Introduction

Statistics for Psychology II

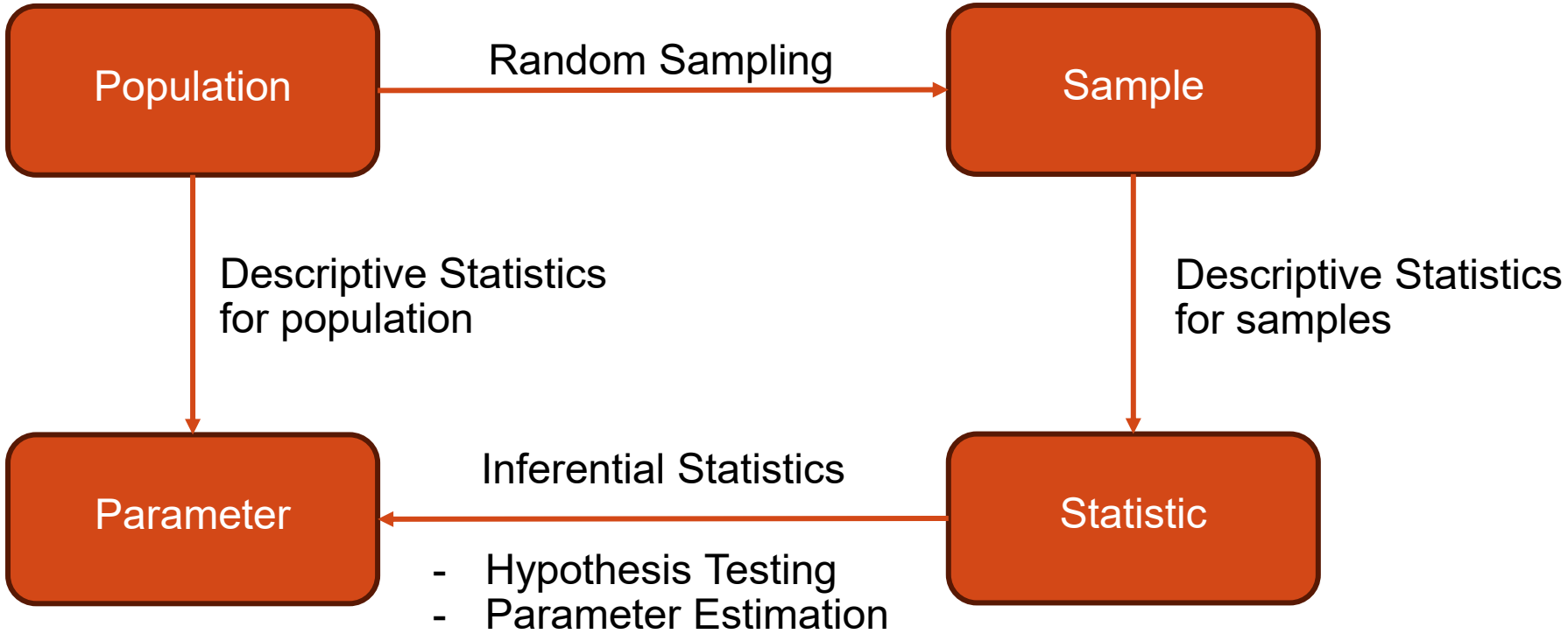
Sunthud Pornprasertmanit

Outline

- Stat I review
- Bootstrapping
- Missing data handling

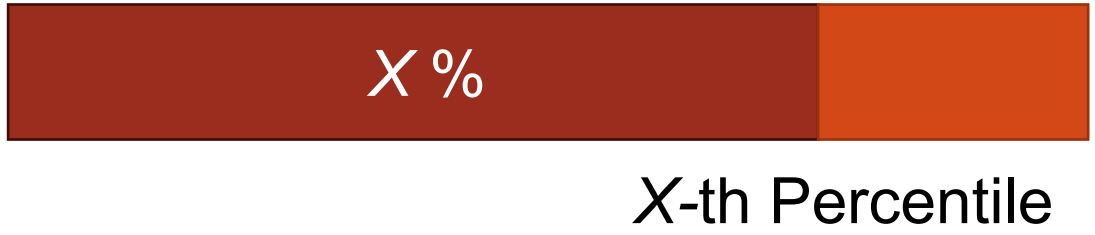
Basic Concepts

- **Descriptive statistics**: Used to identify patterns or summarize the characteristics of a data set (either a sample or the whole population).
- **Inferential statistics**: Used to estimate or make generalizations about the patterns of a larger population based on a smaller sample.


Population
Random Sampling
Sample
Descriptive Statistics
for population
Descriptive Statistics
for samples
Parameter
Inferential Statistics
Statistic
- Hypothesis Testing
- Parameter Estimation

Central Tendency

- **Central tendency** refers to measures that identify the middle point of a data set.
  - **Mean**: The sum of all values in a variable divided by the number of sample
  - **Median**: The value that lies in the middle when all values in the data set are arranged in an ascending (or descending) order.
  - **Mode**: The value that appears the most frequently in the data set.
- **Percentile**



Dispersion

- **Variability** measures the extent to which values in a variable differ or spread out.
  - **Range** = Maximum – Minimum
  - **Interquartile Range (IQR)** = 3rd quartile – 1st quartile = 75th percentile – 25th percentile
  - **Variance**: A measure of how far each value is from the mean (average squared deviation).
  - **Standard deviation**: The square root of variance; represents the average distance from the mean.

Making sense of a score

- A **raw score** is the value obtained directly from a subject.
- On its own, a raw score does not indicate whether it is high or low unless there is a reference for comparison.
- **Criterion referenced**
  - Compares raw scores to a standard established by reliable sources.
  - Example: a TOEFL score of 80 or above is considered adequate for English language proficiency.
- **Norm referenced**
  - Compares a raw score with the raw scores from a target group.

Standard Score or *Z*-score

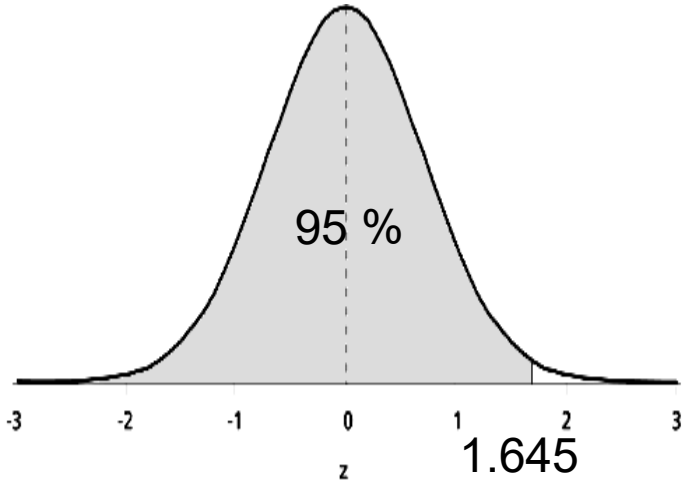
- The formula for calculating a ***z*-score** is

$$z_i = \frac{X_i - M}{SD}$$

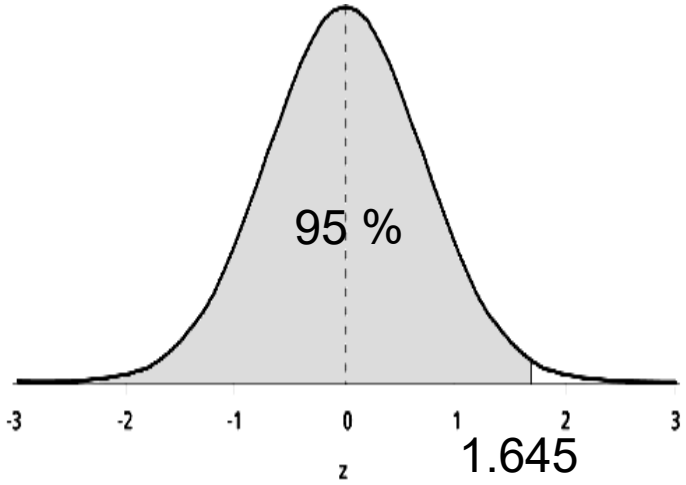
- where X_i: Raw score of variable *X* for the *i*-th subject; z_i: Standard score (*z*-score); *M*: mean and *SD:* standard deviation.
- The *z*-score indicates how far the *i*-th subject's value deviates from the mean, expressed as a multiple of the standard deviation.
- If the data is normally distributed, each *z*-score corresponds with a percentile, which indicates the proportion of subjects with raw scores less than X_i.

Standard Score or *Z*-score

- Standard score -> Percentile
- R: `pnorm(1.645)`
- Excel: `=NORMSDIST(1.645)`



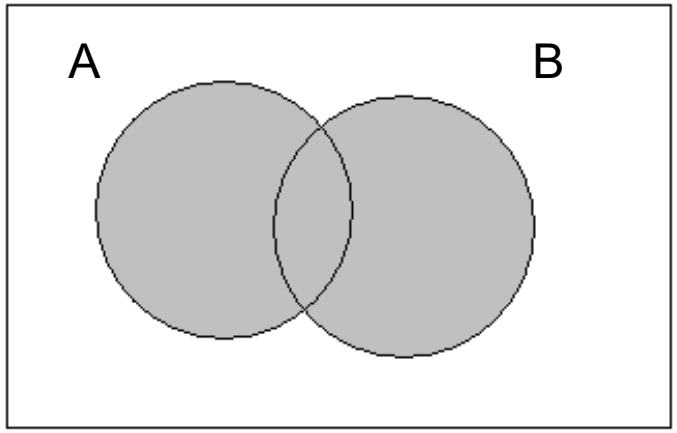
- Percentile -> Standard score
- R: `qnorm(0.95)`
- Excel: `=NORMSINV(0.95)`



Probability

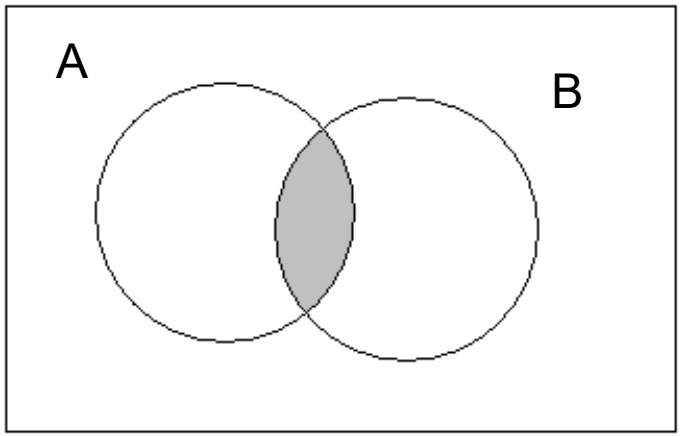
- **Probability** is the likelihood of a target event occurring.
- $p(A)$ represents the probability of event A occurring.

