

Student name:

---

DS 401: Stat modeling for data science  
Fall 2021

Midterm Test  
October. 27, 2021

---

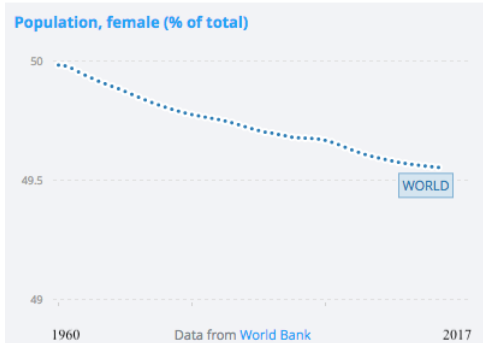
**Instructions:**

- Answer all questions on separate paper (not on this sheet!).
  - This is a regular “closed-book” test, and is to be taken without the use of notes, books, or other reference materials.
  - This test contains questions numbered (1) to (8). It adds up to 38 points.
- 

- (1) [4 pts.] A medical research team, after testing a new drug for treating COVID-19 lung infections, concludes there is statistically significant evidence their drug is more effective than a placebo. They base their conclusion on a significance level of  $\alpha = 0.05$ .
- (a) Would they conclude the same if  $\alpha = 0.01$ ? Justify your answer.
- (b) Suppose their  $P$ -value is 0.005. Explain what the  $P$ -value means in this context.
- (2) [4 pts.] We want to know the true proportion of EC students who love mathematics! We surveyed a group of students and found a 95% confidence interval estimated this proportion to lie in the range [0.73, 0.93].
- (a) What is the margin of error according to this study? Show reasoning.
- (b) Would the interval get narrower, or wider, if we use 80% confidence?
- (c) What would happen to the interval width if the sample size is cut in half?
- (3) [4 pts.] Is there a relationship between the size of sand grains on a beach and its slope? Scatterplot of data from a sample of 20 beaches around the world suggests there may be a linear association. Based on these data, the following linear regression model was constructed to predict grain size (in mm.) from the beach slope (in degrees):
- $$\widehat{\text{median sand diameter}} = 0.16 + 0.053 (\text{beach slope})$$
- Assume all the conditions for linear regression are met.
- (a) Identify the explanatory variable and the response variable, including their units.
- (b) Interpret the meaning of the slope (with units) in this application
- (4) [4 pts.] The distribution of household income in a geographic region is bimodal, with strong right skew. The mean and standard deviation are \$52,000 and \$23,000, respectively.
- (a) Suppose we draw a single random sample of size 10,000 from this population. What is its expected shape, mean and SD?
- (b) What is the expected shape, mean and SD of the sampling distribution of random samples of size 10,000?

- (5) [4 pts.] For each of the following, write the null and alternative hypotheses. Be sure to clearly define the parameter used in your statements:
- An online retail company wants to test whether the average click-through rates (in minutes) on their new website are shorter than the 4.8 minutes that customers spent on their old website before making a purchase.
  - A research group wants to know whether more than half the adults in the U.S. feel they get enough sleep. They find data from a Gallup poll, based on a random sample of 603 U.S. adults, of whom 320 said that they get enough sleep.
- (6) [6 pts.] A travel agency wants to study the association (if any!) between airline fares, and distance traveled. They collect data on fare and distance for a random sample of 34 flights, and find the scatterplot shows an approximately linear association, with a correlation of  $r = 0.53$ . Other summary statistics are given in the following table

<b>Fare</b>	mean= \$148, standard deviation= \$48.03
<b>Distance</b>	mean= 1107 miles, standard deviation= 985.4 miles

- Construct a linear regression model to predict fare from distance.
  - Compute  $R^2$  and explain what it means in this application context.
  - Carry out a hypothesis test to determine whether there is statistically significant evidence of a linear relationship between these variables. The standard error for the slope estimate is 0.09.
- (7) [6 pts.] The World Bank regularly compiles data on various demographic trends in the world's population. One striking trend seen in their data is a steady decline over the past 50+ years in the proportion of females in the world's population (see graph). According to their data, currently about 49.55% of the world is female.
- 
- Consider a sample of 5000 people drawn randomly from the current world population. What is the sampling distribution model for the proportion of females in such samples? Find the probability that at least half the people in a sample of this size will be female.
  - An organization in Southeast Asia claims the proportion of females in their region is higher than the overall global figure of 49.55%. Their claim is based on a random sample of 5000 people from the region, of whom 2548 were female. Carry out a hypothesis test to assess whether their data supports their claim.
- (8) [6 pts.] A public-service organization carried out a survey of the tuition and fees for the 2017-18 school year at a sample of 44 private colleges in the midwestern U.S. The mean and standard deviation (in thousands of dollars) were found to be 32.4 and 7.2, respectively. Carry out a hypothesis test to determine whether these data suggest the mean costs in the midwest differ significantly from the national average of \$34,800 published by the U.S. Department of Education.

