GLIB DOLOTOV (U14734371)
MA214 SPRG 2019

ASSIGNMENT #4

9.70 **Study of armyworm pheromones.** A study was conducted to determine the effectiveness of pheromones produced by two different strains of fall armyworms: the corn-strain and the rice-strain (*Journal of Chemical Ecology*, Mar. 2013). Both corn-strain and rice-strain male armyworms were released into a field containing a synthetic pheromone made from a corn-strain blend. A count of the number of males trapped by the pheromone was then determined. The experiment was conducted once in a corn field and then again in a grass field. The results are provided in the accompanying table.

1. Consider the corn field results. Construct a 90% confidence interval for the difference between the proportions of corn-strain and rice-strain males trapped by the pheromone.
2. Consider the grass field results. Construct a 90% confidence interval for the difference between the proportions of corn-strain and rice-strain males trapped by the pheromone.
3. Based on the confidence intervals, parts **a** and **b**, what can you conclude about the effectiveness of a corn-blend synthetic pheromone placed in a corn field? A grass field?
4. The researchers also want to compare the proportion of corn-strain males trapped in the corn field to the proportion of corn-strain males trapped in the grass field. Carry out this comparison using a hypothesis test (at α=.10
). What inference can you draw from the data?
5. Repeat part **d** for the proportions of rice-strain males trapped by the pheromone.

	• Corn Field	• Grass Field
• Number of corn-strain males released	• 112	• 215
• Number trapped	• 86	• 164
• Number of rice-strain males released	• 150	• 669
• Number trapped	• 92	• 375

Large-Sample $100(1-\alpha)\%$ Confidence Interval for $(p_1 - p_2)$: Normal (z) Statistic

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

(a) $\hat{p}_1 = 86/112 \approx 0.768$

$\hat{p}_2 = 92/150 \approx 0.613$

$(\hat{p}_1 - \hat{p}_2) \approx 0.155$

90% CI

$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

$= 0.155 \pm 2.576\sqrt{\frac{(0.768)(0.232)}{112} + \frac{(0.613)(0.387)}{150}}$

$= 0.155 \pm 2.576(0.053) = \boxed{0.155 \pm 0.137}$

(b) $\hat{p}_1 = 164/215 \approx 0.763$

$\hat{p}_2 = 375/669 \approx 0.561$

$(\hat{p}_1 - \hat{p}_2) \approx 0.202$

90% CI

$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

$= 0.202 \pm 2.576\sqrt{\frac{0.763 \cdot 0.337}{215} + \frac{0.561 \cdot 0.439}{669}}$

$= \boxed{0.202 \pm 0.102}$

(c) We can conclude from the (a) and (b) that greater proportions of corn-strained mice were captured.

(d) $\hat{p}_1 = 86/112 \approx 0.768$

$\hat{p}_2 = 164/215 \approx 0.763$

$(\hat{p}_1 - \hat{p}_2) \approx 0.005$

90% CI

$$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$= 0.005 \pm 2.576\sqrt{\frac{0.768 \cdot 0.232}{112} + \frac{0.763 \cdot 0.237}{215}}$$

$$= \boxed{0.005 \pm 2.262}$$

(e) $\hat{p}_1 = 92/150 \approx 0.613$

$\hat{p}_2 = 375/669 \approx 0.561$

$(\hat{p}_1 - \hat{p}_2) \approx 0.052$

90% CI

$$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$= 0.052 \pm 2.576\sqrt{\frac{0.613 \cdot 0.387}{150} + \frac{0.561 \cdot 0.439}{669}}$$

$$= \boxed{0.052 \pm 2.022}$$

9.107 **Hippo grazing patterns in Kenya.** In Kenya, human-induced land-use changes and excessive resource extraction have threatened the jungle ecosystem by reducing animal grazing areas and disrupting access to water sources. In *Landscape & Ecology Engineering* (Jan. 2013), researchers compared hippopotamus grazing patterns in two Kenyan areas: a national reserve and a community pastoral ranch. Each area was subdivided into plots of land. The plots were sampled (406 plots in the national reserve and 230 plots in the pastoral ranch), and the number of hippo trails from a water source was determined for each plot. Sample statistics are provided in the table. Suppose the researchers want to know if the variability in number of hippo trails from a water source in the National Reserve differs from the variability in number of hippo trails from a water source in the pastoral ranch.

	National Reserve	Pastoral Ranch
Sample size	406	230
Mean number of trails	0.31	0.13
Standard deviation	0.40	0.30

Source: Kanga, E. M., et al. "Hippopotamus and livestock grazing: Influences on riparian vegetation and facilitation of other herbivores in the Mara Region of Kenya." Landscape & Ecology Engineering, Vol. 9, No. 1, Jan. 2013.

