Review on Probability and Statistics

Primus Ji

May 11, 2020

Contents

Chapter 1

Overview and Descriptive Statistics

1.1 Populations and Samples

1.1.1 Facts and Concepts

data collections of facts

population collections of objects

census collecting desired information for all objects in the population

sample subsets of the population

variable any characteristic whose value may change from one object to another in the population

univariate dataset observations on a single variable

bivariate dataset observations on two variables

multivariate dataset observations on more than two variables

1.2 Pictorial and Tabular Methods in Descriptive Statistics

1.2.1 Facts and Concepts

number of classes in histogram $\approx \sqrt{\text{number of observations}}$

density $\dfrac{\text{relative frequency of the class}}{\text{class width}}$

relative frequency $\dfrac{\text{number of times the value appers}}{\text{number of observations in the dataset}}$

unimodal histogram one that rises to a single peak and then declines

bimodal histogram one that has two different peaks

multimodal histogram one that has more than two peaks

symmetric left half of graph is a mirror image of the right half

positively skewed the right or upper tail is stretched out compared with the left or lower tail

negatively skewed the left or lower tail is stretched out compared with the right or upper tail

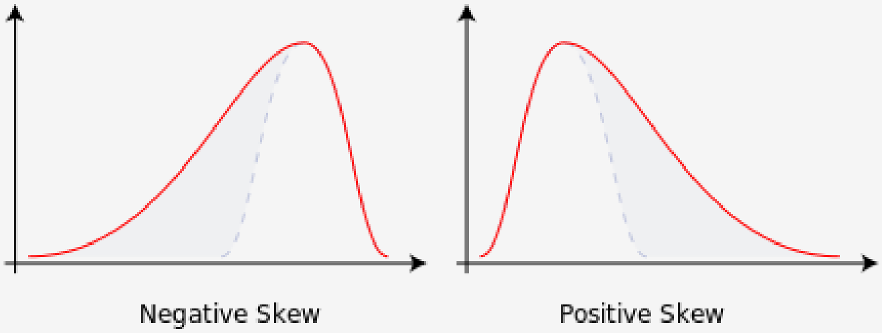

Figure 1.1: Negative and Positive Skew

1.3 Measures of Location

1.3.1 Facts and Concepts

sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

sample median

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

population mean and population median μ and $\tilde{\mu}$
Negative skew if $\mu < \tilde{\mu}$;
Symmetric if $\mu = \tilde{\mu}$;
Positive skew if $\mu > \tilde{\mu}$;

n% trimmed mean average of samples without the smallest $n\%$ and the largest $n\%$

n% percentile samples without the highest $n\%$

1.4 Measures of Variability

1.4.1 Facts and Concepts

sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

An alternative expression for the numerator of s^2 is

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Proof. Because

$$\bar{x} = \frac{\sum x_i}{n},\ n\bar{x}^2 = \frac{n\left(\sum x_i\right)^2}{n^2} = \frac{\left(\sum x_i\right)^2}{n}$$

then

$$\begin{aligned}\sum (x_i - \bar{x})^2 &= \sum \left(x_i^2 - 2\bar{x} \cdot x + \bar{x}^2\right) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum \left(\bar{x}^2\right) \\ &= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n\left(\bar{x}\right)^2 = \sum x_i^2 - n\left(\bar{x}\right)^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\end{aligned}$$ □

---

sample standard deviation

$$s = \sqrt{s^2}$$

population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

population standard deviation

$$\sigma = \sqrt{\sigma^2}$$

fourth spread

$$f_s = \text{upper fourth} - \text{lower fourth}$$

where observations are ordered and separated in half(median in both halves if odd number) and upper fourth is median in the larger half, lower fourth is median in the smaller half

---

Any observation farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.

Chapter 2

Probability

2.1 Sample Spaces and Events

2.1.1 Facts and Concepts

probability randomness and uncertainty

experiment any action or process whose outcome is subject to uncertainty

sample space $\mathcal{S}$ the set of all possible outcomes of that experiment

event any collection(subset) of outcomes contained in the sample space $\mathcal{S}$

An **event** is said to be **simple** if it consists of exactly one outcome and **compound** if it consists of more than one outcome.

---

union $A \cup B$ the event consisting of all outcomes that are either in A or in B or in both events

intersection $A \cap B$ the event consisting of all outcomes that are in both A and B

complement A' the set of all outcomes in $\mathcal{S}$ tha are not contained in A

disjoint, mutually exclusive $A \cap B = \varnothing$ the relationship of having no outcomes in common

2.2 Axioms, Interpretations, and Properties of Probability

AXIOM 2.2.1. For any event A, $P(A) \geq 0$

AXIOM 2.2.2. $P(\mathcal{S}) = 1$

AXIOM 2.2.3. If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, then

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

PROPOSITION 2.2.1. $P(\varnothing) = 0$ where $\varnothing$ is the null event. This in turn implies that the property contained in Axiom 3 is valid for a *finite* collection of events.

PROPOSITION 2.2.2. For any event A, $P(A) = 1 - P(A')$

PROPOSITION 2.2.3. For any event A, $P(A) \leq 1$

PROPOSITION 2.2.4. For any evnets A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2.3 Counting Techniques

DEFINITION 2.3.1. Any ordered sequence of k objects taken from aset of n distinct objects is called a permutation of size k of the objects. The number of permutations of size k that can be constructed from the n objects is denoted by $P_{k,n}$.

$$P_{k,n} = n(n-1)(n-2) \cdot \cdots \cdot (n-k+2)(n-k+1)$$

DEFINITION 2.3.2. For any positive integer m, $m!$ is read "m factorial" and is defined by $m! = m(m-1) \cdot \cdots \cdot (2)(1)$. Also, $0! = 1$.

DEFINITION 2.3.3. Given a set of n distinct objects, any unordered subset of size k of the objects is called a **combination**. The number of combinations of size k that can be formed from n distinct objects will be denoted by $\binom{n}{k}$.(This notation is more common in probability than $C_{k,n}$ which would be analogous to notation for permutations.)

2.4 Conditional Probability

DEFINITION 2.4.1. For any two events A and B with $P(B) > 0$, the **conditional probability of A given that B has occurred** is defined by

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

.

THEOREM 2.4.1 (THE MULTIPLICATION RULE). $P(A \cap B) = P(A|B) \cdot P(B)$

THEOREM 2.4.2 (THE LAW OF TOTAL PROBABILITY). Let $A_1, A_2, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$\begin{aligned} P(B) &= P(B|A_1) \cdot P(A_1) + \cdots + P(B|A_k) \cdot P(A_k) \\ &= \sum_{i=1}^{k} P(B|A_i)P(A_i) \end{aligned}$$

THEOREM 2.4.3 (BAYERS' THEOREM)**.** Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots, k$. Then for any other event B, for which $P(B) > 0$

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum\limits_{i=1}^{k} P(B|A_i)P(A_i)} \quad j = 1, \ldots, k$$

2.5 Independence

DEFINITION 2.5.1. Two events A and B are **independent** if $P(A|B) = P(A)$ and are **dependent** otherwise.

PROPOSITION 2.5.1. A and B are independent if and only If

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

Chapter 3

Discrete Random Variables and Probability Distributions

3.1 Random Variables

DEFINITION 3.1.1. For a given sample space $\mathcal{S}$ of some experiment, a **random variable rv** is any rule that associates a number with each outcome in $\mathcal{S}$. In mathematical language, a random variable is a function whose domain si the sample space and whose range is the set of real numbers.

DEFINITION 3.1.2. Any random variable whose only possible values are 0 and 1 is called a **Bernoulli random variable**.

DEFINITION 3.1.3. A **discrete** random variable is an rv whose possible values either constitute a finite set or else can be listed in an inifinite sequence in which there is a first element, a second element, and so on.

A random variable is **continuous** if *both* of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line(possibly inifinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals(e.g., $[0, 10] \cup [20, 30]$).

2. No possible value of the variable has possitive probability, that is , $P(X = c) = 0$ for any possible value c.

3.2 Probability Distributions for Discrete Random Variables

DEFINITION 3.2.1. The **probability distribution** or **probability mass function**(pmf) of a discrete rv is defined for every number x by $p(x) = P(X = x) = P(\text{all } s \in \mathcal{S} : X(s) = x)$.

