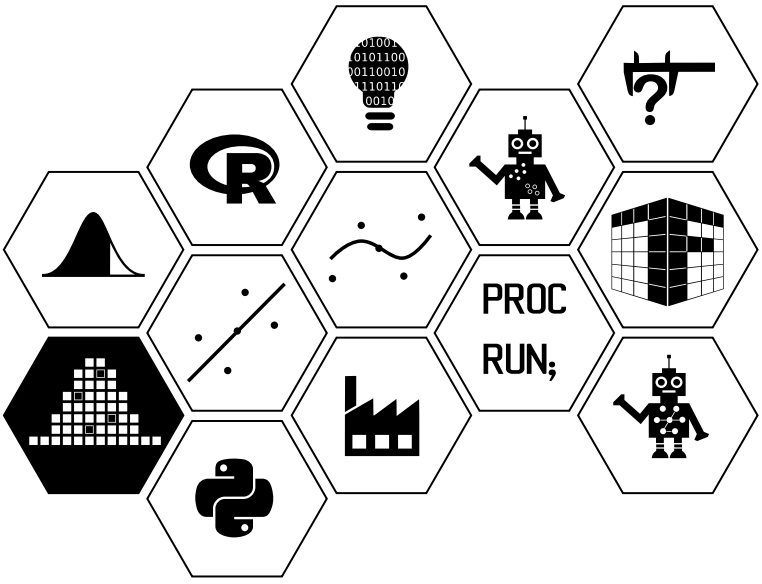
Probability and Sampling Fundamentals

Week 10: Probability Sampling I



DATA ANALYTICS
GLASGOW

Random sampling

In Week 9 we learned about different methods of sampling, with a particular focus on methods of **non-probability sampling**. We also learned that these sampling methods can introduce systematic errors and sampling bias into our results. This week we will introduce some **probability sampling** approaches which allows each unit in the population to have a (known) nonzero probability of being randomly selected.

Week 10 learning material aims

The material in week 10 covers:

- sampling from a finite population
- random number generating;
- general sampling theory;
- sample statistics for SRS without replacement, including the mean, total and proportion;
- confidence intervals for population characteristics for SRS without replacement;
- sample size calculation for SRS without replacement for the mean and proportion.

Some definitions

As discussed in Week 9 one of the most important areas of application for probability is statistical inference. In statistics we generally want to draw conclusions about populations based on information collected from samples and a key link between probability and statistics is made by sampling theory. In particular, in survey research, and in a number of other applications, the statistical objective is generally to estimate the parameters of a **finite** population. Let's remind ourselves of some population notation defined last week.

- y - variable of interest;
- N - number of elements in the population (i.e. population size);
- $y_1, y_2, \ldots, y_N$ - values of the variable in the population;
- y_i - value of the variable in the population;
- θ - population characteristic i.e. a function of $y_1, y_2, \ldots, y_N$ written as $f(y_1, y_2, \ldots, y_N)$;
- $\bar{y}$ - the mean of the values in the population.

Before we begin, let's introduce some new definitions that we will refer to throughout the notes.

- A **population** is a set $U = \{u_1, u_2 \dots u_N\}$ composed by N elementary statistical units.
- A **sample** is a subset $\boldsymbol{s}$ of $n < N$ elements of the population U.
- The **sampling fraction** $f = \frac{n}{N} \in (0, 1)$ indicates the proportion of units of the population that are to be sampled. This quantity is typically small.
- The **sample space** Ω is the set of all the samples of n elements that can be formed from a population on N elements.
- A **sampling design** is a probability distribution p of $\boldsymbol{s}$ over Ω, so that $p(\boldsymbol{s}) \in (0, 1)$ is the probability of extracting sample $\boldsymbol{s}$, and $\sum_{\boldsymbol{s} \in \Omega} p(\boldsymbol{s}) = 1$.

These will all be discussed in more detail when we introduce specific sampling methods.

Simple random sampling without replacement

Since populations are often large and costly to investigate, it is more common to conduct research on a sample. The methods of non-probability sampling discussed in Week 9 would introduce sampling bias, where consciously or subconsciously we sample units from the population that are unrepresentative of the population as a whole. Instead, it is usual for statisticians to recommend that some random process be used. The most straightforward method of this kind is **simple random sampling** (SRS), where every group of n different individuals in the population has the same probability of being included in the sample. Or in other words, every possible subset of n units in the population has the same chance of forming the sample. What this means is that, on average, the samples will be representative of the population and therefore can be used to draw reliable conclusions about the population.

Let's assume that we draw our sample **without replacement**. This means that, once a unit from the population has been chosen for the sample, it cannot be chosen again. In this case, all n sample values must come from different members of the population. Here, the number of possible samples is $\binom{N}{n}$ (look back to Week 1 for a reminder on this notation). Each possible sample therefore has probability $1/\binom{N}{n}$ of being selected.

Random number generation

How can we draw a random sample?

- We give each element in the population a unique number from 1 to N. This is called the **sampling frame**.
- We draw n random numbers out of $1, 2, \dots, N$ (without replicates).
- Each drawn number corresponds to exactly one element in the population, which is included in the sample.

A common method is to use computer software with well performing random number generators. For example using the `sample()` function in `R`.

Example 1

Draw a SRS of sample $n = 3$ from a population of $N = 5$.

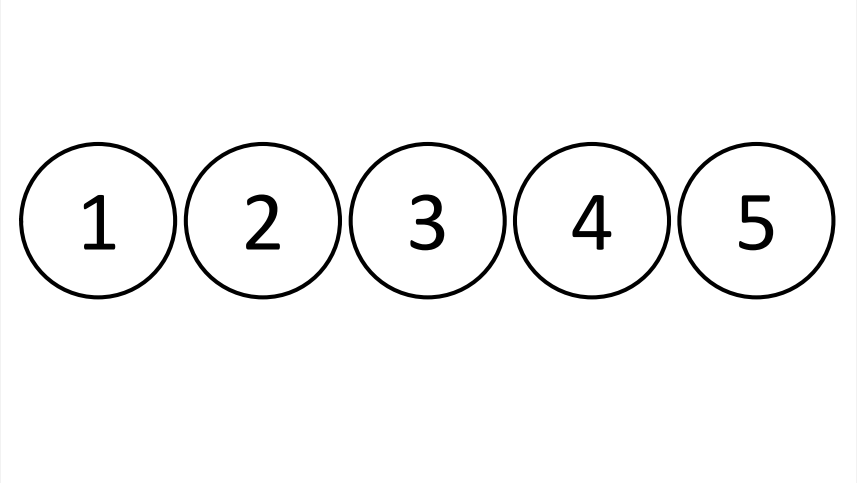


Figure 1

```
sample(1:5, size=3, replace=FALSE)
```

```
R Console
[1] 2 3 1
```

The 10 possible unique samples that we could get in this example are shown below.

