 : m(T) = 0$.

(iii) Can we reject $H_0 : m(T) = 0$ in favor of $H_1 : m(T) < 0$ at significance level $\alpha = 0.05$? What does it imply on the original RVs X and Y?

5. Power of a statistical test

So far, we have derived statistical test based on the following procedure:

(i) Find a good test statistic (e.g., unbiased and sufficient). In most cases sample mean or sample frequency.

(ii) Derive the distribution of the test statistic under the null hypothesis (using normality of population or CLT for known variance).

(iii) By considering the null H_0 and the alternative H_1 hypotheses, form a critical region.

(iv) Compute type I error, and find the largest critical region such that type I error is at most a given significance level.

In this section, we will also consider type II error, which is the probability of not being able to reject H_0 when H_1 is true.

Example 8.5.1 (Power function of a test)**.** Let $X \sim \text{Bernoulli}(p)$ for unknown p. We have the null hypothesis $H_0 : p = p_0$ and the composite alternative $H_1 : p < p_0$. Let $X_1, \cdots, X_n$ be i.i.d. samples of X. We will use the test statistic $Y = \sum_{i=1}^{n} X_i$, which has $\text{Binomial}(n, p)$ distribution. Since $\mathbb{E}[Y] = np$, if Y is much less than np_0, this would support H_1. Hence our critical region will be of the form

$$C = \{(x_1, \cdots, x_n) \,|\, y < c\}. \tag{513}$$

From the argument in Example 8.3.1, we know that

$$\text{Type I error} \le \alpha \quad \Longleftrightarrow \quad c \le np_0 - z_\alpha \sqrt{np_0(1-p_0)}. \tag{514}$$

So for a given significance level α, we will choose our critical region by setting

$$c = c_\alpha := np_0 - z_\alpha \sqrt{p_0(1-p_0)n}. \tag{515}$$

Now, what about the type II error? Since $H_1 : p < p_0$ is a composite hypothesis, we cannot compute the distribution of the test statistic Y under H_1. However, we can compute the probability of not being able to reject $H_0 : p = p_0$ according to the critical region we found above assuming a specific value of $p = p_1 < p_0$. Namely, note that

$$\mathbb{P}(\text{not reject } H_0 \,|\, p = p_1) = \mathbb{P}(Y \ge c_\alpha \,|\, p = p_1) \tag{516}$$

$$= \sum_{c_\alpha \le k \le n} \binom{n}{k} p_1^k (1-p_1)^{n-k}. \tag{517}$$

Notice that the last expression is a function in the assumed value p_1 of the true success probability p. We may as well consider the complement of the above probability:

$$K(p_1) := \mathbb{P}(\text{reject } H_0 \,|\, p = p_1) = \mathbb{P}(Y < c_\alpha \,|\, p = p_1) \tag{518}$$

$$= \sum_{0 \le k \le c_\alpha} \binom{n}{k} p_1^k (1-p_1)^{n-k}. \tag{519}$$

The above function in the assumped value p_1 of p is called the *power function* of the test. Each value $K(p_1)$ is called the *power* of the test at $p = p_1$.

For instance, suppose $n = 20$, and we have the following test

$$\text{Reject } H_0 : p = 1/2 \text{ in favor of } H_1 : p < 1/2 \quad \Longleftrightarrow \quad \{y = x_1 + \cdots + x_n \le 6\}. \tag{520}$$

Then the power function of this test is given by

$$K(p) = \mathbb{P}\left(\text{Binomial}(20, p) \le 6\right) = \sum_{0 \le k \le 6} \binom{20}{6} p^k (1-p)^{20-k} \tag{521}$$

$$\approx \mathbb{P}\left(N(0,1) \le \frac{6 - 20p}{\sqrt{20p(1-p)}}\right), \tag{522}$$

where the last expression comes from the normal approximation. Either using the exact binomial probability or the approximate normal probability, This power function entails type I error, and also type II error for simple alternatives. Namely,

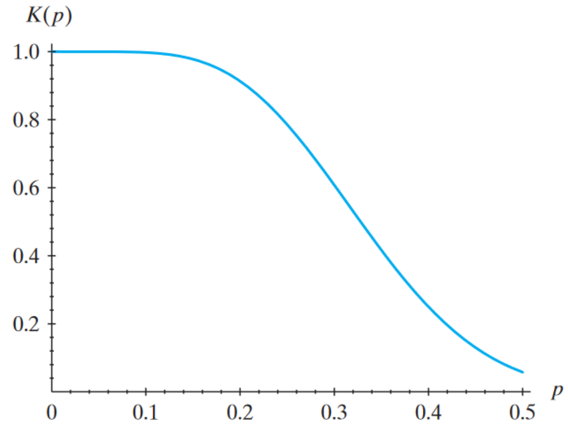


FIGURE 4. Plot of the power function $K(p) = \mathbb{P}(\text{Binomial}(20, p) \le 6; p)$

$$\text{Type I error} = \mathbb{P}(\text{reject } H_0 \,|\, p = 1/2) = K(1/2) = \sum_{0\le k\le 6} \binom{20}{6} (1/2)^k (1/2)^{20-k} = 0.0577. \tag{523}$$

Also, if we were to use the simple alternative $H_1 : p = 1/10$, then

$$\text{Type II error} = \mathbb{P}(\text{not reject } H_0 \,|\, p = 1/10) = 1 - \mathbb{P}(\text{reject } H_0 \,|\, p = 1/10) \tag{524}$$

$$= 1 - K(1/10) = 1 - \sum_{0\le k\le 6} \binom{20}{6} (1/10)^k (9/10)^{20-k} = 1 - 0.9976 = 0.0024. \tag{525}$$

▲

Example 8.5.2 (Determining critical region and sample size from power function)**.** Let $X \sim \text{Bernoulli}(p)$ for unknown p. We have the null hypothesis $H_0 : p = 2/3$ and the composite alternative $H_1 : p = 1/3$. Let $X_1, \cdots, X_n$ be i.i.d. samples of X. We will use the test statistic $Y = \sum_{i=1}^{n} X_i$, which has Binomial(n, p) distribution. Since $\mathbb{E}[Y] = np$, if Y is much less than $2n/3$, this would support H_1. Hence our critical region will be of the form

$$C = \{(x_1, \cdots, x_n) \,|\, y < c\}. \tag{526}$$

There are two parameters in this test; the rejection threshold c and the sample size n. We will choose these parameters such that

$$\text{Type I error} \le 0.05 \qquad \text{and} \qquad \text{Type II error} \le 0.1. \tag{527}$$

From the discussion in Example 8.5.1, the power function of the above test is given by

$$K(p) = \mathbb{P}(\text{reject } H_0 \,;\, p) \tag{528}$$

$$= \mathbb{P}(Y < c;\, p) \tag{529}$$

$$= \mathbb{P}\left(\text{Binomial}(n, p) \le c\right) \tag{530}$$

$$\approx \mathbb{P}\left(N(0, 1) \le \frac{c - np}{\sqrt{np(1-p)}}\right). \tag{531}$$

From the error conditions, we have

$$\begin{cases} 0.05 \geq \text{Type I error} = K(2/3) \approx \mathbb{P}\left(N(0,1) \leq \frac{c-2n/3}{\sqrt{2n/9}}\right) \\ 0.1 \geq \text{Type II error} = 1 - K(1/3) \approx \mathbb{P}\left(N(0,1) > \frac{c-n/3}{\sqrt{2n/9}}\right) \end{cases} \tag{532}$$

Solving these, we get

$$\frac{c-2n/3}{\sqrt{2n/9}} \leq -z_{0.05} \quad \text{and} \quad \frac{c-n/3}{\sqrt{2n/9}} \geq z_{0.1}. \tag{533}$$

An optimal test is achieved if we choose c and n so that the above inequalites are equalities. This gives

$$c \leq (2n/3) - z_{0.05}\sqrt{2n/9} \tag{534}$$

$$c \geq (n/3) + z_{0.1}\sqrt{2n/9}, \tag{535}$$

so n has to satisfy

$$(n/3) + z_{0.1}\sqrt{2n/9} \leq (2n/3) - z_{0.05}\sqrt{2n/9}. \tag{536}$$

Simplifying, we get

$$\frac{n}{3} \geq \sqrt{2n/9}(z_{0.05} + z_{0.1}), \tag{537}$$

so squaring both sides gives

$$n \geq 2(z_{0.05} + z_{0.1})^2 = 2(1.645 + 1.282)^2 = 17.606. \tag{538}$$

Thus the minimial required sample size is $n = 18$. On the other hand, for c this gives

$$(18/3) + 1.282\sqrt{2 \cdot 18/9} \leq (2 \cdot 18/3) - 1.645\sqrt{2 \cdot 18/9}, \tag{539}$$

so

$$8.564 \leq c \leq 8.71. \tag{540}$$

We may choose any value of c within this range, say, $c = 8.6$. Then the test

$$\text{Reject } H_0 : p = 2/3 \text{ in favor of } H_1 : p = 1/3 \quad \iff \quad \{y = x_1 + \cdots + x_{18} \leq 8.6\}. \tag{541}$$

has type I error ≤ 0.05 and type II error ≤ 0.1. ▲

Exercise 8.5.3 (Paired test, known variance of the difference)**.** Let X, Y be RVs. Denote $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_Y$. Suppose we want to test the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ against the alternative hypothesis $H_1 : \mu_X - \mu_Y = -3$. Suppose we have i.i.d. pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the joint distribution of (X, Y). Further assume that we know the value of $\sigma^2 = \mathrm{Var}(X - Y)$.

