

Hypothesis Testing

Gary D. Knott, Ph.D.
Civilized Software, Inc.
12109 Heritage Park Circle
Silver Spring, MD 20906 USA
Tel. (301) 962-3711
Email: csi@civilized.com
URL: www.civilized.com

Hypothesis testing is a major paradigm in statistics. It is closely linked with the computation of specified probability distribution functions. The basic notion is simple. We obtain a sample value v of a random variable T and we ask how probable it is that the sample value v would appear *under a hypothesis H which makes the distribution of T well-defined*. If the probability that, under the hypothesis H , v or a more extreme value of T appears is small, we take this as evidence that the hypothesis H is unlikely to be true. In other words, we conclude that the test of the hypothesis H has not supported H .

The random variable T is called the *test statistic*. If many samples of various random variables are taken, they are often combined in some, possibly quite elaborate, manner to obtain a single sample of a derived test statistic T . In other cases, the test statistic may be a vector-valued random variable with a multivariate distribution function. For example, the test statistic associated with the famous t -test for testing the hypothesis that two normally-distributed random variables have the same mean is the difference between the means, or variance-adjusted means, of two sets of sample values corresponding to the two random variables being studied.

In order to compute the probability p that, under the hypothesis H , the sample value v or a more extreme value of T appears, we must be able to compute $P(T \leq v \mid H)$, which is the distribution of the test statistic T under the hypothesis H . We shall denote a random variable with this distribution by T_H . The hypothesis H must be such that the distribution function of T_H is known; this means that H is often of the form: “there is no difference between two sets of samples”, since it is generally easier to deduce the distribution of T_H in this case. Thus, H is called the null hypothesis, meaning the “no difference” hypothesis.

Suppose that the distribution function of T_H is $G(x) := P(T_H \leq x)$. Also suppose that the density function $dG(x)/dx$ is a unimodal “bell-shaped” curve, so that the extreme sample values of T_H lie toward $+\infty$ and $-\infty$. Suppose the value v is given as a sample value of T . We may compute, for example, $p = P(|T_H - E(T_H)| \geq |v - E(T_H)|)$. This is a particular form of a so-called two-tailed test. p is the probability that the value v or a “more extreme” value occurs as a sample value of T , given H . If p is sufficiently small, we may reject the null hypothesis H as implausible in the face of the “evidence” v . We call such a probability p the *plausibility probability* of H , given v .

If the test statistic T_H were known to be non-negative and the density function $dG(x)/dx$ were a function, such as the exponential density function, which decreases on $[0, \infty)$, then we might use a so-called one-tail test, where we compute the probability $p = P(T_H \geq v)$.

In general, we may specify a particular value α as our criterion of “sufficiently small” and we may choose any subset S of the range of T such that $P(T_H \notin S) = \alpha$. Then if $v \notin S$, the null hypothesis H may be judged implausible. S is called the acceptance set, because, when $v \in S$, the null hypothesis H is not rejected. The value $\alpha = P(T_H \notin S)$ is the probability that we make a mistake if we reject H when $v \notin S$.

How should the acceptance set S be chosen? S should be chosen to minimize the chance of making the mistake of accepting H when H is, in fact, false. But, this can only be done rigorously with respect to an alternative hypothesis H_a such that the distribution of T given H_a is known. We must postulate that H and H_a are the only non-negligible possibilities. Sometimes, $H_a = \neg H$ is a suitable alternate hypothesis, but more often, this is not suitable. Given H_a , the probability we falsely accept H when the alternate hypothesis H_a is true is $P(T_{H_a} \in S) =: \beta$, and we can choose S such that $P(T_H \notin S) = \alpha$ while $P(T_{H_a} \in S) = \beta$ is minimized.

The value $P(T_{H_a} \notin S) = 1 - \beta$ is called the *power* of the test of the null hypothesis H versus the alternate hypothesis H_a . Choosing S to minimize β is the same as choosing S to maximize the power $1 - \beta$.

If we don’t care about achieving the optimal power of the test with respect to a specific alternate hypothesis, but merely wish to compute the plausibility probability that v or a more extreme sample value of T would occur given H , in a fair manner, then we may proceed as follows.

