Statistical Inference

Manish Agarwal

September 11, 2020

These notes are based on the classical book by Casella and Berger, one of the best applied statistics textbook.

1 Probability Theory

1.1 Set Theory

Definition 1.1. *The set, S, of all possible outcomes of a particular experiment is called the sample space for the experiment.*

Sample space can be countable or uncountable.

Definition 1.2. *An event is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).*

The relationship of containment ($A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$), and equality ($A = B \Leftrightarrow A \subset B$ and $B \subset A$) allow us to order and equate sets. Thereafter we define the operations of union ($A \cup B = \{x : x \in A \text{ or } x \in B\}$), intersection ($A \cap B = \{x : x \in A \text{ and } x \in B\}$), and complementation ($A^c = \{x : x \notin A\}$).

Theorem 1.1. *For any three events A, B, and C, defined on a sample space S,*

1. *Commutativity: $A \cup B = B \cup A$, $A \cap B = B \cap A$.*

2. *Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$.*

3. *Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.*

4. *DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.*

The operations of union and intersection can be extended to infinite collections of sets as well.

Definition 1.3. *Two events A and B are disjoint if $A \cap B = \emptyset$. The events $A_1, A_2, \dots$ are pairwise disjoint if $A_{ij} = \emptyset$ for all $i \neq j$.*

Definition 1.4. *If $A_1, A_2, \dots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \dots$ forms a partition of S.*

1.2 Basics of Probability Theory

'Frequency of occurrence' of an event in repeated experiments is one particular interpretation of probability. Another interpretation can be the belief in the chance of an event occurring in a single experiment. With Axiomatic approach to probability we are not concerned with interpretations but only that the probabilities are defined by a function satisfying the axioms.

Definition 1.5. *A collection of subsets of S (sample space) is called a **sigma algebra** or Borel field, denoted by $\mathcal{B}$, if it satisfies the following three properties:*

1. $\emptyset \in \mathcal{B}$
2. *If an event $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$,*
3. *if $A_1, A_2, \ldots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$*

S is always in $\mathcal{B}$. By DeMorgan's Laws it follows that $\mathcal{B}$ is closed under countable intersections as well. Associated with sample space S we have many different sigma algebras. For example, the collection of the two sets $\{\emptyset, S\}$ is a sigma algebra called **trivial sigma algebra**. The only sigma algebra we will be concerned with is the smallest one that contains all of open sets in a given sample space S.

Definition 1.6. *Kolmogorov Axioms of Probability: Given a sample space S and an associated sigma algebra $\mathcal{B}$, a **probability function** is a function P with domain $\mathcal{B}$ that satisfies*

- $P[A] \geq 0$, $\forall A \in \mathbf{B}$.
- $P[S] = 1$.
- *If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint then $P[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$.*

Theorem 1.2. *Let $S = \{s_1, \ldots, x_n\}$ be a countable set. Let $\mathcal{B}$ be any sigma algebra of subsets of S. Let $p_1, \ldots, p_n$ be non-negative numbers that sum to 1. For any $A \in \mathcal{B}$, define $P[A]$ by $P[A] = \sum\limits_{\{i:s_i \in A\}} p_i$. Then P is a probability function on $\mathcal{B}$.*

The physical reality of the experiment might dictate the probabilities assigned. School of statisticians led by deFinetti reject the **Axiom of Countable Additivity**, as it is not simple enough, replacing it by Axiom of Finite Additivity: if $A, B \in \mathcal{B}$ and are disjoint, then $P[A \cup B] = P[A] + P[B]$. This can lead to unexpected complications in statistical theory.

Theorem 1.3. *If P is a probability function and A is any set in $\mathcal{B}$, then*

1. *$P[\emptyset] = 0$, where $\emptyset$ is the empty set;*
2. *$P[A] \leq 1$;*
3. *$P[A^c] = 1 - P[A]$.*

Theorem 1.4. *If P is a probability function and A and B are any sets in $\mathcal{B}$, the*

1. *$P[B \cap A^c] = P[B] - P[A \cap B]$;*

2. *$P[A \cup B] = P[A] + P[B] - P[A \cap B]$;*

3. *If $A \subset B$, then $P[A] \leq P[B]$.*

Bonferroni's Inequality is a special case of the second item above, $PA \cap B \geq P[A] + P[B] - 1$.

Theorem 1.5. *If P is a probability function, then*

1. *$P[A] = \sum_{i=1}^{\infty} P[A \cap C_i]$ for any partition $C_1, C_2, \ldots$;*

2. *$P[\cup_{i=1}^{\infty}] \leq \sum_{i=1}^{\infty} P[A_i]$ for any sets $A_1, A_2, \ldots$, **Boole's Inequality**.*

Methods of counting are used in order to construct probability assignments on finite sample spaces with equally likely outcomes.

Theorem 1.6. *If a job consists of k separate tasks, the i^{th} of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times \ldots \times n_k$ ways.*

Arrangements can be done with and without order and replacement, which we summarize in table 1. The formula for the number of outcomes in the unordered sample space is useful for enumerating the outcomes, but ordered outcomes must be counted to correctly calculate probabilities; i.e. the probabilities should be determined by sampling mechanism, not indistinguishably. In general, if there are k places and we have m different numbers repeated $k_1, k_2, \ldots, k_m$ times, then the number of ordered samples is $\frac{k!}{k_1!k_2!\ldots k_m!}$. This is related to multinomial distribution.

	without replacement	with replacement
Ordered	${}^{n}P_{r}$	n^r
Unordered	${}^{n}C_{r}$	${}^{n+r-1}C_{r}$

Table 1: Number of possible arrangements of size r from n objects.

1.3 Conditional Probability and Independence

Definition 1.7. *If A and B are events in S, and $P[B] > 0$, then the conditional probability of A given B, written $P[A|B]$ is $\frac{P[A \cap B]}{P[B]}$.*

The intuition is that our original sample space, S, has been updated to B. The function $P[.|B]$ satisfies Kolmogorov's Axioms. This gives the law $P[A \cap B] = P[A|B]P[B] = P[B|A]P[A]$.

Theorem 1.7. *Bayes' Rule: Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,*

$$P[A_i|B] = \frac{P[B|A_i]P[A_i]}{\sum_{j=1}^{\infty} P[B|A_j]P[A_j]}.$$

Definition 1.8. *Two events, A and B, are statistically independent if $P[A \cap B] = P[A]P[B]$.*

Theorem 1.8. *If A and B are independent events, then the following pairs are also independent - A and B^c, A^c and B, and A^c and B^c.*

Independence of more than two events can't be defined by $p[A \cap B \cap C] = P[A]P[B]P[C]$, or by requiring all the pairs to be independent. We need a stronger condition.

Definition 1.9. *A collection of events $A_1, \ldots, A_n$ are mutually independent if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have $P\left[\cap_{j=1}^{k} A_{i_j}\right] = \prod_{j=1}^{k} P[A_{i_j}]$.*

1.4 Random Variables

Definition 1.10. *A random variable is a function from the sample space S into the real numbers.*

In defining a random variable, we have also defined a new sample space, the range of the random variable. We must now verify that out probability function, which is defined on the original sample space, can be used for the random variable.

Suppose we have a sample space $S = \{s_1, \ldots, s_n\}$ with a probability function P and we define a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function P_X on $\mathcal{X}$ such that we will observe $X = x_i$ iff the outcome of the random experiment is an $s_j \in S$ such that $X(x_j) = x_i$. Thus, $P_X[X = x_i] = P[\{s_j : X(s_j)\} = x_i]$. P_x is an induced probability function on $\mathcal{X}$, defined in terms of original function P. P_X satisfies Kolmogorov Axioms. Because of the equivalence above we simply write $P[X = x_i]$ rather than $P_X[X = x_i]$.

1.5 Distribution Functions

Definition 1.11. *The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by $F_X(x) = P_X[X \leq x]$, $\forall x$.*

Theorem 1.9. *The function $F(x)$ is a cdf iff the following three conditions hold:*

- $\lim\limits_{x \to -\infty} F(x) = 0$ *and* $\lim\limits_{x \to \infty} F(x) = 1$.
- *$F(x)$ is a non-decreasing function of x.*
- *$F(x)$ is right-continuous, that is for every number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.*

Definition 1.12. *A random variable X is continuous if $F_X(x)$ is a continuous function of x. A random variable X is discrete if $F_X(x)$ is a step function of x.*

F_X completely determines the probability distribution of a random variable X. This is true if $P[X \in A]$ is defined only for events A in $\mathcal{B}^1$, the smallest sigma algebra containing all the intervals of real numbers of the form (a, b), $[a, b)$, $(a, b]$, and $[a, b]$.

Definition 1.13. *The random variables X and Y are identically distributed if, for every set $A \in \mathcal{B}^1$, $P[X \in A] = P[Y \in A]$.*

Note that two random variables that are identically distributed are not necessarily equal.

Theorem 1.10. *The following two statements are equivalent:*

1. *The random variables X and Y are identically distributed.*
2. $F_X(x) = F_Y(x)$ *for every* x.

1.6 Density and Mass Functions

Definition 1.14. *The probability mass function of a discrete random variable X is given by $f_X(x) = P[X = x]$, $\forall x$. The probability density function or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$, $\forall x$.*

For a continuous function $\frac{d}{dx}F_X(x) = f_X(x)$.

Theorem 1.11. *A function $f_X(x)$ is a pdf (or pmf) of a random variable X iff*

1. $f_X(x) \geq 0$, $\forall x$.
2. $\sum_x f_X(x) = 1$ *for a pmf or* $\int_{-\infty}^{\infty} f_X(x)dx = 1$ *for a pdf.*

Actually, thee are continuous random variables for which the integral relationship does not exist for any $f_X(x)$, because $F_X(x)$ may be continuous but not differentiable, e.g. Cantor set.

2 Transformations and Expectations

2.1 Distributions of Functions of a Random Variable

If X is a random variable with cdf $F_X(x)$, then any function of X, say $Y = g(X)$ is also a random variable. For any set A, $P[Y \in A] = P[g(X) \in A]$, showing that the distribution of Y depends on the functions F_X and g. Formally, if we write $y = g(x)$, the function $g(x)$ defines a mapping from the original sample space of X, X, to a new sample space, $\mathcal{Y}$, the sample space of the random variable Y, i.e. $g(x) : \mathcal{X} \to \mathcal{Y}$. We associate with g and inverse mapping, denoted by g^{-1}, which is a mapping from subsets of $\mathcal{Y}$ to subsets of $\mathcal{X}$, defined by $g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$, taking sets into sets. If the random variable Y is now defined by $Y = g(X)$, we can write for any set $A \subset \mathcal{Y}$,

$$P[Y \in A] = P[g(X) \in A] = P[\{x \in \mathcal{X} : g(x) \in A\}] = P[X \in g^{-1}(A)].$$

This defines the probability distribution of Y and satisfies Kolmogorov Axioms.

If X is a discrete random variable, then $\mathcal{X}$ is countable. The sample space for $Y = g(X)$ is $\mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$, which is also a countable set. Thus, Y is also a discrete random variable. Hence,

$$f_Y(y) = P[Y = y] = \sum_{x \in g^{-1}(y)} P[X = x] = \sum_{x \in g^{-1}(y)} f_X(x), fory \in \mathcal{Y},$$

and $f_Y(y) = 0$ for $y \notin \mathcal{Y}$.

If X and Y are continuous random variables, the cdf of $Y = g(X)$ is

$$F_Y(y) = P[Y7 \leq y] = P[g(X) \leq y] = P[\{x \in \mathcal{X} : g(x) \leq y\}] = \int_{\{x \in \mathcal{X} : g(x) \leq y\}} f_X(x)dx.$$

For the transformation from X toe $Y = g(X)$, it is most convenient to use

$$\mathcal{X} = \{x : f_X(x) > 0\} \text{ and } \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

The distribution of the random variable $\mathcal{X}$ is positive only on the set $\mathcal{X}$ and is 0 elsewhere. Such a set is called the *support of a distribution.* It is easiest to deal with functions $g(x)$ that are monotone on the support set. In that case the transformation $y = g(x)$ is one-to-one and onto form $\mathcal{X} \to \mathcal{Y}$. The cumulative distribution function serves us well here.

Theorem 2.1. *Let X have cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x)$ for some $x \in \mathcal{X}\}$.*

1. *If g is an increasing function on $\mathcal{X}$, $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.*

2. *If g is a decreasing function on $\mathcal{X}$ and X is a continuous random variable, $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.*

Theorem 2.2. *Let X have a pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x)$ for some $x \in \mathcal{X}\}$. Suppose $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \textit{otherwise} \end{cases}$$

When g is not monotone the above results will not apply. In that case the pdf of Y is expressed as the sum of pieces, the represent the intervals where g is monotone.

Theorem 2.3. *Let X have pdf $f_X(x)$, let $Y = g(X)$, and define the sample space $\mathcal{X} = \{x : f_X(x) > 0\}$. Suppose there exists a partition $A_0, A_1, \ldots, A_k$, of $\mathcal{X}$ such that $P[X \in A_0] = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exits functions $g_1(x), \ldots, g_k(x)$, defined on $A_1, \ldots, A_k$, respectively, satisfying*

1. *$g(x) = g_i(x)$, for $x \in A_i$,*

2. *$g_i(x)$ is monotone on A_i,*

3. *the set $\mathcal{Y} = \{y : y = g_i(x)$ for some $x \in A_i\}$ is the same for each $i = 1, \ldots, k$, and*

4. *$g_i^{-1}(y)$ has a continuous derivatives on $\mathcal{Y}$, for each $i = 1, \ldots, k$.*

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X(g_i^{-1}(y))\left|\frac{d}{dy}g_i^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \textit{otherwise} \end{cases}.$$

Example 2.1. Let X have the standard normal distribution, $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $-\infty < x < \infty$. Consider $Y = X^2$. The function $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$. The set $\mathcal{Y} = (0, \infty)$. Applying the last theorem, we take $A_0 = \{0\}$; $A_1 = \{-\infty, 0\}$, $g_1(x) = x^2$, $g_1^{-1}(y) = -\sqrt{y}$; $A_2(0, \infty)$, $g_2(x) = x^2$, $g_2^{-1}(y) = \sqrt{y}$. The pdf of Y is $f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-(=\sqrt{y})^2/2}\left| - \frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}}e^{-(\sqrt{y})^2/2}\left|\frac{1}{\sqrt{y}}\right| = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-y/2}$, $0, y, \infty$. The pdf of Y is that of a chi squared random variable with 1 degree of freedom. □

Theorem 2.4. *Probability integral transformation: Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P[Y \leq y] = y$, $0 < y < 1$.*

One application of this theorem is the generation of random samples from a particular distribution. If F_X is strictly increasing, then F_X^{-1} is well defined by $F_X^{-1}(y) = x \iff F_X(x) = y$. However if F_X is constant on some interval, then F_X^{-1} is not well defined by this equivalence. Any x satisfying $x_1 \leq x \leq x_2$ satisfies $F_X(x) = y$. The problem is avoided by defining $F_X^{-1}(y)$ for $0 < y < 1$ by $F_X^{-1}(y) = inf\{x : F_X(x) \geq y\}$. Using this definition, we have $F_X^{-1}(y) = x_1$ for the segment range where x is non increasing. At the endpoints of the range of y, $F_X^{-1}(y)$ can be defined. $F_X^{-1} = \infty$ if $F_X(x) < 1$ for all x, for any F_X, $F_X^{-1}(0) = -\infty$.

2.2 Expected Values

Expected value is the probability weighted average. It is a measure of central tendency.

Definition 2.1. *The expected value or mean of a random variable $g(X)$, denoted by $E[g(X)]$, is*

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \textit{if } X \textit{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P[X = x] & \textit{if } X \textit{ is discrete,} \end{cases}$$

provided that the integral or sum exists. If $|E[g(X)]| = \infty$, we say that $E[g(X)]$ does not exist.

A classical example of a random variable whose expected value does not exist is a Cauchy random variable, that is, one with pdf

$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + x^2}, \quad -\infty < x < \infty.$$

The process of taking expectations is a linear operation, i.e. $E[aX + b] = aE[X] + b$.

Theorem 2.5. *Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

1. $E[ag_1(X) + bg_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c.$

2. *If $g_1(x) \geq 0$ for all x, then $E[g_1(X)] \geq 0$.*

3. *If $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(X)] \geq Eg_2(X)$.*

4. *If $1 \leq g_1(x) \leq b$ for all x, then 1≤E[$g_1(X)$] $\leq b$.*

When evaluating expectations of nonlinear functions of X, we can proceed in one of two ways. We could directly use the definition and calculate $E[g(X)] = \int g(x) f_X(x) dx$. But we could also find the pdf $f_Y(y)$ for $Y = g(X)$ and then calculate $E[g(X)] = E[Y] = \int y f_Y(y) dy$.

2.3 Moments and Moment Generating Functions

Definition 2.2. *For each integer n, the nth moment of X is $\mu_n' = E[X^n]$. The nth central moment of X is $\mu_n = E[(X - \mu)^n]$, where $\mu = \mu_1' = E[X]$.*

Definition 2.3. *The variance of a random variable X is its second central moment, $Var[X] = E[(X - E[X])^2]$. The positive square root of $Var[X]$ is the standard deviation of X.*

The variance gives a measure of the degree of spread of a distribution around its mean.

Theorem 2.6. *If X is a random variable with finite variance, then for any constants a and b, $Var[aX + b] = a^2 Var[X]$.*

Sometimes it is easier to use the alternative formula for variance $Var[X] = E[X^2] - (E[X])^2$.

Definition 2.4. *Let X be a random variable with cdf F_X. The* ***moment generating function****, mgf of X is $M_X(t) = E[e^{tX}]$, provided that the expectations exists for t in some neighborhood of 0.*

Theorem 2.7. *If X has mgf $M_X(t)$, then $E[X^n] = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$. That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.*

Moment generating functions can generate the moments, but more importantly can characterize a distribution. There are, however, some technical difficulties associated. If the mgf exists, it characterizes an infinite set of moments. But characterizing the infinite set of moments uniquely does not uniquely determine a distribution function. There may be two distinct random variable having the same moments. For example $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}$, $0 \leq x < \infty$ and $f_2(x) = f_1(x)[1 + sin(2\pi \log x)]$, $0 \leq x < \infty$, both have same set of moments but different pdfs.

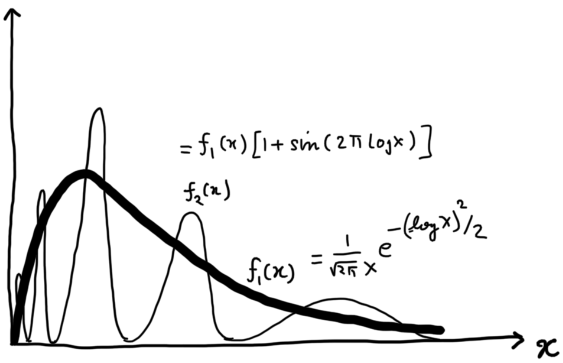

Figure 1: Non unique pdfs for same moments

The problem of uniqueness of moments does not occur if the random variables have bounded support or if mgf exists in a neighborhood of 0, then the distribution is uniquely determined, no matter what the support.

Theorem 2.8. *Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.*

1. *if X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u iff $E[X^r] = E[Y^r]$ for all integers $r = 0, 1, 2, \ldots$.*

2. *If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

Theorem 2.9. *Suppose $\{X_i, i = 1, 2, \ldots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that $\lim_{i\to\infty} M_{X_i}(t) = M_X(t)$, for all t in a neighborhood or 0, and $M_X(t)$ is an mfg. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have $\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$. That is, convergence, for $|t| < h$, of mgfs to an mgf implies convergence of cdfs.*

The proof of this relies on the theory of Laplace transforms and the fact that they are unique. The defining equation for $M_X(t)$, that is

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx,$$

defines a Laplace transform, i.e. $M_X(t)$ is a Laplace transform of $f_X(x)$. If this definition is valid for all t such that $|t| < h$, where h is some positive number, then given $M_X(t)$ there is only one function $f_X(x)$ that satisfies this definition.

If we show that a sequence of moments converges, we will not be able to conclude formally that the random variable converge. To do so, we would have to verify the uniqueness of the moment sequence. However, if the sequence of mgfs converges in a neighborhood of 0, then the random variables converge. Thus, the convergence of mgfs is a sufficient, but not necessary condition for the sequence of random variables to converge.

Example 2.2. *Poisson approximation:* Binomial probabilities can be approximated by Poisson probabilities, which are generally easier to calculate. The Binomial distribution is characterized by two quantities, n and p. When n is large and np is small the approximation is valid. The $Poisson(\lambda)$ pmf is given by $P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$, where λ is a positive constant. The approximation states that $X \sim Binomial(n, p)$ and $Y \sim Poisson(\lambda)$, with $\lambda = np$, then $P[X = x][Y = x]$ for large n and small np. We can show that the mgfs converge. The mfg for binomial is $M_X(t) = [pe^t + (1-p)]^n$ and for poisson is $M_Y(t) = e^{\lambda(e^t-1)}$. If $p = \lambda/n$, then $M_X(t) \to M_Y(t)$ as $n \to \infty$. □

Theorem 2.10. *For any constants a and b, the mgf of the random variable $aX + b$ is given by $M_{aX+b}(t) = e^{bt} M_X(at)$.*

2.4 Differentiating under an integral sign

Theorem 2.11 (Leibnitz's Rule)**.** *If $f(x, \theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x, \theta) dx.$$

If $a(\theta)$ and $b(\theta)$ are constant, we have

$$\frac{d}{d\theta} \int_a^b f(x, \theta) dx = \int_a^b \frac{\partial}{\partial\theta} f(x, \theta) dx.$$

If we have integrals of a differentiable function over a finite range, differentiation of the integral poses no problem. If the range of the integration is infinite, problems can arise. We need to investigate whether limits and integration can be interchanged, since a derivative is a special kind of limit. Using Lebesgue's Dominated Convergence theorem we can state the following theorems.

Theorem 2.12. *Suppose the function $h(x, y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying*

1. $|h(x,y)| \leq g(x)$ *for all* x *and* y,

2. $\int_{-\infty}^{\infty} g(x)dx < \infty$.

Then,

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x,y)dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x,y)dx.$$

If the dominating function $g(x)$ exists with finite integral, we are ensured that the integrals cannot be too badly behaved. Identifying $h(x,y)$ with the difference $\frac{1}{\delta}(f(x,\theta+\delta) - f(x,\theta))$ we state the following theorem.

Theorem 2.13. *Suppose* $f(x,\theta)$ *is differentiable at* $\theta = \theta_0$, *that is,*

$$\lim_{\delta \to 0} \frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta} = \frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0}$$

exists for every x, *and there exists a function* $g(x,\theta_0)$ *and a constant* $\delta_0 > 0$ *such that*

1. $\left|\frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta}\right| \leq g(x,\theta_0)$, *for all* x *and* $|\delta| \leq \delta_0$,

2. $\int_{-\infty}^{\infty} g(x,\theta_0)dx < \infty$.

Then,

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta)dx\Big|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0}\right) dx.$$

The first condition is similar to *Lipschitz condition*, effectively bounding the variability in the first derivative. The statement of the theorem is for only one value of θ. That is, for each value or θ_0 for which $f(x,\theta)$ is differentiable at θ_0 and satisfies the two conditions, the order of integration and differentiation can be interchanged. Typically, $f(x,\theta)$ is differential at all θ, not at just one value of θ_0. Under this case, the first condition can be replaced by

$$\left|\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta'}\right| \leq g(x,\theta) \text{ for all } \theta' \text{ such that } |\theta' - \theta| \leq \delta_0.$$

This can be seen by application of the mean value theorem, it follows that, for fixed x and θ_0, and $|\delta| \leq \delta_0$, $\frac{1}{\delta}(f(x,\theta_0+\delta) - f(x,\theta_0)) = \frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0+\delta^*(x)}$ for some number $\delta^*(x)$, where $|\delta^*(x)| \leq \delta_0$.

Theorem 2.14. *Suppose* $f(x,\theta)$ *is differentiable in* θ *and there exists a function* $g(x,\theta)$ *such that* $\left|\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta'}\right| \leq g(x,\theta)$ *for all* θ' *such that* $|\theta' - \theta| \leq \delta_0$ *is satisfied and* $\int_{-\infty}^{\infty} g(x,\theta)dx < \infty$. *Then* $\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta)dx\Big|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0}\right) dx$ *holds.*

Example 2.3. Let X have the *exponential*(λ) pdf given by $f(x) = \frac{1}{\lambda} e^{-x/\lambda}$, $0 < x < \infty$, and suppose we want to calculate

$$\frac{d}{d\lambda} E[X^n] = \frac{d}{d\lambda} \int_0^{\infty} x^n \frac{1}{\lambda} e^{-x/\lambda} dx$$

for integer $n > 0$. If we could move the differentiation inside the integral, we would have

$$\begin{aligned}
\frac{d}{d\lambda}E[X^n] &= \int_0^\infty \frac{\partial}{\partial\lambda} x^n \frac{1}{\lambda} e^{-x/\lambda} dx \\
&= \int_0^\infty \frac{x^n}{\lambda^2}\left(\frac{x}{\lambda} - 1\right) e^{-x/\lambda} dx \\
&= \frac{1}{\lambda^2}E[X^{n+1}] - \frac{1}{\lambda}E[X^n].
\end{aligned}$$

To justify the interchange of integration and differentiation, we bound the derivative of $x^n \frac{1}{\lambda} e^{-x/\lambda}$. Now,

$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^n e^{-x\lambda}}{\lambda}\right)\right| = \frac{x^n e^{-x/\lambda}}{\lambda^2}\left|\frac{x}{\lambda} - 1\right| \le \frac{x^n e^{-x/\lambda}}{\lambda^2}\left(\frac{x}{\lambda} + 1\right).$$

For some constant δ_0 satisfying $0 < \delta_0 < \lambda$, take

$$g(x,\lambda) = \frac{x^n e^{-x/(\lambda+\delta_0)}}{(\lambda - \delta_0)^2}\left(\frac{x}{\lambda - \delta_0} + 1\right).$$

We then have

$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^n e^{-x/\lambda}}{\lambda}\right)\Big|_{\lambda=\lambda'}\right| \le g(x,\lambda)$$

for all λ' such that $|\lambda' - \lambda| \le \delta_0$. Since the exponential distribution has all its moments, $\int_{-\infty}^{\infty} g(x,\lambda)dx < \infty$ as long as $\lambda - \delta_0 > 0$, so the interchange of integration and differentiation is justified. □

We get a recursive relation for the moments of the exponential distribution

$$E[X^{n+1}] = \lambda E[X^n] + \lambda^2 \frac{d}{d\lambda}E[X^n],$$

making calculations of moments easy. This type of relationship exists for other distributions. In particular, if X is a normal distribution with mean μ and variance 1, then

$$E[X^{n+1}] = \mu E[X^n] - \frac{d}{d\mu}E[X^n].$$

Example 2.4. Let X have a normal distribution with mean μ and variance 1, and consider the mfg of X, $M_X(t) = E[e^{tX}] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-(x-\mu)^2/2}dx$. To calculate moments by differentiating $M_X(t)$ we need to do differentiation under the integral sign. We must find a function $g(x,t)$ with finite integral, that satisfies

$$\frac{\partial}{\partial t}e^{tx}e^{-(x-\mu)^2/2}\Big|_{t=t'} \le g(x,t) \qquad \text{for all } t' \text{ such that} |t' - t| \le \delta_0.$$

we note

$$\left|\frac{\partial}{\partial t}e^{tx}e^{-(x-\mu)^2/2}\right| = \left|xe^{tx}e^{-(x-\mu)^2/2}\right| \le |x|e^{tx}e^{-(x-\mu)^2/2}.$$

We proceed to define $g(x,t)$ as

$$g(x,t) = \begin{cases} |X|e^{(t-\delta_0)x}e^{-(x-\mu)^2/2} & \text{if } x < 0 \\ |X|e^{(t+\delta_0)x}e^{-(x-\mu)^2/2} & \text{if } x \geq 0 \end{cases} .$$

This function clearly satisfies the bounding condition. We need to check that its integral is finite. For $x \geq 0$ we have $g(x,t) = xe^{-(x^2-2x(\mu+t+\delta_0)+\mu^2)/2}$. We complete the square in the exponent and not that for $x \geq 0$,

$$g(x,t) = xe^{-[x-(\mu+t+\delta_0)]^2/2}e^{-[\mu^2-(\mu+t+\delta_0)^2]/2}.$$

Since the last exponential factor in this expression does not depend on x, the integral of this function $\int_0^\infty g(x,t)dx$ is essentially calculating the mean of a normal distribution with mean $\mu + t + \delta_0$, except that the integration is only over $[0,\infty)$. Since the normal has finite mean this integral is finite. A similar argument for $x < 0$ follows. Therefore, we have found an integrable function satisfying boundedness conditions and the interchange of differential and integral is justified. □

Justification for taking the derivative inside the summation is more straightforward than the integration case.

Theorem 2.15. *Suppose that the series $\sum_{x=0}^\infty h(\theta,x)$ converges for all θ in an interval (a,b) of real numbers and $\frac{\partial}{\partial\theta}h(\theta,x)$ is continuous in θ for each x, and $\sum_{x=0}^\infty \frac{\partial}{\partial\theta}h(\theta,x)$ converges uniformly on every closed bounded sub-interval of (a,b). Then*

$$\frac{d}{d\theta}\sum_{x=0}^\infty h(\theta,x) = \sum_{x=0}^\infty \frac{\partial}{\partial\theta}h(\theta,x).$$

A series converges uniformly if its sequence of partial sums converges uniformly.

Example 2.5. Let X be a discrete random variable with geometric distribution $P[X = x] = \theta(1-\theta)^x$, $x = 0, 1, \ldots$, $0 < \theta < 1$. We have that $\sum_{x=0}^\infty \theta(1-\theta)^x = 1$ and, provided that the operations are justified,

$$\begin{aligned} 0 = \frac{d}{d\theta}\sum_{x=0}^\infty \theta(1-\theta)^x &= \sum_{x=0}^\infty \frac{d}{d\theta}\theta(1-\theta)^x \\ &= \sum_{x=0}^\infty \left[(1-\theta)^x - \theta x(1-\theta)^{x-1}\right] \\ &= \frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x - \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x \\ \frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x &= \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x \\ \frac{1}{\theta} &= \frac{1}{1-\theta}E[X] \\ E[X] &= \frac{1-\theta}{\theta}. \end{aligned}$$

To justify differentiation under summation we use the previous theorem and identify $h(\theta, x) = \theta(1-\theta)^x$ and $\frac{\partial}{\partial\theta}h(\theta, x) = (1-\theta)^x - \theta x(1-\theta)^{x-1}$, and verify that $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta}h(\theta, x)$ converges uniformly. Define $S_n(\theta) = \sum_{x=0}^{n} [(1-\theta)^x - (1-\theta)^{x-1}]$. The convergence will be uniform on $[c, d] \subset (0, 1)$ if, given $\epsilon > 0$, we can find an N such that

$$n > N \implies |S_n(\theta) - S_\infty(\theta)| < \epsilon \quad \text{for all } \theta \in [c, d].$$

Applying the sum of geometric series formula

$$\begin{aligned}
\sum_{x=0}^{n}(1-\theta)^x &= \frac{1-(1-\theta)^{n+1}}{\theta} \\
\sum_{x=0}^{n}\theta x(1-\theta)^{x-1} &= \theta\sum_{x=0}^{n} = \frac{\partial}{\partial\theta}(1-\theta)^x \\
&= -\theta\frac{d}{d\theta}\sum_{x=0}^{n}(1-\theta)^x \\
&= -\theta\frac{d}{d\theta}\frac{1-(1-\theta)^{n+1}}{\theta} \\
&= \frac{(1-(1-\theta)^{n+1}) - (n+1)\theta(1-\theta)^n}{\theta}.
\end{aligned}$$

Here we justifiably pull the derivative through the finite sum. Hence we show that $S_n(\theta) = (n+1)(1-\theta)^n$, after some algebra.

For $0 < \theta < 1$, $S_\infty = \lim_{n\to\infty} S_n(\theta) = 0$. Since $S_n(\theta)$ is continuous, the convergence is uniform on any closed bounded interval. Therefore, the series of derivatives converges uniformly and the interchange of differentiation and summation is justified. □

Theorem 2.16. *Suppose the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges uniformly on $[a, b]$ and that, for each x, $h(\theta, x)$ is a continuous function of θ. Then*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x)d\theta = \sum_{x=0}^{\infty}\int_a^b h(\theta, x)d\theta.$$

3 Common families of distributions

3.1 Discrete Distributions

Discrete Uniform Distribution - A random variable X has a discrete uniform $(1, N)$ distribution if

$$P[X = x|n] = \frac{1}{N},\ x = 1, \ldots, N,$$

where N is a specified integer. The expected value is $E[X] = \frac{N+1}{2}$ and variance $V[X] = \frac{N^2-1}{12}$

Hypergeometric distribution - From an urn with M red balls and N-M green balls draw K balls without replacement. The probability of total red balls being x if a hypergeometric distribution given by

$$P[X = x|N, M, K] = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}},\ x = 0, 1, \ldots, K.$$

We have $E[X] = \frac{KM}{N}$ and $V[X] = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$. This has application in acceptance sampling.

Bernoulli Distribution - There are two possible outcomes with success happening with probability p and failure with probability $1-p$. This is also called the indicator function to an event.

$$P[X] = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases},$$

with $0 \leq p \leq 1$. We have $E[X] = p$ and $V[X] = p(1-p)$. This is same as the indicator variable $\mathbb{I}_{success}$.

Binomial Distribution - If there are n trials, the variable to represent y success, Y is Binomial(n, y) such that $Y = \sum_i X_i$, where X_i is the i^{th} Bernoulli trial.

$$P[Y = y|n, p] = \binom{n}{y} p^y (1-p)^{n-y},\ y = 0, 1, \ldots, n.$$

We have $E[Y] = np$ and $V[Y] = np(p-1)$. When $n \to \infty$ and $p \to 0$, such that $np \to \lambda$, we get Poisson approximation. The mgf is $M_X(t) = [pe^t + (1-p)]^n$.

Poisson Distribution - A random variable X, taking values in the non-negative integers, has a Poisson distribution if

$$P[X = x|\lambda] = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, \ldots$$

$E[X] = \lambda$ and $V[X] = \lambda$ with $M_X(t) = e^{\lambda(e^t-1)}$. The recursive relation $P_x = \frac{\lambda}{x} P_{x-1}$ is useful for rapid calculations.

Example 3.1 (Poisson approximation)**.** A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

If setting a word is a Bernoulli trail with success probability $\frac{1}{500}$ and that the trials are independent, then X = number of errors in five pages is $Binomial(1500, \frac{1}{500})$. Thus,

$$\begin{aligned} P[X \leq 2] &= \sum_{x=0}^{2} \binom{1500}{x} \left(\frac{1}{500}\right)^{x} \left(\frac{499}{500}\right)^{1500-x} \\ &= 0.4230 \end{aligned}$$

which is fairly cumbersome calculations. If we use the Poisson approximation with $\lambda = 1500\frac{1}{500} = 3$, we have

$$P[X \leq 2] \approx e^{-3} \left(\frac{3^0}{0!} + \frac{3^1}{1!} + \frac{3^2}{2!}\right) = 0.4232$$

□

Negative Binomial Distribution - We want to count the number of Bernoulli trials required to get a fixed number of successes. X is the trial at which the r^{th} success occurs (the stopping time). Then

$$P[X = x|r,p] = \binom{x-1}{r-1} p^r (1-p)^{x-r},\ x = r, r+1,$$

$E[X] = r(1-p)/p$ and $V[X] = r(1-p)/p^2$. If $r \to \infty$ and $p \to 1$ such that $r(1-p) \to \lambda$, then we get Poisson(λ). Negative binomial distribution can, like the Poisson, be used to model phenomena in which we are waiting for an occurrence. In the negative binomial case we are waiting for a specified number of successes.

Geometric Distribution - If we set $r = 1$ in negative binomial we get geometric distribution

$$P[X = x|p] = p(1-p)^{x-1},\ x = 1, 2, \ldots.$$

$E[X] = \frac{1}{p}$ and $V[X] = \frac{1-p}{p^2}$. It has memoryless property, $P[X > s|X > t] = P[X > s-t]$, for integers $s > t$. Hence, they are not applicable to model lifetimes for which the probability of failure is expected to increase with time.

3.2 Continuous Distributions

Uniform distribution - The uniform distribution is

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}.$$

$E[X] = \frac{a+b}{2}$, and $V[X] = \frac{(b-a)^2}{12}$.

Gamma distribution - A gamma function is defined as $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$, with $\Gamma(\alpha+1) = \alpha\Gamma(\alpha),$ for $\alpha > 0$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. The Gamma distribution is obtained by substituting $X = \beta T$ to get,

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

$E[X] = \alpha\beta$ and $V[X] = \alpha\beta^2$. The mgf is given by $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha$, $t < \frac{1}{\beta}$.

If $\alpha = p/2$, where p is in integer, and $\beta = 2$ we get a **chi squared distribution** with p degree of freedom.

$$f(x|p) = \frac{1}{\Gamma\left(\frac{p}{2}\right)2^{\frac{p}{2}}}x^{\frac{p}{2}-1}e^{-\frac{x}{2}}, \quad 0 < x < \infty.$$

If $\alpha = 1$ we get **exponential distribution** with scale parameter β.

$$f(x|\beta) = \frac{1}{\beta}e^{-\frac{x}{\beta}}, \quad 0 < x < \infty.$$

This is generally used to model lifetimes and has the memoryless property.

If $X \sim \mathit{Exponential}(\beta)$ then $Y = X^{1/\gamma}$ is $Weibull(\gamma,\beta)$. The **Weibull distribution** has the following pdf

$$f_Y(y|\gamma,\beta) = \frac{\gamma}{\beta}y^{\gamma-1}e^{-y^\gamma/\beta}, \quad 0 < y < \infty, \quad \gamma > 0, \quad \beta > 0.$$

This plays an important role in analysis of failure time and modeling hazard functions.

Normal distribution - The pdf is given by

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

$E[X] = \mu$ and $V[X] = \sigma^2$. $Z = (X-\mu)/$ defines a standard normal $\mathcal{N}(0,1)$. Also, $\int_0^\infty e^{-z^2/2}dz = \sqrt{\frac{\pi}{2}}$. Normal distribution is fully defined by the two parameters. This property is shared by a family of pdfs called location-scale families. Normal has a maximum at $x = \mu$ and inflection points (where the curve changes from concave to convex) at $\mu \pm \sigma$. The probability content within 1,2 and 3 σs are 0.68, 0.95 and 0.99, these values are often very useful to remember.

Among the many uses of the normal distribution, an important one is to approximate other distributions, justified by CLT.

Example 3.2 (Normal approximation of the Binomial)**.** Let $X \sim \mathit{Binomial}(25, 0.6)$. We can approximate X with a normal random variable Y, with mean $\mu = 25 \times 0.6 = 15$ and standard deviation $\sigma = \sqrt{25 \times 0.6 \times (1-0.6)} = 2.45$. The suitable conditions for this approximation to hold is that n should be large and p should not be extreme. A conservative rule to follow is that the approximation will be good if $min(np, n(1-p)) \geq 5$, which is applicable here. Thus,

$$p[X \leq 13][Y \leq 13] = P[Z \leq \frac{13-15}{2.45}] = P[Z \leq -0.82] = 0.206,$$

which approximates binomial calculations

$$P[X \leq 13] = \sum_{x=0}^{13} \binom{25}{x} (0.6)^x (1-0.6)^{25-x} = 0.267$$

With continuity correction we instead calculate $P[X \leq 13.5]$, adding back the half to the cutoff point, to obtain 0.271 a much better approximation. □

Beta distribution - The continuous family on $(0, 1)$ given by $beta(\alpha, \beta)$,

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta) x^{\alpha-1}(1-x)^{\beta-1}}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0,$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$ is the beta function. $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. This is often used to model proportions. $E[X^n] = \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)}$, giving $E[X] = \frac{\alpha}{\alpha+\beta}$ and $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. The pdf can be strictly increasing for $\alpha > 1, \beta = 1$, strictly decreasing for $\alpha = 1, \beta > 1$, U-shaped for $\alpha < 1, \beta < 1$, or unimodal for $\alpha > 1, \beta > 1$. For $\alpha = \beta$ we get symmetric distributions. $B(1, 1)$ is Uniform(0,1). It is also related to F distribution through a transformation.

Cauchy distribution - Symmetric, bell-shaped

$$f(x|\theta) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

$E[X] = \infty$. The parameter θ measures the median of the distribution. Ratio of normals is Cauchy and can be ill-behaved.

Lognormal distribution - if $\log X \sim \mathcal{N}(\mu, \sigma^2)$ then X is lognormal.

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \quad 0 < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

$E[X] = e^{\mu + \frac{\sigma^2}{2}}$ and $V[X] = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$.

Double Exponential distribution - Formed by reflecting the exponential distribution around its mean.

$$f(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

$E[X] = \mu$ and $V[X] = 2\sigma^2$. It is symmetric with fat tails.

3.3 Exponential Families

A family of distribution is called an exponential family if it can be expressed as

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right).$$

Here $h(x) \geq 0$ and $t_1(x), \ldots, t_k(x)$ are real-valued functions of the observation x, independent of θ, and $c(\theta) \geq 0$ where $w_1(\theta), \ldots, w_k(\theta)$ are real-values functions of the possibly vector-valued parameter θ, independent of x. Exponential family examples include - normal, gamma, beta, binomial, Poisson and negative binomial.

Theorem 3.1. *If X is a random variable with distribution from an exponential family then*

$$E\left[\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right] = -\frac{\partial}{\partial \theta_j}\log c(\boldsymbol{\theta});$$

$$Var\left[\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right] = -\frac{\partial^2}{\partial \theta_j^2}\log c(\boldsymbol{\theta}) - E\left[\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right]$$

Example 3.3. Let n be a positive integer and consider the *Binomial*(n, p) family with $0 < p < 1$. Then the pmf for this family, for $x = 0, \ldots, n$ and $0 < p < 1$, is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x}(1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right)x\right).$$

Here we define

$$h(x) = \begin{cases} \binom{n}{x} & x = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases} \qquad c(p) = (1-p)^n, \quad 0 < p < 1,$$

$$w_1(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1, \quad \text{and} \quad t_1(x) = x.$$

Then we have $f(x|p) = h(x)c(p)\exp\left[w_1(p)t_1(x) = x\right]$, which is of the form of exponential family with $k = 1$. This gives,

$$\frac{d}{dp}w_1(p) = \frac{d}{dp}\log\left(\frac{p}{1-p}\right) = \frac{1}{p(1-p)}$$

$$\frac{d}{dp}\log c(p) = \frac{d}{dp}n\log(1-p) = \frac{-n}{1-p}$$

and Thus from previous theorem we have $E[\frac{1}{p(1-p)}X] = \frac{n}{1-p}$ and hence $E[X] = np$. □

In general, the set of x values for which $f(x|\theta) > 0$ cannon depend on θ in an exponential family. Then entire definition of the distribution must be incorporated into the standard form. This is most easily accomplished by incorporating the range of x in the expression for $f(x|\theta)$ through the use of an indicator function.

