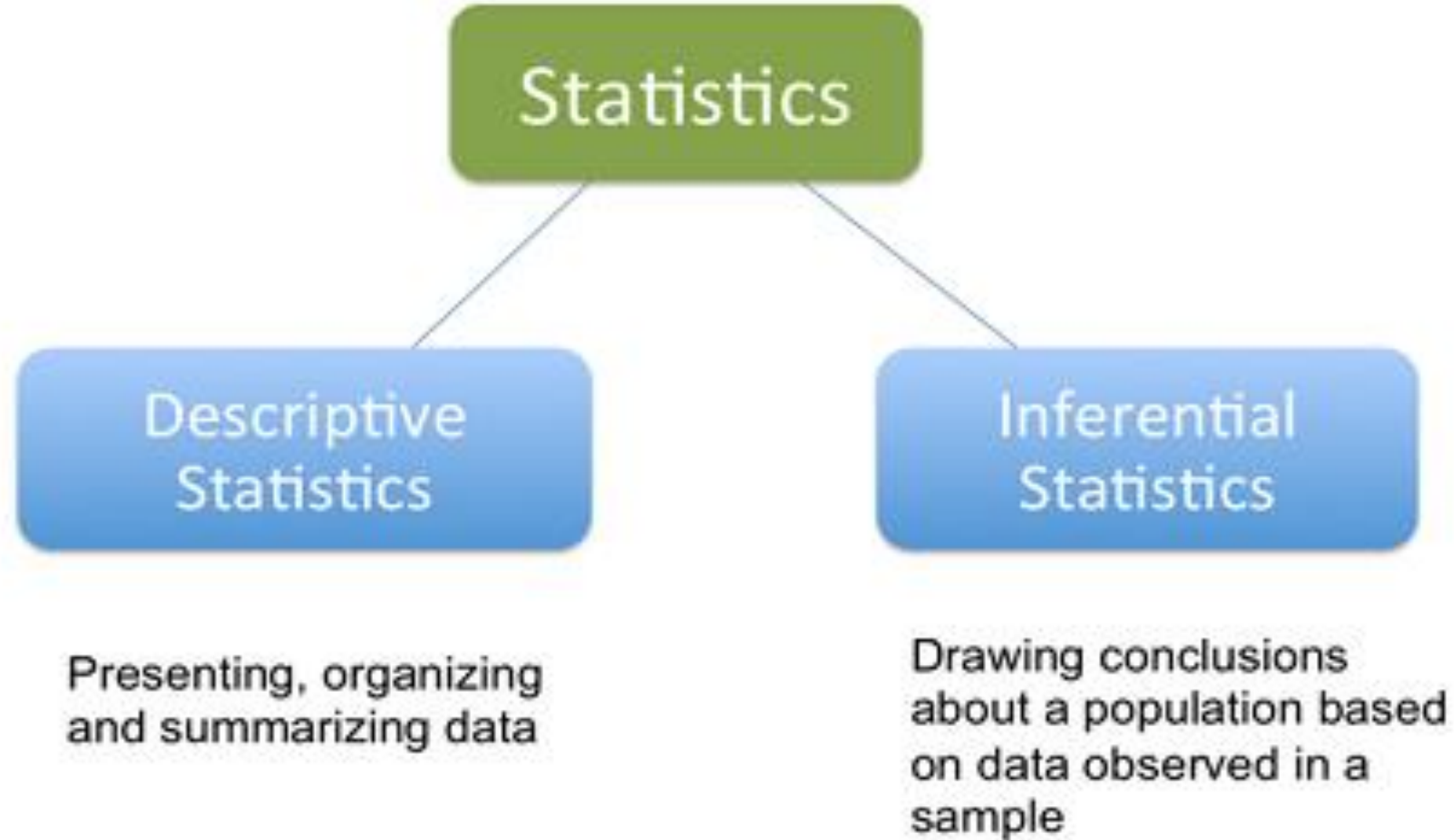
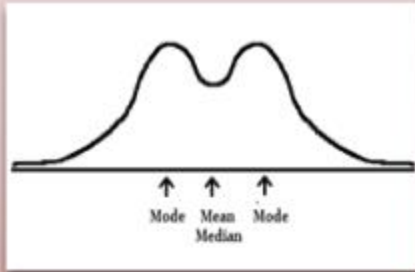




