Augusta University: STAT 7630

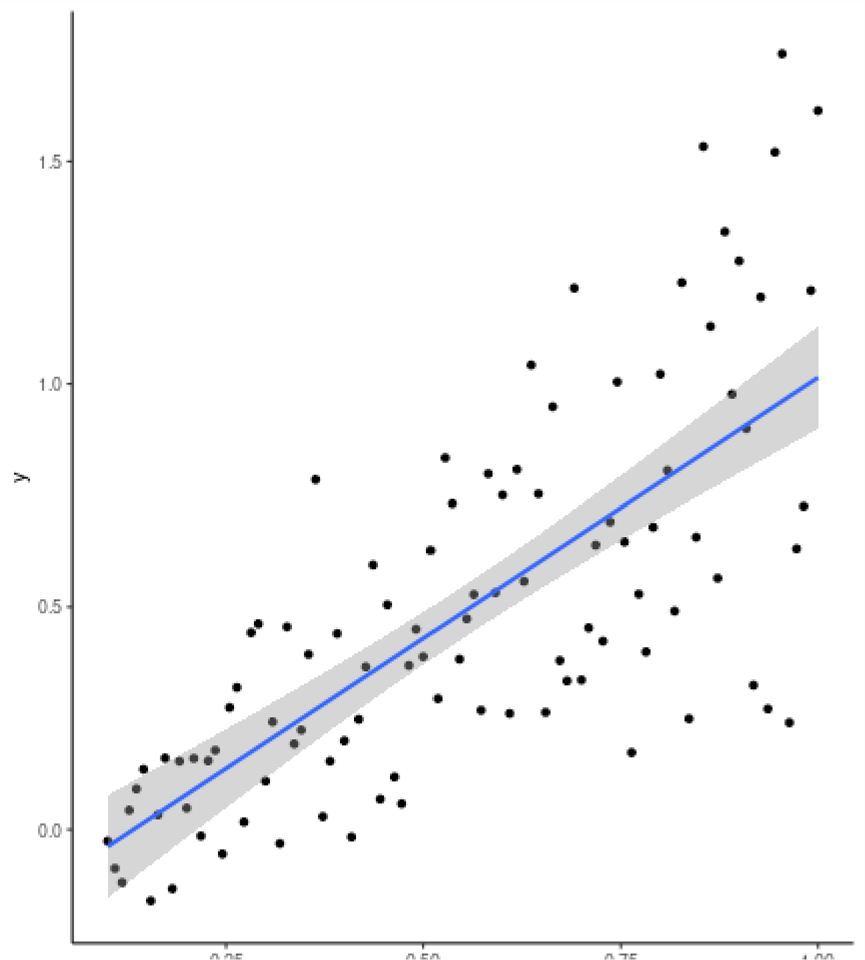
Applied Linear Models

Dustin Pluta

2024 JAN 09

Course Introduction

- **Website:** https://github.com/dspluta/STAT7630_SPRING2024
- **Syllabus:** https://github.com/dspluta/STAT7630_SPRING2024/Syllabus.pdf
- **Instructor email:** dpluta@augusta.edu



Review

Distributions

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

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

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

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

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

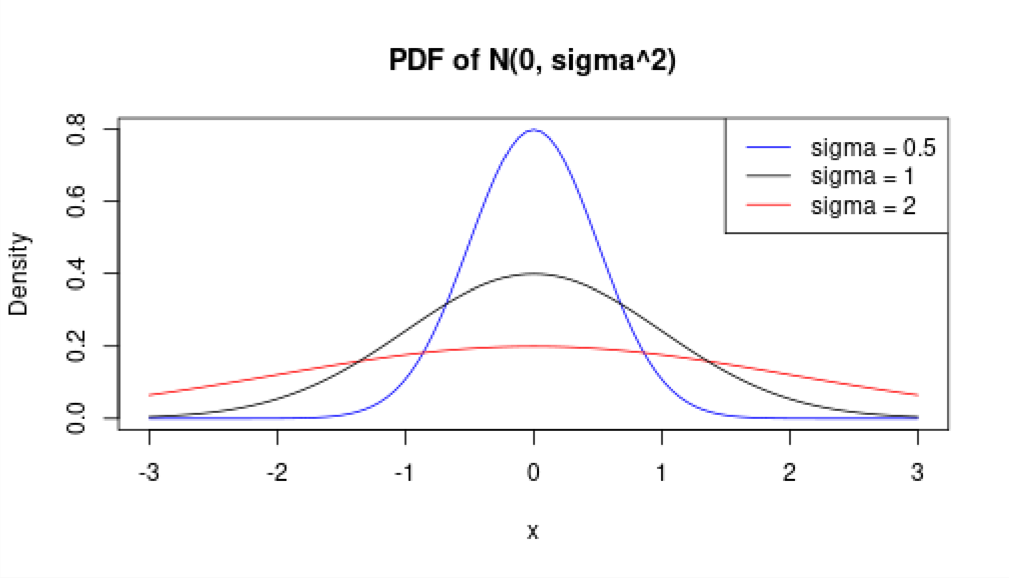
$$\text{Supp}(X) = \{x : f(x) \neq 0\}.$$

