Statistics: A Probabilistic Adventure

Stats

August 30, 2023

E/Ea Thompson(they/them),
Physics and Math Honors

Solo Pursuit of Learning

Contents

Chapter 1

Probability

1.1.0 §Introduction to Probability and Inference

Definition 1.1.1. *An* **experiment** *is the process by which observation is made.*

Definition 1.1.2. *The possible outcomes of an experiment are called* **events***.*

An event which can be decomposed into other events is called a **compound event***.*

Definition 1.1.3. *A* **simple event** *is an event that cannot be decomposed. Each simple event corresponds to one and only one* **sample point***.*

In particular, a simple event is a singleton containing its sample point.

Definition 1.1.4. *The* **sample space** *associated with an experiment is the set consisting of all possible sample points.*

Definition 1.1.5. *A* **discrete sample space** *is one that contains either a finite or a countable number of distinct sample points.*

Remark 1.1.1. All distinct simple events correspond to mutually exclusive sets of simple events, and are thus mutually exclusive events.

Definition 1.1.6. *An* **event** *in a discrete sample space S is a collection of sample points–that is a subset of S.*

Definition 1.1.7. *Probability Model 1 Suppose S is a sample space associated with an experiment. To every event $A \subseteq S$, we assign a number, $P(A)$, called the* **probability of** *A, such that the following axioms hols*

Axiom 1: $0 \leqslant P(A) \leqslant 1$

Axiom 2: $P(S) = 1$

***Axiom 3**: If $A_1, A_2, A_3, ...$ form a sequence of pairwise mutually exclusive events in S (that is, $A_i \cap A_j = \varnothing$ if $i \neq j$), then*

$$P(A_1 \cup A_2 \cup A_3 \cup ...) = \sum_{i=1}^{\infty} P(A_i) \tag{1.1.1}$$

Axiom 3 also applies for a finite sequence $A_1, A_2, ..., A_n$ of pairwise mutually exclusive events:

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) \tag{1.1.2}$$

Definition 1.1.8. *Sample-point Method To find the probability of an event with the sample-point method we proceed as follows:*

1. *Define the experiment and clearly determine how to describe one simple event.*
2. *List the simple events associated with the experiment and test to make certain that it cannot be decomposed. This defines the sample space S.*
3. *Assign reasonable probabilities to the sample points in S, making certain that $0 \leqslant P(E_i) \leqslant 1$ and $\sum P(E_i) = P(S) = 1$.*
4. *Define the event of interest, A, as a specific collection of sample points.*
5. *Find $P(A)$ by summing the probabilities of the sample points in A.*

1.1.1 §Combinatorial Tools

Theorem 1.1.1. *With m elements $a_1, a_2, ..., a_m$ and n elements $b_1, b_2, ..., b_n$, it is possible to form $mn = m \times n$ pairs containing one element from each group.*

Definition 1.1.9. *An ordered arrangement of r distinct objects is called a* **permutation**. *The number of ways of ordering n distinct objects taken r at a time will be designated by the symbol P_r^n.*

Theorem 1.1.2. *For $n \geqslant r$, we have that*

$$P_r^n = \frac{n!}{(n-r)!} \tag{1.1.3}$$

Theorem 1.1.3. *The number of ways of partitioning n distinct objects into k distinct groups containing $n_1, n_2, ..., n_k$ objects, respectively, where each object appears in exactly one group and $\sum_{i=1}^{k} n_i = n$, is*

$$N = \binom{n}{n_1 n_2 ... n_k} = \frac{n!}{n_1! n_2! ... n_k!} \tag{1.1.4}$$

Definition 1.1.10. *The number of* **combinations** *of n objects taken r at a time is the number of subsets, each of size r, that can be formed from the n objects. This number will be denoted by C_r^n or $\binom{n}{r}$.*

Theorem 1.1.4. *The number of unordered subsets of size r chosen from n available objects is*

$$\binom{n}{r} = C_r^n = \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!} \tag{1.1.5}$$

1.2.0 §Conditional Probability

Definition 1.2.1. *The* **conditional probability** *of an event is the probability (relative frequency of occurence) of the event given the fact that one or more events have already occured. In particular, the conditional probability of an event A, given that an event B has occured, is equal to*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.2.1}$$

provided $P(B) > 0$. The symbol $P(A|B)$ is read "probability of A given B."

Definition 1.2.2. *Two events A and B are said to be* **independent** *if any one of the following holds:*

$$\begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B) \\ P(A \cap B) &= P(A)P(B) \end{aligned}$$

Otherwise, the events are said to be **dependent**

1.3.0 §Laws of Probability

Theorem 1.3.1. *Multiplicative Law of Probability The probability of the intersection of two events A and B is*

$$P(A \cap B) = \begin{cases} P(A)P(B|A) \\ P(B)P(A|B) \end{cases} \tag{1.3.1}$$

If A and B are independent, then

$$P(A \cap B) = P(A)P(B) \tag{1.3.2}$$

Theorem 1.3.2. *Additive Law of Probability The probability of the union of two events A and B is*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.3.3}$$

If A and B are mutually exclusive events, $P(A \cap B) = 0$ and

$$P(A \cup B) = P(A) + P(B) \tag{1.3.4}$$

Theorem 1.3.3. *If A is an event, then*

$$P(A) = 1 - P(A^C) \tag{1.3.5}$$

1.4.0 §The Event-Comparison Method

Definition 1.4.1. *Event-Comparison Method The event-comparison method is executed as follows:*

1. *Define the experiment.*
2. *Visualize the nature of the sample points. Identify a few to clarify your thinking.*
3. *Write an equation expressing the event of interest, say A, as a composition of two or more events, using unions, intersections, and/or complements. Make certain that event A and the event implied by the composition represent the same set of sample points.*
4. *Apply the additive and multiplicative laws of probability to the compositions obtained in step 3 to find $P(A)$.*

1.5.0 §Bayes Theorem and The Law of Total Probability

Definition 1.5.1. *For some positive integer k, let the sets $B_1, B_2, ..., B_k$ be such that*

1. $S = B_1 \cup ... \cup B_k$
2. $B_i \cap B_j = \emptyset$ *for* $i \neq j$

Then the collection of sets $\{B_1, ..., B_k\}$ is said to be a **partition** *of S.*

Remark 1.5.1. If $A \subseteq S$ and $\{B_1, ..., B_k\}$ is a partition of S, then A can be decomposed as

$$A = \bigcup_{i=1}^{k} (A \cap B_i)$$

Theorem 1.5.1. *Assume that $\{B_1, ..., B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i \in \{1, ..., k\}$. Then for any event A:*

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Proof. Note any subset A of S can be written as

$$A = A \cap S = \bigcup_{i=1}^{k} (A \cap B_i)$$

Note that because $\{B_1, ..., B_k\}$ is a partition of S, for $i \neq j$

$$(A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset$$

so these events are mutually exclusive. Thus by axiom 3 of probabilities:

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= \sum_{i=1}^{l} P(A|B_i)P(B_i) \end{aligned}$$

as desired. ■

Theorem 1.

Bayes' Rule Assume that $\{B_1, ..., B_k\}$ is a partition of S such that $P(B_i) > 0$ for each $i \in \{1, ..., k\}$. Then

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} \tag{1.5.1}$$

1.6.0 §Numerical Events and Random Sampling

Definition 1.6.1. *A* **random variable** *is a real-valued function for which the domain is the sample space.*

Remark 1.6.1. If y denotes an observed value of the random variable Y, then $P(Y = y)$ is the sum of the probabilities of the sample points E_i for which $Y(E_i) = y$.

Definition 1.6.2. *Let N and n represent the numbers of elements in the population and sample, respectively. If the sampling is conducted in such a way that each of the $\binom{N}{n}$ samples has an equal probability of being selected, the sampling is said to be random, and the result is said to be a* **random sample**.

Chapter 2

Discrete Random Variables

2.1.0 §Basic Definitions: DRV

Recall 2.1.1. Recall that a ***random variable*** is a real-valued function defined over the sample space of the experiment. A random variable can be used to identitfy numerical events that are of interest in an experiment.

Definition 2.1.1. *A random variable Y is said to be* **discrete** *if it can assume only a finite or countably infinite number of distinct values.*

Remark 2.1.2. The collection of probabilities for each value of a random variable is called the ***probability distribution*** of the discrete random variable.

2.2.0 §Probability Distribution Definition for a DRV

Notation 2.2.1. For a random variable Y and a specified observed value y, the expression $(Y = y)$ denotes *the set of all points in S assigned the value of y by the random variable Y.*

Definition 2.2.1. *The probability that Y takes on the value y, $P(Y = y)$, is defined as the* sum of the probabilites of all sample points in S *that are assigned the value of y. We sometimes denote $P(Y = y)$ by $p(y)$, and call P the* probability function *for Y.*

Definition 2.2.2. *The* **probability distribution** *for a discrete variable Y can be represented by a formula, a table, or a graph that provides $p(y) = P(Y = y)$ for all y.*

Theorem 2.2.1. *For any discrete probability distribution, the following must be true:*

1. $0 \leqslant p(y) \leqslant 1$ *for all y.*

2. $\sum_y p(y) = 1$*, where the summation is over all values of y with non-zero probability.*

2.3.0 §Functions of a Random Variable

Definition 2.3.1. *Let Y be a discrete random variable with probability function $p(y)$. Then the* expected value *of Y, $E(Y)$, is defined to be*

$$E(Y) = \sum_y yp(y)$$

This provides the mean of the population with distribution given by $p(y)$. This expected value if the sum is abolutely convergent:

$$\sum_y |y|p(y) < \infty$$

Remark 2.3.1. If $p(y)$ is an accurate characterization of the population frequency distribution, then $E(Y) = \mu$, the population mean.

Theorem 2.3.1. *Let Y be a discrete random variable with probability function $p(y)$ and let $g(Y)$ be a real-valued function of Y. THen the expected value of $g(Y)$ is given by*

$$E[g(Y)] = \sum_{\text{all } y} g(y)p(y)$$

Proof. We prove the case for Y's codomain being finite; $y_1, y_2, ..., y_n$. Because the function $g(y)$ may not be one-to-one, suppose $g(Y)$ takes on values $g_1, g_2, ..., g_m$, where $m \leqslant n$. It follows that $g(Y)$ is a random variable cuh that for $i = 1, 2, ..., m$,

$$P[g(Y) = g_i] = \sum_{\substack{\text{all } y_j \text{ such that} \\ g(y_j) = g_i}} p(y_j) = p^*(g_i)$$

Thus by definition of the expected value of a random variable,

$$\begin{aligned}
E[g(Y)] &= \sum_{i=1}^{m} g_i p^*(g_i) \\
&= \sum_{i=1}^{m} g_i \left\{ \sum_{\substack{\text{all } y_j \text{ such that} \\ g(y_j) = g_i}} p(y_j) \right\} \\
&= \sum_{i=1}^{m} \sum_{\substack{\text{all } y_j \text{ such that} \\ g(y_j) = g_i}} g_i p(y_j) \\
&= \sum_{j=1}^{n} g(y_j)p(y_j)
\end{aligned}$$

■

Definition 2.3.2. *If Y is a random variable with mean $E(Y) = \mu$, the* **varianve** *of a random variable Y is defined to be the expected value of $(Y - \mu)^2$. That is,*

$$V(Y) = E[(Y - \mu)^2]$$

The **standard deviation** *of Y is the positive square root of $V(Y)$.*

Remark 2.3.2. If $p(y)$ is an accurate characterization of the population frequency distribution, then $E(Y) = \mu$, $V(Y) = \sigma^2$, the ***population variance***, and σ is the ***population standard deviation***.

Theorem 2.3.2. *Let Y be a discrete random variable with probability function $p(y)$ and let c be a constant. Then $E(c) = c$.*

Proof. Consider the function $g(Y) \equiv c$. By our previous theorem

$$E(c) = \sum_y cp(y) = c \sum_y p(y)$$

But $\sum_y p(y) = 1$, and hence $E(c) = c$. ■

Theorem 2.3.3. *Let Y be a discrete random variable with probability function $p(y)$, let $g(Y)$ be a function of Y, and let c be a constant. Then*

$$E[cg(Y)] = cE[g(Y)]$$

Proof. By our preceding theorems,

$$E[cg(Y)] = \sum_y cg(y)p(y) = c \sum_y g(y)p(y) = cE[g(Y)]$$

■

Theorem 2.3.4. *Let Y be a discrete random variable with probability function $p(y)$ and $g_1(Y), g_2(Y), ..., g_k(Y)$ be k functions of Y. Then*

$$E[g_1(Y) + g_2(Y) + ... + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + ... + E[g_k(Y)]$$

Proof. Observe that

$$\begin{aligned} E\left[\sum_{i=1}^k g_i(Y)\right] &= \sum_y \left[\sum_{i=1}^k g_i(y)\right] p(y) \\ &= \sum_y \sum_{i=1}^k g_i(y)p(y) \\ &= \sum_{i=1}^k \sum_y g_i(y)p(y) \\ &= \sum_{i=1}^k E[g_i(Y)] \end{aligned}$$

■

Theorem 2.3.5. *Let Y be a discrete random variable with probability function $p(y)$ and mean $E(Y) = \mu$; then*

$$V(Y) = \sigma^2 = E[(Y-\mu)^2] = E(Y^2) - \mu^2$$

Proof. Observe that

$$\begin{aligned} \sigma^2 &= E[(Y-\mu)^2] \\ &= E(Y^2 - 2\mu Y + \mu^2) \\ &- E(Y^2) - E(2\mu Y) + E(\mu^2) \end{aligned}$$

Noting that μ is constant, we find that

$$\sigma^2 = E(Y^2) - 2\mu E(Y) + \mu^2$$

But, $\mu = E(Y)$, so

$$\sigma^2 = E(Y^2) - 2\mu^2 + \mu^2 = E(Y^2) - \mu^2$$

■

2.4.0 §The Binomial Probability Distribution

Definition 2.4.1. *A* **binomial experiment** *possesses the following properties:*

1. *The experiment consists of a fixed number, n, of identical trials.*
2. *Each trial results in one of two outcomes: success, S, or failure, F.*
3. *The probability of success on a single trial is equal to some value p and and remains the same from trial to trial. The probability of failure is equal to $q = (1-p)$.*
4. *The trials are independent.*
5. *The random variable of interest is Y, the number of successes observed during the n trials.*

Definition 2.4.2. *A random variable Y is said to have a* **binomial distribution** *based on n trials with success probability p if and only if*

$$p(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, 2, ..., n \text{ and } 0 \leqslant p \leqslant 1$$

where $q = 1 - p$

Corollary 2.4.1. *If a random variable Y has a binomial distribution based on n trials with success p, then*

$$\sum_y p(y) = \sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y} = (q+p)^n = 1^n = 1$$

Theorem 2.4.2. *Let Y be a binomial random variable based on n trials and success probability p. Then*

$$\mu = E(Y) = np \quad \text{and} \quad \sigma^2 = V(Y) = npq$$

Proof. By definition we have that

$$E(Y) = \sum_y yp(y) = \sum_{y=0}^{n} y\binom{n}{y}p^y q^{n-y}$$

Notice for $y = 0$ we have zero, so we can write

$$\begin{aligned} E(Y) &= \sum_{y=1}^{n} y\frac{n!}{(n-y)!y!}p^y q^{n-y} \\ &= \sum_{y=1}^{n} \frac{n!}{(n-y)!(y-1)!}p^y q^{n-y} \end{aligned}$$

If we factor out np from each term and let $z = y - 1$, we can write

$$\begin{aligned} E(Y) &= np\sum_{y=1}^{n} \frac{(n-1)!}{(n-y)!(y-1)!}p^{y-1} q^{n-y} \\ &= np\sum_{z=0}^{n-1} \frac{(n-1)!}{(n-1-z)!z!}p^{z} q^{n-1-z} \\ &= np\sum_{z=0}^{n-1} \binom{n-1}{z}p^{z} q^{n-1-z} \\ &= np(p+q)^{n-1} \\ &= np \cdot 1^{n-1} \\ &= np \end{aligned}$$

as desired.

Next, we know that $V(Y) = E(Y^2) - E(Y)^2$. Thus, let us first calculate $E(Y^2)$

$$E(Y^2) = \sum_{y=0}^{n} y^2 p(y) = \sum_{y=0}^{n} y^2\binom{n}{y}p^y q^{n-y} = \sum_{y=0}^{n} y^2\frac{n!}{y!(n-y)!}p^y q^{n-y}$$

Next, notice that

$$E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y)$$

and, therefore,

$$E(Y^2) = E[Y(Y-1)] + E(Y)$$

Now, observe that

$$\begin{aligned} E[Y(Y-1)] &= \sum_{y=0}^{n} y(y-1)\frac{n!}{y!(n-y)!}p^y q^{n-y} \\ &= \sum_{y=2}^{n} \frac{n!}{(y-2)!(n-y)!}p^y q^{n-y} \end{aligned}$$

$$\begin{aligned}
&= n(n-1)p^2 \sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!(n-y)!} p^{y-2} q^{n-y} \\
&= n(n-1)p^2 \sum_{z=0}^{n-2} \frac{(n-2)!}{z!(n-2-z)!} p^{z} q^{n-2-z} \\
&= n(n-1)p^2 \sum_{z=0}^{n-2} \binom{n-2}{z} p^{z} q^{n-2-z} \\
&= n(n-1)p^2 (p+q)^{n-2} \\
&= n(n-1)p^2
\end{aligned}$$

Thus, we have that

$$E(Y^2) = E[Y(Y-1)] + E(Y) = n(n-1)p^2 + np$$

and finally

$$V(Y) = E(Y^2) - E(Y)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p) = npq$$

■

2.5.0 §The Geometric Probability Distribution

Definition 2.5.1. *A* **geometric experiment** *possesses the following properties:*

1. *The experiment consists of a fixed number, n, of identical trials.*
2. *Each trial results in one of two outcomes: success, S, or failure, F.*
3. *The probability of success on a single trial is equal to some value p and and remains the same from trial to trial. The probability of failure is equal to $q = (1-p)$.*
4. *The trials are independent.*
5. *The random variable of interest is Y, is the number of the trial on which the first success occurs.*

Definition 2.5.2. *A random variable Y is said to have a* **geometric probability distribution** *if and only if*

$$p(y) = q^{y-1}p, \quad y = 1, 2, 3, ..., 0 \leqslant p \leqslant 1$$

Theorem 2.5.1. *If Y is a random variable with a geometric distribution,*

$$E(Y) = \frac{1}{p} \quad \text{and} \quad V(Y) = \frac{1-p}{p^2}$$

Proof. First note that,

$$E(Y) = \sum_{y=1}^{\infty} yq^{y-1}p = p\sum_{y=1}^{\infty} yq^{y-1}$$

Then observe that

$$\frac{d}{dq}\left(\sum_{y=1}^{\infty} q^y\right) = \sum_{y=1}^{\infty} yq^{y-1}$$

But this is a geometric series with well known summation

$$\sum_{y=1}^{\infty} q^y = \frac{q}{1-q}$$

Hence, we have that

$$E(Y) = p\frac{d}{dq}\left(\frac{q}{1-q}\right) = p\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

Next, for the variance we have

$$V(Y) = E(Y^2) - E(Y)^2 = \sum_{y=1}^{\infty} y^2 q^{y-1} p - \frac{1}{p^2}$$

First, observe that

$$\frac{d}{dq}\left(\sum_{y=1}^{\infty} yq^y\right) = \sum_{y=1}^{\infty} y^2 q^{y-1}$$

Then, that

$$\sum_{y=1}^{\infty} yq^{y-1} = \frac{1}{(1-q)^2}$$

Then it follows that

$$\sum_{y=1}^{\infty} yq^y = q\sum_{y=1}^{\infty} yq^{y-1} = \frac{q}{(1-q)^2}$$

Then it follows that

$$\begin{aligned}
V(Y) &= p\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right) - \frac{1}{p^2} \\
&= p\frac{1+q}{(1-q)^3} - \frac{1}{p^2} \\
&= p\frac{2-p}{p^3} - \frac{1}{p^2} \\
&= \frac{2-p-1}{p^2} \\
&= \frac{1-p}{p^2} \\
&= \frac{q}{p^2}
\end{aligned}$$

as desired. ■

2.6.0 §Negative Binomial Distribution

Definition 2.6.1. *A random variable Y is said to have a* **negative binomial probability distribution** *if and only if*

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r}, \quad y = r, r+1, r+2, ..., 0 \leqslant p \leqslant 1$$

Theorem 2.6.1. *If Y is a random variable with a negative binomial distribution,*

$$\mu = E[Y] = \frac{r}{p} \text{ and } \sigma^2 = V[Y] = \frac{r(1-p)}{p^2}$$

2.7.0 §The Hypergeometric Probability Distribution

Definition 2.7.1. *A random variable Y is said to have a* **hypergeometric probability distribution** *if and only if*

$$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

where y is an integer $0, 1, 2, ..., n$, subject to the restriction $y \leqslant r$ and $n - y \leqslant N - r$.

