CHAPTER 1 - Intro to Stats

Descriptive statistics: Numbers that are used to summarize and describe data. Just descriptive, do not involve generalizing beyond the data at hand. Ex: tabular listing, graph, report

Inferential statistics: Provides a variety of tests and tools for generalizing beyond collections of actual observations. Permits us to use a relatively small collection of actual observations to evaluate. Ex: hypothesis, assertion

Three types of data: Quantitative, Ranked, Qualitative

Quantitative data: consist of numbers that represent an amount or a count. Ex: weights of students

Ranked data: also consists of numbers, but represents relative standing within a group. Ex: if the weights of students are represented as 1 – 100 sorted by lightest to heaviest

Qualitative data: Generally, consists of words, letters, or numerical codes that represent a class / category. Ex: Replies of students

Level of measurement: Specifies the extent to which a number/word/letter represents some attribute. Three levels: nominal, ordinal, and interval / ratio, paired with qualitative, ranked, and quantitative data respectively.

If people are classified as Male / Female, data is qualitative and measure is nominal. Nominal measurement is classification. Ex: classifying mood disorders as manic, bipolar, or depressive

When any single number indicates only relative standing, data is ranked and measurement is ordinal. Ordinal is order. We know the first place in a race is faster than second, but not by how much.

The important properties of interval / ratio are equal intervals and a true zero. Ex: interval between 60 and 70kg is same as 70 and 80kg, while 0kg denotes the absence of a person on the scale. When numbers represent nonphysical characters, interval / ratio is often questionable.

Quantitative variables can be split into two: discrete and continuous. A discrete variable consists of isolated numbers separated by gaps. Ex: number of children in family. A continuous variable consists of numbers whose values have no restrictions (at least in theory). Ex: weight, reaction times

Independent variable: The variable that is manipulated by the experimenter.

Dependent variable: The variable that is affected by the independent variable.

CHAPTER 2 - Plots

Study how to make a plot in Excel biar mempermudah hidup.

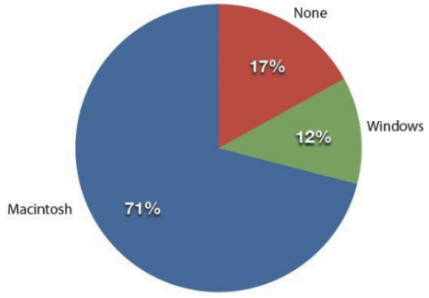
Graphing Qualitative Variables

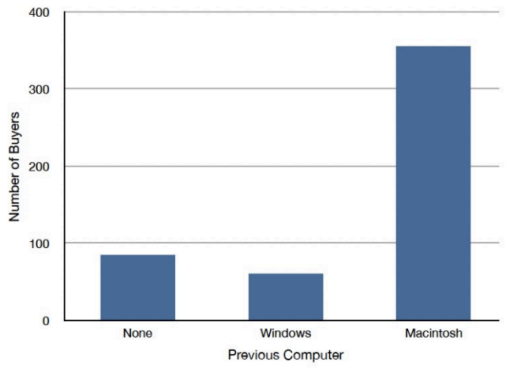
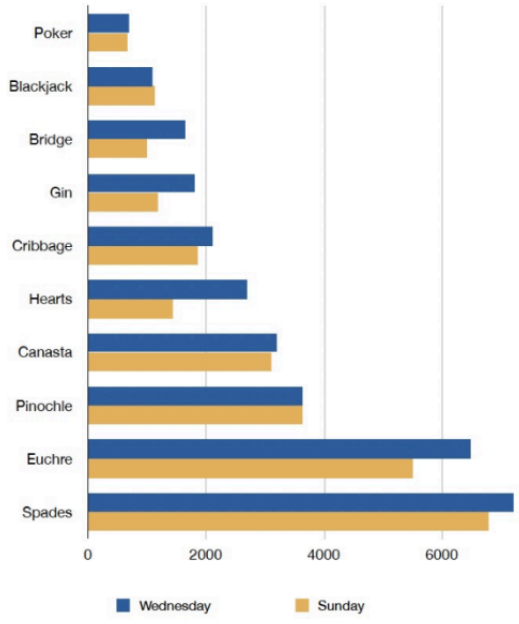
- Frequency Tables

  Example: Frequency Table for the iMac Data

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1

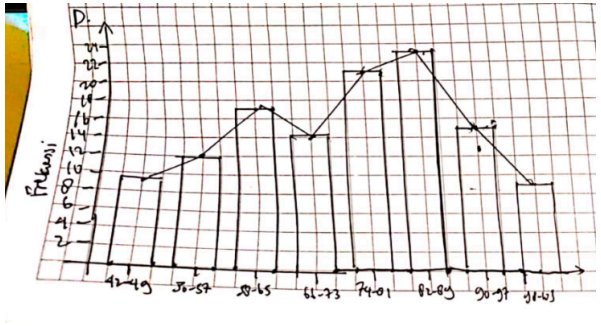
0.17 = 85/500

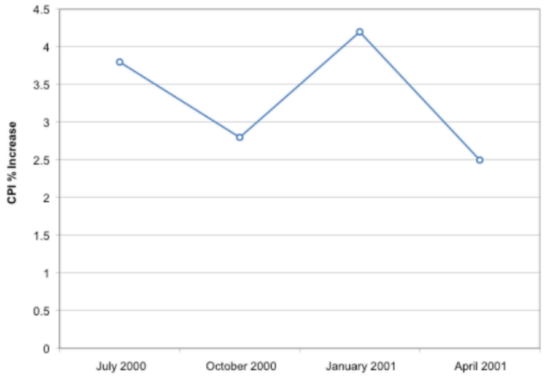
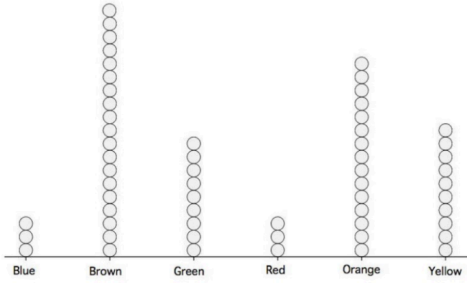
Plot Name	Plot Description	Image
Pie Chart	Pie chart of iMac purchases illustrating frequencies of previous computer ownership	

Bar Chart	Bar charts can also be used to represent frequencies of different categories.	 

Graphing Quantitative Variables

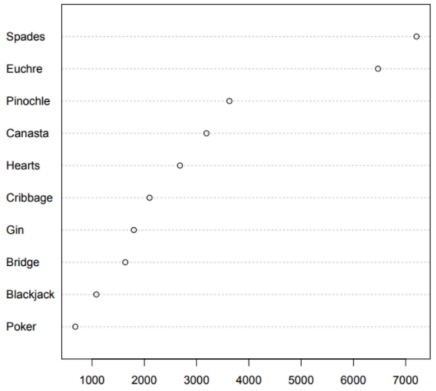
Name	Explanations	Image
Stem & Leaf Displays	• A stem and leaf display of the data is shown in Fig 1. • The left portion of the figure contains the stems. • They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits • The numbers to the right of the bar are leaves, and they represent the 1's digits. • Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem. • We can make our figure (Fig 2) even more revealing by splitting each stem into two parts • The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table. • The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table.	37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6 Sample Data 3|2337 2|001112223889 1|2244456888899 0|69 Stem & Leaf Display (Fig 1) 3|7 3|233 2|889 2|001112223 1|56888899 1|22444 0|69 Stem & Leaf Display (Fig 2)

<table>
<tr>
<td></td>
<td>• Figure 3 compares the numbers of TD passes in the 1998 and 2000 seasons.
• The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data.</td>
<td>
<table>
<tr><td>11</td><td>4</td><td></td></tr>
<tr><td></td><td>3</td><td>7</td></tr>
<tr><td>332</td><td>3</td><td>233</td></tr>
<tr><td>8865</td><td>2</td><td>889</td></tr>
<tr><td>44331110</td><td>2</td><td>001112223</td></tr>
<tr><td>987776665</td><td>1</td><td>56888899</td></tr>
<tr><td>321</td><td>1</td><td>22444</td></tr>
<tr><td>7</td><td>0</td><td>69</td></tr>
</table>
Stem & Leaf display (Fig 3)
</td>
</tr>
<tr>
<td>Histogram</td>
<td>Seorang Guru melakukan survey nilai ujian matematika terhadap 120 siswa kelas 12 SMA Pelita Cemerlang dan mendapatkan hasil sebagai berikut:
<table>
<tr><td>56</td><td>77</td><td>78</td><td>90</td><td>100</td><td>45</td><td>78</td><td>89</td><td>80</td><td>67</td><td>89</td><td>92</td></tr>
<tr><td>78</td><td>65</td><td>85</td><td>48</td><td>59</td><td>77</td><td>52</td><td>87</td><td>55</td><td>75</td><td>46</td><td>51</td></tr>
<tr><td>58</td><td>91</td><td>95</td><td>73</td><td>61</td><td>81</td><td>99</td><td>84</td><td>59</td><td>62</td><td>73</td><td>84</td></tr>
<tr><td>66</td><td>50</td><td>97</td><td>80</td><td>74</td><td>49</td><td>53</td><td>66</td><td>54</td><td>83</td><td>98</td><td>79</td></tr>
<tr><td>61</td><td>58</td><td>99</td><td>71</td><td>79</td><td>82</td><td>64</td><td>70</td><td>88</td><td>83</td><td>72</td><td>49</td></tr>
<tr><td>86</td><td>77</td><td>52</td><td>63</td><td>91</td><td>94</td><td>84</td><td>42</td><td>92</td><td>88</td><td>50</td><td>75</td></tr>
<tr><td>79</td><td>100</td><td>44</td><td>90</td><td>83</td><td>71</td><td>61</td><td>60</td><td>90</td><td>83</td><td>55</td><td>60</td></tr>
<tr><td>66</td><td>71</td><td>82</td><td>93</td><td>87</td><td>76</td><td>48</td><td>54</td><td>61</td><td>100</td><td>95</td><td>85</td></tr>
<tr><td>73</td><td>88</td><td>99</td><td>43</td><td>59</td><td>85</td><td>75</td><td>90</td><td>86</td><td>73</td><td>75</td><td>95</td></tr>
<tr><td>100</td><td>95</td><td>65</td><td>77</td><td>66</td><td>55</td><td>88</td><td>99</td><td>83</td><td>81</td><td>79</td><td>59</td></tr>
</table>
1) Berapa jumlah kelas data yang ideal untuk data diatas?
2) Hitunglah jumlah interval masing-masing kelas!
3) Sajikan data kedalam distribusi frekuensi!
4) Gambarkan diagram histogram, polygon, dan ogive untuk data diatas!
You are allowed to use Excel.
</td>
<td> 
Histogram and Polygon</td>
</tr>
<tr>
<td>Freq Polygon</td>
<td>https://byjus.com/maths/frequency-polygons/</td>
<td></td>
</tr>
</table>

Line Graph	Obvious lah y (kurang anjing)	
Dot Plots	Dot Plots • Dot plots can be used to display various types of information. • Figure uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. • Each dot represents a single M & M. • From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc. 	

