

STAT6171001

Basic Statistics

Summarizing Distributions & Bivariate Data
Session 3

Raymond Bahana
rbahana@binus.ac.id



Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions



People
Innovation
Excellence



GREATER JAKARTA • BANDUNG • MALANG



Topics

- Summarizing Distributions
- Bivariate Data



People
Innovation
Excellence



GREATER JAKARTA • BANDUNG • MALANG



Summarizing Distributions



Introduction

- Descriptive statistics often involves using a few numbers to summarize a distribution.
- One important aspect of a distribution is where its center is located
- A second aspect of a distribution is how spread out it is
- Distributions
 - Symmetric
 - Long tails in just one direction



Central Tendency - Illustration

Imagine this situation:

- Class = 5 students (you and four other students)
- You took the quiz (5-point pop quiz)
- Your score 3/5

How do you react?

Are you happy or disappointed?

How do you decide?

What additional information would you like?



Central Tendency – Illustration (Cont.)

What you will do?

- You will ask your neighbors/ask the instructor

The **additional information** you want is how your quiz score compares to other students' scores

The importance of comparing your score to the class **distribution** of scores



Comparing

- Comparing individual scores to a distribution of scores is fundamental to statistics

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Three possible datasets for the 5-point make-up quiz

Comparing

- Comparing individual scores to a distribution of scores is fundamental to statistics



Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Three possible datasets for the 5-point make-up quiz

Comparing

- Comparing individual scores to a distribution of scores is fundamental to statistics

Your score **below** the *center of the distribution*




Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Three possible datasets for the 5-point make-up quiz

Comparing

- Comparing individual scores to a distribution of scores is fundamental to statistics

Your score **above** the *center of the distribution*



Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Three possible datasets for the 5-point make-up quiz



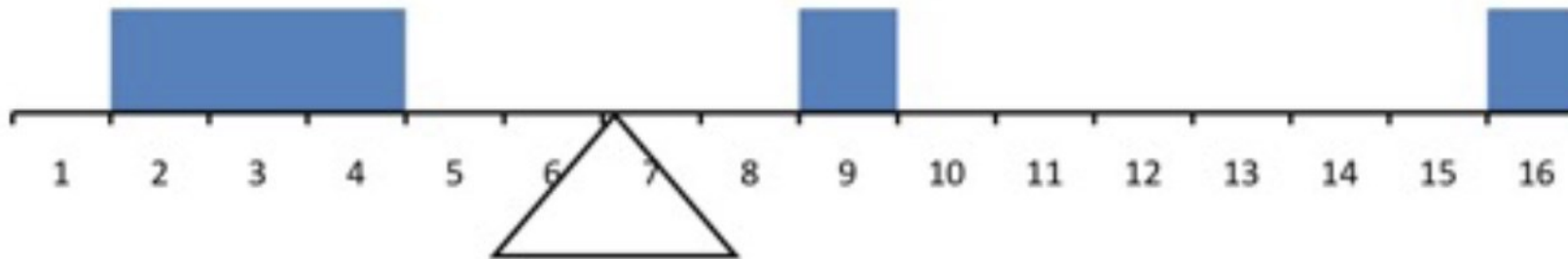
Center

Definitions of Center

- Three different ways of defining the center of a distribution.
 - Balance scale
 - Smallest absolute deviation
 - Smallest squared deviation
- All three are called measures of **central tendency**.

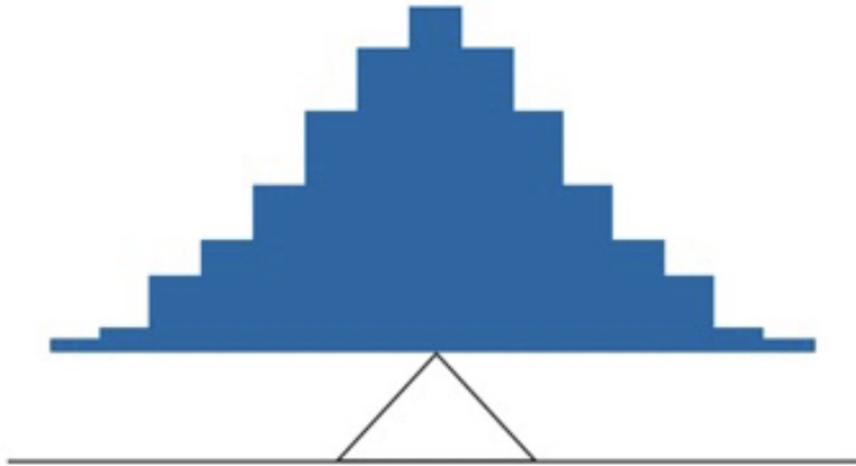
Balance Scale

- One definition of central tendency is the point at which the distribution is in balance

