STA 305s14 Regular Assignment Four[1]

This assignment is preparation for the final exam. Your solutions to these homework problems will not be handed in. Use the formula sheet, which is posted on the course home page. As more material is covered, additional problems will be added at the end of the assignment.

1. This question should have been on Assignment Three. Assume a random sampling (not randomization) model for a completely randomized one-factor design. Define a contrast of the expected responses as $c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p = \mathbf{a}'\boldsymbol{\mu}$, and the corresponding contrast of the sample means as $\widehat{c} = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p = \mathbf{a}'\overline{\mathbf{Y}}$. The "weights" $a_1, \ldots, a_p$ add up to zer.

   (a) Using scalar (not matrix) calculations, show that $\widehat{c}$ is an unbiased estimator of c.

   (b) Calculate $Var(\widehat{c})$ using scalar (not matrix) notation. Denote the sample sizes by $n_1, \ldots, n_p$.

   (c) Suppose that $p = 2$, $a_1 = 1$, $a_2 = -1$ and the total sample size $n = n_1 + n_2$ is fixed, possibly by budgetary constraints. Show that $Var(\widehat{c})$ is minimized (so that the estimate is most accurate on average) when the sample sizes are equal. You did this in assignment three.

   (d) Now assume that there are $p = 3$ experimental conditions, and we are interested in the contrast $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$. Again, let the total sample size $n = n_1 + n_2 + n_3$ be fixed. What choice of n_1, n_2 and n_3 minimizes the variance of the estimated contrast? Show your work.

   I did this by letting $x_1 = \frac{n_1}{n}$ and $x_2 = \frac{n_2}{n}$, and then minimizing a function of x_1 and x_2. I took partial derivaties, and then solved two equations in two unknowns and got a satisfying general answer. Is this answer really the location of the unique minimum, and not a maximum or saddle point? Well yes, but to really prove it you need to calculate the eigenvalues of a matrix of partial derivaties, which is the extension of the second derivative test. It's only a 2×2 matrix, but still it's a lot of work so let it go. This is a job for software; it's easy with R. As an alternative, you can sort of convince yourself that you have located the minimum by playing around with the function if you feel like it.

Lecture Unit 9: Factorial ANOVA

2. Steel is made by heating iron and adding some carbon. A steel company conducted an experiment in which knife blades were manufactured using two different amounts of carbon (Low and High), and three different temperatures (Low, Medium and High). Of course even the Low temperature was very hot. A sample of knife blades was manufactured at each combination of carbon and temperature levels, and then the breaking strength of each blade was measured by a specially designed machine. The response variable is breaking strength.

   (a) In a table with one row for each treatment combination, please make columns giving the coefficients of the contrast or contrasts you would use to test for main effects of Temperature.

   (b) In another table with one row for each treatment combination, please make columns giving the coefficients of the contrast or contrasts you would use to test the Temperature by Carbon Level interaction.

   (c) In one last table with one row for each treatment combination, please make columns showing how you would set up dummy variables for both independent variables, using *effect coding* (that's the scheme with the -1s).

   (d) Write $E(Y|\mathbf{X} = \mathbf{x})$ for the regression model, using the names from your table above. Include the interactions!

[1] Copyright information is at the end of the last page.

(e) Using the β values from your answer to the preceding question, state the null hypothesis you'd use to test whether the effect of carbon level on breaking strength depends on the temperature.

3. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression model for the problem:

$$Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,1} d_{i,2} + \epsilon_i,$$

where $d_{i,1}$ and $d_{i,2}$ are dummy variables.

(a) Make a two-by-two table showing the four treatment means in terms of β values. Use *effect coding* (the scheme with $0,\ 1,\ -1$). In terms of the β values, state the null hypothesis you would use to test for

i. Main effect of the first factor
ii. Main effect of the second factor
iii. Interaction

(b) Make a two-by-two table showing the four treatment means in terms of β values. Use *indicator dummy variables* (zeros and ones). In terms of the β values, state the null hypothesis you would use to test for

i. Main effect of the first factor
ii. Main effect of the second factor
iii. Interaction

(c) Which dummy variable scheme do you like more?

