More Useful Distributions And Test Statistics

Eli Gurarie

StatR 101 - Lecture 10
November 26, 2012

November 26, 2012

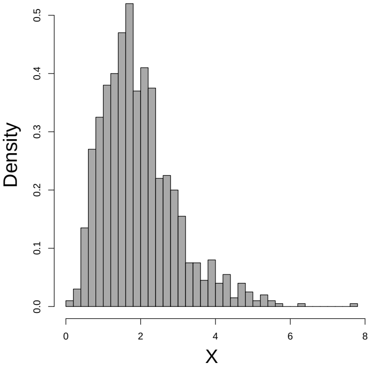
Outline

	Distributions	Comparisons	Statistics
1.	Chi-squared(ν)	Proportions	
2.	$\mathcal{T}(\nu)$	Sample means	$\frac{\overline{X}}{s_x/\sqrt{n}}$
3.	$\mathcal{F}(\nu_1, \nu_2)$	Sample variances	$\frac{s_1^2/n_1}{s_2^2/n_2}$

These three distributions are all derived from the normal distribution and are the most widely used null-distributions for hypothesis testing. You will see them pop up frequently in statistical test output.

Random Variable: X

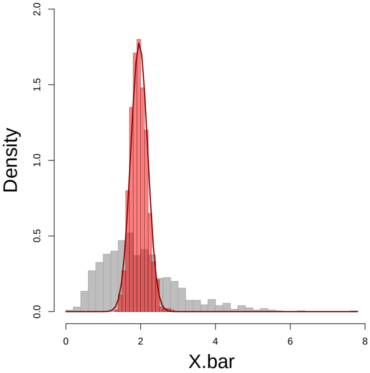
$$X \sim \text{Dist}(\mu, \sigma^2)$$



Sample Mean: $\overline{X}$

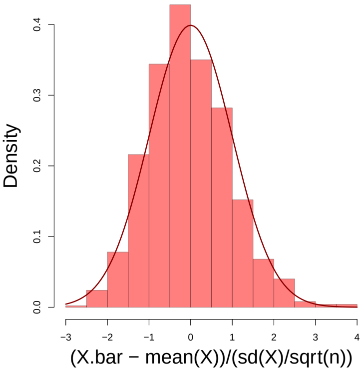
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$



Standardized Sample Mean: Z

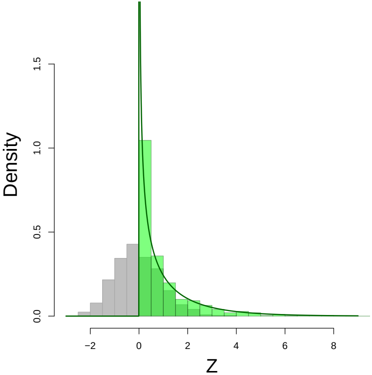
$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0,1)$$



This observation allows us to use th z-statistic to tests of means ($H_0 : \mu_{obs} = \mu_0$) with known variance.

Chi-squared distribution

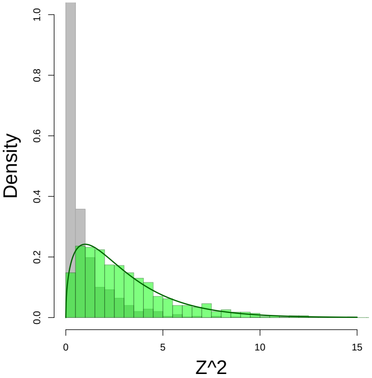
$$Y = Z^2$$
$$\sim \chi^2(\nu = 1)$$
$$f(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}$$



Chi-squared distribution

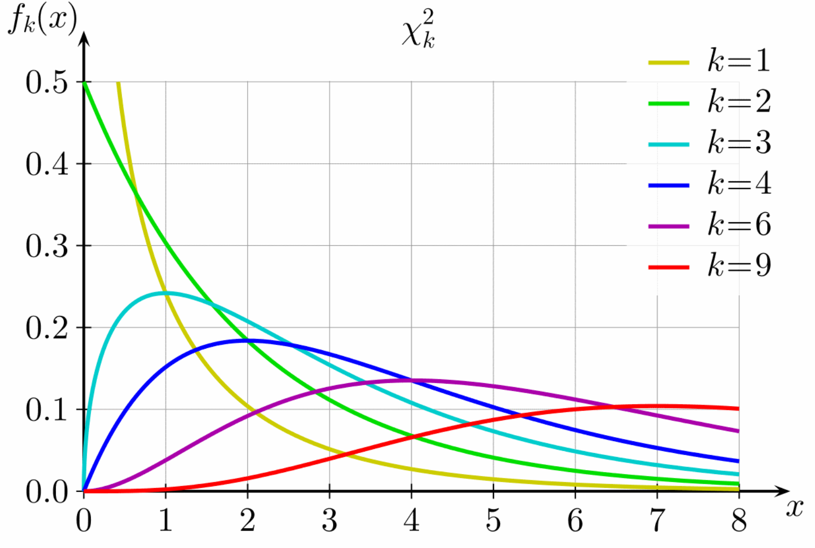
$$Y = Z_1^2 + Z_2^2 + Z_3^2$$
$$\sim \chi^2(\nu = 3)$$

