Describing Data

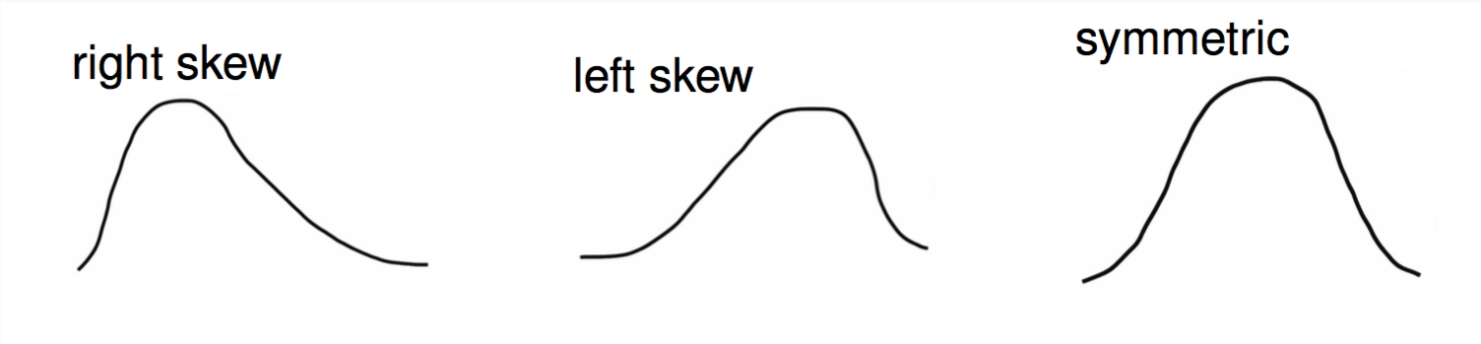
Describing Distributions

- **Shape**: modality, skewness
- **Center**: mean, median, mode
- **Spread**: variance, sd, range, IQR
- **Unusual observations**: outliers

Shape

Modality

Skewness



Shape Q

Which of these variables do you expect to be uniformly distributed?

1. weights of adult females
2. salaries of a random sample of people from Oregon
3. house prices
4. birthdays of classmates (day of the month)

Shape Q

Which of these variables do you expect to be uniformly distributed?

1. weights of adult females
2. salaries of a random sample of people from Oregon
3. house prices
4. **birthdays of classmates (day of the month)**

Center: mean

```
X <- c(8, 11, 7, 7, 8, 11, 9, 6, 10, 7, 9)
```

$$\frac{8+11+7+7+8+11+9+6+10+7+9}{11} = \frac{93}{11} = 8.45$$

Sample mean: the arithmetic mean of the data (vs *pop mean*)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \text{vs.} \qquad \mu$$

```
mean(X)
```

```
## [1] 8.5
```

Center: median

Median: the middle value of a sorted data set.

```
sort(X)
```

```
##  [1]  6  7  7  7  8  8  9  9 10 11 11
```

```
median(X)
```

```
## [1] 8
```

Break ties by averaging middle two if necessary.

Center: mode

Mode: the most frequently observed value in the data set.

```
table(X)
```

```
## X
##  6  7  8  9 10 11
##  1  3  2  2  1  2
```

Spread (on board)

Spread: variance

Sample variance: roughly, the mean squared deviation from the mean.

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

Compare to the *population variance*, σ^2, which divides by n.

Spread: variance

```
X - mean(X)
```

```
##  [1] -0.45  2.55 -1.45 -1.45 -0.45  2.55  0.55 -2.45  1.55
```

```
(X - mean(X))^2
```

```
##  [1] 0.21 6.48 2.12 2.12 0.21 6.48 0.30 6.02 2.39 2.12 0.30
```

```
sum((X - mean(X))^2) / (length(X) - 1)
```

```
## [1] 2.9
```

```
var(X)
```

```
## [1] 2.9
```

Spread: standard deviation

Sample standard deviation: the square root of the variance. Used because units are the same as the data.

$$s = \sqrt{(s^2)}$$

```
sqrt(var(X))
```

```
## [1] 1.7
```

```
sd(X)
```

```
## [1] 1.7
```

Compared to the *population standard deviation*, σ.

