STOR 155 Introductory Statistics

Lecture 4:
Displaying Distributions with Numbers (II)
Numerical Summary for Distributions

• **Center**
  – Mean
  – Median
  – Mode

• **Spread**
  – Quartiles, IQR, Five-number summary and Boxplot
  – Standard Deviation (starting from page 14)
Examples: 2004 Two-Seater Cars

- Highway mileages of the 21 two-seater cars:
  13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32 66
- Q1 = 18
- Q3 = 28
- IQR = Q3 – Q1 = 10
- 1.5*IQR = 15
- Q3+1.5*IQR = 43
- Q1-1.5*IQR = 3
- 66 is a suspected outlier.
The five-number summary

- To get a quick summary of both center and spread, use the following five-number summary:
  
  Minimum  Q1  M  Q3  Maximum
Example: HWY Gas Mileage of 2004 Two-seater/Mini Cars

- **Two-seater**
  - Five-number summary:
    - 13, 18, 23, 27, 32

- **Mini-compact**
  - Five-number summary:
    - 19, 23, 26, 29, 32
Boxplots

• a visual representation of the five-number summary.

• A boxplot consists of
  – A central box spans the quartiles Q1 and Q3.
  – A line inside the box marks the median M.
  – Lines extend from the box out to the smallest and largest observations.
Boxplots of highway/city gas mileages (Two-seaters/minicompacts)
Pros and cons of Boxplots

• Location of the median line in the box indicates symmetry/asymmetry.
• Best used for side-by-side comparison of more than one distribution at a glance.
• Less detailed than histograms or stem plots.
• The box focuses attention on the central half of the data.
Income for different Education Level

![Box plot showing income for different education levels](image)

- No HS
- Some HS
- HS grad
- Some college
- Bachelor’s
- Higher degree

Income in thousands:

- 0
- 40,000
- 80,000
- 120,000
- 160,000
- 200,000
Modified Boxplot

- The current boxplot cannot reveal those possible outliers.

- To modify it,
  - the two lines extend out from the central box only to the smallest and largest observations that are not suspected outliers.
  - Observations more than 1.5*IQR outside the box are plotted as individual points.
## Table 1.1

<table>
<thead>
<tr>
<th>Call length (seconds) for calls to a customer service center</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
</tr>
<tr>
<td>126</td>
</tr>
<tr>
<td>372</td>
</tr>
<tr>
<td>179</td>
</tr>
<tr>
<td>89</td>
</tr>
<tr>
<td>148</td>
</tr>
<tr>
<td>67</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>700</td>
</tr>
<tr>
<td>121</td>
</tr>
</tbody>
</table>

Table 1-1

*Introduction to the Practice of Statistics, Fifth Edition*

© 2005 W.H. Freeman and Company
HG for count in a given time interval

7.6% of all calls are \( \leq 10 \) seconds long

Figure 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2003 W.H. Freeman and Company
Figure 1-18
*Introduction to the Practice of Statistics, Fifth Edition*
© 2005 W.H. Freeman and Company
Sample Variance $s^2$

- **Deviation from mean**: the difference between an observation and the sample mean:
  $$x_i - \bar{x}$$

- **Sample Variance $s^2$**: the average of squares of the deviations of the observations from their mean
  $$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$
  $$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$
Sample Standard Deviation $s$:

- **Sample Standard Deviation** $s$: the square root of the sample variance

\[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \]
Toy Examples

• Data:
  -2, -1, 0, 1, 2
• What is the sample variance and the standard deviation?
• How about this?
  40, 40, 40, 40, 40
Remarks on the definition of Standard Deviation (S.D.)

- The sum of the deviations of the obs from their mean is always 0.

- Why “square the deviations” rather than “absolute deviations”?
  - Mean is a natural center under the “squaring”.
  - S.D. is a natural measure of spread for the normal distributions.
Remarks on S.D.

• Why “S.D.” rather than “variance”?
  – S.D. is natural for measuring spread for normal dist.
  – S.D. is in the original scale.

