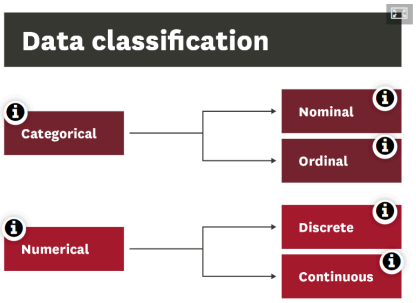
BUSA6004 - CHEAT SHEET
Date: 8 Jun, 2024

## WEEK 1: INTRODUCTION




1. Primary data: Data you collect yourself using surveys or interviews.
2. Secondary data: Data collected by other people.
3. Observation: Value obtained from a sample.
4. Raw data: Data as they are collected, and not yet processed.
5. Frequency: Number of times an observation occurs.
6. Frequency distrib...: Lists the frequency of each observation in a table.

**Mean:** average of a set of values, defined as the arithmetic average.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

**Median:** midpoint, when the data set is sorted. The median value will be higher than 50% of the observations in the data set and lower than 50% of the values in the data set. To determine the median, always organize data from low to high values.

**Mode:** the most frequently occurring value. It corresponds to the highest bar in the histogram.

## WEEK 2: VARIATION MEASURES AND NORMAL DISTRIBUTION

- measure the spread of data
- normal distribution and its properties
- calculate a Z-score

Distribution Shapes (symmetric, skewed, unimodal, bimodal or multimodal)
Range = largest value - smallest value (= the spread of the whole data set)
IQR = Q3 - Q1 (= spread of the middle half of a data set)
Outliers = Q1 - IQR*1.5 or Q3 + IQR*1.5

Deviation = (X - mean)
Mean absolute deviation (MAD) =

$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n}$$

where

Σ means to sum up
| | are vertical bars that mean absolute value
x is each data value
x̄ is the mean
n is the total number of data points in your data set

***Population and sample standard deviation***

Sample: s
Population: σ

**Population standard deviation:**

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

where
N is the count of the entire population
μ is the mean of entire population
xi is a data point
xi - μ is the deviation
(xi - μ)2  is the deviation squared

**Sample standard deviation:**

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

where
x is each data value
x̄ is the mean
n is the total number of data points in your data set
(n-1) denotes degrees of freedom (to be explained later)

**Variance**

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots\ldots + (y_n - \bar{y})^2}{n-1}$$

*always remember:*

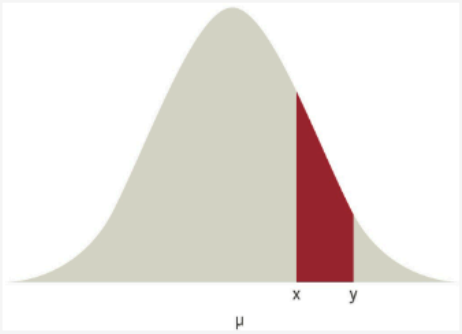
- ❖ The standard deviation is the square root of the variance.
- ❖ The variance is the standard deviation squared.
- ❖ The variance for a sum of two independent variables is found by adding both variances.
- ❖ You can add variances but you cannot add standard deviations.

**The normal distribution (Guassian)**

- Symmetrical about the mean.
- The total area under the curve is defined to be 1.
- It approaches the horizontal axis with about 3 values of σ either side of μ.

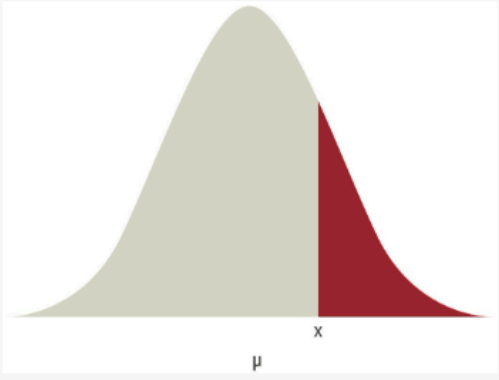
1. The probability of finding a value between two specific points.

The probability that a value from a normal distribution lies between two points x and y is equal to the area under that normal curve between x and y.



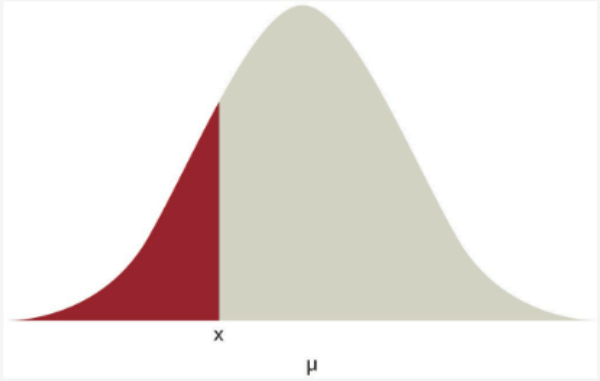

2. The probability of finding a value being below a specific point.

The probability that a value lies above a point x is equal to the area under the curve to the right of x.



3. The probability of finding a value above a specific point.

The probability that a value lies below a point x is equal to the area under the curve to the left of x.




