Introduction to Statistical Inference

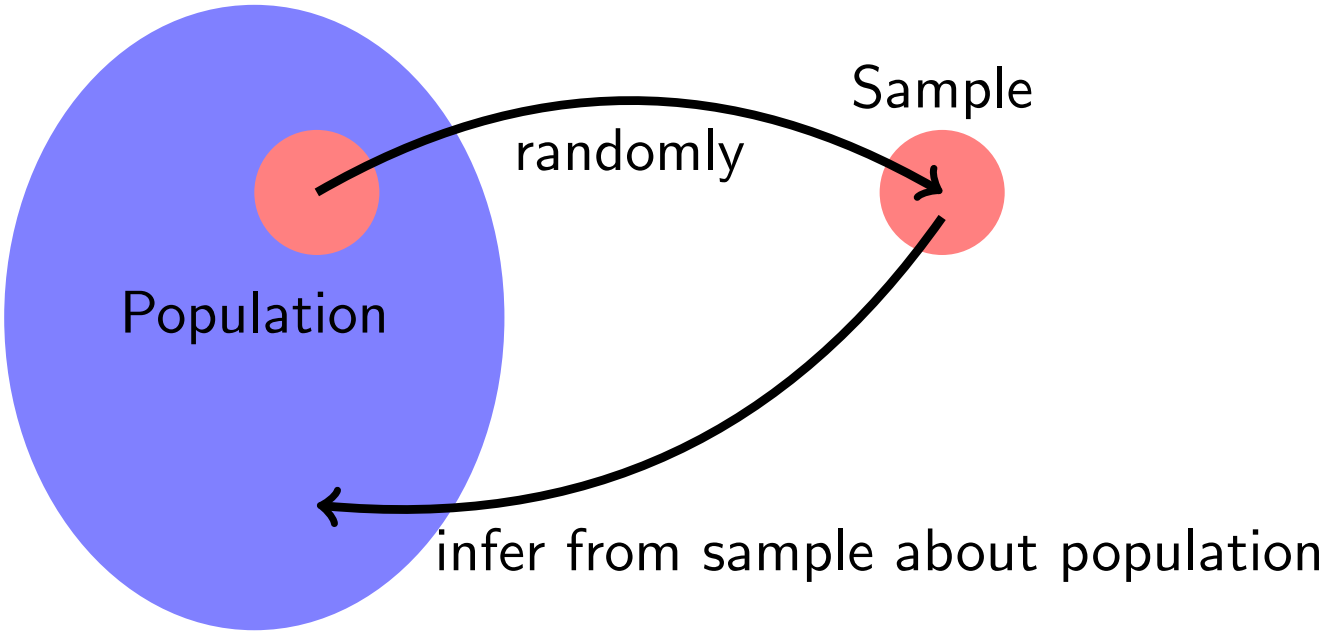
Josemari Sarasola

Statistics for Business

In statistics, we usually seek to know about populations, as a collective and a feature about it (e.g., height of 18 aged people) or as a variable of interest (e.g, daily production in a factory). As that feature is variable, we can take it as a random value, and so we assign that feature a probability distribution, with some unknown parameters (a mean or a proportion, e.g). That probability distribution is a simplified representation of the population, so we also call it a model. Along the next explanations, population, distribution and model will be (almost) interchangeable terms.

Introduction to Statistical Inference

Most times, it's not possible to take data about all elements in a population (too expensive or cannot list all the elements in a population), so we take a sample to get information (infer) about the population. Samples must be *random* in order to be representative about the population.



Known the exact (with exact parameters) probability distribution or model for a population, we can solve many practical problems about it, as we have seen in the previous lessons about concrete probability distributions (Poisson, uniform, exponential, binomial, ...).

But these questions arise right away:

1. How do we set a probability distribution for a given population?
2. How do we quantify the parameters for that distribution?

How do we set a probability distribution for a given population?

At the beginning of the inference process, generic models or distributions can be roughly assumed for a population. E.g., we can assume that sales follow a normal distribution, when after plotting the data, we see a rough bell shaped symmetric curve. Don't worry: these models will be tested at the end of the process.

