How do I get P-values and critical values from R?

2017-03-01

Calculating P-values

We can calculate P P-values in R by using *cumulative distribution functions* and *inverse cumulative distribution functions (quantile function)* of the known sampling distribution.

Cumulative distribution function (CDF)

If the probability density function is given by $f(x)$, then the *cumulative distribution function* is given by $F(x)$ and is defined by

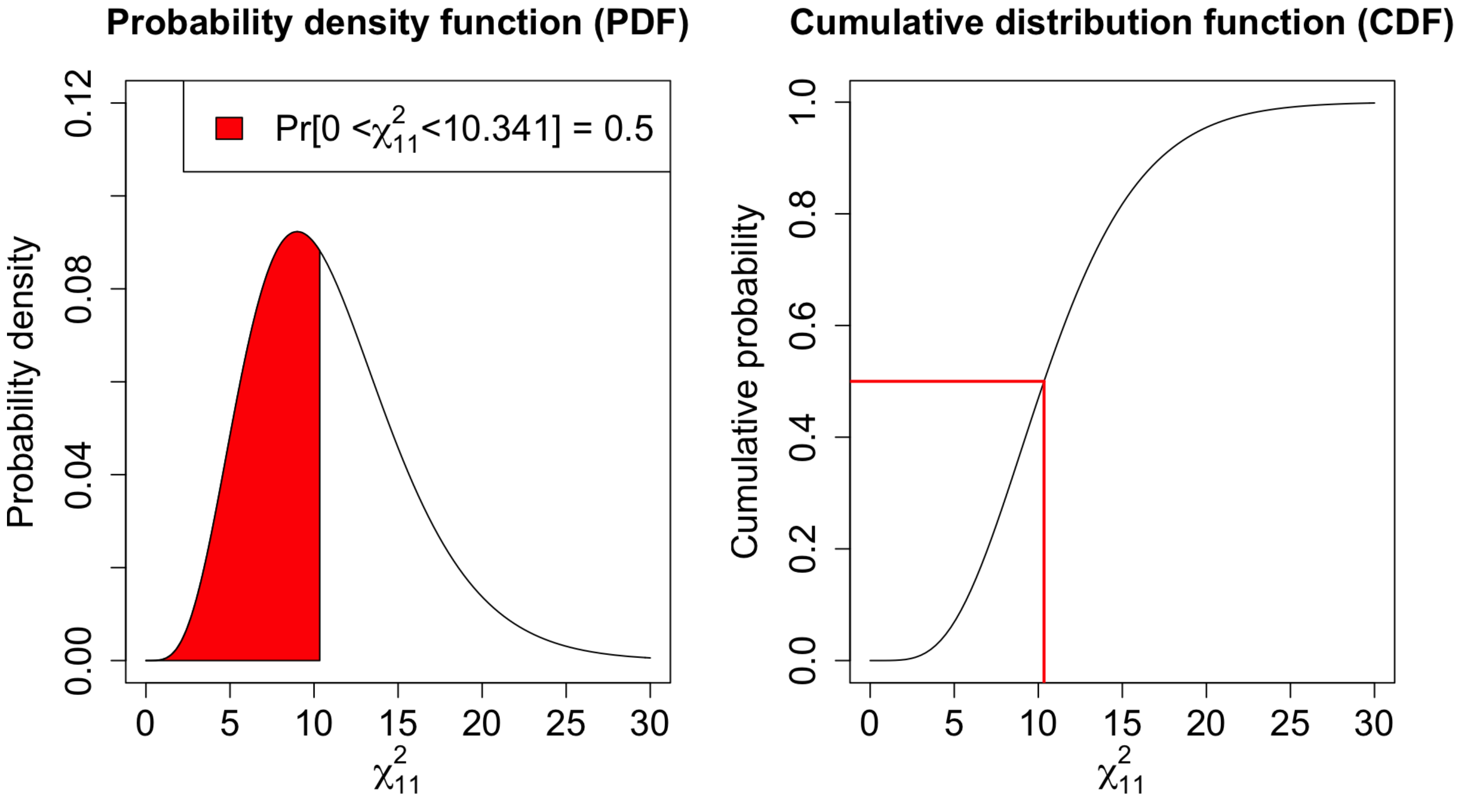
$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

In other words, $F(x)$ asks for the accumulated probability up to x. Remembering calculus, this can be visualized as the area under the probability density curve.