$$f(y) = \frac{y^{(\nu/2)-1}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} e^{-y/2}$$



The "Chi-squared" distribution has one parameter: ν - the 'degrees of freedom', that represents the number of independent standard normal variables that are summed.

Chi-squared distribution



http://en.wikipedia.org/wiki/Chi-squared_distribution

Chi-squared distribution: Card

$X \sim \chi^2(\nu)$

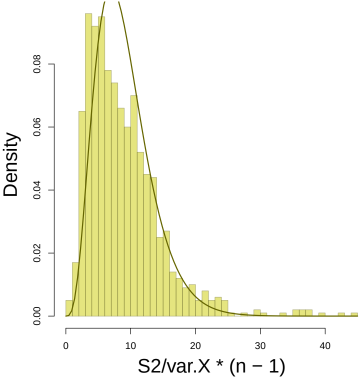
Name:	Chi-squared distribution	Models sum of ν standard normal r.v.'s
Support:	x	$[0, \infty)$
Parameters:	$\nu \in (0, \beta)$	degrees of freedom
pdf:	$\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$	
mean:	$\mathrm{E}\left(X\right)$	ν
variance:	$\mathrm{Var}\left(X\right)$	2ν
`Rcode`:	`*chisq`	argument: `df`

Chi-squared distribution: as a statistic for sample variance

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
\frac{s^2}{\sigma^2} &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\sigma^2}
\end{aligned}
$$

some algebra

$$
\begin{aligned}
(n-1)\frac{s^2}{\sigma^2} &= \sum_{i=1}^{n-1}\frac{(X_i - \mu)^2}{\sigma^2} \\
(n-1)\frac{s^2}{\sigma^2} &\sim \chi^2(\nu = n-1)
\end{aligned}
$$



This means the statistic $(n-1)\frac{s^2}{\sigma^2}$ *could* be used to test $H_0 : \sigma_{obs} = \sigma_0$, i.e. an observed standard deviation against a hypothesized one. But this is not a very common test.

Chi-squared distribution: as a statistic for sample variance

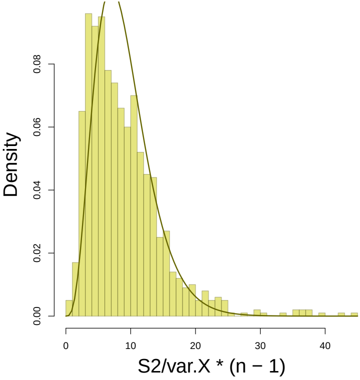
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

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

some algebra

$$(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n-1}\frac{(X_i - \mu)^2}{\sigma^2}$$

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



This means the statistic $(n-1)\frac{s^2}{\sigma^2}$ *could* be used to test $H_0 : \sigma_{obs} = \sigma_0$, i.e. an observed standard deviation against a hypothesized one. But this is not a very common test.

Chi-squared distribution: as a statistic for sample variance

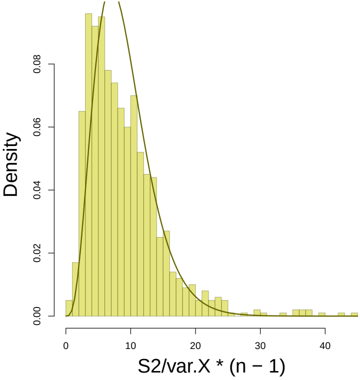
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

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

some algebra

$$(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n-1}\frac{(X_i - \mu)^2}{\sigma^2}$$

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



This means the statistic $(n-1)\frac{s^2}{\sigma^2}$ *could* be used to test $H_0 : \sigma_{obs} = \sigma_0$, i.e. an observed standard deviation against a hypothesized one. But this is not a very common test.

