Module E: Advanced Inference

Slide Deck E3:

Handling One and Two Variances

The section in which we see how to perform hypothesis tests concerning population variances using the R Statistical Environment. This deck will focus not only the code used to perform the analysis, it will also emphasize the analysis process.

Today's Objectives

By the end of this slidedeck, you should

1. understand the theory behind, and test hypotheses about:
   - a single population variance
   - the *ratio* of two population variances
2. better understand the p-value and how to test hypotheses
3. clearly specify how confidence intervals and p-values both give important information about the population parameter

Note that we are moving beyond the general theory of confidence intervals and hypothesis testing. We are looking at how to specifically perform the procedures. It all comes down to the population parameter you are trying to learn about.

One-Parameter Procedures: σ^2

Parametric Procedure: Chi-Square Procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Data generated from Normal distribution
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x)`
- R function: `onevar.test(x)`

One-Parameter Procedures: σ^2

Non-parametric Procedure: Non-Parametric Bootstrap procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Nothing
- R code:

```
st = numeric()
for(i in 1:1e4) {
  x = sample(y, replace=TRUE)
  st[i] = var(x)
}
quantile(st, c(0.025,0.975))
```

The Theory of the Variance

Let us start by assuming the data are generated by a Normal process. That is,

$$X \sim \mathcal{N}\left(\mu;\ \sigma^2\right)$$

Let us define the sample variance as we have in the past

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

It can be shown that

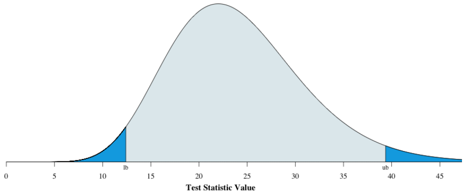
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_\nu$$

Here, the number of degrees of freedom is $\nu = n - 1$.

- With this, we have everything we need to calculate the endpoints of the confidence interval and the p-values.

The Theory of the Variance

The χ^2_{24} distribution with the middle 95% shown.



The Theory of the Variance: Confidence Interval

Recall:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_\nu$$

To calculate the endpoints, we proceed as usual. Define L as the 2.5th quantile. Algebra gives the upper confidence limit:

$$\frac{(n-1)S^2}{\sigma^2} = L$$

$$\frac{(n-1)S^2}{L} = \sigma^2$$

Thus, the upper confidence limit is

$$\frac{(n-1)S^2}{L}$$

The Theory of the Variance: Confidence Interval

Similarly, defining U as the 97.5th quantile gives

$$\frac{(n-1)S^2}{U}$$

as the lower confidence limit.

Putting these together provides the two endpoints:

$$\left(\frac{(n-1)S^2}{U},\ \frac{(n-1)S^2}{L} \right)$$

The Theory of the Variance: p-value

We can also use our definition of p-value to calculate them when testing hypotheses about the population variance, σ^2.

Since we know the distribution of the test statistic, we use that to calculate the p-value.

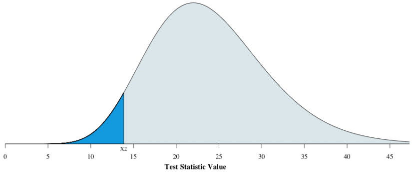
$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_\nu$$

The test statistic is the quantity on the left:

- σ_0^2 is the claimed value
- S^2 is the sample variance
- n is the sample size

The Theory of the Variance: p-value

The χ^2_{24} distribution with the observed value of $X2 = 13.35$ shown.



The p-value (shaded area) is 0.04 for the alternative hypothesis using $<$.

Two-Parameter Procedures: σ_1^2/σ_2^2

Parametric Procedure: Fisher's F-test

- Graphic: Side-by-side box-and-whiskers plot
  `boxplot(x ~ g)`
  `boxplot(x1, x2)`
- Requires: Data generated from Normal distribution — in *each* population
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x ~ g)`
- R function: `var.test(x ~ g)`
- R function: `var.test(x1, x2)`

Example 1: The Risks of IBM

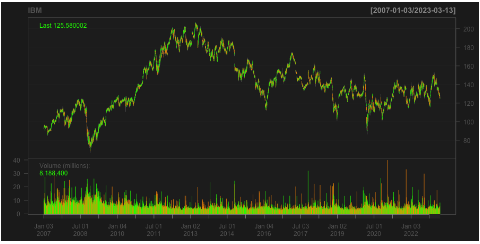
Example

The variance of a stock is frequently used to indicate its level of risk. Higher variances indicate higher risks for the stock. This comes from the idea that it is important to be able to predict the future value of a stock (low volatility).

Estimate the risk of IBM between January 3, 2007, and March 14, 2023.

