ECE286 - Probability and Statistics

What is the probability of me passing this course

Regis Zhao[1]

University of Toronto

regis.zhao@mail.utoronto.ca

[1] *TeX file on GitHub*

Contents

PART

I

Probability

SECTION 1

Introduction

- probability comes from:
  - things we can't model well
  - good models but limited measurements
- uncertainty is unavoidable, but probability helps describe uncertainty

SECTION 2

Sample Space

Definition 1 **Sample Space**: the set of all possible outcomes, S

- e.g. for 1 coin flip: $S = H, T$

- each outcome in a sample space is called an **element** or **member**

SECTION 3

Events

Definition 2 **Event**: a subset of sample space S

- e.g. for a die: each element $\{1, 2, 3, \dots\}$ is an event, rolling even or rolling odd are events

Definition 3 The **complement** of an event A with respect to S : everything in S that isn't in A

- denoted by A'
- e.g. for a die: $\{1, 2\}$ is a complement of $\{3, 4, 5, 6\}$

Definition 4 The **intersection** of two events A and B : everything in A *and* B

- denoted by $A \cap B$
- A and B are mutually exclusive if $A \cap B = \emptyset$ (empty set)

Definition 5 The **union** of two events A and B : everything in A or B

- denoted by $A \cup B$
- $A \cup A' = S$

SECTION 4
Counting

Theorem 1 **Generalized Multiplication Rule**: if an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, and for each of these ..., then the sequence of k operations can be performed in $n_1 n_2 \dots n_k$ ways

- e.g. Menu options: 3 appetizers, 4 mains, 2 deserts
- then there are $3 \cdot 4 \cdot 2 = 24$ options

Definition 6 **Permutation**: an arrangement of all or part of a set of objects

- we can derive formula for permutations using the multiplication rule:
- for example: permutations of three letters a, b, and c
- there are 3 choices for first position, and no matter what you choose, there will be 2 choices for the second, and 1 choice for the third
- therefore: $(3)(2)(1) = 6$ permutations

Theorem 2 The number of permutations of n objects is $n!$.

Theorem 3 The number of permutations of r out of n items is

$$_nP_r = \frac{n!}{(n-r)!}.$$

Theorem 4 The number of permutations of n objects arranged in a circle is $(n-1)!$.

SUBSECTION 4.1
Permutations with Identical Items

Theorem 5 Given m kinds of items, and each kind of item has n_k of them $(k = 1, 2, \dots, m)$, then the number of *distinct* permuations is

$$\binom{n}{n_1, n_2, \dots, n_m} = \frac{n!}{n_1! n_2! \dots n_m!}.$$

SUBSECTION 4.2
Partitions

- partitions divide a set into subsets
- often we want to find the number of possible ways to split a set up into partitions, where in each partition, the order doesn't matter

Theorem 6 Given m partitions of size $n_1, n_2, \dots, n_m$, the number of ways of partitioning the set is

$$\binom{n}{n_1, n_2, \ldots, n_m} = \frac{n!}{n_1! n_2! \ldots n_m!}.$$

- note that this is the same formula as the number of permutations with identical items
- this is because once we put a group of items in a partition, their order doesn't matter anymore and they are essentially identical elements to us

SUBSECTION 4.3

Combinations

- **combinations** are ways of selecting objects without regard to order
- combinations are like permutations except you don't care about order
- you can think of a size r combination as paritioning a set into 2 cells, where one cell has size r and the other is the rest of the set
  - how many ways can you put items from a set into a size r partition
- using the partition formula:

Theorem 7 The number of size r combinations of n distinct objects is

$$\binom{n}{r, n-r} = \frac{n!}{r!(n-r)!},$$

or more commonly written as "n choose r":

$$\binom{n}{r}.$$

SECTION 5

Probability of an Event

- a measure of the likelihood of an event happening – a value ranging from 0 to 1

Definition 7 The **probability** of an event A in sample space S is the sum of the weights of all sample points in A.

$$0 \leq P(A) \leq 1, \quad P(\emptyset) = 0, \quad \text{and} \quad P(S) = 1.$$

- if $A_1, A_2, A_3, \ldots$ is a sequence of mutually exclusive events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots.$$

Subsection 5.1
Additive Rules

Theorem 8 **Additive Rule** (applies to unions of events): if A and B are two events, then

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

- if A and B are mutually exclusive, then $A \cap B = \emptyset$ and

$$P(A \cup B) = P(A) + P(B).$$

- if A and A' are complementary events, then

$$P(A) + P(A') = 1.$$

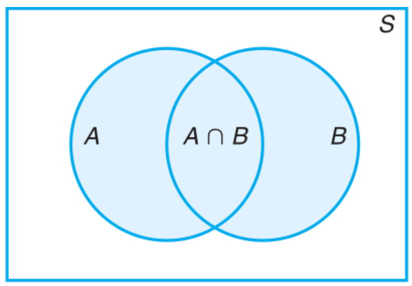

Figure 1. Additive rule of probability

Theorem 9 For three events A, B, and C,

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

Section 6
Conditional Probability, Independence, and the Product Rule

Subsection 6.1
Conditional Probability

- **conditional probability** is the probability of an event B occurring when it is known that some event A has occurred

Definition 8 The **conditional probability** of B, given A, is denoted by $P(B|A)$ and is defined by

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{provided} \quad P(A) > 0.$$

- the probability of B happening, given A, is equal to the probability of their intersection divided by the probability of A happening

SUBSECTION 6.2
Independence

Definition 9 A and B are independent if and only if

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B).$$

- probability of A or B happening doesn't depend on if the other event happened
- otherwise, A and B are **dependent**

- the condition $P(B|A) = P(B)$ implies that $P(A|B) = P(A)$
- note that independence $\neq$ mutually exclusive
  - e.g. head and tails are mutually exclusive but not independent ($P(H|T) = 0$)

SUBSECTION 6.3
Product Rule

- allows us to calculate the probability that two events will both occur

Theorem 10 **Product Rule**: If in an experiment the events A and B can both occur, then

$$P(A \cap B) = P(A)P(B|A), \quad \text{provided} \quad P(A) > 0.$$

Theorem 11 Two events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B).$$

- the probability that two independent events will both occur is equal to the product of their individual probabilities
- notice how this is a special case of the product rule $P(A \cap B) = P(A)P(B|A)$ where $P(B|A) = P(B)$ since A and B are independent

Theorem 12 **Product Rule for two or more events**: if the events A_1, A_2, …, A_k can occur, then

refer to the textbook theorem 2.12 lol.

If the events A_1, A_2, …, A_k are independent, then

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)P(A_2)\ldots P(A_k).$$

SECTION 7
Bayes' Rule

SUBSECTION 7.1
Total Probability

- addresses the problem of finding the total probability of something happening, when you know its conditional probabilities

- for example, what is the probability of a product being defective if it was produced by machines that each have a probability of creating defective products

Theorem 13 If the events B_1, B_2, $\ldots$, B_k constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots k$, then for any event A of S,

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

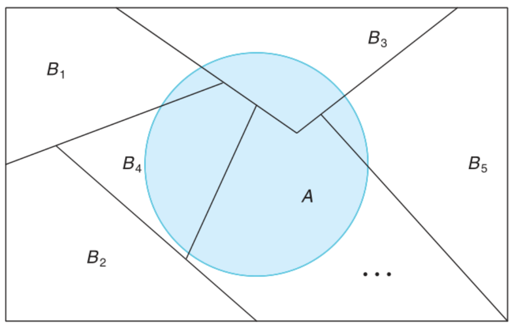

Figure 2. Partitioning the sample space S. The probability of A is the sum of the probabilities of the intersections between the partitions and A.

Subsection 7.2

Bayes' Rule

- addresses the problem of finding conditional probability, $P(B_i|A)$
- for example, what is the probability that a product was created by a certain machine, given that the product is defective
- recall formula for conditional probability, and substitute in the formula for total probability in the denominator:

Theorem 14 **Bayes' Rule**: If the events B_1, B_2, $\ldots$, B_k constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$, then the probability of a cell B_n in the partition, given A, is given by

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

- recall the product rule, rearranging it, we get:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \text{and} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- noting that $P(B \cap A) = P(A \cap B)$, we can rearrange and equate the above equations:

$$P(B|A)P(A) = P(A|B)P(B)$$

or

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

PART II

Random Variables and Probability Distributions

Section 8

Concept of a Random Variable

Definition 10 **Random variable** (RV): a function that maps each element in the sample space to a real number

- we use capital letters, say X, to denote a random variable and use its corresponding small letter, x, for one of its values it can take on

Definition 11 **Discrete RV**: X takes on a finite or countable number of values
Continuous RV: X takes on values in an interval of $\mathbb{R}$

Section 9

Discrete Probability Distributions

Definition 12 The set of ordered pairs (x, f(x)) is a **probability function**, **probability mass function** (PMF), or **probability distribution** of the discrete RV X if, for each possible outcome x,

1. $f(x) \geq 0$
2. $\sum_x f(x) = 1$
3. P(X = x) = f(x)

Definition 13 The **cumulative distribution function** (CDF) $F(x)$ of a discrete random variable X with PMF $f(x)$ is the probability of X being less than or equal to x :

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for} -\infty < x < \infty.$$

Section 10

Continuous Probability Distributions

- for a continuous random variable, the probability of it assuming a specific value *exactly* is 0 since there are infinite values
- but the probability of X assuming a value in an interval is nonzero

Definition 14 The function $f(x)$ is a **probability density function** (PDF) for the continuous RV X, defined over the set of real numbers, if

1. $f(x) \geq 0, \quad$ for all $\quad x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X < b) = \int_a^b f(x)dx$

Definition 15 The **cumulative distribution function** $F(x)$ of a continuous random variable X with PDF $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \quad \text{for} \quad -\infty < x < \infty.$$

Section 11

Joint Probability Distributions

- previous section only focused on 1D sample spaces
- when dealing with the simultaneous occurrence of two RVs:

Definition 16 The function $f(x,y)$ is a **joint probability distribution** or **joint PMF** of the discrete random variables X and Y if

1. $f(x,y) \geq 0$ for all (x,y)
2. $\sum_x \sum_y f(x,y) = 1$
3. $P(X = x, Y = y) = f(x,y)$

For any region A in the xy plane, $P[(X,Y) \in A] = \sum \sum_A f(x,y)$.