4. In a study of math education in elementary school, equal numbers of boys and girls were randomly assigned to one of three training programs designed to improve spatial reasoning. After five school days of training, the students were given a standardized test of spatial reasoning. Score on this test is the dependent variable.

(a) Write $E[Y|\mathbf{X}]$ for a regression model with an intercept. The model includes the possible interaction of Sex by Program, as well as both main effects. You need not say how the dummy variables are defined. You'll do that in the next item.

(b) In the table below, fill in the definitions of the dummy variable)s) for Program, and the dummy variable(s) for Sex. Use *effect coding* (the scheme with $0,\ 1,\ -1$). Show the product terms too. Remember, you *never* include products of the dummy variables for the same factor.

Girls, Program 1	
Girls, Program 2	
Girls, Program 3	
Boys, Program 1	
Boys, Program 2	
Boys, Program 3	

(c) Give the null hypothesis you would test to answer each question below. The answers are in terms of the β parameters in your model. Some of the answers are the same.

Question	*Null Hypothesis*
Averaging the expected values for boys and girls, does program affect test score?	
Does the effect of program on test score depend on the sex of the student?	
Does the effect of sex on test score depend on the which program the student experienced?	
Averaging across expected values for the three programs, is there a sex difference in mean test score?	
Is there a main effect for sex?	
Is there a main effect for program?	
Is there an interaction between sex and program?	
Test both main effects and the interaction, all at the same time; this is one test.	
Are there any differences in average test score among the six treatment combinations? This is one test.	

5. Consider again the math education study of Question 4. Use this notation for the expected test scores.

Girls			Boys		
Program 1	Program 2	Program 3	Program 1	Program 2	Program 3
μ_{11}	μ_{12}	μ_{13}	μ_{21}	μ_{22}	μ_{23}

(a) In terms of the μ_{ij} values, state the null hypotheses you would test to answer the following questions.

i. Averaging the expected values for boys and girls, does program affect test score?
ii. Is there an effect of program type for *either* boys or girls? The negation of this is that there is no effect for boys and no effect for girls. A single test is being requested.
iii. Does the effect of program on test score depend on the sex of the student?
iv. Does the effect of sex on test score depend on the which program the student experienced?
v. Averaging across expected values for the three programs, is there a sex difference in mean test score?
vi. Is there a sex difference in expected test score for any of the three programs? The negation of this is that there is no difference for any program. You are being asked for one test.
vii. Is there a main effect for sex?
viii. Is there a main effect for program?
ix. Is there an interaction between sex and program?
x. Test both main effects and the interaction, all at the same time; this is one test.
xi. Are there any differences in average test score among the six treatment combinations? This is one test.

(b) Now consider contrasts of the form

$$L = c_1\mu_{11} + c_2\mu_{12} + c_3\mu_{13} + c_4\mu_{21} + c_5\mu_{22} + c_6\mu_{23}.$$

For each question below, give the coefficients (c_j quantities) of the contrasts you would test in order to answer the question. In each case, you are testing the null hypothesis that the contrast or set of contrasts equal zero.

i. Averaging the expected values for boys and girls, does program affect test score?
ii. Is there an effect of program type for *either* boys or girls? The negation of this is that there is no effect for boys and no effect for girls. A single test is being requested.
iii. Does the effect of program on test score depend on the sex of the student?
iv. Does the effect of sex on test score depend on the which program the student experienced?
v. Averaging across expected values for the three programs, is there a sex difference in mean test score?
vi. Is there a sex difference in expected test score for any of the three programs? The negation of this is that there is no difference for any program. You are being asked for one test.
vii. Is there a main effect for sex?
viii. Is there a main effect for program?
ix. Is there an interaction between sex and program?
x. Test both main effects and the interaction, all at the same time; this is one test.
xi. Are there any differences in average test score among the six treatment combinations? This is one test.

