Variable is a characteristic or attribute that can assign different values → Data is the values of variables

Population is a collection of all units to be studies

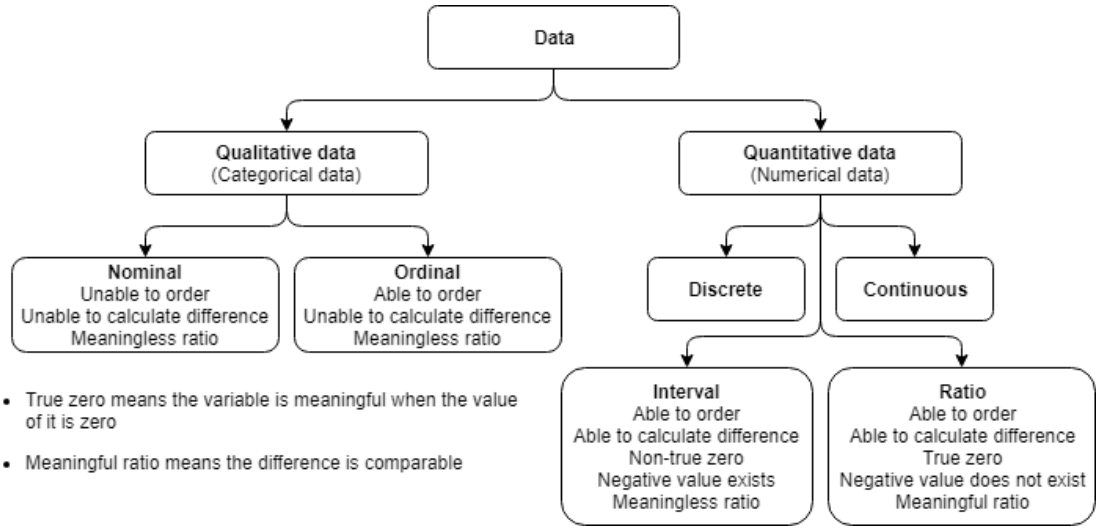
- Parameter is the numerical characteristic of the population; it is a constant and always unknown

Sample is a subset of the population

- Statistic is the numerical characteristic of the sample; it is always used to estimate parameter

Descriptive statistics summarise the given data (sample) with statistics

Inferential statistics use statistics to estimate parameters



Sampling techniques

- Probability sampling
  - Simple random sampling (SRSWR or SRSWOR)
    - Basic probability selection (every unit has an equal probability to be obtained)
  - Systematic random sampling
    - Determine step size $k = \frac{N}{n}$ (always round down)
    - Randomly select the first sample from the first k unit, select every k^{th} unit afterwards
  - Stratified random sampling
    - Take sample from each stratum
    - Units within the same stratum are homogenous (with common characteristics)
  - Cluster random sampling
    - Take sample by selecting clusters (all units within selected clusters will be obtained)
    - Primary or ultimate units may be heterogenous within the same cluster
- Non-probability sampling
  - Convenient
    - Sample is selected by if it is easy to be obtained

Observational study draws conclusions based on observations (has no control over the variables)

Experimental study applies treatments to variables and see the influences and results

Raw data is the original form of data

Measures of central tendency

- ⊙ (Arithmetic) mean

	Ungrouped	Grouped	Weighted
➲ Population	$\mu = \frac{\sum X_i}{N}$	$\mu = \frac{\sum f_{class} \times X_{midpoint}}{N}$	$\mu = \frac{\sum w_i \times X_i}{\sum w}$
➲ Sample	$\bar{x} = \frac{\sum x_i}{n}$	$\bar{x} = \frac{\sum f_{class} \times x_{midpoint}}{n}$	$\bar{x} = \frac{\sum w_i \times x_i}{\sum w}$

  - ☆ Less robust: it is sensitive to outliers
- ⊙ Median (MD)
  - ➲ Odd n: the middle of the ordered data
  - ➲ Even n: average of the middle two of the ordered data
  - ☆ More robust: it is less sensitive to outliers
- ⊙ Mode is a value with highest occur frequency
  - ➲ Either no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal)
  - ➲ The midpoint of the modal class for grouped data
- ⊙ Midrange
  - ➲ $MR = \frac{min+max}{2}$

Measures of variation

- ⊙ Range (R)
  - ➲ $R = max - min$
- ⊙ Variance is the typical squared distance of a data value from the mean
  - ➲ Population $\sigma^2 = \frac{\sum(X_i-\mu)^2}{N} = \frac{N\sum X_i^2-(\sum X_i)^2}{N^2}$
  - ➲ Sample $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{n\sum x_i^2-(\sum x_i)^2}{n(n-1)}$
- ⊙ Standard deviation
  - ➲ Population $\sigma = \sqrt{\sigma^2}$
  - ➲ Sample $s = \sqrt{s^2}$
- ⊙ Coefficient of variation measures relative variation
  - ➲ $CVar = \frac{\text{standard deviation}}{\text{mean}} \times 100\%$
- ⊙ Range rule of thumb finds the approximate maximum and minimum
  - ➲ $\bar{x} \pm 2s$, where $s = \frac{range}{4}$
  - ☆ For unimodal or (approximately) normal distribution only
- ⊙ Chebyshev's theorem is at least K% of data values will fall between a range
  - ➲ $Probability = 1 - \frac{1}{k^2}$, where $k = \frac{x-mean}{\text{standard deviation}}$ and $k > 1$
- ⊙ Empirical rule
  - ➲ $1\sigma \to 34\%, 2\sigma \to 13.5\%, 3\sigma \to 2.35\%$ from the mean
  - ☆ For normal (bell-shaped) distribution only

