Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 6 = More of chapter 3

Agenda:
1) Announce midterm exam (Thursday, July 26)
2) Lecture over more of chapter 3
   (sections 3.3 and 3.2)
Announcement – Midterm Exam:
The midterm exam will be Thursday, July 26

The best thing will be to take it in the classroom (9:00-10:15 AM in Terman 156)

For remote students who absolutely can not come to the classroom that day please email me to confirm arrangements with SCPD

You are allowed one 8.5 x 11 inch sheet (front and back) for notes

No books or computers are allowed, but please bring a hand held calculator

The exam will cover the material that we covered in class from Chapters 1,2,3 and 6
Homework Assignment:

Chapter 3 Homework Part 1 is due Tuesday 7/17

Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

SCPD students may use email or fax or mail.

The assignment is posted at
Chapter 3: Exploring Data
Exploring Data

- We can explore data visually (using tables or graphs) or numerically (using summary statistics)

- Section 3.2 deals with summary statistics

- Section 3.3 deals with visualization

- We will begin with visualization

- Note that many of the techniques you use to explore data are also useful for presenting data
Boxplots (Pages 114-115)

- Invented by J. Tukey

- A simple summary of the distribution of the data

- Boxplots are useful for comparing distributions of multiple attributes or the same attribute for different groups


**Boxplots in R**

- The function `boxplot()` in R plots boxplots

- By default, `boxplot()` in R plots the maximum and the minimum (if they are not outliers) instead of the 10\(^{th}\) and 90\(^{th}\) percentiles as the book describes
Boxplots help you visualize the differences in the medians relative to the variation.

Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a “big” difference?
Boxplots help you visualize the differences in the medians relative to the variation.

Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a “big” difference?

Maybe yes:
Boxplots (Pages 114-115)

- Boxplots help you visualize the differences in the medians relative to the variation.

- Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a “big” difference?

Maybe yes:

Example boxplot showing a smaller difference.

Maybe no:

Example boxplot showing a larger difference.
In class exercise #16:
Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at
www.stats202.com/exams_and_names.csv
In class exercise #16:
Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at
www.stats202.com/exams_and_names.csv

Answer:

data<-read.csv("exams_and_names.csv")

boxplot(data[,2],data[,3],col="blue",
main="Exam Scores",
names=c("Exam 1","Exam 2"),ylab="Exam Score")
In class exercise #16:
Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at www.stats202.com/exams_and_names.csv

Answer:
Visualization in Excel

- Up until now, we have done all the visualization in R.

- Excel also can make many different types of graphs. They are found under the “Insert” menu by selecting “Chart.”

- When using Excel to make graphs which anyone will see other than yourself, I strongly encourage you to change defaults such as the grey background.

- Excel also has a nice tool for making tables and associated graphs called “PivotTable and PivotChart Report” under the “Data” menu.
In class exercise #17:
Use “Insert” > “Chart” > “XY Scatter” to make a scatter plot of the exam scores at
www.stats202.com/exams_and_names.csv
Put Exam 1 on the X axis and Exam 2 on the Y axis.
In class exercise #17:
Use “Insert” > “Chart” > “XY Scatter” to make a scatter plot of the exam scores at
www.stats202.com/exams_and_names.csv
Put Exam 1 on the X axis and Exam 2 on the Y axis.

Answer:
In class exercise #18:
The data [www.stats202.com/more_stats202_logs.txt](http://www.stats202.com/more_stats202_logs.txt) contains access logs from May 7, 2007 to July 1, 2007. Use “Data” > “PivotTable and PivotChart Report” In Excel to make a table with the counts of
GET /lecture2=start-chapter-2.ppt HTTP/1.1 and
GET /lecture2=start-chapter-2.pdf HTTP/1.1 for each date. Which is more popular?
In class exercise #18:
The data [www.stats202.com/more_stats202_logs.txt](http://www.stats202.com/more_stats202_logs.txt) contains access logs from May 7, 2007 to July 1, 2007. Use “Data” > “PivotTable and PivotChart Report” in Excel to make a table with the counts of GET /lecture2=start-chapter-2.ppt HTTP/1.1 and GET /lecture2=start-chapter-2.pdf HTTP/1.1 for each date. Which is more popular?

