Introduction to Statistics

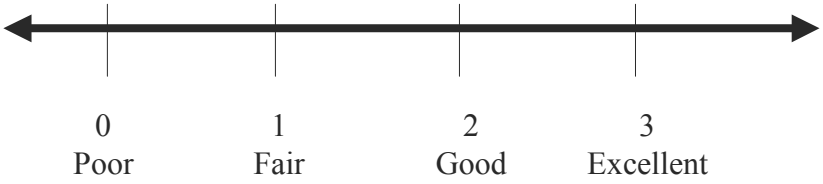
Katherine McKnight, Ph.D.
Pearson Achievement Solutions

AEA Summer Institute
Atlanta, GA
June, 2008

What is measurement?

- Assigning numbers to our observations
- The numbers reflect different properties of those observations (e.g., magnitude, frequency, size, amount)
- Goal: to map our numbers onto the phenomenon of interest as closely as possible

Measuring Health



Health Status Scores

Measuring Income

Please indicate your annual pre-tax income bracket:

_____ Under $30,000
_____ $30,000 to $49,999
_____ $50,000 to $74,999
_____ $75,000 to $99,999
_____ $100,000 to $149,999
_____ $150,000 or more

Measuring …???

How many years have you worked as an evaluator? _____ (years)

Types of Measures

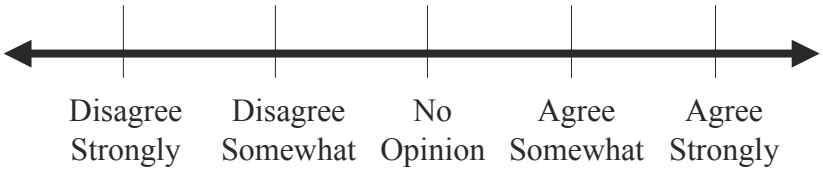
- Quantitative vs. qualitative
- Continuous vs. discrete
- Nominal, ordinal, interval and ratio scales

Nominal Measures

- Used for identifying things
- Numbers represent categories or "types"
  - Program sites (L.A. = 1, Tucson = 2)
  - Numbers on soccer jerseys
  - Males = 0, Females = 1
- Numbers do NOT represent *quantities*, only *qualities*

Ordinal Measures

- Measures magnitude along a continuum



- Rankings or standings
  - College basketball rankings
  - Most livable cities
  - 1st, 2nd and 3rd place

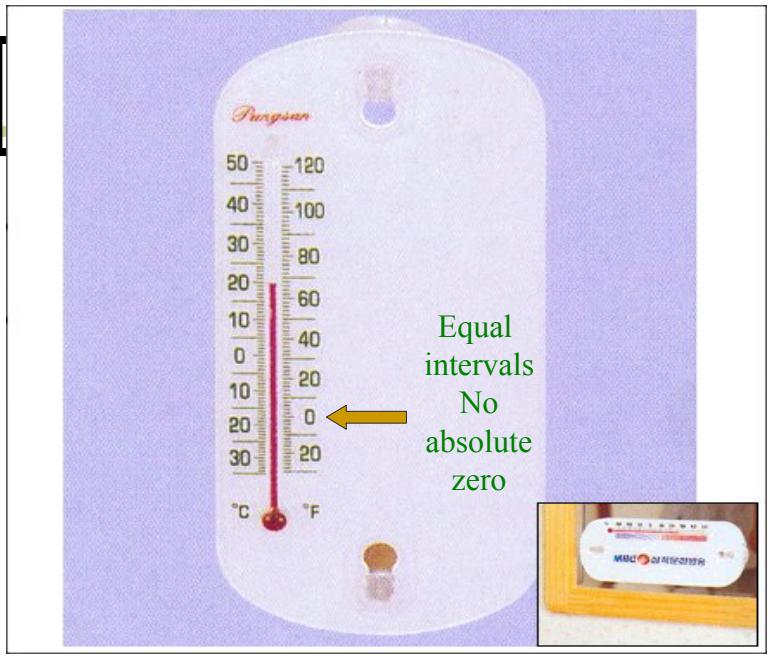
Rankings: unequal intervals

Interval Measures

Measures magnitude in equal intervals

85	100	115	130
low avg.	Average	High avg	Genius

I.Q. Scores



Ratio Measures

- Measures magnitudes in equal intervals with an absolute 0 point
- Absolute 0 means there is ZERO of that quantity
  - time
  - financial support
  - # of times a kid hits another kid
  - scores on Math Achievement test

What to do with our numbers?

- Numbers represent a phenomenon of interest in terms of specific properties (e.g., type, magnitude, frequency, etc.)
- Statistics help us make sense of the numbers
  - frequencies
  - descriptive statistics
  - inferential statistics

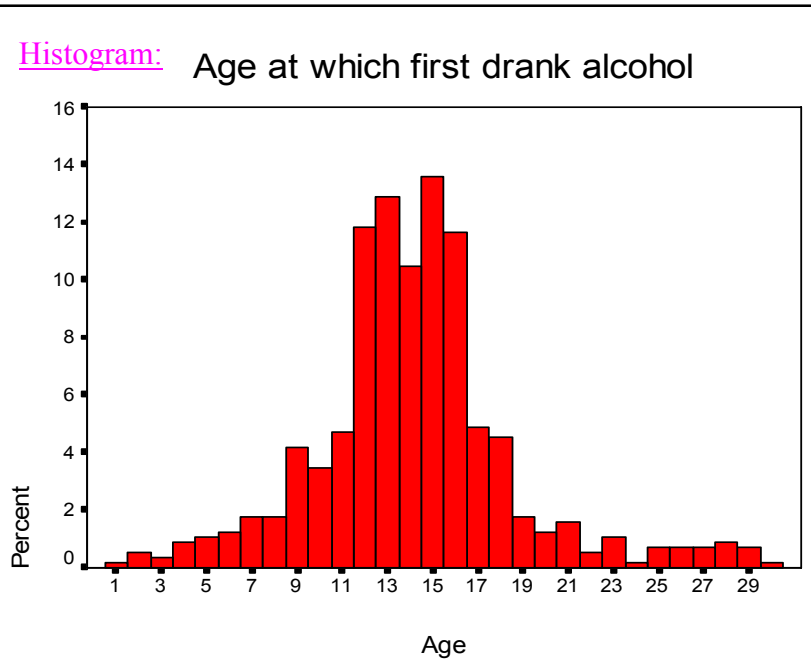
Frequencies

- Counts
- Tells us how much of something we have
- Represent with tables, graphs
- Graphs
  - Bar graphs for nominal, ordinal data (bars don't touch)
  - Histograms for interval, ratio data (bars touch)

