QSCI 381 AB - AUT 22
Homework #7- Vinsensius

The PDF file of the code is attached at the end of the HW

Question 1

In this question we will be using the FlightResponse dataset, downloadable from canvas.
A 1994 study collected data on the effects of air traffic on the behavior of the Pacific Brant (a small migratory goose). The data represent the flight response to helicopter "overflights" to see what the relationship between the proximity of a flight, both lateral and altitudinal, would be to the propensity of the Brant to flee the area. Each case represents a flock of Brant that has been observed during one overflight in the study.
The data contains the following columns

- Obs.ID: Observation (flock) identifier number
- AltCat: Altitude category as an identifier of whether the overflight was at a low or high altitude
- Flight: a binary (0,1) variable indicating whether the flock did (1) or didn't (0) take flight

We will examine differences between high and low altitude overflights

```
low.df <- FlightResponse[FlightResponse$AltCat == "low",]
high.df <- FlightResponse[FlightResponse$AltCat == "high",]
```

Use low.df and high.df to answer the following questions

1a

For low and high overflights, calculate and present the

1. Sample size
2. Number of overflights that lead to Brant fleeing the area
3. Proportion of overflights that lead to Brant fleeing the area

Round your answers to 3 decimal places (3 pts)

	low	high
Sample Size	190	76
Number of overflight leading to fleeing Brant	105	59
Proportion of overflight that leads to fleeing Brant	0.553	0.776

1b

I claim that for Brant, **the probability of disturbance is no different between high overflights and low overflights**.
For this claim write out the null and alternate hypothesis both as statements, and mathematically (i.e. using $=, >, <, \neq, \leq, \geq$), using ph and pl to represent the probability of disturbance for high and low overflights, respectively. In your answer note which hypothesis represents the claim (3 pts)

H_0	$ph = pl$
H_A	$ph \neq pl$

The null hypothesis represents the claim.

1c

Evaluate the conditions for applying a two-sample z-test to test the hypotheses given in (1b). Include the calculation steps, and whether criteria are met, and finally whether this dataset can be analyzed using a two-sample z-test (4 pts)

Let data 1 be the low overflight and data 2 be the high overflight. Then, $\hat{p}_{pooled} = 0.617$.

Criteria	Result
$n_1\hat{p}_{pooled} \geq 10$	met, since the value is 117.
$n_1(1-\hat{p}_{pooled}) \geq 10$	met, since the value is 72
$n_2\hat{p}_{pooled} \geq 10$	met, since the value is 46
$n_2(1-\hat{p}_{pooled}) \geq 10$	met, since the value is 29

Thus, the dataset can be analyzed using a two-sample z-test.

1d

Perform a two sample z-test of the hypotheses given in (1b) for a significance level of $\alpha = 0.01$. Show your working, and in your answer present

- Whether the test is left-, right-, or two-tailed
- The test statistic
- The standard error of the test-statistic
- The standardized test statistic
- The critical region(s) for the given significance level
- The associated p-value

Round all numeric values to 3 decimal places (10 pts)

Type of test	two-tailed test
Test stat $(\hat{p}_1 - \hat{p}_2)$	-0.224
SE	0.066
z	-3.390
z_{crit}	$\pm$2.576
p-val	0.001

1e

Using your answers in (1d) make a decision to reject or fail to reject the null hypothesis. State your answer with reference to the original claim and how overflights affect Brant (3 pts)

Based on (1d), we will reject the null hypothesis since the z that we got is within the rejection region. Thus, there is possibility of difference in the probability of disturbance between high and low overflights, instead being equal as it claimed to be.

Question 2

In this example we will be using the North Carolina births dataset. This data consists of observations from a random selection of births in North Carolina in 2001.
We will be examining birth weights and whether they differ between mothers who were smokers or not. Firstly, we will create two subsets, one consisting of birth records for non-smokers, and one for smokers and we will exclude premature births

```
births <- read.csv(file="north_carolina_births.csv")
smoker <- births[births$habit == "smoker" & births$premie == "full term",]
nonsmoker <- births[births$habit == "nonsmoker" & births$premie == "full term",]
```

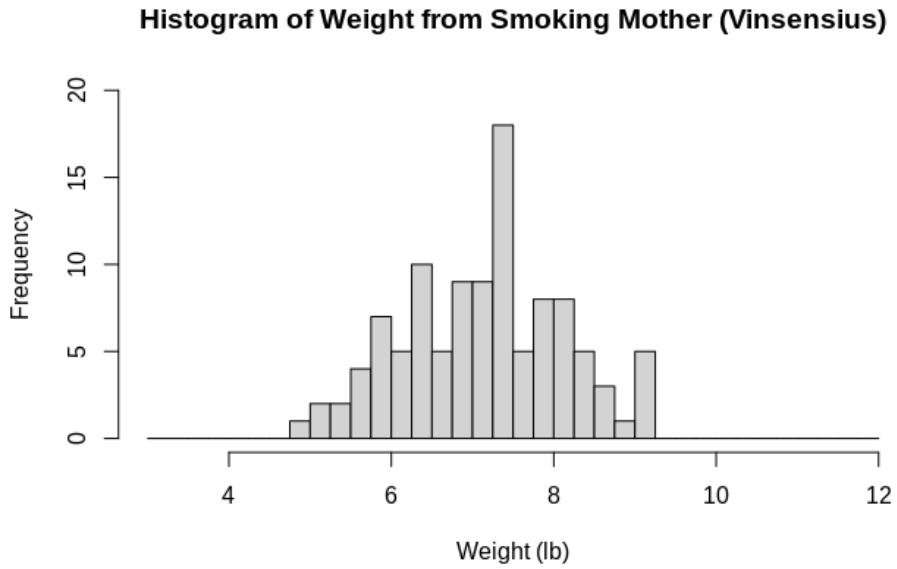
Use the smoker and nonsmoker datasets throughout the rest of the question.