$p(A \cup B)$



Union

$p(A \cap B)$

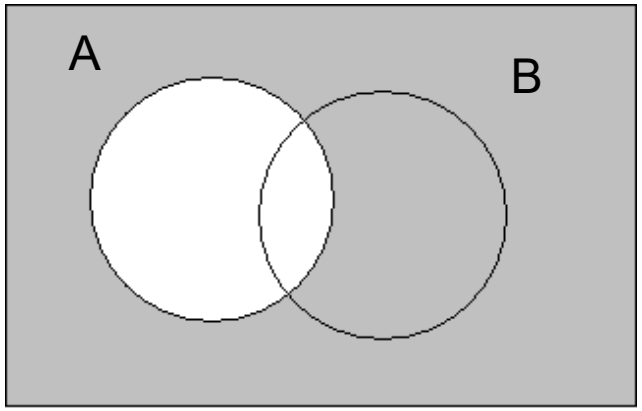


Intersection

Probability

- Probability is the likelihood of a target event occurring.

$$p(A') = 1 - p(A)$$



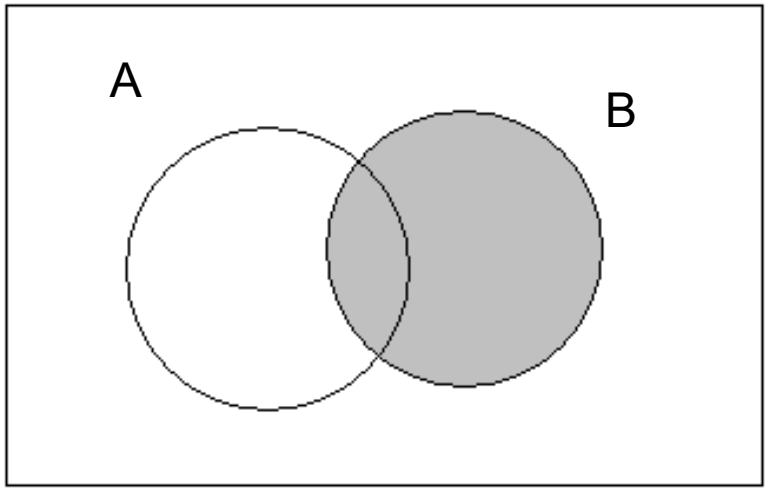
Complement

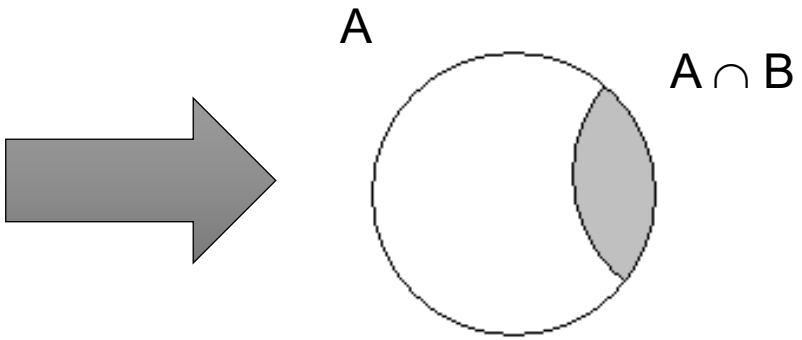
Probability

- **Conditional probability**

$$p(B) = n(B)/n(S)$$

$$p(B|A) = n(A \cap B)/n(A)$$





$$p(B|A) = \frac{n(B \cap A)/n(S)}{n(A)/n(S)} = \frac{p(B \cap A)}{p(A)}$$

Probability

- **Statistical independence** means the probability of event A occurring is not influenced by the outcome of event B.

$$p(A) = p(A|B) = p(A|B')$$

$$p(B) = p(B|A) = p(B|A')$$

Hypothesis Testing

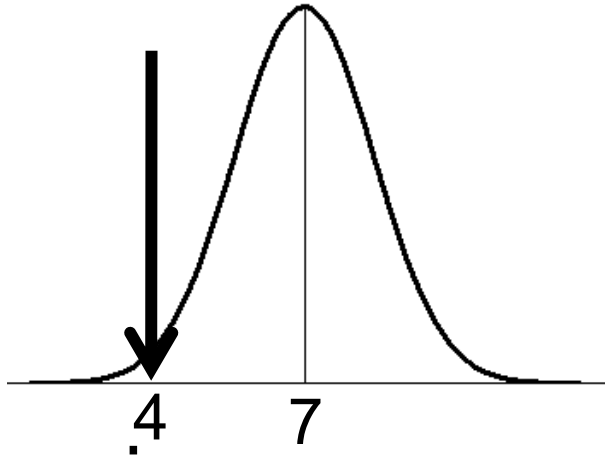
- Check the likelihood of whether the parameters of a population align with expectations. Here is an example.
  - A student's grade is poor, and a teacher hypothesizes that this student might have a learning disability. The student is given a memory test and remembers only four chunks.
  - In the population of this age group, the memory test scores follow a normal distribution with $\mu = 7$ and $\sigma = 1.5$.
  - Research question: Does this student belong to the normal population?

H_0: $\mu = 7$ (The students belongs to the normal population.)

H_1: $\mu < 7$ (The students doesn't belong to the normal population. Additionally, the population mean is less than 7, but we don't know the population mean.)

Hypothesis Testing

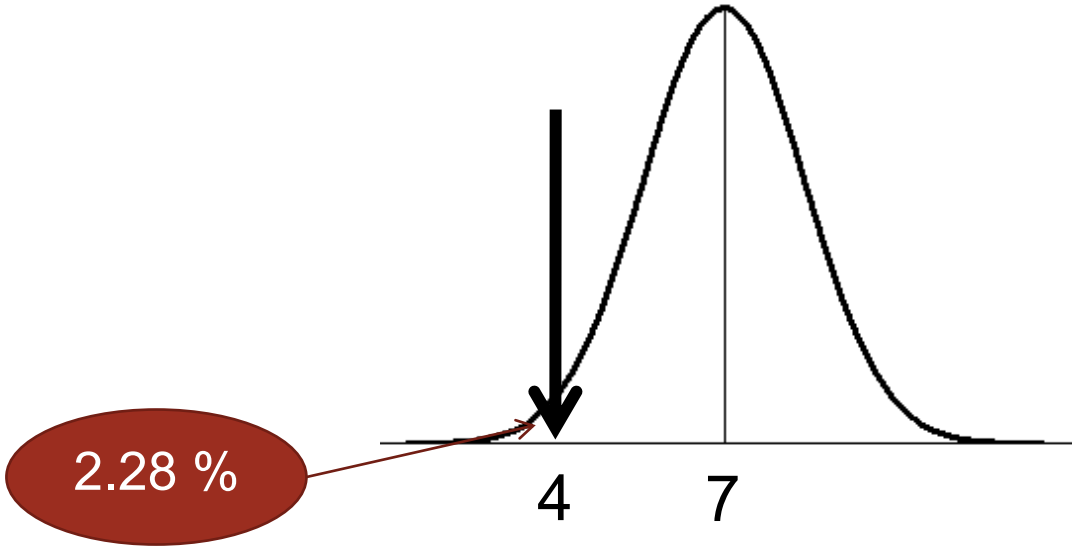
- The null hypothesis assumes the following information
  - $N(7, 1.5)$
  - The student is a sample of the population



- Under the null hypothesis, we can calculate the probability of finding students who remember 4 or fewer chunks.

Hypothesis Testing

- The probability of finding students who remember 4 or fewer chunks can be calculated as follows:
  - Transform the raw score to a *z*-score: (4-7)/1.5 = -2
  - Find the probability for the z-score from -∞ to -2 = 2.28%



Hypothesis Testing

- If we use the criterion that an event is considered rare and unlikely if its probability is less than 5%, we would conclude that this student is unlikely to be normal.
- Since 2.28% < 5%, we decide that this student may have delayed memory development.
- Thus, we reject the null hypothesis (H_0) and conclude that the alternative hypothesis is more likely to be true

H_0: μ = [illegible] ongs to th[illegible])

H_1: μ < 7 (The students doesn't belong to the normal population. Additionally, the population mean is less than 7, but we don't know the population mean.)

Directionality

- Research hypotheses can be formulated in two ways:
  - **One-tailed test**
    - Specifies a particular direction for the hypothesis.
    - Example: A student has delayed memory.
  - **Two-tailed test**
    - Does not specify direction, allowing for either extreme.
    - Example: A student has abnormal development (either too fast or too slow).

# of directions	Null Hypothesis	Alternative Hypothesis
One-tailed	H_0: $\mu = 7$	H_1: $\mu < 7$
Two-tailed	H_0: $\mu = 7$	H_1: $\mu \neq 7$

Significance Level

- The **significance level** is the criterion used to determine when a null hypothesis is unlikely to be true. The commonly used significance level is 0.05.
  - It is often referred to as the **alpha level**.
  - The significance can be adjusted, such as from 5% to 1%, 10%, or other levels.
  - Commonly used significance level are 0.05, 0.01, and 0.001.

