Inferences Based on Two Samples
(Devore Chapter Nine)

MATH-252-01: Probability and Statistics II*

Spring 2018

Contents

Tuesday 13 February 2018

Thus far we've considered inferences (confidence intervals and hypothesis tests) for a sample of data assumed to be drawn from a single distribution. It's often the case, though, that we have multiple samples of data drawn from different distributions, and we're interested in the relationships between those distributions (for example, how the means of the distributions are related). We now turn to so-called two-sample inferences.

1 Two-Sample Z Tests

As a simple starting point, suppose we have a sample of size m drawn from a distribution with mean μ_1 and variance σ_1^2, which we'll call $\{X_i\}$ when thinking of it as a random sample or $\{x_i\}$ when dealing with the actual data values, and another sample of size n, drawn from a distribution with mean μ_2 and variance σ_2^2, which we call $\{Y_j\}$ or $\{y_j\}$. We're interested in inferences about the difference between the two means μ_1 and μ_2. For instance we may be considering the heights in cm of men and women, and want to know, based on a finite sample, how the average heights for men and women compare.

This sort of problem seems like it could be complicated, since the size of the two samples need not be the same, but we know that for one-sample inferences that we can summarize the data using the sample mean, so we construct the means $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$, and consider the behavior of $\overline{X} - \overline{Y}$. We know that $E(\overline{X}) = \mu_1$, $E(\overline{Y}) = \mu_2$, $V(\overline{X}) = \sigma_1^2$, and $V(\overline{Y}) = \sigma_2^2$, and assuming all the random variables in the samples to be independent random variables, the statistics $\overline{X}$ and $\overline{Y}$ should also be independent, which means

$$E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_1 - \mu_2 \tag{1.1a}$$

$$V(\overline{X} - \overline{Y}) = V(\overline{X}) + (-1)^2 V(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \tag{1.1b}$$

1.1 Normal with Known Variances

In the simplest case, we know the samples to be both drawn from normal distributions, and we know the variances. Then we know $\overline{X}$ and $\overline{Y}$ are independent normally-distributed random variables, which means $\overline{X} - \overline{Y}$ is normally distributed with the mean and variance given above, and

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}} \tag{1.2}$$

is standard normal distributed. That means we can do all of the usual standard-normal inferences. For example:

1. A $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ is

$$\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)} \quad \text{to} \quad \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)} \tag{1.3}$$

2. A $100(1-\alpha)\%$ upper bound on $\mu_1 - \mu_2$ is

$$\overline{x} - \overline{y} + z_{\alpha}\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)} \tag{1.4}$$

3. A $100(1-\alpha)\%$ lower bound on $\mu_1 - \mu_2$ is

$$\overline{x} - \overline{y} - z_{\alpha}\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)} \tag{1.5}$$

If we construct the Z-statistic

$$z = \frac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}} \tag{1.6}$$

it can be used to test the hypothesis H_0: $\overline{x} - \overline{y} = \Delta_0$ at significance level (false alarm probability) α:

1. Against the alternative hypothesis H_a: $\overline{x} - \overline{y} > \Delta_0$ with an upper-tailed test which rejects H_0 if $z > z_\alpha$
2. Against the alternative hypothesis H_a: $\overline{x} - \overline{y} < \Delta_0$ with a lower-tailed test which rejects H_0 if $z < -z_\alpha$
3. Against the alternative hypothesis H_a: $\overline{x} - \overline{y} \neq \Delta_0$ with a lower-tailed test which rejects H_0 if $|z| > z_{\alpha/2}$

Likewise, the P value for H_0 is

1. If H_a is $\overline{x} - \overline{y} > \Delta_0$,

$$P = 1 - \Phi\left(\frac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}}\right) \tag{1.7}$$

2. If H_a is $\overline{x} - \overline{y} < \Delta_0$,

$$P = \Phi\left(\frac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}}\right) \tag{1.8}$$

3. If H_a is $\overline{x} - \overline{y} \neq \Delta_0$,

$$P = 2\Phi\left(-\frac{|\overline{x} - \overline{y} - \Delta_0|}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}}\right) \tag{1.9}$$

1.2 Large Samples

If we don't know the variances σ_1^2 and σ_2^2, we can try to estimate them from the data using the sample variances

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2 \tag{1.10a}$$

$$s_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2 \tag{1.10b}$$

As usual, if n and m are large (each $\gtrsim 40$), we can invoke the central limit theorem and find that the statistic

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/m) + (S_2^2/n)}} \tag{1.11}$$

is approximately standard normal, and all of the same inferences go through as before.

