Topic 2.3: basics of statistical inference

The F Distribution

The sampling distribution of sample variance

Sonja Petrović
Created for ITMD/ITMS/STAT 514

Spring 2021.

Sample variance

Motivating question

A manufacturer of car batteries *guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year*. If five of these batteries have lifetimes of $1.9, 2.4, 3.0, 3.5,$ and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

Motivating question

A manufacturer of car batteries *guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year*. If five of these batteries have lifetimes of $1.9, 2.4, 3.0, 3.5,$ and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

Assume that the battery lifetime follows a normal distribution.

... and?? $S^2 \sim \ldots.??$

Motivating question

A manufacturer of car batteries *guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year*. If five of these batteries have lifetimes of $1.9, 2.4, 3.0, 3.5,$ and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

Assume that the battery lifetime follows a normal distribution.

... and?? $S^2 \sim \ldots .??$

Aha!

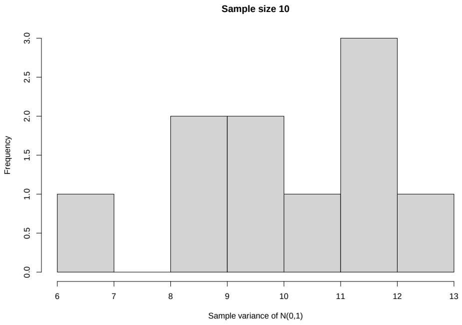
we need to think this through:

- Is the mean of S^2 around σ^2, at least?
- Is $S^2 \sim N(?, ?)$?
- How do we figure out the actual sampling distribution of the random variable S^2?

Discovering the sampling distribution of S^2

Let's make a conjecture using some simulated data.

```
sample.var <- replicate(10, 10* var(rnorm(n=100,mean=mean(c(1.
hist(sample.var, main=paste("Sample size 10"),
     xlab="Sample variance of N(0,1)")
```

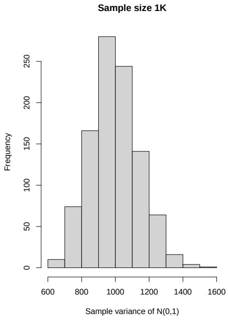


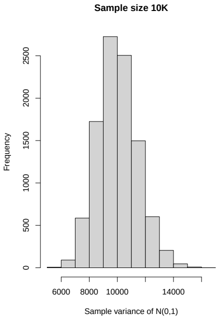
Discovering the sampling distribution of S^2

... was that enough data?!

Discovering the sampling distribution of S^2

... was that enough data?!





Sampling distribution of S^2

Theorem

If S^2 is the variance of a random sample of size n taken from a normal population having the variance $|sigma^2$, then the statistic

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

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

Sampling distribution of S^2

Theorem

If S^2 is the variance of a random sample of size n taken from a normal population having the variance $|sigma^2$, then the statistic

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

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

- The values of the random variable χ^2 are calculated from each sample by the formula

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

Sampling distribution of S^2

Theorem

If S^2 is the variance of a random sample of size n taken from a normal population having the variance $|sigma^2$, then the statistic

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

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

- The values of the random variable χ^2 are calculated from each sample by the formula

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

  Discuss probabilities.

- What does a χ^2 distribution look like?

Back to the example

If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

```
s2 = var(c(1.9, 2.4, 3.0, 3.5,  4.2))
4*s2/1
```

```
[1] 3.26
```

... and?

Back to the example

If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

```
s2 = var(c(1.9, 2.4, 3.0, 3.5,  4.2))
4*s2/1
```

```
[1] 3.26
```

. . . and? What is the probability of seeing a value of '3.26' under the chi-square distribution with 4 degrees of freedom?

Back to the example

If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

```
s2 = var(c(1.9, 2.4, 3.0, 3.5,  4.2))
4*s2/1
```

```
[1] 3.26
```

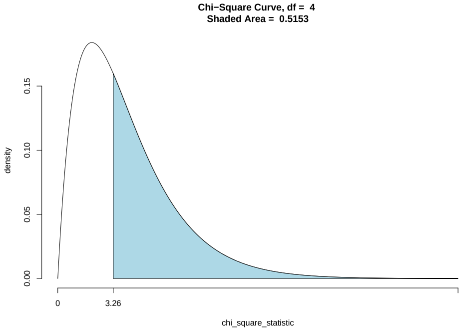
... and? What is the probability of seeing a value of '3.26' under the chi-square distribution with 4 degrees of freedom?

```
1- pchisq(3.26,df=4)
```

```
[1] 0.5152948
```

Back to the example

```
pchisqGC(3.26,region="above",df=4,
         xlab="chi_square_statistic",graph=TRUE)
```




```
[1] 0.5152948
```

Aha!

discuss meaning: what do these values encode? Is ‘3.26‘ expected?