DEFINITION 3.2.2. Suppose $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each differnet value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

DEFINITION 3.2.3. The **cumulative distribution function**(cdf) $F(x)$ of a discrete rv X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

PROPOSITION 3.2.1. For any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = F(b) - F(a-)$$

where $F(a-)$ represents the maximum of $F(x)$ values to the left of a. Equivalently, if a is the limit of values of x approaching from the left, then $F(a-)$ is the limiting value of $F(x)$. In particular, if the only possible values are integers and if a and b are integers, then

$$\begin{aligned} P(a \leq X \leq b) &= P(X = a \text{ or } a+1 \text{ or} \ldots \text{or } b) \\ &= F(b) - F(a-1) \end{aligned}$$

Taking $a = b$ yields $P(X = a) = F(a) - F(a-1)$ in this case.

3.3 Expected Values of Discrete Random Variables

3.3.1 The Expected Value of X

DEFINITION 3.3.1. Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X, is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

This expected value will exist provided that $\sum_{x \in D} |x| \cdot p(x) < \infty$.

3.3.2 The Expected Value of a Function

PROPOSITION 3.3.1. If the rv X has a set of possible values D and pmf $p(x)$, then the expected values of any function $h(X)$, denoted by $E[h(X)])$ of $\mu_{h(X)}$, is computed by

$$E[H(X)] = \sum_{D} h(x) \cdot p(x)$$

assuming that $\sum_{D} |h(x)| \cdot p(x)$ is finite.

PROPOSITION 3.3.2.

$$E(aX + b) = a \cdot E(X) + b$$

(Or, using alternative notation, $\mu_{aX+b} = a \cdot \mu_X + b$.)

3.3.3 The Variance of X

DEFINITION 3.3.2. Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E(X - \mu)^2$$

The **standard deviation**(SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

3.3.4 A Shortcut Formula for σ^2

PROPOSITION 3.3.3.

$$V(X) = \sigma^2 = \left[\sum_{D} x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

3.3.5 Rules of Variance

PROPOSITION 3.3.4.

$$V(aX + b) = \sigma_{aX+b}^2 = a^2 \cdot \sigma_X^2 \textit{ and } \sigma_{aX+b} = |a| \cdot \sigma_X$$

3.4 Moments and Moment Generating Functions

DEFINITION 3.4.1. moments: the expected values of integer powers of X and $X - \mu$.

DEFINITION 3.4.2. measure of departure of symmetry is called **skewness**, derived from third moment about the mean divided by σ^3 to gain scale independence.

$$\frac{E\left[(X-\mu)^3\right]}{\sigma^3} = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

a distribution is negatively skewed if skewness is negative and is positively skewed otherwise.

DEFINITION 3.4.3. The **moment generating function**(mgf) of a discrete rv X is defined to be

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tX} p(x)$$

where D is the set of possible X values. We will say that the moment generating function exists if $M_X(t)$ is defined for an interval of number sthat includes zeor as well as positive and negative values of t (an interval including 0 in its interior).

PROPOSITION 3.4.1. If the mgf exists and is the same for two distributions, then the two distributions are the same. That is, the moment generating function uniquely specified the probability distribution; there is a one-to-one correspondence between distributions and mgf's.

THEOREM 3.4.1. If the mgf exists,

$$E(X^r) = M_X^{(r)}(0)$$

e.g., $M'_X(0) = E(X),\ M''_X(0) = E(X^2)$

PROPOSITION 3.4.2. Let X have the mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{bt} M_X(at)$.

Example 3.34 on page 127 in textbook gives excellent illustation on how linear replacement works in real situations.

3.5 The Binomial Probability Distribution

DEFINITION 3.5.1. An experiment for which Conditions 1-4 is satisfied is called **a binomial experiment**.

1. The experiment consists of a sequence of n smaller experiments called *trials*, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes(dichotomous trials), which we denote by succes(S) or failure(F).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success is constant from trial to trial; we denote this probability by p.

PROPOSITION 3.5.1. Consider sampling without replacement from a dichotomous populatin of size N. If the sample size(number of trials) n is at most 5% of the population size, the experiment can be analyzed as though it were exactly a binomial experiment.

3.5.1 The Binomial Random Variable and Distribution

DEFINITION 3.5.2. Given a binomial experiment consisting of n trials, the **binomial random variable** X accosiated with this experiment is defined as

$$X = \text{the number of } S\text{'s among the } n \text{ trials}$$

NOTATION 3.5.1. We often write $X \sim Bin(n,p)$ to indicate that X is a binomial rb based on n trials with success probability p.

Bacause the pmf of a binomial rb X depends on the two parameters n and p, we denote the pmf by $b(x;n,p)$.

THEOREM 3.5.1.

$$b(x;n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,2,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

3.5.2 Using Binomial Tables

NOTATION 3.5.2. For $X \sim Bin(n,p)$, the cdf will be denoted bu

$$P(X \leq x) = B(x;n,p) = \sum_{y=0}^{x} b(y;n,p) \quad x = 0,1,\ldots,n$$

3.5.3 The Mean and Varaince of X

PROPOSITION 3.5.2. If $X \sim Bin(n,p)$, then $E(X) = np$, $V(X) = np(1-p) = npq$, and $\sigma_X = \sqrt{npq}$(where $q = 1-p$).

3.5.4 The Moment Gnerating Function of X

Obtain mean and variance of binomial rv X.

Proof.

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x \in D} e^{tx} p(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n \end{aligned}$$

Differentiate $M_X(t)$,

$$M'_X(t) = n(pe^t + 1 - p)pe^t$$

$$\mu = M'_X(0) = np$$

$$M''_X(t) = n(n-1)(pe^t + 1 - p)^{n-2} pe^t pe^t + n(pe^t + 1 - p)^{n-1} pe^t$$

$$E(X^2) = M''_X(0) = n(n-1)p^2 + np$$

$$\begin{aligned} \sigma^2 &= V(X) = E(X^2) - [E(X)]^2 \\ &= n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p) \end{aligned}$$

□

3.6 Hypergeometric and Negative Binomial Distributions

3.6.1 The Hypergeometric Distribution

PROPOSITION 3.6.1. If X is the number of S's in a completely random sample of size n drawn from a populatin consisting of M S's and $(N - M)$ F's, then the probability distribution of X, called the **hypergeometric distribution**, is given by

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for x an integer satisfying $\max(0, n - N + M) \leq x \leq \min(n, M)$.

PROPOSITION 3.6.2.
The mean and variance of the hypergeometric rv X having pmf $h(x; n, M, N)$ are

$$E(X) = n \cdot \frac{M}{N} \quad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)$$

Finite populatin correction factor $\dfrac{N-n}{N-1}$ is the difference of $V(X)$ of binomial rv and hypergeometric rv, thus hypergeometric rv has smaller variance than binomial rv.

THEOREM 3.6.1. RULE OF THUMB Let the population size , N, and number of population S's, M, get large with the ratio M/N approaching p. Then $h(x; n, M, N)$ approaches $b(x; n, p)$; so for n/N small, the two are approximately equal provided that p is not too near either 0 or 1.

3.6.2 The Negative Binomial Distribution

PROPOSITION 3.6.3. Negative binomial distribution is based on the following Rules

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success(S) or a failure(F).
3. The probability of success is constant from trial to trial, so $P(S \text{ bon trial } i) = p$ for $i = 1, 2, \ldots$.
4. The experiment continuess(trials are performed) until a total of r successes has been observed, where r is a specified positive integer.

PROPOSITION 3.6.4. The pmf of the negative binomial rv X with parameters r = numbers of S's and $p = P(S)$ is

$$nb(x; r, p) = \binom{x + r - 1}{r - 1} p^r (1-p)^x \quad x = 0, 1, 2, \ldots$$

x denotes the number of failures that precede the rth success.

PROPOSITION 3.6.5. If X is a negative binomial rv with pmf $nb(x; r, p)$, then

$$M_x(t) = \frac{p^r}{[1 - e^t(1-p)]^r} \quad E(X) = \frac{r(1-p)}{p} \quad V(X) = \frac{r(1-p)}{p^2}$$

PROOF NOT ATTACHED

3.7 The Poisson Probability Distribution

DEFINITION 3.7.1. A random varaible X is said to have a **Poison distribution** with parameter $\lambda(\lambda > 0)$ if the pmf of X is

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

3.7.1 The Poisson Distribution as a Limit

PROPOSITION 3.7.1. Suppose that in the binomial pmf b(x;n,p) we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\lambda > 0$. Then $b(x; n, p) \to p(x; \lambda)$.

