Stockholm University. Department of Statistics
Basic Statistics for Economists
Solution to the exam 220114

Problem 1. Twelve students from the same school took the Swedish Scholastic Aptitude Test (Högskoleprovet). The results on this test are normalized to a scale ranging from 0.0 to 2.0 Their scores (ordered lowest to highest) can be found in the table below:

Student	1	2	3	4	5	6	7	8	9	10	11	12
Score	0.1	0.1	0.3	0.7	0.8	0.9	0.9	1.1	1.5	1.5	1.8	2.0

a. Find the Inter Quartile Range of the twelve scores. Choose the alternative closest to your answer. (5p.)

(A) 0.8

(B) 0.9

(C) 1.1

(D) 1.2

(E) 1.3

b. The figure below shows five histograms marked A–E. Find the histogram that correctly represents the twelve scores. (5p.)

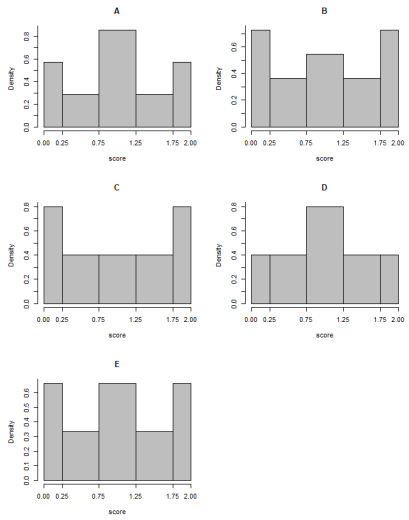
(A) A

(B) B

(C) C

(D) D

(E) E



Solution 1.

a. The Inter Quartile Range —IQR— is defined as $IQR = Q_3 - Q_1$ where Q_1 and Q_3 are, respectively, the first and third quartiles. In turn, Q_1 and Q_3 are defined as the 25th and 75th percentiles, respectively. Let us find the 25th percentile, first. In order to do this we calculate $(n+1)p/100 = (12+1)25/100 = 3.25$, so $a = 3$ and $b = 0.25$, therefore the 25th percentile is $x_{(3)} + 0.25(x_{(4)} - x_{(3)}) = 0.3 + 0.25(0.7 - 0.3) = 0.4$. Now, let us find the 75th percentile: we have $(n+1)p/100 = (12+1)75/100 = 9.75$, so $a = 9$ and $b = 0.75$, therefore the 75th percentile is $x_{(9)} + 0.75(x_{(10)} - x_{(9)}) = 1.5 + 0.75(1.5 - 1.5) = 1.5$. Finally, we have that $IQR = Q_3 - Q_1 = 1.5 - 0.4 = 1.1$.

b. The table below shows the absolute and relative frequencies of each interval.

Interval	0.00 — 0.25	0.25—0.75	0.75—1.25	1.25—1.75	1.75—2.00
Absolute frequency	2	2	4	2	2
Relative frequency	0.1667	0.1667	0.3333	0.1667	0.1667

Recall that it is the *area* (not the height) of the bars in a histogram that should be proportional to their frequencies:

- the first and second categories have the same frequency, therefore their bars should have the same area. As the width of the second category is twice the width of the first one, its height should be halved;
- the same argument applies to the fifth and fourth categories;
- the third category has twice as many observations as the second one, therefore its area should be twice as big. As both have the same width, the third category's height should be twice the second category's height.

Problem 2.

a. A class of fifth-graders has 26 students; 13 are girls and 13 are boys. Three students from the class are chosen at random to help out in the lunch room. Find the probability that the group of three contains both boys and girls. Choose the alternative closest to your answer. (5p.)

(A) 0.72

(B) 0.75

(C) $\boxed{0.78}$

(D) 0.81

(E) 0.84

b. According to experts, the probability that a certain ice-hockey team wins their next match is 50%. The probability that the team wins the match after that is also 50%, and the probability that the team loses both matches is 30%. There will be overtime and penalties, if necessary, to ensure that there are no ties. What is the probability that the team wins exactly one match? Choose the alternative closest to your answer. (5p.)