T-distribution: Ratio of Z to χ^2

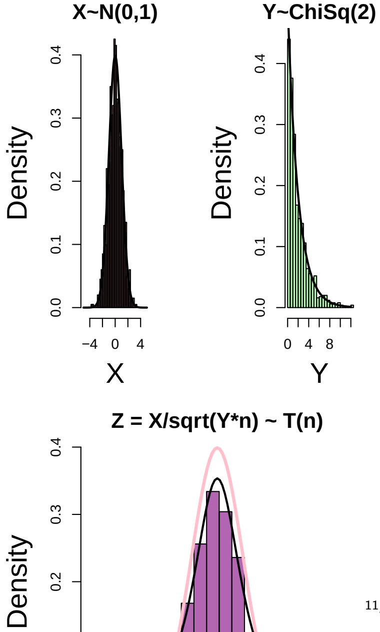
$$T = \frac{Z}{\sqrt{Y_\nu/\nu}}$$

where: $Z \sim \mathcal{N}(0,1)$

and: $Y_\nu \sim \chi^2(\nu)$

then:

$$T \sim \mathcal{T}(\nu)$$



T-distribution: Ratio of Z to χ^2

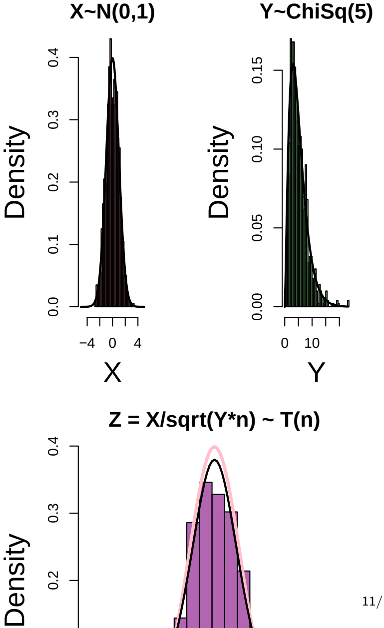
$$T = \frac{Z}{\sqrt{Y_\nu/\nu}}$$

where: $Z \sim \mathcal{N}(0, 1)$

and: $Y_\nu \sim \chi^2(\nu)$

then:

$$T \sim \mathcal{T}(\nu)$$



T-distribution: Ratio of Z to χ^2

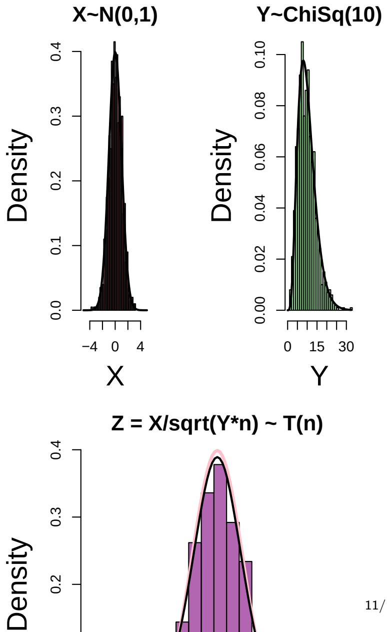
$$T = \frac{Z}{\sqrt{Y_\nu/\nu}}$$

where: $Z \sim \mathcal{N}(0,1)$

and: $Y_\nu \sim \chi^2(\nu)$

then:

$$T \sim \mathcal{T}(\nu)$$



T-distribution: Ratio of Z to χ^2

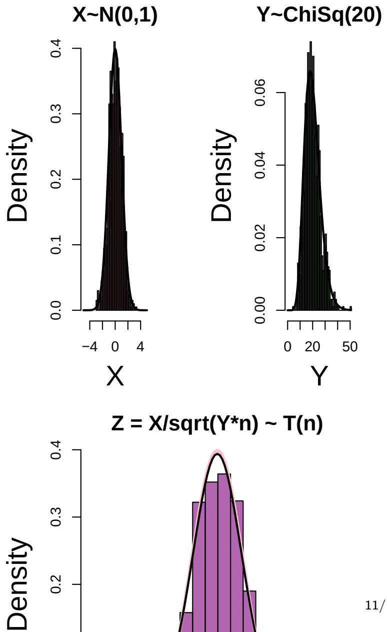
$$T = \frac{Z}{\sqrt{Y_\nu/\nu}}$$

where: $Z \sim \mathcal{N}(0,1)$

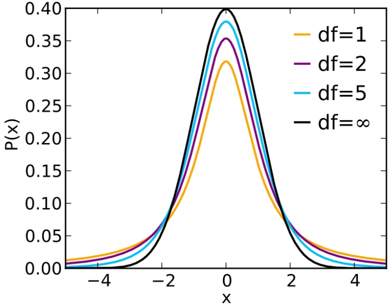
and: $Y_\nu \sim \chi^2(\nu)$

then:

$$T \sim \mathcal{T}(\nu)$$



Student's T distribution



$$t \sim T_\nu : f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

This is **Student's T distribution**.

