Name: ID:

Homework for 4/9 Due 4/16

1. [§13-6] It is conventional wisdom in military squadrons that pilots tend to father more girls than boys. Snyder (1961) gathered data for military fighter pilots. The sex of the pilots' offspring were tabulated for three kinds of flight duty during the month of conception, as shown in the following table.

Father's Activity	Female Offspring	Male Offspring
Flying Fighters	51	38
Flying Transports	14	16
Not Flying	38	46

   a. Is there any significant difference between the three groups? Use $\alpha = 0.05$.

   b. In the United States in 1950, 105.37 males were born for every 100 females. Are the data consistent with this sex ratio? Use $\alpha = 0.05$. (*Hint*: this is similar to the authorship example. We are comparing pilots with general males. Thus we need to combine all pilots in order to make a comparison.)

   a. First, we can find the totals:

Father's Activity	Female Offspring	Male Offspring	Total
Flying Fighters	51	38	89
Flying Transports	14	16	30
Not Flying	38	46	84
Total	103	100	203

   Let the null hypothesis be that there is no difference between these groups. Then if H_0 is true, the expected counts for these groups would be

Father's Activity	Female Offspring	Male Offspring
Flying Fighters	$\frac{89 \cdot 103}{203} = 45.2$	$\frac{89 \cdot 100}{203} = 43.8$
Flying Transports	$\frac{30 \cdot 103}{203} = 15.2$	$\frac{30 \cdot 103}{203} = 14.8$
Not Flying	$\frac{84 \cdot 103}{203} = 42.6$	$\frac{84 \cdot 100}{203} = 41.4$

   Thus the Pearson's χ^2 test statistic is

$$X^2 = \frac{(51-45.2)^2}{45.2} + \frac{(38-43.8)^2}{43.8} + \frac{(14-15.2)^2}{15.2} + \frac{(16-14.8)^2}{14.8} + \frac{(38-42.6)^2}{42.6} + \frac{(46-41.4)^2}{41.4} = 2.712.$$

Moreover, the distribution of X^2 is approximately χ^2_2 ($df = (2-1)(3-1) = 2$). Since $\alpha = 0.05$, the rejection region is $R = \{X^2 > 5.99\}$ ($5.99 = \chi^2_{0.95,2}$).

Since $2.712 < 5.99$, we do not reject H_0. In other words, there is no significance difference between the groups.

b. In this part, we compare pilots with general males. Thus we combine the data for pilots.

Father	Female Offspring	Male Offspring	Total
Pilot	103	100	203
General male	100	105.37	205.37
Total	203	205.37	408.37

Let the null hypothesis be that there is no difference between pilots and general males. Then if H_0 is true, the expected counts for pilots and general males would be

Father	Female Offspring	Male Offspring
Pilot	100.911	102.089
General male	102.089	103.281

And the Pearson's χ^2 test statistic is $X^2 = 0.1709$.

The distribution of the test statistic is approximately χ^2_1 ($df = (2-1)(2-1) = 1$). Since $\alpha = 0.05$, the rejection region is $R = \{X^2 > 3.84\}$ ($3.84 = \chi^2_{0.95,1}$).

Since $0.1709 < 3.84$, we do not reject H_0. In other words, there is no significance difference between pilots and general males, that is, the data is consistent with this sex ratio.

2. [§13-16] A market research team conducted a survey to investigate the relationship of personality to attitude toward small cars. A sample of 299 adults in a metropolitan area were asked to fill out a 16-item self-perception questionnaire, on the basis of which they were classified into three types: cautious conservative, middle-of-the-roader, and confident explorer. They were then asked to give their overall opinion of small cars: favorable, neutral, or unfavorable. Is there a relationship between personality type and attitude toward small cars? Use $\alpha = 0.05$.

	Personality Type		
Attitude	Cautious	Midroad	Explorer
Favorable	79	58	49
Neutral	10	8	9
Unfavorable	10	34	42

We first find the totals.