Definition 17 The function $f(x,y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x,y) \geq 0$, for all (x,y)
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1$
3. $P[(X,Y) \in A] = \int \int_A f(x,y)dxdy$, for any region A in the xy plane

Subsection 11.1

Marginal Distributions

- what if we know the joint distribution of two RVs but only care about one (want to obtain the probability distribution of an individual RV)
- simply integrate/add along the variable to eliminate

Definition 18 The **marginal distributions** of X alone and of Y alone are

- for the discrete case:

$$g(x) = \sum_y f(x,y) \quad \text{and} \quad h(y) = \sum_x f(x,y).$$

- for the continuous case:

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

- idea: marginal distribution is just the 'weighted average' of $f(x,y)$ over all possibilities of x or y

Subsection 11.2

Conditional Distributions

- recall conditional probability
- conditional distributions take very similar form

Definition 19 Let X and Y be two RVs. The **conditional distribution** of RVs (discrete or continuous) is

$$f(y|x) = \frac{f(x,y)}{g(x)}$$
$$f(x|y) = \frac{f(x,y)}{h(y)}.$$

- if we want to find the probability that X falls between a and b, given $Y = y$:

$$P(a < X < b|Y = y) = \sum_{a<x<b} f(x|y)$$
$$P(a < X < b|Y = y) = \int_a^b f(x|y)dx.$$

Subsection 11.3

Statistical Independence

Definition 20 Let X and Y be RVs with joint distribution $f(x,y)$ and marginal distributions $g(x)$ and $h(y)$. X and Y are **statistically independent** if and only if

$$f(x,y) = g(x)h(y).$$

for all (x,y) within their range.

- same idea applies for joint probability distributions of more than 2 RVs – their joint probability distributions are simply the product of the marginal distributions if the RVs are statistically independent

PART III

Mathematical Expectation

Section 12

Mean of a Random Variable

- essentially calculating what value of x is most likely to occur based on the probability distribution $f(x)$

Definition 21 Let X be an RV with probability distribution $f(x)$. The **mean**, or **expected value** of X is

- for discrete case:

$$\mu = E(X) = \sum_x x f(x).$$

- for continuous case:

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

- notice: we mutliply the value of x with its own probability so that values of x with higher probability have a greater influence on what the expected value is

Definition 22 Let X be an RV with probability distribution $f(x)$. We define a new RV as a function of X, $g(X)$. The expectation of the RV $g(X)$ is

- for discrete case:

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x) f(x).$$

- for continuous case:

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

Definition 23 Let X and Y be RVs with joint probability distribution $f(x, y)$. The expectation of the RV $g(X, Y)$ is

- for discrete case:

$$\mu_{g(X,Y)} = E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y).$$

- for continuous case:

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

- generalization of the calculation of mathematical expectations of functions of more than 2 RVs is straightforward

Section 13

Variance and Covariance of Random Variables

Subsection 13.1

Variance

- measures how spread out a distribution is – the variability of an RV

Definition 24 Let X be an RV with probability distribution $f(x)$ and mean $\mu = E(X)$. The **variance** of X is

- if X is discrete:

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x).$$

- if X is continuous:

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

The positive square root of the variance, σ, is called the **standard deviation** of X.

- sometimes, variance is written as $var(X)$
- the $x - \mu$ centers the distribution on the y-axis
- squaring it allows for points at a greater distance from the mean μ to have a larger contribution, therefore measuring how spread out the distribution is

Theorem 15 The variance of an RV X is

$$\sigma^2 = E(X^2) - \mu^2.$$

- proof is in section 4.2 of textbook

Subsection 13.2

Covariance

- measures the joint variability of two variables – the direction of the relationship between two variables
  - if large values of both variables occur together, covariance is positive
  - if large values of one correspond to small values of the other RV, covariance is negative

Definition 25 Let X and Y be RVs with joint probability distribution $f(x, y)$. The **covariance** of X and Y is

- if X and Y are discrete:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y).$$

- if X and Y are continuous:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

- also written as $cov(X, Y)$

Theorem 16 The covariance of two random variables X and Y with means μ_X and μ_Y is given by

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y.$$

Subsection 13.3

Correlation Coefficient

- the sign of covariance provides information about the nature of the relationship between two variables, but the magnitude does not indicated anything about the *strength* of the relationship since covariance isn't scale-free
  - its magnitude depends on the units used to measure X and Y
- the scale-free version of covariance is called the correlation coefficient:

Definition 26 Let X and Y be RVs with covariance σXY and standard deviations σ_X and σ_Y. The **correlation coefficient** of X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- like covariance, but normalized
- magnitude tells us *strength* of relationship
- $-1 \leq \rho_{XY} \leq 1$
- 'uncorrelated' if $\rho_{XY} = 0$
  - since that would mean $\sigma_{XY} = 0$

Section 14

Means and Variances of Linear Combinations of Random Variables

Subsection 14.1

Means of LCs of RVs

- **expectation is linear**

Theorem 17 If a and b are constants, then

$$E(aX + b) = aE(X) + b.$$

Theorem 18 The expectation of the sum or difference of two or more functions of an RV is the sum or difference of the expectations of the functions, i.e.

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

SUBSECTION 14.2

Variances of LCs of RVs

Theorem 19 Let X and Y be two independent RVs. Then

$$E(XY) = E(X)E(Y).$$

- recall formula for covariance:

$$\sigma_{XY} = E(XY) - E(X)E(Y).$$

- if X and Y are independent, then $E(XY) = E(X)E(Y)$, so:

Theorem 20 Let X and Y be two independent RVs. Then $\sigma_{XY} = 0$

- i.e. independence implies uncorrelated
- but uncorrelated does not imply independence
- independece is a stronger property than uncorrelated

Theorem 21 The variance of $aX + bY + c$, where a, b, and c are constants, is

$$\sigma^2_{aX+bY+c} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY}.$$

- notice that c has no effect on the variance – the variance is unchanged if a constant is added or subtracted from an RV
  - it simply shifts the values of the RV left or right, it doesn't change the variability
- also notice that if X and Y are independent, then the last term is 0
- multiplying an RV by a constant scales the variance by the square of the constant, i.e.

Theorem 22 The variance of aX, where a is a constant, is $\sigma^2_{aX} = a^2\sigma^2_X$

PART

IV

Common Discrete Probability Distributions

Section 15

Uniform Distribution

- every element in S has the same probability (e.g. coin flip)
- if $S = 1, 2, \ldots, n$, then $f(k) = \frac{1}{n}$ for $k \in S$

Section 16

Binomial Distribution

Definition 27 **Binomial Distribution**: the probability of x successes in n trials for a binomial experiment:

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Theorem 23 The mean and variance of the binomial distribution are

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p).$$

Section 17

Multinomial Distribution

- like binomial but each trial has more than 2 possibilities, E_1, E_2, $\ldots$, E_k, where k is the number of possibilities a trial can take on

Definition 28 **Multinomial Distribution**: the probability of E_1 happening x_1 times, E_2 happening x_2 times, $\ldots$, E_k happening x_k times, where $x_1 + x_2 + \ldots + x_k = n$:

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

Section 18

Hypergeometric Distribution

Definition 29 **Hypergeometric Distribution**: given that there are a set amount of successes K in a sample space of size N, what is the probability of selecting x successes if you select n times (with replacement)

$$h(x; N, n, K) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \qquad (n - x \leq N - K).$$

- mean and variance:

$$\mu = \frac{nK}{N} \quad \text{and} \quad \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{K}{N}\left(1 - \frac{K}{N}\right).$$

Section 19

Negative Binomial Distribution

Definition 30 **Negative Binomial Distribution**: probability of the k-th success occuring on the x-th trial, where the probability of a success is p

