Module B: Knowing Your Data

Slide Deck B4:

Measures of Uncertainty and Spread

The section in which we learn how to calculate the the variation in the data. This section adds to the previous by providing a measurement of how well the measure of center actually summarizes the data.

Today's Objectives

By the end of this slidedeck, you should

1. calculate the following measures of spread
    - variance and standard deviation (s^2 and s)
    - interquartile range (IQR)
    - coefficient of variation (c_v)
2. determine if the data are "sufficiently skewed" using
    - Hildebrand ratio, H
3. determine what information each measure provides about the data
4. understand the empirical rule and Chebyshev's inequality

Framing Example

Example

I would like to indicate how much my data varies. I should do this because variability gives me insight into the uncertainty of future data values.

Here are my data (scores on a six-point geography quiz):

$$0, 3, 0, 2, 5, 4, 2, 4, 4, 0, 4, 2, 2, 0, 1, 0, 0, 2$$

- What is the typical value?
- How variable are the data?

```
source("https://rfs.kvasaheim.com/stat200.R")
dt = read.csv("https://rfs.kvasaheim.com/data/geography.csv")
attach(dt)
```

Framing Example

When describing the data, one needs to already know something about the data. For instance, what do we know about these data?

- The data are numeric.
- The data are ratio level.

So, which measure of center should be used?

- The mean is 1.944444 `mean(Score)`
- The median is 2.0 `median(Score)`
- the mode is 0 `modal(Score)`

Which of the three is best?

- They all tell a part of the story of the data. The mode is the most frequent/likely value. The median divides the data into two (roughly) equal parts. The mean gives the average score (center of gravity).

Framing Example

Ultimately, we are asked about the variation in the data (uncertainty in a new value; spread of the data).

As expected, there are many ways of measuring variation. They depend on our choices of the measure of center and what we want to do with the measure of spread:

• mean	variance, standard deviation, coefficient of variation
• median	interquartile range

Framing Example

Here are the values, as calculated by `R`:

• variance	2.9967	`var(Score)`
• standard deviation	1.7311	`sd(Score)`
• coefficient of variation	0.8903	`cv(Score)`
• interquartile range	3.75	`IQR(Score)`

Note that the calculations are easy to do using the computer. The remaining question is:

What do these numbers tell us?

Variance

The **variance**, s^2, is an average distance from the values to the center.

- Since it requires the mean, it should only be used if the mean is meaningful
- Mathematicians prefer it because variances add
- Its formula results from its definition:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

`var`

Standard Deviation

The **standard deviation**, s, is also an average distance from the values to the center.

- Since it requires the mean, it should only be used if the mean is meaningful
- Preferred when conveying information because it has the same units as the data
- Its formula results from its definition:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

One more advantage to the standard deviation is...

- if the data are bell-shaped, then about 68% of the data are within one standard deviation of the mean (see: Empirical Rule).

`sd`

Coefficient of Variation

The **coefficient of variation**, c_v, is also an average distance from the values to the center. However, it is scaled by the mean of the data.

- Since it requires the mean, it should only be used if the mean is meaningful
- Preferred when comparing variability between two different variables
- Its formula results from its definition:

$$c_v = \frac{s}{\bar{x}}$$

CV

Interquartile Range

The **interquartile range**, IQR, is the range of the middle 50% of the data. It is calculated as the difference between the first and third quartiles.

- Since it requires the median, it should only be used for numeric data
- Its formula results from its definition:

$$IQR = Q_3 - Q_1$$

IQR

Hildebrand Ratio

The **Hildebrand ratio**, H, is a scaled *measure of skewness* of the data. It relies on the difference between the mean and the median, as scaled by the standard deviation.

- Since it requires the median, it should only be used for numeric data
- Its formula results from its definition:

$$H = \frac{\bar{x} - \tilde{x}}{s}$$

This is used to determine if data are too skewed to use mean-based measures:

- if $H \geq 0.20$, then the data are skewed positive (right)
- if $H \leq -0.20$, then the data are skewed negative (left)
- otherwise, the data are sufficiently symmetric

```
hildebrand.rule
```

Example 1: Geography Quiz

For our first example, let us examine a geography quiz I gave to a previous class.