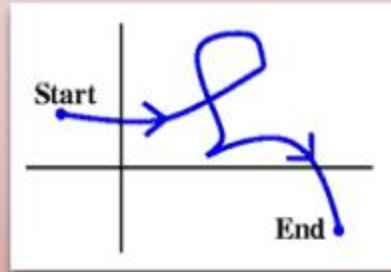
Module 5: Statistics





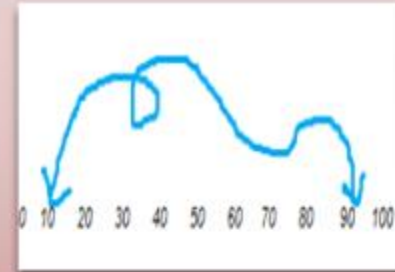
Central Tendency

Mean, Median, Mode, Outliers



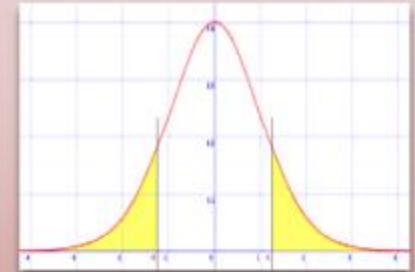
Measures of Spread

Range, Standard deviation, Variance, Quartiles



Percentiles

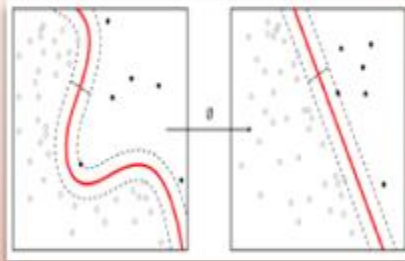
Position of data, percentile rank, percentile range



Probability Distributions:

