Final exam

Contents

Instructions

1. Use the "Visual" markdown version of this document. Click "Visual" in the top-left corner of this document before starting.
2. All necessary packages are loaded in the above code block. Check if they all load by clicking the Run arrow
3. Fill in your answers (R code and text) in each block provided.
   - Only questions that include an empty code block need R code
   - All questions need a text answer. Please delete the text "ENTER YOUR RESPONSE HERE" and replace with your answer.
4. When you are done, click the drop-down menu next to "**Preview**" and select "**Knit to PDF**". Check the pdf to make sure all R code and any associated output is provided.
5. Upload both the pdf to Gradescope.
6. This exam is to be done independently. Please do not consult with others in the class. You can use any external resource available to you. But we will grade based on the material I presented in class.
7. If you have clarification questions:
   1. Check Piazza to see if the question has been addressed already
   2. If not, email the professor or TAs. We will post answers to Piazza so that everyone has access to the clarifications. Please do not post directly so as not to give away answers to others.

8. Read every question **completely** and **carefully**! Below the bold title-text the notes and instructions are very important to answering correctly.

Question 1 - part 1

A yield trial was performed at an almond orchard to evaluate different management treatments for the ground between trees. **3 different ground covers (Cover)** were evaluated. The orchard was divided into **5 strips of trees (Strip)**, and then each Strip was divided into **3 sections (Section)**. Each section of each strip was assigned one of the three ground cover treatments. The **total yield of almonds (yield, kg)** harvested from trees in the middle of each section was measured at the end of the season.

```
almonds = read.csv('almonds_1_field.csv')
str(almonds)
```

```
## 'data.frame':    15 obs. of  4 variables:
##  $ Strip  : chr  "S1" "S1" "S1" "S2" ...
##  $ Cover  : chr  "A" "B" "C" "A" ...
##  $ Section: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ yield  : int  133 128 108 128 128 102 132 120 138 115 ...
```

1.1 Create a design table for this experiment [8 points]

Give a justification for each EU and Block that you specify.

NOTE: Right click on a row to insert new rows if needed. You can also make the table in Excel and paste your table directly here.

Structure	Variable	Type	# levels	Block	EU
Treatment	Cover	Categorical	3	Strip	Section
Design	Strip	Categorical	5		
	Cover:Strip	Categorical	15		
	Section	Categorical	15		
Response	yield	Numeric	15		

My block is `Strip`, since it has one section for each of my three treatment levels, and they are effectively replicates. The EU is `Section`, since each one only receives one level of treatment (as opposed to the three treatments each `Strip` receives).

1.2 Write an appropriate linear model for the analysis [4 points]

Be sure to fix any variables in the data table.

```
almonds$Section = as.factor(almonds$Section)
almodel = lm(yield ~ Cover + Strip, data = almonds)
```

Needed to switch `Section` to something not-numeric. Other than that, used `lm`, since there isn't actually replication of each individual `Cover:Strip`.

1.3 Can you conclude that any of the ground covers affect yield? Which covers appear best? Worst? [6 points]

Use alpha = 0.05. Show the R output used to answer the question and reference the specific values you use in your text.

```
anova(almodel)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Cover      2 1581.73  790.87  68.572 9.229e-06 ***
## Strip      4  629.73  157.43  13.650  0.001197 **
## Residuals  8   92.27   11.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
almeans = emmeans(almodel, spec = 'Cover')
covereffects = contrast(almeans, method = 'pairwise')
coversumm = summary(covereffects, infer = T, alpha = 0.05)
print(almeans)
```

```
##  Cover emmean   SE df lower.CL upper.CL
##  A        126 1.52  8      122      130
##  B        128 1.52  8      125      132
##  C        105 1.52  8      102      109
##
## Results are averaged over the levels of: Strip
## Confidence level used: 0.95
```

```
print(coversumm)
```

```
##  contrast estimate   SE df lower.CL upper.CL t.ratio p.value
##  A - B        -2.2 2.15  8    -8.34     3.94  -1.024  0.5833
```

```
##  A - C        20.6 2.15  8    14.46    26.74   9.591  <.0001
##  B - C        22.8 2.15  8    16.66    28.94  10.615  <.0001
##
## Results are averaged over the levels of: Strip
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
## P value adjustment: tukey method for comparing a family of 3 estimates
```

Yes, judging by an anova, we can confidently say that at least one of the covers affects yield (p-value 9.23e-06). The best performing covers appear to be either A or B, and the worst is C, though A and B are not really distinguishable from each other.

