Point and interval estimates. Testing statistical hypotheses in R

Kaloyan Ganev

2022/2023

Lecture contents

Introductory notes

Types of data and statistical tests

- As we already mentioned many times, data can be either discrete or continuous
- In economics we work with random variables which can be either of the two types
- Therefore, we will split the overview in two respective parts – one on continuous, and one on discrete random variables
- Before that, we still need some theory (at least a refresher)

Some theory

Point estimation

- Suppose we know the type of distribution of the random variable X but we don't know the value of its parameter θ
- We want to estimate it based on the available information set Ω (also called *parameter space*)
- Formally written, the p.d.f. is:

$$f(x;\theta),\ \theta \in \Omega$$

- Note that $f(\cdot)$ is a *family* of distributions – for each value of θ we have a different distribution

Maximum likelihood estimation

- Suppose a sample of size n is drawn, where each drawing is independent from the others and comes from one and the same distribution
- Take the multivariate normal distribution as an example; its p.d.f. is:

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- Suppose that we have a family of normal distributions with variance 1
- Suppose that we have a sample drawn from such a distribution but we do not know which one exactly
- In other words, we don't know the mean (θ) of the distribution and need to estimate it

Maximum likelihood estimation (2)

- In the case where just the mean parameter needs to be estimated (θ), and the variance is known to be equal to 1, form the likelihood function:

$$L(\theta; x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

- Note that because of the fact that $X_1, X_2, \ldots, X_n$ are i.i.d., the covariance matrix $\mathbf{\Sigma}$ has a determinant equal to the product of its diagonal elements, i.e. $(\sigma^2)^n = 1$.
- We aim to find this value of θ which best describes the sample (maximizes the likelihood of observing this sample)
- The normal p.d.f. is concave so setting the first derivative to zero gives us the maximizing value of θ

Maximum likelihood estimation (3)

- It is more convenient to work with natural logs (such a transformation does not change the concavity of the function)

$$\frac{d \ln L(\theta; x_1, x_2, \ldots, x_n)}{d\theta} = \sum_{i=1}^{n} (x_i - \theta) = 0$$

- This gives the result:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The method of moments

- What if the maximization problem cannot be solved analytically?
- One way is to use numerical methods
- Another way – use the method of moments
- Essence: equations relating the population moments to the parameters subject to estimation are derived
- Then the sample moments are calculated from the sample data
- The sample moments are substituted for the population moments and the system of equations is solved with respect to the unknown parameters

The method of moments: A simple example

- We consider the case of only one parameter to be estimated
- Let $X_1, X_2, \ldots, X_k$ be independent random variables drawn from the same distribution having p.d.f. $f(x;\theta)$
- The theoretical mean of the distribution is:

$$\mu_X = \int_{-\infty}^{+\infty} x f(x;\theta) dx = m(\theta)$$

- Asymptotically it is true that the sample mean converges to the true population mean
- If the sample size is large enough, we can approximate then the population mean with the sample mean, i. e.:

$$\overline{X} \approx m(\theta)$$

- The estimator $\widehat{\theta}$ is found by assuming equality and solving the equation with respect to θ

Three definitions

Definition 1

The estimator $\widehat{\theta}$ is **unbiased** if:

$$\mathsf{E}(\widehat{\theta}) = \theta$$

Definition 2

The estimator $\widehat{\theta}$ is **consistent** if:

$$\widehat{\theta} \xrightarrow{p} \theta$$

(Recall that convergence in probability means $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$)

Definition 3

The estimator $\widehat{\theta}$ is called **unbiased minimum variance estimator** if it is unbiased and its variance is less than or equal to the variance of every other unbiased estimator of θ.

A theorem

Theorem 1

Let $X_1, X_2, \ldots, X_n$ be independent random variables with normal distributions $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), \ldots, N(\mu_n, \sigma_n^2)$. Let $k_1, k_2, \ldots, k_n$ be constants. Then the random variable:

$$Y = \sum_{i=1}^{n} k_i X_i$$

has normal distribution with mean $\sum_{i=1}^{n} k_i \mu_i$ and variance $\sum_{i=1}^{n} k_i^2 \sigma_i^2$.

Confidence intervals for means

- In essence, it is an interval estimate, unlike the point one
- Suppose we know the variance of the normal distribution from which samples are drawn but the mean is unknown
- Take the ML estimator of the mean, $\widehat{\mu} = \overline{X}$
- We know that $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- From this follows that:

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

Confidence intervals for means (2)

- Therefore,
$$P\left(-2 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 2\right) = 0.954$$
- This is equivalent to:
$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954$$
- As σ, n and $\overline{X}$ are known, $\overline{X} \pm \frac{2\sigma}{\sqrt{n}}$ are statistics
- Interpretation: Prior to taking the sample, there is a probability of 0.954 that the confidence interval will contain the true mean μ
- In other words, if we make 100 samples, in 95.4% of the cases, the true mean will be in the confidence interval

