Statistical Tests Methods and Limitations

Stephan Neuhaus

Lehrstuhl Softwaretechnik
Universität des Saarlandes, Saarbrücken







Questions To Be Answered By Statistics

- You buy two lightbulbs of the same brand, from different stores. Both burn out within a month after you start using them. Perhaps the brand is no good?







Questions To Be Answered By Statistics

- You buy two lightbulbs of the same brand, from different stores. Both burn out within a month after you start using them. Perhaps the brand is no good?
- You notice that all people of the same age seem to be about the same height. Yet you constantly meet giants or midgets. Does this mean that age and height are not related?

Questions To Be Answered By Statistics

- You buy two lightbulbs of the same brand, from different stores. Both burn out within a month after you start using them. Perhaps the brand is no good?
- You notice that all people of the same age seem to be about the same height. Yet you constantly meet giants or midgets. Does this mean that age and height are not related?
- How do you know whether a treatment for an illness is effective?

Questions To Be Answered By Statistics

- You buy two lightbulbs of the same brand, from different stores. Both burn out within a month after you start using them. Perhaps the brand is no good?
- You notice that all people of the same age seem to be about the same height. Yet you constantly meet giants or midgets. Does this mean that age and height are not related?
- How do you know whether a treatment for an illness is effective?
- How do you know whether one method for automatically finding defects in computer programs is better than another, if you can't prove it?

Questions To Be Answered By Statistics

- You buy two lightbulbs of the same brand, from different stores. Both burn out within a month after you start using them. Perhaps the brand is no good?
- You notice that all people of the same age seem to be about the same height. Yet you constantly meet giants or midgets. Does this mean that age and height are not related?
- How do you know whether a treatment for an illness is effective?
- How do you know whether one method for automatically finding defects in computer programs is better than another, if you can't prove it?

Statistical tests can answer all of these questions.







How a Test Ought to Work

1. You *choose a null hypothesis* that you want to examine. One example of a null hypothesis would be “our method is no better than theirs”.







How a Test Ought to Work

1. You *choose a null hypothesis* that you want to examine. One example of a null hypothesis would be “our method is no better than theirs”.
2. You *choose a confidence level*. That is a real number p between 0 and 1 that gives the probability with which you are willing reject the null hypothesis, even if it’s true. Typical values for p are 0.05 and 0.01 (or 5% and 1%).

How a Test Ought to Work

1. You *choose a null hypothesis* that you want to examine. One example of a null hypothesis would be “our method is no better than theirs”.
2. You *choose a confidence level*. That is a real number p between 0 and 1 that gives the probability with which you are willing reject the null hypothesis, even if it’s true. Typical values for p are 0.05 and 0.01 (or 5% and 1%).
3. You *run the tests and compute a statistic*. For example, you compute the number of 0 bits in a sample of $20,000$ bits.







How a Test Ought to Work

1. You *choose a null hypothesis* that you want to examine. One example of a null hypothesis would be “our method is no better than theirs”.
2. You *choose a confidence level*. That is a real number p between 0 and 1 that gives the probability with which you are willing reject the null hypothesis, even if it’s true. Typical values for p are 0.05 and 0.01 (or 5% and 1%).
3. You *run the tests and compute a statistic*. For example, you compute the number of 0 bits in a sample of $20,000$ bits.
4. You *compute the probability that the statistic has this value* (or is higher, or lower) *if the null hypothesis is true*. If this probability is less than p, you *reject the null hypothesis*.

More About Tests

- You never *accept* the null hypothesis; you only ever *not reject* it.







More About Tests

- You never *accept* the null hypothesis; you only ever *not reject* it.
- In practice, you'll conduct the test first and then later choose the lowest p that will not cause your null hypothesis to be rejected. Therefore, if you see a study that claims that "the hypothesis could be rejected at the 5% level", you can be sure that it *could not have been* rejected at 4%.







More About Tests

- You never *accept* the null hypothesis; you only ever *not reject* it.
- In practice, you'll conduct the test first and then later choose the lowest p that will not cause your null hypothesis to be rejected. Therefore, if you see a study that claims that "the hypothesis could be rejected at the 5% level", you can be sure that it *could not have been* rejected at 4%.
- If you see nonstandard levels (i.e., everything but 10%, 5% or 1%), beware. This is a sure sign of trying to look good.

More About Tests

- You never *accept* the null hypothesis; you only ever *not reject* it.
- In practice, you'll conduct the test first and then later choose the lowest p that will not cause your null hypothesis to be rejected. Therefore, if you see a study that claims that "the hypothesis could be rejected at the 5% level", you can be sure that it *could not have been* rejected at 4%.
- If you see nonstandard levels (i.e., everything but 10%, 5% or 1%), beware. This is a sure sign of trying to look good.
- A null hypothesis that can only be rejected at the 10% level isn't doing particularly well. Insist on 5% or better.

More About Tests

- You never *accept* the null hypothesis; you only ever *not reject* it.
- In practice, you'll conduct the test first and then later choose the lowest p that will not cause your null hypothesis to be rejected. Therefore, if you see a study that claims that "the hypothesis could be rejected at the 5% level", you can be sure that it *could not have been* rejected at 4%.
- If you see nonstandard levels (i.e., everything but 10%, 5% or 1%), beware. This is a sure sign of trying to look good.
- A null hypothesis that can only be rejected at the 10% level isn't doing particularly well. Insist on 5% or better.
- A statistical dependency is not a cause-effect chain!

Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$







Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$ or $|0.5n - k|$







Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$ or $|0.5n - k|$ or $(0.5n - k)^2$







Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$ or $|0.5n - k|$ or $(0.5n - k)^2$ or even $(0.5n - k)^2/0.5n$ (the χ^2 statistic for this case).







Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$ or $|0.5n - k|$ or $(0.5n - k)^2$ or even $(0.5n - k)^2/0.5n$ (the χ^2 statistic for this case).
- That means that generally, large values of the statistic signify large deviations from the distribution that would occur if the null hypothesis were true.





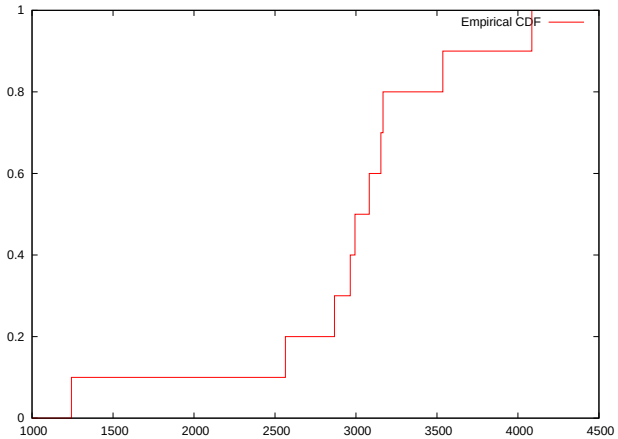


Yet More About Tests

- In general, the statistic that you compute will be some measure of the sample's deviation from the ideal. For example, if you count the number k of 0 bits in a sample of supposedly equidistributed n bits, the statistic could be $0.5n - k$ or $|0.5n - k|$ or $(0.5n - k)^2$ or even $(0.5n - k)^2/0.5n$ (the χ^2 statistic for this case).
- That means that generally, large values of the statistic signify large deviations from the distribution that would occur if the null hypothesis were true.
- Therefore, most tables of statistics are computed to answer the question, "what is the probability of the statistic being this high, or higher, if the null hypothesis is in fact true?"

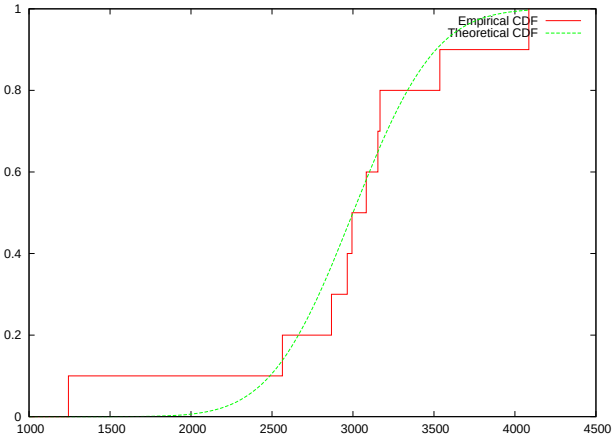
Example: Lightbulbs (1)

The following data is supposed to be from a study of lightbulb lifetime and gives the number of hours a lightbulb shone before burning out: 1243, 2564, 2867, 2965, 2994, 3082, 3154, 3167, 3536, 4086. This is the empirical distribution:



Example: Lightbulbs (1)

We suspect that the lightbulb lifetime might be distributed normally with mean $\mu = 3000$ and standard deviation $\sigma = 400$. So the conjectured theoretical CDF would be



Example: Lightbulbs (3)

If we actually perform a test, we will find that we cannot reject the null hypothesis, not at the 1% level and not at the 5% level. What does this mean?

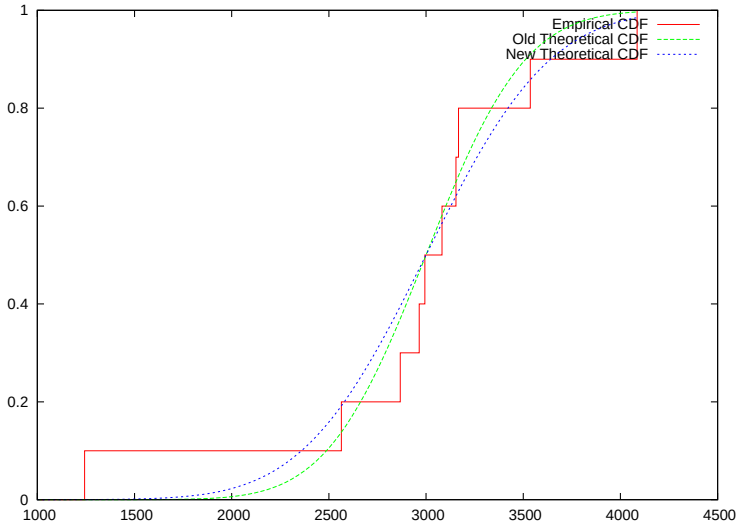






Example: Lightbulbs (3)

If we actually perform a test, we will find that we cannot reject the null hypothesis, not at the 1% level and not at the 5% level. What does this mean? Note, that a larger standard deviation would have given an even better fit.









