Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 6. Statistical estimates. (10 January)
Alla Tambovtseva

Populations vs samples: recall

A **population** is decribed by a random variable with a certain distribution with certain parameters. A **sample** is a set of observed data, values taken from a population (from a random variable with a certain distribution).

Statistical estimates and their distributions

A **statistical estimate** is an (approximate) value of a population parameter calculated based on a sample, on the data given. An estimate of a parameter θ is usually denoted as $\hat{\theta}$ (pronounced as θ hat). A hat here means *approximate*, *evaluated based on the data given.* For instance, μ is a population mean and $\hat{\mu}$ is a sample mean computed to approximate a population mean.

Example 1. A population has a distribution with the expected value μ. For instance, μ can be the mean value of Moscow residents' income that is not known for sure. To get an approximate value of μ, i.e. its estimate, we can take a large sample and calculate a simple average ($\bar{x}$), a weighted average ($\bar{x}_w$) or a sample median (Q_2). All these statistics can be used to evaluate the value of μ, the mean value of the population.

Example 2. The variance of a population distribution is the unknow value σ^2. We can estimate or approximate it by a sample variance s^2.

Example 3. All Moscow residents are either students or not. Let us encode students with 1 and others with 0. Hence, a population is just a set of ones and zeroes. It has a binary distribution with exactly one parameter $p = P(X = 1)$ that is the probability of meeting a student by chance. How to estimate this probability? We can take a random sample, ask people about their status and count the proportion of students in this sample. This proportion $\hat{p}$ is a good natural estimate of the probability $p = P(X = 1)$[1].

A statistical estimate itself is a random variable with its own distribution. An estimate is sometimes called a **statistics**. A distribution of an estimate is called a **sampling distribution** and shows how an estimate varies. A very natural question: why a sample statistics (an average, a sample variance, a sample median) varies? Usually it seems to be a fixed number. For instance, if we choose 20 students and record their marks for the essay, we can calculate the average mark and it will be a single number. Where does the variability come from? The answer is simple: we ask one group of 20 students, another researcher will ask another group of 20 students, one more researcher will choose one more different group of students, and so on. As these samples are different, their means are

[1] Provided that a sample is large enough and not biased.

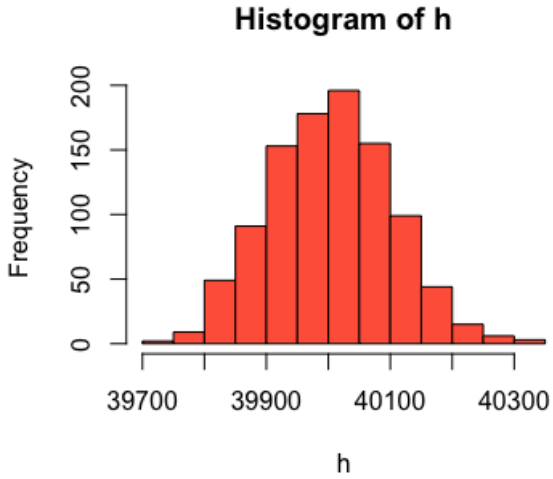
also different, so, we will get a set of different average values that has its own distribution. Let us consider an example.

Example 4. We have a population `pop`, then we take three different samples of size 1000 from this population and calculate their means in R.

```
sample1 <- sample(pop, size = 1000)
sample2 <- sample(pop, size = 1000)
sample3 <- sample(pop, size = 1000)

mean(sample1); mean(sample2); mean(sample3)
[1] 40172.7
[1] 39956.2
[1] 40243.9
```

Then we can take more samples, compute their averages and plot a histogram for them:



Statistical laws

Now we are ready to discuss two major statistical laws that lie behind many principles of statistical inference.

Note: in statistics large enough samples are samples of size $n \geqslant 30$. Most statistical laws work correctly when sample is large enough.

Law of Large Numbers

The more is a sample size n, the closer is a sample mean to a population mean.

Thus, to get a more accurate estimate of a population mean we should take a larger sample. This law sounds good, but it says nothing about the distribution of a sample mean and its dependence on the sample size. So, there is a more powerful theorem.

Central Limit Theorem

Consider a population that has a distribution with the mean value μ and the standard deviation σ. If we take all the possible samples of size n ($n \geqslant 30$) from this population, a set of sample means will approximately have a normal distribution with an expected value μ and a standard deviation $\frac{\sigma}{\sqrt{n}}$.

Why is this theorem important and powerful? It says that no matter which distribution a population has, a mean of a sample taken from this population will always have a normal distribution! And the parameters of this distribution are known as well. We will see an illustration for this theorem while doing practical tasks in R.

Example 5. Suppose a population has a binary distribution with $p = 0.6$. We take all possible samples of size $n = 100$ from this population and calculate their means. What distribution has a set of sample means?

According to the Central limit theorem, the sampling distribution of a mean is normal. How to find the parameters of this distribution, the expected value and the standard deviation? First, we have to calculate the population mean and its standard deviation. Let us write a distribution law for a binary variable with $p = 0.6$.

X	0	1
P	0.4	0.6

$$E(X) = 0 \times 0.4 + 1 \times 0.6 = 0.6$$

$$Var(X) = E(X^2) - [E(X)]^2 = 0.6 - 0.6^2 = 0.24$$

$$sd(X) = \sqrt{0.24} \approx 0.5$$

So, $\mu = 0.6$ and $\sigma = 0.5$. Now we can calculate the expected value and the standard deviation of a sample mean. By theorem, the expected value of a sample mean coincides with the population expected value μ. The standard deviation is $\frac{\sigma}{\sqrt{n}}$, so it is $\frac{0.5}{\sqrt{100}} = 0.05$. Finally, a sample mean has a distribution $N(0.6, 0.05)$.

Properties of estimates

These properties are not widely used in practice (I mean their strict definitions and corresponding proofs), but you can come across them reading about different methods of estimation.

- **Robustness**

  An estimate is called *robust* if it is not affected by outliers.

  Robust estimates: quartiles (including a median).

- **Unbiasedness**

  An estimate is called *unbiased* if its expected value equals the true value of an estimated parameter, so $E(\hat{\theta}) = \theta$.

  Unbiased estimates: a sample mean, a sample proportion, a sample variance (unbiased estimate with $n - 1$ in the denominator).

- **Efficiency**

  An estimate is called *efficient* if it has the smallest possible variance.

  For example, a population mean can be estimated using a sample mean, a weighted mean and a sample median. All this statistics are good enough, all of them have their own distributions, but the variance of a sample mean is the smallest. Thus, it is an effective estimate of a population mean.

  Notes: 1) efficiency is defined only for unbiased estimates; 2) efficient estimates are always non-robust.

  Efficient estimates: a sample mean (an average), a sample variance.

- **Consistency**

  An estimate is called *consistent* if it goes to a true value of a parameter by a probability as a sample size goes to infinity, so $\hat{\theta} \to^p \theta$ as $n \to \infty$.

  Consistent estimates: most known sample statistics.