**Chi-square probability distribution**

The chi-square probability distribution is another frequently used probability distribution in statistics. We will use it when looking at the probability distribution for S2.

The chi-square PDF for a continuous random variable Y is



where ν > 0.

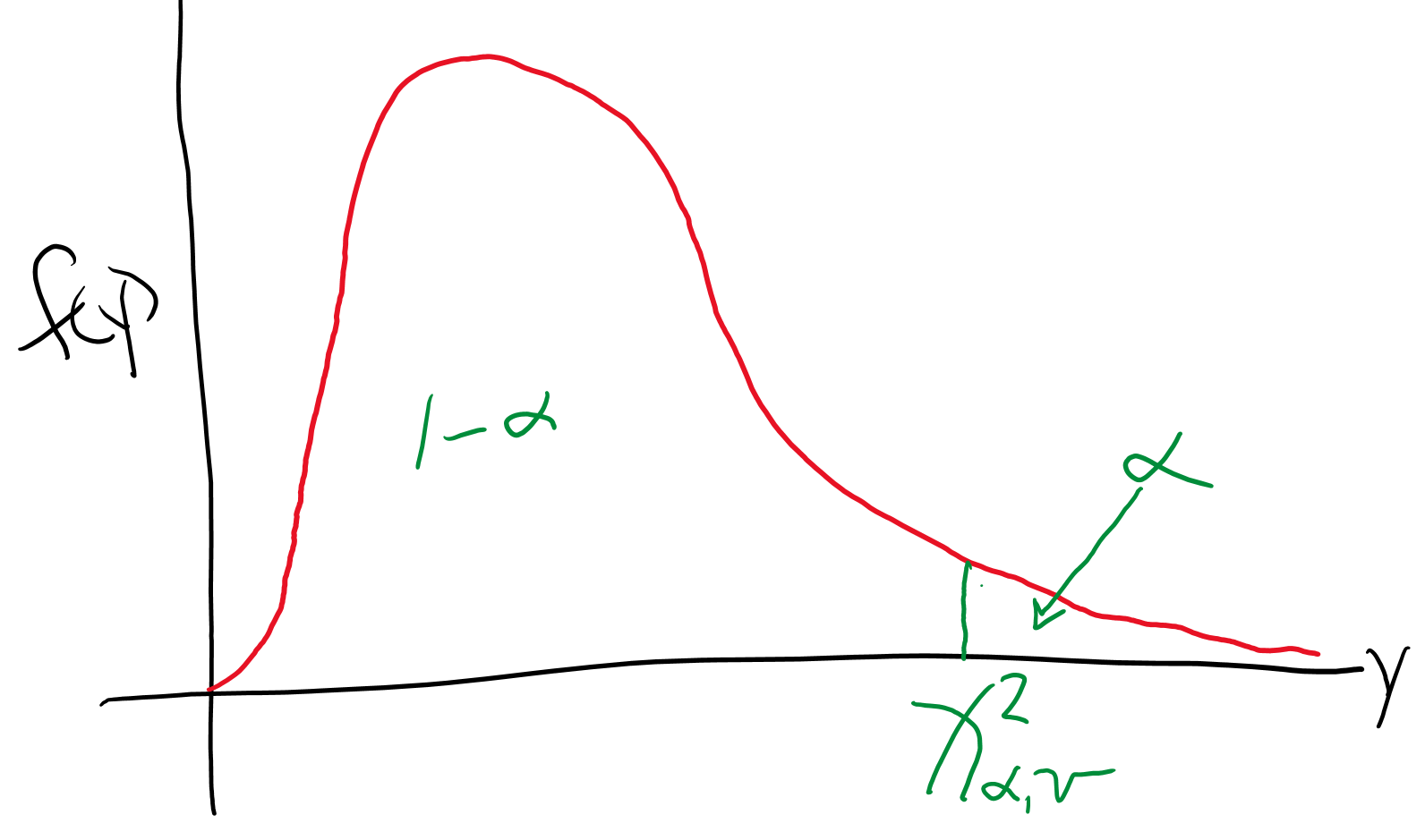
The mean and variance of a random variable with this distribution are

E(Y) = μ = ν and Var(Y) = σ2 = 2ν

Notes:

* ν is a parameter. Different shapes of the probability distribution result from different values of ν.
* ν has a special name: degrees of freedom. A number of PDFs will use this term with a parameter. We will discuss degrees of freedom more later.
* The chi-square PDF is a gamma PDF with α = ν/2 and β = 2. Therefore, one can use this result to show E(Y) and Var(Y).
* The name “chi-square” could equivalently be expressed as χ2.