Practice Problems

9.3, 9.7, 9.15

Thursday 15 February 2018

Review for Prelim Exam One (Chapters 6-8). Please bring questions, and ideally ask them by email before class.

Tuesday 20 February 2018

Prelim Exam One (Chapters 6-8). Closed book, closed notes, but you may bring one handwritten 8.5"×11" (front and back) formula sheet, and also use a scientific calculator.

Thursday 22 February 2018

2 Two-Sample T Tests

As usual, things are somewhat more complicated if the sample size is not large. We can still use the same statistic as before,

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/m) + (S_2^2/n)}} \tag{2.1}$$

but now what we can say is, if the two populations are normal, this will obey a Student t distribution, and we have to use percentiles $t_{\alpha,\nu}$. The number of degrees of freedom ν to use is a somewhat complicated question, however. Before quoting the answer, we consider a slightly modified scenario.

2.1 Pooled T Tests (Student-t Tests)

Suppose we know that the two populations have the same variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but we still don't know what that common variance is. The statistic

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{2.2}$$

would be standard normal, but we can't construct it without the value of σ^2. An obvious estimator of σ^2 which combines all the data would be proportional to

$$\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2 = (m-1)S_1^2 + (n-1)S_2^2 \tag{2.3}$$

We know that this estimator would have expectation value

$$(m-1)\sigma^2 + (n-1)\sigma^2 = (m+n-2)\sigma^2 \tag{2.4}$$

so to get an unbiased estimator of σ^2 we would construct

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} \tag{2.5}$$

and define the "pooled t statistic"

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{2.6}$$

Given the normalization of the estimator, it shouldn't be too surprising that this statistic ends up obeying a Student t distribution with $m+n-2$ degrees of freedom.

2.2 General 2-Sample T tests (Welch-t tests)

If we don't know σ_1 and σ_2 to be the same, we return to the statistic

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/m) + (S_2^2/n)}} \tag{2.1}$$

It turns out this can be approximated as a t-distributed random variable with a number of degrees of freedom ν estimated from the data:

$$\frac{1}{\nu} \approx \frac{\frac{\mathcal{S}_1^4}{\nu_1} + \frac{\mathcal{S}_2^4}{\nu_2}}{(\mathcal{S}_1^2 + \mathcal{S}_2^2)^2} \tag{2.7}$$

where $\nu_1 = m-1$ and $\nu_2 = n-1$ are the degrees of freedom associated with the variance estimates from the first and second samples, and $\mathcal{S}_1^2 = s_1^2/m$ and $\mathcal{S}_2^2 = s_2^2/n$ are the standard errors estimated for $\overline{X}$ and $\overline{Y}$. (Devore says to round down to an integer number of df, which will make things more conservative, but is probably only necessary because there are no t tail tables for non-integer ν.)

We can note some limits here:

- If $\mathcal{S}_1 = \mathcal{S}_2$,
$$\frac{1}{\nu} = \frac{1}{4\nu_1} + \frac{1}{4\nu_2} \tag{2.8}$$
If $\nu_1 = \nu_2$ as well, which means $m = n$ and $s_1 = s_2$, this reduces to $\nu = 2\nu_1 = 2n-2$, which is the same number of degrees of freedom as in a pooled T test.
- If $\mathcal{S}_1 \gg \mathcal{S}_2$,
$$\frac{1}{\nu} \approx \frac{1}{\nu_1} \tag{2.9}$$

In general, the number of degrees of freedom will be less than in the pooled T case, $\nu \le \nu_1 + \nu_2$.