Question 1 - part 2

Given these promising results, you decide to expand the experiment by adding two additional orchards. You replicate the same design over those two additional orchards and add them to your data table.

```
almonds_3_Field = read.csv('almonds_3_field.csv',stringsAsFactors = TRUE)
str(almonds_3_Field)
```

```
## 'data.frame':    45 obs. of  5 variables:
##  $ Field  : Factor w/ 3 levels "F1","F2","F3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Strip  : Factor w/ 5 levels "S1","S2","S3",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ Section: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Cover  : Factor w/ 3 levels "A","B","C": 1 2 3 1 2 3 1 3 2 1 ...
##  $ yield  : int  133 128 108 128 128 102 132 120 138 115 ...
```

1.4 Create a new design table for this larger experiment [8 points]

Give a justification for each EU and Block that you specify.

Structure	Variable	Type	# levels	Block	EU
Treatment	Cover	Categorical	3	Strip, Field	Section
Design	Strip	Categorical	5		
	Field	Categorical	3		
	Strip:Field	Categorical	15		
	Cover:Strip:Field	Categorical	45		
	Section	Categorical	45		
Response	yield	Numeric	45		

Cover is blocked both in `Strip` and `Field`, both receiving the full set of Treatments. The EU is still `Section`, since it is still the only part of the design that receives a single treatment.

1.5 Write an appropriate linear model for the analysis [4 points]

Be sure to fix any variables in the data table.

```
almonds_3_Field$Section = as.factor(almonds_3_Field$Section)
threemod = lmer(yield ~ Field + Cover + (1|Strip), data = almonds_3_Field)
```

Using `lmer`, since `Strip` should be a random variable.

1.6 Estimate the effects of Cover again, and compare the results of this larger experiment to the first experiment. [6 points]

Have your conclusions changed? You've done 3x the work. Have you gained precision in your estimates relative to the first (smaller experiment)? How has the *interpretation* of your treatment effects changed with the new experiment? Show the effects tables used to answer the question and reference the specific values you use in your text.

```
anova(threemod, ddf = 'K')
```

```
## Type III Analysis of Variance Table with Kenward-Roger's method
##       Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
## Field  508.3  254.16     2    36  1.4642 0.2447106
## Cover 3229.6 1614.82     2    36  9.3030 0.0005535 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
threemeans = emmeans(threemod, spec='Cover')
threeeffects = contrast(threemeans, method = 'pairwise')
threesumm = summary(threeeffects, infer = T, level=1-0.05/3)
threesumm$p.value = pmin(1, threesumm$p.value * 3)
print(threemeans)
```

```
##  Cover emmean   SE   df lower.CL upper.CL
##  A        125 4.13 12.4    116.2      134
##  B        126 4.13 12.4    116.7      135
##  C        107 4.13 12.4     98.5      116
##
## Results are averaged over the levels of: Field
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```
print(threesumm)
```

```
##  contrast estimate   SE df lower.CL upper.CL t.ratio p.value
##  A - B      -0.467 4.81 36   -14.45     13.5  -0.097  1.0000
##  A - C      17.733 4.81 36     3.75     31.7   3.686  0.0063
##  B - C      18.200 4.81 36     4.21     32.2   3.783  0.0048
##
## Results are averaged over the levels of: Field
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.983333333333333
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
## P value adjustment: tukey method for comparing a family of 3 estimates
```