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}.$$

- note:

$$b^*(x; k, p) = pb(k-1; x-1, p).$$

Subsection 19.1

Geometric Distribution

Definition 31 **Geometric Distribution**: a special case of the negative binomial distribution, where $k = 1$, i.e. probability of the first success happening on the x^{th} trial

$$g(x; p) = b^*(x; 1, p) = p(1-p)^{x-1}.$$

- mean and variance:

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}.$$

Section 20

Poisson Distribution

- like binomial except number of trials is continuous over some interval ($n \to \infty$)
- properties of a **Poisson Process**:
    1. the number of outcomes in one time interval is independent of the number that occur in any other interval – the Poisson process has no memory
    2. the probability that a single outcome will occur during an interval is proportional to the length of the interval and doesn't depend on the number of outcomes occurring outside this interval

Definition 32

Poisson Distribution: the probability distribution of a Poisson random variable X, representing the number of outcomes occurring in a given time interval or specified region denoted by t, is

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

where λ is the average number of outcomes per unit interval

- mean and variance are both given by

$$\mu = \sigma^2 = \lambda t.$$

- note: the binomial distribution $b(x; n, p)$ becomes the Poisson distribution $p(x; \mu)$ as the sample size $n \to \infty$

PART

V

Continuous Probability Distributions

Section 21

Continuous Uniform Distribution

Definition 33 **Continuous Uniform Distribution**:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{elsewhere} \end{cases}.$$

- the mean and variance are given by

$$\mu = \frac{A+B}{2} \quad \text{and} \quad \sigma^2 = \frac{(B-A)^2}{12}.$$

Section 22

Normal Distribution

- the most important continuous probability distribution in the entire field of statistics
- also known as the **Gaussian distribution**
- a continuous random variable X have the bell-shaped distribution of the normal curve is called a **normal random variable**

Definition 34 **Normal/Gaussian Distribution**:

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

- the mode occurs at the maximum, when $x = \mu$
- it is symmetric about $x = \mu$
- points of inflection occur at $x = \mu \pm \sigma$

Subsection 22.1

Standard Normal Distribution

- calculating areas under the normal curve is important to obtain probabilities, but it's rather dumb to make tables of values for every single value of μ and σ^2
- instead, we are able to transform all observations of any normal RV X into a new set of observations of a normal RV Z with mean 0 and variance 1:

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

Definition 35 **Standard Normal Distribution**: special case of the normal distribution where $\mu = 0$ and $\sigma^2 = 1$

Section 23

Normal Approximation to the Binomial Distribution

Theorem 24 if X is a binomial RV with $\mu = np$ and $\sigma^2 = np(1-p)$, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}}.$$

as $n \to \infty$ is the standard normal distribution $n(z; 0, 1)$

Section 24

Gamma and Exponential Distributions

- the **Gamma function** is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \qquad \text{for } \alpha > 0.$$

- properties of the Gamma function:
  1. $\Gamma(n) = (n-1)(n-2)\dots(1)\Gamma(1)$, for a positive integer n
  2. $\Gamma(n) = (n-1)!$ for a positive integer n
  3. $\Gamma(1) = 1$
  4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Definition 36 **Gamma Distribution**:

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

where $\alpha, \beta > 0$

- mean and variance:

$$\mu = \alpha\beta \quad \text{and} \quad \sigma^2 = \alpha\beta^2.$$

Definition 37 **Exponential Distribution**: special case of Gamma distributiion where $\alpha = 1$

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}.$$

where $\beta > 0$

- mean and variance:

$$\mu = \beta \quad \text{and} \quad \sigma^2 = \beta^2.$$

Section 25

Chi-Squared Distribution

Definition 38 **Chi-Square Distribution**: special case of Gamma distribution where $\alpha = v/2$, $\beta = 2$, and v is a positive integer and is the only parameter, called the **degrees of freedom**

$$f(x; v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}.$$

- mean and variance:

$$\mu = v \quad \text{and} \quad \sigma^2 = 2v.$$

Section 26

Weibull Distribution

Definition 39 **Weibull Distribution**:

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

where $\alpha, \beta > 0$

- its cumulative function is given by

$$F(x) = 1 - e^{-\alpha x^{\beta}}.$$

Functions of Random Variables

PART VI

Section 27

Transformations of Variables

Theorem 25 Suppose X is a **discrete** RV with probability distribution $f(x)$. Let $Y = u(X)$ define a one-to-one transformation between values of X and Y so that we can also write $x = w(y)$. Then the probability distribution of Y is

$$g(y) = f(w(y)).$$

- for the case of joint probability distributions $f(x_1, x_2)$, the probability distribution of Y is

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

Theorem 26 For the case where the RV is **continuous**, the probability distribution of Y is

$$g(y) = f(w(y))|J|.$$

where $J = w'(y)$ is the **Jacobian** of the transformation.

- for joint probability distributions:

$$g(x_1, x_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

  where the Jacobian is the determinant of the Jacobian matrix

Section 28

Moments and Moment-Generating Functions

Definition 40 The r th **moment about the origin** of the RV X is given by

$$\mu'_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx.$$

- we can write mean and variance of a random variable in terms of moments:

$$\mu = \mu'_1 \quad \text{and} \quad \sigma^2 = \mu'_2 - \mu^2.$$

Definition 41 **Moment-generating function** of the RV X: alternative procedure for determining moments

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

- moment-generating functions will exist only if the integral in the above definition converges
- if it exists, a moment-generating function of RV X can be used to generate all the moments of that variable using the below method:

Theorem 27 Let X be a random variable with moment-generating function $M_X(t)$, then

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \mu_r'.$$

Theorem 28 **Uniqueness Theorem**: Let X and Y be two RVs with moment-generating functions $M_X(t)$ and $M_Y(t)$. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the same probability distribution.

Theorem 29

1. $M_{X+a}(t) = e^{at} M_X(t)$
2. $M_{aX}(t) = M_X(at)$

Theorem 30 If $X_1, \ldots, X_n$ are independent RVs with moment-generating functions, and $Y = X_1 + \ldots + X_n$ then

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) \ldots M_{X_n}(t).$$

Subsection 28.1

Linear Combinations of Random Variables

Theorem 31 If $X_1, X_2, \ldots, X_n$ are independent RVs having normal distributions with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then the RV

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n.$$

has a normal distribution with mean

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n.$$

and variance

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2.$$

PART VII

Sampling

Why sample?

- can't measure entire population
- random sampling ensures sample reflects population

Section 29

Measures of Location: Sample Mean and Median

- Given sample data: $x_1, \ldots, x_n$

Definition 42 **Sample Mean**: the numerical average

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Definition 43 **Sample Median**: given that $x_1, x_2, \ldots, x_n$ are arranged in increasing order of magnitude,

$$x_m = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{if } n \text{ even} \end{cases}.$$

- purpose of the sample median is to reflect the central tendency of the sample in a way that is uninfluenced by extreme values or outliers (unlike the mean)

Definition 44 **Mode**: most frequently occurring value

- e.g. 1, 1, 1, 1, 1, 8
- mode = 1

Section 30

Measures of Variability

Subsection 30.1

Sample Range and Sample Standard Deviation

- **sample range** is given by $X_{max} - X_{min}$

Definition 45 **Sample Variance**, denoted by s^2, is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2.$$

The **sample standard deviation**, denoted by s, is the positive square root of s^2:

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

- the quantity $n-1$ is often called the **degrees of freedom associated with the variance**
- this is because in general, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, so the last value in the sample can be determined only using the first $n-1$ values
- this means the computation of sample variance doesn't involve all n independent squared deviations from the mean
- there are only $n-1$ "pieces of information" that produce s^2
- therefore: there are $n-1$ degrees of freedom instead of n when computing sample variance

SECTION 31

Visualization

SUBSECTION 31.1

Histogram

- plots the frequency of each outcome
- also can plot relative frequency:
- dividing each class frequency by the total number of observations, we obtain the relative frequency of each class interval
- we can plot relative frequency in a histogram

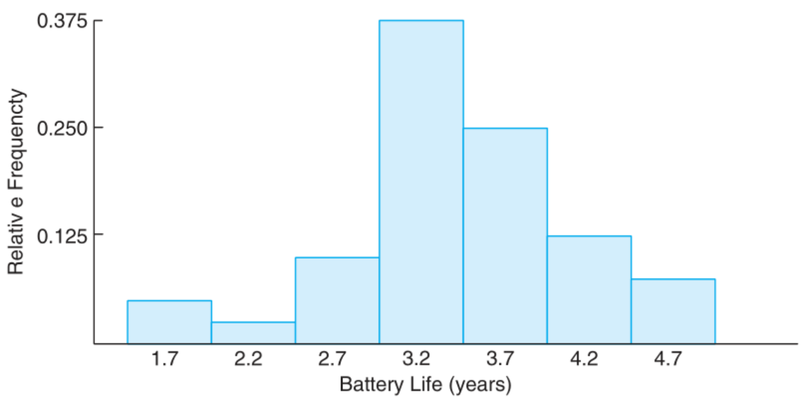

Figure 3. Example of relative frequency histogram.

