STAT 120 C

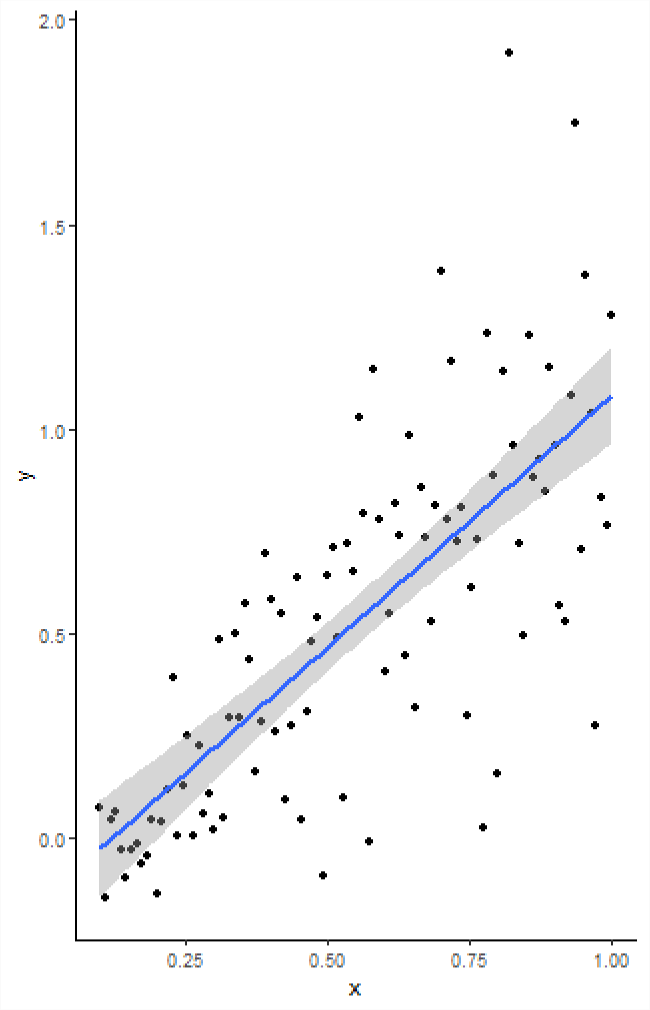
Introduction to Probability and Statistics III

Dustin Pluta
2019/04/01

Intro to STATS 120 C

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 **statistical hypothesis testing, linear regression, analysis of variance, and model checking.**



STAT 120C Course Info

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

STAT 120C Course Info

- **Text**: , 3rd Edition. John Rice. ISBN: 9788131519547.
- **Computing**: Many examples and problems will be given in `R`, which is available at http://www.r-project.org. I also recommend using the development environment RStudio: https://www.rstudio.com/.

STAT 120C Course Info

Grading

- 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)

- **Homework** will be assigned on Monday and **due the following Monday by 5pm** 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.

Week 1

Review of 120 B

- Sample mean and variance of normal distribution
- NHST, reference distributions, rejection regions
- Equivalence of t-test and Likelihood Ratio Test

Review of 120 B

Types of Problems in Statistics

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

Review of 120 B

Hypothesis Testing

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

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

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 .

Review of 120B

Null Hypothesis Significance Testing

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

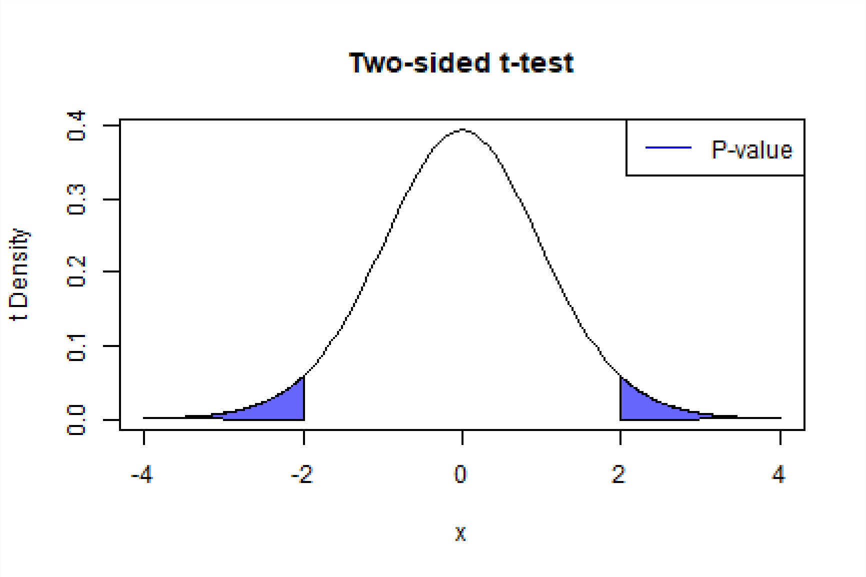
at level of significance α, given a sample $X_1, \ldots, X_n$.