***The standard normal distribution***

Mean = μ = 0

Standard Deviation = σ = 1

***Z-value calculation***

$$\mathrm{Z} = \frac{\mathrm{X}_1 - \mathrm{Mean}}{\mathrm{Standard\ Deviation}} = \frac{X - \mu}{\sigma}$$

**WEEK 3: STATISTICAL INFERENCE**

- calculate the confidence interval
- calculate the expected value
- calculate probabilities
- role of permutations and combinations

***statistical inference*** *= drawing conclusions about a population based on a random sample that comes from the population*

**Population mean and proportion**

*The sample mean provides a point estimate for the population mean, the accuracy of which is measured by its standard error.*

Accuracy of the sample mean as the estimate of the population mean is measured by the standard error of the estimate - SE(x̄):

$$SE(\bar{x}) = \frac{S}{\sqrt{n}}$$

where
s is the sample standard deviation
n is the sample size

The **sample standard deviation (s)** is:

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

x̄ is a good estimate if its standard error is less than 5%.

***Population proportion***

The population proportion Π is best estimated by the sample proportion p. The standard error of this estimate (assuming a sample size of at least 25) is:

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} \qquad z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If the standard error is less than 5% of p, the estimate is reasonably accurate.

**Confidence level**

*Confidence interval for a population mean*

A confidence interval can be constructed if:

- The sample is randomly selected from the population.
- The sample size is at least 25 or the population is normally distributed.

$$\left(\bar{x} - z\frac{s}{\sqrt{n}}, \bar{x} + z\frac{s}{\sqrt{n}}\right)$$

where
z = 1.645 for a 90% confidence interval
z = 1.96 for a 95% confidence interval
z = 2.58 for a 99% confidence interval

You can use Excel to calculate the confidence level if you know:

significance level α
standard deviation s
sample size n

You use the formula: = confidence.norm(alpha,standard deviation, size)

For the soft drinks example above = confidence.norm(0.05,150,90) gives a value of 31. The 95% confidence interval is then (1965-31,1965+31).

*Confidence interval for a population proportion*

The confidence interval for a population proportion can be constructed if the sample size is at least 25. The sample proportion is **p** and the variance of this proportion is **p(1 − p).**

The **confidence interval** for sample proportion is:

$$\left(p - z\sqrt{\frac{p(1-p)}{n}}, p + z\sqrt{\frac{p(1-p)}{n}}\right)$$

**Margin of error**

Another way to express the accuracy of an estimate is to use the margin of error (MOE).

$$MOE = 1.96 \times \frac{s}{\sqrt{n}}$$

**Expected value of a random variable X**

$$\mu = \sum x \Pr(X = x)$$

for the coin toss, the expected value is: 0.5 × 0 + 0.5 × 1 = 0.5

**Risk and chance**

$$\mu = \sum x \Pr(X = x)$$

The expected amount after 12 months under option A
= ($100,000 x 1.10 x 0.65) + ($100,000 x 1.05 x 0.25) + ($100,000x0.90 x 0.10)
= $71500 + $26250 + $9000
= $106,750

***Gambling***
Percentage House Margin
= 100 x (Expected House profit per game / Cost of playing the game)

**Factorials, permutations & combinations**

Factorials

$$n! = n \times (n-1) \times \ldots \times 3 \times 2 \times 1$$

Permutations: $${}_nP_r = \frac{n!}{(n-r)!}$$

Combinations: $${}^nC_r = \frac{n!}{r!(n-r)!}$$

**Number of events before the first success =** $P\ (X = x)\ = \ p(1 - p)^{\ x-1}$

$$P(X = x) = p(1-p)^{x-1}$$

**WEEK 4: HYPOTHESIS TESTING AND ONE-SAMPLE TESTS**

- steps for conducting a hypothesis test
- analyze the results of the hypothesis test
- evaluate the results of the hypothesis test

**Steps conducting a hypothesis testing**

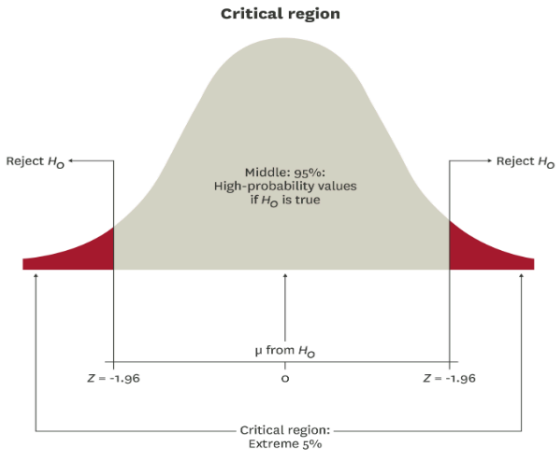
**Step 1:** State a hypothesis

Null hypothesis = no change, no effect, no difference

$H0: \mu = \mu 0$
$H1: \mu \neq \mu 0$

**Step 2:** Determine sample mean distribution

**Step 3:** Determine critical region




**Step 4:** Compare the obtained sample data

- If the *z*-score is located in the critical region, reject the null hypothesis.
- If the *z*-score does not lie in the critical region, then the data does not provide strong evidence that the null hypothesis is wrong.

**Draw conclusions**

## Conclusion: Hypothesis testing vs jury trial

| Steps | Hypothesis testing | Jury trial |
|---|---|---|
| **Null hypothesis** | Treatment has no effect | The defendant is innocent (innocent until proven guilty) |
| **Data collection** | Collect sample data | Police gather evidence |
| **Not enough evidence**<br>**Wrong conclusion** | Fail to reject null hypothesis there is no treatment effect | Fail to find the defendant guilty<br>Defendant is innocent |

**One-tail test**

*"If the direction of the treatment effect could only be in one direction, one-tailed tests are more appropriate."*

***p-value*** = probability of obtaining test results at least as extreme as the observed result in the sample when the null hypothesis is correct.