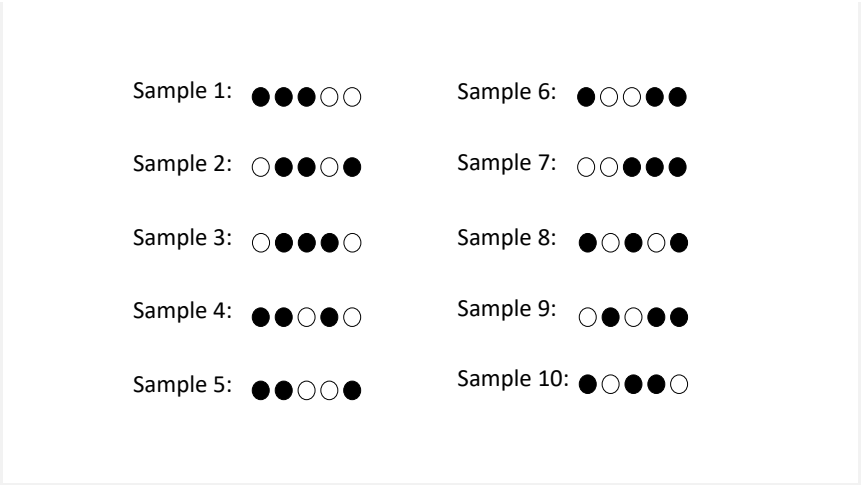


Figure 2

Mathematically, we can write each sample $\boldsymbol{s}$ as follows.

$$\text{Sample 1: } \boldsymbol{s} = [1, 2, 3] \qquad \text{Sample 6: } \boldsymbol{s} = [1, 4, 5]$$

$$\text{Sample 2: } \boldsymbol{s} = [2, 3, 5] \qquad \text{Sample 7: } \boldsymbol{s} = [3, 4, 5]$$

$$\text{Sample 3:} \boldsymbol{s} = [2, 3, 4] \qquad \text{Sample 8: } \boldsymbol{s} = [1, 3, 5]$$

$$\text{Sample 4: } \boldsymbol{s} = [1, 2, 4] \qquad \text{Sample 9: } \boldsymbol{s} = [2, 4, 5]$$

$$\text{Sample 5: } \boldsymbol{s} = [1, 2, 5] \qquad \text{Sample 10: } \boldsymbol{s} = [1, 3, 4]$$

Sampling design

As defined above, under a particular sampling scheme we can write the probability of extracting a particular sample, $\boldsymbol{s}$, as $p(\boldsymbol{s})$. This is referred to as the **sampling design**.

Definition 1

Sampling design (SRS without replacement)

In SRS without replacement we have $\binom{N}{n}$ possible samples we can choose so the **sampling design** is

$$p(\boldsymbol{s}) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } \boldsymbol{s} \text{ has } n \text{ elements} \\ 0, & \text{otherwise,} \end{cases}$$

Example 2

In Example 1 we want to draw a sample $n = 3$ from a population of $N = 5$. The probability of choosing a particular random sample $\boldsymbol{s}$ in this example is

$$p(\boldsymbol{s}) = \binom{N}{n}^{-1} = \binom{5}{3}^{-1} = 10^{-1} = \frac{1}{10}.$$

Really all this is calculating is one over the total number of unique samples that could be drawn. So in this example we have 10 unique ways in which the sample can be drawn, and therefore the probability of each individual sample is just $\frac{1}{10}$.

Indicator variable

To specify which elements are included in the random sample $\boldsymbol{s}$ we can use an **indicator variable**. Here, the inclusion of a given element, i, in a sample is a random event indicated by the random variable $\boldsymbol{I}_i$.

Definition 2

Indicator variable (SRS without replacement)

In the case of simple random sampling without replacement, the **indicator variable** is defined as

$$\boldsymbol{I}_i = \begin{cases} 1, & \text{if } i \in \boldsymbol{s} \\ 0, & \text{if not.} \end{cases}$$

Example 3

In Example 1 find the indicator variable for sample 1.

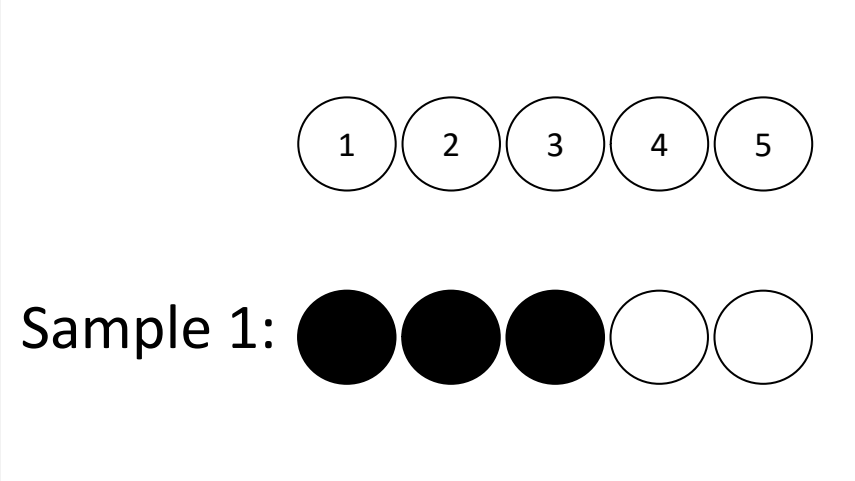


Figure 3

So sample 1 is equal to

$$\boldsymbol{s} = [1, 2, 3],$$

and

$$I_1 = 1, \quad I_2 = 1, \quad I_3 = 1, \quad I_4 = 0, \quad I_5 = 0.$$

So the indicator variable is equal to

$$I_i = (1, 1, 1, 0, 0).$$

Inclusion probability

We can also define the probability that a particular element, or pair of elements are included in the sample. These probabilities are referred to as the **first-order inclusion probability** and the **second-order inclusion probability** and are defined below.

Definition 3

First-order inclusion probability

For a given sampling design p and sample space Ω, we define the **first-order inclusion probability** of element i as

$$\pi_i = \sum_{\boldsymbol{c} \in \Omega_i} p(\boldsymbol{c}).$$

where $\Omega_i \subseteq \Omega$ is the set of the samples that contain element i. $\pi_i \in [0, 1]$ is the probability that element i enters the sample.

Definition 4

Second-order inclusion probability

Similarly, we define the **second-order inclusion probability** for elements i and j as

$$\pi_{ij} = \sum_{\boldsymbol{c} \in \Omega_{ij}} p(\boldsymbol{c}).$$

where $\Omega_{ij} \subseteq \Omega$ is the set of the samples that contain both i and j. $\pi_{ij} \in [0, 1]$ is the probability that pair (i, j) enters the sample.

