14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 2-4

Review of Statistics

Intro to Statistics. Sampling

We will see some data. We want to learn about underlying laws (all potential realizations of data). For example, we may want to figure out whether a given coin is a fair coin, or to find out the mean height of a man in Massachusetts, or to assess whether beautiful people are actually paid more.

We assume that we have data which we treat as a realization of random variables $X_1, ...X_n$. These are the results of experiments.

Let $X_1, ..X_n$ be i.i.d. from a distribution F, where F stands for population. Any functional of F is called a parameter.

For example, in a coin-flip case, let X_i be i.i.d. realizations of a Bernoulli random variable (0-1 random variable), where the probability of success p is a parameter. In the average-height example, we assume that we observe the heights of several randomly chosen men $X_1, ..., X_n$, which are i.i.d. from population with cdf F, and we are interested in parameter $\mu = \int x dF(x)$.

Before the experiment is conducted, we expect to observe random variables $X_1, ..., X_n$. After realization of the experiment we get $(x_1, ..., x_n)$, a set of numbers, our data, which is one realization of the random variables. Induction: observing $(x_1, ..., x_n)$ we want to make a judgement about F. Any function of data $g(X_1, ..., X_n)$ is called statistics. It is a summary, though it might not be a good one. Before the experiment,it is represented by a random variable, and after the realization, you have one number.

Here are a few examples. In the coin toss example, one may report Y = 'number of heads appeared in n tosses' or Y = 'the first toss in which heads appeared'. In the average-height example, one may report Y = 'average height of n men', or Y= 'median (order all men by height, choose the middle one)', or Y = 'average height of all men excluding 10% tallest and 10% shortest'.

Another example, $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is a statistic. Since it is a random variable, it has a distribution. We may try to characterize it.

Statistics has three main tasks:

- Estimation: guess the value of unknown parameters (estimator is a statistics);
- Testing: assess whether a statement is true in a population, produce an answer yes (accept) or no (reject);
- Confidence set: produce a set of values of a parameter that are consistent with data.

For each of these tasks we will have some criteria of quality, that a good procedure should satisfy.

Estimation

Denote θ to be the unknown parameter we are trying to estimate. In the height example $\theta = EX_i$. If we are trying to answer whether beautiful people are paid higher wages on average, assume we observe (X_i, Y_i), where X_i a numeric summary of a beauty of a person, Y_i is the person's salary. The parameter of interest is $\theta = cov(X_i, Y_i)$.

Estimate is a statistic intended for 'guessing' a parameter value based on a sample. It may be written as a rule for how to calculate this guess from the sample, before realization of the experiment it is a random variable:

$$\widehat{\theta} = g(X_1, ..., X_n).$$

After realization of the data set, one gets a number $g(x_1, ..., x_n)$.

Qualities of estimators: unbiasedness

An estimator is called *unbiased* if its mean is equal to the true parameter value:

$$E\widehat{\theta} = \theta.$$

Explanation: $\widehat{\theta}$ is a random variable because of randomness of a sample - for different samples we calculate different values of statistics $\widehat{\theta}$. When we average over all potential values of $\widehat{\theta}$ with proper probabilities, we get the correct θ- that is, on average we are correct.

Example 1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimate for $\theta = EX_i$. Indeed,

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \theta.$$

Example 2. $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimate for variance $\sigma^2 = Var(X_i)$. Prove this by yourself. Suggestions: Define $Z_i = X_i - EX_i$, show that $Var(X_i) = Var(Z_i), EZ_i = 0, s_X^2 = s_Z^2$. Then show the needed statement for s_Z^2.

Qualities of estimators: consistency

An estimator (a set of rules) $\widehat{\theta} = g(X_1, ..., X_n)$ is consistent if

$$\widehat{\theta} \to^p \theta \text{ as } n \to \infty,$$

where the last statement means that for any $\varepsilon > 0$ we have

$$P\{|\widehat{\theta} - \theta| > \varepsilon\} \to 0 \text{ as } n \to \infty.$$

Explanation: we will never know θ for sure, but the estimator with be in a small neighborhood of the true value with probability arbitrarily close to one for large samples.

Example 3. If $E|X_i| < \infty$, then $\bar{X}$ is consistent for EX_i due to Law of Large Numbers.