Review

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

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

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

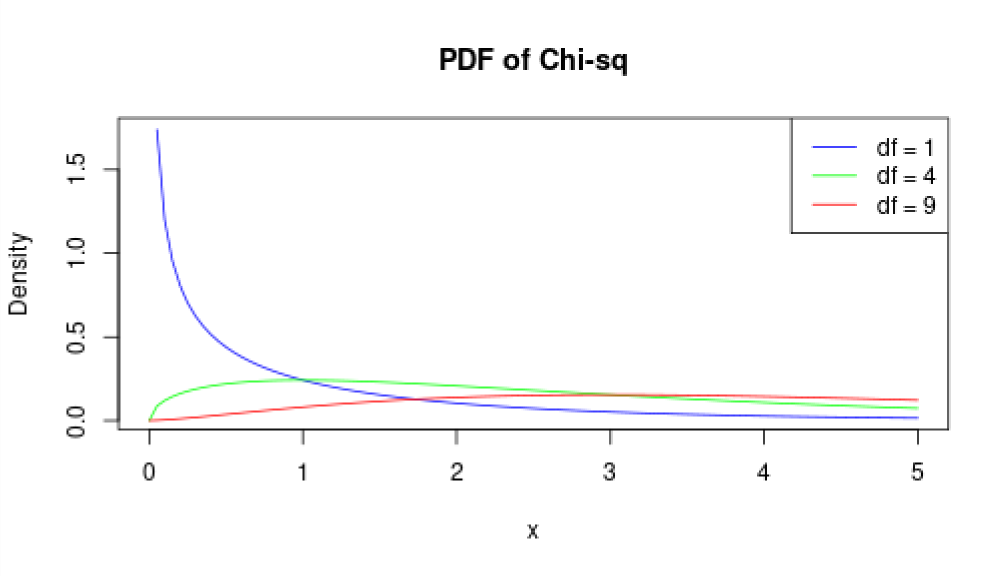
Review

χ^2 Distribution

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

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

The parameter n is the *degrees of freedom* of the distribution.



Review

χ^2 Distribution

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

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

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

.

Review

t Distribution

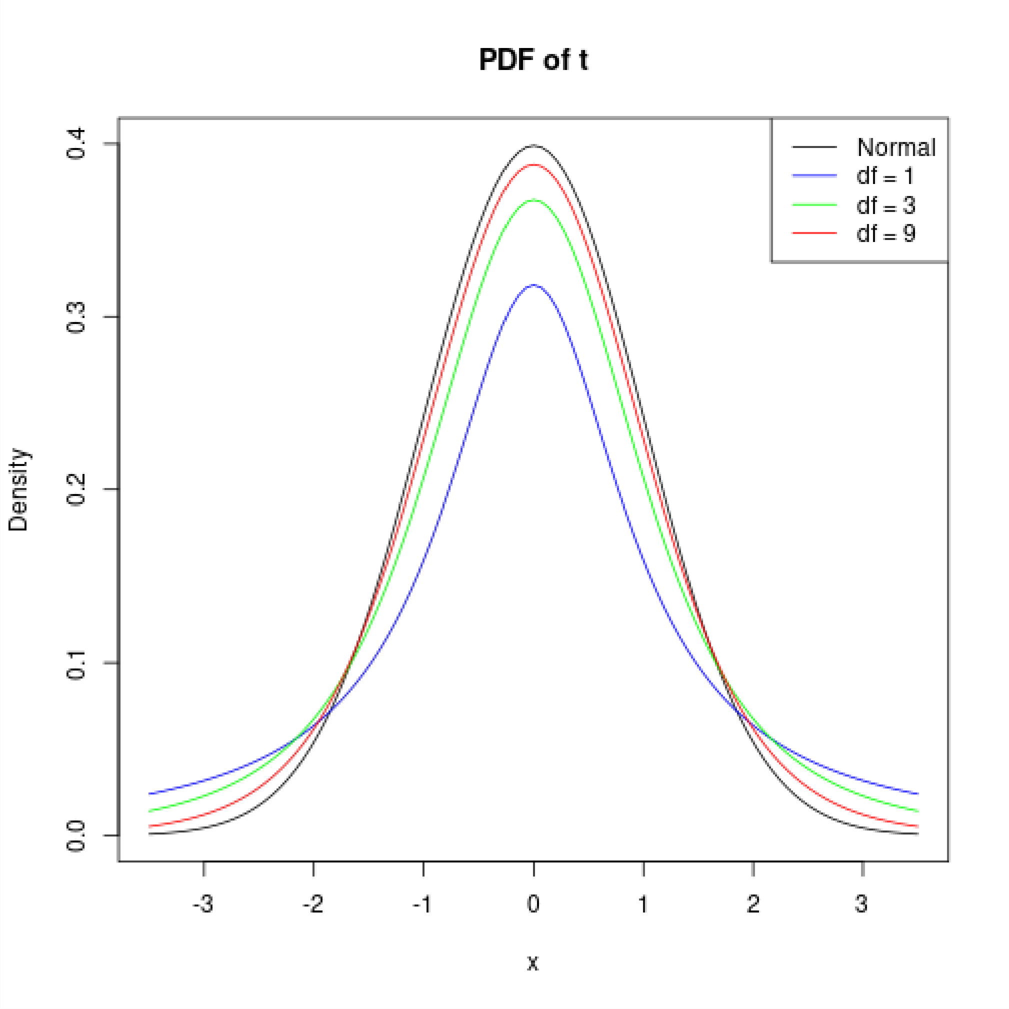
We will define the t distribution as a combination of a standard normal $Z \sim N(0, 1)$, and $V \sim \chi^2_v$:

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

where v is the degrees of freedom of the distribution.

Review

t Distribution



F Distribution

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

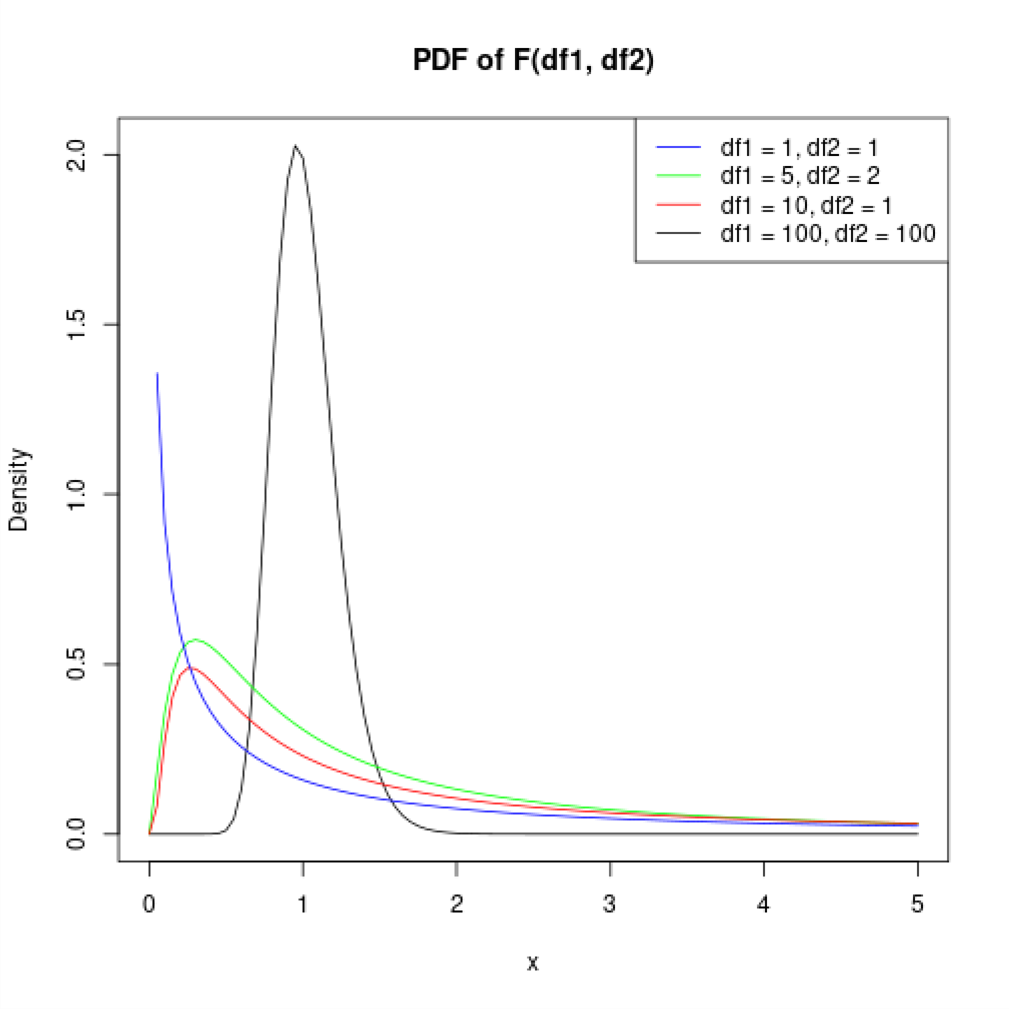
Let $U \sim \chi^2_u$ and $V \sim \chi^2_v$, with U and V independent. Then

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

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

Review

F Distribution



Review

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.

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

Review

Hypothesis Testing

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

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

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 *simple hypothesis*.
- If Ω_0 contains more than one value, H_0 is called a *composite hypothesis*.