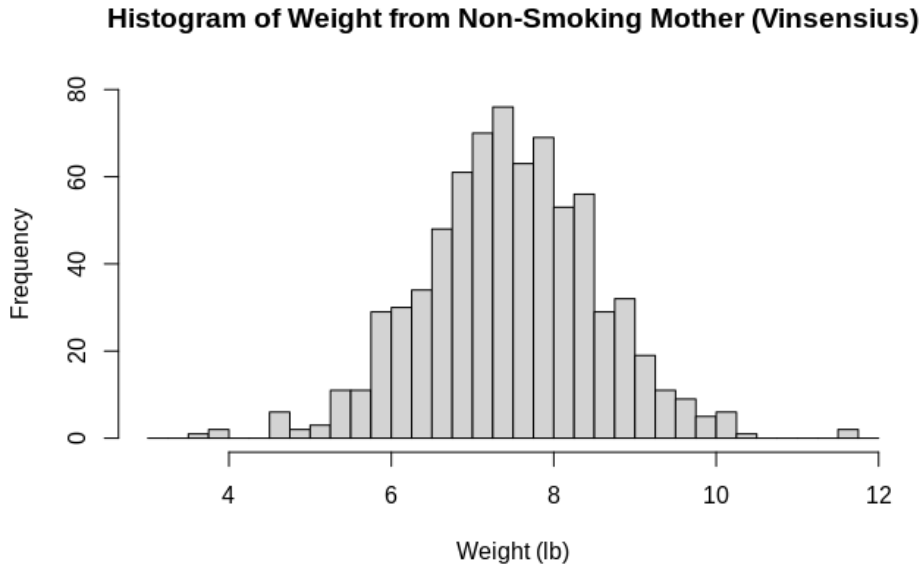
2a

Plot histograms of the birth weights (in pounds; weight) for children from smoking and non-smoking mothers, labelling the axes, and giving each plot a title based on smoking status. Use

... breaks = seq(from=3, to=12, by=0.25) ...

in your hist() call for both.
Based on the histograms, do birth weights appear to be normally distributed? (4 pts)





Both histograms look to be normally distributed. However, there are few outliers on the non-smoking mother histogram.

2b

For both smoking and non-smoking datasets present the sample size, mean, and standard deviation of birth weights. Round your answers to 1 decimal place (3 pts).

	Smoker	Non-smoker
Sample Size	107	739
Sample mean	7.2	7.5
Sample std	1.0	1.1

2c

Perform an F-test on the variance in birth weight between non-smoking mothers and smoking mothers, testing the null hypothesis:
H0: variance in birth weights are equal for babies born to smoking mothers, versus non-smoking mothers; $\sigma_1^2 = \sigma_2^2$
We will test this hypothesis at the 10% significance level, $\alpha = 0.1$. Use the non-smoker variance as the numerator, and the smoker variance as the denominator. In your answer show your working, and present numeric values for

- The F test statistic
- The degrees of freedom of the F distribution
- The critical region for rejection of the null hypothesis given $\alpha = 0.1$

Round your answers to 2 decimal places (8 pts)

F	1.24
df numerator (non-smoker)	738
df denominator (smoker)	106
F_{crit}	1.29

2d

Using your answer in (2c) is there evidence at the $\alpha = 0.1$ level to reject the null hypothesis that variance in birth weight is equal between babies born to smoking mothers, versus non-smoking mothers (2 pts)

There is no evidence to reject the null hypothesis since the F is not in the rejection region. Thus, the claim that the variances are equal can be assumed for further analysis.

2e

Researchers claim that birth weights are lower in smoking mothers than non-smoking mothers. Given this claim, write out the null and alternate hypotheses, and indicate which represents the claim. (3 pts)

H_0	$\mu_{smoke} \geq \mu_{nonsmoke} = \mu_{smoke} - \mu_{nonsmoke} \geq 0$
H_A	$\mu_{smoke} < \mu_{nonsmoke} = \mu_{smoke} - \mu_{nonsmoke} < 0$

The claim is presented by the alternate hypothesis.

2f

Assuming equal variances between groups, perform a two-sample t-test comparing the mean birth weights between babies born to smoking mothers, versus non-smoking mothers, testing the null hypothesis you identified in (2e) at a significance level, $\alpha = 0.01$. Show your working throughout, and present values for the

- The direction of the test
- The t test statistic
- The standard error of the sampling distribution
- The standardized t-test statistic
- The degrees of freedom of the t distribution used to test the standardized test statistic
- The p-value of the standardized test statistic.

