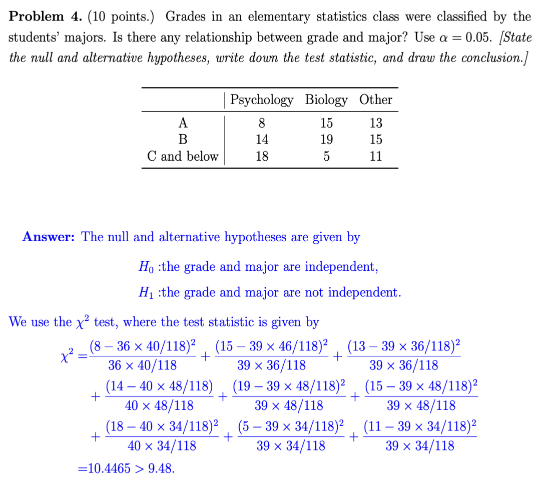
Selection bias: occurs if the method for selecting the participants produces a sample that does not represent the population of interest ; **Nonresponse bias:** occurs when a representative sample is chosen for a survey, but a subset cannot be contacted or does not respond; **Response bias:** occurs when participants provide incorrect info; **simple random sample w/ replacement (SRSWOR):**all possible subsets are equally likely to be chosen for the sample; **Stratified sampling**: population is divided into subgroups (strata), and a simple random sample selected from each subgroup; **Cluster sampling**: population is divided into subgroups called clusters, a random sample of clusters is selected; **Systematic sampling:** population ordered into a list, and list divided into consecutive segments of the same length. A random starting point is selected from the first segment, and the same point is sampled in each successive segment;**Commonly used transformations:** ‣ log / square root (positively skewed data) exponential / square (negatively skewed); **Simpson's paradox:** When effect of confounding variable is strong enough to produce relationships in a different direction from when data are separated into categories according to confounding variable



Problem 4. (10 points.) Grades in an elementary statistics class were classified by the students' majors. Is there any relationship between grade and major? Use $\alpha = 0.05$. *[State the null and alternative hypotheses, write down the test statistic, and draw the conclusion.]*

	Psychology	Biology	Other
A	8	15	13
B	14	19	15
C and below	18	5	11

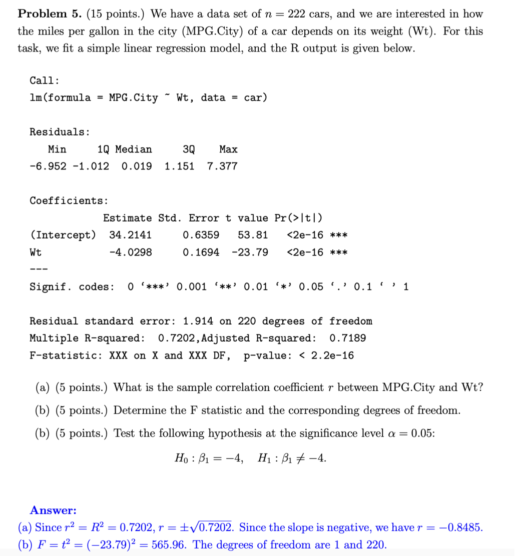
Answer: The null and alternative hypotheses are given by

$$H_0 \text{ :the grade and major are independent,}$$

$$H_1 \text{ :the grade and major are not independent.}$$

We use the χ^2 test, where the test statistic is given by

$$\begin{aligned}\chi^2 =& \frac{(8-36\times 40/118)^2}{36\times 40/118} + \frac{(15-39\times 46/118)^2}{39\times 36/118} + \frac{(13-39\times 36/118)^2}{39\times 36/118}\\ &+ \frac{(14-40\times 48/118)}{40\times 48/118} + \frac{(19-39\times 48/118)^2}{39\times 48/118} + \frac{(15-39\times 48/118)^2}{39\times 48/118}\\ &+ \frac{(18-40\times 34/118)^2}{40\times 34/118} + \frac{(5-39\times 34/118)^2}{39\times 34/118} + \frac{(11-39\times 34/118)^2}{39\times 34/118}\\ =&10.4465 > 9.48.\end{aligned}$$



Problem 5. (15 points.) We have a data set of $n = 222$ cars, and we are interested in how the miles per gallon in the city (MPG.City) of a car depends on its weight (Wt). For this task, we fit a simple linear regression model, and the R output is given below.

```
Call:
lm(formula = MPG.City ~ Wt, data = car)

Residuals:
   Min     1Q Median     3Q    Max
-6.952 -1.012  0.019  1.151  7.377

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.2141     0.6359   53.81   <2e-16 ***
Wt           -4.0298     0.1694  -23.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.914 on 220 degrees of freedom
Multiple R-squared:  0.7202,Adjusted R-squared:  0.7189
F-statistic: XXX on X and XXX DF,  p-value: < 2.2e-16
```

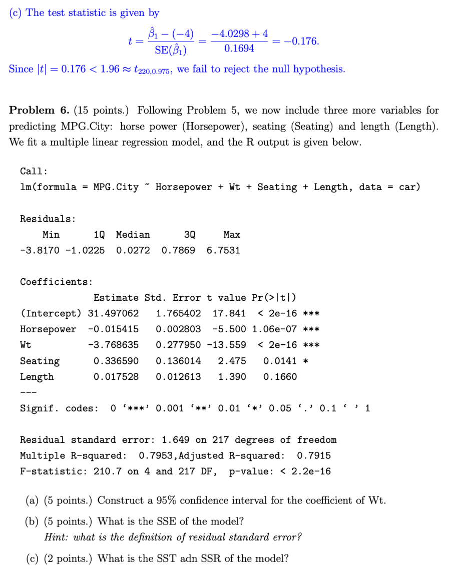
(a) (5 points.) What is the sample correlation coefficient r between MPG.City and Wt?