	Personality Type			
Attitude	Cautious	Midroad	Explorer	Total
Favorable	79	58	49	186
Neutral	10	8	9	27
Unfavorable	10	34	42	86
Total	99	100	100	299

Let the null hypothesis be that there is no relationship, that is, "Personality" and "Attitude" are independent. Then if H_0 is true, the expected counts for pilots and general males would be

	Personality Type		
Attitude	Cautious	Midroad	Explorer
Favorable	$\frac{186 \cdot 99}{299} = 61.6$	$\frac{186 \cdot 100}{299} = 62.2$	$\frac{186 \cdot 100}{299} = 62.2$
Neutral	$\frac{27 \cdot 99}{299} = 8.9$	$\frac{27 \cdot 100}{299} = 9$	$\frac{27 \cdot 100}{299} = 9$
Unfavorable	$\frac{86 \cdot 99}{299} = 28.5$	$\frac{86 \cdot 100}{299} = 28.8$	$\frac{86 \cdot 100}{299} = 28.8$

And the Pearson's χ^2 test statistic is $X^2 = 27.24$.

The distribution of the test statistic is approximately χ^2_4 ($df = (3-1)(3-1) = 4$). Since $\alpha = 0.05$, the rejection region is $R = \{X^2 > 9.49\}$ ($9.49 = \chi^2_{0.95,4}$).

Since $27.24 > 9.49$, we reject H_0. In other words, there is some relationship between personality type and attitude toward small cars.

3. [§14-2] For the following data:

x	.34	1.38	−.65	.68	1.40	−.88	−.30	−1.18	.50	−1.75
y	.27	1.34	−.53	.35	1.28	−.98	−.72	−.81	.64	−1.59

a. Fit a line $y = a + bx$ by the method of least squares.

b. Fit a line $x = c + dy$ by the method of least squares.

a. We have

$$\sum_{i=1}^{n} x_i = -0.46, \quad \sum_{i=1}^{n} x_i^2 = 10.434, \quad \sum_{i=1}^{n} x_i y_i = 9.452,$$
$$\sum_{i=1}^{n} y_i^2 = 8.983, \quad \sum_{i=1}^{n} y_i = -0.75 \quad \text{and} \quad n = 10.$$

Thus, by the method of least squares

$$\begin{aligned} a &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{(10.434)\cdot(-0.75) - (-0.46)\cdot 9.452}{10\cdot 10.434 - (-0.46)^2} \\ &= -0.0334, \\ b &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{10\cdot 9.452 - (-0.46)\cdot(-0.75)}{10\cdot 10.434 - (-0.46)^2} \\ &= 0.904. \end{aligned}$$

Thus the line is

$$y = 0.904x - 0.0334.$$

b. We interchange the role of x and y.

$$\begin{aligned}
c &= \frac{\left(\sum_{i=1}^{n} y_i^2\right)\left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} y_i x_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \\
&= \frac{(8.983)\cdot(-0.46) - (-0.75)\cdot 9.452}{10\cdot 8.983 - (-0.75)^2} \\
&= 0.0331, \\
b &= \frac{n\sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \\
&= \frac{10\cdot 9.452 - (-0.75)\cdot(-0.46)}{10\cdot 8.983 - (-0.75)^2} \\
&= 1.055.
\end{aligned}$$

Thus the line is

$$x = 1.055y + 0.0331.$$

4. [§14-10] Show that the least squares estimates of the slope and intercept of a line may be expressed as

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

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

(*Hint:* begin with $\hat{\beta}_1$ and expand $(x_i - \bar{x})(y_i - \bar{y})$ and $(x_i - \bar{x})^2$.)

We will use the following identities several times:

$$\sum_{i=1}^{n} x_i = n\bar{x} \quad \text{and} \quad \sum_{i=1}^{n} y_i = n\bar{y}.$$

First we have

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
&= \frac{1}{n}\left(n\sum_{i=1}^{n} x_i y_i - (n\bar{x})\cdot(n\bar{y})\right) \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&=\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2 \\
&=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2 \\
&=\frac{1}{n}\left(n\sum_{i=1}^{n}x_i^2-(n\bar{x})^2\right) \\
&=\frac{1}{n}\left[\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right].
\end{aligned}$$

(In fact, we can save the calculation by using the first result and replace y with x.)

