Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

# Lecture 21: Significance Tests, II

**Relevant textbook passages:**

**Larsen–Marx [7]:** Chapter 7, pp. 440–442 in Chapter 8, and Sections 9.1, 9.2, 9.3.

## 21.1 Off-the-shelf modeling

One of the strengths of the classical likelihood-based parametric approach to significance testing is that a number of special cases have been thoroughly analyzed, and there are convenient off-the-shelf solutions to analyzing and testing the data. There are so many of these that I can't possibly discuss them all in this class. I shall give you a few of the most basic tests, those that everyone expects to be covered in an intro stats course. The point is to make sure you know they exist.

When you get to your lab and you have real data to analyze, I recommend consulting a book such as Calvin Dytham's [5] *Choosing and Using Statistics*. It discusses many off-the-shelf models and their subtle points. Better, yet it describes code for a number of different programs and languages. When you are reading the results of someone else's research, they may describe some arcane test or procedure that I haven't covered, or even heard of. In cases like this, Wikipedia is often incredibly useful. It is probably possible to teach a good introductory stats class using Wikipedia as the textbook. Even this approach will probably soon be obsolete as AI (artificially intelligent) statisticians become commonplace. Mary Kennedy told me about a program her lab uses that queries you about your data, then decides on a testing procedure, and analyzes the data for you. In a world where this is common, what is the value of this course? Well, remember the first attempt to use R's numerical optimization function to compute the MLE of $p$ for the coin tossing? It gave 99.9%, not 49.9%. Remember, with any software, or any reference work, "Trust, but verify." In this course, I hope you learn enough to read the manual for your software to have some idea of what the program is doing, and to be able to decide if it makes sense.

The ready availability of off-the-shelf models is also a huge weakness. It tempts you to treat your data as if it fits one of these off-the shelf models even if it doesn't. This problem is rampant in my discipline, economics, and I'm sure in others as well. In the Lecture 23, we'll take up specification testing, which is a step in the right direction. If you have a case where the usual methods seem inappropriate, the notions from likelihood ratio testing can help point you in the right direction. Plus many universities have departments of applied statistics where really smart tooled-up statisticians are always on the lookout for new cases to add to the shelf.

## 21.2 Testing hypotheses based on normality when $\sigma$ is known

I believe it is fair to say that a vast majority of users of hypothesis tests use tests that are based on the assumption that the "error terms" in their data are normally distributed. Indeed, the Central Limit Theorem says that if the errors are the sum of many small independent errors, then the normality assumption is justified. Leo Breiman [4, p. 10] describes this argument as "only a cut above a hopeful appeal to the ghost of Laplace." Nevertheless we shall start with a discussion of tests based on normality, since you will undoubtedly employ them at some point in the analysis of your laboratory data.

In this section we shall make the unreasonable assumption that we are dealing with random variables with a known standard deviation $\sigma$, but an unknown mean $\mu$. This case is easier to understand than the case where $\sigma$ is not known, and it serves as the basis for understanding the more realistic case.

**Larsen–Marx [7]:** Chapter 7

For a sample $X_1, \dots, X_n$ of independent and identically distributed Normal $N(\mu, \sigma^2)$ random variables, setting

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

we have $\bar{X} \sim N(\mu, \sigma^2/n)$, so

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{1}$$

We saw in Lecture 19 that even though we can't observe $Z$ (since $\mu$ is unknown), if we know $\sigma$, then we can use the observed value $\bar{x}$ of $\bar{X}$ to get a confidence interval for $\mu$, the

$$1-\alpha \text{ confidence interval for } \mu \text{ is } \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{2}$$

where $z_{\alpha/2}$ defined by

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2.$$

(See Section 19.3.) The width of the confidence interval depends on the sample size, so we can use (2) to choose the sample size to fix the width of the confidence interval. The width $w$ of the interval is $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ so to get an interval of width $w$ requires

$$n = \frac{4z_{\alpha/2}^2\sigma^2}{w^2}.$$

Equation (2) can also serve as a basis for testing hypotheses about $\mu$.

There are two common classes of null hypotheses and alternative hypotheses regarding $\mu$:

- A **two-sided** hypothesis/test deals with a hypothesis of the form $\mu = \mu_0$, where $\mu_0$ is some fixed value that you wish to test. The alternative is that $\mu \neq \mu_0$. This is called a two-sided alternative because it allows for $\mu < \mu_0$ and also for $\mu > \mu_0$.

- A **one-sided** hypothesis/test deals with a hypothesis of the form $\mu = \mu_0$ and the alternative is that $\mu \geqslant \mu_0$. Or it could be that the alternative is $\mu \leqslant \mu_0$. The point is that you care about only one way that $\mu$ might differ from $\mu_0$.

Why might you care only about one-sided alternatives? Frequently you want to find out if some treatment has a beneficial effect. For instance, $\mu_0$ might be the death rate due to some disease using the standard treatment. You have a new therapy that you hope works better, but costs more. So you care if the death rate is lower than the standard treatment, but not if the death rate is higher, since you do not plan on using the treatment unless the death rate is lower.

On the other hand if your new treatment is cheaper, then you may want to use it unless the death rate is higher, so the other one-sided test may be of interest. Statistics cannot tell you which hypothesis you ought to test, only how to test your hypothesis.

There is another kind of null hypothesis you might want to test in the one-sided case: Namely instead of the jull hypothesis $\mu = \mu_0$ with the alternative $\mu > \mu_0$, the null might be that $\mu \leqslant \mu_0$ with the alternative $\mu > \mu_0$. As long as you have a case with a monotone likelihood ratio (Section 20.9⋆) the likelihood ratio test will be the same in either case.

In what follows I will consider mostly one-sided tests.

### 21.2.1 One-sided alternatives

When the null hypothesis is $\mu \leqslant \mu_0$ and the alternative hypothesis $\mu \geqslant 0$, the likelihood ratio test takes the form: Reject $H_0$ if $\bar{x} > c$ for some appropriate cutoff $c$. To get a significance level of $\alpha$ choose $c$ so that $P_{\mu_0}\left(\bar{X} > c\right) = \alpha$. (Recall Section 19.3.)

Now

$$P_{\mu_0}\left(\bar{X} > c\right) = P_{\mu_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right),$$

and since $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, we need

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

or

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$$

and we

$$\text{Reject } H_0 \text{ if } \bar{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha.$$

## 21.3 Power and sample size

The power of this test is the probability that we reject $H_0$ when the mean is $\mu > \mu_0$, which depends on the value of $\mu$. The graph of the power as a function of $\mu$ is called the the power curve of the test. This probability is

**Larsen–Marx [7]:** pp. 372–374

$$P_\mu\left(\bar{X} > c\right) = P_\mu\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

In order for the power to be equal to $\gamma$ we must have

$$z_\gamma = \frac{c - \mu}{\sigma/\sqrt{n}} \quad \text{or in other words} \quad c = \mu + \frac{\sigma}{\sqrt{n}} z_\gamma.$$

This entails

$$\mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha = c = \mu + \frac{\sigma}{\sqrt{n}} z_\gamma$$

or

$$n = \sigma^2 \left(\frac{z_\alpha - z_\gamma}{\mu - \mu_0}\right)^2.$$

This tells us how large the sample has to be to get the power to be equal to $\gamma$ for a test with significance level $\alpha$. Notice that for $\mu > \mu_0$ (the case of interest) we need to have $z_\alpha > z_\gamma$, which requires $\alpha < \gamma$. This makes sense. The probability of rejecting the null hypothesis when it is true is $\alpha$ and $\gamma$ is the probability of rejecting it when it is false. We want the probability of rejecting $H_0$ when it is false to greater than when it is true.