For the χ^2 distribution, x is always non-negative, so the formula simplifies to

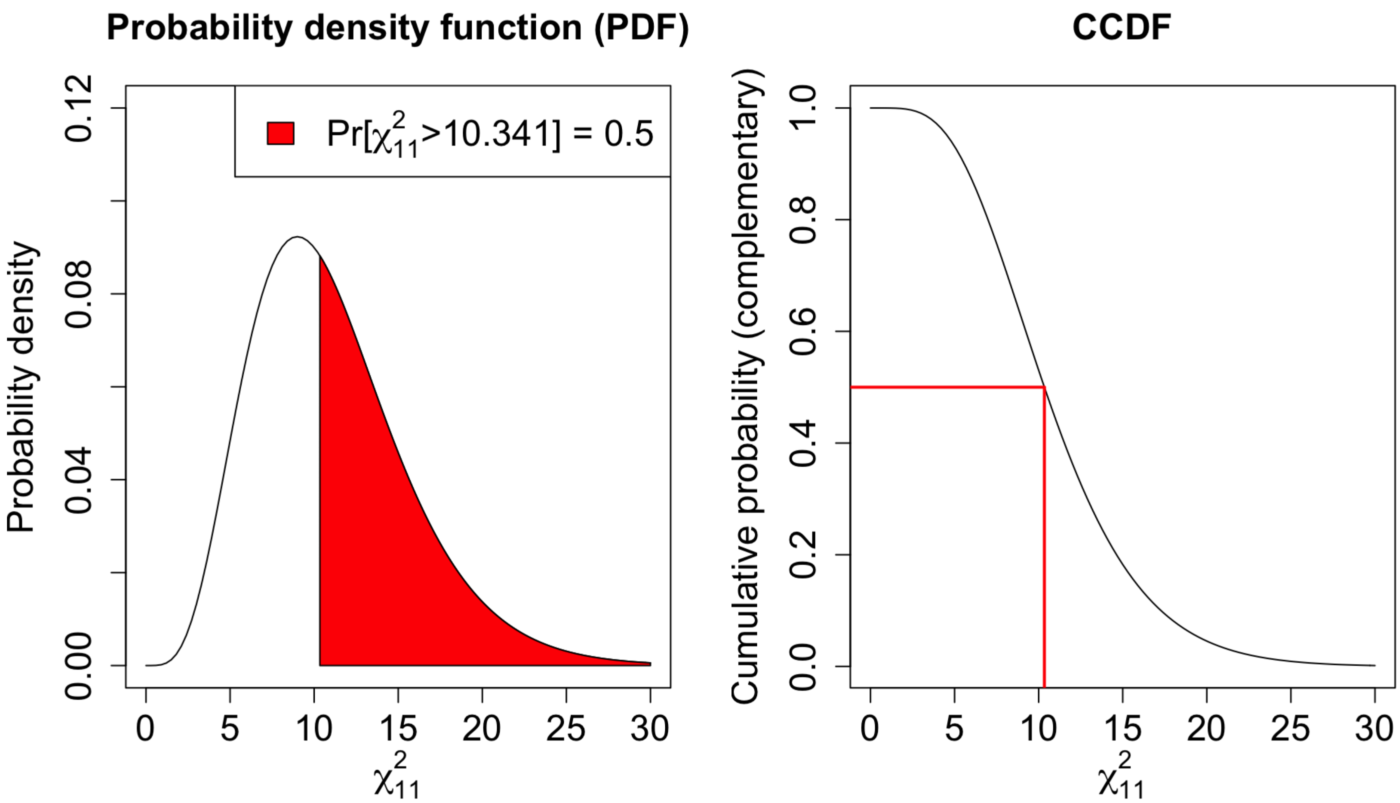
$$F(x) = \int_{0}^{x} f(t)\,dt$$

In the following figure, the left panel corresponds to the χ^2 probability density function $f(x)$, while the right panel corresponds to the CDF $F(x)$.



Complementary cumulative distribution function (CCDF)

The *complementary cumulative distribution function* $G(x)$ G(x) corresponds to the **right-tail** of the probability density function, as can be seen from the following picture.



It can also be defined as follows:

$$G(x) = \int_x^\infty f(t)\,dt = 1 - F(x).$$

G(x)=∫x∞f(t)dt=1−F(x).