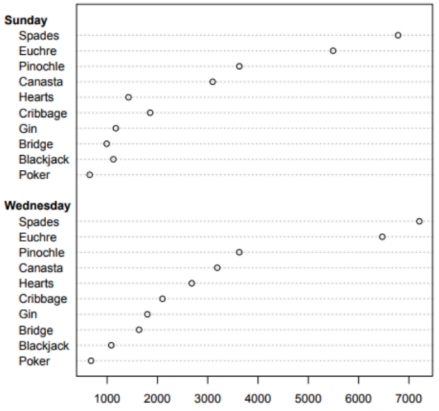
Dot Plots

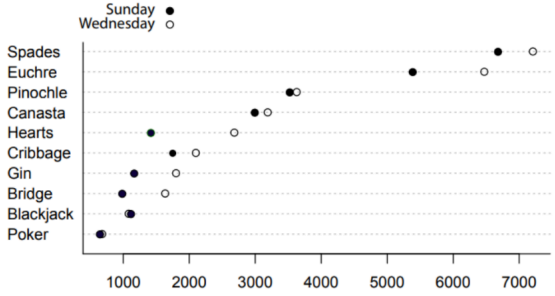
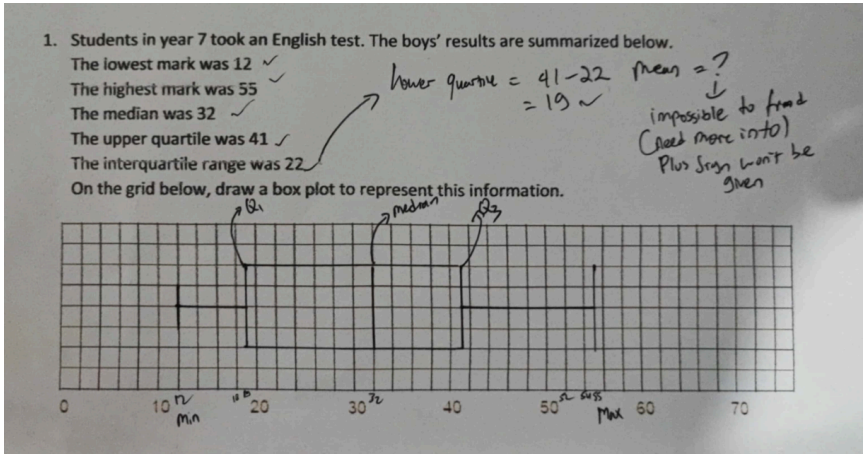
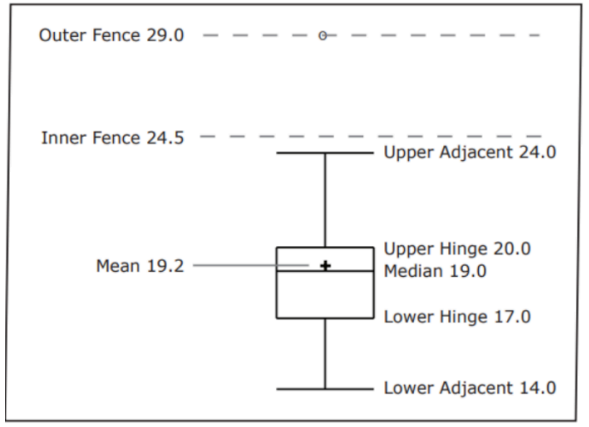
- The dot plot in Figure 2 shows the number of people playing various card games on the Yahoo website on a Wednesday.
- Unlike previous figure, the location rather than the number of dots represents the frequency.



Dot Plots

- The dot plot in this figure shows the number of people playing on a Sunday and on a Wednesday.
- This graph makes it easy to compare the popularity of the games separately for the two days but does not make it easy to compare the popularity of a given game on the two days.



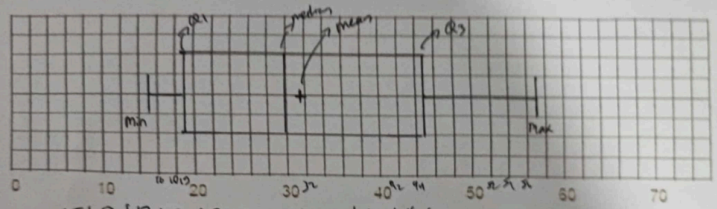
	BINUS UNIVERSITY **Dot Plots** • The dot plot in this figure makes it easy to compare the days of the week for specific games while still portraying differences among games. 	
Box Plot`		

2. 2 Students in year 7 took an English test. The girls' results are summarized in the table below.

min →	15	17	17	21	26	29
	31	35	40	46	49	57 → max

$n = 12$

Use the information in the table above to construct a box plot on the grid below



$\text{Mean} = \frac{15+17+17+21+24+29+31+35+40+46+49+57}{12} = 31,9 \approx 32$

min = 15

max = 57

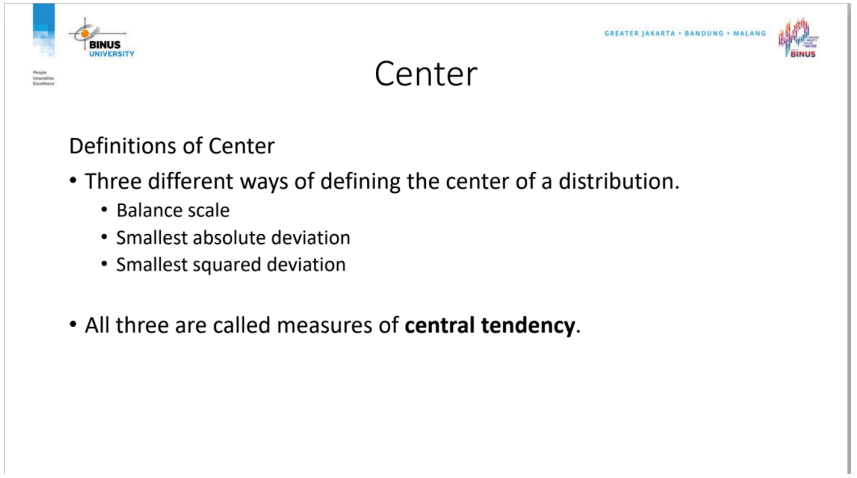
$Q_1 = \frac{1}{4}(n+1) = \frac{1}{4}(13) = 3,25$ → IR = 3, FR = 0,25 · (21 − 17) = 1 → 17 + 2 = 19

median/$Q_2 = \frac{2}{4}(n+1) = \frac{2}{4}(13) = 6,5$ → IR = 6, FR = 0,5 · (31 − 29) = 1 → 29 + 1 = 30

$Q_3 = \frac{3}{4}(n+1) = \frac{3}{4}(13) = 9,75$ → IR = 9, FR = 0,75 · (46 − 40) = 4,5 → 40 + 4,5 = 44,5

