Solutions BSFE Exam 25-01-31

Ulf Högnäs
Department of Statistics,
Stockholm University

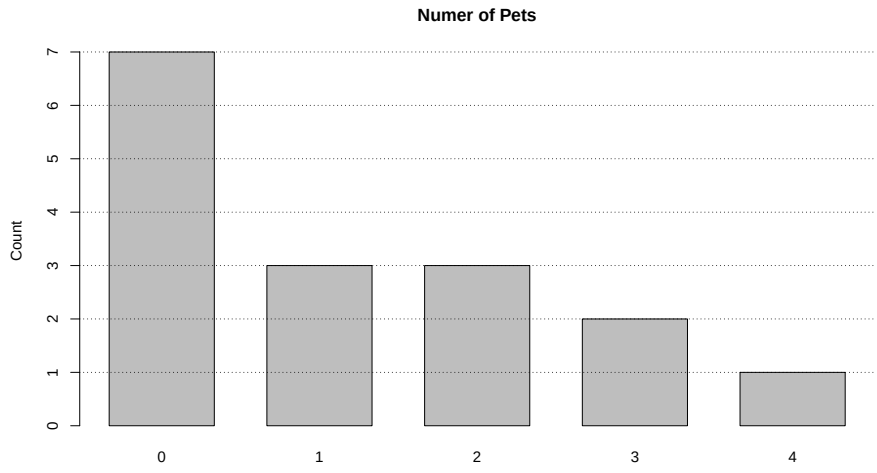
February 12, 2025

This is a draft of the solutions. Please contact ulf.hognas@stat.su.se if you find an error!

Part One – Multiple Choice

Choose the alternative closest to your answer!

1. A statistics student conducts a survey in his apartment building. He is interested in the number of pets in each apartment. He gets sixteen responses and you can find a bar chart describing the number of pets in these sixteen apartments, below. For example, seven apartments has zero pets.



Find the 25th and 75th percentile for the number of pets in the apartments.

(a) 0 and 1

(b) 0 and 2

(c) 0 and 3

(d) 1 and 2

(e) 1 and 3

Solution We order the answers, 16 in total.

0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 4

We then use the method described in the formula sheet,

Let a = integer part of $\frac{(n+1)p}{100}$.

Let b = decimal part of $\frac{(n+1)p}{100}$.

The p-th percentile is given by:

$$\text{Percentile} = x(a) + b \cdot (x(a+1) - x(a)) .$$

So,

$$\frac{(1+16) \cdot 25}{100} = 4.25$$

which gives us

$$a = 4 \quad \text{and} \quad b = 0.25.$$

Since $x(4) = x(5) = 0$ (this is the 4th and 5th observations in the ordered list), the 25th percentile will be

$$0 + 0.25 \cdot (0 + 0) = 0.$$

Using the same method, we find that the 75th percentile is between the 12th and 13th observation in the list, but these are both equal to two. So the answer is 0 & 2.

2. A car rental company knows by experience that 15% of the customers request a sport utility vehicle (SUV) and that the customers' choice is independent of each other.

   What is the probability that out of the next ten customers, exactly one will request an SUV?

   (a) 0.15
   (b) 0.25
   (c) 0.35
   (d) 0.45
   (e) 0.55

   Solution This is the binomial distribution with $n = 10$ and $p = 0.15$,

$$P(X = 1) = \binom{10}{1} p^1 q^{(10-1)} = 10 \cdot 0.15 \cdot 0.85^9 = 0.3474 \approx 0.35.$$

3. From a group of 25 students, three have to present the following Friday. How many different groups of three can be selected from the group of 25 students, assuming order does not matter?

