OKAN ÜNİVERSİTESİ
İSTANBUL

# BBA182 Applied Statistics
# Week 2 (2) Types of Data – (continued)

DR SUSANNE HANSEN SARAL

EMAIL: SUSANNE.SARAL@OKAN.EDU.TR

HTTPS://PIAZZA.COM/CLASS/IXRJ5MMOX1U2T8?CID=4#

WWW.KHANACADEMY.ORG

# NEW IN CLASS?

Send me an email to the following address:

**susanne.saral@okan.edu.tr**

# Activation of piazza.com account

Enter your first and last name

Select : Undergraduate

Select : Economy

Select : Class 1 and add BBA 182 and click “join the class”

OKAN ÜNİVERSİTESİ
İSTANBUL

# *Organizing* categorical data

Categorical data produce **values** that are names, words or codes, but **not** real numbers.

Only calculations based on the **frequency of occurrence** of these names, words or codes are valid.

We count the number of times a certain value occurs and add the frequency in the table.

OKAN ÜNİVERSİTESİ
İSTANBUL

# The Frequency and relative frequency - istribution Table

## **Summarizing** categorical data

A **frequency table** organizes data by recording totals and category names.

The variable we measure here is the number of times a country became world champion in football:

| Year | Champions | Year | Champions |
|---|---|---|---|
| 1930 | Uruguay | 1974 | W. Germany |
| 1934 | Italy | 1978 | Argentina |
| 1938 | Italy | 1982 | Italy |
| 1950 | Uruguay | 1986 | Argentina |
| 1954 | W. Germany | 1990 | W. Germany |
| 1958 | Brazil | 1994 | Brazil |
| 1962 | Brazil | 1998 | France |
| 1966 | England | 2002 | Brazil |
| 1970 | Brazil | 2006 | Italy |
| | | 2010 | Spain |
| | | 2014 | Germany |

| World champion in Football | Number of times |
|---|---|
| Italy | 4 |
| Argentina | 2 |
| France | 1 |
| Uruguay | 2 |
| Brazil | 5 |
| Germany | 4 |
| England | 1 |
| Spain | 1 |
| **Total** | **20** |

OKAN ÜNİVERSİTESİ
İSTANBUL

# Contingency table another type of frequency table

**Contingency tables** list the number of observations for every combination of values for **two** categorical variables

OKAN ÜNİVERSİTESİ
İSTANBUL

# Contingency table

A larger retailer of electronics conducted a survey to determine consumer preferences for various brands of digital cameras. The table summarizes responses by brand and gender:

| Electronics brand | Female | Male | Total |
|---|---|---|---|
| Cannon Power Shot | 73 | 59 | 132 |
| Nikon CoolPix | 49 | 47 | 96 |
| other brands | 86 | 67 | 153 |
| **Total** | **208** | **173** | **381** |

Each cell in a contingency table (any intersection of a row and column of the table) gives the **count** for a combination of values of  two categorical variables

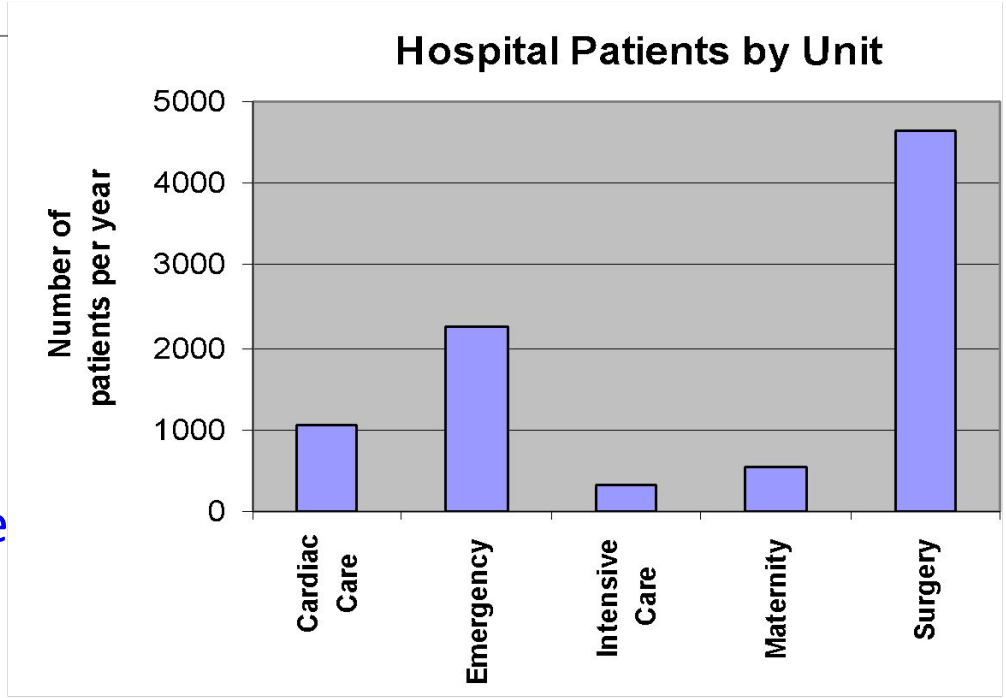
OKAN ÜNİVERSİTESİ
İSTANBUL

# Three Rules of Data Analysis

Rule 1, 2 and 3: Make a picture of the data

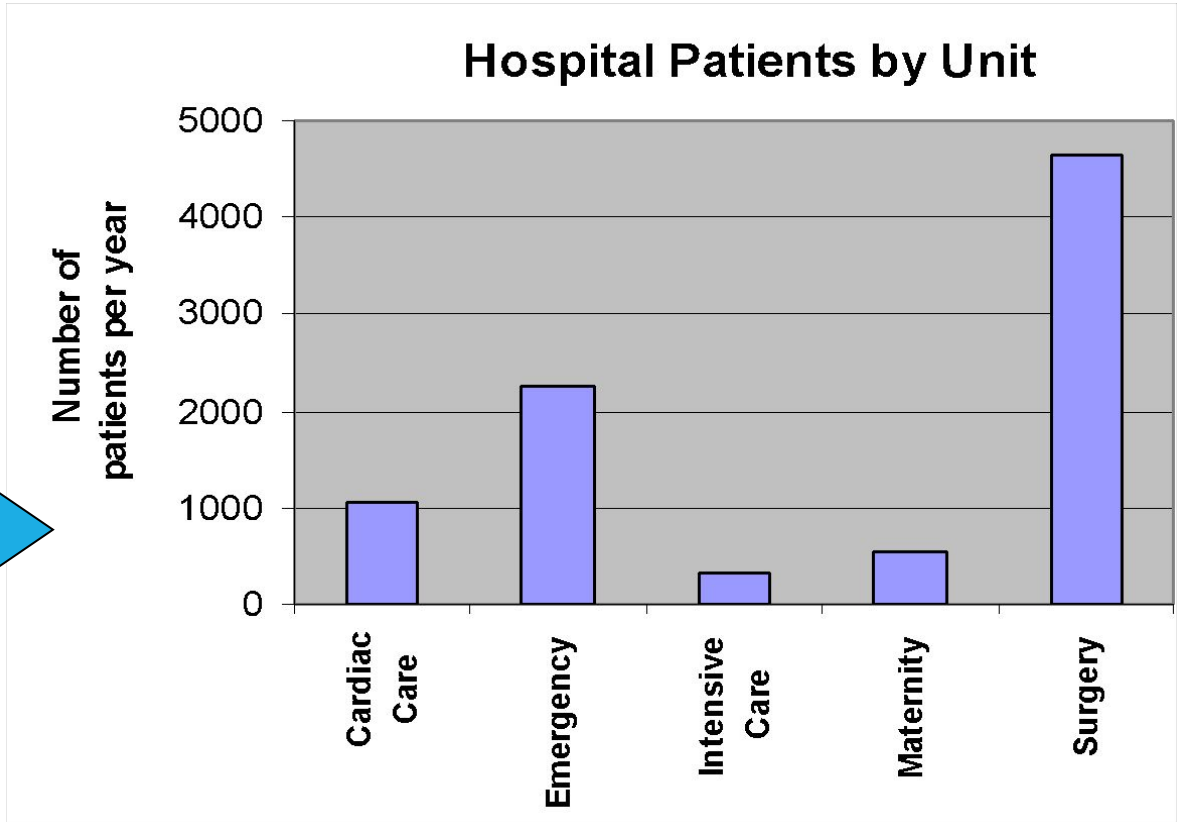
Pictures….

