More on Variability

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

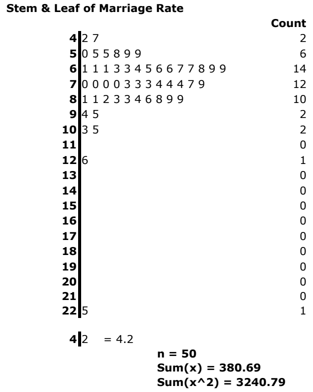
- Continue the discussion of the variance and standard deviation
- Introduce the Coefficient of Variation (CV)
- Revisit Box Plots
- The variance of a proportion
- A brief introduction to Covariance

The Variance is sensitive to outliers

- The variance and the standard deviation are very sensitive to outliers (extreme values)
- When you square large numbers you get much larger numbers
  - $5^2 = 25$
  - $500^2 = 250,000$
- Look what happens when we remove Nevada from the Marriage Rate data

Marriage Data without Nevada

- **Calculate the Variance/Standard Deviation**
- $s^2 = [3240.79 - (380.69)^2/50]/(50-1)$
- $s^2 = [3240.79 - 2898.50]/(49)$
- $s^2 = [342.29]/(49)$
- $\mathbf{s^2 = 6.99}$
- $\mathbf{s = 2.64}$



```
Stem & Leaf of Marriage Rate
                                          Count
 4|2 7                                        2
 5|0 5 5 8 9 9                                6
 6|1 1 1 3 3 4 5 6 6 7 7 8 9 9               14
 7|0 0 0 0 3 3 3 4 4 4 7 9                   12
 8|1 1 2 3 3 4 6 8 9 9                       10
 9|4 5                                        2
10|3 5                                        2
11|                                           0
12|6                                          1
13|                                           0
14|                                           0
15|                                           0
16|                                           0
17|                                           0
18|                                           0
19|                                           0
20|                                           0
21|                                           0
22|5                                          1

 4|2   = 4.2
                  n = 50
                  Sum(x) = 380.69
                  Sum(x^2) = 3240.79
```

Comparisons with and without Nevada

Statistic	W Nevada	W/O Nevada
Sum(x)	441.73	380.69
Sum(x^2)	6967.24	3240.79
Mean	8.66	7.31
Median	7.02	7.00
Mode	7.00	7.00
Minimum	4.20	4.20
Maximum	61.00	22.50
Range	56.80	18.30
Variance	62.82	6.99
Std Dev	7.93	2.64

Excel Commands for Measures of Central Tendency and Variance

Sum	=SUM(B5:B104)	3,699.40
Count	=COUNT(B5:B104)	100.00
Mean	=AVERAGE(B5:b104)	36.99
Minimum	=MIN(B5:B104)	30.00
Maximum	=MAX(B5:B104)	44.90
Median	=MEDIAN(B5:B104)	37.00
Mode	=MODE(B5:B104)	37.00
Range	**subtract the max and min**	14.90
First Quartile	=QUARTILE(B5:B104,1)	35.68
Third Quartile	=QUARTILE(B5:B104,3)	38.33
Inter-Quartile Range	**subtract Q3 minus Q1**	2.65
Variance	=VAR(B5:B104)	5.85
Std Deviation	=STDEV(B5:B104)	2.42

Descriptive Statistics of Marriage Rate data using Excel

- **In Office 2003**
  - Tools
  - Data Analysis
  - Descriptive Statistics
- **In Office 2007**
  - Data
  - Data Analysis
  - Descriptive Statistics

Marriage Rate	
Mean	8.66
Standard Error	1.11
Median	7.02
Mode	#N/A
Standard Deviation	7.93
Sample Variance	62.82
Kurtosis	40.05
Skewness	6.10
Range	56.85
Minimum	4.19
Maximum	61.04
Sum	441.73
Count	51

The Standard Deviation and the Range

- A quick approximation for the standard deviation is the range divided by 4
- It is a crude approximation, but in a symmetric, mound-shaped distribution, it is reasonable
- For the marriage rate
  - With Nevada 56.80/4 = 14.2 compared with 7.93
  - Without Nevada 18.30/4 = 4.58 compared with 2.64
- **It is just an approximation!!!**

Coefficient of Variation

- The Coefficient of Variation
- The ratio of the standard deviation to the absolute value of the mean,
- usually expressed as a percentage (multiply by 100)
- By taking a ratio, we express the std dev relative to the mean
- For the Marriage Rate data, the **CV = 7.93/8.66* 100 = 91.57**

$$CV = \frac{s}{|\bar{x}|}$$

The higher the CV, the more variability in the variable.

