

Statistical Methods for Economists

Lecture 2 & 3

Hypothesis testing:
Parametric and Non-parametric tests
(in marketing and elsewhere)



**SILESIA
UNIVERSITY**

SCHOOL OF BUSINESS
ADMINISTRATION IN KATOWICE

David Bartl

Statistical Methods for Economists
INM/BASTE

Outline of the lecture



- About statistical hypothesis testing
 - PARAMETRIC TESTS
 - t -tests for the means
 - Two sample F -test for the equality of variances
 - 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



- The general outline of a statistical hypothesis test
- The p -value of a test
- Parametric and Non-parametric tests

The general outline of a statistical hypothesis test



- A statistical test consists in the study of the outcomes of a random experiment.
 - We put down a hypothesis about the probability distribution of the outcomes of the random experiment.
 - We also make up a statistic S – a formula, i.e. a mathematical expression – and we **prove** (!) as a mathematical Theorem that, under our hypotheses, the statistic S follows a certain probability distribution.
 - We carry out the random experiment several (or many) times.
 - We put down the results of the experiment, i.e., count positive results, count the negative results, and so on.
-

The general outline of a statistical hypothesis test



- We substitute the results (the counts, and so on) into the mathematical expression-statistic S , which is a random variable thus (its value depends on the results of the random experiment).
- We then choose the **significance level** α – a small probability – such as **$\alpha = 5 \% = 0.05$** (i.e. “one error per twenty trials”).
(Other popular choices include $\alpha = 10 \% = 0.1$ or $\alpha = 1 \% = 0.01$.)
- By using the mathematical Theorem, which we proved (see above), we find the **critical region** $C \subseteq \mathbb{R}$ so that – if our hypotheses are true – then the **probability of the event that $S \in C$ is $\leq \alpha$** .

The general outline of a statistical hypothesis test



- The critical region C is usually a closed interval or the union of two closed intervals.
- Finally, make a statistical conclusion:
- If $S \in C$, then **reject** the hypothesis.
- If the hypotheses are true, then it is quite improbable that $S \in C$; the probability is $\leq \alpha$. So we are making a mistake – type I error, i.e. rejecting a hypothesis which is true – about once per twenty trials, if $\alpha = 5 \%$.

The general outline of a statistical hypothesis test



- If $S \notin C$, then **do not reject** (or **fail to reject**) the hypothesis.
- The fact that we fail to reject the hypothesis is not a confirmation that the hypothesis is true!
- Since the statistic S is a random variable, it may happen by chance that $S \notin C$ even if the hypothesis is false.
- This situation – failing to reject a false hypothesis – is a type II error.

The probability of type II error is β , and this probability is difficult to calculate...

If $\alpha = 5\%$, then the probability β should be $\leq 20\%$. (Is it $\leq 20\%$?)

The probability $1 - \beta$ is the **power of the test**.

The p -value of the test



The above outline of the test is as follows:

- Choose the significance level α (such as $\alpha = 5\%$).
- Depending upon the α , find the critical region $C_\alpha \subseteq \mathbb{R}$ so that –
if the hypothesis is true – then the probability that $(S \in C_\alpha)$ is $\leq \alpha$.
- Carry out the experiment, enumerate the expression S , and see if $S \in C_\alpha$.

Another procedure:

- Carry out the experiment and enumerate the expression S .
- Find the least number $p \in (0, 1)$ such that $S \in C_p$.
- This value p is the **p -value of the test**.

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.
-

PARAMETRIC TESTS

- t -tests for the means
- Two sample F -test
for the equality of variances

t -tests for the means

- One-sample t -test for the population mean
- Paired-sample t -test for the difference of the population means
- Two-sample t -test for the difference of the population means

One-sample t -test for the population mean



Theorem:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$

where

- $\bar{X} = \sum_{i=1}^n X_i / n$

is the sample mean of the random variables

- $\sigma = \sqrt{\sigma^2}$

is the standard deviation of the random variables

- $\mathcal{N}(0, 1)$

is the standard normal distribution

One-sample t -test for the population mean



Theorem:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

where

- $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ is the sample variance of the random variables
- χ_{n-1}^2 is Pearson's χ^2 -distribution with $n - 1$ degrees of freedom

One-sample t -test for the population mean



Theorem – Corollary:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

where

- $\bar{X} = \sum_{i=1}^n X_i / n$ is the sample mean of the random variables
- $s = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)}$ is the sample standard deviation of the random variables
- t_{n-1} is Student's t -distribution with $n - 1$ degrees of freedom

Theorem – Corollary – Proof:



If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{s} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \times \frac{\sqrt{n-1} \sigma/s}{\sqrt{n-1}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}}} \sim t_{n-1}$$

by the definition of Student's t -distribution

$$\frac{Z}{\sqrt{\frac{X_{n-1}^2}{n-1}}} \sim t_{n-1} \quad \text{if } Z \sim \mathcal{N}(0, 1) \text{ and } X_{n-1}^2 \sim \chi_{n-1}^2$$

(having used the preceding two Theorems before).

