Itroduction to Inferential Statistics

Data Science Summer School

Vahe Movsisyan

CBA, AUA, Datamotus LLC

July 3-4, 2019

What are Inferential statistics?

Inferential statistics are statistical methods that use the data collected from a **sample** to draw conclusions about a **population**

Why we need Inferential statistics?

Inferential statistics allow us to draw conclusions from data that might not be immediately obvious.

Inferential statistics help to answer several questions:

- Making inferences about a population from a sample by estimating **unknown parameters of the population**
- Concluding whether a sample is **significantly different** from the population
- If adding or removing a feature from a **statistical model** will help in improving it
- If one statistical model is **significantly different** from the other
- **Hypothesis Testing**

What is a Hypothesis?

A hypothesis is a claim (assertion) about a population parameter.

H_0 - **Null Hypothesis**

H_1 - **Alternative Hypothesis**

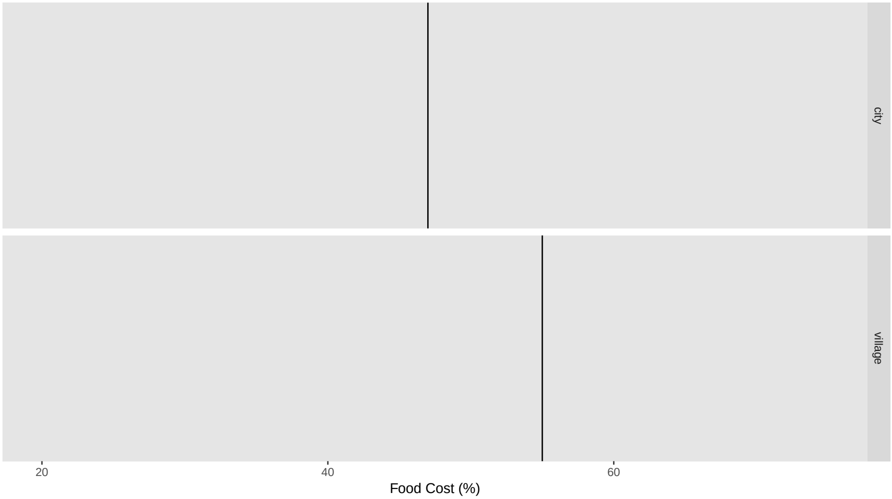
If the sample data is consistent with the null hypothesis, then we do not reject it.

If the sample is inconsistent with the null hypothesis, but consistent with the alternative, then we reject the null hypothesis and conclude that the alternative hypothesis is true.

What is a Hypothesis?

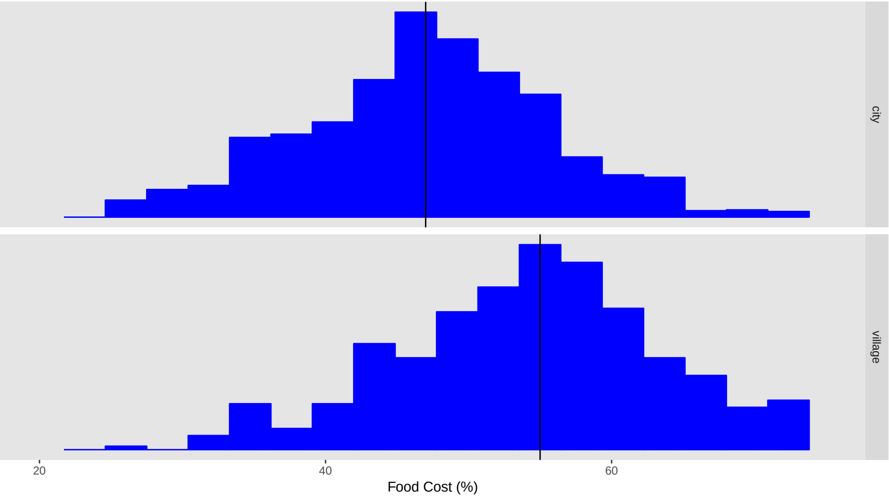
Is there a significant difference in **mean food cost (%)** between rural and urban areas?

region	mean
city	47
village	55



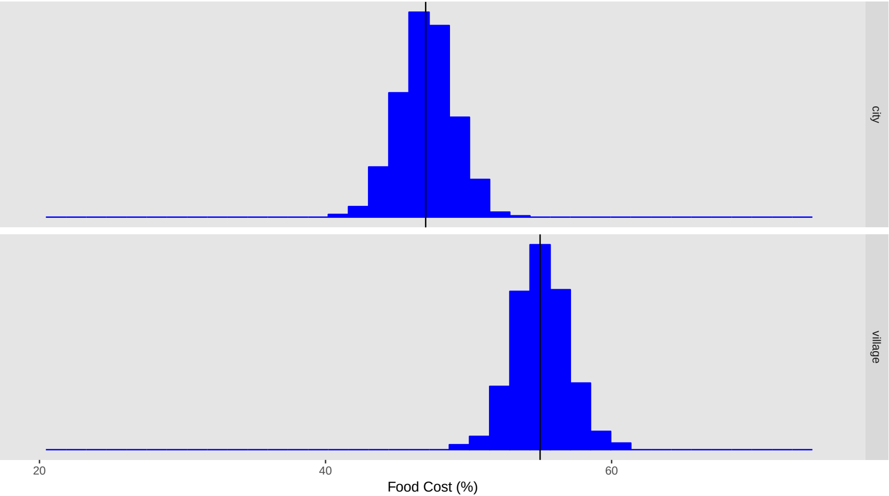
Is there a significant difference in **mean food cost (%)** between rural and urban areas?

region	n	mean	sd	min	Q1	median	Q3	max
city	1582	47	8.7	20	41	47	53	77
village	418	55	10.1	27	48	55	61	88



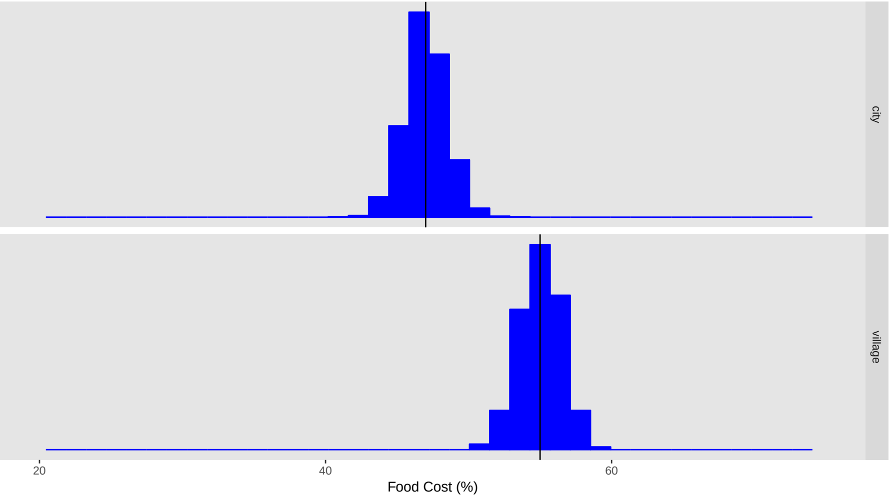
Is there a significant difference in **mean food cost (%)** between rural and urban areas?

region	n	mean	sd	min	Q1	median	Q3	max
city	1582	47	1.9	40.8	45.8	47.1	48.3	54.2
village	418	55	1.9	48.9	53.7	55.0	56.2	60.6



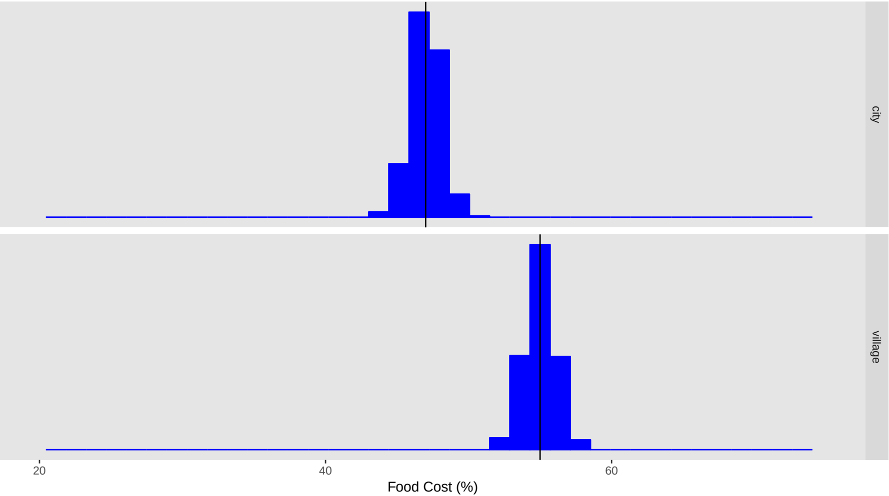
Is there a significant difference in **mean food cost (%)** between rural and urban areas?

region	n	mean	sd	min	Q1	median	Q3	max
city	1582	47	1.5	40.8	46.0	47.0	47.9	53.8
village	418	55	1.5	50.7	54.1	55.1	56.1	59.6



Is there a significant difference in **mean food cost (%)** between rural and urban areas?

region	n	mean	sd	min	Q1	median	Q3	max
city	1582	47	1.1	43.7	46.3	47	47.8	50.5
village	418	55	1.1	51.5	54.3	55	55.7	58.3



Stapes of Hypothesis Testing

1. State the null hypothesis and the alternative hypothesis,
2. Choose the level of significance (α) and the sample size (n),
3. Determine the appropriate test statistic (t, z, F, $\ldots$),
4. Determine the critical values that divide the rejection and non-rejection regions,
5. Collect data and compute the value of the test statistic,
6. If the test statistic falls into the non-rejection region, do not reject the null hypothesis H_0. If the test statistic falls into the rejection region, reject the null hypothesis.

Parametric vs Non-Parametric Hypothesis testing

Most commonly used Parametric tests

- One sample test
- Two independent samples test
- Two related (paired) samples test
- More than 2 samples: One-Way ANOVA test
- Two samples variance test
- Two populations proportions test

Most commonly used Non-Parametric tests

- One sample test: Wilcoxon signed-rank test
- Two independent samples: Mann-Whitney U test
- Two related (paired) samples test: Wilcoxon signed-rank test
- More than 2 samples: Kruskal-Wallis One-Way ANOVA test
- Independence of two categorical variables (Chi-Square test)

Parametric vs Non-Parametric Hypothesis testing

When to use Parametric tests?

- The variable of interest is continuous
- Population distribution in all groups are normal or sample sizes > 30

When to use Non-Parametric tests?

- When the **median** is a better measure of central tendency than the **arithmetic mean**
- The sample size is very small
- The variable of interest is nominal, ordinal, ranked data, or there are outliers which cannot be removed
- Many non-parametric methods convert raw values to ranks and then analyze ranks

Non-Parametric tests are called distribution-free tests because they don't assume that your data follow a specific distribution

Parametric Hypothesis testing

Assumptions:

- Sample is randomly and independently drawn
- Population distribution is normal or sample size > 30

Two Tailed Test

$$H_0 : \mu = c$$

$$H_1 : \mu \neq c$$

Ex: Is there evidence that mean salary in Armenia is **significantly different** from official **160000** dram

One Sample Hypothesis Testing

One Tailed Test (Right Tailed)

$$H_0 : \mu \leq c$$

$$H_1 : \mu > c$$

Ex: Can we prove that average height in Armenia is **significantly higher** than **156** cm

One Tailed Test (Left Tailed)

$$H_0 : \mu \geq c$$

$$H_1 : \mu < c$$

Ex: Is there a **significant evidence** that the average profit of Telco companies is **less** than **one million** dollars

Test Statistic for one sample test

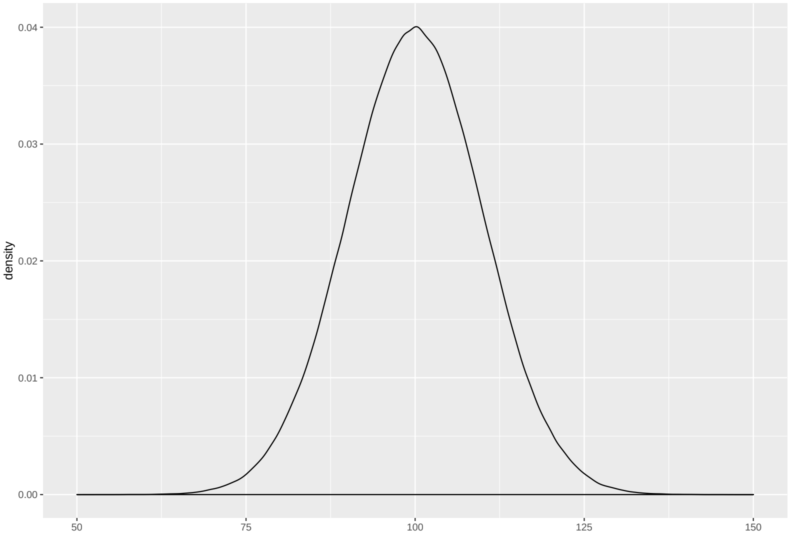
If population standard deviation (σ) is known, then $Z - test$

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

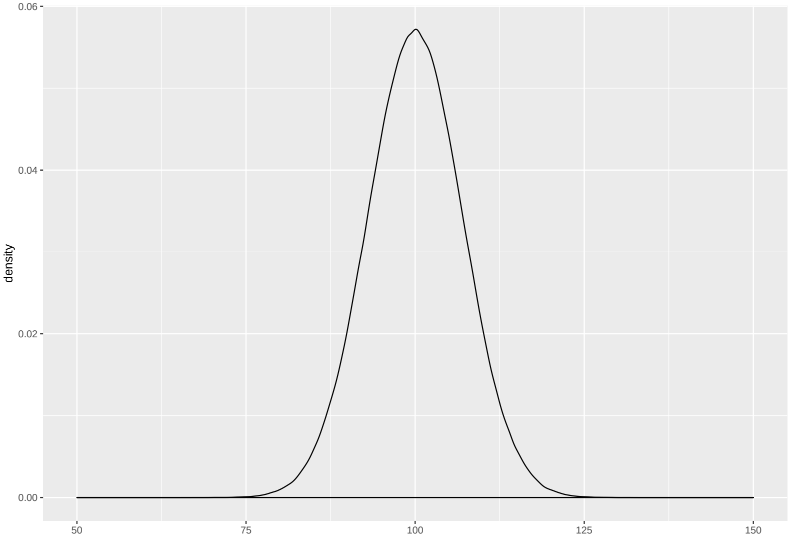
If population standard deviation (σ) is unknown, then $t - test$

