STAT 515 Lec 11

Variance estimation

Karl Gregory

Estimating σ^2 from the sample

Suppose $X_1, \dots, X_n$ are a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown. We consider estimating σ^2 and building a confidence interval for it.

Our estimator of σ^2 is the sample quantity

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

We know that S_n^2 will take a different value every time we draw a sample; if we were to repeat our experiment many times, we would get many different values of S_n^2. Our question is what the distribution of these values would look like.

In building a confidence interval for the mean μ, we began with the assumption

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1),$$

which is satisfied if the population distribution is Normal or approximately satisfied if the sample size is large. This allowed us to write

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

which we could rearrange to get

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

This gave us the $(1 - \alpha)100\%$ confidence interval for μ defined by

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

We will follow similar steps in order to construct a confidence interval for σ^2 based on S_n^2.

Sampling distribution of S_n^2

We need to know the sampling distribution of S_n^2, so that we can write a probability statement involving S_n^2 and the unknown σ^2 which we can rearrange to construct a confidence interval. We will use the following result on the sampling distribution of S_n^2.

Sampling distribution result: Sampling distribution result for S_n^2.

Let $X_1, \ldots, X_n$ be a random sample from a Normal population with variance σ^2 and let $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Then

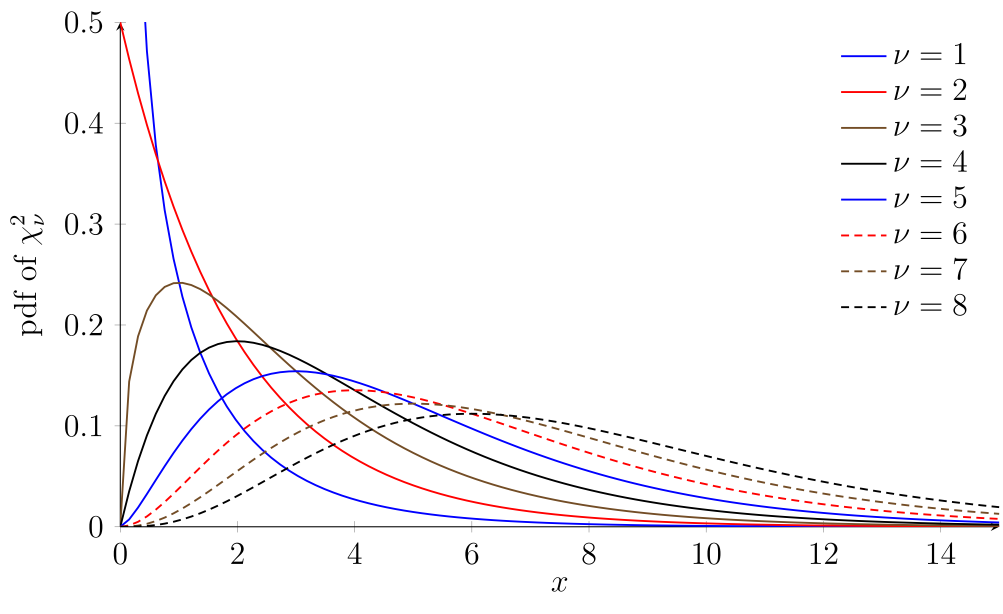
$$\frac{(n-1)S_n^2}{\sigma^2} \quad \text{has the chi-squared distribution with degrees of freedom } n-1.$$

If a random variable W has the chi-squared distribution with degrees of freedom ν, then we write $W \sim \chi^2_\nu$.

