STAT*3110 - WINTER2020

INTRODUCTORY MATHEMATICAL STATISTICS II

Zeny Feng

Department of Mathematics and Statistics,
University of Guelph

November 20, 2019

Contents

Chapter 1

Review of Probability and Distribution (STAT 3100)

1.1 Probability

1.1.1 Probability - frequentist approach

1. Random experiment

    - all possible outcomes can be listed.

    - the outcome is generally uncertain.

2. Sample Space, S, a set of all possible outcomes.

3. Random event, A

   - a possible outcome
   - a subset of sample space

   - event space, $\mathcal{A}$, a collection of all possible events.

4. Probability measure, $P(\cdot)$, a function defined over the sample space mapping to [0,1]

1.1.2 Frequency interpretation

Example: $\mathrm{P}(A) = 0.9$.

1.1.3 Independence of two events

A and B are independent

1.2 Random variables and their distributions

Definition 1.2.1 *A random variable is a function that maps from sample space to the real line.*

Probability distribution function of a random variable X

1. Discrete random variable X:

   - Probability mass function, pmf: $f(x) = \mathrm{P}(X = x)$

   - $\mathrm{P}(X \in A) = \sum_{x \in A} \mathrm{P}(X = x)$

2. Continuous random variable X

   - Probability density function, pdf:

- Note that, $\mathrm{P}(X = x) = 0$.
- In a general form,

$$\mathrm{P}(X \in A) = \int_A f(x)dx$$

1.2.1 Cumulative distribution function (cdf).

$$F(x) = \mathrm{P}(X \leq x), \qquad 0 \leq F(x) \leq 1.$$

$F(x)$ is a non decreasing function:

1. Discrete case: $F(X)$ is a step function.

2. Continuous case:

1.3 Expectation, variance and moments

1.3.1 Expectation: mean, average.

1. Expectation of the function of X.

2. Expectation of X.

3. Expectation of the linear combination.

1.3.2 Moments

1. rth moment.

2. rth central moment.

3. Chebyshev's inequality:
   If $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$, then, $\forall \varepsilon > 0$,

$$\mathrm{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

4. Markov inequality: $\forall \varepsilon > 0$,

$$\mathrm{P}(X \geq \varepsilon) \leq \frac{\mathrm{E}(X)}{\varepsilon}.$$

5. If X and Y are two random variables

$$\text{var}(aX + bY) = a^2\text{var}(X) + 2ab\text{cov}(X, Y) + b^2\text{var}(Y)$$

1.4 Moment generating function (mgf)

Definition 1.4.1 *The **moment generating function** of a random variable X, $M_X(t)$ is given by*

$$M_X(t) = E(e^{tX}), \qquad \text{for } t \in (-\delta, \delta)$$

where δ is a fixed value.

Example 1.4.2 *Find the mgf for $X \sim Bin(n, p)$.*

Example 1.4.3 *Find the mgf for $X \sim Exponential(\lambda)$.*

1. Note that, the mgf does not always exists. The characteristic function $\varphi_X(t) = \mathrm{E}(e^{itX}), i = \sqrt{-1}$, always exists for any random variable.

2. Properties of mgf.

   - $Y = aX + b$, then,

   $$M_Y(t) = e^{bt} M_X(at)$$

   Example: $Z \sim N(0,1), M_Z(t) = e^{t^2/2}$. If $X \sim N(\mu, \sigma^2)$, then

   - X and Y are independent, then

   $$M_{X+Y}(t) = M_X(t) M_Y(t).$$

   - Use mgf to find the rth moment, $E(X^r)$.

   - Uniqueness of mgf.
   If $M_X(t) = M_Y(t)$, then $f_X(x) = f_Y(y), \forall x = y$.

1.5 Common distribution

1.5.1 Discrete random variables

1. Bernoulli (p) and binomial (N, p) distributions

2. Geometric and negative binomial distributions

3. Hypergeometric distribution

- Without replacement

- With replacement

4. Poisson distribution
   A limiting form of binomial distribution when $n \to \infty, p \to 0$, while $\lambda = np$ is constant, the number of successes $X \sim$ Poisson(λ).

1.5.2 Continuous random variables

1. Uniform$[a, b]$, $f(x) = \frac{1}{b-a}, a \leq x \leq b$.

2. $X \sim$ Exponential(θ).

   Memoryless property: $\mathrm{P}(X > t + s | X > s) = \mathrm{P}(X > t)$

3. Normal distribution, $X \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, X \in (-\infty, \infty).$$

4. $X \sim \text{Gamma}(\alpha, \beta), \alpha, \beta > 0,$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

$$\text{E}(X) = \alpha\beta, \qquad \text{var}(X) = \alpha\beta^2$$

- Special case of a Gamma distribution

$$X \sim \text{Exponential}(\beta) = \text{Gamma}(1, \beta)$$

$$X \sim \chi^2_d = \text{Gamma}(\frac{d}{2}, 2), \qquad d = 1, 2, 3, ...$$

5. $X \sim \text{Beta}(\alpha, \beta), \alpha, \beta > 0,$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x, 1$$

$$\text{E}(X) = \frac{\alpha}{\alpha+\beta}, \qquad \text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

1.6 Joint distribution

1.6.1 Discrete case

1. Joint pmf: $f_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y)$,

$$\sum_x \sum_y f_{X,Y}(x,y) = 1, \quad \mathrm{P}((X,Y) \in A) = \sum_{(x,y)} \sum_{\in A} f_{X,Y}(x,y)$$

   Example: $X, Y \sim$ iid Geometric(p)

2. Marginal distribution

3. Conditional distribution

4. Independence: X and Y are independent if $\forall x, y$

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y)$$
$$\text{or}$$
$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Note, if $\exists x_0, y_0$ such that $f_{X,Y}(x_0, y_0) \neq f_X(x_0) f_Y(y_0)$, then

1.6.2 Continuous case

X and Y are jointly continuous distributed if there is a function $f_{X,Y}(x, y)$, $\forall (x, y) \in (-\infty, \infty)$, such that $f_{X,Y}(x, y) \geq 0$, and $\forall a < b, c < d$,

$$\mathrm{P}(a < x < b, c < y < d) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy$$

1. Marginal:

2. Conditional:

3. X and Y are independent

Example 1.6.1 $f_{(X,Y}(x,y) = 6x^2y, 0 \leq x \leq 1, 0 \leq y \leq 1$, *find* $P(X \geq Y)$.

1.7 Sampling distribution

1.7.1 Sample mean and sample variance

Suppose $X_1, X_2, ..., X_n$ is a random sample of size n from a population with mean μ and variance σ^2, then, the sample mean

- By central limit theorem (CLT),

- If $X_1, X_2, ..., X_n$ are iid from $N(\mu, \sigma^2)$, then

If σ^2 is not known, we use the sample variance

- replace the σ^2 by sample variance, we have

Theorem 1.7.1 *If $X_1, ..., X_n$ is a random sample from $N(\mu, \sigma^2)$, let $\bar{X}$ and S^2 be the sample mean and sample variance, then we have*

- *$\bar{X}$ and S^2 are independent.*
- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Theorem 1.7.2 *If S_1^2 and S_2^2 are the sample variances of two random samples of size n_1 and n_2 from two populations with the variances σ_1^2 and σ_2^2, then,*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Chapter 2

Change of Random Variables

In this chapter, we consider the problem of finding the probability distribution function or the probability density function (pdf) of one or more than one variables that is on the basis of other random variable(s). For example, suppose $X_1, ..., X_n$ is a set of random variables with a known joint probability distribution/density function $f_{X_1,...,X_n}(x_1, ..., x_n)$. Now, there is a random variable Y which is a function of X_i's, say,

$$Y = h(X_1, ..., X_n).$$

We want to find the probability distribution/density function for the random variable Y.

There are several techniques of finding the probability distribution/density function in this kind of problems: the distribution function technique, the transformation technique, and the moment generating function technique. Here, we focus more on the distribution function and the transformation technique.

2.1 Distribution function technique

Suppose $X_1, ..., X_n$ is a set of continuous random variables with a known joint pdf $f_{X_1,...,X_n}(x_1, ..., x_n)$, and le

$$Y = h(X_1, ...X_n).$$

Given that $X_1, ..., X_n$ are continuous, Y is continuous, a straightforward way to find the pdf of Y, $f_Y(y)$ is first to find the cumulative density function (cdf) by determining the probability of

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(h(X_1, ..., X_n) \leq y)$$

and then get the pdf of Y by

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y}.$$

Example 2.1.1 *If* $X \sim f_X(x)$ *with*

$$f_X(x) = \begin{cases} 6x(1-x) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

find the pdf of $Y = X^3$.

Example 2.1.2 *If* $X \sim N(0, \sigma^2)$*, find the pdf of* $Y = |X|$.

If $X_1, ..., X_n$ is a set of discrete random variables and $Y = h(X_1, ..., X_n)$, we could work on the pmf of Y directly.

Example 2.1.3 *If $X_1 \sim Poisson(\lambda_1)$, $X_2 \sim Poisson(\lambda_2)$, X_1 and X_2 are independent and $Y = X_1 + X_2$, find the pmf of Y.*

Remark: In the discrete case, we usually work on the probability distribution function directly. In continuous case, it is often easier to find the cumulative distribution function(cdf) and then differentiate the cdf to get the probability density function.

An alternative popular way of finding the probability distribution/density function in the problem of the change of random variables is to use the transformation technique.

2.2 Transformation technique: one variable to one variable

2.2.1 One-to-one correspondent

Suppose $X \sim f_X(x)$, where $f_X(x)$ is the probability density function of X. Let $Y = h(X)$ where their relationship is one-to-one correspondent. Find the probability density function $f_Y(y)$ using the transformation technique.

Theorem 2.2.1 *Suppose X is a continuous random variable with pdf $f_X(x)$ and $Y = h(X)$. If $h(x)$ is differentiable at all values of x and is either a decreasing function or an increasing function such that there is unique inverse function for X that $X = h^{-1}(Y)$, then, the pdf of Y is given by*

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \left| \frac{\partial h^{-1}(y)}{\partial y} \right|$$

provided that $\frac{\partial h(x)}{\partial x} \neq 0$. Otherwise, $f_Y(y) = 0$.

Example 2.2.2 *With reference to Example 2.1.1, if* $X \sim f_X(x)$ *with*

$$f_X(x) = \begin{cases} 6x(1-x) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

use the transformation technique to find the pdf of $Y = X^3$.

Example 2.2.3 *If $F_X(x)$ is the cumulative distribution function of the continuous random variable X, find the probability density of $Y = F_X(x)$.*

Example 2.2.3 tells us that the cumulative distribution function $F_X(x)$ for any distribution follows a uniform [0,1] distribution. This fact is very important not only theoretically, but also facilitates certain simulation procedures for generating random samples that follow a given distribution.

2.2.2 Not one to one corresponding

The conditions underlying Theorem 2.2.1 are often not met. For example, when $Y = X^2$, over the domain of $X \in \mathcal{R}$, the function is concave rather than decreasing or increasing only. In this case, we divide the domain of X in two non-overlapping regions: $A_1 = (-\infty, 0), A_2 = (0, \infty)$. Then, we find the pdf of Y in each of A_1 and A_2 using Theorem 2.2.1.