The F-distribution

Comparison of variability in two populations

Theorem

IF S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

- Discuss: use of the F-distribution, follow-up to Case Study 8.2 (paint drying time).
- Heads-up: this distribution is used in **analysis of variance**, a topic we'll cover soon.

What's next?

Remember that probability calculations for the sample variance rely heavily on the assumption of normality. If the data distribution is not normal, then these probabilities may be way off.

- We will learn about some heuristic tests for normality of the data distribution.

Interlude

[Time permitting . . . Let's talk about importing and selecting from another large dataset with which we may work.]

Appendix

A worksheet on sampling variance

Let us look at more simulations for variances.[1]

We will simulate values of $V^2 := \frac{(n-1)S^2}{\sigma^2}$ from normal data. Assume that the underlying distribution X is distributed as $X \sim N(0, 9)$ and suppose that the sample size, n, is 6.

[1] Examples extracted from sections 5.7.2 and 5.7.2. in this book appendix.

A worksheet on sampling variance

Let us look at more simulations for variances.[1]

We will simulate values of $V^2 := \frac{(n-1)S^2}{\sigma^2}$ from normal data. Assume that the underlying distribution X is distributed as $X \sim N(0, 9)$ and suppose that the sample size, n, is 6.

Step 1. Generate an object named `draws` with 6 rows and 1000 columns of normal observations where the normal observation has mean 0 and standard deviation 3.

```
draws = matrix(rnorm(1000 * 6, 0, 3), 6)
```

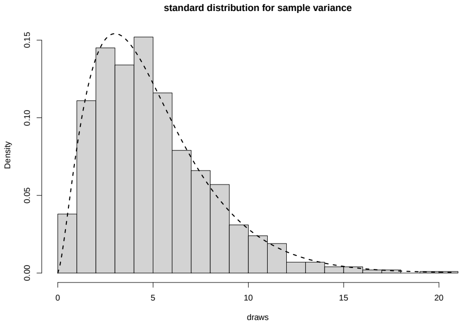
The next line applies the var command to each column using the apply command to create the 1000 values of S^2.

```
drawvar = apply(draws, 2, var)
```

[1] Examples extracted from sections 5.7.2 and 5.7.2. in this book appendix.

Step 2. Present the histogram for these 1000 values of $V^2 := \frac{(n-1)S^2}{\sigma^2}$.

```
draws = 5 * drawvar/9
hist(draws, breaks = 20, prob = TRUE,
     main = "standard distribution for sample variance")
v = seq(0, max(draws), length = 200)
lines(v, dchisq(v, 5), lty = 2, lwd = 2)
```



Not surprisingly, the shape of this simulated distribution is very close to the shape of the theoretical distribution for χ^2 with 5df (overlaid as a dashed lines here by the last two command lines of the code).

Computing Probabilities for the Variance

Suppose you have a sample of size 18 from a population mean of 30 cm and a population variance of 90. What is the probability that S^2 will be less than 160?

```
n = 18
pop.var = 90
value = 160
pchisq((n - 1) * value/pop.var, n - 1)
```

```
[1] 0.9752137
```

Notice where the sample size (n = 18), population variance (pop.var = 90) and value of interest (value = 160) appear in the pchisq command.

As with other probability commands, the upper tail could have been calculated using the option `lower.tail=FALSE`.

Now consider another exmaple about a fruit company, with data about weight of apple sauce in grams having distribution $X \sim N(275, 0.0016)$. Here we want to take a random sample of 9 jars and find the value s^2 so that $Prob(S^2 \leq s^2) = 0.99$.

```
pop.var = 0.0016
n = 9
prob = 0.99
pop.var * qchisq(prob, n - 1)/(n - 1)
```

```
[1] 0.004018047
```

Again notice where the sample size ($n = 9$), probability level ($prob = 0.99$) and population variance ($pop.var = 0.0016$) appear in the calculation. [Why do the variance and sample size appear outside of the command qchisq?]

License

This document is created for ITMD/ITMS/STAT 514, Spring 2021, at Illinois Tech.

While the course materials are generally not to be distributed outside the course without permission of the instructor, all materials posted on this page are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Sonme of the applied examples and case studies in these notes (Topic 2 in general) are taken from one of our reference textbooks.