TESTS OF STATISTICAL HYPOTHESES

A. STATEMENT OF THE PROBLEM AND MAIN CONCEPTS

We have to decide which of two hypotheses is true: one is the "conservative" null hypothesis $H_0$, the other one is called the alternative hypothesis and denoted $H_a$.

When making a decision which of these hypotheses is true using a test we may make one of the following two kinds of errors:

Type 1 error: $H_0$ is rejected when $H_0$ is true.
Type 2 error: $H_0$ is accepted when $H_a$ is true.

We denote: $\alpha$ is the probability of the type 1 error, $\beta$ is the probability of the type 2 error.

$\alpha$ is called the significance level of the test. The number $K = 1 - \beta$ is called the power of the test; $K$ is the probability that we reject $H_0$ when it is false (i.e. when $H_a$ is true).

Usually, when we design the test we first fix a (small) significance level $\alpha$ (e.g. 0.05) and then we find a test with this significance level and with a small $\beta$ and (which is the same) with a large power $K$.

In order to test the hypotheses we need a sample $X_1, X_2, ..., X_n$. A test is a combination of a test statistic $\hat{\Theta}$ and a Rejection Region (RR). The complementary set $AR = \mathbb{R}\backslash RR$ is called the acceptance region. We reject $H_0$ and accept the alternative $H_a$ if the observed value of the statistic $\hat{\theta} \in RR$, and accept $H_0$ if $\hat{\theta} \notin RR$ that is when $\hat{\theta} \in AR$.

We illustrate this general approach with simple examples.

It is clear that $\alpha = P(\hat{\Theta} \in RR; H_0)$ (i.e. when $H_0$ is true); $\beta = P(\hat{\Theta} \notin RR; H_a)$; $K = P(\hat{\Theta} \in RR; H_a)$. 

B. EXAMPLE: TESTS FOR THE MEAN OF A NORMAL R.V.

\((\sigma \text{ is known})\)

**Right-hand alternative**

Let \(X \sim N(\mu, \sigma^2)\); assume \(\sigma\) is known.

We want to test the following hypotheses:
\[
H_0 : \mu = \mu_0 \\
H_a : \mu > \mu_0.
\]

Here the null hypothesis \(H_0\) is a *simple* hypothesis: under this hypothesis the distribution is completely specified. The alternative hypothesis \(H_a\) is a *composite* hypothesis: even if we know it is true, \(\mu\) remains unknown, so the distribution of \(X\) is not specified.

Consider a random sample \(X_1, ..., X_n\).

**Test statistic:** Of course, we consider the MVUE \(\bar{X}\).

**Rejection (Critical) Region.** We consider the RR "in favor" of the alternative \(H_0\). Intuitive motivation: the alternative hypothesis claims that \(\mu\) is "large" and since \(\bar{X}\) is consistent we believe that the observed value \(\bar{x}\) is close to \(\mu\) so it has to be "large", too. So we consider the RR of the form: \(\text{RR} = \{\bar{x} : \bar{x} \geq c\}\) where the *critical threshold* \(c\) is chosen to provide the fixed significance level \(\alpha\). It turns out that this intuitive choice of the RR leads to the smallest \(\beta\) and to the largest power \(K\) for the fixed significance level \(\alpha\).

Let us find \(c\). We want to have:

\[
P(\bar{X} \geq c; \mu = \mu_0) = \alpha.
\]

Recall that if \(H_0\) holds that is \(E(X) = E(\bar{X}) = \mu_0\) then \(Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\) is standard normal.

Sometimes it convenient to consider the test statistic \(Z\) instead of \(\bar{X}\). Then the rejection region is clear: \(P(Z \geq z_\alpha) = \alpha\) so the RR for \(Z\) is \(\{Z \geq z_\alpha\}\).
But let us find \( c \) for the statistic \( \bar{X} \). We have

\[
\alpha = P(Z \geq z_\alpha) = P\left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \right) = P\left( \bar{X} - \mu_0 \geq z_\alpha \frac{\sigma}{\sqrt{n}} \right) = P\left( \bar{X} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \right)
\]

so

\[
c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.
\]

Thus the RR for the statistic \( \bar{X} \) with the significance level \( \alpha \) is \([\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)\).

**Example 1.** Let \( X \) be the strength of a steel bar produced by some process; it is known that \( X \sim N(50, 36) \). A new process is assumed to produce stronger bars, i.e. for the new process \( \mu > 50 \). Before going to the new process we want to check whether this claim is true. So we are going to test the hypotheses:

\[
H_0 : \mu = 50
\]

\[
H_a : \mu > 50.
\]

\( n = 16 \) bars are produced by the new process.

