Jan Długosz University of Częstochowa
Scientific Issues, Mathematics XII, Częstochowa 2007

MATHEMATICAL STATISTICS AND MATHEMATICAL DIDACTICS

Michal Folćan

Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava, Slovak Republic
e-mail: Foltan@yahoo.com

In standings related with didactical experiments we have often need to use statistical check of our hypothesis. Though we consider that our method is correct or that some other method is wrong, we cannot make any claim whiteout mathematical background. Statistical proving of didactical hypothesis enables as to put our standings on mathematical standings. A most of didactical theories in some of their part use statistical proving.

Mathematical statistics enable as this goal to fulfill. Problem on which we could come upon when we start to chose statistical method which is most adequate for our needs, or our experiment needs, is fulfilling conditions which must be satisfied before we use some method. One of the most often conditions is condition than our sample is from normal distribution. This problem can be solved with central limit theorem of mathematical statistics.

In this work I would like to make more observation about cases where is not proper to use central limits theorem. This could be happening from several reasons: maybe sample that we have in not enough big, maybe we have interest, or our didactical experiment demand, knowing preciously to which distribution function belong our sample.

One of the very useful methods is Pearson's χ^2 test of congruence. Benefits of using this test are that this test can be applicable on every distribution function, which make our job much more easier. Other thing is that correspondent statistics are relatively easily countable, actually we have relatively easy calculation. But first take a look at method of maximal likelihood.

1. Method of maximal likelihood

We can estimate unknown parameter, with big size sample, of the numeric value distribution. Procedure of estimating unknown parameter depends on real or asymptotic (limiting) statistics. These statistics are functions of the sample.

Let's assume that theoretical distribution of numeric value is distribution function F, which belongs to set $P = \{F(x,v) : v \in \Theta\}$ acceptable of permitted distributional functions and let sample $(X_1, X_2, \dots, X_n)$ be from distribution F. We can get point estimator of the unknown parameter v from sample $(X_1, X_2, \dots, X_n)$ by following steps: we have to select statistics $T_n = T(X_1, X_2, \dots, X_n)$ and they name is estimator of the unknown parameter v. If realized value of the sample is $(x_1, x_2, \dots, x_n)$, then for approximation of the number v number $T(X_1, X_2, \dots, X_n)$ is taken. There are several methods how to get estimator, but basic is method of exchange. If unknown parameter v can be presented with functional $v = G(F)$, and we can mark with F_n empirical distributional function defined according to sample $(X_1, X_2, \dots, X_n)$. Estimator of the unknown parameter v is statistics $T_n = G(F_n)$.

If we want to have the most optimal properties we need to use the method of maximal veracity.

Let take a look at distribution $L(X)$ of numeric value X, which belongs to the set $P = \{f(x,v) | v \in \Theta\}$ of acceptable distribution. Veracity function is $L(x_1, x_2, \dots, x_n; v) = \prod_{k=1}^{n} f((x_k, v)$, when v is fixed this is density of random vector X. It can be observed that realized values of random vector X are mostly where veracity function take big values. Therefore we take statistics $\hat{v}_n = \hat{v}_n(X)$, for unknown parameter estimator. $\hat{v}_n = \hat{v}_n(X)$ is defined with condition

$$L(X : \hat{v}_n) = \max L(X; v).$$

Because function $\ln : R^+ \to R$ is increasing function $L(x,v)$ and $\ln L(x,v)$ have the same maximal value, we can make maximal veracity estimator from equation

$$\frac{\partial \ln L(X; v)}{\partial v} = 0.$$

Theorem. Let T_n is efficient* estimator of parameter v. Then T_n is the only solution of veracity equation.

Example 1. $(X_1, X_2, \dots, X_n)$ is a sample from $B(1,v)$ distribution. Let's find maximal veracity estimator of unknown parameter v. Distribution of numeric value is defined by $f(x,v) = v^x(1-v)^{1-x}$, where $x \in \{0,1\}$. Function of veracity is

$$L(x_1, x_2, \dots, x_n; v) = v^{x_1 + \dots x_n}(1-v)^{n - x_1 - \dots - x_n}.$$

*Estimator is effecient if stands: $DT_n = G$, where $G = \frac{1}{nE\left(\frac{\partial}{\partial v}\ln(f(x,v))\right)^2}$.