SUBSECTION 31.2

Box-and-Whisker Plot

- displays the center of location, variability, and degree of asymmetry

- encloses the *interquartile range* of the data in a box which also displays the median within
    - essentially encloses the middle 50% of the data
- the extremes of the interquartile range are the 75th percentile (upper quartile) and 25th percentile (lower quartile)
- "whiskers" extend from the sides of the box showing extreme observations
- outliers may be plotted as points as well

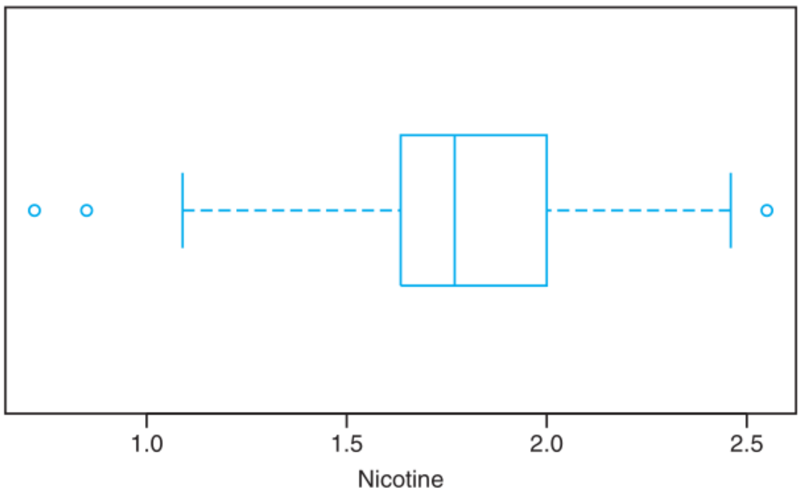

Figure 4. Example of box and whisker plot.

PART VIII

Sampling Distributions

Section 32

Random Sampling

Definition 46 A **population** consists of all possible observations.

Definition 47 A **sample** is a subset of a population.

- we work with samples because it is impractical to observe whole population

Definition 48 Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability as

$$f(x_1, x_2, \ldots x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

Section 33

Statistics and Sampling Distributions

Definition 49 **Statistic**: Any function of the random variables constituting a random sample (e.g. mean, median, variance, etc.).

- a sample is biased if it consistently over- or underestimates a statistic of interest

Definition 50 **Sampling Distribution**: The probability distribution of a statistic.

- the sampling distribution of a statistic depends on the distribution of the population, size of samples, and method of choosing samples
- it is **very important to notice and understand (for pretty much the rest of the course) that the mean and variance of a *statistic* is not the same as the mean and variance for the *population***

Section 34

Sampling Distribution of Means and the Central Limit Theorem

- the sampling distribution of $\overline{X}$ with sample size n is the distribution that results when an experiment is conducted over and over (always with sample size n) and many values of $\overline{X}$ result
- the sampling distribution describes the variability of sample averages around the population mean μ
- if $X_1, \ldots, X_n$ are normal, all with mean μ and variance σ^2, then $\overline{X}$ has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{1}{n}(\mu_1 + \ldots + \mu_n) = \mu.$$

  and variance

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}(\sigma_1^2 + \ldots + \sigma_n^2) = \frac{\sigma^2}{n}.$$

- notice that **the mean and variance of the *statistic* (denoted above as $\mu_{\overline{X}}, \sigma^2_{\overline{X}}$) is not the same as the mean and variance for the *population* (denoted above as μ, σ^2)**

Subsection 34.1

The Central Limit Theorem (CLT)

Theorem 32 **Central Limit Theorem**: Suppose $\overline{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2. Define an RV

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

As $n \to \infty$, the distribution of Z converges to the standard normal distribution, $n(z; 0, 1)$.

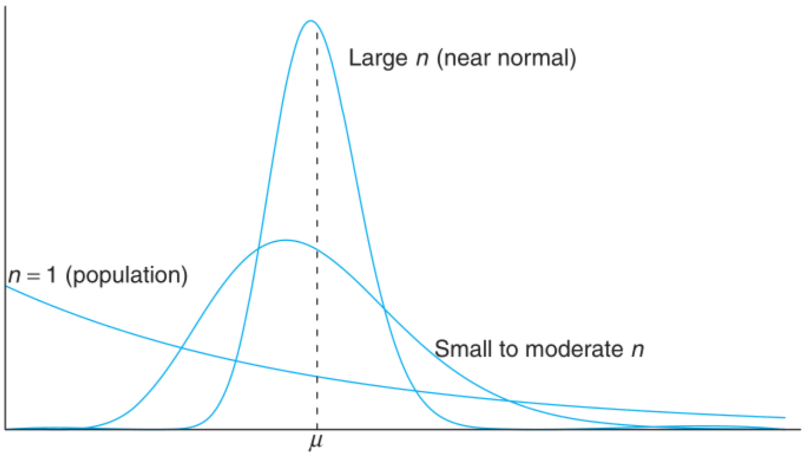

Figure 5. Illustration of the Central Limit Theorem. Note that it shows how the mean of $\overline{X}$ remains μ for any sample size and the variance gets smaller as n increases.

- CLT is widely applicable – works with any distribution as long as observations have same probability distributions with finite variance
- variance of mean shrinks with $\sqrt{n}$ – average becomes more accurate with a bigger sample

Subsection 34.2

Sampling Distribution of the Difference Between Two Means

Theorem 33 If independent samples of size n_1 and n_2 are drawn at random from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Therefore,

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

is approximately a standard normal variable.

Section 35

Sampling Distribution of Variance

Theorem 34 If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sigma^2}.$$

has a chi-squared distribution with $\nu = n - 1$.

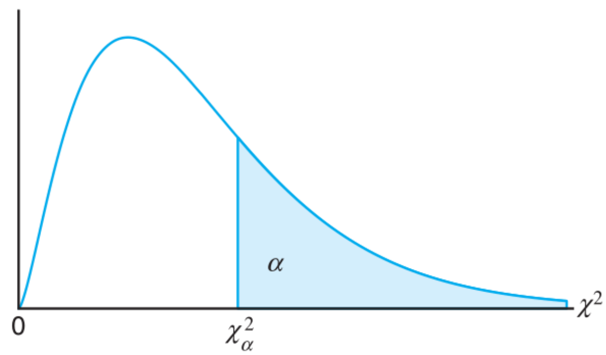

Figure 6. The chi-squared disribution. We let χ^2_α represent the χ^2 value above which we find an area of α.

Subsection 35.1

Degrees of Freedom as a Measure of Sample Information

- recall that

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

has a chi-squared distribution with n *degrees of freedom* while the RV

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sigma^2}.$$

has a chi-squared distribution with $n-1$ *degrees of freedom*

- this is because when μ is not known, i.e. when we are considering the distribution of

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sigma^2}.$$

there is 1 less degree of freedom since a degree of freedom is lost in the estimation of μ (when μ is replaced by $\overline{x}$)

- when μ is known, there are n degrees of freedom, or independent *pieces of information*, in a random sample from a normal distribution

- when data (values in the sample) are used to compute the mean, there is 1 less DOF, 1 less piece of information, used to estimate σ^2

Section 36

t-Distribution

- CLT is for making inferences about the mean μ assuming variance σ^2 is known
- t-distribution is for when σ^2 is not known
- consider the statistic

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

where

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

- if sample size is large ($n \geq 30$), S is close to σ, and T follows a normal distribution
- if sample size is smaller, the values of S^2 fluctuate considerably and the disribution of T deviates much more from the standard normal distribution
- the t-distribution is much more accurate in this case

Definition 51 ***t*-distribution**:

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

- if $X_1, \ldots, X_n$ are independent RVs that are all normal with mean μ and standard deviation σ, and

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

then the RV $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom

- intuition behind the t-distribution:
  - if we knew σ, we'd have a normal distribution
  - instead we only have estimate S^2 – less information, so we expect more variability
- variance of T depends on the sample size n and is always greater than 1
- in the limit that sample size $n \to \infty$ and subsequently $v \to \infty$, the t-distribution becomes the standard normal distribution

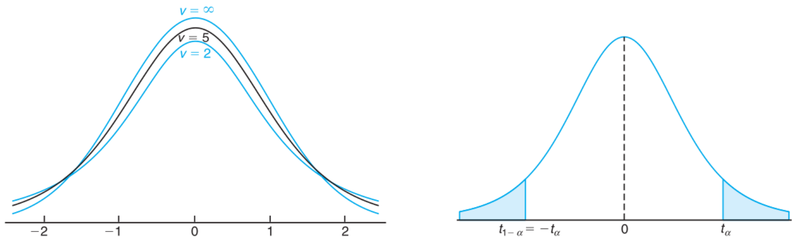

Figure 7. Illustration of the t-distribution. We let t_α represent the t-value above which we find an area equal to α.