Example 2.2.4 *Suppose $X \sim N(0, 1)$, find the pdf for $Y = X^2$.*

2.2.3 Transformation technique: multivariables

Theorem 2.2.1 can be generalized to situations where there are more than one random variables being transformed. For example, suppose we have a set of continuous random variables $X_1, ..., X_n$ with joint probability density function $f_{X_1,...,X_n}(x_1, ..., x_n)$. Now suppose we have another set of random variables $Y_1, ..., Y_n$:

$$\begin{aligned} Y_1 &= h_1(X_1, ..., X_n) \\ Y_2 &= h_2(X_1, ..., X_n) \\ &\vdots \\ Y_n &= h_n(X_1, ..., X_n). \end{aligned}$$

Here, we may want to find the probability density function for each of the Y's or the joint probability density function of all Y's or some Y's. For simplicity, suppose we focus on the case of one-to-one correspondence in the sense that there is an unique set of inverse functions

$$\begin{aligned} X_1 &= g_1(Y_1, ..., Y_n) \\ X_2 &= g_2(Y_1, ..., Y_n) \\ &\vdots \\ X_n &= g_n(Y_1, ..., Y_n). \end{aligned}$$

The below theorem, which is a generalization of the Theorem 2.2.1 to the multivariate case, can be used to find the joint distribution of $Y_1, ..., Y_n$.

Theorem 2.2.5 *Suppose $X_1, ..., X_n$ is a set of continuous random variables with the joint pdf $f_{X_1,...,X_n}(x_1, ..., x_n)$ and let $Y_1 = h_1(X_1, ..., X_n), Y_2 = h_2(X_1, ..., X_n), \ldots, Y_n = h_n(X_1, ..., X_n)$ be another set of random variables. If the $h_i(X_1, ..., X_n)$ functions are differentiable with respect to each of $X_1, ..., X_n$ and are one-to-one correspondent within the range of $X_1, ..., X_n$ for which $f_{X_1,...,X_n}(x_1, ..., x_n) \neq 0$, then, the joint pdf of $Y_1, ..., Y_n$ is given by*

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = f_{X_1,...,X_n}(g_1(y_1, ..., y_n), ..., g_n(y_1, ..., y_n)) \cdot |J|$$

where J is called the ***Jacobian*** *of the transformation and is given by*

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Elsewhere, $f_{Y_1,...,Y_n}(y_1, ..., y_n) = 0$.

Example 2.2.6 *If the joint pdf of X_1 and X_2 is given by*

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} e^{-(x_1+x_2)} & \text{for } x_1 > 0, x_2 > 0 \\ 0 & \text{otherwise,} \end{cases}$$

1. *find the joint pdf of $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_1+X_2}$,*
2. *find the marginal distribution of Y_2.*

From Example 2.2.6, we find that the Theorem 2.2.5 can be used to solve the problem of transforming n random variables to r random variables for $n \geq r$. That is, we only have r Y's or equations for n random variables X's.

Example 2.2.7 *If the joint pdf of* X_1 *and* X_2 *is given by*

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 1 & \textit{for } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \textit{otherwise,} \end{cases}$$

find the pdf of $Y = X_1 + X_2$.

Chapter 3

Order Statistics

Suppose $X_1, ..., X_n$ is a random sample of size n from an infinite population with a continuous pdf. Now, we arrange the values of $X_1, ..., X_n$ in an ascending order and denote them in the form of

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

For a random sample of size n, there are $n!$ possible arrangements, so, there are $n!$ possible configurations of the orders.

Note that, the observations and order statistic are not one-to-one correspondent.

3.1 Distribution of minimum and maximum statistics

Suppose $X_1, ..., X_n \sim$ iid $f_X(x)$ and are continuous. Let

$$X_{(1)} = \min(X_1, X_2, ..., X_n)$$

be the minimum statistic and

$$X_{(n)} = \max(X_1, X_2, ..., X_n).$$

be the maximum statistic. Find the pdf's of $X_{(1)}$ and $X_{(n)}$.

Example 3.1.1 *Suppose* $X_1, ..., X_n \sim$ *iid uniform*$[0, \theta]$, $\theta > 0$, *find the pdf's for the minimum and maximum statistics.*

$$f_X(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta, \\ 0 & otherwise \end{cases}$$

and

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{\theta} & 0 \leq x \leq \theta, \\ 1 & x > \theta. \end{cases}$$

3.2 Distribution of the rth order statistics

Although the sample mean is often used to estimate the population mean in many analysis we carry out, in some cases, it is often better to use the median to describe the "location" of the data. In a random sample $X_1, ..., X_n$ from an infinite population, we denote the sample median as $\tilde{X}$. What is the distribution of $\tilde{X}$? How about the distribution of the other order statistic $X_{(r)}$ for $1 < r < n$?

(1) By the law of mean from calculus, we have

$$\mathrm{P}(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(t)dt = f_X(x)\Delta x$$

for $\Delta x \to 0$.

(2) By (1), for $\Delta x \to 0$, we have

$$\mathrm{P}(x \leq X_{(r)} \leq x + \Delta x) = f_{X_{(r)}}(x)\Delta x$$

(3) The pdf of the rth order statistic $X_{(r)}$ is given by

$$f_{X_{(r)}}(x) = \lim_{\Delta X \to 0} \frac{f_{X_{(r)}}(x)\Delta x}{\Delta x}$$

Putting the results of (2) and (3) together, we can find the pdf of the rth order statistic if $\mathrm{P}(x \leq X_{(r)} \leq x + \Delta x)$ is given. Now, how to find $\mathrm{P}(x \leq X_{(r)} \leq x + \Delta x)$?

Given the $\mathrm{P}(x \leq X_{(r)} \leq x + \Delta x)$ we obtain the pdf of the rth order statistics $f_{X_{(r)}}(x)$.

Apply the above result, we find the sample median $\tilde{X}$ of a random sample $X_1, ..., X_{2n+1}$ of size $2n + 1$ has the pdf of

$$f_{\tilde{X}}(x) =$$

Example 3.2.1 *Suppose $X_1, ..., X_n$ is a random sample from an exponential population with mean θ. Find the distribution of the minimum and the maximum statistics and the distribution of the sample median.*

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0, \quad \theta > 0$$

$$F_X(x) = 1 - e^{-x/\theta}, \quad x > 0, \quad \theta > 0$$

Chapter 4

Point Estimation

In a statistical study, suppose a random sample $X = (X_1, ..., X_n)'$ from a given population is observed. We usually assume a **parametric** model for the population that each observation X_i, for $i = 1, ..., n$, follows a identical distribution with a probability distribution function $f(x, \theta)$. For example, the birthweight of new born baby girls might follow a normal distribution $N(\mu, \sigma^2)$, life times of machine parts might be assumed to follow a Gamma(α, β) distribution, and the daily number of paintings sold in a given art gallery might follow a Poisson(λ) distribution.

In statistics, a parametric model for a given population describes:

1. The general form of the probability distribution/density function $f(x; \theta)$ is known.
2. $f(x; \theta)$ is a member of a distribution family $\{f(x; \theta) : \theta \in \Theta\}$.

   Θ: parameter space, a collection of all possible values of θ.
3. θ is not known.
4. The statistical inference about the population, given the observed data, is essentially inference about θ.

Given a parametric model, there are three main types of inference problems:

1. Point estimation: $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ for estimating the unknown θ, $\hat{\theta}$ is called a point estimator.

   - Construction of $\hat{\theta}$: how to estimate θ.
   - Evaluation of $\hat{\theta}$: unbiasedness, efficiency, consistency and sufficiency.

2. Interval estimation (confidence region of an estimate)

   - Given a point estimate, we want to compute an interval $(\hat{\theta}_1, \hat{\theta}_2)$ for θ such that for some pre-specified probability, say, $(1 - \alpha)$,
   $$\mathrm{P}(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha.$$

3. Hypothesis tests

   - Hypothesis: a statement about the parameter (or the population), e.g.,
   $$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_a : \theta \neq \theta_0.$$
   - Test H_0 and make conclusion if H_0 is true or false.
   - Two types of error:

     Type I error: reject the true H_0.

     Type II error: do not reject the false H_0.
   - Probability of making such errors:
   $$\alpha = \mathrm{P}(\text{type I error}) = \mathrm{P}(\text{reject } H_0 | H_0 \text{ is true})$$
   $$\beta = \mathrm{P}(\text{type II error}) = \mathrm{P}(\text{do not reject } H_0 | H_0 \text{ is false})$$
   $$\text{power} = 1 - \beta = \mathrm{P}(\text{reject } H_0 | H_0 \text{ is false})$$

4.1 Evaluation of point estimations

Suppose a random sample $X_1, ..., X_n$ is collected from a population and it is assumed that $X_i \overset{iid}{\sim} f(x;\theta)$, for $i = 1, ..., n$. Let $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ be a function of X_i's that estimate θ. Then, $\hat{\theta}$ is an estimator of θ.

Properties of a good estimator:

1. Unbiasedness:

2. Efficiency:

3. Consistency:

4. Sufficiency: Does $\hat{\theta}(X_1, ..., X_n)$ utilize all the information contained in the data to estimate θ?

5. Likeliness: $\hat{\theta}(X_1, ..., X_n)]$ is the most likely value of θ given the observed data $X_1, ..., X_n$. (Among all choices for θ, $\hat{\theta}$ is the one with the highest probability for observing the data).

6. Robustness: $\hat{\theta}(X_1, ..., X_n)]$ has a sampling distribution that is not too adversely affected by violations of assumptions made in the model/analysis. (We will not cover this here).

4.1.1 Unbiased estimators

Suppose a random sample $X_1, ..., X_n$ is from a given population with probability distribution function $f(x;\theta)$.

Definition 4.1.1 *A statistic is a function of data $(X_1, ..., X_n)$. Suppose a statistic $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ is used to estimate the parameter θ.*

*We say $\hat{\theta}$ is an **unbiased estimator** of θ if and only if*

$$E[\hat{\theta}(X_1, ..., X_n)] = \theta.$$

Example 4.1.2 *If $X_1, ..., X_n \overset{iid}{\sim}$ Bernoulli(p), is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ an unbiased estimator of p?*

If an estimator $\hat{\theta}$ is biased for θ, the amount of bias is given by

$$\text{bias}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)$$

If $\hat{\theta}$ is unbiased, $\text{bias}(\hat{\theta}) = 0$.

Definition 4.1.3 *The estimator $\hat{\theta}(X_1, ..., X_n)$ is asymptotically unbiased for θ if*

$$\lim_{n\to\infty} bias(\hat{\theta}(X_1, ..., X_n)) = 0.$$

Example 4.1.4 *Let* $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, and let*

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \qquad S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where $\bar{X} = \sum_{i=1}^{n} X_1/n$*. Show that if* S_1^2 *and* S_2 *are unbiased estimators for* σ^2*. If any one of the* S_1^2 *and* S_2^2 *is not unbiased, is it asymptotically unbiased?*

Example 4.1.5 *If* $X_1, ..., X_n \overset{iid}{\sim}$ *uniform*$(0, \theta)$*, show that the maximum statistic is a biased estimator of* θ*. Also, modify this estimator to make it unbiased for* θ*.*

4.1.2 Efficiency

Recall that in the Example 4.1.2, if $X_1, ..., X_n \overset{iid}{\sim}$ Bernoulli(p), $\hat{p}_1 = (\bar{X})$ and $\hat{p}_2 = X_1$ are both unbiased estimators of p. If we are asked to choose one that would be the better one to estimate p, which one should we pick? We usually compare the variances of the two unbiased estimator, the one having the smaller variance will be more precise implying that it will be a better estimator.

Definition 4.1.6 *Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of parameter θ of a given population. $\hat{\theta}_1$ is said to be **relatively more efficient** than $\hat{\theta}_2$ if*

$$var(\hat{\theta}_1) \leq var(\hat{\theta}_2).$$

*The **relative efficiency**, the ratio of the variances of the two estimators*

$$RE = \frac{var(\hat{\theta}_1)}{var(\hat{\theta}_2)},$$

is used to measure the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$.

