BUAD 2060-004 Business Statistics

Instructor:
Zhezhu Wen

The University of Toledo

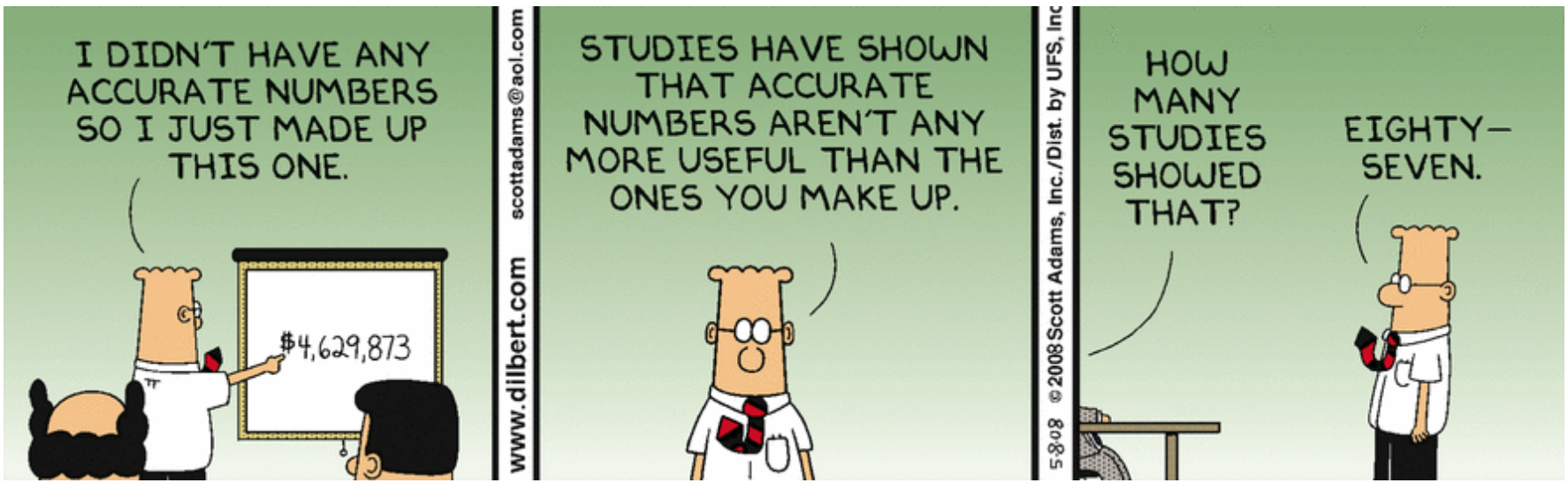
Fall 2020

Outline (1/2)

Outline (2/2)

0. Preface and Motivation

"Statistics" has a Public Relations Problem



What People Talk about Statistics?

> *Stories change people while statistics give them something to argue about.*
>
> *Bernie Siegel*

> *Facts are stubborn, but statistics are more pliable.*
>
> *Mark Twain*

> *Statistics show that of those who contract the habit of eating, very few survive.*
>
> *George Bernard Shaw*

What People Talk about Statistics?

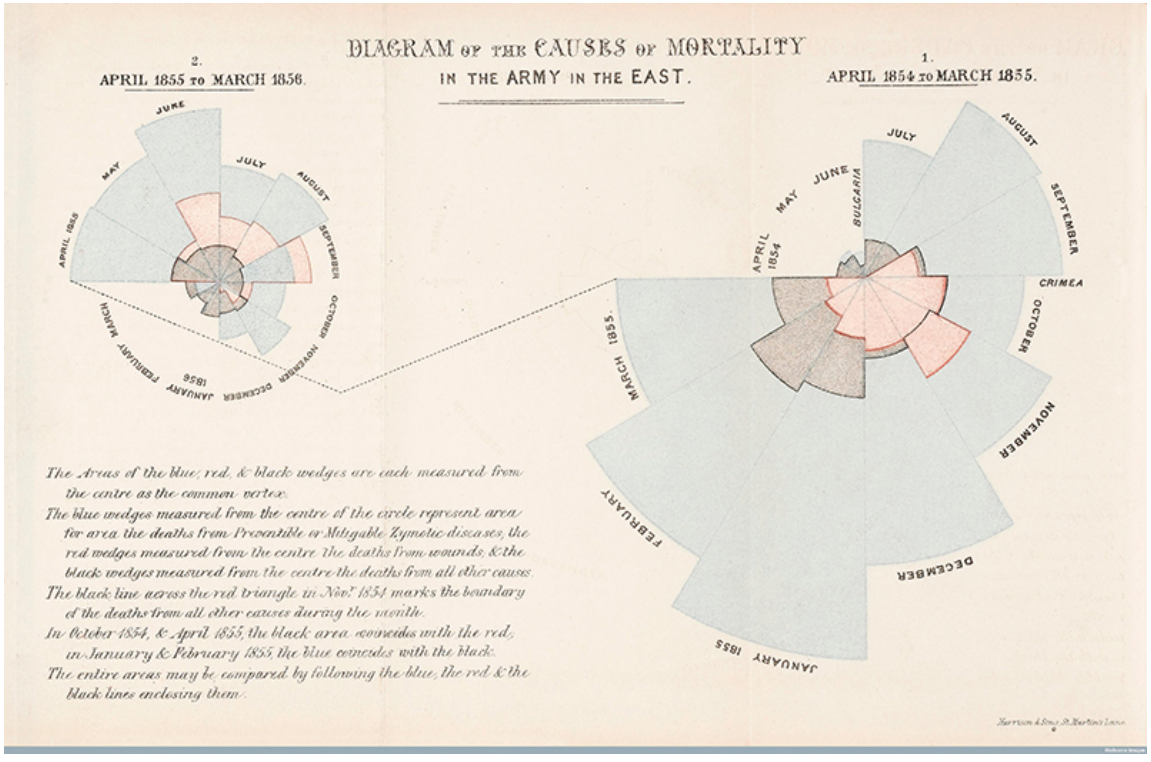
> *In God we trust; all others must bring data.*
>
> *Edwards Deming*

> *Statistics is the grammar of science.*
>
> *Karl Pearson*

> *Statistics are the heart of democracy.*
>
> *Simeon Strunsky*

> *A judicious man looks on statistics not to get knowledge, but to save himself from having ignorance foisted on him.*
>
> *Thomas Carlyle*

People Who Changed the World Using Statistics



By using applied statistical methods, Florence Nightingale (1820 - 1910) made a case for eliminating the practices that contributed to the unsafe and unhealthy environment.

People Who Changed the World Using Statistics (cont.)

Dr. Edwards Deming found great inspiration in the work of Shewhart, the originator of the concepts of *statistical control of processes* and the related technical tool of the control chart, as Deming began to move toward the application of statistical methods to industrial production and management.

What is Statistics?

About the word root:

- The term statistics is ultimately derived from the New Latin statisticum collegium ("council of state") and the Italian word statista ("statesman" or "politician"). - Wikipedia

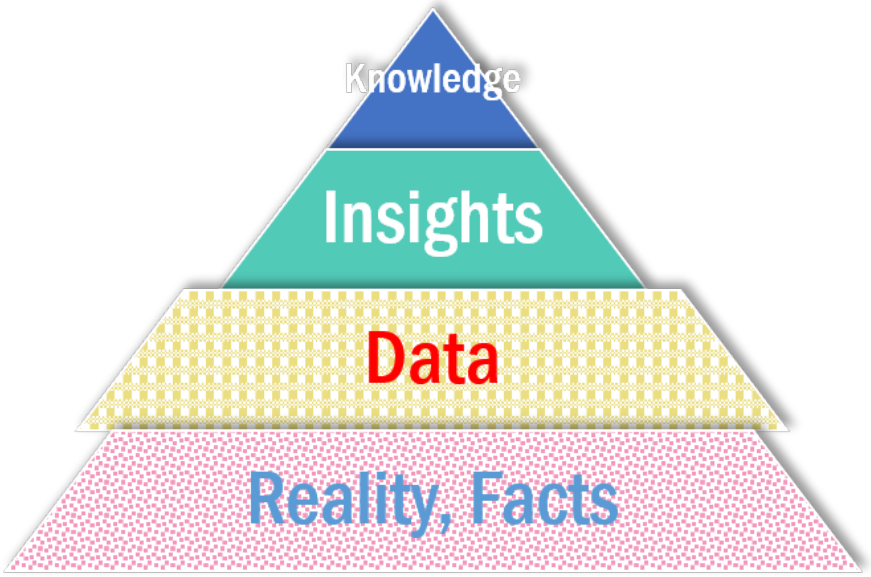
To define statistics:

- Broadly, it refers to *numerical facts* (e.g., mean, medians, percents, and index numbers) that help us understand a variety of business and economic situations.
- Narrowly, it refers to *the art and science* of collecting, analyzing, presenting, and interpreting data.

Why is Statistics a Required Subject in Many Academic Discipline?

- Statistics is the language of scientific exchanges
- Statistics offers the guiding principles for scientific investigations
- Statistics provides the mental frameworks for scientific experiments
- To sum, it is a language, methodology, and framework.
- It is also a prerequisite to many other advanced courses in SCM, Marketing, Management, Finance, etc.

From Data to Knowledge



- Data are the closest possible artifact to the reality.
- The reality is messy and data are spurious.
- Appropriate analysis is the key to extract insights from data.
- Knowledge is nothing but validated insights.

> *It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*
>
> *Arthur Conan Doyle*

Simpson's Paradox - UC Berkerley Gender Bias Case (1973)

At a glance, in the school level, the gender differences in the admission is too large to be considered as a pure chance.

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8,442	**44%**	4,321	35%

Source: Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398-404.

Simpson's Paradox - UC Berkerley Gender Bias Case (cont.)

However, when examining the department level data, more departments appeared to be favoring women applicants.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	35%
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Careers Involving Statistics

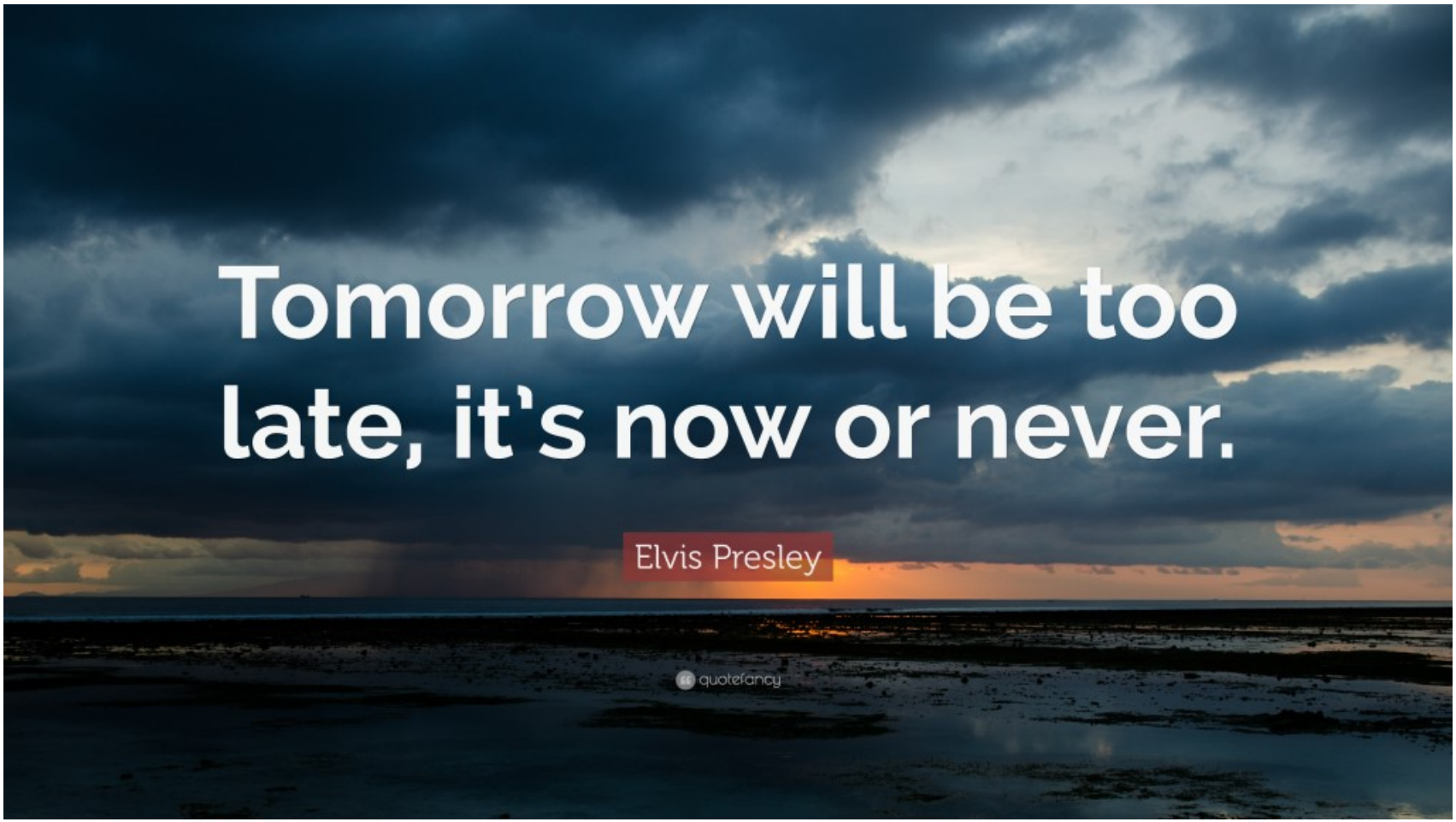
Job Title	Median Salary	Job Growth
Mathematician	$101,900	26%
Market Research Analyst	$63,120	20%
Meteorologist	$94,110	8%
Statistician	$87,780	31%
Operations Research Analyst	$83,390	26%
Financial Analyst	$85,660	6%

* Salary data year 2018; Job growth projection during 2018-2028

* Source: U.S. Bureau of Labor Statistics

* Originally from Study.com

It's Now or Never!



I. Review of Fundamentals

Various Ways of Classifying Data

Based on the Scales of Measurement

- Nominal, Ordinal, Interval, and Ratio

Based on the Nature of the Data

- Categorical (or qualitative) and Quantitative Data

Based on the Continuous Scale

- Binary (or dichotomous), Discrete, and Continuous Data

Based on the Temporality

- Cross-sectional and Time Series Data

The Scales of Measurement

Scale Type	Defined	Examples
Nominal	Data are labels or names used to identify an attribute of the element.	Names: "John Doe", "Jane Doe" Labels: "NY", "OH", "MI"
Ordinal	The data have the properties of nominal data AND the order or rank of the data is meaningful.	Rating: "Good", "Neutral", "Bad" Rank: "First", "Second", "Third"
Interval	The data have the properties of ordinal data AND the interval between observations is expressed in terms of a fixed unit of measure.	Temperature: 0, 3.4, 5.6, 3, 4
Ratio	The data have all the properties of interval data AND the ratio of two values is meaningful.	Height: 5 ft.; Weight: 3 lbs. Debt amount: $345,000

Comparing Different Scales of Measurement

Characteristics	**Nominal**	**Ordinal**	**Interval**	**Ratio**
The order of values is meaningful		Y	Y	Y
"Count" matters	Y	Y	Y	Y
An arithmetic mean has meaning		Y	Y	Y
Can quantify the difference			Y	Y
Can add or subtract values			Y	Y
Can multiply and divide values				Y
"Zero" has meaning				Y

Example of the Data Type

Q: Can you tell what scale of measurement each column uses?

	A	B	C	D	E	F
1	**Nation**	**WTO Status**	**Per Capita GDP**	**Trade Deficit**	**Fitch Rating**	**Fitch Outlook**
2	Armenia	Member	5,400	2,673,359	BB-	Stable
3	Australia	Member	40,800	-33,304,157	AAA	Stable
4	Austria	Member	41,700	12,796,558	AAA	Stable
5	Azerbaijan	Observer	5,400	-16,747,320	BBB-	Positive
6	Bahrain	Member	27,300	3,102,665	BBB	Stable
7	Belgium	Member	37,600	-14,930,833	AA+	Negative
8	Brazil	Member	11,600	-29,796,166	BBB	Stable
9	Bulgaria	Member	13,500	4,049,237	BBB-	Positive
10	Canada	Member	40,300	-1,611,380	AAA	Stable
11	Cape Verde	Member	4,000	874,459	B+	Stable
12	Chile	Member	16,100	-14,558,218	A+	Stable
13	China	Member	8,400	-156,705,311	A+	Stable

Data Source: Textbook Data, Only the first 12 observations are shown.

Data on the Continuous Scale

Data can also classified based on the accuracy of the measurement.

- Binary Scale
  - Only considers two cases or two levels: "Yes" or "No", "Success" or "Failure", "Before" or "After".
  - Convenient in calculating the proportion.
- Discrete Scale
  - Similar with the ordinal scale, the discrete data can be measured using integers.
  - Convenient when outcomes have limited results.
- Continuous Scale
  - Can take any real number (positive, negative, integer, & fraction).
  - Allows many advanced statistical analysis.

Cross-Sectional Data

The cross-sectional and time-series distinction is more about the entire dataset, rather about a specific column.

A cross-sectional dataset provides a snapshot on a specific time point.

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Source: Woodridge (6th Ed.)

Time-series Data

A time-series data observes the same entity overtime.

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Source: Woodridge (6th Ed.)

Check your understanding 1/3

Q: Some hotels ask their guests to rate the hotel's services as excellent, very good, good, and poor. This is an example of the

a. ordinal scale of measurement
b. ratio scale of measurement
c. nominal scale of measurement
d. interval scale of measurement

Check your understanding 2/3

Q: The scale of measurement that has an inherent zero value defined is the

a. ratio scale
b. nominal scale
c. ordinal scale
d. interval scale

Check your understanding 3/3

Q: Data collected over several time periods are

a. time series data
b. time controlled data
c. cross-sectional data
d. time cross-sectional data

Common Math Notations: Series, Index, and Sum

Use **index letter** i or j to enumerate multiple observations from the same series (for example, a column)

$$x_i = \{x_1, x_2, x_3, ...x_I\}$$

In combination with the **sum notation** $\sum$, the indexing makes the calculation of the *series sum* very convenient. (A column total)

$$\sum_{i=1}^{I} x_i = x_1 + x_2 + x_3 + ... + x_I$$

(The summation in plain English: Add the first element from x to Ith element.)

Common Math Notations (cont.)

One example of the series sum where each element of the series is reduced by a constant value k

$$\sum_{i=1}^{I}(x_i - k) = (x_1 - k) + (x_2 - k) + (x_3 - k) + \ldots + (x_I - k)$$

Check your Understanding 1/2

Consider the following example ($k = 5$):

i	x_i	$x_i - k$	$x_i \times k$
1	32	27	160
2	41	36	205
3	57	52	285
Total	130	115	650

- Can you use the summation symbol and index letter to express the summation of each column?

Check your Understanding 2/2

What's the difference: $\sum_{i=1}^{N}(x_i - k)$ vs. $\sum_{i=1}^{N} x_i - k$?

Naming Conventions: True vs. Observed Values

To simplify the expression, we stick to the following naming conventions. There are largely two groups:

- One group is the *true values*. We don't always get to observe these values but we wish to know.
- The other group is the *observed values*. We can always observe these values through samples.

	True Value	Observed Value
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Variance	σ^2	s^2
Proportion	p	$\bar{p}$
Correlation	ρ	r
Size of the Observation	N	n

Check Your Understanding

Can you explain the following expressions verbally?

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

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

Hint: If a calculation involves a series such as x_i, it is convenient to use a tabular approach to solve the problem.

II. Descriptive Statistics

Introduction to Descriptive Statistics

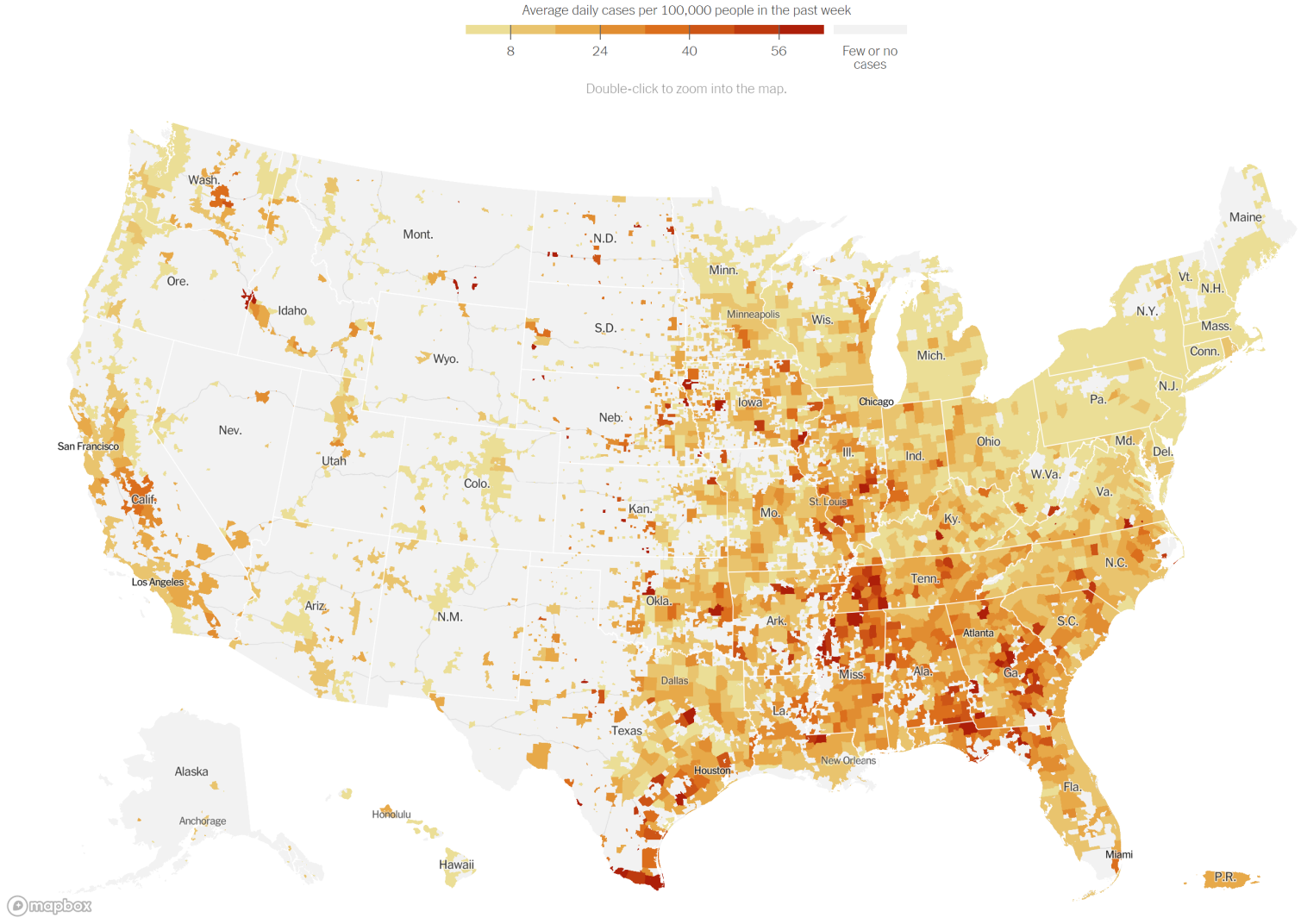
Descriptive statistics is a set of techniques that are being used to summarize a given data.

