Economics 102 A01-A04: Analysis of Economic Data
Cameron Fall 2022 October 13
Department of Economics, U.C.-Davis

First Midterm Exam (Version A)

Compulsory. Closed book. Total of 35 points and worth 20% of course grade.
Read question carefully so you answer the question.
You are to use only **simple calculations (+, -, /, *, square root) and show all workings**.
Use the calculators provided by the department.
For computations final answers should be to at least four significant digits.
You may remove the formula sheet and the Stata output sheet(s) at end of exam.

Question scores

Question	$1a$	$1b$	$1c$	$1d$	$1e$	$1f$	$2a$	$2b$	$2c$	$2d$	$2e$	$2f$	$3a$	$3b$	$3c$	$3d$	$3e$	*Mult.choice*
Points	1	1	1	2	2	2	1	1	1	3	1	3	3	2	1	2	3	5

1.(a) NBA championships have a best-of-seven format where the first team to win four games wins the championship. Suppose you have data on the outcome of the actual number of games in NBA championships for many NBA championships. Which Stata command provides a better graphical presentation of the data: `kdensity` or `histogram, discrete`?

(b) Can the Stata command `use` read in a comma-separated values data set? A simple yes or no will do.

(c) What transformation would you use to make right-skewed data more symmetric?

(d) Suppose a dataset on x_i, $i = 1, ..n$, has mean $\bar{x}$ and sample standard deviation s. What is gained by forming the variable $y_i = (x_i - \bar{x})/s$?

(e) Calculate $\sum_{i=1}^{3} (5 + 2i^2)$. **Show all workings.**

(f) Consider a simple random sample of size 4 with values $1, 2, 3$ and 6.
Compute the sample mean and the sample variance. **Show all workings.**

QUESTION 2 USES STATA OUTPUT GIVEN AT THE END OF THIS EXAM. In some cases the answer may be given directly in the output. In other cases you will need to use the output plus additional computation.

The variable `age` measures age in years for a sample of U.S. individuals aged 25 to 65 years.

2.(a) Give the lower quartile for variable `age`.

(b) Is the distribution of `age` substantially skewed? **Explain your answer.**

(c) Suppose you want to have a visual summary of the key sample statistics for `age`. What sort of graph or plot would you use?

(d) Provide a **90 percent** confidence interval for population mean age.

(e) What Stata command would directly provide a 95 percent confidence interval for mean age? **Provide the complete command.**

(f) The claim is made that mean age of 25-65 year-olds is 39 years. Test this claim at **significance level 0.05**.
State clearly the null and alternative hypotheses and your conclusion.

3.(a) State the three assumptions made about the random variable X_i that are made in the course notes and in class to yield the mean and variance of $\bar{X}$.

(b) Suppose $X = 2$ with probability 0.1, $X = 5$ with probability 0.6 and $X = 6$ with probability 0.3. What is the mean and variance of X? **Show all workings.**

(c) You are given the Stata commands

```
set obs 10000
generate u = runiform()
generate x = 0
replace x = 1 if u < 1/6
```

Give the approximate sample mean for variable `u`.

(d) Suppose for the fair coin toss experiment we form 1000 simple random samples of size 50, from these calculate 1000 95% confidence intervals, and find that 940 of these confidence intervals include the true population mean $\mu = 0.5$. Is this surprising? **Explain.**

(e) Consider a simple random sample of size 25 where X has mean 200 and variance 100. Give the mean, variance and standard deviation of $\overline{X}$.

Multiple Choice Questions (1 point each)

1. Data on annual earnings for a sample of individuals some of whom chose to participate in a training program and some who did not are an example of

a. experimental data

b. time series data

c. observational data.

2. For a simple random sample of variable X with mean μ and variance σ^2, the standard error of the mean is

a. an estimate of the standard deviation of X

b. an estimate of the standard deviation of $\bar{X}$

c. an estimate of the standard deviation of μ

d. none of the above.

3. Suppose X is normally distributed with mean 100 and variance 25. Then in a large random sample we expect around 68% of observations to lie in the range

a. (75, 125)

b. (50, 150)

c. (95, 105)

d. (90, 110)

e. none of the above.