Proof.

$$\begin{aligned} b(x;n,p) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{n \cdot (n-1) \cdot \dots \cdot (n-x+1)}{x!} p^x (1-p)^{n-x} \\ &= \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-x+1}{n} \cdot \frac{(np)^x}{x!} \cdot \frac{(1-p)^n}{(1-p)^x} \end{aligned}$$

$$\lim_{n\to\infty} b(x;n,p) = 1 \cdot 1 \cdot \dots \cdot 1 \cdot \frac{\lambda^x}{x!} \cdot \left(\lim_{n\to\infty} \frac{(1-np/n)^n}{1} \right)$$

$$\begin{aligned} &\because np \to \lambda, \lim_{a_n \to a} \left(1 - \frac{a_n}{n}\right)^n = e^{-a}, \\ &\therefore \lim_{n\to\infty} b(x;n,p) = \frac{\lambda^x}{x!} \cdot \lim_{n\to\infty} \left(1 - \frac{np}{n}\right)^n = \frac{\lambda^x e^{-\lambda}}{x!} = p(x;\lambda) \end{aligned}$$

□

THEOREM 3.7.1. RULE OF THUMB In any binomial experiment for which n is large and p is small, $b(x; n, p) \approx p(x; \lambda)$ where $\lambda = np$. Safe for $n > 50$ **and** $np < 5$.

3.7.2 The Mean, Variance and MGF of X

PROPOSITION 3.7.2. If X has a Poisson distribution with parameter λ, then $E(X) = V(X) = \lambda$.

PROPOSITION 3.7.3. The Poisson moment generating function is

$$M_X(t) = e^{\lambda(e^t-1)}$$

Proof.

$$M_X(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

$$\because \sum_{x=0}^{\infty} \frac{u^x}{x!} = e^u,$$

$$\therefore M_X(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda e^t - \lambda}$$

□

3.7.3 The Poisson Process

PROPOSITION 3.7.4. A **Poisson Process** is based on the following situations:

1. There exists a parameter $\alpha > 0$ such that for any short time interval of length Δt, the probability that the event occurs once is $\alpha \cdot \Delta t + o(\Delta t)$.

2. The probability of event occurring more than once during Δt is $o(\Delta t)$[which, along with Assumption 1, implies that the probability of no event occurs during Δt is $1 - \alpha \cdot \Delta t - o(\Delta t)$].

3. The number of pulses received during the time interval Δt is independent of the number received prior to this time interval.

Let $P_K(t)$ denote the probability that event occurs k times during any particular time interval of length t.

$P_K(t) = \dfrac{e^{-\alpha t}(\alpha t)^k}{k!}$, implies that the times of event occurring in a time interval of length t is a Poisson rv with $\lambda = \alpha t$.

The expected occurring time during a unit time interval is α.

Chapter 4

Continuous Random Variables and Probability Distribution

4.1 Probability Density Functions and Cumulative Distribution Functions

4.1.1 Probability Distributions for Continuous Variables

DEFINITION 4.1.1. Let X be a continuous rv. Then a **probability distribution** or **probability density function**(pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval[a, b] is the area above this interval and under the graph of the density function. The graph of $f(x)$ is often referred to as the *density curve.*

A legitimate pdf $f(x)$ must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx =$ [area under the entire graph of $f(x)] = 1$

DEFINITION 4.1.2. A continuous rv X is said to have a **uniform distribution** on the interval $[A, B]$ if the pdf of X is

$$f(x; A, B) = \begin{cases} \dfrac{1}{B-A} & A \leq X \leq B \\ 0 & \text{otherwise} \end{cases}$$

4.1.2 The Cumulative Distribution Function

DEFINITION 4.1.3. The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(y)dy$$

For each x, $F(x)$ is the area under the density curve to the left of x.

4.1.3 Using $F(x)$ to Compute Probabilities

PROPOSITION 4.1.1. Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

4.1.4 Obtaining $f(x)$ from $F(x)$

PROPOSITION 4.1.2. If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$.

4.1.5 Percentiles of a Continuous Distribution

DEFINITION 4.1.4. Let p be a number between 0 and 1, The $(100p)$th percentile of the distribution of a continuous rv X, denoted by $\eta(p)$, is defiend by

$$p = F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(y)dy$$

DEFINITION 4.1.5. The **median** of a continuous distribution, denoted by $\tilde{\mu}$, is the 50th percentile, so $\tilde{\mu}$ satisfies $.5 = F(\tilde{\mu})$. That is, half the area under the density curve is to the left of $\tilde{\mu}$ and half os tp the right of $\tilde{\mu}$.

4.2 Expected Values and Moment Generating Functions

4.2.1 Expected Values

DEFINITION 4.2.1. The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$.

PROPOSITION 4.2.1. If X is a continuous rv with pdf $f(x)$ and $h(X)$ is any function of X, then

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |h(x)|f(x)dx < \infty$.

4.2.2 The Variance and Standard Deviation

DEFINITION 4.2.2. The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

The **standard deviation**(SD) of X is $\sigma_X = \sqrt{V(X)}$.
Also,

$$V(X) = E(X^2) - [E(X)]^2$$

4.2.3 Approximating the Mean Value and Standard Deviation

PROPOSITION 4.2.2. Suppose $h(x)$ is differentiable and that its derivative evaluated at μ satisfies $h'(\mu) \neq 0$. Then if the variance of X is small, which means the distribution of X is largely concentrated on an interval of values close to μ, the mean value and variance of $Y = h(X)$ can be approximated as follows:

$$E[h(X)] \approx h(\mu), \quad V[h(X)] \approx [h'(\mu)]^2\sigma^2$$

4.2.4 Moment Generating Functions

DEFINITION 4.2.3. The **moment generating function**(mgf) of a continuous random variable X is

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f(x)dx.$$

As in the discrete case, we will say that the moment generating function exists if $M_X(t)$ is defined for an interval of numbers that includes zero in tis interior, which mean that it includes both positive and negative values ot t.

Just as before, when $t = 0$ the value of the mgf is always 1:

$$M_X(0) = E(e^{0X}) = \int_{-\infty}^{\infty} e^{0x} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

PROPOSITION 4.2.3. Two continuous distributions have the same pdf if and only if they have the same moment generating function, assuming that the mgf exists.

PROPOSITION 4.2.4. For continuous rv X, if mgf exists,

$$E(X^r) = M_X^{(r)}(0)$$

A faster version of deriving mean and variance of X if mgf exists is

$$\begin{aligned}\mu = E(X) = R'_X(0)\\ \sigma^2 = V(X) = R''_X(0)\end{aligned}$$

where $R_X(t) = \ln[M_X(t)]$.

PROPOSITION 4.2.5. For continuous rv X and linear function $Y = aX + b$, if mgf $M_X(t)$ exists, then

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

4.3 The Normal Distribution

DEFINITION 4.3.1. A continuous rv X is said to have a **normal distribution** with parameters μ and σ(or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

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

4.3.1 The Standard Normal Distribution

DEFINITION 4.3.2. The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**. A random variable that has a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. The pdf of Z is

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

The cdf of Z is $P(Z \leq z) = \displaystyle\int_{-\infty}^{z} f(y,0,1)dy$, which we will denote by $\Phi(z)$.

4.3.2 Percentiles of the Standard Normal Distribution

NOTATION 4.3.1. z_α will denote the value on the measurement axis for which α of the area under the z curve lies to the right of z_α.

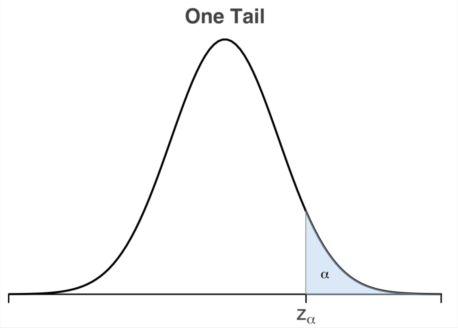

Figure 4.1: One-tail z critical value

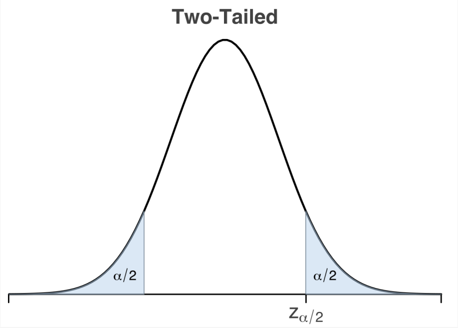

Figure 4.2: Two-tail z critical value

4.3.3 Nonstandard Normal Distribution

PROPOSITION 4.3.1. If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X-\mu}{\sigma}$$

has a standard normal distribution. Thus

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \quad P(X \geq b) = 1 - \Phi(\frac{b-\mu}{\sigma})$$

PROPOSITION 4.3.2. THE EMPIRICAL RULE If the population distribution of a variable is (approximately) nomal, then

1. Roughly 68% of the values are within 1 SD of the mean.
2. Roughly 95% of the values are within 2 SD of the mean.
3. Roughly 99.7% of the values are within 3 SD of the mean.