(A) 20%

(B) 25%

(C) 30%

(D) $\boxed{40\%}$

(E) 50%

Solution 2.

a. Note that the sample space of the random experiment being performed is given by

$$\{bbb\} \quad \{bbg\} \quad \{bgb\} \quad \{bgg\} \quad \{gbb\} \quad \{gbg\} \quad \{ggb\} \quad \{ggg\}$$

where —for instance— $\{bbb\}$ means that only boys were selected in the sample and $\{bbg\}$ means that the first two students selected were boys; and the third one, a girl. We are interested in

$$P(\{bbg\}) + P(\{bgb\}) + P(\{bgg\}) + P(\{gbb\}) + P(\{gbg\}) + P(\{ggb\})$$

which, by the complement rule, is equal to $1 - P(\{bbb\}) + P(\{ggg\})$, which is easier to calculate.

Let us find $P(\{ggg\})$ first. In order to observe the sample $\{ggg\}$, a girl should be selected first, which happens with probability 13/26; then a second girl should be selected, which happens with probability 12/25; finally, a third girl should be selected, which happens with probability 11/24. Therefore, we have

$$P(\{ggg\}) = \frac{13}{26}\frac{12}{25}\frac{11}{24} = \frac{1716}{15600} = 0.11.$$

The same reasoning can be applied to find that $P(\{bbb\}) = 0.11$. Then, the desired probability is $1 - P(\{bbb\}) + P(\{ggg\}) = 1 - 0.11 - 0.11 = 0.78$.

b. Let A be the event "the team wins the next match" and B be the event "the team wins the match after the next one". Therefore $\bar{A}$ denotes the event "the team loses the next match" and $\bar{B}$ denotes the event "the team loses the match after the next one". It is given that $P(A) = P(B) = 0.5$ and $P(\bar{A} \cap \bar{B}) = 0.3$. This information is shown in the following contingency table:

	B	$\bar{B}$	Total
A			0.5
$\bar{A}$		0.3	
	0.5		

- As $P(A) = 0.5$, then $P(\bar{A}) = 1 - P(A) = 1 - 0.5 = 0.5$. By an analogous reasoning, we have $P(\bar{B}) = 0.5$.
- By the law of total probability and taking into account that B and $\bar{B}$ define a partition of the sample space, we have that $P(\bar{A}) = P(\bar{A} \cap B) + P(\bar{A} \cap \bar{B})$ which gives $0.5 = P(\bar{A} \cap B) + 0.3$, therefore $P(\bar{A} \cap B) = 0.5 - 0.3 = 0.2$. By an analogous reasoning, we have $P(A \cap \bar{B}) = 0.2$ and $P(A \cap B) = 0.3$.

Putting all together gives the following contingency table:

	B	$\bar{B}$	Total
A	0.3	0.2	0.5
$\bar{A}$	0.2	0.3	0.5
	0.5	0.5	1

We are interested in the probability that the team wins exactly one match, this is

$$P\left((A \cap \bar{B}) \cup (\bar{A} \cap B)\right) = P(A \cap \bar{B}) + P(\bar{A} \cap B) = 0.2 + 0.2 = 0.4.$$

Where the first equality holds because the events are disjoint, i.e. $P\left((A \cap \bar{B}) \cap (\bar{A} \cap B)\right) = 0$.

Problem 3.

a. A factory produces electrical components. It is known that 5% of all components produced have some defect. A manager draws an i.i.d. sample of 10 components. What is the probability that at least two have some defect? Choose the alternative closest to your answer. (5p.)

(A) 0.07

(B) 0.09

(C) 0.11

(D) 0.13

(E) 0.15

b. A financial analyst uses the following model for the yearly return (change) of two stocks, *Magnavox* and *Pong Inc.* If X is the percentage return of *Magnavox* and Y is the percentage return of *Pong Inc.*, then

$$X \sim N(15\,, 25^2) \qquad Y \sim N(10\,, 20^2) \qquad \rho_{X,Y} = 0.7$$

Find the probability that X yields higher return than Y (over one year). Choose the alternative closest to your answer. **Hint:** Can you find the covariance between X and Y? (5p.)

