STAT 430/830: Assignment 2

DUE: Friday June 19 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

Cash App and ThirdLove both currently have openings for positions that explicitly require expertise in the design and analysis of experiments:

- Cash App: Senior Product Data Scientist
- ThirdLove: Data Scientist - Product Analytics

QUESTION 1 [12 points]

(a) [3 points] Consider the following hypothesis scenario.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ vs. } H_A : \mu_j \neq \mu_k \text{ for some } j \neq k$$

where the number of units in each condition is given by $n_1 = 721, n_2 = 766, n_3 = 791, n_4 = 770$ and the resulting test statistic calculated from the observed data is $t = 2.7$

i. [1 point] State the null distribution of this test statistic.
ii. [1 point] Calculate the appropriate p-value. You should use R for this part, but be sure to state the formula for the p-value in addition to the value calculated in R.
iii. [1 point] Using a 5% significance level, state whether you reject or do not reject H_0.

(b) [3 points] Consider the following hypothesis scenario.

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 \text{ vs. } H_A : \pi_j \neq \pi_k \text{ for some } j \neq k$$

where the number of units in each condition is given by $n_1 = n_2 = n_3 = n_4 = n_5 = 1000$ and the resulting test statistic calculated from the observed data is $t = 8.1$

i. [1 point] State the null distribution of this test statistic.
ii. [1 point] Calculate the appropriate p-value. You should use R for this part, but be sure to state the formula for the p-value in addition to the value calculated in R.
iii. [1 point] Using a 5% significance level, state whether you reject or do not reject H_0.

(c) [3 points] Consider the following hypothesis scenario.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_A : \mu_j \neq \mu_k \text{ for some } j \neq k$$

where the number of units in each condition is given by $n_1 = 10011, n_2 = 9874, n_3 = 10023$ and the resulting test statistic calculated from the observed data is $t = 2.7$

i. [1 point] State the null distribution of this test statistic.
ii. [1 point] State the rejection region associated with this test at a 5% significance level.
iii. [1 point] Based on the observed test statistic and your rejection region from ii., state whether you reject or do not reject H_0.

(d) [3 points] Consider the following hypothesis scenario.

$$H_0 : \pi_1 = \pi_2 = \pi_3 \text{ vs. } H_A : \pi_j \neq \pi_k \text{ for some } j \neq k$$

where the number of units in each condition is given by $n_1 = n_2 = n_3 = 500$ and the resulting test statistic calculated from the observed data is $t = 8.1$

i. [1 point] State the null distribution of this test statistic.
ii. [1 point] State the rejection region associated with this test at a 5% significance level.
iii. [1 point] Based on the observed test statistic and your rejection region from ii., state whether you reject or do not reject H_0.

QUESTION 2 [5 points] (STAT 430 ONLY)

Hypotheses of the form

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

may be tested by the Z-test for proportions or the χ^2-test of independence. Show that these tests are statistically equivalent (i.e., their p-values are identical).

Hint: Recognize that $\chi^2_{(1)} = [\mathrm{N}(0,1)]^2$ and show that

$$\sum_{l=0}^{1}\sum_{j=1}^{2}\frac{(O_{l,j}-E_{l,j})^2}{E_{l,j}} = \left(\frac{\widehat{\pi}_1-\widehat{\pi}_2}{\sqrt{\widehat{\pi}(1-\widehat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right)^2$$

QUESTION 2 [5 points] (STAT 830 ONLY)

Hypotheses of the form

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

may be tested by the Student's t-test or the F-test of overall significance in an appropriately defined linear regression. Show that these tests are statistically equivalent (i.e., their p-values are identical).

Hint: Recognize that $F_{(1,\nu)} = [t_{(\nu)}]^2$ and show that

$$\frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(\overline{y}_{\cdot j}-\overline{y}_{\cdot\cdot}\right)^2}{\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(y_{ij}-\overline{y}_{\cdot j}\right)^2/(n_1+n_2-2)} = \left(\frac{\widehat{\mu}_1-\widehat{\mu}_2}{\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)^2$$

QUESTION 3 [10 points] (STAT 430 ONLY)

Consider the following hypothesis:

$$H_0 : \mu_1 \geq \mu_2 \text{ vs. } H_A : \mu_1 < \mu_2$$

And suppose we test this with a Student's t-test with $n_1 = n_2 = 500$ units in each condition.