PROPOSITION 4.3.3.
(100p)th percentile for normal(μ, σ) = μ + [(100p)th percentile for standard normal]$\cdot\sigma$

4.3.4 Approximating the Binomial Distribution

PROPOSITION 4.3.4. Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. In particular, for $x = a$ possible value of X,

$$\begin{aligned} P(x \leq X) = B(x; n, p) &\approx \text{(area under the normal curve to the left of } x + .5) \\ &= \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right) \end{aligned}$$

In practice, the approximation in adequate provided that both $np \geq 10$ and $nq \geq 10$.

4.3.5 The Normal Moment Generating Function

PROPOSITION 4.3.5. The moment generating function of a normally distributed rv X is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

Proof. See textbook Page 191. □

4.4 The Gamma Distribution and Its Relatives

DEFINITION 4.4.1. For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

Properties of the gamma function are the following:

1. For any $\alpha > 0$, $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$(via integration by parts)
2. For any positive integer n, $\Gamma(n) = (n-1)!$
3. $\Gamma(1/2) = \sqrt{\pi}$

4.4.1 The Family of Gamma Distributions

DEFINITION 4.4.2. A continuous rv X is said to have a **gamma distribution** if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the paramters α and β satisfy $\alpha > 0$, $\beta > 0$. The **standard gamma distribution** has $\beta = 1$, so the pdf of a standard gamma rv is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

PROPOSITION 4.4.1. The moment generating function of a gamma rv is

$$M_X(t) = \frac{1}{(1-\beta t)^\alpha}$$

Proof. See textbook Page 196. □

PROPOSITION 4.4.2. The mean and variance of a rv X having the gamma distribution $f(x;\alpha,\beta)$ are

$$E(X) = \mu = \alpha\beta \quad V(X) = \sigma^2 = \alpha\beta^2$$

PROPOSITION 4.4.3. cdf of standard gamma rv X is

$$F(x;\alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}dy \quad x > 0$$

which is also denoted as **incomplete gamma function**.

PROPOSITION 4.4.4. Let X have a gamma distribution with parameters α and β. Then for any $x > 0$, the cdf of X is given by

$$F(X \leq x) = F(x;\alpha,\beta) = F(\frac{x}{\beta};\alpha)$$

the incomplete gamma function evaluated at x/β.

4.4.2 The Exponential Distribution

DEFINITION 4.4.3. X is said to have an **exponential distribution** with parameter $\lambda(\lambda > 0)$ if the pdf of X is

$$f(x;\lambda) = \left\{ \begin{array}{cc} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{array} \right.$$

PROPOSITION 4.4.5. The mean and variance of an exponential rv X are

$$\mu = \alpha\beta = \frac{1}{\lambda} \quad \sigma^2 = \alpha\beta^2 = \frac{1}{\lambda^2}$$

The cdf of exponential rv X is

$$F(x;\lambda) = \left\{ \begin{array}{cc} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{array} \right.$$

PROPOSITION 4.4.6. Suppose that the number of events occurring in any time interval of length t has a Poisson distribution with parameter αt(where α, the rate of the event process, is the expected number of events occurring in 1 unit of time) and that nubers of occurrences in nonoverlapping intervals are independent of one another. Then the distribution of elapsed tiem between the occurrence of two successive events is exponential with parameter $\lambda = \alpha$.

PROPOSITION 4.4.7. MEMORYLESS PROPERTY The distribution of additional lifetime is exactly the same as the original distribution of lifetime, so at each point in time the component shows no effect of wear. In other words, the distribution of remaining lifetime is independent of current age.

MEMORYLESS PROPERTY

4.4.3 The Chi-Squared Distribution

DEFINITION 4.4.4. Let υ be a positive integer. Then a rv X is said to have a **chi-squared distribution** with parameter υ if the pdf of X is the gamma density with $\alpha = \upsilon/2$ and $\beta = 2$. The pdf of a chi-squared rv is thus

$$f(x;\upsilon) = \begin{cases} \dfrac{1}{2^{\upsilon/2}\Gamma(\upsilon/2)} x^{(\upsilon/2)-1} e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The parameter υ is called the **number of degrees of freedom**(df) of X. The symbol χ^2 is often used in place of "chi-squared".

4.5 Other Continuous Distributions

4.5.1 The Weibull Distribution

DEFINITION 4.5.1. A rv X is said to have a **Weibull distribution** with parameters α and $\beta(\alpha > 0,\ \beta > 0)$ if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The mean and variance of Weibull rv X is

$$\mu = \beta\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \sigma^2 = \beta^2\left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}$$

The cdf of a Weibull rv having parameters α and β is

$$F(x;\alpha,\beta) = \begin{cases} 0 & x < 0 \\ 1 - e^{-(x/\beta)^{\alpha}} & x \geq 0 \end{cases}$$

Frequently, in practical situation, a Weibull model may be reasonable except that the smallest possible X may be some value γ not assumed o be zero(this would also apply to a gamma model). The quantity γ can then be regareded as a third parameter of the distribution. This is equivalent to saying that $X - \gamma$ has the pdf, so that the cdf of X is obtained by replacing x by $x - \gamma$.

4.5.2 The Lognormal Distribution

DEFINITION 4.5.2. A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with paramters μ and σ is

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The μ and σ denoted the mean and variance of $\ln(X)$.
The mean and variance of X can be obtained by

$$E(X) = e^{\mu+\sigma^2/2} \quad V(X) = e^{2\mu+\sigma^2} \cdot (e^{\sigma^2} - 1)$$

When we only have mean and variance of lognormal rv $\ln(X)$ and value table of normal rv X, we obtain the cdf by

$$\begin{aligned} F(x;\mu,\sigma) = P(X \leq x) = P[\ln(X) \leq \ln(x)] = P\left[\frac{\ln(X)-\mu}{\sigma} \leq \frac{\ln(x)-\mu}{\sigma}\right] \\ = P\left[Z \leq \frac{\ln(x)-\mu}{\sigma}\right] = \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right] \end{aligned}$$

4.5.3 The Beta Distribution

DEFINITION 4.5.3. A rv X is said to have a **beta distribution** with parameters α, β(both positive), A and B if the pdf of X is

$$f(x;\alpha,\beta,A,B) = \begin{cases} \dfrac{1}{B-A} \cdot \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \left(\dfrac{x-A}{B-A}\right)^{\alpha-1} \left(\dfrac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

The case $A = 0$, $B = 1$ gives the **standard beta distribution**.
The mean and variance of beta rv X are

$$\mu = A + (B-A) \cdot \frac{\alpha}{\alpha+\beta} \quad \sigma^2 = \frac{(B-A)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

In practice, the parameters A and B often denote the lower and upper bound of X, whereas the cdf has a potential maximum bound of $A \leq X \leq B$. See Example 4.23 in textbook Page 207.

4.6 Probability Plots

DEFINITION 4.6.1. The **probability plot** is used to check a distributional assumption.

4.6.1 A Probability Plot

DEFINITION 4.6.2. A plot of the n pairs

$$([100(i-.5)/n]\text{th } z \text{ percentile, } i\text{th smallest observation})$$

on a two-dimensional coordinate system is called a **normal probability plot**. If the sample observations are in fact drawn from a normal distribution with mean value μ and standard deviation σ, the points should fall close to a straight lien with slope σ and intercept μ. Thus a plot for which the points fall close to some straight line suggests that the assumption of a normal population distribution is plausible.

4.6.2 Transformations of a Random Variable

THEOREM 4.6.1. Let X hae pdf $f_X(x)$ and let $Y = g(X)$, where g is monotonic(either strictly increasing or strictly decreasing) so it has an inverse function $X = h(Y)$. Assume that h has a derivative $h'(y)$. Then $f_Y(y) = f_X(h(y))|h'(y)|$.

when it comes to discrete rv X and Y, the derivative part is not necessary $f_Y(y) = f_X(h(y))$.