Let $m = \text{median}(T_H)$; thus, $P(T_H \geq m) = 0.5$. Now, if $v < m$, choose $r_1 = v$ and r_2 as the value such that $P(m < T_H < r_2) = P(v < T_H < m)$, otherwise choose $r_2 = v$ and choose r_1 as the value such that $P(r_1 < T_H < m) = P(m < T_H < v)$. Then the two-tail plausibility probability

$\alpha = 1 - P(r_1 < T_H < r_2)$. If $v < m$, $\alpha = 2P(T_H \leq v)$, otherwise, if $v \geq m$, $\alpha = 2(1 - P(T_H \leq v))$.

If we know that the only values more extreme than v which we wish to consider as possible are those in the same tail of the density function that v lies in, then we may compute the one-tail plausibility probability as $\alpha = P(T_H \leq v)$ if $v \leq m$ and $\alpha = P(T_H \geq v)$ if $v > m$.

Consider testing the null hypothesis H versus the alternate hypothesis H_a using a sample value v of the test random variable T with the acceptance set S . We have the following outcomes.

	$v \in S$	$v \notin S$
H	accept H prob $1 - \alpha$ correct	reject H prob α rejection error
H_a	accept H prob β acceptance error	reject H prob $1 - \beta$ correct

$$\begin{aligned}
 \alpha &= P(T_H \notin S) = P(\text{we falsely reject } H \mid H) && \text{(rejection error)} \\
 1 - \alpha &= P(T_H \in S) = P(\text{we correctly accept } H \mid H) && \text{(acceptance power)} \\
 \beta &= P(T_{H_a} \in S) = P(\text{we falsely accept } H \mid H_a) && \text{(acceptance error)} \\
 1 - \beta &= P(T_{H_a} \notin S) = P(\text{we correctly reject } H \mid H_a) && \text{(rejection power)}
 \end{aligned}$$

Let Q be the sample-space of the test statistic T . We assumed above that either $H(q) = 1$ for all $q \in Q$ or $H(q) = 0$ for all $q \in Q$, but this universal applicability of H or H_a may be relaxed. Suppose the hypothesis H and the alternate hypothesis H_a may each hold at different points of Q , so that H and H_a define corresponding complementary Bernoulli random variables on Q . Thus $H(q) = 1$ if H holds at the sample point $q \in Q$ and $H(q) = 0$ if H does not hold at the sample point q ; H_a is defined on Q in the same manner.

Let $P(\{q \in Q \mid H(q) = 1\})$ be denoted by $P(H)$ and let $P(\{q \in Q \mid H_a(q) = 1\})$ be denoted by $P(H_a)$. $P(H)$ is called the *incidence* probability of H and $P(H_a)$ is called the *incidence* probability of H_a . As before, we postulate that $P(H) = 1 - P(H_a)$. Often $P(H)$ is 0 or 1 as we assumed above, but it may be that $0 < P(H) < 1$. In this latter case, our test of hypothesis can be taken as a test of whether $H(q) = 1$ or $H_a(q) = 1$ for the

particular sample point q at hand for which $T(q) = v$; the test statistic value v may be taken as evidence serving to increase or diminish the probability of $H(q) = 1$.

Note that we cannot compute the “posterior” probability that $H(q) = 1$ (and that $H_a(q) = 0$), or conversely, *unless* we have the “prior” incidence probability of H being true in the sample-space Q . In particular, if we assume the underlying sample-space point q is chosen at random, then:

$$\begin{aligned} P(H(q) = 1 \quad \& \quad T(q) \in S) &= (1 - \alpha)P(H) \\ P(H(q) = 1 \quad \& \quad T(q) \notin S) &= \alpha P(H) \\ P(H(q) = 0 \quad \& \quad T(q) \in S) &= \beta(1 - P(H)) \\ P(H(q) = 0 \quad \& \quad T(q) \notin S) &= (1 - \beta)(1 - P(H)) \end{aligned}$$

If we take the occurrence of $H(q) = 0$ as being a determination of a “positive state” of the random sample-space point q , then $(1 - \alpha)P(H)$ is the probability of a *true negative* sample, $\alpha P(H)$ is the probability of a *false positive* sample, $\beta(1 - P(H))$ is the probability of a *false negative* sample, and $(1 - \beta)(1 - P(H))$ is the probability of a *true positive* sample.

Now let us look at a particular case of an hypothesis test, namely the so-called F -test for equal variances of two normal populations.

