Statistics

Lecture 10

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

Hypothesis testing:
Non-parametric tests

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- About statistical hypothesis testing:

  Parametric and Non-parametric tests
- Sign test for the median
- Pearson's χ^2-test for the goodness of fit
- χ^2-test of independence of qualitative data items

About statistical hypothesis testing

- Parametric and Non-parametric tests

Parametric and Non-parametric tests

There are two large classes of statistical tests: **parametric** and **non-parametric**.

- The **parametric** tests make assumptions about the probability distributions of the random variables that are subject to the test. It is often assumed that the underlying distribution is normal (Gaussian).

- The **non-parametric** tests do not make such assumptions. The non-parametric tests can be used if the random variables are not normally distributed.

Sign test for the median

- Sign test for the median
- Paired sign test for

  the difference of the medians

Sign test for the median

Motivation:

Let X be a random variable (of any distribution), but assume that its cumulative distribution function F is continuous.

Recall that the median $\tilde{x}$ of the random variable X is the value such that

$$P(X < \tilde{x}) = \frac{1}{2} = P(\tilde{x} < X)$$

We conjecture / we assume / we speculate / we … / that the mean $\tilde{x}$ of the random variable X is equal to some given value $\tilde{x}_0 \in \mathbb{R}$.

We thus formulate the null hypothesis: H_0: $\tilde{x} = \tilde{x}_0$

Sign test for the median

The sign test proceeds as follows:

- Let us have n observations $x_1, x_2, \dots, x_n$ of the random variable X, whose cumulative distribution function F is continuous.
- Considering the null hypothesis $(H_0: \tilde{x} = \tilde{x}_0)$ about the median, calculate the n differences

$$x_1 - \tilde{x}_0, \quad x_2 - \tilde{x}_0, \quad \dots, \quad x_n - \tilde{x}_0$$

- Drop any zero differences (i.e., if $x_i - \tilde{x}_0 = 0$, then drop x_i from the sample).
- We have a sample of m non-zero differences

$$x_{j_1} - \tilde{x}_0, \quad x_{j_2} - \tilde{x}_0, \quad \dots, \quad x_{j_m} - \tilde{x}_0$$

Sign test for the median

- Let

$$Z = \left|\left\{ i : x_{j_i} - \tilde{x}_0 < 0 \right\}\right|$$

  be the number of the negative differences.

<u>**Theorem:**</u>

Under the null hypothesis $(H_0\colon \tilde{x} = \tilde{x}_0)$ that the median $\tilde{x}$ of the random variable X is $\tilde{x}_0$

$$Z \sim \mathrm{Bi}\left(m, \tfrac{1}{2}\right)$$

i.e. the random variable Z follows the binomial probability distribution.

Sign test for the median

Remark: We actually test the hypothesis that the probability

$$P(X < \tilde{x}_0) = P(X \leq \tilde{x}_0) = \frac{1}{2}$$

(We have $P(X < \tilde{x}_0) = P(X \leq \tilde{x}_0)$ because we assume that the cumulative distribution function F is continuous at $\tilde{x}_0$.)

Therefore, we could test in the same manner the null hypothesis that

$\tilde{x}_0$ is the first quartile $\left(P(X < \tilde{x}_0) = P(X \leq \tilde{x}_0) = \frac{1}{4}\right.$, whence $\left.Z \sim \mathrm{Bi}\left(m, \frac{1}{4}\right)\right)$, or that

$\tilde{x}_0$ is the third decile $\left(P(X < \tilde{x}_0) = P(X \leq \tilde{x}_0) = \frac{3}{10}\right.$, whence $\left.Z \sim \mathrm{Bi}\left(m, \frac{3}{10}\right)\right)$, etc.

