ST2334 Summary Notes

AY23/24 Sem 1, github.com/gerteck

1. Basic Probability Concepts

- **Sample Space:** S All possible outcomes of stat. expt.
- **Null Event**: Event that contains no element, empty set, $\varnothing$
- **Axioms of Probability**:
  For any event X, $0 \leq P(X) \leq 1$. $P(S) = 1$.
  If $A \cap B = \emptyset$ (Mut Excl), $P(A \cup B) = P(A) + P(B)$.
- Finite sample space with equally likely outcomes: $P(A)$ = ($\frac{\#samplepointsA}{\#totalsamplepointsS}$). (e.g. birthday problem)

Event Operation & Relationships

- **Event Operations:** Union, Intersection, Complement.
- **Event Relationships:** Contained: $A \subset B$
  Equivalence: $A \subset B$ with $A \supset B \rightarrow A = B$
  Mutually Exclusive: $A \cap B = \emptyset$.
- **De Morgan's Law:** $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$

Counting Methods

- Multiplication Principle: (Sequential Events)
- Addition Principle: (Pairwise Disjoin sets)
- **Permutation**: ${}_nP_r = \frac{n!}{(n-r)!}$
- **Combination**: $\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Conditional Probability

- Understand conditional as reduced sample space.

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

Independence

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

Law of Total Probability

- **Partition:** If $A_1, \cdots, A_n$ mutually exclusive, $\bigcup_{i=1}^{n} A_i = S$, then $A_1, \cdots, A_n$ are partitions.
- If $A_1, \cdots, A_n$ are partitions of S, then for any event B:

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes' Theorem

Let $A_1, \cdots, A_n$ be partitions of S. For any event B:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_k)P(A_i)}$$

For when $n = 2$, $\{A, A'\}$ becomes a partition of S.

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

2. Random Variables

A function X, which assigns a real number to every s $\in$ S is called a random variable.
- **Range space**: $Rx = \{x | x = X(s), s \in S\}$
- Likewise, the set X $\in$ A, for A being a subset of R, is also a subset of $S : s \in S : X(s) \in A$.

Probability Distribution

Two main types of RV used in practice: discrete and continuous.
- Probability assigned to each possible X
- Given RV X with range of R_x:
  Discrete: Numbers in R_x are finite or countable
  Continuous: R_x is interval

(Discrete) Probability Mass Function $f(x)$:

$$f(x) \begin{cases} P(X = x), & \text{for } x \in R_X \\ 0, & \text{for } x \notin R_X \end{cases}$$

1. $f(x_i) = P(X = x_i) \geq 0$ for $x_i \in R_x$
2. $f(x_i) = 0$ for $x_i \notin R_x$
3. $\sum_{i=1}^{\infty} f(x_i) = 1$ (PSum = 1)
4. $\forall B \subseteq \mathbb{R}, P(X \in B) = \sum_{x_i \in B \cap R_x} f(x_i)$

(Continuous) Probability Density Function $f(x)$:

- Given R_x is interval. Quantifies probability that X is in some range.
- $p.f.$ must satisfy:
  1. $f(x) \geq 0$, $f(x) = 0$ for $x \notin R_x$
  2. No need $f(x) \leq 1$ (Concerned with area)
  3. $\int_{R_x} f(x)dx = 1$ (Integration over R_X = 1)
  4. $\forall a, b$ s.t. $a \leq b, P(a \leq X \leq b) = \int_a^b f(x)dx$
- **Note:** $P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$
- Hence, to check if a function is a pdf,
  1. $f(x) \geq 0$ for $x \in R_x$, $f(x) = 0$ for $x \notin R_x$
  2. $\int_{R_x} f(x)dx = 1$.

Cumulative Distribution Function

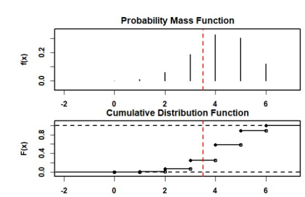
Describes distribution of a RV X: cumulative distribution function (cdf), applicable for discrete or continuous RV.

$$F(x) = P(X \leq x)$$

$F(x)$ is non-decreasing and $0 \leq F(x) \leq 1$
- Probability fn & cumulative distribution fn have one-to-one correspondence. For any probability fn given, the cdf is uniquely determined, vice versa.

CDF Discrete RV: Step Function $F(x)$

$$F(x) = \sum_{t \in R_x; t \leq x} f(t)$$

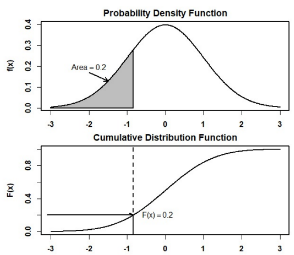


- $P(a \leq X \leq b) = P(X \leq b) - P(X < a)$
- $P(a \leq X \leq b) = F(b) - F(a-)$
- $P(a \leq X \leq b) = F(b) - \lim_{x \to a^-} F(x)$
- $0 \leq f(x) \leq 1$
- c.d.f has to be **right continuous** (• —)