Chapter 3 - Summarizing Distributions & Bivariate Data

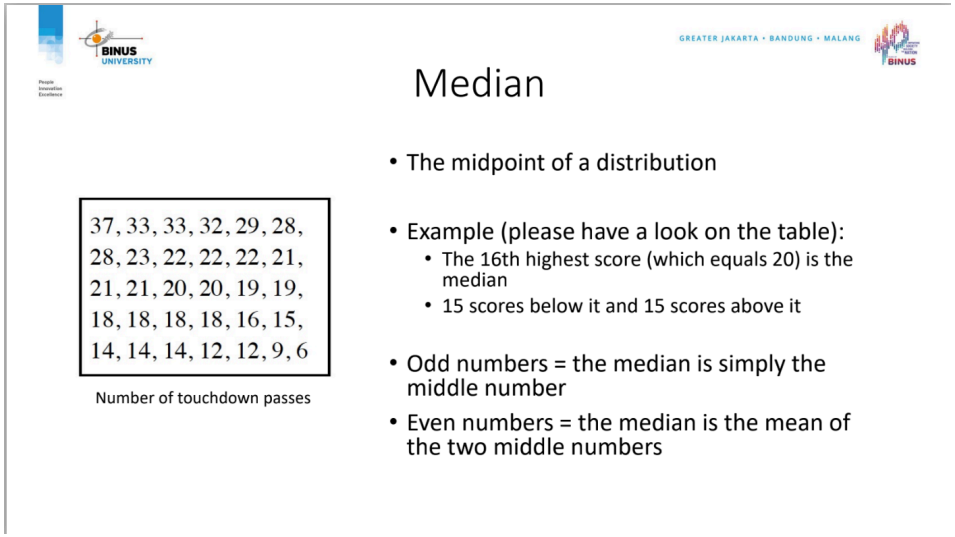
Summarizing Distributions

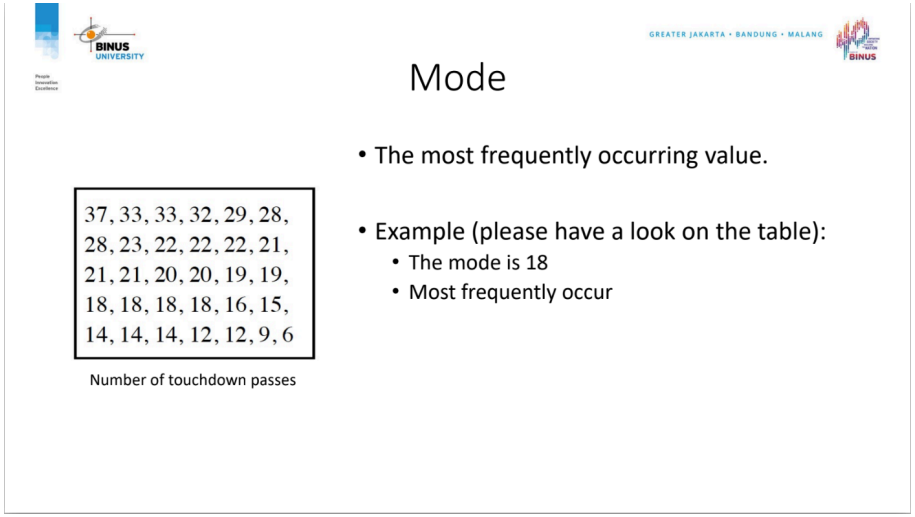


Arithmetic Mean

- The most common measure of central tendency
- Sum of the numbers divided by the number of numbers
- "μ" = the mean of a population. "M"= the mean of a sample (M and μ is essentially identical)
- ΣX = the sum of all the numbers in the population. N = the number of numbers in the population
- The formula for μ is shown below:

$$\mu = \frac{\sum X}{N}$$





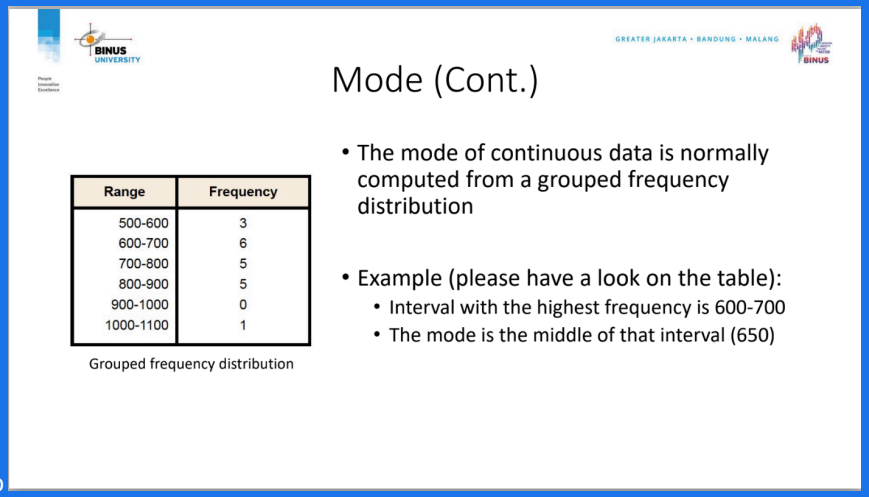
Perform Chi-Square Test Of Independence In Excel (Including P Value!)

youtube.com

8:35

In this tutorial, I will show you how to perform a chi-square test of independence by using Micros...

Replace URL



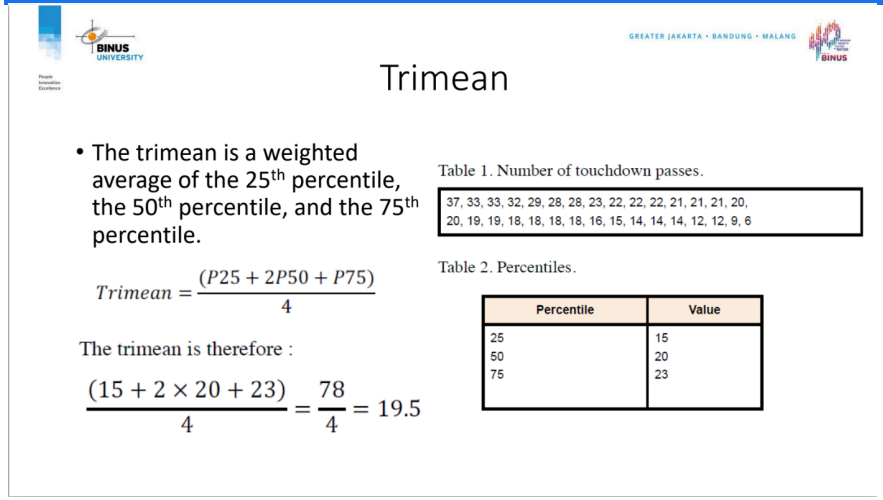
Mode (Cont.)

- The mode of continuous data is normally computed from a grouped frequency distribution
- Example (please have a look on the table):
  - Interval with the highest frequency is 600-700
  - The mode is the middle of that interval (650)

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Grouped frequency distribution

Chip



Trimean

- The trimean is a weighted average of the 25th percentile, the 50th percentile, and the 75th percentile.

$$Trimean = \frac{(P25 + 2P50 + P75)}{4}$$

The trimean is therefore :

$$\frac{(15 + 2 \times 20 + 23)}{4} = \frac{78}{4} = 19.5$$

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

Table 2. Percentiles.

Percentile	Value
25	15
50	20
75	23

Geometric Mean

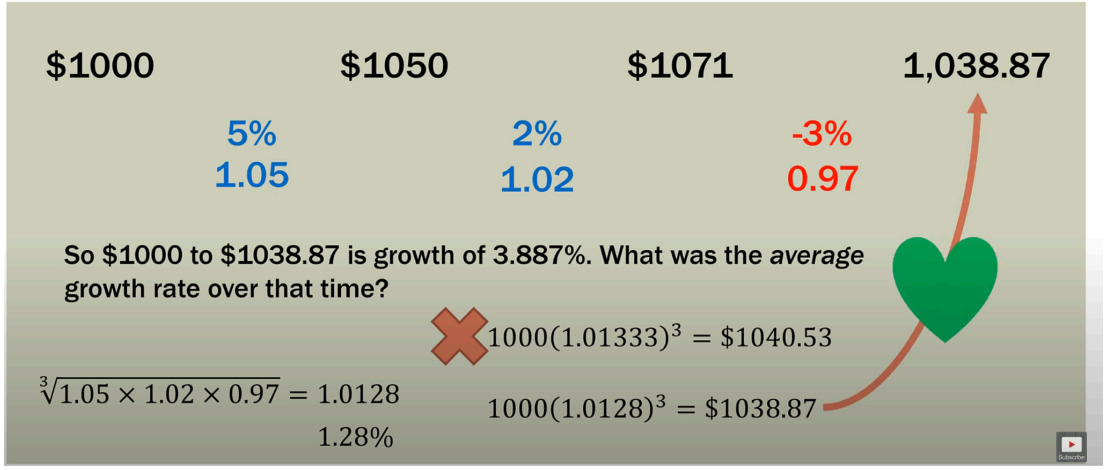
- The geometric mean is computed by multiplying all the numbers together and then taking the n^{th} root of the product.
- For example, for the numbers 1, 10, and 100, the product of all the numbers is: 1 x 10 x 100 = 1,000.
- Since there are three numbers, we take the cubed root of the product (1,000) which is equal to 10.

$$\left(\prod X\right)^{\frac{1}{N}}$$

where the symbol Π means to multiply

Note that the geometric mean only makes sense if all the numbers are positive

FINANCIAL RETURNS



Trimmed Mean

- To compute a trimmed mean, you remove some of the higher and lower scores and compute the mean of the remaining scores.
- A mean trimmed 10% is a mean computed with 10% of the scores trimmed off: 5% from the bottom and 5% from the top.
- A mean trimmed 50% is computed by trimming the upper 25% of the scores and the lower 25% of the scores and computing the mean of the remaining scores.
- The trimmed mean is like the median which, in essence, trims the upper 49+% and the lower 49+% of the scores.
- Therefore, the trimmed mean is a hybrid of the mean and the median.

Example of a Trimmed Mean

- A figure skating competition produces the following scores: 6.0, 8.1, 8.3, 9.1, and 9.9.
- The mean for the scores would equal:

```
((6.0 + 8.1 + 8.3 + 9.1 + 9.9) / 5) = 8.28
```

- To trim the mean by a total of 40%, remove the lowest 20% and the highest 20% of values, eliminating the scores of 6.0 and 9.9.
- Next, we calculate the mean based on the calculation:

```
(8.1 + 8.3 + 9.1) / 3 = 8.50
```

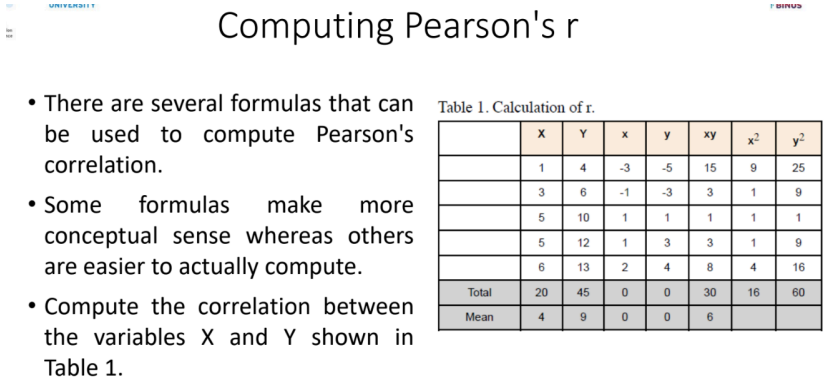
- In other words, a mean trimmed at 40% would equal 8.5 versus 8.28, which reduced the outlier bias and had the effect of increasing the reported average by 0.22 points.