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:
   - $P < \alpha \rightarrow$ **Fail to reject** H_0: There is insufficient evidence to reject the null hypothesis at the α level of significance.
   - $P \geq \alpha \rightarrow$ **Reject** H_0: There is sufficient evidence to reject the null hypothesis (and accept the alternative hypothesis) at the α level of significance.

Review of 120B

Definition: P-value

The **P-value** 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.



Review of 120B

Normal Distribution: PDF

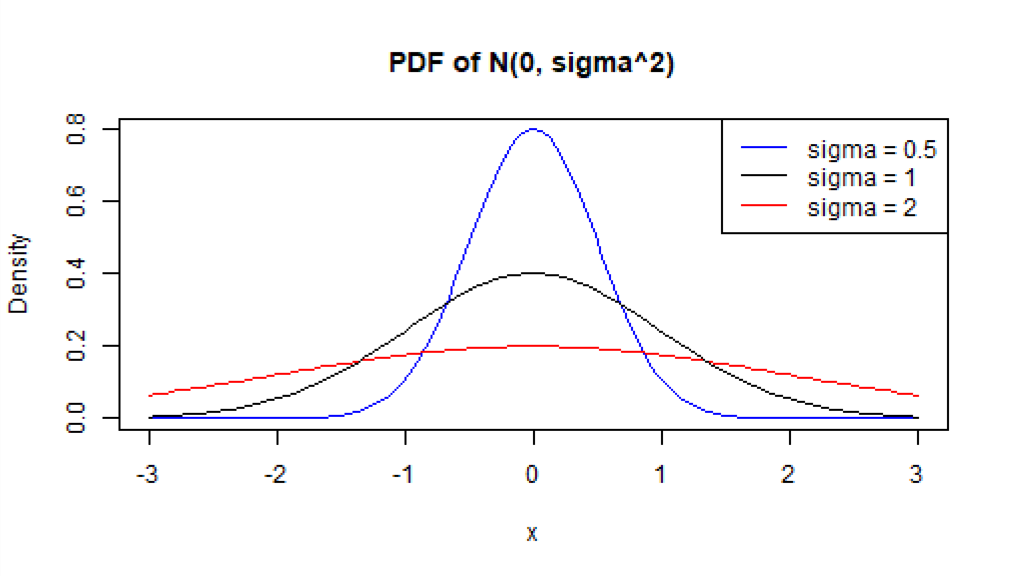
The probability density function of $X \sim \mathcal{N}(\mu, \sigma^2)$ is

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

Review of 120B

Normal Distribution: PDF

```
x <- curve(dnorm(x, mean = 0, sd = 0.5),
           from = -3, to = 3, ylab = "Density",
           col = "blue", main = "PDF of N(0, sigma^2)")
curve(dnorm(x, mean = 0, sd = 2), add = T, col = "red")
curve(dnorm(x, mean = 0, sd = 1), add = T)
legend("topright", legend = c("sigma = 0.5", "sigma = 1", "sigma = 2"),
       col = c("blue", "black", "red"), lty = 1)
```



