STA258H5

Al Nosedal
and Alison Weir

Winter 2017

SAMPLING DISTRIBUTIONS RELATED TO A NORMAL POPULATION

Theorem 7.1

Let $Y_1, Y_2, ..., Y_n$ be a random sample of size n from a Normal distribution with mean μ and variance σ^2. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is Normally distributed with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$.

(or $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$)

Example

Marks on a standardized test are Normally distributed with mean 75, standard deviation 15. What is the probability the class average, for a class of 30, is greater than 76?

χ_1^2 distribution

It can be shown (in STA260) that χ_1^2 distribution is the Gamma($\alpha = 1/2, \beta = 2$) distribution.

$$f(y) = \begin{cases} \frac{1}{\Gamma(1/2)2^{1/2}} y^{1/2-1} e^{-y/2} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

$E(Y) = 1$ $\qquad$ $Var(Y) = 2.$

Example

Let Z be a Normally distributed random variable with mean 0 and variance 1. Use the method of moment-generating functions to find the probability distribution of Z^2.

Solution

$M_{Z^2}(t) = E(e^{tZ^2}) = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$
$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2(\frac{1-2t}{2})} dz$
This integral can be evaluated using an "old trick" (we note that it looks like a Normally distributed random variable).

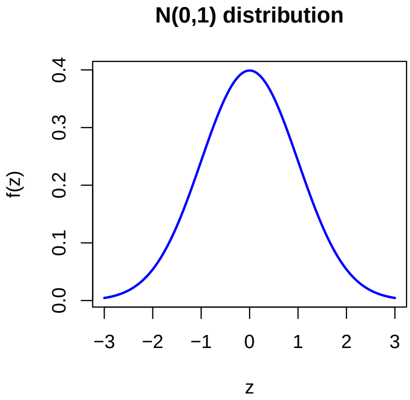
Solution

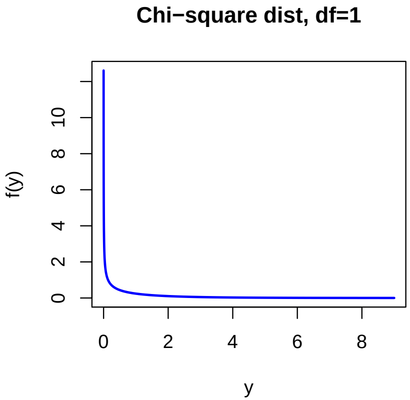
We realize that $e^{-z^2(\frac{(1-2t)}{2})}$ is proportional to a Normal with $\mu = 0$ and $\sigma^2 = 1/(1-2t)$, then

$M_{Z^2}(t) = \frac{\sqrt{2\pi}\sqrt{1/(1-2t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1/(1-2t)}} e^{-z^2(\frac{1-2t}{2})} dz$

$M_{Z^2}(t) = \sqrt{\frac{1}{1-2t}} = (1-2t)^{-1/2}$ (Note. This is valid provided that $t < 1/2$).

$(1-2t)^{-1/2}$ is the moment-generating function for a gamma-distributed random variable with $\alpha = 1/2$ and $\beta = 2$. Hence, Z^2 has a χ^2 distribution with $\nu = 1$ degree of freedom.


N(0,1) distribution
f(z)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
z


Chi−square dist, df=1
f(y)
0
2
4
6
8
10
0
2
4
6
8
y

Example

Suppose that Y_1 and Y_2 are independent, standard Normal random variables. Find the probability distribution of $U = Y_1^2 + Y_2^2$.

Solution

$M_U(t) = E[e^{Ut}] = E[e^{(Y_1^2+Y_2^2)t}]$
$= E[e^{Y_1^2 t}]E[e^{Y_2^2 t}]$
$= M_{Y_1^2}(t)M_{Y_2^2}(t)$
$= [(1-2t)^{-1/2}][(1-2t)^{-1/2}] = (1-2t)^{-2/2}.$
Because moment-generating functions are unique, U has a χ^2 distribution with 2 degrees of freedom.

χ^2_n distribution

Definition. Let $Z_1, Z_2, ..., Z_n$ be iid standard Normal random variables.
Define $V = \sum_{i=1}^{n} Z_i^2$. V has a χ^2_n distribution.
The χ^2_n is the Gamma($\alpha = n/2, \beta = 2$) distribution.

