Types of Data

- Qualitative (Categorical) —
  M/F, Fr/So/Jr/Sr, Dem/Rep/Wrestler, etc.
  - Bar graph (frequency or percent)
  - Pie Chart
  - *Center/Location*: Mode (most frequent)
  - *Spread/Variability*: None!
  - *Later in 36-201*: Contingency Tables
- Quantitative (Discrete or Continuous) —
  Dollars, Age, Test Score, Number of Cars, etc.
  - Stem and leaf
  - boxplot
  - histogram (frequency, percent, density)
  - *Location*: Five number summary:
    Min, Q1, Med, Q3, Max
  - *Center/Location*: Mean, Median, Mode (if discrete)
  - *Spread/Variability*: Range, IQR,
    Standard Deviation (SD; see below!)

What is the best summary for quantitative data???

- Need to know the shape of the data.

Shape of Quantitative Data Distributions

Main distinctions for stem and leaf and histograms:

- Symmetry
  - symmetric
  - skewed left
  - skewed right

- Number of modes
  - unimodal (one hump)
  - bimodal (two humps), trimodal (three), etc.
  - multimodal (two or more humps)
  - rectangular (flat across the top, no humps)

- Outliers
  - Say whether or not there are any
  - If there are:
    - write them down
    - is each one high or low?
    - is each one far from non-outliers or close?

Useful Numerical Summaries

- Unimodal

Shape	Center	Spread
Symmetric, no outliers	Mean = Median = Mode	SD
Symmetric, outliers	Mean = Median	IQR
Skewed left	Mean < Median	IQR
Skewed right	Mean > Median	IQR

- Multimodal
  - Describe the modal humps (where, how wide)
  - Describe the gaps (where, how wide)
  - Maybe give overall Median, IQR

The Standard Deviation

At one food testing laboratory, the egg fat content data was

```
.62   .55   .34   .24   .80   .68   .76
```

$N =$ the number of observations $= 7$.
$x_1, x_2, \ldots, x_N$ name the individual observations:

$$x_1 = .62,\ x_2 = .55,\ x_3 = .34,\ x_4 = .24,\ x_5 = .80,\ x_6 = .68,\ x_7 = .76$$

The Mean

$$\overline{x} = (x_1 + x_2 + \cdots + x_N)/N = 0.57$$

Deviations from the Mean; MAD and Variance

	Observation x	Deviation $x - \overline{x}$	Abs Dev $\lvert x - \overline{x} \rvert$	Sq Dev $(x - \overline{x})^2$
	$x_1 = .62$	0.05	0.05	0.0025
	$x_2 = .55$	-0.02	0.02	0.0004
	$x_3 = .34$	-0.23	0.23	0.0529
	$x_4 = .24$	-0.33	0.33	0.1089
	$x_5 = .80$	0.23	0.23	0.0529
	$x_6 = .68$	0.11	0.11	0.0121
	$x_7 = .76$	0.19	0.19	0.0361
Mean	0.57	0.00	0.1657	0.0380

$$\sqrt{0.0380} = 0.1949$$

The Sample Standard Deviation (SD)

We just calculated

$$\begin{aligned}
\text{Sample Mean } (\overline{x}) &= (x_1 + x_2 + \cdots + x_N)/N \ = \ 0.57 \\
\text{Population Variance} &= \left[(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2\right]/N \\
&= 0.0380 \\
\text{Population SD} &= \sqrt{\text{Population Variance}} \ = \ 0.1949
\end{aligned}$$

For technical statistical reasons ("unbiased estimates") we usually calculate

$$\text{Sample Mean } (\overline{x}) = (x_1 + x_2 + \cdots + x_N)/N \ = \ 0.57$$

$$\text{Sample Variance } (s^2) = \left[(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2\right]/(N-1)$$

$$= 0.0443$$

$$\text{Sample SD } (s) = \sqrt{\text{Sample Variance}} \ = \ 0.2105$$

Comparing Mean/SD with Median/IQR

Among the egg fat measurements

```
.62   .55   .34   .24   .80   .68   .76
```

the largest is

$$x_5 = 0.80$$

What if it were bigger still?

x_5	Mean	SD	Median	IQR
0.8	0.57	0.21	0.62	0.275
0.9	0.58	0.23	0.62	0.275
1	0.60	0.26	0.62	0.275
2	0.74	0.58	0.62	0.275
3	0.88	0.95	0.62	0.275
10	1.88	3.58	0.62	0.275
20	3.31	7.36	0.62	0.275
30	4.74	11.14	0.62	0.275
100	14.74	37.60	0.62	0.275

Binary (Yes/No) Data

Another use for Sample Mean, Sample SD

- Examples…
  - Would You Vote for Smith?
  - Do You Approve of Plan B to Fund New Stadiums in Pittsburgh?
  - Did the Coin Come Up Heads?
- Numerical Representation
  - Yes $= 1$
  - No $= 0$
- We are typically interested in
  - Center: Sample Mean
  - Spread: Sample Variance, SD

Example

32 people were asked: Do you approve of Plan B?

```
1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1

5 Yes's
27 No's
```

Mean: Fraction of Yes's

$$\begin{aligned}\text{Sample Mean } (\overline{x}) &= (x_1 + \cdots + x_N)/N \;=\; \frac{(\#\text{ Yes's})}{N} \\ &= 5/32 \;=\; 0.1563\end{aligned}$$

Variance and SD

$$\text{Sample Variance } (s^2) = \left[(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2\right]/(N-1)$$

$$= \frac{(\#\text{ Yes's}) \times (\#\text{ No's})}{N \times (N-1)} \;=\; 0.1361$$

$$\text{Sample SD } (s) = \sqrt{\text{Sample Variance}} \;=\; 0.3689$$

Notes:

- *Siegel p. 131 has a typo (X = Yes's, Y = No's):*

$$\frac{X+Y}{n \times (n-1)} \quad \textit{should be:} \quad \frac{X \times Y}{n \times (n-1)}$$

- *Many books use*

$$s^2 = \frac{(\#\textit{ Yes's}) \times (\#\textit{ No's})}{N \times N}$$

for the variance of binary data
($s^2 = 0.1318$, $s = 0.3631$).