### *Decision Rule*

Reject $H_0$ if p-value < $\alpha$

DO NOT reject $H_0$ if p-value > $\alpha$

### *Errors in hypothesis testing*

- **Type I error**: reject H0 when it is true
- **Type II error**: not reject H0 when it is false

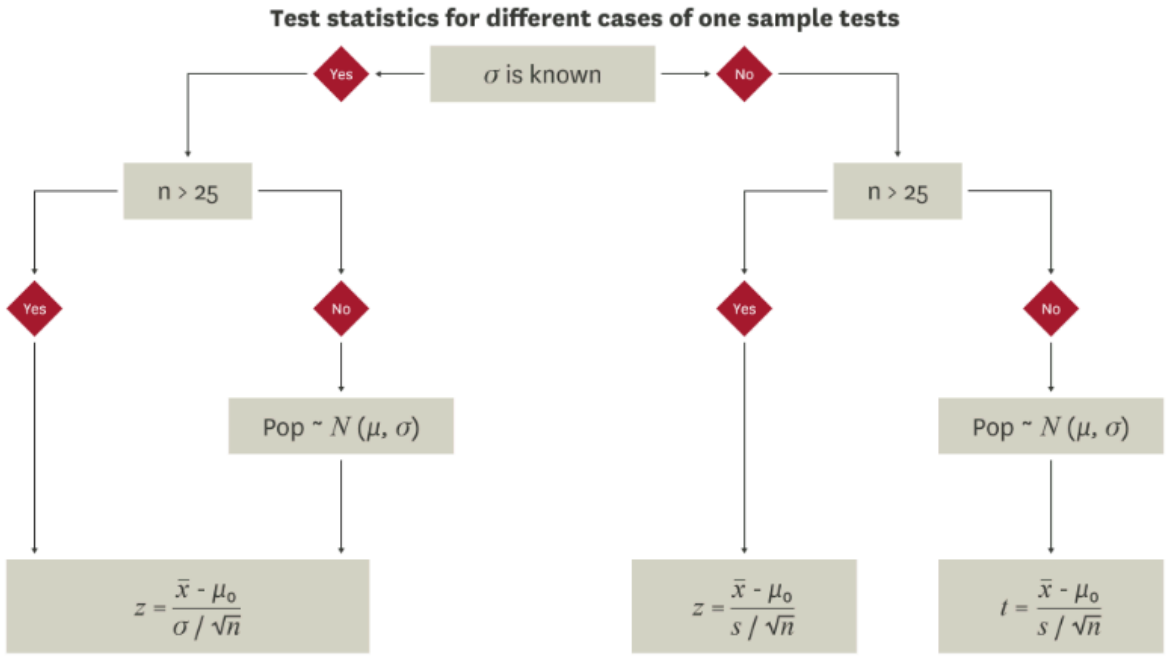
### *Power of a test*

~ the probability of not making type II error.

An increase in sample size will help to magnify the difference between μ0 and μ1 and increase test power.

For μ0 that is close to μ1 or higher than μ1, the test power is small and it is difficult to reject the wrong null hypothesis.

### *Choice of sample tests*



**WEEK 5: CORRELATION AND FORECASTING**

- calculate correlation and regression
- evaluate the results of correlation and regression

**Correlation analysis**

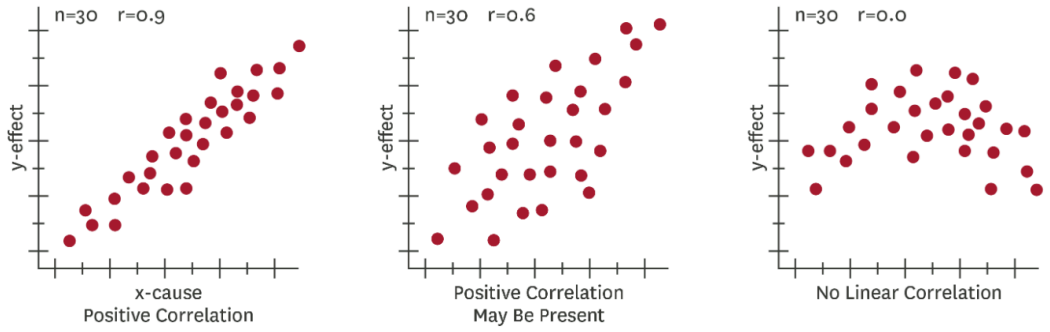
~ a linear relationship

r: sample correlation coefficient
ρ: population correlation coefficient

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

The Excel function is: correl(x,y)

**Test statistics for different cases of one sample tests**




**Test for significance**

- $H_0$: $\rho = 0$
- $H_a$: $\rho = 0$

t-test with test statistic:

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

where t follows a student distribution with n − 2 degrees of freedom

Look up the critical value $t^*_{\alpha,n-2}$. *If* $|t| > t^*_{\alpha,n-2}$ reject $H_0$, otherwise, do not reject $H_0$.