Let's calculate the first and second order probabilities for simple random sampling.

For SRS without replacement, the total number of samples is

$$\binom{N}{n},$$

and the number of samples that will contain the element i is

$$\binom{N-1}{n-1}.$$

The number of samples containing both elements i and k is

$$\binom{N-2}{n-2}.$$

Therefore, the first order inclusion probability is

$$\begin{aligned}
\pi_i &= \sum_{\boldsymbol{c} \in \Omega_i} p(\boldsymbol{c}) \\
&= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \\
&= \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} \\
&= \frac{\frac{(N-1)(N-2)\dots}{(n-1)(n-2)\dots}}{\frac{N(N-1)(N-2)\dots}{n(n-1)(n-2)\dots}} \\
&= \frac{n}{N}
\end{aligned}$$

And, the second order inclusion probability is

$$\begin{aligned}
\pi_{ij} &= \sum_{\boldsymbol{c} \in \Omega_{ij}} p(\boldsymbol{c}) \\
&= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} \\
&= \frac{\frac{(N-2)(N-3)\dots}{(n-2)(n-3)\dots}}{\frac{N(N-1)(N-2)(N-3)\dots}{n(n-1)(n-2)(n-3)\dots}} \\
&= \frac{n(n-1)}{N(N-1)}
\end{aligned}$$

Since we are drawing our sample without replacement $\pi_{ii} = 0$.

Example 4

In Example 1 where $N = 5$ and $n = 3$, we have

$$\pi_i = \frac{3}{5} \quad \text{and} \quad \pi_{ij} = \frac{3}{10}.$$

We can easily check that these are correct for this example by firstly counting up the number of times a particular element (say 1) is included in the sample and dividing this by the total number of samples. Then counting up the number of times a pair of elements (say 1 and 2) are included in the sample and dividing this by the total number of samples.

Looking at the figure below we can count up the samples that include 1. We can see that Samples 1, 4, 5, 6, 8 and 10 all contain 1 and so

$$\pi_1 = \frac{6}{10} = \frac{3}{5}.$$

We can also count up the samples that contain 1 and 2, which gives us Samples 1, 4 and 5. So

$$\pi_{12} = \frac{3}{10}.$$

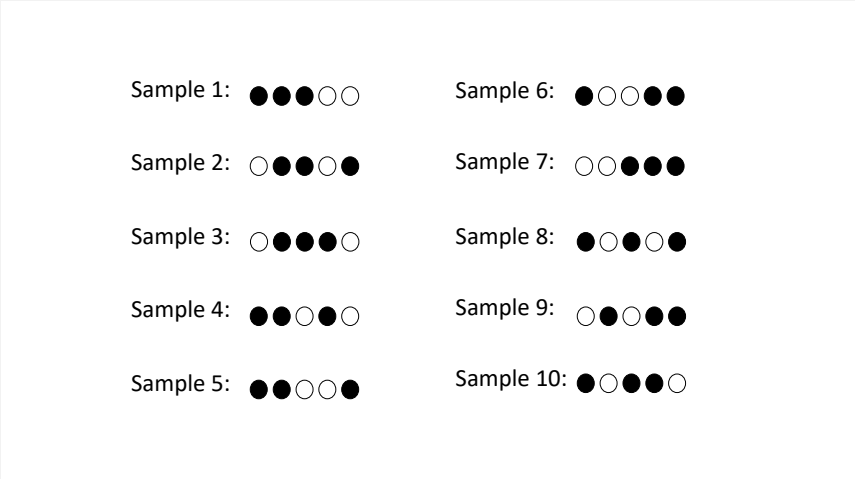


Figure 4

Task 1

Suppose we draw a SRS without replacement of sample $n = 5$ from a population of $N = 10$.

```
sample(1:10, size=5, replace=FALSE)
```

R Console

```
[1] 9 4 7 1 2
```

1. Calculate the sampling design.
2. Find the indicator variable for the sample above.
3. Calculate the first and second order inclusion probability.

Sampling distribution

Recall that we are interested in inferring some population characteristic which is a function of the values of the variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ across the population $\theta = f(y_1, y_2, \ldots, y_N)$. The notation used to refer to the estimate of θ based on the random sample $\boldsymbol{s}$ is

$$\hat{\theta} = \hat{\theta}(\boldsymbol{s}).$$

Definition 5

Sampling distribution

The **sampling distribution** is the distribution of the statistic across an infinite number of samples. Let Ω be the set of all possible samples, then we can define the expectation and variance of $\hat{\theta}_{(\boldsymbol{s})}$ for an unbiased estimate:

$$\mathrm{E}(\hat{\theta}) = \sum_{\boldsymbol{s} \in \Omega} p(\boldsymbol{s}) \hat{\theta}(\boldsymbol{s}),$$

$$\mathrm{Var}(\hat{\theta}) = \sum_{\boldsymbol{s} \in \Omega} p(\boldsymbol{s}) \{\hat{\theta}(\boldsymbol{s}) - \theta\}^2 = \mathrm{E}\left([\hat{\theta}(\boldsymbol{s}) - \theta]^2 \right).$$

Bias and Precision

A good estimator should possess the two qualities of being **unbiased** and **precise**. In Week 9 we defined bias as

$$bias(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta,$$

so an estimator is unbiased is $\mathrm{E}(\hat{\theta}) = \theta$ and so $bias(\hat{\theta}) = 0$.

For an unbiased estimator, the smaller the variance the more **precise/accurate** the estimator is said to be. Here precision is defined as

$$\mathrm{Var}(\hat{\theta})^{-1}.$$

For a biased estimator, we define the mean squared error as

$$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [bias(\hat{\theta})]^2.$$

We expect the precision of an estimator to improve as the sample size increases.

In general, the main aim of sampling theory is to produce sampling schemes which produce **minimum variance unbiased estimators** (MVUE).

You will learn much more about bias and precision in your course Learning from Data next semester.

Population characteristics

In sampling we are generally interested in certain characteristics of the populations such as the population mean:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

the population total:

$$Y = \sum_{i=1}^{N} y_i = N\bar{y},$$

and the population variance:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2.$$

Once we take a sample from our population, we generally have three questions.

1. What is the best estimate of the population mean/total?
2. How confident are we about that estimate?
3. What is our best estimate of the population variance?

We can answer these questions by producing unbiased estimates of these characteristics using our sample.

Note: for the remainder of this week we will usually calculate σ_y^2 using the **corrected** population variance:

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

since it is unlikely we will ever access to the entire population. [1]

Proposition 1

Sample mean

For SRS without replacement the sample mean

$$\hat{\bar{y}} = \frac{1}{n} \sum_{i \in \boldsymbol{s}} y_i$$

is an unbiased estimator of the population mean $\bar{y}$.

