Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2021

Lecture 22: Significance Tests, II

Relevant textbook passages:

Larsen–Marx [14]: Chapter 7, pp. 440–442 in Chapter 8, and Sections 9.1, 9.2, 9.3.

22.1 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[1] 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 [8] *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 flipping coin experiment? It gave 99.99%, not 49.91%. Remember, with any software, or any reference work, "Trust, but verify." In this course, I hope you learn enough to be able 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 they fit one of these off-the shelf models even if they don'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.

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 [7, p. 10] describes this argument as "only a cut above a hopeful appeal to the ghost of Laplace." [2] Nevertheless we shall start with

[1] Some students have asked what the definition of an off-the-shelf model is. What *I* mean is a model hat has been proposed and thoroughly analyzed, and often has techniques for estimating and testing built in to to statistical software. Using an off-the-shelf model makes it easy to compute without thinking.

[2] Here Breiman is alluding to Bishop George Berkeley's 1734 attack [4] on the rigor of the mathematics of his era, where he takes issue with Newton's notion of fluxions. In paragraph 35, he writes, "And what are these fluxions? The velocities of evanescent increments? And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?" [18, p.89]. Once again, I thank Lindsay Cleary for tracking this reference down for me.

a discussion of tests based on normality, since you will undoubtedly employ them at some point in the analysis of your laboratory data.

Here is a list of the off-the-shelf models and their associated hypothesis tests that we shall discuss. They are described in terms of **model equations**, which are a great aid to organizing one's thinking.

22.1.1 One-sample models

This is the case where there is a sample from a single population and the observations are assumed to satisfy the **model equations**

$$X_i = \mu_X + \varepsilon_i, \quad i = 1, \dots, n.$$

Here μ_X is an unknown population **parameter of interest** or the **systematic component**, and the ε_i's are often referred to as **error terms**. The error terms are assumed to be independent and identically distributed Normal$(0, \sigma^2)$. The assumption that the errors are independent and normally distributed is an essential part of the analysis,[3] as is the assumption that the variance of each ε_i is the same. The assumption that each error term has the same variance is known as the case of **homoskedasticity**. (The term **heteroskedasticity** is used when the errors are not homskedastic.) When I say the assumptions are essential, I do not mean that we cannot analyze more complicated cases, only that in other cases, we should analyze the model differently. The assumption that the mean of each error term is zero is not essential, provided the mean of each ε_i is the same. If the mean of the errors is $\mu_\varepsilon \neq 0$, we can add μ_ε to μ_X and subtract it from each ε_i to make the model fit the assumptions.

Get the cite.

There are two subcases that have been analyzed:

1. σ^2 is assumed to be known. This assumption is, in m opinion, not very realistic, but we make it simplify the exposition, and as a launching pad for the analysis when σ^2 is unknown.

2. σ^2 is unknown.

Hypotheses are regarding the mean μ. The three kinds of hypotheses are

1. Two-sided alternative:
$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

2. One-sided alternative:
$$H_0 : \mu \leqslant \mu_0, \quad H_1 : \mu > \mu_0.$$

3. The other one-sided alternative:
$$H_0 : \mu \geqslant \mu_0, \quad H_1 : \mu < \mu_0.$$

22.1.2 Paired samples

In this case there is a population of individuals that are subjected to two "treatments." The pair of treatments generate a pair of measurements X_i and Y_i for each individual i. (Think of before and after measurements on a patient.) We allow for unknown **individual fixed effects**. The model equations are

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

[3] Larsen–Marx [14] argue that the assumption of normality is not crucial, and that the procedures described here "work" in a variety of cases.

where the ε_i and ε'_i are assumed to be independent and identically distributed Normal$(0, \sigma^2)$, and σ^2 is usually assumed to be unknown. The individual fixed effects η_i are also assumed to be unknown, but not random.

Hypotheses are formulated regarding the means μ_X and μ_Y. The three kinds of hypotheses are

1. Two-sided alternative:
$$H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X \neq \mu_Y.$$

2. One-sided alternative:
$$H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X > \mu_Y$$
or
$$H_0 : \mu_X \leqslant \mu_Y, \quad H_1 : \mu_X > \mu_Y.$$

3. The other one-sided alternative:
$$H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X < \mu_Y.$$
or
$$H_0 : \mu_X \geqslant \mu_Y, \quad H_1 : \mu_X < \mu_Y.$$

22.1.3 Two-sample models

In this there are samples of the same measurement from two different populations The model equations are:

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

where the error terms ε are assumed to be independent and identically distributed Normal$(0, \sigma_X^2)$ and Normal$(0, \sigma_Y^2)$.

There are two subcases:

1. σ_X^2 and σ_Y^2 are assumed to be the same, but their value is unknown.
2. σ_X^2 and σ_Y^2 are not assumed to be the same, and their values are unknown.

Hypotheses are regarding the means μ_X and μ_Y. The three kinds of hypotheses are

1. Two-sided alternative:
$$H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X \neq \mu_Y.$$

2. One-sided alternative:
$$H_0 : \mu_X \leqslant \mu_Y, \quad H_1 : \mu_X > \mu_Y.$$

3. The other one-sided alternative:
$$H_0 : \mu_X \geqslant \mu_Y, \quad H_1 : \mu_X < \mu_Y.$$

22.1.4 Test of other parameters; other models

As you can see there is a large number of models and test dealing with the mean. We can also use this same taxonomy to discuss tests regarding variances. For instance, does on population hav a larger variance than another?

