IDS 702 HW 2

Eric Ortega Rodriguez

Exercise 1

Part A: Likelihood function

Normal PDF : $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

↳ multiplying Probability densities •

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2}}$$

$$\boxed{L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}}$$

Part B: Log-Likelihood function

$$\ell\mu = n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\boxed{\ell(\mu) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Part C: Maximum Likelihood Estimator of μ

$$\frac{d}{d\mu} l(\mu) = \frac{d}{d\mu}\left(-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$\hookrightarrow \frac{d}{d\mu} l(\mu) = \frac{1}{2}\cdot 2\sum_{i=1}^{n}(x_i-\mu) = \sum_{i=1}^{n}(x_i-\mu)$$

$$\hookrightarrow n\bar{x} - n\mu = 0$$

$$\hookrightarrow \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\hookrightarrow$ Maximum Likelihood Estimator of μ is $\hat{\mu} = \bar{x}$ (Sample mean)

Part D: Is the MLE of $\hat{\mu}$ a unbiased estimator of μ?

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i)$$

$$\hookrightarrow E\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$\hookrightarrow E\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \mu$$

Therefore, since $E(\hat{\mu}) = \mu$, we can conclude $\hat{\mu}$ is a unbiased estimator of μ.

Exercise 2

You are required to show the code you use to complete each part of exercises 2-5. You must also write your narrative answers below the code.

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
library(tidymodels)
```

```
-- Attaching packages ---------------------------------------- tidymodels 1.2.0 --
v broom        1.0.6     v rsample      1.2.1
v dials        1.3.0     v tune         1.2.1
v infer        1.0.7     v workflows    1.1.4
v modeldata    1.4.0     v workflowsets 1.1.0
v parsnip      1.2.1     v yardstick    1.3.1
v recipes      1.1.0
-- Conflicts ------------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Search for functions across packages at https://www.tidymodels.org/find/
```

```r
library(openintro)
```

```
Loading required package: airports
Loading required package: cherryblossom
Loading required package: usdata

Attaching package: 'openintro'

The following object is masked from 'package:modeldata':
```

```
ames
```

```
data("births14")
```

a. The total number of observations is 1000 and there is a total of 13 variables. The data set contain information about a individual's birth record from the United States. The observational unit is a single birth.

```
dim(births14)
```

```
[1] 1000   13
```

b. The variable which contain missing values are fage (114 missing), visits(56 missing), gained(42 missing), and habit(19 missing).

```
colSums(is.na(births14))
```

```
          fage           mage         mature          weeks         premie 
           114              0              0              0              0 
        visits         gained         weight lowbirthweight            sex 
            56             42              0              0              0 
         habit        marital       whitemom 
            19              0              0 
```

c. The count of of low birth weight babies is 81 babies (indicating a 8.1%). As for not low birth rates, there are 919 babies(indicating a 91.9%).

```
# retrieving the count of low birth rate and not low weight
table(births14$lowbirthweight)
```

```
low not low 
 81     919 
```

```
# in order to calculate %, we must mutiply by 100
prop.table(table(births14$lowbirthweight)) * 100
```

```
low not low 
8.1    91.9 
```

Exercise 3

a.

Null Hypothesis: There would be no difference among the median birth weight of babies born to mothers who are smokers compared to those who are not.

Alternative Hypothesis: There is a difference among the median birth weight of babies born to mothers who are smokers compared to those who are not.

b. In this case, the *observed difference* in the median birth weight between babies born to mothers who smoke and those who do not.

```
births14 <- births14 %>% drop_na(habit)

observed_statistic <- births14 |>
  specify(response = weight,
    explanatory = habit) |>

calculate(stat = "diff in medians", order = c("nonsmoker", "smoker"))

observed_statistic
```

```
Response: weight (numeric)
Explanatory: habit (factor)
# A tibble: 1 x 1
   stat
  <dbl>
1 0.320
```

c. The null hypothesis states that there is no difference in the median birth weight of babies born to mothers who smoke compared to those who do not smoke. After simulating the null distribution and performing the hypothesis test, we found that our p-value (0.02) is below the significance level of 0.05. Therefore, we reject the null hypothesis. This indicates that there is sufficient evidence to conclude that the median birth weight is significantly different when the mother smokes.