You can calculate t∗α,n−2 in excel using the following function: t.inv(1 − α/2, n − 2)

The p-value can also be computed to be compared with α:

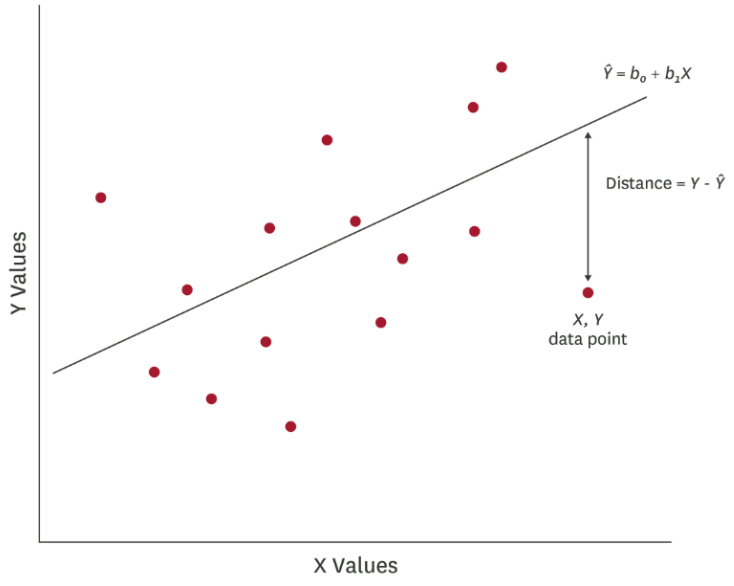
p-value = 2 × t.dist(−abs(t), n − 2,TRUE)

If p-value < α, reject H0, conclude that the correlation is significant.

**Spurious correlation**

The correlation is then spurious (false) since the two variables are not directly related. To check if there are confounding variables, a partial correlation coefficient may be calculated.

**Simple linear regression**




Simple linear regression model is: Y = β0 + β1X + ε

where
y= dependent variable (response variable)
x = independent variable (predictor or explanatory variable)
β0 = intercept (value of y when x = 0)
β1 = slope of the regression line
ε = residual error

If the true intercept and slope are unknown, they are estimated using sample data. Regression equation based on sample data:

$$\hat{y} = b0 + b1x$$

where:
$\hat{y}$ = predicted value of y
b0 = estimate of β0
b1 = estimate of β1

**Prediction interval**

95% confidence interval around the estimate $\bar{x}$ of the population mean is: $\bar{x} \pm MOE$

$$MOE = 1.96 \times \frac{s}{\sqrt{n}}$$

*95% confidence interval around $\hat{Y}$ is: $\hat{Y} \pm MOE$*

where MOE = t × Se × Attributable Fraction (AF); with
t = critical value for two tail t-test with df=n-2
Se = standard deviation of the regression residuals

$$AF = \sqrt{1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{(N-1)S_X^2}}$$

In the regression of student average grade on family income, we know further that N = 14, $\bar{x}$ = 75; sample variance of x is 1667.84; Se = 5.60.

Compute the 95% prediction interval for the average grade of a student coming from family with income $50,000.

For 95% confidence:
Degrees of freedom= N − 2 = 12
t = 2.18
Using the formula for AF provided, AF = 1.05
MOE = 2.18 × 5.60 × 1.05 = 12.82

95% prediction interval for the average grade of a student coming from family with income $50,000 is then:
**83.61 ± 12.82**

**Multiple linear regression**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where
y=dependent variable
X1, X2=explanatory variables
β0=y-intercept
βp=slope coefficients for each explanatory variable
ε =the model's error term (also known as the residuals)

***Coefficient of determination (R-squared)***

~ how much of the variance in the outcome can be accounted for by the change in the independent variables

- R2 close to 1 indicates a good fit, and close to 0 indicates a poor fit.
- For simple regression, R2 is the correlation coefficient squared, i.e R2 =r2.
- R2 increases as more variables are added to the model.

| **T-test**<br>= significance of a variable in the model | **F-test** |
|---|---|
| In a good model, all variables are significant.<br><br>p-value of a variable is greater than 0.05, it should be omitted. | p-value < 0.05 then at least one of the *x* variables is useful in the model at a significance level of 5%.<br>p-value > 0.05 then none of the *x* variables are useful. |

**Partial correlation**

~ detect if a correlation is spurious or not

Method 1: Compare correlation $r_{xy}$ with partial correlation $r_{xy.z}$.

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{\left(1 - r_{XZ}^2\right)\left(1 - r_{YZ}^2\right)}}$$

If $r_{xy}$ is not much different from $r_{xy.z}$, then z is unlikely to be a lurking variable.

Method 2:

A more definite answer can be obtained using Method 2:

- Regress y on z to get residuals $\varepsilon_1$.
- Regress x on z to get residuals $\varepsilon_2$.
- Calculate correlation $r_{\varepsilon_1\varepsilon_2}$ between $\varepsilon_1$ and $\varepsilon_2$.
- Test significance of $r_{\varepsilon_1\varepsilon_2}$.