Univariate data: a single variable, ex: average SAT scores, average weight

Bivariate data: two variables. Ex: relationship between height and weight of people to determine the extent to which taller people weigh more. Bivariate data could also be two sets of items that are dependent on each other, ex: ice cream sales and temperature, traffic accidents and weather

Bivariate analysis: analysis of bivariate data, simplest form of statistical analyses used to find out if there's a relationship between two sets of values. Note: multivariate analysis is the analysis of more than two variables. Bivariate analysis' results can be stored in 2-column table. Bivariate analysis is not the same as two sample data analysis, since in two sample data analysis, the two variables aren't directly related. In bivariate analysis, they are, and there is a Y value for each X.

Pearson Correlation Coefficient

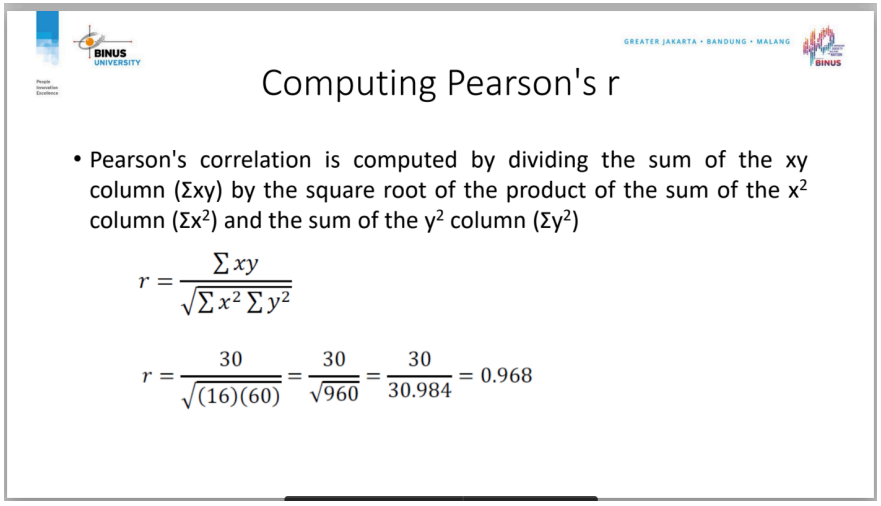


	X	Y	x	y	xy	x^2	y^2
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

Description:
x = X - Mean of X
y = Y - Mean of Y
xy = x*y

Steps to calculate Pearson correlation for two variables X and Y:

1. Compute the mean of X and subtract this mean from all values of X. This new variable is called x.
2. Do the same thing for Y to create the y variable. (x and y are called deviation scores)
3. Create a new column by multiplying x and y (xy)
4. Create two new columns, x^2 and y^2, and fill them appropriately
5. Calculate the sum of the xy column, divided by the square root of the product of the sum of x^2 column and y^2 column



Chapter 4 - Probability

Single event

Probability of a Single Event

- If you roll a six-sided die, there are six possible outcomes, and each of these outcomes is equally likely.
- What is the probability that either a one or a six will come up?
- The two outcomes about which we are concerned (a one or a six coming up) are called favorable outcomes.

$$\text{Probability} = \frac{\text{Possible outcomes}}{\text{Total outcomes}}$$

- In this case there are two favorable outcomes and six possible outcomes.
- So, the probability of throwing either a one or six is 1/3

Here is a more complex example. You throw 2 dice. What is the probability that the sum of the two dice will be 6?

Probability of a Single Event

- The probability is 5/36.
- If you know the probability of an event occurring, it is easy to compute the probability that the event does not occur.
- The probability that the total is not 6 is 1 - 5/36 = 31/36

Table 1. 36 possible outcomes.

Die 1	Die 2	**Total**	Die 1	Die 2	**Total**	Die 1	Die 2	**Total**
1	1	2	3	1	4	5	1	6
1	2	3	3	2	5	5	2	7
1	3	4	3	3	6	5	3	8
1	4	5	3	4	7	5	4	9
1	5	6	3	5	8	5	5	10
1	6	7	3	6	9	5	6	11
2	1	3	4	1	5	6	1	7
2	2	4	4	2	6	6	2	8
2	3	5	4	3	7	6	3	9
2	4	6	4	4	8	6	4	10
2	5	7	4	5	9	6	5	11
2	6	8	4	6	10	6	6	12

Two Independent Event

Probability of Two (or more) Independent Events

- Events A and B are independent events if the probability of Event B occurring is the same whether Event A occurs.

Example:

- A fair coin is tossed two times.
- The probability that a head comes up on the second toss is 1/2 regardless of whether a head came up on the first toss.
- The two events are (1) first toss is a head and (2) second toss is a head.
- So, these events are independent.

Probability of A and B

Probability of A and B

- If events A and B are independent, then the probability of both A and B occurring is:

  `P(A and B) = P(A) x P(B)`

- where `P(A and B)` is the probability of events A and B both occurring, `P(A)` is the probability of event A occurring, and `P(B)` is the probability of event B occurring.
- If you flip two coins, what is the probability that it will come up both heads?

  `1/2 x 1/2 = 1/4`

Probability of A or B

- If Events A and B are independent, the probability that either Event A or Event B occurs is:

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

- In this discussion, when we say "A or B occurs" we include three possibilities:
  1. A occurs and B does not occur
  2. B occurs and A does not occur
  3. Both A and B occur

Probability of A or B

Examples:

- If you flip a coin two times, what is the probability that you will get a head on the first flip or a head on the second flip (or both)?
- Letting Event A be a head on the first flip and Event B be a head on the second flip, then

  $P(A) = 1/2$, $P(B) = 1/2$, and $P(A \text{ and } B) = 1/4$

Therefore,

$P(A \text{ or } B) = 1/2 + 1/2 - 1/4 = 3/4$

Conditional Probabilities

- Often it is required to compute the probability of an event given that another event has occurred.

Example:

- What is the probability that two cards drawn at random from a deck of playing cards will both be aces?
- It might seem that you could use the formula for the probability of two independent events and simply multiply 4/52 x 4/52 = 1/169.
- This would be **incorrect**, because the two events are not independent.
- If the first card drawn is an ace, then the probability that the second card is also an ace would be lower because there would only be three aces left in the deck

Conditional Probabilities

- Once the first card chosen is an ace, the probability that the second card chosen is also an ace is called the conditional probability of drawing an ace.
- In this case, the "condition" is that the first card is an ace.

```
P(ace on second draw | an ace on the first draw)
```

- "The probability that an ace is drawn on the second draw given that an ace was drawn on the first draw."

Conditional Probabilities

- Since after an ace is drawn on the first draw, there are 3 aces out of 51 total cards left. This means that the probability that one of these aces will be drawn is 3/51 = 1/17.
- If Events A and B are not independent, then

  `P(A and B) = P(A) x P(B|A)`
- Applying this to the problem of two aces, the probability of drawing two aces from a deck is

  `4/52 x 3/51 = 1/221`

Birthday Problem

- If there are 25 people in a room, what is the probability that at least two of them share the same birthday.
- If your first thought is that it is 25/365 = 0.068, you will be surprised to learn it is much higher than that.
- This problem requires the application of the sections on P(A and B) and conditional probability.
- This problem is best approached by asking what is the probability that no two people have the same birthday
- If we choose two people at random, what is the probability that they do not share a birthday?

Birthday Problem

- Of the 365 days on which the second person could have a birthday, 364 of them are different from the first person's birthday. Therefore, the probability is 364/365 (P2).
- Now define P3 as the probability that the 3rd person drawn doesn't share a birthday with anyone drawn previously given that there are no previous birthday matches.
- P3 is therefore a conditional probability. Therefore P3 = 363/365.
- In like manner, P4 = 362/365, P5 = 361/365, and so on up to P25 = 341/365.
- Since P(A and B) = P(A)P(B), all we have to do is multiply P2, P3, P4 ...P25 together.
- The result is 0.431. Therefore, the probability of at least one match is 0.569.

Possible of Orders

Possible Orders

- Suppose you had a plate with three pieces of candy on it: one green, one yellow, and one red.
- Pick up these three pieces one at a time.
- In how many different orders can you pick up the pieces?

Table 1. Six Possible Orders.

Number	First	Second	Third
1	red	yellow	green
2	red	green	yellow
3	yellow	red	green
4	yellow	green	red
5	green	red	yellow
6	green	yellow	red

```
Number of orders = n!
```

UNIVERSITY

Multiplication Rule

- A small restaurant whose menu has 3 soups, 6 entrées, and 4 desserts.
- How many possible meals are there?
- The answer is calculated by multiplying the numbers to get

$$3 \times 6 \times 4 = 72$$

Permutations

- Suppose that there were four pieces of candy (red, yellow, green, and brown) and you were only going to pick up exactly two pieces.
- How many ways are there of picking up two pieces?

Table 2. Twelve Possible Orders.

Number	First	Second
1	red	yellow
2	red	green
3	red	brown
4	yellow	red
5	yellow	green
6	yellow	brown
7	green	red
8	green	yellow
9	green	brown
10	brown	red
11	brown	yellow
12	brown	green

- The general formula is:

$$ {}_{n}P_{r} = \frac{n!}{(n-r)!} $$

- where ${}_{n}P_{r}$ is the number of permutations of n things taken r at a time

$$ {}_{4}P_{2} = \frac{4!}{(4-2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 12 $$

Combinations

Combinations

- In counting combinations, choosing red and then yellow is the same as choosing yellow and then red.
- Unlike permutations, order does not count.

$$ {}_{n}C_{r} = \frac{n!}{(n-r)!\,r!} $$

For our example,

$$ {}_{4}C_{2} = \frac{4!}{(4-2)!\,2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = 6 $$

Table 3. Six Combinations.

