**476 Statistics, Spring 2022.**
**Lecturer: Sergey Nadtochiy.**

**Lecture 7. Hypothesis testing: popular large-sample tests, choosing the sample size for a given $\beta$, p-value. (Sections 10.3–10.7)**

# 1 Popular large-sample hypotheses tests

In this section, we consider a sample of i.i.d. $Y_1, \ldots, Y_n$ that is **not** necessarily normal, but we assume that $n$ is large. As before, the distribution of $Y_i$ depends on the unknown parameter $\theta$.

Consider a null hyp. $H_0 = \{\theta = \theta_0\}$, where $\theta_0$ is a known constant, and the alternative is (one of)

$$H_a = \{\theta < \theta_0\}, \{\theta > \theta_0\}, \{\theta \neq \theta_0\}.$$

As the sample is not normal, we have to relax the notion of significance level and consider the **asymptotic significance level** $\alpha$, in the sense that

$$\lim_{n\to\infty} \mathbb{P}^0(U_n \in RR) \to \alpha,$$

where $U_n$ is the test statistic for a sample size $n$, and $RR$ is the rejection region.

To construct a test statistic $U_n$, we start by find an asymptotically normal estimator $\hat{\theta}_n$ of $\theta$: i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \tilde{\sigma}^2),$$

as $n \to \infty$. If the value of $\tilde{\sigma}^2$ under $H_0$ is not known, we find a consistent estimator $\hat{\sigma}_n^2$ of $\tilde{\sigma}^2$: i.e.,

$$\hat{\sigma}_n^2 \to \tilde{\sigma}^2,$$

as $n \to \infty$. (If the value of $\tilde{\sigma}^2$ under $H_0$ is known, we set $\hat{\sigma}_n^2 = \tilde{\sigma}^2$.) Finally, we construct the test statistic

$$U_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\sigma}_n}.$$

From Slutsky's theorem, we know that

$$U_n \to N(0,1), \quad \text{under } H_0,$$

as $n \to \infty$, in the sense of distribution.

In addition, $\hat{\theta}_n$ and hence $U_n$ are more likely to take large values if $\theta$ is large, $U_n$ is more likely to take small values if $\theta$ is small, and $U_n$ are more likely to take values far away from zero if $\theta$ is far away from $\theta_0$.

Thus, $U_n$ is a reasonable choice of a test statistic.

The rejection region of asymptotic level $\alpha$ is chosen as (one of)

$$RR = (-\infty, z_{1-\alpha}],\ [z_\alpha, \infty),\ (-\infty, z_{1-\alpha/2}] \cup [z_{\alpha/2}, \infty),$$

corresponding to the alternative hypotheses.

**Ex 1.** *A manager claims that each of his salespeople establishes, on average, no more than 15 contacts per week. He records the numbers of weekly contacts from $n = 36$ employees, the resulting sample mean is $\bar{y} = 17$ and the sample variance is $s^2 = 9$.*

**Q 1.** *At the asymptotic significance level $0.05$, does the data contradict the manager's claim?*

*$Y_i$ is the number of contacts established by the $i$-th employee per week, with $\mu = \mathbb{E}Y_i$. We test the null hypothesis*

$$H_0 = \{\mu = 15\}$$

*against the alternative*

$$H_a = \{\mu > 15\}.$$

*The values of relevant estimators are*

$$\hat{\theta} = \bar{y} = 17, \quad \hat{\sigma}^2 = s^2 = 9,$$

*and the test statistic takes the value*

$$u = \frac{\sqrt{36}(17 - 15)}{\sqrt{9}} = 4.$$

*The rejection region is*

$$RR = [z_{0.05}, \infty) = [1.645, \infty).$$

*Thus, we reject the null hyp. in favor of $H_a$.*

**Ex 2.** *A machine is in need of repair if more than $10\%$ of items it produces are defective. A sample of size $n = 100$ is collected, resulting in $15$ defective items. A supervisor claims that the machine must be repaired.*

**Q 2.** *Does the data support this claim, at the asymptotic significance level $0.01$?*

*We are dealing with Bernoulli sample with an unknown success probability $p$ (i.e., each $Y_i$ equals one if the $i$-th item is defective, and zero otherwise), and*

$$H_0 = \{p = 0.1\}, \quad H_a = \{p > 0.1\}, \quad \hat{p} = \bar{y} = 0.15, \quad \sqrt{n}(\bar{y} - p) \to N(0, \tilde{\sigma}^2), \quad \tilde{\sigma}^2 = p(1 - p).$$