Definition 3.1 (Indicator function)**.** *The indicator function of a set A, most often denoted by $I_A(x)$, is the function* $I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$. *It is also denoted by $I(x \in A)$.*

Example 3.4. Let $f(x|\mu, \sigma^2)$ be the $\mathcal{N}(\mu, \sigma^2)$ family of distributions, where $\theta = (\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$. Then,

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right).$$

Define $h(x) = 1$ for all x; $c(\theta) = c(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu}{2\sigma^2}\right)$, $-\infty < \mu < \infty$, $\sigma > 0$; $w_1(\mu, \sigma) = 1/\sigma^2$, $\sigma > 0$; $w_2(\mu, \sigma) = \mu/\sigma^2$, $\sigma > 0$; $t_1(x) = -x^2/2$; and $t_2(x) = x$. This gives it the standard form with $k = 2$. The parameter functions are defined only over the range of the parameter. Using an Indicator function we can write it as

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma) \exp[w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)] I_{(-\infty,\infty)}(x).$$

Since the indicator function is a function of only x, it can be incorporated into the function $h(x)$, showing that this distribution is of the general exponential family form. □

Example 3.5. The set of pdfs given y $f(x|\theta) = \frac{1}{\theta} \exp] \left(1 - \frac{x}{\theta}\right)$, $0 < \theta < x < \infty$, is not an exponential family even though we can write $f(x|\theta) = h(x)c(\theta) \exp(w(\theta)t(x))$, where $h(x) = 1$, $c(\theta) = \frac{1}{\theta}$, $w(\theta) = \frac{1}{\theta}$, and $t(x) = -x$. Writing the pdf with indicator functions makes this very clear. We have $f(x|\theta) = \frac{1}{\theta} \exp\left(1 - \frac{x}{\theta}\right) I_{[\theta,\infty]}(x)$. The indicator function can't be incorporated into any functions of the general form since it is not a function of x alone, and can't be expressed as an exponential. Thus, this is not an exponential family. □

An exponential family can be reparametrized as

$$f(x|\eta) = h(x)c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right),$$

with $h(x)$ and $t_i(x)$ the same as in the original parametrization. The set $\mathcal{H} = \{\eta = (\eta_1, \ldots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) dx < \infty\}$ is called the *natural parameter space* for the family. For the values of $\eta \in \mathcal{H}$, we must have $c^*(\eta) = \left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) dx\right]^{-1}$ to ensure that the pdf integrates to 1. Since, the original $f(x|\theta)$ is a distribution, the set $\{\eta = (w_1(\theta), \ldots, w_k(\theta)) : \theta \in \Theta\}$ must be a subset of the natural parameter space. But there may be other values of $\eta \in \mathcal{H}$ also. $\mathcal{H}$ is convex.

Example 3.6. To determine the natural parameter space of the normal family of distributions, we replace $w_i(\mu, \sigma)$ with η_i to obtain

$$f(x|\eta_1, \eta_2) = \sqrt{\frac{\eta_1}{2\pi}} \exp\left(-\frac{\eta_2^2}{2\eta_1}\right) \exp\left(-\frac{\eta_1 x^2}{2} + \eta_2 x\right).$$

The integral will be finite iff the coefficient on x^2 is negative. Thus the natural parameter space is $\{(\eta_1, \eta_2) : \eta_1 >, -\infty < \eta_2 < \infty\}$. We notice that $\eta_1 = 1/\sigma^2$ and $\eta_2 = \mu/\sigma^2$. □

Definition 3.2. *A curved exponential family is a family of densities of the same form as exponential family for which the dimension of the vector θ is equal to $d < k$. If $d = k$, the family is a fully exponential family.*

A normal family can become curved if we assume $\mu^2 = \sigma^2$. For the normal family the full exponential family would have parameter space $(\mu, \sigma^2) = R \times (0, \infty)$, while with the parameter space of the curved family $(\mu, \sigma^2) = (\mu, \mu^2)$ is a parabola.

Example 3.7. If $X_1, \dots, X_n$ is a sample from the $Poisson(\lambda)$ population, then the distribution of $\bar{X} = \sum_i X_i/n$ is approximately $\bar{X} \sim \mathcal{N}(\lambda, \lambda/n)$, a curved exponential family, justified by CLT. Most such CLT approximations will result in a curved normal family. Similarly, if $X_1, \dots, X_n$ are iid $Bernoulli(p)$, then $\bar{X} \sim \mathcal{N}(p, p(1-p)/n)$, approximately is a curved exponential family as well. □

The formula for mean and variance for exponential family is valid for curved exponential families as well. For a large number of data values from a population that has an exponential distribution, only k numbers that can be calculated from the data summarize all the information about θ that is in the data.

3.4 Location and Scale Families

The three families that make modeling convenient are location families, scale families, and location-scale families.

Theorem 3.2. *Let $f(x)$ be any pdf and let μ and $\sigma > 0$ be any given constants. Then the function $g(x|\mu, \sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$ is a pdf (non-negative and integrated to 1 for any μ and σ).*

Definition 3.3. *Let $f(x)$ be any pdf. Then the family of pdfs $f(x - \mu)$, indexed by the parameter μ, $-\infty < \mu < \infty$, is called the location family with standard pdf $f(x)$ and μ is called the location parameter of the family.*

Definition 3.4. *Let $f(x)$ be an pdf. Then for any $\sigma > 0$, the family of pdfs $\frac{1}{\sigma} f(\frac{x}{\sigma})$, indexed by the parameter σ, is called the scale family with standard pdf $f(x)$ and σ is called the scale parameter of the family.*

Definition 3.5. *Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, indexed by the parameter (μ, σ), is called the location-scale family with standard pdf $f(x)$; μ is called the location parameter and σ is called the scale parameter.*

Theorem 3.3. *Let $F(\dot{})$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$ iff there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

The random variable $Z = \frac{X-\mu}{\sigma}$ has pdf $f_Z(x) = f(z)$.

Theorem 3.4. *Let Z be a random variable with pdf $f(z)$. Suppose $E[Z]$ and $Var[Z]$ exist. If X is a random variable with pdf $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, then $E[X] = \sigma E[Z] + \mu$ and $Var[X] = \sigma^2 Var[Z]$. In particular, if $E[Z] = 0$ and $Var[Z] = 1$, then $E[X] = \mu$ and $Var[X] = \sigma^2$.*

3.5 Probability Inequalities and Identities

We start with some probabilistic inequalities which we intend to use later. These inequalities provide some kind of bound given the mean and variance of a distribution.

Theorem 3.5. *Markov's inequality: For a non-negative random variable X and a real number $k > 0$ with finite expected value $E[X] = \mu$,*

$$P[X \geq k\mu] \leq \frac{1}{k}.$$

To prove this we note, by the law of total probability

$$\begin{aligned} E[X] &= P[X < a]E[X|X < a] + P[X \geq a]E[X|X \geq a] \\ &\geq P[X \geq a]E[X|X \geq a] \\ &\geq aP[X \geq a] \end{aligned}$$

This gives the desired result with the change of variables $a \to k\mu$. Markov's inequality doesn't get close to the true value for compact distributions like Gaussian. It can be close to tight for discrete bi-variate distributions.

Theorem 3.6. *Chebyshev's Inequality: Let X be a random variable with finite expected value μ and variance σ^2. Then for every real number $k > 0$*

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}.$$

To prove this we use Markov's inequality. Let $Y = (X - E[X])^2$. Then Y is a non-negative valued random variable with expected value $E[Y] = Var[X]$. By Markov's inequality,

$$P[Y \geq a^2] \leq \frac{E[Y]}{a} = \frac{Var[X]}{a^2}.$$

Now we note that $Y \geq a^2$ is same even as $|X - E[X]| \geq a$. Changing the variable $a \to k\sigma$ proving the inequality. Chebyshev's inequality gives tighter bounds than Markov's inequality and is can be near tight for discrete tri-variate case.

The above two inequalities can be generalized and written as

$$\boxed{P[g(X) \geq r] \leq \frac{E[g(X)]}{r}},$$

where $g(X)$ is a non-negative function and $r > 0$. A similar inequality for the moment generating function is $P[X \geq a] \leq e^{-at}M_X(t)$, which requires the existence of the mgf.

Theorem 3.7. *Let $X_{\alpha,\beta}$ denote a gamma(α, β) random variable with pdf $f(x|\alpha, \beta)$, where $\alpha > 1$. Then for any constants a and b,*

$$P[a < X_{\alpha,\beta} < b] = \beta\left(f(a|\alpha, \beta) - f(b|\alpha, \beta)\right) + P[a < X_{\alpha-1,\beta} < b].$$

Theorem 3.8 (Stein's Lemma)**.** *Let $X \sim \mathcal{N}(\theta, \sigma^2)$, and let g be a differentiable function satisfying $E[|g'(X)|] < \infty$. Then $E[g(X)(X-\theta)] = \sigma^2 E[g'(X)]$.*

This makes calculation of higher-order moments quite easy.

Theorem 3.9. *Let $\mathcal{X}_p^2$ denote a chi squared random variable with p degrees of freedom. For any function $h(x)$,*

$$E[h(\mathcal{X}_p^2)] = pE\left[\frac{h(\mathcal{X}_{p+2}^2)}{\mathcal{X}_{p+2}^2}\right]$$

provided the expectations exist.

Theorem 3.10 (Hwang)**.** *Let $g(x)$ be a faunciton with $-\infty < E[g(X)] < \infty$ and $-\infty < g(-1) < \infty$. Then:*

1. *If $X \sim Poisson(\lambda)$, $E[\lambda g(X)] = E[Xg(X-1)]$.*
2. *If $X \sim$ negative binomial(r,p), $E[(1-p)g(X)] = E\left[\frac{X}{r+X-1}g(X-1)\right]$*

4 Multiple Random Variables

4.1 Joint and Marginal Distributions

Definition 4.1. *An n-dimensional random vector is a function from a sample space S into $\mathbb{R}^n$, n-dimensional Euclidean space.*

Definition 4.2. *Let (X, Y) be a discrete bivariate random vector. Then the function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ defined by $f(x, y) = P[X = x, Y = y]$ is called the joint probability mass function of (X, Y), to stress it it is sometimes denoted by $f_{X,Y}(x, y)$.*

Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) of the discrete random vector (X, Y). Then $g(X, Y)$ is itself a random variable and its expected value $E[g(X, Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x, y)f(x, y)$. The expectation operator continues to hold linearity properties. The joint pmf for any discrete bivariate random vector (X, Y) follows for any (x, y), $f(x, y) \geq 0$ and $\sum_{(x,y)\in\mathbb{R}^2} f(x, y) = 1$. It turns out that any non-negative function from $\mathbb{R}^2$ to $\mathbb{R}$ that is nonzero for at most a countable number of (x, y) pairs and sums to 1 is a joint pmf for some bivariate discrete random vector (X, Y). Thus, by defining $f(x, y)$, we can define a probability model for (X, Y) without ever working with the fundamental sample space S.

Theorem 4.1. *Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = P[X = x]$ and $f_Y(y) = P[Y = y]$, are given by $f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x, y)$, and $f_Y(y) = \sum_{x\in\mathbb{R}} f_{X,Y}(x, y)$.*

The marginal distributions of X and Y, described by the marginal pmfs $f_X(x)$ and $f_Y(y)$, do not completely describe the join distribution of X and Y.

Definition 4.3. *A function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is called a joint probability density function or joint pdf of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$, $P[(X, Y) \in A] = \int\int_A f(x, y)dxdy$.*

If $g(x, y)$ is a real valued function, then the expected value of $g(X, Y)$ is defined to be

$$E[g(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy.$$

The marginal probability density functions of X and Y are also defined as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y), \quad -\infty < x < \infty,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx, \quad -\infty < y < \infty.$$

Any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$ is the joint pdf of some continuous bivariate random vector (X, Y).

The joint probability distribution of (X,Y) can be completely described with the joint cdf rather than with the joint pdf. This is defined as $F(x,y) = P[X \le x, Y \le y]$ for all $(x,y) \in \mathbb{R}^2$. We have the following relationships for cdf and pdf

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)dtds \quad \text{and} \quad \frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y).$$

4.2 Conditional distributions and Independence

Definition 4.4. *Let (X,Y) be a discrete bivariate random vector with joint pmf $f(x,y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P[X=x] = f_X(x) > 0$, the conditional pmf of Y given that $X = x$ is the function of y denoted by $f(y|x) = P[Y = y|X = x] = \frac{f(x,y)}{f_X(x)}$. Similarly, for any y such that $P[Y = y] = f_Y(y) > 0$, the conditional pmf of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by $f(x|y) = P[X = x|Y = y] = \frac{f(x,y)}{f_X(x)}$.*

Definition 4.5. *Let (X,Y) be a continuous bivariate random vector with joint pdf $f(x,y)$ and marginals $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf of Y given that $X = x$ is the function of y denoted by $f(y|x) = \frac{f(x,y)}{f_X(x)}$. For any y such that $f_Y(y) > 0$, the conditional pdf of X given that $Y = y$ is the function of x denoted by $f(x|y) = \frac{f(x,y)}{F_Y(y)}$.*

If $g(Y)$ is a function of Y, then the conditional expected value of $g(Y)$ given that $X = x$ is denoted by $E[g(Y)|x]$ is $\sum_y g(y)f(y|x)$ for the discrete casea and $\int_{-\infty}^{\infty} g(y)f(y|x)dy$ for the continuous case.

Example 4.1. A random vector (X,Y) has joint distribution $f(x,y) = e^{-y}$, $0 < x < y < \infty$. If $x \le 0$, $f(x,y) = 0, \forall y$, so $f_X(x) = 0$. If $x > 0, f(x,y) > 0$ only if $y > x$. Thus,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{x}^{\infty} e^{-y}dy = e^{-x},$$

giving X an exponential distribution. For any $x > 0$ we have

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \begin{cases} \frac{e^{-y}}{e^{-x}} & \text{if } y > x \\ \frac{0}{e^{-x}} & \text{if } y \le x \end{cases} = \begin{cases} e^{x-y} & \text{if } y > x \\ 0 & \text{if } y \le x \end{cases}.$$

Thus, given $x = x$, Y has an exponential distribution, where x is the location parameter in the distribution of Y and $\beta = 1$ is the scale parameter. We can now calculate the conditional value of Y for given $X = x$ as $E[Y|X = x] = \int_x^{\infty} ye^{x-y}dy = 1 + x$. The conditional variance of Y given $X = x$ is given by $E[Y^2|x] - (E[Y|x])^2$. This can be calculated to 1.

The marginal distribution of Y is $gamma(2,1)$ which has $Var[Y] = 2$. Given the knowledge that $X = x$, reduces the variability of Y considerably. □

Joint distributions are sometimes defined by specifying the conditional $f(y|x)$ and the marginal $f_X(x)$ as $f(x,y) = f(y|x)F_X(x)$. When the knowledge of $X = x$ does not give any more information about Y we call the relationship of independence.

Definition 4.6. *Let (X, Y) be a bivariate random vector with joint distribution $f(x, y)$ and marginal distributions $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$, $f9x, y) = f_X(x)f_Y(y)$.*

If X and Y are independent then $f(y|x) = f_Y(y)$.

Theorem 4.2. *Let (X, Y) be a bivariate random vector with join distribution $f(x, y)$. Then X and Y are independent random variables iff there exists functions $g(x)$ and $h(y)$ such that, for every $x \in R$ and $y \in \mathbb{R}$, $f(x, y) = g(x)h(y)$.*

Theorem 4.3. *Let X and Y be independent random variables.*

- *For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P[X \in A, Y \in B] = P[X \in A]P[Y \in B]$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent variables*
- *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.*

Theorem 4.4. *Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by $M_Z(t) = M_X(t)M_Y(t)$.*

Theorem 4.5. *Let $X \sim \mathcal{N}(\mu,^2)$ and $Y \sim \mathcal{N}(\gamma, \tau^2)$ be independent normal random variables. Then the random variable $X = X + Y$ has a $\mathcal{N}(\mu + \gamma, \sigma^2 + \tau^2)$ distribution.*

If $f(x, y)$ is the joint pdf for some continuous random vector (X, Y), $f(x, y) = f_X(x)f_Y(y)$ may fail to hold on a set A of (x, y) values for which $\int\int_A dxdy = 0$. In such a case X and Y are still called independent random variables. This reflect the fact that two distributions that differ only on a set such as A define the same probability distribution for (X, Y).

Example 4.2. For $f(x, y) = e^{-x-y}$, $x > 0, y > 0$, is a pdf for two independent exponential random variables and satisfies $f(x, y) = f_X(x)f_Y(y)$. But, $f^*(x, y)$, which is equal to $f(x, y)$ except that $f^*(x, y) = 0$ if $x = y$, is also the pdf for two independent exponential random variables even though $f(x, y) = f_X(x)f_Y(y)$ is not true on the set $A = \{(x, x) : x > 0\}$. □

4.3 Bivariate Transformations

Let (X, Y) be a bivariate random vector with a known probability distribution. A new bivariate vector (U, V) defined by $U = g_1(X, Y)$ and $V = g_2(X, Y)$, where $g_1(x, y)$ and $g_2(x, y)$ are some specified functions. If B is any subset of $\mathbb{R}^2$, then $(U, V) \in B$ iff $(X, Y) \in A$ where $A = \{(x, y) : (g_1(x, y), g_2(x, y)) \in B\}$. Thus $P[(U, V) \in B] = P[(X, Y) \in A]$ and the probability distribution of (U, V) is completely determined by the probability distribution of (X, Y).

If (X, Y) is a discrete bivariate random vector, then there is only a countable set $\mathcal{A}$ of values for which the joint pmf of (X, Y) is positive. Define $\mathcal{B} = \{(u, v) : u = g_1(x, y), v = g_2(x, y)$ for $(x, y) \in \mathcal{A}\}$. Then $\mathcal{B}$ is the countable set of possible values for the discrete random vector (U, V). If for any $(u, v) \in \mathcal{B}$, A_{uv} is defined to be $\{(x, y) \in \mathcal{A} : g_1(x, y) = u$ and $g_2(x, y) = v\}$, then the joint pmf of (U, V), $f_{U,V}(u, v)$, cam be computed from the joint pmf of (X, Y) by $F_{U,V}(u, v) = P[U = v, V = v] = P[(X, Y) \in A_{uv}] = \sum_{(x,y)\in A_{uv}} f_{X,Y}(x, y)$.

Example 4.3. Let X and Y be independent Poisson random variables with parameters θ and λ, respectively. Thus the joint pmf of (X, Y) is $f_{X,Y}(x, y) = \frac{\theta^x e^{-\theta}}{x!}\frac{\lambda^y e^{-\lambda}}{y!}$, $x = 0, 1, 2, \ldots$, $y = 0, 1, 2, \ldots$. The set $\mathcal{A}$ is $\{(x, y) : x = 0, 1, 2, \ldots$ and $y = 0, 1, 2, \ldots\}$. Now define $U = X + Y$ and $V = Y$. That is, $g_1(x, y) = x + y$ and $g_2(x, y) = y$. The set of all possible (u, v) is $\mathcal{B} = \{(u, v) : v = 0, 1, 2, \ldots,$ and $u = v, v + 1, v + 2, \ldots\}$. For any $(u, v) \in \mathcal{B}$, the only (x, y) value satisfying the required construction is $A_{uv} = \{(u - v, v)\}$. Thus the joint pmf of (U, V) is

$$f_{U,V}(u, v) = f_{X,Y}(u - v, v) = \frac{\theta^{u-v}e^{-\theta}}{(u - v)!}\frac{\lambda^v e^{-\lambda}}{v!}, \quad v = 0, 1, 2, \ldots, \quad u = v, v + 1, v + 2, \ldots$$

The marginal pdf of U can be computed by noting that for any u, $f_{U,V}(u, v) > 0$ only for $v = 0, 1, \ldots, u$. This implies

$$f_U(u) = \sum_{v=0}^{u} \frac{\theta^{u-v}e^{-\theta}}{(u - v)!}\frac{\lambda^v e^{-\lambda}}{v!} = e^{-(\theta+\lambda)} \sum_{v=0}^{u} \frac{\theta^{u-v}}{(u - v)!}\frac{\lambda^v}{v!}, \quad u = 0, 1, 2, \ldots$$

This can be simplified to

$$f_U(u) = \frac{e^{-(\theta+\lambda)}}{u!} \sum_{v=0}^{u} \binom{u}{v} \lambda^v \theta^{u-v} = \frac{e^{-(\theta+\lambda)}}{u!}(\theta + \lambda)^u, \quad u = 0, 1, 2, \ldots$$

This is the pmf of a Poisson random variable with parameter $\theta + \lambda$. □

Theorem 4.6. *If $X \sim$ Poisson(θ) and $Y \sim$ Poisson(λ) and X and Y are independent, then $X + Y \sim$ Poisson$(\theta + \lambda)$.*

If (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of (U, V) can be expressed in terms of $f_{X,Y}(x, y)$. We have $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ and $\mathcal{B} = \{(u, v) : u = g_1(x, y)$ and $v = g_2(x, y), (x, y) \in \mathcal{A})\}$. We assume that the transformation $u = g_1(x, y)$ and $v = g_2(x, y)$ defines a one-to-one transformation of $\mathcal{A}$ onto $\mathcal{B}$. The transformation is onto because of the definition of $\mathcal{B}$. We are assuming that for each $(u, v) \in \mathcal{B}$ there is only one $(x, y) \in \mathcal{A}$ such that $(u, v) = (g_1(x, y), g_2(x, y))$. For such a one-to-one, onto transformation, we can solve the equations $u = g_1(x, y)$ and $v = g_2(x, y)$ for x and y in terms of u and v, denoting the inverse transform by $x = h_1(u, v)$ and $y = h_2(u, v)$. The Jacobian of the transformation is denoted by J, defined by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v} = \frac{\partial h_1(u, v)}{\partial u}\frac{\partial h_1(u, v)}{\partial v} - \frac{\partial h_2(u, v)}{\partial u}\frac{\partial h_2(u, v)}{\partial v},$$

Assuming J is not identically 0 on $\mathcal{B}$, the joint pdf of (U, V) is 0 outside the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|,$$

where $|J|$ is the absolute value of J.

Example 4.4. Let $X \sim beta(\alpha, \beta)$ and $Y \sim beta(\alpha+\beta, \gamma)$ be independent random variables. The joint pdf of (X, Y) is

$$f_{X,Y}(x, y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1 - x)^{\beta-1}\frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha + \beta)\Gamma(\gamma)}y^{\alpha+\beta-1}(1 - y)^{\gamma-1},$$

for $0 < x < 1$ and $0 < y < 1$. Consider $U = XY$ and $V = X$. The set of possible values of V is $0 < v < 1$. For a fixed value of $V = v$, U must be between 0 and v. Thus this transformation maps the set $\mathcal{A}$ onto the set $\mathcal{B} = \{(u,v) : 0 < u < v < 1\}$, with $x = h_1(u,v) = v$ and $y = h_2(u,v) = u/v$. The Jacobian is given by $J = -\frac{1}{v}$. Thus the joint pdf is

$$f_{U,V}(u,v) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} v^{\alpha-1}(1-v)^{\beta-1}\left(\frac{u}{v}\right)^{\alpha+\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{1}{v}, \quad 0 < u < v < 1.$$

The marginal distribution of $V = X$ is, of course, a $beta(\alpha, \beta)$ distribution. The distribution of U can be derived to be $beta(\alpha, \beta+\gamma)$ as follows:

$$\begin{aligned}
f_U(u) &= \int_u^1 f_{U,V}(u,v)dv \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}\int_u^1 \left(\frac{u}{v}-u\right)^{\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{u}{v^2}dv \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\int_u^1 y^{\beta-1}(1-y)^{\gamma-1}dy \quad \text{using } y = \left(\frac{u}{v}-u\right)\frac{1}{1-u} \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)} \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}, \quad 0 < u < 1 \\
&= beta(\alpha, \beta+\gamma)
\end{aligned}$$

□

Example 4.5. Let X and Y be independent, standard normal random variables. For the transformation $U = X + Y$ and $V = X - Y$, we have $U = g_1(X,Y)$ with $g_1(x,y) = x + y$ and $V = g_2(X,Y)$ with $g_2(x,y) = x - y$. The joint pdf of X and Y is, of course, $f_{X,Y}(x,y) = (2\pi)^{-1}\exp(-(x^2+y^2)/2)$, $-\infty < x, y < \infty$. So the set $\mathcal{A} = \mathbb{R}^2$. To determine the set $\mathcal{B}$ for which $f_{U,V}(u,v)$ is positive, we solve for u and v to get $x = h_1(u,v) = (u+v)/2$ and $y = h_2(u,v) = (u-v)/2$. For any $(u,v) \in \mathbb{R}^2$ there is an $(x,y) \in \mathcal{A}$ such that $u = x + y$ and $v = x - y$. So $\mathcal{B}$, the set of all possible (u,v) is $\mathbb{R}^2$. Since the solution is unique, this is also a one-to-one transformation. The Jacobian is easily calculated to be $-1/2$. Hence the joint distribution of U and V can be derived to:

$$\begin{aligned}
f_{U,V}(u,v) &= f_{X,Y}(h_1(u,v), h_2(u,v))|J| \\
&= \frac{1}{2\pi}e^{-\left(\frac{(u+v)}{2}\right)^2/2}e^{-\left(\frac{(u-v)}{2}\right)^2/2}\frac{1}{2} \\
&= \left(\frac{1}{\sqrt{4\pi}}e^{-u^2/4}\right)\left(\frac{1}{\sqrt{4\pi}}e^{-v^2/4}\right) \quad -\infty < u < \infty,\ -\infty < v < \infty
\end{aligned}$$

Since we have factored the joint pdf into a function of u and a function of v, U and V are independent. The marginal distribution of U and V both are $\mathcal{N}(0,2)$. This important fact, that the sums and the differences of independent normal random variables are independent normal random variables is true regardless of the means of X and Y, so long as $Var[X] = Var[Y]$. □

Theorem 4.7. *Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.*

Just as we generalized the univariate method to many-to-one functions, the same can be done in the bivariate case when the transformation of interest is not one-to-one. As before $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$. Suppose $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty, satisfies $P[(X, Y) \in A_0] = 0$. The transformation $U = g_1(X, Y)$ and $V = g_2(X, Y)$ is one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x = h_{1i}(u, v)$ and $y = h_{2i}(u, v)$. This ith inverse gives, for $(u, v) \in \mathcal{B}$, the unique $(x, y) \in A_i$ such that $(u, v) = (g_1(x, y), g_2(x, y))$. Let J_i denote the Jacobian computed from the ith inverse. Then assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the join pdf $f_{U,V}(u, v)$

$$f_{U,V}(u, v) = \sum_{i=1}^{k} f_{X,Y}(h_{1i}(u, v), h_{2i}(u, v))|J_i|.$$

Example 4.6. Let X and Y be independent $\mathcal{N}(0, 1)$ random variables. Consider the transformation $U = X/Y$ and $V = |Y|$. This transformation is not one-to-one. But if we restrict consideration to either positive or negative values of y, then the transformation is one-to-one. Let $A_1 = \{(x, y) : y > 0\}$, $A_2 = \{(x, y) : y < 0\}$, and $A_0 = \{(x, y) : y = 0\}$. A_0, A_1, A_2 form a partition of $\mathcal{A} = \mathbb{R}^2$ and $P[(X, Y) \in A_0] = P[Y = 0] = 0$. For either A_1 or A_2, if $(x, y) \in A_i$, $v = |y| > 0$, and for a fixed value of $v = |y|$, $u = x/y$ can be any real number since x can be any real number. Thus, $\mathcal{B} = \{(u, v) : v > 0\}$ is the image of both A_1 and A_2 under the transformation. The inverse transformation from $\mathcal{B}$ to A_1 and $\mathcal{B}$ to A_2 are given by $x = h_{11}(u, v) = uv$, $y = h_{21(u,v)=v}$, and $x = h_{12}(u, v) = -uv$ and $y = h_{22}(u, v) = -v$. The Jacobians from the two inverses are $J_1 = J_2 = v$. Using $f_{X,Y}(x, y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$, we have

$$\begin{aligned} f_{U,V}(u, v) &= \frac{1}{2\pi}e^{-\frac{(uv)^2+v^2}{2}}|v| + \frac{1}{2\pi}e^{-\frac{(-uv)^2+(-v)^2}{2}}|v| \\ &= \frac{v}{\pi}e^{-\frac{(1+u^2)v^2}{2}}, \quad -\infty < u < \infty, \quad 0 < v < \infty. \end{aligned}$$

This can be used to calculate the marginal of U as

$$\begin{aligned} f_U(u) &= \int_0^\infty \frac{v}{\pi}e^{-\frac{(1+u^2)v^2}{2}} \\ &= \frac{1}{2\pi}\int_0^\infty e^{-\frac{(1+u^2)z}{2}}dz \quad [z = v^2] \\ &= \frac{1}{2\pi}\frac{2}{1+u^2} \quad \text{[exponential kernel]} \\ &= \frac{1}{\pi(1+u^2)}, \quad -\infty < u < \infty \end{aligned}$$

Hence the ratio of two independent standard normal random variables is a Cauchy random variable. □

4.4 Hierarchical models and mixture distributions

The advantage of a hierarchical model is that complicated process may be modelled by a sequence of relatively simple models placed in a hierarchy. Dealing with hierarchy is similar to dealing with conditional and marginal distributions.

Example 4.7. An insect lays a large number of eggs, each surviving with probability p. On the average, how many eggs will survive? Number of eggs laid is a random variable, often taken to be $Poisson(\lambda)$. If we assume that each egg's survival is independent, then we have Bernoulli trials. Therefore, if we let X =number of survivors and Y =number of eggs laid, we have

$$X|Y \sim Binomial(Y, p) \quad Y \sim Poisson(\lambda),$$

a hierarchical model. We can the write,

$$\begin{aligned}
P[X = x] &= \sum_{y=0}^{\infty} P[X = x, Y = y] \\
&= \sum_{y=0}^{\infty} P[X = x|Y = y]P[Y = y] \\
&= \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{(} y - x) \frac{e^{-\lambda}\lambda^y}{y!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^{t}}{t!} \qquad [t = y - x] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} \quad \text{[Poisson kernel]} \\
&= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \\
&= Poisson(\lambda p)
\end{aligned}$$

Thus, any marginal inference on X is with respect to a $Poisson(\lambda p)$ distribution, with Y playing no part at all. To answer the original question we use $E[X] = \lambda p$, so, on an average λp eggs will survive. □

Theorem 4.8. *If X and Y are any two random variables, then provided that the expectations exist*

$$E[X] = E[E[X|Y]].$$

This is more precisely written as

$$E_X[X] = E_Y[E_{X|Y}[X|Y]].$$

Definition 4.7. *A random variable X is said to have a mixture distribution if the distribution of X depends on a quantity that also has a distribution.*

In general, hierarchical models lead to mixture distributions. They can also make calculation easier. A non central chi squared distribution with p degrees of freedom and non centrality parameter has a pdf of

$$f(x|\lambda, p) = \sum_{k=0}^{\infty} \frac{x^{\frac{p}{2}+k-1}e^{-\frac{x}{2}}\lambda^k e^{-\lambda}}{\Gamma(\frac{p}{2}+k)e^{\frac{p}{2}+k}k!}.$$

This is a mixture distribution, made up of $X|K \sim \mathcal{X}^2_{p+2K}$ and $K \sim Poisson(\lambda)$. The marginal distribution of X is then the above expression. Hence, $E[X] = E[E[X|K]] = E[p+2K] = p+2\lambda$. $Var[X]$ can be similarly calculated.

Theorem 4.9. *For any two random variables X and Y,*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]],$$

provided that the expectations exist.

Example 4.8. A generalization of binomial model is to allow the success probability to vary $X|P \sim Binomial(P)$, $i = 1, \dots, n$ and $P \sim beta(\alpha, \beta)$. We can easily see that $E[X] = E[E[X|P]] = E[nP] = n\frac{\alpha}{\alpha+\beta}$. Further, $Var[X] = Var[E[X|P]] + E[Var[X|P]]$ can be calculated to $n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$. □

4.5 Covariance and Correlation

Definition 4.8. *The covariance of X and Y is the number defined by $Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$. The correlation of X and Y is the number defined by $\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X\sigma_Y}$.*

$Cov[X,Y]$ can be any number but ρ_{XY} is a number between 1 and -1.

Theorem 4.10. *For any random variables X and Y, $Cov[X,Y] = E[XY] - \mu_X\mu_Y$.*

Theorem 4.11. *If X and Y are independent random variables, then $Cov[X,Y] = 0$ and $\rho_{XY} = 0$.*

However, if $Cov[X,Y] = 0$, it does not mean X and Y are independent. It is easy to find uncorrelated, dependent random variables. Covariance and correlation measure only a particular kind of linear relationship.

Theorem 4.12. *If X and Y are any two random variables and a and b are any two constants then*

$$Var[aX+bY] = a^2Var[X] + b^2Var[Y] + 2abCov[X,Y].$$

Theorem 4.13. *For any variables X and Y, $-1 \leq \rho_{XY} \leq 1$, and $|\rho_{XY}| = 1$ iff there exist numbers $a \neq 0$ and b such that $P[Y = aX+b] = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.*

If there is a linea $y = ax+b$ with $a \neq 0$, such that the values of (X,Y) have a high probability of being near this line, then the correlation between X and Y will be near 1 or -1. But if no such line exists, the correlation will be near 0. This is an intuitive notion of the linear relationship that is being measured by correlation.

Definition 4.9. *Let $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$ be five numbers. The bivariate normal pdf with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ is the bivariate pdf given by*

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

The marginal distribution of X is $\mathcal{N}(\mu_X, \sigma_X^2)$, the marginal distribution of Y is $\mathcal{N}(\mu_Y, \sigma_Y^2)$. The correlation between X and Y is $\rho_{XY} = \rho$. For any constants a and b, the distribution of $aX + bY$ is $\mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$. All the conditional distributions of Y given $X = x$ and of X given $Y = y$ are also normal distributions. The conditional distribution of Y given $X = x$ is $\mathcal{N}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2))$. All the normal marginal and conditional pdfs are derived from the starting point of bivariate normality. The derivation does not go in the opposite direction, in general. That is, marginal normality does not imply joint normality.

4.6 Multivariate Distributions

The random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a sample space that is a subset of $\mathcal{R}^n$. If $(X_1, \ldots, X_n)$ is a discrete random vector, then the joint pmf of $(X_1, \ldots, X_n)$ is the function defined by $f(\boldsymbol{x}) = f(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n]$ for each $(x_1, \ldots, x_n) \in \mathcal{R}^n$. Then for any $A \subset \mathcal{R}^n$, $P(\boldsymbol{X} \in A) = \sum_{\boldsymbol{x}\in A} f(\boldsymbol{x})$. If $(X_1, \ldots, X_n)$ is a continuous random vector, the joint pdf of $(X_1, \ldots, X_n)$ is a function $(x_1, \ldots, x_n)$ that satisfies $P(\boldsymbol{X} \in A) = \int \ldots \int_A f(\boldsymbol{x})d\boldsymbol{x} = \int \ldots \int_A f(x_1, \ldots, x_n)dx_1, \ldots dx_n$. These integrals are n-fold integrals with limits of integration set so that the integration is over all points $\boldsymbol{x} \in A$.

Let $g(\boldsymbol{x}) = g(x_1, \ldots, x_n)$ be a real valued function defined on the sample space of $\boldsymbol{X}$. Then $g(\boldsymbol{X})$ is a random variable and the expected value of $g(\boldsymbol{X})$ is

$$Eg(\boldsymbol{X}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x} \quad \textit{and} \quad Eg(\boldsymbol{X}) = \sum_{\boldsymbol{x}\in\mathcal{R}^n} g(\boldsymbol{x})f(\boldsymbol{x})$$

is the continuous and discrete cases, respectively. The marginal pdf or pmf of any subset of the coordinates of $(X_1, \ldots, X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates. Thus, for example, the marginal distribution of $(X_1, \ldots, X_k)$, the first k coordinates of $(X_1, \ldots, X_n)$, is given by the pdf or pmf

$$f(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)dx_{k+1} \ldots dx_n$$

or

$$f(x_1, \ldots, x_k) = \sum_{(x_{k+1},\ldots,x_n)\in\mathcal{R}^n} f(x_1, \ldots, x_n).$$

for every $(x_1, \ldots, x_k) \in \mathcal{R}^k$. The conditional pdf or pmf of a subset of the coordinates of $(X_1, \ldots, X_n)$ given the values of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates. Thus, for example, if $f(x_1, \dots, x_n) > 0$, the conditional pdf or pmf of $(X_{k+1}, \dots, X_n)$ given $X_1 = x_1, \dots, X_k = x_k$ is the function of $(x_{k+1}, \dots, x_n)$ defined by

$$f(x_{k+1}, \dots, x_n | x_1, \dots, x_k) = \frac{f(x_1, \dots, x_n)}{f(x_1, \dots, x_k)}.$$

Definition 4.10. *Let n and m be positive integers and let $p_1, \dots, p_n$ be numbers satisfying $0 \leq p_i \leq 1$, $i = 1, \dots, n$, and $\sum_{i=1}^n p_i = 1$. Then the random vector $(X_1, \dots, X_n)$ has a multinomial distribution with m trials and cell probabilities $p_1, \dots, p_n$ if the joint pmf of $(X_1, \dots, X_n)$ is*

$$f(x_1, \dots, x_n) = \frac{m!}{x_1! \dots x_n!} p_1^{x_1} \dots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

on the set of $(x_1, \dots, x_n)$ such that each x_i is a non-negative integer and $\sum_{i=1}^n x_i = m$.

The factor $m!/(x_1! \dots x_n!)$ is called a multinomial coefficient. It is the number of ways that m objects can be divided into n groups with x_1 in the first group, x_2 in the second group, ..., and x_n in the nth group. A generalization of the Binomial theorem is the multinomial theorem.

Theorem 4.14 (Multinomial Theorem)**.** *Let m and n be positive integers. Let $\mathcal{A}$ bet the set of vectors $\boldsymbol{x} = (x_1, , x_n)$ such that each x_i is a non-negative integer and $\sum_{i-1}^n x_i = m$. Then, for any real numbers $p_1, \dots, p_n$,*

$$(p_1 + \ldots + p_n)^m = \sum_{\boldsymbol{x} \in A} \frac{m!}{x_1! \dots x_n!} p_1^{x_1} \dots p_n^{x_n}.$$

X_i should have a *binomial*(m, p_i) distribution for each i. Given $X_n = x_n$, $(X_1, \dots, X_{n-1})$ has a multinomial distribution as well with $m - x_n$ trials and cell probabilities $p_1/(1 - p_n), \dots, p_{n-1}/(1 - p_n)$. In fact, the conditional distribution of any subset of the coordinates of $(X_1, \dots, X_n)$ gives the value of the rest of the coordinates is a multinomial distribution. This suggest that the coordinates of the vector $(X_1, \dots, X_n)$ are related. In particular they are negatively correlated as $Cov[X_i, X_j] = -mp_ip_j$. Thus, the negative correlation is greater for variables with higher success probabilities.

Definition 4.11. *let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be random vectors with join distribution $f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. Let $f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)$ denote the marginal distribution of $\boldsymbol{X}_i$. Then $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are called mutually independent random vectors if, for every $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$,*

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \prod_{i=1}^n f_{\boldsymbol{X}_i}(\boldsymbol{x}_i).$$

If the X_is are all one-dimensional, then $X_1, \dots, X_n$ are called mutually independent random variables.

Theorem 4.15. *Let $X_1, \dots, X_n$ be mutually independent random variables. Let $g_1, \dots, g_n$ be real valued functions such that $g_i(x_i)$ is a function only of x_i, $i = 1, \dots, n$. Then,*

$$E\left[\prod_{i=1}^n g_i(X_i)\right] = \prod_{i=1}^n E[g_i(X_i)].$$

Theorem 4.16. *Let $X_1, \ldots, X_n$ be mutually independent variables with mgfs $M_{x_1}(t), \ldots, M_{X_n}(t)$. Let $Z =_1 + \ldots + X_n$. Then the mfg of Z is $M_Z(t) = \prod_{i=1}^n M_{X_n}(t)$. In particular, if $X_1, \ldots, X_n$ all have the same distribution with mfg $M_X(t)$, then $M_Z(t) = (M_X(t))^n$.*

Example 4.9. Suppose $X_1, \ldots, X_n$ are mutually independent random variables and the distribution X_i is *gamma*(α_i, β). The mgf of a *gamma*(α, β) distribution is $M(t) = (1 - \beta t)^{-}$. Thus, if $Z = X_1 + \ldots + X_n$, the mfg of Z is

$$M_Z(t) = \prod_{i=1}^n M_{x_i}(t) = \prod_{i=1}^n (1 - \beta t)^{-\alpha_i t} = (1 - \beta t)^{-\sum_{i=1}^n \alpha_i t}.$$

This the mgf of a *gamma*$(\sum_{i=1}^n \alpha_i, \beta)$ distribution. □

Theorem 4.17. *Let $X_1, \ldots, X_n$ be mutually independent random variables with moment generation functions $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $a_1, , z_n$ bet $b_1, \ldots, b_n$ be fixed constants. Let $Z = (a_1 X_1 + b_1) + \ldots + (a_n X_n + b_n)$. Then the mgf of Z is $M_Z(t) = e^{t(\sum b_i)} \prod M_{X_i}(a_i t)$.*

Theorem 4.18. *Let $X_1, \ldots, X_n$ be mutually independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then*

$$Z = \sum_{i=1}^n (a_i X_i + b_i) \sim \mathcal{N}\left(\sum_{i=1}^n (a_i \mu_i + b_i), \sum_{i=1}^n a_i^2 \sigma_i^2 \right).$$

Theorem 4.19. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random vectors. Then $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ are mutually independent random vectors iff there exists functions $g_i(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, such that the joint distribution of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ can we written as*

$$f(\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^n g_i(\boldsymbol{x}_i).$$

Theorem 4.20. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors. Let $g_i(\boldsymbol{x}_i)$ be a function only of $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Then the random variables $U_i = g_i(\boldsymbol{X}_i)$, $i = 1, \ldots, n$ are mutually independent.*

Let $(X_1, \ldots, X_n)$ be a random vector with pdf $f_{\boldsymbol{X}}(x_1, \ldots, x_n)$. Let $\mathcal{A} = \boldsymbol{x} : f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$. Consider a new random vector $(U_1, \ldots, U_n)$, defined by $U_i = g_i(\boldsymbol{X})$. Suppose that $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0 which may be empty, satisfies $P[\boldsymbol{X} \in A_0] = 0$. The transformation $(U_1, \ldots, U_n) = (g_1(\boldsymbol{X}), \ldots, g_n(\boldsymbol{X}))$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, \ldots, k$. Then for each i, the inverse function from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $\boldsymbol{x}_1 = h_{1i}(u_1, \ldots, u_n), \boldsymbol{x}_2 = h_{2i}(u_1, \ldots, u_n), \ldots, \boldsymbol{x}_n = h_{ni}(u_1, \ldots, u_n)$. This ith inverse gives, for $(u_1, \ldots, u_n) \in \mathcal{B}$, the unique $(x_1, \ldots, x_n) \in A_i$ such that $(u_1, \ldots, u_n) = (g_1(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n))$. The J_i denote the Jacobian computed from the ith inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_n} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_n} \end{vmatrix},$$

the determinant of an $n \times n$ matrix. Assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_{\boldsymbol{U}}(u_1, \ldots, u_n)$, for $\boldsymbol{u} \in \mathcal{B}$:

$$f_{\boldsymbol{U}}(u_1, \ldots, u_n) = \sum_{i=1}^{k} f_{\boldsymbol{X}}(h_{1i}(u_1, \ldots, U_n), \ldots, h_{ni}(u_1, \ldots, u_n))|J_i|.$$

Example 4.10. If Y_1, Y_2, Y_3, Y_4 are mutually independent exponential ($\lambda = 1$) distributions. Define $X_1 = min(Y_1, Y_2, Y_3, Y_4)$, $X_2 =$ second smallest value of (Y_1, Y_2, Y_3, Y_4), and X_3 as the second largest and X_4 as the largest value. These variables are called ordered statistics. Let (X_1, X_2, X_3, X_4) have joint pdf

$$f_{\boldsymbol{X}}(x_1, x_2, x_3, x_4) = 24e^{-x_1-x_2-x_3-x_4}, 0 < x_1 < x_2 < x_3 < x_4 < \infty.$$

consider the transformation

$$U_1 = X_1,\ U_2 = X_2 - X_1,\ U_3 = X_3 - X_2,\ X_4 = X_4 - X_3.$$

The variables U_2, U_3, U_4 are called the spacings between the order statistic. The transformation maps the set $\mathcal{A}$ onto the set $\mathcal{B} = \boldsymbol{u} : 0 < u_i < \infty, i = 1, 2, 3, 4$. The transformation is one-to-one, so $k = 1$, and the inverse is

$$X_1 = U_1,\ X_2 = U_1 + U_2,\ X_3 = U_1 + U_2 + U + 3,\ X_4 = U_1 + U + 2 + U_3 + U_4.$$

The Jacobian of the inverse can be calculated to be 1. Thus we obtain

$$f_{\boldsymbol{U}}(u_1, u_2, u_3, u_4) = 24e^{-4u_1-3u_2-2u_3-u_4}$$

on $\mathcal{B}$. Form this the marginal pdf of U_1, U_2, U_3, U_4 can be calculated and turn out to be $f_U(u_i) = (5-i)e^{-(5-i)u_i}$, $0 < u_i$; that is $U_i \sim exponential(1/(5-i))$. These are mutually independent as well. The example shows that, for these exponential random variables the spacings between the order statistics are mutually independent and also have exponential distributions. □

4.7 Inequalities

4.7.1 Numerical Inequalities

Let a and b be any positive numbers, and let p and q be any positive numbers (necessarily greater than 1) satisfying $1/p + 1/q = 1$. Then $a^p/p + b^q/q \geq ab$, with equality iff $a^p = b^q$.

