36-309/749 Lecture FR: Final Review 12/8/2015

1. The big ideas: 20 things you should know

a. When answering a scientific question, experimenters must choose what to manipulate and what baseline characteristics to measures. These are the independent variables (explanatory variables). The must also choose what outcome (dependent or response variable) to measure. Both IVs and DVs must be classified as either categorical or quantitative.

b. The Greek goddess Tyche (Roman: Fortuna) and the state of Chaos, through randomness, rule the real world. (C.f. suíjī in Mandarin.) At each combination of settings of the IVs we postulated a ***true (population) distribution of the DV*** rather than a single value of the DV. This is because there is ***random variation*** in the DV due to measurement error, environmental variation, treatment application variation, and subject-to-subject differences (METS). We are only able to detect "signals" that stand above this "noise". Most statistics tests work by comparing signal to noise, e.g., t=statistic/SE(statistic), F= $MS_{between}/MS_{within}$.

c. Therefore statistical analysis is inference about whether or not a ***change*** in the level or value of an explanatory variable changes the population distribution of the outcome.

d. For categorical DVs, we need to look at the full distribution of the outcome (there is no mean) and how it changes with the explanatory variable(s). Chi-square tests are appropriate for a categorical DV and a single categorical IV, with a ***null hypothesis equivalent to no change in DV distribution as the level of the IV changes***.

e. For categorical DVs with only two levels, we use logistic regression. The structural model is a linear relationship between the natural log of the odds of success and a linear combination of IVs and β parameters. The error model is tied to that through the binomial distribution (fair or unfair coin flip). Although the modeled outcome is ***additive*** on the log-odds scale, effects can be explained as ***multiplicative*** changes (times exp(b)) on the odds scale, and predicted probabilities can be calculated for particular combinations of the explanatory variables.

f. For quantitative outcomes, the central limit theorem suggests that it is common to find a Normal distribution of the outcome ***at each combination of levels of the explanatory variables*** i.e. the errors follow a Gaussian distribution. If we also assume equal variance, then "change in distribution" can be simplified to "change in mean", so the focus of analysis is the means (structural) model.

g. For any outcome type, the error model must either assume independent error or model correlated error (usually within each subject only).

h. Using all of the model assumptions ***and*** the null hypothesis, we can figure out the null sampling distribution of some statistic. Using this, we arrange to falsely reject the null hypothesis (make a Type 1 error) only alpha (usually 0.05) of the time ***when studying "ineffective treatments"***. The p-value is the probability of seeing the observed statistic or any even less supportive of H_0 ***if the null hypothesis and assumptions are true.***

i. The Type 1 error rate is guaranteed only when the data match the model behind the statistical analysis. Analysis choice begins by appropriately matching the types of the variables to the model, and by ***thinking about*** the fixed x-assumption for quantitative explanatory variables and about possible correlation of errors (especially within subjects). Model choice is further guided by exploratory data analysis (EDA). After “fitting a model” with a software package, it is critical to do additional assumption checking. For Normal DV’s we can look at residual vs. fit plots and quantile normal plots of residuals. If non-robust assumptions are violated the computer’s results are not reliable.

j. In any experiment we have between a 0 and 95% chance of falsely retaining the null hypothesis (making a Type 2 error, also called a “false negative”) ***when studying effective treatments***, depending on the power of the experiment for a ***specific*** alternate hypothesis, e.g., $|\mu_1-\mu_2|=2$ instead of just $\mu_1\neq\mu_2$.

k. ***Power may be increased*** by increasing sample size, decreasing “METS” variation, increasing the effect size for differences between treatments, using blocking and control variables (effectively making variation matter only at ***each*** level of those variables), and/or by using within-subject designs (effectively eliminating subject-to-subject variation).

l. Conclusions ***generalize*** only to subjects and conditions represented by subjects and conditions in the experiment. This is also called external validity. (Randomization of subject ***selection*** from a defined population is the seldom-achieved ideal, e.g., Department of the Census assisted surveys on drug use, economics, or health).

m. There is danger in substituting ***concepts*** for ***actual*** treatments and outcomes when describing your conclusions due to concern about construct validity.