Equation of veracity is

$$\frac{\partial \ln L}{\partial v} = \frac{\partial}{\partial v}(x_n \ln v + (n - s_n)\ln(1-v)) = \frac{s_n}{v} - \frac{n - s_n}{1 - v} = 0,$$

where $s_n = x_1 + \ldots + x_n$. Solution of this equation is $v_n = \frac{s_n}{n} = \overline{x}_n$. Maximal veracity estimator of the unknown parameter v is $\hat{v}_n = \overline{X}_n$.

Example 2. $(X_1, X_2, \ldots, X_n)$ is a sample from $N(m, \sigma^2)$ distribution, where m and σ^2 are unknown parameters. Let's find maximal veracity estimators of this parameters. Function of veracity is:

$$L(x_1, x_2, \ldots, x_n; m, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - m)^2 \right].$$

Equations of veracity are $\frac{\partial \ln L}{\partial m} = \frac{1}{\sigma^2} \sum\limits_{k=1}^{n} (x_k - m) = 0$, and

$$\frac{\partial \ln\ L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (x_k - m)^2 = 0,$$

and their solutions are $m_k = \frac{1}{n} \sum\limits_{k=1}^{n} x_k = \overline{x}_n$ and $\sigma_n^2 = \frac{1}{n} \sum\limits_{k=1}^{n} (x_k - \overline{x}_n)^2$. From this we derive that maximal veracity estimators of parameters m and σ^2 are

$$\hat{m}_k = \overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x}_n)^2.$$

Note: If in distribution $N(m, \sigma^2)$ parameter σ^2 is known, and parameter m is unknown, then $\hat{m}_k = \overline{X}_n$ is the maximal veracity estimator of unknown parameter m. And if m is known and σ^2 is unknown, then the maximal veracity estimator of unknown parameter σ^2 is given by $\hat{\sigma}_n^2 = \frac{1}{n} \sum\limits_{k=1}^{n} (X_k - m)^2$.

2. Pearson χ^2-test

Hypothesis H_0 is to be tested on the sample $(X_1, X_2, \ldots, X_n)$ from the distribution $L(X)$. Distributional function of distribution X is equal to distributional function F. This can be written as: $H_0 : F_x = F$. Alternative is hypothesis H_1 that distributional function of statistic numeric value X is not equal to F. Significant value is α. Benefits of Pearson χ^2- test are that it can be used on every distributional function, there is no extra condition, and in comparison with other test calculations it is relatively simple.

Based on the sample $(X_1, X_2, \ldots, X_n)$ from distribution $L(X)$ we have to verify hypothesis H_0 that distributional function of distribution X is equal to distributional function F. We can write that as $H_0 : F_x = F$. Alternative hypothesis H_1 stands that distribution function of our population is not equal to F. Level significant value is α. Let $R = S_1 \cup S_2 \cup \ldots \cup S_m$ be fragmenting of set of real numbers. For $k \in \{1, 2, \ldots, m\}$, let M_k be number of elements from sample that have values in set M_k and let $p_k = P\{X \in S_k | H_0\}$. Then $M_k \in B(n, p_n)$. We can define random variable $X_n^2 = \sum_{k=1}^{m} \frac{(M_k - np_k)^2}{np_k}$. This random variable can obtain good view on variation between random variables $M_1, M_2, \ldots, M_m$ and expected values $np_1, np_2, \ldots, np_m$. In this way we can find asymptotic distribution of random variable X_n^2 when hypothesis H_0 is valid.

If hypothesis $H_0 : F_x = F$ is true and if $p_k \in (0, 1)$ for $k \in \{1, 2, \ldots, m\}$, then is $X_n^2 \xrightarrow{D} \chi^2_{m-1}$ for $n \to \infty$.

When we have to solve a specific problem, like testing the match of two distributions, with Pearson χ^2 test, we have to follow next reasoning: from condition $P\{\chi^2_{n-1} \geq \chi^2_{\alpha,m-1} | H_0\} \approx \alpha$ we obtain constant $\chi_{1-\alpha,m-1}$. We have

$$P\left\{\sum_{k=1}^{m} \frac{(M_k - np_k)^2}{np_k} \geq \chi^2_{\alpha,m-1}\right\} \approx \alpha.$$

If for given value of statistics the inequality $\sum_{k=1}^{m} \frac{m_k - np_k}{np_k} \geq \chi_{\alpha,m-1}$ holds true, then we dismiss hypothesis H_0. In practice we can use the following approximations, when $n \geq 50$ and $n \cdot p_k \geq 5$ for $k \in \{1, \ldots, m\}$.

