3.4 The Five Number Summary;
Boxplot

→ Now, descriptive measures based on percentiles (advantage: resistant).

The most commonly used percentiles are quartiles (quantiles):

25th percentile

50th percentile

75th percentile
How do we calculate quartiles?

Let \( n \) denote the \# obs.
Arrange the data in increasing order.

• \( Q_1 \) is at position

• \( Q_2 \) is the median, which is at position

• \( Q_3 \) is at position

If a position is not a whole number,

Note: The textbook calculates percentiles differently. Follow class notes, which are consistent with software.
Ex. City A data: (ordered)

3  4  5  5  5  6  7

First quartile?

Observation at position

Choose

Third quartile?

Observation at position

Choose
Interquartile Range (IQR)

To avoid a single data value overly influencing the measure of dispersion,

IQR =

Ex. City A: $Q_1 = 4, Q_3 = 6$

$IQR_A =$

City B: $Q_1 = -4, Q_3 = 9$

$IQR_B =$

6 Data from City
The median, 1\textsuperscript{st} and 3\textsuperscript{rd} quartiles and smallest and largest observations are useful indicators of the dist’n of a data set.

→ display in

Boxplot - City A
Boxplots are effective for displaying several samples for visual comparison.

Ex.
Outliers

→ Use IQR to identify potential outliers.

Lower limit =
Upper limit =

Obs that lie outside the lower and upper limits are

Ex. The monthly rents -ordered- for 8 one-bedroom apartments, located in one area of the city are

<table>
<thead>
<tr>
<th>525</th>
<th>540</th>
<th>570</th>
<th>580</th>
</tr>
</thead>
<tbody>
<tr>
<td>585</td>
<td>585</td>
<td>625</td>
<td>770</td>
</tr>
</tbody>
</table>
Minitab output:

Descriptive Statistics: Rent

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rent</td>
<td>8</td>
<td>597.5</td>
<td>582.5</td>
<td>597.5</td>
<td>76.0</td>
<td>26.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rent</td>
<td>525.0</td>
<td>770.0</td>
<td>547.5</td>
<td>615.0</td>
</tr>
</tbody>
</table>

First quartile for rent data?
(Need to use linear interpolation)

Obs. at position \((n+1)/4\) =

Hence, \(Q_1\) is in between

\[ Q_1 = \]

↑

known as linear interpolation.

Check: \(Q_3 = 615!\)
\[ \text{IQR} = 615 - 547.5 = 67.5 \]

Lower limit = 
Upper limit =

↑

Useful to identify potential outliers.

Boxplot for rent data:
### 3.5 Descriptive Measures for Populations

<table>
<thead>
<tr>
<th>Statistics (Sample)</th>
<th>Parameters (Population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
</tr>
</tbody>
</table>

↑                     ↑

Note: Sample size: Popul’n size:
Popul’n Mean:

\[ \mu = \]

Popul’n Variance:

\[ \sigma^2 = \]

= 

Popul’n SD:

\[ \sigma = \]
Standardized Variables and z-Scores

For a variable $x$, the variable

$$z = \quad \text{is called the } \text{standardized variable}$$

corresponding to the variable $x$.

Important!!
Example 3.25 (pg 134)

Possible observations for the variable \( x \):

\[
\begin{array}{c}
\text{x} \\
\hline
-1 \\
3 \\
3 \\
3 \\
5 \\
5 \\
\end{array}
\]

a. Assuming the data set is popul’n data, compute the mean and SD.

b. Using your answer in part (a), determine the standardized version of \( x \).

c. Determine the observed value of \( z \) corresponding to an observed value of \( x \) of 5.

d. Obtain all possible observations of \( z \).

e. Find the mean and SD of \( z \) using the popul’n equations.

f. Obtain dotplots of the dist’ns of \( x \) and \( z \).
a. Popul’n mean:

\[ \mu_x = \]

Popul’n SD:

\[ \sigma_x = \sqrt{\frac{\sum x_i^2}{N}} \mu_x^2 \]

\[ \sum_{i=1}^{6} x_i^2 = \]

\[ \rightarrow \sigma_x = \sqrt{\quad} = \]

b. Standardized version of x:

\[ z = \]
c. When $x = 5$, $z =$

d. When $x = -1$, $z =$, and when $x = 3$, $z =$

Hence, the standardized value for each of the observations are:

$$x : \quad -1 \quad 3 \quad 3 \quad 3 \quad 5 \quad 5$$

$$\rightarrow \quad z :$$
e. Popul’n mean of standardized obs.:

\[ \mu_z = \]

Popul’n SD of standardized obs.:

\[ \sigma_z = \sqrt{\frac{\sum z_i^2}{6}} \mu_z^2 \]

\[ \sum_{i=1}^{6} z_i^2 = \]

\[ \rightarrow \sigma_z = \sqrt{\mu_z^2} = \]
f. Dotplot for $x$ and $z$ observations:

Dotplot: $x$, $z$

Already know that

Standardizing **shifts** a dist’n, so that the

Notice:

A lot easier to look at standardized observations!
PROBABILITY CONCEPTS
(Weiss, Chapter 4)

The science of uncertainty is called

4.1 Probability Basics

Experiment

- Is a process that results in one of a number of possible outcomes.
- The outcome that occurs cannot be predicted with certainty.

Example

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flip a coin</td>
<td></td>
</tr>
<tr>
<td>Roll a die</td>
<td></td>
</tr>
<tr>
<td>Dow Jones</td>
<td></td>
</tr>
</tbody>
</table>
Actual outcome can not be determined (predicted) in advance; but can

Definition - Probability

Probability of an event =
Meaning of Probability

Probability is

- a generalization of the

- a numerical measure of

\[
\begin{array}{ccc}
0 & 0.5 & 1 \\
\end{array}
\]

Properties of Probabilities

- 

- \( p = 1 \)

- \( p = 0 \)
Ex. As reported in *Employment and Earnings*, the age dist’n of employed persons 16 years old and over is

<table>
<thead>
<tr>
<th>Age (X)</th>
<th>Frequency (000's)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-19</td>
<td>6,500</td>
</tr>
<tr>
<td>20-24</td>
<td>12,138</td>
</tr>
<tr>
<td>25-34</td>
<td>32,077</td>
</tr>
<tr>
<td>35-44</td>
<td>35,051</td>
</tr>
<tr>
<td>45-54</td>
<td>25,514</td>
</tr>
<tr>
<td>55-64</td>
<td>11,739</td>
</tr>
<tr>
<td>65 &amp; over</td>
<td>3,690</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>126,709</strong></td>
</tr>
</tbody>
</table>
If an employed person is selected at random, find the probability that the person obtained is

a. Between 25 and 34 years old, inclusive.

\[ P(25 \leq X \leq 34) = \]

b. At least 45 years old.

\[ P(X \geq 45) = \]

\[ = 0.3231 \]
c. Between 20 and 34 years old, inclusive.

\[ P(20 \leq X \leq 34) = 0.3489 \]

\[ = 0.3489 \]

d. Under 20 or over 54.

\[ P(X < 20 \text{ or } X > 54) = 0.1731 \]