

ECE286: Probability & Statistics

Introduction (1.1, 2.1-2, 2.4-7)

(1) Sets and Counting

Tips

- If stuck on a question, use a Venn diagram or draw out the region.
- Always check if it's a PMF from the common discrete distributions.
- Think if independent (multiply probabilities), dependent sequential events (multiply probabilities that adjust based on the events changing #'s in probability), mutually exclusive (sum), or both.
 - **Note:** Most times, you are applying the concept of total probability theorem and probability of mutually exclusive events.

L1: Coin Flip, Sets, & Events

1.2 Sample Space S

The set of **all possible outcomes** of a statistical experiment. Denoted S.

- **Types:**
 - **Finite Sample Spaces:** When the number of outcomes is limited.
 - **Infinite Sample Spaces:** When the number of outcomes is unlimited.

1.3 Event

A subset of a sample space S. An event is any outcome or combination of outcomes.

- **Types:**
 - **Simple Event:** An event with a single outcome
 - **Compound Event:** An event with more than one outcome.

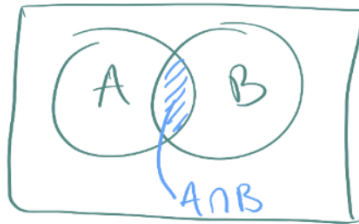
1.5 Complement

The **complement** of an event A w.r.t S is the subset of all elements of S that are not in A. Denote A' .



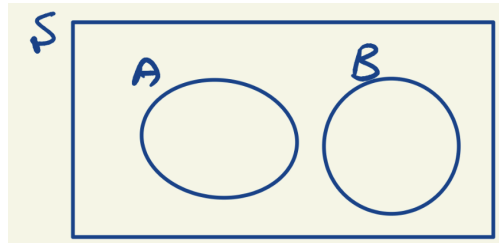
1.6 Intersection

The **intersection** of two events A and B is the event containing all elements that are common to A and B. Denoted $A \cap B$.



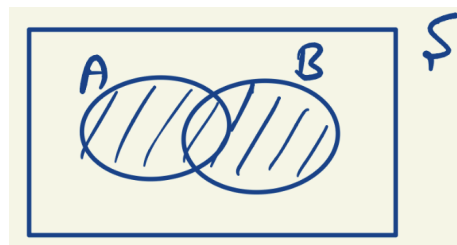
1.7 Mutually Exclusive

Two events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$, that is, if A and B have no elements in common.



1.8 Union

The **union** of the two events A and B is the event containing all the elements that belong to A or B or both. Denoted $A \cup B$.



1.9 Common Set Operations (Nice for proofs)

$A \cap \emptyset = \emptyset$	$S' = \emptyset$
$A \cup \emptyset = A$	$\emptyset' = S$
$A \cap A' = \emptyset$	$(A')' = A$
$A \cup A' = S$	$(A \cap B)' = A' \cup B'$
	$(A \cup B)' = A' \cap B'$
$A \cap S = A$	$A \cap B \cap S = A \cap B$

1.10 Probability of an Event:

The measure of the likelihood that the event will occur, calculated as the ratio of the number of favorable outcomes to the total number of possible outcomes.

L2: Counting

2.0 Fundamental Principle of Counting:

If one event can occur in m ways and a second can occur independently in n ways, the number of ways the events (ie. sequences) can occur in a sequence is $m \times n$.

2.1 Permutations

The number of ways to arrange r distinct objects out of n is given by

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

- n (total number of elements in the set).
- P (the permutation operation).
- r (the number of elements taken from the set).
- **Key:** Order matters
- **Note:** If $r = n$, then ${}_nP_n = n!$ since $0! = 1$.

2.2 Permutations with identical items

If there are m kinds of items and $n_k, k = 1,..., m$ of each kind, then total number of permutations is:

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

where $\sum_{k=1}^m n_k = n$

2.3 Partitions:

The number of ways of partitioning a set of n objects into m partitions of size $n_1,..., n_m$, with

$\sum_{k=1}^m n_k = n$, is

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

- **Note:** Same formula as permutations with identical items.
- **Note:** n and $n_1,..., n_m$ relate, where one refers to the size of each object “pile” and the other refers to the total objects.
- **Note:** We are forcing the splitting of groups here but in permutations with identical items, the split is also present naturally.

2.4 Combinations

The number of ways to choose r objects from n without regard to order is

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

- C is the combination operation.
- **Key:** Order doesn’t matter.

2.5 Distinguishable vs. Indistinguishable, Order vs. Unordered, Labeled vs. Unlabeled

Distinguishable vs. Indistinguishable
Distinguishable means that the same objects can be differentiated. <ul style="list-style-type: none">• Eg. Balls of the same color can be told apart.

Indistinguishable means that the same objects cannot be differentiated.

- Eg. Balls of the same color cannot be told apart, so they have to be in groups.

Order vs. Unordered

Order means that switching the placement of two or more objects does matter.

- Eg. BBBG is not BGGB.

Unordered means that switching the placement of two or more objects doesn't matter.

- Eg. BBBG=BBGB=BGGB=GBBB.

Labeled vs. Unlabeled

Labeled means that groups can be named and so order of the groups matter.

- Eg. Group 1 and Group 2 are not the same as Group 2 and Group 1.

Unlabelled means that groups cannot be named and so order of the groups don't matter.

- Eg. Group 1 and Group 2 are the same as Group 2 and Group 1.

(2) Definitions of Probability

L3: Probability

3.0 Probability of an Event

The **probability of an event A** is the sum of the weights of all sample points in A. Therefore,

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

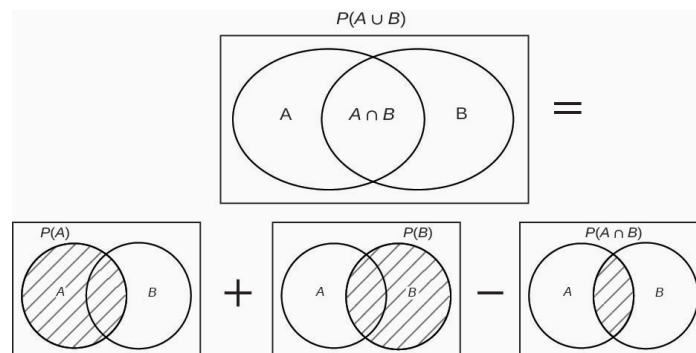
Furthermore, if A_1, A_2, A_3, \dots is a sequence of **mutually exclusive events**, then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

3.1 Additive Rule

For events $A, B \subseteq S$

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



3.2 Additive Rule for More Than Two

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

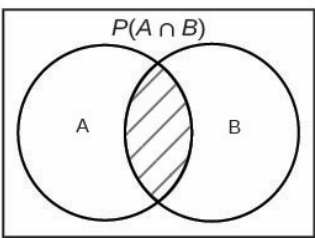
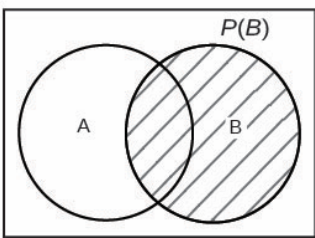
- **Pattern:** Sum their individual probabilities, subtract their pair intersection probabilities, and add intersection between everything.

3.3 Definition of Conditional Probability

Suppose $A, B \in S$. $P(B|A)$ is the probability that B occurs given that A occurs. The conditional probability of B, given A is

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

where $P(A) > 0$.

$$P(A|B) = \frac{\text{Diagram 1}}{\text{Diagram 2}}$$



- **Note:** Since B happened, therefore, it's our new "sample space". Find A is in this new sample space. Then their space is the conditional probability.
- **Note:** $P(B'|A) = 1 - P(B|A)$ | $P(B'|A') = 1 - P(B|A')$.

3.4 Product Rule

If in an experiment the events A and B can both occur, then

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

where $P(A) > 0$.

- **Note:** This comes from conditional probability since $A \cap B \Leftrightarrow B \cap A$.

3.5 Independence**

Two events A and B are **independent** if and only if

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

Assuming the existence of the conditional probabilities. Otherwise, A and B are **dependent**.

- **Key:** Independent is **NOT** mutually exclusive.
- **Key Implication:** For independent events, we can multiply their probabilities to get their intersection (useful for **calculating favorable outcomes**).
 - $P(A \cap B \cap C \cap D) = P(A)P(B)P(C)P(D)$ as long as they are **INDEPENDENT**.
 - **Note:** Can also be applied to a sequence of dependent events as long as you adjust for changing conditions.

(3) Bayes' Rule

3.6 Bayes' Rule

For events with $P(A) > 0$ and $P(B) > 0$:

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

by conditional probability.

Since $P(B \cap A) = P(A \cap B)$, therefore, $P(B|A)P(A) = P(A|B)P(B)$, so:

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

3.7 Mutually Exclusive vs. Independent

- **Independent events** are applicable in scenarios where multiple events are occurring in sequence, and the occurrence of one has no effect on the probability of the others. The probability of independent events occurring together is found by multiplying their probabilities.
- **Mutually exclusive events** are relevant in scenarios where you are considering one event or another happening, but not both. The probability of either mutually exclusive event occurring is the sum of their probabilities
- **Mutually exclusive events** cannot happen at the same time. Their occurrence is exclusive of each other.
- **Independent events** do not influence each other. The occurrence of one does not change the probability of occurrence of the other.

L4: Bayes' Rule

4.0 Partitions

B_1, \dots, B_k is a **partition** if $B_i \cap B_j = \emptyset$ and $B_1 \cup \dots \cup B_k = S$.

4.1 Theorem of Total Probability

Suppose A is an event and B_1, \dots, B_k is a partition. Then

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

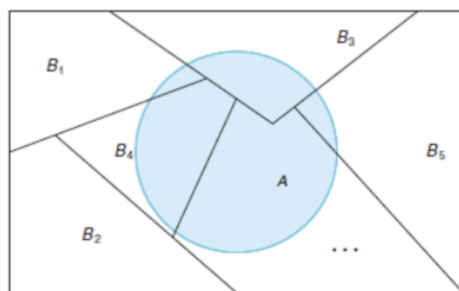


Figure 2.14: Partitioning the sample space S .

4.2 Bayes' Rule with Total Probability

Suppose C_1, \dots, C_k is a partition. Then

$$P(B|A) = \frac{P(B)P(A|B)}{\sum_{i=1}^k P(C_i)P(A|C_i)}$$

- **Note:** This is replacing $P(A)$ with $\sum_{i=1}^k P(C_i)P(A|C_i)$.

Often B is an element of $\{C_1, \dots, C_k\}$, say $B = C_n$. Then

$$P(C_n|A) = \frac{P(C_n)P(A|C_n)}{\sum_{i=1}^k P(C_i)P(A|C_i)}$$

- **Note:** Bayes' rule useful for incorporating limited information.

4.3 Process for Bayes' Rule, Conditional Probability Questions

- Write down all the probabilities.
- Try solving the problem directly using definitions.
 - If given $P(A|B)$ and want $P(B|A)$, then **automatically** use Bayes' Rule.

Random variables, distributions, & expectation (3.1-4, 4.1-3)

(4) RVs & Distributions

L5: Random Variables

5.0 Random Variables (RVs)

A **random variable** (RV) is a function that associates a real number with each element of the sample space. Denote RVs with capital letters (eg. X or Y)

- **Types of Random Variables:**
 - **Discrete RV:** X takes on a finite or countable number of values (eg. $\{1, 2, 3\}$ or \mathbb{Z}).
 - **Continuous RV:** X takes on values in an interval of \mathbb{R} .
- **Notation:** Individual values X can take on with small letters, $X = x$.

5.1 Discrete - Probability Distributions

The probability that a discrete RV takes on each value.

5.2 Discrete - Probability Mass Function (PMF)

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

- $f(x) \geq 0$ for each outcome $X = x$
- $\sum_x f(x) = 1$ (ie. total probability sums to 1)
- $f(x) = P(X = x)$ (ie. probability of each outcome)
- **Drawings:** The values of the PMF will be 0 for the values that aren't defined by definition.

5.3 Discrete - Cumulative Distribution Function (CDF)

Let RV X have a PMF $f(x)$. The CDF of X is

$$P(X \leq x) = F(x) = \sum_{t \leq x} f(t)$$

for $x \in \mathbb{R}$.

- **Note:** $F(x) = P(X \leq x)$, represents the probability that X takes on a value less than or equal to x .
- **Drawings:** For the values that don't contribute to the CDF, it stays at the previous respective value.
- **Note:**
 - $P(a \leq X \leq b) = F(b) - F(a - 1) \mid P(x < a) = F(a - 1)$
 - $P(x > a) = 1 - F(X \leq a) \mid P(x \geq a) = 1 - F(x \leq a - 1)$
- **Key:** CDF only defined for \leq .

5.4 Understanding Continuous Random Variables

1. Probability of an Exact Value

In a reasonable distribution, $P(X = 5) = 0$ because the probability of the variable taking an exact value is infinitesimally small in a **continuous RV**.

2. Probability of a Range (this is what we care about)

The probability of X falling within a range and is a nonzero value for continuous distributions.

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

5.5 Continuous - Probability Density Function (PDF)

Let $f(x)$ be a PDF of the continuous RV X . Then:

- $f(x) \geq 0$ for each outcome $X = x$
- $\int_{-\infty}^{\infty} f(x) dx = 1$ (ie. normalization - the integral of $f(x)$ over the entire real line is 1).
- $\int_a^b f(x) dx = P(a < X < b)$ (ie. the probability that X falls within a certain range $[a, b]$ is given by the integral of $f(x)$ over that range).
- **Note:** The density doesn't give anything meaningful to the probability, but the integral over an interval will give the probability.
- **Note:** Only the **first two bullet points** are conditions for a PDF.
- **Key:** Normalization is key to find unknown parameters.

5.6 Continuous - Cumulative Distribution Function (CDF)

Let X be a continuous RV with PDF $f(x)$. The CDF of X is

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

for $x \in \mathbb{R}$.

- **Key Properties of CDF:**

- $F(x) = P(X \leq x)$ (ie. probability that X takes on a value less than or equal to x).
- $F(\infty) = P(X \leq \infty) = \int_{-\infty}^{\infty} f(t) dt = 1$ (ie. normalization property)
- $P(a < X \leq b) = F(b) - F(a)$ (ie. probability that X falls within the interval $(a, b]$).

- **Geometrically:** Represents the area under the curve of $f(t)$ from $-\infty$ to x .

5.7 Differences between PDFs and PMFs

Continuous PDFs	Discrete PMFs
<ul style="list-style-type: none"> • PDF represented as $f(x)$ • Acts as a 'density' over the sample space S.... $\rho V = m$. • $P(X = x) = 0$ unless there is a Dirac function in the PDF. • Probability over an interval is non-zero that is a defined sample space and $P(a \leq X \leq b)$ is the integral of $f(x)$ from a to b. 	<ul style="list-style-type: none"> • <i>PMF represented as $f(x)$</i> • Each discrete point in the sample space S has a probability 'mass' • $P(X = x) = f(x)$ (ie. direct association between the PMF and probabilities).

5.8 Process to find CDF to PDF and PDF to CDF

1. CDF \rightarrow PDF

$$f(x) = \frac{dF(x)}{dx}$$

2. PDF \rightarrow CDF

- Integrate the first term of the piecewise $f(x)$ from $-\infty$ to x .
- Add the 1st term to 2nd term with defined bounds for 1st term and integrate 2nd term from its lower bound to x .

c. Repeat steps 1-2 until you have gone through all the terms using the same process for 3rd,..., n terms.

d. The final answer should be 1.

