

## A quick review

- What does  $r(23) = -.67$  mean?
- What does  $t(24) = 3.03$ ,  $p < .05$  mean?
- What does  $t(22) = 1.03$ ,  $p = .13$  mean?

## Null results

- One issue from the last topic is that of null results
  - A significant result tells us that something is unlikely to have happened by chance
  - A null result really doesn't tell us anything one way or the other

## Implications of nulls

- For this reason, many journals refuse to publish papers where there are no significant results.
- But this raises issues - if you're testing a therapy, you can never prove that it doesn't work - but if you consistently find no improvement, isn't that important?

## Recent implications

- "The NY state attorney, Eliot Spitzer, recently sued the British drug company GlaxoSmithKline charging that the company had not disclosed the results of clinical trials of their antidepressant drug Paxil that failed to show the drug was effective in treating children and adolescents and that suggested a possible increase of risk of suicide.
- The suit was based on the results of three studies paid for by GSK to see the effect of Paxil on treating major depression in in children and adolescents.

**GSK's Reanalysis of Risks**

GSK study	Final report	Suicidal ideation*	
		on Paxil	on placebo
329, M. Keller <i>et al.</i>	2001	5.8%	1.8%
377	1998-99	4.8%	10.6%
701	2000	18.8%	16.0%

\* Emergent suicidal ideation includes self-injurious remarks or behaviors related to suicidal ideation, suicide attempts, self-inflicted harm, or overdoses.

Taken from CHANCE News 13.05, June - Nov. 2004, Copyright 2004 Laurie Snell

## Recent implications

**GSK's Reanalysis of Risks**

GSK study	Final report	Suicidal ideation*	
		on Paxil	on placebo
329, M. Keller <i>et al.</i>	2001	5.8%	1.8%
377	1998-99	4.8%	10.6%
701	2000	18.8%	16.0%

\* Emergent suicidal ideation includes self-injurious remarks or behaviors related to suicidal ideation, suicide attempts, self-inflicted harm, or overdoses.

- Only study 329 was published. The suit claimed that none of the results were significant and charges GSK with publicizing the apparent favorable study 377 and suppressing information about the other two studies that suggested that the drug had no effect on behavior related to suicides. ...
- The suit was settled in September with the company paying 2.5 million dollars and agreeing to post online both negative and positive results from its clinical drug trials."

Taken from CHANCE News 13.05, June - Nov. 2004, Copyright 2004 Laurie Snell

## When you have three independent groups...

- You cannot simply do all combinations of t-tests
  - That would increase the chance of a Type 1 error
- Need to compare all 3 groups at the same time: one-way Analysis of variance (ANOVA)

## 1-way ANOVA

- Compares the amount of variability **between** groups to the amount of variability **within** groups.
  - You always have some variability among members of the same group.
  - If the three groups are really the same, the variability between the groups should be similar to that within the groups.
  - If the groups differ, you should have more variability across them than within them.
  - That is, you'd have the variation caused by the systematic differences between groups added to the differences caused by random noise.

## Why the name?

- “ANOVA” because you are analyzing the variance.
- “1-way” because the groups differ along one dimension.
- An “F” test.

## Similarities to t-test

- You still need to pick your alpha level in advance.
  - We typically use .05
- The greater the number of subjects, the lower the difference between groups needed for significance.

## Differences from t-test

- You have two different df that get reported
  - df between the groups (the number of groups – 1)
  - df within groups (the # of subjects –1 for each group, summed)
  - $F(df,df)=***, p<***$
- Since it is a ratio of the amount of variability between groups compared to within groups, no difference results in F approaching 1.
  - With t-tests, null effect approached zero.

## Differences from t-test, cont.

- An F test tells you whether the groups differ in some way. It does not tell you which one is bigger.
  - That is, an F test says there is a difference, but it doesn't say anything about the source of the difference.
- Thus, a significant ANOVA requires follow-up tests to see WHICH groups differ from one another.
- Since the F doesn't say which group is bigger, there is no directionality -- F is always positive.
  - So you can not do the equivalent of a 1-tailed test

## Repeated measures

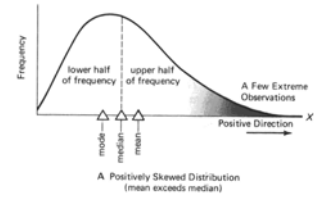
- You still need to distinguish whether your groups are independent from one another (between-subjects analysis) or matched (within-subjects analysis).
- For the latter, you do a repeated measures ANOVA.
  - This is interpreted the same way, with the same tables.

### Parametric vs. nonparametric tests

- t-tests & ANOVA are parametric tests.
- These can only be done when the data is integer or ratio -- not with rankings.
- These tests also assume that the population is normally distributed

### Parametric vs. nonparametric tests

- In a normally distributed population, median & mean are equal
- In a skewed distribution, they aren't.



But parametric tests assume they are.

Image: R. S. Witte, Statistics, 4th ed.

### How much TV?

HI	Non-HI	
11	15	• Each group has one really extremely high value, suggesting the populations are positively skewed.
3	42	
5	1	• When you have small sample sizes, violations of the normality assumption could increase the risk of error.
20	0	
10	2	
4	8	
53	12	

### Nonparametric tests

- Can be used on ranked data
- For skewed data, can rank order the scores, and then treat as ranked.
  - This avoids errors caused by the extreme values, but then also requires the use of one of these nonparametric tests.