While the Shapiro-Wilk test concludes that the data are not from a Normal distribution (p-value $\ll 0.0001$), the sample size is large enough ($n = 4076$) so that the Central Limit Theorem ensures that the sample variances are approximately chi-squared (via Slutsky's Theorem).

Example 1: The Risk of IBM



Example 1: The Risks of IBM

This code

```
onevar.test(IBMvals, s2=750)
```

results in this output

```
  One-Sample Variance Test

data:  IBMvals
X2 = 4328.5, df = 4075, p-value = 0.005804
alternative hypothesis: true variance is not equal to 750
95 percent confidence interval:
 763.1571 832.4027
sample estimates:
 variance of IBMvals
            796.6475
```

Example 1: The Risks of IBM

If you are not comfortable relying on the Central Limit Theorem, we can use bootstrapping:

```
ts = numeric()
for(i in 1:1000) {
  xx = sample(IBMvals, replace=TRUE)
  ts[i] = var(xx)
}

quantile(ts, c(0.025,0.975))
```

This code results in this output:

```
    2.5%       97.5%
767.3672    827.5502
```

What can one conclude from this?

Example 1: The Risks of IBM

Conclusion: We would like to estimate the volatility of IBM using its stock prices between January 3, 2007, and March 13, 2023. While the Shapiro-Wilk test indicates that the data were not generated from a Normal process, the sample size of 4076 suggests that the sample variances closely follow a chi-square distribution.

According to the Chi-square variance test, we are 95% confident that the variance of IBM is between 763 and 832. To support this estimate, the bootstrap suggests that the 95% confidence interval is between 767 and 828.

Example 2: Microsoft vs. Apple

Example

I have saved up some money to invest in the stock market. I would like to invest it in either Apple or Microsoft. I will choose the one that has lower risk.

While the Shapiro-Wilk test concludes that neither data are from a Normal distribution (p-value $\ll 0.0001$), the sample sizes are large enough ($n = 4076$) so that the Central Limit Theorem ensures that the sample variances are approximately chi-squared (via Slutsky's Theorem).

Example 2: Microsoft vs. Apple

Example 2: Microsoft vs. Apple

Example 2: Microsoft vs. Apple

This code

```
var.test(MSFTvals, AAPLvals)
```

results in this output

```
  F test to compare two variances

data:  MSFTvals and AAPLvals
F = 3.2842, num df = 4076, denom df = 4076, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.088639 3.492248
sample estimates:
ratio of variances
          3.284249
```

Example 2: Microsoft vs. Apple

If you are not comfortable relying on the Central Limit Theorem, we can use bootstrapping:

```
ts = numeric()
for(i in 1:1000) {
  xx = sample(MSFTvals, replace=TRUE)
  yy = sample(AAPLvals, replace=TRUE)
  ts[i] = var(xx) / var(yy)
}

quantile(ts, c(0.025,0.975))
```

This code results in this output:

```
    2.5%       97.5%
3.067615    3.535548
```

What can one conclude from this?

Example 2: Microsoft vs. Apple

Conclusion: We would like to determine which of the two stock, Microsoft and Apple, are more volatile— ad by how much. While the Shapiro-Wilk test indicates neither set of stock prices arose from a Normal process, the sample size is sufficient to allow us to use Fisher's F-test.

According to the F-test, we are 95% confident that the variance of Microsoft is between 3.1 and 3.5 times greater than that of Apple. To support this estimate, the bootstrap *also* suggests that the 95% confidence interval is between 3.1 and 3.5.

Thus, I will invest my savings in Apple Computers, because it has a lower risk (volatility) than does Microsoft— by a factor of more than 3!

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. understand the theory behind, and test hypotheses about:
   - a single population variance
   - the *ratio* of two population variances
2. better understand the p-value and how to test hypotheses
3. clearly specify how confidence intervals and p-values both give important information about the population parameter

Today's R Functions

Here is what we used the following `R` functions:

- `shapiroTest(x)` performs the Shapiro-Wilk test for Normality
- `onevar.test(x, mu)` performs the one-sample t-test
- `var.test(x, y)` performs the two-sample t-test

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning population means.

- SCA 9a
- SCA 9b
- SCA 13
- SCA 23

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCAs, there are **Laboratory Activity E** (confidence intervals) and **Laboratory Activity F** (hypothesis testing).

Source: `https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of the material covered in this slidedeck:

- Hawkes Learning: Chapters 10 and 11
- Intro to Modern Statistics: None
- R for Starters: Chapters 5 and 6
- Wikipedia: Confidence Intervals; Hypothesis Testing

Please do not forget to use the `allProcedures` document that lists all of the statistical procedures we will use in R.