STAT251 Elementary Statistics

Kelvin Hsu

April 20, 2019

1 Introduction to Statistic

Statistic is a science involving the design of studies, data collection , summarizing data and analyzing data, interpreting results and drawing conclusions.

Key Statistics Concepts

- Population: all subject of interest in a study
- Sample: a subset of population
- Parameter: a descriptive measure of a population
- Statistic: a descriptive measure of a sample
- Census: collecting data for the entire population
- Sample Survey: collecting data for a sample

Classification of Variables

- Categorical: gender, major, blood type...
- Quantitative: can be discrete or continuous

Inferential VS Descriptive Statistic

- Inferential Statistic: Methods for making decision or prediction about a population based on data collected on a sample.
- Descriptive Statistic: Methods of summarizing data (graphs and numbers)

Types of Graphs

For qualitative data,

- Pie Chart
- Bar Graph

For quantitative data,

- Dot Plot
- Stem and Leaf Plot
- Histogram

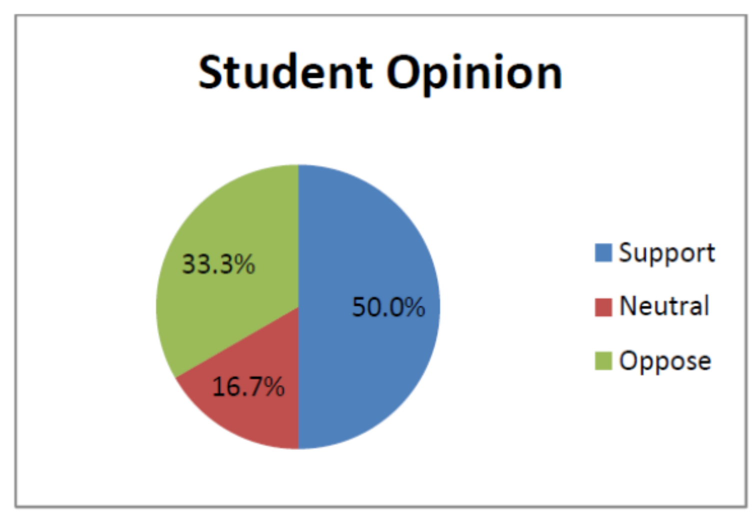

Figure 1: Pie Chart

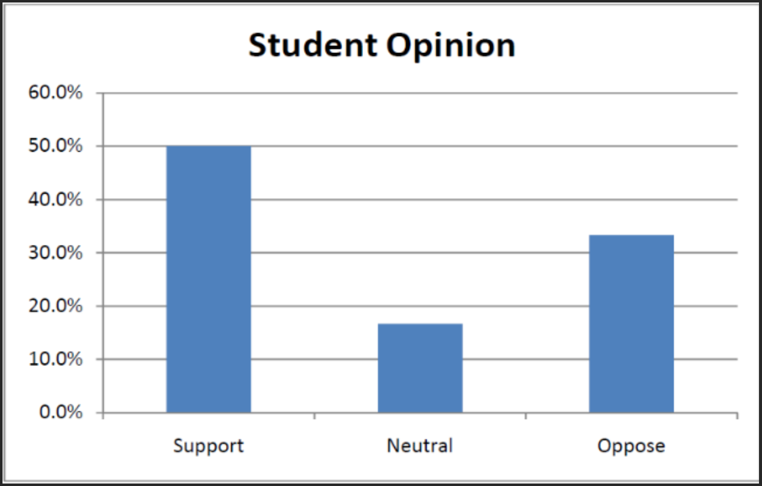

Figure 2: Bar Graph

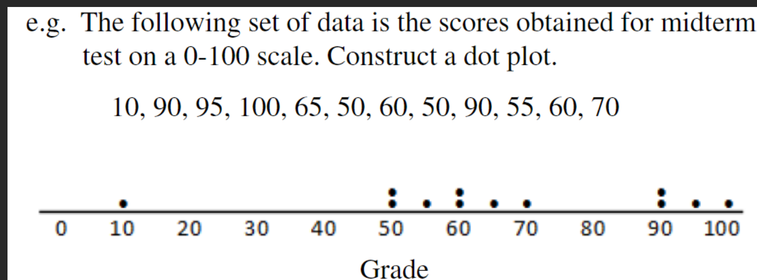

Figure 3: Dot Plot

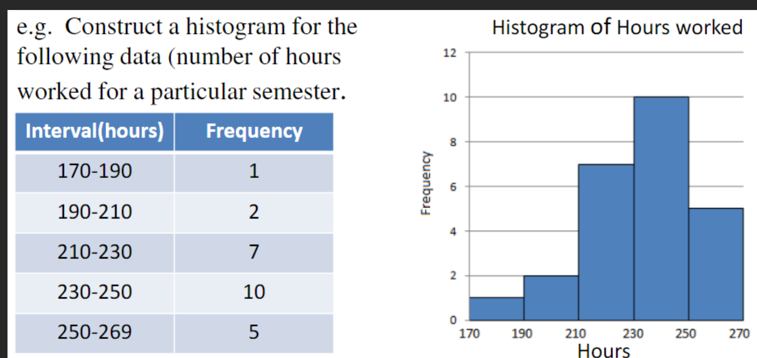

Figure 4: Histogram

Describing A Distribution

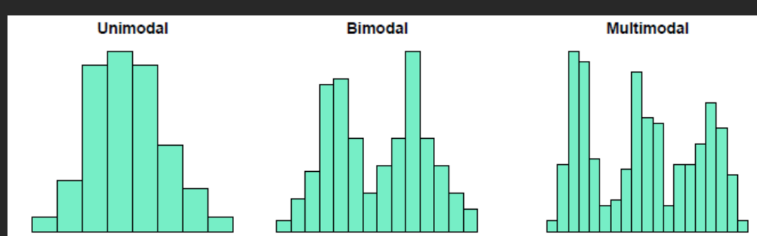

Figure 5: Types Of Distribution

Shape

- Skewness: whether the distribution is symmetrical
- Center: where do the observation cluster about?

- Spread: Assess the spread of a distribution
- Outlier: observations that fall far from the rest of the data
- Mode: the most frequent observation

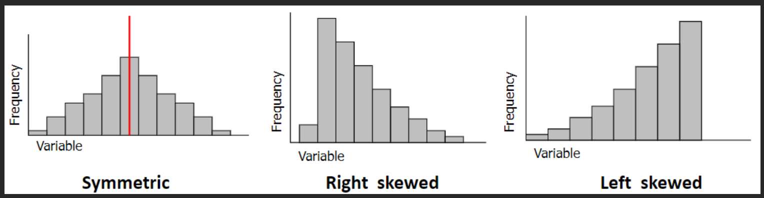

Figure 6: Skewness

Measures of Center

Mean

Mean is the average of the observation.