Suppose $X_{11}, X_{12}, \dots, X_{1n_1}$ are independent identically-distributed random variables distributed as $N(\mu_1, \sigma_1^2)$, and $X_{21}, X_{22}, \dots, X_{2n_2}$ are independent identically-distributed random variables distributed as $N(\mu_2, \sigma_2^2)$. The corresponding sample-variance random variables are

$$S_1^2 = \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2 / (n_1 - 1) \quad \text{and} \quad S_2^2 = \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2 / (n_2 - 1),$$

where $\bar{X}_1 = \sum_{j=1}^{n_1} X_{1j} / n_1$ and $\bar{X}_2 = \sum_{j=1}^{n_2} X_{2j} / n_2$.

Let R denote the sample variance ratio S_1^2 / S_2^2 . Then $R \sim (\sigma_1^2 / \sigma_2^2) F_{n_1-1, n_2-1}$, where F_{n_1-1, n_2-1} is a random variable having the F -distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

We take the null hypothesis H to be $\sigma_1 / \sigma_2 = 1$, so that, given H , the test statistic R is known to be distributed as F_{n_1-1, n_2-1} . In order to determine the acceptance region S with maximal power for α fixed, we take the alternate hypothesis H_a to be $\sigma_1 / \sigma_2 = a$. Then S is the interval $[r_1, r_2]$ where $P(r_1 / a^2 \leq F_{n_1-1, n_2-1} \leq r_2 / a^2) = \beta$ is minimal, subject to $1 - P(r_1 \leq F_{n_1-1, n_2-1} \leq r_2) = \alpha$.

Let $G(z) = P(F_{n_1-1, n_2-1} \leq z)$, the distribution function of F_{n_1-1, n_2-1} , and let $g(z) = G'(z)$, the probability density function of F_{n_1-1, n_2-1} . Then,

we have $r_1 = \text{root}_z[g(z)g(h(z)/a^2) - g(h(z))g(z/a^2)]$ and $r_2 = h(r_1)$, where $h(z) = G^{-1}(1 - \alpha + G(z))$.

A simplified, slightly less powerful, way to choose the acceptance region S is to take $S = [r_1, r_2]$ where r_1 is the value such that $P(F_{n_1-1, n_2-1} \leq r_1) = \alpha/2$ and r_2 is the value such that $P(F_{n_1-1, n_2-1} \geq r_2) = \alpha/2$. Another way to select the acceptance region is to take $S = [1/r, r]$, where r is the value such that $P(1/r \leq F_{n_1-1, n_2-1} \leq r) = 1 - \alpha$. When $n_1 = n_2$, the acceptance region $[r_1, r_2]$ and the acceptance region $[1/r, r]$ are identical.

The foregoing clearly exemplifies the fact that there is a trade-off among the acceptance error probability β , the rejection error probability α , and the sample sizes (n_1, n_2) . If we wish to have a smaller α , then we must have a greater β or greater values of n_1 and n_2 . Similarly, β can only be reduced if we allow α or n_1 and n_2 to increase. In general, given any two of the test parameters α , β , or (n_1, n_2) , we can attempt to determine the third, although a compatible value need not exist. Actually, in most cases, a fourth variable representing the distinction between the null hypothesis and the alternate hypothesis, such as the value a above, enters the trade-off balancing relations.

For the simplified two-tailed F -test with $S = [r_1, r_2]$, the relations among α , β , a , n_1 , n_2 , r_1 , and r_2 are listed below.

$$\begin{aligned} P(F_{n_1-1, n_2-1} \leq r_1) &= \alpha/2, \\ P(F_{n_1-1, n_2-1} \geq r_2) &= \alpha/2, \\ P(r_1 \leq a^2 F_{n_1-1, n_2-1} \leq r_2) &= \beta. \end{aligned}$$

For the alternate case with $S = [1/r, r]$, the relations among α , β , a , n_1 , n_2 , and r are:

$$\begin{aligned} P(1/r \leq F_{n_1-1, n_2-1} \leq r) &= 1 - \alpha \\ P(1/r \leq a^2 F_{n_1-1, n_2-1} \leq r) &= \beta. \end{aligned}$$

In order to reduce the number of unknowns, we may postulate that n_1 and n_2 are related as $n_2 = \theta n_1$, where θ is a fixed constant.

The value of β is determined above by the distribution of $a^2 F_{n_1-1, n_2-1}$, because for the F -test, the test statistic, assuming the alternate hypothesis $\sigma_1/\sigma_2 = a$, is the random variable $a^2 F_{n_1-1, n_2-1}$ whose distribution function

is just the F -distribution with the argument scaled by $1/a^2$. In other cases, the distribution of the alternate hypothesis test statistic T_{H_a} is more difficult to obtain.