Theorem 4.21 (Holder's Inequality)**.** *Let X and Y be any two random variables, and let p and q satisfy $1/p + 1/q = 1$. Then*

$$|E[XY]| \leq E[|XY|] \leq E[|X|^p]^{\frac{1}{p}} E[|Y|^q]^{\frac{1}{q}}.$$

The most famous case is for $p = 2$.

Theorem 4.22 (Cauchy-Schwarz Inequality)**.** *For any two random variables X and Y,*

$$|E[XY]| \leq E[|XY|] \leq \sqrt{E[|X|^2]E[|Y|^2]}.$$

Theorem 4.23 (Liapounov's Inequality)**.** *If* $\infty > s > r > 1$ *then*

$$(E[|X|^r])^{\frac{1}{r}} \leq (E[|X|^s])^{\frac{1}{s}}.$$

Theorem 4.24 (Minkowski's Inequality)**.** *Let* X *and* Y *be any two random variables. Then for* $1 \leq p < \infty$*,*

$$(E|X+Y|^p)^{\frac{1}{p}} \leq (E|X|^p)^{\frac{1}{p}} + (E|Y|^p)^{\frac{1}{p}}$$

4.7.2 Functional Inequalities

Definition 4.12. *A function* $g(x)$ *is convex if* $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$*, for all* x *and* y*, and* $0 < \lambda < 1$*. The function* $g(x)$ *is concave if* $-g(x)$ *is convex.*

Theorem 4.25 (Jensen's Inequality)**.** *For any random variable* X*, if* $g(x)$ *is a convex function, then*

$$E[g(X)] \geq g(E[X]).$$

Equality holds iff, for every line $a + bx$ *that is tangent to* $g(x)$ *at* $x = E[X]$*,* $P[g(X) = a + bX] = 1$*.*

Jensen's inequality immediately implies that $EX^2 \geq (EX)^2$, and $E(1/X) \geq 1/EX$.

Theorem 4.26 (Covariance Inequality)**.** *Let* X *be any random variable and* $g(x)$ *and* $h(x)$ *any functions such that* $E[g(X)]$*,* $E[h(X)]$*, and* $E[g(X)h(X)]$ *exist. If* $g(x)$ *is an non-decreasing function and* $h(x)$ *is an non-increasing function, then* $E[g(X)h(X)] \leq E[g(X)]E[h(X)]$*. If* $g(X)$ *and* $h(X)$ *are either both non-decreasing or both non-increasing then the direction is reversed.*

5 Properties of a Random Sample

We describe the model for random sampling here.

Definition 5.1. *(Sampling from an infinite population) The random variables $X_1, \ldots, X_n$ are called random sample of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called independent and identically distributed random variables with pdf or pmf f(x). This is commonly abbreviated to iid random variables.*

In the n observations the value of one observation has no effect on any other observations. From def. 5.1, the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)(x_n) = \prod_{i=1}^{n} f(x_i).$$

If the pdf or pmf is parametrized by a parameter θ, $f(x|\theta)$, the joint distribution is

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

If the true parameter is unknown, then a random sample from this population has a joint distribution in the above form, and can be used to to study random sample behavior under different populations, parametrized by θ. When sampling with replacement from a finite population the sampling is iid. When sampling without replacement (also called simple random sampling) there is no independence (X_i is slightly negatively correlated to X_j, for all $i \neq j$) but they are identically distributed, i.e. the marginal distribution of each sampled variable is the same. However, if the population size is large compared to sample size n, the variables are nearly independent.

Sampling can also be ordered or unordered. A generalization of iid random variables is exchangeable random variables (deFinetti), i.e. where the order is immaterial.

Definition 5.2. *The random variables $X_1, \ldots, X_n$ are exchangeable if any permutation of any subset of them of size k ($k \leq n$) has the same distribution.*

An infinite sequence of exchangeable random variables is a mixture of iid random variables. An iid sequence is exchangeable but a finite sequence of exchangeable random variables may not be iid.

5.1 Statistic of a Random Sample

Definition 5.3. *Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.*

A statistic can't be a function of a parameter. Two statistics that are often used and provide good summaries of the sample are - sample mean ($\bar{X} = \frac{1}{n}\sum X_i$), sample variance ($S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$). Notice that sample mean ($\bar{X}$) and sample variance (S^2) are random variables and different from the scalars - population mean, μ, and population variance, σ^2.

For random samples $X_1, \ldots, X_n$ and any reasonable function $g(x)$ (finite mean and variance of $g(X)$) writing $E\left[\sum g(X_i)\right] = nE[g(X)]$ does not need independence but only identical distribution. Whereas, proclaiming $Var\left[\sum g(X_i)\right] = nVar[g(X)]$ needs both independence and identical distribution assumption.

Example 5.1. The sample mean is defined as $\bar{X} = \frac{1}{n}\sum X_i$ and sample variance is defined as $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$. The population mean and standard deviation is given as μ and $\sigma < \infty$. Find $E[\bar{X}]$, $Var[\bar{X}]$ and $E[S^2]$.

$$E[\bar{X}] \underset{def.}{=} E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \underset{linearity}{=} \frac{1}{n}\sum_{i=1}^{n} E[X_i] \underset{identical}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

$$Var[\bar{X}] \underset{def.}{=} Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \underset{linearity}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] + \frac{2}{n^2}\underset{independence}{\sum_{i\neq j}\cancelto{0}{Var[X_iX_j]}} = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \underset{identical}{=} \frac{\sigma^2}{n}$$

$$\begin{aligned}
E[S^2] &\underset{def.}{=} E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \underset{linearity}{=} \frac{1}{n-1}\sum_{i=1}^{n} E[X_i^2 + \bar{X}^2 - 2\bar{X}X_i] \\
&\underset{independence}{=} \frac{1}{n-1}\sum_{i=1}^{n}\Big(E[X_i^2] + E[\bar{X}^2] - 2E[\bar{X}]E[X_i]\Big) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\Big(E[X_i^2] - E[\bar{X}^2]\Big) \underset{identical}{=} \frac{n}{n-1}\Big(E[X_i^2] - E[\bar{X}^2]\Big) \\
&= \frac{n}{n-1}\left(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right) = \sigma^2.
\end{aligned}$$

The last step in the third expression use the fact that $Var[Y] = E[Y^2] - E[Y]^2$. This shows that, $\bar{X}$ is an unbiased estimator or μ and S^2 is an unbiased estimator of σ^2 (hence, justifying the $n-1$ in the denominator of the definition) - relating the sample properties to the properties of the population. □

5.2 Distribution of Sample Statistics

After analyzing the two important statistic, we now turn to the sampling distribution of $\bar{X}$. For finite sample, the distribution of the mean will depend on the underlying distribution. If the underlying distribution does not have finite variance the variance may not decrease as the sample size increases (e.g. Cauchy distribution). If the underlying distribution is normal we can make very precise comments about the sample statistics.

Theorem 5.1. *Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and let $\bar{X} = \frac{1}{n}\sum X_i$ and $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$. Then*

- $\bar{X}$ *and* S^2 *are independent random variables.*
- $\bar{X}$ *has a* $\mathcal{N}(\mu, \sigma^2/n)$ *distribution.*
- $(n-1)S^2/\sigma^2$ *has a chi squared distribution with* $n-1$ *degrees of freedom,* $\mathcal{X}^2_{n-1}$.

For normal variables, 0 covariance and Independence are equivalent for learn functions of these random variables. If Z is a $\mathcal{N}(0,1)$, standard normal variable, then $Z^2 \sim \mathcal{X}^2_1$. Further, if $X_1, \ldots, X_n$ are independent and $X_i \sim \mathcal{X}^2_{p_i}$, then $X_1 + \ldots + X_n \sim \mathcal{X}^2_{p1+\ldots+p_n}$.

If $X_1, \ldots, X_n$ are a random sample from a $\mathcal{N}(\mu, \sigma^2)$, then the quantity

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable. Most of the times σ is unknown along with μ. We can then look at next obvious thing

$$\frac{\bar{X}-\mu}{S/\sqrt{n}},$$

a quantity that could be used as a basis for inference about μ when σ is unknown. The distribution can be understood using the following manipulation.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}.$$

The numerator is a standard normal random variable, and the denominator is $\sqrt{\mathcal{X}^2_{n-1}/(n-1)}$, independent of the numerator. The distribution $U/\sqrt{V/p}$, where U is standard normal, V is $\mathcal{X}^2_p$, and U and V are independent gives us the Student's t distribution. Notice that if $p = 1$, then it is Cauchy distribution, with occurs for samples of size 2. Student's t distribution does not have moment of all order. In fact, if there are p degrees of freedom, then there are only $p-1$ moments, i.e. t_1 has no mean, t_2 has no variance, etc. For a random variable $T_p \sim t_p$ we have $E[T_p] = 0$, if $p > 1$, and $Var[T_p] = \frac{p}{p-2}$, if $p > 2$.

Another important derived distribution is the F distribution, which arises naturally as the distribution of a ratio of variances, even when the populations are not normal. For n random samples from $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$ population, and m random samples from $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$ population, the information about ${}^2_X/\sigma^2_Y$ can be derived from S^2_X/S^2_Y. The F-distribution represents it using $n-1$ and $m-1$ degrees of freedom via

$$F \sim \frac{S^2_X/S^2_Y}{\sigma^2_X/\sigma^2_Y} = \frac{S^2_X/\sigma^2_X}{S^2_Y/\sigma^2_Y}.$$

$F \sim \frac{U/p}{V/q}$, where $U \sim \mathcal{X}^2_p$ is independent of $V \sim \mathcal{X}^2_q$. Also, $E[F_{n-1,m-1}] = \frac{m-1}{m-3}$, for $m > 3$. This ratio goes to 1 for large m. If $X \sim F_{p,q}$ then $1/X \sim F_{q,p}$. Also, if $X \sim t_q$, then $X^2 \sim F_{1,q}$.

Definition 5.4. *The order statistics of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$. The order statistics are random variables that satisfy $X_1 \leq \ldots \leq X_{(n)}$. Specifically, $X_{(1)} = \underset{1 \leq i \leq n}{min} X_i, \ldots, X_{(n)} = \underset{1 \leq i \leq n}{max} X_i$.*

The sample range, $R = X_{(n)} - X_{(1)}$ is a measure of dispersion in population. The sample median, M

$$M = \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd} \\ (X_{n/2} + X_{n/2+1})/2 & \text{if } n \text{ is even} \end{cases}$$

is a measure of central tendency like mean. Median is less affected by extreme observations. Other robust statistic is α-trimmed mean defined based on order statistic. Sample percentile can be, similarly, defined based on order statistics. The lower and upper quartiles are also used to determine interquartile range, as another measure of dispersion. The usual technique to derive these distributions is to construct the cumulative distribution of the statistics and then take a derivative to get the pdf.

5.3 Convergence Concepts

The notion of an infinite sample is a theoretical artifact, but often provide us with some useful approximations for the finite sample case. We are mainly interested in the behavior of $\bar{X}_n$, the mean of n observations, as $n \to \infty$.

5.3.1 Convergence in Probability

Definition 5.5. *A sequence of random variables $X_1, X_2, \ldots$ (not necessarily iid), converges in probability to a random variable X if, for every $\epsilon > 0$,*

$$\lim_{n \to \infty} P[|X_n - X| < \epsilon] = 1.$$

Intuitively, the distribution of X_n changes as the subscript changes, and the convergence in probability describe different ways in which the distribution of X_n converges to some limiting distribution as $n \to \infty$.

Theorem 5.2. *Weak Law of Large Numbers: Let $X_1, X_2, \ldots$ be iid random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. For $\bar{X}_n = \frac{1}{n} \sum X_i$, for every $\epsilon > 0$,*

$$\lim_{n \to \infty} P[|\bar{X}_n - \mu| < \epsilon] = 1$$

i.e., $\bar{X}_n$ converges in probability to μ.

Chebyshev's inequality can be used to prove WLLN. The property that a sequence of the 'same' sample quantity approaches a constant as $n \to \infty$, is known as consistency. Similarly, we can prove a WLLN for the sample variance S^2, provided $Var[S_n^2] \to 0$ as $n \to \infty$. This idea can be extended to functions of random variables as well, provided the functions are continuous. That is, suppose that $X_1, X_2, \dots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \dots$ converges in probability to $h(X)$. Utilizing this we can say that if S_n^2 is a consistent estimator of σ^2, then the sample deviation S_n is a consistent estimator of σ. Note that S_n is, in fact, a biased estimator of σ ($E[S] \leq \sigma$ due to convexity of the function) but the bias disappears asymptotically.

5.3.2 Almost Sure Convergence

Definition 5.6. *A sequence of random variables $X_1, X_2, \dots$, converges almost surely to a random variable X if, for every $\epsilon > 0$,*

$$p\left[\lim_{n\to\infty} |X_n - X| < \epsilon\right] = 1.$$

A random variable is a real-valued function defined on a sample space $\mathcal{S}$. If this sample space has elements $s \in \mathcal{S}$, then $X_n(s)$ and $X(s)$ are functions defined on $\mathcal{S}$. The definition above states that X_n converges to X almost surely if the functions $X_n(s)$ converges to $X(s)$ for all $s \in \mathcal{S}$ except perhaps for $s \in \mathcal{W}$, where $\mathcal{W} \subset \mathcal{S}$ and $P[\mathcal{W}] = 0$. This is a much stronger condition than convergence in probability.

Example 5.2. For uniform distribution on $[0, 1]$, $X_n(s) = s + s^n$ converges to $X(s) = s$ as $n \to \infty$ almost surely, because the convergence occurs on the set $[0, 1)$ and $P[[0, 1)] = 1$.

Define the sequence, $X_1(s) = s + \mathbf{I}_{[0,1]}(s)$, $X_2(s) = s + \mathbf{I}_{[0,\frac{1}{2}]}(s)$, $X_2(s) = s + \mathbf{I}_{[\frac{1}{2},1]}(s)$, $X_4(s) = s + \mathbf{I}_{[0,\frac{1}{3}]}(s)$, $X_5(s) = s + \mathbf{I}_{[\frac{1}{3},\frac{2}{3}]}$, $X_6(s) = s + \mathbf{I}_{[\frac{2}{3},1]} \dots$. Also, let $X(s) = s$. It is clear that X_n converges to X in probability - as $n \to \infty$, $P[|X_n - X| \geq \epsilon]$ is equal to the probability of an interval of s values whose length is going to be 0. However, X_n does not converge to X almost surely. In fact, there is no value of $s \in \mathcal{S}$ for which $X_n(s) \to s$. For every s, the value of $X_n(s)$ alternates between s and $s + 1$ infinitely often. □

Convergence almost surely, being the strong criterion, implies convergence in probability. The converse, is not true. However, if a sequence converges in probability, it is possible to find a subsequence that converges almost surely.

Theorem 5.3. *Strong Law of Large Numbers: Let $X_1, X_2, \dots$ be iid random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, and define $\bar{X}_n = \frac{1}{n}\sum X_i$. Then, for every $\epsilon > 0$,*

$$P\left[\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right] = 1$$

that is, $\bar{X}_n$ converges almost surely to μ.

Both the weak and strong laws of large numbers hold without the assumption of finite variance. The only moment condition needed is that $E[X_i] < \infty$.

5.3.3 Convergence in Distribution

Definition 5.7. *A sequence of random variables, $X_1, X_2, \ldots$, converges in distribution to a random variable X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

Theorem 5.4. *If the sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable X, the sequence also converges in distribution to X.*

This means that convergence in distribution is also implied by almost sure convergence.

Theorem 5.5. *The sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a constant μ if and only if the sequence also converges in distribution to μ.*

Theorem 5.6. *Central Limit Theorem: Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mean, μ, and variance, σ^2, are finite. Then, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.*

With just independent and finite variance we end up with normality! This normality comes from sums of 'small' finite variance, independent disturbances. However, the goodness of this approximation is a function of the original distribution.

Theorem 5.7. *Slutsky's Theorem: If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then $Y_n X_n \to aX$ in distribution, and $X_n + Y_n \to X + a$ in distribution.*

Using Slutsky's theorem we can state as $n \to \infty$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathcal{N}(0, 1).$$

5.3.4 The Delta Method

When we are interested in the the function of the the random variable we utilize the delta method to obtain reasonable approximate answers. Using Taylor series approximations for the mean and variance, we get the generalization of the Central Limit Theorem, called the Delta method. Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $T = (T_1, \ldots, T_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. For a differential function $g(T)$ we define

$$g_i'(\theta) = \frac{\partial}{\partial t_i} g(t)\bigg|_{t_1=\theta_1,\ldots,t_k=\theta_k}.$$

Taylor expansion gives

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i).$$

Taking the expectation we get

$$E_\theta[g(T)] \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta) E_\theta[T_i - \theta_i] = g(\theta).$$

We can also take variance on both side to get

$$\begin{aligned}
Var_\theta[g(T)] &\approx E_\theta[(g(T) - g(\theta))^2] \\
&\approx E_\theta\left[\left(\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i)\right)^2\right] \\
&= \sum_{i=1}^{k}[g_i'(\theta)]^2 Var_\theta[T_i] + 2\sum_{i>j} g_i'(\theta)g_j'(\theta)Cov_\theta[T_i, T_j].
\end{aligned}$$

Theorem 5.8. *Delta Method: Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2[g'(\theta)]^2),$$

in distribution.

To prove Deta Method one does Taylor's expansion and applies Slutsky's theorem.

Example 5.3. X is a random variable with $E_\mu[X] = \mu \neq 0$. If we want to estimate the function $g(\mu)$, a first order approximation will get $g(X) = g(\mu) + g'(\mu)(X - \mu)$. If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$E_\mu[g(X)] \approx g(\mu)$$

$$Var_\mu[g(X)] \approx [g'(\mu)]^2 Var_\mu[X].$$

For $g(\mu) = 1/\mu$, we estimate $1/\mu$ with $1/X$, can we can say

$$E_\mu\left[\frac{1}{X}\right] \approx \frac{1}{\mu}$$

$$Var_\mu\left[\frac{1}{X}\right] \approx \left(\frac{1}{\mu}\right)^4 Var_\mu[X].$$

For the mean of the random sample $\bar{X}$ we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathcal{N}\left(0, \left(\frac{1}{\mu}\right)^4 Var_\mu[X]\right)$$

in distribution. Further, if we don't know the variance of X, to use the above approximation requires an estimate, say S^2. Moreover, μ in the term $1/\mu$ is also unknown. We can estimate everything, which gives us the approximate variance

$$\hat{Var}\left[\frac{1}{\bar{X}}\right] \approx \left(\frac{1}{\bar{X}}\right)^4 S^2.$$

Furthermore, as both $\bar{X}$ and S^2 are consistent estimators, we can again apply Slutsky's theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \mathcal{N}(0, 1)$$

in distribution. □

This way of writing mean difference in the numerator and estimated standard deviation in the denominator is the standard way to estimate any parameters in the limiting distribution.

In the event that $g'(\mu) = 0$ and we are interested in estimating the variance we take one more term in the Taylor expansion to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n -)^2 + \ldots$$

This, using the fact $g'(\theta) = 0$, gives us

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \ldots$$

The square of a $\mathcal{N}(0,1)$ is a $\mathcal{X}_1^2$ which implies

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \to \mathcal{X}_1^2$$

in distribution.

Theorem 5.9. *Second-order Delta Method: Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then*

$$\frac{g(Y_n) - g(\theta)}{\frac{\sigma^2}{n}\frac{g''(\theta)}{2}} \to \mathcal{X}_1^2,$$

in distribution.

Example 5.4. Moments of a ratio estimator: Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. The parametric function to estimated is $g(\mu_x, \mu_y) = \frac{\mu_X}{\mu_Y}$. We have,

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y} \text{ and } \frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}.$$

The first order Taylor approximation gives

$$E\left[\frac{X}{Y}\right] \approx \frac{\mu_X}{\mu_Y}$$

and

$$Var\left[\frac{X}{Y}\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{Var[X]}{\mu_X^2} + \frac{Var[Y]}{\mu_Y^2} - 2\frac{Cov[X,Y]}{\mu_X \mu_Y}\right).$$

□

Suppose the vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_p)$ has mean $\mu = (\mu_1, \ldots, \mu_p)$ and covariances $Cov[X_i, X_j] = \sigma_{ij}$, and we observe an independent random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and calculate the means $\bar{X}_i = \sum_{k=1}^{n} X_{ik}, i = 1, \ldots, p$. For a function $g(\mathbf{x}) = g(x_1, \ldots, x_p)$

$$g(\bar{x}_1, \ldots, \bar{x}_p) = g(\mu_1, \ldots, \mu_p) + \sum_{k=1}^{p} g'_k(\mathbf{x})(\bar{x}_k - \mu_k).$$

Theorem 5.10. *Multivariate Delta Method: Let $X_1, \ldots, X_n$ be a random sample with $E[X_i] = \mu_i$ and $Cov[X_{ik}, X_{jk}] = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\mu = (\mu_1, \ldots, \mu_p)$ for which $\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g(\mu)}{\partial \mu_i} \frac{\partial g(\mu)}{\partial \mu_j} > 0$,*

$$\sqrt{n}\frac{g(\bar{X}_1, , \bar{X}_s) - g(\mu_1, , \mu_p)}{\tau} \to \mathcal{N}(0, 1)$$

in distribution.

5.4 Generating a Random Sample

Simulation involves generating random variables we need, and then, we use a version of the Law of Large Numbers to validate the simulation approximation. Generally, one has uniform random variable generator available and a transformation is needed to achieve desired distribution. There are direct (where closed form function $g(u)$ exist) and indirect methods (where the closed form does not exist) to do it. For direct method one can use the Probability Integral Transform.

$$F_Y^{-1}(u) = y \equiv u = \int_{-\infty}^{y} f_y(t)dt.$$

Hence the inverse transformation solves the problem. If Y is a continuous random variable with cdf F_Y, then $F_Y^{-1}(U)$, where $U \sim uniform(0, 1)$ has a distribution F_Y. This method works great for discrete random variables, in general. There are some fancy transformation that can work for continuous variables, one being the Box-Muller algorithm to generate standard normal variables.

Theorem 5.11. *Box-Muller algorithm: Let U_1 and U_2 be two independent $uniform(0, 1)$ random variables. Set, $R = \sqrt{-2\log U_1}$ and $\theta = 2\pi U_2$. Then, $X = R\cos\theta$ and $Y = R\sin\theta$ are independent $\mathcal{N}(0, 1)$ random variables.*

When no easily found direct transformation is available we require methods like Accept/Reject algorithm, Markov Chain Monte Carlo (MCMC) methods (e.g. Gibbs sampler and the Metropolis Algorithm).

6 Principles of data Reduction

To summarize a sample using a few key features is generally the aim. let $\mathbf{X}$ denote the random variables $X_1, \ldots, X_n$ and $\mathbf{x}$ denotes the sample $x_1, \ldots, x_n$. Any statistic, $T(\mathbf{X})$, defines a form of data reduction or data summary. Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t : t = T(\mathbf{x})\text{for some} x \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Then $T(\mathbf{x})$ partitions the sample space into sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{\mathbf{X} : T(\mathbf{x} = t)\}$. The statistic summarizes the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $T(\mathbf{x}) = t$ or, equivalently, $\mathbf{x} \in A_t$. There are various advantages and consequences of data reduction. The aim is to not discard important information about the unknown parameter θ and throw away irrelevant information as far as gaining knowledge about θ is concerned.

6.1 The sufficiency Principle

A sufficient statistic for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample.

Definition 6.1. *Sufficiency Principle: If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$.*

Definition 6.2. *A statistic $T(\mathbf{X})$ is sufficient statistic for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.*

Theorem 6.1. *If $p(x|\theta)$ is the joint pdf/pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf/pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is independent of θ.*

It turns out that outside of exponential family, it is rare to have a sufficient statistic of smaller dimension than the size of the sample.

Theorem 6.2. *Factorization Theorem: Let $f(\mathbf{x}|\theta)$ denote the joint pdf/pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ iff there exist function $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,*

$$f(\mathbf{x}|\theta) = g(T(x|\theta))h(\mathbf{x}).$$

Example 6.1. Uniform sufficient statistic: let $X_1, \ldots, X_n$ be iid observations from the discrete uniform distribution in the integer set $\mathcal{N}_\theta = \{1, \ldots, \theta\}$, where θ is a positive integer. The pmf of X_i is $f(x|\theta) = \theta^{-1}\mathbb{I}_{\mathcal{N}_\theta}(x)$. Thus the joint pmf of $X_1, \ldots, X_n$ is $f(\mathbf{x}|\theta) = \prod \theta^{-1}\mathbb{I}_{\mathcal{N}_\theta}(x_i) = \theta^{-n}\prod \mathbb{I}_{\mathcal{N}_\theta}(x_i)$. Defining $T(\mathbf{x}) = \underset{i}{max}\ x_i$, we see that

$$\prod_{i=1}^{n} \mathbb{I}_{\mathcal{N}_\theta}(x_i) = \mathbb{I}_{\mathcal{N}_\theta}(T(\mathbf{x}))\left(\prod_{i=1}^{n} \mathbb{I}_{\mathcal{N}}(x_i)\right).$$

Thus we have the factorization

$$f(\mathbf{x}|\theta) = \left(\theta^{-n}\mathbb{I}_{\mathcal{N}_\theta}(T(\mathbf{x}))\right)\left(\prod_{i=1}^{n} \mathbb{I}_{\mathcal{N}}(x_i)\right)$$

The first term is a function $g(T(\mathbf{x});\theta)$, that depends on the sufficient statistic $T(\mathbf{x})$ and parameter θ, while second term is a function $h(\mathbf{x})$, independent of θ. □

There are cases were sufficient statistic can be a vector, e.g. Normal sufficient statistic where both mean and variance are unknown. The definition of a sufficient statistic is model-dependent. An assumption of normality makes the first two moment sufficient, but for a general distribution that may not be the case. For an exponential family distributions it is easy to find sufficient statistic using the Factorization theorem.

6.1.1 Minimal sufficient statistics

By default, the complete sample $\mathbf{X}$, is a sufficient statistic. Similarly, any one to one function of a sufficient statistic is a sufficient statistic. We need some kind of minimal sufficiency here.

Definition 6.3. *A sufficient statistic $T(\mathbf{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$, i.e. if $T'(\mathbf{x}) = T'(\mathbf{y})$, then $T(\mathbf{x}) = T(\mathbf{y})$.*

In terms of the partition set, a minimal sufficient statistic is the coarsest possible partition.

Theorem 6.3. *Let $f(\mathbf{x}|\theta)$ bet the pmf/pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is independent of θ iff $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ if a minimal sufficient statistic for θ.*

Example 6.2. Suppose $X_1, \ldots X_n$ are iid uniform random variables on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Then the joint pdf of X is $f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \mathbb{I}_{\theta<x_i<\theta+1}(x_i)$. This can be written as $f(\mathbf{x}|\theta) = \mathbb{I}_{\underset{i}{max}(x_i)-1<\theta<\underset{i}{min}(x_i)}(\mathbf{x})$. Thus, for two sample points $\mathbf{x}$ and y, the numerator and denominator or the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ will be positive for the same values of θ iff $\underset{i}{min}(x_i) = \underset{i}{min}(y_i)$ and $\underset{i}{max}(x_i) = \underset{i}{max}(y_i)$. If the minimum and maximum are same the ratio is 1. Thus we have the minimal sufficient statistic as $T(\mathbf{X}) = (\underset{i}{min}(X_i), \underset{i}{max}(X_i))$. □

6.1.2 Ancillary statistics

Definition 6.4. *A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic.*

As an example, let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Let $X_{(1)} < \ldots < X_{(n)}$ be the order statistics from the sample. The minimal sufficient statistic is $(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$. The range statistic, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic as it does not depend on θ, since it is a location parameter. Certainly, the ancillary statistic and the minimal sufficient statistic are not independent.

Scale parameters also have certain kind of ancillary statistics. Let $X_1, \ldots, X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma)$, $\sigma > 0$. Then any statistic that depends on the sample only through the $n-1$ values $X_1/X_n, \ldots, X_{n-1}/X_n$ is an ancillary statistic, e.g. $\sum X_i/X_n$. In particular for X_1 and X_2 iid normal $\mathcal{N}(0,\sigma^2)$ observations, X_1/X_2 has a distribution that is the same for every value of σ, in fact it is Cauchy distribution.

Example 6.3. Ancillary precision: Let X_1 and X_2 be iid observations from the discrete distribution that satisfies. $P_\theta[X = \theta] = P_\theta[X = \theta + 1] = P_\theta[X = \theta + 2] = \frac{1}{3}$, where θ is an unknown integer parameter. Let $X_{(1)} < X_{(2)}$ be the order statistics for the sample. The minimal sufficient statistic is (R, M) where $R = X_{(2)} - X_{(1)}$ and $M = (X_{(1)} + X_{(2)})/2$. R is an ancillary statistic, but can give information about θ. Say $M = m$, an integer is given. It means either $\theta = m$, or $\theta = m - 1$ or $\theta = m - 2$, that is all three values of θ are possible. Now suppose we are given that $R = 2$. Then it must be the case that $X_{(1)} = m - 1$ and $X_{(2)} = m + 1$. The only possible value of θ now is $m - 1$. Thus, the knowledge of the value of the ancillary statistic R has increased out knowledge about θ. □

6.1.3 Complete statistics

For many important situations, unlike the previous example, our intuition that a minimal sufficient statistic is independent of any ancillary statistic is correct.

Definition 6.5. *Let $f(t|\theta)$ be a family of pdfs/pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called complete if $E_\theta[g(T)] = 0$ for all θ implies $P_\theta[g(T) = 0] = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a complete statistic.*

Completeness is property of a family of probability distributions, not of a particular distribution. For example, if X has a $\mathcal{N}(\theta, 1)$ distribution, $-\infty < \theta < \infty$, we shall see that no function of X, except one that is 0 with probability 1 for all θ, satisfies $E_\theta[g(X)] = 0$ for all θ. Thus, the family $\mathcal{N}(\theta, 1)$ distributions, $-\infty < \theta < \infty$, is complete.

Theorem 6.4. *Basu's Theorem: if $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

The sufficient statistic $T(\mathbf{X})$ of exponential family is complete as long as the parameter space $\Theta = (\theta_1, \ldots, \theta_k)$ contains an open set in $\mathbb{R}^k$. For example for the curved exponential family distribution like $\mathcal{N}(\theta, \theta^2)$ the points are bounded on a parabola and there is no open set in $\mathbf{R}^2$.

Example 6.4. Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. We want to calculate the expected value of $g(\mathbf{X}) = X_n / \sum X_i$. Being in exponential scale parameter family $g(\mathbf{X})$ is an ancillary statistic. $T(\mathbf{X}) = \sum X_i$ is also a complete statistic and hence a sufficient statistic. Hence, by Basu's theorem, $T(\mathbf{X})$ and $g(\mathbf{X})$ are independent. We also have for an exponential distribution $E_\theta[X_n] = \theta$. This gives, $\theta = E_\theta[X_n] = E_\theta[T(\mathbf{X})g(\mathbf{X})] = E_\theta[T(\mathbf{X})]E_\theta[g(\mathbf{X})] = n_\theta[g(\mathbf{X}]$. Hence, for any θ, $E_\theta[g(\mathbf{X})] = n^{-1}$. □

Theorem 6.5. *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.*

6.2 The Likelihood Principle

A specific, important, statistic called the likelihood function can also be used to summarize data. If certain other principles are accepted, then the likelihood function must be used as a data reduction device.

Definition 6.6. *Let $f(\mathbf{x}|\theta)$ denote the joint pdf/pmf of the sample $\mathbf{X} = (X_1, \dots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by*

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the likelihood function.

The likelihood function is the probability of the sample data under various possible parameter value θ. Though the likelihood look similar to pdf, they are different. When we consider pdf $f(\mathbf{x}|\theta)$, we are considering θ as fixed and $\mathbf{x}$ as the variable; when we consider the likelihood function $L(\theta|\mathbf{x})$, we are considering $\mathbf{x}$ to be observed and θ to be varying. Also, there is no guarantee that $L(\theta|\mathbf{x})$ is a probability distribution. Though it could be normalized to give 'fiducial inference', but most statisticians do not subscribe to it.

Definition 6.7. *Likelihood principle: If x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is, there exists a constant $C(x, y)$ such that $L(\theta|x) = C(x, y)L(\theta|y)$ for all θ, then the conclusions drawn from x and y should be identical.*

6.2.1 Formal Likelihood Principle

Formally we define an experiment E to be a triple $(\mathbf{X}, \theta, \{f(x|\theta)\})$, where $\mathbf{X}$ is a random vector with pmf $f(\mathbf{x}|\theta)$ for some θ in the parameter space Θ. The conclusion of the experiment is denoted by $Ev(E, \mathbf{x})$, which stands for the evidence about θ arising from E and $\mathbf{x}$.

Theorem 6.6. *Formal sufficiency principle: Consider experiment $E = (\mathbf{X}, \theta, \{f(x|\theta)\})$ and suppose $T(\mathbf{X})$ is a sufficient statistic for θ. If $\mathbf{x}$ and $\mathbf{y}$ are sample points satisfying $T(\mathbf{x}) = T(\mathbf{y})$, then $Ev(E, \mathbf{x}) = Ev(E, \mathbf{y})$.*

Theorem 6.7. *Conditionality principle: Suppose two experiments with common unknown parameter θ, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$. Consider the mixed experiment in which the random variable J is observed, where it takes value 1 or 2 with probability 1/2 each. Then the experiment E_J is performed where $E_J = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ and $\mathbf{X} = (j, \mathbf{X}_j)$ and $f(\mathbf{x}|\theta) = f((j, \mathbf{x}_j)|\theta) = \frac{1}{2}f_j(\mathbf{x}_j|\theta)$. Then, $Ev(E, (j, \mathbf{x}_j)) = Ev(E_j, \mathbf{x}_j)$.*

The conditionality principle simply says that if one of the two experiments is randomly chosen and the chosen experiment is done, the information about θ depends only on the experiment performed. The fact that this experiment was performed, rather then some other, has not changed the value of θ.

Theorem 6.8. *Formal likelihood principle: Suppose that we have two experiments, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$, where the unknown parameter θ is the same in both experiments. Suppose $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are sample points from E_1 and E_2 respectively, such that $L(\theta|\mathbf{x}_2^*) = CL(\theta|\mathbf{x}_1^*)$ for all θ and for some constant C that may depend on $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ but not θ. Then $Ev(E_1, \mathbf{x}_1^*) = Ev(E_2, \mathbf{x}_2^*)$.*

Theorem 6.9. *Likelihood principle corollary: If $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ is an experiment, then $Ev(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.*

Theorem 6.10. *Birnbaum's Theorem - The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.*

Many common statistical procedures violate the Formal Likelihood Principle, with different conclusions reached for different experiments. This is related to the weakness of sufficiency or conditionality principle. The Formal sufficiency principle is very model-dependent, and a belief in the principle necessitates belief in the model, something that may not be possible to do. Model checking based on statistics other than a sufficient statistic (e.g. checking residuals) violates the Sufficiency principle and Likelihood principle. Because of these reasons Likelihood principle is not universally accepted by all statisticians. It still remains a useful data reduction technique.

6.3 The Equivariance Principle

The Equivariance Principle describes a data reduction technique in a slightly different way. In any application of this principle, a function $T(\mathbf{X})$ is specified, but if $T(\mathbf{x}) = T(\mathbf{y})$, then the Equivariance Principle states that the inference made if $\mathbf{x}$ is observed should have a certain relationship to the inference made if $\mathbf{y}$ is observed, although the two inferences may not be the same. The Equivariance Principle involves two different equivariance considerations.

- Measurement equivariance: inference should be independent of the measurement scale.
- Formal invariance: With same formal mathematical structure, even though representing different physicality, same inference procedure should be used. The elements of the model that must be same are: Θ, the parameter space; $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$, the set of pdfs for the sample; and the set of allowable inferences and consequences of wrong inferences. The set of allowable inferences could be as simple as a choice of an element of Θ as an estimate of the true θ. This is sometimes difficult to justify.

Theorem 6.11. *Equivariance principle: if $\mathbf{Y} = g(\mathbf{X})$ is a change of measurement scale such that the model of $\mathbf{Y}$ has the same formal structure as the model for $\mathbf{X}$, then an inference procedure should be both measurement equivariant and formally invariant.*

Definition 6.8. *A set of functions $\{g(\mathbf{x} : g \in \mathcal{G})\}$ from the sample space $\mathcal{X}$ onto $\mathcal{X}$ is called a group of transformations of $\mathcal{X}$ if*

- *Inverse: For every $g \in \mathcal{G}$ there is a $g' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.*
- *Composition: For every $g \in \mathcal{G}$ and $g' \in \mathcal{G}$ there exists $g'' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = g''(\mathbf{x})$ for all $x \in \mathcal{X}$.*

Definition 6.9. *Let $\mathcal{F} = \{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ be a set of pdfs for X, and let $\mathcal{G}$ be a group of transformations of the sample space $\mathcal{X}$. Then $\mathcal{F}$ is invariant under the group $\mathcal{G}$ if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $\mathbf{Y} = g(\mathbf{X})$ has the distribution $f(\mathbf{y}|\theta)$ if $\mathbf{X}$ has the distribution $f(\mathbf{x}|\theta)$.*

Example 6.5. Let X have a binomial distribution with sample size n known and success probability p unknown. Let $T(x)$ be the estimate of p that is used when $X = x$ is observed, binomial with parameters (n, p). Similarly, we can use number of failures $Y = n - X$, where Y will also have a binomial distribution with parameters $(n, q = 1 - p)$. Let $T^*(y)$ be the estimate of q when $Y = y$ is observed. This means $1 - T^*(y)$ is the estimate of p when $Y = y$ is observed. When $X = x$ is observed, the estimate of p is $T(x)$.

If there are x successes, then there are $n - x$ failures and $1 - T * (n - x)$ is also an estimate of p. Since X to T is just a change of measurement scale, by measurement equivariance, $T(x) = 1 - T^*(n - x)$. Further, the formal structures of the inference problems are the same, both X and Y have *binomial*(n, θ) distributions, $o \leq \theta \leq 1$. So formal invariance requires $T(z) = T^*(z)$ for all $x = 0, \ldots, n$.

Hence, together, using equivariance principle we require

$$T(x) = 1 - T^*(n - x) = 1 - T(n - x).$$

Instead of specifying all $T(0), \ldots, T(n)$ estimators we can simply consider $T(0), \ldots, T(\lfloor n/2 \rfloor)$, reducing the data.

Two estimators that are equivalent for this problem are $T_1(x) = x/n$ and $T_s(x) = 0.9(x/n) + 0.1(1/2)$, as they satisfy the above condition. However, $T_3(x) = 0.8(x/n) + 0.2(1)$ is not equivalent. Hence, the choice of transformation is critical to the equivariance argument.

For the two transformations we have the set $\mathcal{G} = \{g_1, g_2\}$, with $g_1(x) = n - x$ and $g_2(x) - x$. The Inverse and Composition conditions are easily verified. Under this group $\mathcal{G}$ the set of binomial pmfs are invariant as well. For $\mathbf{X} \sim$ *binomial*(n, p), we have $g_1(X) = n - X \sim$ *binomial*$(n, 1 - p)$ and $g_2(X) = X \sim$ *binomial*(n, p), showing the forms are same. □

All three principles prescribe relationships between inferences at different sample points, restricting the set of allowable inferences and, in that way, simplify the analysis of the problem.

7 Point Estimation

When sampling is from a population described by a distribution $f(x|\theta)$, knowledge of θ yields knowledge of the entire population. Hence, point estimating θ is a natural endeavour. It may also be the case that some function of θ, say $\tau(\theta)$, is of interest.

Definition 7.1. *A point estimator is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator. The realized value of an estimator (a number) is called an estimate, when a sample is actually taken.*

7.1 Methods of Finding Estimators

In some cases it is easy to decide how to estimate a parameter, e.g. using sample analogue like mean. I more complicated models, we need a more methodical way of estimating parameters.

7.1.1 Method of Moments

This is the most simplest of the methods and often sub optimal. Let $X_1, \ldots, X_n$ be a sample from a population with distribution $f(x|\theta_1, \ldots, \theta_k)$. We equate the first k moments to the corresponding k population moments, and solve the resulting system of simultaneous equations. More precisely, $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$, $\mu'_j = E[X^j]$ for $j = 1, \ldots, k$. The population moments μ'_j are typically a function of $\theta_1, \ldots, \theta_k$, say $\mu'_j(\theta_1, \ldots, \theta_k)$. Hence, we get $m_j = \mu'_j(\theta_i, \ldots, \theta_k)$, for $j = 1, \ldots, k$. We solve these k equations to get the parameters.

Example 7.1. To estimate crime rate where total number of occurrences k and true reporting p are unknown, we use $Binomial(k, p)$ distribution. Let $X_1, \ldots, X_n$ be iid samples, i.e., $P[X_i = x|k, p] = \binom{k}{x} p^x (1-p)^{k-x}$, $x = 0, 1, \ldots, k$. Equating the first two sample moments to those of population gives $\frac{1}{n}\sum X_i = kp$ and $\frac{1}{n}\sum X_i^2 = kp(1-p) + k^2p^2$. This can be solved to $\hat{k} = \bar{X}^2/(\bar{X} - \frac{1}{n}\sum(X_i - \bar{X})^2)$, and $\hat{p} = \frac{\bar{X}}{\hat{k}}$. Since the range of estimator does not coincide with the range of the parameter, it is possible to get negative estimates for k and p, which of course should be positive numbers.This method of moment matching, however, serve as the approximations to the distributions of statistics. □

Example 7.2. *Satterthwaite approximation:* If Y_i, $i = 1, \ldots, k$, are independent $\mathcal{X}^2_{r_i}$ random variables. The distribution of $\sum Y_i$ is also chi squared with degrees of freedom equal to $\sum r_i$, where $E[Y_i] = r_i$. The distribution of $Z = \sum a_i Y_i$, where a_is are known constants is difficult to obtain, but is reasonable to assume $Z \sim \mathcal{X}^2_\nu/\nu$, for some value of ν might provide a good approximation. Matching the first moment gives $E[\sum a_i Y_i] = \sum a_i E[Y_i] = \sum a_i r_i = 1$, giving a constraint on the a_is. We match the second moment to get ν.

$$\left(\sum_{i=1}^k a_i Y_i\right)^2 = E\left(\frac{\mathcal{X}^2_\nu}{\nu}\right)^2 = \frac{2}{\nu} + 1 \implies \hat{\nu} = \frac{2}{(\sum a_i Y_i)^2 - 1}.$$

Here again ν can be negative. Satterthwaite customized the method of moments to get positive estimates as follows:

$$\begin{aligned}E[(\sum a_iY_i)^2] &= Var[\sum a_iY_i] + \Big(E[\sum a_iY_i]\Big)^2\\ &= \Big(E[\sum a_iY_i]\Big)^2\left[\frac{Var[\sum a_iY_i]}{(E[\sum a_iY_i])^2}+1\right]\\ &= \left[\frac{Var[\sum a_iY_i]}{(E[\sum a_iY_i])^2}+1\right]\end{aligned}$$

We equate to the second moment to obtain, after dropping the left expectation

$$\nu = \frac{2\,(E[\sum a_iY_i])^2}{Var[\sum a_iY_i]}.$$

Finally we use the iid nature or Y_is to get $Var[\sum a_iY_i] = \sum a_i^2 Var[Y_i] = 2\sum \frac{1}{r_i}a_i^2(E[Y_i])^2$. This, give the expression of ν making it necessarily positive

$$\hat{\nu} = \frac{(\sum a_iY_i)^2}{\sum \frac{a_i^2}{r_i}Y_i^2}.$$

□

7.1.2 Maximum Likelihood Estimators

This is the most popular method. If $X_1,\dots,X_n$ are an iid sample from a population with distribution $f(x|\theta_1,\dots,\theta_k)$, the likelihood function is defined by

$$L(\theta|x) = L(\theta_1,\dots,\theta_k|x_1,\dots,x_n) = \prod_{i=1}^{n} f(x_i|\theta_1,\dots,\theta_k).$$

Definition 7.2. *For each sample point x, let $\hat{\theta}(x)$ be a parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample X is $\hat{\theta}(X)$.*

By construction the range of MLE coincides with the range of the parameter. MLE is the parameter point for which the observed sample is most likely. There are two inherent drawbacks with finding maximums - finding and verifying it is a global maximum, and numerical sensitivity to change in data. If the function is differentiable (in θ_i), the possible candidates for the MLE are the values of $(\theta_1,\dots,\theta_k)$ that satisfies

$$\frac{\partial}{\partial\theta_i}L(\theta|x) = 0,\ \text{for } i = 1,\dots,k.$$

The condition for maximum, e.g. with second derivative should be checked. If the extreme occurs at the boundary they need to be checked separately. Further, we need to find global maximum and not the local one or inflection points. In most cases, especially when differentiation is to be used, it is easier to work with the natural logarithm of $L(\theta|x)$, $\log L(\theta|x)$, the log likelihood. This is reasonable because *log* is a monotonic function on $(0,\infty)$ and the extrema of $L(\theta|x)$ and $\log L(\theta|x)$ coincide.