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

where n is the number of observations.

Median

Median is the midpoint of the observations. If the number of sample n is...

odd:

median is the midpoint $\frac{n+1}{2}$th point in an ordered list

even:

median is the average of the 2 middle observations $(\frac{n}{2}th + \frac{n+1}{2}th)/2$

Comparing Mean and Median

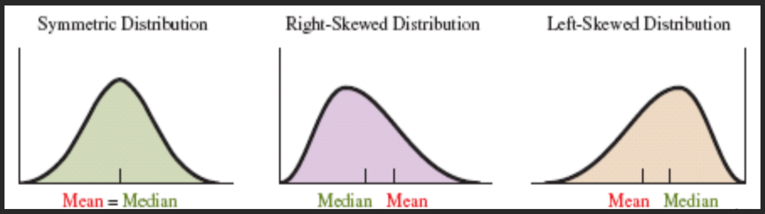

Figure 7: Mean And Median

Measure of Variability

Measure of variation gives information on the spread or variability or dispersion of the data.

Range

$$\text{Range} = X_{max} - X_{min}$$

Variance and Standard Deviation

$$\text{Variance} = S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
$$\text{Standard Deviation} = S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard deviation measures the spread of data and it is zero only when all observations have the same value. As spread increases, s increases. S is sensitive to outliers.

Interquartile Range

$$\text{IQR} = \text{Q3 - Q1}$$

- Q1: first quartile, 25th percentile, the value in the sample that has 25% of data below it
- Q2: median
- Q3: third quartile, 75th percentile

Identifying Outliers

If data is 1.5 IQR below Q1 or above Q3, then it is an outlier.

Percentile

The n-th percentile is a value such that n percent of the observations fall below that value.

Sample Quantile

Let $0 < p < 1$. The sample quantile or order p, Q(p) is a number with the property that approximately $p * 100\%$ of the data are below it.

1. Sort data from smallest to largest
2. let m $= n * p + 0.5$

if m is an integer:

$$Q(p) = x_m$$

else:

$$Q(p) = \frac{x_m + x_{m+1}}{2}$$

Box Plot

Box plot provides a way of looking at data to determine its central tendency, spread, skewness, and the existence of outliers. To construct a box plot,

- largest value within Q3 + IQR
- smallest value within Q1 - IQR
- Q1, Q2, and Q3

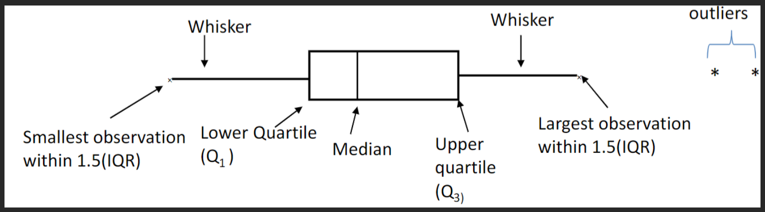

Figure 8: Box Plot

2 Probability

Sample Space

Sample space is a set of all possible outcomes of a random experiment.

Event

An event is a subset of a sample space

Probability of a sample space

Each outcome has a possibility in a sample space. The probability for each outcome is between 0 an 1. The sum of all possibility = 1.

Probability of an event

The probability of an event A, denoted by P(A), is obtained by adding the probability of the individual outcomes in an event.

- $0 \leq P(A) \leq 1$
- P(A) = 1 implies A always occurs
- P(A) = 0 implies A never occurs

When all possible outcomes are equally likely,

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of outcomes in sample space}}$$

