# Handout 1: Types of Data & Review of Probability

**ANSWER KEY**

EC 282: Introduction to Econometrics

Spring 2026

## 1 Types of Data

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

Cross-sectional data consists of observations on multiple units (students) at a **single point in time** (one semester). Each row represents a different individual, and we observe their characteristics at the same moment.

**Examples:**

- Census data: income, education, and demographics of households in a given year
- Survey of firms: employment, revenue, and industry for companies in 2024
- Housing prices: sale prices and characteristics of homes sold in a city during one month

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

Time series data consists of observations on a **single unit** (one state's economy) over **multiple time periods** (24 months). The key feature is tracking the same entity through time.

Observations in time series are often NOT independent because:

- **Serial correlation**: today's unemployment is correlated with yesterday's
- **Trends**: economic variables often trend upward or downward over time
- **Seasonality**: patterns that repeat (e.g., retail sales peak in December)
- **Persistent shocks**: an economic shock affects multiple future periods

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

Panel data has:

- Multiple units (5 students) like cross-sectional data
- Multiple time periods (4 semesters) like time series data
- Each unit is observed in each time period

**Advantages of panel data:**

- Control for unobserved individual heterogeneity (fixed effects)

- Larger sample size than pure time series
- Can study dynamics: how individuals change over time
- Better for causal inference: can compare same person before/after treatment

## 2 Random Variables and Probability Distributions

**Question 2.1:** Define a Bernoulli random variable and identify the parameter $p$.

A **Bernoulli random variable** is a discrete random variable that takes only two values: 0 and 1 (or "failure" and "success"). It has a single parameter $p = P(Y = 1)$, the probability of success.

In this case: $\mathbf{p = 0.04}$ (the probability of having colon cancer)

**Question 2.2:** Calculate population mean and variance.

(a) **Population mean:**

```
pop_mean <- mean(population$colon_cancer)
```

The population mean is exactly **0.04** (equal to $p = 0.04$ by construction)

(b) **Population variance:**

```
pop_var <- var(population$colon_cancer) * (N-1) / N
```

The population variance is exactly **0.0384** (equal to $p(1-p) = 0.04 \times 0.96 = 0.0384$)

**Question 2.3:** Show mathematically that $E[Y] = p$ and $\text{Var}(Y) = p(1-p)$.

**For $E[Y]$:**

$$E[Y] = \sum_y y \cdot P(Y = y) = 0 \cdot (1-p) + 1 \cdot p = \boxed{p}$$

**For Var$(Y)$:**

$$\begin{aligned}\text{Var}(Y) &= E[(Y-\mu)^2] = (0-p)^2(1-p) + (1-p)^2 \cdot p \\ &= p^2(1-p) + (1-p)^2 p = p(1-p)[p + (1-p)] = \boxed{p(1-p)}\end{aligned}$$

**Verification:**

- Theoretical $E[Y] = 0.04$; Calculated $= 0.04$ ✓
- Theoretical $\text{Var}(Y) = 0.04 \times 0.96 = 0.0384$; Calculated $= 0.0384$ ✓

## 3 Joint and Marginal Distributions

**Question 3.1:** Calculate the marginal distributions.

| | $Y = 0$ (Long) | $Y = 1$ (Short) | Marginal of $X$ |
|---|---|---|---|
| $X = 0$ (Rain) | 0.15 | 0.15 | **0.30** |
| $X = 1$ (No Rain) | 0.07 | 0.63 | **0.70** |
| Marginal of $Y$ | **0.22** | **0.78** | 1.00 |

**Marginal distribution of $X$ (Weather):**

- $P(\text{Rain}) = P(X = 0) = 0.15 + 0.15 = \mathbf{0.30}$
- $P(\text{No Rain}) = P(X = 1) = 0.07 + 0.63 = \mathbf{0.70}$

**Marginal distribution of $Y$ (Commute):**

- $P(\text{Long}) = P(Y = 0) = 0.15 + 0.07 = \mathbf{0.22}$
- $P(\text{Short}) = P(Y = 1) = 0.15 + 0.63 = \mathbf{0.78}$

**Question 3.2:** Verify that probabilities sum to 1.

- Sum of joint probabilities: $0.15 + 0.15 + 0.07 + 0.63 = 1.00$ ✓
- Sum of marginal $X$: $0.30 + 0.70 = 1.00$ ✓
- Sum of marginal $Y$: $0.22 + 0.78 = 1.00$ ✓

# 4 Conditional Probability and Bayes' Theorem

**Question 4.1:** Calculate the conditional probabilities:

(a) $P(\text{Short} \mid \text{Rain})$:

$$P(Y = 1 \mid X = 0) = \frac{P(X = 0, Y = 1)}{P(X = 0)} = \frac{0.15}{0.30} = \boxed{0.50}$$

(b) $P(\text{Short} \mid \text{No Rain})$:

$$P(Y = 1 \mid X = 1) = \frac{P(X = 1, Y = 1)}{P(X = 1)} = \frac{0.63}{0.70} = \boxed{0.90}$$

(c) $P(\text{Long} \mid \text{Rain})$:

$$P(Y = 0 \mid X = 0) = \frac{P(X = 0, Y = 0)}{P(X = 0)} = \frac{0.15}{0.30} = \boxed{0.50}$$

**Question 4.2:** Does knowing the weather provide useful information about commute time?

**Yes!** Weather provides valuable information:

- Without weather info: $P(\text{Short}) = 0.78$
- Given rain: $P(\text{Short} \mid \text{Rain}) = 0.50$
- Given no rain: $P(\text{Short} \mid \text{No Rain}) = 0.90$

Knowing it's raining substantially **decreases** the probability of a short commute (from 78% to 50%), while knowing it's not raining **increases** it to 90%. This shows $X$ and $Y$ are **not independent**.