(a) [4 points] The null distribution for this test is $T \sim t_{(998)}$. Thus, if H_0 is true, t should look like it comes from a $t_{(998)}$ distribution.

- Using R, simulate 100,000 such values of t using the `rt` function and for each value of t calculate the p-value associated with the hypothesis above.
- Construct a histogram of the 100,000 p-values. Comment on what you observe.
- Using R, simulate 100,000 observations from the UNIF$(0, 1)$ distribution using the `runif` function.
- Construct a QQ-plot comparing the distribution of the p-values to the distribution of the UNIF$(0, 1)$ values. Comment on what you observe.

(b) [4 points] Let X be a continuous random variable with cumulative distribution function (CDF) $F(x) = \Pr(X \leq x)$. Define $Y = F(X)$ and let $G(y) = \Pr(Y \leq y)$ represent the CDF of Y. Show that $Y \sim \text{UNIF}(0, 1)$. [*Hint:* $W \sim \text{UNIF}(0, 1)$ if and only if $\Pr(W \leq w) = w$].

(c) [2 points] Let P be a random variable representing the p-value associated with the hypothesis above. Using the result from (b), show that $P \sim \text{UNIF}(0, 1)$.

QUESTION 3 [10 points] (STAT 830 ONLY)

Consider the following hypothesis:

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 \text{ vs. } H_A : \pi_j \neq \pi_j \text{ for some } j \neq k$$

And suppose we test this with a χ^2-test of independence.

(a) [4 points] The null distribution for this test is $T \sim \chi^2_{(3)}$. Thus, if H_0 is true, t should look like it comes from a $\chi^2_{(3)}$ distribution.

- Using R, simulate 100,000 such values of t using the `rchisq` function and for each value of t calculate the p-value associated with the hypothesis above.
- Construct a histogram of the 100,000 p-values. Comment on what you observe.
- Using R, simulate 100,000 observations from the UNIF$(0, 1)$ distribution using the `runif` function.
- Construct a QQ-plot comparing the distribution of the p-values to the distribution of the UNIF$(0, 1)$ values. Comment on what you observe.

(b) [4 points] Let X be a continuous random variable with cumulative distribution function (CDF) $F(x) = \Pr(X \leq x)$. Define $Y = F(X)$ and let $G(y) = \Pr(Y \leq y)$ represent the CDF of Y. Show that $Y \sim \text{UNIF}(0, 1)$, and show that $1 - Y \sim \text{UNIF}(0, 1)$ also. [*Hint:* $W \sim \text{UNIF}(0, 1)$ if and only if $\Pr(W \leq w) = w$].

(c) [2 points] Let P be a random variable representing the p-value associated with the hypothesis above. Using the results from (b), show that $P \sim \text{UNIF}(0, 1)$.

QUESTION 4 [13 points]

Consider an experiment with a continuous response and one design factor with m levels. In class we saw that the ANOVA F-test of overall significance in an appropriately defined linear regression model was an appropriate test of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A : \mu_k \neq \mu_l \text{ for some } k \neq l$$

In this question you will develop the same test from the perspective of the following *effects model*:

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

where $j = 1, 2, \ldots, m$ indexes conditions and $i = 1, 2, \ldots, n_j$ indexes replication within conditions. Further, we assume $\varepsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ and we define $\mu_j = \mu + \tau_j$ to be the expected response in condition j where $\mu = \frac{1}{m}\sum_{j=1}^{m} \mu_j$ is interpreted as a global mean. These definitions necessitate the constraint $\sum_{j=1}^{m} \tau_j = 0$ where each τ_j may be viewed as a condition-specific deviation from the global mean μ, and may be interpreted as the *effect* of being in condition j.

In the context of this model, the hypothesis above can be equivalently stated as

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0 \text{ vs. } H_A : \tau_k \neq \tau_l \text{ for some } k \neq l$$

We can formulate the ANOVA F-test by partitioning the *overall* variability in the observed response into two components: *between-condition* variability and *within-condition* variability. In particular we quantify each of these types of variation with the following respective sums of squares:

$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad SS_C = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad SS_E = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{\cdot j}\right)^2$$

where

$$\overline{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \qquad \text{and} \qquad \overline{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j} Y_{ij}$$

and $N = n_1 + n_2 + \cdots + n_m$. The ANOVA table for this model is given by:

Source	SS	df	MS	Test Stat.
Condition	SS_C	$m-1$	$MS_C = \frac{SS_C}{m-1}$	$T = \frac{MS_C}{MS_E}$
Error	SS_E	$N-m$	$MS_E = \frac{SS_E}{N-m}$	
Total	SS_T	$N-1$		

(a) [3 points] Show that $SS_T = SS_C + SS_E$.