Decision Making

- When the null hypothesis is true, the probability of observing the event is called the ***p*-value**.
- If the *p*-value is less than the predesignated significance level, the null hypothesis is rejected.
- The raw score or *z*-score corresponding to the significance level is called the **critical value**.
- The range of values indicating the null hypothesis is unlikely to be true is called the **critical region.**
- If the obtained raw or *z*-score falls within the critical region, the null hypothesis is rejected

Types of Errors in Hypothesis Testing

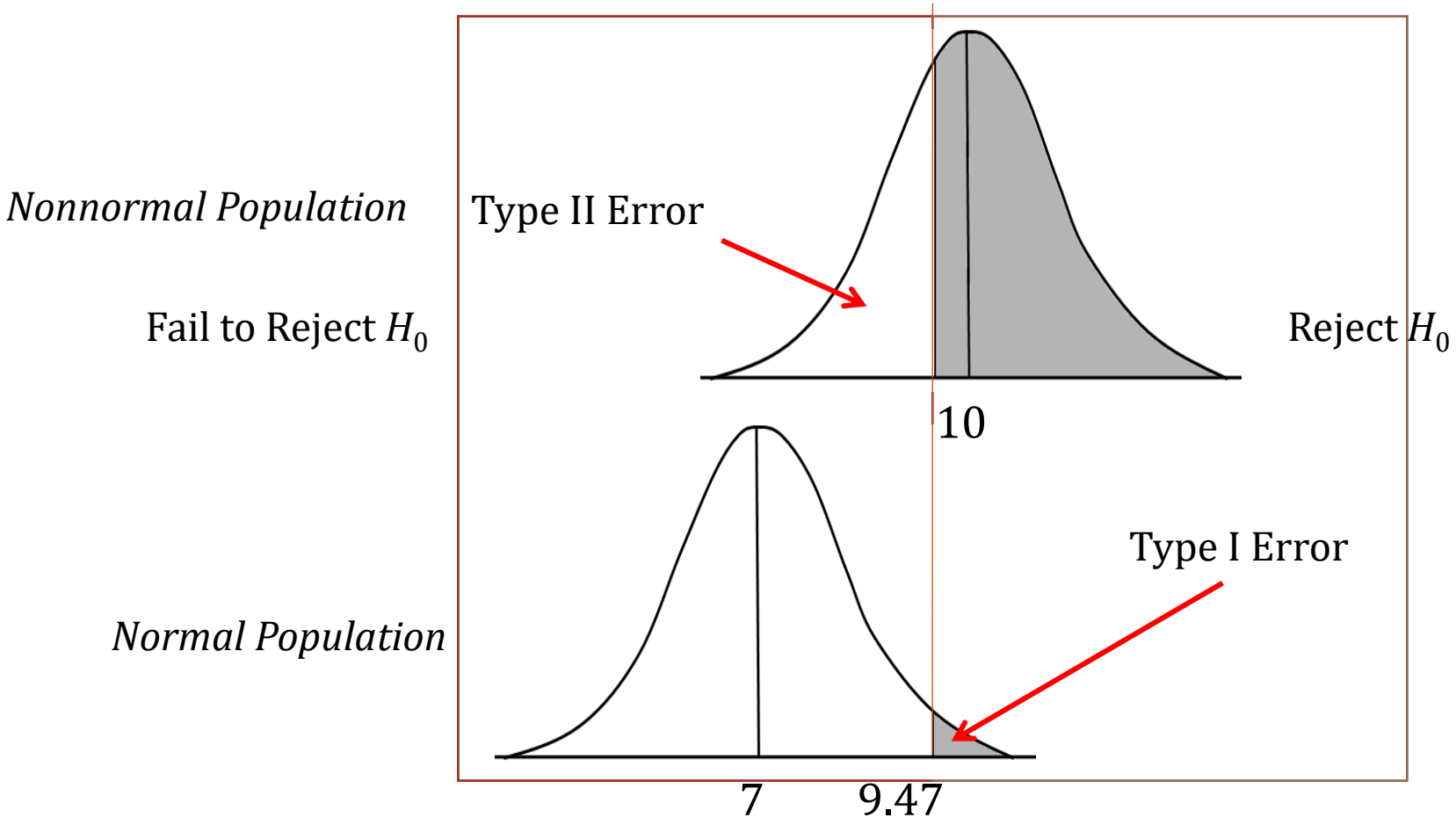
Decision

Truth		Fail to reject H_0	Reject H_0
	H_0 true	✓	Type I error
	H_A true	Type II Error	✓

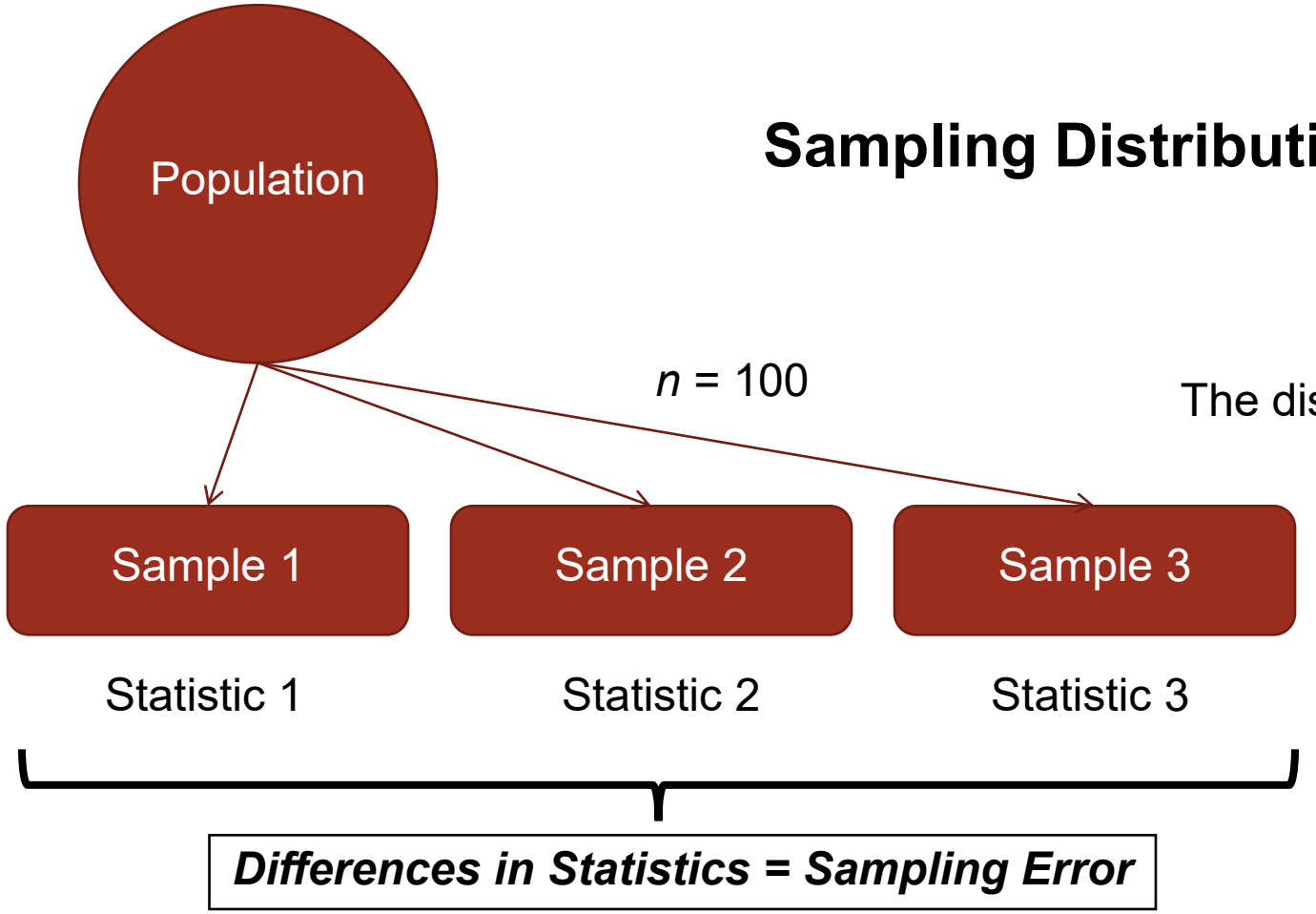
Types of Errors in Hypothesis Testing

- If the obtained value is not in the critical region or the *p*-value is not below the significance level, it does mean the null hypothesis is false. Instead:
  - The null hypothesis may still be true, or
  - You may have made a Type II error (failing to reject a false null hypothesis).

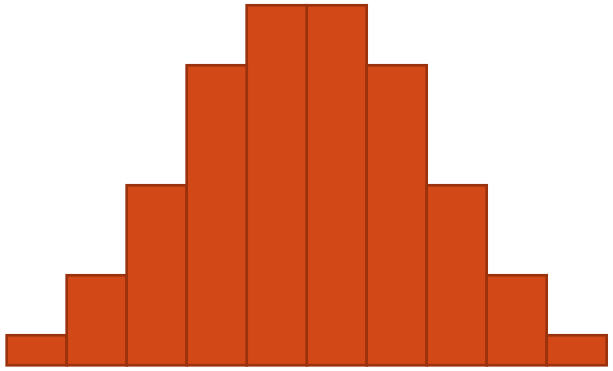
Types of Errors in Hypothesis Testing



Variability from Sampling



Sampling Distribution



The distribution of statistics obtained from repeated samples

. . .

Variability from Sampling

- Let's consider the sampling distribution of means.
- The **Central Limit Theorem** states
  - $\mu_M = \mu$; The mean of the sampling distribution equals the population mean.
  - $\sigma_M^2 = \frac{\sigma^2}{n}$; The variance of the sampling distribution equals the population variance divided by the sample size.

Variability from Sampling

- The sampling distribution of means will approximate a normal distribution if either:
  - The population distribution is normal
  - The sample size is sufficiently large, regardless of the shape of the population distribution (e.g., $n > 30$.)

z test

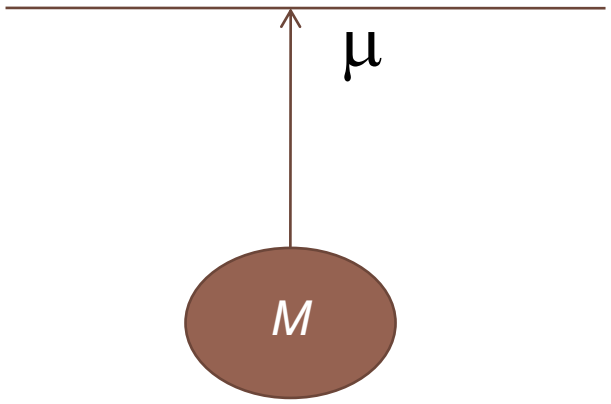
- The ***z*-test** determines whether the sample mean obtained from a sample is likely to come from the target population.
- That involves comparing the sample mean with the sampling distribution of means derived from the target population under the assumption that the null hypothesis is true.
- The formula is

$$z = \frac{M - \mu_M}{\sigma_M} = \frac{M - \mu}{\sigma / N}$$

Parameter Estimation

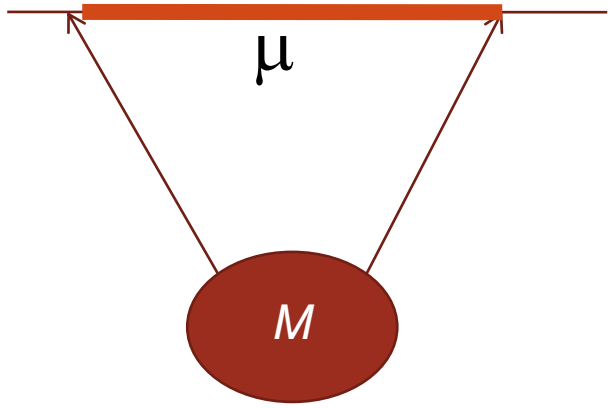
- Two types of parameter estimation:

Point Estimation



Based on the statistic, which parameter value is the most likely to occur?