Measures of position

- z-score/standard score measures the relative position
  - Population $z = \frac{X-\mu}{\sigma}$
  - Sample $z = \frac{x-\bar{x}}{s}$
- Percentile
  - The rank of percentile (Rank is number of values below X in ascending order)
    - $\text{percentile} = \frac{\text{Rank}+0.5}{n} \times 100\%$
  - The value c corresponds to the p^{th} percentile
    - $c = np$
- Decile is a group out of 10 from data
  - $D_n = P_{n\times 10}$
- Quartile
  - $Q_1 = P_{25}, Q_2 = MD, Q_3 = P_{75}$
- Interquartile range
  - $IQR = Q_3 - Q_1$
- Outlier is a value that lie far away from the majority of data
  - Data outside the range of $[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$

Five-number summary

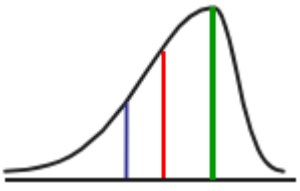
- $\text{Low}, Q_1, MD, Q_3, \text{High}$

Boxplot

- Left shewed: longer tail to the left
- Symmetric: same length of tails
- Right shewed: longer tail to the right

Shape of a distribution

$\text{Mean} < \text{Median} < \text{Mode}$

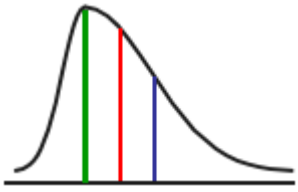


Left skewed

$\text{Mean} = \text{Median} = \text{Mode}$

Symmetric

$\text{Mean} > \text{Median} > \text{Mode}$



Right skewed

Probability is the chance of occurring an event

- Classical probability
  - Assume all outcomes in the sample space are equally likely to occur
  - $P(E) = \frac{N_{E_i}}{N} = \frac{\#\text{ of outcomes satisfy the event}_i}{\text{total \# of outcomes in the sample space}}$
- Empirical probability
  - The proportion of an occurred event
  - $P(E) = \frac{f}{n} = \frac{\text{frequency of desired class}}{\text{sum of all frequencies}}$
- Subjective probability
  - An individual judgment or opinion about the probability of occurrence

A probability experiment is a trial to produce outcomes

An outcome is a result of a trial in a probability experiment

A sample space is a set of all possible outcomes of a probability experiment

An event consists of outcomes

Probability rules

- $0 \leq P(E) \leq 1$
- $P(S) = 1$

Types of events

- Simple event is an event with one characteristic
- Joint event is an event with two or more characteristics

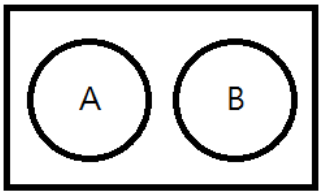
Mutually exclusive is event A and event B have no common outcomes

- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B)$

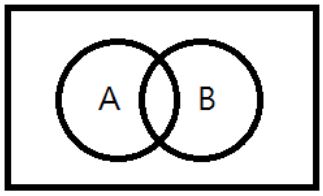
Complement of an event

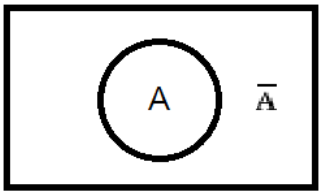
- $P(\overline{A}) = 1 - P(A)$

Mutually exclusive

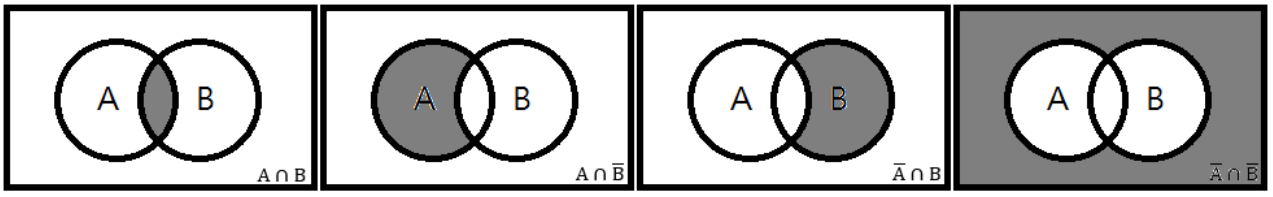


Not mutually exclusive

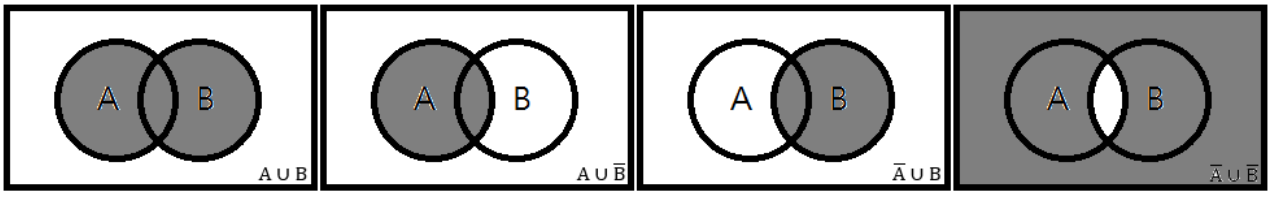




Intersection of events



Union of events



Addition rules

- $P(A \cup B) = P(A) + P(B)$, for mutually exclusive
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, for non mutually exclusive
  - Since $P(A \cap B)$ is duplicated, it has to be reduced

