Mathematical statistics
Week 2/a: Populations, samples and estimators

Péter Elek and Ádám Reiff

24th September, 2013

1. Populations and samples

Statistical inference.

- Statistical inference: to learn something about a "population" given a sample from that population.
- e.g. point estimation, interval estimation, hypothesis testing
- Close relationship with probability theory

Population.

- Population: a well-defined group of subjects (e.g. individuals, firms)
  - We are interested in some characteristic of this population
- More formally: the population is the distribution of an observation from a random experiment.
- Can be infinite $\longrightarrow$ a continuous pdf given
  - Example: body height of Hungarian males is normally distributed with $\mu = 180$ cm and $\sigma = 12$ cm.
- Can be finite $\longrightarrow$ the list of possible values given (with a discrete pdf)
  - Example: all CEU students, possible outcomes of a dice throw

Sample: set of observations drawn from the population.

- The underlying population pdf (which we would like to draw inference from) is called the true/actual pdf.
- i.i.d. (independent, identically distributed) sample of size $n : X_1, X_2, \ldots, X_n$ observations in the sample are independent from each other, and have the same true distribution $f(x)$.
  - Then, the joint distribution is given by $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.
  - Examples: i.i.d. sample from an infinite population, sampling *with replacement* from a finite population

Sample, continued.

- Another type of sampling: sampling *without replacement* from a finite population.
  - Then, the sample elements are identically distributed, but of course not independent!
  - This is called simple random sampling.
- (There are other types of random samples and also non-random samples, we do not deal with these.)
- Example: we throw a dice 100 times. This is a sample with $n = 100$ elements.
  - What is the population? What is the true probability distribution? Is this an i.i.d. random sample?

True population pdf.

- Parametric statistics (covered in this course): unknown true pdf but its type is assumed
  - Examples: Bernoulli with some p (e.g. Democrat or Republican); normal with some (μ, σ) (e.g. body height)
  - We use the notation $f(x, \theta)$ for the pdf (where θ is the unknown parameter).
- Nonparametric statistics: even the type of the pdf is unknown.

2. Parameters, sample statistics, estimators

Parameter.

- Parameter: any unknown constant of the true population pdf
- Example: p in Bernoulli, μ, σ in normal, λ in Poisson, θ in exponential
- Generally we draw samples exactly because we would like to infer the values of the parameters

Sample statistics.

- Sample statistic: a numerical descriptive measure *calculated from the sample*
- Any function of sample elements $x_1, x_2, \ldots, x_n$ is a sample statistic
- Example 1: sample mean: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Example 2: sample median: $m = Med(x_1, x_2, \ldots, x_n)$
- Example 3: sample variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$
- Note the difference: parameter refers to the true distribution (or population), but the sample statistic is calculated from sample.

The sample statistic is a random variable!.

- each element of the sample x_i is a random variable
- the sample statistic is a function of sample elements
- $\Longrightarrow$ the sample statistic itself is a random variable!
- We can highlight this with the following notation: $W = T(X_1, X_2, \ldots, X_n)$.
  - In case of an actual realization, the notation may be $w = T(x_1, x_2, \ldots, x_n)$.
- (in contrast, parameters are numbers)

Estimators and estimates.

- As we said: we draw samples because we want to infer (or estimate) the values of the parameters in the true probability distribution.
- A *point estimator* is any function of the random sample (any *sample statistic*), which we use to estimate a parameter: $W = h(X_1, X_2, \ldots, X_n)$.

- A *point estimate* is obtained when a particular set of numbers is plugged into the function: $w = h(x_1, x_2, \ldots, x_n)$.
- Examples:
  - Sample mean as an estimator of the expected value
  - Sample median can be another estimator of the expected value or median
  - Sample variance: estimator of the variance

Properties of estimators.

- How good are these estimators?
- Is the sample mean or sample median the "better" estimator for the expected value of the population distribution?
- These depend on the properties of the estimators.
- Properties:
  - Finite sample properties (i.e. properties for fixed sample size n)
  - Asymptotic properties (i.e. properties when sample size $n \to \infty$)

3. Sampling distributions

Sampling distributions.

- *Sampling distribution* of a sample statistic (or estimator): the probability distribution of the sample statistic
- Example: draw a sample of $n = 10$ elements by throwing a dice 10 times, and calculate the sample mean.
- Then repeat this 1000 times.
- Then for the 1000 realizations, the sample means will not be identical, but there will be a distribution of sample means.
- This will be the observed sampling distribution of the sample mean.

Derivation of sampling distributions.

- True/theoretical/exact sampling distribution of a sample statistic: the true (and NOT the observed) sampling distribution of a sample statistic
- In certain cases we can derive it analytically, in other cases we can simulate it.
- Simulation:
  - Draw "infinitely many" samples
  - Calculate the sample statistics for all of these
  - The observed sampling distribution can be regarded as (and in fact is very close to) the true sampling distribution – there is a mathematical theorem saying this.

Example.

- Example: the population consists of three elements $\{0, 3, 12\}$, each with probability $p = \frac{1}{3}$, and we draw samples of $n = 3$ elements from this population.
- What is the true probability distribution of the population?
- List all the possible samples that we could draw, and calculate the sample mean and sample median for all of these!
- Find the true sampling distribution of the sample mean and the sample median!
- Calculate the expected value of the sampling distributions of the sample mean and sample median!