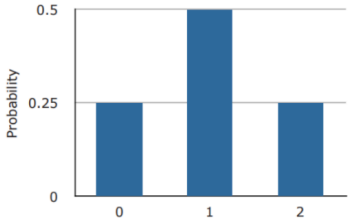
Number	First	Second
1	red	yellow
2	red	green
3	red	brown
x	yellow	red
4	yellow	green
5	yellow	brown
x	green	red
x	green	yellow
6	green	brown
x	brown	red
x	brown	yellow
x	brown	green

Binomial Distribution

- The 4 possible outcomes can be classified in terms of the number of heads that come up.
- The number could be two (Outcome 1), one (Outcomes 2 and 3) or 0 (Outcome 4).
- Since two of the outcomes represent the case in which just one head appears in the two tosses, the probability of this event is equal to 1/4 + 1/4 = 1/2.

Table 2. Probabilities of Getting 0, 1, or 2 Heads.

Number of Heads	Probability
0	1/4
1	1/2
2	1/4



The Formula for Binomial Probabilities

- The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.

$$P(x) = \frac{N!}{x!\,(N-X)!}\pi^{x}(1-\pi)^{N-x}$$

- where P(x) is the probability of x successes out of N trials, N is the number of trials, and π is the probability of success on a given trial

The Formula for Binomial Probabilities

- Applying this to the coin flip example

$$P(0) = \frac{2!}{0!\,(2-0)!}(.5^0)(1-.5)^{2-0} = \frac{2}{2}(1)(.25) = 0.25$$

$$P(1) = \frac{2!}{1!\,(2-1)!}(.5^1)(1-.5)^{2-1} = \frac{2}{1}(.5)(.5) = 0.50$$

$$P(2) = \frac{2!}{2!\,(2-2)!}(.5^2)(1-.5)^{2-2} = \frac{2}{2}(.25)(1) = 0.25$$

Cumulative Probabilities

- Toss a coin 12 times. What is the probability that we get from 0 to 3 heads?
- The answer is found by computing the probability of exactly 0 heads, exactly 1 head, exactly 2 heads, and exactly 3 heads.
- The probability of getting from 0 to 3 heads is then the sum of these probabilities.
- The probabilities are: 0.0002, 0.0029, 0.0161, and 0.0537.
- The sum of the probabilities is 0.073

Mean and Standard Deviation of Binomial Distributions

- In general, the mean of a binomial distribution with parameters N (the number of trials) and π (the probability of success on each trial) is:

$$\mu = N\pi$$

where μ is the mean of the binomial distribution.

- The variance of the binomial distribution is:

$$\sigma 2 = N\pi(1-\pi)$$

where σ2 is the variance of the binomial distribution

Mean and Standard Deviation of Binomial Distributions

- The coin was tossed 12 times, so N = 12.
- A coin has a probability of 0.5 of coming up heads.
- Therefore, $\pi = 0.5$.
- The mean and variance can therefore be computed as follows:

$$\mu = N\pi = (12)(0.5) = 6$$

$$\sigma 2 = N\pi(1-\pi) = (12)(0.5)(1.0 - 0.5) = 3.0$$

Poisson Distribution

Poisson Distribution

- Can be used to calculate the probabilities of various numbers of "successes" based on the mean number of successes.
- The term "success" doesn't really mean success in the traditional positive sense; it just means that the outcome in question occurs.

$$p = \frac{e^{-\mu}\mu^{x}}{x!}$$

e is the base of natural logarithms (2.7183)
μ is the mean number of "successes"
x is the number of "successes" in question

Poisson Distribution

Example:

- Suppose you knew that the mean number of calls to a fire station on a weekday is 8.
- What is the probability that on a given weekday there would be 11 calls?

$$p = \frac{e^{-8} 8^{11}}{11!} = 0.072$$

since the mean is 8 and the question pertains to 11 fires

Multinomial Distribution

Multinomial Distribution

Example:

- Suppose that two chess players had played numerous games, and it was determined that the **probability** that **Player A would win is 0.40**, the probability that **Player B would win is 0.35**, and the probability that the game would **end in a draw is 0.25**.
- The multinomial distribution can be used to answer questions such as: “If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?”

Multinomial Distribution

- The following formula gives the probability of obtaining a specific set of outcomes when there are k possible outcomes for each event:

$$p = \frac{n!}{(n_1!)(n_2!)\ldots(n_k!)} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

- where:
  - p is the probability
  - n is the total number of events
  - n_1 is the number of times Outcome 1 occurs,
  - n_2 is the number of times Outcome 2 occurs,
  - n_k is the number of times Outcome k occurs,
  - p_1 is the probability of Outcome 1
  - p_2 is the probability of Outcome 2, and
  - p_k is the probability of Outcome k

Multinomial Distribution

- For the chess example,
  - n = 12 (12 games are played),
  - n_1 = 7 (number won by Player A),
  - n_2 = 2 (number won by Player B),
  - n_3 = 3 (the number drawn),
  - p_1 = 0.40 (probability Player A wins)
  - p_2 = 0.35(probability Player B wins)
  - p_3 = 0.25(probability of a draw)

$$p = \frac{12!}{(7!)(2!)(3!)} .40^{.7} .35^{.2} .25^{3} = 0.0248$$

Hypergeometric Distribution

- The hypergeometric distribution is used to calculate probabilities when sampling without replacement.
- Example, suppose you first randomly sample 1 card from a deck of 52.
- Then, without putting the card back in the deck you sample a second and then a third.
- Given this sampling procedure, what is the probability that exactly two of the sampled cards will be aces (4 of the 52 cards in the deck are aces).

Hypergeometric Distribution

- Formula based on the hypergeometric distribution

$$p = \frac{{}_{k}C_{x} \;\; {}_{(N-k)}C_{(n-x)}}{{}_{N}C_{n}}$$

- Where:
  - k is the number of “successes” in the population
  - x is the number of “successes” in the sample
  - N is the size of the population
  - n is the number sampled
  - p is the probability of obtaining exactly x successes
  - ${}_{k}C_{x}$ is the number of combinations of k things taken x at a time

- In this example, k = 4 (four aces in the deck), x = 2 (probability of getting two aces), N = 52 (52 cards in a deck), and n = 3 because 3 cards were sampled.
- Therefore,

$$p = \frac{{}_4C_{2\,(52-4)}C_{(3-2)}}{{}_{52}C_3}$$

$$p = \frac{\frac{4!}{2!2!}\frac{48!}{47!1!}}{\frac{52!}{49!3!}} = 0.013$$

Hypergeometric Distribution

- The mean and standard deviation of the hypergeometric distribution are:

$$\text{mean} = \frac{(n)(k)}{N}$$

$$\text{sd} = \sqrt{\frac{(n)(k)(N-k)(N-n)}{N^2(N-1)}}$$

Bayes' Theorem

- This same result can be obtained using Bayes' theorem.
- Bayes' theorem considers both the prior probability of an event and the diagnostic value of a test to determine the posterior probability of the event.

For the current example

- The event is that you have Disease X. Let's call this Event D.
- Since only 2% of people in your situation have Disease X, the prior probability of Event D is 0.02 or, more formally, P(D) = 0.02.
- If D' represents the probability that Event D is false, then

```
P(D') = 1 - P(D) = 0.98
```

Bayes' Theorem

- To define the diagnostic value of the test, we need to define another event: that you test positive for Disease X → Event T.
- The diagnostic value of the test depends on the probability you will test positive given that you have the disease, written as P(T|D), and the probability you test positive given that you do not have the disease, written as P(T|D')

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')}$$

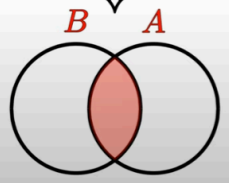
```
P(T|D)  = 0.99
P(T|D') = 0.09
P(D)    = 0.02
P(D')   = 0.98
```

Bayes' Theorem

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

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

$P(B \cap A)$



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

$$P(B|A) \cdot P(A) = P(A \cap B)$$

$$P(A \cap B) = P(B \cap A)$$

Ex You've been planning a picnic for your family. You're trying to decide whether to postpone due to rain. The chance of rain on any day is 15%. The morning of the picnic, it's cloudy. The prob. of it being cloudy is 25% and on days where it rains, it's cloudy in the morning 80% of the time. Should you postpone the picnic?

$P(\text{rain}) = 0.15$

$P(\text{cloudy}) = 0.25$

$P(\text{cloudy}|\text{rain}) = 0.80$

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

$$P(\text{rain}|\text{cloudy}) = \frac{P(\text{cloudy}|\text{rain}) \cdot P(\text{rain})}{P(\text{cloudy})}$$

$$= \frac{0.8 \cdot 0.15}{0.25}$$

$$\boxed{P(\text{rain}|\text{cloudy}) = 0.48}$$

$$P(D|T) = \frac{(0.99)(0.02)}{(0.99)(0.02) + (0.09)(0.98)} = 0.1833$$

Chapter 5 - Research Design

Any complete set of observations can be categorized as a population. Any subset of observations from the population can be categorized as a sample. The sample size is usually relatively small relative to the population size.

Validity: the extent that results represent reality

Face validity: whether the test appears to measure what it's supposed to measure. Most informal and subjective.

Content validity: How well a survey measures the construct it sets out to measure, and how it does not measure unrelated constructs

Construct validity: Can be established by showing a test has both convergent and divergent (discriminant) validity. Convergent validity means it behaves similarly to other similar tests; discriminant validity means it behaves differently when compared to different tests.

Internal validity / cause-and-effect validity: How well the independent variables in a study were identified and controlled for, results have to be due to manipulation of independent variables and nothing else.

External validity: how well the study applies to the real world.

Statistical conclusion validity: the degree to which conclusions about the relationship among variables based on the data are correct or reasonable.

