10) Hypothesis Testing

Tuesday, 8 November 2022 6:24

One and Two Sided Tests

For an unknown population parameter θ (e.g. μ) and a fixed value θ_0 (e.g. 5), the following three cases have to be distinguished:

Case	Null hypothesis	Alternative hypothesis	
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	Two-sided test problem
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	One-sided test problem
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	One-sided test problem

Remark 10.2.1 Note that we have not considered the following situation: H 0 : θ =/θ 0 , H 1 : θ = θ 0 . In general, with the tests described in this chapter, we cannot prove the equality of a parameter to a predefined value and neither can we prove the equality of two parameters, as in H 0: θ 1=/θ 2, H 1: θ 1= θ 2.

Type And Type 2 Error

- The hypothesis H_0 is true but is rejected; this error is called **type I error**.
- The hypothesis H_0 is not rejected although it is wrong; this is called **type II error**.

' The **significance** level is the probability of type I error, P(H 1 | H 0) = α, which is the probability of rejecting H 0(accepting H 1) if H 0is true. '

P-Value: ' It can be interpreted as the probability of observing results equal to, or more extreme than those actually observed if the null hypothesis was true. Then, the decision rule is,
H 0 is rejected if the p-value is smaller than the prespecified significance level α. Otherwise, H 0 cannot be rejected.

Example 10.2.2 Assume that we are dealing with a two-sided test and assume further that the test statistic T (x) is N(0, 1)-distributed under H 0 . The significance level is α = 0.05. If we observe, for example, t = 3, then the p-value is
P H 0 (| T | ≥ 3). This can be calculated in R as
2*(1-pnorm(3)). We have to multiply with two because we are dealing with a two-sided hypothesis.

Test for the Mean When the Variance is Known (One-Sample Gauss Test)

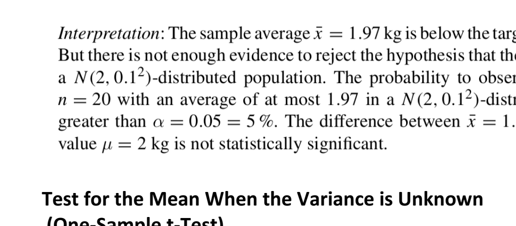
Z-test

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0}\sqrt{n}.$$

Area of extreme values for the test statistic; sufficient enough to reject H_0

Case	H_0	H_1	Critical region K	Reject H_0, if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$	$\lvert t(x)\rvert > z_{1-\alpha/2}$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, \infty)$	$t(x) < z_\alpha$
(c)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, z_\alpha = -z_{1-\alpha})$	$t(x) > z_{1-\alpha}$

For case (a) with H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$, we are interested in extreme values of the test statistic on both tails: very small values and very large values of the test statistic give us evidence that H_0 is wrong (because the statistic is mainly driven by the difference of the sample mean and the test value μ_0 for a fixed variance), see Fig. 10.1. In such a two-sided test, when the distribution of the test statistic is symmetric, we divide the critical region into two equal parts and assign each region of size $\alpha/2$ to the left and right tails of the distribution. For $\alpha = 0.05$, 2.5 % of the most extreme values towards the right end of the distribution and 2.5 % of the most extreme values towards the left end of the distribution give us enough evidence that H_0 is wrong and can be rejected and that H_1 is accepted. It is also clear why α is



Example 10.3.1 A bakery supplies loaves of bread to supermarkets. The stated selling weight (and therefore the required minimum expected weight) is μ = 2 kg. However, not every package weighs exactly 2kg because there is variability in the weights. It is therefore important to find out if the average weight of the loaves is significantly smaller than 2kg. The weight X (measured in kg) of the loaves is assumed to be normally distributed. We assume that the variance σ 0 2 = 0.1 2 is known from experience. A supermarket draws a sample of n = 20 loaves and weighs them. The average weight is calculated as ¯x = 1.97 kg. Since the supermarket wants to be sure that the weights are, on average, not lower than 2kg, a one-sided hypothesis is appropriate and is formulated as H 0 : μ ≥ μ 0 = 2 kg versus H 1 : μ < μ 0 = 2 kg. The significance level is specified as α = 0.05, and therefore, z 1−α = 1.64. The test statistic is calculated as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0}\sqrt{n} = \frac{1.97 - 2}{0.1}\sqrt{20} = -1.34.$$

The null hypothesis is not rejected, since $t(x) = -1.34 > -1.64 = -z_{1-0.05} = z_{0.05}$.

Interpretation: The sample average $\bar{x} = 1.97$ kg is below the target value of $\mu = 2$ kg. But there is not enough evidence to reject the hypothesis that the sample comes from a $N(2, 0.1^2)$-distributed population. The probability to observe a sample of size $n = 20$ with an average of at most 1.97 in a $N(2, 0.1^2)$-distributed population is greater than $\alpha = 0.05 = 5\,\%$. The difference between $\bar{x} = 1.97$ kg and the target value $\mu = 2$ kg is not statistically significant.