How do we quantify the parameters for that population (more exactly the model set for that population?

The parameters of the model or distribution must be inferred or quantified from data.

Main problem

So, the main problem in **statistical inference** is inferring the parameters of a population defined by means of a random model or probability distribution from a sample or subset of data taken from the whole population. At the end, we will test the overall model (generic model plus inferred parameters).

Steps in inference process

Hence, briefly, these are the main steps in inference process:

1. Drawing a random sample from the population.
2. Choosing and assuming a suitable model for the population.
3. Think about how to quantify (infer) parameter values from sample: we must choose estimators.
4. Apply estimators and consequently quantify parameters.
5. Validate model+quantified parameters and other assumptions.

First step: drawing the sample

The sample must be random to be representative about the population. We will test at the end that the sample is really random. We must take samples because analyzing all elements in a population is difficult, expensive or because we cannot list all the elements (we call that an infinite population).

- Infinite populations: no. of customers entering a shop every hour, daily maximum tenperatures in March.
- Finite populations: students in Faculty of Economics, families in Gipuzkoa.

We will assume that our populations are infinite. Inference in finite populations must be learnt apart.

Second step: model choosing

Taken the sample, we must set a distribution or model for those data:

- looking at the histogram or other kind of plot for data (flat histogram $\rightarrow$ uniform distribution)
- looking at the nature of data: customers arrivals are usually random and independent, so we can take for those a Poisson model.

Third step: applying an estimator to data

- To estimate or quantify the parameters we set an estimator. An estimator is just a formula applying to data, that is calculated to approximate the value of a given parameter. For example, the aritmethic mean, or the biggest data.
- Generally, we denote a parameter by θ or other greek letter, an an estimator for that parameter as $\hat{\theta}$.
- For example, to estimate μ, the population mean, we usually apply $\hat{\mu} = \overline{x}$, the sample mean. That kind of intuitive estimators are called *natural estimators*.

Fourth step: quantifying parameters

Having calculated the estimator, we have two ways to quantify the unknown parameters:

- we may take the result in the estimator directly as an estimation for the parameter, that is to say, to make a point estimation; for example: $\hat{u} = \overline{x} = 4.5$.
- and we may also take that result as a basis to perform a statistical test (for example, $H_0 : \mu = 4$, taking as evidence: $\overline{x} = 5$).

Differences between estimators and parameters

Parameters	Estimators
Notation: θ	Notation: $\hat{\theta}$ (θ's estimator)
Corresponds to population	Corresponds to sample
Constants	Changing form one sample to other
Usually unknown	Calculated from data
θ unique	several $\hat{\theta}$ available
E.g.: μ (population mean)	E.g.: $\hat{\mu}_1 = \overline{x}$, $\hat{\mu}_2 = Me$

Fifth step: validation

Some assumptions are usually made when we apply classical statistical inference:

- data are really random, and don't show a tendency;
- data are fit for the assumed model at the beginning of the inference process (e.g. Poisson, uniform, ..); that is we have to validate the *goodness of fit*.
- data are homogenous, from an unique distribution or population (e.g., when we assume male and female data have the same features), so we can put them all together in one sample.

So we must validate (1) randomness (that is, independence) (2) model+quantified parameters, and (3) homogeneity.

Validation

Test for randomness (independence): Wald-Wolfowitz runs test

- We can apply this test to dichotomous and quantitative variables.
- A run is a consecutive sequence of data with the same value.
- When data are quantitative, a run denotes a sequence of data below or above the median.
- Testing statistic is the number of runs: e.g., into the XX0XX000XX sequence number of runs is $R = 5$.
- Very important!: runs must be counted in the order the data have been collected.

Validation

Test for randomness: Wald-Wolfowitz runs test

- XXXXXOOOOO: R=2(few) $\rightarrow$ no random $\rightarrow$ dependence
- XOXOXOXOXO: R=10(many) $\rightarrow$ no random $\rightarrow$ dependence
- XXOOOXOXXO: R=6 (neither few, nor many) $\rightarrow$ data are random, and therefore independence.