(A) 0.44

(B) 0.55

(C) 0.56

(D) 0.61

(E) 0.65

c. Tom and Kelly are both fighter pilots in the Swedish air force. They would like their baby daughter to be a fighter pilot too. They know the air force has a height requirement: to be a pilot, you have to be between 160 cm and 190 cm tall. Taking into account their own heights, they consider that the height of her daughter as an adult can be modeled as the outcome of a normal distribution with mean 170 cm and standard deviation 7 cm. Find the probability that their daughter will be between 160 cm and 190 cm tall, according to the parents' model. Choose the alternative closest to your answer. (5p)

(A) 90%

(B) 92%

(C) 94%

(D) 96%

(E) 98%

Solution 3.

a. Let X =number of defective components in the sample. X follows a binomial distribution with parameters $n = 10$ and $p = 0.05$, $X \sim Bin(10, 0.05)$. We are interested in $P(X \geq 2)$. That probability cannot be directly found in Table 7, then we need to rewrite the desired probability in a convenient way that can be found in that table: using the complement rule, we have that $P(X \geq 2) = 1 - P(X < 2)$; and taking into account that the binomial distribution is discrete, we have $1 - P(X < 2) = 1 - P(X \leq 1)$. This probability can be found in the table: we have $1 - P(X \leq 1) = 1 - 0.91386 = 0.08614$.

b. We are looking for $P(X > Y)$ which is equivalent to $P(X - Y > 0)$. Let $W = X - Y$, we know that:

- if X and Y are normally distributed then $X - Y$ is also normally distributed;
- $E(X - Y) = E(X) - E(Y) = 15 - 10 = 5$;
- $V(X - Y) = V(X) - 2Cov(X, Y) + V(Y) = 25^2 - 2Cov(X, Y) + 20^2$.

So, we need to find $Cov(X,Y)$. We know that the correlation between X and Y is $\rho_{X,Y} = 0.7$. Using the definition of covariance, we have

$$0.7 = \rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X,Y)}{\sqrt{25^2 20^2}} = \frac{Cov(X,Y)}{500} \longrightarrow Cov(X,Y) = 0.7 \times 500 = 350.$$

Using this, we get

$$V(X-Y) = 25^2 - 2 \times 350 + 20^2 = 325.$$

Putting all together we have that $W = X - Y \sim N(5, 325)$. The desired probability is $P(W > 0)$. In order to be able to use Table 1 we need to keep working

$$\begin{aligned} P(W>0) = P\left(\frac{W-5}{\sqrt{325}} > \frac{0-5}{\sqrt{325}}\right) = & \quad \text{(Transforming to the standard normal)} \\ P(Z > -0.28) = & \quad (Z \text{ denotes the standard normal}) \\ P(Z < 0.28) = & \quad \text{(by symmetry)} \\ 0.61026 & \quad \text{(using Table 1)} \end{aligned}$$

c. Let X =height of the daughter as an adult. It is assumed that $X \sim N(170, 7^2)$. We are interested in $P(160 < X < 190)$, we have

$$\begin{aligned} P(160 < X < 190) = & \\ P\left(\frac{160-170}{7} < \frac{X-170}{7} < \frac{190-170}{7}\right) = & \quad \text{(transforming to the standard normal)} \\ P(-1.43 < Z < 2.86) = & \quad (Z \text{ denotes the standard normal}) \\ P(Z < 2.86) - P(Z < -1.43) = & \quad \text{(probability of a range)} \\ P(Z < 2.86) - 1 + P(Z < 1.43) = & \quad \text{(by the complement rule and by symmetry)} \\ 0.99788 - 1 + 0.92364 = & \quad \text{(using Table 1)} \\ 0.92152 & \end{aligned}$$

Problem 4.

a. A biologist collects an i.i.d. random sample of 10 apples from an apple orchard. She measures and weighs each apple. The mean weight of the 10 apples is 185 grams and the sample standard deviation is 35 grams. Assume that the apples' weights follow a normal distribution. Find a 90% confidence interval for the mean weight of an apple from the orchard. Choose the alternative closest to your answer. (5p)