With sample $(X_1, X_2, \ldots, X_n)$ from distribution $L(X)$ we want to test hypothesis H_0 that distribution function of statistical numeric value belongs to set $\{F(x, \theta) | \theta \in \Theta\}$. In this case we can make the same conclusion as we had with testing hypothesis that distribution function of numeric value belongs to given distribution function. But here we have that probabilities $p_k(\theta) = P\{X \in S_k | H_0\}$, $k \in \{1, 2, \ldots, m\}$ depend on parameter θ. For all $\theta \in \Theta$ is valid $p_1(\theta) + \ldots + p_m(\theta) = 1$. Similarly, we can define statistic X_n^2, it depends also on parameter θ:

$$X_n^2(\theta) = P\left\{\sum_{k=1}^{m} \frac{(M_k - np_k(\theta))^2}{mp_k(\theta)}\right\}.$$

Let the parameter θ be r-dimensional: $\theta = (\theta_1, \ldots, \theta_r)$, and $r \leq m - 1$.

Then it can be proved: if $\hat{\theta}_n$ is maximal veracity estimator of the unknown parameter θ, defined with sample dimension n, and if hypothesis H_0 is true, then $X_n^2(\hat{\theta}_n) \xrightarrow{D} \chi^2_{n-r-1}$, $n \to \infty^*$ is valid. By this we can, similar to the case

*Def: Series of the random variables X_n is *converging in distribution* to the random variable X if is valid $\lim_{n\to\infty} F_n(X) = F(X)$ *for* $\forall x$ (functions are continous).

without parameter, to test hypothesis H_0. From the condition $P\{\chi^2_{m-r-1} \geq \chi^2_{\alpha,m-r-1}\} = \alpha$ we can find the constant $\chi^2_{\alpha,m-r-1}$. If test statistics, found out the value $X^2_n(\hat{\theta}_n)$, is bigger than the constant $\chi^2_{\alpha,m-r-1}$, then we dismiss hypothesis H_0. In the opposite case we accept hypothesis H_0.

Example 3. At one faculty had been made test from mathematics. At the test student can achieve at most 500 points. On the sample of 750 student the following results are found:

Table 1

Grade	5	4	3	2	1
Number of points	[0,100)	[100,200)	[200,300)	[300,400)	[400,500]
Number of students, which have appropriate number of points	15	140	370	190	35

On given sample, with significant value $\alpha = 0.01$ test hypothesis that number of students, which have appropriate number of points, on this test, has normal $N(\hat{m}, 85^2)$ distribution (or is given sample from population with $N(\hat{m}, 85^2)$ distribution).

Solution:

In the case when we have to test hypothesis that some sample has or doesn't have given distribution, we can use Pearson χ^2-test. But before that, we must stop at unknown parameter $\hat{m}$.

If there was no parameter $\hat{m}$, if $\hat{m}$ is constant, we could apply Pearson χ^2–test immediately. $\hat{m}$ is unknown, so we have, first, to estimate them. We have to do that with maximal veracity method. Veracity function $L(x_1, x_2, \ldots, x_n; \hat{m}) = \prod_{k=1}^{n} f(x_k \hat{m})$, in this case is

$$L(x_1, x_2, \ldots, x_n; \hat{m}) = \prod_{i=1}^{n} \left(\frac{1}{\delta\sqrt{2\pi}}\right) \cdot \exp\left[-\frac{(X_i - \hat{m})^2}{2\delta^2}\right] = \left(\frac{1}{\delta\sqrt{2\pi}}\right)^n \cdot \exp\left[\frac{-\sum_{i=1}^{n}(X_i - \hat{m})^2}{2\delta^2}\right],$$

further

$$\ln L = n \cdot \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{\sum_{i=1}^{n}(X_i - \hat{m})^2}{2\delta^2} \quad \text{and} \quad \frac{\partial \ln\, L}{\partial \hat{m}} = -\frac{\left(-2\sum_{i=1}^{n} X_i + n\hat{m}\right)}{2\delta^2} = 0,$$

$$\sum_{i=1}^{n} X_i - \hat{m}n = 0, \quad \hat{m} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}_n.$$

We can calculate $\overline{X}_n$:

$\overline{X}_n = \frac{1}{750}(15 \cdot 50 + 140 \cdot 150 + 370 \cdot 250 + 190 \cdot 350 + 35 \cdot 350) = 262.$

Maximal veracity estimator of parameter m is 262. Now we can apply Pearson χ^2-test. In calculation of the value $\sum_{k=1}^{m} \frac{m_k - np_k}{np_k}$ it is often practical to put values in table. In our case the table would be as follows:

Table 2

	$[0, 100)$	$[100, 200)$	$[200, 300)$	$[300, 400)$	$[400, 500]$
m_k	15	140	370	190	35
np_k	21.75	153.41	329.78	206.38	39.18
$\frac{(m_k-np_k)^2}{np_k}$	1.84	1.17	4.90	1.30	0.48

Probabilities p_k $k = 1, \dots, 5$ are probabilities defined on adequate intervals of the function which have $N(\hat{m}, 85^2)$ distribution, e.g. $p_1 = P\{-\infty < x \leq 100\}$, because x has distribution $N(\hat{m}, 85^2)$, $p_1 = P\left\{\frac{-\infty - 262}{85} < x \leq \frac{100 - 262}{85}\right\}$ has $N(0, 1)$ distribution, because $\hat{m} = 262$, further $p_1 = \Phi\left(\frac{100-262}{85}\right) - \Phi\left(\frac{-\infty-262}{85}\right)$ and at the end $p_1 = \Phi(-1.9058) - \Phi(-\infty) = 1 - \Phi(1.90) - 0 = 1 - 0.9713 = 0.029$. Number of students n, in sample is 750, so $n \cdot p_k = 21.75$. Similarly,

$$\begin{aligned} p_2 &= P\left\{\tfrac{100-262}{85} < x \leq \tfrac{200-262}{85}\right\} \\ p_2 &= \Phi\left(\tfrac{200-262}{85}\right) - \Phi\left(\tfrac{100-262}{85}\right) = \Phi(-0.7294) - \Phi(-1.9058) = \\ &= 1 - \Phi(0.7294) - (1 - \Phi(1.9058)) = -\Phi(0.7294) + \Phi(1.9058)) = \\ &= -0.7666 + 0.9713 = 0.2045 \end{aligned}$$

and $n \cdot p_k = 153.41$ and so on.

When we summarize the last line we get $\sum_{k=1}^{m} \frac{m_k - np_k}{np_k} = 9.66$.

Now there is only to calculate $\chi^2_{\alpha, m-r-1}$, $m = 5$ (number of columns), $r = 1$ (number of unknown parameters, which are estimated), $\alpha = 0.01$, from table for χ^2 we find value $\chi^2_{0.01,3} = 11.345$. We see that $\sum_{k=1}^{m} \frac{m_k - np_k}{np_k} = 9.66 < 11.345 = \chi^2_{0.01,3}$, it means, that we can accept hypothesis H_0. Given sample belongs to population which has $N(\hat{m}, 85^2)$ distribution. We can say that students' point numbers have normal $N(\hat{m}, 85^2)$ distribution.

Note 1. Note that all values in the third line in table (values for np_k) are bigger then 5, so condition $n \cdot p_k \geq 5$ is fulfilled. If this wouldn't be the case, then we would have to use another test or we could modify table in the following way: column where value $n \cdot p_k$ is smaller then 5 will be joined with column near by, values m_k and m_{k+1} would be summarized, and probability p_k would be calculated from the beginning of the interval in k column to the end interval in column $k + 1$. Situation is similar for $k - 1$. This should be repeated in all columns until condition $n \cdot p_k \geq 5$ is not fulfilled.

Note 2. Interval $[0, 500]$ defined additionally on R, we need function f on whole R so we can calculate probabilities p_i $i = 1, \ldots, m$.

Example 4. We have given results from entering exam in 2006 at the Faculty of Mathematics, Informatics and Physics, Comenius University in Bratislava. Entering exam attended 862 students. Maximal number of points was 20. Let's divide interval $(0, 20)$ into 10 parts:

$$(-\infty, 2), [2, 4), [4, 6), [6, 8), [8, 10), [10, 12), [12, 14), [14, 16), [16, 18), [18, +\infty).$$

With given sample on size 71 with significant value $\alpha = 0.01$ test hypothesis that numbers of students that have adequate number of points are selected from population whit normal distribution.

Solution:

In this case we have two unknown parameters m and σ. We will estimate these two parameters with maximal veracity method. In the previous example we have, that maximal veracity estimator of unknown parameter $\hat{m}$ is $\hat{m} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}_n$. Now we have to find maximal veracity estimator for σ.