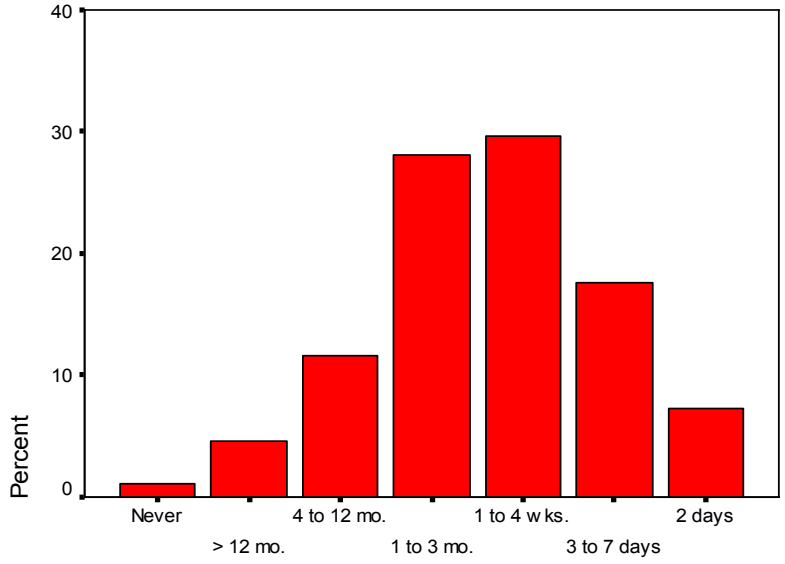
Age first time used alcohol

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	.2	.2	.2
	2	3	.5	.5	.7
	3	2	.3	.3	1.0
	4	5	.9	.9	1.9
	5	6	1.0	1.0	3.0
	6	7	1.2	1.2	4.2
	7	10	1.7	1.7	5.9
	8	10	1.7	1.7	7.7
	9	24	4.1	4.2	11.8
	10	20	3.4	3.5	15.3
	11	27	4.6	4.7	20.0
	12	68	11.6	11.8	31.9
	13	74	12.7	12.9	44.8
	14	60	10.3	10.5	55.2
	15	78	13.4	13.6	68.8
	16	67	11.5	11.7	80.5
	17	28	4.8	4.9	85.4
	18	26	4.5	4.5	89.9
	19	10	1.7	1.7	91.6
	20	7	1.2	1.2	92.9
	21	9	1.5	1.6	94.4
	22	3	.5	.5	94.9
	23	6	1.0	1.0	96.0
	24	1	.2	.2	96.2
	25	4	.7	.7	96.9
	26	4	.7	.7	97.6
	27	4	.7	.7	98.3
	28	5	.9	.9	99.1
	29	4	.7	.7	99.8
	40	1	.2	.2	100.0
	Total	574	98.3	100.0	
Missing	System	10	1.7		
Total		584	100.0		

Histogram:



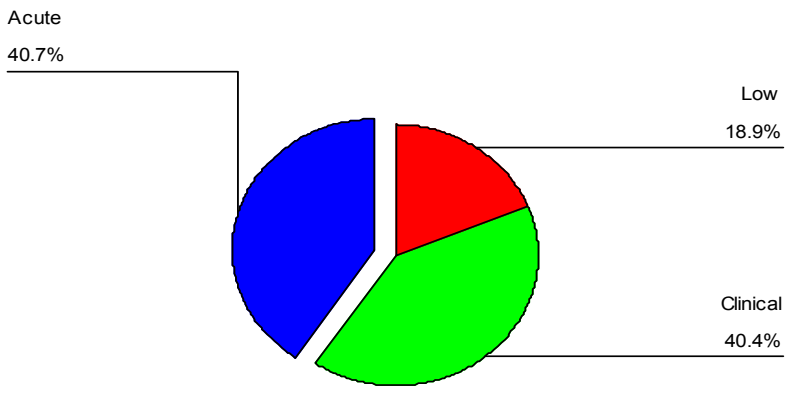
Bar graph: Recency of any kind of ETOH use



Pie chart:

30 month follow up

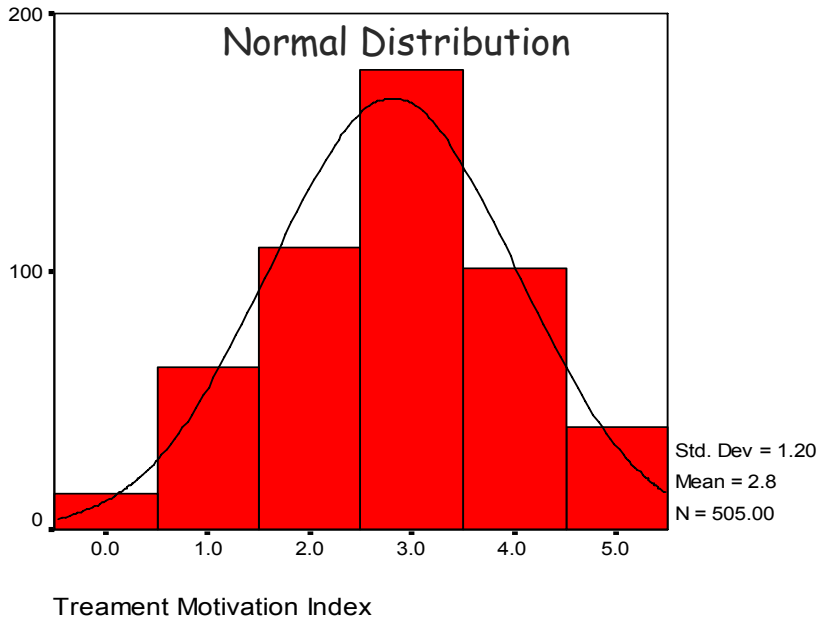
substance use diagnosis



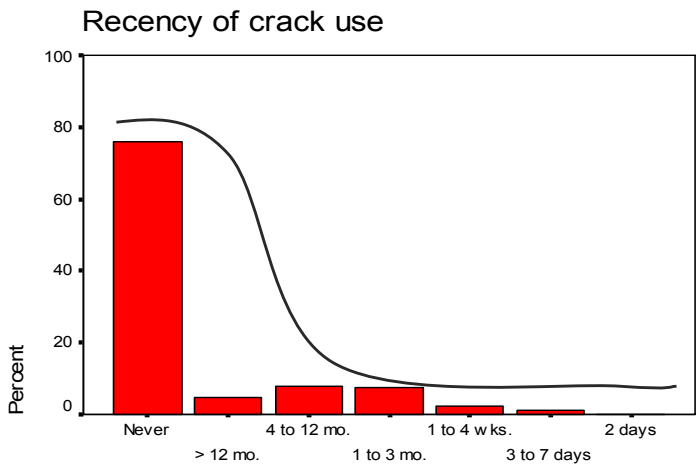
Data distributions:
How scores spread out

Many types, such as…

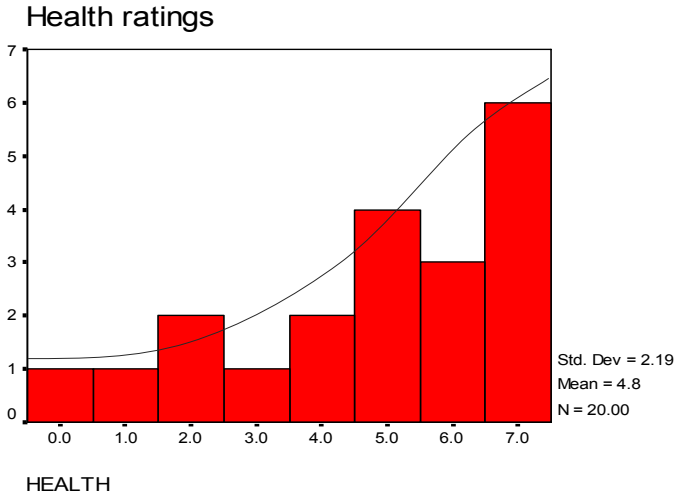
- Gaussian normal (bell-shaped curve)
- Skewed
  - Positive skew (scores “pile up” at low end)
  - Negative skew (scores “pile up” at high end)
- Bimodal