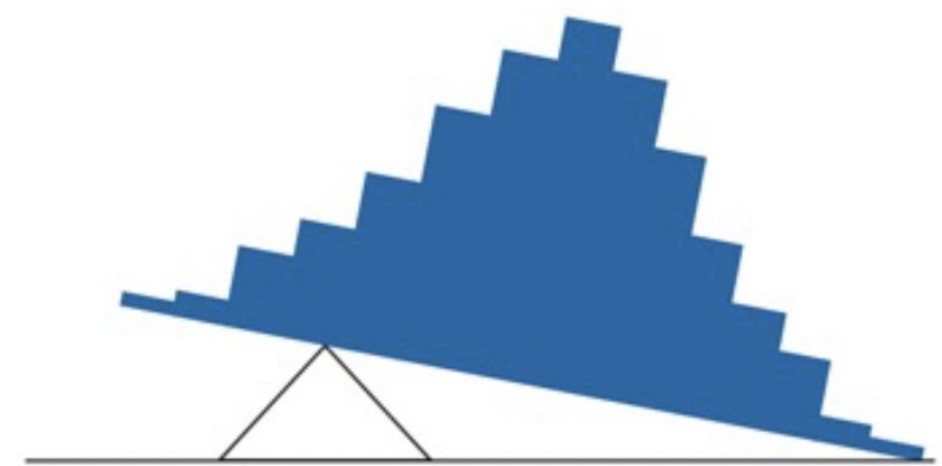


A balance scale

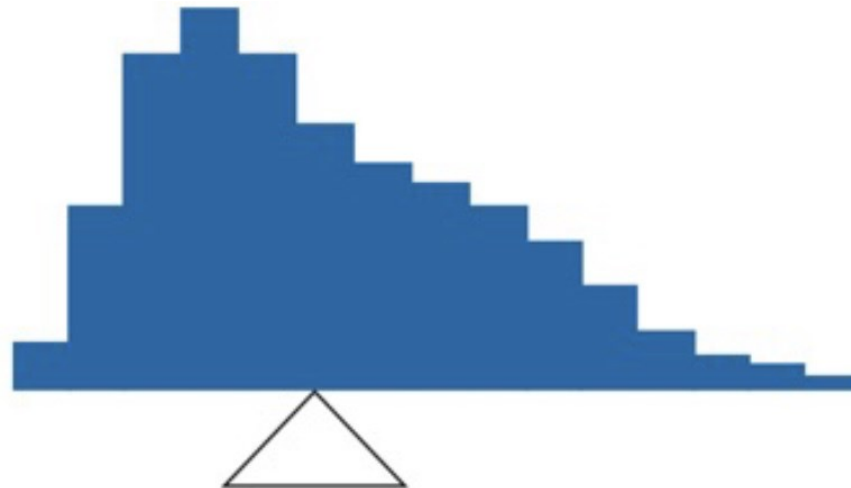
Balance Scale (Cont.)



A distribution balanced on the tip of a triangle



The distribution is not balanced



An asymmetric distribution balanced on the tip of a triangle

The balance point defines one sense of a distribution's center.



Smallest Absolute Deviation

- Smallest Absolute Deviation is another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences).
- Consider the distribution made up of the five numbers 2, 3, 4, 9, 16.
- Let's see how far the distribution is from 10 (picking a number arbitrarily)

Smallest Absolute Deviation

- Based on the concept of the sum of the absolute deviations (differences).
- So, the sum of the absolute deviations from 10 is 28.
- Likewise, the sum of the absolute deviations from 5 equals $3 + 2 + 1 + 4 + 11 = 21$
- So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

An example of the sum of absolute deviations



Smallest Absolute Deviation

- The center of a distribution is the number for which the sum of the absolute deviations is smallest.
- As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21.
- Is there a value for which the sum of the absolute deviations is even smaller than 21? **Yes.**
- For these data, there is a value for which the sum of absolute deviations is only 20.
- See if you can find it.