(b) (5 points.) Determine the F statistic and the corresponding degrees of freedom.

(b) (5 points.) Test the following hypothesis at the significance level $\alpha = 0.05$:

$$H_0 : \beta_1 = -4, \quad H_1 : \beta_1 \neq -4.$$

Answer:
(a) Since $r^2 = R^2 = 0.7202$, $r = \pm\sqrt{0.7202}$. Since the slope is negative, we have $r = -0.8485$.
(b) $F = t^2 = (-23.79)^2 = 565.96$. The degrees of freedom are 1 and 220.



(c) The test statistic is given by

$$t = \frac{\hat\beta_1 - (-4)}{\mathrm{SE}(\hat\beta_1)} = \frac{-4.0298 + 4}{0.1694} = -0.176.$$

Since $|t| = 0.176 < 1.96 \approx t_{220,0.975}$, we fail to reject the null hypothesis.

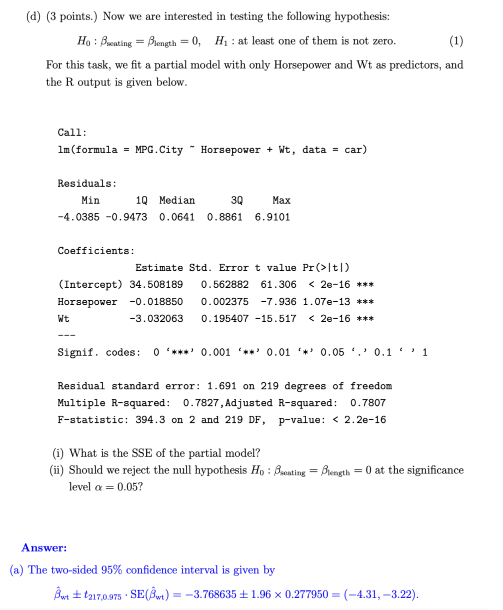
Problem 6. (15 points.) Following Problem 5, we now include three more variables for predicting MPG.City: horse power (Horsepower), seating (Seating) and length (Length). We fit a multiple linear regression model, and the R output is given below.

```
Call:
lm(formula = MPG.City ~ Horsepower + Wt + Seating + Length, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8170 -1.0225  0.0272  0.7869  6.7531

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.497062   1.765402  17.841  < 2e-16 ***
Horsepower  -0.015415   0.002803  -5.500 1.06e-07 ***
Wt          -3.768635   0.277950 -13.559  < 2e-16 ***
Seating      0.336590   0.136014   2.475   0.0141 *
Length       0.017528   0.012613   1.390   0.1660
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.649 on 217 degrees of freedom
Multiple R-squared:  0.7953,Adjusted R-squared:  0.7915
F-statistic: 210.7 on 4 and 217 DF,  p-value: < 2.2e-16
```

(a) (5 points.) Construct a 95% confidence interval for the coefficient of Wt.

(b) (5 points.) What is the SSE of the model?
Hint: what is the definition of residual standard error?

(c) (2 points.) What is the SST adn SSR of the model?



(d) (3 points.) Now we are interested in testing the following hypothesis:

$$H_0 : \beta_{\text{seating}} = \beta_{\text{length}} = 0, \quad H_1 : \text{at least one of them is not zero.} \tag{1}$$

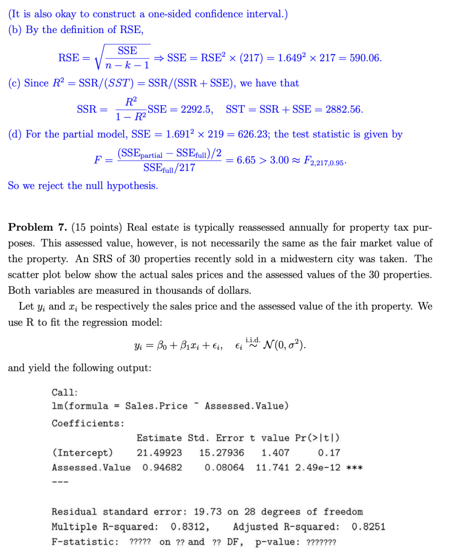
For this task, we fit a partial model with only Horsepower and Wt as predictors, and the R output is given below.

```
Call:
lm(formula = MPG.City ~ Horsepower + Wt, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0385 -0.9473  0.0641  0.8861  6.9101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.508189   0.562882  61.306  < 2e-16 ***
Horsepower  -0.018850   0.002375  -7.936 1.06e-13 ***
Wt          -3.032063   0.195407 -15.517  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.691 on 219 degrees of freedom
Multiple R-squared:  0.7827,Adjusted R-squared:  0.7807
F-statistic: 394.3 on 2 and 219 DF,  p-value: < 2.2e-16
```

(i) What is the SSE of the partial model?
(ii) Should we reject the null hypothesis $H_0 : \beta_{\text{seating}} = \beta_{\text{length}} = 0$ at the significance level $\alpha = 0.05$?