n. The key source of internal validity, which makes experiments superior to observational studies is ***random assignment*** of subjects to different levels of one or more IVs (or vice versa). With good internal validity, the observed changes in the distribution of the DV can be said to be caused by the changes in the IVs. For non-randomly assigned explanatory variables, only association can be claimed, which may be due to any variables that differ among the treatment groups (confounders) rather being due to the intended explanatory variables. Experiments can be ruined (lose internal validity) by lack of blinding or lack of treatment randomization.

o. When there are two or more explanatory variables, interaction (***between*** the IVs in their effects ***on*** the outcome) can occur in the means model. In the absence of interaction (an additive model) between explanatory variables A and B, the effect of varying the level of either A or B ***on the outcome*** (say, Y) can be described without mentioning the other. In the presence of interaction, the effect of varying the level of A on outcome Y ***depends*** on the particular level of B and vice-versa, so complicated explanation and often complex contrast testing are needed. Interaction is ***not*** about the effects of one IV on another IV.

p. **ANOVA** methods test the overall null hypothesis of no (relative) effects of changes in level of a categorical explanatory variable, i.e., a factor, on a quantitative DV. Variances are used in the calculations of the F statistic ($MS_{between}/MS_{within}$), but the null hypotheses are about equality of population means or about "no interaction", i.e., a parallel (additive) pattern of population means. ANOVA p-values alone are ***not scientifically sufficient*** if H_0 is rejected. For k=2 levels, report which level of the IV corresponds to the higher mean of the DV and ideally a 95% CI for the magnitude of the difference. Contrast tests are needed for follow-up if there are more than two levels of the factor. For every overall null hypothesis for a factor with $k \geq 3$ levels, k-1 planned contrasts should be specified in advance. In addition, post-hoc contrasts (multiple comparisons, data snooping) may be conducted, but you must take a ***penalty*** (e.g., Tukey, Bonferroni, etc.) to prevent excess Type 1 error.

q. **Regression** methods allow modeling of a linear effect of a quantitative explanatory variable (after transformation if needed) including appropriate interpolation. They also allow categorical explanatory variables if coded as indicator variables. Null hypotheses are that various β's in the means model equal zero. You should write out and/or plot model and/or prediction equations for each combination of levels of the categorical variables (and possibly, say, quartiles of quantitative variables). The term **ANCOVA** applies when the main interest is in effects of a categorical IV on the DV correcting for the quantitative IVs.

r. ***Correlated errors*** are likely for within-subjects factors which have multiple measurements per subject. This is in contrast to between-subjects factors for which each subject experiences only one level.

s. ***Repeated measures ANOVA*** models the correlation of within-subjects data as either equally correlated ("spherical", using the univariate method) or unstructured (multivariate method).

t. ***Mixed (hierarchical) models*** model correlation in the form of random effects which are per-subject "personal" intercepts and/or slopes (varying around an average intercept and/or slope) with or without additional correlation structure modeling. Unequal spacing and missing data are handed appropriately in contrast to repeated measures ANOVA.

2. Hints for taking the exam

a. Read everything carefully; answer what is asked.

b. Don’t lose sight of the scientific question(s). Whenever possible refer to meaningful outcome, factor, and level names rather than “Y”, “x”, or level codes.

c. Show your work to get partial credit.

d. When it is appropriate, refer to the specific area of the output that supports your conclusions. E.g. “there is strong evidence (F=15.6, df=2,34, p<0.0005) that differences in time of contact cause changes in rat biting behavior.”

e. State conclusions so that they will answer the questions that someone interested in the topic would naturally have:

   i. Include the direction of an effect when discussing contrasts or main effects with two levels, and for regression coefficients.

   ii. Include magnitude of effects (either point estimates or confidence intervals) where appropriate.

f. Use words like “association” for observational studies, and reserve “causation” for experiments.

g. Avoid the word “prove”; use “supports the hypothesis that” or “provides evidence for” or even “suggests that”. Use “non-significant” rather than “insignificant” to describe high p-values. Reserve “insignificant” to describe ***substantively*** small changes.

h. Round final answers to 3 significant figures. ***Do not write p=0.000***. Show that you understand that such an output by a computer package is a programming weakness: write “p<0.0005”.

3. Choosing Tests

Identify the experimental units. Identify the outcome (response, dependent variable). Identify the explanatory variables (independent variables). Classify variables as quantitative vs. categorical. Classify factors as between-subjects or within-subjects.

