Name: ______________________

ID: ________________________

Stat 421, Test 2, Fall, Nov. 16, 2016; Marzban

10 + 14

ONLY a half-size "cheat sheet" is allowed

Multiple choice: Circle all the correct answers; there is wrong-answer penalty

For rest, SHOW answer & work; NO CREDIT for correct answer without explanation

Points

lect 7

1 **1.** In testing the equality of three population means, what is the single most important advantage of performing the ANOVA F-test as compared to performing 3 pairwise t-tests?

[✓ a) The F-test makes less Type I error.] c) The F-test makes fewer assumptions.

b) The F-test is more convenient to code. d) There is no advantage.

lect 10

1 **2.** You make a random sample of 100 observations of y from one population, and another random sample of 50 observations of y from another population. Circle all of the appropriate tests for the equality of means.

→ Because n's are diff, we can't do $d_j = x_{ij} - x_{ij}$.

[✓ a) unpaired, pooled t-test] [✓ b) unpaired, Welch t-test] c) paired t-test [✓ d) an F test in an RCBD]

we didn't cover unbalanced; so, I ignored d.

lect 10

1 **3.** For a given data set involving multiple factors, consider fitting two models: Model 1 has more terms than model 2. Then (circle the correct answer)

[✓ a) SSE1 ≤ SSE2] c) SSE1 ≥ SSE2

More terms ⇒ More params ⇒ smaller SSE

b) SSE1 = SSE2 d) There is no specific relation between SSE1 and SSE2.

hw-Summary, Lab, class

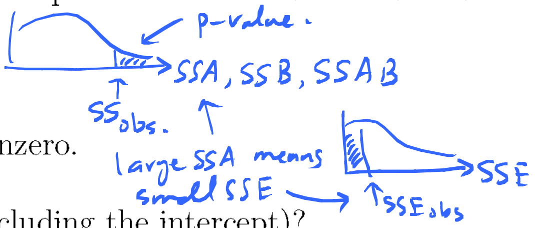
2 **4.** In performing the randomization test in a full model involving two treatment factors ($y = A + B + AB$) which of the following is/are correct? The empirical randomization distribution

[✓ a) of SSA can be used to test the effect of A]

[✓ b) of SSB can be used to test the effect of B]

[✓ c) of SSAB can be used to test the interaction effect AB]

[✓ d) of SSE can be used to test if any of the effects are nonzero.]



1 **5.** In a 2^k design/model, how many effects are there (excluding the intercept)?

a) 2^k b) 2^{k-1} [✓ c) $2^k - 1$] d) 2^{k-2}

k factors, selected 1 at a time + 2 at a time + ... = $2^k - 1$ $\sum_{i=1}^{k} \binom{k}{i} = 2^k$

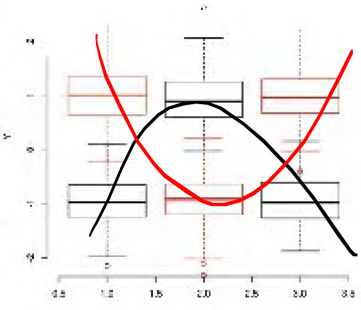
1 **6.** In 2^k design/models, Daniel's conjecture (about using a normal qqlot of the effects to identify significant effects) is useful when the number of effects is

[✓ a) large] b) small c) small or large .

with a small number of effects one doesnot have enough points on The qq plot to see what's normal, and what's not.

lect 13

7. Consider a problem involving a response Y observed at each of 3 levels of factor A. The data are shown in the adjacent figure. The black (red) boxplot denotes the data collected on day 1 (2).



1 **a)** Without doing any tests, does it appear that factor A has an effect on the response? Explain in 1 sentence.

Yes; A has an effect, but The effect depends on day.

or

No (or we don't know), because ∃ too much overlap.

1 **b)** If you were to run a test with the model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ what kind of p-value would you expect to get? Circle: Small [✓ Large]

← The p-value would be large, because if we ignore Day, Then all The treatment means are comparable.