Practice Problems

9.17, 9.19, 9.33

Tuesday 27 February 2018

3 Inferences from Paired Data

So far, we've considered cases where the samples $X_1, \ldots X_m$ and $Y_1, \ldots Y_n$ were independent of each other (and could have different sample sizes). In terms of a joint distribution, this meant

$$f(x_1, \ldots, x_m, y_1, \ldots, y_m) = f_1(x_1) \cdots f_1(x_m) f_2(y_1) \cdots f_2(y_n) \tag{3.1}$$

Now we consider *paired data*, in which we have two samples–$X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$–of the same size, in which the corresponding members of the two samples need not be independent of each other. I.e., X_1 is not independent of Y_1, X_2 is not independent of Y_2, etc. We're really thinking of a sample of size n

from a *bivariate* distribution with joint pdf $f(x,y)$, so that

$$f(x_1,\dots,x_m,y_1,\dots,y_m) = f(x_1,y_1)\cdots f(x_n,y_n) \tag{3.2}$$

For example, suppose we are trying to test the hypothesis that the right arms of most people are better developed than the left arms. We could collect measurements of the bicep circumference in centimeters, but the variability in the arm sizes of different people is likely to be a lot bigger than the typical difference between left arm and right arm sizes. So rather than throwing all the left arm measurements together and all the right arm measurements together and pretending they're uncorrelated, it makes a lot more sense to keep track of the left and right arm measurements from each person. And it's a more accurate description to think about the data as being represented by correlated pairs of random variables.

Thinking about the properties of the joint distribution $f(x,y)$, we can talk about the means $\mu_1 = E(X_i)$ and $\mu_2 = E(Y_i)$, variances $\sigma_1^2 = V(X_i)$, $\sigma_2^2 = V(Y_i)$, and covariance $\rho\sigma_1\sigma_2 = \operatorname{Cov}(X_i,Y_i)$.[1] If we're interested in $\mu_1-\mu_2$, we should work with the differences of the paired random variables from the two samples, $D_i = X_i - Y_i$, which has expectation value

$$\mu_D = E(D_i) = E(X_i - Y_i) = E(X_i) - E(X_i) = \mu_1 - \mu_2 \tag{3.3}$$

and variance

$$\begin{aligned}\sigma_D^2 = V(D) &= V(X_i - Y_i) = V(X_i) + V(Y_i) - 2\operatorname{Cov}(X_i,Y_i) \\ &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\end{aligned} \tag{3.4}$$

For the inferences we're going to do, we could really just treat $D_1,\dots D_n$ as the random sample and forget about $\{X_i\}$ and $\{Y_i\}$, but it's nice to touch base with the original variables from time to time to compare the paired procedure to the old two-sample procedures. With that in mind, our statistics should be built around the sample mean

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{3.5}$$

(which happens to equal $\overline{X}-\overline{Y}$ because everything is linear) and the sample variance

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 \tag{3.6}$$

(which cannot be written in terms of S_1^2 and S_2^2 alone). The inferences are all then based on the statistic

$$T = \frac{\overline{D}-\mu_D}{\sqrt{S_D^2/n}}\ ; \tag{3.7}$$

If the distribution $f(x,y)$ is a bivariate Gaussian[2], then T will be Student-t distributed with $n-1$ degrees of freedom; if $n \gtrsim 40$, then T will be approximately standard normal distributed.

3.1 Paired vs 2-Sample Inference

Note that the paired procedure attempts to estimate

$$\operatorname{Var}(\overline{D}) = \operatorname{Var}(\overline{X}-\overline{Y}) = \frac{\sigma_D^2}{n} = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\rho\frac{\sigma_1}{\sqrt{n}}\frac{\sigma_2}{\sqrt{n}} \tag{3.8}$$

[1]Note, it's important that the index on both random variables be the same in $\operatorname{Cov}(X_i,Y_i)$. We have $\operatorname{Cov}(X_3,Y_3)=\rho\sigma_1\sigma_2$ but $\operatorname{Cov}(X_2,Y_5)=0$.

