1. Survey Design and Experiments

1.1. Table of Contents

1.2. Survey Design

When we collect data for research - we should try to ensure that our data is **representative** and **random**.

> Consider a sample $X_1, \ldots, X_n$ from a random variable X which has probability or density function $f_X(x)$
>
> The sample is said to be **representative** of the population if:
>
> $$f_{X_i}(x) = f_X(x)$$

1.3. Random Samples

> A random sample of size n is a set of random variables
>
> $$X_1, X_2, \ldots, X_n$$
>
> with the following properties:
>
> 1. the X_i's each have the same probability distribution, $f_{X_i}(x) = f_X(x)$ for all $i = 1, \ldots, n$.
> 2. the X_i's are independent
>
> We often say that the X_i are **iid** (independently and identically distributed) when these two properties hold

> A **simple random sample** of size n is a set of subjects sampled in such a way that all possible samples of size n are equally likely.

> A simple random sample does *NOT* consist of iid random variables - they are identitically distributed, but they are dependent. However, the dependence is very weak when the population size N is large compared to n (e.g if $N > 100n$).

To obtain a random sample in `RStudio`:

- Obtain a list of all subjects in the population, and assign each subject a number from 1 to N
- Use `sample(N, n)` to take a simple random sample of size n.

```
N = 100
n = 10
sample(N, n, replace = F)
```

1.3.1. Statistics calculated from samples

> Importantly, statistics of samples are **random variables** - this means we can apply statistics on them; for example $E(\bar{X})$.

- **Sample mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Sample variance:** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- **Sample median:** $\bar{X}_{0.5}$

> If $X_1, \ldots, X_n$ is a random sample from a variable of mean μ and variance σ^2, then the sample mean X satisfies:
>
> $$E(\bar{X}) = \mu \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

1.3.1.1. Sample proportions

If given a sample proportion $\hat{p}$ of a random variable X, we have that:

- $\hat{p} = \frac{X}{n}$
- $\mathbb{E}(\hat{p}) = p$
- $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$

Commonly, we fix $p = 0.5$, this lets us find the "maximum" variance; which lets us find the minimum sample size required for some variance σ^2. It also means we make zero assumption.

1.4. Methods of survey sampling

We must consider **efficiency** and **effort** for a sampling scheme. We want to get a good estimate of our population with respect to how much effort it takes to get the sample.

> Consider two unbiased alternative statistics, denoted as $g(X_1, \ldots X_n)$ and $h(Y_1, \ldots, Y_m)$
>
> We say that g is more efficient than h if:
>
> $$\mathrm{Var}(g) \leq \mathrm{Var}(h)$$

To improve our research regarding the samples, we can either:

1. Use a different statistic (which will be discussed in later chapters)
2. Sample differently

Here are three common ways of sampling:

> **Simple random sample:** Weaknesses are that it is difficult to implement in practice, requires high effort and can be inefficient
>
> **Stratified random sample:** If the populations can be broken into subpopulations, or **strata**, which differ from each other in the variable of interest, it is more efficient to sample separately in each subpopulation.
>
> An example of subpopulations are average taxable income, age, post code etc.
>
> **Cluster sampling:** This is useful when subjects in the population arise in clusters, and it takes less effort to sample within clusters than across clusters. Effort-per-subject can be reduced by sampling clusters than measuring all subjects within a cluster.
>
> An example of this is to interview 100 NSW household owners; its easier to sample ten households in ten postcodes rather then to randomly find 100 household owners in NSW.

1.5. Methods of experimental design

> **Randomised comparative experiment:** Define k treatment groups (each with different levels of the variable X) and randomly assign subjects to each group
>
> **Randomised blocks design:** If there is some "blocking" variable known to be important to the response variable, break subjects into blocks according to this variable and randomise allocation of subjects to treatment groups separately within each block. This controls for the effects of the blocking variable (for e.g sex, gender, height, etc.)
>
> **Matched pairs design:** A common sepcial case of a randomised blocks design, where the blocks come in pairs. Common examples are "before-after" experiments

Example: We want to consider the effects of vitamins on illness.

Experiment A randomly assigns subjects to a placebo and non-placbeo group. This experiment lasts 3 months. Number of illnesses are recorded over the study period.

Experiment B randomly assigns 3 months blocks of tablets (randomly chosen as placebo or non-placebo); which they taken for two blocks (so 6 months total.) Number of illnesses are recorded and compared over the two periods.

We are interested in the mean diffence in number of illnesses between takers of vitamin takers and takers of a placebo, estimated using the sample mean difference $\bar{Y}_v - \bar{Y}_p$.

Assume $\mathrm{Var}(Y_v) = \mathrm{Var}(V_p) = \sigma^2$

1. What type of experiment is done in experiment A and B?

A is a randomised comparative experiment, whereas B is a matched pairs design - where the first block is "before" and the second block is "after".

2. Find $\mathrm{Var}(\bar{Y}_v - \bar{Y}_p)$ for experiment A.

We can assume that the two random variables are random.

$$\begin{aligned}\mathrm{Var}(Y_v - Y_p) &= \mathrm{Var}(Y_v) + \mathrm{Var}(Y_p) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}\end{aligned}$$

3. Assuming the correlation across the two study periods is 0.5, find $\mathrm{Var}(\bar{Y}_v - \bar{Y}_p)$ for experiment B (Note that this is no longer independent between Y_v and Y_p)

$$\begin{aligned}\frac{\mathrm{Cov}(\bar{Y}_v, \bar{Y}_p)}{\ldots} &= 0.5 \\ \mathrm{Cov} &= 0.5 \cdot \mathrm{sd}(\bar{Y}_v) \cdot \mathrm{sd}(\bar{Y}_p) \\ &= \frac{0.5\sigma^2}{2n}\end{aligned}$$

Further noting that $\mathrm{Var}(\bar{Y}_p) = \frac{\sigma^2}{2n}$

$$\begin{aligned}\mathrm{Var}(\bar{Y}_v - \bar{Y}_p) &= \mathrm{Var}(\bar{Y}_v) + \mathrm{Var}(\bar{Y}_p) - 2\mathrm{Cov}(\ldots) \\ &= \frac{\sigma^2}{2n} + \frac{\sigma^2}{2n} - \frac{\sigma^2}{2n} \\ &= \frac{\sigma^2}{2n}\end{aligned}$$

Therefore, Experiment A is less efficient then Experiment B; and Experiment B is a better experiment (functionally). Do note that the experiment takes twice as longer.