Unit 1

$$\text{Sample fraction} = \frac{\text{Sample size}}{\text{Study population}}$$

$$\text{Interval size} = \frac{\text{Study population}}{\text{Sample size}}$$

Unit 2

Measures of central tendency:

Mean; Median; Mode

Sensitive to extreme values

1) Mean: مجموع القيم على عددها

- For population: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

- For Sample: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Commonly Used Symbols

For a Sample

$\bar{x}$	sample mean
s^2	sample variance
s	sample standard deviation

For a Population

μ	population mean
σ^2	population variance
σ	population standard deviation

Not affected by extreme values

2) Median: middle value of the ordered data

لازم نرتب البيانات

median (n+1)/2 = ترتيب القيمة المطلوبه

$\Sigma = |X_i - \text{Median}| = \text{minimum}$

Example:

10 20 30 40 50 60

Note: Data has been ordered from lowest to highest. Since n is even (n=6), the median is the (n+1)/2 ordered observation, or the 3.5th observation, *i.e.*, the average of observation 3 and observation 4.

Answer: The median is 35.

$\frac{6+1}{2} = \frac{7}{2} = 3.5^{th} = \left(\frac{3^{th} + 4^{th}}{2}\right) = \frac{30+40}{2} = 35$

★ For more than one sample:

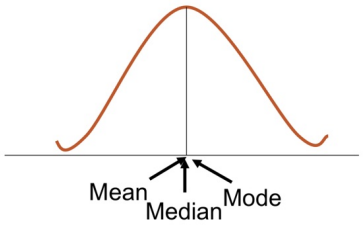
$$\bar{x}_{pooled} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

3) Mode : القيمه الاكثر تكراراً

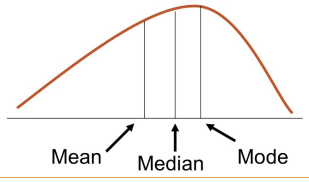
In a normal distribution, the mode , the median, and the mean have the same value .

Skewness

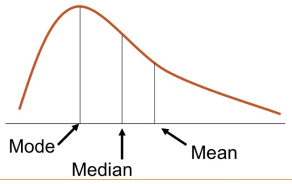
Symmetric (Right and left sides are mirror images)
- Left tail looks like right tail
- Mean = Median = Mode



Left skewed (negatively skewed)
- Long left tail
- Mean < Median



Right skewed (positively skewed)
- Long right tail
- Mean > Median



Quartiles:

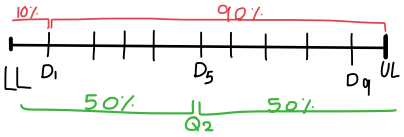
Q1: 25% of the observations are smaller than Q1 and 75% of the observations are larger

Q2: 50% of the observations are smaller than Q2 and 50% of the observations are larger.
Same as the Median.

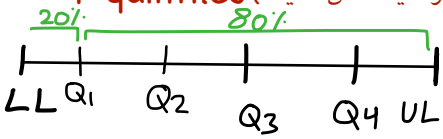
Q3: 75% of the observations are smaller than Q3and 25% of the observations are larger

Other Quantiles

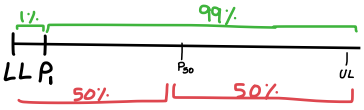
9-Deciles (كل جزء فيه عُشر العينه)



4-quintiles (كل جزء فيه خُمس العينه)

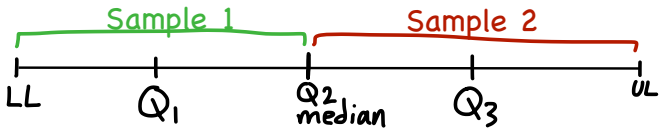


99-percentiles



P20 = D2 = Q1
P50 = D5 = Q2= Median
P75 = Q3
P80 = D8 =Q4
P90 = D9

حساب Q1، Q2،Q3



١) نطلع ال median
٢) نتخيل الجزء يللي على اليمين عينه جديده ونطلع ال median تبعه (يللي هو Q3)
٣) ويللي على اليسار كمان ونطلع ال median تبعها (يللي هو Q1)
٤) بعدين نحسب IQR

Measures of Dispersion

1) Range :

$$Range = UL - LL$$

2) Standard Deviation :

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \Longrightarrow \text{3) Variance: } S^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

- كل ما كان SD ↑ ⟵ كانت القيم متباعدة
- كل ما كان ال SD ↓ ⟵ كانت القيم متقاربة

* Take a look on page 54 in slides .