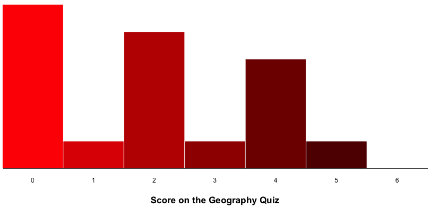
Example

The data are the `geography` quiz data. Calculate and interpret the measures of center and measures of spread.

```
source("https://rfs.kvasaheim.com/stat200.R")
dt = read.csv("https://rfs.kvasaheim.com/data/geography.csv")
attach(dt)

barplot(Score)
```

Example 1: Geography Quiz



Example 1: Geography Quiz

Measures of Center:

- The mean is 1.944 `mean(Score)`
- The median is 2.000 `median(Score)`
- The mode is 0 `modal(Score)`

Measures of Spread:

- The standard deviation is 1.7311 `sd(Score)`
- The IQR is 3.75 `IQR(Score)`
- The cv is 0.8903 `cv(Score)`

Measure of Skewness:

- The Hildebrand ratio is -0.03 `hildebrand.rule(Score)`

Example 2: School Enrollment in 1990

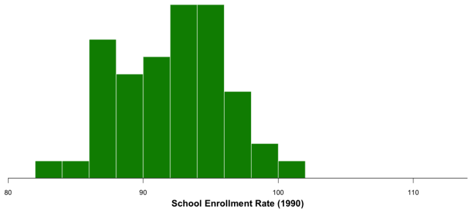
Example

The data are the crime data. The variable of interest is the school enrollment percentages for 1990 (enroll90). Calculate and interpret the measures of center and spread.

```
source("https://rfs.kvasaheim.com/stat200.R")
dt = read.csv("https://rfs.kvasaheim.com/data/crime.csv")
attach(dt)

hist(enroll90)
```

Example 2: School Enrollment in 1990



Example 2: School Enrollment in 1990

Measures of Center:

- The mean is 92.06 `mean(enroll90)`
- The median is 92.60 `median(enroll90)`
- The mode is 92.60 `modal(enroll90)`

Measures of Spread:

- The standard deviation is 3.92 `sd(enroll90)`
- The IQR is 5.55 `IQR(enroll90)`
- the cv is 0.43 `cv(enroll90)`

Measure of Skewness:

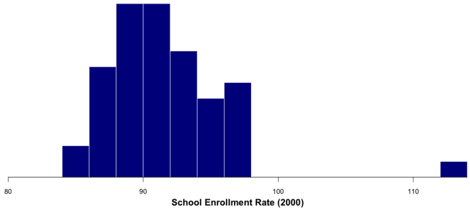
- The Hildebrand ratio is H = -0.14 `hildebrand.rule(enroll90)`

Example 3: School Enrollment in 2000

Example

The data are the `crime` data. The variable of interest is the school enrollment percentages for 2000 (`enroll00`). Calculate and interpret the measures of center and spread.

Example 3: School Enrollment in 2000



Example 3: School Enrollment in 2000

Measures of Center:

- The mean is 91.75 `mean(enroll00)`
- The median is 91.00 `median(enroll00)`
- These data are multimodal `modal(enroll00)`

Measures of Spread:

- The standard deviation is 4.53 `sd(enroll00)`
- The IQR is 5.10 `IQR(enroll00)`
- the cv is 0.05 `cv(enroll00)`

Measure of Skewness:

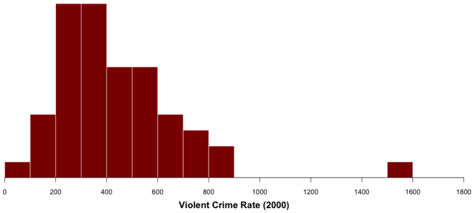
- The Hildebrand ratio is $H = 0.165$ `hildebrand.rule(enroll00)`

Example 4: Violent Crime Rate in 2000

Example

The data are the `crime` data. The variable of interest is the violent crime rate in 2000 (`vcrime00`). Calculate and interpret the measures of center and spread.