Review

Null Hypothesis Significance Testing

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

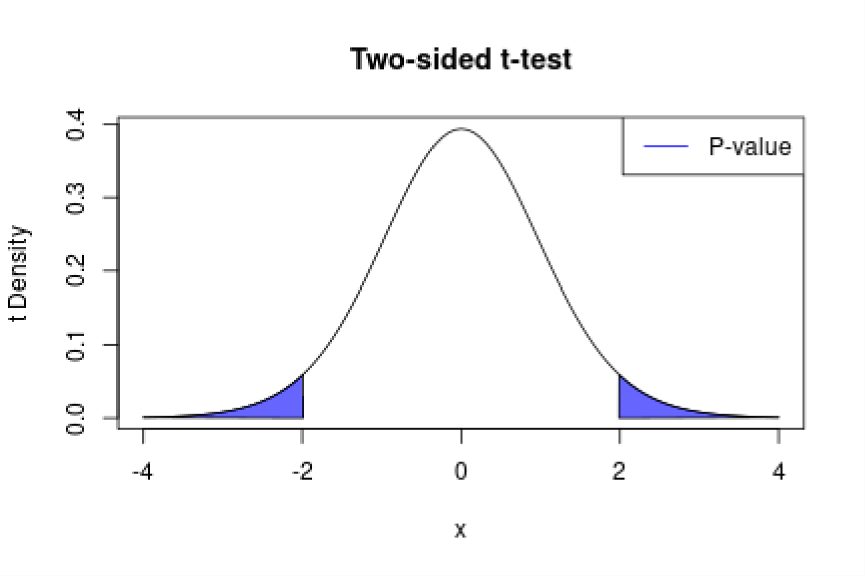
at level of significance α, given a sample $X_1, \ldots, X_n$.

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

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

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

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0.$$

Review

One Sample z-test

Step 2

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

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

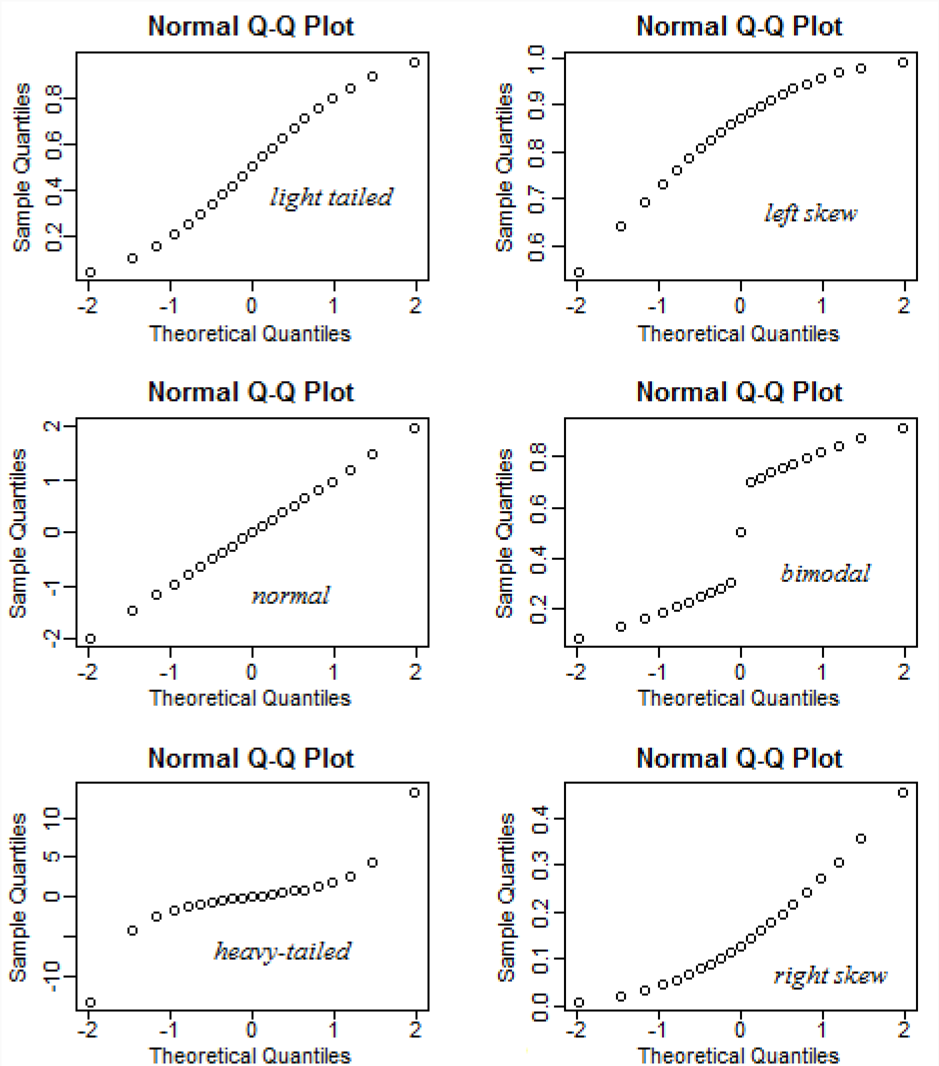
Null Distribution: $\overline{X} \overset{H_0}{\sim} \mathcal{N}(\mu_0, \sigma^2/n)$

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

- The *minimum variance unbiased linear estimator* for μ
- The *Maximum Likelihood Estimator* for μ
- More on this later...

