



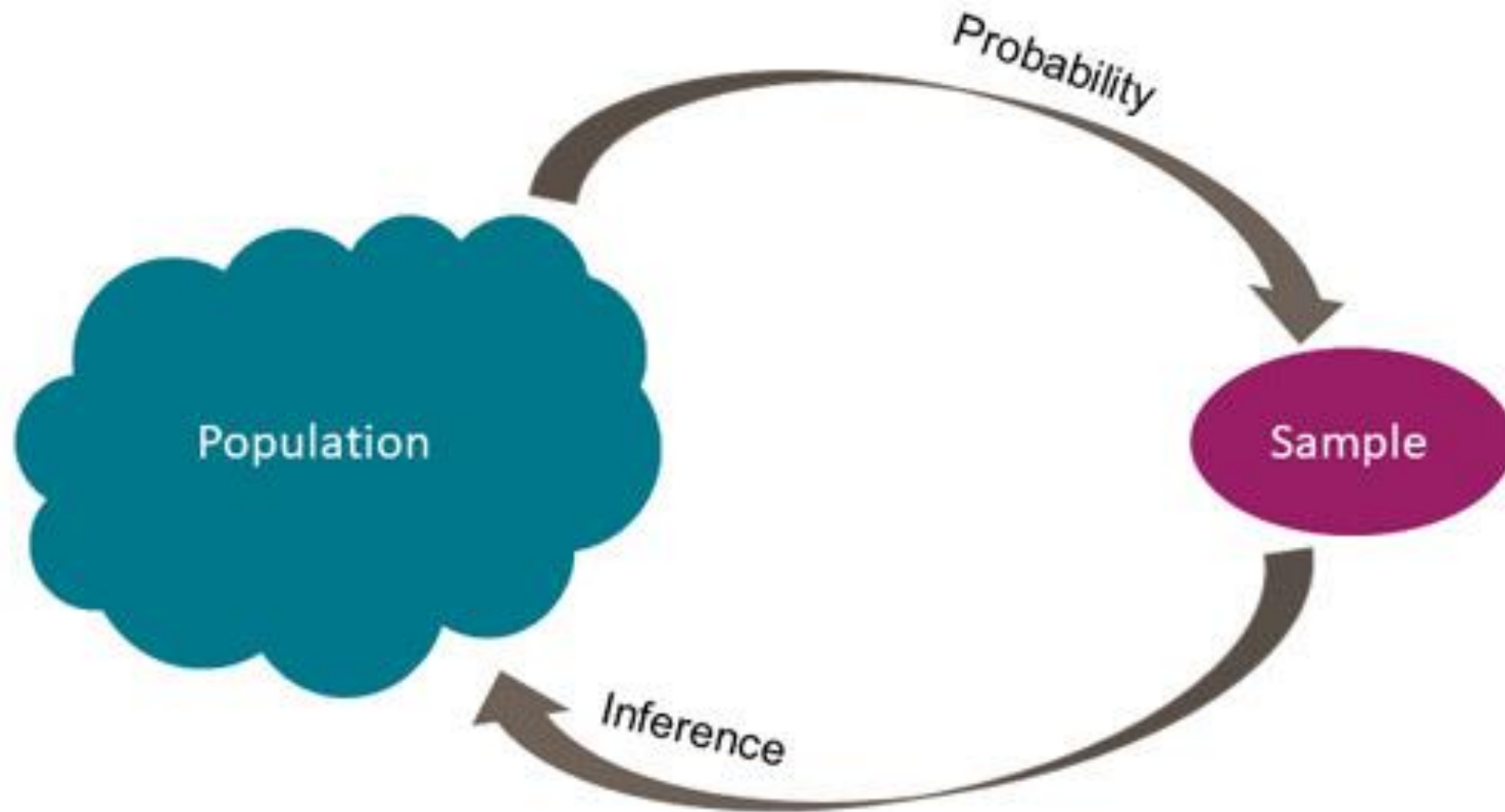
Module 3: Inferential Statistics

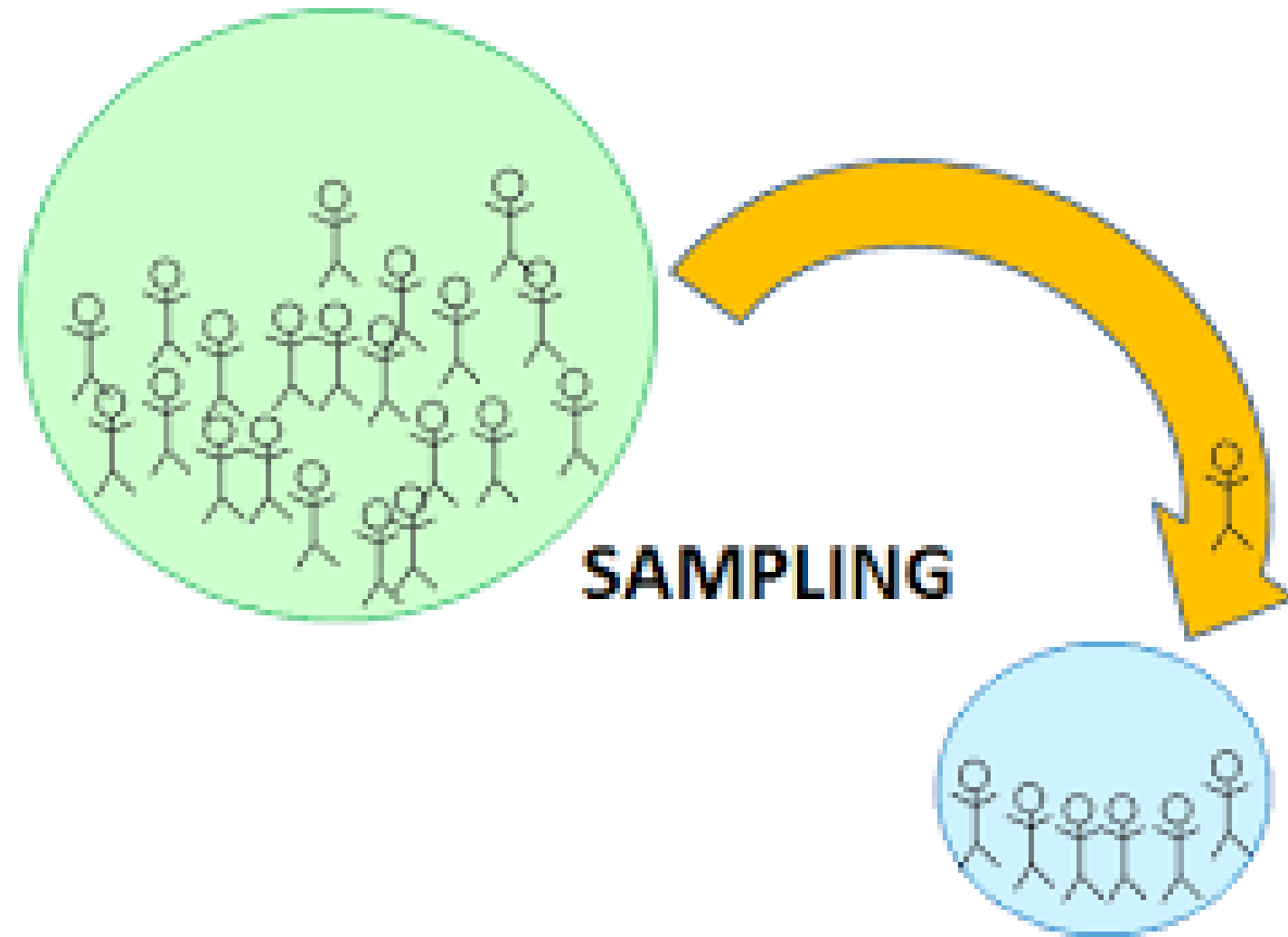
The area of *descriptive statistics* is concerned with meaningful and efficient ways of presenting data.

