Hypothesis Testing

Guy Lebanon

May 16, 2006

We have seen so far two types of statistical estimation frameworks: confidence intervals and point estimation. Hypothesis testing is a third inference framework that is concerned with choosing a hypothesis supported by available data out of a number of competing alternatives. We start with the basic definitions and then proceed to describe a few important cases.

We assume that we have data $X_1, \ldots, X_n$ sampled from a distribution characterized by a parameter θ. A hypothesis is a set of possible values for θ. We will consider cases where there are two competing hypothesis: the null hypothesis H_0 and the alternative hypothesis H_A, with $H_0 \cap H_A = \emptyset$. The two hypothesis are not treated in a symmetric manner. The null hypothesis usually is used to describe a set of standard or believable values. The alternative hypothesis is an alternative to the null. It is sometimes called the research hypothesis since it may describe a research statement that one wishes to examine.

A hypothesis test is executed by observing the values of a test statistic $T(X_1, \ldots, X_n)$. If it lies in a set called the rejection region (RR) then the null hypothesis is rejected and the alternative is accepted. Otherwise, the null is accepted and the alternative hypothesis is rejected

$$T(X_1, \ldots, X_n) \in RR \Rightarrow H_A \text{ is true} \qquad T(X_1, \ldots, X_n) \notin RR \Rightarrow H_0 \text{ is true.}$$

A hypothesis test is thus composed of a sampling distribution, a parameter of interest θ, a null and alternative hypotheses, a test statistics and a rejection region.

A test can be good or bad depending on whether it leads often to the right decision or not. Formally, we have two types of errors that could be made. A type-1 error is made if H_0 is rejected while it is true and a type-2 error is made if H_0 is accepted while it is not true. The corresponding probabilities of these two errors are denoted by α and β:

$$\alpha = P(T(X_1, \ldots, X_n) \in RR | \theta \in H_0) \qquad \beta = P(T(X_1, \ldots, X_n) \notin RR | \theta \in H_A).$$

The two probabilities α, β are actually defined for particular values of θ and thus are function of θ (the true value of the parameter).

Example: A company wishes to test whether a new drug helps to prevent a certain disease. The presence or absence of the disease in people treated with the new drug is a sequence Bernoulli RV with parameter θ. Suppose the rate of contracting the disease for people not taking the drug is 0.3. The null hypothesis then could be that for people taking the drug $H_0 : \theta = 0.3$. The alternative or research hypothesis could be that the drug helps thus $\theta < 0.3$. Note that $H_0 \cup H_A$ does not have to cover the entire set of possible values for θ. The test statistics is $\bar{X}$ and the rejection region is $[0, c]$ for some value of c. The quality of the test is captured by α, β with

$$\alpha = P(\bar{X} \leq c | \theta = 0.3) = P\left(\sum_{i=1}^{n} X_i \leq nc \middle| \theta = 0.3\right) = \sum_{k=0}^{\lfloor nc \rfloor} \binom{n}{k} 0.3^k 0.7^{n-k}.$$

The computation of H_0 was possible since we knew precisely what is the value of θ (since H_0 contained only one value). Since H_A contain several values, we can't compute β unless we are given the specific value. For example, assuming that $\theta = 0.2$ we have

$$\beta = P(\bar{X} > c | \theta = 0.2) = P\left(\sum_{i=1}^{n} X_i > nc \middle| \theta = 0.2\right) = \sum_{k=\lceil nc \rceil}^{n} \binom{n}{k} 0.2^k 0.8^{n-k}.$$

To ensure a good test, we would like both α and β to be low. However, there is a trade-off between α, β. For example, a test for which $RR = \mathbb{R}$ always rejects the null and hence $\beta = 0$ but $\alpha = 1$ while a test for which $RR = \emptyset$, $\alpha = 0$ but $\beta = 1$. From the non-symmetric definitions of H_0 and H_A, α is sometimes considered more important. For the test in the example above, $RR = [0, c]$ and as we increase c from 0 to 1, the probability of rejecting the null increases, thus α increases. On the other hand, since we reject the null more often, we accept H_A more and β decreases.

Large Sample Tests for Means

We turn now to a specific type of hypothesis tests for means where the sampling distribution is unknown but we use the central limit theorem (CLT) to approximte the distribution of the test statistic. This test is sometimes called the z-test. The framework is as follows. We have an iid sample $X_1, \ldots, X_n$ from an unknown distribution with a known variance σ^2 but unknown means $\mathsf{E}(X_i) = \mu$. We want to test $H_0 : \mu = \mu_0$ vs. (i) $H_A : \mu > \mu_0$ or (ii) $H_A : \mu < \mu_0$ or (iii) $H_A : \mu \neq \mu_0$ (the first two alternatives are called one-sided alternatives and the last one is called two-sided alternative). The test statistic is $\bar{X}$, the rejection region is $[c, \infty)$ for case (i), $(-\infty, c]$ for case (ii) and $(-\infty, c_1] \cup [c_2, \infty)$ for case (iii). The problem we are facing is connecting the precise shape of the rejection region (i.e. c or c_1, c_2) with the error probabilities α, β.