(i) Noting that $X_1 - Y_1, \cdots, X_n - Y_n$ are i.i.d. with finite variance σ^2, show using CLT that, as $n \to \infty$,

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma/\sqrt{n}} \Longrightarrow N(0,1). \tag{542}$$

(ii) Our critical region for rejecting the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ in favor of $H_1 : \mu_X - \mu_Y = -3$ will be of the form

$$C = \{(x_1, \cdots, x_n, y_1, \cdots, y_m) \,|\, \bar{x} - \bar{y} \leq c\}, \tag{543}$$

for some rejection threshold $c > 0$. Show that the power function of this test is given by

$$K(\mu_D) = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y \,;\, \mu_X - \mu_Y = \mu_D) \tag{544}$$

$$= \mathbb{P}(\bar{X} - \bar{Y} \leq c \,;\, \mu_X - \mu_Y = \mu_D) \tag{545}$$

$$= \mathbb{P}\left(Z \leq \frac{c - \mu_D}{\sigma/\sqrt{n}} \,;\, \mu_X - \mu_Y = \mu_D\right) \approx \mathbb{P}\left(N(0,1) \leq \frac{c - \mu_D}{\sigma/\sqrt{n}}\right). \tag{546}$$

(iii) Conclude that (approximately for large n)

$$\text{Type I error} \le \alpha \iff \frac{c}{\sigma/\sqrt{n}} \le -z_\alpha, \tag{547}$$

$$\text{Type II error} \le \beta \iff \frac{c+3}{\sigma/\sqrt{n}} \ge z_\beta. \tag{548}$$

(iv) Suppose we want both the type I and II errors to be at most 0.05. Find conditions for c and n that would guarantee this.

Exercise 8.5.4 (Paired test, known normality of the difference)**.** Let X, Y be RVs. Denote $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_Y$. Suppose we want to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X - \mu_Y = -3$. Suppose we have i.i.d. pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the joint distribution of (X, Y). Further assume that we know the $X - Y$ follows a normal distribution.

(i) Noting that $X_1 - Y_1, \cdots, X_n - Y_n$ are i.i.d. with normal distribution, show that (exactly)

$$T := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S/\sqrt{n}} \sim t(n-1), \tag{549}$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n ((X_i - Y_i) - (\bar{X} - \bar{Y}))^2$ denotes the sample variance.

(ii) Our critical region for rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$ in favor of $H_1 : \mu_X - \mu_Y = -3$ will be of the form

$$C = \left\{ (x_1, \cdots, x_n, y_1, \cdots, y_m) \,\middle|\, \frac{\bar{x} - \bar{y}}{s/\sqrt{n}} \le c \right\}, \tag{550}$$

for some rejection threshold $c < 0$. Show that the power function of this test is given by

$$K(\mu_D) = \mathbb{P}(\text{reject } H_0 : \mu_X = \mu_Y\,;\, \mu_X - \mu_Y = \mu_D) \tag{551}$$

$$= \mathbb{P}\left(\frac{\bar{X} - \bar{Y}}{S/\sqrt{n}} \le c\,;\, \mu_X - \mu_Y = \mu_D \right) \tag{552}$$

$$= \mathbb{P}\left(\frac{(\bar{X} - \bar{Y}) - \mu_D}{S/\sqrt{n}} \le c - \frac{\mu_D}{S/\sqrt{n}}\,;\, \mu_X - \mu_Y = \mu_D \right) = \mathbb{P}\left(t(n-1) \le c - \frac{\mu_D}{S/\sqrt{n}} \right). \tag{553}$$

(iii) Conclude that

$$\text{Type I error} \le \alpha \iff c \le -t_\alpha(n-1), \tag{554}$$

$$\text{Type II error} \le \beta \iff c \ge t_\beta(n-1) - \frac{3}{s/\sqrt{n}}. \tag{555}$$

(iv) Suppose we want both the type I and II errors to be at most 0.05. Find (numerically) conditions for c and n that would guarantee this.

6. Best critical regions

Suppose we want to design a test for the null hypothesis H_0 and the alternative hypothesis H_1 on an unknown RV of interest X, based on n i.i.d. samples $X_1, X_2, \cdots, X_n$ of X. Our decision rule will be of the form

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } (X_1, \cdots, X_n) \in C, \tag{556}$$

where C is a critical region of our choice. Recall that the type I and type II errors of this test are given by

$$\text{Type I error} = \mathbb{P}((X_1, \cdots, X_n) \in C; H_0), \tag{557}$$

$$\text{Type II error} = \mathbb{P}((X_1, \cdots, X_n) \notin C; H_1). \tag{558}$$

Suppose we want significance level of α. In other words, we need to choose the critical region C such that

$$\mathbb{P}((X_1, \cdots, X_n) \in C; H_0) = \alpha. \tag{559}$$

However, there could be many choices of such critical region C. Among such, it would desirable to choose the one that gives the smallest type II error, that is, $\mathbb{P}((X_1,\cdots,X_n)\in C;H_1)$, which is also called the *size* of C, is *maximized.* This leads us to the following definition of a *best critical region.*

Definition 8.6.1 (Best critical region)**.** Consider the test of the null hypothesis H_0 against the simple alternative hypothesis H_1. Let C be a critical region of size α. Then C is a *best critical region of size* α if, for every other critical region D of size α, we have

$$\mathbb{P}((X_1,\cdots,X_n)\in C;H_1)\geq\mathbb{P}((X_1,\cdots,X_n)\in D;H_1). \tag{560}$$

That is, when H_1 is true, the probability of rejecting H_0 with the use of the critical region C is at least as great as the corresponding probability with the use of any other critical region D of size α.

Thus, a best critical region of size α is the critical region that has the greatest power among all critical regions of size α. The Neyman-Pearson lemma gives sufficient conditions for a best critical region of size α.

Theorem 8.6.2 (Neyman-Pearson Lemma)**.** *Let X be a RV of pdf or pmf $f_X(x;\theta)$, depending on an unknown parameter $\theta\in\{\theta_0,\theta_1\}$. Let $X_1,\cdots,X_n$ be i.i.d. samples of X. Denote the joint likelihood function*

$$L(\theta)=L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^n f_X(x_i;\theta). \tag{561}$$

Suppose that there exists a positive constant k and a subset C of the sample space such that

(a) $\mathbb{P}((X_1,\cdots,X_n)\in C;\theta=\theta_0)=\alpha\in(0,1)$,

(b) $\dfrac{L(\theta_0)}{L(\theta_1)}\leq k$ *for* $(x_1,\cdots,x_n)\in C$, *and*

(c) $\dfrac{L(\theta_0)}{L(\theta_1)}\geq k$ *for* $(x_1,\cdots,x_n)\notin C$.

Then C is a best critical region of size α for testing the simple null hypothesis $H_0:\theta=\theta_0$ against the simple alternative hypothesis $H_1:\theta=\theta_1$.

PROOF. (Optional*) Denote $\mathbf{X}=(X_1,\cdots,X_n)$. Fix any critical region D of size α. We would like to show that

$$\mathbb{P}(\mathbf{X}\in C;\theta=\theta_1)=\mathbb{P}(\mathbf{X}\in D;\theta=\theta_1). \tag{562}$$

We first write

$$\mathbb{P}(\mathbf{X}\in C;\theta)=\mathbb{P}(\mathbf{X}\in C\cap D;\theta)+\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta) \tag{563}$$

$$\mathbb{P}(\mathbf{X}\in D;\theta)=\mathbb{P}(\mathbf{X}\in D\cap C;\theta)+\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta). \tag{564}$$

Since both critical regions C and D have size α, we have

$$\begin{aligned}0=\alpha-\alpha&=\mathbb{P}(\mathbf{X}\in C;\theta=\theta_0)-\mathbb{P}(\mathbf{X}\in D;\theta=\theta_0) &(565)\\ &=\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_0)-\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_0). &(566)\end{aligned}$$

Also, from condition (a), $L(\theta_0)\leq kL(\theta_1)$ for all points in C, and hence in $C\cap D^c$. So we get

$$\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_0)\leq k\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_1). \tag{567}$$

On the other hand, from condition (b), $L(\theta_0)\geq kL(\theta_1)$ for all points not in C, and hence in $D\cap C^c$. So

$$\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_0)\geq k\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_1). \tag{568}$$

Now combining the above equation and inequalities above,

$$\begin{aligned}\mathbb{P}(\mathbf{X}\in C;\theta=\theta_1)-\mathbb{P}(\mathbf{X}\in D;\theta=\theta_1)&=\mathbb{P}(\mathbf{X}\in C\cap D;\theta=\theta_1)+\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_1) &(569)\\ &\geq\mathbb{P}(\mathbf{X}\in C\cap D;\theta=\theta_1)+\frac{1}{k}\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_0). &(570)\end{aligned}$$