**21.3.1 Example** Larsen–Marx [7, Example 6.4.1, pp. 373–374] ask fro the sample size needed to achieve a power of $\gamma = 0.6$, for a $\alpha = 0.05$ level test when $\sigma = 14$, and $\mu - \mu_0 = 3$. In this case, $z_\alpha = 1.96$ and $z_\gamma = -0.25$, so

$$n = \left(14\frac{2.21}{3}\right)^2 = 78.$$

□

**21.3.2 Example** Here is a numerical example for the case $\mu_0 = 0$, $\mu = 0.1$, $\sigma = 1$, $\alpha = 0.025$, and $\gamma = 0.975$. In this case, $z_\alpha = 1.96$ and $z_\gamma = -1.96$, $\mu - \mu_0 = 0.1$, so

$$n = \left(\frac{3.92}{.1}\right)^2 = 1536.64,$$

so a sample size of 1537 is needed to get a power of 0.975 at $\mu = 0.1$. □

## 21.4 ⋆ Detectability thresholds

Leo Breiman [4, Chapter 5] discusses the notion of **detectability**. In the context of a hypothesis test with null hypothesis $H_0 : \theta \in \Theta_0$ with significance level $\alpha$, we say that the parameter value $\theta \notin \Theta_0$ can be **detected at level $\alpha$** by the test if

$$P_\theta\,(\text{accepting } H_0) \leqslant \alpha$$

or

$$\text{Power}(\theta) \geqslant 1 - \alpha.$$

We say that $\Delta$ is the **detectability threshold** for the test if

$$\text{distance}(\theta, \Theta_0) \geqslant \Delta \implies \text{Power}(\theta) \geqslant 1 - \alpha.$$

That is, the detectability is the minimum distance the parameter has to be from the null hypothesis in order for the probability of a Type II error to be no greater than the probability of a Type 1 error.

So following the analysis of the previous section, for a one-sided text of significance $\alpha$ of the hypothesis $\mu = \mu_0$ versus $\mu > \mu_0$, the detectability threshold $\Delta\mu = \mu - \mu_0$ satisfies

$$\frac{\Delta\mu}{\sigma} = \frac{z_{1-\alpha} - z_\alpha}{\sqrt{n}} = \frac{2z_{1-\alpha}}{\sqrt{n}}.$$

For a two-sided test with significance level $\alpha$, we have

$$\frac{\Delta\mu}{\sigma} = \frac{z_{1-\alpha/2} - z_{\alpha/2}}{\sqrt{n}} = \frac{2z_{1-\alpha/2}}{\sqrt{n}}.$$

We can use these to figure out the sample size need for a given detectability threshold. For instance, for a two-sided test at level 0.05, we have $z_{1-\alpha/2} = z_{0.975} = 1.96$, so

$$\frac{\Delta\mu}{\sigma} = \frac{3.9}{\sqrt{n}}.$$

## 21.5 What if $\sigma$ is unknown?

The problem with the analysis above is that we seldom know $\sigma$. In Lecture 18, we derived the Maximum Likelihood Estimators for $\mu$ and $\sigma^2$ as

$$\hat{\mu}_{\text{MLE}} = \bar{x} \qquad \text{and} \qquad \hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n},$$

where $\bar{x} = \sum_{i=1}^n x_i/n$. We also showed that $\hat{\sigma}^2_{\text{MLE}}$ is biased, so the unbiased estimator $S^2$ is often used instead:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

The question is, what is the distribution of

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}?$$

It turns out it is *not* a standard Normal random variable. In order to describe the distribution of this statistic, we first examine some related distributions.

## 21.6 The chi-square distribution

Recall that the **chi-square($m$)** or **$\chi^2$-distribution with $m$ degrees of freedom** is the distribution of the sum $Z_1^2 + \cdots + Z_m^2$ of squares of $m$ independent standard normal random variables [7, Theorem 7.3.1, p. 389]. It is also a Gamma$(\frac{m}{2}, \frac{1}{2})$ distribution. See Figure 21.1 for the shape of the density.

**21.6.1 Fact** *[7, Theorem 7.3.2, p. 390] If $X_1, \ldots, X_n$ are independent and identically distributed Normal $N(\mu, \sigma^2)$ random variables, then*

1. *$\bar{X}$ and $S^2$ are independent.*
2. *$\bar{X} \sim N(\mu, \sigma^2/n)$.*
3. *$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 \sim$ chi-square$(n-1)$*

(We'll return to this in a later lecture on chi-square tests.)

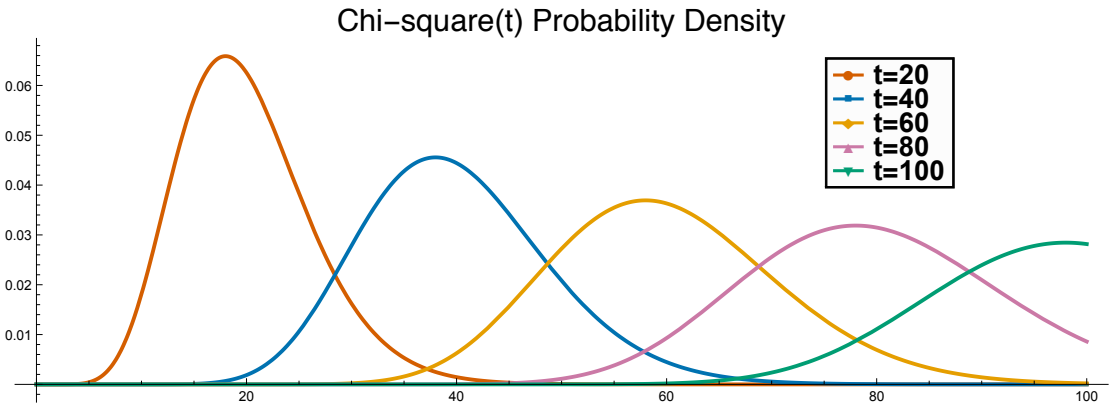


Figure 21.1. Chi-square pdfs.

## 21.7 The $F$-distribution

Let $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$ be independent. Then the random variable

$$\frac{V/m}{U/n}$$

has an **$F_{m,n}$-distribution with $m$ and $n$ degrees of freedom**. The $F$ distribution is also known as the **Snedecor $F$ distribution**, although Larsen and Marx assert that the $F$ is for Fisher.