Sign test for the median

Having stated the **null hypothesis** about the median

$$H_0: \quad \tilde{x} = \tilde{x}_0 \qquad \text{or} \qquad H_0: \quad P(X < \tilde{x}_0) = p_0 = \frac{1}{2}$$

we also state the **alternative hypothesis:**

- two-sided: $H_1: \quad \tilde{x} \neq \tilde{x}_0$ or $H_1: \quad P(X < \tilde{x}_0) \neq p_0$
- one-sided: $H_1: \quad \tilde{x} > \tilde{x}_0$ or $H_1: \quad P(X < \tilde{x}_0) < p_0$
- one-sided: $H_1: \quad \tilde{x} < \tilde{x}_0$ or $H_1: \quad P(X < \tilde{x}_0) > p_0$

The test then proceeds as the binomial test (or z-test approximately) for the

Sign (binomial) test for the median

Consider the first case $(H_1\!: \tilde{x} \neq \tilde{x}_0)$ first. We have:

$$H_0\!:\ P(X < \tilde{x}_0) = p_0 = 1/2$$
$$H_1\!:\ P(X < \tilde{x}_0) \neq p_0$$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find the **critical values** $K, L \in \{0, 1, \dots, m\}$ so that

  K is the largest number and L is the least number such that

$$\sum_{k=0}^{K} \binom{m}{k} p_0^k q_0^{m-k} = \sum_{k=0}^{K} \binom{m}{k} \frac{1}{2^m} \leq \frac{\alpha}{2} \quad \text{and} \quad \sum_{k=L}^{m} \binom{m}{k} p_0^k q_0^{n-k} = \sum_{k=L}^{m} \binom{m}{k} \frac{1}{2^m} \leq \frac{\alpha}{2}$$

- if $Z \in \{0, \dots, K\} \cup \{L, \dots, n\}$, **the critical region**, then **<u>reject</u>** the null hypothesis
- if $Z \in \{K+1, \dots, L-1\}$, then **<u>do not reject</u>** (or <u>fail to reject</u>) the null hypothesis

Sign (binomial) test for the median

Consider now the second case $(H_1\colon \tilde{x} > \tilde{x}_0)$. We have: $H_0\colon\ P(X < \tilde{x}_0) = p_0 = 1/2$

$H_1\colon\ P(X < \tilde{x}_0) < p_0$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find **the critical value** $K \in \{0, 1, \dots, m\}$ so that K is the largest number such that

$$\sum_{k=0}^{K} \binom{m}{k} p_0^k q_0^{m-k} = \sum_{k=0}^{K} \binom{m}{k} \frac{1}{2^m} \le \alpha$$

- if $Z \in \{0, \dots, K\}$, **the critical region**, then **reject** the null hypothesis
- if $Z \in \{K+1, \dots, m\}$, then **do not reject** (or **fail to reject**) the null hypothesis

Sign (binomial) test for the median

Consider finally the third case $(H_1\colon \tilde{x} < \tilde{x}_0)$. We have:

$$H_0\colon\ P(X < \tilde{x}_0) = p_0 = 1/2$$
$$H_1\colon\ P(X < \tilde{x}_0) > p_0$$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find the **critical value** $L \in \{0, 1, \dots, m\}$ so that L is the least number such that

$$\sum_{k=L}^{m} \binom{m}{k} p_0^k q_0^{m-k} = \sum_{k=L}^{m} \binom{m}{k} \frac{1}{2^m} \le \alpha$$

- if $Z \in \{L, \dots, m\}$, **the critical region**, then **reject** the null hypothesis
- if $Z \in \{0, \dots, L-1\}$, then **do not reject** (or fail to reject) the null hypothesis

Sign (z-) test for the median

It is inconvenient to calculate the sums $\sum_{k=0}^{K}\binom{m}{k}\frac{1}{2^m}$ and $\sum_{k=L}^{m}\binom{m}{k}\frac{1}{2^m}$ if m is large. It is more convenient then to approximate the sums by using the de Moivre-Laplace Central Limit Theorem (for $p = q = 1/2$):

It holds, whenever $-\infty \le a < b \le +\infty$, that

$$\frac{\sum_{k=A_m}^{B_m}\binom{m}{k}\frac{2}{2^m} - n}{\sqrt{m}} \rightarrow \underbrace{\int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t}_{\Phi(b)-\Phi(a)} \qquad \text{as} \quad m \rightarrow \infty$$

where $A_m = \lceil (m + a\sqrt{m})/2 \rceil \ge 0$ and $B_m = \lfloor (m + b\sqrt{m})/2 \rfloor \le m$ if $m \ge \max(a^2, b^2)$.

Moreover, the convergence is uniform with respect to a and b.

