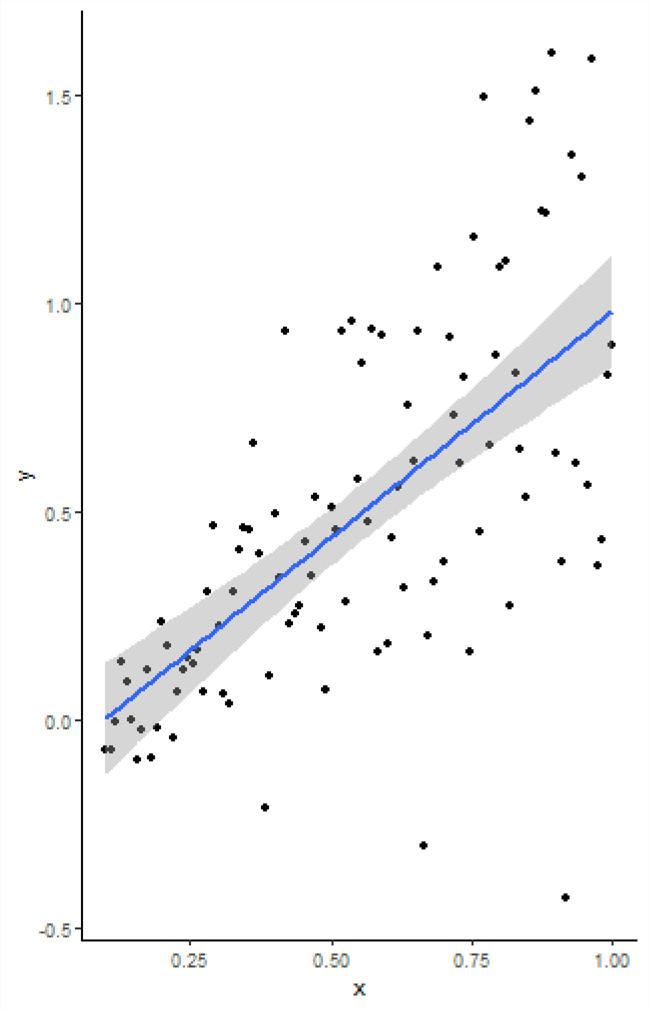
Introduction to Probability and Statistics III

Dustin Pluta

2019/04/01

STATS 120C is the last of a three-quarter series on introduction to probability and statistics. The goal of this course is to introduce basic principles of probability and statistical inference, and learn how these methods are applied to real world problems.

Topics that will be covered include



- Email: dpluta@uci.edu
- Website: https://github.com/dspluta/STAT120C
- Class Times: MWF 10 - 10:50am
- Room: ICS 174
- Office Hours: M 11:30am - 12:30pm; Th 2 - 3pm in DBH 2032
- TA Office Hours: W & Th 11am - 12pm in DBH 2013
- Discussion: MSTB 124
- Discussion Hours: W 5 - 5:50pm; 6 - 6:50pm

- : , 3rd Edition. John Rice. ISBN: 9788131519547.

- : Many examples and problems will be given in , which is available at http://www.r-project.org. I also recommend using the development environment RStudio: https://www.rstudio.com/. Please download and install and RStudio before the next class meeting.

- 30%: Eight (8) homework assignments
- 5%: Two (2) in-class quizzes
- 30%: Midterm Exam (Week 5)
- 35%: Final Exam (June 10th, 10:30am - 12:30pm)
- will be assigned on Monday and in the dropbox located near DBH 2013.

- Late homework will not be accepted!
- Exam make-ups will not be given except in case of emergency.
- One page of notes will be allowed for the midterm exam.
- Two pages of notes will be allowed for the final exam.
- Calculators will not be needed nor allowed.

- Review Distributions: Normal, t, chi-squared, F
- NHST, reference distributions, rejection regions
- Equivalence of t-test and Likelihood Ratio Test

- We will mainly focus on continuous distributions in this course.
- The of a random variable X is denoted $F(x)$, and is defined as the probability that $X < x$.

$$F(x) = P(X < x)$$

- The is denoted $f(x)$, and is defined as the rate of change of the cumulative probability at x,

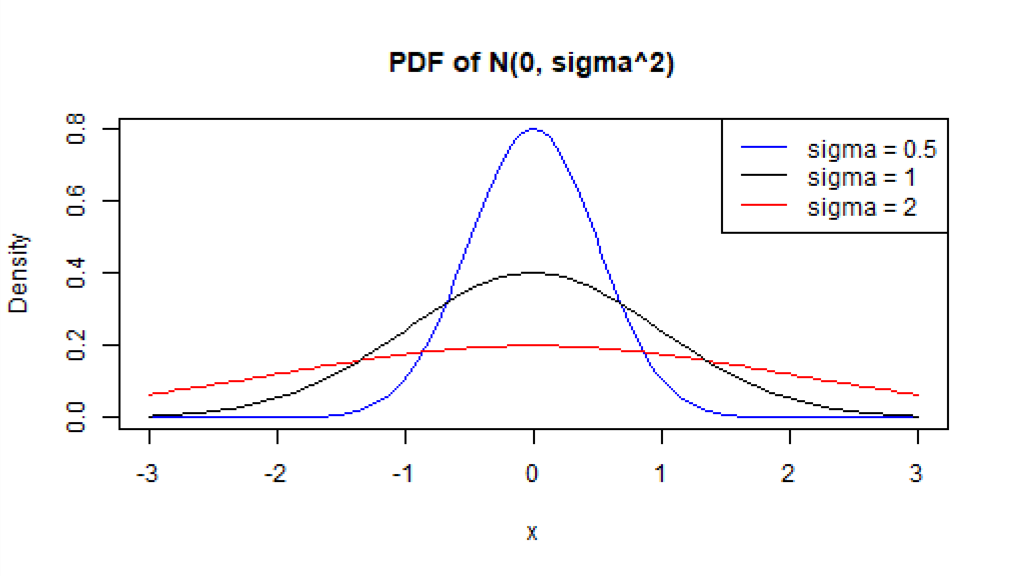
$$f(x) = F'(x).$$

- The of a random variable X is the set of all values for which $f(x) \neq 0$

$$\mathrm{Supp}(X) = \{x : f(x) \neq 0\}.$$

The probability density function of $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$



Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and let a, b be real constants.

