ST2334

AY22/23 Sem 1

github.com/jasonqiu212

01. Basic Concepts of Probability

Event Operations

- **Mutually Exclusive** - $A \cap B = \emptyset$
- **Contained** - $A \subset B$
- **Equivalence** - $A \subset B$ and $A \supset B \rightarrow A = B$
- **Distributive** - $A \cap (B \cup C) = (A \cup B) \cup (A \cup C)$
- **DeMorgan's** - $(A \cup B)' = A' \cap B'$
- $A = (A \cap B) \cup (A \cap B')$

Counting Methods

- **Multiplcation Principle** - Given r experiments performed **sequentially** and each has $n_1, n_2, \cdots, n_r$ outcomes. After r experiments, there are $n_1 n_2 \cdots n_r$ outcomes.
- **Addition Principle** - Given experiment can be done in k different ways and each has $n_1, n_2, \cdots, n_r$ ways. There are $n_1 + n_2 + \cdots + n_k$ total ways.
- **Permutation** - ${}_nP_r = \frac{n!}{(n-r)!}$
- **Combination** - $\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Probability

Axioms of Probability

1. For any event A, $0 \leq P(A) \leq 1$
2. P(S) = 1
3. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

- $P(A') = 1 - P(A)$
- $P(A) = P(A \cap B) + P(A \cap B')$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $A \subset B$, then $P(A) < P(B)$

Finite Sample Space with Equally Likely Outcomes

Given sample space $S = \{a_1, \cdots, a_k\}$ and all outcomes are **equally likely**, i.e. $P(a_1) = \cdots = P(a_k)$:

$$\text{For any event } A \subset S, P(A) = \frac{\text{No. of sample points in A}}{\text{No. of sample points in S}}$$

Conditional Probability

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

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 $A_1, \cdots, A_n$ are mutually exclusive events and $\bigcup_{i=1}^n A_i = S$, 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:

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

02. Random Variables

- Motivation: Assign value to outcome of experiment
- **Random Variable** - Let S be sample space. Function X which maps $\mathbb{R}$ to every $s \in S$

Probability Distribution

- 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 Distribution

- **Probability Function** - Given $R_x = \{x_1, \cdots\}$. For each x_i, there's some probability that $X = x_i$:

$$f(x) = P(X = x)$$

- $p.f.$ must satisfy:
  1. $f(x_i) = P(X = x_i)$ for $x_i \in R_x$
  2. $f(x_i) = 0$ for $x_i \notin R_x$
  3. $\sum_{i=1}^\infty f(x_i) = 1$
  4. $\forall B \subseteq \mathbb{R}, P(X \in B) = \sum_{x_i \in B \cap R_x} f(x_i)$
- **Probability Distribution** - Collection of pairs $(x_i, f(x_i))$

Continuous Probability Distribution

- **Probability Function** - Given R_x is interval. Quantifies probability that X is in some range.
- $p.f.$ must satisfy:
  1. $f(x) \geq 0$
  2. $f(x) = 0$ for $x \notin R_x$
  3. $\int_{R_x} f(x)dx = 1$
  4. $\forall a, b$ s.t. $a \leq b, P(a \leq X \leq b) = \int_a^b f(x)dx$
- Note: $P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$

Cumulative Distributive Function

Given RV X, which can be discrete or continuous:

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

- $F(x)$ is non-decreasing and $0 \leq F(x) \leq 1$
- **For discrete RV**: Step function

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

  - $P(a \leq X \leq b) = F(b) - \lim_{x \to a^-} F(x)$
  - $0 \leq f(x) \leq 1$
- **For continuous RV**:

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

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

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

Expectation of Random Variable

- **Mean of discrete RV**:

$$\mu = E(X) = \sum_{x \in R_x} x_i f(x_i) = \sum_{i=1}^\infty P(X \geq i)$$

  - Let g be some function. $E(g(x)) = \sum_{x \in R_x} g(x)f(x)$
- **Mean of continuous RV**:

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

  - Let g be some function. $E(g(x)) = \int_{x \in R_x} g(x)f(x)dx$
- $E(aX + b) = aE(X) + b$
- Linearity of expectation: $E(X + Y) = E(X) + E(Y)$

Variance of Random Variable

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

- **Variance of discrete RV**:

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

- **Variance of continuous RV**:

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

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

03. Joint Distributions

- Motivation: What if interested in more than 1 RV simultaneously?
- Given sample space S. Let X and Y be functions mapping $s \in S \rightarrow \mathbb{R}$:

$$(X, Y) \text{is 2D random vector}$$

$$\text{Range space: } R_{X,Y} = \{(x, y) | x = X(s), y = Y(s), s \in S\}$$

- **Discrete 2D RV** - If no. of possible values of $(X(s), Y(s))$ are finite or countable
- **Continuous 2D RV** - If no. of possible values of $(X(s), Y(s))$ can be any value in Euclidean space $\mathbb{R}^2$
- If both X and Y are discrete/continuous, then (X, Y) is discrete/continuous.

