*Math 408 - Mathematical Statistics*

# Lecture 36. Summarizing Data - III

April 29, 2013

# Agenda

- Measures of Location
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates
- Measures of Dispersion
  - Sample Standard Deviation
  - Interquartile Range (IQR)
  - Median Absolute Deviation (MAD)
- Boxplots
- Summary

# Measures of Location

In Lectures 34 and 35, we discussed data analogues of the CDFs and PDFs, which convey visual information about the shape of the distribution of the data.

Next Goal: to discuss simple numerical summaries of data that are useful when there is not enough data for construction of an eCDF, or when a more concise summary is needed.

- A **measure of location** is a measure of the center of a batch of numbers.
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates

Example: If the numbers result from different measurement of the same quantity, a measure of location is often used in the hope that it is more accurate than any single measurement.

# The Arithmetic Mean

The most commonly used measure of location is the arithmetic mean,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A common statistical model for the variability of a measurement process is the following:

$$x_i = \mu + \varepsilon_i$$

- $x_i$ is the value of the $i^{\text{th}}$ measurement
- $\mu$ is the true value of the quantity
- $\varepsilon_i$ is the random error, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

The arithmetic mean is then:

$$\overline{x} = \mu + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i, \qquad \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{n})$$

# The Median

The main drawback of the arithmetic mean is it is sensitive to outliers. If fact, by changing a single number, the arithmetic mean of a batch of numbers can be made arbitrary large or small. For this reason, measures of location that are robust, or insensitive to outliers, are important.

## Definition

If the batch size is an odd number, $x_1, \ldots, x_{2n-1}$, then the **median** $\tilde{x}$ is defined to be the middle value of the ordered batch values:

$$x_1, \ldots, x_{2n-1} \quad \rightsquigarrow \quad x_{(1)} < \ldots < x_{(2n-1)}, \quad \boxed{\tilde{x} = x_{(n)}}$$

Important Remark:

Moving the extreme observations does not affect the sample median at all, so the median is quite robust.

# The Trimmed Mean

Another simple and robust measure of location is the **trimmed mean** or **truncated mean**.

## Definition

The $100\alpha\%$ trimmed mean is defined as follows:

1. Order the data: $x_1, \ldots, x_n \quad \rightsquigarrow \quad x_{(1)} < \ldots < x_{(n)}$
2. Discard the lowest $100\alpha\%$ and the highest $100\alpha\%$
3. Take the arithmetic mean of the remaining data:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

where $[s]$ denotes the greatest integer less than or equal to $s$.

Remarks:

- It is generally recommended to use $\alpha \in [0.1, 0.2]$.
- Median can be considered as a 50% trimmed mean.

# M Estimates

Let $x_1, \ldots, x_n$ be a batch of numbers. It is easy to show that

- The mean
$$\overline{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} (x_i - y)^2$$
Outliers have a great effect on mean, since the deviation of $y$ from $x_i$ is measured by the square of their difference.
- The median
$$\tilde{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} |x_i - y|$$
Here, large deviations are not weighted as heavily, that is exactly why the median is robust.

In general, consider the following function:

$$f(y) = \sum_{i=1}^{n} \Psi(x_i, y),$$

where $\Psi$ is called the weight function. **M estimate** is the minimizer of $f$:

$$y^* = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$$

# Measures of Dispersion

A measure of dispersion, or scale, gives a numerical characteristic of the "scatteredness" of a batch of numbers. The most commonly used measure is the sample standard deviation $s$, which is the square root of the sample variance,

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

Q: Why $\frac{1}{n-1}$ instead of $\frac{1}{n}$?

A: $s^2$ is an unbiased estimate of the population variance $\sigma^2$. If $n$ is large, then it makes little difference whether $\frac{1}{n-1}$ or $\frac{1}{n}$ is used.

Like the mean, the standard deviation $s$ is sensitive to outliers.

# Measures of Dispersion

Two simple robust measures of dispersion are the interquartile range (IQR) and the median absolute deviation (MAD).

- **IQR** is the difference between the two sample quartiles:

$$\mathrm{IQR} = Q_3 - Q_1$$

  - $Q_1$ is the first (lower) quartile, splits lowest 25% of batch
  - $Q_2 = \tilde{x}$, cuts batch in half
  - $Q_3$ is the third (upper) quartile, splits highest 75% of batch

  How to compute the quartile values (one possible method):

  1. Find the median. It divides the ordered batch into two halves. Do not include the median into the halves.
  2. $Q_1$ is the median of the lower half of the data. $Q_3$ is the median of the upper half of the data.

- **MAD** is the median of the numbers $|x_i - \tilde{x}|$.