1. $aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.
2. In particular, for $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$, we have

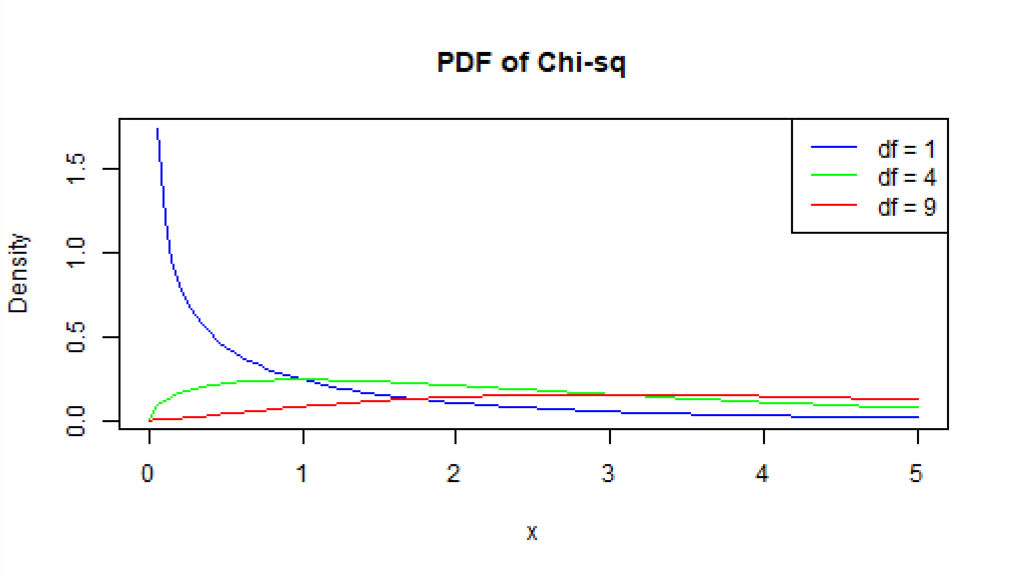
$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

χ^2

$X \sim \chi^2_n$ has pdf

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

The parameter n is the   of the distribution.



χ^2

The following is a key property of the χ^2 distribution that we will use repeatedly throughout the course:

For $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$,

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

.

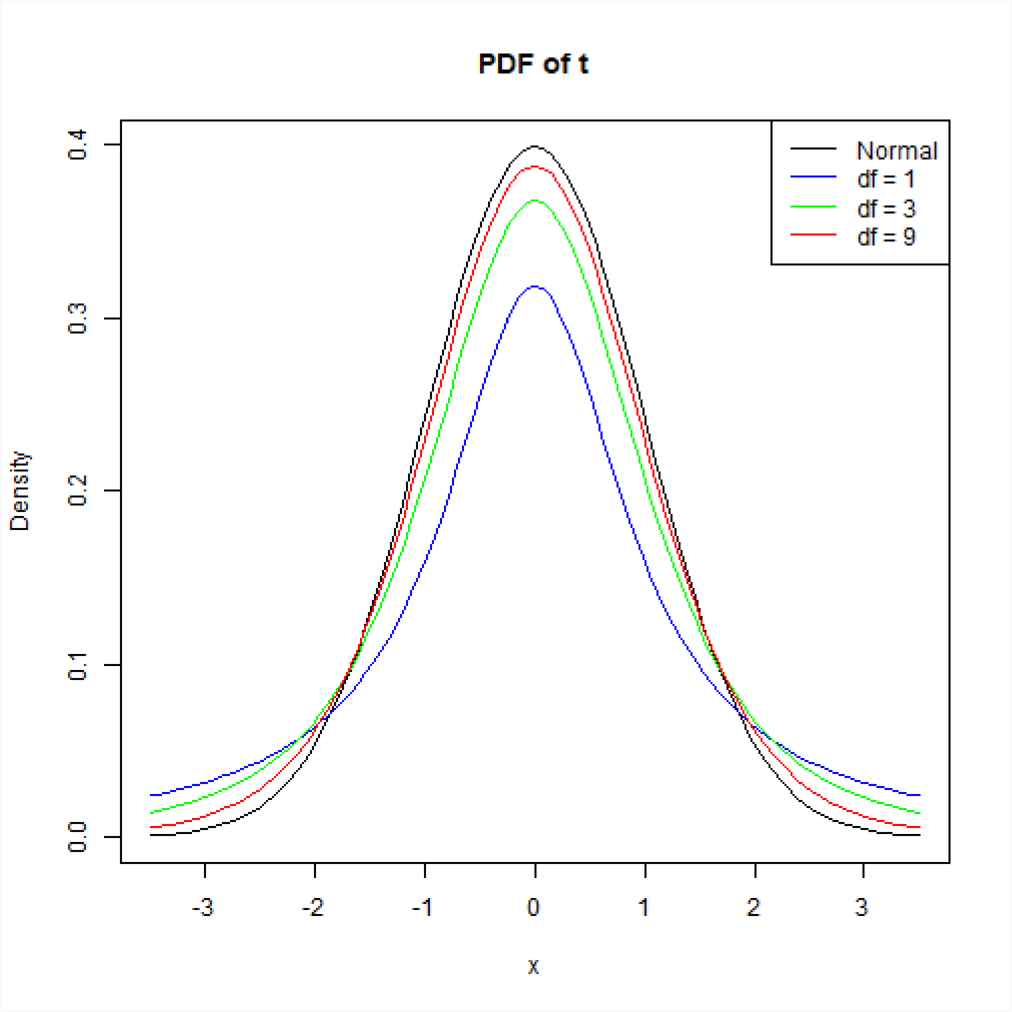
t

We will define the t distribution as a combination of a standard normal $Z \sim N(0, 1)$, and $V \sim \chi^2_v$:

$$T = \frac{Z}{\sqrt{V/v}} \sim t(v),$$

where v is the degrees of freedom of the distribution.

