

Hypothesis Testing

Vincent A. Voelz

E-mail: vvoelz@gmail.com

[†]*Math Bio Boot Camp 2006*

University of California at San Francisco, San Francisco, CA 94143

August 30, 2006

Contents

1	Introduction to Statistical tests	2
1.1	Hypothesis Testing	2
1.2	p -values	4
2	Parametric Tests	5
2.1	Parametric versus Non-parametric Tests	5
2.2	The χ^2 -test	5
2.3	The ‘Student’ t -test	6
2.4	The two-sample t -test	7
2.5	Likelihood Ratios	7
2.6	The Kolmogorov-Smirnov test	8
3	Non-parametric tests	9
3.1	Wilcoxon rank-sum test	10
3.2	Permutation tests	10

3.3	Bootstrapping: resampling with replacement	10
3.4	Jackknifing	11

1 Introduction to Statistical tests

The underlying concept in statistics is quite easy to understand. A **statistic** is simply a computed value, $\theta(\{x_1, x_2, \dots\})$, that characterizes a particular data sample $\{x_1, x_2, \dots\}$. If we have some model of the theoretical probability distribution of this statistic, $P(\theta)$, we can assess its significance by considering the likelihood of our statistic's value.

This process of assessment is surprisingly arbitrary. We can set up any number of conditions to evaluate whether our statistic is significant, and by how much. For example, it is common to declare that if a statistic's value is $\pm 2\sigma$ away from the hypothetical mean value expected by random chance, then it is significant. Perhaps $\pm 3\sigma$ is a safer tolerance; it's all in how we define "significance".

In classical statistics, the hypothesized distribution of a statistic has usually been derived using the *de facto* assumption that underlying variations are normally distributed (a good assumption, it turns out, according to the Central Limit Theorem). The assumption of normally distributed variables leads to the celebrated χ^2 - and t -distributions. Other kinds of statistics have different assumptions, which lead to other kinds of distributions. We may also want to consider the case where we have no model of the underlying distribution of our statistics, and instead wish to empirically construct a probability distribution directly from the sampled data.

1.1 Hypothesis Testing

Hypothesis testing is the rational framework for applying statistical tests.

The main question we usually wish to extract from a statistic is whether the sample data is *significant* or not. For example, let's say we have a hat with two kinds of numbers in it: some of the numbers are drawn from a standard normal distribution (i.e. $\sigma^2 = 1$) with mean $\mu = 0$, and some of the numbers are drawn from a standard normal distribution with unknown mean.

Now let's say we take a number out of the hat. There are two hypotheses that are possible:

- H_0 : **the null hypothesis**. The number is from a standard normal distribution with $\mu = 0$.
- H_A : **the alternative hypothesis**. The number is *not* from a standard normal distribution with $\mu = 0$.

The art of statistics is in finding good ways of formulating criteria, based on the value of one more statistics, to either *accept* or *reject* the null hypothesis H_0 .

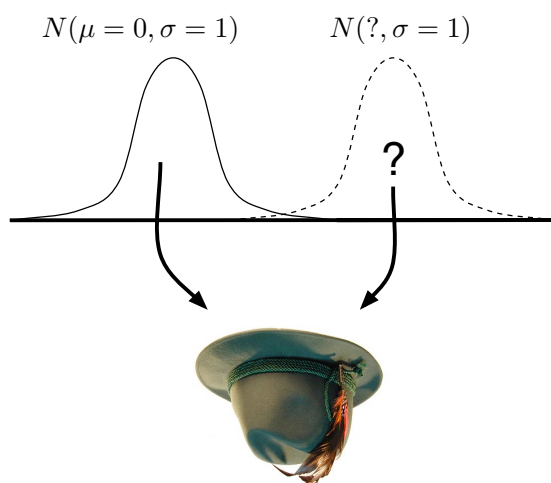


Figure 1: Numbers drawn from two different standard normal distributions are thrown into John Chodera's hat.

It should be noted that H_0 and H_A can be almost anything, and as complicated or as simple as we wish. If a hypothesis is stated such that it specifies the entire distribution, we call it a **simple** hypothesis. Otherwise, we call it a **composite** hypothesis. As you might imagine, more rigorous tests can be done with simple hypotheses, because they specify the entire distribution, from which probability values can be computed.