When it comes to *inferential statistics*, though, our goal is to make some statement about a characteristic of a population based on what we know about a sample drawn from that population.

S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

Inferential statistics





Ex. Dimension of Matrix or Data Frame

```
> set.seed(8212)
```

```
> N <- 500
```

```
# Set Seed for reproducibility
```

```
# Sample size
```

```
> x1 <- round(rnorm(N, 1, 20))
```

```
> x2 <- round(runif(N, 5, 10))
```

```
> x3 <- round(runif(N, 1, 4), 1)
```

```
> x4 <- round(runif(N, 5, 50))
```

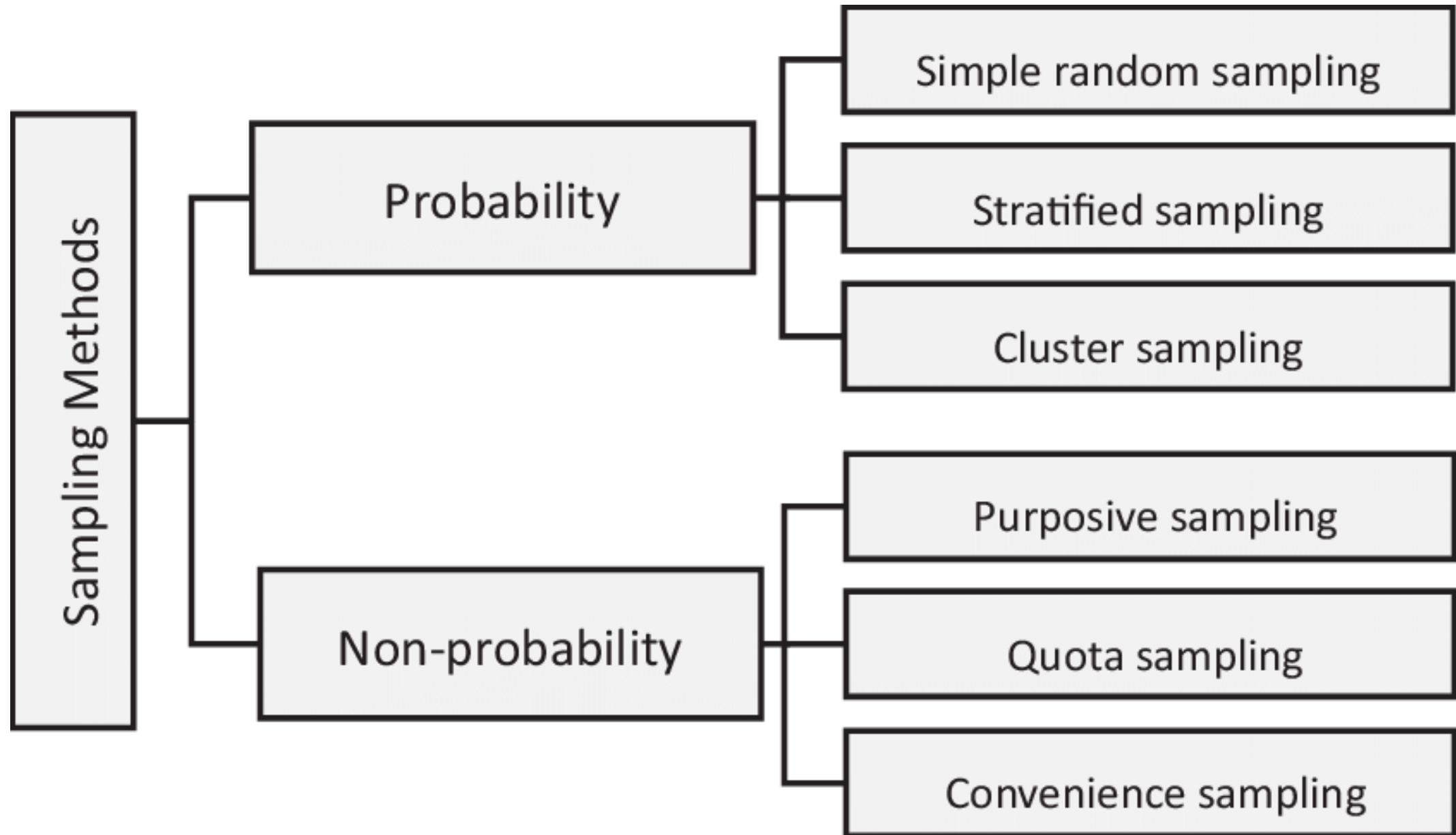
```
> x5 <- rpois(N, 5)
```

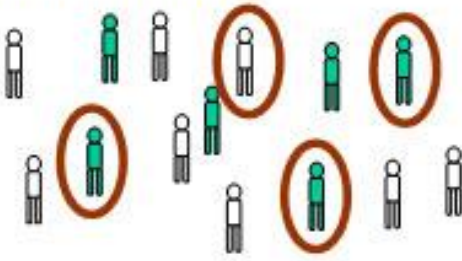
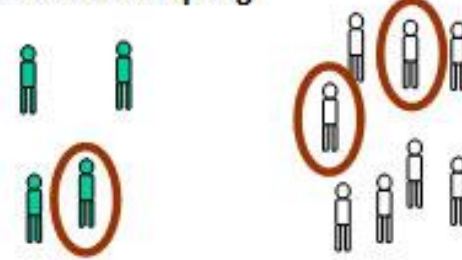
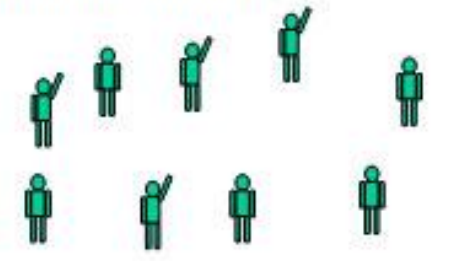
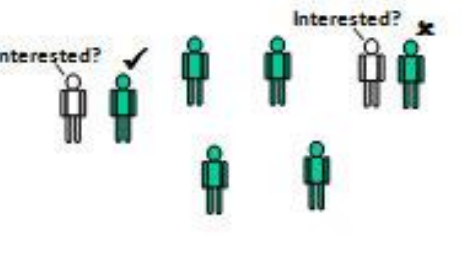
```
# Create 5 random variables
```