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

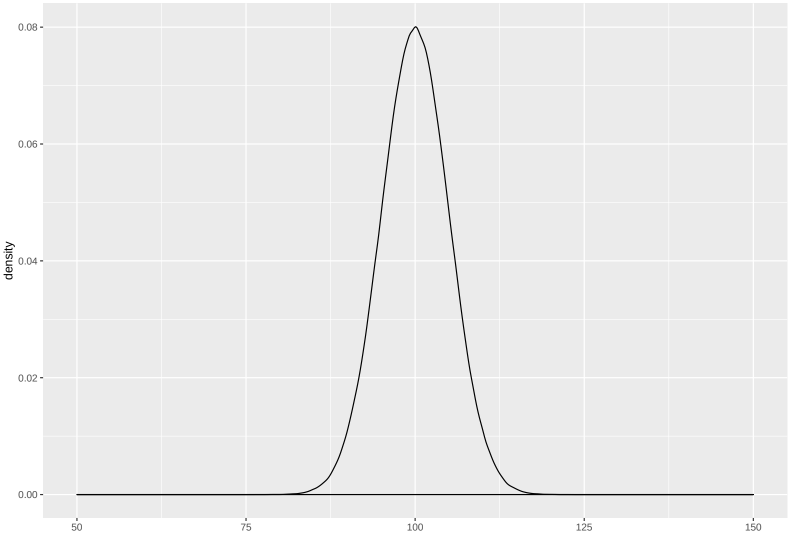
Normal Distribution (bell shaped and symmetric): mean = 100, sd = 10



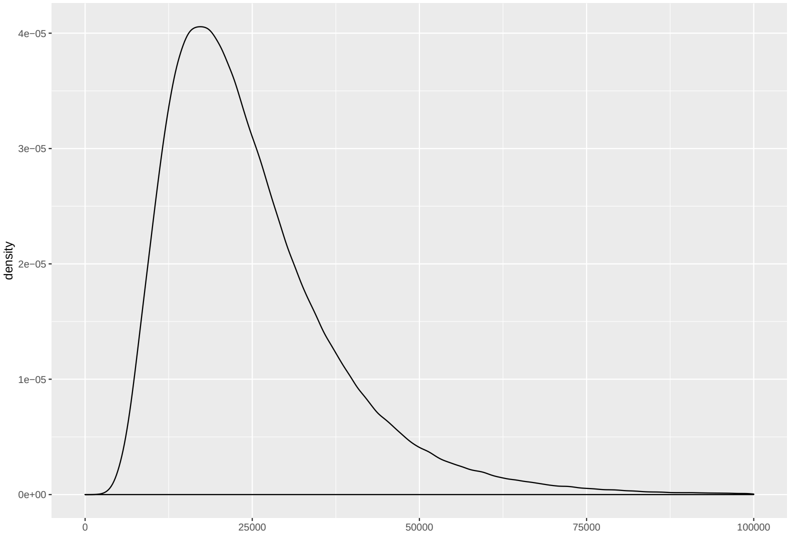
Normal Distribution (bell shaped and symmetric): mean = 100, sd = 7



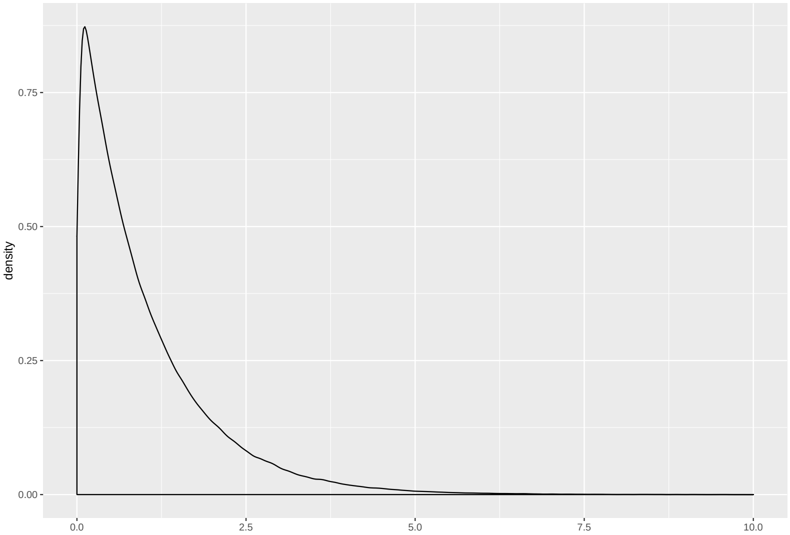
Normal Distribution (bell shaped and symmetric): mean = 100, sd = 5



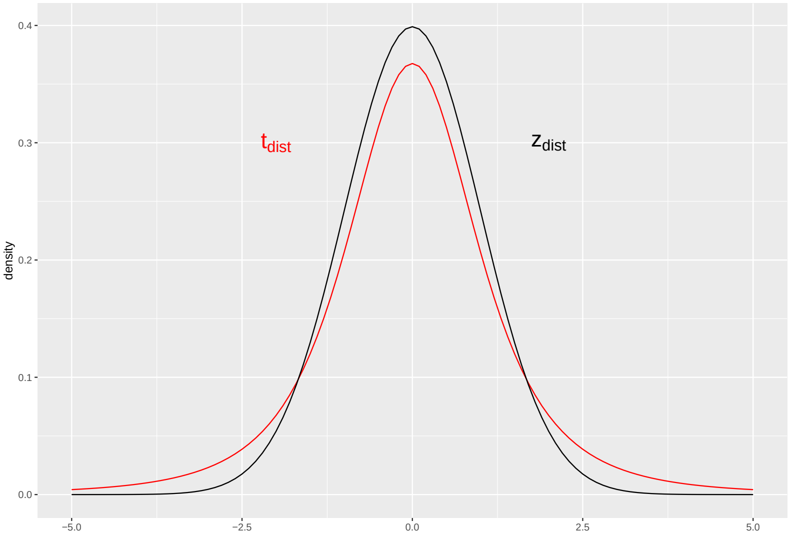
Log-Normal Distribution (right skewed)



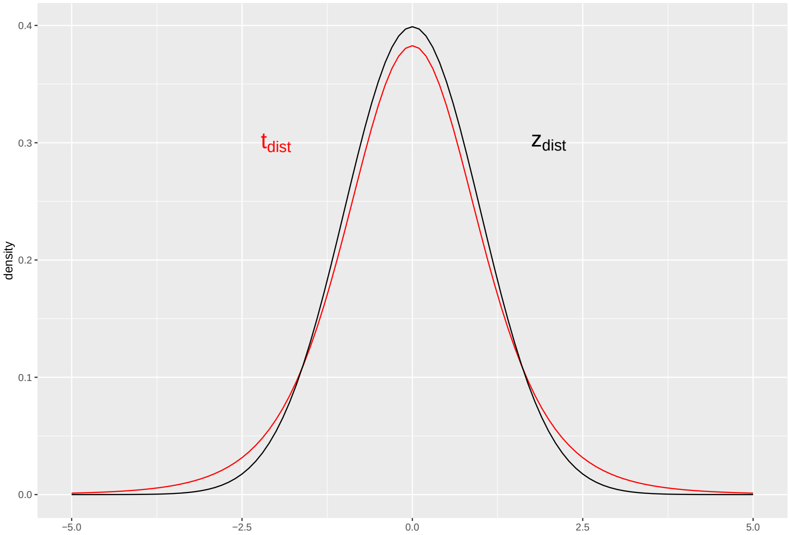
Exponential Distribution (right skewed)



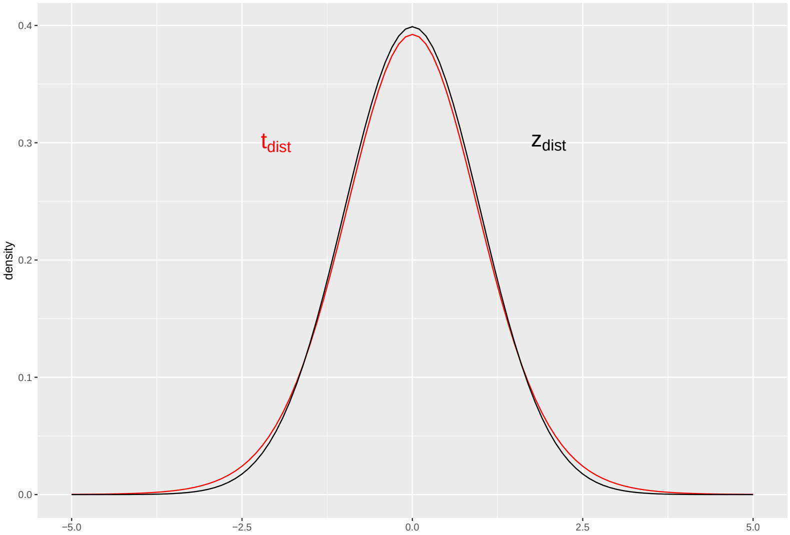
Standardized Normal (z) and Student (t, df = n-1 = 3) Distributions



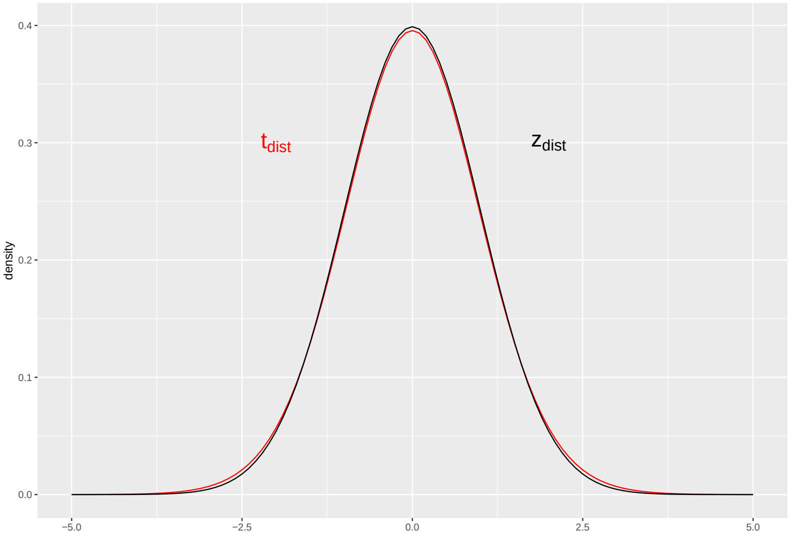
Standardized Normal (z) and Student (t, df = n-1 = 6) Distributions



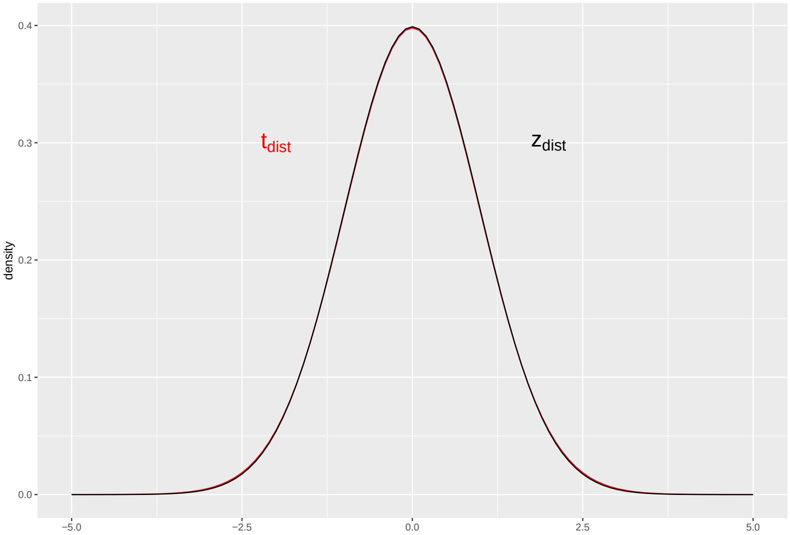
Standardized Normal (z) and Student (t, df = n-1 = 15) Distributions



Standardized Normal (z) and Student (t, df = n-1 = 30) Distributions



Standardized Normal (z) and Student (t, df = n-1 = 100) Distributions



One Sample Hypothesis Testing

Two Tailed Test

$$H_0 : \mu_{age} = 45$$

$$H_1 : \mu_{age} \neq 45$$

One Tailed Test (Right Tailed)

$$H_0 : \mu_{age} \leq 45$$

$$H_1 : \mu_{age} > 45$$

One Tailed Test (Left Tailed)

$$H_0 : \mu_{age} \geq 45$$

$$H_1 : \mu_{age} < 45$$

Risks in Decision Making Using Hypothesis Testing

Using hypothesis testing involves the risk of reaching an incorrect conclusion.

