

A decorative vertical bar on the left side of the slide, featuring a gradient from light orange to white, with several thin vertical lines and five orange circles of varying sizes. The largest circle is at the top, and the others are arranged in a descending pattern towards the bottom.

REVIEW NOTES

What the heck did we learn again?

5700 TOPICS

- Cumulative, adaptive nature of science
- Reliability and validity
- Initial Exploration of data
- Samples, Sampling distributions (Theoretical, Empirical)
- Confidence intervals
- Effect size, CIs for
- Difference and Equivalence
- Correlation
- The General Linear Model
- Simple regression
- NHST and other problems
- Thinking robustly



NATURE OF SCIENCE

- Science is one of many means by which to gain knowledge
- Some of the distinguishing features of science are its openness, testable ideas, weighing of evidence, and progression/change in the face of new evidence
- It ranges from mere observation and description to explicit causal modeling of events and prediction



CUMULATIVE, ADAPTIVE NATURE OF SCIENCE

- How do ideas become accepted?
 - Logical positivism: focus on verification
 - Popper: falsification
 - Kuhn: social nature of science
 - Lakatos: protective belt
 - Feyerabend: anything goes
- Generating sound ideas/theories to test is very difficult, as is judging evidence in support or against those ideas
- Once evidence is provided, old ideas tend more to incorporate and adapt (or are taken into the new theory) rather than whither away
- Science is not just falsification or verification, induction or deduction but instead uses whatever tool is viable to progress our knowledge in some area



RELIABILITY

- *Reliability* of a measure can refer to many things but in general it is the precision of measurement, i.e. more reliability = less measurement error
 - In terms of psychometric measures, it can refer to the amount of true score present in a measure
 - Observed score = True score + Error
 - In terms of individuals, reliability is a measure of how well a measure distinguishes one person from the next
- While extremely important for a scientific endeavor, it is for some reason often ignored by some either outright (not reported or checked) or in the sense of using measures reported to have poor reliability
 - This leads to: inaccurate estimates, attenuated effects, inefficient estimates, muddled and conflicting results etc. and in short a big waste of research time and money
- *Validity* refers more to the accuracy of the measure
 - Even if it is reliable is it measuring what it's designed to or perhaps some other construct?
- As we will see later, bias estimates can help with model validity issues, but construct, face etc. should be adequately assessed in the construction process



INITIAL EXPLORATION OF DATA

- Every data set worth collecting should be examined ‘from head to toe’ and ‘leave no stone unturned’¹
- A great deal of this should be of a graphical nature, histogram/density plots, scatterplots etc.
- Numerically it involves examining measures of central tendency, variability, covariability, reliability etc.
- Robust checks should also begin at this point
 - Median vs mean, MAD vs. sd etc.
- Tests of model assumptions typically *do not* occur at this point.
 - The model itself is usually analyzed
- The gist is one should know everything about the variables of interest that will be used in subsequent inferential analysis and realize that surprising discoveries of worth can happen during this phase of analysis



SAMPLES, SAMPLING DISTRIBUTIONS

- Recall that the population will rarely if ever be available to us in full, but it must be explicit in the research design and analysis
 - Depressed college students? Depressed Americans? Depressed women?
- In order to speak about a population of interest we must take an appropriate sample of it
- From the sample we infer something about the population
 - Inferential statistics (vs. Descriptive)
 - Inductive logic



SAMPLING DISTRIBUTIONS

- However statistical analysis makes use of not just the sample statistics, which by themselves only speak about the sample, not the population
- It is the sampling distribution that allows for inference
 - Mean = sample statistic
 - CI or t-statistic regarding the mean = inferential statistic
- In the past, we always assumed a theoretical distribution, e.g. that was normal or approximated normal, in order to perform our analysis



SAMPLING DISTRIBUTIONS

- Modern analysis can easily implement other distributions if more appropriate, or simply use the data itself to construct an *empirical* sampling distribution
- In either case once we have a sampling distribution specified, we now have probabilities that we can attach to certain events (means, regression coefficients etc.)



