Math 484 Exam 3 is on Wednesday, Nov. 30 and covers ch. 2, 3, 4, 5, homeworks 1-10 and quizzes 1-9. You are allowed 10 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

More emphasis is on quizzes 6-9 and HW 5-10. Everything from Exam 1 is fair game. From Exam 2 handout, points 15), 16), 17), 18), 19) (SLR) will not be on Exam 3.

Types of problems likely to appear on Exam 3:

1) – 13) on Exam 1 review. 14), 20), 22), 23), 27) – 34) on Exam 2 review.

The **fixed effects one way Anova** model has one qualitative explanatory variable called a **factor** and a quantitative response variable Y_{ij}. The factor variable has p levels, $E(Y_{ij}) = \mu_i$ and $V(Y_{ij}) = \sigma^2$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. **Experimental units** are randomly assigned to the treatment levels.

42) Let $n = n_1 + \cdots + n_p$. In an **experiment**, the investigators use randomization to randomly assign n units to treatments. Draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units to treatment 1, the next n_2 units to treatment 2, ..., and the final n_p units to treatment p. Use $n_i \equiv m = n/p$ if possible. Randomization washes out the effect of lurking variables. See HW8 D, Q9?

43) The 4 step fixed effects one way Anova F test has steps
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{p-1,n-p} > Fo$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

See HW9 Bb, Cb, Dh, Q9.

44) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data. See HW9 Ba, Q9.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

Shown is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Boxplots and dot plots for each level are useful for this test. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. If all of the

boxplots or dot plots are about the same, then probably the Anova F test will fail to reject Ho. If Ho is true, then $Y_{ij} = \mu + e_{ij}$ where the e_{ij} are iid with 0 mean and constant variance σ^2. Then $\hat{\mu} = \overline{Y}_{00}$ and the factor doesn't help predict Y_{ij}.

Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(y)$. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2. The one way fixed effects normal ANOVA model is the special case where $Y|(W = a_i) \sim N(\mu_i, \sigma^2)$.

The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way Anova model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each $n_i \geq 10$. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

The *residual plot* is a plot of $\hat{Y}$ versus residual $r = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $r = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $r_{i,1}, ..., r_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

Rule of thumb: If $\max(S_1, ..., S_p) \leq 2 \min(S_1, ..., S_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable.

45) Be able to judge whether the model is good or whether the constant variance assumption or the above rule of thumb is violated by looking at the response and residual plots. See HW9 Cc (nonconstant variance), and the good models Dfg, Eef.

In an **experiment**, the investigators assign treatments to experimental units. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples from p populations (the levels). The effects of lurking variables are present in observational studies.

The **cell means model** for the fixed effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$

and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $r_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares SSTO $= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares SSTR $= \sum_{i=1}^{p} n_i (\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares SSE $= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. The Anova table is the same as that for multiple linear regression, except that SSTR replaces the regression sum of squares and that SSTO, SSTR and SSE have $n-1$, $p-1$ and $n-p$ degrees of freedom.

Know that for the **random effects one way Anova**, the levels of the factor are a random sample of levels from some population of levels Λ_F. Assume the μ_i are iid with mean μ and variance σ_μ^2. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The sample size $n = n_1 + \cdots + n_p$ and often $n_i \equiv m$ so $n = pm$. The μ_i and e_{ij} are independent. The e_{ij} have mean 0 and variance σ^2. The $Y_{ij}|\mu_i \sim f(y - \mu_i)$, a location family with variance σ^2 while $e_{ij} \sim f(y)$. In the test below, if $H_0 : \sigma_\mu^2 = 0$ is true, then the Y_{ij} are iid with pdf $f(y - \mu)$, so the F statistic $\approx F_{p-1,n-p}$ if the n_i are large.

46) **Know** that the 4 step random effects one way Anova test is
i) H_0 $\sigma_\mu^2 = 0$ H_A $\sigma_\mu^2 > 0$
ii) $F_0 = MSTR/MSE$ is usually obtained from output.
iii) The pvalue $= P(F_{p-1,n-p} > F_0)$ is usually obtained from output.
iv) If pvalue $< \delta$ reject Ho, conclude that $\sigma_\mu^2 > 0$ and that the mean response depends on the level of the factor. Otherwise, fail to reject Ho, conclude that $\sigma_\mu^2 = 0$ and that the mean response does not depend on the level of the factor.

See HW10 Fb).

47) Know how to tell whether the experiment is a fixed or random effects one way Anova. (Were the levels fixed or a random sample from a population of levels?) See HW10 Fb.

Response transformations for DOE (design of experiments) is very similar to that of MLR, except the ladder of powers has 5 members: Assume that **all** of the values of the "response" Z are **positive**. A **power transformation** has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

48) A graphical method for response transformations computes the fitted values $\hat{W}$ from the DOE model using $W = t_\lambda(Z)$ as the "response" for each of the five values of $\lambda \in \Lambda_L$. Let $\hat{T} = \hat{W} =$ TZHAT and plot TZHAT vs $t_\lambda(Z)$ for $\lambda \in \{-1, -1/2, 0, 1/2, 1\}$. These plots are called **transformation plots**. The residual or error degrees of freedom used to compute the MSE should not be too small. Choose the transformation $Y = t_{\lambda^*}(Z)$ that has the best plot. Consider the one way Anova model with $n_i > 4$ for $i = 1, ..., p$. i) The dot plots should spread about the identity line with similar shape and spread. ii) Dot plots that are approximately symmetric are better than skewed dot plots. iii) Spread that increases or decreases with TZHAT (the shape of the plotted points is similar to a right or left opening megaphone) is bad.

See HW10 B)

The transformation plot for the selected transformation is also the response plot for that model (eg for the model that uses $Y = \log(Z)$ as the response). Make all of the usual checks on the DOE model (residual and response plots) after selecting the response transformation.

49) For DOE, the log and ladder rules for the response are nearly the same as the MLR rules: so the **log rule** says try $Y = \log(Z)$ if $\max(Z)/\min(Z) > 10$ where $Z > 0$ and the subscripts have been suppressed (so $Z \equiv Z_{ij}$ for the one way Anova model).

50) For the one way Anova, the fitted values $\hat{Y}_{ij} = \overline{Y}_{i0}$ and the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. **Graphical Anova** for the one way Anova model makes a dot plot of scaled treatment deviations (effects) above a dot plot of the residuals. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Decide that an effect is significant if its scaled deviation lies outside the range of the residuals. From Graphical Anova, be able to tell which effects are significant and which means are similar. See HW 10 D).

A contrast $C = \sum_{i=1}^{p} k_i \mu_i$ where $\sum_{i=1}^{p} k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0}$.

51) Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^{p} k_i \mu_i = 0$ where $\sum_{i=1}^{p} k_i = 0$ and the k_i may satisfy other constraints $\}$. Let δ_S denote the probability of a type I error for a single test from the family. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$ P(of making at least one type I error among the family of contrasts) where a type I error is a false rejection.

Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

52) **Know** how to interpret output for multiple comparisons procedures. Underlined means or blocks of letters besides groups of means indicates that the group of means are not significantly different. See HW10 E).