How to calculate P-value

For our example from lecture (Feline High Rise Syndrome), we performed the following chi-squared test.

```
mydata <- read.csv("http://whitlockschluter.zoology.ubc.ca/wp-content/data/chapter08/chap08q21FallingCatsByMonth.csv") %>%
  tbl_df()
(chi2 <- chisq.test(table(mydata), p = rep(1/12, 12)))
```

```


    Chi-squared test for given probabilities

data:  table(mydata)
X-squared = 20.664, df = 11, p-value = 0.03703
```

```
chi2 <- chi2$statistic
```

For our example from lecture (Feline High Rise Syndrome), we got the test statistic

$$\chi^2_{11} = 20.66$$

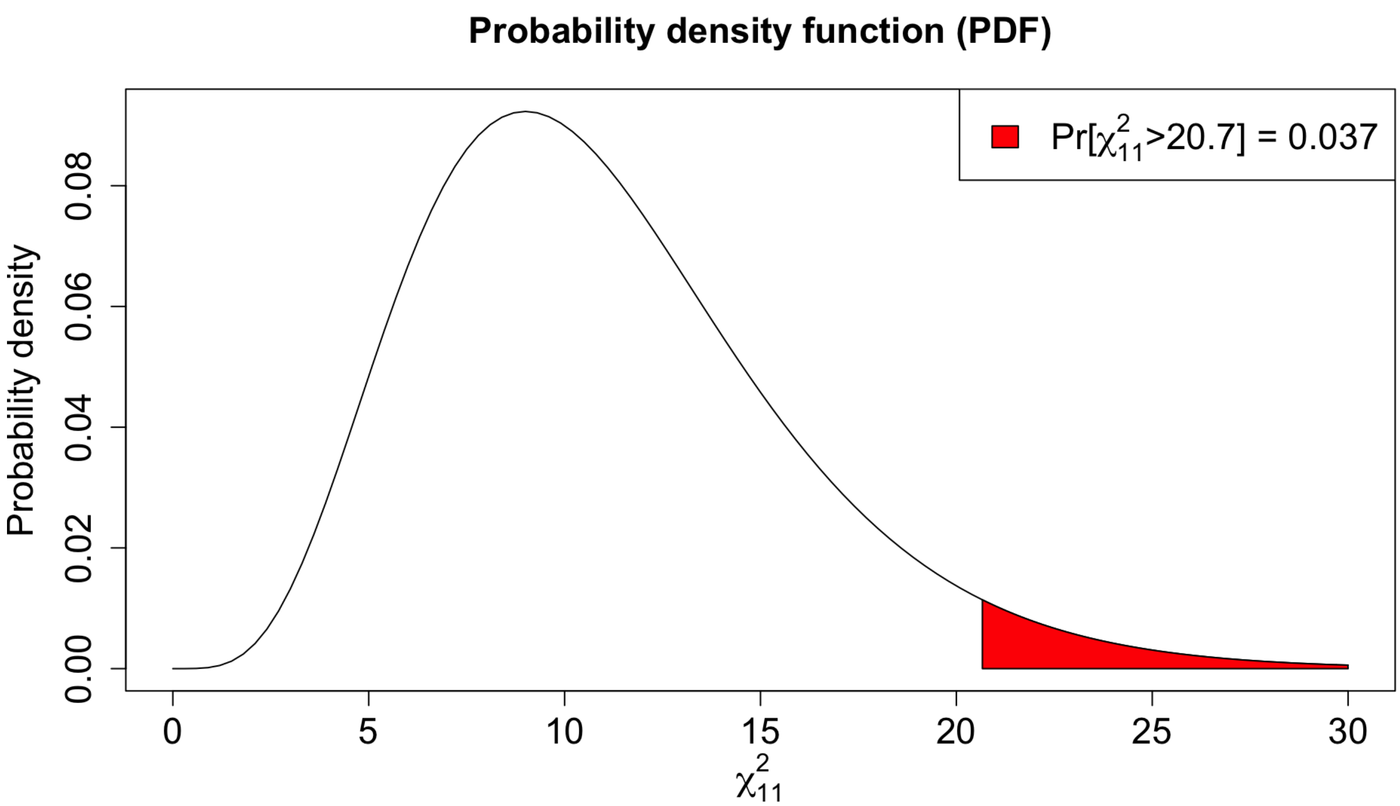
χ112=20.66

So how do we calculate the P-value? The definition of the P-value tells us that

$$P-\text{value} = \Pr[\chi^2_{11} \geq 20.66],$$

P−value=Pr[χ112≥20.66],

as the P-value is the probability of getting your observed test statistic *or worse* in the null distribution. The formula above tells you that the P-value can be calculated by evaluating the CCDF of the χ^2_{11} random variable! Here's the corresponding plot of the area under the PDF that represents the P-value:



So what are the functions in R corresponding to the PDF, CDF and CCDF of a χ^2 random variable?

Name	R command	Uses
PDF	`dchisq(x, df)`	-
CDF	`pchisq(q, df, lower.tail=TRUE)`	-
CCDF	`pchisq(q, df, lower.tail=FALSE)`	Compute P-values

In these commands, `d` stands for `d` ensity, while `p` stands for the cumulative `p` robability distribution.

