8 The statistical inferences based on one sample and two independent samples from Bernoulli distribution

Theorem 8.1
Let $X_1, \ldots, X_n$ be a random sample from Bernoulli distribution $A(\vartheta)$ and let the condition $n\vartheta(1-\vartheta) > 9$ is true. Let $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ be a sample mean (sometimes referred as to sample proportion). Then the statistic
$U = \frac{M-\vartheta}{\sqrt{\frac{M(1-M)}{n}}} \approx N(0,1)$. It has to be read: The statistic U follows asymptotic standard normal distribution.
Thus $100(1-\alpha)\%$ asymptotic confidence limits for the parameter ϑ are:
$d = m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}$
$h = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}$

Remark 8.2
It is essential to realize what is the interpretation of the mean M. The random variable X takes the values: one and zero, where one stands for success. Then $\sum_{i=1}^{n} X_i$ stands for the number of successes in n independent trials and the fracture $\frac{1}{n}\sum_{i=1}^{n} X_i$ stands for the proportion of successes. The sample proportion is the statistic which estimates the parameter ϑ of success probability.

Example 8.3
The marketing department of a particular company analyzes competition market share of the same product that is manufactured by considered company. Drawing randomly 100 consumers it was found out that 34 of them use competitor's product, the rest of them use product of considered company. Find the 95% confidence interval for the proportion of competitor's product in the market.

Solution
Let X_i be a random variable, which takes value 1 if the $i-$th consumer uses the competitor's product and the value 0 otherwise; $i = 1, 2, \ldots, 100$.
Then $X_i \sim A(\vartheta)$ and $X_1, \ldots, X_n$ is a random sample from Bernoulli distribution. The task is to construct the confidence interval for the parameter ϑ of this distribution.
$n = 100 \quad m = \frac{34}{100} \quad u_{1-\alpha/2} = u_{0,975} = 1,96$
Since the parameter ϑ in approximating condition $n\vartheta(1-\vartheta) > 9$ is unknown it should be replaced by its estimate m.
$100.0,34.0,66 = 22,44 > 9$. Thus the estimate m satisfies the condition. Then:

$$d = 0,34 - \sqrt{\tfrac{0,34.0,66}{100}}.1,96 = 0,2472 \qquad h = 0,34 + \sqrt{\tfrac{0,34.0,66}{100}}.1,96 = 0,4328$$

Thus $0,2472 < \vartheta < 0,4328$ with the probability approximately 0,95.
[ϑ is a probability, that the randomly drawn consumer uses competitor's product; this probability lies within the limits of interval (0,2472;0,4328). The confidence, that this interval contains the true parameter, is roughly 95%.]

Theorem 8.4
Let $X_1, \ldots, X_n$ be a random sample from $A(\vartheta)$, $c \in (0,1)$, M be a sample mean and let the condition $n\vartheta(1-\vartheta) > 9$ is true.
At the asymptotic confidence level α the null hypothesis $H_0 : \vartheta = c$ is rejected in favour of the

alternative hypothesis H_1, if the realization of the test statistic
$T_0 = \frac{M-c}{\sqrt{\frac{c(1-c)}{n}}}$ falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

two-tailed test	$H_1 : \vartheta \neq c$	$W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
left-tailed test	$H_1 : \vartheta < c$	$W = (-\infty,\ -u_{1-\alpha}\rangle$
right-tailed test	$H_1 : \vartheta > c$	$W = \langle u_{1-\alpha},\ \infty)$

[If H_0 is true, then $T_0 \approx N(0,1)$.]

Remark 8.5
The test statistic is derived using Moivre-Laplace theorem. $T_0 = \frac{M-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \approx N(0,1)$

Remark 8.6
The pivotal statistic, which is instrumental towards construction of confidence interval, differs from the test statistic stated in previous theorem!

Example 8.7
Manufacturing some components, the manufacturer declares, that the probability of manufactured defective product is $\vartheta = 0,01$. The sample consisting of 1000 products was drawn randomly and it was found that 16 products were defective. At the asymptotic significance level 0.05 test the hypothesis $H_0 : \vartheta = 0,01$ against $H_1 : \vartheta \neq 0,01$.

