Statistics Review

- PSY 450W
- Dr. Schuetze

Two central ways of using numbers.

- Descriptive Statistics:
  - Simple quantitative description or summary.
    - Batting average in baseball
    - Grade-point average
- Inferential Statistics:
  - Conduct analyses on samples
    - Compare groups (experimental v. control…)
  - Use statistical operations to generalize the results to a population.

Describing data

We characterize the general trend or character of data using two key statistics:

1. Central tendency or general "drift" of the scores.
   - Mode → most common score
   - Median → middle of the distribution
   - Mean → average score
2. Variance: how diverse the scores are (how much *vary* from each other).
   - Range → …from the highest to lowest score
   - Standard deviation → "average" amount the scores vary from the Mean score

Mode

- Most frequent score in the distribution
- Example: scores = 16, 20, 21, 20, 36, 15, 25, 15, 12

Score	Frequency	% of cases
12	1	11
15	3	33
20	2	22
21	1	11
25	1	11
36	1	11

- 15 is most common = mode

Mode

- Characteristics
  - Used for all numerical scales, particularly nominal.
  - Insensitive to extreme values or range of scores.
  - Unstable: sensitive to small shifts in number of cases.

Median

- Mid-point of a distribution of scores
  - List scores in numerical order
    - 12, 15, 15, 15, 20, 20, 21, 25, 36
  - Locate the score in the center of the sample
    - 12, 15, 15, 15, 20, 20, 21, 25, 36
    - The middle (5^{th} out of 9) score = 20.

Median

- Characteristics:
  - Sensitive to the range of scores
  - More stable than the mode
  - Not sensitive to extreme scores (e.g., changing highest score (36) to 100 would not change the median.

Mean (*M*)

- The "average" score in a sample

Most common measure of central tendency

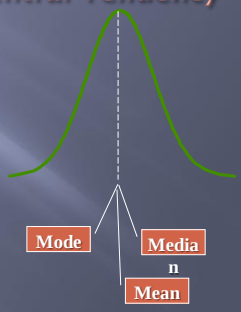
- Total all scores: 12+15+20+21+20+36+15+25+15 = 179
- Divide by "n" of scores: 179 / 9 = 19.9

Mean

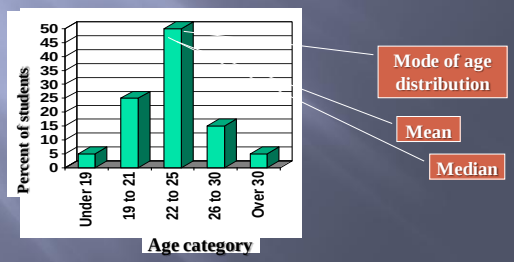
- Characteristics:
  - Good for Ratio or interval scales
  - sensitive to all observed values
  - highly stable; with larger n is insensitive to subtle changes in values
  - Can be highly sensitive to extreme values (particularly in smaller samples).

Measures of Central Tendency

- For a normal distribution the *mean, mode,* and *median* are all same -- the center of the distribution
- **Most variables in nature (and science) are normally distributed**



Age is a good example of a variable that is *normally distributed*



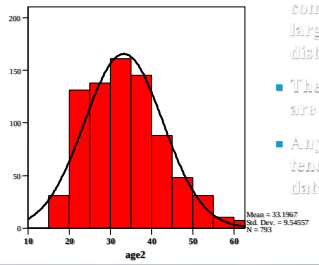
Mode of age distribution

Mean

Median

Age, Chicago community sample

N	Valid	793
	Missing	24
Mean		33.1967
Median		32.0000
Mode		30.00
Std. Deviation		9.54557
Variance		91.118
Skewness		.633
Range		50.00



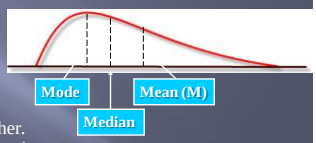
Measures of Central Tendency: A normal distribution

- Scores for age from a large community sample form a largely symmetrical distribution.
- The Mean, Median, and mode are similar.
- Any measure of central tendency well represents the data.

Central tendency: Skewed Distributions

A skewed distribution has extreme scores in one direction.

The extreme scores make the median higher than the mode. *(The high scores to the right move the 50% point that direction…).*



The Mean gets pulled even higher. *(Adding in some very high scores raises the average…).*

Common examples:

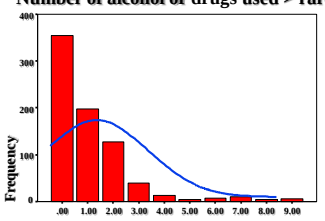
- Behaviors such as alcohol or drug use:
  - Most people use none or moderate
  - A diminishing number use higher levels
- Demographic variables such as income

Positive skew example

Example of typical strong positive skew; Drug & alcohol use (Community survey sample)

NDRGALC Number of alcohol or drugs used > rarely

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	355	43.5	46.3	46.3
	1.00	198	24.2	25.8	72.2
	2.00	128	15.7	16.7	88.9
	3.00	40	4.9	5.2	94.1
	4.00	13	1.6	1.7	95.8
	5.00	5	.6	.7	96.5
	6.00	7	.9	.9	97.4
	7.00	10	1.2	1.3	98.7
	8.00	4	.5	.5	99.2
	9.00	6	.7	.8	100.0
	Total	766	93.8	100.0	
Missing	System	51	6.2		
Total		817	100.0		