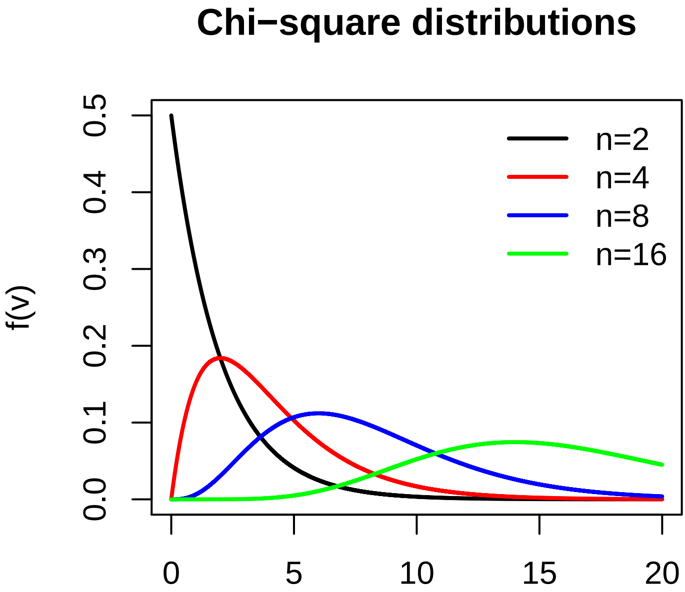
Proof.
By uniqueness of moment generating functions.

χ^2_n distribution

$$f(v) = \begin{cases} \frac{1}{\Gamma(n/2)v^{n/2}} v^{n/2-1} e^{-v/2} & v > 0 \\ 0 & v \leq 0 \end{cases}$$

$E(V) = n \qquad Var(V) = 2n.$

The subscript n is called the degrees of freedom; it is the number of "free" variables in the sum.


Chi−square distributions
f(v)
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
n=2
n=4
n=8
n=16

Theorem 7.2

Let $Y_1, Y_2, ..., Y_n$ be defined as in Theorem 7.1. Then $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent, standard Normal random variables, $i = 1, 2, ..., n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

It can be shown (see STA437) that

If $X_1, X_2, ..., X_n \sim$ iid $N(\mu, \sigma^2)$, then $V = \sum_{j=1}^{n} \left(\frac{X_j - \bar{X}}{\sigma} \right) \sim X_{n-1}^2$

Question: Why do we loose 1 degree of freedom, when we use $\bar{X}$ instead of μ?

Answer: If we fix the sample mean to be a specific number, then X_1 is random (it can be any number), the same holds for $X_2, ..., X_{n-1}$. But X_n is not random, it has to be whatever value it needs to force the sample mean to be a specific number. There are only $n-1$ "free" variables in the sum.

Theorem 7.3

Let $Y_1, Y_2, ..., Y_n$ be a random sample from a Normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a χ^2 distribution with $(n-1)$ df.

Since, if $Y_1, Y_2, ..., Y_n \sim$ iid $N(\mu, \sigma)$, then $V = \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 \sim \chi^2_{n-1}$

It follows that, if $Y_1, Y_2, ..., Y_n \sim$ iid $N(\mu, \sigma)$ and

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, then $V = \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

$E(S^2) = \sigma^2$ and $Var(S^2) = \frac{2\sigma^4}{n-1}$.

Example

The Environmental Protection Agency is concerned with the problem of setting criteria for the amounts of certain toxic chemicals to be allowed in freshwater lakes and rivers. A common measure of toxicity for any pollutant is the concentration of the pollutant that will kill half of the test species in a given amount of time (usually 96 hours for fish species). This measure is called LC50 (lethal concentration killing 50% of test species). In many studies, the values contained in the natural logarithm of LC50 measurements are Normally distributed, and, hence, the analysis is based on ln(LC50) data.

Example

Suppose that $n = 20$ observations are to be taken on ln(LC50) measurements and that $\sigma^2 = 1.4$. Let S^2 denote the sample variance of the 20 measurements.

a. Find a number b such that $P(S^2 \leq b) = 0.975$.

b. Find a number a such that $P(a \leq S^2) = 0.975$.

c. If a and b are as in parts a) and b), what is $P(a \leq S^2 \leq b)$?

Solution

These values can be found by using percentiles from the chi-square distribution.