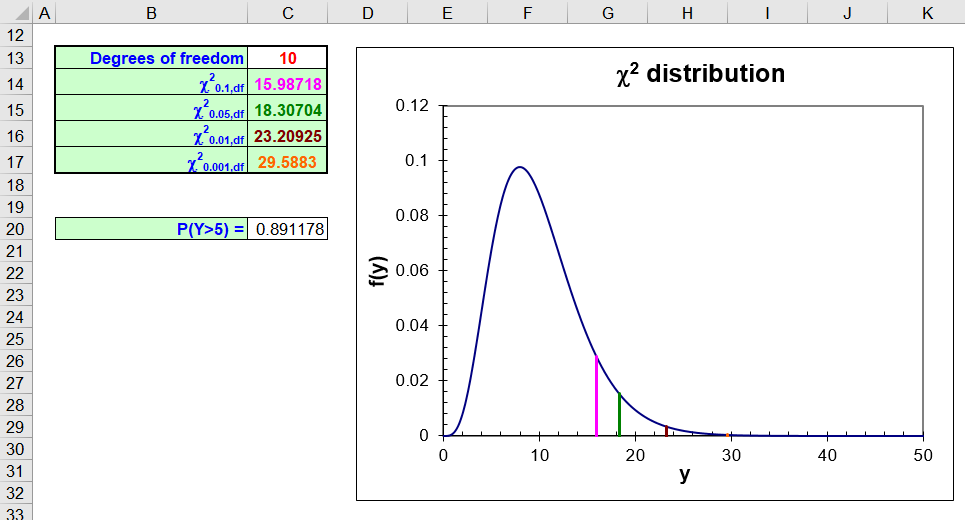
Notation:  denotes the 1 – α quantile from a chi-square distribution with ν degrees of freedom.



Then P(Y < ) = 1 – α.

Example: Chi-square probability distribution plot (chi\_square\_dist.xlsx)

This is an interactive file which allows you to see the probability distribution for different degrees of freedom.



Example: Finding probabilities and quantiles from a chi-square distribution (chi\_square\_dist.R)

To find P(Y < 3.84) with ν = 1, we could use integration:



Instead, it is easier to use the pchisq() function:

> pchisq(q = 3.84, df = 1)

[1] 0.94996

To find the 1 – α quantile from a chi-square distribution, we could use integration:



where we would solve for c in the above equation. Instead, we will use the qchisq() function:

> alpha <- 0.05

> qchisq(p = 1 - alpha, df = 1)

[1] 3.8415

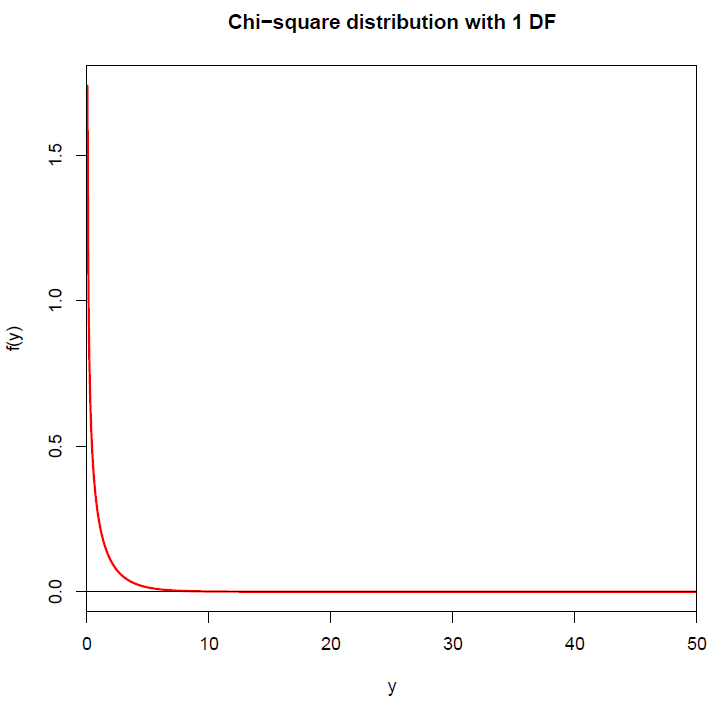
The dchisq() function allows us to evaluate f(y) so that we can plot the distribution:

