THEOMATICS

Author:

Del Washburn

Portland, Oregon
USA

Rendered in German,
summarized and testetd by

Kurt Fettelschoss

July 2004

File: theomatics01

11.1 A short trip into statistics theory (1)

P Definitions:

GG	population
SP	sample
N, n	size of the GG, SP (= quantity of elements in the GG, SP)
X, x	random variables of the GG, SP
p_0, p	expected (hypothecial) or basic probability, probability

P In general, statistics deals with the quantity of features of certain elements of a defined quantity (= GG) and with the distribution of these features. The foundamental principles of statistics come from the theory of sets.

P In most cases the population is such that it cannot be analyzed with reasonable effort. The analysis must then be restricted to a partial quantity of the population in a form of a sample.

P There are certain requirements for the elements of a sample, in order to draw the right conclusions for the GG from the sample results:

- The sample elements must be selectable with the same probability for each element in a random selection process (equal probability model, simple sample).
- The random variables must be randomly independent from each other.
- The defined features must be exclusive (not accumulative).

P Due to the random nature of selection the features of the sample elements do not have absolute values. They are random variables with assigned probabilities.

P Therefore, a result derived from a sample is no absolute result, there is always an error probability tied to it. A central task of statistics is to make the error predictable. A basic element for this calculation is the probability distribution of the sample elements compared to the frequency distribution in the population.

P Distributions (here: of discrete random variables) are in general be characterized by the following parameters:

- mean value (Expectation): $\mu = E(X)$
- deviation from mean value (variance = standard deviation^2): $\sigma^2 = Var(X) = E((X - \mu)^2)$
- Sample (in case of same distribution for all x): $E(x) = \mu$ $var(x) = Var(X)/n$

P There are 2 sample models for frequency distributions:

- Sample “**without** putting back” the elements into the population, i.e. the population changes with every selection of an element out of the population. No double counts are allowed. The basic probabilty for an element to be selected for the sample is given by the selection possibilities and is $p = 1/\text{binomial coefficient } (N \text{ over } n) = 1/(N!/(n!*(N-n)!))$. The probability distribution follows a hypergeometric distribution (e.g. Lotto 6 out of 49).
- Sample “**with** putting back” the elements into the population, i.e. the population remains unchanged with every selection of an element out of the population. Double counts are allowed. The basic probability for an element to be selected for the sample is given by the selection possibilities and is $p = 1/N$. The probability distribution follows a binomial distribution.

11.1 A short trip into statistics theory (2)

P Relating to a sample, the important parameters for the calculation of probabilities are the sample size “n” and the quantity of the observed features of the random variable “x”.

P Both the hypergeometric and the binomial distribution are difficult to handle because of the high number values which can result from the factorials “N!” or “n!”. Therefore, in practice, these distributions will often be approximated as follows:

- ▸ The hypergeometric by the binomial distribution (for $n/N \leq 0{,}05$; as a rule of thumb)
- ▸ The binomial by the normal distribution; for large values of “n” and small values of “p” the exactness of the results is normally not satisfying. For the following combination of parameters, the observed absolute errors will be less than 0.01:
  - – $n \geq 30$, $0.38 \leq p \leq 0.62$
  - – $n \geq 100$, $0.28 \leq p \leq 0.72$
  - – $n \geq 200$, $0.22 \leq p \leq 0.78$
  - – $n \geq 300$, $0.15 \leq p \leq 0.85$
- ▸ The binomial by the Poisson distribution for small values of “n” and especially for small values of “p”:
  - – $p \leq 0.031$ with the expectation value $E(X) = \mu = n*p$
  - – $p \geq 0.969$ with the expectation value $E(X) = \mu = n*(1-p)$
- ▸ The Poisson distribution by the normal distribution for large values of “μ”. The maximum error of approximation is smaller than 0.01, in case $\mu \geq 100$.

P If there is a variance existing in the population, then the sample mean values follow approximately a normal distribution for large values of “n” (central limit theorem of probability calculation). The normal distribution can be brought into a standardized form with $E(X^*) = 0$ and $Var(X^*) = 1$ for a random variable X^* as follows:
$X^* = (X - \mu)/\sigma$.

P The probability for the occurance of certain features can be calculated from one of the above mentioned probability distributions.