What does the chi-squared distribution look like? There is a chi-squared distribution for every positive whole number $\nu = 1, 2, 3, \ldots$, and the whole number with which a chi-squared distribution is associated is called its *degrees of freedom*. The chi-squared distributions are all right-skewed distributions. The probability density function of the χ^2_ν distribution is given by

$$f(x) = \frac{1}{\Gamma(\nu/2)2^{\nu/2}} x^{\nu/2-1} \exp\left(-\frac{x}{2}\right), \quad x > 0,$$

where $\Gamma(z) = \int_0^\infty u^{z-1} e^{-z} dz$ for $z > 0$. The pdfs of the chi-squared distributions with degrees of freedom $\nu = 1, \ldots, 8$ are plotted here:

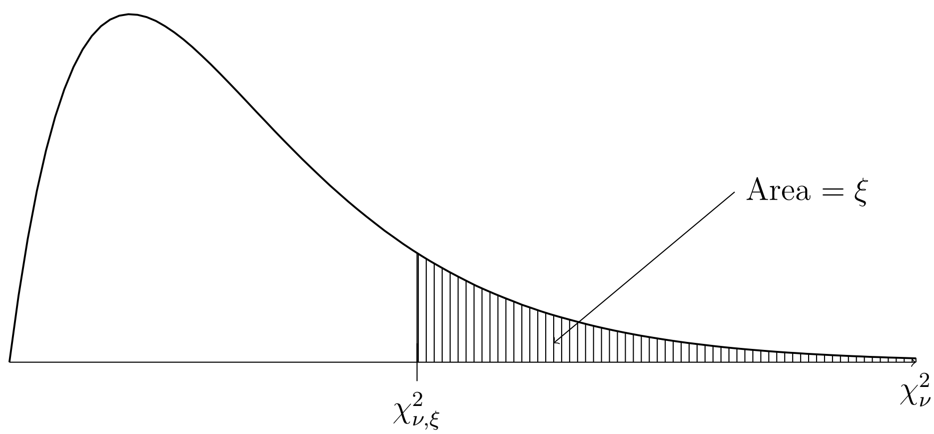


Confidence interval for σ^2

In order to give an expression for a $(1-\alpha)100\%$ confidence interval for σ^2, we define, for any number $0 < \xi < 1$, the quantity $\chi^2_{\nu,\xi}$ to be the value such that

$$P(W > \chi^2_{\nu,\xi}) = \xi,$$

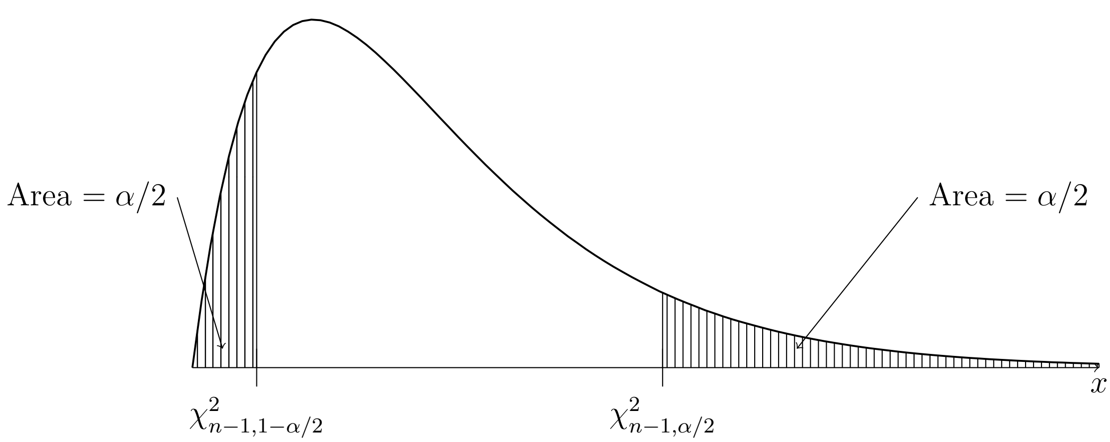
where W is a random variable having the chi-squared distribution with degrees of freedom equal to ν. The value $\chi^2_{\nu,\xi}$ thus admits the depiction



Now we may write the probability statement

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

which corresponds to the picture



We can rearrange the previous probability statement to leave σ^2 in the middle:

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

Thus, a $(1-\alpha)100\%$ confidence interval for σ^2 is given by

$$\left(\frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\right).$$

Note that the interval is not "symmetric" around the estimator S_n^2, that is, it is not of the form $S_n^2 \pm$ something. This is because the sampling distribution of S_n^2 is not symmetric.

