Practice Problems Notebook

MAT241 Class

In this notebook, we'll work through several practice problems from our Textbook and MyOpenMath.

Introduction to Inference (MoM Week 5 HW)

Problem 1 Unsolved

Problem 1: For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

- In a survey, one hundred college students are asked how many hours per week they spend on the Internet.
- In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"
- In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.
- In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.
- In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

Problem 2 Unsolved

Problem 2: A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter, and the standard error for this estimate was 2.4%. Conduct a hypothesis test at the $\alpha = 0.05$ level of significance to determine whether a majority of US adult Twitter users get at least some news on Twitter.

Problem 5 Unsolved

Problem 5: A store randomly samples 603 shoppers over the course of a year and finds that 142 of them made their visit because of a coupon they'd received in the mail. Construct a 95% confidence interval for the fraction of all shoppers during the year whose visit was because of a coupon they'd received in the mail.

Problem 6 Unsolved

Problem 6: A tutoring company would like to understand if most students tend to improve their grades (or not) after they use their services. They sample 200 of the students who used their service in the past year and ask them if their grades have improved or declined from the previous year. Of the 200 sampled, 185 said that their grades had improved. Determine whether the data provides evidence to suggest that the companies tutoring services suggest that over 90% of customers report improved grades after using the tutoring services. Use the $\alpha = 0.1$ level of significance.

Problem 11 Unsolved

Problem 11: A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter (Pew, 2013). The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

Inference on One and Two Proportions (MoM Week 6 HW)

Problem 1 Unsolved

Problem 1: About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.

- The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.
- The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.
- A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.
- A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

Problem 3 Unsolved

Problem 3: Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school. Calculate a 90% confidence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

Problem 4 Unsolved

Problem 4: A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. Construct a 95% confidence interval for $(p_{\text{college grad}} - p_{\text{HS or less}})$, where p is the proportion of those who watch The Daily Show

Problem 5 Unsolved

Problem 5: Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period). Conduct a hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism. (Schmidt, 2011)

	Autism	Typical Development	Total
No vitamin	111	70	181
Vitamin	143	159	302
Total	254	229	483

Chi-Square Goodness of Fit and Independence (Multiple Proportions)

Problem 1 Unsolved

Problem 1 (Open Source Textbook): A professor using an open source introductory statistics book predicts that 20% of the students will purchase a hard copy of the book, 5% will print it out from the web, and 75% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 19 said they bought a hard copy of the book, 9 said they printed it out from the web, and 98 said they read it online. Conduct a test to determine whether the professor's predictions were inaccurate.

Problem 2 Unsolved

Problem 2 (Barking Deer): Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated Grassplot	Deciduous Forests	Other	Total
4	16	61	345	426

Conduct a test to determine whether barking deer prefer to forage in certain habitats over others.

Problem 3 Unsolved

Problem 3 (Full-Body Scan): The table below summarizes a data set we first encountered in Exercise 6.26 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Conduct a test to determine whether an individual's party affiliation and their support of full-body scans are independent of one another. It may be useful to first add on an extra column for row totals before proceeding with the computations.

	Republican	Democrat	Independent
Should	264	299	351
Should Not	38	55	77
Don't Know / No Answer	16	15	22
Total	318	369	450

Problem 4 Unsolved

Problem 4 (Offshore Drilling): The table below summarizes a data set we first encountered in Exercise 6.23 that examines the responses of a random sample of college graduates and non-graduates on the topic of oil drilling. Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

	College Grad	Not College Grad
Support Offshore Drilling	154	132
Oppose Offshore Drilling	180	126
Do not know	104	131
Total	438	389

Inference on One and Two Means

Problem 1 Unsolved

Problem 1 (Sleep Habits of New Yorkers): New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

Problem 2 Unsolved

Problem 2 (Diamond Pricing): Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.

Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

	0.99 carats	1 carat
Mean	\$44.51	\$56.81
SD	\$13.32	\$16.13
n	23	23

Problem 3 Unsolved

Problem 3 (Fuel Efficiency of Cars): The table provides summary statistics on highway fuel economy of 52 cars. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.