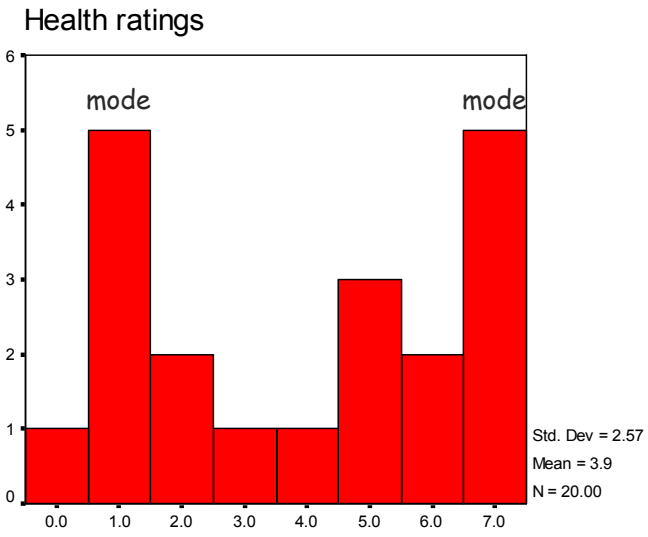
Positively Skewed Distribution



Negatively Skewed Distribution



Bimodal Distribution

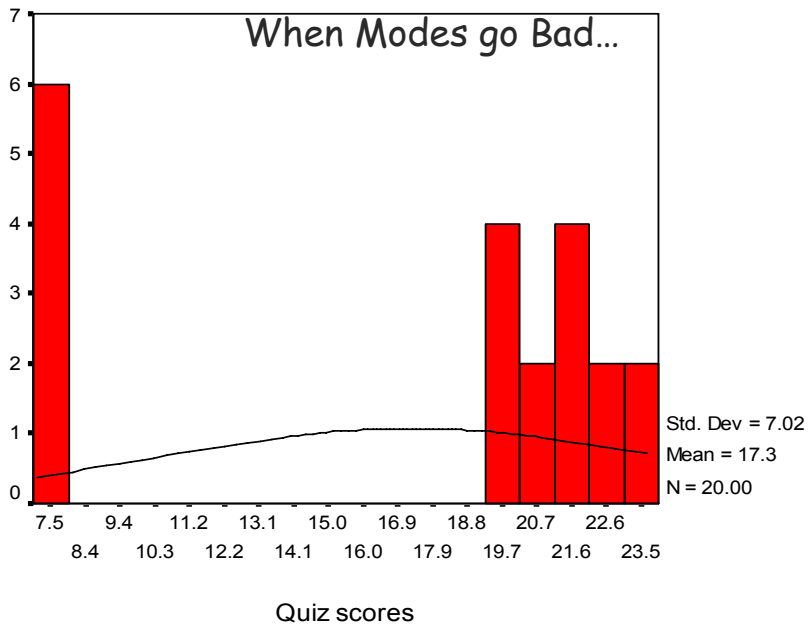


Descriptive Statistics: I. Measures of Central Tendency

- Indicate where the center of the distribution tends to be
  - Mode—most frequent score
  - Median—50th percentile
  - Mean—average score, the mathematical center of a distribution

The Mode

- Useful with nominal data
- In skewed distributions, problematic—may not reflect where the majority of scores are
- Reflects limited information about the data



Median

- 50% of scores are above and 50% are below this value
- Usually located where most of the distribution occurs (improvement on the mode)
- Preferred with ordinal scores and skewed interval and/or ratio scores
  - U.S. Gov't reports median income from Census data

The Median saves the day…

MODE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7.00	6	30.0	30.0	30.0
	20.00	4	20.0	20.0	50.0
	21.00	2	10.0	10.0	60.0
	22.00	4	20.0	20.0	80.0
	23.00	2	10.0	10.0	90.0
	24.00	2	10.0	10.0	100.0
	Total	20	100.0	100.0	

Statistics

MODE

N	Valid	20
	Missing	0
Mean		17.3000
Median		20.5000
Mode		7.00

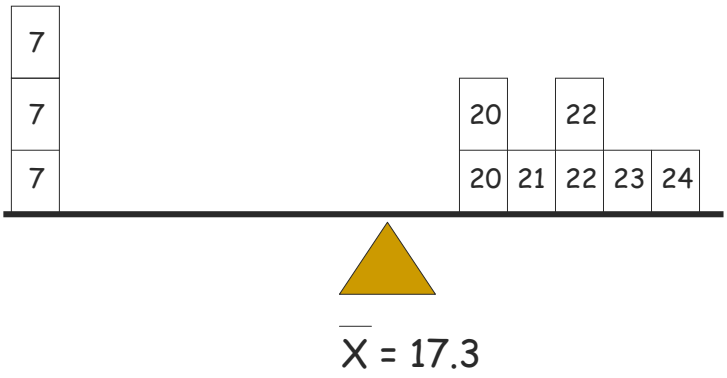
The Median can be insensitive…

- Income for 3 individuals:
  - $0, $5000, $X
    - $5000 = median, so what is X?
  - Median is simply the middle value, so X can be anything (eg, $10K, $100K, $1M, etc)

The Mean

- The mathematical center (balance point) of a distribution
- Useful with interval & ratio data, symmetrical distributions
- To “balance” a skewed distribution, the mean is pulled toward the tail (thus it’s not where MOST of the scores are located)

The Mean has good balance



$\overline{X} = 17.3$

Measures of Central Tendency: Rules of Thumb

- Compute the mode with nominal data or distinctly bimodal distribution of any type of scores
- Compute the median with ordinal data or very skewed distribution of interval/ratio data
- Compute the mean with a symmetrical, unimodal distribution of interval or ratio data

Descriptive Statistics
II. Measures of Variability

- How large are the differences among the individual scores?
- Provides more complete description of the data than merely focusing on central tendency

3 Different Distributions, Same mean

Sample A	Sample B	Sample C
0	8	6
2	7	6
6	6	6
10	5	6
12	4	6
Mean = 6	Mean = 6	Mean = 6

Deviation

- Distance of a score from the mean (raw score minus the mean)
- Size and sign matter
- Sum of all deviations around the mean = 0 (it's the mathematical center!)
  - Basic rule of stats: if we can't perfectly describe every score, we need a number that can, that over- and underestimates equally. In the long run, "errors" using the mean cancel out

Deviation

Sample A	Sample B	Sample C
0 – 6 = -6	8 – 6 = 2	6
2 – 6 = -4	7 – 6 = 1	6
6 – 6 = 0	6 – 6 = 0	6
10 – 6 = 4	5 – 6 = -1	6
12 – 6 = 6	4 – 6 = -2	6
Mean = 6	Mean = 6	Mean = 6

Range

- Communicates the spread of the data
- Subtract lowest from highest score (24 – 7 = 17)
- Compare range between groups
  - Math scores ranged 20 points for School A & 5 points for school B
  - 100 cyclists finishing in 1st – 5th places vs. 1st – 75th places

Variance

- Measures distance from the mean
- Calculated as the squared average deviation from the mean (e.g., 4 *squared years*)
- Larger values = more variability (i.e. scores are more spread out)

Formula for Sample Variance

$$S^2_x = \frac{\Sigma(X - \bar{X})^2}{N}$$

Variance: based on deviation

Sample A	Sample B	Sample C
$0 - 6 = -6^2 = 36$	$8 - 6 = 2$	6
$2 - 6 = -4^2 = 16$	$7 - 6 = 1$	6
$6 - 6 = 0^2 = 0$	$6 - 6 = 0$	6
$10 - 6 = 4^2 = 16$	$5 - 6 = -1$	6
$12 - 6 = 6^2 = 36$	$4 - 6 = -2$	6
Mean = 6 Variance = 104/5 = 20.8	Mean = 6 Variance = ?	Mean = 6 Variance = ?