Example 7.3. Let $X_1, \ldots, X_n$ be iid $(\theta, 1)$, where $\theta \geq 0$. With no restrictions on θ we can find the log likelihood function $L(\theta|x)$

$$\log L(\theta|x) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2.$$

The equation $(d/d\theta)\log L(\theta|x) = 0$ reduces to

$$\sum_{i=1}^{n}(x_i - \theta) = 0,$$

which has the solution $\hat{\theta} = \bar{x}$. Hence, $\bar{x}$ is a candidate for MLE. Since, $\frac{d^2}{d\theta^2}L(\theta|x)\|_{\theta=\bar{x}} < 0$ and boundaries are sub optimal, we conclude that $\bar{x}$ is the MLE.

However if $\bar{X}$ is negative, it will be outside the range of the parameter. If $\bar{X}$ is negative, it is easy to check that the likelihood function $L(\theta|x)$ is decreasing in θ for $\theta \geq 0$ and is maximized at $\hat{\theta} = 0$. Hence, the MLE for θ is $\hat{\theta} = \bar{X}\mathbb{I}_{\bar{X}\geq 0}$. □

Theorem 7.1. *Invariance property of MLEs: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

If the function τ is one-to-one, i.e. invertible there is no problem. If we let $\eta = \tau(\theta)$, then the inverse function $\tau^{-1}(\eta) = \theta$ is well defined and the likelihood function of $\tau(\theta)$, written as a function of η, is given by

$$L^*(\eta|x) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|x)$$

and

$$\underset{\eta}{sup}L^*(\eta|x) = \underset{\eta}{sup}L(\tau^{-1}(\eta)|x) = \underset{\theta}{sup}L(\theta|x).$$

Thus, the maximum of $L^*(\eta|x)$ is attained at $\eta = \tau(\theta) = \tau(\hat{\theta})$, showing that the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$. If the function is not one-to-one we proceed by defining for $\tau(\theta)$ the induced likelihood function L^*, given by

$$L^*(\eta|x) = \underset{\{\theta:\tau(\theta)=\eta\}}{sup} L(\theta|x).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|x)$ will be called the MLE of $\eta = \tau(\theta)$, and the maxima of L^* and L coincide. It can then be shown that $L^*(\hat{\eta}|x) = L(\hat{\theta}|x)$.

The invariance property also hold in multivariate case. However, in this case, check for maximum using second derivative can be tedious. Successive maximization techniques are used in that case, sometimes.

The maximization problem is sensitive to the data. If we calculate $\hat{\theta}$ based on $L(\theta|x)$, we might be interested in $L(\theta|x + \epsilon)$, for small ϵ. We expect these to be close to each other, but that may not be the case. This can happen when the likelihood function is vary flat in the neighbourhood of its maximum or when there is no finite maximum. It is often useful to spend some time investigating the stability of the solution.

7.1.3 Bayes Estimators

In classical statistics θ is an unknown, but fixed quantity. Observations are done to get inference on this parameter. In Bayesian statistic θ is a random variable. Prior distribution is transformed to posterior conditional distribution given the observations. The mean of the posterior could be seen as the point estimate. For any sampling distribution, there is a natural family of prior distributions, called the conjugate family.

Definition 7.3. *Let $\mathcal{F}$ denote the class of distributions $f(x|\theta)$. A class $\prod$ of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class $\prod$ for all $f \in \mathcal{F}$, all priors in $\prod$, and all $x \in \mathcal{X}$.*

The updating of priors takes the form of updating its parameters, for a conjugate family, making it mathematically very convenient. Whether or not a conjugate family is a reasonable choice for a particular problem is a question left to the experimenter.

7.1.4 The EM Algorithm

Expectation-Maximization Algorithm is a way to find MLEs. This algorithm is guaranteed to converge to the MLE. It is based on the idea that replacing one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem. It is particularly suited to 'missing data' problems. In using EM algorithm we consider two different likelihood problems - incomplete data problem (problem of interest) and complete data problem (easier problem we solve).

If $Y = (Y_1, \ldots, Y_n)$ are the incomplete data and $X = (X_1, \ldots, X_m)$ the augmented data, making (Y, X) the complete data, the densities $g(.|)$ for Y and $f(.|)$ for (Y,X) have the relationship $g(y|\theta) = \int f(y, x|\theta) d\theta$. If we turn these into likelihoods, $L(\theta|y) = g(y|\theta)$ is the incomplete-data likelihood and $L(\theta|y, x) = f(y, x|\theta)$ is the complete-likelihood.

This implies the the distribution of augmented data is $k(x|\theta, y) = \frac{f(y,x|\theta)}{g(y|\theta)}$. This gives $logL(\theta|y) = logL(\theta|y, x) - logk(x|\theta, y)$. As x is missing data and hence not observed, we take expectation under $k(x|\theta, y)$ to get $logL(\theta|y) = E\left[logL(\theta|y, X)|\theta', y\right] - E\left[logk(X|\theta, y)|\theta', y\right]$. From an initial point $\theta^{(0)}$ we create a sequence $\theta^{(r)}$

$$\theta^{(r+1)} = \underset{\theta}{argmax} E\left[logL(\theta|y, X)|\theta^{(r)}, y\right],$$

The 'E-step' calculates the expected log likelihood and the 'M-step' finds its maximum. An example will clarify the steps.

Example 7.4. We observe $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, all mutually independent, where $Y_i \sim Poisson(\beta\tau_i)$ and $X_i \sim Poisson(\tau_i)$. This could model, for instance, the incidence of a disease, Y_i, where the underlying rate is a function of an overall effect β (base contagion rate) and and additional factor τ_i (e.g. population density). We get information about τ_i through X_i. The join distribution is

$$f((x_1, y_1), \ldots, (x_n, y_n)|\beta, \tau_1, \ldots, \tau_n) = \prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}.$$

The likelihood estimators, using differentiation are $\hat{\beta} = \frac{\sum y_i}{\sum x_i}$ and $\hat{\tau}_i = \frac{x_i + y_i}{\hat{\beta}+1}$. This is a complete-data likelihood and the data $((x_1, y_1), \ldots, (x_n, y_n))$ is called the complete data. Missing data makes this estimation more difficult.

Suppose that the value of x_1 was missing. We could discard y_1 and proceed with a sample of size $n-1$, but this is ignoring the information in y_1. Using this information would improve our estimation. The distribution now with x_1 missing is simply marginalizing the variable x_1 by summation, giving

$$\sum_{x_1=0}^{\infty} f((x_1, y_1), \ldots, (x_n, y_n)|\beta, \tau_1, \ldots, \tau_n).$$

This likelihood is the incomplete-data likelihood. This is what we want to maximize. For our example summing over x_1 we get

$$L(\beta, \tau_1, \ldots, \tau_n | y_1, (x_2, y_2), \ldots, (x_n, y_n)) = \left[\prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\right] \left[\prod_{i=2}^{n} \frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}\right],$$

ans $(y_1, (x_2, y_2), \ldots, (x_n, y_n))$ is the incomplete data. Differentiating leads to the MLE equations

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\tau}_i}, \ y_1 = \hat{\tau}_1 \hat{\beta}, \text{ and } x_j + y_j = \hat{\tau}_j(\hat{\beta}+1), \text{ for } j = 2, \ldots, n,$$

which we now solve with EM algorithm.

For our problem let (x, y) denote the full data and $(x_{(-1)}, y)$ denote the incomplete data. The expected complete data log likelihood is

$$\begin{aligned}
&E\left[logL(\beta, \tau_1, \ldots, \tau_n|(x,y)|\tau^{(r)}, (x_{(-1)}, y))\right] \\
&= \sum_{x_1=0}^{\infty} log\left(\prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\right) \frac{e^{-\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!} \\
&= \sum_{i=1}^{n} [-\beta\tau_i + y_i(log\beta + log\tau_i) - logy_i!] + \sum_{i=2}^{n} [-\tau_i + x_i \log\tau_i - \log x_i!] \\
&+ \sum_{x_1=0}^{\infty} [-\tau_1 + x_1 log\tau_1 - \log x_1!] \frac{e^{-\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!} \\
&= \left(\sum_{i=1}^{n} [-\beta\tau_i + y_i(\log\beta + \log\tau_i)] + \sum_{i=2}^{n} [-\tau_i + x_i \log\tau_i] + \sum_{x_1=0}^{\infty} [-\tau_1 + x_1 \log\tau_1] \frac{e^{-\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!}\right) + \ldots \\
&= \left(\sum_{i=1}^{n} [-\beta\tau_i + y_i(\log\beta + \log\tau_i)] + \sum_{i=2}^{n} [-\tau_i + x_i \log\tau_i] - \tau_1 + \tau_1^{(r)} \log\tau_1\right) + \ldots
\end{aligned}$$

where the dots represent term that do not depend on β and τ_i. In the last equation we also use the fact that the probability distribution sums to 1 and the expected value of x_1 is $\tau_1^{(r)}$.

This can now be solved to

$$\hat{\beta}^{(r+1)} = \frac{\sum_{i=1}^{n} y_i}{\tau_1^{(r)} + \sum_{i=2}^{n} x_i}; \ \hat{\tau}_1^{(r+1)} = \frac{\hat{\tau}_1^{(r)} + y_1}{\hat{\beta}^{(r+1)} + 1}; \ \hat{\tau}_j^{(r+1)} = \frac{x_j + y_j}{\hat{\beta}^{(r+1)} + 1}, j = 2, \ldots, n.$$

This defines both the E-step and M-step. The properties of EM algorithm give us assurance that the sequence $(\hat{\beta}^{(r)}, \hat{\tau}_1^{(r)}, \ldots, \hat{\tau}_n^{(r)})$ converges to the incomplete-data MLE as $r \to \infty$. □

Theorem 7.2. *Monotonic EM sequence: The sequence $\{\hat{\theta}^{(r)}\}$ defined by*

$$\theta^{(r+1)} = \underset{\theta}{argmax} E\left[logL(\theta|y, X)|\theta^{(r),y}\right]$$

satisfies

$$L(\hat{\theta}^{(r+1)}|y) \geq L(\hat{\theta}^{(r)}|y)$$

with equality holding iff successive iterations yield the same value of the maximized expected complete-data log likelihood.

7.2 Methods of Evaluating Estimators

Different methods of estimations can yield different results. We, hence, need to evaluate the statistical performance of the estimators. This is formalized by decision theory.

7.2.1 Mean Squared Error

Definition 7.4. *The mean squared error, MSE, of an estimator W of a parameter θ is the function of θ defined by $E_\theta[(W - \theta)^2]$.*

$E_\theta[|W - \theta|]$, and any increasing function of $|W - \theta|$, is a reasonable alternative, but MSE has some advantages. First, it is quite tractable analytically and, second it has the interpretation

$$E_\theta[(W - \theta)^2] = Var_\theta[W] + (E_\theta[W] - \theta)^2 = Var_\theta[W] + (Bias_\theta[W])^2.$$

Definition 7.5. *The bias of a point estimator W of a parameter θ is $E_\theta[W - \theta]$. An estimate with 0 bias is called unbiased and satisfies $E_\theta[W] = \theta$ for all θ.*

The MSE incorporates both variability (precision) and bias (accuracy). A good MSE has small combined variance and bias. For an unbiased estimator the MSE is equal to its variance.

Example 7.5. Normal MSE. Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\mu,^2)$. The statistic $\bar{X}$ and S^2 are both unbiased estimators since $E[\bar{X}] = \mu$ and $E[S^2] = \sigma^2$; this is true without the normality assumption. The MSEs of these estimates are $E(\bar{X} - \mu)^2 = Var[\bar{x}] = \frac{\sigma^2}{n}$, and $E(S^2 - \sigma^2)^2 = Var[S^2] = \frac{2\sigma^4}{n-1}$. The second expression need normality assumption. □

It is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a large decrease in variance, resulting in improved MSE.

Example 7.6. An alternative estimator for σ^2 is the MLE $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. We see that $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$, so $\hat{\sigma}^2$ is a biased estimator of σ^2. The variance of $\hat{\sigma}^2$ can be calculated as $Var[\hat{\sigma}^2] = Var[\frac{n-1}{n}S^2] = \frac{2(n-1)\sigma^4}{n^2}$. We, hence, see the MSE is $E(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4$. Also $E(S^2 - \sigma^2)^2 = \left(\frac{2}{n-1}\right)\sigma^4$. Hence, clearly $E(\hat{\sigma}^2 - \sigma^2)^2 < E(S^2 - \sigma^2)^2$. Thus, by trading off variance for bias, the MSE is improved. □

MSE, while a reasonable criterion for location parameters, is not reasonable for scale parameters. MSE penalizes equally for overestimation and underestimation, which is fine in the location case. In the scale case, however, 0 is natural lower bound, so the estimation problem is not symmetric. Use of MSE in this case tends to be forgiving the underestimation. In general, there will not be one 'best' estimator, for full parameter space.

7.2.2 Best Unbiased Estimators

One way to make the problem of finding a 'best' estimator tractable is to limit the class of estimators, e.g. considering only unbiased estimators. Since for unbiased estimators MSE are equal to their variance, the 'best' estimator here would be the one with smallest variance.

Definition 7.6. *An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta[W^*] = \tau(\theta)$ for all θ and, for any other estimator W with $E_\theta[W] = \tau(\theta)$, we have $Var_\theta[W^*] \leq Var_\theta[W]$ for all θ. $W*$ is also called the uniform minimum variance unbiased estimator, UMVUE for $\tau(\theta)$.*

Best unbiased estimators may not be unique. A more formal approach is to find the lower bound on the variance of any unbiased estimator.

Theorem 7.3. *Cramer-Rao Inequality: Let $X_1, \ldots, X_n$ be sample with joint distribution $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be an estimator satisfying*

$$\frac{d}{d\theta}E_\theta[W(\mathbf{X})] = \int_{\mathbf{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x}|\theta)]dx, \text{ and } Var_\theta[W(\mathbf{X})] < \infty.$$

Then

$$Var_\theta[W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right)^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]}.$$

The above theorem can be proved elegantly using Cauchy-Schwartz inequality. For the additional assumption of iid, the above bound simplifies.

Theorem 7.4. *Cramer-Rao Inequality, iid case: If the assumptions of Cramer-Rao Inequality are satisfied and, additionally, if the samples are iid then*

$$Var_\theta[W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right)^2}{nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

The quantity $E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]$ is called the Fisher information of the sample. The information number gives a bound on the variance of the best unbiased estimator of θ. As the information number gets bigger we have smaller bounds on the variance.

Theorem 7.5. *If $f(x|\theta)$ satisfies (true for an exponential family)*

$$\frac{d}{d\theta}E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)f(x|\theta)\right]dx$$

then,

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

It is important to remember that a key assumption in the Cramer-Rao Theorem is the ability to differentiate under the integral sign.

Example 7.7. Let $X_1,\ldots,X_n$ be iid with distribution $f(x|\theta) = 1/\theta$, $0 < x <$. Since $\frac{\partial}{\partial\theta}\log f(x|\theta) = -1/\theta$, we have $E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] = 1/\theta^2$. The Cramer-Rao theorem would then imply that if W is any unbiased estimator of θ, then $Var_\theta[W] \geq \frac{\theta^2}{n}$. For the sufficient statistic $Y = max(X_1,\ldots,X_n)$ the pdf of Y is $f_Y(y|\theta) = ny^{n-1}/\theta^n$, $0 < y < \theta$, so $E_\theta[Y] = \int_0^\theta \frac{ny^n}{\theta^n}dy = \frac{n}{n+1}\theta$, showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ. Further, $Var_\theta\left[\frac{n+1}{n}Y\right] = \frac{1}{n(n+2)}\theta^2$, which is uniformly smaller than θ^2/n. This indicates that the Cramer-Rao theorem is not applicable to this pdf. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable. We can verify the failure of required condition via Leibnitz s Rule

$$\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)dx = \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}dx = \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx \neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}f(x|\theta)dx,$$

unless $h(\theta)/\theta = 0$ for all θ. □

A further shortcoming is how to find if the bound is sharpe, i.e. is the value of Cramer-Rao lower bound strictly smaller than the variance of any unbiased estimator. The conditions for attainment of the Cramer-Rao lower bound are quite simple.

Theorem 7.6. *Attainment of Cramer-Rao bound: Let $X_1,\ldots,X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramer-Rao theorem. Let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1,\ldots,X_n)$ is an unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao lower bound iff*

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial}\log L(\theta|\mathbf{x})$$

for some function $a(\theta)$.

Example 7.8. Let $X_1, \dots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$, and consider estimation of σ^2, where μ is unknown. The normal pdf satisfies the assumption of the Cramer-Rao theorem so we have

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\right) = \frac{1}{2\sigma^2} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$-E\left[\frac{\partial^2}{\partial(\sigma^2)^2} \log f(\mathbf{X}|\mu, \sigma^2)|\mu, \sigma^2\right] = -E\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}|\mu, \sigma^2\right] = \frac{1}{2\sigma^4}$$

Thus, any unbiased estimator, W, of σ^2 must satisfy

$$Var[W|\mu, \sigma^2] \geq \frac{2\sigma^4}{n}.$$

Further we know, $Var[S^2|\mu, \sigma^2] = \frac{2\sigma^4}{n-1}$, which does not attain the Cramer-Rao lower bound. Here we have

$$L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2}$$

, and hence we have

$$\frac{\partial}{\partial\sigma^2} \log L(\mu, \sigma^2|\mathbf{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} - \sigma^2\right).$$

Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of σ^2 is $\sum_{i=1}^{n}(x_i - \mu)^2/n$, when mu is known. If μ is unknown, the bound cannot be attained. □

What can we do if $f(x|\theta)$ does not satisfy the assumptions of the Cramer-Rao theorem? What if the bound is unattainable by allowable estimators? There are more sharper bounds by Chapman and Robbins. We will, however, continue this study using the concept of sufficiency.

7.2.3 Sufficiency and Unbiasedness

We now make use of sufficiency. We will make use of $E[X] = E[E[X|Y]]$ and $Var[X] = Var[E[X|Y]] + E[Var[X|Y]]$.

Theorem 7.7. *Rao-Blackwell theorem: Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = E[W|T]$. Then $E_\theta[\phi(T)] = \tau(\theta)$ and $Var_\theta[\phi(T)] \leq Var_\theta[W]$ for all θ; i.e., $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.*

Therefore, conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only statistics that are function of a sufficient statistic in our search for best unbiased estimators. One must make sure the resulting quantity does not depend on θ for it be a valid estimator.

Example 7.9. Let X_1, X_2 be iid on $\mathcal{N}(0, 1)$. The statistic $\bar{X} = \frac{1}{2}(X_1 + X_2)$ has $E_\theta[\bar{X}] = \theta$ and $Var_\theta[\bar{X}] = \frac{1}{2}$. If we condition on X_1, which is not sufficient we let $\phi(X_1) = E_\theta[\bar{X}|X_1]$. Hence, $E_\theta[\phi(X_1)] = E_\theta[E_\theta[\bar{X}|X_1]] = E_\theta[\bar{X}] = \theta$ and $Var_\theta[\phi(X_1)] = Var_\theta[E_\theta[\bar{X}|X_1]] = Var_\theta[\bar{X}] - E_\theta[Var_\theta[\bar{X}|X_1]] \leq Var_\theta[\bar{X}]$. So $\phi(X_1)$ is better than $\bar{X}$. However, $\phi(X_1) = E_\theta[\bar{X}|X_1] = \frac{1}{2}E_\theta[X_1|X_1] + \frac{1}{2}E_\theta[X_2|X_1] = \frac{1}{2}(X_1 + \theta)$, thus making it a function of θ and, hence, not a valid estimator. □

Theorem 7.8. *If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.*

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? Suppose that W satisfies $E_\theta[W] = \tau(\theta)$, and we have another estimator, U, that satisfies $E_\theta[U] = 0$ for all θ, that is, U is an unbiased estimator of 0. The estimator $\phi_a = W + aU$ has $E_\theta[\phi_a] = \tau(\theta)$ and is unbiased. Further, $Var_\theta[\phi_a] = Var_\theta[W] + a^2 Var_\theta[U] + 2aCov_\theta[W, U]$. It's relation to $Var_\theta[W]$ depending on the covariance term. Hence, the relationship of W with unbiased estimators of 0 is crucial in evaluating whether W is best unbiased.

Theorem 7.9. *If $E_\theta[W] = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ iff W is uncorrelated with all unbiased estimators of 0.*

An unbiased estimator of 0 is nothing more than random noise. If an estimator could be improved by adding noise to it, the estimator probably is defective. This theorem is practically unusable, because we can't verify against all unbiased estimators of 0. But, it is useful in determining that an estimator is not best unbiased.

Characterization of the unbiased estimators of zeros requires conditions on the distribution which we are working with. If a family of distributions $f(x|\theta)$ has the property that there are no unbiased estimators of zero (other than zero itself), then our search would be ended, since any statistic W satisfies $Cov_\theta[W, 0] = 0$. The property of completeness of the family of distributions of the sufficient statistic, guarantees such a situation.

Theorem 7.10. *Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be an estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.*

When there are no obvious candidates for an unbiased estimator of a function $\tau(\theta)$, in presence of completeness, we can find any unbiased estimator and convert it to best unbiased estimator. If T is a complete sufficient statistic for a parameter θ and $h(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $\phi(T) = E[h(X_1, \ldots, X_n)|T]$ is the best unbiased estimator of $\tau(\theta)$.

Example 7.10. Let $X_1, \ldots, X_n$ be iid $Binomial(k, \theta)$. We want to estimate the probability of exactly one success, i.e. $\tau(\theta) = P_\theta[X = 1] = k\theta(1-\theta)^{k-1}$. Now $\sum X_i \sim Binomial(kn, \theta)$ is a complete statistic, but not an unbiased estimator. The simplest estimator we can come up with is $h(X_1) = \mathbb{I}_{X_1=1}$. Here,

$$E_\theta[h(X_1)] = \sum_{x_1=0}^{k} h(x_1)\binom{k}{x_1}\theta^{x_1}(1-\theta)^{k-x_1} = k\theta(1-\theta)^{k-1}$$

and hence is an unbiased estimator of $\tau(\theta)$. The estimator $\phi(\sum X_i) = E\left[h(X_1)|\sum X_i\right]$ is the best unbiased estimator of $\tau(\theta)$. Suppose $\sum X_i = t$. Then

$$\begin{aligned}
\phi(t) &= E\left[h(X_1)|\sum_{i=1} *nX_i = t\right] \\
&= P\left[X_1 = 1|\sum_{i=1}^{n} X_i = t\right] \\
&= \frac{P_\theta\left[X_1 = 1, \sum_{i=1}^{n} X_i = t\right]}{P_\theta\left[\sum_{i=1}^{n} X_i = t\right]} \\
&= \frac{P_\theta\left[X_1 = 1, \sum_{i=2}^{n} X_i = t-1\right]}{P_\theta\left[\sum_{i=1}^{n} X_i = t\right]} \\
&= \frac{P_\theta[X_1 = 1]P[X_1 = 1, \sum_{i=1}^{n} X_i = t]}{P_\theta\left[\sum_{i=1}^{n} X_i = t\right]}
\end{aligned}$$

Now $X_1 \sim Binomial(k, \theta)$, $\sum_{n=2}^{n} \sim Binomial(k(n-1),)$, and $\sum_{n=1}^{n} X_i \sim Binomial(kn, \theta)$, giving us

$$\phi(t) = \frac{k\theta(1-\theta)^{k-1}\binom{k(n-1)}{t-1}^{t-1}(1-\theta)^{k(n-1)-(t-1)}}{\binom{kn}{t}\theta^t(1-\theta)^{kn-t}} = k\frac{\binom{k(n-1)}{t-1}}{\binom{kn}{t}}.$$

We note that the above expression is independent of θ as it must be since $\sum X_i$ is sufficient. This gives us our best unbiased estimator or $\tau(\theta)$. □

7.2.4 Loss Function Optimality

Mean squared error is a special case of a function called a loss function, which we study under decision theory. After the data $\mathbf{X} = \mathbf{x}$ has been observed, where $X \sim f(\mathbf{x}|\theta)$, $\theta \in \Theta$, a decision from allowable action space $\mathcal{A}$ has to be made. Often in point estimation problems $\mathcal{A}$ is equal to Θ, the parameter space, but this will change in other problems, e.g. hypothesis testing. The loss function is a non-negative function that generally increases as the distance between the action a and the true parameter θ increases. Two commonly used loss functions are absolute error loss ($|a-\theta|$) and squared error loss ($(a-\theta)^2$). Squared error loss gives relatively more penalty for large discrepancies, and absolute error loss gives relatively more penalty for small discrepancies. In general, the experimenter must consider the consequences of various errors in estimation for different values of θ and specify a loss function that reflects these consequences.

The quality of an estimator is quantified in its risk function. For an estimator $\delta(\mathbf{x})$ of θ, the risk function, is $R(\theta, \delta) = E_\theta[L(\theta, \delta(\mathbf{X}))]$. At a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(\mathbf{x})$ if used. Usually, one can't find estimators which are uniformly better than others for all θs, it becomes a case of judgement then to simultaneously minimize bias and variance.

We can also use a Bayesian approach to the problem of loss function optimality. For a given prior $\pi(\theta)$ we compute an average risk (Bayes risk) as

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta.$$

The estimator that minimizes Bayes risk is called the Bayes rule, denoted by δ^π. For $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi$ the Bayes risk of a decision rule δ can be written as

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \left(\int_\mathcal{X} L(\theta,\delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x}\right)\pi(\theta)d\theta = \int_\mathcal{X} \left(\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta\right)m(\mathbf{x})d\mathbf{x}.$$

We use $f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$, where $\pi(\theta|\mathbf{x})$ is the posterior distribution and $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$. The quantity in the brackets in the last term is the expected value of the loss function with respect to the posterior distribution, called the posterior expected loss. It is a function independent of θ and depends only on $\mathbf{x}$. Thus, for each $\mathbf{x}$, if we choose the action $\delta(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk. These integrals can be calculated analytically or numerically. For squared error loss $\delta^\pi(\mathbf{x}) = E[\theta|x]$ and for absolute error loss $\delta^\pi = median(\pi(\theta|\mathbf{x}))$.

8 Hypothesis Testing

Definition 8.1. *A hypothesis is a statement about a population parameter. The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.*

If θ denotes a population parameter, the general format of the null and alternative hypothesis is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. After the observation one has to either accept H_0 as true or reject H_0.

Definition 8.2. *A hypothesis test is a rule that specifies:*

- *Acceptance region (R^c): For which sample values the decision is made to accept H_0 as true.*
- *Rejection region (R): For which sample values H_0 is rejected and H_1 is accepted as true.*

Typically, a hypothesis test is specified in terms of a test statistic $W(X_1, \ldots, X_n) = W(\mathbf{X})$, a function of the sample.

8.1 Methods of Finding Tests

The Likelihood ratio test is almost always applicable and also optimal in some cases. The likelihood function is $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$, where $f(x|\theta)$ is the distribution of $X_1, \ldots, X_n$ iid samples, with Θ denoting the entire parameter space.

Definition 8.3. *The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is*

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\Theta_0} L(\theta|\mathbf{x})}{\sup\limits_{\Theta} L(\theta|\mathbf{x})}.$$

A likelihood ratio test, LRT, is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \le c\}$, where c is any number satisfying $0 \le c \le 1$.

Suppose $\hat{\theta}$, an MLE of θ for unrestricted maximization of $L(\theta|\mathbf{x})$. Similarly, we do restricted maximization over Θ_0 of θ, calling it $\hat{\theta}_0$. That is $\hat{\theta}_0 = \hat{\theta}_0(\mathbf{x})$ if the value of $\theta \in \Theta_0$ that maximizes $L(\theta|\mathbf{x})$. Then the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

Theorem 8.1. *If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.*

Example 8.1. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\theta, 1)$ population. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Here θ_0 is a number fixed by the experimenter prior to the experiment. The numerator of $\lambda(\theta)$ is $L(\theta_0|\mathbf{x})$. The unrestricted MLE of θ is $\bar{X}$, the sample mean. Thus the denominator of $\lambda(\mathbf{x})$ is $L(\bar{x}|\mathbf{x})$. So the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2}\exp\left[-\sum(x_i-\theta_0)^2/2\right]}{(2\pi)^{-n/2}\exp\left[-\sum(x_i-\bar{x})^2/2\right]} = \exp\left[\frac{1}{2}\left(-\sum_{i=1}^{n}(x_i-\theta_0)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\right] = e^{-\frac{n}{2}(\bar{x}-\theta_0)^2}.$$

An LRT is a test that rejects H_0 for small values of $\lambda(\mathbf{x})$. The rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) \le c\}$, can be written as $\{\mathbf{x} : |\bar{x}-\theta_0| \ge \sqrt{-2(\log c)/n}\}$. As c is between 0 and 1, $\sqrt{-2(\log c)/n}$ ranges between 0 and ∞. Thus the LRTs are just those tests that reject $H_0 : \theta = \theta_0$ if the sample mean differs from the hypothesized value θ_0 by more than a specified amount.

Without going into the calculations we could have recognized that $\bar{X}$ is a sufficient statistic for θ. We can then simply use the likelihood function associated with $\bar{X} \sim \mathcal{N}(\theta, \frac{1}{n})$ to more easily reach the above conclusion that a likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ rejects H_0 for large values of $|\bar{X}-\theta_0|$. □

Likelihood ratio tests are also useful in situations were there are nuisance parameters (parameters not of inferential interest). They do not effect the LRT construction method but can lead to different test.

Example 8.2. Suppose $X_1, \ldots, X_n$ are a random sample from $\mathcal{N}(\mu, \sigma^2)$, and we are interested only in inferences about μ, specifically the hypothesis test $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$. Then the parameter σ^2 is a nuisance parameter. The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{\underset{\{\mu,\sigma^2:\mu\le\mu_0,\sigma^2\ge0\}}{max} L(\mu,\sigma^2|\mathbf{x})}{\underset{\{\mu,\sigma^2:-\infty<\mu<\infty,\sigma^2\ge0\}}{max} L(\mu,\sigma^2|\mathbf{x})} = \frac{\underset{\{\mu,\sigma^2:\mu\le\mu_0,\sigma^2\ge0\}}{max} L(\mu,\sigma^2|\mathbf{x})}{L(\hat{\mu},\hat{\sigma}^2|\mathbf{x})},$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the MLEs of μ and σ^2. Furthermore, if $\hat{\mu} \le \mu_0$, then the restricted maximum is the same as the unrestricted maximum, while if $\hat{\mu} > \mu_0$, the restricted maximum is $L(\mu_0, \hat{\sigma}_0^2|\mathbf{x})$, where $\hat{\sigma}_0^2 = \sum(x_i-\mu_0)^2/n$. Thus,

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\mu} \le \mu_0 \\ \frac{L(\mu_0,\hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu},\hat{\sigma}^2|\mathbf{x})} & \text{if } \hat{\mu} > \mu_0. \end{cases}$$

It can be shown that the test based on $\lambda(\mathbf{x})$ is equivalent to a test based on Student's t statistic. □

Hypothesis testing problems may also be formulated in a Bayesian model. All inferences about the parameter θ are based on the posterior distribution $\pi(\theta|\mathbf{x})$, which are constructed from the prior $\pi(\theta)$ and the sampling distribution $f(\mathbf{x}|\theta)$. The posterior probabilities $P[\theta \in \Theta_0|\mathbf{x}] = P[H_0 \text{ is true}|\mathbf{x}]$ and $P[\theta \in \Theta_0^c|\mathbf{x}] = P[H_1 \text{ is true}|\mathbf{x}]$ may be computed. For classical statistic these probabilities don't make sense, because they are, classically, either true or false. One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if $P[\theta \in \Theta_0|\mathbf{X}] \ge P[\theta \in \Theta_0^c|\mathbf{X}]$ and to reject H_0 otherwise. Alternatively, if we want to guard against falsely rejecting H_0, we may decide to reject H_0 only if $P[\theta \in \Theta_0^c|\mathbf{X}]$ is greater than some large number, 0.99 for example.

Example 8.3. Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\theta, \sigma^2)$ and let the prior distribution on θ be $\mathcal{N}(\mu, \tau^2)$, where σ^2, μ, and τ^2 are known. Consider testing $H_0 : \theta \leq \theta_0$, versus $H_1 : \theta > \theta_0$. The posterior $\pi(\theta|\bar{x})$ is normal with mean $(n\tau^2\bar{x} + \sigma^2\mu)/(n\tau^2 + \sigma^2)$ and variance $\sigma^2\tau^2/(n\tau^2 + \sigma^2)$.

If we decide to accept H_0 if $P[\theta \in \Theta_0|\mathbf{X}] = P[\theta \leq \theta_0|\mathbf{X}]$, then we will accept H_0 if

$$\frac{1}{2} \leq P[\theta \in \Theta_0|\mathbf{X}] = P[\theta \leq \theta_0|\mathbf{X}].$$

Since $\pi(\theta|\mathbf{x})$ is symmetric, this is true iff mean of $\pi(\theta|\mathbf{x})$ is less than or equal to θ_0. Hence, H_0 will be accepted as true if $\bar{X} \leq \theta_0 + \sigma^2(\theta_0 - \mu)/n/\tau^2$ and H_1 will be accepted as true otherwise. In the special case of $\mu = \theta_0$ then H_0 will be accepted as true if $\bar{x} \leq \theta_0$ and H_1 accepted otherwise. □

The union-intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say $H_0 : \theta \in \bigcap_{\gamma\in\Gamma} \Theta_\gamma$. Here Γ is an arbitrary index set that may be finite or infinite. Suppose the tests are available for each of the problems of testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Say the rejection region for the test of $H_{0\gamma}$ is $\{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$. Then the rejection region for the union-intersection test is $\bigcup_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$. For an example, suppose that each of the individual tests has a rejection region of the form $\{\mathbf{x} : T_\gamma(\mathbf{x}) > c\}$, where c does not depend on γ. The rejection region for the union-intersection test can be expressed as $\bigcup_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) > c\} = \{\mathbf{x} : \sup_{\gamma\in\Gamma} T_\gamma(\mathbf{x}) > c\}$. Thus the test statistic for testing H_0 is $T(\mathbf{x}) = \sup_{\gamma\in\Gamma} T_\gamma(\mathbf{x})$.

Example 8.4. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ population. Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, where μ_0 is a specified number. We can write H_0 as the intersection of two sets, $H_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}$. The LRT of $H_{0L} : \mu \leq \mu_0$ versus $H_{1L} : \mu > \mu_0$ is - reject H_{0L} if $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \geq t_L$. Similarly, the LRT of $H_{0U} : \mu \geq \mu_0$ versus $H_{1U} : \mu < \mu_0$ is - reject H_{0U} if $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \leq t_U$. Thus the union-intersection test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ formed from these two LRT is reject H_0 if $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \geq t_L$ or $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \leq t_U$. If $t_L = -t_U \geq 0$, the union-intersection test can be more simply expressed as - reject H_0 if $\frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} \geq t_L$. This is also the LRT of this problem and is called the two-sided t test. □

Another method, intersection-union method, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis $H_0 : \theta \in \bigcup_{\gamma\in\Gamma} \Theta_\gamma$. Suppose that for each $\gamma \in \Gamma, \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Then the rejection region for the intersection-union test of H_0 versus H_1 is $\bigcap_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$. H_0 is false iff all of the $H_{0\gamma}$ are false, so H_0 can be rejected iff each of the individual $H_{0\gamma}$ can be rejected. The rejection regions are of the form $\{\mathbf{x} : T_\gamma(\mathbf{x}) \geq c\}$, with c independent of γ. The rejection region for H_0, thus, is $\bigcap_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) \geq c\} = \{x : \inf_{\gamma\in\Gamma} T_\gamma(\mathbf{x})\}$. Hence, the intersection-union test statistic is $T(\mathbf{x}) = \inf_{\gamma\in\Gamma} T_\gamma(\mathbf{x})$, and the test rejects H_0 for large values of this statistic.

Example 8.5. For a product two parameters θ_1 and θ_2 should be above 50 and 0.95 respectively for the test to pass. The right hypothesis test would be $H_0 : \{\theta_1 \leq 50 \text{ or } \theta_2 \leq 0.95\}$ versus $H_1 : \{\theta_1 > 50 \text{ and } \theta_2 > 0.95\}$. Suppose $X_1, \ldots, X_n$ are measurement from $\mathcal{N}(\theta_1, \sigma^2)$. The LRT of $H_{01} : \theta \leq 50$ will reject H_{01} if $(\bar{X} - 50)/(S/\sqrt{n}) > t$. Similarly, we also have $Y_1, \ldots, Y_m$ from $Bernoulli(\theta_2)$. The LRT will reject $H_{02} : \theta_2 \leq 0.95$ if $\sum Y_i > b$. The rejection region for the intersection-union test is $\{(\mathbf{x}, \mathbf{y}) : \frac{\bar{x}-50}{s/\sqrt{n}} > t \text{ and } \sum_{i=1}^{m} y_i > b\}$. Thus the intersection-union test decides the product is acceptable, that is H_1 is true, if and only if it decides that each of the individual parameters meets its standard, that is H_{1i} is true. If more than two parameters define a product's quality, individual tests for each parameter can be combined, by means of the inter section-union method, to yield an overall test of the product's quality. □

8.2 Methods of Evaluating Tests

Hypothesis tests are evaluated and compared through their probabilities of making mistakes.

8.2.1 Error Probabilities and Power Function

A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ might make one of the two types of errors as shown in table 2.

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	correct decision	Type 1 error
	H_1	Type 2 error	correct decision

Table 2: Two types of errors in hypothesis testing.

We notice for R as the rejection region for a test,

$$P_\theta[\mathbf{X} \in R] = \begin{cases} P[\text{Type-1 error}] & \text{if } \theta \in \Theta_0 \\ 1 - P[\text{Type-2 error}] & \text{if } \theta \in \Theta_0^c. \end{cases}$$

Definition 8.4. *The power function of a hypothesis test with rejection region R is a function of θ defined by $\beta(\theta) = P_\theta[\mathbf{X} \in R]$.*

The ideal power function is close to 0 for all $\theta \in \Theta_0$ and close to 1 for all $\theta \in \Theta_0^c$.

Example 8.6. Let $X \sim Binomial(5, \theta)$ Consider testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$. Consider the test that rejects H_0 iff all successes are observed. The power function for this test is $\beta_1(\theta) = P_\theta[X \in R] = P_\theta[X = 5] = \theta^5$ (fig. 2). The two types of error are shown in the red curve. Clearly for $\theta \leq \frac{1}{2}$ type 1 error is low but for $\theta > \frac{1}{2}$ type 2 error is quite high. We might consider using the test that rejects H_0 if $X = 3, 4, 5$. The power function of this test is $\beta_2(\theta) = P_\theta[X = 3, 4, 5] = \binom{5}{3}\theta^3(1-)^2 + \binom{5}{4}\theta^4(1 = \theta) + \binom{5}{5}\theta^5(1-\theta)^0$. The blue curves shows the type 1 and type 2 error. Now for $\theta > \frac{1}{2}$ the type 2 error is reduced, but at the expense of increasing the type 1 error for $\theta \leq \frac{1}{2}$. The researcher must decide which error structure is more acceptable. □

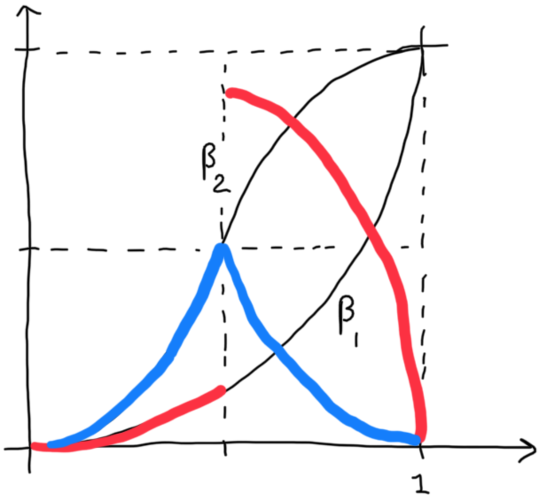

Figure 2: Type 1 and Type two errors for Binomial example.

Typically, the power function of a test will depend on the sample size n. If n can be chosen by the experimenter, consideration of the power function might help determine what sample size is appropriate in an experiment.

Example 8.7. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\theta, \sigma^2)$ population, σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$. The power function for this test is

$$\beta(\theta) = P_\theta\left[\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right] = P_\theta\left[\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right] = P\left[Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right],$$

where Z is a standard normal random variable, since $(\bar{X} - \theta)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0, 1)$. $\beta(\theta)$ is an increasing function of θ as shown in fig. 3. Notice that $\beta(\theta_0) = \alpha = P[Z > c]$.

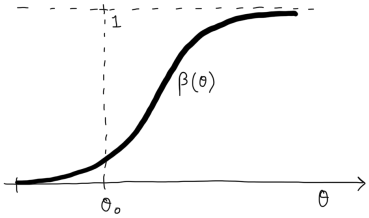

Figure 3: Power function of Normal.

Suppose we want to have the maximum Type 1 error probability of 0.1 and the maximum Type 2 error probability of 0.2 if $\theta \geq \theta_0 + \sigma$. This requirement is met if $\beta(\theta_0) = 0.1$ and $\beta(\theta_0 + \sigma) = 0.8$. For $\beta(\theta_0) = P[Z > 1.28] = 0.1$ gives us $c = 1.28$. Now we also see $\beta(\theta_0 + \sigma) = P[Z > 1.28 - \sqrt{n}] = 0.8$. For $P[Z > -0.84] = 0.8$, so setting $1.28 - \sqrt{n} = -0.84$ gives $n = 4.49$ and hence $n = 5$ yields the error probabilities controlled as specified by the experimenter. □

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. We then search for a good test which restricts type-1 error probability at a specified level. Within this class of tests we then search for the tests that minimizes type-2 errors.

Definition 8.5. *For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta\in\Theta_0}\beta(\theta) = \alpha$, and is a level α test if $\sup_{\theta\in\Theta_0}\beta(\theta) \leq \alpha$.*

The set of level α tests contains the set of size α tests. Experimenters commonly specify the level of the test they wish to use, with typical choices being α =0.01, 0.05, and 0.10. In fixing the level of the test, we are controlling only for Type-1 error probabilities, not Type-2 error. If this approach is taken, the experimenter should specify the null and alternative hypotheses so that it is most important to control the Type-1 error probability. If one expects an experiment to give support to a particular hypothesis, but we want to claim it only after the data gives convincing support, we can set the 'research hypothesis' as the alternative hypothesis. By using a level α test with small α, we are guarding against saying the data support the research hypothesis when it is false.

Example 8.8. We have a normal random sample from $\mathcal{N}(\theta, 1)$, with $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. An LRT is a test that reject H_0 for small values of $\lambda(\mathbf{x})$, with the rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$. We choose c such that $\sup_{\theta\in\Theta_0} P_\theta[\lambda(\mathbf{X}) \leq c] = \alpha$. For, the case of normal distribution with $\theta = \theta_0$ and $\sqrt{n}(\bar{X} - \theta_0) \sim N(0, 1)$. So the test is to reject H_0 if $|\bar{X} - \theta_0| \geq z_{\alpha/2}/\sqrt{n}$, where $z_{\alpha/2}$ satisfies $P[Z \geq z_{\alpha/2}] = \alpha/2$ with $Z \sim \mathcal{N}(0, 1)$, is the size α LRT, corresponding to $c = \exp(-z_{\alpha/2}/2)$.