Multiplication rules

- $P(A \cap B) = P(A) \times P(B)$, for independent events
  - If $P(A \cap B) = P(A) \times P(B)$, event A and event B are statistical independent
    - $P(A)$ has no influence on $P(B)$, vice versa
- $P(A \cap B) = P(B) \times P(A|B)$, for dependent events

Conditional probability is the probability of event A given that event B occurred

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

Counting rules

- Fundamental counting rule/multiplication of choices
  - $\prod k_i$
- Factorial
  - $n!$
- Permutation
  - $nPr = \frac{n!}{(n-r)!}$
- Combination
  - $nCr = \binom{n}{r} = \frac{n!}{(n-r)!r!}$

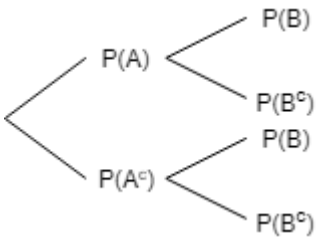
Law of total probability is the sum of subset event A

- $P(A) = \sum P(A \cap B_i)$

Bayes' theorem incorporates new additional information to revise prior probability to posterior probability

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

i.e.,



- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(\bar{A} \cap B)}$
- $P(A|B) + P(\bar{A}|B) = 1$

Random variable is a function maps element to real number

Mean/ expectation

➲ $E(X) = \mu = \sum x_i P(x_i)$

Variance

➲ $Var(X) = \sigma^2 = E(X^2) - E(X)^2$

Binomial expression

➲ $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$

Bernoulli distribution

➢ $X \sim Bernoulli(p)$

➲ $f(x) = \begin{cases} p, & \text{for } x = 1 \text{ or success} \\ p-1, & \text{for } x = 0 \text{ or failure} \end{cases}$

☆ Experiment only results either success or failure (mutually exclusive), where trial is one

Binomial distribution

➢ $X \sim B(n, p)$

➲ $P(x) = \binom{n}{x} p^x (1-p)^{n-x}$, where p is constant probability for success

➲ $\mu = np$

➲ $\sigma^2 = np(1-p)$

☆ Experiment only results either success or failure (mutually exclusive), where trials are fixed

- Consists of a sequence of n identical (independent) Bernoulli trials with replacement

⊙ Exactly x

➲ $P(x)$

⊙ Less than x

➲ $P(X < x) = \sum_{i=0}^{x-1} P(x_i)$

⊙ More than x

➲ $P(X > x) = \sum_{i=x+1}^{n} P(x_i)$

⊙ At least x

➲ $P(X \geq x) = 1 - P(X < x - 1)$

⊙ At most x

➲ $P(X \leq x) = 1 - P(X > x + 1)$

Multinomial distribution

➲ $P(x) = \frac{n!}{\prod x_i!} \times \prod p_i^{x_i}$, where p_i is the probability of $event_i$ and x_i is the frequency of $event_i$

☆ Experiment results more than 2 possible outcomes for each trial

Poisson distribution

➢ $X \sim Pois(\lambda)$

➲ $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, where λ is the mean number of occurrences per unit

☆ Independent variables occur over a period of time where n is large, and p is small

Hypergeometric

➢ $X \sim HyperGeom(N, r, n)$

➲ $P(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, where N is the population size, r is the number of successes in the population, n is the sample size, and x is the number of successes in the sample

☆ Experiment has only 2 outcomes in sampling without replacement

Normal distribution/Gaussian distribution

- ➢ $X \sim N(0, 1)$ for standard normal distribution
- ➢ $X \sim N(\mu, \sigma^2)$ for normal distribution

➲ $f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$, for $x \in \mathbb{R}$ ➲ $F(x) = \int_{-\infty}^{\infty} f(x)\,dx$

☆ The curve is continuous, unimodal, and symmetric about the mean where $\text{mean} = \text{median} = \text{mode}$

☆ $1\sigma \to 34\%, 2\sigma \to 13.5\%, 3\sigma \to 2.35\%$ from the mean

Standardisation of a random variable X using z value

➲ $z = \frac{X-\mu}{\sigma}$

Probability under standard normal distribution

⊙ Left-hand side/to the left probability

➲ $P(X \le x_0) = P\left(\frac{X-\mu}{\sigma} \le \frac{x_0-\mu}{\sigma}\right) = P(Z \le z_0)$

⊙ Right-hand side/to the right probability

➲ $P(Z \ge z_0) = 1 - P(Z < z_0)$

⊙ Between two sides probability

➲ $P(z_0 \le Z \le z_1) = P(|Z| < z_0) = P(Z \le z_1) - P(Z < z_0)$

⊙ Two sides probability

➲ $P(Z \le z_0 \cup Z \ge z_1) = P(|Z| > z_0) = P(Z \le z_0) + P(Z \ge z_1) = 1 - P(z_0 \le Z \le z_1)$