- **Aside:** If we are trying to introduce a variable $X = 2Y$, and find the PDF $f(y)$ given $f(x)$. You must first find the CDF as the PDF is not 1-1, and then plug in $F\left(\frac{y}{2}\right)$ to find the CDF in terms of y , and then differentiate again to get the PDF.

L6: Joint Distributions

6.0 Discrete - Joint PMF

$f(x, y)$ is a joint PMF of the discrete RVs X and Y if

- $f(x, y) \geq 0 \forall (x, y) \in S$
- $\sum_x \sum_y f(x, y) = 1$ (ie. total probability)
- $P(X = x, Y = y) = f(x, y)$

For a subset $A \subset S$. $P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y)$

6.1 Continuous - Joint PDF

$f(x, y)$ is a joint PDF of the continuous RVs X and Y if

- $f(x, y) \geq 0 \forall (x, y) \in S$ (ie. sample space)
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ (ie. the integral of $f(x, y)$ over the entire plane is 1).
- For a subset $A \subseteq S$, $P((X, Y) \in A) = \int_{(x, y) \in A} f(x, y) dx dy$
- **Note:** $P(X < Y) = \int_{-\infty}^{\infty} \int_{-\infty}^y f(x, y) dx dy$

6.2 Marginal Distribution

Given the **joint distribution** of X and Y , $f(x, y)$, find the **marginal distributions**

1. Discrete Variables:

$$g(x) = \sum_y f(x, y), \quad h(y) = \sum_x f(x, y)$$

2. Continuous Variables:

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \mid h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- **Note:** The idea is to take a 'weighted average' of $f(x, y)$ over all possibilities of the other variable, that's why the **integral/sum** is w.r.t the other variable.
- **Note:** When we integrate w.r.t to the other variable, do it as normal, but the bounds of the resultant marginal distribution should be in terms of constants.
 - **Eg.** $0 \leq x \leq y \leq 1$ and want $g(x)$, then $\int_x^1 f(x, y) dy$ and $g(x)$ is from $0 \leq x \leq 1$.
 - **Intuition:** When we integrate w.r.t y , we integrate over the full bound. Since y could be 0, therefore, we make sure to integrate the full bound.
- **Note:** Draw out the region on tests.

6.3 Conditional Distributions (Discrete or Continuous)

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

where $h(y)$ is the marginal distribution of y .

1. Discrete Calculation:

$$P(a \leq X \leq b \mid Y = y) = \sum_{a \leq x \leq b} f(x|y)$$

2. Continuous Calculation:

$$P(a \leq X \leq b \mid Y = y) = \int_a^b f(x|y) dx$$

- **Note:** The definition of $f(x|y)$ is given above, where we replace y with value.

6.4 Independence of Random Variables

X and Y are RVs with joint distribution $f(x, y)$ and marginal distributions $g(x)$ and $h(y)$ are said to be independent if

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

- **Implication: Joint probability distribution** can be the **product** of the individual **marginal distributions**.
- **Note:** Independence means that the **occurrence of one event** (described by one random variable) **does not affect** the probability distribution of the **other**.
- **Note:** Even if it's obvious to separate out $f(x, y)$, you must do the integral way as that's **by definition of marginal distributions**.

(5) Expectation, Variance, & Covariance

L7: Expectation

7.0 Expectation

Let X be an RV with distribution $f(x)$. The expected value or expectation of X is defined as

1. Discrete Case

$$E[X] = \sum_x xf(x)$$

where the sum is taken over all possible values of X .

- **Note:** Think about this intuitively by having some numbers, and multiplying with its probability.

2. Continuous Case

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

where the integral is taken over the entire range of X .

Key Points

- Expectation is a measure of the central tendency of a probability distribution.
- It represents the average or mean value that the RV X is expected to take on.

7.1 Expectation of $g(X)$

Let X be an RV with distribution $f(x)$, and let $g(X)$ be a function of X . The expectation of $g(X)$ is

1. Discrete Case

$$E[g(X)] = \sum_x g(x)f(x)$$

where the sum is over all possible values of X .

2. Continuous Case

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

where the integral is taken over the entire range of X .

Key Points:

- Extends the concept of expectation to any function of the random variable X .
- A general method for calculating expected values of transformations or combinations of RVs.

7.2 Expectation of $g(X,Y)$

Let X and Y be RVs with joint distribution $f(x, y)$, and let $g(X, Y)$ function be a function of X and Y . The expectation of $g(X, Y)$ is

1. Discrete Case

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

where the sum is over all possible values of X and Y .

2. Continuous Case

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

where the integrals are over the entire ranges of X and Y .

Key Points

- A method for calculating expected values of transformations or combinations of two RVs.

7.3 Definition of Variance

Let X be an RV with distribution $f(x)$ and mean $\mu = E[X]$. The variance of X is

1. Discrete Case

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$$

where the sum is over all possible values of X .

2. Continuous Case

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

where the integral is taken over the entire range of X .

Key Points:

- Variance is a measure of variability or spread of an RV.

- The standard deviation, $\sigma = \sqrt{\sigma^2}$

7.4 Useful Formula For Discrete and Continuous RVs:

$$\sigma^2 = E[X^2] - \mu^2$$

Key Points

- This formula simplifies the calculation of variance.
- The same formula applies in the discrete case as well.

7.5 Covariance

Let X and Y be RVs with joint distribution $f(x, y)$ and means μ_x and μ_y . The covariance of X and Y is

$$\sigma_{XY} = \text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

1. Discrete Case

$$\sigma_{XY} = \sum_x \sum_y (x - \mu_x)(y - \mu_y) f(x, y)$$

2. Continuous Case

$$\sigma_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy$$

Interpretation:

- If **positive values** of X and Y **tend to occur together**, then $\sigma_{XY} > 0$.
- If **positive values of X** tend to **occur with negative values of Y** (or vice versa), $\sigma_{XY} < 0$.

7.6 Useful Formula for Covariance

The covariance between RVs X and Y is

$$\sigma_{XY} = E[XY] - \mu_x \mu_y$$

Key Points

- This formula simplifies the calculation of covariance.
- **In words:** Covariance is the difference between the expected value of the product XY and the product of their means $\mu_x \mu_y$.
- Generalization of the variance formula.

7.7 Correlation Coefficient

Let X and Y be RVs with covariance σ_{XY} and standard deviations σ_X and σ_Y . The correlation coefficient of X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

- **Note:** Always calculate the numerator first.

Interpretation of Correlation Coefficient:

- The correlation coefficient is like covariance, but **normalized**.
- $-1 \leq \rho_{XY} \leq 1$
- When $\rho = 0$, X and Y are said to be '**uncorrelated**'.

Key Points

- A correlation coefficient close to **1 or -1** indicates a **strong linear relationship**.
- A correlation coefficient close to **0** suggests a **weak or no linear relationship**.

L8: Linear Combinations of Random Variables

8.0 Linearity

The function $p(x)$ is linear if

1. **Homogeneity:** $p(ax) = ap(x)$ for constant a .
2. **Superposition:** $p(x + y) = p(x) + p(y)$
3. **Combining homogeneity and superposition:** $p(ax + y) = ap(x) + p(y)$

8.1 Linearity of Expectation for Linear Combination of RVs

If X and Y are RVs, then for constants a and b , the expectation operator $E[\cdot]$ is linear:

$$E[aX + bY] = aE[X] + bE[Y]$$

- **Note:** Holds for both continuous and discrete cases.
- **Note:** LS would use the joint distribution $f(x, y)$, while RS would use the marginals $g(x)$ and $h(y)$.
- **Useful Implications:**

- $E[aX + b] = aE[X] + b$ (since $E[b] = b \int_{-\infty}^{\infty} f(x)dx = b$ by normalization)
- $E[g(X, Y) + h(X, Y)] = E[g(X, Y)] + E[h(X, Y)]$
- $E[aX + Y] = aE[X] + E[Y]$

8.2 Variance and Independence of Random Variables

For independent RVs X and Y, then $f(x, y) = g(x)h(y)$. Observe

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

- **Note:** $f(x, y)$ is the joint distribution, while $g(x)$ and $h(y)$ are the marginal distributions.
- **Key:** Independence implies uncorrelated ($\sigma_{XY} = E[XY] - E[X]E[Y] = 0$), but uncorrelated **does not imply independence** (ie. independence stronger).

Key Points

- Uncorrelated random variables can still have a **dependency structure** not captured by linear correlation.

8.3 Useful Formula

Consider RVs X and Y. The variance of $aX + bY + c$ is

$$\sigma_{aX+bY+c}^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab\sigma_{XY}^2$$

Key Points

- c drops out because it does not contribute to the variance.
- If X and Y are independent, σ_{XY} is zero: $\sigma_{aX+bY+c}^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2$
- This formula is crucial when dealing with linear combinations of RVs.

8.4 More Useful Formulas

$$\begin{aligned}\sigma_{aX+bY}^2 &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab\sigma_{XY}^2 \\ \sigma_{aX-bY}^2 &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 - 2ab\sigma_{XY}^2 \\ \sigma_{aX+bY-cZ}^2 &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 + c^2 \sigma_Z^2 + 2ab\sigma_{XY}^2 - 2ac\sigma_{XZ}^2 - 2bc\sigma_{YZ}^2\end{aligned}$$

8.5 Integration by Parts on Crack

1. Given $f(x)$ and $g(x)$, take the derivative of $f(x)$ until $f^n(x) = 0$.
 2. Multiply on the diagonals starting with a positive, then a negative (alternate).
 - a. Eg. $f(x)g'(x) - f'(x)g''(x)$ if $f''(x) = 0$.
- **Note:** If $f(x)$ is a sinusoidal function, then only do it twice only.

$f(x)$	$g(x)$
$f'(x)$	$(+1)g'(x)$
$f''(x)$	$(-1)g''(x)$
\dots	\dots
$f^n(x) = 0$	$g^n(x)$

Common distributions (5.1-5, 6.1-7, 6.10)

(6) Common Discrete Distributions

L9: Common Discrete Distributions

9.0 Parameters vs. Statistics

Parameters	Statistics
<p>Parameters are numerical characteristics of the population.</p> <ul style="list-style-type: none"> They are fixed values, although often unknown in practice. Eg. population mean (μ), population variance (σ^2), and population proportion (p). 	<p>Statistics are numerical characteristics of a sample.</p> <ul style="list-style-type: none"> They are estimates of the population parameters and can vary from sample to sample. Eg. sample mean (\bar{x}), sample variance (s^2), and sample proportion (\hat{p})
Key Differences	
<ul style="list-style-type: none"> Parameters describe the entire population and are usually fixed numbers. 	<ul style="list-style-type: none"> Statistics describe a sample of the population and are used to estimate parameters.

9.1 Discrete Probability Distributions

What is a Discrete Probability Distribution?
<p>Describes the probability of occurrence of each value of a discrete random variable.</p> <ul style="list-style-type: none"> A discrete random variable is one that has countable values, such as 0,1,2,...
Characteristics
<ul style="list-style-type: none"> Each probability is between 0 and 1, inclusive. The sum of all probabilities equals 1.
Mathematical Representation

- Probability Mass Function (PMF), $f(x)$, where $f(x) = P(X = x)$
- Cumulative Distribution Function (CDF), $F(x)$, where $F(x) = P(X \leq x)$.

9.2 Binomial Distribution

Overview:
Describes the number of successes in a fixed number of independent Bernoulli trials.
Bernoulli RV
<ul style="list-style-type: none"> • Consists of only two outcomes, like the coin flip (ie. success/failure) • $S = \{0, 1\}$ • Probability of success: $P(X = 1) = f(1) = p$
Binomial Process
<ul style="list-style-type: none"> • A sequence of n repeated, independent trials with an identical probability distribution (idd) p of success. <ul style="list-style-type: none"> ◦ iid \Rightarrow identical, independent, and distributed (ie. same type of trial, not using different coins) • RV X: # of 1's that occur (ie. number of successes).
Probability Mass Function:
<p>Binomial distribution gives the probability of exactly x successes:</p> $P(X = x) = f(x) = b(x; n, p)$ $b(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}.$ <ul style="list-style-type: none"> • x (# of successes) • n (# of trials) • p (probability)
Conditions:

1. Bernoulli RV (ie. 2 outcomes).
2. Trials are a fixed number that are independent, identical, and distributed (iid)
3. Probability is constant for each trial.

Statistics

1. Mean with RV X

$$E[X] = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x}$$

Think of process as sum of n Bernoulli trials,

$$X = Y_1 + \dots + Y_n$$

By linearity of expectation,

$$\mu = E[X] = np$$

2. Variance

Since each trial is independent, the covariance $\sigma_{Y_j Y_k} = 0$ for $k \neq j$. So

$$\sigma_X^2 = np(1-p)$$

- **Note:** The variance σ_X^2 for a binomial random variable X is found by considering $X = Y_1 + \dots + Y_n$.

Derivation of Binomial Distribution

1. Probability of x 1's (ie. successes) and $n - x$ 0's (ie. failures) in **some particular order**:

$$p^x (1-p)^{n-x}.$$

- **Note:** Multiply probability to get the outcomes of probability of successes and failures as they are independent.
- **Note:** p^x is probability of successes | $(1-p)^{n-x}$ is probability of failures.

2. **Number of ways** to have x 1's (ie. successes) in n trials:

$$\binom{n}{x}$$

- **Note:** Count combinations to get all the different orders of successes and failures
 - **Eg.** 3 coin tosses with 1 as heads and 0 as tails, and trying to get 2 heads is 110, 101, and 011.
3. Binomial probability mass function (combining steps 1-2)

9.3 Multinomial Distribution (Generalization of Binomial)

Overview

Extends the binomial distribution to cases where **each trial can have m outcomes** instead of 2.

- Outcomes E_1, \dots, E_m with probabilities of $P(E_i) = p_i$ and $\sum_{i=1}^m p_i = 1$.
- The probability E_1 occurring x_1 times, E_2 occurring x_2 times, ... where $x_1 + \dots + x_m = n$.

Probability Mass Function:

Probability of a specific arrangement where E_i happens x_i times ($i = 1, \dots, m$):

$$f(x_1, \dots, x_m; p_1, \dots, p_m, n) = \binom{n}{x_1, x_2, \dots, x_m} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}$$

- n (# of trials)
- x_i (groups partitioned in n with $\sum_i x_i = n$)
- p_i (probability of each group with $\sum_{i=1}^m p_i = 1$)
- **Note:** The multiplying the different probabilities with the partitions is to get all the different combinations of probability with a **specific arrangement**.
- **Note:** We only care about the group, not how the group is ordered inside of it.
 - As a result, we remove all those combinations with different ordering inside by dividing by their factorial (ie. partitions).

