STAT 430/830: Assignment 1

DUE: Friday May 29 by 11:59pm EST

INSTRUCTIONS

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via Crowdmark. This means that your responses for different questions should begin on separate pages.

Your solutions should be prepared in a clear and coherent manner. For written responses and derivations, you may upload solutions produced by LaTeX (or some other equation editing software), or handwritten solutions that have been scanned or photographed. Please note that if you scan or photograph your solutions it is your duty to ensure they are readable and of high quality. If the marking team cannot read your solutions, they simply will not be marked.

For questions that involve computation in R, I highly recommend you produce your solutions as a nicely formatted .pdf file with R Markdown. R Markdown facilitates the seamless combination of written text, LaTeX equations and R code/output. Your submission for these questions should include the code, the corresponding output, and any interpretations where appropriate.

DISCLAIMER

The companies, teams, and problems described in this assignment are real, but the experiments are hypothetical and the data are simulated. These are not real experiments, and so it would be inappropriate to represent them as such. These cases are intended for instructional purposes only.

JOB ADS

Shutterfly, MasterClass and Twitch all currently have openings for positions that explicitly require expertise in the design and analysis of experiments:

- Shutterlfy: Sr. Data Scientist II
- MasterClass: Sr. Data Scientist
- Twitch: Product Analyst

QUESTION 1 [14 points]

Shutterfly, Inc. is a retailer and manufacturer of personalized photography-based products and communications. From photo books, calendars, and stationary, to mugs and home decor, Shutterly offers creative and innovative ways to share memories and stay connected with loved ones.

Shutterlfy's Customer Insights and Analytics (CIA) team is devoted to improving their customer's experiences, and hence revenue, through personalization efforts. In particular, they are interested in developing machine learning algorithms to optimally recommend products to their customers. Suppose that the data scientists on the CIA team are experimenting with four different versions of a deep learning model based on different configurations of the model's hyperparameters. Four thousand users were randomly selected and then randomly assigned to one of the four versions of the model for their session (i.e., 1000 users in each condition). Interest lies in determining whether or not the users purchase one of the items recommended to them by the deep learning model. If a user purchases one or more of the recommended items, this is considered a success for the model. Ultimately, the version of the model with the highest success rate will be put into full production.]

In the sentences below, fill in the blanks with words from the following list of terms. Note that there is exactly one correct answer for each blank space. In your submission, please just state the appropriate words. There is no need to recreate the whole sentences.

alternative; *A/B test*; *blocking*; *causal*; *controllable*; *design*; *experimental*; *explanatory*; *factor*; *factorial*; *F-test*; *levels*; *metric of interest*; *nuisance*; *null*; *observational*; *power*; *randomization*; *replication*; *response*; *sample*; *sample size*; *significance level*; *statistical*; *t-test*; *Type I*; *Type II*; *uncontrollable*; *unit*; *valid*; *Z-test*;

(a) [2 points] The Shutterfly experiment can be colloquially referred to as an _____. More formally, we think of this as an experiment with one factor that has four _____.

(b) [2 points] The number 1000 is called the _____ and this corresponds to the experimental design principle called _____.

(c) [1 point] A power analysis was used to determine that each condition required 1000 users. Such an analysis is used to control _____ error.

(d) [1 point] Each of the Shutterfly users is considered an experimental _____.

(e) [1 point] The manner in which the users were selected for inclusion in the experiment is an example of the experimental design principle called _____.

(f) [1 point] The _____ variable, which indicates whether a user purchases a recommended item, is binary.

(g) [1 point] The algorithm *success rate* is the _____.

(h) [1 point] The hypothesis test that is most appropriate for this experiment is a _____.

(i) [2 points] Suppose the test in (h) is carried out and a p-value is calculated. In order to draw a formal conclusion this p-value must be compared to the _____ whose value is chosen to control _____ error.

(j) [2 points] The benefit of such an experiment, relative to an _____ study, is that it facilitates _____ inference.