Meaning of Tests

So the fact that you can't reject the hypothesis does not mean that it's true!







Meaning of Tests

So the fact that you can't reject the hypothesis does not mean that it's true! Also, if you *can* reject a null hypothesis, it doesn't mean it's wrong!

In fact, the null hypothesis in this case is *wrong*, because I chose the numbers to fake a normal distribution. They weren't actually drawn at random.

So what does a statistic of S and a significance of p mean? (We will have $0 \leq p \leq 1$.) It means that if we do N experiments, we can expect Np of them to have a statistic of S or more. Hence Np significances will be p or less. And *that* means that p itself is uniformly distributed between 0 and 1.

Meaning of Tests

So the fact that you can't reject the hypothesis does not mean that it's true! Also, if you *can* reject a null hypothesis, it doesn't mean it's wrong!

In fact, the null hypothesis in this case is *wrong*, because I chose the numbers to fake a normal distribution. They weren't actually drawn at random.

So what does a statistic of S and a significance of p mean? (We will have $0 \leq p \leq 1$.) It means that if we do N experiments, we can expect Np of them to have a statistic of S or more. Hence Np significances will be p or less. And *that* means that p itself is uniformly distributed between 0 and 1.

And this we can, of course, test.

Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases.







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.

Data is continuous if it is most naturally measured with a real number.







Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.

Data is continuous if it is most naturally measured with a real number. Examples: lifetime of a lightbulb

Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.

Data is continuous if it is most naturally measured with a real number. Examples: lifetime of a lightbulb, dick size

Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.

Data is continuous if it is most naturally measured with a real number. Examples: lifetime of a lightbulb, dick size, height of people

Types of Data

Statistical data comes essentially in two types: binned (a.k.a. discrete) and continuous.

Binned data falls naturally into a (usually rather small) number of discrete cases. Examples: the number of points on the top surface of a die, shoe size, number of students that fit into a telephone booth, number of cases where a method led directly to the defect.

Data is continuous if it is most naturally measured with a real number. Examples: lifetime of a lightbulb, dick size, height of people, running time of a program.

More on Binned/Continuous Data

Data that is discrete but that has a very large number of bins is effectively continuous! (This is often forgotten.)

More on Binned/Continuous Data

Data that is discrete but that has a very large number of bins is effectively continuous! (This is often forgotten.)

You can always bin data that is really unbinned, but that loses information; besides the tests for continuous data are not more difficult that the tests for discrete data, so why bother?

More on Binned/Continuous Data

Data that is discrete but that has a very large number of bins is effectively continuous! (This is often forgotten.)

You can always bin data that is really unbinned, but that loses information; besides the tests for continuous data are not more difficult that the tests for discrete data, so why bother?

Sometimes, you measure more than one attribute at the same time. Some attributes may be continuous, others may be discrete. We won't cover such multivariate analyses here.







PDF and CDF

Let X be a random variable with discrete values, then its *probability density function* (PDF) is defined as

$$\phi(x) = \Pr(X = x).$$

If X is any random variable (continuous or discrete), then the *cumulative distribution function* (CDF) for X is defined as

$$\Phi(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \phi(u)\, du.$$

If X is a continuous random variable with a differentiable CDF, then its PDF is

$$\phi(x) = d\Phi(x)/dx.$$

Interesting Questions

- Do two samples have the same mean?







Interesting Questions

- Do two samples have the same mean?
- Do two samples have the same variance?







Interesting Questions

- Do two samples have the same mean?
- Do two samples have the same variance?
- Does a sample have a specified distribution?







Interesting Questions

- Do two samples have the same mean?
- Do two samples have the same variance?
- Does a sample have a specified distribution?
- Do two samples have the same distribution?







Tests For Same Mean: t Test

Idea behind the test: take two samples A and B of size N_A and N_B, respectively, and see how many "standard errors" the two sample means μ_A and μ_B are apart. (The null hypothesis is that the two means are the same.)

Tests For Same Mean: t Test

Idea behind the test: take two samples A and B of size N_A and N_B, respectively, and see how many "standard errors" the two sample means μ_A and μ_B are apart. (The null hypothesis is that the two means are the same.)

The standard error is a measure of the accuracy with which the sample mean approximates the expected value. In this way, large samples get more significance than small samples.







Tests For Same Mean: t Test

Idea behind the test: take two samples A and B of size N_A and N_B, respectively, and see how many "standard errors" the two sample means μ_A and μ_B are apart. (The null hypothesis is that the two means are the same.)

The standard error is a measure of the accuracy with which the sample mean approximates the expected value. In this way, large samples get more significance than small samples.

If the two distributions have (or are thought to have) the same variance, we can estimate the stadard error of the difference of the means by

$$s_D = \sqrt{\frac{\sum_{i\in A}(x_i - \mu_A)^2 + \sum_{i\in B}(x_i - \mu_B)^2}{N_A + N_B - 2}\left(\frac{1}{N_A} + \frac{1}{N_B}\right)} \qquad (1)$$

Tests For Same Mean

Next, compute the t *statistic* as

$$t = (\mu_A - \mu_B)/s_D. \tag{2}$$

Now, compute the significance of t, i.e., the probability that the statistic should be this large or larger. You can do that by looking the values up in a table, but this is not recommended because you'd have to put the table into your code.