**Question 4.3:** Using Bayes' Theorem, calculate $P(\text{Rain} \mid \text{Long Commute})$.

$$P(\text{Rain} \mid \text{Long}) = \frac{P(\text{Long} \mid \text{Rain}) \cdot P(\text{Rain})}{P(\text{Long})}$$

$$= \frac{0.50 \times 0.30}{0.22} = \frac{0.15}{0.22} \approx \boxed{0.682}$$

**Interpretation:** If someone had a long commute, there's about a 68% chance it was raining.

## 5 Conditional Expected Value and Law of Iterated Expectations

**Question 5.1:** Calculate $E[Y]$ directly.

$$E[Y] = \sum_{i=1}^{6} y_i \cdot P(Y = y_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \boxed{3.5}$$

**Question 5.2:** Calculate conditional expected values:

(a) $E[Y \mid X = 1]$ (given odd):

$$E[Y \mid \text{Odd}] = \frac{1}{3}(1+3+5) = \frac{9}{3} = \boxed{3}$$

(b) $E[Y \mid X = 0]$ (given even):

$$E[Y \mid \text{Even}] = \frac{1}{3}(2+4+6) = \frac{12}{3} = \boxed{4}$$

**Question 5.3:** Law of Iterated Expectations:

$$\begin{aligned} E[Y] &= E[Y \mid X = 0] \cdot P(X = 0) + E[Y \mid X = 1] \cdot P(X = 1) \\ &= 4 \times 0.5 + 3 \times 0.5 \\ &= 2 + 1.5 = \boxed{3.5} \end{aligned}$$

This **matches** our direct calculation of $E[Y] = 3.5$, confirming the Law of Iterated Expectations:

$$E[Y] = E\big[E[Y \mid X]\big]$$

## 6 Independence, Covariance, and Correlation

**Question 6.1:** Test whether $X$ (weather) and $Y$ (commute time) are independent.

For independence, we need $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all $x, y$.

**Check:**

$$P(\text{Rain}) \times P(\text{Short}) = 0.30 \times 0.78 = 0.234$$
$$P(\text{Rain AND Short}) = 0.15$$

Since $0.234 \neq 0.15$, $X$ **and** $Y$ **are NOT independent**. Rain affects commute time.

**Question 6.2:** Covariance and Correlation

(a) Covariance $\approx 27$ (exact value depends on random seed)

(b) Correlation $\approx 0.67$ (exact value depends on random seed)

(c) **Interpretation:**

- Positive correlation: students who study more tend to score higher
- Magnitude around 0.65–0.70 indicates a **strong positive** linear relationship
- Correlation is unitless, so changing hours to minutes wouldn't change $r$

**Question 6.3:** If $\text{Cov}(X, Y) = 0$, does this mean $X$ and $Y$ are independent?

**No!** The covariance and correlation are approximately 0, but $Y = X^2$! They are clearly NOT independent—knowing $X$ tells you exactly what $Y$ is.

This illustrates that covariance/correlation only measure **linear** relationships. $X$ and $X^2$ have a perfect **nonlinear** (quadratic) relationship, but zero linear correlation because positive and negative $X$ values both give positive $Y$ values, canceling out.

**Key takeaway:** Zero covariance $\Rightarrow$ no linear relationship, but does NOT imply independence.

# 7 Sampling and the Law of Large Numbers

**Question 7.1:** Is your sample mean exactly equal to the population mean? Why or why not?

The sample mean is typically **NOT** exactly equal to the population mean.

This happens because:

- Random sampling introduces **sampling variability**
- Each sample is just one possible realization from the population
- The sample mean $\bar{Y}$ is itself a **random variable** with its own distribution

**Question 7.2:** State the Law of Large Numbers in your own words.

**Pattern:** As sample size increases, the sample mean gets closer to the population mean.

**Law of Large Numbers:** As $n \to \infty$, the sample mean $\bar{Y}$ converges in probability to the population mean $\mu_Y$:

$$\bar{Y} \xrightarrow{p} \mu_Y$$

In plain language: **larger samples give more accurate estimates** of the population mean.

**Question 7.3:** Sampling distribution of the mean:

(a) The mean of sample means is very close to the population mean ($= 0.04$). This confirms that $E[\bar{Y}] = \mu_Y$ (the sample mean is an **unbiased estimator**).

(b) The variance of sample means is close to $\sigma^2/n$:

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} = \frac{0.0384}{100} = 0.000384$$

This confirms the theoretical result about the variance of the sampling distribution.

(c) The histogram is approximately **bell-shaped (normal)**. This is due to the **Central Limit Theorem**: regardless of the population distribution (which is Bernoulli here), the sampling distribution of the mean is approximately normal for large samples.

$$\bar{Y} \overset{a}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$