Chapter 5

Joint Probability Distributions

5.1 Jointly Distributed Random Variables

5.1.1 The Joint Probability Mass Function for Two Discrete Random Variable

DEFINITION 5.1.1. Let X and Y be two discrete rv's defined on the sample space $\mathcal{S}$ of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

Let A be any set consisting of pairs of (x, y) values. Then the probability that the random pair (X, Y) lies in A is obtained by summing the joint pmf over pairs in A:

$$P[(X, Y) \in A] = \sum_{(x,y)\in A} \sum p(x, y)$$

DEFINITION 5.1.2. The **marginal mass function** of X and of Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x, y) \quad p_Y(y) = \sum_x p(x, y)$$

5.1.2 The Joint Probability Density Function for Two Continuous Random Variables

DEFINITION 5.1.3. Let X and Y be continuous rv's. Then $f(x,y)$ is the **joint probability density function** for X and Y if for any two-dimensional A

$$P[(X,Y)\in A]=\int_A\int f(x,y)dx\ dy$$

In particular, if A is the two-dimensional rectangle $\{(x,y): a\leq x\leq b, c\leq y\leq d\}$, then

$$P[(X,Y)\in A]=P(a\leq X\leq b, c\leq Y\leq d)=\int_a^b\int_c^d f(x,y)dxdy$$

DEFINITION 5.1.4. The **marginal probability density function** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy \quad \text{for } -\infty<x<\infty$$
$$f_Y(y)=\int_{-\infty}^{\infty} f(x,y)dx \quad \text{for } -\infty<y<\infty$$

5.1.3 Independent Random Variable

DEFINITION 5.1.5. Two random variable X and Y are said to be **independent** if for every pair of x and y values,

$$p(x,y)=p_X(x)\cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x,y)=f_X(x)\cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

X and Y are said to be **dependent** if either of these are not satisfied.

Independence of two rv's is more useful when the description of the experiment under study tells us that X and Y have no effect on each other. Then once the marginal pmf's or pdf's have been specified, the joint pmf or pdf is simply the product of the two marginal functions. It follows that

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b) \cdot P(c \leq Y \leq d)$$

5.1.4 More than Two Random Variables

DEFINITION 5.1.6. If $X_1, X_2, \ldots, X_n$ are all discrete random variables, the **joint pmf** of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If the variables are continuous, the **joint pdf** if $X_1, X_2, \ldots, X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) dx_n \ldots d_1$$

PROPOSITION 5.1.1. Joint pmf of multinomial rv based on n independent and identical trials is

$$p(x_1, \ldots, x_r) \\ = \begin{cases} \dfrac{n!}{(x_1!)(x_2!) \cdot \cdots \cdot (x_r!)} p_1^{x_1} \cdot \cdots \cdot p_r^{x_r} & x_i = 0, 1, 2, \ldots, \text{ with } x_1 + \cdots + x_r = n \\ 0 & \text{otherwise} \end{cases}$$

where x_i denotes the number of successes to be i.

DEFINITION 5.1.7. The rv $X_1, X_2, \ldots, X_n$ are said to be **independent** if for *every* subset $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ of the variables(each pair, each triple and so on), the joint pmf or pdf is equal to the product of the marginal pmf's or pdf's.

5.2 Expected Values, Cvariance, and Correlation

PROPOSITION 5.2.1. Let X and Y be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$ is given by

$$E[h(X, Y)] = \begin{cases} \displaystyle\sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y) dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Let X and Y be continuous independent rv's and suppose $h(X,Y)=XY$, then

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dx\ dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dx\ dy \\ &= \int_{-\infty}^{\infty} yf_Y(y)\left[\int_{-\infty}^{\infty} xf_X(x)dx\right]dy = E(X)E(Y) \end{aligned}$$

5.2.1 Convariance

DEFINITION 5.2.1. The **convariance** between two rv's X and Y is

$$\begin{aligned} \mathrm{Cov}(X,y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= \begin{cases} \sum_X \sum_Y (x-\mu_X)(y-\mu_Y)p(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)dx\ dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \end{aligned}$$

PROPOSITION 5.2.2.

$$\mathrm{Cov}(X,Y) = E(XY) - \mu_X \cdot \mu_Y$$

PROPOSITION 5.2.3. If X, Y, and Z are rv's and a and b are constants then

$$\mathrm{Cov}(aX+bY,Z) = a\mathrm{Cov}(X,Z) + b\mathrm{Cov}(Y,Z)$$

5.2.2 Correlation

DEFINITION 5.2.2. The **correlation coefficient** of X and Y, denoted by $\mathrm{Corr}(X,y)$, or $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\text{Cov(X,Y)}}{\sigma_X \cdot \sigma_Y}$$

PROPOSITION 5.2.4. 1. If X and Y are independent, then $\rho=0$, but $\rho=0$ does not imply independence.

2. $\rho = 1$ or -1 iff $Y = aX+b$ for some numbers a and b with $a \neq 0$.

5.3 Conditional Distributions

DEFINITION 5.3.1. Let X and Y be two discrete rv's with joint pmf $p(x, y)$ and marginal X pmf $p_X(x)$. Then for any x value such that $p_X(x) > 0$, the **conditional probability mass function of** Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}$$

Let X and Y be two continuous rv's with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$. Then for any x value such that $f_X(x) > 0$, the **conditional probability density function of** Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

DEFINITION 5.3.2. Let X and Y be two discrete rv's with conditional pmf $p_{Y|X}(y|x)$. Then the **conditional mean** or **expected value of** Y **given that** $X = x$ is

$$\mu_{Y|X=x} = E(Y|X = x) = \sum_{y \in D_Y} y \; p_{Y|X}(y|x)$$

also, for any function $g(y)$,

$$E(g(Y)|X = x) = \sum_{y \in D_Y} g(y) p_{Y|X}(y|x)$$

Let X and Y be two continuous rv's with conditional pdf $f_{Y|X}(y|x)$. Then

$$\mu_{Y|X=x} = E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

also, for any function $g(y)$,

$$E(g(Y)|X = x) = \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy$$

The **conditional variance of** Y **given** $X = x$ is

$$\begin{aligned}\sigma^2_{Y|X=x} = V(Y|X=x) &= E\left\{[Y - E(Y|X=x)]^2|X=x\right\} \\ &= E(Y^2|X=x) - \mu^2_{Y|X=x}\end{aligned}$$

5.3.1 Independence

PROPOSITION 5.3.1. Let X and Y be two rv's, X and Y are **independent** if

$$p_{Y|X}(y|x) = p_Y(y) \quad \text{or} \quad p(x,y) = p_X(x) \cdot p_Y(y)$$

5.3.2 The Bivariate Normal Distribution

DEFINITION 5.3.3. Let X and Y be two rv's which have a bivariate normal joint distribution:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{[(x-\mu_1)/\sigma_1]^2 + [(y-\mu_2)/\sigma_2]^2 - 2\rho(x-\mu_1)(y-\mu_2)/(\sigma_1\sigma_2)}{2(1-\rho^2)}}$$

and we have marginal distribution:

$$f_X(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{[(x-\mu_1)/\sigma_1]^2}{2}}$$

and conditional mean and conditional variance:

$$\mu_{Y|X=x} = E(Y|X=x) = \mu_2 + \rho\sigma_2\frac{x-\mu_1}{\sigma_1}$$

$$\sigma^2_{Y|X=x} = V(Y|X=x) = \sigma_2^2(1-\rho^2)$$

5.3.3 Regression to the Mean

From the equation

$$\frac{\mu_{Y|X=x} - \mu_2}{\sigma_2} = \rho \cdot \frac{x - \mu_1}{\sigma_1}$$

we can conclude that the correlation coefficient ρ is the factor between standardized conditional mean of Y and standardized X.

5.3.4 The Mean and Variance Via the Conditional Mean and Variance

THEOREM 5.3.1.

a. $E(Y) = E[E(Y|X)]$

b. $V(Y) = V[E(Y|X)] + E[V(Y|X)]$

These implies that $E(Y)$ is a weighted average of the conditional means $E(Y|X = x)$, where the weights are given by the pmf of X.