Sign (z-) test for the median

De Moivre-Laplace Central Limit Theorem (reformulated):

If $X \sim \mathrm{Bi}(m, 1/2)$, whenever $-\infty \leq a < b \leq +\infty$, it then holds

$$P\left(a < \frac{2X - m}{\sqrt{m}} < b\right) \rightarrow \underbrace{\int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t}_{\Phi(b)-\Phi(a)} \qquad \text{as} \quad m \rightarrow \infty$$

and the convergence is uniform with respect to a and b.

Sign (z-) test for the median

Consider the first case $(H_1\colon \tilde{x} \neq \tilde{x}_0)$ first. We have:

$$H_0\colon\ P(X < \tilde{x}_0) = p_0 = 1/2$$
$$H_1\colon\ P(X < \tilde{x}_0) \neq p_0$$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find $c > 0$ so that

$$\int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t = \frac{\alpha}{2} \quad \text{and} \quad \int_{+c}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t = \frac{\alpha}{2}$$

- if $Z \leq (m - c\sqrt{m})/2$ or $(m + c\sqrt{m})/2 \leq Z$, **the critical region**, then **reject** the null hypothesis
- if $(m - c\sqrt{m})/2 < Z < (m + c\sqrt{m})/2$, then **do not reject** (or fail to reject) the null hypothesis

Sign (z-) test for the median

Consider now the second case $(H_1: \tilde{x} > \tilde{x}_0)$. We have: H_0: $P(X < \tilde{x}_0) = p_0 = 1/2$

H_1: $P(X < \tilde{x}_0) < p_0$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find $c > 0$ so that

$$\int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t = \alpha$$

- if $Z \leq (m - c\sqrt{m})/2$, **the critical region**, then **reject** the null hypothesis
- if $(m - c\sqrt{m})/2 < Z$, then **do not reject** (or fail to reject) the null hypothesis

Sign (z-) test for the median

Consider finally the third case $(H_1\colon \tilde{x} < \tilde{x}_0)$. We have:

$$H_0\colon\ P(X < \tilde{x}_0) = p_0 = 1/2$$
$$H_1\colon\ P(X < \tilde{x}_0) > p_0$$

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find $c > 0$ so that

$$\int_{+c}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t = \alpha$$

- if $(m + c\sqrt{m})/2 \le Z$, **the critical region**, then **reject** the null hypothesis
- if $Z < (m + c\sqrt{m})/2$, then **do not reject** (or fail to reject) the null hypothesis

Sign test for the median

Remarks:

- By using another probability (such as $p_0 = 0.25$, $p_0 = 0.3$, etc.) we can test the null hypothesis that $\tilde{x}_0$ is, e.g., the first quartile, the third decile, etc.

- If we know that the distribution of X is symmetric $(F(x) = 1 - F(-x))$, then the mean $\mu = \mathrm{E}[X]$ and the median $\tilde{x}$ of the random variable X coincide ($\tilde{x} = \mu$). Then the sign test for the median can also be used as another test for the mean (H_0: $\mu = \tilde{x}_0$).

Sign test for the median

Remarks:

- More generally, if we know that the mean $\mu = \mathrm{E}[X]$ is the p_0-quantile $(0 < p_0 < 1)$ of the distribution of the random variable X with a continuous cumulative distribution function, then the sign test can also be used as another test for the mean $(H_0\colon\ \mu = \tilde{x}_0$ with $Z = \left|\left\{i : x_{j_i} < \tilde{x}_0\right\}\right| \sim \mathrm{Bi}(m, p_0))$.

- Exercise: Apply the procedure of the sign test to determine the confidence interval for the median, i.e. the interval of values $\tilde{x}_0$ such that the null hypothesis is not rejected for them.

Paired sign test for the difference of the medians

Motivation:

Let us have a sample of n objects, e.g. n patients.

We do two measurements with each of the objects (patients)

— before some treatment

— after the treatment

The purpose it to learn whether the treatment has any effect.

(Hence the null hypothesis: "The treatment has no effect.")

Let $x_1, x_2, \ldots, x_n$ be the values measured before the treatment, and

let $y_1, y_2, \ldots, y_n$ be the values measured after the treatment.

Paired sign test for the difference of the medians

That is, the measurement x_i and y_i is done with the i-th object (patient) before and after the treatment for $i = 1, 2, \ldots, n$.

