Mathematics 231

Lecture 3

Liam O'Brien

Announcements

- Reading
  - Today		M&M 1.2		30-44
  - Next class	M&M 1.2		45-47
  				M&M 1.3		53-62

Numerical Measures of Center

- Mean
- Median
- Mode

Mean

- The mean is what is typically though of as the "average" value:

If there are n observations with values $x_1, x_{2,} \ldots, x_n$
the mean is the sum of these numbers divided by the number of observations: $(x_1 + x_2 + \cdots + x_n) / n$
For example, if the data are 2,4,6,2,2, the mean is (2+4+6+2+2)/5=3.2.

Median

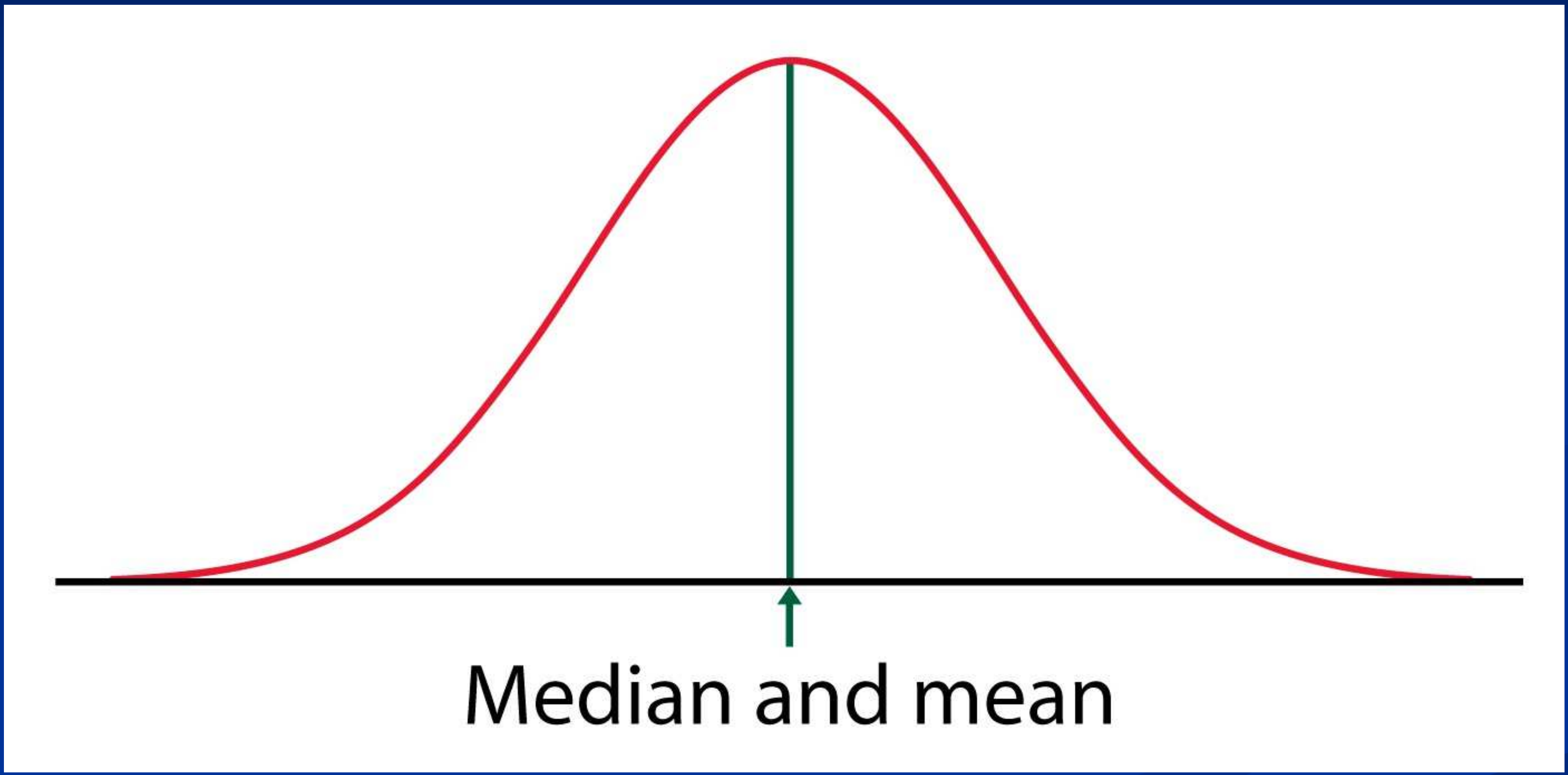
- The median is the "midpoint."
- The median is the point at which 50% of the observations are smaller, and 50% are larger.
- For example, if the data are 2,4,6,2,2, we order them: 2,2,2,4,6 and find the middle value.
- The median is 2.