Example 4.1.7 *Suppose* $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. *It is known that*

$$E(X) = var(X) = \lambda.$$

Determine which one of the two estimators

$$\hat{\lambda}_1 = \bar{X}, \quad \hat{\lambda}_2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

for λ *is more efficient?*

For some unbiased estimators, it might not be trivial to find their variances. Therefore, it is hard to compare their efficiencies in terms of their sizes of variances. Can we find the lower bound for the variance of any estimator? If the variance of a given estimator equals to such a lower bound, it will be the optimal one. Here, if we only focus on the class of all unbiased estimators, the **Cramér-Rao inequality** gives the lower bound of the variances among all unbiased estimators.

Definition 4.1.8 ***Cramér-Rao inequality****: Suppose* $X_1, ..., X_n \overset{iid}{\sim}$ *iid* $f(x;\theta)$*. If* $\hat{\theta}$ *is an unbiased estimator of the parameter* θ*, then the variance of* $\hat{\theta}$ *must satisfy the inequality*

$$var(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]}$$

This bound is called the ***C****ramér-****R****ao* ***l****ower* ***b****ound (CRLB).*

Theorem 4.1.9 *If* $\hat{\theta}$ *is an unbiased estimator for* θ *and the variance of* $\hat{\theta}$ *attains the CRLB, then* $\hat{\theta}$ *is the* ***u****niformly* ***m****inimum* ***v****ariance* ***u****nbiased* ***e****stimator (UMVUE) for* θ*, and* $\hat{\theta}$ *is optimally efficient.*

Example 4.1.10 *If* $X_1, ..., X_n \sim$ *iid Poisson*(λ)*, find the CRLB and compare it with the variance of* $\hat{\lambda}_1$ *in Example 4.1.7.*

Definition 4.1.11 *The **efficiency of an unbiased estimator** of θ is the ratio of the CRLB to the variance of the estimator.*

Example 4.1.12 *If $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, show that $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the UMVUE for μ.*

So far, we focus on finding the UMVUE, the optimal estimators that has the smallest variance among all unbiased estimators. However sometimes, there is a biased estimator that has a smaller variance that the UMVUE. So, which estimator works better? Furthermore, if we have two biased estimators, which one should we take? To compare two estimators that are not necessarily unbiased, we compare their **mean square error**s.

Definition 4.1.13 *The **mean square error** of an estimator $\hat{\theta}$ of the parameter θ is given by*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

An optimal estimator should be the one that minimizes the MSE($\hat{\theta}$).

Definition 4.1.14 *If $\hat{\theta}$ is unbiased for θ, and $\lim_{n\to\infty} var(\hat{\theta}) = CRLB$, then, $\hat{\theta}$ is **asymptotically efficient**.*

4.1.3 Consistency

In general, the probability that the estimate of a parameter θ exactly equals to the true value of θ is 0. The var($\hat{\theta}$) and MSE($\hat{\theta}$) quantify the errors or fluctuations of $\hat{\theta}$ for estimating θ. It is known that if we increase the sample size n, the variance of the estimator will generally decrease. We are then interested in how close $\hat{\theta}$ will be to θ as the sample size increases. To get an idea, we study the asymptotic properties of the estimate $\hat{\theta}$. For example, suppose a population $X \sim N(\mu, \sigma^2)$. The sample mean of a random sample of size n, $\bar{X}$, would be a good estimator in terms of unbiasedness and efficiency for population mean μ. Suppose we can increase the sample size, say let $n \to \infty$ or let $n \to N$ with N being the size of a large but finite population, then, $\bar{X} \to \mu$. The increasing closeness of the estimate to the true value of the parameter θ as the sample size n increases is described by the estimator property, **consistency**.

Definition 4.1.15 *$\hat{\theta}$ is a **consistent** estimator of the parameter θ if and only if for each $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

We generally say that the estimate $\hat{\theta}$ converges to the true value of θ in probability.

Suppose $X_1, ..., X_n$ is a random sample from an infinite population with mean μ and variance σ^2, the sample mean $\bar{X}$ is an consistent estimate of the population mean.

In a similar idea, the Chebyshev's inequality leads to the theorem below.

Theorem 4.1.16 *If $\hat{\theta}$ is an unbiased estimator of the parameter θ and the var$(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is a consistent estimator of θ.*

Example 4.1.17 *Show that for a random sample from $N(\mu, \sigma^2)$, the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent estimator of the population variance σ^2.*

Example 4.1.18 *Let* $X_1, X_2, ..., X_n \overset{iid}{\sim} f(x;\delta)$ *where*

$$f(x,\delta) = \begin{cases} e^{-(x-\delta)} & x \geq \delta, \\ 0, & otherwise \end{cases}$$

Determine if the minimum statistic, $X_{(1)}$ *is unbiased, asymptotically unbiased, and consistent for* δ*.*

First of all, we determine the unbiasedness of $X_{(1)}$. Recall that

$$f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} f(x;\delta)$$

Example 4.1.18 Now, we will show that $X_{(1)}$ is asymptotically unbiased and consistent for δ.

Remark: Theorem 4.1.16 provides a sufficient condition for an estimator being consistent but not a necessary condition. We can alternatively show the consistency by the definition of consistency.

4.1.4 Sufficiency

For a random sample $X_1, ..., X_n$ from a population with pdf $f(x; \theta)$, each observation X_i provides information about the value of θ. If an estimator $\hat{\theta}(X_1, ..., X_n)$ provides all information that the sample $X_1, ..., X_n$ contains for estimating θ, $\hat{\theta}$ is said to be sufficient for θ.

Formally, we determine the sufficiency of an estimator by determining if the conditional joint distribution of $X_1, ..., X_n$ given the estimator $\hat{\theta}$ depends on the parameter θ or not.

The conditional distribution of $X_1, ..., X_n$ given $\hat{\theta}$ is given by

$f(X_1 = x_1, ..., X_n = x_n|\hat{\theta}) =$

If $f(X_1 = x_1, ..., X_n = x_n|\hat{\theta})$ depends on θ,

If $f(X_1 = x_1, ..., X_n = x_n|\hat{\theta})$ is independent on θ,

Definition 4.1.19 *The estimator $\hat{\theta}$ is a* ***sufficient*** *estimator of the parameter θ if and only if the conditional distribution/density of $X_1, ...X_n$ given $\hat{\theta}$ is independent of θ.*

Example 4.1.20 *If $X_1, X_2, X_3 \overset{iid}{\sim}$ Bernoulli(p), then show that, $Y = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$ is not a sufficient estimator of p.*

Sometime it is tedious to check whether an estimator is sufficient for a given parameter by the definition approach. Alternatively, we can use the **factorization theorem** to show the sufficiency of an estimator.

Theorem 4.1.21 *$\hat{\theta}$ is a sufficient estimator of the parameter θ if and only if the joint probability distribution/density of the random sample can be factorized as:*

$$f(X_1 = x_1, ..., X_n = x_n; \theta) = g(\hat{\theta}, \theta) \cdot h(x_1, ..., x_n)$$

where $g(\hat{\theta}, \theta)$ depends only on $\hat{\theta}$ and θ, and $h(x_1, ..., x_n)$ does not depend on θ.

The idea of the factorization theorem is to factorize the joint distribution into two parts:

(1) $g(\hat{\theta}, \theta)$

(2) $h(x_1, ..., x_n)$

Example 4.1.22 *If* $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *show that* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *is sufficient for* μ.

In some situations, we utilize the fact of the dependence of the sample space (domain of the X) on the parameter θ to show the sufficiency of a given estimator.

Example 4.1.23 *If* $X_1, ..., X_n \overset{iid}{\sim}$ *Uniform*$(0, \theta)$*, show that the maximum statistic is a sufficient estimator for the parameter* θ*.*

4.2 Methods of finding point estimators

We have seen that it is perfectly possible for there to be more than one potential estimator for the same parameter. The previous section provides us evaluation criteria of selecting a better or optimal estimators among many. In this section, we focus on different methods of constructing an estimator for a given parameter. We will introduce the simplest method: the **method of moments**, the most popular method (historically): **method of maximum likelihood**, and the increasingly popular methods of **Bayesian estimation**.

4.2.1 The method of moments (MM)

The method of moments is also known as the substitution method. It came with the simple idea of setting the sample moments equaling to the corresponding population moments, a function of parameter(s), to make up equations to solve for the unknown parameter(s).

Definition 4.2.1 *The **rth sample moment** is defined as*

$$m'_r = \frac{\sum_{i=1}^n x_i^r}{n},$$

where we recall that the population moment is defined as

$$\mu_r = E(X^r), \quad r = 1, 2, 3, ...$$

In general, if we have k parameters in the population, we need to have k equations to solve for the k parameters. So, we need k sample moments and k population moments and set them equal.

Single parameter, θ:

Suppose $X_1, ..., X_n \sim$ iid $f(x;\theta)$, let $\bar{X} = \sum_{i=1}^n X_i/n$ and find the first moment $E(X) = g(\theta)$, then

Example 4.2.2 *If* $X_1, ..., X_n \overset{iid}{\sim}$ *Exponetial*(θ)*, find the method of moments estimator (MME) for* θ*.*

Two parameters: θ_1**,** θ_2**.**

Let $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$, and $\bar{X^2} = \sum_{i=1}^{n} \frac{X_i^2}{n}$ and find $\mathrm{E}(X) = g_1(\theta_1, \theta_2), E(X^2) = g_2(\theta_1, \theta_2)$,

Example 4.2.3 *If* $X_1, ..., X_n \overset{iid}{\sim}$ *Gamma*(α, β)*, find the MM estimators for* α *and* β*.*

In some situations, it is convenient to use the variance rather than the second moment. Suppose there are two parameters, θ_1 and θ_2 for a given population. If

$$\mathrm{E}(X) = g_1(\theta_1, \theta_2), \qquad \mathrm{var}(X) = g_2(\theta_1, \theta_2).$$

Then, let $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Setting

$$\begin{aligned} g_1(\theta_1, \theta_2) &= \bar{X}; \text{ and} \\ g_2(\theta_1, \theta_2) &= S_0^2, \end{aligned}$$

we then solve for θ_1 and θ_2.

Comments about MMEs.

- No guarantee that the MM estimators are unbiased.
- Simple and easy to compute.
- Can provide a good initial value for other estimating methods that may need an initial value to start a computational/iterative procedure.

4.2.2 Method of maximum likelihood

Generally, any estimation method attempts to find parameter values that lead to a model (e.g. a distribution; a linear model; etc.) that best fits the data. The maximum likelihood approach does this by finding a value for θ such that it gives the maximum probability of observing the observed data (most likely observed data).

Example 4.2.4 *In an experiment, a coin was tossed 6 times in order to estimate the* $p = P(head)$

Let X be the number of heads among 6 tosses and $X \sim \text{Bin}(6, p)$. Suppose we observed that $X = 2$. Let us use the idea of maximum likelihood to estimate p.

Definition 4.2.5 *If $x_1, ..., x_n$ are observed values of a random sample from a population with the parameter θ, the **likelihood function** of the θ is given by*

$$L(\theta) \equiv L(\theta; x_1, ..., x_n) = f(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

which is the joint probability of distribution of the random sample $(X_1, ..., X_n)' = (x_1, ..., x_n)$ for values of θ within its given domain ($\theta \in \Theta$).

Definition 4.2.6 *The **maximum likelihood estimate** of θ is the value of θ that maximizes the likelihood function $L(\theta)$.*

How to find the maximum likelihood estimate (MLE)?