Round numeric values to 3 decimal places (12 pts)

direction of test	left-tailed test
Test stat $(\bar{x}_1 - \bar{x}_2)$	-0.330
SE	0.111
t	-2.980
df	844
p-val	0.001

2g

Using your answers in (2e) & (2f) make a decision to reject or fail to reject the null hypothesis. State your answer with reference to what you may conclude about the effect of maternal smoking on baby weights (2 pts)

We are rejecting the null hypothesis in favor of the alternate hypothesis since p-val$< \alpha$. The alternate hypothesis is the claim, so the conclusion is that the effect of smoking will lower the baby weights.

```
## Q1

# Preprocessing data
flight_dat <- read.csv(file="FlightResponse.csv")
low_dat <- flight_dat[flight_dat$AltCat == "low",]
high_dat <- flight_dat[flight_dat$AltCat == "high",]

low_flight = low_dat[low_dat$Flight == "1",]
high_flight = high_dat[high_dat$Flight == "1",]

# sample size
low_flight_n = NROW(low_flight)
high_flight_n = NROW(high_flight)
low_dat_n = NROW(low_dat)
high_dat_n = NROW(high_dat)

# calculating the normal dist criterion
p_hat_pooled_1 = (low_flight_n+high_flight_n)/(low_dat_n+high_dat_n)
exp_mean_1 = low_dat_n*p_hat_pooled_1
exp_mean_2 = high_dat_n*p_hat_pooled_1
exp_mean_comp_1 = (1-p_hat_pooled_1)*low_dat_n
exp_mean_comp_2 = (1-p_hat_pooled_1)*high_dat_n

# test statistic
p_hat_1_1 = low_flight_n/low_dat_n
p_hat_1_2 = high_flight_n/high_dat_n
test_stat_1 = p_hat_1_1-p_hat_1_2
SE_1 = sqrt(p_hat_pooled_1*(1-p_hat_pooled_1)
          *(1/low_dat_n+1/high_dat_n))

z_score_1 = (p_hat_1_1-p_hat_1_2)/SE_1

z_crit = qnorm(0.005)
p_val = pnorm(-abs(z_score_1))*2

## Q2
births_dat <- read.csv(file="north_carolina_births.csv")
smoker_dat <- births_dat[births_dat$habit == "smoker" & births_dat$premie == "full term",]
nonsmoker_dat <- births_dat[births_dat$habit == "nonsmoker" & births_dat$premie == "full
term",]

# Histograms
hist(smoker_dat$weight,
     main = "Histogram of Weight from Smoking Mother (Vinsensius)"
     ,breaks = seq(from=3, to=12, by=0.25),ylim = c(0,20),
     xlab = "Weight (lb)")

hist(nonsmoker_dat$weight,
     main = "Histogram of Weight from Non-Smoking Mother (Vinsensius)"
     ,breaks = seq(from=3, to=12, by=0.25),ylim = c(0,80),
     xlab = "Weight (lb)")

# Sample statistics
smoker_n = NROW(smoker_dat)
smoker_mean = mean(smoker_dat$weight)
smoker_std = sd(smoker_dat$weight)

nonsmoker_n = NROW(nonsmoker_dat)
nonsmoker_mean = mean(nonsmoker_dat$weight)
nonsmoker_std = sd(nonsmoker_dat$weight)

smoker_sig2 = var(smoker_dat$weight)
nonsmoker_sig2 = var(nonsmoker_dat$weight)
# non smoker on numerator and smoker in denominator
# assume nonsmoker is data 1 and smoker is data 2
# F test
F_test_2_result = var.test(nonsmoker_dat$weight,smoker_dat$weight,conf.level = 0.9)

F_crit_2 = qf(0.95,nonsmoker_n-1,smoker_n-1)
```

```
# t test, claim smoker weight> smoker weight

test_stat = smoker_mean-nonsmoker_mean


sig_hat_pooled_2 = sqrt((
(nonsmoker_n-1)*nonsmoker_std^2 +(smoker_n-1)*smoker_std^2 )
                        / (nonsmoker_n+smoker_n-2) )

SE_2 = sig_hat_pooled_2*sqrt(1/smoker_n + 1/nonsmoker_n)
t_score_2= (smoker_mean-nonsmoker_mean)/(SE_2)

t_crit_2 = qt(0.01,smoker_n+nonsmoker_n-2)
p_val = pt(t_score_2, (nonsmoker_n+smoker_n-2) )
#t_test_2_result = t.test( smoker_dat$weight,nonsmoker_dat$weight, conf.level = 0.99,
                          #alternative="greater", var.equal=TRUE)

## Q3
#cuckoo_dat <- read.csv(file="cuckoo_eggs.csv")

#cuckoo_anova = aov(species~bird,cuckoo_dat)
#boxplot(cuckoo_dat$Length ~ cuckoo_dat$Bird, main="Box plot of cuckoo egg length as function
of species",
#        xlab="Species", ylab="Egg length (mm)")
```