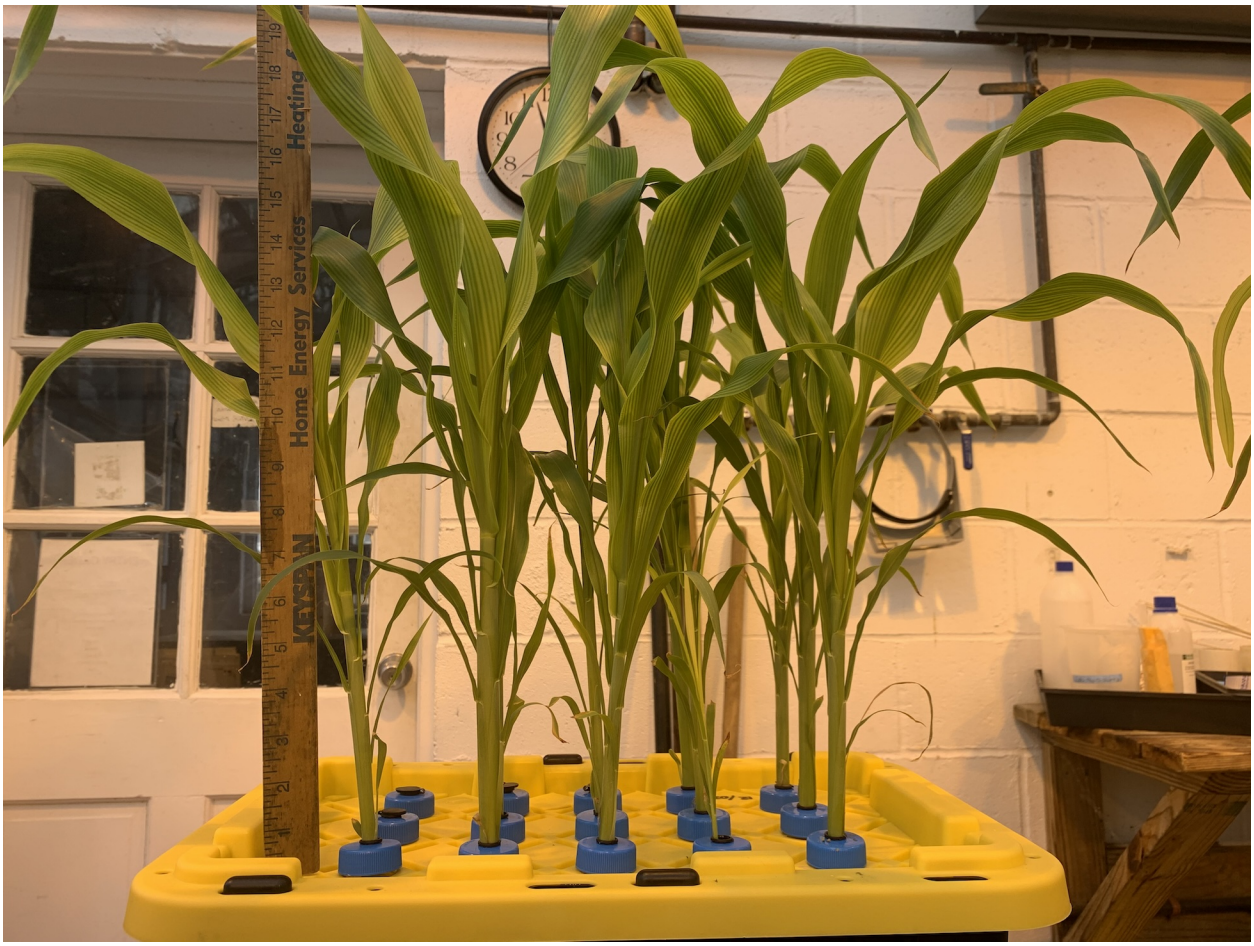
```
no_blocks = lm(yield ~ Cover + Strip, data = almonds_3_Field)
anova(no_blocks)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Cover      2 3229.6 1614.82  9.0811 0.0005977 ***
## Strip      4 1680.2  420.06  2.3622 0.0703641 .
## Residuals 38 6757.2  177.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

My conclusions have not changed, but the context has. The SE of our interaction effects is 4.81 as opposed to 2.15 in the previous experiment, which leads to bigger CIs, and less confidence in the average effect. Truthfully, it's not making much sense to add the two extra fields unless you were explicitly interested to see if they behaved differently (which doesn't seem to be the intention). Theoretically, adding the blocks can indicate if `Field` is interacting with `Cover` and changing treatment effects, but even with pretending as if `Field`s don't exist and we didn't block, the p-value of `Cover` in the anova is nearly the same as with the block structure (0.0005535 vs 0.0005977). Without adding `Field` as a treatment, the replication hasn't improved our precision nor the strength of our interpretation.

However, were we to consider `Field` as a treatment (in addition to a block), we could look make statements about whether it appears that our `Cover` treatment effect is affected by `Field`. Again, when pretending `Field` isn't a block, there isn't a dramatic shift in the F-value or p-value of `Cover` in the anova table. While the blocks ultimately would not have done anything, we would be more confident in generalizing our results to other, similar fields.

Question 2



An experiment was done to test whether the gene PHOSPHORUS-STARVATION TOLERANCE1 (SbPSTOL1) affects Sorghum's ability to take up enough phosphate to maintain root growth under limited phosphate conditions. **Four independent mutants (`MutA`, `MutB`, `MutC`, and `MutD`)** that each independently break the gene SbPSTOL1 in different ways were developed using CRISPR/Cas9 base editing in the r**eference sorghum genotype `Tx430`**. The experiment used hydroponics, meaning that the plants were suspended in a liquid medium where specific concentrations of nutrients could be accurately controlled. Specifically, **10**

tubs (each 20in x 12 in) were filled with Magnavaca modified nutrient solution with either **2.5 uM (low-P) or 250 uM (high-P) phosphate** added as KH2PO4. Plants were germinated on filter paper and then transferred to plugs in the lid of each bin and grown for three weeks. **Each tub housed 15 plants, 7 of the reference genotype `Tx430`, and two each of the four mutants**. The solution was continuously aerated and changed every 5 days. At the end of the experiment each plant was removed, the roots excised, dried, and individually weighed and total dry root mass (g) was recorded for each plant.