Under regular case

The method of maximum likelihood consists of maximizing the likelihood function with respect to θ. In calculus, to find the maxima, we take the derive of the function with respect to the θ, set the derivative to zero and solve for the θ. Usually, we let

$$l(\theta) = \ln L(\theta) = \ln \prod_{i=1}^{n} f(x_i; \theta) = \sum_{i=1}^{n} \ln f(x_i; \theta).$$

Lemma 4.2.7 *The $\hat{\theta}$ maximizes $L(\theta)$ if and only if $\hat{\theta}$ maximizes the $l(\theta)$.*

Example 4.2.8 *If a randome sample* $X_1, ..., X_n \overset{iid}{\sim}$ *Exponential*(θ), *find the maximum likelihood estimator for* θ.

Under irregular case: You can not use the derivative approach.

Example 4.2.9 *If* $X_1, ..., X_n \overset{iid}{\sim}$ *Uniform*$(0, \theta)$, *find the maximum likelihood estimator of* θ.

Comments on the maximum likelihood estimator.

1. If a sufficient statistic exists for θ then the MLE of θ is a function of the sufficient statistic. That is, the MLE of θ is a sufficient statistic.

2. MLE is known to be asymptotically efficient. That is,

$$\lim_{n\to\infty} \text{var}(\hat{\theta}_{MLE}) = CRLB.$$

3. Invariance principle: if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of the function $g(\theta)$.

4. Lack of uniqueness: there could be more than one MLE.

Example 4.2.10 *If* $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, find the MLE of* μ, μ^2, σ^2 *and* σ*.*

Example 4.2.11 *If* $X_1, ..., X_n \overset{iid}{\sim} Uniform(\theta, \theta+1)$, *find the maximum likelihood estimator of* θ.

4.3 Bayesian estimation

In what is termed the 'classical' approach to inference, based upon a frequentist interpretation of probability, we assume that the parameter θ is an unknown constant. We then consider the precision/variance of that parameter estimate by considering its sampling distribution.

In Bayesian estimation, based upon a view of probability as a 'degree of belief', we consider the parameter θ to be a random variable that follows a distribution. We make an assumption about the distribution that the θ follows before we see the data. This distribution is called the **prior distribution**. We denote the prior probability distribution function of θ by $g(\theta)$.

This prior distribution may be based upon expert knowledge about the system, and thus parameters, being studied; the results of an analysis of a previous data set; or, if we have little information about the parameter, we can choose a vague/non-informative prior distribution which has a very large variance.

In Bayesian estimation, we update the information about the population from the prior distribution by that in the data via the likelihood function to obtain the resulting **posterior distribution**.

Prior distribution of θ:

The likelihood of θ given the data:

Posterior distribution of θ:

Example 4.3.1 *If* $X \sim Uniform(0, \theta)$*, if we assume that* $\theta \sim Gamma(\alpha = 2, \beta = 1)$*, find the posterior distribution of* θ*.*

Note that the posterior distribution contains all the distribution information about our parameter, including its location and variance we are therefore unconcerned with the idea of a sampling distribution in Bayesian statistics.

Given the posterior distribution of θ, posterior mean of posterior median are typical **Bayesian estimate** of θ. The posterior mean of θ is given by

$$\hat{\theta}_B = \mathrm{E}(\theta|x_1, ..., x_n)$$

Example 4.3.2 *With reference to Example 4.3.1, find a Bayesian estimator of* θ.

Definition 4.3.3 *A prior distribution that leads to the posterior distribution belonging to the same distribution family is called a* ***conjugate prior****, the distribution family that both prior and posterior distributions belong to is called the* ***conjugate family****.*

The conjugate family has nice mathematical property and convenience in that, the posterior follows a known form of distribution.

Example 4.3.4 *Suppose* $X \sim$ *Binnomial*(n, p)*, we assume that* $p \sim$ *beta*(α, β)*, find the posterior distribution of* p*.*

(Example 4.3.4 continue)

The posterior distribution of p given the data follows the beta distribution:

and we could obtain a Bayesian estimator for p straightly using the mean of a beta distribution as:

Note that, if the prior is informative, then the prior has more impact on the estimator. If the prior is not informative, the estimator depends more on the data.

Finally, although Bayesian methods are useful, in almost all non-standard cases, and many standard cases, it is impossible to do the mathematics analytically in order to obtain a neat posterior distribution. Therefore, most Bayesian inference is carried out using stochastic computational approximation methods, for example, Markov chain Monte Carlo (MCMC) methods, which might require intensive computational cost.

Chapter 5

Interval Estimation

5.1 Introduction

In Chapter 4, we introduced different methods to infer/estimate the value of the parameter θ. As a point estimate $\hat{\theta}$ is a statistic in a form of a single number that seldom equals to the true value of θ. We use the mean square error to quantify the size of error, which includes the bias and the variance of the estimate. Alternatively, we could use an interval to quantify the error of the estimator.

5.1.1 Random interval – Frequentist/classical approach

Let (θ_l, θ_u) be a random interval. With an appropriate probability, say $1 - \alpha$, we want to find values of θ_l and θ_u such that

$$P(\theta_l \leq \theta \leq \theta_u) =$$

We refer to (θ_l, θ_u) as a confidence interval for θ.

Interpretation of a $(1 - \alpha)100\%$ CI:

Confidence coefficient or degree of confidence:

Confidence limits:

Question: how to determine (θ_l, θ_u) ?

5.2 Interval estimation for means

Example 5.2.1 *$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 is known. We are interested in the value of μ.*

The MLE for μ:

We know that,

Then, we want to find values of μ_l and μ_u such that:

$$P(\mu_l \leq \mu \leq \mu_u) = 1 - \alpha,$$

$\Rightarrow \quad \mu_l =$

$\Rightarrow \quad \mu_u =$

Remarks for an CI construction:

- Need to specify a degree of confidence $(1-\alpha)$;
  e.g., let $\alpha = 0.05$.
- Need to know the distribution of our estimator/pivot.
- Under a symmetric distribution, you can split the probability α equally across the tails of the distribution for the estimator.

In the previous example, if $\alpha = 0.05$, we have

$$P(-Z_{0.025} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{0.025}) = 0.95$$

Theorem 5.2.2 *If $\bar{x}$ is the sample mean of a random sample from $N(\mu, \sigma^2)$, σ^2 is known, then*

is a $(1-\alpha)100\%$ confidence interval for the parameter μ.

General rules for finding confidence interval:

- If θ is a location parameter, then the statistic usually involves a difference.

- If θ is a scale parameter, then the statistic usually involves a ratio.

- The MLE or a sufficient statistic is often a good place to start for finding θ_l and θ_u.

Example 5.2.3 *$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 is unknown. Find the 95% CI for μ.*

Example 5.2.4 *$X_1, \ldots, X_n \overset{iid}{\sim} f(x)$ with $E(X) = \mu$ and $var(X) = \sigma^2$. To find the 95% confidence interval for μ, we can use the central limit theorem for when n is large.*

5.3 Interval estimation for the difference between means

Example 5.3.1 *Suppose we have* $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ *and* $Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$. *Find the 95% CI for* $\mu_X - \mu_Y$.

Case 1: σ_X^2 and σ_Y^2 are known.

Case 2: σ_X^2 and σ_Y^2 are unknown.

1). For $n_1 \geq 30, n_2 \geq 30$.

2). For small n_1 and n_2 (< 30), the procedure for constructing CI for the $\mu_X - \mu_Y$ is not straightforward unless we assume $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Then

5.4 Interval estimation for proportions

So far, we have focussed on the problem of estimating population means. In many situations, we might need to estimate proportions, percentiles, or rates, such as the percentage of defected products of a production line, and prevalence of a disease in a given population. To estimate these quantities, we assume we sample from a binomial distribution with size of n with probability p of the event of interest.

Example 5.4.1 *$X \sim Bin(n, p)$, find the 95% CI for p.*

5.5 Interval estimation for the difference between proportions

Example 5.5.1 $X \sim Bin(n, p_1), Y \sim Bin(m, p_2)$.
Find the 95% CI for $p_1 - p_2$.

5.6 Interval estimation for variances

The variance is a scale parameter. It measures the spread of a population. Recall the general rule of interval construction, the interval estimation for the variance involves a ratio.

Review of sampling distribution results.

Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then if $\overline{X} = \sum_{i=1}^n X_i/n$ and $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$:

1. $\overline{X} \sim N(\mu, \sigma^2/n)$,
2. $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2 \sim \chi^2_n$,
3. $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 \sim \chi^2_{n-1}$,
4. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,
5. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Example 5.6.1 *If* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *find the 95% CI for* σ^2.

We consider two cases: μ is known and μ is unknown.

Case 1: μ is known, use $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2 \sim \chi^2_n$.

Case 2: μ is unknown.

- Estimate μ by $\hat{\mu} = \bar{X}$.
- Use $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$, or, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Chapter 6

Hypothesis Testing

6.1 Introduction

Point estimation involves finding an estimate of the parameter, θ, that is as close as possible to the true value of θ based on the observed data.

Interval estimation involves giving a range to the value of θ, such that, the true value of θ would be expected to fall in with a specified degree of confidence. The estimate interval can be used to quantify the error of a estimator.

Point estimation and interval estimation do not necessarily answer specific queries such as:

$$\text{is } \theta = \theta_0?; \qquad \text{is } \theta > \theta_0?; \qquad \text{etc.}$$

where θ_0 is a specified value. The result of such queries may be important in determining some future course of action.

Example 6.1.1 *Conviction of an accused criminal depends on whether there is a DNA match in blood found at the crime scene with the defendant's blood.*

Example 6.1.2 *A biologist may be interested in known if a gene will express differently under two different conditions.*

Example 6.1.3 *A marketing analyst needs to make conclusion about the market sharing of a given product on the basis of sample data.*

These problem can be formulated as **statistical hypothesis testing problems**. We try to answer these problems through statistical hypothesis tests. Suppose $X_1, ..., X_n$ is a random sample from a population that follow a distribution with probability distribution function $f(x, \theta)$. A statistical hypothesis can be a statement about the parameter.

6.1.1 Basic idea and definitions

Example 6.1.4 *Suppose a new cat diet claims to help obese cats reduce weight by more than 2lbs in one month. Previous research suggests that the weight reduction on such a diet would be normally distributed with mean μ and variance 1.2.*

The conjecture $\mu > 2$ is a *statistical hypothesis*. If the conjecture is false then the complementary hypothesis, $\mu \leq 2$ would be true.

To investigate whether the claim valid or not not, a random experiment is conducted to generate data, the above hypothesis is tested based on the observed data.

- Place a random sample of obese cats on new diet for one month and record the weight reductions.

- Assume the claim is not true, i.e, assume $\mu \leq 2$ is true. (Null hypothesis, H_0.)

- Obtain a statistic that estimates μ, say $\hat{\mu}$. Compute the probability that $\hat{\mu} > \mu$ under the H_0.

- Make a decision with regard rejecting or not rejecting H_0.

Definition 6.1.5 *A **statistical hypothesis** is a statement about the population (usually the distribution of a population).*

A statement that fully specifies the distribution of a population is call a *simple hypothesis*. For example, suppose the weight reduction follows $N(\mu, 1.2)$, a statement such as statement $\theta = 2$.

Otherwise, the statement is called a *composite hypothesis*. For example, a statement such as $\mu \leq 2$ (more than one value of μ).

Null hypothesis, H_0:

Alternative hypothesis, H_a:

General setting: null hypothesis vs alternative hypothesis.

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_a : \theta \in \Theta_1$$

where parameter space $\Theta = \Theta_0 \cup \Theta_1$, and $\Theta_0 \cap \Theta_1 = \phi$.

As a result, we can accept one of H_0 or H_a, but not both.

6.2 Testing a statistical hypothesis

The decision regarding which one of H_0 or H_a is accepted is based on the information we have from the data.