[2]The pdf $f(x,y)$ is straightforward but a little too long when written out explicitly, but the statement is equivalent to X and Y being different linear combinations of the same pair of independent Gaussian random variables.

when constructing the T statistic. If we recall the two-sample inference, in the case where $n = m$, the variance of the difference of sample means is assumed to be

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} \tag{3.9}$$

(because the construction assumes there's no correlation). So in the case of positive correlation $\rho > 0$, we're probably overestimating the standard error on the difference of the sample means if we (incorrectly) assume the data to be uncorrelated. Note that the paired procedure also ends up with a different number of degrees of freedom, $n-1$, than the two-sample one, which has

$$\begin{aligned} \nu &= (n-1)\left(\frac{(s_1^2+s_2^2)^2}{s_1^4+s_2^4}\right) = (n-1)\left(\frac{s_1^4+s_2^4+2s_1^2s_2^2}{s_1^4+s_2^4}\right) \\ &= (n-1)\left(1+\frac{2s_1^2s_2^2}{s_1^4+s_2^4}\right) \end{aligned} \tag{3.10}$$

which will be somewhere between 1 and 2 times $n-1$ depending on the relative sizes of the sample variances. So the paired test also has fewer degrees of freedom than the two-sample test, which will make e.g., confidence intervals slightly broader given the same standard error estimate, and thus tend to counteract the underestimation described above.

Practice Problems

9.37, 9.41, 9.45

Thursday 1 March 2018

4 Inferences for the Difference of Population Proportions

As in the one-sample case, one of the more interesting inferences we can consider concerns a comparison between population proportions. In this case we have two binomial random variables $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$, and would like to know something about how the proportions p_1 and p_2 compare based on the observed values of X and Y. We can also think of X and Y as the respective sums of samples drawn from independent Bernoulli distributions, which is why these single binomial random variables are the correct analogues of sample means in other cases. In fact, the respective means of the Bernoulli samples would be unbiased estimators of the two proportions: $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$ satisfy

$$E(\hat{p}_1) = \frac{E(X)}{m} = \frac{mp_1}{m} = p_1 \tag{4.1}$$

and similarly $E(\hat{p}_2) = p_2$. The variances are then

$$V(\hat{p}_1) = \frac{V(X)}{m^2} = \frac{mp_1(1-p_1)}{m^2} = \frac{p_1(1-p_1)}{m} \tag{4.2}$$

and likewise $V(\hat{p}_2) = \frac{p_2(1-p_2)}{n}$

In the interest of simplicity, we consider the case where the normal approximation applies (mp_1, $m(1-p_1)$, np_2, and $n(1-p_2)$ are all $\gtrsim 10$), and construct inferences about the difference $p_1 - p_2$. (E.g., we might want to estimate how much greater a vote share a Republican presidential candidate would have in Arizona than in New Mexico.) We can then treat the estimators $\hat{p}_1$ and $\hat{p}_2$ as approximately normal, and, again using independence of the samples, treating their difference as a normal random variable with mean $p_1 - p_2$ and variance

$$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + (-1)^2 V(\hat{p}_2) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n} \tag{4.3}$$

Thus we can construct an approximately standard normal statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}} \tag{4.4}$$

4.1 Large-Sample Confidence Interval

The most straightforward application of this statistic is as a pivot variable in constructing a coinfidence interval. The most straightforward construction shows that

$$\begin{aligned} 1 - \alpha \approx P\Bigg(&\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}} \\ &< p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}\Bigg) \end{aligned} \tag{4.5}$$

where we have defined $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. Now of course if we knew p_1 and p_2, we wouldn't have to construct a confidence interval, but we can use the estimates $\hat{p}_1 = x/m$ and $\hat{p}_2 = y/n$ in their place, and if we're in the large sample limit, the normal percentiles will still be appropriate, giving us a confidence interval with endpoints

