Math 3070, Applied Statistics

Section 1

August 21, 2019

Lecture Outline, 8/21

- 1.2 Outliers
- 1.3 Measures of Location (Center)
- 1.4 Measures of Spread (Variability)

Outliers, Section 1.2

Outlier: Value which is significantly different from other observations. Mathematical definitions may vary.

Example: 8,3,2,5,4,2,3,5,3,2,1,5,7,7,8,8,9,99,101

Outliers: 99 and 101

Questions?

Measures of Location and Spread, Preface

- **Population parameters**: describe features of the population. Examples: population mean μ, variance σ^2 and standard deviation σ.
- **Sample statistics** : describe features of a sample. Examples: sample mean $\overline{x}$, variance s^2 and standard deviation s.

Measures of Location (Section 1.3)

Goal: Find the center or "middle" of the data.

Tools: **Sample Mean**, **Sample Median** and **Trimmed Mean**

Consideration: outliers

Sample Mean, Definition

Given a set of data denoted as $x_1, x_2, \ldots, x_n$.
Note, sample size $= n$.

Sample Mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

Sample Mean, Example (no outlier)

Data: 8,3,2,5,4,2,3

With previous notation, $x_1 = 8, x_2 = 3, x_3 = 2, \ldots, x_7 = 3$.

Sample Size = n = 7

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{8+3+2+5+4+2+3}{7} = \frac{27}{7} \approx 3.86$$

Sample Mean, Example (one outlier)

Data: 8,3,2,5,4,2,300

With previous notation, $x_1 = 8, x_2 = 3, x_3 = 2, \ldots, x_7 = 300$.

Sample Size $=$ n $= 7$

$$\overline{x} = \frac{8+3+2+5+4+2+300}{7} = \frac{324}{7} \approx 46.28$$

Sample Mean, Comments and Questions

- Sample mean is measure of center.
- Important since it estimates the population mean, used extensively in population models.
- Same units as data.
- Sensitive to outliers. Few large values can pull the sample mean towards their direction.

Questions?

Sample Median, Definition and Method

Given a set of data denoted as $x_1, x_2, \ldots, x_n$.
Note, sample size $= n$.

Sample Median:

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

1. Sort the data from smallest to largest.
2. If n is odd, then the sample median $\tilde{x}$ is the middle observation in the list; if n is even, then $\tilde{x}$ is the average of the two middle observations.

Sample Median, Example (no outlier)

Data: 8,3,2,5,4,2,3

Sorted Data: 2,2,3,3,4,5,8

$$\tilde{x} = 3$$

Sample Median, Example (one outlier)

Data: 8,3,2,5,4,2,300

Sorted Data: 2,2,3,4,5,8,300

$$\tilde{x} = 4$$

Sample Median, Comments and Questions

- Sample median is measure of center.
- Not sensitive to outliers or **robust**.
- Same units as data.
- If $\overline{x}$ is much further right than $\tilde{x}$ ($\overline{x} > \tilde{x}$), then outliers may be pulling the mean to the right. In this case, the distribution is likely right-skewed. Likewise, if $\overline{x}$ is much further left than $\tilde{x}$, the distribution is likely left-skewed.

Questions?

Trimmed Mean, Definition and Method

Median is robust, but sample mean estimates the mean, an important probabilistic quantity. **Trimmed Means** are more robust, but behave similiar to sample means.

Given a set of data denoted as $x_1, x_2, \ldots, x_n$.

1. Order the data, smallest to largest.
2. Discard the largest and smallest α%, α to be chosen.
3. Compute the mean of the remaining numbers. This is the trimmed mean, $\overline{x}_\alpha$

Trimmed Mean, Example (no outlier)

Data: 1,8,3,2,5,4,2,3

Sorted Data: 1,2,2,3,3,4,5,8

For $\alpha = 12$ (the 12% trimmed mean), the first and last numbers are discarded. The remaining are averaged.

$$\overline{x}_{12} = \frac{2+2+3+3+4+5}{6} \approx 3.16$$

$$\tilde{x} = (3+3)/2 = 3, \quad \overline{x} = 3.5$$

Trimmed Mean, Example (one outlier)

Data: 1,8,3,2,5,4,2,300

Sorted Data: 1,2,2,3,3,4,8,300

For $\alpha = 12$, the first and last numbers are discarded. The remaining are averaged.

$$\overline{x}_{12} = \frac{2+2+3+3+4+8}{6} \approx 3.66$$

$$\tilde{x} = (3+3)/2 = 3, \quad \overline{x} = 40.625$$

Trimmed Mean, Comments and Questions

- Trimmed mean is measure of center.
- Can be robust depending on α.
- Not clear how to pick α.