Then there are models with three populations, etc. We shall discuss some of these cases in Lectures 24 and 25 on the standard linear model. In Lecture 26 we discuss testing without normality assumptions.

22.2 One-sample models with known variance

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

Larsen–Marx [14]: 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 18 that even though we can't observe Z (since μ is unknown), if we know σ, then we can use the observed value $\bar{x}$ of $\bar{X}$ to get a confidence interval for μ, 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 18.13.) 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 μ.

There are two common classes of null hypotheses and alternative hypotheses regarding μ:

- A **two-sided** hypothesis/test typically deals with a null hypothesis of the form $H_0\colon \mu = \mu_0$, where μ_0 is some fixed value that you wish to test. The alternative is $H_1\colon \mu \neq \mu_0$. This is called a two-sided alternative because it allows for rejecting the null hypothesis if $\mu < \mu_0$ or if $\mu > \mu_0$.

- A **one-sided** hypothesis/test deals with a null hypothesis of the form $H_0\colon \mu = \mu_0$ and the alternative is that $H_1\colon \mu > \mu_0$. (Or it could be that the null hypothesis is $H_0\colon \mu \geqslant \mu_0$ and the alternative is $H_1\colon \mu < \mu_0$.) The point is that you care about only one direction that μ might differ from μ_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, μ_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 can tell you how to test a hypothesis, but not *which* hypothesis to test.

There is another kind of null hypothesis you might want to test in the one-sided case: Namely instead of the null hypothesis $H_0\colon \mu = \mu_0$ with the alternative $H_1\colon \mu > \mu_0$, the null might be that $H_0\colon \mu \leqslant \mu_0$ with the alternative $H_1\colon \mu > \mu_0$. As long as you have a case with a monotone likelihood ratio (Section 21.11 ⋆) the likelihood ratio test will be the same in either case.

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

22.2.1 One-sided alternatives

When the null hypothesis is $H_0\colon \mu \leqslant \mu_0$ and the alternative hypothesis $h_1\colon \mu > \mu_0$, 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 α choose c so that $P_{\mu_0}\big(\bar{X} > c\big) = \alpha$. (Recall Section 18.13.)

Now

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

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

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \tag{3}$$

and we

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

22.3 On one-sided versus two-sided alternative hypotheses

Here is a good question a student once asked me. If we want to test the null hypothesis $H_0\colon \mu = 0$ versus the two-sided alternative $H_1\colon \mu \neq 0$, at the α level, you say to look up $z_{\alpha/2}$ and reject the null hypothesis if

$$\big|\bar{X}\big| \geqslant z_{\alpha/2}.$$

But if $\bar{X} > 0$, shouldn't we use z_α rather than $z_{\alpha/2}$?

Let's think about this for a moment. Another way to phrase this procedure is this: If $\bar{X} > 0$, then choose the alternative $H_1\colon \mu > 0$, but if $\bar{X} < 0$, then choose the alternative $H_1\colon \mu < 0$, and test the resulting alternative at the α-level of significance. But this procedure does not have significance level (probability of Type I error) α—the probability of a Type I error is actually 2α. To see this, observe that the probability of rejecting H_0 under this procedure is, when $\mu = 0$ is, by the Law of Average Conditional Probability, Proposition 4.5.3, equal to

$$\underbrace{P_0\big(\bar{X} \geqslant z_\alpha \bigm| \bar{X} > 0\big)}_{=2P_0(\bar{X}\geqslant z_\alpha)=2\alpha}\underbrace{P_0\big(\bar{X} > 0\big)}_{=\frac{1}{2}} + \underbrace{P_0\big(\bar{X} \leqslant z_\alpha \bigm| \bar{X} < 0\big)}_{=2P_0(\bar{X}\leqslant z_\alpha)=2\alpha}\underbrace{P_0\big(\bar{X} < 0\big)}_{=\frac{1}{2}} = 2\alpha.$$

That is, looking at the test statistic $\bar{X}$ and then deciding the alternative increases the probability of rejecting the null hypothesis. This makes sense if you think about. You've chosen the alternative to be the one that makes the null hypothesis more likely to be rejected.

This is an example of what econometricians call **data mining**, that is, deciding your hypotheses after exploratory data analysis. In this case data mining changes your significance level, and potentially invalidates your test results.

22.4 Power and sample size

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

Larsen–Marx [14]: pp. 372–374

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

In order for the power to be equal to γ we must have

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

By (3) this entails

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

or

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

This tells us how large the sample has to be to get the power to be equal to γ for a test with significance level α. 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 α and γ 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.

22.4.1 Example Larsen–Marx [14, Example 6.4.1, pp. 373–374] ask for 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

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

□

22.4.2 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 $\mu = 0.1$. □

22.5 ⋆ Detectability thresholds

Should this be moved to Lecture 21?

Leo Breiman [7, Chapter 5] discusses the notion of **detectability**. In the context of a hypothesis test (T, C) for the null hypothesis $H_0 : \theta \in \Theta_0$ with significance level α, we say that the parameter value $\theta \notin \Theta_0$ can be **detected at level $\boldsymbol{\alpha}$** by the test if

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

or

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

We say that Δ 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.

22.5.1 Example So following the analysis of the normal case in the previous section, for a one-sided text of significance α 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 α, 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

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

□

22.6 What if σ is unknown?

The problem with the analysis above is that we seldom know σ. In Lecture 18, we derived the Maximum Likelihood Estimators for μ and σ^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:

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

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.