	x1	x2	x3	x4	x5
1	29	6	2.5	6	8
2	26	8	1.1	47	6
3	-7	6	3.9	38	5
4	-11	6	1.1	33	7
5	17	7	1.7	27	4
6	-11	7	3.0	41	6

The sampling process comprises several stages:

- Define the population of concern
- Specify a [sampling frame](#), a [set](#) of items or events possible to measure
- Specify a [sampling method](#) for selecting items or events from the frame
- Determine the sample size
- Implement the sampling plan
- Sample and collect data



<p>Random sampling</p> 	<p>Every member of a population has an equal chance of being selected</p> <p>E.g. Pulling names out of a hat</p>	<p>For very large samples it provides the best chance of an unbiased representative sample</p>	<p>For large populations it is time-consuming to create a list of every individual.</p>
<p>Stratified sampling</p> 	<p>Dividing the target population into important subcategories</p> <p>Selecting members in proportion that they occur in the population</p> <p>E.g. 2.5% of British are of Indian origin, so 2.5% of your sample should be of Indian origin... and so on</p>	<p>A deliberate effort is made to make the sample representative of the target population</p>	<p>It can be time consuming as the subcategories have to be identified and proportions calculated</p>
<p>Volunteer sampling</p> 	<p>Individuals who have chosen to be involved in a study. Also called self-selecting</p> <p>E.g. people who responded to an advert for participants</p>	<p>Relatively convenient and ethical if it leads to informed consent</p>	<p>Unrepresentative as it leads to bias on the part of the participant. E.g. a daytime TV advert would not attract full-time workers.</p>
<p>Opportunity sampling</p> 	<p>Simply selecting those people that are available at the time.</p> <p>E.g. going up to people in cafés and asking them to be interviewed</p>	<p>Quick, convenient and economical. A most common type of sampling in practice</p>	<p>Very unrepresentative samples and often biased by the researcher who will likely choose people who are 'helpful'</p>

(Simple) Random Sampling

- A sample selected in such a way that every element in the population has a equal probability of being chosen.
- Equivalently, all samples of size n have an equal chance of being selected.
- Obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.
- Inherent in the concept of randomness: the next result (or occurrence) is not predictable.
- Proper procedure for selecting a random sample: use a random number generator or a table of random numbers
- Assumed when performing conventional statistical analyses
- No guarantee of a representative sample
- May not be feasible (e.g., costly, impractical)

Stratified Sampling (Proportional Sample)

- Population gets partitioned into groups based on a factor that may influence the variable that is being measured.
- These groups called strata.
- An individual group is called a stratum.
- To perform **stratified sampling**:
 - Partition the population into groups (strata)
 - Obtain a simple random sample from each group (stratum)
 - Collect data on each sampling unit that was randomly sampled from each group (stratum)
 - Works best when a heterogeneous population is split into fairly homogeneous groups.
- Generally produces more precise estimates of the population percents than estimates that would be found from a simple random sample. More control over representativeness. Allows for intentional oversampling which permits greater statistical precision (i.e., decreases standard errors).
- Must have data on the characteristics of the population in order to select the sample.

Cluster Sampling

- Stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.
 - Divide the population into groups (clusters).
 - Obtain a simple random sample of so many clusters from all possible clusters.
 - Obtain data on every sampling unit in each of the randomly selected clusters.
- Note that, unlike with the strata in stratified sampling, the clusters should be microcosms, rather than subsections, of the population.
- Each cluster should be heterogeneous.
- Statistical analysis often more complicated than stratified sampling.
- Decreases statistical precision (individuals within groups tend to be more similar so we have less unique information)

Syntax for Sample Function in R:

```
sample(x, size, replace = FALSE, prob = NULL)
```

A single roll of a die is a number between one and six

```
> set.seed(1)
```

```
> sample(1:6, 10, replace=TRUE) #Sample with replacement
```

```
> sample(1:6, 10, replace=TRUE)
```

```
> set.seed(123) #Setting a seed
```

```
> index <- sample(1:nrow(iris), 5)
```

```
> index
```

```
> iris[index, ]
```

#Stratified sampling

set.seed(1)

> d1 <- data.frame(ID = 1:100, A = sample(c("AA", "BB", "CC", "DD", "EE"), 100,
replace = TRUE),

B = norm(100), C = abs(round(rnorm(100), digits = 1)),

D = sample(c("CA", "NY", "TX"), 100, replace = TRUE),

E = sample(c("M", "F"), 100, replace = TRUE))

> summary(d1)

> stratified(d1, "A", 0.1) #A 10% sample from all -A- groups in d1

Non-probability Sampling

Sampling types that should be avoided:

- Convenience (accidental) – selected on the basis of availability
- Quota – selected on the basis of availability with “quotas” being selected to represent the distribution in the population.
- Judgmental – researcher selects units he/she thinks are more representative of the population. Every unit is not eligible for inclusion in the sample, personal biases
- Snowball – unit with a desired characteristic is identified. This unit then identifies other units with desired characteristics and so on.... (i.e. social networks)

These are referred to as "sampling disasters".

- Biased samples
- Based on human choice rather than random selection
- Statistical theory cannot explain how they might behave and potential sources of bias are rampant.

Discussion: Variable Selection

How do these limitations impact your selection of data?

- Highly predictive variables — the use of which is prohibited by legal, ethical or regulatory rules.
- Some, some variables might not be available or might be of poor quality during modeling or production stages.
- The business will always have the last word and might insist that only business-sound variables are included or request monotonically increasing or decreasing effects.

Sampling Challenges

- Sampling error - discrepancies between the sample and the population on a certain parameter that are due to random differences; no fault of the researcher.
- *Systematic error* - difference between the sample and the population that is due to a systematic difference between the two rather than random chance alone.
- *Response rate* - sample can become self-selecting, and that there may be something about people who choose to participate in the study that affects one of the variables of interest.
- *Coverage error* - refers to the fact that sometimes researchers mistakenly restrict their sampling frame to a subset of the population of interest

*The more participants a study has, the less likely the study is to suffer from sampling error.