Claim 2.7.1. *I claim that given $r \in \mathbb{Z}^+$, we have*

$$\sum_{i=0}^{n} \binom{r}{i}\binom{N-r}{n-i} = \binom{N}{n}$$

Proof. (To be completed) ■

Theorem 2.7.2. *If Y is a random variable with hypergeometric distribution, the expected value and variance of Y is*

$$\mu = E[Y] = \frac{nr}{N} \text{ and } \sigma^2 = V[Y] = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

2.8.0 §The Poisson Probability Distribution

Construction 2.8.1. *Suppose we have a time in which events of interest occur, and suppose we split up this time into individual intervals where at most one event can occur. Then, if the occurrence of events can be regarded as independent from interval to interval, with equal probability, the number of total event has a binomial distribution.*

Let n be the number of interval divisions and p the probability in a given interval. Note that as n increases p will decrease. Suppose $\lambda = np$ is a fixed constant. Then taking the limit of the *binomial probability* $p(y) = \binom{n}{y} p^y (1-p)^{n-y}$ *as* $n \to \infty$, *we have:*

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{y} p^y (1-p)^{n-y} &= \lim_{n\to\infty} \frac{n(n-1)\dots(n-y+1)}{y!} \left(\frac{\lambda}{n}\right)^2 \left(1-\frac{\lambda}{n}\right)^{n-y} \\
&= \lim_{n\to\infty} \frac{\lambda^y}{y!} \left(1-\frac{\lambda}{n}\right)^{n-1} \frac{n(n-1)\dots(n-y+1)}{n^y} \\
&= \frac{\lambda^y}{y!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-y} \left(1-\frac{1}{n}\right) \\
&\times \left(1-\frac{2}{n}\right) \times \dots \times \left(1-\frac{y-1}{n}\right) \\
&= \frac{\lambda^y}{y!} e^{-\lambda}
\end{aligned}$$

noting that

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

and that all other terms to the right of the limit have a limit of 1. *A random variable possessing this distribution is said to have a Poisson distribution.*

Due to the limiting factor of the Poisson distribution from the binomial distribution, under specific curcumstances one can use the Poisson probabilities to approximate their binomial counterparts.

Definition 2.8.2. *A random variable* Y *is said to have a* **Poisson probability distribution** *if and only if*

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = -, 1, 2, ..., \ \lambda > 0.$$

Theorem 2.8.1. *If* Y *is a random variable possessing a Poisson distribution with parameter* λ, *then*

$$\mu = E[Y] = \lambda \quad \text{and} \quad \sigma^2 = V[Y] = \lambda$$

Proof. By definition,

$$E[Y] = \sum_y y p(y) = \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!}$$

As the first term in this sum is 0, we can write

$$E[Y] = \sum_{y=1}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y-1)!}$$

Factoring out a term of λ and substituting the variable $z = y - 1$, we obtain the sum

$$E[Y] = \lambda \sum_{z=0}^{\infty} \frac{\lambda^z e^{-\lambda}}{z!}$$

where the sum is again the sum of a Poisson distribution over all z, and is hence equal to 1. Thus, $E[Y] = \lambda$.

For the variance, first observe that

$$\begin{aligned} E[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1)\frac{\lambda^y e^{-\lambda}}{y!} \\ &= \lambda^2 \sum_{y=2}^{\infty} \frac{\lambda^{y-2} e^{-\lambda}}{(y-2)!} \\ &= \lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \lambda^2 \end{aligned}$$

Then the variance is found as

$$\begin{aligned} VAR[Y] &= E[Y(Y-1)] + E[Y] - E[Y]^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda \end{aligned}$$

as desired. ■

2.9.0 §Moments and Moment-Generating Functions

In this section we consider a set of numerical descriptive measures that under certain conditions uniquely determine the probability distribution of a random variable.

Definition 2.9.1. *The kth* **moment of a random variable Y taken about the origin** *is defined to be $E[Y^k]$ and is denoted by μ_k'.*

Definition 2.9.2. *The kth* **moment of a random variable Y taken about its mean**, *or the kth* **central moment of Y**, *is defined to be $E[(Y-\mu)^k]$ and is denoted by μ_k.*

Definition 2.9.3. *The* **moment-generating function $m(t)$ for a random variable Y** *is defined to be $m(t) = E[e^{tY}]$. We say that a moment-generating function for Y exists if there exists a positive constant b such that $m(t)$ is finite for $|t| \leqslant b$.*

From a series expansion for e^{ty}, we have

$$e^{ty} = 1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \frac{(ty)^4}{4!} + \dots .$$

Assuming that μ_k' is finite for $k = 1, 2, 3, ...$, we have

$$E[e^{tY}] = \sum_y e^{ty} p(y) = \sum_y \left[1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \dots\right] p(y)$$

$$
\begin{aligned}
&= \sum_y p(y) + t\sum_y yp(y) + \frac{t^2}{2!}\sum_y y^2 p(y) + \frac{t^3}{3!}\sum_y y^3 p(y) + \dots \\
&= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \dots
\end{aligned}
$$

This interchange of summations is possible if the series is convergent, that is $m(t)$ exists. Thus, $E[e^{tY}]$ is a function of all the moments μ_k' about the origin.

Theorem 2.9.1. *If $m(t)$ exists, then for any positive integer k,*

$$\left.\frac{d^k m(t)}{dt^k}\right]_{t=0} = m^{(k)}(0) = \mu_k'$$

In other words, find the kth derivative of $m(t)$ with respect to t and then set $t = 0$ to obtain μ_k'.

Remark 2.9.1. It can be seen that the moment-generating function for a Poisson random variable is

$$m(t) = \sum_{y=0}^{\infty} e^{ty}\frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}$$

The primary application of a moment-generating function is to prove that a random variable possesses a particular probability distribution $p(y)$. If $m(t)$ exists for a probability distribution $p(y)$, it is unique. Also, if the moment-generating functions for two random variables Y and Z are equal (for all $|t| < b$ for some $b > 0$), then Y and Z must have the same probability distribution. It follows that, if we can recognize the moment-generating function of a random variable Y to be one associated with a specific distribution, then Y must have that distribution.

2.10.0 §Tchebysheff's Theorem (Discrete)

Theorem 2.

Tchebysheff's Theorem (Discrete)tch Let Y be a random variable with mean μ and finite variance σ^2. Then, for any constant $k > 0$,

$$P(|Y-\mu| < k\sigma) \geqslant 1 - \frac{1}{k^2} \quad or \quad P(|Y-\mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}$$

Chapter 3

Continuous Random Variables

A random variable is called ***continuous*** if it can take on any value in an interval (i.e. it can take on an uncountable number of different values).

3.1.0 §The Probability Distribution for a Continuous Random Variable

Definition 3.1.1. *Let Y denote any random variable. The* **(cumulative) distribution function** *of Y, denoted by $F(y)$, is such that $F(y) = P(Y \leqslant y)$ for $-\infty < y < \infty$.*

Theorem 3.1.1. *Properties of a Distribution Function If $F(y)$ is a distribution function, then*

1. $F(-\infty) \equiv \lim_{y \to -\infty} F(y) = 0$
2. $F(\infty) \equiv \lim_{y \to \infty} F(y) = 1$
3. *$F(y)$ is a nondecreasing function of y. If $y_1 < y_2$ are any two values, then $F(y_1) \leqslant F(y_2)$.*

Definition 3.1.2. *A random variable Y with distribution function $F(y)$ is said to be* **continuous** *if $F(y)$ is continuous, for $-\infty < y < \infty$. We also require that the first derivative of $F(y)$ exist and be continuous except for, at most, a finite numbr of points in any finite interval.*

It is important to note that if Y is a continuous random variable, then for any real number y,

$$P(Y = y) = 0$$

If this were not true and $P(Y = y_0) = p_0 > 0$, then $F(y)$ would have a discontinuity (jump) of size p_0 at the point y_0, violating the assumption that Y was continuous.

Definition 3.1.3. *Let $F(y)$ be the distribution function for a continuous random variable Y.*

Then $f(y)$, given by

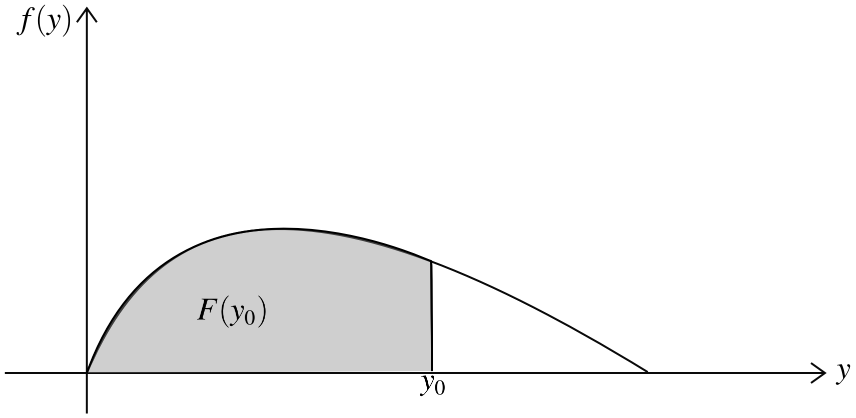
$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

wherever the derivative exists, is called the **probability density function** *for the random variable Y.*

It follows from these previous definitions and the Fundamental Theorem of Calculus that $F(y)$ can be written as

$$F(y) = \int_{-\infty}^{y} f(t)dt$$

The probability density function is a ***theoretical model*** for the frequencey distribution (histogram) of a population of measurements. The relationship between the probability density function and distribution function is given by:



Theorem 3.1.2. *Properties of a Density Function If $f(y)$ is a density function for a continuous random variable, then*

1. $f(y) \geqslant 0$ *for all* y, $-\infty < y < \infty$.
2. $\int_{-\infty}^{\infty} f(y)dy = 1$.

Definition 3.1.4. *Let Y denote any random variable. If $0 < p < 1$, the pth* **quantile** *of Y, denoted by ϕ_p, is the smallest value such that $P(Y \leqslant \phi_p) = F(\phi_p) \geqslant p$. If Y is continuous, ϕ_p is the smallest value such that $F(\phi_p) = P(Y \leqslant \phi_p) = p$.*

A special case is $p = 1/2$, and $\phi_{0.5}$ is the ***median*** of the random variable Y.

Theorem 3.1.3. *If $a < b$, we have that*

$$P(a < Y \leqslant b) = P(Y \leqslant b) - P(Y \leqslant a) = F(b) - F(a) = \int_{a}^{b} f(y)dy$$

But, because $P(Y = a) = 0$, we have the following result: if the random variable Y has density *function $f(y)$ and $a < b$, then the probability that Y falls in the interval $[a,b]$ is*

$$P(a \leqslant Y \leqslant b) = \int_a^b f(y)dy$$

3.2.0 §Expected Values for Continuous Random Variables

Definition 3.2.1. *The expected value of a continuous random variable Y is:*

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy,$$

provided that the integral exists. In particular, we say $E[Y]$ exists if

$$\int_{-\infty}^{\infty} |y|f(y)dy < \infty$$

Theorem 3.2.1. *Let $g(Y)$ be a function of Y; then the expected value of $g(Y)$ is given by:*

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dt,$$

provided that the integral exists.

Theorem 3.2.2. *Let $c \in \mathbb{R}$ be a constant and let $g(Y), g_1(Y), g_2(Y), ..., g_k(Y)$ be functions of a continuous random variable Y. Then the following results hold:*

1. $E[c] = c$
2. $E[cg(Y)] = cE[g(Y)]$
3. $E\left[\sum\limits_{i=1}^{k} g_i(Y)\right] = \sum\limits_{i=1}^{k} E[g_i(Y)]$

3.3.0 §The Uniform Probability Distribution

Intuitively a random variable has a uniform distribution if in some interval the probability is constant, and everywhere else it is zero.

Definition 3.3.1. *If $\theta_1 < \theta_2$, a random variable Y is said to have a continuous* **uniform probability distribution** *on the interval (θ_1, θ_2) if and only if the density function of Y is*

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \leqslant y \leqslant \theta_2 \\ 0, & \textit{elsewhere} \end{cases}$$

Definition 3.3.2. *The constants that determine the specific form of a density function are called* **parameters** *of the density function.*

Theorem 3.3.1. *If $\theta_1 < \theta_2$ and Y is a random variable uniformly distributed on the interval (θ_1, θ_2), then*

$$\mu = E[Y] = \frac{\theta_1 + \theta_2}{2} \quad \text{and} \quad \sigma^2 = VAR[Y] = \frac{(\theta_2 - \theta_1)^2}{12}$$

Proof. By definition:

$$\begin{aligned}
E[Y] &= \int_{-\infty}^{\infty} y f(y) dy \\
&= \int_{\theta_1}^{\theta_2} y \left(\frac{1}{\theta_2 - \theta_1} \right) dy \\
&= \left(\frac{1}{\theta_2 - \theta_1} \right) \frac{y^2}{2} \Big|_{\theta_1}^{\theta_2} \\
&= \frac{\theta_2^2 - \theta_1^2}{2(\theta_2 - \theta_1)} \\
&= \frac{\theta_2 + \theta_1}{2}
\end{aligned}$$

and

$$\begin{aligned}
E[Y^2] &= \int_{-\infty}^{\infty} y^2 f(y) dy \\
&= \int_{\theta_1}^{\theta_2} y^2 \left(\frac{1}{\theta_2 - \theta_1} \right) dy \\
&= \left(\frac{1}{\theta_2 - \theta_1} \right) \frac{y^3}{3} \Big|_{\theta_1}^{\theta_2} \\
&= \frac{\theta_2^3 - \theta_1^3}{3(\theta_2 - \theta_1)} \\
&= \frac{\theta_2^2 + \theta_2\theta_1 + \theta_1^2}{3}
\end{aligned}$$

So we find that

$$\begin{aligned}
VAR[Y] &= E[Y^2] - E[Y]^2 \\
&= \frac{\theta_2^2 + \theta_2\theta_1 + \theta_1^2}{3} - \frac{\theta_2^2 + 2\theta_2\theta_1 + \theta_1^2}{4} \\
&= \frac{\theta_2^2 - 2\theta_2\theta_1 + \theta_1^2}{12} \\
&= \frac{(\theta_2 - \theta_1)^2}{12}
\end{aligned}$$

■

3.4.0 §The Normal Probability Distribution

Definition 3.4.1. *A random variable Y is said to have a* **normal probability distribution** *if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is*

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < y < \infty$$

Theorem 3.4.1. *If Y is a normally distributed random variable with parameters μ and σ, then*

$$E[Y] = \mu \quad \textit{and} \quad VAR[Y] = \sigma^2$$

We note that evaluation of areas under the density function corresponding to $P(a \leqslant Y \leqslant b)$ require evaluating the integral

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$$

which does not have a closed form expression. Thus, numerical integration techniques must be used to determine probabilities.

Definition 3.4.2. *We can transform a normal random variable Y to a standard normal random variable Z by using the relationship:*

$$Z := \frac{Y-\mu}{\sigma}$$

3.5.0 §The Gamma Probability Distribution

A gamma distribution is used when we wish to model continuous random variables which are nonnegative and right-skewed.

Definition 3.5.1. *A random variable Y is said to have a* **gamma distribution** *with parameters $\alpha > 0$ and $\beta > 0$ if and only if the density function of Y is*

$$f(y) = \begin{cases} \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}, & 0 \leqslant y < \infty \\ 0, & \textit{elsewhere} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

is the **gamma function**.

Properties 3.5.2. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ *for all α with $\Re(\alpha) > 0$, and $\Gamma(n) = (n-1)!$ for any positive integer n.*

If α is not an integer, there is no closed form expression for

$$\int_c^d \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}dy$$

for $0 < c < d < \infty$.

Theorem 3.5.1. *If Y has a gamma distribution with parameters shape α and scale β, then*

$$\mu = E[Y] = \alpha\beta \ \text{ and } \ \sigma^2 = VAR[Y] = \alpha\beta^2$$

Proof. By definition we have that

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy = \int_0^\infty y\left(\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right)dy$$

By definition, the gamma density function is such that

$$\int_0^\infty y\left(\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right)dy = 1$$

so we have

$$\int_0^\infty y^{\alpha-1}e^{-y/\beta}dy = \beta^\alpha\Gamma(\alpha)$$

It follows that

$$\begin{aligned}
E[Y] &= \int_0^\infty y\left(\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right)dy \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty y^\alpha e^{-y/\beta}dy \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)}\left[\beta^{\alpha+1}\Gamma(\alpha+1)\right] \\
&= \frac{\beta\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \alpha\beta
\end{aligned}$$

Then, observe that

$$\begin{aligned}
E[Y^2] &= \int_0^\infty y^2\left(\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right)dy \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty y^{\alpha+1}e^{-y/\beta}dy \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)}\left[\beta^{\alpha+2}\Gamma(\alpha+2)\right] \\
&= \frac{\beta^2\alpha(\alpha+1)\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \alpha(\alpha+1)\beta^2
\end{aligned}$$

Then it follows that

$$VAR[Y] = E[Y^2] - E[Y]^2 = \alpha\beta^2$$

■

3.5.1 The Chi-Square Probability Distribution

Definition 3.5.3. *Let ν be a positive integer. A random variable Y is said to have a* **chi-square distribution with ν degrees of freedom** *if and only if Y is a gamma distributed random varaible with parameters $\alpha = \nu/2$ and $\beta = 2$.*

Theorem 3.5.2. *If Y is a chi-square random varaible with ν degrees of freedom, then*

$$\mu = E[Y] = \nu \quad \text{and} \quad \sigma^2 = VAR[Y] = 2\nu$$

3.5.2 The Exponential Probability Distribution

Definition 3.5.4. *A random variable Y is said to have an* **exponential distibution with parameter $\beta > 0$** *if and only if the density function of Y is*

$$f(y) = \begin{cases} \frac{e^{-y/\beta}}{\beta}, & 0 \leqslant y < \infty \\ 0, & elsewhere \end{cases}$$

Theorem 3.5.3. *If Y is an exponential random variable with parameter β, then*

$$\mu = E[Y] = \beta \;\; \text{and} \;\; \sigma^2 = VAR[Y] = \beta^2$$

Properties 3.5.5. *Let $a, b > 0$. Then observe that for an exponential random variable Y,*

$$\begin{aligned} P(Y > a + b | Y > a) &= \frac{P(Y > a + b)}{P(Y > a)} \\ &= \frac{\int_{a+b}^{\infty} \frac{e^{-y/\beta}}{\beta} dy}{\int_{a}^{\infty} \frac{e^{-y/\beta}}{\beta} dy} \\ &= \frac{e^{-(a+b)/\beta}}{e^{-a/\beta}} \\ &= e^{-b/\beta} \\ &= P(Y > b) \end{aligned}$$

This property of the exponential distribution is called the **memoryless property** *of the distribution.*

3.6.0 §The Beta Probability Distribution

The beta density function is a two-parameter density function defined over the closed interval $0 \leqslant y \leqslant 1$: it is often used for modeling proportions.

