# IDS 702 HW 2

**Instructions**:

Use this template to complete your assignment. When you click "Render," you should get a PDF document that contains both your answers and code. **You must show your work/justify your answers to receive credit.**

Submit your rendered PDF file on Gradescope. Remember to render frequently, as this will help you to catch errors in your code before the last minute.

**Add your name in the Author section in the**

## Exercise 1

For this question, you are welcome to write out your answers on paper and include a picture below (Insert > Figure/Image in the toolbar)

Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$, and recall the Normal PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

### 1a

Ques: Write the likelihood function $L(\mu)$

Ans:

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

### 1b

Ques: Write the log-likelihood function $l(\mu)$

Ans: Applying log to part a leads to the log liklihood shown below:

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

### 1c

Ques: Find the maximum likelihood estimator of $\mu$

Ans: Differentiate part b and set it to 0 and solve for MLE estimation of the mean.

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

### 1d

Ques: Determine if the MLE $\hat{\mu}$ is an unbiased estimator of $\mu$

Ans:

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

Questions 2-5 use the **births14** dataset in the openintro package. Codebook

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.

```
#installing all the required pacages here for readability
library(tidyverse)
library(tidymodels)
library(openintro)
data("births14")
```

## Exercise 2

(6 points) First, do some initial data exploration:

### 2a

Ques: How many observations and how many variables are in the dataset? What does each row represent (i.e., what is the observational unit?)

Ans:

```
dim(births14)
```

```
[1] 1000   13
```

There are 1000 observations and 13 variables in the dataset. Each observation represents one baby/pregnancy.

### 2b

Ques: Which variables, if any, contain missing data?

Ans:

```
missing_data <- sapply(births14, function(x) sum(is.na(x)))
missing_vars <- names(missing_data[missing_data > 0])
missing_vars
```

```
[1] "fage"   "visits" "gained" "habit"
```

The following variables contain missing data:

- *fage* - Fathers age in years
- *visits* - Number of hospital visits during pregnancy
- *gained* - weight gained by mother during pregnancy (in pounds)
- *habit* - Status of the mother as a `smoker` or `nonsmoker`

### 2c

Ques: What is the count and percentage of low birth weight vs not low birth weight babies?

Ans:

```
birthweight_counts <- table(births14$lowbirthweight)
birthweight_counts
```

```
low not low
 81     919
```

```
birthweight_percentages <- prop.table(birthweight_counts) * 100
birthweight_percentages
```

```
low not low
8.1    91.9
```

The count of low birth weight vs not low birth weight babies is **81** and **919**.

The percentage of low birth weight vs not low birth weight babies is **8.1%** and **91.9%**.

## Exercise 3

(10 points) Is the median birth weight different when the mother smokes? Conduct a hypothesis test to investigate:

### 3a

Ques: Write the null and alternative hypotheses.

Ans:

Null hypothesis is

The **median weight** for babies from mothers who smoke **is the same** as for babies from mothers who don't smoke.

$H_0 : \eta_{smoker} = \eta_{nonsmoker}$ or $H_0 : \eta_{nonsmoker} - \eta_{smoker} = 0$

Alternative hypothesis can be:

The Alternative hypothesis is that **median weight** for babies from mothers who smoke **is not the same** as for babies from mothers who don't smoke.

$H_1 : \eta_{nonsmoker} \neq \eta_{smoker}$ or $\eta_{nonsmoker} - \eta_{smoker} \neq 0$

### 3b

Ques: Calculate the observed statistic

Ans:

Since we are primarily trying to compare the median birth weights with respect to habit, we have to remove rows that have missing values for habit.

```
# let remove all rows which have missing data for habit
births14_clean_ex3 <- births14 |> filter(!is.na(habit))

obs_stat <- births14_clean_ex3 |>
  specify(response = weight, #response=numeric
          explanatory = habit) |> #exp=categorical
          calculate(stat = "diff in medians",
          order = c("smoker","nonsmoker"))

print(obs_stat)
```

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

The observed statistic is **-0.32**

### 3c

Ques: Simulate the null distribution and calculate the p-value. Interpret the p-value in the context of the problem.