22.7 The chi-square distribution

Recall that the **chi-square(m)** or **χ^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 [14, Theorem 7.3.1, p. 389]. It is also a Gamma($\frac{m}{2}, \frac{1}{2}$) distribution. See Figure 22.1 for the shape of the density.

The next result appears as Corollary 11.5.2. It may also be found in Larsen and Marx[14, Theorem 7.3.2, p. 390].

22.7.1 Fact *If $X_1, \dots, X_n$ 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.)

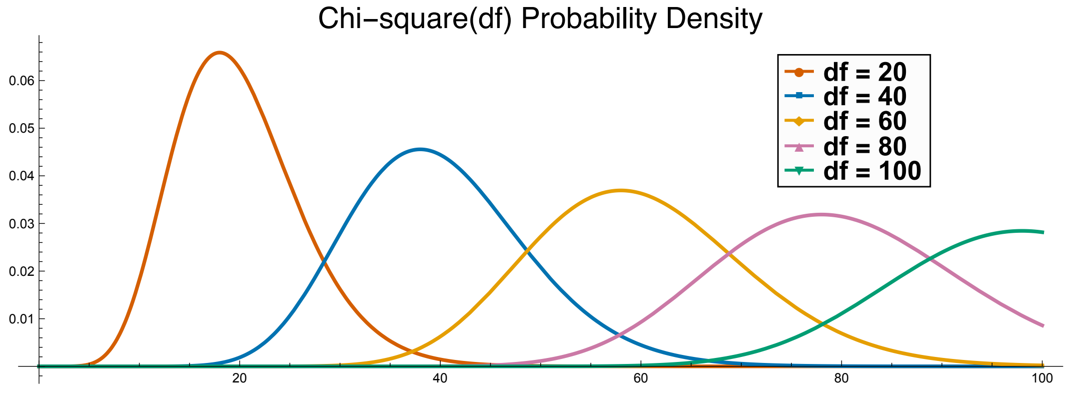


Figure 22.1. Chi-square pdfs.

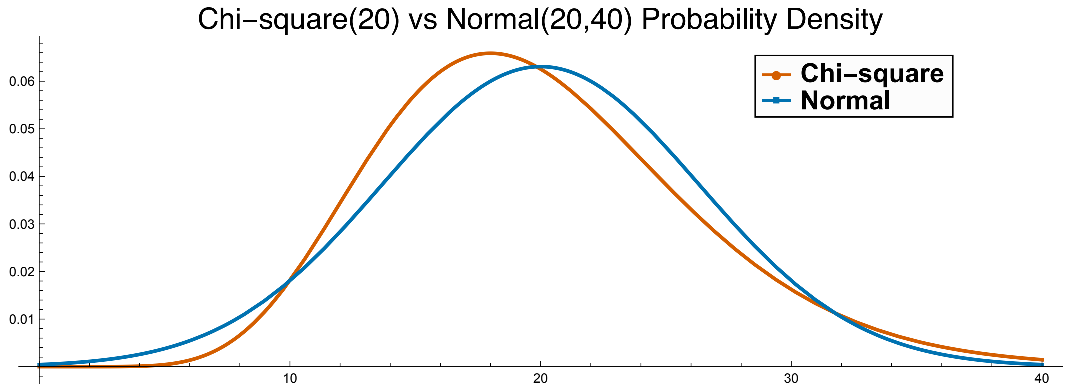


Figure 22.2. Chi-square vs Normal.

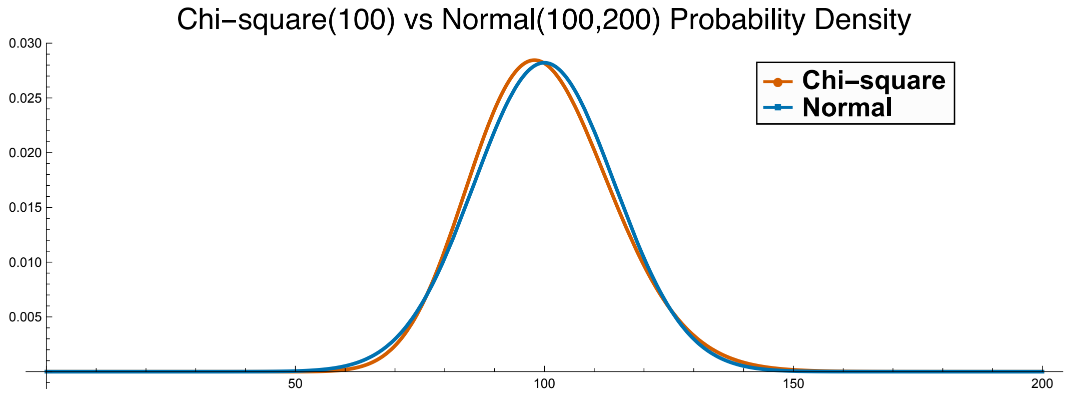


Figure 22.3. Chi-square vs Normal.

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

The $F_{m,n}$ density is given by [14, 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 22.4.