Note:

i) If our decision is "H_0 is true", we say we "accept" or "do not reject H_0", otherwise, we say we "reject H_0".

ii) We have only limited information about the entire population. Therefore we never know conclusively which one of H_0 or H_a is true. Sometime we will make the wrong decision.

iii) A good decision-making rule is desired such that the chance of making the wrong decision is small.

iv) There are two types of errors that we can commit:
type I error and type II error.

Type I and type II error

Two possible outcomes of a hypothesis test: reject or do not reject, H_0 .

Actual fact (in the population): either "H_0 is true" or the "H_0 is false".

	Decision	
Population	reject	do not reject
true H_0	type I error	
false H_0		type II error

Critical region and test function

Given a random sample $(x_1, \ldots, x_n)$, a general procedure for testing the hypothesis is as follows:

- Assume H_0 is true, so that each observation of the random sample follows the distribution described by H_0.

- Partition the sample space $\Omega = \{\text{all possible values of } (x_1, \ldots, x_n)\}$ into two regions: C and its complement C^c.

- Decision rule:

  if $(x_1, \ldots, x_n) \in C$, reject H_0;

  if $(x_1, \ldots, x_n) \in C^c$, do not reject H_0.

Definition 6.2.1 *The test procedure partitions the sample space into two regions: an* ***acceptance region*** *for H_0 and a* ***rejection region*** *for H_0. The rejection region of a test is sometime referred as the* ***critical region****.*

Definition 6.2.2 *When the H_0 is true, the probability of obtaining a value of test statistic that the corresponding random samples fall inside the critical region is called the* ***size of the critical region*** *or the* ***level of significance*** *of the test.*

Test function:

$$\phi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } (x_1, \ldots, x_n) \in C, \\ 0 & \text{if } (x_1, \ldots, x_n) \in C^c. \end{cases}$$

If $\phi(x_1, \ldots, x_n) = 1$, reject H_0; if $\phi(x_1, \ldots, x_n) = 0$, do not reject H_0.

Note: we allow ϕ to take values between 0 and 1.

Refer to the obese cat example (Example 6.1.4), suppose we are testing about the mean of weight lost that

$$H_0 : \mu = 2 \qquad \text{vs} \qquad H_a : \mu = 3.$$

Our decision rule could be: reject H_0 if and only if $\overline{X} > 2.6$.

Questions:

1) What is the probability that we make a false rejection (type I error rate, α)?
2) What is the chance that we do not reject a false H_0 (type II error rate, β)?

We know that the distribution of sample mean $\overline{X}$ is:

Let's compare the probability of type I error (α) and the probability of type II error (β).

$\alpha = P(\text{reject } H_0 | H_0 \text{ is true}) =$

$\beta = P(\text{do not reject } H_0 | H_0 \text{ is false}) =$

If we want to have a small probability of type I error, what should we do and how this affects the probability of a type II error?

If we want to have a small probability of type II error, what should we do and how this affects the probability of a type I error?

Generally, it is felt more important to control type I error. So, we usually choose a decision rule based on a pre-specified level of significance.

For example, we specify $\alpha = 0.05$, then our decision rue is derive based on the pre-specified α?

Remark:

- The critical region C is implemented on the observed values of a statistic, and the decision rule depends on the distribution of that statistic.
- For a discrete distribution, it may be impossible to devise a rule with a specific α, in which case, we usually specify the maximum size of the critical region to be α.

Example 6.2.3 *Suppose X is the observed number of successes in 20 trials with the probability of success, p, for each trial.*

Let's say that to test $H_0 : p = 0.9$ vs $H_a : p = 0.6$, we decide to fail to reject H_0 if $X > 14$; otherwise, we reject H_0.

Find the probabilities of type I error and type II error.

6.3 The Neyman-Pearson theory

We have seen that in a hypothesis test, as α becomes smaller, β becomes larger, and thus, the power (1-β) becomes smaller. As we try to increase the power, we also increase α. To balance the tradeoff between α and power, we usually pre-specify the significance level and then we search for a critical region that maximizes the power of the test. Neyman and Pearson was the first to propose the "first fix α then maximize the power" approach.

Definition 6.3.1 *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} f(x, \theta)$. *To test* $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_1$, *let* C *be the critical region such that:*

If for any other critical region C^* *such that*

we have

then we say C *is the* ***best critical region*** *or the* ***most powerful test****.*

There are many ways to partition the sample space to form a critical region with size α. The goal is to find the one that maximizes the power among all partitions with size α of the test. The Neyman-Pearson Lemma instructs us how to find the best critical region.

Example 6.3.2 *Suppose* $X \sim bin(5, p)$.
We want to test $H_0 : p = \frac{1}{2}$ *vs* $H_a : p = \frac{3}{4}$.

	X	0	1	2	3	4	5
Under H_0:	$f(x; \frac{1}{2})$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
Under H_a:	$f(x; \frac{3}{4})$	$\frac{1}{1024}$	$\frac{15}{1024}$	$\frac{90}{1024}$	$\frac{270}{1024}$	$\frac{405}{1024}$	$\frac{243}{1024}$
$\frac{L_0}{L_1}$	$\frac{f(x;\frac{1}{2})}{f(x;\frac{3}{4})}$	32	$\frac{32}{3}$	$\frac{32}{9}$	$\frac{32}{27}$	$\frac{32}{81}$	$\frac{32}{243}$

Suppose we set $\alpha = \frac{1}{32}$.Then:

Thus, there are two candidate critical regions:

Now, let's compare power:

It is obvious that

Alternatively, we can look at probability (likelihood) ratios,

$$\frac{L_0}{L_1} = \frac{f(x; \frac{1}{2})}{f(x; \frac{3}{4})}.$$

For all points in critical region C with $P(x \in C|H_0 \text{ is true}) = P(x \in C|p = \frac{1}{2}) = \alpha$, we want $f(x; \frac{1}{2})$ to be small in comparison to $f(x; \frac{3}{4})$.

In this example, the ratio is minimized when

Now, suppose we change α to $\frac{6}{32}$. Then, there are four candidate critical regions.

Critical region	Power	Ratio
$C_1 =$		
$C_2 =$		
$C_3 =$		
$C_4 =$		

Lemma 6.3.3 ***(Neyman-Pearson Lemma)*** *If C is a critical region of size α, and k is a constant such that*

$$\frac{L_0}{L_1} = \frac{f(x_1,\ldots,x_n;\theta=\theta_0)}{f(x_1,\ldots,x_n;\theta=\theta_1)} \leq k, \ \ when\ (x_1,\ldots,x_n) \in C,$$

$$\frac{L_0}{L_1} = \frac{f(x_1,\ldots,x_n;\theta=\theta_0)}{f(x_1,\ldots,x_n;\theta=\theta_1)} \geq k, \ \ when\ (x_1,\ldots,x_n) \notin C,$$

then C is the best critical region of size α for testing $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$.

Idea: If H_0 is true, the likelihood under H_0 should be greater than the likelihood under H_a, that is $L_0 > L_1$ and the ratio of $\frac{L_0}{L_1}$ is large. Conversely, if H_0 is false, we expect that $L_0 \leq L_1$, so that the ratio of $\frac{L_0}{L_1}$ is small.

Thus, we want to find a critical region C such that for $(x_1,\ldots,x_n) \in C$, we have a small ratio of $\frac{L_0}{L_1}$. The value of k is determined based on the pre-specified α, such that for $\frac{L_0}{L_1} \leq k$, the ratio is believed to be small enough to reject the H_0.

The Neyman-Pearson Lemma guarantees a most powerful critical region when both the H_0 and H_a are simple hypothesis.

Example 6.3.4 *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. *To test* $H_0 : \mu = \mu_0$ *vs* $H_a : \mu = \mu_1$, *where* $\mu_1 > \mu_0$, *let us use the Neyman-Pearson Lemma to find the most powerful critical region of size* α.

The likelihood ratio is:

Now, we want to find a constant K and a region C such that

In fact, we don't really care what the value of K is, we only care for what value of K^*:

$$\overline{x} \geq K^*, \ \ (x_1, \ldots, x_n) \in C,$$
$$\overline{x} \leq K^*, \ \ (x_1, \ldots, x_n) \notin C.$$

We determine the value of K^* based on the size of test α and the distribution of $\overline{X}$.

Let's consider another application of N-P lemma:

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu, 1)$. Find the most powerful test for $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_1$ $(\mu_1 > \mu_0)$ at the α level of significance.

By the N-P lemma, the most powerful test is given by

$$\phi(x_1, \ldots, \psi_n) = \begin{cases} 1 & \text{if } \Lambda \leq k, \\ 0 & \text{if } \Lambda < k, \end{cases}$$

where

The value of k is determined by the size of test α.

Under $H_0 : \mu = \mu_0$,

Thus, the rejection region $\{\Lambda \leq k\}$ is equivalent to

6.4 The power function of a test

In general, type I error is more series that type II error. Therefore, we control the α at a pre-specified level, then find a critical region, C, based on the given α that maximizes the power. By doing so, the probability of type I error is controlled at α level and the power $(1-\beta)$ is maximized.

The Neyman-Pearson lemma is for testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_a : \theta = \theta_1$. We might want to test, say, $H_0 : \theta < \theta_0$ a composite null hypothesis against $H_a : \theta > \theta_0$, a composite alternative hypothesis, pair of composite hypotheses.

Let us consider a framework:

Definition 6.4.1 *The **power function**, denoted as $\pi(\theta)$, of a test of $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$ is given by*

$$\pi(\theta) = \begin{cases} \alpha(\theta) & \text{for value of } \theta \text{ assumed under } H_0, \\ 1 - \beta(\theta) & \text{for value of } \theta \text{ assumed under } H_a. \end{cases}$$

The power function, $\pi(\theta)$, is in fact, the probability of rejecting the H_0 for a given value of θ:

$$\pi(\theta) = P(\text{reject } H_0 | \theta)$$

Example 6.4.2 *Suppose $X \sim bin(5, \theta)$. We want to test*

$$H_0 : \theta \leq \frac{1}{2} \qquad vs \qquad H_a : \theta > \frac{1}{2}$$

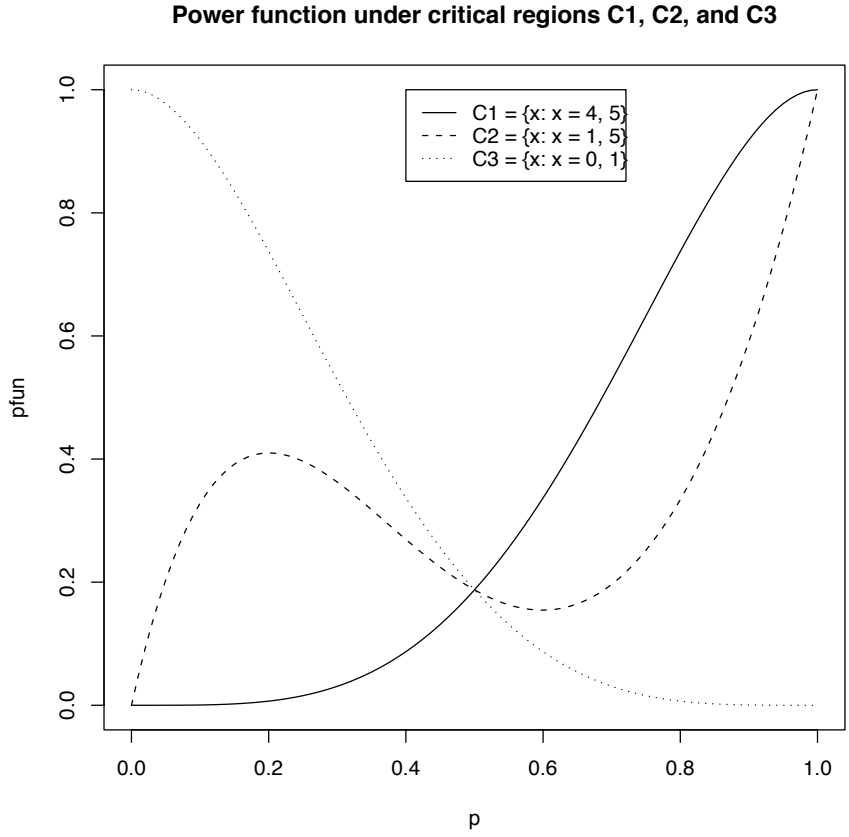
Partitioning of sample space:

Suppose our critical region is $C_1 = \{x : x \in \{4, 5\}\}$. Then the power function is given by:

We sketch the power function for the given critical region $C = \{x : x \in \{4, 5\}\}$.