Uniform, normal (Gaussian), Poisson

Basic Probability and Statistics



Dimensionality reduction

Pruning, PCA



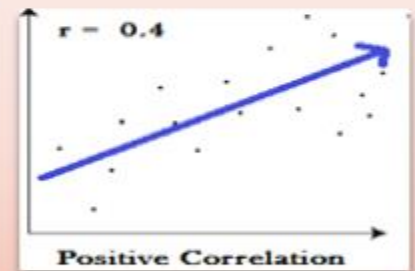
Sampling

SRS, Reservoir, Undersampling, Oversampling,

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

Bayesian statistics

Measuring belief or confidence

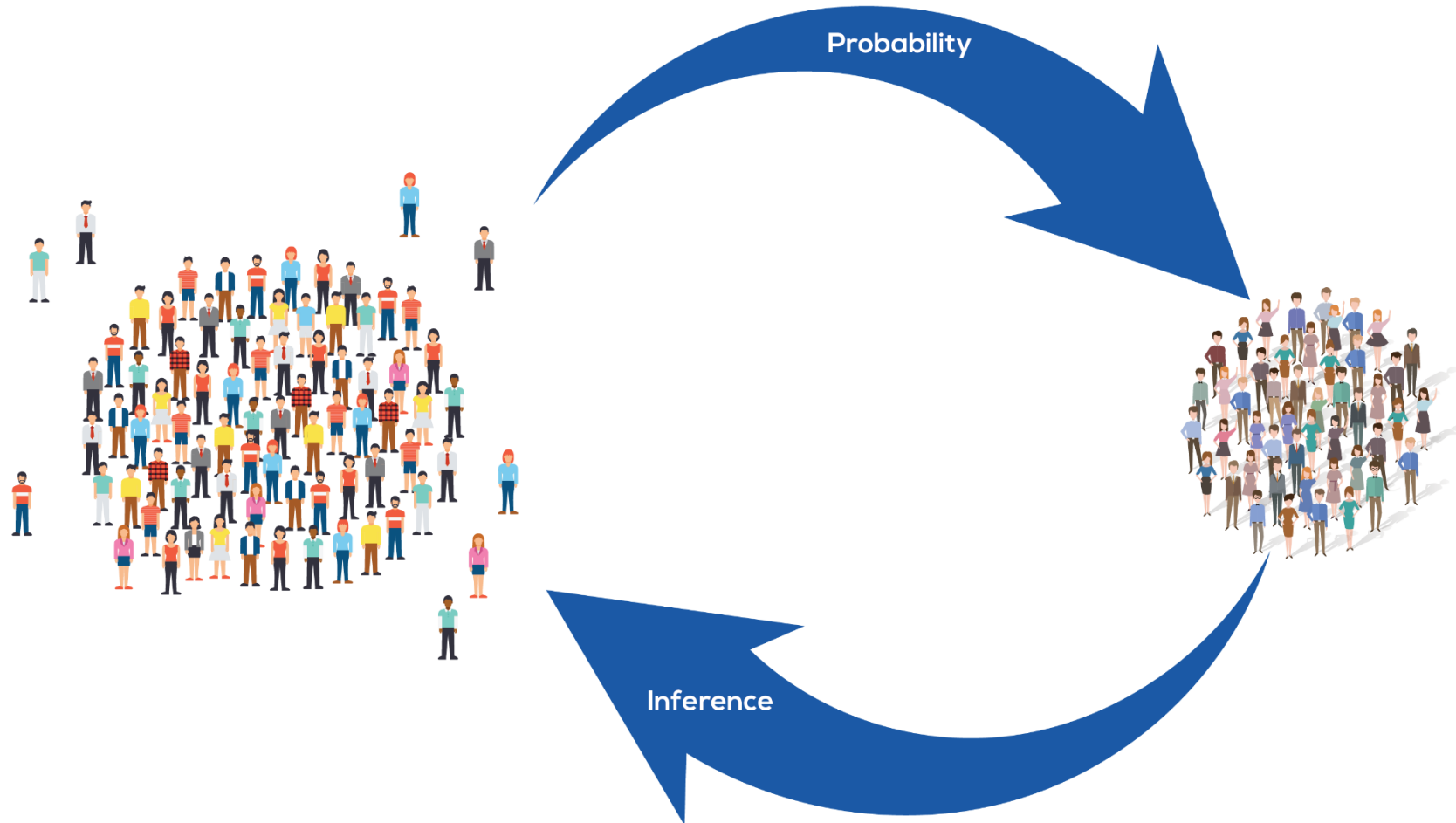


Covariance & correlation

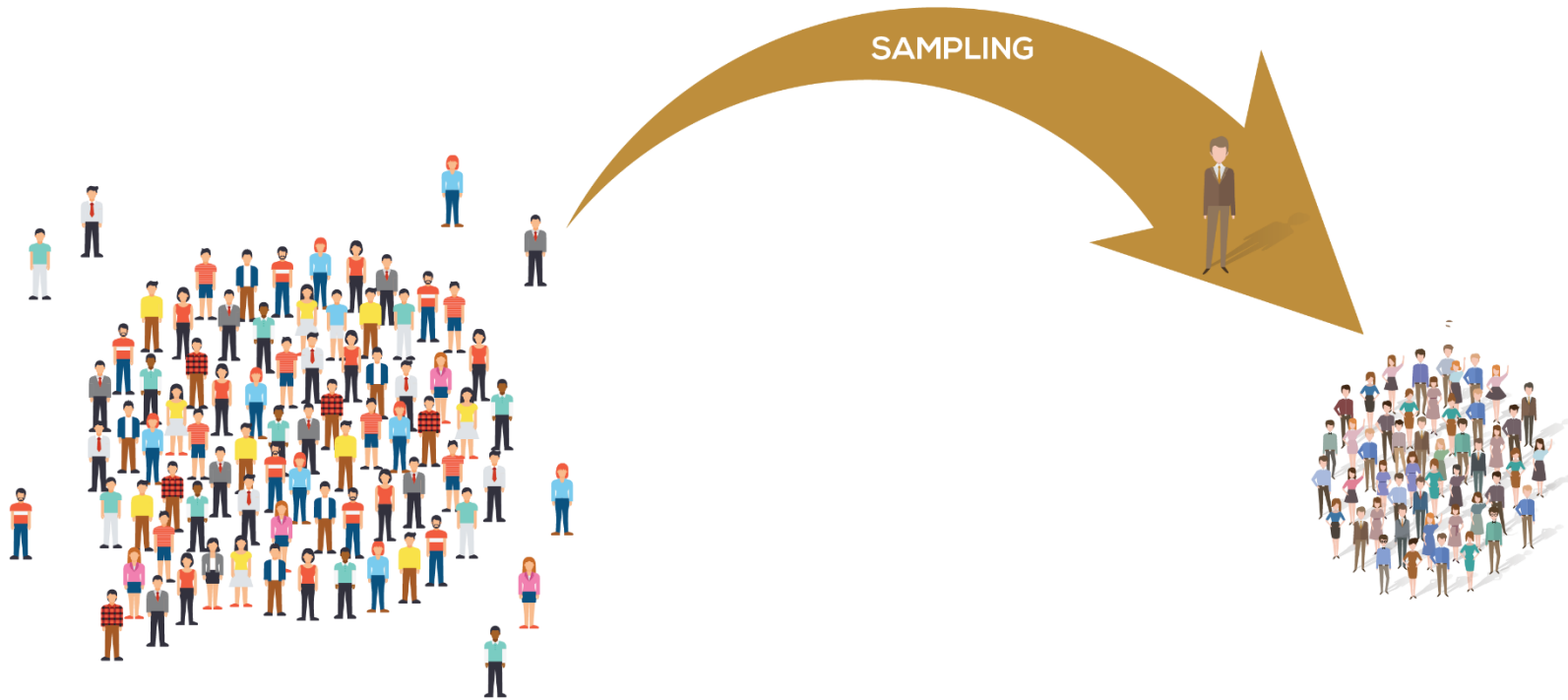
How data is related

More Advanced Probability and Statistics