Table 1: Frequently Used Descriptive Statistics

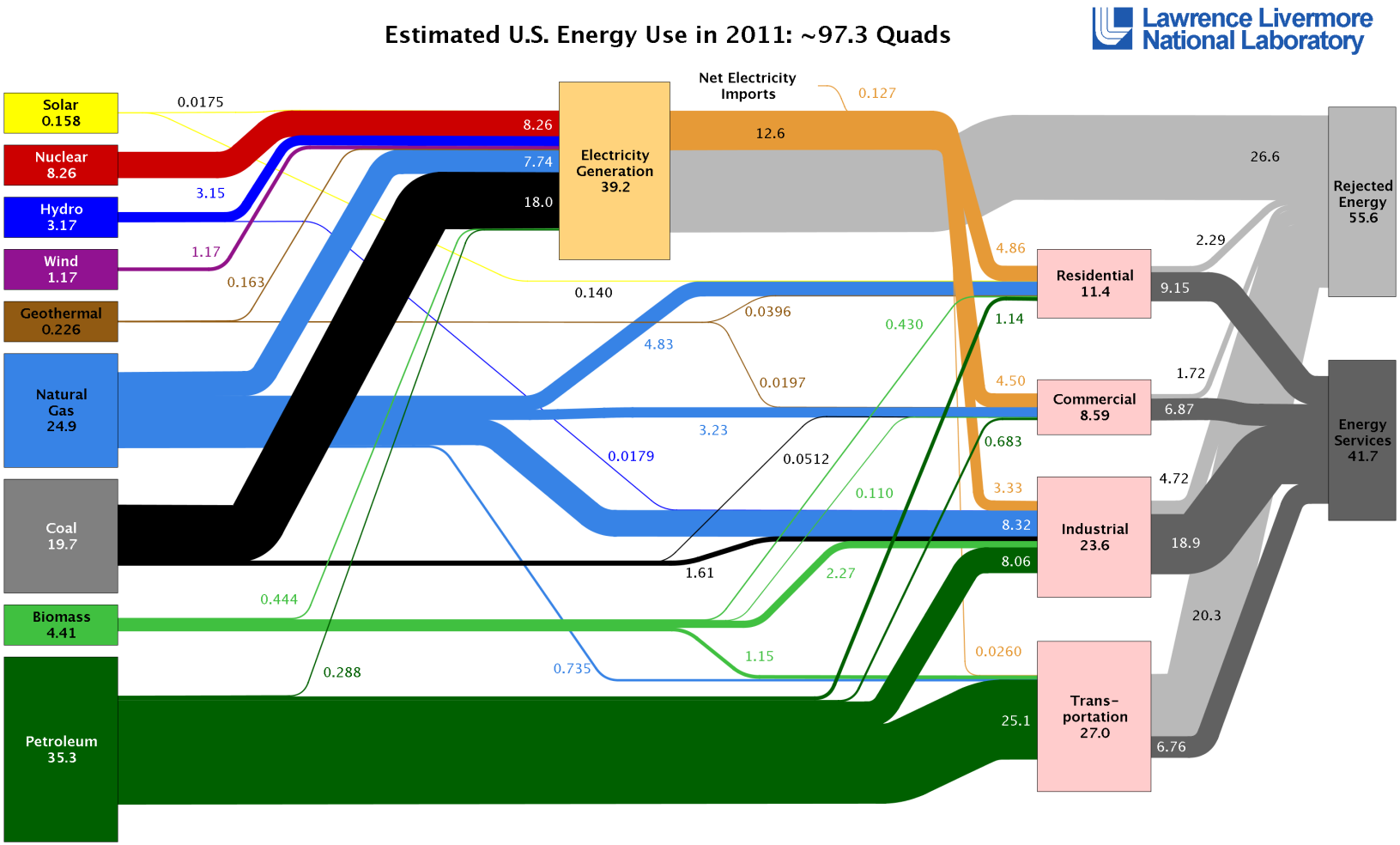
	Graphical	**Tabular**	**Numerical**
Single Categorical	Bar Chart	Frequency Table	
Single Quantitative	Histogram	Frequency Table	Mean, S.d.
Two Categorical	Stacked Bar Chart	Crosstabulation	
Two Quantitative	Scatter Plot	Crosstabulation	Covariance, Correlation Coefficient

There are many more and you can get really creative in terms of using visual elements to communicate.

Some Fancy Visualization Examples (1/2)



Some Fancy Visualization Examples (2/2)



Source: LLNL 2012. Data is based on DOE/EIA-0384(2011), October, 2012. If this information or a reproduction of it is used, credit must be given to the Lawrence Livermore National Laboratory and the Department of Energy, under whose auspices the work was performed. Distributed electricity represents only retail electricity sales and does not include self-generation. EIA reports flows for non-thermal resources (i.e., hydro, wind and solar) in BTU-equivalent values by assuming a typical fossil fuel plant "heat rate." The efficiency of electricity production is calculated as the total retail electricity delivered divided by the primary energy input into electricity generation. End use efficiency is estimated as 80% for the residential, commercial and industrial sectors, and as 25% for the transportation sector. Totals may not equal sum of components due to independent rounding. LLNL-MI-410527

Exhibition of the Data - *Ch2. PelicanStores.xlsx*

(Note: to save space, only showing the first and the last five observations. Originally $n = 100$)

	A	B	C	D	E	F	G	H
1	**Customer**	**Type of Customer**	**Items**	**Net Sales**	**Method of Payment**	**Gender**	**Marital Status**	**Age**
2	1	Regular	1	39.50	Discover	Male	Married	32
3	2	Promotional	1	102.40	Proprietary Card	Female	Married	36
4	3	Regular	1	22.50	Proprietary Card	Female	Married	32
5	4	Promotional	5	100.40	Proprietary Card	Female	Married	28
6	5	Regular	2	54.00	MasterCard	Female	Married	34
96	95	Regular	3	66.00	American Express	Female	Married	46
97	96	Regular	1	39.50	MasterCard	Female	Married	44
98	97	Promotional	9	253.00	Proprietary Card	Female	Married	30
99	98	Promotional	10	287.59	Proprietary Card	Female	Married	52
100	99	Promotional	2	47.60	Proprietary Card	Female	Married	30
101	100	Promotional	1	28.44	Proprietary Card	Female	Married	44

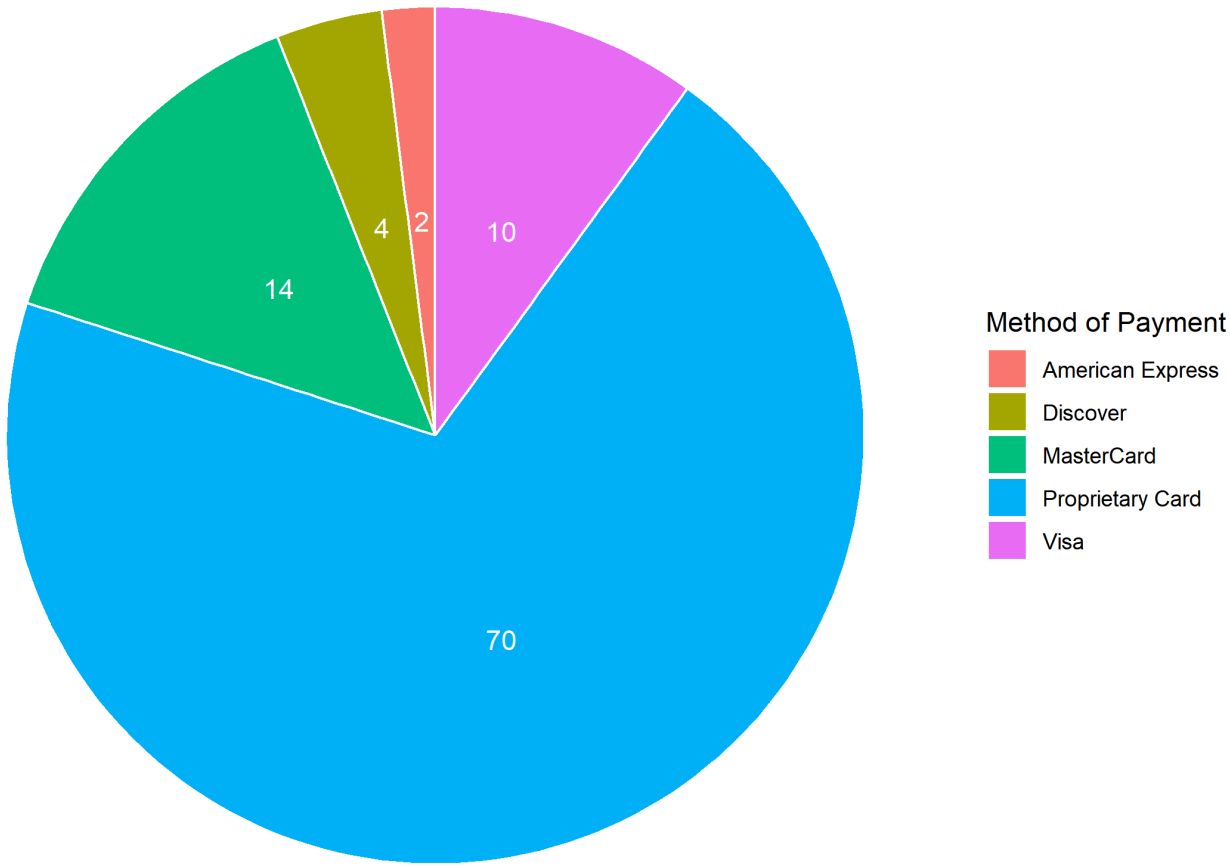
Summarizing One Categorical Variable - Frequency Analysis

Method of Payment Variable (or Column)

Card Type	Frequency	Relative Frequency	Cumulative Relative Frequency
American Express	2	2.00%	2.00%
Discover	4	4.00%	6.00%
MasterCard	14	14.00%	20.00%
Proprietary Card	70	70.00%	90.00%
Visa	10	10.00%	100.00%
Grand Total	**100**	**100.00%**	

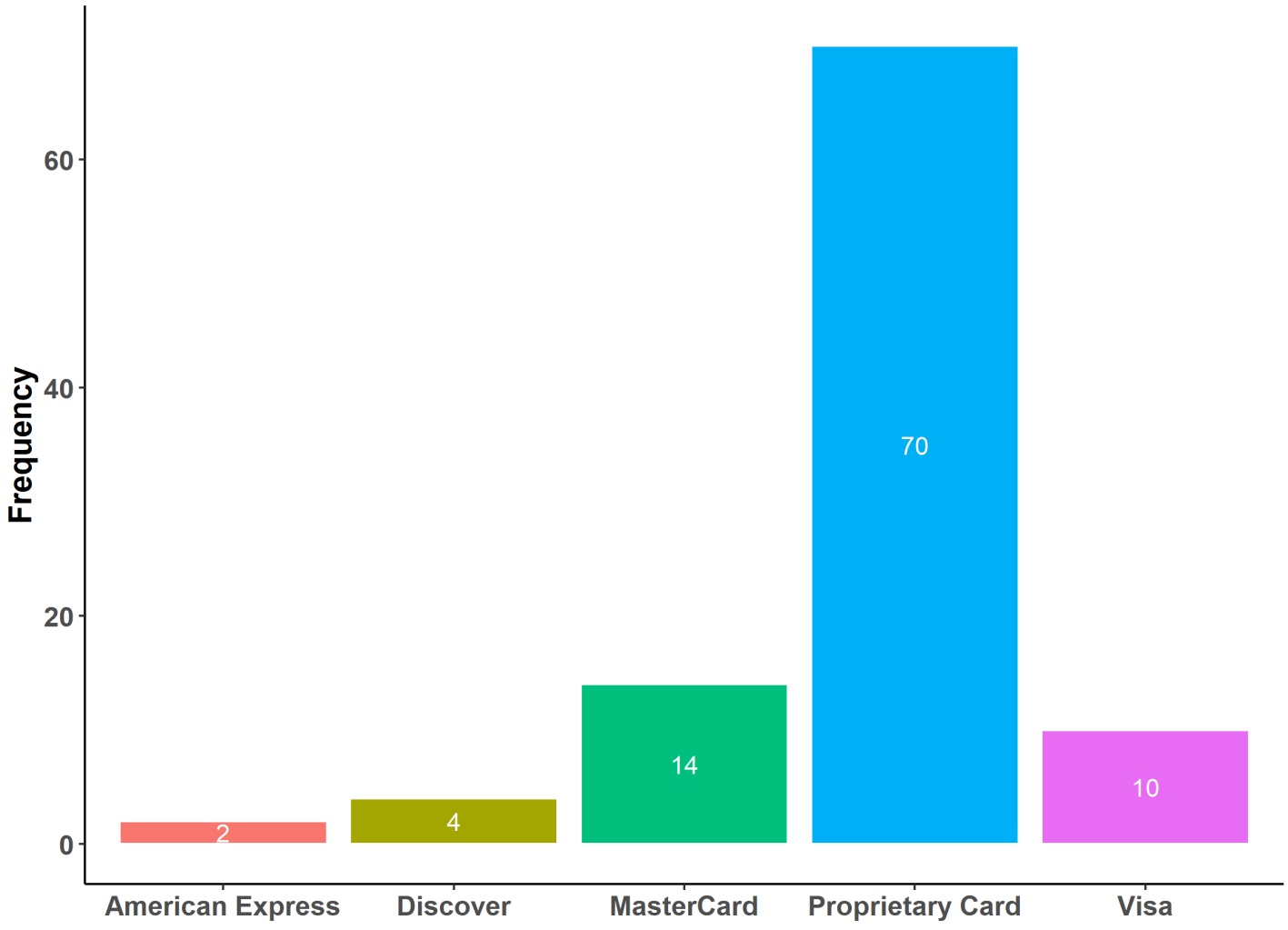
This information will become the basis for many visualization.

Summarizing One Categorical Variable - Pie Chart



The size of the angle for each "pie slice" represents the relative frequency of the category.

Summarizing One Categorical Variable - Bar Chart



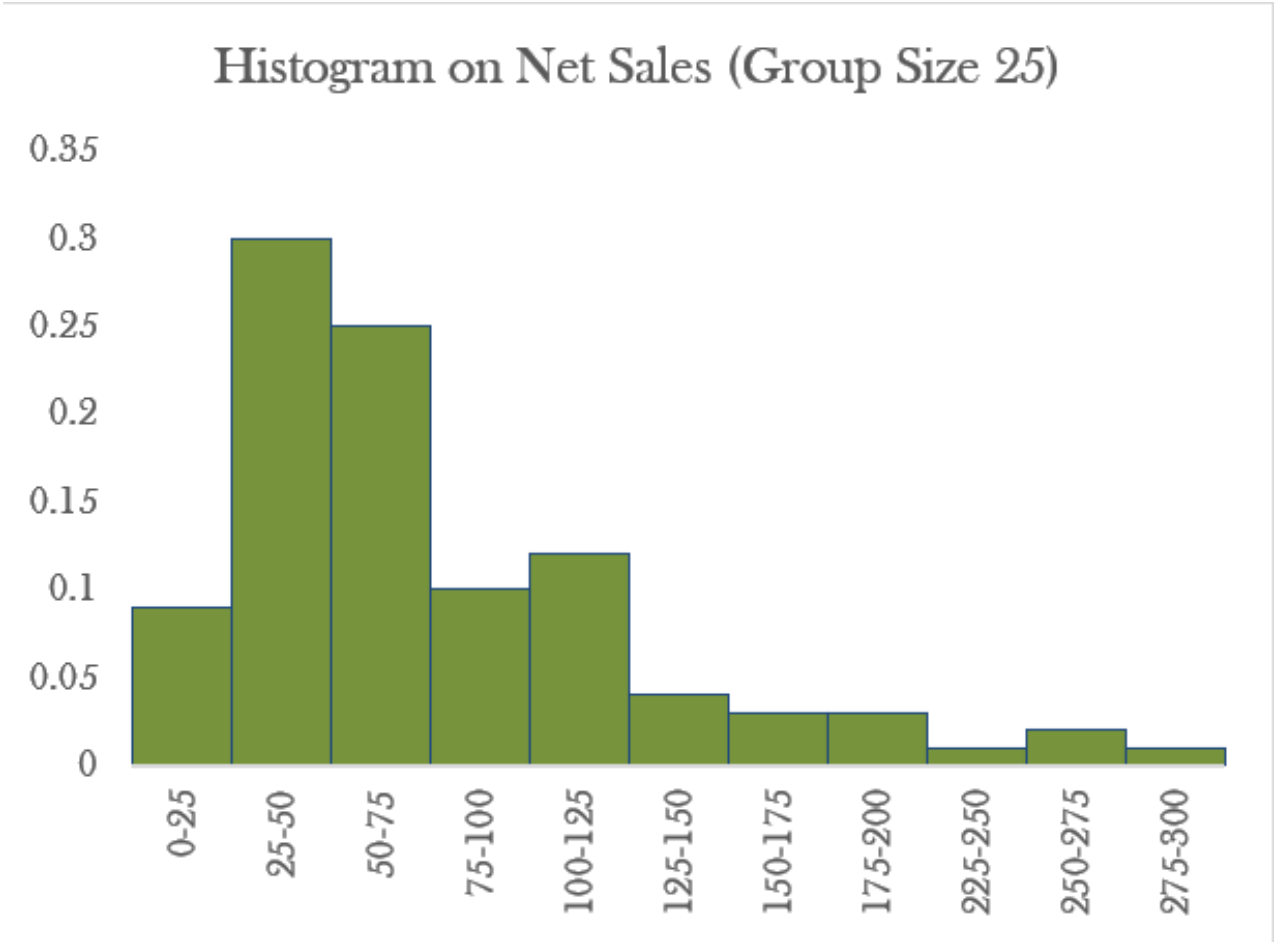
The height of each bar represents the frequency of the category.

Summarizing One Continuous Variable - Frequency Analysis

Net Sales Variable

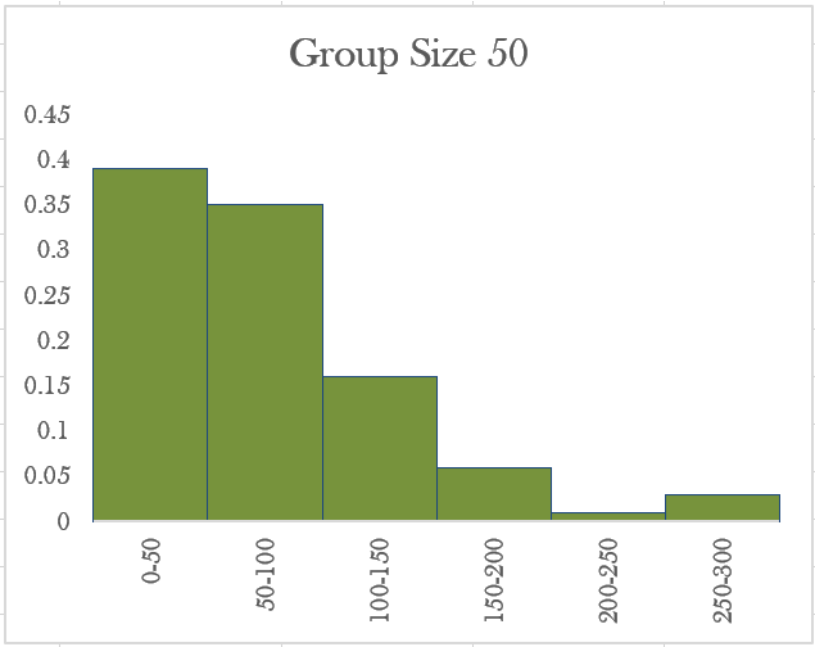
Net Sales	Frequency	Relative Frequency	Cumulative Relative Frequency
0-25	9	9.00%	9.00%
25-50	30	30.00%	39.00%
50-75	25	25.00%	64.00%
75-100	10	10.00%	74.00%
100-125	12	12.00%	86.00%
125-150	4	4.00%	90.00%
150-175	3	3.00%	93.00%
175-200	3	3.00%	96.00%
225-250	1	1.00%	97.00%
250-275	2	2.00%	99.00%
275-300	1	1.00%	100.00%
Grand Total	**100**	**100.00%**	

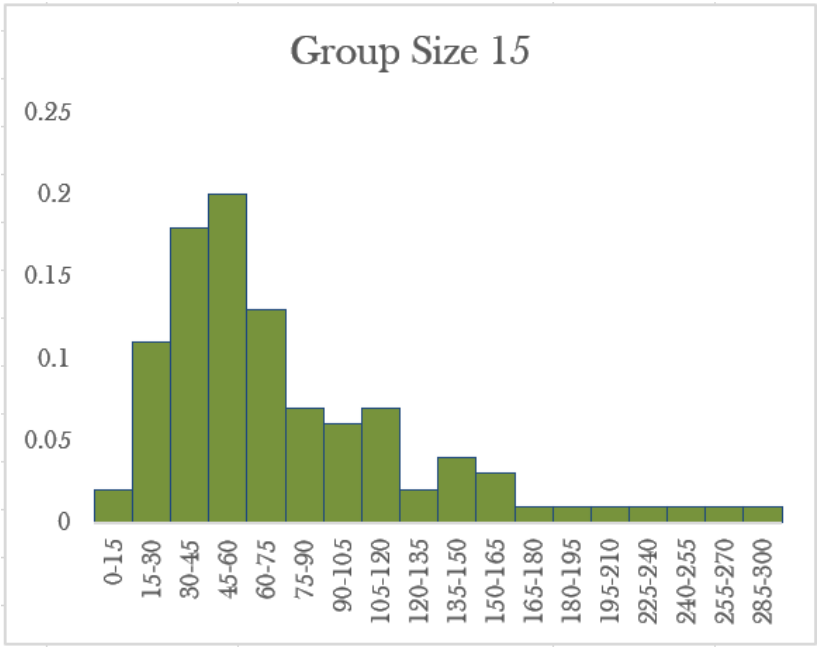
Summarizing One Continuous Variable - Histogram (1/5)



The height of each bar represents the frequency of the group.

Summarizing One Continuous Variable - Histogram (2/5)

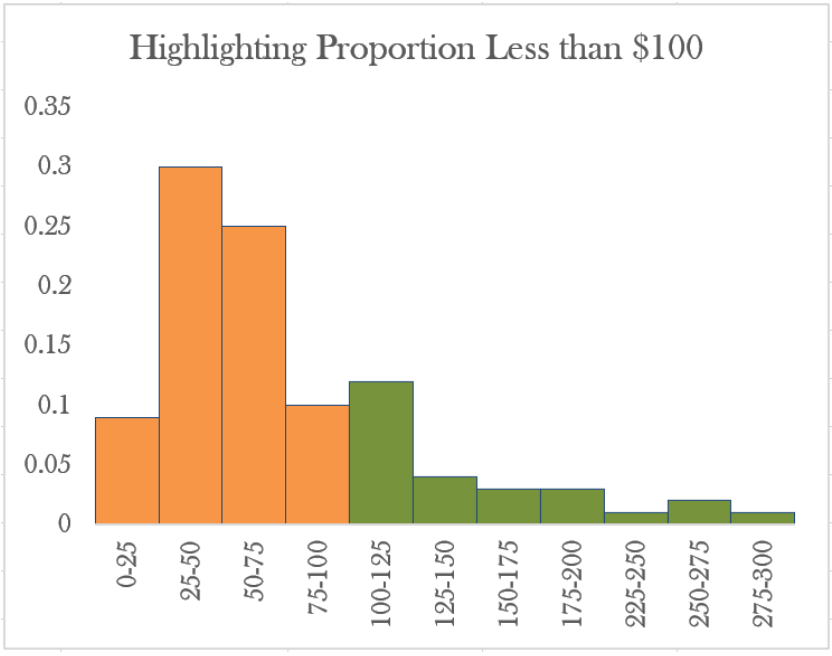


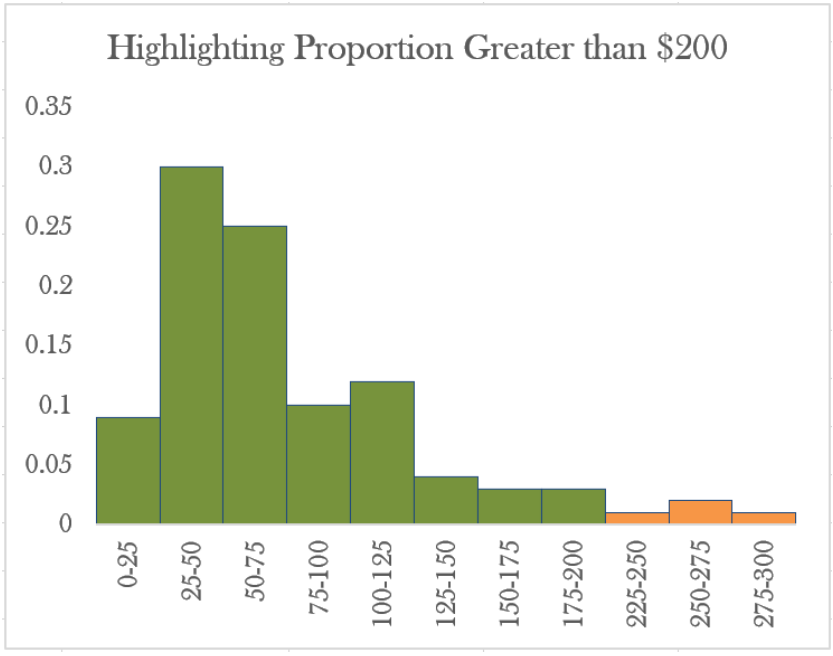


- Depending on the group size of choice, you may ended up obtain very different looking histogram on the same variable.
- The width (or size) of a group is determined by the number of groups you are planning to use.