CDF Continuous RV: $F(x)$

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

$$impt : f(x) = \frac{d(F(x))}{dx}$$



- $P(a \leq X \leq b) = P(a < X < b) = F(b) - F(a)$
- $0 \leq f(x)$.
  e.g. $f(x) = 3x^2$ is a valid $p.f.$ since $\int_{R_x} f(x)dx = 1$

Expectation μ & Variance σ

Expectation of Random Variable: μ

- **Mean of discrete RV**:

$$\mu = E(X) = \sum_{x \in R_x} x_i f(x_i)$$

- **E.g.**: X discrete RV with p.m.f. $f(x)$ and range R_X
  $\mu = E(g(x)) = \sum_{x \in R_x} g(x)f(x)$

- **Mean of continuous RV**:

$$\mu = E(X) = \int_{x \in R_x} xf(x)dx$$

- **E.g.**: X continuous RV with p.d.f. $f(x)$ and range R_X
  $\mu = E(g(x)) = \int_{x \in R_x} g(x)f(x)dx$
- **Properties of Expectation:**
- $E(aX + b) = aE(X) + b$
- Linearity of expectation: $E(X + Y) = E(X) + E(Y)$

Variance of Random Variable: σ

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

- **Variance of discrete RV**:

$$V(X) = \sum_{x \in R_x} (x - \mu_X)^2 f(x)$$

- **Variance of continuous RV**:

$$V(X) = \int_{x \in R_x} (x - \mu_X)^2 f(x)dx$$

- $V(X) \geq 0$ and $V(X) = 0$ when X is a constant
- $V(aX + b) = a^2 V(X)$
- **alt. form:** $V(X) = E(X^2) - (E(X))^2$
- **Standard Deviation:** $\sigma_X = \sqrt{V(X)}$

3. Joint Distributions

- Consider more than 1 RV simultaneously,
- Given sample space S. Let X and Y be functions mapping $s \in S \to \mathbb{R}$: (X, Y) is 2D random vector.

  Range spc: $R_{X,Y} = \{(x, y) | x = X(s), y = Y(s), s \in S\}$
- **Discrete 2D RV:**
  # of possible values of $(X(s), Y(s))$ finite / countable
- **Continuous 2D RV**:
  # of possible values of $(X(s), Y(s))$ assume any value in some region of the Euclidean space $\mathbb{R}^2$
- If both X and Y are discrete/continuous, then (X, Y) is discrete/continuous respectively.

Joint Probability Function

- **Joint Probability (mass) function, 2D discrete RV**:

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

  - $f_{X,Y}(x, y) \geq 0$ for any $(x, y) \in R_{X,Y}$
  - $f_{X,Y}(x, y) = 0$ for any $(x, y) \notin R_{X,Y}$
  - $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} P(X = x_i, Y = y_i) = 1$
  - Let $A \subseteq R_{X,Y}$.
    $P((X, Y) \in A) = \sum \sum_{(x,y) \in A} f_{X,Y}(x, y)$

- **Joint Probability (density) function, 2D cont. RV**:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)dydx$$

  - $f_{X,Y}(x, y) \geq 0$ for any $(x, y) \in R_{X,Y}$
  - $f_{X,Y}(x, y) = 0$ for any $(x, y) \notin R_{X,Y}$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$
    or equivalently:
  - $\int \int_{(x,y) \in R_{X,Y}} f_{X,Y}(x, y)dxdy = 1$

Marginal Probability Function

Marginal distribution of X is individual distribution of X, ignoring the value of Y. "Projection" of 2D function $f_{X,Y}(x, y)$ to 1D function.

Let (X, Y) be 2D RV with joint probability function $f_{X,Y}(x, y)$:

If Y is **discrete**, $f_X(x) = \sum_y f_{X,Y}(x, y)$

If Y is **continuous**, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$

- $f_Y(y)$ defined similarly
- $f_X(x)$ is a $p.f.$, satisfies all properties of prob. fn.

Conditional Distribution

Let (X, Y) be 2D RV with joint probability function $f_{X,Y}(x, y)$. Then $\forall x$ s.t. $f_X(x) > 0$: (X marg prob fn.)
Conditional probability function of Y given X = x:

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

- Intuition: Distribution of Y given $X = x$
- Only defined for x s.t. $f_X(x) > 0$
- $f_{Y|X}(y|x)$ is a $p.f.$ if we fix x, satisfies prop. of prob.fn.
- But, $f_{Y|X}(y|x)$ is not a $p.f.$ for x: No need for sum / integral over x = 1. Hence,
  If $f_X(x) > 0$: $f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x)$
  If $f_Y(y) > 0$: $f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y)$
- **Probability Y $\leq y$, Average Y given $X = x$**
- $P(Y \leq y | X = x) = \int_{-\infty}^{y} f_{Y|X}(y|x)dy$
- $E(Y | X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$

Independent Random Variables

$$X \perp Y : \forall x, y, f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

- Necessary condition: $R_{X,Y}$ must be a product space. i.e. $R_{X,Y} = \{(x, y)|x \in R_X; y \in R_y\} = R_X \times R_Y$ Else, dependent.