t



F

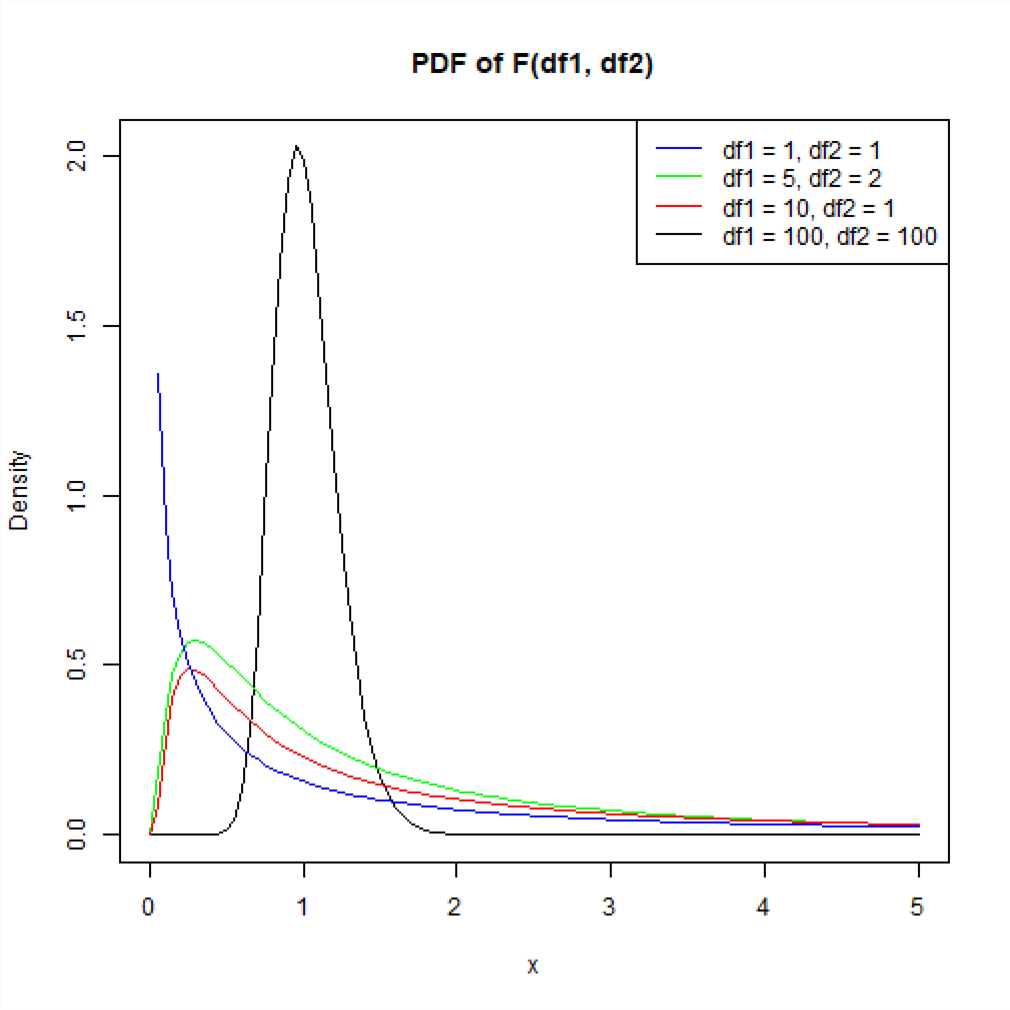
We will encounter the F distribution frequently throughout the course as well.

Let $U \sim \chi^2_u$ and $V \sim \chi^2_v$, with U and V independent. Then

$$X = \frac{U/u}{V/v} \sim F_{u,v},$$

where u and v are the degrees of freedom of the distribution.

F



PDF of F(df1, df2)
df1 = 1, df2 = 1
df1 = 5, df2 = 2
df1 = 10, df2 = 1
df1 = 100, df2 = 100
Density
x
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5

- : Make a binary (Yes/No) decision regarding some unknown quantity.
- : Estimate the value of some unknown quantity, and characterize the uncertainty in the estimate.
- : Predict the values of new observations from existing observations.

We will primarily focus on a review of hypothesis testing this week.

In general, a Null Hypothesis Significance Test (NHST) has the form

$$H_0 : \theta \in \Omega_0, \quad \textbf{(null hypothesis)}$$
$$H_1 : \theta \in \Omega_1, \quad \textbf{(alternative hypothesis)}$$

where $\Omega_0 \subset \mathbb{R}$ is the set of parameter values satisfying the null hypothesis, and similarly for Ω_1.

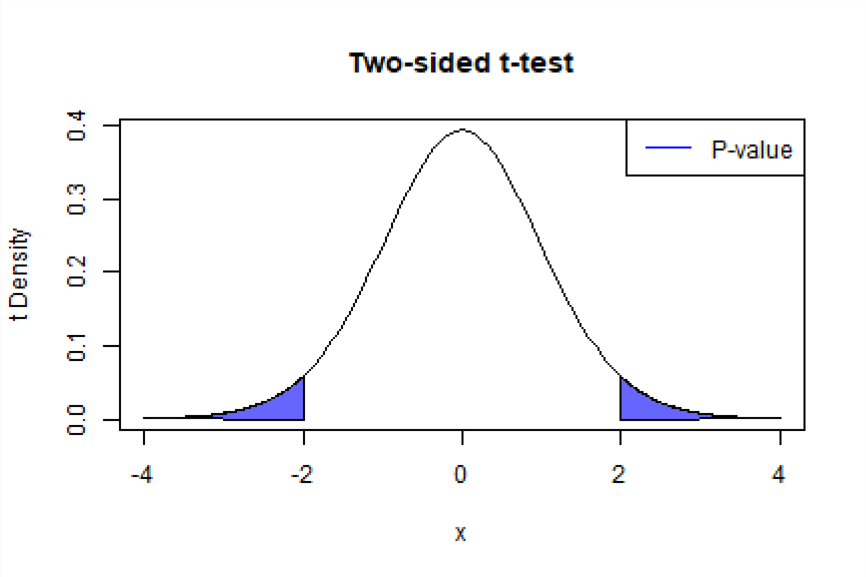
- When $\Omega_0 = \{\theta_0\}$ (contains a single value), then H_0 is $H_0 : \theta = \theta_0$, and is called a .
- If Ω_0 contains more than one value, H_0 is called a .