$$\text{Group Width} = \frac{Max(x_i) - Min(x_i)}{\text{Group Number}}$$

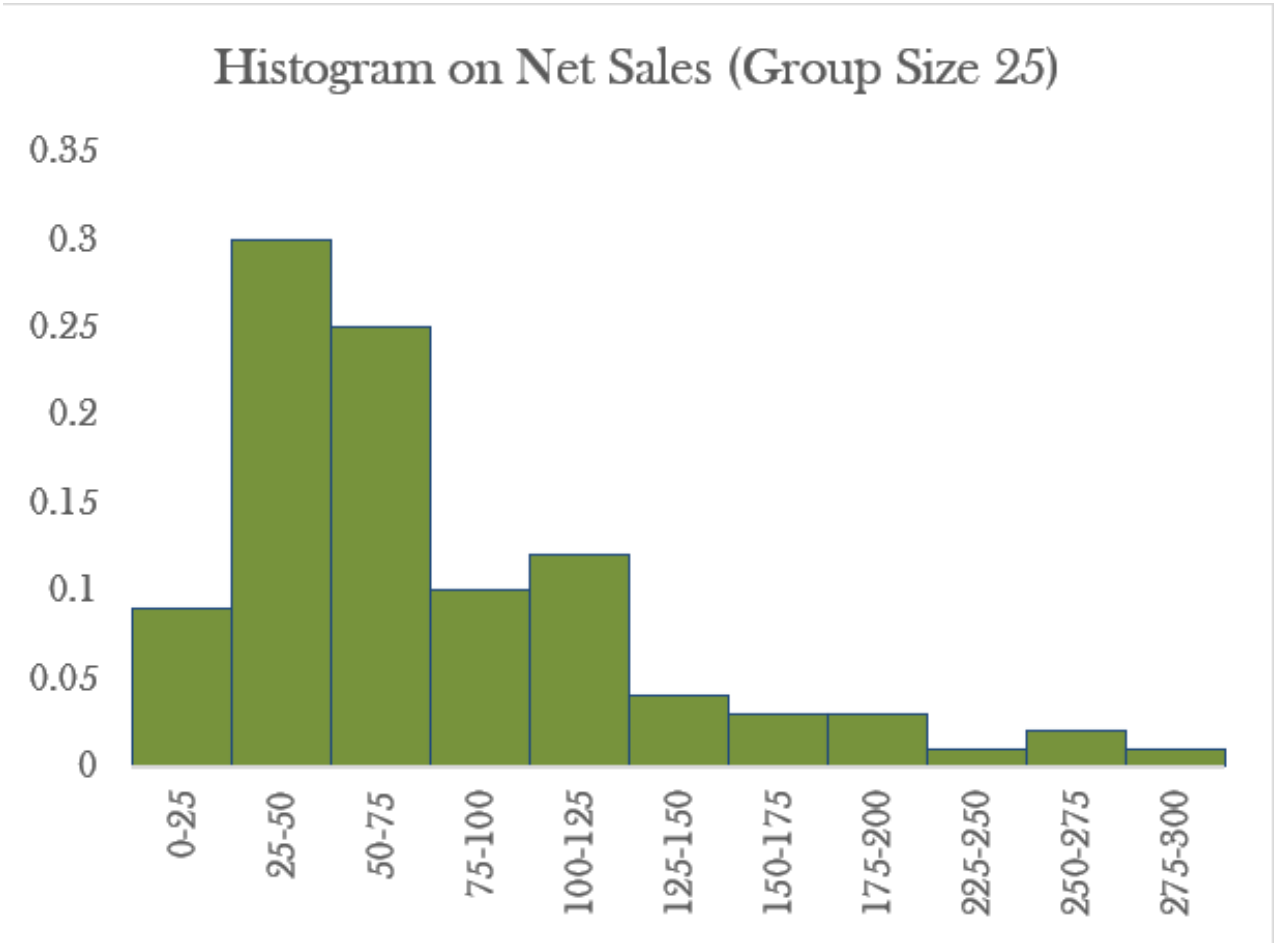
Summarizing One Continuous Variable - Histogram (3/5)





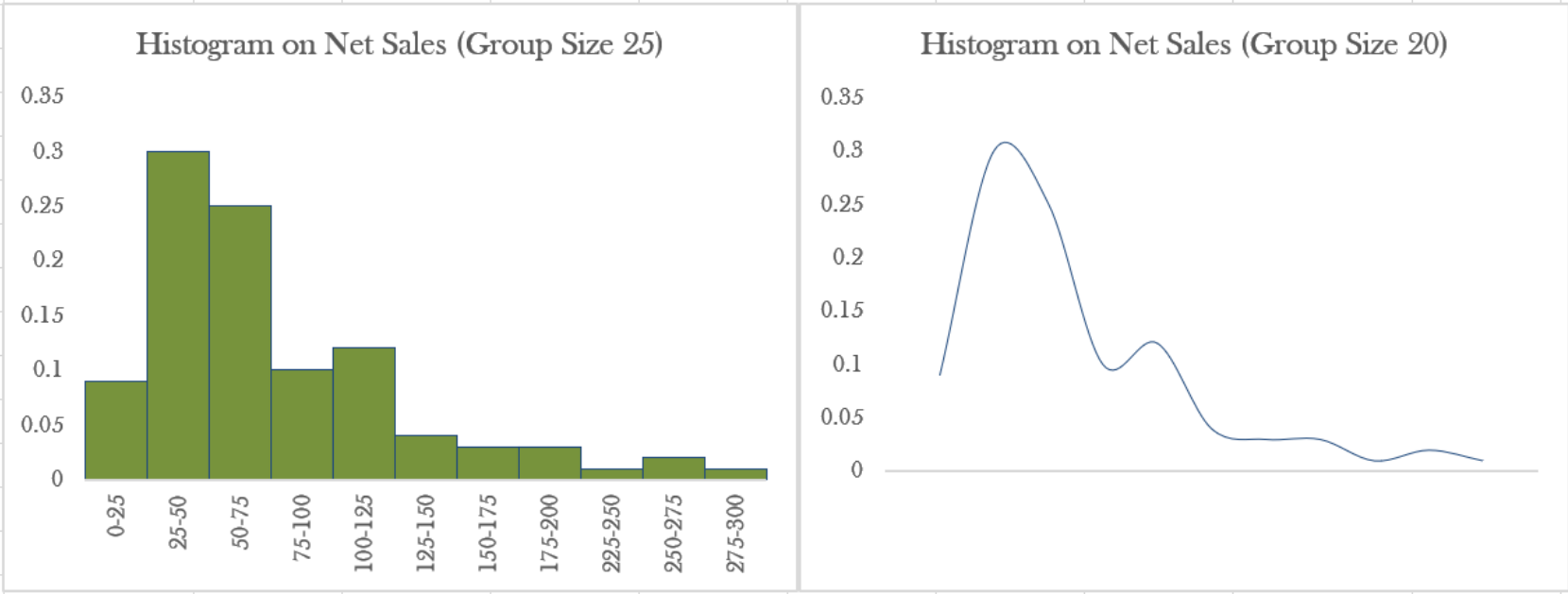
Because all bars are in the same continuous scale, identification of the ***proportion for a sub-group*** *becomes very convenient.*

Summarizing One Continuous Variable - Histogram (4/5)



We can visually identify the point where the most observations are ***concentrated*** *(aka. central tendency), and the pattern of the* ***spread*** *(aka. variability).*

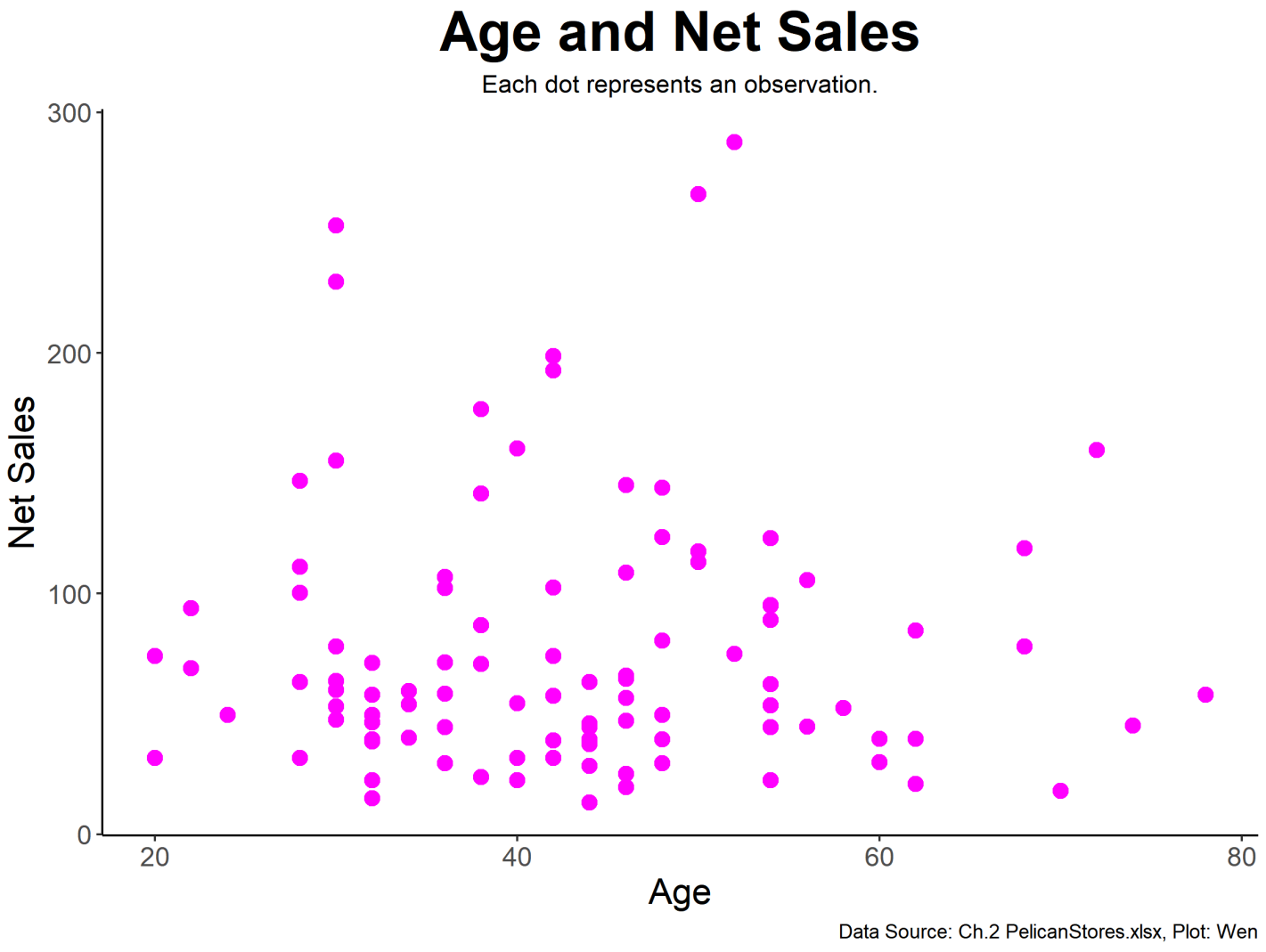
Summarizing One Continuous Variable - Histogram (5/5)



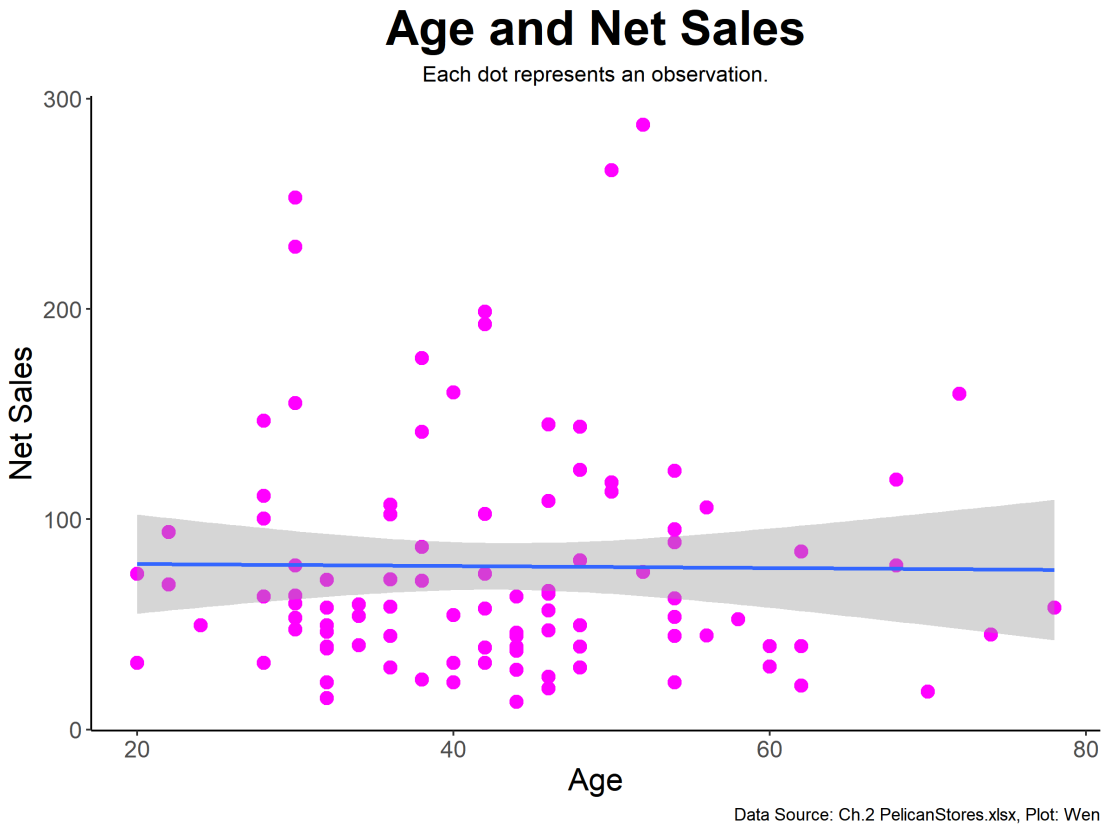
We can also attempt to simplify its appearance by drawing a fit line.

If we assume the total area under the fit line is the same with the histogram, we can borrow calculus to obtain the proportion of a sub-group (Note: Don't worry! No one knows what that means at this point!!!).

Visualizing the Relations Between Two Variables - Scatter Plot



Visualizing the Relations Between Two Variables - Scatter Plot (cont.)

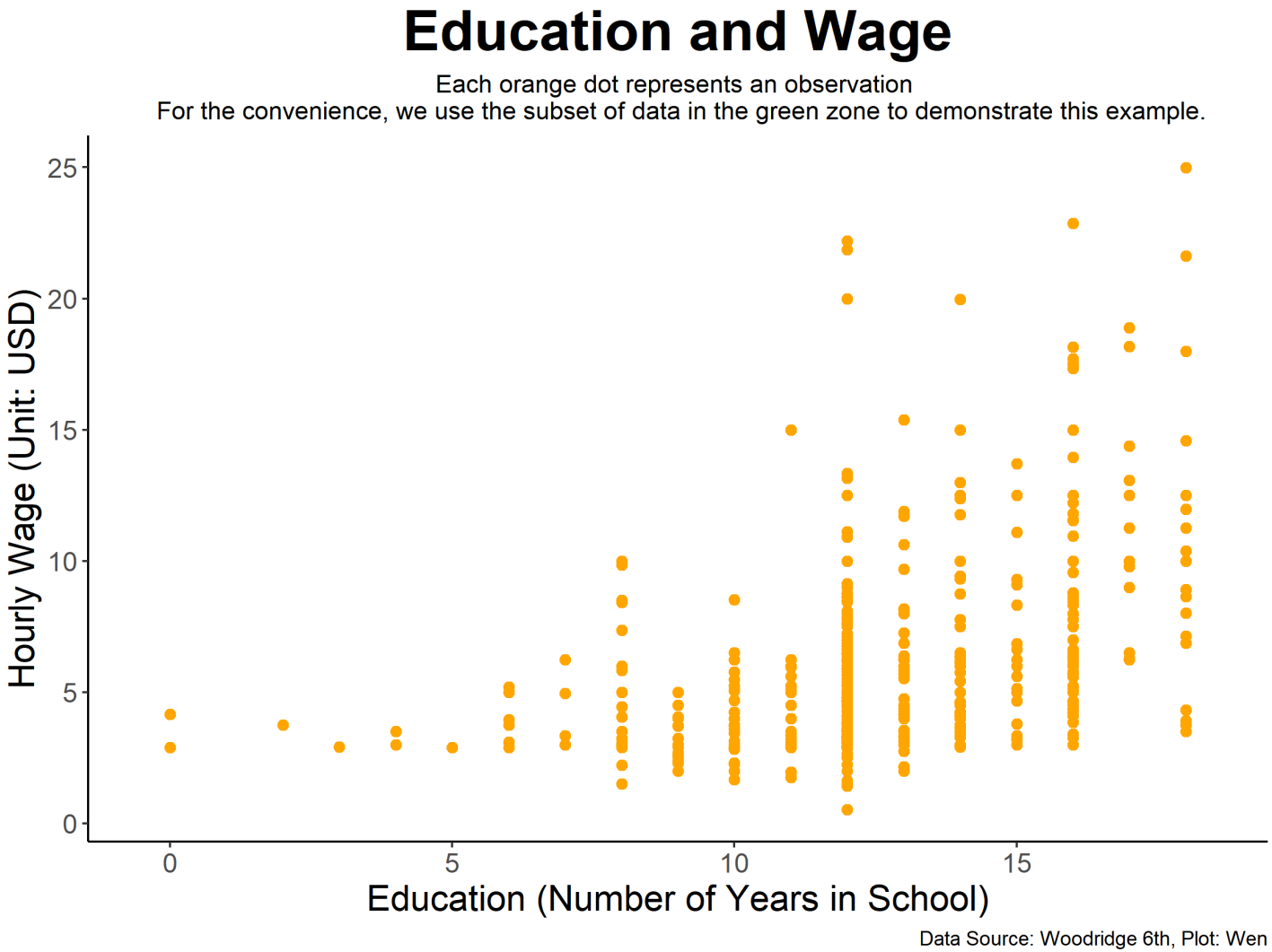


Based on the scatter plot, we can identify a line (regression line) that can represent the relationship between the two variables.

$$\text{Net Sales} = 79.6592 - 0.0478 \cdot \text{Age}$$

Check Your Understanding

How would you evaluate the relationship between the following two variables? Can you identify the regression line?



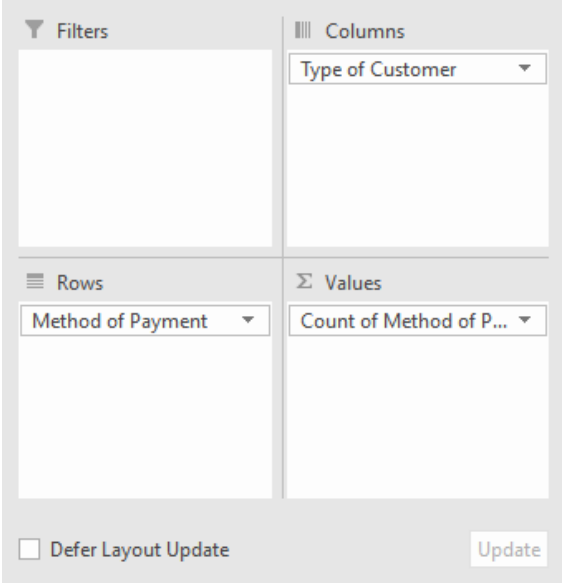
Frequency Analysis for Two Qualitative Variables

Cross-tabulating the *Payment Method* against the *Type of Customers*

Method of Payment	Promotional	Regular
American Express	1	1
Discover	0	4
MasterCard	7	7
Proprietary Card	57	13
Visa	5	5
Grand Total	**70**	**30**

What do you see from the results?

Excel Pivot Table Implementation



Frequency Analysis for One Qualitative and One Continuous Variable

Cross-tabulating the *Payment Method* against the *Net Sales*

Method of Payment	0-50	50-100	100-150	150-200	200-250	250-300
American Express	0	1	0	0	0	1
Discover	1	3	0	0	0	0
MasterCard	6	5	3	0	0	0
Proprietary Card	27	23	12	5	1	2
Visa	5	3	1	1	0	0
Grand Total	**39**	**35**	**16**	**6**	**1**	**3**

What do you see from the results?

Frequency Analysis for Two Continuous Variable

Cross-tabulating the *Age* against the *Net Sales*

Age Group	0-50	50-100	100-150	150-200	200-250	250-300
20-34	11	12	3	1	1	1
35-49	19	13	8	4	0	0
50-64	7	8	4	0	0	2
65-79	2	2	1	1	0	0
Grand Total	**39**	**35**	**16**	**6**	**1**	**3**

What do you see from the results?

Check Your Understanding 1/5

A frequency distribution is

- a. a tabular summary of a data set showing the relative frequency
- b. a graphical form of representing data
- c. a tabular summary of a data set showing the frequency of items in each of several non-overlapping classes
- d. a graphical device for presenting qualitative data

Check Your Understanding 2/5

The sum of the relative frequencies for all classes will always equal

- a. the sample size
- b. the number of classes
- c. one
- d. any value greater than one

Check Your Understanding 3/5

In constructing a frequency distribution for quantitative data, the approximate class width is computed as

- a. (largest data value - smallest data value)/(number of classes)
- b. (largest data value - smallest data value)/(sample size)
- c. (smallest data value - largest data value)/(sample size)
- d. (largest data value)/(number of classes)

Check Your Understanding 4/5

The Numbers of Hours Worked (n = 400)

Number of Hours	Frequency
0 - 9	20
10 - 19	80
20 - 29	200
30 - 39	100

Refer to the frequency distribution above, the proportion (fraction) of people working 19 hours or less is

- a. 100
- b. 0.25
- c. 0.95
- d. 0.05

Check Your Understanding 5/5

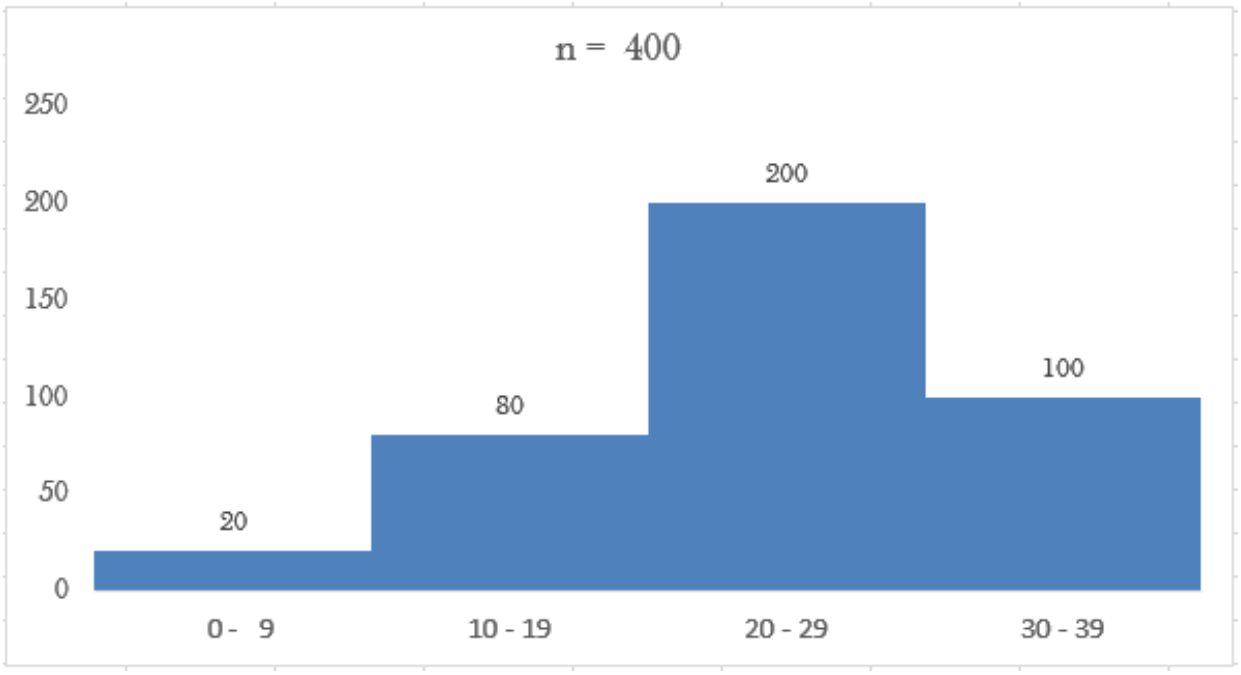
The Numbers of Hours Worked (n = 400)

Number of Hours	Frequency
0 - 9	20
10 - 19	80
20 - 29	200
30 - 39	100

Refer to the frequency distribution above, the cumulative relative frequency for the 20 – 29 class is

- a. is 300
- b. is 0.25
- c. is 0.75
- d. is 0.5

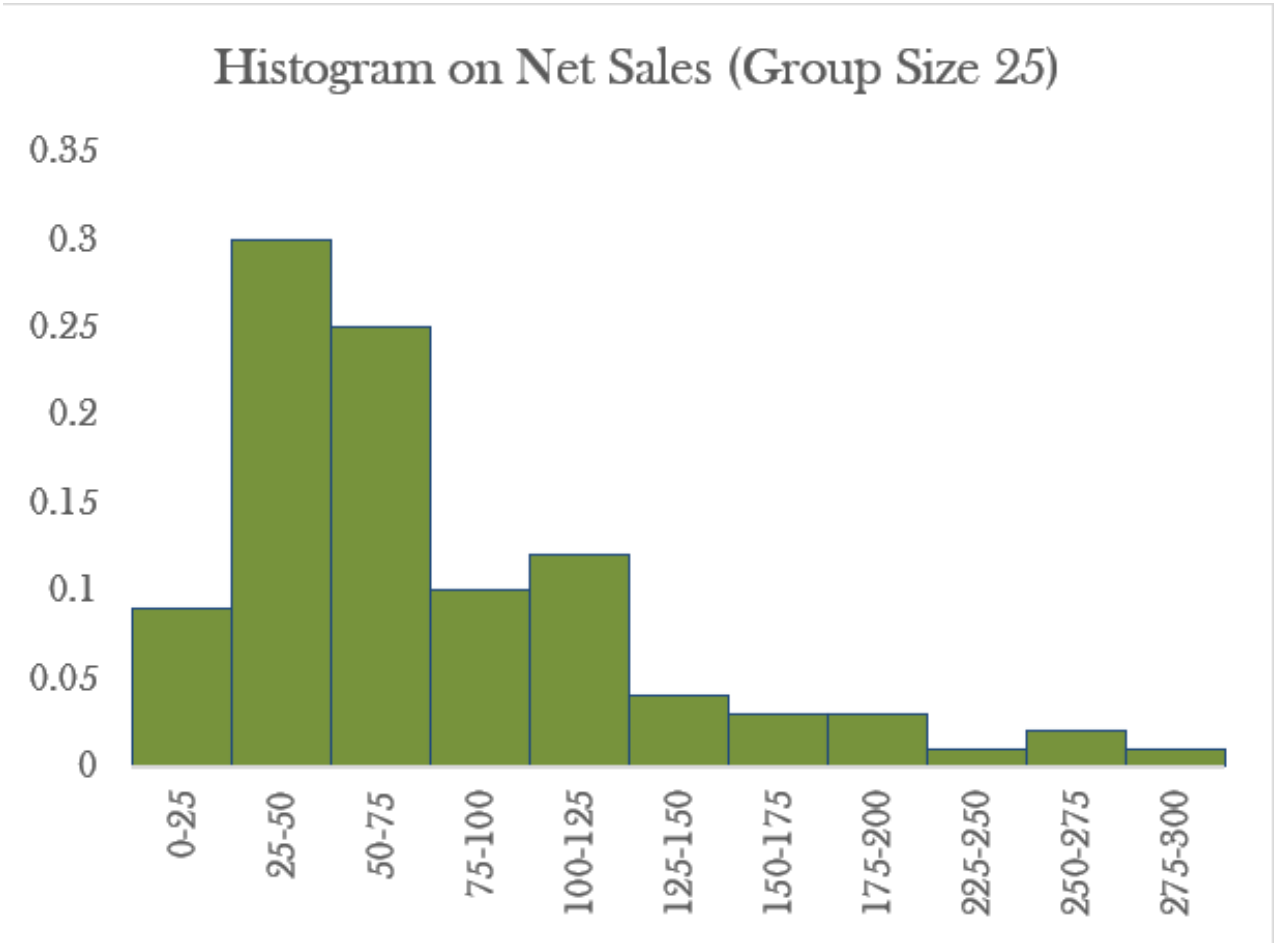
Check Your Understanding: Bonus



Please identify the following section quickly.

