## Statistics One

Lecture 4
Summary Statistics

## Two segments

- Measures of central tendency
- Measures of variability

## Lecture 4 ~ Segment 1

Measures of central tendency

## Wine tasting!

## Example: Wine ratings

- Suppose that 100 wine experts rated the overall quality of different wines on a scale of 1 to 100
  - Higher scores indicate higher quality

## Example: Wine ratings

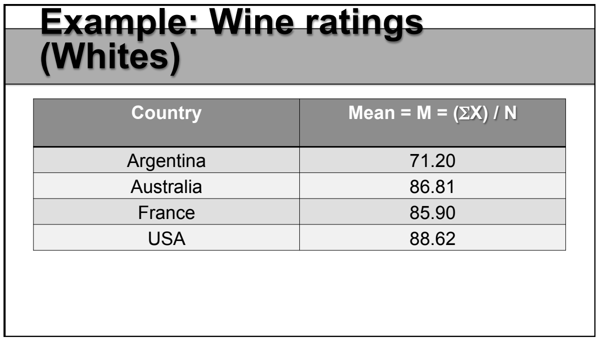
- Consider the red wines, which country had the highest average (mean) rating?

## Example: Wine ratings (Reds)

| Country | Mean = M = (ΣX) / N |
|---|---|
| Argentina | 66.73 |
| Australia | 81.76 |
| France | 70.97 |
| USA | 76.38 |

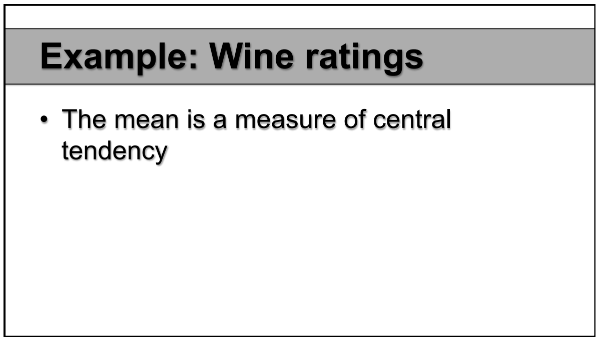
## Example: Wine ratings

- Now consider the white wines, which country had the highest average (mean) rating?



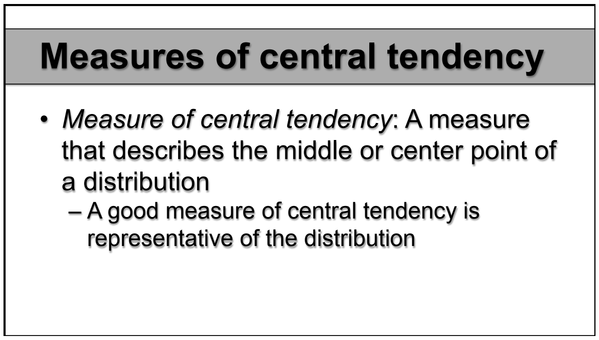
## Example: Wine ratings (Whites)

| Country | Mean = M = (ΣX) / N |
|---|---|
| Argentina | 71.20 |
| Australia | 86.81 |
| France | 85.90 |
| USA | 88.62 |



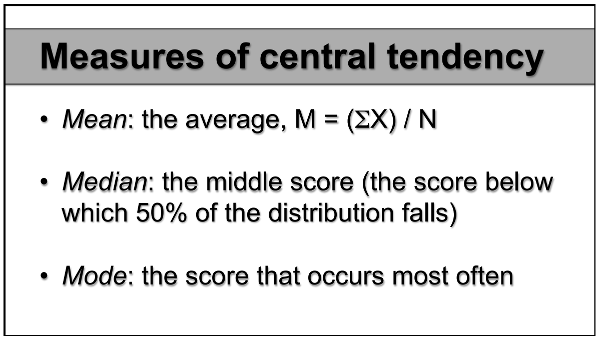
## Example: Wine ratings

- The mean is a measure of central tendency



## Measures of central tendency

- *Measure of central tendency*: A measure that describes the middle or center point of a distribution
  - A good measure of central tendency is representative of the distribution



## Measures of central tendency

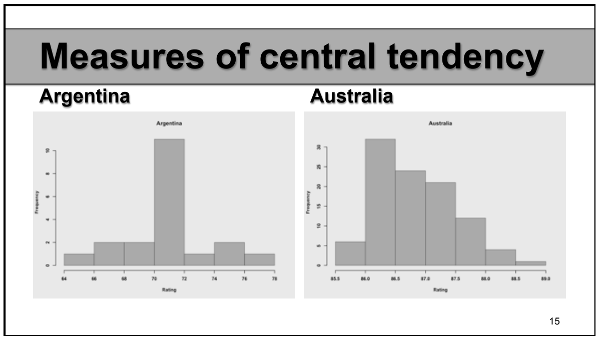
- *Mean*: the average, M = (ΣX) / N
- *Median*: the middle score (the score below which 50% of the distribution falls)
- *Mode*: the score that occurs most often