- Reveal things that cannot be seen in a frequency table
- Show important patterns in the data
- Provide an excellent way for presenting findings to other people



# Bar Chart – Hospital patients

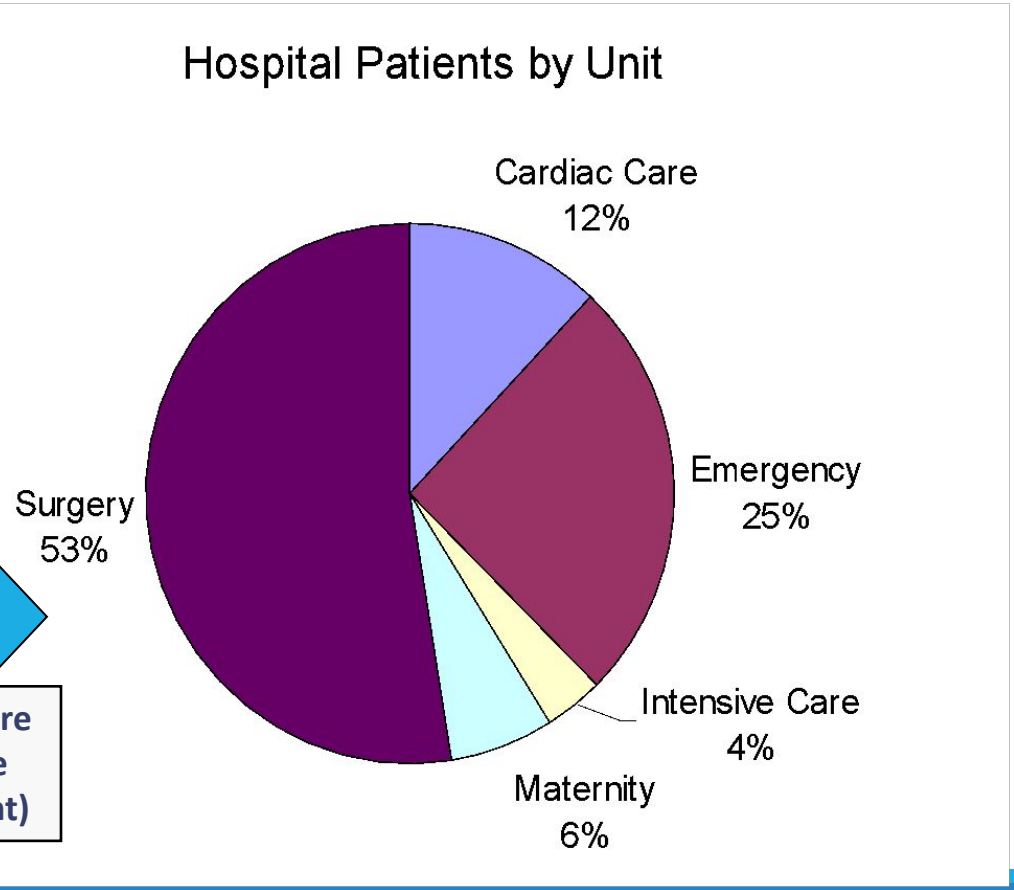
| Hospital Unit | Number of Patients |
|---|---|
| Cardiac Care | 1,052 |
| Emergency | 2,245 |
| Intensive Care | 340 |
| Maternity | 552 |
| Surgery | 4,630 |



# Pie Chart – Hospital patients

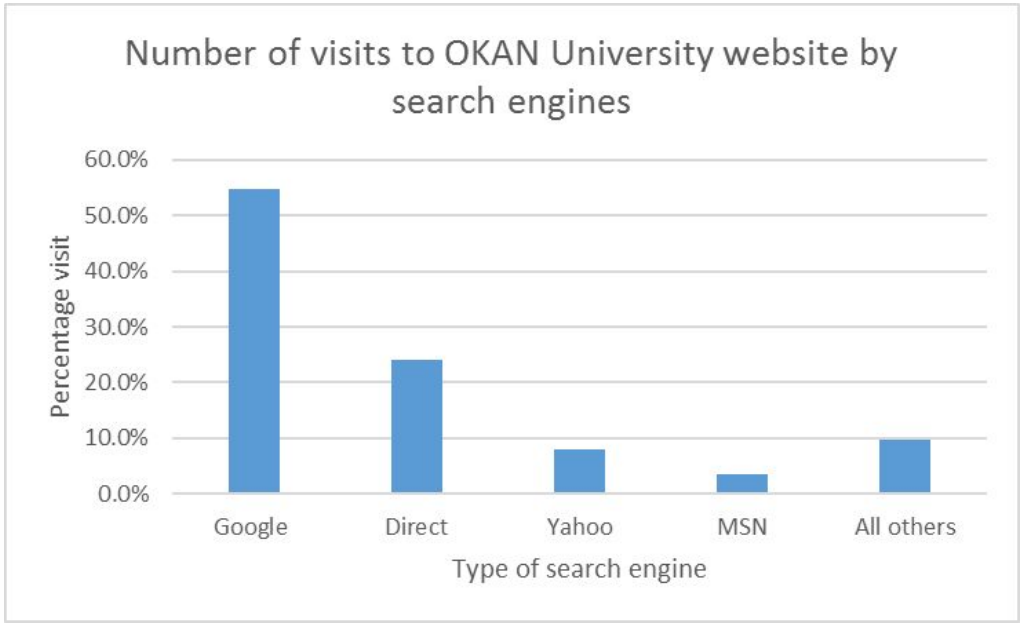
| Hospital Unit | Number of Patients | % of Total |
|---|---|---|
| Cardiac Care | 1,052 | 11.93 |
| Emergency | 2,245 | 25.46 |
| Intensive Care | 340 | 3.86 |
| Maternity | 552 | 6.26 |
| Surgery | 4,630 | 52.50 |

(Percentages are rounded to the nearest percent)