NULL HYPOTHESIS TESTING

- Null hypothesis testing, despite the call for an outright ban by a few statisticians and strong criticism by methodologists for decades¹, is the dominant statistical paradigm
- However there are actually two approaches
 - Fisher: based on the data and focus on the observed p-value, i.e. $p(D | H_0)$, as a measure of disbelief in the null
 - Neyman-Pearson: based on the design of the study, focus on error rates, power, alternative hypothesis
- Philosophically they are incompatible (or at least their developers thought so), and both camps argued vigorously for decades about which was the correct path



NULL HYPOTHESIS TESTING

- Psychology and many other disciplines unfortunately chose to mash the two approaches in an incoherent fashion
- Result:
 - Majority of applied researchers with misunderstanding of the observed p-value
 - Equating p-value to type I error rate
 - Equating p-value to effect size
 - Insertion of Bayesian interpretation etc.
 - Ignoring type II error rate/power
 - Over-emphasis on statistical significance



AVOIDING NHST PROBLEMS

- Most of the problems can be taken care of by:
- Cautious and correct use/interpretation when implemented
 - Difficult given it's not easy to understand even then
- Avoiding use
- Emphasis on effect size, model fit/comparison, prediction/validation etc. rather than statistical significance



FURTHER ISSUES: ASSUMPTIONS

- Inferential analyses come with a variety of assumptions but that will vary depending on the nature of the analysis.
- Examples
 - Homoscedasticity (homogeneity of variance)
 - Normal distribution of residuals
 - Independence of observations
 - Linearity
- In practice these are routinely violated
- Distant past wisdom was that our typical analyses were robust to violations of assumptions
- This was in part reflective of our inability to do anything about it in a practical/computational fashion but also a lot of wishful thinking



FURTHER ISSUES: ASSUMPTIONS

- Recent past and current wisdom is that classical tests are not robust to violations of assumptions
 - They never were robust to outliers, lack of linearity, independence, reliability, rarely are to type II error, might be to type I in some situations
- Ignoring violations has never been acceptable- one at least used the ‘interpret results with caution’ tagline if assumptions were violated
- However with standard univariate techniques this is no longer a viable option, as methods old and new are easily implemented to deal with these issues
 - Studies that do not mention testing assumptions could conceivably be wrong/inaccurate in every sense, and their statistics and any conclusions arising from them are suspect at best.
- Robust methods will at the very least give us a comparison and at best a much more accurate portrayal of reality



ASSUMPTIONS MET

- If we have met what do these inferential statistics tell us?
- Z, t, F, χ^2 , statistics are simply quantiles of a particular probability distribution
- How they are calculated depends on the nature of the test, and so conceptually they may imply different things
 - A t for a regression coefficient is not calculated the same as a t for a planned contrast in ANOVA
 - A Z for a sample mean is not calculated the same as the Z used in a Sobel test for mediation
 - χ^2 for a 2-way frequency table is not calculated the same as that for goodness of fit in an SEM
- It is simply the case that the statistic we are looking at, e.g. ratio of mean between group variance to mean within group variance in ANOVA, adheres to one of those probability distributions with certain degrees of freedom
- Each statistic has an associated probability of occurrence



PROBABILITY

- So what does the observed p-value tell us?
- It tells the probability of observing that test statistic or more extreme if a certain state of affairs is true (H_0)
 - If it is low enough, we reject the null hypothesis
 - Pseudo-falsification approach
- As an example, in regression we get a sense of mean shared variance of the DV and predictors (MS_{model}) and mean squared error in prediction (MS_{error}), leading to an F statistic with associated probability
- Conceptually it tells us the probability of the observed variance accounted for or more extreme (R^2) vs. a H_0 : $R^2 = 0$
- In terms of model comparison, the null model in this case is an intercept only model in which the only prediction is the DV mean and random error.
 - i.e. no relationship between predictors and outcome



PROBABILITY

- What that probability doesn't tell you
 - That what you found is actually meaningful
 - It doesn't tell you type I error rate (that's established before the study via design and is a different probability)
 - It doesn't tell you the strength of association
 - It doesn't tell you the probability that your hypothesis is correct or some other (null or otherwise) is incorrect
 - It doesn't give you any sense of replication
- All it tells you is the probability of your statistic of interest assuming a particular state of affairs (H_0) is true
- The key question is: Does anyone really assume H_0 is viable in the first place?
- Science demands more than a p-value for the progression of knowledge.