Answer:
(a) The two-sided 95% confidence interval is given by

$$\hat\beta_{\text{wt}} \pm t_{217,0.975} \cdot \mathrm{SE}(\hat\beta_{\text{wt}}) = -3.768635 \pm 1.96 \times 0.277950 = (-4.31, -3.22).$$



(It is also okay to construct a one-sided confidence interval.)
(b) By the definition of RSE,

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{SSE}}{n-k-1}} \Rightarrow \mathrm{SSE} = \mathrm{RSE}^2 \times (217) = 1.649^2 \times 217 = 590.06.$$

(c) Since $R^2 = \mathrm{SSR}/(SST) = \mathrm{SSR}/(\mathrm{SSR} + \mathrm{SSE})$, we have that

$$\mathrm{SSR} = \frac{R^2}{1-R^2}\mathrm{SSE} = 2292.5, \quad \mathrm{SST} = \mathrm{SSR} + \mathrm{SSE} = 2882.56.$$

(d) For the partial model, $\mathrm{SSE} = 1.691^2 \times 219 = 626.23$; the test statistic is given by

$$F = \frac{(\mathrm{SSE}_{\text{partial}} - \mathrm{SSE}_{\text{full}})/2}{\mathrm{SSE}_{\text{full}}/217} = 6.65 > 3.00 \approx F_{2,217,0.95}.$$

So we reject the null hypothesis.

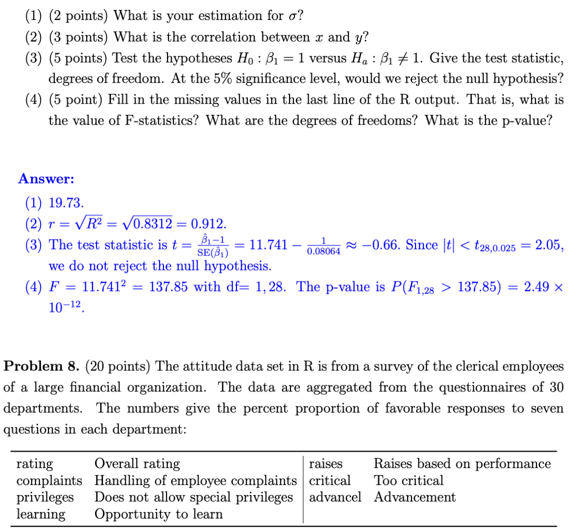
Problem 7. (15 points) Real estate is typically reassessed annually for property tax purposes. This assessed value, however, is not necessarily the same as the fair market value of the property. An SRS of 30 properties recently sold in a midwestern city was taken. The scatter plot below show the actual sales prices and the assessed values of the 30 properties. Both variables are measured in thousands of dollars.

Let y_i and x_i be respectively the sales price and the assessed value of the ith property. We use R to fit the regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

and yield the following output:

```
Call:
lm(formula = Sales.Price ~ Assessed.Value)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    21.49923   15.27936   1.407     0.17
Assessed.Value  0.94682    0.08064  11.741 2.49e-12 ***
---

Residual standard error: 19.73 on 28 degrees of freedom
Multiple R-squared:  0.8312,    Adjusted R-squared:  0.8251
F-statistic:  ?????  on ?? and  ?? DF,  p-value: ???????
```

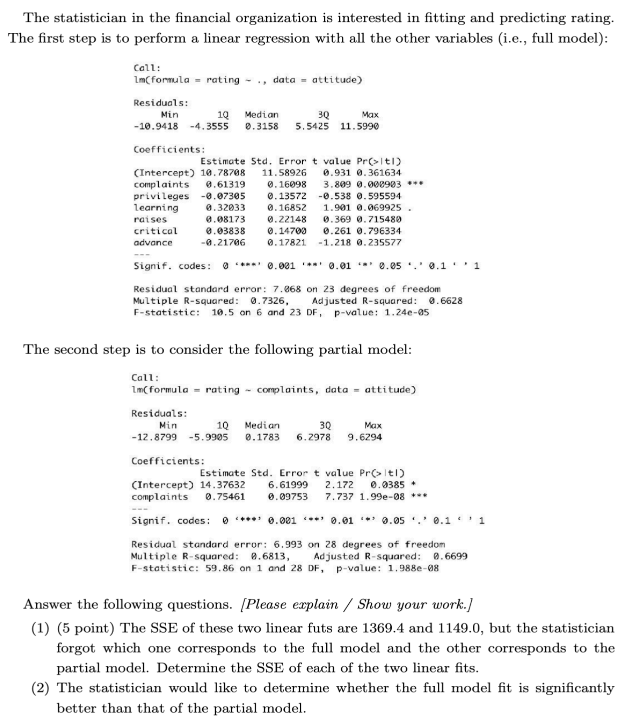


(1) (2 points) What is your estimation for σ?
(2) (3 points) What is the correlation between x and y?
(3) (5 points) Test the hypotheses $H_0 : \beta_1 = 1$ versus $H_a : \beta_1 \neq 1$. Give the test statistic, degrees of freedom. At the 5% significance level, would we reject the null hypothesis?
(4) (5 point) Fill in the missing values in the last line of the R output. That is, what is the value of F-statistics? What are the degrees of freedoms? What is the p-value?