- The parameter ν is called: "the degrees of freedom".
- The larger ν, the closer the T-distribution is to the Z-distribution.

Student's T distribution

Name:	Student's T	Models the distribution of the $t = \frac{\overline{X} - \mu}{s_x / \sqrt{n}}$
Support:	x	$(-\infty, \infty)$
Parameters:	$\nu \in (1, 2, 3, ..., \infty)$	degrees of freedom
pdf	$f(x\|n) =$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
mean	$\mathrm{E}(X)$	0
variance	$\mathrm{Var}(X)$	$\frac{\nu}{\nu-2}$
`R code`	`*t`	argument: `df`

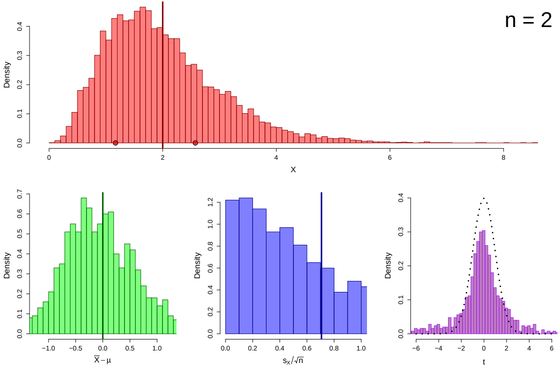
T-distribution: as statistic for comparing means

Testing: $H_0 : \mu_{obs} = \mu$

$$t_{obs} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(\nu = n-1)$$

compare to:

$$z_{obs} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$



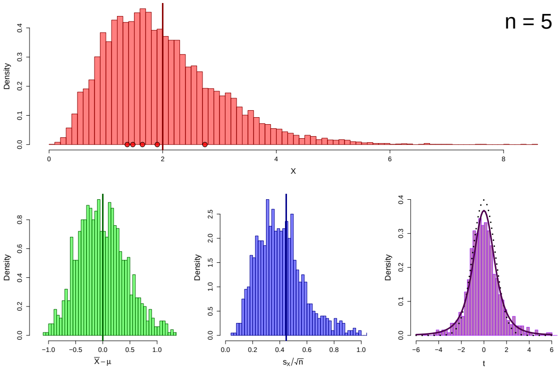
T-distribution: as statistic for comparing means

Testing: $H_0 : \mu_{obs} = \mu$

$$t_{obs} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(\nu = n-1)$$

compare to:

$$z_{obs} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$



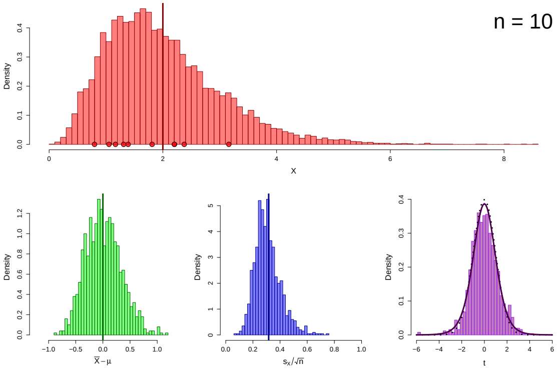
T-distribution: as statistic for comparing means

Testing: $H_0 : \mu_{obs} = \mu$

$$t_{obs} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(\nu = n - 1)$$

compare to:

$$z_{obs} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$



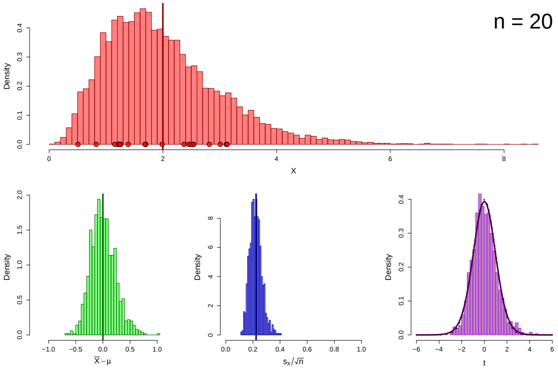
T-distribution: as statistic for comparing means