You could also let a statistics program compute the significance for you, but that is *also* not something I'd recommended, because it's too easy to overlook factors of 2, $\sqrt{N}$, or something when you just copy numbers from one program to the next. If you write your own, (I find) that this happens much more seldom.

Significance of the t Statistic

The significance of the t statistic is given by the incomplete beta function with $d := N_A + N_B - 2$ degrees of freedom:

$$Q(t|d) = 1 - I_{d/(d+t^2)}\left(\frac{d}{2}, \frac{1}{2}\right), \tag{3}$$

where the incomplete beta function is:

$$\begin{aligned} I_x(a,b) &= B_x(a,b)/B(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}\,dt \\ B(a,b) &= \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \end{aligned}$$

which can be found in any “special function” library (usually called `ibeta()` or something like that).

The Gamma Function

The Gamma function is like the factorial function extended to complex numbers that are not negative integers (i.e., $z \neq -1, -2, \ldots$). In Eq. (6), n is a nonnegative integer.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \tag{4}$$

$$\Gamma(z+1) = z\Gamma(z), \tag{5}$$

$$\Gamma(n+1) = n!. \tag{6}$$







Things to Watch Out For

Sometimes, libraries or programs have a function called `ibeta()` that actually computes $B_x(a,b)$, not $I_x(a,b)$. This happens with the incomplete gamma function, too (see below).

In this case, you take $B_x(a,b)$ and multiply that with

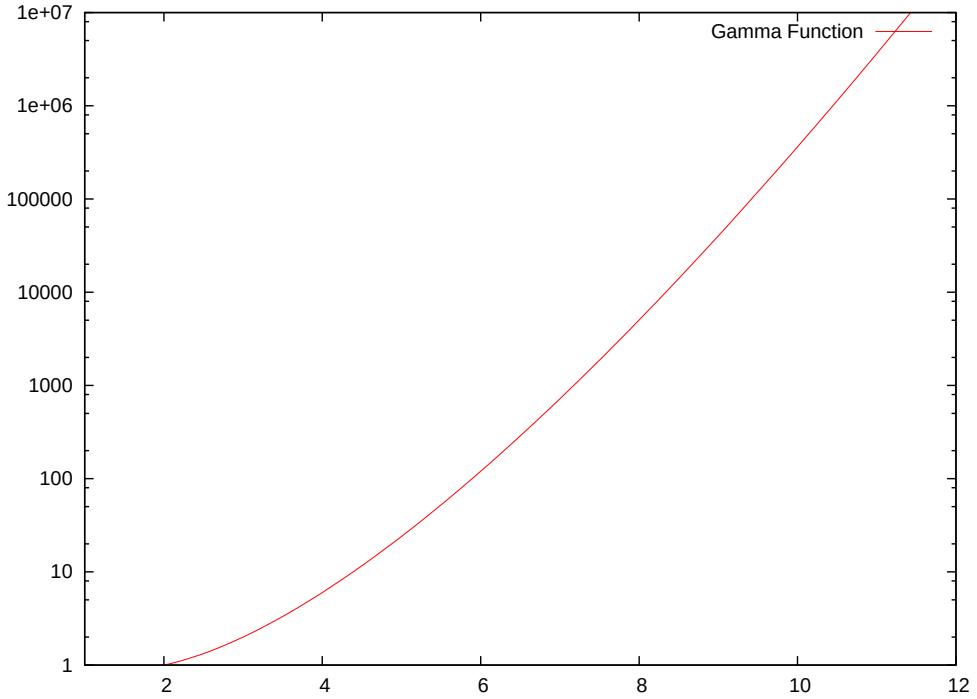
$$\exp(\ln\Gamma(a+b) - \ln\Gamma(a) - \ln\Gamma(b)).$$

Most libraries will have a routine usually called `lgamma()` that gives you $\ln\Gamma(x)$ directly, without calculating $\Gamma(x)$ first.

You do *not* multiply (or even calculate) Γ values that are modified by later operations! (This is because intermediate Γ values will overflow very quickly even if the end result is a rather small number.)

Plot of the Gamma Function

This is a plot of $\Gamma(x)$ vs. x. Note that the y axis is logarithmic!



t Test Summary

Name	Student's t Test
Question	Do two distributions have same mean?
Applicable	Two distributions have same variance.
Statistic	Student's t statistic, see Eq. (2).
Significance	See Eq. (3).

Unequal-Variance t Test

If we already know that the two distributions have unequal variances, we still might want to wish to know whether their means are different.

For example, one method of automated debugging may be totally wrong some of the time and on the spot another time, but we want to know whether it is better than our method *on the average*.