Review

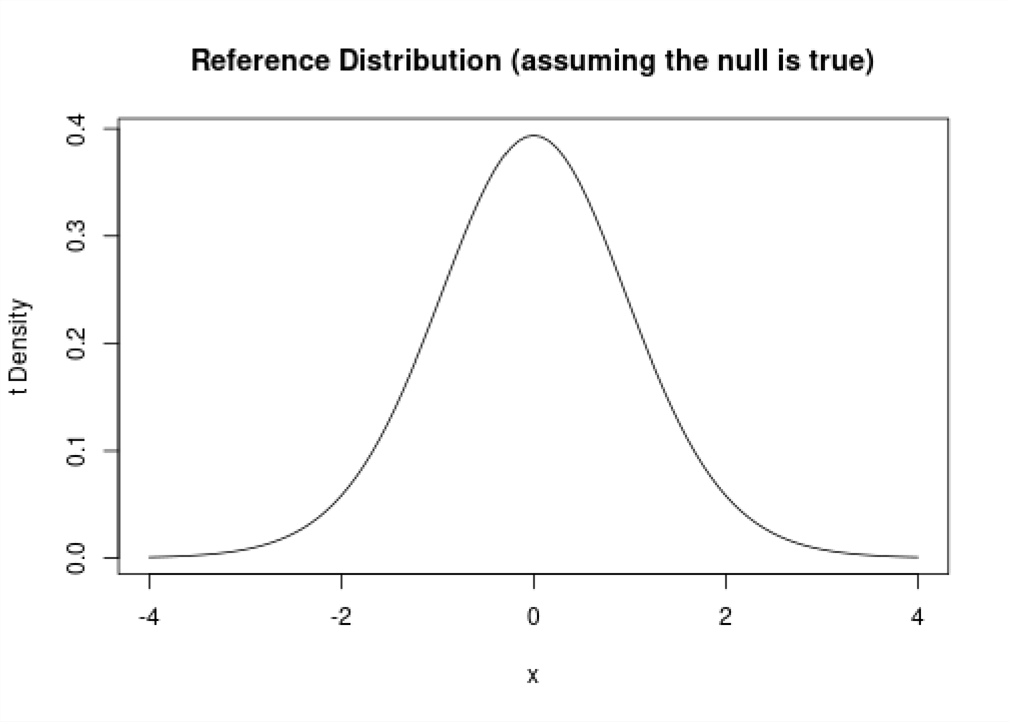
Step 3 Check model assumptions.



Review

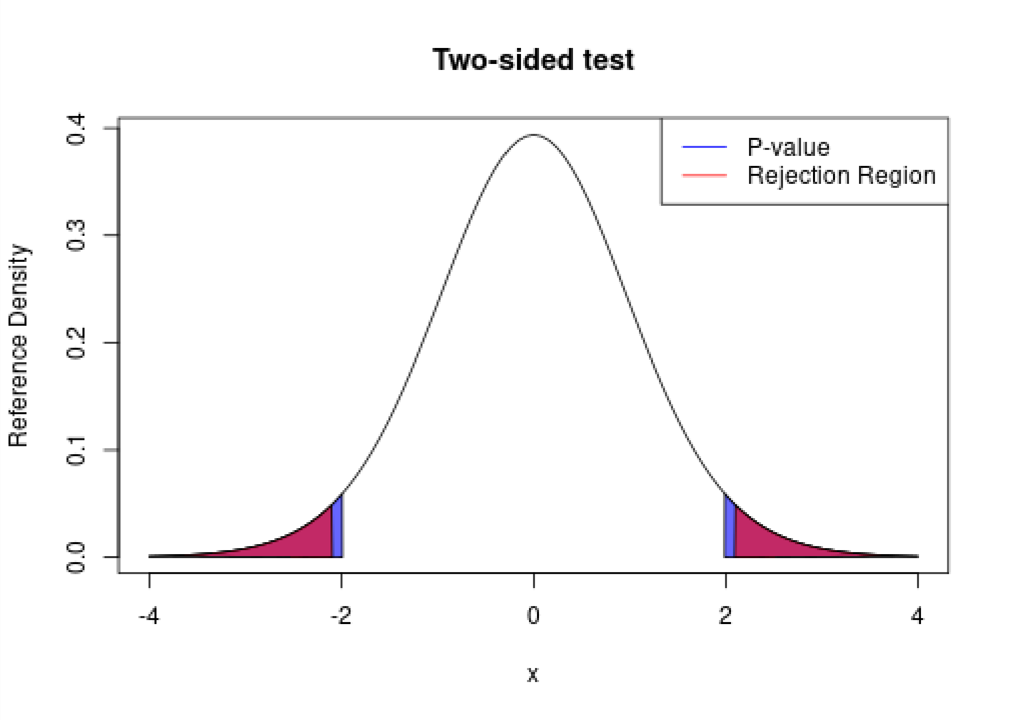
Step 4 Compute reference distribution.