Note that $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

- You divide by N only when you have taken a census and therefore know the population mean. This is rarely the case.
- Normally, we work with a sample and calculate sample measures, like the sample mean and the sample standard deviation:

4) Coefficient of Variation (CV):

$$CV = \frac{S}{\bar{X}} * (100\%)$$

- $S < \bar{X} \longrightarrow CV < 100\%$ ⇒ تباعد قليل
- $S > \bar{X} \longrightarrow Cv > 100\%$ ⇒ تباعد كبير
- $S = \bar{X} \longrightarrow Cv = 100\%$

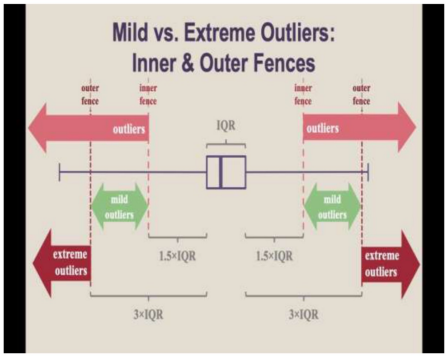
Not sensitive to extreme values

5) The Interquartile range(IQR) :

$$IQR = Q3 - Q1$$

A mild outlier is a data value that lies between 1.5 and 3 times the IQR below Q1 or above Q3.

Extreme outlier is a data value that is more that three times the IQR below Q1 or above Q3.



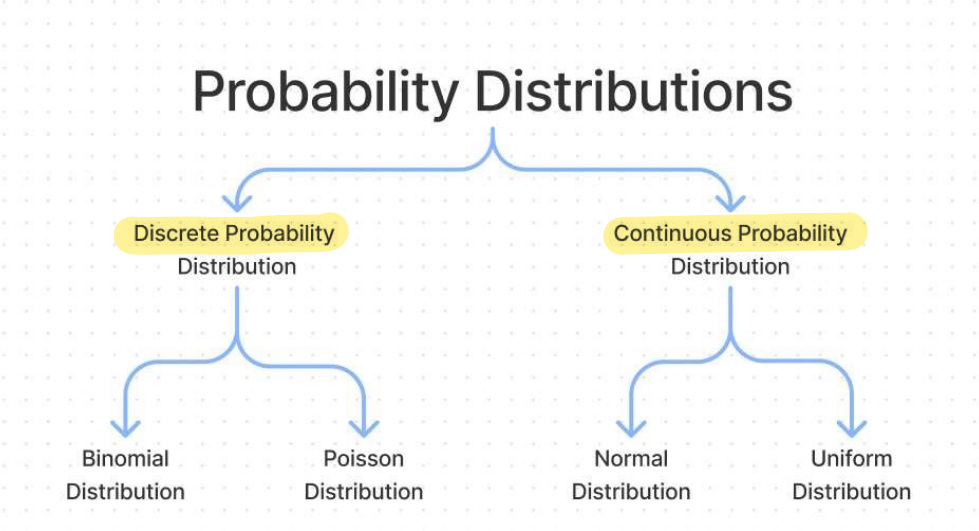
Outliers ⟶ Less than : Q1-1.5* IQR
⟶ More than : Q3+1.5*IQQ

(Q1-1.5* IQR , Q3+1.5*IQR) outlier اي قيمه خارج هذا النطاق تعتبر

* Five Number Summary

Lower limit(Smallest value) , Q1 , Q2(median) , Q3 , Upper limit (Largest value)

Unit 3



Discrete Random Variable

Value of X	x_1	x_2	x_3	⋯	x_k
Probability	p_1	p_2	p_3	⋯	p_k

1. Every probability p_i is a number between 0 and 1.
2. The sum of the probabilities must be 1.

Find the probabilities of any event by adding the probabilities of the particular values that make up the event.