(A) (160 , 210)

(B) $\boxed{(164\,,205)}$

(C) (167 , 203)

(D) (172 , 198)

(E) (179 , 191)

b. In a medical study, scientists compared the efficiency of two nicotine patches of two different strengths, 22 mg and 44 mg. Nicotine patches are used to help tobacco smokers to not smoke. A sample of 100 volunteer subjects were divided into two groups: 50 were given the 22 mg patch and 50 were given the 44 mg patch. After six months, the subjects were asked if they had smoked since the study started or if they had stopped completely. The result can be found in the table below:

Group	44 mg	22mg
n	50	50
Stopped smoking	16	11

Find a 95% confidence interval for the difference in proportion between the two groups (smokers using the 44 mg patch and smokers using the 22 mg patch). Choose the alternative closest to your answer. (5p.)

(A) $(-0.07\,,0.27)$

(B) $(-0.04\,,0.24)$

(C) $(-0.02\,,0.22)$

(D) $(0.01\,,0.19)$

(E) $(0.02\,,0.18)$

c. Assume that the natural lifespan of a lab mouse kept in a cage has a known population standard deviation of 80 days and that the lifespan is normally distributed. A researcher wants to estimate the population lifespan of such a mouse, with a 95% confidence interval. Find the minimum sample size that the researcher needs to use to guarantee that the margin of error is at most 5 days. Choose the alternative closest to your answer. (5p.)

(A) 40

(B) 246

(C) 385

(D) 984

(E) 1537

Solution 4.

a. Let X = weight of a randomly sampled apple. We want to construct a 90% confidence interval for μ_x. As the sample size is small, X is normally distributed and the variance of X is unknown, the confidence interval is of the form

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}}.$$

It is given that $n = 10$, $\bar{x} = 185$ and $s_x = 35$. The desired confidence level is $100(1-\alpha) = 0.9$, which means that $\alpha = 0.1$. All that remains to be found is the corresponding quantile of the t distribution, we have

$$t_{n-1,\alpha/2} = t_{9,0.05} = 1.833.$$

This yields

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}} = 185 \pm 1.833\frac{35}{\sqrt{10}} = 185 \pm 20 = (165\,, 205)$$

b. Let P_x and P_y be the proportion of subjects that stop smoking when using the 44 mg patch and the 22 mg patch, respectively. We want to construct a 95% confidence interval for $P_x - P_y$. As we have two independent samples that can be considered as large, the confidence interval is of the form

$$(\hat{p}_x - \hat{p}_x) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}.$$

It is given that $n_x = n_y = 50$. The desired confidence level is $100(1-\alpha) = 0.95$, which means that $\alpha = 0.05$. The sample proportions are

$$\hat{p}_x = \frac{16}{50} = 0.32 \qquad \text{and} \qquad \hat{p}_y = \frac{11}{50} = 0.22.$$

All that remains to be found is the corresponding quantile of the normal distribution, we have

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

This yields

$$(\hat{p}_x - \hat{p}_x) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} =$$

$$(0.32 - 0.22) \pm 1.96\sqrt{\frac{0.32(1-0.32)}{50} + \frac{0.22(1-0.22)}{50}} =$$

$$0.10 \pm 0.17 = (-0.07\,, 0.27)$$

c. Let X =lifespan of a lab mouse. As X is normally distributed and its standard deviation is known, a confidence interval is of the form

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}},$$

where the second term is the margin of error —ME—. This means that we are looking for the minimum value of n such that $ME = z_{\alpha/2}\sigma_x/\sqrt{n} \leq 5$. First, let us find the value of n that makes the margin of error to be exactly equal to 5:

$$z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}} = 5 \longrightarrow n = z^2_{\alpha/2}\frac{\sigma^2_x}{5^2}.$$

It is known that $\sigma_x = 80$. The desired confidence level is $100(1-\alpha) = 0.95$, which means $\alpha = 0.05$, therefore the corresponding quantile of the standard normal distribution is $z_{\alpha/2} = z_{0.025} = 1.96$. This yields

$$n = 1.96^2\frac{80^2}{5^2} = 983.45.$$

Which means that we would need a sample of 983.45 mice in order to obtain a confidence interval with the desired margin of error. Of course this is not possible as the sample should be an integer. As the margin of error is a decreasing function of the sample size, any sample size larger than 983.45 would yield a ME smaller than 5. Therefore the minimum sample size is 984.