Let's work out a complete example: Body Mass Index for 24 subjects

- **Mean**
- **Median**
- **Mode**
- **Minimum**
- **Maximum**
- **Range**
- **Variance**
- **Standard Deviation**
- **CV**

BMI Stem and Leaf Plot

STEM	LEAF
17	7
18	
19	6 6
20	6
21	4 5
22	0 7
23	2 5 8 8
24	5 6
25	2 2 4
26	
27	5 8
28	1
29	1 9
30	
31	4
32	
33	5

Stem is the whole number

Sum X	593.6
Sum X^2	15040.4

Let's work out a complete example: Body Mass Index for 24 subjects

- **Mean = 593.6/24 = 24.73**
- **Median is average of 12th and 13th positions = (23.8 +24.5)/2 = 24.15**
- **Mode is either 19.6, 23.8, or 25.2**
- **Minimum = 17.7**
- **Maximum = 33.5**
- **Range = 33.5 - 17.7 = 15.80**
- **Variance = (15040.4-$(593.6)^2$/24)/23 = 15.60**
- **Standard Deviation = SQRT(15.60) = 3.95**
- **CV = 3.95/24.73*100 = 15.97**

BMI Stem and Leaf Plot

STEM	LEAF
17	7
18	
19	6 6
20	6
21	4 5
22	0 7
23	2 5 8 8
24	5 6
25	2 2 4
26	
27	5 8
28	1
29	1 9
30	
31	4
32	
33	5

Stem is the whole number

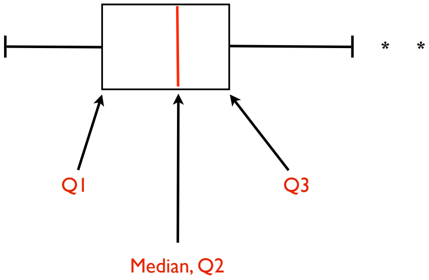
Sum X	593.6
Sum X^2	15040.4

Let's Revisit Box Plots

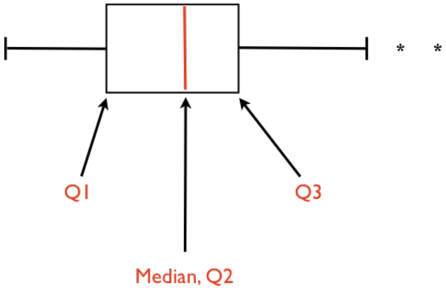
- Box plots are a way to show the distribution of a variable relative to the median, showing shape, skew and outliers
- Box plots highlight extreme values in data
- Can be graphed for a small or large sample size
- Five number summary
  - Minimum
  - Q1
  - Median
  - Q3
  - Maximum
- This gives us the extremes, the middle, the range, and the Inter-Quartile Range

Box Plot Fundamentals

Let's look at what a Box Plot is, step by step.



Box Plots often appear horizontal



Stem and Leaf Plot of Health Data - the WEIGHT of the subject

Statistics	
Sample Size	80
Mean	159.385
Median	161
Std. Deviation	34.87689
Minimum	94.3
Maximum	255.9