* Take a look on page 10 in slides.

Normal Distribution

Mean = 0 , SD = 1

-The normal curve is unimodal and symmetric about its mean (m).
-In this distribution the mean, median and mode are all identical.
-The standard deviation (s) specifies the amount of dispersion around the mean.
-The two parameters m and s completely define a normal curve.

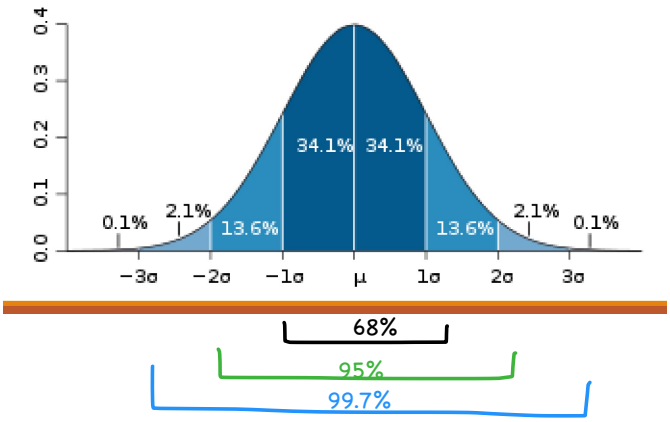
-The probability that X = any particular value is 0.
-Consequently, we talk about intervals. The probability is = to the area under the curve.
-The area under the whole curve = 1.

※ Properties of a Normal Distribution

1. It is symmetrical about m .
2. The mean, median and mode are all equal.
3. The total area under the curve above the x-axis is 1 square unit. Therefore 50% is to the right of m and 50% is to the left of m.
4. Perpendiculars of:
   ±1 s contain about 68%;
   ±2 s contain about 95%;
   ±3 s contain about 99.7%

of the area under the curve.

The normal distribution



The standard z score is obtained by creating a variable z whose value is

$$Z = \frac{X - \overline{X}}{s}$$

نستعملها لما يكون ال mean لا يساوي 0
وال SD لا يساوي 1 .
نحول قيم X ل Z وناخذ القيم من الجدول .

Standardizing Data: Z-Scores

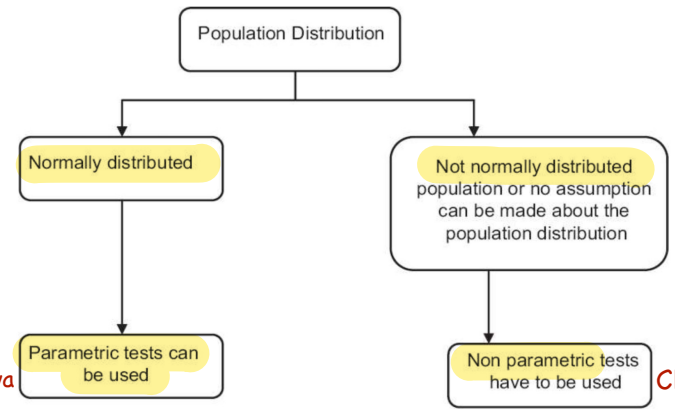
- We can convert the original scores to new scores with $\overline{X} = 0$ and $s = 1$.
- This will give us a pure number with no units of measurement.
- Any score below the mean will now be negative.
- Any score at the mean will be 0.
- Any score above the mean will be positive.

السلايد فيه امثلة على موضوع ال z-score ،
شوفوهم 👌

Central Limit Theorem

شوفوها بالسلايد افضل 😊

Unit 4



t-test/one way Anova

Chi-square

Important

The chi square test can only be used on data that has the following characteristics:

- The data must be in the form of frequencies
- The frequency data must have a precise numerical value and must be organised into categories or groups.
- The expected frequency in any one cell of the table must be greater than 5.
- The total number of observations must be greater than 20.

