Chapter 4: Estimation

Dr. Chennaf Bouchra

Contents

1 Estimation

1.1 Estimators

Estimating a parameter means finding an approximate value using results obtained on a sample.

Example 1.1. *Estimate the population mean from the empirical mean obtained on a sample drawn from that population.*

Definition 1.1. *An estimator $\hat{\theta}$ of an unknown parameter θ is a function that maps a sequence of observations to an approximate value $\hat{\theta}$ of θ, called the estimate:*

$$\hat{\theta} : (x_1, \ldots, x_n) \mapsto \hat{\theta} = f(x_1, \ldots, x_n).$$

```r
# Sample wheat yields (q/ha)
sample_yield <- c(45, 47, 44)
mean(sample_yield)   # point estimator for population mean
```

Listing 1: R code: Example computing a point estimator (mean) for wheat yield

An estimator $\hat{\theta}$ is itself a random variable. We can compute its expectation $\mathbb{E}(\hat{\theta})$ and variance $\mathrm{Var}(\hat{\theta})$. Those quantities allow us to judge the quality of $\hat{\theta}$.

Different estimators may exist for the same parameter. For example, to estimate the population mean one can use the arithmetic mean, the median, etc.

Definition 1.2. *An estimator $\hat{\theta}$ is* unbiased *if*

$$\mathbb{E}(\hat{\theta}) = \theta.$$

Otherwise $\hat{\theta}$ is biased. *The bias of $\hat{\theta}$ is defined by*

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Remark 1.1. *Lack of bias alone is not sufficient to guarantee that an estimator is good. A parameter θ may have several unbiased estimators. One then compares them by their variances: an estimator with large variance may take values far from the true parameter value even if it is unbiased.*

Definition 1.3. *An unbiased estimator $\hat{\theta}_1$ is called* efficient *if it has the smallest variance among unbiased estimators. Formally, if $\hat{\theta}_2$ is any unbiased estimator with $\mathbb{E}(\hat{\theta}_1) = \mathbb{E}(\hat{\theta}_2) = \theta$, then*

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2).$$

Definition 1.4. *An estimator $\hat{\theta}$ is* consistent *(or convergent) if its distribution concentrates around θ as $n \to \infty$, equivalently if*

$$\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0.$$

1.1.1 Common estimators

(A) Quantitative case Let X be a random variable with mean μ and standard deviation σ defined on a mother population Ω. Let $(X_1, \ldots, X_n)$ be an n-sample.

Proposition 1.1. *(a) $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased and consistent estimator of μ:* $\mathbb{E}(\bar{X}) = \mu$.

(b) $V = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator of the variance σ^2.

(c) $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \dfrac{n}{n-1}V$ is an unbiased and consistent estimator of σ^2.

(B) Qualitative case For a qualitative characteristic with population proportion p, the sample proportion F is an unbiased and consistent estimator of p.

1.2 Confidence intervals

Rather than providing a single estimated value for a parameter θ, we often seek an interval in which θ lies with high probability.

Definition 1.5. *Let X be a random variable whose distribution depends on a parameter θ. An interval estimator (confidence interval) of risk (or significance level) α for the parameter θ, obtained from different n-samples $(x_1, \ldots, x_n)$, is a random interval $[a(x_1, \ldots, x_n),\, b(x_1, \ldots, x_n)]$ such that a proportion $1 - \alpha$ of these intervals contain θ:*

$$P\Big(a \leq \theta \leq b\Big) = 1 - \alpha.$$

Remark 1.2. *(1) The quantity $1 - \alpha$ is called the* confidence level.

(2) In practice one typically has a single observed sample which yields the observed confidence interval $[a, b]$.

(3) Parameters commonly estimated are the mean or variance in the quantitative case, and the proportion in the qualitative case.

In the following we focus on symmetric confidence intervals such that

$$P\Big(\hat{\theta} < a\Big) = \frac{\alpha}{2}, \qquad P\Big(\hat{\theta} > b\Big) = \frac{\alpha}{2}.$$

We then determine random variables A_n and B_n (functions of the sample) such that

$$P\Big(A_n \leq \theta \leq B_n\Big) = 1 - \alpha,$$

and the realized interval is $[a, b]$ with $a = A_n(x_1, \ldots, x_n)$ and $b = B_n(x_1, \ldots, x_n)$.

1.2.1 Confidence interval for a mean

We consider two cases: either the parent distribution is normal with parameters μ, σ, or the distribution is unknown but $n > 30$ so that the sample mean is approximately normal by the central limit theorem.