1. Find an interval estimate of $(\sigma_1)^2/(\sigma_2)^2$, the ratio of the variances associated with the two areas. Use a 90% confidence level.
2. Can the researchers reliably conclude that the variability in number of hippo trails from a water source in the National Reserve differs from the variability in number of hippo trails from a water source in the pastoral ranch? Explain.
3. Explain why a test of hypothesis at α=.10 will result in the same inference, then carry out the test to verify your results.

(1) $\frac{\sigma_1^2}{\sigma_2^2}$ $\sigma_1 \leftarrow$ national reserve ; $s_1 = 0.4$ $\quad \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{L,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{U,\alpha/2}$

$\sigma_2 \leftarrow$ pastoral ranch ; $s_2 = 0.3$

> find F values using stat software

$F_{L,\alpha/2}$: $\alpha/2 = 0.05$, df numerator $= n_1 - 1 = 405$, df denominator $= n_2 - 1 = 229$

$= \frac{s_1^2 / s_2^2}{F_{L,(1-\alpha/2)}}$

$F_{U,\alpha/2}$: $\alpha/2 = 0.05$, df numerator $= n_2 - 1 = 229$, df denominator $= n_1 - 1 = 405$

```
> qf(0.95, df1 = 405, df2 = 229)
[1] 1.216095   = F_{L,α/2}
> qf(0.95, df1 = 229, df2 = 405)
[1] 1.208856   = F_{U,α/2}
> s <- 0.4^2 / 0.3^2
> s                    } ← s1^2 / s2^2
[1] 1.777778
> s / qf(0.95, df1 = 405, df2 = 229)
[1] 1.461874   ← s1^2/s2^2 · 1/F_{L,α/2}
> s * qf(0.95, df1 = 229, df2 = 405)
[1] 2.149078   ← s1^2/s2^2 · F_{U,α/2}
```

(1.462, 2.149)

(2) Yes. If they were the same, the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ would equal 1. However, 1 falls outside the confidence interval, implying that the variabilities differ.

(3) a hypothesis test is equivalent:

A test statistic falling into a rejection region happens iff. the null hypoth value falls outside the confidence int.

	$H_0: \sigma_1^2 = \sigma_2^2$
	$H_a: \sigma_1^2 \neq \sigma_2^2$
Test statistic:	$F_c = \dfrac{\text{Larger } s^2}{\text{Smaller } s^2}$
Rejection region:	$F_c > F_{\alpha/2}$
Numerator df:	$\nu_1 = n - 1$ for larger s^2
Denominator df:	$\nu_2 = n - 1$ for smaller s^2
p-value:	$P(F^* < 1/F_c) + P(F > F_c)$

$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

Larger s^2: $(0.4)^2$
$\Rightarrow \nu_1 = 406 - 1 = 405$

Smaller s^2: $(0.3)^2$
$\Rightarrow \nu_2 = 230 - 1 = 229$

$F_c = \frac{0.16}{0.09} = 1.778 \qquad F_{0.05}$ ($df_1 = 405$, $df_2 = 229$) $= 1.216$ (part 2)

$F_c = 1.778 > F_{\alpha/2} = 1.216 \Rightarrow$ reject null hypothesis, accept alternate: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$.

HANDSHK 9.112 **Hygiene of handshakes, high fives, and fist bumps.** Refer to the *American Journal of Infection Control* (Aug. 2014) study of the hygiene of hand greetings, **Exercise 9.24** (p. **451**). The number of bacteria transferred from a gloved hand dipped into a culture of bacteria to a second gloved hand contacted by either a handshake, high five, or fist bump was recorded. Recall that the experiment was replicated five times for each contact method and the data used to compare the mean percentage of bacteria transferred for any two contact methods. The data are repeated in the table.

1. What assumption about the variability of the data is required to make the mean comparisons for this small-sample study?
2. Conduct a test to determine if the assumption, part **a,** is reasonably satisfied when comparing the handshake to the fist bump. Test using α=.05
.
3. How does your answer, part **b,** affect the validity of the inferences made in **Exercise 9.24**?
4. Find the 95% confidence interval you found in Exercise 9.24 using this new information.

Handshake:	131	74	129	96	92
High five:	44	70	69	43	53
Fist bump:	15	14	21	29	21

(1)

Conditions Required for Inferences about $\mu_1 - \mu_2$

Large samples:
1. Independent random samples
2. $n_1 \geq 30, n_2 \geq 30$

Small samples:
1. Independent random samples
2. Both populations normal
3. $\sigma_1^2 = \sigma_2^2$

(3) It calls into question the validity of those inferences

(4)

Approximate Small-Sample Procedures when $\sigma_1^2 \neq \sigma_2^2$

1. **Equal Sample Sizes** $(n_1 = n_2 = n)$

Confidence interval: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{(s_1^2 + s_2^2)/n}$

Test statistic for H_0: $(\mu_1 - \mu_2) = 0$: $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{(s_1^2 + s_2^2)/n}$

where t is based on $\nu = n_1 + n_2 - 2 = 2(n-1)$ degrees of freedom.