Motivation . . .

The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favor of a certain belief, or hypothesis, about a parameter.

It is a 7-step process:

1. Make Assumptions.
2. Take an initial position.
3. Determine the alternate position.
4. Set acceptance criteria
5. Conduct fact based tests.
6. Evaluate results. Does the evaluation support the initial position?
Are we confident that the result is not due to chance?
7. Reach one of the following conclusion: Reject the original position in favor of alternate position or fail to reject the initial position.

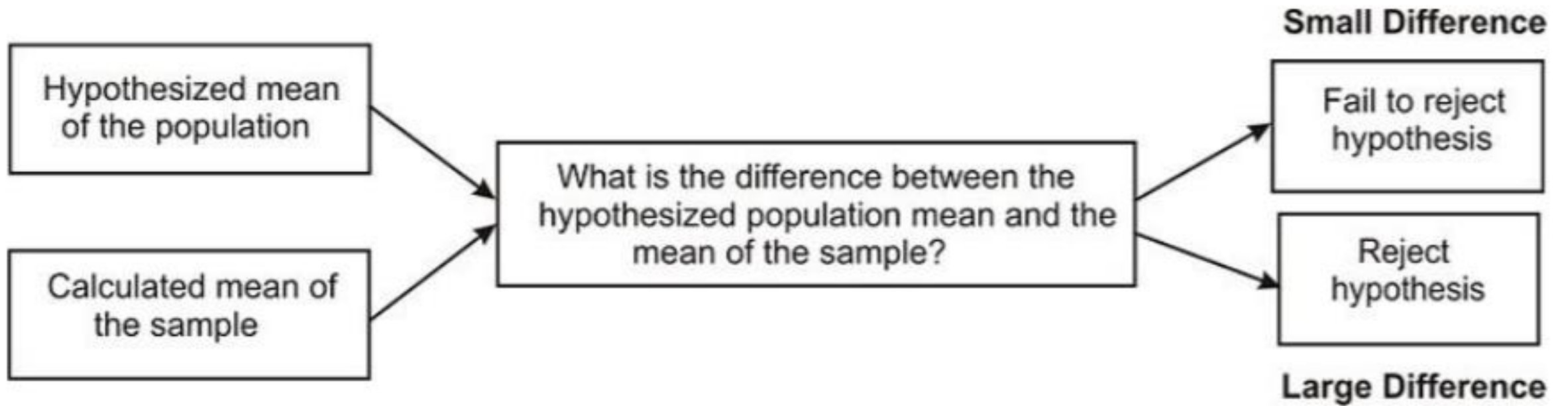
Example

Consider a person on trial for a “criminal” offense in the United States. Under the US system a jury (or sometimes just the judge) must decide if the person is innocent or guilty while in fact the person may be innocent or guilty.

Person Is:

		Innocent	Guilty
		No Error	Error
Jury Says:	Innocent	No Error	Error
	Guilty	Error	No Error

There is a favored assumption, an initial bias. The jury is instructed to assume the person is innocent.



Recall the **Central Limit Theorem**:

- Using this, we determine if our assumption for the null hypothesis (**H0**) is reasonable or not. If it is unlikely, by the **Rare Event rule**, our hypothesis is probably incorrect (i.e. reject **H0**).

In general, we use the following:

- If the test sample yields an unlikely result, it is probably incorrect (reject **H0**)
- If the test sample yields a likely result, it is probably correct (fail to reject **H0**)

- There is an extremely close relationship between confidence intervals and hypothesis testing.
- When a 95% confidence interval is constructed, all values in the interval are considered plausible values for the parameter being estimated. Values outside the interval are rejected as relatively implausible.
- The confidence interval tells you **more than just the possible range around the estimate**. It also tells you about **how stable the estimate is**.
- If exact p-value is reported, then the relationship **between confidence intervals and hypothesis testing** is very close. However, the objective of the two methods is different: **Hypothesis testing** relates to a single conclusion of statistical significance vs. no statistical significance.

- Probability theory: Allows us to calculate the exact *probability* that chance was the real reason for the relationship.
- Probability theory allows us to produce test statistics (using mathematical formulas)
- A test statistic is a number that is used to decide whether to accept or reject the null hypothesis.
- The most common statistical tests include:
 - Chi-square
 - T-test
 - ANOVA
 - Correlation
 - Linear Regression

Create Sample Data in R

#Define Sample 1

```
smp2014 <- c(222, 823, 1092, 400, 948, 836)
```

#Define Sample 2

```
smp2019 <- c(910, 650, 700, 892, 229, 1051)
```

#Two sample T-test

```
t.test(smp2014, smp2018, var.equal=FALSE)
```

#What is the p-value?

#Run Welch's T-test of Equal Variance

```
t.test(smp2014, smp2019, var.equal=TRUE)
```


Conditions for a 1 sample z-test

A one sample z test is one of the most basic types of hypothesis test.

1. Normality -normal population. Data roughly fits a [bell curve](#) shape.
Central Limit Theorem ($n \geq 30$)

Graphing - Qplots/box plots/normal probability

2. Independence - population is greater than 10 times the sample size ($N \geq 10n$)

Null hypothesis for a 1 sample z-test

- The mean of a population is equal to the sample mean
 $\mu = \mu_0$

Alternative hypotheses for a 1 sample z-test

- The mean of a population is (not equal to/less than/greater than) the sample mean
 $\mu \neq \mu_0$ OR $\mu < \mu_0$ OR $\mu > \mu_0$

Conditions for a 2 sample z-test

Both (two) populations are normally distributed

Null hypothesis for a 2 sample z-test

- The difference between the means of 2 populations is zero (equal means)

$$\mu_1 - \mu_2 = 0 \text{ OR } \mu_1 = \mu_2$$

Alternative hypotheses for a 2 sample z-test

- Population 1 mean is (not equal to/greater than/less than) the population 2 mean

Confusion Matrix

		Prediction Result	
		Attrition	No Attrition
Actual State	Attrition	True Positive	False Negative
	No Attrition	False Positive	True Negative