Test for the Mean When the Variance is Unknown (One-Sample t-Test)

unbiased estimator of σ^2 is the sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The test statistic is therefore

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S_X}\sqrt{n},$$

which follows a t-distribution with $n - 1$ degrees of freedom if H_0 is true, as we know from Theorem 8.3.2.

Critical regions and test decisions
Since $T(\mathbf{X})$ follows a t-distribution under H_0, the critical regions refer to the regions of the t-distribution which are unlikely to be observed under H_0:

Case	H_0	H_1	Critical region K	Reject H_0, if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}, \infty)$	$\lvert t(x)\rvert > t_{n-1;1-\alpha/2}$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha})$	$t(x) < -t_{n-1;1-\alpha}$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (t_{n-1;1-\alpha}, \infty)$	$t(x) > t_{n-1;1-\alpha}$

DF

The hypothesis H_0 is rejected if the realized test statistic, i.e.

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n},$$

falls into the critical region. The critical regions are based on the appropriate quantiles of the t-distribution with $(n - 1)$ degrees of freedom, as outlined in Table 10.2.

Example 10.3.2 We again consider Example 10.3.1. Now we assume that the variance of the loaves is unknown. Suppose a random sample of size $n = 20$ has an arithmetic mean of $\bar{x} = 1.9668$ and a sample variance of $s^2 = 0.0927^2$. We want to test whether this result contradicts the two-sided hypothesis H_0: $\mu = 2$, that is case (a). The significance level is fixed at $\alpha = 0.05$. For the realized test statistic $t(x)$, we calculate

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n} = \frac{1.9668 - 2}{0.0927}\sqrt{20} = -1.60.$$

In R: t.test()

Comparing the Means of Two Independent Samples

1) **Case 1: The variances are known (two-sample Gauss test).**

$$t(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

2) **Case 2: The variances are unknown, but equal (two-sample t-test).**
We denote the unknown variance of both distributions as σ^2 (i.e. both the populations are assumed to have variance σ^2). We estimate σ^2 by using the pooled sample variance where each sample is assigned weights relative to the sample size:

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}. \tag{10.3}$$

The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \tag{10.4}$$

with S as in (10.3) follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if H_0 is true. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{s}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \tag{10.5}$$

3) **Case 3: The variances are unknown and unequal (Welch test).**
We test H_0: $\mu_X = \mu_Y$ versus H_1: $\mu_X \neq \mu_Y$ given $\sigma_X^2 \neq \sigma_Y^2$ and both σ_X^2 and σ_Y^2 are unknown. This problem is also known as the Behrens–Fisher problem and is the most frequently used test when comparing two means in practice. The test statistic can be written as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}, \tag{10.6}$$

which is approximately t-distributed with v degrees of freedom:

$$v = \left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2 \Big/ \left(\frac{(s_x^2/n_1)^2}{n_1 - 1} + \frac{(s_y^2/n_2)^2}{n_2 - 1}\right) \tag{10.7}$$

Example 10.3.3 A small bakery sells cookies in packages of 500 g. The cookies are handmade and the packaging is either done by the baker himself or his wife. Some customers conjecture that the wife is more generous than the baker. One customer does an experiment: he buys packages of cookies packed by the baker and his wife on 16 different days and weighs the packages. He gets the following two samples (one for the baker, one for his wife).

Weight (wife) (X)	512	530	498	540	521	528	505	523
Weight (baker) (Y)	499	500	510	495	515	503	490	511

We want to test whether the complaint of the customers is justified. Let us start with the following simple hypotheses:

$$H_0 : \mu_x = \mu_y \quad \text{versus} \quad H_1 : \mu_x \neq \mu_y,$$

i.e. we only want to test whether the weights are different, not that the wife is making heavier cookie packages. Since the variances are unknown, we assume that case 3 is the right choice. We calculate and obtain $\bar{x} = 519.625$, $\bar{y} = 502.875$, $s_X^2 = 192.268$, and $s_Y^2 = 73.554$. The test statistic is:

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{|519.625 - 502.875|}{\sqrt{\frac{192.268}{8} + \frac{73.554}{8}}} \approx 2.91.$$

The degrees of freedom are:

$$v = \left(\frac{192.268}{8} + \frac{73.554}{8}\right)^2 \Big/ \left(\frac{(192.268/8)^2}{7} + \frac{(73.554/8)^2}{7}\right) \approx 11.67 \approx 12.$$

