Confidence Intervals

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

You should have downloaded studio11.zip and unzipped it into your 18.05 working directory.

Confidence Intervals Applet

Open the applet:

http://ocw.mit.edu/ans7870/18/18.05/s14/applets/confidence-jmo.html

1. Play around with the applet. Make sure you understand how it measures if a confidence interval is correct.

2. Read the help page.

3. What is random each time you click the 'Run N trials' button?

4. Fix the parameter settings and run many trials.

(a) Does the confidence interval contain the true mean the correct percentage of the time?

(b) What can you say about the size of the z and t-intevals over repeated trials?

5. How does increasing c change the confidence intervals? Why?

6. How does increasing n, μ or σ change the intervals? Why?

Answers on next slide.

Answers

3. The confidence interval is random. Each new set of random data produces a new confidence interval.

4(a) The percentage correct should be close to the confidence setting for both z and t confidence intervals.

4(b) The z-interval is always the same width. This is because its width is $2 * z_{\alpha/2}/\sqrt{n}$, which depends only on the fixed parameter settings.

The t-interval's width changes with each new sample of data. This is because its width also depends on the sample variance s, which is random.

5. Increasing c increases the width of the confidence intervals because widening an interval increases the probability it contains the true mean.

The answer to 6 is on the next slide.

Answers continued.

6. Increasing n decreases the size of the intervals. You can see this in the $\sqrt{n}$ in the denominator of the formulas for confidence intervals:

$$\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n} \qquad \overline{x} \pm t_{\alpha/2}s/\sqrt{n}$$

It also makes sense because more data will give a better estimate of the mean.

Increasing μ shifts the center of the intervals but does not affect their width.

Increasing σ results in wider intervals. Again, you can see this by the σ in the formula for the z-interval and the s in the formula for the t-interval. (Increasing σ will tend to increase s.) It also makes sense because a bigger σ will tend to spread out the data making the location of the mean harder to resolve.

Review: $\chi^2(df)$ confidence intervals for σ^2

- Range: $[0, \infty]$
- Parameter: df = degrees of freedom

Data: $x_1, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, where μ and σ are unknown.

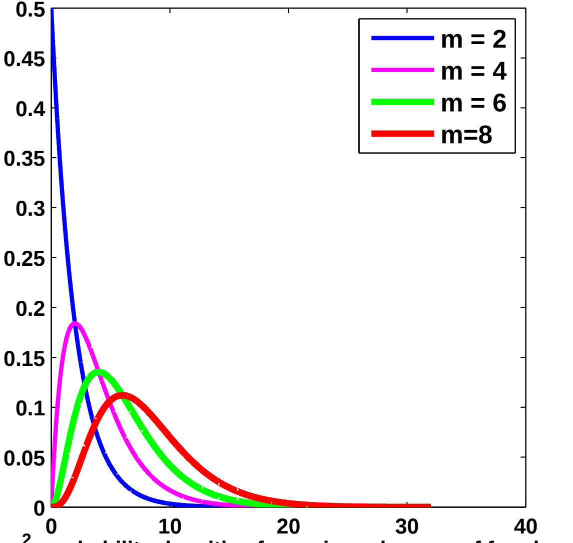
Test statistic: $r = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

$1 - \alpha$ confidence interval for σ^2:

$$\left[\frac{r}{c_{\alpha/2}}, \frac{r}{c_{1-\alpha/2}}\right],$$

$c_{\alpha/2}$ is the right-tail critical value.

χ^2



χ^2_m **probability densities for various degrees of freedom**

R Problem 1: Confidence intervals for σ^2

Write R code that:

(a) Simulates sampling 17 samples from a N(2, 3^2) distribution.

(b) Computes the 90% confidence interval for σ^2 from the sample.

See studio11-sol.r for the code.

Stock market volatility

Data file for studio: `studio11SP500data.csv`

• Contains the daily percentage change in the *Standard and Poors 500* stock index over the 14 years.

Volatility:

- Let σ^2 be the variance of the daily percentage change.
- By definition **volatility** $= \sigma$.
- High volatility implies large, fast changes in the value of the index.

Question: Is the volatility of the stock market independent of the day of the week, or are there certain weekdays when volatility tends to be higher?

R Problem 2: Stock market volatility

1. Use the code in `studio11.r` to load the percentage change data for Mondays and Fridays.

(This code also does a little data exploration using plots and a table.)

2. Let σ_M^2 be the true variance of the percent returns on Mondays. Likewise σ_F^2 for Fridays.

3. Use `?var.test` to learn about the function `var.test()`

4. Use `var.test()` to compute a 95% confidence interval for the ratio of the variances. Use the result to decide if one of Mondays or Fridays is more volatile than the other.