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



Review

Step 5 Make conclusion.



Review

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

One-sample z-test

- 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

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

Review

One-sample z-test

- Suppose $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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?

Review

One-sample z-test (cont'd)

- 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

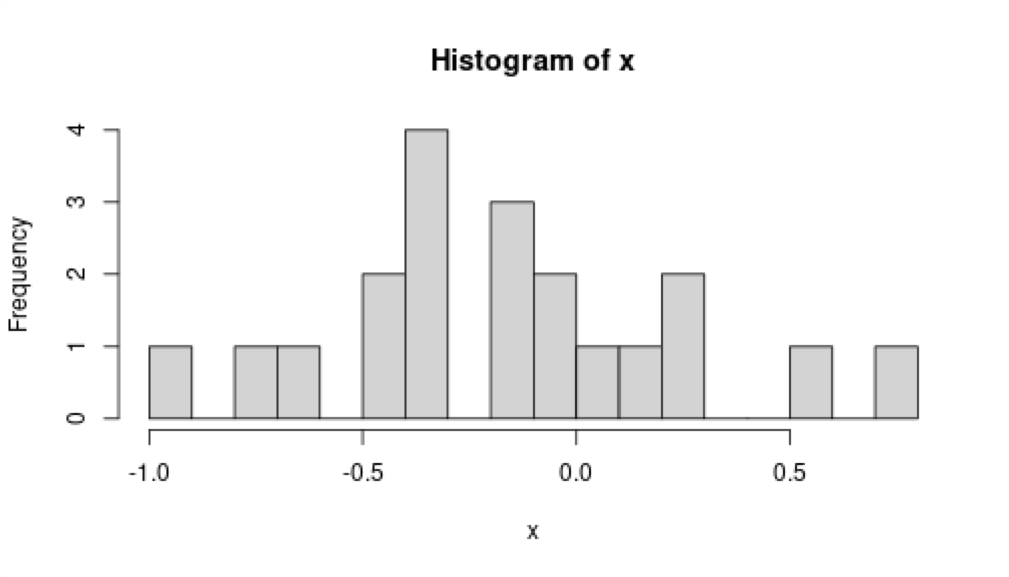
$$T = \frac{Z}{\sqrt{V/v}},$$

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

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

Example: One-sample t-test

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



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

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

Likelihood Ratio Test

Review

Likelihood Ratio Test

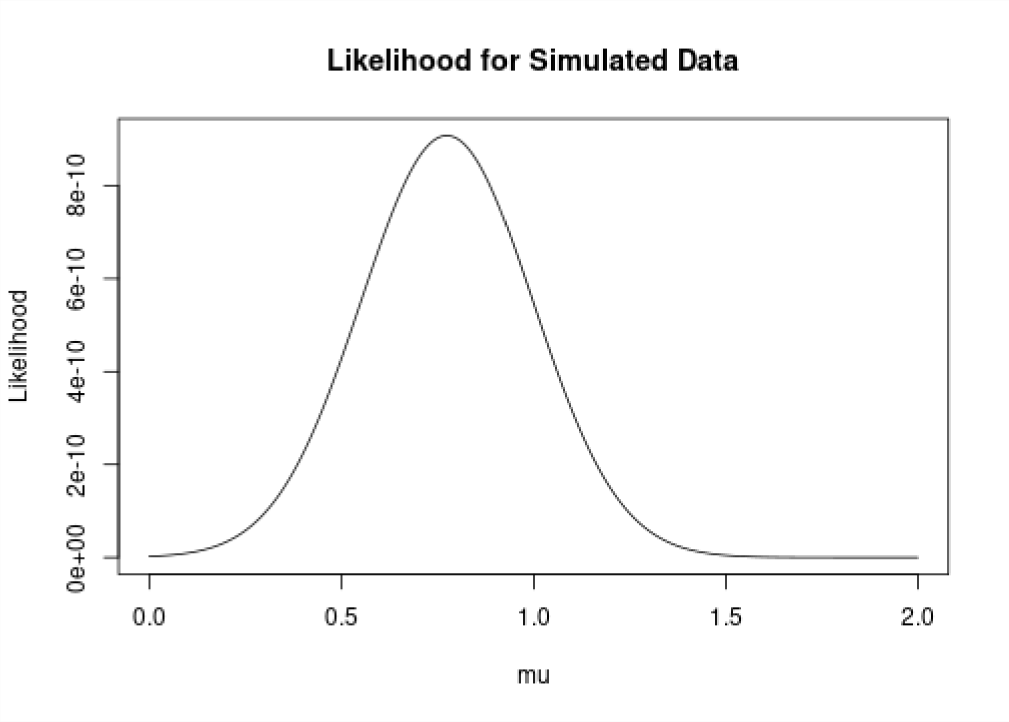
```
set.seed(1234)
n ← 20
mu ← 0.9
sigma ← 0.5
x ← rnorm(n, mu, sigma)

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)
```