⊙ Probability of negative standard score

➲ $P(Z < -z_0) = 1 - P(Z < z_0)$

➲ $P(Z > -z_0) = P(Z < z_0)$

☆ Calculator FMLA 04 for reference checking ONLY

– Input z value → result in cumulative probability

Probability to X value

➲ $X = \mu + z\sigma$

Probability corresponds to standard score

➲ $P(Z < z_0) = P(x)$ ➲ $P(|Z| < z_0) = 1 - \frac{1-P(x)}{2}$

➲ $P(Z > z_0) = 1 - P(x)$ ➲ $P(|Z| > z_0) = \frac{1-P(x)}{2}$

Normality checking

⊙ Histogram

☆ The diagram is approximately bell-shaped

⊙ Pearson's index of skewness

➲ $PI = \frac{3(\bar{x}-MD)}{s}$ left skewed $\text{skewness} < 0$ symmetric $\text{skewness} = 0$ right skewed $\text{skewness} > 0$

☆ Data are not significantly skewed; $-1 < \text{skewness} < 1$

⊙ Outliers

➲ Data outside the range of $[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$, where $IQR = Q_3 - Q_1$

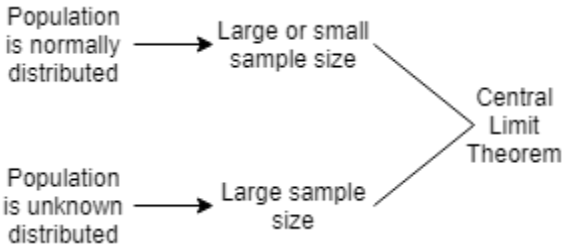
☆ No outliers

Central limit theorem (CLT) states that when a large number of simple random samples are selected from the population and the mean is calculated for each then the distribution of these sample means will assume the normal probability distribution

Theory of CLT

1) If the population is $X \sim N(\mu, \sigma^2)$, a sample with size n (either small or large); or
2) If the population is not normal distributed, a sample with size n (large)

Then, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$



- ➲ $E(\overline{X}) = \mu_{\overline{X}} = \mu$
- ➲ $Var(\overline{X}) = \sigma^2_{\overline{X}} \neq \sigma^2$

Sampling distribution uses sample statistic to contract a distribution

- ➢ All the possible random samples of size n that are drawn from the same population

Sampling error is a fact that unrepresentative of sample; the difference between parameters and statistics

Standard error

- ⊙ Sampling with replacement
  - ➲ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- ⊙ Sampling without replacement
  - ➲ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
- ⊙ Finite population correction factor
  - ➲ $\sqrt{\frac{N-n}{N-1}}$

☆ Sample size n increase, standard error $\sigma_{\bar{x}}$ decrease; vice versa

Standardisation under central limit theorem

- ⊙ Sampling with replacement
  - ➲ $z = \frac{\bar{x} - \mu_{\overline{X}}}{\sigma/\sqrt{n}}$
- ⊙ Sampling without replacement
  - ➲ $z = \frac{\bar{x} - \mu_{\overline{X}}}{\sigma/\sqrt{n} \times \sqrt{(N-n)/(N-1)}}$

Normal approximation to the binomial distribution

– If $X \sim B(n, p)$ and $np \geq 5$ and $n(1-p) \geq 5$, then random variable X is approximately $N(np, npq)$

➲ Continuity correction transfers discrete distribution into continuous distribution

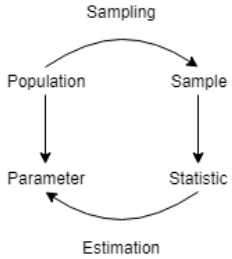
- ⊙ $P(X = x_0) \rightarrow P(x_0 - 0.5 < X < x_0 + 0.5)$
- ⊙ $P(X \geq x_0) \rightarrow P(X > x_0 - 0.5)$
- ⊙ $P(X > x_0) \rightarrow P(X > x_0 + 0.5)$
- ⊙ $P(X \leq x_0) \rightarrow P(X < x_0 + 0.5)$
- ⊙ $P(X < x_0) \rightarrow P(X < x_0 - 0.5)$

Point estimation is a process of estimating parameter by using point estimator

Point estimator is a random variable that used in statistical inference (point estimation)

Point estimate is the value of an estimated parameter

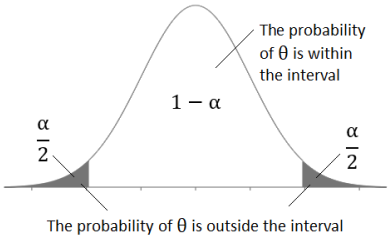
$$\text{estimator} \begin{Bmatrix} \overset{\text{estimation}}{\bar{x} \rightarrow \mu} \\ s^2 \rightarrow \sigma^2 \\ \hat{p} \rightarrow p \\ \hat{p}\hat{q} \rightarrow pq \end{Bmatrix} \text{estimate}$$



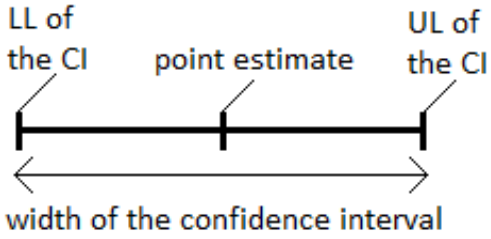
- Unbiased: the expected value of estimator should equal to the value of parameter i.e., $E(\hat{\theta}_n) = \theta$
- Consistent: the estimate approaches to the parameter when sample size increase i.e., $\hat{\theta}_n \rightarrow \theta$ as $n \rightarrow \infty$
- Relatively efficient: the standard error should be small i.e., $Var(\hat{\theta}_n) \rightarrow 0$ as $n \rightarrow \infty$