Answer:

<table>
<thead>
<tr>
<th>Date</th>
<th>GET /lecture2=start-chapter-2.pdf HTTP/1.1</th>
<th>GET /lecture2=start-chapter-2.ppt HTTP/1.1</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-Jun-07</td>
<td>150</td>
<td>17</td>
<td>167</td>
</tr>
<tr>
<td>28-Jun-07</td>
<td>247</td>
<td>29</td>
<td>276</td>
</tr>
<tr>
<td>29-Jun-07</td>
<td>253</td>
<td>53</td>
<td>306</td>
</tr>
<tr>
<td>30-Jun-07</td>
<td>77</td>
<td>9</td>
<td>86</td>
</tr>
<tr>
<td>1-Jul-07</td>
<td>50</td>
<td>7</td>
<td>57</td>
</tr>
<tr>
<td>Grand Total</td>
<td>777</td>
<td>115</td>
<td>892</td>
</tr>
</tbody>
</table>
In class exercise #19:
The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Use “Data” > “PivotTable and PivotChart Report” In Excel to make a table with the counts of the rows for each date in May.
In class exercise #19:

The data [www.stats202.com/more_stats202_logs.txt](http://www.stats202.com/more_stats202_logs.txt) contains access logs from May 7, 2007 to July 1, 2007. Use “Data” > “PivotTable and PivotChart Report” In Excel to make a table with the counts of the rows for each date in May.

<table>
<thead>
<tr>
<th>Date</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>May-7</td>
<td>88</td>
</tr>
<tr>
<td>May-8</td>
<td>88</td>
</tr>
<tr>
<td>May-9</td>
<td>65</td>
</tr>
<tr>
<td>May-10</td>
<td>179</td>
</tr>
<tr>
<td>May-11</td>
<td>47</td>
</tr>
<tr>
<td>May-12</td>
<td>67</td>
</tr>
<tr>
<td>May-13</td>
<td>47</td>
</tr>
<tr>
<td>May-14</td>
<td>59</td>
</tr>
<tr>
<td>May-15</td>
<td>58</td>
</tr>
<tr>
<td>May-16</td>
<td>107</td>
</tr>
<tr>
<td>May-17</td>
<td>64</td>
</tr>
<tr>
<td>May-18</td>
<td>93</td>
</tr>
<tr>
<td>May-19</td>
<td>66</td>
</tr>
<tr>
<td>May-20</td>
<td>104</td>
</tr>
<tr>
<td>May-21</td>
<td>123</td>
</tr>
<tr>
<td>May-22</td>
<td>75</td>
</tr>
<tr>
<td>May-23</td>
<td>85</td>
</tr>
<tr>
<td>May-24</td>
<td>81</td>
</tr>
<tr>
<td>May-25</td>
<td>49</td>
</tr>
<tr>
<td>May-26</td>
<td>60</td>
</tr>
<tr>
<td>May-27</td>
<td>78</td>
</tr>
<tr>
<td>May-28</td>
<td>66</td>
</tr>
<tr>
<td>May-29</td>
<td>64</td>
</tr>
<tr>
<td>May-30</td>
<td>69</td>
</tr>
<tr>
<td>May-31</td>
<td>46</td>
</tr>
</tbody>
</table>
In class exercise #20:
Use “Insert” > “Chart” > “Line” In Excel to make a graph on the number of rows versus the date for the previous exercise.
In class exercise #20:
Use “Insert” > “Chart” > “Line” In Excel to make a graph on the number of rows versus the date for the previous exercise.

Answer:
Using Color in Plots

- In R, the graphing parameter “col” can often be used to specify different colors for points, lines etc.

- Some advantages of color:
  - provides a nice way to differentiate
  - makes it more interesting to look at

- Some disadvantages of color:
  - Some people are color blind
  - Most printing is in black and white
  - Color can be distracting
  - A poor color scheme can make the graph difficult to read (example: yellow lines in Excel)
3-Dimensional Plots

- 3D plots can sometimes be useful

- One example is the 3D scatter plot for plotting 3 attributes (page 119)

- The function `scatterplot3d()` makes fairly nice 3D scatter plots in R
  - this is not in the base package so you need to do:
    ```r
    install.packages("scatterplot3d")
    library(scatterplot3d)
    ```

- However, it may be better to show the 3rd dimension by simply using a 2D plot with different plotting characters (page 119)
3-Dimesional Plots

Never use the 3\textsuperscript{rd} dimension in a manner that conveys no extra information just to make the plot look more impressive
3-Dimesional Plots

- Never use the 3rd dimension in a manner that conveys no extra information just to make the plot look more impressive

Examples:
In class exercise #21:
Not only does the 3rd dimension fail to provide any information in the previous two examples, but it can also distort the truth. How?
Do’s and Don’ts (Page 130)

- Read the ACCENT Principles
- Read Tufte’s Guidelines
Compressing Vertical Axis

Bad Presentation

Good Presentation

Quarterly Sales

Quarterly Sales

$0

$200

$100

$0

Q1 Q2 Q3 Q4

Q1 Q2 Q3 Q4

$0

$50

$25

$0

Q1 Q2 Q3 Q4
No Zero Point On Vertical Axis

Bad Presentation

Good Presentations

Graphing the first six months of sales
No Relative Basis

Bad Presentation
A’s received by students.

