STAT 301 Final Exam Preparation Winter 2011

Logistics:

- Tues Mar 15 or Wed Mar 16 or Fri Mar 18
- 10:10am-1:00pm
  - But exam will be less than twice as long as a midterm
- Open-book, open-notes
  - With exception of a small, conceptual, multiple-choice section at beginning
- Bring calculator
- Some computer (applet/R/Minitab) use
- Partially cumulative
  - Roughly one-half to two-thirds on material since last exam
    1. Wed Feb 23 – Wed Mar 9
    2. HW15-18
    3. Quiz 12
  - Roughly one-third to one-half on (bigger ideas from) earlier material
  - See "Overview of Procedures" at end of this document
- Extra office hours
  - Mon Mar 14, 12-2pm
  - Tues Mar 15, 8:30-9:30am, 2-3pm
  - Wed Mar 16, 8:30-9:30am, 5-6pm
- Materials available online
  - This (and previous) exam preparation document
  - Exam solutions
  - Day-by-day notes
  - HW solutions

Overview (since last exam):
We have analyzed studies that involve a *quantitative* response variable, both with comparing two groups and a single quantitative variable. We have studied graphical and numerical summaries for such data. We have again used inference methods based on simulation/randomization and also approximate methods based on the *t*-distribution.

Outline (since last exam):

- Simulating randomization test for comparing two groups with quantitative response
- Approximate and exact randomization distribution, approximate and exact p-value
- Histogram, five-number summary, boxplot, outlier test
- Shape of distribution, resistance, variability, standard deviation, interquartile range
- Effects of within-group variability, sample size, difference in group means
- Two-sample *t*-test for comparing means, pooled and unpooled
- Two-sample *t*-interval for difference in population means
- Data transformation
- Matched pairs design, randomization test for paired data, paired *t*-test
- Paired *t*-confidence interval, prediction interval

Advice:

- Organize notes for efficient retrieval of information/formulas
- Don’t plan to use text, notes too much
  - Prepare as if exam were closed book/notes
  - Focus on understanding, not memorization
- Be prepared to think/explain/interpret
  - Do not just plug into formulas
  - Be ready to explain process of how you would do calculations
- Take advantage of information provided
  - Perhaps including output
- Relate conclusions to context
- Practice
  - Re-work in-class examples
  - Re-work HW questions
  - Re-work quiz questions
  - Re-work previous exam questions

OVERVIEW OF STATISTICAL PROCEDURES

<u>Selecting the inferential procedure</u>**:** Start with 3 questions

1. Does the research question need a confidence interval or a test of significance?
2. Is the question dealing with a mean (quantitative response) or a proportion (categorical response)?
3. How many (independent) populations do I have: Am I comparing two groups or analyzing one group?

One sample	**Mean (Quantitative – matched pairs)**	**Proportion (Categorical)**
	Descriptive Statistics	
Graphical summary	Histogram, boxplot, dotplot	Bar graph
Numerical summary	$\bar{x}$, s, n	$\hat{p}$, n
	Inferential Statistics	
Null hypothesis	H_0: $\mu = \mu_0$ μ_0 = hypothesized mean	H_0: $\pi = \pi_0$ π_0 = hypothesized population proportion or process probability
Simulation (tactile)	Flip coin for each pair to determine sign of difference	Flip coin (with probability π) for each observational unit
Simulation (technology)	R simulation	Binomial distribution applet
Exact probability model		Binomial
Approximate probability model	One sample *t* procedure	One sample *z* procedure Wald adjustment for 95% CI
Sample size check	Normal population or $n \geq 30$	Wald ci: $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$ tos: $n\pi_0$ and $n(1-\pi_0) \geq 10$ Adjusted Wald: $n \geq 5$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ df=n-1	$z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$
Confidence interval	for μ: $\bar{x} \pm (t^*_{n-1})(s/\sqrt{n})$	for π: $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$ Adjusted: $\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$
R Commander	*Statistics > Means > Paired t-test*	*Statistics > Proportions > Single-sample proportion test (with raw data)*
TOS Applet	One Mean	One Proportion
Simulation applet	Matched Pairs Randomization	Binomial distribution applet
Prediction interval	$\bar{x} \pm (t^*_{n-1})s\sqrt{1+1/n}$ With normal population	

Two independent samples or Randomized experiment	**Means (Quantitative response)**	**Proportions (Categorical response)**
	Descriptive Statistics	
Graphical summary	As above but on same scale	Segmented bar graph for each group
Numerical summary	$\bar{x}_1$, $\bar{x}_2$, s_1, s_2, n_1, n_2	$\hat{p}_1$ - $\hat{p}_2$ or $\hat{p}_1 / \hat{p}_2$ (rel risk), $\hat{\tau}$ (odds ratio)
	Inferential Statistics	
Null hypothesis	H_0: μ_1-μ_2 = 0	H_0: π_1- π_2 = 0
Simulation (tactile)	Index cards with response values	Playing cards with response S, F
Simulation (technology)	Randomization test applet	Two-way table simulation applet
Exact probability model	Randomization test	Fisher's Exact Test (hypergeometric)
Approximate probability model	Two sample *t* procedure	Two sample *z* procedure Wilson adjustment for 95% CI
Sample size check	normal populations or $n_1, n_2 \geq 20$	At least 5 successes and failures in each sample
Test Statistic	$t = \dfrac{\bar{x}_1 - \bar{x}_2 - hypoth\,diff}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ (unpooled) approx df = min(n_1-1, n_2-1)	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ $\hat{p}$ = (total # of successes)/(n_1+n_2)
Confidence Interval	$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ approx df=min(n_1-1,n_2-1)	$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
R	*Statistics > Means > Independent samples t-test*	*prop.test (two-way table)* *fisher.test (two-way table, nrow=2)*
TOS Applet	Two means	Two Proportions
Simulation applet	Randomization Test	Two-way Table Simulation

Note: For case-control studies with a binary response can construct a confidence interval for population relative risk: $\exp[\ln(\hat{p}_1/\hat{p}_1) \pm z^* \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}]$ and for the population odds ratio: $\exp[\ln(\hat{\tau}) \pm z^* \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]$.
Can assess whether 1 is a plausible value for π_1/π_2 or for τ.

Note: Can also use simulation for other statistics, e.g., medians