For the case where the variance is unknown too, the problem of finding a size α union-intersection test involves finding constants t_L and t_U such that

$$\sup_{\theta\in\Theta_0} P_\theta\left[\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_L \text{ or } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_U\right] = \alpha.$$

But for any $(\mu, \sigma^2) = \theta \in \Theta_0$, $\mu = \mu_0$ and thus $(\bar{X} - \mu_0)/S/\sqrt{n}$ has a Student's t distribution with $n - 1$ degrees of freedom. So any choice of $t_U = t_{n-1,1-\alpha_1}$ and $t_L = t_{n-1,\alpha_2}$, with $\alpha_1 + \alpha_2 = \alpha$, will yield a test with Type-1 error probability of exactly α for all $\theta \in \Theta_0$. The usual choice is $t_L = -t_U = t_{n-1,\alpha/2}$. □

In the above $z_{\alpha/2}$ is used to denote the point having probability $\alpha/2$ to the right of it for a standard normal pdf. We would also like a test to be more likely to reject H_0 if $\theta \in \Theta_0^c$ than if $\theta \in \Theta_0$, giving the unbiased property.

Definition 8.6. *A test with power function $\beta(\theta)$ is unbiased if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$.*

In most problems there are many unbiased tests. Likewise, there are many size α test, likelihood ratio tests, etc. We will now discuss criteria for selecting one out of a class of tests.

8.2.2 Most Powerful Tests

Definition 8.7. *Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.*

For our analysis the class $\mathcal{C}$ is the class of all level α tests. UMP may not exist at all, but if they do we would want to identify them.

Theorem 8.2. *Neyman-Pearson Lemma: Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the distribution corresponding to θ_i is $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies $\mathbf{x} \in R$ if $f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$ and $\mathbf{x} \in R^c$ if $f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0)$, for some $k \geq 0$, and $\alpha = P_{\theta_0}[\mathbf{X} \in R]$. Then,*

- *Sufficiency: Any test that satisfies these conditions is a UMP level α test.*
- *Necessity: If there exists a test satisfying these conditions with $k > 0$, then every UMP level α test is a size α test and every UMP level α test satisfies these conditions, except perhaps on a set A satisfying $P_{\theta_0}[\mathbf{X} \in A] = P_{\theta_1}[\mathbf{X} \in A] = 0$.*

Example 8.9. Let $X \sim \mathit{Binomial}(2, \theta)$. We want to test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta = \frac{3}{4}$. The possible values of $f(x|\theta_1)/f(x|\theta_0)$ are $1/4, 3/4, 9/4$ for $x = 0, 1, 2$ respectively. If we choose $\frac{3}{4} < k < \frac{9}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if $X = 2$ is the UMP level $\alpha = P[X = 2| = \frac{1}{2}] = \frac{1}{4}$ test. If we choose $\frac{1}{4} < k\frac{3}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if $X = 1$ or 2 is the UMP level $\alpha = P[X = 1 \text{ or } 2|\theta = \frac{1}{2}] = \frac{3}{4}$ test. Choosing $k < \frac{1}{4}$ or $k > \frac{9}{4}$ yields the UMP level $\alpha = 1$ or $\alpha = 0$ test. □

Theorem 8.3. *Consider the previous hypothesis problem. Suppose $T(\mathbf{X})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the distribution of T corresponding to $\theta_i, i = 0, 1$. Then any test based on T with rejection region S, a subset of the sample space of T, is a UMP level α test if it satisfies $t \in S$ if $g(t|\theta_1) > kg(t|\theta_0)$ and $t \in S^c$ if $g(t|\theta_1) < kg(t|\theta_0)$, for some $k \geq 0$, where $\alpha = P_{\theta_0}[T \in S]$.*

Hypotheses that specify only one possible distribution for the same X, like above, are called simple hypotheses. Hypotheses with more than one possible distribution for the sample are called composite hypotheses. Similarly we have one sided ($\theta \geq \theta_0$ or $\theta < \theta_0$) and two-sided hypotheses ($\theta \neq \theta_0$).

Definition 8.8. *A family of distributions $\{g(t|\theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a monotone likelihood ratio, MLR, if, for every $\theta_2 > \theta_1$, $g(t|\theta_2)/g(t|\theta_1)$ is a monotone function of t on $\{t : g(t|\theta_1) > 0$ or $g(t|\theta_2) > 0\}$.*

A large class of problems that admit UMP level α test involve one-sided hypotheses and distributions with the monotone likelihood ratio property. Any regular exponential family with $g(t|\theta) = h(t)c(\theta)e^{w(\theta)t}$ has an MLR if $w(\theta)$ is a non-decreasing function. For example, the normal (known variance), Poisson, and Binomial all have an MLR.

Theorem 8.4. *Karlin-Rubin: Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of distributions $\{g(t|\theta) : \theta \in \Theta\}$ of T has an MLR. Then for any t_0, the test that rejects H_0 iff $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}[T > t_0]$.*

By an analogous argument, the test that rejects $H_0 : \theta \geq \theta_0$ in favor of $H_1 : \theta < \theta_0$ iff $T < t_0$ is a UMP level $\alpha = P_{\theta_0}[T < t_0]$ test.

Example 8.10. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\theta, \sigma^2)$ population with σ^2 known. The sample mean $\bar{X}$ is a sufficient statistic for θ. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$. The conditions of Neyman-Pearson Lemma, $g(\bar{x}|\theta_1) > kg(\bar{x}|\theta_0)$, is equivalent to $\bar{x} < (2\sigma^2 \log k/n - \theta_0^2 + \theta_1^2)/(\theta_1 - \theta_0)/2$. The right hand side increases from $-\infty$ to ∞ as k increases from 0 to ∞. Thus, the test with rejection region $\bar{x} < c$ is the UMP level α test, where $\alpha = P_{\theta_0}[\bar{x} < c]$. If a particular α is specified, then the UMP test rejects H_0 if $\bar{X} < c = -\sigma z_\alpha/\sqrt{n} + \theta_0$.

Now consider testing $H_0' : \theta \geq \theta_0$ versus $H_1' : \theta < \theta_0$ using the test that rejects H_0' if $\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0$. As $\bar{X}$ is sufficient and its distribution has an MLR, it follows that the test is a UMP level α test in this problem. As the power function of this test, $\beta(\theta) = P_\theta[\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0]$, is a decreasing function of θ, the value of α is given by $\sup_{\theta \geq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$. □

When no UMP test exists because the class of level α tests if so large that no one test dominates all the others in terms of power, we restrict attention to unbiased tests and find the best test.

Example 8.11. Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\theta, \sigma^2)$, σ^2 known. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. For a specified value of α, a level α test in this problem is any test that satisfies $P_{\theta_0}[\text{reject } H_0] \leq \alpha$. For any $\theta_1 < \theta_0$, among all tests the test that rejects H_0 $\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0$ has the highest possible power at θ_1. Call this Test 1. Now consider Test 2, which rejects H_0 if $\bar{X} > \sigma z_\alpha/\sqrt{n} + \theta_0$. This is also a level α test. It can be shown that for $\beta_i(\theta)$ denoting the power function of Test i, for $\theta_2 > \theta_0$, $\beta_2(\theta_2) > \beta_1(\theta_2)$, showing the tests are not UMP.

When no UMP level α test exists within the class of all tests, we might try to find a UMP level α test within the class of unbiased tests. The power function $\beta_3(\theta)$, of Test 3, which rejects H_0 iff $\bar{X} > \sigma z_{\alpha/2}/\sqrt{n} + \theta_0$ or $\bar{X} < -\sigma z_{\alpha/2}/\sqrt{n} + \theta_0$ is shown in figure 4. Test 3 is actually a UMP unbiased level α test. □

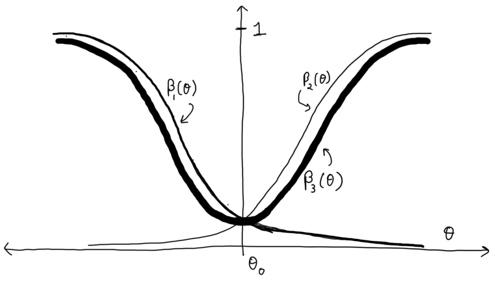

Figure 4: Power functions for three tests.

8.2.3 Sizes of Union-Intersection and Intersection-Union Tests

Theorem 8.5. *We are testing a null hypothesis of the form $H_0 : \theta \in \Theta_0$, where $\Theta_0 =_{\gamma\in\Gamma} \Theta_\gamma$. Let $\lambda_\gamma(\mathbf{x})$ be the LRT statistic for testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$, and let $\lambda(\mathbf{x})$ be the LRT statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. We define $T(\mathbf{x}) = \inf_{\gamma\in\Gamma} \lambda_\gamma(\mathbf{x})$, and form the union-intersection tests (UIT) with rejection region $\{\mathbf{x} : \lambda_\gamma(\mathbf{x}) < c$ for some $\gamma \in \Gamma\} = \{\mathbf{x} : T(\mathbf{x}) < c\}$. The usual LRT has the rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) < c\}$. Then*

1. *$T(\mathbf{x}) \geq \lambda(\mathbf{x})$ for every $\mathbf{x}$.*

2. *If $\beta_T(\theta)$ and $\beta_\lambda(\theta)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$ for every $\theta \in \Theta$.*

3. *If the LRT is a level α test, then the UIT is a level α test.*

Theorem 8.6. *For intersection-union tests (IUT) the null hypothesis is expressible as a union, $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 = \bigcup_{\gamma\in\Gamma} \Theta_\gamma$. The rejection region is of the form $R = \bigcap_{\gamma\in\Gamma} R_\gamma$, where R_γ is the rejection region for the test $H_{0\gamma} : \theta \in \Theta_\gamma$. Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region R_γ. Then the IUT with rejection region $R = \bigcap_{\gamma\in\Gamma} R_\gamma$ is a level $\alpha = \sup_{\gamma\in\Gamma} \alpha_\gamma$ test.*

The IUT is a level α test and may be very conservative as stated above.

Theorem 8.7. *Consider testing $H_0 : \theta \in \bigcup_{j=1}^{k} \Theta_j$, where k is a finite positive integer. For each $j = 1, \ldots, k$, let R_j be the rejection region of a level α test of H_{0j}. Suppose that for some $i = 1, \ldots, k$ there exists a sequence of parameter points, $\theta_l \in \Theta_i, l = 1, 2, \ldots$, such that*

1. *$\lim_{t\to\infty} P_{\theta_l}[\mathbf{X} \in R_i] = \alpha$.*

2. *for each $j = 1, \ldots, k$, $j \neq i$, $\lim_{l\to\infty} P_{\theta_l}[\mathbf{X} \in R_j] = 1$.*

Then, the IUT with rejection region $R = \bigcap_{j=1}^{k} R_j$ is a size α test.

8.2.4 p-Values

One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. If α is small, the decision to reject H_0 is very convincing. Another way of reporting the results of a hypothesis test is to report the test statistic called a p-value.

Definition 8.9. *A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$, $P_\theta[p(\mathbf{X}) \leq \alpha] \leq \alpha$.*

The test that rejects H_0 iff $p(\mathbf{X}) \leq \alpha$ is a level α test.

Theorem 8.8. *Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define $p(\mathbf{x}) = \underset{\theta\in\Theta_0}{sup} P_\theta[W(\mathbf{X}) \geq W(\mathbf{x})]$. Then. $p(\mathbf{X})$ is a valid p-value.*

Example 8.12. Let $X_1, \ldots, X_n$ be a normal sample from a $\mathcal{N}(\mu, \sigma^2)$ population. For $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, the LRT rejects H_0 for large values of $W(\mathbf{X}) = |\bar{X} - \mu_0|/(S/\sqrt{n})$. If $\mu = \mu_0$, regardless of the value of σ, $W(\mathbf{X})$ has a Student's t distribution with $n - 1$ degrees of freedom. Thus, $p(\mathbf{x}) = \sup_\sigma P_\sigma[W(\mathbf{X}) \geq W(\mathbf{x})]$ is independent of σ. Thus, the p-value for this two-sided t test is $p(\mathbf{x}) = 2P[T_{n-1} > |\bar{x} - \mu_0|/(S/\sqrt{n})]$, where T_{n-1} has a Student's t distribution with $n - 1$ degree of freedom. □

Example 8.13. Again consider the normal model, but testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. The LRT rejects H_0 for large values of $W(\mathbf{X}) = (\bar{X} - \mu_0)/(S/\sqrt{n})$. For this statistic, the supremum always occurs at the parameter (μ_0, σ), and the value of σ does not matter. Consider any $\mu \leq \mu_0$ and any σ, then

$$\begin{aligned} P_{\mu,\sigma}[W(\mathbf{X}) \geq W(\mathbf{x})] &= P_{\mu,\sigma}\left[\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq W(\mathbf{x})\right] \\ &= P_{\mu,\sigma}\left[\frac{\bar{X} - \mu}{S/\sqrt{n}} \geq W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right] \\ &= P_{\mu,\sigma}\left[T_{n-1} \geq W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right] \\ &\leq P[T_{n-1} \geq W(\mathbf{x})]. \end{aligned}$$

The last line follows from the fact that $\mu_0 \geq \mu$ and $(\mu_0 - \mu)/(S/\sqrt{n})$ is a non negative random variable, and the probability does not depend on (μ, σ). Furthermore,

$$P[T_{n-1} \geq W(\mathbf{x})] = P_{\mu_0,\sigma}\left[\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq W(\mathbf{x})\right] = P_{\mu_0,\sigma}[W(\mathbf{X}) \geq W(\mathbf{x})],$$

Thus, the p-value for this one-sided t test is $p(\mathbf{x}) = P[T_{n-1} \geq W(\mathbf{x})] = P[T_{n-1} \geq (\bar{x} - \mu_0)/(s/\sqrt{n})]$. □

Another way to define p-value is to condition on a sufficient statistic. Suppose $S(\mathbf{X})$ is a sufficient statistic for the model $\{f(\mathbf{x}|\theta) : \theta \in \Theta_0\}$, with S being sufficient only for the null model, not the entire model. If the null hypothesis is true, the conditional distribution of $\mathbf{X}$ given $S = \mathbf{s}$ does not depend on θ. let $W(\mathbf{X})$ denote a test statistic for which large values give evidence that H_1 is true. Then for each sample point $\mathbf{x}$ define $p(\mathbf{x}) = P[W(\mathbf{x}) \geq W(\mathbf{x})|S = S(\mathbf{x})]$. We see for any $0 \leq \alpha \leq 1$, $P[p(\mathbf{X}) \leq \alpha|S = s] \leq \alpha$. Then for any $\theta \in \Theta_0$, unconditionally we have $P_\theta[p(\mathbf{X}) \leq \alpha] = \sum_s P[p(\mathbf{X}) \leq \alpha|S = s]P_\theta[S = s] \leq \sum_s \alpha P_\theta[S = s] \leq \alpha$. Thus $p(\mathbf{X})$ is a valid p-value.

8.2.5 Loss Function Optimality

A decision theoretic analysis may also be used to compare hypothesis tests. In a hypothesis testing problem, only two actions are allowable, a_0 ='accept H_0' or a_1 ='reject H_0', the action space $\mathcal{A} = \{a_0, a_1\}$. A decision rule $\delta(\mathbf{x})$ is a function of $\mathcal{X}$ that takes on only two values a_0, and a_1. The set $\{\mathbf{x} : \delta(\mathbf{x} = a_0)\}$ is the acceptance region for the test and the set

$\{\mathbf{x} : \delta(\mathbf{x} = a_1)\}$ is the rejection region. The loss function should minimize the losses $L(\theta, a_0)$, type 1 and $L(\theta, a_1)$, type 2. A generalized 0-1 loss is given by

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_2 & \theta \in \Theta_0^c \end{cases} \text{ and } L(\theta, a_1) = \begin{cases} c_1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}.$$

Only the ratio c_2/c_1 matters. The risk function is used to evaluate a hypothesis testing procedure, which is closely related to the power function. Let $\beta(\theta)$ be the power function of the test based on the decision rule δ. The risk function can be written as

$$R(\theta, \delta) = \begin{cases} c_1\beta(\theta) & \text{if } \theta \in \Theta_0 \\ c_2(1 - \beta(\theta)) & \text{if } \theta \in \Theta_0^c \end{cases}.$$

The expected loss is $R(\theta, \delta) = L(\theta, a_0)(1 - \beta(\theta)) + L(\theta, a_1)\beta(\theta)$. Hence, in decision theoretic analysis, weights given by the loss function, apart from the power function is also important.

Sometimes some wrong decisions are more serious than others and the loss function should reflect this. When we test $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, it is the type-1 error to reject H_0 if θ is slightly bigger than θ_0, but it may not be a very serious mistake. The adverse consequences of rejecting H_0 may be much worse if θ is much larger than θ_0. A loss function that reflects this is

$$L(\theta, a_0) = \begin{cases} 0 & \theta \geq \theta_0 \\ b(\theta_0 - \theta) & \theta < \theta_0 \end{cases} \text{ and } L(\theta, a_1) = \begin{cases} c(\theta - \theta_0)^2 & \theta \geq \theta_0 \\ 0 & \theta < \theta_0 \end{cases},$$

where b and c are positive constants.

9 Interval Estimation

Definition 9.1. *An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_a, \dots, x_n)$ and $U(x_a, \dots, x_n)$ of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. If $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an interval estimator.*

By giving up some precision of point estimate, we have gained some confidence with an interval.

Definition 9.2. *For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ, the coverage probability of $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that the random interval covers the true parameter θ, i.e. $P_\theta[\theta \in$[L(X), U(X)]]. The confidence coefficient of $[L(\mathbf{X}), U(\mathbf{X})]$ is the infimum of the coverage probabilities, $\inf_\theta P_\theta[\theta \in$[L(X), U(X)]].*

The interval is a random quantity, not a parameter, hence these probability statement refer to $\mathbf{X}$ and not θ. Interval estimators, together with confidence coefficient are known as confidence intervals. Since we do not know the true value of θ, we can only guarantee a coverage probability equal to the confidence coefficient.

Example 9.1. Let $X_1, \dots, X_n$ be a random sample from $Uniform(0, \theta)$ population and let $Y = max\{X_1, \dots, X_n\}$. We consider two interval estimators $[aY, bY]$, $1 \leq a < b$, and $[Y+c, Y+d]$, $0 \leq c < d$, where a, b, c, and d are specified constants. Note that θ is necessarily larger than y. For the first interval we have

$$P_\theta[\theta \in [aY, bY]] = P_\theta[aY \leq \theta \leq bY] = P_\theta\left[\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right] = P_\theta\left[\frac{1}{b} \leq T \leq \frac{1}{a}\right],$$

where $T = Y/\theta$. Now $f_Y(y) = ny^{n-1}/\theta^n$, $0 \leq y \leq \theta$, so the pdf of T is $f_T(t) = nt^{n-1}$, $0 \leq t \leq 1$. We therefore have $P_\theta\left[\frac{1}{b} \leq T \leq \frac{1}{a}\right] = \int_{1/b}^{1/a} nt^{n-1}dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$. This coverage probability is independent of the value of θ, and thus $\frac{1}{a^n} - \frac{1}{b^n}$ is the confidence coefficient of the interval.

For the other interval, for $\theta \geq d$ a similar calculation yields $P_\theta[\theta \in [Y+c, Y+d]] = P_\theta[Y + c \leq \theta \leq Y + d] = P_\theta\left[1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right] = \int_{1-d/\theta}^{1-c/\theta} nt^{n-1}dt = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$. This coverage probability depends on θ. But noting that $\lim\limits_{\theta \to \infty} \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n = 0$, shows that the confidence coefficient of this interval estimator is 0. □

9.1 Methods of Finding Interval Estimators

9.1.1 Inverting a Test Statistic

There is a strong correspondence between hypothesis testing and interval estimation, every confidence set corresponds to a test and vice versa.

Example 9.2. Let $X_1, \dots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. For a fixed α level, the most powerful unbiased test has rejection region

$\{x : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$. Hence, H_0 is accepted for sample points with $\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$. Since the test is size α, this means $P[H_0 \text{is rejected}|\mu = \mu_0] = \alpha$, or $P[H_0 \text{is accepted}|\mu = \mu_0] = 1 - \alpha$. This and the fact that it is true for every μ_0, gives,

$$P_\mu \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha.$$

The interval $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$, obtained by inverting the acceptance region of the level α test, is a $1 - \alpha$ confidence interval. The correspondence between the testing and interval estimation for the two-sided normal problem is illustrated in fig. 5. The hypothesis test fixes the parameter and asks what sample values are consistent with that fixed value. The confidence set fixes the sample value and asks what parameter values make this sample value most plausible. □

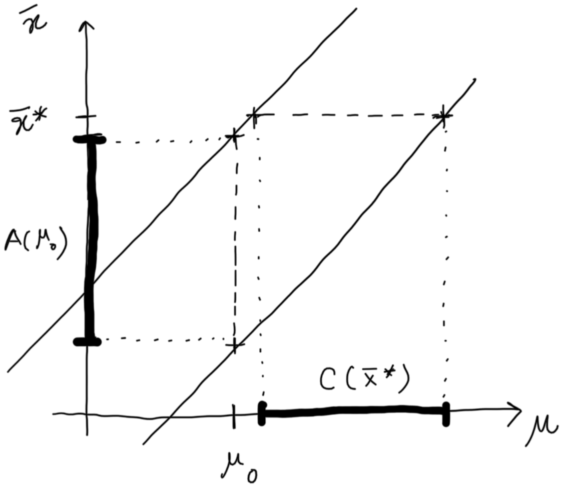

Figure 5: Relationship between confidence intervals and acceptance regions for tests.

Theorem 9.1. *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\mathbf{x} \in \mathcal{X}$, define a set $C(\mathbf{x})$ in the parameter space by $C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}$. Then the random set $C(\mathbf{X})$ is a $1 - \alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define $A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$. Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.*

All that is required of the acceptance region is based on H_0, i.e. $P_\theta[\mathbf{X}(\theta_0)] \geq 1 - \alpha$. We will also have in mind an alternative hypothesis such as $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$. The alternative with dictate the form of $A(\theta_0)$ that is reasonable, and the form of $A(\theta_0)$ will determine the shape of $C(\mathbf{x})$. There is no guarantee that the confidence set would be an interval, though it is often the case. Unbiased tests, when inverted, will produce unbiased confidence intervals. We can confine our attention to sufficient statistics when looking for a good test, this carries over to good confidence set too.

Example 9.3. For an exponential with mean λ we can obtain a confidence interval by inverting a level α test of $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$. For a random sample $X_1, \ldots, X_n$, the LRT statistic is given by

$$\frac{\sup\limits_{\lambda=\lambda_0} \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}}}{\sup\limits_{\lambda} \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}}} = \left(\frac{\sum x_i}{n\lambda_0}\right)^n e^{n - \frac{\sum x_i}{\lambda_0}}.$$

For a fixed λ_0, the acceptance region is given by

$$A(\lambda_0) = \left\{\mathbf{x} : \left(\frac{\sum x_i}{\lambda_0}\right)^n e^{-\sum x_i/\lambda_0} \geq k^*\right\},$$

where k^* is chosen to satisfy $P_{\lambda_0}[\mathbf{X} \in A(\lambda_0)] = 1 - \alpha$. Inverting this acceptance region gives the $1 - \alpha$ confidence set

$$C(\mathbf{x}) = \left\{\lambda : \left(\frac{\sum x_i}{\lambda}\right)^n e^{-\sum x_i/\lambda} \geq k^*\right\}.$$

The expression defining $C(\mathbf{x})$ depends on $\mathbf{x}$ only through $\sum x_i$. So the confidence interval can be expressed in the form

$$C(\textstyle\sum x_i) = \left\{\lambda : L(\textstyle\sum x_i) \leq \lambda \leq U(\textstyle\sum x_i)\right\},$$

where L and U are functions determined by the requirement that acceptance region has probability $1 - \alpha$ and

$$\left(\frac{\sum x_i}{L(\sum x_i)}\right)^n e^{-\sum x_i / L(\sum x_i)} = \left(\frac{\sum x_i}{U(\sum x_i)}\right)^n e^{-\sum x_i / U(\sum x_i)}.$$

This can be solved numerically to generated acceptance and confidence set as shown in fig. 6.

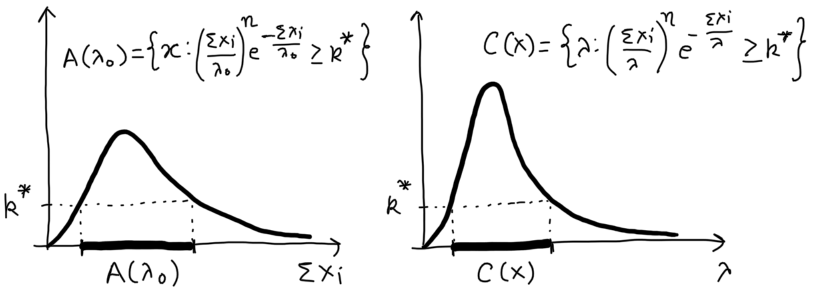
$A(\lambda_0)=\{x:\left(\frac{\Sigma x_i}{\lambda_0}\right)^n e^{-\frac{\Sigma x_i}{\lambda_0}} \geq k^*\}$
k^*
$A(\lambda_0)$
Σx_i
$C(x)= \{\lambda: \left(\frac{\Sigma x_i}{\lambda}\right)^n e^{-\frac{\Sigma x_i}{\lambda}} \geq k^*\}$
k^*
$C(x)$
λ

Figure 6: Acceptance region and confidence interval for the exponential distribution with mean λ and hypothesis $H_0 : \lambda = \lambda_0$ and $H_1 : \lambda \neq \lambda_0$.

Example 9.4. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population. To construct a $1-\alpha$ upper confidence bound on μ of the form $(-\infty, U(\mathbf{x})]$, we invert the one-sided tests of $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$. The size α LRT of H_0 versus H_1 rejects H_0 if $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{n-1,\alpha}$. Thus the acceptance region for this test is $A(\mu_0) = \{\mathbf{x} : \bar{x} \geq \mu_0 - t_{n-1,\alpha}s/\sqrt{n}\}$ and $\mathbf{x} \in A(\mu_0) \iff \bar{x} + t_{n-1,\alpha}s/\sqrt{n}$. We then define $C(\mathbf{x}) = \{\mu_0 : \mathbf{x} \in A(\mu_0)\} = \{\mu_0 : \bar{x} + t_{n-1,\alpha}s/\sqrt{n} \geq \mu_0\}$. The random set $C(\mathbf{X}) = (-\infty, \bar{X} + t_{n-1,\alpha}S/\sqrt{n}]$ is a $1-\alpha$ confidence set for μ. □

9.1.2 Pivotal Quantities

Definition 9.3. *A random variable $Q(\mathbf{X}, \theta) = Q(X_1, \ldots, X_n, \theta)$ is a pivotal quantity (or pivot) if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. That is, if $\mathbf{X}(\mathbf{x}|\theta)$, then $Q(\mathbf{X}, \theta)$ has the same distribution for all values of θ.*

It is relatively easy task to find pivots for location or scale parameters. In general, differences are pivotal for location problems, while ratios are pivotal for scale problems. For example, for a $\mathcal{N}(\mu, \sigma^2)$, the t statistic $(\bar{X} - \mu)/(S/\sqrt{n})$ is a pivot because the t distribution does not depend on the parameters μ and σ^2.

Example 9.5. Gamma pivot: Suppose that $X_1, \ldots, X_n$ are iid exponential(λ). Then $T =_i$ is a sufficient statistic for λ and $T \sim gamma(n, \lambda)$. The pdf $\frac{t^{n-1}e^{-t/\lambda}}{\Gamma(n)\lambda^n}$ is a scale family. Thus, if $Q(T, \lambda) = 2T/\lambda$, then $Q(T, \lambda) \sim gamma(n, 2) = \mathcal{X}^2_{2n}$, which does not depend on λ. The quantity $2T/\lambda$ is a pivot with $\mathcal{X}^2_{2n}$ distribution. □

In general, suppose the pdf of a statistic T, $f(t|\theta)$, can be expressed in the form $f(t|\theta) = g(Q(t, \theta))|\frac{\partial}{\partial t}Q(t, \theta)|$ for some function g and some monotone function Q, then $Q(T, \theta)$ is a pivot. If $Q(\mathbf{X}, \theta)$ is a pivot, then for a specified value of α we can find numbers a and b, which do not depend on θ, to satisfy, $P_\theta[a \leq Q(\mathbf{X}, \theta) \leq b] \geq 1 - \alpha$. Then, for each $\theta_0 \in \Theta$, $A(\theta_0) = \{\mathbf{x} : a \leq Q(\mathbf{x}, \theta_0) \leq b\}$ is the acceptance region for a level α test of $H_0 : \theta = \theta_0$. We invert these tests to obtain $C(\mathbf{x}) = \{\theta : a \leq Q(\mathbf{x}, \theta) \leq b\}$, and $C(\mathbf{X})$ is a $1-\alpha$ confidence set for θ. If θ is a real-valued parameter and if, for each $\mathbf{x} \in \mathcal{X}$, $Q(\mathbf{x}, \theta)$ is a monotone function or θ, then $C(\mathbf{x})$ will be an interval. In fact, if $Q(\mathbf{x}, \theta)$ is an increasing function of θ, then $C(\mathbf{x})$ has the form $L(\mathbf{x}, a) \leq \theta \leq U(\mathbf{x}, b)$. If $Q(\mathbf{x}, \theta)$ is a decreasing function of θ then $C(\mathbf{x})$ has the form $L(\mathbf{x}, b) \leq \theta \leq U(\mathbf{x}, a)$.

Example 9.6. If we have a sample $X_1, \ldots, X_n$ from an exponential(λ) pdf and the hypothesis $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$, we can define $T =_i$ and $Q(T, \lambda) = 2T/\lambda \sim \mathcal{X}^2_{2n}$. If we choose constants a and b to satisfy $P[a \leq \mathcal{X}^2_{2n} \leq b] = 1 - \alpha$, then $P_\lambda\left[a \leq \frac{2T}{\lambda} \leq b\right] = 1 - \alpha$. Inverting the set $A(\lambda) = \{t : a \leq \frac{2t}{\lambda} \leq b\}$ gives $C(t) = \{\lambda : \frac{2t}{b} \leq \lambda \leq \frac{2t}{a}\}$, which is the $1-\alpha$ confidence interval. □

Example 9.7. Normal pivot interval: If $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, then $Z = (\bar{X} - \mu)(\sigma/\sqrt{n}) \sim \mathcal{N}(0, 1)$ is a pivot. If σ^2 is known, we can use this pivot to calculate a confidence interval of μ. For any constant a, $P[-a \leq Z \leq a] = 1-\alpha$ can be inverted to get the confidence interval $\{\mu : \bar{x} - a\sigma/\sqrt{n} \leq \mu \leq \bar{x} + a\sigma/\sqrt{n}\}$. If σ^2 is unknown, we can use the location-scale pivot $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T_{n-1}$, giving $P[-a \leq T_{n-1} \leq a]$. For any given α, we take $a = t_{n-1,\alpha/2}$, we find that a $1-a\alpha$ confidence interval is given by $\{\mu : \bar{x} - t_{n-1,\alpha/2}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{n-1,\alpha/2}s/\sqrt{n}\}$, which is the classic $1-\alpha$ confidence interval for μ based on Student's t distribution.

To estimate the interval for σ, we use the pivot $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$. If we choose a and b to satisfy $P[a \leq \mathcal{X}^2_{n-1} \leq b] = 1-\alpha$, we can invert this to get the $1-\alpha$ confidence interval $\{\sigma : \sqrt{\frac{(n-1)}{b}}s \leq \sigma \leq \sqrt{\frac{(n-1)}{a}}s\}$. A symmetric choice of $a = \mathcal{X}^2_{n-1,1-\alpha/2}$ and $b = \mathcal{X}^2_{n-1,\alpha/2}$ splits the probability equally, putting $\alpha/2$ in each tail, but is not optimal for a skewed distribution.

To construct the confidence on μ and σ simultaneously we use Bonferroni Inequality, $P[A \cap B] \geq P[A] + P[B] - 1$. Let $A = P[\text{interval covers } \mu]$ and $B = P[\text{interval covers } \sigma^2]$. Using the above intervals with $t_{n-1,\alpha/4}$ to get $1-\alpha/2$ interval for μ, and using $b = \mathcal{X}^2_{n-1,\alpha/4}$ and $a = \mathcal{X}^2_{n-1,1-\alpha/4}$ to get a $1-\alpha/2$ confidence interval for σ we get the combined confidence interval as

$$C_\alpha(x) = \left\{(\mu,\sigma^2) : \left(\bar{x} - \frac{t_{n-1,\alpha/4}}{\sqrt{n}}s \leq \mu \leq \bar{x} + \frac{t_{n-1,\alpha/4}}{\sqrt{n}}s\right) \text{ and } \left(\sqrt{\frac{(n-1)}{\mathcal{X}^2_{n-1,\alpha/4}}}s \leq \sigma \leq \sqrt{\frac{(n-1)}{\mathcal{X}^2_{n-1,1-\alpha/4}}}s\right)\right\}$$

and $P[C_\alpha(\mathbf{X})]$ covers (μ,σ^2), such that $P[A \cap B] \geq P[A] + P[B] - 1 = 2(1-\alpha/2) - 1 = 1-\alpha$. □

9.1.3 Pivoting the CDF

Test inversion method, without extra conditions on the exact types of acceptance regions used, can fail to return continuous interval. We, hence, base our confidence interval construction for a parameter θ on a real-valued statistic T with cdf $F_T(t|\theta)$. By probability Integral Transformation, which tells us that the random variable $F_T(T|\theta)$ is $Uniform(0,1)$, a pivot. Thus, if $\alpha_1 + \alpha_2 = \alpha$, an α-level acceptance region of the hypothesis $H_0 : \theta = \theta_0$ is $\{t : \alpha_1 \leq F_T(t|\theta_0) \leq 1-\alpha_2\}$, with associated confidence set $\{\theta : \alpha_1 \leq F_T(t|\theta) \leq 1-\alpha_2\}$. To guarantee that the confidence set is an interval, we need to have $F_T(t|\theta)$ to be monotone in θ. This is true as as a family of cdfs $F(t|\theta)$ is stochastically increasing (decreasing) in θ if, for each $t \in \mathcal{T}$, the sample space of T, $F(t|\theta)$ is a decreasing (increasing) function of θ, i.e. F is a monotone in θ.

Theorem 9.2. *Pivoting a continuous cdf: let T be a statistic with continuous cdf $F_T(t|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, if $F_T(t|\theta)$ is a decreasing function of θ for each t, we have $F_T(t|\theta_U(t)) = \alpha_1$ and $F_T(t|\theta_L(T)) = 1-\alpha_2$; if $F_T(t|\theta)$ is an increasing function of θ for each t, we have $F_T(t|\theta_U(T)) = 1-\alpha_2$ and $F_T(t|\theta_L(t)) = \alpha_1$. Then the random interval $[\theta_L(T), \theta_U(T)]$ is a $1-\alpha$ confidence interval for θ.*

Example 9.8. Location exponential interval: If $X_1, \ldots, X_n$ are iid with exponential pdf $f(x|\mu) = e^{-(x-\mu)}\mathbf{I}_{[\mu,\infty)}(y)$, then $Y = \min\{X_1, \ldots, X_n\}$ is a sufficient statistic for μ with pdf $f_Y(y|\mu) = ne^{-n(y-\mu)}\mathbf{I}_{[\mu,\infty)}(y)$. Fix α and define $\mu_L(y)$ and $\mu_U(y)$ to satisfy $\int_{\mu_U(y)}^{y} ne^{-n(u-\mu_U(y))} = \frac{\alpha}{2}$, $\int_y^\infty ne^{-n(u-\mu_L(y))du} = \frac{\alpha}{2}$. These integrals can be evaluated to give the equations $1 - e^{-n(y-\mu_U(y))} = \frac{\alpha}{2}$, $e^{-n(y-\mu_L(y))} = \frac{\alpha}{2}$, which gives us the solutions $\mu_U(y) = y + \frac{1}{n}\log\left(1-\frac{\alpha}{2}\right)$,

$\mu_L(y) = y + \frac{1}{n}\log\left(\frac{\alpha}{2}\right)$. Hence, the $1-\alpha$ confidence interval for μ is $C(Y) = \Big\{\mu : Y + \frac{1}{n}\log\left(\frac{\alpha}{2}\right) \leq \mu \leq Y + \frac{1}{n}\log\left(1-\frac{\alpha}{2}\right)\Big\}$. □

Theorem 9.3. *Pivoting a discrete cdf: let T be a statistic with discrete cdf $F_T(t|\theta) = P[T \leq t|\theta]$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, if $F_T(t|\theta)$ is a decreasing function of θ for each t, we have $P[T \leq t|\theta_U(t)] = \alpha_1$ and $P[T \geq t|\theta_L(T)] = 1-\alpha_2$; if $F_T(t|\theta)$ is an increasing function of θ for each t, we have $P[T \geq t|\theta_U(T)] = \alpha_1$ and $P[T \leq t|\theta_L(t)] = \alpha_2$. Then the random interval $[\theta_L(T), \theta_U(T)]$ is a $1-\alpha$ confidence interval for θ.*

Example 9.9. Poisson interval estimator: Let $X_1, \ldots, X_n$ be a random sample from a Poison population with parameter λ and define $Y = \sum X_i$. Y is sufficient for λ and $Y \sim Poisson(n\lambda)$. Applying the above method with $\alpha_1 = \alpha_2 = \alpha/2$, if $Y = y_0$ is observed, we are led to solve for λ in the equations $\sum_{k=0}^{y_0} e^{-n\lambda}\frac{(n\lambda)^k}{k!} = \frac{\alpha}{2} = P[Y \leq y_0|\lambda] = P[\mathcal{X}^2_{2(y_0+1)} > 2n\lambda]$ and $\sum_{k=y_0}^{\infty} e^{-n\lambda}\frac{(n\lambda)^k}{k!} = \frac{\alpha}{2} = P[Y \geq y_0|\lambda] = P[\mathcal{X}^2_{2y_0} < 2n\lambda]$. Thus the solution to get $1-\alpha$ confidence interval for λ is $\Big\{\lambda : \frac{1}{2n}\mathcal{X}^2_{2y_0,1-\alpha/2} \leq \lambda \leq \frac{1}{2n}\mathcal{X}^2_{2(y_0+1),\alpha/2}\Big\}$. □

9.1.4 Bayesian Intervals

With the confidence interval we only mean that we know that $(1-\alpha)\%$ of the sample points of the random interval cover the true parameter, the true parameter is fixed. In contrast, Bayesian setup allows us to say that the parameter is inside the confidence interval with probability $1-\alpha$, the parameter is a random variable. For this reason they are called *credible sets* instead of confidence sets.

If $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ given $\mathbf{X} = \boldsymbol{x}$, then for any set $A \subset \Theta$, the credible probability of A is $P[\theta \in A|\boldsymbol{x}] = \int_A \pi(\theta|\boldsymbol{x})d\theta$, and A is a credible set for θ. The clarity of construction here comes at the cost of additional assumptions and more required inputs than the classical model.

Example 9.10. Let $X_1, \ldots, X_n$ be iid $Poisson(\lambda)$ and assume that λ has a gamma prior $\lambda \sim gamma(a, b)$. The posterior of λ is $\pi(\lambda|\sum X = x) = gamma(a+\sum x, (n+\frac{1}{b})^{-1})$. One simple way to form a credible set for λ is to split α equally between the upper and lower endpoints. Since $\frac{2(nb+1)}{b}\lambda \sim \mathcal{X}^2_{2(a+\sum x_i)}$ the $1-\alpha$ credible interval is $\Big\{\lambda : \frac{b}{2(nb+1)}\mathcal{X}^2_{2(a+\sum x),1-\alpha/2} \leq \lambda \leq \frac{b}{2(nb+1)}\mathcal{X}^2_{2(a+\sum x),\alpha/2}\Big\}$. This credible set is different from the confidence set we calculated in a previous example. The credible set here has somewhat shorter intervals, and the upper endpoints are closer to 0. This reflects the prior, which is pulling the intervals towards 0. □

It is important not to confuse credible probability with converge probability, as they are very different entities, with different meanings and interpretations. Credible probability reflects the experimenter's subjective beliefs, as express in the prior distribution and updated with the data to the posterior distribution. Coverage probability, on the hand, reflects the uncertainty in the sampling procedure, getting its probability from the objective mechanism of repeated experimental trials.

These solutions may be quite different. A Bayes solution is often not reasonable under classical evaluations and vice versa.

Example 9.11. Let $X_1, \dots, X_n$ be iid $\mathcal{N}(\theta, \sigma^2)$, and let θ have the prior pdf $\mathcal{N}(\mu, \tau^2)$, where μ, σ, and τ are all known. The posterior is $\pi(\theta|\bar{x}) \sim \mathcal{N}(\delta^B(\bar{x}), Var(\theta|\bar{x}))$, where $\delta^B(x) = \frac{\sigma^2}{\sigma^2+n\tau^2}\mu + \frac{n\tau^2}{\sigma^2+n\tau^2}\bar{x} = \frac{\gamma\mu+\bar{x}}{1+\gamma}$, and $Var(\theta|\bar{x}) = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2} = \frac{\sigma^2}{n(1+\gamma)}$, where we substituted $\gamma = \sigma^2/(n\tau^2)$. It follows that under the posterior distribution, $\frac{\theta-\delta^B(\bar{x})}{\sqrt{Var(\theta|\bar{x})}} \sim \mathcal{N}(0,1)$, a $1-\alpha$ credible set for θ is given by

$$\left\{\theta : \delta^B(\bar{x}) - z_{\alpha/2}\sqrt{Var(\theta|\bar{x})} \leq \theta \leq \delta^B(\bar{x}) + z_{\alpha/2}\sqrt{Var(\theta|\bar{x})}\right\}.$$

To calculate the coverage probability of the Bayesian region above, we note that under classical model $\bar{X}$ is the random variable and θ is fixed with $\bar{X} \sim \mathcal{N}(\theta, \sigma^2/n)$. This gives the coverage probability as

$$\begin{aligned}
&P_\theta\left[|\theta - \delta^B(\bar{X})| \leq z_{\alpha/2}\sqrt{Var(\theta|\bar{X})}\right] \\
&= P_\theta\left[\left|\theta - \left(\frac{\gamma}{1+\gamma}\mu + \frac{1}{1+\gamma}\bar{X}\right)\right| \leq z_{\alpha/2}\sqrt{\frac{\sigma^2}{n(1+\gamma)}}\right] \\
&= P_\theta\left[-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}}\right],
\end{aligned}$$

where we use the fact $\sqrt{n}(\bar{X}-\theta)/\sigma = Z \sim \mathcal{N}(0,1)$. For $\theta \neq \mu$, $\gamma = 1$ and $n \to \infty$ it is easy to see that the above probability goes to 0.

On the other hand, the usual $1-\alpha$ confidence set for θ is $\{\theta : |\theta - \bar{x}| \leq z_{\alpha/2}\sigma/\sqrt{n}\}$ The credible probability for this set (now $\theta \sim \pi(\theta|\bar{x})$) is given by

$$\begin{aligned}
P_{\bar{x}}\left[\right] &= P_{\bar{x}}\left[\left|[\theta - \delta^B(\bar{x})] + [\delta^B(\bar{x}) - \bar{x}]\right| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\
&= P_{\bar{x}}\left[-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}/\sqrt{n}} \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\sigma/\sqrt{n}},\right]
\end{aligned}$$

where we use the fact $\frac{\theta-\delta^B(\bar{x})}{\sqrt{Var(\theta|\bar{x})}} \sim \mathcal{N}(0,1)$. Yet again, if $\bar{x} \neq \mu$, $\gamma = 1$, and $n \to \infty$ we see the credible probability going to 0. □

9.2 Methods of Evaluating Interval Estimators

There can be different confidence sets for the same problem, motivating us to choose a best one. The size and coverage probability vie against each other - we want our set to have small size and large coverage probability. The coverage probability is a function of the parameter, in general, and hence usually measured by confidence coefficient, the infimum of the coverage probabilities. By size we mean the length of the interval of the volume of the of a multidimensional set.

9.2.1 Size and Coverage Probability

For a given specified converge probability we want to find the confidence interval with the shortest length. The strategy of splitting α equally is not always optimal.

Theorem 9.4. *Let $f(x)$ be a unimodal pdf. If the interval $[a, b]$ satisfies*

- $\int_a^b f(x)dx = 1 - \alpha$,
- $f(a) = f(b) > 0$, *and*
- $a \le x^* \le b$, *where x^* is a mode of $f(x)$,*

then $[a, b]$ is the shortest among all intervals that satisfies the first item.

The same construction yields an optimal Bayesian region. A symmetric unimodal pdf will be optimal at equal α split.

