# HYPOTHESIS TEST

Assistant Professor Dr. Pornpimol Chaiwuttisak

# Definition of Hypotheses

- A hypothesis is a statement of the researcher's idea or guess.
- It is usually concerned with the parameters of the population
- Example: The hospital administrator may want to test the hypothesis that the average length of stay of patients admitted to the hospital in 5 days

# Statistical Hypotheses

- They are hypothesis that are stated in such a way that they may be evaluated by appropriate statistical techniques.
- To test a hypothesis the first thing we do is write down a statement – called the null hypothesis.
- The null hypothesis is often the opposite of the researcher's guess.
- A **null hypothesis $H_0$** is a statistical hypothesis that contains a statement of equality such as $\leq$, =, or $\geq$.
- A **alternative hypothesis $H_a$** is the complement of the null hypothesis. It is a statement that must be true if $H_0$ is false and contains a statement of inequality such as >, $\neq$, or <.

# Stating a Hypothesis

**Example:**

Write the claim as a mathematical sentence. State the null and alternative hypotheses and identify which represents the claim.

the average length of stay of patients admitted to the hospital in 5 days.

$\mu = 5$

Condition of equality

$H_0$: $\mu = 5$ (Claim)

$H_a$: $\mu \neq 5$

Complement of the null hypothesis

# Stating a Hypothesis

**Example:**

Write the claim as a mathematical sentence. State the null and alternative hypotheses and identify which represents the claim.

A manufacturer claims that its rechargeable batteries have an average life of at least 1,000 charges.

$\mu \geq 1000$

→ Condition of equality

$H_0$: $\mu \geq 1000$ (Claim)

$H_a$: $\mu < 1000$

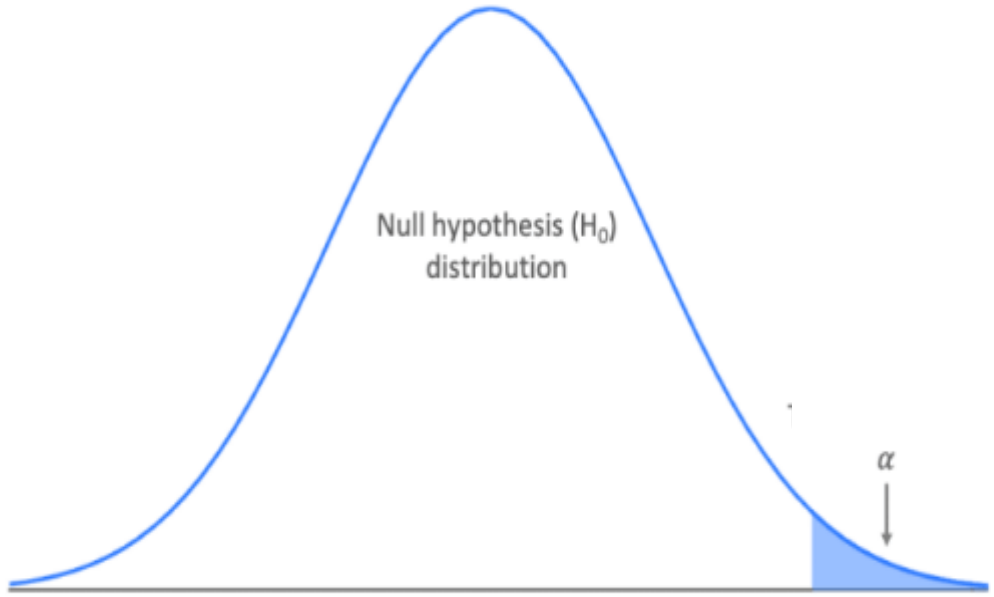
→ Complement of the null hypothesis

# Significance (1)

- Before carrying out any test we have to decide on a **significance level** which lets us determine at what point to **reject** the null hypothesis and accept the alternative hypothesis.
- The Level of Significance is a **probability of rejecting a true null hypothesis**.
- Conclusion. If $H_0$ is rejected, we conclude that $H_A$ is true. If $H_0$ is no rejected, we conclude that $H_0$ may be true

# Significance (2)

- Significance is based on the probability of a particular result.
- Statisticians have calculated the probability of all possible 'chance' events occurring.



# Critical Region

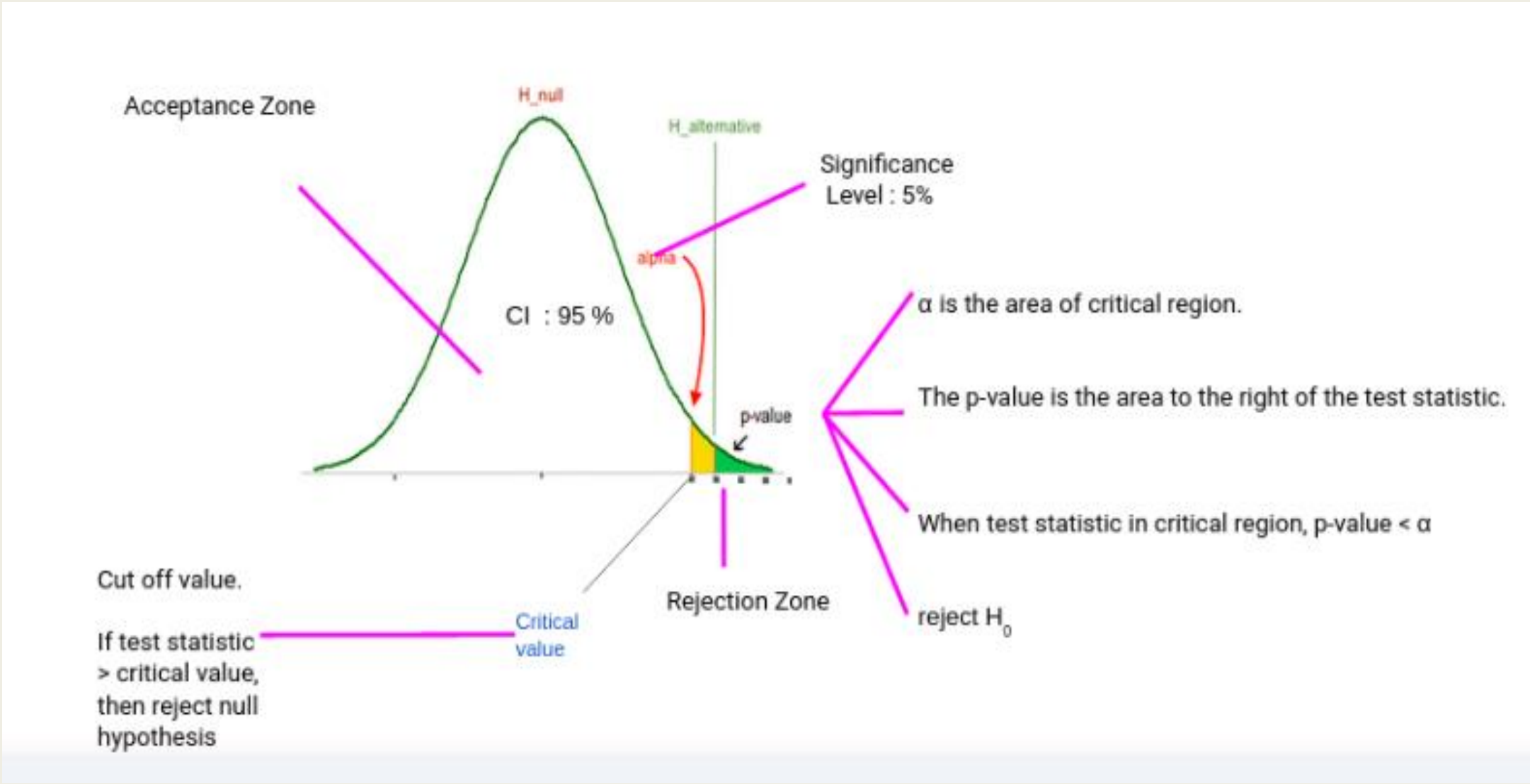
- Critical region is the part of the sample space that corresponds to the rejection of the null hypothesis, i.e. the set of possible values of the test statistic which are better explained by the alternative hypothesis.
- The significance level is the probability that the test statistic will fall within the critical region when the null hypothesis is assumed.

# P-Value

- P-value (the probability value) is the value p of the statistic used to test the null hypothesis.

  If $p < \alpha$ then we reject the null hypothesis.

  If $p \geq \alpha$ then we do not reject the null hypothesis

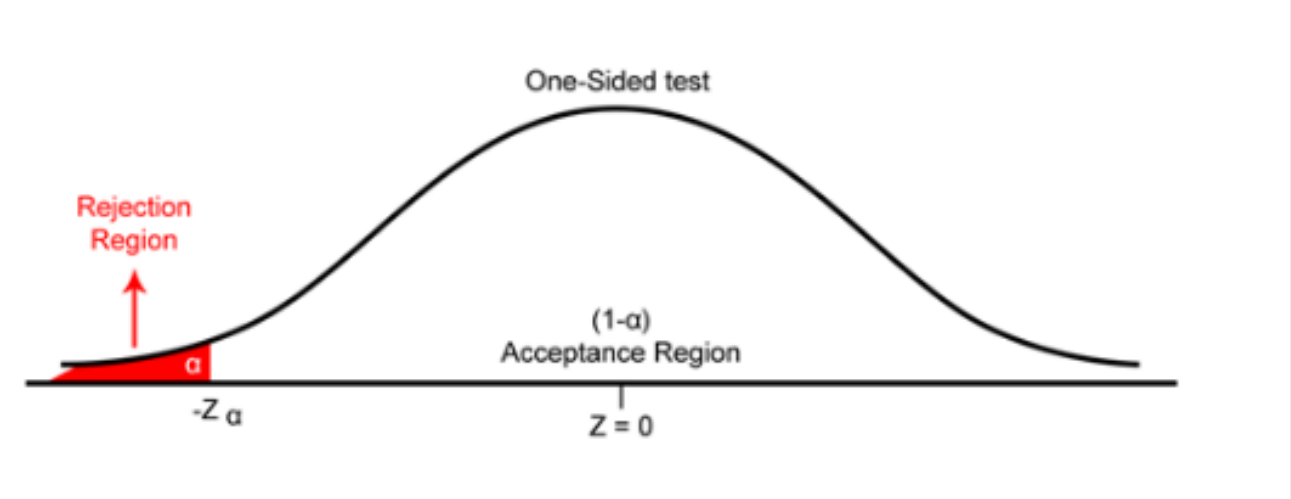

Acceptance Zone
H_null
H_alternative
Significance
Level : 5%
alpha
CI : 95 %
α is the area of critical region.
The p-value is the area to the right of the test statistic.
p-value
When test statistic in critical region, p-value < α
Cut off value.
Rejection Zone
reject H0
Critical
value
If test statistic
> critical value,
then reject null
hypothesis

# One-tailed hypothesis testing

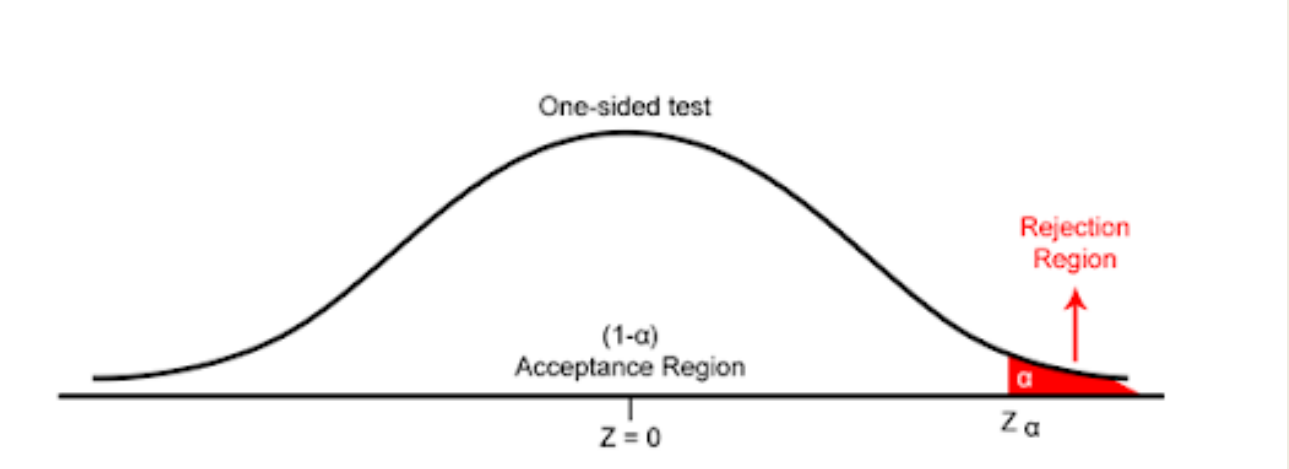
**One-tailed hypothesis testing** specifies a direction of the statistical test. For example to test whether cloud seeding increases the average annual rainfall in an area which usually has an average annual rainfall of 20 cm, we define the null and alternative hypotheses as follows, where $\mu$ represents the average rainfall after cloud seeding.

- $H_0$: $\mu \leq 20$ (i.e. average rainfall does not increase after cloud seeding)
- $H_1$: $\mu > 20$ (i.e. average rainfall increases after cloud seeding

# One-tailed hypothesis testing




Left-tailed test



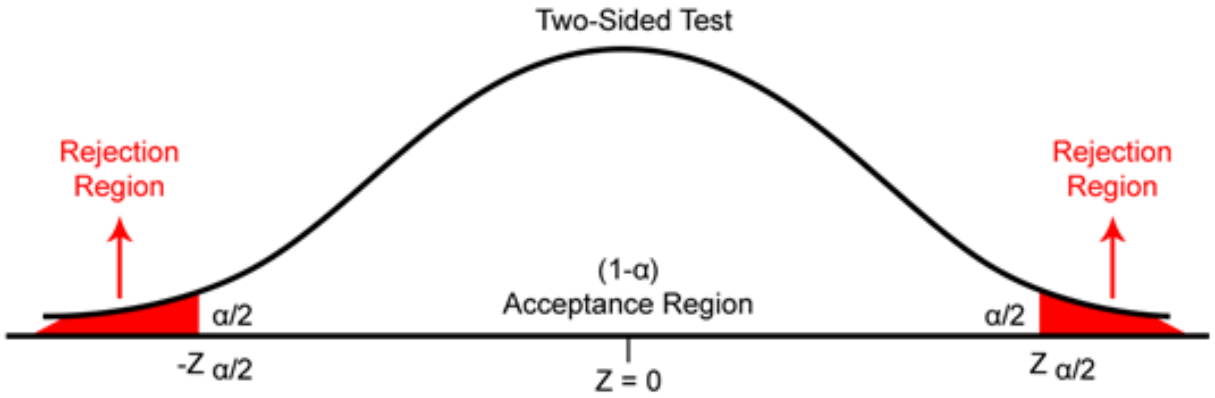

Right-tailed test

# Two-tailed hypothesis testing

**Two-tailed hypothesis testing** doesn't specify a direction of the test. For the cloud seeding example, it is more common to use a two-tailed test. Here the null and alternative hypotheses are as follows.