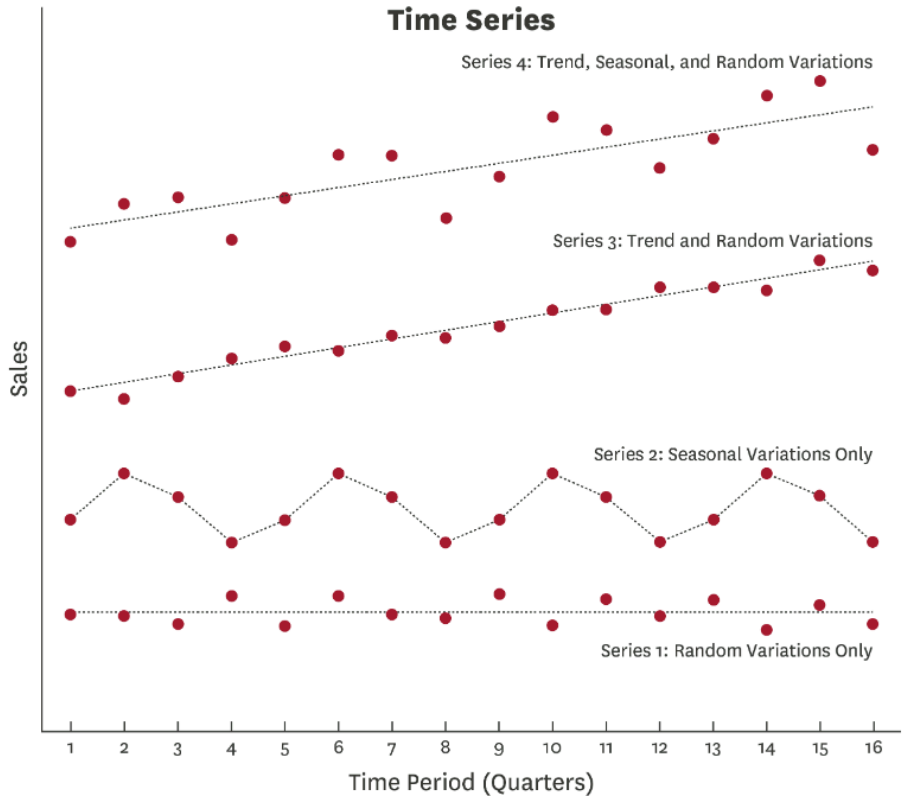
A researcher is examining the relationship between salary (y) and blood pressures (x). It is suspected that age (z) is a lurking variable. The correlations between variables are:
rxy = 0.968
rxz = 0.960
ryz = 0.984

Calculate the partial correlation between x and y, taking into account z and compare it with rxy to see if z is a lurking variable.

rxy.z = 0.469
Since there is a large difference between rxy.z and rxy, z is likely to be a lurking variable.

**Forecasting**

~ four possible series patterns are trend, seasonal, cyclical and random.




***Moving average***

The n-period moving average forecast Ft+1, for time period t + 1 is given by:

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-n+1}}{n}$$

***Exponential smoothing***

$$F_{t+1} = Y_t + \alpha(F_t - Y_t)$$

where
Yt is the actual value in period t
Ft is the previous forecast (for period t)
α is constant, 0 ≤ α ≤ 1, called the 'damping factor'.

The initial forecast is set equal to the last period actual value: F1 = Y0.

## WEEK 6: ANALYSIS OF CATEGORICAL DATA

- conduct a chi-square goodness of fit test
- evaluate the results of the goodness of fit test

### Contingency tables

| Netflix | 0 | 1 | 2 | 3 or more |
|---|---|---|---|---|
| Yes | 48 | 37 | 86 | 36 |
| No | 72 | 53 | 54 | 14 |
| Total | 120 | 90 | 140 | 50 |

### Chi-square goodness of fit test

~ used as an analysis of a single variable problem.

H0: (null hypothesis) the variable follows a hypothesized distribution.
HA: (alternative hypothesis) the variable does not follow a hypothesized distribution.

⇒ the test comparing the observed frequencies of occurrence of the categories and the expected frequencies

$$T = CHI^2 = \sum_{i=1}^{N} \frac{(OF_i - EF_i)^2}{EF_i}$$

where
N is the number of categories
OFi is the observed frequency of category i
EFi is the 'expected' frequency of the category, meaning the frequencies predicted by the hypothesis.

## Example

Review the calculation of the test statistic below.

| Outcome (i) | 1 | 2 | 3 | 4 | 5 | 6 | total |
|---|---|---|---|---|---|---|---|
| $EF_i$ | 50 | 50 | 50 | 50 | 50 | 50 | |
| $OF_i$ | 48 | 57 | 60 | 42 | 44 | 49 | |
| $\frac{(OF_i - EF_i)^2}{EF_i}$ | 0.08 | 0.98 | 2.00 | 1.28 | 0.72 | 0.02 | TS=5.08 |

The **degrees of freedom (df)** for our test statistic TS is N-1=6-1+5.

Decision Rule: If T > CV you will take this as evidence that the hypothesis used to generate the expected frequencies is doubtful and the probability that you would get this value of the test statistic from the data using that hypothesis is low, e.g. less than α = 5%. Therefore, you will reject the hypothesis.

Null Hypothesis (H0): There is no association between gender and preference for a particular soft drink brand. In other words, the preferences for soft drink brands are independent of gender.

Alternative Hypothesis (H1): There is an association between gender and preference for a particular soft drink brand. In other words, the preferences for soft drink brands are not independent of gender.

In this example, the null hypothesis assumes that there is no relationship between the two categorical variables (gender and soft drink brand preference), while the alternative hypothesis suggests that there is a relationship between them.

The chi-square test will calculate a test statistic based on the observed and expected frequencies in the contingency table. If the test statistic is large enough (exceeds a critical value based on the chosen significance level), the null hypothesis is rejected, indicating that there is evidence of an association between gender and soft drink brand preference. If the test statistic is small, the null hypothesis cannot be rejected, suggesting that there is no evidence of an association between the two variables.

It's important to note that the chi-square test does not provide information about the strength or direction of the association; it only determines whether an association exists or not.

**Testing the hypothesis of independence**

## Contingency tables for testing the hypothesis of independence

As previously mentioned, you can generalise the previous chi-square technique to the case where two variables are involved.

| | | Variable B | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | row total |
| Variable A | Level 1 | $F_{1,1}$ | $F_{1,2}$ | $F_{1,1} + F_{1,2}$ |
| | Level 2 | $F_{2,1}$ | $F_{2,2}$ | $F_{2,1} + F_{2,2}$ |
| | column total | $F_{1,1} + F_{2,1}$ | $F_{1,2} + F_{2,2}$ | GRANDTOTAL |

$$GRANDTOTAL = F_{1,1} + F_{1,2} + F_{2,1} + F_{2,2}$$

### *Sample size for means*

Based on the above equations, what value would you use for *s*, the estimated standard deviation of the data?