Answer:
(1) 19.73.
(2) $r = \sqrt{R^2} = \sqrt{0.8312} = 0.912$.
(3) The test statistic is $t = \frac{\hat\beta_1 - 1}{\mathrm{SE}(\hat\beta_1)} = 11.741 - \frac{1}{0.08064} \approx -0.66$. Since $|t| < t_{28,0.025} = 2.05$, we do not reject the null hypothesis.
(4) $F = 11.741^2 = 137.85$ with df$= 1, 28$. The p-value is $P(F_{1,28} > 137.85) = 2.49 \times 10^{-12}$.

Problem 8. (20 points) The attitude data set in R is from a survey of the clerical employees of a large financial organization. The data are aggregated from the questionnaires of 30 departments. The numbers give the percent proportion of favorable responses to seven questions in each department:

rating	Overall rating	raises	Raises based on performance
complaints	Handling of employee complaints	critical	Too critical
privileges	Does not allow special privileges	advancel	Advancement
learning	Opportunity to learn		



The statistician in the financial organization is interested in fitting and predicting rating. The first step is to perform a linear regression with all the other variables (i.e., full model):

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

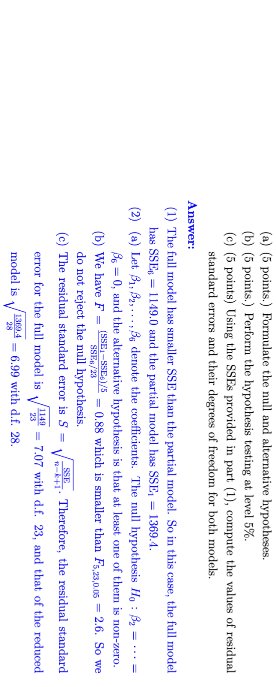
The second step is to consider the following partial model:

```
Call:
lm(formula = rating ~ complaints, data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8799  -5.9905   0.1783   6.2978   9.6294

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.37632    6.61999   2.172   0.0385 *
complaints   0.75461    0.09753   7.737 1.99e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.993 on 28 degrees of freedom
Multiple R-squared:  0.6813,    Adjusted R-squared:  0.6699
F-statistic: 59.86 on 1 and 28 DF,  p-value: 1.988e-08
```

Answer the following questions. *[Please explain / Show your work.]*

(1) (5 point) The SSE of these two linear futs are 1369.4 and 1149.0, but the statistician forgot which one corresponds to the full model and the other corresponds to the partial model. Determine the SSE of each of the two linear fits.

(2) The statistician would like to determine whether the full model fit is significantly better than that of the partial model.



(a) (5 points.) Formulate the null and alternative hypotheses.

(b) (5 points.) Perform the hypothesis testing at level 5%.

(c) (5 points) Using the SSEs provided in part (1), compute the values of residual standard errors and their degrees of freedom for both models.

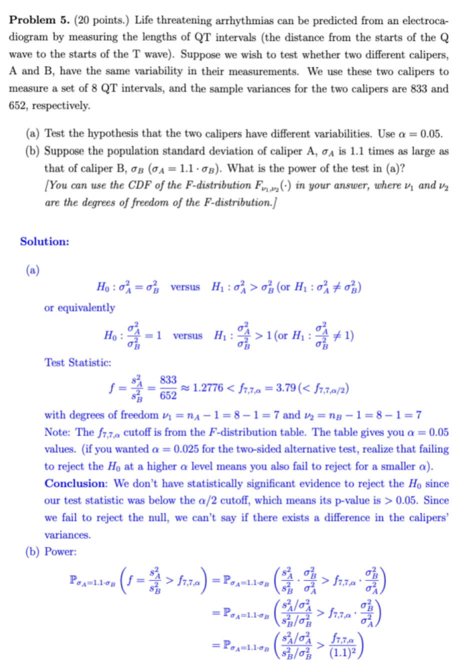
Answer:

(1) The full model has smaller SSE than the partial model. So in this case, the full model has $\mathrm{SSE}_6 = 1149.0$ and the partial model has $\mathrm{SSE}_1 = 1369.4$.

(2) (a) Let $\beta_1, \beta_2, \ldots, \beta_6$ denote the coefficients. The null hypothesis $H_0 : \beta_2 = \cdots = \beta_6 = 0$, and the alternative hypothesis is that at least one of them is non-zero.

