Chapter 3: Sampling

Dr. Chennaf Bouchra

Contents

1 Sampling

1.1 Notion of sampling

Definition 1.1. *Consider a population Ω of size N. A* sample *is a subset of this population. A sample of size n is therefore a list of n individuals $(\omega_1, \omega_2, \ldots, \omega_n)$ taken from the mother population.*

Example 1.1. ***Sampling wheat plots.***

A researcher has a small experimental farm with 5 wheat plots and wants to estimate average yield (quintals per hectare). The plot yields are:

$$45,\ 40,\ 47,\ 52,\ 44 \quad \textit{(q/ha)}.$$

She will draw samples of size n (plots) from the population of plots to estimate the mean yield.

```r
# Wheat plot yields (q/ha) - population of 5 plots
population <- c(45, 40, 47, 52, 44)
N <- length(population)

# Draw a random sample of size n = 2 (without replacement)
set.seed(123)
n <- 2
```

```r
sample_plots <- sample(population, n, replace = FALSE)
sample_plots
```

Listing 1: R code: Draw a random sample of wheat plots

Definition 1.2. Sampling *is the drawing of samples. The ratio t of the sample size n to the population size N from which it was drawn is called the sampling rate or sampling fraction, i.e.*

$$t = \frac{n}{N}.$$

Example 1.2. ***Computing the sampling fraction.***

If from the 5 plots above the agronomist samples $n = 2$ plots, then the sampling fraction is $t = 2/5 = 0.4$.

```r
N <- length(population)
n <- 2
t <- n / N
t
```

Listing 2: R code: Compute sampling fraction

Definition 1.3. *A* random sampling *is a selection of n individuals from a mother population such that all possible combinations of n individuals have the same probability of being selected.*

Example 1.3. ***Random sampling with and without replacement.***

In a large field, the researcher may either sample plots without replacement (each sampled plot removed from the pool) or with replacement (allowing repeated picks, useful when modeling or bootstrapping). For finite small experiments choose without replacement; for large populations or simulations replacement sampling is common.

```r
# Without replacement
sample(population, 3, replace = FALSE)

# With replacement (useful for bootstrap or simulation)
sample(population, 3, replace = TRUE)
```

Listing 3: R code: Sampling with and without replacement

There are other forms of sampling, but in this chapter we only consider random sampling.

Remark 1.1. *We aim to describe a characteristic C, qualitative or quantitative, present in a population Ω by studying the results obtained on a sample of size n.*

Example 1.4. *(1) Given a population, we may be interested in quantitative characteristics such as weight, height, yield, nitrogen concentration, etc.*

(2) Given a population, we may be interested in qualitative characteristics such as disease presence/absence, germination success, or variety classification.

(3) In the initial student example, the characteristic studied is the weekly time devoted to studying statistics. In agronomy, the analogous characteristic could be 'yield per hectare' or 'seed germination (yes/no)'.

Definition 1.4. *Let C be a quantitative characteristic defined on a mother population Ω. C is the realization of a random variable X defined on Ω:*

$$X : \Omega \to \mathbb{R}, \qquad \omega_i \mapsto X(\omega_i) = x_i.$$

The n-sample of values of X is the list of observed values $(x_1, x_2, \ldots, x_n)$ taken by X on a sample $(\omega_1, \ldots, \omega_n)$ of Ω. These coordinates can be considered as realizations of the random vector $(X_1, \ldots, X_n)$ called the n-sample of X, where the X_i's are identically distributed and independent (i.i.d.).

Definition 1.5. *A* statistic *is any random variable that can be written using the sample random variables $X_1, \ldots, X_n$.*

Example 1.5. *The individual random variables X_i, and the sample mean*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

are statistics.

If we extract several samples of fixed size n, the results we obtain vary because they depend on the considered sample. This is called *sampling fluctuations*. We therefore study the probability laws that govern these fluctuations to draw valid conclusions about the mother population.