Consider that you can take two different approaches. What would they be?

**Typical approach 1:** You can use the population standard deviation $\sigma$ if you know it (but usually you don't).
**Typical approach 2:** You can perform a small pilot study with $n = 25$ and estimate *s* from that.

For higher accuracy expand the pilot study to give more accurate results.
**Which approach is more likely to be encountered in practice?**
Approach 2 is more likely to be encountered in practice.
To apply it, you will need to:
• have made an estimate of s
• decide on how confident you want to be in the level of accuracy.

## WEEK 7: TWO-SAMPLE TESTS

- conduct two-sample t-tests and paired t-tests

- evaluate the results of two-sample t-tests and paired t-tests

    ~ used to determine whether or not two population means are equal.

**Test Objectives**

1. **Testing if a treatment is effective or not.**
   Paired data: for each patient, we have both measurements before and after taking the drug. Types of test: **1) paired t-test; 2) sign test**.
    a. data ~ normal distributions: use **paired t-test**.
    b. data not ~ normal distribution: use the **sign test**.
        *(**sign test** has lower power than the two-sample t-test, works best for large samples (at least 25), and is also used for small samples)*

2. **Testing if the two populations have different means.**
    Types of test: **1) two sample z-test; 2) two sample t-test; 3) two sample Wilcoxon test.**
    a. Use the **z-test** if i) the two samples are from normal distributions; ii) the distributions have the same standard deviation that is known; iii) the samples are large.

    b. Use **t-test** if i) the two samples are from normal distributions; ii) the standard deviations are unknown. Condition i) is often taken for granted.

    c. Use the **Wilcoxon test** if the two samples are from non-normal distributions.

**Paired data**

1. ***Paired t-test***

**Step 1:** Hypothesis construction
H0: there is no difference between the two samples, $\mu d = 0$
Ha: there is a difference, $\mu d \neq 0$

where
$\mu d$ is the population mean the difference between the pairs of observations.

**Step 2:** Calculate the actual differences between the values in the two samples.
**Step 3:** Calculate the mean x and the standard deviations of the values of the differences d.
**Step 4:** Calculate the value of the paired t-test statistic:

$$t = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$$

**Step 5:** Look up the critical value for the degree of freedom *n-1*, and the designed significance level.

> ***Note:*** *the paired t-test is actually a one-sample test applied to the difference between the two samples. The only difference is that you use t-statistic no matter what the sample size is.*
>
> *In the case of one sample test, use z-statistic when the sample size is large and only use t-statistic when the sample size is small. The assumption of normal distribution for the population is also required for one sample test when the sample size is small.*

### 2. *Sign test*

**Step 1:** Construct null hypothesis

H0: μ1 = μ2 | Ha: μ1 ≠ μ2

**Step 2:** Subtract, pairwise, the values in sample 2 from the corresponding value in sample 1. Record a '+' sign if this difference is positive and a '-' sign if the difference is negative. If two observations are equal, record ½ '+' and ½ '-'.

Let **x = number of "+" sign**.

**Step 3:** Calculate the value of the z-statistic:

$$z = \frac{x - n/2}{\sqrt{n}/2}$$

**Step 4:** If |z|> critical value, reject H0.

## Independent data

*The two-sample z-test assumes that the two populations have the same standard deviation and the standard deviation is known. When this is not the case, you use t-test.*

*Wilcoxon test is a non-parametric test that relies on only the ranks of the observations. It does not require that the data are from normal distributions, or the samples are large. It is sometimes referred to as Mann-Whitney test.*

## 1. *Two-sample z-test*

Step 1: The hypotheses for the z-test are

**H0: μ1 = μ2 | Ha: μ1 ≠ μ2**

Step 2: You use the following formula to calculate the test statistic:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Step 3: Finally, you compare the z-value with CV and draw a conclusion.

## 2. *Two-sample t-test*

Step 1: The t-test assumes that the two samples are from normal distributions. The hypotheses for the t-test are:

H0: μ1 = μ2 | Ha: μ1 ≠ μ2

Step 2: The test statistic can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

p s is the pooled standard deviation given by:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}.$$

Step 3: After calculating the TS, look up the CV in Table 4 Critical values for two-sided t distribution. Use n1 + n2 - 2 to determine the degree of freedom at the desired level of significance.

Step 4: Finally, compare t-statistics with CV and draw a conclusion.

### *3. Two-sample Wilcoxon test*

Suppose you have two samples x and y. The sizes of the two samples are *nx* and *ny;* and the combined sample has a size of N (sum of *nx* and *ny*).

The steps for the Wilcoxon test:

**Step 1:** Form a hypothesis:

- H0: there is no difference between the two samples
- Ha: there is a difference between the two samples

**Step 2:** Rank the pooled data set (*x* and *y* samples combined). For tied observations, use the average of the ranks they take up.