Problem 5. The marketing department of an online gaming site creates two alternative two-minute videos to market their latest game. They call the videos *Version A* and *Version B*. They design the website so that visitors to the site are randomly shown either *A* or *B*. They then record the number of visitors and the number of downloads, for each version.

Version	Downloads	Total visitors
A	120	995
B	105	1005

Test at the 5% level whether the proportion of downloads per visitor is greater for *Version A* than for *Version B*. Call the test variable z_{obs}.

a. Find the decision rule of the test. (5p)

(A) Reject H_0 if $z_{obs} < 1.6449$
(B) Reject H_0 if $|z_{obs}| < 1.96$
(C) Reject H_0 if $z_{obs} > 1.6449$
(D) Reject H_0 if $|z_{obs}| > 1.96$
(E) Reject H_0 if $|z_{obs}| > 1.6449$

b. Find the value of the test variable. Choose the alternative closest to your answer. (5p)

(A) 0.83
(B) 0.98
(C) 1.07
(D) 1.14
(E) 1.28

Solution 5. Let P_A and P_B the proportion of downloads of *Version A* and *Version B*, respectively. The hypothesis system is of the form

$$H_0 : P_A = P_B \qquad \text{vs.} \qquad H_1 : P_A > P_B.$$

Taking into account that we have two independent and large samples, the test variable is of the form

$$z_{obs} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \qquad \text{with} \qquad \hat{p}_0 = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A+n_B}, \tag{1}$$

where $\hat{p}_A$ and $\hat{p}_B$ are, respectively, the proportion of downloads of *Version A* and *Version B*; and n_A and n_B, the sample sizes of *Version A* and *Version B*, respectively.

a. If the null hypothesis is true, z_{obs} follows approximately a standard normal distribution. Therefore, taking into account that we have a one-sided test where the alternative is "larger than..." we would reject the null hypothesis if z_{obs} is "too large", i.e. we would reject the null hypothesis if $z_{obs} > z_\alpha$. As $\alpha = 0.05$, looking at Table 2 we obtain $z_{0.05} = 1.6449$.

b. Let us find the value of z_{obs} given in equation (1): we have $n_A = 995$, $n_B = 1005$, $\hat{p}_A = 120/995$ and $\hat{p}_B = 105/1005$. Thus

$$\hat{p}_0 = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B} = \frac{995 \cdot \frac{120}{995} + 1005 \cdot \frac{105}{1005}}{995 + 1005} = \frac{225}{2000}$$

and

$$z_{obs} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{\frac{120}{995} - \frac{105}{1005}}{\sqrt{\frac{225}{2000}(1 - \frac{225}{2000})\left(\frac{1}{995} + \frac{1}{1005}\right)}} = 1.14.$$

Problem 6. A Swedish sociologist studying discrimination conducts an experiment. She sends out 200 job applications to 200 different advertised entry-level jobs in Sweden. In half of the 200 job applications, she uses a fictive Swedish name (like Karl Andersson) and in the other half, she uses a fictive foreign sounding first name (like Shady Gamhour). Otherwise, she takes great care to make all applications as similar as possible to each other (same CV, same personal letter, same time sent relative to the deadline, and so on). The results can be found in the table below:

	Interview	No interview
Swedish name	36	64
Foreign name	25	75

Test at the 1% level whether interview/no interview is independent of type of name used.

a. State hypotheses and the test variable. (5p.)

b. State the critical value and decision rule. (5p.)

c. Calculate the test statistic and draw conclusion. (5p.)

d. Briefly explain what a p-value is for this test. You do not have to estimate the p-value. (Tip: you can draw a picture) (5p.)