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}} \tag{4.6}$$

4.1.1 Comment on a Comment

Devore makes the offhand comment that "recent research has shown" that better results can be obtained by replacing the estimate

$$\hat{p}_1 - \hat{p}_2 = \frac{x}{m} - \frac{y}{n} \tag{4.7}$$

with

$$\frac{x+1}{m+2} - \frac{y+1}{n+2} \tag{4.8}$$

basically pretending that each experiment had had an extra success and failure. This is actually a very old point of debate predating classical statistics itself. If you consider just a single binomial experiment leading to random variable $Y \sim \text{Bin}(n, p)$, the obvious unbiased estimator

$$\hat{p} = \frac{Y}{n} \tag{4.9}$$

has the problem that it evaluates to exactly zero if the results happened to include no successes, or one if they include no failures. But that sort of absolutism is a bad idea when predicting future outcomes, and for instance causes estimates of the standard error to go to zero. (Although $p(1-p)/n$ presumably doesn't vanish, $\hat{p}(1-\hat{p})/n$ will in this case.) Pierre-Simon Laplace suggested the $\frac{y+1}{n+2}$ estimate back in the early 19th century, and it's often known as the *Bayes-Laplace Rule of Succession*. In the Bayesian framework, it's the most likely value for p after the experiment if you assume a priori that all values are equally likely.

4.2 Large-Sample Hypothesis Tests

The need to estimate the variance of the estimator $\hat{p}_1 - \hat{p}_2$ complicates the standard approach to hypothesis testing. One could follow the same approach as in confidence interval construction, and define a test statistic corresponding to a null hypothesis Δ_0

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}}} \qquad \text{if } \Delta_0 \neq 0 \tag{4.10}$$

However, it is often the case that the null hypothesis is Δ_0 (the two proportions are the same) and in this case we can do a little better estimating the denominator of (4.4). We're interested in a statistic which is approximately standard normal when $p_1 = p_2 = p$. In that scenario, $X \sim \text{Bin}(m,p)$ and $Y \sim \text{Bin}(n,p)$ which means that $X+Y \sim \text{Bin}(m+n,p)$, i.e., we can combine the two success counts into the results of a $m+n$ binomial trials. Thus the best estimator is

$$\hat{p} = \frac{X+Y}{m+n} = \frac{m\hat{p}_1 + n\hat{p}_2}{m+n} \tag{4.11}$$

i.e., a weighted average of the two estimators. This we can use the statistic value

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} \qquad \text{if } \Delta_0 = 0 \tag{4.12}$$

This can be used to construct P-values and perform tests at a specified significance.

4.2.1 False Dismissal Probability

To get the approximate probability of a type II error (false dismissal) in the case where the null hypothesis H_0 is not satisfied, we need to specify assumed values for p_1 and p_2 (not just their difference). We have to wave our hands a little even to get the expectation value of the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} \tag{4.13}$$

We'll assume that for large samples we can replace

$$\hat{p} = \frac{m\hat{p}_1 + n\hat{p}_2}{m+n} \tag{4.14}$$

in the denominator of Z with its expectation value

$$\overline{p} = \frac{mp_1 + np_2}{m+n} \tag{4.15}$$

and thus write

$$E(Z) \approx \frac{p_1 - p_2}{\sqrt{\overline{p}\,\overline{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} \tag{4.16}$$

by the same token, we assume

$$V(Z) \approx \frac{V(\hat{p}_1) + V(\hat{p}_2)}{\overline{p}\,\overline{q}\left(\frac{1}{m}+\frac{1}{n}\right)} = \frac{\frac{p_1q_1}{m} + \frac{p_2q_2}{n}}{\overline{p}\,\overline{q}\left(\frac{1}{m}+\frac{1}{n}\right)} \tag{4.17}$$