Good Presentation
A’s received by students.

FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior
Chart Junk

Bad Presentation

Minimum Wage

1960: $1.00
1970: $1.60
1980: $3.10
1990: $3.80

Good Presentation

Minimum Wage


$
Final Touches

- Many times plots are difficult to read or unattractive because people do not take the time to learn how to adjust default values for font size, font type, color schemes, margin size, plotting characters, etc.

- In R, the function `par()` controls a lot of these

- Also in R, the command `expression()` can produce subscripts and Greek letters in the text
  - example: `xlab=expression(alpha[1])`

- In Excel, it is often difficult to get exactly what you want, but you can usually improve upon the default values
Exploring Data

- We can explore data visually (using tables or graphs) or numerically (using summary statistics).

- Section 3.2 deals with summary statistics.

- Section 3.3 deals with visualization.

- We will begin with visualization.

- Note that many of the techniques you use to explore data are also useful for presenting data.
Summary Statistics (Section 3.2, Page 98):

You should be familiar with the following elementary summary statistics:

- Measures of Location: Percentiles (page 100)
  - Mean (page 101)
  - Median (page 101)

- Measures of Spread: Range (page 102)
  - Variance (page 103)
  - Standard Deviation (page 103)
  - Interquartile Range (page 103)

- Measures of Association: Covariance (page 104)
  - Correlation (page 104)
Measures of Location

- Terminology: the “mean” is the average

- Terminology: the “median” is the 50\textsuperscript{th} percentile

- Your book classifies only the mean and median as measures of location but not percentiles

- More commonly, all three are thought of as measures of location and the mean and median are more specifically measures of center

- Terminology: the 1\textsuperscript{st}, 2\textsuperscript{nd} and 3\textsuperscript{rd} quartiles are the 25\textsuperscript{th}, 50\textsuperscript{th} and 75\textsuperscript{th} percentiles respectively
Mean vs. Median

- While both are measures of center, the median is sometimes preferred over the mean because it is more robust to outliers (=extreme observations) and skewness.

- If the data is right-skewed, the mean will be greater than the median.

- If the data is left-skewed, the mean will be smaller than the median.

- If the data is symmetric, the mean will be equal to the median.
San Francisco median home price tops $600K

Housing sales up even while area struggles to emerge from high-tech slump

REUTERS

Updated: 6:14 p.m. ET July 16, 2005

SAN FRANCISCO - The median price of a home sold in the San Francisco Bay area topped $600,000 for the first time in June amid a near record month of sales, analysts said Tuesday.

The median price paid for a home in the region rose to $610,000 in June, up 18.2 percent from a year earlier and 2.5 percent from May, according to real estate information service DataQuick Information Systems.
Measures of Spread:

- The *range* is the maximum minus the minimum. This is not robust and is extremely sensitive to outliers.

- The *variance* is
  \[ \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} \]
  where \( n \) is the sample size and \( \bar{X} \) is the sample mean. This is also not very robust to outliers.

- The *standard deviation* is simply the square root of the variance. It is on the scale of the original data. It is roughly the average distance from the mean.

- The *interquartile range* is the 3\(^{rd}\) quartile minus the 1\(^{st}\) quartile. This is quite robust to outliers.
In class exercise #22:
Compute the standard deviation for this data by hand:

2 10 22 43 18

Confirm that R and Excel give the same values.
Measures of Association:

- The **covariance** between x and y is defined as
  \[
  \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}
  \]
  where \( \bar{X} \) is the mean of x and \( \bar{Y} \) is the mean of y and \( n \) is the sample size. This will be positive if x and y have a positive relationship and negative if they have a negative relationship.

- The **correlation** is the covariance divided by the product of the two standard deviations. It will be between -1 and +1 inclusive. It is often denoted \( r \). It is sometimes called the coefficient of correlation.

- These are both very sensitive to outliers.
Correlation ($r$):

- $r = -1$
- $r = -0.6$
- $r = +1$
- $r = +0.3$
In class exercise #23:
Match each plot with its correct coefficient of correlation.

Choices: $r=-3.20, r=-0.98, r=0.86, r=0.95, r=1.20, r=-0.96, r=-0.40$
In class exercise #24:
Make two vectors of length 1,000,000 in R using `runif(1000000)` and compute the coefficient of correlation using `cor()`. Does the resulting value surprise you?
In class exercise #25:
What value of $r$ would you expect for the two exam scores in www.stats202.com/exams_and_names.csv which are plotted below. Compute the value to check your intuition.