QUESTION 2 [12 points]

(a) In each of the following questions, calculate the appropriate test statistic given the null hypothesis and the relevant data summaries. You may use R for this question, but make sure to show your work.

i. [1 point] $H_0 : \mu_1 = \mu_2$ (assuming $\sigma_1 = \sigma_2$ are unknown)

- $n_1 = 500$, $\widehat{\mu}_1 = \overline{y}_1 = 30$, $\widehat{\sigma}_1 = s_1 = 5$
- $n_2 = 500$, $\widehat{\mu}_2 = \overline{y}_2 = 33$, $\widehat{\sigma}_2 = s_2 = 4$

ii. [1 point] $H_0 : \mu_1 \leq \mu_2$ (assuming $\sigma_1 \neq \sigma_2$ are unknown)

- $n_1 = 300$, $\widehat{\mu}_1 = \overline{y}_1 = 110$, $\widehat{\sigma}_1 = s_1 = 10$
- $n_2 = 300$, $\widehat{\mu}_2 = \overline{y}_2 = 100$, $\widehat{\sigma}_2 = s_2 = 12$

iii. [1 point] $H_0 : \pi_1 \geq \pi_2$

- $n_1 = 1000$, $\widehat{\pi}_1 = \overline{y}_1 = 0.03$
- $n_2 = 1000$, $\widehat{\pi}_2 = \overline{y}_2 = 0.05$

iv. [1 point] $H_0 : \sigma_1^2 = \sigma_2^2$

- $n_1 = 300$, $\widehat{\mu}_1 = \overline{y}_1 = 110$, $\widehat{\sigma}_1 = s_1 = 10$
- $n_2 = 300$, $\widehat{\mu}_2 = \overline{y}_2 = 100$, $\widehat{\sigma}_2 = s_2 = 12$

(b) Suppose we perform an experiment with two conditions containing $n_1 = 200$ and $n_2 = 100$ units, respectively. For each of the hypotheses and test statistics below, state the null distribution and calculate the appropriate p-value. Note that in the case of Welch's t-test you may use the approximate degrees of freedom $\min(n_1, n_2) - 1$. You may use R for this question, but make sure to show your work.

i. [2 points] $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$ (assuming $\sigma_1 = \sigma_2$ are unknown)

- $t = 0.5$

ii. [2 points] $H_0 : \mu_1 \geq \mu_2$ vs. $H_A : \mu_1 < \mu_2$ (assuming $\sigma_1 \neq \sigma_2$ are unknown)

- $t = -1.8$

iii. [2 points] $H_0 : \pi_1 \leq \pi_2$ vs. $H_A : \pi_1 > \pi_2$

- $t = 2.1$

iv. [2 points] $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_A : \sigma_1^2 \neq \sigma_2^2$

- $t = 0.9$

QUESTION 3 [6 points]

STAT 430 ONLY

Suppose we are interested in using the Z-test for proportions to test the following hypothesis:

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

where π_j is the expected response in condition j assuming $Y_j \sim \text{BIN}(1, \pi_j)$, $j = 1, 2$.

(a) [3 points] Show that the sample size n (in each condition) required to test this hypothesis at a significance level α and power $1 - \beta$ is given by

$$n = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2 \left[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right]}{(\pi_1 - \pi_2)^2}$$

where z_c is the value in the standard normal distribution such that $\Pr(Z \geq z_c) = c$.

(b) [2 points] Based on the equation in part (a), write a sample size determination function in R called `ssd_z_test_prop` which takes as input `p1`, `p2`, `sig.level`, and `power`, and outputs the required integer sample size `n`, where

- `p1`: the best guess of π_1
- `p2`: the best guess of π_2
- `sig.level`: the significance level α
- `power`: the power $1 - \beta$

(c) [1 point] Using your function from part (b) determine the sample size required in each condition to test the hypothesis above, assuming Type I and Type II error rates of 1%, and assuming $\pi_1 = 0.05$ and $\pi_2 = 0.06$.

STAT 830 ONLY

Suppose we are interested in using the Student's t-test to test the following hypothesis:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

where μ_j is the expected response in condition j assuming $Y_j \sim \text{N}(\mu_j, \sigma^2)$, $j = 1, 2$.

