Summarizing and Displaying Measurement Data/Understanding and Comparing Distributions

Histograms, Mean, Median, Five-Number Summary and Boxplots, Standard Deviation
Thought Questions

1. If you were to read the results of a study showing that daily use of a certain exercise machine resulted in an average 10-pound weight loss, what more would you want to know about the numbers in addition to the average?

   (Hint: Do you think everyone who used the machine lost 10 pounds?)
Thought Questions

2. In a diet study, the weight loss of a particular subject was considered an “outlier.”

- What do you think this means?
Always start the summary of data by making a picture

- Histograms
  - First, slice up the entire span of values covered by the measurement variable into equal-width piles called bins.
  
  - The bins and the counts in each bin give the distribution of the quantitative variable.
Histogram for quantitative variable

- The height values are grouped into bins. X axis plots the bins.
- Y axis plots the frequency (aka counts).
- The height of each bar represents the number of the subjects whose heights fall into a specific bin, e.g, between 170cm and 173cm.
Bin width for histogram

Narrower bins, width<0.5cm

Wider bins, width=20cm
A histogram plots the bin counts as the heights of bars (like a bar chart). It displays the distribution at a glance. Here is a histogram of earthquake magnitudes:
A relative frequency histogram displays the percentage of cases in each bin instead of the count.

Here is a relative histogram of earthquake magnitudes:
Use of histogram

- The shape of the histogram tells how quantitative values are distributed.

1. What ranges of values occur more frequently and what ranges of value occur less frequently?
2. Whether the distribution is symmetric
3. The modality of the distribution: is there one major hump (unimodal) or more (bimodal or multimodal)
4. Check outliers: any data points far from the rest? Odd-balls or errors?
Various shapes of Histograms
Humps

1. Does the histogram have a single, central hump or several separated bumps?
   - Humps in a histogram are called **modes**.
   - A histogram with one main peak is dubbed **unimodal**; histograms with two peaks are **bimodal**; histograms with three or more peaks are called **multimodal**.
Humps

- A bimodal histogram has two apparent peaks:
Humps

- A histogram that doesn’t appear to have any mode and in which all the bars are approximately the same height is called **uniform**: 

![Histogram Example]

- **Counts**
- **X-axis Values**
Symmetry

2. Is the histogram symmetric?
   - If you can fold the histogram along a vertical line through the middle and have the edges match pretty closely, the histogram is symmetric.
Symmetry

- The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

- In the figure below, the histogram on the left is said to be skewed left, while the histogram on the right is said to be skewed right.
Anything Unusual?

3. Do any unusual features stick out?

- Sometimes it’s the unusual features that tell us something interesting or exciting about the data.
- You should always mention any stragglers, or **outliers**, that stand off away from the body of the distribution.
- Are there any **gaps** in the distribution? If so, we might have data from more than one group.
Anything Unusual?

- The following histogram has outliers—there are three cities in the leftmost bar:
There are many hundreds of useful tools—statistical methods—for analyzing data and drawing conclusions.

Like all tools, the effectiveness of the statistical methods depends on using them appropriately.

They are often concerned with summarizing data so that we can draw some conclusions without looking at the data in detail.

Examples of such tools of summarization are mean, median, standard deviation (a measure of the scatter, or dispersion, of the data)
Quantitative Data – Typical Value

- The most commonly used statistical summary measure is a typical value for a set of data.

- **Why would someone want a typical value for a set of data?**
  - An athlete might want to know the typical time for a particular knee injury to heal.
  - A researcher might want to know the average cholesterol reduction of a particular drug.
  - An investor might want to know the typical annual return of mutual funds in an industry sector.

- We also think of a typical value as a measure of **central tendency**, showing where the data tend to cluster.
Central Tendency – The Mean

- The data below are the annual salaries of 10 business executives (in thousands of dollars):

890
1,110
1,460
1,420
2,000
1,430
1,520
1,110
2,400
1,680

- The arithmetic mean, usually called the mean or the average, is the sum of all data values divided by the number of such values.