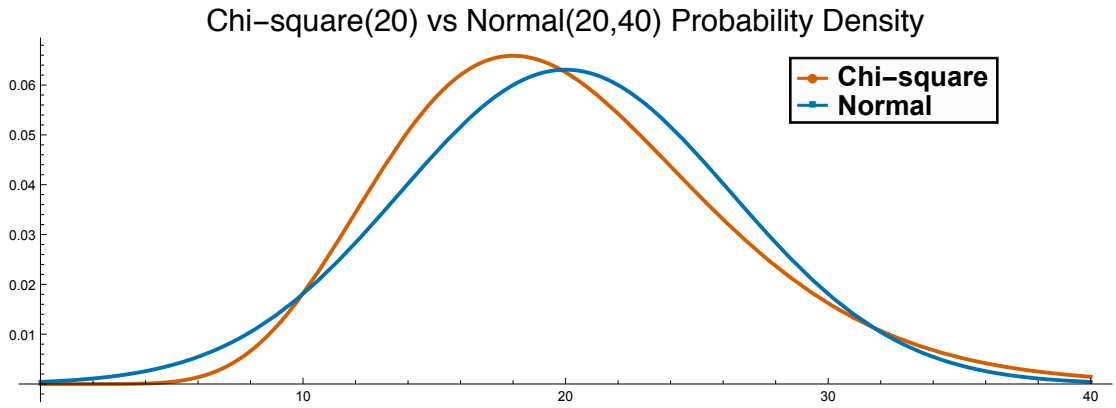


Figure 21.2. Chi-square vs Normal.

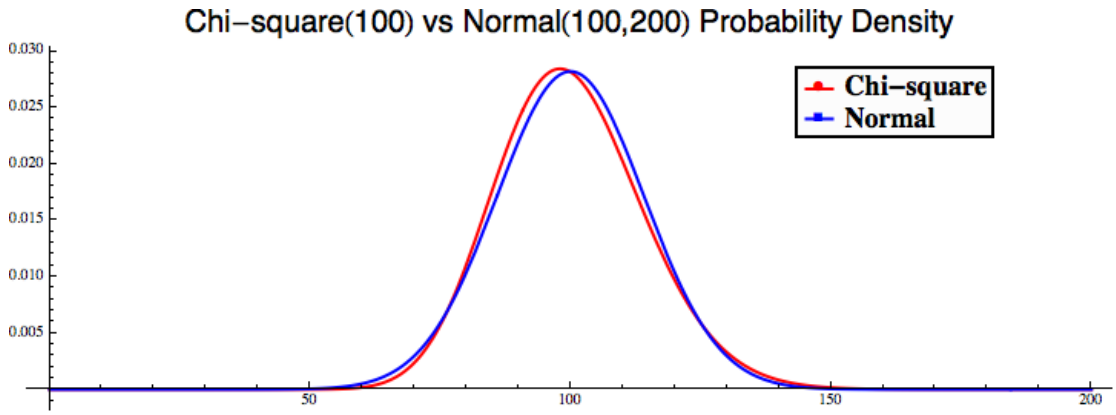


Figure 21.3. Chi-square vs Normal.

The $F_{m,n}$ density is given by [7, Theorem 7.3.3, p. 390]

$$f(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\, m^{m/2} n^{n/2}\, \frac{x^{(m/2)-1}}{(n+mx)^{(m+n)/2}} \qquad (x \geqslant 0).$$

See Figure 21.4.

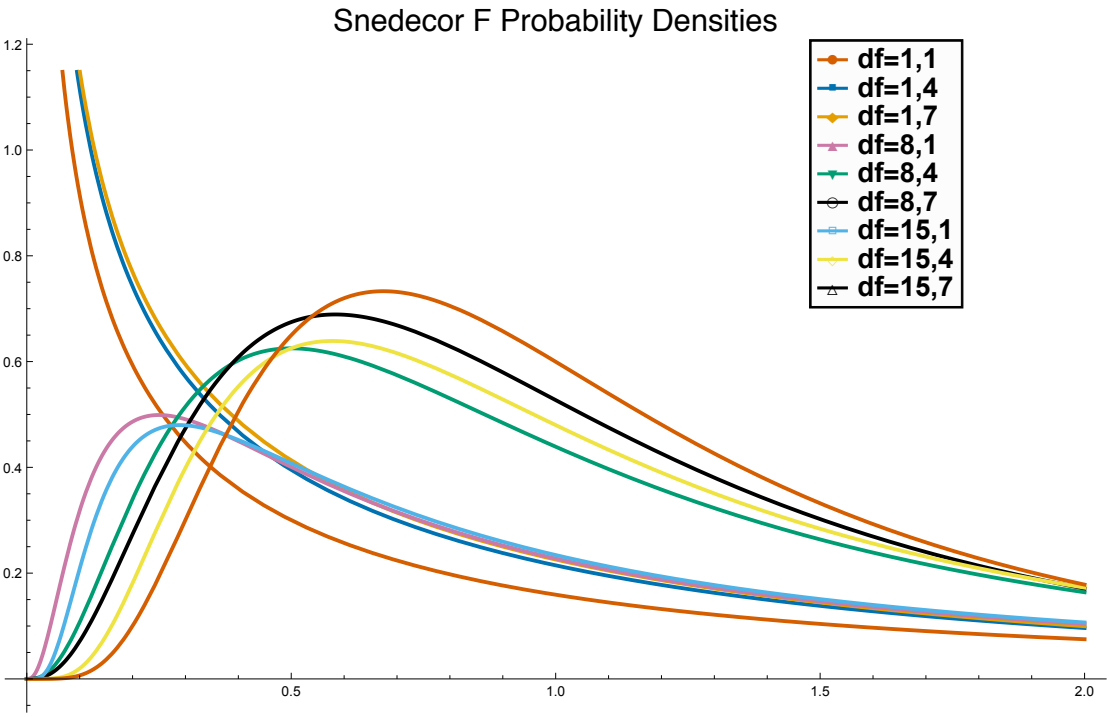


Figure 21.4. Snedecor $F$ pdfs.

## 21.8 The Student $t$-distribution

Let $Z \sim N(0,1)$ and $U \sim \chi^2(n)$ be independent, then the random variable

$$T_n = \frac{Z}{\sqrt{\frac{U}{n}}}$$

has the **Student $t$-distribution with $n$ degrees of freedom**.

(The $t$-distribution was first calculated by William Sealy Gossett in 1908, who, because he was violating a nondisclosure agreement with his employer, the Guinness Brewery (makers of Guinness Stout and the original publisher of the *Guinness Book of Records*), published it under the pseudonym Student [12]. See Larsen–Marx [7, pp. 386–387].)

The density is given by [7, Theorem 7.3.4, p. 390]

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{2}\right)^{-(n+1)/2} \qquad (x \in \boldsymbol{R}).$$

Note that this is symmetric about zero. See Figure 21.5.