*End of test*

## DS 401: Fall 2021: Midterm solutions

- (1) [4 pts.] A medical research team, after testing a new drug for treating COVID-19 lung infections, concludes there is statistically significant evidence their drug is more effective than a placebo. They base their conclusion on a significance level of  $\alpha = 0.05$ .  
(a) Would they conclude the same if  $\alpha = 0.01$ ? Justify your answer.

**Solution:** The conclusion could change if  $\alpha$  is lowered to 0.01. Because all we know is  $P\text{-value} < 0.05$ . It doesn't necessarily mean the  $P\text{-value} < 0.01$ .

(b) Suppose their  $P\text{-value}$  is 0.005. Explain what the  $P\text{-value}$  means in this context.

**Solution:** The  $P\text{-value}$  means: If the research team's drug is not more effective than the placebo, then there is a 0.005 chance of observing the results their sample produced (or even more extreme results).

**Grade:** (a)=1.5 point, (b)=2.5 pt.

For (a): 1 pt =correct answer; 0.5 pt = reason.

For (b): No partial credit, unless it is mostly correct.

- (2) [4 pts.] We want to know the true proportion of EC students who love mathematics! We surveyed a group of students and found a 95% confidence interval estimated this proportion to lie in the range [0.73, 0.93].

(a) What is the margin of error according to this study? Show reasoning.

**Solution:** The margin of error =  $(0.93 - 0.73)/2 = 0.1$ .

The ME is just half the width of the confidence interval.

(b) Would the interval get narrower, or wider, if we use 80% confidence?

**Solution:** Narrower. If all else remains the same, lower confidence level will decrease the ME, and the width of the confidence interval.

(c) What would happen to the interval width if the sample size is cut in half?

**Solution:** It will become wider, since the ME contains a  $\sqrt{n}$  term in the denominator. So, when  $n$  is cut in half, the ME increases by  $\sqrt{2}$ .

**Grade:** (a)=1.5 pt, (b)=1 pt, (c)=1.5 pt.

For (a) and (c): 1 pt =correct answer; 0.5 pt = reason.

- (3) [4 pts.] Is there a relationship between the size of sand grains on a beach and its slope? Scatterplot of data from a sample of 20 beaches around the world suggests there may be a linear association. Based on these data, the following linear regression model was constructed to predict grain size (in mm.) from the beach slope (in degrees):

$$\widehat{\text{median sand diameter}} = 0.16 + 0.053 (\text{beach slope})$$

Assume all the conditions for linear regression are met.

(a) Identify the explanatory variable and the response variable, including their units.

**Solution:** Explanatory var. = beach slope in degrees. Response var. = median sand diameter in mm.

(b) Interpret the meaning of the slope (with units) in this application.

**Solution:** For each degree that the beach slope increases, the median sand diameter is predicted to increase by 0.053 mm, on average.

**Grade:** 2 points each.

For (a): 0.5+0.5pt for each correct answer + units.

For (b): No partial credit, unless it is substantially correct.

- (4) [4 pts.] The distribution of household income in a geographic region is bimodal, with strong right skew. The mean and standard deviation are \$52,000 and \$23,000, respectively.

(a) Suppose we draw a single random sample of size 10,000 from this population. What is its expected shape, mean and SD?

**Solution:** A single random sample of that size would be expected to resemble the population from which it was drawn. Thus it is expected to be bimodal, right skewed, with mean and standard deviation close to \$52,000 and \$23,000, respectively.

(b) What is the expected shape, mean and SD of the sampling distribution of random samples of size 10,000?

**Solution:** Since the sample size is large, and the samples are randomly selected, the central limit theorem applies. Thus, the sampling distribution is given by the normal model  $N(52000, 230)$ . The standard deviation comes from the CLT formula:  $23,000/\sqrt{10,000} = 230$ .

**Grade:** 2 points each. Generally: 0.5 pt each for correct shape+mean+SD+units.

- (5) [4 pts.] For each of the following, write the null and alternative hypotheses. Be sure to clearly define the parameter used in your statements:

(a) An online retail company wants to test whether the average click-through rates (in minutes) on their new website are shorter than the 4.8 minutes that customers spent on their old website before making a purchase.

**Solution:** Let  $\mu$  = true mean click-through rate (in minutes) on their new website.

Null hypothesis  $H_0 : \mu = 4.8$

Alt hypothesis  $H_A : \mu < 4.8$

(b) A research group wants to know whether more than half the adults in the U.S. feel they get enough sleep. They find data from a Gallup poll, based on a random sample of 603 U.S. adults, of whom 320 said that they get enough sleep.