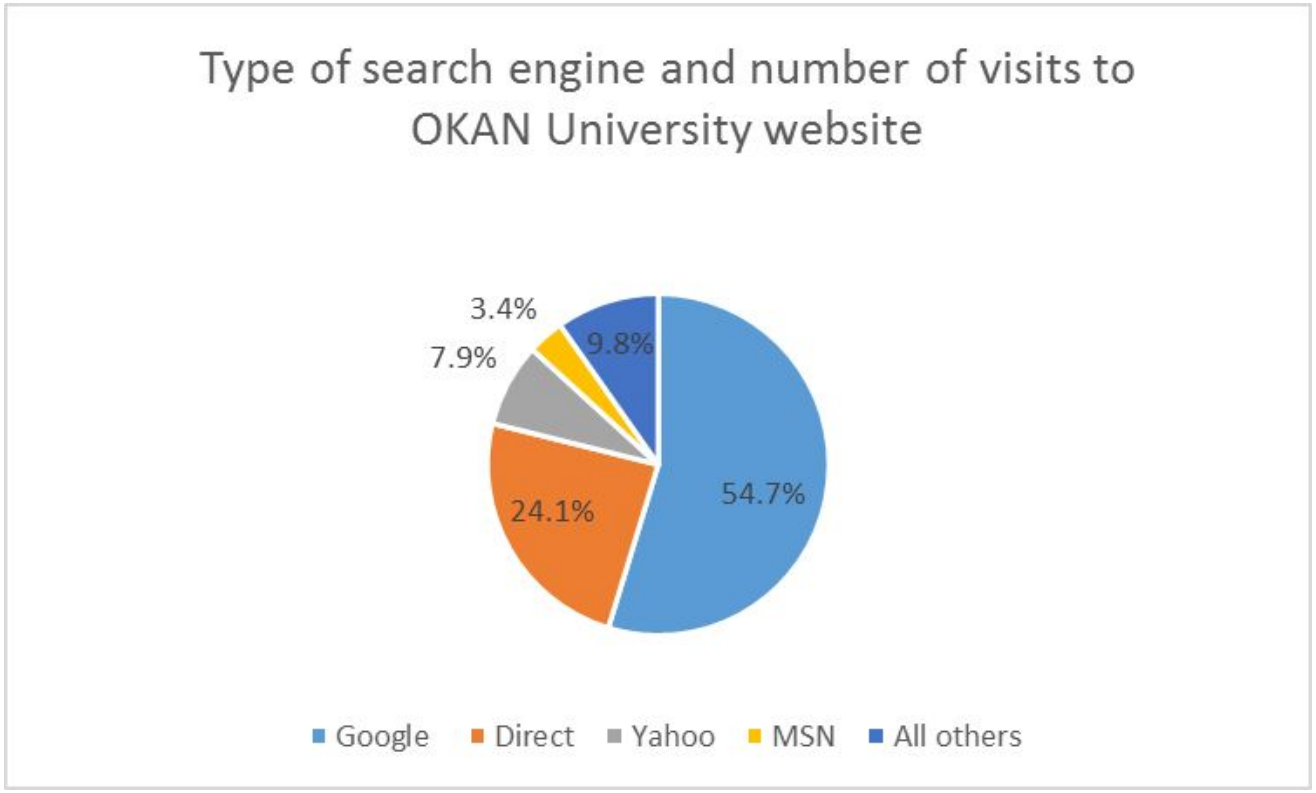
# Bar-chart
# Number of visits to OKAN University

| Search engine | Frequency (# of visits) | Relative frequency |
|---|---|---|
| Google | 50269 | 54.7% |
| Direct | 22173 | 24.1% |
| Yahoo | 7272 | 7.9% |
| MSN | 3166 | 3.4% |
| All others | 8967 | 9.8% |
| **Total** | **91847** | **100.0%** |



# Pie-chart
# Number of visits to OKAN University

| Search engine | Frequency (# of visits) | Relative frequency |
|---|---|---|
| Google | 50269 | 54.7% |
| Direct | 22173 | 24.1% |
| Yahoo | 7272 | 7.9% |
| MSN | 3166 | 3.4% |
| All others | 8967 | 9.8% |
| **Total** | **91847** | **100.0%** |



# Graphing
# Multivariate Categorical Data *(continued)*

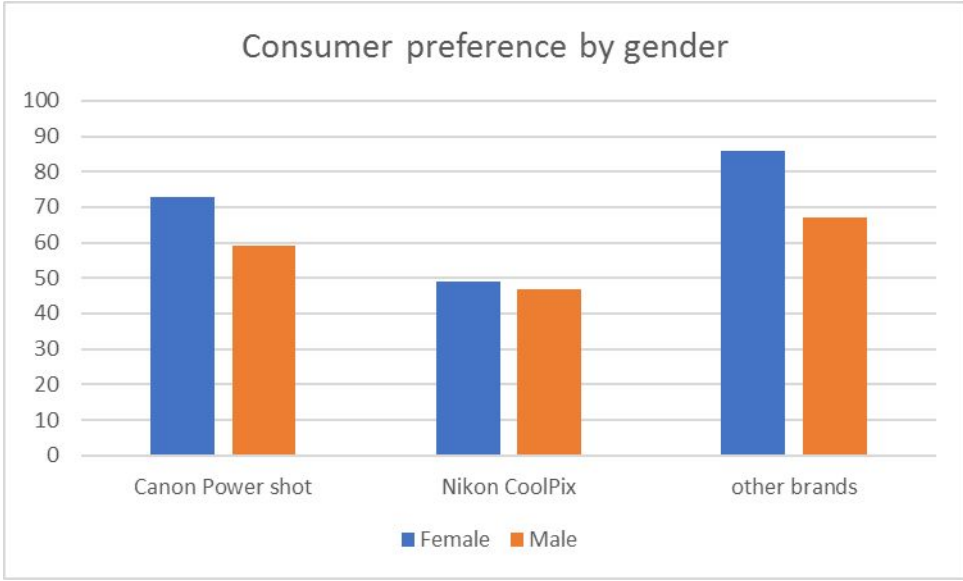
MULTIVARIATE= MORE THAN ONE VARIABLE

## Why multivariate?

We are investigating more than one variable:

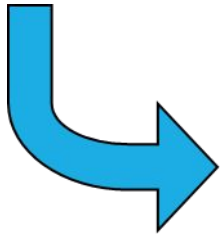
(1) Gender: Female and male

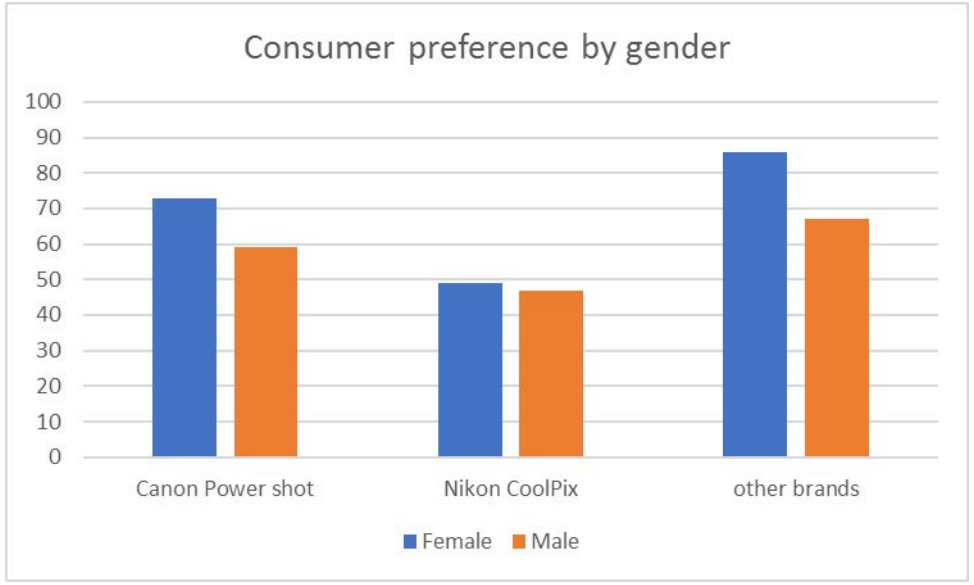
(2) Camera brand: Canon Powershot, Nikon CoolPix, other brands