Joint Probability Function

- **For discrete**:

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

  - $f_{X,Y}(x,y) \geq 0$ 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 $A \subseteq R_{X,Y}$. $P((X,Y) \in A) = \sum\sum_{(x,y)\in A} f_{X,Y}(x,y)$
- **For continuous**:

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

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

Marginal Probability Function

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

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

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

- $f_Y(y)$ defined similarly
- Intuition: Marginal distribution for X ignores presence of Y
- $f_X(x)$ is a $p.f.$

Conditional Distribution

Let (X,Y) be a 2D RV with joint probability function $f_{X,Y}(x,y)$. Then $\forall x$ s.t. $f_X(x) > 0$:

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

- 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
- But, $f_{Y|X}(y|x)$ is not a $p.f.$ for 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

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

- Necessary condition: $R_{X,Y}$ must be a product space. Else, dependent.

Properties

Suppose X, Y are independent RV:

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

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

- $g_1(X)$ and $g_2(Y)$ are independent
- Independence is related with conditional distribution:

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

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

Quick way to check independence

1. $R_{X,Y}$ is a product space. i.e. R_X does not depend on Y and vice versa.
2. $f_{X,Y}(x,y)$ can be written as $cg_1(x)g_2(y)$ where g_1 depends on x only and g_2 depends on y only.
3. For discrete: $f_X(x) = \frac{g_1(x)}{\sum_{t\in R_X} g_1(t)}$
4. For continuous: $f_X(x) = \frac{g_1(x)}{\int_{t\in R_X} g_1(t)dt}$

Expectation

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

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

$$\text{If } (X,Y) \text{ is continuous, } E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dydx$$

- $E(XY) = E(X)E(Y)$ if $X \perp Y$

Covariance

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

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

$$\text{If } (X,Y) \text{ is cont., } cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)dxdy$$

- $cov(X,Y) = E(XY) - E(X)E(Y)$
- $X \perp Y \rightarrow cov(X,Y) = 0$. But converse is not always true.
- $cov(aX + b, cY + d) = (ac)cov(X,Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abcov(X,Y)$
- $X \perp Y \rightarrow V(X + Y) = V(X) + V(Y)$

04. Special Probability Distributions

Discrete Uniform Distribution

- If X has values $x_1, x_2, \cdots, x_k$ with **equal probability**
- $p.f.$: $f_X(x) = \frac{1}{k}$ where $x = x_1, \cdots, x_k$ and 0 otherwise
- Expectation: $\mu_X = E(X) = \sum_{i=1}^{k} x_i f_X(x_i) = \frac{1}{k}\sum_{i=1}^{k} x_i$
- Variance: $\sigma_X^2 = V(X) = E(X^2) - (E(X))^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - \mu_X^2$

Bernoulli

- **Bernoulli Trial** - Random experiment with 2 possible outcomes (success and failure)

Bernoulli Random Variable

- Number of successes in Bernoulli trial (Either 1 or 0)
- 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
- Notation: $X \sim Ber(p)$ and $q = 1 - p$
- $\mu_X = E(X) = p$ and $\sigma_X^2 = V(X) = p(1 - p)$

Bernoulli Process

- Sequence of repeatedly performed **independent and identical** Ber. trials
- Generates sequence of **independet and identically distributed (i.i.d.)** Ber. RVs: $X_1, X_2, \cdots$

Binomial Distribution

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

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

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

Negative Binomial Distribution

- Let X = Number of i.i.d. Bernoulli(p) trials until kth success occurs

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

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

Geometric Distribution

- Let X = Number 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 happening in **fixed period of time or fixed region**

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

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

Poisson Process

- Continuous time process, where we count number of occurrences within some interval of time
- Given Poisson process with rate parameter α:
  - Expected number of occurrences in interval of length T is αT
  - No simultaneous occurrences
  - Number of occurrences in disjoint intervals are independent
- Number of occurrences in any interval T of Poisson process follows $Poisson(\alpha T)$ distribution