**Solution:** Let  $p$  = true proportion of adults in the U.S. who feel they get enough sleep.

Null hypothesis  $H_0 : p = 0.5$

Alt hypothesis  $H_A : p > 0.5$

**Grade:** 2 points each. Generally: 1 pt = correct  $H_0$ ; 1 pt = correct  $H_A$ .  
-0.5 pt each for missing or unclear parameter in hypotheses.

- (6) [6 pts.] A travel agency wants to study the association (if any!) between airline fares, and distance traveled. They collect data on fare and distance for a random sample of 34 flights, and find the scatterplot shows an approximately linear association, with a correlation of  $r = 0.53$ . Other summary statistics are given in the following table

<b>Fare</b>	mean= \$148, standard deviation= \$48.03
<b>Distance</b>	mean= 1107 miles, standard deviation= 985.4 miles

(a) Construct a linear regression model to predict fare from distance.

**Solution:**

Let us first check the conditions for constructing a linear model:

- (i) Is the relationship approximately linear? Yes. The question says the scatterplot shows an approximately linear association.
- (ii) Are the observations independent? Yes, it is reasonable to assume, since the sample is random and 34 flights is  $< 10\%$  of all flights.

- (iii) and (iv) require information about residuals, for which we need the actual data. Since that is not given, we can't check whether the residuals have constant variability and are normally distributed.

Since we want to predict fare from distance:

$x$ -variable = distance (in miles)

$y$ -variable = fare (in \$)

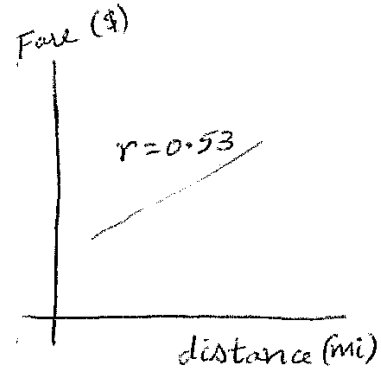
Given summary statistics are:

$$\bar{x} = 1107, s_x = 985.4 \text{ mi}, r = 0.53$$

$$\bar{y} = 148, s_y = 48.03$$

Model looks like:  $\hat{y} = b_0 + b_1x$ .

$$b_1 = r \frac{s_y}{s_x} = 0.53 \left( \frac{48.03}{985.4} \right) = 0.0258 \frac{\$}{\text{mile}}$$



Then we have:  $\hat{y} = b_0 + 0.0258x$ .

Find  $b_0$  by plugging in  $(\bar{x}, \bar{y})$ :

$$148 = b_0 + 0.0258(1107) \Rightarrow b_0 = 148 - 0.0258 \times 1107 = 119.4 \text{ \$}$$

Equation of regression line is:

$$\hat{y} = 119.4 + 0.0258x \quad \text{OR} \quad \boxed{\widehat{\text{Fare}} = 119.4 + 0.0258 \text{ Distance}}$$

(b) Compute  $R^2$  and explain what it means in this application context.

**Solution:**

$$\text{Since } r = 0.53, R^2 = (0.53)^2 \times 100\% = \boxed{28.09\%}$$

The  $R^2$  value of 28.09% means that about 28% of the variability in fares is explained by the variability in distance traveled.

(c) Carry out a hypothesis test to determine whether there is statistically significant evidence of a linear relationship between these variables. The standard error for the slope estimate is 0.09.

**Solution:**

Let  $\beta_1$  denote the true slope of a possible linear relationship between airfares and distance traveled. The hypotheses are:

Null hypothesis  $H_0 : \beta_1 = 0$

Alt hypothesis  $H_A : \beta_1 \neq 0$

I will use a significance level of 10% ( $\alpha = 0.1$ ).

Since the conditions have been checked, we can proceed to the computations.

$$t\text{-score} = \frac{b_1 - \beta_1}{SE} = \frac{0.0258 - 0}{0.09} = 0.287$$

Here  $n = 34$ ,  $df = 32$ . The  $P$ -value for ( $t_{32} = 0.287$ ) is over 20%. Since this is larger than  $\alpha$ , we retain the null hypothesis and conclude there is no statistically significant evidence of a linear relationship between airfares and distance.

**Grade:** 6 points. Distribution: (a)=3 pt, (b)=1 pt, (c)=2 pt.  
 For (a): 0.5 pt = check conditions; 2 pt = correctly pick  $x, y$  + compute slope + find intercept + include units at least somewhere; 0.5 pt = regression equation.  
 For (b): 0.5 pt each for correct answer + correct interpretation.  
 For (c): 0.5 pt each for hypotheses,  $t$ -score computation,  $P$ -value, conclusion.

(7) [6 pts.] The World Bank regularly compiles data on various demographic trends in the world's population. One striking trend seen in their data is a steady decline over the past 50+ years in the proportion of females in the world's population (see graph). According to their data, currently about 49.55% of the world is female.