Veracity function $L(x_1, x_2, \ldots, x_n; \hat{m}, \hat{\sigma}) = \prod_{k=1}^{n} f(x_k, \hat{m}, \hat{\sigma})$ is

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \hat{m}, \hat{\sigma}) &= \prod_{i=1}^{n} \left(\tfrac{1}{\hat{\delta}\sqrt{2\pi}}\right) \cdot \exp\left[-\tfrac{(X_i - \hat{m})^2}{2\hat{\delta}^2}\right] = \\ &= \left(\tfrac{1}{\hat{\delta}\sqrt{2\pi}}\right)^n \cdot \exp\left[\frac{-\sum_{i=1}^{n}(X_i - \hat{m})^2}{2\hat{\delta}^2}\right], \end{aligned}$$

further

$$\ln L = n \cdot \ln \frac{1}{\hat{\sigma}\sqrt{2\pi}} - \frac{\sum_{i=1}^{n}(X_1 - \hat{m})^2}{2\hat{\delta}^2} \quad \text{and} \quad \frac{\partial \ln L}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} | \frac{1}{2\hat{\sigma}^4} \sum_{k=1}^{n} (x_k - \hat{m})^2 = 0$$

and at the end we have parameter estimator $\hat{\sigma}_n^2 = \overline{S}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$. Then we get $\overline{X}_n = 10.464$ and $\overline{S}_2 = 32.69652$. We will test that sample is from $N(10.464; 32.69652^2)$ distribution. Similar to the previous example, we have to calculate numbers m_k (numbers of students that have points number in adequate interval), probabilities p_k and finish table filing

$p_{(2,4)} = P\{2 < x < 4\} = F\left(\frac{4 - 10.465}{32.6965}\right) - F\left(\frac{2 - 10.465}{32.6965}\right) = F(-0.1977) - F(-0.2589) = 1 - F(0.1977) - (1 - F(0.2589)) = 1 - 0.578 - 1 + 0.601 = 0.023$, etc.

Then we obtain the table:

Table 3

	$(-\infty, 2)$	$[2, 4)$	$[4, 6)$	$[6, 8)$	$[8, 10)$	$[10, 12)$
m_k	0	4	4	11	13	13
p_k	0.339	0.023	0.025	0.025	0.028	0.0195
np_k	24.069	1.633	1.775	1.775	1.988	1.3845
$\frac{(m_k-np_k)^2}{np_k}$						

$[12, 14)$	$[14, 16)$	$[16, 18)$	$[18, 20)$	$[20, +\infty)$
13	9	3	1	0
0.023	0.025	0.026	0.024	0.448
1.633	1.775	1.846	1.704	31.808

Condition $n > 50$ is fulfilled, but condition $n \cdot p_k \geq 5$ doesn't hold true in columns 2,3,4,5,6,7,8,9,10. Table has to be changed in the way that condition $n \cdot p_k \geq 5$, $k \in \{1, \ldots, m\}$ is true. We will join columns 2,3,4,5,6,7,8,9,10, values m_k and np_k will be summarized and other date must be calculated again. Then we obtain the next table:

Table 4

	$(-\infty, 2)$	$(2, 8]$	$(8, 14]$	$(14, 20]$	$(20, +\infty)$
m_k	0	19	39	13	0
p_k	0.339	0.073	0.0705	0.075	0.443
np_k	24.069	5.183	5.005	5.325	31.808
$\frac{(m_k-np_k)^2}{np_k}$	24.07	36.83	230.87	11.06	31.81

Both conditions are fulfilled, so we can continue. We had two unknown parameters during calculation, so when we look for $\chi^2_{\alpha,m-r-1}$, $r = 2$. $\chi^2_{0.01,5-2-1} = \chi^2_{0.01,2} = 9.21$. In the other side $\displaystyle\sum_{k=1}^{m} \frac{m_k - np_k}{np_k} = 334.6$.

We have $\sum\limits_{k=1}^{m} \frac{m_k-np_k}{np_k} = 334.6 > 9.21 > \chi^2_{0.01;5-2-1} = \chi^2_{0.01;2}$.
It means that we dismiss hypothesis H_0. Given sample is not from population which has normal $N(10.464; 32.69652^2)$ distribution, with significance level $\alpha = 0.01$. With given sample we can not claim that number of points at the exam has normal distribution.

References

[1] M. Holender, M. Wulf. Neparametryčeskije metody statistiki, Finansy i Statistika, Moskva, 1983.

[2] P. Mladenović. Verovatnoća i statistika, Vesta Matematički Fakultet Beograd, 1995.

[3] A. Huťa. Materiály z Didaktického semináru, Bratislava, 2006.