Example 9.12. Suppose $X \sim \mathit{gamma}(k, \beta)$. The quantity $Y = X/\beta$ is a pivot, with $Y \sim \mathit{gamma}(k, 1)$, so we can get a confidence pivot interval by finding constants a and b to satisfy $P[a \le Y \le b] = 1 - \alpha$. Choosing a, and b satisfying $f_Y(a) = f_Y(b)$ will not give an optimal solution for β though. This is because, the interval on β is of the form $\{\beta : x/b \le \beta \le x/a\}$, so the length of the interval is $(1/a - 1/b)x$. We can write b as a function of a as $b(a)$. To solve this we need to solve the optimization problem

$$\begin{aligned} &\underset{a}{\text{minimize}} \quad \frac{1}{a} - \frac{1}{b(a)} \\ &\text{subject to} \quad \int_a^{b(a)} f_Y(y)dy = 1 - \alpha. \end{aligned}$$

Differentiating the first equation gives $db/da = b^2/a^2$ and for the constraint gives $f(b)b^2 = f(a)a^2$. Equations like these also arise in interval estimation of the variance of a normal distribution. □

9.2.2 Test-Related Optimality

Since there is one to one correspondence between confidence sets and tests of hypotheses, there is some correspondence between optimality of tests and optimality of confidence sets. They generally relative to the probability of the set covering false values, instead of size of the set. The probability of false coverage, indirectly measures the size of a confidence set. For $\boldsymbol{X} \sim f(\boldsymbol{x}|\theta)$, we construct a $1-\alpha$ confidence set for θ, $C(\boldsymbol{x})$, by inverting the acceptance region $A(\theta)$. The probability of coverage of $X(\boldsymbol{x})$ is the probability of true coverage, is a function of θ given by $P_\theta[\theta \in C(\boldsymbol{X})]$. The probability of false coverage is a function of θ (true parameter) and θ', defined as the probability of covering θ'

$$\begin{aligned} &P_\theta[\theta' \in C(\boldsymbol{X})], \theta' \ne \theta, \text{ if } C(\boldsymbol{X}) = [L(\boldsymbol{X}), U(\boldsymbol{X})], \\ &P_\theta[\theta' \in C(\boldsymbol{X})], \theta' < \theta, \text{ if } C(\boldsymbol{X}) = [L(\boldsymbol{X}), \infty), \\ &P_\theta[\theta' \in C(\boldsymbol{X})], \theta' > \theta, \text{ if } C(\boldsymbol{X}) = (-\infty, U(\boldsymbol{X})]. \end{aligned}$$

A $1 - \alpha$ confidence set that minimizes the probability of false coverage over a class of $1 - \alpha$ confidence sets is called a uniformly most accurate (UMA) confidence set.

Theorem 9.5. *Let $\boldsymbol{X} \sim f(\boldsymbol{x}|\theta)$, where θ is a real-valued parameter. For each $\theta_0 \in \Theta$, let $A^*(\theta_0)$ be the UMP level α acceptance region of a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Let $C^*(\boldsymbol{x})$ be the $1 - \alpha$ confidence set formed by inverting the UMP acceptance regions. Then for any other $1 - \alpha$ confidence set C, $P_\theta[\theta' \in C^*(\boldsymbol{x})] \leq P_\theta[\theta' \in C(\boldsymbol{X})]$ for all $\theta' < theta$.*

For a normal with known variance, the one sided $1 - \alpha$ confidence bound $\{\mu : \mu \geq \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\}$ is UMA as it is obtained by inverting the UMP test which is $H_0 : \mu = \mu_0$ verses $H_1 : \mu > \mu_0$. The more common two sided interval $\{\mu : \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\}$ is not UMA, since it is obtained by inverting a two sided acceptance region from the test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, for which no UMP test exists. Here, it is reasonable to restrict consideration to unbiased confidence sets. Unbiased confidence sets can be obtained by inverting unbiased sets.

Definition 9.4. *A $1 - \alpha$ confidence set $C(\boldsymbol{x})$ is unbiased if $P_\theta[\theta' \in C(\boldsymbol{X})] \leq 1 - \alpha$ for all $\theta' \neq \theta$.*

The two sided normal interval seen above is an unbiased interval. It can be obtained by inverting the unbiased test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Similarly, the interval in the case of unknown variance based on t distribution is also an unbiased interval, since it can be obtained by inverting a unbiased test.

Sets that minimize the probability of false coverage are also called Neyman-shortest. The expected length of $C(x)$ is equal to the sum of the probabilities of false coverage.

Theorem 9.6. *Let X be a real valued random variable with $X \sim f(x|\theta)$, where θ is a real valued parameter. Let $C(x) = [L(x), U(x)]$ be a confidence interval for θ. If $L(x)$ and $U(x)$ are both increasing functions of x, then for any value θ^*, $E_\theta[length(C(\boldsymbol{X}))] = \int_{\theta \neq \theta^*} P_{\theta^*}[\theta \in C(\boldsymbol{X})] d\theta$.*

The theorem shows that there is a formal relationship between the length of a confidence interval and its probability of false coverage. In the two-sided case, this implies that minimizing the probability of false coverage carries along some guarantee of length optimally. In the one-sided case, however, the analogy does not quite work. In that case, intervals that are set up to minimize the probability of false coverage are concerned with parameters in only a portion of the parameter space and length optimality may no obtain.

9.2.3 Bayesian Optimality

For a posterior distribution $\pi(\theta|\boldsymbol{x})$, we would like a set $C(\boldsymbol{x})$ that satisfies $\int_{C(\boldsymbol{x})} \pi(\theta|\boldsymbol{x}) d\boldsymbol{x} = 1 - \alpha$ and Size($C(\boldsymbol{x})$)$\leq$Size($C'(\boldsymbol{x})$) for any set $C'(\boldsymbol{x})$ satisfying $\int_{C'(\boldsymbol{x})} \pi(\theta|\boldsymbol{x}) d\boldsymbol{x} \geq 1 - \alpha$.

Theorem 9.7. *If the posterior density $\pi(\theta|\boldsymbol{x})$ is unimodal, then for a given value of α, the shortest credible interval for θ is given by $\{\theta : \pi(\theta|\boldsymbol{x}) \geq k\}$ where $\int_{\{\theta:\pi(\theta|\boldsymbol{x}) \geq k\}} \pi(\theta|\boldsymbol{x}) d\theta = 1 - \alpha$. This is called the highest posterior density, HPD, region.*

Example 9.13. The $1 - \alpha$ credible set for a Poisson parameter. The HPD for this region is given by $\{\lambda : \pi(\lambda| \sum x)\} \geq k$, where k is chosen so that $1 - \alpha = \int_{\{\lambda:\pi(\lambda|\sum x) \geq \lambda\}} \pi(\lambda|\sum x) d\lambda$.

The posterior pdf of λ is $gamma(a + \sum x, (n + \frac{1}{b})^{-1})$, we we need to find λ_L and $_U$ such that $\pi(\lambda_L | \sum x) = \pi(\lambda_U | \sum x)$ and $\int_{\lambda_L}^{\lambda_U} \pi(\lambda | \sum x) d\lambda = 1 - \alpha$. For $a = b = 1$, the posterior for λ given $\sum x$ can be expressed as $2(n+1)\lambda \sim \mathcal{X}^2_{2(\sum x+1)}$ and, if $n = 10$ and $\sum x = 6$, the 90% HPD credible set for λ is given by $[0.253, 1.005]$. □

9.2.4 Loss Function Optimality

It is possible to put the requirement of minimum coverage probability and the shortest interval in the loss function and use decision theory to search for an optimal estimator. A decision rule $\delta(\boldsymbol{x})$ simply specifies, for each $\boldsymbol{x} \in \mathcal{X}$, which set $C \in \mathcal{A}$ will be used as an estimate of θ if $\boldsymbol{X} = \boldsymbol{x}$ is observed. The loss function in an interval estimation problem usually includes two quantities: a measure of whether the set estimate correctly includes true value θ and a measure of the size of the set estimate. The natural measure of size is $Length(C)$. To express the correctness measure we use the indicator function $I_C(\theta)$ which is 1 if $\theta \in C$. The loss function can be written as $L(\theta, C) = bLength(C) - I_C(\theta)$, where b is a positive constant that reflects the relative weight that we want to give to the two criteria. The risk function associated is given by $R(\theta, C) = bE_\theta[Length(C(\boldsymbol{X}))] - P_\theta[\theta \in C(\boldsymbol{X})]$.

For $b = 0$, only coverage probability matters and hence $C = (-\infty, \infty)$ is the best decision. For $b = \infty$ point sets become optimal.

Example 9.14. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and assume σ^2 is known. For $c \geq 0$, define an interval estimator for μ by $C(x) = [x - c\sigma, x + c\sigma]$. Let the loss function be $L(\mu, C) = bLength(C) - I_C(\mu)$. The first term in the risk expression is $b2c\sigma$. The second term is $P_\mu[\mu \in C(X)] = P_\mu[X - c\sigma \leq \mu \leq X + c\sigma] = P_\mu[-c \leq \frac{X-\mu}{\sigma} \leq c] = 2P[X \leq c] - 1$. Thus the risk function is $R(\mu, C) = 2bc\sigma + 1 - 2P[Z \leq c]$. Since this is independent of μ, we get the best interval by minimizing this expression for c.

if $b\sigma > 1/\sqrt{2\pi}$, then the risk is minimized at $c = 0$. This corresponds to point estimator $C(x) = [x, x]$. But if $b\sigma \leq 1/\sqrt{2\pi}$, then the risk is minimized at $c = \sqrt{-2\log(b\sigma\sqrt{2\pi})}$. This is same as the usual $1 - \alpha$ confidence interval. □

10 Asymptotic Evaluations

When we let the sample size become infinite, calculations simplify.

10.1 Point Estimation

10.1.1 Consistency

Consistency requires that the estimator converges to the correct value as the sample size becomes infinite.

Definition 10.1. *A sequence of estimators $W_n = W_n(X_1, \dots, X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$, $\lim\limits_{n\to\infty} P_\theta[|W_n - \theta| < \epsilon] = 1$.*

While this looks similar to the definition of convergence in probability, which deals with one particular probability structure, this definition deals with an entire family of probability structures, indexed by θ.

For an estimator W_n, Chebychev's Inequality states $P_\theta\left[|W_n - \theta| \geq \epsilon\right] \leq \frac{E_\theta[(W_n-\theta)^2]}{\epsilon^2}$, so if, for every $\theta \in \Theta$, $\lim\limits_{n\to\infty} E_\theta[(W_n - \theta)^2] = 0$, then the sequence of estimators is consistent. We know that $E_\theta[(W_n - \theta)^2] = Var_\theta[W_n] + (Bias_\theta[W_n])^2$.

Theorem 10.1. *If W_n is a sequence of estimators of a parameter θ satisfying*

1. $\lim\limits_{n\to\infty} Var_\theta[W_n] = 0$,
2. $\lim\limits_{n\to\infty} Bias_\theta[W_n] = 0$,

for every $\theta \in \Theta$, then W_n is a consistent sequence of estimators of θ.

Theorem 10.2. *Consistency of MLEs: Let $X_1, X_2, \dots,$ be iid $f(x|\theta)$, and let $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\theta)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions on $f(x|\theta)$ and, hence, $L(\theta|\boldsymbol{x})$, for every $\epsilon > 0$ and every $\theta \in \Theta$, $\lim\limits_{n\to\infty} P_\theta[|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon] = 0$. That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.*

To prove the above we first show that $\frac{1}{n}\log L(\hat{\theta}|\boldsymbol{x})$ converges almost surely to $E_\theta[\log f(X|\theta)]$ for every $\theta \in \Theta$. Under some conditions on $f(x|\theta)$, this implies that $\hat{\theta}$ converges to θ in probability and, hence $\tau(\hat{\theta})$ converges to $\tau(\theta)$ in probability.

10.1.2 Efficiency

Efficiency concerns with the asymptotic variance of an estimator. Intuitively, given an estimator T_n based on sample of size n, we can calculate the finite-sample variance $Var[T_n]$, and then evaluate $\lim\limits_{n\to\infty} k_n Var[T_n]$, where k_n is some normalizing constant. In many cases $Var[T_n] \to 0$ as $n \to \infty$, so we need a factor k_n to force it to a limit.

Definition 10.2. *For an estimator T_n, if $\lim_{n\to\infty} k_n Var[T_n] = \tau^2 < \infty$, where $\{k_n\}$ is a sequence of constants, then τ^2 is called the limiting variance.*

Example 10.1. For the mean $\bar{X}_n$ of n iid normal observations with $E[X] = \mu$ and $Var[X] = \sigma^2$, if we take $T_n = \bar{X}_n$, then $\lim_{n\to\infty} \sqrt{n} Var[\bar{X}_n] = \sigma^2$ is the limiting variance of T_n. To estimate $1/\mu$ we take $T_n = 1/\bar{X}_n$, and find that the variance is $Var[T_n] = \infty$ as the *limiting variance.* But using delta method we get $E[\frac{1}{\bar{X}_n}] \approx \frac{1}{\mu}$, and $Var[\frac{1}{\bar{X}_n}] \approx \frac{1}{\mu^4} Var \bar{X}_n$. Hence, $Var[T_n] \approx \frac{\sigma^2}{n\mu^4} < \infty$, as the *asymptotic variance.* □

Definition 10.3. *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \to \mathcal{N}(0, \sigma^2)$ in distribution. The parameter σ^2 is called the asymptotic variance of T_n.*

Limiting variance sometimes fails us as in the above example so we use asymptotic variance. Asymptotic variance is always smaller than limiting variance.

Example 10.2. For $Y_n | W_n = w_n \sim \mathcal{N}(0, w_n + (1 - w_n)\sigma_n^2)$, where $W_n \sim Bernoulli(p_n)$ is a mixture model. We have, $Var[X] = E[Var[X|Y]] + Var[E[X|Y]]$, giving $Var[Y_n] = p_n + (1 - p_n)\sigma_n^2$. The limiting variance of Y_n is finite only if $\lim_{n\to\infty}(1 - p_n)\sigma_n^2 < \infty$. The asymptotic distribution of Y_n can be directly calculated using $P[Y_n < a] = p_n P[Z < a] + (1 - p_n)P[Z < a/\sigma_n]$. If we let $p_n \to 1$ and $\sigma_n \to \infty$ in such a way that $(1 - p_n)\sigma_n^2 \to \infty$. It then follows that $P[Y_n < a] \to P[Z < a]$, that is $Y_n \to \mathcal{N}(0, 1)$, and we have the limiting variance as ∞ and asymptotic variance as 1. □

Definition 10.4. *A sequence of estimators W_n is asymptotically efficient for a parameter $\tau(\theta)$ if $\sqrt{n}[W_n - \tau\theta] \to \mathcal{N}(0, v(\theta))$ in distribution and*

$$v(\theta) = \frac{[\tau'(\theta)]^2}{E_\theta[(\frac{\partial}{\partial\theta} \log f(X|\theta))^2]},$$

that is, the asymptotic variances of W_n achieves the Cramer-Rao Lower Bound.

Theorem 10.3. *Asymptotic efficient of MLEs: Let $X_1, X_2, \ldots,$ be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions on $f(x|\theta)$ and, hence, $L(\theta|\boldsymbol{x})$, $\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to \mathcal{N}(0, v(\theta))$, where $v(\theta)$ is the Cramer-Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.*

Efficiency is defined only when the estimator is asymptotically normal and, asymptotic normality implies consistency.

10.1.3 Calculations and Comparisons

One of the regularity condition which must be valid, it must be the case that the support of the distribution, hence likelihood function, must be independent of the parameter.

If an MLE is asymptotically efficient, it is the same as the Delta method variance. Thus, we can use Cramer-Rao Lower bound as an approximation to the true variance of the MLE. Suppose $X_1, \ldots, X_n$ are iid $f(x|\theta)$, $\hat{\theta}$ is the MLE of θ, and $I_n(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{X})\right)^2\right]$ is the information number of the sample, with $Var[\hat{\theta}] = \frac{1}{I_n(\theta)}$. From the Delta Method and asymptotic efficiency of MLEs the variance of $h(\hat{\theta})$ can be approximated by $Var[h(\hat{\theta})|\theta] \approx \frac{h'(\theta)^2}{I_n(\theta)}$. Now, using the fact $E_\theta\left[\left(\frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right]$ and the result (Efron and Hinkley, 78) that observed information number is superior to the expected information number, we get

$$Var[h(\hat{\theta})|\theta] \approx \frac{[h'(\theta)]^2\Big|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\Big|_{\theta=\hat{\theta}}}.$$

This is a consistent estimator of $Var_\theta[h(\hat{\theta})]$ till $h(\theta)$ is a monotonic function, which otherwise could cause underestimation. Further, since this is already based on Cramer-Rao Lower Bound, it is probably already an underestimate. Notice that the variance estimation process is a two-step procedure. To estimate $Var_\theta[h(\hat{\theta})]$, first we approximate $Var_\theta[h(\hat{\theta})]$; then we estimate the resulting approximation by substituting $\hat{\theta}$ for θ.

Example 10.3. For random sample $X_1, \dots, X_n$ from $Bernoulli(p)$ population. The MLE for p is $\hat{p} = \frac{1}{n}\sum X_i$. By direct calculation we know $Var_p[\hat{p}] = \frac{1}{n}p(1-p)$, and direct substitution, hopefully reasonable, gives estimate of $Var_p[\hat{p}]$ is $\hat{Var}_p[\hat{p}] = \frac{1}{n}\hat{p}(1-\hat{p})$. Applying the previous approximation with $h(x) = p$ and noting that $\log L(p|\boldsymbol{x}) = n\hat{p}\log(p) + n(1-\hat{p})\log(1-p)$, we get an estimate of $Var_p[\hat{p}]$ as $\hat{Var}_p[\hat{p}] = \frac{\hat{p}(1-\hat{p})}{n}$, which is the same as guessed before. We can hence, assert asymptotic efficiency of $\hat{p}$ and, in particular, that $\sqrt{n}(\hat{p}-p) \to \mathcal{N}(0, p(1-p))$ in distribution. Applying Slutsky's Theorem we can conclude that $\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} \to \mathcal{N}(0,1)$.

To get the variance of $\hat{p}/(1-\hat{p})$, an estimates of the odds, we use Delta Method and final step substitution

$$\hat{Var}\left[\frac{\hat{p}}{1-\hat{p}}\right] = \frac{\left[\frac{\partial}{\partial}\left(\frac{p}{1-p}\right)\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})\Big|_{p=\hat{p}}} = \frac{\hat{p}}{n(1-\hat{p})^3}.$$

This estimator is asymptotically efficient.

Further, say we want to estimate the variance of the Bernoulli distribution, $p(1-p)$. The MLE of this variance is given by $\hat{p}(1-\hat{p})$, and an estimate of the variance of this estimator can be obtained by the approximation

$$\hat{Var}[\hat{p}(1-\hat{p})] = \frac{\left[\frac{\partial}{\partial p}p(1-p)\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{p^2}\log L(p|\boldsymbol{x})\Big|_{p=\hat{p}}} = \frac{\hat{p}(1-\hat{p})(1-2\hat{p})^2}{n},$$

which can be 0 if $\hat{p} = \frac{1}{2}$, a clear underestimate of the variance of $\hat{p}(1-\hat{p})$. The fact that the function $p(1-p)$ is not monotone is a cause of this problem. Hence, our estimate is asymptotically efficient as long as $p \neq \frac{1}{2}$. If $p = \frac{1}{2}$ we need to use a second-order approximation. □

Definition 10.5. *If two estimators W_n and V_n satisfy $\sqrt{n}[W_n - \tau(\theta)] \to \mathcal{N}(0, \sigma_W^2)$ and $\sqrt{n}[V_n - \tau(\theta)] \to \mathcal{N}(0, \sigma_V^2)$ in distribution, the asymptotic relative efficiency (ARE) of V_n with respect to W_n is $ARE(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}$.*

Example 10.4. Suppose that $X_1, X_2,$ are iid $Poisson(\lambda)$, and we are estimating the 0 probability event. $P[X = 0] = e^{-\lambda}$, and a maive estimator comes from defining $Y_i = \mathbf{1}_{X_i=0}$ and using $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. The Y_is are $Bernoulli(e^{-\lambda}$, and hence follows that $E[\hat{\tau}] = e^{-\lambda}$ and $Var[\hat{\tau}] = \frac{1}{n}e^{-\lambda}(1 - e^{-\lambda})$. Alternatively, the MLE of $e^{-\lambda}$ is $e^{-\hat{\lambda}}$, where$\hat{} = \frac{1}{n}\sum_i X_i$ is the MLE of λ. Using Delta Method approximations we have, $E[e^{-\hat{\lambda}}] \approx e^{-\lambda}$ and $Var[e^{-\hat{\lambda}}] \approx \frac{1}{n}^{-2\lambda}$. Since, $\sqrt{n}(\hat{\tau} - e^{-\lambda}) \to \mathcal{N}(0, e^{-\lambda}(1 - e^{-\lambda}))$ and $\sqrt{n}(e^{-\hat{\lambda}} - e^{-\lambda}) \to \mathcal{N}(0,^{-2\lambda})$ in distribution, the ARE of $\hat{\tau}$ with respect to the MLE $e^{-\hat{\lambda}}$ is

$$ARE(\hat{\tau}, e^{-\hat{\lambda}}) = \frac{\lambda}{e^{\lambda} - 1}.$$

It is a strictly decreasing function with maximum value of 1 at $\lambda = 0$ and tailing off rapidly to asymptote to 0 as $\lambda \to \infty$. □

Sine the MLE is typically asymptotically efficient, another estimator can't hope to beat its asymptotic variance. However, other estimators may have other desirable properties like ease of calculation, robustness to underlying assumptions. In such cases efficiency of MLE becomes a calibration for the alternative estimator.

10.1.4 Bootstrap Standard Errors

Bootstrap provides and alternative means of calculating standard errors. The bootstrap helps us learn about the sample characteristics, and hence population characteristics, by taking resamples. Like the Delta Method, the variance formula from bootstrap is applicable to virtually any estimator. For n^n resamples we take B samples. For any estimator $\hat{\theta}(x) = \hat{\theta}$ we have

$$Var_B^*[\hat{\theta}] = \frac{1}{B-1}\sum i = 1^B(\hat{\theta}_i^* - \bar{\hat{\theta}}^0)^2,$$

where $\hat{\theta}_i^*$ is the estimator calculated from the $i$the resample and $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$, the mean of the B resampled values. Delta method estimates are based on lower bound and hence underestimates, while bootstrap based variance are closer to true values and automatically takes the second order effect into account. Here, we assumed no functional form for the population pdf or cdf and hence it is called *nonparametric bootstrap.*

Suppose we have a sample $X_1, \ldots, X_n$ from a distribution with pdf $f(x|\theta)$, where θ may be a vector of parameters. We can estimate θ with $\hat{\theta}$, the MLE, and draw samples $X_1^*, \ldots, X_n^* \sim f(x|\hat{\theta})$. If we take B such samples, we can estimate the variance of $\hat{\theta}$. This is called *parametric bootstrap.* Not that these samples are not resamples of data, but actual random samples drawn from $f(x|\hat{\theta})$, which is sometimes called the *plug-in distribution.*

We have an all-purpose method for computing standard errors but we need to know if it is a good method. Delta Method based on MLE will typically produce consistent estimators. In many cases, the bootstrap does provide us with a reasonable estimator that is consistent. Using Law of Large Numbers one can establish that

$$Var^*_B[\hat{\theta}] \overset{B\to\infty}{\to} Var^*[\hat{\theta}].$$

Typically consistency will be obtained in iid sampling, but in more general situations it may not occur, i.e. for iid samples

$$Var^*[\hat{\theta}] \overset{n\to\infty}{\to} Var[\hat{\theta}].$$

10.2 Robustness

If the assumed underlying model is not correct, then we can't be guaranteed of the optimality of our estimators. Even after careful consideration of the model, we might be concerned about small deviations from our assumed model. This leads to the consideration of robust estimators, which comes at the expense of giving up some optimality for the assumed model in exchange for reasonable performance if the assumed model is not the true model. Any statistical procedure should posses the following desirable features:

- It should have a reasonably good, optimal or near optimal, efficiency at the assumed model.
- It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly.
- Somewhat larger deviations from the model should not cause a catastrophe.

10.2.1 The Mean and the Median

Example 10.5. Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$. The variance of $\bar{X}$ is σ^2/n, which is the Cramer-Rao Lower Bound. Hence, $\bar{X}$ is near optimal in terms of attaining best variance at the assumed model. For small deviations from the model, a common interpretation is to use an δ-contamination model; that is, for small δ, assume that we observe

$$X_i \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & \text{with probability } 1-\delta \\ f(x) & \text{with probability } \delta \end{cases},$$

where $f(x)$ is some other distribution. Suppose we take $f(x)$ to be any density with mean θ and variance τ^2. Then $Var[\bar{X}] = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \delta(1-\delta)\frac{(\theta-\mu)^2}{n}$. If $\theta \approx \mu$ and $\sigma \approx \tau$, then $\bar{X}$ will be near optima.If we perturb the model a little more, things can get quite bad. Consider $f(x)$ is a Cauchy distribution, then $Var[\bar{X}] = \infty$.

Definition 10.6. *Let $X_{(1)} < \ldots < X_{(n)}$ be an ordered sample of size n, and let T_n be a static based on this sample. T_n has breakdown value b, $0 \le b \le 1$, if, for every $\epsilon > 0$,*

$$\lim_{X_{(\{(1-b)n\})}\to\infty} T_n < \infty \textit{ and } \lim_{(\{(1-(b+\epsilon))n\})\to\infty} T_n = \infty.$$

The breakdown value of $\bar{X}$ is 0; that is, if any fraction of the sample is driven to infinity, so is the value of $\bar{X}$. On the other hand sample median has a breakdown value of 50%. To find how better a particular estimator is we make use of asymptotic relative efficiency. To compute the ARE of the median with respect to mean, we establish the asymptotic normality of the median.

Example 10.6. Let $X_1, \dots, X_n$ be a sample from a population with pdf f and cdf F (assumed to be differentiable), with $P[X_i \leq \mu] = \frac{1}{2}$, so μ is the population median. Let M_n be the sample median. To calculate $\lim_{n\to\infty} P[\sqrt{n}(M_n - \mu) \leq a]$, for some a we define the random variables Y_i by $Y_i = \begin{cases} 1 & \text{if } X_i \leq \mu + a/\sqrt{n} \\ 0 & \text{otherwise} \end{cases}$, it follows that Y_i is a Bernoulli random variable with success probability $p_n = F(\mu + a/\sqrt{n})$. To avoid complications, we will assume that n is odd and thus the event $\{M_n \leq \mu + a/\sqrt{n}\}$ is equivalent to the event $\{\sum_i Y_i \geq (n+1)/2\}$. Now, after some algebra we get

$$P[\sqrt{n}(M_n - \mu) \leq a] = P\left[\frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right].$$

Now $p_n \to p = F(\mu) = \frac{1}{2}$. A limit calculation shows that

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} \to -2aF'(\mu) = -2af(\mu).$$

This shows that $P[\sqrt{n}(M_n - \mu) \leq a][Z \geq -2af(\mu)]$, where $Z = \frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}}$ and thus $\sqrt{n}(M_n - \mu)$ is asymptotically normal with mean 0 and variance $1/[2f(\mu)]^2$.

As there are simple expressions for the asymptotic variances of mean and median, the ARE is easily computed. The ARE for normal, logistic and double exponential distributions are 0.64, 0.82 and 2, respectively. As the tails get heavier, the ARE gets bigger, i.e. the performance of the median improves. □

10.2.2 M-Estimators

Huber considered a compromise between the mean and the median. We minimize a criterion function $\sum_{i=1}^n \rho(x_i - a)$, where ρ is given by $\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| \geq k \end{cases}$, a continuous, differentiable function. For smaller values of tuning parameter k we get more 'median-like' estimator. For a general function ρ, we call the estimator minimizing $\sum_i \rho(x_i - \theta)$ and M-estimator. If ρ is chosen to be negative log likelihood, $-\ell(\theta|x)$, then the M-estimator is the usual MLE. We define $\psi = \rho'$ and hence M-estimator is a solution to $\sum_{i=1} *n\psi(x_i - \theta) = 0$. For symmetric $\rho(x)$, and its derivative $\phi(x)$ monotone increasing we write the Taylor expansion of ψ around the value θ_0 as

$$\sum_{i=1}^n \psi(x_i - \theta) = \sum_{i=1}^n \psi(x_i - \theta_0) + (\theta - \theta_0)\sum_{i=1}^n \psi'(x_i - \theta_0) + \dots$$

For $\hat{\theta}_M$ as the solution, i.e. $\sum_{i=1}^n \psi(x_i - \hat{\theta}_M) = 0$ and substituting this for θ and ignoring higher order terms, we get

$$0 = \sum_{i=1}^n \psi(x_i - \theta_0) + (\hat{\theta}_M - \theta_0) \sum_{i=1}^n \psi'(x_i - \theta_0).$$

Rearranging this gives

$$\sqrt{n}(\hat{\theta}_M - \theta_0) = \frac{-\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i - \theta_0)}{\frac{1}{n} \sum_{i=1}^n \psi'(x_i - \theta_0)}.$$

Further, let's define θ_0 by $E_{\theta_0}[\psi(X - \theta_0)] = 0$, it then follows

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i - \theta_0) = \sqrt{n}\left[-\frac{1}{n} \sum_{i=1}^n \psi(X_i - \theta_0)\right] \to \mathcal{N}(0, E_{\theta_0}[\psi(X - \theta_0)^2])$$

in distribution, and the Law of Large Numbers yields

$$\frac{1}{n} \sum_{i=1}^n \psi'(x_i - \theta_0) \to E_{\theta_0}[\psi'(X - \theta_0)]$$

in probability. This gives

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \to \mathcal{N}\left(0, \frac{E_{\theta_0}[\psi(X - \theta_0)^2]}{[E_{\theta_0}[\psi'(X - \theta_0)]]^2}\right).$$

Huber estimator is asymptotically normal as well. Huber estimator behaves similar to the mean for the normal and logistic distributions and is an improvement on the median. For the double exponential it is an improvement over mean but not as good as median.

We sett that an M-estimator is a compromise between robustness and efficiency. The denominator of the asymptotic variance contains the term $E_{\theta_0}[\psi'(X - \theta_0)]$, which can be shown to be equal to $E_{\theta_0}[\psi(X - \theta)\ell'(\theta|X)]$. To compare the asymptotic variance of an M-estimator to that of the MLE we write

$$ARE(\hat{\theta}_M, \hat{\theta}) = \frac{[E_\theta[\psi(X - \theta_0)\ell'(\theta|X)]]^2}{E_\theta[\psi(X - \theta)^2] E_\theta[\ell'(\theta|X)^2]} \leq 1$$

by virtue of the Cauchy-Swartz Inequality. Thus an M-estimator is always less efficient than MLE, and matches in efficiency only if ϕ is proportional to ℓ'. This loss of efficiency is made up by the increase in robustness.

10.3 Hypothesis Testing

10.3.1 Asymptotic Distributions of LRTs

The test statistic in the likelihood ratio method is $\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})}$, and the rejection region region is $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$. The likelihood is completely defined after the observation $\boldsymbol{X} = \boldsymbol{x}$ has been made, though the supremum may not have a closed form. To define a level α test, the constant c must be chosen such that $\sup_{\theta \in \Theta_0} P_\theta[\lambda(\boldsymbol{X}) \leq c] \leq \alpha$. If asymptotic are appealed to we can get an approximate answer.

Theorem 10.4. *For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, suppose $X_1, \dots, X_n$ are iid $f(x|\theta)$, $\hat{\theta}$ is the MLE of θ, and $f(x|\theta)$ satisfies the regularity conditions, then under H_0, as $n \to \infty$, $-2 \log \lambda(\boldsymbol{X}) \to \mathcal{X}_1^2$ in distribution.*

Example 10.7. For testing $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ based on observing $X_1, \dots, X_n$ iid $Poisson(\lambda)$, we have

$$-2 \log \lambda(\boldsymbol{x}) = -2 \log \left(\frac{e^{-n\lambda_0} \lambda_0^{\sum x_i}}{e^{-n\hat{}} \hat{\lambda}^{\sum x_i}} \right) = 2n \left[(\lambda_0 - \hat{}) - \hat{\lambda} \log(\lambda_0 / \hat{\lambda}) \right],$$

where $\hat{\lambda} = \sum x_i / n$ is the MLE of λ. We would reject H_0 at level α if $-2 \log \lambda(\boldsymbol{x}) > \mathcal{X}_{1,\alpha}^2$. Even with $n = 25$ asymptotic matches quite well. □

Theorem 10.5. *Let $X_1, \dots, X_n$ be random samples from a distribution $f(x|\theta)$. Under the regularity conditions if $\theta \in \Theta_0$, then the distribution of the statistic $-2 \log \lambda(\boldsymbol{X})$ converges to a chi squared distribution (degree of freedom is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$) as $n \to \infty$.*

Rejection of $H_0 : \theta \in \Theta_0$ for small values of $\lambda(\boldsymbol{X})$. Thus, H_0 is rejected iff $-2 \log \lambda(\boldsymbol{X}) \geq \mathcal{X}_{\nu,\alpha}^2$, where ν is the degree of freedom as specified above. The Type-1 error probability will be approximately α if $\theta \in \Theta_0$ and the sample size is large. Most often Θ can be represented as a subset of q-dimensional Euclidian space that contains an open subset in $\mathcal{R}^q$, and Θ_0 can be represented as a subset of p-dimensional Euclidian space that contains an open subset in $\mathcal{R}^p$, where $p < q$. Then $\nu = q - p$ is the degree of freedom for the test statistic.

Example 10.8. Let $\theta = (p_1, p_2, p_3, P_4, p_5)$, where the p_js are non-negative and sum to 1. Suppose $X_1, \dots, X_n$ are iid discrete random variables and $P_\theta[X_i = j] = p_j$, $j = 1, \dots, 5$. Thus the distribution of X_i is $f(j|\theta) = p_j$ and the likelihood function is $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5}$, where $y_j =$ number of $x_1, \dots, x_n$ equal to j. For testing $H_0 : p_1 = p_2 = p_3$ and $p_4 = p_5$ versus $H_1 : H_0$ is not true. The full parameter space is a four dimensional set, $q = 4$. There is only free parameter in the set specified y H_0, $p = 1$. The degree of freedom $\nu = 3$. The MLE is determined by $\frac{\partial}{\partial p_j} \log L(\theta|\boldsymbol{x}) = 0$, for each of $j = 1, \dots, 4$. The MLE of p_j under Θ is $\hat{p}_j = \frac{y_j}{n}$. Under H_0 the likelihood function is $L(\theta|\boldsymbol{x}) = p_1^{y_1+y_2+y_3} \left(\frac{1-3p_1}{2}\right)^{y_4+y_5}$. This gives the MLE for Θ_0 as $\hat{p}_{10} = \hat{p}_{20} = \hat{p}_{30} = (y_1 + y_2 + y_3)/(3n)$ and $\hat{p}_{40} = \hat{p}_{50} = (1 - 3\hat{p}_{10})/2$. Substituting these we get

$$\lambda(\boldsymbol{x}) = \left(\frac{y_1 + y_2 + y_3}{3y_1} \right)^{y_1} \left(\frac{y_1 + y_2 + y_3}{3y_2} \right)^{y_2} \left(\frac{y_1 + y_2 + y_3}{3y_3} \right)^{y_3} \left(\frac{y_4 + y_5}{2y_4} \right)^{y_4} \left(\frac{y_4 + y_5}{2y_5} \right)^{y_5}.$$

Thus the test statistic is $-2 \log \lambda(\boldsymbol{x}) = 2 \sum_{i=1}^5 y_i \log \left(\frac{y_i}{m_i} \right)$, where $m_1 = m_2 = m_3 = (y_1 + y_2 + y_3)/3$ and $m_4 = m_5 = (y_4 + y_5)/2$. The asymptotic size α test rejects H_0 if $-2 \log \lambda(\boldsymbol{x}) \geq \mathcal{X}_{3,\alpha}^2$. □

10.3.2 Other Large-Sample Tests

Suppose we wish to test a hypothesis about a real-valued parameter θ, and $W_n = W(\boldsymbol{X})$ is a point estimator of θ, based on sample size n. For example, W_n might be the MLE of θ. If σ_n^2 denotes the variance of W_n and if we can use some form of CLT to show that, as $n \to \infty$, $(W_n - \theta)/\sigma_n$ converges in distribution to a standard normal random variable, then $(W_n - \theta)/\sigma_n$ can be compared to a $\mathcal{N}(0, 1)$ distribution. In some instances σ_n may depends on unknown parameters. In such a case, we look for an estimate S_n of σ_n with the property that σ_n/S_n converges in probability to 1. Then, using Slutsky's theorem we can deduce that $(W_n - \theta_0)/S_n$ also converges in distribution to a standard normal distribution.

Suppose we wish to test the two-sided hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Based on the statistic $Z_n = (W_n - \theta_0)/S_n$ we reject H_0 if and only if $Z_n < -z_{\alpha/2}$ or $Z_n > z_{\alpha/2}$. If H_0 is true, then $\theta = \theta_0$ and Z_n converges in distribution to $Z \sim \mathcal{N}(0, 1)$. Thus, the type-1 error probability, $P_{\theta_0}[Z_n < -z_{\alpha/2}$ or $Z_n > z_{\alpha/2}] \to P[Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}] = \alpha$, and this is an asymptotically size α test.

Now consider an alternative parameter value $\theta \neq \theta_0$. We can write $Z_n = \frac{W_n - \theta}{S_n} = \frac{W_n - \theta_0}{S_n} + \frac{\theta - \theta_0}{S_n}$. No matter what the value of θ, the term $(W_n - \theta)/S_n \to \mathcal{N}(0, 1)$. Typically $\sigma_n \to 0$ are $n \to \infty$. Thus, S_n will converge in probability to 0 and the term $(\theta - \theta_0)/S_n$ will converge to $+\infty$ or $-\infty$ in probability, depending on sign of $(\theta - \theta_0)$. Thus, Z_n will converge to $+\infty$ or $-\infty$ in probability and $P_\theta[\text{reject } H_0] = P_\theta[Z_n < -z_{\alpha/2}$ or $Z_n > z_{\alpha/2}] \to 1$ as $n \to \infty$. Hence, a test with asymptotic size α and asymptotic power 1 can be constructed.

For a one sided hypothesis $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, a similar test might be constructed. Again, the test statistic $Z_n = (W_n - \theta_0)/S_n$ would be used and the test would reject H_0 if and only if $Z_n > z_\alpha$. Similar to above, we could conclude that that that power function of this test converges to 0, α, or 1 according as $\theta < \theta_0$, $\theta = \theta_0$, or $\theta > \theta_0$. Thus this test too has a reasonable asymptotic power properties.

A **Wald test** is a test based on a statistic of the form $Z_n = \frac{W_n - \theta_0}{S_n}$, where θ_0 is the hypothesized value of the parameter θ, W_n is an estimator of θ, and S_n is the standard error for W_n, an estimate of the standard deviation of W_n. If W_n is the MLE of θ, then $1/\sqrt{I_n(W_n)}$ is a reasonable standard error for W_n. Alternatively, $1/\sqrt{\hat{I}_n(W_n)}$, where

$$\hat{I}_n(W_n) = -\frac{\partial^2}{\partial\theta^2} \log L(\theta|\boldsymbol{X})\Big|_{\theta = W_n}$$

is the observed information number, is often used.

Example 10.9. Let $X_1, \ldots, X_n$ be a random sample from *Bernoulli*(p) population. Consider testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, where $0 < p_0 < 1$ is a specified value. The MLE of p is $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. CLT states that for any p, $p < p < 1$, $\frac{\hat{p}_n - p}{\sigma_n}$ converges to a standard normal variable, with $\sigma_n = \sqrt{p(1-p)/n}$. A reasonable estimate of $_n$ is $S_n = \sqrt{\hat{p}_n(1 - \hat{p}_n)/n}$, which is same as $1/\sqrt{I_n(\hat{p}_n)}$, and σ_n/S_n converges in probability to 1.

Thus, $\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \to \mathcal{N}(0,1)$. The Wald test statistic Z_n is defined by replacing p by p_0 and the large-sample Wald test rejects H_0 if $Z_n > z_\alpha$.

For two sided hypothesis $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, where $0 < p_0 < 1$ is a specific value, We can use the above method again. Alternatively, by CLT for any p, $0 < p < 1$, $\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \to \mathcal{N}(0,1)$. Therefore if the null hypothesis is true, the statistic $Z'_n = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{N}(0,1)$, approximately. The approximate level α test rejects H_0 if $|Z'_n| > z_{\alpha/2}$. In cases where both tests are applicable, preference is unclear as the actual power functions cross each other. For approximate tests continuity corrections are used to maintain the level α. □

This motivates the large-sample **score test**. The score statistic is defined to be

$$S(\theta) = \frac{\partial}{\partial\theta} \log f(X|\theta) = \frac{\partial}{\partial\theta} \log L(\theta|X).$$

We have seen that $E_\theta[(\theta)] = 0$. For $H_0 : \theta = \theta_0$ and if H_0 is true, then $S(\theta_0)$ has mean 0. Further, $Var_\theta[S(\theta)] = E_\theta\left[(\frac{\partial}{\partial\theta} \log L(\theta|X))^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta|X)\right] = I_n(\theta)$; the information number is the variance of the score statistic. The test statistic for the score test is $Z_S = S(\theta_0)/\sqrt{I_n(\theta_n)}$. If H_0 is true, Z_s has mean 0 and variance 1 and converges to a standard normal. Thus, the approximate level α score test rejects H_0 if $|Z_s| > z_{\alpha/2}$. If H_0 is composite, then $\hat{\theta}_0$, an estimate of θ assuming H_0 is true, replaces θ_0 in Z_S. For restricted MLE, $\hat{\theta}_0$ can be obtained by Lagrange multipliers. The score test is sometimes called the Lagrange multiplier test.

Finally, robust tests can be considered. If $X_1, \ldots, X_n$ are iid from a location family and $\hat{\theta}_M$ is an M-estimator, then $\sqrt{n}(\hat{\theta}_M - \theta_0) \to \mathcal{N}(0, Var_{\theta_0}[\hat{\theta}_M])$, where $Var_{\theta_0}[\hat{\theta}_M] = \frac{E_{\theta_0}[\psi(X-\theta_0)^2]}{(E_{\theta_0}[\psi'(X-\theta_0)])^2}$ is the asymptotic variance. The generalized score statistic is

$$Z_G = \sqrt{n}\frac{\hat{\theta}_M - \theta_0}{\sqrt{Var_{\theta_0}[\hat{\theta}_M]}},$$

or a generalized Wald statistic,

$$Z_{GW} = \sqrt{n}\frac{\hat{\theta}_M - \theta_0}{\sqrt{\hat{Var}_{\theta_0}[\hat{\theta}_M]}},$$

where $\hat{Var}_{\theta_0}[\hat{\theta}_M]$ can be any consistent estimator. For example, we could use a boot-strap estimate of standard error of simply substitute an estimator in the expression of $Var_{\theta_0}[\hat{\theta}_M]$ and use

$$\hat{Var}[\hat{\theta}_M] = \frac{\frac{1}{n}\sum_{i=1}^n [\psi(x_i - \hat{\theta}_M)]^2}{\left[\frac{1}{n}\sum_{i=1}^n \psi'(x_i - \hat{\theta}_M)\right]^2}.$$

Example 10.10. If $X_1, \ldots, X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, then the Huber M-estimator using the ρ function and the ψ function, we have an asymptotic variance

$$\frac{\int_{-k}^{k} x^2 f(x)dx + k^2 P_0[|X| > k]}{(P_0[|X| \leq k])^2}.$$

Based on the asymptotic normality of M-estimator, we can test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ at level α by rejecting H_0 if $|Z_G S| > z_{\alpha/2}$. Practically, we use the approximate test with statistic Z_{GW}, with variance estimate

$$\hat{Var}_w[\hat{\theta}_M] = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_M)^2 I(|x_i - \hat{\theta}_M| < k) + k^2 \left(\frac{1}{n}\sum_{i=1}^{n} I(|x_i - \hat{\theta}_M| > k)\right)}{\left(1 - \frac{1}{n}\sum_{i=1}^{n} I(|x_i - \hat{\theta}_M| < k)\right)^2}.$$

For a naive test we add Z_N that uses the simple variance estimate

$$\hat{Var}_n[\hat{\theta}_M] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_M)^2.$$

Simulation wise, the $z_{\alpha/2}$ cutoffs are generally too small, neglecting to account for variation in the variance estimates. The actual size is typically greater than the nominal size. However, there is consistency across a range of distributions, with double exponential beign the best case. □

10.4 Interval Estimation

10.4.1 Approximate Maximum Likelihood Intervals

If $X_1, \ldots, X_n$ are iid $f(x|\theta)$ and $\hat{\theta}$ is the MLE of θ, then the variance of a function $h(\hat{\theta})$ can be approximated by

$$\hat{Var}[h(\hat{\theta})|\theta] \approx \frac{(h'(\theta))^2\Big|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{x})\Big|_{\theta=\hat{\theta}}}.$$

Further for a fixed but arbitrary value of θ, we have shown that

$$\frac{h(\hat{theta}) - h(\theta)}{\sqrt{\hat{Var}[h(\hat{\theta})|\theta]}} \to \mathcal{N}(0,1),$$

giving the approximate confidence interval

$$h(\hat{\theta}) - z_{\alpha/2}\sqrt{\hat{Var}[h(\hat{\theta})|\theta]} \leq h(\theta) \leq h(\hat{\theta}) + z_{\alpha/2}\sqrt{\hat{Var}[h(\hat{\theta})|\theta]}.$$

Example 10.11. We have a random sample $X_1, \ldots, X_n$ from a *Bernoulli*(p) population. To estimate the odds ratio $p/(1-p)$ by its MLE $\hat{p}/(1-\hat{p})$ and that this estimate has approximate variance $\hat{p}/n/(1-\hat{p})^3$. We therefore can construct the approximate confidence interval

$$\frac{1}{1-\hat{p}}\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}}{n(1-\hat{p})}}\right) \leq \frac{p}{1-p} \leq \frac{1}{1-\hat{p}}\left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}}{n(1-\hat{p})}}\right).$$

□

A more restrictive form of the likelihood approximation, when applicable, gives better intervals is based on the score statistic. The random quantity

$$Q(X|\theta) = \frac{\frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})}{\sqrt{-E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right]}}$$

has a $\mathcal{N}(0,1)$ distribution asymptotically as $n \to \infty$. Thus, the set $\{\theta : |Q(\boldsymbol{x}|\theta)| \le z_{\alpha/2}\}$ is an approximate $1-\alpha$ confidence set. The expectation of this quantity is 0 and the variance is 1, and so this approximation exactly matches the first two moments of a $\mathcal{N}(0,1)$ random variable. These intervals have an asymptotic optimal shortest interval in certain class.

