STAT 420: Homework 07

Fall 2020, D. Unger

Due: Tuesday, October 27 by 11:30 PM CT

Contents

Directions

Students are encouraged to work together on homework. However, sharing, copying or providing any part of a homework solution or code is an infraction of the University's rules on Academic Integrity. Any violation will be punished as severely as possible.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

- Your assignment must be submitted through the submission link on **Compass 2g.** You are required to attach one `.zip` file, named `hw07_yourNetID.zip`, which contains:
  - Your RMarkdown file which should be saved as `hw07_yourNetID.Rmd`. For example `hw07_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw07_yourNetID.html`. For example `hw07_dunger.html`.
  - Any raw data supplied by me for the assignment. For example `nutrition.csv`.
- Your resulting `.html` file will be considered a "report" which is the material that will determine the majority of your grade. Be sure to visibly include all `R` code and output that is relevant to answering the exercises. (You do not need to include irrelevant code you tried that resulted in error or did not answer the question correctly.)
- You are granted an unlimited number of submissi • ons, but only the last submission *before* the deadline will be viewed and graded.
- If you use this `.Rmd` file as a template, be sure to remove the directions section. Consider removing `eval = FALSE` from any code chunks provided in the template, if you would like to run that code as part of your assignment.

- Your `.Rmd` file should be written such that, if it is placed in a folder with any data your are asked to import, it will knit properly without modification.
- Unless otherwise stated, you may use `R` for each of the exercises.
- Be sure to read each exercise carefully!
- Include your Name and NetID in the final document, not only in your filenames.

Assignment

Exercise 1 (Brand Rankings)

For this exercise we will use the data stored in `cookies.csv`. In order to determine which of three recipes (`A`, `B`, and `C`) to use, a cookie manufacturer divided 18 individuals at random into three groups and asked each one of them to rate one recipe on a scale from 0 to 100.

Consider the model $y_{ij} = \mu + \alpha_i + e_{ij}$ where $\sum \alpha_i = 0$ and $e_{ij} \sim N(0, \sigma^2)$. Here, $\mu + \alpha_i$ represents the mean of group (recipe) i.

Create side-by-side boxplots of the ratings of the three recipes. Test for a difference among the three recipes. If there is a difference, which recipes are different? Use $\alpha = 0.10$ for all tests. Which recipe would you use?

Exercise 2 (Concrete Strength)

An engineer is investigating the strength of concrete beams made from four types of cement and employing three curing processes. For each cement-curing combination, three beams are made and their breaking strength is measured. (A 4 × 3 randomized factorial design with 3 replicates.)

Consider the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

$$i = 1, \cdots I \quad j = 1, \cdots J \quad k = 1 \cdots, K$$

and ϵ_{ijk} are $N(0, \sigma^2)$ random variables.

With constraints:

$$\sum \alpha_i = 0 \qquad \sum \beta_j = 0.$$

Additionally:

$$(\alpha\beta)_{1j} + (\alpha\beta)_{2j} + (\alpha\beta)_{3j} = 0(\alpha\beta)_{i1} + (\alpha\beta)_{i2} + (\alpha\beta)_{i3} + (\alpha\beta)_{i4} = 0$$

for any i or j.

Let α_i represent the main effect for cement, which has four levels.

Let β_j represent the main effect for curing process, which takes three levels.

The data can be found in `concrete.csv`. Test for interaction between the two factors. If necessary, test for main effects. Use $\alpha = 0.05$ for all tests. State the final model you choose. Also, create an interaction plot. Does this plot make sense for the model you chose? With the model you chose (and then fit), create a table that shows the estimated mean for each of the 4×3 factor level combinations.

Exercise 3 (Weight Gain)

A total of 60 rats were used in an experiment about the effects of protein quantity and source on weight gain. The experiment used a 2×3 randomized factorial design with 10 replicates. (For each of the 6 treatments, 10 rats were randomly chosen.)

Each rat was fed a `low` or `high` protein diet from one of three sources: `beef`, `cereal`, or `pork`. After a period of time, the weight `gain` (the response, y) of each was measured in grams.