Answers

The function `var.test()` performs an *F*-test to compare the variances of two normal distributions.

In `studio11-sol.r` the command

```
        var.test(pctReturn.monday, pctReturn.friday,
                 alternative="two-sided")
```

gave the ratio of the sample variances $s_M^2/s_F^2 = 1.59$, with a confidence interval [1.37, 1.85].

Since the 95% confidence interval is strictly above 1, we conclude that $\sigma_M > \sigma_F$, i.e. that Mondays are more volatile than Fridays.

Understanding `var.test()`

Notation: $F(\text{df1}, \text{df2}) = F$ distribution with (df1, df2) degrees of freedom.

Theorem. If $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are independent samples from normal distributions with the **same variance** then the ratio of sample variances follows an F distribution:

$$F = \frac{\text{var}(x_i)}{\text{var}(y_j)} \sim F(n-1, m-1).$$

- Now assume that the normal distributions have **different variances**, σ_x^2, σ_y^2.

Problem: (a) Use the F statistic, critical values of the F distribution and the theorem to determine the $1-\alpha$ confidence interval for the ratio of variances σ_x^2/σ_y^2.

(b) Code your answer in R and show you get the same results as we did using `var.test(x, y)`.

Solution

Standardization. The key is the same as for previous confidence intervals: we need a standardized statistic that follows a known distribution. Here it is:

$$r = \frac{s_x^2/s_y^2}{\sigma_x^2/\sigma_y^2} \sim F(n-1, m-1)$$

We will show this below. Let's first use r to find the $1-\alpha$ confidence interval for σ_x^2/σ_y^2. All it takes is a bit of algebra.

Let $c_{\alpha/2}$ be the right $\alpha/2$ critical value for $F(n-1, m-1)$. Since r follows an $F(n-1, m-1)$ distribution, we have

$$P(c_{1-\alpha/2} < r < c_{\alpha/2}\,|\sigma_y, \sigma_y) = 1-\alpha$$

(As usual, we emphasize that σ_x and σ_y are fixed, not random, values by explicitly making the probability conditional on them.)

The following sequence of algebraic manipulations leads to the confidence interval.

Solution continued

Substitute the formula for r:

$$P\left(c_{1-\alpha/2} < \frac{s_x^2/s_y^2}{\sigma_x^2/\sigma_y^2} < c_{\alpha/2} \,|\, \sigma_x, \sigma_y\right) = 1-\alpha.$$

Do some algebraic manipulation:

$$P\left(\frac{s_x^2/s_y^2}{c_{\alpha/2}} < \sigma_x^2/\sigma_y^2 < \frac{s_x^2/s_y^2}{c_{1-\alpha/2}} \,|\, \sigma_x, \sigma_y\right) = 1-\alpha.$$

Use the definition of the F-statistic $F = s_x^2/s_y^2$:

$$P\left(\frac{F}{c_{\alpha/2}} < \sigma_x^2/\sigma_y^2 < \frac{F}{c_{1-\alpha/2}} \,|\, \sigma_x, \sigma_y\right) = 1-\alpha.$$

This give us the $1-\alpha$ confidence interval for σ_x^2/σ_y^2:

$$\left[\frac{F}{c_{\alpha/2}},\ \frac{F}{c_{1-\alpha/2}}\right]$$

Solution continued

All that's left is to show that the standardized statistic r follows an $F(n-1, m-1)$ distribution.

Since $x_i \sim \mathrm{N}(\mu_x, \sigma_x^2)$ and $y_j \sim \mathrm{N}(\mu_y, \sigma_y^2)$, we know that

$$\frac{x_i}{\sigma_x} \sim \mathrm{N}\left(\frac{\mu_x}{\sigma_x}, 1\right) \qquad \text{and} \qquad \frac{y_i}{\sigma_y} \sim \mathrm{N}\left(\frac{\mu_y}{\sigma_y}, 1\right).$$

Since x_i/σ_x and y_j/σ_y have the same variance, i.e. 1, the above theorem says that

$$\frac{\mathrm{var}(x_i/\sigma_x)}{\mathrm{var}(y_i/\sigma_y)} \sim F(n-1, m-1).$$

By the linearity rules for variance we know that $\mathrm{var}(x/\sigma) = \mathrm{var}(x)/\sigma^2 = s_x^2/\sigma_x^2$. Therefore

$$\frac{\mathrm{var}(x_i/\sigma_x)}{\mathrm{var}(y_i/\sigma_y)} = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = \frac{s_x^2/s_y^2}{\sigma_x^2/\sigma_y^2} = r. \qquad \text{QED}$$