P The probability distribution of a random variable in a sample normally differs from the probability distribution of the population, because of the nature of randomness. From the amount of the deviation it can be checked with a certain error probability whether the deviation between sample and population is significant or not. This requires a suitable check measure.

P The testing of a statistical hypothesis is based upon this consideration. In such a test it is checked, whether a predefined numerical value of a parameter (= hypothesis) is confirmed with high probability by the sample result or not.

P The testing of a statistical hypothesis is generally performed with “rejection levels or areas”. Therefore, it makes sense to claim in a test hypothesis H_0 the opposite of what should be proven (hypothesis H_1), and then to reject the test hypothesis H_0 with a certain error probability.

Author:

Del Washburn

Portland, Oregon
USA

Rendered in German,
summarized and testetd by

Kurt Fettelschoss

July 2004

File: theomatics01

THEOMATICS

Author:

Del Washburn

Portland, Oregon
USA

Rendered in German, summarized and testetd by

Kurt Fettelschoss

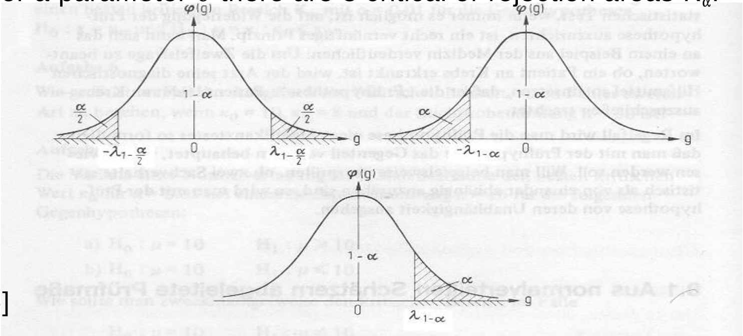
July 2004

File: theomatics01

11.1 A short trip into statistics theory (3)

- A statistical test which is based on the rejection of a test hypothesis is called a **significance level test**. For the performance of a significance level test it is essential to identify a critical significance level “α” which is suitable for the test hypothesis and a given check measure.
- A suitable check measure “z” for a normal distributed random variable is e.g. the deviation from the mean value as follows: $z = (\mu_{SP} - \mu_{GG})/\sigma_{SP}$
  The standard deviation σ_{SP} or the variance of the sample can be determined form the variance of the population as (see above): $\sigma_{SP}^2 = \sigma_{GG}^2/n$.
  If the numerical value of the check measure “z” is beyond the rejection level, which is defined by the significance level “α”, then the test hypothesis can be rejected with the error probability “α” (see the following).
  Note: In case the variance of the population is not known, it can be estimated out of the sample. This leads to a check measure “t”, which is defined similar as “z”, and which follows a t or student distribution with “ν” degrees of freedom. The degrees of freedom result from the sample size “n” reduced by the quantity of estimated parameters. For large values of “n” the t distribution can be approximated by the normal distribution ($\nu > 120$).
- There are two error possibilities in the testing of a test hypothesis H_0:
  1. The test hypothesis H_0 is correct:
     A. H_0 is accepted
     B. H_0 is rejected ⇒ alpha error / error probability **“α”** (rejection level, significance level)
  2. The test hypothesis H_0 is not correct:
     A. H_0 is accepted ⇒ beta error / error probability **“β”**
     B. H_0 is rejected
- The error probabilities α and β are dependent on each other. A small value for α means a large value for β and vice versa. In a significance level test α is less critical than β. In practice α is normally given as e.g. $\alpha = 0.05$ or $\alpha = 0.01$ and 1-β is determined as a measure for the test quality with a value as high as possible for 1-β. The aim is to minimize the error for the acceptance of the test hypothesis in case the test hypothesis is not correct.
- Dependent on the test hypothesis H_0 for a parameter θ there are 3 critical or rejection areas K_α:
  - 1. H_0: $\theta = \theta_0$ K_α: $|g| > \lambda_{1-\alpha/2}$
  - 2. H_0: $\theta > \theta_0$ K_α: $g < \lambda_\alpha = -\lambda_{1-\alpha}$
  - 3. H_0: $\theta < \theta_0$ K_α: $g > \lambda_{1-\alpha}$