H0: $\mu = 20$

H1: $\mu \neq 20$



# Type of Error

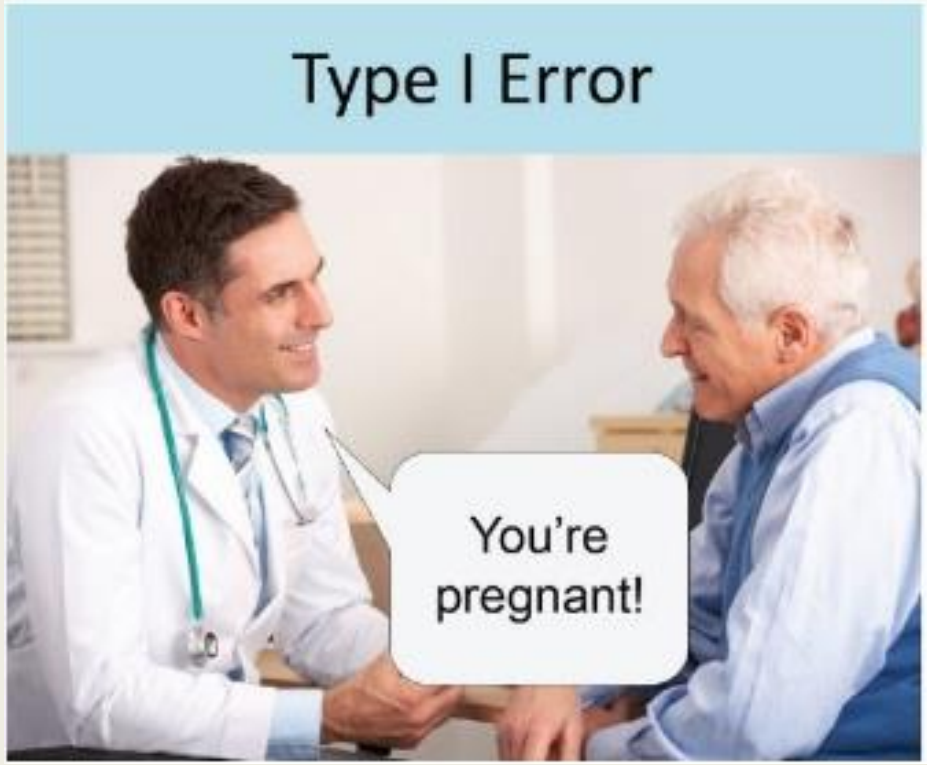
- When we make a conclusion from a statistical test there are two types of errors that we could make. They are called: Type I and Type II Errors
  - Type I error – reject $H_0$ when $H_0$ is true.
  - Type II error – do not reject $H_0$ when $H_0$ is false.
- Results of a statistical test:

| | $H_0$ is True | $H_0$ is False |
|---|---|---|
| **Reject $H_0$** | Type I Error (False Positive) Probability = $\alpha$  | Correct Decision (True Positive) Probability = $1-\beta$  |
| **Do not Reject $H_0$** | Correct Decision (True Negative) Probability = $1-\alpha$  | Type II Error (False Negative) Probability = $\beta$  |

# Type of Error-Example

$H_0$: you are not pregnant

$H_1$: you are pregnant

| $H_0$ is true | $H_0$ is false |
| --- | --- |
|  |  |

# Type of Error-Example

$H_0$: you don't have coronavirus

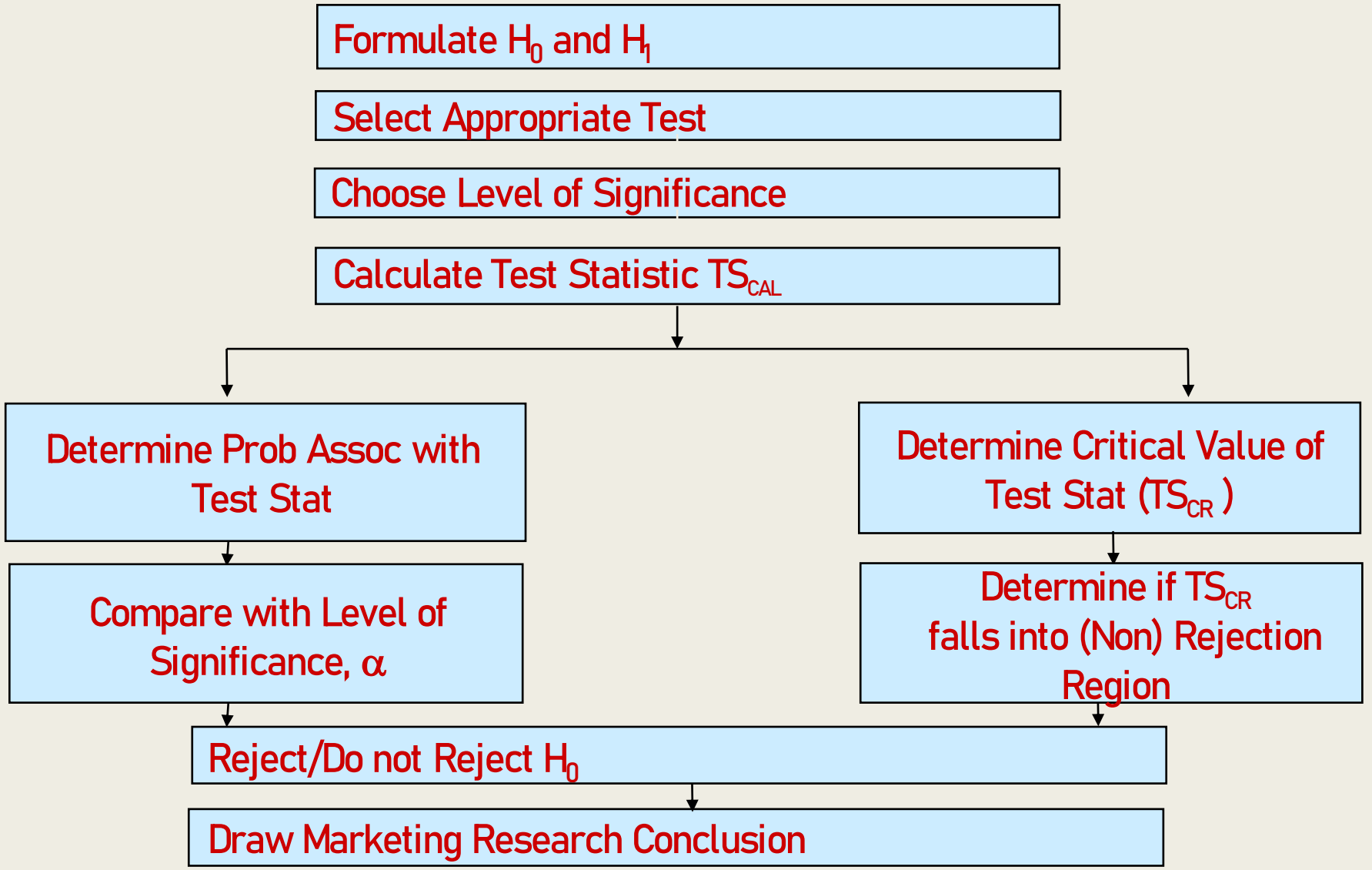
$H_1$: you have coronavirus

**Example: Type I vs Type II error**

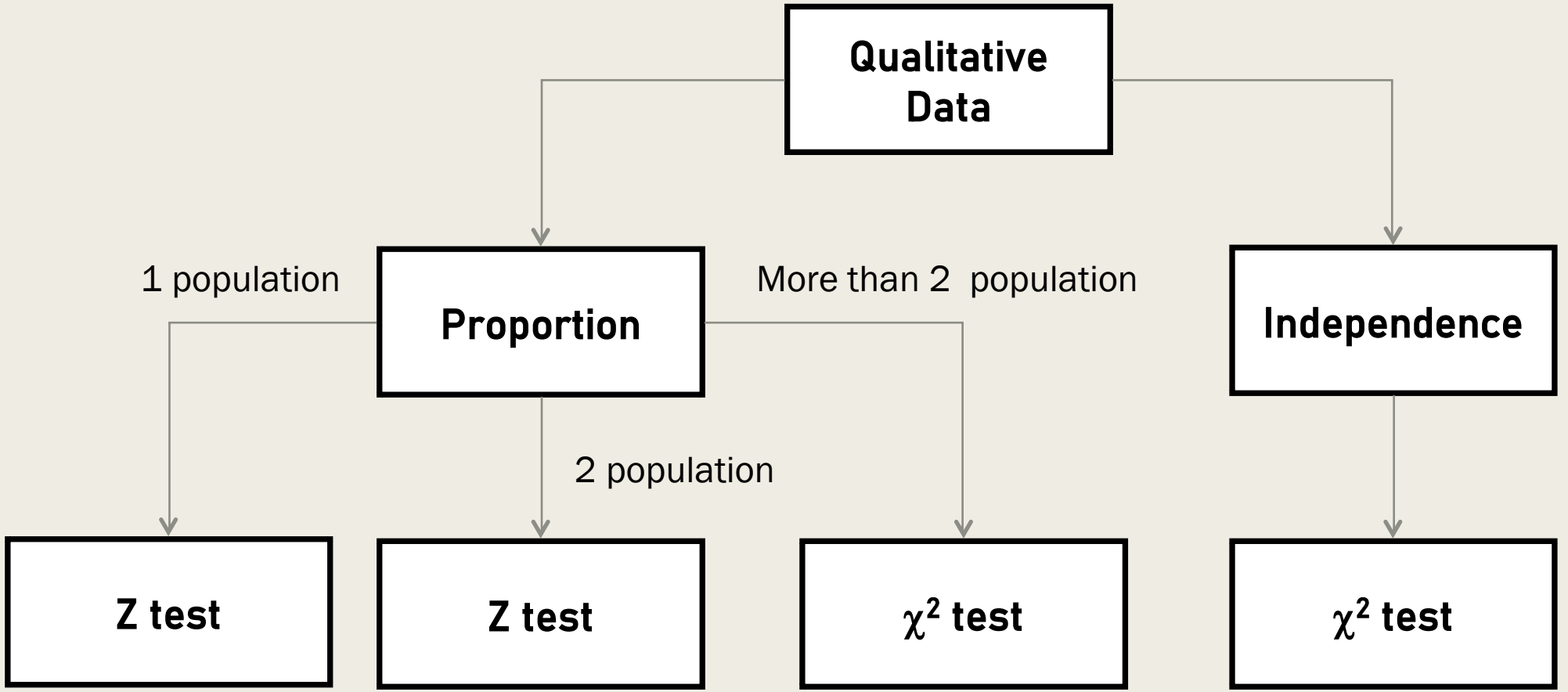
You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

- Type I error (false positive): the test result says you have coronavirus, but you actually don't.
- Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

# Steps for Hypothesis Testing



# Hypothesis Tests : Qualitative Data



# Hypotheses for Two Proportions

| Hypothesis | Research Question | | |
|---|---|---|---|
| | No Difference<br>Any Difference | Pop1 ≥ Pop2<br>Pop 1 < Pop2 | Pop 1 ≤ Pop 2<br>Pop1 > Pop2 |
| $H_0$ | $p_1 - p_2 = 0$ | $p_1 - p_2 \geq 0$ | $p_1 - p_2 \leq 0$ |
| $H_a$ | $P_1 - p_2 \neq 0$ | $P_1 - p_2 < 0$ | $P_1 - p_2 > 0$ |

# Z Test for Difference in Two Proportions

1. Assumptions
   - Populations are Independent
   - Populations follow Binomial Distribution
   - Normal Approximation can be used for large samples (All Expected Counts $\geq 5$)
2. Z-Test Statistic for Two Proportions

$$Z \cong \frac{\left(\hat{p}_1 - \hat{p}_2\right) - \left(p_1 - p_2\right)}{\sqrt{\hat{p} \cdot \left(1 - \hat{p}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

# Example: Titanic


- The ship Titanic sank in 1912 with the loss of most of its passengers
- 809 of the 1,309 passengers and crew died = 61.8%
- **Research question:** Did class (of travel) affect survival?

# Chi squared Test?

- **Null:** There is **NO** association between class and survival ($\rho = 0$)
- **Alternative:** There **IS** an association between class and survival ($\rho \neq 0$)

3 x 2 contingency table

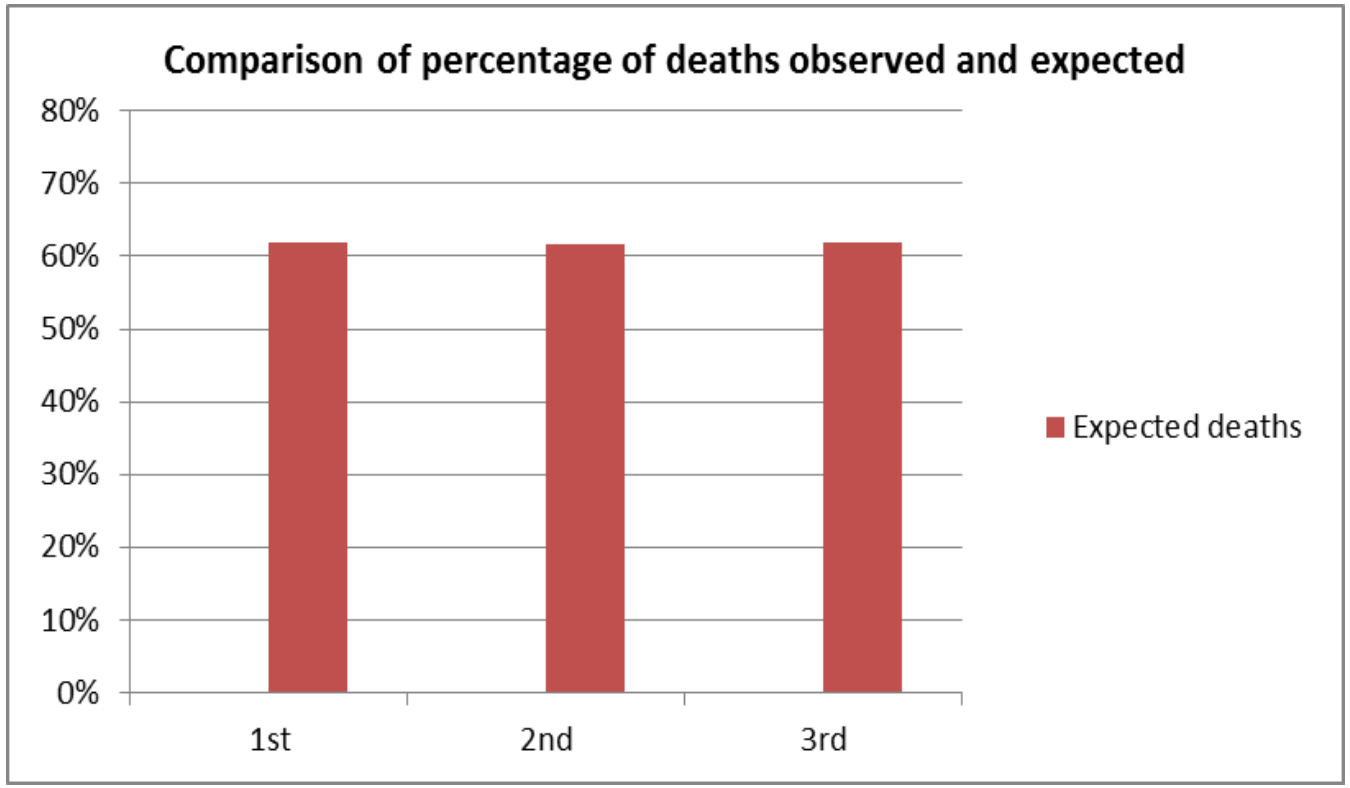
**Class * Survived? Crosstabulation**

Count

| | | Survived? | | |
|---|---|---|---|---|
| | | Died | Survived | Total |
| Class | 1st | 123 | 200 | 323 |
| | 2nd | 158 | 119 | 277 |
| | 3rd | 528 | 181 | 709 |
| Total | | 809 | 500 | 1309 |