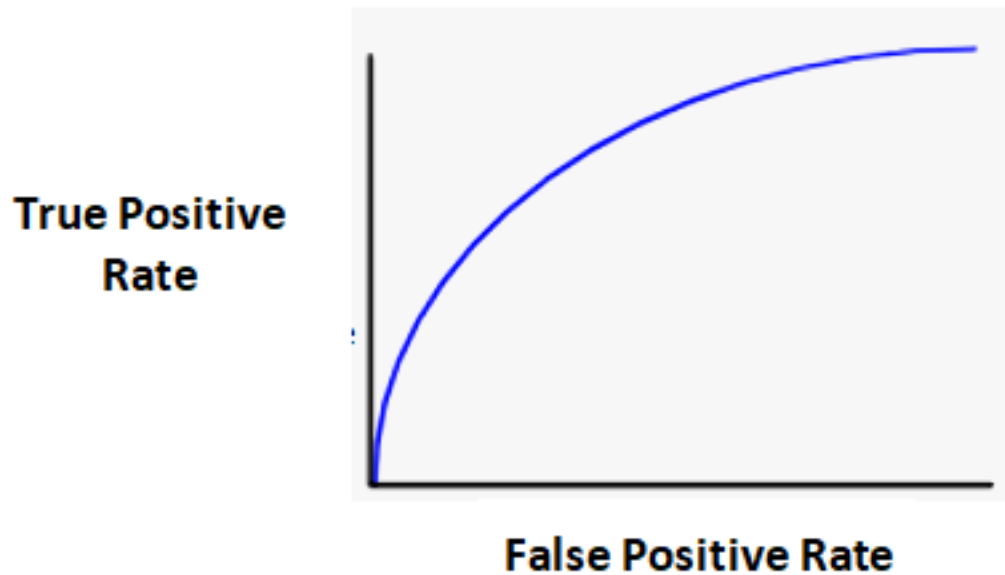
Precision = True Positive / (True Positive + False Positive)

- True Positive : Number of customers who actually attrited whom we correctly predicted as attritors.
- False Positive : Number of customers who actually did not attrite whom we incorrectly predicted as attritors.

Recall = True Positive / (True Positive + False Negative)

- True Positive : Number of customers who actually attrited whom we correctly predicted as attritors.
- False Negative: Number of customers who actually attrited whom we incorrectly predicted as non-attritors.

Precision Recall Curve: popular model performance metrics to evaluate binary classification model



Advantage of using AUPRC over ROC

Cutoff	0.9	0.75	0.6	0.5	0.4
Recall (X)	0.14	0.41	0.69	0.88	0.90
Precision (Y)	0.90	0.85	0.82	0.81	0.50

```
#Area under the precision recall area
recall=c(0.14, 0.41, 0.69, 0.88, 0.90)
precision=c(0.90, 0.85, 0.82, 0.81, 0.50)
i = 2:length(recall)
recall = recall[i] - recall[i-1]
precision = precision[i] + precision[i-1]
(AUPRC = sum(recall * precision)/2)
```

Were you right ? ...

The Truth (Based on Entire Population)			
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

Which of the two errors is more serious? Type I or Type II ?

- Since *Type I is the more serious error* (usually), rationale is stick to the status quo or default assumption, at least you're not making things *worse*. But it depends...
- Note: alpha is not a Type I error. Alpha, ' α ', is the *probability of committing* a Type I error. (i.e. $\alpha = .05$; 5% is the level of reasonable doubt that you are willing to accept when statistical tests are used to analyze the data after the study is completed.). Likewise beta, ' β ', is the *probability of committing* a Type II error. A Type II error relates to the concept of "power,"; having enough power depends on whether the sample size is sufficiently large to detect a difference when it exists.
- Although type I and type II errors can never be avoided entirely, the you can reduce the likelihood of occurrence by increasing the sample size (the larger the sample, the lesser is the likelihood that it will differ substantially from the population).
- False-positive and false-negative results can also occur because of bias (observer, instrument, recall, etc.). (Errors due to bias, however, are not referred to as type I and type II errors.) Such errors are troublesome, since they may be difficult to detect and cannot usually be quantified.

Conclusions are sentence answers which include whether there is enough evidence or not (based on the decision), the level of significance, and whether the original claim is supported or rejected.

Conclusions are based on the original claim, which may be the null or alternative hypotheses. The decisions are always based on the null hypothesis

Decision	Original Claim	
	H_0 "REJECT"	H_1 "SUPPORT"
Reject H_0 "SUFFICIENT"	There is sufficient evidence at the alpha level of significance to reject the claim that (insert original claim here)	There is sufficient evidence at the alpha level of significance to support the claim that (insert original claim here)
Fail to reject H_0 "INSUFFICIENT"	There is insufficient evidence at the alpha level of significance to reject the claim that (insert original claim here)	There is insufficient evidence at the alpha level of significance to support the claim that (insert original claim here)

① One-tailed (directional)

$$H_A: \rho > 0$$

$$H_A: \rho < 0$$



② Two-tailed (non directional)

$$H_A: \rho \neq 0$$



Distinction between NON-DIRECTIONAL and DIRECTIONAL: Research hypotheses

Non-directional hypotheses

- Only state that one group differs from another on some characteristic, i.e., it does NOT specify the DIRECTION of the difference
- Example: H0 -- Michigan students differ from the college population in ideological attitudes

Directional hypotheses

- Specifies the nature of the difference, i.e., that one group is higher, or lower, than another group on some attribute
- Example:
- UM students are more conservative than other students = H1
- UM students are more liberal than other students = H2