## Measures of central tendency

- Mean (average) is the best measure of central tendency when the distribution is normal
  - Red wine ratings
  - Another example: Grade Point Average (GPA)

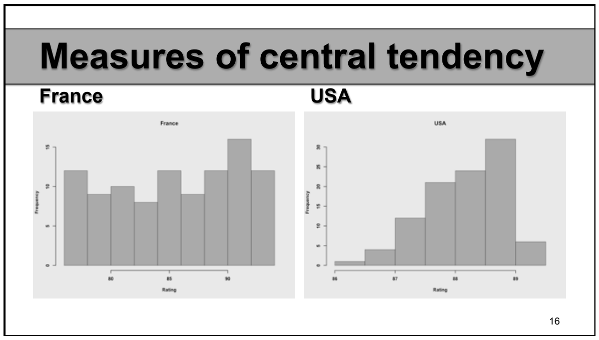
## Measures of central tendency

- Median (middle score) is preferred when there are extreme scores in the distribution
  - White wine ratings?
  - Another example: Household income in USA



## Measures of central tendency

**Argentina** **Australia**



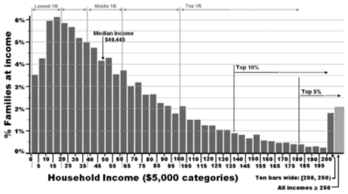
## Measures of central tendency

**France** **USA**

## Example: Wine ratings (Whites)

| Country | Mean = M = (ΣX) / N | Median |
|---|---|---|
| Argentina | 71.20 | 71.00 |
| Australia | 86.81 | 86.68 |
| France | 85.90 | 86.00 |
| USA | 88.62 | 88.65 |

## Measures of central tendency



Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

18

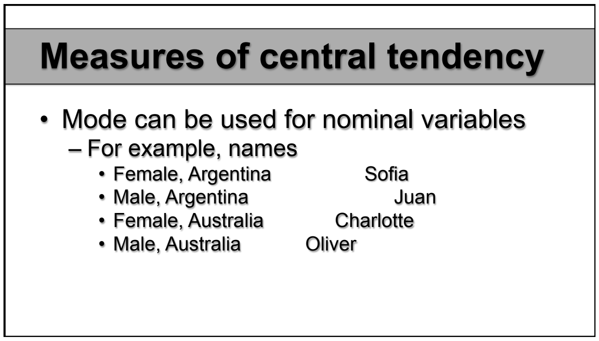
## Measures of central tendency

- Mode is the score that occurs most often
  - – The peak of a histogram
  - – The rating that occurred the most
    - For example, the Argentina white, Mode = 70 – 72

## Measures of central tendency

- Mode can be used for nominal variables
  - – For example, names

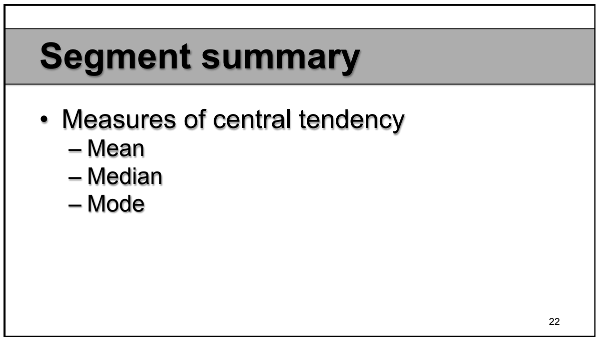
| | |
|---|---|
| Female, USA | Sophia |
| Male, USA | James |
| Female, France | Emma |
| Male, France | Nathan |



## Measures of central tendency

- Mode can be used for nominal variables
  - For example, names

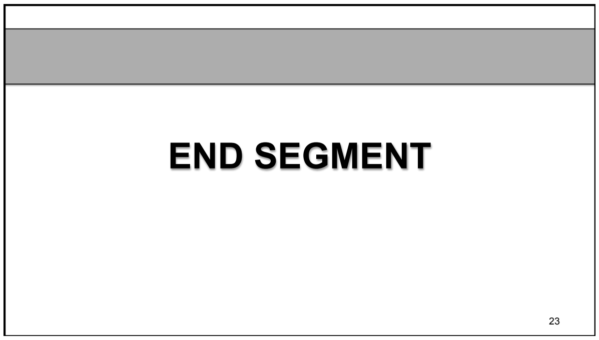
| | |
|---|---|
| • Female, Argentina | Sofia |
| • Male, Argentina | Juan |
| • Female, Australia | Charlotte |
| • Male, Australia | Oliver |



## Segment summary

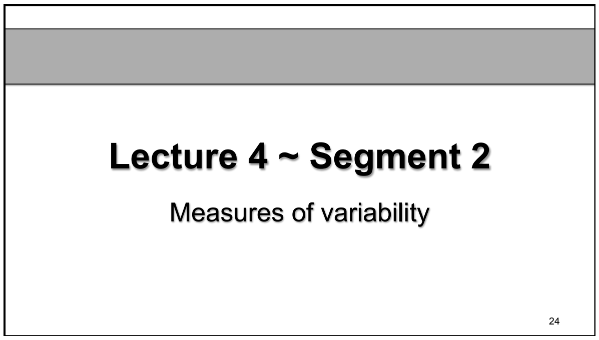
- Measures of central tendency
  - Mean
  - Median
  - Mode