One-sample t -test for the population mean



Theorem – Corollary:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

Example or motivation:

Assume that $X \sim \mathcal{N}(\mu, \sigma^2)$ is a random variable following the normal probability distribution with some mean $\mu \in \mathbb{R}$ and with some variance $\sigma^2 \in \mathbb{R}_0^+$.

Knowing neither the variance σ^2 nor the true value of the population mean $\mu \in \mathbb{R}$,
we conjecture / we assume / we ... / that the population mean $\mu = \mu_0$, i.e.
the (unknown) population mean μ is equal to some prescribed value $\mu_0 \in \mathbb{R}$.

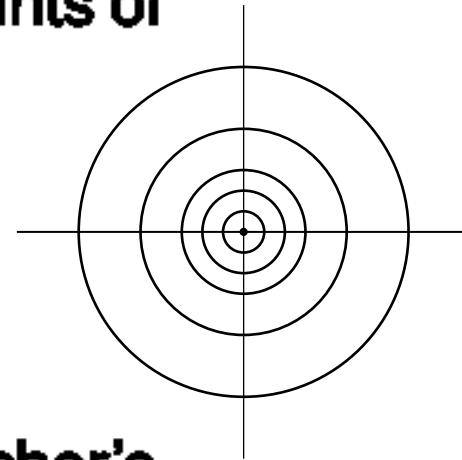
One-sample t -test for the population mean



Example: An archer shoots an arrow against the plane.

The sample space $\Omega = \mathbb{R}^2 = \{ [x, y] : x, y \in \mathbb{R} \}$ is the set of all the points of the plane. The random variable X is the x -coordinate of the hit, i.e.

$$X(\omega) = X([x, y]) = x.$$



We do not know the archer's variance σ^2 and we do not know the archer's intention, i.e. we do not know the point which the archer intends to hit, i.e. we do not know the archer's mean μ .

We conjecture that the archer's intention is to hit the origin, i.e. $\mu = 0$.

One-sample t -test for the population mean



Let $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, ..., $x_n = X(\omega_n)$ be the numerical results of n trials of a random experiment, where $X \sim \mathcal{N}(\mu, \sigma^2)$, such as the x -coordinates of the archer's n hits.

We do not know the variance σ^2 and we do not know the mean μ .

We state the null hypothesis (about the mean):

$$H_0: \mu = \mu_0$$

where $\mu_0 \in \mathbb{R}$ is some number such that we conjecture that the true mean could equal the μ_0 .

One-sample t -test for the population mean



The meaning of the null hypothesis (such as $H_0: \mu = \mu_0$ in our example) is that

- the observed distinct values are caused by the randomness only
(according to the assumed distribution, such as $X \sim \mathcal{N}(\mu, \sigma^2)$ in our example)
- there are no other factors causing the distinct values
- everything is all right, no need to reconfigure anything
- all factors under the consideration are equivalent (have the same effect)

One-sample t -test for the population mean



Having stated the **null hypothesis**

$$H_0: \mu = \mu_0$$

we also state the **alternative hypothesis**:

- **two-sided:** $H_1: \mu \neq \mu_0$
- **one-sided:** $H_1: \mu < \mu_0$
- **one-sided:** $H_1: \mu > \mu_0$

One-sample t -test for the population mean



Which alternative hypothesis ($\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$) do we choose?

→ That depends upon our knowledge of the situation.

In our example:

- If we suspect that the archer's intention is to hit a point different from the given point (such as the origin), we choose $H_1: \mu \neq \mu_0$.
- If we conjecture that the archer's intention is to hit a point to the left of the given point (such as the origin), we choose $H_1: \mu < \mu_0$.
- If we conjecture that the archer's intention is to hit a point to the right of the given point (such as the origin), we choose $H_1: \mu > \mu_0$.

One-sample t -test for the population mean



Under our assumptions $(x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2))$ are independent and $\mu = \mu_0$, it follows by the Theorem that

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Thus, having the n measurements $x_1, x_2, \dots, x_n$, we calculate the statistic

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

We know (or assume) that $T \sim t_{n-1}$.

One-sample t -test for the population mean



We have $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$.

Then, if $-\infty \leq a < b \leq +\infty$, the probability that $a < T < b$ is

$$P(a < T < b) = \int_a^b f(x) dx$$

where

$$f(x) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right) \sqrt{\pi(n-1)}} \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

is the density of Student's t -distribution with $n - 1$ degrees of freedom.

The gamma function



$$\Gamma(z) = \int_0^{+\infty} x^z e^{-x} dx \quad \text{for } z \in \mathbb{C} \text{ such that } \operatorname{Re}(z) > 0$$

It is easy to calculate:

$$\Gamma(1) = 1$$

$$\Gamma(z + 1) = z\Gamma(z)$$

Therefore:

$$\Gamma(n + 1) = n! \quad \text{for } n = 0, 1, 2, 3, \dots$$

The gamma function – another definition (due to Euler)



$$\Gamma(z) = \frac{1}{z} \prod_{n=1}^{\infty} \frac{\left(1 + \frac{1}{n}\right)^z}{1 + \frac{z}{n}} \quad \text{for} \quad z \in \mathbb{C} \setminus \{0, -1, -2, -3, \dots\}$$

One-sample t -test for the population mean



Consider the first case ($H_1: \mu \neq \mu_0$) first. We have:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Knowing that $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, the probability

$$P(-c < T < +c)$$

is quite high,

so having $-c < T < +c$ accords with H_0 , if $c > 0$ is large enough.