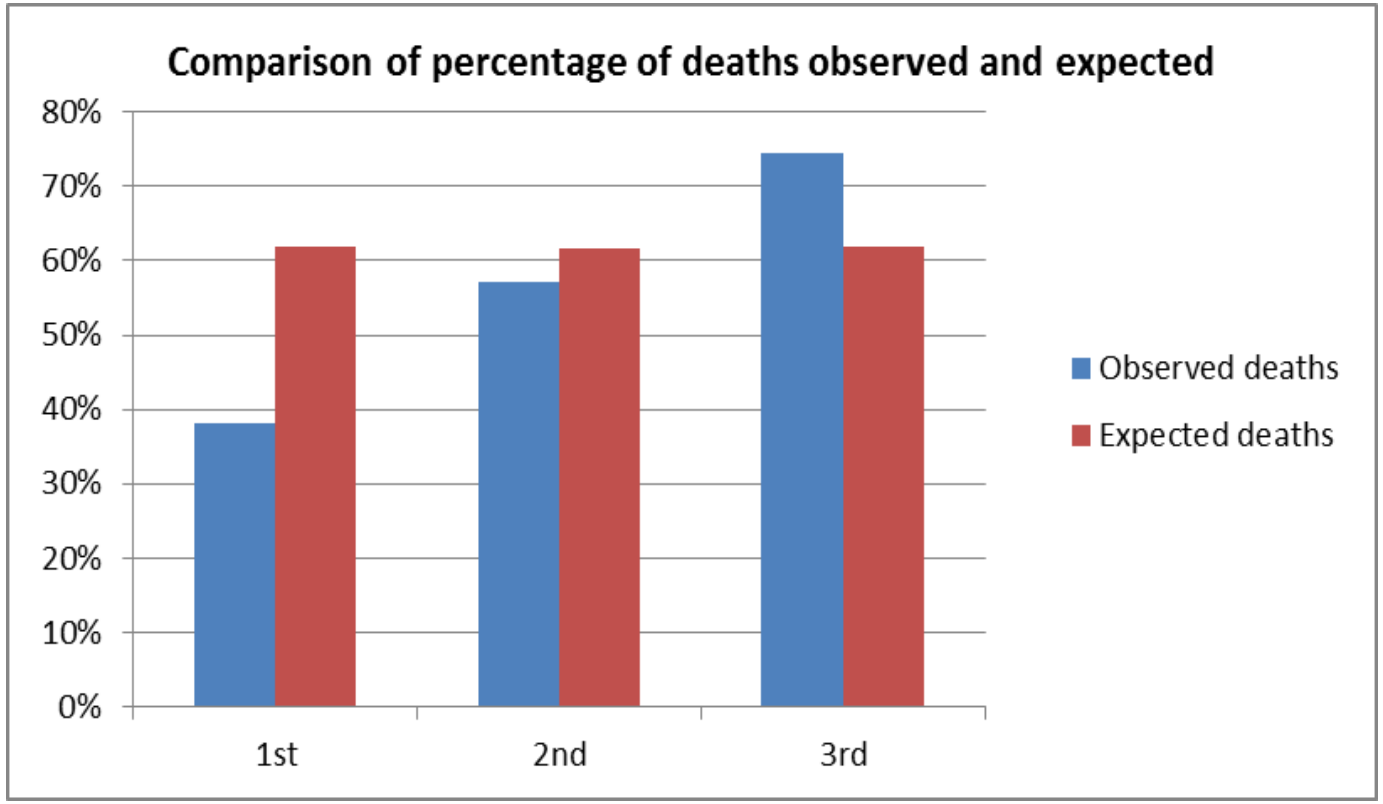
# What would be expected if the null is true?

- Same proportion of people would have died in each class!
- Overall, 809 people died out of 1309 = 61.8%



# What would be expected if the null is true?

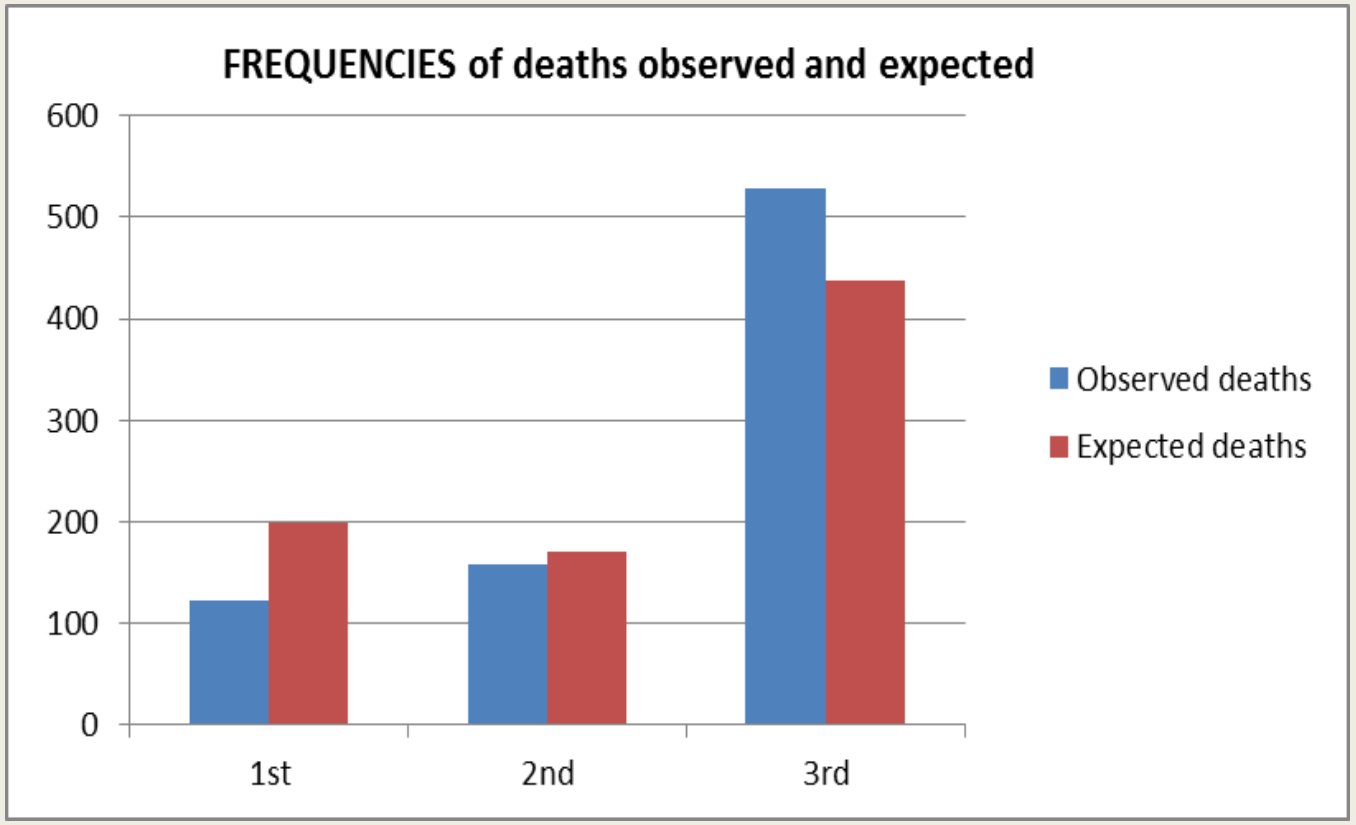
- Same proportion of people would have died in each class!
- Overall, 809 people died out of 1309 = 61.8%



# Expected Value

- The expected value is the average result we expect.
- It is the product of the probability times the number of observations $E = p \times n$
- A really useful case is the where the probability of all cases is equal.
- For example, in fair coin tosses, the probability of Heads =1/2. If we flip a coin 24 times, we EXPECT ½ x 24 = 12 Heads
- Equal probability cases are usually the basis of the "Null Hypothesis"

# Chi-Squared Test Actually Compares Observed and Expected Frequencies




Expected number dying in each class = 0.618 * no. in class

# Chi-squared test statistic

- The chi-squared test is used when we want to see if two categorical variables are related
- The test statistic for the Chi-squared test uses the sum of the squared differences between each pair of observed (O) and expected values (E)

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

# Using SPSS

Analyse → Descriptive Statistics → Crosstabs

Click on 'Statistics' button & select Chi-squared

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 127.859[a] | 2 | .000 |
| Likelihood Ratio | 127.765 | 2 | .000 |
| Linear-by-Linear Association | 127.709 | 1 | .000 |
| N of Valid Cases | 1309 | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 105.81.

**Test Statistic = 127.859**

**p- value**
**p < 0.001**

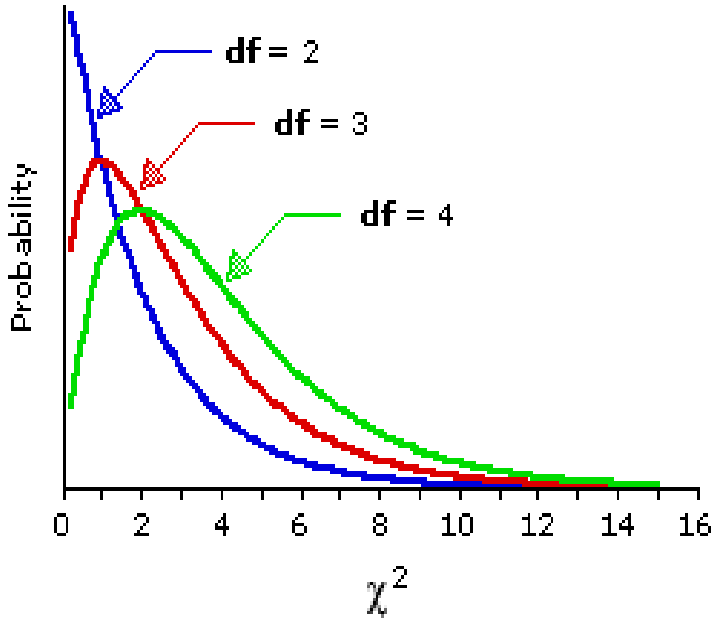
Note: Double clicking on the output will display the p-value to more decimal places

# Hypothesis Testing: Decision Rule

- We can use statistical software to undertake a hypothesis test e.g. SPSS
- One part of the output is the p-value (P)
- If P < **0.05 reject** $H_0$ => **Evidence** of $H_A$ being true (i.e. **IS** association)
- If P > **0.05 do not** reject $H_0$ (i.e. **NO** association)

# Chi squared distribution

- The p-value is calculated using the Chi-squared distribution for this test
- Chi-squared is a skewed distribution which varies depending on the degrees of freedom




Testing relationships between 2:
v = degrees of freedom
(no. of rows – 1) x (no. of columns – 1)

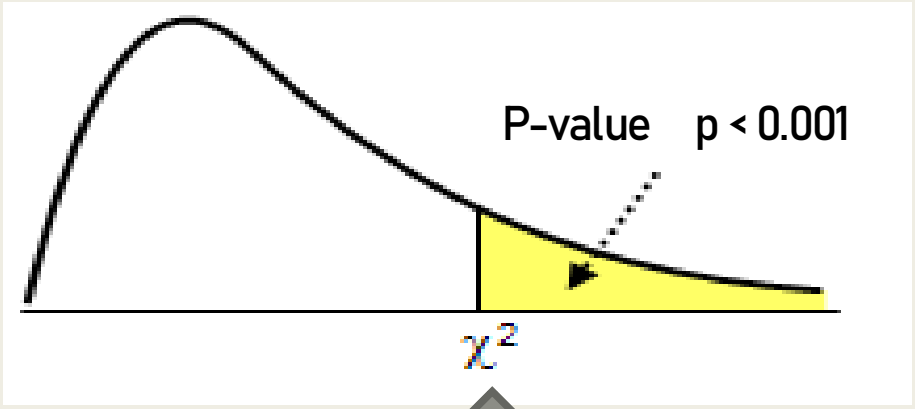
*Note: One sample test:*
*v = df = outcomes – 1*

# What's a p-value?
# The technical answer!

Probability of getting a test statistic at least as extreme as the one calculated **if the null is true**

In Titanic example, the probability of getting a test statistic of 127.859 or above **(if the null is true)** is < 0.001

Distribution of test statistics



# Contingency tables

Which percentages are better for investigating whether class had an effect on survival?