- In this case, the total for all the salaries is $15 million; divided by 10 you get a mean executive salary of $1.5 million.
Central Tendency – The Mean

The arithmetic mean has the most meaning when the values are closely centered, with few exceptional values and tending to symmetry about the mean.

Salary Example

- But suppose that the one executive who earned $1,460,000 has had a profit-sharing bonanza one year and earned $5 million more for a total salary of $6,460,000 instead of $1,460,000.

- While most of the executive salaries are still around $1.5 million and only one other makes more than $2 million, the mean has jumped from $1.5 million to $2 million, an increase in the value of the mean of more than 30%.
Central Tendency – The Mean

- Arithmetic mean of a set of observations $x_1, x_2, \ldots, x_n$

\[
\bar{x} = \frac{1}{n} (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

- Population mean $\mu$
- Sample mean $\bar{x}$ is an estimate of population mean $\mu$
The mean: pros and cons

Pros:
- Easy to understand and calculate
- Uses all observations
- Stable across different samples from the same population

Cons:
- Not applicable to qualitative variables
- Problems when some observations are missing
- VERY MUCH AFFECTED BY EXTREME VALUES
Central Tendency – The Median

The median is that value that about half the population have values below and half have values above.

Salary Example

- To get the value of the median, take all the numbers you have collected, and order them by increasing value.
- Once the numbers have been ordered, the median is the middle value (if the number of values is odd) or the average of the two middle values (if the number of values is even).
Central Tendency – The Median

- To get the median of the salaries, order the values as shown below:
  890
  1,110
  1,110
  1,420
  1,430
  1,460
  1,520
  1,680
  2,000
  2,400

- Then find the middle value (or as in this case, the average of the middle two values) to get a median executive salary of $1,445,000 ($1,430,000 + $1,460,000 divided by 2).
Central Tendency – The Median

**Salary Example**

- Note that in the original data set, the median of $1,445,000 is only a little less than the arithmetic mean $1.5 million.

- But when the one executive's $1,460,000 salary is increased to $6,460,000, the median does not change.

- At $1,445,000, the median is still typical of the executive salaries. The mean does change, however, and the new mean of $2 million is not a typical value.
The Median

- Order dataset values
  \[ x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k)} \]

- If number of observations in the dataset, \( k \), is odd, then
  \[ \tilde{x} = x_{\left(\frac{k+1}{2}\right)} \]

- If number of observations in the dataset, \( k \), is even, then
  \[ \tilde{x} = \frac{1}{2} \left( x_{\left(\frac{k}{2}\right)} + x_{\left(\frac{k}{2}+1\right)} \right) \]
The median: pros and cons

Pros
- Easily defined
- Easy to calculate
- Stable, not affected by outliers

Cons
- Stability has its downside – The median is not based on all observations
Mean and Median on a Histogram

![Histogram showing mean and median with positively skewed distribution.](image)

- Median
- Mean

Positively skewed

Mean = 29.17
Std. Dev. = 7.906
N = 94
What Is Middle Class in Manhattan? – NY Times, Jan 18\textsuperscript{th}, 2013

“The average sale price of a home in Manhattan last year was $1.46 million, according to a recent Douglas Elliman report, while the average sale price for a new home in the United States was just under $230,000.”

The average of $1.46 million they were referring to was the mean sales price.

The median sales price of a home in Manhattan was $837,500.
“The commissioner of the Food and Drug Administration on Friday revoked the approval of the drug Avastin as a treatment for breast cancer, ruling on an emotional issue that pitted the hopes of some desperate patients against the statistics of clinical trials.

Many breast cancer specialists say that Avastin does appear to work very well for some patients, and some advocates have said the drug should be left on the market for the sake of those patients.”
Question

In each of the following cases, would the mean or median probably be higher, or would they be about equal:

1. Ages at which residents of a sunburban city die, including everything from infant deaths to the most elderly
2. Heights of all 7-year-old children in a large city
3. Shoes sizes of adult women
How Spread Out is the Distribution?