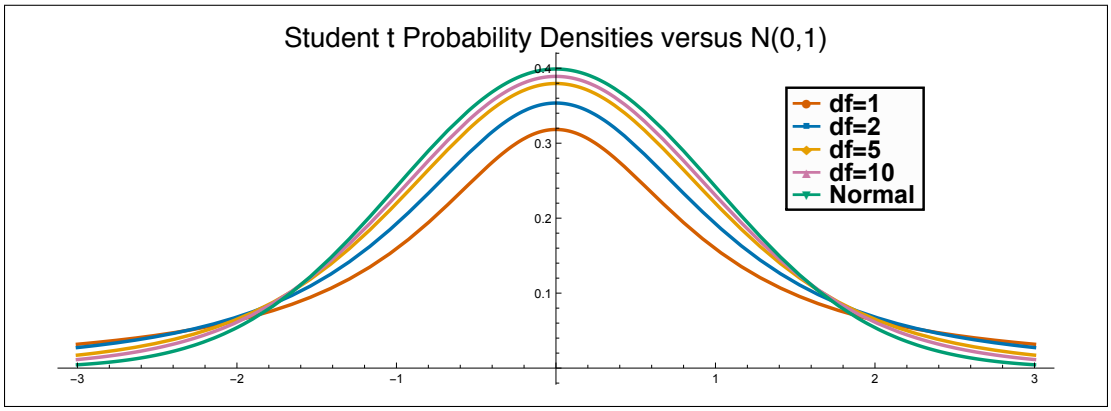


Figure 21.5. Student $t$-pdfs and the Standard Normal. The $t$ densities are more spread out.

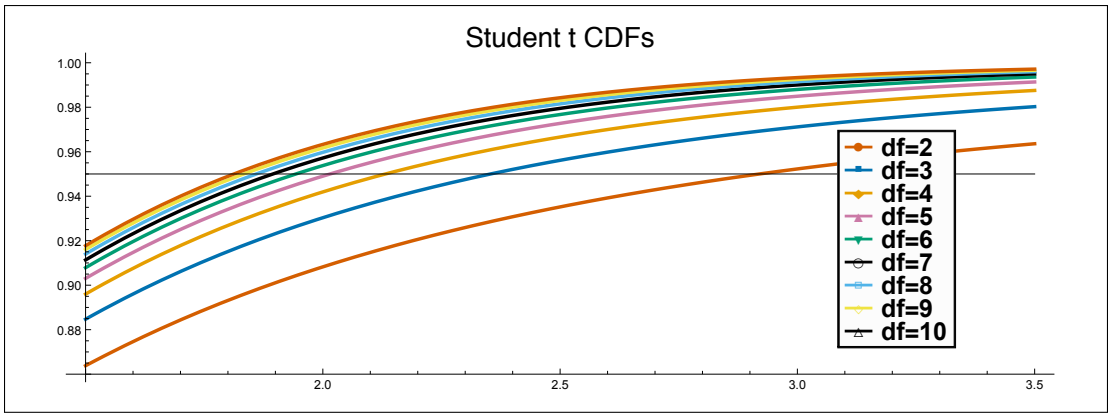


Figure 21.6. The tail of the CDF for various Student $t$-distributions.

## 21.9 A test statistic for the mean with estimated $\sigma$

**21.9.1 Theorem** *[7, Theorem 7.3.5, p. 393] For a sample $X_1, \dots, X_n$ of independent and identically distributed Normal $N(\mu, \sigma^2)$ random variables, the test statistic*

$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

*has a Student t-distribution with $n-1$ degrees of freedom.*

## 21.10 Confidence interval for $\mu$

**21.10.1 Definition ($t$-distribution cutoffs)** *Larsen–Marx [7, p. 395] define $t_{\alpha,n}$ by*

$$P\left(T_n \geqslant t_{\alpha,n}\right) = \alpha,$$

*where $T_n$ has the Student t-distribution with $n$ degrees of freedom.*

Then

$$P\left( -t_{\alpha/2,n-1} \leqslant \frac{\bar{X} - \mu}{S/\sqrt{n}} \leqslant t_{\alpha/2,n-1} \right) = 1 - \alpha$$

or equivalently

$$P\left(\bar{X} - t_{\alpha/2,n-1} S/\sqrt{n} \leqslant \mu \leqslant \bar{X} + t_{\alpha/2,n-1} S/\sqrt{n}\right) = 1 - \alpha.$$

In other words,

given the sample values $x_1, \dots, x_n$ from $n$ independent and identically distributed draws from a normal distribution, a $1-\alpha$ confidence interval for $\mu$ is the interval

$$\left(\bar{x} - t_{\alpha/2,n-1} s/\sqrt{n},\, \bar{x} + t_{\alpha/2,n-1} s/\sqrt{n}\right).$$

Figure 21.7 shows the result of using this procedure 100 times to construct a symmetric 95% confidence interval for $\mu$, based on (pseudo-)random samples of size 5 drawn from a standard normal distribution. Note that in this instance, 5 of the 100 intervals missed the true mean 0.

Compare this figure to Figure 19.1, where it was assumed that the variance was known. In that case, all the confidence intervals had the same width. When the variance is estimated from the sample, this is no longer the case. Also recall that the sample mean and he sample variance are independent, so a short confidence interval is not necessarily a "better" confidence interval.

## 21.11 $t$-quantiles versus $z$-quantiles

The value $z_{0.025} = 1.96$, which is used to construct a 95% confidence interval is based on knowing the standard deviation $\sigma$, can be very misleading for small sample sizes, when $\sigma$ is estimated by the unbiased version of the MLE estimate. The following table gives $t_{.025,n}$ for various values of $n$. This also shows how the critical value of a test changes with the number of degrees of freedom.

| df | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|
| $t_{.025,\mathrm{df}}$ | 12.71 | 4.3 | 2.78 | 2.31 | 2.12 | 2.04 | 2.0 | 1.98 | 1.97 | 1.96 |