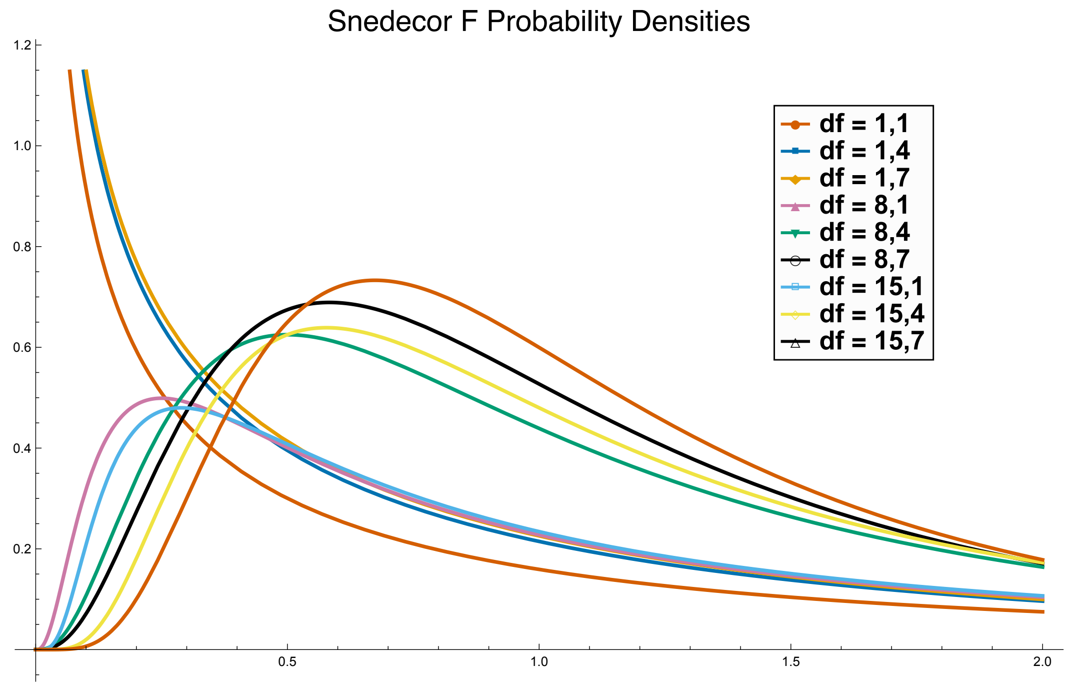


Figure 22.4. Snedecor F pdfs.

22.9 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 $\boldsymbol{t}$-distribution with $\boldsymbol{n}$ degrees of freedom**. This distribution figures in testing hypotheses about means for small samples, when the variance is unknown, and must be estimated form the data.

Aside: The t-distribution was first calculated by William Sealy Gossett (1876-1937) in 1908. He spent his working life as an employee of Arthur Guinness, Son & Co., Ltd., brewers of stout at the St. James Gate Brewery in Dublin. Because of his employer's obsession with secrecy, Gossett was allowed to publish his scientifc work only if he used a pseudonym, and he chose the *nom de plume* Student [26]. (See Larsen–Marx [14, pp. 386–387], and Pearson [23, pp. 5, 17].) There is a story (reported by Pearson's son [23, p. 73]) that Karl Pearson, when asked by Gossett for advice in dealing with small sample sizes, jokingly remarked, "Only naughty brewers deal in small samples."

The density is given by [14, 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 22.5.

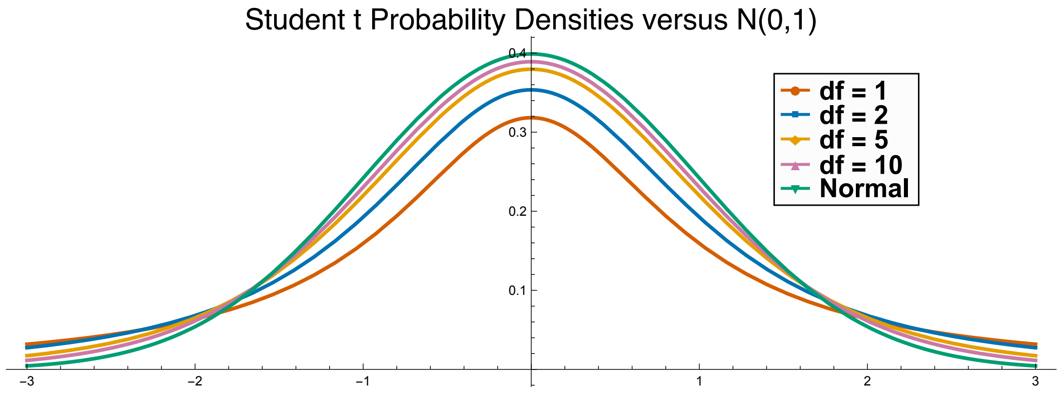


Figure 22.5. Student t-pdfs and the Standard Normal. The t densities are more spread out.

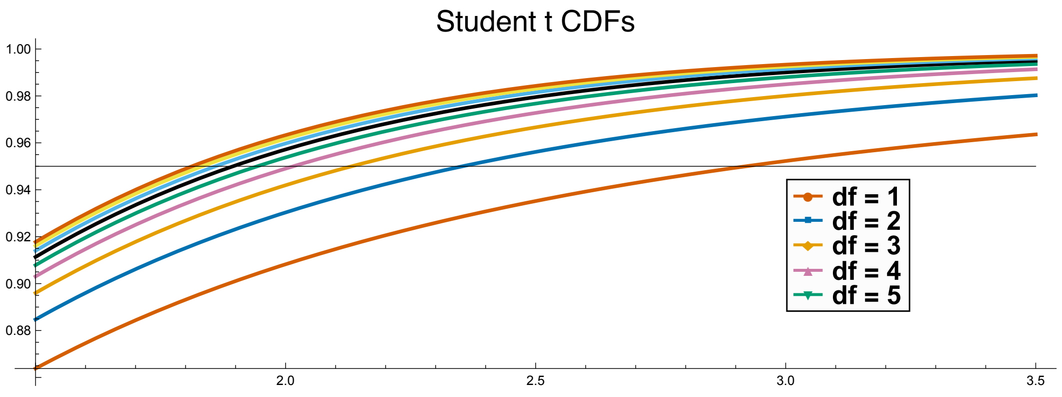


Figure 22.6. The tail of the CDF for various Student t-distributions.