Interval estimate is a range of estimates; it is a random interval since it depends on the random sample

– Confidence level is the probability that the interval estimate will contain the parameter



– Confidence interval is a specific interval estimate of a parameter



☆ The true value of parameter is possible that not within the interval

Confidence interval for population mean

X~N		X~unknown	
known $\sigma, n \geq 1$	unknown $\sigma, n < 30$	known $\sigma, n \geq 30$	unknown $\sigma, n \geq 30$
$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\bar{x} \pm t_{\alpha/2;(n-1)}\left(\frac{s}{\sqrt{n}}\right)$	$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
☆ CLT is applied	☆ No theorem is applied	☆ CLT is applied	☆ $\sim t \rightarrow \sim N$ as $n \rightarrow \infty$

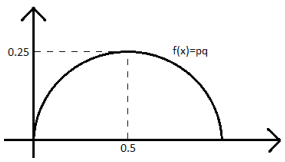
Confidence interval for population proportion

➲ $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

☆ Confidence interval attains maximum when $\hat{p} = 0.5$

☆ Assume $np \geq 5 \cap nq \geq 5 \cap n > 30$

☆ Assume asymptotic distribution where CLT is always applied



Confidence interval for population variance

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

☆ Two values of chi-square differ

☆ Assume X~N with unknown μ and unknown σ

Confidence interval for population standard deviation

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

Margin of error is the maximum error of the estimate

➲ $E = z_{\alpha/x}\left(\frac{\sigma}{\sqrt{n}}\right)$ ➲ $E = z_{\alpha/x}\left(\frac{\hat{p}\hat{q}}{\sqrt{n}}\right)$ Inflection in margin of error = $\begin{cases}\sigma\downarrow \cup n\uparrow \cup \alpha\uparrow \Longrightarrow E\downarrow \\ \sigma\uparrow \cup n\downarrow \cup \alpha\downarrow \Longrightarrow E\uparrow\end{cases}$

☆ The width of the interval is twice the size of the margin of error

Minimum sample size needed for an interval estimate

➲ $n = \sigma^2\left(\frac{z_{\alpha/2}}{E}\right)^2$ ➲ $n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$

☆ Always round up the answer to ensuring the minimum size required

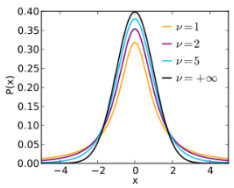
☆ Assume the sampling distribution is normally distributed

– Always use z value, do not use t value!

Probability distribution

⊙ Student's t distribution

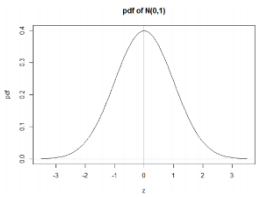
➲ $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$



↓

$n \to \infty$

↓



- Symmetric about the mean, median, and mode; these statistics are equal to 0
- The variance of t distribution is greater than 1
- The curve approaches the standard normal when sample size increase
- Based on the concept of degree of freedom (d.f.), which is related to sample size
  - Degree of freedom is subtracted by number of independent variables
  - The distribution curve changes as the sample size changes

⊙ Standard normal distribution

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

- Symmetric about the mean, median, and mode at the central of the distribution

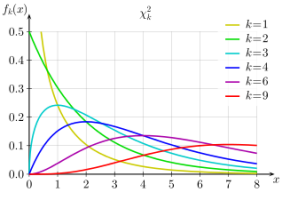
⊙ Chi-Square distribution

➲ $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

↓

$\sum Z^2$

↓



- The distribution is right skewed
  - The probability of two tails differ
- Based on the concept of degree of freedom (d.f.), which is related to sample size
  - Degree of freedom is subtracted by number of independent variables
  - The distribution curve changes as the sample size changes

☆ Assume $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

Probability under chi-squared distribution

⊙ Left-hand side/to the left probability

➲ $P(S^2 < s^2) = P\left(\chi^2 < \frac{(n-1)s^2}{\sigma^2}\right) = 1 - P(\chi^2_{n-1})$

⊙ Right-hand side/to the right probability

➲ $P(S^2 > s^2) = P\left(\chi^2 > \frac{(n-1)s^2}{\sigma^2}\right) = P(\chi^2_{n-1})$

Hypothesis testing is a method of using statistical evidence to decide whether reject or not about a hypothesis

☆ It does not give evidence about a hypothesis is true, it rejects hypothesis only

Statistical hypothesis is a conjecture about a parameter

⊙ Null hypothesis (H_0) is the expectation of a parameter

⊙ Alternative hypothesis (H_1) is a claim of conjecturing a parameter

☆ Two hypotheses must be mutually exclusive

Errors of hypothesis testing are the errors due to random samples; either one of the errors must happen

⊙ Type I error is the error of rejecting null hypothesis if it is true; or reject true null hypothesis

➲ $P(\text{type I error}) = \alpha$, where $\alpha > 0$

⊙ Type II error is the error of not rejecting null hypothesis if it is false; or not reject false null hypothesis

➲ $P(\text{tpye II error}) = \beta$, where $\beta > 0$

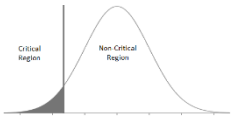
	True H_0	False H_0
Reject H_0	Type I error	Correct decision
Not reject H_0	Correct decision	Type II error