**Step 3:** Calculate *W* which is the sum of the ranks belonging to *x* observations.

**Step 4:** Calculate *T = W - nx (N-1)/2.*

**Step 5:** Calculate z-statistic:

$$z = \frac{T}{\sqrt{n_x n_y (N+1)/12}}$$

**Step 6:** Look up the critical value in Table 5 if N ≤ 12 or the usual critical z values if N>12.

**Step 7:** Compare z-statistic with CV and draw conclusions.

## WEEK 8: ANALYSIS OF VARIANCE

- conduct one-way and two-way ANOVA
- evaluate the results of one-way and two-way ANOVA

The null hypothesis for one-way ANOVA is H0: There is no difference between the n groups being compared:

**H0 = μ1 = μ2 = .... = μn**

The alternative hypothesis is Ha: There is no difference between the n groups being compared:

H0 = μ1, μ2 ,..., μn **are not equal**

***Assumptions for ANOVA***

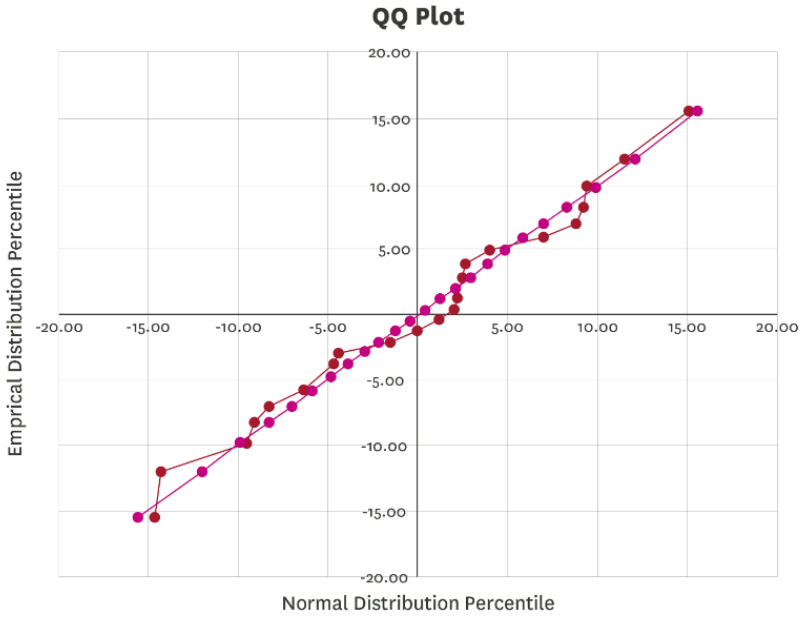
- The populations from which the samples were drawn are normally distributed.
- The populations all have the same standard deviation.
- The samples selected from these populations are independent and random.

⇒ A test statistic (TS), we denote with F.

***The process for testing the null hypothesis:***

1. Compute the test statistic (TS) from the sample.
2. Identify the parameters df1 and df2 of the F distribution, this test statistic would have a distribution if the null hypothesis is true.
3. Decide on the level of significance alpha for the test.
4. Compute the critical value (CV) for the F distribution that corresponds to the significance level alpha (Excel has a function that can do this).
5. Check whether the TS is higher than the CV and if so reject the null hypothesis and if not do not reject it.
6. Alternatively, compute the p-value for the TS obtained from the F distribution (Excel has a built-in function that can do this).
7. Check if the p-value is lower than the significance level: if so, reject the null hypothesis; if not then do not reject it.

**Normality Testing**




The null hypothesis is that data follow a normal distribution.
If the p-value is lower than 5%, you reject the null, otherwise, you do not.

**WEEK 9: BINARY LOGISTIC REGRESSION**

~ In Excel (Assignment 2)

(b) To predict the probability that a 7-year-old wagon in average condition will sell at auction using the logistic regression model constructed above, we have the formula for the probability of selling at auction as follows.

$$\text{P (Y=1, sold at auction)} = \frac{e^{\beta_0+\beta_1\times Age+\beta_2\times Condition}}{1+ e^{\beta_0+\beta_1\times Age+\beta_2\times Condition}}$$

As constructed in question 1a, we have:
$\beta_0 \approx 0.7587$
$\beta_1 \approx -0.4868$
$\beta_2 \approx 3.2106$

Given:
Age of the wagon = 7 (in years)
Condition = 0 (for average condition)

Thus, substitute to the equation above, we have:

$$\text{P (Y=1)} \approx \frac{e^{0.7587 - 0.4868\times 7 + 3.2106\times 0}}{1+ e^{0.7587 - 0.4868\times 7 + 3.2106\times 0}} \approx 0.0661$$

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 50 | | | *coeff b* | *s.e.* | *Wald* | *p-value* | *exp(b)* | *lower* | *upper* |
| 51 | | Intercept | 0.75866104 | 1.562302952 | 0.2358115 | 0.6272477 | 2.13542 | | |
| 52 | | X1 (Age) | -0.486756233 | 0.376927637 | 1.6676583 | 0.1965725 | 0.61462 | 0.2936 | 1.28661 |
| 53 | | X2 (Condition) | 3.210648662 | 1.447464045 | 4.9200555 | 0.0265467 | 24.7952 | 1.45306 | 423.107 |
| 54 | | | | | | | | | |
| 55 | | Odds Ratio (X1) = | | 0.6146 | = EXP(C52) | | | | |
| 56 | | Odds Ratio (X2) = | | 24.7952 | =EXP(C53) | | | | |
| 57 | | | | | | | | | |
| 58 | (b) | **P (Y=1) =** | | 0.0661 | =EXP(C51+C52*7+C53*0)/(1+EXP(C51+C52*7+C53*0)) | | | | |
| 59 | | | | 6.61% | | | | | |

*Figure 5: Compute the probability of selling at auction of the given wagon (in Excel)*

The estimated probability that a 7-year-old wagon in average condition will sell at auction is approximately 6.61%. Using the cut-off at 0.5, we can infer that this wagon is unlikely to be sold at the auction. However, considering the model's 90.91% accuracy in predicting unsuccessful cases, this outcome is highly probable.