Set Theory for Event using Venn Diagram

Complement of an Event

$$P(A^c) + P(A) = 1$$

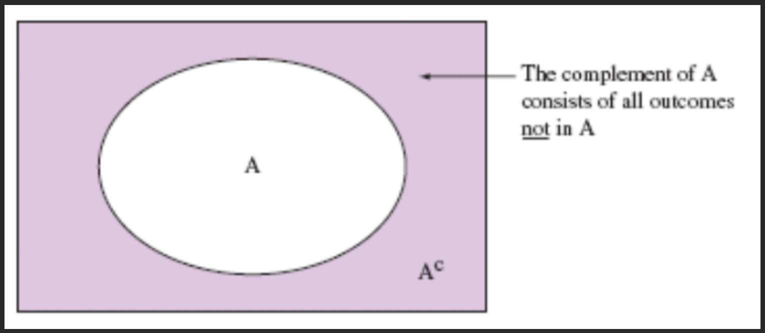

Figure 9: Complement of an event

Intersection of 2 Events

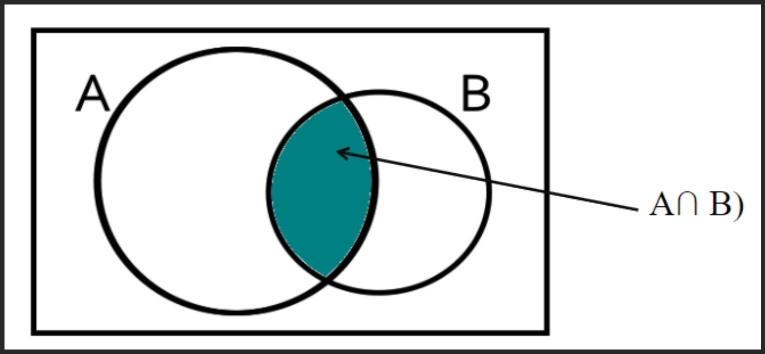

Figure 10: Intersection of 2 Events

Disjoint or Mutually Exclusive Event

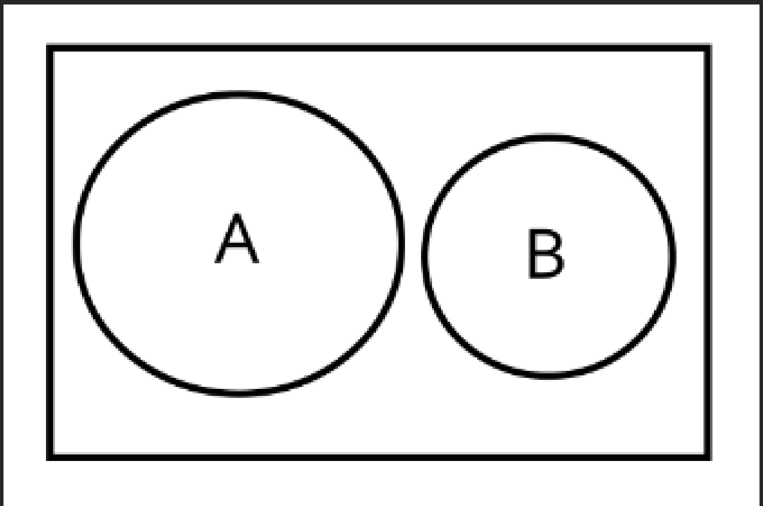

Figure 11: Disjoint Event

Union of 2 Events

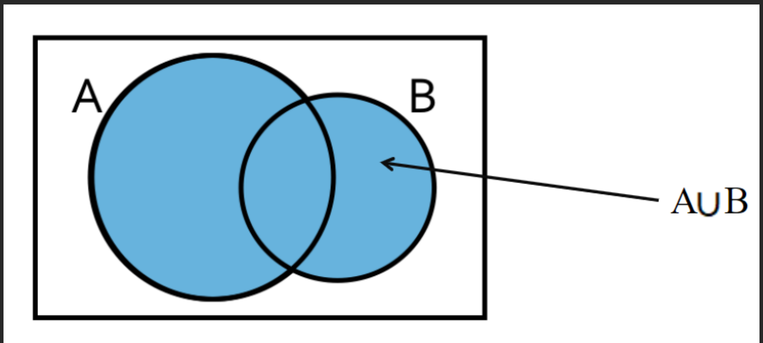

Figure 12: Union of 2 Events

Property of Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(B) \text{ if A and B are disjoint}$$

Conditional Probability

Conditional probability is used to determine how 2 events are related.

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

Independent Event

Event A and Event B are independent if knowing one event occurs does not change the probability of the other.

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$
$$P(A \cap B) = P(A)P(B) \text{ if A and B are independent}$$

If A and B are independent, A^c and B, A and B^c, A^c and B^c are also independent.

Tree Diagram

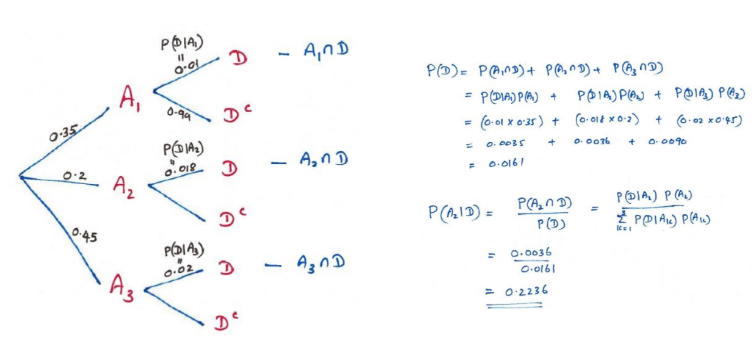

Figure 13: Tree Diagram

3 Random Variable

A random variable X is a function defined on the sample space S, assigning a number $x = X(\omega)$ to each outcome ω in the sample space. Random variables are defined by upper case letters such as X, Y, Z... Lower case values x, y, z represents a possible value of the random variable

Types of Random Variable

- Discrete
- Continuous

Note

- For a continuous random variable X, $P(X = c) = 0$ for any c.
- Any random variable whose possible values are either 1 or 0 is called Bernoulli's Random Variable

Discrete Random Variable

Probability Mass Function (pmf)

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

where f(x) gives the probability of each possible value x of X. It has the following properties.

$$f(x) \geq 0 \text{ for all x}$$
$$\sum f(x) = 1$$

Cumulative Distribution Function (cdf)

$$F(x) = P(X \leq x) = \sum_{d \leq x} f(k) \text{ for all real x}$$

Mean and Variance of a Discrete Random Variable

Mean(Expected Value) of a discrete random variable with pmf f(x) is

$$\mu = R[X] = \sum xf(x)$$

Variance of a discrete random variable is

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

The expected value of some function g(x) corresponding to the random variable X with pmf f(x) is

$$E[g(x)] = \sum g(x)f(x)$$

Continuous Random Variable

Probability Density Function

Let X be a continuous random variable, then the probability density function of X is a function f(x) such that for any 2 numbers a and b and $a < b$.

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$
$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Distribution Function

$$F(x) = P(X \leq x) = \int_{-infty}^{x} f(t)dt \text{ for all x}$$
$$P(X > a) = 1 - F(a)$$
$$P(a < X < b) = F(b) - F(a)$$

Mean and Variance of Continuous Random Variable

Mean of f(x) is

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

Variance is

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

Properties of Mean and Variance

1. $E(aX + b) = aE(x) + b$
2. $E(X + Y) = E(X) + E(Y)$
3. $E(XY) = E(X)E(Y)$ if X and Y are independent
4. $Var(aX + b) = a^2 Var(X)$
5. $Var(X + Y) = VAR(X) + VAR(Y)$ if X and Y are independent
6. $Var(X_Y) = VAR(X) - VAR(Y)$ if X and Y are independent

Some Continuous Random Variable Model

Uniform Distribution

If X is a uniform random variable, $X \sim U(a, b)$, X is evenly distributed on the interval [a,b].

$$pdf = f(x) = \frac{1}{b-a}$$
$$\mu = E[X] = \frac{a+b}{2}$$
$$\sigma^2 = \frac{(b-a)^2}{12}$$

Exponential Distribution

Exponential random variables are often used to model the time until an event occur. A random variable $Xexp(\lambda)$ has,

$$pdf = f(x) = \lambda e^{-\lambda x}; x \geq 0; \lambda > 0$$
$$\mu = E[X] = \frac{1}{\lambda}$$
$$\sigma^2 = \frac{1}{\lambda^2}$$

where λ is called the rate.

Sum and Average of Independent Random Variable

Let $X_1, X_2, X_3 ... X_n$ be n independent random variables. $Y = a_1X_1 + a_2X_2 ... a_nX_n$.

$$E[Y] = a_1E[X_1] + a_2E[X_2]...a_nE[X_n]$$
$$Var[Y] = a_1^2Var[X_1] + a_2^2Var[X_2]... + a_n^2Var[X_n]$$

If all a are 1, then

$$\bar{X} = (X_1 + ...X_n)/n$$

Max and Min of Independent Random Variables

The maximum(V) and the minimum($\wedge$) of a sequence of n independent random variables can be used to model a number of random quantities.