```
set.seed(150) #random number
# Simulating the null distribution
null_dist <- births14 %>%
  specify(formula = weight ~ habit) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
```

```
  calculate(stat = "diff in medians", order = c("nonsmoker", "smoker"))

# Getting out p-value
p_value <- null_dist %>%
  get_p_value(obs_stat = observed_statistic, direction = "both")

p_value
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1    0.02
```

d. After calculating our bootstrap confidence interval, we see that zero lies within the interval. Given this and our p-value, we fail to reject the null hypothesis, which states that there is no difference in median birth weights between the two groups. The confidence interval indicates that there is insufficient evidence to prove that a mother's smoking status is associated with a difference in median birth weight.

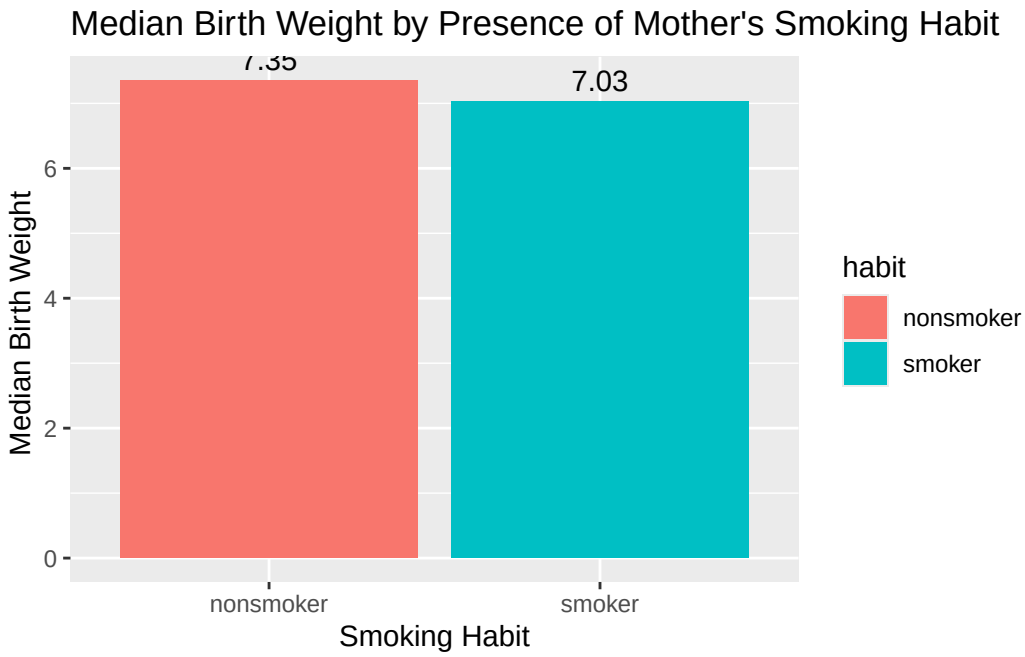
```
set.seed(150)
  boot_dist <- births14 |>
  specify(response = weight,explanatory = habit) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in medians",
  order=c("nonsmoker", "smoker"))

get_confidence_interval(boot_dist, level=0.95,
type="percentile")
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1  -0.0351     0.64
```

e. Plot illustrating the median birth weight when the mother smokes vs does not smoke can be seen below.

```
median_birthweight <- births14 %>%
  group_by(habit) %>%
  summarize(median_weight = median(weight), .groups = 'drop')
```

```
ggplot(median_birthweight, aes(x = habit, y = median_weight, fill = habit)) +
  geom_bar(stat = "identity", position = "dodge") + geom_text(aes(label = round(median_weight
    x = "Smoking Habit", y = "Median Birth Weight ")
```



Exercise 4

a.Table showing premature birth and low birth weight can be seen below

```
maturity_table <- births14 %>%
    select(premie, lowbirthweight) %>%
    table()

