476 Statistics, Spring 2022.

Lecturer: Sergey Nadtochiy.

Lecture 6. Hypothesis testing: definitions, common tests for normal sample. (Sections 10.1–10.2, 10.8–1.9)

Main difference between **hypothesis testing** and estimation: instead of finding the true value of the parameter (among many alternatives), we only aim to find out if it belongs to one subset or to the other (i.e., **choose one out of two given possibilities**).

Another difference is that we **treat the two possibilities asymmetrically**: one is taken as a default option (**null hypothesis**), unless it has to be rejected in favor of the other one (**alternative hypothesis**).

We only reject the null hypothesis if the data contradicts it significantly: i.e. if a given realization of the sample "is very unlikely". To make this precise, we choose a **test statistic** and a **rejection region**, s.t., under the null hypothesis the probability that the test statistic takes values in the rejection region is $\alpha \in (0,1)$, where $\alpha << 1$ is the chosen (confidence) level.

The smaller is α, the stronger is our a priori belief in null hypothesis: i.e., the less likely we are to reject the null hypothesis if it is correct. **Note that the terminology is counter-intuitive: we aim to make the "confidence level" α small.**.

The **input** needed to construct a hypothesis test is:

1. The nature of unknown parameter θ (i.e., mean, variance, etc.).
2. A null hypothesis H_0, which is a subset of the possible values Θ of θ.
3. An alternative hypothesis H_a, which is another subset of Θ ($H_0 \cap H_a = \emptyset$).
4. A significance level $\alpha \in (0,1)$.

The **hypothesis test** itself is given by:

1. A test statistic U.
2. A rejection region (RR), which is a subset of $\mathbb{R}$ (assuming that U is one-dimensional).

If the test statistic takes its value in the RR, we reject H_0 in favor of H_a. Otherwise, we fail to reject H_0. We only reject H_0 if it is very unlikely to observe such a value of U under the hypothesis H_0:

$$\mathbb{P}^0(U \in RR) = \alpha << 1.$$

The choice of H_a is dictated by our intuition about the goal of the test: i.e., if H_a is true, it is very important to us. The choice of H_0 is slightly different: we either choose smth. that has been used before, or smth. that is neutral for us.

The test statistic U and the rejection region RR are chosen by (i) common sense, and (ii) by computational convenience. It is important that the **distribution of U is known under H_0**.

Ex 1. *A poll is conducted to estimate the prob. p that a randomly chosen person votes for candidate Jones in the upcoming elections. The sample size is $n = 15$.*

The unknown parameter is p, which may take any value in $\Theta = (0,1)$. It is natural to choose between two hypotheses: $H_0 = \{0.5\}$ and $H_a = [0, 0.5)$.

Q 1. *If the test statistic is $U = \sum_{i=1}^{15} Y_i$ and RR=$[0, 2]$, what is the corresponding significance level α?*

Recalling that $U \sim Bin(0.5, 15)$ under H_0, we obtain

$$\alpha = \mathbb{P}^0 \left(\sum_{i=1}^{15} Y_i \leq 2 \right) \approx 0.004$$

It is clear that small α is good: it means that the probability of rejecting a correct null hypothesis is small. We can make α smaller by reducing RR. However, then, one may end up with a RR that is too conservative (i.e., too small), so that it becomes useless. To control for this, we consider β.

Def 1. *Denote the test statistic by U and the rejection region by RR. If H_0 is a singleton (i.e., H_0 is simple), then the* **type I error** *is*

$$\alpha := \mathbb{P}^0(U \in RR).$$

If H_a is a singleton (i.e., H_a is simple), then the **type II error** *is*

$$\beta := \mathbb{P}^a(U \notin RR).$$

We are typically more concerned about type I error, hence, we fix α a priori. For this reason, we typically choose H_0 as a singleton, while H_a may contain multiple values (in the latter case, type II error cannot be computed as defined above).

Ex 2. *A poll is conducted to estimate the prob. p that a randomly chosen person votes for candidate Jones in the upcoming elections. The sample size is $n = 15$.*

Assume $U = \sum_{i=1}^{15} Y_i$, $RR = [0, 2]$, $H_0 = \{0.5\}$ and $H_a = \{0.3\}$.

