

Numerical Descriptive Measures for Quantitative data III

Dr. Tom Ilvento
FREC 408

Interpreting the Standard Deviation

- We can use the standard deviation to express the proportion of cases that might fall within one or 2 standard deviations from the mean.
- We can use two theorems to help
 - **Chebyshev's Rule** (Tchebysheff's theorem in book, p148)
 - **Empirical Rule** (p148)

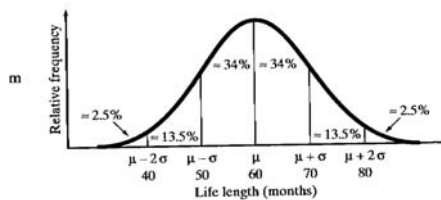
Chebyshev's Rule (Tchebysheff's Theorem P148)

- Is based on a mathematical theorem for any data
- At least $\frac{3}{4}$ of the measurements will fall within ± 2 standard deviations from the mean
- At least $\frac{8}{9}$ of the measurements will fall within ± 3 standard deviations from the mean

Empirical Rule (P148)

- Based on a symmetrical distribution where the mean, median, and the mode are similar – the EPA mpg data fits this

A Symmetrical Curve



Empirical Rule

- Approximately 68% of the measurements will be ± 1 standard deviation from the mean
- Approximately 95% of the cases fall between ± 2 standard deviations from the mean

Empirical Rule cont.

- Approximately 99.7% of the cases will fall within ± 3 standard deviations from the mean
 - This means it will be very rare to be more than 3 standard deviations from the mean when dealing with a symmetrical distribution

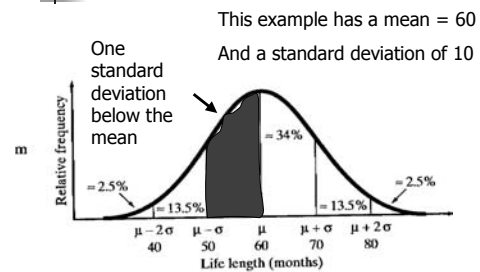
Empirical Rule

- For the EPA mpg data we would expect that 68% of the cases would fall between
 - 36.99 ± 2.42 or
 - Between **34.57 to 38.41**

Empirical Rule and EPA mpg Data

- 1 Std Dev**
 - 36.99 ± 2.42 34.57 to 38.41
- 2 Std Dev**
 - 36.99 ± 4.84 32.15 to 41.83
- 3 Std Dev**
 - 36.99 ± 7.26 29.73 to 44.25

A Symmetrical Curve



Auto Batteries Example

- Grade A Battery **Average Life** is **60 Months**
- Guarantee is for 36 months
- Standard Deviation **s = 10 months**
- Frequency distribution is mound-shaped and symmetrical

Battery example

- What percent of the Grade A Batteries will last **more than 50 months?**
 - Start with finding how many standard deviations 50 months is from the mean
 - Draw it out
 - Figure out the probability from the Empirical Rule



Battery example

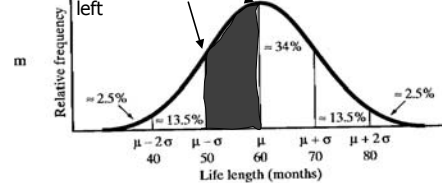


- 50 months is one standard deviation to the left of the mean
- This represents **34%** of the cases
- Because ± 1 std deviation = 68%, so -1 std deviation = **34%**
- To the right of the mean (60 months or more) represents **50%** of the cases
- Answer: $34 + 50 = 84\%$**

Battery Example – more than 50 months

With a mean = 60 and $s = 10$

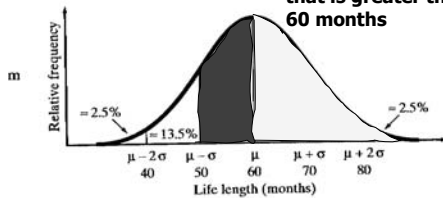
Here's the part that is one std deviation to the left



Battery Example – more than 50 months

With a mean = 60 and $s = 10$

And here's the part that is greater than 60 months



Battery example



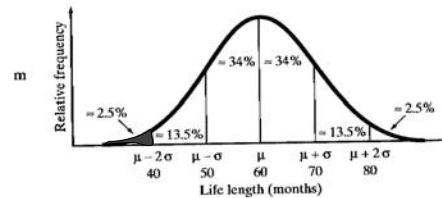
- Approximately what percentage of the batteries will last less than 40 months?
 - Start with finding how many standard deviations 40 months is from the mean
 - Draw it out
 - Figure out the probability

Battery Example

- 40 is 2 standard deviations** from the mean
- ± 2 standard deviations = 95% of the cases
- So, less than 40 is $\frac{1}{2}$ of the 5% remaining
- So it represents **2.5%** of the cases

Battery Example – less than 40 months

With a mean = 60 and $s = 10$



Battery Example

- Suppose your battery lasted 37 months. What could you infer about the manufacturer's claim?



