Parameter Estimation

N. Torelli, L. Egidi, G. Di Credico

Spring 2020

University of Trieste

Point estimation

The aim of point estimation

Given a model for the data **y**, with parameter θ, **point estimation** is concerned with finding a reasonable parameter estimate from the data.

There are several methods for doing this, and the problem can be simply stated as *finding the parameter value most consistent with the data*, a definition that leads to the method of **maximum likelihood estimation**.

We will delve into the details of maximum likelihood estimation in due time, but here we focus on some general aspects of point estimation.

Example: sample mean and sample variance

A very simple model assumes that the data are a random sample from a normal distribution namely they are the observations of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

Straightforward estimates of μ and σ^2 are given by **the sample mean**

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and by the **sample variance**

$$\widehat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2\,.$$

Such estimates are actually sensible anytime we are interested in estimating the mean and variance of an i.i.d. sample.

Estimation properties

To figure out what could be a good estimate, we need to consider *repeated estimation* under *repeated replication of the data-generating process*.

This makes fully sense whenever the available data are a random sample obtained from a large population, like in industrial or social surveys, so that it would perfectly possible to iterate the sampling and obtain further data with the same structure of $\mathbf{y}$.

However, we apply the same logic even when repetition is just the result of an idealization, like in the case of the temperatures recorded in New Haven of the previous lecture.

The point is: what do we expect to find if we repeat the same analysis to many data sets generated from the same model?

Unbiasedness

If we replicate the random data and we repeat the estimation process, the result will be a different value of $\widehat{\boldsymbol{\theta}}$ for each replicate.

The values are observations of a random vector, the **estimator** of $\boldsymbol{\theta}$, which is usually also denoted by $\widehat{\boldsymbol{\theta}}$ (the context will make clear whether we are referring to the estimator or to the estimate for a given sample).

Since, the estimator is a r.v., it makes fully sense to compute its mean.

For an **unbiased** estimator

$$E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}\,.$$

Unbiasedness is a desirable property, and we would also like the estimator to have **low variance**.

Mean Squared Error

There is *tradeoff* between unbiasedness and low variance, so we usually seek to get both (to some extent): ideally we would target a small **Mean Squared Error (MSE)**

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}) = E\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\}\,.$$

With some algebra, we obtain

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}) = \{E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\}^2 + \mathrm{var}(\widehat{\boldsymbol{\theta}}) = \text{Squared bias} + \text{Variance}\,.$$

Example: normal random sample

For a normal random sample, it is straightforward to verify that

$$E(\overline{Y}) = \mu\,, \qquad \mathrm{var}(\overline{Y}) = \frac{\sigma^2}{n} = \mathrm{MSE}(\overline{Y})\,.$$

For the sample variance, we use the property that

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

to obtain

$$E(S^2) = \sigma^2\,, \qquad \mathrm{var}(S^2) = \frac{2\,(n-1)\,\sigma^4}{n^2} = \mathrm{MSE}(S^2)\,.$$

The unbiasedness of the sample mean and variance is a general property, holding also for non-normal samples.

Consistency

A (scalar) estimator is said to be **(weakly) consistent** if, for any $\epsilon > 0$

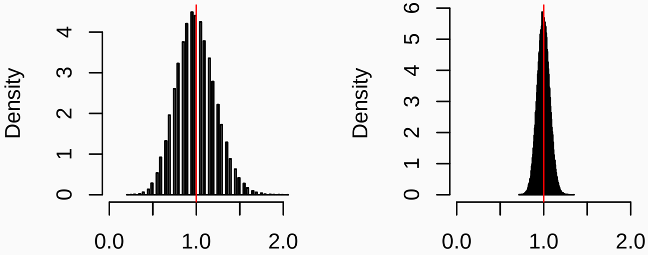
$$\Pr(|\widehat{\theta} - \theta| > \epsilon) \to 0\,, \qquad \text{as } n \to \infty\,.$$

A sufficient condition for this is that $MSE(\widehat{\theta}) \to 0$ for large samples, which requires that both bias and variance become negligible.