- Greater than 9
- Less than 19
- Greater than 10 but less than 29
- Equal to 30

Before the Numerical Measures - Review of Histogram



We can visually identify the point where the most observations are ***concentrated*** *(aka. central tendency), and the pattern of the* ***spread*** *(aka. variability).*

Numerical Measures: The Measures of Central Tendency

Mean (aka. Mathematical Mean or Average)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{or} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Excel Formula: =AVERAGE(RANGE)

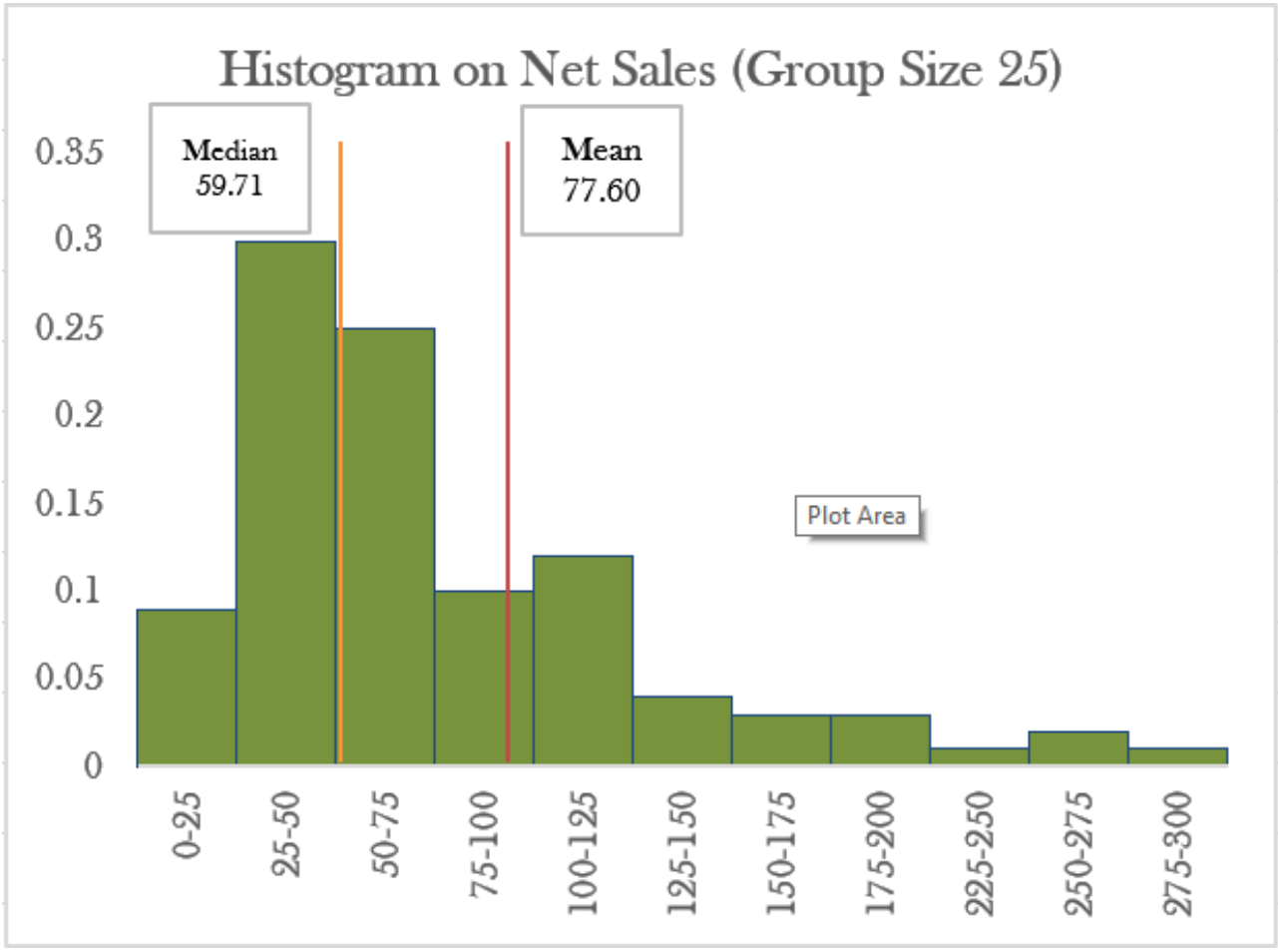
Median

The middle number

(when ranked from the lowest to the highest)

Excel Formula: =Median(RANGE)

Numerical Measures: The Measures of Central Tendency



To represent the middle, the mean can be sensitive to the extreme values. Therefore, when extreme values are clearly present, we typically use median.

Numerical Measures: The Measures of Variability

Range

$$range = max(x_i) - min(x_i)$$

Excel Formula: =Max(RANGE)-Min(RANGE)

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \quad \text{or} \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Excel Formula: =Var.P(RANGE) or *=Var.S(RANGE)*

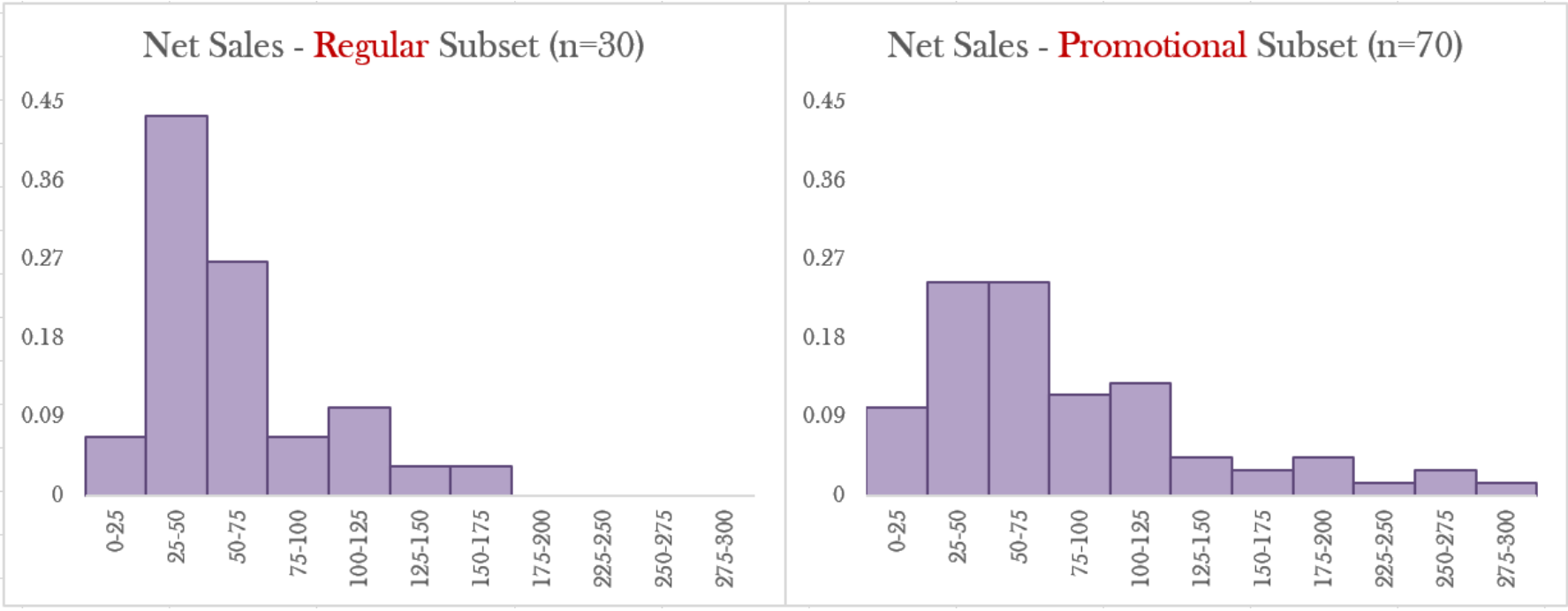
Standard Deviation (In short, s.d.)

$$\sigma = \sqrt{\sigma^2} \quad \text{or} \quad s = \sqrt{s^2}$$

Excel Formula: =Stdev.P(RANGE) or *=Stdev.S(RANGE)*

The Measure of Variability Visualized

How would you comment on the level of spread that each group shows?



	Mean	Median	Range	Variance	S.D.
Regular	61.992	51.000	137.250	1,229.761	35.068
Promotional	84.290	63.420	274.360	3,777.614	61.462

Let's Get Closer to the Variance and S.D.

Let's Think About the Meaning of the Variance

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

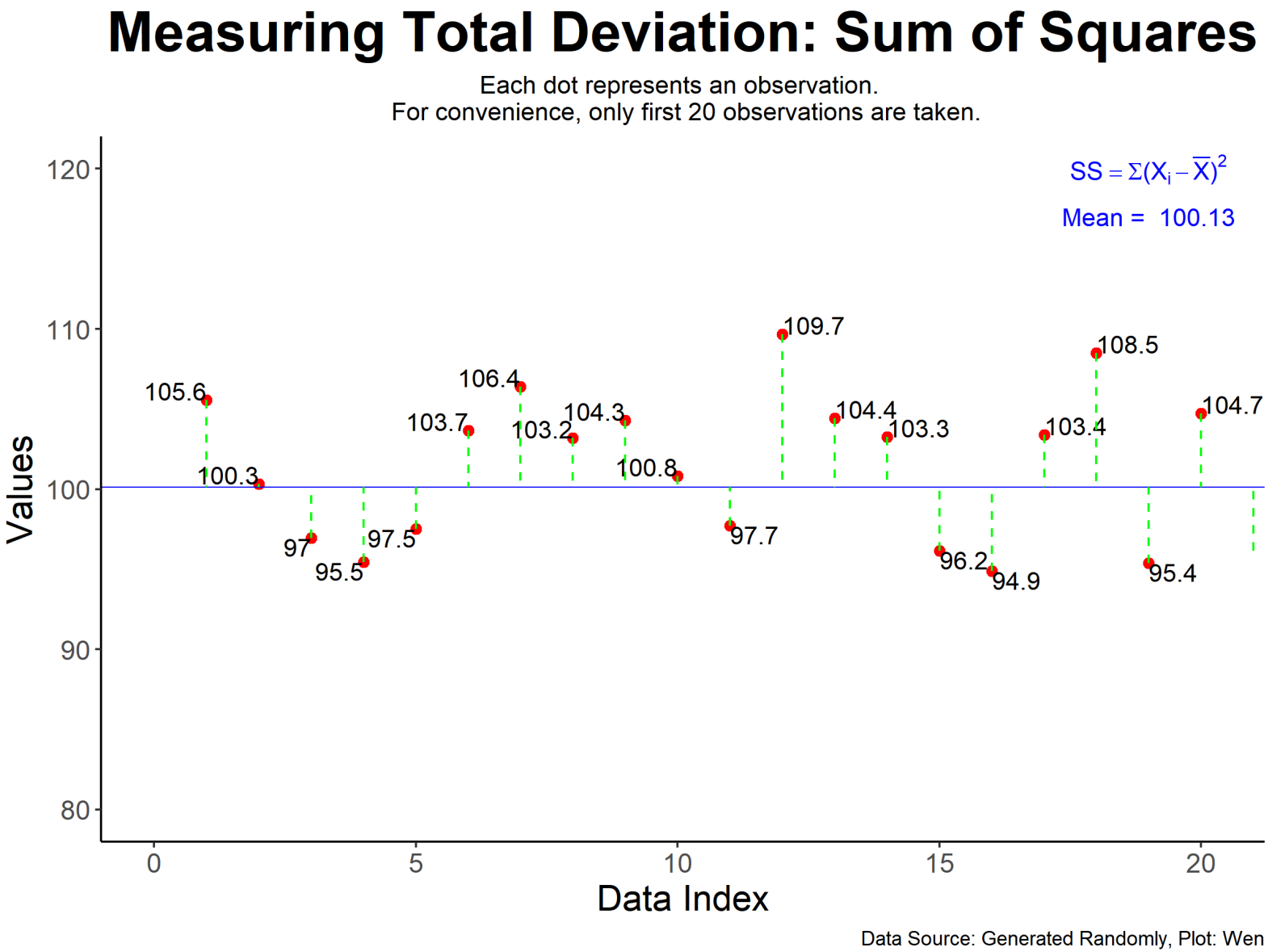
Its numerator is some kind of calculation based on μ:

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

Its denominator divides the quantity above by the following number of chucks:

$$N$$

Visualizing the Sum of Squares



Thinking about the Variance and Standard Deviation

- Variance? Variation? Variability? What is the difference?
- What does the variance intended to measure? The extent of deviation of each point with respect to the mean of the data.
- Why was the deviation squared before it was added together? $\sum_{i=1}^{N}(x_i - \mu)^2$
- By square rooting the variance, the standard deviation is much easier to interpret because it maintains the same unit with the original data.
- From the way they are calculated, they can never take negative values. When there is no variation in the data set, the variance will be 0, so is the s.d.

Check Your Understanding (1/2)

The sum of deviations of the individual data values from their mean is

- a. always greater than zero
- b. always less than zero
- c. sometimes greater than and sometimes less than zero, depending on the data values
- d. always equal to zero

Check Your Understanding (2/2)

Variable A has the variance of 4 and variable B has the s.d. of 2, then

- a. variable A has a greater variation
- b. variable B has a greater variation
- c. both have the same level of variation
- d. not enough information

Descriptive Statistics on Single Variable

Do you see how the descriptive statistics can be used to evaluate the data?

	Mean	Median	Range	Variance	S.D.
Regular	61.992	51.000	137.250	1,229.761	35.068
Promotional	84.290	63.420	274.360	3,777.614	61.462

The Measures of Association Between Two Variables

Oftentimes, we wish to measure and evaluate the relationship between two quantitative variables.

For example, we wish to determine the **direction** of the relationship:

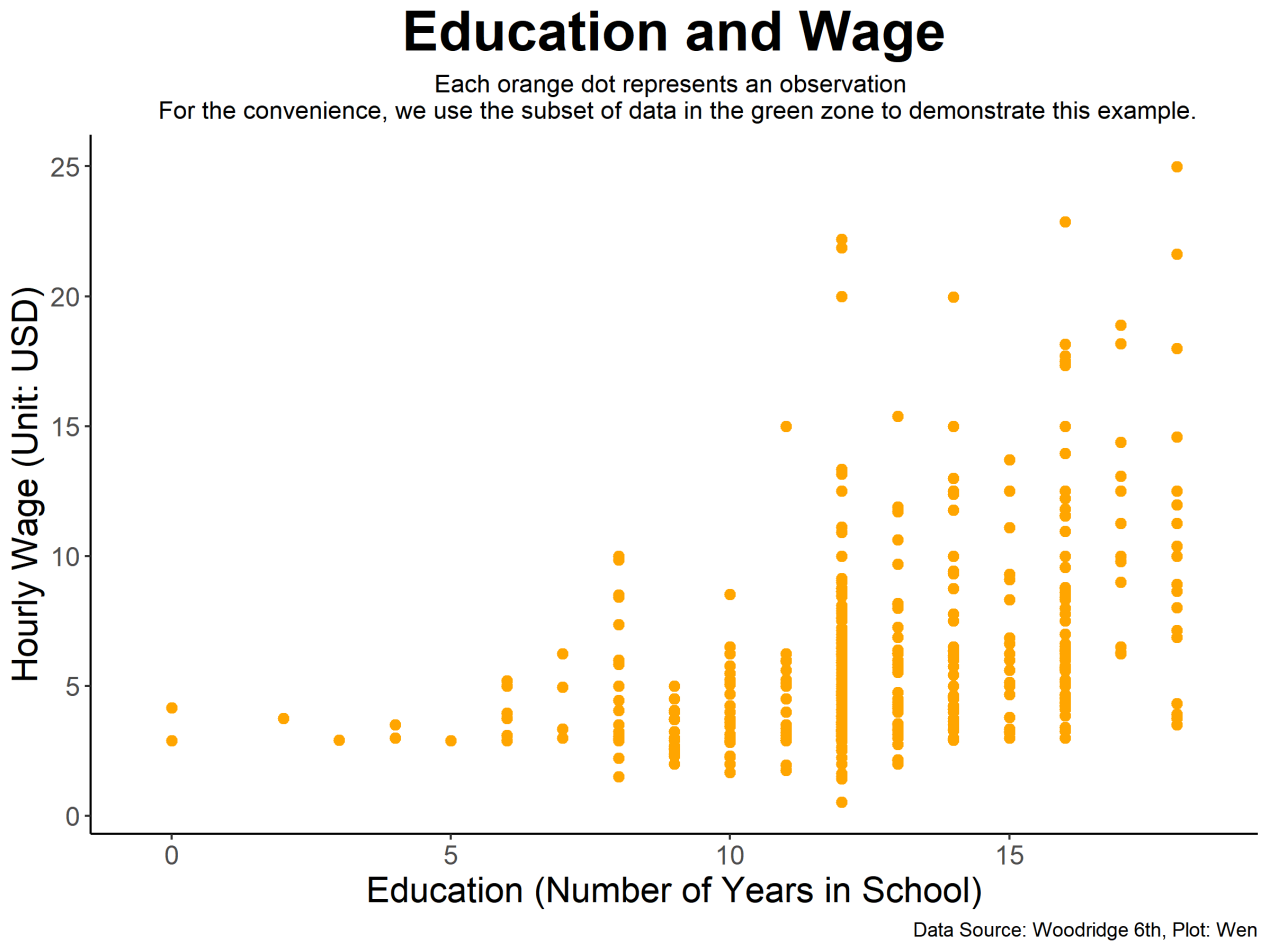
- When A increases, B also increases. (positively correlated)
- When B increases, B decreases. (negatively correlated)
- An increase in A has nothing to do with that in B. (no correlation)

We also wish to know how strong (**magnitude**) is such relationship.

- Always / almost always / in general / seldom / not at all
- Very strong / strong / moderate / weak / very weak

Using the Scatter Plot to Explore a Correlation

Do you observe a positive or a negative correlation between the education and the hourly wage?



Numerical Measures of Correlation - Covariance

Covariance

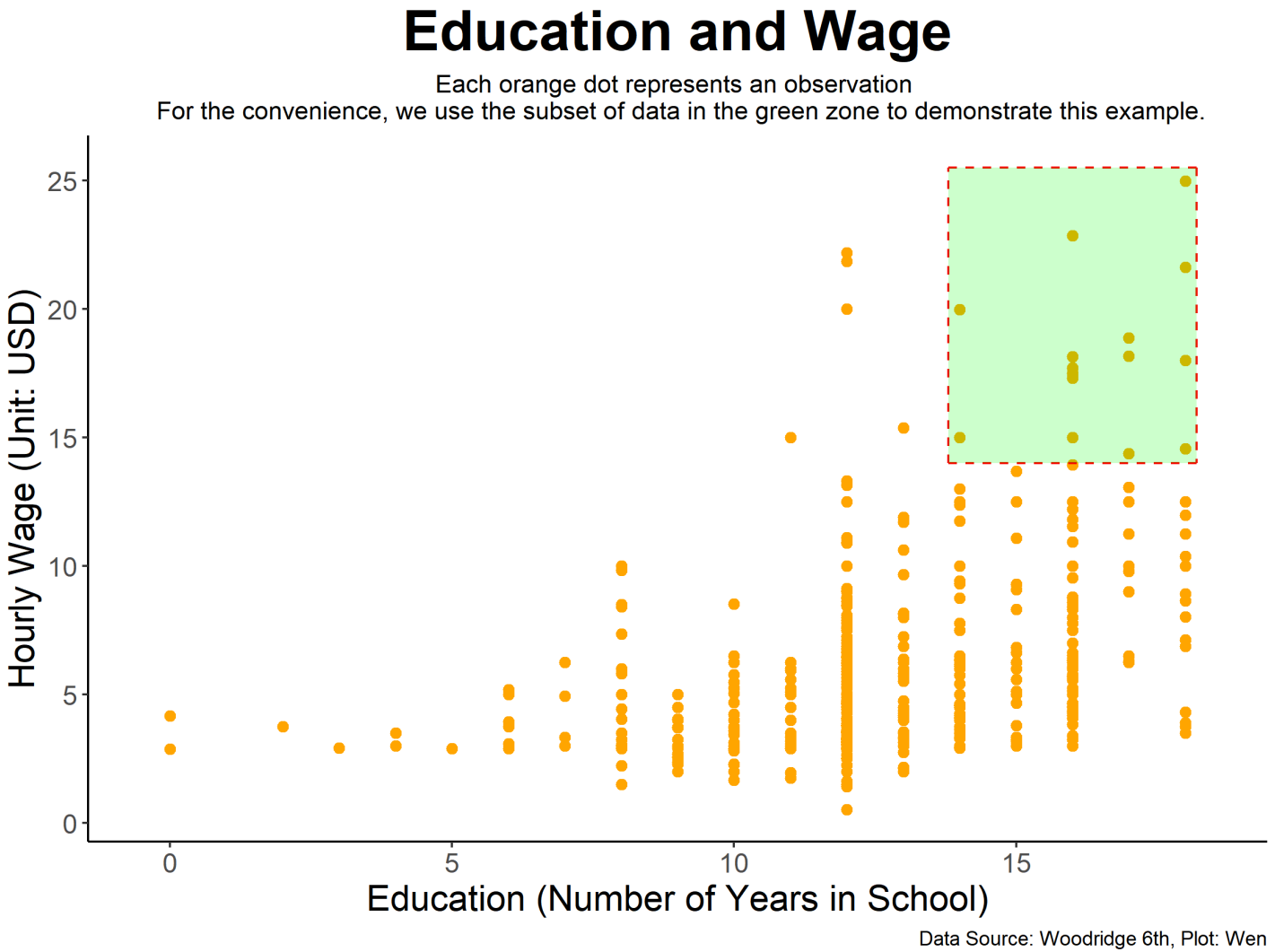
$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N} \quad \text{or}$$

$$s_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

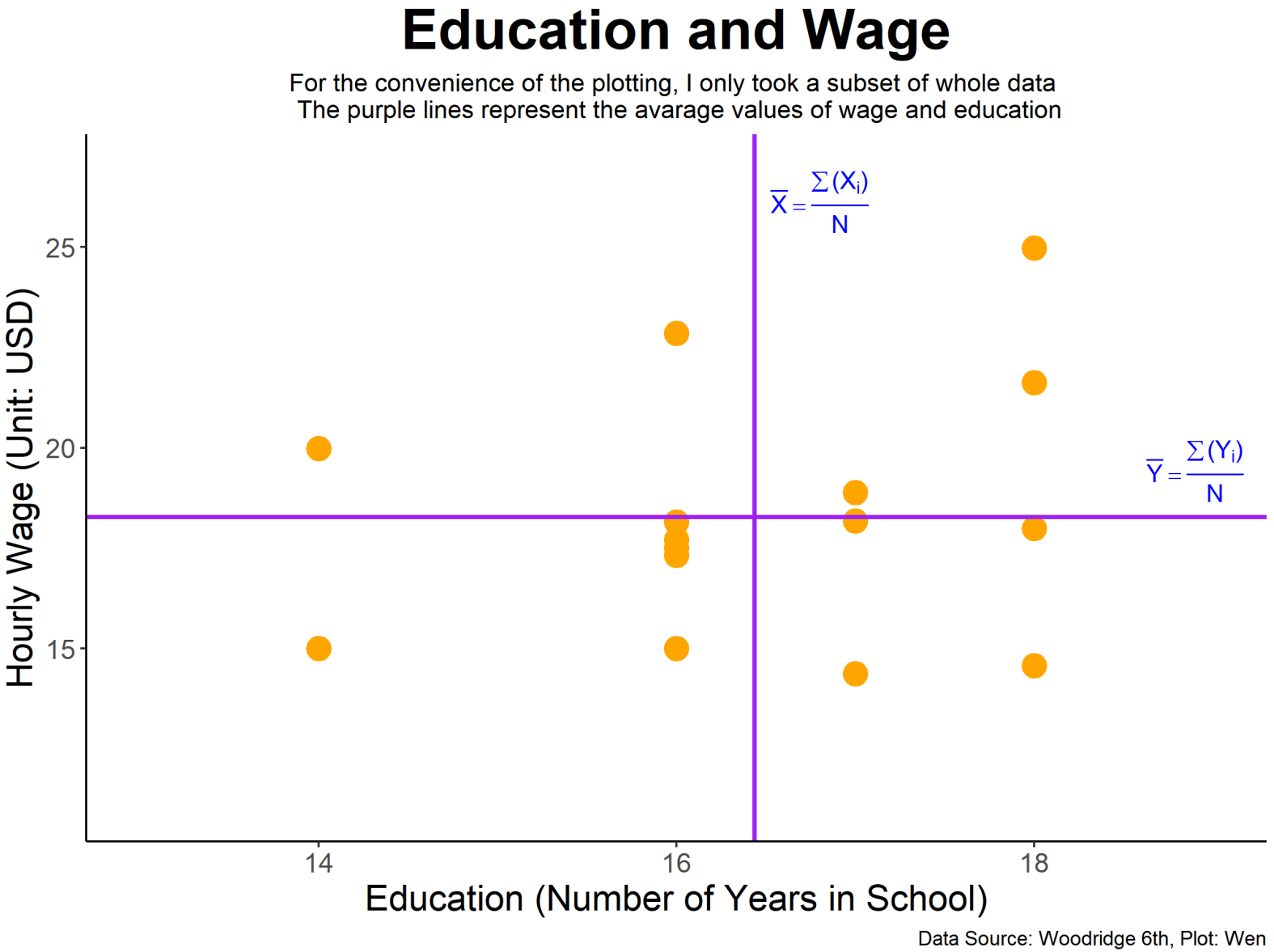
EXCEL Formula: =Covariance.P(var1_Range, var2_Range)
EXCEL Formula: =Covariance.S(var1_Range, var2_Range)

