Department of Economics UN3412
Columbia University Fall 2021

SOLUTIONS to Problem Set 1
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw
For all sections

"Calculator" was once a job description. This problem set gives you an opportunity to do some calculations on the relation between smoking and lung cancer, using a (very) small sample of five countries. The purpose of this exercise is to illustrate the mechanics of ordinary least squares (OLS) regression. You will calculate the regression "by hand" using formulas from class and the textbook. For these calculations, you may relive history and use long multiplication, long division, and tables of square roots and logarithms; or you may use an electronic calculator or a spreadsheet.
The data are summarized in the following table. The variables are per capita cigarette consumption in 1930 (the independent variable, "*X*") and the death rate from lung cancer in 1950 (the dependent variable, "*Y*"). The cancer rates are shown for a later time period because it takes time for lung cancer to develop and be diagnosed.

Observation #	Country	Cigarettes consumed per capita in 1930 (*X*)	Lung cancer deaths per million people in 1950 (*Y*)
1	Switzerland	530	250
2	Finland	1115	350
3	Great Britain	1145	465
4	Canada	510	150
5	Denmark	380	165

Source: Edward R. Tufte, *Data Analysis for Politics and Management*, Table 3.3.

1. (40p) Use a calculator, a spreadsheet, or "by hand" methods to compute the following; refer to the textbook for the necessary formulas. (*Note*: please do not use functions like 'Correl' in Excel, we want you to calculate sample slope with the formula, if you use a spreadsheet, attach a printout)

(a) (5p) The sample means of *X* and *Y*, $\bar{X}$ and $\bar{Y}$.
$\bar{X} = 736,\ \bar{Y} = 276$

(b) (6p) The standard deviations of *X* and *Y*, s_X and s_Y.
$s_X = 364.41,\ s_Y = 132.35$

(c) (6p) The correlation coefficient, *r*, between *X* and *Y*.
$r = 0.92$

(d) (6p) $\hat{\beta}_1$, the OLS estimated slope coefficient from the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$
$\hat{\beta}_1 = 0.336418$

(e) (5p) $\hat{\beta}_0$, the OLS estimated intercept term from the same regression.
$\hat{\beta}_0 = 28.39656$

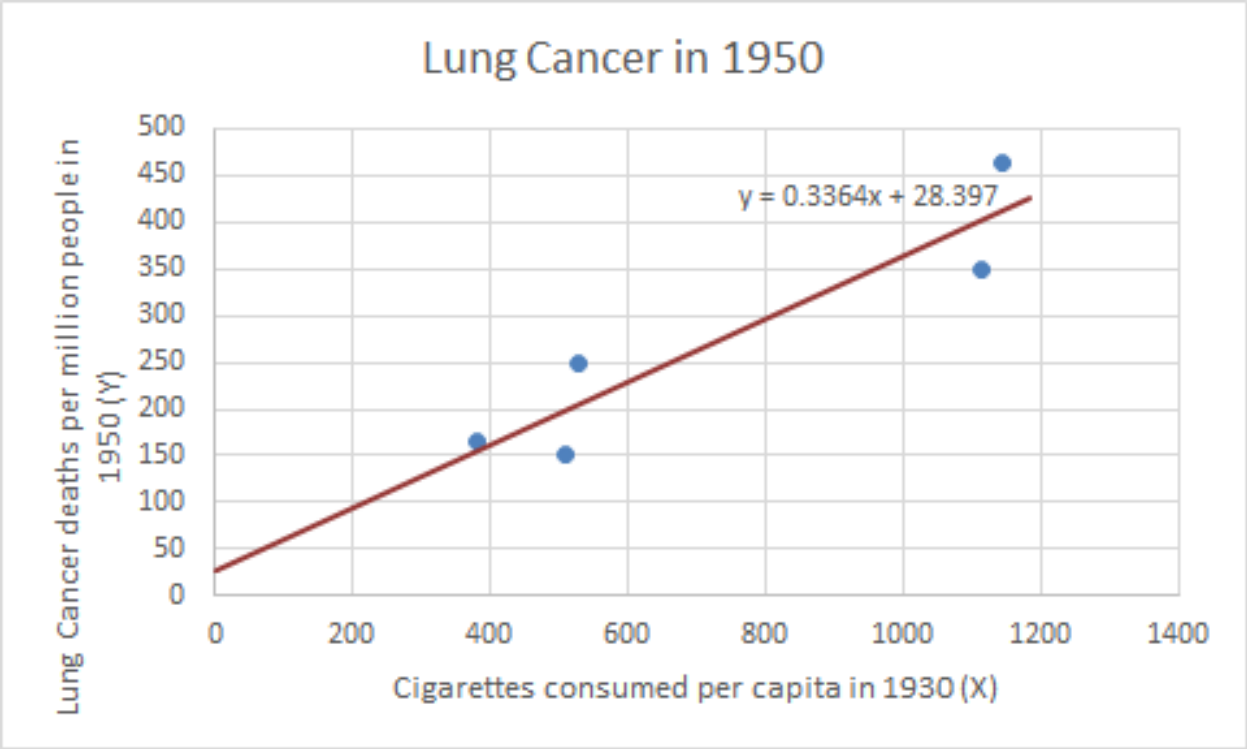
(f) (6p) $\hat{Y}_i$, $i = 1,\ldots, n$, the predicted values for each country from the regression