Therefore,

$$\begin{aligned}
\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} &= \frac{\frac{1}{n}\left[\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\right]}{\frac{1}{n}\left[\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]} \\
&= \frac{\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\hat{\beta}_1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\hat{\beta}_0 &= \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\left(\sum_{i=1}^n y_i\right) + \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\left[\left(\sum_{i=1}^n x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right]\left(\sum_{i=1}^n y_i\right) - \left[\left(\sum_{i=1}^n x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)\right]\left(\sum_{i=1}^n x_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\left[\left(\sum_{i=1}^n x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right](n\bar{y}) - \left[\left(\sum_{i=1}^n x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)\right](n\bar{x})}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \bar{y} - \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\bar{x} \\
&= \bar{y} - \hat{\beta}_1\bar{x}.
\end{aligned}$$

(One can also go backwards.)

Name: ID:

Homework for 4/11 Due 4/16

1. [§13-17] Let X and Y be random variables with

$$\begin{aligned} &\mathbb{E}[X] = \mu_x && \mathbb{E}[Y] = \mu_y \\ &\mathrm{Var}[X] = \sigma_x^2 && \mathrm{Var}[Y] = \sigma_y^2 \\ &\mathrm{Cov}[X, Y] = \sigma_{xy} \end{aligned}$$

Consider predicting Y from X as $\hat{Y} = \alpha + \beta X$, where α and β are chosen to minimize $\mathbb{E}[(Y - \hat{Y})^2]$, the expected squared prediction error.

a. Show that the minimizing values of α and β are

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2} \qquad \alpha = \mu_y - \beta\mu_x$$

(*Hint*: $\mathbb{E}[(Y - \hat{Y})^2] = (\mathbb{E}[Y] - \mathbb{E}[\hat{Y}])^2 + \mathrm{Var}[Y - \hat{Y}]$. Section 4.3 may be helpful. Especially, Theorem A, Corollary A, and Corollary B.)

b. Show that for this choice of α and β

$$\frac{\mathrm{Var}[Y] - \mathrm{Var}[Y - \hat{Y}]}{\mathrm{Var}[Y]} = r_{xy}^2,$$

where r_{xy} is correlation between X and Y:

$$r_{xy} = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

a. First we have

$$\begin{aligned} \mathbb{E}[\hat{Y}] &= \mathbb{E}[\alpha + \beta X] = \alpha + \beta\mathbb{E}[X] = \alpha + \beta\mu_x, \\ \mathrm{Var}[\hat{Y}] &= \mathrm{Var}[\alpha + \beta X] = \beta^2\mathrm{Var}[X] = \beta^2\sigma_x^2, \quad \text{and} \\ \mathrm{Cov}[Y, \hat{Y}] &= \mathrm{Cov}[Y, \alpha + \beta X] = \beta\mathrm{Cov}[Y, X] = \beta\sigma_{xy}. \end{aligned}$$

Thus

$$\mathrm{Var}[Y - \hat{Y}] = \mathrm{Var}[Y] + \mathrm{Var}[\hat{Y}] - 2\mathrm{Cov}[Y, \hat{Y}] = \sigma_y^2 + \beta^2\sigma_x^2 - 2\beta\sigma_{xy}.$$

In order to minimize $\mathbb{E}[(Y - \hat{Y})^2]$, we notice that

$$\begin{aligned} \mathbb{E}[(Y - \hat{Y})^2] &= (\mathbb{E}[Y] - \mathbb{E}[\hat{Y}])^2 + \mathrm{Var}[Y - \hat{Y}] \\ &= (\mu_y - \alpha - \beta\mu_x)^2 - (\sigma_y^2 + \beta^2\sigma_x^2 - 2\beta\sigma_{xy}), \\ &= f(\alpha, \beta). \end{aligned}$$

Furthermore,

$$\begin{aligned}
\frac{\partial f}{\partial \alpha} &= -2(\mu_y - \alpha - \beta\mu_x), \\
\frac{\partial f}{\partial \beta} &= -2\mu_x(\mu_y - \alpha - \beta\mu_x) - (2\sigma_x^2\beta - 2\sigma_{xy}) \\
\frac{\partial^2 f}{\partial \alpha^2} &= 2, \qquad \frac{\partial^2 f}{\partial \beta^2} = 2\mu_x^2 - 2\sigma_x^2, \quad \text{and} \\
\frac{\partial^2 f}{\partial \alpha \partial \beta} &= \frac{\partial^2 f}{\partial \beta \partial \alpha} = 2\mu_x.
\end{aligned}$$