Similarly, we have

$$\mathbb{P}(\mathbf{X}\in D;\theta=\theta_1)-\mathbb{P}(\mathbf{X}\in D;\theta=\theta_1)=\mathbb{P}(\mathbf{X}\in D\cap C;\theta=\theta_1)+\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_1) \tag{571}$$

$$\leq \mathbb{P}(\mathbf{X}\in D\cap C;\theta=\theta_1)+\frac{1}{k}\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_0). \tag{572}$$

Taking the difference, we obtain

$$\mathbb{P}(\mathbf{X}\in C;\theta=\theta_1)-\mathbb{P}(\mathbf{X}\in D;\theta=\theta_1)\geq\frac{1}{k}\left(\mathbb{P}(\mathbf{X}\in C\cap D^c;\theta=\theta_0)-\mathbb{P}(\mathbf{X}\in D\cap C^c;\theta=\theta_0)\right)=0, \tag{573}$$

as desired. □

Example 8.6.3 (Excerpted from [HTZ77])**.** Let $X_1, X_2, \cdots, X_n$ be a random sample from the normal distribution $N(\mu, 36)$. We shall find the best critical region for testing the simple hypothesis $H_0 : \mu = 50$ against the simple alternative hypothesis $H_1 : \mu = 55$. Using the ratio of the likelihood functions, namely, $L(50)/L(55)$, we shall find those points in the sample space for which this ratio is less than or equal to some positive constant k.

We first compute the ratio of the likelihood functions:

$$\frac{L(50)}{L(55)}=\frac{(2\pi\cdot 36)^{-n/2}\exp\left[-\frac{1}{2\cdot 36}\sum_{i=1}^n(x_i-50)^2\right]}{(2\pi\cdot 36)^{-n/2}\exp\left[-\frac{1}{2\cdot 36}\sum_{i=1}^n(x_i-55)^2\right]} \tag{574}$$

$$=\exp\left[-\frac{1}{72}\sum_{i=1}^n(x_i-50)^2-(x_i-55)^2\right] \tag{575}$$

$$=\exp\left[-\frac{1}{72}\sum_{i=1}^n(2x_i-105)(5)\right]. \tag{576}$$

Then note that

$$\frac{L(50)}{L(55)}\leq k \iff \log\frac{L(50)}{L(55)}\leq\log k \iff -\frac{5}{72}\sum_{i=1}^n(2x_i-105)\leq\log k. \tag{577}$$

The last condition is equivalent to

$$\sum_{i=1}^n x_i\geq\frac{105n}{2}-\frac{36}{5}\log k. \tag{578}$$

Thus we have shown that

$$\frac{L(50)}{L(55)}\leq k \iff \bar{x}\geq c:=\frac{105}{2}-\frac{36}{5n}\log k. \tag{579}$$

Hence according to the Neyman-Pearson lemma, a best critical region is given by

$$C=\{(x_1,\cdots,x_n)\,|\,\bar{x}\geq c\}, \tag{580}$$

where c is chosen so that the size of the critical region (i.e., type I error) is α.

For instance, suppose $n = 16$ and $\alpha = 0.05$. Then we will choose c so that

$$0.05\geq\mathbb{P}(\bar{X}\geq c;\mu=50) \tag{581}$$

$$=\mathbb{P}\left(\frac{\bar{X}-50}{6/\sqrt{16}}\geq\frac{c-50}{6/\sqrt{16}};\mu=50\right)=\mathbb{P}\left(N(0,1)\geq\frac{c-50}{3/2}\right). \tag{582}$$

Thus we would want $(c-50)/(3/2)\geq z_{0.05}=1.645$. This gives $c\geq 50+(3/2)(1.645)\approx 52.47$. According to the previous discussion, the following critical region

$$C=\{(x_1,\cdots,x_n)\,|\,\bar{x}\geq 52.47\} \tag{583}$$

is a best critical region for rejecting $H_0 : \mu = 50$ against $H_1 : \mu = 55$ at significance level $\alpha = 0.05$. ▲

Exercise 8.6.4. Let $X_1, \cdots, X_n$ be i.i.d. samples from Poisson(λ). We want to test the null hypothesis $H_0 : \lambda$=2 against the alternative hypothesis $H_1 : \lambda = 5$.

(i) Compute the ratio $L(2)/L(5)$ of joint likelihood functions. Show that

$$\frac{L(2)}{L(5)} \le k \quad \Longleftrightarrow \quad \left(\sum_{i=1}^{n} x_i\right) \log(2/5) + 3n \le \log k. \tag{584}$$

(ii) Using the Neyman-Pearson lemma, conclude that a best critical region C for testing $H_0 : p = 2$ against $H_1 : p = 5$ is given by

$$C = \left\{ (x_1, \cdots, x_n) \,\middle|\, \sum_{i=1}^{n} x_i \ge c \right\}. \tag{585}$$

(iii) Let $n = 30$. Use CLT and (ii) and obtain a best critical region C for testing $H_0 : p = 2$ against $H_1 : p = 5$ at significance level $\alpha = 0.05$.

Definition 8.6.5. A test defined by a critical region C of size α is a *uniformly most powerful test* if it is a most powerful test against each simple alternative in H_1. The critical region C is called a uniformly most powerful critical region of size α.

Example 8.6.6 (Excerpted from [HTZ77])**.** Let $X_1, X_2, \cdots, X_n$ be a random sample from the normal distribution $N(\mu, 36)$. We shall find the best critical region for testing the simple hypothesis $H_0 : \mu = 50$ against the simple alternative hypothesis $H_1 : \mu = \mu_1$, where $\mu_1 > 50$.

We first compute the ratio of the likelihood functions:

$$\frac{L(50)}{L(\mu_1)} = \frac{(2\pi \cdot 36)^{-n/2} \exp\left[-\frac{1}{2\cdot 36}\sum_{i=1}^{n}(x_i - 50)^2\right]}{(2\pi \cdot 36)^{-n/2} \exp\left[-\frac{1}{2\cdot 36}\sum_{i=1}^{n}(x_i - \mu_1)^2\right]} \tag{586}$$

$$= \exp\left[-\frac{1}{72}\sum_{i=1}^{n}(x_i - 50)^2 - (x_i - \mu_1)^2\right] \tag{587}$$

$$= \exp\left[-\frac{1}{72}\sum_{i=1}^{n}(2x_i - 50 - \mu_1)(50 - \mu_1)\right]. \tag{588}$$

Then note that

$$\frac{L(50)}{L(\mu_1)} \le k \quad \Longleftrightarrow \quad \log\frac{L(50)}{L(\mu_1)} \le \log k \quad \Longleftrightarrow \quad -\frac{50 - \mu_1}{72}\sum_{i=1}^{n}(2x_i - 50 - \mu_1) \le \log k. \tag{589}$$

Since $\mu_1 > 50$, the last condition is equivalent to

$$\sum_{i=1}^{n} x_i \ge \frac{(50 + \mu_1)n}{2} - \frac{36}{50 - \mu_1}\log k. \tag{590}$$

Thus we have shown that

$$\frac{L(50)}{L(\mu_1)} \le k \quad \Longleftrightarrow \quad \bar{x} \ge c := \frac{50 + \mu_1}{2} - \frac{36}{(50 - \mu_1)n}\log k. \tag{591}$$

Hence according to the Neyman-Pearson lemma, a best critical region is given by

$$C = \{(x_1, \cdots, x_n) \,|\, \bar{x} \ge c\}, \tag{592}$$

where c is chosen so that the size of the critical region (i.e., type I error) is α. It is important to notice that the chosen value of c depend only on α, although the value of k depends on μ_1.

For instance, suppose $n = 16$ and $\alpha = 0.05$. Then we will choose c so that

$$0.05 \ge \mathbb{P}(\bar{X} \ge c; \mu = 50) \tag{593}$$

$$= \mathbb{P}\left(\frac{\bar{X} - 50}{6/\sqrt{16}} \ge \frac{c - 50}{6/\sqrt{16}}; \mu = 50\right) = \mathbb{P}\left(N(0,1) \ge \frac{c - 50}{3/2}\right). \tag{594}$$

Thus we would want $(c-50)/(3/2) \geq z_{0.05} = 1.645$. This gives $c \geq 50 + (3/2)(1.645) \approx 52.47$. According to the previous discussion, the following critical region

$$C = \{(x_1, \cdots, x_n) \,|\, \bar{x} \geq 52.47\} \tag{595}$$

is a best critical region for rejecting $H_0 : \mu = 50$ against $H_1 : \mu = \mu_1$ at significance level $\alpha = 0.05$. Moreover, this holds for all $\mu_1 > 50$. Hence C is a uniformly most powerful critical region of size α. ▲

Exercise 8.6.7. Let $X_1, \cdots, X_n$ be i.i.d. samples from Bernoulli(p). We want to find a uniformly most powerful test for the null hypothesis $H_0 : p = p_0$ against the one-sided alternative hypothesis $H_1 : p = p_1 > p_0$.