Switzerland	*206.6981*
France	*403.5026*
GreatBritain	*413.5952*
Canada	*199.9697*
Denmark	*156.2354*

(g) (6p) $\hat{u}_i$, the OLS residual for each country.

Switzerland	*43.3019*
France	*-53.5026*
GreatBritain	*51.40483*
Canada	*-49.9697*
Denmark	*8.7646*

2. (10p) On graph paper or using a spreadsheet, graph the scatterplot of the five data points and the regression line. Be sure to label the axes, the data points, the residuals, and the slope and intercept of the regression line.



3. (30p) Suppose that you were told that average hourly wages in State A and State B are different. However, you have your own suspicion that it could be the same. You suggest to your brother, who is attending a school in State B to collect data on hourly wage as part of your project in ECON3412. You do the same for State A as your school is located in State A. The accompanying table shows your and your brother's findings.

Average Hourly Wages in State A and State B, in dollar

State A			State B		
$\overline{Y}_{State\ A}$	$s_{State\ A}$	$n_{State\ A}$	$\overline{Y}_{State\ B}$	$s_{State\ B}$	$n_{State\ B}$
57.8	3.9	155	58.7	4.2	157

(a) (6p) Let your null hypothesis be that there is no difference in the average hourly wage in the two states. Specify the alternative hypothesis.
(b) (6p) Find the difference in average hourly wage and the standard error of the difference.
(c) (6p) Generate a 95% confidence interval for the difference in average hourly wage.
(d) (6p) Calculate the *t*-statistic for comparing the two means. Is the difference statistically significant at the 1% level? Which critical value did you use? Why would this number be smaller if you had assumed a one-sided alternative hypothesis? What is the intuition behind this?
(e) (6p) What about at 5% level?

Solution:

(a) $H_0 = \mu_{State\ A} - \mu_{State\ B} = 0$ *and* $H_0 = \mu_{State\ A} - \mu_{State\ B} \neq 0$

(b) $Y_{State\ A}$ - $Y_{State\ B}$ = -0.9 , SE($Y_{State\ A}$ - $Y_{State\ B}$) = $\sqrt{\frac{3.9^2}{155} + \frac{4.2^2}{157}}$ = 0.459.

(c) -0.9 ± 1.96 × 0.459 = (-0.00036, -1.79964).

(d) (d) $t = \frac{\bar{Y}_{State\ A} - \bar{Y}_{State\ B}}{SE(\bar{Y}_{State\ A} - \bar{Y}_{State\ B})}$ = -1.9617, so $|t|$< 2.58, which is the critical value at the 1% level. Hence you cannot reject the null hypothesis. That means you are right to have suspicion.

- The critical value for the one-sided hypothesis would have been 2.33.
- Assuming a one-sided hypothesis implies that you have some information about the problem at hand, and, as a result, can be more easily convinced than if you had no prior expectation.

(e) At 5% significance level, the critical value is $|t|$ < 1.96. Hence you can reject the null hypothesis. In this case your suspicion is wrong as wage in State B is significantly higher than in State A.