Criterion-related validity: a method of testing the correlation of a variable to a concrete outcome. Two types: predictive validity, models the likelihood of an outcome, concurrent validity, confirms whether one measure is equal or better than another accepted measure when testing the same thing at the same time.

Sampling bias: refers to the method of sampling, not the sample itself.

Self-selection bias: happens when the non-random component occurs after the potential subject has enlisted in the experiment

Undercoverage bias: sample too few observations, happens when a significant entity goes unselected

Survivorship bias: observations recorded at the end are a non-random set of those present at the beginning of the investigation.

Experimental designs:

Between-subjects design: each subject is assigned one treatment condition, results based on group differences between participants in various conditions. Requires larger samples.

Within-subject design: All participants exposed to every treatment condition. Less samples required, but may result in fatigue.

Complex design: Combination of the previous two

Chapter 6 - Normal Distribution & Estimation (Z-Test)

Variance

- The average of the squared differences from the Mean.
- To calculate the variance, follow these steps:
  1. Work out the Mean
  2. Then for each number: subtract the Mean and square the result
  3. Then work out the average of those squared differences.

Variance - Example

- You and your friends have just measured the heights of your dogs (in millimeters):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

- Find out:
  - the Mean
  - the Variance
  - the Standard Deviation

Variance - Example

- To calculate the Variance, take each difference, square it, and then average the result.

$$\text{Variance} = \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$\sigma^2 = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} = 21704$$

- And the Standard Deviation is just the square root of Variance, so:

$$\sigma = \sqrt{21704}$$
$$= 147.32277$$
$$= 147 \text{ (to the nearest mm)}$$

Standard Deviation

- That example is for a Population (the 5 dogs are the only dogs we are interested in).
- But if the data is a Sample (a selection taken from a bigger Population), then the calculation **changes**!

- The Population: divide by N when calculating Variance
- A Sample: divide by N-1 when calculating Variance

Z-Test (One Sample)

Standard Normal Distribution

- A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution**.
- A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

$$Z = (X - \mu)/\sigma$$

- As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26?
- Applying the formula:

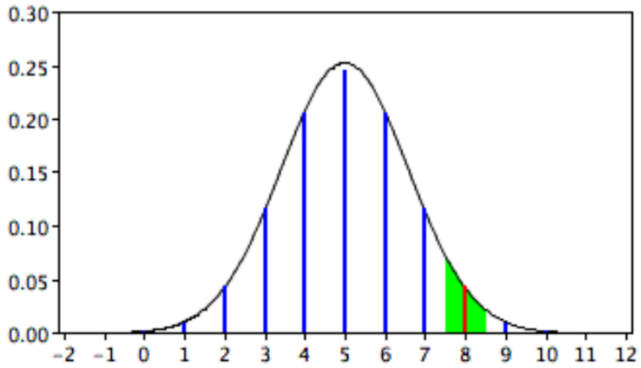
$$Z = (26 - 50)/10 = -2.4$$

Normal Approximation to the Binomial

- Assume you have a fair coin and wish to know the **probability** that you would get **8 heads out of 10 flips**.
- The binomial distribution has a mean of $\mu = N\pi = (10)(0.5) = 5$ and a variance of $\sigma^2 = N\pi(1-\pi) = (10)(0.5)(0.5) = 2.5$
- The standard deviation is therefore $\sigma = 1.5811$
- **Binomial distribution** is a **discrete probability distribution**, whereas the **normal distribution** is a **continuous distribution**.
- The solution is to consider any value from 7.5 to 8.5 to represent an outcome of 8 heads.

Normal Approximation to the Binomial

- Using this approach, figure out the area under a normal curve from 7.5 to 8.5.



Normal Approximation to the Binomial

- You could find the solution using a table of the standard normal distribution (a Z table) as follows:
  1. Find a Z score for 8.5 using the formula Z = (8.5 - 5)/1.5811 = 2.21.
  2. Find the area below a Z of 2.21 = 0.98645.
  3. Find a Z score for 7.5 using the formula Z = (7.5 - 5)/1.5811 = 1.58.
  4. Find the area below a Z of 1.58 = 0.94295.
  5. Subtract the value in step 4 from the value in step 2 to get 0.0435

Calculating the Confidence Interval

- **Step 1**: start with
  - the number of observations n
  - the mean X
  - and the standard deviation s

- Using our example:
  - number of observations n = 40
  - mean X = 175
  - standard deviation s = 20

Calculating the Confidence Interval

- **Step 2:** decide what Confidence Interval we want: 95% or 99% are common choices.
- Then find the "Z" value for that Confidence Interval
- For 95% the Z value is 1.960

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Calculating the Confidence Interval

- **Step 3**: use that Z value in this formula for the Confidence Interval

$$\bar{X} \pm Z\frac{s}{\sqrt{n}}$$

- Where:
  - X is the mean
  - Z is the chosen Z-value from the table above
  - s is the standard deviation
  - n is the number of observations
- And we have: $175 \pm 1.960 \times \frac{20}{\sqrt{40}}$ → 175cm ± 6.20 cm
- In other words: from 168.8cm to 181.2 cm

Discrete Distributions

- A discrete distribution has a range of values that are countable

Example

- 3 pool balls, each with a number on it. Suppose 2 of the balls are selected randomly (with replacement) and the average of their numbers is computed.

Outcome	Ball 1	Ball 2	Mean
1	1	1	1
2	1	2	1.5
3	1	3	2
4	2	1	1.5
5	2	2	2
6	2	3	2.5
7	3	1	2
8	3	2	2.5
9	3	3	3

Discrete Distributions

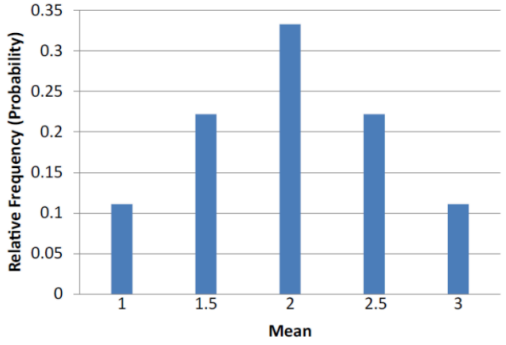
- Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0
- The frequencies of these means are shown in the next table.
- Frequencies of means for N = 2.
- The relative frequencies are equal to the frequencies divided by 9 because there are 9 possible outcomes.

Mean	Frequency	Relative Frequency
1	1	0.111
1.5	2	0.222
2	3	0.333
2.5	2	0.222
3	1	0.111

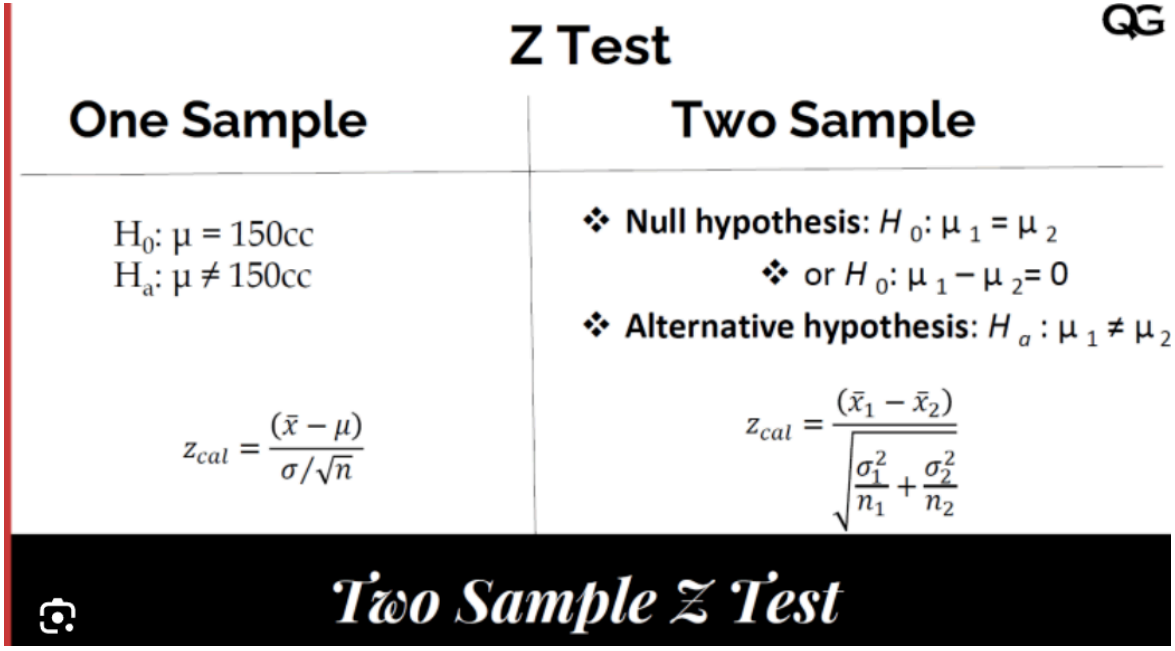
,

Discrete Distributions

- Distribution of means for N = 2.
- The distribution shown is called the sampling distribution of the mean.
- For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions.
- The pool balls have only the values 1, 2, and 3, and a sample mean can have 1 of only 5 values.



Z-Test (Two Samples)



Contoh Soal
https://courses.washington.edu/psy315/tutorials/z_test_tutorial.pdf

Chapter 7 - Hypothesis Testing & T-Test

The Null Hypothesis

- The null hypothesis is a characteristic arithmetic theory suggesting that no statistical relationship and significance exists in a set of given, single, observed variables between two sets of observed data and measured phenomena.
- The hypotheses play an important role in testing the significance of differences in experiments and between observations.
- H_0 symbolizes the null hypothesis of no difference.
- It presumes to be true until evidence indicates otherwise.