(i) Compute the ratio $L(p_0)/L(p_1)$ of joint likelihood functions, where $p_1 > p_0$. Show that

$$\frac{L(p_0)}{L(p_1)} \leq k \quad \Longleftrightarrow \quad \left[\frac{p_0(1-p_1)}{p_1(1-p_0)}\right]^{\sum_{i=1}^n x_i} \left[\frac{1-p_0}{1-p_1}\right]^n \leq k \tag{596}$$

$$\Longleftrightarrow \quad \sum_{i=1}^n x_i \geq c := \frac{\log k - n \log[(1-p_0)/(1-p_1)]}{\log[p_0(1-p_1)/p_1(1-p_0)]}. \tag{597}$$

(ii) Using the Neyman-Pearson lemma, conclude that a best critical region C for testing $H_0 : p = p_0$ against $H_1 : p = p_1 > p_0$ is given by

$$C = \left\{ (x_1, \cdots, x_n) \,\middle|\, \sum_{i=1}^n x_i \geq c \right\}. \tag{598}$$

(iii) Let $n = 30$. Use CLT and (ii) and obtain a best critical region C for testing $H_0 : p = 0.2$ against $H_1 : p = 0.5$ at significance level $\alpha = 0.05$.

Remark 8.6.8. If a sufficient statistic $Y = u(X_1, \cdots, X_n)$ exists for an unknown paramter θ, then by the factorization theorem (Thm 5.2.1), we have

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{\phi\big(T(x_1, \cdots, x_n); \theta_0\big)\, g(x_1, \cdots, x_n)}{\phi\big(T(x_1, \cdots, x_n); \theta_1\big)\, g(x_1, \cdots, x_n)} \tag{599}$$

$$= \frac{\phi\big(T(x_1, \cdots, x_n); \theta_0\big)}{\phi\big(T(x_1, \cdots, x_n); \theta_1\big)}. \tag{600}$$

Thus $L(\theta_0)/L(\theta_1) \leq k$ provides a critical region that is a function of the observations $x_1, x_2, \cdots, x_n$ only through the observed value of the sufficient statistic $y = u(x_1, x_2, \cdots, x_n)$. Hence, best critical and uniformly most powerful critical regions (at least the ones given by the Neyman-Pearson lemma) are based upon sufficient statistics when they exist.

7. Likelihood ratio tests

In this section, we develop a general framework of deriving statistical tests for composite (both null and alternative) hypothesis.

Consider an unknown RV X with distribution $f_X(x;\theta)$ that has unknown parameter(s). Let Ω denote the space of all possible values of θ (e.g., set of all weights in a deep neural network). Suppose we are given a sample $(x_1, x_2, \cdots, x_n)$ of X. For each subset $\Omega' \subseteq \Omega$ of the parameter space, we define its *maximum likelihood* as

$$L(\Omega') = L(x_1, \cdots, x_n; \Omega') := \sup_{\theta' \in \Omega'} L(x_1, \cdots, x_n; \theta = \theta'). \tag{601}$$

Namely, this is the largest likelihood of the sample $(x_1, \cdots, x_n)$ assuming the parameters can take values only inside the subset Ω'. The key observation here is that

$$L(x_1, \cdots, x_n; \Omega') \text{ large} \quad \Longleftrightarrow \quad \Omega' \text{ can well-explain the given sample} \tag{602}$$

We now introduce the likelihood ratio test. We take two disjoint subsets Ω_0 and Ω_1 of the parameter space Ω. Our null and alternative hypotheses are then formulated as

$$H_0 : \theta \in \Omega_0, \qquad H_1 : \theta \in \Omega_1. \tag{603}$$

We define the *likelihood ratio* associated with the subset Ω' and sample $(x_1, \cdots, x_n)$ by

$$\lambda = \frac{L(\Omega_0)}{L(\Omega_1)} = \frac{L(x_1, \cdots, x_n; \Omega')}{L(x_1, \cdots, x_n; \Omega)}. \tag{604}$$

The critical region for the likelihood ratio test is given by

$$\text{Reject } H_0 : \theta \in \Omega_0 \text{ in favor of } H_1 : \theta \in \Omega_1 \quad \Longleftrightarrow \quad \lambda = \frac{L(\Omega_0)}{L(\Omega_1)} \le k, \tag{605}$$

where k is some constant that will be chosen so that the associate type I error is within a desired confidence level. The rationale behind this likelihood ratio test is the following. If Ω_0 is a much less likely set of values of parameters than Ω_1 (that has weaker explanation power of the given sample), then we should have $L(\Omega_0) \ll L(\Omega_1)$, or, equivalently, $\lambda \ll 1$. This is why we form the critical region of the form $\{\lambda \le k\}$.

In this section, we will mostly consider the case of 'complementary hypotheses', $\Omega_1 = \Omega \setminus \Omega_1$. (For instance, $\mu = 0$ v.s. $\mu \neq 0$). In this case, the critical region for the likelihood ratio test simplifies to

$$\text{Reject } H_0 : \theta \in \Omega_0 \text{ in favor of } H_1 : \theta \notin \Omega_0 \quad \Longleftrightarrow \quad \lambda = \frac{L(\Omega_0)}{L(\Omega)} \le k. \tag{606}$$

See Exercise 8.7.1 for a justification.

Exercise 8.7.1 (Likelihood ratio test for complementary hypotheses)**.** Fix $\Omega_0 \subseteq \Omega$ and let $\Omega_1 = \Omega \setminus \Omega_0$. Fix a constant $0 < k < 1$. Show the following equivalences:

$$\frac{L(\Omega_0)}{L(\Omega_1)} \le k \quad \Longleftrightarrow \quad L(\Omega_0) \le kL(\Omega \setminus \Omega_0) \tag{607}$$

$$\Longleftrightarrow \quad L(\Omega_0) \le kL(\theta') \text{ for some } \theta' \notin \Omega_0 \quad \Longleftrightarrow \quad L(\Omega_0) \le kL(\Omega) \quad \Longleftrightarrow \quad \frac{L(\Omega_0)}{L(\Omega)} \le k \tag{608}$$

Example 8.7.2 (Excerpted from [HTZ77])**.** Let $X_1, X_2, \cdots, X_n$ be a random sample from the normal distribution $N(\mu, 5)$. We shall find the best critical region for testing the simple hypothesis $H_0 : \mu = 162$ against the composite alternative hypothesis $H_1 : \mu \neq 162$. Here we have a single parameter μ, so our parameter space is $\Omega = \mathbb{R}$. To compute the maximum likelihood function of the null hypothesis, note that

$$L(162) = (2\pi \cdot 5)^{-n/2} \exp\left[-\frac{1}{2 \cdot 5} \sum_{i=1}^{n} (x_i - 162)^2 \right]. \tag{609}$$

On the other hand, the full maximum likelihood is given by

$$L(\Omega) = \sup_{\theta \in \mathbb{R}} (2\pi \cdot 5)^{-n/2} \exp\left[-\frac{1}{2 \cdot 5} \sum_{i=1}^{n} (x_i - \theta)^2 \right]. \tag{610}$$

By a computation we made in Example 3.1.6, we know that the full maximum likelihood in this Gaussian case is achieved when $\theta = \bar{x} = n^{-1}(x_1 + \cdots + x_n)$. Hence

$$L(\Omega) = L(\bar{x}) = (2\pi \cdot 5)^{-n/2} \exp\left[-\frac{1}{2 \cdot 5} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]. \tag{611}$$

Thus the likelihood ratio is given by

$$\lambda = \frac{L(162)}{L(\bar{x})} = \frac{(2\pi \cdot 5)^{-n/2} \exp\left[-\frac{1}{2 \cdot 5} \sum_{i=1}^{n} (x_i - 162)^2 \right]}{(2\pi \cdot 5)^{-n/2} \exp\left[-\frac{1}{2 \cdot 5} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]} \tag{612}$$

$$= \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(x_i-162)^2-(x_i-\bar{x})^2\right] \tag{613}$$

$$= \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(162-\bar{x})^2\right] = \exp\left[-\frac{n}{10}(162-\bar{x})^2\right], \tag{614}$$

where we have used the 'Pythagorian theorem'

$$\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-a)^2 = \left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)+n(a-\bar{x})^2 \tag{615}$$

with $a = 162$.

On the one hand, a value of x close to 162 would tend to support H_0, and in that case λ is close to 1. On the other hand, an $\bar{x}$ that differs from 162 by too much would tend to support H_1. See Figure 5 for an illustration.

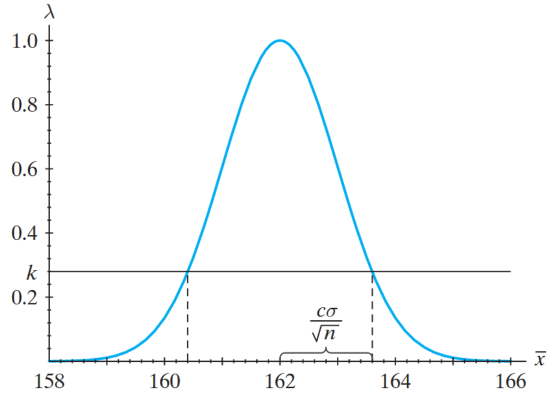


FIGURE 5. The Likelihood ratio test for testing $H_0 : \mu = 162$ against $H_1 : \mu \neq 162$.