One-sample t -test for the population mean



On the other hand, if H_0 is true, then it is quite improbable that $T \notin (-c, +c)$.

Therefore, if we observe that

$$T \leq -c \quad \text{or} \quad +c \leq T$$

then we may conclude that H_0 is probably not true,

i.e. we reject the null hypothesis H_0 .

Therefore, the statistical test proceeds as follows:
(see below)

One-sample t -test for the population mean



Statistical one-sample t -test with two-sided alternative hypothesis ($\mu \neq \mu_0$):

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

$$\int_{-\infty}^{-c} f(x) dx + \int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $T \in (-c, +c)$, then **do not reject** (or **fail to reject**) the null hypothesis

Type I and Type II error



There are exactly four possibilities when testing the null hypothesis H_0 :

- **the null hypothesis (H_0) is actually true & we do not reject it — OK**
- **the null hypothesis (H_0) is actually true & we reject it — type I error**
- **the null hypothesis (H_0) is actually not true & we do not reject it — type II error**
- **the null hypothesis (H_0) is actually not true & we reject it — OK**

The purpose is that the probability of the type I error and that of the type II error is as little as possible.

Type I and Type II error



**What is the probability of type I error
(the null hypothesis (H_0) is actually true & we reject it)?**

The probability is equal to the significance level α , usually $\alpha = 5\%$.

Recall: The null hypothesis H_0 is rejected if and only if

$T \in (-\infty, -c] \cup [+c, +\infty)$, i.e. if and only if $|T| \geq c$.

**The critical value c is such that – if H_0 holds true – then $P(|T| \geq c) = \alpha$,
i.e. the probability of the type I error (rejecting H_0 when it is true) is α .**

Type I and Type II error



**What is the probability of type II error
(the null hypothesis (H_0) is actually false & we fail to reject it)?**

The probability of type II error is denoted by β .

The power of the test is the probability $1 - \beta$

It is much more difficult to calculate the probability β of type II error.

It must be calculated for each test separately.

Type I and Type II error



To calculate the probability β of type II error, consider that the null hypothesis

H_0 is not true ($\mu \neq \mu_0$) and we fail to reject it ($|T| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| < c$).

By the Theorem then, we have $\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{x - \mu_0 + (\mu_0 - \mu)}{s/\sqrt{n}} \sim t_{n-1}$.

Then the probability of the type II error (H_0 not true & fail to reject it) is:

$$\begin{aligned} P\left(-c < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < +c\right) &= P\left(-c < \frac{\bar{x} - \mu + (\mu_0 - \mu)}{s/\sqrt{n}} < +c\right) = \\ &= P\left(\frac{\mu - \mu_0}{s/\sqrt{n}} - c < \frac{\bar{x} - \mu}{s/\sqrt{n}} < \frac{\mu - \mu_0}{s/\sqrt{n}} + c\right) = \\ &= \beta = \int_{(\mu - \mu_0)/(s/\sqrt{n}) - c}^{(\mu - \mu_0)/(s/\sqrt{n}) + c} f(x) dx \end{aligned}$$

Type I and Type II error



Notice that, if the true μ is close to the hypothesized μ_0 ($\mu \approx \mu_0$), then $\frac{\mu - \mu_0}{s/\sqrt{n}} \approx 0$,
hence

$$\beta = \int_{(\mu - \mu_0)/(s/\sqrt{n}) - c}^{(\mu - \mu_0)/(s/\sqrt{n}) + c} f(x) dx \approx \int_{-c}^{+c} f(x) dx = 1 - \alpha = 95 \%$$

if $\alpha = 5 \%$, say.

It is recommended that β should be $\leq 20 \%$.

Therefore, if we wish to have $\beta \approx 20 \%$ or $\beta \leq 20 \%$,

then we must not consider the true μ close to the hypothesized μ_0 .

Type I and Type II error: Summary



There are exactly four possibilities when testing the null hypothesis H_0 :

- the null hypothesis (H_0) is actually true & we do not reject it — OK
- the null hypothesis (H_0) is actually true & we reject it — type I error
- the null hypothesis (H_0) is actually not true & we do not reject it — type II error
- the null hypothesis (H_0) is actually not true & we reject it — OK

The probability of the type I error is the **significance level** α

The probability of the type II error is β

The **power of the test** is the probability $1 - \beta$

One-sample t -test for the population mean



Consider now the second case ($H_1: \mu < \mu_0$). We have:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Knowing that $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$ and assuming that $c > 0$ is large enough,

what is the probability that $-c < T$?

If H_0 is true, then it is quite improbable that $T \notin (-c, +\infty)$.

Therefore, if we observe that $T \leq -c$, then we may conclude that

H_0 is probably not true, i.e. we reject the null hypothesis H_0 .