# Graphing
# Multivariate Categorical Data

| Electronics brand | Female | Male | Total |
|---|---|---|---|
| Cannon Power Shot | 73 | 59 | 132 |
| Nikon CoolPix | 49 | 47 | 96 |
| other brands | 86 | 67 | 153 |
| **Total** | **208** | **173** | **381** |



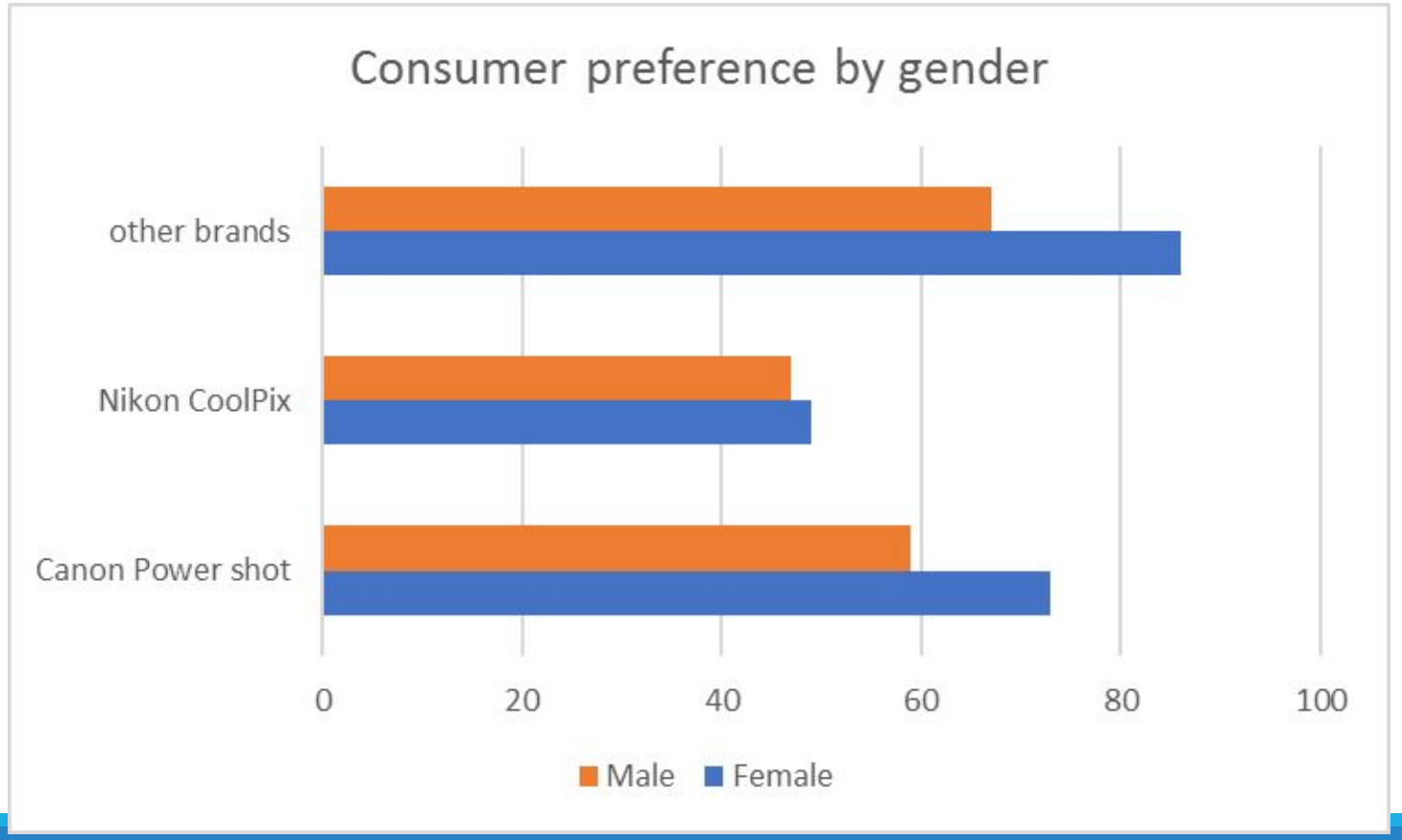


# Graphing Multivariate Categorical Data

*(continued)*

- **Side by side horizontal bar chart**

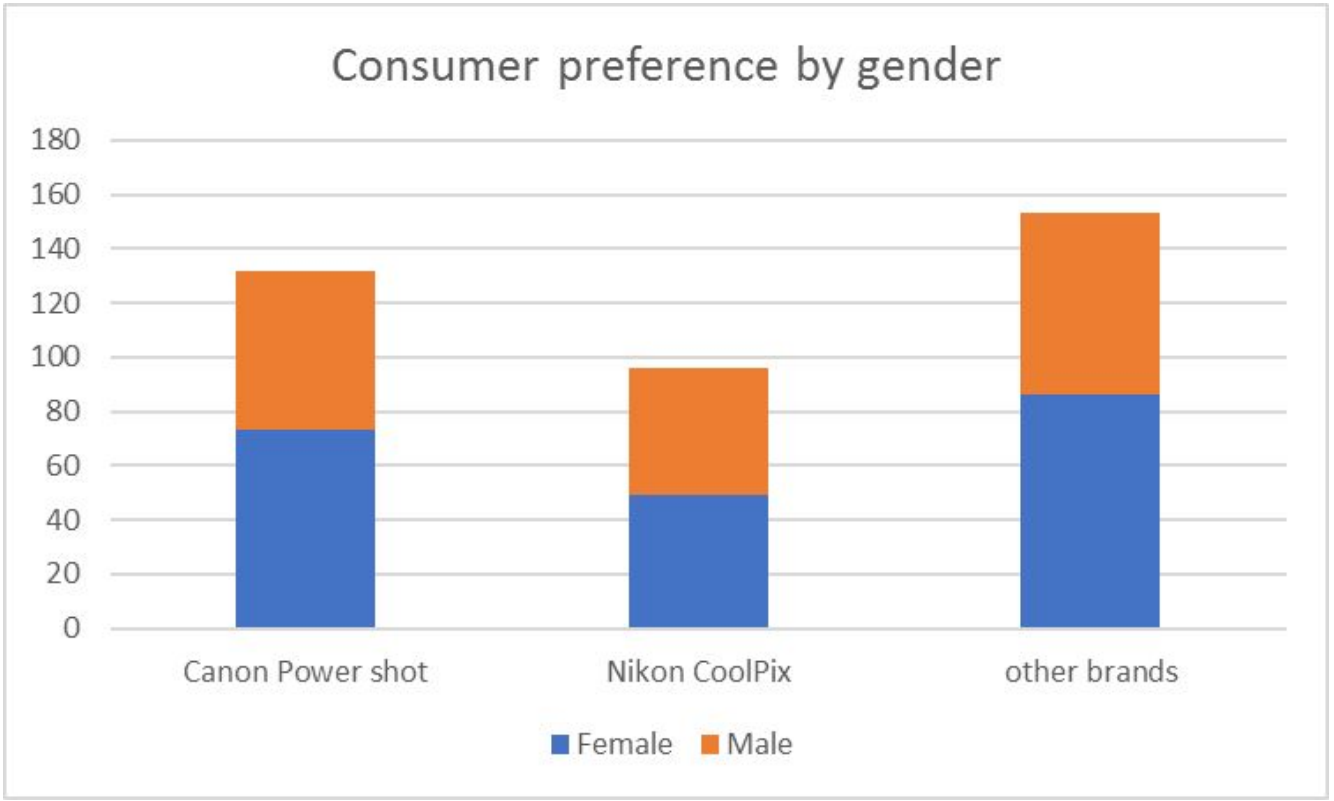
| Electronics brand | Female | Male | Total |
|---|---|---|---|
| Cannon Power Shot | 73 | 59 | 132 |
| Nikon CoolPix | 49 | 47 | 96 |
| other brands | 86 | 67 | 153 |
| **Total** | **208** | **173** | **381** |