4. (20p) Suppose that the following table is the joint probability distribution of two random variables X and Y:

		X			
		x = -2	**x = 0**	**x = 2**	**x = 3**
Y	**y = 2**	**0.27**	**0.08**	**0.16**	**0.2**
	y = 5	**0.1**	**0.04**	**0.1**	**0.05**

(a) (4p) Find the marginal PDF of X when x=-2, 0, 2, and 3.
(b) (4p) Find the marginal PDF of Y when y=2 and 5.
(c) (4p) Find the conditional PDF of x=-2 and 3 given that y=2 has occurred.
(d) (4p) Find the conditional PDF of y=2 and 5 given that x=3 has occurred.
(e) (4p) State conditional PDF in terms of Joint and marginal PDFs.

Solution:

a. $f(x=-2)=\sum_y f(x,y)=0.27+0.1=0.37$; $f(x=0)=0.12$; $f(x=2)=0.26$; $f(x=3)=0.25$
b. $f(y=2)=\sum_x f(x,y)=0.27+0.08+0.16+0.2=0.71$;
$f(y=5)=\sum_x f(x,y)=0.1+0.04+0.1+0.05=0.29$
c. $f(x|y)=\frac{f(x,y)}{f(y)}=f(x=-2|y=2)=\frac{f(x=-2,y=2)}{f(y=2)}=\frac{0.27}{0.71}=0.38$; $f(x=3|y=2)=\frac{f(x=3,\ y=2)}{f(y=2)}=\frac{0.20}{0.71}=0.28$
d. $f(y|x)=\frac{f(x,y)}{f(x)}=f(y=2|x=3)=\frac{f(x=3,\ y=2)}{f(x=3)}=\frac{0.20}{0.25}=0.80$; $f(y=5|x=3)=\frac{f(x=3,\ y=5)}{f(x=3)}=\frac{0.05}{0.25}=0.20$
e. The Conditional PDF of one variable can be expressed as the ratio of the Joint PDF to the Marginal PDF of another (conditioning) variable.

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

1. [Practice question, not graded] SW 2.3

	Rain (X=0)	No Rain (X=1)	Total
Long Commute (Y=0)	0.15	0.07	0.22
Short Commute (Y=1)	0.15	0.63	0.78
Total	0.30	.70	1.00

Using the random variables X and Y from Table 2.2 (given above), consider two new random variables W = 3 + 6X and V = 20 – 7Y. Compute:

a) E(W) and E(V).

b) σ^2_W and σ^2_V.

c) $\sigma_{W,V}$ and Corr(W,V).

Solution:

(a) E(V) = E(20-7Y) = 20 – 7E(Y) = 20 – 7 × 0.78 = 14.54
E(W) = E(3+6X) = 3 + 6E(X) = 3 + 6 × 0.70 = 7.2
(b) Var(W) = var(3+6X) = 6^2 Var(X) = 36 × 0.21 = 7.56
Var(V) = var(20-7Y) = $(-7)^2$ Var(Y) = 49 × 0.1716 = 8.4084
(c) Cov(W,V) = Cov(3+6X,20-7Y) = 6 (-7)Cov(X,Y) = -42 × 0.084 = -3.528
*Corr(W,V) = -3.528/√(7.56*8.4084)*

2. [Practice question, not graded] SW 2.6

The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working age US population, based on the 1990 US Census.

	Unemployed (Y=0)	Employed (Y=1)	Total
Non-college grads (X=0)	0.045	0.709	0.754
College grads (X=1)	0.005	0.241	0.246
Total	0.050	0.950	1.000

a. Compute E(Y).

b. The unemployment rate is the fraction of the labor force that is unemployed. Show that the unemployment rate is given by 1-E(Y).

c. Calculate the E(Y|X=1) and E(Y|X=0).

d. Calculate the unemployment rate for (i) college graduates and (ii) non-college graduates.

e. A randomly selected member of this population reports being unemployed. What is the probability that this worker is a college graduate? A non-college graduate?

f. Are educational achievement and employment status independent? Explain.