Covariance Visualized (1/3)

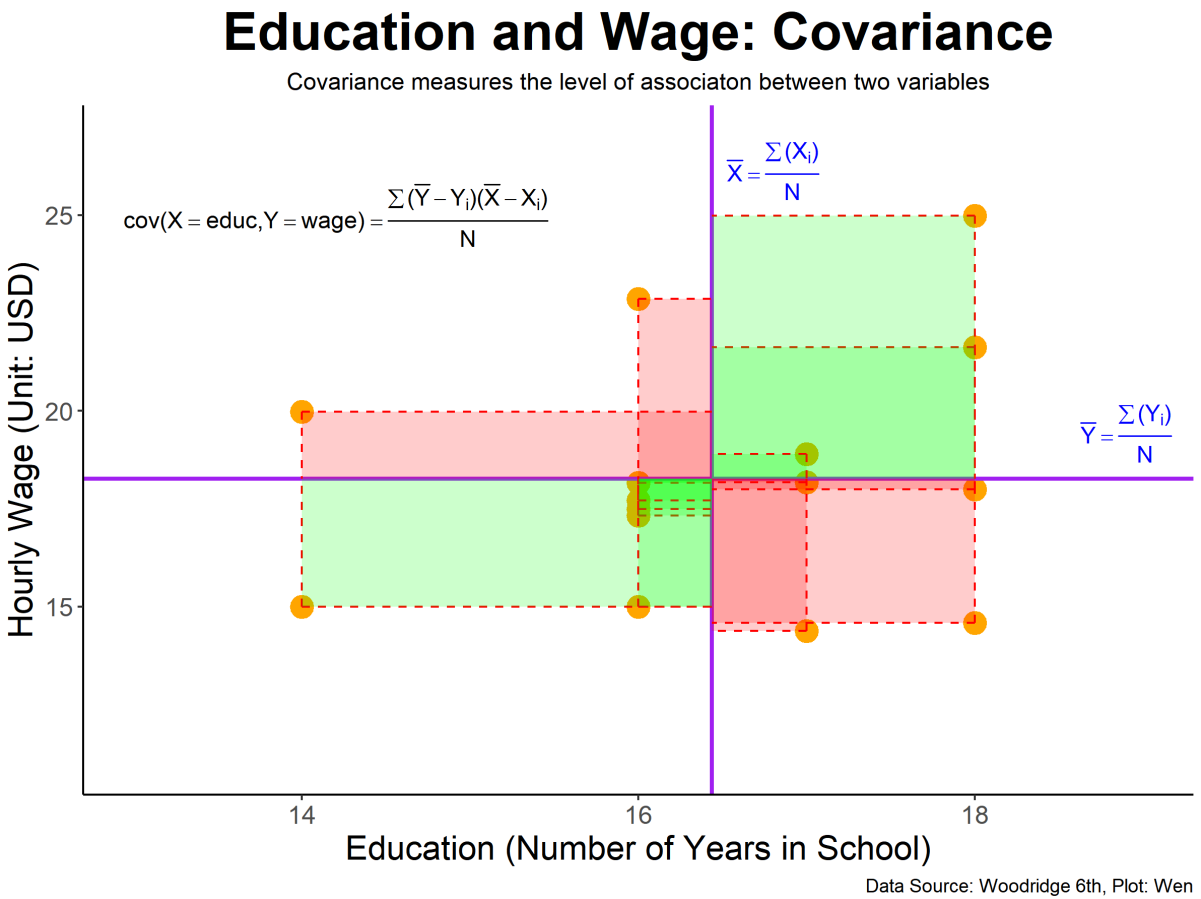
We use the subset of the dataset (the green area) to illustrate the concept of the covariance.



Covariance Visualized (2/3)



Covariance Visualized (3/3)



Theoretically, while the green rectangles will have positive dimensions, the red ones will have negative dimensions.

Interpretation of the Covariance

- If the total combined area of the green rectangles are larger than that of the red rectangles, two variables can be said *positively* correlated. (Covariance > 0)
- If the total combined area of the red rectangles are larger than that of the green rectangles, two variables can be said *negatively* correlated. (Covariance < 0)
- If the total combined area of the rectangles from each diagonal are equal, then, two variables are said to have no correlation. (Covariance = 0)

A Problem with the Covariance

- A larger positive (or negative) value for the covariance indicates a strong positive (or negative) relationship.
- However, one problem with using covariance as the strength of the relationship is that the value of the covariance depends on the units of measurement for x and y.
- Consider the following two cases, which will have a higher covariance?

	Variable A	Variable B
Case 1	Height in Feet	Height in Feet
Case 2	Height in Inch	Height in Inch

Therefore, the interpretation of the covariance should be cautious. We can also consider using a standardized measure ρ

Correlation Coefficient ρ

Correlation Coefficient (standardized covariance)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{or} \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

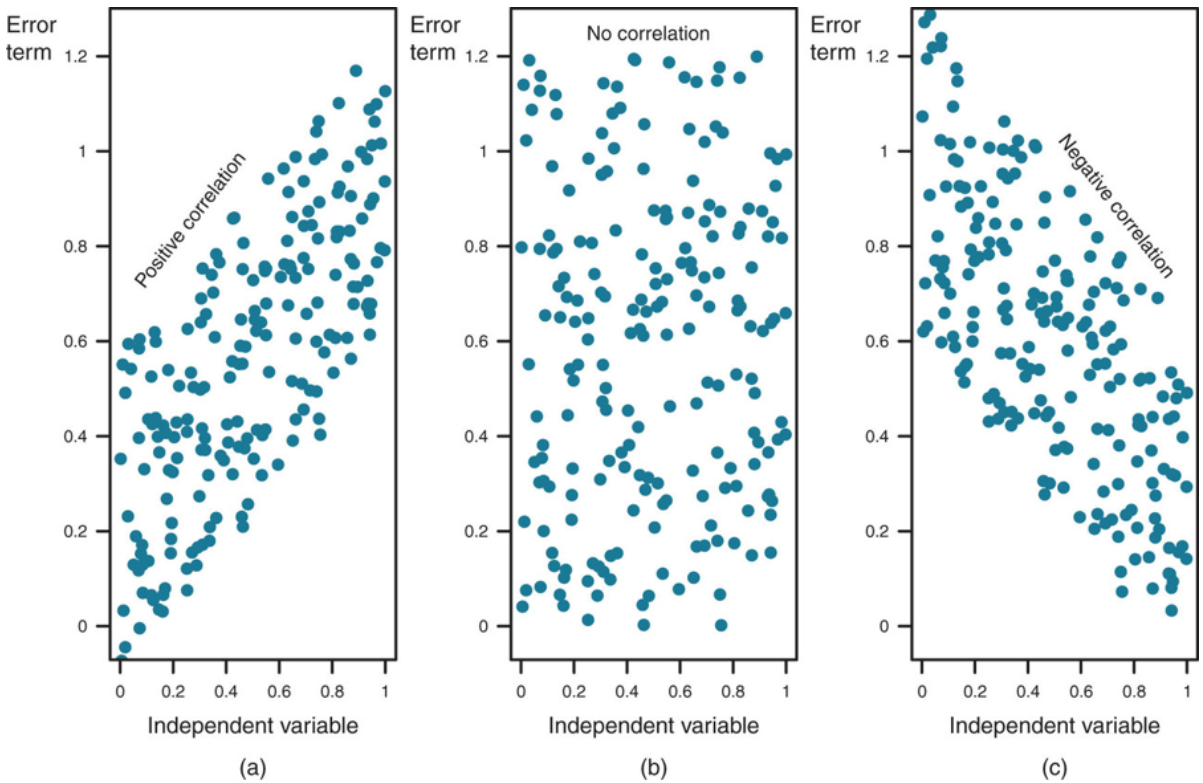
Excel Formula: =Correl(var1_Range, var2_Range)

- This measure will be ranging between -1.00 and 1.00.

Size of the Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
0 to .30 (0 to -.30)	Negligible correlation

Check Your Understanding (1/3)

Q: How would you (verbally) describe the level of association between the two variables in each of the following cases?



Check Your Understanding (2/3)

Guessing the correlation coefficient

- Click to play this simulation game
- (Another version)

Check Your Understanding (3/3)

A measure of linear association between two quantitative variables is the

- a. variance
- b. coefficient of variation
- c. correlation coefficient
- d. standard deviation

Check Points Before Move On

- From data to a frequency distribution to a histogram to a fitting line
- From a frequency distribution to a cumulative relative frequency distribution
- Mean and median as the measure of the central tendency
- Variance and standard deviation as the measure of the spread
- Two areas of assessment in evaluating the linear association between two variables.
- Covariance and correlation coefficient as the measure of the linear association between two variables

Other Measures of Central Tendency

Weighted Mean

$$\mu = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \quad \text{or} \quad \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$(w_i = \text{the weight of the observation } i\,)$$

In the mathematical mean, the weight of the each item is considered the same as $1/N$. But in the weighted mean the weight is different for each observation.

Excel Formula: =SUMPRODUCT(x_i, w_i)/SUM(w_i)

Mode

The most frequent observations

Excel Formula: =Mode.mult(Range) or *=Mode.sngl(Range)*

Check Your Understanding

A college sophomore has completed so far 3 courses. He received an A for a 5 credit hour course, a B for a 4 credit hour course, and a C for a 3 credit hour course. What is his GPA?

- a. 2.83
- b. 3.50
- c. 3.00
- d. 3.17

Please show the steps in Excel.

Other Measures of Variability using Standard Deviation

Coefficient of Variation or Relative Standard Deviation

$$\frac{\sigma}{\mu} \quad \text{or} \quad \frac{s}{\bar{x}}$$

Advantage

- Understand the variations in relation to the mean.
- Making the comparison of the variations between variables with different unit and size.

Disadvantage

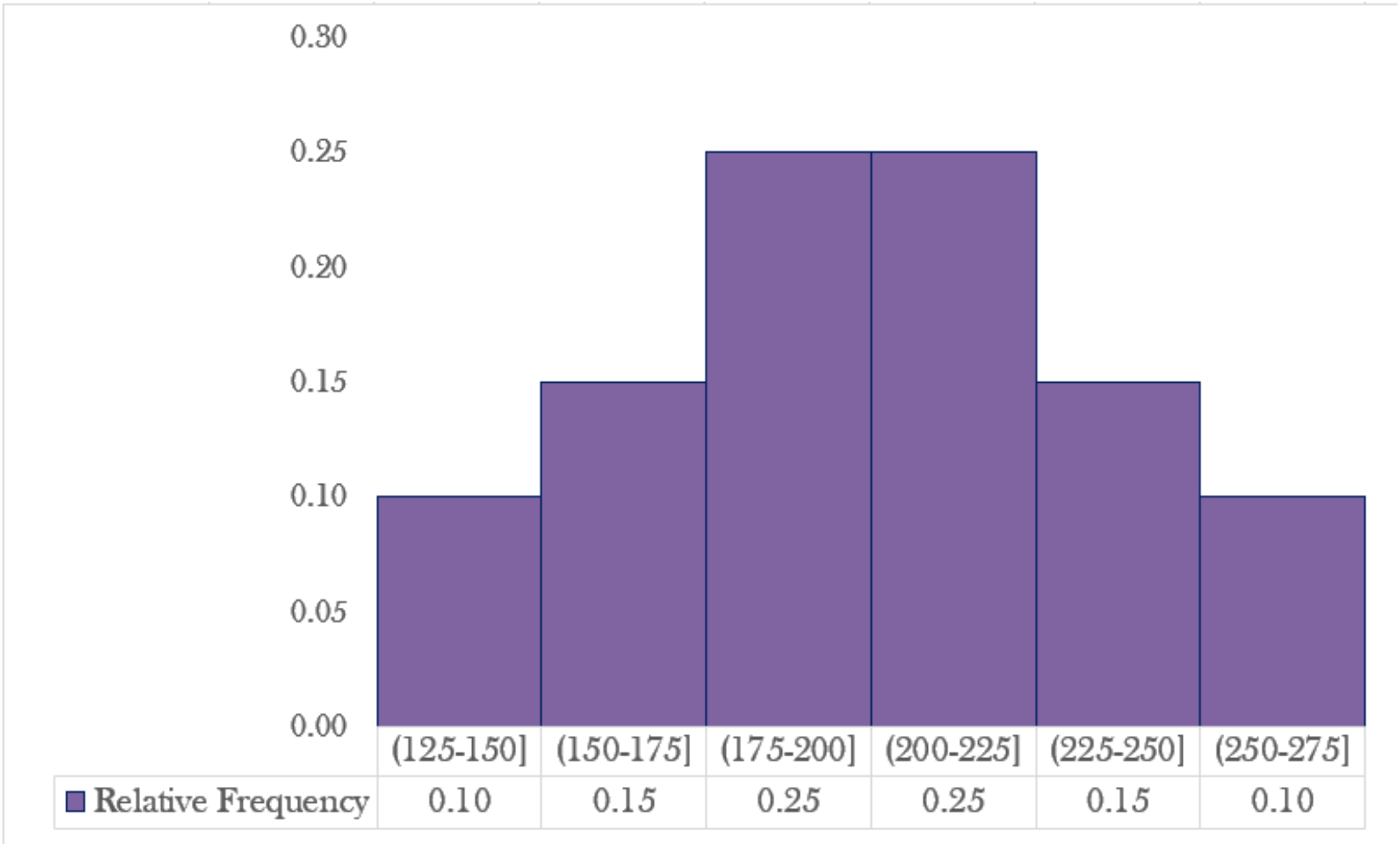
- Cannot use it when the mean is zero.

Check Your Understanding

Please compare the level of variations among the following stock price data. (Average stock price observed sequentially.) Explain why the use of coefficient of variation is more useful in this case.

	A	B	C	D	E	F	G
1	ID	A1	B1	C1	A2	B2	C2
2	1	70	66	71	209787	196240	183883
3	2	69	64	74	206711	191961	221979
4	3	71	65	70	212750	194933	209911
5	4	75	65	74	224929	194900	221810
6	5	71	67	73	212782	200962	218906
7	6	73	65	77	218853	194842	230805
8	7	73	64	75	218950	191756	224841
9	8	72	68	74	215988	176683	221898
10	9	66	65	68	197777	194844	203949
11	10	68	69	73	203725	206855	218884
12	11	68	66	75	203890	197814	224725
13	12	72	66	71	215742	197803	212842
14	13	68	66	76	203797	197911	227813
15	14	75	65	75	224847	194878	224893
16	15	69	68	72	206993	203957	215832
17	16	66	69	73	197875	206946	218777
18	17	68	68	74	203727	203865	221945
19	18	74	66	74	221789	197701	221940
20	19	73	68	73	218909	203792	218918
21	20	72	68	71	215824	203910	212903

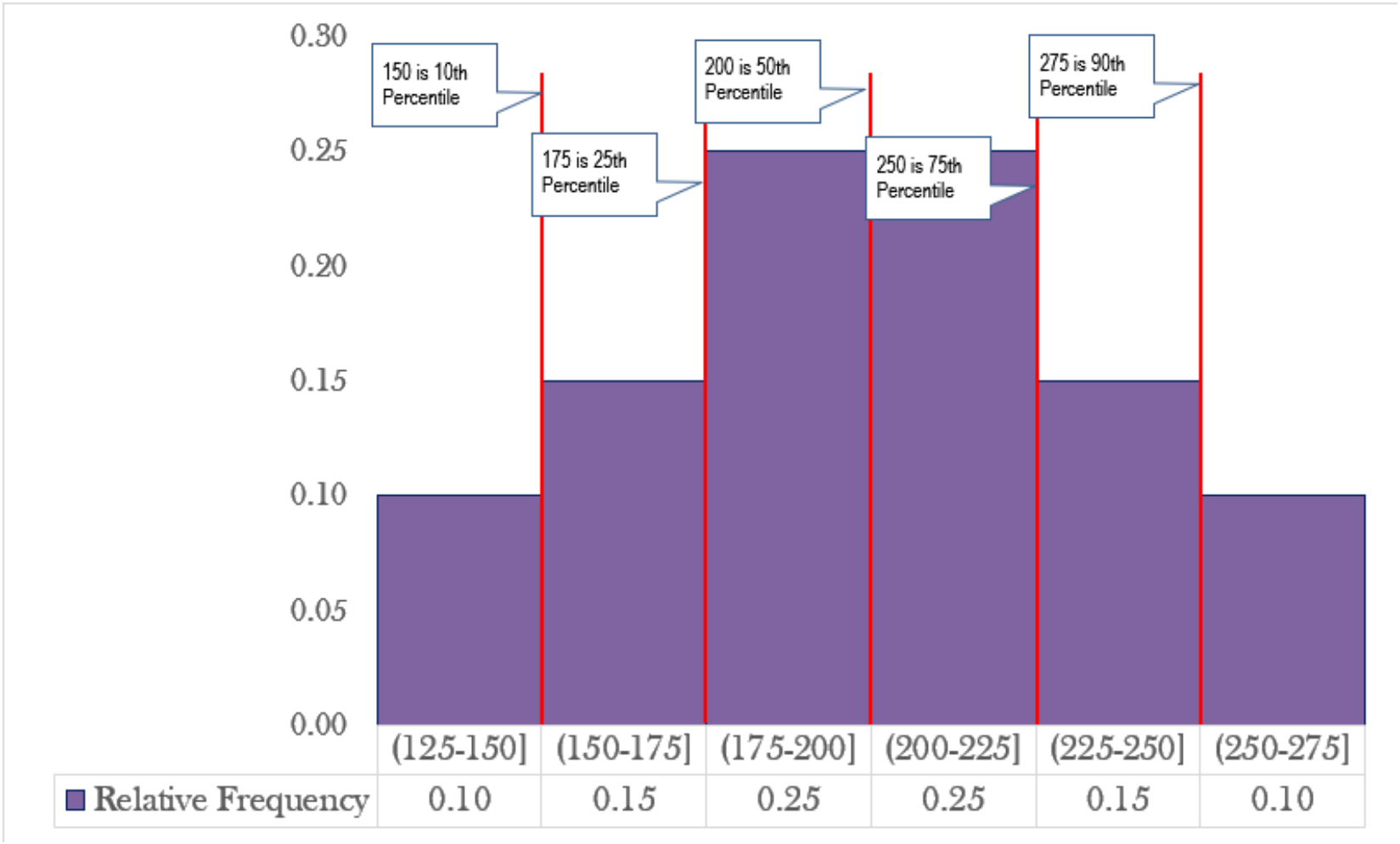
The Measures of Location: Percentiles and Quartiles (1/5)



	(125-150]	(150-175]	(175-200]	(200-225]	(225-250]	(250-275]
Relative Frequency	0.10	0.15	0.25	0.25	0.15	0.10

What do you see and what can you discuss about the data/histogram?

The Measures of Location: Percentiles and Quartiles (2/5)



	(125-150]	(150-175]	(175-200]	(200-225]	(225-250]	(250-275]
Relative Frequency	0.10	0.15	0.25	0.25	0.15	0.10

The Measures of Location: Percentiles and Quartiles (3/5)

Steps to find a **Percentile**

- Step 1: identify the location of the pth percentile

$$L_p = \frac{p}{100}(n+1)$$

- Step 2: the pth percentile can be identified through a series of well defined steps (aka. algorithm) using the L_p identified above.
- (I will only ask you to find a percentile using the Excel formula.)

Excel Formula: =Percentile.exc(DataRange, Percentile)

The Measures of Location: Percentiles and Quartiles (4/5)

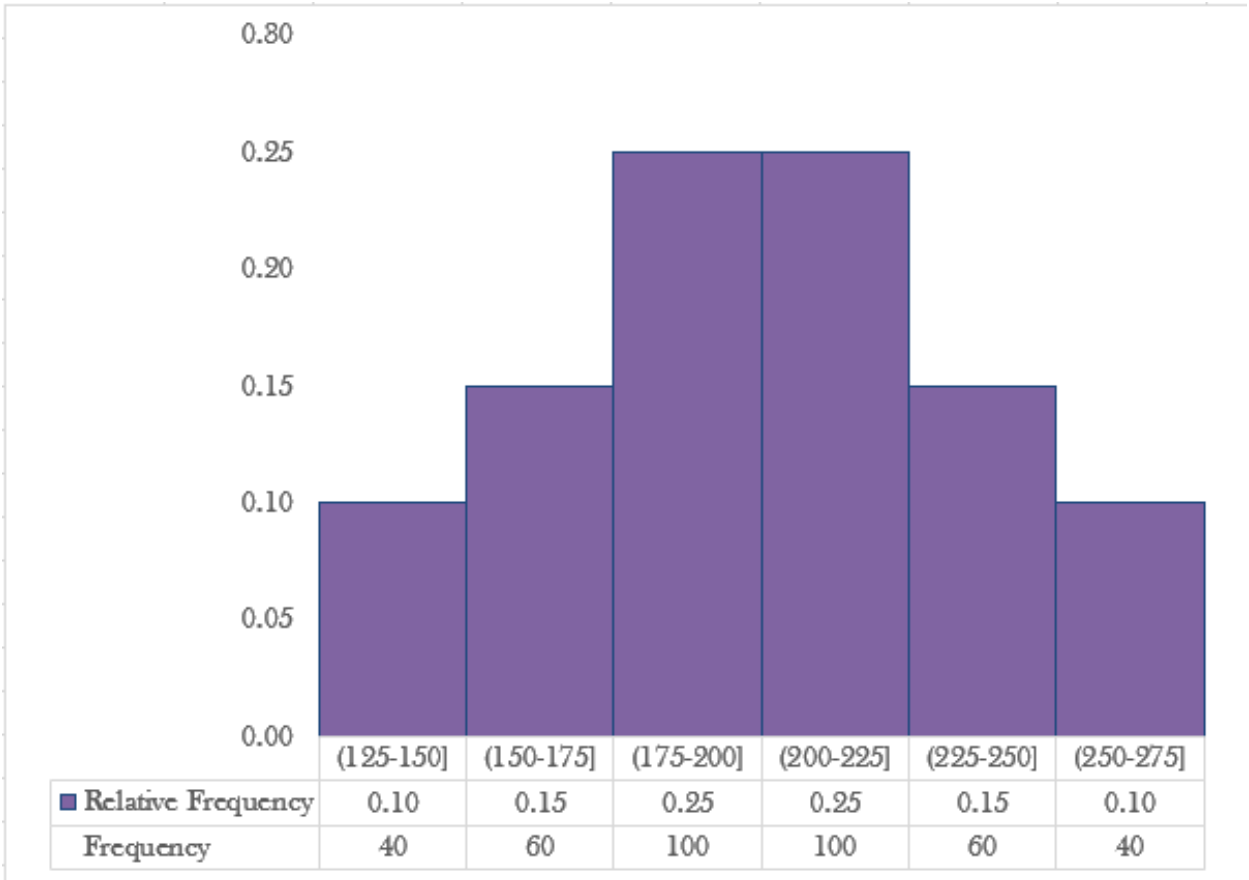
Quartiles divide the entire data into four equal size groups.

- 25th Percentiles (or Q_1 = **first quartile**)
- 50th Percentiles (or Q_2 = **second quartile**, or the **median**)
- 75th Percentiles (or Q_3 = **third quartile**)

Excel Formula: =Quartile.exc(DataRange, Quart)

(The *quintiles* and the *deciles* are similar concepts.)

The Measures of Location: Percentiles and Quartiles (5/5)



	(125-150]	(150-175]	(175-200]	(200-225]	(225-250]	(250-275]
Relative Frequency	0.10	0.15	0.25	0.25	0.15	0.10
Frequency	40	60	100	100	60	40

Can you find all three quartiles from this distribution?