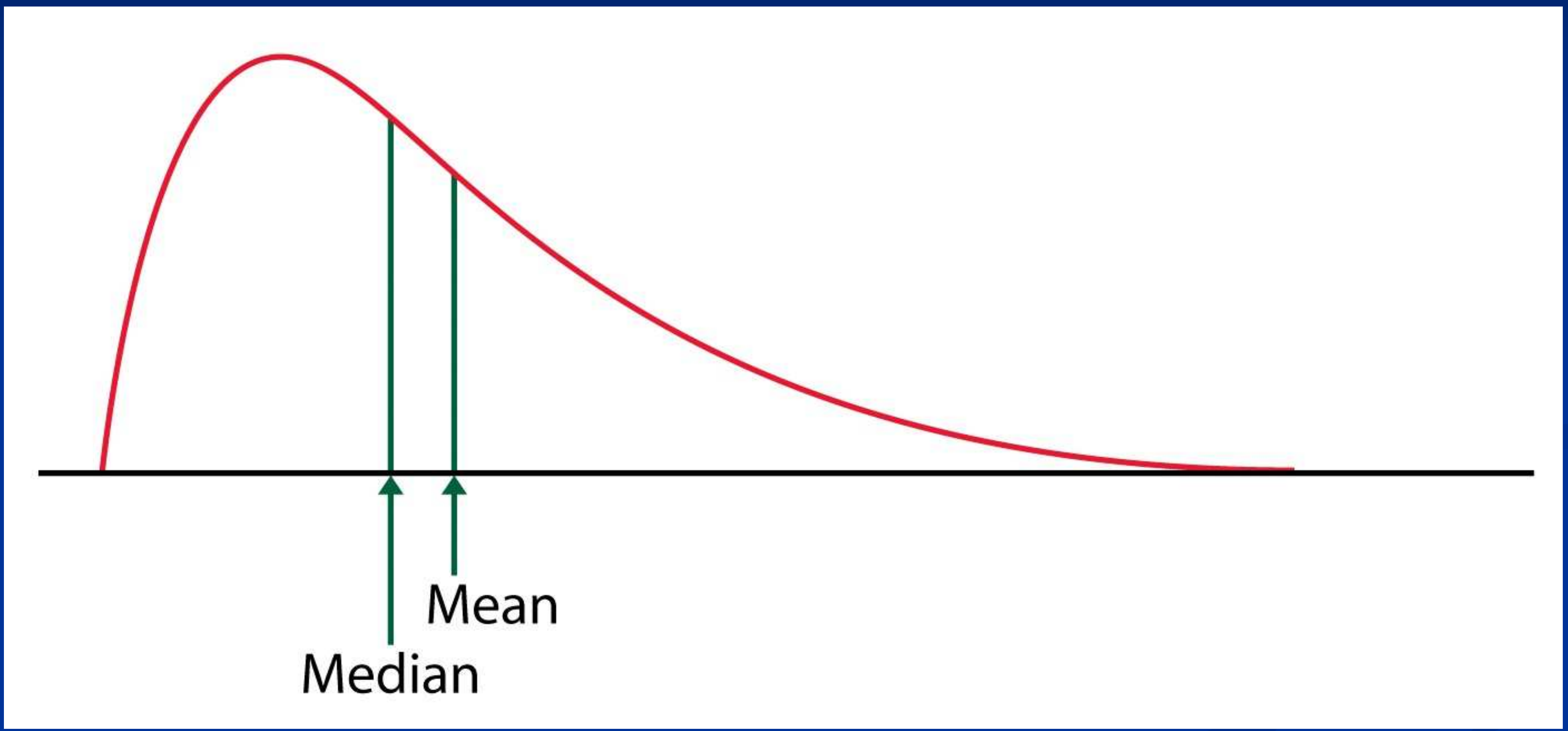
Mode

- The mode is the value that is observed the most often.
- The mode need not be unique – we can have two or more values that are observed the same (but most frequent) number of times.
- For example, if the data are 2,2,2,4,6, the mode is 2 because it occurs the most often.