Let the observed sample be $(x_1, \ldots, x_n)$. Denote

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

(A) Known σ We know that

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

Let $z_{1-\alpha/2}$ denote the $1-\alpha/2$ quantile of the standard normal distribution. Then

$$P\Big(-z_{1-\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\Big) = 1-\alpha,$$

which yields the confidence interval

$$\text{CI} = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Example 1.2. ***95% CI for mean maize yield (known sigma approximation).***

Assume a researcher knows the historical field standard deviation for maize yield is approx. $\sigma = 0.35$ *t/ha. She samples* $n = 9$ *plots and obtains sample mean* $\bar{x} = 6.7$ *t/ha. The 95% CI (using normal quantile) is:*

$$6.7 \pm 1.96 \times \frac{0.35}{\sqrt{9}}.$$

```r
xbar <- 6.7
sigma <- 0.35
n <- 9
z <- qnorm(0.975)
se <- sigma / sqrt(n)
c(xbar - z * se, xbar + z * se)
```

Listing 2: R code: CI for mean with known sigma (approximation)

(B) Unknown σ If σ is unknown, replace σ by s and use Student's t distribution with $n-1$ degrees of freedom. Let $t_{1-\alpha/2,n-1}$ denote the corresponding quantile. Then

$$\text{CI} = \left[\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right].$$

Remark: If $n > 30$, often $t_{1-\alpha/2,n-1} \approx z_{1-\alpha/2}$.

Example 1.3. ***95% CI for mean wheat yield (unknown sigma).***

A sample of $n = 8$ *wheat plots yields:*

$$6.2,\ 6.5,\ 6.8,\ 7.1,\ 6.4,\ 6.7,\ 7.0,\ 6.6 \quad \textit{(t/ha)}.$$

Compute the 95% t-based CI for the mean.

```r
yield <- c(6.2,6.5,6.8,7.1,6.4,6.7,7.0,6.6)
t.test(yield, conf.level = 0.95)
```

Listing 3: R code: t-based 95% CI for mean yield

1.2.2 Confidence interval for a variance

Assume X is normally distributed with parameters μ, σ^2.

(A) Known μ If μ is known, define

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then a $1-\alpha$ confidence interval for σ^2 is

$$\text{CI} = \left[\frac{nv}{\chi^2_{1-\alpha/2}(n)}, \quad \frac{nv}{\chi^2_{\alpha/2}(n)}\right],$$

where $\chi^2_p(n)$ denotes the p-quantile of the chi-squared distribution with n degrees of freedom.

(B) Unknown μ If μ is unknown, let

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

Then a $1-\alpha$ confidence interval for σ^2 is

$$\text{CI} = \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \quad \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right].$$

Example 1.4. ***95% CI for variance of soil nitrogen.***

Using the soil nitrogen sample from above (10 observations), compute a 95% CI for the population variance.

```r
nitrogen <- c(14.8,15.2,13.9,16.1,15.0,15.4,14.7,16.0,14.9,15.1)
n <- length(nitrogen)
s2 <- var(nitrogen)          # sample variance (denominator n-1)
alpha <- 0.05
chi_low <- qchisq(1 - alpha/2, df = n - 1)
chi_high <- qchisq(alpha/2, df = n - 1)
ci_lower <- (n - 1) * s2 / chi_low
ci_upper <- (n - 1) * s2 / chi_high
c(ci_lower, ci_upper)
```

Listing 4: R code: 95% CI for variance (soil nitrogen)

Remark 1.3. *For large n, one can use approximations to the chi-squared quantiles. For instance (approximation), one may use:*

$$\chi^2_\alpha(n-1) \approx \frac{1}{2}\Big(t_\alpha + \sqrt{2n-3}\Big)^2$$

under certain approximations, noting also the symmetry relation for the Student quantiles $t_{\alpha/2} = -t_{1-\alpha/2}$.

1.2.3 Confidence interval for a proportion

Recall that the sample proportion F can be approximated by a normal distribution $\mathcal{N}(p, \sqrt{pq/n})$ for sufficiently large n, where $q = 1 - p$. Replacing p by the observed proportion f and using the standard normal quantile $z_{1-\alpha/2}$, we get the approximate $1-\alpha$ confidence interval

$$\text{CI} = \left[f - z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}, \quad f + z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \right].$$

Example 1.5. ***95% CI for germination proportion.***

From $n = 200$ soybean seeds, $K = 180$ germinated. Compute the approximate 95% CI for the germination proportion.

```r
K <- 180
n <- 200
f <- K / n
z <- qnorm(0.975)
se <- sqrt(f * (1 - f) / n)
c(f - z * se, f + z * se)
```

Listing 5: R code: 95% CI for proportion (approximate)