$$
\begin{aligned}
H_0 &: \theta \in \Omega_0, \quad \textbf{(null hypothesis)} \\
H_1 &: \theta \in \Omega_1, \quad \textbf{(alternative hypothesis)}
\end{aligned}
$$

at level of significance α, given a sample $X_1, \ldots, X_n$.

1. State the null and alternative hypotheses, the assumed sampling distribution of the data.
2. Choose an appropriate test statistic $T(X)$ for the null hypothesis.
3. Check model assumptions. (e.g. QQ-plot, histogram, scatterplot)
4. Compute the reference distribution and corresponding P-value for the test statistic.
5. Conclude one of:
    - $P \geq \alpha \rightarrow$ H_0: There is insufficient evidence to reject the null hypothesis at the α level of significance.
    - $P < \alpha \rightarrow$ H_0: There is sufficient evidence to reject the null hypothesis (and accept the alternative hypothesis) at the α level of significance.

The of a NHST is the probability of seeing a test statistic as extreme or more extremem than the observed test statistic, assuming the null hypothesis is true.



z

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ^2 known.

We wish to test the hypothesis

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0. \end{aligned}$$

z

We will test H_0 with test statistic $T(X) = \frac{\overline{X} - \mu_0}{\sigma}$.

"True" Distribution: $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$

Null Distribution: $\overline{X} \overset{H_0}{\sim} \mathcal{N}(\mu_0, \sigma^2/n)$

z

We will test H_0 with test statistic $T(X) = \frac{\overline{X} - \mu_0}{\sigma}$.

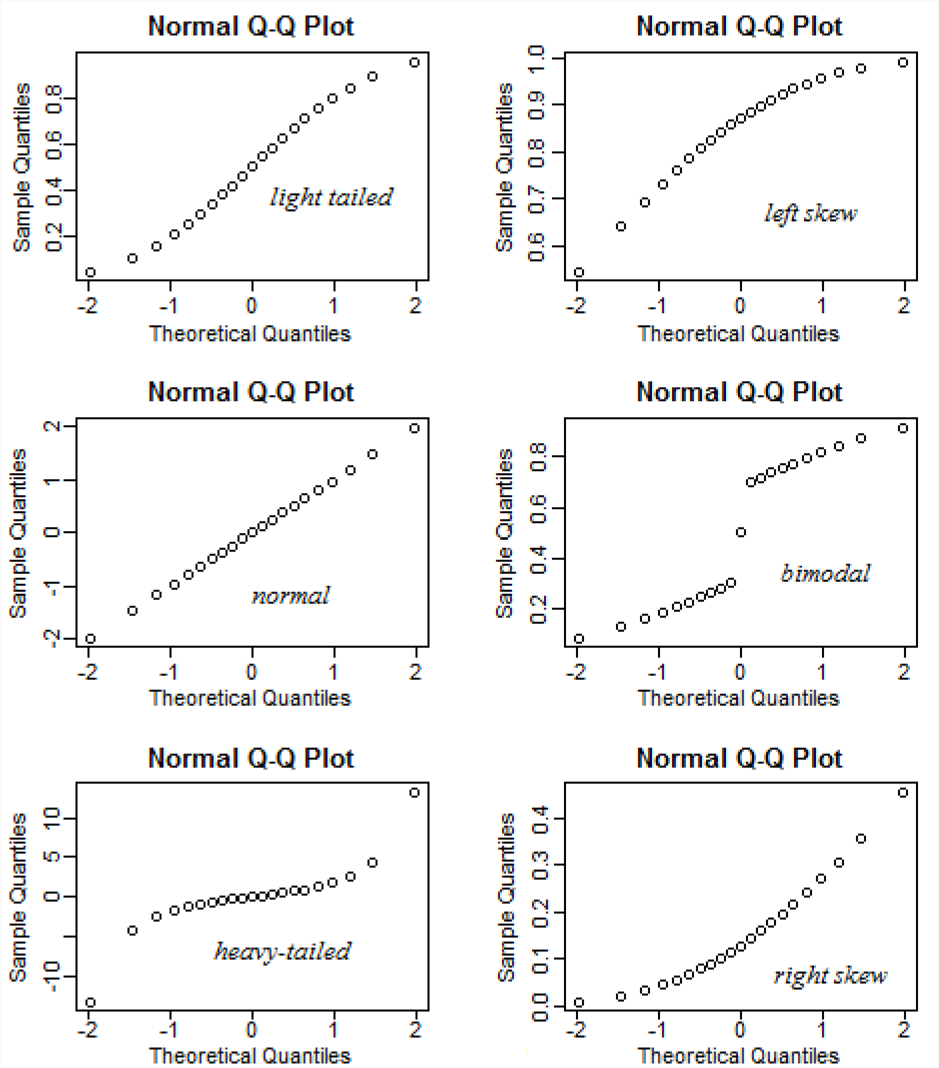
"True" Distribution: $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$

Null Distribution: $\overline{X} \overset{H_0}{\sim} \mathcal{N}(\mu_0, \sigma^2/n)$

A statistic based on $\overline{X}$ is a natural choice, and is also theoretically motivated, since it is