1 **c)** If you were to run a test with the model $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ with β corresponding to the Day factor, what kind of p-value would you expect to get? Circle: [✓ Small] Large

Here The p-value would be small, because The model would see That an effect does exist, within each block.

hw-AV

~~~2~~ 1.5 **8.** Consider data collected from a completely randomized factorial design involving three 4-level factors. A research group uses a full factorial model of the data and finds that factor A is highly significant. A different research group decides to use a model based on an LSD. Is it possible that they will find factor A to be non-significant? [✓ Yes]/No AND provide some explanation.

Because LSD has less data, it is more likely to yield non-significant results. This does not contradict The significant result found in The full model, because a large p-value simply means "we don't know." But This doesn't mean That LSD is useless, because it can find significant results, and with fewer runs.

~~~2~~ 1.5 **9.** Suppose you have collected data according to an LSD involving three p-level factors. Give an argument in terms of the number of observations and the number of parameters in a model (or equivalently in terms of the df of SS terms), to explain what will happen if you try to fit the following model to the data: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijk}$

p^2-1 df in data

This term alone has $(p-1)(p-1)$ df in params

So, There are too many params Than data, and So some of The params will not be estimable. And, of course, SSE will be zero, too.

FYI: This is one reason why one cannot estimate interactions in LSD.

10. Consider the "reduced model" $y_{ijk} = \mu + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$. It can be shown that $\hat{\mu} = \bar{y}_{...}$, $\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$, and $\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{.j.}$

~~~2~~ 1.5 a) Write the expression for the residuals (estimated errors) in terms of conditional means of y_{ijk}.

$$\hat{\epsilon}_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \hat{\mu} - \hat{\beta}_j - \widehat{(\alpha\beta)}_{ij} = y_{ijk} - \cancel{\bar{y}_{...}} - \cancel{\bar{y}_{.j.}} + \cancel{\bar{y}_{...}} - \bar{y}_{ij.} + \cancel{\bar{y}_{.j.}}$$

$$= y_{ijk} - \bar{y}_{ij.}$$

lect 10, p.3 (as an example)

~~~3~~ 2.5 b) Note (do not prove) that the estimate of β_j is the same as that of the full model $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$. Also recall (do not prove) that the estimate of the residuals in the full model is $y_{ijk} - \bar{y}_{ij.}$. Show that $\text{SSAB}_{reduced} = \text{SSA}_{full} + \text{SSAB}_{full}$
Hint: use the anova decomposition for the two models; DO NOT use the formulas for parameter estimates in the formulas for SS because it will be very time consuming.

$SSE_{full} = SSE_{red.}$

$SSB_{red} = SSB_{full}$

$$\begin{array}{ccccccccc}
\boxed{\begin{array}{c} SST \\ \parallel \\ SST \end{array}} & = & SSA_{full} & + & \boxed{\begin{array}{c} SSB_{full} \\ \parallel \\ SSB_{red.} \end{array}} & \begin{array}{c} + \\ \\ + \end{array} & \begin{array}{c} SSAB_{full} \\ \\ SSAB_{red.} \end{array} & \begin{array}{c} + \\ \\ + \end{array} & \boxed{\begin{array}{c} SSE_{full} \\ \parallel \\ SSE_{red} \end{array}}
\end{array}$$

$\therefore SSA_{full} + SSAB_{full} = SSAB_{red}$

df: $abn-1$ $\quad ab-1 \quad abn-ab = ab(n-1)$

~ 2 Lect 9 **11.** Consider the model $y_{ijk} = \mu + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $i = 1-a, j = 1-b, k = 1-n$. Starting from the expression for SSE, take appropriate derivatives, impose appropriate constraints, and find the estimates for the parameters μ, and $(\alpha\beta)_{ij}$.