(c) Suppose you rejected the null hypothesis of no sex by program interaction, and you interpreted this as meaning that the sex differences were not the same for the three programs. Next, you need to follow this up, to determine where the effect came from.

i. How about testing all pairwise differences between differences? Give the null hypotheses you would test, in term of μ_{ij} quantities.
ii. In this situation, people often look at the usual tests for pairwise differences between expected values to see where the interaction came from. Are these null hypotheses implied by the null hypothesis of the test we are following up?

Lecture Unit 10: Analysis of Covariance

6. Suppose that an experiment has just one treatment condition and one control condition. For experimental units exposed to the control condition, the expected response is $\beta_0 + \beta_1 x$, where x is the value of a covariate. Under the assumption of unit-treatment additivity, what is the expected value in the treatment condition? What is the connection to the concept of an interaction?

7. In a study of agricultural productivity, small apple farms are randomly assigned to use one of three Pesticides (Type A, B or C) and one of three Fertilizers (Type 1, 2 or 3). The dependent variable is total crop yield in kilograms, and there are two covariates: number of trees on the farm, and crop yield last year.

(a) In the table below, fill in the definitions of the dummy variables for Pesticide (p_1 and p_2), and the dummy variables for Fertilizer (f_1 and f_2). Use *effect coding* (the scheme with $0,\ 1,\ -1$).

Pesticide	*Fertilizer*	p_1	p_2	f_1	f_2
A	1				
A	2				
A	3				
B	1				
B	2				
B	3				
C	1				
C	2				
C	3				

(b) Write $E[Y|\mathbf{X}]$ for a model that includes a possible Pesticide by Fertilizer interaction as well as their main effects. Denote the covariates by X_1 and X_2. Of course the vector $\mathbf{X}$ includes p_1, p_2 and so on as well as X_1 and X_2. There are no interactions between the two covariates, or between covariates and factors.

(c) Give the null hypothesis you would test to answer each question below. The answers are in terms of the β parameters in your model. Some of the answers are the same. Except for the last one, assume that each question begins with "Controlling for number of trees and crop yield last year . . ." .

Question	*Null Hypothesis*
Averaging across fertilizer types, does type of pesticide affect average crop yield?	
Does the effect of fertilizer type on crop yield depend on the type of pesticide used?	
Does the effect of pesticide type on crop yield depend on the type of fertilizer used?	
Averaging across pesticide types, does fertilizer type affect average crop yield?	
Is there a main effect for pesticide type?	
Is there a main effect for fertilizer type?	
Is there an interaction between fertilizer type and pesticide type?	
Test both main effects and the interaction, all at the same time.	
Test both covariates simultaneously, controlling for the main effects and the interaction.	

8. In the *Eating Study*, pairs of university students came to a Psychology laboratory to eat a meal together. They were either friends or strangers, they ate from either small or large plates, and the food was in either a common bowl or separate bowls. Before the meal, they rated how hungry they were. The variable `Hunger` in the data is the mean rating to the two subjects who are eating together. The total amount of food they served out onto their plates and the total amount of food they actually ate were recorded, in grams. Here is the SAS program, followed by part of the list file.

```
/* eating.sas: Pliner's Yummy data */
options linesize=79 pagesize=500 noovp formdlim='_';
title "Eating Data";
```

```
proc format;
     value ffmt 1 = 'Friends' 2 = 'Strangers';
     value pfmt 1 = 'Large Plate'  2 = 'Small Plate';
     value sfmt 1 = 'Common Bowl' 2 = 'Separate bowls';

data chowtime;
     infile 'Eating.data' firstobs=2;
     input Friend Plate Share Hunger  FoodSrv FoodEat;
     format Friend ffmt.;
     format Plate  pfmt.;
     format Share  sfmt.;

proc glm;
     class Friend Plate Share;
     model FoodSrv FoodEat = hunger Friend|Plate|Share;
     lsmeans Friend|Plate|Share;

----------------------------------------

                                  Eating Data                                1

                               The GLM Procedure

                            Class Level Information

              Class         Levels    Values

              Friend             2    Friends Strangers

              Plate              2    Large Plate Small Plate

              Share              2    Common Bowl Separate bowls


                    Number of Observations Read          57
                    Number of Observations Used          57


-------------------------------------------------------------------------------
                                  Eating Data                                2

                               The GLM Procedure

Dependent Variable: FoodSrv

                                             Sum of
Source                      DF          Squares     Mean Square    F Value    Pr > F

Model                        8      247157.8287      30894.7286       2.97    0.0087
```