Proof.

$$E(Y) = E[E(Y|X)]$$

$$\begin{aligned} E[E(Y|X)] &= \sum_{x\in D_x} E(Y|X = x)p_X(x) = \sum_{x\in D_X}\sum_{y\in D_Y} y p_{Y|X}(y|x)p_X(x) \\ &= \sum_{x\in D_x}\sum_{y\in D_Y} y\frac{p(x,y)}{p_X(x)}p_X(x) = \sum_{y\in D_Y} y \sum_{x\in D_X} p(x,y) = \sum_{y\in D_Y} y p_Y(y) = E(Y) \end{aligned}$$

□

5.4 Transformations of Random Variables

5.4.1 The Joint Distribution of Two New Random Variables

THEOREM 5.4.1.

$$\det(M) = \left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right|$$

$$g(y_1, y_2) = f(x_1, x_2) \cdot \det(M)$$

where $\det(M)$ is called the Jacobian.

5.4.2 The Joint Distribution of More than Two New Variables

The case of two new variables can be extended to more than two variables with Jacobian.

5.5 Order Statistics

DEFINITION 5.5.1. The **order statistics** from a random sample are the rv's $Y_1, \dots, Y_n$ given by

$$\begin{aligned}&Y_1 = \text{the smallest among } \{X_i\}\\&\cdots\\&Y_n = \text{the largest among } \{X_i\}\end{aligned}$$

Then we get $Y_1 < Y_2 < \cdots < Y_n$.

5.5.1 The Distributions of Y_n and Y_1

PROPOSITION 5.5.1. Let Y_1 and Y_n denote the smallest and largest order statistics, respectively, based on a random sample from a continuous distribution with cdf $F(x)$ and pdf $f(x)$. Then the cdf and pdf of Y_n are

$$G_n(y) = [F(Y < y)]^n = [F(y)]^n \quad g_n(y) = n[F(y)]^{n-1} \cdot f(y)$$

The cdf and pdf of Y_1 are

$$G_1(y) = 1 - [F(Y \geq y)]^n = 1 - [1 - F(y)]^n \quad g_1(y) = n[1 - F(y)]^{n-1} \cdot f(y)$$

5.5.2 The Joint Distribution of the n Order Statistics

PROPOSITION 5.5.2. Let $g(y_1, \dots, y_n)$ denotes the joint pdf of the order statistics $Y_1, \dots, Y_n$ resulting from a random sample of X_i's from a pdf $f(x)$. Then

$$g(y_1, \dots, y_n) = \begin{cases} n! f(y_1) \cdot f(y_2) \cdot \cdots \cdot f(y_n) & y_1 < y_2 < \cdots < y_n \\ 0 & \text{otherwise} \end{cases}$$

n independent uniform rv's $(0, B)$ shows that the probability that all values are separated by at least d is

$$P(\text{all values are separated by more than } d) \quad = \begin{cases} [1-(n-1)d/B]^n & 0 \leq d \leq B/(n-1) \\ 0 & d > B/(n-1) \end{cases}$$

5.5.3 The Distribution of a Single Order Statistic

PROPOSITION 5.5.3. To get the probability distribution of y_i in continuous ordered random sample $\{y_i\}$ alone, we have

$$\begin{aligned} g(y_i) &= \int_{y_i}^{\infty} \cdots \int_{y_{n-1}}^{\infty} \int_{-\infty}^{y_i} \cdots \int_{-\infty}^{y_2} n! f(y_1) \cdot \cdots \cdot f(y_n) dy_1 \cdots dy_{i-1} dy_n \cdots dy_{i+1} \\ &= n! \left[\int_{y_i}^{\infty} \cdots \int_{y_{n-1}}^{\infty} f(y_{i+1}) \cdots f(y_n) dy_n \cdots dy_{i+1} \right] \cdot \left[\int_{-\infty}^{y_i} \cdots \int_{-\infty}^{y_2} f(y_1) \cdots f(y_{i-1}) dy_1 \cdots dy_{i-1} \right] \cdot f(y_i) \end{aligned}$$

Thanks to

$$\begin{aligned} &\int [F(x)]^k f(x) dx = \frac{1}{k+1} [F(x)]^{k+1} + c && \text{let } u = F(x) \\ &\int [1-F(x)]^k f(x) dx = -\frac{1}{k+1} [1-F(x)]^{k+1} + c && \text{let } u = 1-F(x) \end{aligned}$$

we get

$$g(y_i) = \frac{n!}{(i-1)! \cdot (n-i)!} [F(y_i)]^{i-1} [1-F(y_i)]^{n-i} f(y_i) \quad -\infty < y_i < \infty$$

5.5.4 The Joint Distribution of Two Order Statistics

To get the probability distribution of $Y = \{y_i, y_j\}$, we apply the same method to form three-part continuous integrals ranged from $y_1 \sim y_i$, $y_i \sim y_j$, and $y_j \sim y_n$.

Chapter 6

Statistics and Sampling Distributions

6.1 Statistics and Their Distributions

Any sample mean can be regarded as a *point estimate* of the population mean μ.

DEFINITION 6.1.1. A **statistic** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letterl; a lowercase letter is used to represent the calculated or observed value of the statistic.

6.1.1 Random Samples

DEFINITION 6.1.2. The rv's $X_1, \ldots, X_n$ are said to form a (simple) **random sample** of size n if

1. The X_i's are independent rv's.
2. Every X_i has the same probability distribution.

Thus, X_i's are **independent and identically distributed**(idd).

The conditions are satisfied if sampling are with replacement, from an infinite population, or without replacement yet the sample size n and population size N satisfies $n/N \leq .05$(at most 5% of the population is sampled).

6.2 The Distribution of the Sample Mean

PROPOSITION 6.2.1. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ, then

1. $E(\bar{X}) = \mu_{\bar{X}} = \mu$

2. $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
3. $E(T_o) = n\mu$
4. $V(T_o) = n\sigma^2$ and $\sigma_{T_o} = \sqrt{n}\sigma$

where $T_o = X_1 + \cdots + X_n$.

6.2.1 The Case of a Normal Population Distribution

PROPOSITION 6.2.2. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and stardard deviation σ. Then for any n, $\bar{X}$ is normally distributed(with mean μ and standard deviation $\sigma/\sqrt{n}$), as is T_o(with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$).

6.2.2 The Central Limit Theorem

THEOREM 6.2.1. THE CENTRAL LIMIT THEOREM(CLT) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit as $n \to \infty$, the standardized versions of $\bar{X}$ and T_o have the standard normal distribution. That is,

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P\left(\frac{T_o - n\mu}{\sqrt{n}\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

where Z is a standard normal rv. $\bar{X}$ and T_o are **asymptotically normal**.

Practical use of CLT: when n is large and we wish to calculate a probability such as $P(a \leq \bar{X} \leq b)$, we need only to pretend that $\bar{X}$ is normal, standardize it and use the normal table. The resulting answer will be approximatedly correct, as shown in Figure 6.1.

If $n > 30$, the CLT can be used.

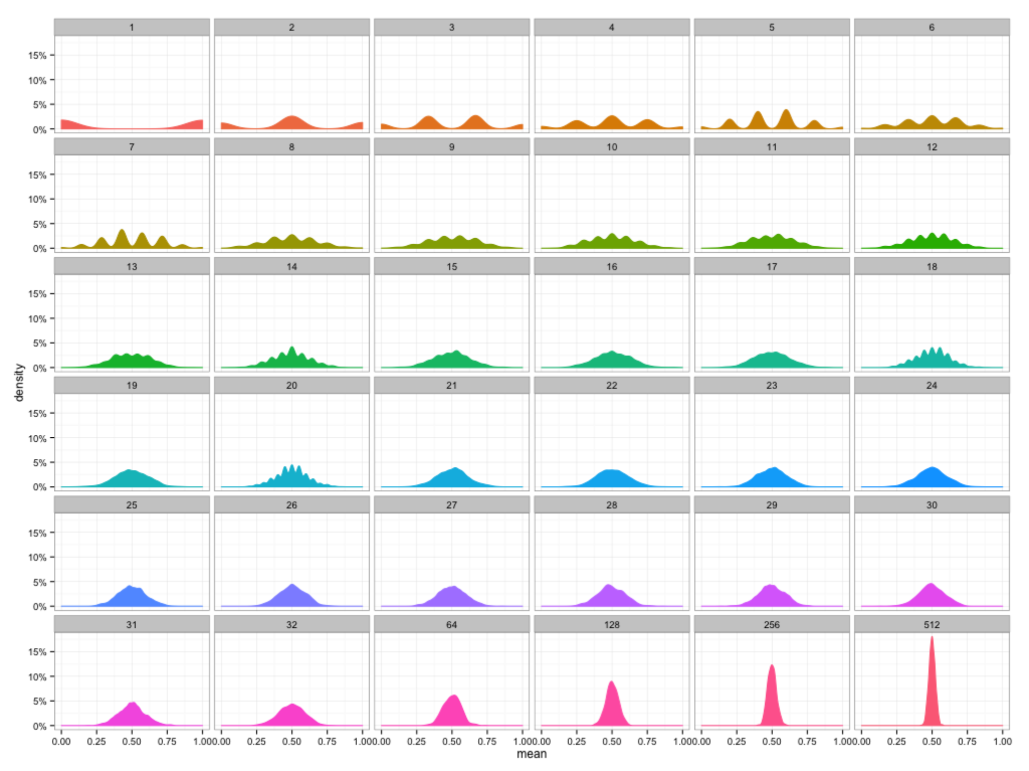

Figure 6.1: A simulation using the binomial distribution. Random 0s and 1s were generated, and then their means calculated for sample sizes ranging from 1 to 512. Note that as the sample size increases the tails become thinner and the distribution becomes more concentrated around the mean.(From Wikipedia.org)