Examples of a Null Hypothesis

- A school principal claims that students in her school score an average of seven out of 10 in exams.
- The null hypothesis is that the population mean is 7.0.
- To test this null hypothesis, we record marks of 30 students (sample) from the entire student population of the school (300) and calculate the mean of that sample.
- We can then compare the (calculated) sample mean to the (hypothesized) population mean of 7.0 and attempt to reject the null hypothesis.

Significance Testing

- A low probability value casts doubt on the null hypothesis.
- So, how low must the probability value be in order to conclude that the null hypothesis is false?
- Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05 (< 0.05).
- More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01 (< 0.01).

Significance Testing

- When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis.
- The probability value below which the null hypothesis is rejected is called the α level or simply **α** (alpha).
- It is also called the significance level.

Steps in Hypothesis Testing

1. Specify the null hypothesis.
2. Specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.
3. Compute the probability value (also known as the p value).
4. Compare the probability value with the α level.

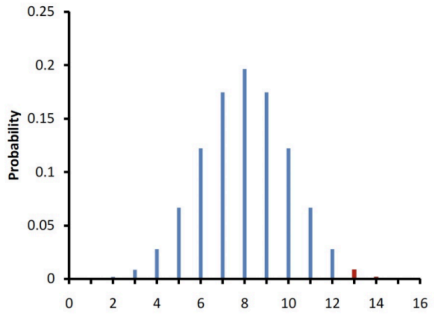
Error in Statistical Decision-Making

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

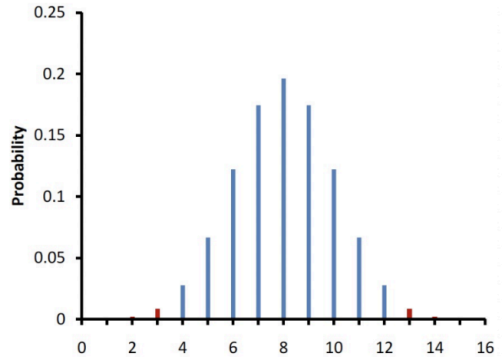
Scribbr

One-Tailed Test



- From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106
- The red bars show the values greater than or equal to 13.
- The probabilities are calculated for the upper tail of the distribution
- A probability calculated in only one tail of the distribution is called a “one-tailed probability.”

Two-Tailed Test



- "What is the probability of getting a result as extreme or more extreme than the one observed"?
- Since the chance expectation is 8/16, a result of 3/13 is equally as extreme as 13/16.
- Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously.
- Therefore, $p = 0.0212$
- A probability calculated in both tails of a distribution is called a two-tailed probability

T-Test One Sample

Example

- A company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction.
- After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15.
- Did the training work? Test your hypothesis at a 5% alpha level.

- **Step 1**: Write your null hypothesis statement.
- The accepted hypothesis is that there is no difference in sales, so:

  H0: $\mu = \$100$

- **Step 2**: Write your alternate hypothesis.
- This is the one you're testing in the one sample t test.
- You think that there is a difference (that the mean sales increased), so:

  H1: $\mu > \$100$

- **Step 3**: Identify the following pieces of information you'll need to calculate the test statistic.
- The question should give you these items:
  - The sample mean($\bar{x}$) → This is given in the question as $130.
  - The population mean(μ) → Given as $100 (from past data).
  - The sample standard deviation(s) = $15.
  - Number of observations(n) = 25.

- **Step 4**: Insert the items from above into the t score formula.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

t = (130 – 100) / ((15 / √(25))
t = (30 / 3) = 10
- This is your calculated t-value.

- **Step 5:** Find the t-table value. You need two values to find this:
  1. The alpha level: given as 5% in the question.
  2. The degrees of freedom, which is the number of items in the **sample** (n) minus 1: 25 – 1 = 24.

Critical values of *t* for one-tailed tests

Significance level (α)

Degrees of freedom (*df*)	.2	.15	.1	.05	.025	.01	.005	.001
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450

- **Step 6**: Compare Step 4 to Step 5.
- The value from Step 4 does not fall into the range calculated in Step 5, so we can reject the null hypothesis.
- The value of 10 falls into the rejection region (the left tail).
- In other words, it's highly likely that the mean sale is greater.
- The one sample t test has told us that sales training was probably a success.

https://courses.washington.edu/psy315/tutorials/t_test_tutorial.pdf

T-Test Two Independent Sample

Pooled Variance

https://courses.washington.edu/psy315/tutorials/t_test_2_independent_means_tutorial.pdf

Example

- For each group, we need the average, standard deviation and sample size.

Table 2: Average, standard deviation and sample size statistics grouped by gender

Group	Sample Size (n)	Average (X-bar)	Standard deviation (s)
Women	10	22.29	5.32
Men	13	14.95	6.84

- We start by calculating our test statistic.
- This calculation begins with finding the difference between the two averages:

  22.29 − 14.95 = 7.34

- This difference in our samples estimates the difference between the population means for the two groups.
- Next, calculate the pooled standard deviation.

Example

- This builds a combined estimate of the overall standard deviation.
- First, we calculate the pooled variance:

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

$$s_p^2 = \frac{((10-1)5.32^2)+((13-1)6.84^2)}{(10+13-2)}$$

$$= \frac{(9\times 28.30)+(12\times 46.82)}{21}$$

$$= \frac{(254.7+561.85)}{21}$$

$$= \frac{816.55}{21} = 38.88$$

$$\sqrt{38.88} = 6.24$$

Example

- Formula: $$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{\text{difference of group averages}}{\text{standard error of difference}} = \frac{7.34}{(6.24 \times \sqrt{(1/10+1/13)})} = \frac{7.34}{2.62} = 2.80$$

- The population mean($\mu_1 - \mu_2$) = 0

Example

- The t value with α = 0.05 and 21 degrees of freedom is 2.080.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	[illegible]	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074

Compare the value of our statistic (2.80) to the *t* value.
Since 2.80 > 2.080, we reject the null hypothesis that the mean body fat for men and women are equal and conclude that we have evidence body fat in the population is different between men and women.

Different Variance

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

T-Test Two Related Samples

Paired t Tests

$H_0: \mu_{before} = \mu_{after}$
$H_a: \mu_{before} \neq \mu_{after}$

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

Patient	Before	After	difference
1	120	122	-2
2	122	120	2
3	143	141	2
4	100	109	-9
5	109	109	0

- Example: Before and after medicine BP was measured. Is there a **difference** at 95% confidence level?
- $\bar{d}$ = -1.4 , s = 4.56 , n = 5
- $t_{cal.}$ = 1.4/2.04 = - 0.69

Paired t Test

Example:

Table 1: Exam scores for each student

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	6
Nina	65	65	0
Tim	56	62	6
Kate	100	91	-9
Alonzo	88	78	-10
Jose	83	87	4
Nikhil	77	79	2
Julia	92	88	-4
Tohru	90	85	-5
Michael	84	92	8
Jean	68	69	1
Indra	74	81	7
Susan	87	84	-3
Allen	64	75	11
Paul	71	84	13
Edwina	88	82	-6

- The average score difference is:

$$\overline{x_d} = 1.31$$

- Next, we calculate the standard error for the score difference. The calculation is:

$$\text{Standard Error} = \frac{s_d}{\sqrt{n}} = \frac{7.00}{\sqrt{16}} = \frac{7.00}{4} = 1.75$$

- In the formula above, *n* is the number of students – which is the number of differences.

- Calculate our test statistic as:

$$t = \frac{\text{Average difference}}{\text{Standard Error}} = \frac{1.31}{1.75} = 0.750$$

Comparing T vs T's T-Table

- To find this value, we need the significance level (α = 0.05) and the degrees of freedom.
- The degrees of freedom (df) are based on the population size.
- For the exam score data, this is:
  df = n = 16

Example

- The t value with α = 0.05 and 16 degrees of freedom is 2.120.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	[illegible]	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074

We compare the value of our statistic (0.750) to the t value. Because 0.750 < 2.120, we can reject our idea that the mean score difference is zero.

Chapter 8 - Regression

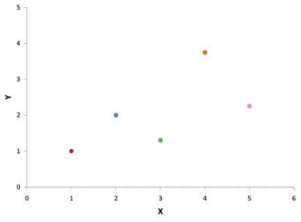
Computing Regression Lines

Linear Regression

- The example data in table are plotted in the figure.
- You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

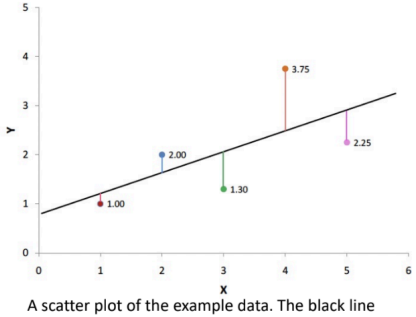
Example data



A scatter plot of the example data

Linear Regression

- Linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a regression line.
- The black diagonal line in the figure is the regression line and consists of the predicted score on Y for each possible value of X.
- The vertical lines from the points to the regression line represent the errors of prediction.
- As you can see, the red point is very near the regression line; its error of prediction is small.
- By contrast, the orange point is much higher than the regression line and therefore its error of prediction is large.



A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

Computing the Regression Line

- The calculations are based on the statistics shown in Table below.
- M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the standard deviation of Y, and r is the correlation between X and Y.