*Note that*

$$\hat{\sigma}^2 = \bar{y}(1 - \bar{y})$$

*is a consistent estimator of $\tilde{\sigma}^2$. However, under $H_0$, we know the exact value of $\tilde{\sigma}^2$:*

$$\tilde{\sigma}^2 = 0.1 \cdot (1 - 0.1) = 0.09.$$

*Clearly, it is more efficient to use the above value than it estimate. Thus, we obtain the test statistic value:*

$$u = \frac{\sqrt{100}(0.15 - 0.1)}{\sqrt{0.09}} = \frac{0.5}{0.3} \approx 1.67, \quad RR = [z_\alpha, \infty) = [2.33, \infty).$$

*Thus, we accept (or, fail to reject) $H_0$, which means that the data does not support the supervisor's claim at the given significance level.*

*Note that we could use $\hat{\sigma}^2$ as an approximation of $\tilde{\sigma}^2$. Strictly speaking it would be correct (and, in this case, would yield the same answer), but it is clearly less efficient.*

Next, assume that we need to compare the means $\mu_1$ and $\mu_2$ of two independent samples $\{Y_i\}_{i=1}^{n_1}$ and $\{Z_i\}_{i=1}^{n_2}$, with the null hypothesis

$$H_0 = \{\mu_1 - \mu_2 = D_0\}$$

and with the alternatives $H_a = \{\mu_1 - \mu_2 < D_0\}, \{\mu_1 - \mu_2 > D_0\}, \{\mu_1 - \mu_2 \neq D_0\}$.

A natural choice of the test statistic is

$$U = \frac{\bar{Y} - \bar{Z} - D_0}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}} \to N(0,1),$$

where the limit is taken as $n_1, n_2 \to \infty$, and where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are consistent estimators of $\sigma_1^2 = V(Y_i)$ and $\sigma_2^2 = V(Z_i)$.

We choose RR accordingly: $RR = (-\infty, -z_\alpha], [z_\alpha, \infty), (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$.

**Rem 1.** *Just remember that we also need to assume that*

$$\frac{|n_1 - n_2|}{\sqrt{n_1} + \sqrt{n_2}}$$

*remains bounded.*

**Ex 3.** *A study compares reaction times (in seconds) of men and women to a certain stimulus. Two samples $\{Y_i\}$ and $\{Z_i\}$ of sizes $n_1 = 50$ and $n_2 = 50$, of men and women respectively, are taken. The realized statistics' values are $\bar{y} = 3.6$, $s_1^2 = 0.18$, $\bar{z} = 3.8$, $s_2^2 = 0.14$.*

**Q 3.** *At the (asymptotic) level $0.05$, does the data provide sufficient evidence that $\mu_1 \neq \mu_2$?*

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

$$u = \frac{3.6 - 3.8}{\sqrt{0.18/50 + 0.14/50}} = -2.5, \quad RR = (-\infty, -z_{0.025}] \cup [z_{0.025}, \infty) = (-\infty, -1.96) \cup (1.96, \infty)$$

*Thus, $H_0$ is rejected, which means that the data does provide sufficient evidence that $\mu_1 \neq \mu_2$ at the given significance level.*

## 2 Choosing sample size for a given $\beta$

Recall that the type-I error, measured by the significance level $\alpha$, is typically fixed. On the other hand, if $H_a$ is simple, we may be interested in **decreasing the type-II error** $\beta$. One way to achieve this is by increasing the sample size $n$.

Consider $H_0 = \{\theta = \theta_0\}$ and $H_a = \{\theta = \theta_a\}$, where $\theta_0 < \theta_a$ are known constants. Let $\hat{\theta}_n$ be an asymptotically normal estimator of $\theta$, with

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \tilde{\sigma}^2).$$

Assume also that $\hat{\sigma}_n^2$ is a consistent estimator of $\tilde{\sigma}^2$. Then, we can use the test statistic $U$ and the rejection region RR given by

$$U_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\sigma}_n}, \quad RR = (z_\alpha, \infty).$$

Note that $U_n \to N(0,1)$, hence the chosen test has asymptotic significance level (i.e., the type-I error) $\alpha$.