	Automatic Highway mpg	Manual Highway mpg
Mean	22.92	27.88
SD	5.29	5.01
n	26	26

Problem 4 Unsolved

Problem 4 (Forest Management): Forest rangers wanted to better understand the rate of growth for younger trees in the park. They took measurements of a random sample of 50 young trees in 2009 and again measured those same trees in 2019. The data below summarize their measurements, where the heights are in feet:

	2009	2019	Differences
$\bar{x}$	12.0	24.5	12.5
s	3.5	9.5	7.2
n	50	50	50

Construct a 99% confidence interval for the average growth of (what had been) younger trees in the park over 2009-2019.

Sample Size Problems

The following problems ask us to estimate a required sample size. As a reminder, when estimating sample sizes, we need to round *up* to the next whole number to ensure that the resulting sample size is large enough to achieve our desired level of confidence and our desired margin of error.

Problem 1 Unsolved

Problem 1 (Legalization of Marijuana): As discussed in Exercise 6.10, the General Social Survey reported a sample where about 61% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey?

Problem 2 Unsolved

Problem 2 (Spring Break Spending): A marketing research firm wants to estimate the average amount a student spends during the Spring break. They want to determine it to within \$120 with 90% confidence. The first does some research which allows it to roughly say that Spring Break expenditure ranges from \$100 to \$1700 per student. They use the approximation $\frac{\text{range}}{4}$ for σ. How many students should they sample?

Comparing Many Means with Analysis of Variance (ANOVA)

In this section, we explore the use of Analysis of Variance to compare multiple (more than two) group means.

Problem 1 Unsolved

Problem 1 (Taylor Swift Songs): Take a look at the Taylor Albums data frame from the `{taylor}` R package. Compare one of the song metrics (`duration`, `danceability`, `speechiness`, `loudness`, etc.) across Taylor's released albums.

```r
library(taylor)

taylors_version <- taylor_album_songs %>%
  filter(str_detect(album_name, "Taylor's Version"))

taylors_version %>%
  select(-lyrics) %>%
  head() %>%
  kable() %>%
  kable_styling(bootstrap_options = c("hover", "striped"))
```

album_name	ep	album_release	track_number	track_name
Fearless (Taylor's Version)	FALSE	2021-04-09	1	Fearless (Taylor's Version)

Fearless (Taylor's Version)	FALSE	2021-04-09	2	Fifteen (Taylor's Version)
Fearless (Taylor's Version)	FALSE	2021-04-09	3	Love Story (Taylor's Version)
Fearless (Taylor's Version)	FALSE	2021-04-09	4	Hey Stephen (Taylor's Version)
Fearless (Taylor's Version)	FALSE	2021-04-09	5	White Horse (Taylor's Version)
Fearless (Taylor's Version)	FALSE	2021-04-09	6	You Belong With Me (Taylor's Versi

Problem 2 Unsolved

Problem 2 (Penguin Physiology): Take a look at the `penguins` data frame from the `{palmerpenguins}` R package. Compare one of the body measurements (`body_mass_g`, `flipper_length_mm`, `bill_depth_mm`, etc.) across the different penguin species, islands, or observation years.

```r
library(palmerpenguins)
```

```

Attaching package: 'palmerpenguins'


The following objects are masked from 'package:datasets':

    penguins, penguins_raw
```

```r
penguins %>%
  head() %>%
  kable() %>%
  kable_styling(bootstrap_options = c("hover", "striped"))
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	ye
Adelie	Torgersen	39.1	18.7	181	3750	male	20
Adelie	Torgersen	39.5	17.4	186	3800	female	20
Adelie	Torgersen	40.3	18.0	195	3250	female	20
Adelie	Torgersen	NA	NA	NA	NA	NA	20
Adelie	Torgersen	36.7	19.3	193	3450	female	20
Adelie	Torgersen	39.3	20.6	190	3650	male	20

Problem 3 Unsolved

Problem 3 (NFL Rush Yards by Direction): Explore the `rush_21` data frame, containing data on running plays from the 2021 NFL season. Conduct a test to determine whether the average yards gained on a rushing play is associated with the direction (*left*, *middle*, *right*) of the running play.

```r
#Running this code may crash a free-tiered Posit Cloud workspace
library(nflfastR)

pbp_21 <- load_pbp(2021)

rush_21 <- pbp_21 %>%
  filter(play_type == "run")

rush_21 %>%
  select(rusher, down, qtr, rushing_yards, run_location) %>%
  head() %>%
  kable() %>%
  kable_styling(bootstrap_options = c("hover", "striped"))
```