Now we compute the critical region, which is given by

$$\lambda = \exp\left[-\frac{n}{10}(162-\bar{x})^2\right] \leq k \iff -\frac{n}{10}(162-\bar{x})^2 \leq \log k \tag{616}$$

$$\iff |162-\bar{x}| \geq c := \sqrt{-\frac{10}{n}\log k}. \tag{617}$$

In order to guarantee significance level α, we compute the Type I error of this test, using the fact that

$$\bar{X} \sim N(\mu, 5/n) \overset{H_0}{=} N(162, 5/n), \tag{618}$$

noting that the i.i.d. samples $X_1, \cdots, X_n$ are drawn from a normal distribution $N(\mu, 5)$ and H_0 assumes $\mu = 162$. This gives

$$\text{Type I error} = \mathbb{P}\left(\text{reject } H_0; H_0\right) \tag{619}$$

$$= \mathbb{P}\left(|\bar{X}-162| \geq c\right) \tag{620}$$

$$= \mathbb{P}\left(\frac{|\bar{X}-162|}{\sqrt{5/n}} \geq \frac{c}{\sqrt{5/n}}\right) = \mathbb{P}\left(|N(0,1)| \geq \frac{c}{\sqrt{5/n}}\right). \tag{621}$$

Thus we have

$$\text{Type I error} \leq \alpha \iff \frac{c}{\sqrt{5/n}} \geq z_{\alpha/2}. \tag{622}$$

In summary, the critical region of the likelihood ratio test for testing $H_0 : \mu = 162$ against $H_1 : \mu \neq 162$ at significance level α is given by

$$C = \left\{ (x_1, \cdots, x_n) : |\bar{x} - 162| \geq \sqrt{5/n} z_{\alpha/2} \right\}. \tag{623}$$

This is precisely the z-test we had before in Algorithm 2. ▲

Example 8.7.3 (Testing mean for unknown variance)**.** Let $X_1, X_2, \cdots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Here we have two parameters μ and σ^2, so our parameter space is $\Omega = \mathbb{R} \times [0, \infty)$. We shall find the best critical region for testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$. We set

$$\Omega_0 = \{(\mu, \sigma^2) \in \Omega \,|\, \mu = \mu_0\}, \qquad \Omega_0 = \{(\mu, \sigma^2) \in \Omega \,|\, \mu \neq \mu_0\} = \Omega \setminus \Omega_0. \tag{624}$$

Note that because σ^2 is a variable (unknown), both of the hypotheses are composite.

To compute the maximum likelihood function of the null hypothesis, note that

$$\begin{aligned} L(\Omega_0) &= \sup_{(\mu,\sigma^2)\in\Omega_0} (2\pi \cdot \sigma^2)^{-n/2} \exp\left[-\frac{1}{2 \cdot \sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] & (625) \\ &= \left(2\pi \cdot n^{-1} \sum_{i=1}^n (x_i - \mu_0)^2 \right)^{-n/2} \exp\left[-\frac{1}{2 \cdot n^{-1} \sum_{i=1}^n (x_i - \mu_0)^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right] & (626) \\ &= \left(2\pi \cdot n^{-1} \sum_{i=1}^n (x_i - \mu_0)^2 \right)^{-n/2} \exp\left[-\frac{n}{2} \right], & (627) \end{aligned}$$

where we have used the fact that the MLE of σ^2 for the normal density $N(\mu_0, \sigma^2)$ is $\hat{\sigma^2} = n^{-1} \sum_{i=1}^n (x_i - \mu_0)^2$ (check this by differentiating $L(\Omega_0)$ by σ^2 and finding the value of σ^2 that makes the derivative zero).

On the other hand, the full maximum likelihood is given by

$$\begin{aligned} L(\Omega) &= \sup_{(\mu,\sigma^2)\in\Omega} (2\pi \cdot \sigma^2)^{-n/2} \exp\left[-\frac{1}{2 \cdot \sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] & (628) \\ &= \left(2\pi \cdot n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-n/2} \exp\left[-\frac{1}{2 \cdot n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] & (629) \\ &= \left(2\pi \cdot n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-n/2} \exp\left[-\frac{n}{2} \right], & (630) \end{aligned}$$

where we have used the fact that the unconstrained MLE of μ and σ^2 for the normal density $N(\mu, \sigma^2)$ are given by $\hat{\mu} = \bar{x}$ and $\hat{\sigma^2} = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ (see Example 3.1.6).

Thus the likelihood ratio is given by

$$\lambda = \frac{L(\Omega_0)}{L(\Omega)} = \frac{\left(2\pi \cdot n^{-1} \sum_{i=1}^n (x - \mu_0)^2 \right)^{-n/2}}{\left(2\pi \cdot n^{-1} \sum_{i=1}^n (x - \bar{x})^2 \right)^{-n/2}} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right)^{n/2}. \tag{631}$$

Using the 'Pythagorian theorem' (615) for $a = \mu_0$, we can rewrite the likelihood ratio as

$$\lambda = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] + n(\bar{x} - \mu_0)^2} \right)^{n/2} = \left(\frac{1}{1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)^{n/2} = \left(1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{-n/2}. \tag{632}$$

Note that λ is close to 1 if $\bar{x} \approx \mu_0$, and λ is small when $\bar{x}$ and μ_0 differ by a lot.

Now we compute the critical region, which is given by

$$\lambda = \left(1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{-n/2} \leq k \quad \Longleftrightarrow \quad \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \geq k^{-2/n} - 1 \tag{633}$$

$$\iff \quad \frac{(\bar{x}-\mu_0)^2}{\left(\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2\right)/n} \geq (n-1)(k^{-2/n}-1) \tag{634}$$

$$\iff \quad \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} \geq c := \sqrt{(n-1)(k^{-2/n}-1)}, \tag{635}$$

where s denotes the sample standard deviation.

In order to guarantee significance level α, we compute the Type I error of this test, we recall that (see Proposition 7.3.5)

$$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t(n-1). \tag{636}$$

This gives

$$\begin{aligned} \text{Type I error} &= \mathbb{P}\left(\text{reject } H_0; H_0\right) && (637)\\ &= \mathbb{P}\left(|T| \geq c; H_0\right) && (638)\\ &= \mathbb{P}\left(|t(n-1)| \geq c\right). && (639)\end{aligned}$$

Thus we have

$$\text{Type I error} \leq \alpha \quad \iff \quad c \geq t_{\alpha/2}(n-1). \tag{640}$$

In summary, the critical region of the likelihood ratio test for testing $H_0 : \mu = 162$ against $H_1 : \mu \neq 162$ for unknown σ^2 at significance level α is given by

$$C = \left\{ (x_1, \cdots, x_n) : \left| \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \right| \geq t_{\alpha/2}(n-1) \right\}. \tag{641}$$

This is precisely the t-test we had before in Algorithm 2. ▲

Example 8.7.4 (testing variance for unknown mean)**.** Let $X_1, X_2, \cdots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Here we have two parameters μ and σ^2, so our parameter space is $\Omega = \mathbb{R} \times [0, \infty)$. We shall find the best critical region for testing the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against the alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$. We set

$$\Omega_0 = \{(\mu, \sigma^2) \in \Omega \,|\, \sigma^2 = \sigma_0^2\}, \qquad \Omega_1 = \{(\mu, \sigma^2) \in \Omega \,|\, \sigma \neq \sigma_0^2\} = \Omega \setminus \Omega_0. \tag{642}$$

Note that because μ is a variable (unknown), both of the hypotheses are composite.

To compute the maximum likelihood function of the null hypothesis, note that

$$L(\Omega_0) = \sup_{(\mu,\sigma^2)\in\Omega_0} (2\pi \cdot \sigma^2)^{-n/2} \exp\left[-\frac{1}{2\cdot\sigma^2} \sum_{i=1}^n (x_i-\mu)^2 \right] \tag{643}$$

$$= \left(2\pi\cdot\sigma_0^2\right)^{-n/2} \exp\left[-\frac{1}{2\cdot\sigma_0^2} \sum_{i=1}^n (x_i-\bar{x})^2 \right] \tag{644}$$

where we have used the fact that the MLE of μ for the normal density $N(\mu, \sigma_0^2)$ is the sample mean $\bar{x}$. On the other hand, as in the previous example, the full maximum likelihood is given by

$$L(\Omega) = \sup_{(\mu,\sigma^2)\in\Omega} (2\pi \cdot \sigma^2)^{-n/2} \exp\left[-\frac{1}{2\cdot\sigma^2} \sum_{i=1}^n (x_i-\mu)^2 \right] \tag{645}$$

$$= \left(2\pi \cdot n^{-1} \sum_{i=1}^n (x_i-\bar{x})^2 \right)^{-n/2} \exp\left[-\frac{n}{2} \right], \tag{646}$$

Thus the likelihood ratio is given by

$$\lambda = \frac{L(\Omega_0)}{L(\Omega)} = \frac{\left(2\pi\cdot\sigma_0^2\right)^{-n/2} \exp\left[-\frac{1}{2\cdot\sigma_0^2}\sum_{i=1}^n (x_i-\bar{x})^2 \right]}{\left(2\pi\cdot n^{-1}\sum_{i=1}^n (x_i-\bar{x})^2\right)^{-n/2} \exp\left[-\frac{n}{2}\right]} = \left(\frac{w}{n}\right)^{n/2} \exp\left[-\frac{w}{2} + \frac{n}{2} \right], \tag{647}$$

where we denote

$$w = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{648}$$

Now we compute the critical region as

$$\lambda = \left(\frac{w}{n}\right)^{n/2}\exp\left[-\frac{w}{2}+\frac{n}{2}\right] \le k \quad \Longleftrightarrow \quad n\log(w/n) - w + n \le 2\log k \tag{649}$$

$$\Longleftrightarrow \quad w \le c_1 \text{ or } w \ge c_2, \tag{650}$$

where c_1 and c_2 are constants that depends on n and k (see Exercise 8.7.5).