Maximum

$$V = max\{X_1, X_2, ...X_n\}$$

V can be used to model

1. The lifetime of a system of n components connected in parallel
2. The completion time of a project made up of n subprojects which can be pursued simultaneously
3. The maximum flood level of a river in the next n years

$$\begin{aligned} F_V(v) = P(V \leq v) &= P(X_1 \leq v, X_2 \leq v \dots X_n \leq v) \\ &= F_1(v)F_2(v)...F_n(v) \end{aligned}$$

where $P(x_i \leq v) = F_i(v)$

Minimum

$$\wedge = min\{X_1, X_2, ...X_n\}$$

$\wedge$ can be used to model

1. The lifetime of a system of n components connected in series
2. The completion time of a project independently pursued by n competing teams
3. The minimum flood level of a river in the next n years

$$\begin{aligned} F_\wedge(u) = P(\wedge \leq v) = 1 - P(\wedge > u) &= 1 - P(X_1 \geq u, X_2 \geq u \dots X_n \geq u) \\ &= 1 - ((1 - F_1(u))(1 - F_2(u))...(1 - F_n(u)) \end{aligned}$$

4 Normal Distribution

Standard Normal

When $a = 1/\sigma$ and $b = -\mu/\sigma$,

$$Z = \frac{x - \mu}{\sigma}$$
$$\mu = 0$$
$$\sigma^2 = 1$$

The standard normal density is denoted by ϕ, the standard normal distribution is denoted by Φ.

$$\phi(z) = \frac{1}{2\pi} e^{\frac{-z^2}{2}}$$
$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{2\pi} e^{\frac{-t^2}{2}} dt$$
$$\Phi(-z) = 1 - \Phi(Z)$$

Normal density cannot be integrated in close form, therefore tables are used.

$$F(x) = P(X \leq x) = P(z < \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$$

The 68-95-99.7 Rule

In the normal distribution with mean μ and standard deviation σ,

- approximately 68% of observation falls within σ of μ
- approximately 95% of observation falls within 2σ of μ
- approximately 99.7% of observation falls within 3σ of μ

Properties of normal distribution

1. If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.
2. Suppose $X_1 \dots X_n$ are independent normal random variables. $X_i \sim N(\mu_i, \sigma_i^2)$, If Y is a linear combination of X, then $Y \sim N(a_1\mu_1 + ... a_n\mu_n, a_1^2\sigma_1^2 + \dots a_n^2\sigma_n^2)$.
3. If $X_1 \dots X_n$ are independent, identically distributed random variable with mean μ and std σ.Then $\bar{X} = N(\mu, \sigma^2/n)$. Where $\bar{X}$ is a linear combination of all X.

Some Probability Model

Bernoulli's Ditribution

Bernoulli's random variable is a random variable that takes value of 1 in case of success and 0 in case of failure.

$$X(success) = 1$$
$$X(failure) = 0$$

$$X \sim Bernoulli(P)$$
$$P(X = x) = P^x(1 - P)^{1-p}; x = 0, 1 \ldots$$

where p is the probability of success.

$$E[X] = P$$
$$Var[X] = p(1 - p)$$

Binomial Distribution

A binomial random variable is the number of success for n independent trials.

$X \sim Bin(n, p)$ if the discrete random variable X has the pmf

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x};\ x = 1, 2, 3, ...n$$

where n is the number of trials and p is the probability of success.

$$E[X] = np$$
$$Var[X] = np(1 - p)$$

Motivation for binomial

Consider performing an experiment n times where the probability of success in every trial is p and the n trials are independent. We are interested in the probability of X successes. The probability of getting X successes and n-x failures in the specific order is $P^x(1 - p)^{n-x};\ x = 0, 1, 2, n$. Since there are $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ ways the trial outcome can be ordered, $P(xsuccesses) = \binom{n}{x} p^x (1 - p)^{n-x}$.

Poisson's Distribution

Poisson's Distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occurs with a known constant rate and independently.

X - the number of occurances in a given inteval or space

$X \sim Poisson(\lambda)$

$$pmf = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}\ x = 1, 2, 3, \ldots$$

where λ is the rate of occurrences.

$$e^k = 1 + k + k^2/(2!)... = \sum_0^\infty \frac{k^x}{x!}$$
$$E(X) = Var(X) = \lambda$$

Rules

- The probability of an event within a certain interval does not change over other intervals.
- The umber of occurrences of an event in an interval is proportional to the size of the inteval.
- Events cannot happen simultaneously

Poisson's Distribution to estimate binomial

If $X \sim Bin(n,p)$ with large n($n \geq 20$) and small p($np < 5$), then we can use Poisson's random variable with rate $\lambda = np$ to approximate binomial.

$$Bin(n,p) = Poisson(np)$$

Exponential Random Variable

Let T be the time between two consecutive occurrences of an event. Then T is a continuous rv and it has exponential density.

$$T \sim exp(\lambda)$$
$$f(t) = \lambda e^{-lambdat}$$
$$F(t) = 1 - e^{-\lambda t}$$

T is also the waiting time until the first event.

Geometric Distribution

A geometric random variable counts the number of independent trials needed until the first success occurs where each trial has a probability of success.

$$X \sim Geo(p)$$
$$f(x) = P(X = x) = P(1 - p^{x-1}); \; x = 1, 2, \ldots, n$$
$$F(x) = 1 - (1-p)^x; \; x = 1, 2, dots, n$$

$$E[X] = \frac{1}{p}$$
$$Var(X) = \frac{1-p}{p^2}$$

Normal Probability Approximations

Population and Sample

- Population: The entire collection of individuals we want to study.
- Sample: A subset of individuals selected from the population.

Statistical techniques are used to make conclusions about the population based on the sample.

Statistic and Parameter

- Statistic: A numerical summary of the sample. Ex. Sample mean, sample standard deviation.
- Sample: A numerical summary of the population. Ex. Population mean, population standard deviation.

Note that,

- Values of the parameter cannot be determined in practice.
- Due to sampling variability a statistic takes on different values for different samples.
- Parameters are estimated using sample date. Statistics is used to estimate parameters.

Sampling Distributions

The sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Sampling distributions describe the variability that occurs from study to study using statistics to estimate population parameters.

Sampling distributions help to predict how close a statistic falls to the parameter it estimates.

If $X = [X_1, X_2, ...X_n]$ is a sample from a normal population with mean μ and standard deviation σ, then

$$\bar{X} = \frac{X_1 + ...X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

However, samples do not always follow a normal distribution. Suppose a random sample of n observations is taken from a population with mean μ and standard deviation σ, then the mean of the mean of the samples is μ and the standard deviation of the mean of the samples is $\frac{\sigma}{\sqrt{n}}$. The standard deviation of the samples mean is called the standard error.

Central Limit Theorem

The CLT states that when an infinite number of successive random samples are taken from a population, the "sampling distribution of the means of those samples will become approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Suppose we draw a random sample of size n, $X_1, X_2...X_n$ from a population random variable that is distributed with mean μ and standard deviation σ. Do this repeatedly and then calculate the mean of each sample. As the sample size increases, the distribution of the mean of the drawn samples will approach a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{X} = \frac{X_1 + ...X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

The central limit theorem describes how the population mean and standard deviation are related to the mean and the standard deviation of the mean of the samples.

Then,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Example

Closing prices of stocks have a right-skewed distribution with mean of $\$25$ and standard deviation of $\$20$. What is the probability that mean of a random sample of 40 stocks will be less than $\$20$?