Error	48	498824.3092	10392.1731		
Corrected Total	56	745982.1379			

R-Square	Coeff Var	Root MSE	FoodSrv Mean
0.331319	23.36760	101.9420	436.2537

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Hunger	1	64064.75136	64064.75136	6.16	0.0166
Friend	1	58889.57844	58889.57844	5.67	0.0213
Plate	1	12511.69338	12511.69338	1.20	0.2780
Friend*Plate	1	9950.85369	9950.85369	0.96	0.3327
Share	1	39207.81294	39207.81294	3.77	0.0580
Friend*Share	1	232.07441	232.07441	0.02	0.8818
Plate*Share	1	50525.55846	50525.55846	4.86	0.0323
Friend*Plate*Share	1	11775.50606	11775.50606	1.13	0.2924

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Hunger	1	8242.14175	8242.14175	0.79	0.3776
Friend	1	75213.88643	75213.88643	7.24	0.0098
Plate	1	13230.41227	13230.41227	1.27	0.2648
Friend*Plate	1	6940.34667	6940.34667	0.67	0.4178
Share	1	47995.34574	47995.34574	4.62	0.0367
Friend*Share	1	157.24354	157.24354	0.02	0.9026
Plate*Share	1	55315.60357	55315.60357	5.32	0.0254
Friend*Plate*Share	1	11775.50606	11775.50606	1.13	0.2924

The GLM Procedure

Dependent Variable: FoodEat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	215056.3060	26882.0383	2.57	0.0203
Error	48	502109.4096	10460.6127		
Corrected Total	56	717165.7157			

R-Square	Coeff Var	Root MSE	FoodEat Mean
0.299870	24.11728	102.2771	424.0825

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Hunger	1	67061.23143	67061.23143	6.41	0.0147
Friend	1	48734.93228	48734.93228	4.66	0.0359
Plate	1	5124.99555	5124.99555	0.49	0.4873
Friend*Plate	1	11886.26519	11886.26519	1.14	0.2918
Share	1	42825.98220	42825.98220	4.09	0.0486
Friend*Share	1	540.51108	540.51108	0.05	0.8211
Plate*Share	1	34885.13845	34885.13845	3.33	0.0740
Friend*Plate*Share	1	3997.24985	3997.24985	0.38	0.5394

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Hunger	1	16310.92742	16310.92742	1.56	0.2178
Friend	1	58466.31824	58466.31824	5.59	0.0222
Plate	1	4413.24880	4413.24880	0.42	0.5191
Friend*Plate	1	9396.68815	9396.68815	0.90	0.3480
Share	1	50998.56677	50998.56677	4.88	0.0321
Friend*Share	1	336.44991	336.44991	0.03	0.8584
Plate*Share	1	37062.36817	37062.36817	3.54	0.0659
Friend*Plate*Share	1	3997.24985	3997.24985	0.38	0.5394

The GLM Procedure
Least Squares Means

Friend	FoodSrv LSMEAN	FoodEat LSMEAN
Friends	476.288291	459.474409
Strangers	399.704253	391.952903

Plate	FoodSrv LSMEAN	FoodEat LSMEAN

Large Plate	421.637672	416.265683
Small Plate	454.354872	435.161629

Friend	Plate	FoodSrv LSMEAN	FoodEat LSMEAN
Friends	Large Plate	471.172417	463.108264
Friends	Small Plate	481.404164	455.840553
Strangers	Large Plate	372.102926	369.423101
Strangers	Small Plate	427.305579	414.482705