Bayesian

THE PROBABILITY OF "B"
BEING TRUE GIVEN THAT
"A" IS TRUE

↓

THE PROBABILITY
OF "A" BEING
TRUE

↙

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

↑

THE PROBABILITY
OF "A" BEING TRUE
GIVEN THAT "B" IS
TRUE

↖

THE PROBABILITY
OF "B" BEING
TRUE

t -test statistic

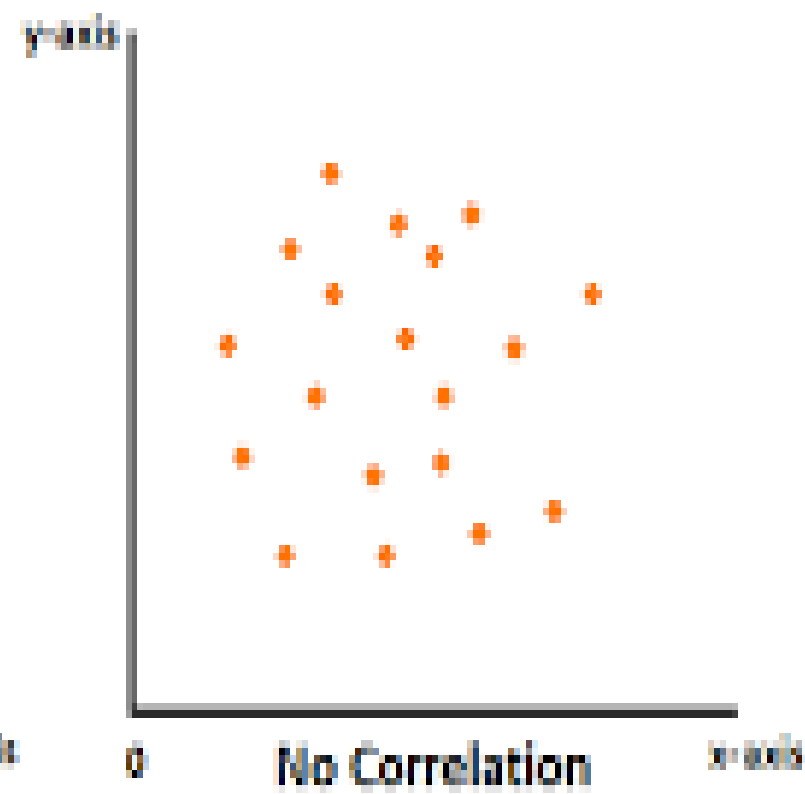
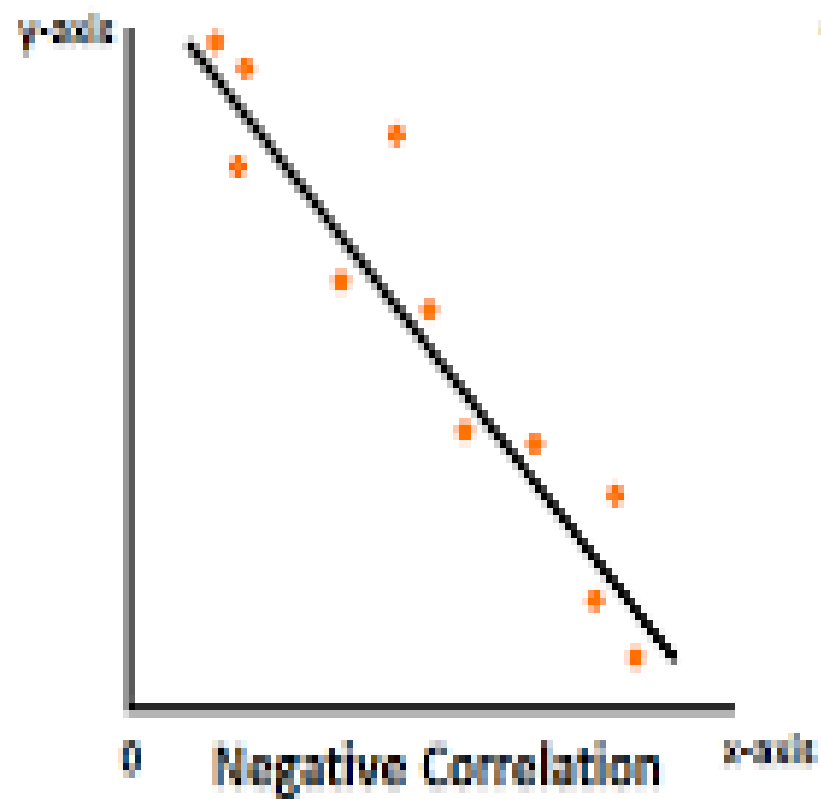
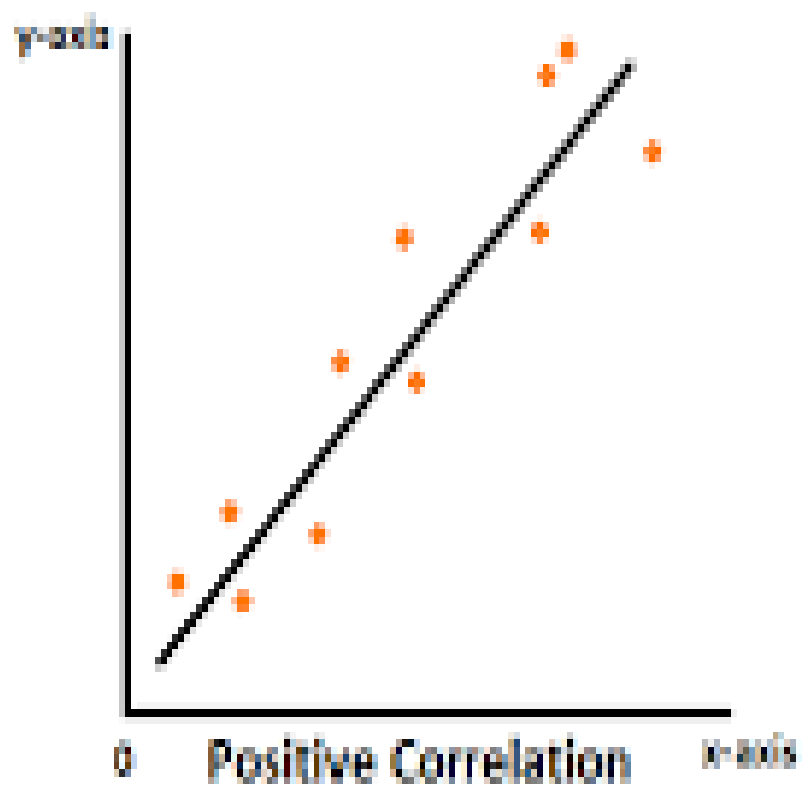
- The t statistic allows researchers to use sample data to test hypotheses about an unknown population mean.
- The t statistic is mostly used when a researcher wants to determine whether or not a treatment intervention causes a significant change from a population or untreated mean.
- The goal for a hypothesis test is to evaluate the *significance* of the observed discrepancy between a sample mean and the population mean.
- Therefore, the t statistic requires that you use the sample data to compute an estimated standard error of M .
- A large value for t (a large ratio) indicates that the obtained difference between the data and the hypothesis is greater than would be expected if the treatment has no effect.

Student's t-test: Commands and explanation

- `t.test(data.1, data.2)` – The basic method of applying a t-test is to compare two vectors of numeric data.
- `var.equal = FALSE` – If the `var.equal` instruction is set to `TRUE`, the variance is considered to be equal and the standard test is carried out. If the instruction is set to `FALSE` (the default), the variance is considered unequal and the Welch two-sample test is carried out.
- `mu = 0` – If a one-sample test is carried out, `mu` indicates the mean against which the sample should be tested.
- `alternative = "two.sided"` – It sets the alternative hypothesis. The default value for this is `"two.sided"` but a greater or lesser value can also be assigned. You can abbreviate the instruction.
- `conf.level = 0.95` – It sets the confidence level of the interval (default = 0.95).
- `paired = FALSE` – If set to `TRUE`, a matched pair T-test is carried out.
- `t.test(y ~ x, data, subset)` – The required data can be specified as a formula of the form `response ~ predictor`. In this case, the data should be named and a subset of the predictor variable can be specified.
- `subset = predictor %in% c("sample.1", sample.2")` – If the data is in the form `response ~ predictor`, the two samples to be selected from the predictor should be specified by the `subset` instruction from the column of the data.