The starting point is the construction of a pivot that is approximately normal (by CLT) $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$. We can then compute α by manipulating the probability of rejecting the probabilities related to the (approximately) normal pivot and then consulting a table of normal values. For example, for case (i) we have

$$\alpha = P(\bar{X} \geq c | \mu_0) = P\left(\bar{X} - \mu_0 \geq c - \mu_0\right) = P\left(\frac{\bar{X}-\mu_0}{\sigma} \geq \frac{c-\mu_0}{\sigma}\right) = P\left(\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} \geq \sqrt{n}\frac{c-\mu_0}{\sigma}\right)$$
$$\approx P\left(Z \geq \sqrt{n}\frac{c-\mu_0}{\sigma}\right)$$

where $Z \sim N(0,1)$ and hence $\sqrt{n}\frac{c-\mu_0}{\sigma} = z_\alpha$ whose solution gives $c = \mu_0 + z_\alpha \sigma/\sqrt{n}$. Such a rejection region $[\mu_0 + z_\alpha\sigma/\sqrt{n}, \infty)$ will ensure the corresponding α (approximately). Since our analysis depends on the CLT, it will be a good approximation for large n. For this reason such tests are called large sample tests. Note that we did not assume that we know the distribution of X_i. The derivation of case(ii) is very similar to case (i). In case (iii) we have

$$1 - \alpha = P(c_1 < \bar{X} < c_2 | \mu_0) = P\left(\sqrt{n}\frac{c_1-\mu_0}{\sigma} < \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} < \sqrt{n}\frac{c_2-\mu_0}{\sigma}\right) \approx P\left(\sqrt{n}\frac{c_1-\mu_0}{\sigma} < Z < \sqrt{n}\frac{c_2-\mu_0}{\sigma}\right).$$

Assuming symmetry we can set $\sqrt{n}\frac{c_1-\mu_0}{\sigma} = -z_{\alpha/2}$ and $\sqrt{n}\frac{c_2-\mu_0}{\sigma} = z_{\alpha/2}$ (there also other solution for non-symmetric RR) yields the solutions $c_1 = \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$, $c_2 = \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$.

Notice that for all cases (i), (ii), (iii) we do not know the precise value of μ under H_A and therefore it is more difficult to compute β than α (for the above test). If we specify a particular value of μ, μ_A under H_A we can solve for β in a similar way as we did for α. An interesting application of this is when we can choose n as well as the rejection region (we can sample additional data as we please). In this case we can treat both n and c (or c_1, c_2) as variables and solve for their values that yield specific α as well as β. For example for case (i) we get

$$\beta = P(\bar{X} < c | \mu_A) = P\left(\sqrt{n}\frac{\bar{X}-\mu_A}{\sigma} < \sqrt{n}\frac{c-\mu_A}{\sigma}\right) \approx P\left(Z < \sqrt{n}\frac{c-\mu_A}{\sigma}\right).$$

This leads to the following two equations (that may be solved for n, c): $z_\alpha = \sqrt{n}\frac{c-\mu_0}{\sigma}$, $z_\beta = -\sqrt{n}\frac{c-\mu_A}{\sigma}$.

If the variance σ^2 of the sampling distribution is not known, we can use an approximation for it such as $\hat{\sigma}^2 = S^2$. To connect the RR and α we now use the CLT together with Slutsky's theorem which proves that the pivot $\sqrt{n}\frac{\bar{X}-\mu_0}{S}$ is approximately $N(0,1)$ (this is proven in the note on consistency). All the formulas carry over with σ being replaced by S.

Small Sample Tests for Means and Variances

If we know that the sampling distribution of $X_1, \ldots, X_n$ is $N(\mu, \sigma^2)$ (without necessarily knowing μ, σ^2) we can construct precise or small sample tests. We start by focusing on a test for the mean, which is known as the t-test.

The test statistic and rejection region shape is as before. However, the connection between c (or c_1, c_2) the α (or α, β and n) is different since we are using an exact rather than a precise pivot. If the variance σ^2 is known we use the pivot $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ which is a linear combination of normal distribution and therefore it has a normal distribution. Its mean is 0 and variance is $\mathsf{Var}\,\sqrt{n}\frac{\bar{X}-\mu}{\sigma} = \frac{n}{\sigma^2}\mathsf{Var}\,\bar{X} = 1$ and hence the pivot is $N(0,1)$. If σ^2 is unknown we can use the pivot $\sqrt{n}\frac{\bar{X}-\mu}{S}$ which has a t-distribution with $n-1$ degrees of freedom (this was proved in the note on sampling distributions).