Properties of Independent RV

For X, Y independent RV:

- If $A, B \subseteq \mathbb{R}$, then events $X \in A$ and $Y \in B$ are independent:

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

$$P(X \leq x; Y \leq y) = P(X \leq x)P(Y \leq y)$$

- Then, $g_1(X)$ and $g_2(Y)$ are **independent**, for arbitrary g.
- **Conditional distribution** given Independence:

$$f_X(x) > 0 \rightarrow f_{Y|X}(y|x) = f_Y(y)$$

$$f_Y(y) > 0 \rightarrow f_{X|Y}(x|y) = f_X(x)$$

To check independence

1. $R_{X,Y}$ is a product space. i.e. R_X does not depend on Y, vice versa. (e.g. $0 < y < x$ is NOT a product space)
2. Additionally, $f_{X,Y}(x, y)$ **= some** $C * g_1(x)g_2(y)$ **where** g_1 **depends on** x **only and** g_2 **depends on** y **only.**

Marginal Distribution under Independence

- Since, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for independent RV, we derive marginal distribution by standardising $g_1(x)$ and $g_2(y)$.
- For discrete: $f_X(x) = \frac{g_1(x)}{\sum_{t \in R_X} g_1(t)}$
- For continuous: $f_X(x) = \frac{g_1(x)}{\int_{t \in R_X} g_1(t)dt}$

Expectation of a Random Vector

Given **2 variable function** $g(x, y)$**:**

- If (X, Y) is discrete:

$$E(g(X, Y)) = \sum_x \sum_y g(x, y) f_{X,Y}(x, y)$$

- If (X, Y) is continuous:

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

- If $X \perp Y$:

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

- (If $X \perp Y$, follows that $cov(X, Y) = 0$). However, converse not always true.

Covariance

- For random variables X, Y:

$$cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

- If (X, Y) both **discrete**:

$$cov(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y)$$

- If (X, Y) both **continuous**:

$$cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x, y) dx dy$$

- **Alt:** $cov(X, Y) = E(XY) - E(X)E(Y)$
- **Hence, for**$X \perp Y \rightarrow cov(X, Y) = 0$. (However, converse not always true).
- **Properties of covariance:**
- $cov(aX + b, cY + d) = (ac)cov(X, Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab * cov(X, Y)$
- $X \perp Y \rightarrow V(X \pm Y) = V(X) + V(Y)$

4.1 Special Probability Distributions

- **Discrete Distributions**: Study whole classes of discrete RVs that arise frequently in applications.

Discrete Uniform Distribution

- If X has values $x_1, x_2, \cdots, x_k$ with **equal probability**

$$f(x) \begin{cases} \frac{1}{k}, & \text{for } x = x_1, x_2, ..., x_k \\ 0, & \text{otherwise} \end{cases}$$

- **Expectation**:
$\mu_X = E(X) = \sum_{i=1}^{k} x_i f_X(x_i) = \frac{1}{k} \sum x_i$
- **Variance**:
$\sigma_X^2 = V(X) = E(X^2) - (E(X))^2 = \frac{1}{k} \sum x_i^2 - \mu_X^2$

Bernoulli, $Ber(p)$

- **Bernoulli Trial**: Random experiment has 2 possible outcomes (success and failure).
- **Bernoulli Random Variable**: X represents number of success in a single Bernoulli Trial. X has only two possible values: 1, or 0.
- **Probability mass function:** Let $0 \leq p \leq 1$ be the probability of success in Bernoulli trial

$$f_X(x) = P(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & otherwise \end{cases}$$

- $f_X(x) = p^x(1-p)^{1-x}$ for $x = 0$ or 1
- Bernoulli distr. is case of binomial distr. where $n = 1$.
- **Notation:** $X \sim Ber(p)$ and $q = 1 - p$

$$f_x(1) = p, f_x(0) = q$$

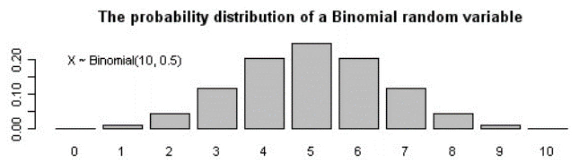
- **Expectation:** $\mu_X = E(X) = p$
- **Variance:** $\sigma_X^2 = V(X) = p(1-p)$
- **Bernoulli Process**: Sequence of repeatedly performed independent and identical Ber. trials.
- Generates sequence of **independent and identically distributed (i.i.d.)** Ber. RVs: $X_1, X_2, \cdots$

Binomial Distribution, $B(n, p)$

- **Binomial RV:** counts **number of successes** in n trials in a Ber. process.
- Given n independent trials with each trial having same probability p of success:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- **Notation**: $X \sim B(n, p)$
- $E(X) = np$, $V(X) = np(1-p)$



Negative Binomial Distribution, $NB(k, p)$ (k^{th} success)

- Let X = no. of independent identical distributed Bernoulli(p) trials until k^{th} **success** occurs.
- **Probability mass function of X**:

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

- **Notation**: $X \sim NB(k, p)$
- $E(X) = \frac{k}{p}$ and $V(X) = \frac{(1-p)k}{p^2}$