Review of 120B

Sums of Normally Distributed Variables

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

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

Review of 120B

One Sample z-test

Step 1

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

We wish to test the hypothesis

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

Review of 120B

One Sample z-test

Step 2

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

Review of 120B

One Sample z-test

Step 2

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

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

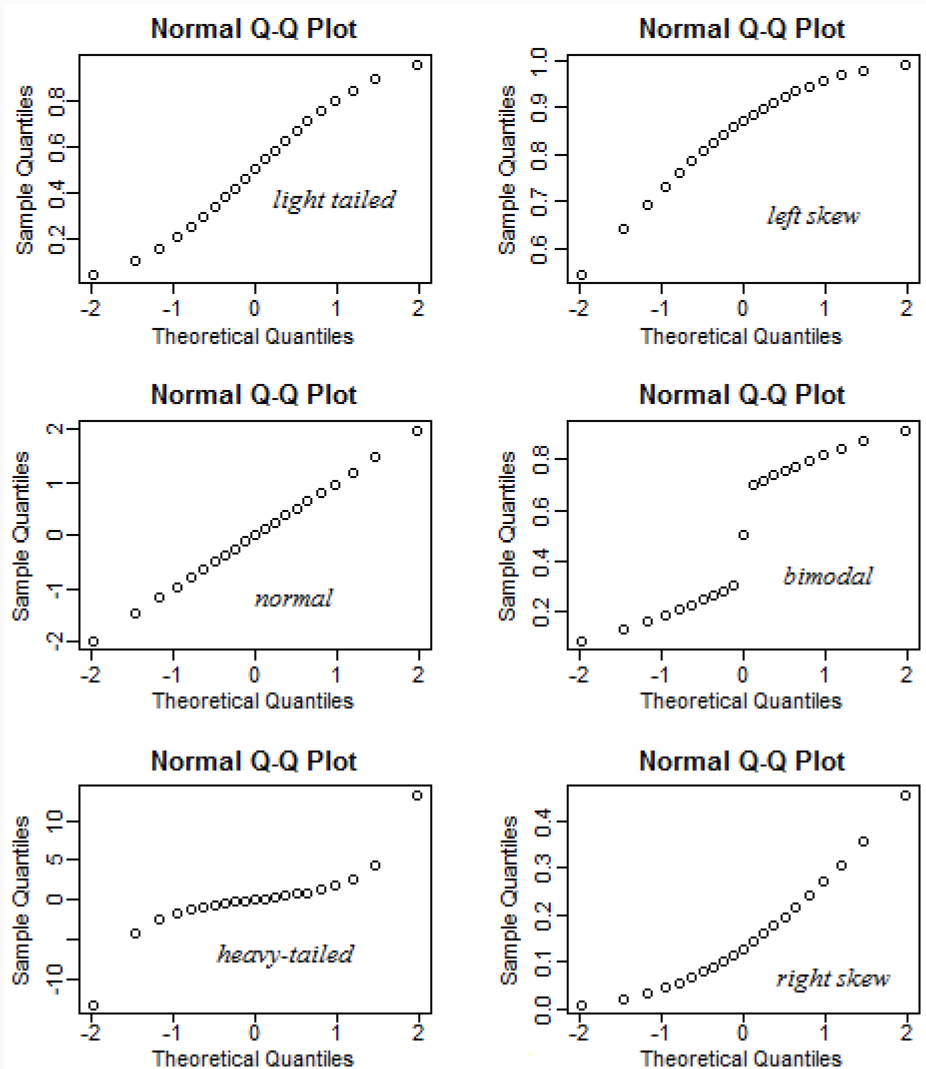
Null Distribution: $\overline{X} \overset{H_0}{\sim} \mathcal{N}(\mu_0, \sigma^2/n)$

Note: 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...

Review of 120B

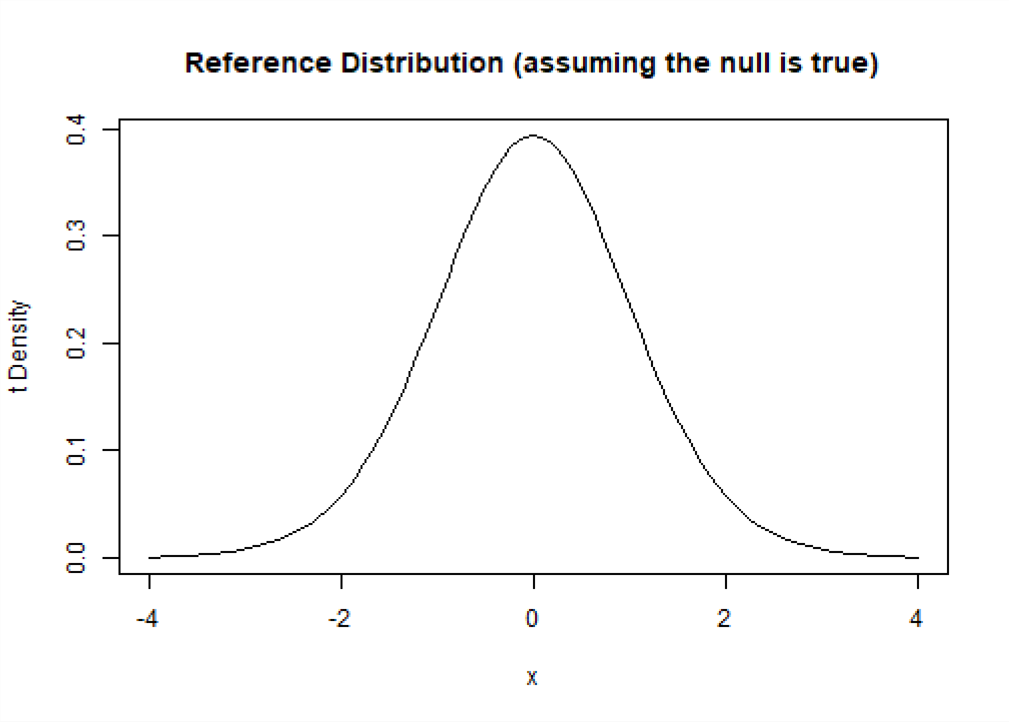
Step 3 Check model assumptions.