(b) [10 points] **STAT 430 ONLY** The form of the test statistic may justified by recognizing that

$$\mathrm{E}[MS_E] = \sigma^2 \qquad \text{and} \qquad \mathrm{E}[MS_C] = \sigma^2 + \frac{\sum_{j=1}^{m} n_j \tau_j^2}{m-1}$$

Clearly when $H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0$ is true, $\mathrm{E}[MS_C] = \mathrm{E}[MS_E] = \sigma^2$ and when it's false $\mathrm{E}[MS_C] > \mathrm{E}[MS_E]$. Thus the ratio MS_C/MS_E is a sensible test statistic: values close to 1 provide evidence in favor of the null hypothesis and values much larger than 1 provide evidence against it.

Prove the two expectation results concerning $\mathrm{E}[MS_C]$ and $\mathrm{E}[MS_E]$.

(c) [10 points] **STAT 830 ONLY** That the test statistic $T = MS_C/MS_E$ follows an $F_{(m-1,N-m)}$ distribution relies on the fact that

$$\frac{SS_C}{\sigma^2} \sim \chi^2_{(m-1)} \qquad \text{and} \qquad \frac{SS_E}{\sigma^2} \sim \chi^2_{(N-m)}$$

when $H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0$ is true. Prove these two distributional results. [*Hint:* Consider using moment generating functions or Cochran's Theorem.]

QUESTION 5 [19 points]

Cash App is a mobile finance application developed by Square Inc. (a "fintech" venture founded by Jack Dorsey, the founder of Twitter). What began simply as a peer-to-peer mobile payment service has grown into a robust alternative to the traditional banking system, with features like direct deposit, debit cards, stock trading and more. This expansion into additional financial services has coincided with incredible user growth; in 2020 Cash App reached 24 million monthly active users and in 2019 their net revenue exceeded $1 billion.

A chief contributor to the success of Cash App has been a series of unique viral marketing campaigns via Twitter, Instagram, and TikTok as well as partnerships with celebrities and other brands such as Burger King, Spotify, and Nintendo's Animal Crossing. Cash App boasts "more engagement with millions of followers across social media in a day than most brands see in a year". It is clear that social engagement is a key component of their customer acquisition and retention strategy.

Suppose that the data scientists on the Growth team are experimenting with three different engagement incentives for new users. In particular, new users were randomly assigned to one of three conditions:

1. No engagement incentive
2. If you send/receive money with 5 different people in your first 14 days, $10 will be added to your Cash App balance.
3. If you send/receive money with 10 different people in your first 14 days, $25 will be added to your Cash App balance.

The data scientists are interested in determining which incentive maximizes the 30-day retention rate (i.e., the proportion of new users that remain active users 30-days after the app's installation).

(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) [1 points] What constitutes an experimental unit in this experiment?

(d) [1 point] State the null and alternative hypotheses for a test of overall equality in the context of this experiment. Use the notation π_j which represents the metric of interest in condition $j = 1, 2, 3$.

(e) [4 points] The file `cashapp.csv` contains observations for 4500 new users (1500 in each of the three experimental conditions). For each new user, a binary indicator records whether or not they are still an active user 30-days post-installation (recorded as 1 if they are and 0 if they're not). Using this data, test the hypothesis in (d) at a 5% significance level. Clearly state your conclusion in the context of the problem. Note that although you may use R to test this hypothesis, be sure to:

- State the formula and the value of the test statistic
- State the null distribution
- State the formula and the value of the p-value

(f) [3 points] As a follow-up to part (e), calculate the p-values associated with each of the following four hypotheses. Note that you need only state the resulting values themselves. You may find the `pairwise.prop.test` function in R useful.

$$H_0 : \pi_1 \geq \pi_2 \text{ versus } H_A : \pi_1 < \pi_2$$
$$H_0 : \pi_1 \geq \pi_3 \text{ versus } H_A : \pi_1 < \pi_3$$
$$H_0 : \pi_2 \geq \pi_3 \text{ versus } H_A : \pi_2 < \pi_3$$
$$H_0 : \pi_2 = \pi_3 \text{ versus } H_A : \pi_2 \neq \pi_3$$

(g) [6 points] Using the p-values from (f), in this question you will identify the incentive that maximizes 30-day retention rate while controlling the family-wise error rate. Note that you may use `p.adjust()` where appropriate.

i. [2 points] Calculate the Bonferroni-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.05$.

ii. [2 points] Calculate the Šidák-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.05$.

iii. [2 points] Calculate the Holm-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.05$.