Section 37

F-Distribution

- define a statistic F to be the ratio of two independent chi-squared RVs U and V, each divided by its number of degrees of freedom, v_1 and v_2:

$$F = \frac{U/v_1}{V/v_2}.$$

Definition 52

***F*-distribution**: sampling distribution of F is given by

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0 \\ 0, & f \leq 0 \end{cases}.$$

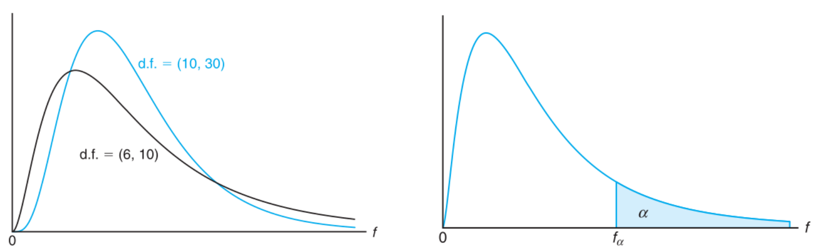

Figure 8. Typical F-distributions. We let f_α be the f-value above which we find an area equal to α.

Subsection 37.1

The F-Distribution with Two Sample Variances

Theorem 35

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom

Section 38

Quantile and Probability Plots

- quantile plots depict (in sample form) the cumulative distribution function

Definition 53 A **quantile** of a sample, denoted by $q(f)$, is a value for which a specified fraction of f of the data values is less than or equal to $q(f)$.

- a **quantile plot** plots $q(f)$ versus f, where $q(f)$ is on the y-axis
- to sketch:
  - rank sample in increasing order, $x_1, \ldots, x_n$
  - for each data point $i = 1, \ldots, n$, plot

$$\left(\frac{i - 3/8}{n + 1/4}, x_i\right).$$

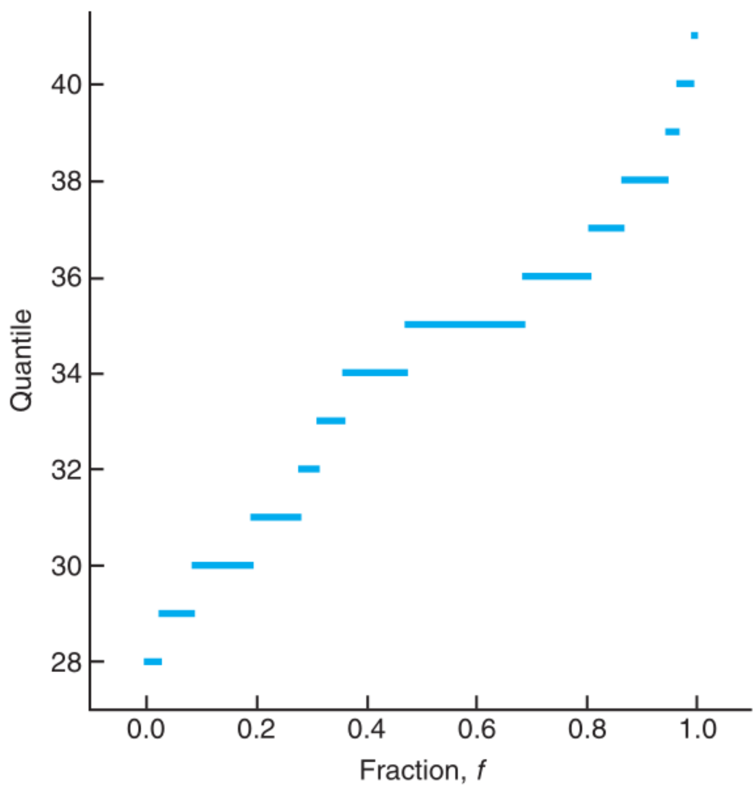

Figure 9. Example of a quantile plot.

- notice in Figure 9:
  - $q(0.5)$ is the sample median
  - lower quartile (25th percentile) is $q(0.25)$
  - upper quartile (75th percentile) is $q(0.75)$
  - flat areas indicate clusters of data
  - steep areas indicate sparsity of data

Subsection 38.1

Normal Quantile-Quantile Plot

- we often want to know how close data is to a normal distribution since we understand normal distributions very well and many tools (t and F distributions) assume normality
- the expression for the quantile of a normal distribution is very complicated but can be approximated as

$$q_{\mu,\sigma}(f) = \mu + \sigma(4.91(f^{0.14} - (1-f)^{0.14})).$$

Definition 54 **Normal quantile-quantile plot**: a plot of $y_{(i)}$ (ordered observations) against $q_{0,1}(f_i)$, where $f_i = \frac{i-3/8}{n+1/4}$.

- if the curve is straight, the data is roughly normal

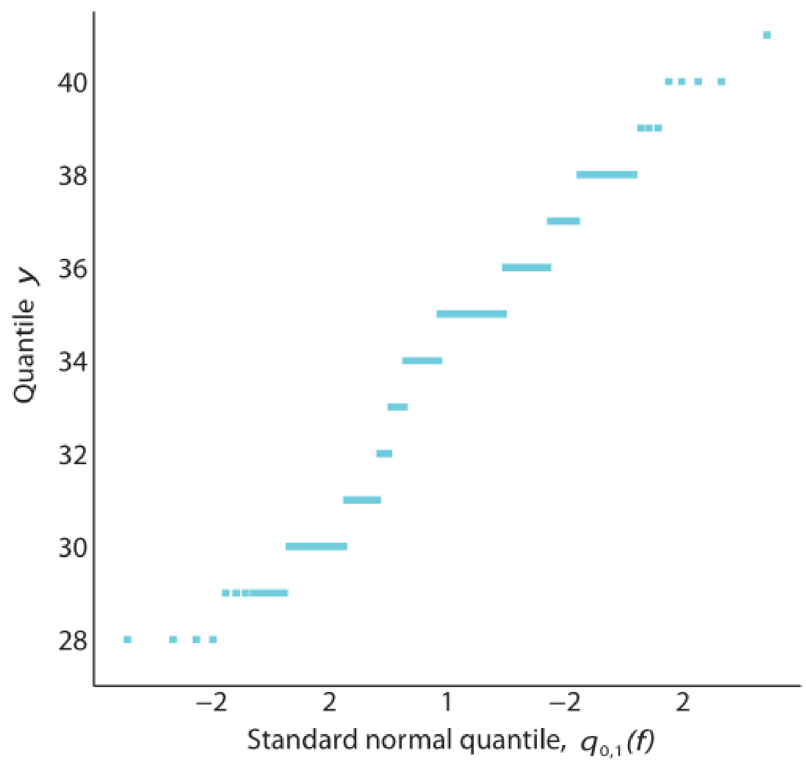

Figure 10. Example of a normal quantile-quantile plot.

PART

IX

Estimation

SECTION 39

Classical Methods of Estimation

- in general, given a sample, we write
  - θ is the true parameter of the population (like μ)
  - $\hat{\theta}$ is the observed value from the sample (like $\overline{x}$)
  - $\hat{\Theta}$ is the sample statistic (like $\overline{X}$)

Definition 55 A **point estimate** of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

- for example: the value $\overline{x}$ of the statistic $\overline{X}$, computed from a sample of size n, is a point estimate of the population parameter μ.

SUBSECTION 39.1

Unbiased Estimator

Definition 56 A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

SUBSECTION 39.2

Variance of a Point Estimator

Definition 57 Considering all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the **most efficient estimator** of θ.

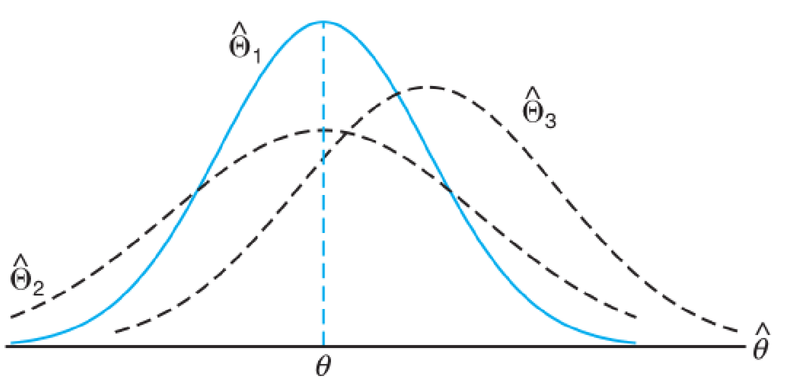

Figure 11. Sampling distributions of different estimators of θ.