When are the Mean and Median Similar?

- When the shape of the distribution is symmetric, the mean and median are similar.
- When the distribution is "skewed" the mean is farther out in the "tail" than the median.
- It has been said that the "mean follows the tail."
- The median is much less sensitive to extreme observations (sometimes called "outliers").


Median and mean


Mean
Median

Example: Sample Data

- If the data are 2,2,2,4,6, consider replacing the 6 with 60.
- The mean changes to 15, but median is still 2.
- Which is more representative?

Example: Harvard Salary Survey

- In 1998, the entering class of 1973 was surveyed.
- Interested in determining the typical salary for a graduate of the big H 25 years after graduation.
- Mean salary: $750,000
- Median salary: $175,000
- Why such a large discrepancy?

Example: Expected Salary

- Mean = \$372k

- Median = \$100k

- How much variability is there? A lot? A little? How can we quantify it?

Percentiles

- If *k* marks the ***p*** th **percentile**, then *p* percent of the data are less than or equal to *k*.
- Two common percentiles:
  - **25**$^{\text{th}}$ **percentile**: sometimes called the 1$^{\text{st}}$ (or lower) quartile, $\mathbf{Q_1}$
  - **75**$^{\text{th}}$ **percentile**: sometimes called the 3$^{\text{rd}}$ (or upper) quartile, $\mathbf{Q_3}$

Finding the Quartiles

1. Sort the observations in numerical order.
2. Q_1 = median of the lower half of the list.
3. Q_3 = median of the upper half of the list.

- We already know how to find the 2nd quartile, Q_2 – it's just the median.
- Note that if there are an odd number of observations, then don't include the median in the lists used in steps (1) and (2).

5-Number Summary

- We can now find the 5-number summary of a dataset. This is often used as a basic way to look at the distribution of the data.
- The 5 numbers are:
  1. Smallest value
  2. 1^{st} quartile
  3. Median
  4. 3^{rd} quartile
  5. Largest value

Inter-quartile Range (IQR)

- The IQR is the spread (or range) in the middle half of the data; distance between the 1st and 3rd quartiles: $IQR=(Q_3-Q_1)$.
- This not only tells us something about the spread, but it can also help identify outliers.
- An observation is defined as an outlier if it falls more than 1.5*IQR above Q_3 below Q_1.

Example: Expected Salary

- 1st quartile = \$100k
- Median = \$100k
- 3rd quartile = \$200k

- IQR = \$200k - \$100k = \$100k

- Does this indicate a lot of variability?

Variance and Standard Deviation

- Consider how we might measure the spread in terms of the distance of each observation from the mean:

$$x_i - \overline{x}$$

- What if we summed these distances for all observations?

$$\sum_i (x_i - \overline{x}) = 0$$

Variance and Standard Deviation

- The **variance** is defined as (approximately) the average squared distance of the observations from the mean:

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

- Here, *n* is the number of observations, and $x_1, x_2, \ldots, x_n$ are the observations themselves.

Variance and Standard Deviation

- The standard deviation is usually denoted by *s*, and the simply the square root of the variance.

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

- Note that the standard deviation is in term of the original measurement units, but the variance is not.

Example: Expected Salary

- Variance = 2121929k (in squared dollars)

- Standard deviation = $1457k

Numerical Summaries in Stata

```
. summ  salary, d

                       Expected salary
-------------------------------------------------------------
      Percentiles      Smallest
 1%            8              8
 5%           45             18
10%           55             45       Obs                  46
25%          100             45       Sum of Wgt.          46

50%          100                      Mean           372.3043
                        Largest       Std. Dev.      1456.684
75%          200            450
90%          400            500       Variance        2121929
95%          500            650       Skewness       6.481538
99%        10000          10000       Kurtosis       43.34826
```