Inferential Statistics

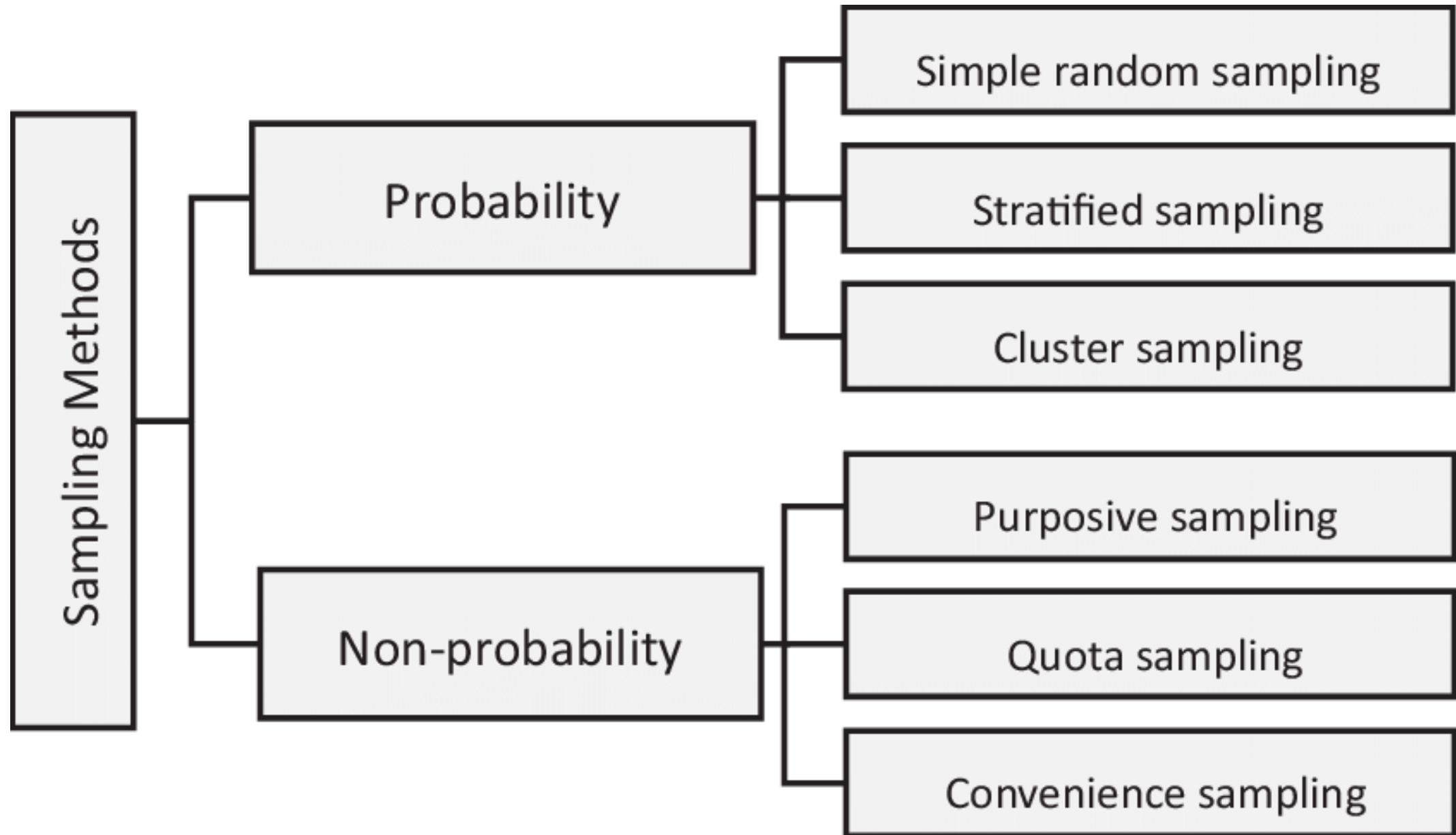


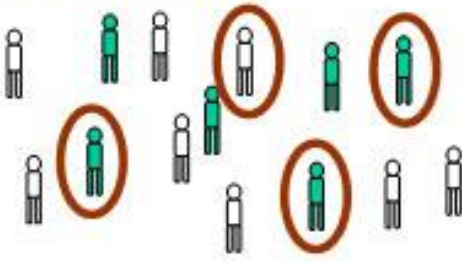
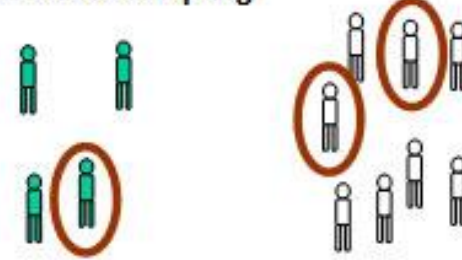
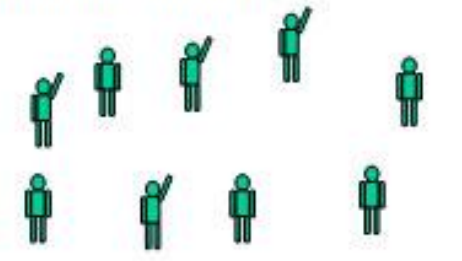
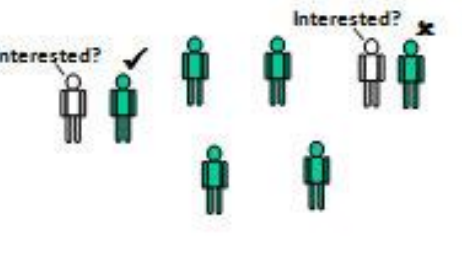
Sampling