The law of large samples implies that the sample mean is a consistent estimator for the true mean in random samples.

R lab: consistency of the sample mean

```r
M <- 100000; n1 <- 20; n2 <- 200; y1 <- y2 <- rep(NA, M)
for(i in 1:M) {y1[i] <- mean(rpois(n1, 1))
               y2[i] <- mean(rpois(n2, 1))}
par(mfrow=c(1,2))
hist.scott(y1, xlim=c(0,2), main="", xlab=""); abline(v=1,col=2)
hist.scott(y2, xlim=c(0,2), main="", xlab=""); abline(v=1,col=2)
```



Efficiency

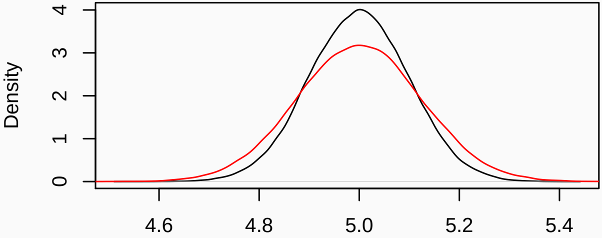
An **efficient estimator** is an estimator that estimates the parameter of interest in some *optimal* manner.

Among estimators with negligible bias, efficiency is associated to small variance. Since this is the case of consistent estimators, they are usually compared in terms of their variance.

R lab: efficiency of the sample mean

For a normal random sample, both the sample mean and sample median are consistent estimators of μ. The mean is more efficient.

```r
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {y <- rnorm(n, 5)
               mat.y[i,] <- c(mean(y), median(y))}
plot(density(mat.y[,1]), type="l", main="")
lines(density(mat.y[,2]), col=2)
```



Standard Error

An important quantity defined for a (scalar) estimator is given by its **standard error**, defined as

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta})}\,.$$

Once a sample is observed, and a numerical estimate of θ obtained, then the estimated standard error is obtained by replacing θ by $\widehat{\theta}$.

An example is the **standard error of the mean** $\mathrm{SE}(\overline{Y}) = \sigma/\sqrt{n}$, which is estimated by $s/\sqrt{n}$.

In applications, the estimated standard error is routinely reported along with the estimate, since it quantifies the **estimation precision**.

The delta method

Suppose that we are interested in a parameter which is a function of a scalar parameter θ, namely

$$\psi = g(\theta)\,, \qquad \text{for a continuous and differentiable function } g\,.$$

If $\widehat{\theta}$ is a consistent estimator of θ, then the **continuous mapping theorem** ensures that $g(\widehat{\theta})$ is consistent for ψ.

Its standard error is provided by the **delta method**, stating that

$$\mathrm{SE}(\widehat{\psi}) \doteq \mathrm{SE}(\widehat{\theta})\,|g'(\theta)|\,,$$

with the approximation becoming more accurate for larger samples.

The result can be extended to settings with multiple parameters.

Robust estimation

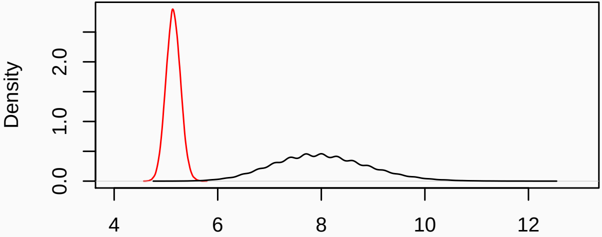
A **robust** estimator has good performances across a wide range of statistical models for the data.

The **sample median** is a robust estimation of location, not affected by possible outlying data, quite the opposite of the sample mean.

Robust estimation trades some efficiency with resistance to outliers, and they are often a sensible choice for semi-automatic data analyses.