Example 4. If $EX_i^2 < \infty$, then s^2 is consistent for σ^2. Sketch of the proof:

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

Next we notice that by the Law of Large Numbers $\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 \to^p E[X^2]$, $\frac{n}{n-1} \to 1$ and $\bar{X} \to^p E[X]$. Then we use a continuous mapping theorem.

Qualities of estimators: efficiency

If we have two unbiased estimators, that is, estimators centered correctly, we should prefer the one with smaller variance (or as we say 'higher efficiency'). If we consider $\widehat{\theta} = \bar{X}$ as an estimator for EX_i, we know it is unbiased. Efficiency is measured by

$$MSE = E(\widehat{\theta} - \theta)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Let us consider a set of linear estimators for $\theta = EX_i$, it means we would construct an estimator as

$$\widehat{\theta} = \sum_{i=1}^{n} \omega_i X_i,$$

with some non-random weights ω_i. What weights would bring us to an unbiased estimator? The ones such that $\sum_{i=1}^{n} \omega_i = 1$ (prove it). Which linear unbiased estimator is the most efficient one? For this we have to solve the following optimization problem:

$$Var(\widehat{\theta}) = Var(\sum_{i=1}^{n} \omega_i X_i) = \sigma^2 \sum_{i=1}^{n} \omega_i^2 \to \min_{\omega_i},$$

subject to the constraint that $\sum_{i=1}^{n} \omega_i = 1$. The solution to this problem is $\omega_i = 1/n$, which gives the sample average as an answer. Thus we have the following statement: $\bar{X}$ is the best linear unbiased estimator or, in short, BLUE.

Hypothesis testing

Hypothesis testing answers the following question: given a sample, is there enough evidence to deny some assertions about a population? That is, we have some statement about a population (unknown distribution of the data), the validity of which we are trying to determine based on our sample.

Example 5. Assume we have an i.i.d. sample $X_1, ..., X_n$ from unknown distribution F, and the parameter of interest is $EX_i = \theta$. The null hypothesis (statement in question), $H_0 : \theta \in \Theta_0$ (where Θ_0 is a pre-specified set, say, $(-\infty, 8]$). The alternative hypothesis is the negation of the null, thus, $H_1 : \theta \notin \Theta_0$.

A test is a decision rule from a set of potential realizations of the sample $(x_1, ..., x_n)$ to 1 (reject) or 0 (accept). By making such a decision we will make mistakes from time to time, they are classified as being Type 1 or Type 2. That is, a Type 1 mistake is the one of rejecting the correct null hypothesis, while Type

	H_0 is true	H_1 is true
H_0 is accepted	correct	Type 2
H_0 is rejected	Type 1	correct

2 is the one of accepting the wrong null hypothesis. There is a trade-off between the two types of mistakes. Probability of a Type-1 mistake is called *Size*.

$$size(\theta_0) = P_{\theta_0}\{H_0 \text{ is rejected}\}, \text{ defined for } \theta_0 \in \Theta_0.$$

Sometimes people make a distinction between size and level, where the latter stay for composite null (case when more than one probability distribution belongs to the null)

$$level = 1 - \sup_{\theta \in \Theta_0} size(\theta_0).$$

Notice, that size (level) is calculated 'under the null', that is, assuming that the null hypothesis is correct. Probability of the same event, but calculated under the alternative is called power and corresponds to correct decision made while rejecting the null:

$$power(\theta) = P_\theta\{H_0 \text{ is rejected}\}, \text{ defined for } \theta \notin \Theta_0.$$

Power is connected to the Type-2 mistake:

$$power(\theta) = 1 - P_\theta\{Type\ 2\ mistake\}$$

Example 6. There was an election with two candidates: A received 42% of votes, B received 58%. Candidate A became convinced that election was rigged, so he hired an agency that randomly sampled 100 voters, 53 of whom said that they voted for A. Should A conclude that election fraud occurred?