Battery Example – 37 months

- **37 months is more than 2 standard deviations** from the mean
- **Less than 2.5%** of the batteries would fail within 37 months if the claims were true
- **It's possible you just got a bad one...do you feel lucky?**
- **Or unlucky??????**

Z-Scores

- This is a method of transforming the data to reflect relative standing of the value
- We subtract the mean and divide by the standard deviation

$$z_i = \frac{(x_i - \bar{x})}{s}$$

Z-Scores

- The result represents the distance between a given measurement x and the mean, expressed in standard deviations
 - **distance between a value and the mean**
 - **expressed in standard deviations**

Z-Scores

- A positive z-score means that that measurement is larger than the mean
- A negative z-score means that it is smaller than the mean

Demonstration of z-score

- EPA MPG Data
 - Mean = 37 (rounded off)
 - $s = 2.4$
 - One value is 34.0
 - z-score is
 - $(34.0 - 37.0)/2.4 = -1.25$
 - **This value of 34 is 1.25 standard deviations below the mean**

You try it

- Create a z-score for the following values (mean = 37, s = 2.4)
- 30
- 42
- 38

Z-Scores

- If we were to convert an entire variable to z-scores...
 - This means create a new variable by taking each value, subtracting the mean, and dividing by the standard deviation
- This is called a **data transformation**
- The new variable would have
 - **Mean = 0**
 - **Standard deviation = 1**

Empirical Rule and Z-Scores

- Approximately **68%** of the measurements will have a z-score **between -1 and 1**
- Approximately **95%** of the measurements will have a z-score **between -2 and 2**
- **Almost all** the measurements (99.7%) will have a z-score **between -3 and 3**

Data Example

- A female bank employee believes her salary is low as a result of sex discrimination. **Her salary is \$27,000**
- She collects information on salaries of male counterparts. **Their mean salary is \$34,000** with a **standard deviation of \$2,000.**
- Does this information support her claim?

How to begin to examine this issue

- What is her salary in relation to the mean male salary?
- Create a z-score for her salary to see how far below the mean her salary is in standard deviations

Solve for the z-score

$$z = \frac{\$27,000 - \$34,000}{\$2,000} = -3.5$$

Rare-Event Approach

- Her salary is 3.5 standard deviations below that of her male counterparts
- If her salary is part of the same distribution as the males in her bank, a value of -3.5 would be very rare

Rare Event Approach

- Perhaps her salary does not come from the same distribution, and we might conclude there is something different about her salary
- One conclusion could be discrimination
- But it could also be related to **performance, or time on the job, or some other factors**

Rare Event Approach

- **What if the woman's salary was only 1 standard deviation below her male counterparts?**

The Rare Event Approach

- We hypothesize a frequency distribution to describe a population of measurements
- We draw a sample from the population
- Compare the sample statistic to the hypothesized frequency distribution
- And see how **likely or unlikely** the sample came from the hypothesized distribution

Box Plots

- The book covers quartiles and box plots on page 158 and page 162
- I want you to look this material over, but I won't require you draw a box plot
- Box plots are a way to show the distribution of a variable relative to the median
- Box plots highlight extreme values in data

Box Plots and 5 number summary

- Five number summary
 - Lowest
 - Q1
 - Median
 - Q3
 - Highest number
- This gives us the extremes, the middle, the range, and the Inter-Quartile Range

Standard Box Plot

- The box is proportional to the data and has the median in the middle, and Q 1 and Q3 on either end

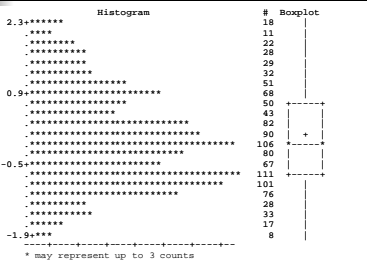


- Plus whiskers that go to the two extreme values

Modified Box Plot

- A more advanced Box Plot use the Inter-Quartile Range to construct an Inner and Outer Fence
 - Inner Fence = $1.5 \times \text{IQR}$
 - Outer Fence = $3 \times \text{IQR}$
- To better identify mild and extreme outliers