2) $H_0: \frac{\sigma^2_{Handshake}}{\sigma^2_{Fbump}} = 1$ $\quad H_a: \frac{\sigma_H^2}{\sigma_F^2} \neq 1$

```
> # Sample variances
> var(data$HShake)
[1] 615.3
> var(data$FBump)
[1] 36

> # Test statistic
> s = var(data$HShake) / var(data$FBump)
> s
[1] 17.09167
> qf(0.975, df1 = 4, df2 = 4)
[1] 9.60453
```

$F_c = 17.09 > F_{\alpha/2} = 9.60$

$\Rightarrow$ reject null, $\frac{\sigma^2_{Hand}}{\sigma^2_{Fbump}} \neq 1$

$\Leftrightarrow \sigma^2_{Hand} \neq \sigma^2_{Fbump}$.

```
> mean(data$HShake)
[1] 104.4
> mean(data$FBump)
[1] 20
```

$CI = (104.4 - 20) \pm t_{0.025,\ df=8} \cdot \sqrt{(615.3+36)/10}$

$= 84.4 \pm 2.306 \cdot 8.070$

$= \boxed{84.4 \pm 18.610}$

11.14 **Forecasting movie revenues with Twitter.** A study presented at the 2010 *IEEE International Conference on Web Intelligence and Intelligent Agent Technology* investigated whether the volume of chatter on **Twitter.com** could be used to forecast the box office revenues of movies. For each in a sample of 24 recent movies, opening weekend box office revenue (in millions of dollars) was measured as well as the movie's *tweet rate* (the average number of tweets referring to the movie one week prior to the movie's release).

1. In this study, identify the dependent and independent variables.
2. Explain why a probabilistic model is more appropriate than a deterministic model.
3. Write the equation of the straight-line, probabilistic model.

1. dependent: opening weekend box-office revenue
   independent: movie "tweet rate"

2. Because there are countless other factors that might affect our dependent variable that we are not collecting data for. i.e., the relationship between tweet-rate & opening is not 1-to-1 and direct.

3. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = \{1, 2, 3, \ldots, 24\}$

FCAT 11.30 **FCAT scores and poverty.** In the state of Florida, elementary school performance is based on the average score obtained by students on a standardized exam called the Florida Comprehensive Assessment Test (FCAT). An analysis of the link between FCAT scores and sociodemographic factors was published in the *Journal of Educational and Behavioral Statistics* (Spring 2004). Data on average math and reading FCAT scores of third graders, as well as the percentage of students below the poverty level, for a sample of 22 Florida elementary schools are listed in the table on p. **601**.

a. Propose a straight-line model relating math score (y) to percentage (x) of students below the poverty level.
b. Use the method of least squares to fit the model to the data in the FCAT file.
c. Graph the least squares line on a scatterplot of the data. Is there visual evidence of a relationship between the two variables? Is the relationship positive or negative?
d. Interpret the estimates of the y-intercept and slope in the words of the problem.
e. Now consider a model relating reading score (y) to percentage (x) of students below the poverty level. Repeat parts **a–d** for this model.

(a) $y = \beta_0 + \beta_1 x + \varepsilon_i \quad , \quad \beta_0, \beta_1 \in \mathbb{R}, \quad \varepsilon_i \sim N(\mu, \sigma^2)$

(b)

$$\text{Slope: } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$y\text{-intercept: } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