Here we observe a random sample $X_1, ..., X_n$ of size $n = 100$ from Bernoulli with probability of 1 (voting for A), equal to unknown parameter θ. We wish to test

$$H_0 : \theta \leq 0.42 \qquad vs \qquad H_1 : \theta > 0.42.$$

We wish to use statistic $Y = \sum_{i=1}^{n} X_i$ (which in given realization equal to 53). The decision here is to reject if Y is above some cut-off c. The question is how to choose c? Assume the null (results of the election are correct and $\theta = 0.42$), then

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - 0.42\right) \approx N(0,1).$$

This statement is due to the Central Limit Theorem. If we want to get a test of size 5% (that is, reject when the null is true in at most 5% of cases), we should choose a cut-off which is high enough that only 5% of gaussian draws in the approximation above fall above the cut-off. The 95% quantile of the standard normal is 1.65. Thus, we should calculate quantity $\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - 0.42\right)$ and compare it to 1.65, and if it is above, then reject the null. In our case $\sigma^2 = \theta(1-\theta) = 0.42 \cdot 0.58$

$$z = \frac{0.53 - 0.42}{\sqrt{0.42 \cdot 0.58}}\sqrt{100} \approx 2.229 > 1.65.$$

Thus, we reject the null that the results of elections are correct. □

Above we created a test of asymptotic size 5%. The choice of the size is somewhat arbitrary- in social sciences we often use the 5% size, other commonly used significance levels are 90% and 99%. A good way to report results without attaching oneself to a specific significance level is to report the p-value.

Definition. P-value is the probability under the null of drawing a sample with the value of statistics at least as adverse to the null as the value computed from your sample.

Example 6 (continued). For the introduced above z-statistics, the values that are more adverse to the null are above 2.229. Thus,

$$p - value = P\{N(0,1) > 2.229\} = 0.011.$$

If p-value is less than 0.05, then we reject at 95% level.

Power. The test of 5% size we used is equivalent to rejecting whenever

$$\frac{1}{100}\sum_{i=1}^{100} X_i > 0.42 + 1.65\frac{\sqrt{0.42 \cdot 0.58}}{\sqrt{100}}.$$

Let us calculate the rejection probability if the true parameter $\theta > 0.42$:

$$power(\theta) = P\{\frac{1}{100}\sum_{i=1}^{100} X_i > 0.42 + 1.65\frac{\sqrt{0.42 \cdot 0.58}}{\sqrt{100}}\} =$$

$$= P\left\{\frac{\frac{1}{100}\sum_{i=1}^{100} X_i - \theta}{\sqrt{\theta(1-\theta)}}\sqrt{n} > \frac{0.42 - \theta}{\sqrt{\theta(1-\theta)}}\sqrt{n} + 1.65\frac{\sqrt{0.42 \cdot 0.58}}{\sqrt{\theta(1-\theta)}}\right\} =$$

$$= 1 - \Phi\left(\frac{0.42 - \theta}{\sqrt{\theta(1-\theta)}}\sqrt{n} + 1.65\frac{\sqrt{0.42 \cdot 0.58}}{\sqrt{\theta(1-\theta)}}\right).$$

Here we used the Central Limit Theorem again, but now assume that θ is the true value. One can draw the power curve and find that it is increasing in θ. One way of choosing the sample size is to require what

power one wants to achieve at a given value of θ (say $\theta = 0.5$) for the 5% test. Indeed, for $\theta = 0.5$ we have $\frac{0.42-\theta}{\sqrt{\theta(1-\theta)}} \approx -0.16$ and $\frac{\sqrt{0.42 \cdot 0.58}}{\sqrt{\theta(1-\theta)}} \approx 1$, then

$$power(0.5) \approx 1 - \Phi(-0.16\sqrt{n} + 1.65) \to 1 \quad (\text{ as } n \to \infty).$$

If one has the ability to create an experiment (choose the sample size), power consideration is a main guide. Unfortunately, in observational studies we are given a sample and often have no control over the sample size. □

Example 7. Does granting to a city the status of an enterprize zone have any effect on investment? Assume we have a sample of cities that has such a status, and the observed data is on $Y_i =$ 'percentage change in investment from the year before status to the year after status'. The hypothesis of interest is

$$H_0 : EY_i = 0 \qquad vs \qquad H_1 : EY_i \neq 0.$$

Let's assume that we are agnostic about the possible sign of the effect and want to test a two-sided alternative. Denote for simplicity $\mu_Y = EY_i$, the parameter of interest.