Example: Assume that σ^2 is unknown and we have $H_A = \mu > \mu_0$, $RR = [c, \infty)$ and the test statistic $\bar{X}$. Then the connection between α and c is as follows.

$$\alpha = P(\bar{X} \geq c|\mu_0) = P\left(\sqrt{n}\frac{\bar{X}-\mu_0}{S} \geq \sqrt{n}\frac{c-\mu_0}{S}\right) = P\left(T \geq \sqrt{n}\frac{c-\mu_0}{S}\right)$$

where T has a t-distribution with $n-1$ dof. Note that instead of the approximation sign as before we have now equality. The connection is thus precise for small as well as for large n. Solving $\sqrt{n}\frac{c-\mu_0}{S} = t_\alpha$ (where t_α is the α-critical value of the t-distribution - obtained from a table) we have $c = \mu_0 + t_\alpha \sigma/\sqrt{n}$.

Assuming that the data is normally distributed $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (μ, σ^2 are unknown), we can also test for the variance $H_0 : \sigma^2 = \sigma_0^2$ vs. (i) $H_A : \sigma^2 > \sigma_0^2$ or (ii) $H_A : \sigma^2 < \sigma_0^2$ or (iii) $H_A : \sigma^2 \neq \sigma_0^2$. The test statistic is S^2 and the rejection region is $[c, \infty)$ in case (i), $[0, c]$ in case (ii) and $[0, c_1] \cup [c_2, \infty)$ in case (iii). To connect the parameters c, c_1, c_2 to the error probabilities we construct the pivot $(n-1)\frac{S^2}{\sigma^2}$ whose distribution is χ^2 with $n-1$ dof. For example, in case (i) we have

$$\alpha = P(S^2 \geq c|\sigma_0) = P\left((n-1)\frac{S^2}{\sigma_0^2} \geq (n-1)\frac{c}{\sigma_0^2}\right) = P\left(W \geq (n-1)\frac{c}{\sigma_0^2}\right)$$

leading to $(n-1)\frac{c}{\sigma_0^2} = \chi^2_{(\alpha)}$ or $c = \chi^2_{(\alpha)}\sigma_0^2/(n-1)$ where $\chi^2_{(\alpha)}$ is the critical α-value of the χ^2 distribution. In the case (iii) we have

$$1-\alpha = P(c_1 < S^2 < c_2|\sigma_0) = P\left((n-1)\frac{c_1}{\sigma_0^2} < (n-1)\frac{S^2}{\sigma_0^2} < (n-1)\frac{c_1}{\sigma_0^2}\right) = P\left((n-1)\frac{c_1}{\sigma_0^2} < W < (n-1)\frac{c_1}{\sigma_0^2}\right)$$

whose symmetric solution for c_1, c_2 is $(n-1)\frac{c_1}{\sigma_0^2} = \chi^2_{(1-\alpha/2)} \Rightarrow c_1 = \chi^2_{(1-\alpha/2)}\sigma_0^2/(n-1)$ and $(n-1)\frac{c_2}{\sigma_0^2} = \chi^2_{(\alpha/2)} \Rightarrow c_2 = \chi^2_{(\alpha/2)}\sigma_0^2/(n-1)$. Note that the two-sided solution is different here than the t or z tests since the χ^2 distribution is not symmetric (like the t or normal distributions) and the pdf is non-zero only for positive values.

Two-Sample Tests

A variant of the tests above is when we have two samples from different distributions $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and we want to compare whether the two means (variances) are the same or not.

We start with tests for means: $H_0 : \mu_X = \mu_Y$ vs. (i) $H_A : \mu_X > \mu_Y$ or (ii) $H_A : \mu_X < \mu_Y$ or (iii) $H_A : \mu_X \neq \mu_Y$. Instead we tests for the following equivalent hypotheses $H_0 : \mu_X - \mu_Y = 0$ vs. (i) $H_A : \mu_X - \mu_Y > 0$ or (ii) $H_A : \mu_X - \mu_Y < 0$ or (iii) $H_A : \mu_X - \mu_Y \neq 0$. The test statistic is $\bar{X} - \bar{Y}$ and the rejection regions have the same form as for the z or t tests above.

Example 1: σ_X, σ_Y are known but the precise distributions are unknown. We use the following approximately $N(0,1)$ pivot $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}$. For example in case (i) this leads to

$$\alpha = P(\bar{X} - \bar{Y} \geq c|\mu_X - \mu_Y = 0) = P\left(\frac{\bar{X}-\bar{Y}-0}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}} \geq \frac{c-0}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}\right) \approx P\left(Z \geq \frac{c}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}\right)$$

which yields the solution $z_\alpha = \frac{c}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}} \Rightarrow c = z_\alpha\sqrt{\sigma_X^2/n + \sigma_Y^2/m}$.