Consider the sampling distribution of sample mean. By CLT,

$$\bar{X} \sim N(25, \frac{20^2}{40})$$

$$\begin{aligned} P(\bar{X} < 20) &= P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{20 - 25}{\frac{20}{\sqrt{40}}}) \\ &= P(Z < -1.58) \\ &= P(Z > 1.58) \\ &= 1 - P(Z > 1.58) \\ &= 1 - P(Z < 1.58) \\ &= \boxed{0.0571} \end{aligned}$$

Example

The time taken by a randomly selected applicant for a mortgage to fill out a certain form has a normal distribution with mean of 10 minute and standard deviation of 2 minute. If five individuals fill out a form on one day, what is the probability that the sample average amount of time taken on that day is at most 11 min?

$$\mu = 10$$
$$\sigma = 2$$
$$n = 5$$
$$P(\bar{X} \leq 11) = ?$$

$$\begin{aligned} P(\bar{X} \leq 11) &= P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{11 - 10}{\frac{2}{\sqrt{5}}}) \\ &= (Z \leq 1.12) \\ &= 0.8686 \end{aligned}$$

TODO!!!!!!!!!!!!

Normal Approximation to the Binomial Distribution

Let $X \sim Bin(n, p)$. When n is large so that both $np \geq 5$, we can use the normal distribution to get an approximate answer.

$$X \sim N(np, np(1-p))$$

* When we use normal approximation to the Binomial distribution, the **continuity correction** should be used because we are approximating a discrete random variable with a continuous random variable.

Example

Let $X \sim Bin(10, 0.5)$, obtain $P(x \leq 2)$,

1. exactly
2. using the normal approximation
3. using the normal approximation with a continuity correction

1.

$$X \sim Bin(10, 0.5)$$
$$n = 10$$
$$p = 0.5$$

$$
\begin{aligned}
P(X \leq 2) &= P(x=0) + P(x=1) + P(x=2) \\
&= \binom{10}{0}(0.5)^0(0.5)^{10} + \binom{10}{1}(0.5)^1(0.5)^9 + \binom{10}{2}(0.5)^2(0.5)^8 \\
&= 0.0547
\end{aligned}
$$

2.

$$
\begin{aligned}
n &= 10 \\
p &= 0.5 \\
np &= 5 \geq 5 \\
n(1-p) &\geq 5
\end{aligned}
$$

Therefore $X \sim N(5, 2.5)$,

$$
\begin{aligned}
P(x \leq 2) &= P(\frac{x-\mu}{\sigma} \leq \frac{2-5}{\sqrt{2.5}}) \\
&= P(Z \leq -1.9) = 0.0287
\end{aligned}
$$

This is not so good because the exact answer is 0.0547.

3.

$$
\begin{aligned}
P(x \leq 2) = P(x \leq 2 + 0.5) &= P(\frac{x-\mu}{\sigma} \leq \frac{2.5-5}{\sqrt{2.5}}) \\
&= P(Z \leq -1.58) \\
&= 0.0571
\end{aligned}
$$

0.5 is the continuity correction. The end result is closer to the exact answer.

Continuity Correction

If x is a discrete random variable and y is a continuous random variable.

$$
\begin{aligned}
P(x > 4) &= P(x \geq 5) = P(y \geq 4.5) \\
P(x \leq 4) &= P(y \leq 4.5) \\
P(x < 4) &= P(x \leq 3) = P(y \leq 3.5) \\
P(x \leq 4) &= P(y \leq 4.5) \\
P(x = 4) &= P(4 - 0.5 \leq y \leq 4 + 0.5)
\end{aligned}
$$

Normal Approximation To The Poisson Distribution

If $X \sim Poisson(\alpha)$ where α is the expected number of counts. When α is large (≥ 20), then normal distribution can be used to approximate the Poisson distribution.

$$X \sim N(\alpha, \alpha)$$

Example

A radioactive element disintegrates such that it follows a Poisson distribution. The mean number of particles emitted is recoded in 1 second interval is 55. Find the probability of

(a) more than 60 particles are emitted in 1 second

(b) between 50 and 65 particles inclusive are emitted in 1 second

Let $X = \#$ of particles emitted in 1 second.

$$X \sim Poisson(55)$$

Since λ is large(≥ 20), $X \sim N(55, 55)$.

(a)

$$\begin{aligned} P(X > 60) &= P(X \geq 61) \\ &= P(X \geq 60.5) \\ &= P(\frac{x-\mu}{\sigma} \geq \frac{60.5-55}{\sqrt{55}}) \\ &= P(Z \geq 0.74) \\ &= 1 - P(Z < 0.74) \\ &= 1 - 0.7704 \\ &= 0.2296 \end{aligned}$$

(b)

$$\begin{aligned} P(50 \leq x \leq 65) &= P(49.5 \leq x \leq 65.5) \\ &= P(\frac{49.5-55}{\sqrt{55}} \leq \frac{x-\mu}{\sigma} \leq \frac{65.5-55}{\sqrt{55}}) \\ &= P(-0.74 \leq Z \leq 1.41) \\ &= 0.6911 \end{aligned}$$

Sum of Random Samples With CLT

Consider a random sample $X_1, X_2 ... X_n$ from a distribution with mean μ and variance σ^2. When n is large, by CLT

$$\bar{X} = \frac{X_1 + ... X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

If a question is about sum instead of an average, CLT can still be used. Let $T = X_1 + X_2 + ... X_n$,

$$E[T] = n\mu$$
$$Var[T] = n\sigma^2$$
$$T \sim N(n\mu, n\sigma^2)$$

Statistical Modeling and Inference

Statistical Inference

Method of making decisions or predictions about a population based on information obtained from a sample.

The objective of estimation is to determine the approximate value of a population parameter on the bases of a sample statistic.

Sample mean ($\bar{X}$) is used to estimate the population mean(μ).

Two Types of Estimators

- Point Estimator - draws inferences about a population by estimating the value of an unknown parameter using a single value or point.
- Interval Estimator - draws inferences about a population by estimating the value of an unknown parameter using an interval. The population parameter of interest is between some lower and upper bounds.

Point Estimate vs Interval Estimate

A point estimate does not tell us how close the estimate is likely to be to the parameter. An interval estimate is usually more useful.

Point Estimators

A good estimator has a sampling distribution that is centered at the parameter(unbiasedness).

An unbiased estimator of a population parameter is an estimator whose expected value is equal to that parameter.

$$E[\hat{\theta}] = \theta$$

where θ is the parameter and $\hat{\theta}$ is a point estimator.

A good estimator has a small standard error compared to the other estimators.

Example

Suppose we want to estimate the mean summer income of a class of statistics students. For a sample of 30 students sample mean($\bar{X}$) is calculated to be $\$500$ per week.