Spread: IQR

Inner Quartile Range: the range of the middle 50% of the data.

$$IQR = Q3 - Q1$$

```
sort(X)
```

```
##  [1]  6  7  7  7  8  8  9  9 10 11 11
```

```
IQR(X)
```

```
## [1] 2.5
```

Spread: range

Range: the range of the full data set.

$$range = max - min$$

```
max(X) - min(X)
```

```
## [1] 5
```

```
range(X)
```

```
## [1]  6 11
```

Spread Q

Which measure(s) of spread would be sensitive to the presence of outliers?

1. variance
2. standard deviation
3. IQR
4. Range

Spread Q

```
X
```

```
##  [1]  8 11  7  7  8 11  9  6 10  7  9
```

```
Y
```

```
##  [1] 37 11  7  7  8 11  9  6 10  7  9
```

```
var(X)
```

```
## [1] 2.9
```

```
var(Y)
```

```
## [1] 77
```

Spread Q

```
IQR(X)
```

```
## [1] 2.5
```

```
IQR(Y)
```

```
## [1] 3.5
```

```
range(X)
```

```
## [1]  6 11
```

```
range(Y)
```

```
## [1]  6 37
```

Spread Q

Which measure(s) of spread would be sensitive to the presence of outliers?

1. **variance**
2. **standard deviation**
3. IQR
4. **Range**

Center Q

Which measure(s) of center would be sensitive to the presence of outliers?

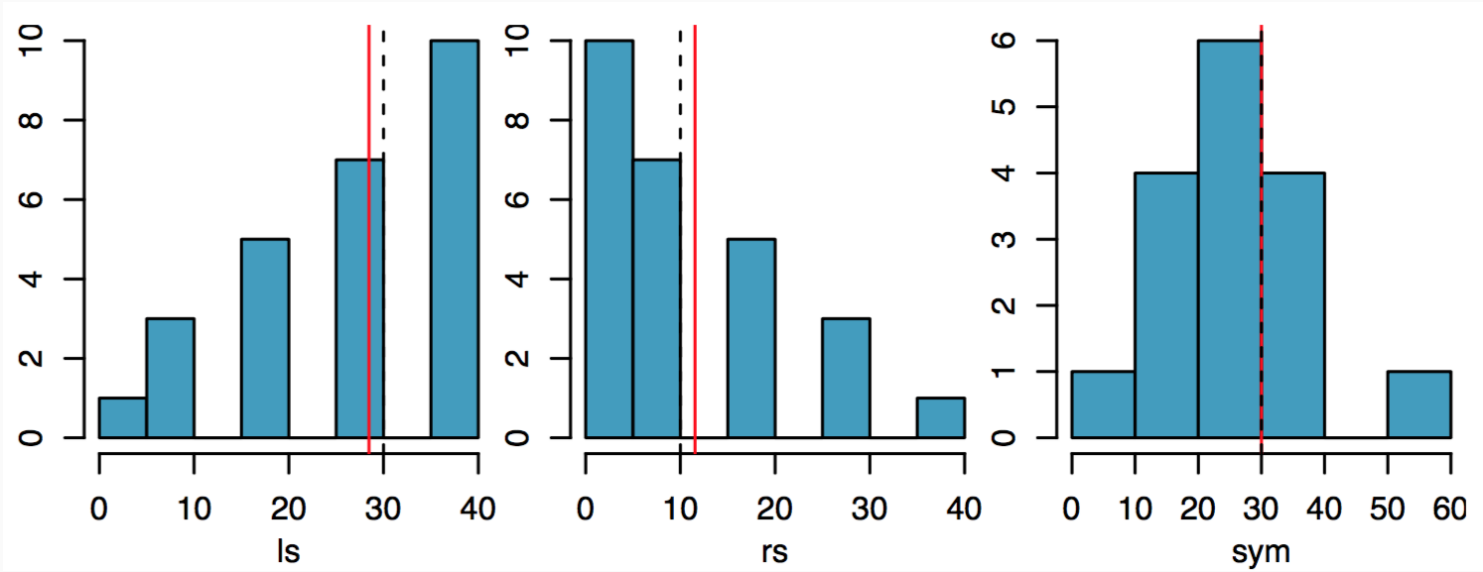
1. mean
2. median
3. mode

Center Q

Which measure(s) of center would be sensitive to the presence of outliers?

1. **mean**
2. median
3. mode

Mean vs median



The mean (red line) is sensitive to extreme values, so it gets pulled towards the tail. The median (dashed line) is less sensitive.

For symmetric dists, use *mean* and *sd*.

For skewed dists, use *median* and *iqr*.