R lab: robustness of the sample median

```
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { x <- rbinom(n, size = 1, prob = 0.9)
                y <- x * rnorm(n, 5) + (1 - x) * rnorm(x, 35)
                mat.y[i,] <- c(mean(y), median(y))}
plot(density(mat.y[,2]), type="l", main="", xlim=c(4, 13),
     col = 2)
lines(density(mat.y[,1]), col=1)
```



Interval estimation

The aim of interval estimation

Confidence intervals provide more satisfactory estimation results than point estimates alone, giving an entire set of values to estimate the model parameter.

They are built by considering a single parameter at a time.

Extensions to multidimensional *confidence regions* exist, but they are seldom used in practice.

Pivots

Confidence intervals make suitable usage of **pivots**, which are **functions of the data and the parameter whose distribution is known**.

A notable example is the following one for a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, when the parameter of interest is the mean μ, and σ^2 is not known (so that $\boldsymbol{\theta} = (\mu, \sigma^2)$):

$$T(\mu) = \frac{\overline{Y} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}\,, \qquad \forall \mu \in \mathbb{R}\,, \sigma^2 > 0$$

Obtaining a confidence interval

In the normal random sample example, from the previous pivot property it follows that (for $0 < \alpha < 1$)

$$\Pr\left(t_{n-1;\alpha/2} \leq T(\mu) \leq t_{n-1;1-\alpha/2}\right) = 1 - \alpha\,,$$

where $t_{n-1;\alpha}$ is the α quantile of a t_{n-1} distribution; due to symmetry of the latter, $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$.

With some simple algebra, the previous property is equivalent to

$$\Pr\left(\overline{Y} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}} \leq \mu \leq \overline{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1 - \alpha\,.$$

Definition of confidence interval

Hence the *random interval* with endpoints

$$\overline{Y} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}\,, \qquad \overline{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}$$

contains μ with probability $(1-\alpha)$.

This interval is called a $(1-\alpha) \times 100\%$ **confidence interval**.

Common choices are $(1-\alpha) = 0.95$ or $(1-\alpha) = 0.99$.

Interpretation

Given a particular set of data $y_1, \ldots, y_n$ we calculate the confidence interval by replacing $\overline{Y}$ and S^2 with their observed values $\overline{y}$ and s^2

$$\overline{y} - t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n}}, \qquad \overline{y} + t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n}}$$

This interval *either does or does not contains the true value of* μ.

The probability interpretation previously introduced refers to an *hypothetical sequence of sets of data* generated from the statistical model.

R lab: confidence interval

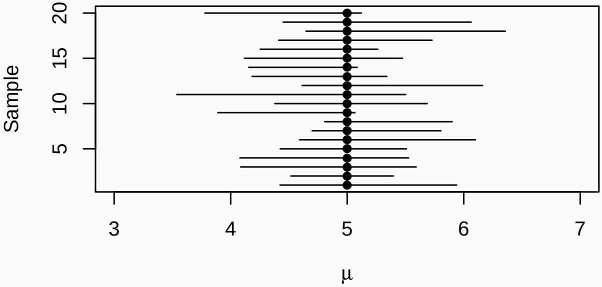
```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                se_t <- sqrt(var(y) / n)  * qt(0.975, n-1)
                mat.ci[i,] <- mean(y) + se_t * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
```

```
## [1] 0.95032
```

R lab: visualizing confidence intervals

We can visualize the first 20 simulated confidence intervals, expecting that (on average) 19 out of 20 will include the true μ

```
plot(rep(5, 20), 1:20, pch = 16, ylab="Sample",
     xlab=expression(mu))
for(i in 1:20) segments(mat.ci[i,1],i,mat.ci[i,2],i)
```



One-sided confidence intervals

If we lift the equi-tailed condition, we can define infinitely many intervals such that

$$\Pr\left(\overline{Y} - t_{n-1;1-\alpha_1}\sqrt{\frac{S^2}{n}} \leq \mu \leq \overline{Y} + t_{n-1;1-\alpha_2}\sqrt{\frac{S^2}{n}}\right) = 1 - \alpha\,,$$

where $\alpha_1 + \alpha_2 = \alpha$.