What this means is that the sample mean is our best estimate of the population mean, which answers question 1 above.

However, what we have to remember here is that it is not the y_i values that are random, what is random is the sample that is drawn. The y_i values can be thought of as a fixed quantity, e.g. a person's height, which will not change when we take our sample. The randomness is introduced in the sample that is taken, e.g. the people we select and record the heights of to make our sample. This means that if we were to draw a second sample and calculate the sample mean again, our estimate of the population mean would change - i.e. there is variability in this estimate. This eludes to question 2 above - how confident are we about our estimate?

In order to quantify this, we can find an expression for the **variance of the sample mean**.

Proposition 2

Variance of sample mean

The variance of the sample mean is given by

$$\mathrm{Var}(\hat{\bar{y}}) = \left(\frac{1-f}{n} \right) \sigma_y^2,$$

where f is the sampling fraction, $\frac{n}{N}$.

This value quantifies how confident we are about our estimate of the population mean. Intuitively, as the sampling fraction, $f = \frac{n}{N}$ increases towards 1 (or in other words, as our sample size n approaches

the population size N), the variance of the sample mean will decrease. This makes sense since the closer the sample size is to the population size, the more information we have about the population.

This also requires that the population variance is known and in practice this will not be the case. As such we may estimate it by

$$\hat{\mathrm{Var}}(\hat{\bar{y}}) = \left(\frac{1-f}{n}\right)\hat{\sigma}_y^2, \quad \text{where } \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\bar{y}})^2.$$

This quantity gives us an answer to question 3 above.

Note: We now use the corrected sample variance (we divide by (n-1) when calculating $\hat{\sigma}_y^2$) since we are estimating the population mean with the sample mean.

This video introduces simple random sampling from a finite population.

Video

SRS from a finite population

Duration 9:38

Similarly, we can find an expression for the sample total.

Proposition 3

Estimate of population total

An unbiased estimator of the population total Y is given by

$$\hat{Y} = \frac{N}{n}\sum_{i\in s} y_i,$$

with corresponding variance

$$\mathrm{Var}(\hat{Y}) = N^2\left(\frac{1-f}{n}\right)\sigma_y^2.$$

As before, the population variance is usually unknown and is estimated by

$$\hat{\mathrm{Var}}(\hat{Y}) = N^2 \left(\frac{1-f}{n}\right) \hat{\sigma}_y^2, \quad \text{where } \hat{\sigma}_y^2 = \frac{1}{n-1} \sum_{i \in \boldsymbol{s}} (y_i - \hat{\bar{y}})^2.$$

Note: the sample variance can be rearranged to $\hat{\sigma}_y^2 = \frac{1}{n-1}\left(\sum_{i \in \boldsymbol{s}} y_i^2 - n\hat{\bar{y}}^2\right)$ which may be easier for calculations.

Task 2 (Optional)

Show that

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{\bar{y}})^2 = \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2 \right),$$

and hence

$$\hat{\sigma}_y^2 = \frac{1}{n-1} \left(\sum_{i \in \boldsymbol{s}} y_i^2 - n\hat{\bar{y}}^2 \right).$$

Confidence in the sample

Using the central limit theorem (Week 8) we can assume that our sampling distribution is approximately normally distributed with a sufficiently large n. Consequently, we may produce confidence intervals (CIs) for our sample statistics.

Remember from Week 6 that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$\mathrm{P}(X < x) = \mathrm{P}\left(Z < \frac{x-\mu}{\sigma}\right) = \alpha.$$

Then by the central limit theorem we have

$$\hat{\bar{y}} \sim \mathrm{N}(\mu = \bar{y}, \sigma^2 = \mathrm{Var}(\bar{y})) = \mathrm{N}\left(\bar{y}, \left(\frac{1-f}{n}\right)\sigma_y^2\right).$$

Then using the standard normal transformation

$$\mathrm{P}\left(Z < \frac{\bar{y} - \hat{\bar{y}}}{\sigma_y \sqrt{(1-f)/n}}\right) = \alpha.$$

Proposition 4

Confidence Interval for mean

A $100(1-\alpha)\%$ confidence interval (CI) for the mean is given by

$$100(1-\alpha)\%\text{CI}_{\hat{\bar{y}}} = \left(\hat{\bar{y}} \pm z_{1-\frac{\alpha}{2}}\sigma_y\sqrt{(1-f)/n}\right)$$

Or, when the population variance is not known,

$$100(1-\alpha)\%\text{CI}_{\hat{\bar{y}}} = \left(\hat{\bar{y}} \pm z_{1-\frac{\alpha}{2}}\hat{\sigma}_y\sqrt{(1-f)/n}\right)$$

Similarly for the total we find

Proposition 5

Confidence Interval for total

$$100(1-\alpha)\%\text{CI}_{\hat{Y}} = \left(\hat{Y} \pm z_{1-\frac{\alpha}{2}}N\sigma_y\sqrt{(1-f)/n}\right)$$

Or, when the population variance is not known,

$$100(1-\alpha)\%\text{CI}_{\hat{Y}} = \left(\hat{Y} \pm z_{1-\frac{\alpha}{2}}N\hat{\sigma}_y\sqrt{(1-f)/n}\right)$$

In these intervals, $z_{1-\frac{\alpha}{2}}$ must be looked up in the standard normal tables which were discussed in Week 6, or they can be calculated in `R` using `qnorm()`. One choice that must be made first is the confidence level. For example if we are interested in a 95% CI, then $\alpha = 5\% = 0.05$, therefore

$$1 - \frac{0.05}{2} = 0.975, \quad z_{0.975} = 1.96.$$

```
qnorm(0.975)
```

```
R Console
[1] 1.959964
```

Or for a 99% CI, $\alpha = 1\% = 0.01$ so

$$1 - \frac{0.01}{2} = 0.995, \quad z_{0.995} = 2.58.$$

```
qnorm(0.995)
```

```
R Console
[1] 2.575829
```

In order for our confidence intervals to be valid, we require the sample size to be large (i.e. n>40). For sample sizes less than 40, a better allowance for the unknown population variance σ_y^2 involves forming our CI using students t-distribution, which we met in Week 6.

$$\text{Mean:} \quad 100(1-\alpha)\%\text{CI}_{\hat{\bar{y}}} = \left(\hat{\bar{y}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sigma_y\sqrt{(1-f)/n}\right)$$

$$\text{Total:} \quad 100(1-\alpha)\%\text{CI}_{\hat{Y}} = \left(\hat{Y} \pm t_{1-\frac{\alpha}{2}}(n-1)N\sigma_y\sqrt{(1-f)/n}\right)$$