Type I and Type II errors. In any testing situation, two kinds of error could occur:

- **Type I (false positive).** We reject the null hypothesis when it's actually true.
- **Type II (false negative).** We accept the null hypothesis when it's actually false.

There is no way to completely eliminate both kinds of errors. For instance, say we draw the number 4 out of the hat, we may reject the null hypothesis when actually it was generated from the $N(0, 1)$ distribution hypothesized in H_0 , and just happens to be $+4\sigma$ away from the mean. This would be a *Type I error*. Similarly, we might draw the number 0 out of the hat, and incorrectly accept the null hypothesis when actually it was generated from a $N(\mu = 3, 1)$ distribution, and just happens to be -3σ away from the mean. This would be a *Type II error*.

The probability of committing a Type I error is typically denoted α , and the probability of a Type II error is denoted β .

- α : the probability of making a Type I error (false positive).
- β : the probability of making a Type II error (false negative).

α is often called a *significance level* or *sensitivity*. Typically, we try to fix an accepted level, α of Type I error, and go on to find ways of minimizing the level of Type II error, β .

The statistical **power** of a test is defined as $(1 - \beta)$. We usually want to maximize the power of our test in order to detect as many significant signals from our data as we possibly can.

1.2 p -values

Let's say we have a statistic x , and the null hypothesis is that x comes from a standard normal probability distribution $P(x) = N(\mu = 0, \sigma^2 = 1)$. We want to choose a critical value of x , call it x^* , that we can use as a criterion rejecting the null hypothesis:

- *CRITERION*: Reject the null hypothesis if $x \geq x^*$.

To reach a target significance level α that we have chosen beforehand, we need to choose x^* such that the probability of observing our statistic's value, or anything greater, is α :

$$\int_{x^*}^{\infty} P(x) dx = \alpha$$

For a standard normal distribution, it turns out that if you choose a critical value of $x^* \approx +2\sigma$, a $\alpha = 0.05$ will result.

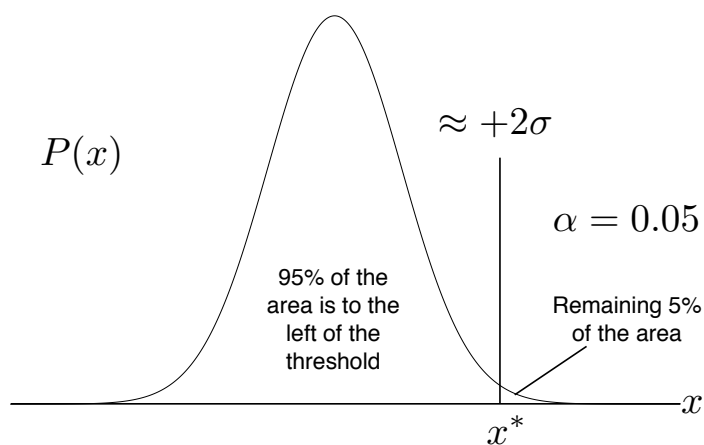


Figure 2: For a null hypothesis that assumes that our statistic x is drawn from a standard normal distribution, choosing a threshold value of $x^* = +2\sigma$ gives a significance level of about $\alpha = 0.05$.

The **p -value** is the way to test against this significance level α . Suppose you measure your statistic to be x^\dagger . Then the p -value for your statistic is:

$$p_{x^\dagger} = P(x \geq x^\dagger | H_0) = \int_{x^\dagger}^{\infty} P(x) dx$$

If the p -value is less than α , then you might want to consider rejecting the null hypothesis. If the p -value is *considerably* less than α , then the statistic is highly significant (at least compared to the significance of the hypothesis test).

(*Note:* the p -value is *not* the probability of observing this data, $P(x^\dagger)$, rather it is the probability of observing a value of x that is at least x^\dagger or larger, $P(x \geq x^\dagger)$.)

This is the general way most statistical tests are applied. Each statistical test consists of a *statistic* that you compute, and (usually) a known distribution of this statistic, which is the null hypothesis. (I say “usually” because in some case you have to build this distribution empirically from the data). *More about this in the next section.*) You first:

1. pick an appropriate significance level, α , and then
2. calculate (or look up from a table) the p -value of your observed statistic.
3. If the p -value is less than α , then reject the null hypothesis.