Subsection 39.3

Interval Estimation

- a point estimate $\hat{\theta}$ is rarely exactly θ
- it's useful to have an interval, $\hat{\theta_L} \leq \theta \leq \hat{\theta_U}$
- $\hat{\theta_L}$ and $\hat{\theta_U}$ depend on the value of the statistic $\hat{\Theta}$ for a particular sample and also on the sampling distribution of $\hat{\Theta}$
- we want to make a statement of the form

$$P(\hat{\Theta_L} < \theta < \hat{\Theta_U}) = 1 - \alpha$$

  for $0 < \alpha < 1$
- the interval $\hat{\theta_L} \leq \theta \leq \hat{\theta_U}$, computed from the selected sample, is called a $100(1-\alpha)\%$ **confidence interval**
  - e.g. if $\alpha = 0.05$, we have a 95% confidence interval
- the fraction $1-\alpha$ is called the **confidence coefficient** or **degree of confidence**
- the endpoints of the interval are called the lower and upper **confidence limits**

Section 40

Single Sample: Estimating the Mean

- Setup:
  - n samples
  - observed mean $\overline{x}$
  - known variance σ^2
  - and the statistic $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$
- then

$$1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}),$$

  where

$$z_\beta = -\Phi^{-1}(\beta)$$

  and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Theorem 36 If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

- i.e.

$$\overline{X}_L = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \overline{X}_U = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Theorem 37 If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specificed amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Subsection 40.1

One-Sided Confidence Intervals

- if we want a confidence interval of the form

$$1-\alpha = P(Z \leq z_\alpha)$$

  we set

$$z_\alpha = -\Phi^{-1}(\alpha).$$

- then:

$$\overline{X}_U = \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Subsection 40.2

Estimates with Unknown σ

- samples from normal distribution with unknown σ and any n, we use t-distribution

Theorem 38 If $\overline{x}$ and s are the mean and standard deviation of a random sample from a normal population with an unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

- one-sided confidence intervals (upper and lower $100(1-\alpha)\%$ confidence intervals) for μ with unknown σ are

$$\overline{x} + t_\alpha \frac{s}{\sqrt{n}} \quad \text{and} \quad \overline{x} - t_\alpha \frac{s}{\sqrt{n}}.$$

Section 41

Standard Error of a Point Estimate

- given samples $X_1, \ldots, X_n$ drawn from an unknown distribution with variance σ
- as $n \to \infty$, the distribution of $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ approaches $n(z;0,1)$, the standard normal
- this implies that the standard deviation of Z is around $\frac{\sigma}{n}$

Definition 58 The **standard error** of an estimator is its standard deviation.

- e.g. the standard of error of $\overline{X}$ is $\sigma/\sqrt{n}$

- recall the $100(1-\alpha)\%$ confidence intervals for the mean:

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{is written as} \quad \overline{x} \pm z_{\alpha/2}\text{s.e.}(\overline{x})$$

  where "s.e." is the standard error

- the width of confidence intervals depends on the confidence and the standard error

SECTION 42

Prediction Intervals

- so far we've been give samples and $\overline{X}$ and characterized the error/uncertainty of $\overline{X}$
- we now want to predict the value of a future observation
- suppose we have normal samples $X_1, \dots, X_n$, each with known variance σ and a sample mean of $\overline{X}$
- $\overline{X}$ is a good point estimate of a single new sample X_0
- the error of the point estimate is $X_0 - \overline{X}$
- due to independence, the variance of the error is $\sigma^2 + \sigma^2/n$
- we define the statistic

$$Z = \frac{X_0 - \overline{X}}{\sigma\sqrt{1 + 1/n}}.$$

- the distribution of Z is $n(z; 0, 1)$
- therefore we can write the probability statement

$$1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}).$$

Theorem 39 For a normal distribution of measurements with unknown mean μ and known variance σ^2, a $100(1-\alpha)\%$ **prediction interval** of a future observation x_0 is

$$\overline{x} - z_{\alpha/2}\sigma\sqrt{1+1/n} < x_0 < \overline{x} + z_{\alpha/2}\sigma\sqrt{1+1/n},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Theorem 40 For a normal distribution of measurements with unknown mean μ and unknown variance σ^2, a $100(1-\alpha)\%$ **prediction interval** of a future observation x_0 is

$$\overline{x} - t_{\alpha/2}s\sqrt{1+1/n} < x_0 < \overline{x} + t_{\alpha/2}s\sqrt{1+1/n},$$

where $t_{\alpha/2}$ is the t-value with $v = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

- **outlier detection**: if a new observation is outside the prediction interval, we can declare it an outlier

SECTION 43

Tolerance Limits

- we want to define bounds that cover a fixed proportion of the measurements

Definition 59 For a normal distribution of measurements with unknown mean μ and unknown standard deviation σ, **tolerance limits** are given by $\overline{x} \pm ks$, where k is determined such that one can assert with $100(1-\gamma)\%$ confidence that the given limits contain at least the proportion of $1-\alpha$ of the measurements.

- values of k given in a provided table

SUBSECTION 43.1

Comparison of Intervals

- Confidence Intervals:
  - setup: independent observations of RVs, $x_1, \ldots, x_n$, and the mean is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - there is a $100(1-\alpha)\%$ chance the true mean μ is in an interval around $\overline{x}$
  - use CLT to compute interval when we know σ for each observation or n is large
  - use t-distribution when σ is unknown
- Prediction Intervals:
  - same setup
  - $100(1-\alpha)\%$ chance the next observation x_0 is in an interval around $\overline{x}$
  - compute from t or normal distribution
- Tolerance Limits:
  - same setup
  - $100(1-\gamma)\%$ of measurements will be in an interval $\overline{x} \pm ks$
  - compute from table

SECTION 44

Two Samples: Estimating the Difference between Two Means

- if we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2
- then a point estimator of the difference between μ_1 and μ_2 is given by the statistic $\overline{X}_1 - \overline{X}_2$
- therefore to obtain a point estimate of $\mu_1 - \mu_2$, we select independent random sample from each population of sizes n_1 and n_2 and compute $\overline{x}_1 - \overline{x}_2$ (the difference of the sample means)
- the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normally distributed with mean $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

- therefore, we define a standard normal variable

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

- and assert that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Theorem 41 **Confidence Interval for $\mu_1 - \mu_2$, Known Variances**: If $\overline{x}_1$ and $\overline{x}_2$ are means of independent random samples of sizes n_1 and n_2 from populations with known variances σ_1^2 and σ_2^2, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

SUBSECTION 44.1

Two Samples with Unknown Variance

44.1.1 Equal Variances

Definition 60 **Pooled Estimate of Variance**: the sample size-weighted average of S_1^2 and S_2^2

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

- define the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}.$$

- then T has the t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom
- we have

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

Theorem 42 **Confidence Interval for $\mu_1 - \mu_2$, Equal but Unknown Variances**: a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2}s_p\sqrt{1/n_1 + 1/n_2} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2}s_p\sqrt{1/n_1 + 1/n_2},$$

where s_p is the pooled estimate of the population standard deviation and $t_{\alpha/2}$ is the t-value with $v = n_1 + n_2 - 2$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

44.1.2 Different Variances

- if the variances are unknown and different, we use the statistic

$$T' = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

- then T' approximately has a t-distribution with

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

degrees of freedom

Section 45

Paired Observations

- previously we considered two sample populations of different sizes and all measurements were independent
- now we consider two sample populations of the same size and pairs of observations have one measurement from each population
  - e.g. measuring before and after observations on n people – each "before" measurement is paired with an "after" measurement
- consider paired samples $(X_i, Y_i), i = 1, \ldots, n$ with statistics $\mu_X, \sigma_X, \ldots$ (same for Y_i's)
- we are interested in the difference $D_i = X_i - Y_i$
- the variance of the difference:

$$\text{var}(D_i) = \text{var}(X_i - Y_i) = \sigma_X^2 + \sigma_Y^2 - 2\text{cov}(X_i, Y_i).$$

- we expect $\text{cov}(X_i, Y_i) \geq 0$, e.g. the before and after weights for one person is likely both above μ_X and μ_Y
- pairing helps reduce variance
- we can then apply usual CLT or t-distribution confidence intervals to sample D_i

Section 46

Single Sample: Estimating a Proportion

- a point estimator of the proportion p in a binomial experiment is given by the staistic $\hat{P} = X/n$
- X represents the number of successes in n trials
- the sample proportion $\hat{p} = x/n$ is used as the point estimate of the parameter p
- by CLT, for n sufficiently large, $\hat{P}$ is approximately normally distributed with mean p and variance $\frac{pq}{n}$
- we can use the statistic

$$Z = \frac{\hat{P} - p}{\sqrt{pq/n}}.$$

Theorem 43 **Large Sample Confidence Intervals for** p: If $\hat{p}$ is the proportion of successes in a random sample of size n and $\hat{q} = 1 - \hat{p}$, an approximate $100(1-\alpha)\%$ confidence interval for the binomial parameter p is given by (method 1)

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

or by the limits (method 2)

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

Theorem 44 If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

Subsection 46.1

Choice of Sample Size

- suppose we want to be $100(1-\alpha)\%$ confident that the error is less than some specified amount e

Theorem 45 If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will be less than a specified amount e when the sample size is approximately

$$n = \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}.$$

- but there's a catch: $\hat{p}$ depends on n
- if we want a safe lower bound for n :

Theorem 46 If $\hat{p}$ is used as an estimate of p, we can be **at least** $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \frac{z^2_{\alpha/2}}{4e^2}.$$

Section 47

Single Sample: Estimating the Variance

- an interval estimate of σ^2 can be established using the statistic

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

Theorem 47 **Confidence Interval for σ^2:** If s^2 is the variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2-values with $v = n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$ to the right.