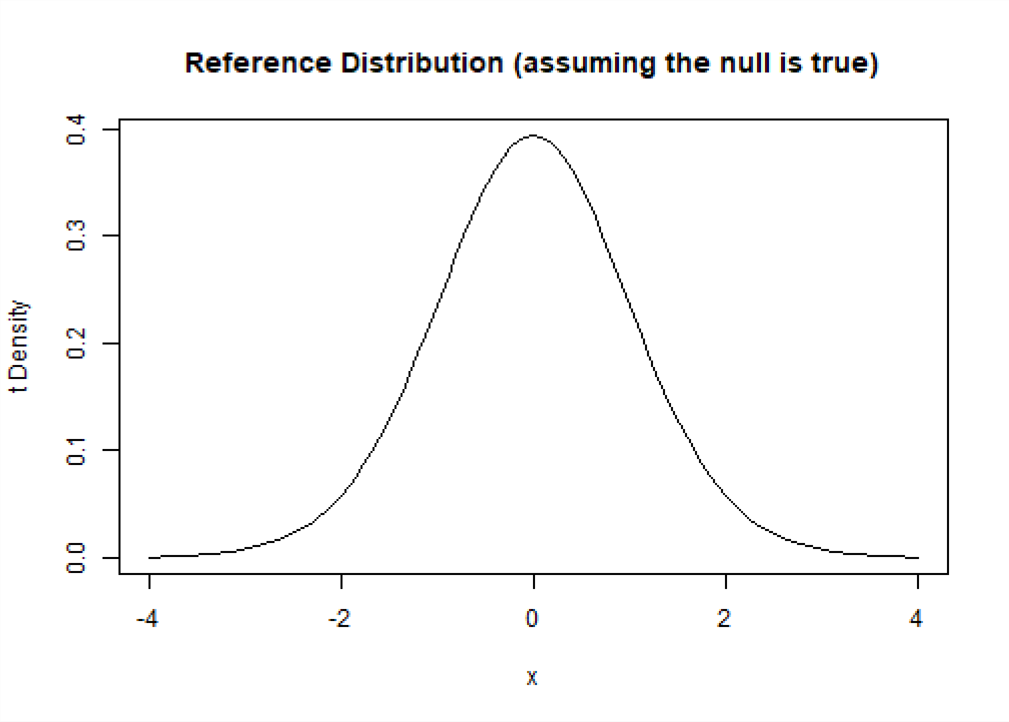
- The for μ
- The for μ
- More on this later...

Check model assumptions.

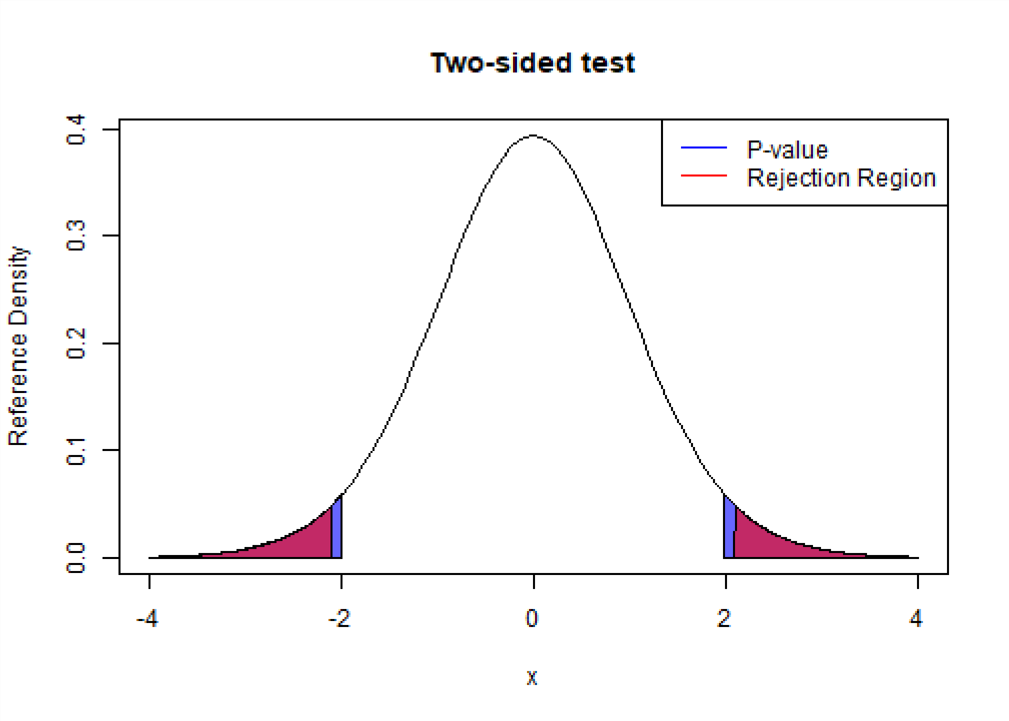


Compute reference distribution.

- Reference Distribution: $T(X) \overset{H_0}{\sim} \mathcal{N}(0,1)$



Make conclusion.



- $: \alpha = P(\text{Reject } H_0 | H_0 \text{ is True})$
- $: \beta = P(\text{Fail to reject } H_0 | H_0 \text{ is False})$
- $: 1 - \beta = P(\text{Reject } H_0 | H_0 \text{ is False})$

- In the NHST framework, α is selected by the researcher.
- Power is determined by the choice of α, as well as the sample size and the size of the effect being tested.