Statistics

Number of alcohol or drugs used > rarely

N	Valid	766
	Missing	51
Mean		1.11
Median		1.00
Mode		.00
Std. Deviation		1.59
Variance		2.54
Skewness		2.46
Minimum		.00
Maximum		9.00

Measures of Variability

- Variability: amount of fluctuation in data.
- 20, 30, 40, 50, 60, 70, 80
- 47, 48, 49, 50, 51, 52, 53

Measures of Variability

- **Range:** Difference between highest and lowest scores.
- **Variance:** deviation from the mean of the scores. How much scores are spead out or dispersed around mean.
- **Standard Deviation:** Squared root of variance.

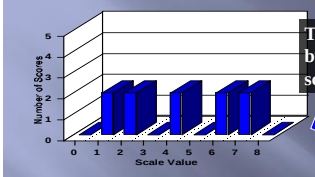
Variance: Standard Deviation

Estimates of *Variance*:

2. The Standard deviation (*S*) of scores around the Mean

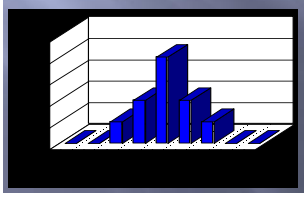
- Similar to the "*average*" amount that each score deviates from the *M* of the sample.
- "Standardizes" scores to a normal curve, allowing basic statistics to be used.
- More accurate & detailed than range:
  - *A few extremely high or low scores ("outliers") may make the range inaccurate*
  - *S assesses the deviation of all scores in the sample from the mean*

Comparing Scores: Variance & Standard Deviation



The data sets have the same M, but differ in how widely their scores vary (their variance).

- ✓ **High variance**
- ✓ **S = 2.4**



- ✓ **Less variance**
- ✓ **S = 1.15**

Scales of Measurement

- **Nominal Scale:** observations are labeled and categorized (qualitative).
- **Ordinal Scale:** observations are ranked in terms of size/ magnitude they are in relation to each other (qualitative).
- **Interval Scale:** equal differences (intervals) between numbers on the scale reflect equal differences in magnitude (quantitative).
- **Ratio Scale:** ratios of numbers do reflect ratios of magnitude (quantitative).

Normal Distribution

- **Characteristics**
  - Symmetrical
  - Three measures of central tendency are same value
  - Most scores fall close to mean
- **Parametric Statistics:** inferential stats used to analyze normally distributed interval/ratio scores.
- **Nonparametric Statistics:** inferential statistics used to analyze interval/ratio scores not normally distributed.

Testing Hypotheses

- **Statistical Hypothesis:** restatement of research hypothesis into two different hypotheses.
  - **Alternative Hypothesis:** statistical term for research hypothesis (H_o).
  - **Null Hypothesis:** Predicted relationship does not exist in the population (H_1).

Statistical Hypothesis Testing

Null Hypothesis. **All scores differ from the *M* by chance alone.**

- *Statistical Question (alternate hypothesis):*
  - Does this score differ from the *M* by > chance?
- *Using the Normal Distribution*
  - More extreme scores have a lower probability of occurring by chance alone
  - The # of standard deviation units ('Z' score) = the % of cases above or below the observed score (its "extremity")

"Statistical significance"

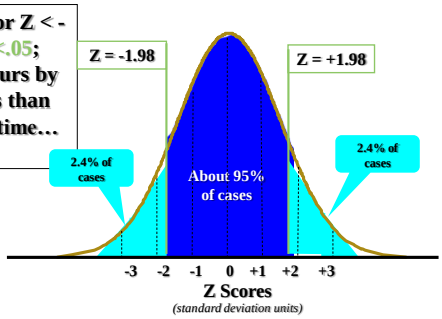
Statistical significance

- ✓ By convention, we assume that a score with less than 5% probability of occurring [i.e., higher or lower than 95% of the other scores…$p < .05$] has not occurred by chance alone.
- ✓ $p < .05$ corresponds to $Z = 1.98$; Z tells us if we can consider the effect (the distance from the *M*) to be "Statistically Significant."
- ✓ if $Z > 1.98$ we consider the score to be "significantly" different from the mean

Statistical significance & areas under the normal curve

95% of scores are between Z = -1.98 and Z = +1.98.

Z > +1.98 or Z < -1.98 has p<.05; a.k.a it occurs by chance less than 5% of the time…



One-tailed vs. Two-tailed

- Nondirectional hypothesis → two-tailed test
- Directional hypothesis → one-tailed test

Errors in Hypothesis Testing

- **Type I Error:** the null hypothesis has been mistakenly rejected when it is actually true.
- **Type II Error:** the null hypothesis has been mistakenly accepted when it is actually false.

Chi Square

- Nonparametric test: determines whether the frequencies of responses in our sample represent frequencies expected in the population.
- Contingency table
- Compares obtained frequencies with expected frequencies

Chi Square

	Complied	Refused	Row Totals
Group	6 (15.5) {7.5}	74 (64.5) {9.3}	80
Alone	25 (15.5) {31.3}	55 (64.5) {68.75}	80
Column Totals	31	129	160