   (a) 8
   (b) 27
   (c) 2300
   (d) 13800
   (e) 15625

   Solution This is "without replacement, order does not matter"

$$\binom{25}{3} = \frac{25!}{3!(25-3)!} = \frac{25 \cdot 24 \cdot 23 \cdot \cancel{22} \cdot \cdots \cdot \cancel{1}}{3 \cdot 2 \cdot 1 \cdot \cancel{22} \cdot \ldots \cancel{1}} = \frac{25 \cdot 24 \cdot 23}{3 \cdot 2 \cdot 1} = 2300.$$

4. A copier service company has identified three types of errors that cause some kind of action on their behalf and have also determined the probabilities of these events occurring for a randomly selected customer during a given month.

$$A = \text{“software related”} \quad P(A) = 0.05$$
$$B = \text{“mechanical failure”} \quad P(B) = 0.10$$
$$C = \text{“user error (no real error)”} \quad P(C) = 0.20$$

It is assumed that A is independent of both B and C. It is also assumed that B and C are disjoint events.

Which of the following statements is false?

(a) $P(A \cap B) = 0.005$
(b) $P(A \cup C) = 0.24$
(c) $P(C \mid A) = 0.20$
(d) $P(B \cap C) = 0$
(e) B and C are independent

Solution The problem states that “B and C are disjoint events”. Disjoint events are never independent†, so the answer is (e). If we do not know this, could instead check the definition of independence. Recall that two events B and C are independent if

$$P(B \cap C) = P(B)P(C).$$

Since B and C are disjoint, they cannot happen at once, so the left hand side of the equation is zero, i.e. $P(B \cap C) = 0$ (and this means that (d) is true). What about the right hand side?

$$P(B)P(C) = 0.10 \cdot 0.20 = 0.02 \neq 0.$$

Hence B and C are not independent.

† There is an exception to this. If at least one of the two events have zero probability, they are independent, but this is not the case here.

5. A manager at a textile factory estimates that the number of t-shirts produced per work day is approximately normally distributed with mean 3000 and standard deviation 200. Each t-shirt brings in $3 in revenue.

Find the probability that the revenue from t-shirts will be at least $8250 tomorrow, according to the manager’s model. Choose the alternative closest to your answer.

(a) 5%

(b) 20%

(c) 63%

(d) 89%

(e) 95%

Solution If X is the number of t-shirts produced, then $X \sim N(3000, 200^2)$. If Y is the revenue, then $Y = 3 \cdot X$.

$$E[3X] = 3 \cdot 3000 = 9000$$
$$Var(3X) = 3^2 \cdot Var(X) = 9 \cdot 200^2$$

so $Y \sim N(9000, 9 \cdot 200^2)$ and the standard deviation is $\sqrt{9 \cdot 200^2} = 600$. Now we standardize and solve

$$\begin{aligned} P(Y \geq 8250) &= P\left(\frac{Y - 9000}{600} \geq \frac{8250 - 9000}{600}\right) \\ &= P\left(Z \geq -1.25\right) \\ &\text{[draw and use symmetry]} \\ &= P\left(Z < 1.25\right) = 0.89435. \end{aligned}$$

6. A biologist estimates that the weight of a randomly chosen male red fox is normally distributed with a mean of 8 kg and a standard deviation of 2 kg.

   Find the probability that a randomly chosen male fox weighs more than 11 kg, according to the biologist's model. Choose the alternative closest to your answer.

   (a) 0.012

   (b) 0.017

   (c) 0.025

   (d) 0.038

   (e) 0.067

   Solution This is an easier version of the last problem. Let X be the weight of a random male red fox,

$$\begin{aligned} P(X \geq 11) &= P\left(\frac{X-8}{2} \geq \frac{11-8}{2}\right) \\ &= P\left(Z \geq 1.5\right) \\ &\text{[draw and use complement rule]} \\ &= 1 - P\left(Z < 1.5\right) = 1 - 0.93319 = 0.06681. \end{aligned}$$

7. Wenjun, a very strong chess player, is participating in a chess tournament. The table below gives the *joint probabilities* for the outcomes (Loss, Draw, or Win) of her first two games. For example, the joint probability that she wins the first game and draws (ties) the second game is 0.12.