Exercise - Calculate P-value

Given this table, compute the P-value for a χ^2_{11} test statistic of 20.66.

script.R

```
## Compute the P-value for an 11 degree-of-f
#  chi-squared test statistic of 20.66.
```

R Console

```
>
```

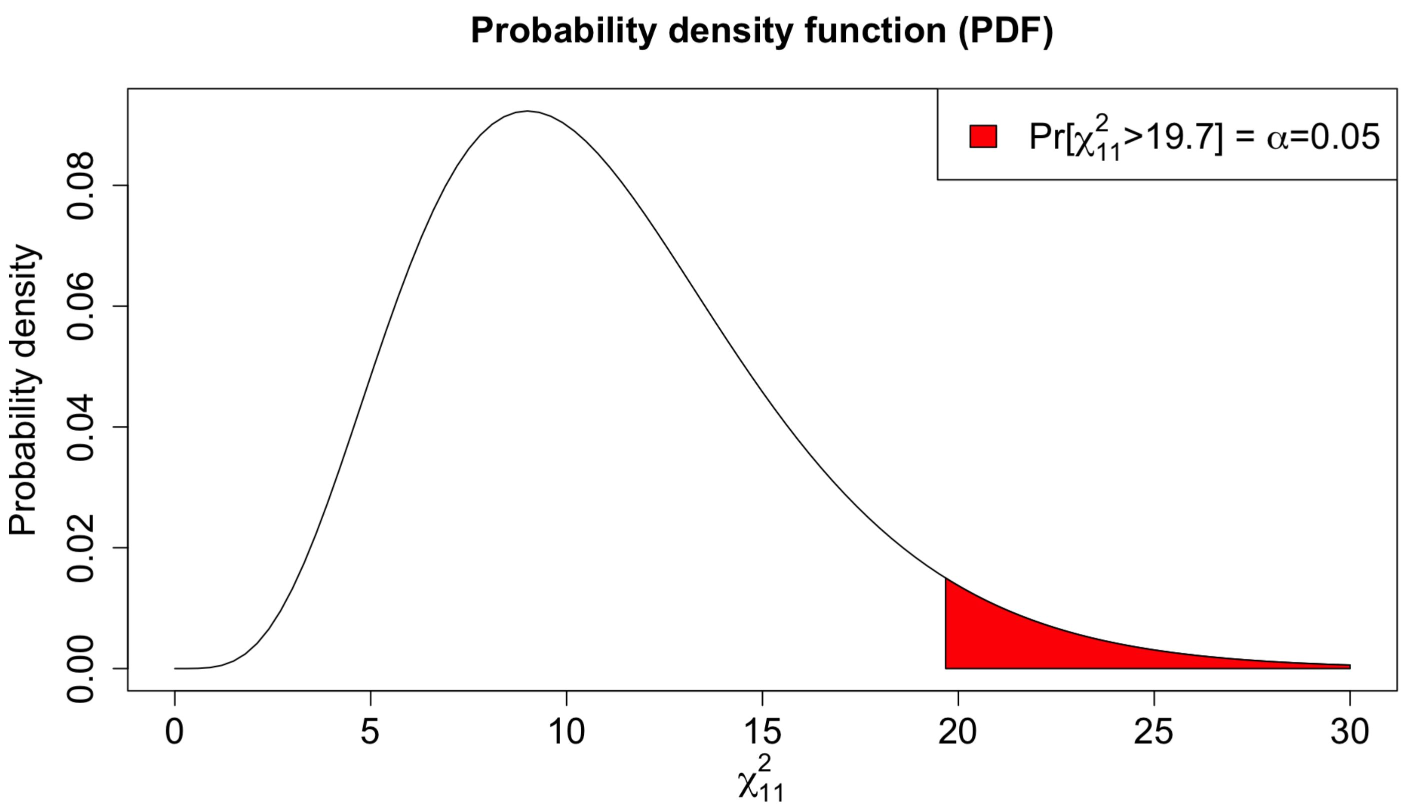
Solution Run

Powered by DataCamp

Calculating critical values

Calculating critical values is the reverse process of calculating P P-values. To calculate P P-values, we have a test statistics as input, and want the area under the curve (probability) as output. Now, for critical values, we start with area under the curve as input and want to know as output at what value of the test statistic do we achieve that given probability.

More specifically, the probability that we use as input is the significance level α α (in many cases 0.05). The critical value associated with that is denoted $\chi^2_{df,\alpha}$ χdf,α2 and is shown in the following figure.