Conditions
<ol style="list-style-type: none"> 1. Fixed number of trials. 2. Multiple outcomes (ie. more than 2). 3. Constant probabilities.
Derivation of Multinomial Distribution
<ol style="list-style-type: none"> 1. Chance of a particular instance of E_i happening x_i times $p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}$ <ul style="list-style-type: none"> • $i = 1, \dots, m$ 2. Combination formulas: Number of ways to partition n trials into m groups of sizes x_1, x_2, \dots, x_m: $\binom{n}{x_1, x_2, \dots, x_m} = \frac{n!}{x_1! x_2! \dots x_m!}$ <ul style="list-style-type: none"> • Note: Binomial is special case of partitions of size $x_1 = x$ and $x_2 = n - x$ since there are two groups: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ 3. Multinomial probability mass function (combining steps 1-2)

L10: More Discrete Distributions

10.0 Differences with Distributions:

1. **Binomial:** Flipping a coin n times, where each outcome (e.g., heads) remains possible in subsequent trials.
2. **Multinomial:** Drawing a card, noting its value, and replacing it before the next draw.
3. **Hypergeometric:** Used when items are not replaced, as in drawing cards without replacement.

10.1 Hypergeometric Distribution

Overview
<p>Models the number of successes x in a random sample of size n without replacement from a finite population of size N containing exactly K successes.</p> <ul style="list-style-type: none">• N total objects• Sample n times.• K of the N are successes.• Chance of x successes and $n - x$ failures.
Probability Distribution Function
<p>The probability of observing x successes out of K successes in n draws from a finite population of size N without replacement:</p> $h(x; N, n, K) = \frac{\binom{K}{x} \binom{N - K}{n - x}}{\binom{N}{n}}$ <p>with $\max\{0, n - (N - K)\} \leq x \leq \min\{n, K\}$</p> <ul style="list-style-type: none">• x (# of successes in the sample)• K (# of successes in the population)• N (size of population)• n (size of sample)
Conditions
<ol style="list-style-type: none">1. Fixed population size and sample size.2. Two types of outcomes in population (ie. successes/failures)3. Sampling is done without replacement.4. Independence not required.
Range of Random Variable X :

- Determined by the binomial coefficients in the function, where x and $n - x$ can be no more than K and $N - K$, respectively.
- X typically ranges from 0 to n when K and $N - K$ are larger than the sample size n .

Statistics

1. Mean

$$\mu = \frac{nK}{n}$$

2. Variance

$$\sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N} \left(1 - \frac{K}{N}\right)$$

10.2 Binomial vs. Negative Binomial

- Binomial:** Gives the probability of the number of k successes in n trials.
 - Key:** Focuses on the number of successes in a fixed the number of trials
 - Eg.** Chance of k heads out of n coin flips:
- Negative binomial:** Gives the probability that the k th success occurs on the n th trial.
 - Key:** Focuses on how many trials it takes to achieve a fixed number of successes.
 - Eg.** Chance the k th head occurs on the n th coin flip

10.3 Negative Binomial Distribution

Overview

Describes the number of trials X needed to achieve k successes in repeated, independent trials, with success probability p and failure probability $q = 1 - p$.

- Repeated trials with probability p of success and $1 - p$ of failure.
- RV X is the trial on which the k th success occurs.

Probability Distribution Function

The probability that the k th success occurs on the x th trial is given by:

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

where $x = k, k + 1, k + 2, \dots$

- k (# of successes)
- x (# of trials)
- p (probability)

- **Note:** The coefficient $\binom{x-1}{k-1}$ represents the number of ways to arrange $k - 1$ successes in the first $x - 1$ trials.

Conditions

1. **Bernoulli Trials:** A process consisting of repeated, independent trials with two possible outcomes (success or failure).
2. Constant probabilities.
3. Fixed number of successes.

Statistics

1. **Mean:** The expected number of trials required to achieve k successes

$$\mu = \frac{k}{p}$$

- **Note:** Provides a measure of how many trials on average one can expect to perform before achieving the desired number of successes.

2. Variance

$$\sigma^2 = \frac{k(1-p)}{p^2}$$

- **Note:** Tells us the dispersion or variability of the trials around the expected number μ .

Interpretation:

- If the probability of success p is low, the mean and variance will be higher, indicating a longer and more uncertain process to achieve k successes.

10.4 Geometric Distribution

Overview

A special case of the Negative Binomial distribution where the **number of successes k is one**.

- **Note:** Models the probability of observing the first success on the xth trial in a sequence of independent Bernoulli trials.

- **Note:** Success on a single trial occurs with probability p , and failure with probability $q = 1 - p$.
- Repeated trials with probability of success p .
- RV X is the trial on which the first success occurs.

Probability Distribution Function

$$g(x; p) = p(1 - p)^{x-1}$$

for $x = 1, 2, 3, \dots$

- x (# of trials)
- p (probability)

Conditions

1. **Bernoulli Trials:** A series of independent trials with two outcomes: success or failure.
2. Constant probabilities.
3. Fixed number of successes as 1.

Statistics

1. **Mean:** Representing the expected number of trials until the first success

$$\mu = \frac{1}{p}$$

2. **Variance:** Measures the dispersion of the number of trials

$$\sigma^2 = \frac{1-p}{p^2}$$

Significance:

- These statistics help understand the average wait for an event and the variability around this wait.
- In practice, a lower p (probability of success) increases both the mean and variance, indicating a longer and more variable wait for a success.

L11: Poisson Distribution and the Poisson Process

11.0 Poisson Distribution

Overview:
The Poisson distribution is used to model the number of times an event occurs in a defined interval of time or space.
Poisson Process
A random process that models the occurrence of events that happen independently over a continuous interval. <ul style="list-style-type: none">• Note: Often used to describe the random nature of events scattered in time or space.
Key Properties:
<ul style="list-style-type: none">• Independence: The number of events occurring in any interval is independent of the number occurring in any other non-overlapping interval. (ie. no memory)• Proportionality: The probability of a single event occurring within a very short interval is proportional to the length of that interval (eg. duration or length).• Rare Events: The number of events in non-overlapping intervals follows a Poisson distribution
Probability Mass Function
<p>The probability of observing x events in a fixed interval t is:</p> $p(x; \lambda) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$ <p>for $x = 0, 1, 2, \dots$</p> <ul style="list-style-type: none">• $\lambda \left[\frac{\text{average \# of occurrences}}{\text{unit time}} \right]$ (<i>rate at which events occur</i>)• λt (<i>represents the average number of occurrences per interval</i>)• x (# of events)
Conditions
<ol style="list-style-type: none">1. Discrete Events: Countable, distinct occurrences (e.g., emails, calls).2. Constant Rate: The average number of events per interval is constant.

3. **Independent Events:** Events occur independently of each other.
4. **Rare Events:** Suitable for rare events in large populations or areas.
5. **Defined Interval:** Events are counted within a specific time, area, or volume.
6. **Single Events:** Events occur singly, not simultaneously.
7. **Random Occurrence:** Events happen randomly.

Statistics

1. Mean

$$\mu = \lambda t$$

2. Variance

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

Poisson Distribution as Limit of Binomial Distribution

$$\lim_{n \rightarrow \infty, p \rightarrow 0} b(x; n, p) = p(x; \lambda t)$$

- **Key:** As $n \rightarrow \infty$, $p \rightarrow 0$, $\lambda t = np$.
- **Note:** Useful for a large number of trials or events that are distributed continuously over space or time.
- **Simplifies:** Simplifies calculations, especially in cases involving a very large number of trials.
- **Note:** Aligns with the principles of the Poisson process.

11.1 Chebyshev's Theorem (Discrete or Continuous RVs)

The probability that X is within k standard deviations of the mean is at least

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

holds for all $k > 0$.

- **What does it do?** Provides a bound on the probability that the value of a RV X deviates from its mean μ by more than k standard deviations σ .
- **Condition:** It is applicable to any probability distribution, regardless of its shape, provided the mean and variance are known.

(7) Common Continuous Distributions

L12: Uniform and Normal Distributions

Reminder:

For continuous distributions:

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$$

12.0 Uniform Distribution

Probability Density Function	
<p>The uniform distribution is a constant PDF over an interval [A,B] with continuous uniform RV:</p> $f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{otherwise} \end{cases}$ <ul style="list-style-type: none"> Key: Every outcome in the interval [A,B] is equally likely. 	
Statistics	
1. Mean (i.e. midpoint of the interval)	$\mu = \frac{A+B}{2}$
2. Variance	$\sigma^2 = \frac{(B-A)^2}{12}$

12.1 Normal (Gaussian) Distribution

Probability Density Function	
<p>A normal RV X with mean μ and variance σ^2 has PDF of</p> $n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \text{ for } -\infty < x < \infty$	
Statistics	
1. Mean	$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \mu$

2. Variance:

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

- **Key:** For the mean and variance, we are just proving why we can use these in the definition of the normal PDF by showing by definition that LS=RS.

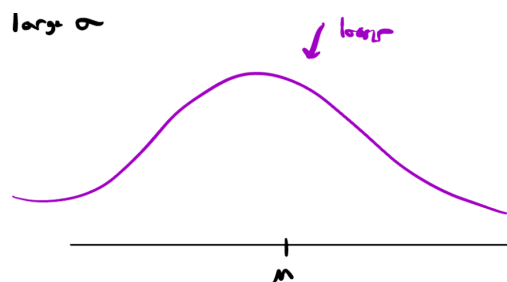
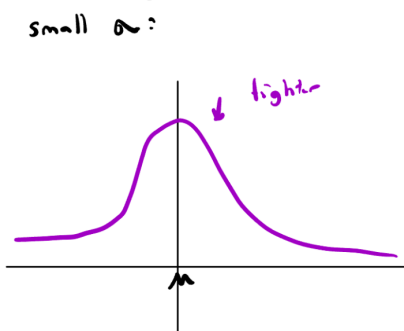
Intuition

Characterized by 2 parameters: mean (μ) and variance (σ^2):

- **Mean (μ):** Determines the center of the distribution. It's the value around which the data is symmetrical.
 - A change in μ shifts the distribution left/right on the horizontal axis.
- **Standard Deviation (σ):** Defines the spread of the distribution.
 - A small σ means data is tightly clustered around the mean.
 - A large σ indicates the data is spread out over a wider range of values.

Visual Representation:

Visual Representation:



12.2 Standard Normal Distribution (ie. Normal Distribution we will use)

Probability Density Function

$$n(x; \mu = 0, \sigma = 1)$$

Cumulative Distribution Function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{t^2}{2}\right\} dt$$

Probability Between a Range

For a standard normal RV X , then the probability that X falls within the interval $[A, B]$ is

$$P(A \leq X \leq B) = \Phi(B) - \Phi(A)$$

- **Note:** Use a table to compute the values of Φ .

How to use $n(x; 0, 1)$ and $\Phi(x)$ for general μ and σ ?

Given a RV X with a normal distribution $n(x; \mu, \sigma)$, then define the **standardized variable**.

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

- **Note:** This centers everything around 0 and with the same spread.
- **Result:** Z has the standard normal distribution $n(x; \mu = 0, \sigma = 1)$ (ie. we can use the table).

CDF:

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

- **Note:** This transformation allows us to use the standard normal table to find probabilities for any normal random variable (**ie. same probability**).

Calculating Probabilities Using the Standard Normal Distribution

The probability that X falls between two values A and B is:

$$P(A \leq X \leq B) = P\left(\frac{A - \mu}{\sigma} \leq Z \leq \frac{B - \mu}{\sigma}\right) = \Phi\left(\frac{B - \mu}{\sigma}\right) - \Phi\left(\frac{A - \mu}{\sigma}\right)$$

Relationship Between the PDFs of X and Z

$$n(x; \mu, \sigma) = \frac{n\left(\frac{x - \mu}{\sigma}; 0, 1\right)}{\sigma}$$

How to Use The Standard Normal Table?

1. Convert the normal distribution to the standard normal distribution by using the standardized variable.
2. Match the Z values with the 1st column that gives the first decimal point and the 1st row that gives the second decimal point and add those together.
3. The value that you converge on will be the probability.
 - a. **Note:** You might also be asked to go backwards, where you have the probability, but are looking the value of k , where $P(Z < k) = P_0$

Practical Use:

- Given precomputed values of $\Phi(x)$, $x \in \mathbb{R}$.
- Given normal RV X with PDF $n(x; \mu, \sigma)$

<ul style="list-style-type: none"> • Compute statistics of X by evaluating standard normal PDF and CDF for $Z = \frac{X-\mu}{\sigma}$.
Random:
<ul style="list-style-type: none"> • $F(1) \neq F(-1)$

L13: A Few More Continuous Distributions

13.0 Normal Approximation of Binomial PMF

Overview
<ul style="list-style-type: none"> • A binomial random variable X with parameters n and p can be approximated by a normal distribution for large n. • This approximation assumes X has a mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$.
Approximation Formula
<p>The probability that X is less than or equal to x is given by</p> $P(X \leq x) \approx P\left(Z \leq \frac{x+0.5-np}{\sqrt{npq}}\right)$ <ul style="list-style-type: none"> • Z (standard normal RV) • Note: As $n \rightarrow \infty$, limiting PDF of Z is $n(x; 0, 1)$. • Note: In the textbook, they subtract, but that's fine.
Conditions
<ol style="list-style-type: none"> 1. $np \geq 5$ 2. $n(1 - p) \geq 5$

13.1 Gamma Distribution

Gamma Function
$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \alpha > 0$ <ul style="list-style-type: none"> • Properties

- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ for $\alpha = \frac{1}{2}$.
- $\Gamma(n) = (n - 1)!$ for $n \in \mathbb{N}$.
- It generalizes the factorial function to non-integer values.

Probability Density Function

RV X has gamma distribution with parameters $\alpha > 0$ and $\beta > 0$:

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

Statistics

1. Mean

$$\mu = \alpha\beta$$

2. Variance

$$\sigma^2 = \alpha\beta^2$$

13.2 Chi-Squared Distribution

Probability Density Function

A continuous RV X has chi-squared (χ^2) distribution with v degrees of freedom

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

- **Relationship with Gamma Distribution:** Special case with $\alpha = \frac{v}{2}$ and $\beta = 2$.

Statistics

For a Chi-Squared RV X with degrees of freedom v :

1. Mean

$$\mu = v$$

2. Variance

$$\sigma^2 = 2v$$

13.3 Exponential Distribution

Overview

Describes the time between events in a Poisson process, where events occur continuously and independently at a constant average rate.

- **Note:** Exponential distribution is used to model the time until an event occurs.

Probability Density Function

RV X has exponential distribution with a scale parameter $\beta > 0$:

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

- β is the inverse of the rate parameter (i.e. scale parameter) ($\beta = \frac{1}{\lambda}$).
- **Relationship with Gamma Distribution:** Special case with $\alpha = 1$

Statistics:

For an exponential RV X with parameter β

1. Mean

$$\mu = \beta$$

2. Variance .

$$\sigma^2 = \beta^2$$

Memoryless Nature of Exponential Distribution

The memoryless property signifies that the distribution of the remaining lifetime is independent of any elapsed duration.

Given RV X for time until an event with PDF $f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$ for $x \geq 0$.

$$P(X \geq s + t \mid X \geq s) = P(X \geq t)$$

Functions of random variables (7.1-3)

(8) Functions of RVs

L14: Functions of Random Variables

14.0 Transformations of Discrete RVs from X to Y for a PMF

Setup
<ul style="list-style-type: none"> A discrete random variable X with PMF $f(x)$. Define $Y = u(X)$, where u is a bijective (one-to-one and onto) function. <ul style="list-style-type: none"> Note: $X = u^{-1}(y)$ For each value of X, there is a unique corresponding value of Y.
PMF of Y:
<p>The PMF of Y, $g(y)$, is</p> $g(y) = f(u^{-1}(y))$ <ul style="list-style-type: none"> Condition: Holds when X has PMF $f(x)$, $Y = u(x)$, and u is invertible.
Process
<ol style="list-style-type: none"> Find $u(x)$: The relationship of Y expressed in terms of X (ie. $u = \text{fcn of } x$) Find $u^{-1}(y)$: The relationship of X expressed in terms of Y (ie. $u^{-1} = \text{fcn of } y$) Plug $u^{-1}(y)$ into corresponding PMF for f to find f in terms of y <ol style="list-style-type: none"> Key: The bounds need to change as well and y can only take on discrete values.