Review of 120B

Step 4 Compute reference distribution.

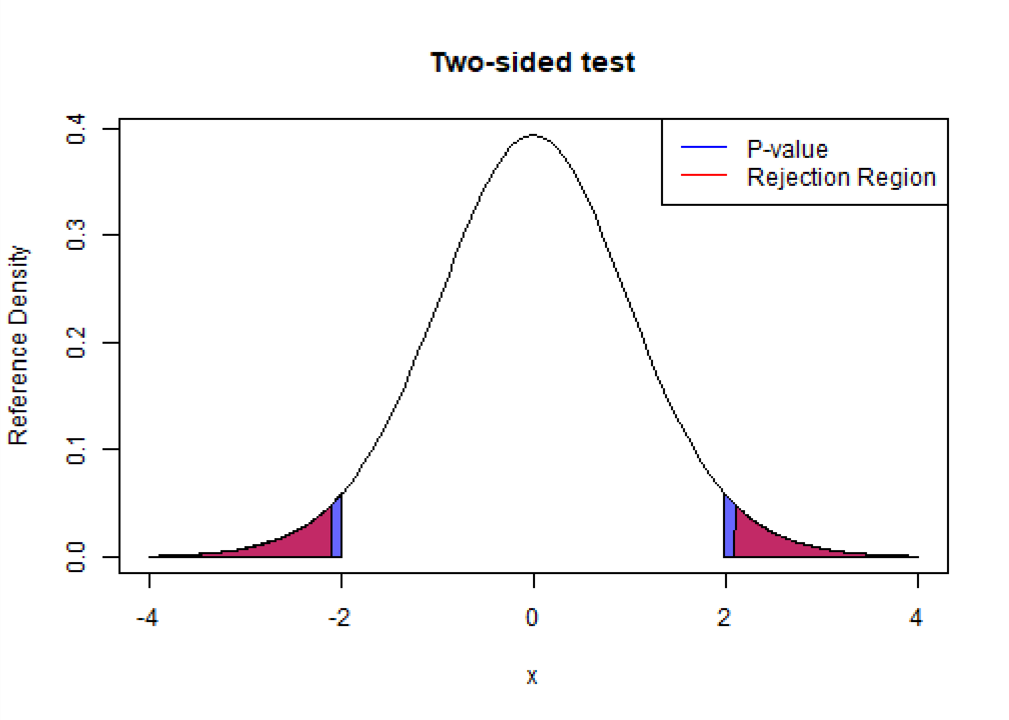
- Reference Distribution: $T(X) \overset{H_0}{\sim} \mathcal{N}(0,1)$



Review of 120B

Step 5 Make conclusion.

- $P = 1 - \Phi\left(\frac{\overline{X} - \mu_0}{\sigma}\right)$



Review of 120B

Hypothesis Testing Terminology

- **Type I Error**: $\alpha = P(\text{Reject } H_0 | H_0 \text{ is True})$
- **Type II Error**: $\beta = P(\text{Fail to reject } H_0 | H_0 \text{ is False})$
- **Power**: $1 - \beta = P(\text{Reject } H_0 | H_0 \text{ is False})$

Remarks

- 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.