# Example

Let the ordered batch be $\{x_i\} = \{1, 2, 5, 6, 9, 11, 19\}$

- $Q_2 = \tilde{x} = 6$
- $Q_1 = 2$
- $Q_3 = 11$

$$\mathrm{IQR} = 9$$

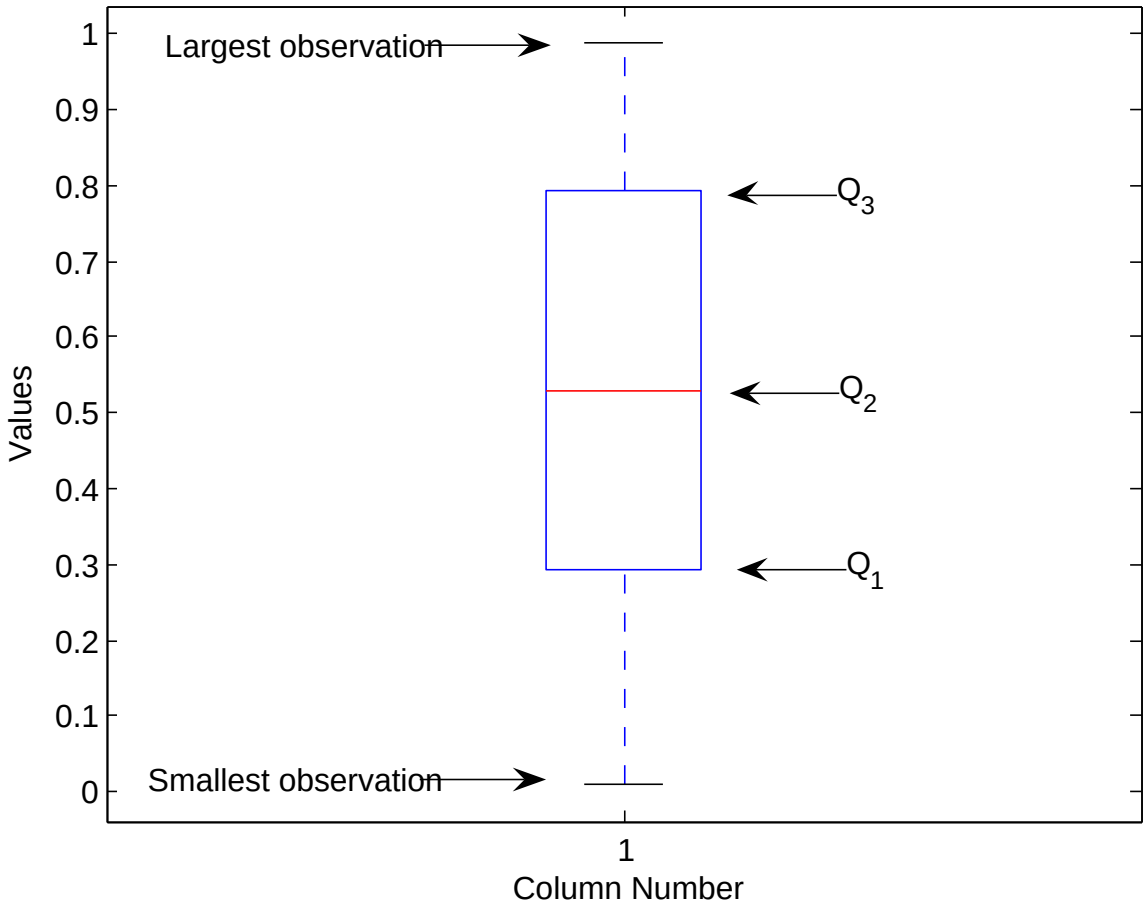
- $\{|x_i - \tilde{x}|\} = \{5, 4, 1, 0, 3, 5, 13\}$

$$\mathrm{MAD} = 4$$

# Boxplots

A boxplot is a graphical display of numerical data that is based on five-number summaries: the smallest observation, lower quartile ($Q_1$), median ($Q_2$), upper quartile ($Q_3$), and largest observation.

Example: $x_1, \ldots, x_n \sim U[0,1]$, $n = 100$



# Summary

- Measures of Location
  - Arithmetic Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers)
  - Median: the middle value of the ordered batch values $\tilde{x} = Q_2$
  - Trimmed Mean:
$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$
  - M estimate: $y^* = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$
    - ⋆ if $\Psi(x_i, y) = (x_i - y)^2$, then $y^* = \overline{x}$
    - ⋆ it $\Psi(x_i, y) = |x_i - y|$, then $y^* = \tilde{x}$
- Measures of Dispersion
  - Sample Standard Deviation (sensitive to outliers):
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
  - Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$
  - Median Absolute Deviation: $\mathrm{MAD} =$ median of the numbers $|x_i - \tilde{x}|$
- Boxplots are useful graphical displays.