Intro to hypothesis testing

With material from Howard Seltman, Blase Ur
BACKGROUND: WHAT IS HYPOTHESIS TESTING?
“Classical” hypothesis testing

• Frame a mathematical model describing relationship between input and output variables
• Specify null and alternative hypotheses within this model
• Choose a statistic that (hopefully) can discriminate between the null and alternative hypotheses (probabilistically)
Running example: Password meters

• Do different password meters help users to create better passwords?
Framing a model and hypotheses

• In general: DV varies with IV
  – Do you expect a direction?
  – Categorical vs. numeric inputs and outputs

• Null hypothesis: IV *does not* influence DV

• Alt. hypothesis: IV *does* influence DV
  – As framed in model
(Running example)

- Password meter that changes colors will produce stronger passwords than all-green meter
  - Has a direction: color $\gg$ green
  - IV: categorical; DV: numeric (guess score)
- H0: No difference in strength between meters
- H1: Multicolor $\gg$ Green

- What does it mean for one set to be “stronger”? 
What does a statistical test do?

• Compares tendencies in the data set
  – Central tendency: mean, median, etc.
  – Start w/ mean b/c simplest to talk about

• We have samples; they have errors
  – Measurement errors
  – Sampling errors
  – Random noise, variation in people, etc.
Running example

• $H_0$: $m_M = m_G$
• $H_1$: $m_M > m_G$
What does a statistical test do?

• Find evidence to **reject** the null
  – Or not!

• Does not find evidence to **support** the null
  – In practice: Evidence things are different, but not finding evidence of difference != evidence that things are the same
What kind of evidence?

• Calculate theoretical sample distribution for the null

• $p$-value = area under curve that is *more extreme* than your observed sample

• Decision rule: $p < \alpha$ (typically $\alpha = 0.05$)
Example: Independent samples t-test

- Assumes DV is normally distributed for each condition
- Assumes common variance between conditions
- Assumes means mu, mu+delta
  - H0: delta = 0, or mu_A = mu_B
T-test continued

• Calculate sample distribution under H0

• T-statistic (for null hyp):
  – \((m_A - m_B) / \sqrt{\text{var}1/n1 + \text{var}2/n2}\)
  – Denominator based on variance, sample size
  – Follows t-distribution based on assumptions, degrees of freedom (N-1)

• P-value = area under curve more extreme than observed T-stat
  – Tailedness (in our running example)
Interpreting p-values

• Model assumptions matter!
• NO ASSUMPTIONS about chance that H0 is true – comes from dist. assuming H0 is true!
• Type I error: We found a difference that isn’t real
• Type II error: We failed to find a real difference
• P-values ONLY BOUND TYPE I
Interpreting p-values

• Small p-value is evidence that H1 is likely
  – Otherwise, bad luck of wonky sample
• Large p-value: H0 is true, or type 2 error
  – Can use power analysis to interrogate this a bit
  – “Statistical power”: prob. of rejecting null if you should (more later)

• P-values don’t PROVE anything
Interpreting p-values

• Generally, reject null w/ \( p < 0.05 \)

• A p-value is not magic, just probability, and the threshold is arbitrary

• But, reported TRUE or FALSE: You don’t say something is “more significant” because the p-value is lower
Defining significance

• Statistically significant: It would be unlikely to observe this data if the underlying distributions were the same (e.g. if they had the same mean)

• This doesn’t mean the difference is meaningful!
  – Effect size == strength of the effect
  – Sufficiently large samples can find real but small effects
Running example

- We find that mM >> mG (H0 is rejected)
  - Attacker has to guess 2 million more passwords?
  - Attacker has to guess 2 more passwords?
  - Etc.
P values and multiple testing

• P-values bound Type I error (false positive)
  – You expect this to happen 5% of the time if $\alpha = 0.05$

• What happens if you conduct a lot of statistical tests in one experiment?

• Your cumulative probability of a Type I error can increase dramatically!

• You can correct for this – more details later
p-values and confidence intervals

• 95% CI: over many repeated experiments, the parameter of interest (e.g. mean) will be within the CI 95% of the time
  – Calculated from your sample, with assumptions
• Related to p-value: if CIs do not overlap, $p < 0.05$
• Human judgment about whether this is narrow/wide and how to interpret result
CHOOSING THE RIGHT TEST
Planning

• Choose the test(s) before you collect the data
  – Especially before you look at it!

• Interplay between: what question am I asking? How could I demonstrate a result? What data must be collected for that to work? Etc.

• But, do exploratory analysis / visualization to sanity check your plan and results
What kind of data do you have?

• For input and outcome variables

• Quantitative
  – Discrete (Number of caffeine pills taken by each pony)
  – Continuous (Weight of each pony)

• Categorical
  – Binary (Is it or isn’t it a pony?)
  – Nominal: No order (Color of the pony)
  – Ordinal: Ordered (Is the pony super cool, cool, partly cool, or uncool)

• How many of each?!
What kind of data do you have?

- Does your dependent data follow a normal distribution? (You can calculate this!)

- Choice of test depends on normality
  - If so, use parametric tests.
  - If not, use non-parametric tests.
What kind of data do you have?

• Are your data independent?
  – Within vs. between subjects; group effects; time series
  – If not, repeated-measures, mixed models, etc.
  – Can you make them independent? E.g., after – before
  – Independence is usually the least robust assumption!

• Other assumptions (and robustness) about distribution, errors, variance, etc.
What is your hypothesis?

- One-tailed vs. two-tailed
  - Which side the p-value area is calculated under
  - Doesn’t apply to all tests
  - Make sure your observed data goes in the predicted direction
  - No cheating – must select this *ahead of time*
Reporting a statistical test

• Make it clear what the IV and DVs were
  – Fails surprisingly often
• Be clear which test was used and why
• Supply:
  – p-value
  – Test statistic (sometimes redundant but people like it)
  – df / sample size
  – Effect size whenever possible
From here, for a while

- Different experimental designs and tests that go with them
- Assumptions and assumption checking