If we take the dichotomy $\sigma_1/\sigma_2 = 1$ vs. $\sigma_1/\sigma_2 = a$ as the *only* two possibilities, then a one-tailed F -test is most appropriate. Suppose $a > 1$. Then we take $S = [-\infty, r_2]$, and we have the relations: $P(F_{n_1-1, n_2-1} \geq r_2) = \alpha$, and $P(a^2 F_{n_1-1, n_2-1} \leq r_2) = \beta$. If $a < 1$, then with the null hypothesis $\sigma_1/\sigma_2 = 1$ and the alternate hypothesis $\sigma_1/\sigma_2 = a$, we should take $S = [r_1, \infty]$. Then, $P(F_{n_1-1, n_2-1} \leq r_1) = \alpha$, and $P(a^2 F_{n_1-1, n_2-1} \geq r_1) = \beta$.

Generally, hypothesis testing is most useful when a decision is to be made. Instead, for example, suppose we are interested in the variance ratio $(\sigma_1/\sigma_2)^2$ between two normal populations for computational purposes. Then it is preferable to use estimation techniques and confidence intervals to characterize (σ_1/σ_2) , rather than to use a hypothesis test whose only useful outcome is “significantly implausible”, or “not significantly implausible” with the significance level α (which is the same as the rejection error probability).

Let r_1 satisfy $P(F_{n_1-1, n_2-1} \leq r_1) = \alpha_1$, and let r_2 satisfy $P(F_{n_1-1, n_2-1} \leq r_2) = 1 - \alpha_2$, with $\alpha_1 + \alpha_2 = \alpha < 1$. Then $P((\sigma_1/\sigma_2)^2 r_1 > R \text{ or } R > (\sigma_1/\sigma_2)^2 r_2) = \alpha_1 + \alpha_2$, and $P((\sigma_1/\sigma_2)^2 r_1 < R \text{ and } R < (\sigma_1/\sigma_2)^2 r_2) = 1 - \alpha_1 - \alpha_2 = P((\sigma_1/\sigma_2)^2 < R/r_1 \text{ and } R/r_2 < (\sigma_1/\sigma_2)^2) = P(R/r_2 < (\sigma_1/\sigma_2)^2 < R/r_1)$.

Thus, $[R/r_2, R/r_1]$ is a $(1 - \alpha)$ -confidence interval which is an interval-valued random variable that contains the true value (σ_1/σ_2) with probability $1 - \alpha$. The length of this interval is minimized for $n_2 > 2$ by choosing α_1 and α_2 , subject to $\alpha_1 + \alpha_2 = \alpha$, such that $G^{-1}(\alpha_1)^2 g(G^{-1}(1 - \alpha_2)) - G^{-1}(1 - \alpha_2)^2 g(G^{-1}(\alpha_1)) = 0$, where $G(x) = P(F_{n_1-1, n_2-1} \leq x)$ and where $g(x) = G'(x)$, the probability density function of F_{n_1-1, n_2-1} . Then $\alpha_1 = \text{root}_z(G^{-1}(z)^2 g(G^{-1}(1 + \alpha - z)) - G^{-1}(1 - \alpha + z)^2 g(G^{-1}(z)))$, and $\alpha_2 = \alpha - \alpha_1$, $r_1 = G^{-1}(\alpha_1)$, and $r_2 = G^{-1}(1 - \alpha_2)$.

Let v denote the observed sample value of R . Then $[v/r_2, v/r_1]$ is a sample $(1 - \alpha)$ -confidence interval for $(\sigma_1/\sigma_2)^2$.

The MLAB mathematical and statistical modeling system contains functions for various statistical tests and also functions to compute associated power and sample-size values. Let us consider an example focusing on the simplified F -test discussed above. We are given the following data:

x1: -1.66, 0.46, 0.15, 0.52, 0.82, -0.58, -0.44, -0.53, 0.4, -1.1

x2: 3.02, 2.88, 0.98, 2.01, 3.06, 2.95, 3.4, 2.76, 3.92, 5.02, 4, 4.89, 2.64, 3.08