This can then be used to estimate the false dismissal probability for a test of false alarm probability α as a function of p_1 and p_2. For instance, if we have an upper-tailed test which rejects H_0: $p_1 = p_2$ when $Z > z_\alpha$, this will fail to reject H_0 with probability

$$\begin{aligned}\beta(p_1,p_2) = P(Z \le z_\alpha) &= \Phi\left(\frac{z_\alpha - E(Z)}{\sqrt{V(Z)}}\right) \\ &= \Phi\left(\frac{z_\alpha\sqrt{\overline{p}\,\overline{q}\left(\frac{1}{m}+\frac{1}{n}\right)} - (p_1-p_2)}{\sqrt{\frac{p_1q_1}{m} + \frac{p_2q_2}{n}}}\right)\end{aligned} \tag{4.18}$$

Practice Problems

9.49, 9.55, 9.57

Tuesday 6 March 2018

5 Two-Sample F Tests

As our final class of two-sample inference, we consider the comparison of the variances of two distributions given samples drawn from each. Specifically, suppose $\{X_1, \ldots, X_m\}$ is a sample drawn from a $N(\mu_1, \sigma_1^2)$ distribution and $\{Y_1, \ldots, Y_n\}$ is a sample drawn from a $N(\mu_2, \sigma_2^2)$ distribution, where μ_1 and μ_2 are unknown but not of interest, and we want to know about σ_1^2 and σ_2^2, e.g., whether they're equal or not. Recall from our study of confidence intervals for a single population variance that

$$U_1 = \sum_{i=1}^{m} \frac{(X_i - \overline{X})^2}{\sigma_1^2} = (m-1)S_1^2/\sigma_1^2 \tag{5.1}$$

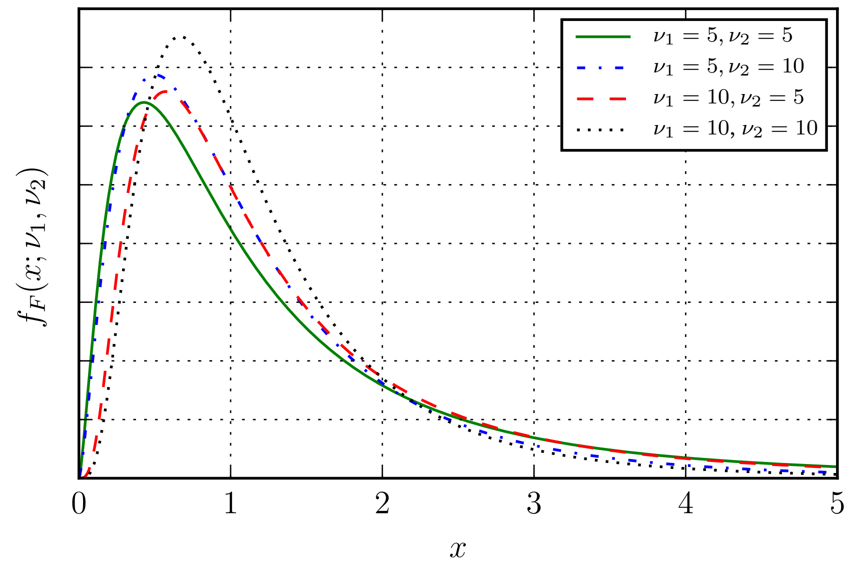
is a chi-square random variable with $\nu_1 = m - 1$ degrees of freedom, and likewise $U_2 = (n-1)S_2^2/\sigma_2^2 \sim \chi^2(\nu_2)$ where $\nu_2 = n-1$. For once, we're going to do things a bit differently and *not* try to produce an estimator for $\sigma_1^2 - \sigma_2^2$. The reason is because σ is what we call a *scale parameter* in the normal distribution, which means that it's most natural to multiply and divide by it rather than adding and subtracting. So we're more interested in inferences concerning σ_1^2/σ_2^2 (or σ_1/σ_2). It turns out that if we construct the statistic