In the sections below we will describe off number of different kinds of statistical tests. Keep in mind that they are all used according to this same general recipe.

2 Parametric Tests

2.1 Parametric versus Non-parametric Tests

In general, there are two kinds of statistical tests. *Classical* statistics mostly deals with **parametric tests**. These are tests which assume some sort of model for the underlying distribution that the sample data is drawn from. Many of the statistical distributions used in these tests assume that the data is drawn from a normal (Gaussian) distribution. Given this assumption, much can be derived about the distribution of the observations themselves.

Non-parametric tests do not assume any kind of underlying probability distribution. This can be very useful in cases where it would be very hard to justify that the data are normally-distributed (or if we know it’s just plain not true). Many non-parametric tests can quite powerful simply by considering the *rank order* of the observations.

2.2 The χ^2 -test

The χ^2 -test is a useful statistic for calculating “goodness of fit”. Suppose we have N independent variables, x_i , each normally distributed with mean μ_i and variance σ_i^2 . The random variable

$$Z = \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

Parametric Tests	Non-Parametric Tests
Assumes a model for the underlying distribution of a statistic (often Gaussian)	No assumptions about underlying distribution
Typically deals with distributions of actual observations	Typically deals with the <i>rank ordering</i> of observations

Figure 3: Some differences between parametric and non-parametric tests.

is distributed according to a χ^2 -distribution with N degrees of freedom. (See the *Statistics Primer* notes for a detailed description of this.) If you have a set of observations O_i , and a set of expected values for each of your observations E_i , the following value also has a distribution that is a very good approximation to a χ^2 -distribution:

$$Z = \sum_i^N \frac{(O_i - E_i)^2}{E_i}$$

You can estimate the statistical significance of your results by looking up p -values for the χ^2 -distribution.

2.3 The ‘Student’ t -test

The t -distribution is especially useful for testing the statistical significance of *sample means* and *sample variances*. Suppose we have a set of N independent, normally-distributed variables x_i with mean μ and variance σ^2 .

Now, if we *knew* the true variance σ , then it is straightforward to show that the renormalized variable

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$$

is normally distributed with mean 0 and variance 1. But if all we have is a sample of data, the best guess we have for the variance is the *sample variance*, s^2 :

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2 \quad (1)$$

In this case, the quantity

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$

is not quite distributed as a standard normal. Instead, it is distributed according to a **t -distribution with $(N - 1)$ degrees of freedom** (Again, see the *Statistics Primer* notes for the functional form of this distribution.)

The ***t*-test** is a statistical test that uses the sample mean and sample variance to determine whether or not a given sample comes from a normal distribution. To use the test, you calculate the statistic Z and find the p -value for it according to the t -distribution.

2.4 The two-sample t -test

Perhaps the most widely-used application of the t -distribution is to test whether or not two samples come from a distribution with the same mean. A similar reasoning as above is used here:

Consider two sets of sample data: a set of N independent, normally-distributed variables x_i with mean μ_x and variance σ_x^2 , and a set of M independent, normally-distributed variables y_i with mean μ_y and variance σ_y^2 .

Again, if we *knew* the true variance σ , then it is straightforward to show that the renormalized variable

$$Z = \frac{(\bar{x} - \mu_x) - (\bar{y} - \mu_y)}{\sigma \sqrt{1/N + 1/M}}$$

is distributed as a standard normal distribution. However, since we don't know this, we must estimate it using the *pooled sample variance*, S_p^2 :

$$S_p^2 = \frac{(N-1)s_x^2 + (M-1)s_y^2}{n+m-2}$$

where s_x^2 and s_y^2 are the sample variances for x_i and y_i , respectively.

It can then be shown that the quantity

$$Z = \frac{(\bar{x} - \mu_x) - (\bar{y} - \mu_y)}{S_p \sqrt{1/N + 1/M}}$$

is distributed as a t -distribution with $(n+m-2)$ degrees of freedom. Like before, to use the test, you calculate the statistic Z and find the p -value for it according to the t -distribution.

Trivia. The 'Student' t -statistic is sometimes called the Guinness t -test, because its inventor, William Sealy Gosset, worked for the Guinness brewery in Dublin, and formulated it to measure the statistical consistency across batches of brewed beer. Because he was publishing trade secrets, he was forced to use a pen name, 'Student'.