Point estimate for population mean(μ) income per week is $\hat{\mu} = \bar{X} = \$500$(point estimate). An alternative statement is: "The mean income is between $\$400 \sim \$600$ per week(interval estimate).

Example

Suppose that $X_1, X_2 ... X_n$ is a random sample from a population with mean μ and variance σ^2.

- $\bar{X}$ is an unbiased estimator of μ

$$E[\bar{X}] = E[\frac{\sum X_i}{n}] = \mu$$

- S^2 is an unbiased estimator of σ^2

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

Confidence Interval for μ

Consider the absolute estimation error $|\bar{Y} - \mu|$ (sample mean - population mean). We wish to find a value d such that there is a large probability (0.95 or 0.99) that the absolute estimation error is below d.

$$\text{Confidence Level} = P(|\bar{Y} - \mu| < d) = 1 - \alpha$$

where α is typically 0.05(95$\%$ confidence) or 0.01(99$\%$ confidence). Assume 0.005 if not defined.

The resulting d can be added or subtracted from the observed average $\bar{y}$ to obtain the upper and lower limits of an interval called $(1 - \alpha)$ 100% confidence interval.

$$(\bar{y} - d, \bar{y} + d)$$

Example

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
$$P(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95$$
$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Therefore, 95$\%$ confidence interval for μ is

$$[\bar{X} - Z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{0.025}\frac{\sigma}{\sqrt{n}}]$$

Example

The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6g per ml. Find the 95% and 99% confidence interval for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3g/ml.

$$\bar{X} = 2.6$$
$$n = 36$$
$$\sigma = 0.3$$

95$\%$ CI for μ

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$[2.507, 2.693]$$

99% **CI for μ**

$$[2.471, 2.729]$$

CI when σ is unknown

σ is estimated by the sample standard deviation S. This estimation introduces extra error. To account for this error, z-score is replaced by a slightly larger score called the t-score.

Interpreting a CI for μ

If we repeatedly obtain samples of size n and construct the corresponding 95% confidence interval for μ, on average, 95% of these intervals will include the value of μ.

Example

A reporter is writing an article on the cost of off-campus housing. A sample of 16 studio apartments within 3 km of campus resulted in a sample mean $\$1000/month$ and sample standard deviation of $\$150$. Assuming population to be normal, calculate 95% CI for the population mean rent per month.

$$\begin{aligned} n &= 16 \\ \bar{X} &= 1000 \\ S &= 150 \end{aligned}$$

$$\begin{aligned} \bar{X} \pm t_{0.025,n-1}\frac{s}{\sqrt{n}} \\ 1000 \pm 2.131\frac{150}{\sqrt{16}} \\ [920, 1080] \end{aligned}$$

Testing Hypothesis About μ

Hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

Null and Alternative Hypothesis

The null hypothesis, denoted by H_o, is a tentative assumption about a population parameter.

The alternative hypothesis, denoted by H_a is the opposite of what is stated in the null hypothesis. The alternative hypothesis is what the test is attempting to establish.

The equality part of the hypothesis always appears in the null hypothesis.

Hypothesis test about the value of a population mean μ must take one of the 3 forms.

- $H_o : \mu \geq \mu_o$ vs $H_a : \mu < \mu_o$
- $H_o : \mu \leq \mu_o$ vs $H_a : \mu > \mu_o$

- $H_o : \mu = \mu_o$ vs $H_a : \mu \neq \mu_o$

where μ_o is the hypothesized value of the population mean.

The hypothesis should be formulated before viewing or analyzing the data.

Test Procedures

A test procedures is specified by the following.

- Test Statistic: a function of the sample data on which the decision (Rejected H_o or do not reject H_o) is to be based.
- Rejection Region: the set of all test statistic values for which H_o will be rejected.

A test statistic is constructed assuming the null hypothesis is correct.

Case 1: σ is known test statistic is,

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

Case 2: σ is unknown test statistic is,

$$t = \frac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

When n is large,

$$t = \frac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \sim N(0, 1)$$

The null hypothesis will be rejected if and only if the observed or computed test statistic value falls in the rejection region.

We also use the test statistic to assess the evidence against the null hypothesis by giving a probability, p-value.

P-Value

The P-Value summarizes the evidence. It describes how unusual the data would be if H_o were true.

P-Value is defined as the probability of observing a result as extreme or more extreme than what we observed given that H_o is true.

Significance Level (α)

The Significance level is a predetermined number such that we reject H_o if the P-Value is less than or equal to that number.

In practice, the most common significance level is $\alpha = 0.05$.

When we reject H_o, we say the results are statistically significant.

- if $p - value \leq \alpha =>$ Reject H_o
- if $p - value > \alpha =>$ Do not reject H_o

Steps of Hypothesis Testing

1. Develop the null and alternative hypothesis
2. Specify the level of significance α
3. Collect the sample data and compute the test statistic

Then, 2 approaches can be used.

p-value approach

1. Use the value of the test statistic to compute the p-value
2. Reject H_o if $p - value \leq \alpha$
3. Conclusion

Critical Value Approach

1. Use the level of significance to determine the critical value and the rejection rule
2. Use the value of the test statistic and rejection rule to determine whether to reject H_o
3. Conclusion

Decisions and Types of Errors in Hypothesis Testing

- H_o is true / Reject $H_o =>$ type I error
- H_o is false / Do not reject $H_o =>$ type II error

What is a good test?
A test that rarely makes type I and type II errors.

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

We can control the probability of type I error by our choice of the significance level α.

It is difficult to control the probability of making type II error.

Statisticians avoid the risk of making a type II error by using "do not reject H_o" and not "accept H_o"

$1-\beta$ is referred to as the power of a test.The greater the power, the less likely type II error occurs.

α and β are test properties, and they are independent of data.

Power of a Test

Power is the probability of correctly rejecting the null hypothesis H_o, when H_o is false.

Tail Test

Use two tail test when $\mu = \mu_o$; use one tail test when $\mu \geq \mu_o$.

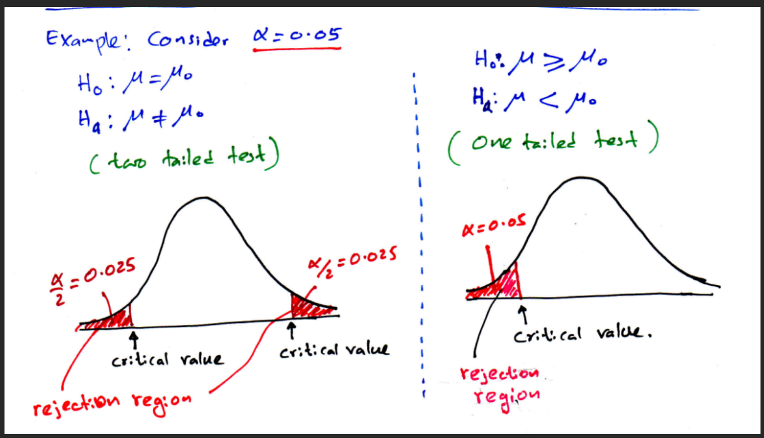

Figure 14: Tail Tests

Example

A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is more then \$170.