(b) We have $F = \frac{(\mathrm{SSE}_1 - \mathrm{SSE}_6)/5}{\mathrm{SSE}_6/23} = 0.88$ which is smaller than $F_{5,23,0.05} = 2.6$. So we do not reject the null hypothesis.

(c) The residual standard error is $S = \sqrt{\frac{\mathrm{SSE}}{n-k+1}}$. Therefore, the residual standard error for the full model is $\sqrt{\frac{1149}{23}} = 7.07$ with d.f. 23, and that of the reduced model is $\sqrt{\frac{1369.4}{28}} = 6.99$ with d.f. 28.



Problem 5. (20 points.) Life threatening arrhythmias can be predicted from an electrocadiogram by measuring the lengths of QT intervals (the distance from the starts of the Q wave to the starts of the T wave). Suppose we wish to test whether two different calipers, A and B, have the same variability in their measurements. We use these two calipers to measure a set of 8 QT intervals, and the sample variances for the two calipers are 833 and 652, respectively.

(a) Test the hypothesis that the two calipers have different variabilities. Use $\alpha = 0.05$.

(b) Suppose the population standard deviation of caliper A, σ_A is 1.1 times as large as that of caliper B, σ_B ($\sigma_A = 1.1 \cdot \sigma_B$). What is the power of the test in (a)? *[You can use the CDF of the F-distribution $F_{\nu_1,\nu_2}(\cdot)$ in your answer, where ν_1 and ν_2 are the degrees of freedom of the F-distribution.]*

Solution:

(a)

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad \text{versus} \quad H_1 : \sigma_A^2 > \sigma_B^2 \,(\text{or } H_1 : \sigma_A^2 \neq \sigma_B^2)$$

or equivalently

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1 \quad \text{versus} \quad H_1 : \frac{\sigma_A^2}{\sigma_B^2} > 1 \,(\text{or } H_1 : \frac{\sigma_A^2}{\sigma_B^2} \neq 1)$$

Test Statistic:

$$f = \frac{s_A^2}{s_B^2} = \frac{833}{652} \approx 1.2776 < f_{7,7,\alpha} = 3.79 \,(< f_{7,7,\alpha/2})$$

with degrees of freedom $\nu_1 = n_A - 1 = 8 - 1 = 7$ and $\nu_2 = n_B - 1 = 8 - 1 = 7$
Note: The $f_{7,7,\alpha}$ cutoff is from the F-distribution table. The table gives you $\alpha = 0.05$ values. (if you wanted $\alpha = 0.025$ for the two-sided alternative test, realize that failing to reject the H_0 at a higher α level means you also fail to reject for a smaller α).
Conclusion: We don't have statistically significant evidence to reject the H_0 since our test statistic was below the $\alpha/2$ cutoff, which means its p-value is > 0.05. Since we fail to reject the null, we can't say if there exists a difference in the calipers' variances.

(b) Power:

$$\begin{aligned}\mathbb{P}_{\sigma_A=1.1\cdot\sigma_B}\left(f = \frac{s_A^2}{s_B^2} > f_{7,7,\alpha}\right) &= \mathbb{P}_{\sigma_A=1.1\cdot\sigma_B}\left(\frac{s_A^2}{s_B^2}\cdot\frac{\sigma_B^2}{\sigma_A^2} > f_{7,7,\alpha}\cdot\frac{\sigma_B^2}{\sigma_A^2}\right)\\ &= \mathbb{P}_{\sigma_A=1.1\cdot\sigma_B}\left(\frac{s_A^2/\sigma_A^2}{s_B^2/\sigma_B^2} > f_{7,7,\alpha}\cdot\frac{\sigma_B^2}{\sigma_A^2}\right)\\ &= \mathbb{P}_{\sigma_A=1.1\cdot\sigma_B}\left(\frac{s_A^2/\sigma_A^2}{s_B^2/\sigma_B^2} > \frac{f_{7,7,\alpha}}{(1.1)^2}\right)\end{aligned}$$

Example: airline revenue

- Airlines use sampling to estimate the mean of the revenue for passengers traveling between A to B
- Suppose the revenues are normally distributed with unknown mean μ and known standard deviation $\sigma = 50$
- To estimate the mean revenue per ticket, the airline uses a sample of 400 tickets with sample mean $\bar{X} = 175.60$ dollars

The two-sided 95% CI for the mean share μ is

$$\left[175.6 - 1.96\frac{50}{\sqrt{400}}, 175.6 + 1.96\frac{50}{\sqrt{400}}\right] = [170.70, 180.50]$$

Example: CI for the density of the earth

- The following table gives 29 measurements of the density of earth made in 1798 by the British scientist Henry Cavendish
- Expressed as multiples of the density of water, i.e., in grams/cc.
- Estimate the density of earth from these measurements using a 95% CI

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.86	5.85	

Problem 4. (15 points.) Answer the following questions and explain your work.

(a) Two data sets have correlations +0.3 and −0.7 respectively. Which one exhibits stronger linear relationship? The one with correlation +0.3 or the one with correlation −0.7?

(b) Is the following statement true or false:
when Y is usually less than X, the correlation between X and Y is negative since .