4. A key feature of the approach to statistical inference for univariate data that are independent and identically distributed is that in different samples

a. the population mean is unchanging but the sample mean changes

b. the sample mean is unchanging but the population mean changes

c. both the population mean and the sample mean change

d. both the population mean and the sample mean are unchanging.

5. For the 1880 Census example studied in class and in notes

a. individual age is approximately normally distributed

b. average age in repeated samples of size 25 is approximately normally distributed

c. both of the above

d. neither of the above.

SOME USEFUL FORMULAS

Univariate Data

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

$\bar{x} \pm t_{\alpha/2;n-1} \times (s_x/\sqrt{n})$ *and* $t = \dfrac{\bar{x} - \mu^*}{s/\sqrt{n}}$

ttail$(df, t) = \Pr[T > t]$ where $T \sim t(df)$

$t_{\alpha/2}$ such that $\Pr[|T| > t_{\alpha/2}] = \alpha$ is calculated using invttail$(df, \alpha/2)$.

Bivariate Data

$r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \dfrac{s_{xy}}{s_x \times s_y}$ [Here $s_{xx} = s_x^2$ and $s_{yy} = s_y^2$].

$\widehat{y} = b_1 + b_2 x_i$ $b_2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $b_1 = \bar{y} - b_2\bar{x}$

TSS $= \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ ResidualSS $= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ ExplainedSS = TSS - ResidualSS

$R^2 = 1-$ ResidualSS/TSS

$b_2 \pm t_{\alpha/2;n-2} \times s_{b_2}$

$t = \dfrac{b_2 - \beta_2}{s_{b_2}}$ $s_{b_2}^2 = \dfrac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$

$y|x = x^* \in b_1 + b_2 x^* \pm t_{\alpha/2;n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_i (x_i - \bar{x})^2} + 1}$

$\mathrm{E}[y|x = x^*] \in b_1 + b_2 x^* \pm t_{\alpha/2;n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$

Multivariate Data

$\widehat{y} = b_1 + b_2 x_{2i} + \cdots + b_k x_{ki}$

$R^2 = 1-$ ResidualSS/TSS $\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$

$b_j \pm t_{\alpha/2;n-k} \times s_{bj}$ *and* $t = \dfrac{b_j - \beta_{j0}}{s_{b_j}}$

$F = \dfrac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k)$

$F = \dfrac{(\text{ResSS}_r - \text{ResSS}_u)/(k-g)}{\text{ResSS}_u/(n-k)} \sim F(k-g, n-k)$

Ftail$(df1, df2, f) = \Pr[F > f]$ where F is $F(df1, df2)$ distributed.

F_α such that $\Pr[F > f_\alpha] = \alpha$ is calculated using invFtail$(df1, df2, \alpha)$.

```
. summarize age, detail

                        Age in years
-------------------------------------------------------------
      Percentiles      Smallest
 1%           25             25
 5%           26             25
10%           27             25       Obs               1,194
25%           30             25       Sum of wgt.       1,194

50%           37                      Mean           39.63233
                        Largest       Std. dev.        10.586
75%           49             61
90%           55             61       Variance       112.0634
95%           59             61       Skewness       .3885907
99%           61             62       Kurtosis       1.902251

. di _n "Square Root of " N " = " sqrt(N) _n   ///
>       "Square Root of " N-1 " = " sqrt(N-1) _n

Square Root of 1194 = 34.554305
Square Root of 1193 = 34.539832


. di _n "KEY CRITICAL VALUES FOR THIS EXAM" _n _n       ///
>       "t_" dof ",.005 = " %5.3f invttail(dof,.005) _n ///
>       "t_" dof ",.01  = " %5.3f invttail(dof,.01) _n  ///
>       "t_" dof ",.025 = " %5.3f invttail(dof,.025) _n ///
>       "t_" dof ",.05  = " %5.3f invttail(dof,.05) _n  ///
>       "t_" dof ",.10  = " %5.3f invttail(dof,.10) _n

KEY CRITICAL VALUES FOR THIS EXAM

t_1193,.005 = 2.580
t_1193,.01  = 2.329
t_1193,.025 = 1.962
t_1193,.05  = 1.646
t_1193,.10  = 1.282
```