Solution
Since the parameter ϑ is unknown the condition of normal approximation $n\vartheta(1-\vartheta) > 9$ should be replaced by the condition $nm(1-m) > 9$.
$1000.\frac{16}{1000} \cdot \frac{984}{1000} = 15.744 > 9$, thus the normal approximation is possible.
The realization of the test statistic follows: $t_0 = \frac{16/1000-0,01}{\sqrt{\frac{0,01\cdot 0,99}{1000}}} = 1,907$
The critic region is expressed: $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty) = (-\infty,\ -1,96\rangle \bigcup \langle 1,96,\ \infty)$.
Since $1,907 \notin W$, $\quad H_0$ is not rejected at the asymptotic significance level 0,05.
[Based on the values of the random sample there is no reason to doubt about declared probability 0.01 of manufacturing the defective product.]

Theorem 8.8
Let us consider two independent samples. Let $X_{11}, \ldots, X_{1n_1}$ be a random sample from Bernoulli distribution $A(\vartheta_1)$ and $X_{21}, \ldots, X_{2n_2}$ be a random sample from $A(\vartheta_2)$. Let the conditions $n_i\vartheta_i(1-\vartheta_i) > 9,\ i = 1,2$ are true. Let $M_1,\ M_2$ be sample means. Then the statistic

$$U = \frac{(M_1-M_2)-(\vartheta_1-\vartheta_2)}{\sqrt{\frac{M_1(1-M_1)}{n_1}+\frac{M_2(1-M_2)}{n_2}}} \approx N(0,1).$$

Thus $100(1-\alpha)\%$ asymptotic confidence limits for the parametric function $\vartheta_1 - \vartheta_2$ are:
$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2}$
$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2}$

Example 8.9
The supermarket management advertised the week of prices reduction. The aim was to find out if the prices reduction does impact the proportion of the heavy shopping (over 500 Kč). During the week without reductions it was drawn randomly 200 customers and 97 of them had done heavy shopping. During the week with reductions the size of the random sample was 300 and the number of heavy

shopping was 162. Determine the 95% asymptotic confidence interval for the difference between the probabilities of heavy shopping during the week without reductions and week with reductions.

Solution

The random variable $X_{1,i}$ takes the value 1, if during the week without reduction in prices the $i-$th randomly drawn customer realizes heavy shopping and the value 0 otherwise, $i = 1, \ldots, 200$. The random variables $X_{1,1}, \ldots, X_{1,200}$ form the random sample from distribution $A(\vartheta_1)$. Further the random variable X_{2i} takes the value 1, if during the week with reduction in prices the $i-$th randomly drawn customer realizes heavy shopping and the value 0 otherwise, $i = 1, \ldots, 300$. The random variables $X_{2,1}, \ldots, X_{2,300}$ form the random sample from distribution $A(\vartheta_2)$ and this sample is independent from the previous one.

$n_1 = 200, \quad n_2 = 300, \quad m_1 = 97/200, \quad m_2 = 162/300.$

To verify the conditions $n_i\vartheta_i(1-\vartheta_i) > 9,\ i = 1, 2$ of normal approximation the unknown parameters ϑ_i should be replaced by their estimates m_i. Thus this estimates meet the conditions:

$200 \cdot 97/200 \cdot 103/200 = 49,955 > 9, \ \ 300 \cdot 162/300 \cdot 138/300 = 74,52 > 9.$

Thus the $100(1-\alpha)\%$ asymptotic confidence limits for parametric function $\vartheta_1 - \vartheta_2$ follow:

$$d = m_1 - m_2 - \sqrt{\tfrac{m_1(1-m_1)}{n_1} + \tfrac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2} =$$
$$= 97/200 - 162/300 - \sqrt{\tfrac{97/200(1-97/200)}{200} + \tfrac{162/300(1-162/300)}{300}} \cdot 1,96 =$$
$$= -0,1443$$
$$h = m_1 - m_2 + \sqrt{\tfrac{m_1(1-m_1)}{n_1} + \tfrac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2} =$$
$$= 97/200 - 162/300 + \sqrt{\tfrac{97/200(1-97/200)}{200} + \tfrac{162/300(1-162/300)}{300}} \cdot 1,96 =$$
$$= 0,0343$$

Hence the parametric funktion

$\vartheta_1 - \vartheta_2 \in (-0,1443\ ,\ 0,0343)$ with the probability approximately 0.95.