# Graphing
# Multivariate Categorical Data *(continued)*

## Stacked bar chart

| Electronics brand | Female | Male | Total |
|---|---|---|---|
| Cannon Power Shot | 73 | 59 | 132 |
| Nikon CoolPix | 49 | 47 | 96 |
| other brands | 86 | 67 | 153 |
| **Total** | **208** | **173** | **381** |



# Class exercise

The ~~following raw data show responses to the question "What is your primary source for news?"~~

from **a sample of college students**:

Internet Newspaper Internet TV Internet Newspaper TV Internet Internet TV

Newspaper TV TV Newspaper TV Internet Internet Internet Internet Internet

TV Internet Internet TV TV

a. Prepare a frequency table for these data. How many students were sampled?
b. Prepare a relative frequency table for these data.
c. Based on the frequencies, construct a bar chart manually.
d. What is the variable we are measuring?

# Class exercise

A cable company surveyed its customers and asked how likely they were to bundle other services, such as phone and Internet, with their cable TV subscription. The following raw data show the responses:

| | | | |
|---|---|---|---|
| Very Likely | Unlikely | Unlikely | Very Likely |
| Likely | Unlikely | Likely | Likely |
| Unlikely | Unlikely | Likely | Likely |
| Very Likely | Unlikely | Unlikely | Very Likely |
| Unlikely | Unlikely | Unlikely | Likely |

a. Prepare a frequency table for these data. How many customers were sampled?

b. Prepare a relative frequency table for these data.

c. Based on frequencies, construct a bar chart manually

d. What is the variable we are measuring?

OKAN ÜNİVERSİTESİ
İSTANBUL

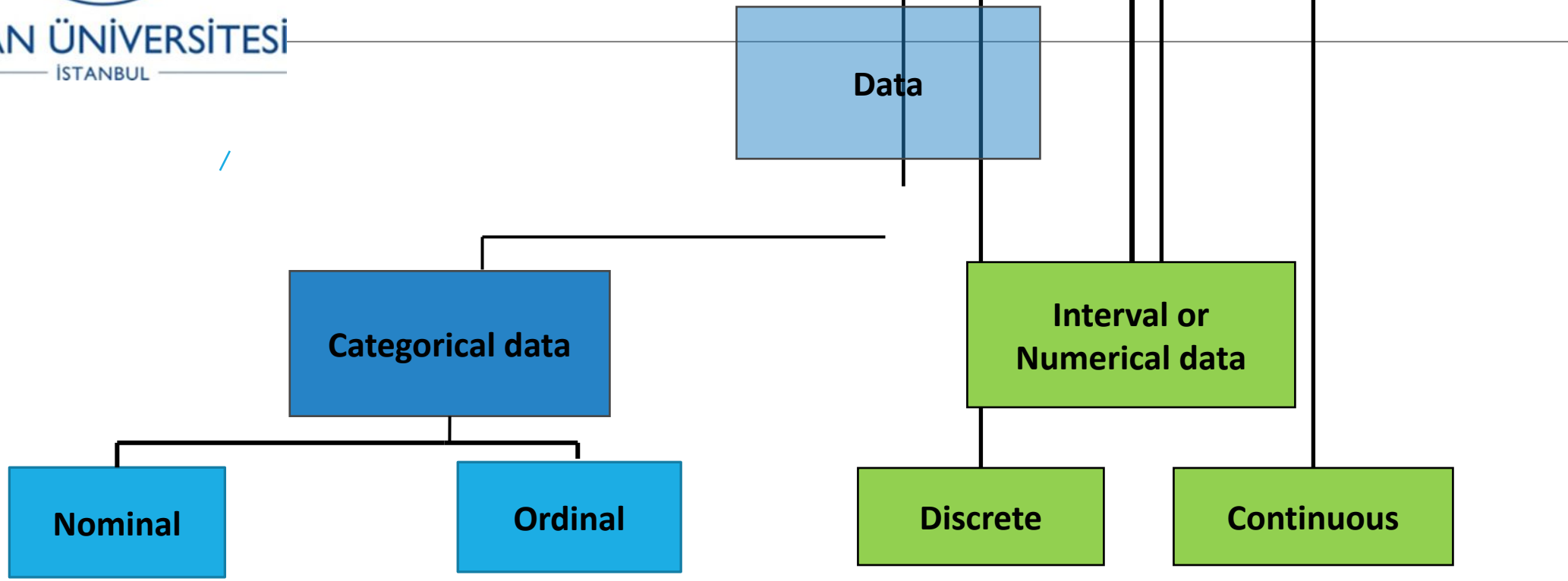
# Week 2 (2) How to organize and illustrate numerical data

DR SUSANNE HANSEN SARAL

EMAIL: SUSANNE.SARAL@OKAN.EDU.TR OR

SUSANNEHANSENSARAL@GMAIL.COM

# Classification of ariables




**Examples:**

- **# of goals in a football match**
- **# of subscriptions**
- **# of meals sold in a restaurant (Counted items)**

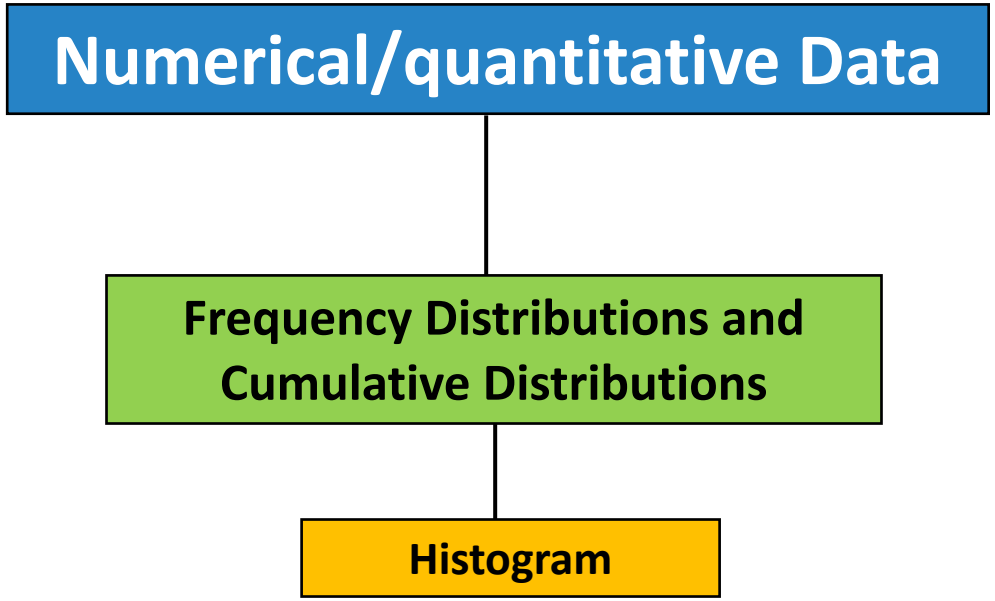
**Examples:**

- **Weight**
- **Volume**
- **Size**

**(Measured in units)**

OKAN ÜNİVERSİTESİ
İSTANBUL

# Tables and Graphs to Describe Numerical Variables



# Enron Corporation - energy trading company

Energy trading company from 1985 – 2001 (then went bankrupt):