	Loss 2nd	Draw 2nd	Win 2nd
Loss 1st	0.06	0.04	0.10
Draw 1st	0.06	0.04	0.10
Win 1st	0.08	0.12	0.40

Find the conditional probability that she wins the second game, given that she drew the first game.

(a) 0.10

(b) 0.20

(c) 0.40

(d) 0.50

(e) 0.60

Solution We will use the definition of conditional probability. Let W_2 = "wins 2nd" and D_1 = "draws 1st".

$$P(W_2|D_1) = \frac{P(W_2 \cap D_1)}{P(D_1)}$$

We do not have $P(D_1)$ but it is the sum of the middle row, by the law of total probability.

$$P(D_1) = 0.06 + 0.04 + 0.10$$

So,

$$\frac{P(W_2 \cap D_1)}{P(D_1)} = \frac{0.10}{0.20} = 0.50.$$

8. A scientist carries out a lab experiment on rats. She has a treatment group ("Trt") and a control group ("Ctrl"); she measures the difference in life average life span between the two groups. Her hypothesis is that the treatment group will have longer average life span.

$$H_0 : \mu_{Trt} - \mu_{Ctrl} = 0$$
$$H_A : \mu_{Trt} - \mu_{Ctrl} > 0$$

She calculates the test statistic and it is equal to 1.65.
Find an approximate p-value for the test. Assume that her test statistic follows a normal distribution.

(a) 0.01

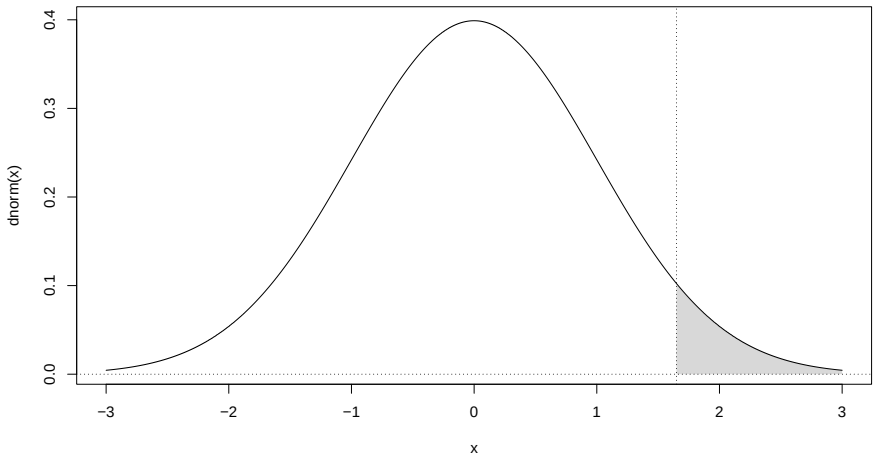
(b) 0.05

(c) 0.10

(d) 0.20

(e) Not enough information

Solution This is a one-sided test with "greater than", so the p-value will be the probability of getting a test statistic greater than 1.65, if we repeat the experiment, and given that the null is true. Here is an illustration.



We are assuming that the test statistic follows a normal distribution, so we can use table 1 in the formula sheet. We find that $P(Z < 1.65) = 0.95053$, which means that $P(Z > 1.65) = 1 - 0.95053 = 0.04947 \approx 0.05$.

9. In 2019, a Swedish survey of drug use among second-year high-school students included the question

   "have you used cannabis in the last 12 months?"

   In a random sample of 400 students, 11% answered "yes".

   Find a 90% confidence interval for the percentage of students that would answer "yes" in the population of second-year high-school students.

   (a) (8.4, 13.6)
   (b) (7.9, 14.1)
   (c) (7.5, 14.5)
   (d) (7.0, 15.0)
   (e) (6.5, 15.5)

   Solution This is a confidence interval for proportion,

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

   We just need to find the values in the formula,

$\hat{p}$	n	α	$\alpha/2$	$z_{\alpha/2}$
0.11	400	0.1	0.05	1.645

   so

$$0.11 \pm 1.645\sqrt{\frac{0.11(1-0.11)}{400}} = 0.11 \pm 1.645 \cdot 0.01564.$$

   This gives us $(0.0842722, 0.1357278) \approx (0.0843, 0.136)$.

