Data Analyses Stats & Decisions

- ANOVA & Decision Outcomes
- H0: & RH: -- Not always an “either – or”
- Decision Errors vs. Hypothesis Disconfirmation
- p-values “vs” effect sizes

ANOVA

Between Groups (Independent Samples, etc.)

•H0: Populations represented by the IV conditions have the same mean DV.

•degrees of freedom (df) numerator = 1, denominator = N - 2

•Range of values 0 to ∞

•Reject Ho: If $F_{obtained} > F_{critical}$ or If $p < .05$

Within-groups (Dependent Samples, etc.)

•H0: Populations represented by the IV conditions have the same mean DV.

•degrees of freedom (df) numerator = 1, denominator = N - 1

•Range of values 0 to ∞

•Reject Ho: If $F_{obtained} > F_{critical}$ or If $p < .05$

When doing NHST, we are concerned with making statistical decision errors -- we want our research results to represent what’s really going on in the population.

Traditionally, we’ve been concerned with two types of statistical decision errors:

Type I Statistical Decision Errors

- rejecting H0: when it should not be rejected
- deciding there is a relationship between the two variables in the population when there really isn’t
- a False Alarm
- how’s this happen?
  - sampling variability (“sampling happens”)
  - nonrepresentative sample (Ext Val)
  - confound (Int Val)
  - poor measures/manipulations of variables (Msr Val)
  - Remember the decision rule is to reject H0: if $p < .05$ -- so we’re going to make Type I errors 5% of the time!

Type II Statistical Decision Errors

- retaining H0: when it should be rejected
- deciding there is not a relationship between the two variables in the population when there really is
- a Miss
- how’s this happen?
    - sampling variability (“sampling happens”)
    - nonrepresentative sample (Ext Val) poor
    - confound (Int Val)
    - poor measures/manipulations of the variables (Msr Val)
    - if the sample size is too small, the “power” of the statistical test might be too low to detect a relationship that is really there (much more later…)

This is what we referred to as “statistical conclusion validity” in the first part of the course.

- Whether or not our statistical conclusions are valid / correct ??

These are the two types of statistical decision errors that are traditionally discussed in a class like this. Summarized below...

	in the target population	
	H0:	~ H0:
our statistical decision	variables not related	variables are related
p > .05 -- decide to retain H0:	Correct Retention of H0:	Type II error “Miss”
p < .05 -- decide to reject H0:	Type I error “False Alarm”	Correct Rejection of H0:

However,there is a 3rd kind of statistical decision error that I want you to be familiar with, that is cleverly called a ...

Type III statistical decision errors

- correctly rejecting H0:, but mis-specifying the relationship between the variables in the population
- deciding there is a certain direction or pattern of relationship between the two variables in the population when there really is different direction or pattern of relationship
- a Mis-specification
- how’s this happen?
    - sampling variability (“sampling happens”)
    - nonrepresentative sample (Ext Val)
    - confound (Int Val)
    - poor measures/manipulations of variables (Msr Val)

To summarize…

Type I error -- "false alarm" - finding a significant mean difference between the conditions in the study when there really ***isn't*** a difference between the populations

Type II error -- "miss" - finding no difference between the conditions of the study when there really ***is*** a difference between the populations

Type III error -- "misspecification" - finding a difference between the conditions of the study that ***is different from*** the the difference between the populations

Correctly retained H0: -- finding no difference between the conditions of the study when there really ***is no difference*** between the populations

Correctly rejected H0: -- finding a difference between the conditions of the study ***that is the same as*** the the difference between the populations

What makes all of this troublesome, is that we'll never know the "real" relationship between the variables in the population

- we can't obtain data from the entire target population (that's why we *have* sampling - duh!)
- if we knew the population data, we'd not ever have to make NHSTs, make statistical decisions , etc (double duh!)

The best we can do is...

- replicate our studies
  - using different samplings from the target population
  - using different measures/manipulations of our variables
- identify the most consistent results
- use these consistent results as our best guess of what's going on in the target population

Practice with statistical decision errors evaluated by comparing our finding with "other research" …

We found that those in the Treatment group performed the same as those in the Control group. However, the other 10 studies in the field found the Treatment group performed better,

Type II

We found that those in the Treatment group performed better than those in the Control group. This is the same thing the other 10 studies in the field have found.

Correct Reject

We found that those in the Treatment group performed poorer than those in the Control group. But all of the other 10 studies in the field found the opposite effect.

Type III

We found that those in the Treatment group performed better than those in the Control group. But none of the other 10 studies in the field found any difference.

Type I

We found that those in the Treatment group performed the same as those in the Control group. This is the same thing the other 10 studies in the field have found.

Correct retain

Another practice with statistical decision errors ...

We found that students who did more homework problems tended to have higher exam scores, which is what the other studies have found.

Correct Rejection

We found that students who did more homework problems tended to have lower exam scores. Ours is the only study with this finding.

Can't tell -- what DID the other studies find?

We found that students who did more homework problems tended to have lower exam scores. All other studies found the opposite effect.

Type III

We found that students who did more homework problems and those who did fewer problems tended to have about the same exam scores, which is what the other studies have found.

Correct H0:

We found that students who did more homework problems tended to have lower exam scores. Ours is the only study with this finding, other find no relationship.

Type I

We found that students who did more homework problems and those who did fewer problems tended to have about the same exam scores. Everybody else has found that homework helps.

Type II

Keep in mind that rejecting H0: does **not** guarantee support for the research hypothesis?

Why not ???

- The direction of the mean difference might be opposite that of the RH: ? ? ? ?