We may read this data into two vectors in MLAB and test whether the two data sets x_1 and x_2 have equal variances by using the MLAB F -test function /QFT/, which implements the $[1/r, r]$ simplified F -test specified above. The MLAB dialog to do this is exhibited below

```
* x1 = read(x1file); x2 = read(x2file);
* qft(x1,x2)
```

```
[F-test: are the variances v1 and v2 of 2 normal populations
plausibly equal?]
```

```
null hypothesis H0: v1/v2 = 1. Then v1/v2 is F-distributed with
(n1-1,n2-1) degrees of freedom. n1 & n2 are the sample sizes.
The sample value of v1/v2 = 0.577562, n1 = 10, n2 = 15
```

```
The probability P(F < 0.577562) = 0.205421
This means that a value of v1/v2 smaller than 0.577562 arises about
20.542096 percent of the time, given H0.
```

```
The probability P(F > 0.577562) = 0.794579
This means that a value of v1/v2 larger than 0.577562 arises about
79.457904 percent of the time, given H0.
```

```
The probability: 1-P(0.577562 < F < 1.731416) = 0.377689
This means that a value of v1/v2 more extreme than 0.577562 arises
about 37.768896 percent of the time, given H0.
```

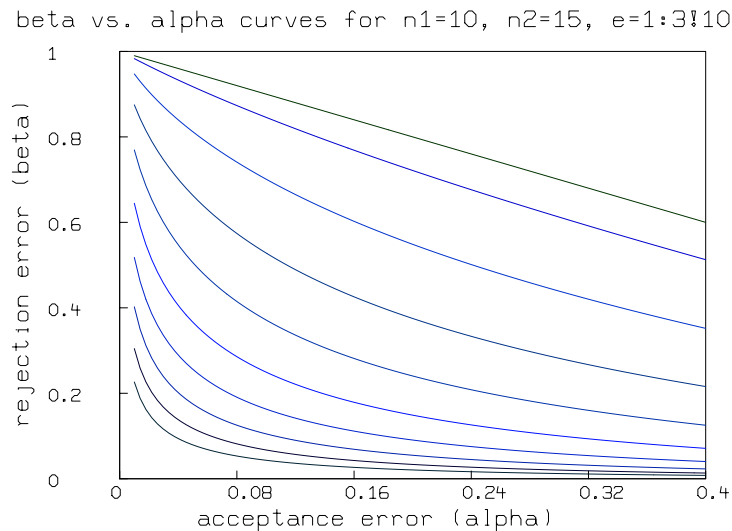
The $\alpha = .05$ simplified F -test acceptance set of the form $[1/r, r]$ can be computed directly as follows. /QFF/ is the name of the F -distribution function in MLAB.

```
* n1 = nrow(x1)-1; n2 = nrow(x2)-1;
* fct rv(a) = root(z,.001,300,qff(1/z,n1,n2)+1-qff(z,n1,n2)-a)
* r = rv(.05)
* type 1/r,r
    = .285471776
    R = 3.50297326
```

Thus, a sample of $F_{9,14}$ will lie in $[.2855, 3.503]$ with probability .95.

The rejection error probability β can be plotted as a function of the acceptance error probability α for the sample sizes 10 and 15 by using the builtin function /QFB/ as follows. The function /QFB/(α, n, θ, e) returns the rejection error probability value β that corresponds to the sample sizes n and θn , with the acceptance error probability α and the alternate hypothesis variance ratio e .

```
* fct b(a) = qfb(a,10,3/2,e)
* for e = 1:3!10 do {draw points(b,.01:.4!100)}
* left title "rejection error (beta)"
* bottom title "acceptance error (alpha)"
* top title "beta vs. alpha curves for n1=10, n2=15, e=1:3!10"
* view
```



Suppose we want to take n samples from each of two populations to be used to test whether these populations have the variance ratio 1 versus the variance ratio e , with acceptance error probability $\alpha = .05$ and rejection error $\beta = .05$. We can use the builtin function /QFN/ to compute the sample size n as a function of e as follows. The function /QFN/(α, β, θ, e) returns the sample size n that corresponds to the variance ratio numerator sample size, assuming the denominator

sample size θn , and given that the acceptance error probability is α , the rejection error probability is β , and the alternate hypothesis variance ratio value is e .

```
* fct n(e) = qfn(.05,.05,1,e)
* draw points(n,.1:2.5!50)
* top title "sample size vs. variance ratio (with a=b=.05,t=1)"
* left title "sample size (n)"
* bottom title "variance ratio (e)"
* view
```