10. The table below show mean and sample standard deviation for the body weights of a random sample of nine Chinstrap penguins (the penguins were weighed, not harmed).

Statistic	**Value**
Sample Size	9
Mean	3600
Standard Deviation	450

Find a 90% confidence interval for the true mean weight of this penguin population. Assume a normal distribution of weight in the population.

(a) $(3321, 3879)$

(b) $(3306, 3894)$

(c) $(3353, 3845)$

(d) $(3265, 3934)$

(e) $(3182, 4019)$

Solution

- *Is the sample size small?* Yes, it is less than 30.
- *Is the population normally distributed?* Yes.
- *Is the variance known?* No, we are given "sample standard deviation" in the problem, which suggests that the true population variance is unknown.

Hence, we have a t-distribution and now it is another direct application of a formula.

$$\bar{x} \pm t_{\alpha/2;\nu}\frac{s}{\sqrt{n}}$$

$\bar{x}$	s	n	ν	α	$\alpha/2$	$t_{\alpha/2;\nu}$
3600	450	9	8	0.10	0.05	1.860

$$3600 \pm 1.860\frac{450}{\sqrt{9}}$$

Which gives us $(3321, 3879)$.

11. A group of university students collect a random sample of 10 cans of *White Lightning Cola* from stores around Sweden. According to the packaging, each container is supposed to contain 330 ml of soda. The amount of soda, measured in milliliters, in each of the 10 cans can be found in the table below:

i	1	2	3	4	5	6	7	8	9	10
volume (ml)	331	329	328	328	328	329	331	329	328	329

Find the sample standard deviation of the sample.

(a) 1.10

(b) 1.15

(c) 1.20

(d) 1.33

(e) 1.50

Solution The easy way is to use this formula for the variance

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

We find the mean $\bar{x} = 329$, which we subtract from every observation. Then we square the differences.

i	1	2	3	4	5	6	7	8	9	10
volume (ml)	331	329	328	328	328	329	331	329	328	329
$x_i - \bar{x}$	2	0	-1	-1	-1	0	2	0	-1	0
$(x_i - \bar{x})^2$	4	0	1	1	1	0	4	0	1	0

When we sum the last row, we get 12. This is the numerator of the variance formula.

$$s_x^2 = \frac{12}{10-1} = \frac{4}{3}.$$

This is the variance, so we take the square root to find the standard deviation

$$\sqrt{\frac{4}{3}} \approx 1.15.$$

12. A child called Tiger is told that in a bag of m & m's, each color is equally likely and that the probabilities for each pair of candies are independent. Tiger get his parents to agree to let him collect a random sample of 10 bags, at total of 360 candies. He counts each color and compares the count to the count under the null hypothesis: that the expected number of candies is 60 of each of the six colors.

    Find the critical value for the test, if the level of significance is 5%.
    (Hint: what is the distribution of the test variable and what is the degrees of freedom?)

    (a) 1.145

    (b) 1.645

    (c) 1.960

    (d) 11.070

    (e) 16.750

    Solution This is a goodness-of-fit test. The test variable follows a χ^2-distribution, with ν degrees of freedom. For the goodness-of-fit test, ν is the number of categories minus one. Since there are six colors, $\nu = 5$.

    We go to table 4, row $\nu = 5$ and the column $\alpha = 0.05$. We find 11.070.

Part Two – Complete Solutions

13. In a sleep study, a group of scientists recorded the number of hours slept by eight randomly selected (and consenting) patients during one night. A week later, the same eight patients were given a sleep-inducing drug before their sleep was recorded once more. The table shows the hours slept by each patient, with and without the drug. Test at the 5% level whether patients sleep longer under the influence of the sleep-inducing drug.