We calculate

$$t = \frac{\mu_A - \mu_B}{\sqrt{\mathrm{Var}(A)/N_A + \mathrm{Var}(B)/N_B}} \tag{7}$$

Degrees of Freedom

This statistic is distributed *approximately* as Student's t with this number of degrees of freedom:

$$d = \frac{\left(\frac{\mathrm{Var}(A)}{N_A} + \frac{\mathrm{Var}(B)}{N_B}\right)^2}{\frac{\left(\mathrm{Var}(A)/N_A\right)^2}{N_A - 1} + \frac{\left(\mathrm{Var}(B)/N_B\right)^2}{N_B - 1}} \tag{8}$$

This number is not an integer, but Eq. (3) works for noninteger values of d, too.

Unequal-Variance t Test Summary

Name	Student's t Test for unequal variances.
Question	Do two distributions have same mean?
Applicable	Two distributions have unequal variances.
Statistic	Student's t statistic, see Eqs. (7) and (8).
Significance	See Eq. (3).

Paired Samples

We have N faulty programs that we submit to their and our method in turn. The question is: Is our method or their method better on the average? We compute:

$$\mathrm{Cov}(A,B) = \left(\sum_{j=1}^{N} (x_{A_j} - \mu_A)(x_{B_j} - \mu_B) \right) \Big/ (N-1), \tag{9}$$

$$s_D = \sqrt{\frac{\mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A,B)}{N}}, \tag{10}$$

$$t = (\mu_A - \mu_B)/s_D, \qquad d = N-1. \tag{11}$$

(I'm doubtful about the formulas' numerical stability; for example, in Eq. (10), we could take the square root of a negative number, due to rounding errors in Eq. (9).)

Paired Sample t Test Summary

Name	Student's t Test for paired samples.
Question	Do two distributions have same mean?
Applicable	Same sample submitted to two methods.
Statistic	Student's t statistic, see Eqs. (9)–(11).
Significance	See Eq. (3).

F-Test for Different Variances

Some of the t tests work only if we know that the variances differ significantly or that the variances are not significantly different. The F-Test can find that out. If we let $V_A = \mathrm{Var}(A)$ and $V_B = \mathrm{Var}(B)$, the null hypothesis is that "$V_A = V_B$", and the statistic is simply

$$F = V_A / V_B \tag{12}$$

Significance for F-Statistic

Its significance, i.e., the probability that the statistic should be as large as F or larger if the null hypothesis is true, is again an incomplete beta function. First compute

$$S = 2I_{V_B/(V_B+V_AF)}\left(\frac{V_A}{2}, \frac{V_B}{2}\right) \tag{13}$$

Then the significance is

$$Q(F|V_A, V_B) = \begin{cases} S, & \text{if } S \leq 1 \\ 2 - S, & \text{if } S > 1. \end{cases} \tag{14}$$

(The second case can happen if the null hypothesis is strongly supported by the data. Eq. (13) sums up two tails of the F distribution, which can overlap.)

F-Test for Different Variances Summary

Name	F-test for significantly different variances.
Question	Do two distributions have same variance?
Applicable	Always.
Statistic	F-statistic, see Eq. (12).
Significance	See Eqs. (13)–(14).







F-Test for Greater Variances

If we want to know whether one variance is significantly greater than the other, we compute F as in Eq. (12), but use a "one-tailed" distribution:

$$Q(F|V_A, V_B) = I_{V_B/(V_B+V_AF)}\left(\frac{V_A}{2}, \frac{V_B}{2}\right) \tag{15}$$

This cannot be greater than 1.

F-Test for Greater Variances Summary

Name	F-test for significantly different variances.
Question	Do two distributions have same variance?
Applicable	Always.
Statistic	F-statistic, see Eq. (12).
Significance	See Eq. (15)







Are Two Distributions Different?

We have one sample A of which we suspect we know its distribution. Do the numbers bear this out?







Are Two Distributions Different?

We have one sample A of which we suspect we know its distribution. Do the numbers bear this out?

Or: we have two samples A and B. Do they come from the same distribution?

We'll look at the top question first and consider binned (discrete) data first.

χ^2 Test if We Know the Distribution

In this setting, we suppose that of all (independent) observations that we have made, N_i are in category (bin) i. If we know the distribution, we can compute the expected number of observations in bin i. We call that number n_i (this need not be an integer). The null hypothesis is then that "$N_i = n_i$ for all i".

We then compute the mean square deviation of N_i and n_i and call that the χ^2 statistic:

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i} \tag{16}$$

Warning: Omit any term with $n_i = N_i = 0$. On the other hand, if $n_i = 0$ but $N_i \neq 0$, make $\chi^2 = \infty$ and reject the null hypothesis.

Significance of χ^2

The probability that the statistic should be as large as χ^2 or larger if the null hypothesis is true is

$$Q(\chi^2|d) = 1 - P(d/2, \chi^2/2) \tag{17}$$

$$P(a,x) = \frac{\gamma(a,x)}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1}\, dt, \tag{18}$$

where $\gamma(a,x)$ is an incomplete gamma function. This function is often present in libraries as `igamma()`, but beware: Some libraries compute $\gamma(a,x)/\Gamma(a)$ under this name, or even $\Gamma(a,x)/\Gamma(a) = 1 - \gamma(a,x)/\Gamma(a)$.

Degrees of Freedom

If we have K categories (bins) and if $\sum_i n_i = \sum_i N_i$, then we have $d = K - 1$. Otherwise, $d = K$.