•The RH: might be that's there's no difference (RH: = H0:)

Also … replication of findings is important, even when you get what you expect !!

RH: The 4th graders will have higher geography scores than the 3rd graders

Results #1 4th = 62% 3rd = 58% F(1,48) = 4.3, p = .02

Reject H0: -- mean dif in correct direction

Results #2 4th = 62% 3rd = 60% F(1,18) = 2.3, p = .16

Retain H0: -- no support for RH:

Results #3 4th = 62% 3rd = 68% F(1,28) = 5.3, p = .01

Reject H0: -- mean dif in wrong direction

The whole process goes like this…

1. Determine the RH:
   – specific direction/pattern or H0:
2. Test RH:, based on …
   a. Evaluate p-value from significance test
   b. Examine data pattern
3. If results from similar other studies are available, evaluate possibility of a Statistical Decision Error
   – If reject H0: check for Type I or Type III errors
   – If retain H0: check for Type II error

A VERY important distinction!!!

Type III Statistical Decision Error

- When our significant findings have a direction or pattern different from that found in the population
- A difference between “the effect we found” and “the effect we should have found”

“Results contrary to our RH:”

- When our findings have a direction or pattern different from what we had hypothesized
- A difference between “the effect we found” and “the effect we hypothesized”

A result can be BOTH!!!!! (Or neither, or one, or the other !!!)

2-group outcomes & “truth” ...

In the population there are only three possibilities...

… and three possible statistical decisions

	In the Population		
Decisions	G1 < G2	G1 = G2	G1 > G2
G1 < G2	Correctly rejected H0:	Type I error	Type III error
G1 = G2	Type II error	Correctly retained H0:	Type II error
G1 > G2	Type III error	Type I error	Correctly rejected H0:

Please note that this is a different question than whether the results “match” the RH: This is about whether the results from the sample are “correct” – whether the results are “right.” This is about statistical conclusion validity

2-group RH: and outcomes BG & WG...

There are only three possible Research Hypotheses

… and three possible statistical outcomes

	Research Hypotheses		
Outcomes	G1 < G2	G1 = G2	G1 > G2
G1 < G2	☺	☹	? ? 😮 ? ?
G1 = G2	☹	☺	☹
G1 > G2	? ? 😮 ? ?	☹	☺

So, there are only 9 possible combinations of RH: & Outcomes …

… of 3 types "effect as expected" ☺

"unexpected null/effect" ☹

"backward effect" ? ? 😮 ? ?

RH:, statistical conclusions & statistical decision errors ...

☺ Results supported ☹ Results not supported

Statistical Decision	RH: + direction/pattern	 H0:	 - direction/pattern
+ direction/pattern (p < .05)	☺ Correct rejection Type I or Type III	☹ Correct rejection Type I or Type III	☹ Correct rejection Type I or Type III
H0: (p > .05)	☹ Correct retention or Type II	☺ Correct retention or Type II	☹ Correct retention or Type II
- direction/pattern (p < .05	☹ Correct rejection Type I or Type III	☹ Correct rejection Type I or Type III	☺ Correct rejection Type I or Type III

Consider the following three pieces of information…

Our RH: is that there will be a positive correlation between how much a person likes performing practical jokes and the number of close friends a person reports.

We found r (58) = -.30, p = .02.

These results are "contrary to our RH:" -- a significant, relationship in the opposite direction from the RH:

A literature review revealed 12 other studies of these two variables, each of which found a correlation between -.25 and -.32 (all p < .05).

The consistent findings of these other studies suggests that our finding was correct – it was our hypothesis that was wrong!!!

How'd we not know the results of the other 12 studies!!

Try this one …

Our RH: is that there will be a negative correlation between the severity of depression at the beginning of therapy and the amount of improvement a patient shows during the first six weeks of therapy.

We found r (63) = .27, p = .035.

These results are “contrary to our RH:” -- a significant, relationship in the opposite direction from the RH:

A literature review revealed 34 other studies of these two variables, each of which found a correlation between -.33 and -.41 (all p < .05).

The consistent findings of these other studies suggests that our finding was a Type III error – what we found “doesn’t describe the relationship between these variables in the population”. Our RH: was correct, but not our data!!!

Information from p-values “vs.” Effect Sizes

- The p-value (value range 1.0 – 0) tells the probability of making a Type I error if you reject the H0: based on the data from this sample
  - e.g., p = .10 means “if we reject H0: based on these data there is a 10% chance that there really is no relationship between the variables in the population represented by the sample”
  - The usual “acceptable risk” is less than 5% or p < .05
- Effect size estimates (value range 0 – 1.0) tell how much of the variability in the DV is “accounted for” (“predicted from” or “caused by”) the IV
  - e.g., r = .30 means “we estimate that $.30^2$ or 9% of the variability in the DV is accounted for by the IV
  - “large enough to be interesting” effect sizes vary with research topics and design types, but a common guideline is .1 = small, .3 = medium and .5 = large

p-values “vs.” Effect Sizes

For 2-group ANOVA (BG or WG) $\mathbf{r} = \sqrt{F / (F + df_{error})}$

	Effect Size	
Significance Test	“large enough to be interesting”	“too small to be interesting”
p < .05	“Best case” “big enough” & “probably really there”	Be careful about dismissing these – many “small effects” have turned out to be important
p > .05	Which to “believe”? Rem - w/ small N comes lowered confidence in the replicability of r Easier to believe r if it replicates earlier research – then the large p-value is probably small N	“Best case” “too small to care about” & “probably not really there”