Patient	No Drug	With Drug
1	8.5	9.5
2	6.6	7.6
3	7.8	7.8
4	7.2	8.2
5	7.6	9.6
6	7.6	9.6
7	8.7	8.7
8	7.0	8.0

Note: the numbers in this problem have been chosen to make most of the calculations easy. They may not look natural or random.

(a) State your assumptions, hypotheses, test statistic, critical value, and decision rule. (8p.)

Solution We are measuring the same person twice, for each pair of data. Hence, this is paired data and we find "for the difference $\mu_D = \mu_X - \mu_Y$ (paired dependent samples $D_i = X_i - Y_i$)" in the formula sheet.

i. **Hypotheses** We test if the subject sleep **longer** with the drug. That would make $\mu_{Drug} - \mu_{No}$ greater than zero so,

$$H_0 : \mu_{Drug} - \mu_{No} = 0$$
$$H_A : \mu_{Drug} - \mu_{No} > 0$$

ii. **Test statistic**

$$t_{n-1} = \frac{\bar{d} - \mu_0}{s_{\bar{d}}/\sqrt{n}}$$

iii. **Critical value** We have a sample size of 8, so $\nu = 8 - 1 = 7$. This is a one-sided test and $\alpha = 0.05$, so we find the column for $\nu = 8 - 1 = 7$ and $\alpha = 0.05$

$$t_{0.05;7} = 1.895.$$

iv. **Decision Rule** This is a one-sided test with "greater than", so we will reject test statistic is greater than the critical value. We reject the null if $t_{obs} > 1.895$.

(b) Finish your calculations, state your conclusions, and give a verbal interpretation. (6p.)

Solution

i. **Calculations** We need the differences. Since we are going to find the standard deviation of the differences, we also calculate the squared differences. Here, the column d is "With Drug" minus "No Drug".

Patient	No Drug	With Drug	d	d^2
1	8.5	9.5	1	1
2	6.6	7.6	1	1
3	7.8	7.8	0	0
4	7.2	8.2	1	1
5	7.6	9.6	2	4
6	7.6	9.6	2	4
7	8.7	8.7	0	0
8	7.0	8.0	1	1

The formula for variance

$$s_{\bar{d}}^2 = \frac{\sum d_i^2 - n\bar{d}^2}{n-1}.$$

We need $\bar{d} = \frac{1+1+0+1+2+2+0+1}{8} = 1$ and $\sum d_i^2 = 12$ (this is the sum of the last column).

$$s_{\bar{d}}^2 = \frac{12 - 8 \cdot 1^2}{8-1} = \frac{4}{7}.$$

Note that mu_0 in the formula for the test variable is the difference in the null hypothesis, which is zero. We are now ready to calculate t_{obs},

$$t_{obs} = \frac{\bar{d} - \mu_0}{s_{\bar{d}}/\sqrt{n}} = \frac{1-0}{\sqrt{\frac{4}{7}}/\sqrt{8}} = \frac{1-0}{\sqrt{\frac{4}{7 \cdot 8}}} \approx 3.742.$$

ii. **Conclusion and Verbal Interpretation** Since $t_{obs} = 3.742 > 1.895$, we reject the null. We have found significant evidence, at the 5% level, for the hypothesis that the drug makes people sleep longer, on average.

(c) Briefly explain the difference between a Type I error and a Type II error in statistics. Give a brief example of a Type I error. You may use the study in this problem to construct your example. (6p.)

Solution A Type I error is rejecting the null when the null is true. A Type II error is failing to reject the null when the null is false.

Example: In the sleep study, suppose that the drug is completely inefficient, so the null is true. If we conduct a a lab experiment and then reject the null, this would be an example of a Type I error.