6.2.3 Other Applications of the Central Limit Theorem

PROPOSITION 6.2.3. Let $X_1, \dots, X_n$ be a random sample from a distribution for which only positive values are possible($P(X_i > 0) = 1$). Then if n is sufficiently large, the product $Y = X_1 X_2 \cdot \dots \cdot X_n$ ahs approximately a lognormal distribution; that is, $\ln(Y)$ has a normal distribution.

6.2.4 The Law of Large Number

THEOREM 6.2.2. If $X_1, \dots, X_n$ is a random sample from a distribution with mean μ and variance σ^2, then $\bar{X}$ converges to μ

a. In mean square: $E[(\bar{X} - \mu)]^2 \to 0$ as $n \to \infty$

b. In probability: $P[|\bar{X} - \mu| \geq \epsilon \to 0]$ as $n \to 0$ for any $\epsilon > 0$

6.3 The Mean, Variance, and MGF for Several Variables

DEFINITION 6.3.1. Given a collection of n rv's $X_1, \dots, X_n$ and n numerical constants $a_1, \dots, a_n$, the rv

$$Y = a_1 X_1 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a **linear combination** of the X_i's.

PROPOSITION 6.3.1. Let $X_1, \dots, X_n$ have mean values $\mu_1, \dots, \mu_n$ respectively, and variances $\sigma_1^2, \dots, \sigma_n^2$ respectively.

1. Whether or not the X_i's are independent,

$$\begin{aligned} E(a_1 X_1 + \cdots + a_n X_n) &= a_1 E(X_1) + \cdots + a_n E(X_n) \\ &= a_1 \mu_1 + \cdots + a_n \mu_n \end{aligned}$$

2. If $X_1, \dots, X_n$ are independent,

$$\begin{aligned} V(a_1 X_1 + \cdots + a_n X_n) &= a_1^2 V(X_1) + \cdots + a_n^2 V(X_n) \\ &= a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2 \end{aligned}$$

and

$$\sigma_{a_1 X_1 + \cdots + a_n X_n} = \sqrt{a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2}$$

3. For any $X_1, \dots, X_n$,

$$V(a_1 X_1 + \cdots + a_n X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$$

6.3.1 The Difference Between Two Random Variables

PROPOSITION 6.3.2. $E(X_1 - X_2) = E(X_1) - E(X_2)$ and, if X_1 and X_2 are independent, $V(X_1 - X_2) = V(X_1) + V(X_2)$.

6.3.2 The Case of Normal Random Variables

PROPOSITION 6.3.3. If $X_1, \ldots, X_n$ are independent, normally distributed rv's, then any linear combination of the X_i's also has a normal distribution. In particular, the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

PROPOSITION 6.3.4. Let U and V be linear combinations of the independent normal rv's $X_1, \ldots, X_n$. Then the joint distribution of U and V is bivariate normal. The converse is also true.

6.3.3 Moment Generating Functions for Linear Combinations

PROPOSITION 6.3.5. Let $X_1, \ldots, X_n$ be independent rv's with moment generating functions $M_{X_1}(t), \ldots, M_{X_n}(t)$, respectively. Define $Y = a_1X_1 + \cdots + a_nX_n$, where $a_1, \ldots, a_n$ are constants. Then

$$M_Y(t) = \prod_{k=1}^{n} M_{X_k}(a_k t)$$

In the special case that $a_1 = \cdots = a_n = 1$,

$$M_Y(t) = \prod_{k=1}^{n} M_{X_k}(t)$$

which denotes the logarithm form of normal distribution convolution operations.

6.4 Distributions Based on a Normal Random Sample

Purpose:

1. Normal sample distribution is based on multiple normal rv's.
2. Normal sample variance → normal rv's variance → distribution for sums of normal rv's squares → χ^2 distribution
3. Normal sample standard deviation → combine normal rv with square root of χ^2 distribution → t distribution
4. Comparison of two normal sample variance → ratio of two χ^2 distributions → F distribution

6.4.1 The Chi-Squared Distribution

Recall: χ^2 distribution

χ^2 distribution is the special case of the gamma distribution with $\alpha = \upsilon/2$ and $\beta = 2$.

The pdf of χ^2 distribution is

$$f(x) = \begin{cases} \dfrac{1}{2^{\upsilon/2}\Gamma(\upsilon/2)} x^{(\upsilon/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where υ is degrees of freedom.

The mean, variance and degrees of freedeom are

$$\mu = \alpha\beta = \upsilon \quad \sigma^2 = \alpha\beta^2 = 2\upsilon \quad M_X(t) = (1-2t)^{\upsilon/2}$$

PROPOSITION 6.4.1. If Z has a standard normal distribution and $X = Z^2$, then the pdf of X is

$$f(x) = \begin{cases} \dfrac{1}{2^{1/2}\Gamma(1/2)} x^{(1/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

X is chi-squared with 1 df, or $X \sim \chi^2_1$.

The conclusion can be derived by differentiating cdf of $X = Z^2$.

PROPOSITION 6.4.2. If $X_1 \sim \chi^2_{\upsilon_1}$, $X_2 \sim \chi^2_{\upsilon_2}$, and they are independent, then $X_1 + X_2 \sim \chi^2_{\upsilon_1+\upsilon_2}$.

The conclusion can be derived by MGF products.

PROPOSITION 6.4.3. The df of idd is add-ups of idd df's.

If $Z_1, \ldots, Z_n$ are independent standard normal distributions, then $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n$.

$\chi_{\alpha,\upsilon} = c$ means that $P(\chi^2_\upsilon > c) = \alpha$(compared to $Z_\alpha = z \leftrightarrow P(Z \leq z) = \alpha$).

PROPOSITION 6.4.4. If X_i's are a random sample from a normal distribution, then $\bar{X}$ and S^2 are independent.

The conclusion can be derived by Covariance of the two items.

PROPOSITION 6.4.5. If $X_3 = X_1 + X_2$, and $X_1 \sim \chi^2_{\upsilon_1}$, $X_3 \sim \chi^2_{\upsilon_3}$, $\upsilon_3 > \upsilon_1$, and X_1 and X_2 are independent, then $X_2 \sim \chi^2_{\upsilon_3-\upsilon_1}$.

The conclustion can be dirived by MGF products.

PROPOSITION 6.4.6. If X_i's are a random sample from a normal distribution, then

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

The df $n-1$ is based on the fact that for X_i, s_i^2 is based on the first $n-1$ x_i's rather than all x_i's.

6.4.2 The t Distribution

THEOREM 6.4.1. If $X_1, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, then variable has the t distribution with $n-1$ df, t_{n-1}, is

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{X/\upsilon}}$$

The pdf of a rv T with t distribution and υ df is

$$f(t) = \frac{1}{\sqrt{\pi\upsilon}} \frac{\Gamma[(\upsilon+1)/2]}{\Gamma(\upsilon/2)} \frac{1}{(1+t^2/\upsilon)^{(\upsilon+1)/2}} \quad -\infty < t < \infty$$

The conclusion can be derived by differentiating the cdf of t distribution.