22.10 One-sample models with unknown variance

22.10.1 Theorem *[14, 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.

22.10.1 Confidence interval for $\boldsymbol{\mu}$

Recall the following Definition 18.12.2. Define $\boldsymbol{t_{\alpha,n}}$ by

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

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

Then

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

or equivalently

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

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 μ is the interval

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

22.10.2 $\boldsymbol{t}$-quantiles versus $\boldsymbol{z}$-quantiles

The values $z_\alpha = 1.96$, which are used to construct a $(1 - \alpha\%)$ confidence intervals based on knowing the standard deviation σ, can be very misleading for small sample sizes, when σ is estimated by the unbiased version of the MLE estimate. The following Table 22.1 gives z_α and $t_{\alpha,n}$ for various values of α and n. This shows how the critical value of a test changes with the number of degrees of freedom.

22.10.3 The "$\boldsymbol{t}$-test" for a one-sample model

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

	degrees of freedom n										
	1	2	4	8	16	32	64	128	256	512	
α	$t_{\alpha,n}$										z_α
0.10	3.08	1.89	1.53	1.4	1.34	1.31	1.29	1.29	1.28	1.28	1.28
0.05	6.31	2.92	2.13	1.86	1.75	1.69	1.67	1.66	1.65	1.65	1.64
0.025	12.71	4.3	2.78	2.31	2.12	2.04	2.	1.98	1.97	1.96	1.96
0.01	31.82	6.96	3.75	2.9	2.58	2.45	2.39	2.36	2.34	2.33	2.33
0.005	63.66	9.92	4.6	3.36	2.92	2.74	2.65	2.61	2.6	2.59	2.58

Table 22.1. $t_{\alpha,n}$ compared to z_α for various degrees of freedom n and significance levels α.

To test the Null Hypothesis

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

versus the one-sided alternative

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

at the α 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 [14, 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 =` μ_0`)` returns a detailed report on the two-sided test of the hypothesis $\mu = 0$ including a confidence interval for μ. (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 [8] has excellent sample code for a number of programs including R, SPSS, Minitab, and even Excel, but not Mathematica.

22.10.4 ⋆ 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 σ 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 α. We want to compute the power at μ, which is just

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

The problem is that if each $X_i \sim N(\mu, \sigma^2)$, then the test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ 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 [9] 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 μ 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 Φ depends only on the value s^2/σ^2, which we know has a χ^2 distribution. One can compute the expected value to get the power of the test at μ. Ferris, et. al. do this and report the operating characteristic (1 minus the power) graphically. Their graph is reproduced in Breiman [7, p. 147].

Breiman recommends using the same rule for detectability thresholds with unknown σ as for known σ 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 σ, we cannot use this to compute a sample size ex ante. We first have to estimate σ, and then use that as a guide to deciding whether to collect a larger sample to increase the power of the test.

22.11 Two-sample model, same variances

Larsen–Marx [14]: Section 9.2

Given $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ 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 [14].

The model equations are

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

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

The test statistic uses the pooled sample to estimate the variance for the following t-statistic:

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

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.

22.12 Two-sample model, potentially different variances

Larsen–Marx [14]: 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

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

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 test statistic proposed by Welch [27] is

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

Larsen–Marx [14]: p. 466

The Welch statistic has a distribution that is approximately a t-distribution with ν degrees of freedom, where ν is the integer nearest to

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

But on Mathematica, `TTest[data1, data2]` does it all for you. In R, use `t.test(data1,data2)`. See Dytham [8, pp. 103–110].

22.13 Difference of means, Paired samples

Larsen–Marx [14]: 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 ε and ε' are independent and identically distributed Normal$(0, \sigma^2)$, and σ^2 is unknown. 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.

22.14 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 ([14, Theorem 7.3.2, p. 390] discussed in Lecture 21) that

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

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

The χ^2-distribution is not symmetric (see Figure 22.7), so for a two-sided test, you need two different critical values.

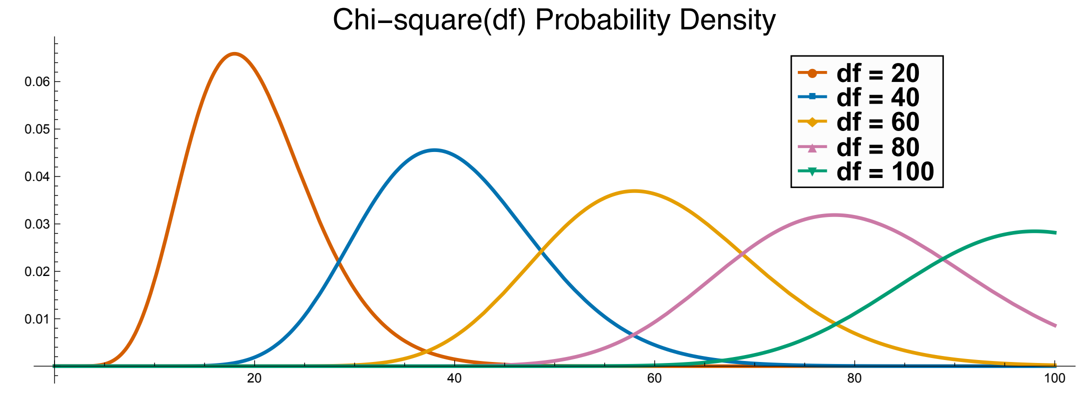


Figure 22.7. Chi-square pdfs.