- Variation matters, and Statistics is about variation.
- Are the values of the distribution tightly clustered around the center or more spread out?

- The **range** of the data is the difference between the maximum and minimum values:

  \[
  \text{Range} = \text{max} - \text{min}
  \]

- A disadvantage of the range is that a single extreme value can make it very large and, thus, not representative of the data overall.
Spread: The Interquartile Range

- The **interquartile range (IQR)** lets us ignore extreme data values and concentrate on the middle of the data.

- To find the IQR, we first need to know what quartiles are…

- The difference between the quartiles is the **interquartile range (IQR)**, so

\[ \text{IQR} = \text{upper quartile} - \text{lower quartile} \]
Spread: The Interquartile Range

- The lower and upper quartiles are the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles of the data, so...
- The IQR contains the middle 50\% of the values of the distribution, as shown in figure:
Example

Find Lower Quartile (Q1) and Upper Quartile (Q3):

Data: 850, 900, 1400, 1200, 1050, 1000, 750, 1250, 1050, 565

Order dataset: 565, 750, 850, 900, 1000, 1050, 1050, 1200, 1250, 1400

Q1: use left part of the data 565, 750, 850, 900, 1000
median of this part = Q1 = 850

Q3: use right part of the data 1050, 1050, 1200, 1250, 1400
median of this part = Q3 = 1200

IQR = 1200 – 850 = 350
The Five-Number Summary

Example: Systolic Blood Pressure

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>200</td>
</tr>
<tr>
<td>Q3</td>
<td>138</td>
</tr>
<tr>
<td>Median</td>
<td>132</td>
</tr>
<tr>
<td>Q1</td>
<td>121</td>
</tr>
<tr>
<td>Min</td>
<td>108</td>
</tr>
</tbody>
</table>

The **five-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum).
Systolic Blood Pressure (SBP) - BOXPLOT

Outlier is an observation that is located farther than 1.5 IQR from the closest quartile (Q1 or Q3). Outlier is extreme if it is more than 3 IQR from the closest quartile (Q1 or Q3).
11. **Heart attack stays.** The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospitals in New York during one year with a primary diagnosis of acute myocardial infarction (heart attack).

![Histogram of heart attack stays](image)

a) From the histogram, would you expect the mean or median to be larger? Explain.

b) Write a few sentences describing this distribution (shape, center, spread, unusual features).

c) Which summary statistics would you choose to summarize the center and spread in these data? Why?
Comparing Histograms and Boxplots

Compare the histogram and boxplot for daily wind speeds:

- Median: 1.90
- IQR: 1.78

Histogram showing the distribution of average wind speeds with bars indicating the frequency of days for each speed range. Boxplot on the right with median, interquartile range, and outliers marked.
Comparing Groups - Internal Radiation Exposure After the Fukushima Nuclear Power Plant Disaster

Figure. Histograms of Cesium Concentration (Bq/kg) in Exposed Children and Adults

Frequency of exposure measured in children in 1 Bq/kg increments and in adults in 2 Bq/kg increments. Nonexposed adults and children were excluded; therefore, the numbers of individuals included are 235 children and 3051 adults.
Comparing Groups - Relationship of Collegiate Football Experience and Concussion With Hippocampal Volume and Cognitive Outcomes

Figure 1. Smaller Hippocampal Volumes in Collegiate Football Athletes Relative to Healthy Controls
Alternative Measure of Spread: The Standard Deviation

- A more powerful measure of spread than the IQR is the standard deviation, which takes into account how far each data value is from the mean.