Solution:

(a) E(Y) = 0 × Pr(Y=0) + 1× Pr(Y=1) = 0 × 0.05 + 1 × 0.095 = 0.95
(b) Unemployment Rate = #(unemployed)/ # (labor force)
= Pr(Y=0) = 1 - Pr(Y=1) = 1 - EY
(c) We calculate the conditional probabilities first:
Pr(Y=0|X=0) = Pr(X=0 & Y=0)/Pr(X=0) = 0.045/0.754 = 0.0597
Pr(Y=1|X=0) = Pr(X=0 & 1=0)/Pr(X=0) = 0.709/0.754 = 0.9403
Pr(Y=0|X=1) = Pr(X=1 & Y=0)/Pr(X=1) = 0.005/0.246 = 0.0203
Pr(Y=1|X=1) = Pr(X=1 & Y=1)/Pr(X=1) = 0.241/0.246 = 0.9797
The conditional expectations are:
E(Y|X=1) = 0 × Pr(Y=0|X=1) + 1 × Pr(Y=1|X=1)
= 0 × 0.0203 + 1 × 0.9797 = 0.9797
E(Y|X=0) = 0 × Pr(Y=0|X=0) + 1 × Pr(Y=1|X=0)
= 0 × 0.0597+ 1 × 0.9403 = 0.9403
(d) Use the Solution to part (b)
Unemployment rate for college grads
= 1 – E(Y | X=1) = 1-0.9797 = 0.0203
Unemployment rate for non-college grads
= 1 – E(Y | X=0) = 1-0.9403= 0.0597
(e) The probability that a randomly selected workers who is reported being unemployed is a college graduate is
Pr(X=1|Y=0) = Pr(X=1 & Y=0)/Pr(Y=0) = 0.005/0.050 = 0.1
The probability that this worker is a non college graduate is
Pr(X=0|Y=0) = 1 – Pr(X=1|Y=0) = 1 - 0.1 = 0.9

(f) Educational achievement and employment status are not independent because they do not satisfy that, for all values of x and y,

$$\Pr(Y = y|X = x) = \Pr(Y = y).$$

For example,

$$\Pr(Y = 0|X = 0) = 0.0597 \neq \Pr(Y = 0) = 0.050.$$

3. [Practice question, not graded] SW 2.14 [Hint: Use SW Appendix Table 1.]

In a population E[Y] = 100 and Var(Y) = 43. Use the central limit theorem to answer the following questions:

a. In a random sample of size n = 100, find Pr($\overline{Y}$ ≤101)

b. In a random sample of size n = 165, find Pr($\overline{Y}$ >98)

c. In a random sample of size n = 64, find Pr(101 ≤ $\overline{Y}$ ≤103)

Solution:

a. In a random sample of size n = 100, find Pr(Ā<=101)
Var(Ā) = 43/100 = 0.43 so
Pr(Ā<=101) = Pr(Ā -100/√0.43 ≤ (101-100)/ √0.43)
= Φ(1.525) = 0..9364
b. In a random sample of size n = 165, find Pr(Ā>98)
Var(Ā) = 43/165 = 0.2606
Pr(Ā>98) = 1 - Pr(Ā≤98) = 1- Pr(Ā -100/√0.2606 ≤ (98-100)/ √0.2606)
= 1 - Φ(-3.9178) = Φ(3.9178) ≈ 1
c. In a random sample of size n = 64, find Pr(101 <= Ā <=103)
Var(Ā) = 43/64 = 0.6719
Pr(101 <= Ā <=103)
= Pr ((101-100)/√0.6719 ≤ (Ā-100)/ √0.6719 ≤ (101-100)/√0.6719)
≈ Φ(3.6599) - Φ(1.22) = 0.9999 – 0.8888 = 0.1111

4. [Practice question, not graded] SW 3.12

To investigate possible gender discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of the resulting monthly salaries are:

	Avg. Salary ($\overline{Y}$)	Stand Dev (of Y)	n
Men	$3100	$200	100
Women	$2900	$320	64

a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that wages of men and women are different? (To answer

this question, first state the null and alternative hypothesis; second, compute the relevant t-statistic; and finally, use the p-value to answer the equation.)

b. Do these data suggest that the firm is guilty of gender discrimination in its compensation politics? Explain.

Solution:

The standard error of $\bar{Y}_1 - \bar{Y}_2$ *is* $SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{200^2}{100} + \frac{320^2}{64}} = 44.721.$

(a) The hypothesis test for the difference in mean monthly salaries is

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 \neq 0.$$

The t-statistic for testing the null hypothesis is

$$t^{act} = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{SE}(\bar{Y}_1 - \bar{Y}_2)} = \frac{3100 - 2900}{44.721} = 4.4722.$$

Use Equation (3.14) in the text to get the p-value:

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-4.4722) = 2 \times (3.8744 \times 10^{-6}) = 7.7488 \times 10^{-6}.$$

The extremely low level of p-value implies that the difference in the monthly salaries for men and women is statistically significant. We can reject the null hypothesis with a high degree of confidence.