Definition 3.6.1. *A random variable Y is said to have a* **beta probability distribution with shape parameters $\alpha > 0$ and $\beta > 0$** *if and only if the density function of Y is*

$$f(y) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, & 0 \leqslant y \leqslant 1 \\ 0, & elsewhere, \end{cases}$$

where

$$B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The beta density function can apply to a random variable defined on the interval $c \leqslant y \leqslant d$ by defining a new variable $y^* = (y-c)/(d-c)$ so that $0 \leqslant y^* \leqslant 1$.

Definition 3.6.2. *The cdf for a beta random varaible is called the* **incomplete beta function** *and is denoted by:*

$$F(y) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}dt = I_y(\alpha,\beta)$$

If α and β are both integers, then

$$F(y) = \sum_{i=\alpha}^{n} \binom{n}{i} y^i(1-y)^{n-i}$$

for $n = \alpha+\beta-1$.

Theorem 3.6.1. *If Y is a beta-distributed random variable with parameters $\alpha > 0$ and $\beta > 0$, then*

$$\mu = E[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \sigma^2 = VAR[Y] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Proof. By definition we have that

$$\begin{aligned}
E[Y] &= \int_{-\infty}^{\infty} yf(y)dy \\
&= \int_0^1 y\left[\frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\right]dy \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}dy \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

Similarly,

$$\begin{aligned}
E[Y^2] &= \int_{-\infty}^{\infty} y^2 f(y)dy \\
&= \int_0^1 y^2\left[\frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\right]dy
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha+1}(1-y)^{\beta-1}dy \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\alpha(\alpha+1)\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)(\alpha+\beta+1)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}
\end{aligned}
$$

so

$$
\begin{aligned}
VAR[Y] &= E[Y^2] - E[Y]^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta-(\alpha^3+\alpha^2\beta+\alpha^2)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

as claimed. ■

3.7.0 §Tchebysheff's Theorem (Continuous)

Theorem 3.

Tchebysheff's Theorem (Continuous)tchcont Let Y be a random variable with mean μ and finite variance σ^2. Then, for any constant $k > 0$,

$$
P(|Y-\mu| < k\sigma) \geqslant 1 - \frac{1}{k^2} \quad or \quad P(|Y-\mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}
$$

Proof. Let $f(y)$ denote the density function of Y. Then

$$
\begin{aligned}
V[Y] &= \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy \\
&= \int_{-\infty}^{\mu-k\sigma} (y-\mu)^2 f(y)dy + \int_{\mu-k\sigma}^{\mu+k\sigma} (y-\mu)^2 f(y)dy \\
&+ \int_{\mu+k\sigma}^{\infty} (y-\mu)^2 f(y)dy
\end{aligned}
$$

TNote that $(y-\mu)^2 \geqslant k^2\sigma^2$ for all values of y between the limits of integration for the first and third integrals. If we replace the second integral by 0 and substitute $k^2\sigma^2$ for $(y-\mu)^2$ in the first and thid integrals to obtain the inequality

$$
V[Y] \geqslant \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(y)dy + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(y)dy
$$

Then

$$\sigma^2 \geqslant k^2\sigma^2 \left[\int_{-\infty}^{\mu-k\sigma} f(y)dy + \int_{\mu+k\sigma}^{\infty} f(y)dy \right]$$

or

$$\sigma^2 \geqslant k^2\sigma^2[P(Y \leqslant \mu - k\sigma) + P(Y \geqslant \mu + k\sigma)] = k^2\sigma^2 P(|Y - \mu| \geqslant k\sigma)$$

Dividing by $k^2\sigma^2$ we obtain

$$P(|Y - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}$$

or, equivalently,

$$P(|Y - \mu| < k\sigma) \geqslant 1 - \frac{1}{k^2}$$

■

Chapter 4

Multivariate Probability Distributions

4.1.0 §Bivariate and Multivariate Probability Distributions

Definition 4.1.1. *Let Y_1 and Y_2 be discrete random variables. The* **joint (or bivariate) probability function** *for Y_1 and Y_2 is given by:*

$$p(y_1,y_2) = P(Y_1 = y_1, Y_2 = y_2), \quad -\infty < y_1 < \infty, -\infty < y_2 < \infty$$

Note that the joint probability function for discrete random variables assigns nonzero probabilities to only finitely many or countably infinitely many pairs of values. Furthermore, the nonzero probabilities must sum to 1.

Theorem 4.1.1. *If Y_1 and Y_2 are discrete random variables with joint probability function $p(y_1,y_2)$, then:*

1. *$p(y_1,y_2) \geqslant 0$ for all y_1, y_2.*
2. *$\sum_{y_1}\sum_{y_2} p(y_1,y_2) = 1$, where the sum is over all values (y_1,y_2) that are assigned nonzero probabilities.*

Definition 4.1.2. *For any random variables Y_1 and Y_2, the joint (bivariate) distribution function $F(y_1,y_2)$ is:*

$$F(y_1,y_2) = P(Y_1 \leqslant y_1, Y_2 \leqslant y_2), \quad -\infty < y_1 < \infty, -\infty < y_2 < \infty$$

If both variables are discrete we have that:

$$F(y_1,y_2) = \sum_{t_1 \leqslant y_1} \sum_{t_2 \leqslant y_2} p(t_1,t_2)$$

Two random variables are said to be ***jointly continuous*** if their joint distribution function $F(y_1,y_2)$ is continuous in both arguments.

Definition 4.1.3. *Let Y_1 and Y_2 be continuous random variables with joint distribution function $F(y_1,y_2)$. If there exists a nonnegative function $f(y_1,y_2)$, such that*

$$F(y_1,y_2) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f(t_1,t_2)dt_2dt_1$$

for all $-\infty < y_1 < \infty, -\infty < y_2 < \infty$, then Y_1 and Y_2 are said to be **jointly continuous random variables**. *The function $f(y_1,y_2)$ is called the* **joint probability density function**.

Theorem 4.1.2. *If Y_1 and Y_2 are random variables with joint distribution function $F(y_1,y_2)$, then*

1. $F(-\infty,-\infty) = F(-\infty,y_2) = F(y_1,-\infty) = 0$
2. $F(\infty,\infty) = \lim_{y_1\to\infty}\lim_{y_2\to\infty} F(y_1,y_2) = 1$
3. *If $y_1^* \geqslant y_1$ and $y_2^* \geqslant y_2$, then*

$$F(y_1^*,y_2^*) - F(y_1^*,y_2) - F(y_1,y_2^*) + F(y_1,y_2) \geqslant 0$$

which follows as this expression is equal to $P(y_1 < Y_1 \leqslant y_1^*, y_2 < Y_2 \leqslant y_2^*) \geqslant 0$.

Theorem 4.1.3. *If Y_1 and Y_2 are jointly continuous random varaibels with a joint density function given by $f(y_1,y_2)$, then*

1. $f(y_1,y_2) \geqslant 0$ *for all y_1, y_2.*
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y_1,y_2)dy_1dy_2 = 1$

Note that volumes under the $f(y_1,y_2)$ surface correspond to probabilities.

This concepts can be straight-forwardly be extended to the case of n joint random variables.

4.2.0 §Marginal and Conditional Probability Distributions

Definition 4.2.1.

1. *Let Y_1 and Y_2 be jointly discrete random variables with probability function $p(y_1,y_2)$. THen the* **marginal probability functions** *of Y_1 and Y_2, respectively, are given by:*

$$p_1(y_1) = \sum_{\text{all } y_2} p(y_1,y_2) \;\; \text{and} \;\; p_2(y_2) = \sum_{\text{all } y_1} p(y_1,y_2)$$

2. *Let Y_1 and Y_2 be jointly continuous random variables with joint density function $f(y_1,y_2)$. Then the* **marginal density functions** *of Y_1 and Y_2, respectively, are given by:*

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1,y_2)dy_2 \;\; \text{and} \;\; f_2(y_2) = \int_{-\infty}^{\infty} f(y_1,y_2)dy_1$$

Definition 4.2.2. *If Y_1 and Y_2 are jointly discrete random variables with joint probability function $p(y_1, y_2)$ and marginal probability functions $p_1(y_1)$ and $p_2(y_2)$, respectively, then the* **conditional discrete probability function** *of Y_1 given Y_2 is*

$$p(y_1|y_2) = P(Y_1 = y_1|Y_2 = y_2) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p(y_1, y_2)}{p_2(y_2)}$$

provided that $p_2(y_2) > 0$.

Definition 4.2.3. *If Y_1 and Y_2 are jointly continuous random variables with joint density function $f(y_1, y_2)$, then the* **conditional distribution function** *of Y_1 given $Y_2 = y_2$ is*

$$F(y_1|y_2) = P(Y_1 \leqslant y_1|Y_2 = y_2)$$

which has the property that

$$F(y_1) = \int_{-\infty}^{\infty} F(y_1|y_2) f_2(y_2) dy_2$$

Moreover, we have the following:

$$\begin{aligned} F(y_1) &= \int_{-\infty}^{y_1} f_1(t_1) dt_1 = \int_{-\infty}^{y_1} \left[\int_{-\infty}^{\infty} f(t_1, y_2) dy_2 \right] dt_1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} f(t_1, y_2) dt_1 dy_2 \end{aligned}$$

Equating these expressions we find that

$$F(y_1|y_2) f_2(y_2) = \int_{-\infty}^{y_1} f(t_1, y_2) dt_1$$

or

$$F(y_1|y_2) = \int_{-\infty}^{y_1} \frac{f(t_1, y_2)}{f_2(y_2)} dt_1$$

Definition 4.2.4. *Let Y_1 and Y_2 be jointly continuous random variables with joint density function $f(y_1, y_2)$ and marginal densities $f_1(y_1)$ and $f_2(y_2)$, respectively. For any y_2 such that $f_2(y_2) > 0$, the conditional density of Y_1 given $Y_2 = y_2$ is given by:*

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$

and, for any y_1 such that $f_1(y_1) > 0$, the conditional density of Y_2 given $Y_1 = y_1$ is given by:

$$f(y_2|y_1) = \frac{f(y_1, y_2)}{f_1(y_1)}$$

4.3.0 §Independent Variables

Definition 4.3.1. *Let Y_1 have distribution function $F_1(y_1)$, Y_2 have distribution function $F_2(y_2)$, and Y_1 and Y_2 have joint distribution function $F(y_1, y_2)$. Then Y_1 and Y_2 are said to be* **independent** *if and only if*

$$F(y_1, y_2) = F_1(y_1) F_2(y_2)$$

for every pair of real numbers (y_1, y_2).

If Y_1 *and* Y_2 *are not independent, they are said to be* **dependent**.

Theorem 4.3.1. *If* Y_1 *and* Y_2 *are discrete random variables with joint probability function* $p(y_1, y_2)$ *and marginal probability functions* $p_1(y_1)$ *and* $p_2(y_2)$*, respectively, then* Y_1 *and* Y_2 *are independent if and only if*

$$p(y_1, y_2) = p_1(y_1)p_2(y_2)$$

for all pairs of real numbers (y_1, y_2).

If Y_1 *and* Y_2 *are continuous random variables with joint density function* $f(y_1, y_2)$ *and marginal density functions* $f_1(y_1)$ *and* $f_2(y_2)$*, respectively, then* Y_1 *and* Y_2 *are independent if and only if*

$$f(y_1, y_2) = f_1(y_1)f_2(y_2)$$

for all pairs of real numbers (y_1, y_2).

Theorem 4.3.2. *Let* Y_1 *and* Y_2 *have a joint density function* $f(y_1, y_2)$ *that is positive if and only if* $a \leqslant y_1 \leqslant b$ *and* $c \leqslant y_2 \leqslant d$*, for constants* a, b, c *and* d*; and* $f(y_1, y_2) = 0$ *otherwise. Then* Y_1 *and* Y_2 *are independent random variables if and only if*

$$f(y_1, y_2) = g(y_1)h(y_2)$$

where $g(y_1)$ *is a nonnegative function of* y_1 *along and* $h(y_2)$ *is a nonnegative function of* y_2 *along.*

These definitions can be analogously generalized to n random variables.

4.4.0 §The Covariance of Two Random Variables

Definition 4.4.1. *If* Y_1 *and* Y_2 *are random variables with means* μ_1 *and* μ_2*, respectively, the* **covariance** *of* Y_1 *and* Y_2 *is*

$$COV(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

Note that positive values indicate that Y_1 increases as Y_2 increases, and negative valus indicate that Y_1 decreases as Y_2 increases. A zero value of covariance indicates that the variables are ***uncorrelated*** and that there is no linear dependence between Y_1 and Y_2.

Definition 4.4.2. *We can standardize this value using the* **correlation coefficient**, ρ*, defined as*

$$\rho := \frac{COV(Y_1, Y_2)}{\sqrt{VAR[Y_1]VAR[Y_2]}}$$

Evidently, the correlation coefficient satisfies the inequality $-1 \leqslant \rho \leqslant 1$.

$\rho = +1$ implies perfect correlation, with all points falling on a straight line with positive slope. A value of $\rho = 0$ implies zero covariance and no correlation. Finally, a $\rho = -1$ implies perfect correlation, with all points falling on a straight line with negative slope.

Theorem 4.4.1. *If Y_1 and Y_2 are random variables with means μ_1 and μ_2, respectively, then*

$$COV(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = E[Y_1Y_2] - E[Y_1]E[Y_2]$$

Proof.

$$\begin{aligned} COV(Y_1, Y_2) &= E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= E[Y_1Y_2 - \mu_1Y_2 - \mu_2Y_1 + \mu_1\mu_2] \\ &= E[Y_1Y_2] - E[Y_1]E[Y_2] - E[Y_1]E[Y_2] + E[Y_1]E[Y_2] \\ &= E[Y_1Y_2] - E[Y_1]E[Y_2] \end{aligned}$$

■

Definition 4.4.3. *Let $g(Y_1, Y_2, ..., Y_k)$ be a function of the discrete random variables, $Y_1, Y_2, ..., Y_k$, which have probability function $p(y_1, y_2, ..., y_k)$. Then the* **expected value of** $g(Y_1, Y_2, ..., Y_k)$ *is*

$$E[g(Y_1, Y_2, ..., Y_k)] = \sum_{all\ y_k} \cdots \sum_{all\ y_2} \sum_{all\ y_1} g(y_1, y_2, ..., y_k)p(y_1, y_2, ..., y_k)$$

If $Y_1, Y_2, ..., Y_k$ are continuous random variables with joint density function $f(y_1, y_2, ..., y_k)$, then

$$E[g(Y_1, ..., Y_k)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, ..., y_k)f(y_1, ..., y_k)dy_1...dy_k$$

Remark 4.4.1. We now show our definition of $E[Y_1]$ lines up with this definition when considering two random variables Y_1 and Y_2 with density function $f(y_1, y_2)$. In particular, we wish to find the expected value of $g(Y_1, Y_2) = Y_1$. Then from this definition we have

$$\begin{aligned} E[Y_1] &= \int_{-\infty}^{\infty} y_1 f(y_1, y_2)dy_2dy_1 \\ &= \int_{-\infty}^{\infty} y_1 \left[\int_{-\infty}^{\infty} f(y_1, y_2)dy_2\right] dy_1 \\ &= \int_{-\infty}^{\infty} y_1 f_1(y_1)dy_1 \end{aligned}$$

which agrees with our previous definition.

Theorem 4.4.2. *Let Y_1 and Y_2 be independent random variables and $g(Y_1)$ and $h(Y_2)$ be functions of only Y_1 and Y_2, respectively. Then*

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)]$$

provided the expectations exist.

Proof. We shall prove the result for the continuous case. Let $f(y_1,y_2)$ denote the joint density function of Y_1 and Y_2. The product $g(Y_1)h(Y_2)$ is a function of Y_1 and Y_2. Hence,

$$\begin{aligned}
E[g(Y_1)h(Y_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f(y_1,y_2)dy_2dy_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f_1(y_1)f_2(y_2)dy_2dy_1 \\
&= \int_{-\infty}^{\infty} g(y_1)f_1(y_1)\left[\int_{-\infty}^{\infty} h(y_2)f_2(y_2)dy_2\right]dy_1 \\
&= \int_{-\infty}^{\infty} g(y_1)f_1(y_1)E[h(Y_2)]dy_1 \\
&= E[h(Y_2)]\int_{-\infty}^{\infty} g(y_1)f_1(y_1)dy_1 \\
&= E[h(Y_2)]E[g(Y_1)]
\end{aligned}$$

■

Theorem 4.4.3. *If Y_1 and Y_2 are independent random variables, then*

$$COV(Y_1,Y_2) = 0$$

Thus, indepedent random variables must be uncorrelated.

Proof.

$$\begin{aligned}
COV(Y_1,Y_2) &= E[Y_1Y_2] - E[Y_1]E[Y_2] \\
&= E[Y_1]E[Y_2] - E[Y_1]E[Y_2] \\
&= 0
\end{aligned}$$

■

4.5.0 §The Expected Value and Variance of Linear Functions of Random Variables

Theorem 4.5.1. *Let $Y_1,...,Y_n$ and $X_1,...,X_m$ be random variables with $E[Y_i] = \mu_i$ and $E[X_j] = \xi_j$. Define*

$$U_1 = \sum_{i=1}^{n} a_iY_1 \quad and \quad U_2 = \sum_{j=1}^{m} b_jX_j$$

for constants $a_1,a_2,...,a_n$ and $b_1,b_2,...,b_m$. Them the following hold:

1. $E[U_1] = \sum_{i=1}^{n} a_i\mu_i$

2. $V[U_1] = \sum_{i=1}^{n} a_i^2 V[Y_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j COV(Y_i, Y_j)$, *where the double sum is over all pairs* (i, j) *with* $i < j$.