One-sample t -test for the population mean



Statistical one-sample t -test with one-sided alternative hypothesis ($\mu < \mu_0$):

- 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

$$\int_{-\infty}^{-c} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in (-\infty, -c]$, the **critical region**, then reject the null hypothesis
- if $T \in (-c, +\infty)$, then do not reject (or fail to reject) the null hypothesis

One-sample t -test for the population mean



Consider finally the third case ($H_1: \mu > \mu_0$). We have:

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Knowing that $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$ and assuming that $c > 0$ is large enough,

what is the probability that $T < +c$?

If H_0 is true, then it is quite improbable that $T \notin (-\infty, +c)$.

Therefore, if we observe that $+c \leq T$, then we may conclude that

H_0 is probably not true, i.e. we reject the null hypothesis H_0 .

One-sample t -test for the population mean



Statistical one-sample t -test with one-sided alternative hypothesis ($\mu > \mu_0$):

- 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

$$\int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $T \in (-\infty, +c)$, then **do not reject** (or **fail to reject**) the null hypothesis

Paired-sample t -test for the difference of the pop.means



Example or 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 is to learn whether the treatment has any effect.

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

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

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

Paired-sample t -test for the difference of the pop.means



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

We assume that $X \sim \mathcal{N}(\mu^{\text{before}}, \sigma_X^2)$, i.e. the random variable of the measurement before the treatment follows the normal distribution, and that $Y \sim \mathcal{N}(\mu^{\text{after}}, \sigma_Y^2)$, i.e. the random variable of the measurement after the treatment also follows the normal distribution, for some $\mu^{\text{before}}, \mu^{\text{after}} \in \mathbb{R}$ and for some $\sigma_X^2, \sigma_Y^2 \in \mathbb{R}_0^+$.

We do not know the true values of the population means μ^{before} and μ^{after} , and we do not know the true values of the variances σ_X^2 and σ_Y^2 .

Paired-sample t -test for the difference of the pop.means



We formulate the **null hypothesis**:

the treatment has no effect, i.e. the population means are the same

$$H_0: \mu^{\text{before}} = \mu^{\text{after}}$$

Recall that we do not know the true population means μ^{before} and μ^{after} . We only test the hypothesis by having done a sample of n pairs of measurements.

Formulate the alternative hypothesis:

- two-sided: $H_1: \mu^{\text{before}} \neq \mu^{\text{after}}$ (the treatment has some effect)
- one-sided: $H_1: \mu^{\text{before}} < \mu^{\text{after}}$ (the treatment increases / ...
- one-sided: $H_1: \mu^{\text{before}} > \mu^{\text{after}}$... / decreases the quantity)

Paired-sample t -test for the difference of the pop.means



Recall the theorem:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

Notice also:

If $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu^{\text{before}}, \sigma_X^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu^{\text{after}}, \sigma_Y^2)$ are independent, then the differences

$$X_1 - Y_1, X_2 - Y_2, \dots, X_n - Y_n \sim \mathcal{N}(\mu^{\text{before}} - \mu^{\text{after}}, \sigma_X^2 + \sigma_Y^2)$$

Now, the hypothesis $\mu^{\text{before}} = \mu^{\text{after}}$ is equivalent to that the mean of the difference $X - Y$ is $\mu = \mu_0 = 0$.

Paired-sample t -test for the difference of the pop.means



We have thus

reduced

the paired-sample t -test for the difference of the population means

to

the one-sample t -test for the population mean,

which we already know.

Paired-sample t -test for the difference of the pop.means



Having the n pairs of the measurements $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$, calculate the statistic

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s^2}/\sqrt{n}} \quad \text{or} \quad T = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{s^2}/\sqrt{n}}$$

where

- $\bar{x} = (\sum_{i=1}^n x_i)/n$ is the sample mean of the measurements before the treatment
- $\bar{y} = (\sum_{i=1}^n y_i)/n$ is the sample mean of the measurements after the treatment
- $\mu_0 = 0$ for no difference of the means ($\mu^{\text{before}} = \mu^{\text{after}}$)
- $\mu_0 = \text{const.}$ for a general difference of the means ($\mu^{\text{before}} = \mu^{\text{after}} + \text{const.}$)
- $s^2 = (\sum_{i=1}^n (x_i - y_i - \bar{x} + \bar{y})^2)/(n - 1)$ is the sample variance of the differences

Paired-sample t -test for the difference of the pop.means



In the first case ($H_1: \mu^{\text{before}} \neq \mu^{\text{after}}$), we have:

$$H_0: \mu^{\text{before}} = \mu^{\text{after}}$$

$$H_1: \mu^{\text{before}} \neq \mu^{\text{after}}$$

Knowing that

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s^2/\sqrt{n}}} \sim t_{n-1}$$

we fail to reject H_0 iff

$$-c < T < +c$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that

$P(-c < T < +c) = 1 - \alpha$ where the probability α of type I error is small.