Smallest Squared Deviation

- Based on the concept of the sum of squared deviations (differences)
- The target that minimizes the sum of squared deviations provides another useful definition of central tendency
- Challenging to find the value that minimizes this sum

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

An example of the sum of squared deviations

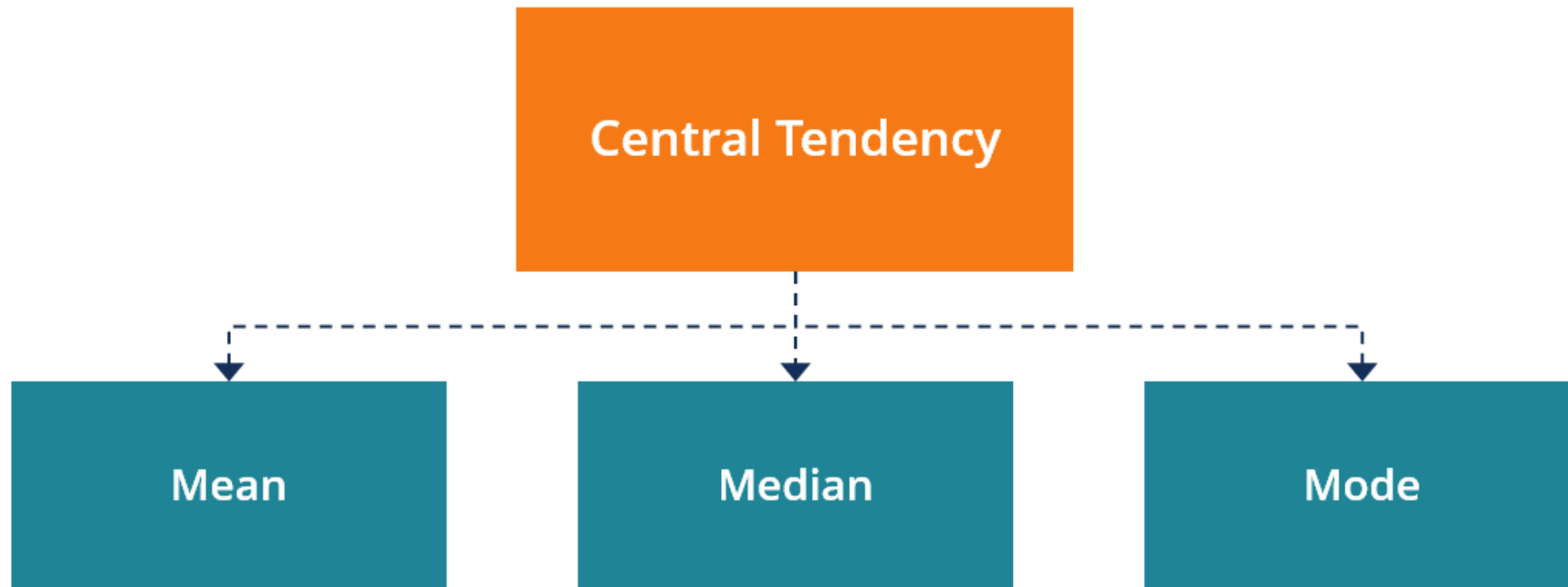


Smallest Squared Deviation

- The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth.
- Changing the target from 10 to 5, calculate the sum of the squared deviations from 5 as $9 + 4 + 1 + 16 + 121 = 151$.
- So, the sum of the squared deviations from 5 < the sum of the squared deviations from 10.
- Is there a value for which the sum of the squared deviations is even smaller than 151?
- **Yes**, it is possible to reach 134.8.



Measures of Central Tendency





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:

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



Median

- The midpoint of a distribution
- Example (please have a look on the table):
 - The 16th highest score (which equals 20) is the median
 - 15 scores below it and 15 scores above it
- Odd numbers = the median is simply the middle number
- Even numbers = the median is the mean of the two middle numbers

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

Number of touchdown passes



Mode

- The most frequently occurring value.
- Example (please have a look on the table):
 - The mode is 18
 - Most frequently occur

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

Number of touchdown passes



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

Trimean

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

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

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



What is the purpose of a Trimean?

- The trimean is almost as resistant to extreme scores as the median and is less subject to sampling fluctuations than the arithmetic mean in extremely skewed distributions.
- It is less efficient than the mean for normal distributions.
- The trimean is a good measure of central tendency and is probably not used as much as it should be.



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: $1 \times 10 \times 100 = 1,000$.
- 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 \prod means to multiply

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

Geometric Mean

- The geometric mean is an appropriate measure to use for averaging rates.
- For example, consider a stock portfolio that began with a value of \$1,000 and had annual returns of 13%, 22%, 12%, -5%, and -13%.
- The question is how to compute average annual rate of return.

Table 4. Portfolio Returns

Year	Return	Value
1	13%	1,130
2	22%	1,379
3	12%	1,544
4	-5%	1,467
5	-13%	1,276



Geometric Mean

- The answer is to compute the geometric mean of the returns.
- Instead of using the percent, each return is represented as a multiplier indicating how much higher the value is after the year.
- This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss.
- The multipliers for this example are 1.13, 1.22, 1.12, 0.95, and 0.87.
- The geometric mean of these multipliers is 1.05.

$$\sqrt[5]{1.276142448} = 1.0499771090332$$



Geometric Mean

- Therefore, the average annual rate of return is 5%.
- Table 5 shows how a portfolio gaining 5% a year would end up with the same value (\$1,276) as shown in Table 4.