3. $COV(U_1, U_2) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j COV(Y_i, X_j)$

Proof. 1. follows directly from the linearity of the expected value function. To proceed with 2. we appeal to the definition of variance and write:

$$\begin{aligned}
V[U_1] = E[(U_1 - E(U_1))^2] &= E\left[\left(\sum_{i=1}^{n} a_i Y_i - \sum_{i=1}^{n} a_i \mu_i\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} a_i (Y_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n} a_i^2 (Y_i - \mu_i)^2 + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} a_i a_j (Y_i - \mu_i)(Y_j - \mu_j)\right] \\
&= \sum_{i=1}^{n} a_i^2 E[(Y_i - \mu_i)^2] + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} a_i a_j E[(Y_i - \mu_i)(Y_j - \mu_j)]
\end{aligned}$$

By definition of variance and covariance, we have

$$V[U_1] = \sum_{i=1}^{n} a_i^2 V[Y_i] + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} a_i a_j COV(Y_i, Y_j)$$

Because $COV(Y_i, Y_j) = COV(Y_j, Y_i)$, we can write

$$V[U_1] = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j COV(Y_i, Y_j)$$

We now apply similar steps to find 3.:

$$\begin{aligned}
COV(U_1, U_2) = E[(U_1 - E(U_1))(U_2 - E(U_2))] \\
&= E\left[\left(\sum_{i=1}^{n} a_i Y_i - \sum_{i=1}^{n} a_i \mu_i\right)\left(\sum_{j=1}^{m} b_j X_j - \sum_{j=1}^{m} b_j \xi_j\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n} a_i (Y_i - \mu_i)\right)\left(\sum_{j=1}^{m} b_j (X_j - \xi_j)\right)\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j (Y_i - \mu_i)(X_j - \xi_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j E[(Y_i - \mu_i)(X_j - \xi_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j COV(Y_i, X_j)
\end{aligned}$$

On observing $COV(Y_i, Y_i) = V(Y_i)$, we see that 2. is a special case of 3.. ■

Claim 4.5.2. *Let $Y_1, .., Y_n$ be independent random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$. Define*

$$\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

then $E[\overline{Y}] = \mu$ and $V[\overline{Y}] = \sigma^2/n$.

Proof. Since $\overline{Y}$ is a linear function of random variables we have that

$$E[\overline{Y}] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}n\mu = \mu$$

Next for the variance we have that

$$V[\overline{Y}] = \sum_{i=1}^{n} \frac{1}{n}^2 V[Y_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{1}{n}\frac{1}{n}COV(Y_i, Y_j)$$

But each Y_i and Y_j are independent for $i \neq j$, so $COV(Y_i, Y_j) = 0$ for all $i \neq j$. Thus

$$V[\overline{Y}] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

■

4.6.0 §Bivariate Normal Distribution

Definition 4.6.1. *Two continuous random variables Y_1 and Y_2 follow a bivariate normal distribution if their joint density function is:*

$$f(y_1, y_2) = \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \quad -\infty < y_1 < \infty, -\infty < y_2 < \infty,$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right]$$

The marginal distributions of Y_1 and Y_2 are normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. Moreover, $COV(Y_1, Y_2) = \rho\sigma_1\sigma_2$.

Chapter 5

Functions of Random Variables

5.1.0 §Probability Distributions of Functions of Random Variables

In this chapter we assume that the populations are large in comparison to the sample size so that the random variables obtained through a random sample are in fact independent of one another.

In the discrete case this implies that the joint probability function for $Y_1, ..., Y_n$, all sampled from the same population, is given by

$$p(y_1, ..., y_n) = p(y_1)p(y_2)...p(y_n)$$

In the continuous case the joint density function is

$$f(y_1, ..., y_n) = f(y_1)f(y_2)...f(y_n)$$

The statement "$Y_1, ..., Y_n$, is a random sample from a population with density $f(y)$" will mean that the random variables are independent with common density function $f(y)$.

In the upcoming sections we will explore three methods for finding the probability distribution for a function of random variables, and one for finding the joint distribution of several functions of random variables.

5.2.0 §The Method of Distribution Functions

Remark 5.2.1. Consider random variables $Y_1, Y_2, ..., Y_n$ and a function $U(Y_1, ..., Y_n)$. The method of distribution functions is typically used when the Y's have continuous distributions. First, find the distribution function for U, $F_U(u) = P(U \leqslant u)$, by using previous methods. To do so, we must find the region in the $y_1, y_2, ..., y_n$ space for which $U \leqslant u$ and then find $P(U \leqslant u)$ by integrating $f(y_1, y_2, ..., y_n)$ over this region. The density function for U is then obtained by differentiating the distribution function, $F_U(u)$.

Example 5.2.1. Suppose Y is a random variable with density function

$$f(y) = \begin{cases} 2y, & 0 \leqslant y \leqslant 1, \\ 0, & elsewhere. \end{cases}$$

Define $U(Y) = 3Y - 1$. To employ the distribution function approach, we must find:

$$F_U(u) = P(U \leqslant u) = P(3Y - 1 \leqslant u) = P\left(Y \leqslant \frac{u+1}{3}\right)$$

If $u < -1$, then $(u+1)/3 < 0$ and, therefore, $F_U(u) = P(Y \leqslant (u+1)/3) = 0$. Also, if $u > 2$, then $(u+1)/3 > 1$, and $F_U(u) = P(Y \leqslant (u+1)/3) = 1$. However, if $-1 \leqslant u \leqslant 2$, the probability can be written as an integral of $f(y)$, and

$$P\left(Y \leqslant \frac{u+1}{3}\right) = \int_0^{(u+1)/3} 2y dy = \left(\frac{u+1}{3}\right)^2$$

Thus the distribution of the random variable U is given by

$$F_U(u) = \begin{cases} 0, & u < -1, \\ \left(\frac{u+1}{3}\right)^2, & -1 \leqslant u \leqslant 2, \\ 1, & u > 2, \end{cases}$$

and the density function for U is

$$f_U(u) = \frac{d}{du}(F_U(u)) = \begin{cases} (2/9)(u+1), & -1 \leqslant u < 2, \\ 0, & elsewhere. \end{cases}$$

In the bivariate case, let Y_1 and Y_2 be random variables with joint density $f(y_1, y_2)$ and let $U(Y_1, Y_2)$ be a function of Y_1 and Y_2. Then for every point (y_1, y_2), there corresponds one and only one value of U. If we can find the region of values (y_1, y_2) such that $U \leqslant u$, then the integral of the joint density function $f(y_1, y_2)$ over this region equals $P(U \leqslant u) = F_U(u)$.

Definition 5.2.1. *Distribution Function Method Let U be a function of the random variables $Y_1, Y_2, ..., Y_n$.*

1. *Find the region $U = u$ in the $(y_1, y_2, ..., y_n)$ space.*
2. *Find the region $U \leqslant u$.*
3. *Find $F_U(u) = P(U \leqslant u)$ by integrating $f(y_1, y_2, ..., y_n)$ over the region $U \leqslant u$.*
4. *Find the density function $f_U(u)$ by differentiating $F_U(u)$. Thus, $f_U(u) = \frac{d}{du} F_U(u)$.*

In certain instances it is possible to transform a random variable with a uniform distribution on $(0, 1)$ into a random variable with some other specified distribution function.

Example 5.2.2. Let U be a uniform random variable on $(0, 1)$, we wish to find G such that $G(U)$ possesses an exponential distribution with mean β.

The distribution function of U is

$$F_U(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \leqslant u \leqslant 1, \\ 1, & u > 1. \end{cases}$$

Let Y denote a random variable with an exponential distribution with mean β:

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ 1 - e^{-y/\beta}, & y \geqslant 0 \end{cases}$$

Let $0 < u < 1$, then there is a unique y such that $F_Y(y) = u$. Thus, $F_Y^{-1}(u)$, $0 < u < 1$, is well defined. In particular $F_Y^{-1}(u) = -\beta \ln(1-u)$. Consider the random variable $F_Y^{-1}(U) = -\beta \ln(1-U)$ and observe that if $y > 0$,

$$\begin{aligned} P(F_Y^{-1}(U) \leqslant y) &= P[-\beta \ln(1-U) \leqslant y] \\ &= P[\ln(1-U) \geqslant -y/\beta] \\ &= P(U \leqslant 1 - e^{-y/\beta}) \\ &= 1 - e^{-y/\beta} \end{aligned}$$

Also, $P[F_Y^{-1}(U) \leqslant y] = 0$ if $y \leqslant 0$. Thus, $F_Y^{-1}(U) = -\beta \ln(1-U)$ possesses an exponential distribution with mean β, as desired.

This technique is useful in conjunction with the methods computers often use to generate random numbers.

5.3.0 §The Method of Transformations

Remark 5.3.1. If we are given the density function of a random variable Y, the method of transformations results in a general expression for the density of $U = h(Y)$ for an increasing or decreasing function $h(y)$. Then if Y_1 and Y_2 have a bivaraite distribution, we can use the univariate result to find the joint density of Y_1 and $U = h(Y_1, Y_2)$. By integrating over y_1, we find the marginal probability density function of U, which is our objective.

Suppose that $h(y)$ is an increasing function of y and that $U = h(Y)$, where Y has density function $f_Y(y)$. Then $h^{-1}(u)$ is an increasing function of u: if $u_1 < u_2$, then $h^{-1}(u_1) = y_1 < y_2 = h^{-1}(u_2)$. The set of points y such that $h(y) \leqslant u_1$ is precisely the same set as those y such that $y \leqslant h^{-1}(u_1)$. Therefore,

$$P(U \leqslant u) = P(h(Y) \leqslant u) = P(Y \leqslant h^{-1}(u))$$

so $F_U(u) = F_Y(h^{-1}(u))$. Then differentiating with respect to u, we have

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du} F_Y(h^{-1}(u)) = f_Y(h^{-1}(u)) \frac{d}{du}(h^{-1}(u))$$

If $h(y)$ is a decreasing function of y, then $h^{-1}(u)$ is a decreasing function of u. That is, if $u_1 < u_2$, then $h^{-1}(u_1) = y_1 > y_2 = h^{-1}(u_2)$. The set of points y such that $h(y) \leqslant u_1$ is then the same as the set of points such that $y \geqslant h^{-1}(u_1)$.

It follows that for $U = h(Y)$:

$$P(U \leqslant u) = P(Y \geqslant h^{-1}(u))$$

so $F_U(u) = 1 - F_Y(h^{-1}(u))$. If we differentiate with respect to u we obtain:

$$f_U(u) = -f_Y(h^{-1}(u))\frac{d}{du}(h^{-1}(u))$$

Definition 5.3.1. *The set of points $\{y : f_Y(y) > 0\}$ is called the* **support** *of the density function $f_Y(y)$.*

We only require that $h(\cdot)$ be increasing or decreasing on the support of $f_Y(y)$ in order to be able to apply this method.

Theorem 5.3.1. *Let Y have a probability density function $f_Y(y)$. If $h(y)$ is increasing for all y such that $f_Y(y) > 0$, then $U = h(Y)$ has density function*

$$f_U(u) = f_Y(h^{-1}(u))\frac{d}{du}(h^{-1}(u))$$

and if $h(y)$ is decreasing the density function is

$$f_U(u) = -f_Y(h^{-1}(u))\frac{d}{du}(h^{-1}(u))$$

Definition 5.3.2. *Transformation Method Let $U = h(Y)$, where $h(y)$ is either an increasing or decreasing function of y for all y such that $f_Y(y) > 0$.*

1. *Find the inverse function, $y = h^{-1}(u)$,*
2. *Evaluate $\frac{d}{du}(h^{-1}(u))$,*
3. *Find $f_U(u)$ by*

$$f_U(u) = f_Y(h^{-1}(u))\left|\frac{d}{du}(h^{-1}(u))\right|$$

Remark 5.3.2. If $Z = h(X, Y)$, and there exists an inverse such that $Y = g_Y^{-1}(x, z)$, then we have that

$$f_{XZ}(x, z) = f_{XY}(x, g_Y^{-1}(x, z))\left|\frac{d}{dz}g_Y^{-1}(x, z)\right|$$

Then we have that

$$f_Z(z) = \int_{Support\ X} f_{XZ}(x, z)dx$$

5.4.0 §Method of Moment-Generating Functions

Remark 5.4.1. This method is based on the uniqueness theoerm, which states that, if two random variables have identical moment-generating functions, the two random variables possess the same probability distributions. To use this method, we must find the moment-generating function for U and compare it with the moment-generating functions for the common discrete and continuous distributions.

Theorem 5.4.1. *Let $m_X(t)$ and $m_Y(t)$ denote the moment-generating functions of random variables X and Y, respectively. If both moment-generating functions exist and $m_X(t) = m_Y(t)$ for all values of t, then X and Y have the same probability distribution.*

Recall 5.4.2. Recall that

$$M_X(t) = E[e^{xt}] = \begin{cases} \int_{all\ x} e^{xt} f_X(x) d & \text{for continuous RV} \\ \sum_{all\ x} e^{xt} P_X(X = x) & \text{for discrete RV} \end{cases}$$

The first step in using this theorem is to find the moment-generating function of U: $m_U(t) = E[e^{tU}]$.

Theorem 5.4.2. *Let $Y_1, Y_2, ..., Y_n$ be independent random variables with moment-generating functions $m_{Y_1}(t), m_{Y_2}(t), ..., m_{Y_n}(t)$, respectively. If $U = Y_1 + Y_2 + \ldots + Y_n$, then*

$$m_U(t) = m_{Y_1}(t) \cdot m_{Y_2}(t) \cdot ... \cdot m_{Y_n}(t)$$

Proof. Recall that since the random variables $Y_1, Y_2, ..., Y_n$ are independent,

$$\begin{aligned} m_U(t) &= E[e^{t(Y_1+...+Y_n)}] \\ &= E[e^{tY_1} e^{tY_2} ... e^{tY_n}] \\ &= E[e^{tY_1}] \cdot E[e^{tY_2}] \cdot ... \cdot E[e^{tY_n}] \end{aligned}$$

Thus by definition of moment-generating functions:

$$m_U(t) = m_{Y_1}(t) \cdot m_{Y_1}(t) \cdot ... \cdot m_{Y_n}(t)$$

■

The method of moment-generating functions can also be used to establish interesting and useful results about the distributions of functions of normally distributed random variables.

Theorem 5.4.3. *Let $Y_1, Y_2, ..., Y_n$ be independent normally distributed random variables with $E[Y_i] = \mu_i$ and $V[Y_i] = \sigma_i^2$, for $i \in \{1, 2, ..., n\}$, and let $a_1, a_2, ..., a_n$ be constants. If*

$$U = \sum_{i=1}^{n} a_i Y_i$$

then U is a normally distributed random variable with

$$E[U] = \sum_{i=1}^{n} a_i \mu_i$$

and

$$V[U] = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

Proof. Because Y_i is normally distributed with mean μ_i and variance σ_i^2, Y_i has moment-generating function given by

$$m_{Y_i}(t) = \exp\left\{\mu_i t + \frac{\sigma_i^2 t^2}{2}\right\}$$

Therefore, $a_i Y_i$ has moment-generating function given by

$$m_{a_iY_i}(t) = E[e^{ta_iY_i}] = m_{Y_i}(a_i t) = \exp\left\{\mu_i a_i t + \frac{a_i^2\sigma_i^2 t^2}{2}\right\}$$

Because the random variables Y_i are independent, the random variables a_iY_i are independent, for $i \in \{1, 2, ..., n\}$, and the previous theorem implies that

$$\begin{aligned} m_U(t) &= m_{a_1Y_1}(t) \cdot ... \cdot m_{a_nY_n}(t) \\ &= \exp\left\{\mu_1 a_1 t + \frac{a_1^2\sigma_1^2t^2}{2}\right\} \cdot ... \cdot \exp\left\{\mu_n a_n t + \frac{a_n^2\sigma_n^2t^2}{2}\right\} \\ &= \exp\left(t\sum_{i=1}^n a_i\mu_i + \frac{t^2}{2}\sum_{i=1}^n a_i^2\sigma_i^2\right) \end{aligned}$$

Thus, U has a normal distribution with mean $\sum\limits_{i=1}^n a_i\mu_i$ and variance $\sum\limits_{i=1}^n a_i^2\sigma_i^2$. ■

Theorem 5.4.4. *Let $Y_1, Y_2, ..., Y_n$ be independent normally distributed random variables with $E[Y_i] = \mu_i$ and $V[Y_i] = \sigma_i^2$, for $i \in \{1, 2, ..., n\}$. Define Z_i by*

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i}, \quad i \in \{1, 2, ..., n\}$$

Then $\sum\limits_{i=1}^n Z_i^2$ has a χ^2 distribution with n degrees of freedom.

Proof. Note each Z_i is normally distributed with mean 0 and variance 1. It can be shown that Z_i has a χ^2 distribution with 1 degree of freedom. That is,

$$m_{Z_i^2}(t) = (1 - 2t)^{-1/2}$$

and with $V = \sum\limits_{i=1}^n$,

$$\begin{aligned} m_V(t) &= m_{Z_1^2}(t) \cdot ... \cdot m_{Z_n^2}(t) \\ &= (1 - 2t)^{-1/2} \cdot ... \cdot (1 - 2t)^{-1/2} = (1 - 2t)^{-n/2} \end{aligned}$$

Because moment-generating functions are unique, V has a χ^2 distribution with n degrees of freedom. ■

Definition 5.4.1. *Moment-Generating Function Method Let U be a function of the random variables $Y_1, ..., Y_n$.*

1. *Find the moment-generating function for U, $m_U(t)$.*
2. *Compare $m_U(t)$ with other well-known moment-generating functions. If $m_U(t) = m_V(t)$ for all values of t, uniqueness implies that U and V have identical distributions.*

5.5.0 §Multivariate Transformations

Definition 5.5.1. *Method If $U = h_1(X,Y)$ and $W = h_2(X,Y)$ are functions of joint continuous random variables with inverses $X = g_X^{-1}(u,w)$ and $Y = g_Y^{-1}(u,w)$. Then the joint density function of U and W is given by*

$$f_{UW}(u,w) = f_{XY}(g_X^{-1}(u,w), g_Y^{-1}(u,w)) \left| \frac{\partial(g_X^{-1}, g_Y^{-1})}{\partial(u,w)} \right|$$

This extends for k functions of k random variables.

5.6.0 §Order Statistics

Remark 5.6.1. We often order ovserved random variables according to their magnitudes. The resulting ordered variabels are called ***order statistics***.