Confidence intervals for means (3)

- Let in the same problem the variance σ^2 be also unknown
- In this case the variance we know is the sample one, s^2
- First, let's see what the distributions of the sample mean and variance are
- With respect to the sample mean, using the Theorem, we find that it has a normal distribution with mean:

$$\sum_{i=1}^{n} \frac{1}{n}\mu = \mu$$

and variance:

$$\sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}$$

Confidence intervals for means (4)

- Concerning the sample variance, recall that the joint p.d.f. is written as:

$$f(x;\mu,\sigma^2) = \left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

- The numerator of the exponent argument can be written as:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2 = \\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2 + 2(\overline{x}-\mu)\sum_{i=1}^{n}(x_i-\overline{x}) + \sum_{i=1}^{n}(\overline{x}-\mu)^2\end{aligned}$$

Confidence intervals for means (5)

- Taking into account that $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, the latter becomes:

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2$$

- Thus, the p.d.f. can be rewritten as:

$$f(x;\mu,\sigma^2) = \left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{2\sigma^2} - \frac{n(\overline{x} - \mu)^2}{2\sigma^2}\right) \qquad (*)$$

Confidence intervals for means (6)

- Now let's consider the joint distribution of $Y_1 = \overline{X}, Y_2 = X_2, \ldots, Y_n = X_n$
- It can be proved that it is:

$$n\left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\right)^n \quad \exp\left(-\frac{(ny_1 - y_2 - \ldots - y_n - y_1)^2}{2\sigma^2} - \right.$$

$$\left. - \frac{\sum_2^n (y_i - y_1)^2}{2\sigma^2} - \frac{n(y_1 - \mu)^2}{2\sigma^2}\right) \quad (**)$$

- After dividing $(**)$ by $(*)$, it can then be proved that $ns^2/\sigma^2 \sim \chi^2(n-1)$

Tests

Comparing means: The t-test

- Basic assumption: data are independent drawings from a normal distribution
- Suppose we want to test the hypothesis that the empirical mean equals some pre-specified value μ_0
- Therefore, our null hypothesis is $H_0 : \mu = \mu_0$, and the alternative is $H_1 : \mu \neq \mu_0$
- Define the standard error of the MLE of the mean as the square root of the sample variance:

$$SE(\overline{x}) = \frac{\sigma}{\sqrt{n}}$$

- The test statistic is:

$$t = \frac{\overline{x} - \mu_0}{SE(\overline{x})}$$

Comparing means: The t-test (2)

- Define the significance level to 5%
- This is the probability of committing the so-called **Type I error**: to reject the null when it is true
- (**Type II error:** to accept the null when it is false)
- Let's see how the test is implemented in R
- We will consider two cases: the sample mean checked against a specified number, and two sample means checked against each other

Comparing means: The t-test (3)

- Load `t-test.csv`:

```
tt <- read.csv("t-test.csv")
attach(tt)
```

- Run the test:

```
t.test(var1, mu = 3)
```

- The p-value that we get signifies that we cannot reject the null
- Note that R also reports the confidence interval for the true mean and the sample mean itself
- The confidence interval is calculated as follows:

$$\overline{x} - t_{0.975} \cdot s < \mu < \overline{x} + t_{0.975} \cdot s$$

Comparing means: The t-test (4)

- To compare the means of the two variables, you just issue:

```
t.test(var1, var2)
```

- There are additional options in the test specification, e.g. to specify a significance level other than 5%
- For example, if you want a 99% confidence interval for the true mean, write:

```
t.test(var1, mu = 3, conf.level = 0.99)
```

The Wilcoxon signed-rank test

- If you wish to avoid making the assumption that data comes from a normal distribution, this is a suitable alternative to the t-test
- In other words, the Wilcoxon signed-rank test is ‘distribution-free’
- Essence: subtract the theoretical mean from all sample values; exclude all zero differences
- Then, rank the results according to their *absolute* value starting with the smallest one
- Sum all ranks multiplied by their respective signs and take the absolute value of the result – this is your W statistic
- Use the table to determine significance

The Wilcoxon signed-rank test (2)

- In R, for the one-sample case this is implemented via:

```
wilcox.test(var1, mu = 3)
```

Note: For the one-sample test, R reports the V statistic which is the sum of *only positive-signed ranks*!

