The Empirical Rule, z-Scores, and the Rare Event Approach

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

Overview

- Look at Chebyshev's Rule and the Empirical Rule
- Explore some applications of the Empirical Rule
- How to calculate and use z-scores
- Introducing the "Rare Event" strategy for inference

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**
  - **Empirical Rule**

Chebyshev's Rule

- Based on a mathematical theorem for any data, regardless of the distribution of the variable.
- The percentage of observations that are contained within distances of k standard deviations around the mean must be:
  - $(1-1/k^2) * 100\%$
  - Example: k=2 $(1-1/2^2) *100 = 75\%$
  - At least 3/4 of the measurements will fall within ± 2 standard deviations from the mean
- At least 8/9 (88.89%) of the measurements will fall within ± 3 standard deviations from the mean

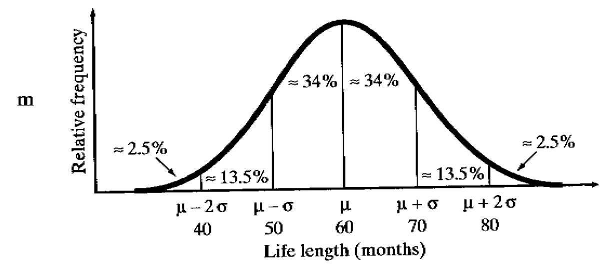
The Empirical Rule

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

Stem-and-Leaf Display for MPG
Stem unit: Whole number

30	0
31	8
32	5 7 9 9
33	1 2 6 8 9 9
34	0 2 4 5 8 8
35	0 1 2 3 5 6 6 7 8 9 9
36	0 1 2 3 3 4 4 5 5 6 6 7 7 7 8 8 8 9 9 9
37	0 0 0 0 1 1 1 2 2 3 3 4 4 5 6 6 7 7 8 9 9
38	0 1 2 2 3 4 5 6 7 8
39	0 0 3 4 5 7 8 9
40	0 1 2 3 5 5 7
41	0 0 2
42	1
43	
44	9

A Symmetrical Mound-Shaped Distribution



Empirical Rule

- Approximately **68%** of the measurements will be **± 1 standard deviations** from the mean
- Approximately **95%** of the cases fall between **± 2 standard deviations** from the mean
- Approximately **99.7%** of the cases will fall within **± 3 standard deviations** from the mean

MPG Car Data

MPG	
Mean	36.99
Standard Error	0.24
Median	37.00
Mode	37.00
Standard Deviation	2.42
Sample Variance	5.85
Coefficient of Variation	6.54%
Kurtosis	0.77
Skewness	0.05
Range	14.90
Minimum	30.00
Maximum	44.90
Sum	3699.40
Count	100

Stem-and-Leaf Display for MPG
Stem unit: Whole number

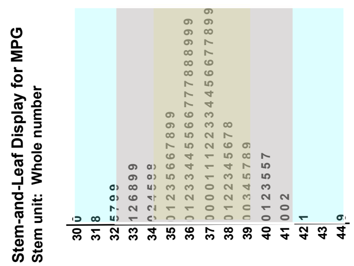
30	0
31	8
32	5 7 9 9
33	1 2 6 8 9 9
34	0 2 4 5 8 8
35	0 1 2 3 5 6 6 7 8 9 9
36	0 1 2 3 3 4 4 5 5 6 6 7 7 7 8 8 8 9 9 9
37	0 0 0 0 1 1 1 2 2 3 3 4 4 5 6 6 7 7 8 9 9
38	0 1 2 2 3 4 5 6 7 8
39	0 0 3 4 5 7 8 9
40	0 1 2 3 5 5 7
41	0 0 2
42	1
43	
44	9

MPG Data

We would expect that 68% of the values would fall between

- **36.99 ± 2.42**
- **34.57 to 39.41**

- We expect that 95% of the values would fall between
  - **36.99 ± 2*2.42**
  - **32.15 to 41.83**
- We expect that 99% of the values would fall between
  - **36.99 ± 3*2.42**
  - **29.73 to 44.25**

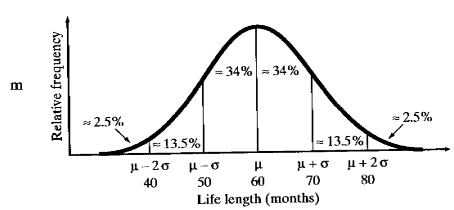


Auto Batteries Example

- 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
- What percent of the Grade A Batteries will last more than 50 months?