Definition 5.6.1. *Let $Y_1, ..., Y_n$ denote independent continuous random variables with distribution function $F(y)$ and density function $f(y)$, we denote the ordered variables by*

$$Y_{(1)} \leqslant Y_{(2)} \leqslant ... \leqslant Y_{(n)}$$

Then

$$Y_{(1)} = \min_{1 \leqslant i \leqslant n} Y_i, Y_{(n)} = \max_{1 \leqslant i \leqslant n} Y_i$$

Definition 5.6.2. *Note, $(Y_{(n)} \leqslant y)$ will occur if and only if $Y_i \leqslant y$ for all i, so we have*

$$P(Y_{(n)} \leqslant y) = P(Y_1 \leqslant y, ..., Y_n \leqslant y)$$

and as the Y_i are independent

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leqslant y) = \prod_{i=1}^{n} P(Y_i \leqslant y) = [F(y)]^n$$

so

$$f_{Y_{(n)}}(y) = n[F(y)]^{n-1} f(y)$$

Definition 5.6.3. *Note, $(Y_{(1)} > y)$ occurs if and only if $Y_i \geqslant y$ for all i, so*

$$F_{Y_{(1)}}(y) = P(Y_{(1)} \leqslant y) = 1 - P(Y_{(1)} > y) = 1 - P(Y_1 \geqslant y, ..., Y_n \geqslant y)$$

Then since the Y_i are independent

$$F_{Y_{(1)}}(y) = 1 - \prod_{i=1}^{n} P(Y_i > y) = 1 - [1 - F(y)]^n$$

so

$$f_{Y_{(1)}}(y) = n[1 - F(y)]^{n-1} f(y)$$

Definition 5.6.4. *Let $Y_1,...,Y_n$ be i.i.d. random variables, then the joint density of $Y_{(1)},...,Y_{(n)}$ is*

$$f_{(1)(2)...(n)}(y_1,y_2,...,y_n) = \begin{cases} n!f(y_1)f(y_2)...f(y_n) & y_1 \leqslant y_2 \leqslant ... \leqslant y_n \\ 0, & elsewhere \end{cases}$$

Theorem 5.6.1. *Let $Y_1,...,Y_n$ be i.i.d continuous random variables with common CDF $F(y)$ and common pdf $f(y)$. If $Y_{(k)}$ denotes the kth order statistic the pdf of $Y_{(k)}$ is*

$$f_{(k)}(y_k) = \frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1}[1-F(y_k)]^{n-k}f(y_k)$$

If $1 \leqslant j \leqslant k \leqslant n$, the joint pdf of $Y_{(j)}$ and $Y_{(k)}$ is

$$f_{(j)(k)}(y_j,y_k) = \frac{n!}{(j-1)!(k-1-j)!(n-k)!}[F(y_j)]^{j-1}[F(y_k)-F(y_j)]^{k-1-j}[1-F(y_k)]^{n-k}f(y_j)f(y_k)$$

for $-\infty < y_j < y_k < \infty$.

Example 5.6.1. $X_1,...,X_n$ i.i.d uniform distributed random variables between 0 to 1. THen

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k} = \frac{\Gamma(n-k+k+1)}{\Gamma(k)\Gamma(n-k+1)}x^{k-1}(1-x)^{(n-k+1)-1}$$

for $0 \leqslant x \leqslant 1$ so $X_{(k)} \sim beta(\alpha = k, \beta = n-k+1)$

Chapter 6

Central Limit Theorem and Sampling

6.1.0 §Sampling Distributions Related to the Normal Distribution

Definition 6.1.1. *Let $Y_1, ..., Y_n$ be independent random variables with the same distribution. We estimate the population mean these random variables come from using the following random variable:*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

This random variable is an example of a *statistic*.

Definition 6.1.2. *A* **statistic** *is a function of the observable random variables in a sample and known constants.*

Statistics are used to make inferences (estimates or decisions) about unknown population parameters. All statistics are random variables, and hence have probability distributions called their ***sampling distributions***.

Theorem 6.1.1. *Let $Y_1, ..., Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$.

Proof. Because $Y_1, ..., Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2, Y_i ($i \in \{1, 2, ..., n\}$) are independent, normally distributed variables, with $E[Y_1] = \mu$ and $VAR[Y_i] = \sigma^2$. Further,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is a linear combination of the Y_i, and hence by a preceeding theorem, we can conclude that $\overline{Y}$ is normally distributed with

$$E[\overline{Y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and

$$VAR[\overline{Y}] = VAR\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \sigma^2/n$$

since the Y_i are independent. That is, the sampling distribution of $\overline{Y}$ is normal with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$. ■

Corollary 6.1.2. $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim normal(\mu = 0, \sigma = 1)$

Theorem 6.1.3. *Let $Y_1, ..., Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, with $i \in \{1, 2, ..., n\}$, and*

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2$$

has a χ^2 distribution with n degrees of freedom.

Proof. Since the Y_i is a random sample from a normal distribution with mean μ and variance σ^2, each of the $Z_i = (Y_i - \mu)/\sigma$ have a standard normal distribution. Further, the random variables Z_i are independent because the random variables Y_i are independent. The fact that $\sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with n degrees of freedom follows from a previous theorem. ■

Theorem 6.1.4. *Cochran's Theorem Let $Z_1^2, ..., Z_n^2$ be a i.i.d chi-square random variables with their own respective degrees of freedom $\nu_1, ..., \nu_n$. Then*

$$\sum_{i=1}^{n} Z_i^2 \sim chi - squared(\nu = \sum_{i=1}^{n}\nu_i)$$

Definition 6.1.3. *The sample variance for a random sample $Y_1, Y_2, ..., Y_n$ is given as follows*

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

and it is an unbiased estimator of σ^2.

Theorem 6.1.5. *Let $Y_1, ..., Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

has a χ^2 distribution with $(n-1)$ degrees of freedom. Also, $\overline{Y}$ and S^2 are independent random variables.

Definition 6.1.4. *Let Z be a standard normal random variable and let W be a χ^2-distributed variables with v degrees of freedom. Then, if Z and W are independent,*

$$T = \frac{Z}{\sqrt{W/v}}$$

is said to have a t **distribution** *with v degrees of freedom. T has a density function*

$$f_T(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}, -\infty < t < \infty$$

If $Y_1, ..., Y_n$ constitute a random sample from a normal population with mean μ and variance σ^2, $Z = \frac{(\overline{Y}-\mu)}{\sigma/\sqrt{n}}$ has a standard normal distribution. From a previous theorem we have that $W = (n-1)S^2/\sigma^2$ has a χ^2 distribution with $v = n-1$ degrees of freedom and that Z and W are independent. Therefore, by our definition,

$$T = \frac{Z}{\sqrt{W/v}} = \frac{\sqrt{n}(\overline{Y}-\mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{n}\left(\frac{\overline{Y}-\mu}{S}\right)$$

has a t distribution with $(n-1)$ degrees of freedom.

Remark 6.1.1. For $v > 2$, the T distributed random variable has $E[T] = 0$ and $VAR[T] = \frac{v}{v-2}$.

Definition 6.1.5. *Let W_1 and W_2 be independent χ^2-distributed random variables with v_1 and v_2 degrees of freedom, respectively. Then*

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom. Then F has the following density function

$$f_F(x) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)}\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2}-1}\left(1+\frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}}, x > 0$$

Considering two indepenent random samples from normal distributions, we know that $W_1 = (n_1-1)S_1^2/\sigma_1^2$ and $W_2 = (n_2-1)S_2^2/\sigma_2^2$ have independent χ^2-distributions with $v_1 = (n_1-1)$ and $v_2 = (n_2-1)$ degrees of freedom, respectively. Thus, by definition we have that

$$F = \frac{W_1/v_1}{W_2/v_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with (n_1-1) numerator degrees of freedom and (n_2-1) denominator degrees of freedom.

Remark 6.1.2. If $v_2 > 2$, then $E[F] = \frac{v_2}{v_2-2}$, and if $v_2 > 4$, then

$$V[F] = \frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$$

Corollary 6.1.6. *Let $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ for $\overline{X}$ normally distributed with mean μ and standard deviation σ. Then $T = \frac{Z}{\sqrt{W/(n-1)}} \sim T_{df=n-1}$ where $W \sim \chi^2_{df=n-1}$. It then follows that*

$$T^2 = \frac{Z^2/1}{W/(n-1)} \sim F_{dfnum=1,dfden=n-1}$$

6.2.0 §The Central Limit Theorem

Theorem 4.

Central Limit Theorem Let $Y_1, Y_2, ..., Y_n$ be indepedent and identically distributed random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2 < \infty$. Define

$$U_n = \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Then the distribution function for U_n converges to the standard normal distribution function as $n \to \infty$. That is,

$$\lim_{n\to\infty} P(U_n \leqslant u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

for all u.

Theorem 6.2.1. *Let Y and $Y_1, Y_2, Y_3, ...,$ be random variables with moment-generating functions $m(t)$ and $m_1(t), m_2(t), m_3(t), ...,$ respectively. If*

$$\lim_{n\to\infty} m_n(t) = m(t)$$

for all real t, then the distribution function of Y_n converges to the distribution function of Y as $n \to \infty$.

We now sketch a proof of the Central Limit Theorem:

Proof. Write

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^n Y_i - n\mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$$

where $Z_i = \frac{Y_i-\mu}{\sigma}$. Because the random variables Y_i's are independent and identically distributed, Z_i, $i \in \{1, 2, ..., n\}$, are independent, and identically distributed with $E[Z_i] = 0$ and $V[Z_i] = 1$.

Since the moment-generating function of the sum of independent random variables is the product of their individual moment-generating functions,

$$m_{\sum Z_i}(t) = m_{Z_1}(t) \cdot ... \cdot m_{Z_n}(t) = [m_{Z_1}(t)]^n$$

and

$$m_{U_n}(t) = m_{\sum Z_i}(t/\sqrt{n}) = \left[m_{Z_1}(t/\sqrt{n})\right]^n$$

By Taylor's theorem, with remainder,

$$m_{Z_1}(t) = m_{Z_1}(0) + m'_{Z_1}(0)t + m''_{Z_1}(\xi)\frac{t^2}{2},$$

where $\xi \in (0,t)$, and because $m_{Z_1}(0) = E[e^{0Z_1}] = E[1] = 1$, and $m_{Z_1}(0) = E[Z_1] = 0$,

$$m_{Z_1}(t) = 1 + \frac{m''Z_1(\xi)}{2}t^2,$$

Therefore,

$$m_{U_n}(t) = \left[1 + \frac{m''_{Z_1}(\xi_n)}{2}(t/\sqrt{n})^2\right]^n = \left[1 + \frac{m''_{Z_1}(\xi_n)t^2/2}{n}\right]^n$$

for $\xi_n \in (0, t/\sqrt{n})$. Notice that as $n \to \infty$, $\xi_n \to 0$ and $m''_{Z_1}(\xi_n)t^2/2 \to m''_{Z_1}(0)t^2/2 = E[Z_1^2]t^2/2 = t^2/2$ because $E[Z_1^2] = V[Z_1] = 1$. Recall that if

$$\lim_{n\to\infty} b_n = b, \quad then \quad \lim_{n\to\infty}\left(1 + \frac{b_n}{n}\right)^n = e^b$$

Finally,

$$\lim_{n\to\infty} m_{U_n}(t) = \lim_{n\to\infty}\left[1 + \frac{m''_{Z_1}(\xi_n)t^2/2}{n}\right]^n = e^{t^2/2}$$

the moment-generating function for a standard normal random variable. Applying the previous theorem, we conclude that U_n has a distribution function that converges to the distribution function of the standard normal random variable. ■

Let $Y \sim binomial(n,p)$. Then Y can be viewed as the sum $Y = \sum_{i=1}^{n} X_i$, where each X_i is an independent Bernoulli random variable with probability of success p. That is $E[X_i] = p$ and $V[X_i] = p(1-p)$ for each i, so consequently when n is large, the sample fraction of successes

$$\frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

possesses an approximately normal sampling distribution with mean $E[X_i] = p$ and variance $V[X_i]/n = p(1-p)/n$. It then follows that for large n, $Y \sim norm(\mu = np, \sigma^2 = np(1-p))$.

Thus, using the Central Limit Theorem we establish that if Y is a binomial random variable with parameters n and p, and if n is large, then Y/n has approximately a normal distribution with mean $\mu = p$ and variance $\sigma^2 = p(1-p)/n$. Equivalently, for large n, we can think of Y as having approximately a normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

For the CLT to be used on proportions we require that either $n > 9\left(\frac{\text{larger of p and q}}{\text{smaller of p and q}}\right)$, or from a sample we have at least 10 of each response/category.

Definition 6.2.1. *When we approximate $Y \sim norm(\mu = np, \sigma^2 = np(1-p))$ for Y binomial for n sufficiently large, the* **continuity correction** *takes $P(Y \leqslant d + 0.5)$ and $P(Y \geqslant d - 0.5)$ when approximating $P(Y \leqslant d)$ and $P(Y \geqslant d)$ for $d \in \mathbb{Z}$.*

Chapter 7

Estimation

Recall that populations are characterized by numerical descriptive measures, called ***parameters***, so the objective of many statistical investigations is to estimate the value of one or more relevant parameters.

In general, we call the parameter of interest in an experiment the ***target parameter***.

Definition 7.0.1. *A* **point estimate** *of a parameter is a single value estimate.*

Definition 7.0.2. *An* **interval estimate** *of a parameter is an open interval (a,b) in which it is intended that the parameter of interest is enclosed in the interval.*

Definition 7.0.3. *An* **estimator** *is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.*

For example, the sample mean

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is one possible point estimator of the population mean μ.

7.1.0 §The Bias and Mean Square Error of Point Estimators

First, note that we cannot evaluate the "goodness" of a point estimation procedure on the basis of the value of a single estimate. Rather, we must observe the results when the estimation procedure is used many, many times.

Now, suppose we wish to specify a point estimate for a population parameter θ. The estimator of θ will be indicated by the symbol $\hat{\theta}$. It is highly desirable for the sampling distribution of the estimator, i.e. the distribution of estimates, to cluster about the target parameter. In other words, we would like the mean or expected value of the distribution of estimates to equal the parameter estimated: $E(\hat{\theta}) = \theta$.

Definition 7.1.1. *Point estimators, $\hat{\theta}$, that satisfy this property, $E(\hat{\theta}) = \theta$, are said to be* **unbiased**. *If $E(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ is said to be* **biased**.

Definition 7.1.2. *The* **bias** *of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.*

Definition 7.1.3. *Point estimators, $\hat{\theta}$, that satisfy $E(\hat{\theta}) > \theta$ are said to be* **positively biased**.

We also want the variance of the distribution of the estimator, $V(\hat{\theta})$, to be as small as possible.

Definition 7.1.4. *The* **mean square error** *of a point estimator $\hat{\theta}$ is*

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

The mean square of an estimator $\hat{\theta}$, $MSE(\hat{\theta})$, is a function of both its variance and its bias. Indeed, if $B(\hat{\theta})$ denotes the bias of the estimator $\hat{\theta}$, then we have that

$$\begin{aligned} MSE(\hat{\theta}) &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= V[\hat{\theta}] + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= V[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 \\ &= V[\hat{\theta}] + [B(\hat{\theta})]^2 \end{aligned}$$

7.2.0 §Common Unbiased Point Estimators

First, we denote the variance of the sampling distribution of the estimator $\hat{\theta}$ by $\sigma^2_{\hat{\theta}}$, and the standard deviation $\sigma_{\hat{\theta}}$ of the sampling distribution of the estimator $\hat{\theta}$ is called the ***standard error*** of the estimator.

Then for independent samples, the following are point estimators of often desirable quantities:

Table 7.1: Expected values and standard errors of common point estimators

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Parameter, θ	Estimator/statistic $\hat{\theta}$
$E[X] = \mu$ the population mean	Sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, numeric data
$V[X] = \sigma^2$, population variance	Sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(n-1)}$, numeric data
$V[X] = \sigma^2$, population variance	Sample variance, $\frac{p(1-p)}{n}$, for categorical data
$E[X] = p$, population proportion	Sample proportion $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for categorical data
$E[X] = \mu$, population mean	Sample mode $\dot{X}$ - most frequent data
$E[X] = \mu$, population mean	Sample median $\widetilde{X}$ center of ordered data
Population min	$X_{(1)} = \min(X_1, ..., X_n)$
Population max	$X_{(n)} = \max(X_1, ..., X_n)$

The expected values and standard errors for $\overline{Y}$ and $\overline{Y}_1 - \overline{Y}_2$ are valid regardless of the distribution of the population(s) from which the sample(s) is (are) taken. Moreover, all four estimators possess probability distributions that are approximately normal for large samples. The central limit theorem guarantees this for $\overline{Y}$ and $\hat{p}$, and the others are justified by similar theorems.

Definition 7.2.1. *The following is an unbiased estimator for the sample variance:*

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

and the following is a biased estimator for the sample variance:

$$S'^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}$$

7.3.0 §Evaluating the Goodness of a Point Estimator

Definition 7.3.1. *The* **error of estimation** *ε is the distance between an estimator and its target parameter. That is, $\varepsilon = |\hat{\theta} - \theta|$.*

Since $\hat{\theta}$ is a random variable, so is ε. If we take b as a "probabilistic bound" on the error of estimation, then $P(\varepsilon < b) = P(|\hat{\theta} - \theta| < b)$ provides a measure of the goodness of a single estimate (if b can be regarded as small from some practical point of view).

To determine the exact b such that $P(\varepsilon < b) = p$ for some $0 < p < 1$ requires knowledge of the probability distribution or density function of $\hat{\theta}$. However, even if we don't know this we can find an approximate bound on ε by expressing b as a multiple of the standard error of $\hat{\theta}$. For example, if $k \geqslant 1$ and we let $b = k\sigma_{\hat{\theta}}$, then we know from Tchebysheff's theorem that ε will be less that $k\sigma_{\hat{\theta}}$ with probability at least $1 - 1/k^2$.

7.4.0 §Confidence Intervals

Definition 7.4.1. *An* **interval estimator** *is a rule specifying the method for using the sample measurements to calculate two numbers that form the endpoints of the interval. Ideally, the resulting interval will have two properties: First, it will contain the target parameter θ; second, it will be relatively narrow.*

Interval estimators are also referred to as **confidence intervals**.

Definition 7.4.2. *The upper and lower endpoints of a confidence interval are called the* **upper** *and* **lower confidence limits**, *respectively.*

Definition 7.4.3. *The probability that a random confidence interval will enclose the fixed quantity θ is called the* **confidence coefficient**.

Definition 7.4.4. *Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the random lower and upper confidence limits, respectively, for a parameter θ. Then if*

$$P(\hat{\theta}_L \leqslant \theta \leqslant \hat{\theta}_U) = 1 - \alpha$$

the probability $(1 - \alpha)$ is the **confidence coefficient**. *The resulting random interval defined by $[\hat{\theta}_L, \hat{\theta}_U]$ is called a* **two-sided confidence interval**.

Definition 7.4.5. *A* **one-sided confidence interval** *is an interval $[\hat{\theta}_L, \infty)$ or $(-\infty, \hat{\theta}_U]$, where*

$$P(\hat{\theta}_L \leqslant \theta) = 1 - \alpha$$

or

$$P(\theta \leqslant \hat{\theta}_U) = 1 - \alpha$$

respectively.

Definition 7.4.6. *The* **pivotal method** *depends on finding a pivotal quantity that possesses two characteristics:*

1. *It is a function of the sample measurements and the unknown parameter θ, where θ is the only unknown quanitty.*

2. *Its probability distribution does not depend on the parameter θ.*

If the probability distribution of the pivotal quantity is known, the following logic can be used to form the desired interval estimate: If Y is any random variable, $c > 0$ is a constant, and $P(a \leqslant Y \leqslant b) = p$ for some $p \in (0, 1)$, then certainly $P(ca \leqslant cY \leqslant cb) = p$. Similarly, $P(a + c \leqslant Y + c \leqslant b + c) = p$. Thus if we know the probability distribution of a pivotal quantity, we may be able to use operations like these to form the desired interval estimator.

7.5.0 §Large-Sample Confidence Intervals

Recall that the parameters $\mu, p, \mu_1 - \mu_2, p_1 - p_2$ have approximately normal sampling distributions with standard errors for large samples. Then for large samples if θ is one of these parameters, then

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

possesses approximately a standard normal distribution. Consequently, Z forms approximately a pivotal quantity.