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No Error Probability $1 - \alpha$	Type II Error Probability β
Reject H_0	Type I Error Probability α	No Error Probability $1 - \beta$

Decision Rule

1. Critical Value Approach

If the value of test-statistic falls in Non-Rejection Region (between two critical values in case of two tailed test), then do not reject H_0, otherwise reject H_0

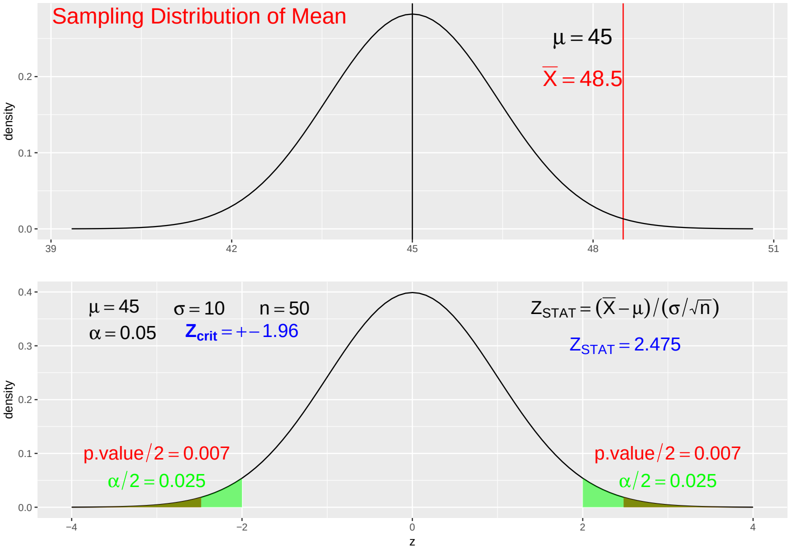
2. P-value Approach

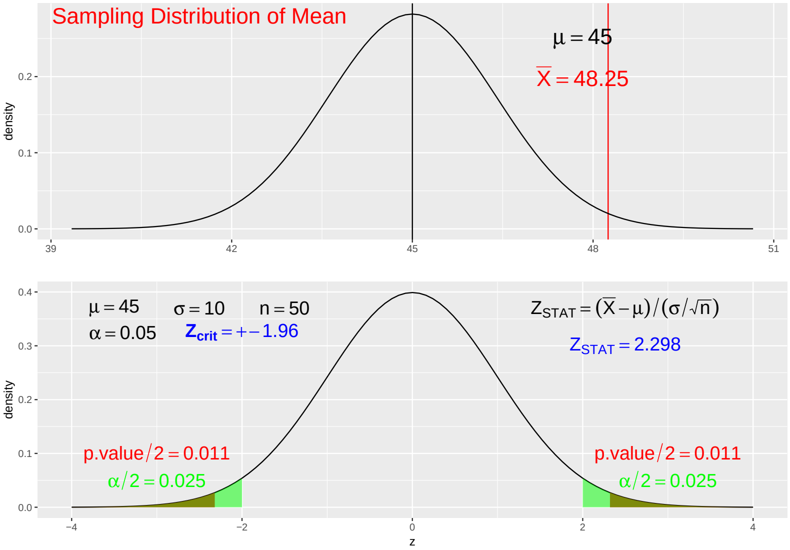
Reject the Null Hypothesis if p-value $<$ alpha

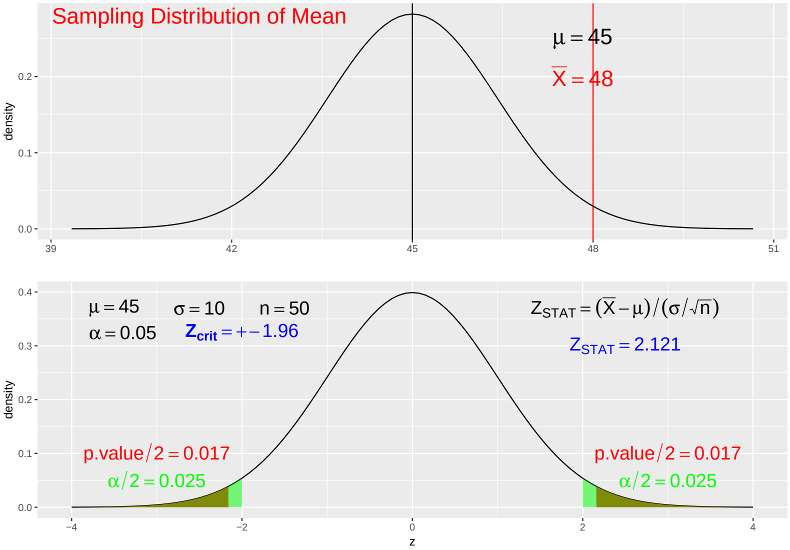
The P value is a probability of finding the observed, or more extreme results when the null hypothesis is true.

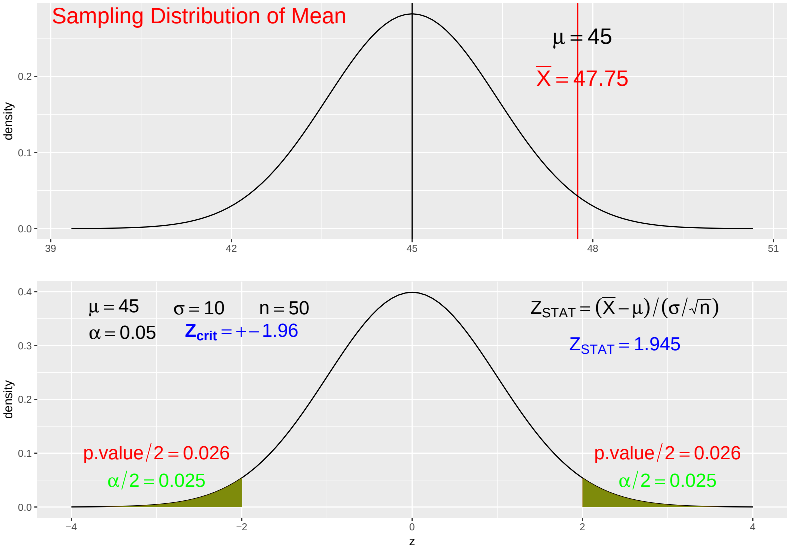
The **level of significance (alpha)** is used to refer to a hypothetical **Type I error**

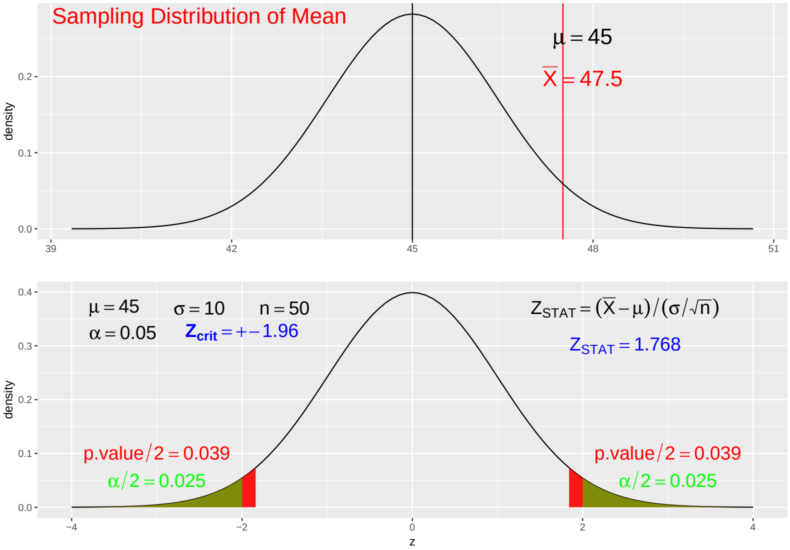
The **P value** is used to indicate a probability that you calculate after a given study, so it can be interpreted as an **Observed alpha**.

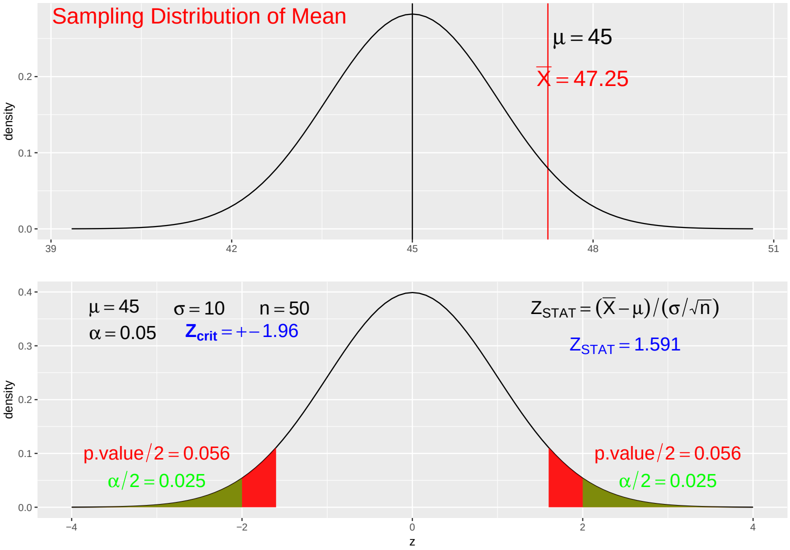

Sampling Distribution of Mean
μ = 45
X̄ = 48.5
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 2.475
p.value/2 = 0.007
α/2 = 0.025
p.value/2 = 0.007
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

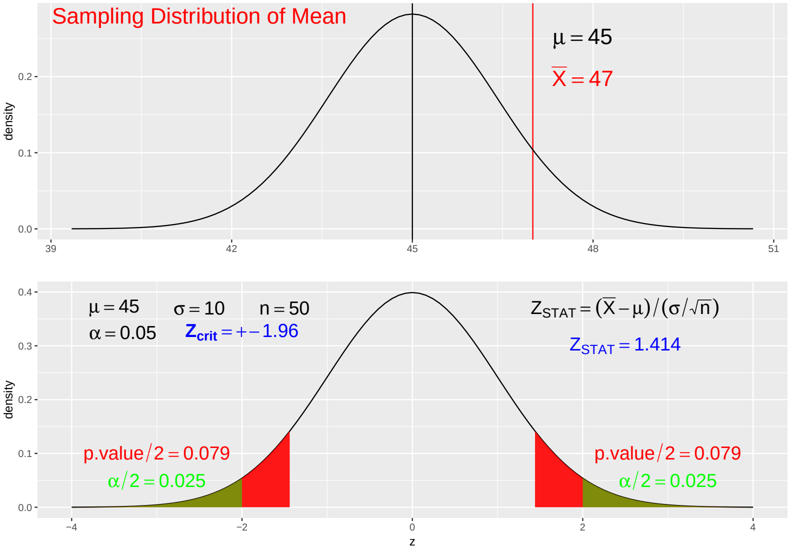

Sampling Distribution of Mean
μ = 45
X̄ = 48.25
density
0.2
0.1
0.0
39
42
45
48
51
0.4
0.3
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 2.298
p.value/2 = 0.011
α/2 = 0.025
p.value/2 = 0.011
α/2 = 0.025
−4
−2
0
2
4
z

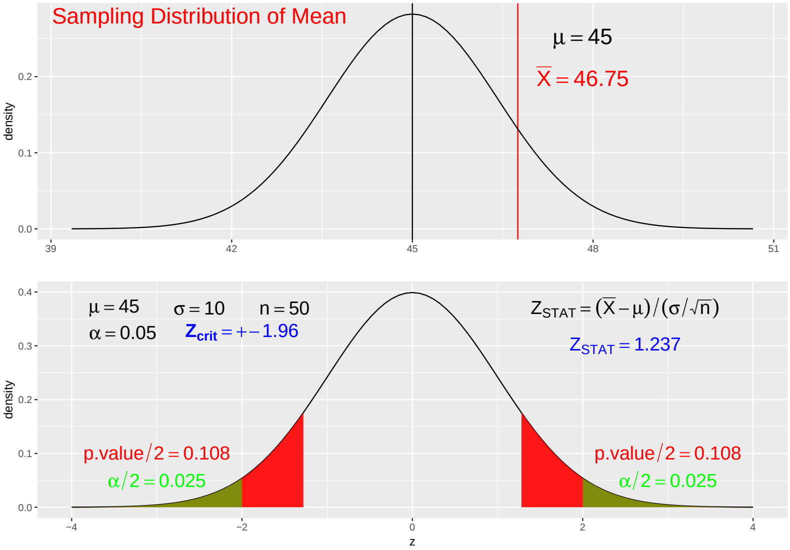

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 48$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 2.121$
$p.value/2 = 0.017$
$\alpha/2 = 0.025$
$p.value/2 = 0.017$
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

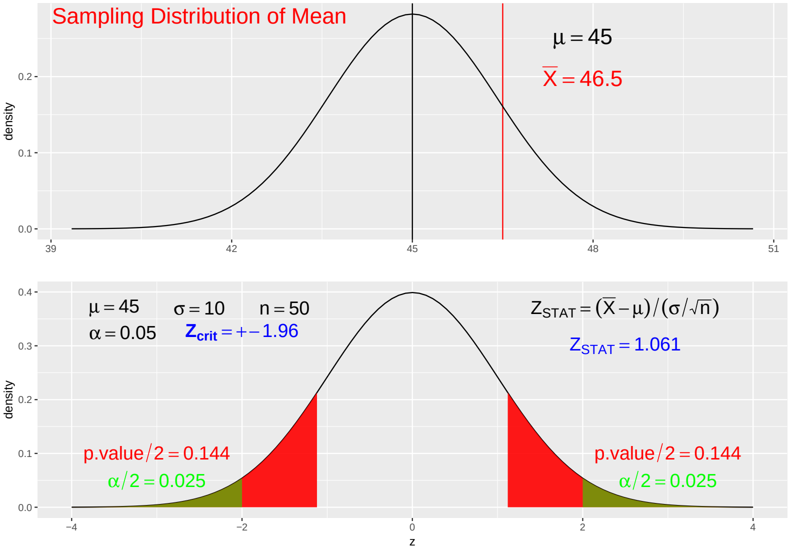

Sampling Distribution of Mean
μ = 45
X̄ = 47.75
density
0.2
0.1
0.0
39
42
45
48
51
0.4
0.3
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.945
p.value/2 = 0.026
α/2 = 0.025
p.value/2 = 0.026
α/2 = 0.025
−4
−2
0
2
4
z