For multiple quantitative outcome measurements on the same subject: use paired t-test or the repeated measures version of ANOVA or mixed models, otherwise use this table:

	Quantitative Outcome	**Categorical Outcome**
Categorical Explanatory	**ANOVA**	**Chi Square Test of Independence**
Quantitative Explanatory	**Regression**	**Logistic Regression**
Both	**Regression (ANCOVA)**	**Logistic Regression**

Follow-up rejected composite (overall) hypotheses (e.g. H_0: $\mu_1=\mu_2=\mu_3$) with planned comparisons (contrasts) were possible. Supplement with post-hoc tests (multiple comparisons, e.g. Tukey) where appropriate.

4. Alyssa Agitator

Alyssa Agitator studies agitation by testing how long blindfolded subjects can hold their hands in a bucket of mayonnaise while listening to a recording describing one of three gross substances (pus, vomit, or feces).

Source	**SS**	**df**	**MS**	**F**	**Sig.**
Substance	896	2	448	14	0.001
Male	192	1	192	6	0.032
Substance*male	48	1	48	1.5	0.246
Error	480	11	32		
Total	1616	15			

What are the variables and their types and roles? Is there any reason to worry about non-independent error?

Consider EDA, type of analysis, usefulness of df values and MS values, and how p-values are derived from the F values. What is the useful null hypothesis for this table? What is your decision concerning that null hypothesis? What would you do next based on your decision? What patterns of relationship between the IVs and DV are possible? How do we test, e.g., H_0: $\mu_P=\mu_F$?

What assumptions need to be met for the p-value to be "correct", i.e., in the long run falsely reject H_0 5% of the time if the treatments really are useless? How can you test those assumptions? Which assumptions are least robust? What can you do if the assumptions are strongly violated?

Comment on internal validity, external validity, and power. Comment on the conclusion.

5. Brian Pickle

Brian studies the effects of preparation methods on crunchiness of pickles. His outcome variable is a quantitative measure of crunch. One IV is brine concentration and the other is cucumber type (A, B, or C).

Source	**B**	**SE**	**t**	**Sig.**	**95% CI**
(Constant)	22.0	2.51	8.75	0.000	[17.0, 27.0]
Brine (gm/mL)	1.2	0.42	2.86	0.005	[-2.03, -0.37]
A	2.0	0.45	4.44	0.000	[1.11, 2.89]
C	-1.0	0.45	2.22	0.015	[-1.89, -0.11]
Brine*A	-0.5	0.16	-3.12	0.002	[-0.82, -0.18]
Brine*C	-0.5	0.16	-3.12	0.002	[-0.82, -0.18]

What type of analysis was performed? What EDA should have been done? How would we use the F-change statistic here?

What is the meaning of the intercept? Is the p-value scientifically meaningful?

Which groups of subjects have their own model equations? What are the model equations?

Which specific scientific claims can we make?

What are the assumptions and how do we check them?

6. The Great Santilli

Caroline is a basketball trainer who teaches players how to not get psyched out during free throws. She randomizes subjects to training method “A” or “B” (coded as in indicator variable for method “A”). The free throw average for the past year is used as an additional IV (covariate). LOFTA is the log odds of the free throw average. The outcome is a successful free throw.

Dependent Variable Encoding:

Original Value	Internal Value
Free throw in	1
Free throw miss	0

Categorical Variable Codings:

		Freq	Parameter (1)
Method A	A	30	1.000
	B	20	0.000

Source	B	SE	Wald	Sig.	Exp(B)	95% CI
(Constant)	-0.04	0.04	1.00	0.162	0.96	[0.89, 1.04]
Method A	0.50	0.22	5.16	0.025	1.65	[1.06, 2.56]
LOFTA	0.96	0.05	368.64	0.000	2.61	[2.26, 2.89]

What is the analysis and what EDA could have been performed? What are the model assumptions? Why do we need to look at the first two tables carefully in SPSS?

What is the scale of the DV in human terms? What is the scale of the DV for the model?

What is the meaning of the intercept? What is the null hypothesis for the intercept?

What is the interpretation of the Method A line?

What is the interpretation of the LOFTA line?

What additional model should be run? What test can be used to partially check the assumptions? What concerns (and fixes) do you have concerning independent errors?