(a) [3 points] Show that the power associated with this test is

$$1 - \beta = \Pr\left(X \geq t^* - \frac{\delta}{\widehat{\sigma}}\sqrt{\frac{n}{2}}\right) + \Pr\left(X \leq -t^* - \frac{\delta}{\widehat{\sigma}}\sqrt{\frac{n}{2}}\right)$$

where δ is the minimum detectable effect, $\widehat{\sigma}$ is an estimate of the common standard deviation, $X \sim t_{2(n-1)}$, and where $\Pr(X \geq t^*) = \alpha/2$.

(b) [2 points] Based on the equation in part (a), write a sample size determination function in R called `ssd_t_test` which takes as input `delta`, `sd`, `sig.level`, and `power`, and outputs the required integer sample size `n`, where

- `delta`: the minimum detectable effect δ
- `sd`: the best guess of $\widehat{\sigma}$
- `sig.level`: the significance level α
- `power`: the power $1 - \beta$

(c) [1 point] Using your function from part (b) determine the sample size required in each condition to test the hypothesis above, assuming Type I and Type II error rates of 1%, and assuming $\delta = 0.5$ and $\widehat{\sigma} = 1.5$.

QUESTION 4 [5 points]

STAT 430 ONLY

Assume $Y_{ij} \overset{iid}{\sim} \mathrm{N}\left(\mu_j, \sigma^2\right)$ for $i = 1, 2, \ldots, n_j$, $j = 1, 2$. Then we know that

$$\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \sim \mathrm{N}\left(\mu_j, \frac{\sigma^2}{n_j}\right) \qquad \text{and} \qquad W_j = \frac{(n_j - 1)\widehat{\sigma}_j^2}{\sigma^2} \sim \chi^2_{(n_j-1)}$$

where

$$\widehat{\sigma}_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2.$$

Furthermore, we know $\overline{Y}_1, \overline{Y}_2, W_1, W_2$ are all independent of eachother. Using these facts, prove the following two distributional results.

(a) [3 points]

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

(b) [2 points]

$$T = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F_{(n_1-1, n_2-1)}$$

STAT 830 ONLY

The problem of *post selection inference* is to ensure the validity of inferences one makes *after* some selection event has taken place. This issue commonly arises in online A/B tests. Typically two experimental conditions are being compared and we wish to determine the "winner", i.e., the condition that optimizes some metric of interest. In order to do so, a hypothesis such as

$$H_0 : \mu_T \leq \mu_C \text{ versus } H_A : \mu_T > \mu_C$$

is tested, where μ_T and μ_C represent the metric of interest in the treatment and control conditions, respectively. If H_0 is rejected, it suggests that the treatment condition significantly out performs the control. Then, once this decision has been made, the true treatment effect $\delta = \mu_T - \mu_C$ is estimated by $\widehat{\delta} = \widehat{\mu}_T - \widehat{\mu}_C = \overline{y}_T - \overline{y}_C$.

Ordinarily, $\widehat{\delta}$ would be an unbiased estimate of the true treatment effect δ, but this unbiasedness property does not account for the fact that the null hypothesis H_0 has already been rejected.

Define $\widetilde{\delta} = \widetilde{\mu}_T - \widetilde{\mu}_C = \overline{Y}_T - \overline{Y}_C$ to be the treatment effect estimator and assume that due to the normality of the underlying data, or due to the central limit theorem, we have $\widetilde{\delta} \sim \mathrm{N}(\delta, \sigma^2)$, where $\sigma^2 = \mathrm{Var}\left[\overline{Y}_T - \overline{Y}_C\right]$.

(a) [3 points] In this question you will prove that the estimated treatment effect which is only estimated *after* H_0 is rejected (at an $\alpha \times 100\%$ significance level), is biased. In particular, show that

$$\mathrm{E}\left[\widetilde{\delta} \mid H_0 \text{ is rejected}\right] = \delta + \sigma\frac{\phi\left(z_\alpha - \frac{\delta}{\sigma}\right)}{1 - \Phi\left(z_\alpha - \frac{\delta}{\sigma}\right)}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal probability density and cumulative distribution functions.