Geometric Distribution, $G(p)$ (till 1^{st} success)

- Let X = no. of i.i.d. Bernoulli(p) trials until 1st success occurs.

$$P(X = x) = p(1-p)^{x-1}$$

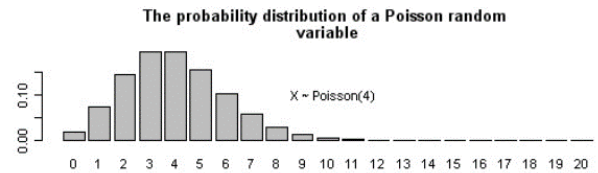
- **Notation**: $X \sim G(p)$
- $E(X) = \frac{1}{p}$ and $V(X) = \frac{1-p}{p^2}$

Poisson Distribution

- **Poisson RV**: Denotes number of events occuring in **fixed period of time or fixed region**, k = no. of occurences.

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

- **Notation**: $X \sim Poisson(\lambda)$ where $\lambda > 0$ is expected number of occurrences during given period/region
- $E(X) = \lambda$ and $V(X) = \lambda$



The number of infections X in a hospital each week has been shown to follow a Poisson distribution with a mean of 3 infections per week. What is the probability that

(a) there is *no* infection for a week?
(b) there are *less than* 4 infections for a week?

We are given than $X \sim \text{Poisson}(3)$. Then required probabilities are

(a) $P(X = 0) = e^{-3}$.
(b) $P(X \leq 3) = e^{-3}\left(1 + 3 + \frac{3^2}{2!} + \frac{3^3}{3!}\right)$.

Poisson Process

- Continuous time process, count number of occurrences within some interval of time. (given **rate** α)
- Properties of **Poisson process with rate parameter** α:
  - Expected no. of occurrences in interval length T: αT
  - No simultaneous occurrences, and no. of occurrences in disjoint intervals independent.
- **Number of occurrences in any interval** T of Poisson process follows $Poisson(\alpha T)$ distribution.
(**Apply** X $Poisson(\alpha T)$ **directly)**

Poisson Approximation of Binomial Distribution

- Let $X \sim B(n, p)$. Suppose $n \to \infty$ and $p \to 0$ s.t. $\lambda = np$ remains constant.
- Then, approximately, $X \sim Poisson(\lambda)$.

$$\lim_{p \to 0; n \to \infty} P(X = x) = \frac{e^{-np}(np)^x}{x!}$$

- Approximation is good when ($n \geq 20$ and $p \leq 0.05$), or ($n \geq 100$ and $np \leq 10$)

4.2 Special Probability Distributions

- **Continuous Distributions**: Many "natural" RVs whose set of possible values **uncountable**. Model with classes of continuous random variables.

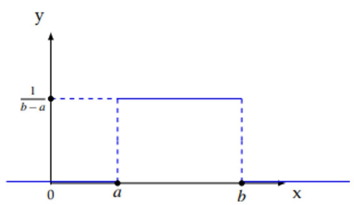
Continuous Uniform Distribution, $U(a,b)$

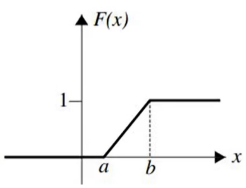
RV X follows uniform distribution over interval (a,b) if $p.d.f.$ given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

- **Notation**: $X \sim U(a,b)$
- $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$ (derive by integration).
- **Cumulative distr. func.** $c.d.f.$ is given by:

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$





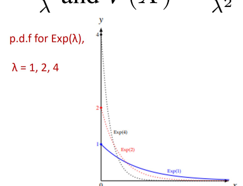
Exponential Distribution, $Exp(\lambda)$

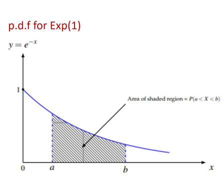
- Continuous counterpart to **geometric distribution**.

X follows exponential distribution, with parameter $\lambda > 0$ if $p.f.$ is given by:

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- **Notation**: $X \sim Exp(\lambda)$
- $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$





- We can **derive λ from mean / expectation of X**, since $E(X) = \frac{1}{\lambda}$.
- $c.d.f.$ is given by:

$$F_X(x) = P(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- Additionally, $P(X > x) = e^{-\lambda x}$, for $x > 0$.
- **Exponential distribution "Memoryless"**: Suppose X has exponential distribution with $\lambda > 0$. Then for any positive numbers s and t, we have:

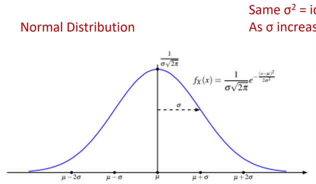
$$P(X > s + t | X > s) = P(X > t)$$

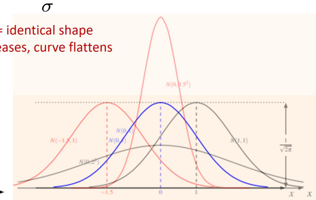
Normal Distribution, $N(\mu, \sigma^2)$

X said to follow normal distribution with mean μ and variance σ^2 if $p.f.$ given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