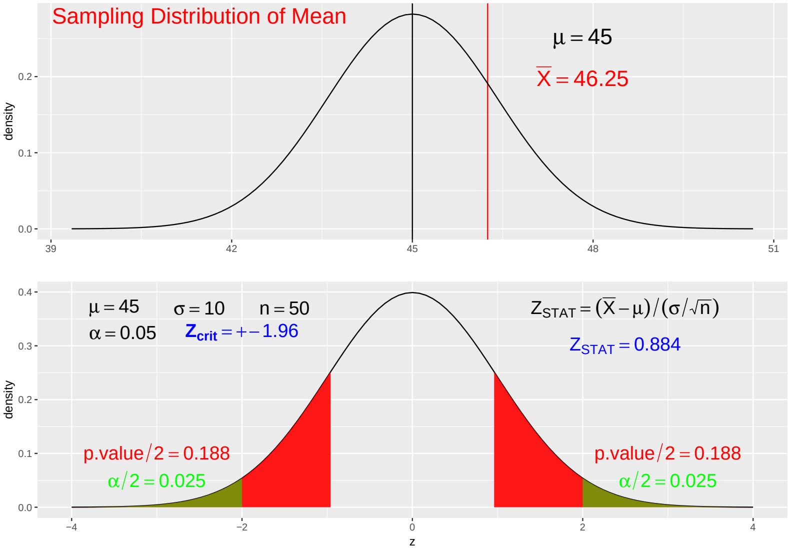

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 47.5$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 1.768$
p.value/2 = 0.039
$\alpha/2 = 0.025$
p.value/2 = 0.039
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

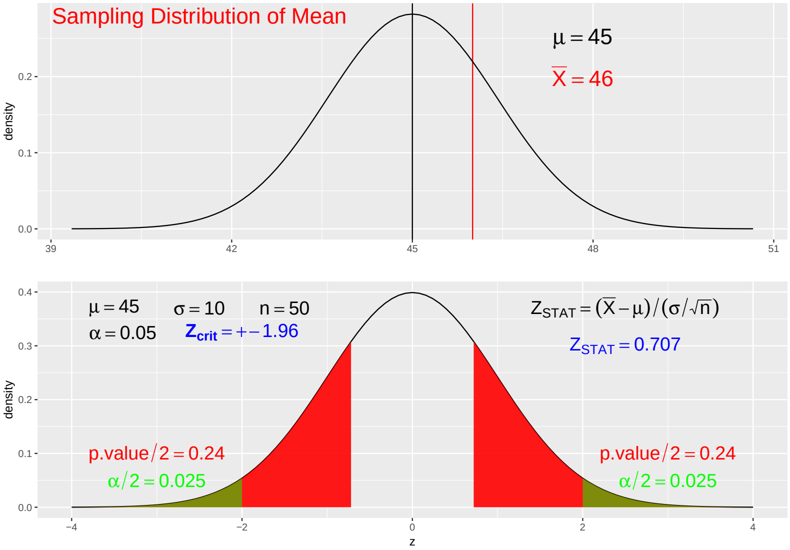

Sampling Distribution of Mean
μ = 45
X̄ = 47.25
density
0.2
0.1
0.0
39
42
45
48
51
0.4
0.3
μ = 45
σ = 10
n = 50
α = 0.05
Z_crit = +−1.96
Z_STAT = (X̄ − μ)/(σ/√n)
Z_STAT = 1.591
p.value/2 = 0.056
α/2 = 0.025
p.value/2 = 0.056
α/2 = 0.025
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 47$
density
0.2
0.1
0.0
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 1.414$
p.value/2 = 0.079
$\alpha/2 = 0.025$
p.value/2 = 0.079
$\alpha/2 = 0.025$
density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
z

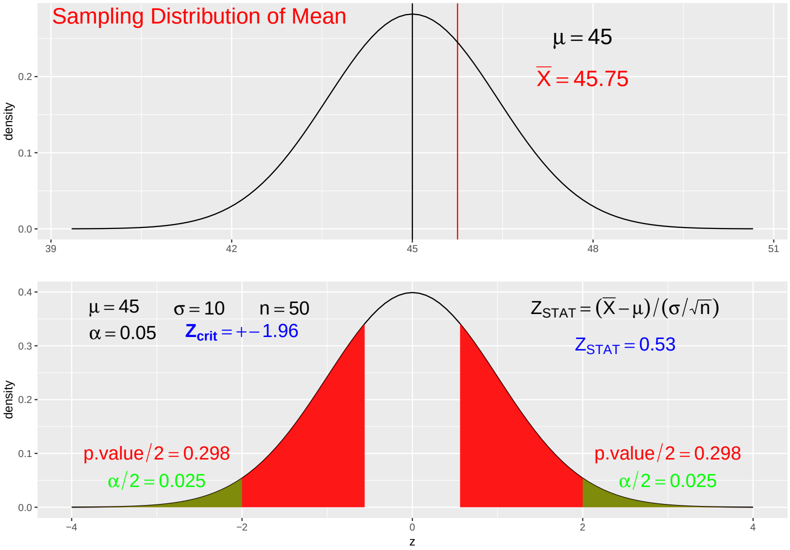

Sampling Distribution of Mean
μ = 45
X̄ = 46.75
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Z_crit = +−1.96
Z_STAT = (X̄ − μ)/(σ/√n)
Z_STAT = 1.237
p.value/2 = 0.108
α/2 = 0.025
p.value/2 = 0.108
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

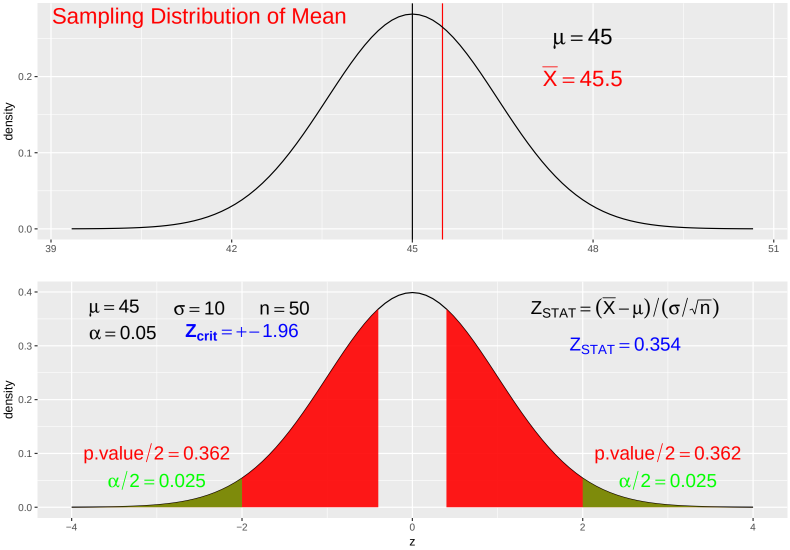

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 46.5$
density
0.2
0.1
0.0
39
42
45
48
51
0.4
0.3
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 1.061$
$p.value/2 = 0.144$
$\alpha/2 = 0.025$
$p.value/2 = 0.144$
$\alpha/2 = 0.025$
−4
−2
0
2
4
z

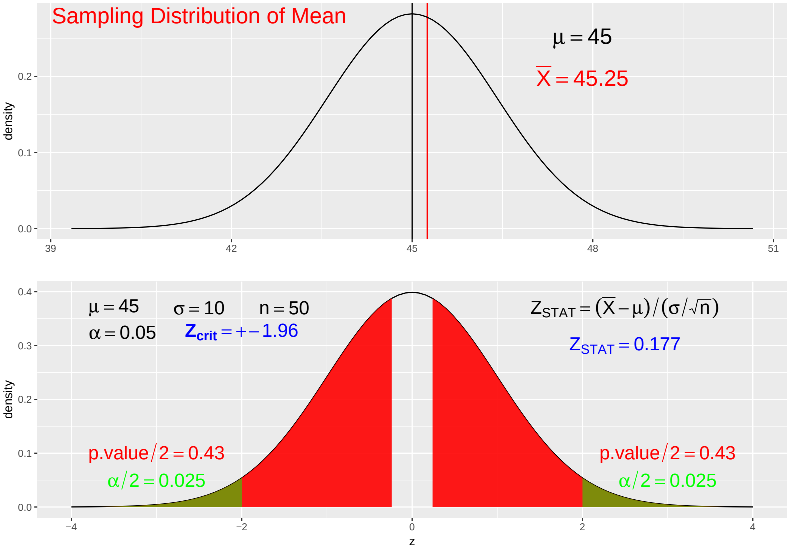

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 46.25$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.884$
p.value/2 = 0.188
$\alpha/2 = 0.025$
p.value/2 = 0.188
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

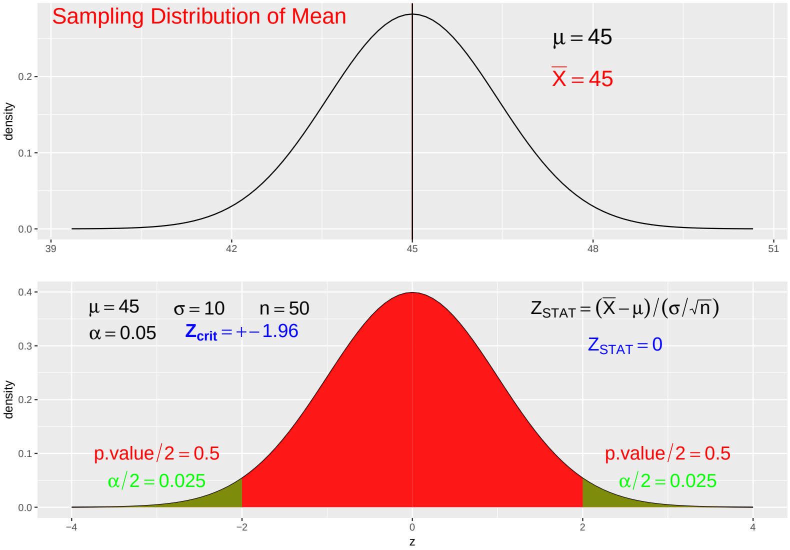


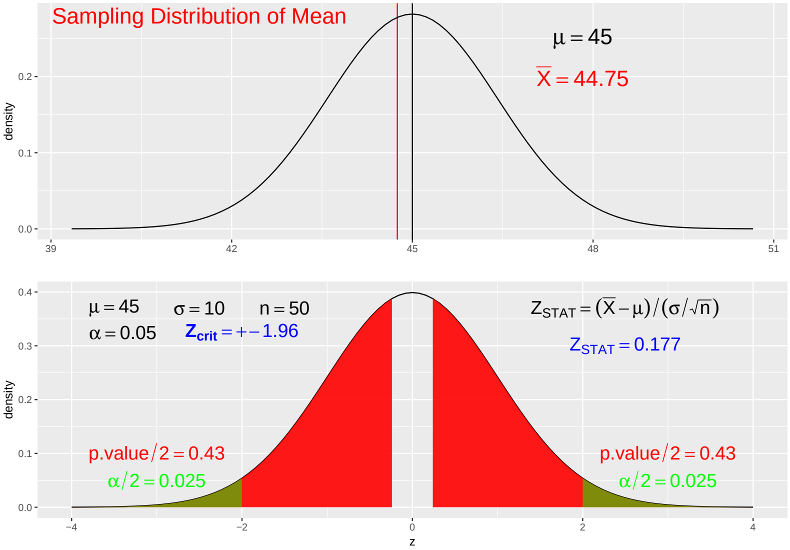
Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 46$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.707$
$p.value/2 = 0.24$
$\alpha/2 = 0.025$
$p.value/2 = 0.24$
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 45.75$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.53$
$p.value/2 = 0.298$
$\alpha/2 = 0.025$
$p.value/2 = 0.298$
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

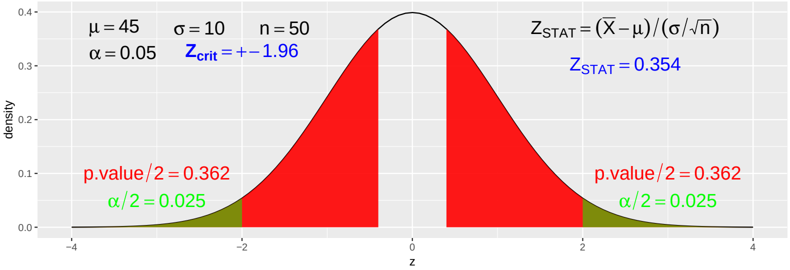

Sampling Distribution of Mean
μ = 45
X̄ = 45.5
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 0.354
p.value/2 = 0.362
α/2 = 0.025
p.value/2 = 0.362
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

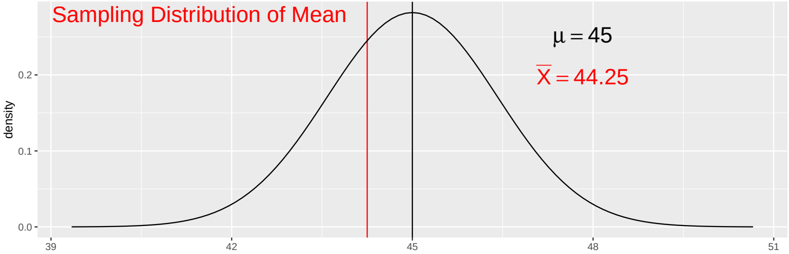

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 45.25$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = + - 1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.177$
p.value/2 = 0.43
$\alpha/2 = 0.025$
p.value/2 = 0.43
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 45$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0$
p.value/2 = 0.5
$\alpha/2 = 0.025$
p.value/2 = 0.5
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

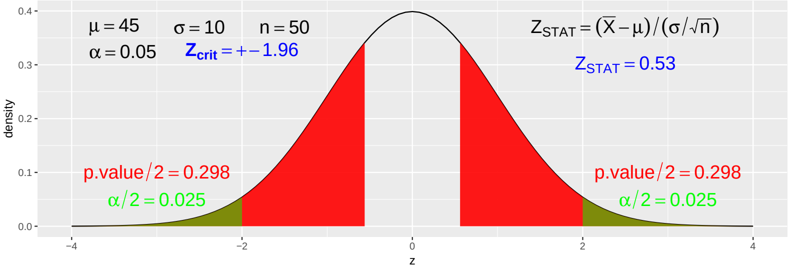

Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 44.75$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.177$
p.value/2 = 0.43
$\alpha/2 = 0.025$
p.value/2 = 0.43
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 44.5$
density
0.0
0.1
0.2
39
42
45
48
51

