MSAN 631: Design and Analysis of Experiments

Cumulative-Rolling-Take-Home-Final Examination

Due Date: Thursday June 29th by 5:00pm

Description:

The list of problems below is intended to assess your competency with both technical and non-technical experimental design concepts. These problems also test your ability to analyze experiment data both by hand and by R. You may complete these problems at your own pace, so long as they are completed and submitted by the due date indicated above. You are expected to complete this examination independently and submit your own work.

Problems:

1. Discuss the merits and drawbacks of an experiment.

2. Discuss the importance of replication, randomization, and blocking in the context of an experiment.

3. Briefly explain the philosophy and mechanics of an "A/B" test.

4. Suppose that an online retailer is interested in determining whether the average purchase price is larger among customers in "condition A" than among customers in "condition B". Formally:

   $$H_0: \mu_1 \leq \mu_2 \text{ versus } H_A: \mu_1 > \mu_2$$

   To test this hypothesis the retailer randomizes $n = 1000$ customers to each condition and records their purchase prices. Data for this experiment can be found in the file `retail.txt`. Use R to analyze this data and hence conclude whether condition "A" results, on average, in significantly larger purchase prices than condition "B", at a 5% level of significance. Be sure to assess the validity of any assumptions your analysis makes.

   Assuming an effect size of $\delta = \mu_1 - \mu_2 = 1$, and a common standard deviation of 5, calculate the power of this hypothesis test, and determine the sample size that would be sufficient to achieve 90% power.

5. Suppose your colleague wishes to compare two groups but they are unwilling to use parametric hypothesis tests such as *t*-tests for means or *z*-tests for proportions. Instead they wish to use a non-parametric randomization test, and they ask you for your help. Unfortunately you are busy, so you only have time to write pseudocode for such a test. Provide that pseudocode here.

6. Suppose we are interested in testing the following hypothesis:

$$H_0: \pi_1 = \pi_2 \text{ versus } H_A: \pi_1 \neq \pi_2$$

where π_i is the proportion of individuals that conform to some condition in population $i = 1, 2$. Derive a formula for the sample size n that each sample would need to have to be able to this hypothesis a $100\times\alpha\%$ significance level and a power of $1 - \beta$.

Using this formula, calculate the sample size necessary for testing this hypothesis assuming Type I and II error rates of 5%, and assuming $p_1 = 0.78$ and $p_2 = 0.84$. Repeat this calculation using the `power.prop.test()` function in R.

7. Online retailers often use banner ads to drive traffic to their websites. Nike is trying to determine which of five online banner ads is related to the largest click-through-rate. To investigate which is best, the experimentation team plans to display each of the 5 ads to various US Facebook users in their newsfeeds. Because ad impressions are expected to differ by geographic region, each of the 5 ads is displayed 5000 times in Facebook newsfeeds for IP addresses originating in each of the four major US geographic regions: Northeast, Midwest, South, and West.

The data may be described as follows:

$$y_{ijk} = \begin{cases} 1 & if\ ad\ i\ was\ clicked\ by\ person\ j\ in\ region\ k \\ 0 & otherwise \end{cases}$$

where $i = 1, \ldots, 5$ indexes ads, $j = 1, \ldots, 5000$ indexes subjects and $k = 1, 2, 3, 4$ corresponds to geographic regions Northeast, Midwest, South and West. The data for this experiment can be found in the file `nike.txt`.

Let π_{ik} denote the click-through-rate for ad i in region (block) k. Use R to conduct a χ^2 test of the hypothesis

$$H_0: \pi_{1k} = \cdots = \pi_{5k} \text{ versus } H_A: \pi_{ik} \neq \pi_{lk} \text{ for some } i \neq l$$

in each of the four geographic regions. Use a series of pairwise tests in each region to decide which ads are associated with the highest and lowest click-through-rates per region. For purposes of this question, you may ignore the "multiple testing" problem.

Discuss the relevance of blocking here, and why pooling data across geographic regions may be misleading.

8. Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the color bomb. Users in each condition receive (for free) 5 boosters corresponding to their condition and interest lies in

evaluating the effect of these different boosters on the length of time a user plays the game. Let μ_i represent the average length of game play (in minutes) associated with booster $i = 1, 2, 3$. Use R and the data found in `candycrush.txt` to conduct an *F*-test of the hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ versus } H_A: \mu_i \neq \mu_j \text{ for some } i \neq j$$

Use a series of pairwise *t*-tests to decide which booster is associated with the largest average length of game play. For each of the pairwise tests use a significance level in accordance with the 'Bonferroni correction' to help mitigate the "multiple testing" problem.