Share	FoodSrv LSMEAN	FoodEat LSMEAN
Common Bowl	467.487383	456.113444
Separate bowls	408.505160	395.313867

Friend	Share	FoodSrv LSMEAN	FoodEat LSMEAN
Friends	Common Bowl	507.470277	492.347545
Friends	Separate bowls	445.106304	426.601272
Strangers	Common Bowl	427.504490	419.879343
Strangers	Separate bowls	371.904016	364.026462

Plate	Share	FoodSrv LSMEAN	FoodEat LSMEAN
Large Plate	Common Bowl	483.349917	473.039931
Large Plate	Separate bowls	359.925426	359.491434
Small Plate	Common Bowl	451.624849	439.186958
Small Plate	Separate bowls	457.084894	431.136300

Friend	Plate	Share	FoodSrv LSMEAN	FoodEat LSMEAN
Friends	Large Plate	Common Bowl	549.411222	530.999536
Friends	Large Plate	Separate bowls	392.933613	395.216993
Friends	Small Plate	Common Bowl	465.529332	453.695554
Friends	Small Plate	Separate bowls	497.278996	457.985551
Strangers	Large Plate	Common Bowl	417.288613	415.080326
Strangers	Large Plate	Separate bowls	326.917240	323.765875
Strangers	Small Plate	Common Bowl	437.720366	424.678361
Strangers	Small Plate	Separate bowls	416.890792	404.287049

Now please answer these questions. Please remember to ignore the tests based on Type I Sums of squares. Type III corresponds to the general linear test.

(a) This is an analysis of covariance. What is the covariate?

(b) I believe it's possible (though not guaranteed) that the covariate was influenced by one of the the factors. Which one any why?

(c) Controlling for reported hunger and averaging across size of plate and whether they were serving from a common bowl or separate bowls, did the amount the subjects ate depend on whether they were eating with a friend? Give the value of the F statistic, the p-value, and whether you reject H_0 at $\alpha = 0.05$. In plain language, what do you conclude?

(d) Controlling for reported hunger and averaging across size of plate and whether they were eating with a friend or a stranger, did the amount the subjects ate depend on whether they were served from a common bowl or separate bowls? Give the value of the F statistic, the p-value, and whether you reject H_0 at $\alpha = 0.05$. In plain language, what do you conclude?

(e) Controlling for reported hunger and averaging across whether they were eating with a friend or a stranger, it looks like the amount of food served onto the plate might depend on the *combination* of the size of plate and whether they were served from the same bowl. Give the value of the F statistic and the p-value. Describe the interaction in plain language.

9. In a study of fuel efficiency, mid-sized SUVs were randomly assigned to receive one of three fuel additives (call them A, B and C), and then their fuel consumption in liters per 100 kilometers was assessed by driving over a standard course. Weight of the vehicle was a covariate.

   For this problem, you will use a regression model with an intercept. You will use *effect coding* (the scheme with $0,\ 1,\ -1$). The model will allow you to test for a possible interaction of fuel additive by vehicle weight. That's right, just multiply the dummy variables by the covariate. It's not so obvious why this is a good idea with effect coding. The point of this question is to see how it works.

   (a) Make a table with one row for each experimental condition, showing how the dummy variables are defined. Make a wider row in which you show $E[Y|\mathbf{X}]$ for each experimental condition.

   (b) Make another table in which you simplify our expressions for $E[Y|\mathbf{X}]$, giving the formulas for the regression lines in slope-intercept form.

   (c) Averaging across the three experimental treatments, what is the average (arithmetic mean) intercept? Give the answer in terms of the β parameters of your model.

   (d) What is the average slope?

   (e) What is the difference between the slope for additive A and the average slope?

   (f) What is the difference between the intercept for additive B and the average intercept?

   (g) What is the difference between the slope for additive C and the average slope?

   (h) What is the difference between the intercept for additive C and the average intercept?

   (i) In terms of the β parameters of your model, what null hypothesis would you test to compare the slopes of additive A and additive C? Simplify.

   (j) In terms of the β parameters of your model, what null hypothesis would you test to check the equal slopes assumption?