Geometric Mean

Table 5. Portfolio Returns

Year	Return	Value
1	5%	1,050
2	5%	1,103
3	5%	1,158
4	5%	1,216
5	5%	1,276



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.



People
Innovation
Excellence

GREATER JAKARTA • BANDUNG • MALANG



Bivariate Data



Describing Univariate Data

- When you conduct a study that looks at a single variable, that study involves univariate data.
- For example, you might study a group of college students to find out their average SAT scores or you might study a group of diabetic patients to find their weights.



Describing Bivariate Data

- Bivariate data is when you are studying two variables.
- A dataset with two variables contains what is called bivariate data, the relationship between two variables.
- For example, you may wish to describe the relationship between the heights and weights of people to determine the extent to which taller people weigh more
- Other example, if you are studying a group of college students to find out their average SAT score and their age, you have two pieces of the puzzle to find (SAT score and age).



Describing Bivariate Data

- Or if you want to find out the weights and heights of diabetic patients, then you also have bivariate data.
- Bivariate data could also be two sets of items that are dependent on each other.
- For example:
 - Ice cream sales compared to the temperature that day.
 - Traffic accidents along with the weather on a particular day.



What is Bivariate Analysis?

- Bivariate analysis means the analysis of bivariate data.
- It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values.
- It usually involves the variables X and Y.
- Univariate analysis is the analysis of one (“uni”) variable.
- Bivariate analysis is the analysis of exactly two variables.
- Multivariate analysis is the analysis of more than two variables.



What is Bivariate Analysis?

- The results from bivariate analysis can be stored in a two-column data table.
- For example, you might want to find out the relationship between caloric intake and weight (of course, there is a pretty strong relationship between the two).
- Caloric intake would be your independent variable, X and weight would be your dependent variable, Y.

Caloric Intake X	Weight Y
3500	250lbs
2000	225lbs
1500	110lbs
2250	145lbs
4500	380lbs



What is Bivariate Analysis?

- Bivariate analysis is not the same as two sample data analysis.
- With two sample data analysis, the X and Y are not directly related.
- You can also have a different number of data values in each sample; with bivariate analysis, there is a Y value for each X.
- Let's say you had a caloric intake of 3,000 calories per day and a weight of 300lbs.
- You would write that with the x-variable followed by the y-variable: (3000,300).



What is Bivariate Analysis?