14.1 Transformations of Continuous RVs from X to Y for a PDF

Setup
<ul style="list-style-type: none"> Let X be a continuous random variable with PDF $f(x)$. Define a new random variable $Y = u(X)$, where u is a bijective function.

- **Note:** $X = u^{-1}(y)$

PDF of Y:

The CDF of Y, $G(Y)$, is

$$G(y) = P(Y \leq y) = \int_{-\infty}^{x=u^{-1}(y)} f(t)dt$$

The PDF of Y, $g(y)$, is

$$g(y) = \frac{d}{dy} G(y) = f(u^{-1}(y)) \cdot \left| \frac{du^{-1}(y)}{dy} \right|$$

- **Note:** The absolute value is used because $u(X)$ can be an increasing or decreasing function (but we don't care about that).

Process

1. Find $u(x)$:

- a. **Key:** Be careful to choose the bounds of x to fulfill the bijective requirement.

The relationship of Y expressed in terms of X (ie. $u = \text{fcn of } x$)

2. Find $u^{-1}(y)$: The relationship of X expressed in terms of Y (ie. $u^{-1} = \text{fcn of } y$)

3. Find the Jacobian.

- a. **Note:** The old variable is the numerator, while the denominator will be the new variable.

4. Plug $u^{-1}(y)$ into corresponding PMF for f to find f in terms of y

- a. **Key:** The bounds need to change as well and y can only take on discrete values.

(9) Moments

14.2 Moment and Moment-Generating Functions

Moments
<p>The rth moment about the origin of the RV X is</p> $\mu'_r = E[X^r]$ $= \begin{cases} \sum x^r f(x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x^r f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$ <ul style="list-style-type: none"> • Observations <ul style="list-style-type: none"> ○ The mean is equivalent to the first moment about the origin: $\mu = \mu'_1$ ○ The variance can be derived from the second moment about the origin and the mean: $\sigma^2 = \mu'_2 - \mu^2$ ○ Higher-order moments can be used to describe the shape of the distribution, such as skewness (μ'_3) and kurtosis (μ'_4) • Key: Hard to calculate these moments directly so we need a better way (ie. moment-generating functions).
Moment-Generating Functions
<p>The moment-generating function (MGF) of a RV X is</p> $M_X(t) = E[e^{tX}] = \begin{cases} \sum e^{tx} f(x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$ <ul style="list-style-type: none"> • Properties <ul style="list-style-type: none"> ○ MGFs uniquely determine the probability distribution of a random variable, provided they exist. ○ MGFs can simplify the process of finding moments, especially in the context of sums of independent random variables.

****Relationship between rth moment about the origin and MGF****

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

Process for using MGF to find moment (ie. w/o directly calculating the moment):

1. Find the MGF using the definition.
 - a. **Key:** Jim Davis' Calculus comes back in which you have to recognize the Maclaurin series.
2. Take rth derivative w.r.t t to find what you are looking for (i.e. mean, etc).
3. Sub $t = 0$.

L15: More on Moment Generating Functions

15.0 MGF of a Normal RV

Given a normal RV X with mean μ and variance σ^2 , the MGF, $M_X(t)$ is

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

15.1 Linear Combination of RVs

Given an RV X with PDF $f(x)$, the PDF of $Y = aX$, $h(y)$ is

1. Discrete RV X

$$h(y) = f\left(\frac{y}{a}\right)$$

- **Note:** This is going from $f(x) \rightarrow h(y)$

2. Continuous RV X

$$h(y) = \frac{1}{|a|} f\left(\frac{y}{a}\right)$$

- **Note:** This is going from $f(x) \rightarrow h(y)$

15.2 MGFs for Linear Combinations of RVs

Given X with MGF $M_X(t)$, the MGF of $Y = aX$ is:

$$M_{Y=aX}(t) = M_X(at)$$

- **Note:** This holds for both discrete and continuous cases.

15.3 PMFs and PDFs for Sum of RVs

Overview
<p>For independent RVs X and Y with marginal distributions $f(x)$ and $g(y)$, the marginal distribution of $Z = X + Y$</p> <p>1. Discrete Case</p> $h(z) = \sum_{w=-\infty}^{\infty} f(w)g(z - w)$ <p>2. Continuous Case</p> $h(z) = \int_{-\infty}^{\infty} f(w)g(z - w)dw$ <ul style="list-style-type: none"> • Note: We set $x = w$ (ie. <i>fixed value</i>) $\rightarrow z = x + y \rightarrow y = z - w$ • Note: This is starting off with the marginal distributions, not the joint distribution.

15.5 MGF of a Sum of Two RVs

Given two independent RVs X and Y with MGFs $M_X(t)$ and $M_Y(t)$ respectively, the MGF of $Z = X + Y$:

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

- **Extension:** The MGF of $aX + bY$ is $M_X(at)M_Y(bt)$.

Sampling (1.2-6)

(10) Sampling

L16: Sampling

16.0 Why do we sample?

Population
The entire set of individuals or items of interest in a statistical study. <ul style="list-style-type: none">• Note: Each observation represents an outcome of an RV.
Sampling
A subset of the population, selected for study to provide statistical information about the population. <ul style="list-style-type: none">• Why is sampling necessary? Full population measurements are often impractical or impossible to study an entire population.• What does a random sample do? A random sample aims to reflect the characteristics of the population from which it is drawn.
Importance of Random Sampling
<ul style="list-style-type: none">• Ensures that each member of the population has an equal chance of being included in the sample.• Helps to avoid bias and makes the results of the study more generalizable to the population.

16.1 Measures of Locations

Random Sample (Data Context)
<ul style="list-style-type: none">• Sample data: x_1, \dots, x_n<ul style="list-style-type: none">◦ Note: A random sample consists of n observations (L17).• Each x_i (i.e. observation i) is a realization of an independent RV, X_i for $i = 1, \dots, n$.

- The **joint distribution** of these observations is given by the **product of their individual distributions** (from L17):

$$f(x_1, \dots, x_n) = f(x_1) \cdot \dots \cdot f(x_n)$$

Sample mean

1. **Empirical value of the mean:**

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

2. **Random variable representing the sample mean**

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

Sample median

- Data ordered in increasing sequence $x_{(1)}, \dots, x_{(n)}$.

$$\text{Median} = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even,} \\ x_{((n+1)/2)} & \text{if } n \text{ is odd.} \end{cases}$$

Sample mode

The value that appears most frequently in the sample data.

16.2 Measures of Variability

Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- **Note:** This is the unbiased estimator of σ^2 as it compensates for the **loss of a degree of freedom** when using the **sample mean**, \bar{x} as an estimator for the **population mean**, μ .
 - **Note:** This is done using the $n - 1$

Understanding Sample Variance (i.e. this proves the unbiased estimator):

$$E[S^2] = \sigma^2$$

- **Explanation:**

- The expectation of the sample variance S^2 , $E[S^2]$ equals the population variance σ^2 , **making S^2 an unbiased estimator.**
- The derivation considers the variance of the sample mean and the variance of individual observations, demonstrating the need for the $n - 1$ denominator.

Alternate Formula:

If S^2 is the variance of a random sample of size n , then

$$S^2 = \frac{1}{n(n-1)} \left[n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 \right]$$

- **Interpretation on 2nd Formula:** Emphasizes the difference between the sum of squares of the observations and the square of the sum of the observations, normalized by the sample size, n and $n - 1$.

Sample Standard Deviation

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

Sample Range

The difference between the maximum and minimum values:

$$Range = \max(x_i) - \min(x_i)$$

16.3 Histograms

Overview

A graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.

Purpose

To give a visual impression of the distribution of data.

How it Works:

- Data is divided into intervals, known as bins.
- Frequency or count of data points within each bin is calculated.
- Each bin is represented by a bar, where the height reflects the frequency.

Interpretation:

- The shape of the histogram can reveal much about the underlying data distribution – whether it is normal, skewed, has outliers, etc.
- The spread of the histogram indicates the variability of the data.

16.4 Box-and-Whisker Plots (Box Plot)

Overview

Standardized way of displaying the distribution of data based on a five-number summary:

1. Minimum (i.e. left whisker)
 2. First quartile (Q1)
 3. Median (Q2)
 4. Third quartile (Q3)
 5. Maximum (i.e. right whisker)
- **Note:** After ordering, Q_i is the value in position $(n + 1) \times \frac{i}{4}$, for $i = 1, 2, 3$.

Components

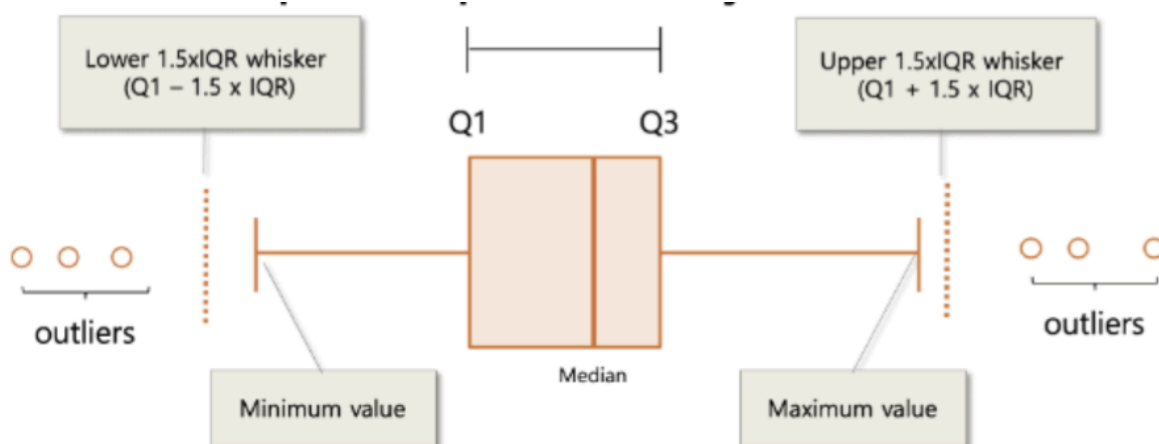
Box: The body of the plot representing the interquartile range (IQR) which is the distance between Q1 and Q3.

Whiskers: Lines extending from the box to the highest and lowest values, excluding outliers.

Median: A line across the box indicating the median value.

Outliers: Individual points plotted beyond the whiskers.

Diagram



Mistake: The right upper whisker should be $Q3 + 1.5 \times IQR$ and the median is $Q2$.

Steps

1. Set up the data points in increasing order.
2. Define $Q1 @ 0.25$, $Q2 @ 0.5$, and $Q3 @ 0.75$ using the formula: $(n + 1) \times \frac{i}{4}$, for $i = 1, 2, 3$ to find the position and therefore, the value.
3. Plot the box plot with $Q1$, $Q2$, and $Q3$.
4. Calculate $IQR = Q_3 - Q_1$
5. Calculate the whiskers (i.e. minimum value and maximum value)
 - a. $Q1 - 1.5 \cdot IQR$ (i.e. minimum value) and $Q3 + 1.5 \cdot IQR$ (i.e. maximum value)
6. Plot the whiskers, which is an estimation of the variability.

Sampling distributions (8.1-8)

(11) Sampling Distributions

L17: Sampling Distributions

17.0 Statistic

Overview
<p>A statistic is a function of the sample observations X_i for $i = 1, \dots, n$.</p> <ul style="list-style-type: none">• Eg. Sample mean, median, and variance.• Note: A sample is considered biased if it does not accurately represent the statistic of the population.
Sampling Distribution
<p>The probability distribution of a statistic.</p>
Key Facts
<ol style="list-style-type: none">1. If X_1 and X_2 are independent normal variables with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, then $X_1 + X_2$ has a normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.2. If X is normal with mean μ and variance σ^2, then $\frac{X}{n}$ is normal with mean $\frac{\mu}{n}$ and variance $\frac{\sigma^2}{n^2}$.3. If X_1, \dots, X_n are independent normals with mean μ and variance σ^2, then the sample mean \bar{X} is also normal with mean μ and variance $\frac{\sigma^2}{n}$. <ul style="list-style-type: none">• Note: This relationship doesn't hold for other distributions.

17.1 Central Limit Theorem (CLT)

Overview
<ul style="list-style-type: none">• Consider a sample X_1, \dots, X_n from a population.

- Each X_i is an independent and identically distributed (IID) random variable.
 - **Note:** X_i can be with any distribution.
- All X_i have a common mean μ and finite variance σ^2 .
- Focuses on the distribution of the **sample mean** \bar{X} (from L19).

Central Limit Theorem

If \bar{X}_n is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2 , then the limiting form of the distribution of

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

- **Note:** As $n \rightarrow \infty$, is the standard normal distribution $n(z; \mu = 0, \sigma = 1)$, regardless of the distribution of the original random variables.

Statistics

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \mid \mu_{\bar{X}} = \mu \mid \sigma_{\bar{X}} = s$
 - **1st Term:** The standard deviation of the sample mean is equal to the population variance normalized by the root of n .
 - **2nd Term:** The mean of the sample mean is equal to the population mean.
 - **Note:** Proof is in W8 HW.
 - **Key:** σ should be fixed among other samples.

Difference between normal Z and average Z:

- Normal Z has variance $\sigma^2 \rightarrow$ standard deviation is σ , while this average Z has variance $\frac{\sigma^2}{n} \rightarrow$ standard deviation is $\frac{\sigma}{\sqrt{n}}$. Hence, the denominators are changed.

Difference between μ and \bar{X}_n :

- μ is the population mean, and \bar{X}_n is the sample mean.

Process
<ol style="list-style-type: none"> 1. Recognize if you have to apply the central limit theorem. 2. Find the mean (i.e. average), \bar{X}_n. <ol style="list-style-type: none"> a. Note: For averages that involve the linear combination of each other (e.g. $\bar{X}_1 + \bar{X}_2$), you can derive the mean and variance from first principles. <ol style="list-style-type: none"> i. Eg. $E[\bar{X}_1 + \bar{X}_2]$ and $var(\bar{X}_1 + \bar{X}_2)$ and just plug in necessary formulas. ii. Key: You don't have to memorize formulas. 3. Apply the central limit theorem. 4. Go to the standard normal distribution table.
Discussion on the Central Limit Theorem
<ul style="list-style-type: none"> • Applicability: CLT is widely applicable to any distribution, as long as the observations are independent and identically distributed (IID) with finite variance. • Variance of Mean: The variance of the sample mean decreases as the sample size increases, scaling with \sqrt{n}. <ul style="list-style-type: none"> ◦ Note: This supports the common wisdom that averaging over a larger sample size yields a more accurate estimate of the mean. • Estimating μ: The CLT is particularly useful for estimating the population mean, μ. • Limitations: One limitation of CLT is the need to know the population standard deviation, σ, to apply it directly. <ul style="list-style-type: none"> ◦ This limitation is addressed by using the t-distribution when σ is not known.