Types of hypothesis

⊙ Left-tailed test

➲ $H_0: \theta = c$

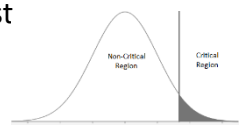
➲ $H_1: \theta < c$



⊙ Right-tailed test

➲ $H_0: \theta = c$

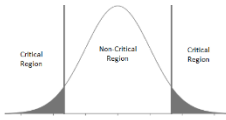
➲ $H_1: \theta > c$



⊙ Two-tailed test

➲ $H_0: \theta = c$

➲ $H_1: \theta \neq c$



Critical value is a value corresponding to the level of significance that separate critical and non-critical region

⊙ Critical/ rejection region indicates a significant difference of a parameter and a specific value

⊙ Non-critical/ non-rejection region indicates an insignificant difference of a parameter and a specific value

Decision of hypothesis testing

➢ Reject null hypothesis when the test value falls in the critical region

	Left-tailed test	Right-tailed test	Two-tailed test
Decision rule	$\theta < -\theta_\alpha$	$\theta > \theta_\alpha$	$\lvert\theta\rvert > \theta_{\alpha/2}$

➢ The power of a test measures

➲ $\text{The power of a test} = P(\text{decision rule of } \theta)$

Hypothesis testing using probability value/ p-value, which is the probability of test value

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \theta = c$ $H_1: \theta < c$	$H_0: \theta = c$ $H_1: \theta > c$	$H_0: \theta = c$ $H_1: \theta \neq c$
P-value	$P(\theta < \text{test value})$	$P(\theta > \text{test value})$	$2P(\theta > \lvert\text{test value}\rvert)$
Level of significance	α		
Decision	Reject H_0 if p-value $< \alpha$		

z test for a mean with unknown μ and known σ

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

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \mu = c$ $H_1: \mu < c$	$H_0: \mu = c$ $H_1: \mu > c$	$H_0: \mu = c$ $H_1: \mu \neq c$
Test value	$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$		
Critical value(s)	$-z_\alpha$	z_α	$\pm z_{\alpha/2}$
Decision	Reject H_0 if $z < -z_\alpha$	Reject H_0 if $z > z_\alpha$	Reject H_0 if $\lvert z\rvert > z_{\alpha/2}$

☆ Assume sample data is independent, identical distributed to normal; if not, require $n \geq 30$ to apply CLT

t test for a mean with unknown μ and unknown σ

- $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \mu = c$ $H_1: \mu < c$	$H_0: \mu = c$ $H_1: \mu > c$	$H_0: \mu = c$ $H_1: \mu \neq c$
Test value	$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$		
Degree of freedom	$n-1$		
Critical value(s)	$-t_{\alpha;(n-1)}$	$t_{\alpha;(n-1)}$	$\pm t_{\alpha/2;(n-1)}$
Decision	Reject H_0 if $t < -t_{\alpha;(n-1)}$	Reject H_0 if $t > -t_{\alpha;(n-1)}$	Reject H_0 if $\lvert t\rvert > t_{\alpha/2;(n-1)}$

☆ Assume $\bar{X} \sim_{iid} N$; if not, require $n \geq 30$ to apply CLT

z test for a proportion with unknown p

- $\frac{\hat{p}-p}{\sqrt{pq/n}} \sim N(0,1)$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: p = c$ $H_1: p < c$	$H_0: p = c$ $H_1: p > c$	$H_0: p = c$ $H_1: p \neq c$
Test value	$z = \frac{\hat{p}-p}{\sqrt{pq/n}}$		
Critical value(s)	$-z_\alpha$	z_α	$\pm z_{\alpha/2}$
Decision	Reject H_0 if $z < -z_\alpha$	Reject H_0 if $z > z_\alpha$	Reject H_0 if $\lvert z\rvert > z_{\alpha/2}$

☆ Assume $X \sim$ Bernoulli and $X_1, \ldots, X_n \sim$ Bio

☆ Require $np \geq 5$ and $nq \geq 5$ to apply normal approximation from binomial distribution, such that $\bar{X} \sim N$

– If the requirement is not fulfilled, then use p-value method instead of traditional method

χ^2 test for a variance with unknown μ and unknown σ

➢ $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{\alpha/2;(n-1)}$

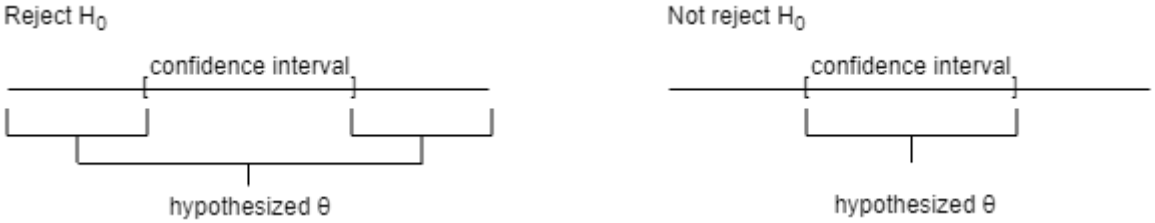
	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \sigma^2 = c$ $H_1: \sigma^2 < c$	$H_0: \sigma^2 = c$ $H_1: \sigma^2 > c$	$H_0: \sigma^2 = c$ $H_1: \sigma^2 \neq c$
Test value	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$		
Degree of freedom	$n-1$		
Critical value(s)	$\chi^2_{1-\alpha;(n-1)}$	$\chi^2_{\alpha;(n-1)}$	$\chi^2_{1-\alpha/2;(n-1)} \cap \chi^2_{\alpha/2;(n-1)}$
Decision	Reject H_0 if $\chi^2 < \chi^2_{1-\alpha;(n-1)}$	Reject H_0 if $\chi^2 > \chi^2_{\alpha;(n-1)}$	Reject H_0 if χ^2 is outside $(\chi^2_{1-\alpha;(n-1)}, \chi^2_{\alpha;(n-1)})$