MORE PROBABILITY: ERROR RATES

- As mentioned, the observed p-value previously discussed is the conditional probability of the data given the null hypothesis is true
 - $P(D | H_0)$
- Type I error rate is the probability of rejecting the null when it is actually true
 - $P(\text{reject } H_0 | H_0)$
- They neither look or sound the same and in fact are not
- The *observed* (hint hint) p-value comes from actual observation of behavior or whatever is of interest
- Type I error rate is what is controlled/kept at a specified minimum if you design the study a certain way.
- The observed p-value after conducting the study might be .37, .19, .05, .0000001. The probability of making a type I error, if the study is conducted as specified beforehand and assumptions met, is whatever value was decided upon in the design phase, typically .05.



PROBLEMS WITH PROBABILITIES

- A typical and incorrect method of reporting probabilities is a symptom of the philosophical mishmash noted earlier.
- Focus on observed p-values adheres to the Fisherian perspective, it was Neyman and Pearson that saw the key probability in terms of error rates
- If an alpha significance level of .05 is chosen during design, if we reject H_0 only when the observed p is less than .05 the error rate is maintained
- It makes no difference what the actual p-value is, only whether its associated statistic is within the region of rejection
- Furthermore, it makes no sense to report $p < .05$, $p < .01$, $p < .001$. Type I error rate is *fixed* via design *before* data is collected
 - Though not meeting assumptions may alter it
- In the N-P system it also makes no sense to report both the observed p-value and alpha significance level, only whether the statistic falls into the region of rejection or not e.g. $p < .05$
- It also is incorrect to use phrases such as ‘marginal significance’ or a ‘trend’¹ toward significance and similar.
- If you want to play the game in this fashion (focus on error rates) you either reach the cutoff or not, there is no ‘close enough’.



WHY THE CONFUSION?

- Probabilities are confusing
- NHST is really confusing
- Put them together and it's not surprising problems result
- Add to this, poor journal practices that saw statistical significance as the primary, if not only, measure of meaningfulness of a study, which forced people to make a big deal about anything that either reached the cutoff or got close, and equated really small p-values as 'highly' significant¹, i.e. even more important
- Also people tend see $p < .05$, $p < .001$ etc. as just description when in fact there is much more behind it
 - And if you want to be informative report the observed p-value; people can tell clearly what it's less than



PROBLEMS WITH PROBABILITIES

- Along with the observed p-value/alpha confusion, many ignore type II error rate/power concerns, which often are more pressing issue (assuming an over-reliance of statistical significance)
- As an example, a simple model for regression with 5 predictors would require $N = 122$ to find an $R^2 = .10$ statistically significant if $\alpha = .05$. *AND* if assumptions have been met (N required could easily double with heteroscedasticity, lack of normality etc.).
- Typical psych regularly reports small effects (at least partly due to lack of reliability in measures)
- Typical psych research often does not meet assumptions
- As a result, typical psych research doesn't have adequate sample sizes



OTHER ISSUES: POWER

- Power is a concern of the Neyman-Pearson framework
 - Given an effect size...
 - Given an alpha...
 - Given a sample size N ...
 - What is the probability I will correctly reject a nil null? Or alternatively, what is the probability of making a type II error, i.e. not rejecting the null when it is false?
- Used correctly, we decide what power/type II error rate we want and estimate sample size needed to maintain that and type I error before a study is conducted
- Unfortunately
 - Less than perfectly reliable measures
 - Outliers
 - Unbalanced design
 - Different effect sizes
 - Using different alpha levels after collection
 - Unmet assumptions...
- And a host of other things will all affect your observed power in varying degrees, many of which cannot be accounted for before the study
- Furthermore, if you are not adhering the strict N-P approach of $p = .051$ not reject, $.049$ reject then you have also rendered the a priori power analysis void because you have not kept alpha constant



THINGS TO NOTE

- Despite certain beliefs to the contrary, p-values are not required for scientific reporting
- Statistical significance \neq practical significance
- Unimportant differences in coefficients can have dramatic differences in the observed p-value
- Testing one theory against one of no relationship only to say 'unlikely' is largely uninformative
- APA has requested past practices be done away with and/or changed dramatically (and their recommendations were light compared to what many methodologists suggest)



WHAT CAN BE DONE

- Relegate observed p-values to minor import or simply don't report
 - E.g. Interval estimates provide the same functionality and more
- Shift focus to effect size, model fit in comparison to other viable models, accuracy in prediction, size and meaningfulness of coefficients
- Use methods that will result in more accurate probabilities and intervals when desired