Again, in these intervals, $t_{1-\frac{\alpha}{2}}(n-1)$ must be looked up in the standard **t-distribution tables**, similar to the normal table discussed in week 6. Or they can be calculated in `R` using `qt()`. E.g. suppose $\alpha = 0.05$ and we have a sample size of 30.

$$1 - \frac{0.05}{2} = 0.975, \quad t_{0.975}(29) = 2.0452.$$

```
qt(0.975,29)
```

```
R Console
[1] 2.04523
```

Example 5

Heights

Suppose there is a population of 25 statistics students in a class and we are interested in finding the average height of these students. Suppose we randomly sample 5 students and we get the following data

$$\text{Sample: } 168\text{cm}, 191\text{cm}, 188\text{cm}, 175\text{cm}, 165\text{cm}$$

1. Compute the value of the unbiased estimator of the mean.
2. Give a 95% confidence interval for the mean.

Answer: Mean:

$$\begin{aligned}\hat{\bar{y}} &= \frac{1}{n}\sum_{i\in \boldsymbol{s}} y_i \\ &= \frac{1}{5}\cdot 887 \\ &= 177.4\end{aligned}$$

Sample variance:

$$\begin{aligned}\hat{\sigma}_y^2 &= \frac{1}{n-1}\left(\sum_{i\in \boldsymbol{s}} {y_i}^2 - n\hat{\bar{y}}^2\right) \\ &= \frac{1}{4}(157899 - 5(177.4)^2) \\ &= 136.3\end{aligned}$$

95% CI:

$$\begin{aligned}\hat{\bar{y}} \pm t_{1-\frac{\alpha}{2}}(n-1)\hat{\sigma}_y\sqrt{(1-f)/n} \\ \text{i.e. } \hat{\bar{y}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{(1-f)}{n}\cdot\hat{\sigma}_y^2} \\ \text{i.e. } 177.4 \pm 2.7764\sqrt{\frac{(1-\frac{1}{5})}{5}\cdot 136.3} \\ \text{i.e. } 177.4 \pm 12.9655 \\ \text{i.e. } (164.43, 190.37)\end{aligned}$$

Therefore it is highly likely that the average height of students in this statistics class lies between 164.43 and 190.37.

N.b. here our sample size is very small (mainly for illustrative purposes) so we have used the t-distribution to form our confidence interval as discussed above.

Task 3

Look back at Example 5, suppose now we randomly sample 15 students with

$$\sum_{i \in \boldsymbol{s}} Y_i = 2594\text{cm} \quad \text{and} \quad \sum_{i \in \boldsymbol{s}} Y_i^2 = 449700\text{cm}^2.$$

1. Compute the value of the unbiased estimator of the mean.
2. Give a 95% confidence interval for the population mean.
3. Compare this to the CI for the sample of 5 students. What do you notice?

Choice of sample size for the sample mean

It is clear from Example 5 and the following task that the sample size affects the width of the CI for the population parameter. An important question in sampling is: how large should the sample size be, for our estimates to be reliable? Furthermore: what does *reliable* even mean?

Consider an unbiased estimator $\hat{\theta}$ for θ. We can choose the minimum sample size based on the desired *precision* (i.e. how precise we want our estimate to be) which ensures:

$$\mathrm{P}(|\hat{\theta} - \theta| > d) \leq \alpha$$

where d is called the **margin of error** and is a function of $\mathrm{Var}(\hat{\theta})$.

Assuming a normal approximation to the sample distribution of the mean, we can standardise the above expression and obtain

$$\frac{d}{\sigma_y \sqrt{\frac{1-f}{n}}} \geq z_{1-\alpha/2}.$$

If we specify d along with α and rearrange the formula above then we can calculate the minimum required sample size.

Supplement 1

Here, as supplementary material, is an outline derivation of the expression above.

$$
\begin{aligned}
\mathrm{P}(|\hat{\theta}-\theta|>d) = \mathrm{P}(\hat{\theta}-\theta<-d) + \mathrm{P}(\hat{\theta}-\theta>d) &\leq \alpha \\
1-\mathrm{P}(\hat{\theta}-\theta<d) + 1 - \mathrm{P}(\hat{\theta}-\theta<d) &\leq \alpha \\
2-2\mathrm{P}(\hat{\theta}-\theta<d) &\leq \alpha \\
1-\mathrm{P}(\hat{\theta}-\theta<d) &\leq \alpha/2 \\
\mathrm{P}(\hat{\theta}-\theta<d) &\geq 1-\alpha/2 \\
\mathrm{P}\left(\frac{\hat{\theta}-\theta}{\sigma_y\sqrt{\frac{1-f}{n}}} < \frac{d}{\sigma_y\sqrt{\frac{1-f}{n}}}\right) &\geq 1-\alpha/2 \\
\frac{d}{\sigma_y\sqrt{\frac{1-f}{n}}} &\geq z_{1-\alpha/2}
\end{aligned}
$$

The second last line holds since $\hat{\theta}-\theta \sim \mathrm{N}\left(0, \sigma_y^2\left(\frac{1-f}{n}\right)\right)$ since $\hat{\bar{y}}$ is an unbiased estimator of the sample mean. Details of this are shown below.

For an unbiased estimator $\mathrm{E}(\hat{\theta}) = \theta$ and $\mathrm{E}(\hat{\theta}-\theta) = 0$.

Also $\mathrm{E}[(\hat{\theta}-\theta))] = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$,

since $\mathrm{bias}(\hat{\theta}) = 0$ for an unbiased estimator.

Also

$$
\begin{aligned}
\mathrm{Var}[(\hat{\theta}-\theta)] &= \mathrm{E}[(\hat{\theta}-\theta)^2] - \mathrm{E}[(\hat{\theta}-\theta)]^2 \\
&= \mathrm{Var}(\hat{\theta}) - 0^2 \\
&= \mathrm{Var}(\hat{\theta})
\end{aligned}
$$

and $\mathrm{Var}(\hat{\bar{y}}) = \sigma_y^2\left(\frac{1-f}{n}\right)$.

Proposition 6

Sample size

When choosing the sample size for the sample mean under SRS without replacement we require

$$n \geq \frac{N\sigma_y^2}{\sigma_y^2 + N(d/z_{1-\frac{\alpha}{2}})^2},$$

As previously mentioned, it is often necessary to approximate σ_y^2 with $\hat{\sigma}_y^2$.

The figure below shows the relationship between the sample size (n) and the confidence level ($1-\alpha$) for different levels of the margin of error (d) for a population of $N = 100$ and a variance of 16. Firstly it is clear that as d decreases, the required sample size increases. This is to be expected since smaller levels of d are associated with higher precision in the estimates. Secondly, as the confidence level ($1-\alpha$) increases the sample size also increases gradually, until $1-\alpha = 1$, where the sample size increases to the population size.