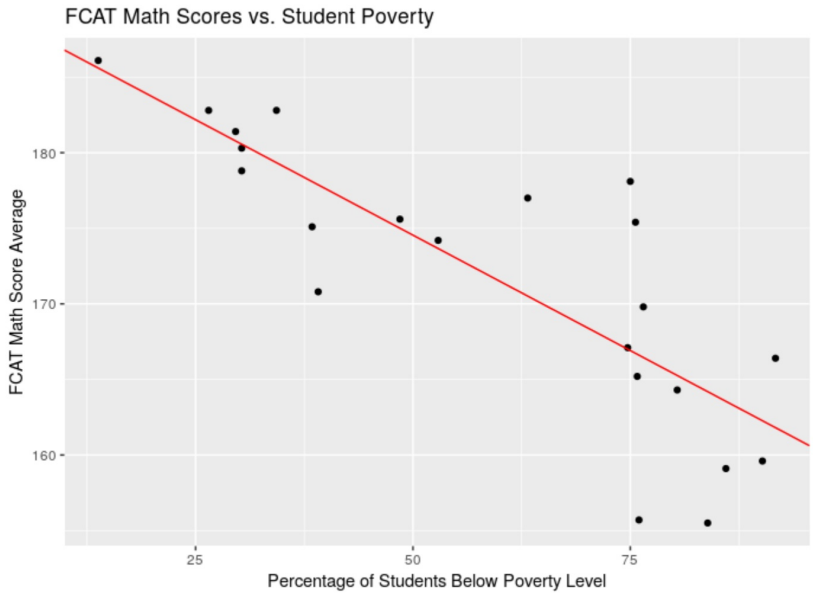
```
> # 5)
> data <- read.csv("FCATtxt.csv")
> xbar <- mean(data$POVERTY)
> ybar <- mean(data$MATH)
> SSXY <- sum((data$POVERTY - xbar) * (data$MATH - ybar))
> SSXX <- sum((data$POVERTY - xbar)^2)
> b_1 = SSXY / SSXX
> b_0 = ybar - b_1 * xbar
> c(b_0, b_1)
[1] 189.8158233  -0.3054445
> lm(data$MATH ~ data$POVERTY)

Call:
lm(formula = data$MATH ~ data$POVERTY)

Coefficients:
 (Intercept)  data$POVERTY
    189.8158       -0.3054
```

(c)

```
ggplot(data, aes(x = data$POVERTY, y = data$MATH)) +
  geom_point() +
  geom_abline(slope = b_1, intercept = b_0, color = "red") +
  xlab("Percentage of Students Below Poverty Level") +
  ylab("FCAT Math Score Average") +
  ggtitle("FCAT Math Scores vs. Student Poverty")
```



(d) $\hat{\beta}_0$: math score avg expected if poverty were at 0

$\hat{\beta}_1$: how much scores decrease per unit increase in poverty percentage.

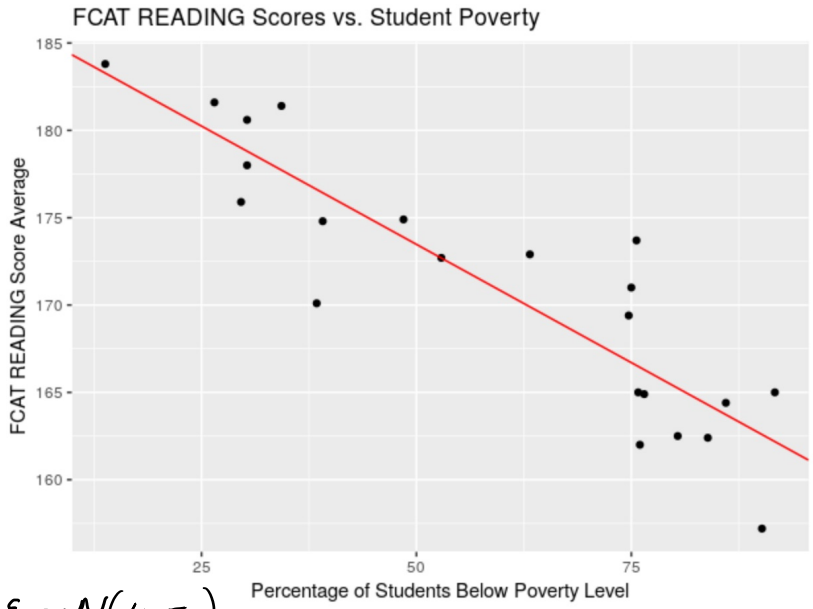
(e)

```
> xbar <- mean(data$POVERTY)
> ybar <- mean(data$READING)
> SSXY <- sum((data$POVERTY - xbar) * (data$READING - ybar))
> SSXX <- sum((data$POVERTY - xbar)^2)
> b_1 = SSXY / SSXX
> b_0 = ybar - b_1 * xbar
> c(b_0, b_1)
[1] 187.0126192  -0.2708112
> lm(data$READING ~ data$POVERTY)

Call:
lm(formula = data$READING ~ data$POVERTY)

Coefficients:
 (Intercept)  data$POVERTY
    187.0126       -0.2708
```

```
ggplot(data, aes(x = data$POVERTY, y = data$READING)) +
  geom_point() +
  geom_abline(slope = b_1, intercept = b_0, color = "red") +
  xlab("Percentage of Students Below Poverty Level") +
  ylab("FCAT READING Score Average") +
  ggtitle("FCAT READING Scores vs. Student Poverty")
```



$y = \beta_0 + \beta_1 x + \varepsilon_i \quad , \quad \beta_0, \beta_1 \in \mathbb{R}, \quad \varepsilon_i \sim N(\mu, \sigma^2)$

$\hat{\beta}_0$: math score avg expected if poverty were at 0

$\hat{\beta}_1$: how much scores decrease per unit increase in poverty percentage.