Note: this problem is inspired by scientific studies of discrimination by the late sociologist Devah Pager, but the context and numbers here are made up. Similar studies have been conducted in Sweden. See for example Bursell et al. (2021)

Solution 6.

a. Test for independence between two categorical random variables using the χ^2-method.

Hypotheses: H_0: interview/no interview and name used are independent
H_1: interview/no interview and name used are dependent

Test variable:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} = \chi^2_1$$

where O_{ij} are the observed frequencies, $E_{ij} = R_iC_j/n$ are the expected frequencies under H_0 (i.e. under the assumption that H_0 is true), R_i and C_j are the totals across rows and columns respectively, $r = 2$ is the number of rows and $c = 2$ is the number of columns. Thus, the test variables is χ^2distributed with $(r-1)(c-1) = (2-1)(2-1) = 1$ degrees of freedom.

b. Critical value:

$$\chi^2_{crit} = \chi^2_{(r-1)(c-1);\alpha} = \chi^2_{(2-1)(2-1);0.01} = \chi^2_{1;0.01} \underset{\text{Table 4}}{=} 6.635$$

Decision rule: We reject the null hypothesis at the 1% significance level if

$$\chi^2_{obs} > \chi^2_{crit}$$

c. Calculations:
Observed frequencies O_{ij}

	Interview	No interview	Total
Swedish name	36	64	100
Foreign name	25	75	100
Total	61	139	200

Expected frequencies E_{ij}

	Interview	No interview	Total
Swedish name	$(61 \cdot 100)/200 = 30.5$	$(139 \cdot 100)/200 = 69.5$	100
Foreign name	$(61 \cdot 100)/200 = 30.5$	$(139 \cdot 100)/200 = 69.5$	100
Total	61	139	200

Differences $O_{ij} - E_{ij}$

	Interview	No interview	Total
Swedish name	36 - 30.5 = 5.5	64 - 69.5 = -5.5	0
Foreign name	25 - 30.5 = -5.5	75 - 69.5 = 5.5	0
Total	0	0	0

Squared differences $(O_{ij} - E_{ij})^2$

	Interview	No interview
Swedish name	$(5.5)^2 = 30.25$	$(-5.5)^2 = 30.25$
Foreign name	$(-5.5)^2 = 30.25$	$(5.5)^2 = 30.25$

$(O_{ij} - E_{ij})^2/E_{ij}$

	Interview	No interview	Total
Swedish name	$30.25/30.5 = 0.992$	$30.25/69.5 = 0.435$	1.427
Foreign name	$30.25/30.5 = 0.992$	$30.25/69.5 = 0.435$	1.427
Total	1.984	0.87	2.854

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 2.854$$

Conclusion: Since $\chi^2_{obs} = 2.854 < \chi^2_{crit} = 6.635$ we cannot reject H_0 at the 1% significance level. We conclude that there is no statistical evidence to suggest that whether there is interview or not depends on the name used on the application.

d. In general, the p-value is interpreted as the probability of observing something as "extreme" or more "extreme" than what we observe. If this probability is too small, and taking into account that the test is constructed by assuming that the null hypothesis is true, we have evidence for concluding that the null hypothesis is quite unlikely, therefore we reject it.

Problem 7. A student wants to examine the relationship between median income and median rent, in U.S. states. To make calculations easier, we have limited the data to a sample of 11 states. You can find data, plus the products of the variable rent and income in the table below:

State	Income	Rent	Population	Income×Rent
California	29	1360	39.8	39440
Connecticut	35	1120	3.6	39200
Vermont	29	940	0.6	27260
Wisconsin	30	810	5.8	24300
Virginia	33	1170	8.5	38610
Delaware	32	1080	1	34560
Texas	28	950	28.7	26600
Oregon	27	990	4.2	26730
Kansas	29	800	2.9	23200
Alaska	33	1200	0.7	39600
Alabama	24	750	4.9	18000
Sum	329	11170	100.7	337500

Income: the median income in the state, in thousands of dollars per year

Rent: the median rent in the state, in dollars per month

Income: the population of the state, in millions

For parts a. and b, you should consider Model 1:

$$rent = \beta_0 + \beta_1 income + \epsilon$$

a. Find the variance of income and the covariance between income and rent. (5p.)

