Economics 640 San Diego State University
LIANG Fall 2025

Econ 640 - Problem Set 1

Instructions: *Please submit all answers as a* ***physical*** *copy in a separate document. For Stata output, ensure that your do-files and results are in one place. Full credit will only be awarded to answers that demonstrate clear reasoning and a solid understanding of the problem. This problem set is due at the* ***beginning of class on Thursday, September 18th****.*

1. (Discrete) Consider a discrete random variable X with support $\{0, 1, 2, 3, 4\}$ and pdf

x	0	1	2	3	4
$f_X(x)$	0.10	0.20	0.30	0.25	0.15

   (a) Compute $\mathbb{E}[X]$ and $Var(X)$.
   (b) Let $Y = \mathbf{1}\{X \geq 2\}$. Compute $\mathbb{E}[Y]$ and (Y).
   (c) Compute $\mathbb{E}[XY]$ and $\mathbb{E}[X]\,\mathbb{E}[Y]$ and compare. Why do they differ?
   (d) Find $P(X = 3 \mid Y = 1)$ and $P(X = 3, Y = 1)$.
   (e) Suppose you take i.i.d. samples of X and compute the sample mean $\bar{X}_n$. Is $\bar{X}_n$ an unbiased estimator of $\mathbb{E}[X]$?
   (f) Consider the alternative estimator X_1 (the first observation). Is X_1 unbiased for $\mathbb{E}[X]$? Explain.
   (g) Are $\bar{X}_n$ and X_1 consistent? Which is more efficient?
   (h) (Computation) Use Stata to demonstrate the efficiency difference by simulating the sampling distributions of $\bar{X}_n$ and X_1.

2. (Continuous) Let X be a continuous random variable on $(0, 1)$ with pdf

$$f_X(x) = 2x, \qquad 0 < x < 1.$$

   (a) Compute $\mathbb{E}[X]$ and $Var(X)$.
   (b) Let $Y = \mathbf{1}\{X > 0.6\}$. Compute $\mathbb{E}[Y]$ and $Var(Y)$.

3. **(2024 U.S. election context; using simulated counts)** Analysts noted meaningful differences in candidate support by *voting method* (mail/early vs. Election Day) in 2024.[1] Suppose you survey $N = 1000$ voters and record their age group, voting method, and presidential vote between **Harris** and **Trump**. You obtain the following counts:

Mail / Early voters $(n = 500)$

Age Group	Harris	Trump	Total
18–29	65	35	100
30–44	80	60	140
45–64	85	95	180
65+	30	50	80
Total	260	240	500

Election Day voters $(n = 500)$

Age Group	Harris	Trump	Total
18–29	35	45	80
30–44	50	70	120
45–64	55	95	150
65+	35	115	150
Total	175	325	500

(a) **Marginals.** Compute the marginal distribution of support for each candidate overall (i.e., % Harris vs. % Trump among all 1000).

(b) **Age structure.** Collapse across method and compute, *within each age group*, the share voting for Harris and for Trump. Briefly compare to the marginal distribution—what heterogeneity do you see?

(c) **Voting method effect.** Using the 2×2 table (Method $\in${Mail/Early, Election Day} $\times$ Candidate $\in${Harris, Trump}), compute: (i) row and column percentages, (ii) the *odds ratio* of voting Trump for Election Day vs. Mail/Early.

(d) **Visualization & simulation in Stata.** Simulate a dataset that mirrors the *cell probabilities* in the tables (not necessarily the exact counts), then produce: (i) a stacked bar chart of candidate share by method; (ii) a stacked bar chart of candidate share by age group within each method; (iii) a logistic regression of $\mathbb{1}\{\text{Trump vote}\}$ on method, age groups, and their interaction; use `margins` and `marginsplot` to visualize Pr(Trump) by age $\times$ method.

Starter Stata code:

```
clear all
set seed 20250908
set obs 1000
```

[1]Synthetic table below is stylized but consistent with public summaries from AP VoteCast and election-process reports.

```
* 0 = Mail/Early, 1 = Election Day
gen method = (runiform()>=0.5)
label define method 0 "Mail/Early" 1 "Election Day"
label values method method

* Assign age conditional on method to match table row shares
gen age = .
gen u = runiform()
replace age = 1 if method==0 & u<0.20
replace age = 2 if method==0 & u>=0.20 & u<0.48
replace age = 3 if method==0 & u>=0.48 & u<0.84
replace age = 4 if method==0 & u>=0.84
gen v = runiform()
replace age = 1 if method==1 & v<0.16
replace age = 2 if method==1 & v>=0.16 & v<0.40
replace age = 3 if method==1 & v>=0.40 & v<0.70
replace age = 4 if method==1 & v>=0.70
label define age 1 "18-29" 2 "30-44" 3 "45-64" 4 "65+"
label values age age

* Harris probability by stratum (from tables)
gen w = runiform()
gen harris = .
replace harris = (w<0.6500) if method==0 & age==1
replace harris = (w<0.5714) if method==0 & age==2
replace harris = (w<0.4722) if method==0 & age==3
replace harris = (w<0.3750) if method==0 & age==4
replace harris = (w<0.4375) if method==1 & age==1
replace harris = (w<0.4167) if method==1 & age==2
replace harris = (w<0.3667) if method==1 & age==3
replace harris = (w<0.2333) if method==1 & age==4
//... ...
```

4. This exercise should help you understanding the properties of *summations*. Remember that

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_{n-1} + X_n$$

Consider the following sequences of variables

$$\begin{array}{lll} X_1 = 1 & Y_1 = 1 & Z_1 = 3 \\ X_2 = 0 & Y_2 = 2 & Z_2 = 3 \\ X_3 = 2 & & Z_3 = 3 \\ & & Z_4 = 3 \\ & & Z_5 = 3 \end{array}$$

You should show each of the following things two ways, first using the formulas and then using the actual numbers.

(a) Show that

$$\sum_{i=1}^{5} Z_i = 5Z_1$$

(b) Show that

$$\sum_{i=1}^{3}\sum_{j=1}^{2} X_i Y_j = \left(\sum_{i=1}^{3} X_i\right)\left(\sum_{j=1}^{2} Y_j\right) = \sum_{j=1}^{2}\left(Y_j \sum_{i=1}^{3} X_i\right)$$

(c) Show that

$$\sum_{i=1}^{3}\left(\frac{X_i}{\sum_{j=1}^{2} Y_j}\right) = \frac{\sum_{i=1}^{3} X_i}{\sum_{j=1}^{2} Y_j}$$

(d) Show that

$$\sum_{i=1}^{3}\sum_{j=1}^{4} X_i Z_j = 12 \sum_{i=1}^{3} X_i$$