L18: More Sampling Distributions

18.0 Distribution of Sample Variance and Chi-squared Distribution

Chi-squared Distribution
A RV Y with v degrees of freedom:

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

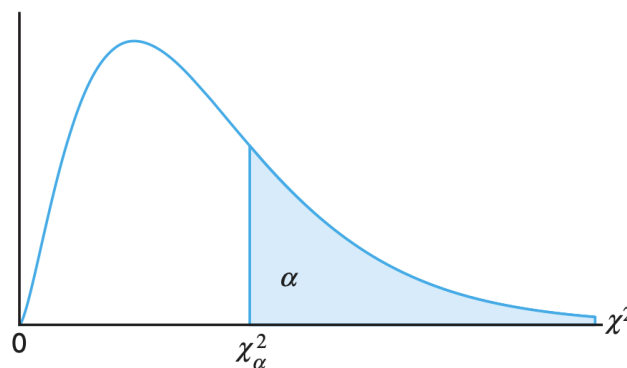
χ^2 - Chi-Squared Distribution

For samples X_1, \dots, X_n from a normal distribution with variance σ^2 :

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

- $v = n - 1$ (degrees of freedom)
- **Note:** With known μ , χ^2 follows a chi-squared distribution with $v = n$.
- **Note:** Using \bar{X} instead of μ reduces v by 1, leading to a higher variance in the chi-squared distribution.
- **Key:** The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the curve to the right of this value.
- **Key:** The probability of the chi-squared distribution gives the **probability of the sample variance**.

Pictorial Representation



- **Note:** Let χ^2_α represents the χ^2 value above which we find an area of α .

Process

1. Figure out whether or not you have to apply the Chi-squared distribution by seeing if you are finding the probability of the sample variance.
2. If so, apply $S^2 \Rightarrow \chi^2 = \frac{(n-1)S^2}{\sigma^2}$ inside the probability range.

a. **Note:** The probability will be of the form $P(\chi_{\alpha_1}^2 < \chi^2 < \chi_{\alpha_2}^2)$, in which the **bounds are the x-axis values** in the chi-squared, and we are looking for the probability.

b. **Note:** $P(\chi^2 > \chi_{\alpha}^2) = \alpha$, where χ_{α}^2 is the x-axis value, and α is the area under the curve.

3. Look in the table for the appropriate value given the **degrees of freedom** and χ_{α}^2 or α .

a. **Note:** To navigate the table quicker, look for the degrees of freedom first.

4. Draw the distribution to see what area you are looking for, and do simple geometric math to figure out how to get the area.

18.1 t-Distribution (Student t-distribution)

Overview
<ul style="list-style-type: none"> Used when the population variance (σ^2) is unknown. Particularly useful for small sample sizes ($n < 30$). When $n \geq 30$, the t-distribution approximates the normal distribution. For smaller samples, it provides a more accurate reflection of the uncertainty in the estimate of μ.
Intuition
<ul style="list-style-type: none"> If the population standard deviation σ were known, T would follow a normal distribution. Since σ is unknown, we use S as an estimate. This introduces more variability, hence the t-distribution accounts for this by having heavier tails (i.e. wider tails) than the normal distribution. <ul style="list-style-type: none"> The less observations we have, the more broad the tails will be in the t-distribution. The more observations as it gets to n, then the closer it will approximate to the normal distribution.
Sampling Statistic T:
<ul style="list-style-type: none"> Consider a sample with n observations: X_1, \dots, X_n. The sample mean and variance are denoted as \bar{X} and S^2 respectively. $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

- $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ (sample standard deviation)
- $v = n - 1$ (degrees of freedom)

Process

1. Figure out whether or not you have to apply the T distribution.
2. If so, apply $\bar{X} \Rightarrow T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ inside the probability range.
 - a. **Note:** $P(T > t_\alpha) = \alpha$, where α is the area under the curve, and t_α is the x-axis value at which the area to the right is α .
3. Look in the table for the appropriate value given the **degrees of freedom** and α or t_α .
 - a. **Note:** To navigate the table quicker, look for the degrees of freedom first.
4. Draw the distribution to see what area you are looking for, and do simple geometric math to figure out how to get the area.
 - a. **Key:** The table is based on the upper-tail probability (i.e. area under the curve), so when you are trying to find the t-Value, be careful.
 - i. If it's on the **right side**, then the t-value is given in the table.
 - ii. If it's on the **left side**, then the t-value will be the negative t-Value.
 - b. **Note:** $P(t > t_\alpha) = \alpha$ and $P(t < -t_\alpha) = \alpha$ (i.e. symmetric distribution about the y-axis).
 - c. **Two Different Cases:**
 - i. **Case 1:** If you care about both tails, alpha will be divided in the two intervals $\left(\frac{\alpha}{2}\right)$.
 - ii. **Case 2:** If you only care about one tail, you will focus on one alpha.
 - d. **Note:** The middle values are t_α .
 - e. **Note:** The area under the whole T distribution is 1 since it is normalized.

Miscellaneous

- **Difference between T and Z:**
 - For t-value problems, we have a probability and degrees of freedom and are finding the critical value T (ie. x-axis value) that gives the area under the curve to the left or right.
 - For Z-value problems, we have the Z value and were looking for the probability.
- **Difference between normal and t-distributions:** The tails widen out more for the

t-distribution.

18.2 F-Distribution

Sampling Statistic F

- Consider two independent samples of sizes n_1 and n_2 .
 - Each sample comes from normal distributions with variances σ_1^2 and σ_2^2 respectively.
 - Sample variances are S_1^2 and S_2^2 .
- $$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$
- The ratio follows an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

t, F-distributions (FROM L19)

Overview

t and F-distributions for **smaller samples** or when the **population variance is unknown**

Stuff

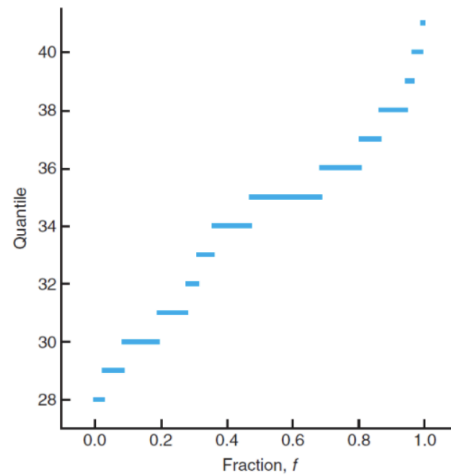
- Applicable for **any sample size n**, not just large samples.
- Assumes samples are from identical normal distributions.
- **Does not assume prior knowledge of variance σ^2 .**
- Degrees of freedom (n) represent the amount of information in the sample.
- **Uses sample variance S^2** instead of population variance σ^2 for the t-statistic.

(12) Quantiles

L19: Quantile and Probability Plots

19.1 Quantile and Quantile Plot

Quantile Overview
<p>A quantile of a sample, $q(f)$, is a value for which a specified fraction f of the data values is less than or equal to $q(f)$.</p> <ul style="list-style-type: none">• Key: The fraction indicates the percentage for the data value being equal to or less than to that specific data point compared to the sample data.<ul style="list-style-type: none">◦ Note: The fraction is an approximation, so it may not be accurate. As it should add up to 1, but may be a bit below.• Eg. The median value is $q(0.5)$, which makes sense since the median value would be in the middle.
Quantile Plot Overview
<p>A quantile plot displays data values against the proportion of observations they exceed.</p> <ul style="list-style-type: none">• The fraction for the i-th data point is $f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$ with n being the sample size.<ul style="list-style-type: none">◦ $i = 1, 2, \dots, n$• Ordered observations $y(i)$ (i.e. quantiles) are plotted against f_i.
Qualitative Observations of Quantile Plot
<p>This plot type helps visualize the sample's empirical distribution function.</p> <ul style="list-style-type: none">• Note: Flat areas indicate clusters of data.• Note: Steep areas indicate sparsity of data.



- **Note:** So on the x-axis, the fraction is what we calculate, and the lines are determined by the frequency of the same data.
- **Note:** If we switch the axes, the graph represents the CDF.

Process

1. Arrange the sample in increasing order, denoted as x_1, \dots, x_n .
2. Calculate the fractions for each index $i = 1, \dots, n$, where each index indicates a data point x_1, \dots, x_n .
3. For each individual point indexed by i , plot the pair, where the x-axis will be the fraction, and the y-axis will be the data points.
4. For the sample data that are the same, create a line to connect the points.

19.2 Relation of Quantile Function to CDF

Overview

Let X be a continuous RV with PDF $f(x)$ and CDF $F(x)$.

- $F(x)$ represents the probability that an outcome is less than or equal to x .
- Consider a 'theoretical' quantile: $q(f) = x$ such that a fraction f of the data is less than or equal to x .

Relationship

- If F is continuous and strictly increasing, then the quantile function q is the inverse of F , denoted as $q = F^{-1}$
 - **Note:** This is a simple axis swap (**i.e. switching the axis for quantile and fraction**).
 - **Note:** F being continuous or not strictly increasing makes it more complex.

19.3 Normal Quantile-Quantile Plots

Overview
<p>A normal quantile-quantile (Q-Q) plot is a graphical tool to compare a sample's distribution to a normal distribution.</p> <ul style="list-style-type: none"> • Note: The plot displays the ordered sample values against the theoretical quantiles from a standard normal distribution.
Approximating Quantiles
<p>For a normal RV X with mean μ and standard deviation σ, the quantile $q_{\mu,\sigma}(f)$ can be approximated by</p> $q_{\mu,\sigma}(f) = \mu + \sigma \{4.91[f^{0.14} - (1 - f)^{0.14}]\}$
Quantile of the Standard Normal $N(0, 1)$ RV
<p>The term $\{4.91[f^{0.14} - (1 - f)^{0.14}]\}$ provides a good approximation for the quantile of the standard normal $N(0,1)$ RV:</p> $q_{0,1}(f) = 4.91[f^{0.14} - (1 - f)^{0.14}]$
How to Use Normal Quantile-Quantile Plots
<ul style="list-style-type: none"> • This approximation is used in constructing Q-Q plots. • It simplifies the process of comparing empirical data quantiles with theoretical quantiles from the normal distribution. • The plot is constructed by plotting observed values against $q_{0,1}(f_i)$, where $f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$ for the i-th ordered observation $y(i)$.
Interpretation

- A **nearly straight-line pattern** indicates that the **sample distribution** is **approximately normal**.
 - **Note:** Deviations from straight lines suggest deviations from normality.
- The **slope** of the line in the Q-Q plot estimates the **standard deviation σ** , and the **intercept** estimates the **mean μ** .

Qualitative Observations of Quantile-Quantile Plot

- The sample quantiles come from 19.1, while the normal quantiles come from the normal distribution.

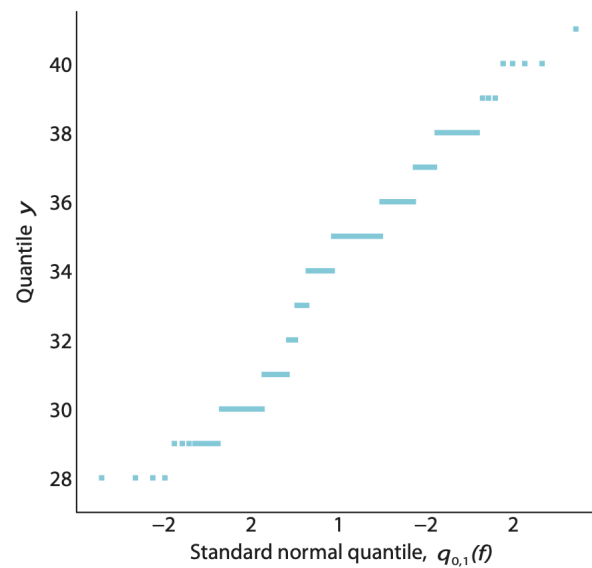


Figure 8.16: Normal quantile-quantile plot for paint data.

Estimation (9.1-14)

(13) Point Estimates

L20: Point Estimates and Confidence Intervals

20.0 Point Estimates

Overview
<ul style="list-style-type: none">Consider observed data values x_1, x_2, \dots, x_n.Assume they are realization of IID random variables X_1, X_2, \dots, X_n.
Point Estimates
<p>Purpose: Point estimates like the sample mean are used to infer properties of the population from the sample data.</p> <ol style="list-style-type: none">θ denotes the true parameter of the population, analogous to μ or σ^2.$\hat{\theta}$ is the point estimate observed from the sample, akin to \bar{x} or s^2.$\hat{\Theta}$ is the statistical estimator, similar to \bar{X} or S^2.
Unbiased Estimators
<p>An estimator $\hat{\Theta}$ is unbiased for a parameter θ if the expected value of $\hat{\Theta}$ is equal to θ, that is,</p> $E[\hat{\Theta}] = \theta$ <ul style="list-style-type: none">E.g. The sample variance, S^2, is an unbiased estimator of the population variance σ^2 (i.e. $E[S^2] = \sigma^2$).E.g. The sample mean is an unbiased estimator of μ since $E[\bar{X}] = \mu$.E.g. Each X_i is an unbiased estimator of μ since $E[X_i] = \mu$.
Efficient Estimators
<p>Among all unbiased estimators of a parameter θ, the one with the smallest variance is considered the most efficient estimator of θ.</p> <ul style="list-style-type: none">Intuition behind efficiency:

- An estimator $\hat{\theta}$ is a statistic and therefore has a probability distribution.
- An efficient estimator has the distribution with the **smallest spread**, meaning the smallest variance.
- **Example of why we pick the most efficient estimator (Why choose the sample mean in CLT?)**
 - The **variance decreases at a rate proportional to \sqrt{n}** , improving the precision of \bar{X} as an estimator for μ with larger samples.

(14) Types of Intervals

20.1 Interval Estimation and Confidence Intervals

Interval Estimation
<ul style="list-style-type: none"> Motivation: A point estimate, $\hat{\theta}$, is a single best guess at the true parameter value, θ, but it is rarely exact. <ul style="list-style-type: none"> Therefore, we use an interval estimate that provides a range, $\theta_L \leq \theta \leq \theta_U$, which likely contains θ.
Confidence Interval
$P(\theta_L \leq \theta \leq \theta_U) = 1 - \alpha$ <ul style="list-style-type: none"> θ_L and θ_U (the lower and upper bounds of a confidence interval are based on sample data) α (level of uncertainty we are willing to accept) <ul style="list-style-type: none"> Note: We choose this value. $1 - \alpha$ (confidence level)

20.2 Two-Sided Confidence Intervals for the Mean with Known Variance

Given
<ul style="list-style-type: none"> Based on IID observations: x_1, \dots, x_n Under the CLT, the underlying distribution must be normal or large n.
Statistic
$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
Confidence Intervals for the Mean with known σ^2
<p>A $100(1 - \alpha)\%$ chance that the true mean, μ, is within the interval around \bar{x} (for a two-sided case):</p> $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$ <ul style="list-style-type: none"> $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (margin error)

○ **Graphically:** This is the x-axis value on the standard normal distribution.