b. Estimate the coefficients of Model 1. Clearly state the estimated model. (5p)

For part c. and d, you should consider Model 2:

$$rent = \beta_0 + \beta_1 income + \beta_2 population + \epsilon$$

Part of the output from model 2 can be found below.

c. Use a formal test to test whether Model 2 is better than Model 1 (i.e., test whether population should be included in the model). (5p.)

d. Briefly explain what a residual is. Explain why it is a good idea to plot the residuals of your linear regression model. (5p.)

	df	SS	MS
Regression	2	231725.89	115862.95
Residual	8	129746.84	16218.35
Total	10	361472.73	

	Coefficients	Standard error
Intercept	-315.40	401.02
Income	41.90	13.10
Population	8.48	3.20

Solution 7.

a. Let X = Income and Y = Rent. Then the variance of income is:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

and the covariance between income and rent

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{n-1}$$

State	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	y_i	$(x_i \cdot y_i)$
California	29	(29-29.91) = -0.91	$(-0.91)^2$= 0.8281	1360	39440
Connecticut	35	(35-29.91) = 5.09	5.09^2 = 25.9081	1120	39200
Vermont	29	(29-29.91) = -0.91	$(-0.91)^2$= 0.8281	940	27260
Wisconsin	30	(30-29.91) = 0.09	0.09^2 = 0.0081	810	24300
Virginia	33	(33-29.91) = 3.09	3.09^2 = 9.5481	1170	38610
Delaware	32	(32-29.91) = 2.09	2.09^2 = 4.3681	1080	34560
Texas	28	(28-29.91) = -1.91	$(-1.91)^2$= 3.6481	950	26600
Oregon	27	(27-29.91) = -2.91	$(-2.91)^2$= 8.4681	990	26730
Kansas	29	(29-29.91) = -0.91	$(-0.91)^2$= 0.8281	800	23200
Alaska	33	(33-29.91) = 3.09	3.09^2 = 9.5481	1200	39600
Alabama	24	(24-29.91) = -5.91	$(-5.91)^2$= 34.9281	750	18000
Sum	329	0	98.9091	11170	337500

The mean of income is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{329}{11} = 29.91$$

and the variance of income is

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{98.9091}{11-1} = \frac{98.9091}{10} = 9.89.$$

The mean of rent is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{11170}{11} = 1015.45.$$

The covariance between income and rent

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{n-1} = \frac{337500 - 11 \cdot 29.9 \cdot 1015.45}{11-1} = \frac{337500 - 334093.2}{10} = 340.6$$

b. The estimate of the slope β_1 is

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{340.6}{9.89} = 34.44$$

The estimate of the intercept β_0 is

$$b_0 = \bar{y} - b_1\bar{x} = 1015.45 - 34.44 \cdot 29.91 = -14.65$$

So the estimated model is

$$\widehat{rent} = -14.65 + 34.44 \cdot income$$

c. Hypotheses: $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$:

Test variable:

$$t = \frac{b_2 - \beta_2^*}{s_{b_2}}$$

Critical value: We will take $\alpha = 5\%$

$$t_{crit} = t_{n-K-1;\alpha/2} = t_{11-2-1;0.05/2} = t_{8;0.025} \underset{\text{Table 3}}{=} 2.306$$

Decision rule: As we have a two-sided test we reject the null hypothesis at the 5% significance level if

$$|t_{obs}| > t_{crit}$$

Calculations:

$$t_{obs} = \frac{8.48 - 0}{3.20} = 2.65$$

Conclusions: Since $|t_{obs}| = 2.65 > t_{crit} = 2.306$ we reject H_0 at the 5% significance level and conclude that *population* should be included in the model.

d. A residual is the difference between an observation and its value on the estimated regression line. Residual plots are useful for validating the model's assumptions: the residuals can be seen as a sample from a normal distribution, they are independent and they all have the same variance.

References

Moa Bursell, Magnus Bygren, and Michael Gähler. Does employer discrimination contribute to the subordinate labor market inclusion of individuals of a foreign background? *Social Science Research*, page 102582, 2021.