Standard deviation

- Square root of variance, measured in the same units as the raw scores
  - (e.g., "scores differ from the mean age by an average of 2 *years*")
- Larger values = more variability in scores
- More variability = scores are farther from the mean → measure of how accurately mean summarizes scores
- Tells us
  - *Average deviation from the mean*
  - *Consistency in scores*
  - *How far scores spread out around the mean*

Formula for Sample Standard Deviation

$$S_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

Variance: based on deviation

Sample A	Sample B	Sample C
0 – 6 = $-6^2 = 36$	8 – 6 = 2	6
2 – 6 = $-4^2 = 16$	7 – 6 = 1	6
6 – 6 = $0^2 = 0$	6 – 6 = 0	6
10 – 6 = $4^2 = 16$	5 – 6 = -1	6
12 – 6 = $6^2 = 36$	4 – 6 = -2	6
Mean = 6 Variance = 20.8 Std Dev = 4.56	Mean = 6 Variance = ? Std Dev = ?	Mean = 6 Variance = ? Std Dev = ?

Descriptive Statistics

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
How old are you today	523	7.00	12.00	19.00	15.8298	1.12942	1.276
P90 - Most days gone without using AOD	521	90.00	.00	90.00	26.6276	25.42046	646.200
General Mental Distress Index	523	21.00	.00	21.00	8.3098	5.03927	25.394
Valid N (listwise)	521						

Let's check out our survey…

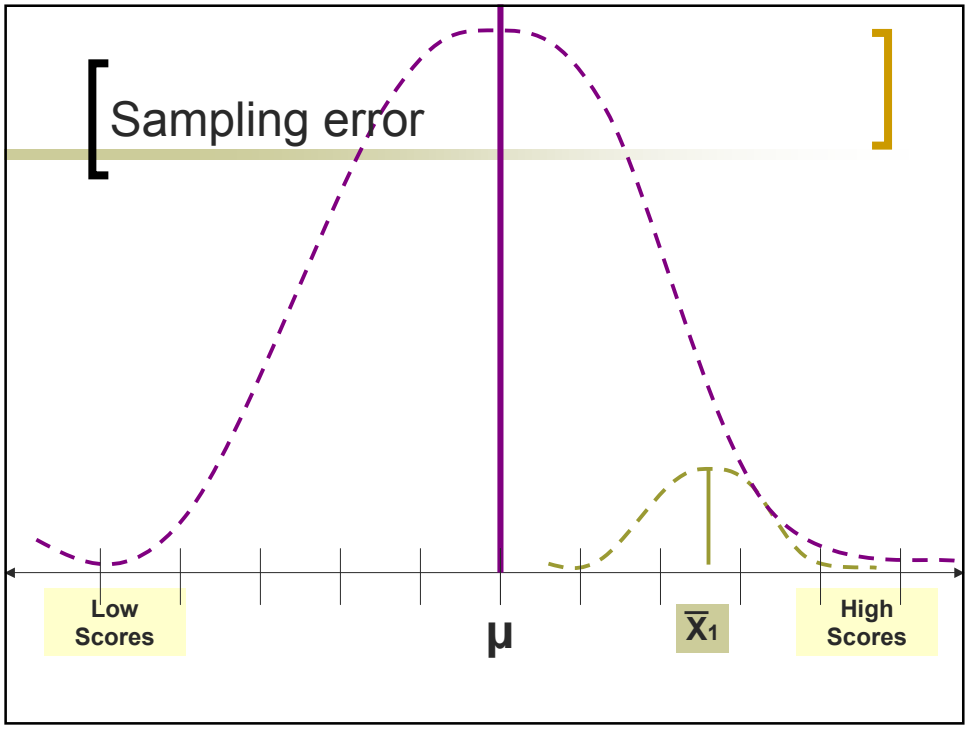
Inferential statistics: Samples & Populations

- Goal of research: infer from sample to populations
- No guarantee that our sample accurately reflects the population
  - *Representative sample* = characteristics of individuals & scores accurately reflect those found in the population
    - Must select via random sampling
- Any sample may not be perfectly representative, so sample *statistics* ≠ population *parameters* they represent

Sampling error

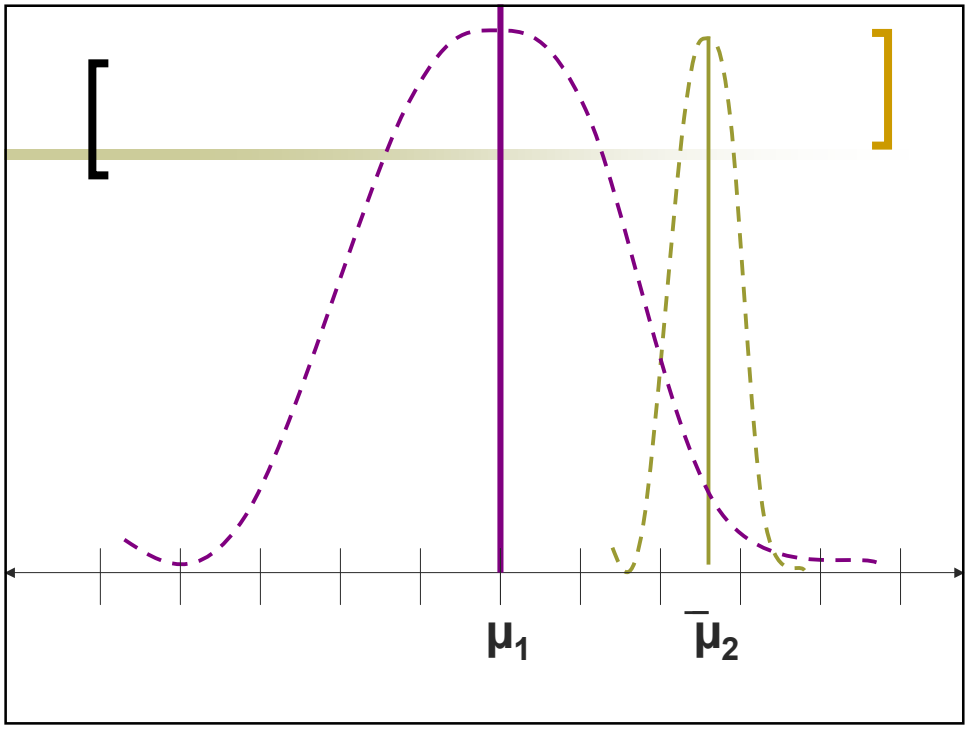
- Occurs when random chance produces a sample statistic not equal to the population parameter
- Summarized as a statistic quantifying the probability that chance produced an unrepresentative sample

Sampling error



Interpreting sampling error

- The reason a sample mean is different from μ is because by chance, the sample is unrepresentative
  - "luck of the draw" produced sample with too many high or too many low scores
- With sampling error, can't be sure which population our sample represents
  - (1) poorly represents one population due to sampling error
  - (2) accurately represents another population