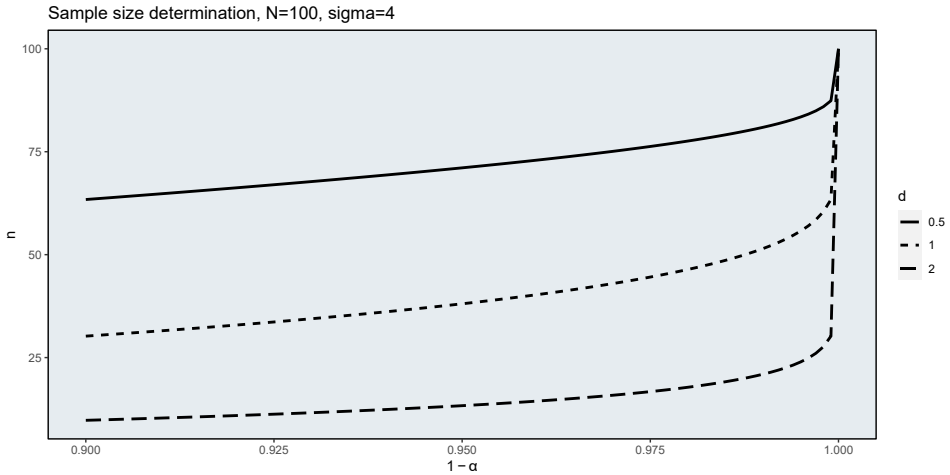


Figure 5

Example 6

Compute the minimum sample size required to produce an estimate of the mean height of the students within 5cm of the true height in the lecture example ensuring a 95% confidence.

Answer:

Using $\hat{\sigma}_y^2 = 136.3$ from Example 5 as an estimate of the population variance, we have

$$N = 25, \quad \hat{\sigma}_y^2 = 136.3, \quad d = 5, \quad \alpha = 0.05, \quad z = 1.96.$$

So,

$$\begin{aligned} n &\geq \frac{N\sigma_y^2}{\sigma_y^2 + N(d/z_{1-\frac{\alpha}{2}})^2} \\ &\geq \frac{25 \cdot 136.3}{136.3 + 25(\frac{5}{1.96})^2} \\ &\geq 11.396 \end{aligned}$$

So we would need a sample size of 12 to be within 5cm of the population mean.

Population proportion

Recall that we may label our population by $U = \{u_1, \ldots, u_N\}$, then we may be interested in a subset of U, i.e.

$$U_d \subset U, \quad \text{where } N_d = \text{size of } U_d.$$

The function of interest here is the population proportion

$$P_d = \frac{N_d}{N}.$$

For example, we may be interested in

- The proportion of the UK population who are going to vote for the Labour party in the next election.
- The proportion of cattle that are pregnant on a large farm.

It is often assumed that the population size (N) is known but N_d is unknown. In this case we can define the variable:

$$y_{di} = \begin{cases} 1, & \text{if } i \in U_d, \quad i = 1, \ldots, N \\ 0, & \text{otherwise,} \end{cases}$$

In the example of pregnant cattle we would define pregnant = 1 and not pregnant = 0.

The population total may be written as

$$\sum_{i=1}^{N} y_{di} = N_d,$$

and the population mean may be written as

$$\bar{y}_d = \frac{1}{N} \sum_{i=1}^{N} y_{di} = \frac{N_d}{N} = P_d.$$

We can also define the population variance as

$$\sigma_d^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_{di} - P_d)^2 = \frac{NP_d(1-P_d)}{N-1}.$$

Let $Q_d = 1 - P_d$ so

$$\sigma_d^2 = \frac{NP_dQ_d}{N-1}.$$

Note: we use the subscript d to indicate that we are interested in the population proportion.

Estimating a population proportion

As before we can use our sample to estimate a population proportion. In this case the sample total is

$$\hat{Y}_d = \sum_{i \in \boldsymbol{s}} y_{di} = n_d,$$

which can be used to calculate the sample mean as follows.

Proposition 7

The sample mean for a proportion is given by

$$\hat{\bar{y}}_d = \frac{1}{n}\sum_{i\in \boldsymbol{s}} y_{di} = \frac{n_d}{n} = \hat{P}_d.$$

This is therefore the sample proportion. The variance of the sample proportion is equal to

$$\begin{aligned}\mathrm{Var}(\hat{P}_d) &= \left(\frac{1-f}{n}\right)\sigma_d^2 \\ &= \frac{N-n}{N-1}\frac{P_d Q_d}{n}.\end{aligned}$$

Since the population variance is usually unknown we can estimate it by

$$\hat{\mathrm{Var}}(\hat{P}_d) = \left(\frac{1-f}{n}\right)\hat{\sigma}_d^2, \quad \text{where } \hat{\sigma}_d^2 = \frac{n\hat{P}_d\hat{Q}_d}{n-1}.$$

So the variance estimator of the proportion is now

$$\hat{\mathrm{Var}}(\hat{P}_d) = \frac{(1-f)\hat{P}_d\hat{Q}_d}{n-1}.$$

Confidence intervals for proportions

Consider the case of sampling from a population that has only two types of elements. Suppose there are N elements in the population,

1. M of them are of type 1,
2. $N-M$ of them are type 2.

Now suppose you draw a sample of n of these elements from the population **with replacement** (that is after selecting one of the elements you replace it and it could theoretically be sampled again). As you sample with replacement each time you randomly draw an element the probability of obtaining one of type 1 is $\mathrm{Bern}(\theta = M/N)$ and all draws are independent. Hence if X denotes the random variable representing the number of type 1 elements drawn from n samples with replacement, you have that $X \sim \mathrm{Bi}(n, \theta = M/N)$.

Now suppose you sample **without replacement** (that is once an element is sampled it is not returned to the population and cannot be sampled again) from the population, then the binomial model no longer holds. This is because the probability of the second element being of type 1 depends on whether the first element was type 1 or type 2.

- If the first element was type 1 then the probability of the second element being type 1 is $\theta = (M-1)/(N-1)$.

- If the first element was of type 2 then the probability changes to $\theta = M/(N-1)$.

Thus the trials do not have constant probability and are not independent.

Let X denote the number of type 1 objects chosen when n objects are sampled without replacement from a population containing M objects of type 1 and $N-M$ objects of type 2. Before we define the probability mass function for this situation, what is the sample space of X?

- With a sample size of n elements to choose and a total population of M elements of type 1, X cannot be greater than $\min\{n, M\}$.

- Also, X must be at least zero. In addition, if the number to be sampled n is greater than the number of type 2 elements $N-M$, then at least $n-(N-M)$ type 1 elements must be sampled as all the type 2 elements will already have been sampled. Thus we have that X cannot be less than $\max\{0, n-(N-M)\}$.