Loblolly pine trees example

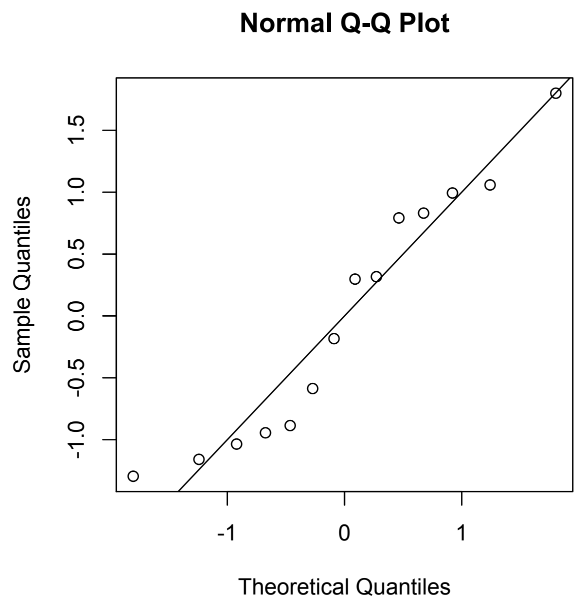
Exercise. Using the data set `Loblolly` in R, which one can access by entering `data(Loblolly)` into the console, build a 95% confidence interval for the variance σ^2 of the height of Loblolly pines which are ten years old.

Answer: Execute the command

```
x <- Loblolly$height[Loblolly$age==10]
```

in R. This stores the desired values in the vector `x`. We can compute S_n^2 by typing `var(x)`, which gives $S_n^2 = 2.365095$.

To make sure the data are Normally distributed (which is necessary in order to construct a confidence interval based on a chi-squared distribution), we make a Normal QQ plot with the commands `qqnorm(scale(x))` and `abline(0,1)`. This produces the plot



The points in the plot deviate somewhat from a straight line, but it seems pretty safe to assume that the data have come from a Normal distribution.

The sample size is $n = 14$, which we can get by entering `length(x)` into the console. The relevant chi-squared distribution is thus the chi-squared distribution with degrees of freedom equal to $14 - 1 = 13$. We can retrieve quantiles of the chi-squared distributions using the `qchisq()` function in R or by consulting the tables on pages 818 and 819 of the textbook. We find

$$\chi^2_{13,.975} = \texttt{qchisq(.025,13)} = 5.00874 \quad \text{and} \quad \chi^2_{13,.025} = \texttt{qchisq(.975,13)} = 24.7356.$$

A 95% confidence interval for σ^2 is thus given by

$$\left(\frac{(14-1)2.365095}{24.7356}, \frac{(14-1)2.365095}{5.00874}\right) = (1.242995, 6.138517).$$

Where do the chi-squared distributions come from?

Let $Z_1, \dots, Z_n$ be a random sample from the $Z \sim \text{Normal}(0, 1)$ distribution. If we define

$$W_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2,$$

we find that $W_n \sim \chi^2_n$. We can write

$$\begin{aligned} \frac{(n-1)S_n^2}{\sigma^2} &= \frac{n-1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \\ &= \left(\frac{X_1 - \bar{X}}{\sigma}\right)^2 + \left(\frac{X_2 - \bar{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \bar{X}}{\sigma}\right)^2, \end{aligned}$$

which looks a lot like a sum of Z values, just with μ replaced by $\bar{X}$. A theorem called Cochran's theorem can be used to conclude that the effect of having $\bar{X}$ instead of μ is a reduction in the degrees of freedom by 1. So

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