Some final thoughts/comments

Patrick Breheny

May 5

The canonical problem

- The basic issue we have looked at in this course is: how does X affect Y?
  - How does a drug affect the lung function of cystic fibrosis patients?
  - How does lead exposure affect neurological development?
  - How does vaccination affect polio risk?
- As we have seen, the way to address that question and estimate the effect of X on Y depends on the nature of how X and Y are measured (i.e., categorical vs. continuous)

A grid

We can put the analyses we have covered in a grid:

	Outcome (Y)	
Groups (X)	Continuous	Categorical
1	One-sample (paired) t-test Wilcoxon signed rank test	z-test Binomial test
2	Two-sample t-test: Student's Two-sample t-test: Welch's Wilcoxon rank sum test	χ^2 test Fisher's exact test
3+	ANOVA	χ^2 test Fisher's exact test
Continuous	Linear regression	*Logistic regression*

Grid: Comments

A few comments on the preceding grid:

- The grid gives the name of tests – keep in mind that each test has an associated method for estimating a confidence interval for the effect of X on Y
- I didn't have room to include it on the slide, but time-to-event outcomes are another important category, and of course have their own methods (Kaplan-Meier curves, log-rank tests, Cox regression) that depend on the nature of X
- Obviously, we did not cover all boxes in the grid in equal depth:
  - We covered ANOVA briefly and logistic regression not at all (I'm just including it for the sake of completeness)
  - Conversely, we spent a lot of time on 1- and 2-sample studies
- As a consequence, I recommend focusing your studying efforts for the final on the top two rows of the grid

Final: Points breakdown

- The approximate distribution of points to topics on the final is as follows:
  - 60 points: 1- and 2-sample studies
  - 15 points: New topics since quiz 4 (multiple comparisons, time-to-event data)
  - 25 points: Assorted topics from the rest of the course
- "Assorted topics" will focus on main ideas (population vs. sample, p-values vs. confidence intervals, central limit theorem, probability, etc.) that recurred throughout the course

Reflecting on course goals

- Think statistically – to understand the importance of collecting data and using appropriate statistical methods in order to test hypotheses, estimate unknown quantities, and conduct research
- Analyze data using basic statistical methods
- Recognize the strengths and limitations of those methods
- Better comprehend journal articles containing statistical analyses
- Have the necessary background to enroll in BIOS:5120

Final remarks

Some main ideas from the course as a whole that I want to re-emphasize:

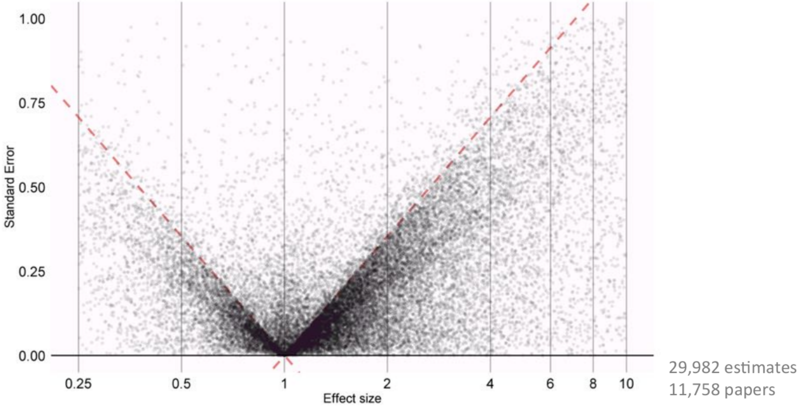
- Think about the study design: Was the study a controlled experiment or an observational study? What population did the sample(s) come from? Is it possible for hidden bias/confounding to explain the results?
- Look at your data: When conducting a study, look at graphs and observe trends, outliers, patterns; don't jump straight to the analysis

Final remarks: p-values

Keep in mind that p-values, although they seem seductively easy to interpret, have some limitations that often lead to over-interpretation:

- They strongly depend on sample size, and tests can be highly significant analyses even when the effect size is small
- A non-significant p-value *does not* mean that there is any evidence that the null hypothesis is true
- They are distorted by multiple comparisons

Science is clearly biased by p-value obsession



Final remarks: Confidence intervals

Confidence intervals do not have these problems, and provide more information than p-values, but you do need to pay attention to what it's a confidence interval *for*

- A ratio? (Null is 1)
- A difference? (Null is 0)
- A proportion? (Null is probably 0.5)