- **Notation**: $X \sim N(\mu, \sigma^2)$
- $E(X) = \mu$ and $V(X) = \sigma^2$
- $p.f.$ is **bell-shaped curve and symmetric** about $x = \mu$
- Total area under curve is 1
- 2 normal curves are identical in shape if they have same σ^2. They differ in location by $\mu_1 - \mu_2$.
- As σ increases, curve becomes more spread out
- If $X \sigma N(\mu, \sigma^2)$ and let $Z = \frac{X-\mu}{\sigma}$





Standardized Normal Distribution, $Z = N(0,1)$

If $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0,1)$:

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

- $E(Z) = 0$ and $V(Z) = 1$
- $p.f$ **of Z is given by**:

$$\phi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- **Standardizing normal distribution** allows us to use tables to find probabilities:
- For $X \sim N(\mu, \sigma^2)$, compute $P(x_1 < X < x_2)$ by standardization:

$$x_1 < X < x_2 \leftrightarrow \frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}$$

- Then, $P(z_1 < Z < z_2)$, **use $f_Z(z)$ table to calculate**.
- **Cumulative d.f. of standard Normal**:

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

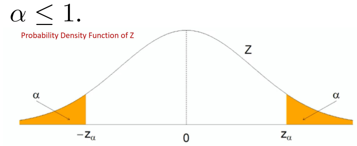
- $P(Z \geq 0) = P(Z \leq 0) = \phi(0) = 0.5$
- For any z, $\Phi(z) = P(Z \leq z) = P(Z \geq -z) = 1 - \phi(-z)$
- $-Z \sim N(0,1)$
- If $Z \sim N(0,1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$

Quantile

- **Upper Quantile:** x_α that satisfies:

$$P(X \geq x_\alpha) = \alpha$$

- where $0 \leq \alpha \leq 1$.



e.g. The 0.05th (upper) quantile of $Z \sim N(0,1)$ is 1.645, i.e. $z_{0.05} = 1.645$.

- $P(Z \geq z_\alpha) = P(Z \leq -z_\alpha) = \alpha$
- Upper z_α = Lower $z_{1-\alpha}$

Normal Approximation to Binomial Distribution

Let $X \sim B(n,p)$, then as $n \to \infty$:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- Approximation is good when $np > 5$ and $n(1-p) > 5$

5. Sampling, Sampling Distributions

Population and Sample

- **Statistical Inference**: Infer about population w. sample.
- **Population:** Totality of all possible obsv / outcomes.
- **Sample:** Subset of population
- Observation can be **numerical or categorical**
- Population can be **Finite or Infinite.**

Random Sampling

- Motivation: Often know what distribution population belongs to, but we not the parameters of distribution. Hence, use sample to estimate the parameters.

Single Random Sample

- **Simple Random Sample (SRS):** Sample of size n. Every subset of n observations (total $\binom{N}{n}$) equal chance of selection.

SRS for Infinite Population

- **For X be RV with certain** $p.f.\ f_X(x)$:
- Let $X_1, X_2, \cdots, X_n$ be n independent RV with same distribution as X. Then $X_1, \cdots, X_n$ is a **simple random sample** of size n.
- **Joint probability function of $X_1, \cdots, X_n$: (product)**

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n)$$

Sampling with Replacement (as Infinite)

- **Sampling with replacement** from finite population is considered as sampling from **infinite population**.
- Sample is random if:
  - Every element in population has same probability
  - Successive draws are independent

Sample Distribution of Sample Mean

- **Statistic:** Suppose random sample of n observations is $X_1, \cdots, X_n$. A **statistic** is a function of $X_1, \cdots, X_n$
- **Sample Mean**

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

- **Sample Variance**

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

- **Statistics are random variables**. If values in random sample observed, calculate **realization** of the statistic. Meaningful to consider distribution of statistics.

Sampling Distribution

Distribution of a statistic

- Mean and variance of $\bar{X}$:

$$E(\bar{X}) = \mu \text{ and } V(\bar{X}) = \frac{\sigma_X^2}{n}$$

  μ_X is unknown constant. $\bar{X}$ serves as valid estimator for μ_X. As n increases, accuracy of $\bar{X}$ increases.
- **Standard Error:** Standard deviation of sampling distribution (e.g. $\sigma_{\bar{X}}$), describes how much $\bar{X}$ tends to vary from sample to sample of size n.
- **Law of Large Numbers:** As n increases, $\bar{X}$ converges to μ_X. i.e. For any $\epsilon \in \mathbb{R}$:

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

  As n increases, probability that sample mean differs from population mean goes to zero.

Central Limit Theorem

$\bar{X}$, **mean of random sample of size** n from population with mean μ and variance σ^2, then as $n \to \infty$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \text{ approximately}$$

- For large n, $\bar{X}$ is approximately normally distributed.
- If random sample is from normal population, $\bar{X}$ is normally distributed no matter value of n
- If very skewed, CLT may not hold even with large n.