FIRST, assume that only two outcomes are possible:

- $x_i < y_i$ (improvement)
- $x_i > y_i$ (worsening)

Objects with $x_i = y_i$ are dropped from the sample.

We then can test the null hypothesis that the treatment has no effect, i.e.

$$Z = |\{i : x_i < y_i\}| \sim \mathrm{Bi}\left(m, \tfrac{1}{2}\right)$$

etc. (Finish the details of the test analogously as above as an exercise.)

Paired sign test for the difference of the medians

That is, the measurement x_i and y_i is done with the i-th object (patient) before and after the treatment for $i = 1, 2, \dots, n$.

SECOND, assume that $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ are the numerical outcomes of the random variable X and Y, respectively, with a continuous cumulative distribution function F_X and F_Y, respectively.

<u>Theorem:</u> The median $\tilde{x}_0$ of the difference $X - Y$ of the random variables is

$$\tilde{x}_0 = \tilde{x} - \tilde{y}$$

Paired sign test for the difference of the medians

Thus, we can test the null hypothesis that the median $\tilde{x}$ of the random variable X (before the treatment) is the same as the median $\tilde{y}$ of the random variable Y (after the treatment), i.e. their difference is $\tilde{x}_0 = \tilde{x} - \tilde{y} = 0$.

(More generally, we can test that the difference $\tilde{x} - \tilde{y}$ is equal to some prescribed value $\tilde{x}_0 \in \mathbb{R}$.)

(Complete the details of the test analogously as above as an exercise.)

χ^2-test for goodness of fit

- Pearson's χ^2-test for the goodness of fit

Pearson's χ^2-test for the goodness of fit

Let X be a random variable (discrete or continuous) and let F be the cumulative distribution function of the random variable X.

We do not know the cumulative distribution function F.

We have the numerical results $x_1 = X(\omega_1),\ x_2 = X(\omega_2),\ \dots,\ x_N = X(\omega_N)$ of N trials of the corresponding random experiment.

Let F_0 be some cumulative distribution function. We conjecture / we assume / we speculate / we … / that $F = F_0$, i.e. the random variable X follows the probability distribution with the cumulative distribution function $F = F_0$.

Pearson's χ^2-test for the goodness of fit

More generally, let $\mathcal{F}_0$ be a class of cumulative distribution functions (c.d.f.'s) of a certain type, such as

- the collection of all c.d.f.'s of $\mathcal{U}(a, b)$ for various $a, b \in \mathbb{R},\ a < b$
- the collection of all c.d.f.'s of $\mathcal{N}(\mu, \sigma^2)$ for various $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_0^+$
- the collection of all c.d.f.'s of $\mathrm{Exp}(\lambda)$ for various $\lambda \in \mathbb{R}^+$
- etc.

Having the numerical results $x_1 = X(\omega_1),\ x_2 = X(\omega_2),\ \ldots,\ x_N = X(\omega_N)$ of N trials of a random experiment, we conjecture / we assume / we speculate / we … / that $F \in \mathcal{F}_0$, i.e. the random variable X follows the probability distribution

Pearson's χ^2-test for the goodness of fit

Having the numerical results $x_1 = X(\omega_1),\ x_2 = X(\omega_2),\ \ldots,\ x_N = X(\omega_N)$ of the N trials of the random experiment and having the class $\mathcal{F}_0$ of the cumulative distribution functions – <u>first of all</u> – find the cumulative distribution function $F_0 \in \mathcal{F}_0$ that best fits the experimental data:

- if $\mathcal{F}_0 = \{F_0\}$, then the c.d.f. F_0 is given; the number of parameters is $\nu = 0$
- if $\mathcal{F}_0$ is the collection of all c.d.f.'s of $\mathcal{N}(\mu, \sigma^2)$, then put

$$\mu = \bar{x} \qquad \text{and} \qquad \sigma^2 = s^2$$

(the sample mean and the sample variance); the number of parameters is $\nu = 2$

Pearson's χ^2-test for the goodness of fit

- if $\mathcal{F}_0$ is the collection of all c.d.f.'s of $\mathrm{Exp}(\lambda)$, then put

$$\text{either} \quad \lambda = \frac{1}{\bar{x}} \qquad \text{or} \quad \lambda = \sqrt{\frac{1}{s^2}}$$

  the number of parameters is $\nu = 1$

  (recall: if $X \sim \mathrm{Exp}(\lambda)$, then $\mathrm{E}[X] = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$)

- if $\mathcal{F}_0$ is the collection of all c.d.f.'s of $\mathcal{U}(a,b)$, then consider the German Tank Problem (see previous lectures); the number of parameters is $\nu = 2$
- etc.