- A deviation is the distance that a data value is from the mean.
Alternative Measure of Spread: The Standard Deviation

- The **variance**, notated by $s^2$, is found by summing the squared deviations and (almost) averaging them:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

- The variance will play a role later in our study, but it is problematic as a measure of spread—it is measured in \textit{squared} units!
Alternative Measure of Spread: The Standard Deviation

- The **standard deviation**, \( s \), is just the square root of the variance and is measured in the same units as the original data.

\[
s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}
\]
Sample Distribution of Men's Height in U.S.
Sample Size = 200

Sample mean is an estimate of the population mean
Sample mean = sum of men's heights in sample divided by the sample size

Sample standard deviation is an estimate of the population standard deviation

Calculating the sample standard deviation

1. Find the mean.
2. Find the deviation of each value from the mean.
   Deviation = value – mean.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the sum by (the number of values) – 1, resulting in the variance.
5. Take the square root of the variance.
6. The result is the sample standard deviation.

Sample mean (\( \bar{x} \)) = 69.65
Sample standard deviation(s) = 3.36

We use a theoretical distribution, instead of the sample distribution. This allows us to answer certain questions using only the mean and standard deviation of the data.
## Sample Distribution of Men’s Height

Calculating the Sample Standard Deviation

<table>
<thead>
<tr>
<th>value (height in inches)</th>
<th>Deviation = value – mean</th>
<th>Square the Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.00</td>
<td>-9.65</td>
<td>93.1225</td>
</tr>
<tr>
<td>60.00</td>
<td>-9.65</td>
<td>93.1225</td>
</tr>
<tr>
<td>61.00</td>
<td>-8.65</td>
<td>74.8225</td>
</tr>
<tr>
<td>62.00</td>
<td>-7.65</td>
<td>58.5225</td>
</tr>
<tr>
<td>62.00</td>
<td>-7.65</td>
<td>58.5225</td>
</tr>
<tr>
<td>63.00</td>
<td>-6.65</td>
<td>44.2225</td>
</tr>
<tr>
<td>63.00</td>
<td>-6.65</td>
<td>44.2225</td>
</tr>
<tr>
<td>63.00</td>
<td>-6.65</td>
<td>44.2225</td>
</tr>
<tr>
<td>64.00</td>
<td>-5.65</td>
<td>31.9225</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>76.00</td>
<td>6.35</td>
<td>40.3225</td>
</tr>
<tr>
<td>76.00</td>
<td>6.35</td>
<td>40.3225</td>
</tr>
<tr>
<td>77.00</td>
<td>7.35</td>
<td>54.0225</td>
</tr>
<tr>
<td>77.00</td>
<td>7.35</td>
<td>54.0225</td>
</tr>
<tr>
<td>78.00</td>
<td>8.35</td>
<td>69.7225</td>
</tr>
<tr>
<td><strong>0.00</strong></td>
<td><strong>0.00</strong></td>
<td><strong>2243.5</strong></td>
</tr>
</tbody>
</table>
Example: Metabolic Rates

The following data consist of the metabolic rates (cal./24hr.) of 7 men from a dieting study:

\[ 1792 \quad 1666 \quad 1362 \quad 1614 \quad 1460 \quad 1867 \quad 1439 \]

First, compute the sample mean:

\[ \bar{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \]

\[ = \frac{11,200}{7} \]

\[ = 1600 \]
Example: Metabolic Rates

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations</th>
<th>Squared deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1792</td>
<td>1792–1600 = 192</td>
<td>(192)^2 = 36,864</td>
</tr>
<tr>
<td>1666</td>
<td>1666–1600 = 66</td>
<td>(66)^2 = 4,356</td>
</tr>
<tr>
<td>1362</td>
<td>1362–1600 = -238</td>
<td>(-238)^2 = 56,644</td>
</tr>
<tr>
<td>1614</td>
<td>1614–1600 = 14</td>
<td>(14)^2 = 196</td>
</tr>
<tr>
<td>1460</td>
<td>1460–1600 = -140</td>
<td>(-140)^2 = 19,600</td>
</tr>
<tr>
<td>1867</td>
<td>1867–1600 = 267</td>
<td>(267)^2 = 71,289</td>
</tr>
<tr>
<td>1439</td>
<td>1439–1600 = -161</td>
<td>(-161)^2 = 25,921</td>
</tr>
<tr>
<td>sum</td>
<td>0</td>
<td>sum = 214,870</td>
</tr>
</tbody>
</table>
Example: Metabolic Rates

\[ s^2 = \frac{214,870}{7-1} = 35,811.67 \]

\[ s = \sqrt{35,811.67} = 189.24 \text{ calories} \]
Sample variance $s^2 = 11.9359/(11-1) = 1.19359$