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to test $H_0 : \mu = \mu_0$. Assume σ^2 is known.
- Since $\bar{X}$ is an unbiased sufficient statistic for μ, we can use this estimator to construct our test statistic.
- We want to standardize the statistic to make it easy to compute the P-value.
- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2)$, so

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} \mathcal{N}(0, 1).$$

- Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to test $H_0 : \mu = \mu_0$. Assume σ^2 is known.
- We can again use $\overline{X}$ to construct our test statistic, but we must now also estimate σ^2.
- Use the sample variance estimator, which is unbiased for σ^2:

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

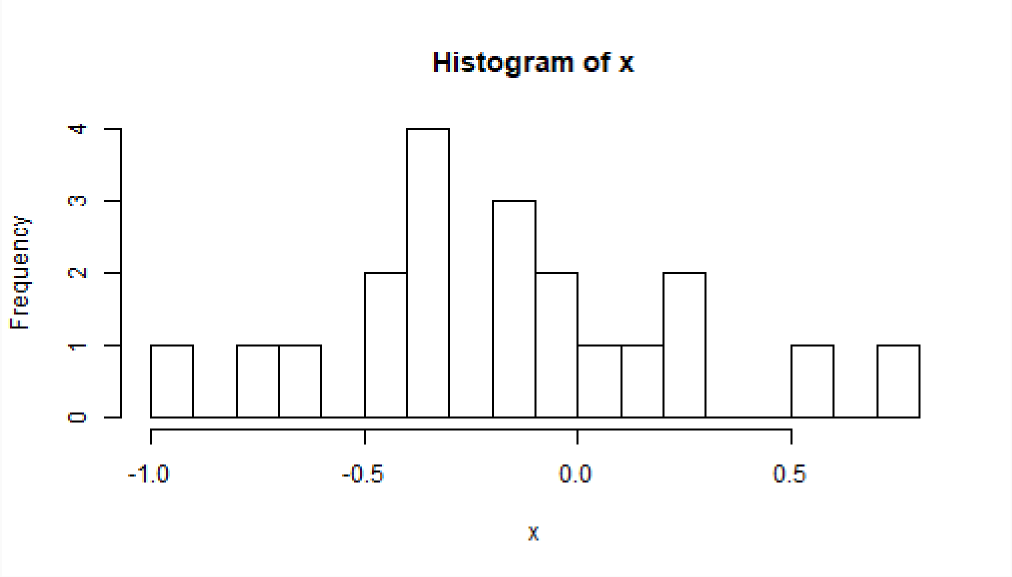
- Now consider test statistic $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. What is the reference distribution?

- What is the reference distribution of $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$?
- Recall that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$.
- We can rewrite T as

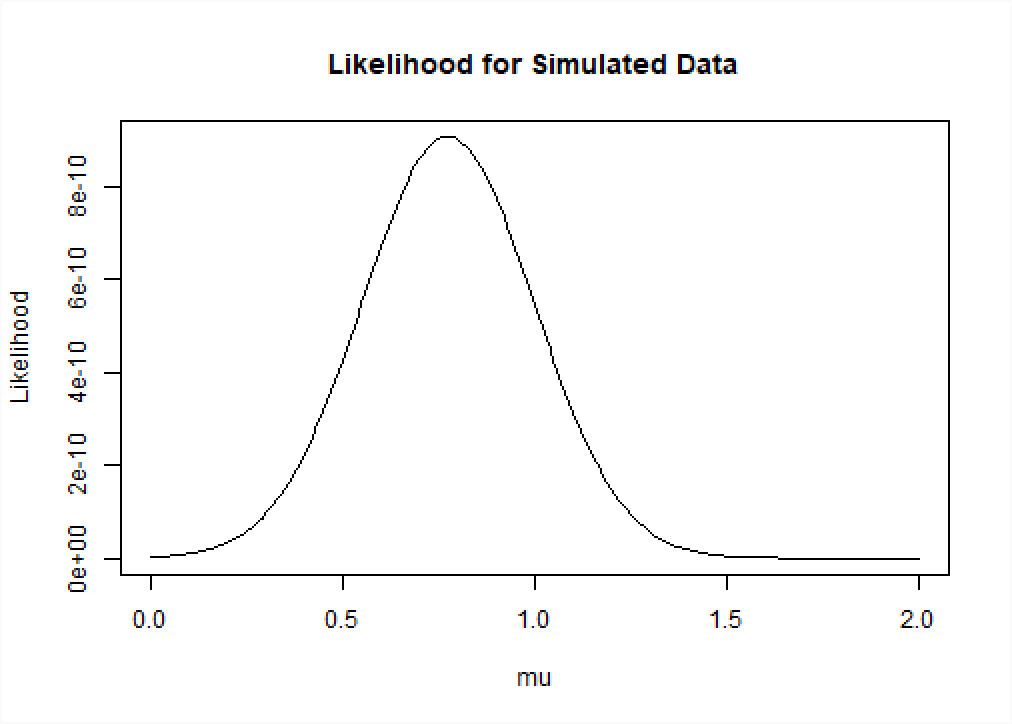
$$T = \frac{Z}{\sqrt{V/v}},$$

where $Z = \frac{(\overline{X} - \mu_0)}{\sigma}$ and $V = \frac{(n-1)s^2}{\sigma^2}$.

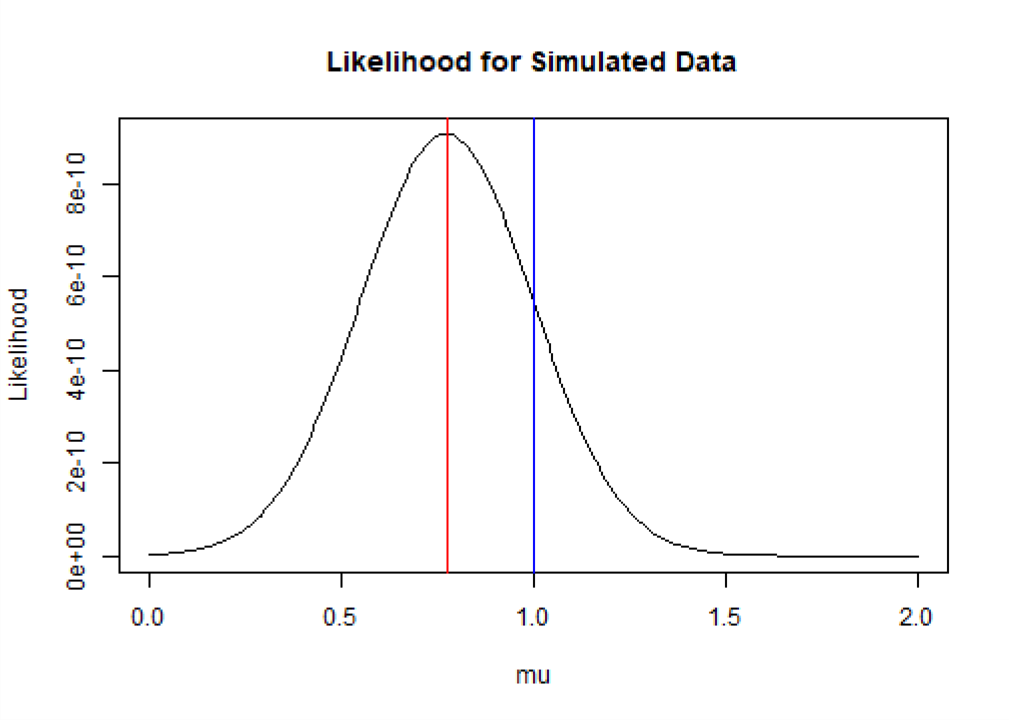
- Thus, $T \overset{H_0}{\sim} \chi^2_{n-1}$.