Since $|t(x)| = 2.91 > 2.18 = t_{12;0.975}$, it follows that H_0 is rejected. Therefore, H_1 is statistically significant. This means that the mean weight of the wife's packages is different from the mean weight of the baker's packages. Let us refine the hypothesis and try to find out whether the wife's packages have a higher mean weight. The hypotheses are now:

$$H_0 : \mu_x \leq \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y.$$

- **T-tests in R**
  Any kind of t-test can be calculated with the t.test command: for example, the two-sample t-test requires to specify the option var.equal=TRUE while the Welch test is calculated when the (default) option var.equal=FALSE is set. We can also conduct a one-sample t-test. Suppose we are interested in whether the mean

Test for Comparing the Means of Two Dependent Samples (Paired t-Test)

- They could be dependent because we measure the same variable twice on the same subjects at different times.
- Since the same variable is measured twice on the same subject, it makes sense to calculate a difference between the two respective values.
- Let D = X − Y denote the random variable "difference of X and Y"

$$T(\mathbf{X}, \mathbf{Y}) = T(\mathbf{D}) = \frac{\bar{D}}{S_D}\sqrt{n} \tag{10.9}$$

is t-distributed with $n - 1$ degrees of freedom. The sample mean is $\bar{D} = \sum_{i=1}^{n} / D_i n$ and the sample variance is

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n - 1}$$

(i) We compare the test statistic (t = −3.35) with the critical value (1.83, obtained via qt(0.95,9)).

(ii) We evaluate whether the p-value (0.008468) is smaller than the significance

level α = 0.05.

(iii) We evaluate whether the confidence interval for the mean difference covers "0" or not.

F-Test (Testing Variances)

The test statistic

$$T(\mathbf{X}) = \frac{(n-1)S_X^2}{\sigma_0^2}$$

follows a χ^2_{n-1}-distribution under H_0. The critical region is constructed by taking the $\alpha/2$- and $(1 - \alpha/2)$ quantile as critical values; i.e. H_0 is rejected, if

$$t(x) < c_{n-1;\alpha/2}$$

or if

$$t(x) > c_{n-1;1-\alpha/2},$$

where $c_{n-1;\alpha/2}$ and $c_{n-1;1-\alpha/2}$ are the desired quantiles of a χ^2-distribution. In R, the test can be called by the `sigma.test` function in the `TeachingDemos` library or the `varTest` function in library `EnvStats`. Both functions also return a confidence interval for the desired confidence level. Note that the test is biased. An unbiased level α test would not take $\alpha/2$ at the tails but two different tail probabilities α_1 and α_2 with $\alpha_1 + \alpha_2 = \alpha$.

***F*-Test for Comparing Two Variances.** Comparing variances can be of interest when comparing the variability, i.e. the "precision" of two industrial processes; or when comparing two medical treatments with respect to their reliability. Consider two populations characterized by two independent random variables X and Y which follow normal distributions:

$$X \sim N(\mu_X, \sigma_X^2), \qquad Y \sim N(\mu_Y, \sigma_Y^2).$$

For now, we distinguish the following two hypotheses:

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ versus } H_1 : \sigma_X^2 \neq \sigma_Y^2, \quad \text{two-sided}$$
$$H_0 : \sigma_X^2 \leq \sigma_Y^2 \text{ versus } H_1 : \sigma_X^2 > \sigma_Y^2, \quad \text{one-sided.}$$

The third hypothesis with $H_1 : \sigma_X^2 < \sigma_Y^2$ is similar to the second hypothesis where X and Y are replaced with each other.

Test Statistic

Let $(X_1, X_2, \ldots, X_{n_1})$ and $(Y_1, Y_2, \ldots, Y_{n_2})$ be two independent random samples of size n_1 and n_2. The test statistic is defined as the ratio of the two sample variances

$$T(\mathbf{X}, \mathbf{Y}) = \frac{S_X^2}{S_Y^2}, \tag{C.10}$$

which is, under the null hypothesis, F-distributed with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, see also Sect. 8.3.3.

χ 2-Goodness-of-Fit Test

The χ 2 -goodness-of-fit test is one of the most popular tests for testing the goodness of fit of the observed data to a distribution.
χ 2-test only works properly if k is fixed and n is large(therefore for continuous variable it is practice to group into k classes.

Test statistic.
The test statistic is defined as

$$T(\mathbf{X}) = t(x) = \chi^2 = \sum_{i=1}^{k}\frac{(N_i - np_i)^2}{np_i}. \tag{10.17}$$

Frequency in class 'i' Probability for class 'i'

Test decision.
For a significance level α, H_0 is rejected if $t(x)$ is greater than the $(1 - \alpha)$-quantile of the χ^2-distribution with $k - 1 - r$ degrees of freedom, i.e. if

$$t(x) = \chi^2 > c_{k-1-r,1-\alpha}.$$

Note that r is the number of parameters of $F_0(x)$, if these parameters are estimated from the sample. The χ^2-test statistic is only asymptotically χ^2-distributed under H_0.