	θ	$\pi(\theta)$
$\theta \in \Theta_0$	0.1	$\alpha(0.1) =$
	0.2	$\alpha(0.2) =$
	...	...
	0.5	$\alpha(0.5) =$
$\theta \in \Theta_1$	0.6	$1 - \beta(0.6) =$
	...	...
	0.9	$1 - \beta(0.9) =$

We sketch the power function for other critical regions $C_2 = \{x : x \in \{1, 5\}\}$ and $C_3 = \{x : x \in \{0, 1\}\}$.



Definition 6.4.3 *Given a pre-specified significance level α, if a test*

$$\phi(x_1, \ldots, \psi_n) = \begin{cases} 1 & \text{if } (x_1, \ldots, x_n) \in C, \\ 0 & \text{if } (x_1, \ldots, x_n) \notin C, \end{cases}$$

satisfies $P(\text{reject } H_0 | H_0 \text{ is true}) \leq \alpha$, then the test is called an α ***level significant test****.*

Definition 6.4.4 *An α level significant test with the smallest β (or the greatest power) is called the* ***uniformly most powerful test*** *(UMPT).*

Remarks:

1. There could be multiple tests (rejection regions) at a given α level; we want the one that maximizes the power.
2. Unfortunately, uniformly most powerful tests rarely exist when testing a simple null hypothesis versus a composite alternative hypothesis, e.g.,

3. When testing a simple null hypothesis versus a simple alternative hypothesis, e.g.,

   the N-P lemma gives the uniformly most powerful test.

6.5 The likelihood ratio tests

The Neyman-Pearson lemma provides a method for constructing the most powerful critical region for testing:

We now present a general method, the likelihood ratio test (LRT), for constructing critical regions for the hypothesis tests that consist of composite hypothesis such as:

LRTs are generalization of the Neyman-Pearson lemma, but they are not necessarily uniformly most powerful. LRTs compare the maximum likelihood under H_0 with the unrestricted maximum likelihood for all values in the parameter space, that is $\theta \in \Theta$.

Suppose we have a random sample $(X_1, \ldots, X_n) \overset{iid}{\sim} f(x;\theta)$. The maximum likelihood under H_0 is given by

The maximum likelihood for all values of $\theta \in \Theta$, is given by

Then, their ratio

$$\Lambda = \frac{\max L_0}{\max L}$$

is referred to the **likelihood ratio statistic**.

Suppose we have

$$\max L_0 = L(\tilde{\theta}) \quad \leq \quad \max L = L(\hat{\theta})$$

where $\tilde{\theta}$ is

and $\hat{\theta}$ is

The equality holds iff $\tilde{\theta} = \hat{\theta}$.

- There are two scenarios to consider:
  - If H_0 is true, we expect:

– If H_0 is false, we expect:

- The ratio
$$\Lambda = \frac{\max L_0}{\max L}$$
is bounded between 0 and 1.

- If $\Lambda \approx 0$, we would like to reject H_0;
if $\Lambda \approx 1$, we would like to accept H_0.

Definition 6.5.1 *If $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$, and if*

$$\Lambda = \frac{\max L_0}{\max L} = \frac{L(\tilde{\theta})}{L(\hat{\theta})},$$

then the critical region

$$\Lambda \leq k$$

*where $0 < k < 1$, is a **likelihood ratio test** for testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$.*

Example 6.5.2 *Suppose we have a random sample $(X_1, \ldots, X_n)$ from a $N(\mu, \sigma^2)$. Find the critical region of the likelihood ratio test for testing*

$$H_0 : \mu = \mu_0 \qquad vs \qquad H_a : \mu \neq \mu_0.$$

Since the only choice for μ under H_0 is μ_0, we have

Further, we know the MLE of μ is $\hat{\mu} = \overline{X}$, so have

The likelihood ratio statistic becomes

Hence the critical region of the likelihood ratio test can be derived as

We determine the critical region by the size of the test, α:

We know $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$ under H_0, so we have:

In Example 6.5.2, when the random sample is from a normal distribution, it is relatively easy to find the critical region for the test, since we know the distribution of the pivotal quantity for estimating the parameter. This means we don't have to derive the distribution of Λ. However, the distribution of Λ is often difficult to derive, and thus, it is often difficult to determine the critical value k. In this case, we can use the following approximation.

Theorem 6.5.3 *For a large sample size, n,*

$$-2\ln\Lambda = -2\ln\left(\frac{\max L_0}{\max L}\right) \quad \sim \quad \chi^2_1.$$

With reference to Example 6.5.2, we can find critical region using Theorem 6.5.3.

Chapter 7

Test of Hypothesis of Means, Variances and Proportions

7.1 Introduction

In general, we can formulate a hypothesis test in the form of: $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$, for $\Theta_0 \cap \Theta_1 = \emptyset$, and $\Theta_0, \Theta_1 \subseteq \Theta$. In many situations, a simple null hypothesis against a composite alternative hypothesis can be formulated in different forms.

Two-sided alternative:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta \neq \theta_0,$$

Intuitively, we compare the point estimate $\hat{\theta}$ with θ_0, and our decision of the test is based on

- we don't want to reject H_0 if $\hat{\theta}$ is 'close to' θ_0;
- we would like to reject H_0 if $\hat{\theta}$ is 'much smaller' or 'much larger' than θ_0.

Then, we consider both tails of the distribution of $\hat{\theta}$ when constructing the critical region.

With reference to the Example 6.5.2, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$,

and we test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0.$$

Using the likelihood ratio statistic, the critical region for an α level test is then given by

$$|\bar{x} - \mu_0| \geq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

We refer this type of tests as a **two-sided test**.

One-sided alternative:
Case 1:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta < \theta_0.$$

- we don't want to reject H_0 if $\hat{\theta}$ is 'close to' θ_0;
- we would like to reject H_0 if $\hat{\theta}$ is 'much smaller' than θ_0.

Then, we only look at the tail of the distribution of $\hat{\theta}$ that takes on small values when constructing the critical values.

With reference to the Example 6.5.2, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and we test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu < \mu_0.$$

Case 2:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta > \theta_0.$$

- we don't want to reject H_0 if $\hat{\theta}$ is 'close to' θ_0;
- we would like to reject H_0 if $\hat{\theta}$ is 'much larger' than θ_0.

Then, we only look at the tail of the distribution of $\hat{\theta}$ that takes on large values when constructing the critical values.

With reference to the Example 6.5.2, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and we test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu > \mu_0.$$

General procedure for constructing tests:

1. Formulate H_0 and H_a, and specify the size of the test α.

2. Find an appropriate test statistic – generally a pivot with a distribution not dependant upon the parameter(s) under the H_0.

3. Determine the critical region based on the size of test α.

4. Compute the value of the observed test statistic based on the sample data.

5. Make decision about rejecting or failing to reject (accepting) H_0 based on the following equivalent criteria:

   (a) if the observed value of test statistic falls in critical region, reject H_0; otherwise, do not to reject H_0.

   (b) if the p-value (the probability of observing a value that is more extreme than the observed statistic under H_0) is less than or equal to α, reject H_0; otherwise, do not reject H_0.

Concept of p-value:

$\alpha =$

p-value $=$

7.2 Tests involving means

Example 7.2.1 *(Miller and Miller's Text) The specifications for a certain kind of ribbons require a mean breaking strength of 185 pounds. Five pieces of ribbon were randomly selected form different rolls having breaking strengths of 171.6, 191.8, 178.3, 184.9 and 189.1 pounds. Use this data to test* $H_0 : \mu = 185$ *against* $H_a : \mu < 185$ *at the 0.05 level of significance.*

Alternatively, we compare the p-value with the significance level.

7.3 Tests for the difference between means

Many studies involve the comparison between two populations. For example, we may be interested in knowing if women perform some general computational tasks at the same rate as men or not. Or, we want to know if Canadians spend more time (in hours) in watching TV than American on average? This type of population comparison problems can be formulated as a hypothesis test for the difference between two population means.

Let $X_1, \cdots, X_{n_1}$ be a random sample from a $N(\mu_1, \sigma_1^2)$ distribution and $Y_1, \cdots, Y_{n_2}$ be a random sample from a $N(\mu_2, \sigma_2^2)$ distribution. At the 0.05 level of significance, we want to test

$$H_0 : \mu_1 - \mu_2 = \delta \quad \text{vs} \quad H_a : \mu_1 - \mu_2 \neq \delta.$$

The population variances σ_1^2 and σ_2^2 are known.

The population variances σ_1^2 and σ_2^2 are unknown.

We consider two cases:

1. Sample sizes of n_1 and n_2 are large ($n_1, n_2 \geq 30$).

2. Either or both sample sizes are small (≤ 30).

Example 7.3.1 *(Miller and Miller's Text) The coverage (in square feet) of two brands of cans of paint are compared. Four 1-gallon cans of one brand cover on average 546 square feet with a a sample standard deviation of 31 square feet, whereas four 1-gallon cans of another brand cover on average 492 feet with a sample standard deviation of 26 square feet. Assuming that cans from each brand are sampled from two normal population with equal variances, test if two population have the same mean or not at the 0.05 level of significance.*

7.4 Tests for variances

In hypothesis tests involving variances, similar to the hypothesis test of means, we might have one sample problem that we want to test whether the variance of a given population (σ^2) equals to a specific value (σ_0^2) or not. In two samples problem, we might want to test whether the two populations have equal variances or not. For example, in the previous section, when we test whether the two population means are equal or not and the two population variances are unknown, to use the pooled variance method, we assume that the two variances $\sigma_1^2 = \sigma_2^2$. In this case, to validate the assumption of equal variances, we test whether σ_1^2 equals σ_2^2 or not.

Recall the general procedure for constructing tests, in step 2, we need to find an appropriate test statistic, a pivot statistic, for a hypothesis test involving variance. We can use the results 2, 3, and 4 of sampling distribution in Section 5.6. for one sample problems.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then if $\overline{X} = \sum_{i=1}^n X_i/n$ and $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$:

2. $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2 \sim \chi_n^2$,
3. $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 \sim \chi_{n-1}^2$,
4. $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$,

7.4.1 Test for one variance

Given a random sample $X_1, \cdots, X_n$ from a $N(\mu, \sigma^2)$ distribution, we want to test if $\sigma^2 = \sigma_0^2$ or not at an α level of significance.

Case 1: μ is known.

Case 2: μ is uknown.

7.4.2 Test for comparing two variances

Just as we might be interested in comparing two population means, we may also be interested in comparing two population variances: σ_1^2 and σ_2^2.

Suppose we have two random samples: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$. How do we test if $\sigma_1^2 = \sigma_2^2$ or not. What test statistic should we use?