**Column**

**Class * Survived? Crosstabulation**

| | | | Survived? | | |
|---|---|---|---|---|---|
| | | | Died | Survived | Total |
| Class | 1st | Count | 123 | 200 | 323 |
| | | % within Survived? | 15.2% | 40.0% | 24.7% |
| | 2nd | Count | 158 | 119 | 277 |
| | | % within Survived? | 19.5% | 23.8% | 21.2% |
| | 3rd | Count | 528 | 181 | 709 |
| | | % within Survived? | 65.3% | 36.2% | 54.2% |
| Total | | Count | 809 | 500 | 1309 |
| | | % within Survived? | 100.0% | 100.0% | 100.0% |

**Row**

**Class * Survived? Crosstabulation**

| | | | Survived? | | |
|---|---|---|---|---|---|
| | | | Died | Survived | Total |
| Class | 1st | Count | 123 | 200 | 323 |
| | | % within Class | 38.1% | 61.9% | 100.0% |
| | 2nd | Count | 158 | 119 | 277 |
| | | % within Class | 57.0% | 43.0% | 100.0% |
| | 3rd | Count | 528 | 181 | 709 |
| | | % within Class | 74.5% | 25.5% | 100.0% |
| Total | | Count | 809 | 500 | 1309 |
| | | % within Class | 61.8% | 38.2% | 100.0% |

65.3% of those who died were in 3rd class  74.5% of those in 3rd class died

# Did class affect survival? Solution

%'s within each class are preferable due to different class frequencies

**pclass * survived Crosstabulation**

| | | | survived | | Total |
|---|---|---|---|---|---|
| | | | Died | Survived | |
| pclass | 1st | Count | 123 | 200 | 323 |
| | | % within pclass | 38.1% | 61.9% | 100.0% |
| | 2nd | Count | 158 | 119 | 277 |
| | | % within pclass | 57.0% | 43.0% | 100.0% |
| | 3rd | Count | 528 | 181 | 709 |
| | | % within pclass | 74.5% | 25.5% | 100.0% |
| Total | | Count | 809 | 500 | 1309 |
| | | % within pclass | 61.8% | 38.2% | 100.0% |

# Did class affect survival? Solution

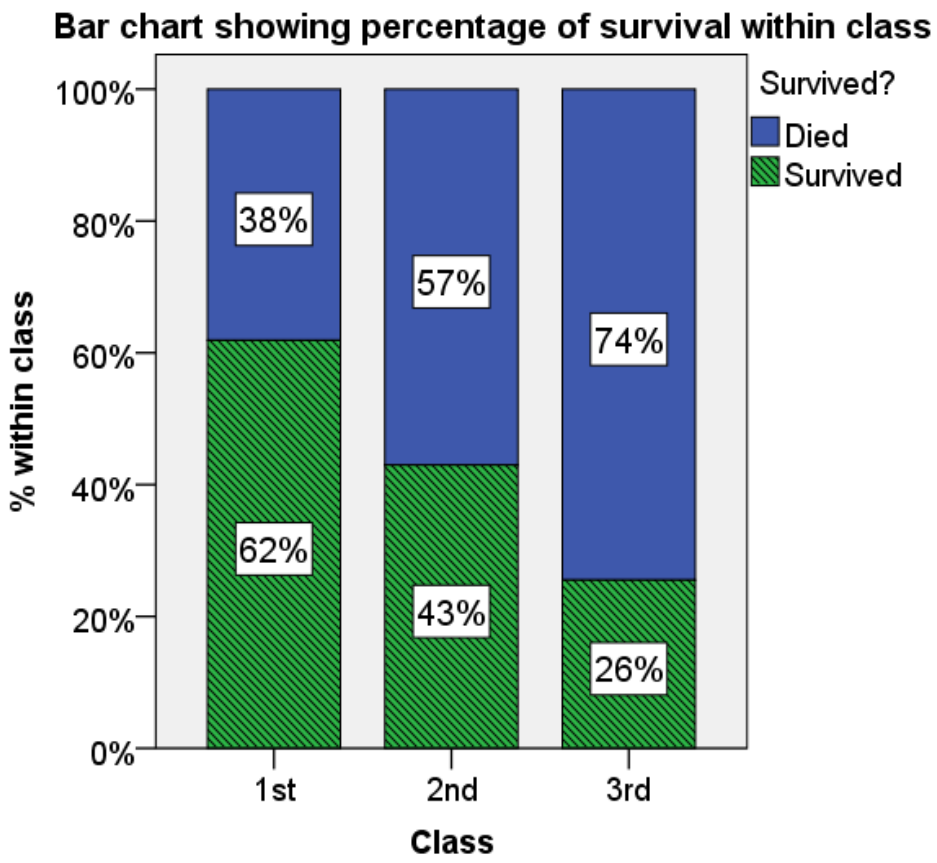



*Figure 1: Bar chart showing % of passengers surviving within each class*

Data collected on 1309 passengers aboard the Titanic was used to investigate whether class had an effect on chances of survival. There was evidence ($\chi^2_2$=127.86, $p < 0.001$) to suggest that there is an association between class and survival.

*Figure 1* shows that class and chances of survival were related. As class decreases, the percentage of those surviving also decreases from 62% in 1st Class to 26% in 3rd Class.

# Low EXPECTED Cell Counts with the Chi-squared test

| | Died | Survived | Total |
|---|---|---|---|
| 1st Class | 200 | 123 | 323 |
| 2nd Class | 171 | 106 | 277 |
| 3rd Class | 438 | 271 | 709 |
| Total | 809 | 500 | 1,309 |

We have no cells with expected counts below 5

SPSS Output

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 127.859[a] | 2 | .000 |
| Likelihood Ratio | 127.765 | 2 | .000 |
| Linear-by-Linear Association | 127.709 | 1 | .000 |
| N of Valid Cases | 1309 | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 105.81.

# Test Statistics

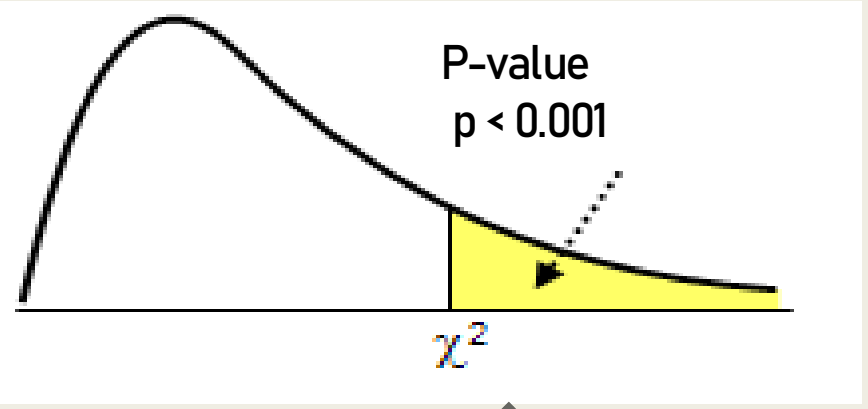
- To check the null hypothesis we calculate a figure known as a test statistic, which is based on data from our samples.
- Different types of problems require different test statistics. Values for comparison to our data have all been put into statistical tables.
- All we need to do is to calculate our value and compare it with the value in the table to get our answer.

# Interpretation

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 127.859[a] | 2 | .000 |

Since $p < 0.05$ we reject the null

There is evidence ($\chi^2_2$=127.86, $p < 0.001$) to suggest that there is an association between class and survival



# Low Cell Counts with the Chi-squared test

- Check no. of cells with EXPECTED counts less than 5
- SPSS reports the % of cells with an expected count <5
  - *If more than 20% then the test statistic does not approximate a chi-squared distribution very well*
  - *If any expected cell counts are <1 then cannot use the chi-squared distribution*
- In either case if have a 2x2 table use **Fishers' Exact test** (SPSS reports this for 2x2 tables)
- In larger tables (3x2 etc.) combine categories to make cell counts larger (providing it's meaningful)

# Brief historical interlude

- Karl Pearson (b. London1857; d. London 1936)
- Considered the founder of mathematical statistics
- Developed the chi-square test (published July 1900)
- Coined the term "standard deviation"
- Developed the product-moment correlation
- Now you know who to blame

# Chi-Square as a Statistical Test

- **Chi-square test**: an inferential statistics technique designed to test for significant relationships between two variables organized in a bivariate table.

- Chi-square requires **no assumptions** about the shape of the population distribution from which a sample is drawn.

# Additional Uses of Chi-Square Distribution

we will cover two additional testing procedures by using the Chi-Square distribution.

- **Goodness of Fit Test**
- **Test of Independence**

# Additional Uses of Chi-Square Distribution

## Goodness of Fit Test

**How close are sample results to the expected results?**

**Example:** In tossing a coin, you expect half heads and half tails. You tossed a coin 100 times. You expected 50 heads and 50 tails. However, you obtained 48 heads and 52 tails. Are 48 heads and 52 tails close enough to call the coin fair?

# Additional Uses of Chi-Square Distribution

## Test of Independence

Are two variables of interest independent of each other?

## Examples:

- Is starting salary of fresh graduates independent of graduates' field of study?
- Is beer preference independent of the gender of the beer drinker?

# Hypothesis (Goodness of Fit) Test for Proportions of a Multinomial Population

1. Set up the null and alternative hypotheses.

2. Select a random sample and record the observed frequency, $f_i$ , for each of the k categories.

3. Assuming $H_0$ is true, compute the expected frequency, ei , in each category by multiplying the category probability by the sample size.

# Hypothesis (Goodness of Fit) Test for Proportions of a Multinomial Population

4. Compute the value of the test statistic.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

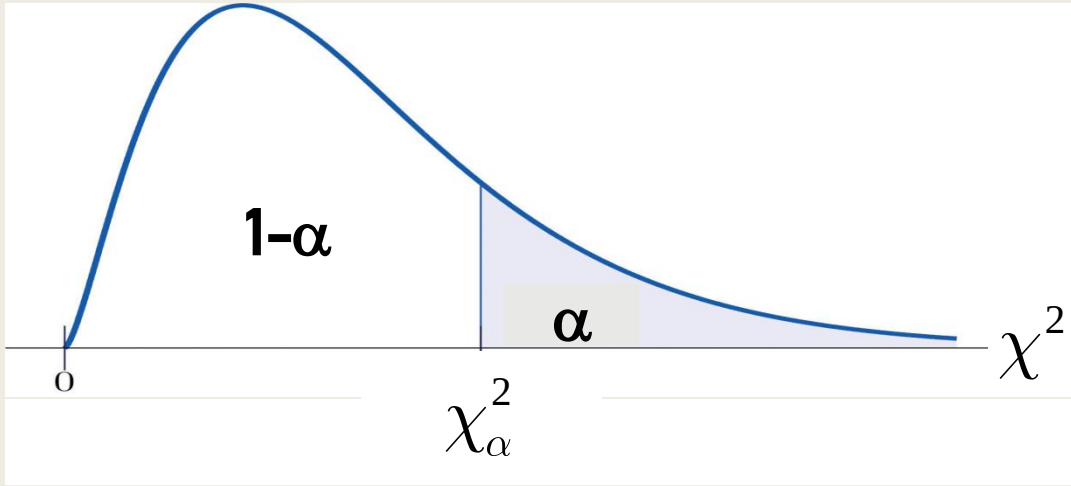
where:

$f_i$ = observed frequency for category i
$e_i$ = expected frequency for category i
k = number of categories

**Note: The test statistic has a chi-square distribution with k – 1 df provided that the expected frequencies are 5 or more for all categories.**

# Chi-square

- Here's the distribution:




The value of the chi-square random variable $\chi^2$ with df=k that cuts off a right tail of area $\alpha$ is denoted $\chi^2_\alpha$ and is called **a critical value**

# Hypothesis (Goodness of Fit) Test for Proportions of a Multinomial Population

5. Rejection rule:

**p-value approach:** **Reject $H_0$ if p-value $< \alpha$**

**Critical value approach:** **Reject $H_0$ if $\chi^2 \geq \chi^2_\alpha$**

where $\alpha$ is the significance level and
there are k – 1 degrees of freedom

# Multinomial Distribution Goodness of Fit Test

**Example: Finger Lakes Homes (A)**

Finger Lakes Homes manufactures four models of prefabricated homes, two-story colonial, a log cabin, a split-level, and an A-frame. To help in production planning, management would like to determine if previous customer purchases indicate that there is a preference in the style selected.

# Multinomial Distribution Goodness of Fit Test

**Example: Finger Lakes Homes (A)**

The number of homes sold of each model for 100 sales over the past two years is shown below.

| Model | Colonial | Log | Split Level | A Frame |
|---|---|---|---|---|
| # Sold | 30 | 20 | 35 | 15 |

# Multinomial Distribution Goodness of Fit Test

- Hypotheses

$H_0$: $p_C = p_L = p_S = p_A = .25$

$H_a$: The population proportions are not

$p_C = .25$, $p_L = .25$, $p_S = .25$, and $p_A = .25$

where:

$p_C$ = population proportion that purchase a colonial

$p_L$ = population proportion that purchase a log cabin

$p_S$ = population proportion that purchase a split-level

$p_A$ = population proportion that purchase an A-frame

# Hypotheses

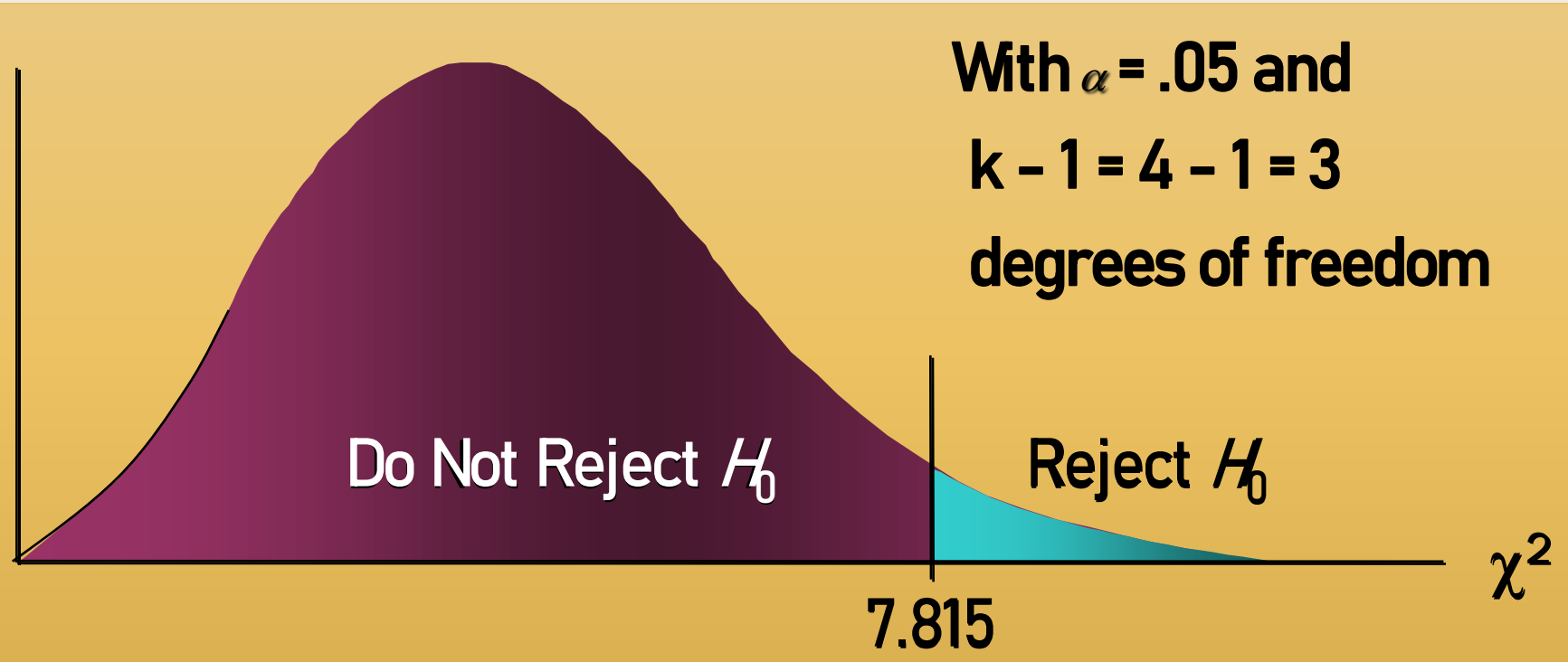
$H_0$ : There is no preference in the home styles or all home styles have equal preferences

$H_a$ : All home styles do not have equal preferences

# Multinomial Distribution Goodness of Fit Test

- Rejection Rule

Reject $H_0$ if p-value < .05 or $\chi^2 > 7.815$.