Paired-sample t -test for the difference of the pop.means



Statistical paired-sample t -test for the difference of the population means with two-sided alternative hypothesis ($\mu^{\text{before}} \neq \mu^{\text{after}}$):

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

$$\int_{-\infty}^{-c} f(x) dx + \int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then reject the null hypothesis
- if $T \in (-c, +c)$, then do not reject (or fail to reject) the null hypothesis

Paired-sample t -test for the difference of the pop.means



In the second case ($H_1: \mu^{\text{before}} < \mu^{\text{after}}$), we have:

$$H_0: \mu^{\text{before}} = \mu^{\text{after}}$$

$$H_1: \mu^{\text{before}} < \mu^{\text{after}}$$

Knowing that

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s^2}/\sqrt{n}} \sim t_{n-1}$$

we fail to reject H_0 iff

$$-c < T$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that $P(-c < T) = 1 - \alpha$ where the probability α of type I error is small.

Paired-sample t -test for the difference of the pop.means



Statistical paired-sample t -test for the difference of the population means with one-sided alternative hypothesis ($\mu^{\text{before}} < \mu^{\text{after}}$):

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

$$\int_{-\infty}^{-c} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in (-\infty, -c]$, the **critical region**, then **reject** the null hypothesis
- if $T \in (-c, +\infty)$, then **do not reject** (or **fail to reject**) the null hypothesis

Paired-sample t -test for the difference of the pop.means



In the third case ($H_1: \mu^{\text{before}} > \mu^{\text{after}}$), we have:

$$H_0: \mu^{\text{before}} = \mu^{\text{after}}$$

$$H_1: \mu^{\text{before}} > \mu^{\text{after}}$$

Knowing that

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s^2}/\sqrt{n}} \sim t_{n-1}$$

we fail to reject H_0 iff

$$T < +c$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that $P(T < +c) = 1 - \alpha$ where the probability α of type I error is small.

Paired-sample t -test for the difference of the pop.means



Statistical paired-sample t -test for the difference of the population means with one-sided alternative hypothesis ($\mu^{\text{before}} > \mu^{\text{after}}$):

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

$$\int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $n - 1$ degrees of freedom

- if $T \in [+c, +\infty)$, **the critical region**, then reject the null hypothesis
- if $T \in (-\infty, +c)$, then do not reject (or fail to reject) the null hypothesis

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Motivation:

We have two unknown random variables X and Y . We ask (test the hypothesis) whether the population means of both random variables are the same.

We assume that both random variables are normal, i.e. $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, for some $\mu_X, \mu_Y \in \mathbb{R}$ and for some $\sigma_X^2, \sigma_Y^2 \in \mathbb{R}_0^+$.

Although we do not know the means μ_X, μ_Y nor the variances σ_X^2, σ_Y^2 , we assume that

$$\text{||| ||| |||} \quad \sigma_X^2 = \sigma_Y^2 \quad \text{!!! !!! !!!}$$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Having the m samples $x_1, x_2, \dots, x_m$ of the random variable $X \sim \mathcal{N}(\mu_X, \sigma^2)$ and having the n samples $y_1, y_2, \dots, y_n$ of the random variable $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, we formulate the **null hypothesis**:

both samples come from the same population:
the values of the population means are the same

$$H_0: \mu_X = \mu_Y$$

Recall that we do not know the true population means μ_X and μ_Y . **We only test the hypothesis by the means of two samples of m and n measurements with the same variance.**

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Having the m samples $x_1, x_2, \dots, x_m$ of the random variable $X \sim \mathcal{N}(\mu_X, \sigma^2)$,
the n samples $y_1, y_2, \dots, y_n$ of the random variable $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, and

$$H_0: \mu_X = \mu_Y$$

formulate the alternative hypothesis:

- two-sided: $H_1: \mu_X \neq \mu_Y$ (the means are different)
- one-sided: $H_1: \mu_X < \mu_Y$ (the first mean < the second mean)
- one-sided: $H_1: \mu_X > \mu_Y$ (the first mean > the second mean)

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



By the Theorem:

If $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma^2)$ are independent, then

$$\frac{\bar{X} - \mu_X}{\sigma/\sqrt{m}} \sim \mathcal{N}(0, 1) \quad \text{and} \quad \frac{\bar{Y} - \mu_Y}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

equivalently

$$\bar{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma^2}{m}\right) \quad \text{and} \quad \bar{Y} \sim \mathcal{N}\left(\mu_Y, \frac{\sigma^2}{n}\right)$$

Therefore:

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right)$$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



We have shown:

If $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma^2)$ are independent, then

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right)$$

equivalently

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0, 1)$$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



The above Theorem says:

If $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma^2)$ are independent, then

$$\frac{(m-1)s_X^2}{\sigma^2} \sim \chi_{m-1}^2 \quad \text{and} \quad \frac{(n-1)s_Y^2}{\sigma^2} \sim \chi_{n-1}^2$$

Therefore:

$$\frac{(m-1)s_X^2 + (n-1)s_Y^2}{\sigma^2} \sim \chi_{m+n-2}^2$$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Recall also that, if

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0, 1)$$

and

$$Y = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{\sigma^2} \sim \chi_{m+n-2}^2$$

then

$$T = \frac{Z}{\sqrt{\frac{Y}{m+n-2}}} \sim t_{m+n-2}$$

by the definition of Student's t -distribution.

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Therefore:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{\sigma^2}}{m+n-2}}} \sim t_{m+n-2}$$

where

$$s_X^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2}{m-1} \quad \text{and} \quad s_Y^2 = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2}{n-1}$$

are the sample variances.