☆ Assume $\overline{X} \sim_{iid} N$; if not, require $n \geq 30$ to apply CLT

Hypothesis testing using confidence interval

⊙ Reject null hypothesis if the confidence interval does not contain the hypothesized θ

⊙ Not reject null hypothesis if the confidence interval contains the hypothesized θ



Interpretation for traditional method

Since test value (test value) is greater than/ less than/ within/ outside critical value(s) (critical value(s)), it is in critical region/ non-critical region and we will/ will not reject H_0 at α level of significance. We do/ do not have evidence to conclude that topic of θ is less than/ greater than/ not c unit.

Interpretation for p-value method

Since p-value (p-value) is greater than /less than α, we will/ will not reject H_0 at α level of significance. We do/ do not have evidence to conclude that topic of θ is less than/ greater than/ not c unit.

Interpretation for confidence interval method

Since hypothesized θ is within/ outside confidence interval, we will/ will not reject H_0 at α level of significance. We do/ do not have evidence to conclude that topic of θ is less than/ greater than/ not c unit.

Hypothesis testing for the difference of parameters from two identical (independent) populations

➢ $\overline{X}-\overline{Y}\sim N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}\right)$

➲ Normality assumption is required in the two populations
  – If the populations are unknown distributed, it requires the sample to be large in order to apply CLT

➲ Sample data are identical, independent and distributed to normal

Hypothesis testing for the difference between two means

z test for the difference between two means with unknown μ_X,μ_Y and known σ_X,σ_Y

➢ $\frac{(\bar{x}-\bar{y})-c}{\sqrt{\sigma_X^2/n_X+\sigma_Y^2/n_Y}}\sim N(0,1)$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y<c$	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y>c$	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y\neq c$
Test value	$z=\frac{(\bar{x}-\bar{y})-c}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$		
Critical value(s)	$-z_\alpha$	z_α	$\pm z_{\alpha/2}$
Decision	Reject H_0 if $z<-z_\alpha$	Reject H_0 if $z>z_\alpha$	Reject H_0 if $\lvert z\rvert>z_{\alpha/2}$

☆ Assume $\overline{X}\sim_{iid}N$ and $\overline{Y}\sim_{iid}N$; if not, require $n\geq 30$ to apply CLT

t test for the difference between two means with unknown μ_X,μ_Y and unknown σ_X,σ_Y

➢ $\frac{(\bar{x}-\bar{y})-c}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}}\sim t_{(n-1)}$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y<c$	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y>c$	$H_0:\mu_X-\mu_Y=c$ $H_1:\mu_X-\mu_Y\neq c$
Test value	$t=\frac{(\bar{x}-\bar{y})-c}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}}$		
Degree of freedom	The smallest value of either n_X-1 or n_Y-1		
Critical value(s)	$-t_{\alpha;(n-1)}$	$t_{\alpha;(n-1)}$	$\pm t_{\alpha/2;(n-1)}$
Decision	Reject H_0 if $t<-t_{\alpha;(n-1)}$	Reject H_0 if $t>-t_{\alpha;(n-1)}$	Reject H_0 if $\lvert t\rvert>t_{\alpha/2;(n-1)}$

☆ Assume $\overline{X}\sim_{iid}N$ and $\overline{Y}\sim_{iid}N$; if not, require $n\geq 30$ to apply CLT

☆ Assume unequal variance where $\sigma_X\neq\sigma_Y$; to ensure the overall level of significance by not using F test

Hypothesis testing for the difference between two proportions

z test for the difference between two proportions with unknown p_X, p_Y and unknown p_Xq_X, p_Yq_Y

➢ $\frac{(\hat{p}_X-\hat{p}_Y)-c}{\sqrt{\frac{p_Xq_X}{n_X}+\frac{p_Yq_Y}{n_Y}}} \sim N(0,1)$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: p_X - p_Y = c$ $H_1: p_X - p_Y < c$	$H_0: p_X - p_Y = c$ $H_1: p_X - p_Y > c$	$H_0: p_X - p_Y = c$ $H_1: p_X - p_Y \neq c$
Test value	$z = \frac{(\hat{p}_X - \hat{p}_Y) - c}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$, where $\bar{p} = \frac{X+Y}{n_X+n_Y}$ and $\bar{q} = 1-\bar{p}$ and $\hat{p}_X = \frac{X}{n_X}$ and $\hat{p}_Y = \frac{Y}{n_Y}$		
Critical value(s)	$-z_\alpha$	z_α	$\pm z_{\alpha/2}$
Decision	Reject H_0 if $z < -z_\alpha$	Reject H_0 if $z > z_\alpha$	Reject H_0 if $\lvert z\rvert > z_{\alpha/2}$