Testing: $H_0 : \mu_{obs} = \mu$

$$t_{obs} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(\nu = n-1)$$

compare to:

$$z_{obs} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$



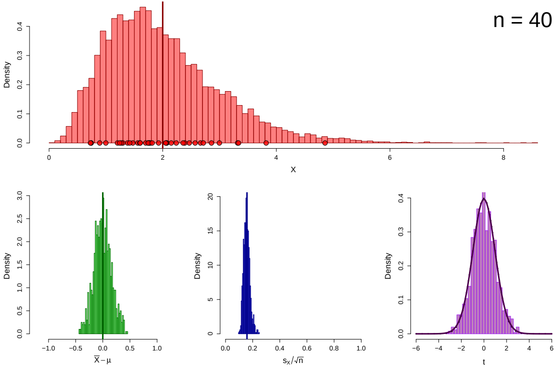
T-distribution: as statistic for comparing means

Testing: $H_0 : \mu_{obs} = \mu$

$$t_{obs} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(\nu = n - 1)$$

compare to:

$$z_{obs} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$



- As n increases, the distribution of $s_x/\sqrt{n}$ becomes very close to $\sigma_x/\sqrt{n}$, and the t-statistic becomes very similar to the z-statistic.
- This comparison of means can be easily extended to $H_0 : \mu_{obs} = \mu_0$, $H_0 : \mu_1 = \mu_2$ (equivalent to $H_0 : \mu_1 - \mu_2 = 0$), and paired T-tests

Examples to come

F-distribution: Ratio of χ^2

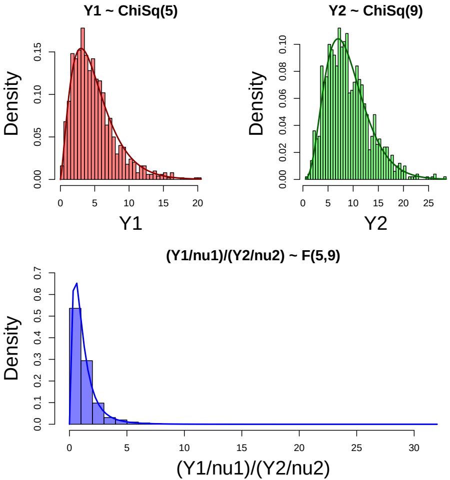
$$F = \frac{Y_{\nu_1}/\nu_1}{Y_{\nu_2}/\nu_2}$$

where: $Y_{\nu_1} \sim \chi^2(\nu_1)$

and: $Y_{\nu_2} \sim \chi^2(\nu_2)$

then:

$$F \sim \mathcal{F}(\nu_1, \nu_2)$$



where $\mathcal{F}(\nu)$ is Fisher's F-distribution with ν_1 and ν_2 degrees of freedom.

F-distribution

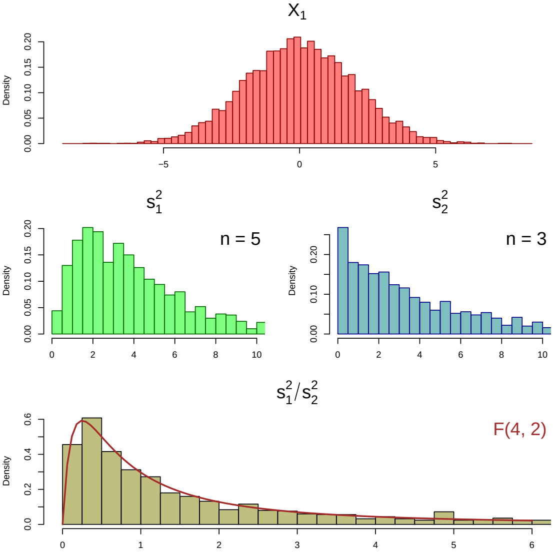
Name:	Fisher's F	Models: $F = s_1^2/s_2^2$
Support:	x	$(0, \infty)$
Parameters:	$\nu_1, \nu_2 \in (1, 2, 3, ..., \infty)$	degrees of freedom
pdf	$f(x\|\nu_1, \nu_2) =$	$\frac{\sqrt{\frac{(\nu_1\, x)^{\nu_1}\ \ \nu_2^{\nu_2}}{(\nu_1\, x+\nu_2)^{\nu_1+\nu_2}}}}{x\,\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}$
mean	$\mathrm{E}\,(X)$	$\frac{\nu_2}{\nu_2-2}$
variance	$\mathrm{Var}\,(X)$	$\frac{2\,\nu_2^2\,(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$
`R code`	`*f`	argument: `df1, df2`

