± 6

Ten Curious Questions

Ten statistical questions that require careful thinking

What is a p-value?

- Define a p-value
- P-values and Tommy John
- 100’s of p-values in a medical study?
- Find articles about “p-hacking”
- Read *Statistics in the Courtroom* by George Cobb
- ASA’s official position on p-value use

What do p-values not tell us?

- The ***correct decision*** to make…

  Whether to ride your bike on busy or scenic route… (from *What is a P-Value, Anyway?)*

  Whether to convict or not…

- The ***cause*** of what we see…

  Does the smoke detector ad campaign cause these results?

  Did Kristen Gilbert cause these deaths?

Can association suggest causation?

Or a better question: What do you think about the following stories?

- "New findings show that even children mildly affected by lead poisoning as infants or toddlers run a higher risk of being suspended from school by fourth grade."

  Greg Toppo, USA TODAY 12:13 p.m. EDT August 14, 2013

- "CT Scans Linked to Increased Cancer Risk"

- Spurious Correlations: http://www.tylervigen.com

What are lurking variables?

- Lurking is NOT an AP topic, but it IS pretty descriptive…

  Typically used in bivariate analysis as a possible explanation for the observed relationship

- Confounding IS on the AP syllabus…

  “confounding” means “mixed together”

What is confounding?

- See 2016 Exam, #3, especially the rubric on Part (c)
- See lurking and confounding articles at noblestatman.com
- See College Board article: Planning and Conducting a Study, especially the article by Roxy Peck on confounding (apcentral.com)
- Confounding IS on the AP syllabus…

  "confounding" means "mixed together"

Is a statistical test always necessary?

- Well…obviously not, or I wouldn't have asked the question…
- See story: “Clinical Trials vs. Big Data” (on noblestatman.com)
- See story about Cockpit Flaw of Averages (on noblestatman.com)
- See story: “Do Clinical Trials Work?” (on noblestatman.com)

How is it possible for overall wages in the US have risen, but wages within every subgroup have fallen?

- high school dropouts, HS grads with no college, people with some college, and people with Bachelor's or higher degrees
- Baseball player A can be better against both left- and right-handed pitchers, but player B can have a higher batting average
- **Simpson's Paradox**

If you test positive for a disease, what is the probability you actually have the disease?

- Possibly lower than you might think...
- It depends on how rare the disease is...
- AND, the error rates of the test (false positives and false negatives)
- Find a good example that is NOT already in a textbook…

Is it normal for data distributions to be normal?

- Find lots of commonly (naturally) occurring data and see what you find…

If a residual plot is curved, then is a linear model appropriate?

- Maybe...
- How big are the residuals?
- What is the purpose of the model? Predicting? Or fitting?
- Always THINK about the context.

What does it mean to not reject the H_o?

- The p-value must have been high.
- You do NOT have strong evidence of anything.
- Suppose a newly found manuscript cannot be shown by statistical analysis to NOT be Shakespeare.
- Therefore, there is no evidence that it's NOT Shakespeare.
- But that is NOT the same as saying that is IS Shakespeare.

What do confidence intervals measure?

- From a medical study: confidence interval for the age of males in study: 61-63 years old.
- What is 95% ***confidence***?
- What is the correct interpretation of a ***single*** interval?
- What conclusion can be drawn about a claim of p = 0.21 when the interval is (0.12, 0.33)?
- What is the conclusion if the claim was p = 0.325?
- What conclusion can be drawn about a claim of p = 0.41 when the interval is (0.12, 0.33)?

Why is the best fit model not always the best model?

- “I tend to think that if one is only attempting to predict, the best FIT is arguably the best MODEL. But if one wishes to ***understand***, the model must consider the known (or suspected) science plus some common sense.” --Chris Olsen (listserve 4/4/14)
- “Ballistic motion should probably be modeled as quadratic, irrespective of the data, unrestrained population growth should probably be exponential (restrained probably logistic), etc.”

(Why is the best fit model not always the best model?)