Check Your Understanding

A data was collected about starting salaries of all graduates from a small local university who were hired within three months after graduation. The starting salary of John, a graduate from the business college, is at the 77% percentile in the data. It means that approximately

- a. 77% of the graduates were offered more than John and 23% less than John
- b. 77% of the graduates were offered less than John and 23% more than John
- c. John's salary is the average of the top 77% of the starting salaries
- d. John's salary is the average of the bottom 77% of the starting salaries

One Important Application of the Quartiles - Boxplot(1/4)

Inter-Quartile Range (IQR) is an *interval* where the mid 50 percent of the data points are concentrated.

$$IQR = Q_3 - Q_1$$

It excludes the smallest 25% and largest 25% of the data.

Check Your Understanding (1/2)

The median of a data set is

- a. the second quartile
- b. the 50th percentile
- c. the middle value when the number of values is odd and they are arranged in ascending order
- d. the average of the two middle values when the number of values is even and they are arranged in ascending order
- e. all the above

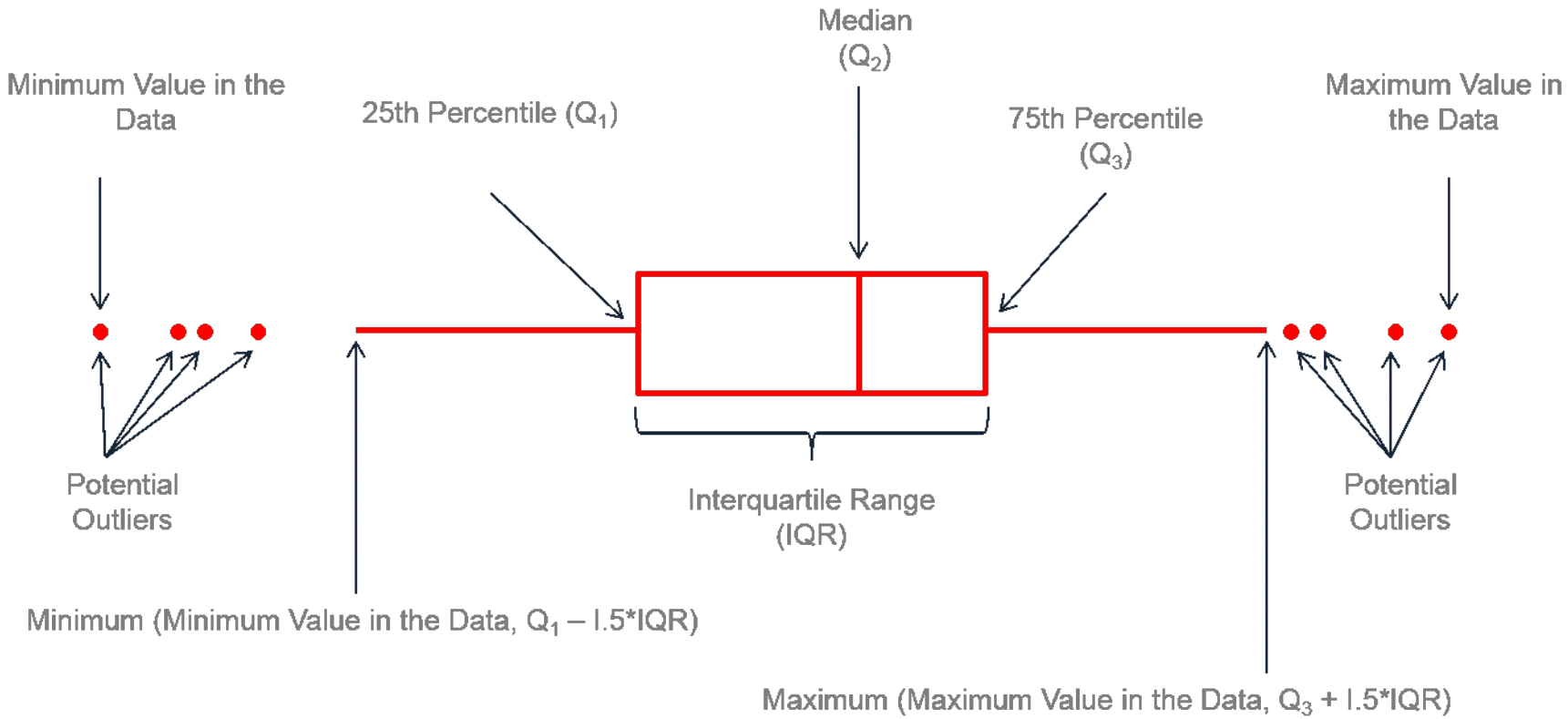
Check Your Understanding (2/2)

The interquartile range is used as a measure of variability to overcome what disadvantage of the range?

- a. the range is typically too short
- b. the range is difficult to compute
- c. the range is influenced too much by extreme values
- d. the range is never negative

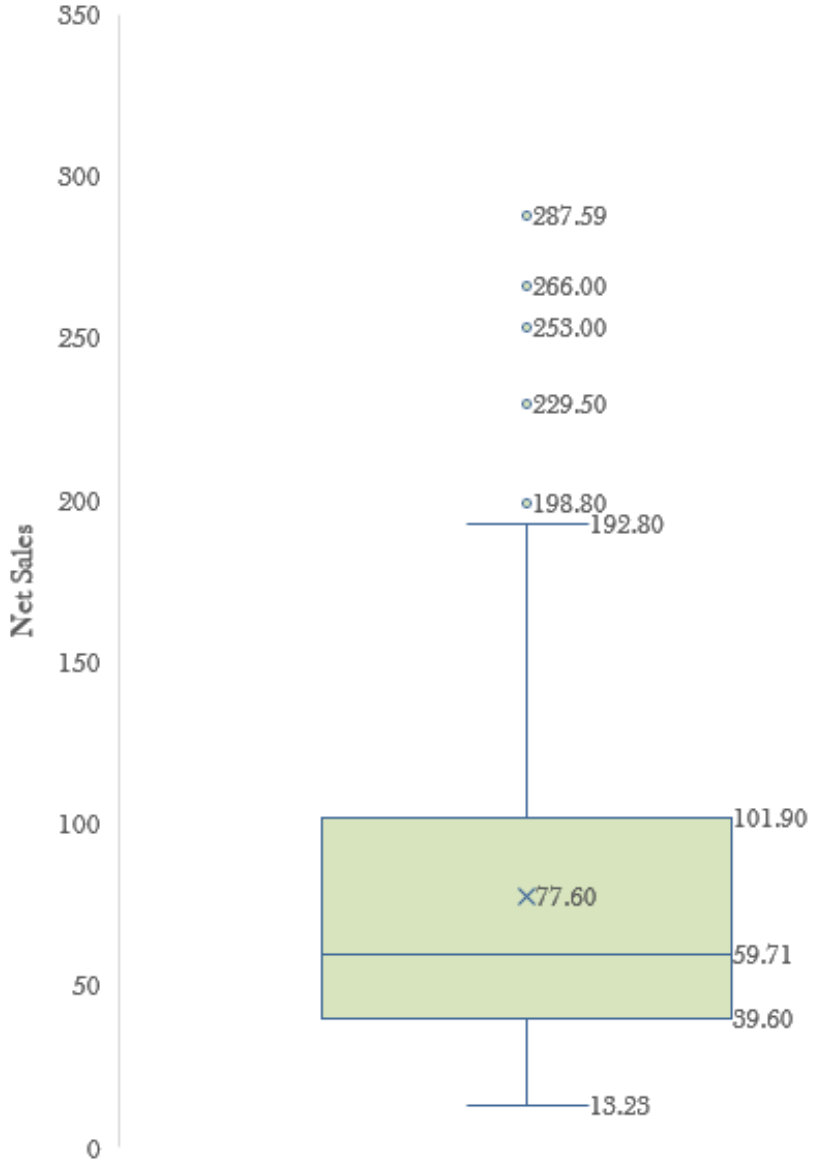
One Important Application of the Quartiles - Boxplot(2/4)

The Anatomy of a Boxplot



Reference: https://www.r-graph-gallery.com/boxplot.html

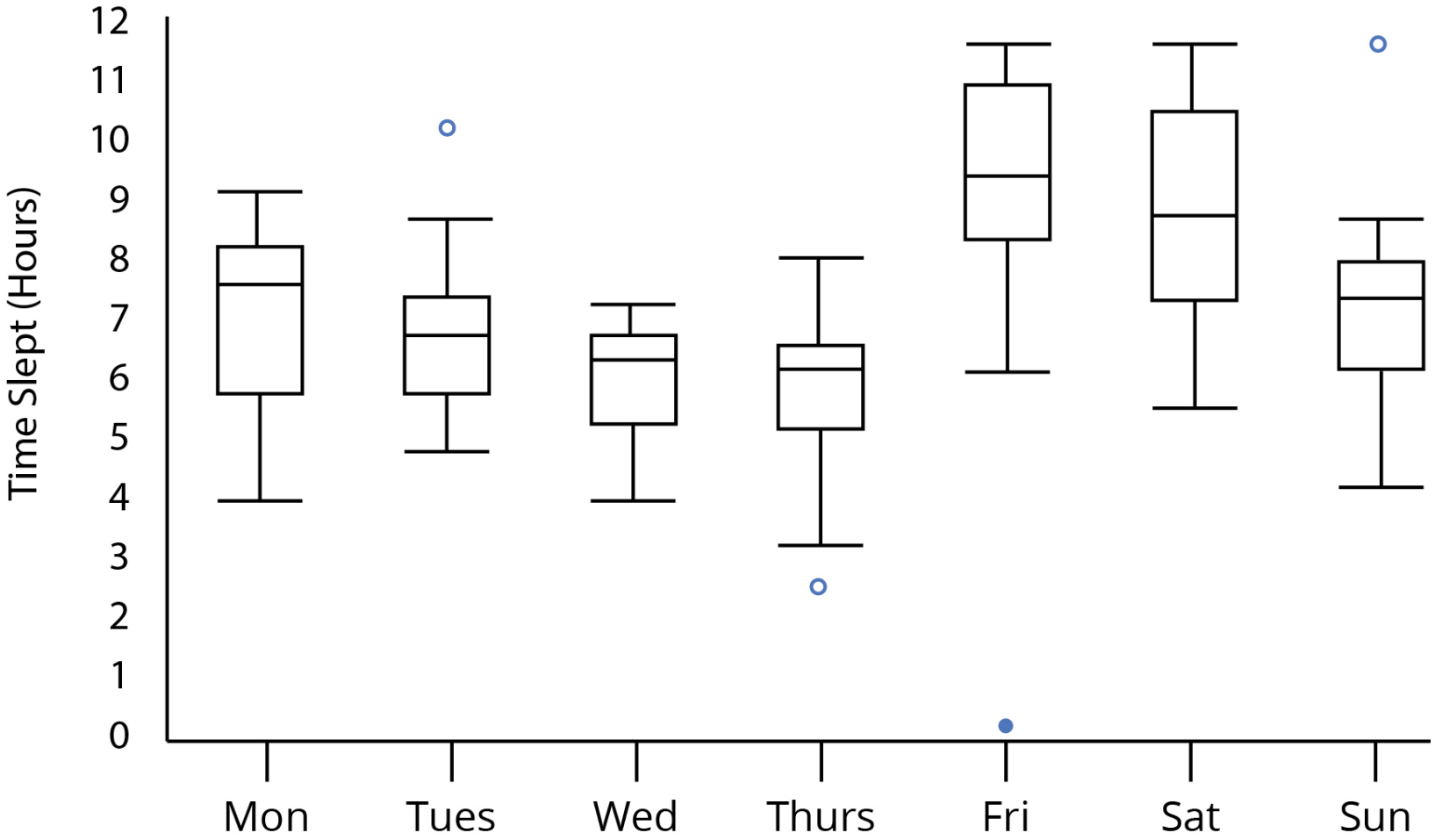
One Important Application of the Quartiles - Boxplot(3/4)



Five Number Summary	
Minimum	13.23
1st Quartile	39.60
2nd Quartile	59.71
3rd Quartile	101.90
Maximum	287.59
Boxplot Boundaries	
IQR	62.30
Whisker Top	195.35
Whisker Bottor	-53.85
Summary Statistics	
Mean	77.60
Median	59.71
Range	274.36
Variance	3098.59
S.D.	55.66

One Important Application of the Quartiles - Boxplot(4/4)

Boxplot makes it easier to compare multiple groups



The Measures of Location Using Standard Deviation (1/4)

Standard deviation as the ruler: **z-score** (standardized of x_i)

$$z_i = \frac{x_i - \mu}{\sigma} \quad \text{or} \quad z_i = \frac{x_i - \bar{x}}{s}$$

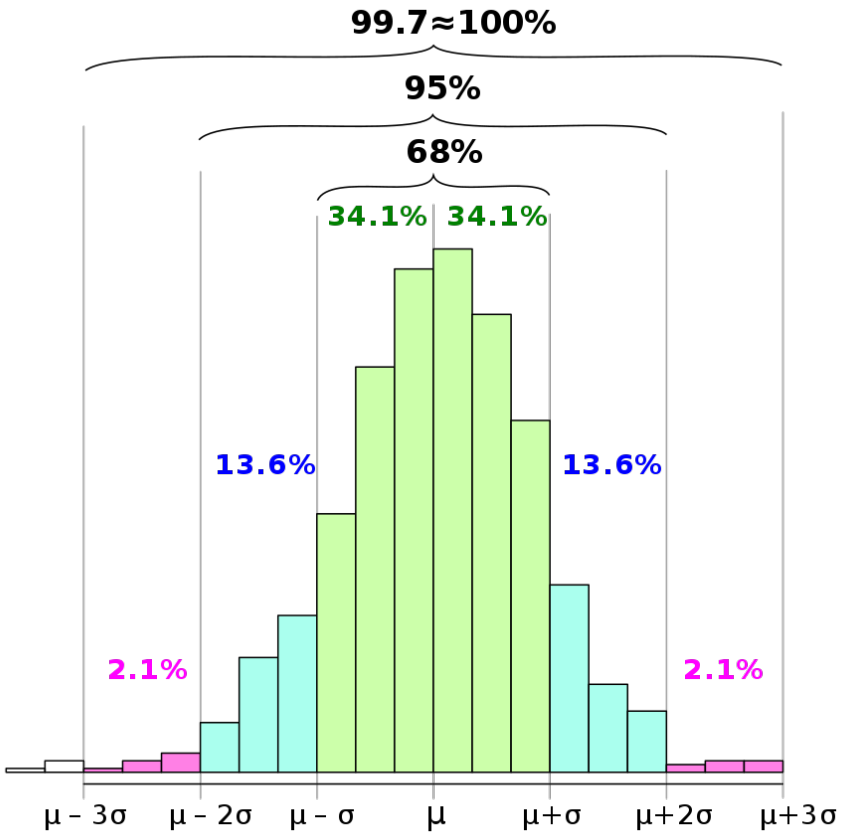
Excel Formula: =Standardize(x, mu, sd)

For example, in the *net sales* variable,

- it has $\bar{x} = 77.60$, $s = 55.66$

Value of x	Steps to z	z-score	Interpretation
40.78	(40.78 - 77.60)/55.66	-.66	.66 s.d. less than the average.
77.60	(77.60 - 77.60)/55.66	.00	The same with the average
133.26	(133.26 - 77.60)/55.66	1.00	1 s.d. larger than the average.

The Measures of Location Using Standard Deviation (2/4)



The **empirical rule** shows the *approximate* fraction of observations concentrated within one, two, and three s.d. from the mean of a normal distribution.

The Measures of Location Using Standard Deviation (3/4)

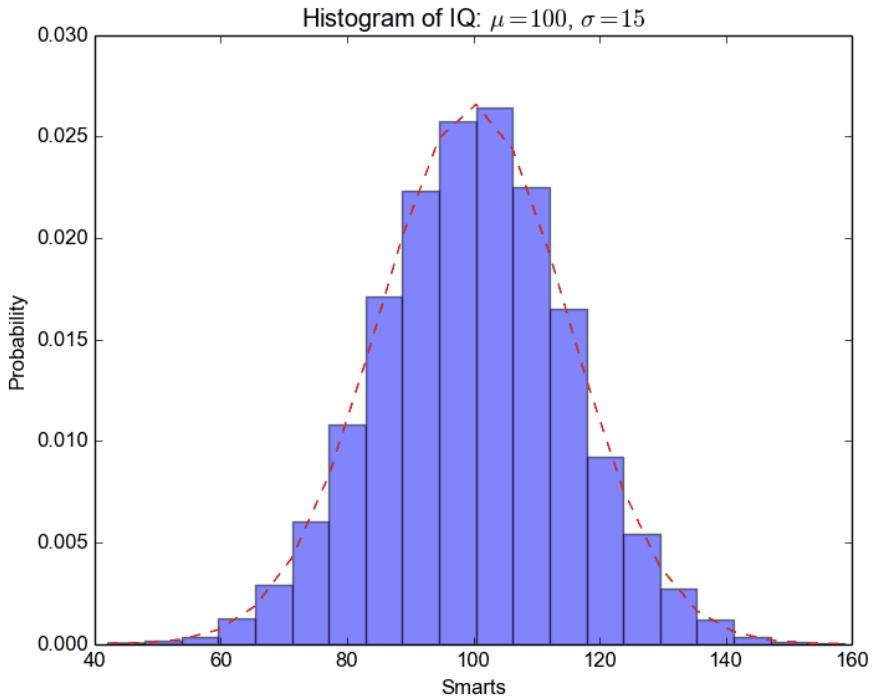
Detection of the **Outliers**

- Since the empirical rule tells that about 99.7% of the observations are concentrated within the 3 s.d. distances.
- Outside of the boundaries, there will be only about 0.3% observations.
- Therefore, those fell outside are considered rare occurrences.
- So they are considered "outliers".
- Outliers will have z-score > 3 or z-score < -3.

The Measures of Location Using Standard Deviation (4/4)

- This usage of standard deviation, mean, and the distributional characteristics are extremely important.
- The standardized value of x, or the z-score,
  - tells whether it is larger than the average
  - tells how common the value is when compared with other observations in the data
  - can be easily related to the percentile. For example,
    - If z-score $= -2$, then the number is likely a 16 percentile.
    - If z-score $= 2$, then the number is likely a 98 percentile.

Histogram and the Fit Line (1/2)



The fit line that approximates the histogram is defined as:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Histogram and the Fit Line (2/2)

The comparisons of Excel solutions on finding the percentile and the fraction

	Percentile	Fraction before X
Histogram	=Percentile.exc(Range, p)	Many steps...
Fit Line	=Norm.inv(p, mean, sd)	=Norm.dist(x, mean, sd, TRUE)
Fit Line (Z)	=Norm.s.inv(p)	=Norm.dist(z, TRUE)

III. Introduction to Probability

Why We Learn Probability (1/2)

In many cases, to make a decision, we consider many "variables", or many factors involved in the decision making process.

- $\textit{Budget Needed} = \textit{Planned Spending} - \textit{Cash at Hand}$
- $\textit{Inventory to Prepare} = \textit{Upcoming Demand} - \textit{On-hand Inventory}$

The quantities such as *Planned Spending* and *Upcoming Demand* are considered **probabilistic**, because we cannot predict those with accuracy.

The quantities such as *Cash at Hand* and *On-hand Inventory* are considered **deterministic**, because these are quantities we know and/or can exercise control with a great certainties.

Why We Learn Probability (2/2)

For a probabilistic quantity, if available, we can consider the various possible outcomes with the probabilities associated with each outcome.

Demand	Probability
25,000	10%
26,000	15%
27,000	25%
28,000	25%
29,000	15%
30,000	10%
Total	**100%**

A weighted average can help us determine a most likely outcome:

$$Expected\ Demand = \sum Demand_i \cdot Probability_i = 27,500$$

In the long run, assuming the probability assigned to each demand scenario does not change, this quantity will be the most likely demand.

Defining Probability

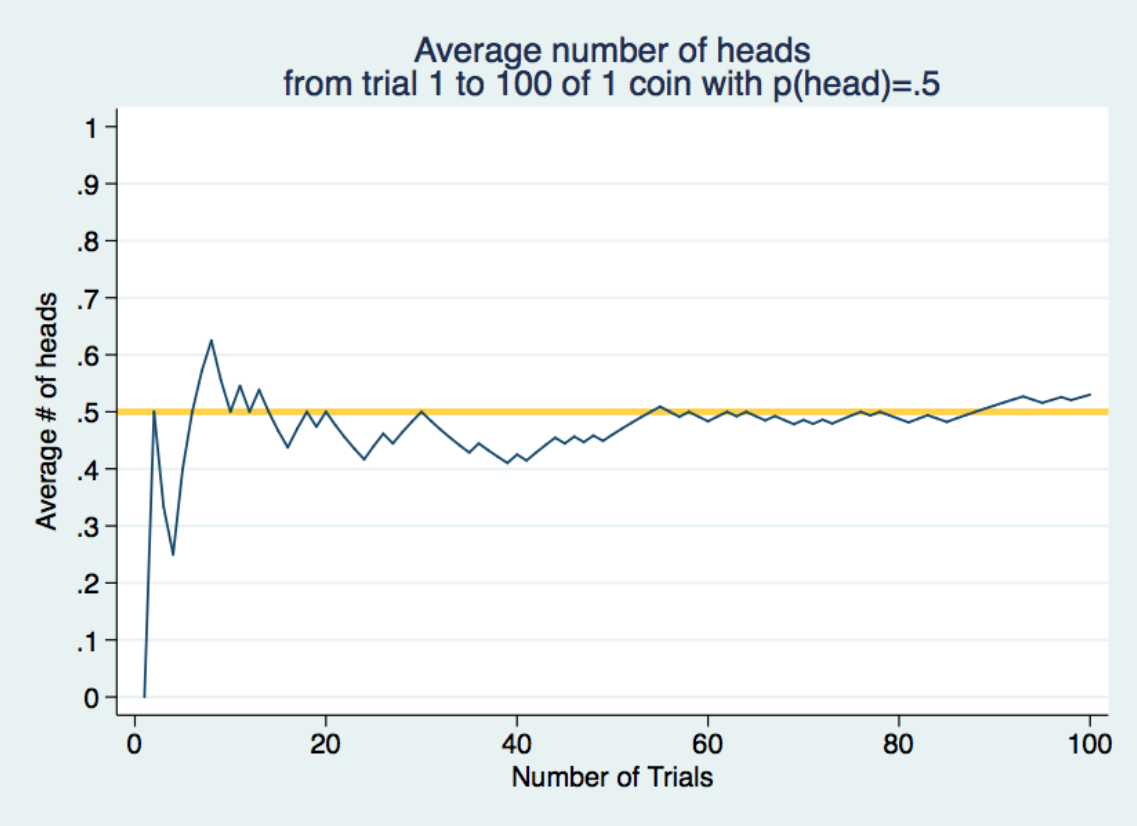
Probability Definition

Probability is the quantitative expression of the chance that an event will occur.

- The probability of an event A is usually written as $P(A)$ or $Pr(A)$.
- If $P(A) = 0$, the event has no chance of occurring.
- If $P(A) = 1$, the event will certainly happen with no doubt.
- A probability can be obtained through the observation (e.g., past record), logical analysis, and subjective determination.

More on Probability

The probability of an outcome is interpreted as the long-run proportion of the time that the outcome would occur, if the experiment were repeated indefinitely. That is, probability is **long-term relative frequency**. [1]



[1] citing from: http://www.math.wsu.edu/faculty/djohnson/

Experiments and Sample Point

- An **experiment** is a process that generates well-defined outcomes. For examples,
- An experiment outcome is called a **sample point**.

Experiment	All Sample Points
Toss a coin	Head, tail
Inspection a part	Defective, non-defective
Roll a die	1, 2, 3, 4, 5, 6
Class Attendance	0, 1, 2, 3, ... Max

Sample Space as a Set

An exhaustive set of sample points of an experiment is called a **sample space**. Therefore, a sample space is all possible experimental outcomes.

$$Payment\ method = \{Amex, Discover, Visa, Master, Others\}$$

$$Type\ of\ customer = \{Regular, Promotional\}$$

$$Number\ of\ family\ members = \{1, 2, 3, 4, 5, 6, 7, 8, above\ 8\}$$

Assigning Probabilities

The probability assigned to each experiment outcome must be between 0 and 1

$$0 \leq P(E_i) \leq 1 \text{ for all } i$$

The sum of the probability for all the sample point must equal to 1

$$\sum_{i=1}^{n} P(E_i) = 1$$

This is the same with the relative frequency distribution...

Assigning Probabilities Using Relative Frequency

The relative frequency information can be utilized as the basis of the probability. For example,

Card Type	Frequency	Relative Frequency	Probability
American Express	2	2.00%	2.00%
Discover	4	4.00%	4.00%
MasterCard	14	14.00%	14.00%
Proprietary Card	70	70.00%	70.00%
Visa	10	10.00%	10.00%
Grand Total	**100**	**100.00%**	**100.00%**

What is the probability of someone who will use a non-proprietary card for shopping? Please use the notation to show the steps.

Check Your Understanding

Consider the experiment of selecting a playing card from a deck of 52 playing cards. Each card corresponds to a sample point with a 1/52 probability.

- List the sample points in the event an ace is selected.
- List the sample points in the event a club is selected.
- List the sample points in the event a face card (jack, queen, or king) is selected.
- Find the probabilities associated with each of the events in above three questions.

Events and Their Probabilities

An **event** is a collection of sample points (or experiment outcome).

The **probability of an event** is the sum of the probability of the sample points in the event.