  1. double sided test
  2. single sided test, lefthand
  3. single sided test, righthand



Reference: [5]

THEOMATICS

11.1 A short trip into statistics theory (4)

P Sample (related to the testing of theomatics):
The parameter p of a binomial distribution shall be $p = p_0$ with p_0 [0,1]. The observed value in a sample is "p_{SP}". The hypothesis $H_0 : p = p_0$ shall be tested. The check measure **z = (p_{SP} -p_0)/square root((p_0 - (1 -p_0))*square root (n)** follows a standardized normal distribution for sufficiently large values of "n".
The quality of the approximation can be checked by the approximate definition of the acceptance levels as follows: p_0 - $\lambda_{1-\alpha/2}$*square root(p_0*(1-p_0)/n) $\leq p_{SP} \leq p_0$ + $\lambda_{1-\alpha/2}$*square root(p_0*(1-p_0)/n). The exact limits can be taken out of a diagram for the confidence limits of the binomial distribution.
In case a test hypothesis is to be rejected, the best test quality for a given significance level "α" will be achieved for those observed values of p_{SP}, which show the largest distance from the acceptance area.

P Some typical probabilities for falling below the critical areas of a standardized normal distribution are as follows:

α:	0.050	0.025	0.010	0.005	0.001
1-α:	0.950	0.975	0.990	0.995	0.999
$\lambda_{1-\alpha}$:	1.6449	1.9600	2.3263	2.5758	3.0902

P A statistical test dealing with a hypothesis concerning the shape of a distribution is called a **matching test**. A common matching test is the **χ^2-Test**. By applying the χ^2-Test it can be tested, whether the sample results belong to a population which shows a distribution in accordance with the test hypothesis.

P In case of one feature only, the check measure is: $\chi^2 = \Sigma((x_i\text{-}n{*}p_i)^2/n{*}p_i)$. With $\chi^2 = 0$ both distributions are identical. For testing a frequency distribution it is assumed, that there are "i" frequency categories defined which are mutually exclusive. I.e. it is assumed that each element belongs to one of these frequency categories only with respect to the observed feature. Theomatics deals with i = 5 hit / feature categories (0 / -1 / 1 / -2 / 2).

P For a sufficiently acceptable approximation to the χ^2 distribution the observed quantity in each feature category must be $x_i \geq 10$ (rule of thumb). If this is not the case, the feature categories must be summarized in a suitable way.

P The acceptance level for the check measure χ^2 is determined by $P(\chi^2 \leq \chi^2_{1-\alpha;\upsilon}$ I $H_0) \doteq 1 -\alpha$, i.e. the test hypothesis has to be rejected with an error probability "α" if $\chi^2 > \chi^2_{1-\alpha;\nu}$. The parameter "$\upsilon$" stands for the degrees of freedom of the check measure (υ = quantity of feature categories "i" - 1 - quantity of estimated parameters). Critical values / limits for the check measure $\chi^2_{1-\alpha;\upsilon}$ are contained in tables for the χ^2 distribution or can be calculated approximately for large degrees of freedom ($\upsilon > 100$). Some typical values are:
υ **= 1**: $\alpha = 0.05$: $\chi^2_{0.95;1} = 3.84$, $\alpha = 0.01$: $\chi^2_{0.99;1} = 6.63$ / υ **= 2**: $\alpha = 0.05$: $\chi^2_{0.95;2} = 5.99$, $\alpha = 0.01$: $\chi^2_{0.99;2} = 9.21$

P Finally, there are the following statistical tests for testing the observed phenomenons of **Theomatics**:
- Significance level test to check, whether the observed quantity of theomatic multiples differs significantly from a quantity which could be expected from randomness. I.e. double sided test in order to reject the test hypothesis H_0, that the hits observed in the sample do not differ from the expected quantities from randomness, i.e.
  H_0: $\theta = \theta_0$ and K_α: $|z| > \lambda_{1-\alpha/2}$
- Matching test (χ^2 test) to check, whether the observed clustering hit distribution differs significantly from a frequency distribution, that could be expected from randomness (see above).

In both cases the random variable is the frequency of the theomatc features.

Author:

Del Washburn

Portland, Oregon
USA

Rendered in German,
summarized and testetd by

Kurt Fettelschoss

July 2004

File: theomatics01