(b) From part (a), there is overwhelming statistical evidence that mean earnings for men differ from mean earnings for women. To examine whether there is gender discrimination in the compensation policies, we take the following one-sided alternative test

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 > 0.$$

With the t-statistic $t^{act} = 4.4722$, *the p-value for the one-sided test is:*

$$p\text{-value} = 1 - \Phi(t^{act}) = 1 - \Phi(4.4722) = 1 - 0.999996126 = 3.874 \times 10^{-6}.$$

With the extremely small p-value, the null hypothesis can be rejected with a high degree of confidence. There is overwhelming statistical evidence that mean earnings for men are greater than mean earnings for women. However, by itself, this does not imply gender discrimination by the firm. Gender discrimination means that two workers, identical in every way but gender, are paid different wages. The data description suggests that some care has been taken to make sure that workers with similar jobs are being compared. But, it is also important to control for characteristics of the workers that may affect their productivity (education, years of experience, etc.). If these characteristics are systematically different between men and women, then they may be responsible for the difference in mean wages. (If this is true, it raises an interesting and important question of why women tend to have less education or less experience than men, but that is a question about something other than gender discrimination by this firm.) Since these characteristics are not controlled for in the statistical analysis, it is premature to reach a conclusion about gender discrimination.

5. [Practice question, not graded] SW 2.10 [Hint: Use SW Appendix Table 1.]

Compute the following probabilities:

a. If Y is distributed N(1,4), find Pr(Y≤3).

b. If Y is distributed N(3,9), find Pr(Y>0).

c. If Y is distributed N(50,25), find Pr(40≤Y≤52).

d. If Y is distributed N(5,2), find Pr(6≤Y≤8)

Solution:

Using the fact that if $Y \sim N(\mu_Y, \sigma_Y^2)$ *then* $\frac{Y-\mu_Y}{\sigma_Y} \sim N(0,1)$ *and Appendix Table 1, we have*

(a)

$$\Pr(Y \le 3) = \Pr\left(\frac{Y-1}{2} \le \frac{3-1}{2}\right) = \Phi(1) = 0.8413.$$

(b)

$$\begin{aligned}\Pr(Y > 0) &= 1 - \Pr(Y \le 0) \\ &= 1 - \Pr\left(\frac{Y-3}{3} \le \frac{0-3}{3}\right) = 1 - \Phi(-1) = \Phi(1) = 0.8413.\end{aligned}$$

(c)

$$\begin{aligned}\Pr(40 \le Y \le 52) &= \Pr\left(\frac{40-50}{5} \le \frac{Y-50}{5} \le \frac{52-50}{5}\right) \\ &= \Phi(0.4) - \Phi(-2) = \Phi(0.4) - [1 - \Phi(2)] \\ &= 0.6554 - 1 + 0.9772 = 0.6326.\end{aligned}$$

(d)

$$\begin{aligned}\Pr(6 \le Y \le 8) &= \Pr\left(\frac{6-5}{\sqrt{2}} \le \frac{Y-5}{\sqrt{2}} \le \frac{8-5}{\sqrt{2}}\right) \\ &= \Phi(2.1213) - \Phi(0.7071) \\ &= 0.9831 - 0.7602 = 0.2229.\end{aligned}$$

6. [Practice question, not graded] SW 3.3

In a survey of 400 likely voters, 215 responded that they would vote for the incumbent and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely

voters that preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents that preferred the incumbent.

a. Use the survey results to estimate p.

b. Use the estimator of the variance of $\hat{p}$, $\hat{p}$ (1 - $\hat{p}$)/n to calculate the standard error of your estimator.

c. What is the p-value for the test H0: p=0.5 vs. H1:p≠0.5?

d. What is the p-value for the test H0: p=0.5 vs. H1:p>0.5?

e. Why do the results from (c) and (d) differ?

f. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

Solution:

(a) $\hat{p} = \frac{215}{400} = 0.5375.$

(b) $\text{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.5375 \times (1-0.5375)}{400} = 6.2148 \times 10^{-4}$. *The standard error is SE* $(\hat{p}) = (\text{var}(\hat{p}))^{\frac{1}{2}} = 0.0249.$

(c) The computed t-statistic is

$$t^{act} = \frac{\hat{p} - \mu_{p,0}}{\text{SE}(\hat{p})} = \frac{0.5375 - 0.5}{0.0249} = 1.506.$$

Because of the large sample size $(n = 400)$, *we can use Equation (3.14) in the text to get the p-value for the test* $H_0 : p = 0.5$ *vs.* $H_1 : p \neq 0.5$:

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-1.506) = 2 \times 0.066 = 0.132.$$

(d) Using Equation (3.17) in the text, the p-value for the test $H_0 : p = 0.5$ *vs.* $H_1 : p > 0.5$ *is*

$$p\text{-value} = 1 - \Phi(t^{act}) = 1 - \Phi(1.506) = 1 - 0.934 = 0.066.$$

(e) Part (c) is a two-sided test and the p-value is the area in the tails of the standard normal distribution outside ± (calculated t-statistic). Part (d) is a one-sided test and the p-value is the area under the standard normal distribution to the right of the calculated t-statistic.

(f) For the test $H_0 : p = 0.5$ *vs.* $H_1 : p > 0.5$, *we cannot reject the null hypothesis at the 5% significance level. The p-value 0.066 is larger than 0.05. Equivalently the calculated t-statistic* 1.506 *is less than the critical value 1.645 for a one-sided test with a 5% significance level. The test suggests that the survey did not contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey.*

7. Consider two events A and B with Pr(A) = 0.5 and Pr(B) = 0.9. Determine the maximum and minimum values of Pr(A ∪ B).

Maximum value that $Pr(A \cup B)$ can be is 1.

Based on relation that $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$, the minimum value of $Pr(A \cup B)$ occurs when maximum value of $Pr(A \cap B)$ occurs. Maximum value of $Pr(A \cap B)$ is when $A \subseteq B$ such that $A \cap B = A$. So, minimum is $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A) = 0.9$

$$0.9 \leq Pr(A \cup B) \leq 1$$

8. Assume that events A and B^c are independent. That is, $Pr(A \cap B^c) = Pr(A)Pr(B^c)$. Are events A and B also independent?

   We know that
   $Pr(A \cap Bc) = Pr(A)Pr(Bc)$,
   $Pr(B) + Pr(BC) = 1, and$
   $Pr(A \cap Bc) + Pr(A \cap B) = Pr(A)$.

   $Pr(A) * Pr(Bc) = Pr(A) * (1 - Pr(B)) = Pr(A) - Pr(A) * Pr(B)$
   From fact 3, $Pr(A \cap Bc) = Pr(A) - Pr(A \cap B)$

   Plugging these into our original equality, we get that
   $Pr(A) - Pr(A \cap B) = Pr(A) - Pr(A) * Pr(B)$, *which simplifies to*
   $Pr(A \cap B) = Pr(A) * Pr(B)$, *which proves that A and B are independent.*

9. Let X and Y denote two random variables.
   (i) Show that if at least one of X or Y has expectation equal to zero, then cov(X, Y) = E[XY].

   $Cov(X,Y) = E(XY) - E(X)E(Y)$, *so if either* $E(X)$ *or* $E(Y)$ *equals zero, then*
   $Cov(X,Y) = E(XY)$

   (ii) Show that $cov(\alpha X + a, \beta Y + b) = \alpha\beta cov(X, Y)$ for constants (α, β, a, b).

   $$\begin{aligned} cov(\alpha X + a, \beta Y + b) &= E(a + \alpha X - \mu a + \alpha X)(b + \beta Y - \mu b + \beta Y) \\ &= E(a + \alpha X - \alpha\mu X - a)(b + \beta Y - \beta\mu Y - b) \\ &= E(\alpha X - \alpha\mu X)(\beta Y - \beta\mu Y) \\ &= \alpha\beta E(X - \mu X)(Y - \mu Y) \\ &= \alpha\beta cov(X,Y) \end{aligned}$$

   $$cov(\alpha X + a, \beta Y + b) = \alpha\beta cov(X,Y)$$

   (iii) Show that $var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$.