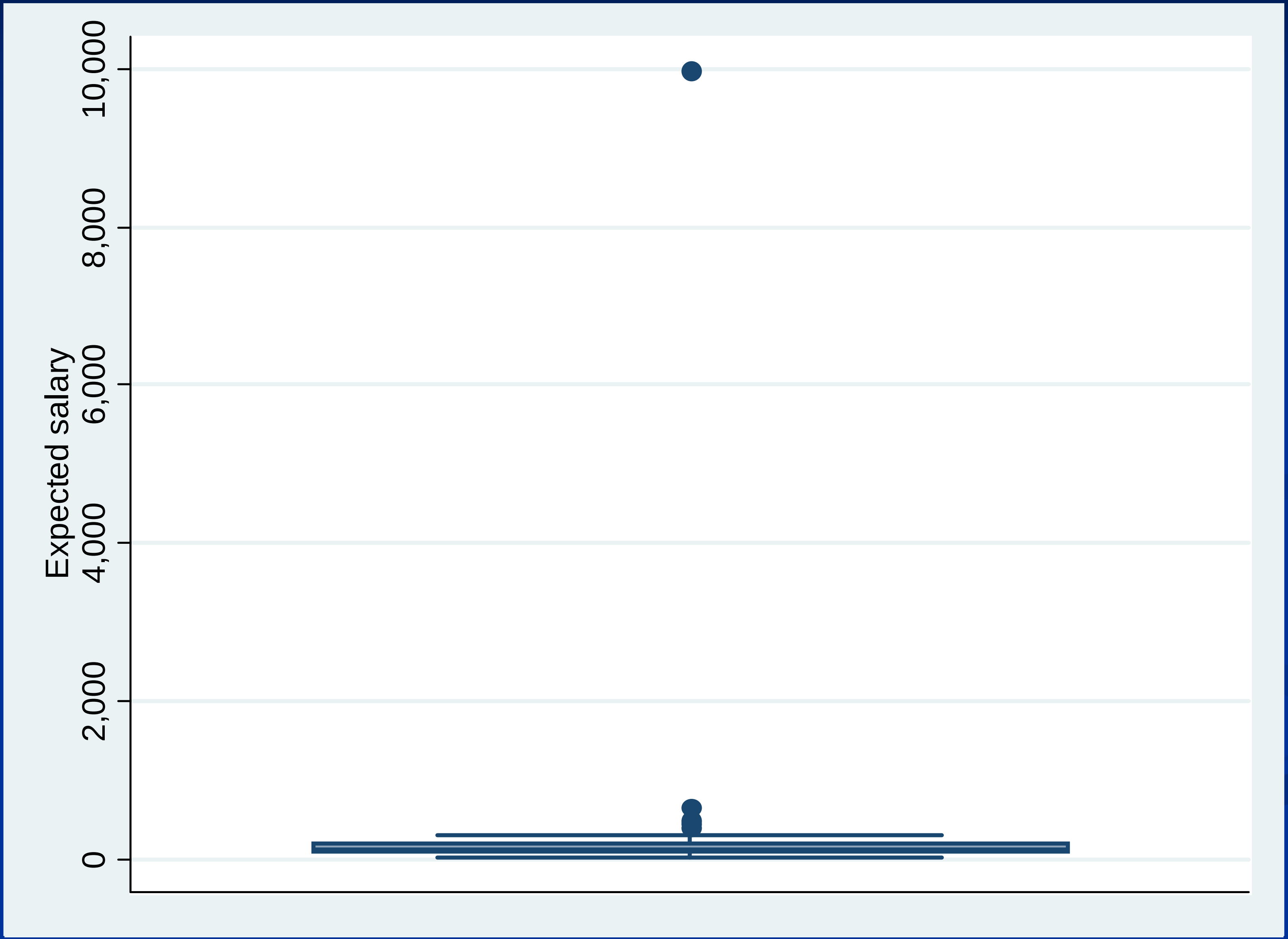
Boxplots

- A **boxplot** graphically displays several important features of a distribution, including the median, quartiles, and outliers.
- Boxplots are often useful for comparing the distributions for two or more groups (e.g., males vs. females).