A random sample of 400 monthly accounts is drawn, from which the sample mean is $\$178$. The accounts are approximately normally distributed with $\sigma = \$65$.

(a) Can we conclude that the new system will be cost effective? (use $\alpha = 0.05$)?

(b) Describe what type I and type II errors in the content of this problem situation

(c) Considering the test procedure, find the rejection region of $\bar{X}$

(d) When $\mu = 180$, find the probability of type II error.

(e) Evaluate the power of the test when $\mu = 180$

(a)

$$\begin{aligned} n &= 400 \\ \bar{X} &= 170 \\ \sigma &= 65 \end{aligned}$$

$$\begin{aligned} H_o &= \mu \leq 170 \rightarrow \text{ use right tail test} \\ \mu_o &= 170 \\ H_a &= \mu > 170 \end{aligned}$$

$$\begin{aligned} Z_{obs} &= \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \\ &= \frac{178 - 170}{\frac{65}{\sqrt{400}}} = 2.46 \end{aligned}$$

Method1:Critical Value Approach Since $Z_{obs} = 2.46 > Z_{0.05} = 1.645$, we reject H_o. The conclusion is that the new system is cost-effective.

Method2:P-value Approach P-Value = P(observing data as extreme or more extreme than what we observed, given H_o is true)

$$\begin{aligned} &P(\bar{X} \geq 178 \text{ when } \mu = 170) \\ &= P(Z \geq 2.46) = 0.0069 \end{aligned}$$

Since $0.0069 < \alpha = 0.05$, H_o is rejected, the conclusion is that the new system is cost effective.

(b)

Type I error: reject H_o when H_o is true.

Conclude that the new billing system is cost-effective when it is not(i.e. true mean > 170).

Type II error: do not reject H_o when H_o is false.

Conclude that the new billing system is not cost-effective when it is.

(c)

Reject when $Z > 1.645$.

$$Z = \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$
$$\bar{X} > 175.35$$

Therefore, rejection region is $\bar{X} > 175.35$.

(d)

$$\mu = 180$$

When $\mu = 180$, $\bar{X}$ has a normal distribution with mean 180 and sigma of $\frac{65}{\sqrt{400}}$.

$$\begin{aligned}
\beta = \text{P(type II error)} &= \text{P(do not reject } H_o \text{ when } H_o \text{ is false)} \\
&= P(\bar{X} < 175.35 \text{ when } \mu = 180) \\
&= P(\frac{\bar{X} - 180}{\frac{65}{\sqrt{400}}} < \frac{175.35 - 180}{\frac{65}{\sqrt{400}}}) \\
&= P(Z < -1.43) \\
&= 0.0764
\end{aligned}$$

(e)

$$\begin{aligned}
\text{Power} &= \text{P(Reject } H_o \text{ when } H_o \text{ is false)} \\
&= 1 - \beta \\
&= 1 - 0.0764 \\
&= 0.9236
\end{aligned}$$

This is a very powerful test since it makes the correct decision 92.36$\%$ of the time when $\mu = 180$.

Two Sample Problems

Compare the means of two independent populations, assuming equal population standard deviations.

*Suppose we draws a random sample from each of the two independent populations with means μ_1, μ_2, and standard deviations of σ_1 and σ_2.

Hypotheses take one of the following 3 forms.

- Left-Tailed

$$H_o : \mu_1 - \mu_2 \geq \Delta_o$$
$$H_a : \mu_1 - \mu_2 < \Delta_o$$

- Right-Tailed

$$H_o : \mu_1 - \mu_2 \leq \Delta_o$$
$$H_a : \mu_1 - \mu_2 > \Delta_o$$

- Two-Tailed

$$H_o : \mu_1 - \mu_2 = \Delta_o$$
$$H_a : \mu_1 - \mu_2 \neq \Delta_o$$

Example

If the Hypotheses are

$$H_o : \mu_1 \geq \mu_2 \rightarrow \mu_1 - \mu_2 \geq 0$$
$$H_a : mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0$$

In this case, $\Delta = 0$.

Assumptions

- random samples from each of the population is drawn
- the sample individuals are independent of each other
- both populations are normal or we need reasonably large samples to validate using CLT
- both population distributions have equal variance ($\sigma_1^2 = \sigma_2^2$)

Test Statistic

We select a simple random sample of size n_1, from population 1 and a simple random sample of size n_2.

Let $\bar{X}_1$ be the mean of sample 1 and $\bar{X}_2$ be the mean of sample 2.

Test statistic t is.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_o}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where S_p is the pooled standard deviation.

The Pooled Standard Deviation

This method requires the assumption that population variances are equal.

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

S_p, the pooled standard deviation estimates the common value σ.

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

$(1-\alpha)100\%$ confidence interval for the difference between two population means (i.e. $\mu_1 - \mu_2$).

- Point Estimator of $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$
- CI $\rightarrow$ point estimate $\pm$ margin of error.
- $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example

Average densities of two types of brick are compared. (Type A and Type B).

a)

Using the following sample data, test the claim that the true mean densities are equal.

b)

Use a 0.05 significance level and assume normality of the two density distribution and equal population variances.

c)

Calculate 95% CI for $\mu_A - \mu_B$.

Type A	Type B
$n_A = 8$	$n_B = 10$
$\bar{X_A} = 22.7$	$\bar{X_B} = 21.5$
$S_A = 0.8$	$S_B = 0.6$

Let μ_A be the true average density for type A and μ_B be the true density for Type B.

Hypotheses

$$
\begin{aligned}
H_o &: \mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0 \\
H_a &: \mu_A \neq \mu_B \rightarrow \mu A = \mu_B \neq 0
\end{aligned}
$$

Use two tail test and $\alpha = 0.05$.