- Thus the sample space for X is the set of integers between

$$\mathcal{S} = \max\{0, n-(N-M)\}, \ldots, \min\{n, M\}.$$

Definition 6

A **hypergeometric** random variable is denoted by $X \sim \text{Hyp}(n, N, M)$ and has a probability mass function given by

$$p(x) = \text{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x = \max\{0, n-(N-M)\}, \ldots, \min\{n, M\}.$$

It can be shown that the estimator of the population total is distributed as a **hypergeometric distribution**. In this scenario, the estimator of the population total may be written as:

$$n_d \sim \text{Hyp}(n, N_d, N).$$

It is possible to construct confidence intervals for the proportion by making exact probability statements about n_d. However, this requires use of cumulative probabilities for the hypergeometric.

Instead the hypergeometric can be approximated by the binomial distribution when

$$n \ll N_d \quad \text{and} \quad n \ll N,$$

where $\ll$ means 'much smaller than'. In other words, when the sample being drawn (n) is much smaller than the population size of the variable we are interested in (N_d) and the total population size (N), then the probability of success for a binomial distribution, $\theta = N_d/N$ will not change much when we remove n samples and so, $n_d \sim \text{Hyp}(n, N_d, N)$ can be approximated by

$$n_d \sim \text{Bi}(n, N_d/N).$$

Task 4

Sweet experiment

Suppose you have a bag of N sweets in a bag. M are milk chocolate and $N - M$ are not. We draw 1 sweet from the bag noting if it's milk chocolate or not and then put the sweet back in the bag. We repeat this process n times.

What is the probability that m of the n sweets are milk chocolate?

Although it is possible to construct confidence intervals for the proportion by making exact probability statements about n_d, again this requires use of cumulative probabilities for the binomial. Instead, the binomial can be approximated by the normal distribution when

$$n \ll N_d \quad \text{and} \quad n \ll N \quad \text{and} \quad \min(nP_d, nQ_d) > 30.$$

We learned in Week 8 that $X \sim \text{Bin}(n, \theta)$ may be approximated by $X \sim \text{N}(n\theta, n\theta(1-\theta))$. However, to allow for the lack of replacement we use a different approximation here.

Definition 7

CI for proportion

The estimator of the population proportion may be approximated as

$$\mathrm{P}\left(\frac{|P_d - \hat{P}_d|}{\sqrt{\mathrm{Var}(\hat{P}_d)}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Replacing $\sqrt{\mathrm{Var}(\hat{P}_d)}$ with $\sqrt{\hat{\mathrm{Var}}(\hat{P}_d)}$ above gives

$$100(1-\alpha)\%\mathrm{CI}_{\hat{P}_d} = \hat{P}_d \pm z_{1-\frac{\alpha}{2}}\sqrt{(1-f)\hat{P}_d\hat{Q}_d/(n-1)}.$$

Example 7

In Example 5, suppose we also have collected the sex of the students in the class with 0-female and 1-male. Suppose $N = 25$, $n = 10$ and

$$\mathrm{Sample}: 1, 1, 1, 1, 0, 1, 0, 1, 1, 1.$$

1. Compute the value of the unbiased estimator of the proportion of male students.
2. Give a 95% CI for the population proportion*.

Answer:

1. We have

$$n_d = \sum_{i \in \boldsymbol{s}} y_i = 8, \quad \text{so } \hat{P}_d = \frac{n_d}{n} = \frac{8}{10} = 0.8$$

2. From part 1, we have $\hat{Q}_d = 1 - \hat{P}_d = 0.2$ so

$$\hat{P}_d \pm z_{1-\frac{\alpha}{2}}\sqrt{(1-f)\hat{P}_d\hat{Q}_d/(n-1)}$$

$$\text{i.e. } 0.8 \pm 1.96\sqrt{(1-\frac{10}{25})\cdot(0.8\cdot 0.2)/9}$$

$$\text{i.e. } 0.8 \pm 0.202$$

$$\text{i.e. } (0.598, 1.000)$$

Note: we round down to 1 because we have a proportion here.

*Although the sample size requirements above don't hold for this example, we form our CI using the normal distribution for illustrative purposes.

This video discusses estimating population proportions when sampling from a finite population and illustrates how to calculate a 95% confidence interval for a population proportion.

Video

Sampling from a finite population - population proportions

Duration 12:32

Choice of sample size for population proportion

Like before, the width of the confidence interval for the population proportion is affected by the sample size. We can choose the minimum sample size based on the desired precision which ensures:

$$\mathrm{P}(|\hat{P}_d - P_d| > d) \leq \alpha.$$

Assuming a normal approximation to the sample distribution of the proportion, we can standardise the above expression:

$$\mathrm{P}\left(\frac{|\hat{P}_d - P_d|}{\sqrt{\mathrm{Var}(\hat{P}_d)}} > \frac{d}{\sqrt{\mathrm{Var}(\hat{P}_d)}}\right) \leq \alpha.$$

Hence using the normal approximation, we require

$$\frac{d}{\sqrt{\mathrm{Var}(\hat{P}_d)}} \geq z_{1-\alpha/2}.$$

Equivalently we can express this in terms of the variance:

$$\mathrm{Var}(\hat{P}_d) = \frac{N-n}{N-1}\frac{P_d Q_d}{n} \leq \left(\frac{d}{z_{1-\alpha/2}}\right)^2 = D.$$

So we require:

$$n \geq \frac{P_d Q_d}{D}\left[1 + \frac{1}{N}\left(\frac{P_d Q_d}{D} - 1\right)\right]^{-1},$$

where $D = \left(\frac{d}{z_{1-\alpha/2}}\right)^2$.

It is important to note that if P_d is unknown, we choose the most conservative value, i.e. 0.5, in determining the sample size. For very large populations the required sample size is equal to

$$n \geq \frac{P_d Q_d}{D}.$$

Learning outcomes for week 10

By the end of week 10, you should be able to:

- calculate the sampling design, indicator variable and inclusion probability for SRS without replacement;
- compute estimates and calculate 95% CIs for the mean, total and population proportion for SRS without replacement;
- calculate the sample size for a given precision for SRS without replacement for the mean and population proportion.

A summary of the most important concepts and written answers to all tasks are provided overleaf.