   (k) In terms of the β parameters of your model, what null hypothesis would you test to see whether the three regression lines were the same? (This is sometimes called "equal regressions.")

Lecture Unit 12: Randomized block designs

10. Sprinters on High School track teams are to be randomly assigned to different doses of anabolic steroids. Briefly explain why team would be a good blocking variable.

11. Suppose there are m treatments, arranged in a complete randomized block design with k blocks. This means each treatment appears exactly once within each block.

    (a) How many experimental units are required?

    (b) In how many ways can the units within each block be assigned to experimental treatments?

    (c) In how many total ways can the experimental units be assigned to treatments?

    (d) What is the maximum number of values in the permutation distribution of the test statistic?

12. As in the last question, suppose there are m treatments, arranged in a complete randomized block design with k blocks, so that each treatment appears exactly once within each block. Consider a regression model for this, with effect coding for the dummy variables and products for the interaction of Block and Treatment.

    (a) How many dummy variables are there for Treatment?

    (b) How many dummy variables are there for Block?

    (c) How many products of dummy variables are in the regression model?

    (d) How many regression coefficients are there in total?

    (e) What is the sample size n?

    (f) What are the error degrees of freedom for this model?

    (g) Looking at the formula sheet, why does this tell you that you can't test any hypotheses with the general linear F-test?

    So now you see why models for randomized block designs have no interaction between block and treatment. However, you should know that the interaction is testable another way – one due to our old friend Mr. Tukey.

13. A study is designed to compare two contact lens cleaning solutions in a randomized block design. Block is the person. For each person, one eye is randomly assigned to each treatment. The dependent variable is amount of redness in the eye as rated by a nurse after two weeks. For a small study with only five subjects,

    (a) Write a regression model with an intercept. Begin with "$Y_i = \ldots$" and so on.

    (b) Make a table showing how your dummy variables are defined.

    (c) In terms of the β parameters of your model, what null hypothesis would you test to see which contact lens solution worked better?

14. The lecture slide show has an example of a Latin Square design with four treatments.

    (a) Write down a Latin Square design with three treatments.

    (b) For your design with three treatments, write a regression model with an intercept. Begin with "$Y_i = \ldots$" and so on.

    (c) For your design with three treatments, make a table showing how your dummy variables are defined.

    (d) In terms of the β parameters of your model, what null hypothesis would you test to see whether there were any treatment effects?

Lecture Unit 14: Choosing sample size

15. Assume a random sampling model for a completely randomized design with p experimental treatments. Let

$$\begin{aligned}\ell &= a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p \\ \hat{\ell} &= a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p\end{aligned}$$

(a) Calculate the expected value and variance of $\hat{\ell}$.

(b) In terms of a regression model with cell means coding (indicator dummy variables and no intercept) what is $\hat{\ell}$?

(c) The ordinary $(1-\alpha)100\%$ confidence interval for ℓ is $\hat{\ell}$ plus or minus something. What is that something? Use the formula sheet.

(d) Why might it be reasonable to assume that the distribution of $\hat{\ell}$ is approximately normal even if the Y_{ij} data values are not?

(e) Derive a formula for $Pr\{|\hat{\ell} - \ell| < m\}$, where m is some desired margin of error. Express your answer in terms of the function $\Phi(x)$, the cumulative distribution function of a standard normal random variable.

(f) Using $z_{\alpha/2}$ to denote the number satisfying $\Phi(z_{\alpha/2}) = 1 - \alpha/2$, What value of the sample size n is required for $Pr\{|\hat{\ell} - \ell| < m\} = 1 - \alpha$? Show your work.

(g) In a study with just an experimental group and a control group, the experimenter uses intuition to decide $n_2 = 2n_1$, because units in the control group require about half as much time (and therefore expense) to process. What should the minimum total sample size be so that $\mu_1 - \mu_2$ can be estimated to within $\frac{\sigma}{10}$, with probability 0.95? For comparison, the (corrected) lecture slides indicate that with $n_1 = n_2$, the required sample size is $n = 1,538$. Note that the formula sheet now has critical values of the standard normal distribution.