14. A business student wants to calculate the *Beta coefficient* for the stock MGM. The Beta coefficient is a measure of the expected change of a stock in proportion to movements of the rest of the stock market. He uses data from 255 trading days and the movement of the S&P 500 (an American stock index) as proxy for the stock market. He then estimates the following regression model

$$MGM = \beta_0 + \beta_1 \cdot SP500 + \varepsilon. \tag{1}$$

Where $SP500$ is the daily returns of the S&P 500 and MGM is the daily returns of the stock MGM. The coefficient β_1 is the Beta coefficient.

"The daily return" is the proportional change from one day to the next. For example, if the stock increased one percent, the return was 0.01.

Below, you are given some sums for $i = 1, \ldots, n$. Use these together with the formula sheet to solve the problems below. The values have been rounded to make the calculations easier.

$$\begin{aligned} n &= 255 \\ \sum x_i &= 0.145 \\ \sum x_i^2 &= 0.0166 \\ \sum y_i &= 0.125 \\ \sum y_i^2 &= 0.129 \\ \sum x_i y_i &= 0.0286 \end{aligned}$$

where x_i is the ith daily returns of S&P500 and y_i is the ith daily return of MGM.

(a) Find the sample covariance between x and y. (2p.)

Solution We need to $\bar{x}$ and $\bar{y}$,

$$\begin{aligned} \bar{x} &= \frac{0.145}{255} \\ \bar{y} &= \frac{0.125}{255} \end{aligned}$$

$$\begin{aligned} s_{xy} = \mathrm{Cov}(x, y) &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} \\ &= \frac{0.0286 - 255 \cdot 0.145/255 \cdot 0.125/255}{255 - 1} \approx 0.0001123. \end{aligned}$$

(b) Find the sample variance of x. (2p.)

Solution

$$s_x^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{0.0166 - 255 \cdot (0.145/255)^2}{255-1} \approx 6.502 \cdot 10^{-5}$$

(c) Find the slope coefficient b_1 of the regression model, i.e. find the Beta coefficient. (3p.)

Solution Here, we will use rounded values from (a) and (b), instead of the full expressions, to make the solutions readable. You are allowed to do this too, but do note that it is better to save the answers with maximal decimal places and used the saved values in (c).

$$b_1 = \frac{\text{Cov}(x,y)}{s_x^2} \approx \frac{0.0001123}{6.502 \cdot 10^{-5}} \approx 1.727.$$

It is easy to mess up calculations with decimal form with lot of zeros, so it might be easier to use the scientific form here, e.g. $6.502 \cdot 10^{-5}$ instead of 0.00006502.

(d) Interpret b_1 in words. (3p.)

Solution If we increase the return of the S&P500 by one percentage point, the expected return of MGM increases by 1.727 percentage points, compared to the a day with with a one percent lower return of the S&P500.

Note that these models have an intercept, so it is not exactly correct to say "if the return of the S&P500 is one percentage point, the expected return of MGM is 1.727 percentage points".

You can say "if the return of the S&P500 is one percentage point, the expected return of MGM is 1.727 percentage points, plus the intercept".

(e) Briefly explain what the difference between a confidence interval and prediction interval is, in the context of linear regression. (5p.)

Solution A confidence interval describes the mean of a population, conditional on specific values of the independent variables.

A prediction interval is an interval that is supposed to predict individual outcomes, again conditional on some specific values of the independent variables.

Let us take the model in this problem as example. A 95% confidence interval would give you an interval for the mean return of MGM, given some return of the S&P500, e.g. 0%.

A 95% prediction interval given a 0% return of the S&P500, would give us an interval that captures the return of the MGM stock that day, with "95% confidence".

(f) Briefly explain the difference between b_1, the value that you calculated in (c), and β_1, the slope in the equation marked (1). (5p.)

Solution In the equation

$$MGM = \beta_0 + \beta_1 \cdot SP500 + \varepsilon.$$

describes the "true", and unknown, relationship between return of the MGM stock and the return of the S&P500. We estimate this relationship when we fit the linear regression model. The coefficient b_1 is our estimate of β_1, just like $\bar{x}$ is the estimate of the population parameter μ, when we draw a random sample to estimate a mean.