- and for the two-sample test it is:

```
wilcox.test(var1, var2)
```

Comparison of variances

- Needed for two-sample t-tests
- Null hypothesis: The ratio of the two variances is 1
- Since each sample variance is chi-square distributed, the ratio of the two variances is F-distributed
- Implemented in R through the following function:

```
var.test(var1, var2)
```

- How to do it manually:

```
F_stat <- var(var1)/var(var2)
p_val <- 2*pf(F_stat, 24, 24) # Two-sided test
```

Comparison of variances (2)

- Concerning the confidence interval, the hypothesized ratio $\frac{\sigma_1^2}{\sigma_2^2}$ is 1
- The confidence interval is constructed by analogy to the Gaussian case:

$$\frac{s_1^2}{s_2^2} \cdot F_{n_1-1,n_2-1,\alpha/2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \cdot F_{n_1-1,n_2-1,1-\alpha/2},$$

  where s_1^2 and s_2^2 are the estimated (sample) variances, and α is the selected level of significance
- To find the lower and upper bounds of the confidence interval in R,

```
lbound <- F_stat*qf(0.025, 24, 24)
ubound <- F_stat*qf(0.975, 24, 24)
```

Paired t-tests

- Suppose that you have two samples and you have direct correspondence between the elements in the two samples
- For example, you measure weight of babies before and after breast-feeding (done in paediatric care when babies have issues in gaining weight)
- If we want to measure the average effect, we use the paired t-tests
- Essentially, this test boils down to a one-sample test
- First, the differences between paired observations are calculated
- Then, the mean of the differences is computed
- Calculate the t-statistic (has $n - 1$ df) of the difference and check it against the tabulated distribution
- R implementation (compare results to the unpaired test):

```
t.test(var1, var2, paired = TRUE)
```

Paired Wilcoxon test

- Same as the one-sample Wilcoxon test but applied to the differences of paired observations
- Implementation:

```
wilcox.test(var1, var2, paired = TRUE)
```

Proportion test

- Suppose there are several groups of data
- Suppose you are aiming at estimating the probability of success in each group (or any other analogical ratio)
- The proportions test checks the validity of the null hypothesis that the ratios (probabilities of success) in the chosen groups of data are the same
- Also it can check whether they are equal to some pre-specified values

Proportion test (2)

- An example: Check whether the proportion of male smokers in a sample is the same as the proportion of female smokers
- R implementation

```
sexsmoke <- matrix(c(70,120,65,140),ncol=2,byrow=TRUE)
rownames(sexsmoke) <- c("male","female")
colnames(sexsmoke) <- c("smoke","nosmoke")
prop.test(sexsmoke)
```

Binomial test

- Used in cases where an experiment consists of a series of identical trials
- Each trial has two outcomes (a Bernoulli trial; e. g. tossing a fair coin)
- The distribution of the success ('heads') is given by the binomial distribution
- Example: You have a coin for which you expect a probability of success to be 0.5. You toss the coin 1000 times and get 486 occurrences of heads
- Test the hypothesis that your assumed probability of success of 0.5 is true
- In R it is done as follows:

```
binom.test(486,1000)
```

A concept: Contingency tables

- Called also ‘crosstabs’ or ‘cross tabulations’
- A table (matrix) which displays the multivariate frequency distribution of the variables
- An example:

	BA	Economics	Total
Males	27	24	51
Females	25	24	49
Total	52	48	100

The Fisher test

- Used for categorical data (nominal variables) that is obtained after classifying variables in two different ways (contingency tables)
- Purpose: to test whether the association between the two types of classification is significant (in other words – whether the two classifications are independent)[1]
- In the current example contingency table, test whether gender is independent from major (null hypothesis)
- In other words, we test whether there is some kind of nonrandom association between the two variables

[1] 'Contingency' in this context means 'chance'.

The Fisher test (2)

- The null hypothesis formalized: the relative proportions of one variable are independent from the relative proportions of another variable
- The test **does not** use a mathematical function estimating p-value of a test statistic
- Instead, it calculates the probability of getting the data observations that we have
- Fisher showed that this probability is calculated using the hypergeormetric p.d.f.

The Fisher test (3)

- Assume that the data is in general displayed in a contingency table as follows:

	Category B1	Category B2	Total
Category A1	a	b	$a+b$
Category A2	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

- If we denote $n = a+b+c+d$, then:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

The Fisher test (4)

- The null hypothesis is rejected if the calculated p-value is less than the specified level of significance
- In R the test is implemented via:

```
fisher.test(conttable)
```

- The Fisher test is very useful for small samples but works equally well for large samples
- The point is that it requires a lot of computations for large datasets
- In such cases, chi-square tests are used

Chi-square (χ^2) tests

- Most often, the Pearson chi-square test is meant[2]
- The name of the test follows from the fact that the sampling distribution of the test statistic is χ^2
- Assume that f_{ij} is the observed frequency count of events which belong to the ith category of the x variable and to the jth category of the y variable
- Suppose that the frequency count which corresponds to the case of independence between the variables is e_{ij}

[2] For other variants, see for example http://en.wikipedia.org/wiki/Chi-squared_test.

Chi-square (χ^2) tests (2)

- The following statistic is calculated:

$$\chi^2 = \frac{\sum_{i,j}(f_{ij} - e_{ij})^2}{e_{ij}}$$

- The null hypothesis is the same as in the Fisher test
- If the calculated p-value is lower than the pre-specified level of significance, then the null is rejected
- In R:

```
chisq.test(conttable)
```