Ans:

```r
set.seed(1)

null_dist <- births14_clean_ex3  |>
  specify(response = weight,
          explanatory = habit) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in medians",
            order = c("smoker","nonsmoker"))

# get p-value
p_value <- null_dist |>
  get_p_value(obs_stat, direction = "two-sided")

print(p_value)
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.034
```

The calculated p_value for this problem is **0.034** which is lower than commonly accepted significance level of ~0.05. Therefore, we can **reject the null hypothesis** that the median weight of babies is same for mother who smokes vs mothers who don't smoke.

### 3d

Ques: Generate a bootstrap confidence interval and interpret the interval.

Ans:

```r
boot_dist <- births14_clean_ex3 |>
  specify(response = weight,
          explanatory = habit) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in medians",
            order=c("smoker", "nonsmoker"))

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

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1    -0.62   0.0500
```

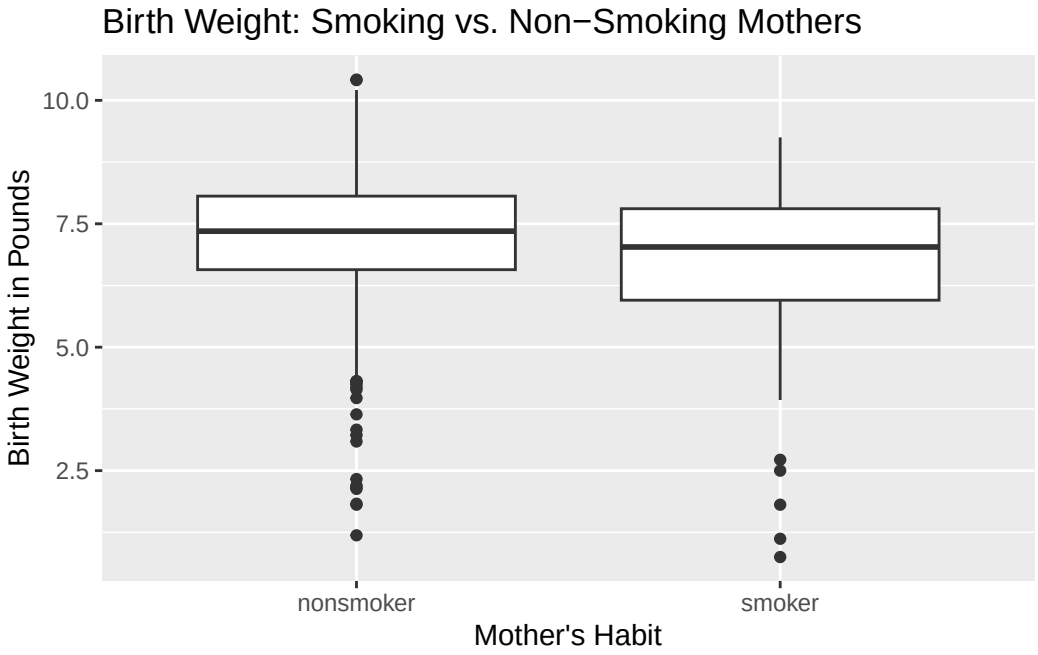
The 95% confidence interval is between **-0.62 and 0.05**. This means that, based on the bootstrap samples, we are 95% confident that the **true (or population level) difference in medians** of birth weights between mothers who smoke and mothers who don't smoke lies between -0.62 and 0.05. In other words, in 95% of repeated samples from the population, the difference in medians will fall within this range.

### 3e

Ques: Generate a plot to illustrate the median birth weight when the mother smokes vs doesn't smoke.