# Multinomial Distribution Goodness of Fit Test

- **Expected Frequencies**

$e_1 = .25(100) = 25$ $e_2 = .25(100) = 25$

$e_3 = .25(100) = 25$ $e_4 = .25(100) = 25$

- **Test Statistic**

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(35-25)^2}{25} + \frac{(15-25)^2}{25}$$

$$= 1 + 1 + 4 + 4$$

$$= 10$$

# Multinomial Distribution Goodness of Fit Test

- Conclusion Using the p-Value Approach

| Area in Upper Tail | .10 | .05 | .025 | .01 | .005 |
|---|---|---|---|---|---|
| $\chi^2$ Value (df = 3) | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |

Because $\chi^2$ = 10 is between 9.348 and 11.345, the area in the upper tail of the distribution is between .025 and .01.

**The p-value < a . We can reject the null hypothesis.**

# Multinomial Distribution Goodness of Fit Test

- Conclusion Using the Critical Value Approach

$$\chi^2 = 10 > 7.815$$

**We reject, at the .05 level of significance, the assumption that there is no home style preference.**

# Sample Problem

Problem

$H_0$ : Viewing audience proportions are the same

$H_a$ : Viewing audience proportions are not the same

| Category | Frequencies Observed | Frequencies Expected | Differences Squared/Expected Frequencies |
|---|---|---|---|
| ABC | 95 | 87 | 0.74 |
| CBA | 70 | 84 | 2.33 |
| NBC | 89 | 75 | 2.61 |
| Independents | 46 | 54 | 1.19 |

# Sample Problem Continued

Decision Rule: Reject Null Hypothesis if $\chi^2 > \chi^2_{0.05, 3}$

$$\chi^2 < 6.87 > \chi^2_{0.05, 3} = 7.815$$

Decision: Do not reject the null hypothesis

Interpretation: Viewing audience proportions have not changed.

# Chi-squared Test of Independence

- Two random variables x and y are called **independent** if the probability distribution of one variable is not affected by the presence of another.
- Assume $f_{ij}$ is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume $e_{ij}$ to be the corresponding expected count if x and y are independent.
- The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level $\alpha$.

# Test of Independence: Contingency Tables

1. Set up the null and alternative hypotheses.
2. Select a random sample and record the observed frequency, $f_{ij}$ , for each cell of the contingency table.
3. Compute the expected frequency, $e_{ij}$ , for each cell.

$$e_{ij} = \frac{(\text{Row } i \text{ Total})(\text{Column } j \text{ Total})}{\text{Sample Size}}$$

# Test of Independence: Contingency Tables

4. Compute the test statistic.

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

5. Determine the rejection rule.

Reject $H_0$ if p -value $< \alpha$ or $\chi^2 \geq \chi^2_\alpha$

where $\alpha$ is the significance level and, with n rows and m columns, there are $(n - 1)(m - 1)$ degrees of freedom.

# Contingency Table (Independence) Test

Example: Finger Lakes Homes (B)

Each home sold by Finger Lakes Homes can be classified according to price and to style. Finger Lakes' manager would like to determine if the price of the home and the style of the home are independent variables.

# Contingency Table (Independence) Test

Example: Finger Lakes Homes (B)

The number of homes sold for each model and price for the past two years is shown below. For convenience, the price of the home is listed as either $99,000 or less or more than $99,000.

| Price | Colonial | Log | Split-Level | A-Frame |
|---|---|---|---|---|
| ≤ $99,000 | 18 | 6 | 19 | 12 |
| > $99,000 | 12 | 14 | 16 | 3 |

# Contingency Table (Independence) Test

Hypotheses

$H_0$: Price of the home is independent of the style of the home that is purchased

$H_a$: Price of the home is not independent of the style of the home that is purchased

# Contingency Table (Independence) Test

## Expected Frequencies

| Price | Colonial | Log | Split-Level | A-Frame | Total |
|---|---|---|---|---|---|
| ≤ $99K | 18 | 6 | 19 | 12 | 55 |
| > $99K | 12 | 14 | 16 | 3 | 45 |
| Total | 30 | 20 | 35 | 15 | 100 |

# Contingency Table (Independence) Test

## Rejection Rule

With $\alpha = .05$ and $(2 - 1)(4 - 1) = 3$ d.f., $\chi^2_{0.05} \geq 7.815$

Reject $H_0$ if p-value < .05 or $\chi^2 \geq 7.815$

## Test Statistic

$$\chi^2 = \frac{(18 - 16.5)^2}{16.5} + \frac{(6 - 11)^2}{11} + \ldots + \frac{(3 - 6.75)^2}{6.75}$$

$$= .1364 + 2.2727 + \ldots + 2.0833 = 9.149$$

# Contingency Table (Independence) Test

- Conclusion Using the p-Value Approach

| Area in Upper Tail | .10 | .05 | .025 | .01 | .005 |
|---|---|---|---|---|---|
| $\chi^2$ Value (df = 3) | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |

Because $\chi^2 = 9.145$ is between 7.815 and 9.348, the area in the upper tail of the distribution is between .05 and .025.

The p-value < $\alpha$ . We can reject the null hypothesis.

# Contingency Table (Independence) Test

- Conclusion Using the Critical Value Approach

$$\chi^2 = 9.145 \geq 7.815$$

We reject, at the .05 level of significance, the assumption that the price of the home is independent of the style of home that is purchased.

# Testing independence with chi-square

- In a certain town, there are about 1 million voters. An SRS of 10,000 was chosen to study the relationship between gender and participation.

| | Men | Women | Total |
|---|---|---|---|
| Voted | 2792 | 3591 | 6383 |
| Didn't vote | 1486 | 2131 | 3617 |
| Total | 4278 | 5722 | 10000 |

- Are gender and voting independent?
- We can answer this with a chi-square test.
- First, we need the expected values for each cell.

# Expected values

- The expected value for each cell is simply:

$$E = \frac{\text{row total} \times \text{column total}}{\text{total}}$$

- For men who voted, this is:

$$E = \frac{6383 \times 4278}{10000}$$
$$= 2730.6$$

- Thus, we have the following:

# Observed and expected values

| | Observed | | Expected | | Difference | |
|---|---|---|---|---|---|---|
| | Men | Women | Men | Women | Men | Women |
| Vote | 2792 | 3591 | 2730.6 | 3652.4 | 61.4 | -61.4 |
| Didn't vote | 1486 | 2131 | 1547.4 | 2069.6 | -61.4 | 61.4 |

We calculate the test statistic in the same way as before:

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

# The test statistic

$$\chi^2 = \frac{(61.4)^2}{2730.6} + \frac{(-61.4)^2}{3652.4} + \frac{(-61.4)^2}{1547.4} + \frac{(61.4)^2}{2069.6}$$
$$\approx 6.7$$

The degrees of freedom are:

$$d = (\#\text{row} - 1) \times (\#\text{columns} - 1)$$
$$= (2-1) \times (2-1)$$
$$= 1$$

# The Interpretation

- Because $\chi 2$ (Cal.) = 6.7 > $\chi 2$ (Table)= 3.84

| Degrees of freedom | 99% | 95% | 90% | 70% | 50% | 30% | 10% | 5% | 1% |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00016 | 0.0039 | 0.016 | 0.15 | 0.46 | 1.07 | 2.71 | 3.84 | 6.64 |
| 2 | 0.020 | 0.10 | 0.21 | 0.71 | 1.39 | 2.41 | 4.60 | 5.99 | 9.21 |
| 3 | 0.12 | 0.35 | 0.58 | 1.42 | 2.37 | 3.67 | 6.25 | 7.82 | 11.34 |
| 4 | 0.30 | 0.71 | 1.06 | 2.20 | 3.36 | 4.88 | | | |

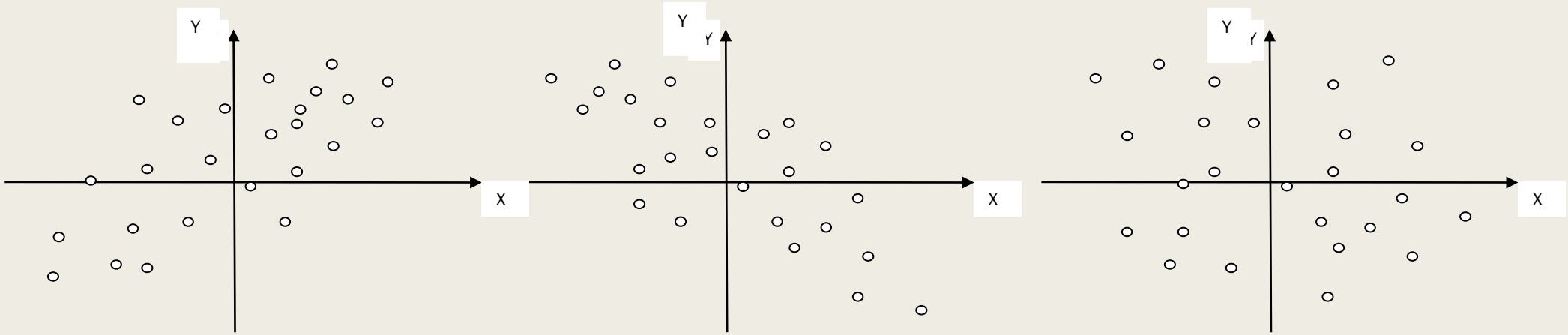
- We reject the null hypothesis and conclude that voting and gender are not independent. or, in other words, men and women (in the population) don't vote the same way.

# INVESTIGATING RELATIONSHIPS

# Correlation Coefficient

- Correlations between variables play an important role in a descriptive analysis.
- A correlation measures the relationship between two variables, that is, how they are linked to each other. In this sense, a correlation allows to know which variables evolve in the same direction, which ones evolve in the opposite direction, and which ones are independent.

# Scatter




Positive correlation

Negative correlation

No correlation

# Variance vs Covariance

- Do two variables change together?

## Variance:

- Gives information on variability of a single variable.

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

## Covariance:

- Gives information on the degree to which two variables vary together.
- Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

# Covariance

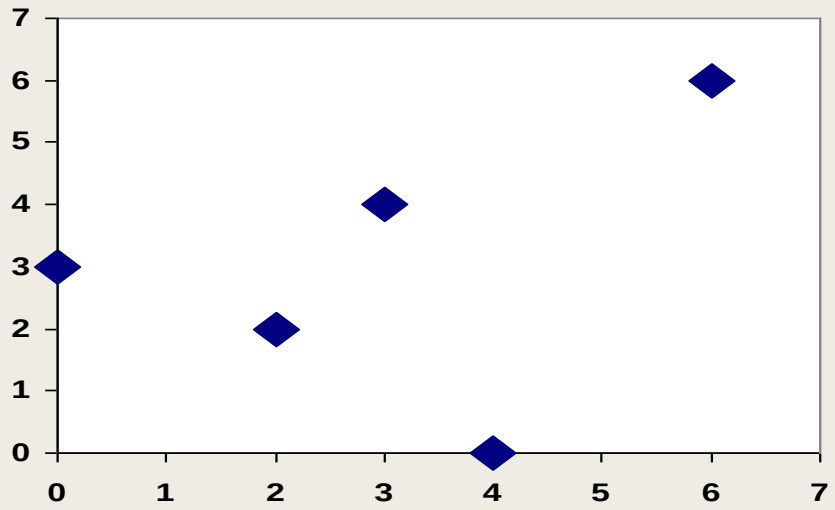
$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

When X↑ and Y↑ : cov (x,y) = pos.

When X↓ and Y↑ : cov (x,y) = neg.

When no constant relationship: cov (x,y) = 0

# Example of Covariance




| x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 0 | 3 | -3 | 0 | 0 |
| 2 | 2 | -1 | -1 | 1 |
| 3 | 4 | 0 | 1 | 0 |
| 4 | 0 | 1 | -3 | -3 |
| 6 | 6 | 3 | 3 | 9 |
| $\bar{x}=3$ | $\bar{y}=3$ | | | 7 |

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

# Problem with Covariance

- The value obtained by covariance is dependent on the size of the data's standard deviations: if large, the value will be greater than if small... *even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.*

# Example of how covariance value relies on variance

| | High variance data | | | | Low variance data | | |
|---|---|---|---|---|---|---|---|
| Subject | x | y | x error * y error | | x | y | X error * y error |
| 1 | 101 | 100 | 2500 | | 54 | 53 | 9 |
| 2 | 81 | 80 | 900 | | 53 | 52 | 4 |
| 3 | 61 | 60 | 100 | | 52 | 51 | 1 |
| 4 | 51 | 50 | 0 | | 51 | 50 | 0 |
| 5 | 41 | 40 | 100 | | 50 | 49 | 1 |
| 6 | 21 | 20 | 900 | | 49 | 48 | 4 |
| 7 | 1 | 0 | 2500 | | 48 | 47 | 9 |
| Mean | 51 | 50 | | | 51 | 50 | |
| Sum of x error * y error : | | | 7000 | | Sum of x error * y error : | | 28 |
| Covariance: | | | 1166.67 | | Covariance: | | 4.67 |

# Solution: Pearson's r

- Covariance does not really tell us anything
  - *Solution: standardise this measure*
- Pearson's R: standardises the covariance value.
- Divides the covariance by the multiplied standard deviations of X and Y:

$$r_{xy} = \frac{cov(x,y)}{s_x s_y}$$

# Spearman Rank r

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

| | Marks | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 56 | 75 | 45 | 71 | 61 | 64 | 58 | 80 | 76 | 61 |
| Maths | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

# Spearman Rank r

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) | d | $d^2$ |
|---|---|---|---|---|---|
| 56 | 66 | 9 | 4 | 5 | 25 |
| 75 | 70 | 3 | 2 | 1 | 1 |
| 45 | 40 | 10 | 10 | 0 | 0 |
| 71 | 60 | 4 | 7 | 3 | 9 |
| 62 | 65 | 6 | 5 | 1 | 1 |
| 64 | 56 | 5 | 9 | 4 | 16 |
| 58 | 59 | 8 | 8 | 0 | 0 |
| 80 | 77 | 1 | 1 | 0 | 0 |
| 76 | 67 | 2 | 3 | 1 | 1 |
| 61 | 63 | 7 | 6 | 1 | 1 |

# Spearman Rank r

Where d = difference between ranks and $d^2$ = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