- “...consider the shrinking times in Olympic races. One might be able to model these with linear functions and get a best FIT, but the context would suggest rather a best MODEL of the k/x ^n + b or e^(-kx) + b variety. (--Chris Olsen...)

Is comparing two 1-sample CI's useful to determine if means are different?

- If the CI's do NOT overlap, then it is reasonable to conclude that the means are different.
- BUT...if the CI's DO OVERLAP, the means could still be statistically different...
- Plus...comparing two CI's is mathematically equivalent to saying: $\sqrt{a^2+b^2} = a+b$
- See Fathom demo... (1int vs. 2int)

Do students have to check the 10% condition?

- Students: yes. Statisticians: no.
- Background: if n > 10% of population (i.e. the population is not large enough, i.e. it is *finite*), statisticians use the Finite Population Correction Factor to "adjust" the calculations.
- So in *practice*, statisticians do not worry if the sample is "too large," they just correct it with the FPCF.
- Since the FPCF is not in the AP curriculum, students should check this condition (although some years' rubrics did not require this condition to be checked).

Does the Chi-Square distribution include zero?

- Justify your answer.
- (The answer is “no.”)

What is a Type III Error?

- Google…Siri…

What are degrees of freedom?

Consider a univariate data set containing nothing but "responses"--i.e., y-values, with no predictor variables x. Perhaps you want to model these data as being normally distributed around some common but unknown mean mu. If you had only a single observation (a datum!), you would have a way to estimate mu--albeit a very crude way!--because the observation itself would be an unbiased estimator of mu. But you would have no way to estimate sigma. That's because you'd have "used up" the information that was in your observation estimating mu, and there would be no information left to estimate variability around mu.

Now consider a bivariate data set containing responses (y-values), and a predictor variable x associated with each one. If you wanted to predict y's from x's, you might reasonably fit a line to your data. But what if you only had two data values? You'd be able to fit a line, but you would know in advance (regardless of what the two data pairs actually are) that the fit would appear perfect. You would have used up the two pieces of information that live in your data estimating the slope and intercept of the line. There is no information left over to estimate variability around the line.

Now consider a model that tries to predict y's from x's using a quadratic model: yhat=Ax^2+Bx+C. If you have three (x,y) pairs you can estimate A, B, and C... but as in both of the last two situations, you would have no way to estimate how much variabiltiy the y's in the population have above/below the model's predictions. Three pieces of information--degrees of freedom--resided in the three data points. You used them all up estimating A, B, and C, and have no information left to estimate variability. (Three points will exactly determine a quadratic.)

(What are degrees of freedom?)

Degrees of freedom can be thought of as pieces of information. An initial data set has as many degrees of freedom as there are observed independent responses, the y's. You "use up" a degree of freedom every time you estimate a parameter in the model E(Y)=f(x; A, B, C, ...), where A, B, and C and so on are model parameters that need to be estimated. Any time you have exactly as many y's as you have parameters to estimate, you are in a situation where you can estimate the parameters, but you'll have no way to estimate the variability in y's above/below their model-predicted values, since the model will by necessity perfectly predict the y's.

What we need is "leftover information". And for that we need more responses y than we have parameters needing to be estimated. Ideally, a whole lot more! But even one more data value will be minimally acceptable.

That's the gist of it. There is information "living" in a set of n observed responses y. If the observations are independent, then we have n degrees of freedom. When we construct mathematical models to predict y's (the information) we are really rearranging the information so that some of it is exclusively used in the model (estimating parameters) and the rest of it is exclusively used in the noise (deviations between observations and predictions). Only the latter is useful for estimating variability. And you need to have some information there, or you can't estimate variability at all. The "degrees of freedom" associated with an inferential technique such as a one-sample t-test or linear regression is actually the number of degrees of freedom that reside in the noise. And that's the number of degrees of freedom in the data minus the degrees of freedom in the model. Or: it's the sample size minus the number of estimated model parameters.

If all of this is as clear as mud, it's okay. This is probably the most difficult concept for many, many students of statistics, from those in Intro Stats 101 to those in graduate school studying advanced multivariate methods. It takes many, many "passes" for most people before it makes any sense at all.

--Floyd Bullard