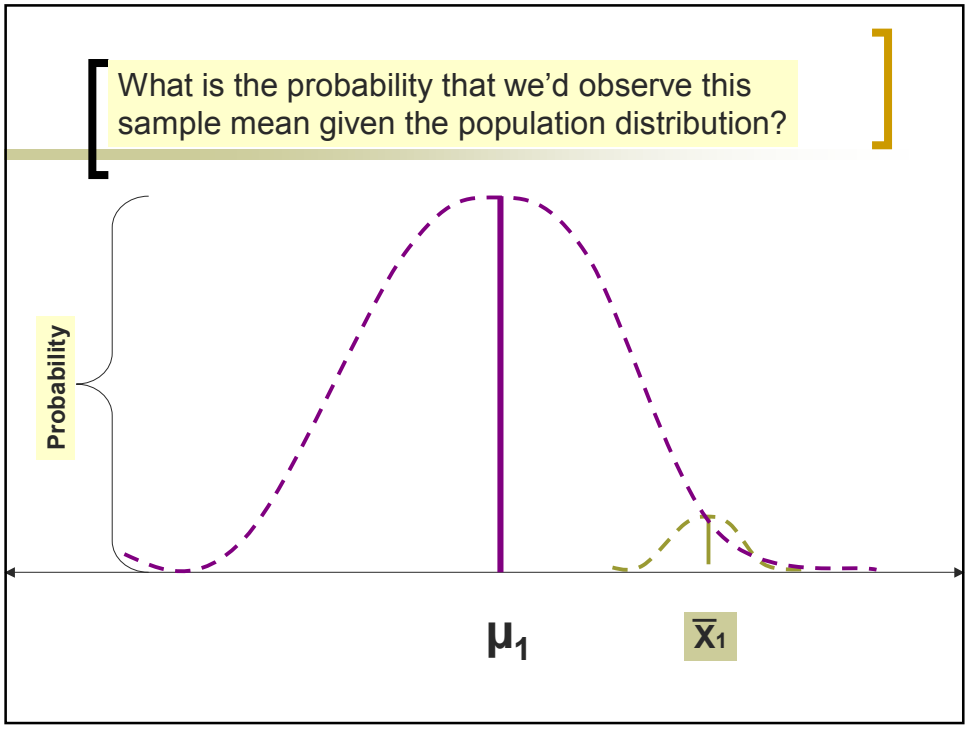


Inferential statistics

- Used to make the decision about the population being represented by our sample
  - Example: a sample mean is higher than the known population mean (μ). Is it because of sampling error, or is the sample different (e.g., more intelligent) than the population?

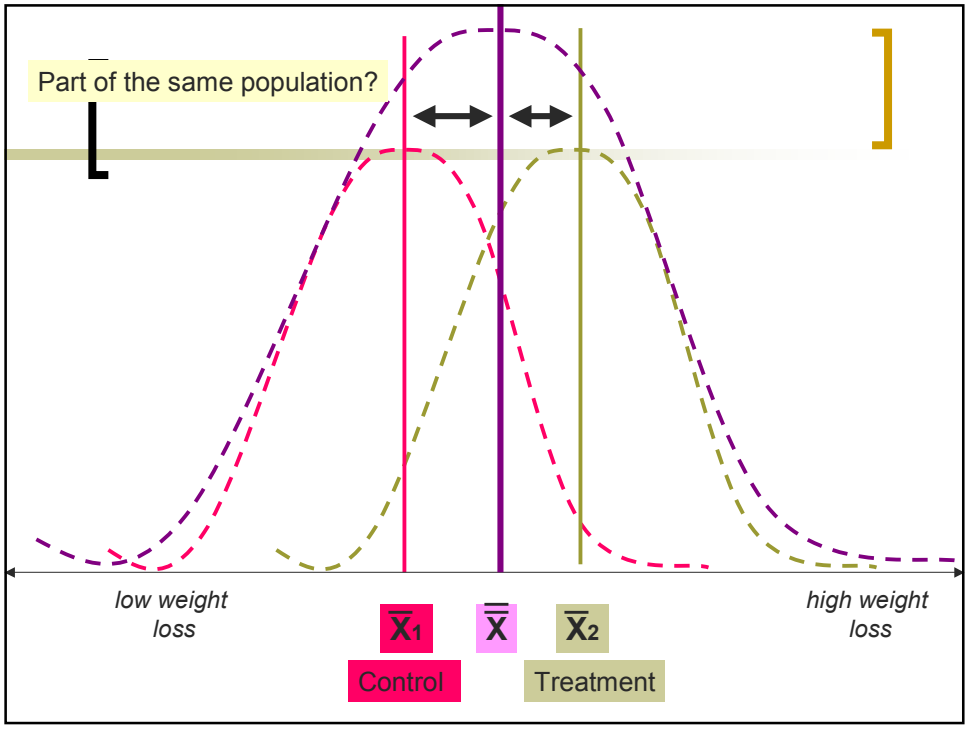
Probability & inferential statistics

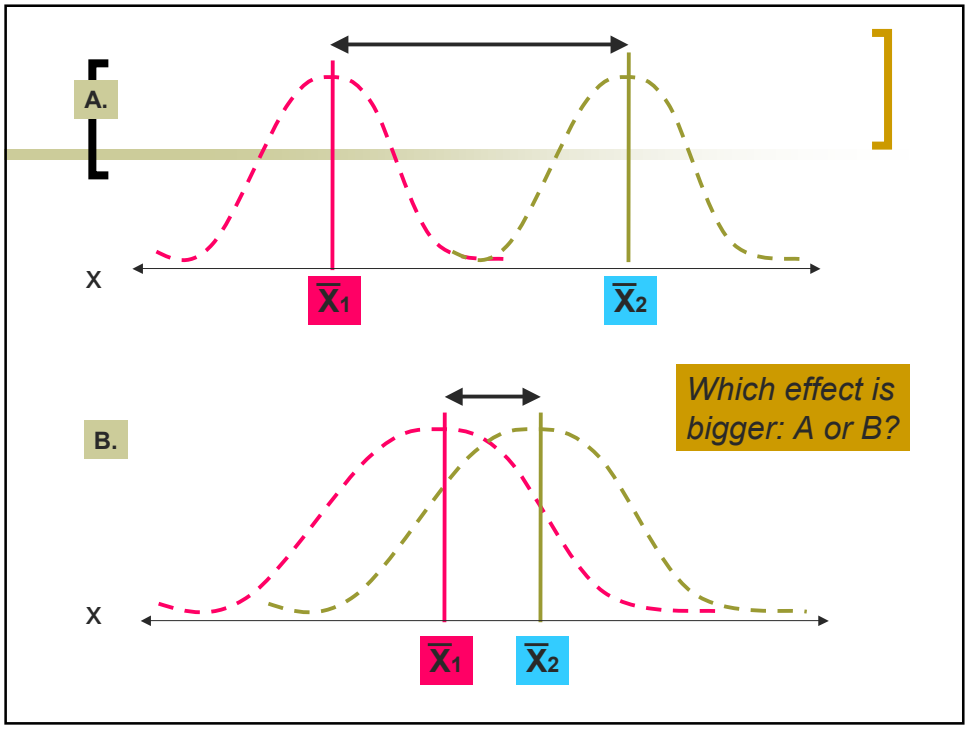
- Probability helps us decide whether our sample is likely to come from (and thus represent) a population
  - If likely, we decide we have a representative sample
  - If unlikely, we decide the sample represents some other population
  - We do not expect a *perfectly* representative sample (e.g., sample mean = population mean)