```
sorghum = read.csv('sorghum_root.csv')
str(sorghum)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Tub     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Plant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Phos    : chr  "250uM" "250uM" "250uM" "250uM" ...
##  $ Geno    : chr  "MutC" "Tx430" "Tx430" "Tx430" ...
##  $ RootMass: num  59 64 55.5 60.8 65.7 57.2 61.3 62.3 57.6 65.5 ...
```

2.1 Prepare a Design Table for this experiment [8 points]

The focal treatment is `Geno`. Give a justification for each EU and Block that you specify.

NOTE: Right click on a row to insert new rows if needed. You can also make the table in Excel and paste your table directly here.

Structure	Variable	Type	N. levels	Block	EU
Focal	Geno	Categorical	5	Phos, Tub	Plant
Moderator	Phos	Categorical	2		Tub
Combo	Geno:Phos	Categorical	10	Tub(Incomplete)	Plant
Design	Tub	Categorical	10		
	Plant:Phos	Categorical	30		
	Plant	Categorical	150		
Response	RootMass	Numeric	150		

> Okay, so with a focal treatment of `Geno`, our EU is Plant, since that's the part of the design that receives only one level of that treatment. `Geno` is blocked both in `Tub`, since each one contains every level of the treatment, as does each level of `Phos` (which is almost uninteresting. Of course the focal treatment is blocked in the moderator). Our EU for `Phos` is `Tub`, for a similar reason as before. Each `Tub` receives only one level of `Phos`. There is no block for `Phos`. Our combo variable, `Geno:Phos`, has an EU of plant, once again, since that is the only part of the design that receives only one level of the treatment. It is blocked incompletely by `Tub`. While each `Tub` does not receive the full suite of treatments, pairs of hi-P and low-P treated `Tub`s represent the full range.
>
> `Phos:Tub` and `Tub` are aliased, and since `Tub` is an EU, I'm going to keep it over the other.

2.2 Write an appropriate linear model for the analysis [4 points]

Be sure to fix any variables in the data table. Be careful about nested terms.

```
sorghum$Tub = as.factor(sorghum$Tub)
sorghum$Plant = as.factor(sorghum$Plant)
sorgmod = lmer(RootMass ~ (1|Tub) + (1|Plant:Phos) + Geno + Phos + Geno:Phos, data = sorghum)
```

> Using both `Tub` and `Plant:Phos` as my random terms in the model.

2.3 Evaluate whether any of the mutants alter root growth in either Phosphorous treatment relative to Tx430. Use $\alpha = 0.05$. Do not use an ANOVA. [6 points]

Show the effects table used to answer the question and reference the specific values you use in your text.

```
sormeans = emmeans(sorgmod, spec = 'Geno', by = 'Phos')
sorcomp = contrast(sormeans, method = 'trt.vs.ctrl', ref = 'Tx430', name = 'Gen_effect')
sorsumm = summary(sorcomp, infer = T, level = 1-0.05/2, as.df = TRUE)
sorsumm$p.value = pmin(1, sorsumm$p.value * 2)
print(sorsumm)
```

```
## Phos = 2.5uM:
##  Gen_effect   estimate   SE  df lower.CL upper.CL t.ratio p.value
##  MutA - Tx430  -17.377 1.60 132   -21.76   -12.99 -10.882  <.0001
##  MutB - Tx430    2.152 1.60 132    -2.24     6.54   1.345  0.9313
##  MutC - Tx430   -9.568 1.61 132   -13.98    -5.15  -5.949  <.0001
##  MutD - Tx430   -0.948 1.59 131    -5.31     3.41  -0.597  1.0000
##
## Phos = 250uM:
##  Gen_effect   estimate   SE  df lower.CL upper.CL t.ratio p.value
##  MutA - Tx430   -1.058 1.57 129    -5.38     3.26  -0.672  1.0000
##  MutB - Tx430    1.311 1.56 127    -2.98     5.61   0.839  1.0000
##  MutC - Tx430   -7.641 1.55 124   -11.91    -3.37  -4.919  <.0001
##  MutD - Tx430    1.557 1.60 132    -2.84     5.95   0.972  1.0000
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.975
## Conf-level adjustment: dunnettx method for 4 estimates
## P value adjustment: dunnettx method for 4 tests
```

> Yes, we can see that there are at least a few mutants that later root growth relative to Tx430 (when using a Bonferroni multiple comparison correction at 2, since that is how many levels our moderator has), MutA and MutC at alpha = 0.05. The p-value for both at low-P is below 0.0001, and at the hi-P MutC also appears to alter root growth with a p-value == 0.0001.

2.4 Evaluate which mutant effects on root growth changed in low Phosphorous vs High Phosphorous conditions. Use $\alpha = 0.05$. [6 points]

Show the effects table used to answer the question and reference the specific values you use in your text.