In principle, we can compute Type I error of the above test by noting that

$$W = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)(S/\sigma_0)^2 \overset{H_0}{\sim} \chi^2(n-1), \tag{651}$$

which we have shown in the proof of Proposition 7.3.5. In practice, most statisticians tend to choose c_1 and c_2 so that both regions $\{W \le c_1\}$ and $\{W \ge c_2\}$ get equal probability $\alpha/2$. For this choice, the resulting critical region will be

$$\text{Reject } H_0 : \sigma^2 = \sigma_0^2 \text{ in favor of } H_1 : \sigma^2 \ne \sigma_0^2 \text{ if } w \le \chi^2_{1-\alpha/2}(n-1) \text{ or } w \ge \chi^2_{\alpha/2}(n-1). \tag{652}$$

▲

Exercise 8.7.5. Let $X_1, X_2, \cdots, X_n$ be a random sample from the normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. We set

$$\Omega_0 = \{(\mu,\sigma^2) \in \Omega \,|\, \sigma^2 = \sigma_0^2\}, \qquad \Omega_0 = \{(\mu,\sigma^2) \in \Omega \,|\, \sigma \ne \sigma_0^2\} = \Omega \setminus \Omega_0. \tag{653}$$

According to Example 20.4, we know that

$$\lambda = \frac{L(\Omega_0)}{L(\Omega)} = \left(\frac{w}{n}\right)^{n/2}\exp\left[-\frac{w}{2}+\frac{n}{2}\right], \tag{654}$$

where

$$w = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{655}$$

(i) For $n = 30$, draw the graph of the function $f(x) = n\log(x/n) - x + n$. (Using graphic calculator or software package)

(ii) Draw the regions of $\{f(x) \le -2\} = \{w \le c_1\} \cup \{w \ge c_2\}$ and approximately compute c_1 and c_2. Using the fact that $W = (1/\sigma_0^2)\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(29)$, compute the probabilities,

$$\mathbb{P}(W \le c_1), \qquad \mathbb{P}(W \ge c_2), \tag{656}$$

which add up to the Type I error of the critical region $\{2\log\lambda \le -2\} = \{\lambda \le 1/e\}$. Are they the same?

(iii) Assume that $n = 30$, $w = 10$, and $\sigma^2 = 5$, and $\alpha = 0.05$. Using the following approximate (and symmetric) critical region, determine whether we can reject $H_0 : \sigma^2 = \sigma_0^2$ in favor of $H_1 : \sigma^2 \ne \sigma_0^2$ at significance level $\alpha = 0.05$.

$$\text{Reject } H_0 : \sigma^2 = \sigma_0^2 \text{ in favor of } H_1 : \sigma^2 \ne \sigma_0^2 \text{ if } w \le \chi^2_{1-\alpha/2}(n-1) \text{ or } w \ge \chi^2_{\alpha/2}(n-1). \tag{657}$$

8. Chi-square goodness-of-fit tests

In this section, we will study how to test a hypothesis about the entire distribution of an unknown random variable, not just on its mean or variance. Roughly speaking, we can see how our hypothetical distribution 'fits' to the empirical distribution given by the sample values of the random variable. If they differ by a lot, then we reject the hypothetical distribution.

8.1. Chi-square goodness-of-fit test. Let X be a discrete RV taking r distinct values from $\{x_1, x_2, \cdots, x_r\}$. We would like to formulate a hypothesis about its entire PMF, which we will denote by f_X:

$$f_X = [\mathbb{P}(X = x_1), \mathbb{P}(X = x_2), \cdots, \mathbb{P}(X = x_r)] = [p_1, p_2, \cdots, p_r]. \tag{658}$$

Consider the following complementary hypotheses

$$H_0 : f_X = [p_1^\circ, p_2^\circ, \cdots, p_r^\circ] \tag{659}$$

$$H_1 : f_X \neq [p_1^\circ, p_2^\circ, \cdots, p_r^\circ]. \tag{660}$$

In order to test the above hypotheses, we first draw some i.i.d. samples, $X_1, X_2, \cdots, X_n$ of X. This will give us the following *empirical distribution*

$$\hat{f}_X = \left[\frac{N_1}{n}, \frac{N_2}{n}, \cdots, \frac{N_r}{n} \right], \tag{661}$$

where for each $1 \leq i \leq r$, N_i denotes the number of samples of value x_i:

$$N_i = \sum_{k=1}^{n} \mathbf{1}(X_k = x_i). \tag{662}$$

Since X_k's are i.i.d. and each X_k take value x_i with probability p_i,

$$N_i \sim \text{Binomial}(n, p_i). \tag{663}$$

In particular,

$$\mathbb{E}[N_i] = np_i, \qquad \text{Var}(N_i) = np_i(1 - p_i). \tag{664}$$

Furthermore, according to CLT we also have the following normal approximation

$$\frac{N_i - np_i}{\sqrt{np_i(1 - p_i)}} \Longrightarrow N(0, 1) \tag{665}$$

as the sample size n tends to infinity.

However, the above 'large sample approximation' only applies for each N_i separately, whereas we would like to approximate the entire empirical distribution $\hat{f}_X$ as $n \to \infty$. The difficulty here is that the counts $N_1, \cdots, N_r$ are not independent, since they satisfy the following deterministic condition

$$N_1 + N_2 + \cdots + N_r = n. \tag{666}$$

The following classical theorem due to Pearson overcomes this issue and provides the basis of 'goodness-of-fit' tests. See the end of this section for a proof this result.

Theorem 8.8.1 (Pearson, 1900). *Let $N_1, \cdots, N_r$ be as before. Then as the sample size $n \to \infty$,*

$$\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \Longrightarrow \chi^2(r - 1), \tag{667}$$

where $\Longrightarrow$ denotes convergence in distribution and $\chi^2(r-1)$ denotes the chi-square distribution with $r - 1$ degrees of freedom.

In light of Pearson's theorem, we formulate the chi-square goodness-of-fit test as follows.

Example 8.8.2 (Chi-square goodness-of-fit test)**.** Let X be a discrete RV of interest, taking r distinct values from $\{x_1, x_2, \cdots, x_r\}$. Consider the following complementary hypotheses on the PMF f_X of X:

$$H_0 : f_X = [p_1^\circ, p_2^\circ, \cdots, p_r^\circ] \tag{668}$$

$$H_1 : f_X \neq [p_1^\circ, p_2^\circ, \cdots, p_r^\circ]. \tag{669}$$

Form the following test statistic

$$T = \sum_{i=1}^{r} \frac{(N_i - np_i^\circ)^2}{np_i^\circ} \tag{670}$$

Clearly, $T = 0$ if H_0 were true, and large values of T will indicate that H_0 is far from the truth. Hence we may form the critical region as

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } T \geq \theta, \tag{671}$$

where $\theta \geq 0$ is a thresholding parameter that we will choose in response to a desired significance level α.

By using Pearson's Theorem (Theorem 8.8.1), we can approximately compute the type I error as

$$\begin{aligned} \text{Type I error} &= \mathbb{P}(\text{Reject } H_0; H_0) & (672)\\ &= \mathbb{P}(T \geq \theta; H_0) & (673)\\ &\approx \mathbb{P}(\chi^2(r-1) \geq \theta). & (674) \end{aligned}$$

Hence, approximately as $n \to \infty$,

$$\text{Type I error} \leq \alpha \quad \Longleftrightarrow \quad \mathbb{P}(\chi^2(r-1) \geq \theta) \leq \alpha \quad \Longleftrightarrow \quad \theta \geq \chi^2_\alpha(r-1). \tag{675}$$

Thus the optimal choice of θ is $\theta = \chi^2_\alpha(r-1)$. ▲

Example 8.8.3 (Excerpted from [HTZ77]). If persons are asked to record a string of random digits, such as

$$3 \quad 7 \quad 2 \quad 4 \quad 1 \quad 9 \quad 7 \quad 2 \quad 1 \quad 5 \quad 0 \quad 8 \quad \cdots \tag{676}$$

we usually find that they are reluctant to record the same or even the two closest numbers in adjacent positions. Indeed, if the digits are i.i.d. uniform from $\{0,1,\cdots,9\}$, then the probability of the next digit being the same as the preceding one is $q_1 = 1/10$, the probability of the next being only one away from the preceding (assuming that 0 is one away from 9) is $q_2 = 2/10$, and the probability of all other possibilities is $q_3 = 7/10$.