When $\theta = \mu$ is the target parameter, then $\hat{\theta} = \overline{Y}$ and $\sigma^2_{\hat{\theta}} = \sigma^2/n$, where σ^2 is the population variance. If σ^2 is unknown, and n is large, then one can use the estimator s^2 to substitute for σ^2 in estimating the confidence interval. Similarly, if σ_1^2 and σ_2^2 are unknown and both n_1 and n_2 are large, s_1^2 and s_2^2 can be substituted for these values in the finding the confidence interval for $\theta = \mu_1 - \mu_2$.

When $\theta = p$ is the target parameter, then $\hat{\theta} = \hat{p}$ and $\sigma_{\hat{p}} = \sqrt{pq/n}$. Because p is the unknown target parameter, $\sigma_{\hat{p}}$ cannot be evaluated. If n is large and we substitute $\hat{p}$ for p and $\hat{q} = 1 - \hat{p}$ for q in the formula for $\sigma_{\hat{p}}$, then the resulting confidence interval will have approximately the desired confidence coefficient. For n_1 and n_2 large, similar statements hold when $\hat{p}_1$ and $\hat{p}_2$ are used to estimate p_1 and p_2 respectively in the formula for $\sigma^2_{\hat{p}_1 - \hat{p}_2}$.

Proposition 7.5.1. *Suppose σ is known and $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim norm(0,1)$, where X_i are i.i.d sample and either $X_i \sim norm(\mu, \sigma)$ or n is large enough for the central limit theorem to reasonably approximate $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$. Then the $(1-\alpha)100\%$ confidence interval for μ is*

$$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $P(Z \leqslant z_{\alpha/2}) = \alpha/2$ and $P(Z \geqslant z_{1-\alpha/2}) = \alpha/2, Z \sim norm(0,1)$.

Proposition 7.5.2. *If X_i are an i.i.d sample and either $X_i \sim norm(\mu, \sigma)$ or n is large enough for the CLT to reasonably approximate that $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$, and if σ is unknown, $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim T_{df=n-1}$ is the pivotal quantity for obtaining the following $(1-\alpha)100\%$ confidence interval for μ:*

$$\overline{X} \pm t_{\alpha/2, df=n-1} \frac{S}{\sqrt{n}}$$

Remark 7.5.1. Suppose $Y_1, ..., Y_n$ represent a sample normal population with μ and σ^2 unknown. Suppose n is small. Then

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim T_{df=n-1}$$

T serves as a pivotal quantity for μ. Then we obtain the confidence interval

$$\overline{Y} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

for μ.

Remark 7.5.2. Suppose we have two normal populations, one with mean μ_1 and variance σ_1^2, and the other with mean μ_2 and variance σ_2^2. Suppose $\sigma_1 = \sigma_2$. Then

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a pivotal quantity for $\mu_1 - \mu_2$. In this case $Z \sim norm(0,1)$.

Definition 7.5.1. *For two populations $Y_{1,i}$ and $Y_{2,j}$, $1 \leqslant i \leqslant n_1, 1 \leqslant j \leqslant n_2$, with common variance σ^2, the* **pooled estimator** S_p^2 *is*

$$S_P^2 = \frac{\sum_{i=1}^{n_1}(Y_{1,i} - \overline{Y}_1)^2 + \sum_{j=1}^{n_2}(Y_{2,j} - \overline{Y}_2)^2}{n_1 + n_2 + 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2}{n_1 + n_2 - 2}$$

Then

$$W = \frac{(n_1 + n_2 - 2)S_P^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

is the sum of two χ^2 distributed random variables with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom, respectively. Thus $W \sim \chi^2_{df=n_1+n_2-2}$. THen

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is $T_{df=n_1+n_2-2}$

Corollary 7.5.3. *The confidence interval for $\mu_1 - \mu_2$ above is*

$$(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2} S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Remark 7.5.3. Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are an i.i.d sample with $X_i \sim bernoulli(p)$ and n is large enough for the CLT to reasonably approximate that $\hat{p} \sim norm(\mu = p, \sigma = \sqrt{p(1-p)/n})$. Note if n is large enough for CLT to approximate $\hat{p} \approx p$, then

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim norm(0,1)$$

is a pivotal quantity for p, and the $(1 - \alpha)100\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

7.6.0 §Selecting the Sample Size

Definition 7.6.1. *The* **sampling procedure** *or* **experimental design** *affects the quantity of information per measurement. This together with the sample size n controls the total amount of relevant information in a sample.*

Example 7.6.1. An experimenter wishes to compare the effectiveness of two methods of training industrial employees to perform an assembly operation. The selected employees are to be divided into two groups of equal size, the first receiving training method 1 and the second reveivving training method 2. After training, each employee will perform the assembly operation, and the length of assembly time will be recorded. The experimenter expects the measurements for both groups to have a range of approximately 8 minutes. If the estimate of the difference in mean assembly times is to be correct to within 1 minute with probability 0.95, how many workers must be included in each training group?

The manufacturer specified $1 - \alpha = 0.95$, so $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$. We want

$$1.96\sigma_{(\overline{Y}_1 - \overline{Y}_2)} = 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1$$

Note we want $n_1 = n_2 = n$. The variability of each method of assembly is approximately the same, so $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and because the range is 8 minutes, we have $4\sigma \approx 8$, so $\sigma \approx 2$. Substituting in these values and solving for n we obtain $n \approx 30.73$, so each group should contain $n = 31$ workers.

7.7.0 §Small-Sample Confidence Intervals for Means and Mean Differences

The confidence intervals for a population mean μ that we discuss in this section are based on the assumption that the experimenter's sample has been randomly selected from a normal population.

We assume that $Y_1, Y_2, ..., Y_n$ represent a random sample selected from a normal population, and we let $\overline{Y}$ and S^2 represent the sample mean and sample variance, respectively. We wish to construct a confidence interval for the population mean when $V[Y_i] = \sigma^2$ is unknown and the sample size is too small to permit use of large-sample techniques from the previous section. Under these assumptions recall that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

has a t distribution with $(n - 1)$ degrees of freedom. The quantity T serves as the pivotal quantity that we will use to form a confidence interval for μ. We can find values t_α such that

$$P(-t_{\alpha/2} \leqslant T \leqslant t_{\alpha/2}) = 1 - \alpha$$

Equivalently, we have that

$$P(\overline{Y} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \overline{Y} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

providing the desired confidence interval for μ. Using similar methods we can find one sided confidence intervals.

Suppose we want to compare the means of two normal populations, one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2. If the samples are independent, the confidence intervals for $\mu_1 - \mu_2$ based on a t-distributed random variable can be constructed if we assume that the two populations have a common but unknown variance, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

If $\overline{Y}_1$ and $\overline{Y}_2$ are the respective sample means obtained from independent random samples from normal populations, the large-sample confidence interval for $(\mu_1 - \mu_2)$ is developed using

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where Z has a standard normal distribution since the sampled populations are assumed to be normally distruted.

Let $Y_{11}, Y_{12}, ..., Y_{1n_1}$ denote the random sample of size n_1 from the first population and let $Y_{21}, Y_{22}, ..., Y_{2n_2}$ denote an independent random sample of size n_2 from the second population. The unbiased estimator of the common variance σ^2 is obtained by pooling the sample data to obtain the ***pooled estimator*** S_p^2:

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

If $n_1 = n_2$, S_p^2 is the average of S_1^2 and S_2^2. Otherwise, it is a weighted average of the sample variances, with larger weight given to the sample variance associated with the larger sample size. Further,

$$W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2}{\sigma^2} + \frac{\sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)^2}{\sigma^2}$$

is the sum of two independent χ^2-distributed random variables with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom, respectively. Thus W has a χ^2-distribution with $\nu = (n_1 - 1) + (n_2 - 1)$ degrees of freedom. We can use the χ^2-distributed variable W and the independent standard normal quantity Z defined in the previous paragraph to form a pivotal quantity:

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

a quantity that by construction has a t distribution with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom. It follows that the confidence interval for $(\mu_1 - \mu_2)$ has the form

$$\left((\overline{Y}_1 - \overline{Y}_2) - t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\overline{Y}_1 - \overline{Y}_2) + t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

where $t_{\alpha/2}$ is determined from the t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Definition 7.7.1. *The method of Small-Sample Confidence Intervals for Means of Normal Distributions with Unknown Variance(s) is summarized as follows:*

Parameter	Confidence Interval	Degrees of Freedom
μ	$\left(\overline{Y} - t_{\alpha/2}S/\sqrt{n}, \overline{Y} + t_{\alpha/2}S/\sqrt{n}\right)$	$\nu = n-1$
$\mu_1 - \mu_2$	$\left((\overline{Y}_1 - \overline{Y}_2) - t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\overline{Y}_1 - \overline{Y}_2) + t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$	$\nu = n_1 + n_2 - 2$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

and we require that $\sigma_1^2 = \sigma_2^2$.

When sample sizes are large the confidence intervals produced with these methods agree closely with those of the previous section. There is considerable empirical evidence that these intervals maintain their nominal confidence coefficient as long as the populations sampled have roughly mound-shaped distributions. If $n_1 \approx n_2$, the intervals for $\mu_1 - \mu_2$ also maintain their nominal confidence coefficients as long as the population variances are roughly equal.

It is crucial that we have ***independence of the samples*** in order to use the confidence intervals developed in this section to compare two population means.

7.8.0 §Large Normally Distributed Data and Two Sample Confidence Intervals

Remark 7.8.1. Let X, Y be independent χ^2 distributed random variables with ν and δ degrees of freedom respectively. Then

$$F = \frac{X/\nu}{Y/\delta}$$

is F distributed with ν numerator degrees of freedom and δ denominator degrees of freedom.

$$f_F(x) = \frac{\Gamma\left(\frac{\nu+\delta}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\delta}{2}\right)}\left(\frac{\nu}{\delta}\right)^{\frac{\nu}{2}} x^{\frac{\nu}{2}-1}\left(1+\frac{\nu}{\delta}t\right)^{-\frac{\nu+\delta}{2}}, x > 0$$

It can be shown $E[F] = \frac{\delta}{\delta-2}$ for $\delta > 2$, $V[F] = \frac{2\delta(\delta+\nu-2)}{\nu(\delta-2)^2(\delta-4)}$ for $\delta > 4$. Then

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

Remark 7.8.2. Suppose $S_j^2 = \frac{\sum_{i=1}^{n_j}(X_{i,j}-\overline{X}_j)^2}{n_j-1}$ where $X_{i,j}$ is an i.i.d sample from group $j \in \{1,2\}$, and either $X_{i,j} \sim norm(\mu_j, \sigma_j)$ or n is large enough for the CLT to reasonably approximate $\overline{X} \sim norm(\mu_j, \sigma_j/\sqrt{n_j})$. Then

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

is a pivotal quantity for σ_1^2/σ_2^2, and the $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 is

$$F_{\alpha/2,n_2-1,n_1-1}S_1^2/S_2^2 \leqslant \sigma_1^2/\sigma_2^2 \leqslant F_{1-\alpha/2,n_2-1,n_1-1}S_1^2/S_2^2$$

When our interval contains 1, there is no significant evidence to say σ_1^2 does not equal σ_2^2

When our interval does NOT contain 1, we have significant evidence to say $\sigma_1^2 \neq \sigma_2^2$.

Note 7.8.3. $F_{1-\alpha/2,n_1-1,n_2-1} = \frac{1}{F_{\alpha/2,n_2-1,n_1-1}}$

Proposition 7.8.1. *Let $X_{1,i}$, $1 \leqslant i \leqslant n_1$, be a random sample from a population with mean μ_1 and variance σ^2. Let $X_{2,j}$, $1 \leqslant j \leqslant n_2$, be a random sample from a population with mean μ_2 and variance σ^2, where both populations have the same distribution which is either normal, or n is high enough for it's mean to be approximately normally distributed by the CLT. A $(1-\alpha)100\%$ confidence interval estimate for the difference $\mu_1 - \mu_2$ between population means is*

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,df=n_1+n_2-2} \sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Even if $\sigma_1^2 = \sigma_2^2$, if the ratio of the variances has a 95% confidence interval containing 1, then this method can still be used.

Proposition 7.8.2. *Let $X_{1,i}$, $1 \leqslant i \leqslant n_1$, be a random sample from a population with mean μ_1 and variance σ_1^2. Let $X_{2,j}$, $1 \leqslant j \leqslant n_2$, be a random sample from a population with mean μ_2 and variance σ_2^2, where both populations have the same distribution which is either normal, or n is high enough for it's mean to be approximately normally distributed by the CLT. A $(1-\alpha)100\%$ confidence interval estimate for the difference between the two population means $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$ is*

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,df} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$

where

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Proposition 7.8.3. *Let $\hat{p}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{j,i}$, $j = 1,2$, where $X_{j,i}$ is an i.i.d sample with $X_{j,i} \sim$ bernoulli(p_j) and n_j is large enough for the CLT to reasonably approximate $\hat{p}_j \sim norm(\mu = p_j, \sigma = \sqrt{p_j(1-p_j)/n_j})$ Note n_j is large enough for the CLT to approximate $\hat{p}_j \approx p_j$. THen using the pivotal quantity $\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim norm(0,1)$ so the $(1-\alpha)100\%$ confidence interval of $p_1 - p_2$ is*

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where by CLT $\hat{p}_1 - \hat{p}_2 \sim norm(\mu = p_1 - p_2, \sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

7.9.0 §Confidence Intervals for Variance

First we require a pivotal quantity. Assume that we have a random sample $Y_1, Y_2, ..., Y_n$ from a normal distribution with mean μ and variance σ^2, both unknown. We know that

$$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

has a χ^2-distribution with $(n-1)$ degrees of freedom. We can then proceed by the pivotal method to find χ^2_L and χ^2_U such that

$$P\left[\chi^2_L \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2_U\right] = 1 - \alpha$$

for any confidence coefficient $(1-\alpha)$. Since χ^2 does not have a symmetric density function, we have some freedom in choosing χ^2_L and χ^2_U. We would like to find the shortest interval that includes σ^2 with probability $(1-\alpha)$. Generally this requires trial and error, so we will compromise with points that cut off equal tail areas:

$$P\left[\chi^2_{1-(\alpha/2)} \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2_{\alpha/2}\right] = 1 - \alpha$$

and a reordering of the inequality in the probability statement gives:

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right] = 1 - \alpha$$

Proposition 7.9.1. *It follows that the* $100(1-\alpha)\%$ *confidence interval for* σ^2 *is*

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right)$$

If the sampled population is not normally distributed, the intervals for σ^2 presented in this section can have confidence coefficients that differ markedly from the nominal level if the sampled population is not normally distributed.

Remark 7.9.1. Let $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ where X_i is an i.i.d sample and either $X_i \sim norm(\mu, \sigma)$ or n is large enough for the CLT to reasonably approximate $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$. Then with pivotal quantity

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{df=n-1}$$

we can construct the $(1-\alpha)100\%$ confidence interval for σ, σ^2:

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,df=n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2_{\alpha/2,df=n-1}} \iff \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,df=n-1}}} \leqslant \sigma \leqslant \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,df=n-1}}}$$

Chapter 8

Properties of Point Estimators and Methods of Estimation

8.1.0 §Relative Efficiency

Definition 8.1.1. *Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ, with variances $V[\hat{\theta}_1]$ and $V[\hat{\theta}_2]$, respectively, then the* **efficiency** *of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted $eff(\hat{\theta}_1, \hat{\theta}_2)$, is defined to be the ratio*

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

If $eff(\hat{\theta}_1, \hat{\theta}_2) > 1$, then $V(\hat{\theta}_2) > V(\hat{\theta}_1)$ and $\hat{\theta}_1$ is a better unbiased estimator than $\hat{\theta}_2$.

8.2.0 §Consistency

Definition 8.2.1. *The estimator $\hat{\theta}_n$ is said to be a* **consistent estimator of** θ *if, for any positive number ε,*

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

or equivalently,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

Theorem 8.2.1. *An unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator of θ if*

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0$$

Proof. If Y is a random variable with $E[Y] = \mu$ and $V[Y] = \sigma^2 < \infty$, and if k is a nonnegative constant, Tchebysheff's Theorem implies that

$$P(|Y - \mu| > k\sigma) \leqslant \frac{1}{k^2} \iff P(|Y - \mu| \leqslant k\sigma) \geqslant 1 - \frac{1}{k^2}$$

Because $\hat{\theta}_n$ is an unbiased estimator for θ, $E[\hat{\theta}_n] = \theta$. Then if $\sigma_{\hat{\theta}_n} = \sqrt{V(\hat{\theta}_n)}$, then

$$P(|\hat{\theta}_n - \theta| > k\sigma_{\hat{\theta}_n}) \leqslant \frac{1}{k^2}$$

Let n be fixed. For any $\varepsilon > 0$, let $k = \frac{\varepsilon}{\sigma_{\hat{\theta}_n}} > 0$. Then

$$P(|\hat{\theta}_n - \theta| > \varepsilon) = P(|\hat{\theta}_n - \theta| > \frac{\varepsilon}{\sigma_{\hat{\theta}_n}}\sigma_{\hat{\theta}_n}) \leqslant \frac{\sigma^2_{\hat{\theta}_n}}{\varepsilon^2} = \frac{V(\hat{\theta}_n)}{\varepsilon^2}$$

If $\lim\limits_{n\to\infty} V(\hat{\theta}_n) = 0$, then

$$0 = \lim_{n\to\infty} 0 \leqslant \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) \leqslant \lim_{n\to\infty} \frac{V(\hat{\theta}_n)}{\varepsilon^2} = 0$$

Thus $\hat{\theta}_n$ is a consistent estimator of θ. ■

Theorem 8.2.2. *Suppose $\hat{\theta}_n$ converges in probability to θ and $\hat{\theta}'_n$ converges in probability to θ':*

- *$\hat{\theta}_n + \hat{\theta}'_n$ converges in probability to $\theta + \theta'$*
- *$\hat{\theta}_n \cdot \hat{\theta}'_n$ converges in probability to $\theta \cdot \theta'$*
- *If $\theta' \neq 0$, $\hat{\theta}_n/\hat{\theta}'_n$ converges in probability to θ/θ'*
- *If $g : \mathbb{R} \to \mathbb{R}$ is a real valued function that is continuous at θ, then $g(\hat{\theta}_n)$ converges in probability to $g(\theta)$.*

Theorem 8.2.3. *Suppose that U_n has a distribution function that converges to a standard normal distribution function as $n \to \infty$. If W_n converges in probability to 1, then the distribution function U_n/W_n converges to a standard normal distribution function.*

8.3.0 §The Method of Moments

Recall 8.3.1. Recall the kth moment of a random variable taken about the origin is

$$\mu'_k = E(Y^k)$$

The corresponding kth sample moment is the average

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

Process 8.3.1. *Method of Moments Choose as estimates those values of the parameters that are solutions of the equations $\mu'_k = m'_k$ for $k = 1, 2, ..., t$, where t is the number of parameters to be estimated.*

Definition 8.3.2.

