Introduction to Statistical Inference

Josemari Sarasola

Statistics for Business

In statistics, we usually seek to know about populations (e.g., 18 aged people in a country), and more concretely about some variable feature among them (e.g, height). As that feature is variable, we can take it as a random value, and so we assign that feature a probability distribution, with some parameters. That probability distribution is a simplified representation of the population, so we also call it a model. Along the next explanations, *population, and distribution and model* for that population will be (almost) interchangeable terms.

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 have an idea about the population. Samples must be *random* in order to be representative about the population.

Introduction to Statistical Inference

Known the exact (with exact parameters) probability distribution 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 this question arises right away: How do we set a probability distribution for a given population? How do we quantify the parameters for that distribution?

- Answer for the 1st question: At the beginning, generic models or distributions can be assumed for a population. E.g., we can assume sales follow a normal distribution, when plotted the data, we see a bell shaped symmetric curve.
- Answer for the 2nd question: But generally we cannot assume given values for the parameters of an assumed distribution. The parameters must be inferred or quantified from data.

Main problem

So, the main problem in **statistical inference** is inferring the parameters of a population that follows a given distribution from a sample or subsets of data from the the whole population.

First step: drawing the sample

The sample must be random to be representative about the population. 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).

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, $H_0 : \mu = 4$, 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 statistical inference:

- data are drawn randomly;
- 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 must validate (1) randomness (2) model, and (3) homogeneity.

Validation

Test for randomness: 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 $R = 5$.

Validation

Test for randomness: Wald-Wolfowitz runs test

- XXXXXOOOOO: R=2(scarce) $\rightarrow$
  no randomness or indep.
- XOXOXOXOXO: R=10 (much) $\rightarrow$
  no randomess or indep
- XXOOOXOXXO: R=6 (neither scarce, nor much) $\rightarrow$
  randomness, and therefore independence.

Hence, [H_0:randomness/independence] is rejected we runs number is big or small enough.

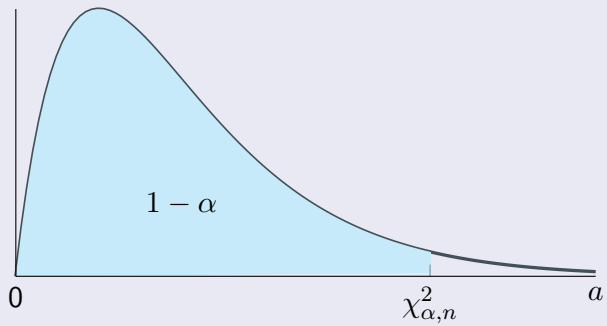
Test for randomness: Wald-Wolfowitz runs test

- Data must be taken always in the order we collected them.
- Test is two-sided, as we reject H_0 when R is very big and very small.
- Critical values are tabulated for small samples.
- 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 or index.
- 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 comnpare X^2 statics 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.

Test for homogeneity: Wilcoxon rank sum test

- For quantitative data, but distinguishable about a dichotomous feature.
- H_0: feature doesn't have influence, that is, homogeneity
- Sort 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:

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