Interval Estimation



Based on the statistic, what is the range of parameter values with the highest probability?

Point Estimation

- The statistic used to estimate a parameter should have two characteristics:
  1. Precision
     - The mean of the statistic from all possible samples should be equal to the target parameter.
     - $\mu_{\text{Statistic}} = \text{PARAMETER}$ or $E(\text{Statistic}) = \text{PARAMETER}$
  2. Efficiency
     - The variability of the statistic from all possible samples should be as small as possible.
     - $\sigma_{\text{Statistic}}$ should be as small as possible

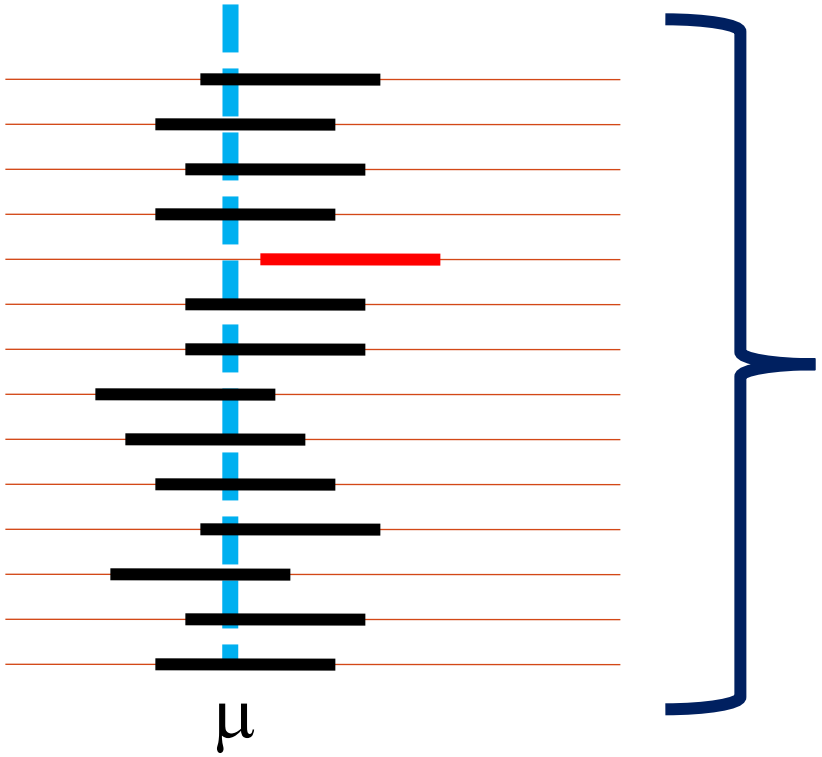
Interval Estimation

- Interval estimation is a method to determine the range that is likely to contain the parameter of interest
- This range is known as the confidence interval
- For example, the confidence interval of mean.

$$CI_{1-\alpha} = M \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Interval Estimation

- What does a 95% confidence interval mean?
- Imagine taking random samples (n = 1000) 100 times



From 100 samples, we construct 100 confidence intervals at the 95% confidence level, each based on the sample mean. Out of these, 95% of the confidence intervals (approximately 95 out of 100) are expected to contain the target parameter.

Interval Estimation

- Can we say there is a 95% probability that the parameter lies within the confidence interval?
  - No.
  - This is because the parameter is either within the confidence interval (100%) or not (0%).

Effect Size

- Statistical significance indicates that the null hypothesis is unlikely to be true. For example, it suggests that:
  - A person is not "normal" and might have depression.
- However, statistical significance does not measure the strength or magnitude of differences or correlations. For instance:
  - Stating that someone is not "normal" does not quantify how far from normal they are.
  - Suggesting that someone might have depression does not indicate the severity of the condition.
- The magnitude or strength of these differences or relationships is quantified by the **effect size**.

Effect Size

- To compare means, the effect size can be expressed as the difference between two means or as the difference in terms of standard deviations.
- The latter is referred to as the **standardized effect size** or **Cohen's *d***, which provides a standardized measure of the magnitude of the difference.

** Use standard deviation not standard Error

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

$$d = \frac{M_1 - \mu_0}{\sigma}$$

In an experiment, **Cohen's *d*** is calculated as the mean of the experimental group minus the mean of the control group, divided by the standard deviation of the control group.

Confidence Interval of Effect Size

- Use interval estimation to determine the population effect size or the confidence interval of effect size:

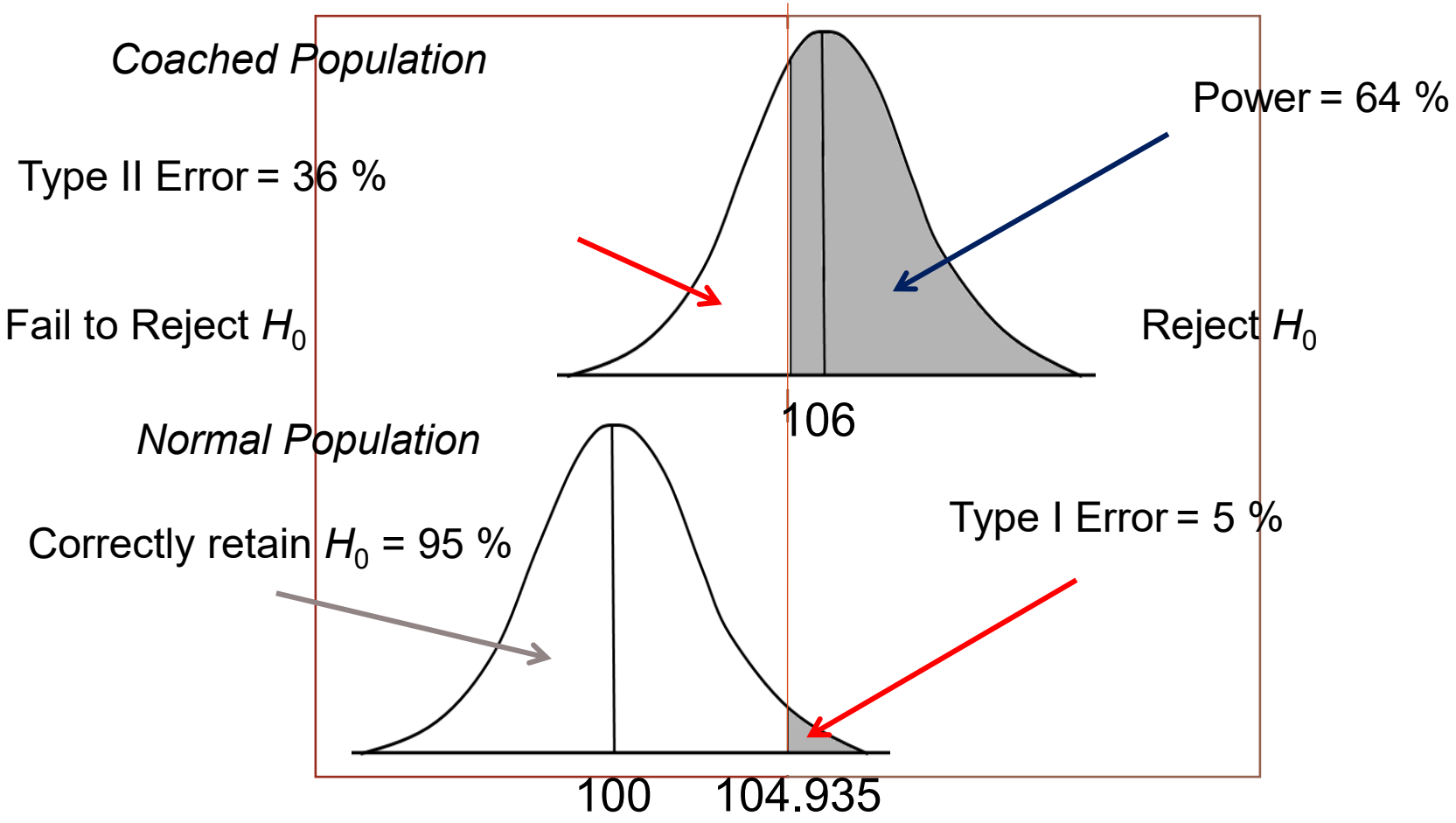
$$CI_{1-\alpha} \text{of } \delta = d \pm \frac{z_{\alpha/2}}{\sqrt{n}}$$

Statistical Power

- Although the alternative hypothesis may be true, a hypothesis test might fail to reject the null hypothesis, leading to a Type II error. Sampling error can result in obtaining a sample mean that is very close to the population mean under the null hypothesis.
- The probability of obtaining statistically significant results when the alternative hypothesis is true is referred to as **statistical power**, or simply **power**.