With $\sigma^2 = 1.4$ and $n = 20$,

$\frac{n-1}{\sigma^2}S^2 = \frac{19}{1.4}S^2$ has a chi-square distribution with 19 degrees of freedom.

a. $P(S^2 \leq b) = P\left(\frac{n-1}{\sigma^2}S^2 \leq \frac{(n-1)b}{\sigma^2}\right) = P\left(\frac{19}{1.4}S^2 \leq \frac{19b}{1.4}\right) = 0.975$

$\frac{19b}{1.4}$ must be equal to the 97.5%-tile of a chi-square with 19 df, thus

$\frac{19b}{1.4} = 32.8523$ (using Table 6). An so, $b = 2.42$

Solution

b. Similarly, $P(S^2 \geq a) = P\left(\frac{n-1}{\sigma^2}S^2 \geq \frac{(n-1)a}{\sigma^2}\right) = 0.975$. Thus, $\frac{19a}{1.4} = 8.90655$, the 2.5%-tile of this chi-square distribution, and so $a = 0.656$.

c. $P(a \leq S^2 \leq b) = P(0.656 \leq S^2 \leq 2.42) = 0.95$.

HW?

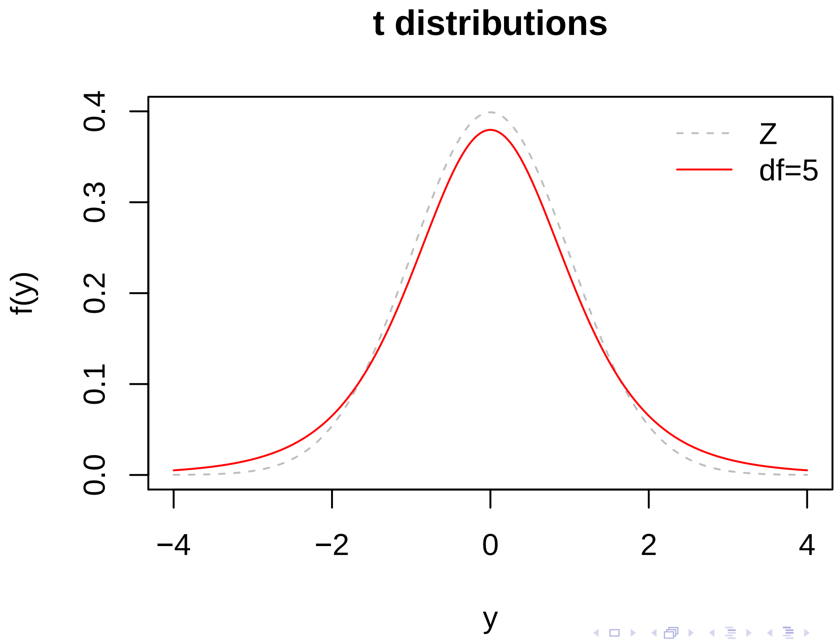
Marks on a standardized test are Normally distributed with mean 75, standard deviation 15. What is the probability the class sample standard deviation, for a class of 31, is greater than 16?

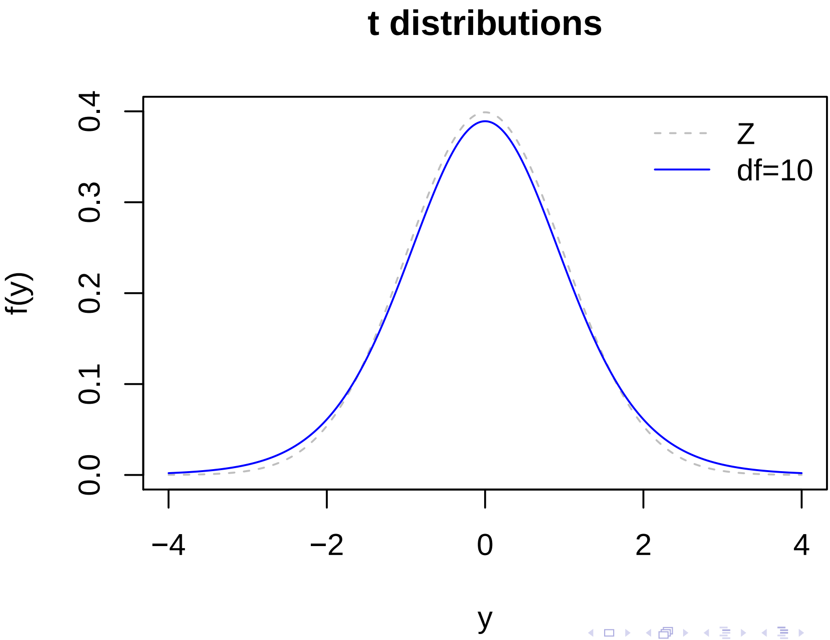
Definition 7.2 (t Distribution)