For example, for a six-dimensional dice, the sample space is:

$$\textit{Roll a dice} = \{1, 2, 3, 4, 5, 6\}$$

Each sample outcome has the probability of 1/6. Then, the probabilities for the following events are:

$$P(\textit{Results are even}) = P(\{2, 4, 6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

$$P(\textit{Results are odd}) = P(\{1, 3, 5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Multiphase or Multistage Experiment (1/4)

Sometimes an experiment is more complicated as it can involve multiple steps, multiple phases, or multiple con-current activities.

$$\textit{Rolling two dices} = \{(1,1),(1,2),(1,3)...(6,6)\}$$

$$\textit{Completion time of a two-phase project} = \{(2,7),(2,8),(3,7),...(4,8)\}$$

$$\textit{Color combination of two poker cards} = \{(R,B),(R,R),(B,R),(B,B)\}$$

Multiphase or Multistage Experiment (2/4)

TABLE 4.1 EXPERIMENTAL OUTCOMES (SAMPLE POINTS) FOR THE KP&L PROJECT

Completion Time (months)			
Stage 1 Design	**Stage 2 Construction**	**Notation for Experimental Outcome**	**Total Project Completion Time (months)**
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

Multiphase or Multistage Experiment (3/4)

TABLE 4.2 COMPLETION RESULTS FOR 40 KP&L PROJECTS

Completion Time (months)			
Stage 1 Design	Stage 2 Construction	Sample Point	Number of Past Projects Having These Completion Times
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
		Total	40

Multiphase or Multistage Experiment (4/4)

TABLE 4.3 PROBABILITY ASSIGNMENTS FOR THE KP&L PROJECT BASED ON THE RELATIVE FREQUENCY METHOD

Sample Point	Project Completion Time	Probability of Sample Point
(2, 6)	8 months	$P(2, 6) = 6/40 = .15$
(2, 7)	9 months	$P(2, 7) = 6/40 = .15$
(2, 8)	10 months	$P(2, 8) = 2/40 = .05$
(3, 6)	9 months	$P(3, 6) = 4/40 = .10$
(3, 7)	10 months	$P(3, 7) = 8/40 = .20$
(3, 8)	11 months	$P(3, 8) = 2/40 = .05$
(4, 6)	10 months	$P(4, 6) = 2/40 = .05$
(4, 7)	11 months	$P(4, 7) = 4/40 = .10$
(4, 8)	12 months	$P(4, 8) = 6/40 = .15$
		Total 1.00

Can you find out the followings?

$$P(\textit{Take less than 10 month}) = ?$$

$$P(\textit{Take more than 10 month}) = ?$$


Life begins at the end of your comfort zone.
Neale Donald Walsch
BrainyQuote

Complement of an Event (Or Collectively Exhaustive Events)

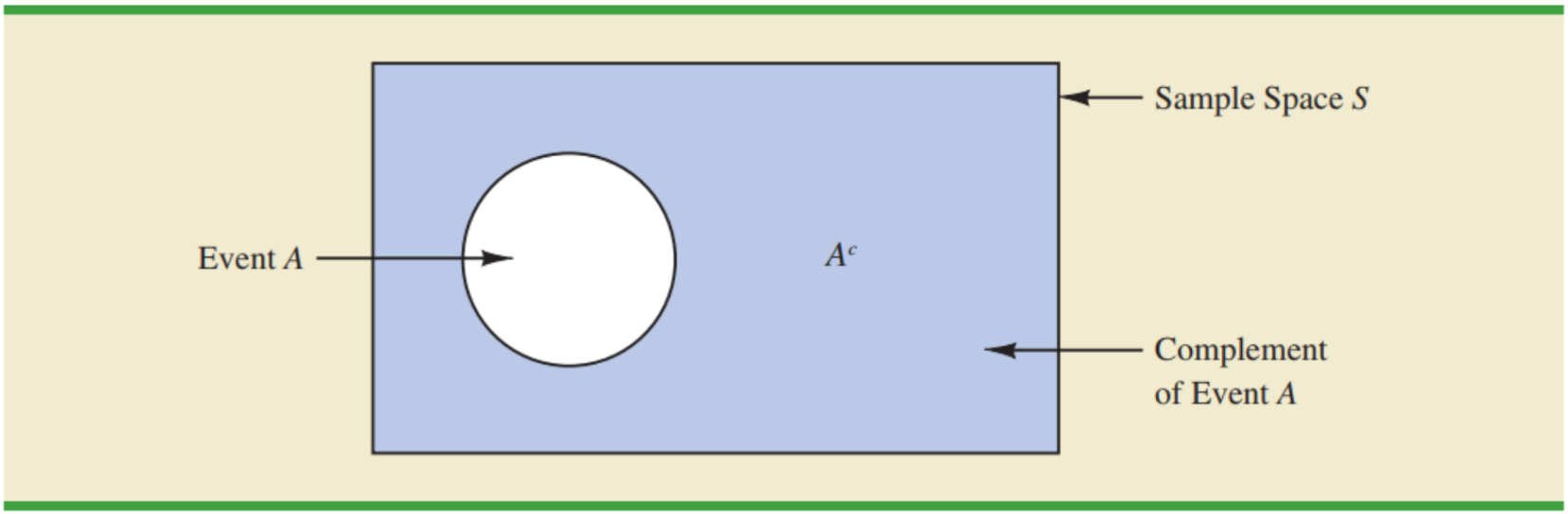
Given an event A, the **complement** of A is defined to be the event consisting of all sample points that *are not* in A. Therefore,

$$P(A) + P(A^c) = 1$$

or

$$P(A) = 1 - P(A^c)$$

FIGURE 4.4 COMPLEMENT OF EVENT *A* IS SHADED

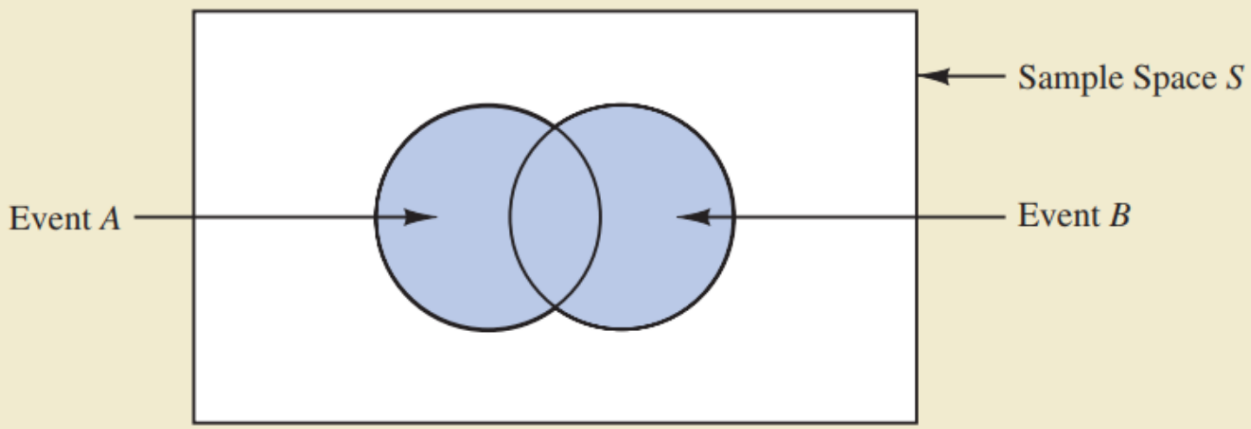


Addition Law - Union of Multiple Events

The **union** of A and B is the event containing all sample points belonging to A or B or both. Therefore, to add both,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

FIGURE 4.5 UNION OF EVENTS A AND B IS SHADED

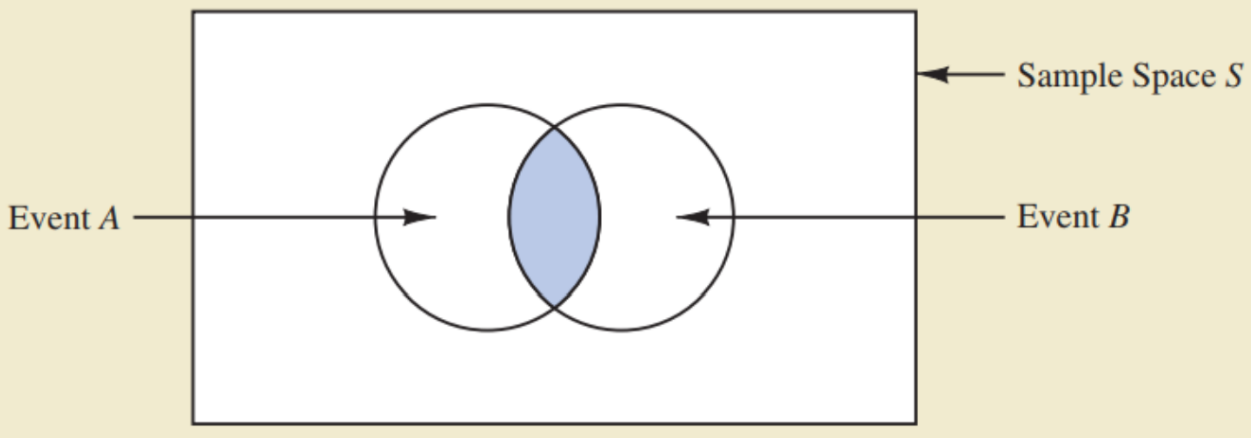


Intersection of Multiple Events

Given two events A and B, the **intersection** of A and B is the event containing the sample points belonging to both A and B.

It is denoted as $P(A \cap B)$

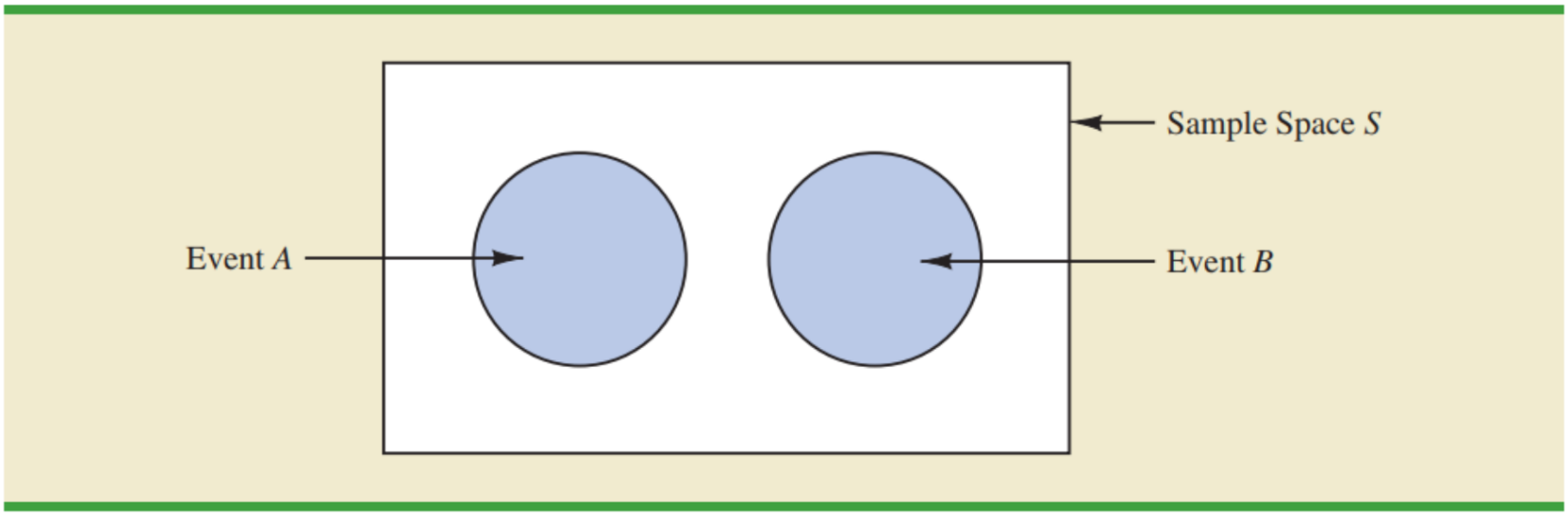
FIGURE 4.6 INTERSECTION OF EVENTS *A* AND *B* IS SHADED



Addition Law for Mutually Exclusive Events

Two events are said to be **mutually exclusive** if the events have no sample points in common. or $P(A \cap B) = 0$

FIGURE 4.7 MUTUALLY EXCLUSIVE EVENTS



Therefore, the addition for mutually exclusive events are:

$$P(A \cup B) = P(A) + P(B)$$

Check Your Understanding (1/2)

The U.S. Census Bureau provides data on the number of young adults, ages 18–24, who are living in their parents' home. Let

M = the event a male young adult is living in his parents' home
F = the event a female young adult is living in her parents' home

If we randomly select a male young adult and a female young adult, the Census Bureau data enable us to conclude P(M)=.56 and P(F)=.42 (The World Almanac, 2006). The probability that both are living in their parents' home is .24

- What is the probability at least one of the two young adults selected is living in his or her parents' home?
- What is the probability both young adults selected are living on their own (neither is living in their parents' home)?

Check Your Understanding (2/2)

Please see the following exhibit:

Promotion Status of Police Officers

	Men	Women	Total
Promoted	288	36	324
Not Promoted	672	204	876
Total	960	240	1,200

Please answer the following questions:

- Please identify at least two pairs of collectively exhaustive events.
- Find the probability of a policeman, regardless of the gender, got the promotion.
- Find the intersection event of women and promoted. Also calculate the probability.
- How to calculate the probability of someone gets promoted given this person is a men?

Marginal and Joint Probability

Illustration: Promotion Status of Police Officers

Promoted	M	F	Total
Yes	288	36	324
No	672	204	876
Total	960	240	1,200

Promoted	M	F	Total
Yes	.24	.03	.27
No	.56	.17	.73
Total	.80	.20	1.00

Marginal Probability is the probability of the event itself. For example,

$$P(Yes) = .27 \quad \text{or} \quad P(No) = .73$$

$$P(M) = .80 \quad \text{or} \quad P(F) = .20$$

Joint Probability is the probability of the intersection of the two events. For example,

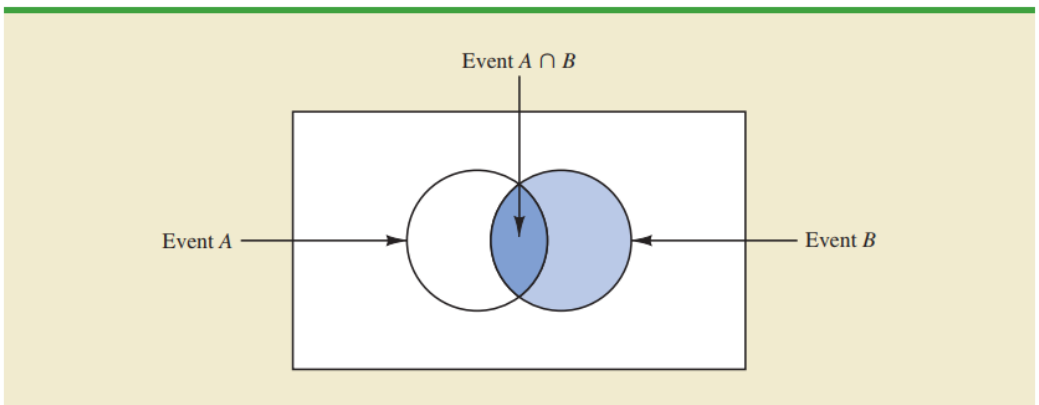
$$P(Yes \cap M) = \frac{288}{1200} = .24 \quad \text{or} \quad P(No \cap F) = \frac{204}{1200} = .17$$

Conditional Probability (1/5)

Conditional Probability is the probability of an event occurring given that another event has occurred.

- Conceptually, the conditional probability is limited or conditioned by the probability of the precursor event.
- For example, the figure below shows the conditional probability as the A within B. The probability of B determines the conditional probability.

FIGURE 4.8 CONDITIONAL PROBABILITY $P(A \mid B) = P(A \cap B)/P(B)$



$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{or} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Conditional Probability (2/5)

Illustration: Promotion Status of Police Officers

Promoted	M	F	Total
Yes	288	36	324
No	672	204	876
Total	960	240	1,200

Promoted	M	F	Total
Yes	.24	.03	.27
No	.56	.17	.73
Total	.80	.20	1.00

The probability of a promotion given the gender is male.

$$P(Yes|M) = \frac{P(Yes \cap M)}{P(M)} = \frac{288}{\mathbf{960}} = \frac{.24}{.80} = .30$$

Not to be confused with the probability of a promotion and the male.

$$P(Yes \cap M) = \frac{288}{\mathbf{1200}} = .24$$

Conditional Probability (3/5)

Illustration: Promotion Status of Police Officers

Promoted	M	F	Total
Yes	288	36	324
No	672	204	876
Total	960	240	1,200

Promoted	M	F	Total
Yes	.24	.03	.27
No	.56	.17	.73
Total	.80	.20	1.00

When compare the probability in connection with the gender, we found the followings:

The promotion chance for **male** officers

$$P(Yes|M) = \frac{P(Yes \cap M)}{P(M)} = \frac{288}{960} = \frac{.24}{.80} = .30$$

The promotion chance for **female** officers

$$P(Yes|F) = \frac{P(Yes \cap F)}{P(F)} = \frac{36}{240} = \frac{.03}{.20} = .15$$

The conclusion: A person's gender limits the promotion probability in this particular example.

Conditional Probability (4/5) - Check Your Understanding

Please find out all the marginal and joint probabilities. And fill in the results on the table with the blanks.

Illustration: Promotion Status of Police Officers (Branch K)

Promoted	M	F	Total
Yes	125	25	150
No	780	156	936
Total	905	181	1,086

Promoted	M	F	Total
Yes			
No			
Total			1.00

- Please evaluate the conditional probability on the chance of the promotion given someone's gender. What is $P(Yes|M)$ and $P(Yes|F)$ respectively?
- What would you comment on the gender equality in the promotion in this particular branch?

Conditional Probability (5/5) - Check Your Understanding

Let's also review the following case. Please compare the $P(Ace)$ with $P(Ace|Club)$.

Illustration: Poker Cards (Jokers Excluded)

	Club	$Club^C$	Total
Ace	1	3	4
Ace^C	12	36	48
Total	13	39	52

- What can you comment about the probability of picking an ace from a deck of cards and that of picking an ace from the club?
- Does the choice of the flower interfere with the chance of picking an ace?

Independent Events

When the followings are true, then we call A and B are independent events. That is, neither has influence over the chance of the other.

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

Q: If A and B are independent events and $P(A|B) = P(A)$, is this always true? $P(B|A) = P(B)$

Multiplication Law

The multiplication law says the following:

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

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

Remember this? Instead of trying to memorize the law, transform it from the conditional probability.

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

For independent events, the multiplication law is simply

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

(Because $P(A|B) = P(A)$ or $P(B|A) = P(B)$)

Check Your Understanding

An experiment A has two mutually exclusive sample points. Let's call these A_1 and A_2. The conditional probability of B given each of the A's sample points are known as below. Please complete the rest of the table.

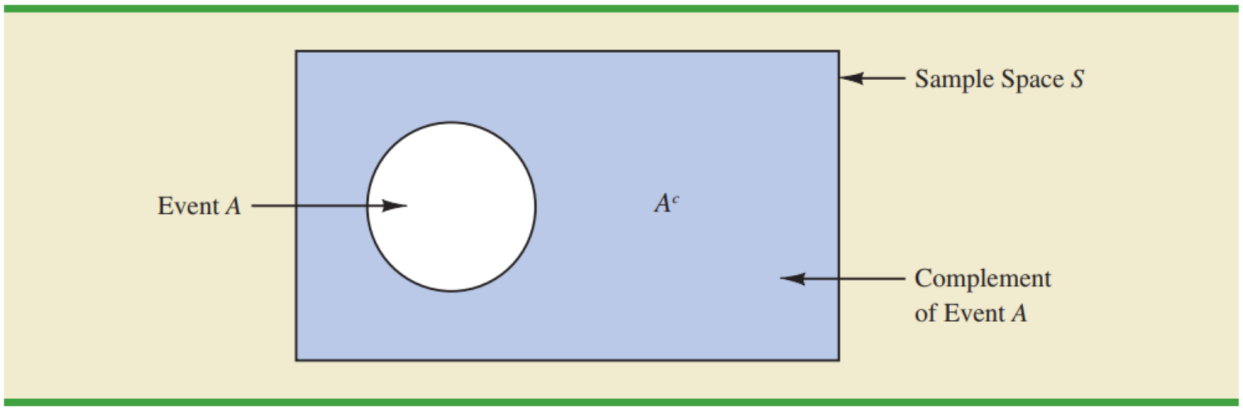
Events	**Marginal**	**Conditional**	**Joint**	**Updated**
A_1	$P(A_1) = .65$	$P(B\|A_1) = .02$	$P(A_1 \cap B) = ?$	$P(A_1\|B) = ?$
A_2	$P(A_2) = ?$	$P(B\|A_2) = .05$	$P(A_2 \cap B) = ?$	$P(A_2\|B) = ?$
	1.00		$P(B) = ?$	

(Note: The original marginal probability is updated based on the newly obtained evidences provided through the conditional probability. This is a simple example of the Bayes Theorem.)

Bonus - Graphical Tools to visualize the Probability (1/2)

Addition law visualization - **Venn's Diagram**

FIGURE 4.4 COMPLEMENT OF EVENT *A* IS SHADED



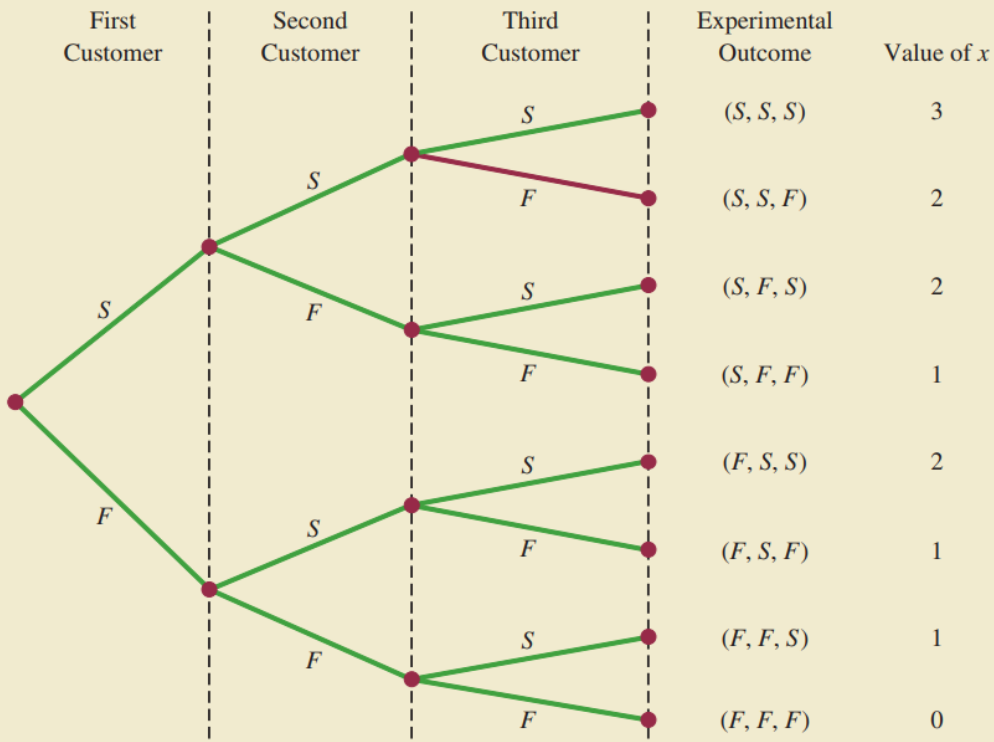
Joint probability visualization - **Matrix Table** (or Cross Tabulation)

Promoted	M	F	Total
Yes	.24	.03	.27
No	.56	.17	.73
Total	.80	.20	1.00

Bonus - Graphical Tools to visualize the Probability (2/2)

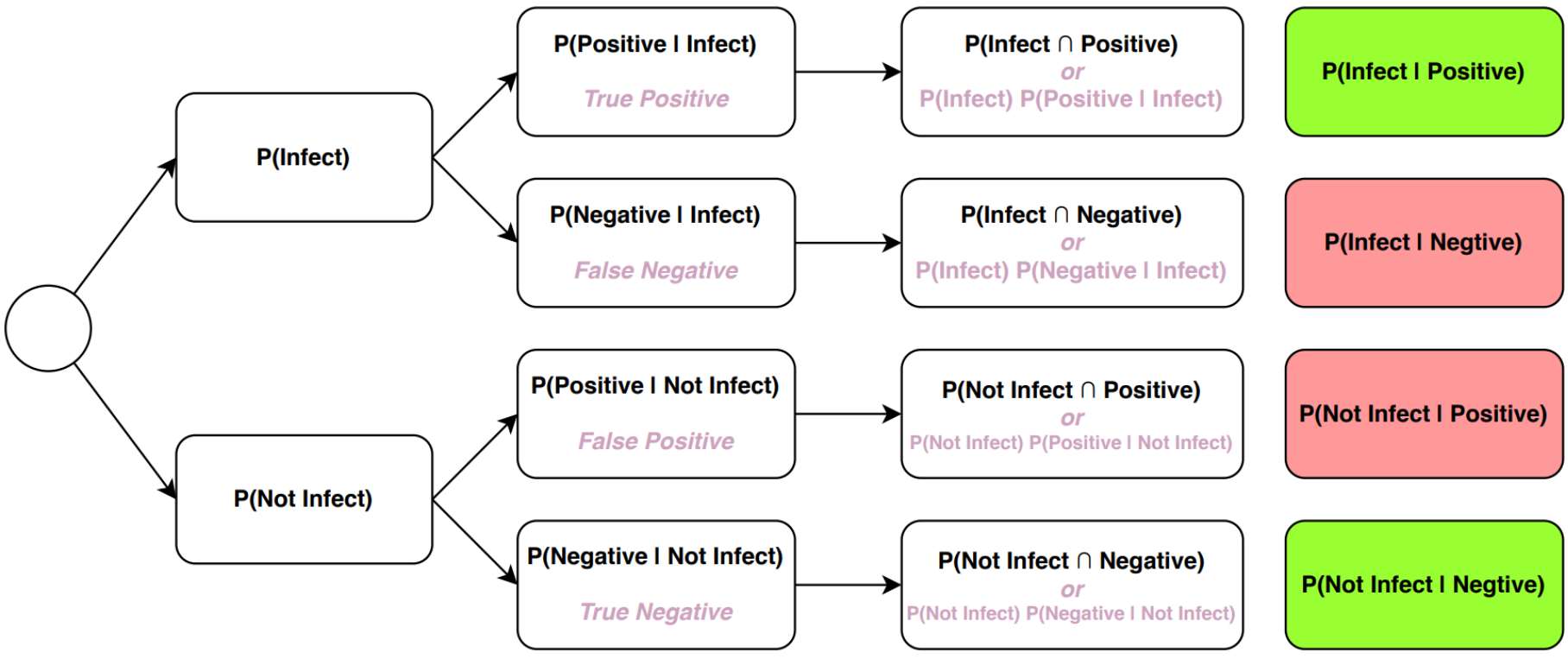
Multiplication law visualization - **Tree Diagram**

FIGURE 5.3 TREE DIAGRAM FOR THE MARTIN CLOTHING STORE PROBLEM



S = Purchase
F = No purchase
x = Number of customers making a purchase

Bonus: Saliva Based COVID19 Testing - A Bayesian Framework



IV. Probability Distributions

Overview: What We Know about Probability So Far

What we have learned in the previous section - Probability:

- An event can be deterministic or probabilistic;
- When the uncertainties are involved, by assigning probability to a specific outcome, we allow the solutions to be more inclusive;
- The term fraction, proportion, relative frequency, and probability are interchangeable;
- When a probability is assigned to all experimental outcomes, it forms a probability distribution.