Other than the standard choice $\alpha_1 = \alpha_2 = \alpha/2$, other notable choices are $\alpha_1 = 0$ (which makes the lower limit equal to $-\infty$) or $\alpha_2 = 0$ (which makes the upper limit equal to ∞).

They are called **one-sided confidence intervals**, and are sometimes employed in applications.

Approximate confidence intervals & coverage probability

Exact pivots are scarce, but approximate ones are easy to find.

A common one is the **Wald pivot** for a generic parameter of interest ψ, based on a consistent estimator which is approximately normally distributed for large samples

$$Z(\psi) = \frac{\widehat{\psi} - \psi}{\mathrm{SE}(\widehat{\psi})} \,\dot{\sim}\, \mathcal{N}(0,1)\,, \qquad \forall \psi \in \Psi$$

The corresponding confidence interval is

$$\widehat{\psi} - z_{1-\alpha/2}\,\mathrm{SE}(\widehat{\psi})\,, \qquad \widehat{\psi} + z_{1-\alpha/2}\,\mathrm{SE}(\widehat{\psi})$$

The Central Limit Theorem provides such a solution for random samples, when ψ corresponds to the mean of each variable.

R lab: approximate confidence intervals

```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                se_z <- sqrt(var(y) / n)  * qnorm(0.975)
                mat.ci[i,] <- mean(y) + se_z * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
```

```
## [1] 0.91861
```

```
M <- 100000; n <- 100; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                se_z <- sqrt(var(y) / n)  * qnorm(0.975)
                mat.ci[i,] <- mean(y) + se_z * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)
```

```
## [1] 0.94776
```

Confidence interval for a proportion

The method for approximate intervals can be readily used for confidence intervals on a proportion π, the success probability of a random sample of n binary variables,

$$Y_i \sim \mathcal{B}(1, \pi)\,, \qquad\qquad i = 1, \ldots, n\,.$$

Here the pivot is

$$Z(\pi) = \frac{\overline{Y} - \pi}{\sqrt{\dfrac{\overline{Y}\,(1 - \overline{Y})}{n}}} \;\dot{\sim}\; \mathcal{N}(0, 1)\,, \qquad\qquad \forall \pi \in (0, 1)\,,$$

since $\widehat{\pi} = \overline{Y}$ and $\mathrm{SE}(\widehat{\pi}) = \sqrt{\dfrac{\pi\,(1 - \pi)}{n}}$, which is estimated by plugging-in $\widehat{\pi}$ in place of π.

R lab: confidence interval for a proportion

```r
M <- 100000; n <- 50; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rbinom(n, size = 1, prob = 0.25)
                p.hat <- mean(y)
                se_z <- sqrt(p.hat * (1 - p.hat) / n)
                se_qz <- se_z * qnorm(0.975)
                mat.ci[i,] <- mean(y) + se_qz * c(-1, 1)}
mean(mat.ci[,1] < 0.25 & mat.ci[,2] > 0.25)
```

```
## [1] 0.94005
```

Confidence interval for a difference of means

An important application concerns the computation of the confidence interval for the difference between two means $\delta = \mu_X - \mu_Y$.

For two independent (and large) random samples, the approximate normal pivot is

$$Z(\delta) = \frac{\widehat{\delta} - \delta}{\mathrm{SE}(\widehat{\delta})} \,,$$

with $\widehat{\delta} = \overline{X} - \overline{Y}$ and $\mathrm{SE}(\widehat{\delta}) = \sqrt{\mathrm{SE}(\overline{X})^2 + \mathrm{SE}(\overline{Y})^2}$.

Again, for normal samples, exact solutions exist, both for the case of equal variances and for the case of unequal variances.