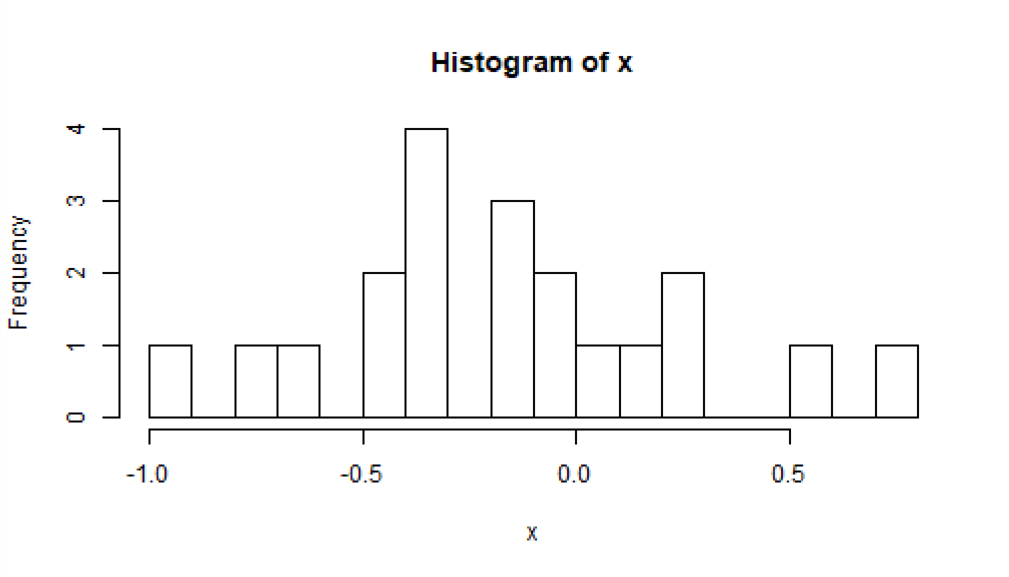
Review of 120B

Deriving the one-sample t-test

What about the case when σ^2 is unknown?

Example: One-sample t-test

```
set.seed(12)
n <- 20
mu <- 0
sigma <- 0.5
x <- rnorm(n, mu, sigma)
hist(x, breaks = 20)
```



Review of 120B

Example: One-sample t-test

```
x_bar <- sum(x) / n
s <- sqrt(sum((x - x_bar)^2) / (n - 1))
print(x_bar)

## [1] -0.1656045

print(s)

## [1] 0.4334393

test_stat <- (x_bar - 0) / (s / sqrt(n))
print(test_stat)

## [1] -1.708673
```

Review of 120B

Example: One-sample t-test

```
pnorm(test_stat)
```

```
## [1] 0.04375576
```

```
pt(test_stat, df = n - 1)
```

```
## [1] 0.05189839
```

- 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.

Review of 120B

Likelihood Ratio Test

Review of 120B

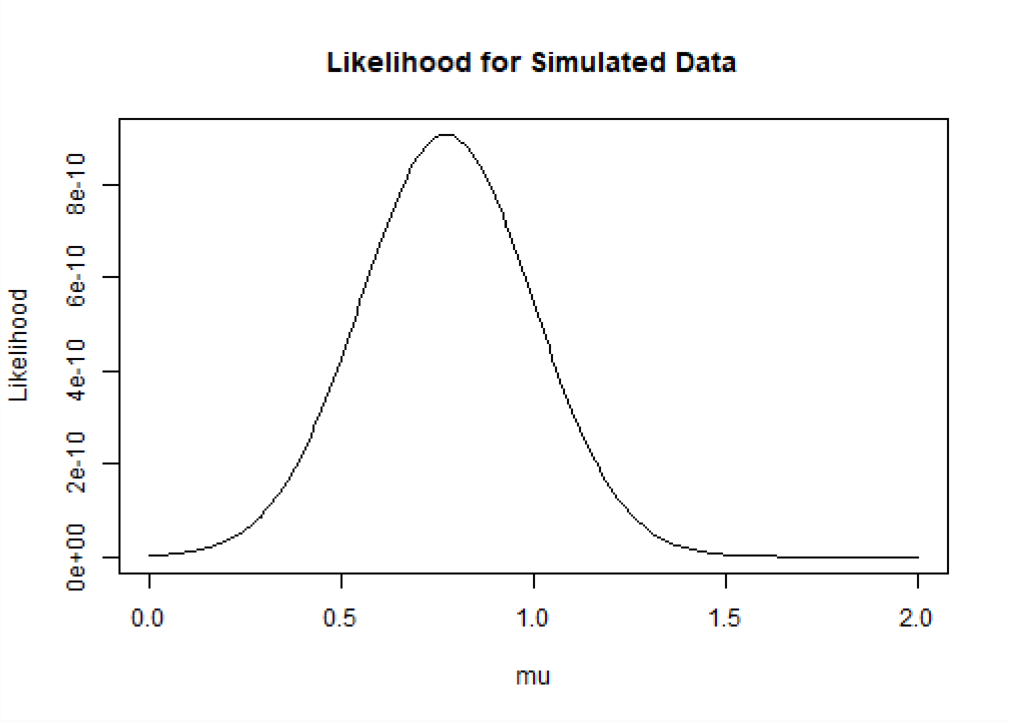
Likelihood Ratio Test

```r
set.seed(1234)
n <- 20
mu <- 0.9
sigma <- 0.5
x <- rnorm(n, mu, sigma)

lik <- function(mu, sigma = 1) {
  (2 * pi * sigma^2)^(-n / 2) * exp(- 1 / (2 * sigma^2) * sum((x - mu)^2))
}
mu_seq <- seq(0, 2, 0.01)
lik_vals <- sapply(X = mu_seq, FUN = lik)
```