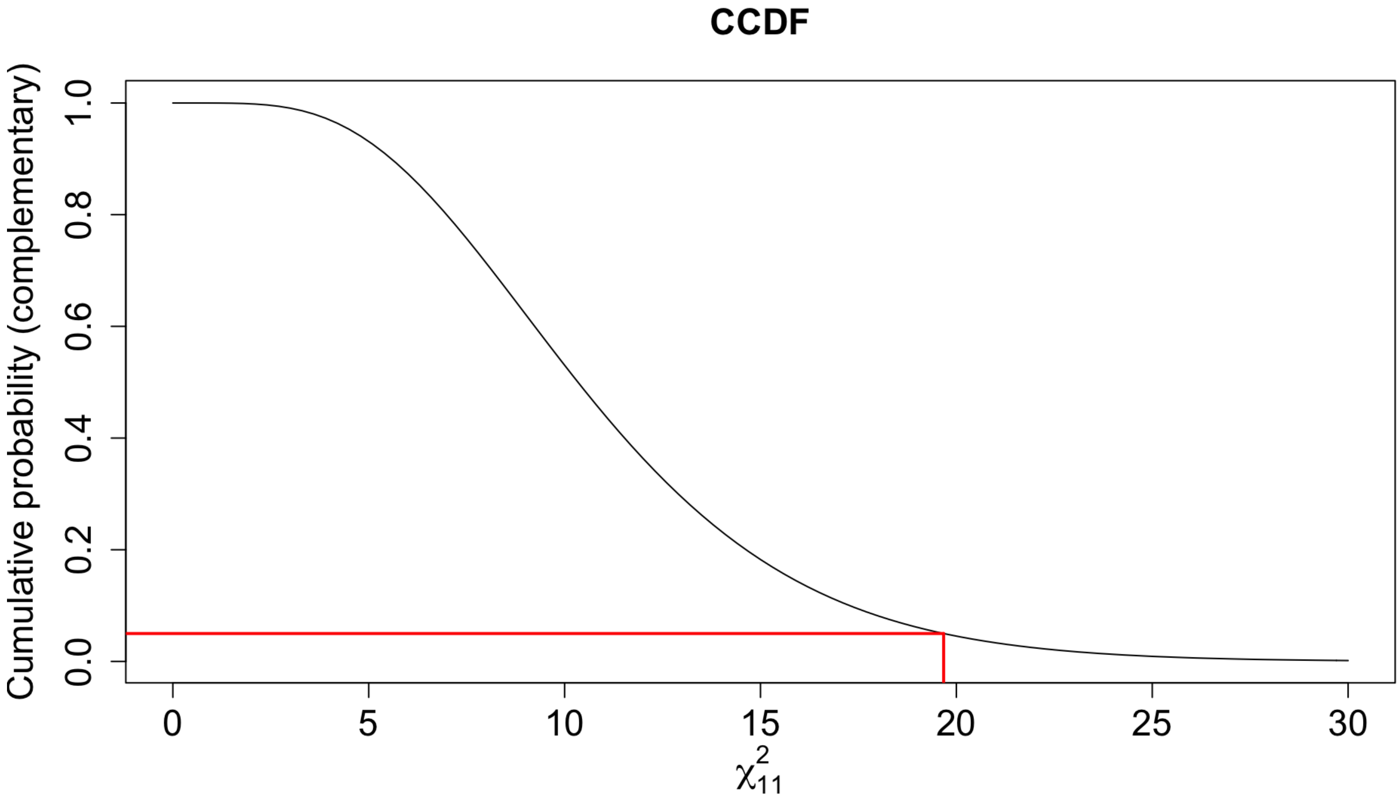


In this case, the critical value is given by

$$\chi^2_{11,0.05} = 19.6751376$$