Type of test	Level of measurement	Sample characteristics					Correlation
		One sample	Two sample		K samples (i.e., >2)		
			Independent	Dependent	Independent	Dependent	
Parametric	Interval or ratio	Z-test or <i>t</i> -test	Independent sample <i>t</i> -test	Paired sample <i>t</i> -test	One-way ANOVA	Repeated measure ANOVA	Pearson's test
Nonparametric	Categorical or nominal	Chi-square test	Chi-square test	Mc-Nemar test	Chi-square test	Cochran's Q	Spearman's rho
	Rank or ordinal	Chi-square test	Mann-Whitney U-test	Wilcoxon signed rank test	Kruskal-Wallis	Friedman's ANOVA	

ANOVA: Analysis of variance



Correlations

- A simplified format is **cor**(x, use=, method=)

Correlations/covariances among numeric variables in

data frame mtcars.

Use listwise deletion of missing data.

```
> cor(mtcars, use="complete.obs", method="kendall")
```

```
> cov(mtcars, use="complete.obs")
```

#Use corrgram() to plot correlograms #Visualizing Correlations

Chi-square

- The function used for performing chi-Square test is `chisq.test()`. The syntax is `Chisq.test(data)`.

```
> install.library("MASS") #install package
```

```
> library ("MASS") #load library
```

```
# Create a data frame from the main data set.
```

```
> car.data <- data.frame(Cars93$AirBags, Cars93$Type)
```

```
# Create a table with the needed variables.
```

```
> car.data = table(Cars93$AirBags, Cars93$Type)
```

```
> print(car.data)
```

```
# Perform the Chi-Square test.
```

```
> print(chisq.test(car.data))
```


Analysis of Variance (ANOVA)

One Way Anova (Completely Randomized Design)

```
>fit <- aov(y ~ A, data=mydataframe)
```

Randomized Block Design (B is the blocking factor)

```
>fit <- aov(y ~ A + B, data=mydataframe)
```

Two Way Factorial Design

```
>fit <- aov(y ~ A + B + A:B, data=mydataframe)
```

```
>fit <- aov(y ~ A*B, data=mydataframe) # same thing
```

Analysis of Covariance

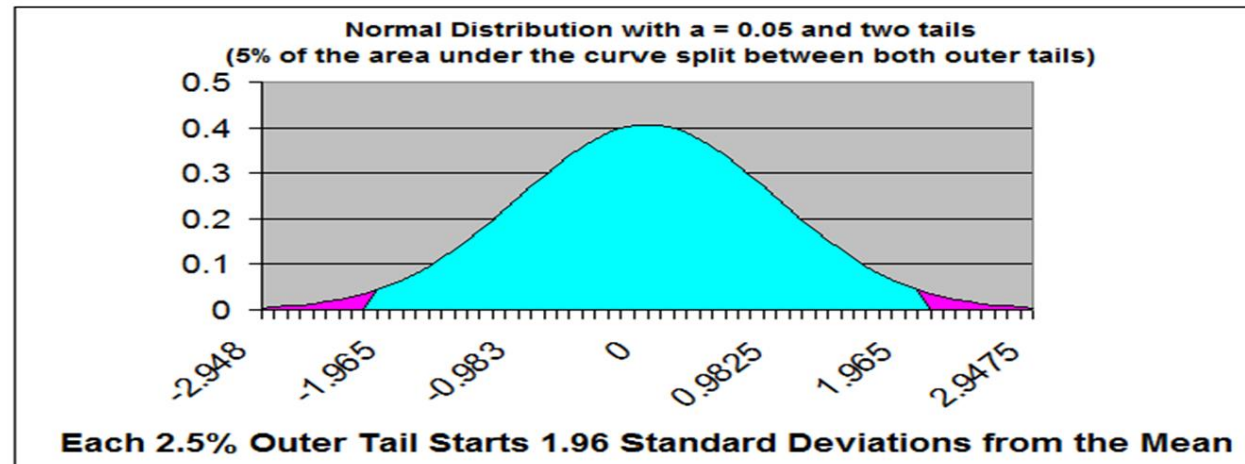
```
>fit <- aov(y ~ A + x, data=mydataframe)
```

#Diagnostic **plots** provide checks for heteroscedasticity, normality, and influential observations.

	Outcome variable						
Input Variable		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
	Nominal	χ^2 or Fisher's	χ^2	χ^2 -trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank ^a	Student's <i>t</i> test
	Categorical (2>categories)	χ^2	χ^2	Kruskal-Wallis ^b	Kruskal-Wallis ^b	Kruskal-Wallis ^b	Analysis of variance ^c
	Ordinal (Ordered categories)	χ^2 -trend or Mann-Whitney	e	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression ^d
	Quantitative Discrete	Logistic regression	e	e	Spearman rank	Spearman rank	Spearman rank or linear regression ^d
	Quantitative non-Normal	Logistic regression	e	e	e	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	e	e	e	Linear regression ^d	Pearson and linear regression

Significance

- If the test statistic produced by the statistical test (using a mathematical formula) falls within a specified *rejection region* on the normal distribution, then we can conclude that the relationship between the independent and dependent variables *is unlikely to be due to chance*. (rejection = rejection of the NULL hypothesis)
- The rejection region is determined by the researcher prior to conducting the statistical test and is called the *alpha level*.



p Values

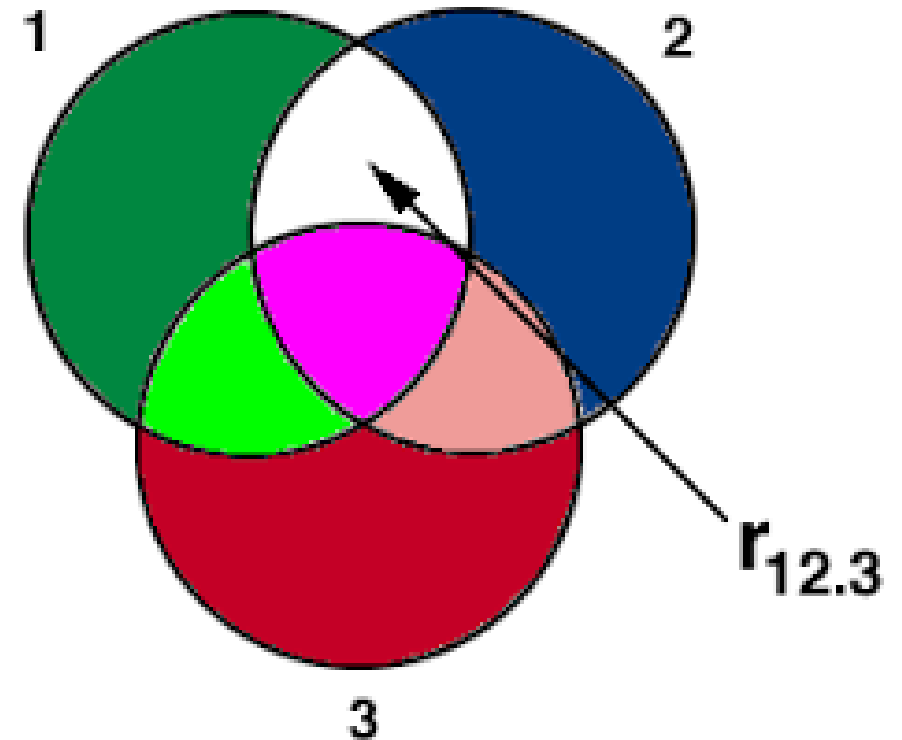
- Each test statistic has a p value (a probability value) associated with it.
- When you plot a test statistic on the normal distribution, the location of the test statistic on the normal distribution is associated with a p value, or a probability.
- If the p value produced by the test statistic is within the rejection region on the normal distribution, then you reject the null hypothesis and conclude that there is a relationship between the independent and the dependent variables. This shows statistical significance.