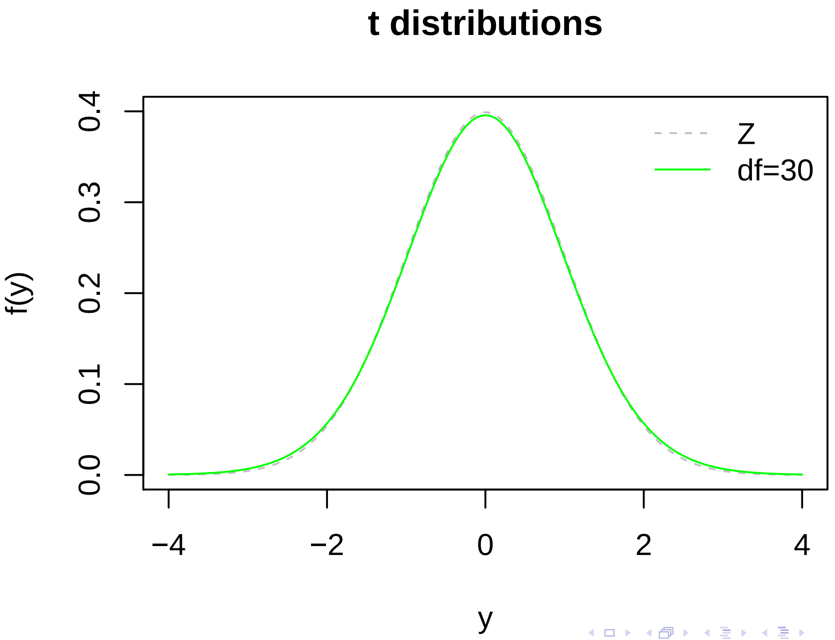
Let Z be a standard Normal random variable and let W be a χ^2-distributed variable with ν df. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a t distribution with ν df.


t distributions
Z
df=5
0.0
0.1
0.2
0.3
0.4
f(y)
−4
−2
0
2
4
y


t distributions
Z
df=10
0.0
0.1
0.2
0.3
0.4
f(y)
−4
−2
0
2
4
y



t distributions
Z
df=30
f(y)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
y

Properties of the t distribution, $T \sim t_\nu$

- Support - the real numbers
- $E(T) = 0$
- $Var(T) = \frac{\nu}{\nu-2} > 1$ $Var(T) \to 1$ as $v \to \infty$
- The t_ν is shorter and fatter than the $N(0,1)$ distribution. The tails of the t_ν decrease to zero slower than those for the $N(0,1)$.
- $t_\nu \to Z$ (in distribution) as $\nu \to \infty$

Question: Why is the t_ν wider and more tail heavy than the $N(0,1)$.
Answer:
$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ contains variability from $\bar{X}$.
$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ contains variability from $\bar{X}$ and from S.

Very important use of the t distribution

If $Y_1, Y_2, ..., Y_n \sim$ iid $N(\mu, \sigma^2)$, then

1. $\bar{Y} \sim N(\mu, \sigma^2/n)$ or $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$
3. $\bar{X}$ and S^2 are independent (see STA437)

Thus $\frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n-1}{\sigma^2} S^2/(n-1)}} = \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

Example

The Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

Solution

$X_1, X_2, ..., X_{15}$ iid random variables from a $N(\mu = 300, \sigma = 15)$.

$$
\begin{aligned}
P(\bar{X} \leq 290) &= P\left(\frac{\bar{X}-\mu}{s/\sqrt{n}} \leq \frac{290-300}{50/\sqrt{15}}\right) \\
&= P(t_{14} \leq -0.7745) \text{ (since T distributions are symmetric)} \\
&= P(t_{14} \geq 0.7745) \text{ (using our T distribution table)}
\end{aligned}
$$

The best that we can do using table is:

$0.20 < P(\bar{X} \leq 290) < 0.25$

```r
# pt = CDF of a t distribution;

pt(-0.7745,14);

## [1] 0.2257589
```