On the other hand, as $n \to \infty$. the type-II error is

$$\beta = \mathbb{P}^a(U_n \notin RR) = \mathbb{P}^a\left(\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\sigma}_n} \leq z_\alpha\right) = \mathbb{P}^a\left(\frac{\sqrt{n}(\hat{\theta}_n - \theta_a)}{\hat{\sigma}_n} \leq z_\alpha - \frac{\sqrt{n}(\theta_a - \theta_0)}{\hat{\sigma}_n}\right)$$

$$= \mathbb{P}^a\left(U_n \leq z_\alpha - \frac{\sqrt{n}(\theta_a - \theta_0)}{\hat{\sigma}_n}\right) \approx \Phi\left(z_\alpha - \frac{\sqrt{n}(\theta_a - \theta_0)}{\hat{\sigma}_n}\right),$$

where $\Phi$ is the cdf of $N(0,1)$.

**Rem 2.** *The above derivation is purely heuristic – it only holds under additional technical assumptions. But, in this course, we assume that these assumptions are always satisfied.*

If we want to achieve a given value $\beta$ of the type-II error, we need

$$\Phi\left(z_\alpha - \frac{\sqrt{n}(\theta_a - \theta_0)}{\hat{\sigma}_n}\right) = \beta,$$

which leads to

$$z_\alpha - \frac{\sqrt{n}(\theta_a - \theta_0)}{\hat{\sigma}_n} = z_{1-\beta} = -z_\beta,$$

$$n = \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}_n^2}{(\theta_a - \theta_0)^2}.$$

**Rem 3.** *Note that $\hat{\sigma}_n^2$ still depends on $n$, so, strictly speaking, we haven't solved the equation for $n$. In practice, we cheat a bit: we run a preliminary estimate (on a possibly smaller sample), to compute $\hat{\sigma}^2$, then, use this estimate in the above formula to get $n$ for the new sample.*

**Ex 4.** *A manager wants to test $H_0$ that his salespeople establish on average 15 contacts per week (each) against the alternative $H_a$ of 16 contacts per week, with errors $\alpha = \beta = 0.05$. Assume that the true variance of the number of weekly contacts made by an employee, denoted $\sigma^2$, has been approximated (via a consistent estimator) by $9$.*

**Q 4.** *What is the size of the sample needed to achieve the desired accuracy of the test?*

$$n = \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}_n^2}{(\theta_a - \theta_0)^2} = \frac{(1.645 + 1.645)^2 \cdot 9}{(16 - 15)^2} = 97.4$$

# 3 p-value

**What significance level $\alpha$ should be used for a test?** This choice is somewhat arbitrary.

Instead of fixing $\alpha$ and searching for "acceptance/rejection" answer, let us search for the smallest $\alpha$ that gives us the "rejection" answer.

**Def 1.** *Consider a test statistic $U(Y_1, \ldots, Y_n)$ and a family of rejection regions $RR(\alpha)$, for all $\alpha \in (0,1)$. Then, the p-Value of this test is the smallest $\alpha$ s.t.*

$$u \in RR(\alpha),$$

*where $u$ is the realized value of $U$.*

Note that p-Value contains more information than a simple "acceptance/rejection" answer for a given level. Indeed, p-value is the smallest significance level at which we can reject the null hypothesis: i.e., the null hyp. is rejected for any $\alpha \geq p$ and accepted for any $\alpha < p$.

If the null hyp. is rejected (i.e. if the test is conclusive), the smaller is the p-Value, the more compelling is the evidence for rejection. Heuristically, "one minus the p-value" is the confidence with which we can reject the null (although it is not a very precise statement).

**Ex 5.** *Recall the example of a political poll, whose goal is to make a statistical inference about the unknown probability $p$ that a randomly chosen voter is in favor of candidate Jones. In a sample of size $n = 15$ there are $3$ people who are in favor of Jones.*