Recall Theorem 1.7.2, if S_1^2 and S_2^2 are the sample variances of two random samples of size n_1 and n_2 from two populations with the variances σ_1^2 and σ_2^2, then,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Example 7.4.1 *(Miller and Miller's Text) Assume we have two random samples sampled from two normal populations with unknown mean and unknown variances. The first sample with sample size* $n_1 = 13$ *yields a sample variance of* $s_1^2 = 19.2$*, and the second sample with sample size* $n_2 = 16$ *yields a sample variance of* $s_2^2 = 3.5$*. We want to test if* $\sigma_1^2 = \sigma_2^2$ *at the significance level of* $\alpha = 0.02$*.*

7.5 Tests concerning proportions

Suppose we have $X \sim$ Bin (n, θ), and we want to test

$$H_0 : \theta = \theta_0, \quad \text{vs} \quad H_a : \theta \neq \theta_0$$

Here, X is discrete random variable taking values of 0, 1, ..., n, and so, its cdf is a step function. To construct critical region for a two-sided alternative test at the α level of significance, we follow the below procedure.

For a two-sided test, we want to define the critical regions as such, if

$$X \geq K\left(\frac{\alpha}{2}\right) \quad \text{or} \quad X \leq K'\left(\frac{\alpha}{2}\right)$$

we reject H_0 .

For one-sided test: $H_0 : \theta = \theta_0$, against $H_a : \theta > \theta_0$, we define the critical region for an α level test as:

Example 7.5.1 *(Miller and Miller's Text) If $x = 4$ of $n = 20$ patient suffered serious side effects from a new medication, test $H_0 : \theta = 0.5$ against the alternative hypothesis $H_A : \theta \neq 0.5$ at the $\alpha = 0.05$ level of significance. Here θ is the true proportion of patients suffering serious side effects from the new medication.*

Two approaches to test the hypothesis:

1. Compute the p-value.

2. Construct the critical region.

For small value of n (≤ 20), it is feasible to compute the p-value or to identify the critical region. For large n, we can use the normal approximation to the binomial distribution. To test

$$H_0 : \theta = \theta_0, \quad \text{vs} \quad H_a : \theta \neq \theta_0$$

at the α level, we construct the test statistic by using the central limit theorem.

7.6 Test concerning differences among k proportions

Suppose we observe $X_1, \cdots, X_k$ where each observation X_i from an independent binomial trial (distribution):

$$X_i \sim \text{ Bin}(n_i, \theta_i); \quad i = 1, 2, \ldots, k.$$

We want to test

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = (\theta_0) \quad \text{vs} \quad H_a : \exists i \neq j, \quad \text{s.t} \quad \theta_i \neq \theta_j.$$

- There are k independent populations.

- If the n_i are sufficiently large for each population, we construct k independent test statistics

- By central limit theorem, we have

- Recall that if $Z_1, ...Z_k \overset{iid}{\sim} N(0, 1)$, then,

  and therefore we can construct a summary statistic for our test:

- Under H_0:

- Thus, we reject H_0 if

Let us consider an alternative formulation of the chi-square statistic, χ^2, that lends itself more rapidly to other applications. Suppose with observed $X_1, \ldots, X_k$ from k independent trials of size $n_1, \ldots, n_k$, respectively. We summarize the observed data in the following $k \times 2$ table:

	successes	failures
sample 1	x_1	$n_1 - x_1$
sample 2	x_2	$n_2 - x_2$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
sample k	x_k	$n_k - x_k$

Let f_{ij} be the observed frequencies; equal to the value in the ith row and jth column in the table. The χ^2 statistic summarizes the difference between the observed frequencies f_{ij} and the expected frequencies E_{ij} of each cell in the table. Under the hypothesis that $\theta_1 = \theta_2 = \cdots = \theta_k = \theta_0$, the expected cell frequencies E_{ij}'s are given by:

Example 7.6.1 *(Miller and Miller's Text) Based on the data summarized in the table below, test whether the true proportions of shoppers favouring detergent A and detergent B are the same in all three cities at the* $\alpha = 0.05$ *level of significance.*

	# favoring A	# favoring B	Total (n_i)
Los Angeles	232	168	400
San Diego	260	240	500
Fresno	197	203	400
Total	689	611	1300

7.7 The analysis of an $k \times c$ table

Suppose instead of have only two possible outcomes (e.g. yes or no), we have $c \geq 2$ possible outcomes. For example, in Example 7.6.1, we may ask the shopper whether he/she prefers detergent A, detergent B, or has no preference. Then we may collect and summarize data as follows:

	# favoring A	# favoring B	# no preference	Total ($n_{i.}$)
Los Angeles	174	93	133	400
San Diego	196	124	180	500
Fresno	148	105	147	400
Total ($n_{.j}$)	518	322	460	1300

In this more general case, we might want to test wheter the distributions of shoppers preferring A, B, or having no preference, are the same among the three cities.

$$\begin{aligned} H_0: &\quad \theta_{11} = \theta_{21} = \theta_{31}, \text{ and } \theta_{12} = \theta_{22} = \theta_{32} \\ &\qquad \text{vs} \\ H_a: &\quad \theta_{11}, \theta_{21}, \theta_{31} \text{ are not all equal, or,} \\ &\qquad \theta_{12}, \theta_{22}, \theta_{32} \text{ are not all equal, or both.} \end{aligned}$$

Under the $H_0 : \theta_{11} = \theta_{21} = \theta_{31} = \theta_1,$ and $\theta_{12} = \theta_{22} = \theta_{32} = \theta_2.$

Here, we present more application examples of using χ^2 test using a $k \times c$ table. Suppose we interviewed 40 boys and 40 girls for which one is his/her biggest fear among the snake, blood, and darkness. Data are summarized in the following table.

	Biggest fear			
Gender	Snakes	Blood	Dark	Total ($n_{i.}$)
Boys	12	19	9	40
Girls	15	13	12	40
Total ($n_{.j}$)	27	32	21	80

Notation:

X_{ij} = the number of observations in the ith population with the jth outcome, $i = 1, 2$ and $j = 1, 2, 3$.

θ_{ij} = the true proportion of subjects with the jth outcome in the ith population , $i = 1, 2$ and $j = 1, 2, 3$.

$n_{i.}$ = row total for population i,
i.e., total number of observations in population i.

$n_{.j}$ = column total for outcome j,
i.e., total number of subjects expressing j^{th} outcome.

We are typically interested in answering two questions about the system described by our contingency table.

Question 1: Is the distribution of biggest fears for boys the same as that of girls?

$$H_0 : \theta_{11} = \theta_{21} \text{ and } \theta_{12} = \theta_{22}$$
$$\text{vs}$$
$$H_a : \theta_{11} \neq \theta_{21}, \text{ or } \theta_{12} \neq \theta_{22} \text{ or both.}$$

Question 2: Are gender and biggest fear independent?

In general, we can interpret the joint probability of two events as

$$\begin{aligned} \pi_{ij} &= \text{P}(j^{th} \text{ outcome and in population } i) \\ &= \text{P}(\text{a sample is from population } i \text{ and expresses } j\text{th outcome}). \end{aligned}$$

7.8 Test of the goodness of fit

The χ^2 tests discussed so far can be thought of as testing whether the data comes from a specified distribution, e.g., binomial distribution, multinomial distibution. If that is the case, the observed data should be close to what we would expect under the given distribution.

With this principle in mind, we could move one step further. Given a set of observed data, we could assume that the observed data come from a specified distribution. Then, a hypothesis test is often conducted to attempt to validate our assumption.

Essentially, a **goodness-of-fit test** is used to test how well a proposed model (distribution) fits the observed data. The χ^2 test is often a common choice for carrying out a goodness-of-fit test.

Example 7.8.1 *(Miller and Miller's Text) Based on the data shown in the table below, we want to test whether the number of errors a compositor makes in setting a galley of type is a random variable having a Poisson distribution.*

Number of errors	Observed frequencies f_i	Poisson probability with $\lambda = 3$	Expected frequencies E_i
0	18		
1	53		
2	103		
3	107		
4	82		
5	46		
6	18		
7	10		
8	2		
9	1		

Step 1. Estimate the Poisson parameter λ.

Step 2. Compute the probability for each observation under the Poisson distribution (λ).

Step 3. Compute the expected frequencies.

Step 4. Test the goodness-of-fit of the Poisson distribution to the observed data.

If Poisson distribution is an adequate distribution that fits the observed data well, we expect to see that the f_i are close to the E_i. Then, $(f_i - E_i)^2$ would be small and, thus, the statistic $\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)^2}{E_i}$ would be small.

$$H_0 : X \sim \text{Poisson}(\lambda), \quad \text{against} \quad H_a : X \not\sim \text{Poisson}(\lambda).$$

Chapter 8

Non-parametric Tests

8.1 Introduction

In Chapter 6 and 7, we introduced hypothesis testing procedures that are applicable to a wide range of practical problems. However, a major drawback is that these testing procedures are derived based on an assumption about the distribution of the random variables (observations) of the study. That is, we assume that the data collected is a random sample from a given population with a known distribution that depends on the unknown parameters. The whole testing procedure is then focusing on the parameter(s), for example, we test whether an unknown parameter equals a specified value.

As these hypothesis tests depend heavily on the correctness of distribution assumption of a given population, the violation of the distribution assumption might seriously affect the significance level, i.e., the control of the type I error. Violations of the distribution assumption may due to improper assumption of the distribution of the population, or could be due to, say, undetected outliers. Here is an example where the distribution assumption is violated.

Example 8.1.1 *When estimating the average income (μ) of a given population, a random sample from a normal population of mean μ and*

variance σ^2 is assumed. Consider a hypothesis test of

$$H_0 : \mu = \mu_0 \quad vs \quad H_a : \mu > \mu_0.$$

An often used test statistic and critical region are given as:

However, in reality, distributions of income are often highly skewed.

Since the underlying distribution of income is not normal, the resulting t statistics will not follow a t distribution. Then an α level critical region defined based on the symmetric t distribution may result in a test with a smaller significance level of $\alpha^* < \alpha$. In other words, we use a more stringent threshold value than we mean to, and the resulting test will be less powerful.

In statistical inference, another concept of "robustnss" was in introduced in Chapter 4 in the connection with the evaluation of the parameter estimate. We here provide a definition to "robustness".

Definition 8.1.2 *An estimator is said to be robust if its sampling distribution is not seriously affected by the violation of underlying assumptions made.*

It is difficult to ascertain or determine whether an estimator is robust or not. Extend this concept to the problem of hypothesis test, a test is robust if the sampling distribution of the test statistic is not seriously affected by the violation of underlying assumptions made. In general, we have the following consequence:

Violation of model assumption $\Rightarrow$
A wrongly specified distribution of test-statistic $\Rightarrow$
A wrongly specified the critical region for a given level of significance.

For the purpose of conducting a robust hypothesis test, an alternative is to use **non-parametric methods** which are based upon test statistics that are **parametric distribution free** (i.e. make no assumption about the population in a parametric form).

In general, non-parametric tests are robust compared to equivalent parametric tests. For example, if we thought that the data came from a normal population, but in fact it did not, the power and significance level of a non-parametric test would be unaffected; not necessarily the case for parametric hypothesis.

Question: If we don't know the underlying distribution, then what aspect of the data provides us information about the underlying distribution for, say, comparative purposes, in a non-parametric setting.

Suppose we wish to compare two samples collected from 2 populations:

$$X_1, X_2, \cdots, X_n \qquad \text{and} \qquad Y_1, Y_2, \ldots, Y_m.$$

Use this data, we want to test if the two population have the same distribution.

Or we may wish to test if one distribution is shifted somewhat relative to the other.

8.2 Sign test

Suppose we have a random sample $(X_1, X_2, \ldots, X_n)$ and we want to test if the median of the population $\tilde{\mu} = \tilde{\mu}_0$ or not.