# Spearman Rank r

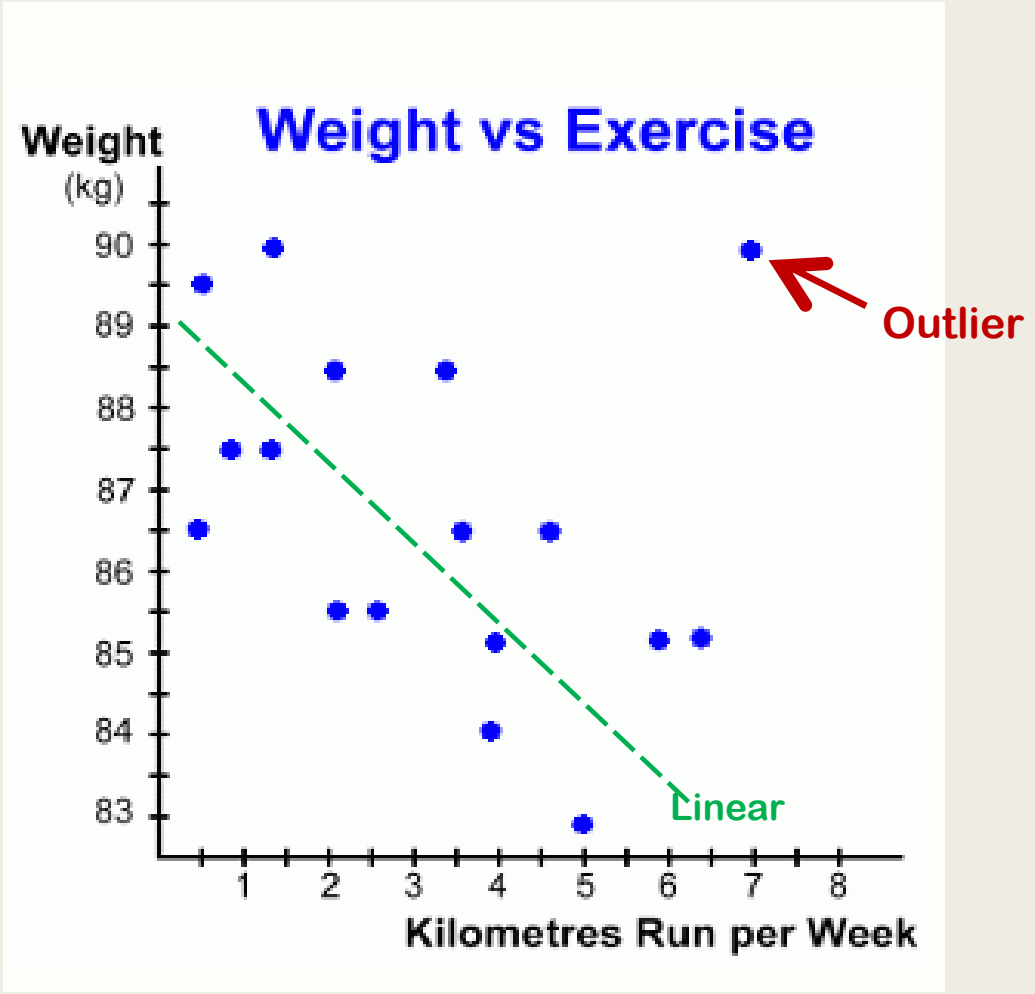
as n = 10. Hence, we have a ρ (or rs) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

# Scatterplot

## Relationship between two scale variables:

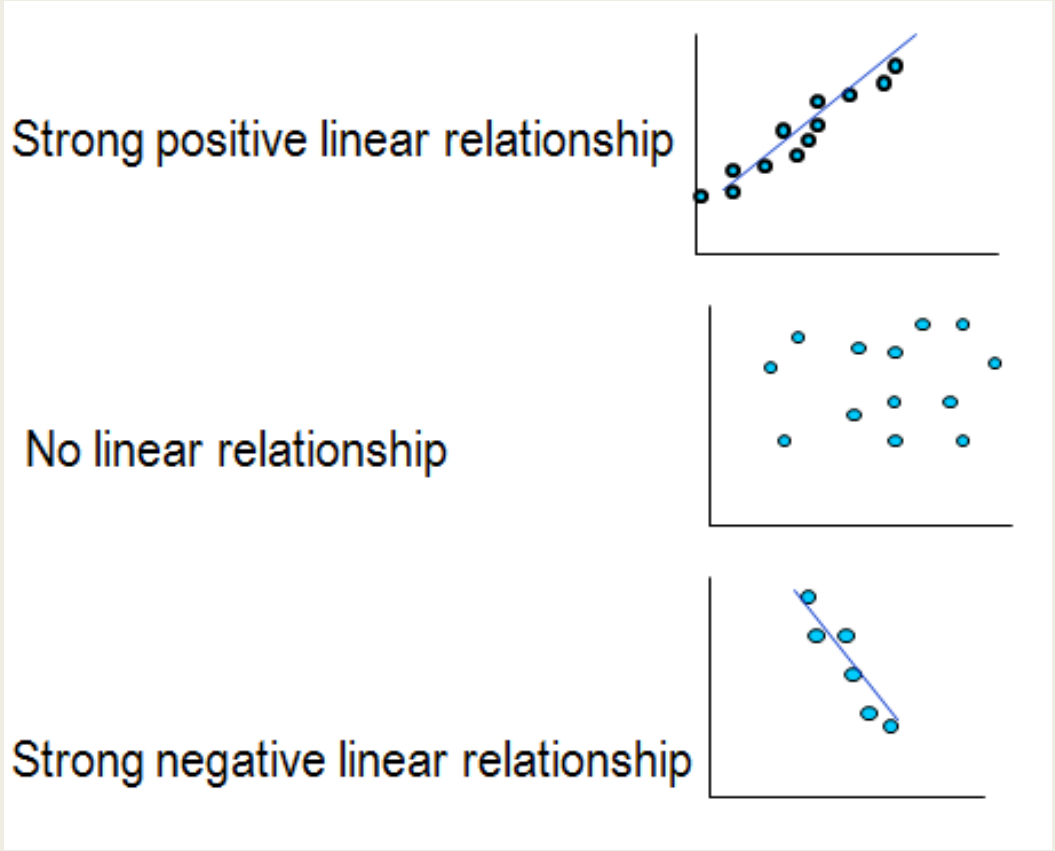
- Explores the way the two co-vary: (correlate)
  - Positive / negative
  - Linear / non-linear
  - Strong / weak
- Presence of outliers
- Statistic used:

r = correlation coefficient



# Correlation Coefficient r

- **Measures strength of a relationship between two continuous variables** $-1 \leq r \leq 1$




r = 0.9

r = 0.01

r = -0.9

# Correlation Interpretation

An interpretation of the size of the coefficient has been described by Cohen (1992) as:

| Correlation coefficient value | | | Relationship |
|---|---|---|---|
| -0.3 to +0.3 | | | Weak |
| -0.5 to -0.3 | or | 0.3 to 0.5 | Moderate |
| -0.9 to -0.5 | or | 0.5 to 0.9 | Strong |
| -1.0 to -0.9 | or | 0.9 to 1.0 | Very strong |

*Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159*

# Two Categorical Variables

Are boys more likely to prefer maths and science than girls?

Variables:

- Favourite subject (Nominal)
- Gender (Binary/ Nominal)

Summarise using %'s/ stacked or multiple bar charts

**Test: Chi-squared**

Tests for a relationship between two categorical variables

# Does chocolate make you clever or crazy?

A paper in the New England Journal of Medicine claimed a relationship between chocolate and Nobel Prize winners

r = 0.791

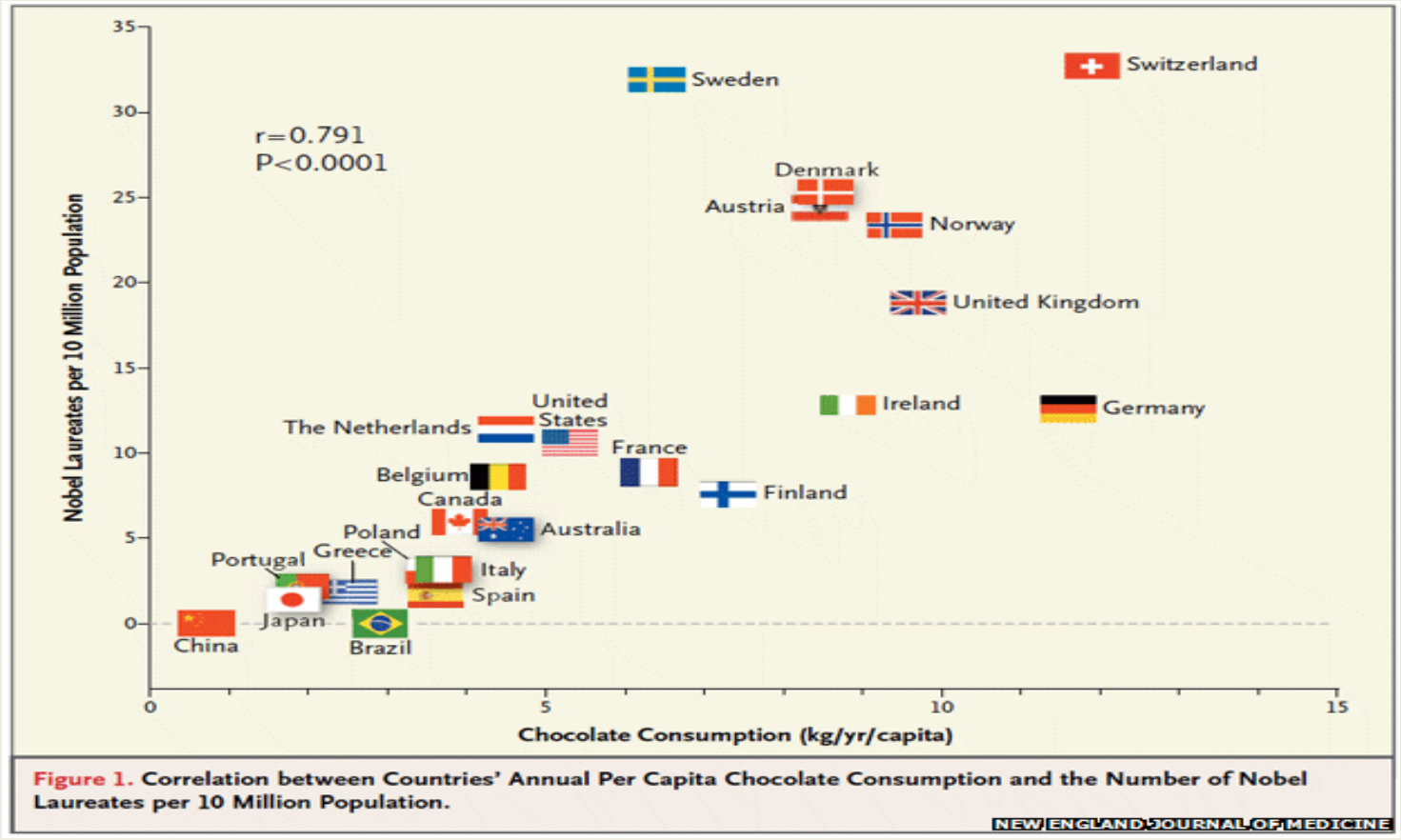


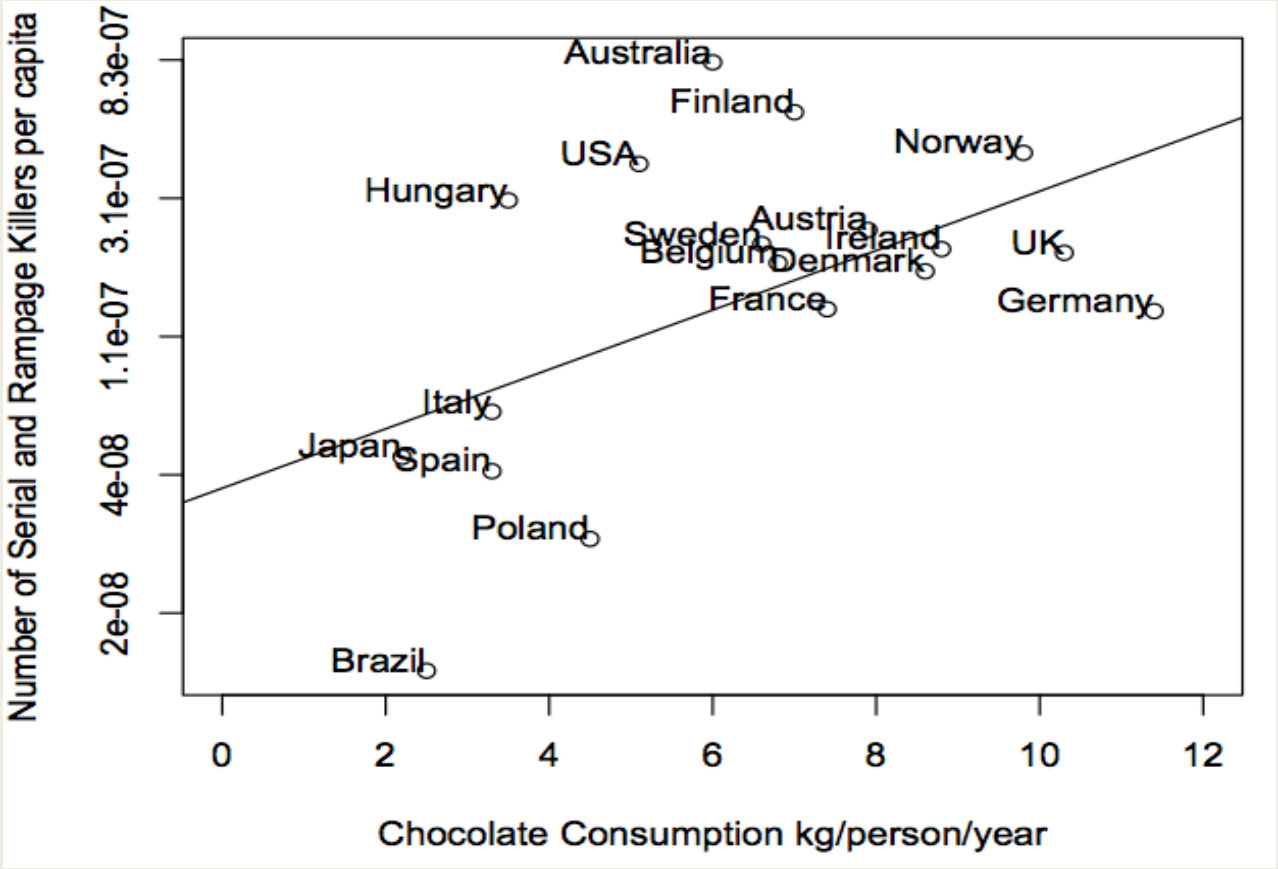

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

http://www.nejm.org/doi/full/10.1056/NEJMon1211064

# Chocolate and serial killers

- What else is related to chocolate consumption?




r = 0.52

http://www.replicatedtypo.com/chocolate-consumption-traffic-accidents-and-serial-killers/5718.html

# Hypothesis tests for r

Tests the null hypothesis that the population correlation r = 0 NOT that there is a strong relationship!

It is highly influenced by the number of observations e.g. sample size of 150 will classify a correlation of 0.16 as significant!