Q 2. *Calculate the type II error.*

Recalling that $U \sim Bin(0.3, 15)$ under H_a, we obtain

$$\beta = \mathbb{P}^a \left(\sum_{i=1}^{15} Y_i > 2 \right) \approx 0.873.$$

It is very high, which indicates that the upper boundary 2 of RR is too low. Indeed, even if Jones loses 3 to 12 in the poll, we still do not reject the null. Of course, increasing the upper bound of RR we decrease β, but at the expense of increasing α.

Ex 3. *A poll is conducted to estimate the prob. p that a randomly chosen person votes for candidate Jones in the upcoming elections. The sample size is $n = 15$.*

Assume $U = \sum_{i=1}^{15} Y_i$, $RR = [0, 2]$, $H_0 = \{0.5\}$ and $H_a = \{0.1\}$.

Q 3. *Calculate the type II error.*

Recalling that $U \sim Bin(0.1, 15)$ under H_a, we obtain

$$\beta = \mathbb{P}^a \left(\sum_{i=1}^{15} Y_i > 2 \right) \approx 0.184$$

It is not too high, which indicates that the upper boundary 2 may be okay in this case. The fact that β is smaller than in the previous example, with α being the same in both cases, indicates that it is "easier" to distinguish between the two hypotheses in the present case. This is because the two hypotheses are "farther apart" in the present case.

The values of α and β can be balanced via the choice of RR and are inversely related. Both can be reduced simultaneously (i) if we choose a more efficient U, or (ii) if we use a larger sample size.

Ex 4. *A poll is conducted to estimate the prob. p that a randomly chosen person votes for candidate Jones in the upcoming elections. The sample size is $n = 15$.*

Assume $U = \sum_{i=1}^{15} Y_i$, $RR = [0, 5]$, $H_0 = \{0.5\}$ and $H_a = \{0.1\}$.

Q 4. *Calculate type I and II errors.*

$$\alpha = \mathbb{P}^0\left(\sum_{i=1}^{15} Y_i \le 5\right) \approx 0.151, \quad \beta = \mathbb{P}^a\left(\sum_{i=1}^{15} Y_i > 5\right) \approx 0.278.$$

This may be a better balance (although both errors are still large), although there is no canonical choice of the "best balance".

1 Popular tests for a normal sample

Assume that the sample $Y_1, \dots, Y_n$ consists of i.i.d. $N(\mu, \sigma^2)$ r.v.'s.

Q 5. *How to test $H_0 = \{\mu = \mu_0\}$ against $H_a = \{\mu < \mu_0\}$, $\{\mu > \mu_0\}$, $\{\mu \neq \mu_0\}$?*

Note that μ_0 **is a known constant**. Then, a relevant statistic is:

$$U = \sqrt{n}\frac{\bar{Y} - \mu_0}{S} \sim T(n-1), \text{ under } \mathbb{P}^0.$$

The natural choices of RR (for each choice of H_a, respectively) are

$$RR = (-\infty, -t_\alpha],\ [t_\alpha, \infty),\ (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty).$$

Ex 5. *Consider a sample of daily number of visits to a website: $n = 8$, $\bar{y} = 2959$, $s = 39.1$. The marketing agency which manages the website guarantees the average number of daily visits to be at least $3,000$.*

Q 6. *Assuming that the sample is normal, shall we accept or reject this claim at 0.025 significance level?*

$$H_0 = \{\mu = 3000\}, \quad H_a = \{\mu < 3000\},$$
$$u = \sqrt{8}\frac{2959 - 3000}{39.1} \approx -2.966 \in RR = (-\infty, -2.365).$$

Thus, we reject (or, fail to accept) H_0 and hence reject the claim.

Assume that we have two indep. samples $\{Y_i\}_{i=1}^{n_1}$ and $\{Z_i\}_{i=1}^{n_2}$ from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$.