**Q 5.** *For the hypotheses $H_0 = \{p = 0.5\}$, $H_a = \{p < 0.5\}$, construct a test and find its p-value.*

*As before, the sample $\{Y_i\}$ consists of i.i.d. r.v.'s with $Bernoulli(p)$ distribution ($Y_i = 1$ corresponds to the $i$-th person being in favor of Jones) we use*

$$U = \sum_{i=1}^{15} Y_i$$

*as a test statistic, and use the rejection region $RR(\alpha) = (0, b_{1-\alpha}]$, where $b_{1-\alpha}$ is the quantile of level $\alpha$ of $Bin(0.5, 15)$. Then, we compute*

$$u = \sum_{i=1}^{15} y_i = 3, \quad RR(\alpha) = (0, b_{1-\alpha}].$$

*To compute the p-value, we need to find the smallest $\alpha \in (0,1)$ s.t.*

$$3 \leq b_{1-\alpha}.$$

*The above is equivalent to finding the smallest $\alpha \in (0,1)$ s.t.*

$$\mathbb{P}\left(\sum_{i=1}^{15} Y_i \leq 3\right) \leq \mathbb{P}\left(\sum_{i=1}^{15} Y_i \leq b_{1-\alpha}\right) = \alpha,$$

*which gives the p-value*

$$p = \mathbb{P}\left(\sum_{i=1}^{15} Y_i \leq 3\right) \approx 0.018.$$

*Thus, according to the data, the null hyp. $H_0 = \{p = 0.5\}$ is rejected if our target significance level is at or above* $0.018$. *In other words, this null hyp. can only be accepted if the significance level is below* $0.018$.

The above example motivates the following general method. Assume that $RR(\alpha)$ has the form

$$(-\infty, q_{1-\alpha}), (q_\alpha, \infty), (-\infty, q_{1-\alpha/2}) \cup (q_{\alpha/2}, \infty),$$

with $q_\alpha$ being the quantile of the test statistic $U$ at level $1 - \alpha$, under $H_0$. Then, the p-value is

$$p = F_U(u),\ 1 - F_U(u),\ 2\min\left(F_U(u), 1 - F_U(u)\right),$$

where $F_U$ is the cdf of $U$ under $H_0$, and $u$ is the realized value of the test statistic.

Indeed, let us consider the first case, $RR = (-\infty, q_{1-\alpha})$. Then, the p-value is the smallest $\alpha$ s.t.

$$u \leq q_{1-\alpha}.$$

The above is equivalent to finding the smallest $\alpha$ s.t.

$$F_U(u) \leq F_U(q_{1-\alpha}) = \alpha.$$

Clearly, the smallest value of $\alpha$ that satisfies the above is $F_U(u)$, which is the desired p-value. The other cases are treated similarly.

If the sample and the unknown parameter are "nice" (e.g., the sample is normal), we may be able to find the distribution of $U$ under $H_0$ explicitly.

If the sample is not normal, the test statistic $U$ is typically chosen to be asymptotically normal under $H_0$ (e.g., via an asymptotically normal estimator of the unknown parameter). In such a case, $F_U$ is replaced by the asymptotic (normal) distribution of $U$ under $H_0$.

**Ex 6.** *A study compares the reaction times (measured in sec.) of men and women to a certain stimulus. Two samples (the first one of men, and the second one of women) are collected:* $n_1 = 50$, $\bar{y} = 3.6$, $s_1^2 = 0.18$, $n_2 = 50$, $\bar{z} = 3.8$, $s_2^2 = 0.14$.

**Q 6.** *Find a p-value for*

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

*with*

$$U = \frac{\bar{Y} - \bar{Z}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \quad RR = (-\infty, -r) \cup (r, \infty).$$

*Recall that, as $n_1, n_2 \to \infty$ (with $n_1$ and $n_2$ being not to different), we have:*

$$\frac{\bar{Y} - \bar{Z} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \to N(0, 1).$$

*Then, under $H_0$, we have*

$$U = \frac{\bar{Y} - \bar{Z}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \to N(0, 1).$$

*As the test statistic has a known (asymptotically) distribution under $H_0$, and since the rejection interval is of the prescribed type, we can use the above general method to obtain*

$$u = -2.5, \quad p = 2\min(\Phi(-2.5), 1 - \Phi(-2.5)) = 2\Phi(-2.5) = 2 \cdot 0.0062 = 0.0124.$$

*One has to have a very small significance level (below the above p-value) to accept the null hyp.*

**Rem 4.** *Two-sided tests and confidence intervals based on an asymptotically normal estimator $\hat{\theta}$ are closely related. Indeed, a confidence interval is centered around $\hat{\theta}$, while the acceptance region of the test has the same length and is centered around $\theta_0$ (the value of $\theta$ under $H_0$). In the former case, we know that the true parameter value is in the confidence interval with high probability. In the latter case, under $H_0$, the observed value of $\hat{\theta}$ must lie in the acceptance region with high probability.*