The F-statistic for comparing sample variances

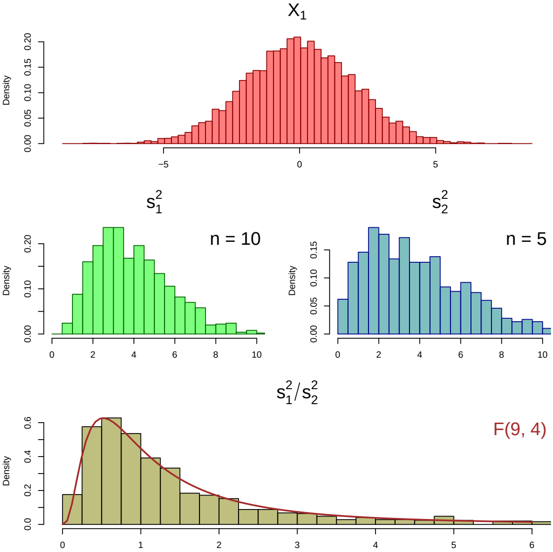
$$
\begin{aligned}
F_{obs} &= \frac{s_1^2}{s_2^2} \\
&\sim \mathcal{F}(\nu_1 = n_1 - 1, \nu_2 = n_2 - 1)
\end{aligned}
$$

The F-statistic allows us to compare two *sample variances*.

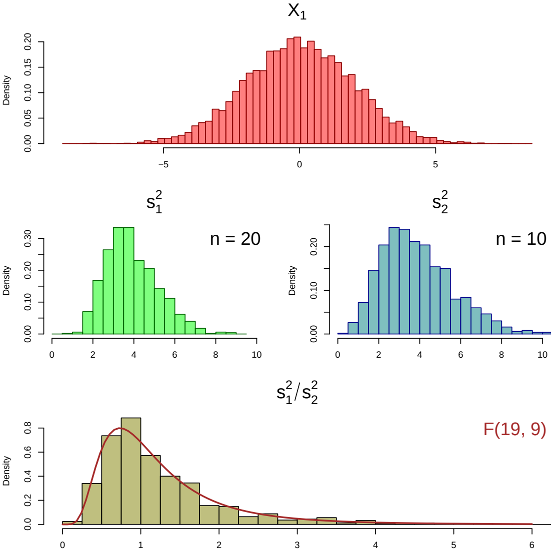
The F-statistic



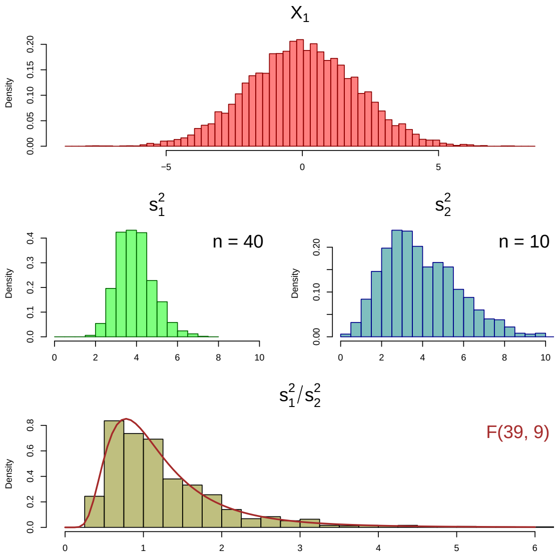
The F-statistic



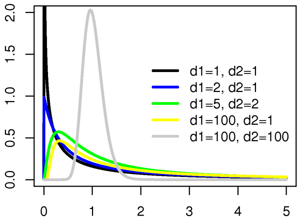
The F-statistic



The F-statistic

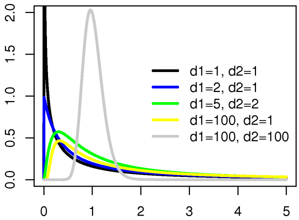


F-statistic



- F-distributions are right-skewed
- If $s_1 = s_2$, peak is near one (esp. with large sample sizes)
- Values very far from 1 are evidence against H_0

F-statistic



- F-distributions are right-skewed
- If $s_1 = s_2$, peak is near one (esp. with large sample sizes)
- Values very far from 1 are evidence against H_0