(c) Fill in the blank in the following table to make the correlation between X and Y exactly 1.

X	1	3	3	4
Y	1	4	4	

Answer:

(a) The one with correlation −0.7 exhibits stronger linear relationship since the correlation coefficient is larger in absolute value.

(b) False. The correlation between X and Y can be positive, e.g., $Y = X/2$ for $X \geq 0$.

(c) When $r = 1$, the points are on a line. Using the observed poins, we can see the line is $Y = 1.5X - 0.5$. With $X = 4$, $Y = 4 \times 1.5 - 0.5 = 5.5$.

Problem 5. (25 points) Suppose $X_1, X_2, \ldots, X_n$ ($n \geq 3$) are i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ denote the sample mean estimator and the sample variance estimator repectively.

(a) Show that $\hat{\mu}_1 = \frac{X_1+X_2}{2}$ and $\hat{\mu}_2 = 0.3X_1 + 0.4X_2 + 0.3X_3$ are both unbiased estimators for μ. *[Show your work.]*

(b) Compute the MSE of $\hat{\mu}_1$ and $\hat{\mu}_2$, and compare with the MSE of $\bar{X}$. *[Show your work.]*

(c) Show that $(\bar{X})^2$ is a biased estimator for μ^2 and compute its bias. *[Show your work.]*

(d) Propose an unbiased estimator for μ^2. *[Show your work.]*

(e) Suppose further that $X_1, X_2, \ldots, X_n$ are from a normal distribution. Calculate the MSE of $\rho\sum_{i=1}^n (X_i - \bar{X})^2$ for estimating σ^2 and find the ρ that minimizes the MSE. *[Show your work.]*

[Hint: the expectation and variance of a χ^2 r.v. with r df is r and $2r$, respectively.]

Example

A computer network manager wants to model how access time in milliseconds (y) for data files varies with the number of simultaneous users (x) accessing the files. Based on 50 paired observations, we obtain the following summary statistics:

$$\bar{x} = 8.7,\ s_x = 2.5,\ \bar{y} = 15.3,\ s_y = 4.8,\ r = 0.8$$

$\hat{\beta}_1 = r\frac{s_y}{s_x} = 0.8 \times 4.8/2.5 = 1.536$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 15.3 - 1.536 \times 8.7 = 1.937$

LS line: y = 1.937 + 1.536 x Predicted value: y = 1.937 + 1.536 x 12 = 20.368

(a) Using the linearity of expectation, we have

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}[\frac{X_1+X_2}{2}] = \frac{1}{2}(\mathbb{E}[X_1] + \mathbb{E}[X_2]) = \mu$$
$$\mathbb{E}[\hat{\mu}_2] = \mathbb{E}[0.3X_1 + 0.4X_2 + 0.3X_3] + 0.3\mathbb{E}[X_1] + 0.4\mathbb{E}[X_2] + 0.3\mathbb{E}[X_3] = \mu.$$

(b) Since both $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased estimators,

$$\mathrm{MSE}(\hat{\mu}_1) = \mathrm{Var}(\hat{\mu}_1) = \frac{1}{4}\mathrm{Var}(X_1 + X_2) = \frac{1}{4}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)) = \frac{\sigma^2}{2}$$
$$\mathrm{MSE}(\hat{\mu}_2) = \mathrm{Var}(\hat{\mu}_2) = 0.09\sigma^2 + 0.16\sigma^2 + 0.09\sigma^2 = 0.34\sigma^2.$$

Recall that $\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) = \sigma^2/n$ and $n \geq 3$; there is $\mathrm{MSE}(\hat{\mu}_1) > \mathrm{MSE}(\bar{X})$ and $\mathrm{MSE}(\hat{\mu}_2) > \mathrm{MSE}(\bar{X})$.

(c) $\mathbb{E}[(\bar{X})^2] = \mathrm{Var}(\bar{X}) + (\mathbb{E}[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2$. Therefore, $\bar{X}^2$ is not an unbiased estimator for μ^2 and its bias is $\frac{\sigma^2}{n}$.

(d) $\hat{\theta} = (\bar{X})^2 - \frac{S^2}{n}$ is an unbiased estimator for μ^2:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[(\bar{X})^2] - \frac{1}{n}\mathbb{E}[S^2] = \frac{\sigma^2}{n} + \mu^2 - \frac{\sigma^2}{n} = \mu^2.$$

(e) Let $Y = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2$. Then $Y \sim \chi^2_{n-1}$.

$$\begin{aligned}\mathrm{MSE}(\rho\sum_{i=1}^n (X_i - \bar{X})^2) =& \mathrm{MSE}(\rho\sigma^2 Y) = \mathrm{Var}(\rho\sigma^2 Y) + (\mathbb{E}[\rho\sigma^2 Y] - \sigma^2)^2\\ =& \rho^2\sigma^4\mathrm{Var}(Y) + (\rho\sigma^2\mathbb{E}[Y] - \sigma^2)^2\\ =& \rho^2\sigma^4 2(n-1) + (\rho\sigma^2(n-1) - \sigma^2)^2\\ =& \sigma^4\{2\rho^2(n-1) + (\rho(n-1) - 1)^2\}.\end{aligned}$$

Taking derivative w.r.t. ρ and set it to zero, we have

$$\begin{aligned}&\sigma^2\{4\rho(n-1) + 2(n-1)(\rho n - \rho - 1)\} = 0\\ \Rightarrow& 2\sigma^4(n-1)(\rho + \rho n - 1) = 0\\ \Rightarrow& \rho = \frac{1}{n+1}.\end{aligned}$$

It can be checked that the minimum is achieved when $\rho = \frac{1}{n+1}$.