Let us try the RR=\( \{x : x \geq 53\} \).

\[
\alpha = P(\bar{X} \geq 53; H_0) = P\left( \frac{\bar{X} - 50}{\sqrt{36/16}} \geq \frac{53 - 50}{\sqrt{36/16}} \right) = P(Z \geq \frac{53-50}{2}) = P(Z \geq 2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228.
\]

So \( \alpha \) is small. But what is \( \beta \) and \( K \)?

\[
\beta = P(\bar{X} < 53; H_a).
\]

Since \( H_a \) is a composite hypothesis we cannot calculate \( \beta \) and \( K \) without specifying a value of \( \mu > 50 \).

First observe that \( K(50) = P(\bar{X} \geq 53; H_0) = \alpha \).

It is always so: \( K(H_0) = P(\bar{X} \in RR; H_0 = \alpha) \).

Let \( \mu = 54 \).

\[
\beta(54) = P(\bar{X} < 53; \mu = 54) = P\left( \frac{\bar{X} - 54}{\sqrt{36/16}} \leq \frac{53 - 54}{2} \right) = P(Z \leq -0.67) = 1 - \Phi(0.67) = 1 - 0.7486 = 0.2514.
\]

\[
K(54) = 1 - \beta(54) = 0.7486.
\]

\[
\beta(56) = P(\bar{X} < 53; \mu = 56) = P(Z < \frac{53 - 56}{\sqrt{36/16}}) = P(Z < -2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228.
\]
\[ K(56) = 1 - \beta(56) = 0.9772. \]

Let us construct the RR with \( \alpha = 0.05 \). We want to find the critical threshold \( c \) such that
\[
P(\bar{X} \geq c; \mu = 50) = 0.05.
\]
\[
P\left( \frac{\bar{X} - 50}{\frac{3}{2}} \geq z_{0.05} \right) = 0.05. \text{ Thus } P(\bar{X} \geq 50 + z_{0.05} \times \frac{3}{2} = 52.47. \text{ Hence we obtained:} \]
\[
c_{0.05} = 52.47. \text{ Of course we could use formula (0.1) to avoid this calculations!}
\]
Now we consider the RR=\{ \( x \) : \( x \geq 52.47 \) \} with \( \alpha = 0.05 \). So the new RR has a larger \( \alpha \) (still small enough!) but it has smaller \( \beta \) and larger power \( K \).

For example, similar calculations as above (with 52.47 instead of 53) give us:
\[
K[54] = 0.8461; \ K[56] = 0.9907.
\]
\[
\beta[54] = 0.1539; \ \beta[56] = 0.0093.
\]
Compare! The smaller is \( \alpha \), the smaller is the RR and therefore the smaller is \( K \).

See Figure 1 below. Note that in both cases \( K(50) = \alpha \) (recall that 50 = \( \mu_0 \) in our example).

![Figure 1: THE POWER FUNCTIONS OF THE TESTS WITH \( \alpha = 0.05 \) (\( c = 52.47 \) (GREEN) AND \( \alpha = 0.0228 \) (\( c = 53 \) (RED))](image)

Performing the test. Now assume that the 16 trials produced \( \bar{x} = 52.7 \). This value
is in RR=\{\bar{x} \geq 52.47\} but not in RR=\{\bar{x} \geq 53\}. So we reject \(H_0\) with the significance level 0.05 but do not reject it with the significance level 0.0228 (recall that the latter test provides a smaller probability to reject \(H_0\) when it is true - we are more careful, do not want to risk!).

The tests with the RR=\(\{x \geq c_\alpha\}\) have the greatest power, if \(\alpha\) is fixed (this follows from a general theorem).

Let us compare the RR with \(\alpha = 0.05\) (green graph in Figure 1) with two other RRs also with \(\alpha = 0.05\), but, of course, with smaller power.

1) RR=\(\bar{X} \in [51, 51.25]\). \(\alpha = 0.05\); see Figure 2 (note: \(K(50) = 0.05\)).

![Figure 2: THE POWER FUNCTIONS OF THE TEST WITH RR=[51,51.25](\alpha = 0.05 )](Image)

2) RR=\(\{x : x \leq 50 - 2.47 = 47\}\); this is a "crazy" RR: it is a one-sided interval but it is in the wrong direction: "against" the alternative hypothesis \(H_a\) (although \(\alpha = 0.05\) too). See Figure 3.
Figure 3: THE POWER FUNCTIONS OF THE "CRAZY" TEST WITH RR=$\{x : x \leq 47.53\}$ ($\alpha = 0.05$)