- $\theta_L = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (lower bound)
- $\theta_U = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (upper bound)

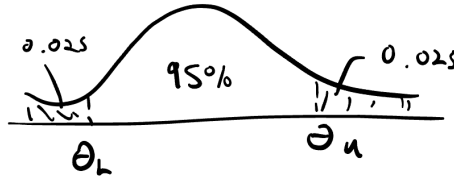


Figure: $1 - \alpha = 0.95$, $\alpha = 0.05$, and $\alpha/2 = 0.025$.

L21: Confidence Intervals and Prediction Intervals

21.0 One-Sided Confidence Intervals for the Mean with Known Variance

Overview
<ul style="list-style-type: none"> • Sample size: n • Observed sample mean: \bar{x} • Known population variance: σ^2
Statistic
$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
One-Sided Confidence Interval for the Mean
<p>For a one-sided interval to the right:</p> $P(Z \leq -z_{\alpha}) = 1 - \alpha$ <ul style="list-style-type: none"> • $z_{\alpha} = -\Phi^{-1}(\alpha)$ ($\Phi(x)$ is the CDF of the standard normal distribution).
Upper Bound for the One-Sided Interval
$\bar{X}_U = \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$ <ul style="list-style-type: none"> • Question: Why not \bar{X} (i.e. capitalized?)

Process
<ol style="list-style-type: none"> 1. See if the question applies 2. Calculate the values necessary using formulas you know. 3. Calculate z_α by going backwards from the inside (i.e. α) to the outside of the table. <ol style="list-style-type: none"> a. Key: The distribution is defined as a less than and equal to area.

21.1 Confidence Intervals for the Mean with Unknown Variance

Overview
<ul style="list-style-type: none"> • When σ is unknown and the distribution is normal: use the t-distribution.
Statistic
$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ with } n - 1 \text{ degrees of freedom.}$
Two-Sided Confidence Interval
$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = P\left(\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$ <ul style="list-style-type: none"> • $1 - \alpha$ (confidence level) • $t_{\alpha/2}$ (critical value is determined from the t-distribution's CDF [i.e. x-axis value]).

21.2 Standard Errors

Overview
<ul style="list-style-type: none"> • When we draw samples X_1, \dots, X_n from a distribution with unknown mean μ, variance σ^2, and the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. • As n grows large, by the CLT, the distribution of the statistic: $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ approaches a standard normal distribution $n(z; 0, 1)$.
Standard Error
$\frac{\sigma}{\sqrt{n}}$

- **Note:** It quantifies the variability of the sample mean \bar{X} as an estimate of the population mean μ .
- **Note:** $z_{\alpha/2}(\text{standard error}) = \text{margin error}$

21.3 Prediction Intervals With Known Variance

Overview
<ul style="list-style-type: none"> • With normal samples RV X_1, \dots, X_n, each with variance σ^2, we have a sample mean \bar{X}. • For a new observation X_0 (prediction), we use \bar{X} as a point estimate. • The error $X_0 - \bar{X}$ has variance $\sigma^2 + \frac{\sigma^2}{n}$ due to independence.
Given
<ul style="list-style-type: none"> • Based on IID observations: x_1, \dots, x_n
Statistic
$Z = \frac{X_0 - \bar{X}}{\sigma \sqrt{1 + \frac{1}{n}}}$ <ul style="list-style-type: none"> • Note: Z follows a standard normal distribution $n(z; 0, 1)$
Prediction Interval
<p>A $100(1 - \alpha)\%$ chance that the next observation, x_0, will fall within the interval around \bar{x} (for two-sided case):</p> $\bar{x} - z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}} \leq x_0 \leq \bar{x} + z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}}$ <ul style="list-style-type: none"> • Note: Outliers can be detected if a new observation falls outside this interval.

L22: Tolerance Limits and Two Samples

22.0 Tolerance Intervals for Normal Distributions with unknown mean and standard deviation

Overview
<p>Tolerance limits define a range within which we expect a certain proportion of the population to fall.</p> <ul style="list-style-type: none">• Difference with confidence/prediction intervals: Unlike confidence or prediction intervals, tolerance limits <u>do not shrink with increasing sample size</u> but are intended to encompass a <u>fixed percentage of the population</u>.
Given
<ol style="list-style-type: none">1. IID observations: x_1, \dots, x_n2. Sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$3. Sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Tolerance Interval
<p>Assert with $100(1 - \gamma)\%$ confidence that $100(1 - \alpha)\%$ of all future measurements are expected to fall within the interval $\bar{x} \pm ks$ (for a two-sided case):</p> $\bar{x} \pm ks$
Selection of k
<p>The multiplier k is chosen such that a specified percentage (e.g., $100(1 - \alpha)\%$) of the population lies within the tolerance limits.</p> <ul style="list-style-type: none">• Table A.7: Statistical table of k.

22.1 Comparing the Mean of Two Independent Samples

Givens
<ol style="list-style-type: none">1. Sample sizes: n_1 and n_2

2. Sample means: \bar{X}_1 and \bar{X}_2
3. Population means: μ_1 and μ_2
4. Population variances: σ_1^2 and σ_2^2

Estimating the Difference Between Two Means

1. **Point estimate for $\mu_1 - \mu_2$:** $\bar{X}_1 - \bar{X}_2$
2. **Mean:** $\mu_1 - \mu_2$
3. **Variance:** $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Statistic:

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

- **Key:** This is always the same format with \bar{X} , μ , and σ , but in different expressions.
- **Note:** The R.V. $\bar{X}_1 - \bar{X}_2$ is approximately normal because each sample mean is approximately normal due to the CLT and the difference of two normally distributed variables is also normally distributed.

22.2 Confidence Intervals with Two Samples

1. Estimating the difference between two means with known σ .

Given

1. \bar{x}_1 and \bar{x}_2 are means of independent random samples
2. Sizes n_1 and n_2 from populations.
3. Known variances σ_1^2 and σ_2^2

Confidence Interval

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

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

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

2. Estimating the difference between two means with unknown (but equal) σ .

Conditions

1. Assume $\sigma_1 = \sigma_2$, but both are unknown.
2. Samples sizes $n_1, n_2 < 30$, *normally distributed*.

Test Statistic

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

- $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ (Pooled estimate of variance, where S_p^2 is the sample size-weighted average of S_1^2 and S_2^2).

Degrees of Freedom

$$v = n_1 + n_2 - 2$$

Two-Sided Confidence Intervals

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

$$\left(\bar{x}_1 - \bar{x}_2\right) - t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < \left(\bar{x}_1 - \bar{x}_2\right) + t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Derivation

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

$$P\left(-t_{\alpha/2} \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) \leq t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

$P\left(-\left(\bar{x}_1 - \bar{x}_2\right) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq -\left(\mu_1 - \mu_2\right) \leq -\left(\bar{x}_1 - \bar{x}_2\right) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$ $P\left(\left(\bar{x}_1 - \bar{x}_2\right) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \geq \left(\mu_1 - \mu_2\right) \geq \left(\bar{x}_1 - \bar{x}_2\right) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$ $P\left(\left(\bar{x}_1 - \bar{x}_2\right) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\bar{x}_1 - \bar{x}_2\right) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$
3. Estimating the difference between two means with unknown (and different) σ.
Given
<ul style="list-style-type: none"> σ_1/σ_2 are unknown and unlikely to be equal.
Test Statistic
$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Degrees of Freedom
$v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ <ul style="list-style-type: none"> Note: Rounded down to the nearest integer, no matter what.
Confidence Intervals
$P\left(-t_{\alpha/2} \leq T' \leq t_{\alpha/2}\right) \approx 1 - \alpha$ <ul style="list-style-type: none"> Note: Pooled variance combines the variances of two samples to estimate a common variance, which is used when the sample variances are assumed to be equal.

L23: Paired Observations and Estimating the Binomial Parameter

23.0 Confidence Intervals for Paired Observations

Overview
Comparison of two related samples, with one-to-one correspondence between measurements.

Given
<ol style="list-style-type: none"> Let (X_i, Y_i) represent the paired samples for $i = 1, \dots, n$. Assume X_i and Y_i are normally distributed with means μ_X, μ_Y and standard deviations σ_X, σ_Y. $D_i = X_i - Y_i$ for each pair. Variance of D_i: $Var(D_i) = \sigma_X^2 + \sigma_Y^2 - 2Cov(X_i, Y_i)$ and $S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$ <ol style="list-style-type: none"> Note: The covariance may have to be used since the variables x and y are not independent.
Definition
<p>A $100(1 - \alpha)\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is</p> $\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}}$ <ul style="list-style-type: none"> $t_{\alpha/2}$ (t-value with $v = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right). \bar{d} and s_d are the mean and standard deviation, respectively, of the normally distributed differences of n random pairs of measurements.
Derivation
$P(\bar{d}_L < \mu_d < \bar{d}_U) = 1 - \alpha$ $P\left(t_{\alpha/2} < \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$ $P\left(t_{\alpha/2} \frac{s_d}{\sqrt{n}} < \bar{d} - \mu_d < t_{\alpha/2} \frac{s_d}{\sqrt{n}}\right) = 1 - \alpha$ $P\left(-\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}} < -\mu_d < -\bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}}\right) = 1 - \alpha$ $P\left(\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}}\right) = 1 - \alpha$

23.1 Confidence Intervals Estimating a Proportion (single sample)

Overview

A point estimator of the proportion p in a binomial experiment is given by the statistic $\hat{P} = \frac{X}{n}$, where X represents the number of successes in n trials.

Therefore, the sample proportion $\hat{p} = \frac{x}{n}$ will be used as the point estimate of the parameter p .

Confidence Intervals Estimating a Proportion

By the CLT, for n sufficiently large, \hat{P} is approximately normally distributed with mean $\mu_{\hat{p}} = p$ and variance $\sigma_{\hat{p}}^2 = \frac{pq}{n}$, and:

$$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

- When n is large, very little error is introduced by substituting the point estimate $\hat{p} = \frac{x}{n}$ for the p under the radical sign. Then we can write the confidence interval for p .

Can we estimate $p = P(Y_i = 1)$

- Estimator: $\hat{P} = \frac{X}{n}$

1. If n is large, replace p with $\hat{p} = \frac{x}{n}$ in the standard error:

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

23.2 Choice of Sample Size

Overview

Suppose we want to be $100(1 - \alpha)\%$ confident that the estimation error is less than δ , then

$$n = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{\delta^2}$$

- **Note:** This comes from the margin error formula: $\delta = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- **Note:** Always round up.

Since $\hat{p}(1 - \hat{p}) \leq 0.25$, therefore,

$$n \geq \frac{z_{\alpha/2}^2}{4\delta^2}$$

- **Note:** This is to get a safe lower bound for n.
- **Note on testable information:** All these confidence intervals cannot just be used to find the confidence interval, but can be used to theoretically get every variable.
- **Note:** No chi-squared/F-distribution.

Flow Chart

1. Mean (\bar{X}):
 - a. Is σ known?
 - i. For all n values, use the normal distribution.
 - b. Is s known?
 - i. Is $n \gg 30$?
 1. Use the normal distribution
 - ii. Is $n < 30$?
 1. Use the t-distribution.
2. CLT:
 - a. Are you looking for probability of \bar{X} ?
 - i. Finite, known variance σ^2 (i.e. population variance is known)?
 1. Yes, are the data values IID?
 - a. Yes, are the values either normal?
 - i. Use CLT.
 - b. Yes, is $n > 30$?
 - i. Use CLT.

L24: Estimating Variance, Maximum Likelihood Estimation

24.0 Estimating Variance

Overview

If sample of size n is drawn from a **normal population**, then the statistic S^2 is an estimator of σ^2 .

- **Note:** Need to work with chi-squared no matter the sample size.

Statistic

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

Degrees of Freedom

$$v = n - 1$$

Probability

$$P\left(\chi_{1-\alpha/2}^2 < X^2 < \chi_{\alpha/2}^2\right) = 1 - \alpha$$

$$P\left(\chi_{1-\alpha/2}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2}^2\right) = 1 - \alpha$$

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

Two-Sided Confidence Interval

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

- $\chi_{\alpha/2}^2$ (leaving an area of $\alpha/2$ to the right)
- $\chi_{1-\alpha/2}^2$ (leaving an area of $1 - \alpha/2$ to the right)

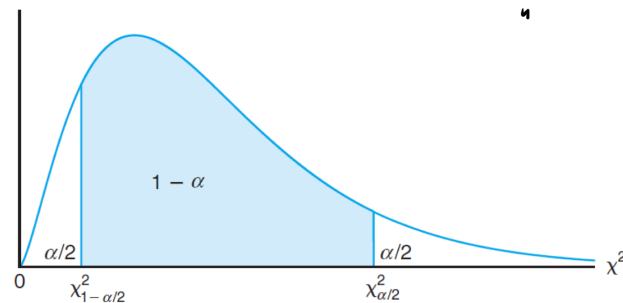


Figure 9.7: $P(\chi_{1-\alpha/2}^2 < X^2 < \chi_{\alpha/2}^2) = 1 - \alpha$.

(15) Maximum Likelihood Estimation

24.1 Maximum Likelihood Estimator

Overview
MLE provides a method for estimating parameters by maximizing the likelihood function.
Likelihood Function
Given independent observations x_1, \dots, x_n from a PDF (continuous case) or PMF (discrete case): $L(x_1, \dots, x_n; \theta) = f(x_1, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^n f(x_i; \theta)$
MLE
The maximum likelihood estimator of θ is $\hat{\theta}$ such that it maximizes the likelihood function: $\hat{\theta} \text{ s. t. } \frac{\partial L}{\partial \theta} = 0$

L25: Maximum Likelihood Estimation

25.0 Motivation for the Log-Likelihood

Taking the logarithm of the likelihood function simplifies the process:

- Removes the exponential, resulting in a simpler function.
- Use the law of logarithms.
- Since the logarithmic function is strictly increasing, the maximum of the likelihood is preserved when taking the log:
 - For any two numbers, if $x < y$, then $\log(x) < \log(y)$
 - Therefore, the argument that maximizes the likelihood also maximizes the log-likelihood.

25.1 Steps to Find the MLE

1. Find the likelihood function.

2. **Log-likelihood function:** Take the logarithm of the likelihood function to use logarithm properties of summation.
3. Take the derivative and set equal to 0 to find the maximum likelihood estimator.
 - a. **Note:** If you are finding the maximum likelihood for finding estimates of two parameters. Take the derivative w.r.t each variable and find both.
 - b. **Note:** Once you set the derivative to 0, you are using estimators for all the parameters.