Degree	Frequency	Relative Frequency	Percentage
High School	2	0.050	5.0
Bachelor's	7	0.175	17.5
MBA	20	0.500	50.0
Master's	3	0.075	7.5
Law	4	0.100	10.0
PhD	4	0.100	10.0
	40		

$$X^2 = \frac{\Sigma (O - E)^2}{E}$$

X^2 = The value of chi square
O = The observed value
E = The expected value
$\Sigma (O - E)^2$ = all the values of $(O - E)$ squared then added together

* we use chi-square when both dependent and independent variable are "nominal" ⟶ Categorical

For example :gender, type of car and yes or no

Steps in test of hypothesis:

1)Determine the appropriate test

dependent and independent variable or nominal

2)Establish the level of significance: α

acceptable margin of error (type I) = .05(5%)

level of confidence(1- α) = 95%

usually .05 , but it can be 0.01 or 0.001

3)formulate the statical hypothesis :

Ha: P1 ≠ P2 ⟶ There is a relationship between IV and DV

Ho: P1 = P2 ⟶ There is no relationship between IV and DV

4) Calculating test statistic:

X calc. ⟶ Calculated value of chi-squared

$$\chi^2 = \sum \left[\frac{(F_o - F_e)^2}{F_e} \right]$$

Observed frequencies; Expected frequency; Expected frequency

$$\text{Expected frequency} = \frac{\text{row total x column total}}{\text{Grand total}}$$

5) Determine the degree of freedom :

df (r) = (R-1) (C-1)

no. of levels in row — no. of levels in column

6) Compare the computer test statistic X calculated against a table/ critical value

If X^2 calc. > X^2 table ⟶ reject Ho // "it was successful to prove my own hypothesis and rejected and no hypothesis which mean there was an statical significant relationship between the 2 variables"

If X^2 calc. < X^2 table ⟶ Keep Ho

* Decision and interpretation

- When X calc. X table

We are 95% confidence that there statistically significant relationship between independent and dependent variable on the population level.

- When X calc. X table

There's no relationship.

Unit 5

Logically, The larger the difference in means, the more likely to find a significant t test.

But, recall:

1.Variability

More variability = less overlap = larger difference

2. Sample size

Larger sample size = less variability (pop) = larger difference

Assumption

1. Dependent variable should be continuous (I/R)
2. The groups should be randomly drawn from normally distributed and independent populations

e.g. Male X Female //Dentist X Physician // Manager X Staff

NO OVER LAP

3. the independent variable is categorical with two levels
4. Distribution for the two independent variables is normal
5. Equal variance (homogeneity of variance)
6. large variation = less likely to have sig t test = accepting null hypothesis (fail to reject) = Type II error = a threat to power

$$t_{\bar{X}1-\bar{X}2} = \frac{\bar{X}1 - \bar{X}2}{SE_{diff}} \longrightarrow SE_{diff} = \sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}$$

Independent sample step 1

1- set alpha equal 0.5
2- State hypothesis
Null Ho : $\mu_1 = \mu_2$
Alternative is H1: $\mu_1 \neq \mu_2$

3- Calculate test static
4- determine the critical value(df = N1+ N2 - 2)
5- state decision rule
if |t calc. | > t table , reject the null.
6- conclusion: reject the null. population means are different. there is a relationship between two variables.

Dependent sample t-test

$$t_{\bar{X}_D} = \frac{\bar{D}}{SE_{diff}} \qquad SE_{diff} = \frac{SD_D}{\sqrt{n_{Pairs}}} \qquad \text{Another way} \longrightarrow t_{\bar{X}_D} = \frac{\bar{D}}{SD_D / \sqrt{n_{Pairs}}}$$

1- Set alpha = .05
2- Null hypothesis:
H0: $\mu_1 = \mu_2$ ⟶ لا فرق /لا علاقة
Alternative is H1: $\mu_1 \neq \mu_2$ ⟶ هناك فرق/ هناك علاقة
3- Calculate the test statistic:

4- Determine the critical value of t. Alpha =.05
df = N(pairs)-1
5- Decision rule: is absolute value of sample value larger than critical value?
6. Conclusion.

-يَوْمًا مَا ستشكر نَفْسِك أَنَّك لَمْ تستسلم .

Good luck 💛
By : Lara Saqer