1.2 Sampling distributions

1.2.1 Sample mean – sample variance

Definition 1.6. *Consider a population Ω whose elements have a quantitative characteristic C that is the realization of a random variable X with expectation μ and standard deviation σ. Assume the population is infinite or sampling is done with replacement.*

Take a sample $(X_1, \ldots, X_n)$ of X with observed values $(x_1, \ldots, x_n)$. The sample mean is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is the value taken by the random variable

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

called the sample mean. *Similarly the (uncorrected) sample variance v of the observed sample is*

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

which is the value taken by the random variable

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We define the corrected sample variance S^2 (the unbiased estimator of the population variance) by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}V.$$

Example 1.6. ***Sample mean and variance for maize yield.***

A technician measures maize yield (tons/ha) on 4 sampled plots and records:

$$6.2,\ 6.5,\ 6.8,\ 7.1 \quad \textit{(t/ha)}.$$

Compute the sample mean and sample variance as estimators of the field mean and variability.

```r
# Maize yields in sample (t/ha)
yield <- c(6.2, 6.5, 6.8, 7.1)
mean_yield <- mean(yield)
var_yield <- var(yield)   # sample variance (denominator n-1)
mean_yield
var_yield
```

Listing 4: R code: Maize yield sample mean and variance

Proposition 1.1. *(1) For any distribution of X,*

$$\mathbb{E}(\bar{X}) = \mu, \qquad \mathbb{E}(V) = \frac{n-1}{n}\sigma^2, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \mathbb{E}(S^2) = \sigma^2.$$

(2) If $X \sim \mathcal{N}(\mu, \sigma^2)$ (normal distribution with mean μ and variance σ^2), then:

(i) If σ is known,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

(ii) If σ is unknown,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} denotes Student's t distribution with $n-1$ degrees of freedom.

(iii)

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where χ^2_{n-1} is the chi-squared distribution with $n-1$ degrees of freedom.

Example 1.7. ***Sample variance for soil nitrogen.***

Ten soil samples (mg/kg) are collected from an experimental plot:

$$14.8,\ 15.2,\ 13.9,\ 16.1,\ 15.0,\ 15.4,\ 14.7,\ 16.0,\ 14.9,\ 15.1.$$

Compute the sample mean and sample standard deviation to estimate the field average nitrogen content and its variability.

```r
nitrogen <- c(14.8,15.2,13.9,16.1,15.0,15.4,14.7,16.0,14.9,15.1)
mean(nitrogen)    # mean nitrogen (mg/kg)
sd(nitrogen)      # sample standard deviation
```

Listing 5: R code: Soil nitrogen mean and standard deviation

1.2.2 Sample proportion

Sometimes the characteristic to estimate is qualitative. In that case we estimate the proportion p of individuals with that characteristic using a sample.

Definition 1.7. *Let K be the random variable counting the number of occurrences of the characteristic in an n-sample. The sample proportion is*

$$F = \frac{K}{n}.$$

By definition $K \sim \text{Binomial}(n, p)$, hence

$$\mathbb{E}(K) = np, \qquad \text{Var}(K) = npq,$$

with $q = 1-p$. Therefore

$$\mathbb{E}(F) = p, \qquad \text{Var}(F) = \frac{pq}{n}.$$

Example 1.8. ***Seed germination proportion.***

A seed technologist tests $n = 100$ seeds and observes $K = 92$ seedlings after 14 days. The sample proportion is $\hat{p} = 92/100 = 0.92$. We can compute its estimated standard error and a confidence interval.

```r
n <- 100
K <- 92
p_hat <- K / n
se_p <- sqrt(p_hat * (1 - p_hat) / n)
p_hat
se_p
```

Listing 6: R code: Seed germination proportion and standard error

Remark 1.2. *For $n \geq 30$, and if $np \geq 15$ and $nq \geq 15$, the distribution of F can be approximated by a normal distribution $\mathcal{N}\left(p, \sqrt{pq/n}\right)$.*