25.2 Common Product Notation Rules

1. $\prod_{i=1}^n a = a \cdot \dots \cdot a = a^n$
2. $\prod_{i=1}^n ab = \prod_{i=1}^n a \cdot \prod_{i=1}^n b = a^n \cdot b^n$
3. $\prod_{i=1}^n e^{x_i} = e^{\sum_{i=1}^n x_i}$
4. $\prod_{i=1}^n x_i^b = \left(\prod_{i=1}^n x_i \right)^b$

25.3 Common Summation Notation Rules

1. $\sum_{i=1}^n a = an$
2. $\sum_{i=1}^n a_i + b_i = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i$
3. $\ln\left(\prod_{i=1}^n x_i\right) = \ln(x_1 \cdots x_n) = \sum_{i=1}^n \ln(x_i)$

25.4 Laws of Logarithms

1. $\ln(mn) = \ln(m) + \ln(n)$
2. $\ln\left(\frac{m}{n}\right) = \ln(m) - \ln(n)$
3. $\ln(m^k) = k\ln(m)$

Hypothesis testing (10.1-14)

(16) Type 1 and Type 2 Errors

L26: Hypothesis Testing

26.0. Hypothesis Terminology

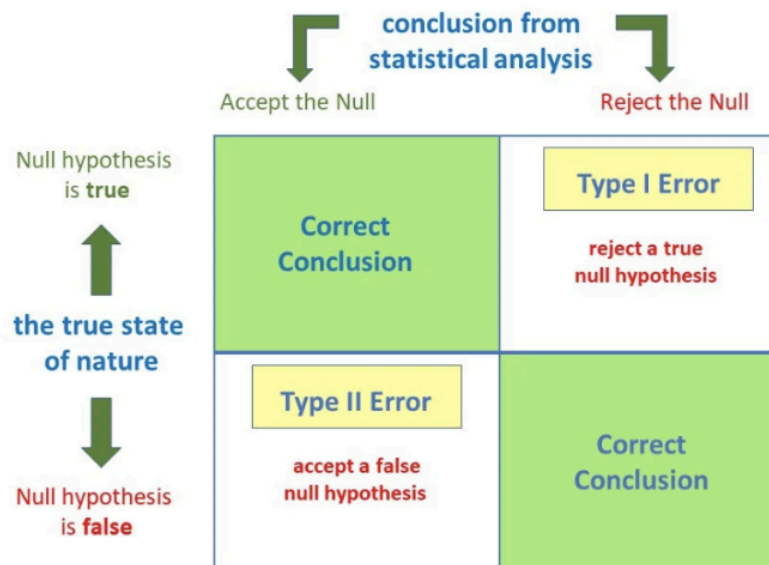
Overview
A hypothesis is a specific claim regarding a population(s) parameter .
Null/Alternative Hypothesis
<ol style="list-style-type: none">1. Null Hypothesis H_0: Represents the “status quo” or a default position.2. Alternative Hypothesis H_1: Represents a contrary claim to be tested against the null hypothesis.
Testing Process
Based on the sample x_1, \dots, x_n we decide whether to: <ol style="list-style-type: none">1. Reject H_0 in favor of H_1.<ol style="list-style-type: none">a. Note: If strong evidence against the null hypothesis.2. Fail to reject H_0, maintaining the status quo<ol style="list-style-type: none">a. Evidence doesn't prove H_0, but indicates insufficient evidence to prove H_1.
Type I Error (False Positive) [i.e. falsely rejected when true]
<ul style="list-style-type: none">• Occurs when the null hypothesis H_0 is incorrectly rejected when it is actually true.• α (probability of committing a Type I error, a.k.a level of significance).
Type II Error (False Negative) [i.e. failed to notice that its false]
<ul style="list-style-type: none">• Occurs when the null hypothesis H_0 is not rejected when it is actually false.• Denoted by β (probability of committing a Type II error).

Critical Region (Rejection Region)

The set of all possible values of the test statistic for which the null hypothesis is rejected based on the α (i.e. probability of rejecting the null hypothesis when it is actually true (a Type I error))

- **Compare this to the p-value.**

Summary



- **Note:** $\alpha + \beta \neq 1$
- **Correct conclusion:** In the **bottom right** is associated with **statistical power**.
- **Question:** How to read this diagram?

26.1 How to apply hypothesis testing?

Given

1. Null hypothesis H_0
2. Alternative hypothesis H_1
 - a. **Note:** H_1 can be \neq , $<$, and $>$.
3. Sample with mean \bar{x} .

Process

1. Assume the **null distribution** is normal centered around μ for H_0 and **assume the null**

hypothesis is true.

2. Create arbitrary bounds for the null hypothesis on $X_L < \bar{x} < X_U$, where we don't reject H_0 .
3. Otherwise, the **critical region** is the area outside of these bounds where we **reject** H_0 .

a. **Note:** The area is α (i.e. probability of committing a Type I error).

b. **Note:** This means it's a false positive, where we are incorrectly rejecting since everything is **assumed to be true in the null hypothesis**.

4. **Type I Error:** Use CLT to determine the **critical region** probability, α , which is the complement of the bounds (i.e. range leading to rejection of H_0):

$$1 - P(X_L < \bar{x} < X_U) = 1 - P(Z_L < Z < Z_U) = \alpha$$

a. **Key:** Use the parameter of the **null hypothesis** (e.g. μ)

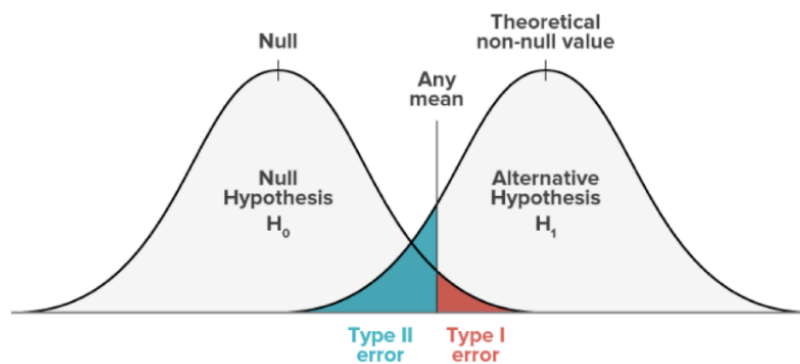
5. Assume the **alternative distribution** is normal centered around μ for H_1 and **assume the alternative hypothesis is true**.

6. **Type II Error:** Use CLT to determine the region in which it falls in the range of $X_L < \bar{x} < X_U$ but under the area of the **alternative hypothesis** (i.e. failing to reject the null hypothesis when it is false):

$$P(X_L < \bar{x} < X_U) = P(Z_L < Z < Z_U) = \beta$$

a. **Key:** Use the parameter of the **alternative hypothesis** (e.g. μ of the alternative hypothesis)

Recap



- **Note:** This doesn't relate to the process above, however, "Any mean" relates to X_U , but there is no lower bound in this case.

How to change α and β ?

1. Increasing the sample size (n) typically decreases both α & β .

2. A larger critical region will reduce α but may increase β .
3. A more distinct H_1 that greatly differs from H_0 will usually reduce β but might increase α .

26.2 Statistical Power

The **power** of a test is the probability that the test **correctly rejects** the **null hypothesis** (H_0) when a specific **alternative hypothesis** (H_1) is **true**:

$$1 - \beta$$

- β (probability of committing a Type II error).

(17) Hypothesis Testing Terminology

L27: More Hypothesis Testing

27.0 One- and Two- Tailed Tests

Overview

To determine which of three options you have for H_1 , there are two options:

1. **Hard way:** The alternative hypothesis is determined by seeing what we want to gather evidence about.
2. **Easy way:** From the question, you can see where the observed value is, and just pick the interval closest to the observed value.

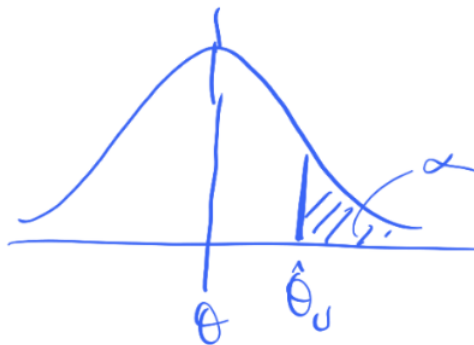
One-Tailed Test

A test of any statistical hypothesis where the alternative is one-sided.

1st Option:

$$H_0: \theta = \theta_0$$

$$H_1: \theta > \theta_0$$



2nd Option:

$$H_0: \theta = \theta_0$$

$$H_1: \theta < \theta_0$$

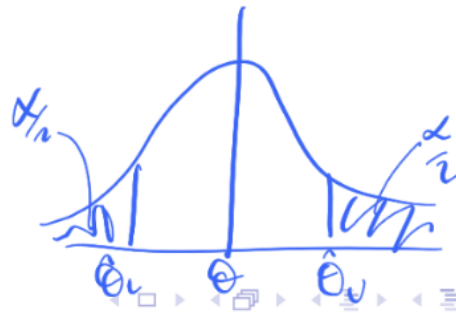


Two-Tailed Test

A test where the alternative hypothesis is two-sided.

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$



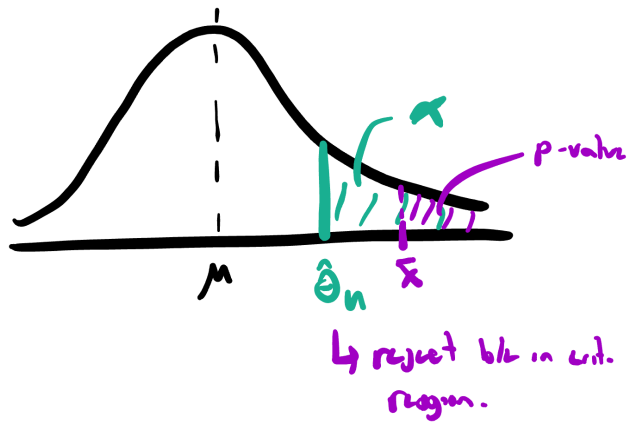
27.1 P-Values in Hypothesis Testing

Definition of P-Value

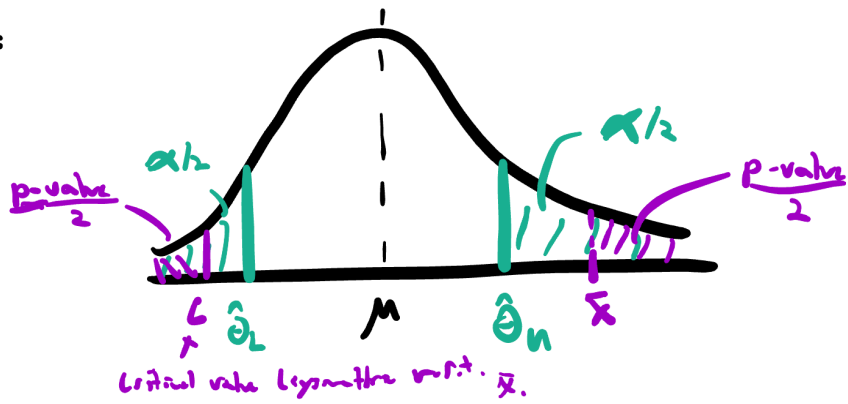
Given H_0 is true, the probability of observing a test statistic is more extreme than the observed value.

- **Usage:** This quantifies the strength of evidence against H_0 to see whether or not to reject H_0 .

Example of One-Sided



Example of Two-Sided



(18) Hypothesis Testing

27,0 Hypothesis Testing for the Mean (two-sided)

Hypothesis Statement:
$H_0: \mu = \mu_0 \mid H_1: \mu \neq \mu_0$
Given
<ul style="list-style-type: none"> • Sample: x_1, \dots, x_n • Known variance: σ^2 • Type I error probability: α
Statistic (if CLT holds)
$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$
Non-rejection Region for H_0
$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$
Calculating P-Value (if σ known)
<p>1. Two-Sided</p> $P - value = 2P(Z > z)$ <p>2. Lower One-Sided for $H_1: \theta < \theta_0$</p> $P - value = P(Z < z)$ <p>3. Upper One-Sided for $H_1: \theta > \theta_0$</p> $P - value = P(Z > z)$ <ul style="list-style-type: none"> • Note: The z value is calculated with the sample mean from the question. • Note: If σ unknown, use the t-distribution.
Decision Rule:
<p>1. Probabilities:</p> <p>a. $P_{crit.} \leq \alpha$: reject the null</p>

b. $P_{crit.} > \alpha$: fail to reject the null.

- **Note:** We are comparing the probabilities of the tails.

2. Z Values:

a. $Z_{crit.} > Z_{\alpha}$: reject the null.

b. $Z_{crit.} < Z_{\alpha}$: fail to reject the null.

27.1 Hypothesis Testing for the Mean with Unknown σ

Statistic:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

27.2 Hypothesis Testing for Two Means

Overview

- Null hypothesis:
 - $H_0: \mu_1 - \mu_2 = d_0$
- Alternative hypothesis (two-sided)
 - $H_1: \mu_1 - \mu_2 \neq d_0$
- Alternative hypothesis (one-sided)
 - $H_1: \mu_1 - \mu_2 > d_0$
 - $H_1: \mu_1 - \mu_2 < d_0$

Known σ_1 and σ_2

Test Statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- \bar{X}_1, \bar{X}_2 (sample means)
- σ_1^2, σ_2^2 (population variances)
- n_1, n_2 (sample sizes)

Unknown $\sigma_1 = \sigma_2$	
Test Statistic	
Test Statistic	$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ <ul style="list-style-type: none"> • $s_p^2 = \frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1 + n_2 - 2}$ • \bar{X}_1, \bar{X}_2 (sample means) • s_1^2, s_2^2 (sample variances) • n_1, n_2 (sample sizes)
Degrees of Freedom	$v = n_1 + n_2 - 2$
Unknown $\sigma_1 \neq \sigma_2$	
Test Statistic	$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
Degrees of Freedom	$v \approx \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$

27.3 Hypothesis Testing for Paired Observations

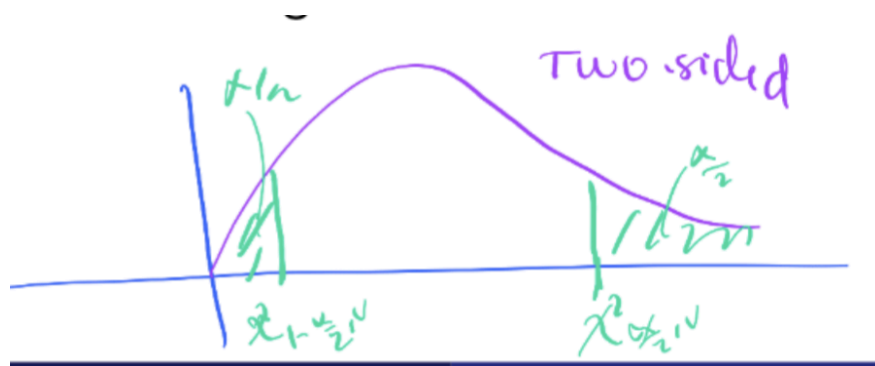
Overview
<p>When data are in the form of paired observations, the test for two means can be approached as a one-sample problem with the computed differences of the paired observations. Assuming the observations are normal, the hypothesis for the paired observations is given by</p> <ul style="list-style-type: none"> • Null hypothesis

<ul style="list-style-type: none"> ○ $H_0: \mu_D = \mu_1 - \mu_2 = d_0$ ● Alternative hypothesis (two-sided) <ul style="list-style-type: none"> ○ $H_1: \mu_D \neq d_0$ ● Alternative hypothesis (one-sided) <ul style="list-style-type: none"> ○ $H_1: \mu_D > d_0$ ○ $H_1: \mu_D < d_0$
Statistic
<ul style="list-style-type: none"> ● Test statistic: <ul style="list-style-type: none"> ○ $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$ ● Degrees of freedom <ul style="list-style-type: none"> ○ $v = n - 1$