Suppose we have an algorithm that generates strings of digits, and we would like to test if it generates random strings. Let

$$\begin{aligned} p_1 &= \text{frequeny that two consecutive digits generated by the algroithm are the same} & (677)\\ p_2 &= \text{frequeny that two consecutive digits generated by the algroithm differ by 1} & (678)\\ p_3 &= \text{frequeny that two consecutive digits generated by the algroithm differ by} \geq 2. & (679) \end{aligned}$$

By comparing the actual random number generator, we form the following null hypothesis

$$H_0 : [p_1, p_2, p_3] = [q_1, q_2, q_3] = [1/10, 2/10, 7/10]. \tag{680}$$

Our test statistic is

$$T = \sum_{i=1}^{3} \frac{N_i - nq_i}{nq_i}, \tag{681}$$

where

$$\begin{aligned} N_1 &= \#\text{ of times that the next digit is the same as before} & (682)\\ N_2 &= \#\text{ of times that the next digit is one away from before} & (683)\\ N_3 &= \#\text{ of times that the next digit is more than two away from before.} & (684) \end{aligned}$$

The critical region for $\alpha = 0.05$ significance level is $t \geq \chi^2_{0.05}(2) = 5.991$.

Suppose the observed sequence of 51 digits is as follows:

$$583194679263087513621954803714604382739856187035252. \tag{685}$$

By examining this sequence, we find

$$N_1 = 0, \qquad N_2 = 8, \qquad N_3 = 42. \tag{686}$$

There are $n = 50$ consecutive digits in this string. The computed chi-square statistic is

$$t = \frac{(0-50(1/10))^2}{50(1/10)} + \frac{(8-50(2/10))^2}{50(2/10)} + \frac{(42-50(7/10))^2}{50(7/10)} = 6.8 > 5.991 = \chi^2_{0.05}(2). \tag{687}$$

Thus, we conclude that the algorithm does not seem to be generating random numbers. ▲

The proof of Pearson's theorem uses a random vector version of CLT, which is called the 'multivariate CLT'. We state this result without proof.

Theorem 8.8.4 (Multivariate CLT). *Let $\mathbf{X} = (X_1, \cdots, X_r)$ be a random vector so that each entry has finite variance. Let Σ denote the covariance matrix of $\mathbf{X}$, that is, the $r \times r$ matrix whose (i, j) entry is given by*

$$\Sigma[i,j] = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]. \tag{688}$$

Let $(\mathbf{X}_i)_{i\ge 1}$ be a sequence of i.i.d. copies of $\mathbf{X}$. Then as $n \to \infty$,

$$\mathbf{Z}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]) \Longrightarrow N(\mathbf{0}, \Sigma), \tag{689}$$

where $\Longrightarrow$ denotes convergence in distribution and $N(\mathbf{0}, \Sigma)$ denotes the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ.

Proof of Theorem 8.8.1. (Optional*) Recall that we have an unknown RV X that takes r values $x_1, \cdots, x_r$, and we have i.i.d. samples $X_1, \cdots, X_n$ of X. Every time we draw a sample X_k, we observe the r-dimensional row vector of indicators

$$\mathbf{X}_k = [\mathbf{1}(X_k = x_1), \mathbf{1}(X_k = x_2), \cdots, \mathbf{1}(X_k = x_r)]. \tag{690}$$

In words, if X_k takes the jth value x_j, then the jth entry of $\mathbf{X}_j$ becomes 1 and its all other entries are zero. To compute its covariance matrix Σ of $\mathbf{X}_k$, observe that

$$\begin{aligned} \mathrm{Cov}(\mathbf{1}(X_k = x_i)\mathbf{1}(X_k = x_j)) &= \mathbb{E}\left[\mathbf{1}(X_k = x_i)\mathbf{1}(X_k = x_j)\right] - \mathbb{E}\left[\mathbf{1}(X_k = x_i)\right]\mathbb{E}\left[\mathbf{1}(X_k = x_j)\right] && (691) \\ &= \mathbb{E}\left[\mathbf{1}(X_k = x_i)\mathbf{1}(X_k = x_j)\right] - p_i p_j && (692) \\ &= \begin{cases} p_i(1-p_i) & \text{if } i = j \\ -p_i p_j & \text{if } i \neq j. \end{cases} && (693) \end{aligned}$$

Thus the covariance matrix Σ is given as

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & -p_1 p_3 & \cdots & -p_1 p_r \\ -p_2 p_1 & p_2(1-p_2) & -p_2 p_3 & \cdots & -p_2 p_r \\ \vdots & & \ddots & & \vdots \\ -p_r p_1 & -p_2(1-p_2) & \cdots & -p_r p_{r-1} & p_r(1-p_r) \end{bmatrix}. \tag{694}$$

Hence according to the mutivariate CLT, as $n \to \infty$,

$$\mathbf{Z}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]) = \left[\frac{N_1 - np_1}{\sqrt{n}}, \frac{N_2 - np_2}{\sqrt{n}}, \cdots, \frac{N_r - np_r}{\sqrt{n}}\right] \Longrightarrow N(0, \Sigma). \tag{695}$$

A slight modification will show that

$$\left[\frac{N_1 - np_1}{\sqrt{np_1}}, \frac{N_2 - np_2}{\sqrt{np_2}}, \cdots, \frac{N_r - np_r}{\sqrt{np_r}}\right] \Longrightarrow N(0, \Sigma'), \tag{696}$$

where the new covariance matrix Σ' is given by

$$\Sigma' = \begin{bmatrix} 1-p_1 & -\sqrt{p_1 p_2} & -\sqrt{p_1 p_3} & \cdots & -\sqrt{p_1 p_r} \\ -\sqrt{p_2 p_1} & 1-p_2 & -\sqrt{p_2 p_3} & \cdots & -\sqrt{p_2 p_r} \\ \vdots & & \ddots & & \vdots \\ -\sqrt{p_r p_1} & -\sqrt{p_2(1-p_2)} & \cdots & -\sqrt{p_r p_{r-1}} & 1-p_r \end{bmatrix}. \tag{697}$$

This[1] shows that

$$\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \Longrightarrow \sum_{i=1}^{r} Z_i^2, \tag{698}$$

where the Gaussian random vector $[Z_1, \cdots, Z_r]$ follows the multivariate normal distribution $N(0, \Sigma')$.

To finish up, we use the result of Exercise 8.8.5, which states that

$$Z_1^2 + \cdots + Z_{r-1}^2 + Z_r^2 \overset{d}{=} \widetilde{Z}_1^2 + \cdots + \widetilde{Z}_{r-1}^2, \tag{699}$$

where $[\widetilde{Z}_1, \cdots, \widetilde{Z}_{r-1}]$ is multivariate normal with mean zero and covariance matrix I_{r-1}. Since the right hand side follows $\chi^2(r-1)$, the assertion then follows. □

Exercise 8.8.5. Let $\mathbf{g} = [g_1, \cdots, g_r] \sim N(0, I_r)$ and let $\mathbf{p} = [\sqrt{p_1}, \cdots, \sqrt{p_r}]$. Note that $\|\mathbf{p}\| = 1$. Using the standard basis $[\mathbf{e}_1, \cdots, \mathbf{e}_r]$, write

$$\mathbf{g} - (\mathbf{g} \cdot \mathbf{p})\mathbf{p} = Z_1 \mathbf{e}_1 + \cdots + Z_{r-1} \mathbf{e}_{r-1} + Z_r \mathbf{e}_r. \tag{700}$$

(i) Show that $[Z_1, \cdots, Z_r] \sim N(0, \Sigma')$, where the covariance matrix Σ' is given by (697).

(ii) By changing the standard basis $[\mathbf{e}_1, \cdots, \mathbf{e}_r]$ to an orthonormal basis $[\mathbf{q}_1, \cdots, \mathbf{q}_1, \mathbf{p}]$, write

$$\mathbf{g} = \widetilde{Z}_1 \mathbf{q}_1 + \cdots + \widetilde{Z}_{r-1} \mathbf{q}_{r-1} + (\mathbf{g} \cdot \mathbf{p})\mathbf{p}. \tag{701}$$

Using the fact that the covariance matrix of a multivariate normal does not change under change of orthonormal basis, deduce that $[\widetilde{Z}_1, \cdots, \widetilde{Z}_{r-1}] \sim N(0, I_{r-1})$.

(iii) From (i) and (ii), show that

$$Z_1^2 + \cdots + Z_{r-1}^2 + Z_r^2 = \|\mathbf{g} - (\mathbf{g} \cdot \mathbf{p})\mathbf{p}\|^2 = \widetilde{Z}_1^2 + \cdots + \widetilde{Z}_{r-1}^2. \tag{702}$$

Conclude that

$$Z_1^2 + \cdots + Z_{r-1}^2 + Z_r^2 \sim \chi^2(r-1). \tag{703}$$

[1] In fact, we need to appeal to a functional version of the multivariate CLT to deduce that the convergence of a sequence of random vector in distribution transfers to a function of their coordinates.