Likelihood Ratio Test

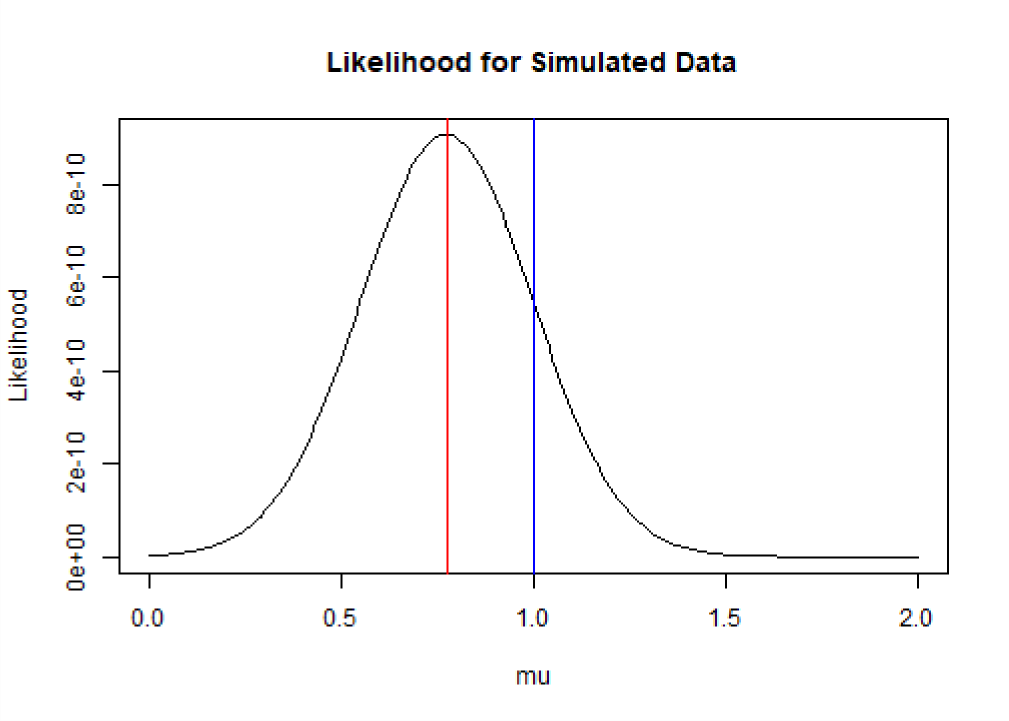
```
plot(mu_seq, lik_vals, ty = "l", ylab = "Likelihood", xlab = "mu",
     main = "Likelihood for Simulated Data")
```



Likelihood Ratio Test

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

```
plot(mu_seq, lik_vals, ty = "l", ylab = "Likelihood", xlab = "mu",
     main = "Likelihood for Simulated Data")
abline(v = 1, col = "blue")
abline(v = mean(x), col = "red")
```



Review of 120B

Likelihood Ratio Test

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

```
set.seed(1234)
n ← 20
mu ← 0.9
sigma ← 0.5
x ← rnorm(n, mu, sigma)

mu_0 ← 1
s_0 ← sqrt(1 / n * sum((x - mu_0)^2))
mu_hat ← mean(x)
s ← sqrt(1 / (n - 1) * sum((x - mu_hat)^2))

F_stat ← n * (mu_hat - mu_0)^2 / s^2
P_val ← pf(F_stat, df1 = 1, df2 = n - 1, lower.tail = FALSE)
P_val
```

```
## [1] 0.06141741
```