You may ask, how can we have $\sum_i n_i \neq \sum_i N_i$ if n_i is supposed to be the expected number of events in bin i? This is indeed a rare case when we determine all of the n_i beforehand and have no additional constraints on the N_i.

Normally, we renormalize the n_i so that $\sum_i n_i = \sum_i N_i$, and in that case, we have $d = K - 1$.







Caveats

The distribution of χ^2 as an incomplete gamma function has been worked out only if the individual bins are normally distributed. This is clearly not always the case.

But if either K is large of the number of observations in each bin is large, then Eq. (17) is a good approximation to the distribution of χ^2 in case of the null hypothesis.

Consequence 1: If your (non-normally distributed) sample *only just* fails the test at 5%, perhaps you should reconsider.

Consequence 2: Don't bother computing significances to an exorbitant number of significant digits. Make sure that the first few leading digits are good, though.

χ^2 Test for Known Distribution Summary

Name	χ^2 Test for Known Distribution.
Question	Does a sample have a given distribution?
Applicable	Binned data.
Statistic	χ^2-statistic, see Eq. (16).
Significance	See Eq. (17)

χ^2 *for Two Samples*

We have samples A consisting of K bins and a sample B, consisting of the same K bins. Do A and B have the same distribution? To find out, compute

$$\chi^2 = \sum_i \frac{(A_i - B_i)^2}{A_i + B_i} \tag{19}$$

If the data was collected in such a way that necessarily $\sum A_i = \sum B_i$, use $d = K - 1$, as usual. If that's not the case, use $d = K$ and

$$\chi^2 = \sum_i \frac{\left(\sqrt{B/A}\, A_i - \sqrt{A/B}\, B_i\right)^2}{A_i + B_i} \qquad A = \sum_i A_i, \quad B = \sum_i B_i \tag{20}$$

Examples for Degrees of Freedom

Example 1: We submit the same 100 programs to our and their bug finder. If the bins are “pinpoints the bug” and “does not pinpoint the bug”, we have $K = 2$ and therefore $d = 1$.

Example 2: We take our test programs and put them through our bug finder. Then we take their test programs and put them through their bug finder. In this case we have $d = 2$.

Examples for Degrees of Freedom

Example 1: We submit the same 100 programs to our and their bug finder. If the bins are “pinpoints the bug” and “does not pinpoint the bug”, we have $K = 2$ and therefore $d = 1$.

Example 2: We take our test programs and put them through our bug finder. Then we take their test programs and put them through their bug finder. In this case we have $d = 2$.

But then we test only if our bug finder is better with our programs than their bug finder is with theirs

Examples for Degrees of Freedom

Example 1: We submit the same 100 programs to our and their bug finder. If the bins are “pinpoints the bug” and “does not pinpoint the bug”, we have $K = 2$ and therefore $d = 1$.

Example 2: We take our test programs and put them through our bug finder. Then we take their test programs and put them through their bug finder. In this case we have $d = 2$.

But then we test only if our bug finder is better with our programs than their bug finder is with theirs, which we probably believe anyway!

χ^2 Test for Two Samples Summary

Name	χ^2 Test for Two Samples.
Question	Do two samples have the same distribution?
Applicable	Binned data.
Statistic	χ^2-statistic, see Eqs. (19) and (20).
Significance	See Eq. (17).







Continuous Data: KS Test

So far we have worked with data in bins and have used the expected and observed numbers in each bin. In other words, so far we have used the PDF.







Continuous Data: KS Test

So far we have worked with data in bins and have used the expected and observed numbers in each bin. In other words, so far we have used the PDF.

The KS test works on the CDF, since for a continuous distribution, $\Pr(X = x)$ is always zero.

Example: I draw a real number between 0 and 1 at random. The probability to hit any given number is 0, even though I actually do hit one every time. This is not a contradiction.







Empirical CDF

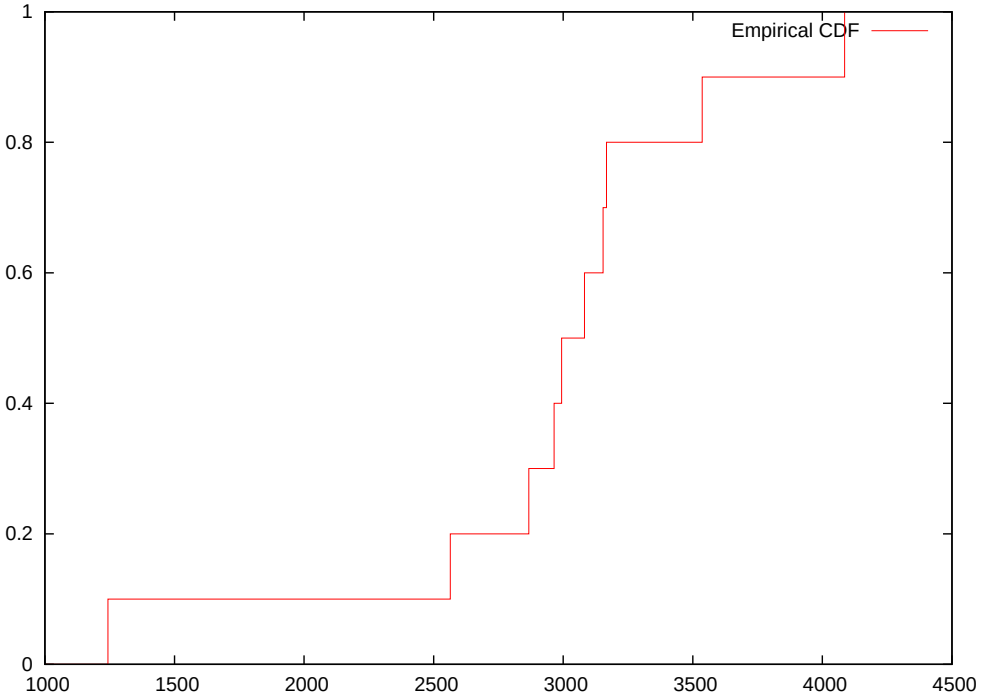
If we have made N samples and sort them into increasing order such that $x_i \leq x_j$ for $1 \leq i < j \leq N$, then the empirical CDF is