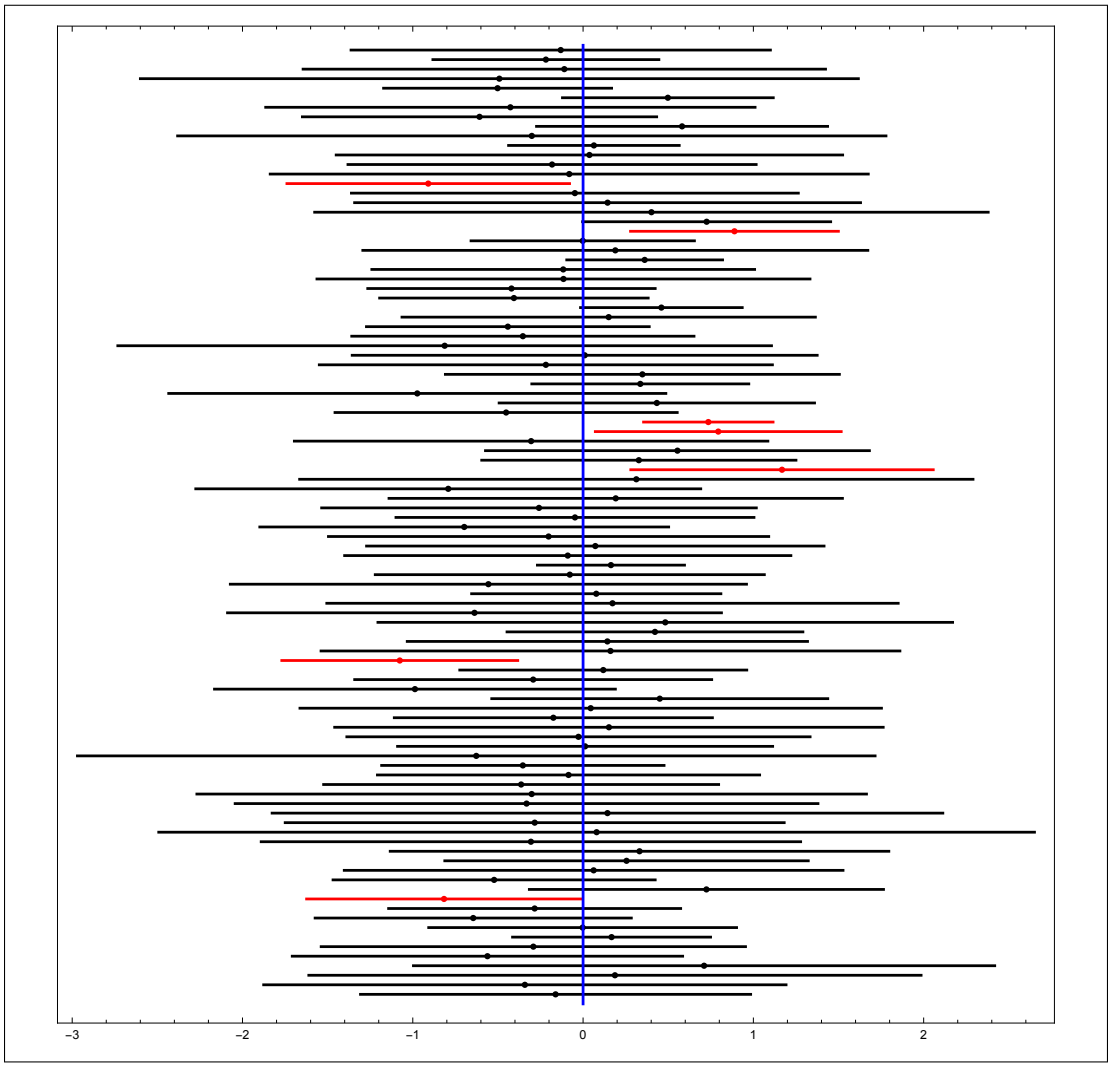


Figure 21.7. Here are one hundred 95% confidence intervals for the mean from a Monte Carlo simulation of a sample of size 5 independent standard normals. The confidence interval is based on the estimated standard deviation, so not all intervals are the same length. The intervals that do not include 0 are shown in red.

## 21.12 The "*t*-test"

This also forms the basis for a hypothesis test, called a ***t*-test**.

To test the Null Hypothesis

$$H_0 \colon \mu = \mu_0$$

versus the one-sided alternative

$$H_1 \colon \mu > \mu_0$$

at the $\alpha$ significance level, compute the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Reject $H_0$ if $t > t_{\alpha,n-1}$.

See [7, Theorem 7.4.2, p. 401] for the related two-sided or the other one-sided test.

With modern software, performing "$t$-tests" is trivial. In Mathematica, you find the $p$-value of $t$ with `CDF[StudentTDistribution[n], t]`, where $n$ is the degrees of freedom. Or simpler yet, if your sample is the array `data`, the command `TTest[data, m]` returns the $p$-value of $t$ under the null hypothesis $\mu = m$, against the two-sided alternative $\mu \neq m$. See the documentation for more options. In R, if your sample is in the array `data`, the command `t.test(data, mu=`$\mu_0$`)` returns a detailed report on the two-sided test of the hypothesis $\mu = 0$ including a confidence interval for $\mu$. (To test the hypothesis $\mu = m$, use: `t.test(data-m, mu=m)`.

By the way, *Choosing and Using Statistics: A Biologist's Guide* by Calvin Dytham [5] has excellent sample code for a number of programs including R, SPSS, Minitab, and even Excel, but not Mathematica.

## 21.13 ⋆ On the power of the *t*-test

It is not straightforward to compute the power of the $t$ test. We start with a sample of size $n$ of independent random variables $X_i$, distributed as Normal$(0, \sigma^2)$, where $\sigma$ is unknown. We have the null hypothesis $H_0 : \mu = \mu_0$ with the alternative $H_1 : \mu > \mu_0$. The test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ is computed. Under the null hypothesis, the test statistic has a $t$ distribution with $n - 1$ degrees of freedom. The null hypothesis is rejected if $t > t_\alpha$ for test with significance level $\alpha$. We want to compute the power at $\mu$, which is just

$$P_{\mu,\sigma}\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t_\alpha\right).$$

The problem is that if each $X_i \sim N(0, \sigma^2)$, then the test statistic is not distributed according to a $t$ distribution. We need to transform the problem into something we can cope with.

To that end, let's follow Ferris, Grubbs, and Weaver [6] and recast the problem like this. Let

$$\rho = \frac{\mu - \mu_0}{\sigma}.$$

Then we can rewrite the null hypothesis as $H_0 : \rho = 0$ versus the alternative $H_1 : \rho > 0$.

Note that for any constant $c$,

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > c \iff \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > \frac{s}{\sigma}c \iff \frac{\sqrt{n}\big(\bar{X} - \mu - (\mu_0 - \mu)\big)}{\sigma} > \frac{s}{\sigma}c$$

$$\iff \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{s}{\sigma}c - \sqrt{n}\rho.$$

Now when $\mu$ is the mean, the quantity $\sqrt{n}(\bar{X}-\mu)/\sigma$ is a Standard Normal random variable, so

$$P_{\mu,\sigma}\left(\frac{\bar{X}-\mu_0}{s/\sqrt{n}} > c\right) = P_{\mu,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > \frac{s}{\sigma}c - \sqrt{n}\rho\right) = \Phi\left(\frac{s}{\sigma}c - \sqrt{n}\rho\right).$$

The problem is that $s$ is a random variable, so this gives the probability conditional on the value of $s$. But the argument of $\Phi$ depends only on the value $s^2/\sigma^2$, which we know has a $\chi^2$ distribution. One can compute the expected value to get the power of the test at $\mu$. Ferris, et. al. do this and report the operating characteristic (1 minus the power) graphically. Their graph is reproduced in Breiman [4, p. 147].

Breiman recommends using the same rule for detectability thresholds with unknown $\sigma$ as for known $\sigma$ for the $t$-test with moderately large sample sizes.

$$\frac{\Delta\mu}{\sigma} = \frac{z_{1-\alpha/2} - z_{\alpha/2}}{\sqrt{n}} = \frac{2z_{1-\alpha/2}}{\sqrt{n}}.$$

Note that since this depends on the unknown standard deviation $\sigma$, we cannot use this to compute a sample size ex ante. We first have to estimate $\sigma$, and then use that as a guide to deciding whether to collect a larger sample to increase the poser of the test.