```
#now comparison
regroupsor = update(sorcomp, by = 'Gen_effect')
sorinter = contrast(regroupsor, method = 'pairwise')
sorintersumm = summary(sorinter, infer = T, level = 1-0.05/4, as.df = TRUE)
sorintersumm$p.value = pmin(1, sorintersumm$p.value * 4)
print(sorintersumm)
```

```
## Gen_effect = MutA - Tx430:
##  contrast      estimate   SE  df lower.CL upper.CL t.ratio p.value
##  2.5uM - 250uM   -16.32 2.24 131   -22.00   -10.64  -7.280  <.0001
##
## Gen_effect = MutB - Tx430:
##  contrast      estimate   SE  df lower.CL upper.CL t.ratio p.value
##  2.5uM - 250uM     0.84 2.24 131    -4.82     6.51   0.376  1.0000
##
## Gen_effect = MutC - Tx430:
```

```
##  contrast      estimate   SE  df lower.CL upper.CL t.ratio p.value
##  2.5uM - 250uM    -1.93 2.24 131    -7.59     3.74  -0.862  1.0000
##
## Gen_effect = MutD - Tx430:
##  contrast      estimate   SE  df lower.CL upper.CL t.ratio p.value
##  2.5uM - 250uM    -2.50 2.26 132    -8.22     3.21  -1.110  1.0000
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.9875
```

Again using a Bonferroni correction (This time at 4, since we have 4 comparisons across each specific effect), we can see that the only mutant effects on root growth that really changed was MutA, with a p.value of below 0.001.

2.5 Discuss what the researchers learned about the role of SbPSTOL1 on root growth. [5 points]

Address the following:

1. You should have found one mutant with a strong effect in both conditions, and one mutant with an effect that changed considerably between conditions. Which mutant is more interesting for follow-up work? Why?
2. Given that each mutation can alter the genome in other ways in addition to "breaking" a gene, and that each mutation breaks the gene in different ways, is it reasonable to conclude from this work that breaking the gene SpPSTOL1 will affect phosphorous responses in all Sorghum varieties? Why or why not? What would you recommend to the researchers to gain further evidence that this gene is actually involved?

Note: A figure might help you explain your answer to #1.

1. The researchers found both MutA and MutC. However, MutC's effect didn't seem to be affected by the levels of Phos, whereas MutA was. This makes MutA a little more interesting than MutC, since it is a **conditional knockout**. A knockout can answer questions about what happens when a gene breaks completely, but conditional knockouts have more uses beyond that. They can be used to model effects of scarcity, as a method of establishing a control vs treatment relationship without having to create two separate transgenic plant lines, and more. Furthermore, it is likely that effect on the structure of the protein for knockout of MutA is a little more subtle than the full knockout of MutC. Even if the *opposite* is true, this could provide some interesting insights into the protein structure of SbPSTOL1 and its relationship with function. (i.e., which domain in the protein has the most effect on phosphate uptake, etc).

2. No. Off target effects are off target effects, and can change depending on what strain you are doing your experiment in. There is no real replication or manipulation of the, essentially, `Mutation:PlantLine` combination. These effects, therefore, can only really be stated with confidence that they occur in the Tx430 line. Each individual strain of sorghum will have its own unique genetic landscape (it's why they're different strains to begin with); in order to make the statement that knocking out/down SbPSTOL1 will affect phosphorous responses in all Sorghum varieties, they would need to repeat this experiment in different strains, then compare the specific effects of a SbPSTOL1 k/o on each strain.

2.6 Speculate on why the researcher allocated more plants to the Tx430 genotype in each tub than to each of the mutants. Was this a reasonable choice to make? [5 points]

I have two guesses here. The first is that creating transgenic plant lines is not a simple task. It is time-consuming and potentially expensive and error prone. The second is that more control treated plants means your estimation of the baseline interaction between `Phos` and `Geno` is more confident; i.e. you have a better idea of the baseline effect that switching phosphorous concentrations does. This in turn means we can make more confident predictions about the effect of different Mutants with less total plants assigned to those treatments. I think this was reasonable, assuming they were limited in space, time, and funds.