Q 7. *How to test $H_0 = \{\mu_1 - \mu_2 = D_0\}$ against $H_a = \{\mu_1 - \mu_2 < D_0\}$, $\{\mu_1 - \mu_2 > D_0\}$, $\{\mu_1 - \mu_2 \neq D_0\}$?*

Note again that D_0 **is a known constant**. Note also that, **for the method described below, it is important that** $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

To construct a test, we recall the pooled estimator

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2},$$

and that

$$U = \frac{\bar{Y} - \bar{Z} - D_0}{S_p\sqrt{1/n_1 + 1/n_2}} \sim T(n_1+n_2-2), \quad \text{under } \mathbb{P}^0.$$

Note that U is more likely to take large values if $\mu_1 - \mu_2$ is large, and vice versa. Since the distribution of U is also known under H_0, it is natural to use U as the test statistic.

The connection between U and $\mu_1 - \mu_2$, mentioned above, leads to the following rejection intervals (for each choice of H_a):

$$RR = (-\infty, -t_\alpha],\ [t_\alpha, \infty),\ (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty).$$

Ex 6. *The performance of an employee is measured by the number of minutes he/she takes to complete a specific task. A new training method of employees (method 2) is tested against the old one (method 1). A sample* $\{y_i\}$ *(each* y_i *denotes the number of minutes it takes the* i*-th employee to complete the task) of size* $n_1 = 9$ *is collected from the employees trained according to method 1, and a sample* $\{z_i\}$ *of size* $n_2 = 9$ *is collected from the employees trained according to method 2. The following values of relevant statistics are observed:* $\bar{y} = 35.22$, $\bar{z} = 31.56$, $s_1^2 = 24.445$, $s_2^2 = 20.028$.

Q 8. *Assuming that the samples are normal, is there sufficient evidence to conclude that the new method is significantly different from the old one, at the* 0.05 *significance level?*

$$H_0 = \{\mu_1 - \mu_2 = 0\}, \quad H_a = \{\mu_1 - \mu_2 \neq 0\},$$

$$s_p = \sqrt{\frac{8s_1^2 + 8s_2^2}{16}} \approx 4.716, \quad u = \frac{\bar{y} - \bar{z}}{s_p\sqrt{1/9 + 1/9}} \approx 1.65,$$

$$RR = (-\infty, -t_{0.025}] \cup [t_{0.025}, \infty) = (-\infty, -2.12] \cup [2.12, \infty).$$

Thus, $u \notin RR$*, hence we accept* H_0*, and hence there is no significant evidence that the new method is significantly different from the old one (at the given confidence level).*

Assume that we need to test a hypothesis on the value of the true variance σ^2 of a normal distr.

Q 9. *How to test* $H_0 = \{\sigma^2 = \sigma_0^2\}$ *against* $H_a = \{\sigma^2 < \sigma_0^2\},\ \{\sigma^2 > \sigma_0^2\},\ \{\sigma^2 \neq \sigma_0^2\}$*?*

A natural candidate for a test statistic is

$$U = \frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1), \text{ under } \mathbb{P}^0.$$

Indeed, as S^2 is a good (e.g., unbiased, consistent) estimator of σ^2, we conclude that U is more likely to take large values if σ^2 is large, and vice versa. In addition, the distribution of U is known under H_0.

In view of the above, the rejection region (for each H_a) is

$$RR = (0, \chi_{1-\alpha}^2],\ [\chi_\alpha^2, \infty),\ (0, \chi_{1-\alpha/2}^2] \cup [\chi_{\alpha/2}^2, \infty).$$

Ex 7. *A factory machine produces a part of electric engine. The diameter of the produced parts is supposed to have variance no larger than* 0.0002 *(inches squared). A sample of size* $n = 10$ *produced* $s^2 = 0.0003$*.*

Q 10. *At* 0.05 *level, test* $H_0 = \{\sigma^2 = 0.0002\}$ *against* $H_a = \{\sigma^2 > 0.0002\}$*.*

The test statistic is

$$u = \frac{10-1}{0.0002}s^2 = \frac{9}{0.0002}0.0003 = 13.5.$$

And the rejection region is

$$RR = [\chi^2_{0.05}, \infty) = [16.919, \infty).$$

Thus, H_0 *is accepted (or, we fail to reject* H_0 *based on the data we have).*