## 21.14 Model equations

The structure of an experiment and the nature of the data it generates is often captured by its **model equations**. The simplest way to explain them is via examples.

## 21.15 Difference of means, same variances

**Larsen–Marx [7]:** Section 9.2

Given $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ normal with same variance, but possibly different means, there is a $t$-test for the null hypothesis $\mu_X = \mu_Y$. See Section 9.2 of Larsen and Marx [7].

The model equations are

$$X_i = \mu_X + \varepsilon_{Xi},\ i = 1, \ldots, m \qquad Y_j = \mu_Y + \varepsilon_{Yj},\ j = 1, \ldots, m,$$

where $\varepsilon_X$ and $\varepsilon_Y$ are independent and identically distributed Normal$(0, \sigma^2)$.

The test statistic is

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where

$$s_p = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{j=1}^m (y_j - \bar{y})^2}{n + m - 2}.$$

It has a $t$ distribution with $n + m - 2$ degrees of freedom under the null hypothesis.

## 21.16 Difference of means, potentially different variances

**Larsen–Marx [7]:** Section 9.2, p. 466

What if we don't know that the variance is the same for each sample? This is known as the **Behrens–Fisher Problem**.

The model equations are

$$X_i = \mu_X + \varepsilon_{Xi},\ i = 1, \ldots, m \qquad Y_j = \mu_Y + \varepsilon_{Yj},\ j = 1, \ldots, m,$$

where $\varepsilon_X \sim N(0, \sigma_X^2)$ and $\varepsilon_X \sim N(0, \sigma_Y^2)$.

The typical null hypothesis is $H_0\colon \mu_X - \mu_Y = 0$. The approximate test statistic is

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

It is approximately a $t$-distribution with $\nu$ degrees of freedom, where $\nu$ is the integer nearest to **Larsen–Marx [7]:** p. 466

$$\frac{\left(\frac{s_x^2}{s_y^2} + \frac{n}{m}\right)^2}{\frac{1}{(n-1)}\frac{s_x^2}{s_y^2} + \frac{1}{(m-1)}\left(\frac{n}{m}\right)^2}$$

But on Mathematica, `TTest[data1, data2]` does it all for you. In R, use `t.test(data1,data2)`. See Dytham [5, pp. 103–110].

## 21.17 Difference of means, Paired data

**Larsen–Marx [7]:** pp. 440–442

Sometimes there is a special structure to the data that simplifies the test of differences of mean. That is when the data are **paired data**. Typically one element of the pair is called the **control**. Then for each pair $(X_i, Y_i)$ the model equations are

$$X_i = \mu_X + \eta_i + \varepsilon_i, \qquad Y_i = \mu_Y + \eta_i + \varepsilon_i', \qquad i = 1, \dots, n,$$

where $\varepsilon$ and $\varepsilon'$ are independent and identically distributed Normal$(0, \sigma^2)$. Then

$$X_i - Y_i = (\mu_X - \mu_Y) + (\varepsilon_i - \varepsilon_i').$$

This can be tested as a simple $t$-test with $n - 1$ degrees of freedom.

## 21.18 Tests of Variance

Why might you care about variance? Suppose your laboratory has two microtomes. It is important for you to slice your tissue samples as uniformly as possible. Each machine has a tiny variation in the thicknesses it produces. You would like to use the one with the smaller variance. Hence the desire to test the difference of two variances.

Recall ([7, Theorem 7.3.2, p. 390] discussed in Lecture 20) that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

has a $\chi^2$-distribution with $n - 1$ degrees of freedom.

The $\chi^2$-distribution is not symmetric (see Figure 21.8), so for a two-sided test, you need two different critical values.

> The symbol $\chi^2_{\alpha,n}$ represents the $\alpha$ quantile of the $\chi^2$-distribution with $n$ degrees of freedom. That is, if $Q \sim \chi^2(n)$,
>
> $$P\left(Q \leqslant \chi^2_{\alpha,n}\right) = \alpha.$$
>
> N.B. This is different from the notation for $z_\alpha$ and $t_{\alpha,n}$. $(P(Z > z_\alpha) = \alpha.)$

Confidence intervals of $\sigma^2$: **Larsen–Marx [7]:** p. 412

$$P\left(\chi^2_{\alpha/2,n-1} \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2_{1-(\alpha/2),n-1}\right) = 1 - \alpha,$$

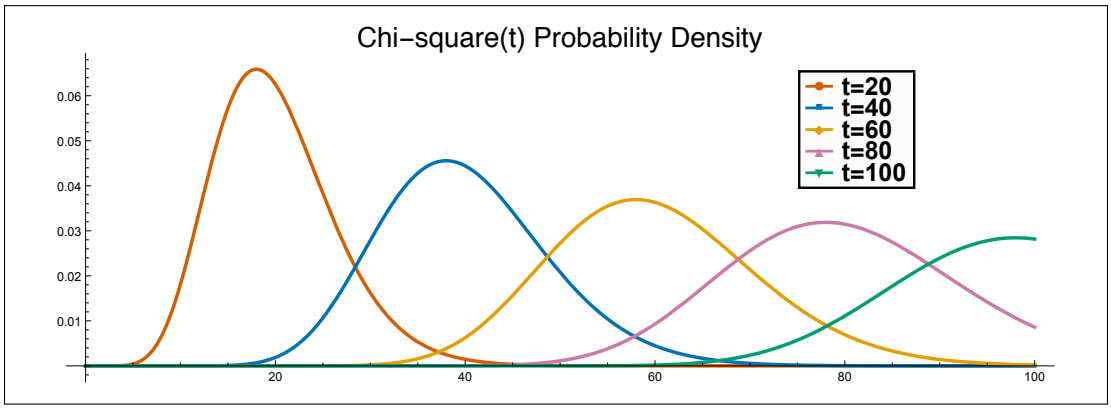


Figure 21.8. Chi-square pdfs.

so

the $1-\alpha$ confidence interval for $\sigma^2$ is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2),n-1}},\ \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$$

Table 5.7.2 in Larsen–Marx [7, p. 414] gives some useful values. You can use Mathematica or R to construct your own such table, and it serves as a useful check.

## 21.19 Confidence intervals and hypothesis test

We can turn the confidence interval into a hypothesis test. The following is Theorem 7.5.2 in Larsen–Marx [7, p. 415].

Let $X_1, \ldots, X_n$ be independent and identically distributed Normal$(\mu, \sigma^2)$. Let $s^2$ denote the unbiased sample variance estimate,

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

To test the null hypothesis

$$H_0\colon \sigma^2 = \sigma_0^2,$$

compute the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

a. Against the one-sided alternative $H_1\colon \sigma^2 > \sigma_0^2$ at the $\alpha$-level of significance, reject $H_0$ if

$$\chi^2 \geqslant {\chi^2}_{1-\alpha,n-1}.$$

b. Against the one-sided alternative $H_1\colon \sigma^2 < \sigma_0^2$ at the $\alpha$-level of significance, reject $H_0$ if

$$\chi^2 \leqslant {\chi^2}_{\alpha,n-1}.$$

c. Against the two-sided alternative $H_1\colon \sigma^2 \neq \sigma_0^2$ at the $\alpha$-level of significance, reject $H_0$ if

$$\text{either } \chi^2 \leqslant {\chi^2}_{\alpha/2,n-1} \text{ or } \chi^2 \geqslant {\chi^2}_{1-(\alpha/2),n-1}.$$

## 21.20 Testing Difference of Variances, *F* tests

How do we test the hypothesis that two sets of measurements come from normals with the same variance?