Review of 120B

Likelihood Ratio Test

```
t.test(x = x, mu = 1)
```

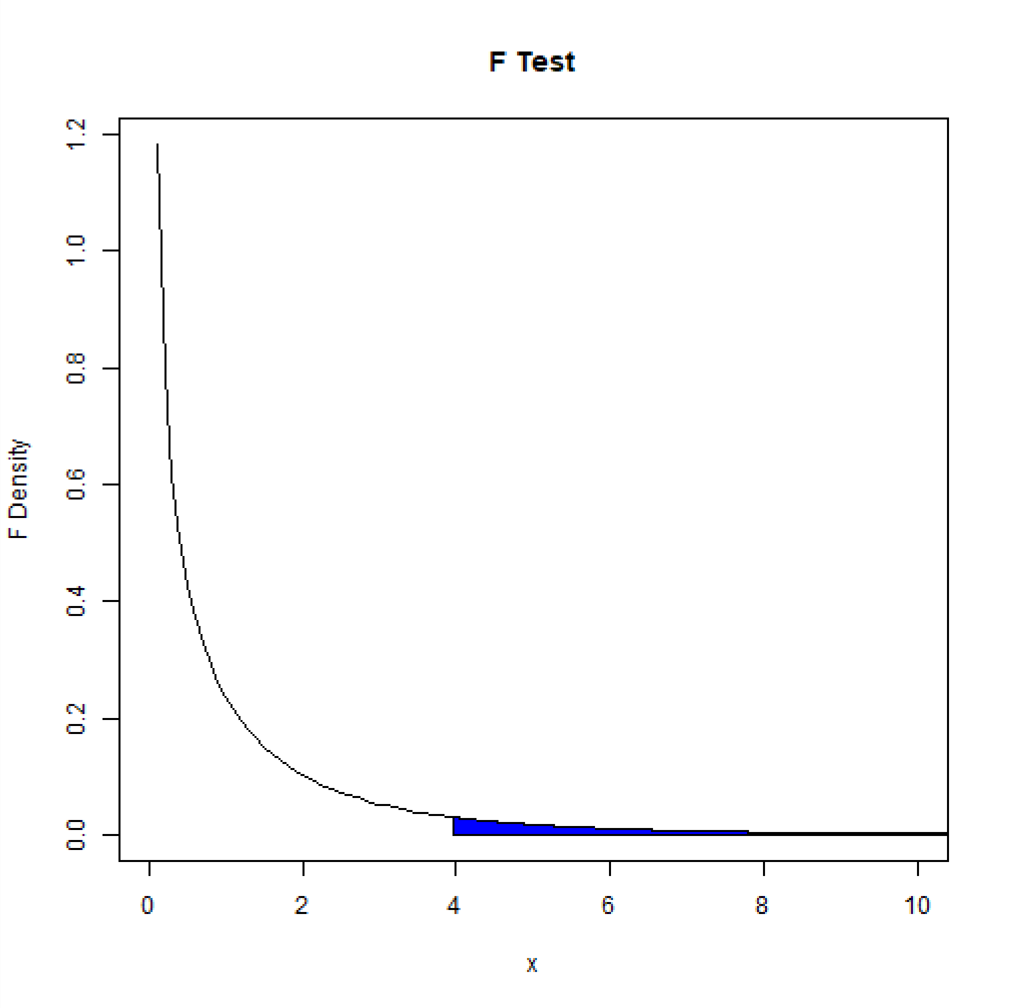
```
##
##     One Sample t-test
##
## data:  x
## t = -1.988, df = 19, p-value = 0.06142
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  0.5374297 1.0119063
## sample estimates:
## mean of x
##  0.774668
```

```
T_stat <- (mu_hat - mu_0) / sqrt(s^2 / n)
2 * pt(T_stat, df = n - 1, lower.tail = T)
```

```
## [1] 0.06141741
```