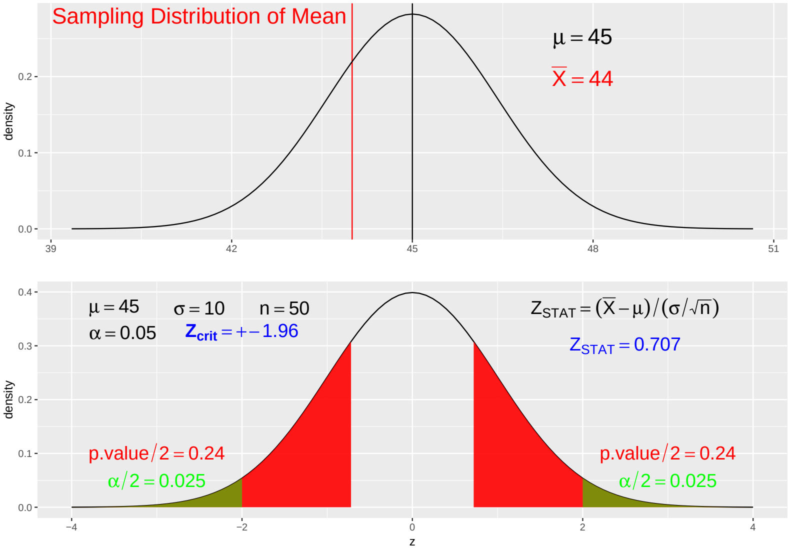

$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.354$
p.value/2 = 0.362
$\alpha/2 = 0.025$
p.value/2 = 0.362
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

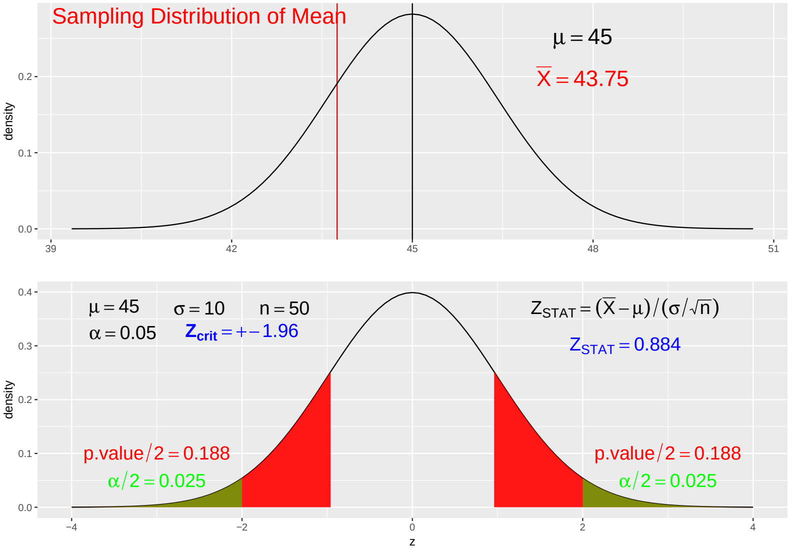


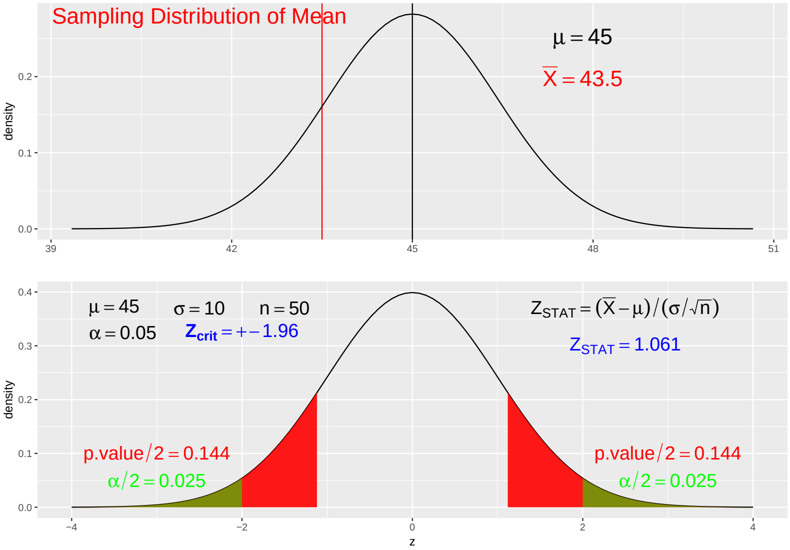
Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 44.25$
density
0.0
0.1
0.2
39
42
45
48
51

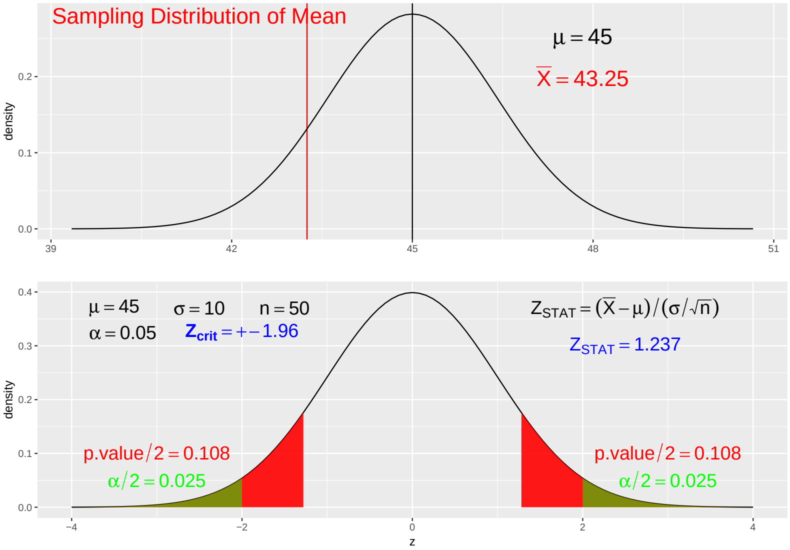


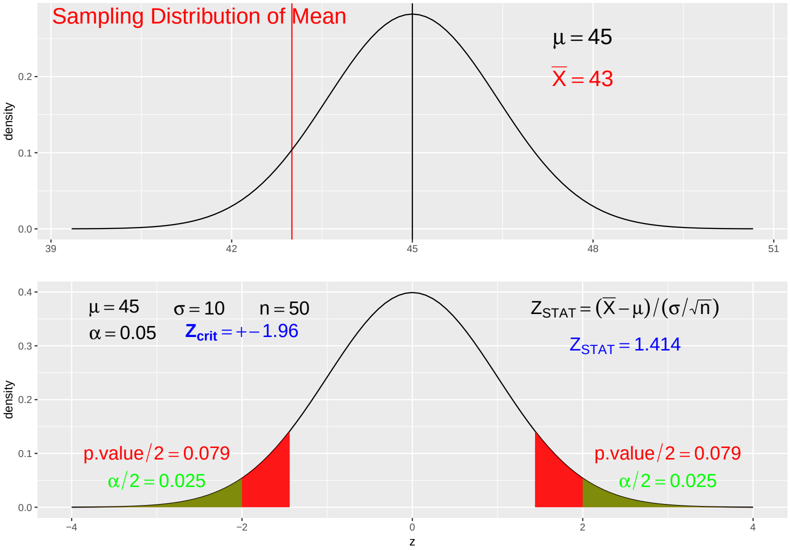
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 0.53$
$p.value/2 = 0.298$
$\alpha/2 = 0.025$
$p.value/2 = 0.298$
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
μ = 45
X̄ = 44
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Z_crit = +−1.96
Z_STAT = (X̄ − μ)/(σ/√n)
Z_STAT = 0.707
p.value/2 = 0.24
α/2 = 0.025
p.value/2 = 0.24
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

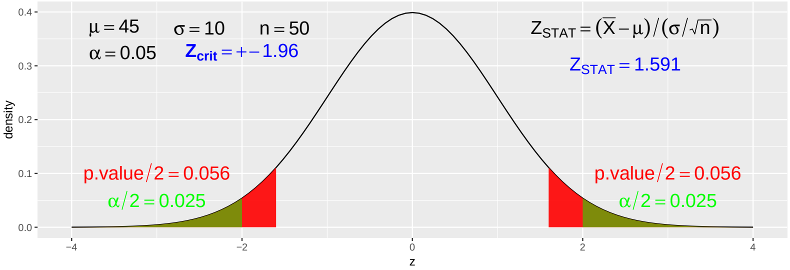

Sampling Distribution of Mean
μ = 45
X̄ = 43.75
density
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 0.884
p.value/2 = 0.188
α/2 = 0.025
p.value/2 = 0.188
α/2 = 0.025
z

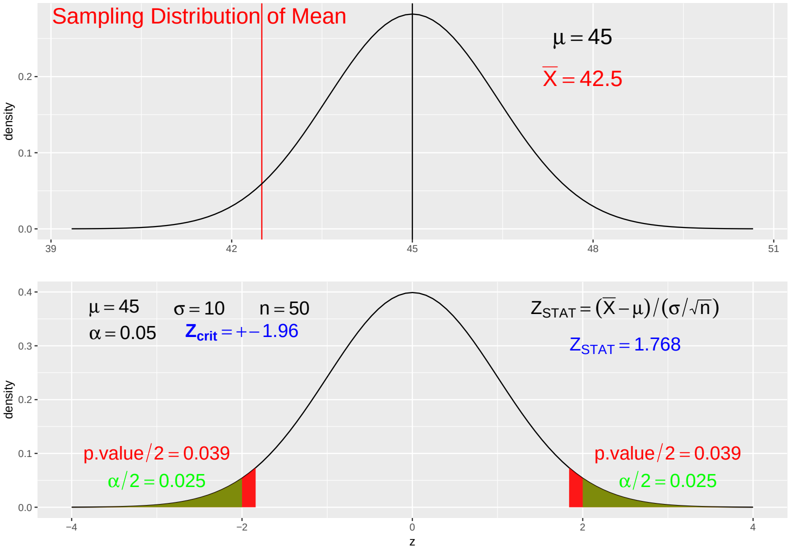

Sampling Distribution of Mean
μ = 45
X̄ = 43.5
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.061
p.value/2 = 0.144
α/2 = 0.025
p.value/2 = 0.144
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

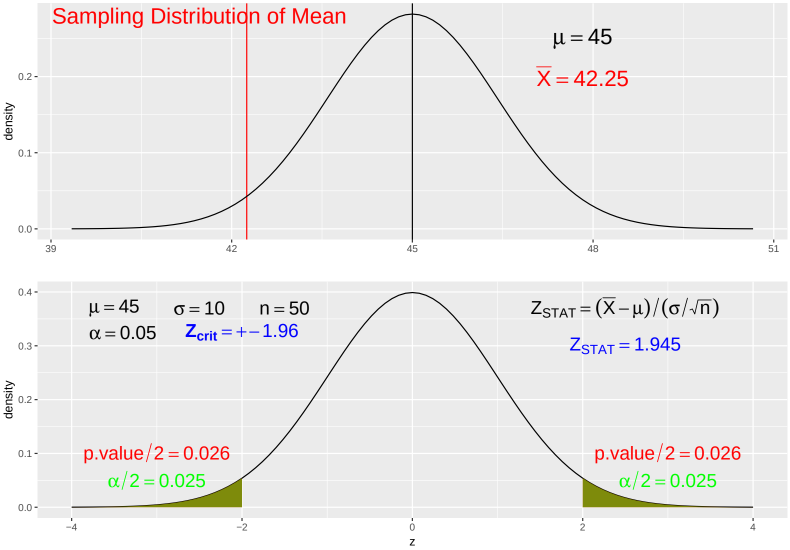

Sampling Distribution of Mean
μ = 45
X̄ = 43.25
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.237
p.value/2 = 0.108
α/2 = 0.025
p.value/2 = 0.108
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
μ = 45
X̄ = 43
density
0.2
0.1
0.0
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.414
p.value/2 = 0.079
α/2 = 0.025
p.value/2 = 0.079
α/2 = 0.025
density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
z

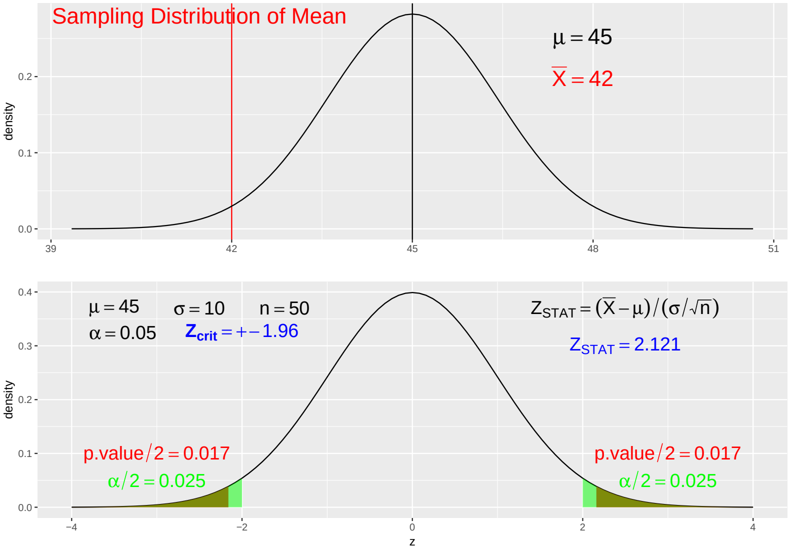

Sampling Distribution of Mean
μ = 45
X̄ = 42.75
density
0.0
0.1
0.2
39
42
45
48
51