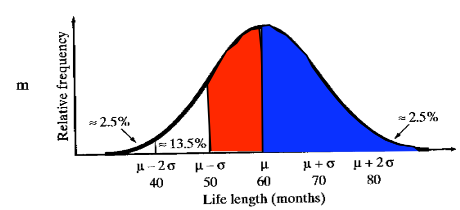
What percent of the Grade A Batteries will last more than 50 months?

- Start with finding how many standard deviations 50 months is from the mean
- Draw it out
- Figure out the probability from the Empirical Rule



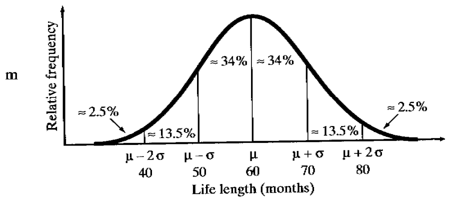
What percent of the Grade A Batteries will last more than 50 months?

- 50 months is one standard deviation to the left of the mean - **(60-10) = 50**
- 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%



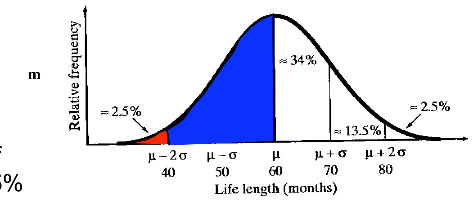
Approximately what percentage of the batteries will last less than 40 months?

- Start with finding how many standard deviations 40 months is from the mean
- Draw it out
- Figure out the probability

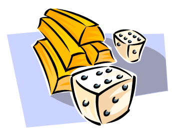


What percent of the Grade A Batteries will last less than 40 months?

- 40 is 2 standard deviations from the mean, and ± 2 standard deviations = 95% of the cases
- I am interested in the part less than 40 months
- ½ of the 95% for two standard deviations, the left hand side of the distribution, is equal to 47.5%
- But I want the part in the left hand tail of the distribution
- Which is 50% - 47.5% = 2.5%
- So it represents 2.5% of the cases



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



- 37 months is more than 2 standard deviations below 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
- **For a value, X, Subtract the mean and divide by the std dev**
- The result represents the distance between a given measurement X and the mean, expressed in standard deviations
  - A positive z-score means that the measurement is larger than the mean
  - A negative z-score means that it is smaller than the mean

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

z-score example from MPG data

- Mean = 36.99
- s = 2.42
- One value is 33.2
- The z-score is **(33.2-36.99)/2.42 = -1.57**
- This value of -1.57 is **1.57 standard deviations below the mean**

Create a z-score for the following values for a variable with a mean = 2.0007 and s = .0446

- 1.894
- 2.05
- 2.11

- **z = (1.894 - 2.0007)/.0446 = -2.392**
- **z = (2.050 - 2.0007)/.0446 = 1.105**
- **z = (2.110 - 2.0007/.0446 = 2.45**

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**

A 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 think about this problem?

- What is her salary in relation to the mean male salary?
- To find out, calculate a z-score for her salary to see how far below the mean her salary is in standard deviations
- 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

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

The Rare-Event Approach

- This approach is called a **Rare-Event Approach**
  - Express the problem in terms of a known distribution - the distribution of males
  - And see how rare it was to observe the value you have - the woman's salary
- Based on a z-score of -3.5, we might doubt that her salary comes 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

The Rare Event Approach

- **What if the woman's salary was only one standard deviation below the mean?**
  - That would not be such an unusual thing
  - In the distribution were symmetrical, we might expect 34% to fall between the mean and one standard deviation below.
  - And more importantly, 16% would be more than one standard deviation below the mean

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
- The decision would focus on how many standard deviations our sample statistic is from the hypothesized value

A definition of an outlier

- One way to determine what is an outlier is to calculate the z-score
- If a value is more than three standard deviations from the mean, it is relatively far away
- Later, we will express this in a probability framework via the normal or t-distribution
- For now we will say that any value that is more than three standard deviations from the mean is unusual, and an outlier.

Summary

- This concludes the basic description section of the course
- From graphs to central tendency to variability – all are ways to describe data
- The z-score approach is a way to express a data point as being so many standard deviations from the mean
- The rare event approach is away to start to make inferences – do you lucky?