SAS will do a Stem & Leaf (or Histogram) and a Box Plot



SAS Univariate Example

Measures based on the mean

Measures based on the median and position

Extreme Values

```

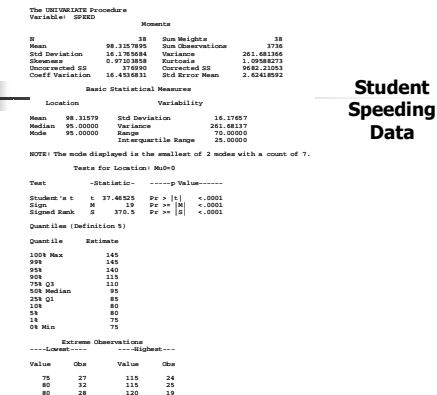
The SAS System
Univariate Procedure
Variable= Poultry Grower Satisfaction

Moments
N          1151  Sum Wgts    1151
Mean       0          Sum       0
Std Dev    0.941405  Variance    0.886244
Skewness   0.410211  Kurtosis   -0.53125
USS        1019.18  CSS        1019.18
CV         0.027748  Std Mean   1.027748
T:Mean=0   0  Pr>|T|    1.0000
M:Sum=0    0  Pr>|M|    0.524
N(Sign)    -51.5  Pr=|S|    0.0026
Sign Rank  -15772  Pr=|S|    0.1821
W:Normal    0.954257  Pr=0       0.0001

Quantiles(Def=5)
100% Max  2.248864  99%  2.248864
75% Q3    0.643514  95%  1.712166
50% Med   -0.099024  90%  1.35261
25% Q1    -0.728338  10%  -1.099005
0% Min    -1.81793    5%  -1.43535

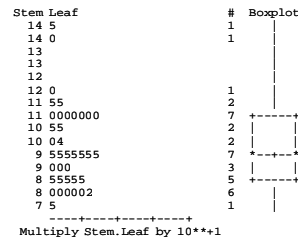
Range     4.06659   1%  -1.71886
Q3-Q1    1.37189
Mode      -1.00449

Extremes
Lowest   Obs      Highest  Obs
-1.81793  831  2.248864  1262
-1.81793  814  2.248864  1005
-1.81793  793  2.248864  1124
-1.81793  501  2.248864  1127
-1.81793  431  2.248864  1202
    
```



Student Speeding Data

Student Speeding Data



Multiply Stem.Leaf by 10***1

Alternative Stem and Leaf

Speed Stem and Leaf for Speed	
Stem unit: 10	
7	5
8	0 0 0 0 2 5 5 5 5
9	0 0 5 5 5 5 5 5
10	0 4 5 5
11	0 0 0 0 0 0 5 5
12	0
13	
14	0 5

Exam I example

Length of Hospital Stay Data	
0	2 2 3 3 3
0+	4 4 4 4 4 4 5 5 5 5 5
0++	6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7
0++++	8 8 8 8 9 9 9 9 9
1	0 1 1
1+	2
1++	5

The stems break the tens digit into parts designated by using "+" where 0++ stands for no tens, values of 4 and 5.

Sum of $x = 327$	Sum of $x^2 = 2477$	
Q2 = 6	n = 50	

Calculate:

1. Mean
2. Standard Deviation
3. Median
4. Mode
5. Z-score for a value of 15

Exam I example

Length of Hospital Stay Data		Calculate:
0	2 2 3 3 3	1. Mean = $327/50 = 6.54$
0+	4 4 4 4 4 4 5 5 5 5 5	2. Std Dev =
0++	6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7	$[(2477 - (327^2/50))/49]^{.5} = 2.63$
0++++	8 8 8 8 9 9 9 9 9	1. Median = Q2 = 6
1	0 1 1	2. Mode = 6
1+	2	3. Z-score for a value of 15 =
1++	5	$(15 - 6.54)/2.63 = 3.22$

The stems break the tens digit into parts designated where 0++ stands for no tens, values of 4 and 5.

Sum of $x = 327$	Sum of $x^2 = 2477$	
Q2 = 6	n = 50	

Alternative Stem and Leaf

Stem-and-Leaf Display for Length	
Stem unit: 1	
2	0 0
3	0 0 0
4	0 0 0 0 0 0
5	0 0 0 0 0
6	0 0 0 0 0 0 0 0 0
7	0 0 0 0 0 0 0 0
8	0 0 0 0
9	0 0 0 0 0
10	
11	0
12	
13	
14	
15	0