Other Sampling Distributions

$\chi^2(n)$ (Chi) Distribution

- Let $Z_1, \cdots, Z_n$ be n independent and identically distributed standard normal RVs.
- A χ^2 RV with n **degrees of freedom** is defined as a RV with same distribution as $Z_1^2 + \cdots + Z_n^2$
- **Notation:** $\chi^2(n)$ with n degrees of freedom
- If $Y \sim \chi^2(n)$, then $E(Y) = n$ and $V(Y) = 2n$
- **For large** n, $\chi^2(n)$ is approximately $N(n, 2n)$
- If Y_1 and Y_2 are independent χ^2 RVs with m **and** n **degrees of freedom respectively,** then $Y_1 + Y_2$ is $\chi^2(m+n)$
- χ^2 distribution is a family of curves. All density functions have long right tail.

Sampling Distribution of S^2

- $E(S^2) = \sigma^2$

Sampling Distribution of $\frac{(n-1)S^2}{\sigma^2}$

If S^2 is variance of random sample of size n from normal population of variance σ^2, then:

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

has $\chi^2(n-1)$ distribution

Suppose 6 random samples are drawn from a normal population $N(\mu, 4)$. Define the sample variance

$$S^2 = \frac{1}{5}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Find c such that $P(S^2 > c) = 0.05$.

Solution:
We know that $\frac{5S^2}{4} \sim \chi^2(5)$. Hence,

$$\begin{aligned} & P(S^2 > c) = 0.05 \\ \Leftrightarrow\ & P(5S^2/4 > 5c/4) = 0.05 \\ \Leftrightarrow\ & 5c/4 = \chi^2(5; 0.05) = 11.07 \\ \Leftrightarrow\ & c = 8.86. \end{aligned}$$

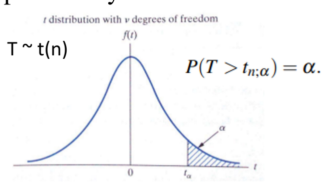
t-Distribution $t(n)$

Suppose $Z \sim N(0,1)$, $U \sim \chi^2(n)$. If Z, U independent:

$$T = \frac{Z}{\sqrt{U/n}} \sim t(n)$$

where $t(n)$ is t-distribution with n degrees of freedom

- **t-Distribution approaches $N(0,1)$ as $n \to \infty$.** When $n \geq 30$, t-dist approx normal, replace by $N(0,1)$.
- **Expectation, Variance:** If $T \sim t(n)$, then $E(T) = 0$ and $V(T) = \frac{n}{n-2}$ for $n > 2$
- Symmetric about vertical axis and resembles standard normal distribution
- **Critical value for t-distribution** $t_{n;\alpha}$: number with right hand tail probability of α.



- If $X_1, \cdots, X_n$ are independent and identically distributed normal RVs with mean μ and variance σ^2, then:

$$t.value = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

i.e. follows **t distribution** with **n-1 degrees of freedom**.

L-EXAMPLE 5.12
A manufacturer of light bulbs claims that his light bulbs will burn on the average $\mu = 500$ hours. To maintain this average, he tests 25 bulbs each month.

If the computed t value, $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, falls between $-t_{24;0.05}$ and $t_{24;0.05}$, he is satisfied with his claim.

What conclusion should be drawn from a sample that has a mean ≡518 hours and a standard deviation $s = 40$ hours? Assume that the distribution of burning times in hours is approximately normal.

Solution:
From the t-table or software, $t_{24;0.05} = 1.711$.

Therefore, the manufacturer is satisfied with his claim if a sample of 25 bulbs yields a t-value between -1.711 and 1.711.

If $\mu = 500$, then

$$t = \frac{518 - 500}{40/5} = 2.25 > 1.711.$$

Note that if $\mu > 500$, then the value of t computed from the sample would be more reasonable. Hence the manufacturer is likely to conclude that his bulbs are a better product than he thought.

F-Distribution $F(m,n)$

Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ independent:

$$F = \frac{U/m}{V/m} \sim F(m,n)$$

i.e. **F-distribution with (m,n) degrees of freedom**

- If $X \sim F(m,n)$, then **mean**:

$$E(X) = \frac{n}{n-2} \text{ for } n > 2$$

and **variance**:

$$V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ for } n > 4$$

- Values of the F-distribution can be found in the statistical tables or software. The values of interests are $F(m,n;\alpha)$ such that

$$P(F > F(m,n;\alpha)) = \alpha,$$

where $F \sim F(m,n)$.
- It can be shown that

$$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

- If $F \sim F(m,n)$, then $1/F \sim F(n,m)$

L-EXAMPLE 5.15
Let S_1^2 and S_2^2 be the sample variances of independent random samples of sizes $n_1 = 25$ and $n_2 = 31$, taken from normal populations with variances $\sigma_1^2 = 10$ and $\sigma_2^2 = 15$ respectively. Find $P(S_1^2/S_2^2 > 1.26)$.

Solution:
Note that

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

which gives

$$\frac{S_1^2/10}{S_2^2/15} \sim F(24,30).$$

Thus

$$P\left(\frac{S_1^2}{S_2^2} > 1.26\right) = P\left(\frac{S_1^2/10}{S_2^2/15} > 1.26 \times \frac{15}{10}\right)$$
$$= P(F > 1.89) = 0.05.$$

Note that here $F \sim F(24,30)$.