Section 48
Maximum Likelihood Estimation

- so far, we've used intuitive sampling statistics:
  - $\overline{X}$ for μ, S^2 for σ^2, $\hat{P}$ for p
- but sometimes it is not obvious what the proper estimator for parameters should be
  - e.g. degrees of freedom, α and β in gamma distribution, etc.
- one of the most important approaches to estimation in statistical inference: **method of maximum likelihood**

Subsection 48.1
The Likelihood Function

- the method of maximum likelihood is that for which the likelihood function is maximized
- main philosophy: the reasonable estimator of a parameter based on sample information is the parameter value that produces the largest probability of obtaining the sample – given a sample, what was the parameter value that most likely produced it
- the likelihood of a sample for a certain value of a parameter is simply the joint distribution of the random variables for a certain value of that parameter, i.e.

$$P(X_1 = x_1, \ldots, X_n = x_n | \theta) = f(x_1, \theta) \ldots f(x_n, \theta).$$

Definition 61 Given independent observations $x_1, x_2, \ldots, x_n$ from a probability density function (continuous case) or probability mass function (discrete case) $f(\boldsymbol{x}; \theta)$, the maximum likelihood estimator (MLE) $\hat{\theta}$ is that which maximizes the **likelihood function**

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

i.e.

$$\hat{\theta} = \max_{\theta} L(x_1, \ldots, x_n; \theta).$$

- it is often convenient to work with the natural log of the likelihood function in finding its maximum
- for example, given an RV X with a gamma distribution and a sample $x_1, \ldots, x_n$, how can we estimate α and β?
- using MLE:

$$\hat{\alpha} = \max_{\alpha} \prod_{i=1}^{n} f(x_i; \alpha, \beta),$$

and same applies to β

PART

X

Hypothesis Testing

SECTION 49

Statiscal Hypotheses: General Concepts

Definition 62 **Statistical Hypothesis**: an assertion or conjecture concerning one or more populations (a special case of more general hypotheses)

- rejection of a hypothesis implies that the sample evidence refutes it – there is a extremely small probability of obtaining the sample information observed if the hypothesis was true (therefore it isn't)
- typically, a contention is reached via a rejection of an opposing hypothesis

SUBSECTION 49.1

The Null and Alternative Hypotheses

- **null hypothesis**: any hypothesis we wish to test, denoted by H_0
- the rejection of H_0 leads to the acceptance of the **alternative hypothesis**, denoted by H_1
- the alternative hypothesis H_1 usually represents the *question to be answered or the theory to be tested* (its specification is crucial)
- the null hypothesis nullifies or opposes the alternative hypothesis (they are often logical complements)
- a data analyst arrives at one of two conclusions:
  1. **reject** H_0 in favour of H_1 because of sufficient evidence in the data, or
  2. **fail to reject** H_0 because of insufficient evidence in the data
- for example: innocent until proven guilty:
  - H_0 is innocent, H_1 is guilty
  - if there is strong enough evidence pointing to guilty, we reject H_0 in favour of H_1
  - if evidence is weak, we fail to reject H_0
  - **key point**: we are not proving innocence, we are *failing to reject* innocence
- **hypothesis testing** is using confidence intervals and logic to draw these conclusions

Section 50

Testing a Statistical Hypothesis

Definition 63 **Type I error**: rejection of the null hypothesis when it's actually true (false positive)

- **level of significance**: probability of committing a type I error

$$\alpha = P(\text{type I error}).$$

Definition 64 **Type II error**: nonrejection of the null hypothesis when it's actually false (false negative)

- **level of significance**: probability of committing a type II error

$$\beta = P(\text{type II error}).$$

- the probability of committing both types of error can be reduced by increasing the sample size

Example: mean weight of male students in a college

- our null and alternative hypotheses: H_0 is $\mu = 68$ kg, H_1 is $\mu \neq 68$ kg
- we encounter a first issue: $P(H_0) = P(\mu = 68) = 0$
  - then H_0 will almost always be rejected
- to solve this we use a **critical region** – a range leading to rejection of H_0:
  - if $67 < \overline{x} < 69$, don't reject H_0
  - critical region is the complement of [67, 69]

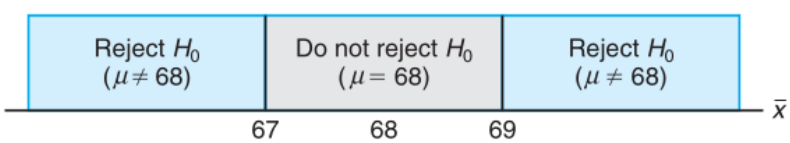

Figure 12. Critical region shown in blue.

- we now calculate the probabilties of committing type I and type II errors
- assume sample size of $n = 36$
- we assume that standard deviation of the population of weights is $\sigma = 3.6$
  - for larger samples, we may substitute s for σ if no other estimate of σ is available
- our decision statistic will be $\overline{X}$, the most efficient estimator of μ
- from the CLT, we know the sampling distribution of $\overline{X}$ is approximately normal with standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$

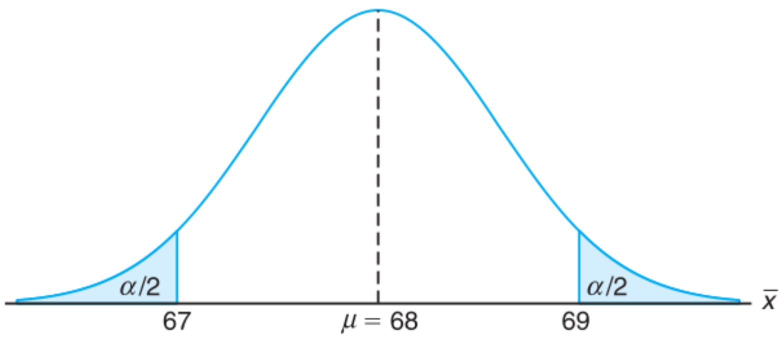

Figure 13. Probability of a type I error.

- the probability of committing a type I error is given by

$$\alpha = P(\overline{X} < 67) + P(\overline{X} > 69) \quad \text{when } \mu = 68.$$

- converting to z-values and looking at tables for normal distribution, we find that $\alpha = 0.0950$
- interpretation: 9.5% of all samples of size 36 would lead us to reject $\mu = 68$ kg when in fact it is true
- to reduce α, we can increase sample size or widen the fail-to-reject region
  - if we increase sample size to 64, then repeating the calculations, we obtain $\alpha = 0.0264$
- but reduction in α is not sufficient by itself to guarantee good testing procedure, we must also evaluate β for various alternative hypotheses
- if it's important to reject H_0 when the true mean is some value $\mu \geq 70$ or $\mu \leq 66$, then the probability of committing a type II error should be calculated for alternatives $\mu = 66$ and $\mu = 70$
  - due to symmetry, it's only necessary to consider one case
- a type II error occurs when $67 < \overline{x} < 69$ when H_1 is true:

$$\beta = P(67 \leq \overline{X} \leq 69 \text{ when } \mu = 70).$$

- by calculating z-values and looking at tables, we obtain $\beta = 0.0132$ (same result if the true value of μ was 66)
- again, the value of β can be decreased if sample size n is increased
- the probability of committing a type II error increases rapidly when the true value of μ approaches (but is not equal to) the hypothesized value
  - for example, if the alternative hypothesis $\mu = 68.5$ is true:

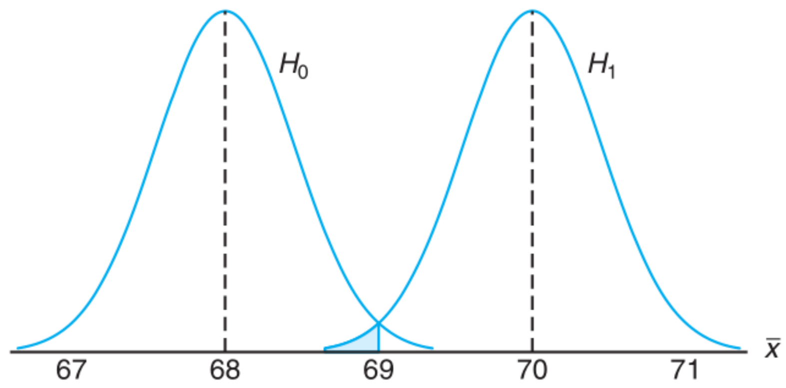

Figure 14. Probability of type II error for testing $\mu = 68$ versus $\mu = 70$.