Theorem 8.10

Let us consider two independent samples. Let $X_{11}, \ldots, X_{1n_1}$ be a random sample from Bernoulli distribution $A(\vartheta_1)$ and $X_{21}, \ldots, X_{2n_2}$ be a random sample from $A(\vartheta_2)$. Let the conditions $n_i\vartheta_i(1-\vartheta_i) > 9,\ i = 1, 2$ are true. Let $M_1,\ M_2$ be sample means.

At the asymptotic level α the null hypothesis $H_0 : \vartheta_1 - \vartheta_2 = c$ is rejected in favour of the alternative hypothesis if the realization of the test statistic

$T_0 = \dfrac{(M_1-M_2)-c}{\sqrt{\frac{M_1(1-M_1)}{n_1}+\frac{M_2(1-M_2)}{n_2}}}$ falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

two-sided test	$H_1 : \vartheta_1 - \vartheta_2 \neq c$	$W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
left-sided test	$H_1 : \vartheta_1 - \vartheta_2 < c$	$W = (-\infty,\ -u_{1-\alpha}\rangle$
right-sided test	$H_1 : \vartheta_1 - \vartheta_2 > c$	$W = \langle u_{1-\alpha},\ \infty)$

[If H_0 is true, then $T_0 \approx N(0,1)$.]

Remark 8.11

In the case of $H_0 : \vartheta_1 - \vartheta_2 = 0$ $(c = 0)$ the test statistic T_0 is preferable,

$T_0 = \dfrac{M_1-M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, where $M_* = \frac{n_1M_1+n_2M_2}{n_1+n_2}$.

[If H_0 is true, then $T_0 \approx N(0,1)$.]

Example 8.12

Using the data from exercise 8.9 and at the asymptotic significance level 0.05 test the hypothesis,

that the week of prices reductions does not increase the probability of heavy shopping.

Solution

We are running the left tailed test $H_0 : \vartheta_1 - \vartheta_2 = 0$ versus $H_1 : \vartheta_1 - \vartheta_2 < 0$ at asymptotic $\alpha = 0,05$.
$n_1 = 200,\ n_2 = 300,\ m_1 = 97/200,\ m_2 = 162/300,\ m_* = (97 + 162)/500 = 0,518.$
The assumptions of normal approximation have been verified in 8.9

ad a) Using confidence interval method:

For the left-tailed test we use right-sided confidence interval:

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha} =$$
$$= 97/200 - 162/300 + \sqrt{\frac{97/200(1-97/200)}{200} + \frac{162/300(1-162/300)}{300}} \cdot 1,645 =$$
$$= 0,02$$

Since the value $c = 0$ is within the interval $(-\infty\ ;\ 0,02),\ H_0$ is not rejected at the asymptotic $\alpha = 0,05$, thus the week of prices reductions does not increase the probability of heavy shopping.

ad b) Using classical method:

The test statistic follows:

$$T_0 = \frac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad \text{kde } M_* = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$$

$$m_* = \frac{200.97/200 + 300.162/300}{200+300} = 0,518$$

$$t_0 = \frac{97/200 - 162/300}{\sqrt{0,518(1-0,518)\left(\frac{1}{200} + \frac{1}{300}\right)}} = -1,2058$$

The critical region follows:
$W = (-\infty\ ,\ -u_{1-\alpha}\rangle = (-\infty\ ,\ -u_{0,95}\rangle = (-\infty\ ,\ -1,645\rangle.$
Since $t_0 \notin W,\ H_0$ is not rejected at the asymptotic $\alpha = 0,05$