22.14.1 Definition *The symbol $\chi^2_{\alpha,n}$ represents the α quantile of the χ^2-distribution with n degrees of freedom. That is, if $Q \sim \chi^2(n)$,*

$$P\big(Q \leqslant \chi^2_{\alpha,n}\big) = \alpha.$$

N.B. This is different from the notation for z_α and $t_{\alpha,n}$. ($P(Z > z_\alpha) = \alpha$.)

Confidence intervals of σ^2:

Larsen–Marx [14]: p. 412

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

so

the $1-\alpha$ confidence interval for σ^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 [14, 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. (See Section 22.20.)

22.14.1 Confidence intervals and hypothesis tests

We can turn the confidence interval into a hypothesis test. The following is Theorem 7.5.2 in Larsen–Marx [14, p. 415].

Let $X_1, \dots, X_n$ be independent and identically distributed Normal(μ, σ^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

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

at the α-level of significance, compute the test statistic

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

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

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

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

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

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

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

The cutoffs are not symmetric, because the χ^2 distribution is not symmetric.

22.15 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 [14]: Section 9.3

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

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

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

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

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

Larsen–Marx [14, Theorem 9.3.1, pp. 471–472]

22.15.2 Theorem (*F*-test) *To test*

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

at the α 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*

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

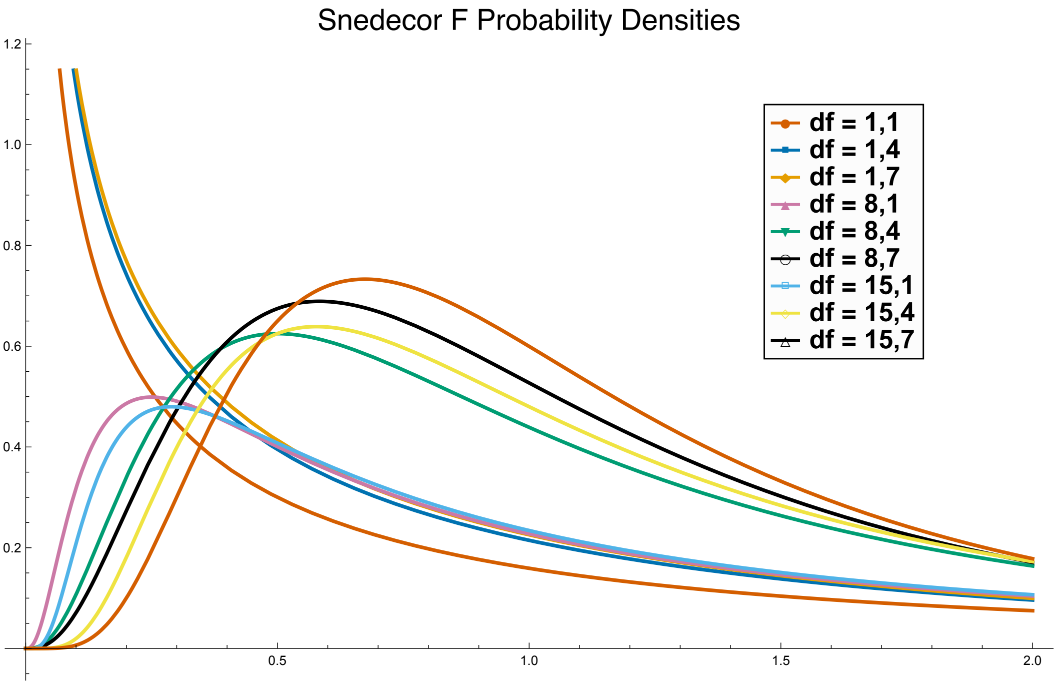


Figure 22.8. *F*-pdfs.

22.16 Some recent developments

In 2012, Zhang, Xu, and Chen [28] developed a test for the difference of two Normal samples.

Let $x_1, \dots, x_n$ and $y_1, \dots, y_m$ be independent samples from normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. To test

$$\begin{gathered} H_0\colon \mu_1 = \mu_1 \text{ and } \sigma_1^2 = \sigma_2^2 \\ \text{versus} \\ H_1\colon \mu_1 \neq \mu_1 \text{ and/or } \sigma_1^2 \neq \sigma_2^2, \end{gathered}$$

Write out the derivation.

form the likelihood ratio test statistic

$$\lambda = \frac{\left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}\right]^{n/2} \left[\frac{\sum_{j=1}^m (y_j - \bar{y})^2}{m}\right]^{m/2}}{\left[\frac{\sum_{i=1}^n (x_i - \bar{u})^2 + \sum_{j=1}^m (y_j - \bar{u})^2}{m+n}\right]^{(m+n)/2}}$$

where $\bar{u}$ is the overall sample average. The Null hypothesis is rejected if this likelihood ratio is too small.

Earlier, Pearson and Neyman [20] showed that as $n, m \to \infty$, then the distribution of λ under the null hypothesis is approximately Uniform$(0, 1)$. Zhang et al find the exact distribution (depending on n and m) and give an implementation of their test in R.