- Two sample data analysis

Sample 1: 100,45,88,99

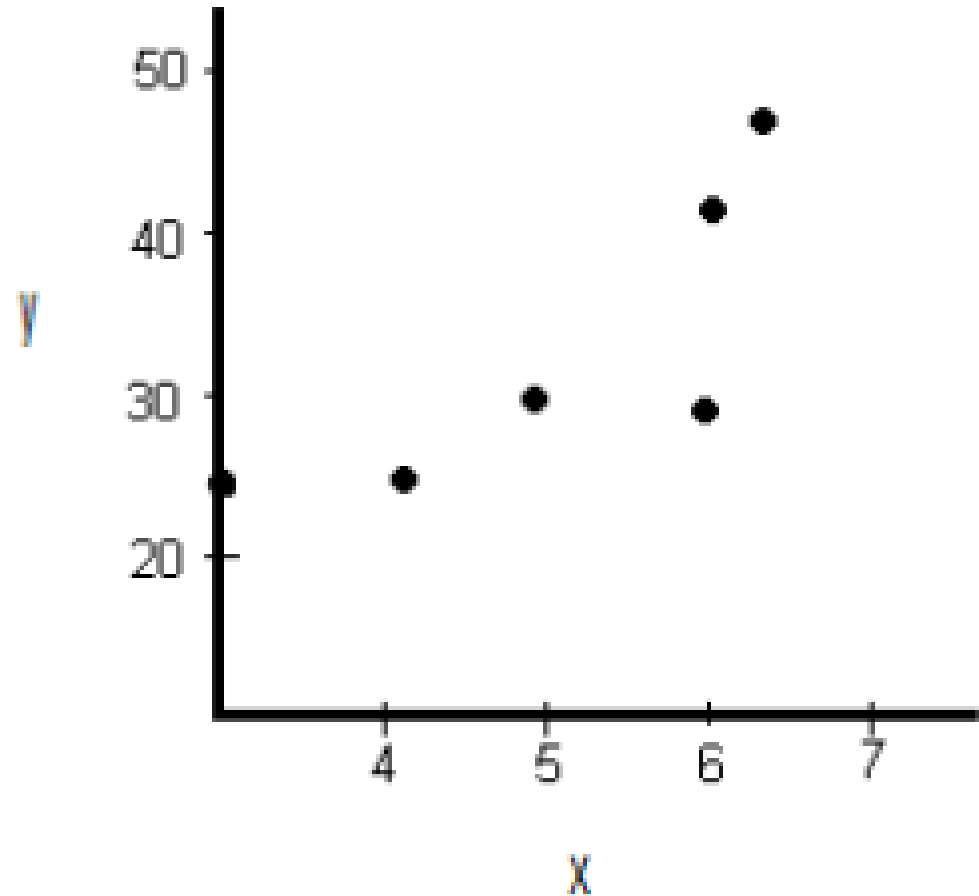
Sample 2: 44,33,101

- Bivariate analysis

$(X,Y)=(100,56),(23,84),(398,63),(56,42)$

Types of Bivariate Analysis

- Common types of bivariate analysis include:
 1. Scatter plots
- These give you a visual idea of the pattern that your variables follow.





Types of Bivariate Analysis

2. Regression Analysis

- Regression analysis is a catch all term for a wide variety of tools that you can use to determine how your data points might be related.
- In the previous image, the points look like they could follow an exponential curve (as opposed to a straight line).
- Regression analysis can give you the equation for that curve or line. It can also give you the correlation coefficient.



Types of Bivariate Analysis

3. Correlation Coefficients

- Calculating values for correlation coefficients are usually performed on a computer, although you can find the steps to find the correlation coefficient by hand here.
- This coefficient tells you if the variables are related.
- Basically, a zero means they aren't correlated (i.e. related in some way), while a 1 (either positive or negative) means that the variables are perfectly correlated (i.e. they are perfectly in sync with each other).

Introduction to Bivariate Data

- Consists of two quantitative variables for each individual.
- Example: people tend to marry other people of about the same age?

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

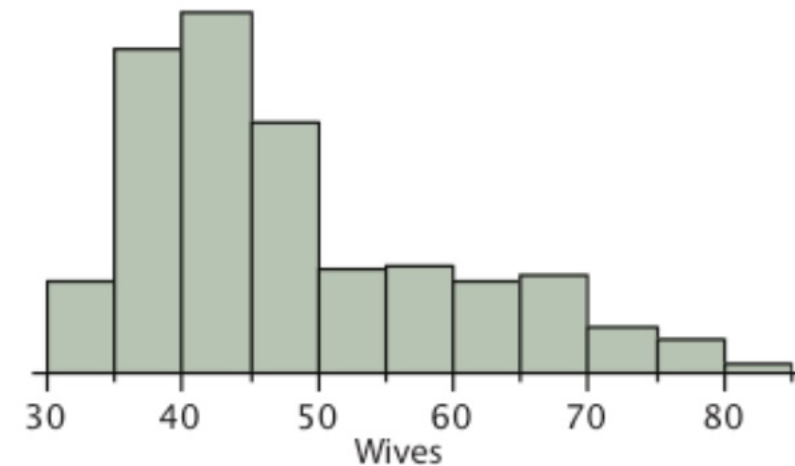
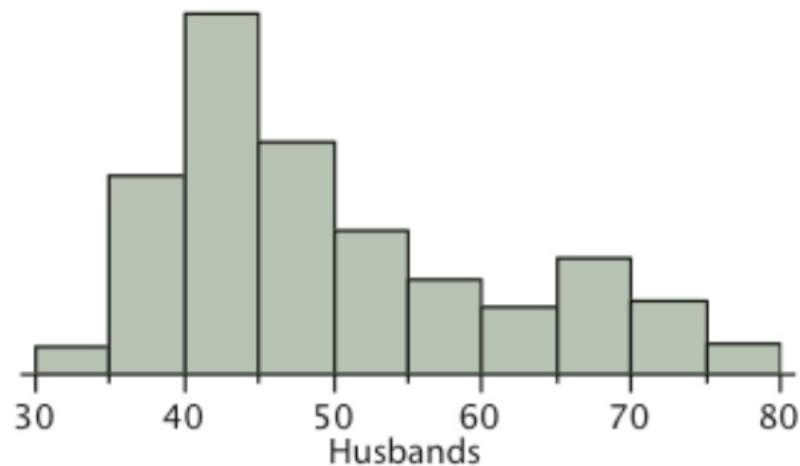
Sample of spousal ages of 10 White American Couples (from dataset consisting of 282 pairs)

Introduction to Bivariate Data

- Use table? TOO MANY
- How to summarize the 282 pairs of ages?