Example 10.12. For a Binomial sampling, if $Y = \sum_{i=1}^n X_i$, where each X_i is an independent *Bernoulli*(p) random variable, we have

$$\begin{aligned} Q(Y|p) &= \frac{\frac{\partial}{\partial p}\log L(p|Y)}{\sqrt{-E_p\left[\frac{\partial^2}{\partial p^2}\log L(p|Y)\right]}} \\ &= \frac{\frac{y}{p} - \frac{n-y}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} \\ &= \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}, \end{aligned}$$

where $\hat{p} = y/n$. Hence, an approximate $1-\alpha$ confidence interval is given by

$$\Big\{p : \Big|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\Big| \le z_{\alpha/2}\Big\}.$$

This is the interval that results from inverting the score statistic, which needs solving a quadratic in p. □

Another likelihood test, we encountered, is based on the fact that $-2\log\lambda(\boldsymbol{X})$ has an asymptotic chi squared distribution. This suggest that if $X_1, \ldots, X_n$ are iid $f(x|\theta)$ and $\hat{\theta}$ is the MLE of θ, then the set

$$\Big\{\theta : -2\log\left(\frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}\right) \le \mathcal{X}^2_{1,\alpha}\Big\}$$

is an approximate $1-\alpha$ confidence interval. This is indeed the case and gives us another approximate likelihood interval. This corresponds to the highest likelihood region.

Example 10.13. For $Y = \sum_{i=1}^n X_i$, where each X_i is an independent *Bernoulli*(p) random variable, we have the approximate $1-\alpha$ confidence set

$$\Big\{p : -2\log\left(\frac{p^y(1-p)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}}\right) \le \mathcal{X}^2_{1,\alpha}\Big\}$$

□

10.4.2 Other Large-Sample Intervals

Most approximate confidence intervals are based on either finding approximate or asymptotic pivots or inverting approximate level α test statistics. For Wald-type interval, if we have any statistic W and V and a parameter θ such that, as $n \to \infty$, $(W-\theta)/V \to \mathcal{N}(0,1)$, then we can form the approximate confidence interval for θ given by $W - z_{\alpha/2}V \leq \theta \leq W + z_{\alpha/2}V$. Direct application of CLT with Slutsky's Theorem will usually give an approximate confidence interval.

Example 10.14. if $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2, then, from CLT we have $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$. Moreover, from Slutsky's Theorem, if $S^2 \to \sigma^2$ in probability, then $\frac{\bar{X}-\mu}{S/\sqrt{n}} \to \mathcal{N}(0,1)$, giving the approximate $1-\alpha$ confidence interval $\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$. If the interval is pivotal, the coverage property does not depend on the parameter value. There is optimism here as we do not account for the variability in S for finite sample case. As the sample size increases, the approximation will improve. □

By providing the form of the sampling distribution we can improve the approximation.

Example 10.15. If $X_1, \ldots, X_n$ are iid $Poisson(\lambda)$, then we know that $\frac{\bar{X}-\lambda}{S/\sqrt{n}} \to \mathcal{N}(0,1)$. However, this is true even if we did not sample from a Poisson population. For Poisson $Var[X] = \lambda = E[\bar{X}]$ and $\bar{X}$ is a good estimator of λ. Thus, using the Poisson assumption we get the equivalent interval from inverting Wald test using $\frac{\bar{X}-\lambda}{}\sqrt{X/n} \to \mathcal{N}(0,1)$. We can us the Poisson assumption in another way to get the interval corresponding to the score test, which is also the likelihood interval by using $\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \to \mathcal{N}(0,1)$. □

Parameters are fixed and don't produce any variability into the approximation, while each statistic brings more variability along with it. Hence, a reasonable rule of thumb is to use as few estimates and as many parameters as possible in an approximation.

Example 10.16. For a random sample $X_1, \ldots, X_n$ from a $Bernoulli(p)$ population, we saw that both $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ and $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ converge in distribution to a standard normal, where $\hat{p} = \sum x_i/n$. Tests based on the first approximation is called the Wald test while the latter the score test. The score test approximation, with fewer statistics and more parameters values with give the interval which is the asymptotically optimal one. $\left\{p : \left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right| \leq z_{\alpha/2}\right\}$ is the better approximate interval. The exact interval can be obtained by solving the quadratic in p by squaring the two sides of the inequality and noting that the quadratic opens upward and then p lies between the two roots of the quadratic. This interval can be further improved by using a continuity correction by solving two separate quadratics. For larger root or upper interval endpoint $\left|\frac{\hat{p}+\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right| \leq z_{\alpha/2}$ and for the smaller root or lower interval endpoint $\left|\frac{\hat{p}-\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right| \leq z_{\alpha/2}$. At the endpoints there are approximation to satisfy the structural constraints. □

Comparing the three intervals - Wald, score and LRT, we find the score interval is the longest. The continuity corrected score interval, although longer, is the interval of choice for small n as it maintains optimal coverage probability. The LRT and Wald procedure intervals may be too short for small n, with the Wald interval also suffering from endpoint maladies.

Example 10.17. If $X_1, \ldots, X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, we have the approximate interval for θ, $\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{Var[\hat{\theta}_M]}{n}}$. The overoptimism of the $z_{\alpha/2}$ cutoff gives worse results than simply using mean and variance, but only slightly so. □

Example 10.18. Let $X_1, \ldots, X_n$ be iid negative binomial(r, p). r is known and we want a CI on p. Using the fact that $Y = \sum X_i \sim$ negative binomial(nr, p) we can form intervals in a number of ways. Using a variation of the binomial-F distribution relationship, we can from an exact CI. Or we can use a normal approximation. Interestingly, as $p \to 0$, $2pY \to \mathcal{X}^2_{2nr}$ in distribution. So, for small p, $2pY$ is a pivot. Using this fact, we can construct a pivotal $1-\alpha$ confidence interval, valid for small p: $\left\{p : \frac{\mathcal{X}^2_{2nr,1-\alpha/2}}{2y} \leq p \leq \frac{\mathcal{X}^2_{2nr,\alpha/2}}{2y}\right\}$.

□

11 Analysis of Variance and Regression

ANOVA and regression analysis are based on underlying assumption of a linear relationship between and independent random variable and covariates. ANOVA concerns with analyzing variation in means, by partitioning variation. In simple linear regression, the mean of a random variable Y, is modelled as a function of another observable variable x, by the relationship $E[Y] = \alpha + \beta x$, a linear population regression function.

11.1 Oneway Analysis of Variance

In the oneway analysis of variance we assume that data Y_{ij} are observed according to a model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, k \quad j = 1, \ldots, n_i,$$

where θ_i are unknown parameters and the ϵ_{ij} are the error random variables. Schematically, the data y_{ij} from oneway ANOVA will look like this table. Note we do not assume that there are equal numbers of observations in each treatment group. Without loss of generality we can assume $E[\epsilon_{ij}] = 0$, since if not, we can rescale the ϵ_{ij} and absorb the leftover mean into θ_i. Thus, $E[Y_{ij}] = \theta_i$, $j = 1, \ldots, n_i$. The θ_is are usually referred to as *treatment means.* There is an alternative model called overparametrized model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, where again $E[\epsilon_{ij}] = 0$. It follows that $E[Y_{ij}] = \mu + \tau_i$, where μ is the grand mean and parameters τ_i are deviations from the mean level that is caused by the treatment. However, we cannot estimate both τ_i and μ separately, because there are problems with identifiability.

1	2	3	...	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	y_{k3}
		y_{3n_3}		$\vdots$
y_{1n_1}				
	y_{2n_2}			y_{kn_k}

Definition 11.1. *A parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct distributions, i.e. if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$.*

Identifiability is a property of the model, not of an estimator or estimation procedure. If a model is not identifiable, then there is difficulty in doing inference, as both θ and θ' would give the same likelihood value. Problems with identifiability can usually be solved by redefining the model. For example, here we have $k + 1$ parameters, $(\mu, \tau_1, \ldots, \tau_k)$ but only k means $E[Y_{ij}]$, $i = 1, \ldots, k$. We can add the restriction $\sum_{i=1}^{k} \tau_i = 0$, which effectively reduces the number of parameters to k and makes the model identifiable. The restriction also has the effect of giving the τ_is an interpretation as deviations from an overall mean level.

11.1.1 Model and Distribution Assumptions

The minimum assumption that is needed before any estimation can be done is that $E[\epsilon_{ij}] = 0$ and $Var[\epsilon_{ij}] < \infty$ for all i, j. Under these assumptions we can do some estimation of the θ_is. However, to do any confidence interval estimation or testing, we need distributional assumptions.

Definition 11.2. *Oneway ANOVA assumptions: Random variables Y_{ij} are observed according to the model*

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, k \quad j = 1, \ldots, n_i$$

where

1. *$E[\epsilon_{ij}] = 0$, $Var[\epsilon_{ij}] = \sigma_i^2 < \infty$, for all i, j. $Cov[\epsilon_{ij}, \epsilon_{i'j'}] = 0$ for all i, i', j, j', unless $i = i'$ and $j = j'$.*

2. *The ϵ_{ij} are independent and normally distributed.*

3. *$\sigma_i^2 = \sigma^2$ for all i, equality of variance, also known as homoscedasticity.*

If we assume some distribution other than normal, intervals and tests can be difficult but still possible to derive. With reasonable sample sizes and populations that are not too asymmetric, we have CLT to rely on. The equality of variance assumption is linked to the normality assumption. If it is suspected that the data badly violates the ANOVA assumptions, we can try to transform the data nonlinearly, e.g. using Box-Cox family of power transformations. The problem of estimating means when variances are unequal is known as **Behrens-Fisher problem**.

11.1.2 The Classic ANOVA Hypothesis

The classic ANOVA test is a test of the null hypothesis $H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$ versus $H_1 : \theta_i \neq \theta_j$, for some i, j. The real interest in ANOVA is not in testing but in estimation. Using the connection between testing and interval estimation, we can derive confidence regions by deriving, then inverting, appropriate tests. To find a statistical description of the θ_is we can break down the ANOVA hypotheses into smaller, more easily describable pieces using the union-intersection method. The ANOVA null is an intersection of more easily understood univariate hypotheses expressed in terms of *contrasts*.

Definition 11.3. *Let $t = (t_1, \ldots, t_k)$ be a set of variables, either parameters or statistics, and let $a = (a_1, \ldots, a_k)$ be known constants. The function $\sum_{i=1}^{k} a_i t_i$ is called linear combination of the t_is. Further, if $\sum a_i = 0$, it is called a* ***contrast****.*

Contrasts can be used to compare treatment means. For example, if we have the means $\theta_1, \ldots, \theta_k$ and constants $a = (1, -1, 0, \ldots, 0)$, then $\sum a_i\theta_i = \theta_1 - \theta_2$ is a contrast that compares θ_1 to θ_2.

Theorem 11.1. *Let $\theta = (\theta_1, \ldots, \theta_k)$ be arbitrary parameters. Then*

$$\theta_1 = \theta_2 = \ldots = \theta_k \iff \sum_{i=1}^{k} a_i\theta_i = 0 \textit{ for all } a \in \mathcal{A},$$

where $\mathcal{A}$ is the set of constants satisfying $\mathcal{A} = \{a = (a_1, \ldots, a_k) : \sum a_i = 0\}$; that is, all contrasts must satisfy $\sum a_i\theta_i = 0$.

The ANOVA null can be expressed as a hypothesis about contrasts. That is, the null hypothesis is true iff the hypothesis $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ for all $(a_1, \ldots, a_k)$ such that $\sum_{i=1}^k a_i = 0$ is true. The ANOVA alternative, $H_1 : \sum_{i=1}^k a_i\theta_i \neq 0$ for some $(a_i, \ldots, a_k)$ such that $\sum_{i=1}^k a_i = 0$, means at least one contrast is nonzero. The real gain is that the use of the contrasts allows us to operate in a univariate manner.

11.1.3 Inferences Regarding Linear Combinations of Means

Most interesting inferences in an ANOVA can be expressed as contrasts or sets of contrasts. Working under the oneway ANOVA assumptions, we have that $Y_{ij} \sim \mathcal{N}(\theta_i, \sigma^2)$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Therefore, $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim \mathcal{N}(\theta_i, \sigma^2/n_i)$, $i = 1, \ldots, k$. A replaced $\cdot$ (dot) means that the subscript has been summed over, e.g. $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ and $Y_{\cdot j} = \sum_{i=1}^{k} Y_{ij}$. The addition of a 'bar' indicates that a mean is taken. $b\bar{a}rY$ mean grand mean over both subscripts.

For any constants $a = (a_1, \ldots, a_k)$, $\sum_{i=1}^k a_i\bar{Y}_{i\cdot}$ is also normal with $E\left[\sum_{i=1}^k a_i\bar{Y}_{i\cdot}\right] = \sum_{i=1}^k a_i\theta_i$ and $Var\left[\sum_{i=1}^k a_i\bar{Y}_{i\cdot}\right] = \sigma^2\sum_{i=1}^k \frac{a_i^2}{n_i}$, and furthermore

$$\frac{\sum_{i=1}^k a_i\bar{Y}_{i\cdot} - \sum_{i=1}^k a_i\theta_i}{\sqrt{\sigma^2\sum_{i=1}^k \frac{a_i^2}{n_i}}} \sim \mathcal{N}(0, 1).$$

When σ is to be inferred we replace σ with sample variance $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$, $i = 1, , k$ and $(n_i - 1)S_i^2/\sigma^2 \sim \mathcal{X}^2_{n_i-1}$. Under the ANOVA assumptions, since each S_i^2 estimates the same σ^2, we can improve the estimator by combining them. We thus use the pooled estimator of σ^2, S_p^2, given by $S_p^2 = \frac{1}{N-k}\sum_{i=1}^k(n_i - 1)S_i^2 = \frac{1}{N-k}\sum_{i=1}^k\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$. Also, $(N - k)S_p^2/\sigma^2 \sim \mathcal{X}^2_{N-k}$. Also, S_p^2 is independent of each $Y_{i\cdot}$, thus we have a Student's t with $N - k$ degrees of freedom.

$$\frac{\sum_{i=1}^k a_i\bar{Y}_{i\cdot} - \sum_{i=1}^k a_i\theta_i}{\sqrt{S_p^2\sum_{i=1}^k a_i^2/n}} \sim t_{N-k}.$$

To test $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ versus $H_1 : \sum_{i=1}^k a_i\theta_i \neq 0$ at level α, we would reject H_0 if

$$\left|\frac{\sum_{i=1}^k a_i\bar{Y}_{i\cdot}}{\sqrt{S_p^2\sum_{i=1}^k a_i^2/n_i}}\right| > t_{N-k,\alpha/2}.$$

Furthermore, since we have a pivot here that can be inverted to give an estimator of $\sum a_i\theta_i$. With probability $1 - \alpha$,

$$\sum_{i=1}^k a_i\bar{Y}_{i\cdot} - t_{N-k,\alpha/2}\sqrt{S_p^2\sum_{i=1}^k \frac{a_i^2}{n_i}} \;\leq\; \sum_{i=1}^k a_i\theta_i \;\leq\; \sum_{i=1}^k a_i\bar{Y}_{i\cdot} + t_{N-k,\alpha/2}\sqrt{S_p^2\sum_{i=1}^k \frac{a_i^2}{n_i}}.$$

Example 11.1. Take $a = (1, -1, 0, \ldots, 0)$ to compare treatment 1 and 2. To test $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, we would reject H_0 if

$$\left|\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right| > t_{N-k,\alpha/2}.$$

Note, this is different from two-sample t test, in the sense that here information from treatments $3, \ldots, k$, as well as treatment 1 and 2, is used to estimate σ^2. To compare treatment 1 to average of treatment 2 and 3, we would take $a = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \ldots, 0)$. Thus, we have a way of testing or estimating any linear combination in the ANOVA. For example, if we look at the contrasts $\theta_1 - \theta_2$, $\theta_2 - \theta_3$, and $\theta_1 - \theta_3$, we can learn something about the ordering of the θ_is. To take care of overall α level when doing a number of tests or intervals, we can use Bonferroni Inequality. □

11.1.4 The ANOVA F Test

The ANOVA hypothesis can be written as $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ for all $a \in \mathcal{A}$ versus $H_1 : \sum_{i=1}^k a_i\theta_i \neq 0$ for some $a \in \mathcal{A}$, where $\mathcal{A} = \{a = (a_1, \ldots, a_k) : \sum_{i=1}^k = 0\}$. To see this as union-intersection test, define, for each a, the set $\Theta_a = \{\theta = (\theta_1, \ldots, \theta_k) : \sum_{i=1}^k a_i\theta_i = 0\}$. Then, by the previous theorem we have $\theta \in \{\theta :_1 = \theta_2 = \ldots = \theta_k\} \iff \theta \in \Theta_a$ for all $a \in \mathcal{A} \iff \theta \in \underset{a\in\mathcal{A}}{\cap} \Theta_a$, showing that the ANOVA null can be written as an intersection. We would reject $H_0 : \theta \in \underset{a\in A}{\cap} \Theta_a$ if we reject $H_{0_a} : \theta \in \Theta_a$ versus $H_{1_a} : \theta \notin \Theta_a$ for any a. We test H_{0_a} with the t statistic

$$T_a = \left|\frac{\sum_{i-1}^k a_i\bar{Y}_{i\cdot} - \sum_{i=1}^k a_i\theta_i}{\sqrt{S_p^2 \sum_{i=1}^k a_i^2/n_i}}\right|.$$

When then reject H_{0_a} if $T_a > k$ for some constant k. If we could reject for any a, we could reject for the a that maximizes T_a. Thus, the union-intersection test of the ANOVA null is to reject H_0 if $\sup_a T_a > k$, where k is chosen so that $P_{H_0}[\sup_a T_a > k] = \alpha$. Calculation of $\sup_a T_a$ is done by solving for a constrained maximum by using Cauchy-Schwarz Inequality.

Theorem 11.2. *Let $(v_1, \ldots, v_k)$ be constants and let $(c_1, \ldots, c_k)$ be positive constants. Then, for $\mathcal{A} = \{a = (a_1, \ldots, a_k) : \sum a_i = 0\}$,*

$$\max_{a\in\mathcal{A}}\left[\frac{(\sum_{i=1}^k a_iv_i)^2}{\sum_{i=1}^k a_i^2/c_i}\right] = \sum_{i=1}^k c_i(v_i - \bar{v}_c)^2$$

where $\bar{v}_c = \sum c_iv_i / \sum c_i$. The maximum is attained at any a of the form $a_i = Kc_i(v_i - \bar{v}_c)$, where K is a nonzero constant.

Theorem 11.3. *For T_a,*

$$\sup_{a:\sum a_i=0} T_a^2 = \frac{1}{S_p^2}\left(\sum_{i=1}^k n_i\left[(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta})\right]^2\right),$$

where $\bar{\bar{Y}} = \sum n_i \bar{Y}_{i\cdot} / \sum n_i$ *and* $\bar{\theta} = \sum n_i \theta_i / \sum n_i$. *Furthermore, under the ANOVA assumptions,*

$$\frac{1}{k-1} \sup_{a:\sum a_i = 0} T_a^2 \sim F_{k-1,N-k},$$

where $N = \sum n_i$.

If $H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$ is true, $\theta_i = \bar{\theta}$ for all $i = 1, \ldots, k$ and the $\theta_i - \bar{\theta}$ terms drop out. Thus, for an α level test of the ANOVA hypotheses $H_0 : \theta_1 = \theta_2, \ldots = \theta_k$ versus $H_1 : \theta_i : \theta_i \neq \theta_j$ for some i, j we reject H_o if

$$F = \frac{1}{S_p^2} \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 / (k-1) > F_{k-1,N-k,\alpha},$$

and is called the ANOVA F statistic.

11.1.5 Simultaneous Estimation of Contrasts

In ANOVA, we often want to make more than one inference and the simultaneous inference from many α level tests is not necessarily at level α.

Example 11.2. *Pairwise differences:* If an ANOVA has means $\theta_1, \ldots, \theta_k$, there may be interest in interval estimates of $\theta_1 - \theta_2, \theta_2 - \theta_3, \theta_3 - \theta_4, \ldots$. For,

$$C_{ij} = \left\{ \theta_i - \theta_j : \theta_i - \theta_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_{N-k,\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \right\}.$$

Then $P[C_{ij}] = 1 - \alpha$ for each C_{ij}, but, for example, $P[C_{12} \text{ and } C_{23}] < 1 - \alpha$. With the Bonferroni Inequality, we can build a simultaneous inference statement. Recalling, $P[\cap_{i=1}^{n} A_i] \geq \sum_{i=1}^{n} P[A_i] - (n-1)$. In this case we want to bound $P[\cap_{i,j} C_{ij}]$ by constructing m confidence sets to be level γ, where $1 - \alpha = \sum_{i=1}^{m} \gamma - (m-1)$, implying $\gamma = 1 - \frac{\alpha}{m}$. A slight generalization is possible by not requiring each individual inference at the same level. We can construct each confidence set to be of level γ_i, where $1 - \alpha = \sum_{i=1}^{m} \gamma_i - (m-1)$. In an ANOVA with k treatments, simultaneous inference on all $k(k-1)/2$ pairwise differences can be made with confidence $1 - \alpha$ if each t interval has confidence $1 - 2\alpha/[k(k-1)]$. □

For testing small number of contrasts Bonferroni bounds are preferred. Alternatively for large number of contrasts, Scheffe's **S method** allows for simultaneous confidence intervals on all contrasts, setting up confidence coefficient that will be valid for all contrast intervals simultaneously.

Theorem 11.4. *Under the ANOVA assumptions, if* $M = \sqrt{(k-1)F_{k-1,N-k,\alpha}}$, *then the probability is* $1 - \alpha$ *that*

$$\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - M \sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \leq \sum_{i=1}^{k} a_i \theta_i \leq \sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} + M \sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

simultaneously for all $a \in \mathcal{A} = \{a = (a_1, \ldots, a_k) : \sum a_i = 0\}$.

S methods allows legitimate 'data snooping' because intervals or tests are valid for *all* contrasts, whether already observed in the data or not. We pay for all the inferential power the longer intervals. If we compare t and F distributions $t_{\nu,alpha/2} \leq \sqrt{(k-1)F_{k-1,\nu,\alpha}}$ for any ν, α, and k. Hence, Scheffe intervals are always wider than single-contrast intervals.

11.1.6 Partitioning Sums of Squares

Theorem 11.5. *For any numbers y_{ij}, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$,*

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2,$$

where $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_j y_{ij}$ and $\bar{\bar{y}} = \sum_i n_i \bar{y}_{i\cdot}.\sum_i n_i$.

These are called the sums of squares and measure variations in the data ascribing to different sources. The total variation is a sum of variations between treatments and variations within treatments. Under normality, the corrected sums of squares are chi squared random variables and can be added together to get new chi squared variable.

If $Y_{ij} \sim \mathcal{N}(\theta_i, \sigma^2)$, then $\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \mathcal{X}^2_{N-k}$, because for each $i = 1, \ldots, k$, $\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \mathcal{X}^2_{n_i-1}$, all independent, and for independent chi squared random variables, $\sum_{i=1}^{k}\mathcal{X}^2_{n_i-1} \sim \mathcal{X}^2_{N-k}$. Furthermore, if $\theta_i = \theta_j$ for every i, j then $\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \sim \mathcal{X}^2_{k-1}$ and $\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\bar{Y}})^2 \sim \mathcal{X}^2_{N-1}$. Thus, under $H_0 : \theta_1 = \ldots = \theta_k$, the sum of squares partitioning is a partitioning of chi squared random variables, the left side of the equation when scaled is distributed as a $\mathcal{X}^2_{N-1}$, and the right hand side sum of two independent random variables $\mathcal{X}^2_{k-1}$ and $\mathcal{X}^2_{N-k}$. This partitioning is true only if the two terms on the right hand side are independent, which follows due to the assumptions of ANOVA. Cochran's Theorem has been generalized to the extent that necessary and sufficient conditions are known for the distribution of, not necessarily iid, squared normals to be chi squared.

In general, it is possible to partition a sum of squares into sums of squares of uncorrelated contrasts, each with 1 degree of freedom. If the sum of sqaures has ν degrees of freedom and is $\mathcal{X}^2_{\nu}$, it is possible to partition it into ν independent terms, each of which is $\mathcal{X}^2_1$. The quantity $(\sum a_i \bar{Y}_{i\cdot})^2/(\sum a_i^2/n_i)$ is called the contrast sum of squares for a treatment contrast $\sum a_i \bar{Y}_{i\cdot}$. In a oneway ANOVA it is always possible to find sets of constants $a^{(l)} = (a_1^{(l)}, \ldots, a_k^{(l)})$, $l = 1, \ldots, k-1$, to satisfy

$$\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 = \frac{\sum a_i^{(l)}\bar{Y}_{i\cdot}^2}{\sum(a_i^{(l)})^2/n_i} + \ldots + \frac{\sum a_i^{(k-1)}\bar{Y}_{i\cdot}^2}{\sum(a_i^{(k-1)})^2/n_i}$$

and $\sum \frac{a_i^{(l)}a_i^{(l')}}{n_i} = 0$ for all $l \neq l'$. Thus, the individual contrast sums of squares are all uncorrelated and hence independent under normality. This equation when suitably normalized, the left-hand side is distributed as $\mathcal{X}^2_{k-1}$ and the right hand side is $k-1$ $\mathcal{X}^2_1$s. These are called orthogonal contrasts.

It is common to summarize the results of an ANOVA F test in a standard form, called an ANOVA table. The sum of the columns adds - that is $SSB + SSW = SST$. Similarly, degrees of freedom columns adds. The MSW is the usual pooled, unbiased estimator of σ^2, S_p^2.

Source of	Degrees of	Sum of	Mean	F
variation	freedom	squares	square	statistic
Between treatment groups	$k-1$	SSB = $\sum n_i(\bar{y}_i - \bar{\bar{y}})^2$	MSB = $SSB/(k-1)$	$F = \frac{MSB}{MSW}$
Within treatment groups	$N-k$	SSW = $\sum\sum(y_{ij} - \bar{y}_{i\cdot})^2$	MSW = $SSW/(N-k)$	
Total	$N-1$	SST = $\sum\sum(y_{ij} - \bar{\bar{y}})^2$		

Table 3: ANOVA table for oneway classification.

11.1.7 Randomized complete block designs

In Oneway classification of the data there was only one categorization (treatment) in the experiment. In general, the ANOVA allows for many types of categorization, the other most common being Randomized Complete Block (RCB) ANOVA. A block is categorization that is in an experiment for the express purpose of removing variation. In contract to treatment, there is usually no interest in finding block differences. Random variables Y_{ij} are observed according to the model $Y_{ij}|\boldsymbol{b} = \mu + \tau_i + b_j + \epsilon_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, r$, where the random variables $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, r$; and the random variables $B_1, \ldots, B_r$, whose realized values are the blocks $b_1, \ldots, b_r$, are iid $\mathcal{N}(0, \sigma_B^2)$ and are independent of $_{ij}$ for all i, j. The mean and variance of Y_{ij} are $E[Y_{ij}] = \mu + \tau_i$ and $Var[Y_{ij}] = \sigma_B^2 + \sigma^2$. Although the Y_{ij}s are conditionally uncorrelated, there is a correlation in the blocks unconditionally. THe correlation between Y_{ij} and $Y_{i'j}$ in block j, with $i \neq i'$, is $\frac{\sigma_B^2}{\sigma_B^2 + \sigma^2}$, a quantity called the intra-class correlation, implying positive correlation. Even though the Y_{ij}s are not independent, the intra class correlation structure still results in an analysis of variance where ratios of mean squares have the F distribution.

11.2 Simple Linear Regression

In a simple linear regression we have a relationship of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where Y_i is a random variable and x_i is another observable variable. The quantities α and β are *intercept* and *slope* of the regression, are assumed to be fixed and unknown parameters and ϵ_i is, necessarily, a random variable. It is common to suppose $E[\epsilon_i] = 0$, otherwise we could just rescale the excess into α. So we have $E[Y_i] = \alpha + \beta x_i$, *the population regression* function here is linear. Prediction is the main purpose here, and interpreted as saying Y_i depends on x_i. It is common to refer to Y_i as the response variable and to refer to x_i are the *predictor* variable. To emphasize the fact that our inferences about the relationship between Y_i and x_i assume knowledge of x_i, we could write $E[Y_i|x_i] = \alpha + \beta x_i$, to emphasize the conditional aspect. The implicit assumption of linearity may not be justified, because there may not be any underlying theory to support a linear relationship. We might still want to assume that the regression of Y on X can be adequately approximated by a linear function for convenience. Thus we really hope that $E[Y_i|x_i]\alpha + \beta x_i$ is a reasonable approximation. However, if we start from the rather strong assumption that the pair (X_i, Y_i) has a bivariate normal distribution, it immediately follows that the regression of Y on X is linear, and hence $E[Y|x]$ is linear in parameters.

When we do regression analysis, there are two steps.

1. Data-oriented step: Here we attempt to summarize the observed data. This is not a matter of statistical inference as we don't make any assumptions about parameters. In simple linear regression problem, we observe data consisting of n pairs of observations $(x_i, y_1), \dots, (x_n, y_n)$. The sample means are $\bar{x} = \frac{1}{n}\sum x_i$, and $\bar{y} = \frac{1}{n}\sum y_i$. The sum of squares are $S_{xx} = \sum(x_i - \bar{x})^2$ and $S_{yy} = \sum(y_i - \bar{y})^2$, and the sum of cross-products is $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$.

2. Statistical: Here we attempt to infer conclusions about the relationship in the population, i.e. about the population regression function. We need to make assumptions about the population, e.g. we need to assume that there are parameters that correspond to slope and intercept. We will consider a number of different models for the data, entailing different assumptions about whether x or y or both are observed values of random variables X or Y. In each we will be interested in the linear relationship between x and y. We will find many different approaches lead us to the same line, specifically the estimates, $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$, and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

11.2.1 Least Squares: A Mathematical Solution

We make no statistical assumptions about the observations (x_i, y_i) here and simply want to fit a line that is closest to the data cloud. For any line $y = c + dx$, the residual sum of squares (RSS) is defined to be $RSS = \sum_{i=1}^{n}(y_i - (c + dx_i))^2$. It measures the vertical distance for each data point to the line $c + dx$ and then sums the squares of these distances. The **least squares estimates** of α and β are defined to be those values a and b such that the line $a + bx$ minimizes RSS. That is, the least squares estimates, a and b, satisfy $\min\limits_{c,d} \sum_{i=1}^{n}(y_i - (c + dx_i))^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$. We can solve for c and d by taking partial differential with respect to c and d and equating it to zero to get

$$\frac{\partial}{\partial c}\sum(y_i - c - dx_i)^2 = -2\sum(y_i - c - dx_i) = 0 \implies c = \bar{y} - d\bar{c}$$

$$\frac{\partial}{\partial d}\sum(y_i - c - dx_i)^2 = -2\sum x_i(y_i - c - dx_i) = 0 \implies \sum x_i y_i - d\sum x_i^2 - nc\bar{x} = 0.$$

These solve to $d = \frac{S_{xy}}{S_{xx}}$ and $c = \bar{y} - d\bar{x}$. RSS is only one of the many possible ways to measure the distance from the line $c + dx$ to the data points. For example, we could use vertical distance as well. This can be done by interchanging x and y. We find the least squares line is $x = a' + b'y$, where $b' = \frac{S_{xy}}{S_{yy}}$ and $a' = \bar{x} - b'\bar{y}$. Rearranging we get, $y = -(a'/b') + (1/b')x$. Usually the line obtained by considering horizontal distances is different from the line obtained by considering vertical distances. The ratio of the two slopes $b/(1/b') = bb' = \frac{S_{xy}^2}{S_{xx}S_{yy}} \leq 1$, using Holder's Inequality.

If x is the predictor variable and y is the response variable, and we think of predicting y from x, then the vertical distance measured in RSS is a reasonable choice, as it measures the distance between the true value y_i from the predicted value $\hat{y}_i = c + dx_i$. But if we do not make this distinction between x and y, then it is unsettling that another reasonable criterion, horizontal distance, gives a different line!

These least square solutions are simply mathematical in nature and are not 'estimates' derived from a statistical model. We have no basis for constructing confidence intervals or testing hypotheses, in this section, as we have not used any statistical model for the data. But, as we shall see, these least squares solutions have optimality properties in certain statistical models.

11.2.2 Best Linear Unbiased Estimators: A Statistical Solution

Assume that the values $x_1, \ldots, x_n$ are known, fixed values. The values $y_1, \ldots, y_n$ are observed values of uncorrelated random variables $Y_1, \ldots, Y_n$. The linear relationship assumed between the xs and the ys is $E[Y_i] = \alpha + \beta x_i$, $i = 1, \ldots, n$, where we assume $Var[Y_i] = \sigma^2$, that is, all Y_is have the same unknown variance. These assumptions about the first two moments of Y_is are the only assumptions we need to make some inference, for this model. The same model can be expressed by stating that $Y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated random variables with $E[\epsilon_i] = 0$ and $Var[\epsilon)i] = \sigma^2$. The $\epsilon_1, \ldots, \epsilon_n$ are called the random errors.

To derive the estimators for the parameters α and β, we restrict attention to the class of linear estimators, one which is of the form $\sum d_i Y_i$, where $d_1, \ldots, d_n$ are known, fixed constants. Among the class of linear estimators, we further restrict attention to the unbiased estimators, restricting the values of $d_1, \ldots, d_n$ that can be used. An unbiased estimator of the slope β must satisfy $E\left[\sum d_i Y_i\right] = \beta$, regardless of the true values of the parameters α and β. This implies that

$$\beta = E\left[\sum d_i Y_i\right] = \sum d_i E[Y_i] = \sum d_i(\alpha + \beta x_i) = \alpha \sum d_i + \beta \sum d_i x_i.$$

This equality is true for all α and β iff $\sum d_i = 0$ and $\sum d_i x_i = 1$, which are the required conditions on $d_1, \ldots, d_n$ in order for the estimator to be an unbiased estimator of β.

Further, an estimator is the **best linear unbiased estimator (BLUE)** if it is the linear unbiased estimator with the smallest variance. Because $Y_1, \ldots, Y_n$ are uncorrelated with equal variance σ^2, the variance of any linear estimator is given by $Var[\sum d_i Y_i] = \sum d_i^2 Var[Y_i] = \sum d_i^2 \sigma^2 = \sigma^2 \sum d_i^2$. The BLUE of β, therefore have a minimum value of $\sum d_i^2$. Using theorem 11.2 we set $d_i = K(x_i - \bar{x})$, $i = 1, \dots, n$, to maximize $\frac{(\sum d_i x_i)^2}{\sum d_i^2}$ among all $d_1, \dots, d_n$ and choose the one that satisfies $\sum d_i = 0$ and $\sum d_i x_i = 1$. Furthermore, since $\{(d_1, \dots, d_n) : \sum d_i = 0,_i x_i = 1\} \subset \{(d_1, \dots, d_n) : \sum d_i = 0\}$, if d_is we get also satisfy the two required conditions, they certainly fit the bill. Now, we have $\sum d_i x_i = \sum K(x_i - \bar{x}) x_i = K S_{xx}$. This means $K = \frac{1}{S_{xx}}$. Therefore, with $d_1, \dots, d_n$ defined by $d_i = \frac{x_i - \bar{x}}{S_{xx}}$, $i = 1, \dots, n$, both constraints are satisfied and this set of d_is produces the maximum. Finally, note that for all $d_1, \dots, d_n$ that satisfies the two required conditions $\frac{(\sum d_i x_i)^2}{\sum d_i^2} = \frac{1}{\sum d_i^2}$. Thu, for $d_1, \dots, d_n$ that satisfies the two conditions, maximization of left hand term is equivalent to the minimization of $\sum d_i^2$. Hence, we conclude that the d_is as defined above give the minimum value of $\sum d_i^2$ among all d_is that satisfy the required unbiasness conditions, and the linear unbiased estimator defined by these d_is, namely, $b = \sum \frac{(x_i - \bar{x})}{S_{xx}} y_i = \frac{S_{xy}}{S_{xx}}$, is the BLUE of β.

Theorem 11.6. *Gauss-Markov Theorem: Least squares estimates for linear regression are BLUE under the set of following conditions:*

1. *Linearity: The parameters we are estimating using the OLS method must be themselves linear.*

2. *Random: Our data must have been randomly sampled from the population, with error term having mean zero, and no autocorrelation.*

3. *Homoscedasticity: No matter what the values of our regressors might be, the error of the variance is constant.*

4. *Exogeneity: The regressors aren't correlated with the error term.*

5. *Non-Collinearity: The regressors begin calculated aren't perfectly correlated with each other.*

No assumptions on the distribution has been made here, except the existence of first two moments.

11.2.3 Models and Distribution Assumptions

To obtain the inference on the parameters we did not have to specify a complete probability model for the data, only assumptions about the first two moments. We were able to obtain a general optimality property under these minimal assumptions, but the optimality was restricted only to linear unbiased estimators. We were not able to derive exact tests and confidence intervals under this model because the model does not specify enough about the probability distribution of the data. We now present two statistical models that completely specify the probabilistic structure of the data.

Conditional normal model: The values of the predictor variable $x_1, \dots, x_n$ are considered to be known, fixed constants. The values of the response variable $y_1, \dots, y_n$ are observed values of random variable $Y_1, \dots, Y_n$, independent. The distribution of Y_is is normal $Y_i \sim$

$\mathcal{N}(\alpha + \beta x_i, \sigma^2)$, $i = 1, \dots, n$. This can be expressed as $Y_i = \alpha +_i +\epsilon_i$, $i = 1, \dots, n$, where $\epsilon_1, \dots, \epsilon_n$, are iid $\mathcal{N}(0, 1)$ independent random variables. Notice that we have strengthened the uncorrelatedness of $Y_1, \dots, Y_n$ to independence and instead of just two moments we now specify the exact probability distribution. Using independence, the joint pdf of $Y_1, \dots, Y_n$ is given by

$$f(\boldsymbol{y}|\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} f(y_i|\alpha, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\Big\{ - \frac{1}{2\sigma^2} \sum (y_i - \alpha - \beta x_i)^2) \Big\}$$

which can be used to develop the statistical procedures, e.g. finding MLE of α, β, and σ^2.

Bivariate normal model: Experimenters certainly do no choose the values of x. In the bivariate normal model the data (x_i, y_i) is observed value of the bivariate random vector (X_i, Y_i), where $i = 1, \dots, n$. The random vectors are independent and the join distribution is bivariate, $(X_i, Y_i) \sim \mathcal{N}_2(\mu_X, \mu_Y, ^2_X, \sigma^2_Y, \rho)$.

The prediction problem lends itself naturally to the conditional distribution of Y given $X = x$. For a bivariate normal model this conditional random variable is normal as well. The population regression function is now a true conditional expectation,

$$E[Y|x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X + \rho\frac{\sigma_Y}{\sigma_X}x.$$

The bivariate normal model *implies* that the population regression is a linear function of x, and we need not assume it as in the previous models. Here, $E[Y|x] = \alpha + \beta x$, where $\beta = \rho\frac{\sigma_Y}{\sigma_X}$ and $\alpha = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$. Also, as in the conditional normal model, the conditional variance of the response variable Y does not depend on x, $Var[Y|x] = \sigma^2_Y(1 - \rho^2)$.

In simple linear regression we do not use the fact of bivariate normality except to define the conditional distribution, and the marginal of X is of no consequence. Inference based on point estimators, intervals, or tests is the same for the two models.

11.2.4 Estimation and Testing with Normal Errors

We find the MLE of the three parameters α, β and σ^2. Using the joint pdf we see that the log likelihood function is

$$\log L(\alpha, \beta, \sigma^2|\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{n}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

For any fixed value of σ^2, $\log L$ is maximized as a function of α and β by those values, $\hat{\alpha}$ and $\hat{\beta}$, that minimize $\sum(y_i - \alpha - \beta x_i)^2$. But this function is just the RSS from the previous section and is minimized by $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \widehat{beta}\bar{x}$. Thus, the least square estimators of α and β are also the MLEs of α and β. Now, substituting in the log likelihood, to find the MLE of σ^2 we need to maximize

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Differentiating and equating to 0 gives $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$, the RSS, evaluated at the least square line, divided by the sample size.

$\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β. However, $\hat{\sigma}^2$ is not an unbiased estimator of σ^2.