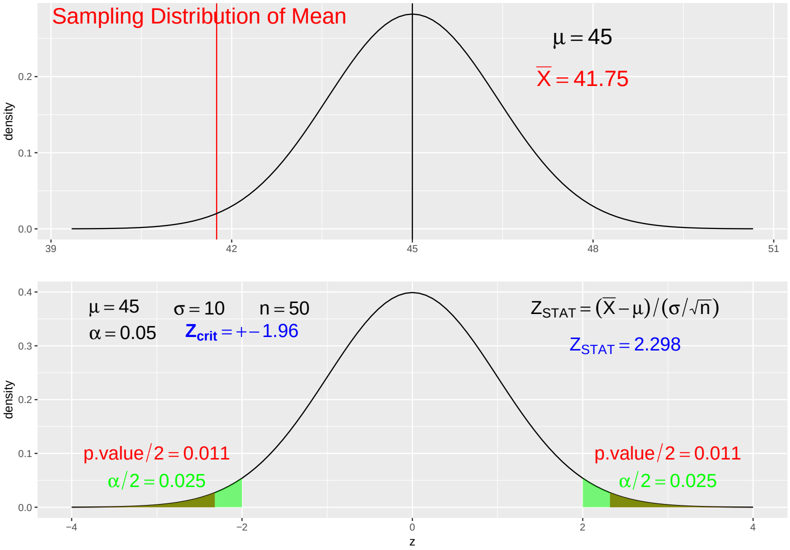

μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.591
p.value/2 = 0.056
α/2 = 0.025
p.value/2 = 0.056
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 42.5$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 1.768$
p.value/2 = 0.039
$\alpha/2 = 0.025$
p.value/2 = 0.039
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

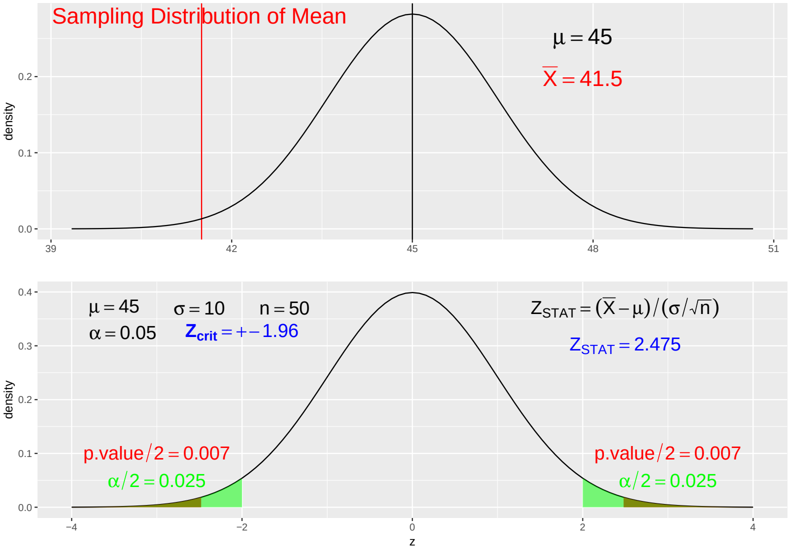


Sampling Distribution of Mean
μ = 45
X̄ = 42.25
density
0.0
0.1
0.2
39
42
45
48
51
μ = 45
σ = 10
n = 50
α = 0.05
Zcrit = +−1.96
ZSTAT = (X̄ − μ)/(σ/√n)
ZSTAT = 1.945
p.value/2 = 0.026
α/2 = 0.025
p.value/2 = 0.026
α/2 = 0.025
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z


Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 42$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 2.121$
p.value/2 = 0.017
$\alpha/2 = 0.025$
p.value/2 = 0.017
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z



Sampling Distribution of Mean
$\mu = 45$
$\overline{X} = 41.75$
density
0.2
0.1
0.0
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$\mathbf{Z_{crit}} = +-1.96$
$Z_{STAT} = (\overline{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 2.298$
p.value/2 = 0.011
$\alpha/2 = 0.025$
p.value/2 = 0.011
$\alpha/2 = 0.025$
density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
z



Sampling Distribution of Mean
$\mu = 45$
$\bar{X} = 41.5$
density
0.0
0.1
0.2
39
42
45
48
51
$\mu = 45$
$\sigma = 10$
$n = 50$
$\alpha = 0.05$
$Z_{crit} = +-1.96$
$Z_{STAT} = (\bar{X} - \mu)/(\sigma/\sqrt{n})$
$Z_{STAT} = 2.475$
p.value/2 = 0.007
$\alpha/2 = 0.025$
p.value/2 = 0.007
$\alpha/2 = 0.025$
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
z

Hypothesis testing with R using "Credit.csv"

```
credit <- read.csv("Credit.csv", stringsAsFactors = F)
str(credit)
```

```
## 'data.frame':    908 obs. of  17 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender  : chr  "Female" "Male" "Female" "Male" ...
##  $ age     : int  20 22 67 26 52 44 47 59 33 44 ...
##  $ agecat  : chr  "18-24" "18-24" "65+" "25-34" ...
##  $ ed      : int  15 17 14 16 14 16 11 19 8 10 ...
##  $ edcat   : chr  "Some college" "College degree" "High school degree" "Some colle
##  $ jobcat  : chr  "Managerial and Professional" "Sales and Office" "Sales and Offi
##  $ empcat  : chr  "Less than 2" "Less than 2" "More than 15" "Less than 2" ...
##  $ retire  : chr  "No" "No" "No" "No" ...
##  $ income  : int  31 15 35 23 77 97 84 47 19 73 ...
##  $ inccat  : chr  "$25 - $49" "Under $25" "$25 - $49" "Under $25" ...
##  $ debtinc: num  11.1 18.6 9.9 1.7 1.9 14.4 4.1 8.6 0.9 2.8 ...
##  $ jobsat  : chr  "Highly dissatisfied" "Highly dissatisfied" "Somewhat satisfied"
##  $ marital: chr  "Unmarried" "Unmarried" "Married" "Married" ...
##  $ homeown: chr  "Rent" "Own" "Own" "Rent" ...
##  $ card    : chr  "Mastercard" "Visa" "Visa" "Discover" ...
##  $ default: chr  "Yes" "Yes" "No" "No" ...
```

R Packages

```r
library(ggplot2)

library(gridExtra)

library(FSA)

library(multcomp)

library(car)

library(fitdistrplus)
```

One sample t-test in R: Two Tailed Test ($\alpha = 0.05$)

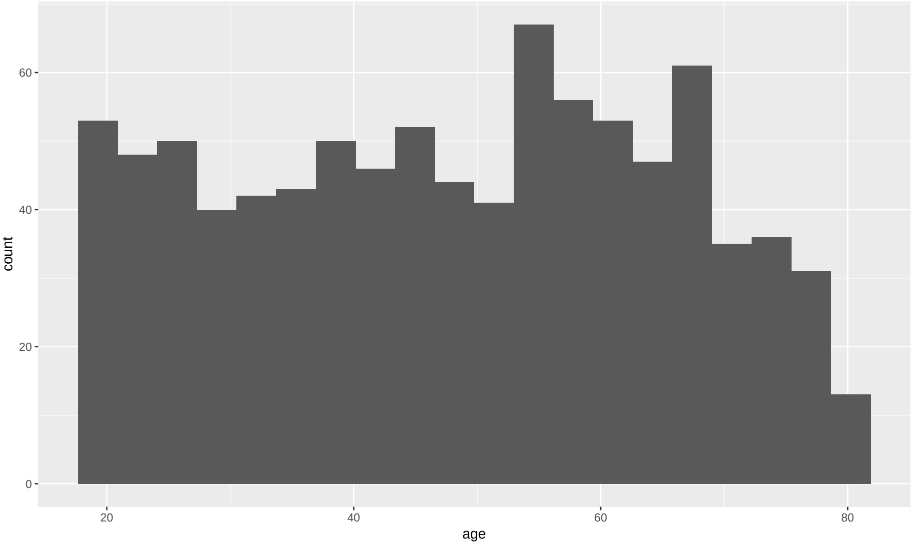
Test whether mean age of borrowers is 49 years old

$$H_0 : \mu_{age} = 49$$
$$H_1 : \mu_{age} \neq 49$$

```r
summary(credit$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   49.00   48.04   62.00   79.00
```

```r
ggplot(credit, aes(x = age)) + geom_histogram(bins = 20)
```


count
60
40
20
0
20
40
60
80
age

One sample t-test in R: Two Tailed Test ($\alpha = 0.05$)

$$H_0 : \mu_{age} = 49$$
$$H_1 : \mu_{age} \neq 49$$

```
t.test(x = credit$age, mu = 49)
```

```
##
##  One Sample t-test
##
## data:  credit$age
## t = -1.6651, df = 907, p-value = 0.09624
## alternative hypothesis: true mean is not equal to 49
## 95 percent confidence interval:
##  46.91011 49.17139
## sample estimates:
## mean of x
##  48.04075
```

if $p - value < \alpha$ then reject the H_0, otherwise do not reject H_0

Conclusion: There is no enough evidence to reject the H_0 and claim that the mean age is not 49 years old

One sample t-test in R: Upper Tailed Test ($\alpha = 0.05$)

Test the claim that mean age of borrowers is greater than 49 years old

$$H_0 : \mu_{age} \leq 49$$
$$H_1 : \mu_{age} > 49$$

```
t.test(x = credit$age, mu = 49, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  credit$age
## t = -1.6651, df = 907, p-value = 0.9519
## alternative hypothesis: true mean is greater than 49
## 95 percent confidence interval:
##  47.09218      Inf
## sample estimates:
## mean of x
##  48.04075
```

if $p - value < \alpha$ then reject the H_0, otherwise do not reject H_0

Conclusion: There is no enough evidence to reject the H_0 and claim that the mean age is greater than 49 years old

One sample t-test in R: Lower Tailed Test ($\alpha = 0.05$)

Test the claim that mean age of borrowers is less than 49 years old

$$H_0 : \mu_{age} \geq 49$$
$$H_1 : \mu_{age} < 49$$

```
t.test(x = credit$age, mu = 49, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  credit$age
## t = -1.6651, df = 907, p-value = 0.04812
## alternative hypothesis: true mean is less than 49
## 95 percent confidence interval:
##      -Inf 48.98932
## sample estimates:
## mean of x
##  48.04075
```

if $p - value < \alpha$ then reject the H_0, otherwise do not reject H_0

Conclusion: There is enough evidence to reject the H_0 and claim that the mean age is less than 49 years old

Two Samples Hypothesis

Assumptions:

- Samples are randomly and independently drawn from two populations
- Populations distributions are normal or both sample size > 30

1. The ***means*** of two ***independent*** populations

$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 > \mu_2$

2. The ***means*** of two ***related or paired*** populations

$H_0 : \mu_D = 0$	$H_0 : \mu_D \geq 0$	$H_0 : \mu_D \leq 0$
$H_1 : \mu_D \neq 0$	$H_1 : \mu_D < 0$	$H_1 : \mu_D > 0$

3. The ***proportions*** of two ***independent*** populations

$H_0 : \pi_1 = \pi_2$	$H_0 : \pi_1 \geq \pi_2$	$H_0 : \pi_1 \leq \pi_2$
$H_1 : \pi_1 \neq \pi_2$	$H_1 : \pi_1 < \pi_2$	$H_1 : \pi_1 > \pi_2$

4. The ***variances*** of two ***independent*** populations

$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$

1. The ***means*** of two ***independent*** populations

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$Z_{STAT} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad t_{STAT} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2. The ***means*** of two ***related or paired*** populations

$H_0 : \mu_D = 0$
$H_1 : \mu_D \neq 0$

$$Z_{STAT} = \frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}} \qquad t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

μ_D - hypothesized mean difference

σ_D - population standard dev. of differences

s_D - sample standard dev. of differences

n - the sample size (number of pairs)

3. The ***proportions*** of two ***independent*** populations

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 \neq \pi_2$

$$Z_{STAT} = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$, $\quad p_1 = \frac{X_1}{n_1}$, $\quad p_2 = \frac{X_2}{n_2}$

Assumptions

$n_1 p_1 \geq 5$, $\quad n_1(1 - p_1) \geq 5$
$n_2 p_2 \geq 5$, $\quad n_2(1 - p_2) \geq 5$

Two Samples Hypothesis: Test Statistic

4. The ***variances*** of two ***independent*** populations

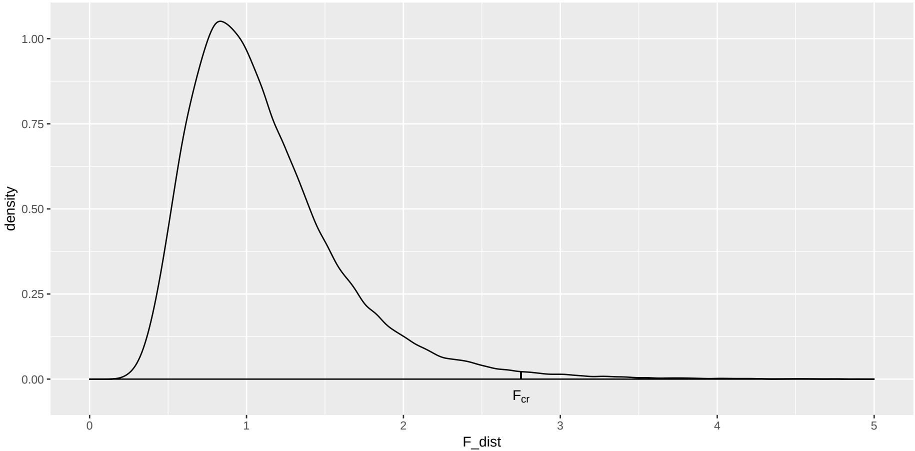
$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F_{Stat} = \frac{S_1^2}{S_2^2}$$

S_1^2 - sample variance of the first sample (the larger sample variance)