	Mean	Standard Deviation
Husbands	49	11
Wives	47	11

Means and standard deviations of spousal ages



Histograms of spousal ages

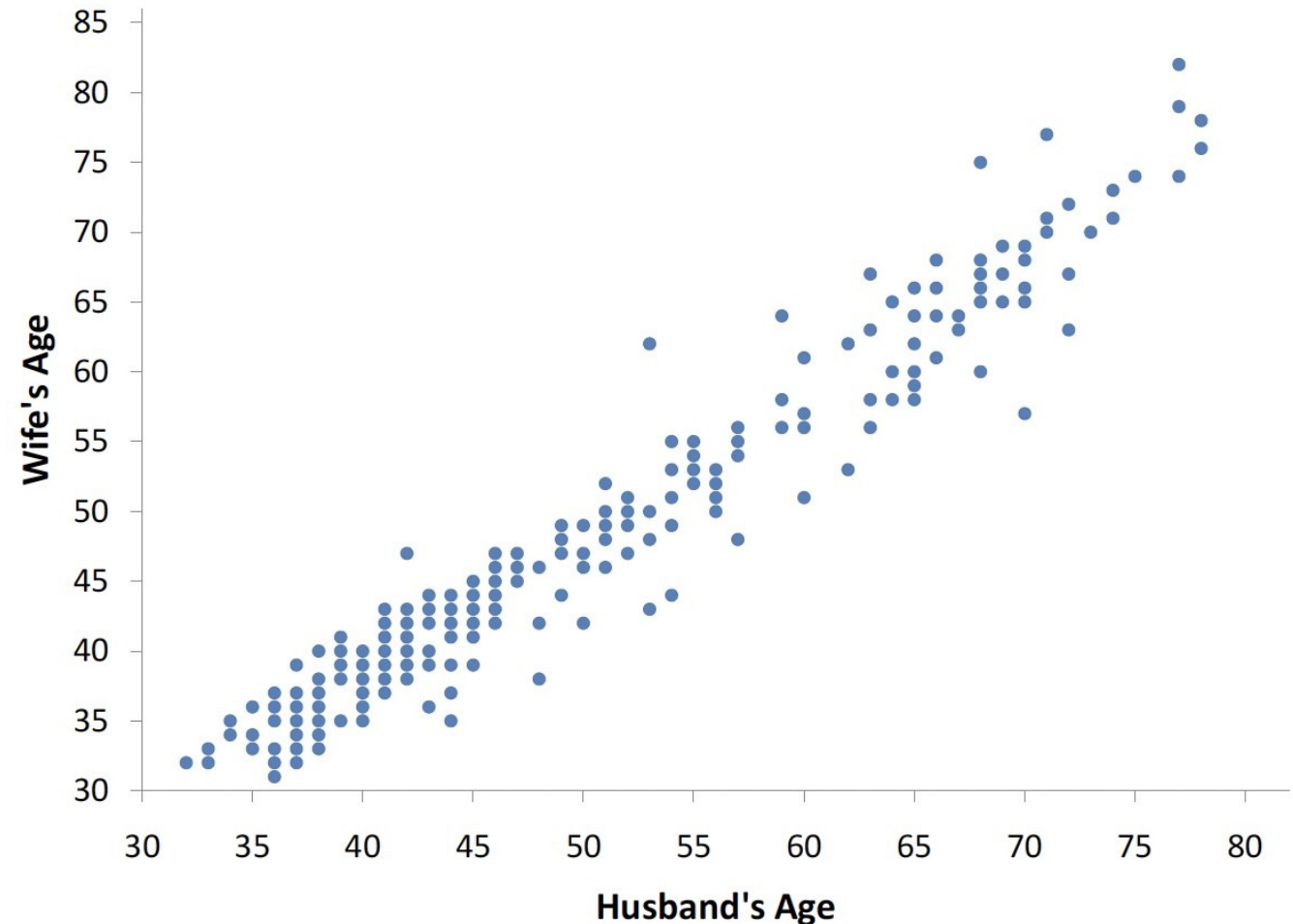


Introduction to Bivariate Data

- Each distribution is fairly skewed with a long right tail.
- From 1st table we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables.
- We cannot say, for example, based on the means alone what percentage of couples has younger husbands than wives. We have to count across pairs to find this out.

Introduction to Bivariate Data

- Displaying using scatter plot
- Example: Scatter plot showing wife's age as a function of husband's age.





Values of the Pearson Correlation

- The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables.
- It is referred to as Pearson's correlation or simply as the correlation coefficient.
- If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.
- The symbol for Pearson's correlation is " ρ " when it is measured in the population and " r " when it is measured in a sample.



Values of the Pearson Correlation

- Pearson's r can range from -1 to 1.
- An r of -1 indicates a perfect negative linear relationship between variables.
- An r of 0 indicates no linear relationship between variables.
- An r of 1 indicates a perfect positive linear relationship between variables

Values of the Pearson Correlation

- Figure 1 shows a scatter plot for which $r = 1$.