Questions?

Measures of Spread (Section 1.4)

Goal: Find the spread or variability of the data.

Tools: **Sample Variance and Standard Deviation**, **Range** and **Interquartile Range** (IQR or the 'fourth spread' in the book), box plots

Consideration: outliers

Sample Variance, Standard Deviation and Range, Definition

Given a set of data denoted as $x_1, x_2, \ldots, x_n$.

$n =$ sample size

$$\textbf{Range} := \max(x_i) - \min(x_i)$$

$$\textbf{Sample Variance} := s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

$$\textbf{Sample Standard Deviation} := s = \sqrt{s^2}$$

s^2 and s are important since the estimate population variance and standard deviation. Again, used in probabilistic models.

$\overline{x}$ is in the definition of s^2 and s, shouldn't expect them to be robust. And, the Range depends solely depends on the value of the outliers, shouldn't expect it to be robust either.

Sample Variance, Standard Deviation and Range, Example (no outlier)

Data: 1,8,3,2,5,4,2,3

Range $= 8 - 1 = 7$

sample size $=$ 8, $\overline{x} = 3.5$ from eariler

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - 3.5)^2}{8-1} \approx 4.85$$

$$s = \sqrt{s^2} \approx 2.2$$

Sample Variance, Standard Deviation and Range, Example (one outlier)

Data: 1,8,3,2,5,4,2,300

Range $= 300 - 1 = 299$

sample size $= 8$, $\overline{x} = 40.625$ from eariler

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - 40.625)^2}{8-1} \approx 10,989$$

$$s = \sqrt{s^2} \approx 104.83$$

Sample Variance, Standard Deviation and Range, Comments and Questions

- Sample Variance, Standard Deviation and Range are sensitive to outliers.
- Sample variance is measured in squared units while range and standard deviation have the same units as the data.

Questions?

Interquartile Range, Method

Given a set of data denoted as $x_1, x_2, \ldots, x_n$.

1. Separate the data into the lower half and upper half. (Include the median $\tilde{x}$ in both halves if n is odd.)
2. The median of the lower half is called the **first quartile**, or lower fourth.
3. The median of the higher half is called the **third quartile**, or upper fourth.
4. Their difference is the **interquartile range** (IQR), or fourth spread:

$$\text{IQR} = \text{third quartile} - \text{first quartile}$$

Interquartile Range, Example (no outlier)

Data: 1,8,3,2,5,4,2,3,7

Sorted Data: 1,2,2,3,3,4,5,7,8

Step 1: Separate the data into the lower half and higher half.

Lower Half: 1,2,2,3,3 Upper Half: 3,4,5,7,8

Steps 2 and 3: median of upper half is the third quartile; median of lower half is the first quartile

first quartile $= 2$ third quartile $= 5$

Step 4: Positive difference is the IQR

$$\text{IQR} = 5 - 2 = 3$$

Interquartile Range, Example (one outlier)

Data: 1,8,3,2,5,4,2,3,799

Sorted Data: 1,2,2,3,3,4,5,8,799

Step 1: Separate the data into the smaller half and larger half.

Lower Half: 1,2,2,3,3 Upper Half: 3,4,5,8,799

Steps 2 and 3: median of upper half is the third quartile; median of lower half is the first quartile

first quartile $= 2$ third quartile $= 5$

Step 4: Positive difference is the IQR

$$\mathrm{IQR} = 5 - 2 = 3$$

Sample Variance, Standard Deviation and Range, Comments and Questions

- IQR is a robust measure of spread.
- IQR has the same units as the data.

Questions?

Measures of Location and Spread, Summary

- Sample mean and median are measures of location.
- Median is robust; mean is not.
- Mean $<<$ median indicates likely left-skew. Mean $>>$ median indicates likely right-skew.
- Sample variance, standard deviation, IQR and range are measures of spread.
- Only the IQR is robust.
- $\overline{x}$, s^2 and s are important estimators of population parameters which will impact probabilistic models.