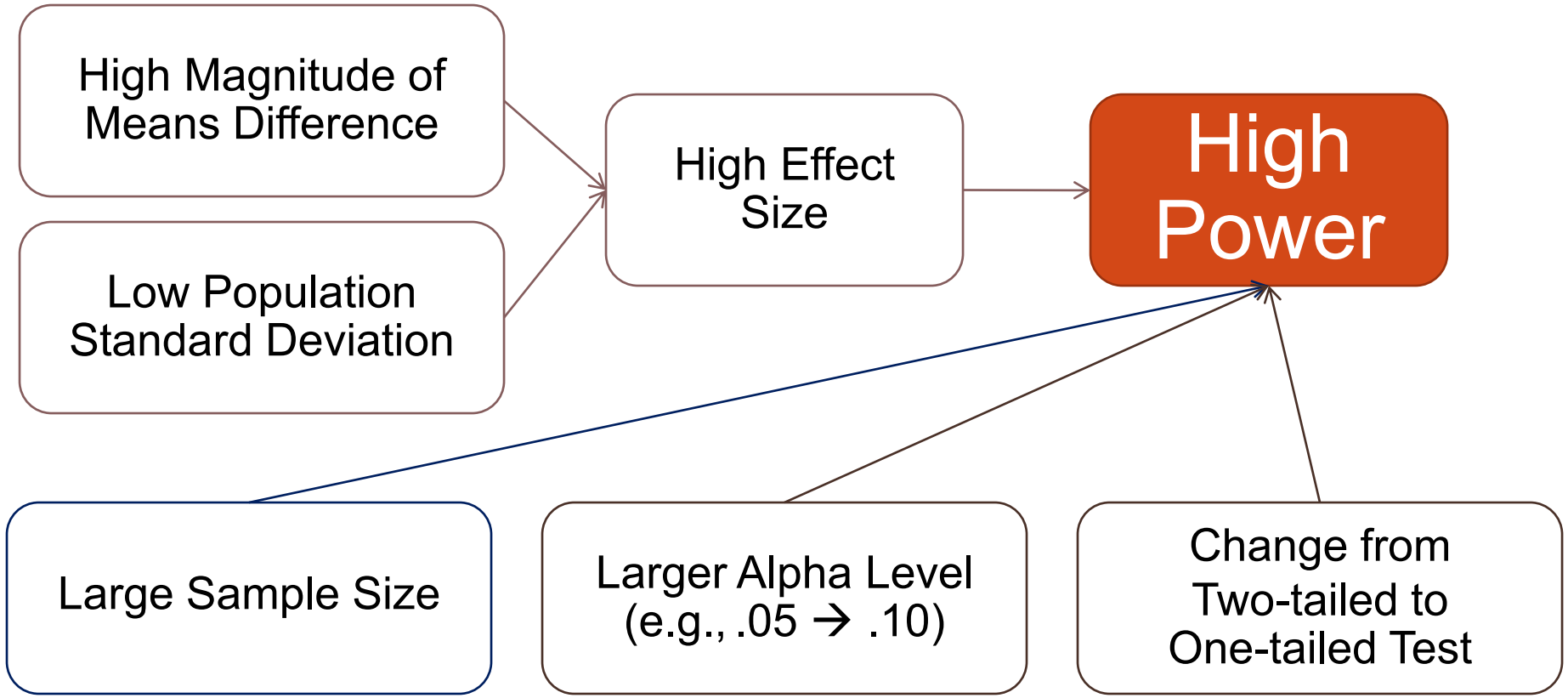
		Decision	
		Fail to reject H_0	**Reject H_0**
Truth	H_0 true	✓	Type I Error
	H_A true	Type II Error	**Power**

Statistical Power



Statistical Power

- Factors influence statistical power:



Sample Size Estimation

- Before data collection, researchers must ensure that the sample size is large enough to minimize the risk of a Type II error (i.e., to achieve sufficient statistical power). Statistical power levels of 0.8 or 0.9 are generally considered adequate.
- To determine the required sample size, researchers use the desired statistical power and the anticipated magnitude of the effect size.

Sample Size Estimation

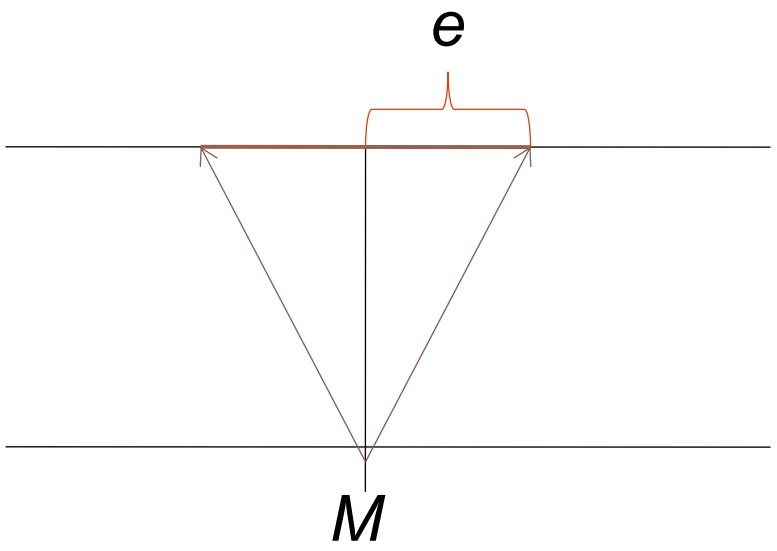
- From the z-test, the formula for calculating sample size is as follows:

$$n = \left(\frac{\sigma(z_{H1} - z_{H0})}{\mu_{H0} - \mu_{H1}} \right)^2$$

Sample Size Estimation

- Another approach is to determine the sample size required to achieve a desired level of error in interval estimation.
- For estimating the population mean, the formula to calculate the required sample size is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$



One-sample *t*-test

- The **one-sample *t*-test** is used to compare the mean of a sample to a known or hypothesized population mean.
- While the one-sample *t*-test is similar to the one-sample *z*-test, it differs in that the *t*-test uses the sample standard deviation instead of the population standard deviation, which is often unknown.
- In statistical software, you can perform a one-sample *t*-test by specifying the variable of interest and the target population mean. The software will compute the *p*-value and provide the confidence interval for the comparison.

One-sample *t*-test

- The formula for effect size in a one-sample *t*-test is similar to that of the one-sample *z*-test.

$$d = \frac{M_1 - \mu_0}{SD}$$

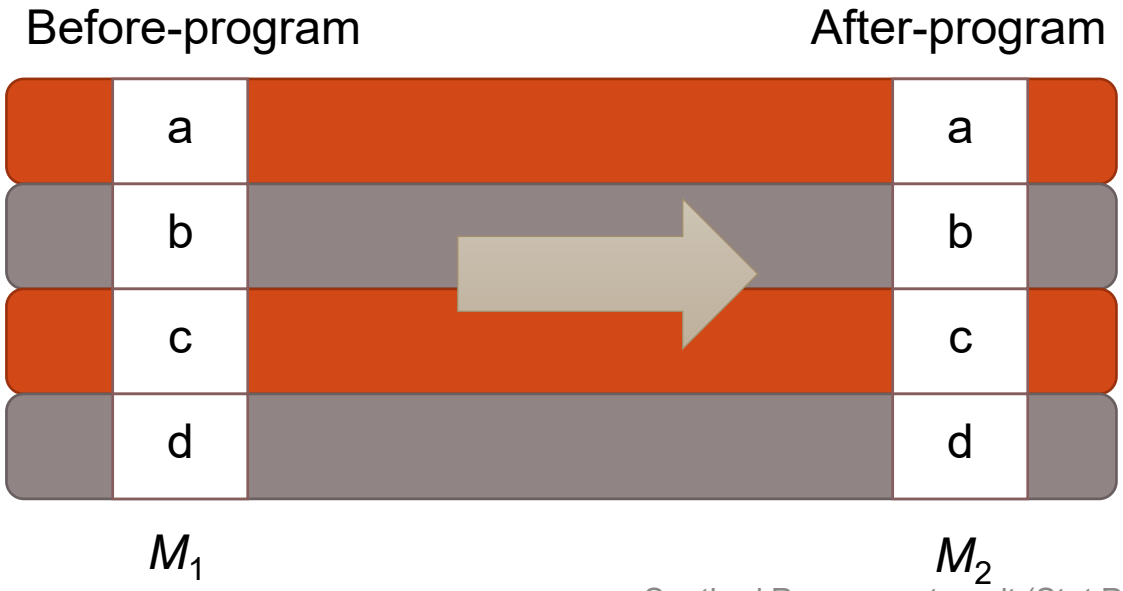
- Sample size estimation can be performed using various programs. Among the most popular is **G*Power**, while the **Jamovi** software offers a great alternative through its **Pamlj** package.

Two Means Comparison

- Researchers can compare two groups to determine whether their population means differ. While the sample means may show a difference, statistical tests are used to assess whether the parameter means are significantly different.
- There are two types of statistical tests for comparing group means:
  - **Dependent groups**: A score in one group is directly related to a specific score in the other group (e.g., paired observations). The two groups must have equal sample sizes.
  - **Independent groups**: A score in one group is unrelated to any data point in the other group. The two groups may have equal or unequal sample sizes.

Two Means Comparison

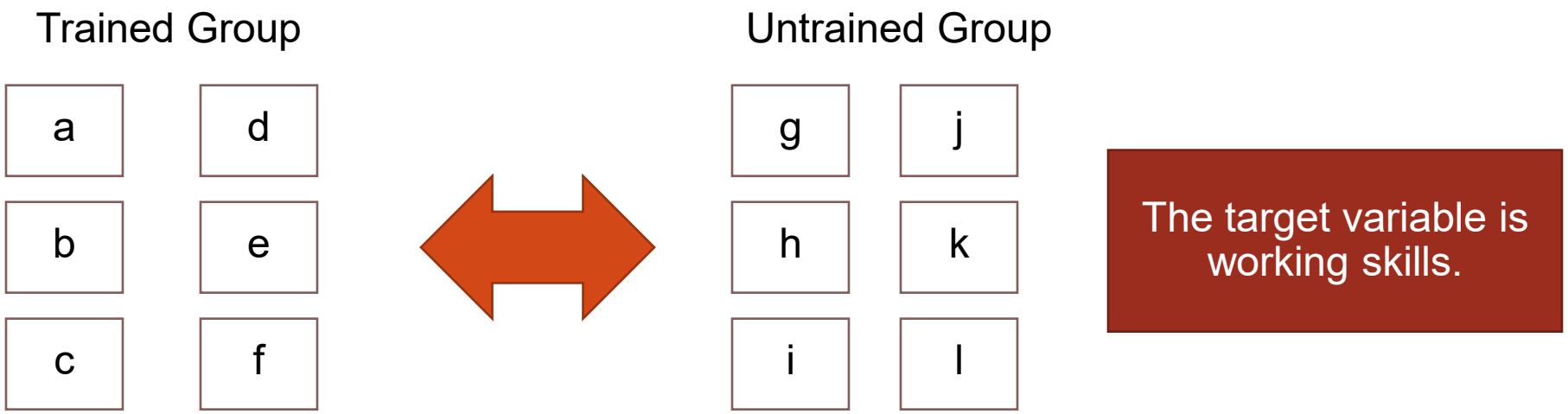
- Example: A weight comparison between before- and after-program from a sample group of subjects
- The before- and after-program weights are related because both come from the same persons.



The target variable is weights.

Two Means Comparison

- Example of Independent Groups: A comparison between trained and untrained groups on the working skills
- Individuals in the trained group cannot be coupled with any individuals in the untrained group.



Two Means Comparison

- The dependent groups are generally shown in 4 formats:
  - Repeated-measures from the same subjects
  - Matching or randomized block design
  - Twin studies
  - Natural pairs such as direct competitors, spouses, relatives
- The independent groups are generally shown in 2 formats:
  - Drawn from different populations, such as biological sex
  - (Random) assignment by researchers

Independent *t*-test

- **Independent *t*-test** is used for comparing means between two independent groups.
- Statistical programs can test and find confidence interval of the mean difference.
- The main assumption is homogeneity of variance or equal population variances between two groups.
  - If sample sizes between groups are equal, traditional independent *t*-test is robust for the assumption violation.
  - If the sample sizes are tremendously different, the robustness is compromised when the variance ratio is over 10:1. That is, the obtained *p*-value and confidence interval is not accurate.