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>

Generating random data

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

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

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

> set.seed(123)
> index <- sample(1:nrow(iris), 5)
> index
> iris[index, ]
```

Simple random distributions:

1. Normal (aka Gaussian): bell-shaped, and has two parameters: a mean and a standard deviation.

- *# 6 samples from a Normal dist with mean = 0, sd = 1*
rnorm(n = 6, mean = 0, sd = 1)
- *# 4 samples from a Normal dist with mean = -10, sd = 15*
rnorm(n = 4, mean = -10, sd = 20)

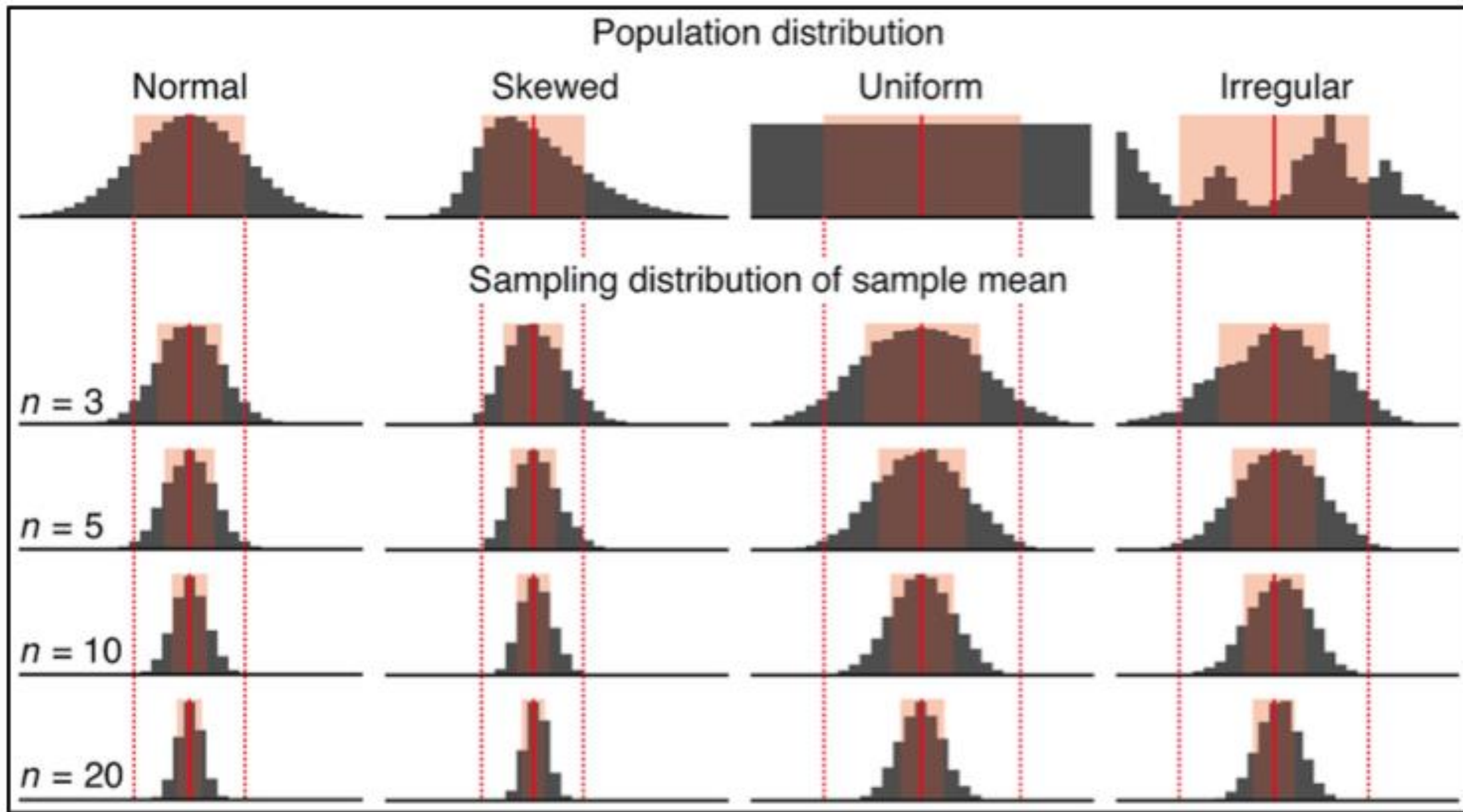
2. Uniform: everything between its lower and upper bounds are equally likely to occur.

- *# 5 samples from Uniform dist with bounds at 0 and 1*
runif(n = 5, min = 0, max = 1)
- *# 10 samples from Uniform dist with bounds at -75 and +75*
runif(n = 5, min = -75, max = 75)

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.

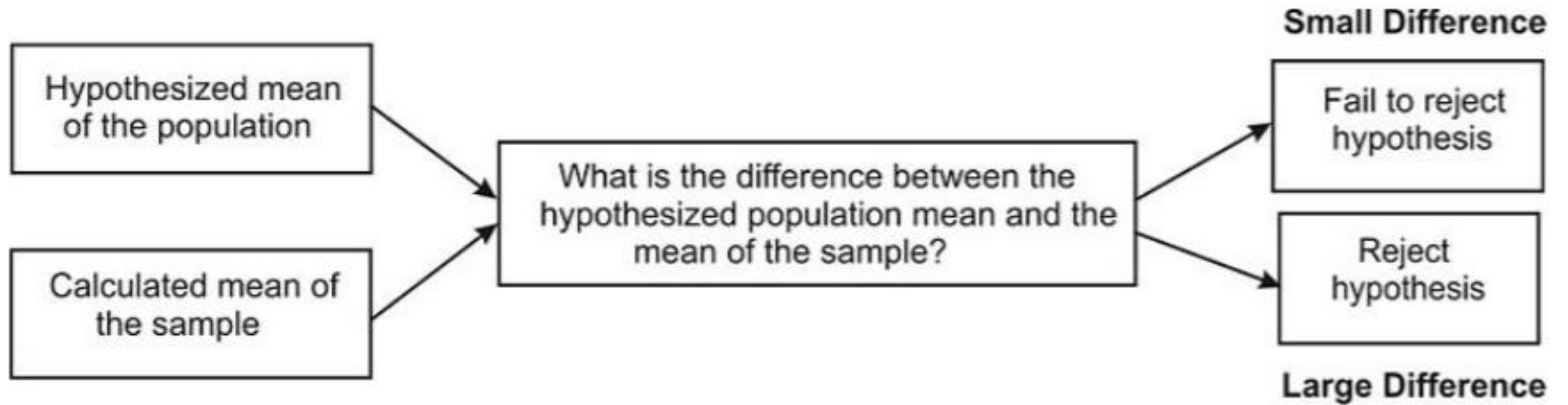


Motivation . . .

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.



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**).

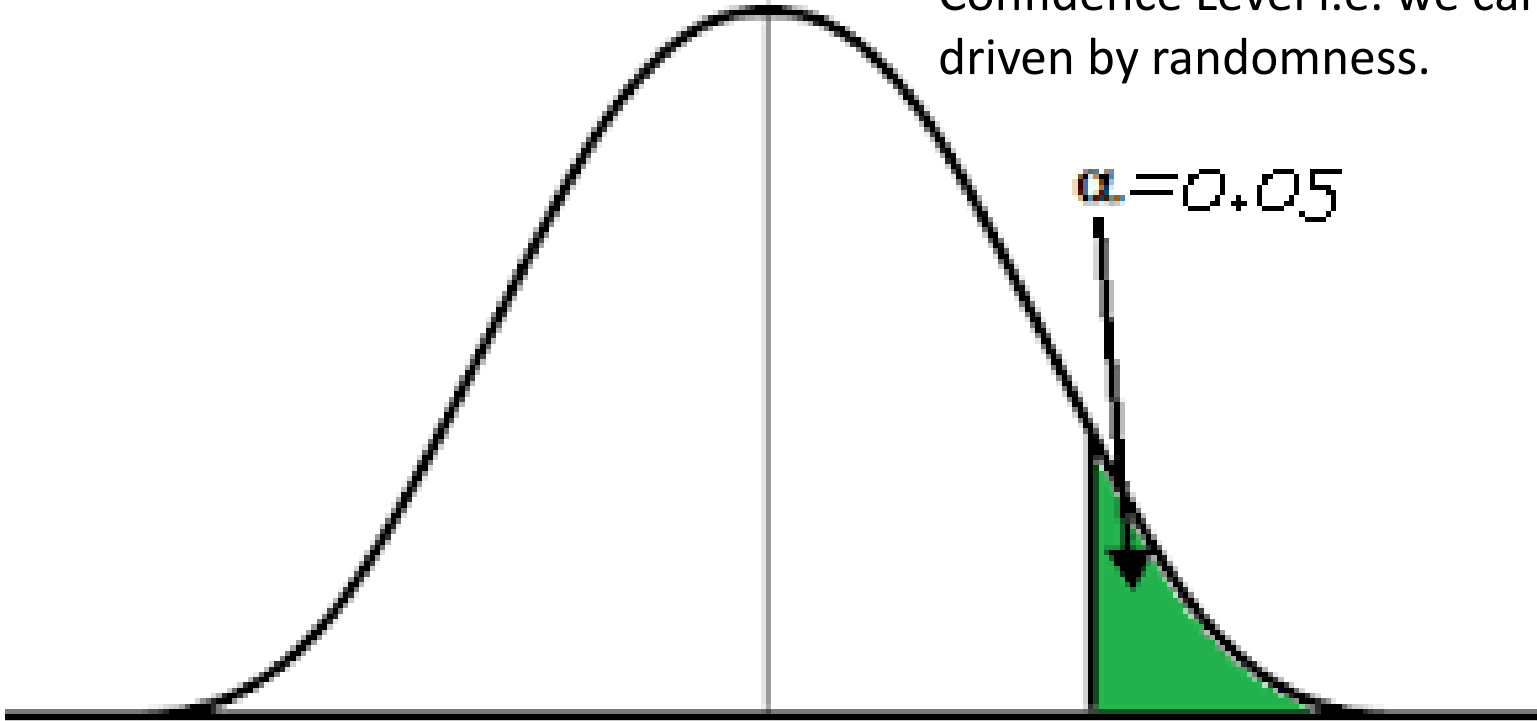
In general, we use the following: Source: corporatefinanceinstitute.com