**Larsen–Marx [7]:** Section 9.3

Given $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ normal $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, then

$$\frac{\frac{(m-1)S_Y^2}{\sigma_Y^2}}{\frac{(n-1)S_X^2}{\sigma_X^2}} \sim F_{m-1,n-1}.$$

The symbol $F_{\alpha,m,n}$ represents the $\alpha$ quantile of the $F(m,n)$-distribution. That is, if $X \sim F(m,n)$,

$$P\left(X \leqslant F_{\alpha,m,n}\right) = \alpha.$$

N.B. This agrees with the convention for ${\chi^2}_{\alpha,n}$, but is different from the notation for $z_\alpha$ and $t_{\alpha,n}$. ($P\left(Z > z_\alpha\right) = \alpha$.)

Larsen–Marx [7, Theorem 9.3.1, pp. 471–472]

**21.20.1 Theorem** *To test*

$$H_0 \colon \sigma_X^2 = \sigma_Y^2$$

*at the $\alpha$ level of significance,*

1. *versus $H_1 \colon \sigma_X^2 > \sigma_Y^2$, reject $H_0$ if*

$$\frac{s_Y^2}{s_X^2} \leqslant F_{\alpha,m-1,n-1}.$$

2. *versus $H_1 \colon \sigma_X^2 < \sigma_Y^2$, reject $H_0$ if*

$$\frac{s_Y^2}{s_X^2} \geqslant F_{1-\alpha,m-1,n-1}.$$

3. *versus $H_1 \colon \sigma_X^2 \neq \sigma_Y^2$, reject $H_0$ if*

$$\frac{s_Y^2}{s_X^2} \leqslant F_{(\alpha/2),m-1,n-1} \quad \text{or} \quad \frac{s_Y^2}{s_X^2} \geqslant F_{1-(\alpha/2),m-1,n-1}.$$

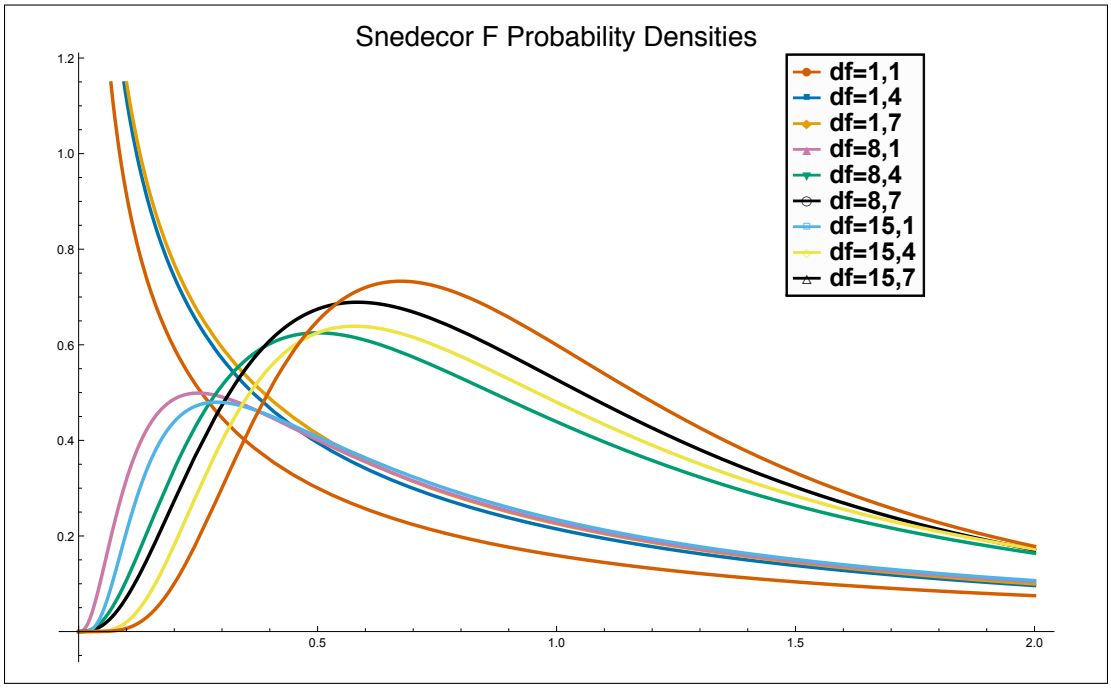


Figure 21.9. $F$-pdfs.

## 21.21 A caveat on hypothesis testing

Consider the coin tossing experiment. We want to test the Null Hypothesis that the probability $p$ of Tails is 1/2, $H_0\colon p = 0.5$. Now real coins are manufactured, and so subject to various imperfections, so it is hardly likely that the probability is exactly 1/2. In fact, it is reasonable to suppose the probability of a value exactly 1/2 is zero. The Strong Law of Large Numbers says that the MLE of $\#Tails/\#Tosses$ will converge with probability one to the true value, which is not 1/2, as the sample size gets large. Since the critical region is shrinking to zero with the sample size, with probability one we shall reject the Null Hypothesis if we get enough data. And we know this before we start! So why bother?

There are two responses to this question. The first is that if the coin is grossly biased, a hypothesis test with even a small sample size may reveal it. That is, hypothesis testing is an important part of data exploration.

The second response is that we are naïve to formulate such a restrictive hypothesis. We should restrict attention to null hypotheses such as the probability of Tails line in an interval $(0.5-\varepsilon, 0.5+\varepsilon)$, $H_0\colon p \in (0.5-\varepsilon, 0.5+\varepsilon)$, where $\varepsilon$ is chose small enough so that we don't care.

## 21.22 The significance of statistical significance

Many of the statistical tests that are performed are designed to examine either a correlation or the difference of two means. For example, does a particular treatment decrease the mean severity of a disease or increase the average longevity? Is there a correlation between certain seismic readings and the presence of oil? Is the measured velocity of light different when the it is moving with the aether drift or against it?

A typical null hypothesis is that two means are the same (or equivalently that their difference is zero), or perhaps that two variables have zero correlation. So the typical null hypothesis is of the form $\theta = 0$. Data are gathered and a test statistic $T$ is computed, and the null hypothesis is rejected if $T > t_\alpha$, where $t_\alpha$ is chosen so that *if* indeed $\theta = 0$, then $P(T > t_\alpha) = \alpha$. That is, you reject the null hypothesis if the test statistic is "significantly different from zero."