Consider the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where

$$i = 1, \cdots I \quad j = 1, \cdots J \quad k = 1 \cdots, K$$

and ϵ_{ijk} are $N(0, \sigma^2)$ random variables.

With constraints:

$$\sum \alpha_i = 0 \qquad \sum \beta_j = 0.$$

Additionally:

$$(\alpha\beta)_{1j} + (\alpha\beta)_{2j} + (\alpha\beta)_{3j} = 0 (\alpha\beta)_{i1} + (\alpha\beta)_{i2} + (\alpha\beta)_{i3} + (\alpha\beta)_{i4} = 0$$

for any i or j.

Let α_i represent the main effect for protein quantity, which has two levels; `high` and `low`.

Let β_j represent the main effect for protein source, which takes three levels; `beef`, `cereal` and `pork`.

The data can be found in `rat_wt.csv`. Test for interaction between the two factors. If necessary, test for main effects. Use $\alpha = 0.10$ for all tests. State the final model you choose. Also, create an interaction plot. Does this plot make sense for the model you chose? With the model you chose (and then fit), create a table that shows the estimated mean for each of the 2×3 factor level combinations.

Exercise 4 (Sample Size, Power)

Now that we're performing experiments, getting more data means finding more test subjects, running more lab tests, etc. In other words, it will cost more time and money.

We'd like to design our experiment so that we have a good chance of detecting an interesting effect size, without spending too much money. There's no point in running an experiment if there's only a very low chance that it has a significant result **that you care about**. (Not all statistically significant results have practical value.)

Suppose we will run an experiment that compares three treatments: `A`, `B`, and `C`. From previous study, we believe the shared variance could be $\sigma^2 = 1$.

Consider the model $y_{ij} = \mu_j + e_{ij}$ where $e_{ij} \sim N(0, \sigma^2)$. Here $j = 1, 2, 3$, for `A`, `B`, and `C`.

The null hypothesis of the test we will run is:

$$H_0 : \mu_A = \mu_B = \mu_C$$

Suppose that we're interested in an alternative where

$$\mu_A = -1, \mu_B = 0, \mu_C = 1$$

Mostly, we've used simulation to verify results. Now, we'll use simulation to save money (in place of some rather difficult mathematics)!

Use simulation to determine the *minimum* sample size that has *at least* a 90% chance to reject the null hypothesis when that alternative is true and $\alpha = 0.05$. That is, find the sample size which gives a **power** of at least 0.90 for the stated alternative. Consider only balanced designs, which have the same number of replications in each group. For each sample size, use at least 250 simulations. (More simulations will give a better estimate of the power and will create a smoother resulting curve.)

Plot your results. What sample size do you choose?

Before performing the simulations, set a seed value equal to **your** birthday, as was done in the previous homework assignments.

```r
birthday = 18760613
set.seed(birthday)
```

Exercise 5 (Balanced Design, Power)

Why do we use a balanced (equal number of replicates in each group) design? To maximize power. Let's verify this with simulation.

Consider a simple example with 2 groups A and B and a *total* sample size of $N = 10$. Where should we place these samples (replicates) between A and B? Obviously, at least one replicate needs to be in each, but after that, we can choose.

Consider the model $y_{ij} = \mu_j + e_{ij}$ where $e_{ij} \sim N(0, \sigma^2 = 1)$. Here $j = 1, 2$, for `A` and `B`.

The null hypothesis of the test we will run is:

$$H_0 : \mu_A = \mu_B$$

Suppose that we're interested in an alternative where

$$\mu_A = 0, \mu_B = 2$$

Calculate the power for each of the possible placements of the replicates with $\alpha = 0.05$. (Essentially, for $n_a = 1, 2, \ldots 9$.) For each possibility, use at least 500 simulations. Plot the results. Does balance provide the best power?

Before performing the simulations, set a seed value equal to **your** birthday, as was done in the previous homework assignments.

```
birthday = 17770430
set.seed(birthday)
```