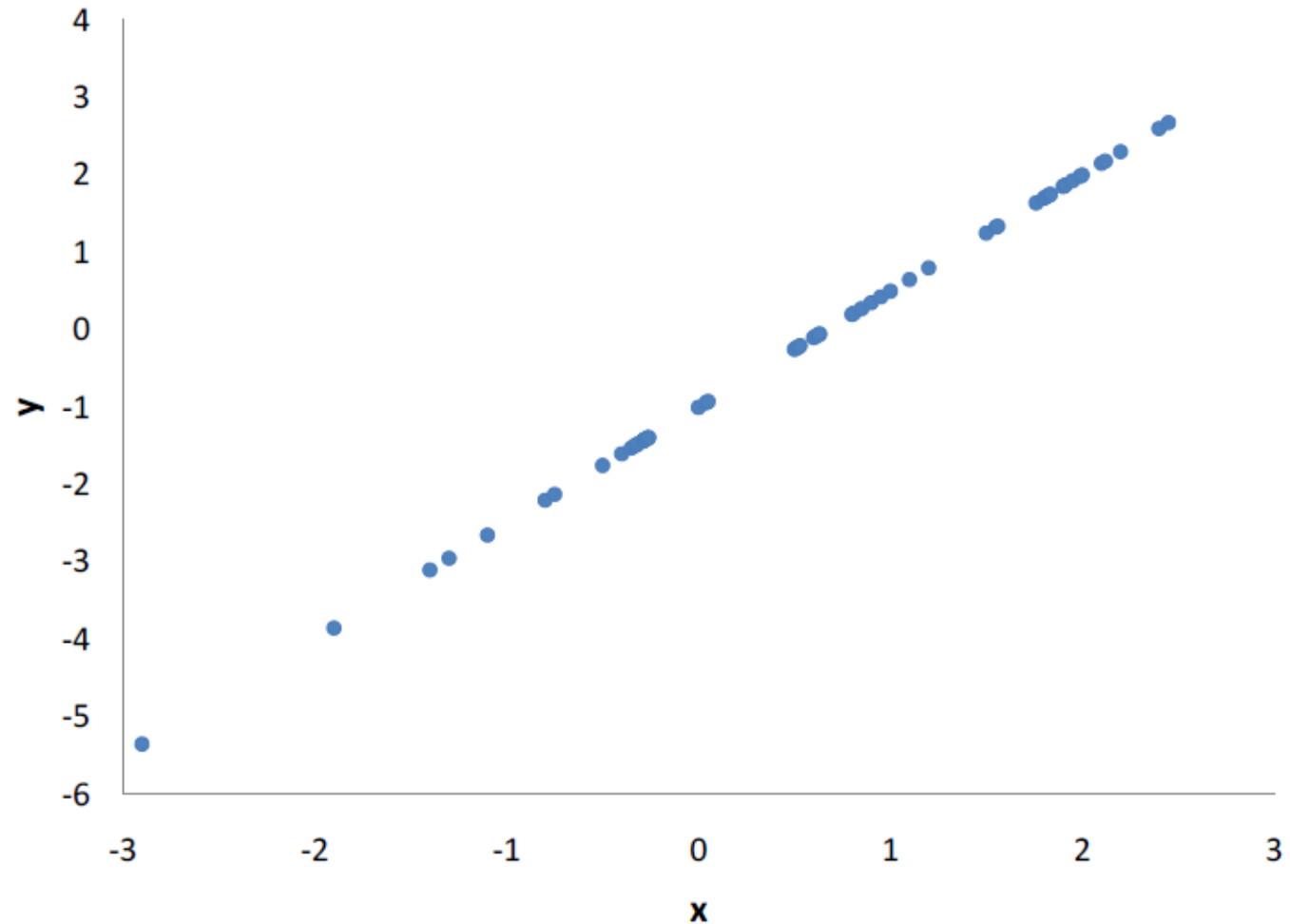
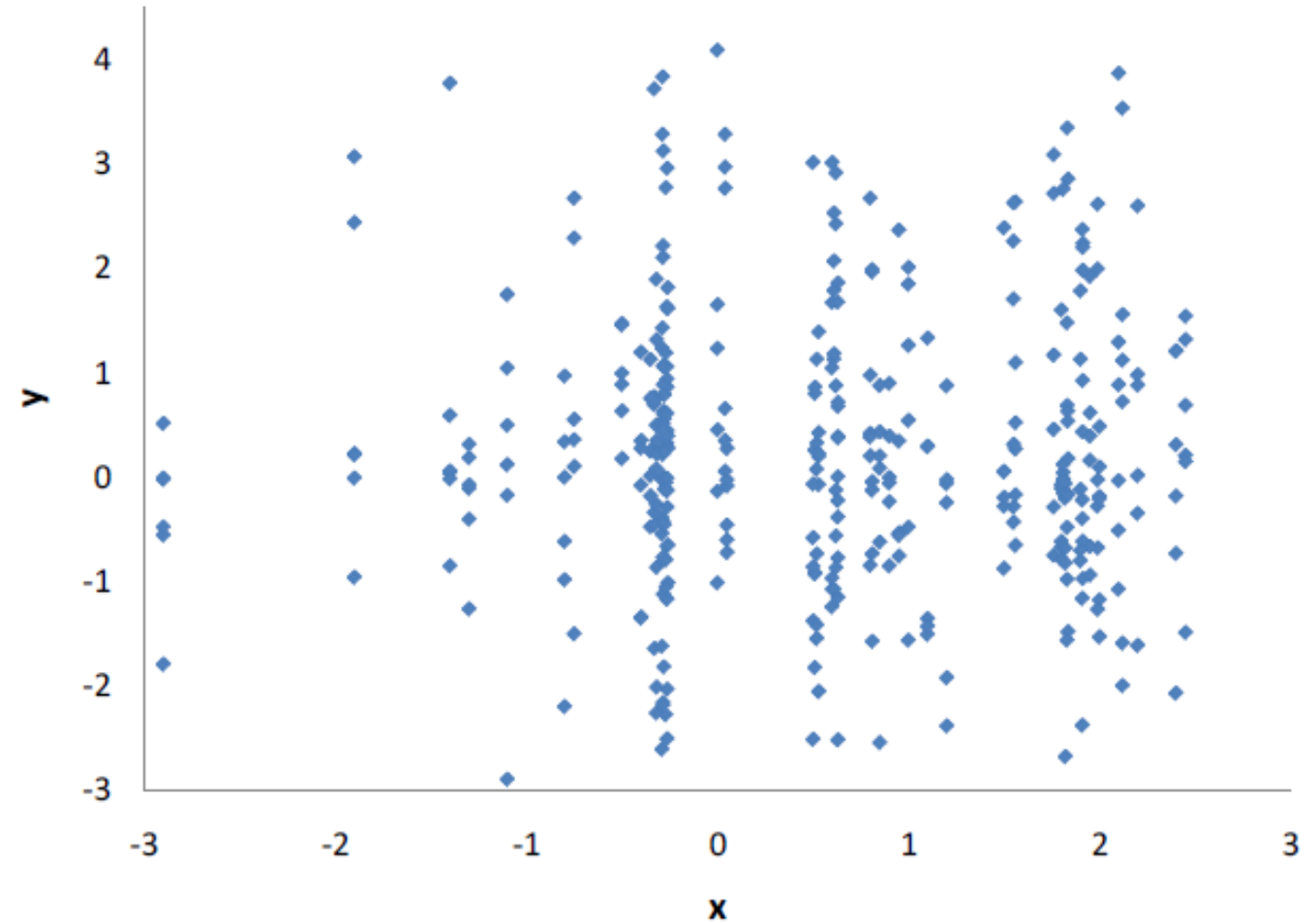