Sample standard deviation $s = (1.19359)^{1/2} = 1.0925 \text{ mm}$
ods graphics on;

title "Demonstrating PROC UNIVARIATE";
proc univariate data=example.Blood_Pressure;
   id Subj;
   var SBP DBP;
   histogram;
   probplot / normal(mu=est sigma=est);
run;
The UNIVARIATE Procedure
Variable: SBP

Moments

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>56</td>
</tr>
<tr>
<td>Sum Weights</td>
<td>56</td>
</tr>
<tr>
<td>Mean</td>
<td>130.535714</td>
</tr>
<tr>
<td>Sum Observations</td>
<td>7310</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>10.911152</td>
</tr>
<tr>
<td>Variance</td>
<td>119.053247</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.1155991</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.5354758</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>960764</td>
</tr>
<tr>
<td>Corrected SS</td>
<td>6547.92857</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>8.35874876</td>
</tr>
<tr>
<td>Std Error Mean</td>
<td>1.45806407</td>
</tr>
</tbody>
</table>

Basic Statistical Measures

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>130.5357</td>
</tr>
<tr>
<td>Median</td>
<td>132.0000</td>
</tr>
<tr>
<td>Mode</td>
<td>134.0000</td>
</tr>
<tr>
<td>Variability</td>
<td></td>
</tr>
<tr>
<td>Std Deviation</td>
<td>10.91115</td>
</tr>
<tr>
<td>Variance</td>
<td>119.05325</td>
</tr>
<tr>
<td>Range</td>
<td>48.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>17.00000</td>
</tr>
</tbody>
</table>

Tests for Location: Mu0=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student's t</td>
<td>t</td>
<td>Pr &gt;</td>
</tr>
<tr>
<td>Sign</td>
<td>M</td>
<td>Pr &gt;=</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S</td>
<td>Pr &gt;=</td>
</tr>
</tbody>
</table>

Quantiles (Definition 5)

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>156</td>
</tr>
<tr>
<td>99%</td>
<td>156</td>
</tr>
<tr>
<td>95%</td>
<td>148</td>
</tr>
<tr>
<td>90%</td>
<td>144</td>
</tr>
<tr>
<td>75% Q3</td>
<td>138</td>
</tr>
<tr>
<td>50% Median</td>
<td>132</td>
</tr>
<tr>
<td>25% Q1</td>
<td>121</td>
</tr>
<tr>
<td>10%</td>
<td>114</td>
</tr>
<tr>
<td>5%</td>
<td>112</td>
</tr>
<tr>
<td>1%</td>
<td>108</td>
</tr>
<tr>
<td>0% Min</td>
<td>108</td>
</tr>
</tbody>
</table>
PROC UNIVARIATE - HISTOGRAM

Distribution of SBP

<table>
<thead>
<tr>
<th>SBP</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>116</td>
<td>10</td>
</tr>
<tr>
<td>124</td>
<td>15</td>
</tr>
<tr>
<td>132</td>
<td>25</td>
</tr>
<tr>
<td>140</td>
<td>25</td>
</tr>
<tr>
<td>140</td>
<td>10</td>
</tr>
<tr>
<td>156</td>
<td>5</td>
</tr>
</tbody>
</table>
title "Demonstrating MIDPOINT= Histogram Option";
proc univariate data=example.Blood_Pressure;
   id Subj;
   var SBP;
   histogram / normal midpoints=100 to 170 by 5;
   probplot / normal(mu=est sigma=est);
run;
PROC UNIVARIATE - HISTOGRAM

Distribution of SBP

Curve  Normal(Mu=130.54 Sigma=10.911)