The solution to

$$\begin{cases} -2(\mu_y - \alpha - \beta\mu_x) = 0 \\ -2\mu_x(\mu_y - \alpha - \beta\mu_x) - (2\sigma_x^2\beta - 2\sigma_{xy}) = 0 \end{cases}$$

is

$$\alpha = \mu_y - \beta\mu_x \qquad \beta = \frac{\sigma_{xy}}{\sigma_x^2}.$$

Since

$$\begin{vmatrix} \frac{\partial^2 f}{\partial \alpha^2} & \frac{\partial^2 f}{\partial \alpha \partial \beta} \\ \frac{\partial^2 f}{\partial \beta \partial \alpha} & \frac{\partial^2 f}{\partial \beta^2} \end{vmatrix} = \begin{vmatrix} 2 & 2\mu_x \\ 2\mu_x & 2\mu_x^2 - 2\sigma_x^2 \end{vmatrix} = 4\sigma_x^2 > 0,$$

we see that $f(\alpha, \beta)$ achieve its minimum at

$$\alpha = \mu_y - \beta\mu_x \qquad \beta = \frac{\sigma_{xy}}{\sigma_x^2}.$$

b. From (a), we immediately have

$$\begin{aligned}
\frac{\text{Var}[Y] - \text{Var}[Y - \hat{Y}]}{\text{Var}[Y]} &= \frac{\sigma_y^2 - (\sigma_y^2 + \beta^2\sigma_x^2 - 2\beta\sigma_{xy})}{\sigma_y^2} = -\beta^2\frac{\sigma_x^2}{\sigma_y^2} + 2\beta\frac{\sigma_{xy}}{\sigma_y^2} \\
&= -\left(\frac{\sigma_{xy}}{\sigma_x^2}\right)^2 \cdot \frac{\sigma_x^2}{\sigma_y^2} + 2 \cdot \frac{\sigma_{xy}}{\sigma_x^2} \cdot \frac{\sigma_{xy}}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} \\
&= r_{xy}^2.
\end{aligned}$$

2. [§13-18] Suppose that

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, n$$

where the e_i are independent and normally distributed with mean zero and variance σ^2. Find the mle's of β_0 and β_1 and verify that they are the least squares estimates. (Hint: Under these assumptions, the Y_i are independent and normally distributed with means $\beta_0 + \beta_1 x_i$ and variance σ^2. Write the joint density function of the Y_i and thus the likelihood.)

Since e_i's are i.i.d. $N(0,1)$ random variables, we have $Y_i \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$ and Y_i's are independent. Let $f_i(y_i)$ be the pdf of Y_i. We have

$$f_i(y_i|\beta_0, \beta_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}.$$

Since Y_i's are independent, the joint pdf of $Y_1, \ldots, Y_n$ is

$$f(y_1, \ldots, y_2|\beta_0, \beta_1) = \prod_{i=1}^{n} f_i(y_i|\beta_0, \beta_1).$$

Correspondingly, the likelihood function and log-likelihood function are

$$\begin{aligned}
\text{lik}(\beta_0, \beta_1) &= \prod_{i=1}^{n} f_i(\beta_0, \beta_1|Y_i), \quad \text{and} \\
l(\beta_0, \beta_1) &= \log \text{lik}(\beta_0, \beta_1) = \sum_{i=1}^{n} \log f_i(\beta_0, \beta_1|Y_i) \\
&= \sum_{i=1}^{n} \left(-\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}[Y_i - (\beta_0 + \beta_1 x_i)]^2\right) \\
&= -\sum_{i=1}^{n} \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2 \\
&= -\sum_{i=1}^{n} \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} S(\beta_0, \beta_1),
\end{aligned}$$

where $S(\beta_0, \beta_1) = \sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2$. It follows that the minimizer of $S(\beta_0, \beta_1)$ is the maximizer of $l(\beta_0, \beta_1)$. Therefore, the mle for β_0 and β_1

are the least square estimates:

$$\begin{cases} \hat{\beta}_0 &= \dfrac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \hat{\beta}_1 &= \dfrac{n\sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{cases}.$$