Example 4: Violent Crime Rate in 2000



Example 4: Violent Crime Rate in 2000

Measures of Center:

- The mean is 441.6 `mean(vcrime00)`
- The median is 383.8 `median(vcrime00)`

Measures of Spread:

- The standard deviation is 241.45 `sd(vcrime00)`
- The IQR is 268.25 `IQR(vcrime00)`
- the cv is 0.547 `cv(vcrime00)`

Measure of Skewness:

- The Hildebrand ratio is H = 0.239 `hildebrand.rule(vcrime00)`

The Empirical Rule

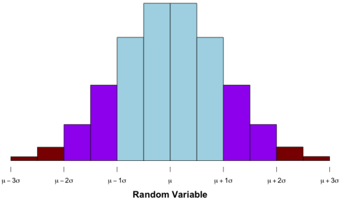
The standard deviation gives us information about intervals that contain "a certain amount" of the data. One of the most useful is the **Empirical Rule**. According to the empirical rule, *if* the data are bell-shaped, then the following approximations hold.

percent of the data	is within	that is
approximately 68%	one standard deviation of the mean	between $\mu - 1\sigma$ and $\mu + 1\sigma$
approximately 95%	two standard deviations of the mean	between $\mu - 2\sigma$ and $\mu + 2\sigma$
approximately 99.7%	three standard deviation of the mean	between $\mu - 3\sigma$ and $\mu + 3\sigma$

Note: These are approximations.

The Empirical Rule

Graphical illustration:



The Empirical Rule Results

Let us see how the data we looked at today fits with the empirical rule. The following gives the theoretical and the actual proportion of the data within *one* standard deviation of the mean.

Variable	Empirical Rule	Actual
`Score`	0.68	0.39
`enroll90`	0.68	0.63
`enroll00`	0.68	0.78
`vcrime00`	0.68	0.76

Note: As expected, the approximation is better for data distributions that are more bell shaped. The less bell-shaped, the worse the approximation.

Chebyshev's Inequality

While the empirical rule gives approximate bounds for specified proportions of the data, Chebyshev's inequality gives absolute bounds (bounds that must be met). According to the Chebyshev's inequality, at least

$$1 - \frac{1}{k^2}$$

of the data are within k standard deviations of the mean. For instance:

percent of the data	is within	that is
at least 0%	one standard deviation of the mean	between $\mu - 1\sigma$ and $\mu + 1\sigma$
at least 75%	two standard deviations of the mean	between $\mu - 2\sigma$ and $\mu + 2\sigma$
at least 88.89%	three standard deviation of the mean	between $\mu - 3\sigma$ and $\mu + 3\sigma$

Note: These are guaranteed bounds.

Chebyshev's Inequality Results

Let us see how the data we looked at today fits with the empirical rule. The following gives the theoretical and the actual proportion of the data within *two* standard deviation of the mean.

Variable	Empirical Rule	Chebyshev	Actual
`Score`	0.95	0.75	1.00
`enroll90`	0.95	0.75	0.96
`enroll00`	0.95	0.75	0.98
`vcrime00`	0.95	0.75	0.98

Note: Again, the Chebyshev bounds are guaranteed (at least 75% of the data are within two standard deviations of the mean). The empirical rule bounds are approximate (approximately 95% of the data are within two standard deviations of the mean).

Take-Aways about These Approximations

Here are some things to take away from this discussion.

- The empirical rule is an approximation
- Chebyshev's inequality is guaranteed

- The empirical rule says *approximately* 95% of the data are within 2σ of μ
- Chebyshev's inequality says *at least* 75% of the data are within 2σ of μ

- The empirical rule is better when the data are bell-shaped
- Chebyshev's inequality is guaranteed, regardless of the data distribution

- The empirical rule is useful in applied statistics
- Chebyshev's inequality is useful in proofs

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. calculate the following measures of spread
   - variance and standard deviation (s^2 and s)
   - interquartile range (IQR)
   - coefficient of variation (c_v)

2. determine if the data are "sufficiently skewed" using:
   - Hildebrand ratio, H

3. determine what information each measure provides about the data

4. understand the empirical rule and Chebyshev's inequality

Today's R Functions

In this slide deck, we saw the following R functions:

- `var`
- `sd`
- `cv`
- `IQR`
- `hildebrand.rule`

In the script accompanying this slidedeck, we used

- `isBetween`

Supplemental Activities

The following activities may be of interest to you in terms of today's topics:

- SCA 2a is for measures of center
- SCA 2b is for measures of position
- SCA 2c is for measures of **spread**

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of measures of position (and calculating them in `R`):

Hawkes Learning:	Section 3.2
Intro to Modern Statistics:	Chapter 5
`R` for Starters:	Section 4.3

- Wikipedia: Empirical rule
- Wikipedia: Chebyshev's inequality