How many observations do we expected to be greater than the median of a population?

Under the null hypothesis, how many observations do we expected to be greater than $\tilde{\mu}_0$?

Let D be the number of observations greater than $\tilde{\mu}_0$. What is the distribution of D?

Thus, to test

$$H_0 : \tilde{\mu} = \tilde{\mu}_0 \qquad \text{vs} \qquad H_a : \tilde{\mu} \neq \tilde{\mu}_0,$$

at the α level of significance, we use:

test statistic:

critical region:

This is called the **sign-test**.

Example 8.2.1 *The following data represent the number of hours that a rechargeable hedge trimmer operates before a recharge is required:*

$$1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.$$

At the 0.05 level of significance, test that this particular trimmer has between charge intervals that follow a distribution with a median of 1.8 hours duration.

Now, say we are interested in testing the population mean. Can we still use the sign test?

8.3 Paired-sample sign test

The sign test is often used when we have paired data. Suppose we have n pairs of observations: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. If the X's have mean μ_1 and the Y's have mean μ_2, we may want to test if $\mu_1 - \mu_2 = d_0$, where the d_0 is the hypothetical difference between the two means under the null.

Example 8.3.1 *(Miller and Miller's Text) A taxi company is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Sixteen cars are equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars are then equipped with regular belted tires and driven once again over the test course. The gasoline consumption (km/L) for each run is given in the following table. We want to test if the car equipped with radial tires has a better fuel economy than those equipped with regular belted tires?*

Car	1	2	3	4	5	6	7	8
Radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0
Belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8
Car	9	10	11	12	13	14	15	16
Radial	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Belted	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

8.4 The signed-rank test

In a one sample test, the sign test only utilize the plus and minus sign of the differences (greater or smaller relationship) between the observations and the hypothetical median $\tilde{\mu}_0$ or mean μ_0. Alternatively, in a paired-sample test, the plus and minus signs of the differences between the paired observations are used. However, it does not use the information contained in the magnitude of the differences.

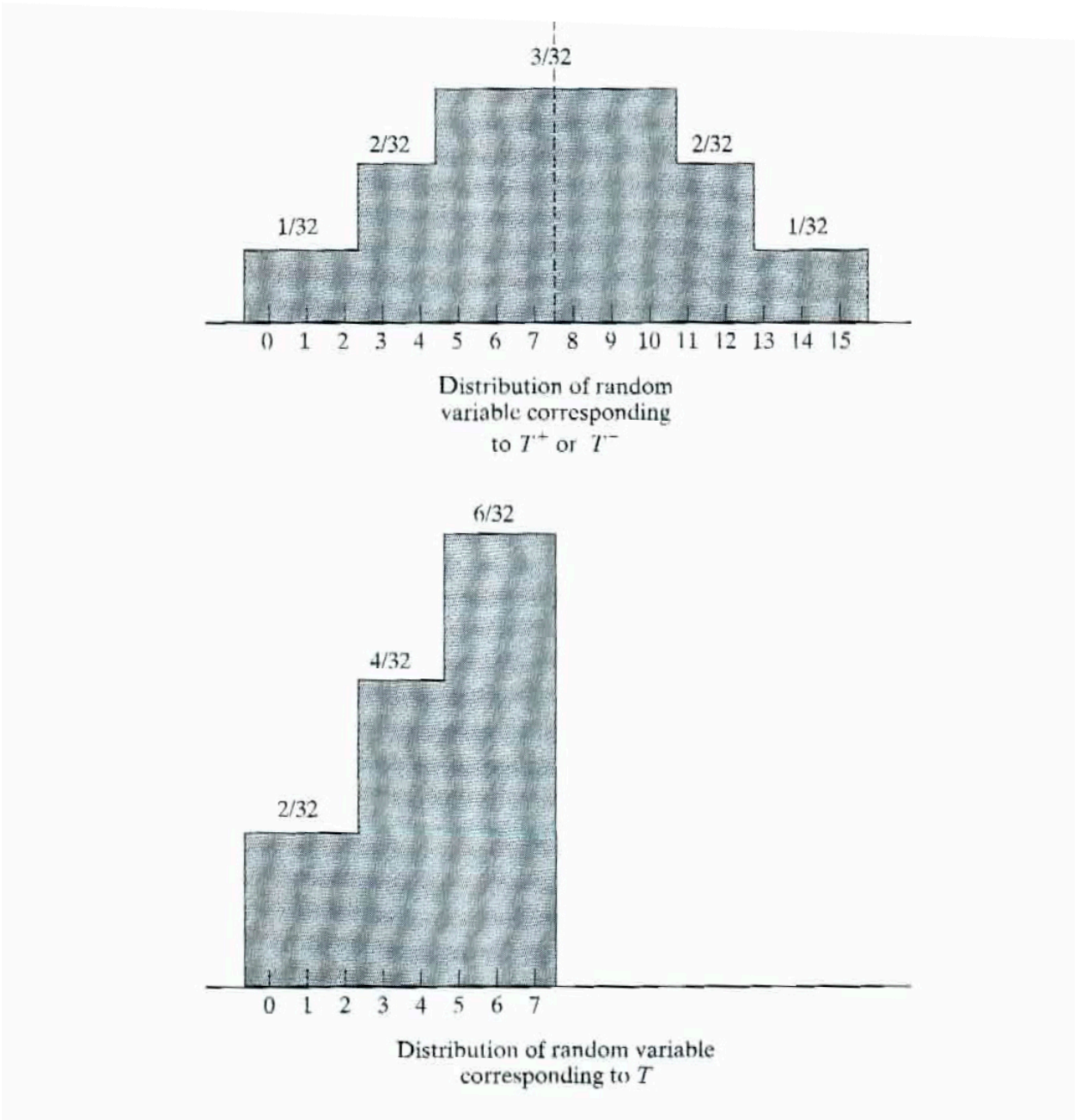
In 1945, Frank Wilcoxon proposed the Wilcoxon signed-rank test that does utilize both the information in direction and magnitude of the differences.

The Wilcoxon signed-rank test can be applied when we have a symmetric continuous distribution. Under this condition, we can test $H_0 : \mu = \mu_0$ for one sample, and $H_0 : \mu_1 = \mu_2$ for a paired-sample.

Suppose we have a random sample $X_1, \cdots, X_5$ from a population with mean μ. We assume that the population has a symmetric distribution. At the α level of significance, we want to test $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$.

The distribution of T^+ (or T^-) is summarized in this table:

T^+	possible outcomes	# of outcomes	probability
0			
1			
2			
3			
4			
5			
$\vdots$			
15			
total			



Distribution of random variable corresponding to T^+ or T^-

Distribution of random variable corresponding to T

Example 8.4.1 *(Miller and Miller's Text) The first column of the following table contains the 15 measurements of the octane rating of a certain kind of gasoline. Let us use the signed-rank test at the 0.05 level of significant to test whether the mean octane rating of the given kind of gasoline is 98.5.*

measurement	d_i	rank	D_i
97.5	-1.0	4	
95.2	-3.3	12	
97.3	-1.2	6	
96.0	-2.5	10	
96.8	-1.7	7	
100.3	1.8	8	
97.4	-1.1	5	
95.3	-3.2	11	
93.2	-5.3	14	
99.1	0.6	2	
96.1	-2.4	9	
97.6	-0.9	3	
98.2	-0.3	1	
98.5	0.0		
94.9	-3.6	13	

Theorem 8.4.2 *Under the assumptions required by the signed-rank test, T^+ is a random variable with the mean*

$$E(T^+) = \frac{n(n+1)}{4}$$

and the variance

$$var(T^+) = \frac{n(n+1)(2n+1)}{24}.$$

This result holds for for T^- as well.

Proof: The assumption required by the signed-rank test is that: 1) for the one sample test, a random sample is from a population with a symmetric distribution; and 2) for the paired-sample test, the paired differences are from a population with a symmetric distribution.

The result of Theorem 8.4.2 can be applied to the Example 8.3.1 concerning a paired-sample test with a normal approximation to the test statistic T^+.

Car	1	2	3	4	5	6	7	8
Radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0
Belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8
Car	9	10	11	12	13	14	15	16
Radial	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Belted	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

8.5 Wilcoxon rank-sum test: the U test

The sign-test and Wilcoxon's signed-rank test are two alternative procedures for one-sample tests or equivalently paired tests. Now, we consider the problem where we have two independent (i.e. no pairing) random samples from two populations with continuous distributions that are obviously not normal. The **Wilcoxon rank-sum test** is an alternative nonparametric procedure to the two-sample t-test, in which, the normality assumption of the population distribution is not required.

Suppose we have a random sample $X_1, \ldots, X_{n_1}$ of size n_1 from a population with mean μ_1 and another random sample $X_{n_1+1}, \ldots, X_{n_1+n_2}$ of size n_2 from a population with mean μ_2. We want to test $H_0 : \mu_1 = \mu_2$ against a suitable alternative. In general, $n_1 \leq n_2$. Then the Wilcoxon rank-sum test is proceeded as follows:

Sample 1	Rank	Sample 2	Rank
X_1	R_1	X_{n_1+1}	R_{n_1+1}
X_2	R_2	X_{n_1+2}	R_{n_1+2}
X_3	R_3	X_{n_1+3}	R_{n_1+3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{n_1}	R_{n_1}	X_{2n_1}	R_{2n_1}
		$\vdots$	$\vdots$
		$X_{n_1+n_2}$	$R_{n_1+n_2}$
	$W_1 = \sum_{i=1}^{n_1} R_i$		$W_2 = \sum_{i=n_1}^{n_1+n_2} R_i$

The total ranking $W_1 + W_2$ depends on the total sample size $n_1 + n_2$. In general,

Once we have determined W_1, W_2 can be determined easily by

The principle of rejection rules:

In actual practice, our rejection rule based on the following statistics: for one-sided tests:

$U_1 =$

$U_2 =$

for two-sided tests:

$U =$

Properties of U_1 and U_2:

1. $U_1 + U_2 = n_1 n_2$, and $U_1, U_2 \in [0, n_1 n_2]$.

Properties of U_1 and U_2:

2. Under the assumption required by U test, U_1 and U_2 are random variables with mean
$$\mu = \frac{n_1 n_2}{2}$$
and the variance
$$\sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Example 8.5.1 *The nicotine content of two brands of cigarettes, measured in milligrams, was found to be as follows:*

Brand A	*2.1*	*4.0*	*6.3*	*5.4*	*4.8*	*3.7*	*6.1*	*3.3*		
Brand B	*4.1*	*0.6*	*3.1*	*2.5*	*4.0*	*6.2*	*1.6*	*2.2*	*1.9*	*5.4*

Test the hypothesis, at the 0.05 level of significance, that the median nicotine contents of the two brands are equal against the alternative that they are unequal.

Brand A	Rank	Brand B	Rank
2.1		4.1	
4.0		0.6	
6.3		3.1	
5.4		2.5	
4.8		4.0	
3.7		6.2	
6.1		1.6	
3.3		2.2	
		1.9	
		5.4	
	$W_1 =$		$W_2 =$

Normal approximation for two-sample test
When both sample sizes exceed 8, i.e., $n_1 > 8$ and $n_2 > 8$, the sampling distribution of U_1 (or U_2) approaches the normal distribution with mean and variance as

$$E(U_1) = \frac{n_1 n_2}{2}, \qquad var(U_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

We can thus use the normal approximation to form a test statistic

for the U test, where the critical region is defined as:

Bibliography

[1] Miller I and Miller M, John E. Fruend's Mathematical Statistics with applications, 8th edition, Pearson 2014.