Independent *t*-test

- Welch test is an alternative for the violation.
- However, Welch test strictly assumes population distribution to be normal.
- Bootstrapping or nonparametric tests are also alternatives.
- To use Levene test and decide to use traditional or Welch tests is inaccurate because the traditional *t*-test is generally robust in most cases.

Independent *t*-test

- Effect size or **Cohen's *d***:

$$d = \frac{M_1 - M_2}{SD_{\text{Pooled}}}$$

- Standard deviation can be either the *SD* from a control group (SD_C) or pooled *SD*:

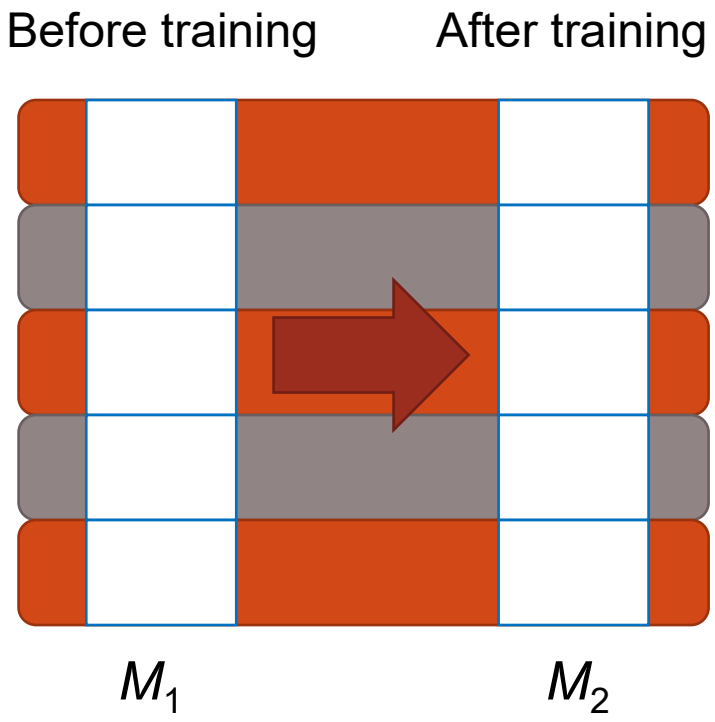
$$SD_{\text{Pooled}} = \sqrt{\frac{SS_1 + SS_2}{df_1 + df_2}}; SS_k = df_k(SD_k^2)$$

Independent *t*-test

- G*Power and `Pamlj` can run power analysis and sample size estimation.

Dependent *t*-test

- To run the test, the difference score is calculated for each pair.



d_1
d_2
d_3
d_4
d_5
M_d

Difference score (d) = After - Before
(or Before – After)

One-sample *t*-test is used to test whether M_d is statistically different from 0.

$M_d = M_2 - M_1$

Dependent *t*-test

- This procedure is referred to as the **dependent *t*-test**.
- Statistical programs can run the hypothesis test and find the confidence interval of the group difference.
- Effect size is quite complicated because three types of *SD* can be used:
  - Standard deviation of the control group (or unintervened group)
  - Pooled standard deviation from two groups
  - Standard deviation of the difference scores

Dependent *t*-test

- Control-group *SD* and pooled *SD* provide a similar effect size. They tell the magnitude of the differences in the unit of natural standard deviation.
  - Example: Weight-loss program can reduce the weight in the multiple of standard deviation of the original weights.
  - This effect size is similar to one from the independent *t*-test.
- Difference-score *SD* is hard to interpret but it is used in the power analysis program, such as G*Power and `Pamlj`.

Comparing Dependent and Independent *t*-tests

- The benefits of repeated measures:
  - Use a lower number of subjects
  - Higher statistical power, especially when two sets of scores are highly correlated
- The drawbacks of repeated measures:
  - Longer
  - Subjects dropout
  - Carry-over effect

Correlation

- **Pearson's correlation** measures the degree to which two interval variables increases or decreases simultaneously.

$$r_{XY} = \frac{s_{XY}}{SD_X SD_Y}$$

- The correlation ranges from -1 to 1.
- The sign shows the direction of the correlation.
- The absolute value of the correlation indicates the magnitude of correlation: .1 as low, .3 as medium, and .5 as high.

Correlation

- If there are more than two variables, correlations can be displayed in a correlation matrix:

	A	*B*	*C*
A	1	.50	-.30
B	.50	1	-.45
C	-.30	-.45	1

Symmetric Matrix

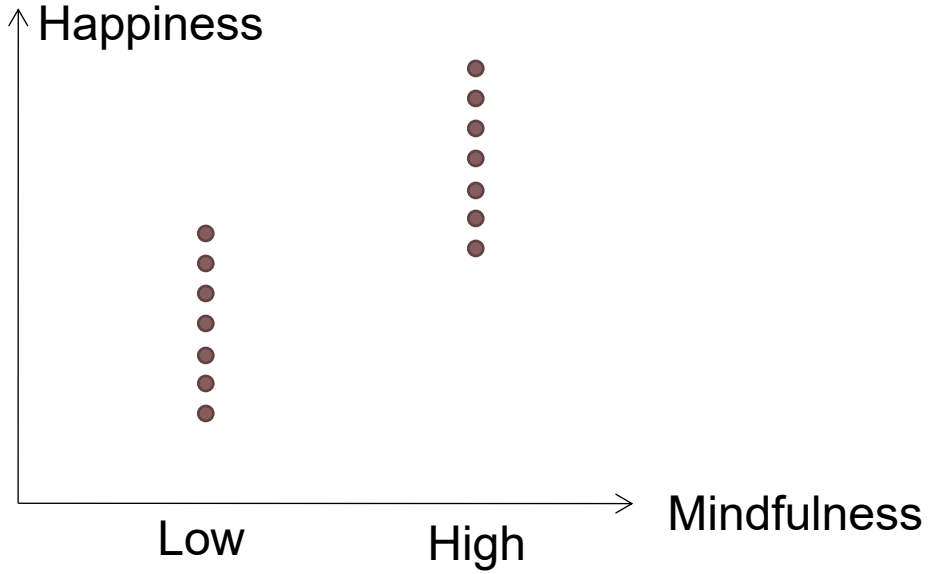
Correlation with itself equals 1

Correlation

- Generally, hypothesis testing checks whether the population correlation equals 0 or not.
- If researchers want to check whether the population correlation is a specific value rather than 0, the confidence interval of correlation will be used.
  - Usually involves with Fisher's *z* transformation or bootstrap
- Power analysis and sample size estimation can be done by G*Power and `Pamlj`.

Difference and Correlations

- By definition, correlation means when one variable changes, the other variable changes too.
- For example, mindfulness is correlated with happiness.
- The higher mindfulness, the happier the person.



Difference and Correlation

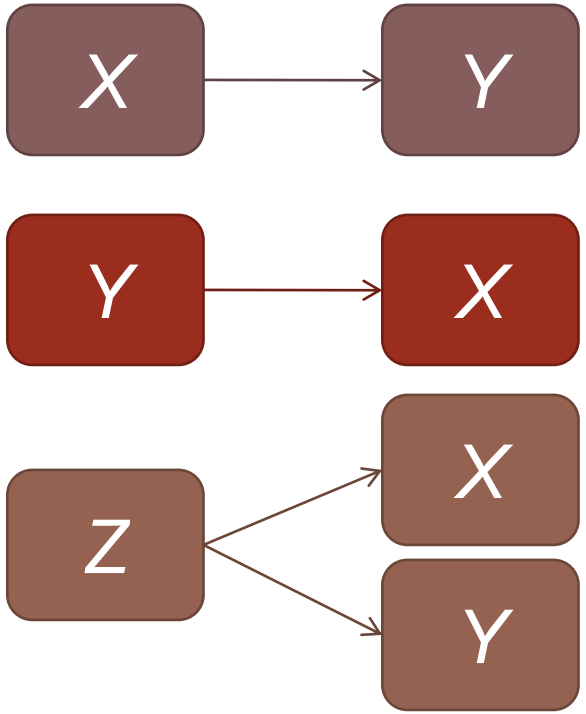
- From the picture, low- and high-mindfulness groups are different in happiness.
- But from the definition of correlation, it means that mindfulness and happiness are correlated.
- Thus, group differences in a variable mean correlation.
- E.g., ethnicity is correlated with SAT scores.

Factors Influencing Correlation

- Spurious correlation, such as the number of fire trucks is correlated with the fire damage level on properties
- Range restriction
- Outliers

Causal Relationship

- If you see differences between groups or correlation between variables, it doesn't necessarily mean one causes the other.
- If *X* is related to *Y*, there are three possible causal relationships:
  - *X* causes *Y*.
  - *Y* causes *X*.
  - Other variables cause both *X* and *Y*.



Causal Relationship

- If *X* causes *Y*, four things should be satisfied:
  - *X* is correlated to *Y*.
  - *X* precedes *Y* in time.
  - The change in *Y* must not be explained by other variables.
  - Have theories and explanations why *X* causes *Y*

Statistics for Proportions

- **Proportion** is the rate of an event occurring such as the proportion of males in this class is .45
- **Percentage** is the rate where an event occurs or proportion times 100 such as the percentage of miscarriage within two-month pregnancy after fertilization is 20%.
- **Ratio** is the comparison between two things occurring such as the stay and quit ratio of working in a company for one year is 2 to 3.
- **Odd** is the ratio where the last number is 1 such as the patients with heart problems and exercise have odds of survival in one year of 3 over 1.

Chi-square Test for Goodness-of-fit

- The **Chi-square goodness-of-fit test** is used for comparing proportions of each category in samples with proportions defined at the population level.
  - E.g., Is the market share of personal cars different from the market share in the previous year? Is the dice fair?
- Confidence interval of the proportion can be done by the binomial test.

Chi-square Goodness-of-fit Test

- The effect sizes of comparing proportions are in many forms:
  - Proportion difference, risk difference
  - Risk ratio
  - Odd ratio
  - Cohen's *w*
- Power analysis and sample size estimation can be done in G*Power and `Pamlj`.

Chi-square Goodness-of-fit Test

- One form of sample size estimation is based on the error of parameter estimation.
- This is Yamane's formula that is popular in Thai literature.
- Most psychological research does not estimate population parameters but still uses Yamane's formula.
- Actually, Yamane is the author of the book. He did not make the formula.