	Average	SD
Husband height (inches)	68.2	3.5
Wife height (inches)	62.8	2.0

The sample correlation is $r = 0.25$. Answer the following questions.

(a) (10 points.) Write down the regression line with y being the husband's height and x being the wife's height.

(b) (5 points.) Use the regression line to predict the height of a husband when the height of his wife is 65 inches.

Solution:

(a) The regression line is

$$\frac{y - 68.2}{3.5} = 0.25 \times \frac{x - 62.8}{2.0}.$$

Rearranging the terms, we get $y = 0.4375x + 40.725$.

(b) Plugging in $x = 65$, we get $y = 69.1625$, i.e., the predicted husband height is 69.1625.

Problem 6. (10 points) The city of Philadelphia aims to find out its residents' opinion on making Daylight Saving Time permanent. Suppose that 50% of the Philadelphia residents are in favor of having Daylight Saving Time permanent. A polling company randomly interviewd 40 residents of Philadelphia and asked their opinion. Suppose the interviewees' opinions are independent of each other. The goal is to find out the probability that a majority of the sample (21 or more) will favor the idea of making Daylight Saving Time permanent.

(a) Calculate the probability using the normal approximation without the continuity correction. *[Show your work.]*

(b) Calculate the probability using the normal approximation with the continuity correction. *[Show your work.]*

[You can present your answers using $\Phi(\cdot)$, the CDF of the standard normal distribution.]

Let $X_1, X_2, \ldots, X_{40}$ denote the opinions of the 40 interviewees, where $X_i = 1$ if the i-th interviewee is in favor of making Daylight Saving Time permanent and $X_i = 0$ otherwise. $\sum_{i=1}^{40} X_i \sim \mathrm{Binom}(20, 10)$.

(a) Using normal approximation without continuity correction, the probablity is

$$\mathbb{P}\Big(\sum_{i=1}^{40} X_i \geq 21\Big) = \mathbb{P}\Big(\frac{\sum_{i=1}^{40} X_i - 20}{\sqrt{10}} \geq \frac{21 - 20}{\sqrt{10}}\Big) \approx 1 - \Phi\Big(\frac{1}{\sqrt{10}}\Big).$$

(b) Using normal approximation with continuity correction, the probablity is

$$\mathbb{P}\Big(\sum_{i=1}^{40} X_i \geq 21\Big) \approx \mathbb{P}(\mathcal{N}(20, 10) \geq 20.5) = \mathbb{P}\Big(\mathcal{N}(0, 1) \geq \frac{0.5}{\sqrt{10}}\Big) = 1 - \Phi\Big(\frac{0.5}{\sqrt{10}}\Big).$$

Problem 7. (15 points) A real estate office wants to investigate how far on average the head of household in a neighborhood has to commute for work. To do so, a simple random sample of 30 households are chosen. It is found that on average, the heads of the sample households commute 10.6 miles to work, and the standard deviation of the sample is 7.8 miles. Suppose the commuting distance of each head of household is from $\mathcal{N}(\mu, \sigma^2)$ and independent of each other.

(a) Find a 95% two-sided confidence interval for the average commute distance in the neighborhood using large-sample approximation. *[Show your work.]*

(b) Find an exact 95% two-sided confidence interval for the average commute distance in the neighborhood without using large-sample approximation. *[Show your work.]*

(c) Find a 95% two-sided confidence interval for the variance of the commute distance in the neighborhood. *[Show your work.]*

(a) The approximate 95% two-sided confidence interval for μ is

$$\Big[\bar{X} - z_{0.025}\frac{S}{\sqrt{30}}, \bar{X} + z_{0.025}\frac{S}{\sqrt{30}}\Big] = [7.81, 13.39].$$

(b) The exact 95% two-sided confidence interval for μ is

$$\Big[\bar{X} - t_{29,0.025}\frac{S}{\sqrt{30}}, \bar{X} + t_{29,0.025}\frac{S}{\sqrt{30}}\Big] = [7.68, 13.52].$$

(c) The 95% two-sided confidence interval for σ^2 is

$$\Big[\frac{29 \cdot S^2}{\chi^2_{29,0.025}}, \frac{29 \cdot S^2}{\chi^2_{29,0.975}}\Big] = [38.59, 109.95].$$

Problem 6. (15 points.) To evaluate the accuracy of thermometers purchased from a medical supply vendor, the quality control department of a hospital conducts a test on a sample of thermometers at a controlled temperature of 98.6°F (normal body temperature). At this temperature, the standard deviation of the thermometer readings is 0.5°F. The readings are assumed to follow a normal distribution. Define the bias of these thermometers to be the (true) mean of the thermometer readings minus 98.6°F. How many readings need to be tested to have a 95% power to detect a bias of 0.3°F using a 0.05-level one-sided test (the alternative corresponds to positive bias)?