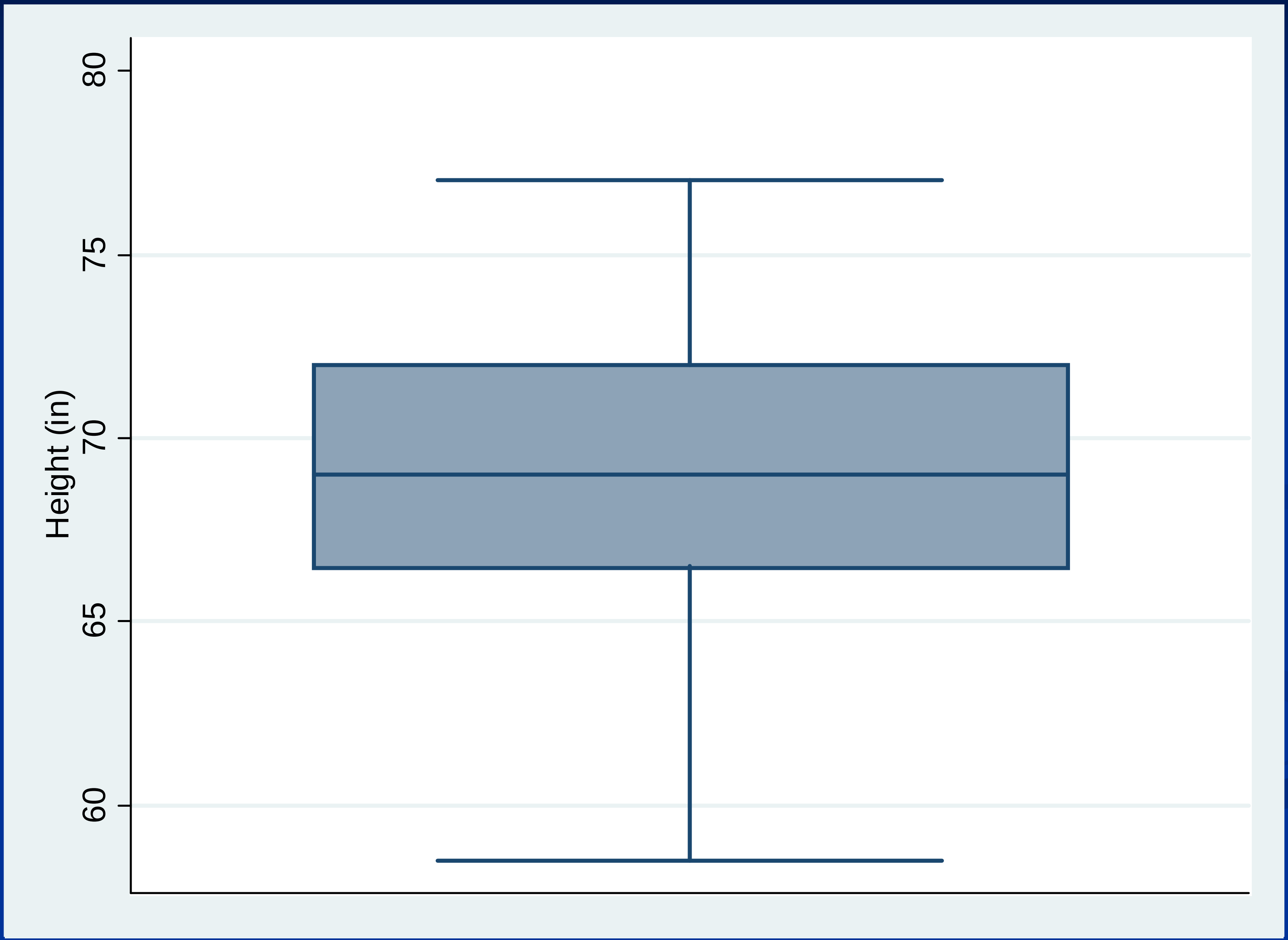
Constructing a Boxplot

- Draw a box whose ends are at the 1st and 3rd quartiles (the width of the box is equal to the IQR).
- Draw a line through the box at the median.
- Any observations that are greater than Q_3+1.5*IQR or less than Q_1-1.5*IQR are considered to be outliers and are individually plotted.
- Draw lines from the ends of the box to the most extreme values that aren’t outliers.