Example 2: It is known that $X_1,\ldots,X_n \sim N(\mu_X, \sigma_X^2)$, $Y_1,\ldots,Y_n \sim N(\mu_Y,\sigma_Y^2)$ with σ_X^2, σ_Y^2 known. The test statistic is $\bar{X} - \bar{Y}$ and the shape of the rejection regions are as before. We use the precise (small sample) pivot $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}} \sim N(0,1)$. This is exactly the same pivot as example 1 and therefore the answers will be the same, only this time they will be precise rather than approximate.

Example 3: It is known that $X_1,\ldots,X_n \sim N(\mu_X,\sigma_X^2)$, $Y_1,\ldots,Y_n \sim N(\mu_Y,\sigma_Y^2)$ but $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ is not-known (but the two variances are equal). The test statistic is $\bar{X} - \bar{Y}$ and the shape of the rejection regions are as before. To construct a pivot recall that if $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$ then $\frac{Z}{\sqrt{W/\nu}}$ has a t distribution with ν dof. For Z we use $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}} \sim N(0,1)$ and for W we use $(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2}$ which is a sum of $\chi^2_{(n-1)}$ RV and a $\chi^2_{(m-1)}$ and therefore is a $\chi^2_{(n+m-2)}$ RV. Putting this all together our $t_{(n+m-2)}$ pivot is

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}} \Big/ \sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{\sigma^2(n+m-2)}} = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{1/n+1/m}} \cdot \sqrt{\frac{(n+m-2)}{(n-1)S_1^2+(m-1)S_2^2}}$$
$$= \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p\sqrt{1/n+1/m}}$$

where S_p is the pooled variance estimator (a weighted average of the two variance estimators) $S_p = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$. For example in case (iii) this leads to

$$1-\alpha = P(c_1 < \bar{X}-\bar{Y} < c_2 | \mu_X - \mu_Y = 0) = P\left(\frac{c_1 - 0}{S_p\sqrt{1/n+1/m}} < \frac{\bar{X}-\bar{Y}-0}{S_p\sqrt{1/n+1/m}} < \frac{c_2-0}{S_p\sqrt{1/n+1/m}}\right)$$
$$= P\left(\frac{c_1}{S_p\sqrt{1/n+1/m}} < t_{(n+m-2)} < \frac{c_2}{S_p\sqrt{1/n+1/m}}\right)$$

and solving for c_1, c_2 in terms of the critical values of the t-distribution yields $c_1 = -t_{\alpha/2}S_p\sqrt{1/n+1/m}$, $c_2 = t_{\alpha/2}S_p\sqrt{1/n+1/m}$.

We now consider the case where $X_1,\ldots,X_n \sim N(\mu_X,\sigma_X^2)$, $Y_1,\ldots,Y_n \sim N(\mu_Y,\sigma_Y^2)$ but we wish to test for the variances $H_0 : \sigma_X^2 = \sigma_Y^2$ vs. (i) $H_A : \sigma_X^2 > \sigma_Y^2$ (ii) $H_A : \sigma_X^2 < \sigma_Y^2$ or (iii) $H_A : \sigma_X^2 \neq \sigma_Y^2$. Instead we test the equivalent hypotheses $H_0 : \sigma_X^2/\sigma_Y^2 = 1$ vs. (i) $H_A : \sigma_X^2/\sigma_Y^2 > 1$ or (ii) $H_A : \sigma_X^2/\sigma_Y^2 < 1$ or (iii) $H_A : \sigma_X^2/\sigma_Y^2 \neq 1$. The test statistic is S_1^2/S_2^2 and the rejection region is $[c,\infty)$ (in case (i)), $[0,c]$ (in case (ii)) or $[0,c_1] \cup [c_2,\infty)$ (in case (iii)). The pivot is the following ratio of two independent (normalized by their dof) chi^2 distributions which is therefore an F-distribution RV (see note on sampling distributions)

$$\frac{(n-1)S_1^2}{\sigma_X^2(n-1)} \Big/ \frac{(m-1)S_2^2}{\sigma_Y^2(m-1)} = \frac{S_1^2\sigma_Y^2}{S_2^2\sigma_X^2} \sim F_{m-1}^{n-1}.$$

For example, in case (i) we have

$$\alpha = P(S_1^2/S_2^2 \geq c | \sigma_X^2/\sigma_Y^2 = 1) = P\left(\frac{S_1^2}{S_2^2}\cdot 1 \geq c \cdot 1\right)$$

which leads to $c = F_{m-1,\alpha}^{n-1}$ (the critical α-value of the F_{m-1}^{n-1} distribution).