Lecture 2
Segment 2

Summary statistics

Summary statistics

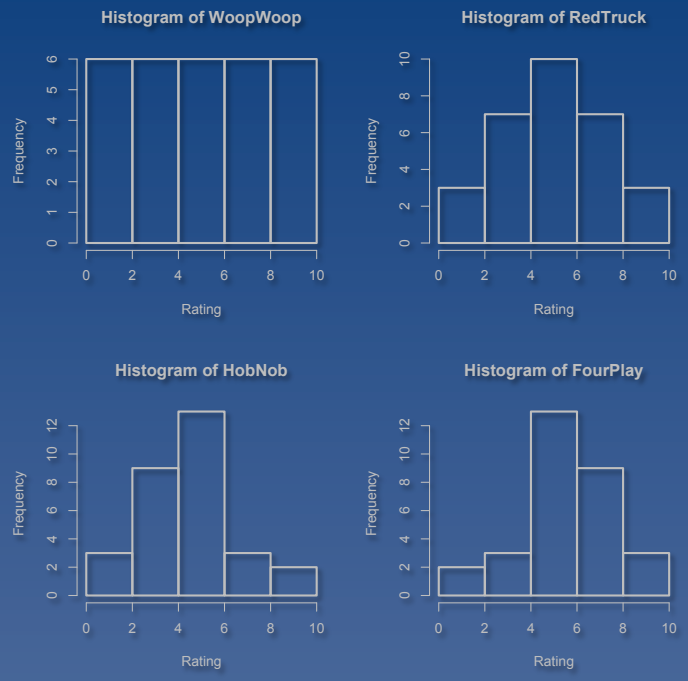
- Important concepts
  - Central tendency (mean, median, mode)
  - Variability (standard deviation and variance)
  - Skew
  - Kurtosis

Wine tasting!

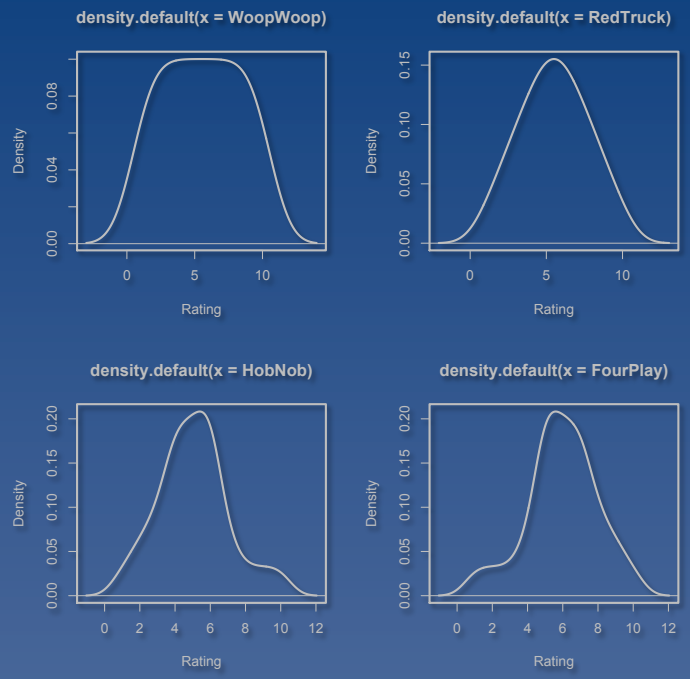
Example

- Suppose that 30 wine experts rated the overall quality of 4 different wines on a scale of 1-10
  - Higher scores indicate higher quality

Four histograms



Four density plots



Summary statistics from R

```
> # Descriptive statistics for the variables in the dataframe called ratings
> describe(ratings)
          var  n mean   sd median trimmed  mad min max range  skew kurtosis   se
RedTruck    1 30 5.50 2.26    5.5    5.50 2.22   1  10     9  0.00    -0.81 0.41
WoopWoop    2 30 5.50 2.92    5.5    5.50 3.71   1  10     9  0.00    -1.34 0.53
HobNob      3 30 5.03 2.01    5.0    4.96 1.48   1  10     9  0.33    -0.01 0.37
FourPlay    4 30 5.97 2.01    6.0    6.04 1.48   1  10     9 -0.33    -0.01 0.37
>
```

Measures of central tendency

- Measure of central tendency: a measure that describes the center point of a distribution. A good measure of central tendency is representative of the entire distribution.

Measures of central tendency

- Mean: the average, $M = (\Sigma X)/ N$
- 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
  - Average GPA
  - Average SAT
  - Average rating (e.g., wine ratings)

Measures of central tendency

- Median (middle score) is preferred when there are extreme scores in the distribution
  - Median household income
  - Median reaction time
  - Median GPA?
    - May be best representative of overall performance if distribution of grades is skewed

Measures of central tendency

- Mode is the score that occurs most often
  - The peak of a histogram
  - The most “popular” score
  - Again, GPA?
    - “I mostly got As in college”

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

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

Variance

- Variation is natural and observed in all species and that's good! See Darwin:
  - *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

```
> # 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

```
> # 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)
  - Skew
  - Kurtosis

Summary statistics: Review

- Descriptive statistics (formulae to know)
  - $M = (\Sigma X) / N$
  - $SD^2 = [\Sigma(X - M)^2] / N$
  - $SD^2 = [\Sigma(X - M)^2] / (N - 1)$

Image in slide 3 was retrieved from
http://www.delawareonline.com/blogs/secondhelpings/uploaded_images/redwinegl-758416.JPG

Image in slide 15 is from Nathaniel S. Butler, NBAE/Getty Images.

Image in slide 19 is from John Angelillo, API.