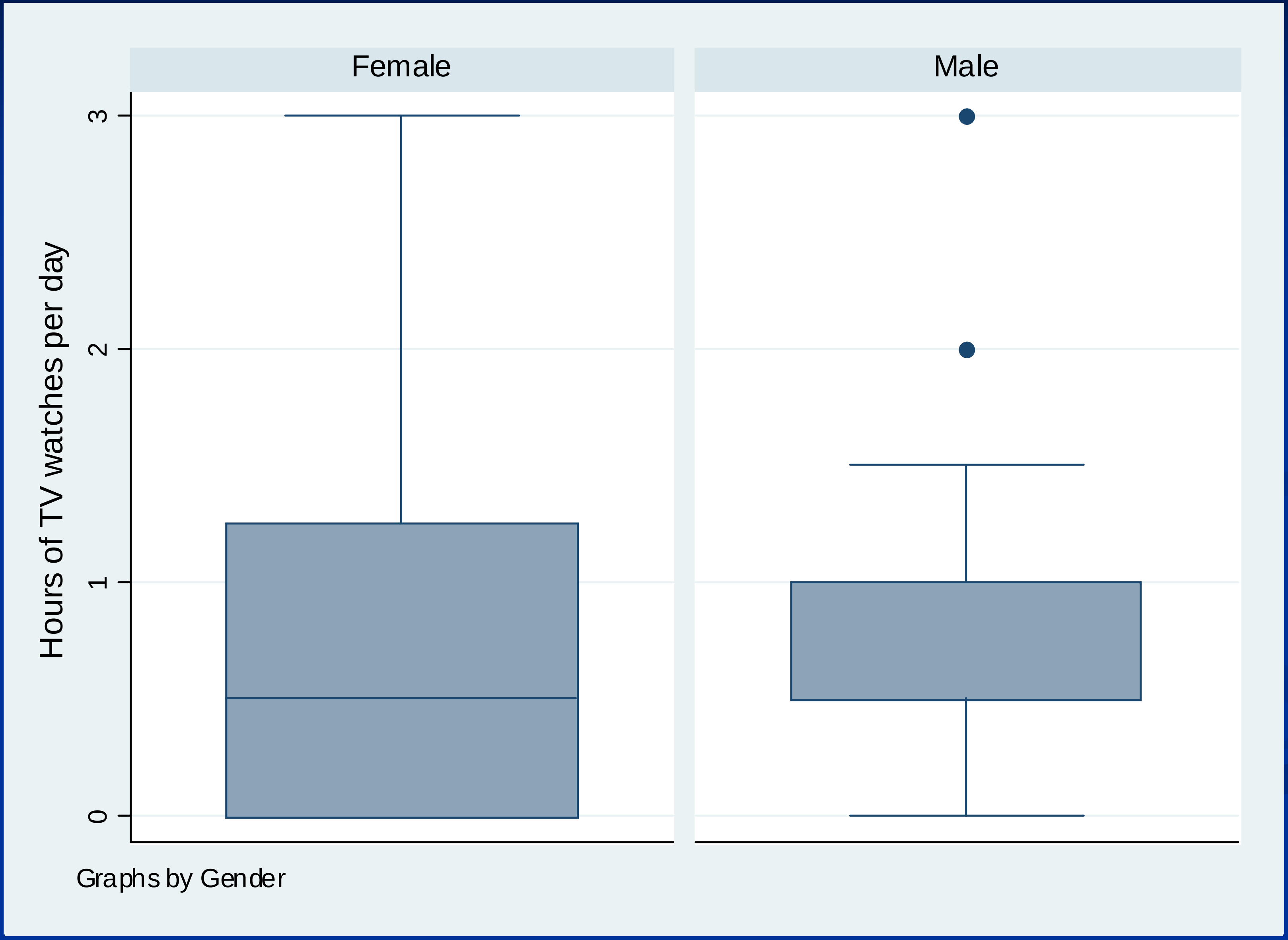
Example: Expected Salary



Example: Heights



Example: TV Viewing By Gender



Which Summary Measures to Use?

- **Mean and standard deviation:** These are sensitive to outliers and skewness and are more appropriate when the data distribution is fairly symmetric.
- **Median and IQR:** Far less sensitive to outliers, and less sensitive to skewness.