Bibliography

[HTZ77] Robert V Hogg, Elliot A Tanis, and Dale L Zimmerman, *Probability and statistical inference*, vol. 993, Macmillan New York, 1977.

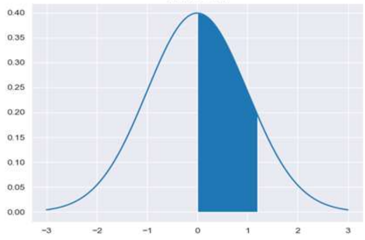
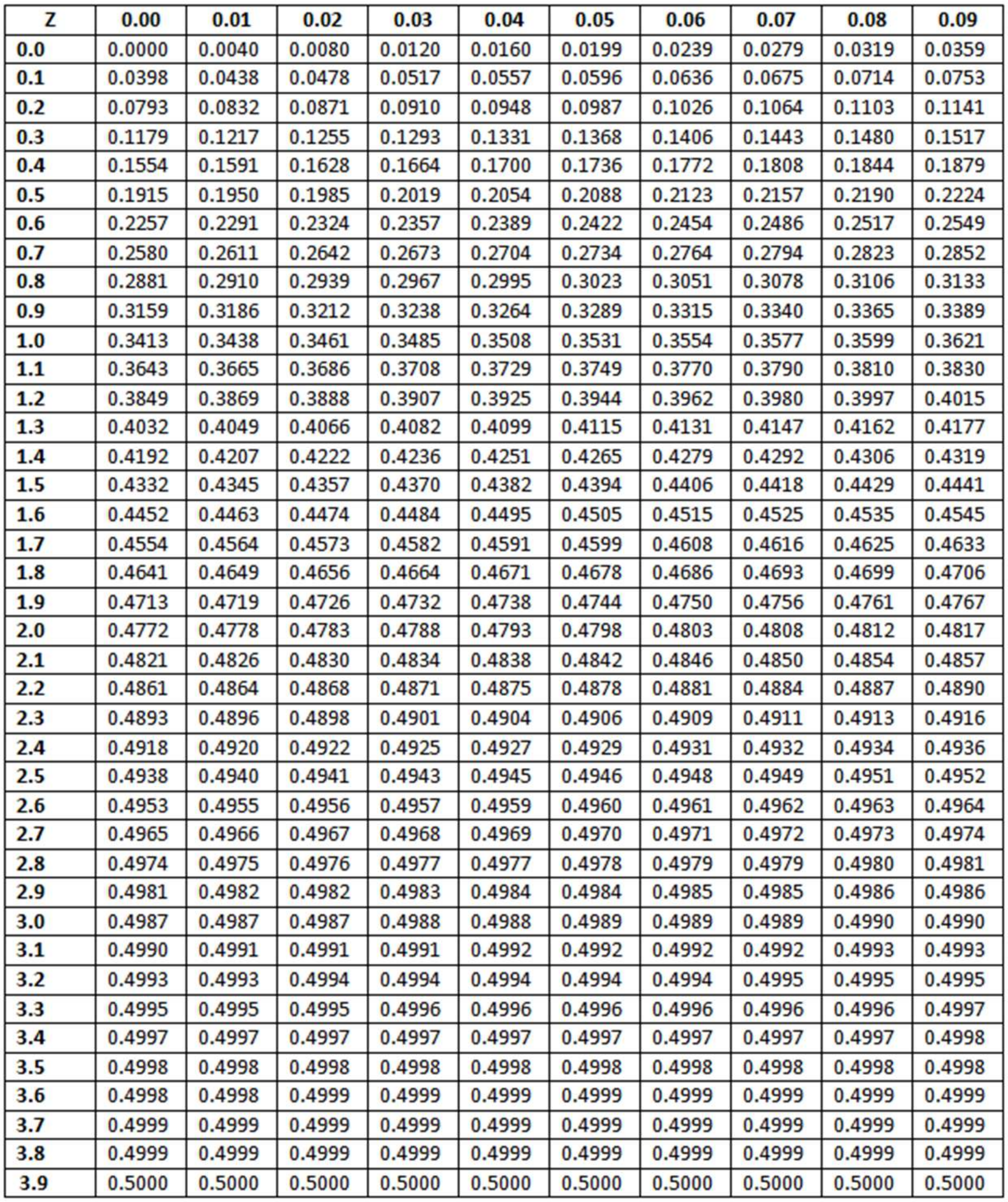


Table of standard normal probabilities $\mathbb{P}(0 \leq Z \leq z)$, where $Z \sim N(0,1)$.

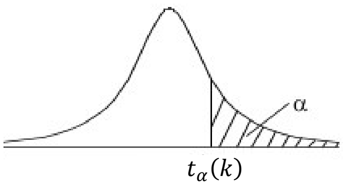


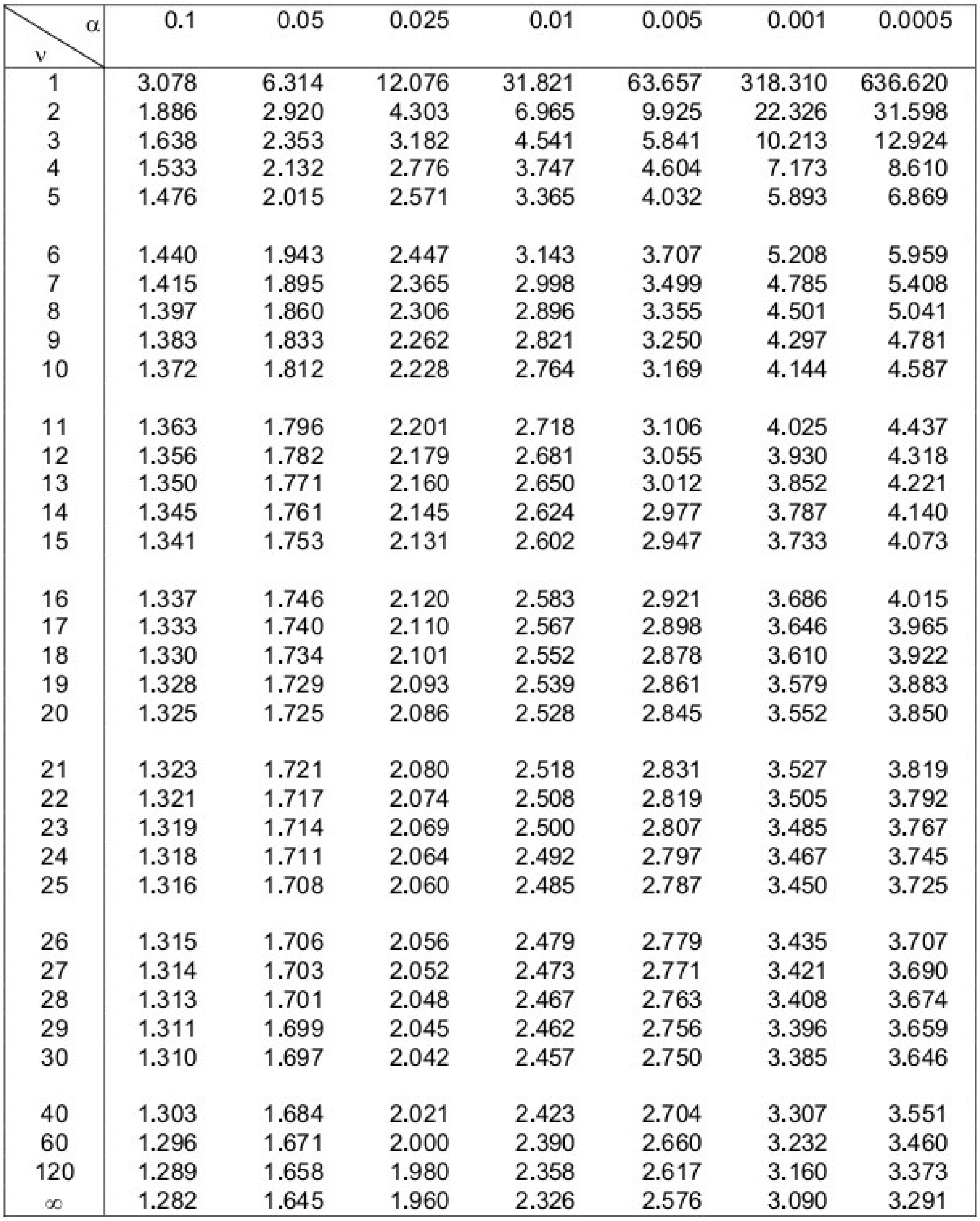
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

TABLE 1. Standard normal table

Table of the Student's *t*-distribution

Table of t-scores $t_\alpha(k)$, which is the unique value such that $\mathbb{P}\big(T > t_\alpha(k)\big) = \alpha$ for $T \sim t(k)$.





α ν	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

TABLE 2. Table of t-scores