Linear Transformations

Goal: Relate the sample mean, variance and standard deviation of data after linear transformations, scaling then shifting ($ax + c$).

Linear Transformations and the Sample Mean

Given data, $x_1, x_2, \ldots, x_n$

Linearly transform the data: $y_i = ax_i + c$. a, c are constants.

Relate the sample means of each data.

$$\overline{y} = \frac{(\sum_{i=1}^{n} ax_i + c)}{n} = \frac{(\sum_{i=1}^{n} ax_i) + (\sum_{i=1}^{n} c)}{n} = \frac{a(\sum_{i=1}^{n} x_i) + nc}{n}$$

$$= \frac{a\sum_{i=1}^{n} x_i}{n} + \frac{nc}{n} = a\frac{\sum_{i=1}^{n} x_i}{n} + c = a\overline{x} + c$$

$$\overline{y} = a\overline{x} + c$$

Linearly transformations of data transform the mean in the exact same way.

Linear Transformations and the Sample Variance and Standard Deviation

Given data, $x_1, x_2, \ldots, x_n$

Linearly transform the data: $y_i = ax_i + c$. a, c are constants.

The sample mean of both data is related, $\overline{y} = a\overline{x} + c$.
Relate the variances of each data, s_y^2 and s_x^2.

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(ax_i + c - (a\overline{x} + c))^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} a^2(x_i - \overline{x})^2}{n-1} = a^2 s_x^2$$

$$s_y^2 = a^2 s_x^2, \quad s_y = |a| s_x$$

Sample variance and standard deviation ignore shifts. Variance squares scales as the square of the scaling and standard deviation scales by the absolute value.

Linear Transformations, Summary

Given data, $x_1, x_2, \ldots, x_n$, if the data is is linearly transformed: $y_i = ax + c$, a, c are constants, then the following hold.

- $\overline{y} = a\overline{x} + c$
- $s_y^2 = a^2 s_x^2$
- $s_y = |a| s_x$

Linear Transformations, Example

Denote the previous data, $1, 8, 3, 2, 5, 4, 2, 3$ by x_i. The sample mean, variance and standard deviation were calculated and will be denoted as

$$\overline{x} = 3.5, s_x^2 \approx 4.85 \text{ and } s_x \approx 2.2.$$

Multiply the data by -2 and add 100, $y_i = -2x_i + 100$.

$$\overline{y} = -2\overline{x} + 100 = -2(3.5) + 100 = 93$$

$$s_y^2 = (-2)^2 s_x^2 \approx 4(4.85) = 19.4$$

$$s_y = |-2| s_x \approx 2(2.2) = 4.4$$

Linear Transformations, Comments and Questions?

- May not work with other transformations.
- Formulas will be paralleled in probabilistic models.

Questions?

Five-number summary and Boxplot

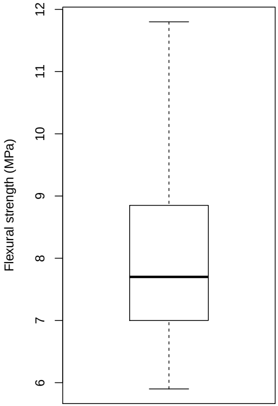
Five-number summary

1. Minimum observation
2. First quartile
3. Median
4. Third quartile
5. Maximum observation

For the concrete strength data:

$$5.90, 7.00, 7.70, 8.85, 11.80$$

This is shown graphically in a Boxplot: The bottom and top of the box show the first and third quartiles; the horizontal line inside the box shows the median; the whiskers (dotted lines) extend to the minimum and maximum observations.



Sample Proportion, Definition

Goal: Estimate the proportion of times a specified outcome of a categorical variable is observed.
Sucess: Number of times a specified outcome of a categorical variable is observed.

$$\textbf{Sample Proportion} := \hat{p} = \frac{\text{Successes}}{\text{Sample Size}}$$

Sample Proportion, Example

Data: 153 heads are observed in 300 variables.

$$\text{successes} = 153, \quad \text{sample size} = 304$$

$$\hat{p} = \frac{153}{304} \approx 0.503$$

The proportion of heads in the sample is roughly 0.503.

Sample Proportion, Questions

Questions?