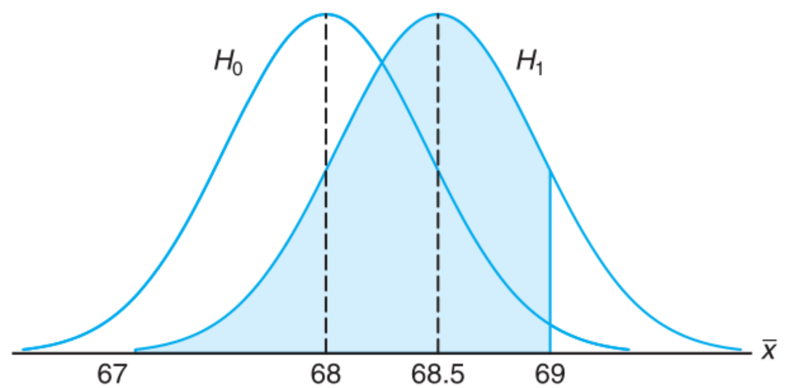

Figure 15. Probability of type II error for testing $\mu = 68$ vs $\mu = 68.5$

Theorem 48 **Important Properties of a Hypothesis Test**:

1. type I and type II error are related (decrease in probability of one generally results in an increase in the probability of the other)
2. size of the critical region (and therefore the probability of committing a type I error) can always be reduced by adjusting the critical values
3. increase in sample size will reduce α and β
4. if the null hypothesis H_0 is false, β is maximized when the true value of a parameter approaches the hypothesized value (the greater the distance between the true and hypothesized value, the smaller it will be)

Definition 65 **Power** of a test: the probability of rejecting H_0 given that a specific alternative is true.

- computed as $1 - \beta$

Subsection 50.1

One- and Two-Tailed Tests

- **one-tailed test**: a test of a statistical hypothesis where the alternative is **one sided**
  - for example:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0.$$

- **two-tailed test**: a test of a statistical hypothesis where the alternative is **two sided**
  - for example:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0.$$

SECTION 51

P-Values

- so far, we are either in or out of a predetermined critical region
- but it is also important to know the probability of an outcome occurring, or something else that is equal or even rarer, given that H_0 is true
- it gives the analyst an alternative (in terms of a probability) to a mere 'reject' or 'do not reject' conclusion

Definition 66 **P-value**: the probability of generating the observed data or something else that is equal or rarer, given that H_0 is true

- tells us the probability of a test statistic being as extreme or more extreme than the measured value
- textbook definition: the lowest level (of significance) at which the observed value of the test statistic is significant

SUBSECTION 51.1

P-Values vs Classic Hypothesis Testing

- there are differences in approach and philosophy of these two methods
- when using P-values, there is no fixed α determined and conclusions are drawn on the basis of the size of the P-value together with subjective judgement of the analyst
- their approaches are summarized below

Theorem 49 **Approach to Hypothesis Testing with Fixed Probability of Type I Error**:

1. State null and alternative hypotheses
2. choose a fixed significance level α
3. Choose an appropriate test statistic and establish the critical region based on α

4. Reject H_0 if the computed test statistic is in the critical region, otherwise don't reject
5. Draw conclusions

Theorem 50 **Significance Testing (P-value) Approach**:

1. state null and alternative hypotheses
2. Choose an appropriate test statistic
3. Compute P-value based on the computed value of the test statistic
4. Use jedgement based on the P-value and knowledge of the scientific system

Example:

- hypothesis: H_0 is $\mu = 5$, H_1 is $\mu \neq 5$
- Sample data:
  - $n = 40$ samples
  - $\overline{x} = 5.5$
  - $s \approx \sigma = 1$
- using **classic hypothesis testing**
  - use a fixed probability of a type I error $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$
  - compute
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = 3.16.$$
  - since this is outside of [-1.96, 1.96], we reject H_0
- using **P-value approach**
  - the P-value is the probability of something equally or more rare occurring:
$$P = 2P(Z > 3.16) = 0.0016.$$
  - so H_0 is very unlikely

SECTION 52

Goodness-of-Fit Test

- so far we have only looked at testing statistical hypotheses about single population parameters such as μ and σ^2
- now we consider a test to determine if a population has a specified theoretical distribution
- the test is based on how good a fit there is between the frequency of occurrence of observations in a sample and the expected frequencies obtained from the hypothesized distributions

Definition 67 **Goodness-of-Fit Test**:

- setup:
  - discrete RV with possible outcomes $i = 1, \ldots, k$
  - n trials
  - $e_i = nP(i)$ is the expected frequency of outcome $i = 1, \ldots, k$
  - o_i is the observed frequency of i (O_i is the RV)

Let

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}.$$

- distribution of χ^2 is approximated very closely by the chi-squared distribution with $v = k - 1$ degrees of freedom
- **small χ^2 indicates a good fit** (large is bad fit)
- number of degrees of freedom is equal to $k - 1$ since there are only $k - 1$ freely determined frequencies (the last frequency is determined by the others)

- since large values of χ^2 indicates a poor fit which leads to rejection of H_0, the critical region will fall in the right tail of the chi-squared distribution
- for a level of significance of α, we find the critical value χ^2_α from textbook Table A.5, then $\chi^2 > \chi^2_\alpha$ is the critical region
- note: this decision criterion shouldn't be used unless each of the expected frequencies is ≥ 5

PART

XI

Linear Regression and Correlation

Section 53

Function Approximation

- Basic setup:
  - input/output pairs: $(x_i, y_i), i = 1, \ldots, n$
  - we want a function $y = f(x)$ that minimizes errors $e_i = y_i - f(x_i)$
- types of function approximators:
  - linear: $y = ax + b$
  - nonlinear (kernel regression, splines, neural networks)
  - classification
    * $x \in \mathbb{R}, y \in 0, 1$
    * support vector machine

Section 54

Linear Regression with Least Squares

- we want to fit a linear function $y = ax + b$ to the data
- the errors are $e_i = y_i - ax_i - b, i = 1, \ldots, n$
- the total sqaured error is given by

$$\mathcal{E} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

- we want to minimize total squared error, so we solve for a and b that minimize $\mathcal{E}$:
  - we differentiate with respect to a and b and set equal to 0:

$$\frac{d\mathcal{E}}{da} = \sum_{i=1}^{n} \frac{d}{da}(y_i - ax_i - b)^2 = 0$$
$$\frac{d\mathcal{E}}{db} = \sum_{i=1}^{n} \frac{d}{db}(y_i - ax_i - b)^2 = 0.$$

- rearranging these equations, we get the 'normal equations' which can be solved to yield computing formulas for a and b:

Theorem 51 **Estimating the Regression Coefficients:** Given the sample $(x_i, y_i); i = 1, \ldots, n$, the least squares estimates of the regression coefficients a and b are computed from the formulas

$$a = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$b = \frac{\sum_{i=1}^n y_i - a\sum_{i=1}^n x_i}{n} = \overline{y} - a\overline{x}.$$

Interpretation:

- the errors are the vertical deviations from the line to each data point

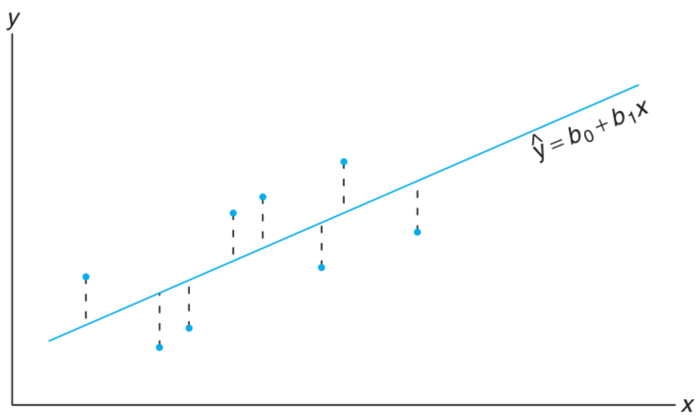

Figure 16. Errors/residuals as vertical deviations.

- Deming regression uses geometric distance – better for independent errors in x_i and y_i
- for linear regression we don't need to know σ, but for Deming we need to know the ratio of variances of x and y errors

Example with Maximum Likelihood:

- recall MLE
- in this case, each error e_i is a realization of a normal RV E_i with $\mu = 0$ and variance σ^2
  - this implies that each y_i is also an RV with a mean of $ax_i - b$ and variance σ^2
- parameters: $\theta = (a, b)$
- likelihood function:

$$L(e_1, \ldots, e_n; a, b) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{e_i^2/2\sigma^2}.$$

- maximizing this function over (a, b) gives the least squares solution

SECTION 55

Properties of the Least Squares Estimators

Conclusion:

- if we assume $y_i = ax_i + b + e_i$
- e_i is a realization of the normal RV E_i with $\mu = 0$ and variance σ^2
- y_i is also a realization of RV Y_i and is a function of E_i
- then a and b are realizations of RVs A and B
- we see that A and B are **unbiased estimators of the true coefficients α and β**

SUBSECTION 55.1

Estimating the Error

- if $y_i = ax_i + b + e_i$ and e_i is a realization of an RV E_i with variance σ^2, then σ^2 reflects random variation or experimental error variation around the regression line
- the total squared error is
$$\sum_{i=1}^{n} e_i^2.$$
- we define the statistic
$$S^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-2}.$$
- this is an unbiased estimator of σ^2
- if we denote the regression estimate as $\hat{y}_i = ax_i + b$ then $e_i = y_i - \hat{y}_i$
- we can write a realization of this as the following:

Theorem 52 An unbiased estimate of σ^2 is
$$\begin{aligned} s^2 &= \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} \\ &= \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \alpha \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-2}. \end{aligned}$$