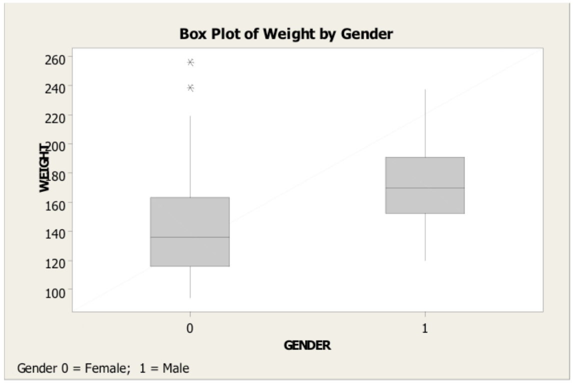
```
Stem & Leaf Plot of Weight

Stem unit:  10

 9| 4
10| 1 5 7 8 9
11| 0 2 3 5 9 9
12| 3 4 4 4 7
13| 0 0 5 5 6 6 7 7 9
14| 1 4 4 8 9
15| 0 1 1 2 3 6 6
16| 0 0 2 2 2 3 3 4 5 6 7 9
17| 0 0 3 3 4 5 5 6 7 9
18| 0 1 2 7 8 9
19| 1 4 6 8
20| 1 5 9
21| 3 4 9
22| 1
23| 7 8
24|
25| 6
```

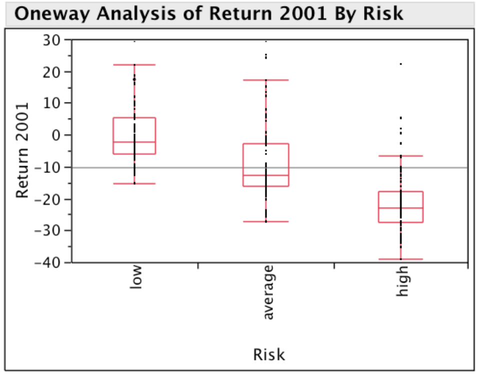
This is Bi-Modal

Mintab and JMP Box Plots



Gender 0 = Female; 1 = Male

Box Plots of Stock Returns by Level of Risk



Excluded Rows 21

Dealing with the Mean and Variance of a Proportion

- Sometimes our data deals with a dichotomous variable
  - Yes or No
  - Male or Female
  - Treatment or Control
- If we code the variable as a zero/one dichotomy, it is called a **dummy variable.**
- The mean of the dummy variable is the **proportion** of the attribute coded as one
- And the variance is very easy to compute

Coding Strategy, Let 1=Yes, 0 = No

- Just to be clear, this is what I mean by using a coding strategy
- I will code the response as dummy variable
  - 1 = Yes
  - 0 = No

Do you support candidate A?	
Response	**Code**
Yes	1
Yes	1
No	0
No	0
Yes	1
No	0
No	0
No	0
Yes	1
No	0

Proportions

Do you support	
Response	**Code**
Yes	1
Yes	1
No	0
No	0
Yes	1
No	0
No	0
No	0
Yes	1
No	0

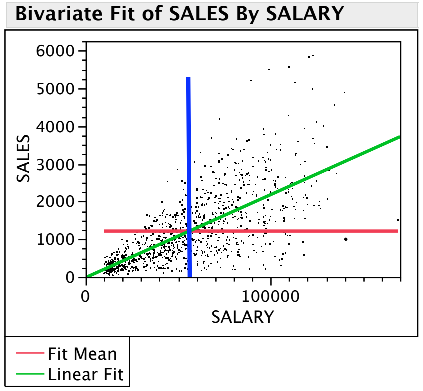
- Let **p** = Number of Successes/Total
  - Example: # Yes/n = 4/10 = .4
- And **q** = (**1-p**)
- The mean = **p**
  - **Σx/n = (1+1+0+0+1+0+0+0+1+0)/10 = .4**
- The variance of a proportion is given by
  - s^2 = **p*q**
  - s = **(p*q)**$^{.5}$
  - **s^2 = .4*.6 = .24   s = .4899**

Covariance

- Covariance looks at how two variables vary about their means together, on average (divided by n)
- The Variance is the covariance of a variable with itself!

$$Cov_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n}$$

The Scatterplot is a picture of Covariance



Summary

- The Variance and the Standard Deviation are influenced by outliers
- The Coefficient of Variation allows us to compare the variability of different variables
- Box and Whisker Plots allow us to see the spread of data and compare different groups
- Proportions via dummy variables
- Covariance is related to the variance - it shows how two variables vary about their means together