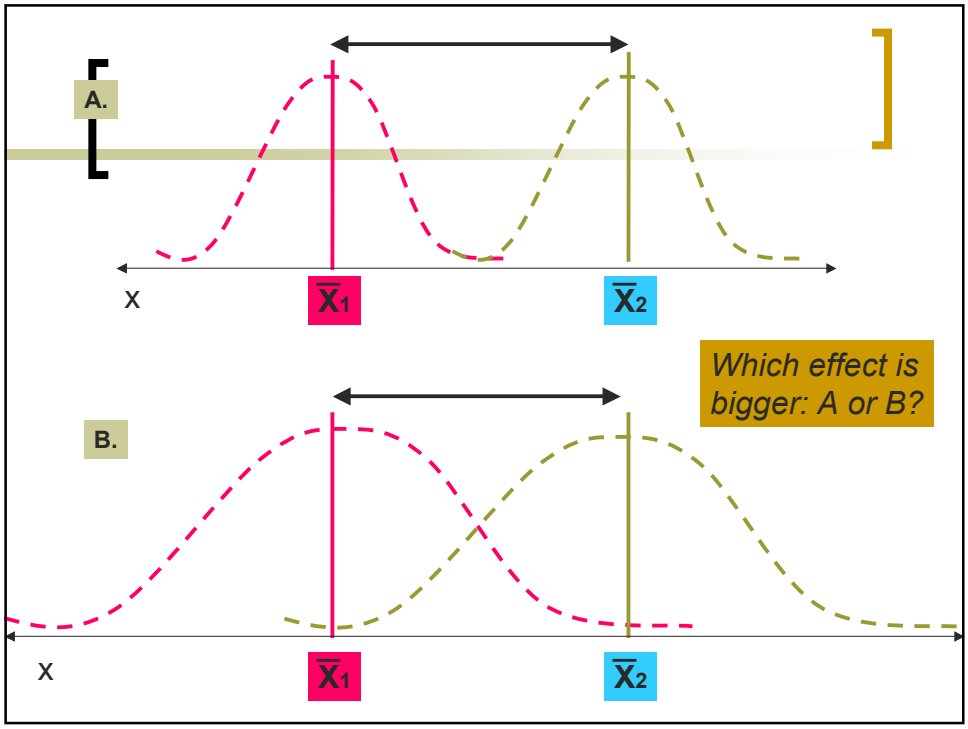


The role of inferential statistics in experimental research

- Experiments: we expect change in IV produces consistent change in DV
- If scores in each condition differ in our study, we'd expect to see that in nature
- Due to sampling error, the true relationship between IV and DV might be different in nature
- Use inferential stats to decide whether sample data represent a particular relationship in the population


Part of the same population?
low weight loss
high weight loss
$\bar{X}_1$
$\bar{\bar{X}}$
$\bar{X}_2$
Control
Treatment


A.
X
$\bar{X}_1$
$\bar{X}_2$
B.
X
$\bar{X}_1$
$\bar{X}_2$
Which effect is bigger: A or B?



Parametric & Nonparametric

- Both used to decide if the data reflect a relationship in nature or if sampling error is misleading us
- Parametric: Require specific assumptions about raw scores of the population being represented
  - Population dependent scores approx. normally distributed
  - Interval or ratio scores
- Nonparametric: no stringent assumptions
  - Nominal or ordinal scores
  - Skewed interval or ratio distributions

Setting up inferential procedures

- Create the study hypothesis
- Design and conduct the study to test the hypotheses (via data)
- Translate the study hypotheses into statistical hypotheses
- Perform the statistical test of the statistical hypotheses

Null & Alternative Hypotheses

- **Alternative (H_1):** describes the population parameters the sample data represent *if the predicted relationship exists*.
- **Null (H_0):** describes the population parameters that the sample data represent *if the predicted relationship DOES NOT exist*.

  Maintains that the sample data may contain sampling error

Back to our survey…

What questions might we test with our survey using inferential statistics?

Interpreting significant results

- Our results are unlikely to occur if the predicted relationship does NOT exist in the population. Thus, we can reject H_0
  - We did not PROVE that H_0 is false
  - We have not proven our IV produced the DV: must consider confounds or alternative explanations
  - If H_1 is correct, the population parameters not *exactly* what we observed with our sample

Interpreting Nonsignificant results

- The relationships found were likely due to chance, sampling error, without there being a relationship in nature.
  - We did not PROVE that H_1 is false: we have failed to find convincing evidence to support it.
  - The only certainty: sampling error *could* have produced our data, thus 2 viable hypotheses:
    - (1) H_0 is correct, there's no relationship in nature
    - (2) H_1 is correct and we have failed to support it due to sampling error

Inferential Statistics
Evaluation Goal: Compare Groups
Independent Samples t-test

- 2 samples are independent when we select participants for a sample without regard to who else has been selected for either sample
- Scores in 1 sample not influenced (i.e. independent) by other sample

Hypotheses

- H_1: there is a relationship between school program and achievement; so mean score (achievement test) for program clients > mean for students in the old program
- H_0: there is no relationship; the means for both groups will be the same

Example: Starting salaries

- H_1: there is a relationship between gender and starting salary; so avg starting salaries for males > avg starting salaries for females
- H_0: there is no relationship; the means for both groups will be the same
  - Remember, "the same" doesn't mean exactly the same—values fall within upper & lower limits of each other

Example: Independent Samples T-test: Do males & females differ on starting salary?

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Starting Salary	Female	469	24769.51	6895.765	318.417
	Male	631	27026.51	6870.097	273.494

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Starting Salary	Equal variances assumed	.034	.854	-5.380	1098	.000	-2257.00	419.517	-3080.142	-1433.850
	Equal variances not assumed			-5.377	1006.360	.000	-2257.00	419.748	-3080.678	-1433.314

Reporting results

$t(1098) = -5.38, p < .05$

Back to our survey…

What questions might we test with our survey using independent samples t-test?

Inferential statistics
Evaluation Goal: check association between 2 variables - *Pearson's r*

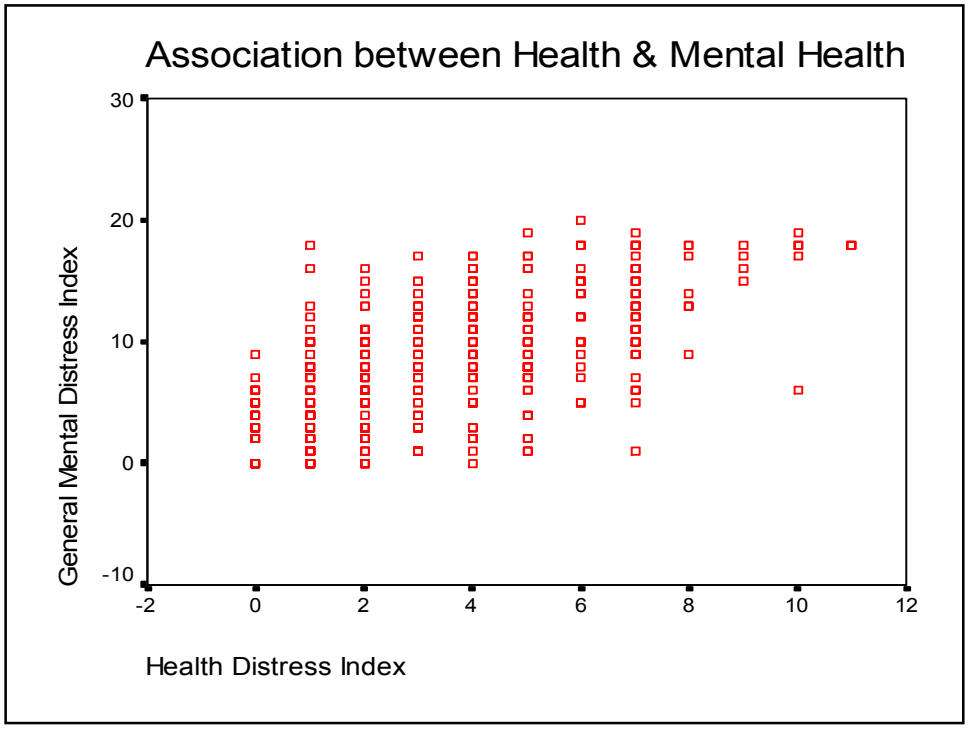
- In a relationship, as the scores in 1 variable change, there is a consistent pattern of change in the scores on the other variable
- In research, we want to describe and summarize the relationship
- Correlation coefficients summarize relationships