2.5 Likelihood Ratios

In the case that both the null hypothesis H_0 and the alternative hypothesis H_A are *simple* (i.e. they specify the entire probability distribution $P(x)$), then it can be shown that hypothesis tests using likelihood ratios give the most amount of statistical **power**, $(1 - \beta)$.

The likelihood ratio (LR) for a statistic x is defined as:

$$LR = \frac{p(x|H_A)}{p(x|H_0)}$$

In the example shown in Figure 4, when this ratio surpasses $LR > 1$, we reject the null hypothesis. This can be adjusted, depending on the desired level of significance, α .

Since the logarithm is a monotonic function, often times it is more practical to consider the *log-likelihood ratio* (LLR). When combined with Bayesian statistics, LR tests can be quite powerful.

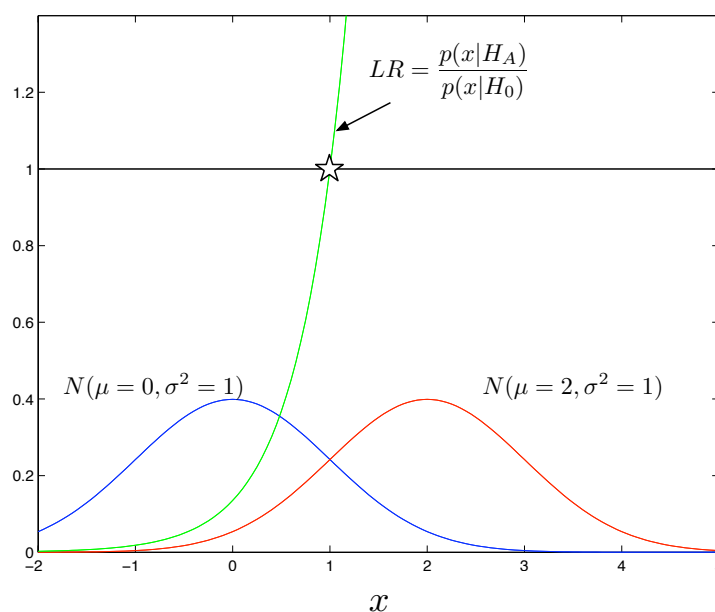


Figure 4: Here, H_0 is that a number picked from our hat is a standard normal $N(0,1)$ (blue), and H_A is that the number is from $N(2,1)$ (red). We've chosen the point where the *likelihood ratio* (LR) ≥ 1 as the best threshold for rejecting the null hypothesis.

2.6 The Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (K-S) test is used to detect *any* difference between two distributions. It is somewhat limited, because it only has statistical *power* in the context of very small sample sizes.

In the K-S test, the null hypothesis H_0 is that a sample variable x comes from a parent distribution $P(x)$, and the alternative hypothesis H_A is that it does *not* come from $P(x)$.

The K-S statistic is defined as:

$$D = \max_{-\infty < x < \infty} |S_N(x) - P(x)|$$

where $S_{N_1}(x)$ is the *cumulative probability distribution* (CDF) built up from the sampled data (see Figure 5), and $P(x)$ is the cumulative distribution hypothesized for H_0 .

Alternatively, one may wish to compare two samples directly, by calculating the maximum difference between the two different cumulative distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$. The same statistic applies:

$$D = \max_{-\infty < x < \infty} |S_{N_1}(x) - S_{N_2}(x)|$$

To a useful approximation, the K-S-distribution of the D statistic can be calculated, and the p -value for a measured D^\dagger is given by:

$$P(D \geq D^\dagger) = Q_{KS} \left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e} \right] D \right)$$

where $N_e = N$ for the case of a single sample, and $N_e = N_1 N_2 / (N_1 + N_2)$ for the two-sample case, and Q_{KS} is defined as:

$$Q_{KS}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}$$

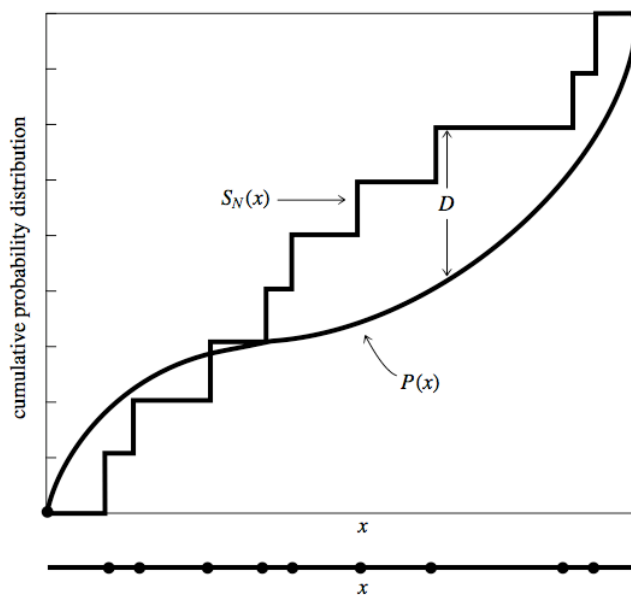


Figure 5: The K-S statistic D measures the maximum difference between the sample's CDF, $S_N(x)$, and a theoretical CDF, $P(x)$.

3 Non-parametric tests

Suppose we have gene expression data for 1000 genes suspected to be involved in breast cancer from two groups of patients. The first group is 41 patients who metastasized, and the second group is 59 patients who didn't. It would be nice to apply a parametric test like the two-sample t -test for each gene, in order to determine which gene are differentially expressed in the disease.

The problem is that gene expression data is notoriously non-normal and skewed such that we can't trust the t -test – the assumption of normally-distributed data is not valid. This is especially true if the data set contains outliers (greatly affecting the sample variance).

This is the kind of problem that may benefit from non-parametric tests. We'll use this example to illustrate the following non-parametric tests.

3.1 Wilcoxon rank-sum test

The procedure for calculating the U -statistic for the Wilcoxon rank-sum test (also known as the Mann Whitney U -test) is very conceptually simple. Consider two sets of data: N samples of x_i and M samples of y_i .

1. First, rank all of the values in the combined set of all $N + M$ samples, from 1 to $(N + M)$.
2. Choose the sample for which the ranks seem to be smaller (the choice is relevant only to ease of computation). Call this "sample 1", and call the other sample "sample 2".
3. Taking each observation in sample 1, count the number of observations in sample 2 that are smaller than it.
4. The total of these counts is the U -statistic.

The U -test is included in most modern statistical packages. Just like the other tests, you must look up the p -value of your measured U statistic from a table compiled from the U -distribution.

3.2 Permutation tests

Consider just one particular gene's expression data for the two groups of patients, $\{x_i\}, 1 \leq i \leq N$ and $\{y_j\}, 1 \leq j \leq M$. Suppose we want to test the null hypothesis:

- H_0 : the sample means of the expression data for the two patient groups are drawn from the same distribution.

One way to test this is to construct a probability distribution that corresponds to the null hypothesis, but which is built empirically from the data. This can be done by repeatedly permuting the category labels and compiling a histogram of the sample means \bar{x} and \bar{y} .

3.3 Bootstrapping: resampling with replacement

Sometimes we are very cautious about making any assumptions about the underlying distribution of sampled data. However, one recourse is to make a very weak assumption and construct a hypothetical distribution from the data itself.

Again, consider just one particular gene's expression data for the two groups of patients, $\{x_i\}, 1 \leq i \leq N$ and $\{y_j\}, 1 \leq j \leq M$. Suppose we want to construct some statistical test based on the sample mean \bar{x}

across the group of patients. A bootstrapped distribution $P(\bar{x})$ can be constructed by directly sampling *with replacement* from the set of values $\{x_i\}$; likewise $P(\bar{y})$ can be constructed directly by sampling with replacement from $\{y_j\}$. These can then be compared to null hypothesis distributions $P_0(\bar{x})$ and $P_0(\bar{y})$ which have been constructed by sampling with replacement from $\{x_i\} \cup \{y_j\}$.

(You can see why they call it “bootstrapping” – it’s almost like you’re getting something for nothing, like you’re “picking yourself up by your bootstraps”.)

3.4 Jackknifing

Jackknifed statistics are similar to bootstrapping, but instead of resampling, various subsets of the data are systematically removed, and the distribution of the resulting variations are studied. This is especially useful in characterizing how robust a given set of data is in supporting a conclusion. The robustness of calculated expectation values over time series is a good example of this.