### Spearman's Rho ( $\rho$ ) or rank-order

- Rationale:
  - If you rank order a group of individuals on two different variables, a perfect correlation would be when each person had the same ranking on both.
  - Given that, the difference in rankings can be taken as an indication of disparity between the two sets of rankings.

### So...

- Pearson's product-moment
  - Both scales are continuous, interval or ratio, and linear
- Spearman's rho (or Spearman's rank-order)
  - Both scales are rank ordered

## Other correlations you might see

- Point-Biserial ( $r_{pb}$ )
  - One scale is dichotomous (i.e., nominal; ex: gender, race) and the other continuous (interval or ratio)
- Biserial ( $r_{bi}$ )
  - One scale is continuous and the other dichotomous, but artificially so (for ex.; pass/fail, where there really is an underlying continuous variable)
- Phi ( $\phi$ ) or tetrachoric ( $r_{tet}$ )
  - Both scales are dichotomous (phi - naturally, tetrachoric - artificially)
- Gamma
  - One scale nominal, one ordinal
- Eta ( $\eta$ ) coefficient
  - For nonlinear relationships

From Brown, J. D. (2001). Statistics Corner. Questions and answers about language testing statistics: Point-biserial correlation coefficients. *Journal of Testing & Evaluation SIG Newsletter*, 3(3), 12-15. Also at [http://www.ielt.org/test/bro\\_12.htm](http://www.ielt.org/test/bro_12.htm) and from <http://www.gesis.acfa.edu/courses/ed230/notes3/cor3.html>

## Mann-Whitney U-test

- Used for two independent groups
  - Nonparametric analog to independent samples t-test
- Measures the number of times that members of one group outrank members of the other group
- A score either very high or very low indicates larger differences
- Can be one-tailed or two-tailed.
- Why not use it all the time? It isn't as powerful if the distribution really is Gaussian.

## Wilcoxin T-test

- Used for two paired groups
  - Nonparametric analog to paired t-test
- Measures whether the scores in one group consistently “outrank” the scores in the other group.
- If sum of positive ranks and sum of negative ranks are very different, the groups differ.

## Kruskal-Wallis H test

- Used for 3 or more groups
- Compares whether the rankings of the different groups are similar.
- Scores in this one are more like the F-test
  - a higher score is more likely to be significant
  - the test itself is nondirectional.

## When to use different tests

	Ordinal	Interval/Ratio
2 paired groups	Wilcoxin T-test	Paired t-test
2 independent groups	Mann-Whitney U-test	Unpaired t-test
3 or more independent groups	Kruskal-Wallis H-test	1-way ANOVA (F-test)

## The main point

- The important thing to remember here is that the more standard tests can only be used when certain assumptions are met.
- When those assumptions are not met, they are not the right tests to be using.

## Chi-squared test ( $\chi^2$ )

- Works on nominal data
- Compares the proportions of people per group in your sample to the proportions you expect based on chance
- The greater the chi-square, the larger the difference
- Does not work well if expected frequencies are very low (<5%)

## Binomial test

- Sometimes referred to as a sign test
- Used for nominal data, when only two choices and only one group
- Determines whether the probability is different from chance
  - Ex: Did my penny flip heads more often than chance?
  - Ex: Does hearing & speech really attract significantly more women than men

## When to use different tests

		Interval/Ratio	Ordinal	Nominal
1 group		One sample t-test		Binomial test
	matched groups	Paired t-test	Wilcoxin T-test	
2 groups	independent groups	Unpaired t-test	Mann-Whitney U-test	Chi-squared test
	matched groups	Repeated measures ANOVA (F)	Friedman Test	
3 or more groups	independent groups	1-way ANOVA (F-test)	Kruskal-Wallis H-test	Chi-squared test

## Sample study 1

- Researcher's hypothesis: women who are experiencing symptoms of menopause will have more difficulties accessing words (more TOTs)
- Subjects: two groups of women, those who are in menopause and those who are not
- Measure: how many times they make mistakes on a naming test (fail to access the word)

## Sample study 2

- Researcher's hypothesis: degree of hearing loss may influence the number of friends someone has and how happy they are with these friends
- Subjects: individuals who are mildly, moderately, severely, or profoundly HI
- Measure: rankings based on results of survey about the quantity and quality of a person's friends

## Sample study 3

- Researcher's hypothesis: pronunciation changes in fluent speech will influence word recognition
- Subjects: normal individuals
- Task: cross-modal priming
  - people hear a word, then see a word or nonword
  - have to decide if the item on the screen is a real word or not.
- Measure: response times to "today" after hearing "yesterday", "yestday" or "telephone"

### Sample study 4

- Researcher's hypothesis: kids whose parents stutter are themselves more likely to stutter
- Measure: The researcher knows that stuttering occurs by chance in X% of the population; compares this to the proportion of people who stutter whose kids also stutter

### Sample study 5

- Researcher's hypothesis: being familiar with a voice encourages infants to listen to that voice
- Subjects: two groups of infants
  - Infants in one group hear their own mothers' voice as the target voice.
  - Infants in the other group each hear the mom of one of the infants in the first group.
- Measure: how long the infants listen to the target voice