(h) Continuing with the last item, which experiment will cost more to achieve the same precision, the one with equal sample sizes, or the one with unequal sample sizes?

(i) In an experiment with $p = 3$ treatment conditions, suppose we are interested in the contrast $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$.

   i. With equal sample sizes, what total sample size $n = n_1 + n_2 + n_3$ is required so that the contrast can be estimated to within $\frac{\sigma}{4}$ of its true value, with 90% probability?

   ii. In Problem 1d, you found that the variance was least with $n_1 = 2n_2$ and $n_2 = n_3$. How much total sample size would you save on the last problemn if you used these optimal relative sample sizes?

(j) Specialize the formula on the formula sheet for the case of estimating a single expected value based on one random sample.

(k) No matter what he does, the approval rating of our Mayor seems to hover around 40%. We are planning a survey to estimate it again.

   i. Assuming $Y_1, \ldots, Y_n$ are Bernoulli and the approval rating does not change a whole lot, this gives us a pretty good idea of σ^2. What is a good guess of σ^2?

   ii. We want to be able to say the usual "These results are expected to be accurate within three percentage points, 19 times out of 20." If the approval rating goes up, we might have to say something like "accurate to within 3.5 percentage points," but nobody will really mind. What is the required sample size?

16. If a random variable W has moment-generating function

$$M_W(t) = (1-2t)^{-\frac{\nu}{2}} e^{\frac{\lambda t}{1-2t}}$$

then we say that W has non-central chi-square random variable with degrees of freedom $\nu > 0$ and non-centrality parameter $\lambda \geq 0$, and we write $W \sim \chi^2(\nu, \lambda)$.

Let $W_1, \ldots, W_n$ be independent random variables, with $W_i \sim \chi^2(\nu_i, \lambda_i)$, and let $W = \sum_{i=1}^n W_i$. Find the distribution of W; show your work.

17. Recall from lecture that if Z is normal with variance one, then $Z^2 \sim \chi^2(1, (E(Z))^2)$.

    (a) First the univariate version: Let $Y \sim N(\mu, \sigma^2)$ Show $\frac{Y^2}{\sigma^2} \sim \chi^2(1, \frac{\mu^2}{\sigma^2})$.

    (b) Now the multivariate version: Let $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show $\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \sim \chi^2(p, \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$.

18. For the general linear test, you know that whether $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is true or not, $SSE \sim \chi^2(n-p)$ and SSE is independent of $\widehat{\boldsymbol{\beta}}$. Thus, to prove that F^* on the formula sheet has a non-central F distribution, all you need to show is that

$$\frac{1}{\sigma^2}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{t})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{t})$$

has a non-central chi-squared distribution. Show it, and also verify that the formula for the non-centrality parameter on the formula sheet is correct.

19. For the special case of testing the difference between a treatment and control condition,

    (a) Show that the non-centrality parameter may be written $\lambda = nf(1-f)d^2$, where $f = \frac{n_1}{n}$ and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$.

    (b) You know that the greater the non-centrality parameter, the greater the power. Prove that for any total sample size n and any effect size d, the power is greatest when $n_1 = n_2$.

20. In an experiment with three conditions, the investigator plans to test $H_0 : \mu_1 = \mu_2 = \mu_3$, and decides on equal sample sizes. The investigator would like to be able to reject the null hypothesis with high probability when $\mu_1 = 45$, $\mu_2 = 50$, $\mu_3 = 55$ and $\sigma^2 = 100$. The non-centrality parameter λ may be written as the sample size n, multiplied by a quantity that could be called "effect size." Find the effect size for this problem. The answer is a number. This is something you can do by hand, but you can also look at the code on the lecture slides and see how to check your answer with SAS.

    You can invert a 2×2 matrix by hand if you need, to, but to make it easier you might want to use orthogonal contrasts. The ones I used are called *orthogonal polynomials*, and if you look at the coefficients you can see why. Here they are:

-1	0	1
1	-2	1