Better to use Cohen's interpretation

# Exercise

- Interpret the following correlation coefficients using Cohen's and explain what it means

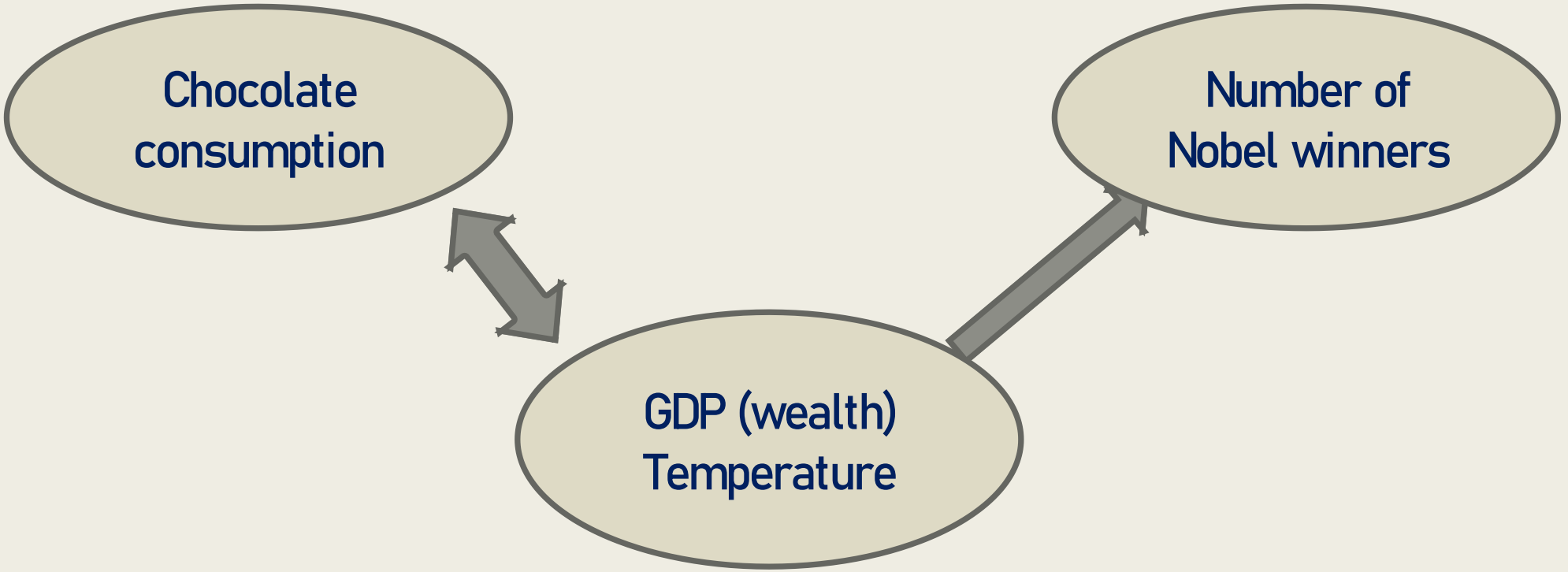
| Relationship | Correlation |
|---|---|
| Average IQ and chocolate consumption | 0.27 |
| Road fatalities and Nobel winners | 0.55 |
| Gross Domestic Product and Nobel winners | 0.7 |
| Mean temperature and Nobel winners | -0.6 |

# Exercise - solution

| Relationship | Correlation | Interpretation |
|---|---|---|
| Average IQ and chocolate consumption | 0.27 | Weak positive relationship. More chocolate per capita = higher average IQ |
| Road fatalities and Nobel winners | 0.55 | Strong positive. More accidents = more prizes! |
| Gross Domestic Product and Nobel winners | 0.7 | Strong positive. Wealthy countries = more prizes |
| Mean temperature and Nobel winners | -0.6 | Strong negative. Colder countries = more prizes. |

# Confounding

Is there something else affecting both chocolate consumption and Nobel prize winners?



# The relationship between x and y

- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?

- CORRELATION ≠ CAUSATION
  - *In order to infer causality: manipulate independent variable and observe effect on dependent variable*

# Regression: Association between two variables

- Regression is useful when we want to

a) *look for significant relationships* between two variables

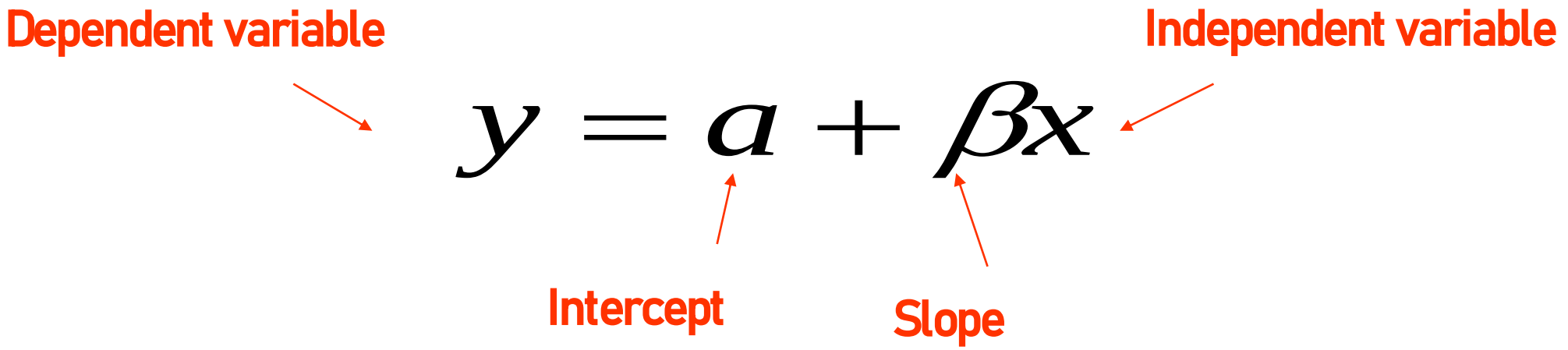
b) *predict* a value of one variable for a given value of the other

It involves estimating the line of best fit through the data which minimises the sum of the squared residuals

What are the residuals?

# Regression

**Simple linear regression looks at the relationship between two Scale variables** by producing an equation for a straight line of the form




Which uses the independent variable to predict the dependent variable

# Multiple regression

What affects the number of Nobel prize winners?

Dependent: Number of Nobel prize winners

Possible independents: Chocolate consumption, GDP and mean temperature

- Chocolate consumption is significantly related to Nobel prize winners in simple linear regression
- Once the effect of a country's GDP and temperature were taken into account, there was no relationship

# Logistic regression

- Logistic regression has a binary dependent variable
- The model can be used to estimate probabilities
- Example: insurance quotes are based on the likelihood of you having an accident
- Dependent = Have an accident/ do not have accident
- Independents: Age (preferably Scale), gender, occupation, marital status, annual mileage
- Ordinal regression is for ordinal dependent variables

# CHOOSING THE RIGHT TEST

# Choosing the right test

- One of the most common queries in statistics support is 'Which analysis should I use'
- There are several steps to help the student decide
- When a student is explaining their project, these are the questions you need answers for

# Choosing the right test

1) A clearly defined research question
2) What is the dependent variable and what type of variable is it?
3) How many independent variables are there and what data types are they?
4) Are you interested in comparing means or investigating relationships?
5) Do you have repeated measurements of the same variable for each subject?

# Logic of data analysis

- Univariate analysis
  - *One variable at a time (descriptive)*
- Bivariate analysis
  - *Two variables at a time (testing relationships)*
- Multivariate analysis
  - *More than two variables at a time (testing relationships and controlling for other variables)*

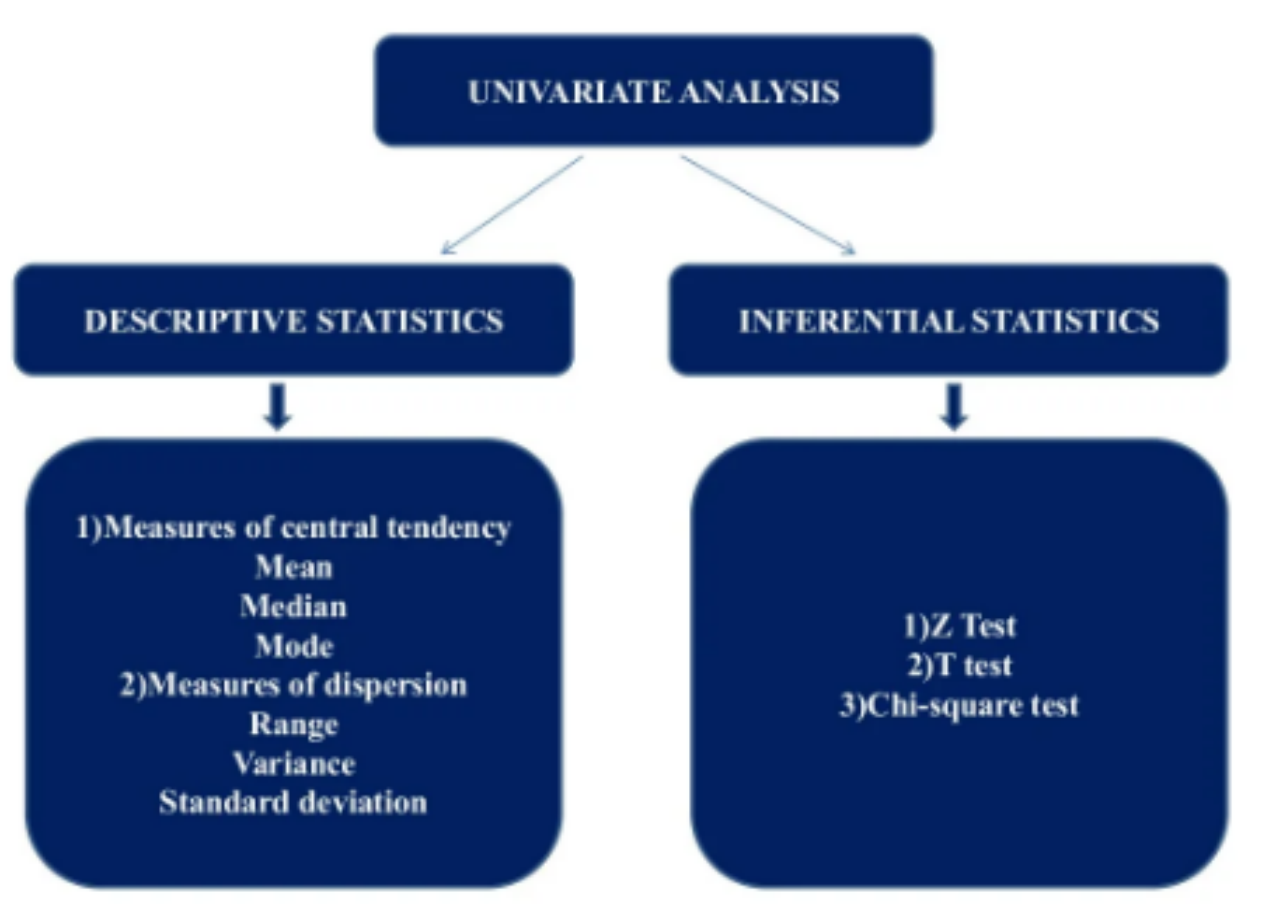
# Type of Analysis

- Univariate Analysis
  - *The examination of distribution of cases on only one variable at a time(e..g., weight of college students*
- Bivariate Analysis
  - *The Examination of two variables simultaneously (e.g., the relation between gender and weight of college students*
- Multivariate Analysis
  - *The examination of more than two variables simultaneously (e.g., the relationship between gender, race and weight of college students)*

# Purpose of Type of Analysis

- Univariate Analysis
  - *Purpose: mainly description*
- Bivariate Analysis
  - *Purpose: determining the empirical relationship between the two variables*
- Multivariate Analysis
  - *Purpose: determining the empirical relationship among multiple variables*

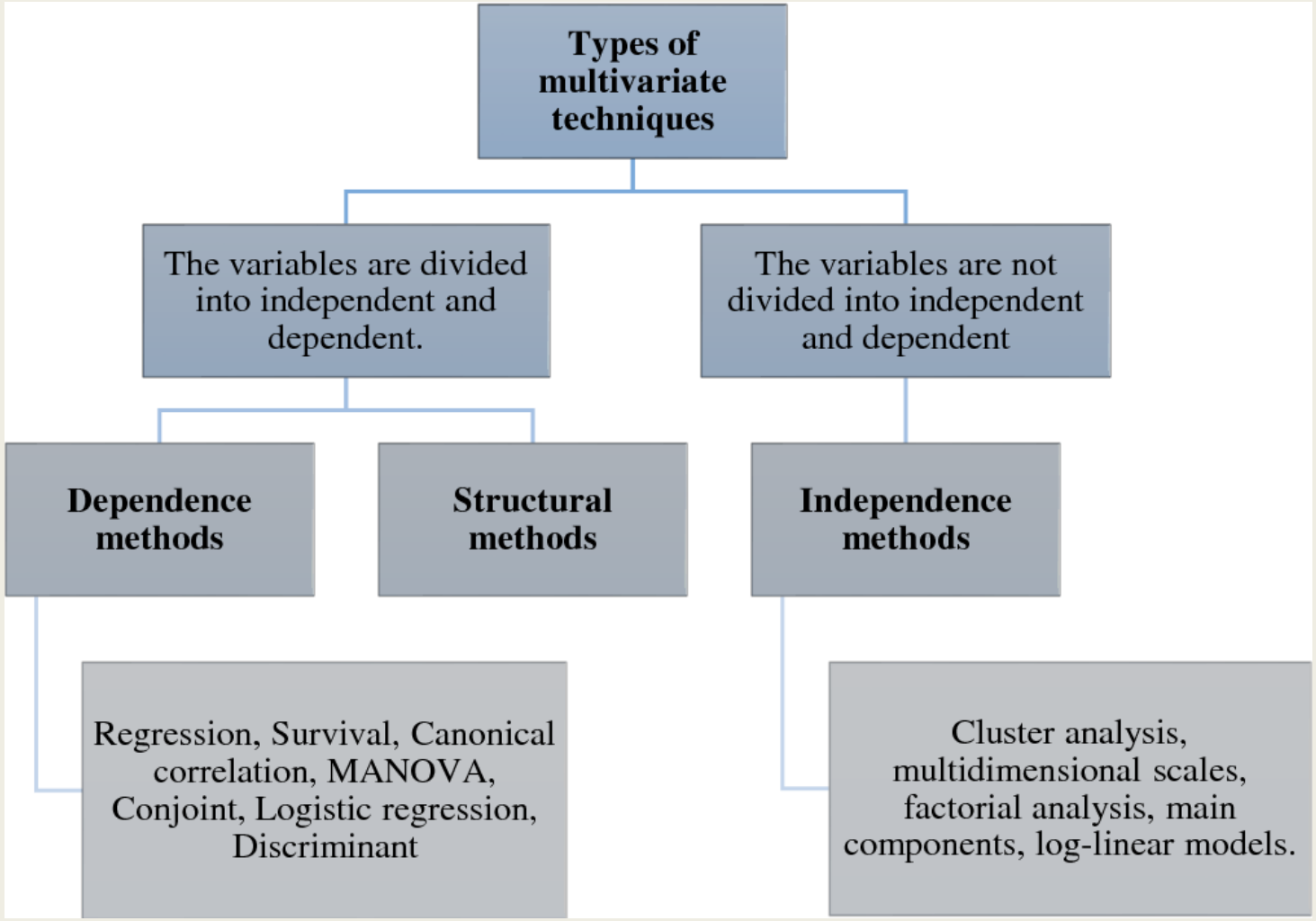
# Summary of Statistical Methods for Univariate Analysis



# Summary of Statistical Methods for Bivariate Analysis

| Type of Variable | Nominal or Ordinal | Interval or Ratio |
|---|---|---|
| Nominal or Ordinal | Chi-Square Test<br>Cross-Tab | Analysis of Variance (ANOVA) |
| Interval or Ratio | Chi-Square Test | Pearson Correlation |