• Why “n-1” rather than “n”?
  – Intuitively speaking, S.D. is not defined for n=1.
  – Sum of deviations is always 0, which means “if we
    know (n-1) of them, we know the last one”.
  – Only (n-1) deviations can change freely.
  – n-1: degrees of freedom.
Properties of the standard deviation (S.D.) $s$

- $s$ measures the spread about the mean;
- $s$ should be used only when the mean is chosen to measure the center;
- $s=0$ if and only if there is no spread;
  - When?
- $s>0$ almost always, increases with more spread;
- $s$, like the mean, is not resistant, i.e. sensitive to outliers.
Examples: 2004 Two-seater Cars

Highway mileages of the 21 two-seater cars:
13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32 66

• Gasoline-powered cars
  – Mean: 22.6
  – S.D.=5.3

• All cars
  – Mean: 24.7
  – S.D.=10.8
Three measures of spread

• The *range* is the spread of all the observations;

• The *interquartile range* is the spread of (roughly) the middle 50% of the observations;

• *S.D.* is a measure of the distance from sample mean. S.D. can be regarded as a “typical” distance of the observations from their mean.
The five-number summary vs Mean and S.D.

- The five-number summary is preferred for a skewed distribution or a distribution with strong outliers.
- $\bar{x}$ and $s$ are preferred for reasonably symmetric distributions that are free of outliers.

- Always plot your data first.
- Use boxplots.
Changing the unit of measurement

- The same variable can be recorded in different units of measurement.

- **Distance:**
  - Miles (US) vs Kilometers (Elsewhere)
  - 1 mile = 1.6 km
  - 1 km = ? mile

- **Temperature**
  - Fahrenheit (US) vs Celsius (Elsewhere)
  - 0 F = -17.8 C
  - 100 F = 37.8 C
  - 212 F = 100 C
Boiled Billy

• An Australian student Billy has recently been on a trip to the States. Soon after he arrived there, he caught a cold and had a fever.

• He went to see Doctor Z. Doctor Z measured his body temperature and told Billy, “Just relax! No big deal! It’s only a little above 100 degree!”

• “100!!!”, Billy yelled, “How can you say it’s not a big deal? I am boiled…”
Linear Transformation

• A linear transformation changes the original variable $x$ into a new variable $x_{new}$ according to the following equation,

\[ x_{new} = a + bx. \]

• Temperature: Celsius vs Fahrenheit
  – $x$ in Celsius, $x_{new}$ in Fahrenheit,

\[ x_{new} = 32 + \frac{9}{5} x. \]

  – How about the inverse transformation?
Effects of Linear Transformation

• The shape of a distribution remains unchanged, except that the direction of the skewness might change.
  – When?

• Measures of center and spread change.
  – Multiplying each obs by a positive number $b$ multiplies both measures of center and spread by $b$;
  – Adding the same number $a$ to each obs adds $a$ to measures of center and to percentiles, but does not change measures of spread.
Example: Salary Raise

• A sample was taken of the salaries of 20 employees of a large company. Suppose everyone will receive a $3000 increase, then
• how will the standard deviation of the salaries change?
• How about the mean?
• How about the median?
• How about Q1 and Q3?
What is the effect of \( x_{\text{new}} = a + bx \)?

- mean of \( X_{\text{new}} = a + b \) (mean of \( X \))
- median of \( X_{\text{new}} = a + b \) (median of \( X \))
- SD of \( X_{\text{new}} = |b| \) (SD of \( X \))
- Variance of \( X_{\text{new}} = b^2 \) (Variance of \( X \))
- IQR of \( X_{\text{new}} = |b| \) (IQR of \( X \))
Take Home Message

- Boxplot, modified boxplot, side-by-side boxplot
- Sample variance and sample standard deviation
- Remarks on the definition of S.D.
  - Why “squaring”?  
  - Why S.D. instead of Variance?  
  - Why n-1?  
- Properties of sample standard deviation
- 3 measures of spread
- Five-number summary vs Mean and S.D.
- Linear transformation and its effects on shape, center and spread