Experimental Outcomes and Random Variable

We use the term **random variable** to numerically describe the sample outcomes.

- A variable can be understood as a container that can store values;
- The name *random variable* implies that we can store any values in the variable.
- The type of the variable can limit the way the experimental outcomes are being described.

Check Your Understanding

Consider the experiment of tossing a coin twice.

- List the experimental outcomes;
- Define a random variable x that represents the number of heads occurring on the two tosses;
- Show what value the random variable x would assume for each of the experimental outcomes.

Discrete Random Variable

A **discrete random variable** assumes either a finite number of values or an infinite sequence of values. For example,

TABLE 5.1 EXAMPLES OF DISCRETE RANDOM VARIABLES

Experiment	Random Variable (x)	Possible Values for the Random Variable
Contact five customers	Number of customers who place an order	0, 1, 2, 3, 4, 5
Inspect a shipment of 50 radios	Number of defective radios	0, 1, 2, . . . , 49, 50
Operate a restaurant for one day	Number of customers	0, 1, 2, 3, . . .
Sell an automobile	Gender of the customer	0 if male; 1 if female

The probability of a specific outcome can be obtained and it is meaningful.

Continuous Random Variable

A **continuous random variable** assumes any numerical value in an interval or collection of intervals. For example,

TABLE 5.2 EXAMPLES OF CONTINUOUS RANDOM VARIABLES

Experiment	Random Variable (x)	Possible Values for the Random Variable
Operate a bank	Time between customer arrivals in minutes	$x \geq 0$
Fill a soft drink can (max = 12.1 ounces)	Number of ounces	$0 \leq x \leq 12.1$
Construct a new library	Percentage of project complete after six months	$0 \leq x \leq 100$
Test a new chemical process	Temperature when the desired reaction takes place (min 150° F; max 212° F)	$150 \leq x \leq 212$

The probability of a specific outcome cannot be obtained and it is meaningless. We are interested in the probability of a section.

Check Your Understanding

Here is a list of experiments. Please make up your own random variable and discuss whether the variable is discrete or continuous.

- Take a 20-question examination
- Observe cars arriving at a tollbooth for 1 hour
- Audit 50 tax returns
- Observe an employee's phone calls to potential buyers
- Weight a shipment of goods

Components of a Discrete Probability Distribution

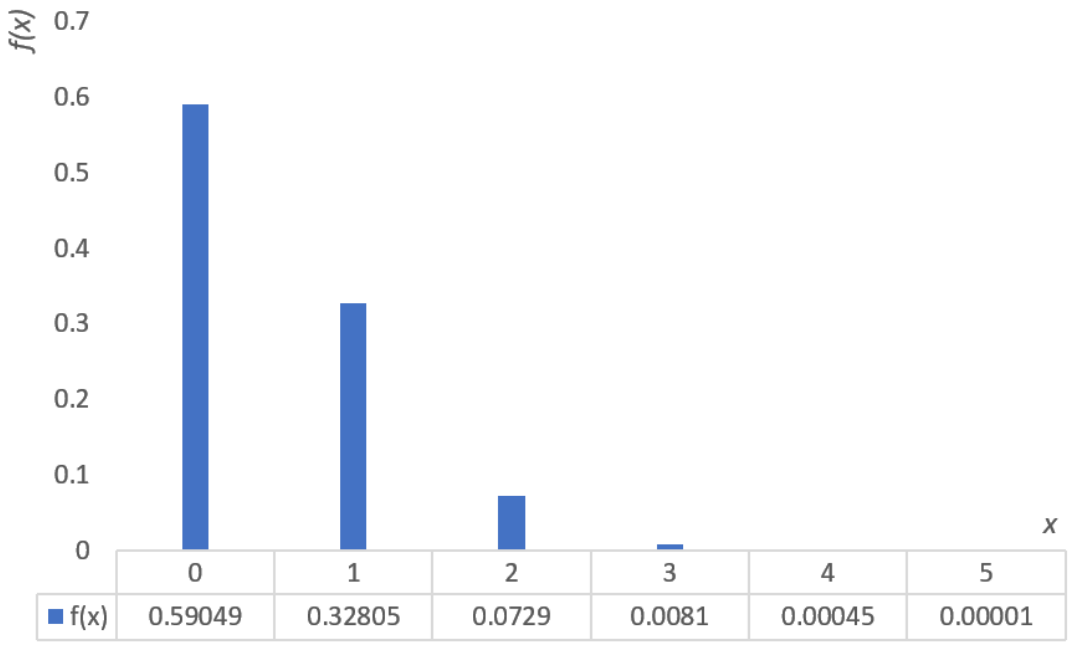
See below for a typical discrete probability distribution:

Prob. Dist. for Number of Broken Parts on a Five-Part Machine

x	$f(x)$
0	.5905
1	.3281
2	.0729
3	.0081
4	.0005
5	1.0E-5

- The **first column** is all the possible values of x. The x also represents the all the possible outcomes of the experiment.
- Instead of using $P(x)$, the **second column** uses $f(x)$, implies the probability is some function of the x.

Visualization of a Discrete Probability Distribution



We have seen similar graphs in the past. Also,

$$f(x_i) \geq 0$$

$$\sum_{i=0}^{N=5} f(x_i) = 1$$

Empirical Discrete Distribution

Please note that x here is NOT a nominal scale measure. It implies the value of x can become the basis of calculations.

When a relative frequency distribution is used to obtain the values of $f(x)$, we call such probability distribution as **empirical discrete distribution**.

The term **empirical** means "*based on, concerned with, or verifiable by observation or experience rather than theory or pure logic.*"

In this case, we map x with $f(x)$ using the past records. That is, $x \rightarrow$ **past records** $\rightarrow f(x)$.

Empirical Discrete Distribution: Expected Value

The **expected value**, or mean, of a random variable is a measure of the central tendency for the random variable [2].

For an empirical discrete distribution, the expected value can be calculated as the following:

$$E[x] = \mu = \sum x_i \cdot f(x_i)$$

Because the $f(x_i)$ part functions much like the weight of each x_i, it can be seen as the weighted mean.

Throwback: Remember the *=SUMPRODUCT(valueRange, weightRange)* formula from Excel?

[2] It can also be understood as the average level of output in long-run, when the experiment is repeated under the same condition.

Empirical Discrete Distribution: Variance

The level of the spread or the variability of the random variable [3] can be measured as the following:

$$Var(x) = \sigma^2 = \sum (x_i - \mu)^2 \cdot f(x_i)$$

(Note: The formula is much similar to that of the variance we learned previously. The previous one used $\frac{1}{N}$, while this one is using the weight.

[3] Essentially, this will be the level of variability for all the outcomes when the experiment is repeated under the same condition over time.

A Worked Example: Variance of an Empirical Discrete Distribution

(Note: Since it is an empirical discrete distribution, the x_i and $f(x_i)$ columns are provided based on the frequency.)

x_i	$f(x_i)$	$x_i - f(x_i)$	$(x_i - f(x_i))^2$	$(x_i - f(x_i))^2 \cdot f(x_i)$
0	0.5905	-0.5000	0.2500	0.1476
1	0.3281	0.5000	0.2500	0.0820
2	0.0729	1.5000	2.2500	0.1640
3	0.0081	2.5000	6.2500	0.0506
4	0.0004	3.5000	12.2500	0.0055
5	0.0000	4.5000	20.2500	0.0002

Expected Value: $E[x] = 0.50$ **Variance**: $Var[x] = 0.45$

Check Your Understanding

Interpretation Challenge!

Please go back to the previous slide. Please make sense of the expected value and variance by reviewing the distribution.

- What is x_i? Sample outcome or just unintelligible index number?
- The expected value is 0.50. Is it a probability or what? What is the unit of the value?
- The variance is 0.45. Is it a probability or what? What is the unit of the value?

Other Ways to Map x with $f(x)$ - Discrete Uniform

When the logical relationship is clear, we can simply rely on math formula to map x with $f(x)$.

For example, consider the example of a single fair dice throwing. The probability of the each side can be expressed as the follows:

$$f(x_i) = 1/6$$

where

$$x = \{1, 2, 3, 4, 5, 6\}$$

1. Please find out the expected value and the variance.
2. Simulate the experiment using excel and validate the results from step 1. (Hint: use *=RANDBETWEEN(min, max)* formula)

Other Ways to Map x with $f(x)$ - Binomial Distribution (1/10)

When an experiment satisfies a certain condition, we can conveniently exploit the characteristics of a binomial distribution.

Here are the conditions for a **binomial experiment**:

- The experiment consists of a **sequence** of n identical trials;
- **Two outcomes** are possible on each trial. (One *success*, the other *failure*. Thus, bi-nomial.)
- The probability of success, $P(S) = p$, is **consistent** in each trial. (Thus, $P(F) = 1 - p$)
- Each subsequent trial is **independent** from the previous trial.

Binomial Probability Distribution (2/10)

Examples of binomial experiments:
(Note: See if you can identify the 1) multi-steps, 2) two outcomes, 3) constant p in each step, and 4) independence assumptions)

- A sales person makes 10 phone calls to make a sale;
- Throwing 24 dices at the same time to see how many 1s are facing upward;
- 10 identical machines work at the same time to prevent a service disruption;
- Answering 10 multiple choice questions each with 4 answer choices;

Binomial Probability Distribution (3/10)

Let's put the following problem under the microscope to do a thorough study on the binomial distribution.

Q: The probability of 2 out of three customers make purchases, when each customer's chance of making purchase is determined at 0.30 and we are assuming the following conditions:

- $p = 0.30$ is constant from customer to customer;
- each customer's purchasing decision is independent.

Then, based on the theory (or the formula) which we are about to discuss in a great length, the answer is:

$$f(x = 2) = \binom{3}{2} \cdot 0.3^2 \cdot 0.7^1 = 0.189$$

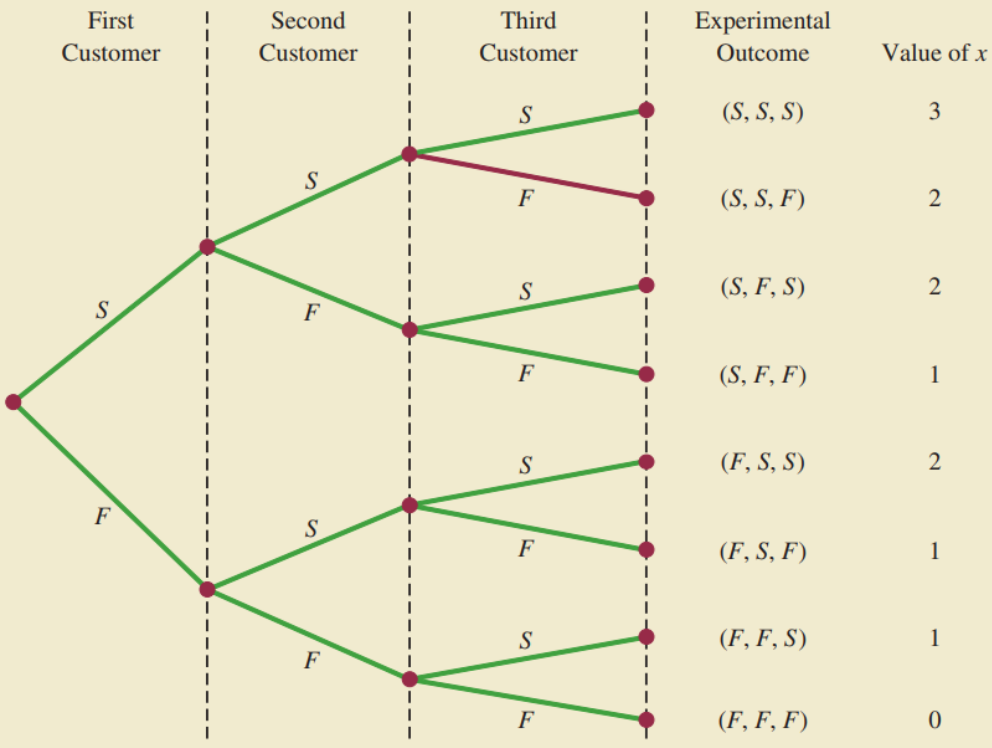
"Wait, what just happened?"

Binomial Probability Distribution (4/10)

First, why is this even a binomial experiment?

A Tree Diagram View

FIGURE 5.3 TREE DIAGRAM FOR THE MARTIN CLOTHING STORE PROBLEM



Binomial Probability Distribution (5/10)

Second, what is the probability of 2 out of 3 customers make purchase?

Because each customer's purchasing decision is independent, based on the multiplication law:

$$P(\text{2 out 3 make purchases}) = 0.3 \cdot 0.3 \cdot 0.7 = 0.063$$

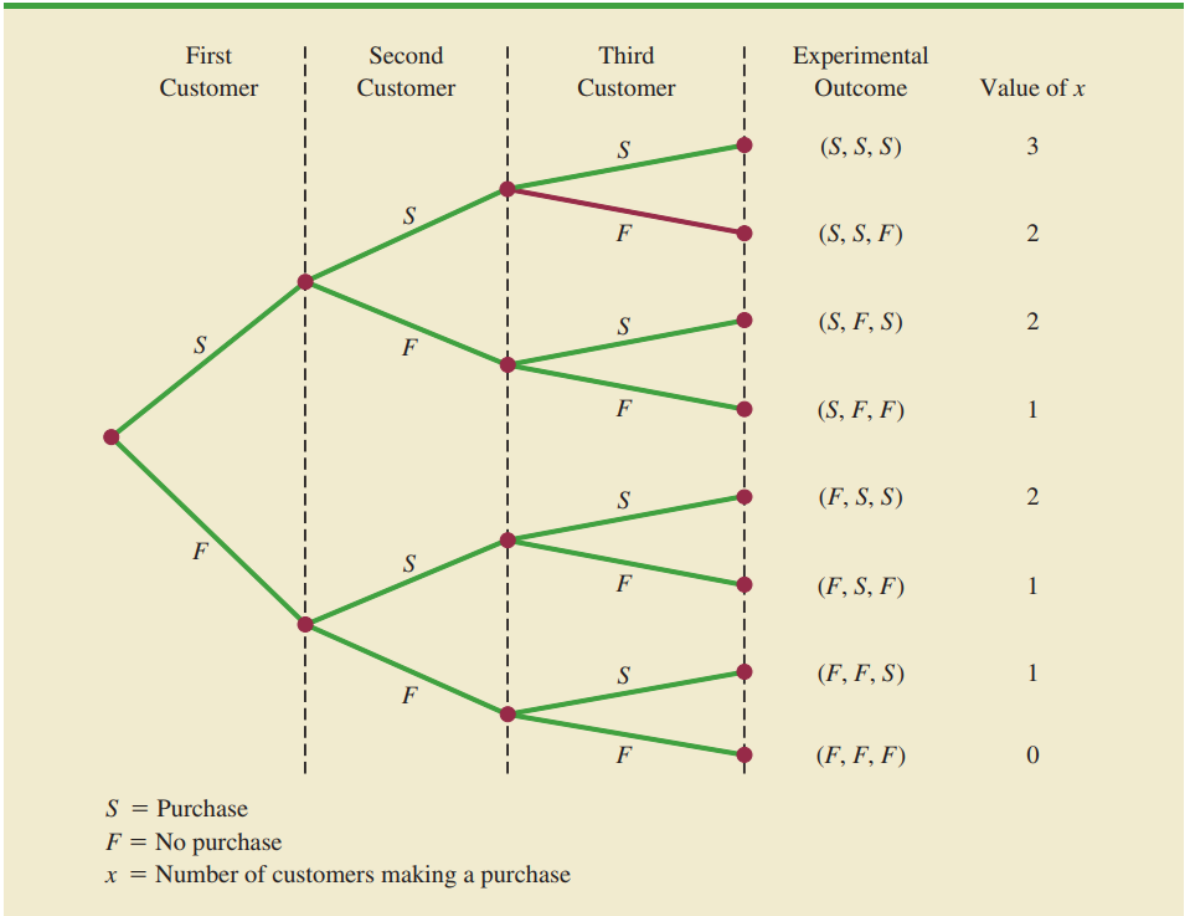
Then, why earlier it was 0.189? Which is the three times of 0.063?

Binomial Probability Distribution (6/10)

Third, how many experimental outcomes satisfy "2 out of 3"? What was the event this problem is looking for?

A Tree Diagram View

FIGURE 5.3 TREE DIAGRAM FOR THE MARTIN CLOTHING STORE PROBLEM



Binomial Probability Distribution (7/10)

To summarize, here is how we got 0.189.

$$P(x = 2) = \underbrace{3 \text{ counts}}_{\text{3 occasions}} \cdot \underbrace{0.3 \cdot 0.3}_{\text{2 successes}} \cdot \underbrace{0.7}_{\text{1 failure}} = 3 \cdot 0.063 = 0.189$$

Binomial Probability Distribution (8/10)

Here is the formula for **the probability of the event** x in a binomial distribution.

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

where,

- n total number of trials;
- x number of success we'd like to see;
- p the probability of the success event;
- $n - x$ number of failures;
- $1 - p$ the probability of the failure event;

Binomial Probability Distribution (9/10)

What does each component of the formula mean?

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

The meaning of each component:

- $\binom{n}{x}$ the number of experimental outcome that qualifies
- p^x or multiplying p for x times, the probability of x number of consecutive successes
- $(1-p)^{(n-x)}$ the probability of $(n-x)$ number of consecutive failures

The Excel solution

- Non-cumulative (P of x) = $BINOM.DIST(x, n, p, FALSE)$
- Cumulative (P up until x) = $BINOM.DIST(x, n, p, TRUE)$

Binomial Probability Distribution (10/10) - Distributional Characteristics

Shape

- When $n \cdot p \geq 5$, it resembles a normal distribution (aka. a symmetric bell curve shape).

Expected Value

$$E[x] = n \cdot p$$

Variance

$$Var[x] = n \cdot p \cdot (1 - p)$$

A Worked Example: Mary's Bakery Shop (1/4)

Problem: Mary runs a bakery shop business. Each day she calls 6 customers randomly to make a sale. If the probability of make purchase is consistent for each customer at 10 percent, what is the probability of closing a day's business with 4 sales?

Perspective: This is a very classic example of a binomial probability case. "She calls 6 customers" can be seen as an experiment with 6 consecutive steps, "each customer has the same 10 percent chance of purchase" satisfies the constant ,independent, and binary outcome conditions.

Can you identify all the components that goes into the formula:

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

A Worked Example: Mary's Bakery Shop (2/4)

Known parameters:

- $n = 6$
- $p = 0.10$
- $x = 4$

Solution:

$$f(x = 4) = BINOM.DIST(x, n, p, FALSE)$$

$$= BINOM.DIST(4, 6, .10, FALSE) = 0.001215$$

Let's explore further with this problem...

A Worked Example: Mary's Bakery Shop (3/4)

We can develop a full distribution for Mary's each day sale:

x	$f(x)$	$f(\leq x)$
0	0.53144	0.53144
1	0.35429	0.88574
2	0.09842	0.98415
3	0.01458	0.99873
4	0.00122	0.99995
5	0.00005	1.00000
6	0.00000	1.00000

$$E[x] = np = 6 \times 0.10 = .60$$

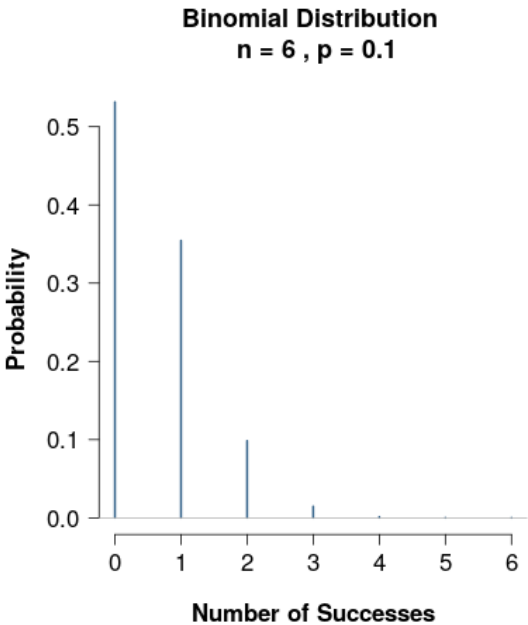
$$Var[x] = np(1-p) = 6 \times 0.10 \times 0.90 = .54$$

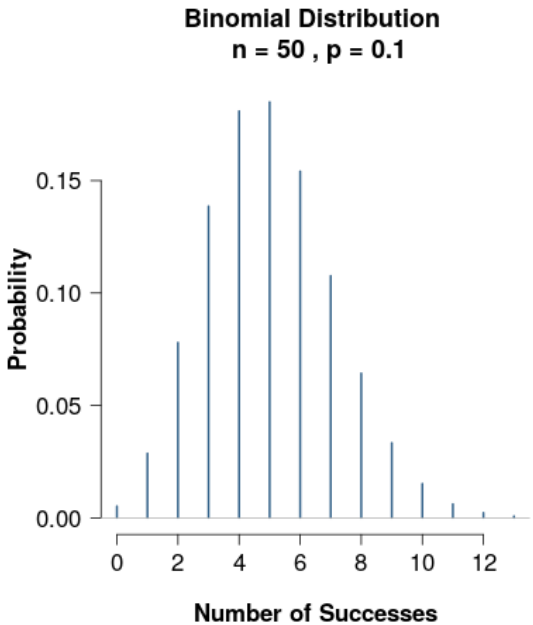
A Worked Example: Mary's Bakery Shop (4/4)

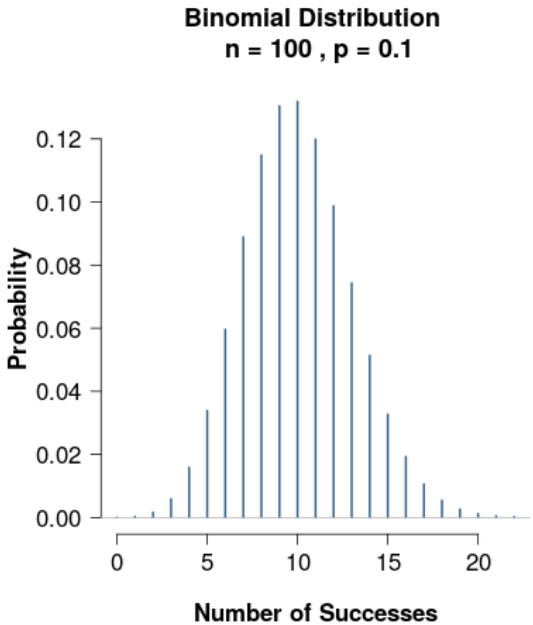
The insights from this exercise:

- Since the p and n are both low, Mary's daily sales will be low in general.
- In long run, we can expect to see 0.6 count of daily sales (based on the expected value).
- We will also see a very little surprise in every day business (low variance) with mostly 0 or 1 transaction.
- To improve the sales, she will need to improve the n or p or both.

What Happens When $np \geq 5$?







It is getting closer to a bell curve. If it is a bell curve, we can conveniently exploit the properties of the bell curve, which we are going to learn in the next section.