# "You Do It" Solution

Data Types, Selection Bias, Non-response Bias, Sampling Variability, and Response Bias

Prepared as a classroom-ready solution set for ECOM5005

## Exercise 1: Classifying Variables

**Question**

Identify whether each variable is **numerical** or **categorical**. If it is numerical, state whether it is **discrete** or **continuous**.

(a) Your age

(b) Income

**Detailed Solution**

**(a) Your age: numerical and continuous.**
Age is a quantitative variable because it measures an amount of time. In principle, a person's age can take any value on a continuum. For example, a person may be

$$20.0, \quad 20.5, \quad 20.75, \quad \text{or } 20.753 \text{ years old.}$$

Therefore, when age is measured precisely, it is classified as a **continuous numerical variable**.
**Important qualification.** If a survey records age only as the number of completed years—for example, 18, 19, 20, and so on—then the *recorded* variable is discrete. The underlying concept of age, however, remains continuous. Unless the question explicitly says "age in completed years," the standard answer is **continuous**.

**(b) Income: numerical and continuous.**
Income is quantitative because it is measured numerically in dollars, such as \$45,000 or \$72,350.50. It can vary over a very large range and, conceptually, may take any value within that range. It is therefore normally treated as a **continuous numerical variable**.
Although income is often recorded to the nearest dollar or cent, this is a limitation of measurement precision rather than a reason to classify the underlying variable as discrete.

**Answer Summary**

| **Variable** | Classification | Reason |
|---|---|---|
| **Age** | Numerical, continuous | Time can be measured to arbitrarily fine precision. |
| **Income** | Numerical, continuous | Monetary income can take a wide range of values and is conventionally modelled as continuous. |

# Exercise 2: Diagnosing Possible Biases

**Question**

A simple random sample of $n = 300$ full-time employees is selected from a company list containing all $N = 5{,}000$ full-time employees to evaluate job satisfaction.
Give an example of possible:

(a) selection bias;

(b) non-response bias;

(c) sampling bias;

(d) response bias.

## Preliminary Observation

Under a correctly implemented simple random sample, each employee has the same inclusion probability:

$$\Pr(\text{employee is selected}) = \frac{n}{N} = \frac{300}{5{,}000} = 0.06.$$

Thus, each employee has a 6% chance of selection. Because the sampling frame is stated to contain all full-time employees, the *design itself* is not subject to selection bias. Bias can nevertheless arise if the procedure is implemented incorrectly, if selected employees fail to respond, or if answers are distorted.

### (a) Possible Selection Bias

**Example: selected employees who are on leave, working remotely, or without regular access to the company's email system are not contacted and are replaced by easier-to-reach employees.**
This creates selection bias because the final group invited to participate no longer follows the original simple random selection. Employees who are physically present at head office or who regularly check company email become more likely to be included than other selected employees. If work location or leave status is related to job satisfaction, the resulting estimate will systematically differ from the true population value. For example, if remote employees are more satisfied but are underrepresented, the estimated average satisfaction will be biased downward.
**Key point:** with the complete list and a genuine random draw, selection bias should not occur. The example arises from a departure from the stated sampling design during implementation.

### (b) Possible Non-response Bias

**Example: dissatisfied employees are less likely to complete the survey because they distrust management or believe that participation will not lead to meaningful change.**
Suppose the 300 selected employees are invited, but only 180 respond. If highly dissatisfied employees respond at a lower rate than satisfied employees, the respondents will not be representative of the selected sample.
The calculated mean job-satisfaction score will then tend to be **too high**. The problem is not that dissatisfied employees had no chance of being selected; rather, they were selected but did not provide data.
The opposite direction is also possible. If dissatisfied employees are especially motivated to complain and therefore respond at a higher rate, the survey may understate average satisfaction. What matters is whether response probability is systematically related to job satisfaction.

### (c) Possible "Sampling Bias" / Sampling Variability

**Example: purely by chance, the random sample contains too many head-office professionals and too few warehouse or shift workers.**
Even when every employee has an equal chance of selection, one particular sample may differ from the population. If head-office employees have systematically higher job satisfaction than warehouse employees, this chance imbalance may produce a sample mean that is higher than the true population mean.
However, the statistically precise term for this phenomenon is **sampling variability** or **sampling error**, not bias. Random samples naturally differ from one another. Unlike selection bias, the error is not systematically in the same direction over repeated samples.
A larger sample, stratification by department or employment location, and appropriate weighting can reduce this variability.

### (d) Possible Response Bias

**Example: employees overstate their satisfaction because they fear that management may identify their responses.**
Even if the survey promises confidentiality, employees may worry that critical answers could affect promotion, work allocation, or job security. They may therefore select more favourable responses than they genuinely believe. This would tend to bias the estimated satisfaction level upward.
Response bias may also be caused by leading wording. For example:

> "How satisfied are you with the company's excellent employee-support initiatives?"

The positive wording encourages favourable responses. A more neutral question would be:

> "Overall, how satisfied or dissatisfied are you with your job at this company?"

Using anonymous data collection, neutral wording, and an independent survey administrator can reduce response bias.

### Comparison of the Four Issues

| Issue | Source of the problem | Illustrative consequence |
|---|---|---|
| **Selection bias** | Some employees are systematically more likely to enter the realised sample. | Remote or leave-taking employees are underrepresented. |
| **Non-response bias** | Selected employees do not respond, and response is related to satisfaction. | Dissatisfied non-respondents make satisfaction appear too high. |
| **Sampling variability** | A valid random sample differs from the population by chance. | One department is accidentally overrepresented. |
| **Response bias** | Recorded answers differ systematically from true views. | Fear of identification causes inflated satisfaction scores. |

### Terminology Note

Some introductory materials use the phrase *sampling bias* to describe differences among samples of the same size. In formal statistics, this is more accurately called **sampling variability**. Bias is systematic; sampling variability is random and can occur even under a perfectly designed simple random sample.

### Concise Answer Key

1. **Age:** numerical and continuous; if recorded only in completed years, the recorded values are discrete.
2. **Income:** numerical and continuous.
3. **Selection bias:** selected employees who are difficult to contact are omitted or replaced by easier-to-reach employees.
4. **Non-response bias:** dissatisfied selected employees respond less often, causing satisfaction to be overstated.
5. **Sampling variability:** the random sample accidentally overrepresents a department with unusually high or low satisfaction.
6. **Response bias:** employees give overly positive answers because they fear identification or because questions are leading.