- (a) Consider a sample of 5000 people drawn randomly from the current world population. What is the sampling distribution model for the proportion of females in such samples? Find the probability that at least half the people in a sample of this size will be female.

**Solution:**

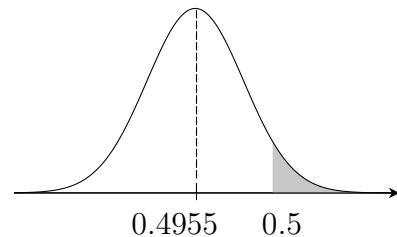
Check conditions:

- (i) Since the sample is random and 5000 people is  $< 10\%$  of the world population, it is reasonable to assume independence.
- (ii) The sample is large enough since there are more than 10 successes and failures ( $5000 \times 0.4955$  and  $5000 \times (1 - 0.4955)$  are way larger than 10).

According to the central limit theorem the sampling distribution follows the model

$$N(0.4955, \sqrt{\frac{0.4955(1 - 0.4955)}{5000}})$$

$$= N(0.4955, 0.0071)$$



To find the probability that at least half will be female:

$$z = \frac{0.5 - 0.4955}{0.0071} = 0.635. \text{ From } z\text{-table, the probability} = 0.261$$

The probability that at least half the sample is female = 0.261.

- (b) An organization in Southeast Asia claims the proportion of females in their region is higher than the overall global figure of 49.55%. Their claim is based on a random sample of 5000 people from the region, of whom 2548 were female. Carry out a hypothesis test to assess whether their data supports their claim.

**Solution:**

Let  $p$  denote the true proportion of females in the region of interest.

Null hypothesis  $H_0 : p = 0.4955$

Alt hypothesis  $H_A : p > 0.4955$

I will use a significance level of 10% ( $\alpha = 0.1$ ).

The sample is random + same size as question (a). So the conditions are met.

The sampling distribution model is the same as well:  $N(0.4955, 0.0071)$

Data from the sample:  $\hat{p} = 2548/5000 = 0.5096$

$$z = \frac{0.5096 - 0.4955}{0.0071} = 1.99. \text{ From } z\text{-table, the } P\text{-value} = 0.0233$$

Conclusion: Since the  $P$ -value is less than  $\alpha$ , we reject the null hypothesis and conclude the sample provides statistically significant evidence that the proportion of females in that region is higher than the overall global figure.

**Grade:** 6 points. Distribution: (a)=(b)=3 points.  
For (a): 1 pt=correct conditions check; 1.5 pt=correct shape+mean+SD of model;  
0.5 pt=correct computation of required probability.  
For (b): 1 pt=correct hypotheses + clear indication of parameter; 0.5 pt=correct  
sampling distribution model; 1 pt= $z$ -score +  $P$ -value computation;  
0.5pt=conclusion.

- (8) [6 pts.] A public-service organization carried out a survey of the tuition and fees for the 2017-18 school year at a sample of 44 private colleges in the midwestern U.S. The mean and standard deviation (in thousands of dollars) were found to be 32.4 and 7.2, respectively. Carry out a hypothesis test to determine whether these data suggest the mean costs in the midwest differ significantly from the national average of \$34,800 published by the U.S. Department of Education.

**Solution:**

Let  $\mu$  = true mean tuition and fees at private colleges in the midwestern U.S.

Null hypothesis  $H_0 : \mu = 34.8$  thousand dollars

Alt hypothesis  $H_A : \mu \neq 34.8$  thousand dollars

I will use a significance level of 10% ( $\alpha = 0.1$ ).

Check conditions:

- (i) It is not clear whether the sample is random, or at least representative. Thus, it may not meet the independence condition.  
(ii) There is no indication whether data in the sample is normally distributed. However, since the sample size is 44, this condition is less critical.

Sampling distribution model:

Sample statistics are:  $n = 44, \bar{x} = 32.4, s = 7.2$  thousand dollars

The sampling distribution follows:  $t_{43}(34.8, \frac{7.2}{\sqrt{44}}) = t_{43}(34.8, 1.0854)$

$$t\text{-score} = \frac{32.4 - 34.8}{1.0854} = -2.211$$

From  $t$ -table, the closest lower  $df$  is 40, and the  $P$ -value is between 0.02 and 0.05.

Conclusion: Since the  $P$ -value is less than  $\alpha$ , we reject the null hypothesis and conclude the sample provides statistically significant evidence that the mean tuition and fees at private colleges in the midwestern U.S. is different from the national average. We note, however, that the sample may not have met the independence condition, which makes this conclusion unreliable.

**Grade:** 6 points. Distribution:  
1.5 pt=correct hypotheses + clear indication of parameter;  
1 pt=correct conditions check;  
2 pt=correct sampling distribution model ( $df$ , shape, mean, SD);  
1 pt= $t$ -score +  $P$ -value computation;  
0.5pt=conclusion.