They recommend the following procedure:

1. First test the null hypothesis that the x and y samples come from the same distribution, using their test. If the null hypothesis is not rejected, you're done.

2. If H_0 is rejected, you should next test the null hypothesis

$$\begin{gathered} H_0^\sigma : \sigma_1^2 = \sigma_2^2 \\ \text{versus} \\ H_1^\sigma : \sigma_1^2 \neq \sigma_2^2, \end{gathered}$$

using the F-test (Theorem 22.15.2).

If H_0^σ is not rejected, use the pooled t-test (Section 22.11) to test the null hypothesis

$$\begin{gathered} H_0^\mu : \mu_1 = \mu_1 \\ \text{versus} \\ H_1^\mu : \mu_1 \neq \mu_1. \end{gathered}$$

3. If H_0^σ is rejected, use the Welch approximate t-test of Section 22.12 to test the null hypothesis

$$\begin{gathered} H_0^\mu : \mu_1 = \mu_1 \\ \text{versus} \\ H_1^\mu : \mu_1 \neq \mu_1. \end{gathered}$$

4. Finally, you should test the hypothesis that the data are normal. We will get to that in a couple of lectures.

22.17 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?

Ask Richard Roll for the name of this paradox.

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 ε is chose small enough so that we don't care.

22.18 Abuses of *p*-values

Significance versus insignificance

Gelman and Stern [10] and Nieuwenhuis et al. [22], among others, point out that many published studies use the following sort of logic. Two different treatments are tested. Treatment A is effective at a stringent level of significance and Treatment B is not. Therefore Treatment A is more effective than Treatment B. Gelman and Stern [10] give the following simple example that shows that this argument may not be valid. (And it is not because the *p*-values are close but on opposite sides of significance, e.g., 4.9% versus 5.1%.)

22.18.1 Example (Significance versus insignificance) For simplicity, let's assume normality and equal sample sizes, and the null hypothesis is that the data have mean 0, the one-sided alternative is that the mean is greater than 0. Data for Treatment A have mean 25 and standard error 10, while data for Treatment B have mean 10 and standard error 10. The one-sided p-value for A is 0.6% (highly significant) while that for B is 15.9%. The difference in p-values is nowhere near to being close. Yet, if we test the difference, the difference has mean 15 and standard error $= \sqrt{10^2 + 10^2} \approx 14.1$, so the difference is $15/14.1 = 1.06$ standard deviations away from zero, which has a two-sided p-value equal to 28.9%, and would be judged insignificant by almost any standard. □

Selection on the basis of *p*-values

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 measurement apparatus 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_α 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 α, 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.

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 [15, 16] has some suggestions that seem to have failed to catch on. Simmons et al. [25] 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[4] 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 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, Kou, and Monod [1, 2]. (Bartz is a recent Caltech alumnus.)

Aside: Craig Bennet, *et al.* [3] 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 22.9. (Incidentally, the salmon was dead.)

Publication bias

****

22.19 The Bonferroni correction

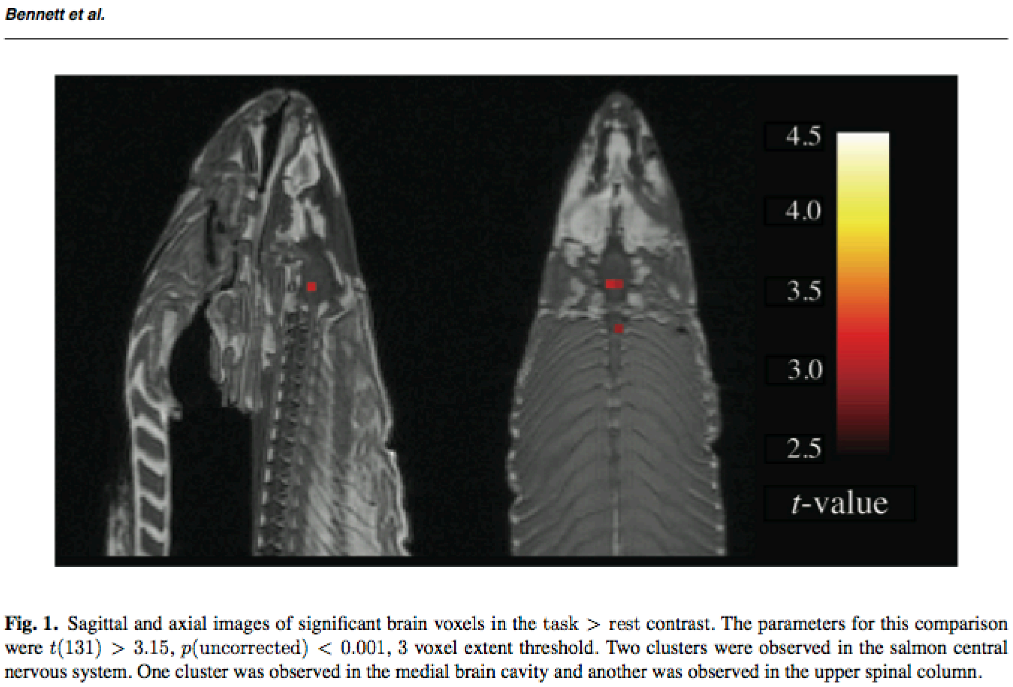
Carlo Bonferroni [5] proposed the following crude antidote for the **multiple comparisons** problem. Suppose you have n measurements, and want to test a hypothesis H_0 about each one. If each test is conducted at the significance level α/n, then the probability that at least one of the n tests rejects the null is no more than α. This crude upper bound is based on the very crude Boole's Inequality: $P\left(\bigcup\limits_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i)$. The paper by Naiman and Wynn [19] offers more details.