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \tag{5.2}$$

it obeys something called an F-distribution with degree-of-freedom parameters $\nu_1 = m - 1$ and $\nu_2 = n - 1$. The PDF of the F-distribution has the form

$$f(t; \nu_1, \nu_2) \propto t^{\frac{\nu_1}{2}-1} \left(\frac{\nu_1}{\nu_2} t\right)^{-\left(\frac{\nu_1+\nu_2}{2}\right)} \tag{5.3}$$

and looks like this



but what's important is that the percentiles and cdf values of this distribution are available in statistical computing programs as well as to a limited extent in statistical tables. The percentiles are defined by the notation F_{α,ν_1,ν_2}, so that

$$P(F > F_{\alpha,\nu_1,\nu_2}; \nu_1, \nu_2) = \alpha \tag{5.4}$$

This is tabulated for assorted values of α, ν_1 and ν_2. The distribution is not symmetric, so in particular $F_{1-\alpha,\nu_1,\nu_2} \neq -F_{\alpha,\nu_1,\nu_2}$. However, if we look at the definition (5.2), we see that if F is an F-distributed random variable with ν_1 and ν_2 degrees of freedom, then $1/F$ will be an F-distributed random variable with ν_2 and ν_1 degrees of freedom. Thus

$$\begin{aligned}\alpha &= P(F < F_{1-\alpha,\nu_1,\nu_2}; \nu_1, \nu_2) = P\left(\frac{1}{F} < F_{1-\alpha,\nu_1,\nu_2}; \nu_2, \nu_1\right) \\ &= P\left(F > \frac{1}{F_{1-\alpha,\nu_1,\nu_2}}; \nu_2, \nu_1\right) = P\left(F > F_{\alpha,\nu_2,\nu_1}; \nu_2, \nu_1\right)\end{aligned} \tag{5.5}$$

Thus $F_{1-\alpha,\nu_1,\nu_2} = 1/F_{\alpha,\nu_2,\nu_1}$. So for example if we want the 5th percentile $F_{.95,5,10}$, this is $1/F_{.05,10,5} \approx 1/4.74 \approx .211$ where 4.74 comes from table A.9 in Devore. We can confirm this using the SciPy stats package:

```
In [1]: from scipy import stats

In [2]: print stats.f.isf(.95,5,10)
0.211190428782
```

So if we wish to test the null hypothesis H_0: $\sigma_1 = \sigma_2$, we can construct the test statistic

$$F = \frac{S_1^2}{S_2^2} \tag{5.6}$$

If we want to get a test of H_0 versus H_a: $\sigma_1 > \sigma_2$ with confidence level α, we reject H_0 if $s_1^2 s_2^2 > F_{\alpha,m-1,n-1}$. For instance, if $m = 6$ and $n = 11$, so $\nu_1 = 5$ and $\nu_2 = 10$ as above, we reject H_0 at the 5% level if the ratio of the sample variances is greater than $F_{.05,5,10} \approx 3.20$. On the other hand, we can reject it at the 1% level if $s_1^2 s_2^2 > F_{.01,5,10} \approx 5.32$. We could also ask, suppose the ratio of the sample variances is 4.00 (so the ratio of the sample standard deviations is 2; what is the P-value. This should be $P(F > 4.00; 5, 10)$, but the practical problem is that the cdf is not tabulated. So from the tables in Devore, all we can say is that P is between 1% and 5%. This is kind of a silly distinction, though, since after all, statistical software packages have the cdf avaiable:

```
In [3]: print stats.f.sf(4,5,10)
0.0296752952221
```

so in fact the P-value is 3.0%.

5.1 Confidence Intervals

We can also use the F distribution to set a confidence interval on the ratio of the variances, using the fact that

$$\begin{aligned}
1-\alpha &= P\left(F_{1-\alpha/2,m-1,n-1} < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2,m-1,n-1}\right) \\
&= P\left(\frac{1}{F_{\alpha/2,n-1,m-1}} < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2,m-1,n-1}\right) \\
&= P\left(\frac{1}{F_{\alpha/2,m-1,n-1}}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{\alpha/2,n-1,m-1}\frac{S_1^2}{S_2^2}\right)
\end{aligned} \tag{5.7}$$

Up Next

Note that we will skip chapters 10 and 11 on the Analysis of Variance (ANOVA) and proceed next week with chapter 12, on regression.

Practice Problems

9.59, 9.61, 9.65