   $Var(X + Y) = E[(X + Y)(X + Y)] - E2[X + Y]$

$$= E[X2 + 2XY + Y2] - (\mu X + \mu Y)2$$
$$= E[X2 + 2XY + Y2] - \mu X2 - 2\,\mu X\,\mu Y - \mu Y2$$
$$= (E[X2] - \mu X2) + (E[Y2] - \mu Y2) + 2(E[XY] - \mu X\,\mu Y)$$
$$= Var(X) + Var(Y) + 2Cov(X,Y)$$

by definitions of $Var(X), Var(Y),$ *and* $Cov(X,Y)$

(iv) Show that the population correlation coefficient is always between [−1, 1].

The Cauchy-Schwarz inequality states that $Cov^2(X,Y) \le Var(X)Var(Y)$.
Thus,

$$-\sqrt{Var(X)Var(Y)} \le Cov(X,Y) \le \sqrt{Var(X)Var(Y)}$$

And dividing both sides by $\sqrt{Var(X)Var(Y)}$ *gives*

$$-1 \le \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \le 1$$

Since $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$,

$$-1 \le \rho(X,Y) \le 1 \ \checkmark$$

10. The following admission data are for the graduate program in the six largest majors at the University of California at Berkeley for the fall 1973 quarter.

Graduate	Male Applicants		Female Applicants	
Program	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

(a) What is the overall probability of being admitted for males? For females? What is the standard deviation for males and for females?

Variance of Bernoulli $= p(1-p)$
Standard deviation $= \sqrt{p(1-p)}$

	Male	*Female*
Overall Probability	*44.52%*	*30.35%*
Standard Deviation within Departments	*49.70%*	*45.98%*

(b) How would you write down the null and alternative hypotheses in order to test that the overall probability of admission is higher for men than for women?

$$H_0: P_{men} - P_{women} \leq 0$$
$$H_1: P_{men} - P_{women} > 0$$

We want to test if $P_{men} > P_{women}$

(c) Conduct a t-test of the hypothesis from part (b) and report the p-value.

Pooled Variance $S_p^2 = \frac{(n-1)S_{Men}{}^2 + (m-1)S_{women}{}^2}{n+m-2}$

Standard error $= S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$

Using these to calculate a t-value gives
t-statistic = 9.70
p-value < .00001

(d) Is the result significant at the 5% level? Does it provide evidence of discrimination?

Yes it is significant at the 5% level as .00001<.05

(e) Committee chairpersons claim they are more likely to admit women than men. Is this claim true? Compute acceptance rates for men and women by graduate program.

	Males					Females					Statistics	
Program	Admitted	Rejected	Total Applicants	Percent Accepted	Standard Deviation	Admitted	Rejected	Total Applicants	Percent Accepted	Standard Deviation	Standard Error	T-Test for Difference
A	512	313	825	62.06%	48.52%	89	19	108	82.41%	38.08%	4.85%	-4.19
B	353	207	560	63.04%	48.27%	17	8	25	68.00%	46.65%	9.85%	-0.50
C	120	205	325	36.92%	48.26%	202	391	593	34.06%	47.39%	3.29%	0.87
D	138	279	417	33.09%	47.05%	131	244	375	34.93%	47.68%	3.37%	-0.55
E	53	138	191	27.75%	44.78%	94	299	393	23.92%	42.66%	3.82%	1.00
F	22	351	373	5.90%	23.56%	24	317	341	7.04%	25.58%	1.84%	-0.62
Total	1198	1493	2691	44.52%	49.70%	557	1278	1835	30.35%	45.98%	1.46%	9.70

Almost every committee, except C and E, admit more women than men.
This claim is only true (at a statistically significant level), however, for department A. Department A has a t-statistic of -4.19 < -1.645 which is significant at 5% level for one-sided test. All the other firms either hire more men, or the amount by which they hire more women is statistically insignificant at the 5% level.

(f) Do these data suggest that the university is guilty of gender discrimination in its admission policy? Explain briefly.

There is NO evidence to suggest that the university is guilty of gender discrimination against women. There seems evidence for gender discrimination in the aggregate statistic, but admission records at the departmental level do not indicate evidence of discrimination against women. This is an example of so-called Simpson's paradox.