Histogram of x
Frequency
0
1
2
3
4
-1.0
-0.5
0.0
0.5
x

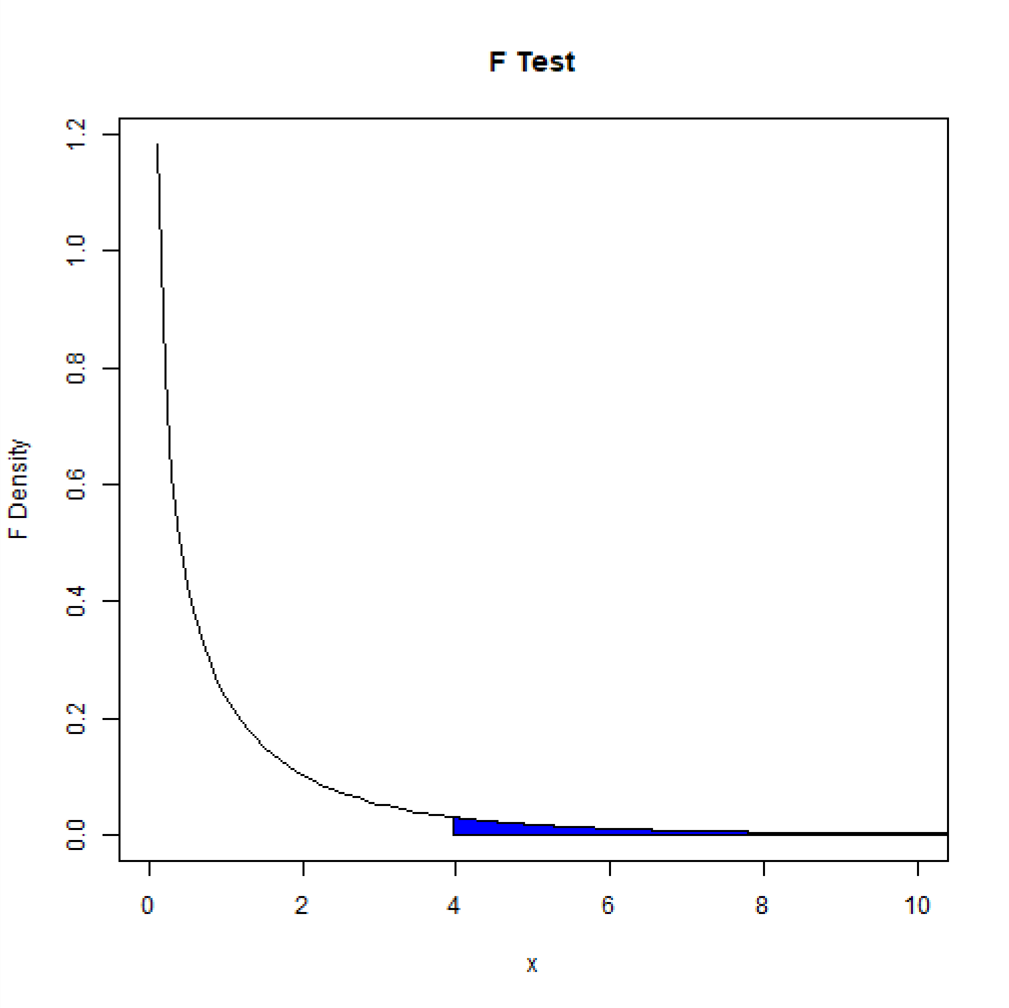
- We see that the t-test gives a larger P-value than what one would get from the normal distribution.
- If one incorrectly applies a z-test instead of a t-test, the Type I error will be inflated, especially for small sample sizes.