Partial Correlation

- Partial Correlation measures relationship between two variables (X,Y) while eliminating influence of a third variable (Z).
- Called “correlation” but it is actually regression. It requires estimation of variances.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

```
# Partial correlation  
library(ppcor)  
with(mydata, pcor.test(Height,Weight,Age))
```



Semi-partial Correlation

- Semi-partial correlation measures the strength of linear relationship between variables X1 and X2 holding X3 constant for just X1 or just X2. It is also called part correlation.

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}} \text{ and } r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}}$$

```
#Semipartial Correlation Coefficient
```

```
with(mydata, spcor.test(Height,Weight,Age))
```

```
#Semi partial correlation - Age constant for Weight only
```

Testing for Normality

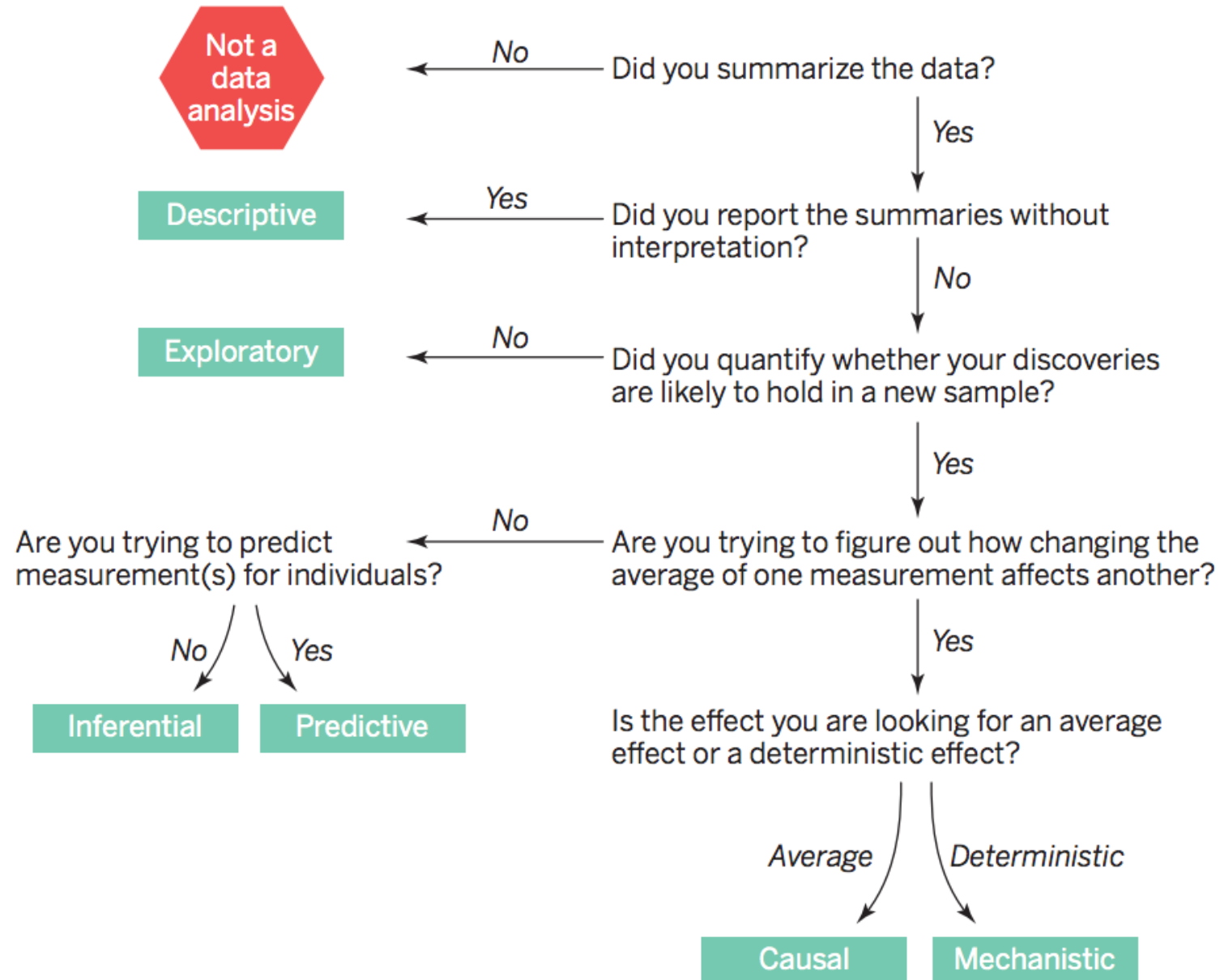
- Plotting returns in R `plot(x$s) or ggplot`
- Kolmogorov-Smirnov test in R `ks.test(x$s, "pnorm", mean=mean(x$r), sd=sd(x$r))`
- Shapiro-Wilk test in R `shapiro.test(x$columnname)`
- Jarque-Bera test in R `install.packages("tseries")
library(tseries)
jarque.bera.test(x$columnname)`

`#import the data`

`View(rtns)`

`#data wrangle and select a column from a dataframe (use select())`

Data analysis flowchart



Variable selection method	Examples
Supervised variable selection outside predictive models (Figure 3)	Information value Chi-square statistics Gini index
Unsupervised variable selection/extraction outside predictive models	Correlation analysis Cluster analysis Principal component analysis Neural networks
Supervised variable selection inside predictive models	Recursive feature selection: forward, backward and stepwise Regularisation techniques (for example, AIC/BIC, lasso, ridge) Ensemble techniques (for example, random forest and gradient boosting) Cross validation