QUESTION 6 [22 points]

ThirdLove is a San Francisco-based bra manufacturer and retailer whose mission is to empower women to feel comfortable and confident in their everyday lives. ThirdLove is known for taking a very data-driven approach to product development and marketing; this has helped them capture significant market share in the $100 billion bra industry and become one of the fastest growing consumer brands in the US.

One e-shopping feature that is hypothesized to improve the user experience and ultimately sell more product is the "See on Other Sizes" feature. When browsing a particular garment, this feature allows the customer to view images of the same garment worn by different sized models. The Product Data Science Team is interested in determining whether increasing the number of "Other Sizes" a customer can view will lead to an increase in the average purchase size (i.e, the average number of items a customer purchases). To investigate this they conduct an experiment with four conditions:

- Condition 1: The feature is turned off (i.e., a customer sees the garment worn by just one model).
- Condition 2: The feature allows customers to see the garment worn by two different sized models.
- Condition 3: The feature allows customers to see the garment worn by three different sized models.
- Condition 4: The feature allows customers to see the garment worn by four different sized models.

(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) [1 points] What constitutes an experimental unit in this experiment?

(d) [1 point] State the null and alternative hypotheses for a test of overall equality in the context of this experiment. Use the notation μ_j which represents the metric of interest in condition $j = 1, 2, 3, 4$.

(e) [4 points] The file `thirdlove.csv` contains observations for 2000 customers (500 in each of the three experimental conditions). For each customer, the number of items they purchased is recorded. Using this data, test the hypothesis in (d) at a 5% significance level. Clearly state your conclusion in the context of the problem. Note that although you may use R to test this hypothesis, be sure to:

- State the formula and the value of the test statistic
- State the null distribution
- State the formula and the value of the p-value

(f) [3 points] As a follow-up to part (e), calculate the p-values associated with each of the following seven hypotheses. Note that you need only state the resulting values themselves. You may find the `pairwise.t.test` function in R useful.

$$H_0 : \mu_1 \geq \mu_2 \text{ versus } H_A : \mu_1 < \mu_2$$
$$H_0 : \mu_1 \geq \mu_3 \text{ versus } H_A : \mu_1 < \mu_3$$
$$H_0 : \mu_1 \geq \mu_4 \text{ versus } H_A : \mu_1 < \mu_4$$
$$H_0 : \mu_2 \geq \mu_3 \text{ versus } H_A : \mu_2 < \mu_3$$
$$H_0 : \mu_2 \geq \mu_4 \text{ versus } H_A : \mu_2 < \mu_4$$
$$H_0 : \mu_3 \geq \mu_4 \text{ versus } H_A : \mu_3 < \mu_4$$
$$H_0 : \mu_3 = \mu_4 \text{ versus } H_A : \mu_3 \neq \mu_4$$

(g) [5 points] Using the p-values from (f), sort them and plot them versus their ranks. Add to this plot:

- the Bonferroni threshold (in red), assuming $\alpha^\star = 0.1$
- Holm's threshold (in blue), assuming $\alpha^\star = 0.1$
- the Benjamini-Hochberg threshold (in purple), assuming $\alpha^\star = 0.1$
- a legend distingishing these three lines

(h) [4 points] Based on the plot you constructed in (g) determine whether or not each of the seven null hypotheses in part (f) would be rejected based on each correction method. In particular, complete the following table by filling in "REJECT" or "ACCEPT", as appropriate, in each cell. Finally, draw a conclusion about the number of "Other Sizes" that maximizes average purchase size.

Hypothesis	Bonferroni	Holm	Benjamini-Hochberg
$H_0 : \mu_1 \geq \mu_2$			
$H_0 : \mu_1 \geq \mu_3$			
$H_0 : \mu_1 \geq \mu_4$			
$H_0 : \mu_2 \geq \mu_3$			
$H_0 : \mu_2 \geq \mu_4$			
$H_0 : \mu_3 \geq \mu_4$			
$H_0 : \mu_3 = \mu_4$			