22



## END SEGMENT

23



## Lecture 4 ~ Segment 2

Measures of variability

24

## Variability

- A measure that describes the range and diversity of scores in a distribution
  - *Standard deviation* (SD): the average deviation from the mean in a distribution
  - *Variance* = $SD^2$

## Variability

- *Variance =* $SD^2$

$$SD^2 = [\Sigma(X - M)^2] / N$$

## Variance

- Variation is natural and observed in all species and that's good:
  - *On the Origin of Species (1859)*
  - *Variation Under Domestication (1868)*

## Linsanity!

## Jeremy Lin (10 games)

| Points per game | (X-M) | $(X-M)^2$ |
|---|---|---|
| 28 | 5.3 | 28.09 |
| 26 | 3.3 | 10.89 |
| 10 | -12.7 | 161.29 |
| 27 | 4.3 | 18.49 |
| 20 | -2.7 | 7.29 |
| 38 | 15.3 | 234.09 |
| 23 | 0.3 | 0.09 |
| 28 | 5.3 | 28.09 |
| 25 | 2.3 | 5.29 |
| 2 | -20.7 | 428.49 |
| M = 227/10 = 22.7 | M = 0/10 = 0 | M = 922.1/10 = 92.21 |

## Results

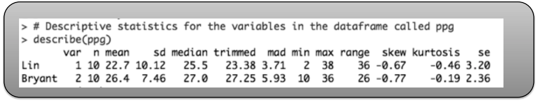
- M = Mean = 22.7
- SD = Standard Deviation = 9.6
- $SD^2$ = Variance = 92.21

## Notation

- M = Mean
- SD = Standard Deviation
- $SD^2$ = Variance (also known as MS)
  - MS stands for Mean Squares
  - SS stands for Sum of Squares

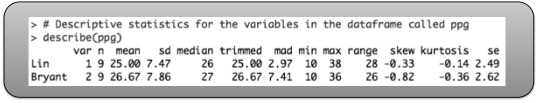
## Lin vs. Kobe

## 10 games, R output



```
> # Descriptive statistics for the variables in the dataframe called ppg
> describe(ppg)
       var  n mean    sd median trimmed  mad min max range  skew kurtosis   se
Lin      1 10 22.7 10.12   25.5   23.38 3.71   2  38    36 -0.67    -0.46 3.20
Bryant   2 10 26.4  7.46   27.0   27.25 5.93  10  36    26 -0.77    -0.19 2.36
```

## 9 games, R output



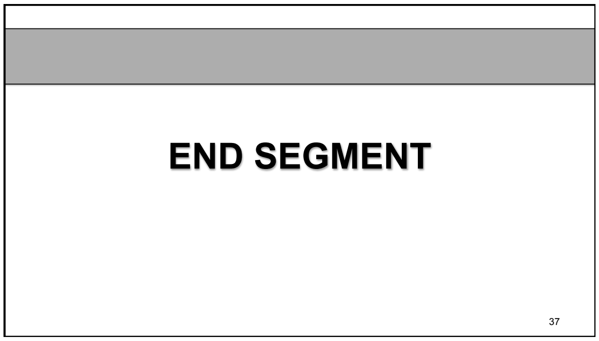
```
> # Descriptive statistics for the variables in the dataframe called ppg
> describe(ppg)
       var n  mean   sd median trimmed  mad min max range  skew kurtosis   se
Lin      1 9 25.00 7.47     26   25.00 2.97  10  38    28 -0.33    -0.14 2.49
Bryant   2 9 26.67 7.86     27   26.67 7.41  10  36    26 -0.82    -0.36 2.62
```

## Summary statistics: Review

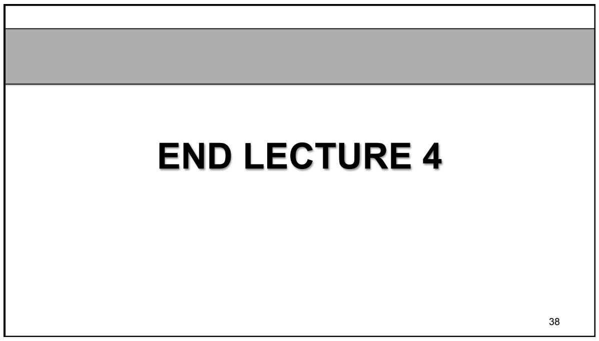
- Important concepts
  - Central tendency (mean, median, mode)
  - Variability (standard deviation and variance)

## Summary statistics: Review

- Summary statistics (formulae to know)
  - $M = (\Sigma X) / N$
  - $SD^2 = [\Sigma(X - M)^2] / N$
    - Used for descriptive statistics
  - $SD^2 = [\Sigma(X - M)^2] / (N – 1)$
    - Used for inferential statistics



END SEGMENT



END LECTURE 4