Data 407 - Assignment 1:

Samira Almuallim, 62197256, Data 407 2025WT2

1. (2 marks) Regarding measurement error, which of the following types are true?

Answers: a, b, c, & d

- **a. Respondent error** - is when respondent gives incorrect or inaccurate answers
- **b. Interviewer error** - is when the interviewer influences or records responses incorrectly
- **c. Instrument error** - is when faulty or poorly calibrated measurement tools are used
- **d. Operator error** - is when human error in using the instrument or recording data happens
- **e. Sampling error** - is **NOT** a measurement error; it arises from random sampling variability.

Thus: the answers are all except E

Answers: a, b, c, & d

2. (2 marks) Regarding selection bias, which of the following types are true?

Answers: a, b, c, & d

- **a. Sampling bias** - is when the sample is not representative of the population
- **b. Selective survival** - is when only certain units remain observable (survivorship bias).
- **c. Volunteer bias** - is when participants differ systematically from non-participants
- **d. Non-response bias** - is when respondents differ from non-respondents
- **e. Sampling error** - is **NOT** random variation, not bias

Thus: the answers are all except E

Answers: a, b, c, & d

3. (2 marks) Regarding simple random sampling without replacement (SRS) from a population of N units, what is the probability that the first unit will appear in the sample of size n ?

Answer: a. $\frac{n}{N}$

Explanation:
In simple random sampling without replacement, each unit has the same inclusion probability:

$$P(\text{unit included}) = \frac{n}{N}$$

The phrase “first unit” does not change this probability

And the replacement component is irrelevant here since we are only considering the 1 case

4. (2 marks) Regarding simple random sampling without replacement (SRS) from a population of N units, what is the probability that the first two units will appear in the sample of size n?

Answer: e. $\frac{n(n-1)}{N(N-1)}$

Explanation:

- Probability first unit is included

$$\frac{n}{N}$$

- Given that, the probability the second unit is also included

$$\frac{n-1}{N-1}$$

- the joint prob is

$$\frac{n}{N} \cdot \frac{n-1}{N-1} = \frac{n(n-1)}{N(N-1)}$$

- Thus, the Answer is

$$\frac{n(n-1)}{N(N-1)}$$

5. (2 marks) In SRS, is the sample mean _ an unbiased estimate

$$\bar{y} = \frac{1}{n} \sum_{i \in S} y_i$$

Answer: Yes

Under simple random sampling without replacement $E(\bar{y}) = \bar{Y}$

And since $E(\bar{y}) = avg(y_i)$

The sample mean is an unbiased estimator of the population mean

Thus: **Ans: a. Yes**

6. (2 marks) Regarding SRS from a population of N units with population variance S_U^2, what is the variance of the sample mean $\bar{y} = \frac{1}{n} \sum_{i \in S} y_i$?

Answer: b. $\left(1 - \frac{n}{N}\right) \frac{S_U^2}{n}$

For SRS without replacement, the variance of the sample mean includes the finite population correction:
$B.Var(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S_U^2}{n}$
$D.Var(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S_U^2}{N}$

And we know the denominator is n not N

Thus
Ans: $b.Var(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S_U^2}{n}$

7. (3 marks) In SRS, $N = 1000$, $n = 100$, $\bar{y} = 1$, $s = 0.5$

a. Find an approximate 95% CI for $\bar{y}_U$ using the normal approximation:

For SRS without replacement, the SE of sample mean is

$$SE(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

We Substitute the given values:

$$SE(\bar{y}) = \sqrt{\left(1 - \frac{100}{1000}\right)\frac{(0.5)^2}{100}} = \sqrt{0.9 \cdot \frac{0.25}{100}} = \sqrt{0.00225} \approx 0.0474$$

95% CI implies `0.95+0.025+0.025`
Using the normal approximation with $z_{0.95+0.025} = z_{0.975} = 1.96$,
$\text{Margin of error} = 1.96 \times 0.0474 \approx 0.093$
(3sf)

95% CI for $\bar{y}_U$:
$1 \pm 0.093 = (0.907,\ 1.093)$

Ans: $(0.907,\ 1.093)$

b. Find an approximate 95% CI for $N\bar{y}_U$ using the normal approximation

The estimator of the population total is

$$\widehat{Y_U} = N\bar{y}$$

The standard error scales by N:

$$SE(N\bar{y}) = N \cdot SE(\bar{y}) = 1000 \times 0.0474 = 47.4$$

So the error margin at 95% CI is

$$1.96 \times 47.4 \approx 93.0$$

95% CI for $N\bar{y}_U$:
$1000 \pm 93 = (907,\ 1093)$

Ans: $(907,\ 1093)$

8. (3 marks)

We consider random variables $X_1, \ldots, X_n$ with dependence structure

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} \sigma^2, & i = j, \\ 0.5\,\sigma^2, & |i-j| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

References used: https://en.wikipedia.org/wiki/Central_limit_theorem#CLT_under_weak_dependence

A. Verify that

$$Var\left(\sum_{i=a+1}^{a+n} X_i\right) \approx nA^2 \quad \text{uniformly in } a\ (n \to \infty),\ A^2 > 0$$

First lets take the Var of whats inside the sum

$$Var\left(\sum_{i=a+1}^{a+n} X_i\right) = \sum_{i=a+1}^{a+n} Var(X_i) + 2 \sum_{a+1 \le i < j \le a+n} Cov(X_i, X_j)$$

Then Lets split this into two parts:

- **Diagonal terms:**

$$\sum_{i=a+1}^{a+n} Var(X_i) = n\sigma^2$$

- **Off-diagonal terms:**
  //Only adjacent indices contribute

$$\sum_{a+1 \le i < j \le a+n} Cov(X_i, X_j) = \sum_{i=a+1}^{a+n-1} 0.5\sigma^2 = (n-1)\,0.5\sigma^2.$$

Including the factor of 2

$$2 \times (n-1)\, 0.5\sigma^2 = (n-1)\sigma^2.$$

Thus,

$$Var\left(\sum_{i=a+1}^{a+n} X_i\right) = n\sigma^2 + (n-1)\sigma^2 = (2n-1)\sigma^2$$

As $n \to \infty$,

$$Var\left(\sum_{i=a+1}^{a+n} X_i\right) \sim 2n\sigma^2$$

Hence,

$$A^2 = 2\sigma^2 > 0,$$

and the approximation holds **uniformly in a** .

b. Suppose $E(X_i) = 0$. Use Serfling's Theorem to show that

$$\frac{\sum_{i=1}^{n} X_i}{\sqrt{2n\sigma^2}} \xrightarrow{d} N(0,1)$$

From part A

$$Var\left(\sum_{i=1}^{n} X_i\right) \sim 2n\sigma^2.$$

The sequence $\{X_i\}$ is **stationary**, **short-range dependent**, and has:

- finite second moments,
- absolutely summable covariances:

$$\sum_{k=-\infty}^{\infty} |Cov(X_0, X_k)| = \sigma^2 + 2(0.5\sigma^2) < \infty$$

Therefore, the conditions of **Serfling's CLT for dependent sequences** are satisfied.

Hence,

$$\frac{\sum_{i=1}^{n} X_i - E(\sum_{i=1}^{n} X_i)}{\sqrt{Var(\sum_{i=1}^{n} X_i)}} \xrightarrow{d} N(0,1)$$

Since $E(X_i) = 0$,

$$\frac{\sum_{i=1}^{n} X_i}{\sqrt{2n\sigma^2}} \xrightarrow{d} N(0,1)$$

Therefore, proven