Exercise Sheet: Generating and Measuring Data

Data, Data Storage, Data Collection

Exercise 1 – Remote Working and Employee Stress

A research team wants to investigate the following question:

> *Does working from home reduce employees' stress during the working day?*

The team proposes the following study:

1. Recruit volunteers through an advertisement posted on the company's internal website.
2. Ask each volunteer whether they prefer to work from home or from the office.
3. Ask participants to complete a stress questionnaire at the end of each working day, specifying whether they worked from home or in the office on that day.

(**Q1**) What is the target population, what is the access frame, and what is the sample?
(**Q2**) Is the collected data observational, experimental or simulated? Explain.
(**Q3**) Identify at least two distinct sources of error or limitation in this study. For each, explain why collecting more participants would or would not solve the problem.

Exercise 2 – Personalized Product Recommendation

An online retailer wants to know whether a new *personalized product recommendation* feature increases average order value. The data team lets customers opt in to a beta program to try the new feature, then compares the average order value of beta users against non-beta users over the following month.

(**Q1**) Is the collected data observational, experimental or simulated? Explain.
(**Q2**) Beta users spend, on average, 40% more per order than non-beta users. Identify the independent and dependent variables. Propose one plausible confounding variable.
(**Q3**) Redesign the study so that it supports a causal claim. What, concretely, would the retailer need to change? Does it change the status (observational, experimental or simulated) of the data?
(**Q4**) Suppose the retailer randomly assigns the feature to customers who have already made at least one purchase in the last year. How well will the study generalize to all customers? Explain.

Exercise 3 – Fitness Tracker for Heart Rate

A fitness tracker measures heart rate by sampling once every 2 seconds.

(**Q1**) A brief, genuine spike in heart rate lasting under one second (e.g., from a sudden startle) occurs. Will the tracker record it? Is this a bias, a precision problem, or something else?
(**Q2**) What is the sampling frequency of the monitoring system? What is the highest heart-rate-related frequency this tracker can faithfully capture, according to the Nyquist–Shannon sampling theorem?

(**Q3**) A signal $\cos(2\pi f_0 t)$ of true frequency $f_0 = 13$ Hz is sampled at $f_s = 10$ Hz. Identify the frequency $f_{\text{alias}} < f_s/2$ of a cosine that produces the exact same sampled values.

(**Q4**) Generalize: for a signal of frequency f_0 sampled at rate f_s, propose a formula for the apparent frequency f_{alias} in terms of f_0 and f_s, valid when $f_s/2 < f_0 < f_s$. Check your formula against the previous question.

Exercise 4 – Monte Carlo: How Many Draws Do We Need?

In this exercise, we assume known Chebyshev's inequality

$$\Pr\left(|X - \mathbb{E}[X]| \geq \epsilon\right) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

One standard proof relies on Markov's inequality.

(**Q1**) Let $X_1, \ldots, X_n$ be independent Bernoulli random variables of parameter p, and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Using Chebyshev's inequality, show that, for any $\epsilon > 0$,

$$\Pr\left(|\hat{p} - p| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}.$$

(**Q2**) Get a bound for $|\hat{p} - p|$ that holds with probability at least $1 - \frac{1}{t^2}$. What can you conclude for the order of the error? What hypothesis from the lecture was not needed?

(**Q3**) You want to estimate π using the disk method from class, with an error of at most $\epsilon = 0.01$, and you want this guarantee to hold with probability at least 95% (i.e., the bound above should be at most 0.05). How many draws n do you need?

(**Q4**) Your colleague argues: "Since Chebyshev only requires independence and finite variance, and gives a valid bound for *any* n, we should always use it instead of a normal approximation, since the normal approximation requires n to be large." Identify one cost of following this advice.

Recall that approximately 95% of a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ lies within 1.96 standard deviations of its mean:

$$\Pr\left(|X - \mu| \leq 1.96\sigma\right) \approx 0.95.$$

Exercise 5 – A Confounded Comparison (Simpson's Paradox)

A hospital compares two treatments, A and B, for a condition, testing each on patients classified as having either a Mild or Severe case.

	Mild cases		Severe cases	
	Treated	Recovered	Treated	Recovered
Treatment A	180	171	20	6
Treatment B	20	19	180	90

(**Q1**) Compute the recovery rate for Treatment A and for Treatment B, separately within Mild cases and within Severe cases. Which treatment looks better in *each* subgroup?

(**Q2**) Now compute the *overall* recovery rate for each treatment, ignoring the Mild/Severe distinction. Which treatment now looks better?

(**Q3**) Explain this reversal using the vocabulary of confounding introduced in class. Identify the independent and dependent variables, as well as the confound.

(**Q4**) Using the idea of random assignment from class, explain precisely what change to this study's design would prevent this reversal from being possible.

Exercise 6 – A Monte Carlo Simulation for Hashing

A student wants to estimate the probability that inserting n random keys into a hash table with m buckets produces at least one collision (two keys hashed to the same bucket), by simulation:

```
table = set()

def estimate_collision_prob(n, m, trials):
    collision_count = 0
    for t in range(trials):
        collided = False
        for i in range(n):
            h = random_int(0, m - 1)
            if h in table:
                collided = True
            table.add(h)
        if collided:
            collision_count += 1
    return collision_count / trials
```

Run with $n = 20$, $m = 1000$, and `trials` = 10,000, this code returns an estimated collision probability noticeably *higher* than the true value. A classmate points out a bug.

(**Q1**) Identify the bug. Propose a corrected version.

(**Q2**) Is this a bias problem, a variance problem, both, or something the bias/variance framework doesn't capture at all? Justify.

(**Q3**) Now that the bug is fixed, the student claims: "Now that each trial is correct and independent, using more trials will always shrink my error at the same $1/\sqrt{n}$ rate we saw in class." Is there a hidden assumption in this claim that could still fail?