Pooled Standard Deviation

$$
\begin{aligned}
S_p^2 &= \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} \\
&= \frac{(8-1)(0.8)^2 + (10-1)(0.6)^2}{8 + 10 - 2} \\
&= 0.4825 \\
S_p &= 0.695
\end{aligned}
$$

Test Statistic

$$
\begin{aligned}
t_{obs} &= \frac{(\bar{X_A} - \bar{X_B}) - 0}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_1+n_2-2} \\
&= \frac{22.7 - 21.5}{0.695\sqrt{\frac{1}{8} + \frac{1}{10}}} \\
&= 3.64
\end{aligned}
$$

Critical Value Approach

$$\alpha = 0.05$$
$$\alpha/2 = 0.025$$
$$t_{0.025,16} = 2.12$$

Calculated test statistic value is in the rejection region.

$$|t_{obs}| = 3.64 > t_{0.025,16} = 2.12 \rightarrow \text{Reject } H_o \text{ at } \alpha = 0.05$$

Conclusion

At the significant level 0.05, we conclude that the true mean densities of two types of brick are not equal.

c)

$$(\bar{X_A} - \bar{X_B}) \pm t_{\frac{\alpha}{2}, n_A+n_B-2} \cdot S_p \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$
$$(22.7 - 21.5) \pm 2.12(0.695)\sqrt{\frac{1}{8} + \frac{1}{10}}$$
$$1.2 \pm 0.33$$
$$[0.87, 1.53]$$

Hypotheses

$$H_o : \mu_A - \mu_B = 0$$
$$H_a : \mu_A - \mu_B \neq 0$$

The calculated confidence interval does not contain the hypothesized value. Therefore we can reject the null hypothesis, H_o.

Comparison of Several Means

Analysis of Variance ANOVA

ANOVA is a statistical method that tests the equality of three or more population means by analyzing sample variances or variation in the data.

he simplest ANOVA problem is referred to variously as a single-factor, single-classification, or one-way ANOVA.

Example

1. An experiment to study the effect of five different brands of gasoline on automobile engine operating efficiency (mpg)

2. An experiment to study the effect of the presence of three different sugar solutions on bacterial growth.

One-Way ANOVA

One-way ANOVA focuses on a comparison of 3 or more population or treatment means.

Let k be the number of populations or treatments being compared.

μ_1 is the mean of population 1 or the true average response when treatment 1 is applied.

.

.

.

μ_k is the mean of population k or the true average response when treatment k is applied.

Hypotheses

$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_a : \text{at least two of the } \nu_i \text{ are different}$$

Reject H_o here means that at least two population means have different values.

Assumptions for ANOVA

For each population, the response variable is normally distributed.

The variance of the response variable, denoted σ^2, is the same for all the populations.

The observations must be independent.

Notation

k random samples observed. y_{ij} is the j_{th} observed value from the i^{th} population.

Treatment:	1	2	...	i	...	k
	y_{11}	y_{21}		y_{i1}		y_{k1}
	y_{12}	y_{22}		y_{i2}		y_{k2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n_1}	y_{2n_2}		y_{in_i}		y_{kn_k}
Total	$y_{1\cdot}$	$y_{2\cdot}$		$y_{i\cdot}$		$y_{k\cdot}$
Mean	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$		$\bar{y}_{i\cdot}$		$\bar{y}_{k\cdot}$

k^{th} treatment mean.

where $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \frac{y_{i\cdot}}{n_i}$

where,

$$\bar{y_i} = \frac{\sum_{j=1}^{ni} y_{ij}}{n_i} = \frac{y_{i0}}{n_i}$$

$$\text{Total Sample Size } = n = n_1 + ...n_k$$

$$\text{Grand Total } = y_{oo} = \sum_{k}^{i=1}\sum_{n_i}^{j=1} y_{ij}$$

$$\text{Grand Mean } = \bar{y_{oo}} = \frac{y_{oo}}{n} = \frac{\sum_{k}^{i=1}\sum_{n_i}^{j=1} y_{ij}}{n}$$

Let Y_{ij} be the random variable that denotes the j^{th} measurement taken from the i^{th} treatment.

Then y_{ij} is the observed value of Y_{ij}.

$$E[\bar{y_i}] = \mu_i$$
$$Var(\bar{y_i}) = \frac{\sigma}{n_i}$$

For k random samples, we can find calculate the sample variances.

$$S_1^2, S_2^2...S_k^2$$

$S_1^2, S_2^2...S_k^2$ are k different unbiased estimates for the common variance σ^2.

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

These k estimates can be combined to obtain an unbiased estimate for σ^2.

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{n - k} = MSE$$

where

$$S_i^2 = \frac{\sum_{i=1}^{n_i}(y_{ij} - \bar{y_{ij}})^2}{n_i - 1}$$

H_o is true

Sample means are close together because there is only one sampling distribution.

H_o is false

Sample means comes from different sampling distributions and are not close together.

4.1 Total Variation In The Data (SSE - total sum of squares)

Comes from 2 sources.

- Variation between groups/treatments. (SSTr- Treatment sum of squares)
- Variation within groups/treatments.(SSE - Error sum of squares)

$$SST = SSTr + SSE$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_{oo}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}y_{oo}^2$$
$$SSTr = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{io} \bar{-} \bar{y_{oo}})^2 = \sum_{i=1}^{k}\frac{1}{n_i}y_i^2 - \frac{1}{n}y_{oo}^2$$
$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_{io}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{y_{io}^2}{n_i} = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

Degrees of Freedom

$$df(SST) = n - 1$$
$$df(SSE) = n - k$$
$$df(SSTr) = k - 1$$

Mean Squares

$$\text{Mean Square Treatment} = MSTr = \frac{SSTr}{k-1}$$
$$\text{Mean Square Error} = MSE = \frac{SSE}{n-k}$$

*MSE is a measure of with-sample variation.

ANOVA Test Procedure

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a : \mu_i \neq \mu_j \text{ for } i \neq j$$

Test Statistic

$$F_{obs} = \frac{MSTr}{MSE} \sim F_{\gamma_1, \gamma_2}$$

Under H_o, F_{obs} follows the F-distribution with degrees of freedom,

$$\gamma_1(\text{numerator df}) = df(SSTr) = k - 1$$
$$\gamma_2(\text{Denominator df} = df(SSE) = n - k$$

F-distribution

Reject H_o if $F_{obs} \geq F_\alpha$.

The ANOVA TABLE

Source of Variation	df	Sum of Squares	Mean Squre	F-ratio
Treatment	$k-1$	$SSTr$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		

The ANOVA Model

The assumptions of single-factor ANOVA can be described sufficiently by means of the "model equation".

Each measurement will be represented as the sum of two terms, as unknown constant, μ_i, and a random variable, ϵ_{ij}.

$$Y_{ij} = \mu_i + \epsilon_{ij}$$
$$i = 1, 2, \ldots, k$$
$$j = 1, 2, \ldots, n_i$$

where ϵ_{ij} represents a random deviation from the population or true treatment mean μ_i.

The model assumptions are:

1. Independence: The random variables ϵ_{ij} are independent (implying that X_{ij} are also).
2. Constant Treatment Means: $E(\epsilon_{ij}) = 0$ for all i and j
3. Constant Variance: $Var(\epsilon_{ij}) = \sigma^2$ for all i and j
4. Normality: The variables ϵ_{ij} are normal.