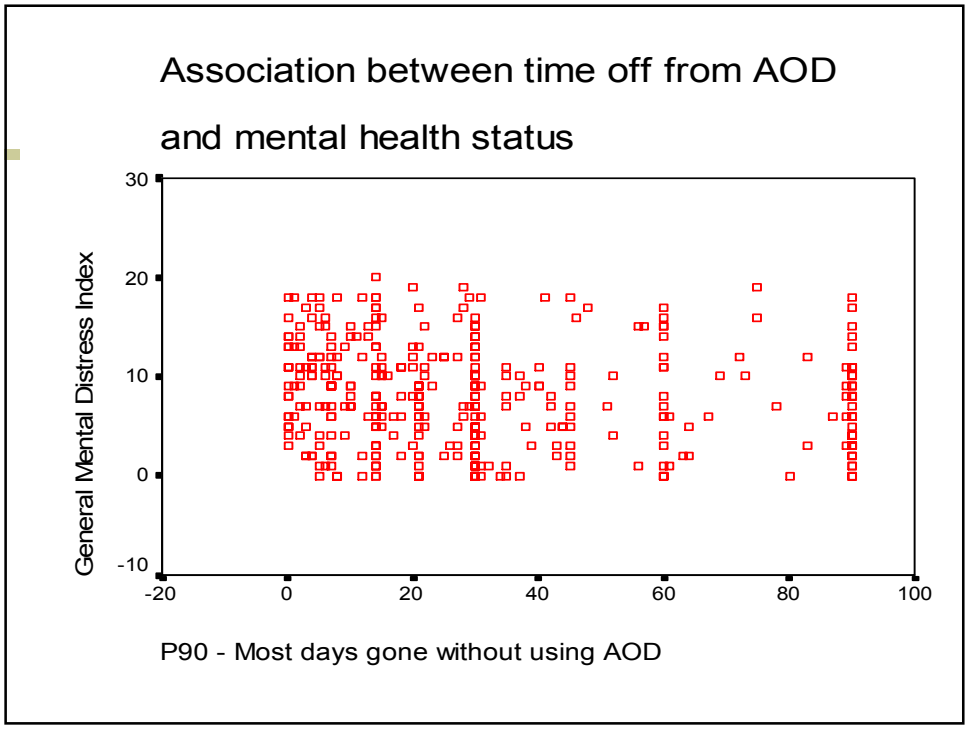
Correlation coefficient

- A number that describes a relationship
- Size – tells us strength of the relationship (how consistently scores change together). Ranges from -1 to 1.
- Sign –tells us the direction (positive vs negative) of a relationship.
  - Example: the correlation between parental and offspring IQ is approximately 0.50

Scatterplot

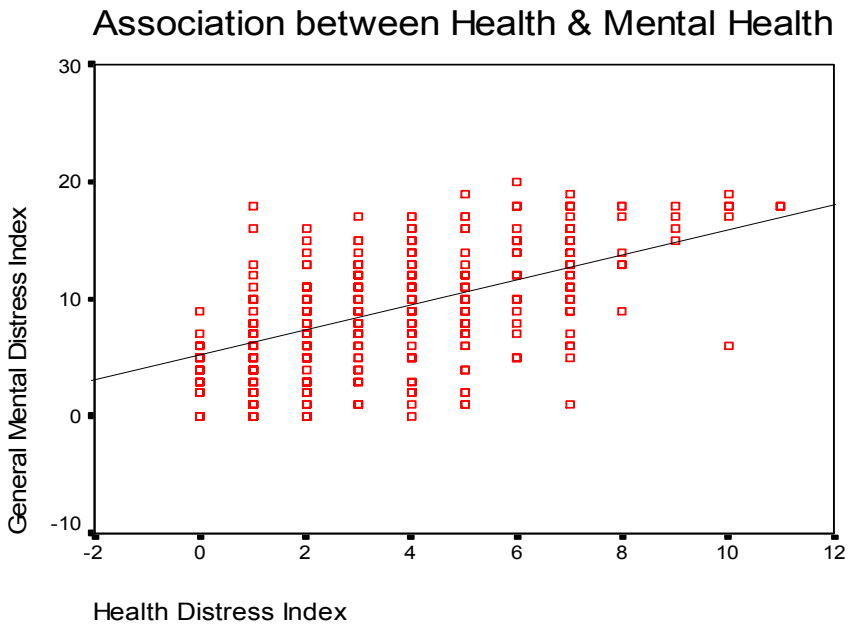
- A graph that shows the location of each data point formed by a pair of X-Y scores
- 2 types of relationships:
  - Linear: summarized by a line; as scores on X increase, scores on Y tend to change in only 1 direction (increase or decrease)
  - Nonlinear: cannot be summarized by a *straight* line. As X scores change, Y scores not only increase or decrease, they alter direction of change at some point

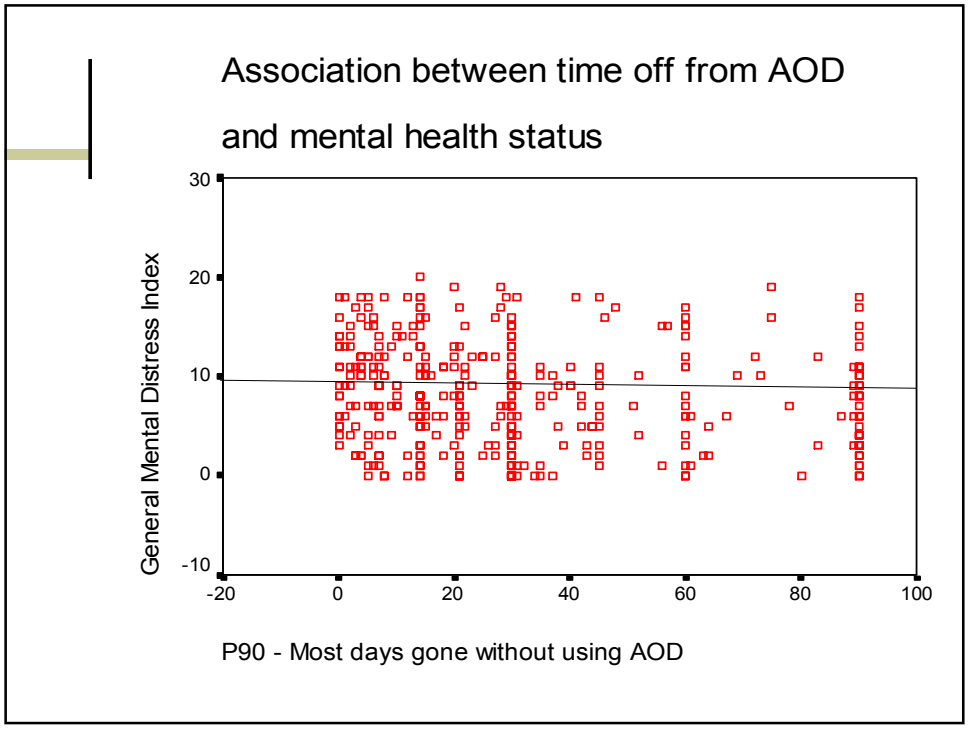

Association between Health & Mental Health
General Mental Distress Index
30
20
10
0
-10
-2
0
2
4
6
8
10
12
Health Distress Index