Pearson's χ^2-test for the goodness of fit

Having the sample data $x_1, x_2, \dots, x_N$ of the random variable X and the cumulative distribution function $F_0 \in \mathcal{F}_0$ that best fits the sample.

Now – as the second step – choose n intervals

$$(t_0, t_1], \quad (t_1, t_2], \quad (t_2, t_3], \quad \dots, \quad (t_{n-2}, t_{n-1}], \quad (t_{n-1}, t_n]$$

with

$$t_0 < t_1 < t_2 < t_3 < \cdots < t_{n-2} < t_{n-1} < t_n$$

as well as

$$t_0 < \min\{x_1, \dots, x_N\} \qquad \text{and} \qquad \max\{x_1, \dots, x_N\} \leq t_n$$

so that

— there are at least 5 outcomes in each of the intervals

Pearson's χ^2-test for the goodness of fit

Formulate <u>the null hypothesis</u>: The random variable X follows the probability distribution with the cumulative distribution function $F = F_0$:

$$H_0: \quad F = F_0$$

Next – <u>as the third step</u> – assume the null hypothesis H_0 and calculate the theoretical probability that $t_{i-1} < X \leq t_i$, i.e.

$$p_i = P(t_{i-1} < X \leq t_i) =$$

$$= F_0(t_i) - F_0(t_{i-1}) \qquad \text{for} \quad i = 1, 2, \ldots, n$$

Pearson's χ^2-test for the goodness of fit

Since p_i is the expected probability (under the null hypothesis H_0) that $X \in (t_{i-1}, t_i]$ and we have a sample $x_1, x_2, \dots, x_N$ of N observations, we should find about

$$E_i = N \times p_i$$

observations in the interval $(t_{i-1}, t_i]$ for $i = 1, 2, \dots, n$.

Let

$$O_i = \left|\{j : x_j \in (t_{i-1}, t_i]\}\right|$$

be the true number of the observations found in the interval $(t_{i-1}, t_i]$ for $i = 1, 2, \dots, n$.

Pearson's χ^2-test for the goodness of fit

Theorem: If the null hypothesis $H_0\colon F = F_0$ is true, then the statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-\nu-1} \quad \textit{approximately} \quad \text{as} \quad N \to \infty$$

where

- n is the number of the intervals $(t_{i-1}, t_i]$
- ν is the number of the parameters that have been determined when finding the cumulative distribution function F_0 $(\nu = 0, 1, 2, \ldots)$
- O_i is the number of the results found (observed) in the i-th interval $(t_{i-1}, t_i]$
- E_i is the number of the results expected (if H_0 is true) in the interval $(t_{i-1}, t_i]$

Pearson's χ^2-test for the goodness of fit

Now, finish Pearson's χ^2-test for the goodness of fit $(H_0: F = F_0)$ as follows:

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- find the **critical value** $c > 0$ so that

$$\int_c^{+\infty} f(x)\, \mathrm{d}x = \alpha$$

  where f is the density of the χ^2-distribution with $n - \nu - 1$ degrees of freedom
- if $X^2 \geq c$, **the critical region**, then **reject** the null hypothesis
- if $X^2 < c$, then **do not reject** (or fail to reject) the null hypothesis

Example: Tests for population proportion

Tossing a coin repeatedly, we ask whether the coin is fair.

More generally, we consider a Bernoulli trial, with the probability of the success being $p \in (0,1)$, and with the probability of the failure being $q = 1 - p$.
<u>We do not know the true probability</u> p.

We conjecture / We assume / We … / that the probability $p = p_0$, i.e.
the (unknown) probability p is equal to some prescribed value $p_0 \in (0,1)$,
e.g., in the case of the coin, conjecture that $p_0 = 50\ \%$ (meaning the coin is fair).

Example: Tests for population proportion

We now know three statistical tests to test the null hypothesis that $p = p_0$:

- the binomial test for the population proportion
- the *z*-test for the population proportion
- Pearson's χ^2-test for the goodness of fit

The binomial test is exact and the *z*-test is an approximation of it.