(b) [2 points] Based on the result in part (a), explain why the bias represents an overestimation of δ and hence an exaggeration of the good performance of the "winning" condition.

QUESTION 5 [10 points]

MasterClass is an online learning platform in which students engage with lectures and tutorials taught by experts and "reknowned personalities in their respective fields". For instance, Samuel L. Jackson can teach you to act, Christina Aguilera can teach you to sing, and Steph Curry can teach you to dribble and shoot. For $20/month MasterClass offers an "all-access" subscription which unlocks for users the entire catalogue of classes which span genres such as "Music & Entertainment", "Writing", "Design, Photography & Fashion", "Culinary Arts", "Film & TV" and "Sport & Games".

Suppose the Data Team at MasterClass is interested in determing whether to showcase Gordon Ramsey or Wolfgang Puck on their "Culinary Arts" landing page. In particular, they would like to determine whether there is a difference in subscription rates between the two celebrity chefs. To investigate this, two experimental conditions are devised: one in which visitors to the "Culinary Arts" page are automatically shown a video preview of Gordon Ramsey's master class and another in which visitors are shown a video preview of Wolfgang Puck's master class.

(a) [2 points] What is the metric of interest and what is the corresponding response variable?

(b) [2 points] What is the design factor and what are its levels?

(c) [2 points] State the null and alternative hypotheses for this experiment. Be sure to define any notation you use.

(d) [2 points] The file `masterclass.csv` contains observations on whether or not a given visitor to the site initiated a subscription (1=subscribed, 0=not). Note that the experiment consisted of 5000 users (2500 in each condition). Using this observed data, test the hypothesis in (c) at a 5% significance level. Clearly state your conclusion in the context of the problem. You may use R for this question, but make sure to show your work.

(e) [2 points] Provide a point estimate and a 95% confidence interval for the true difference in subscription rates between the two conditions. You may use R for this question, but make sure to show your work.

QUESTION 6 [17 points]

Twitch is a video live streaming service whose primary focus is video game live streaming, where broadcasters stream a video feed of themselves and their game's screen while they play. Although Twitch is dominated by gamers every once in a while you may find a statistics lecture or two. With over 2 million monthly broadcasters and 15 million daily active users, Twitch is the leading streaming service of its kind.

The Mobile team at Twitch regularly uses experimentation as a means to guide product changes in their iOS and Android apps. For instance, consider an experiment on their "Live Channels" tab. This tab consists of a vertically scrollable list that displays all of the channels that are streaming live at any given moment in time. Twitch makes money when users watch live streams, so it's in Twitch's best interest to entice a user to select a stream and begin watching as quickly as possible. The experiment consists of two conditions: one in which each live stream is previewed with a static image, and another in which the streams are previewed with a video clip. Interest lies in determining whether the average time until selection is lower in the video condition than in the static image condition.

(a) [2 points] What is the metric of interest and what is the corresponding response variable?

(b) [2 points] What is the design factor and what are its levels?

(c) [2 points] State the null and alternative hypotheses for this experiment. Be sure to define any notation you use.

(d) [2 points] The file `twitch.csv` contains observations on the time until selection (in seconds) for 2000 users (1000 in each condition). Plot two histograms of time until selection, one for each condition. On each histogram indicate the mean and median time until selection with red and blue vertical lines, respectively. Be sure to add a title and axis labels. Comment on which condition appears to minimize selection times.

(e) [2 points] Using the observed data, test the hypothesis in (c) at a 5% significance level. Clearly state your conclusion in the context of the problem. You may use R for this question, but make sure to show your work.

(f) [2 points] Provide a point estimate and a 95% confidence interval for the true difference in average time until selection between the two conditions. You may use R for this question, but make sure to show your work.

(g) [5 points] Suppose that instead of average time until selection, the Mobile team wishes to compare the two conditions on the basis of their median time until selection. Using a randomization test (with N=10,000), determine whether video previews significantly ($\alpha = 0.05$) decreases the median time until selection relative to static image previews. Be sure to state the hypothesis being tested, define any notation used, and clearly state your conclusions in the context of the problem. You must use R for this question.