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

- r is the correlation between X and Y

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

	X	Y	x	y	xy	x^2	y^2
	1	1	-2	-1.06	2.12	4	1.1236
	2	2	-1	-0.06	0.06	1	0.0036
	3	1.3	0	-0.76	0	0	0.5776
	4	3.75	1	1.69	1.69	1	2.8561
	5	2.25	2	0.19	0.38	4	0.0361
∑	15	10.3	0.00	0.00	4.25	10.00	4.60
Mean	3	2.06					

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

- The slope (b) can be calculated as follows:

$$b = r\frac{s_y}{s_x}$$

$$b = (0.627)\frac{1.072}{1.581} = 0.425$$

- and the intercept (A) can be calculated as

  $A = M_Y - b\,M_X$

  $A = 2.06 - (0.425)(3) = 0.785$

- The formula for a regression line is

  $Y' = bX + A$

- where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.
- The equation for the line in previous Figure is

  $Y' = 0.425X + 0.785$

- For X = 1,

  $Y' = (0.425)(1) + 0.785 = 1.21.$

- For X = 2,

  $Y' = (0.425)(2) + 0.785 = 1.64$

Partitioning Sum Of Squares

- One aspect of regression: it can divide the variation in Y into 2 parts: the variation of the predicted scores and the variation in the errors of prediction.
- The variation of Y is called the **sum of squares Y**.
- In the population, the formula is

$$SSY = \sum (Y - \mu_Y)^2$$

- When computed in a sample, use the sample mean, M:

$$SSY = \sum (Y - M_Y)^2$$

How the SSY is partitioned

- The data in this Table are reproduced from the introductory section.
- The column X has the values of the predictor variable, and the column Y has the criterion variable.

X	Y	y	y^2	Y'	y'	y'^2	Y-Y'	$(Y-Y')^2$
1	1	-1.06	1.1236	1.21	-0.85	0.7225	-0.21	0.044
2	2	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3	1.3	-0.76	0.5776	2.06	0	0	-0.76	0.578
4	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.6
5	2.25	0.19	0.0361	2.91	0.85	0.7225	-0.66	0.436
15	10.3	0	4.597	10.3	0	1.806	0	2.791

How the SSY is partitioned

- The 3rd column, y, the differences between the column Y and the mean of Y.
- The 4th column, y^2, is simply the square of the y column.
- The column Y' contains the predicted values of Y.

  $Y' = 0.425X + 0.785$
- The column y' contains deviations of Y' from the mean of Y' and y'^2 is the square of this column.
- The next to last column, Y-Y', contains the actual scores (Y) minus the predicted scores (Y').
- The last column contains the squares of these errors of prediction.

How the SSY is partitioned

- Recall that SSY is the sum of the squared deviations from the mean, therefore the sum of the y^2 column (4.597).
- SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE).
- SSY' is the sum of the squared deviations of the predicted scores from the mean predicted score → the sum of y'^2 column (1.806).
- SSE is the sum of the squared errors of prediction, therefore, the sum of the $(Y-Y')^2$ column (2.791).
- This can be summed up as:

  $SSY = SSY' + SSE$

  $4.597 = 1.806 + 2.791$

How the SSY is partitioned

- The SSY is the total variation, the SSY' is the variation explained, and the SSE is the variation unexplained.
- Therefore, the proportion of variation explained can be computed as:

$$\text{Proportion explained} = \frac{SSY'}{SSY}$$

- Similarly, the proportion not explained is:

$$\text{Proportion not explained} = \frac{SSE}{SSY}$$

Pearson's Correlation

- There is an important relationship between the proportion of variation explained and Pearson's correlation: **r^2 is the proportion of variation explained**.
- Therefore, if r = 1, then, naturally, the proportion of variation explained is 1; if r = 0, then the proportion explained is 0.
- Other example: for r = 0.4, the proportion of variation explained is 0.16

Standard Error of the Estimate

The Standard Error

- The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

- where σ_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores.

Example

	X	Y	Y'	Y-Y'	$(Y-Y')^2$
	1	1	1.21	-0.21	0.044
	2	2	1.635	0.365	0.133
	3	1.3	2.06	-0.76	0.578
	4	3.75	2.485	1.265	1.6
	5	2.25	2.91	-0.66	0.436
Sum	15	10.3	10.3	0	2.791

- The standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747$$

- There is a version of the formula for the standard error in terms of Pearson's correlation:

$$\sigma_{est} = \sqrt{\frac{(1-\rho^2)SSY}{N}}$$

- where ρ is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2$$

- For the data in Table 1, my = 10.30, SSY = 4.597 and r = 0.6268. Therefore,

$$\sigma_{est} = \sqrt{\frac{(1-0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747$$

Significance Test for the Slope (b)

- Recall the general formula for a t test:

$$t = \frac{statistics - hypothesized\ value}{estimated\ standard\ error\ of\ the\ statistic}$$

- The degrees of freedom for this test are:

df = N-2

- where N is the number of pairs of scores.

UNIVERSITY

Significance Test for the Slope (b)

- The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}}$$

- where s_b is the estimated standard error of b, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X.

Significance Test for the Slope (b)

- SSX is calculated as

$$SSX = \sum (X - M_x)^2$$

- where Mx is the mean of X.
- The standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

Example

	X	Y	x	x^2	y	y^2
	1	1	-2	4	-1.06	1.1236
	2	2	-1	1	-0.06	0.0036
	3	1.3	0	0	-0.76	0.5776
	4	3.75	1	1	1.69	2.8561
	5	2.25	2	4	0.19	0.0361
Sum	15	10.3	0	10	0	4.597

- The computation of the standard error of the estimate (s_{est}) for these data is shown in the section on the standard error of the estimate.

  s_{est} = 0.964
- SSX is the sum of squared deviations from the mean of X.
- It is, therefore, equal to the sum of the x^2 column.

  SSX = 10.00
- The standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

- As shown previously, the slope (b) is 0.425.
- Therefore:

$$t = \frac{0.425}{0.305} = 1.39$$

df = N-2 = 5-2 = 3.

- The p value for a two-tailed t test is 0.26 → the slope is not significantly different from 0

Chapter 9 - ANOVA

One way ANOVA

https://www.youtube.com/watch?v=q48uKU_KWas&ab_channel=ArmstrongPSYC2190

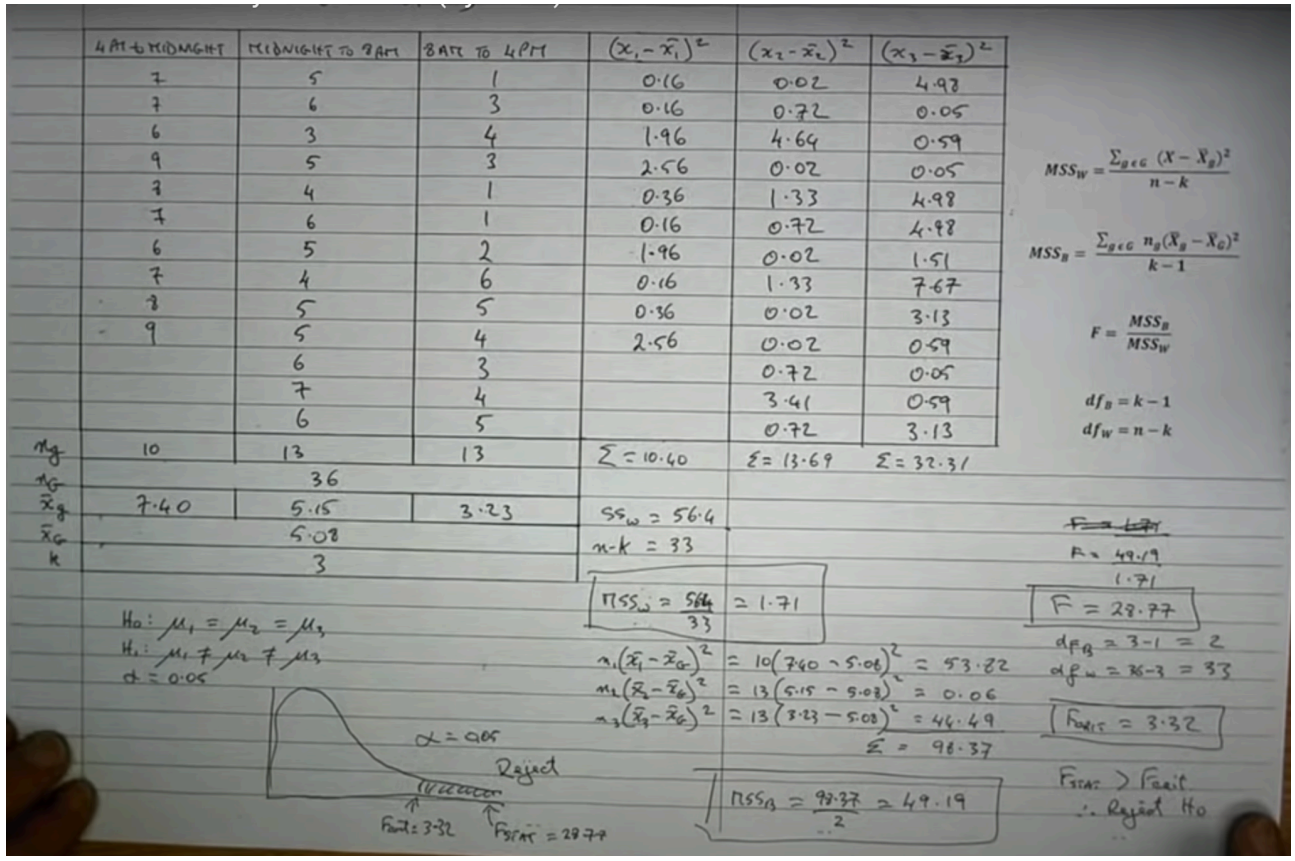
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$	$df_w = k-1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$df_b = n-k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$	$df_t = n-1$		

https://www.people.vcu.edu/~wsstreet/courses/314_20033/Examples.ANOVA.pdf

Two ways ANOVA

https://www.youtube.com/watch?v=0K-bfzLTRiY&ab_channel=VectorsAcademy

https://www.youtube.com/watch?v=KS5kPf-CMuA&ab_channel=Time2Study



Chapter 10 - Chi-Square

https://www.youtube.com/watch?v=NDhmMH25AC4&ab_channel=StevenBradburn

https://www.youtube.com/watch?v=HKDqlYSLt68&t=235s&pp=ygULQ2hpIFNxdWFyZSA%3D

https://www.bisd303.org/cms/lib3/wa01001636/centricity/domain/587/chi-squarepractice.pdf