The point is that usually you want to reject the null hypothesis. You set things up so that your experiment is a success if it rejects the null hypothesis. Rejecting the null hypothesis means you have found something that significantly improves the mean, or is significantly correlated.

Hypothesis tests are predicated on the assumption that you already have in mind a hypothesis that you want to test, you set $\alpha$, gather your data, and then test for significance.

But this is rarely the way science is done. There may be hundreds of different drug-disease combinations that you want to test for efficacy. If you have computed your tests properly, and perform a hundred different experiments, then *even if the null hypothesis is always true*, 5% of your test statistics will be significantly different from zero at the 5% level of significance. You really should be looking at the distribution of the $100^{\text{th}}$ order statistic, not an individual test statistic.

Or maybe you look over the data to decide which correlations to test, or which variables to include in your analysis, and you discard those for which no correlations are found.

In other words, if there is "exploratory data analysis," or worse yet "data mining," then the fact that a significant test statistic is found is not significant. It is not clear what to do about this, but Ed Leamer [8, 9] has some suggestions that seem to have failed to catch on. Simmons et al. [11] have some concrete suggestions as well.

An important counter-example is in neuroscience, where a typical fMRI brain-imaging study divides the brain into about 60,000 "voxels," and looks for differences in the BOLD signal[1] in two different circumstances at different times. Deciding whether two brains are different involves literally millions of $t$-tests. Competent neuroscientists often apply the so-called Bonferroni

[1] This stands for blood-oxygen-level-dependent signal [10].

correction (see below) and use a significance level on the order of $\alpha = 1.5 \times 10^{-6}$ for each stand-alone $t$ test.)

But this approach seems wrong too. It is quite likely that voxels are spatially correlated, not independent and we are throwing away valuable information that is in the data. New techniques, based on random field theory are being explored. See, for example, Adler, Bartz, and Kou [1]. (Bartz is a recent Caltech alumnus.)

**Aside**: Craig Bennet, *et al.* [2] report on what can happen if a multiple comparison correction is not performed. They analyzed the effect of showing photos of humans in various kinds of social situations to a salmon, and found an area of the salmon brain and an area of the spinal column that responded. See Figure 21.10. (Incidentally, the salmon was dead.)

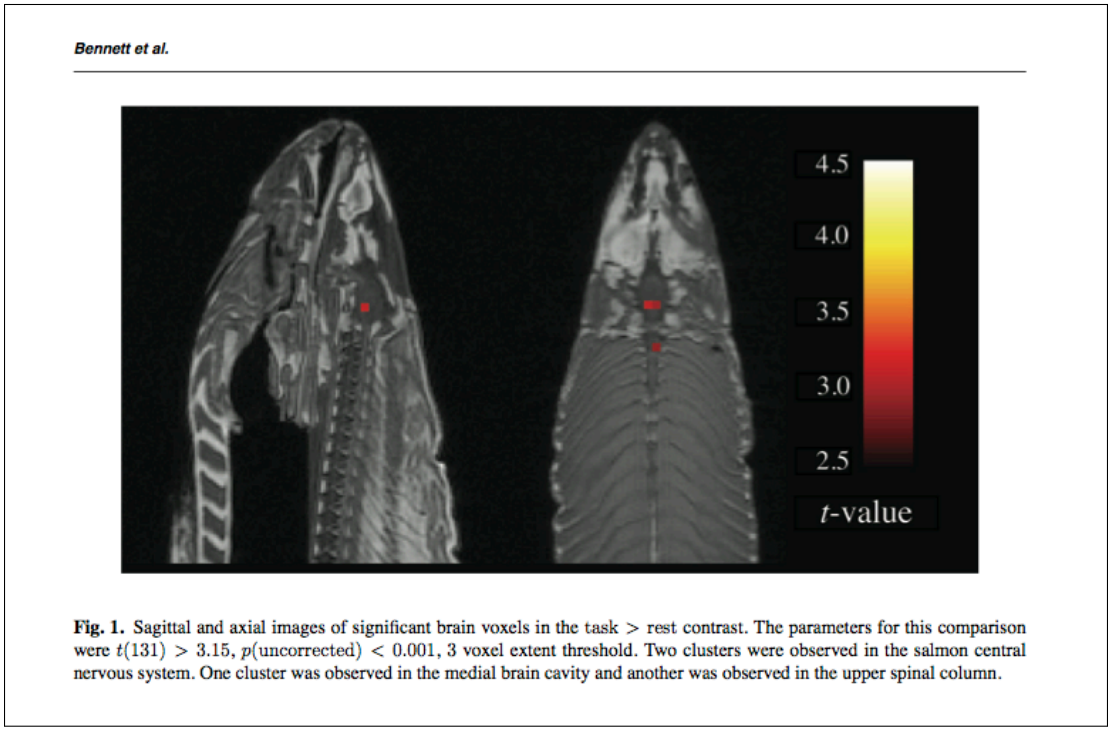


Figure 21.10. fMRI results for a postmortem Atlantic salmon.

## 21.23 The Bonferroni correction

Carlo Bonferroni [3] proposed the following crude antidote for the **multiple comparisons** problem. Suppose you have $n$ measurements, and want to test a hypothesis $H_0$ about each one. If each test is conducted at the significance level $\alpha/n$, then the probability that at least one of the $n$ tests rejects the null is no more than $\alpha$. This crude upper bound is based on the very crude Boole's Inequality: $P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i)$.

## Bibliography

[1] R. Adler, K. Bartz, and S. Kou. 2011. Estimating thresholding levels for random fields via Euler characteristics. Manuscript. http://www.kevinbartz.com/uploads/field/paper.pdf

[2] C. M. Bennett, A. A. Baird, M. B. Miller, and G. L. Wolford. 2010. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results* 1(1):1–5. The original web site (http://www.jsur.org/v1n1p1) has disappeared.

[3] C. E. Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8:3–62.

[4] L. Breiman. 1973. *Statistics: With a view toward applications*. Boston: Houghton Mifflin Co.

[5] C. Dytham. 2011. *Choosing and using statistics: A biologist's guide*, 3d. ed. Wiley–Blackwell.

[6] C. D. Ferris, F. E. Grubbs, and C. L. Weaver. 1946. Operating characteristics for the common statistical tests of significance. *Annals of Mathematical Statistics* 17(2):178–197. DOI: 10.2307/2236037

[7] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[8] E. E. Leamer. 1974. False models and post-data model construction. *Journal of the American Statistical Association* 69(345):122–131. http://www.jstor.org/stable/2285510

[9] ——— . 1975. "Explaining your results" as access-biased memory. *Journal of the American Statistical Association* 70(349):88–93. http://www.jstor.org/stable/2285382

[10] N. K. Logothetis and B. A. Wandell. 2004. Interpreting the BOLD signal. *Annual Review of Physiology* 66(1):735–769. PMID: 14977420 DOI: 10.1146/annurev.physiol.66.082602.092845

[11] J. P. Simmons, L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11):1359–1366. DOI: 10.1177/0956797611417632

[12] Student. 1908. The probable error of a mean. *Biometrika* 6(1):1–25. http://www.jstor.org/stable/2331554