Poisson Approximation of Binomial Distribution

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

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

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

Continuous Uniform Distribution

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

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

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

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

Exponential Distribution

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

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

- Notation: $X \sim Exp(\lambda)$
- $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$
- $c.d.f.$ is given by:

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

- Suppose X has exponential distribution with parameter $\lambda > 0$. Then for any positive numbers s and t, we have:

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

Normal Distribution

X follows normal distribution with mean μ and variance σ^2 if $p.f.$ is 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

If $X \sim N(\mu, \sigma^2)$ and let

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

then $Z \sim N(0, 1)$

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

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

- Importance of standardizing normal distribution is that it allows us to use tables to find probabilities
- Let $X \sim N(\mu, \sigma^2)$. We can compute $P(x_1 < X < x_2)$ by standadization:

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

- $c.d.f.$ is given by:

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

- $P(Z \ge 0) = P(Z \le 0) = \phi(0) = 0.5$
- For any z, $\phi(z) = P(Z \le z) = P(Z \ge -z) = 1 - \phi(-z)$
- $-Z \sim N(0, 1)$
- If $Z \sim N(0, 1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$
- **Upper Quantile** - Given α is upper-tail percentage. The αth upper quantile is x_α that satisfies:

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

  e.g. The 0.05th (upper) quantile of $Z \sim N(0, 1)$ is 1.645, i.e. $z_{0.05} = 1.645$.
- $P(Z \ge z_\alpha) = P(Z \le -z_\alpha) = \alpha$
- 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 $n(1 - p) > 5$

05. Sampling Distributions

Population and Sample

- Motivation: Infer something about population using sample
- **Population** - All possible observations of survey
- **Sample** - Subset of population
- Every observation can be numerical or categorical
- Finite population vs. Infinite population

Random Sampling

- Motivation: We usually know what distribution the population belongs to, but we don't know the parameters of the distribution. We can use sample to estimate the parameters.
- **Simple Random Sample** - Given sample of size n, each sample has equal chance of being selected

SRS for Infinite Population

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

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

Sampling with Replacement

- 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 $X_1, X_2, \cdots, X_n$. 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**. We can look at distribution of statistics.
- **Sample Distribution** - Distribution of a statistic
- Mean and variance of $\bar{X}$:

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

  Intuition: μ_X is some unknown constant. $\bar{X}$ guesses that. As n increases, accuracy of $\bar{X}$ increases.
- **Standard Error** - Standard deviation of sampling distribution (e.g. $\sigma_{\bar{X}}$)
- **Law of Large Numbers** - As n increases, $\bar{X}$ converges to μ_X. i.e. For any $\epsilon \in \mathbb{R}$:

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

Central Limit Theorem

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

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

- Intuition: For a 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

χ^2 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 is 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

t-Distribution

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

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

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

- t-Distribution approaches $N(0,1)$ as $n \to \infty$
- When $n \geq 30$, t-distribution is approximately normal
- 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
- If $X_1, \cdots, X_n$ are independent and identically distributed normal RVs with mean μ and variance σ^2, then:

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

F-Distribution

Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$. If U_1 and U_2 are independent, then:

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

is called F-distribution with (m,n) degrees of freedom

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

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

and

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

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

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:

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

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

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:

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

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.

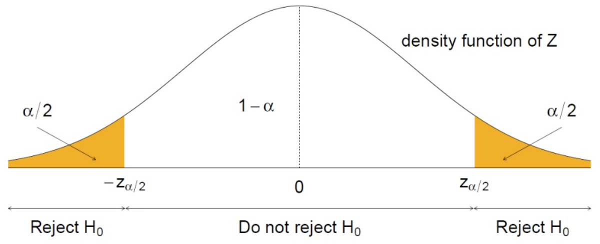
$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

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

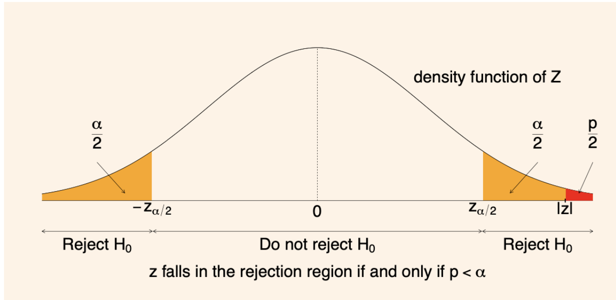
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

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
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \lvert z \rvert)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\lvert z \rvert)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

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

$$\int u dv = uv - \int v du$$

- How to choose u? LIPET