Chapter 2: Descriptive Statistics

- Measures of Location

Observe $X_1, X_2, \ldots, X_n$

Sample of numeric values

subscript indexes the units

Eg. X_i = Birthweight for baby i

- Measure of location = center of a sample (statistic) or a population (parameter)

- Arithmetic Mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- Example: $X_1 = 2$, $X_2 = 5$, $X_3 = -4$

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 2 + 5 + -4$$

$$\sum_{i=2}^{3} X_i = X_2 + X_3 = 5 + -4$$

$$\sum_{i=2}^{2} X_i = X_2 = 5$$

$$\bar{X} = \frac{1}{3}\sum_{i=1}^{3} X_i = \frac{1}{3}(2 + 5 + -4) = \frac{1}{3} \cdot 3 = 1$$

- $\bar{X}$ is sensative to extreme observations

$X_4 = 3997$

$$\frac{1}{4}\sum_{i=1}^{4} X_i = \frac{1}{4}(2 + 5 + -4 + 3997) = \frac{1}{4} \cdot 4000 = 1000$$

- <u>Median:</u>

  n odd $\Rightarrow$ $\left(\frac{n+1}{2}\right)$th largest observation

  n even $\Rightarrow$ Average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations

- <u>Example:</u> $x_1 = 2,\ x_2 = 5,\ x_3 = -4 \Rightarrow -4, 2, 5$

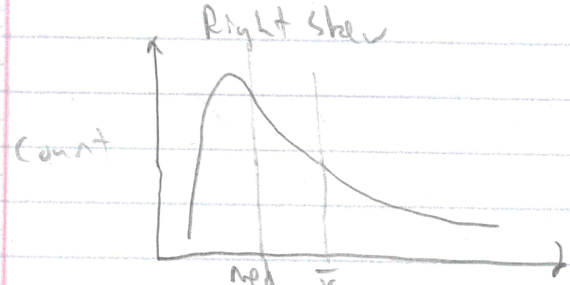
  $\text{Median}(x) = 2$

  $x_4 = 3997$

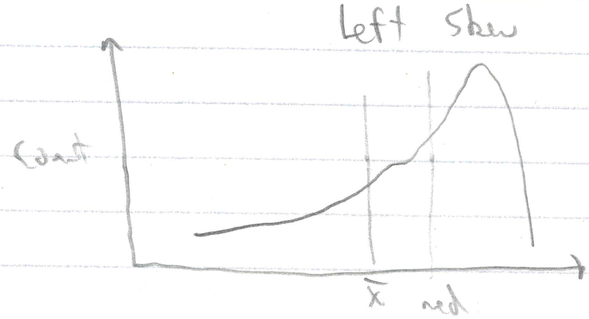
  $\Rightarrow \text{Median}(x) = \frac{2+5}{2} = 3.5$

- If distribution is symmetric, $\text{median}(x) \approx \bar{x}$

  Mean chases skew of distribution



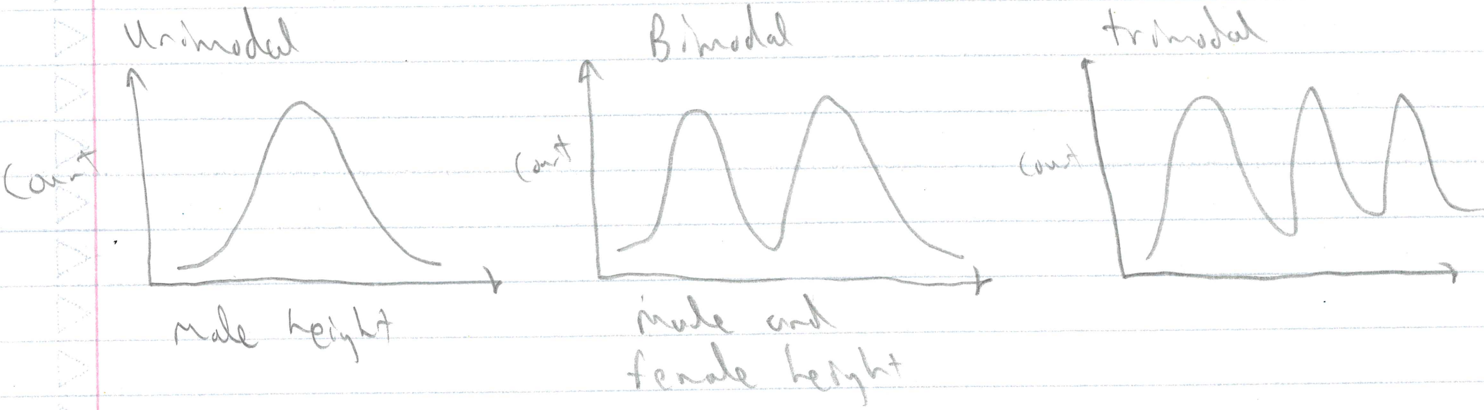
Years oral contraception use



Relative humidity in Summer Ohio

- Use mean if total is important
- Use median if lots of skew

- A <u>mode</u> is a frequently occurring value



- The mode is typically not used as a real measure of center, but rather as a way to describe distributions.