1. $\mu'_k = E[X^k]$ *is the kth (theoretical) moment of the distribution about the origin, for* $k = 1, 2, ...$

2. $\mu^*_k = E[(X - \mu)^k]$ *is the kth (theoretical) moment of the distribution about the mean, for* $k = 1, 2, ...$

3. $M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ *is the kth sample moment about the origin, for* $k = 1, 2, ...$

4. $M^*_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$ *is the kth sample moment about the mean, for* $k = 1, 2, ...$

Process 8.3.3. *Method Suppose you have k parameters,* $\theta_1, ..., \theta_k$ *you want to estimate. Then*

1. *Equate the lth sample moment aout the origin,* $M'_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l$ *to the lth theoretical moment* $E[X^l]$ *for all* $1 \leqslant l \leqslant k$ *moments*

2. *Solve for the parameters*

Another form of the method is as follows:

1. *Equate the first sample moment about the origin* $M'_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ *to the first theoretical moment* $E[X]$

2. *Equate the second sample moment about the mean* $M^*_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *to the second theoretical moment about the mean* $E[(X - \mu)^2] = V[X]$

3. *Continue equating sample moments about the mean* M^*_K *with the corresponding theoretical moments about the mean* $E[(X - \mu)^k]$, $k = 3, 4, ...$ *until you have as many equations as you have paramaters*

4. *Solve for the parameters*

8.4.0 §The Method of Maximum Likelihood

Process 8.4.1. *Method Suppose that the likelihood function depends on k parameters* $\theta_1, ..., \theta_k$. *Choose as estimates those values of the parameters that maximize the likelihood* $L(y_1, .., y_n|\theta_1, ..., \theta_k)$. *This gives* **maximum-likelihood estimators** *or* **MLEs**

Definition 8.4.2. *The range of possible values for a parameter* θ *is called the* **parameter space**, Ω.

Definition 8.4.3. *THe function of* $X_1, ..., X_n$ *that is the statistic* $u(X_1, ..., X_n)$, *used to estimate* θ *is called a* **point estimator** *of* θ.

Remark 8.4.1. Suppose we have a random sample $X_1, ..., X_n$ whose assumed probability distribution depends on some unknown parameter θ. Our primary goal is to find a point estimator $u(X_1, ..., X_n)$ such that $u(x_1, ..., x_n)$ is a "good" estimate of θ, where $x_1, ..., x_n$ are the observed values of the random sample.

Note 8.4.2. We aim to get an estimator of the unknown parameter by finding the value of θ that maximizes the probabiliy/likelihood of getting the data we observed.

Suppose $X_1, ..., X_n$ is a random sample with pdf $f(x_i|\theta)$ for each X_i. Then the joint pdf of $X_1, ..., X_n$ is

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

We wish to find the "value" of θ which maximizes $L(\theta)$.

Note 8.4.3. Because the log function is a monotonically increasing function, the maximum of $L(\theta)$ with respect to θ occurs for the same θ as the maximum of $\log(L(\theta))$.

Definition 8.4.4. *Let $X_1, ..., X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, ..., \theta_m$ with pdf $f(x_i|\theta_1, ..., \theta_m)$. Suppose that $(\theta_1, ..., \theta_m)$ is restricted to a given parameter space Ω. Then:*

1. *When regarded as a function of $\theta_1, ..., \theta_m$, the joint pdf of $X_1, ..., X_n$*

$$L(\theta_1, ..., \theta_m) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_m)$$

   $(\theta_1, ..., \theta_m) \in \Omega$, the **likelihood function**

2. *If $[u_1(x_1, ..., x_n), ..., u_m(x_1, ..., x_n)]$ is the m-tuple that maximizes the likelihood function, then:*

$$\hat{\theta}_i = u_i(x_1, ..., x_n)$$

   is the **maximum-likelihood estimator** *for θ_i, $1 \leqslant i \leqslant m$.*

3. *The corresponding observed values of the statistics, $u_i(x_1, ..., x_n)$, are called the* **maximum-likelihood estimates** *of θ_i, $1 \leqslant i \leqslant m$.*

Chapter 9

Hypothesis Testing

Support for one theory is obtained by showing lack of support for its converse-like a proof by contradiction.

9.0.1. *Elements of a Statistical Test*

- *The null hypothesis H_0*
- *The alternative hypothesis H_a (converse of the null)*
  - *Hypothesis to be accepted in case H_0 is rejected*
- *Test statistic*
  - *Function of the sample measurements*
- *Rejection region*
  - *Specifies the values of the test statistic for which the null hypothesis is to be rejected in favour of the alternative hypothesis. i.e. Sample falls in RR, reject H_0 and accept H_a, sample doesn't fall in RR accept H_0*

Definition 9.0.2. *A* **Type I error** *is made if H_0 is rejected when H_0 is true. The* **probability of a Type I error** *is denoted by α. The value of $|alpha$ is called the* **level of significance** *of the test.*

A **Type II error** *is made if H_0 is accepted when H_a is true. The* **probability of a type II error** *is denoted by β.*

Remark 9.0.1. Hypothesis testing is a formal procedure that enables us to choose between two hypotheses when we are uncertain about our measurements.

Definition 9.0.3. *The* **null hypothesis**, *H_0, is the conservative status quo statement about a population parameter. Often represents no change, no effect, or no difference, in the context of researching new ideas.*

Definition 9.0.4. *The* **alternative hypothesis**, *H_a, is the research hypothesis. It is usually a statement about the value of a parameter that we hope to demonstrate is true.*

Process 9.0.5.

1. *Start with a pair of hypotheses: In a formal hypothesis test, hypotheses are always statements about population parameters, θ_0. We use statistics as evidence to justify our hypothesis. Hypotheses come in mutually exclusive pairs:*
   - ***Simple:***
     - *is, $=$, versus is not, $\neq$*
   - ***Composite:***
     - *Is or more, $\geqslant$, versus less $<$*
     - *Is or less, $\leqslant$, versus more $>$*

   Note that the equality is reserved for null hypotheses.
2. *Forming the hypothesis*

Remark 9.0.2. The null hypothesis is by default assumed true. This means when no substantial evidence is provided, the null is "not dethroned" - this does *not* guarantee the truthfulness of the null, rather we just can't discount it at the moment.

Definition 9.0.6. *If W is a test statistic, the* **p-value**, *or* **attained significance level**, *is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.*

The smaller the p-value the more compelling the evidence that the null hypothesis should be rejected. It is the smallest value of α for which the null hypothesis can be rejected.

Remark 9.0.3. If our RR for a null H_0 and test statistic W is $\{w \leqslant k\}$, the p-value associated with an observed value of w_0 of W is

$$p - value = P(W \leqslant w_0 | H_0 \text{ is true})$$

(i.e. $H_0 : W > k$, and $H_a : W \leqslant k$). If the RR is $\{w \geqslant k\}$, then the p-value is

$$p - value = P(W \geqslant w_0 | H_0 \text{ is true})$$

If H_0 is of the form $\theta = \theta_0$ and $H_a : \theta \neq \theta_0$, then the $p - value$ for rejecting H_0 given w_0 is

$$p - value = P(|W| \geqslant |w_0||H_0 \text{ is true})$$

Remark 9.0.4. If H_0 is rejected for a small value of α (or for a small $p - value$) this occurrence does not imply that the null hypothesis is wrong by a large amount. It does mean that the null hypothesis can be rejected based on a predure that incorrectly rejects the null hypothesis with a small probability.

Remark 9.0.5. If we consider $H_a : \theta > \theta_0$, $\alpha = \max_{\theta \leqslant \theta_0} P(\text{test statistic in RR})$ typically occurs when $\theta = \theta_0$. Similarly, if $H_a : \theta < \theta_0$, $\alpha = \max_{\theta \geqslant \theta_0} P(\text{test statistic in RR})$ typically occurs when $\theta = \theta_0$.

Note 9.0.6. The null must always contain the '$=$'.

Remark 9.0.7. If we do not have sufficient evidence against the null, we say we "fail to reject the null"

Remark 9.0.8. The ***level of significance*** α represents an amount of allowable error/tolerance to deviation. We aggregate the info from the sample into a test statistic which is converted into a p-value that we can compare to the significance level, α. p-value refers to the probability of repeating the sampling process and comparing the theoretical results to what did happen:

$$p - value = P(f_X(x)[(\neq, >, <)\text{sign in the alternative hypothesis}]X|H_0 \text{ is true})$$

where X is the test statistic.

If $p - value \geqslant \alpha$, we conclude the null hypothesis is supported we fail to reject H_0. Implies if H_0 is true we got a typical response. If $p - value < \alpha$ we conclude the null hypothesis is not supported, we reject H_0. Implies if H_0 is true something rare has occurred.

Often we take $\alpha = 0.05$.

Definition 9.0.7. *A* **Type I error** *is when we reject H_0 when it is actually true.*

Definition 9.0.8. *A* **Type II error** *is when we fail to reject H_0 when it is actually false.*

Remark 9.0.9. Conclusion Based on Truth H_0 true:

- RH_0,

$$\begin{aligned} \alpha &= \text{significance level} \\ &= P(RH_0|H_0\ true) \\ &= P(Type\ I\ error) \end{aligned}$$

- FRH_0,

$$\begin{aligned} 1 - \alpha &= \text{confidence level} \\ &= P(FRH_0|H_0\ true) \\ &= P(no\ error) \end{aligned}$$

H_0 is false:

- RH_0,

$$\begin{aligned} 1 - \beta &= Power \\ &= P(RH_0|H_0\ false) \\ &= P(no\ error) \end{aligned}$$

- FRH_0,

$$\begin{aligned}\beta &= P(FRH_0|H_0\, false)\\ &= P(Type\ II\ error)\end{aligned}$$

Remark 9.0.10. We are free to choose any value for α. But, as α gets small, $1-\beta$ gets smaller too. Although α and β are inversely related, $1-\alpha \neq \beta$. We can lower both α and β simultaneously if we increase the sample size n. Bigger samples should reduce the variance of consistent estimators and decrease rick of error.

Usually we set α, then from α derive the corresponding decision rule for the rejection region.

9.1.0 §Common Large-Sample Tests

Let θ be a parameter we are using hypothesis testing on, and θ_0 is a specific value:

If we are testing $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ for test statistic $\hat{\theta}$ and rejection region $RR : \{\hat{\theta} > k\}$ for some k determined by α. If we are testing $H_0 : \theta = \theta_0$ against $H_a : \theta < \theta_0$ our rejection region is $RR : \{\hat{\theta} < k\}$ for some k determined by α. In testing $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_a$, we have rejection region $RR : \{\hat{\theta} < k_1 \cup k_2 < \hat{\theta}\}$ where both cut off $\alpha/2$ in probability.

Summary 9.1.1.

- $H_0 : \theta = \theta_0$
-
$$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative)}\\ \theta < \theta_0 & \text{(lower-tail alternative)}\\ \theta \neq \theta_0 & \text{(two-tailed alternative)}\end{cases}$$
- Test statistic: $Z = \frac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}$
- Rejection Region:
$$RR : \begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR)}\\ \{z < -z_\alpha\} & \text{(lower-tail RR)}\\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR)}\end{cases}$$

For the test $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_a$, we can calculate β only for specific values for θ in H_a. Suppose we fix $\theta_a > \theta_0$ and our RR is of the form

$$RR = \{\hat{\theta} : \hat{\theta} > k\}$$

Then

$$\beta = P(\hat{\theta} \leqslant k|\theta = \theta_a) = P\left(\frac{\hat{\theta}-\theta_a}{\sigma_{\hat{\theta}}} \leqslant \frac{k-\theta_a}{\sigma_{\hat{\theta}}}\middle|\theta = \theta_a\right)$$

If $\theta = \theta_a$, $(\hat{\theta}-\theta_a)/\sigma_{\hat{\theta}} \sim norm(0,1)$. Consequently, β can be determined approximately by finding the area under a standard normal.

Claim 9.1.1. *For desired α, β, the needed sample size for an upper-tail α-level test is*

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_a - \mu_0)^2}$$

9.2.0 §Common Confidence Intervals

Remark 9.2.1.

- $(1-\alpha)100\%$ confidence interval for μ when σ is known, we use pivotal quantity: $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim norm(0,1)$ with either $X_i \sim norm(\mu,\sigma)$, or n is large enough so CLT approximates $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$.
- $(1-\alpha)100\%$ confidence interval for μ when σ is unknown, we use pivotal quantity: $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim T(df = n-1)$ with either $X_i \sim norm(\mu,\sigma)$, or n is large enough so CLT approximates $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$.
- $(1-\alpha)100\%$ confidence interval for p when p is unknown, we use pivotal quantity: $Z = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim norm(0,1)$ with either $X_i \sim Bern(p)$ and n is large enough so CLT approximates $\hat{p} \sim norm(\mu = p, \sigma = \sqrt{\frac{p(1-p)}{n}})$.
- $(1-\alpha)100\%$ confidence interval for σ, we use pivotal quantity: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(df = n-1)$ with either $X_i \sim norm(\mu,\sigma)$, or n is large enough so CLT approximates $\overline{X} \sim norm(\mu, \sigma/\sqrt{n})$.

Remark 9.2.2. For hypothesis testing, $H_0 : \mu = \mu_0$ (for estimating μ), $H_0 : p = p_0$ (when estimating p), and $H_0 : \sigma = \sigma_0$ (when estimating σ)

Remark 9.2.3.

- $Z_{calc} = \frac{\overline{X}_0-\mu_0}{\sigma/\sqrt{n}} \sim norm(0,1)$
  - $H_a : \mu < \mu_0$:

    $$p - val = P(\overline{X} \leqslant \overline{X}_0 | \mu = \mu_0) = P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \leqslant \frac{\overline{X}_0-\mu_0}{\sigma/\sqrt{n}}\right) = P(Z \leqslant Z_{calc})$$

  - $H_a : \mu > \mu_0$:

    $$p - val = P(Z \geqslant Z_{calc})$$

  - $H_a : \mu \neq \mu_0$:

    $$p - val = \begin{cases} 2P(Z \leqslant Z_{calc}), & if\ Z_{calc} < 0 \\ 2P(Z \geqslant Z_{calc}), & if\ Z_{calc} > 0 \end{cases}$$

- $T_{calc} = \frac{\overline{X}_0-\mu_0}{S/\sqrt{n}} \sim T(df = n-1)$

- $H_a : \mu < \mu_0$:
$$p - val = P(T \leqslant T_{calc})$$
- $H_a : \mu > \mu_0$:
$$p - val = P(T \geqslant T_{calc})$$
- $H_a : \mu \neq \mu_0$:
$$p - val = \begin{cases} 2P(T \leqslant T_{calc}), & if\ T_{calc} < 0 \\ 2P(T \geqslant T_{calc}), & if\ T_{calc} > 0 \end{cases}$$

- $Z_{calc} = \frac{\hat{p}_0 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim norm(0, 1)$
  - $H_a : p < p_0$:
$$p - val = P(Z \leqslant Z_{calc})$$
  - $H_a : p > p_0$:
$$p - val = P(Z \geqslant Z_{calc})$$
  - $H_a : p \neq p_0$:
$$p - val = \begin{cases} 2P(Z \leqslant Z_{calc}), & if\ Z_{calc} < 0 \\ 2P(Z \geqslant Z_{calc}), & if\ Z_{calc} > 0 \end{cases}$$

Recall 9.2.4. $(1 - \alpha)100\%$ confidence interval for σ_1^2/σ_2^2, pivotal quantity: $\frac{S_2^2\sigma_1^2}{S_1^2\sigma_2^2} \sim F_{n_2-1,n_1-1}$, and either $X_{i,j} \sim norm(\mu_j, \sigma_j)$ or n is large enough for CLT to approximate $\overline{X} \sim norm(\mu_j, \sigma_j/\sqrt{n_j})$

$$\frac{S_1^2}{F_{1-\alpha/2,n_1-1,n_2-1}S_2^2} \leqslant \frac{\sigma_1^2}{\sigma_2^2} \leqslant \frac{S_1^2}{F_{\alpha/2,n_1-1,n_2-1}S_2^2}$$

If $H_0 : \sigma_1^2 = \sigma_2^2$, so under the assumption of H_0 $\frac{S_2^2\sigma_1^2}{S_1^2\sigma_2^2} \sim F_{n_2-1,n_1-1}$ and

$$F_{calc} = \frac{S_1^2}{S_2^2}$$

- $H_a : \sigma_1^2 < \sigma_2^2$:
$$p - val = P(F_{n_1-1,n_2-1} \leqslant S_1^2/S_2^2)$$
- $H_a : \sigma_1^2 > \sigma_2^2$:
$$p - val = P(F_{n_1-1,n_2-1} \geqslant S_1^2/S_2^2)$$
- $H_a : \sigma_1^2 \neq \sigma_2^2$:
$$p - val = 2\min(P(F_{n_1-1,n_2-1} \leqslant S_1^2/S_2^2), P(F_{n_1-1,n_2-1} \geqslant S_1^2/S_2^2))$$

Remark 9.2.5. Test Statistics for Differences

- $\mu_1 - \mu_2$, when $\sigma_1^2 = \sigma_2^2 = \sigma^2$
$$T_{calc,df=n_1+n_2-2} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$

$$T_{calc,df} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

- $p_1 - p_2$ when $p_1 - p_2 = 0$ is the null:

$$Z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim norm(0,1)$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + x_2}$$

- $p_1 - p_2$ when $p_1 - p_2 = d$

$$Z_{calc} = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p - 2)}{\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}} \sim norm(0,1)$$

$$\hat{p}_i = \frac{x_i}{n_i}$$

9.3.0 §The Power of a Statistical Test

Definition 9.3.1. *A* **Type I error** *occurs if we reject the null hypothesis H_0 (in favour of the alternate hypothesis H_a) when the null hypothesis H_0 is true. We denote it by $\alpha = P(RH_0|H_0\ true)$.*

Definition 9.3.2. *A* **Type II error** *occurs if we fail reject the null hypothesis H_0 when the alternate hypothesis H_a is true. We denote it by $\beta = P(FRH_0|H_0\ false)$.*

Definition 9.3.3. **Power** *is defined as*

$$1 - \beta = power = P(RH_0|H_0\ false)$$

The power of a test measures the test's ability to detect that the null hypothesis is false.

9.4.0 §Power of Tests and the Neyman-Pearson Lemma

A ***power curve*** is a graph of $power(\theta)$.

Definition 9.4.1. *If a random sample is taken from a distribution with parameter θ, a hypothesis is said to be a* **simple hypothesis** *if that hypothesis* **uniquely specifies** *the distribution of the population from which the sample is taken. Any hypothesis that is not simple is called a composite hypothesis.*

Theorem 5.

The Neyman-Pearson Lemma Suppose that we wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ versus the simple alternative hypothesis $H_a : \theta = \theta_a$, based on a random sample $Y_1, ..., Y_n$ from a distribution with parameter θ. Let $L(\theta)$ denote the likelihood of the sample when the value of the parameter is θ. Then, for a given α, the test that maximizes the power at θ_a has a rejection region determined by

$$\frac{L(\theta_0)}{L(\theta_a)} < k$$

for k to be determined by α. SUch a test is called a **most powerful α-level test for H_0 versus H_a**

Remark 9.4.1. When a test obtained by this theorem maximizes the power for every value of θ greater (resp. lesser) than θ_0, it is said to be a ***uniformly most powerful test*** for $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ (resp. $H_a : \theta < \theta_0$)

Remark 9.4.2. THe goodness of a test is measured by how small α and β are.