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Having the m measurements $x_1, x_2, \dots, x_m$ and n measurements $y_1, y_2, \dots, y_n$, recall that

- $\bar{x} = \sum_{i=1}^m x_i / m$ is the sample mean of the first sample
- $\bar{y} = \sum_{j=1}^n y_j / n$ is the sample mean of the second sample
- $s_x^2 = \sum_{i=1}^m (x_i - \bar{x})^2 / (m - 1)$ is the sample variance of the first sample
- $s_y^2 = \sum_{j=1}^n (y_j - \bar{y})^2 / (n - 1)$ is the sample variance of the second sample
- m is the size of the first sample
- n is the size of the second sample

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Having the m measurements $x_1, x_2, \dots, x_m$ and n measurements $y_1, y_2, \dots, y_n$, calculate the statistic

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

for no difference of the means ($\mu_X = \mu_Y$)

We know (or assume) that $T \sim t_{m+n-2}$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Or, having the m measurements $x_1, x_2, \dots, x_m$ and n measurements $y_1, y_2, \dots, y_n$, calculate the statistic

$$T = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

for a general difference of the means ($\mu_X = \mu_Y + \mu_0$)

We know (or assume) that $T \sim t_{m+n-2}$

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



In the first case ($H_1: \mu_X \neq \mu_Y$), we have:

$$H_0: \mu_X = \mu_Y$$

$$H_1: \mu_X \neq \mu_Y$$

Knowing that

$$T \sim t_{m+n-2}$$

we fail to reject H_0 iff

$$-c < T < +c$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that $P(-c < T < +c) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Statistical two-sample t -test for the difference of the population means with two-sided alternative hypothesis ($\mu_X \neq \mu_Y$) and with the same variances:

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

$$\int_{-\infty}^{-c} f(x) dx + \int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $m + n - 2$ degrees of freedom

- if $T \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $T \in (-c, +c)$, then **do not reject** (or **fail to reject**) the null hypothesis

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



In the second case ($H_1: \mu_X < \mu_Y$), we have:

$$H_0: \mu_X = \mu_Y$$

$$H_1: \mu_X < \mu_Y$$

Knowing that

$$T \sim t_{m+n-2}$$

we fail to reject H_0 iff

$$-c < T$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that $P(-c < T) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Statistical two-sample t -test for the difference of the population means with one-sided alternative hypothesis ($\mu_X < \mu_Y$) and with the same variances :

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

$$\int_{-\infty}^{-c} f(x) dx = \alpha$$

where f is the density of the t -distribution with $m + n - 2$ degrees of freedom

- if $T \in (-\infty, -c]$, the **critical region**, then reject the null hypothesis
- if $T \in (-c, +\infty)$, then do not reject (or fail to reject) the null hypothesis

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



In the third case ($H_1: \mu_X > \mu_Y$), we have:

$$H_0: \mu_X = \mu_Y$$

$$H_1: \mu_X > \mu_Y$$

Knowing that

$$T \sim t_{m+n-2}$$

we fail to reject H_0 iff

$$T < +c$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that $P(T < +c) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample t -test for the diff. of the pop. means // $\sigma_X = \sigma_Y$



Statistical two-sample t -test for the difference of the population means with one-sided alternative hypothesis ($\mu_X > \mu_Y$):

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

$$\int_{+c}^{+\infty} f(x) dx = \alpha$$

where f is the density of the t -distribution with $m + n - 2$ degrees of freedom

- if $T \in [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $T \in (-\infty, +c)$, then **do not reject** (or **fail to reject**) the null hypothesis

Two-sample t -test for the diff. of the pop. means // $\sigma_X \neq \sigma_Y$



Consider two normal random variables $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$,
for some $\mu_X, \mu_Y \in \mathbb{R}$ and for some $\sigma_X^2, \sigma_Y^2 \in \mathbb{R}_0^+$.

We ask (test the hypothesis) whether the population means of both random variables are the same.

Once $\sigma_X^2 = \sigma_Y^2$ is not assumed, the things get complicated.

We have an approximate result only.

Theorem (Satterthwaite's approximation):



If the random variables $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then the statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \sim t_v \quad \text{approximately}$$

where

$$v = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{s_X^4}{m^2(m-1)} + \frac{s_Y^4}{n^2(n-1)}}$$

Two-sample t -test for the diff. of the pop. means // $\sigma_x \neq \sigma_y$



Exercise:

Use the last Theorem (Satterthwaite's approximation) to formulate a statistical two-sample t -test for the difference of the population means with two-sided / one-sided alternative hypothesis (not assuming the same variance).

Two sample F -test for the equality of variances

Motivation



We have two unknown random variables X and Y . We ask (test the hypothesis) whether the population variances of both random variables are the same.

We assume that both random variables are normal, i.e. $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, for some $\mu_X, \mu_Y \in \mathbb{R}$ and for some $\sigma_X^2, \sigma_Y^2 \in \mathbb{R}_0^+$.