Chi-square Contingency Tables

- The chi-square goodness-of-fit can be adapted to compare proportions between two or more groups.
- The **chi-square contingency table** is similar to the independent *t*-test but used to compare categorical variables.
  - E.g., whether the proportion of customers who prefer Som Tum differs between dine-in and takeaway orders.
  - Are there any differences in color preference between age groups?
- Hypothesis testing can be done in most statistical programs. Confidence interval is usually implemented in combination with effect sizes.

Chi-square Contingency Table

- Effect sizes are as follows:
  - Proportion difference or risk difference
  - Risk ratio
  - Odd ratio
  - Cohen's *w*
- Power analysis and sample size estimation can be done in G*Power and `Pamlj`.

McNemar Test

- **McNemar test** is to compare proportions between two dependent groups like the dependent *t*-test.
- McNemar test can be used for only two groups and the target variable can be categorized to only two groups.

Resampling Techniques

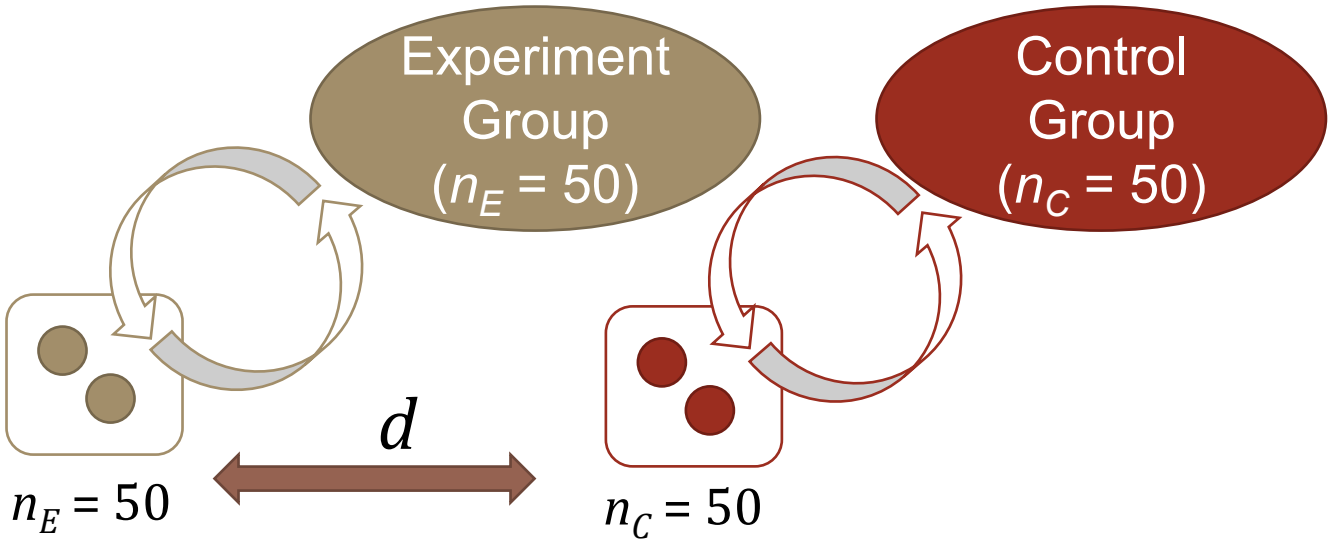
- When conducting hypothesis test or confidence interval, some assumptions must be made to ensure that the predefined reference distribution such as normal, *t*, *F*, chi-square distributions can be used. Sometimes, those assumptions cannot be assumed.
- Some statistics cannot easily find confidence interval (e.g., *r*) because the distribution does not follow the predefined distributions.
- **Bootstrapping techniques** is used to find empirical sampling distribution without using predefined distributions.

Resampling Techniques

- The basic idea of the resampling techniques is to treat the obtained sample as a pseudo-population.
- Then, use sampling with replacement with the sample size equal to the obtained sample.
- Find the desired statistics, such as means, *SD*s, mean difference.
- Repeat the two steps above a large number of times (e.g., 1000 times).
- Collect the statistics to find confidence intervals. For example, 95% bootstrap CI takes the 2.5th and 97.5th percentiles of the collected statistics as the lower and upper bounds.

Resampling Techniques

- For example, bootstrap confidence interval of Cohen's *d* from two independent mean comparison

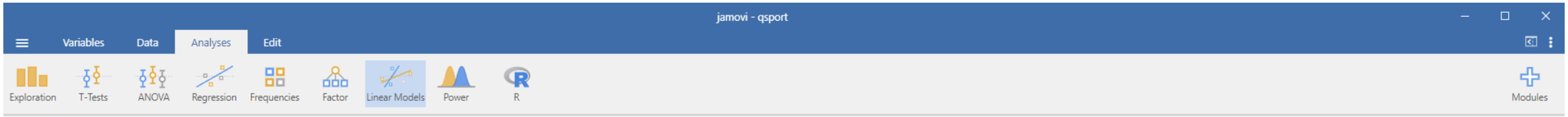


Sampling with Replacement

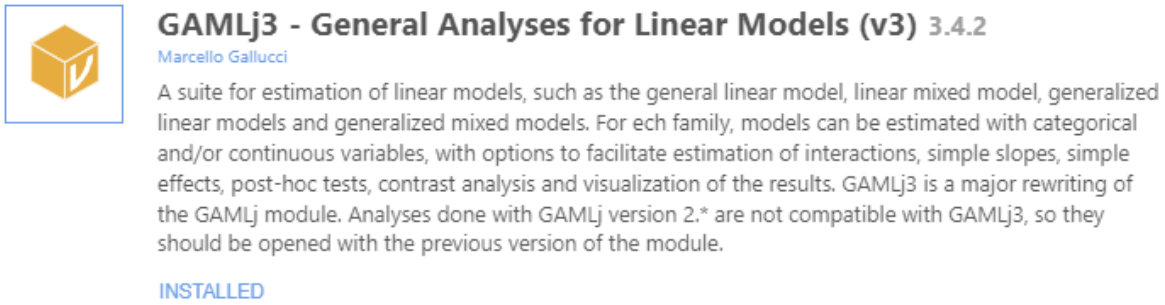
Repeat this process 1,000 times → Find 1,000 Cohen's *d* → Find Confidence Interval

Resampling Techniques

- For example, comparing between A and P types in sports performance:
- Make sure you have the Jamovi version 2.6 or more
- Jamovi -> Analysis tab -> Add Modules

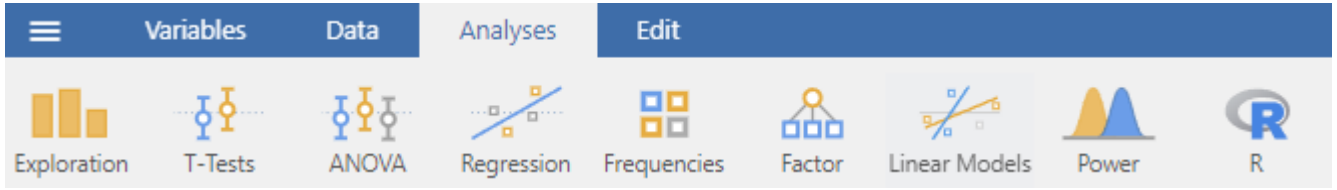


- Jamovi library -> Available tab -> Install `GAMLj3`

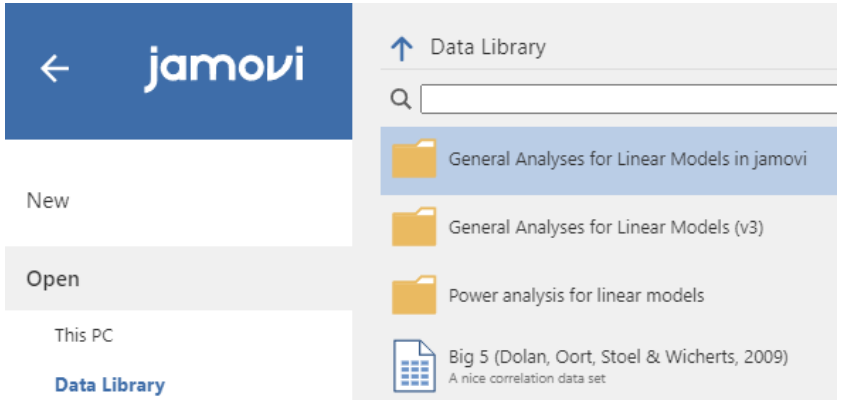


Resampling Techniques

- Make sure you see “Linear Models” in the Analyses tab

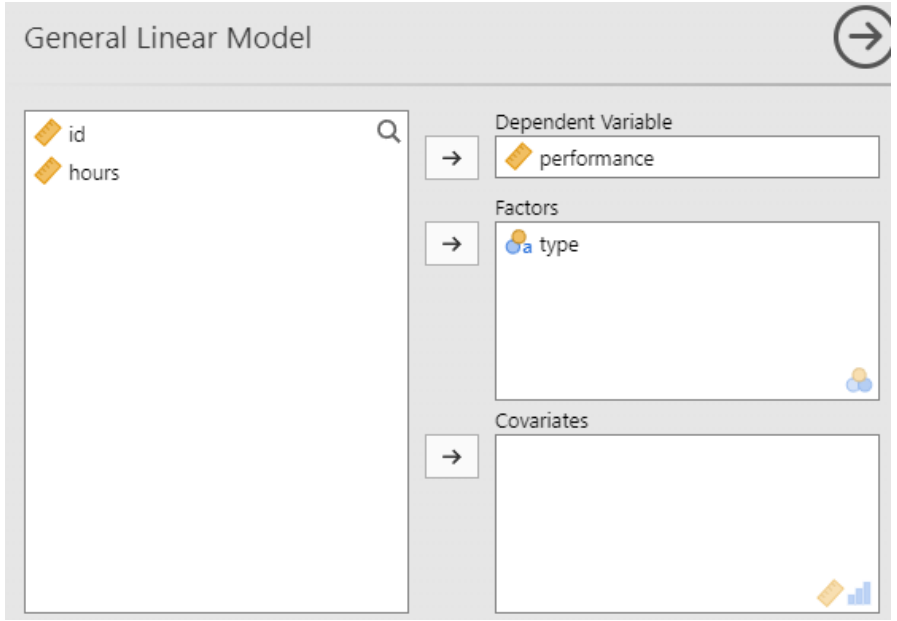


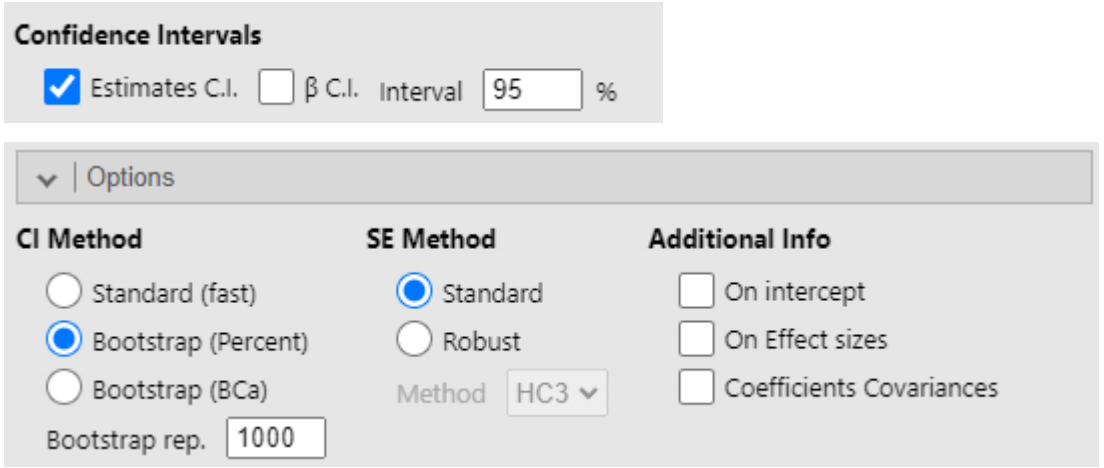
- Load the `qsport` data from the data library of “General Analyses for Linear Model (v3)