9.5.0 §Likelihood Ratio Test

Suppose a random sample is selected from a distribution and that the likelihood function $L(y_1, ..., y_n|\theta_1, ..., \theta_k)$ is a function of k parameters, $\theta_1, ..., \theta_k$. Define $\Theta = (\theta_1, ..., \theta_k)$ and write $L(\Theta)$. Parameters not involved in our testing that are unknown are called ***nuisance parameters***. Suppose the null hypothesis says $H_0 : \Theta \in \Omega_0$, and $H_a : \Theta \in \Omega_a$ such that $\Omega_0 \cap \Omega_a = \varnothing$. Let $L(\hat{\Omega}_0)$ denote the supremum of the likelihood function for all $\Theta \in \Omega_0$. Let $\Omega = \Omega_0 \cup \Omega_a$, and similarly we have $L(\hat{\Omega})$ as the supremum over Ω. If $L(\hat{\Omega}_0) = L(\hat{\Omega})$, then the best explanation for the observed data can be found in Ω_0. Otherwise, if $L(\hat{\Omega}_0) < L(\hat{\Omega})$, the best explanation for the data can be found in Ω_a, and we should consider rejecting H_0 in favour of H_a.

Theorem 6.

Likelihood Ratio Test (text) Define λ by

$$s\lambda = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\max_{\Theta \in \Omega_0} L(\Theta)}{\max_{\Theta \in \Omega} L(\Theta)}$$

A likelihood ratio test of $H_0 : \Theta \in \Omega_0$ versus $H_a : \Theta \in \Omega_a$ employs λ as a test statistic, and the rejection region is determined by $\lambda \leqslant k$.

Theorem 9.5.1. *Let $Y_1, ..., Y_n$ have joint likelihood function $L(\Theta)$. Let r_0 denote the number of free parameters that are specified by $H_0 : \Theta \in \Omega_0$ and let r denote the number of free parameters specified by the statement $\Theta \in \Omega$. Then, for large n, $-2\ln(\lambda)$ has approximately a χ^2 distribution with $r_0 - r$ degrees of freedom.*

This implies that for $RR : \{\lambda < k\} = \{-2\ln(\lambda) > -2\ln(k) = k^*\}$, $k^* \approx \chi^2_\alpha$ if we desire an α level test.

Notation 9.5.1. We'll assume that the probability density (or mass) function of X is $f(x|\theta)$, where θ represents one or more unknown parameters (a vector). Then:

1. Let Ω denote the total possible parameter space of θ, that is the set of all possible values of θ as specified in totality in the null and alternative hypothesis.
2. Let $H_0 : \theta \in \omega$ denote the null hypothesis, where $\omega \subseteq \Omega$
3. Let $H_a : \theta \in \omega'$ denote the alternative hypothesis where $\omega' = \Omega\backslash\omega$ (the complement)

Definition 9.5.1. *Let*

1. *$L(\hat{\omega})$ denote the maximum of the likelihood function with respect to θ when $\theta \in \omega$.*
2. *$L(\hat{\Omega})$ denote the maximum of the likelihood function with respect to θ when $\theta \in \Omega$.*

Then the **likelihood ratio** *is the quotient:*

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$

And to test the null hypothesis $H_0 : \theta \in \omega$ against the alternative hypothesis $H_a : \theta \in \omega'$, the critical region for the likelihood ratio test is the set for which:

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \leqslant k$$

where $0 < \lambda < 1 \implies 0 < k < 1$, and k is selected so that the test has a desired significance level α.

Chapter 10

Linear Models and Estimation by Least Squares

Definition 10.0.1. *We call a random variable Y a* **dependent variable** *if it has a mean that is a function of one or more non-random variables $x_1, ..., x_k$, called* **independent variables**

Definition 10.0.2. *In* **deterministic models***, the output of the model is fully determined by teh parameter values and the initial conditions. In particular, a deterministic model of a random variable Y is*

$$Y = f(x_1, ..., x_n)$$

for some $f : \mathbb{R}^n \to X$, where each x_i are parameter values/independent non-random variables.

Definition 10.0.3. *A* **probabilistic model** *for a random variable Y is*

$$Y = \gamma(x_1, ..., x_n | \varepsilon_1, ..., \varepsilon_k)$$

where the x_i are independent non-random variables while the ε_j are random variables.

10.1.0 §Linear Statistical Models

Definition 10.1.1. *Let Y be a random variable that is a function of k independent variables $x_1, ..., x_k$. let $\beta_0, ..., \beta_k$ be associated unknown parameters. Then if $E[Y]$ is a linear function of the β_i, we call the model a* **linear statistical model***. If $k = 1$, it is a* **simple linear regression model***. Otherwise it is called a* **multiple linear regression model**

Note 10.1.1. The x_i are assumed to be measured without error in an experiment

Definition 10.1.2. *A* **linera statistical model** *relating a random response Y to a set of independent variables $x_1, ..., x_k$ is of the form*

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$$

where $\beta_0, ..., \beta_k$ are unknown parameters, ε is a random variable, and the variables $x_1, ..., x_k$ are known values. We will assume $E[\varepsilon] = 0$, so that

$$E[Y] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

10.2.0 §Method of Least Squares

Suppose we wish to fit the model $E[Y] = \beta_0 + \beta_1 x$ to some set of data. Note in general $x = f(w)$ is a function of some other independent variable(s). That is we postulate $Y = \beta_0 + \beta_1 x + \varepsilon$, where ε possesses some probability distribution with $E[\varepsilon] = 0$. If $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators for β_0 and β_1, then $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is an estimator for $E[Y]$.

Definition 10.2.1. *If $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of the ith y value, then the deviation or error of the observed value of y_i from $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the difference, $y_i - \hat{y}_i$. The sum of squares of deviations is given by*

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

If SSE possesses a min, it will occur for value of $\hat{beta}_0$ and $\hat{beta}_1$ that satisfy the equations

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = 0 \ \ \& \ \ \frac{\partial SSE}{\partial \hat{\beta}_1} = 0$$

These are the ***least-squares equations***.

Claim 10.2.1. *The minimizers of the SSE for two parameters are:*

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Proposition 10.2.2. *For $Y = \beta_0 + \beta_1 x + \varepsilon$, with $V(Y) = V(\varepsilon) = \sigma^2$:*

- $E[\hat{\beta}_1] = \beta_1$
- $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $E[\hat{\beta}_0] = \beta_0$
- $V(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\overline{x}\sigma^2}{S_{xx}} = \frac{-\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Remark 10.2.1. Since we are using $\hat{Y}_i$ to estimate $E[Y_i]$, we use

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}SSE$$

to estimate σ^2. This is an unbiased estimator.

Note 10.2.2. $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xx}$

Claim 10.2.3. *If $\varepsilon_i \sim norm(0, \sigma^2)$, $Y_i \sim norm(\beta_0 + \beta_1 x_i, \sigma^2)$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed. Moreover,*

$$\frac{(n-2)S_e^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{df=n-2}$$

and the statistic S_e^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$

Let $\hat{Y}$ be the probabilistic model that is closest overall to each sample point, meaning the st of sample points (X_i, Y_i)'s that are respectively closest to $(\hat{X}_i, \hat{Y}_i)$'s. Then we define the difference between Y_i and $\hat{Y}_i$ to be $\varepsilon_i = Y_i - \hat{Y}_i$, called the **residuals/errors/residual errors**. We are setting $\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$, so we can find the model with the least overall error. However, we need to square it to avoid negatives and canceling of errors.

Definition 10.2.2. *The* **sum of squared errors** *is*

$$SSE = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Claim 10.2.4. *The least squares estimate that minimizes SSE for $\hat{\beta}_0$ is*

$$\boxed{\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}}$$

and the least squares estimate for $\hat{\beta}_1$ is

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}$$

10.2.1 Evaluating the Quality of the Linear Relationship

Definition 10.2.3. $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ *is called the* **total sum of squared errors** *(SST)*

Remark 10.2.3. When we have Y_i's that are linearly dependent on X_i's, $(Y_i|X_i)$, a better estimator than simply $\overline{Y}$ is $\hat{Y}_i$ the simple linear regression estimator.

We now have two estimators $\overline{Y}$ and $\hat{Y}_i$, and their respective errors (SST and SSE).

Remark 10.2.4.

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \text{sum of explained errors}$$

which gives

$$SST = SSE + SSR$$

the total sum of squared errors = the sum of squared error still in regression plus the sum of squared error explained by the linear model.

Remark 10.2.5.

$$SST = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

so

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Definition 10.2.4. *The* **coefficient of determination** *is*

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

and

$$r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Note 10.2.6.

$$SST = S_{yy} = S_Y^2(n-1)$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n}\hat{y}_i^2 - n\overline{y}^2$$

$$SSE = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Remark 10.2.7. Least-Squares Conditions/Assumptions

- The response variable Y is normally distributed regardless of the value of the predictor variable X.
- The variation in the response variable Y is the same regardless of the value of the predictor variable X_i. This concept of common variation/standard deviation is called ***homoscedasticity***. This common variance in the response variable Y is represented by σ^2.

Remark 10.2.8. To see if our model satisfies 2., we look at the residuals plot: $(y_i - \hat{y}_i)$ with respect to $\hat{y}_i$. We want randomness, so we look for patterns.

Remark 10.2.9. If our model satisfies 1, the normal $Q - Q$ plot should have the points as close to the dotted line as possible.

Remark 10.2.10. Points that are numbered in Cook's distance could be outliers (plot of residuals versus leverages)

10.3.0 §Inferences Concerning the Parameters

Recall 10.3.1. If ε is normally distributed, $\hat{\beta}_i$ is an unbiased normally distributed estimator of β_i with

$$V(\hat{\beta}_0) = c_{00}\sigma^2, c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$$

and

$$V(\hat{\beta}_1) = c_{11}\sigma^2, c_{11} = \frac{1}{S_{xx}}$$

Proposition 10.3.1. *For a test of $H_0 : \beta_i = \beta_{i0}$ we can use the test statistic*

$$Z = \frac{\hat{\beta}_i - \beta_{i0}}{\sigma\sqrt{c_{ii}}}$$

The rejection region for a two tailed test is $|Z| > z_{\alpha/2}$. If we estimate σ with $S = \sqrt{SSE/(n-2)}$, then

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}} \sim T(df = n-2)$$

Remark 10.3.2. Test Hypothesis for Parameter

- $H_0 : \beta_i = \beta_{i0}$
- $$H_a : \begin{cases} \beta_i > \beta_{i0} & \text{(upper-tail alternative)} \\ \beta_i < \beta_{i0} & \text{(lower-tail alternative)} \\ \beta_i \neq \beta_{i0} & \text{(two-tailed alternative)} \end{cases}$$
- Test statistic: $T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$
- Rejection Region:
$$RR : \begin{cases} \{t > t_\alpha\} & \text{(upper-tail RR)} \\ \{t < -t_\alpha\} & \text{(lower-tail RR)} \\ \{|t| > t_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$$
- where $c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$ and $c_{11} = \frac{1}{S_{xx}}$

Remark 10.3.3. Confidence Interval $100(1-\alpha)\%$ confidence interval for β_i:

$$\hat{\beta}_i \pm t_{\alpha/2} S\sqrt{c_{ii}}$$

where $c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$ and $c_{11} = \frac{1}{S_{xx}}$

10.4.0 §Simple Linear Regression

Suppose we want to make an inference about $\theta = a_0\beta_0 + a_1\beta_1$. Then $\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$ is an unbiased estimator of θ. Then

$$V(\hat{\theta}) = a_0^2 c_{00}\sigma^2 + a_1^2 c_{11}\sigma^2 + 2a_0a_1c_{01}\sigma^2$$

where $c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$, $c_{01} = \frac{-\overline{x}}{S_{xx}}$, and $c_{11} = \frac{1}{S_{xx}}$, so

$$V(\hat{\theta}) = \left[\frac{a_0^2 \frac{\sum_{i=1}^n x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}} \right] \sigma^2$$

For large sampling $\hat{\beta}_0, \hat{\beta}_1 \sim norm$, so $\hat{\theta} \sim norm$, so

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim norm(0,1)$$

so a $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$$

Remark 10.4.1. Test for theta

- $H_0 : \theta = \theta_0$
- $$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative)} \\ \theta < \theta_0 & \text{(lower-tail alternative)} \\ \theta \neq \theta_0 & \text{(two-tailed alternative)} \end{cases}$$
- Test statistic: $T = \frac{\hat{\theta} - \theta_0}{S\sqrt{V(\hat{\theta})}}$
- Rejection Region:
$$RR : \begin{cases} \{t > t_\alpha\} & \text{(upper-tail RR)} \\ \{t < -t_\alpha\} & \text{(lower-tail RR)} \\ \{|t| > t_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$$
- where $c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$ and $c_{11} = \frac{1}{S_{xx}}$

Remark 10.4.2. Confidence Interval $100(1-\alpha)\%$ confidence interval for θ:

$$\hat{\theta} \pm t_{\alpha/2}S\sqrt{\left[\frac{a_0^2 \frac{\sum_{i=1}^n x_i^2}{n} + a_1^2 - 2a_0a_1\overline{x}}{S_{xx}} \right] \sigma^2}$$

where $t_{\alpha/2}$ is based on $n-2$ degrees of freedom.

Remark 10.4.3. Confidence Interval $100(1-\alpha)\%$ confidence interval for $E[Y] = \beta_0 + \beta_1 x^*$:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is based on $n-2$ degrees of freedom.

If we are interested in the value of Y when $x = x^*$, call it Y^*, we could use $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$. If $x = x^*$ the error of predicting a particular value of Y^* using $\hat{Y}^*$ is

$$error = Y^* - \hat{Y}^*$$

For ε normal, $Y^*, \hat{Y}^* \sim norm$. Since $E[\hat{Y}^*] = E[Y^*]$, $E[error] = 0$. Since we are predicting a future value of Y^* that is not employed in the computation of $\hat{Y}^*$, they are independent and $Cov(\hat{Y}^*, Y^*) = 0$. Then

$$V(error) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]$$

Then

$$Z = \frac{Y^* - \hat{Y}^*}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim norm(0,1)$$

and

$$T = \frac{Y^* - \hat{Y}^*}{S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim T(df = n-2)$$

Remark 10.4.4. Predictor Interval $100(1-\alpha)\%$ prediction interval for Y when $x = x^*$:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is based on $n-2$ degrees of freedom.

Remark 10.4.5. F-Test of Linear Appropriateness Consider the null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_2 = ... = 0$$
$$H_a : \text{at least one}\, \beta_i \neq 0, \text{ where } i = 1, 2, ...$$

Here in SLR there is only one slope β_1, so $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. One can test this using variance decomposition in the response variable, encountered in determining the coefficient of determination and standard deviation in the regression, to derive a test for linear appropriateness.

$$SST = SSE + SSR \implies \frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSR}{\sigma^2}$$
$$\frac{SST}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sigma^2} = \frac{(n-1)S_y^2}{\sigma^2} \sim \chi^2_{df=n-1}$$
$$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2} = \frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{df=n-2}$$

Then $\frac{SSR}{\sigma^2} \sim \chi^2_{df=1}$ since $n-1-(n-2) = 1$. Then, note $MSE = \frac{SSE}{(n-2)}$, and let $MSR = \frac{SSR}{1} = SSR$, so $MSi = \frac{SSi}{df_i}$

10.5.0 §Correlation

Suppose (X, Y) has a bivariate normal distribution with $E[X] = \mu_X$, $E[Y] = \mu_Y$, $V[X] = \sigma_X^2$, and $V[Y] = \sigma_Y^2$, and

$$\rho = \frac{Cov(X, Y)}{\sqrt{V[X]V[Y]}}$$

Then $E[Y|X = x] = \beta_0 + \beta_1 x$ where $\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample. The maximum likelihood estimator of ρ is given bye

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{yy}}} \end{aligned}$$

For testing hypotheses in terms of β_1 or ρ, we use

$$T = \frac{\hat{\beta}_1 - 0}{S/\sqrt{S_{xx}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(df = n-2)$$

r^2 is called the ***coefficient of determination***

Proposition 10.5.1.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$

where $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Remark 10.5.1. $S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2$ provides a measure of the total variation among the y-values, ignoring the x's. SSE measures the variation in the y-values that remains unexplained after using the x's to fit the simple linear regression model. Thus SSE/S_{yy} gives the proportion of the total variation in the y_i's that is unexplained by the model. Then

$$r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = 1 - \frac{SEE}{S_{yy}}$$

so r^2 is the proportion of total variation in the y's that is explained by the variable x in a simple linear regression model.

10.6.0 §Simple Linear Regressio: Bivariate Data

Definition 10.6.1. *Bivariate data is the observed values on two distinct/different population variables pertaining to unit/individuals in the population of interest.*

Notation 10.6.1. If X_i is the observed value of a random variable X from subject i, $i = 1, 2, ..., n$, and Y_i the observed value of random variable Y from subject i, $i = 1, 2, ..., n$, or

$$(X_i, Y_i) \in \{(X_1, Y_1), ..., (X_n, Y_n)\}$$

Remark 10.6.2. When an experimental study or random sampling method produces data on two different variables, there are three research questions that are posed:

1. Is there a relationship between the two variables? If there is, what is the direction of the relationship? Is the relationship positive? negative (or inverse)? Does the relationship seem to be linear? non-linear?
2. If a relationship exists between X and Y, how strong is the relationship? IS such a relationship subtle, or strong?
3. If the relationship between X and Y is strong, can the existing relationship be used to predict what will happen in the future? That is, can we create a mathematical function $f(x)$ that will predict the value of one variable based on the value of the other?

Recall 10.6.3.

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

Definition 10.6.2. *The* **sample covariance** *is given by*

$$\frac{S_{xy}}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

and it measures dependence.

Definition 10.6.3. **Pearson's Correlation** *is a scaled version of coveriance:*

$$Cor(X, Y) = \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X, Y)}{\sqrt{V[X]V[Y]}}, -1 \leqslant \rho \leqslant 1$$

Definition 10.6.4. *The* **sample correlation** *is given by:*

$$\begin{aligned} r &= \frac{S_{xy}/(n-1)}{S_X S_Y} \\ &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}} \\ &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, -1 \leqslant r \leqslant 1 \end{aligned}$$

Note that

$$S_X^2 = \frac{S_{xx}}{n-1}$$

Remark 10.6.4. Pearson's Correlation is a measure of the linear quality of the relationship between two variables. It wouldn't represent curvilinear relationships.

Remark 10.6.5. We want to model expected results on the varaible deemed the ***response variable***, Y, based only off the ***predictor variable***, X. Steps:

1. Collect bivariate X and Y variables from historic events. This data set will serve as training to understand the linear relationship that exists between the variables
2. Develop a mathematical expression/equation to transform a particular/hypothetical X into an expected/estimated Y

Remark 10.6.6. In a deterministic model all points fall on the same line.

Definition 10.6.5. *We define the linear equation to follow the deterministic model*

$$Y_i = \beta_0 + \beta_1 X_i$$

where β_0 is the y-intercept and β_1 is the slope.

Remark 10.6.7. In a deterministic model only two sample points are needed to find the model (slope point method).

Definition 10.6.6. *The probabilistic model appears to be the same as the deterministic model, however, it includes an additional ε_i, term:*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim norm(0, \sigma^2)$$

The **probabilistic model** *can also be written this way by admitting the following values are estimates:*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \sim norm(\hat{\beta}_0 + \hat{\beta}_1 X_i, \sigma^2)$$