Review

Likelihood Ratio Test

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

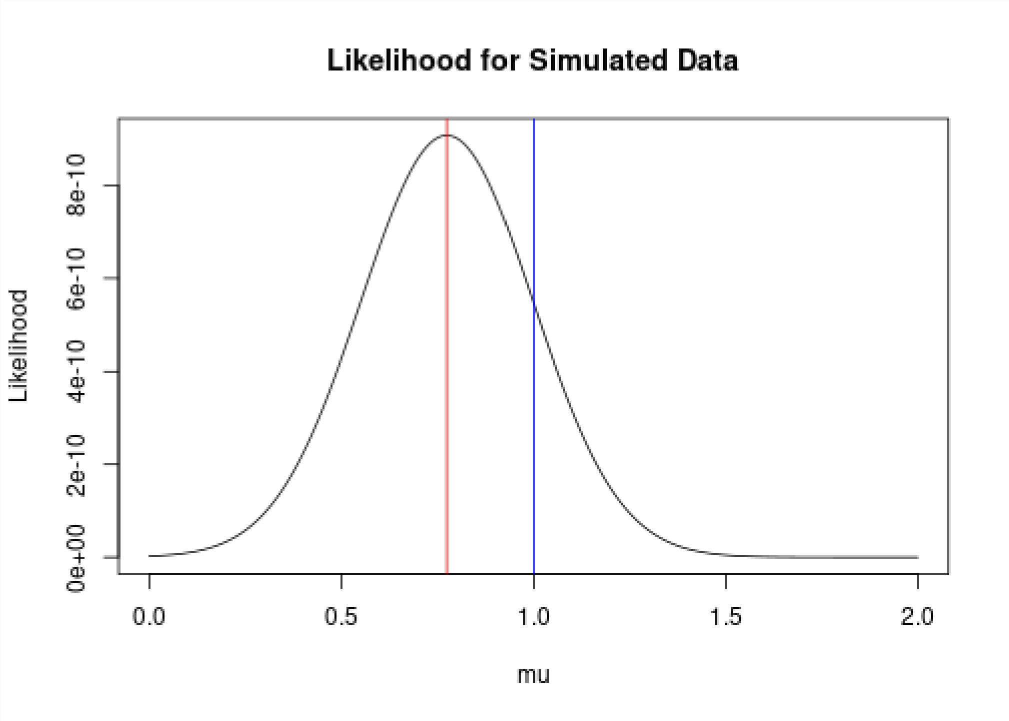


Review

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

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

Likelihood Ratio Test

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

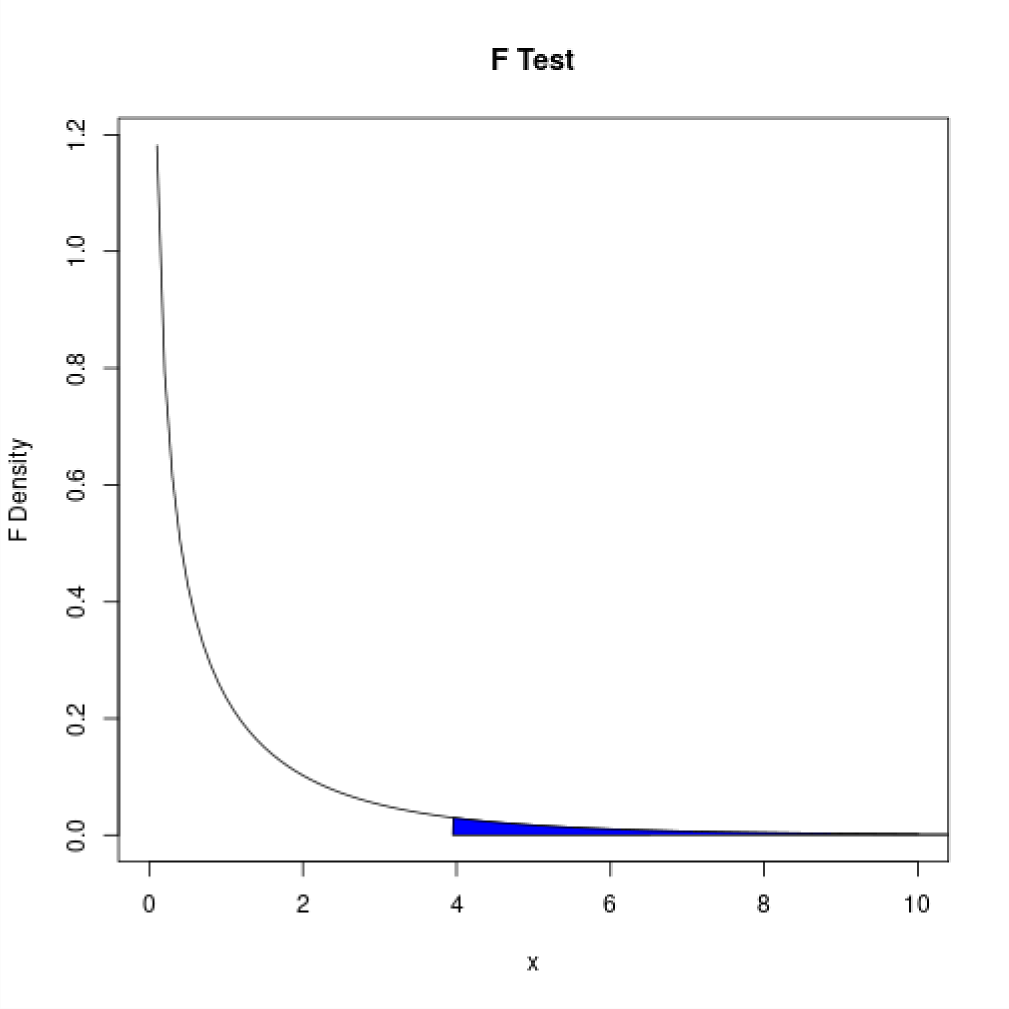
```
##
##      One Sample t-test
##
## data:  x
## t = -1.988, df = 19, p-value = 0.06142
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  0.5374297 1.0119063
## sample estimates:
## mean of x
##  0.774668
```