# Summary of Statistical Methods for Multivariate Analysis

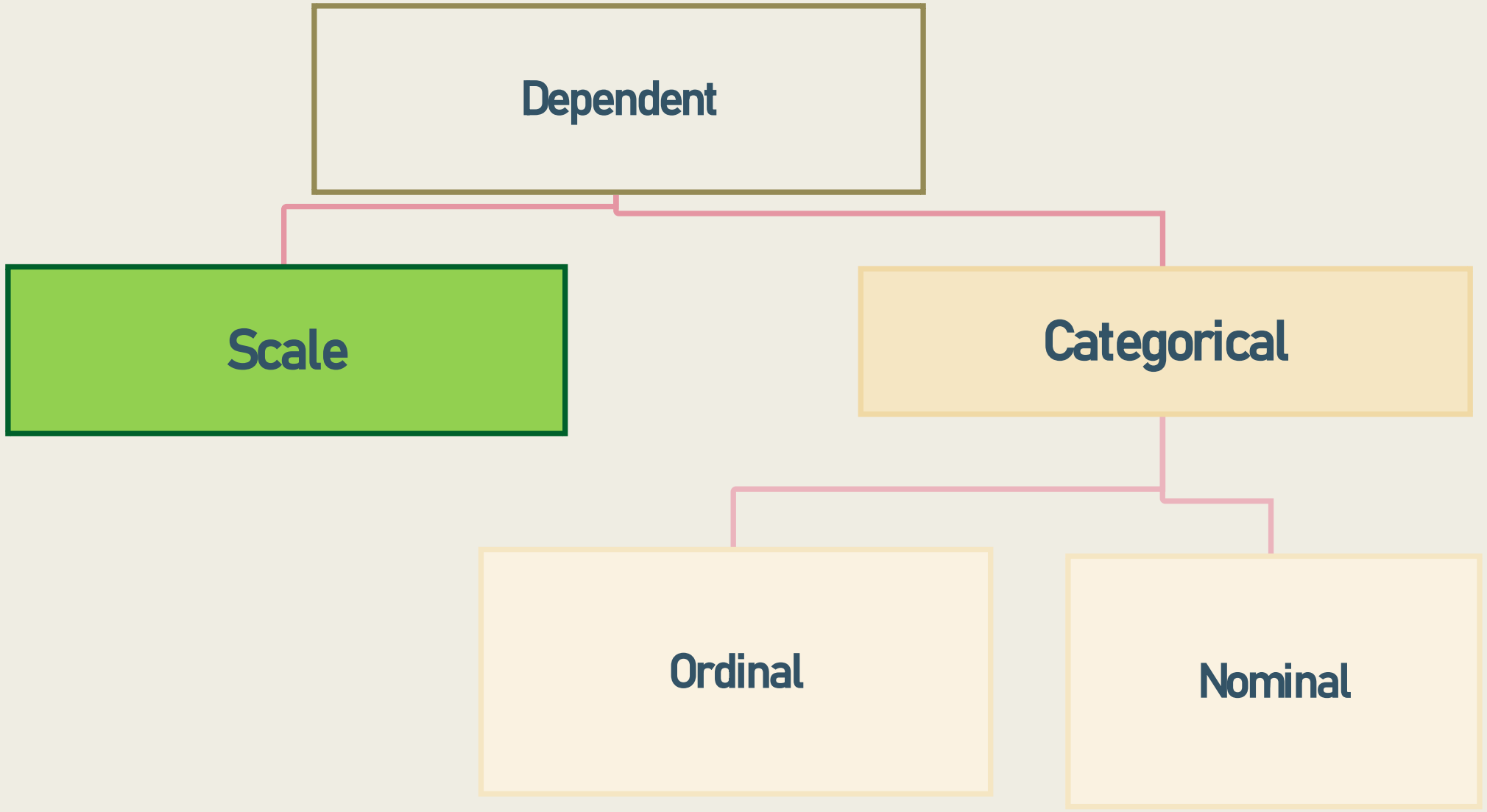


# Dependent variables

# Variables

- Dependent Variable (DV):
  - *What we are trying to predict*
  - *E.g., Candidate preference*
- Independent Variables (IV):
  - *What we are using as predictors*
  - *E.g., Gender, Party affiliation*

# What variable type is the dependent?



# Testing hypothesis for two nominal variables

| Variables | Null hypothesis | Procedure |
| --- | --- | --- |
| Gender<br><br>Pass/Fail | Passing is not related to gender | Chi-square |

# Testing hypothesis for one nominal and one ratio variable

| Variables | Null hypothesis | Procedure |
| --- | --- | --- |
| Gender<br><br>Test score | Score is not related to gender | T-test |

# Testing hypothesis for one nominal and one ratio variable

| Variable | Null hypothesis | Procedure |
| --- | --- | --- |
| Year in school | Score is not related to year in school | ANOVA |
| Test score | | |

- Can be used when nominal variable has more than two categories and can include more than one independent variable

# Testing hypothesis for two ratio variables

| Variable | Null hypothesis | Procedure |
| --- | --- | --- |
| Hours spent studying | Score is not related to hours spent studying | Correlation |
| Test score | | |

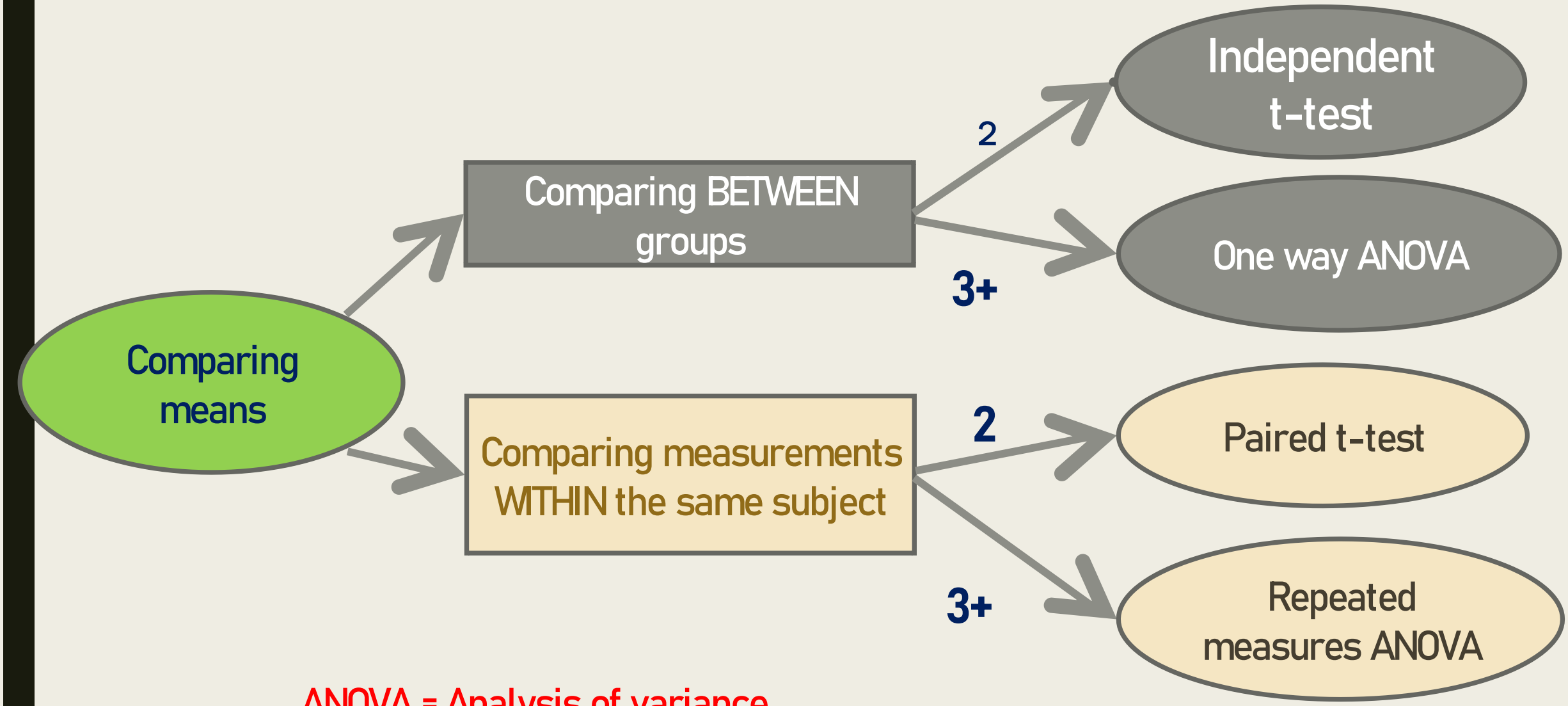
# Testing hypothesis for more than two ratio variables

| Variable | Null hypothesis | Procedure |
| --- | --- | --- |
| Hours spent studying<br><br>Classes missed<br><br>Test score | Score is not positively related to hours spent studying and not negatively related to classes missed | Multiple regression |

# Summary of Statistical Procedures

| Variables | Procedure |
| --- | --- |
| Nominal IV, Nominal DV | Chi-square |
| Nominal IV, Ratio DV | T-test |
| Multiple Nominal IVs, Ratio DV | ANOVA |
| Ratio IV, Ratio DV | Pearson's R |
| Multiple Nominal IVs, Ratio DV with ratio covariates | ANCOVA |
| Multiple ratio | Multiple Regression |

# Comparing means




ANOVA = Analysis of variance

# Question: Are boys better at maths than girls?

# Comparing measurements on the same people

*Also known as within group comparisons or repeated measures.*

Can be used to look at differences in mean score:
(1) over 2 or more time points e.g. 1988 vs 2014
(2) under 2 or more conditions e.g. taste scores

Participants are asked to taste 2 types of cola and give each scores out of 100.
Dependent = taste score
Independent = type of cola

# Data types

| Research question | Dependent/ outcome variable | Independent/ explanatory variable |
|---|---|---|
| **Does attendance have an association with exam score?** | Exam score (scale) | Attendance (Scale) |
| Do women do more housework than men? | Hours of housework per week (Scale) | Gender (Nominal) |

# Exercise:

How would you investigate the following topics? State the dependent and independent variables and their variable types.

| Research question | Dependent/ outcome variable | Independent/ explanatory variable |
|---|---|---|
| Were Americans more likely to survive on board the Titanic? | | |
| Does weekly hours of work influence the amount of time spent on housework? | | |
| Which of 3 diets is best for losing weight? | | |

# Tests investigating relationships

| Investigating relationships between | Dependent variable | Independent variable | Test |
|---|---|---|---|
| 2 categorical variables | Categorical | Categorical | Chi-squared test |
| 2 Scale variables | Scale | Scale | Pearson's correlation |
| Predicting the value of an dependent variable from the value of a independent variable | Scale | Scale/binary | Simple Linear Regression |
| | Binary | Scale/ binary | Logistic regression |

Note: Multiple linear regression is when there are several independent variables

# Exercise: Relationships

| Research question | Dependent variable | Independent variables | Test |
|---|---|---|---|
| Does attendance affect exam score? | Exam score (Scale) | Attendance (Scale) | |
| Do women do more housework than men? | Housework (hrs per week) (scale) | Gender (Binary) Hours worked (Scale) | |
| Were American rich men more likely to survive on board the Titanic? | Survival (Binary) | Nationality , Gender, class | |

# End of Chapter

# CHISQ.TEST() Function in R

- **chisq.test()** function performs chi squared contingency table tests and goodness of fit tests.
- chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)

## Arguments

- x: a  numeric vector or matrix. x and y can also both be factors.
- y: a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of the same length.
- correct logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all |O−E| differences; however, the correction will not be bigger than the differences themselves. No correction is done if simulate.p.value = TRUE.
- p: a vector of probabilities of the same length of x. An error is given if any entry of p is negative.
- rescale.p logical scalar; if TRUE then p is rescaled (if necessary) to sum to 1. If rescale.p is FALSE, and p does not sum to 1, an error is given.
- simulate.p.value: a logical indicating whether to compute p-values by Monte Carlo simulation.
- B: an integer specifying the number of replicates used in the Monte Carlo test.

# GOODNESS-OF-FIT TESTS IN R

**Example:** (Roses) When crossing certain types of red and white roses, one obtains red, white and pink roses. Theory predicts that the proportion of red to white to pink roses is like **3:2:2**. Test the plausibility of this theory when out of a sample of 80 crosses, 35 are red, 31 are white and 14 are pink. (Note: Sampling design is multinomial sampling of one variable and we test to see if the multinomial probabilities are equal to some specified values)

# GOODNESS-OF-FIT TESTS IN R

```
> chisq.test(c(35,31,14), p=c(3,2,2)/7)

        Chi-squared test for given probabilities

data:  c(35, 31, 14)

X-squared = 6.3479, df = 2, p-value = 0.04184
```

**Conclusion: At a 5% significance level, the data provide sufficient evidence (P-value = 0.0418) that the proportion of red to white to pink roses is different from 3:2:2.**

# CONTINGENCY TABLE TESTS IN R

**Example:** (Hair and Eye Color) In a sample of 65 students, we recorded the hair color (categories blond, brown, dark) and eye color (categories bright, dark). The table below summarizes the counts.

**Null hypothesis:** Hair and Eye color are independent.

**Alternative Hypothesis:** Hair and eye color are associated.

# CONTINGENCY TABLE TESTS IN R

```
> table1=matrix(c(12,2,8,25,6,12),ncol=3)
> colnames(table1)=c("blond","brown","dark")
> rownames(table1)=c("bright","dark")
> table1
       blond brown dark
bright    12     8    6
dark       2    25   12
```

# CONTINGENCY TABLE TESTS IN R

```
> chisq.test(table1)

Pearson's Chi-squared test

data: table1

X-squared = 15.938, df = 2, p-value = 0.000346
```

**Conclusion:** There is sufficient evidence (P-value=0.0003) that hair and eye color of students are associated

# GOODNESS OF FIT IN PYTHON

- The Chi-Square Goodness of Fit Test using the chisquare function from the SciPy library.
- Recall that a Chi-Square Goodness of Fit Test uses the following null and alternative hypotheses:
  - **$H_0$: (null hypothesis):** A variable follows a hypothesized distribution.
  - **$H_1$: (alternative hypothesis):** A variable does not follow a hypothesized distribution.

# GOODNESS OF FIT IN PYTHON

```
expected = [50, 50, 50, 50, 50]
observed = [50, 60, 40, 47, 53]
import scipy.stats as stats


#perform Chi-Square Goodness of Fit Test
stats.chisquare(f_obs=observed, f_exp=expected)


(statistic=4.36, pvalue=0.35947)
```

# CONTINGENCY TABLE TESTS IN PYTHON

- The Pearson's Chi-Square statistical hypothesis is a test for independence between categorical variables
- Using Python's SciPy module.
- Starting by defining the null hypothesis (H0) which states that there is no relation between the variables. An alternate hypothesis would state that there is a significant relation between the two.

# CONTINGENCY TABLE TESTS IN PYTHON

```
from scipy.stats import chi2_contingency
# defining the table
data = [[207, 282, 241], [234, 242, 232]]
stat, p, dof, expected = chi2_contingency(data)
# interpret p-value
alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Independent (H0 holds true)')
```