☆ Assume $X \sim$ Bernoulli, $Y \sim$ Bernoulli and $X_1, \dots, X_n \sim$ Bio, $Y_1, \dots, Y_n \sim$ Bio

☆ Require $n_X\bar{p} > 5$, $n_X\bar{q} > 5$ and $n_Y\bar{p}$, $n_Y\bar{q}$ to apply normal approximation, such that $\bar{X} \sim N$ and $\bar{Y} \sim N$

☆ Assume equal variance where $p_Xq_X = p_Yq_Y$; it is based on the claim of $p_X = p_Y$

Hypothesis testing for the ratio between two variances

F test for the ratio of two variances with unknown μ_X, μ_Y and unknown σ_X, σ_Y (unequal parameters)

➢ $\frac{s_X^2}{s_Y^2} \sim F_{n_X-1, n_Y-1}$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1$ $H_1: \frac{\sigma_X^2}{\sigma_Y^2} < c$	$H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1$ $H_1: \frac{\sigma_X^2}{\sigma_Y^2} > c$	$H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1$ $H_1: \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$
Test value	$F = \frac{s_X^2}{s_Y^2}$		
Degrees of freedom	$n_X - 1, n_Y - 1$		
Critical value(s)	$\frac{1}{F_{\alpha;(n_Y-1,n_X-1)}}$	$F_{\alpha;(n_X-1,n_Y-1)}$	$\frac{1}{F_{\alpha/2;(n_Y-1,n_X-1)}} \cap F_{\alpha/2;(n_X-1,n_Y-1)}$
Decision	Reject H_0 if $F < \frac{1}{F_{\alpha;(n_Y-1,n_X-1)}}$	Reject H_0 if $F > F_{\alpha;(n_X-1,n_Y-1)}$	Reject H_0 if F is outside $\left(\frac{1}{F_{\alpha;(n_Y-1,n_X-1)}}, F_{\alpha;(n_X-1,n_Y-1)}\right)$

☆ Assume $\bar{X} \sim_{iid} N$ and $\bar{Y} \sim_{iid} N$; if not, require $n \geq 30$ to apply CLT

☆ Assume X has a larger variance and Y has a smaller variance

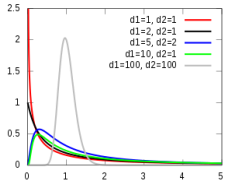
– Degrees of freedom: $n_X - 1$ be the numerator and $n_Y - 1$ be the denominator FOR F_{right} VALUE

Probability distribution

⊙ F distribution

➲ $\frac{\chi_k^2/k}{\chi_m^2/m} \sim F_{k;m}$

- F value cannot be negative (variance always positive) and $\mu \approx 1$
- The distribution is right skewed
- Based on degrees of freedom



Hypothesis testing for the difference of parameters from one population with dependent variable

➢ $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ There are two types of data which come from the same group of samples

➲ Normality assumption is required in the population
  – If the population is unknown distributed, it requires the sample to be large in order to apply CLT

➲ Sample data are identical, independent and distributed to normal

Hypothesis testing for the difference between two means

t test for the difference between two means with unknown μ_D and unknown σ_D

➢ $\frac{\overline{D}-c}{s_D/\sqrt{n}} \sim t_{(n-1)}$

	Left-tailed test	Right-tailed test	Two-tailed test
Hypothesis	$H_0: \mu_D = c$ $H_1: \mu_D < c$	$H_0: \mu_D = c$ $H_1: \mu_D > c$	$H_0: \mu_D = c$ $H_1: \mu_D \neq c$
Test value	$t = \frac{\overline{D} - c}{s_D/\sqrt{n}}$ where $\overline{D} = \frac{\sum D}{n} = \frac{\sum(x_i - x_j)}{n}$ and $s_D = \sqrt{\frac{n\sum D^2 - (\sum D)^2}{n(n-1)}}$		
Degree of freedom	$n - 1$		
Critical value(s)	$-t_{\alpha;(n-1)}$	$t_{\alpha;(n-1)}$	$\pm t_{\alpha/2;(n-1)}$
Decision	Reject H_0 if $t < -t_{\alpha;(n-1)}$	Reject H_0 if $t > -t_{\alpha;(n-1)}$	Reject H_0 if $\lvert t\rvert > t_{\alpha/2;(n-1)}$

☆ Assume $\overline{X} \sim_{iid} N$; if not, require $n \geq 30$ to apply CLT

☆ x_i and x_j are dependent variable which from the same population

Hypothesis testing for the difference parameters using confidence interval

⊙ Confidence interval for means with known σ

  ➲ $(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}$

⊙ Confidence interval for means with unknown σ

  ➲ $(\bar{x} - \bar{y}) \pm t_{\alpha/2;(n-1)}\sqrt{\frac{s^2}{n_X} + \frac{s^2}{n_Y}}$

⊙ Confidence interval for proportions with no equal variance assumption (not suitable for hypothesis test)

  ➲ $(\hat{p}_X - \hat{p}_Y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X\hat{q}_X}{n_X} + \frac{\hat{p}_Y\hat{q}_Y}{n_Y}}$

⊙ Confidence interval for means with dependent variable

  ➲ $\overline{D} \pm t_{\alpha/2;(n-1)}\frac{s_D}{\sqrt{n}}$

	Left-tailed test	Right-tailed test	Two-tailed test
Decision	Reject H_0 if $UL < 0$	Reject H_0 if $LL > 0$	Reject H_0 if 0 is outside CI