Ans:

```
births14_clean_ex3 |>
  ggplot(aes(x = habit, y = weight)) +
  geom_boxplot() +
  labs(title = "Birth Weight: Smoking vs. Non-Smoking Mothers",
       x = "Mother's Habit",
       y = "Birth Weight in Pounds")
```




**Visually we can clearly see that for mothers who smoke, their babies weigh less**. This supports our earlier finding that there may be something special going on between smoking habit and birth wright (rejecting the null hypothesis).

## Exercise 4

(8 points) Is there a relationship between premature birth and low birth weight? Conduct a hypothesis test to investigate:

### 4a

Ques: Create a 2x2 table to show the counts of the two relevant variables

Ans:

```
contingency_table <- table(births14$premie, births14$lowbirthweight)
contingency_table
```

```
          low not low
full term  28     848
```

```
premie     53      71
```

| | Low Birth Weight | Not Low Birth Weight |
|---|---|---|
| Not Premature | 28 | 848 |
| Premature | 53 | 71 |

### 4b

Ques: What are the necessary conditions to conduct a Chi-square test of independence, and are those conditions met?

Ans:

```
chi_test <- chisq.test(contingency_table)
chi_test$expected
```

```
             low not low
full term 70.956 805.044
premie    10.044 113.956
```

The following conditions should be met to conduct a Chi-square test of independence:

1. **Categorical data**: In this case, both `premie` (premature vs. full-term) and `lowbirthweight` (low vs. not low) are categorical, which satisfies this condition.
2. **Randomness and independence**: In this case, the data is randomly sampled from the population, and each observation is independent from each other so it is satisfied.
3. **Expected Frequency**: For expected frequency, we run the above code. As in each shell, we have at least value of 5 for each shell, which satisfies this condition.
4. **Sample Size**: Our sample size is 1000, which is sufficient

### 4c

Ques: Conduct the appropriate test. Write the null and alternative hypotheses, calculate and report the p-value, and interpret it in the context of the problem.

Ans:

- **Null Hypothesis ($H_0$):** There is no relationship between premature birth and low birth weight (i.e., premature birth and low birth weight are independent).

- **Alternative Hypothesis ($H_1$):** There is a relationship between premature birth and low birth weight (i.e., premature birth and low birth weight are not independent).

```
chi_test <- chisq.test(contingency_table)
chi_test
```

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

data:  contingency_table
X-squared = 222.92, df = 1, p-value < 2.2e-16
```

As the p value is smaller than 0.05 (assuming a 5% significance level), we **reject the null hypothesis**, indicating that there is a statistically significant relationship between premature birth and low birth weight.

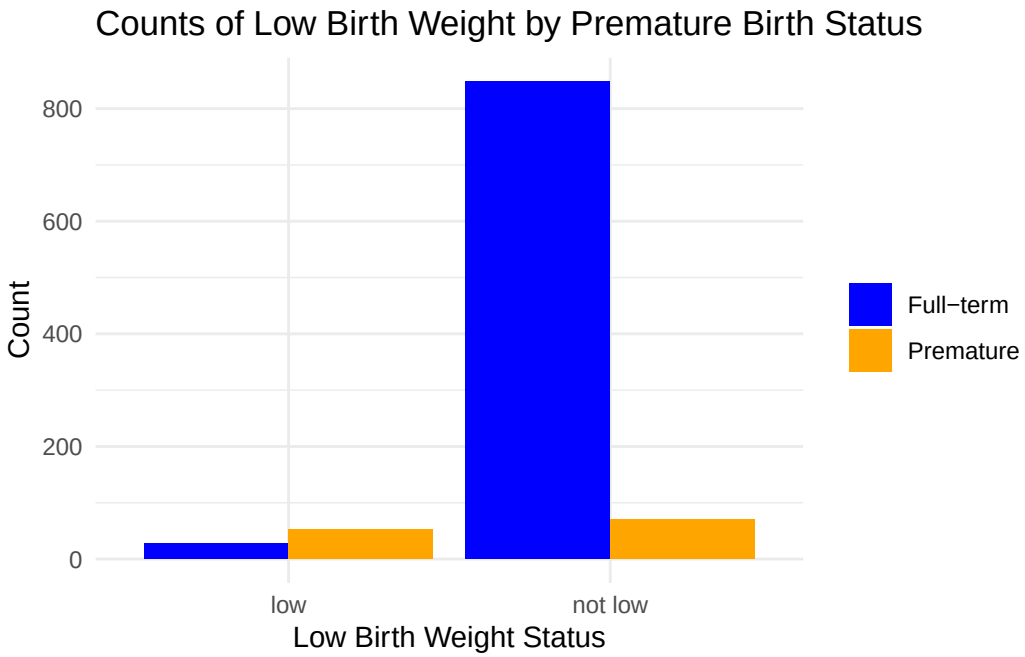
### 4d

Ques: Generate a plot to illustrate the relationship between premature birth and low birth weight.

Ans:

```
data_plot <- as.data.frame(contingency_table)
colnames(data_plot) <- c("Premature", "LowBirthWeight", "Count")

ggplot(data_plot, aes(x = LowBirthWeight, y = Count, fill = Premature)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Counts of Low Birth Weight by Premature Birth Status",
       x = "Low Birth Weight Status",
       y = "Count") +
  scale_fill_manual(values = c("blue", "orange"),
                    labels = c("Full-term", "Premature")) +
  theme_minimal() +
  theme(legend.title = element_blank())
```




## Exercise 5

(10 points) Is there a relationship between hospital visits during pregnancy and mean mother's age? Conduct a hypothesis test to investigate:

### 5a

Ques: Create a new variable that categorizes the number of visits as: 10 or fewer, 11-15, more than 15

Ans:

```
## Creating a new DF with nulls dropped from the columns of interest
births14_clean_ex5 <- births14 %>% drop_na(visits, mage)

## creating a new categorical column for the number of visits
births14_clean_ex5$visits_cat <- cut(births14_clean_ex5$visits,
                                     breaks = c(-Inf, 10, 15, Inf),
                                     labels = c("10 or fewer", "11-15", "more than 15"),
                                     right = TRUE)
```

```r
## Validation of Prior step, Optional and not required
visits_cat_validation <- births14_clean_ex5 %>%
  group_by(visits_cat) %>%
  summarise(
    min_visits = min(visits, na.rm = TRUE),
    max_visits = max(visits, na.rm = TRUE),
    mean_visits = mean(visits, na.rm = TRUE)
  )
visits_cat_validation
```

```
# A tibble: 3 x 4
  visits_cat   min_visits max_visits mean_visits
  <fct>             <dbl>      <dbl>       <dbl>
1 10 or fewer           0         10        7.63
2 11-15                11         15       12.7
3 more than 15         16         30       18.6
```

### 5b

Ques: Calculate the mean of mother's age for each of the visit categories that you created in part a.

Ans:

```r
#calculating mean of mothers age based on prior catagorization
mean_mothers_age_ex5 <- births14_clean_ex5 %>%
  group_by(visits_cat) %>%
  summarise(mean_age = mean(mage, na.rm = TRUE))
mean_mothers_age_ex5
```

```
# A tibble: 3 x 2
  visits_cat   mean_age
  <fct>           <dbl>
1 10 or fewer      28.1
2 11-15            28.5
3 more than 15     28.9
```

The mean age of mothers:

1. Who had less than 10 Visits: 28.06 years
2. Between 11-15 visits: 28.52 years

3. More than 15 visits: 28.86 years

The mean age is similar in all the categories

### 5c

Ques: What is the appropriate test to assess the research question? Write the null and alternative hypotheses.

Ans: The appropriate test to assess this research question is an **ANOVA** test since we are trying to study the relationship between a continuous variable and a categorical variable with more than 2 levels.

**Null Hypothesis**: $H_0$ - There is no relationship between the number of hospital visits and mothers age during pregnancy

i.e. the mean age of the mothers from the three groups will be equal

$$\mu_0 = \mu_1 = \mu_2$$

**Alternative Hypothesis**: $H_1$ - There is a relationship between the number of hospital visits and the mothers age during pregnancy

i.e at least 1 pair of the means of the mothers ages will not be equal

$$\mu_0 \neq \mu_1 \; or \; \mu_1 \neq \mu_2 \; or \; \mu_2 \neq \mu_0$$

### 5d

Ques: Conduct the appropriate test; calculate and report the p-value and interpret it in the context of the problem.

Ans:

(I am mentioning 2 methods to perform ANOVA since the student may use any method as long as the final p-value and interpretation is correct)

Option-1 :