$$\mathrm{cdf}(x) = j/N, \qquad \text{if } x_j \leq x < x_{j+1}. \tag{21}$$

Here we assume the existence of two sentinel values $x_0 = -\infty$ and $x_{N+1} = +\infty$. We also assume w.l.o.g. that all samples are different.

Empirical CDF Example

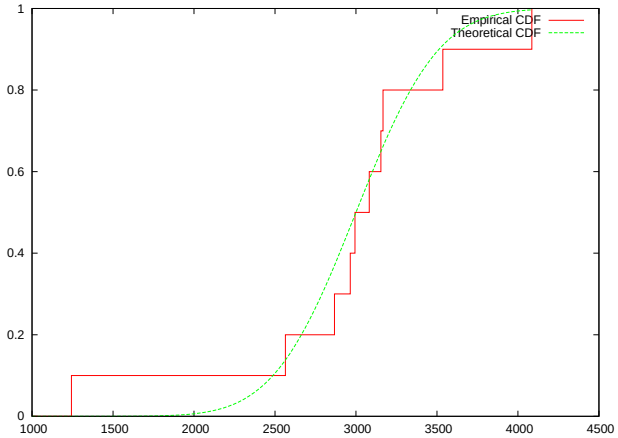
Here are our lightbulbs again:



Theoretical CDF

We suspect that the lightbulb lifetime might be distributed normally with mean $\mu = 3000$ and standard deviation $\sigma = 400$. So the conjectured theoretical CDF would be

$$\Pr(X < x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right)\right) \tag{22}$$



KS Statistics

Now we compute the maximum deviations from above and below between the empirical and the conjectured CDFs:

$$D_N^+ = \max_{1 \le j \le N} \left(S_N(x_j) - \Pr(X < x_j) \right) \quad (23)$$

$$D_N^- = \max_{1 \le j \le N} \left(\Pr(X < x_j) - S_N(x_j) \right) \quad (24)$$

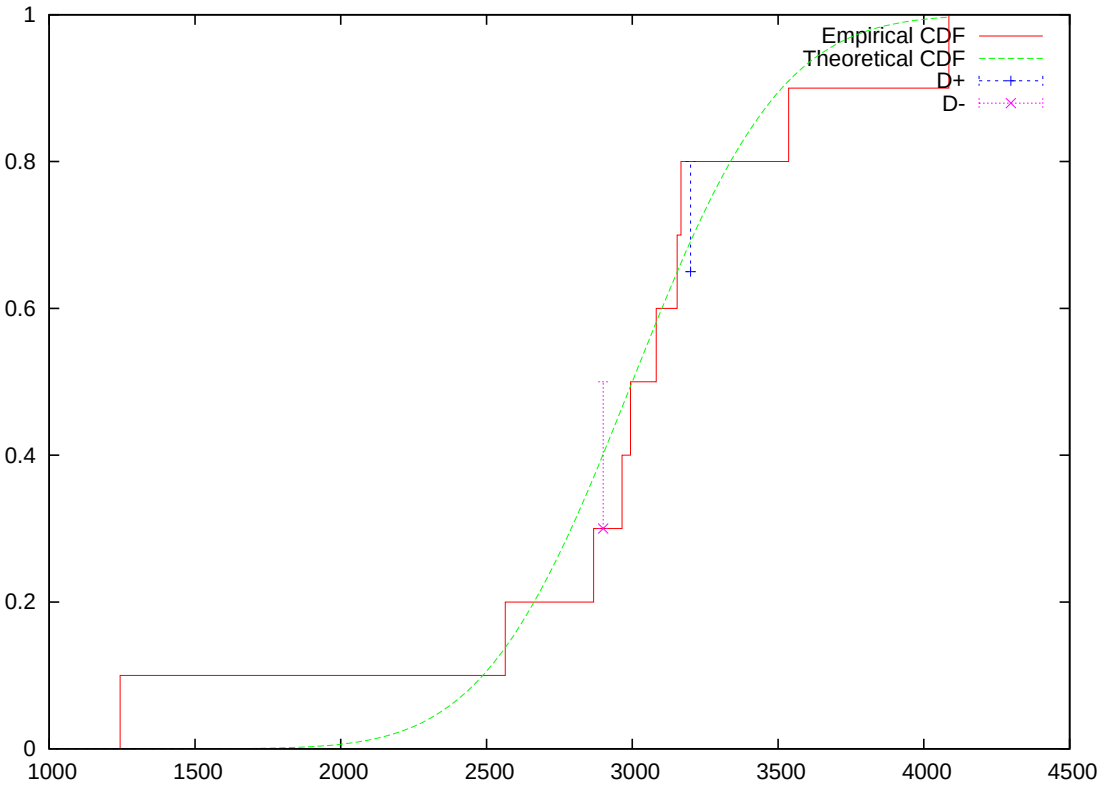
$$D_N = \max_{1 \le j \le N} \left| S_N(x_j) - \Pr(X < x_j) \right| \quad (25)$$

