Handout 1: Types of Data & Review of Probability

EC 282: Introduction to Econometrics

Spring 2026

Instructions: Run the provided R code and answer the questions. Show your work for calculations.

1 Types of Data

Run the code below to generate three different datasets.

```r
library(dplyr)
library(tidyr)
library(ggplot2)
set.seed(282)

# CROSS-SECTIONAL DATA
cross_section <- data.frame(
  student_id = 1:20,
  gpa = round(runif(20, 2.0, 4.0), 2),
  hours_studied = round(rnorm(20, mean = 15, sd = 5), 1),
  employed = sample(c(0, 1), 20, replace = TRUE, prob = c(0.6, 0.4))
)

# TIME SERIES DATA
time_series <- data.frame(
  month = seq(as.Date("2024-01-01"), by = "month", length.out = 24),
  unemployment_rate = round(4.5 + cumsum(rnorm(24, 0, 0.3)), 2)
)

# PANEL DATA
panel_data <- expand.grid(
  student_id = 1:5,
  semester = c("Fall 2024", "Spring 2025", "Fall 2025", "Spring 2026")
) %>%
  arrange(student_id, semester) %>%
  mutate(
    gpa = round(2.5 + 0.1 * as.numeric(factor(semester)) +
          rnorm(n(), 0, 0.3) + rep(rnorm(5, 0, 0.5), each = 4), 2),
    gpa = pmin(pmax(gpa, 0), 4.0)
  )
```

Question 1.1: Examine the `cross_section` data. What makes this "cross-sectional"? Give two other examples of cross-sectional data in economics.

Question 1.2: Examine the `time_series` data. What makes this "time series"? Why might the observations in a time series NOT be independent?

Question 1.3: Examine the `panel_data`. How does panel data combine features of both cross-sectional and time series data? What are the advantages of having panel data?

2 Random Variables and Probability Distributions

Run the code below to create a population with a known disease rate.

```
N <- 100000
p_cancer <- 0.04

population <- data.frame(
  id = 1:N,
  colon_cancer = sample(c(0, 1), N, replace = TRUE,
                        prob = c(1 - p_cancer, p_cancer))
)
```

Question 2.1: The variable `colon_cancer` is a Bernoulli random variable. Define what a Bernoulli random variable is and identify the parameter p in this case.

Question 2.2: Using the population data, calculate:

(a) The population mean $\mu_Y = E[Y]$

(b) The population variance $\sigma_Y^2 = \text{Var}(Y)$

Question 2.3: For a Bernoulli random variable, show mathematically that $E[Y] = p$ and $\text{Var}(Y) = p(1-p)$. Verify that your calculated values are close to the theoretical values.

3 Joint and Marginal Distributions

Consider the relationship between weather (Rain/No Rain) and commute time (Long/Short). The joint distribution is given below:

	$Y = 0$ (Long)	$Y = 1$ (Short)	Marginal of X
$X = 0$ (Rain)	0.15	0.15	?
$X = 1$ (No Rain)	0.07	0.63	?
Marginal of Y	?	?	1.00

```
joint_dist <- matrix(
  c(0.15, 0.15, 0.07, 0.63),
  nrow = 2, byrow = TRUE,
  dimnames = list(
    X = c("Rain (X=0)", "No Rain (X=1)"),
    Y = c("Long (Y=0)", "Short (Y=1)")
  )
)
```

Question 3.1: Calculate the marginal distribution of X (weather) and the marginal distribution of Y (commute time). Fill in the missing values in the table above.

Question 3.2: Verify that all joint probabilities sum to 1 and that each marginal distribution sums to 1.

4 Conditional Probability and Bayes' Theorem

Using the joint distribution from Part 3:

Question 4.1: Calculate the following conditional probabilities:

(a) $P(\text{Short Commute} \mid \text{Rain}) = P(Y = 1 \mid X = 0)$

(b) $P(\text{Short Commute} \mid \text{No Rain}) = P(Y = 1 \mid X = 1)$

(c) $P(\text{Long Commute} \mid \text{Rain}) = P(Y = 0 \mid X = 0)$

Question 4.2: Does knowing the weather provide useful information about commute time? Explain using the conditional probabilities you calculated.

Question 4.3: Using Bayes' Theorem, calculate $P(\text{Rain} \mid \text{Long Commute})$. Show your work.
Hint: Bayes' Theorem states that $P(A \mid B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

5 Conditional Expected Value and Law of Iterated Expectations

Consider rolling a fair six-sided die. Define:

- $Y \in \{1, 2, 3, 4, 5, 6\}$ as the outcome
- $X = 0$ if Y is even, $X = 1$ if Y is odd

Question 5.1: Calculate $E[Y]$ directly using the definition of expected value.

$$E[Y] = \sum_{i=1}^{6} y_i \cdot P(Y = y_i)$$

Question 5.2: Calculate:

(a) $E[Y \mid X = 1]$ (expected value given the roll is odd)

(b) $E[Y \mid X = 0]$ (expected value given the roll is even)

Question 5.3: Use the Law of Iterated Expectations to calculate $E[Y]$:

$$E[Y] = E[Y \mid X = 0] \cdot P(X = 0) + E[Y \mid X = 1] \cdot P(X = 1)$$

Verify that this equals your answer from Question 5.1.

6 Independence, Covariance, and Correlation

Question 6.1: Two random variables X and Y are independent if $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all x, y. Using the commute/rain example from Part 3, test whether X (weather) and Y (commute time) are independent.

Question 6.2: Run the code below to generate data on study hours and exam scores:

```r
n_students <- 500
hours_studied <- rnorm(n_students, mean = 10, sd = 3)
exam_score <- 50 + 3 * hours_studied + rnorm(n_students, 0, sd = 10)

cov(hours_studied, exam_score)  # Covariance
cor(hours_studied, exam_score)  # Correlation
```

(a) What is the covariance between hours studied and exam score?

(b) What is the correlation?

(c) Interpret the sign and magnitude of the correlation.

Question 6.3: If $\text{Cov}(X, Y) = 0$, does this mean X and Y are independent? Run the code below and explain.

```
x_vals <- runif(1000, -1, 1)
y_vals <- x_vals^2

cov(x_vals, y_vals)
cor(x_vals, y_vals)
```

7 Sampling and the Law of Large Numbers

Using the population created in Part 2:

Question 7.1: Draw a random sample of $n = 100$ from the population and calculate the sample mean.

```
sample_100 <- sample(population$colon_cancer, 100)
mean(sample_100)
```

Is your sample mean exactly equal to the population mean? Why or why not?

Question 7.2: Run the code below to see how the sample mean changes as sample size increases:

```
sample_sizes <- c(50, 100, 200, 500, 1000, 5000, 10000, 50000)
pop_mean <- mean(population$colon_cancer)

sample_means <- sapply(sample_sizes, function(n) {
  mean(sample(population$colon_cancer, n))
})

data.frame(n = sample_sizes, sample_mean = sample_means, pop_mean = pop_
    mean)
```

What pattern do you observe? State the Law of Large Numbers in your own words.

Question 7.3: Draw 1000 samples of size $n = 100$ and calculate the mean of each sample:

```
sample_means_dist <- replicate(1000, mean(sample(population$colon_cancer,
    100)))

mean(sample_means_dist)   # Mean of sample means
var(sample_means_dist)    # Variance of sample means
hist(sample_means_dist)   # Histogram
```

(a) How does the mean of the sample means compare to the population mean?

(b) Compare the variance of the sample means to σ^2/n where σ^2 is the population variance. What do you notice?

(c) What shape does the histogram have? Why?