22.20 Quantile cutoffs for classic tests

Here is a summary the various quantiles used as critical values for the classic tests described in this lecture.

Recall the definitions:

[4] This stands for blood-oxygen-level-dependent signal [17].



Bennett et al.

Fig. 1. Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.

Figure 22.9. fMRI results for a postmortem Atlantic salmon.

Definition 18.13.1 z_α is defined by

$$P(Z > z_\alpha) = \alpha, ,$$

where $Z \sim N(0,1)$.

Definition 18.12.2 $t_{\alpha,n}$ is defined by

$$P(T_n > t_{\alpha,n}) = \alpha,$$

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

Definition 22.14.1 $\chi^2_{\alpha,n}$ is defined by

$$P\big(Q \leqslant \chi^2_{\alpha,n}\big) = \alpha,$$

where $Q \sim \chi^2(n)$. Note that the inequality is different from the case of z_α and $t_{\alpha,n}$.

Definition 22.15.1 $F_{\alpha,m,n}$ is defined by

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

where $X \sim F(m,n)$. This agrees with the convention for $\chi^2{}_{\alpha,n}$, but is different from the notation for z_α and $t_{\alpha,n}$.

And here is how to look them up with Mathematica or R. Be sure to replace α, m, and n in the code by their appropriate numeric values.

	Mathematica
z_α	`InverseCDF[NormalDistribution[], 1 - α]`
$t_{\alpha,n}$	`InverseCDF[StudentTDistribution[n], 1 - α]`
$\chi^2_{\alpha,n}$	`InverseCDF[ChiSquareDistribution[n], α]`
$F_{\alpha,m,n}$	`InverseCDF[FRatioDistribution[m,n], α]`

	R
z_α	`qnorm(α, lower.tail = FALSE)`
$t_{\alpha,n}$	`qt(α, n, lower.tail = FALSE)`
$\chi^2_{\alpha,n}$	`qchisq(α, n)`
$F_{\alpha,m,n}$	`qf(α, m, n)`

Bibliography

[1] R. J. Adler, K. Bartz, and S. C. Kou. 2011. Estimating thresholding levels for random fields via Euler characteristics. Manuscript.

[2] R. J. Adler, K. Bartz, S. C. Kou, and A. Monod. 2017. Estimating thresholding levels for random fields via Euler characteristics. Manuscript.
https://arxiv.org/pdf/1704.08562.pdf

[3] 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.

[4] G. Berkeley. 1734. The analyst. In Luce and Jessop [18], pages 65–102.

[5] 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.

[6] R. Border, E. C. Johnson, L. M. Evans, A. Smolen, N. Berley, P. F. Sullivan, and M. C. Keller. 2019. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry* 176(5):367–375. DOI: 10.1176/appi.ajp.2018.18070881

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

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

[9] 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

[10] A. Gelman and H. Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* 60(4):328–331. DOI: 10.1198/000313006X152649

[11] J. P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine* 2(8):e124. DOI: 10.1371/journal.pmed.0020124

[12] ——— . 2018. The proposal to lower p value thresholds to .005. *JAMA* 319(14):1429–1430. DOI: 10.1001/jama.2018.1536

[13] ——— . 2019. What have we (not) learnt from millions of scientific papers with p values? *The American Statistician* 73(sup1):20–25. DOI: 10.1080/00031305.2018.1447512

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

[15] 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`

[16] ——— . 1975. "Explaining your results" as access-biased memory. *Journal of the American Statistical Association* 70(349):88–93. `http://www.jstor.org/stable/2285382`

[17] 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

[18] A. A. Luce and T. E. Jessop, eds. 1951. *The works of George Berkeley Bishop of Cloyne*, volume IV. London: Thomas Nelson and Sons.

[19] D. Q. Naiman and H. P. Wynn. 1992. Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Annals of Statistics* 20(1):43–76. `http://www.jstor.org/stable/2242150`

[20] J. Neyman and E. S. Pearson. 1930. On the problem of two samples. *Bulletin de l'Académie Polonaise des Sciences* pages 73–96. Reprinted in [21, Chapter 3, pp. 99–115].

[21] ——— . 1966. *Joint statistical papers*. Berkeley: University of California Press.

[22] S. Nieuwenhuis, B. U. Forstmann, and E.-J. Wagenmakers. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience* 14(9):1105 – 1107. DOI: 10.1038/nn.2886

[23] E. S. Pearson. 1990. *'Student': A statistical biography of William Sealy Gossett*. Oxford, England: Clarendon Press. Based on writings by E. S. Pearson, edited and augmented by R. L. Plackett with the assistance of G.A. Barnard.

[24] S. K. Perng and R. C. Littell. 1976. A test of equality of two normal population means and variances. *Journal of the American Statistical Association* 71(356):968–971. DOI: 10.1080/01621459.1976.10480978

[25] 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

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

[27] B. L. Welch. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29(3/4):350–362. http://www.jstor.org/stable/2332010

[28] L. Zhang, X. Xu, and G. Chen. 2012. The exact likelihood ratio test for equality of two normal populations. *The American Statistician* 66(3):180–184. DOI: 10.1080/00031305.2012.707083