$$= \max\{D_N^+, D_N^-\} \quad (26)$$

Warning: Some authors (notably Knuth) define $K_N^+ = \sqrt{N} D_N^+$ etc.

Note: The empirical CDF cannot be everywhere below (or above) the conjectured CDF, because $\lim_{x \to -\infty} \Pr(X < x) = 0$ and $\lim_{x \to \infty} \Pr(X < x) = 1$.

KS Statistics



Significance of the KS Statistic

For small N (and also for D_N^-), use

$$\Pr(D_N^+ > x/N) = \frac{x}{N^N} \sum_{x<k\le N} \binom{N}{k} (k-x)^k (x+N-k)^{N-k-1}.$$

Significance of the KS Statistic

For small N (and also for D_N^-), use

$$\Pr(D_N^+ > x/N) = \frac{x}{N^N} \sum_{x<k\le N} \binom{N}{k} (k-x)^k (x+N-k)^{N-k-1}. \quad (27)$$

But don't choose a small N! Instead, choose a large N and use

$$\Pr(D_N^+ > x) \approx e^{-2Nx^2}, \qquad x \ge 0 \quad (28)$$

$$\Pr(D > x) \approx Q\left((\sqrt{N} + 0.12 + 0.11/\sqrt{N}) \cdot x\right) \quad (29)$$

$$Q(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}. \quad (30)$$

A Word of Caution

Eq. (27) is not a computationally effective formula. The value for x/N^N will get extremely small, and since the entire value is going to be a probability and hence between 0 and 1, we know that the sum must be extremely large in order to compensate for this.

So my recommendation is again not to use small values of N.

How large should N be? Some authors tell you to choose $N > 4$, but I wouldn't trust any test that has N as small as 5, even if Eq. (28) is a good approximation to Eq. (27) for $N = 5$.

Knuth says to use $N > 30$, which seems much better advice. In my own tests, I have always used $N > 1000$, which went very well, both numerically and statistically.

And the Lightbulbs...?

In our example, $D_N^+ \approx 0.15$ and $D_N^- \approx 0.2$. Using the formulas for large N, we get

$$\begin{aligned} \Pr(D_N^+ > 0.15) &\approx e^{-2 \cdot 10 \cdot 0.15^2} = e^{-0.45} = 0.64 \\ \Pr(D_N^- > 0.2) &\approx e^{-2 \cdot 10 \cdot 0.2^2} = e^{-0.8} = 0.45 \end{aligned}$$

Since both values are larger than 0.05, we cannot reject the null hypothesis that the lightbulb lifetime is indeed normally distributed with mean 3000 and standard deviation 400, neither at the 1% nor at the 5% levels.

KS Test for Known Distribution Summary

Name	Kolmogorov-Smirnov (KS) Test
Question	Does a sample have a given distribution?
Applicable	Continuous data.
Statistic	KS statistics, see Eqs. (23)–(25).
Significance	See Eqs. (27)–(29).







KS Test for Two Samples

We have two samples A with the empirical CDF S_A and B with the empirical CDF S_B, consisting of continuous data. The KS statistic for two samples is analogous to Eq. (25):

$$D_N = \max \left| S_A(x) - S_B(x) \right| \tag{31}$$

Note that the one-sided statistics of Eqs. (23)–(24) don't make sense here. The significance is calculated as in Eq. (29), only that N is replaced by

$$\frac{N_A N_B}{N_A + N_B} \tag{32}$$

KS Test for Two Samples Summary

Name	KS Test for Two Samples.
Question	Do two samples have the same distribution?
Applicable	Continuous data.
Statistic	KS-statistic, see Eq. (31).
Significance	See Eq. (32).

Libraries and References

The Linux standard math library has all of these special functions already built in.

If Java doesn't have them, I have written well-tested routines in C that should be easy to port to Java. These routines have withstood a decade or so of abuse by people like me.

One of the inspirations for these routines was *Numerical Recipes in C* by Press et.al. This book has concise introductions to the subject, but beware! It also has a few bugs!

No bugs can be found in D.E. Knuth, *The Art Of Computer Programming, Vol 2: Seminumerical Algorithms*, Section 3.3.

More on tests and and testing can be found in my ancient 1993 paper *Statistical Properties of IDEA session keys in PGP*, http://www.artdecode.de/download/pgprtest-long.ps