Both binomial test and *z*-test allow one-sided or two-sided alternative hypothesis.

Pearson's χ^2-test for the goodness of fit allows two-sided alternative hypothesis $(H_1\colon F \neq F_0)$ only.

Example: Tests for population proportion

Pearson's χ^2-test for the goodness of fit proceeds as follows:

- there are two intervals (1 = "success" and 0 = "failure")
- having N observations of the random variable X, we expect (under the null hypothesis that $p = p_0$) that $E_1 = N \times p_0$ and $E_0 = N \times (1 - p_0)$
- let O_1 and O_0 be the observed number of successes and failures, respectively
- the statistic

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_0 - E_0)^2}{E_0} \sim \chi_1^2 \quad \textit{approximately} \quad \text{as} \quad N \to \infty$$

(we have $n = 2$ and $\nu = 0$, therefore $n - \nu - 1 = 1$)

Pearson's χ^2-test for the goodness of fit

Remark: In Pearson's χ^2-test for the goodness of fit, we have

$$X^2 \sim \chi^2_{n-\nu-1}$$

where

- n is the number of the intervals $(t_{i-1}, t_i]$
- ν is the number of the parameters that have been determined when finding the cumulative distribution function F_0 $(\nu = 0, 1, 2, \ldots)$

Notice that one degree of freedom ("−1") must always be subtracted because the observed counts $O_1, O_2, \ldots, O_n$ are bound by the equation

$$O_1 + O_2 + \cdots + O_n = N$$

therefore only $n - 1$ of the counts (such as $O_1, O_2, \ldots, O_{n-1}$, say) are free,

χ^2-test of independence of qualitative data items

- χ^2-test of independence of qualitative data items

χ^2-test of independence of qualitative data items

Consider a dataset where each data unit has two qualitative data items (i.e. two qualitative variables).

Let the qualitative variables under the consideration be denoted by A and B.

Let the variable A can attain up to r ("rows") distinct categories

$$A_1, \quad A_2, \quad \dots, \quad A_r$$

Let the variable B can attain up to s ("columns") distinct categories

$$B_1, \quad B_2, \quad \dots, \quad B_s$$

The counts of the occurrences of all the $r \times s$ combinations of the categories are easily summarized by a contingency table.

Contingency table

the observed counts of the combinations of the categories A_i&B_j for i=1,…,r & j=1,…,s

$A \setminus B$	B_1	B_2	…	B_s	TOTAL
A_1	n_{11}	n_{12}	…	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2s}	$n_{2.}$
…	⋮	⋮	…	⋮	⋮
A_r	n_{r1}	n_{r2}	…	n_{rs}	$n_{r.}$
TOTAL	$n_{.1}$	$n_{.2}$	…	$n_{.s}$	n

marginal totals

marginal totals

the grand total

2 2 contingency table

The 2 2 contingency table is popular.

It is a contingency table with r=2 rows and s=2 columns.

the observed counts of the combinations of the categories A_i&B_j for i=1,2 & j=1,2

A \ B	B_1	B_2	TOTAL
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n_{.1}$	$n_{.2}$	n

marginal totals

marginal totals

the grand total

χ^2-test of independence of qualitative data items

Having all the observed counts of the combinations of the categories A_i & B_j summarized in the contingency table for $i=1,\dots,r$ and for $j=1,\dots,s$, we ask whether the category of the data item (variable) ***B*** depends upon the category of the data item (variable) ***A***, or whether the categories of both data items (variables) ***A*** and ***B*** are independent of each other.

Assume therefore the null hypothesis H_0:

the categories of both data items (variables) ***A*** and ***B*** are independent of each other

χ^2-test of independence of qualitative data items

Having all the observed counts of the combinations of the categories A_i & B_j summarized in the contingency table for $i=1,\dots,r$ and for $j=1,\dots,s$, assume <u>the null hypothesis</u> H_0 that the categories of both data items (variables) ***A*** and ***B*** are independent of each other.

Now – if we choose a data unit randomly:

- What is the probability that the data item ***A*** of the chosen data unit is of category A_i for some $i=1,\dots,r$?
- What is the probability that the data item ***B*** of the chosen data unit is of category B_j for some $j=1,\dots,s$?