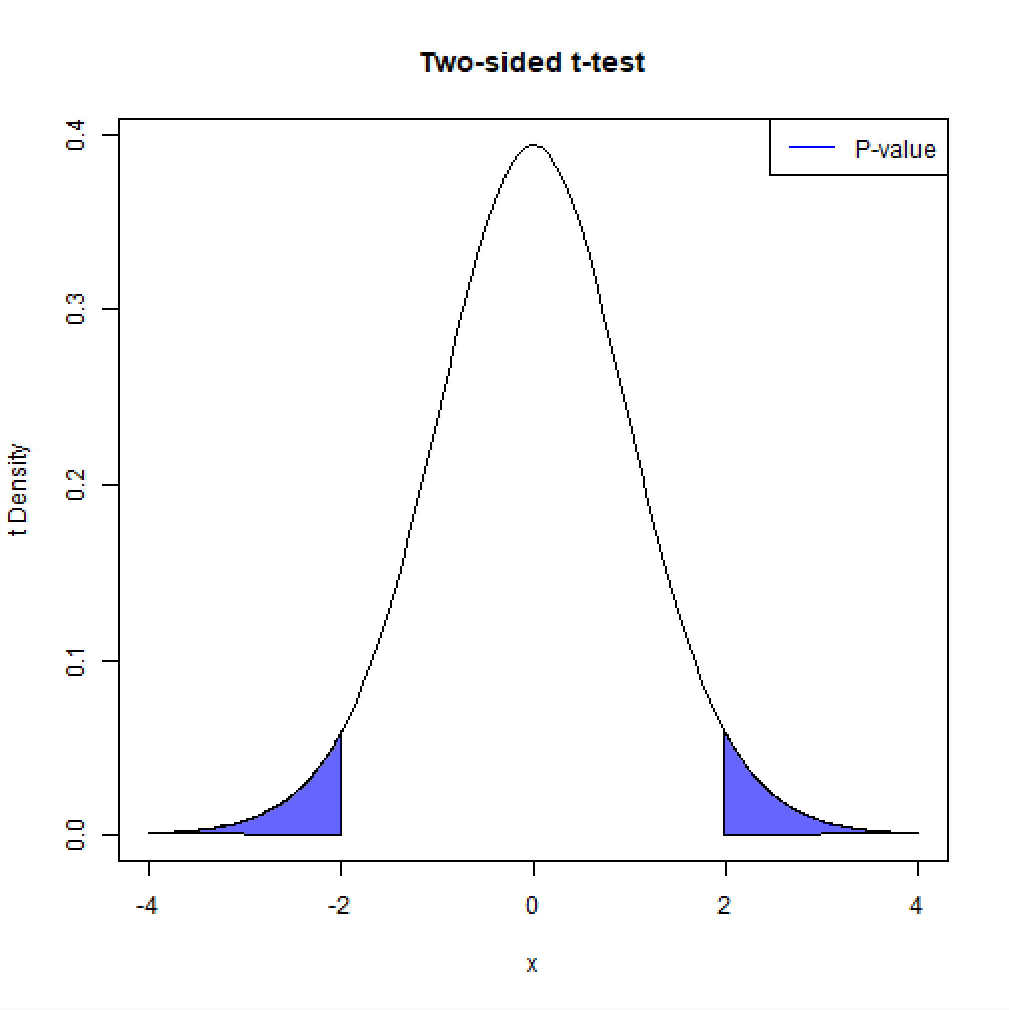

Likelihood for Simulated Data
Likelihood
0e+00
2e-10
4e-10
6e-10
8e-10
0.0
0.5
1.0
1.5
2.0
mu

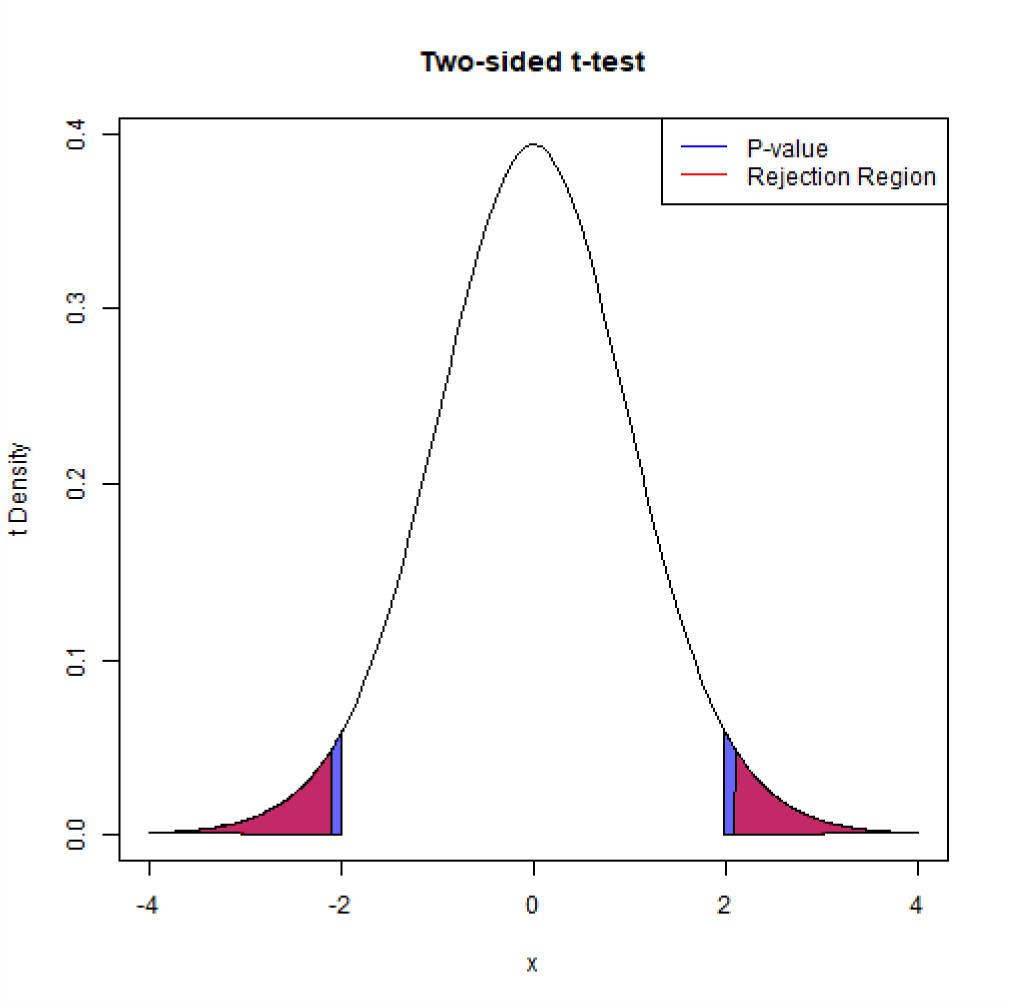
Suppose we want to test $H_0 : \mu = 1$ with the LRT.



Suppose we want to test $H_0 : \mu = 1$ using the LRT.


F Test
F Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
x


Two-sided t-test
P-value
t Density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
x


Two-sided t-test
P-value
Rejection Region
t Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x