9. (a) Show that when k hypothesis tests are being conducted, each at significance level α, the experimentwise Type I error rate is given by:

$$1 - (1 - \alpha)^k$$

(b) Calculate the experimentwise Type I error rate for large k (i.e., as $k \to \infty$) when using the Bonferroni correction, and justify the efficacy of this correction.

10. In the world of Big Data, there is typically little emphasis placed on obtaining an adequate sample size as it is often believed that exceedingly large samples are easily attainable. However, when performing inference via usual hypothesis tests and/or confidence intervals, a sample that is too large can be problematic. Consider a simple one-sample *t*-test of the hypothesis

$$H_0: \mu = \mu_0 \text{ versus } H_A: \mu \neq \mu_0$$

where the test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Notice that as n gets very large, t also becomes very large and hence the associated *p*-value gets infinitesimally small. In other words, if your sample is large enough, you will always find sufficient evidence to reject the null-hypothesis.

Do some light reading on the topic of "equivalence tests". Provide a brief description of this method of hypothesis testing, and discuss some of the advantages it enjoys relative to the standard hypothesis-testing framework.

11. Consider an experiment with a continuous response and one factor with a levels. The *effects model* that describes data arising from a such a design is given by:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $i = 1, \ldots, a$ indexes the levels of the factor, $j = 1, \ldots, n$ indexes replication and we assume the ε_{ij} are i.i.d. $N(0, \sigma^2)$.

In class we showed that the ANOVA F-test is an appropriate test of $H_0: \mu_1 = \mu_2 = \cdots = \mu_a$, and that in the context of this effects model it is equivalent to the test of $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$. The appropriateness of the ANOVA F-test relied on the fact that $E(MS_{Cond}) = E(MS_E)$ if H_0 is true. Here you are required to derive these two expected mean squares. In particular, prove that the following expected-mean-square equivalences are true.

$$E(MS_{Cond}) = \sigma^2 + \frac{n \sum_{i=1}^{a} \tau_i^2}{a - 1}$$

$$E(MS_E) = \sigma^2$$

12. Consider the Instagram factorial experiment from class, in which the influence of two factors on engagement time was investigated. Use R to answer the following questions. The data are available in the `instagram-factorial.txt` file.

    (a) Construct a main effects plot for the factor "prevalence" and explain whether and why you do/do not believe it is a significant factor. Comment on which level is most detrimental to the user experience.

    (b) Construct a main effects plot for the factor "type" and explain whether and why you do/do not believe it is a significant factor. Comment on which level is most detrimental to the user experience.

    (c) Using the factors "prevalence" and "type" as an example, explain what is meant by interaction. Construct an interaction plot for these factors explain whether and why you do/do not believe this effect is significant.

    (d) Fit the full model with all main effects and interaction terms and formally evaluate the statistical significance of each effect. What do you conclude? Do these formal conclusions agree with the informal ones drawn in parts (a)-(c)?

    (e) Based on this analysis, which combination of factor levels appears to maximize engagement? Discuss the practical implications of this choice.

    (f) Assess the residuals of the model in part (d). Do the typical ordinary least squares assumptions appear to be met?

13. In the context of a 2^k factorial design, we typically model factor levels in coded units (± 1) rather than natural units. A model for a single replicate of a 2^2 factorial design can be written as follows:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{12})^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)^T$, $\varepsilon_i \sim N(0, \sigma^2)$ and

$$X = [\mathbf{1} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2} \quad \boldsymbol{x_1 \cdot x_2}] = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix}$$

One consequence of using these coded units is that the columns of the design matrix in such models are orthogonal. There are several convenient by-products of this.

(a) Show that the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are unaffected by the presence/absence of the interaction term.

(b) Show that the standard errors of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_{12}$ are all the same.

(c) Explain why $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_{12}$ are each interpreted as one-half of the corresponding factor's effect estimate.

14. Fractional factorial designs are an efficient approach to investigating the effects of a large number of factors.

(a) Briefly explain the philosophy and mechanics of a 2^{k-p} fractional factorial design.

(b) Briefly explain *aliasing* and *resolution*, and why a resolution-IV design is to be preferred over a resolution-III design.

Note that there are NO more problems to come

Hooray!!