06. Estimation

Point Estimation for Mean

- Single number to estimate population parameter
- **Point Estimator**Formula that describes this calculation
- **Point Estimate**Result of point estimator
- Notation: θ represents parameter of interest. θ can be p, μ, σ, etc.

Unbiased Estimator

Let $\hat{\theta}$ be an estimator of θ. Then $\hat{\theta}$ is unbiased if:

$$E(\hat{\theta}) = \theta$$

Maximum Error of Estimate

- Motivation: Usually $\bar{X} \neq \mu$. So $\bar{X} - \mu$ measures difference between estimator and parameter
- Let z_α be αth upper quantile of standard normal distribution Z. i.e. $P(Z > z_\alpha) = \alpha$

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = P(|\bar{X} - \mu| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- **Maximum Error of Estimate**Given probability $1 - \alpha$:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Determination of Sample Size

Given probability $1 - \alpha$ and maximum error E, what is the minimum sample size n?

$$n \geq (\frac{z_{\alpha/2}\sigma}{E})^2$$

Different Cases

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
II	any	known	large	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2} \cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot s}{E_0}\right)^2$

Confidence Interval for Mean

- **Interval Estimator**Rule for calculating an interval (a, b) in which we are fairly certain the parameter lies
- **Confidence Level**Probability that interval contains parameter. i.e. $1 - \alpha$

$$P(a < \mu < b) = 1 - \alpha$$

- **Confidence Interval**Interval calculated by interval estimator. i.e. (a, b)

Case 1: σ known, data normal

Previously:

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

By rearranging, the $1 - \alpha$ confidence interval is:

$$(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

Other Cases

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

- n is considered large when $n \geq 30$

Comparing 2 Populations

- Goal: Make inference on $\mu_1 - \mu_2$

Experimental Design

- **Independent Samples**Completely randomized
- **Matched Pairs Samples**Randomization between matches pairs

Independent Samples: Known and Unequal Variance

Assumptions:

1. Given: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2
2. 2 samples are independent
3. Population variances are known and $\sigma_1^2 \neq \sigma_2^2$
4. Both populations are normal OR $n_1 \geq 30$ and $n_2 \geq 30$

Let $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ be random samples:

$$E(\bar{X}) = \mu_1 \,,\, V(\bar{X}) = \frac{\sigma_1^2}{n_1} \,,\, E(\bar{Y}) = \mu_2 \,,\, V(\bar{Y}) = \frac{\sigma_2^2}{n_2}$$

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 \,,\, V(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Thus, by normalizing RV $\bar{X} - \bar{Y}$ and using assumption 4:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Thus, the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Independent Samples: Unknown and Unequal Variance

Assumptions:

1. Given: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2
2. 2 samples are independent
3. Population variances are unknown and $\sigma_1^2 \neq \sigma_2^2$
4. $n_1 \geq 30$ and $n_2 \geq 30$

Since σ_1 and σ_2 are unknown, we use the standard error instead:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \,,\, S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$$

Thus, by normalizing RV $\bar{X} - \bar{Y}$ and using assumption 4:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1)$$

Thus, the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Independent Samples: Small n, Unknown and Equal Variance

Assumptions:

1. Given: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2
2. 2 samples are independent
3. Population variances are unknown and $\sigma_1^2 = \sigma_2^2$
4. $n_1 < 30$ and $n_2 < 30$
5. Both populations are normally distributed

Thus, by normalizing RV $\bar{X} - \bar{Y}$ and using assumptions 3 and 4:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where S_p is the pooled sample variance, which estimates σ^2

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

Thus, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Independent Samples: Large n, Unknown and Equal Variance

Assumptions:

1. Given: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2
2. 2 samples are independent
3. Population variances are unknown and $\sigma_1^2 = \sigma_2^2$
4. $n_1 \geq 30$ and $n_2 \geq 30$

By applying CLT on assumption 4, we can replace $t_{n_1+n_2-2;\alpha/2}$ with $z_{\alpha/2}$. Thus, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Paired Data

Assumptions:

1. Given: $(X_1, Y_1), \cdots, (X_n, Y_n)$ are matched pairs, where $X_1, \cdots, X_n$ is random sample from population 1 and $Y_1, \cdots, Y_n$ is random sample from population 2
2. X_i and Y_i are dependent
3. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$

Define $D_i = X_i - Y_i$, $\mu_D = \mu_1 - \mu_2$. We can treat $D_1, \cdots, D_n$ as random sample from single population with μ_D and σ_D^2. Consider the statistic:

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}, \text{ where } \bar{D} = \frac{\sum_{i=1}^n D_i}{n} \text{ and } S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}$$

If $n < 30$ and population is normally distributed:

$$T \sim t_{n-1}$$

Thus, if $n < 30$ and the population is normally distributed, the $100(1-\alpha)\%$ confidence interval for μ_D is:

$$\bar{d} \pm t_{n-1;\alpha/2} \frac{S_D}{\sqrt{n}}$$

If $n \geq 30$:

$$T \sim N(0,1)$$

Thus, if $n \geq 30$, the $100(1-\alpha)\%$ confidence interval for μ_D is:

$$\bar{d} \pm z_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

07. Hypothesis Testing

Steps for Hypothesis Testing

Step 1: Null Hypothesis and Alternative Hypothesis

- **Null Hypothesis**H_0 Statement that parameter takes some value
- **Alternative Hypothesis**H_1 Statement that parameter falls in alt. range
- **2-Sided Test**If H_1 is "Parameter is $\neq$ to value under H_0"
- **Right-Sided Test**If H_1 is "Parameter is $>$ to value under H_0"
- **Left-Sided Test**If H_1 is "Parameter is $<$ to value under H_0"

Step 2: Level of Significance

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

- **Level of Significance**α Probability of rejecting H_0 when it is true. i.e.