Solution: Let μ denote the population mean of the thermometer reading. We are insterested in testing $H_0 : \mu \leq 98.6$ against $H_1 : \mu > 98.6$. Let $X_1, \ldots, X_n$ denote the readings, and we reject H_0 when

$$\frac{\bar{X} - 98.6}{0.5/\sqrt{n}} > z_{0.05}.$$

When $\mu = 98.9$°F, the power is given by

$$\mathbb{P}\Big(\bar{X} - 98.6 \geq z_{0.05}\frac{0.5}{\sqrt{n}}\Big) = \mathbb{P}\Big(\bar{X} - 98.9 \geq -0.3 + z_{0.05}\frac{0.5}{\sqrt{n}}\Big) = 1 - \Phi\Big(z_{0.05} - 0.3\frac{\sqrt{n}}{0.5}\Big).$$

Letting the above $\geq$ 95% and solve for n, we conclude that at least $\lceil(\frac{10}{3} \cdot z_{0.05})^2\rceil = 31$ samples are needed to achieve a 95% power.

	Before	After	After - Before
	24.6	10.1	-14.5
	17.0	5.7	-11.3
	16.0	5.6	-10.4
	10.4	3.4	-7.0
	8.2	6.5	-1.7
	7.9	0.7	-7.2
	8.2	6.5	-1.7
	7.9	0.7	-7.2
	5.8	6.1	0.3
	5.4	4.7	-0.7
	5.1	2.0	-3.1
	4.7	2.9	-1.8
Sample mean	10.1	4.58	-5.53
Sample SD	6.06	2.75	4.78

What can you conclude about the effect of captopril? State the assumptions you made in order to arrive at your conclusion.

Solution:

For this problem, we use a significance level = 0.05. First we state the hypotheses.

Null hypothesis: The treatment has no effect or positive effect on the amounts of urinary protein.

Alternative hypothesis: The treatment has a negative effect on the amounts of urinary protein.

Thus, let us compute the test statistic. Since the sample size is small and the variance is unknown, we use t-distribution here.

$$t = \frac{-5.53 - 0}{4.78/\sqrt{12}} = -4.01$$

Note that we are doing a one-sided test here. By the table for t-distribution, we know the rejection rule is

$$t_{11,0.95} = -1.796$$

Since $-4.01 < -1.796$, we reject the null hypothesis that the treatment has no effect or positive effect on the amounts of urinary protein.

	Control	Peptic Ulcer
Group A	4219	579
Group O	4578	911

Is there a relationship between blood type and propensity to have peptic ulcer? Use $\alpha = 0.05$.

Solution:

H_0: There is **no relationship** between blood type and propensity to have peptic ulcer.
(i.e. Blood type and the propensity to have peptic ulcer are **independent** of one another.)

v.s.

H_1: There **exists** a relationship between blood type and propensity to have peptic ulcer.
(i.e. Blood type and the propensity to have peptic ulcer are **dependent** on one another.)

- Original Table (with row and column totals)

	Control	Peptic Ulcer	Total
Group A	4219	579	4798
Group O	4578	911	5489
Total	8797	1490	10287

- Independent Table's Expected Count

	Control	Peptic Ulcer
Group A	$\frac{4798\times8797}{10287}$	$\frac{4798\times1490}{10287}$
Group O	$\frac{5489\times8797}{10287}$	$\frac{5489\times1490}{10287}$

$\approx$

	Control	Peptic Ulcer
Group A	4103.043	694.957
Group O	4693.957	795.043

- Test Statistic $\chi^2 = \sum\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$

	Control	Peptic Ulcer
Group A	$\frac{(4219-4103.043)^2}{4103.043}$	$\frac{(579-694.957)^2}{694.957}$
Group O	$\frac{(4578-4693.957)^2}{4693.957}$	$\frac{(911-795.043)^2}{795.043}$

$\approx$

	Control	Peptic Ulcer
Group A	3.277	19.348
Group O	2.865	16.912

- $\chi^2 \approx 3.277 + 19.348 + 2.865 + 16.912 = 42.402 \geq \chi^2_{1,\alpha=0.05} = 3.843$ from the χ^2-distribution table with degrees of freedom $(r-1)(c-1) = (2-1)(2-1) = 1$ (for r = # rows and c = # columns).
- Conclusion: We reject the H_0 because our test statistic is above the cutoff (meaning the p-value of this test is $< \alpha = 0.05$). Therefore, there's statistically significant evidence to show there exists *some type of association* between blood type and propensity to have peptic ulcers. (We don't know what the relationship is though.)

- Four possibilities in total

		Decision: Do not reject H_0	Decision: Reject H_0
H_0	True	Correct Decision	Type I Error
	False	Type II Error	Correct Decision

- **Type I error:** reject H_0 when H_0 is true $\rightsquigarrow$ false positive
- **Type II error:** do not reject H_0 when H_0 is false $\rightsquigarrow$ false negative