- 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.

What is Significance Level?

This 5% is called **Significance Level** also known as alpha level (symbolized as α). It means that if random chance probability is less than 5% then we can conclude that there is difference in behavior of two different population. (1- Significance level) is also known as Confidence Level i.e. we can say that I am 95% confident that it is not driven by randomness.



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!

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

Original Claim

Decision	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)

❶ One-tailed (directional)

$$H_A: \rho > 0$$

$$H_A: \rho < 0$$

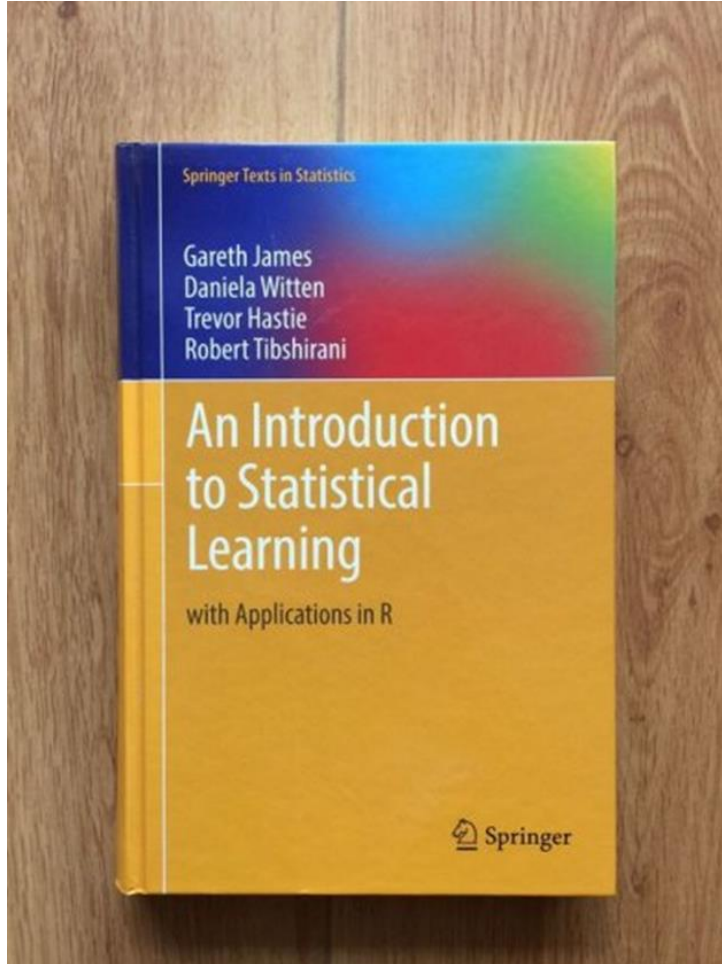


❷ Two-tailed (non directional)

$$H_A: \rho \neq 0$$



Statistical learning emphasizes the models and their interpretability, and precision and uncertainty.



The most common statistical tests include:

- Chi-square
- T-test
- ANOVA
- Correlation
- Linear Regression

Chi Square Test for Goodness of Fit

$$\text{chi square} = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where, O_i is the observed frequency

E_i is the expected frequency

$$E_i = N * p_i$$

degrees of freedom, $df = (r - 1) * (c - 1)$

where r = number of rows

c = number of columns

The R function `chisq.test()` can be used as follow:

`chisq.test(x, p)`

- statistic: the value the chi-squared test statistic.
- parameter: the degrees of freedom
- p.value: the p-value of the test
- observed: the observed count
- expected: the expected count

Analysis of Variance (ANOVA) is a statistical technique, commonly used to study differences between two or more group means.

```
> #medical
> anova(lm(StressReduction ~ Gender, dataMedical))
Analysis of Variance Table
```

```
Response: StressReduction
      Df Sum Sq Mean Sq    F value    Pr(>F)
Gender   1      0 0.00000 2.958e-31      1
Residuals 18     12 0.66667
```

```
> #mental
> anova(lm(StressReduction ~ Gender, dataMental))
Analysis of Variance Table
```

```
Response: StressReduction
      Df Sum Sq Mean Sq    F value    Pr(>F)
Gender   1      5 5.00000    7.5 0.01350 *
Residuals 18     12 0.66667
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> #physical
> anova(lm(StressReduction ~ Gender, dataPhysical))
Analysis of Variance Table
```

```
Response: StressReduction
      Df Sum Sq Mean Sq    F value    Pr(>F)
Gender   1     20 20.0000    30 3.345e-05 ***
Residuals 18     12 0.66667
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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