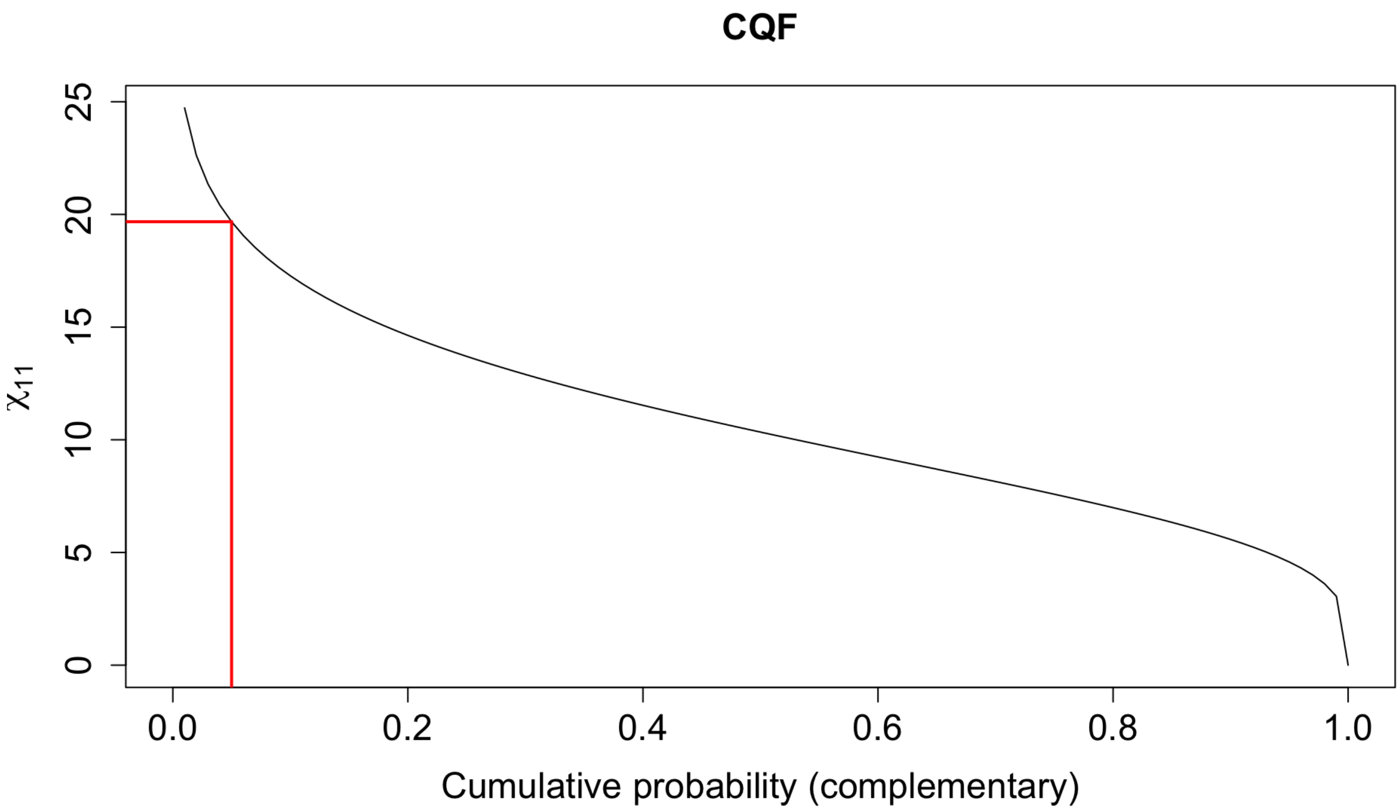
χ11,0.052=19.6751376

Here’s the corresponding CCDF.