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

- **Power of the Test**$1 - \beta = P(\text{Reject } H_0 | H_0 \text{ is false})$ where

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

Step 3: Test Statistic, Distribution, and Rejection Region

- **Test Statistic**Statistic used to see how far away from H_0 the data is

Step 4: Conclusion

Given test statistic, determine if it is in the rejection region:

- If it is, reject H_0 and fail to reject H_1
- Otherwise, fail to reject H_0

Hypotheses for Mean

Case 1: Known Variance

Assumptions:

1. Population variance is known
2. Underlying distribution is normal OR $n \geq 30$

Steps:

1. Set null and alternative hypotheses. e.g.

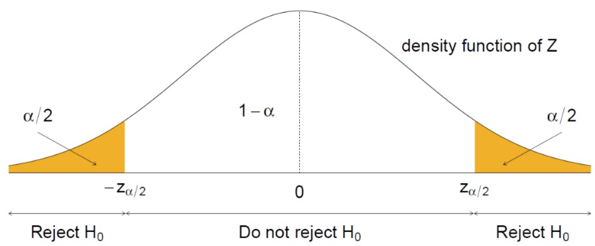
$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

2. Set level of significance
3. With σ^2 known and population normal (or $n \geq 30$), the test statistic is:

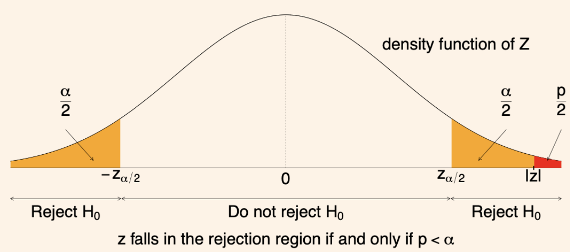
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Rejection region, where we let observed value of Z be z:

- $H_1 : \mu \neq \mu_0$: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
- $H_1 : \mu < \mu_0$: $z < -z_\alpha$
- $H_1 : \mu > \mu_0$: $z > z_\alpha$



- **p-Value**Conditional probability that test statistic is as extreme as observed value, given H_0 is true
- $H_1 : \mu \neq \mu_0$: $p = 2P(Z < -|z|)$
- $H_1 : \mu < \mu_0$: $p = P(Z < -|z|)$
- $H_1 : \mu > \mu_0$: $p = P(Z > |z|)$



4. - Rejection region: If z is inside rejection region, reject H_0. Otherwise do not reject.
   - p-Value: If p is less than α, reject H_0. Otherwise do not reject.

Case 2: Unknown Variance

Assumptions:

1. Population variance is unknown
2. Underlying distribution is normal

- Test statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- Rejection region:
  - $H_1 : \mu \neq \mu_0$: $t < -t_{n-1;\alpha/2}$ or $t > t_{n-1;\alpha/2}$
  - $H_1 : \mu < \mu_0$: $t < -t_{n-1;\alpha}$
  - $H_1 : \mu > \mu_0$: $t > t_{n-1;\alpha}$
- When $n \geq 30$, we can replace t_{n-1} by Z

Comparing Means: Independent Samples

- Motivation: Given 2 independent samples from 2 populations, interested in testing $H_0 : \mu_1 - \mu_2 = \delta_0$

Rejection Regions and p-Values

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \|z\|)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\|z\|)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \|z\|)$

Case 1: Known Variance

Assumptions:

1. Population variances are known
2. Underlying distributions are normal OR $n_1 \geq 30$ and $n_2 \geq 30$

- Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Case 2: Unknown Variance

Assumptions:

1. Population variances are unknown
2. $n_1 \geq 30$ and $n_2 \geq 30$

- Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Case 3: Unknown, Equal Variance

Assumptions:

1. Population variances are unknown but equal
2. Underlying distributions are normal
3. $n_1 < 30$ and $n_2 < 30$

- Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Comparing Means: Paired Data

- Intuition: Get difference, then use methods from single samples
- Define $D_i = X_i - Y_i$. For $H_0 : \mu_D = \mu_{D_0}$, test statistic:

$$T = \frac{\bar{D} - \mu_{D_0}}{S_D/\sqrt{n}}$$

  - If $n < 30$ and population is normally distributed, $T \sim t_{n-1}$
  - If $n \geq 30$, $T \sim N(0,1)$

08. Miscellaneous

Integration by Parts

$$\int udv = uv - \int vdu$$

- How to choose u? LIPET