We ask (test the hypothesis) whether

$$? \quad \sigma_X^2 = \sigma_Y^2 \quad ?$$

Two sample F -test for the equality of variances



Having the m samples $x_1, x_2, \dots, x_m$ of the random variable $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and having the n samples $y_1, y_2, \dots, y_n$ of the random variable $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, we formulate the **null hypothesis**:

both samples come from populations with the same variances:

$$H_0: \sigma_X^2 = \sigma_Y^2$$

Recall that we do not know the true population variances σ_X^2 and σ_Y^2 .

We only test the hypothesis by the means of the two samples of m and n independent measurements.

Two sample *F*-test for the equality of variances



The meaning of the null hypothesis (such as $H_0: \sigma_X^2 = \sigma_Y^2$ in our example) is that

- **the observed distinct values are caused by the randomness only (according to the assumed distribution, such as $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 = \sigma_X^2 = \sigma_Y^2$, in our example)**
- **there are no other factors causing the distinct values**
- **everything is all right, no need to reconfigure anything**
- **all factors under the consideration are equivalent (have the same effect)**

Two sample *F*-test for the equality of variances



Having stated the **null hypothesis**

$$H_0: \sigma_X^2 = \sigma_Y^2$$

we also state the **alternative hypothesis**:

- **two-sided:** $H_1: \sigma_X^2 \neq \sigma_Y^2$
- **one-sided:** $H_1: \sigma_X^2 < \sigma_Y^2$
- **one-sided:** $H_1: \sigma_X^2 > \sigma_Y^2$

Theorem



If the random variables $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent and

$$\sigma_X^2 = \sigma_Y^2$$

then

$$\frac{s_X^2}{s_Y^2} \sim F_{m-1, n-1}$$

where

$F_{m-1, n-1}$ is Fisher's F -distribution with $m - 1$ and $n - 1$ degrees of freedom

$s_X^2 = \sum_{i=1}^m (X_i - \bar{X})^2 / (m - 1)$ is the sample variance of the first sample

$s_Y^2 = \sum_{j=1}^n (Y_j - \bar{Y})^2 / (n - 1)$ is the sample variance of the second sample

Two-sample F -test for the equality of variances



Having the m measurements $x_1, x_2, \dots, x_m$ and n measurements $y_1, y_2, \dots, y_n$, calculate the statistic

$$F = \frac{s_x^2}{s_y^2}$$

where

- $s_x^2 = \sum_{i=1}^m (x_i - \bar{x})^2 / (m - 1)$ is the sample variance of the first sample
- $s_y^2 = \sum_{j=1}^n (y_j - \bar{y})^2 / (n - 1)$ is the sample variance of the second sample
- $\bar{x} = \sum_{i=1}^m x_i / m$ is the sample mean of the first sample
- $\bar{y} = \sum_{j=1}^n y_j / n$ is the sample mean of the second sample
- m and n is the size of the first and second, respectively, sample

Two-sample F -test for the equality of variances



In the first case ($H_1: \sigma_X^2 \neq \sigma_Y^2$), we have:

$$H_0: \sigma_X^2 = \sigma_Y^2$$

$$H_1: \sigma_X^2 \neq \sigma_Y^2$$

Knowing that

$$F = \frac{s_x^2}{s_y^2} \sim F_{m-1, n-1}$$

we fail to reject H_0 iff

$$c < F < d$$

where the critical value $d > c > 0$, under the assumption that H_0 is true, are such that $P(c < F < d) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample F -test for the equality of variances



Statistical two-sample F -test for the equality of the population variances with two-sided alternative hypothesis ($\sigma_X^2 \neq \sigma_Y^2$):

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

$$\int_0^c f(x) dx = \frac{\alpha}{2} \quad \text{and} \quad \int_d^{+\infty} f(x) dx = \frac{\alpha}{2}$$

where f is the density of the F -distribution with $m - 1$ and $n - 1$ d.f.

- if $F \in [0, c] \cup [d, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $F \in (c, d)$, then **do not reject** (or **fail to reject**) the null hypothesis

Two-sample F -test for the equality of variances



In the second case ($H_1: \sigma_X^2 < \sigma_Y^2$), we have:

$$H_0: \sigma_X^2 = \sigma_Y^2$$

$$H_1: \sigma_X^2 < \sigma_Y^2$$

Knowing that

$$F = \frac{s_x^2}{s_y^2} \sim F_{m-1, n-1}$$

we fail to reject H_0 iff

$$c < F$$

where the critical value $c > 0$, under the assumption that H_0 is true, is such that

$P(c < F < +\infty) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample F -test for the equality of variances



Statistical two-sample F -test for the equality of the population variances with one-sided alternative hypothesis ($\sigma_X^2 < \sigma_Y^2$):

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

$$\int_0^c f(x) dx = \alpha$$

where f is the density of the F -distribution with $m - 1$ and $n - 1$ d.f.