C. THE $p$-VALUE OF THE OBSERVED STATISTIC

Example 2. Refer to Example 1. Assume we have to write a report about the result of the 16 trials of the steel bars. Assume that we observed $\bar{x} = 52.2$. We do not reject $H_0$ with $\alpha = 0.0228$ and even with $\alpha = 0.05$ since $\bar{x}$ does not belong to the rejection regions. So we write that we cannot reject $H_0$ with significance level 0.05. But 52.2 seems to be so close to the critical value $c_{0.05} = 52.47$, and the reader may have doubts about accepting $H_0$. May be it would be rejected if we chose a larger RR, i.e. a smaller critical level? So let us put $c = 52.2$ and calculate $\alpha$ for this RR.

$$P(\bar{X} \geq 52.2; H_0) = P\left(\frac{\bar{X} - 50}{\frac{3}{\sqrt{2}}} \geq \frac{52.2 - 50}{\frac{3}{\sqrt{2}}}\right) = P(Z \geq 1.47) = 1 - 0.9292 = 0.0708.$$  

This probability is called the probability value (or, in short, $p$-value) of the observed value $\bar{x} = 52.2$.  

6
We report this value and let the reader decide whether to accept or reject $H_0$ when the probability of a mistaken rejection of $H_0$, $\alpha = 0.07$. In general, the $p$-value for the one-sided right-hand alternative when we observe $\bar{x}$ is

$$p - \text{value}(\bar{x}) = P(\bar{X} \geq \bar{x}; H_0).$$

In other words, the $p$-value is the significance level of the test when the critical value $c = \bar{x}$ (the observed value of $\bar{X}$). Roughly speaking it shows, how likely it is that $H_0$ is true if we have observed the value $\bar{x}$.

**Example 3.** Refer to Example 1. Assume we observed $\bar{x} = 51$. As in the previous example we do not reject $H_0$ with significance $\alpha = 0.05$. But then the reader could reject $H_0$ with significance $\alpha = 0.07$. In this case the discrepancy between the observed value of $\bar{x}$ and $\mu_0 = 50$ is considerably smaller, and it seems that this result is in favor of $H_0$. We can support our intuition with a quantity:

$$p\text{-value}(51) = P(\bar{X} \geq 51; H_0) = P\left(\frac{\bar{X} - 50}{\frac{3}{2}} \geq \frac{51 - 50}{\frac{3}{2}}\right) = P(Z \geq 0.67) = 1 - 0.7486 = 0.2514.$$

We report this $p$-value. I do not think somebody will reject $H_0$ when $\alpha = 0.25$.

In general, when we are testing hypotheses about a parameter $\theta$ using a test statistic $\hat{\Theta}$, and the observed value is $\theta$, we chose the smallest RR that still contains $\theta$, say $\text{RR}(\theta)$.

**Definition 1.** The $p$-value equals the significance level $\alpha$ corresponding to $\text{RR}(\theta)$, that is

$$p\text{-value}(\theta) = P(\hat{\Theta} \in \text{RR}(\theta); H_0).$$

In our examples we had to choose the smallest interval $[c, \infty)$ such that $\bar{x} \in [c, \infty)$; of course $c = \bar{x}$. Thus $p\text{-value}(\bar{x}) = P(\bar{X} \geq \bar{x}; H_0)$. 


D. COMPARISON: USING CIs AND RRs TO TEST HYPOTHESES

When we studied CIs, we used them to test statistical hypotheses. For example, to test the hypotheses about the mean value of a normal random variable $X$

$$H_0 : \mu = \mu_0$$
$$H_a : \mu > \mu_0$$

we considered a random sample $X_1, \ldots, X_n$ and used the CI $\{ x : x \geq \bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \}$. We rejected $H_0$ (and accepted $H_a$) if $\mu_0 \notin CI$. If the confidence coefficient of this CI is $Q = 1 - \alpha$ we have $P(\mu_0 \notin CI; H_0) = 1 - \alpha$ and $P(\mu_0 \notin CI; H_0) = \alpha$ which is, in fact, the probability of the type 1 error ($H_0$ is rejected when it is true). So the CI approach provided a test with significance level $\alpha$. But we could not consider neither $\beta$, nor $K$, and hence could not explain rigorously why we have chosen this one-sided right-hand CI.

Let us note that $\mu_0 \notin CI$, i.e. $\mu_0 \leq \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$ is the same as $\bar{x} \geq \mu_0 - z_{\alpha} \frac{\sigma}{\sqrt{n}}$ or that $\bar{x} \in RR$. So the CI test is essentially the same as the RR test but the RR approach gives us the possibility to explain why it is the one-sided right-hand RR (and the one-sided right-hand CI) it is reasonable to use for testing our hypotheses: as we know, this RR provides the greatest power $K$ and the smallest probability of type 2 error $\beta$ (with the fixed $\alpha$).

This approach is also very efficient when we consider tests of other kinds of statistical hypothesis: it shows how to construct RR with largest powers ("most powerful tests"). This theory will be covered later.