```
oneway.test(mage ~ visits_cat, data=births14_clean_ex5, var.equal = TRUE)
```

```
    One-way analysis of means

data:  mage and visits_cat
F = 1.0602, num df = 2, denom df = 941, p-value = 0.3468
```

Option-2:

```r
anova_ex5 <- aov(mage ~ visits_cat, data= births14_clean_ex5)
summary(anova_ex5)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
visits_cat    2     70   34.87    1.06  0.347
Residuals   941  30945   32.89
```

The p-value is **0.347**

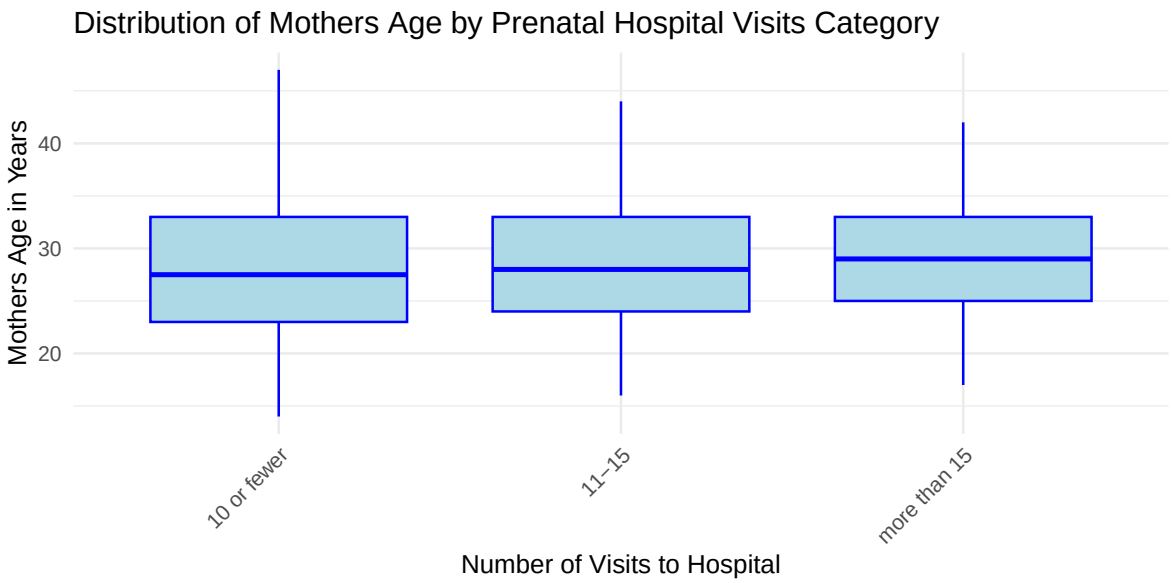
Since the p-value is grater than our general significance level of 0.05, we **fail to reject the null hypothesis** i.e. There is no relation between the age of the mother and the number of hospital visits during pregnancy

### 5e

Ques: Generate a plot to illustrate the relationship between mother's age and the categorized visit variable.

Ans:

```r
ggplot(births14_clean_ex5, aes(x = visits_cat, y = mage)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  labs(title = "Distribution of Mothers Age by Prenatal Hospital Visits Category",
       x = "Number of Visits to Hospital",
       y = "Mothers Age in Years") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




As visible in the graph, the distribution of mothers age is similar in all 3 categories, which confirms our earlier result that there is no relationship between the number of hospital visits and the mothers age during pregnancy