- Company grew steadily over the 15 years
- Stock price in 1985 $ 5/share. By the end of 2000 it was $ 89.75
- At the end of 2000 the company was worth $ 6 billion

At the end of 2001 the stock had fallen to $ 0.25! The company had lost 99% of it's value

**Were there any warning signs in the data?**

# Enron Corporation - energy trading company

Energy trading company from 1985 – 2001:

- Were there any warning signs in the data?

| Monthly stock price change in dollars of Enron stock for the period January 1997 to December 2001 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | |
| | Jan. | Feb. | Mar. | Apr. | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
| 1997 | -1.44 | -1.75 | -0.69 | -0.88 | 0.12 | 0.75 | 0.81 | -1.75 | 0.69 | -0.22 | -0.16 | 0.34 |
| 1998 | 0.78 | 0.62 | 2.44 | -0.28 | 2.22 | -0.5 | 2.06 | -0.88 | -4.5 | 4.12 | 1.16 | -0.5 |
| 1999 | 3.28 | 3.34 | -1.22 | 0.47 | 5.26 | -1.59 | 4.31 | 1.47 | -0.72 | -0.038 | -3.25 | 0.03 |
| 2000 | 5.72 | 21.06 | 4.5 | 4.56 | -1.25 | -1.19 | -3.12 | 8 | 9.31 | 1.12 | -3.19 | -17.75 |
| 2001 | 14.38 | -1.08 | -10.11 | -12.11 | 5.84 | -9.37 | -4.74 | -2.69 | -10.61 | -5.85 | -17.16 | -11.59 |

# Enron Corporation - energy trading company

Energy trading company from 1985 – 2001:

- Were there any warning signs about the fall of the stock price in the data?

Hard to tell from the raw data

Let’s follow the **first rule of data analysis** and make **a picture** of the data

OKAN ÜNİVERSİTESİ
İSTANBUL

# Enron Corporation – frequency distributio

Monthly stock price change in dollars of Enron stock for the period January 1997 to December 2001

| | Jan. | Feb. | Mar. | Apr. | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1997** | -1.44 | -1.75 | -0.69 | -0.88 | 0.12 | 0.75 | 0.81 | -1.75 | 0.69 | -0.22 | -0.16 | 0.34 |
| **1998** | 0.78 | 0.62 | 2.44 | -0.28 | 2.22 | -0.5 | 2.06 | -0.88 | -4.5 | 4.12 | 1.16 | -0.5 |
| **1999** | 3.28 | 3.34 | -1.22 | 0.47 | 5.26 | -1.59 | 4.31 | 1.47 | -0.72 | -0.038 | -3.25 | 0.03 |
| **2000** | 5.72 | 21.06 | 4.5 | 4.56 | -1.25 | -1.19 | -3.12 | 8 | 9.31 | 1.12 | -3.19 | -17.75 |
| **2001** | 14.38 | -1.08 | -10.11 | -12.11 | 5.84 | -9.37 | -4.74 | -2.69 | -10.61 | -5.85 | -17.16 | -11.59 |

| *Price change* | *# of months* |
|---|---|
| -20 | 0 |
| -15 | 2 |
| -10 | 4 |
| -5 | 2 |
| 0 | 24 |
| 5 | 21 |
| 10 | 5 |
| 15 | 1 |
| 20 | 0 |
| More | 1 |

Frequency table for the price change of Enron st

OKAN ÜNİVERSİTESİ
İSTANBUL

# Why Use Frequency Distributions and graphs for numerical data?

- A frequency distribution is a way to summarize numerical data
- **It condenses** the raw data into ranges/intervals
- and allows for a **quick visual interpretation** of the data – a **PICTURE**

**The picture of numerical/quantitative data is called a histogram**

OKAN ÜNİVERSİTESİ
İSTANBUL

# Frequency Distributions

What is a Frequency Distribution for **numerical data**?

- A frequency distribution is a table
- containing ranges/intervals within which the data fall
- and the corresponding frequencies with which data fall within each class or category

| | *Price change* | *# of months* | |
|---|---|---|---|
| | -20 | 0 | |
| | -15 | 2 | |
| | -10 | 4 | |
| | -5 | 2 | |
| | 0 | 24 | |
| | 5 | 21 | |
| | 10 | 5 | |
| | 15 | 1 | |
| | 20 | 0 | |
| | More | 1 | |

Frequency table for the price change of Enron

# Frequency Distributions for numerical data

**Intervals for numerical data** are not as easy to identify as for **categorical data.**

Determining the intervals of a frequency table for numerical data requires answers to the following questions:

- How many intervals should be used?
- How wide should each interval be?

# Raw data
# (sample of 110 employees in a production plant)

OKAN ÜNİVERSİTESİ
İSTANBUL

Not easy to see a picture or pattern!

**Completion Times of a particular task (in seconds) for 110 employees**

271 236 294 252 254 263 266 222 262 278 288

262 237 247 282 224 263 267 254 271 278 263

262 288 247 252 264 263 247 225 281 279 238

252 242 248 263 255 294 268 255 272 271 291

263 242 288 252 226 263 269 227 273 281 267

263 244 249 252 256 263 252 261 245 252 294

288 245 251 269 256 264 252 232 275 284 252

263 274 252 252 256 254 269 234 285 275 263

263 246 294 252 231 265 269 235 275 288 294

263 247 252 269 261 266 269 236 276 248 299

# How to determine the number of intervals/classes

## A quick guide

| Sample size | Number of intervals |
|---|---|
| Fewer than 50 | 5 - 7 |
| 50 to 100 | 7 - 8 |
| 101 to 500 | 8 - 10 |
| 501 to 1,000 | 10 - 11 |
| 1,001 to 5,000 | 11 - 14 |
| More than 5,000 | 14 - 20 |

Use at least 5 intervals but no more than 15-20 otherwise we loose the overview of the data

OKAN ÜNİVERSİTESİ
İSTANBUL

# How to determine the interval width

Each class/interval grouping has to have the same width

Determine the width of each interval by

$$w = \text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$$

- Use at least 5 but no more than 15-20 intervals
- Intervals never overlap
- Round up the interval width to get desirable interval endpoints

OKAN ÜNİVERSİTESİ
İSTANBUL

# Employee completion time

110 employees' time have been recorded and the plant supervisor needs to report to his manager how long on average his employees finish the job.