χ^2-test of independence of qualitative data items

The total number of all data units is n.

The count of the data units of category A_i is $n_{i\cdot}$

Therefore, the probability that a randomly selected data unit is of category A_i is

$$p_{i\cdot} = \frac{n_{i\cdot}}{n}$$

The count of the data units of category B_j is $n_{\cdot j}$

Therefore, the probability that a randomly selected data unit is of category B_j is

$$p_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2-test of independence of qualitative data items

Recall that the probability that a randomly selected data unit is of category A_i and B_j is

$$p_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{and} \quad p_{\cdot j} = \frac{n_{\cdot j}}{n}$$

respectively. If the null hypothesis H_0 (that the categories of A and B are independent of each other) is true, then the (cumulative) probability that a randomly selected data unit is of category A_i and B_j should be

$$p_{ij} = p_{i\cdot} \times p_{\cdot j} = \frac{n_{i\cdot} \times n_{\cdot j}}{n^2}$$

for $i = 1, 2, \ldots, r$ and for $j = 1, 2, \ldots, s$.

χ^2-test of independence of qualitative data items

Once the probability that a randomly selected data unit is of category A_i and B_j is

$$p_{ij} = p_{i\cdot} \times p_{\cdot j} = \frac{n_{i\cdot} \times n_{\cdot j}}{n^2}$$

then we should expect

$$E_{ij} = p_{ij} \times n = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

data units of category A_i and B_j for $i = 1, 2, \dots, r$ and for $j = 1, 2, \dots, s$

if the null hypothesis H_0 (that the categories of A and B are independent

of each other) is true.

χ^2-test of independence of qualitative data items

Expecting

$$E_{ij} = p_{ij} \times n = \frac{n_{i.} \times n_{.j}}{n}$$

and observing

$$O_{ij} = n_{ij}$$

data units of category A_i and B_j for $i = 1, 2, \ldots, r$ and for $j = 1, 2, \ldots, s$,

we apply Pearson's χ^2-test for the goodness of fit to see if the observed counts agree with the expected counts, i.e. if the null hypothesis H_0 (that the categories of A and B are independent of each other) is true.

χ^2-test of independence of qualitative data items

Calculate

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n \times n_{ij} - n_{i.} \times n_{.j})^2}{n_{i.} \times n_{.j}}$$

<u>Theorem:</u>

If the null hypothesis is true, then

$$X^2 \sim \chi^2_{(r-1)(s-1)} \quad \textit{approximately} \quad \text{as} \quad n \to \infty$$

Notice the number of the degrees of freedom

(see below)

χ^2-test of independence of qualitative data items

The number of the degrees of freedom:

The observed counts O_{ij} for $i = 1, \dots, r$ and for $j = 1, \dots, s$

are bound by the system of $r + s$ equations:

$$\sum_{j=1}^{s} O_{ij} = \sum_{j=1}^{s} n_{ij} = n_{i.} \qquad \text{for} \quad i = 1, 2, \dots, r$$

$$\sum_{i=1}^{r} O_{ij} = \sum_{i=1}^{r} n_{ij} = n_{.j} \qquad \text{for} \quad j = 1, 2, \dots, s$$

of which only $r + s - 1$ are linearly independent, i.e. one of the equations depends on the others.

χ^2-test of independence of qualitative data items

The number of the degrees of freedom:

We thus have $r \times s$ observed counts O_{ij} for $i = 1, \ldots, r$ and for $j = 1, \ldots, s$

bound by $r + s - 1$ linearly independent equations, i.e. only

$$r \times s - r - s + 1 \;=\; (r-1) \times (s-1)$$

of the observed counts are free.

Therefore, the number of the degrees of freedom is

$$(r-1)(s-1)$$

χ^2-test of independence of qualitative data items

Now, finish the χ^2-test of independence of qualitative data items

(H_0: the categories of A and B are independent of each other) as follows:

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$
- find **the critical value** $c > 0$ so that

$$\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$$

  where f is the density of the χ^2-distribution with $(r-1)(s-1)$ d.f.

- if $X^2 \geq c$, **the critical region**, then **reject** the null hypothesis
- if $X^2 < c$, then **do not reject** (or fail to reject) the null hypothesis