6.4.3 The F Distribution

DEFINITION 6.4.1. Let X_1, X_2 be independent χ^2 rv's with υ_1 and υ_2 df. The F distribution with υ_1 numerator df and υ_2 denominator df is

$$F_{\upsilon_1,\upsilon_2} = \frac{X_1/\upsilon_1}{X_2/\upsilon_2}$$

Chapter 7

Point Estimation

Purpose:

1. **Point estimation** is a guess for the true value of the parameter from a sample.
2. Two methods of obtaining point estimation: **the method of moments** and **the method of maximum likelihood**
3. **Sufficiency** guarantees no information loss in the chosen statistic.

7.1 General Concepts and Criteria

DEFINITION 7.1.1. A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ.

A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data.

The selected statistic is called the **point estimator** of θ.

The point estimate resulting from a given sample or the estimator of θ are both called $\hat{\theta}$.

7.1.1 Mean Squared Error

DEFINITION 7.1.2. The **mean squared error** of an estimator $\hat{\theta}$ is $E[(\hat{\theta}-\theta)^2]$.

$$\mathrm{MSE} = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 = \text{variance of estimator} + (\text{bias})^2$$

7.1.2 Unbiased Estimators

DEFINITION 7.1.3. A point estimator $\hat{\theta}$ has an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ.

If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

PROPOSITION 7.1.1. When X is a binomial rv with parameters n and p, the sample proposition $\hat{p} = X/n$ is an unbiased estimator of p.

Sample variance S^2 is an unbiased estimator of σ^2.

7.1.3 Estimators with Minimum Variance

DEFINITION 7.1.4. Among all estimators of θ that are unbiased, choose the one that has minimum variance.

The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator**(MVUE) of θ.

Since MSE = variance + (bias)^2, seeking an unbiased estimator with minimum variance is the same as seeking an unbiased estimator with minimum MSE.

THEOREM 7.1.1. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ. Then the estimator $\hat{\mu} = \bar{X}$ is the MVUE for μ.

When estimating a point of symmetry μ of a continuous probability distribution, a trimmed mean with trimming proportion 10% or 20%(from each end of the sample) produces reasonably bahaved estimates over a very wide range of possible models. Thus, a trimmed mean with small trimming percentage is said to be a **robust estimator**.

PROPOSITION 7.1.2. Censoring: see textbook Page 343.

7.1.4 Reporting a Point Estimate: The Standard Error

DEFINITION 7.1.5. The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$.

If the standard error involves unknown parameters, substitution creates the **estimated standard error**(estimated standard deviation) of the estimator.

The estimated standard error can be denoted by either $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$.

7.1.5 The Bootstrap

Bootstrap estimate of $\hat{\theta}$'s standard error equals the sample standard deviation of the $\hat{\theta}_i^*$'s

$$S_{\hat{\theta}} = \sqrt{\frac{1}{B-1} \sum \left(\hat{\theta}_i^* - \bar{\theta^*}\right)^2}$$

7.2 Methods of Point Estimation

7.2.1 The Method of Moments

DEFINITION 7.2.1. Let $X_1, \ldots, X_n$ be a random sample from a pmf or pdf $f(x)$.

For $k = 1, 2, 3, \ldots$, the k**th population moment**, or the k**th moment of the distribution** $f(x)$, is $E(X^k)$.

The k**th sample moment** is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

DEFINITION 7.2.2. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown.

Then the **moment estimators** $\theta_1, \ldots, \theta_m$ are obtained by the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \ldots, \theta_m$.

e.g., $E(X) = \frac{1}{n}\sum X_i = \bar{X}, E(X^2) = \frac{1}{n}\sum X_i^2$

7.2.2 Maximum Likelihood Estimation

DEFINITION 7.2.3. Let $X_1, \ldots, X_n$ have joint pmf or pdf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$ where parameters $\theta_1, \ldots, \theta_m$ have unknown values.

When $x_1, \ldots, x_n$ are the observed sample values and f is regarded as a function of $\theta_1, \ldots, \theta_m$, it is called the **likelihood function**.

The maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that $f(x_1, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m) \geq f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$ for all $\theta_1, \ldots, \theta_m$.

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** result.

7.2.3 Some Properties of MLEs

PROPOSITION 7.2.1. The Invariance Principle

Let $\hat{\theta}_1, \ldots, \hat{\theta}_m$ be the mle's of the parameters $\theta_1, \ldots, \theta_m$. Then the mle of any function $h(\theta_1, \ldots, \theta_m)$ of these parameters is the functuion $h(\hat{\theta}_1, \ldots, \hat{\theta}_m)$, of the mle's.

Intuitively, the principle extends the MLE of parameterθ to MLE of function$h(\theta)$.

7.2.4 Large-Sample Behavior of the MLE

PROPOSITION 7.2.2. Under very general conditions on the joint distribution of the sample, when the sample size is large, the maximum likelihood estimator of any parameter θ is close to θ(cosistency), is approximately unbiased $[E(\hat{\theta}) \approx \theta]$, and has variance that is nearly as small as can be achieved by any unbiased estimator. Stated another way, the mle $\hat{\theta}$ is approximately the MVUE of θ.

7.3 Sufficiency

DEFINITION 7.3.1. A statistic $T = t(X_1, \ldots, X_n)$ is said to be **sufficient** for making inferences about a parameter θ if the joint distribution of $X_1, \ldots, X_n$ given that $T = t$ does not depend upon θ for every possible value t of the statistic T.

7.3.1 The Factorization Theorem

THEOREM 7.3.1. THE NEYMAN FACTORIZATION Theorem See textbook Page 376.

7.4 Information and Efficiency

DEFINITION 7.4.1. The **Fisher information** $I(\theta)$ in a single observation from a pmf or pdf $f(x;\theta)$ is the variance of the random variable $U = \frac{\partial}{\partial\theta}\ln[f(X;\theta)]$:

$$I(\theta) = V\left[\frac{\partial}{\partial\theta}\ln(f(X;\theta))\right]$$

Proof.

$$\begin{aligned}
1 &= \sum_x f(x;\theta) \\
0 &= \frac{\partial}{\partial\theta}\sum_x f(x;\theta) = \sum_x \frac{\partial}{\partial\theta} f(x;\theta) \\
&= \sum_x \frac{\partial}{\partial\theta}[\ln f(x;\theta)]f(x;\theta) = E\left[\frac{\partial}{\partial\theta}\ln(f(X;\theta))\right] = E(U) \\
I(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X;\theta))\right]
\end{aligned}$$

□

7.4.1 Information in a Random Sample

score function

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln f(X_1,\ldots,X_n;\theta) &= \frac{\partial}{\partial\theta}\ln[f(X_1;\theta)\cdot\cdots\cdot f(X_n;\theta)] \\
&= \frac{\partial}{\partial\theta}\ln f(X_1;\theta) + \cdots + \frac{\partial}{\partial\theta}\ln f(X_n;\theta)
\end{aligned}$$

$$E\left[\frac{\partial}{\partial\theta}\ln f(X_1,\ldots,X_n;\theta)\right] = 0$$

$$I_n(\theta) = V\left[\frac{\partial}{\partial\theta}\ln f(X_1,\ldots,X_n;\theta)\right] = nV\left[\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right] = nI(\theta)$$

7.4.2 The Cramer-Rao Inequality

THEOREM 7.4.1. Assume a random sample $X_1, \dots, X_n$ from the distribution with pmf or pdf $f(x;\theta)$ such that the set of possible values does not depend on θ. If the statistic $T = t(X_1, \dots, X_n)$ is an unbiased estimator for the paramter θ, then

$$V(T) \geq \frac{1}{V\left\{\frac{\partial}{\partial\theta}[\ln f(X_1, \dots, X_n; \theta)]\right\}} = \frac{1}{nI(\theta)} = \frac{1}{I_n(\theta)}$$

DEFINITION 7.4.2. Let T be an unbiased estimator of θ. The ratio of the lower bound to the variance of T is its efficiency. Then T is said to be an efficient esitimator if T achieves the Cramer-Rao lower bound(the efficiency is 1). An efficient esitmator is a minimum variance unbiased(MVUE) estimator.

7.4.3 Large Sample Properties of the MLE

THEOREM 7.4.2. Given a random sample $X_1, \dots, X_n$ from a distribution with pmf or pdf $f(x;\theta)$, assume that the set of possible x values does not depend on θ. Then for large n the maximum likelihood estimator $\hat{\theta}$ has approximately a normal distribution with mean θ and variance $\frac{1}{nI(\theta)}$. More precisely, the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is normal with mean 0 and variance $\frac{1}{I(\theta)}$.

Chapter 8

Statistical Intervals Based on a Single Sample