The idea of a test would be to look at $\frac{1}{n}\sum_{i=1}^n Y_i$ and compare it to 0. If the deviation from zero is large enough we would reject and say that found evidence of the mean being non-zero. The big question is how to measure large or small- we have a question of scale. By the Central Limit Theorem:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i - \mu_Y\right) \Rightarrow \sigma N(0,1),$$

where $\sigma^2 = Var(Y_i)$. Notice that in Example 6, we new the variance ($\theta(1-\theta)$), that allowed us to create the so-called $z-$ statistics. Here we do not know the variance, and thus, have to estimate it. As we know a good estimate of a variance is the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. By the Continuous Mapping Theorem:

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n Y_i - \mu_Y}{s_Y} \Rightarrow N(0,1).$$

For testing H_0 we create $t-$ statistics:

$$t = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n Y_i - 0}{s_Y},$$

and compare this to values typical for the standard normal distribution. In particular, given that we are not committed to a specific sign of effect, we reject large positive and large negative values symmetrically. Thus, for the 5% test we cut off 2.5% tail probabilities on both sides, which leads us to reject whenever $|t| > 1.96$. □

General case of testing using t-statistics. In general if we want to test a hypothesis about parameter θ in one of the following cases:

(1) $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$(two-sided);

(2) $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ (one-sided);

(3) $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$ (one-sided).

Assume that we have an estimate $\widehat{\theta}$ which is consistent and asymptotically gaussian, in particular:

$$\sqrt{n}(\widehat{\theta} - \theta) \Rightarrow N(0, \sigma^2).$$

Assume also that we have a consistent estimate of asymptotic variance $\widehat{\sigma}^2$ such that $\widehat{\sigma}^2 \to^p \sigma^2$. Then for testing we can create a t-statistic

$$t = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}/\sqrt{n}}.$$

Then under the null hypothesis (the most binding null hypothesis in case of one-sided) when $\theta = \theta_0$ we have $t \approx N(0,1)$. Notice that here $\frac{\widehat{\sigma}}{\sqrt{n}}$ is the scale, known as the standard errors. In the corresponding null/alternative combinations we will take the following actions:

(1) reject if $|t| > 1.96$; $p - value = 2\Phi(-|t|)$;

(2) reject if $t > 1.65$; $p - value = 1 - \Phi(t)$;

(3) reject if $t < -1.65$; $p - value = \Phi(t)$.

The standard way of reporting statistical estimation results is to report $\widehat{\theta}$ and the corresponding standard errors, so one can calculate the corresponding t-statistics for any hypothesis of interest.

Example 8. (Comparing means). Assume one has two independent data sets: one, $X_1, ..., X_n$, comes from F_X, and $Y_1, ..., Y_m$ is from F_Y. For example, X_i is the income of a college graduate, while Y_j is the income of a person without a college diploma. The hypothesis of interest is

$$H_0 : \mu_X = \mu_Y \qquad vs \qquad H_1 : \mu_X \neq \mu_Y,$$

where $\mu_X = EX_i, \mu_Y = EY_j$.

One suggestion is to formulate the problem in terms of one-dimensional parameter $\theta = \mu_X - \mu_Y$, then the corresponding hypothesis is $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. A natural estimate of θ is $\widehat{\theta} = \bar{X} - \bar{Y}$, consistent and asymptotically gaussian, the question is one of scale:

$$Var(\widehat{\theta}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

A natural guess for standard errors is $s.e. = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$. Indeed, one can show (try to sketch the argument yourself) that under the null (and assuming that both $n, m \to \infty$):

$$t = \frac{\widehat{\theta} - 0}{s.e.} \Rightarrow N(0,1).$$

For the hypothesis of interest we would reject the null if $|t| > 1.96$.

Confidence sets

The confidence set for parameter θ at a given significance level is the set of parameter values consistent with the data, that is, the set of parameters that cannot be rejected at the given significance level. The confidence set characterizes the uncertainty about a parameter value.

Assume again that we have a consistent estimate $\widehat{\theta}$ for θ and can produce standard errors $s.e.(\widehat{\theta})$ such that $\frac{\widehat{\theta} - \theta}{s.e.(\widehat{\theta})} \Rightarrow N(0,1)$.

If one wants to test a specific value θ_0, that is $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, then s/he will accept if

$$\left| \frac{\widehat{\theta} - \theta_0}{s.e.(\widehat{\theta})} \right| < 1.96,$$

then the set of θ_0 for which the corresponding null hypothesis will be accepted is $[\widehat{\theta} - 1.96 \cdot s.e.(\widehat{\theta}), \widehat{\theta} + 1.96 \cdot s.e.(\widehat{\theta})]$. This is a 95% significance level confidence set.