Association between time off from AOD
and mental health status
General Mental Distress Index
30
20
10
0
-10
-20
0
20
40
60
80
100
P90 - Most days gone without using AOD

Regression line (line of "best fit")

- Can summarize a linear relationship by drawing a line through the scatterplot.
  - Line passes thru center of scatterplot so distance of the points above is equal to distance below
- No relationship: no pattern of scores, regression line is (nearly or perfectly) horizontal
- Positive relationship: /
- Negative relationship: \

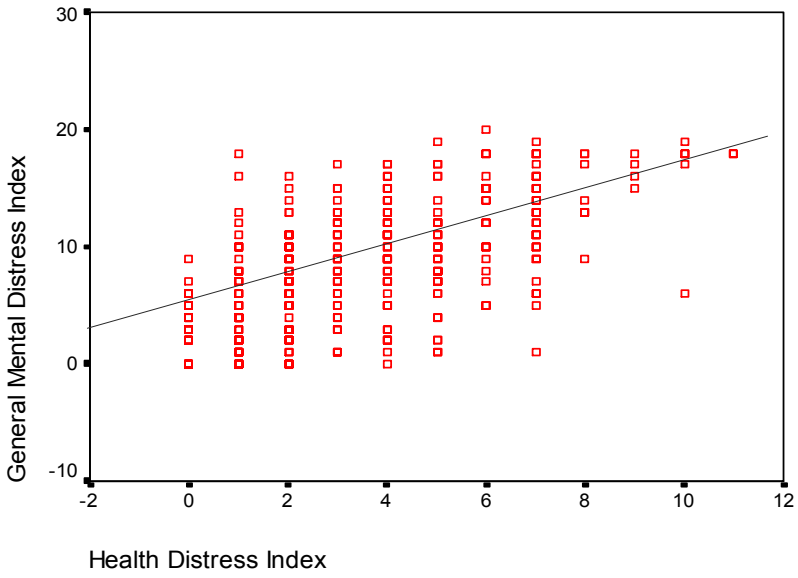




Pearson's *r*

- Describes relationship between 2 interval and/or ratio variables. Compares how consistently each value of Y is paired with each value of X
- Ranges from – 1 to + 1
- Values at or near 0 mean weak to no relationship
- Values at or near -1 or +1 mean very strong to perfect relationship

Association between Health & Mental Health



Correlation between physical & mental health

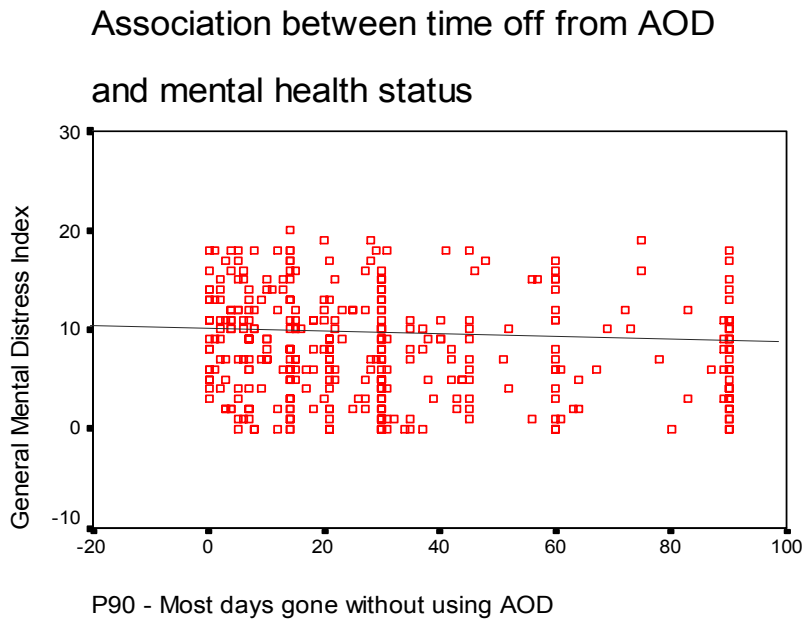
Correlations

		Health Distress Index	General Mental Distress Index
Health Distress Index	Pearson Correlation	1	.621**
	Sig. (2-tailed)	.	.000
	N	377	377
General Mental Distress Index	Pearson Correlation	.621**	1
	Sig. (2-tailed)	.000	.
	N	377	378

**. Correlation is significant at the 0.01 level (2-tailed).

Significance testing for *r*

- Tests that the sample correlation coefficient *r* accurately represents the relationship found in the population (called *rho* or ρ)
- Alternative hypothesis always states that a relationship exists, or H_1: $\rho \neq 0$
- H_0 states that predicted relationship does not exist or H_0: $\rho = 0$



Correlation between mental health & time off From drug use

Correlations

		General Mental Distress Index	P90 - Most days gone without using AOD
General Mental Distress Index	Pearson Correlation	1	-.116*
	Sig. (2-tailed)	.	.024
	N	378	377
P90 - Most days gone without using AOD	Pearson Correlation	-.116*	1
	Sig. (2-tailed)	.024	.
	N	377	377

*. Correlation is significant at the 0.05 level (2-tailed).

Reporting results

$$r\ (377) = -.12,\ p < .05$$

Proportion of variance accounted for (r^2)

- Describes proportion of all differences in Y scores associated with changing the X variable.
- Square Pearson's *r*
- With *r* = -.45 , r^2 = .20, therefore 20% of variance in scores on Y is associated with differences in X (and 80% are not)

Back to our survey…

What questions might we test with our survey using Pearson's *r* ?

Nonparametric statistics

- Inferential statistics, therefore used for deciding whether relationships in the sample accurately represent those in the population
- Use when dependent variable scores are nominal or ordinal, or populations are severely skewed
- If data severely violate assumptions of parametric stats (which are quite robust), Type I error rate increases (i.e. stats tests are less reliable)

Inferential, Nonparametric Statistics
Evaluation Goal: Assess relationship
Contingency Chi Square (a.k.a. "Cross Tabs")

- Assesses relationship between 2 categorical/nominal variables (e.g., smoking (Y/N) & lung cancer (Y/N))
- No distributional assumptions
- Compares expected to observed frequency

Contingency Chi Square: rurality x substance use group

rural (< 50,000) vs. not * Grouped Recency of Any Substance Use Crosstabulation

Count

		Grouped Recency of Any Substance Use			
		low	clinical	acute	Total
rural (< 50,000) vs. not	not rural	14	225	34	273
		6	180	64	250
Total		20	405	98	523

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.404[a]	2	.000
Likelihood Ratio	16.623	2	.000
Linear-by-Linear Association	16.138	1	.000
N of Valid Cases	523		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.56.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.177	.000
	Cramer's V	.177	.000
N of Valid Cases		523	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the hypothesis.

Reporting results

$$\chi^2 (2, N = 523) = 16.4, p < .05$$

Back to our survey…

What questions might we test with our survey using Contingency Chi-Square?