- <u>Properties of $\bar{x}$:</u>

- Suppose you have a frequency table

The intervals between menstrual periods (days)

Value	Freq
24	5
25	10
26	28
27	64
28	185

$n = 5 + 10 + 28 + 64 + 185 = 292$

$$\bar{x} = \frac{1}{292}\Big(\underbrace{x_1 + \cdots + x_5}_{24} + \underbrace{x_6 + \cdots x_{15}}_{25} + \cdots\Big)$$

$$= \frac{1}{292}\left(5\cdot 24 + 10\cdot 25 + 28\cdot 26 + 64\cdot 27 + 185\cdot 28\right)$$

$$= 27.42$$

$$\text{Median}(x) = \frac{146^{th} \text{ and } 147^{th} \text{ values}}{2}$$

$$= \frac{28+28}{2} = 28$$

- Let $y_i = x_i + c$, then $\bar{y} = \bar{x} + c$

Proof: $\bar{y} = \frac{1}{n}\sum(x_i+c) = \frac{1}{n}\sum x_i + \frac{1}{n}\sum c = \bar{x} + \frac{1}{n} nc = \bar{x} + c$ //

- Eg. let y_i = deviation from 28 day cycle
  $y_i = x_i - 28$
  $\bar{y} = 27.42 - 28 = -0.58$

- $\text{Median}(y) = \text{Median}(x) + c$

- Let $y_i = c x_i$, then $\bar{y} = c\bar{x}$

proof: $\bar{y} = \frac{1}{n}\sum c x_i = c \frac{1}{n} \sum x_i = c\bar{x}$

- Eg. Change units from days to weeks
  $y_i = \frac{1}{7} x_i$
  $\bar{y} = \frac{1}{7} \cdot 27.42 \approx 3.92$

- If $y_i = c_1 x_i + c_2$, then $\bar{y} = c_1\bar{x} + c_2$

- Exercise: What is the mean menstrual cycle deviation from 4 weeks

$3.92 - 4 = -0.08$

Measures of Spread:

Spread = how far apart numbers are

- Range = Max − Min (sensitive to extreme values)
- Inter-quartile range (IQR)

75th percentile − 25th percentile

- p^{th} percentile = value V_p such that $p\%$ of points are at or below V_p
- Ex.) Median = 50th percentile
- Quantile = in units of proportions instead of percents

0.75 quantile = 75th percentile

- Ex.) $x_1 = 2$, $x_2 = 5$, $x_3 = -4$

$\frac{1}{3}$ Quantile = −4

$\frac{2}{3}$ Quantile = 2

1 Quantile = 5

What about the 40th percentile

Use some weighted average of −4 and 2, but definitions vary

- Variance = Average of squared deviations

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

↑ $n-1$ because lose some information by estimating $\bar{x}$

- Standard deviation = square root of variance

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

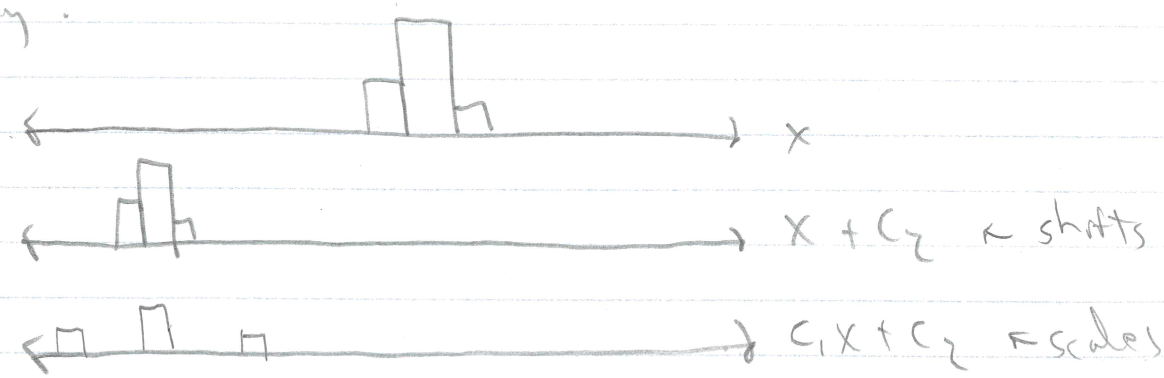
↑ puts this measure of center on same scale as units (e.g. oz instead of oz^2)

- Let $y_i = c_1 x_i + c_2$ then $s^2(y) = c_1^2 s^2(x)$

↑ Only scaling affects variance and SD
↓

$$s(y) = c_1 s(x)$$

- Why?



- **Exercise:** What is $s^2(y)$ when $y_i = c_1 x$

  $c_1^2 s^2(x)$

- $\boxed{\text{R Notebook on Mean / Median / SD / var}}$