Resampling Techniques

- Let's compare performance between type = A and type = P
- Go to Linear Models -> General Linear Model
- Put the variables and click options as follows:





Resampling Techniques

- The 95% percentile bootstrap confidence interval is (-1.501, 3.879). The CI bracketed 0 so the mean difference was not significant.

Parameter Estimates (Coefficients)

				95% Confidence Intervals					
Names	Effect	Estimate	SE	Lower	Upper	β	df	t	p
(Intercept)	(Intercept)	37.880	0.705	36.514	39.290	0.000	98	53.715	< .001
type1	P - A	1.280	1.410	−1.501	3.879	0.182	98	0.908	0.366

- You see that the independent *t*-test is run on regression or general linear model. This topic will be discussed in the regression chapters.

Missing Data Handling

- Missing data is common in survey designs.
- Causes of missing data:
  - Intentionally avoid answering some questions, e.g., income
  - Accidentally skip some questions
  - Some respondents are not eligible to answer certain questions due to the survey design (e.g., non-drivers cannot answer questions about driving behavior).
- Mechanics of missing data: MCAR, MAR, MNAR

Missing Data Handling

- **Missing completely at random (MCAR)**: Missing data process is not systematic, e.g., subjects forget to answer some questions.

Survey	Weight
Online	75
In-person	45
Online	?
Online	65
In-person	55
In-person	50

This respondent forgot to answer this question.

Missing Data Handling

- **Missing at random (MAR)** is that the probability of missing data points is completely based on other variables in the data set. If handled properly, the parameter estimate and standard error are accurate, e.g., the probability of answering age is based on the sources of respondents, which is collected

Survey	Age
Online	75
In-person	45
Online	70
Online	65
In-person	?
In-person	?

In-person respondents tend to avoid answering their ages.

Missing Data Handling

- **Missing not at random (MNAR)**: The probability of missing data points is not completely explained by the variables in the data set. Some variables that explain the missing process are not collected or the probability of missing is based on the variable itself.
- The biases in parameter estimates and standard errors are unavoidable. Some techniques can be used to reduce the impact of nonrandomness.

Missing Data Handling

- For example, the missing of income data is based on the values of income itself. Subjects with high income tend to avoid answering the question.

Survey	Income
Online	?
In-person	15k
Online	20k
Online	15k
In-person	20k
In-person	?

High income tends to not answer the income question.

Missing Data Handling

- In each dataset, each variable may have different missing data processes.
- Unless you plan or know in advance, you never know whether each variable missingness is MCAR, MAR, or MNAR.

Missing Data Handling

- Also, you probably cannot find perfect MAR. MAR and MNAR indicate a degree to which variables in a data set (and a handling method) can recover the missing information. If 100%, then it is MAR. If less than 100%, which is almost always true, it is MNAR.
- Researchers need to find collect all relevant variables that can predict the missing values in the model. For example, if income is missing for high-income subjects, you need to find other relevant variables such as the number of cars, the size of house, the number of travelling abroad each year.

Missing Data Handling

- Traditional methods
  - Listwise (or casewise) deletion
  - Pairwise deletion (This can cause the covariance/correlation matrix to be nonpositive definite.)
  - Mean imputation
  - Regression method
- Some methods provide biases in parameter estimates if the missing process is MCAR. All methods provide biases in standard errors.

Missing Data Handling

- Listwise deletion means if any data point in a case is missing, the entire case is removed

X	*Y*	*Z*
5	7	5
4	8	4
~~3~~	~~?~~	~~1~~
4	7	2
6	9	2

For example, if regression analysis using these three variables, our four rows are used.

The deleted row still have some information in *X* and *Z* but these information is removed. Thus, standard errors are overestimated.

Missing Data Handling

- Pairwise deletion means, if statistics use only two variables, the analysis will use all available pairs. For example, make a correlation matrix

X	*Y*	*Z*
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

Correlation between
- *X* and *Y*: 4 cases
- *X* and *Z*: 5 cases
- *Y* and *Z*: 4 cases

Sometimes, the correlation matrix will be used for further analyses, such as regression, SEM. The correlation matrix may be statistically impossible.

In further analysis, the specified sample size often reflects the total sample size, ignoring the presence of missing data. As a result, the standard errors are underestimated, leading to incorrect conclusions.

Missing Data Handling

- Mean imputation is to substitute the missing value by the mean of the other available information from the variable.

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

Substitute by (7 + 8 + 7 + 9) / 4 = 7.75

This method usually underestimates the relationship between variables and underestimates the variances.

Missing Data Handling

- Regression method is to substitute the missing values by predicted scores by other variables using regression.

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

$$\hat{Y} = 5.611 + 0.583X - 0.194Z$$

$$\hat{Y} = 5.611 + 0.583(3) - 0.194(1)$$
$$= 7.17$$

The regression method overestimates the relationship between variables.

Missing Data Handling

- Stochastic regression methods is to substitute by the predicted score by the regression method added by some error.

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

$$\hat{Y} = 5.611 + 0.583(3) - 0.194(1) = 7.17$$

The substituted value is 7.17 plus a value drawn from *N*(0, standard error of the estimate).

The substituted value is treated as real value. Thus, the standard error is underestimated.

Missing Data Handling

- Within-case mean substitution is to replace the missing value by the mean of other variables within the case.
- Generally, replace by the mean of other items in the same scale.

X	Y_1	Y_2	Y_3	Y_4	Z
5	7	3	1	3	5
4	8	9	5	4	4
3	?	2	7	5	1
4	7	2	7	3	2
6	9	8	5	2	2

Replaced by (2 + 7 + 5) / 3 = 4.67

The assumption is that the average of all items must be equal, which is usually not true.

Also, the replaced value is treated as the real value leading to underestimated standard errors.

Missing Data Handling

- Modern methods
  - Direct maximum likelihood (or full information maximum likelihood)
  - Multiple imputation
- If the missing data process is MAR, the parameter estimates and standard errors are unbiased.

Missing Data Handling

- Direct maximum likelihood or full information maximum likelihood is to estimate statistics by the available data points and skip all missingness. But it is only applicable when missingness occurs on endogenous variables.
- This method is common when using structural equation modeling.

X	*Y*	*Z*
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

Use all available information to estimate statistics. In this case, parameter estimation will use *X*=3 and *Z*=1 too and skip the missing *Y* value. All 5 cases with 14 data points are used in parameter estimation.

Missing Data Handling

- Multiple imputation is steps to impute missing values by predicted scores plus some errors like stochastic regression.
- Instead of having one dataset, the missing values will be replaced multiple times yielding multiple datasets (where the missing values will be imputed by different values).
- If the prediction of a missing value is quite good, the difference of the imputed data point will be less and vice versa.

Missing Data Handling

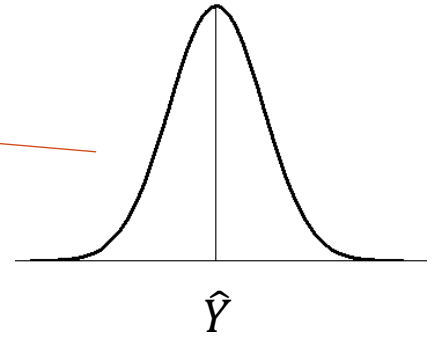
- For example, the imputed missing values are replaced into 30 datasets.
- The target statistics will be computed in each dataset, such as pairwise correlation, regression, and so on.
- Then, 30 sets of statistics will be combined to a single statistics using the Rubin's method to obtain pooled parameter estimate and pooled standard error.

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

$$\hat{Y} = 5.611 + 0.583(3) - 0.194(1)$$
$$= 7.17$$

Assume that $SE(\hat{Y}) = 2$

Random from normal distribution



X	Y	Z
5	7	5
4	8	4
3	10.02	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	3.48	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	6.15	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	10.02	1
4	7	2
6	9	2

$\hat{Y} = 9.8 - 0.03X - 0.53Z$

X	Y	Z
5	7	5
4	8	4
3	3.48	1
4	7	2
6	9	2

$\hat{Y} = 0.2 + 1.37X + 0.24Z$

X	Y	Z
5	7	5
4	8	4
3	6.15	1
4	7	2
6	9	2

$\hat{Y} = 4.1 + 0.8X - 0.1Z$

Imputed No	b_1	$SE(b_1)$
1	-0.028	0.631
2	1.374	0.752
3	0.801	0.425

Using Rubin's Rules

$b_1 = 0.715$
$SE(b_1) = 1.022$

Missing Data Handling

- When missing value rates are less than 5%, traditional methods are not severely biased but modern methods are still better.
- If any variables have missing values over 5%, the modern methods should be considered.
- Multiple imputation has the advantages that researchers can impute data to multiple datasets once. Then, the set of imputed data can be used for any statistical analyses later. But the challenge is that Rubin's method is not applicable in certain models.

Missing Data Handling

- Although Jamovi 2.7 includes add-on modules for handling missing data such as `jYS` and `BSS`, these modules implement stochastic single imputation rather than true multiple imputation.
- In contrast, full multiple imputation procedures are supported in software such as SPSS, R, SAS, and Stata.