1 Constraint

$$SSE = \sum_{ijk} \epsilon_{ijk}^2 = \sum_{ijk} \left(y_{ijk} - \mu - (\alpha\beta)_{ij} \right)^2$$

$$\frac{\partial}{\partial\mu}: \quad \sum_{ijk} \left(y_{ijk} - \hat{\mu} - (\hat{\alpha\beta})_{ij} \right) = y_{...} - abn\hat{\mu} - n\underbrace{\boxed{(\hat{\alpha\beta})_{..}}}_{} = 0 \Rightarrow \underline{\hat{\mu} = \bar{y}_{...}}$$

$$\overset{0}{\boxed{(\hat{\alpha\beta})_{..}}}$$

$$\frac{\partial}{\partial(\alpha\beta)_{ij}}: \quad \sum_{k} \left(y_{ijk} - \hat{\mu} - (\hat{\alpha\beta})_{ij} \right) = y_{ij.} - \boxed{\cancel{a}}\cancel{b}n\hat{\mu} - n(\hat{\alpha\beta})_{ij} = 0$$

FYI, because There is only one constraint The df of SSAB is not $(a-1)(b-1)$; it's $(ab-1)$

$$\therefore \underline{(\hat{\alpha\beta})_{ij} = \bar{y}_{ij.} - \bar{y}_{...}}$$

~ ~~3~~ hw_AC **12.** In a homework problem you considered constraints other than the usual ones (i.e., $\alpha. = \beta. = 0$) and found that the estimates of the parameters $(\alpha_i, \beta_j, \cdots)$ depend on the choice of the constraints. However, you also found out that certain combinations of parameter estimates are unaffected by the choice of the constraint. Such combinations are called uniquely estimable functions. Here, consider the additive model $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, for whom the least-squares normal equations are given below. Show that any zero-sum contrast $\Gamma = \sum_i c_i\mu_i$, is a uniquely estimable function, where $\mu_i = \mu + \alpha_i$ are the treatment means. Hint1: If some combination of parameters can be written in terms of the data (y), then it is uniquely estimable. Hint2: It's sufficient to look at only one of the least squares equations (for convenience, I have left out the ˆ symbols):

2.5

$$\bar{y}_{..} - \mu - \bar{\alpha}. - \bar{\beta}. = 0$$
$$\bar{y}_{i.} - \mu - \alpha_i - \bar{\beta}. = 0$$
$$\bar{y}_{.j} - \mu - \bar{\alpha}. - \beta_j = 0$$

Hint 2.

$$\Gamma = \sum_i c_i\mu_i = \sum_i c_i(\mu + \tau_i) = \sum_i c_i(\bar{y}_{i.} - \bar{\beta}.)$$

$$= \underbrace{\sum_i c_i\bar{y}_{i.}}_{\text{all data no params.}} - \bar{\beta}. \underbrace{\sum_i c_i}_{=0}$$

$$\therefore \underline{\Gamma = \text{uniquely estimable}}$$

(ie. it's estimate does not depend on The Choice of constraints)

~ 2 hw_AZ lect16 **13.** Consider a 2^1 design/model with a treatment factor A, with n replications. Show that SSA, as defined by $\sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2$, is equal to $\frac{1}{2n}(\text{contrast}_A)^2$.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, 2, \quad j = 1, \cdots n.$$

$$SSA = \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 = n\sum_i^2 \left(\frac{1}{n}y_{i.} - \frac{1}{2n}y_{..} \right)^2 = \frac{1}{n}\sum_i \left(y_{i.} - \frac{1}{2}y_{..} \right)^2$$

$$= \frac{1}{n}\left[\left(y_{1.} - \frac{1}{2}y_{..} \right)^2 + \left(y_{2.} - \frac{1}{2}y_{..} \right)^2 \right] = \frac{1}{n}\left[\left(y_{1.} - \frac{1}{2}y_{1.} - \frac{1}{2}y_{2.} \right)^2 + \left(y_{2.} - \frac{1}{2}y_{1.} - \frac{1}{2}y_{2.} \right)^2 \right]$$

$$= \frac{1}{n}\left[\frac{1}{4}(y_{1.} - y_{2.})^2 + \frac{1}{4}(y_{2.} - y_{1.})^2 \right] = \frac{1}{2n}(y_{2.} - y_{1.})^2 = \underline{\frac{1}{2n}(\text{Contr}_A)^2}.$$