maturity_table
```

```
           lowbirthweight
premie      low not low
  full term  27     833
  premie     51      70
```

	Low birth weight	Not low birth weight
Not premature	27	833
Premature	51	70

b. The necessary conditions to conduct a Chi-square test of independence are the following:
1. ***Variables must be categorical*** - data must be in categories, not continuous.
2. ***Independence*** - each observation should be independent of others.
3. ***Expected Frequency*** - The expected frequency in each category should be at least five. This ensures the Chi-Square approximation to the distribution is valid.

Given these conditions, we can determine that the conditions are ***met*** and we can conduct the Chi-square test. Each row is independent from the others since it is a person's birth, data is categorical since it divided by categories, and expected frequencies are at least five.

c.
Null Hypothesis: There is no relationship between premature birth and low birth weight (independent).

Alternative Hypothesis: There is a relationship between premature birth and low birth weight (not independent).

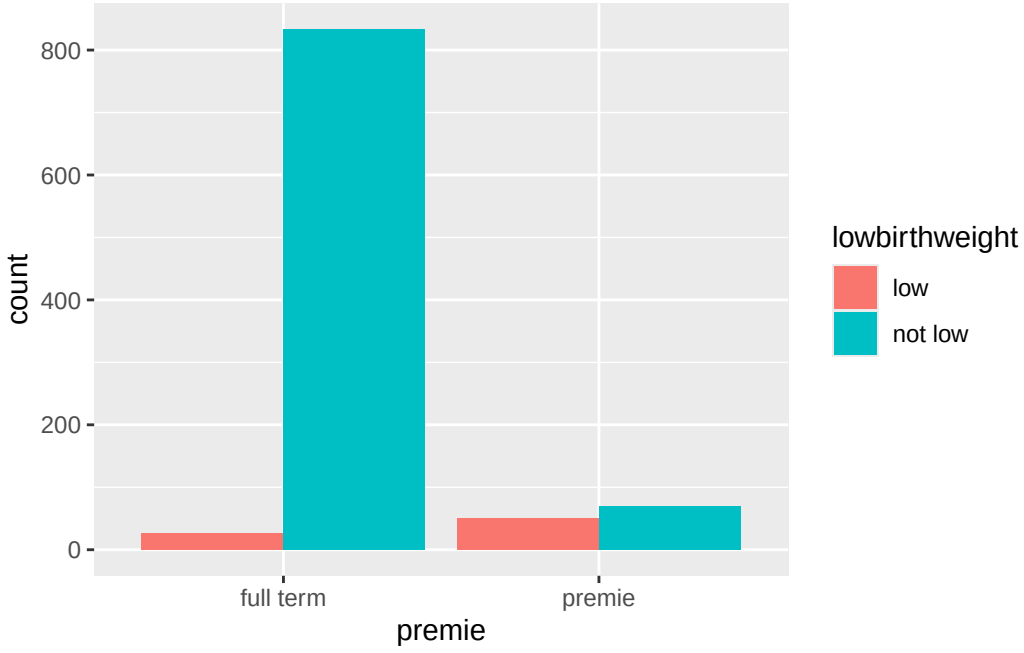
The calculation of the p-value using R can be seen below. The p-value is less than 2.2e-16. This small p-value indicates that we reject the null hypothesis.Thus, indicating that there is a statistically significant correlation between premature birth and low birth weight. More specifically, given the context of the problem the likelihood of a baby being born with low birth weight is correlated with whether the birth was premature.

```
# Chi-square test
chi_testing <- chisq.test(maturity_table)

# Results
chi_testing
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  maturity_table
X-squared = 215.25, df = 1, p-value < 2.2e-16
```

d. The plot highlighting the relationship between premature births and low birth rates can be seen below.

```r
ggplot(births14) +
  aes(x = premie, fill = lowbirthweight) + geom_bar(position = "dodge")
```



Exercise 5

a. Creating a new variable can be seen below:

```r
births14 <- births14 %>%
  mutate(visits_category = case_when(
    visits <= 10 ~ "10 or fewer",
    visits >= 11 & visits <= 15 ~ "11-15",
    visits > 15 ~ "more than 15",
    TRUE ~ NA_character_
  ))
```

b. The mean by number of visits for the following:
10 or fewer visits: 27.98
11-15 visits: 28.49
More than 15 visits: 28.83

```r
mean_category <- births14 %>% group_by(visits_category) %>% summarize(mean_mother_age = mean

mean_category
```

```
# A tibble: 4 x 2
  visits_category mean_mother_age
  <chr>                     <dbl>
1 10 or fewer                28.0
2 11-15                      28.5
3 more than 15               28.8
4 <NA>                       29.8
```

c. The appropriate test to assess this research question would be the Analysis of Variance (ANOVA) test. The conditions for a ANOVA test are fulfilled since both conditions are met. These conditions are that there is independence, normality, and homogeneity of variance, all of which are satisfied.

Null Hypothesis: The mean mother's age is the identical across all three visit categories.

Alternative Hypothesis: At least one of the three means is different than the rest.

d. Conduction of the ANOVA test can be seen below. Our P- value is approximately 0.347. Due to this, we cannot reject our null hypothesis. In the context of our data, this indicated that there is no statistically significant difference in the mean mother's age among our different prenatal visit categories (10 or fewer, 11 to 15, and more than 15.

```r
births14$visits_category <- factor(births14$visits_category, levels = c("10 or fewer", "11-15

# Remove rows with missing data
births14_cleaned <- births14 %>%
  filter(!is.na(visits_category), !is.na(mage))

# ANOVA test
anova_result <- aov(mage ~ visits_category, data = births14_cleaned)

# summary
summary(anova_result)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
visits_category   2     81   40.45   1.236  0.291
Residuals       923  30214   32.73
```

e. A box plot highlighting showing relationship between hospital visits during pregnancy and mean mother's age can be seen below.

```
ggplot(births14, aes(x = visits_category, y = mage, fill = visits_category)) +
  geom_boxplot(outlier.alpha = 0.1, outlier.size = 0.5) +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 3, color = "black", fill = "bla
  labs(
    title = "Mother's Age and Number of Prenatal Visits",
    x = "Prenatal Visit Categories",
    y = "Mother's Age"
  )
```