- if $F \in [0, c]$, the **critical region**, then **reject** the null hypothesis
- if $F \in (c, +\infty)$, then **do not reject** (or **fail to reject**) the null hypothesis

Two-sample F -test for the equality of variances



In the third case ($H_1: \sigma_X^2 > \sigma_Y^2$), we have:

$$H_0: \sigma_X^2 = \sigma_Y^2$$

$$H_1: \sigma_X^2 > \sigma_Y^2$$

Knowing that

$$F = \frac{s_x^2}{s_y^2} \sim F_{m-1, n-1}$$

we fail to reject H_0 iff

$$F < d$$

where the critical value $d > 0$, under the assumption that H_0 is true, is such that $P(0 \leq F < d) = 1 - \alpha$ where the probability α of type I error is small.

Two-sample F -test for the equality of variances



Statistical two-sample F -test for the difference of the population variances with one-sided alternative hypothesis ($\sigma_X^2 > \sigma_Y^2$):

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

$$\int_d^{+\infty} f(x) dx = \alpha$$

where f is the density of the F -distribution with $m - 1$ and $n - 1$ d.f.

- if $F \in [d, +\infty)$, the **critical region**, then **reject** the null hypothesis
- if $Z \in [0, d)$, then **do not reject** (or **fail to reject**) the null hypothesis

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

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 samples $x_1, x_2, \dots, x_n$ 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

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

be the number of the negative differences.

Theorem:

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

$$Z \sim \text{Bi}(m, \frac{1}{2})$$

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 ($P(X < \tilde{x}_0) = P(X \leq \tilde{x}_0) = \frac{1}{4}$, whence $Z \sim \text{Bi}\left(m, \frac{1}{4}\right)$), or that

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

Sign test for the median



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

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

we also state the **alternative hypothesis**:

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

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

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

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

Sign (binomial) 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 the **critical value** $K \in \{0, 1, \dots, m\}$ so that K is the largest number such that

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

- 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 ($H_1: \tilde{x} < \tilde{x}_0$). 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 the **critical value** $L \in \{0, 1, \dots, m\}$ so that L is the least number such that

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

- 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 $-\infty \leq a < b \leq +\infty$, that

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

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

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

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

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

Sign (z-) test for the median



Consider the first case ($H_1: \tilde{x} \neq \tilde{x}_0$) 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 $c > 0$ so that

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

- 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

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

- 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 ($H_1: \tilde{x} < \tilde{x}_0$). 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

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

- if $(m + c\sqrt{m})/2 \leq Z$, 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

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 $\mu = E[X]$ 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 ($H_0: \mu = \tilde{x}_0$).

Remarks:

- More generally, if we know that the mean $\mu = E[X]$ 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 ($H_0: \mu = \tilde{x}_0$ with $Z = |\{i : x_{j_i} < \tilde{x}_0\}| \sim \text{Bi}(m, p_0)$).
- 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 is to learn whether the treatment has any effect.

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

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

let $y_1, y_2, \dots, y_n$ 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 $i = 1, 2, \dots, n$.

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.

$$Z = |\{i : x_i < y_i\}| \sim \text{Bi}(m, \frac{1}{2})$$

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.

Theorem: 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 $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, ..., $x_N = X(\omega_N)$ 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 $\text{Exp}(\lambda)$ for various $\lambda \in \mathbb{R}^+$
- etc.

Having the numerical results $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, ..., $x_N = X(\omega_N)$

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 $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, ..., $x_N = X(\omega_N)$ of the N trials of the random experiment and having the class $\mathcal{F}_0$ of the cumulative distribution functions – first of all – 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

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

(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 $\text{Exp}(\lambda)$, then put

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

the number of parameters is $\nu = 1$

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

- 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

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

as well as

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

so that

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

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



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

$$H_0: F = F_0$$

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

$$\begin{aligned} p_i &= P(t_{i-1} < X \leq t_i) = \\ &= F_0(t_i) - F_0(t_{i-1}) \quad \text{for } i = 1, 2, \dots, n \end{aligned}$$

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

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

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 $H_0: F = F_0$ is true, then the statistic

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

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 ($\nu = 0, 1, 2, \dots$)
- 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

$$\int_c^{+\infty} f(x) dx = \alpha$$

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 $q = 1 - p$.

We do not know the true probability 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 ($H_1: F \neq F_0$) 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

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

(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

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

where

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

Notice that one degree of freedom (“-1”) must always be subtracted

because the observed counts $O_1, O_2, \dots, O_n$ are bound by the equation

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

therefore only $n - 1$ of the counts (such as $O_1, O_2, \dots, O_{n-1}$, 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

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

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

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

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 $i=1, \dots, r$ & $j=1, \dots, s$

$A \setminus B$	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}	$n_{2.}$
...	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
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 $i=1,2$ & $j=1,2$

$A \setminus B$	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) $\mathbf{B}$ depends upon the category of the data item (variable) $\mathbf{A}$, or whether the categories of both data items (variables) $\mathbf{A}$ and $\mathbf{B}$ are independent of each other.

Assume therefore the null hypothesis H_0 :

the categories of both data items (variables) $\mathbf{A}$ and $\mathbf{B}$ 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 the null hypothesis 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

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

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

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

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

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

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

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

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

and observing

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

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

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

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

Theorem:

If the null hypothesis is true, then

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

**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:

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

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 $i = 1, \dots, r$ and for $j = 1, \dots, s$ 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

$$\int_c^{+\infty} f(x) dx = \alpha$$

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