Hence, [H_0:randomness/independence] is rejected when no. of runs is big or small enough (hence, it's a two-sided test).

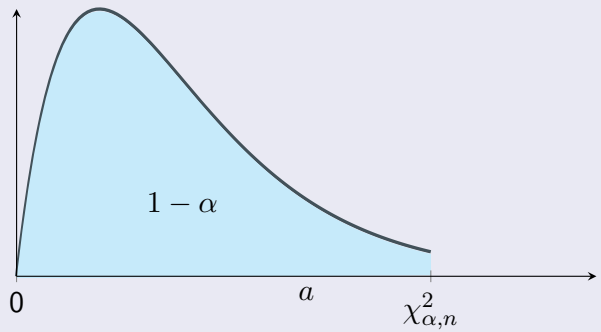
Test for randomness: Wald-Wolfowitz runs test

- Critical values are tabulated for small samples. We reject randomness when R no. of runs is equal or larger than the upwards critical value, or equal or smaller than the downwards critical value.
- For big samples, runs distributes in this manner under H_0:

$$R \sim N\left(\mu = \frac{2n_1n_2}{n_1+n_2} + 1, \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}\right)$$

Test for goodness of fit: chi-square test

Applying chi-square test we will use a new distribution: χ^2_n, named chi-square, with only one parameter: n, named degrees of freedom, taking only integer positive numbers. It's like this:



Chi-square values are tabulated, for given values of $1-\alpha$ probabilities below. E.g.,

- $\chi^2_{0.01,4} = 13.3$
- $\chi^2_{0.25,2} = 2.77$

Test for goodness of fit: chi-square test

- H_0: model is fit or OK for data.
- We calculate observed (O_i) and expected (E_i) frequencies, the latter from theoretical probabilities..
- Calculate $X^2 = \frac{(O_i - E_i)^2}{E_i}$ statistic (a statistics is a formula from data; all estimators are statistics, but not all statistics are used as estimators for parameters).
- X^2 being very big means that observed and expected frequencies are very different, and hence we should reject the assumed model. Hence, chi-square test is one-tailed and the critical region in on the upper side.
- To perform the test, we compare X^2 statistic to the critical value:
  - to $\chi^2_{\alpha,k-1}$ value, k being number of differents values or intervals for data; or,
  - when some parameters are estimated in the assumed model, to $\chi^2_{\alpha,k-e-1}$ value, e being the number of estimated parameters.

Example

We flip a coin and we get 86Xs and 114 Os. Do we have a balanced coin? Significance-level: 10%.
Model: p(o)=p(x)=0.5

Outcomes	Observed (O)	Prob.	Expected (E)	$\frac{(O-E)^2}{E}$
o	86	0.5	$0.5 \times 200 = 100$	1.96
x	114	0.5	$0.5 \times 200 = 100$	1.96
	200		200	$\mathbf{X}^2 = 3.92$

In a chi-square distribution with 2-1=1 degree of freedom, the value leaving above it a probability of 0.1 is 2.71. So, the value of the statistic (the distance observed/expected) is significative, so we reject the model and claim that the coin is not balanced.

Test for homogeneity: Wilcoxon rank sum test

- For quantitative data, but distinguishable about a dichotomous feature (e.g., califications for some men and women)
- H_0: feature (sex) doesn't have influence, that is, homogeneity (hence, all califications may be taken as a unique sample).
- Sort all data from the smallest to the biggest.
- Calculate ranks, distinguishing about the dichotomous feature.
- Calculate W rank sums about both categories in the feature.
- Take smallest W as testing statistic: W_{min}.

Test for homogeneity: Wilcoxon rank sum test

- Very small values for W_{min} statistic mean that both subsets of data are different.
- Test is two-tailed because W_{min} may correspond to either of the categories.
- Critical values are tabulated, for different numbers of data in both categories.
- For big sample sizes, W_1 statistics distributes like this, n_1 being the sample size for the category no. 1:

$$W_1 \sim N\left(\mu = \frac{n_1(n_1+n_2+1)}{2},\ \sigma = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right)$$