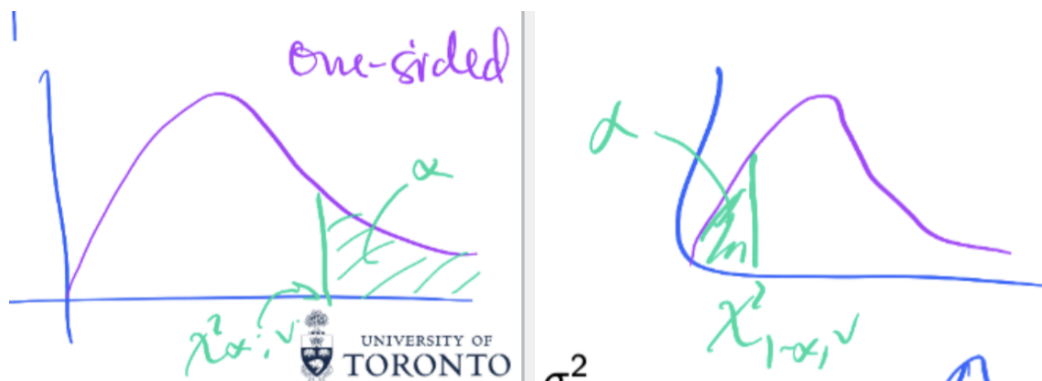
27.4 Hypothesis Testing for One Variance

Hypotheses
Null Hypothesis: $H_0: \sigma^2 = \sigma_0^2$ Alternative Hypothesis (two-sided): $H_1: \sigma^2 \neq \sigma_0^2$ Alternative Hypothesis (one-sided): $H_1: \sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$
Test Statistic
$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \text{ with } v = n - 1$
Applying in problems
<ul style="list-style-type: none"> ● Note: there are two ways to determine failing to reject the null or rejecting the null <ul style="list-style-type: none"> ○ Use the p-value or use the x-axis value (i.e. chi-squared value) ○ If $p\text{-value} > \alpha$, then we fail to reject the null. ● Note: On the bounds, if the alternative hypothesis is $\sigma^2 > \sigma_0^2$, then we are finding an upper bound where $P(\chi^2 > \text{some number})$

Two-Sided



One Sided:



27.5 Hypothesis Testing for Two Variances

Hypotheses

Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Alternative Hypothesis (two-sided): $H_1: \sigma_1^2 \neq \sigma_2^2$

Alternative Hypothesis (one-sided): $H_1: \sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$

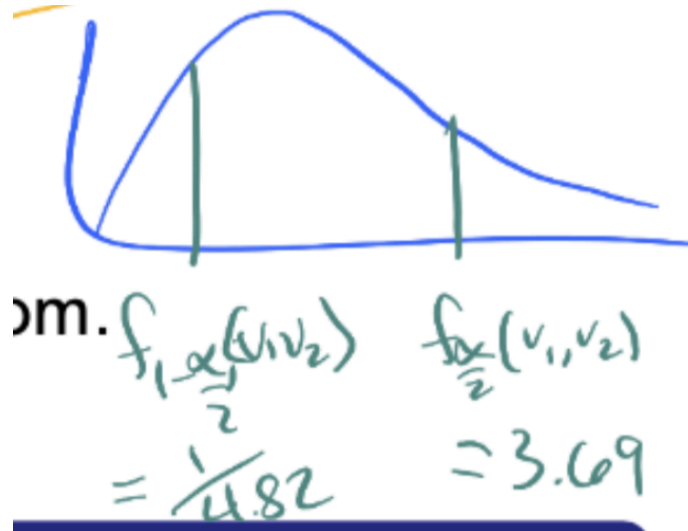
Test Statistic:

$$f = \frac{s_1^2}{s_2^2} \text{ with } v_1 = n_1 - 1 \text{ \& } v_2 = n_2 - 1$$

Left Tail

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$

- This allows us to calculate the left tail since the tables only have $\alpha = 0.05$ or 0.01 , so in the table, it has v_2 along the columns, and v_1 along the rows.
- In this case, the roles of **v_2 and v_1 get switched here**, where $v_1 = v_2$ (rows) and $v_2 = v_1$ (columns)



(19) Goodness of Fit

27.6 Goodness-of-Fit Test

Overview
<ul style="list-style-type: none"> Discrete RV with possible outcomes $i = 1, \dots, k$ n trials with expected frequencies $e_i = nP(i)$ for outcomes $i = 1, \dots, k$ Observed frequencies o_i for outcomes i. O_i is the RV.
Test
$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \text{ with } v = k - 1$ <ul style="list-style-type: none"> χ^2 (value of RV whose sampling distribution is approximated by the chi-squared distribution). $v = k - 1$ (degrees of freedom). <ul style="list-style-type: none"> Note: THIS IS NOT THE NUMBER OF TRIALS, BUT THE NUMBER OF OUTCOMES. o_i (observed frequencies) e_i (expected frequencies) <p>Note: Since we're not talking about a specific parameter, we are not using a hypothesis.</p>
Interpreting Results
<ul style="list-style-type: none"> χ^2-value small \rightarrow good fit b/w observed and expected frequencies. χ^2-value large \rightarrow bad fit. <p>Critical Region Definition:</p> <ul style="list-style-type: none"> The critical region is in the right tail of the chi-squared distribution for a significance level α, found using the critical value from Table A.5. Then, $\chi^2 > \chi_{\alpha}^2$ constitutes the critical region for rejection. Note: Always focused on the right tail here.

Linear regression (11.1-12)

(20) Linear Regression

L28: Regression

Basic Setup
<ul style="list-style-type: none"> Input/output pairs: $(x_i, y_i), i = 1, \dots, n$ Want a function $y = f(x)$ that minimizes errors: $e_i = y_i - f(x_i), i = 1, \dots, n$
Population or True Linear Regression Line
$\mu_{Y x} = Y = \beta_0 + \beta_1 x$ <ul style="list-style-type: none"> Y (response variable) x (regressor variable) β_0 (intercept) β_1 (slope) Deterministic: There is no random or probabilistic component to it in this form.
Simple Linear Regression Model
$Y = \beta_0 + \beta_1 x + \epsilon$ <ul style="list-style-type: none"> β_0 (unknown intercept) β_1 (slope) ϵ (random error term) <ul style="list-style-type: none"> $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2$ <p>What does this model imply?</p> <ul style="list-style-type: none"> Response variable Y is not deterministic due to ϵ. Regressor variable x is assumed to be measured without significant error. Random error ϵ is assumed to have a constant variance (homoscedasticity) <p>What does this linear model ensure?</p> <p>Ensures that the y-values are distributed around the population regression line</p>

$$\mu_{Y|x} = Y = \beta_0 + \beta_1 x$$

- **Note:** This is theoretical (don't know it), w/o randomness model that we are trying to estimate with the fitted regression line.
- **Note:** If the model fits well, it captures the essence of the data, allowing for both positive and negative deviations due to random error.

Graph:

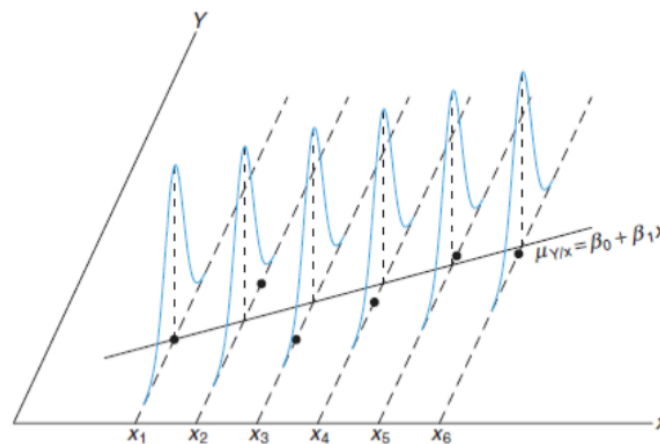


Figure 11.4: Individual observations around true regression line.

Key: The true regression line is plotted, and 6 data points, however, we see there is a normal distribution (blue curves) in this 3D plot that captures where the point is most likely to be at.

- The normal distribution is captured in the error term for the true regression line:
 - $Y = \beta_0 + \beta_1 x + \epsilon$

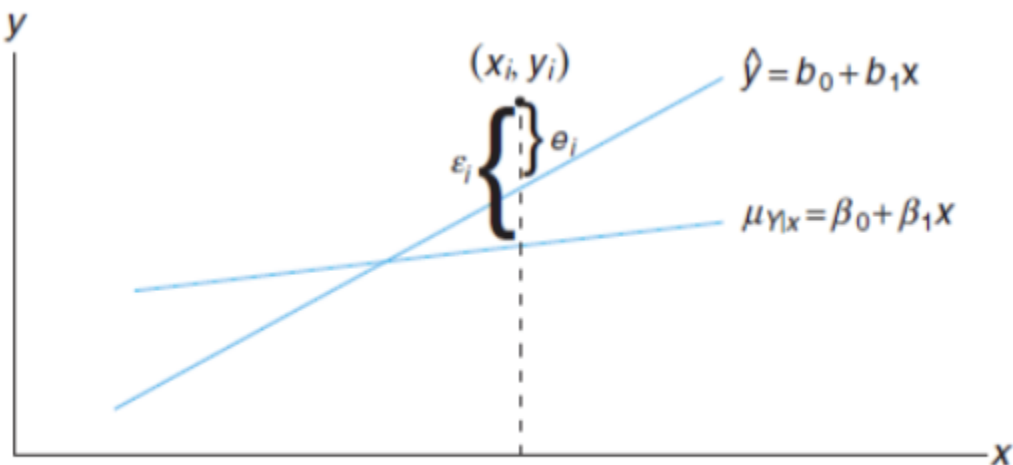
28.0 Fitted Regression Line

Overview:

Since we will never know β_0 and β_1 , therefore, assuming b_0 and b_1 as the estimates for β_0 and β_1 , therefore, the fitted (estimated) regression line is

$$\hat{y} = b_0 + b_1 x$$

- \hat{y} (predicted values on the fitted regression line)
- **Note:** This is fitted with sample data, so there is no error associated with it since the values of the sample are deterministic.
- **Note:** Serves as an estimate of the true regression line, which we aim to approximate.

Residual Error in Regression
<p>Given $\{(x_i, y_i); i = 1, 2, \dots, n\}$ and $\hat{y} = b_0 + b_1 x$, then the residual e_i is</p> $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$ <ul style="list-style-type: none"> Note: Large set of residuals indicates a poor fit, whereas small residuals suggest a good fit.
Useful Relationship
$y_i = b_0 + b_1 x_i + e_i$
Difference between Residual e_i and the unobserved model errors ϵ_i
 <p style="text-align: center;">Figure 11.5: Comparing ϵ_i with the residual, e_i.</p>
<ul style="list-style-type: none"> Note: ϵ_0 is the error from the true regression line to the point. Note: e_i is the error from the fitted regression line to the point.

28.1 Estimating the Regression Coefficients

Intercept
$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \bar{y} - b_1 \bar{x}$
Slope

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

28.2 Mean and Variance of Estimators

Slope

$$\mu_{B_1} = E[B_1] = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1$$

$$\sigma_{B_1}^2 = \sigma_{B_1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_{Y_i}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- **Note:** For the variance of B_1 , $\sigma_{Y_i}^2$ is the variance for the corresponding x_i , which is shown in the true regression line with the normal distributions.
 - The normal distributions are the blue curves with the same variance since they come from the same error, but different means.
 - Therefore, the variances are the same, so $\sigma_{Y_i}^2$ doesn't rely on the sum, so we can take it out and replace it with σ^2 .

Intercept

$$\mu_{B_0} = E[B_0] = \beta_0$$

$$\sigma_{B_0}^2 = \left(\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right) \sigma^2$$

Analogy

$$\bar{x} \rightarrow b_1, b_0$$

$$\bar{X} \rightarrow B_1, B_0$$

$$E[\bar{X}] = \mu \rightarrow E[B_1] = \beta_1, E[B_0] = \beta_0$$

(21) Analysis of Linear Regression

L29: Analysis of Linear Regression

29.0 Sums of Error

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} = \left(\sum_{i=1}^n x_i^2\right) - n(\bar{x})^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = \left(\sum_{i=1}^n y_i^2\right) - n(\bar{y})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n} = \left(\sum_{i=1}^n x_i y_i\right) - n\bar{x}\bar{y}$$

29.1 Sum of Squared Errors (SSE)

$$SSE = S_{yy} - b_1 S_{xy}$$

29.2 Unbiased Estimator for σ^2

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}$$

29.3 Confidence Interval for the Slope

Statistic
$T = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{B_1 - \beta_1}{\sqrt{\frac{S_{yy} - b_1 S_{xy}}{(n-2)S_{xx}}}}$ with $v = n - 2$
Confidence Interval
A 100(1 - α)% confidence interval for the parameter β_1 in the regression line $\mu_{Y x} = \beta_0 + \beta_1 x$ is

$$b_1 - t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} \leq \beta_1 \leq b_1 + t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}$$

- $v = n - 2$ (degrees of freedom)

29.4 Hypothesis Testing for the Slope

Hypotheses
Null Hypothesis: $H_0: \beta_1 = \beta_{1_0}$ Alternative Hypothesis (two-sided): $H_1: \beta_1 \neq \beta_{1_0}$ Alternative Hypothesis (one-sided): $H_1: \beta_1 > \beta_{1_0}$ or $H_1: \beta_1 < \beta_{1_0}$
Test Statistic
$t = \frac{b_1 - \beta_{1_0}}{s \cdot \sqrt{S_{xx}}} \text{ with } v = n - 2$ <ul style="list-style-type: none"> • t (test statistic) • b_1 (estimated slope from the sample) • β_{1_0} (slope under the null hypothesis) • s (estimate of the standard error of the model) • S_{xx} (sum of squares of the deviations of the x-values from their mean)

29.5 Confidence Interval for the Intercept

Statistic
$T = \frac{B_0 - \beta_0}{S \cdot \sqrt{\frac{\sum_{i=1}^n x_i^2}{n S_{xx}}}} \text{ with } v = n - 2$
Confidence Interval
A $100(1 - \alpha)\%$ confidence interval for the parameter β_0 in the regression line $\mu_{Y x} = \beta_0 + \beta_1 x$ is

$$b_0 - t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^n x_i^2} \leq \beta_0 \leq b_0 + t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^n x_i^2}$$

- $v = n - 2$ (degrees of freedom)

29.6 Hypothesis Testing for the Intercept

Hypotheses
<p>Null Hypothesis: $H_0: \beta_0 = \beta_{0_0}$</p> <p>Alternative Hypothesis (two-sided): $H_1: \beta_0 \neq \beta_{0_0}$</p> <p>Alternative Hypothesis (one-sided): $H_1: \beta_0 > \beta_{0_0}$ or $H_1: \beta_0 < \beta_{0_0}$</p>
Test Statistic
$t = \frac{b_0 - \beta_{0_0}}{s \cdot \sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}} \text{ with } v = n - 2$ <ul style="list-style-type: none"> • t (test statistic) • b_0 (estimated intercept from the sample) • β_{0_0} (slope under the null hypothesis) • s (estimate of the standard error of the model) • S_{xx} (sum of squares of the deviations of the x-values from their mean)

29.7 Coefficient of Determination

The coefficient of determination measures the proportion of variability in the response variable that is explained by the regression model.

$$R^2 = 1 - \frac{SSE}{SST}$$

- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ (error sum of squares)
- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ (total corrected sum of squares)

- $R^2 = 1$ (perfect fit \rightarrow meaning that all points lie exactly on the regression line)
- $R^2 \approx 0$ (model fails to accurately model the data)