S_2^2 - sample variance of the second sample



1.00
0.75
0.50
0.25
0.00
density
F_{cr}
0
1
2
3
4
5
F_dist

The *means* of two independent populations using Countries.csv data

```r
countries <- read.csv("Countries.csv", stringsAsFactors = F)
str(countries)
```

```
## 'data.frame':	61 obs. of  20 variables:
##  $ Country.Name          : chr  "Argentina" "Austria" "Bahamas" "Barbados" ...
##  $ Abbr                  : chr  "ARG" "AUT" "BHS" "BRB" ...
##  $ Region                : chr  "America" "Europe" "America" "America" ...
##  $ Property.Rights       : num  32.4 86 45.3 55.5 50.9 83.3 43.5 25.7 55 62.5 ...
##  $ Judical.Effectiveness : num  39.6 81.8 48.7 33 56.3 69.3 48.7 15.4 49.7 38.9 ...
##  $ Government.Integrity  : num  38.2 75.2 38.2 34.3 37.6 71.5 35 32.6 33.4 41.8 ...
##  $ Fiscal.Health         : num  56.4 79.7 42.3 0 92.8 66.3 60.5 81.4 22.8 86.4 ...
##  $ Business.Freedom      : num  57.3 76.9 68.5 69.6 71.3 82 62.7 58.9 61.3 66.7 ...
##  $ Labor.Freedom_2015    : num  46.1 67.6 71.5 67.7 74.6 61.1 53.6 35.8 52.3 68.3 ...
##  $ Labor.Freedom_2016    : num  52.3 65 88.1 79.3 87.7 ...
##  $ Monetary.Freedom      : num  50.9 83.4 77 83.7 60.4 84.9 79.6 66.4 67 83.3 ...
##  $ Trade.Freedom         : num  66.7 87 50.6 62.2 80.6 87 70.1 76 69.4 87 ...
##  $ Investment.Freedom    : int  50 90 50 75 30 85 50 5 50 70 ...
##  $ Financial.Freedom     : int  50 70 60 60 10 70 50 40 50 60 ...
##  $ GDP.Growth.Rate       : num  1.2 0.9 0.5 0.5 -3.9 1.4 1.5 4.8 -3.8 3 ...
##  $ GDP.per.Capita.PPP    : int  22554 47250 25167 16575 17654 43585 8373 6465 15615 19097 ...
##  $ Unemployment          : num  6.7 5.7 14.4 12.3 6.1 8.7 11.8 3.6 7.2 9.8 ...
##  $ Inflation.Perc        : num  26.5 0.8 1.9 0.5 13.5 0.6 -0.6 4.1 9 -1.1 ...
##  $ FDI.Inflow.Millions   : num  11655 3837 385 254 1584 ...
##  $ Public.Debt.Perc.of.GDP: num  56.5 86.2 65.7 103 59.9 ...
```

Source: The Heritage Foundation and The Wall Street Journal

Test whether the average Unemployment Rate is the same in Europe and America at the $\alpha = 0.05$ level of significance

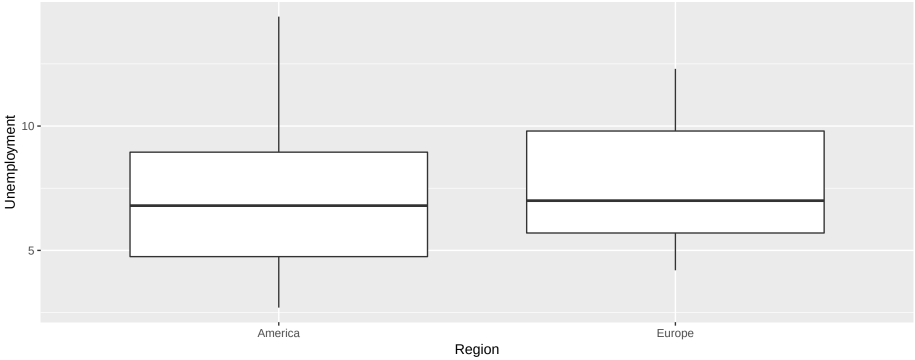
$H_0 : \mu_{Europe} = \mu_{America}$

$H_0 : \mu_{Europe} \neq \mu_{America}$

```
Summarize(data = countries, Unemployment ~ Region, digits = 2)
```

```
##    Region  n mean   sd min   Q1 median   Q3  max
## 1 America 28 7.41 3.45 2.7 4.75    6.8 8.95 14.4
## 2  Europe 33 7.69 2.54 4.2 5.70    7.0 9.80 12.3
```

```
ggplot(countries, aes(x = Region, y = Unemployment)) + geom_boxplot()
```




```r
t.test(data = countries, Unemployment ~ Region)
```

```
##
##  Welch Two Sample t-test
##
## data:  Unemployment by Region
## t = -0.35989, df = 48.803, p-value = 0.7205
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.864859  1.298409
## sample estimates:
## mean in group America  mean in group Europe
##              7.410714              7.693939
```

Test whether the average value of Business Freedom index in Europe is higher than in America at the $\alpha = 0.01$ level of significance

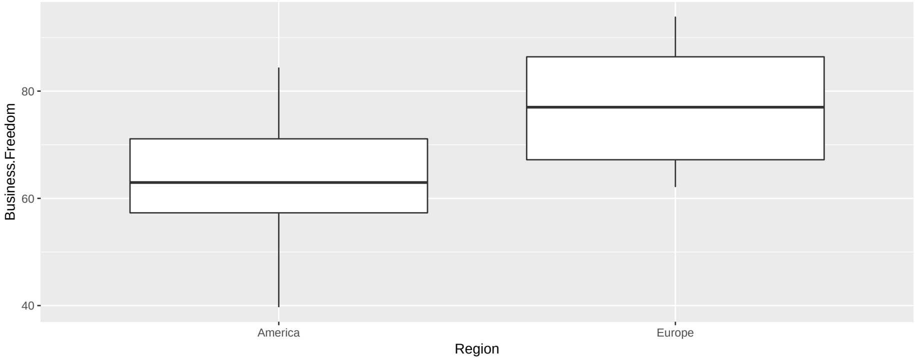
$H_0 : \mu_{America} \geq \mu_{Europe}$

$H_0 : \mu_{America} < \mu_{Europe}$

```
Summarize(data = countries, Business.Freedom ~ Region, digits = 2)
```

```
##    Region  n  mean    sd  min   Q1 median   Q3  max
## 1 America 28 64.34 10.63 39.7 57.3  62.95 71.1 84.4
## 2  Europe 33 76.70  9.85 62.1 67.2  77.00 86.4 93.9
```

```
ggplot(countries, aes(x = Region, y = Business.Freedom)) + geom_boxplot()
```




```r
t.test(data = countries, Business.Freedom ~ Region,  alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Business.Freedom by Region
## t = -4.6841, df = 55.722, p-value = 9.243e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -7.951029
## sample estimates:
## mean in group America  mean in group Europe
##              64.33571              76.70303
```

Is there an evidence that proportion of home owners is different for males and females at the $\alpha = 0.05$ level of significance

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 \neq \pi_2$

```
table(credit$gender, credit$homeown)
```

```
##
##          Own Rent
##   Female 306  171
##   Male   271  160
```

```
prop.test(x = c(306, 171), n = c(577, 331), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(306, 171) out of c(577, 331)
## X-squared = 0.15862, df = 1, p-value = 0.6904
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05378877  0.08121472
## sample estimates:
##    prop 1    prop 2
## 0.5303293 0.5166163
```

p-value > 0.05, so do not reject the null hypothesis

Compare two populations variances

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

```
Summarize(data = countries, Business.Freedom ~ Region, digits = 2)
```

```
##    Region  n  mean    sd  min   Q1 median   Q3  max
## 1 America 28 64.34 10.63 39.7 57.3  62.95 71.1 84.4
## 2  Europe 33 76.70  9.85 62.1 67.2  77.00 86.4 93.9
```

```
var.test(data = countries, Business.Freedom ~ Region)
```

```
##
##  F test to compare two variances
##
## data:  Business.Freedom by Region
## F = 1.1644, num df = 27, denom df = 32, p-value = 0.6749
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5627776 2.4657475
## sample estimates:
## ratio of variances
##           1.164352
```

p-value > 0.05, so do not reject the null hypothesis

One-Way ANOVA test

$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_c$

H_1 : *Not all of the population means are the same*

or

H_1 : *At least one population mean is different*

Assumptions

- Populations are normally distributed or their sample sizes > 30
- Populations have equal variances

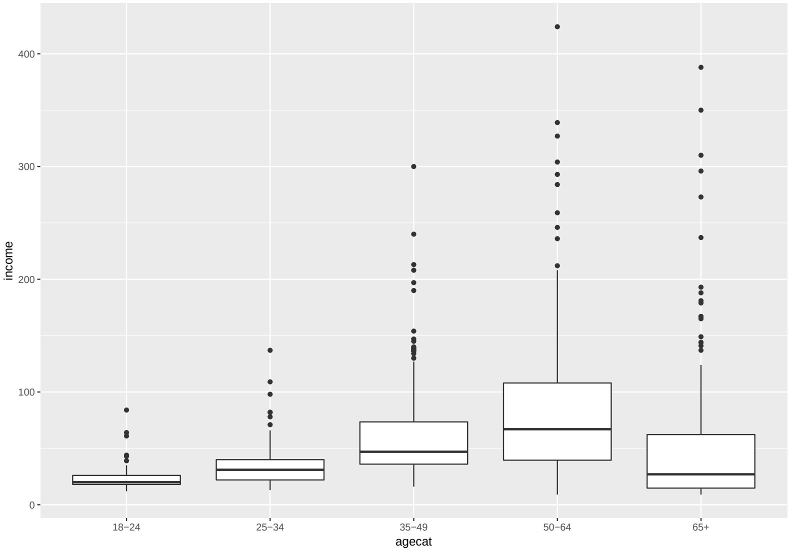
One-Way ANOVA test in R

Test whether the average income differs across age groups at the $\alpha = 0.05$

```
Summarize(income ~ agecat, data = credit, digits = 0)
```

```
##    agecat   n mean sd min Q1 median  Q3 max
## 1   18-24 101   23 10  12 18     20  26  84
## 2   25-34 144   35 18  13 22     31  40 137
## 3   35-49 223   61 41  16 36     47  74 300
## 4   50-64 248   84 65   9 40     67 108 424
## 5     65+ 192   52 63   9 15     27  62 388
```

```
ggplot(credit, aes(x = agecat, y = income)) + geom_boxplot()
```



income
400
300
200
100
0
18−24
25−34
35−49
50−64
65+
agecat

One-Way ANOVA test

Before performing ANOVA test let's check whether the assumption of Homogeneity of Variance is satisfied

Let's use Levene Test for Homogeneity of Variances

```
#library(car)
leveneTest(income ~ agecat, data = credit)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   4  25.674 < 2.2e-16 ***
##       903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P - value < 0.05, so we reject the ull hypothesis and conclude that the Homogeneity of Variance assumption is violated. This means that we have to consider this fact while conducting ANOVA test

One-Way ANOVA test

```r
anova <- aov(income ~ agecat, data = credit)
summary(anova)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)    
## agecat        4  378242   94560   38.26 <2e-16 ***
## Residuals   903 2231745    2471                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
oneway.test(income ~ agecat, data = credit, var.equal=FALSE)
```

```
## 
##  One-way analysis of means (not assuming equal variances)
## 
## data:  income and agecat
## F = 91.092, num df = 4.00, denom df = 445.41, p-value < 2.2e-16
```

p-value is less than alpha (0.05), so we conclude that the mean income of at least one age group is significantly different.

oneway.test() is a function from library(car)

Tukey's Test can be performed to investigate the differences between all age groups