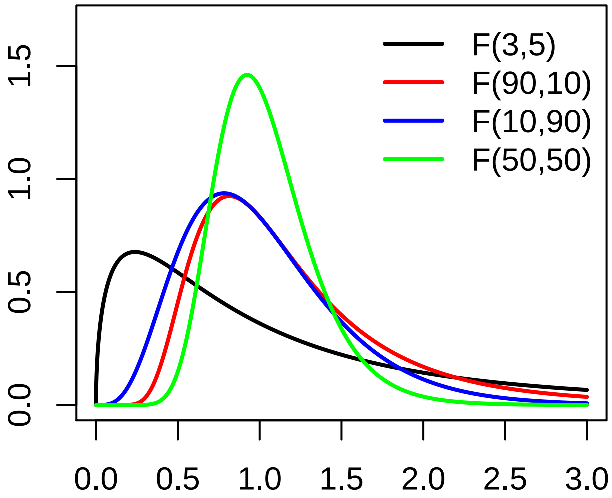
Definition 7.3

Let W_1 and W_2 be independent χ^2-distributed random variables with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F_{\nu 1,\ \nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

F distributions



Properties of F distribution, F_{ν_1,ν_2}

- Support - non-negative real numbers
- Relationship with its reciprocal: If $F \sim F_{\nu_1,\nu_2}$ then $W_1 = 1/F \sim F_{\nu_2,\nu_1}$
- Relationship with the t distribution: If $T \sim t_\nu$ then $W_2 = T^2 \sim F_{1,\nu}$
- Relationship with the exponential distribution: If U_1 and U_2 are independent exponential RVs with the same parameter, then $W_3 = U_1/U_2 \sim F_{2,2}$

Very important use of the F distribution

If

1. $X_1, X_2, ..., X_n \sim$ iid $N(\mu_X, \sigma^2)$, then $\frac{n-1}{\sigma^2} S_X^2 \sim \chi^2_{n-1}$
2. $Y_1, Y_2, ..., Y_m \sim$ iid $N(\mu_Y, \sigma^2)$, then $\frac{m-1}{\sigma^2} S_Y^2 \sim \chi^2_{m-1}$
3. Xs and Ys all independent

then

$$F = \frac{\left(\frac{n-1}{\sigma^2} S_X^2\right)/(n-1)}{\left(\frac{m-1}{\sigma^2} S_Y^2\right)/(m-1)} = \frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}$$

Example

Marks on a standardized test are normally distributed with mean 75, standard deviation 15. The test is given to two classes, one class (A) has 31 students and the other (B) has 16 students.

- What is the probability the ratio of the two sample variances (A/B) is greater 2?
- What is the probability the ratio of the two sample variances (B/A) is less than 2?

Solution (A/B)

$\mu = 75$, $\sigma = 15$, $n = 31$, $m = 16$.

$$
\begin{aligned}
P\left(\frac{S_X^2}{S_Y^2} > 2\right) &= P\left(\frac{\sigma^2}{\sigma^2}\frac{S_X^2}{S_Y^2} > 2\right) \\
&= P\left(\frac{S_X^2/\sigma^2}{S_Y^2/\sigma^2} > 2\right) \\
&= P\left(\frac{\frac{n-1}{\sigma^2}\frac{S_X^2}{n-1}}{\frac{m-1}{\sigma^2}\frac{S_Y^2}{m-1}} > 2\right) \\
&= \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} \\
&= P(F_{n-1,\ m-1} > 2) = P(F_{30,15} > 2)
\end{aligned}
$$

Using our Table, the best that we can do is:

$0.05 < P\left(\frac{S_X^2}{S_Y^2} > 2\right) < 0.10.$

```r
# pf = CDF of F ;

# area to the left of 2;

pf(2,30,15);

## [1] 0.9213152

# area to the right of 2;

1-pf(2,30,15);

## [1] 0.07868481
```

Solution (B/A)

$\mu = 75,\ \sigma = 15,\ n = 31,\ m = 16.$

$$
\begin{aligned}
P\left(\frac{S_Y^2}{S_X^2} > 2\right) &= P\left(\frac{\sigma^2}{\sigma^2}\frac{S_Y^2}{S_X^2} > 2\right) \\
&= P\left(\frac{S_Y^2/\sigma^2}{S_X^2/\sigma^2} > 2\right) \\
&= P\left(\frac{\frac{m-1}{\sigma^2}\frac{S_Y^2}{m-1}}{\frac{n-1}{\sigma^2}\frac{S_X^2}{n-1}} > 2\right) \\
&= \frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} \\
&= P(F_{m-1,\ n-1} > 2) = P(F_{15,30} > 2)
\end{aligned}
$$

```r
# pf = CDF of F ;

# area to the left of 2;

pf(2,15,30);

## [1] 0.948209

# area to the right of 2;

1-pf(2,15,30);

## [1] 0.051791
```