Week 10 summary

RS without replacement

Sampling design

In SRS without replacement we have $\binom{N}{n}$ possible samples we can choose so the sampling design is

$$p(\boldsymbol{s}) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } \boldsymbol{s} \text{ has } n \text{ elements} \\ 0, & \text{otherwise,} \end{cases}$$

Indicator variable

In the case of simple random sampling without replacement, the indicator variable is defined as

$$\boldsymbol{I}_i = \begin{cases} 1, & \text{if } i \in \boldsymbol{s} \\ 0, & \text{if not,} \end{cases}$$

Inclusion probability

The probability that element i is included in the sample, π_i is referred to as the first-order inclusion probability, where

$$\pi_i = \sum_{\boldsymbol{c} \in \Omega_i} p(\boldsymbol{c}).$$

The probability that a pair of elements i and k are included in the sample, π_{ik} is referred to as the second-order inclusion probability, where

$$\pi_{ij} = \sum_{\boldsymbol{c} \in \Omega_{ij}} p(\boldsymbol{c}).$$

Estimates of population characteristics

$$\text{Sample mean:} \quad \hat{\bar{y}} = \frac{1}{n} \sum_{i \in \boldsymbol{s}} y_i$$

$$\text{Variance of the sample mean:} \quad \operatorname{Var}(\hat{\bar{y}}) = \left(\frac{1-f}{n}\right)\sigma_y^2,$$

$$\text{Estimated variance of the sample mean:} \quad \hat{\operatorname{Var}}(\hat{\bar{y}}) = \left(\frac{1-f}{n}\right)\hat{\sigma}_y^2, \quad \text{where } \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i\in\boldsymbol{s}}(y_i - \hat{\bar{y}})^2$$

$$\text{Estimated total:} \quad \hat{Y} = \frac{N}{n}\sum_{i\in\boldsymbol{s}} y_i,$$

$$\text{Variance of estimated total:} \quad \operatorname{Var}(\hat{Y}) = N^2\left(\frac{1-f}{n}\right)\sigma_y^2,$$

Confidence in the sample

A $100(1-\alpha)\%$ CI for the mean is given by

$$100(1-\alpha)\%\text{CI}_{\hat{\bar{y}}} = \left(\hat{\bar{y}} \pm z_{1-\frac{\alpha}{2}}\sigma_y\sqrt{(1-f)/n}\right)$$

A $100(1-\alpha)\%$ CI for the total is given by

$$100(1-\alpha)\%\text{CI}_{\hat{Y}} = \left(\hat{Y} \pm z_{1-\frac{\alpha}{2}}N\sigma_y\sqrt{(1-f)/n}\right)$$

Sample size for population mean

When choosing the sample size for the population mean for SRS without replacement we require

$$n \geq \frac{N\sigma_y^2}{\sigma_y^2 + N(d/z_{1-\frac{\alpha}{2}})^2}.$$

Population proportion

The sample mean for a proportion is given by

$$\hat{\bar{y}}_d = \frac{1}{n}\sum_{i\in\boldsymbol{s}} y_{di} = \frac{n_d}{n} = \hat{P}_d.$$

This is therefore the sample proportion. The variance of the sample proportion is equal to

$$\begin{aligned}\text{Var}(\hat{P}_d) &= \left(\frac{1-f}{n}\right)\sigma_d^2 \\ &= \frac{N-n}{N-1}\frac{P_dQ_d}{n}.\end{aligned}$$

A $100(1-\alpha)\%$ CI for the proportion is given by

$$100(1-\alpha)\%\text{CI}_{\hat{P}_d} = \hat{P}_d \pm z_{1-\frac{\alpha}{2}}\sqrt{(1-f)\hat{P}_d\hat{Q}_d/(n-1)}.$$

Sample size for population proportion

When choosing the sample size for the population proportion for SRS without replacement we require

$$n \geq \frac{P_dQ_d}{D}\left[1+\frac{1}{N}\left(\frac{P_dQ_d}{D}-1\right)\right]^{-1}.$$

Answer 1

1. $$p(s) = \binom{N}{n}^{-1} = \binom{10}{5}^{-1} = \frac{1}{252}.$$

2. $$I_i = \{1,1,0,1,0,0,1,0,1,0\}.$$

3. The first order inclusion probability is

$$\begin{aligned}\pi_i &= \frac{n}{N} \\ &= \frac{5}{10} \\ &= \frac{1}{2}.\end{aligned}$$

And, the second order inclusion probability is

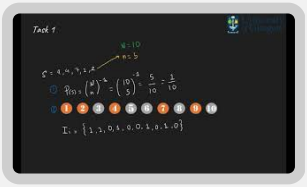
$$
\begin{aligned}
\pi_{ij} &= \frac{n(n-1)}{N(N-1)} \\
&= \frac{5 \cdot 4}{10 \cdot 9} \\
&= \frac{2}{9}
\end{aligned}
$$

Here is a video worked solution.

Video

Week 10 - Task 1

Duration 2:15



Answer 2

$$
\begin{aligned}
\sigma_y^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \\
&= \frac{1}{N-1} \sum_{i=1}^{N} (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - 2 \sum_{i=1}^{N} y_i\bar{y} + N\bar{y}^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - 2N\bar{y}_i\bar{y}_i + N\bar{y}^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2 \right)
\end{aligned}
$$

Hence the sample variance is equal to

$$
\hat{\sigma}_y^2 = \frac{1}{n-1} \left(\sum_{i \in \boldsymbol{s}} y_i^2 - n\hat{\bar{y}}^2 \right)
$$

Answer 3

1. Mean:

$$\begin{aligned}\hat{\bar{y}} &= \frac{1}{n}\sum_{i\in \boldsymbol{s}} y_i \\ &= \frac{1}{15}\cdot 2594 \\ &= 172.9\end{aligned}$$

2. Sample variance:

$$\begin{aligned}\hat{\sigma}_y^2 &= \frac{1}{n-1}\left(\sum_{i\in \boldsymbol{s}} {y_i}^2 - n\hat{\bar{y}}^2\right) \\ &= \frac{1}{14}(449700 - 15(172.9)^2) \\ &= 91.70\end{aligned}$$

95% CI:

$$\begin{aligned}\hat{\bar{y}} \pm t_{1-\frac{\alpha}{2}}(n-1)\hat{\sigma}_y\sqrt{(1-f)/n} \\ \text{i.e. } \hat{\bar{y}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{(1-f)/n\cdot\hat{\sigma}_y^2} \\ \text{i.e. } 172.9 \pm 2.145\sqrt{\frac{1-\frac{15}{25}}{15}\cdot 91.7} \\ \text{i.e. } 172.9 \pm 3.354 \\ \text{i.e. } (169.55, 176.25)\end{aligned}$$

3. This CI is narrower that the CI from the example with 5 students. Hence we can see that by increasing the sample size, the confidence interval tightens and we are more confident in our estimate.

Answer 4

Let $X = \{\text{the number of sweets drawn}\}$. Then

$$X \sim \text{Bi}(n, \theta),$$

where $\theta = \frac{M}{N}$.

Thus the estimator of the population proportion may be written as

$$n_d \sim \mathrm{Bi}(n, p_d).$$

Here is a video worked solution.

Video

Week 10 - Task 4

Duration 1:08

Footnotes

1. See here for more information about the corrected population variance. ↩