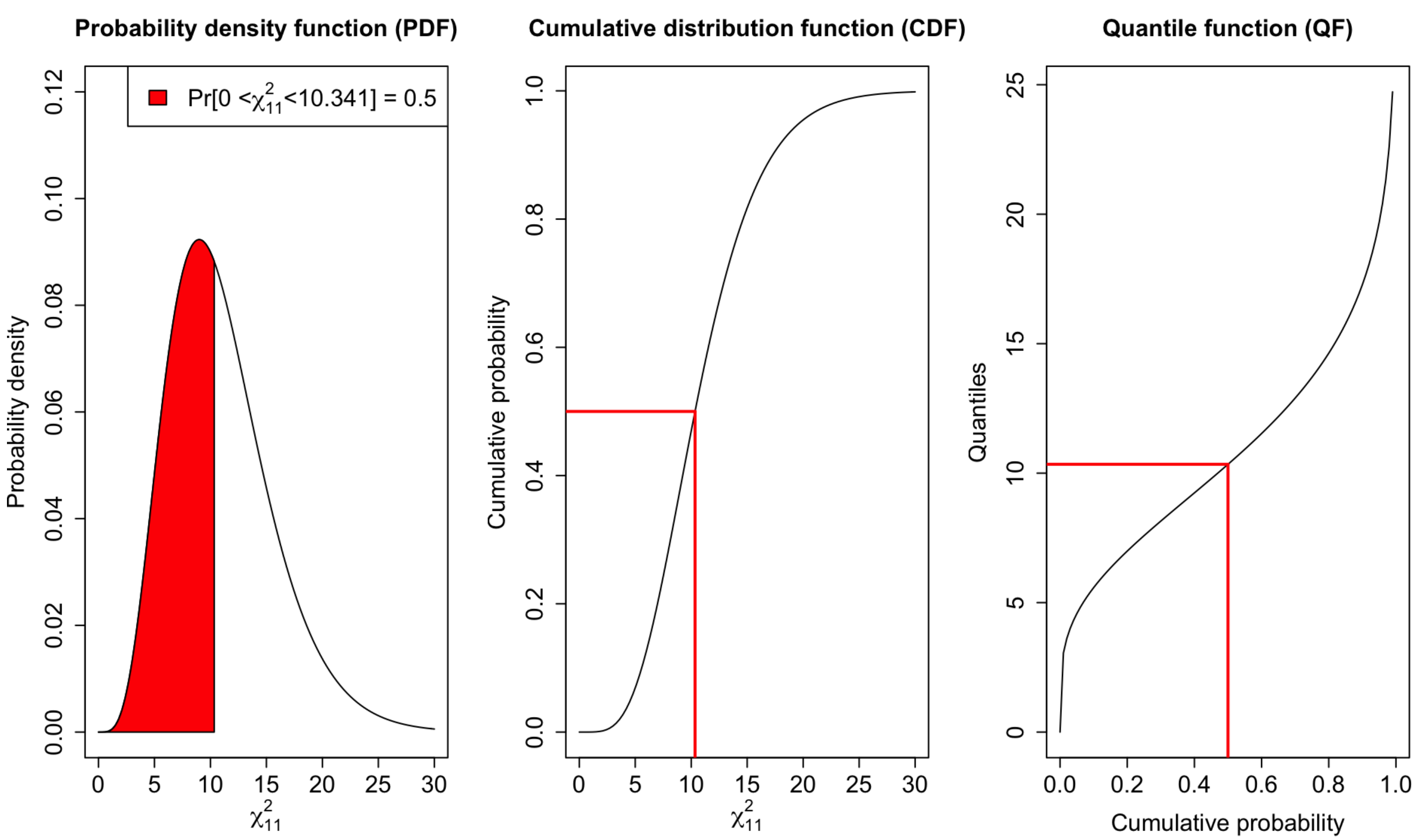


Complementary quantile function (CQF)

By looking at the CCDF, what we notice is we are going *from the vertical axis (probability) to the horizontal axis (test statistic)*. This is the **inverse of the CCDF**, which is called the *complementary quantile function (CQF)*, and is given by the following graph.



We also have the quantile function (QF), which is defined as the *inverse of the cumulative distribution function*, i.e. $Q = F^{-1}$ Q=F−1. The quantile function can be used to calculate **all quantiles** of a distribution, including medians and quartiles. The 3rd panel in the following figure shows the quantile function. Make sure to compare its graph to the 2nd panel, which is the cumulative distribution function.



So, again, what are the functions in R corresponding to the CQF (and quantile function, QF) of a χ^2 random variable?

Name	R command	Uses
QF	`qchisq(p, df, lower.tail=TRUE)`	-
CQF	`qchisq(p, df, lower.tail=FALSE)`	Compute critical values

In these commands, `p` stands for `p` robability (the significance level in this case) and `q` stands for `q` uantile function.

Exercise - Calculate critical value

Given this table, compute the critical value $\chi^2_{11,0.05}$.

