ECN 102: Analysis of Economics Data

Midterm Exam SS1 2025 – Answer Key

Remy Beauregard, Department of Economics, UC Davis

This exam consists of 5 short-answer questions (with sub-parts) and 5 multiple choice questions. You will have a maximum of 100 minutes to complete this exam. No additional time may be taken without prior accommodation. Show all work to receive full credit. You may use only the calculators provided by the instructor. For questions requiring computation, it suffices to express your final answer with 2 decimal places. There is a formula sheet and scratch paper provided at the end of this exam. You may remove these pages and discard them after the exam. If you plan to use any of these extra pages for your final answers, please write your name and student ID at the top of each to ensure they are not lost.

This exam is worth 70 points.

Question 1: Summary Statistics [15 points]

Suppose we are using a dataset that contains one large positive outlier (i.e. one value far greater than all other values). We know this value is not a mistake and so do not want to drop it from our sample. However, we do want to use the proper statistics to describe our data with this outlier:

a) What statistic of central tendency should we use for this data? Why? [3 points]

We should use the median. It is less affected by outlier values than the mean.

b) What statistic of spread should we use for this data? Why? [3 points]

We should use the IQR. It is less affected by outlier values than the range or SD/variance.

c) What can we infer about the skewness of this dataset from the description above? What can we infer about the mean of the data relative to the median? [3 points]

We can infer than $\bar{x} > median$ for this data and that we are right-skewed ($skewness > 0$).

d) Can we expect this dataset to be normally distributed based on the information above? Please justify your answer. [3 points]

With $skew > 0$, we cannot expect this data to be normally distributed. A normal distribution has zero skew (is symmetric).

e) Suppose we really want to work with a normal distribution for our analysis. Is there anything we could do to our dataset above to make it appear more normally distributed? If we are successful, what name would we give the resulting (more normal-looking) distribution? [3 points]

We could try log-transforming our series, as this can make right-skewed distributions appear more normal. If successful, we would call this a *lognormal* distribution.

Question 2: Summation Notation [4 points]

Please rewrite (*but do not simplify*) the following quantities using summation notation ($\sum$) with index i:

a) [2 points]

$$3x_1 + 3x_2 + 3x_3 + 3x_4$$

$$\sum_{i=1}^{4} 3x_i$$

b) [2 points]

$$1 + 3 + 5 + 7 + 9 + 11$$

$$\sum_{i=1}^{6} (2i - 1)$$

Question 3: Univariate Inference [8 points]

Suppose we wish to learn about the population mean of a series but only know the information given below. For each of the following scenarios, please give (1) whether or not we would need to conduct inference to answer our question and (2) the name and formula of the test statistic we would construct (if doing inference):

a) We know $\bar{x}$, $s_{\bar{x}}$, n and hypothesize μ_0. [2 points]

We will need to conduct inference to learn about μ. We will construct a t-statistic:

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

b) We know $\bar{x}$, $\sigma_{\bar{x}}$, n and hypothesize μ_0. [2 points]

We will need to conduct inference to learn about μ. We will construct a z-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

c) We know $\bar{x}$, σ_x, n and hypothesize μ_0. [2 points]

We will need to conduct inference to learn about μ. We will construct a z-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma_x/\sqrt{n}}$$

d) We know $\bar{x}$, σ_x, n, μ and hypothesize μ_0. [2 points]

We will NOT need to conduct inference or compute a test statistic since we know μ already.

Question 4: More Univariate Inference [15 points]

We are given the following output from Stata:

```
(1978 automobile data)


invttail(60,0.05)=1.6706489
invttail(60,0.025)=2.0002978
invttail(59,0.05)=1.671093
invttail(59,0.025)=2.0009954
invttail(58,0.05)=1.6715528
invttail(58,0.025)=2.0017175
```

Suppose we then use the `summarize` command to obtain the following results for the three variables below:

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         60    5907.467    2577.946       3291      13594
         mpg |         60    21.66667    6.001883         12         41
      weight |         60    2923.833    769.0917       1760       4840
```

a) Which of the three variables has the *lowest* dispersion (spread relative to its mean)? [2 points]

$$CV_{price} = \frac{2577.95}{5907.47} = 0.44$$
$$CV_{mpg} = \frac{6}{21.67} = 0.28$$
$$CV_{weight} = \frac{769.09}{2923.83} = 0.26$$

We see that *weight* has the smallest CV and thus lowest dispersion.

b) Provide a 90% CI for the mean of *mpg*. Interpret this range in words. [3 points]

$$21.67 \pm 1.67 \times \frac{6}{\sqrt{60}} = (20.38, 22.96)$$

This range has a 90% chance of including the true population mean of *mpg*.

c) The claim is made that the average price of a car in 1978 is below \$6000. Using our sample, test this claim at the 5% significance level. Clearly state your null and alternate hypotheses, your test statistic, and your conclusion. [5 points]

$$H_0 : \mu_{price} \geq 6000$$
$$H_A : \mu_{price} < 6000$$
$$t = \frac{5907.47 - 6000}{2577.95/\sqrt{60}} = -0.28$$
$$t^*_{60-1,0.05} = 1.67$$
$$|t| < |t^*| \Rightarrow \textbf{ Fail to reject the null}$$

We cannot say that the population mean of car price in 1978 was below \$6000.

d) The claim is made that the average weight of a car in 1978 is different from 3500 pounds. Using our sample, test this claim at the 5% significance level. Clearly state your null and alternate hypotheses, your test statistic, and your conclusion. [5 points]

$$H_0 : \mu_{weight} = 3500$$
$$H_A : \mu_{weight} \neq 3500$$
$$t = \frac{2923.83 - 3500}{769.09/\sqrt{60}} = -5.80$$
$$t^*_{60-1,0.025} = 2$$
$$|t| > |t^*| \Rightarrow \textbf{ Reject the null}$$

We have sufficient evidence to say that the population mean of car weight in 1978 was different from 3500 pounds.

Question 5: Bivariate Inference [18 points]

a) Write down the population model (equation) for a bivariate regression of y on x if our assumptions about linearity, unbiasedness, homoskedasticity, and independence hold. [2 points]

$$y = \beta_1 + \beta_2 x + u$$

b) Assuming our four population assumptions hold, complete the statement (fill in the blanks) for how we expect b_2 to be distributed. [5 points]

$$\frac{b_2 - \beta_2}{\sigma_{b_2}} \left[= \frac{b_2 - \beta_2}{\frac{\sigma_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \right] \sim N(0,1) \textbf{ when } n > 30$$

Suppose we now run a regression and obtain these results:

```
      Source |       SS           df       MS      Number of obs   =        60
-------------+----------------------------------   F(1, 58)        =     15.05
       Model |  80799680.8         1  80799680.8   Prob > F        =    0.0003
    Residual |   311302776        58  5367289.24   R-squared       =    0.2061
-------------+----------------------------------   Adj R-squared   =    0.1924
       Total |   392102457        59  6645804.35   Root MSE        =    2316.7

------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         mpg |  -194.9806   50.25323    -3.88   0.000    -295.5733   -94.38778
       _cons |   10132.05   1129.152     8.97   0.000     7871.802    12392.29
------------------------------------------------------------------------------
```

c) Write down the sample regression model we have estimated (with proper variable names). [3 points]

$$\widehat{price} = 10132.05 - 194.98 mpg$$

d) Suppose we are worried Stata made a mistake computing our R-squared value (0.2061). Verify this calculation using other information from the regression output. [2 points]

$$R^2 = \frac{ExpSS}{TSS} = 1 - \frac{ResSS}{TSS} = \frac{80799680.8}{392102457} = 0.2061$$

e) Suppose we are worried Stata made a mistake computing the default t-statistic for b_2 (-3.88). Verify this calculation using other information from the regression output. [2 points]

$$t = \frac{b_2 - 0}{s_{b_2}} = \frac{-194.98}{50.25} = -3.88$$

f) What will $\sum_{i=1}^{n} e_i$ equal for regression? What property will $\sum_{i=1}^{n} e_i^2$ have? [2 points]

$$\sum_{i=1}^{n} e_i = 0 \textbf{ for all best fit lines; } \sum_{i=1}^{n} e_i^2 \textbf{ is minimized by the OLS process}$$

g) Suppose we standardized both our series X and Y and reran the regression above. If we now obtain a b_2 value of -0.45, what does that tell us about r_{xy}? Interpret this value. [2 points]

BONUS: What is the value of $\frac{s_{price}}{s_{mpg}}$? [+2 points]

From our formula, we know $b_2 = r_{xy}\left(\frac{s_y}{s_x}\right)$. Standardizing x and y sets $s_y, s_x = 1$. Thus, we know $r_{xy} = -0.45$; our series X and Y are weakly negatively correlated.

[BONUS] From the formula, we also know that

$$\frac{s_{price}}{s_{mpg}} = \frac{b_2}{r_{xy}} = \frac{-194.98}{-0.45} = 433.29$$

Multiple Choice [2 points each]

Suppose we know a r.v. $X \sim (16, 16)$, draw 16 samples of size 36 (each), and compute 16 sample means.

MC 1

What would we expect the population **mean** of the distribution of sample means to be equal to?

a) $\sqrt{16} = 4$

b) 16

c) $\frac{16}{16} = 1$

d) $\frac{\sqrt{16}}{16} = \frac{1}{4}$

e) $\sqrt{\frac{16}{16}} = 1$

MC 2

What would we expect the population **variance** of the distribution of sample means to be equal to?

a) 16

b) $\frac{16}{36} = \frac{4}{9}$

c) $\frac{16}{16} = 1$

d) $\sqrt{\frac{16}{16}} = 1$

e) $\sqrt{\frac{16}{36}} = \frac{2}{3}$

MC 3

How would we expect our sample means to be **distributed** and why?

a) Normally distributed because X is normally distributed

b) Normally distributed by the CLT because we drew $16 < 30$ samples

c) NOT normally distributed by the CLT because we drew $16 < 30$ samples

d) Normally distributed by the CLT because our sample size per draw was $36 > 30$

e) NOT normally distributed by the CLT because our sample size per draw was $36 > 30$

MC 4

What **range** would we expect to contain roughly 99.7% of the values of sample means?

a) $16 \pm 3 \times 16$

b) $16 \pm 3 \times \sqrt{16}$

c) $16 \pm 3 \times \frac{16}{16}$

d) $16 \pm 3 \times \frac{16}{36}$

e) $16 \pm 3 \times \sqrt{\frac{16}{36}}$

MC 5

Which of the following is **true** about the population standard deviations of X and $\bar{X}$?

a) $\sigma_{\bar{x}} \geq \sigma_x$ always

b) $\sigma_{\bar{x}} = \sigma_x$ always

c) $\sigma_{\bar{x}} \leq \sigma_x$ **always**

d) The answer will depend on our sample size n

e) None of the above