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

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

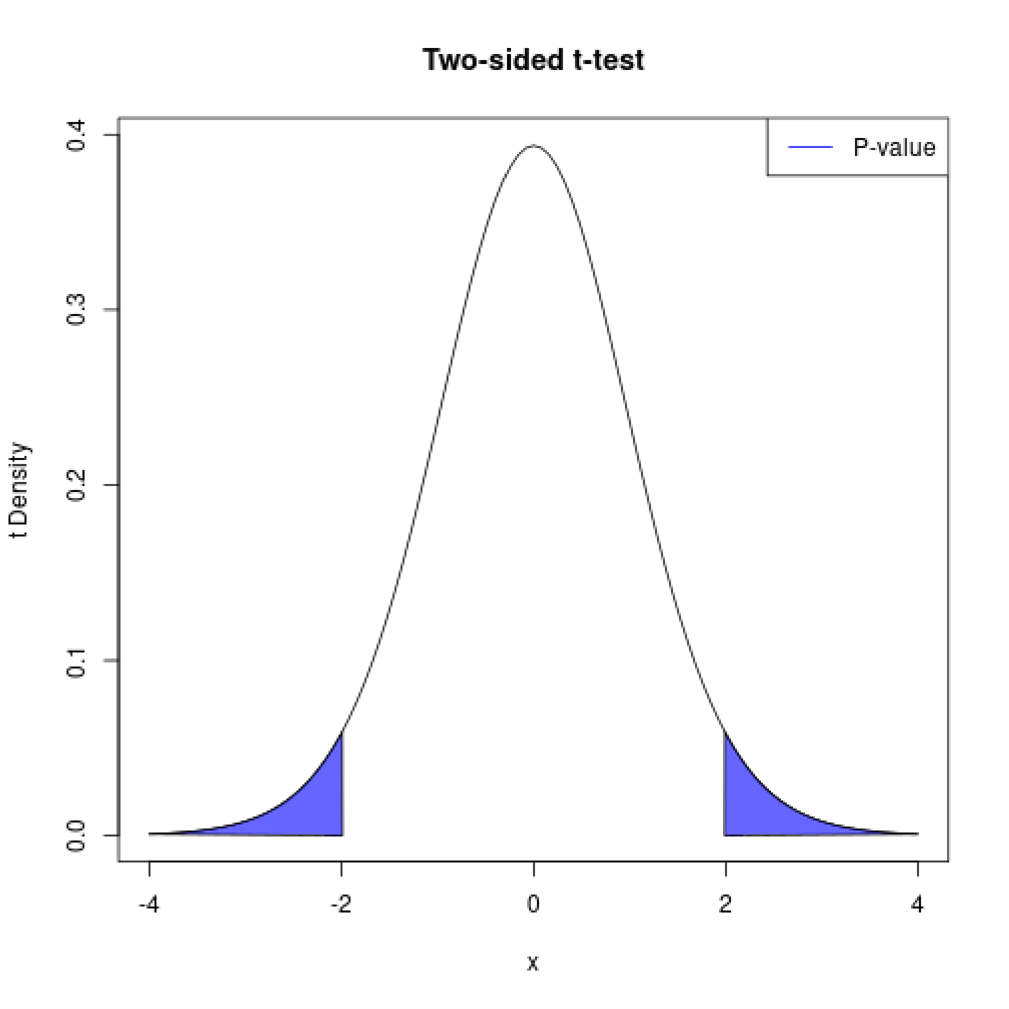
Review

Likelihood Ratio Test



Review

Likelihood Ratio Test



Review

Likelihood Ratio Test