Figure 1. A perfect linear relationship, $r = 1$.

Values of the Pearson Correlation

- A scatter plot for which $r = 0$.
- Notice that there is no relationship between X and Y.



Computing Pearson's r

- There are several formulas that can be used to compute Pearson's correlation.
- Some formulas make more conceptual sense whereas others are easier to actually compute.
- Compute the correlation between the variables X and Y shown in Table 1.

Table 1. Calculation of r .

	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		



Computing Pearson's r

- We begin by computing the mean for X and subtracting this mean from all values of X .
- The new variable is called “ x .” The variable “ y ” is computed similarly.
- The variables x and y are said to be deviation scores because each score is a deviation from the mean.
- Notice that the means of x and y are both 0.
- Next, we create a new column by multiplying x and y .



Computing Pearson's r

- Pearson's correlation is computed by dividing the sum of the xy column ($\sum xy$) by the square root of the product of the sum of the x^2 column ($\sum x^2$) and the sum of the y^2 column ($\sum y^2$)

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

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$



References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at <https://open.umn.edu/opentextbooks/textbooks/459>
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

The background is a solid blue color. On the left side, there are two overlapping circles. The circle in the foreground is a lighter shade of blue and is partially cut off by the left edge of the frame. The circle behind it is a darker shade of blue and is also partially cut off. The text "Thank you" is written in white, sans-serif font, positioned in the lower-left quadrant of the image, overlapping the lighter blue circle.

Thank you