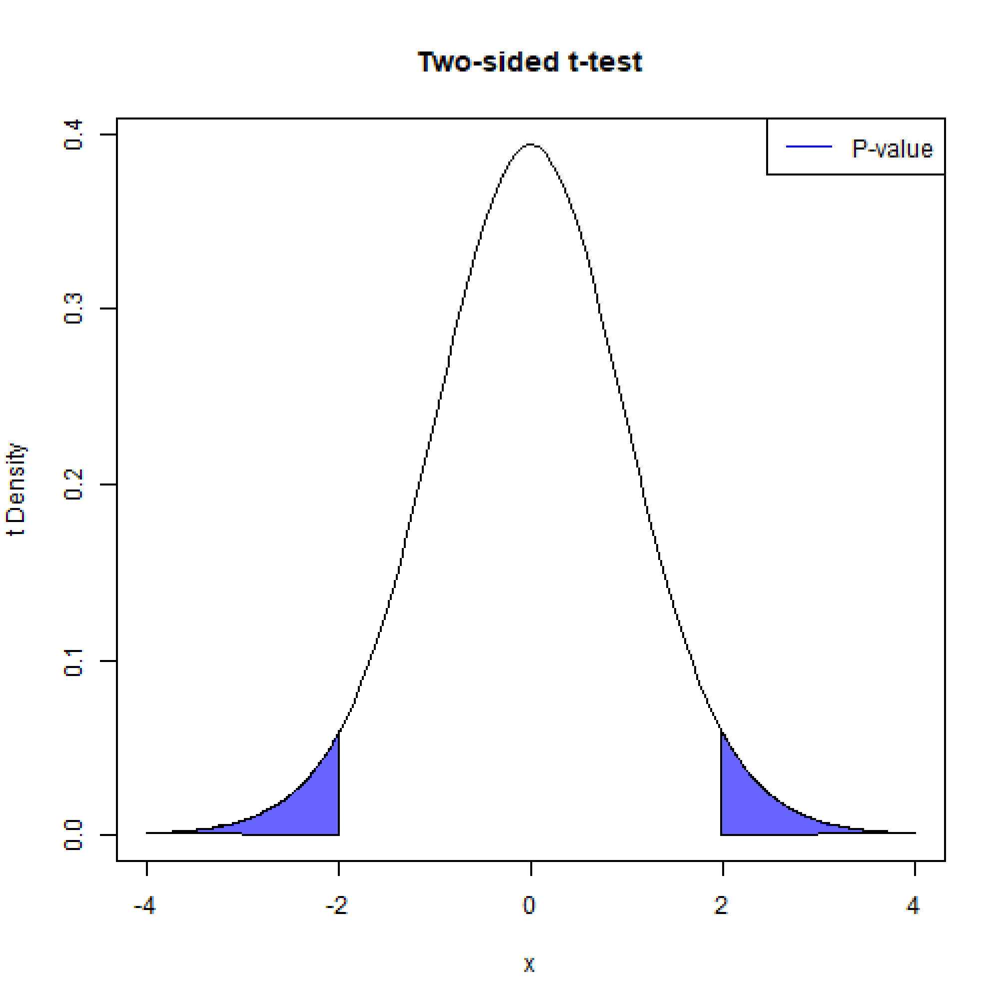
Review of 120B

Likelihood Ratio Test



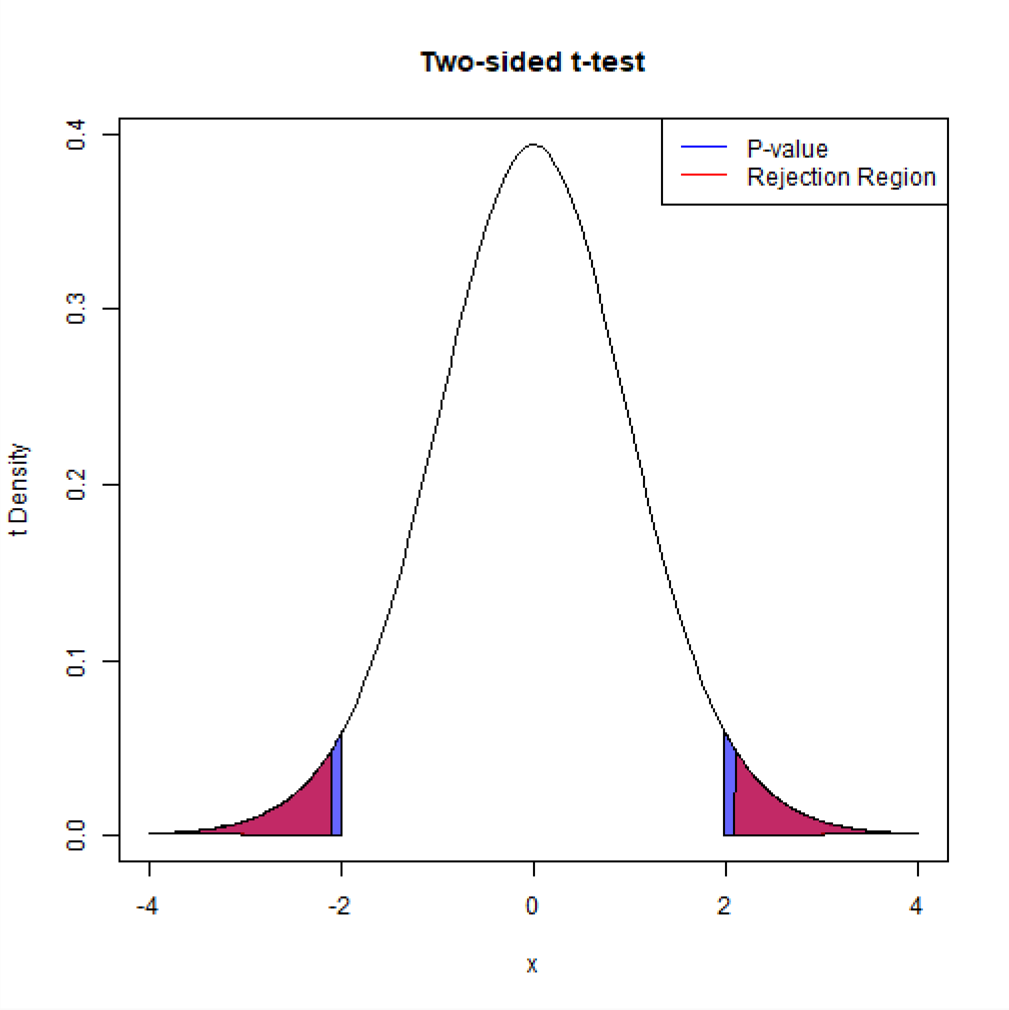
Review of 120B

Likelihood Ratio Test



Review of 120B

Likelihood Ratio Test



Review of 120B

#