## WEEK 10: STATISTICAL DISTRIBUTIONS AND PROBABILITIES

- How to obtain the 95% confidence interval for the probability of an event that occurs for the first time
  **The rule of 3** says that a 95% confidence interval for P is (0.3/N).

*For example, what is a 95% confidence interval for the probability that for the first time, tomorrow the lift will stop working, given that it has been in operation for 1,800 days trouble-free?*

**Solution**

The upper confidence limit is 3/1800.
The lower confidence limit is 0.

- How to calculate the probability of k successes out of N trials (**binomial distribution**)

$$\frac{N!}{K! \times (N-K)!} \times P^K \times (1-P)^{(N-K)}$$

for K = 0, 1, 2, … N

In Excel, the function is BINOM.DIST and is computed as follows:

BINOM.DIST(number_s, trials, probability_s, probability_s_cumulative)

where
number_s = number of successes
trials = total number of trials
probability_s = probability of success on each trial
probability_s_cumulative = TRUE returns the cumulative probability, FALSE returns the exact probability

- How to calculate the probability that k events occur over a given time period when we know the average number of events that occur over the same period (**Poisson distribution**)

Given a rate parameter λ which is the average number of events per unit of time. The probability of getting x events during the time interval is:

$$\Pr(X = x) = e^{-\lambda} \times \lambda^x \times \frac{1}{x!}$$

where
e is Euler's number (e = 2.71828)
! is the factorial function

In Excel the function is POISSON.DIST and is computed as follows:
POISSON.DIST(x, λ, cumulative)

where
x = the number of events
λ = the mean, the expected number of events
Cumulative = TRUE indicates the probability of many events happening between 0 and x or FALSE suggests the probability of the number of events happening the same as the x.

- How to calculate the **conditional probability**
  - Pr(A) = 33.8%
  - Pr(A|male) = Pr(A|B) = 27.1
  - Pr(A|female) = Pr(A| not B) = 40.5%

  **Joint probability**
  - The probability of being male and surviving 30 years is 0.50*0.271=13.53%. This is denoted Pr(*A&B*).
  - The probability of being female and surviving 30 years is 0.50*0.405=20.25%. This is denoted Pr(A & not B).
  - The probability of surviving for 30 years from age 60 is not independent of gender.
  - For any 2 events A and B we have Pr*(A|B)* = Pr*(A&B)* /Pr*(B).*
  - If events A and B are independent, then Pr(*A*|*B*) = Pr(*A*) or in other words event B has no effect on the probability of event A.

- How to evaluate survey reliability (Cronbach's alpha)
  Cronbach's alpha is obtained by:
  1. Computing the sample standard deviations of the scores for q1, q2, q3, q4, q5 and the total.
  2. Computing the sample variances of the scores for q1, q2, q3, q4, q5 and the total.
  3. Computing the difference between total variance and the square sum of 5 questions.

  The formula of Cronbach's alpha:

$$\alpha = \frac{K}{K-1} \times \frac{VAR(TOTAL) - (VAR(Q1) + VAR(Q2) + VAR(Q3) + VAR(Q4) + VAR(Q5))}{VAR(TOTAL)}$$

**WEEK 11: ADVANCED CORRELATION AND FORECASTING**

Correlation coefficient =

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad and \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad amd \; S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$r = \left[\left(\sum_{i=1}^{n} X_i Y_i\right) - n \times (\bar{X} \times \bar{Y})\right] \times \frac{1}{(n-1) \times S_X \times S_Y}$$

**Test for significance:**

- The methodology for testing the hypothesis is
  - 1) Compute the sample correlation coefficient from the data.
  - 2) Compute the test statistic $\boldsymbol{TS} = abs(\boldsymbol{T}) = \frac{\boldsymbol{abs(R)}}{\sqrt{\boldsymbol{1-R^2}}} \times \sqrt{(\boldsymbol{N-2})}$
    - This test statistic has the t distribution with N – 2 degree of freedom
  - 3) Compute the critical value CV from the t distribution (T.INV.2T(α, N-2)) where α is the level of significance
  - 4) Compare the TS and the CV and if TS > CV
  - 5) OR Compute the p-value for the test statistic using the t distribution (T.DIST.ST(TS, N-2)) and compare to α and if p-value < α, we reject H0

Spearman's rank correlation coefficient

- The computation method
  - 1) Set up the data in columns or rows so we have the ranks given by the 2 judges side by side for all N pairs of ranks
  - 2) Compute the differences between the ranks as shown above
  - 3) Compute the square of the differences D
  - 4) Sum the squared differences to get $\boldsymbol{SUMSQDIFF} = \sum_{I=1}^{N} \boldsymbol{D_i^2}$
  - 5) Compute the spearman rank correlation

$$\boldsymbol{R_S = 1 - \frac{6 \times SUMSQDIFF}{N^3 - N}}$$

$$SUMSQDIFF = \sum_{I=1}^{N} D_i^2 = 24 = 1 + 1 + 0 + 9 + 4 + 4 + 4 + 1$$

$$N^3 - N = 8^3 - 8 = 504$$

$$R_S = 1 - \frac{6 \times 24}{504} = 0.7143$$

**WEEK 12: COURSE REVISION**

- Apply the concept of binomial distribution to gambling
- Binomial
    - N trials
    - Each trial has two possible outcomes: success or failure
    - probability of success is p
- As N gets larger, the distribution of the number of successes gets closer to a normal distribution.
- A normal distribution can be used to approximate the distribution of the number of successes
- Normal distribution is convenient to use since we know how to construct confidence intervals from this distribution