Example 10.7.3 Gregor Mendel (1822–1884) conducted crossing experiments with pea plants of different shape and colour. Let us look at the outcome of a pea crossing experiment with the following results:

Crossing result	Round Yellow	Round Green	Edged Yellow	Edged Green
Observations	315	108	101	32

Mendel had the hypothesis that the four different types occur in proportions of 9:3:3:1, that is

$$p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3 = \frac{3}{16},\ p_4 = \frac{1}{16}.$$

The hypotheses are

$$H_0 : P(X = i) = p_i \quad \text{versus} \quad H_1 : P(X = i) \neq p_i, \qquad i = 1, 2, 3, 4.$$

With $n = 556$ observations, the test statistic can be calculated from the following observed and expected frequencies:

The χ^2-test statistic is calculated as

$$t(x) = \chi^2 = \frac{(315 - 312.75)^2}{312.75} + \cdots + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$

Since $\chi^2 = 0.47 < 7.815 = \chi^2_{0.95,3} = c_{0.95,3}$, the null hypothesis is not rejected. Statistically, there is no evidence that Mendel was wrong with his 9:3:3:1 assumption. In R, the test can be conducted by applying the `chisq.test` command:

χ 2-Independence Test

If we are not interested in the strength of association but rather in finding out whether there is an association at all, one can use the χ 2 - independence test.

Example 10.8.1 Consider the following contingency table. Here, X describes the educational level (1: primary, 2: secondary, 3: tertiary) and Y the preference for a specific political party (1: Party A, 2: Party B, 3: Party C). Our null hypothesis is that the two variables are independent, and we want to show the alternative hypothesis which says that there is a relationship between them.

		Y			Total
		1	2	3	
X	1	100	200	300	600
	2	100	100	100	300
	3	80	10	10	100
	Total	280	310	410	1000

For the (estimated) expected frequencies $\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$, we get

		Y		
		1	2	3
X	1	168	186	246
	2	84	93	123
	3	28	31	41

For example: m 11= 600 · 280/1000 = 168.

Test statistic.
Pearson's χ^2-test statistic was introduced in Chap. 4, Eq. (4.6). It is

$$T(\mathbf{X}, \mathbf{Y}) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - m_{ij})^2}{m_{ij}},$$

where $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ (expected absolute cell frequencies under H_0). Strictly speaking, m_{ij} are the true, unknown expected frequencies under H_0 and are estimated by $\hat{m}_{ij} = n\pi_{i+}\pi_{+j}$, such that the realized test statistic equates to

$$t(x, y) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \tag{10.19}$$

In our example:

$$t(x, y) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = \frac{(100 - 168)^2}{168} + \cdots + \frac{(10 - 41)^2}{41} \approx 182.54.$$

Since $\chi^2_{4;0.95} = 9.49 < t(x, y) = 182.54$, H_0 is rejected.

χ 2-test of homogeneity.

formulated as a test for the null hypothesis that the proportions of the binary variable are equal in several (≥ 2) groups, i.e. for a K × 2 (or 2 × K) table.

Example 10.8.2 Consider two variables X and Y, where X is describing the rating of a coffee brand with the categories "bad taste" and "good taste" and Y denotes three age subgroups, e.g. "18–25", "25–35", and "35–45". The observed data is

		Y			
		18–25	25–35	35–45	Total
X	Bad	10	30	65	105
	Good	90	70	35	195
	Total	100	100	100	300

Assume H_0 is the hypothesis that the probabilities $P(X = \text{'good'}|Y = \text{'18–25'})$, $P(X = \text{'good'}|Y = \text{'25–35'})$, and $P(X = \text{'good'}|Y = \text{'35–45'})$ are all equal. Then, we can use the function either `prop.test` or `chisq.test` in R to test this hypothesis:

GoF vs Homogeneity vs Independence Testing:

1. The "goodness-of-fit test" is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't. It answers the question: are the frequencies I observe for my *single categorical variable* consistent with my theory?
2. The "test of homogeneity" is a way of determining whether two or more DIFFERENT POPULATIONS (or GROUPS) share the same distribution of a SINGLE CATEGORICAL VARIABLE. For example, do people of different races have the same proportion of smokers to non-smokers, or do different education levels have different proportions of Democrats, Republicans, and Independent.
   H_0: $p_1 = p_2 = \cdots = p_n$ the proportion of X is the same in all the populations studied.
   H_a: At least one proportion of X is not the same.
3. The "test of independence" is a way of determining whether TWO CATEGORICAL VARIABLES are associated with one another in ONE SINGLE POPULATION. For example, we draw a single group of 200 subjects and record the number of children they have, and the number of colds they each got last year, trying to see if there is a relationship between the having children and getting colds.