Theorem 11.7. *Let $Y_1, \ldots, Y_n$ be uncorrelated random variables with $Var[Y_i] = \sigma^2$ for all $i = 1, \ldots, n$. Let $c_1, \ldots, c_n$ and $d_1, \ldots, d_n$ be two sets of constants. Then $Cov\left[\sum c_iY_i, \sum d_iY_i\right] = \left(\sum c_id_i\right)\sigma^2$.*

To find the bias of σ^2 we have $\epsilon_i = Y_i - \alpha - \beta x_i$. We define the residuals from the regression to be $\hat{\epsilon_i} = Y_i - \hat{\alpha} - \hat{\beta}x_i$, and thus $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\epsilon}_i^2 = \frac{1}{n}RSS$. It can be shown that $E[\hat{\epsilon}_i] = 0$, and $Var[\hat{\epsilon}_i] = E[\hat{\epsilon}_i^2]$ and thus, $E[\hat{\sigma}^2] = \frac{1}{n}\sum E[\hat{\epsilon}_i^2] = \frac{n-2}{n}\sigma^2$. The MLE $\hat{\sigma}^2$ is a biased estimator of σ^2. The more commonly used estimator of σ^2, which is unbiased, is

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}\sum \hat{\epsilon}_i^2.$$

Theorem 11.8. *Under the conditional normal regression model, the sampling distributions of the estimators $\hat{\alpha}$, $\hat{\beta}$ and S^2 are*

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{nS_{xx}}\sum x_i^2\right), \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right),$$

with

$$Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2\bar{x}}{S_{xx}}.$$

Furthermore, $(\hat{\alpha}, \hat{\beta})$ and S^2 are independent and

$$\frac{(n-2)S^2}{\sigma^2} \sim \mathcal{X}_{n-2}^2.$$

Proof: We first show that $\hat{\alpha}$ and $\hat{\beta}$ have indicated normal distribution. The estimators $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of the independent normal random variables $Y_1, \ldots, Y_n$. Specifically,

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

Now, we know that the linear combination of mutually independent normal random variables is a normal itself. To find the mean of this normal distribution we note $\sum(x_i - \bar{x}) = 0$ and hence

$$E[\hat{\beta}] = E\left[\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)y_i\right] = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)E[y_i] = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)(\alpha + \beta y_i) = \beta\sum_{i=1}^{n}\frac{x_i(x_i - \bar{x})}{S_{xx}} = \beta.$$

Further, to calculate the variance for $\hat{\beta}$ we note

$$Var[\hat{\beta}] = Var\left[\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)y_i\right] = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 Var[y_i] = Var[Y_i]\frac{\sum(x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

This shows,

$$\boxed{\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right)}.$$

The estimator $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ can be expressed as a linear combination as well,

$$\hat{\alpha} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) y_i.$$

Again since this is a linear combination of independent normal variables, $\hat{\alpha}$ itself is normally distributed. To find the mean we note after some simple algebra and cancellation of terms,

$$E[\hat{\alpha}] = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) E[y_i] = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) (\alpha + \beta x_i) = \alpha,$$

and

$$Var[\hat{\alpha}] = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \sigma^2 \left(\frac{\sum x_i^2}{nS_{xx}}\right).$$

This shows,

$$\boxed{\hat{\alpha} \sim \mathcal{N}\left(\alpha, \sigma^2 \frac{\sum x_i^2}{nS_{xx}}\right)}.$$

To find the correlation between $\hat{\alpha}$ and $\hat{\beta}$ we use the earlier stated theorem to note

$$\begin{aligned} Cov[\hat{\alpha}, \hat{\beta}] &= Cov\left[\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) y_i, \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) y_i\right] \\ &= \left[\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)\right] \sigma^2 \\ &= -\frac{\sigma^2 \bar{x}}{S_{xx}}. \end{aligned}$$

To show $\hat{\alpha}$ and $\hat{\beta}$ are independent of S^2 we start with the definition of $\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$, which is normally distributed. For $\hat{\alpha} = \sum_j c_j Y_j$ and $\hat{\beta} = \sum_j d_j Y_j$, where $c_j = \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{S_{xx}}$ and $d_j = \frac{(x_j - \bar{x})}{S_{xx}}$. Substituting we get

$$\hat{\epsilon}_i = \sum_j [\delta_{ij} - (c_j - d_j x_i)] Y_j.$$

Hence,

$$\begin{aligned} Cov[\hat{\epsilon}_i, \hat{\alpha}] &= Cov\left[\sum_j [\delta_{ij} - (c_j - d_j x_i)] Y_j, \sum_j c_j Y_j\right] \\ &= \sigma^2 \sum_j \left[c_j(\delta_{ij} - c_j - d_j x_i)\right] \\ &= \sigma^2 \left[c_i - \sum_j c_j^2 - x_i \sum_j c_j d_j\right] = 0. \end{aligned}$$

This uses the facts $c_i = \frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}}$, $\sum_j c_j^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}$, and $x_i \sum_j c_j d_j = -\frac{x_i \bar{x}}{S_{xx}}$. Similarly,

$$\begin{aligned} Cov[\hat{\epsilon}_i, \hat{\beta}] &= Cov\left[\sum_j [\delta_{ij} - (c_j - d_j x_i)] Y_j, \sum_j d_j Y_j\right] \\ &= \sigma^2 \sum_j \left[d_j(\delta_{ij} - c_j - d_j x_i)\right] \\ &= \sigma^2 \left[d_i - \sum_j c_j d_j - x_i \sum_j d_j^2\right] = 0, \end{aligned}$$

where we use $d_i = \frac{(x_i-\bar{x})}{S_{xx}}$, $\sum_j c_j d_j = -\frac{\bar{x}}{S_{xx}}$, and $x_i \sum_j d_j^2 = \frac{1}{S_{xx}}$. Hence, we established that $Cov[\hat{\epsilon}_i, \hat{\alpha}] = Cov[\hat{\epsilon}_i, \hat{\beta}] = 0$, $i = 1, \ldots, n$. Noting the normality of all the variables involved we prove the independence. Further, since $S^2 = \sum \hat{\epsilon}_i^2/(n-2)$ is simply a function of $\hat{\epsilon}_i$ which is independent of $\hat{\alpha}$ and $\hat{\beta}$ we prove independence of for S^2 to $\hat{\alpha}$ and $\hat{\beta}$.

To prove that $(n-2)S^2/\sigma^2 \sim \mathcal{X}_{n-2}^2$, we write $(n-2)S^2$ as a sum of $n-2$ independent random variables, each of which has a $\mathcal{X}_1^2$ distribution. That is, we find constants a_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, n-2$, that satisfy

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{j=1}^{n-2} \left(\sum_{i=1}^{n} a_{ij} Y_i\right)^2,$$

where $\sum_{i=1}^n a_{ij} = 0$, $j = 1, \ldots, n-2$, and $\sum_{i=1}^n a_{ij} a_{ij'} = 0$, $j \neq j'$. The RSS from the linear regression contains information about the worth of a polynomial fit of higher order, over and above a liner fit, which we are ignoring as random variation for a linear fit. A general recursive formula can be found (Robson 1959) to get coefficients for such higher-order polynomial fits, which can be adapted to explicitly find the a_{ij}s. Alternatively, Cochran's Theorem can be used to establish that $\sum \hat{\epsilon}_i^2/\sigma^2 \sim \mathcal{X}_{n-2}^2$. □

Inferences regarding the two parameters α and β are usually based on the two Student's t distributions. Following the previous theorem we have

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{(\sum x_i^2)/(nS_{xx})}} \sim t_{n-2}$$

and

$$\frac{\hat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}.$$

For simultaneous inference of α and β the joint distribution of these two t statistics, called the bivariate Student's t distribution, is used. Usually there is more interest in β than in α. The parameter α is the expected value of Y at $x = 0$, $E[Y|x = 0]$. β is the rate of change or $E[Y|x]$ as a function of x.

If $\beta = 0$, then $E[Y|x] = \alpha$ and $Y \sim \mathcal{N}(\alpha, \sigma^2)$, which does not depend on x. In a well thought out experiment leading to regression analysis we do not expect this to be the case, but we would be interested in knowing this if it were true. The test that $\beta = 0$ is quite similar to the ANOVA test that all treatments are equal. In the ANOVA the null hypothesis states that the treatments are unrelated to the response *in any way*, while in linear regression the null hypothesis $\beta = 0$ states that the treatments (x) are unrelated to the response in a linear way.

To test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ we use the Student's t distribution stated above. We reject H_0 at level α if

$$\Big|\frac{\hat{\beta} - 0}{S/\sqrt{S_{xx}}}\Big| > t_{n-2,\alpha/2} \quad \text{or equivalently} \quad \frac{\hat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2,\alpha}.$$

We note that $RSS = \sum \hat{\epsilon}_i$, giving us

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{S_{xy}^2/S_{xx}}{RSS/(n-2)} = \frac{\text{Regression sum of squares}}{\text{Residual sum of squares/df}}.$$

This formula is summarized in the *regression ANOVA table.* For the above tests the ANOVA table looks like in table 4. The table involves only a hypothesis about β. The parameters α and the estimate $\hat{\alpha}$ play the same role here as the grand mean did in the previous section, they merely serve to locate the overall level of the data and are 'corrected' for in the sum of squares.

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression (slope)	1	Reg. SS= S_{xy}^2/S_{xx}	MS(Reg)= Reg. SS	$F = \frac{MS(Reg)}{MS(Resid)}$
Residual	n-2	$RSS = \sum \hat{\epsilon}_i^2$	MS(Resid)= RSS/(n-2)	
Total	n-1	SST = $\sum(y_i - \bar{y})^2$		

Table 4: ANOVA table for simple linear regression.

the partitioning of the sum of squares of the ANOVA has an analogue in regression.

Total sum of squares = Regression sum of squares + Residual sum of squares.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1} n(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. The RSS measures deviation of the fitted line from the observed values, and the regression sum of squares measures the deviation of predicted values from the grand mean. This summation is valid because of the disappearance of the cross-term. In the expression $\sum(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}$, the right hand side is easier for computation and provides link with the t test, while the left hand side is easily to interpret.

A statistic that is used to quantify how well the fitted line describes the data is the **coefficient of determination** or $\mathbf{R^2}$.

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

The coefficient of determination measures the proportion of the total variation in $y_1, \ldots, y_n$, measured by S_{yy} that is explained by the fitted line, measured by the regression sum of squares. $0 \leq R^2 \leq 1$, is same as the square of the sample correlation coefficient of the n pairs $(y_1, x_1), \ldots, (y_n, x_n)$

The distribution of $\hat{\beta}$ can be used to construct a $100(1-\alpha)\%$ confidence interval for β given by

$$\hat{\beta} - t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}}.$$

Also, a level α test of $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ rejects H_0 if

$$\left|\frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}}\right| > t_{n-2,\alpha/2}.$$

It is common to test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ to determine if there is some linear relationship between the predictor and the response variables. However, the above test is more general, since any value of β_0 can be specified. The regression ANOVA, which is locked into a 'recipe', can test only $H_0 : \beta = 0$.

11.2.5 Estimation and Prediction at a Specified $x = x_0$

Associated with a specified value of the predictor variables, say $x = x_0$, there is a population of Y values distributed as $Y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$. We are interested in quantifying this distribution for prediction purposes. We assume that $(x_1, Y_1), \ldots, (x_n, Y_n)$ satisfy the conditional normal regression model, and based on these n observations we have the estimates $\hat{\alpha}$, $\hat{\beta}$, and S^2. Let x_0 be a specified value of the predictor variable. We start with estimating the mean of the Y population associated with x_0, that is $E[Y|x_0] = \alpha + \beta x_0$. The obvious choice four our point estimator is $\hat{\alpha} + \hat{\beta}x_0$, the plug in estimate. This is an unbiased estimator since $E[\hat{\alpha} + \hat{\beta}x_0] = E[\hat{\alpha}] + E[\hat{\beta}]x_0 = \alpha + \beta x_0$. Further we can calculate the variance as

$$Var[\hat{\alpha} + \hat{\beta}x_0] = Var[\hat{\alpha}] + Var[\hat{\beta}]x_0^2 + 2x_0 Cov[\hat{\alpha}, \hat{\beta}] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right).$$

Finally, since $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of $Y_1, \ldots, Y_n$, so is $\hat{\alpha} + \hat{\beta}x_0$. Thus $\hat{\alpha} + \hat{\beta}x_0$ has a normal distribution, specifically,

$$\hat{\alpha} + \hat{\beta}x_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

We showed already that $\hat{\alpha}$ and $\hat{\beta}$ are independent of S^2. Thus S^2 is also independent of $\hat{\alpha} + \hat{\beta}x_0$ and

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

This can be inverted to get the $100(1-\alpha)\%$ confidence interval for $\alpha+\beta x_0$

$$\hat{\alpha}+\hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} \leq \alpha+\beta x_0 \leq \hat{\alpha}+\hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}.$$

The length of this interval is shorter if x_0 is near $\bar{x}$, i.e. we can estimate more precisely near the center of the data we observed.

A type of inference we have not discussed until now is prediction of an, as yet, unobserved random variable Y.

Definition 11.4. *A* $100(1-\alpha)\%$ *prediction interval for an unobserved random variable* Y *based on the observed data* X *is a random interval* $[L(X), U(X)]$ *with the property that* $P_\theta[L(X) \leq Y \leq U(X)] \geq 1-\alpha$ *for all values of the parameter* θ.

Prediction interval are similar to a confidence interval, but a prediction interval on a random variable, rather than a parameter, and hence is more wider than a confidence interval of the same level. We assume that the new observation Y_0 to be taken at $x=x_0$ has a $\mathcal{N}(\alpha+\beta x_0, \sigma^2)$ distribution, independent of the previous data. The estimators $\hat{\alpha}$, $\hat{\beta}$, and S^2 are calculated form the previous data and, thus, Y_0 is independent of $\hat{\alpha}$, $\hat{\beta}$, and S^2. Hence, we find that $Y_0-(\hat{\alpha}+\hat{\beta}x_0)$ has a normal distribution with mean $E[Y_0-(\hat{\alpha}+\hat{\beta}x_0)] = \alpha+\beta x_0-(\alpha+\beta x_0) = 0$ and variance $Var[Y_0-(\hat{\alpha}+\hat{\beta}x_0)] = Var[Y_0]+Var[\hat{\alpha}+\hat{\beta}x_0] = \sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)$. Using the independence of S^2 and $Y_0-(\hat{\alpha}-\hat{\beta}x_0)$, we see that

$$T = \frac{Y_0-(\hat{\alpha}+\hat{\beta}x_0)}{S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

which can be rearranged in the usual way to obtain the $100(1-\alpha)\%$ prediction interval,

$$\hat{\alpha}+\hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} < Y_0 < \hat{\alpha}+\hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}.$$

11.2.6 Simultaneous Estimation and Confidence Bands

For simultaneous inference, predicting at many x_0s the value $E[Y|x_{0i}]$, $i=1,\ldots,m$ a reasonably good solution is to use the Bonferroni Inequality. We can state the probability is at least $1-\alpha$ that

$$\hat{\alpha}+\hat{\beta}x_{0i} - t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n}+\frac{(x_{0i}-\bar{x})^2}{S_{xx}}} < \alpha+\beta x_{0i} < \hat{\alpha}+\hat{\beta}x_{0i} + t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n}+\frac{(x_{0i}-\bar{x})^2}{S_{xx}}}$$

simultaneously for $i=1,\ldots,m$.

Simultaneous inference in regression can do better. Since the equation $E[Y|x] = \alpha+\beta x$ holds for all x, we should be make inference for all x. Scheffe derived a solution for this problem.

Theorem 11.9. *Under the conditional normal regression model, the probability is at least* $1-\alpha$ *that*

$$\hat{\alpha}+\hat{\beta}x - M_\alpha S\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}} < \alpha+\beta x < \hat{\alpha}+\hat{\beta}x + F_\alpha S\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$$

simultaneously for all x*, where* $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$.

Proof: We want to show that

$$P\left[\max_x \frac{\left((\hat{\alpha}+\hat{\beta}x)-(\alpha+\beta x)\right)^2}{S^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} \le M_\alpha^2\right] = 1-\alpha.$$

We can parametrize the relevant expression using $\hat{\alpha}+\hat{\beta}x = \bar{Y}+\hat{\beta}t$, $\alpha+\beta x = \mu_{\bar{Y}}+\beta t$, where $t = x-\bar{x}$ and $\mu_{\bar{Y}} = E[\bar{Y}] = \alpha+\beta\bar{x}$. We, hence, want to find M_α to satisfy

$$P\left[\max_t \frac{\left((\bar{Y}-\mu_{\bar{Y}})+(\hat{\beta}-\beta)t\right)^2}{S^2\left(\frac{1}{n}+\frac{t^2}{S_{xx}}\right)} \le M_\alpha^2\right] = 1-\alpha.$$

A bit of calculus shows the term is maximized at $t = \frac{(\hat{\beta}-\beta)}{(\bar{Y}-\mu_{\bar{Y}})}\frac{S_{xx}}{n}$. This can be substituted to transform the term $\frac{n(\bar{Y}-\mu_{\bar{Y}})^2+S_{xx}(\hat{\beta}-\beta)^2}{S^2} = \frac{\frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n}+\frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}}{S/\sigma^2}$. This expression is the quotient of independent chi squared random variables, the denominator being divided by its degrees of freedom. The numerator is the sum of two independent random variables which have $\mathcal{X}_1^2$ distribution. Thus the numerator is distributed as $\mathcal{X}_2^2$. Using the previous theorem on the distribution of S^2/σ^2 we can say $\frac{\frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n}+\frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}}{S/\sigma^2} \sim 2F_{2,n-2}$, proving the requirement for $M_\alpha = \sqrt{2F_{2,n-2}}$. □

This confidence interval covers an entire line with a band. Bonferroni inference, necessarily, pertains to fewer intervals but can be wider whenever $t_{n-2,\alpha/(2m)} > 2F_{2,n-2,\alpha}$, which is always true for large enough m.

To get simultaneous prediction intervals, we can follow a similar layout as before. We need to show that

$$P\left[\max_x \frac{\left((\hat{\alpha}+\hat{\beta}x)-Y\right)^2}{S^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} \le M_\alpha^2\right] = 1-\alpha.$$

This can be maximized using similar tricks like before to give the expression $\frac{\frac{n}{n+1}(\bar{Y}-\mu_{\bar{Y}})^2+S_{xx}(\hat{\beta}-\beta)^2}{S^2} = \frac{\frac{1}{n+1}\frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n}+\frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}}{S/\sigma^2}$. The problem, however, is that the resulting statistic does not have a particularly nice distribution. We can use Satterthwaite approximation using moment matching. We can approximate the numerator as $\mathcal{X}_\nu^2$ where $\nu = 1+2\frac{a^2b^2}{a^2+b^2}$, where $a = \frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n}$, and

$b = \frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}$, resulting in $\frac{\frac{1}{n+1}\frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n}+\frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}}{S/\sigma^2} \sim 2F_{\nu,n-2}$. This can now be used to construct the confidence intervals.

With procedures like the Scheffe band, inferences at x values that are outside the range of the observed xs are unwise. This is because the assumption that the population regression function is linear for all x, does not hold well on extrapolation, as there is no data outside the range to check.

12 Regression Models

12.1 Regression with Errors in Variables

Regression with errors in variables, EIV, also known as the measurement error model, is fundamentally different from the simple linear regression we saw in the previous section. regression with EIV is a generalization of simple linear regression in that we work with the model of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

but now we do not assume that xs are known. Instead, it is a random variable. In the general EIV model we assume that we observe pairs (x_i, y_i) sampled from random variables (X_i, Y_i) whose means satisfy the linear relationship $E[Y_i] = \alpha + [X_i]$, where $E[Y_i] = \eta_i$, and $E[X_i] = \xi_i$. The variables η_i and ξ_i are sometimes called the *latent variables*, because they can't be directly measured. There is no distinction between X and Y, but for regression there should be a reason for choosing Y as the response and X as the predictor. We define the *measurement error model.* Observe independent pairs (X_i, Y_i), $i = 1, \dots, n$ according to

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2), \\ X_i &= \xi_i + \delta_i, \quad \delta_i \sim \mathcal{N}(0, \sigma_\delta^2). \end{aligned}$$

The assumption of normality, although common, is not necessary. If $\delta_i = 0$, then we get the simple linear regression model. If $\alpha = 0$ with possibly $\sigma_\delta^1 \neq \sigma_\epsilon^2$ we get the Behrens-Fisher problem.

12.1.1 Functional and Structural Relationships

Linear functional relationship model: We have a random variable X_i and Y_i, with $E[X_i] = \xi_i$, $E[Y_i] = \eta_i$, and we assume the functional relationship $\eta_i = \alpha | \beta\xi_i$. We observe pairs (X_i, Y_i), $i = 1, \dots, n$ according to $Y_i = \alpha + \beta\xi_i + \epsilon_i$, $\epsilon_i \sim (N)(0, \sigma_\epsilon^2)$, and $X_i = \xi_i + \delta_i$, $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, where the ξ_is are fixed, unknown parameters and the ϵ_is and δ_is are independent. The parameter of interest are α and β, and inference on these parameters is made using the joint distribution of $((X_1, Y_1), \dots, (X_n, Y_n))$, **conditional** on $\xi_1, \dots, \xi_n$.

Linear structural relationship model: This can be thought of as an extent ion of the functional relationship model, through the following hierarchy. As in the functional relationship model, we have random variables X_i and Y_i, with $E[X_i] = \xi_i$ and $E[Y_i] = \eta_i$, and we assume the functional relationship $\eta_i = \alpha + \beta\xi_i$. But now we assume that the parameters $\xi_1, \dots, \xi_n$ are themselves a random sample from a common population $\mathcal{N}(\xi, \sigma_\xi^2)$. Thus, conditional on $\xi_1, \dots, \xi_n$ we observe pairs (X_i, Y_i), $i = 1, \dots, n$ according to $Y_i = \alpha + \beta\xi_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $X_i = \xi_i + \delta_i$, $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, and also $\xi_i \sim$ iid $\mathcal{N}(\xi, \sigma_\xi^2)$. As before, the ϵ_is and δ_is are independent and they are also independent for the ξ_is. The parameters of main interest are α and β. However, the inference on these parameters is made using the joint distribution of $((X_1, Y_1), \dots, (X_n, Y_n))$, **unconditionally** on $\xi_1, \dots, \xi_n$.

Estimators that are consistent in the functional model are also consistent in the structural model. The converse implication is false. However, if a parameter is not identifiable in the structural model, it is also not identifiable in the functional model. It is easier to do statistical theory in the structural model, which the functional model often seems to be more reasonable for many situations.

12.1.2 A Least Squares Solution

In simple linear regression we assume x_is to be fixed and hence it made sense to consider minimization of vertical distances, resulting in *ordinary least squares.* OLS for EIV models give inconsistent coefficients. One way to account for errors in X_is is to do *orthogonal least squares*, that is, find the line that minimizes orthogonal distance, instead of vertical distances. This is called the **method of total least sqaures**. For a particular data point (x', y') the point on a line $y = a+bx$ that is closest is given by $\hat{x}' = \frac{by'+x'-ab}{1+b^2}$ and $\hat{y}' = a+\frac{b}{1+b^2}(by'+x'-ab)$. The squared distance between an observed point (x_i, y_i) and the closest point on the line $y = a + bx$ is $(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2$. The total least square problem is to minimize, over all a and b, the quantity $\sum_{i=1}^n \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2\right]$. Substituting the values and taking derivatives with respect to a and b and equating them to 0 gives, $a = \bar{y} - b\bar{x}$, where

$$b = \frac{1}{2S_{xy}}\left(-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}\right),$$

where $S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$, and $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$. This line always lies between the ordinary regression of y on x and the ordinary regression of x on y. In simple linear regression we saw that, under normality, the ordinary least squares solution to α and β were the same as the MLEs. Here, the orthogonal least squares solution is not equal to the MLE in general.

12.1.3 Maximum Likelihood Estimation

Functional linear relationship model: With the normality assumption, the functional relationship model can be expressed as $Y_i \sim \mathcal{N}(\alpha + \beta\xi_i, \sigma_\epsilon^2)$ and $X_i \sim \mathcal{N}(\xi_i, \sigma_\delta^2)$, $i = 1, \ldots, n$, where X_is and Y_is are independent. Given observations $(\boldsymbol{xy}) = ((x_1, y_1), \ldots, (x_n, y_n))$, the likelihood function is

$$L(\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma_\delta^2, \sigma_\epsilon^2|(\boldsymbol{x}, \boldsymbol{y})) = \frac{1}{(2\pi)^n}\frac{1}{(\sigma_\delta^2\sigma_\epsilon^2)^{n/2}}\exp\left[-\sum_{i=1}^n \frac{(x_i - \xi_i)^2}{2\sigma_\delta^2}\right]\exp\left[-\sum_{i=1}^n \frac{(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\epsilon^2}\right].$$

Since we are free to choose ξ_is we can choose it such that $\xi_i = x_i$ and then let $\sigma_\delta^2 \to 0$. This makes the likelihood goes to infinity. However, if we make the reasonable assumption that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, where $\lambda > 0$ is fixed and known, the problem is alleviated. This is one of the least restrictive assumptions which makes the model well behaved. This simplifies the likelihood to

$$L(\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma_\delta^2|(\boldsymbol{x}, \boldsymbol{y})) = \frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left[-\sum_{i=1}^n \frac{(x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\delta^2}\right].$$

which we can now maximize. We can minimize $\sum_{i=1}^{n}\left[(x_i-\xi_i)^2+\lambda(y_i-(\alpha+\beta\xi_i))^2\right]$ at $\xi_i^*=\frac{x_i+\lambda\beta(y_i-\alpha)}{1+\lambda\beta^2}$ for given α, β, and σ_δ, giving

$$L(\alpha,\beta,\sigma_\delta^2|(\boldsymbol{x},\boldsymbol{y}))=\frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left[-\frac{1}{2\sigma_\delta^2}\left(\frac{\lambda}{1+\lambda\beta^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2\right)\right].$$

To maximize with respect to α and β, we define $\alpha^*=\sqrt{\lambda}\alpha$, $\beta^*=\sqrt{\lambda}\beta$, $y_i^*=\sqrt{\lambda}y_i$, $i=1,\dots,n$. This cane be minimized for $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$ and $\hat{\beta}=\frac{-(S_{xx}-\lambda S_{yy})+\sqrt{(S_{xx}-\lambda S_{yy})^2+4\lambda S_{xy}^2}}{2\lambda S_{xy}}$. This is similar to orthogonal least squares for $\lambda=1$, treating the variance of x and y equally. For $\lambda=0$ we obtain the ordinary least squares case. We now have the likelihood as

$$L(\sigma_\delta^2|(\boldsymbol{x},\boldsymbol{y}))=\frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left[-\frac{1}{2\sigma_\delta^2}\frac{\lambda}{1+\lambda\hat{\beta}^2}\sum_{i=1}^{n}\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)^2\right)\right].$$

The resulting MLE of σ_δ^2 is $\hat{\sigma}_\delta^2=\frac{1}{2n}\frac{\lambda}{1+\lambda\hat{\beta}^2}\sum_{i=1}^{n}(y_i-(\hat{\alpha}+\hat{\beta}x_i))^2$. And from the properties of the MLEs, the MLE of σ_ϵ^2 is $\hat{\sigma}_\epsilon^2=\hat{\sigma}_\delta^2/\lambda$ and $\hat{\xi}_i=\hat{\alpha}+\hat{\beta}x_i$. Here, $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators, $\hat{\sigma}_\delta^2$ is not. More precisely, as $n\to\infty$, $\hat{\sigma}_\delta^2\to\frac{1}{2}\sigma_\delta^2$.

Linear structural relationship model: We observe pairs (X_i,Y_i), $i=1,\dots,n$, according to $Y_i\sim\mathcal{N}(\alpha+\beta\xi_i,\sigma_\epsilon^2)$, $X_i\sim\mathcal{N}(\xi_i,\sigma_\delta^2)$, $\xi_i\sim\mathcal{N}(\xi,\sigma_\xi^2)$, where the ξ_is are independent and, given the ξ_is the X_is and Y_is are independent. If we integrate out ξ_i, we obtain the marginal distribution of $(X_i,Y_i)\sim\mathcal{N}(\xi,\alpha+\beta\xi,\sigma_\delta^2+\sigma_\xi^2,\sigma_\epsilon^2+\beta^2\sigma_\xi^2,\beta\sigma_\xi^2)$. This is similar to RCB ANOVA, where conditional on the blocks, the observations are uncorrelated, but unconditionally, they show positive intra class correlation. The likelihood function is that of bivariate normal and hence the likelihood estimators could be found by equating sample quantities to population quantities. Hence we solve,

$$\begin{aligned}\bar{y}&=\hat{\alpha}+\hat{\beta}\hat{\xi}\\ \bar{x}&=\hat{\xi}\\ \frac{1}{n}S_{yy}&=\hat{\sigma}_\epsilon^2+\hat{\beta}^2\hat{\sigma}_\xi^2\\ \frac{1}{n}S_{xx}&=\hat{\sigma}_\delta^2+\hat{\sigma}_\xi^2\\ \frac{1}{n}S_{xy}&=\hat{\beta}\hat{\sigma}_\xi^2.\end{aligned}$$

Notice that int he functional relationship model we write $Var[X_i]=\sigma_\delta^2$, while in the structural model we write $Var[X_i]=\sigma^2-\delta+\sigma_\xi^2$. These equations can be used to constraint $\hat{\beta}$ as $\frac{|S_{xy}|}{S_{xx}}\le|\hat{\beta}|\le\frac{S_{yy}}{|S_{xy}|}$. With 5 equations and 6 unknowns the system is indeterminate. To make it identifiable we use the same constraint as before $\sigma_\delta^2=\sigma_\epsilon^2$, where λ is known. This leads to the same MLE estimates for $\hat{\alpha}$ and $\hat{\beta}$ but different for the variances given by

$$\hat{\sigma}_\delta^2 = \frac{1}{n}\left(S_{xx} - \frac{S_{xy}}{\beta}\right)$$
$$\hat{\sigma}_\epsilon^2 = \frac{\hat{\sigma}_\delta^2}{\lambda} = \frac{1}{n}(S_{yy} - \hat{\beta}S_{xy})$$
$$\hat{\sigma}_\xi^2 = \frac{1}{n}\frac{S_{xy}}{\hat{\beta}}.$$

Unlike functional relationship model, these estimators are all consistent in the linear structural relationship model.

12.1.4 Confidence Sets

Construction of confidence sets in the EIV model is difficult so we concentrate on β here for structural relationship case of the EIV model. However, the confidence set results are valid in both the structural and functional cases and the formulas remain the same. We continue to assume that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, where λ is known. An approximate confidence interval for β can be constructed by using the fact that the estimator

$$\hat{\sigma}_\beta^2 = \frac{(1+\lambda\hat{\beta}^2)^2(S_{xx}S_{yy} - S_{xy}^2)}{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}$$

is a consistent estimator of σ_β^2, the true variance of $\hat{\beta}$. Hence, using the CLT together with Slutsky's Theorem, we can show that the interval

$$\hat{\beta} - \frac{z_{\alpha/2}\hat{\sigma}_\beta}{\sqrt{n}} \leq \beta \leq \hat{\beta} + \frac{z_{\alpha/2}\hat{\sigma}_\beta}{\sqrt{n}}$$

is an approximate $1-\alpha$ confidence interval for β. But it has finite length and cannot maintain $1-\alpha$ confidence level for all values of parameters, and hence have a confidence coefficient equal to 0 (Gleser and Hwang). A slight modification of this interval gives

$$\hat{\beta} - \frac{t_{n-2,\alpha/2}\hat{\sigma}_\beta}{\sqrt{n-2}} \leq \beta \leq \hat{\beta} + \frac{t_{n-2,\alpha/2}\hat{\sigma}_\beta}{\sqrt{n-2}}$$

again at $1-\alpha$ confidence. This too has finite length, and cannot maintain $1-\alpha$ convergence for all parameter values. This can be quantified using the parameter $\tau^2 = \frac{\sigma_\xi^2}{\sigma_\delta^2}$ as determining the amount of information potentially available in the data to determine the slope β. As $\tau^2 \to 0$, the coverage probability of any finite-length confidence interval on β must also go to 0, since if ξ_is don't vary it would be impossible to fit a unique straight line. They propose, $\tau^2 \geq 0.25$, for high values of τ^2 or n the performance of the test improves.

The Creasy-Williams confidence set is based on the fact that if $\sigma_\delta^2 = \sigma_\epsilon^2$, then $Cov[\beta\lambda Y_i + X_i, Y_i - \beta X_i] = 0$, with exact confidence set which has infinite length. Define $r_\lambda(\beta)$ to be the sample correlation between $\beta_i + X_i$, and $Y_i = \beta X_i$. Since $(\beta\lambda Y_i + X_i, Y_i - \beta X_i)$ is a bivariate normal with correlation 0, it follows that

$$\frac{\sqrt{n-2}r_\lambda(\beta)}{\sqrt{1-r_\lambda^2(\beta)}} \sim t_{n-2}$$

for any value of β. Thus we have identified a pivotal quantity and we conclude that the set

$$\left\{\beta : \frac{(n-2)r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} \leq F_{1,n-2,\alpha}\right\}$$

is a $1-\alpha$ confidence set for β. This set however suffers with defect similar to those of Fieller's intervals with two minima, where the function is 0. The confidence set, hence, consists of disjoint intervals. Furthermore, for every value of β, $-r(\beta) = r_\lambda(-1/(\lambda\beta))$ so that if β is in the confidence set, so is $-1/(\lambda\beta)$. We, thus, can't distinguish β from $-1/(\lambda\beta)$ and this confidence set always contains both positive and negative values, making the determination of sign of slope impossible.

12.2 Logistic Regression

A GLM describes a relationship between the mean of a response variable Y and an independent variable x. But the relationship may be more complicated than the $E[Y_i] = \alpha + \beta x_i$. Logistic regression model is a kind of GLM.

12.2.1 The Model

A GLM consists of three components:

1. The response variables $Y_1, \ldots, Y_n$ are the *random component.* They are assumed to be independent random variables, each with a distribution from a specified exponential family. The Y_is are not identically distributed, but they each have a distribution from the same family.

2. The *systematic component* is the model. It is the function of the predictor variable x_i, linear in the parameters, that is related to the mean of Y_i. So the systematic component could be $\alpha + \beta x_i$ or $\alpha + \beta/x_i$, for example.

3. The *link function* $g(\mu)$ links the two components by asserting that $g(\mu_i) = \alpha + \beta x_i$, for example, where $\mu_i = E[Y_i]$.

The conditional normal regression model is an example of GLM, which has an identity link function $g(\mu) = \mu$. In the logistic regression model, the responses $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \textit{Bernoulli}(\pi_i)$ with $E[Y_i] = \pi_i = P[Y_i = 1]$. We assume that

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i.$$

The left hand side is the log odds of success of Y_i, called *logit*. The Bernoulli pmf can be written in exponential family form as

$$\pi^y(1-\pi)^{1-y} = (1-\pi)\exp\left[y\log\left(\frac{\pi_i}{1-pi_i}\right)\right].$$

The term $\log(\pi/(1-\pi))$ is the natural parameter of this exponential family and the link function $g(\pi) = \log(\pi/(1-\pi))$ is used. When the natural parameter is used in this way, it is called the *canonical link*.

The logit expression can be written as $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\eta x}}$. We see that $0 < \pi(x) < 1$, which is appropriate because $\pi(x)$ is a probability. But, if it is possible that $\pi(x) = 0$ or 1 for some x, then this model is not appropriate. β is the change in the log-odds of success corresponding to a one-unit increase in x. Also,

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\beta}\frac{\pi(x)}{1-\pi(x)},$$

that is, e^{β} is the multiplicative change in the odds of success corresponding to a one unit increase in x. $F(w) = e^w/(1+e^w)$ is the cdf of a logistic(0,1) distribution. In logistic regression we assumed $\pi(x) = F(\alpha+\beta x)$. We can define other models for $\pi(x)$ by using other continuous cdfs. If $F(w)$ is the standard normal cdf, the model is called the probit regression. If a Gumbel cdf is used, the link function is called the log-log link.

12.2.2 Estimation

With $Y_i \sim \mathit{Bernoulli}(\pi_i)$, we no longer have a direct connection between Y_i and $\alpha + \beta x_i$ making least squares unusable. We instead use maximum liklihood to estimate the model. For a general model we have $Y_i \sim \mathit{Bernoulli}(\pi_i)$, where $\pi(x) = F(\alpha+\beta x)$. If we let $F_i = F(+\beta x_i)$ we get the log likelihood as

$$\log L(\alpha,\beta|y) = \sum_{i=1}^{n}\left\{\log(1-F_i) + y_i\log\left(\frac{F_i}{1-F_i}\right)\right\}.$$

We let $dF(w)/dw = f(w)$, and let $f_i = f(\alpha+\beta x_i)$. By differentiating with respects to the parameters α and β we get

$$\frac{\partial}{\partial\alpha}\log L(\alpha,\beta|y) = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)} \quad\text{and}\quad \frac{\partial}{\partial\beta}\log L(\alpha,\beta|y) = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)}x_i.$$

For logistic regression $F(w) = e^w/(1+e^w)$ and $f_i/(F_i(1-Fi)) = 1$. These differentials are set to 0 to obtain $\hat{\alpha}$ and $\hat{\beta}$, which are solved numerically. For logistic and probit regression, the log likelihood is strictly concave, hence there is a unique solution.

In situations where there are multiple Bernoulli observations at each value of x we use the binomial distribution. Suppose there are J different values of the predictor x in the data set $x_1,\ldots,x_J$. Let n_j denote the number of Bernoulli observations at x_j, and let Y_j^* denote the number of successes in these n_j observations. Thus, $Y_j^* \sim binomial(n_j, \pi(x_j))$. Then the likelihood is

$$L(\alpha, \beta | y^*) = \prod_{j=1}^{J} F_j^{y_j^*}(1 - F_j)^{n_j - y_j^*},$$

and the likelihood equations are

$$\sum_{j=1}^{J}(y_j^* - n_j F_j)\frac{f_j}{F_j(1 - F_j)} = 0 \quad \text{and} \quad \sum_{j=1}^{J}(y_j^* - n_j F_j)\frac{f_j}{F_j(1 - F_j)}x_j = 0.$$

We can also approximate variances using MLE asymptotics. The information matrix can be calculated as

$$I(\theta_1, \theta_2) = \begin{bmatrix} -\frac{\partial^2}{\partial\theta_1^2}\log L & -\frac{\partial^2}{\partial\theta_1\partial\theta_2}\log L \\ -\frac{\partial^2}{\partial\theta_1\partial\theta_2}\log L & -\frac{\partial^2}{\partial\theta_2^2}\log L \end{bmatrix}.$$

For logistic regression, the information matrix is given by

$$I(\alpha, \beta) = \begin{bmatrix} \sum_{j=1}^{J} n_j F_j(1 - F_j) & \sum_{j=1}^{J} x_j n_j F_j(1 - F_j) \\ \sum_{j=1}^{J} x_j n_j F_j(1 - F_j) & \sum_{j=1}^{J} x_j^2 n_j F_j(1 - F_j) \end{bmatrix}.$$

We then use the approximation $Var[h(\hat{\theta})|\theta] \approx I(\hat{\boldsymbol{\theta}})^{-1}\left[h'(\boldsymbol{\theta})^T h'(\boldsymbol{\theta})\right]$ to obtain variance of a general function of the parameters. The estimates of the variances $[se(\hat{\alpha})]^2$ and $[se(\hat{\beta})]^2$ are diagonal elements of the inverse $I(\hat{\alpha}, \hat{\beta})$.

It is most common to test the hypothesis $H_0 : \beta = 0$, to test no relationship between predictor and response variables. The Wald test statistic, $X = \hat{\beta}/se(\hat{\beta})$, has approximately a standard normal distribution if H_0 is true and the sample size is large. Thus, H_0 can be rejected if $|Z| > z_{\alpha/2}$. Alternatively, H_0 can be tested with the log LRT statistic $-2\log\lambda(y^*) = 2[\log L(\hat{\alpha}, \hat{\beta}|y^*) - L(\hat{\alpha}_0, 0|y^*)]$, where $\hat{\alpha}_0$ is the MLE of α assuming $\beta = 0$. With standard binomial arguments we can show that $\hat{\alpha}_0 = \sum_{i=1}^{n} y_i/n = \sum_{j=1}^{J} y_j^* / \sum_{j=1}^{J} n_j$. Therefore, under H_0, $-2\log\lambda$ has an approximate $\mathcal{X}_1^2$ distribution, and we can reject H_0 at level α if $-2log\lambda \geq \mathcal{X}_{1,\alpha}^2$.

12.3 Robust Regression

When observing $x_1, \ldots, x_n$, we can define the mean and the median as minimizers of the following quantities:

$$\text{mean} : \min_m \left\{ \sum_{i=1}^{n}(x_i - m)^2 \right\} \quad \text{median} : \min_m \left\{ \sum_{i=1}^{n} |x_i - m| \right\}.$$

The least square satisfies

$$\min_{a,b} \left\{ {}_{i=1}^{n} \; (y_i - (a +_i))^2 \right\},$$

and, analogously, we define the least absolute deviation, LAD, regression estimate by

$$\min_{a,b} \left\{ \sum_{i=1}^{n} \left\| y_i - (a + bx_i) \right\| \right\}.$$

The LAD estimate may not be unique. We know that the least squares estimator b with variance $\sigma^2/\sum(x_i-\bar{x})^2$ is the BLUE of β. To investigate how b performs for small deviation we assume that

$$Var[\epsilon_i] = \begin{cases} \sigma^2 & \text{with probability } 1-\delta \\ \tau^2 & \text{with probability } \delta \end{cases}$$

Writing $b = \sum d_i Y_i$, with $d_i = (x_i - \bar{x})/\sum(x_i-\bar{x})^2$, we now have

$$Var[b] = \sum_{i=1}^{n} d_i^2 Var[\epsilon_i] = \frac{(1-\delta)\sigma^2 + \delta\tau^2}{\sum(x_i-\bar{x})^2}.$$

This shows that, as with sample mean, for small perturbations b performs pretty well. The behavior of least squares intercept a is similar. An introduction of catastrophic observation has much less effect on LAD then on OLS. This is similar to the breakdown values of 0% for mean and 50% for median. However, this comes at the cost of losing efficiency.

Example 12.1. We consider the simplified model $Y_i = \beta x_i + \epsilon_i$. The LAD estimator is obtained by minimizing $\sum_{i=1}^n \rho(y_i - \beta x_i) = \sum_{i=1}^n |y_i - \beta x_i| = \sum_{i=1}^n (y_i \beta x_i)\mathbf{1}_{y_i>\beta x_i} - (y_i - \beta x_i)\mathbf{1}_{y_i<\beta x_i}$. We calculate $\psi = \rho'$ and solve $\sum_i \psi(y_i - \beta x_i) = 0$ for β, where $\psi(y_i - \beta x_i) = x_i \mathbf{1}_{\_y_i>\beta x_i} - x_i\mathbf{1}_{y_i<\beta x_i}$. If $\hat{\beta}_L$ is the solution, we expand ϕ in a Taylor series around β:

$$\sum_{i=1}^{n} \psi(y_i - \hat{\beta}_L x_i) = \sum_{i=1}^{n} \psi(y_i - \beta x_i) + (\hat{\beta}_L - \beta)\frac{d}{d\hat{\beta}_L}\sum_{i=1}^{n}\psi(y_i - \hat{\beta}_L x_i)\Big|_{\hat{\beta}_L=\beta} + \dots$$

We assume that the left hand side approaches 0 as $n \to \infty$. Rearranging we get

$$\sqrt{n}(\hat{\beta}_L - \beta) = \frac{-\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(y_i - \beta x_i)}{\frac{1}{n}\frac{d}{d\hat{\beta}_L}\sum_{i=1}^n \psi(y_i - \hat{\beta}_L x_i)\Big|_{\hat{\beta}_L=\beta}}$$

for the numerator we have $E_\beta[\psi(Y_i - \hat{\beta}x_i)] = 0$ and $Var[\psi(Y_i - \hat{\beta}_L x_i)] = x_i^2$. It follows that the numerator $\to \mathcal{N}(0, \sigma_x^2)$, where $\sigma_x^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n x_i^2$. The denominator we approach as follows

$$\begin{aligned}\frac{1}{n}\frac{d}{d\beta_0}\sum_{i=1}^{n}\psi(y_i - \beta_0 x_i) &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\beta_0}E_\beta[\psi(Y_i - \beta_0 x_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\beta_0}[x_i P[Y_i > \beta_0 x_i] - x_i P[Y_i < \beta_0 x_i]] \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 f(\beta_0 x_i - \beta x_i) + x_i^2 f(\beta_0 x_i - \beta x_i).\end{aligned}$$

Evaluating it at $\beta_0 = \beta$ we get $\approx 2f(0)\frac{1}{n}\sum x_i^2$. Putting this together we get

$$\sqrt{n}(\hat{\beta}_L - \beta) \to \mathcal{N}(0, \frac{1}{4f(0)^2\sigma_x^2})$$

Finally, for $\alpha = 0$ the OLS estimator $\hat{\beta} = \sum x_i y_i / \sum x_i^2$ satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}(0, \frac{1}{\sigma_x^2})$$

So that the asymptotic relative efficiency of $\hat{\beta}_L$ to $\hat{\beta}$ is

$$ARE(\hat{\beta}_L, \hat{\beta}) = \frac{1/\sigma_x^2}{1/(4f(0)^2\sigma_x^2)} = 4f(0)^2,$$

which generally shows that the LAD estimator gives up a good bit of efficiency with respect to least squares.

□

The LAD alternative to OLS seem to loose too much in efficiency if the errors are truly normal. The compromise, once again, is an M-estimator. We can construct one by minimizing a function $\sum_i \rho_i(\alpha, \beta)$, where

$$\rho_i(\alpha, \beta) = \begin{cases} \frac{1}{2}(y_i - \alpha - \beta x_i)^2 & \text{if } |y_i - \alpha - \beta x_i| \leq k \\ k|y_i - \alpha - \beta x_i| - \frac{1}{2}k^2 & \text{if } |y_i - \alpha - \beta x_i| > k, \end{cases}$$

where k is a tuning parameter. The M-estimator turns out to be somewhat more resistant than the least squares lines, behaving more like the LAD fit when there are outliers. We expect the ARE of the M-estimator to be better than that of the LAD, which is the case, albeit the calculation is messy.