We have 110 values ranging **from 222 seconds to 299**

We need to determine the ***number of intervals***:

| Sample size | Number of intervals |
|---|---|
| Fewer than 50 | 5 - 7 |
| 50 to 100 | 7 - 8 |
| 101 to 500 | 8 - 10 |
| 501 to 1,000 | 10 - 11 |
| 1,001 to 5,000 | 11 - 14 |
| More than 5,000 | 14 - 20 |

OKAN ÜNİVERSİTESİ
İSTANBUL

# Employee completion time

Determine width of interval:

$$\text{Interval width} = \frac{\textbf{largest number} - \textbf{smallest number}}{\textbf{number of intervals}}$$

$$\text{Interval width} = \frac{299 - 222}{8} = 9.6 \text{ - rounded up to } 10$$

Interval width = 10

OKAN ÜNİVERSİTESİ
İSTANBUL

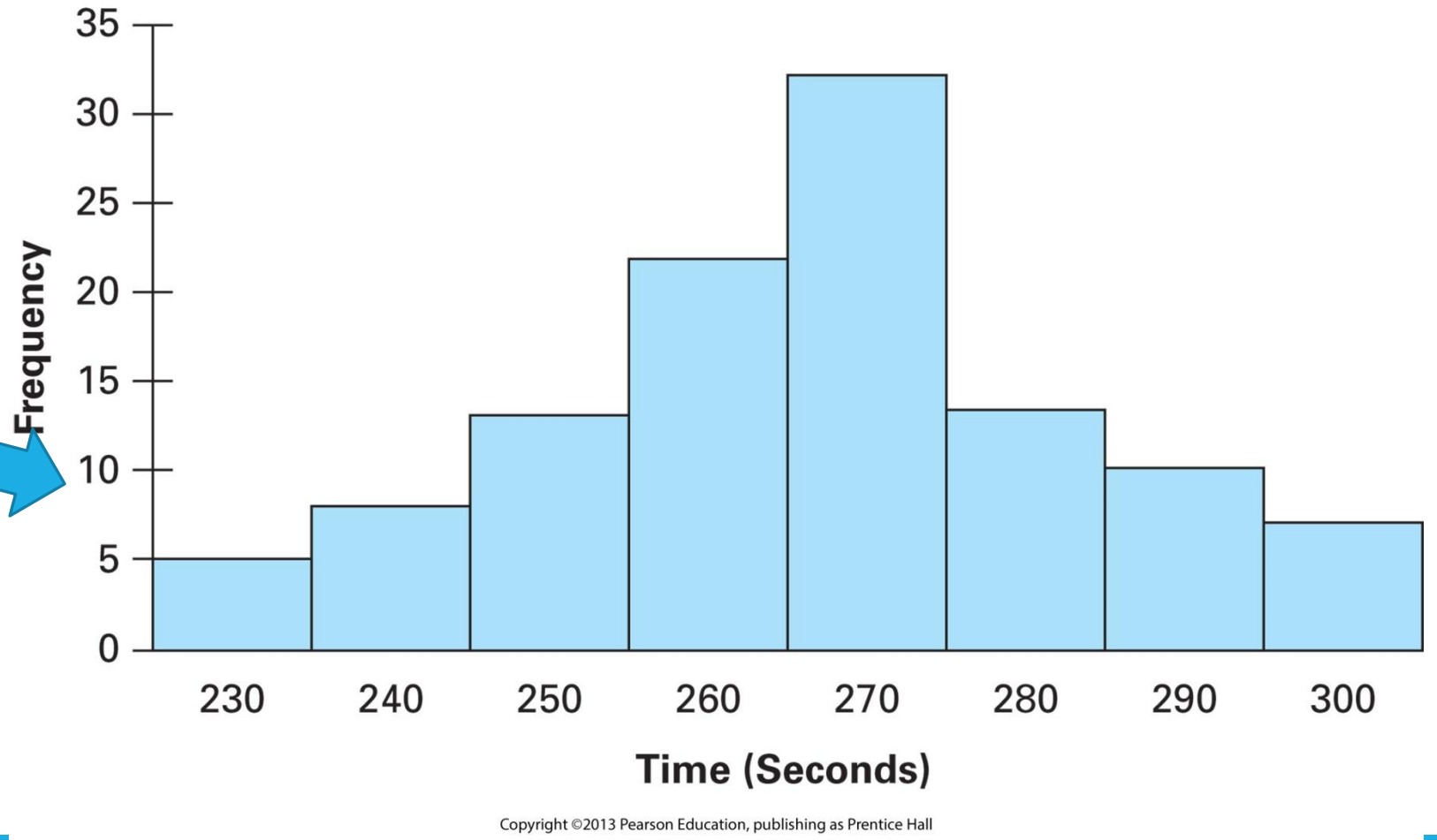
# Employee completion time

| Completion time (in seconds) | Frequency | Relative frequency % |
|---|---|---|
| 220 – 229 | 5 | 4.5 |
| 230 – 239 | 8 | 7.3 |
| 240 – 249 | 13 | 11.8 |
| 250 – 259 | 22 | 20.0 |
| 260 – 269 | 32 | 29.1 |
| 270 – 279 | 13 | 11.8 |
| 280 – 289 | 10 | 9.1 |
| 290 – 300 | 7 | 6.4 |
| Total | 110 | 100 % |

OKAN ÜNİVERSİTESİ
İSTANBUL

# Histogram of employee completion tim
# Absolute frequency

| Interval (sec.) | Frequency |
|---|---|
| 230 | 5 |
| 240 | 8 |
| 250 | 13 |
| 260 | 22 |
| 270 | 32 |
| 280 | 13 |
| 290 | 10 |
| 300 | 7 |