> curve(expr = dchisq(x = x, df = 1), xlim = c(0,5), col

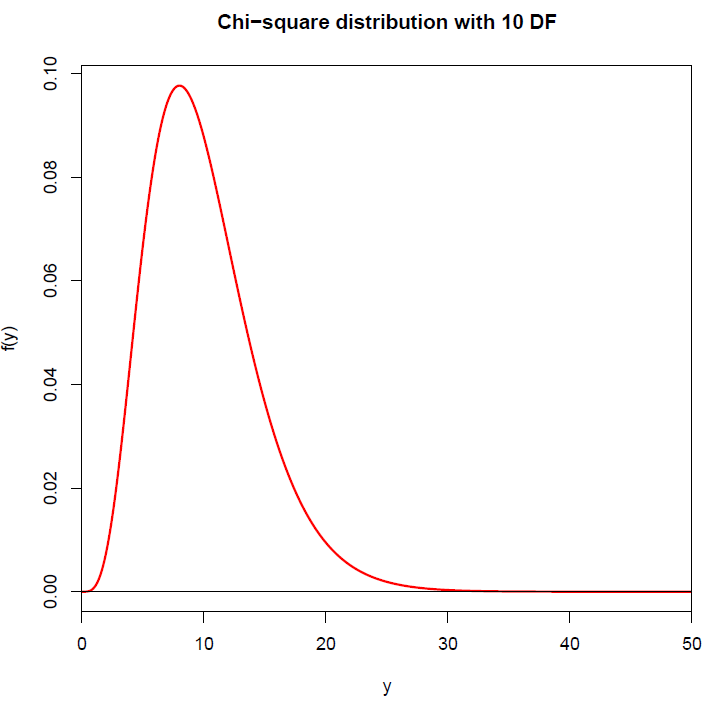
= "red", lwd = 2, main = "Chi-square distribution with

1 DF", ylab = "f(x)", xlab = "y", n = 1000)

> abline(h = 0)



Obviously, the distribution is quite skewed for ν = 1 degree of freedom. Below is another plot of the distribution, but with ν = 10 degrees of freedom:



Some introduction to statistics books provide probabilities corresponding to a particular degrees of freedom in a table format. We will not use a table in this course.

Probability distribution involving S2

In addition to obtaining the probability distribution for , we may want a probability distribution for S2. In order to do this, we need to make the assumption that Y1, Y2, …, Yn are a random sample from a population characterized by a normal probability distribution with E(Yi) = μ and Var(Yi) = σ2. With this assumption, one can show that



has a chi-square probability distribution with ν = n – 1 degrees of freedom.

Proof: If interested, please see Casella and Berger (2002, p. 218) or Wackerly et al. (2008, p. 358).

Notes:

* What does this result mean?
* What are the limitations of this result in comparison to the central limit theorem?

Unbiased estimator

An unbiased estimator is a statistic with an expected value equal to the corresponding parameter it is estimating.

Suppose T is a statistic and θ is a parameter. T is an unbiased estimator for T if E(T) = θ.

We saw previously that E() = μ, so  is an unbiased estimator of μ.

Using unbiased estimator very desirable!

S2 is an unbiased estimator for σ2

proof:

Remember that for two independent random variables X and Y and constants a and b,

* E(aX + bY) = aE(X) + bE(Y)
* Var(Y) = E(Y2) – [E(Y)]2; this means that E(Y2) = Var(Y) + [E(Y)]2 = Var(Y) + μ2
* E() = μ
* Var() = σ2/n and Var() = E(2) – [E()]2; this means that E(2) = σ2/n + [E()]2 = σ2/n + μ2

Suppose Y1, …, Yn are random variables with E(Yi) = μ and Var(Yi) = σ2. Then











Therefore, E(S2) = σ2.

This is why we divide  by n – 1 rather than n!

What are degrees of freedom?

Suppose the sum of three numbers is 6. In order to know all of the three numbers, you only need to know 2 of them and the sum. For example,

X1 = 1 (pick)

X2 = 2 (pick)  
X3 = **3** (cannot Vary)  
Sum = 6

The degrees of freedom are 2 for the sum. Similarly, the degrees of freedom are n - 1 for  and S2. Many probability distributions build these values into their mathematical functions as parameters. Typically, the degrees of freedom will be known because we know the sample size. Thus, these probability distributions can be easier to work with.