script.R

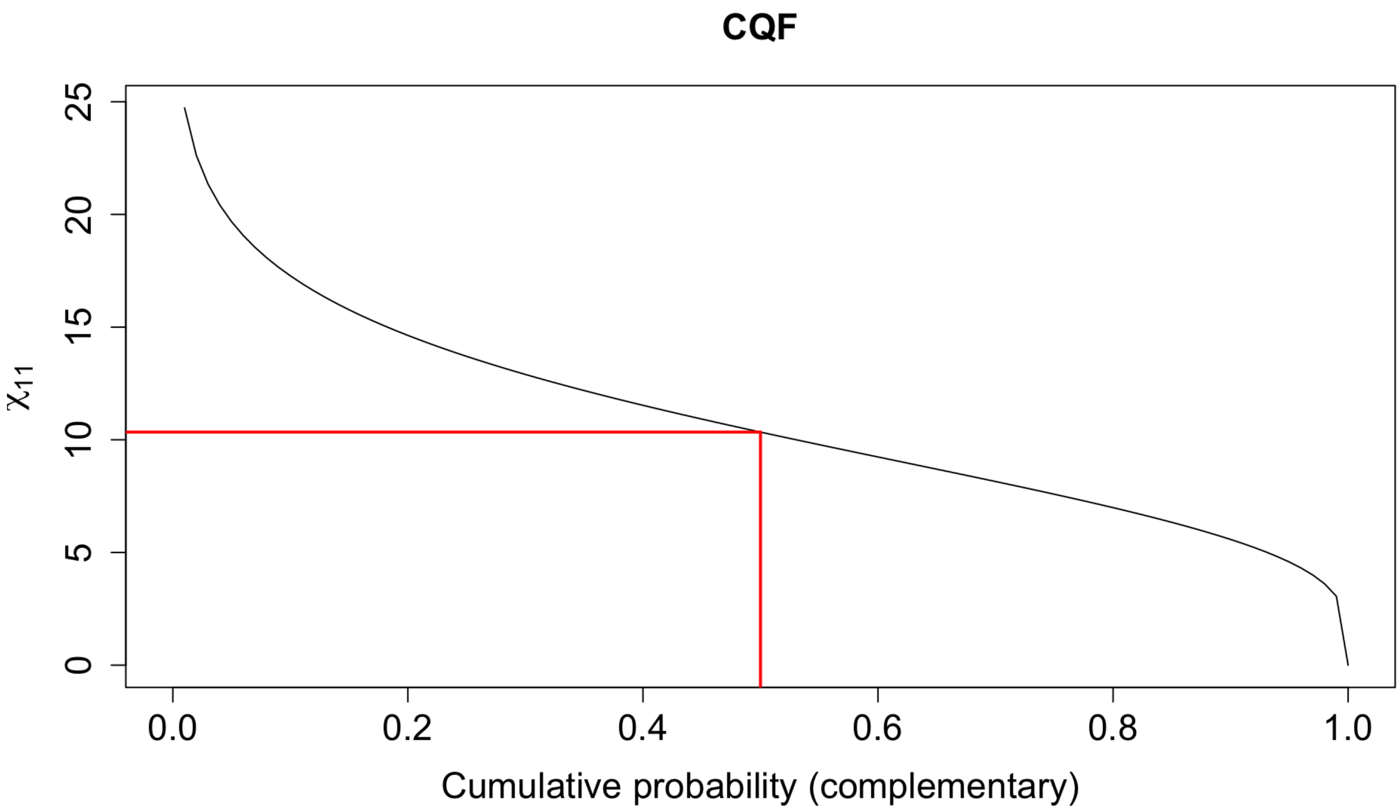
```
## Compute the critical value for an 11 degr
#  chi-squared distribution with alpha = 0.0
```

R Console

```
>
```

Solution Run

Here’s the graph associated with the previous calculation.



Exercise - Calculate quantiles

script.R

```
## Compute the median of a chi-squared distr
11 degrees of freedom


## Compute the 1st quartile of a chi-squared
with 11 degrees of freedom


## Compute the 3rd quintile of a chi-squared
with 11 degrees of freedom
```

R Console

```
>
```

Solution Run

Powered by DataCamp

Summary of functions for chi-squared

Name	R command	Uses
PDF	`dchisq(x, df)`	Compute density

Name	R command	Uses
CDF	`pchisq(q, df, lower.tail=TRUE)`	Compute cumulative probability
CCDF	`pchisq(q, df, lower.tail=FALSE)`	Compute P P-values
QF	`qchisq(p, df, lower.tail=TRUE)`	Compute quantiles
CQF	`qchisq(p, df, lower.tail=FALSE)`	Compute critical values

Chi-squared tests

Here's a quick example of how to do a χ^2χ2 goodness-of-fit test the quick way. This is based on Whitlock & Schluter, Chapter 8, Question #21.

```
mydata <- read.csv("http://whitlockschluter.zoology.ubc.ca/wp-content/data/chapter08/chap08q21Fall
ingCatsByMonth.csv")
str(mydata)
```

```
'data.frame':	119 obs. of  1 variable:
 $ month: Factor w/ 12 levels "April","August",..: 5 5 5 5 4 4 4 4 4 4 ...
```

Let's construct the table.

```
(mytable <- table(mydata))
```

```
mydata
    April    August  December  February   January      July      June 
       10        13         5         6         4        19        14 
    March       May  November   October September 
        8         9         7        12        12 
```

Now perform the test using the `chisq.test` function.

```
chisq.test(mytable, p = rep(1/12, 12))
```

```

	Chi-squared test for given probabilities

data:  mytable
X-squared = 20.664, df = 11, p-value = 0.03703
```

The argument `p` needs to be a probability distribution (i.e. sum to 1). In other words, `p` is NOT the expected frequencies - `p` represents the *expected relative frequencies*.