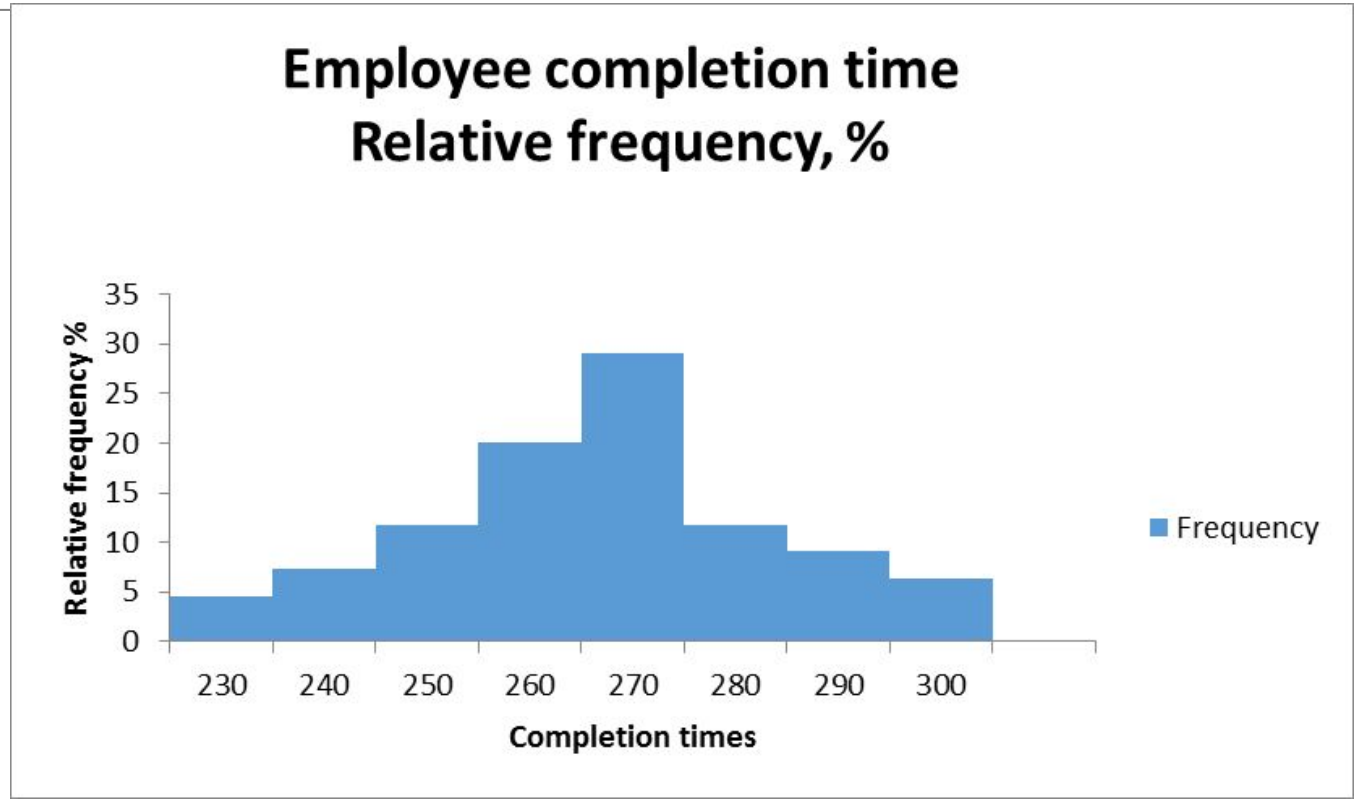


OKAN ÜNİVERSİTESİ
İSTANBUL

# Histogram of employee completion times Relative frequency

same graph as absolute frequency

| *Completion time in sec.* | *Relative frequency* |
|---|---|
| 230 | 4.5 |
| 240 | 7.3 |
| 250 | 11.8 |
| 260 | 20 |
| 270 | 29.1 |
| 280 | 11.8 |
| 290 | 9.1 |
| 300 | 6.4 |



OKAN ÜNİVERSİTESİ
İSTANBUL

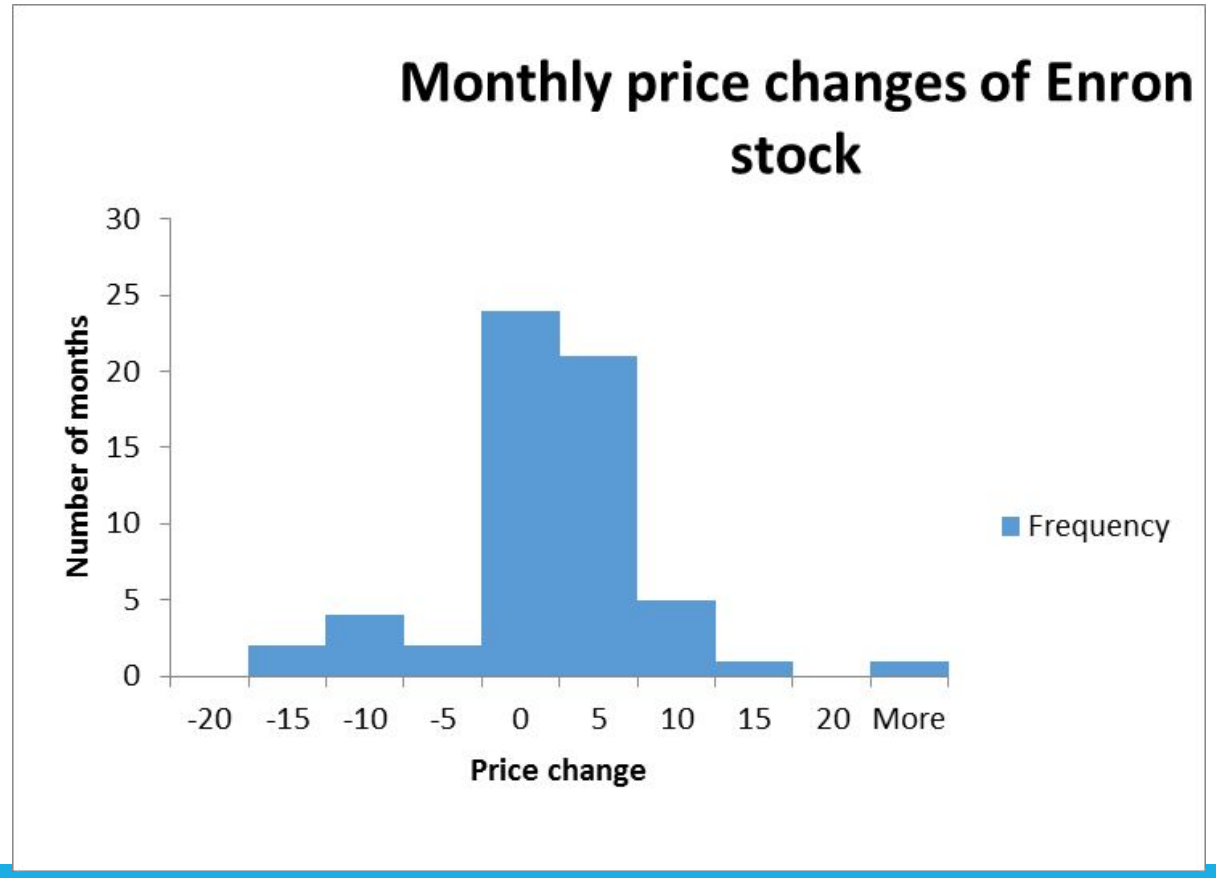
# Employee completion time Cumulative frequency

| *Intervals (sec.)* | *Frequency* | *Relative frequency* | *Cumulative %* | |
|---|---|---|---|---|
| 230 | 5 | 4.5% | 4.5% | 4.5 |
| 240 | 8 | 7.3% | 11.8% | 4.5 +7.3 = 11.8 |
| 250 | 13 | 11.8% | 23.6% | 11.8 + 11.8 = 23.6 |
| 260 | 22 | 20.0% | 43.6% | 23.6 + 20 = 43.6 |
| 270 | 32 | 29.1% | 72.7% | 43.6 + 29.1 = 72.7 |
| 280 | 13 | 11.8% | 84.5% | 72.7 + 11.8 = 84.5 |
| 290 | 10 | 9.1% | 93.6% | 84.5 + 9.1 = 93.6 |
| 300 | 7 | 6.4% | 100.0% | 93.6 + 6.4 = 100 |
| N = | 110 | | | |

# Histogram – Absolute frequency
# Enron: Change in stock price

| | *Price change* | *# of months* |
|---|---|---|
| | -20 | 0 |
| | -15 | 2 |
| | -10 | 4 |
| | -5 | 2 |
| | 0 | 24 |
| | 5 | 21 |
| | 10 | 5 |
| | 15 | 1 |
| | 20 | 0 |
| | More | 1 |

Frequency table for the price change of Enron stock



# Histogram – Relative frequency
# Enron: Change in stock price

| *Change in stock price* | *Relative frequencyin months* |
|---|---|
| -20 | 0 |
| -15 | 3.3 |
| -10 | 6.7 |
| -5 | 3.3 |
| 0 | 40 |
| 5 | 35 |
| 10 | 8.3 |
| 15 | 1.7 |
| 20 | 0 |
| More | 1.7 |
| | 100 |