```r
#library(multcomp)
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = income ~ agecat, data = credit)
##
## $agecat
##                    diff        lwr        upr     p adj
## 25-34-18-24  11.51458  -6.121677  29.150830 0.3832111
## 35-49-18-24  37.90010  21.602466  54.197738 0.0000000
## 50-64-18-24  60.99621  44.956685  77.035729 0.0000000
## 65+-18-24    29.13263  12.429891  45.835373 0.0000214
## 35-49-25-34  26.38553  11.858905  40.912146 0.0000081
## 50-64-25-34  49.48163  35.245198  63.718064 0.0000000
## 65+-25-34    17.61806   2.638359  32.597752 0.0117831
## 50-64-35-49  23.09611  10.556109  35.636102 0.0000057
## 65+-35-49    -8.76747 -22.145317   4.610377 0.3792217
## 65+-50-64   -31.86358 -44.925739 -18.801412 0.0000000
```

anova object is created with **aov()** function

p-values of some paired comparisons are less than alpha (0.05), so the mean income of corresponding age groups are significantly different from each other.

How to check for Normality?

Not all continuous random variables are normally distributed

It is important to evaluate how well the data set is approximated by a normal distribution

Construct charts or graphs

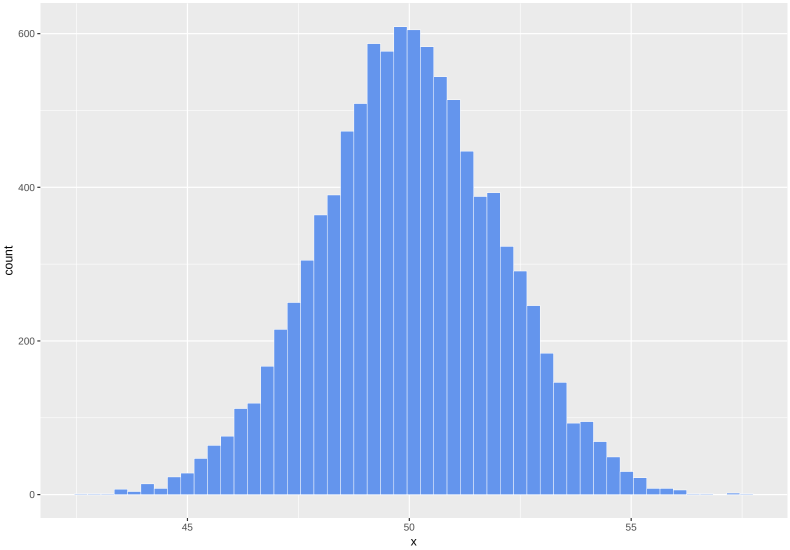
- For small or moderate-sized data sets, do box-and-whisker plot look symmetric?
- For large data sets, does the histogram appears bell-shaped?

Compute descriptive summary measures

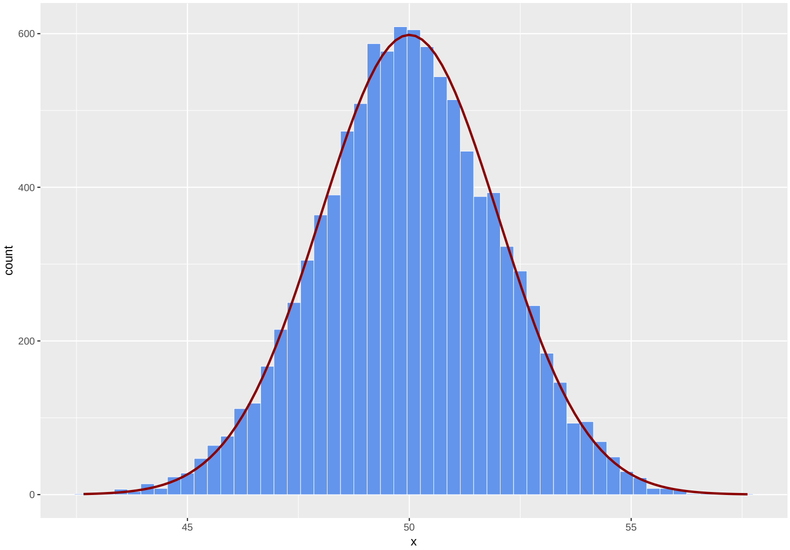
- Do the mean, median and mode have similar values?
- Is the range approximately 6σ?

Use formal hypothesis testing

- Kolmogorov-Smirnov
- Shapiro-Wilk



600
400
200
0
count
45
50
55
x


600
400
200
0
count
45
50
55
x

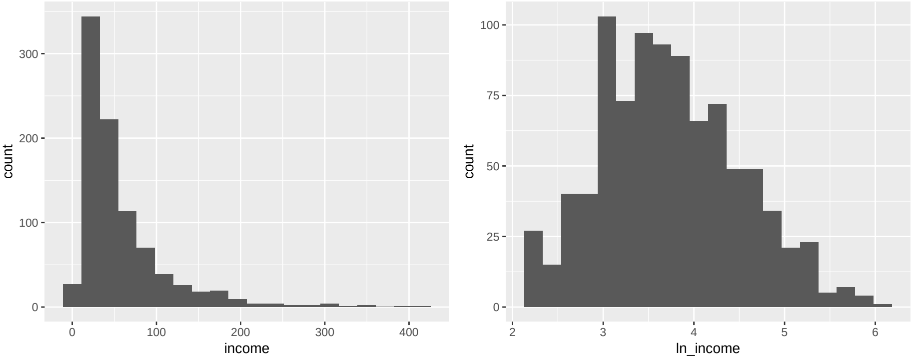
How to check for Normality?

Is the distribution of income normal? maybe log-normal?

Generate a log of income variable and add to existing credit data

```
credit$ln_income = log(credit$income)

grid.arrange(
ggplot(credit, aes(x = income)) + geom_histogram(bins=20),
ggplot(credit, aes(x = ln_income)) + geom_histogram(bins=20),
ncol=2)
```


count
300
200
100
0
0
100
200
300
400
income
count
100
75
50
25
0
2
3
4
5
6
ln_income

How to check for Normality?

Shapiro-Wilk test

H_0 : The distribution is Normal

```
shapiro.test(credit$income)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  credit$income
## W = 0.73629, p-value < 2.2e-16
```

```
shapiro.test(credit$ln_income)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  credit$ln_income
## W = 0.9876, p-value = 5.99e-07
```

Fitting Distributions in R with **fitdistrplus** package

Best theoretical distribution for **Business Freedom index**

```
dist1  <- fitdist(countries$Business.Freedom, "norm")
dist2  <- fitdist(countries$Business.Freedom, "lnorm")
dist3  <- fitdist(countries$Business.Freedom, "logis")
dist4  <- fitdist(countries$Business.Freedom, "weibull")
dist5  <- fitdist(countries$Business.Freedom, "gamma")
distributions <- c("norm", "lnorm", "logis", "Weibull", "gamma")

# Estimated parameters for normal distributions
dist1
```

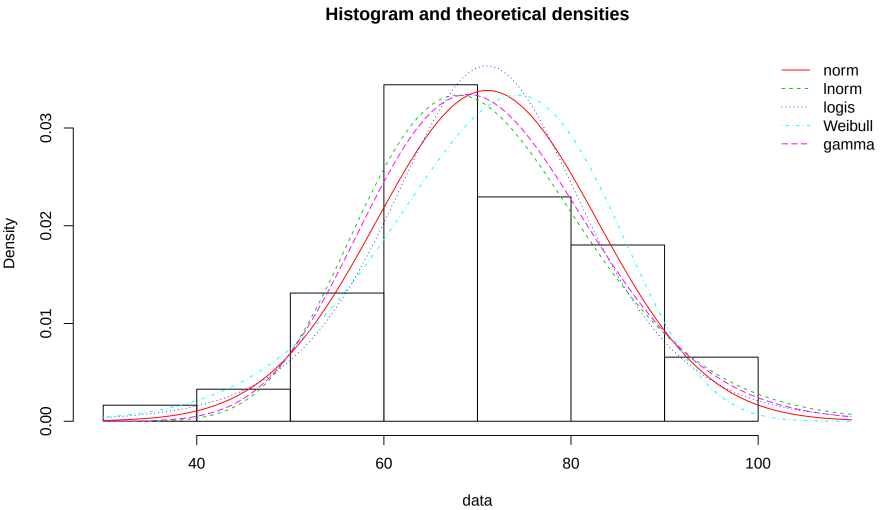
```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters:
##      estimate Std. Error
## mean 71.02623   1.508597
## sd   11.78252   1.066739
```

Goodness of fit measures

```r
gofstat(list(dist1, dist2, dist3, dist4, dist5),
        fitnames = distributions)
```

```
## Goodness-of-fit statistics
##                                    norm      lnorm      logis    Weibull
## Kolmogorov-Smirnov statistic 0.06603402 0.08056946 0.07597859 0.09572056
## Cramer-von Mises statistic   0.04078499 0.04111627 0.05614199 0.07433658
## Anderson-Darling statistic   0.27486665 0.34864405 0.36072733 0.45389158
##                                   gamma
## Kolmogorov-Smirnov statistic 0.07630093
## Cramer-von Mises statistic   0.03504950
## Anderson-Darling statistic   0.28495736
##
## Goodness-of-fit criteria
##                                    norm    lnorm    logis  Weibull
## Akaike's Information Criterion 478.0378 481.6509 480.6054 478.4688
## Bayesian Information Criterion 482.2596 485.8727 484.8272 482.6905
##                                   gamma
## Akaike's Information Criterion 479.8543
## Bayesian Information Criterion 484.0761
```

```r
# Plot empirical and fitted distributions
denscomp(list(dist1, dist2, dist3, dist4, dist5),
        legendtext = distributions, xlim = c(30, 110))
```



According to Goodness of fit measures the most appropriate distribution for the Business Freedom index is the Normal distribution.

Non-Parametric Hypothesis testing

One-sample Wilcoxon signed rank test: Two Tailed Test

Test whether the median value of Public Dept to GDP ratio of American countries is 45%

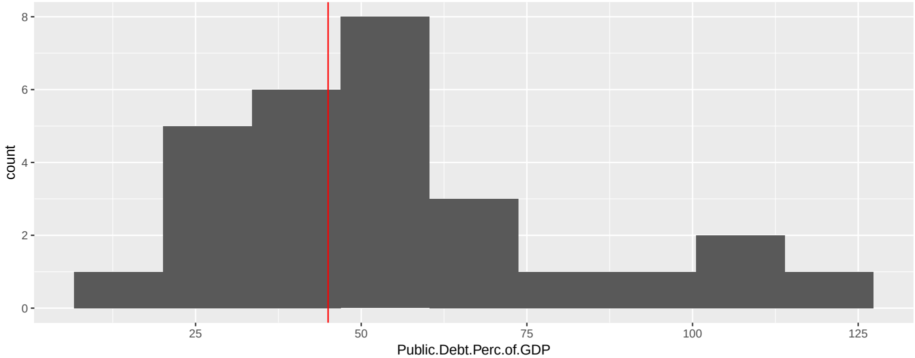
H_0 : $median = 45$

H_1 : $median \neq 45$

```
America <- countries[countries$Region == "America", ]
summary(America$Public.Debt.Perc.of.GDP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.10   34.45   48.80   53.57   62.77  124.30
```

```
ggplot(America, aes(x =  Public.Debt.Perc.of.GDP)) +
  geom_histogram(bins = 9)+
  geom_vline(xintercept = 45, col = "red")
```




```r
wilcox.test(America$Public.Debt.Perc.of.GDP, mu = 45,
            alternative = "two.sided")
```

```
## Warning in wilcox.test.default(America$Public.Debt.Perc.of.GDP, mu = 45, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  America$Public.Debt.Perc.of.GDP
## V = 252, p-value = 0.2694
## alternative hypothesis: true location is not equal to 45
```

p-value > alpha (0.05), so we fail to reject the null hypothesis: The median value of Public Dept to GDP ratio doesn't differ significantly from 45%

How One-sample Wilcoxon signed rank test works

```
df <- America[, c("Country.Name", "Public.Debt.Perc.of.GDP")]
df$Diff <- df$Public.Debt.Perc.of.GDP - 45
df$Rank <- rank(abs(df$Diff))
df$SR <- sign(df$Diff)*df$Rank
df$SR_plus <- ifelse(df$SR > 0, df$SR, NA)
df$SR_minus <- ifelse(df$SR <= 0, -1*df$SR, NA)
head(df)
```

```
##   Country.Name Public.Debt.Perc.of.GDP Diff Rank SR SR_plus SR_minus
## 1    Argentina                    56.5 11.5   13 13      13       NA
## 3      Bahamas                    65.7 20.7   19 19      19       NA
## 4     Barbados                   103.0 58.0   26 26      26       NA
## 7       Belize                    76.3 31.3   24 24      24       NA
## 8      Bolivia                    39.7 -5.3    7 -7      NA        7
## 9       Brazil                    73.7 28.7   23 23      23       NA
```

```
Wilc_test <- min(colSums(df[, c("SR_plus","SR_minus")], na.rm=T))
Wilc_test
```

```
## [1] 154
```

One-sample Wilcoxon signed rank test: Right Tailed Test

H_0 : $median \leq 45$

H_1 : $median > 45$

```
wilcox.test(America$Public.Debt.Perc.of.GDP, mu = 45,
            alternative = "greater")
```

```
## Warning in wilcox.test.default(America$Public.Debt.Perc.of.GDP, mu = 45, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  America$Public.Debt.Perc.of.GDP
## V = 252, p-value = 0.1347
## alternative hypothesis: true location is greater than 45
```

p-value > alpha, Do not reject the null hypothesis: There is no enough evidence to claim that the median value of Public Dept to GDP ratio is significantly higher than 45%

One-sample Wilcoxon signed rank test: Left Tailed Test

H_0 : $median \geq 45$

H_1 : $median < 45$

```
wilcox.test(America$Public.Debt.Perc.of.GDP, mu = 45,
            alternative = "less")
```

```
## Warning in wilcox.test.default(America$Public.Debt.Perc.of.GDP, mu = 45, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  America$Public.Debt.Perc.of.GDP
## V = 252, p-value = 0.8702
## alternative hypothesis: true location is less than 45
```

p-value > 0.05, so do not reject the null hypothesis

Two-samples Mann-Whitney (or Wilcoxon–Mann–Whitney) Test

H_0 :

- The **medians** of two populations are equal
- The **distributions** of two populations are equal
- The **mean ranks** of two populations are equal

H_1

- The **medians** of two populations are not equal
- The **distributions** of two populations are not equal
- The **mean ranks** of two populations are not equal

Two-samples Mann-Whitney (or Wilcoxon–Mann–Whitney) Test: Two Tailed Test

Is there an evidence that the median debt to GDP ratio is different between European and American countries

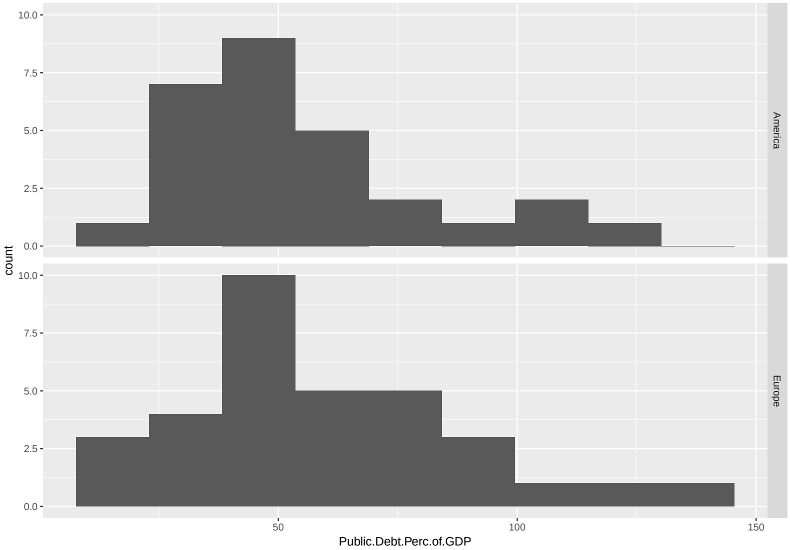
$H_0 : median_{America} = median_{Europe}$

$H_1 : median_{America} \neq median_{Europe}$

```
Summarize(data = countries, Public.Debt.Perc.of.GDP ~ Region, digits = 1)
```

```
##    Region  n mean   sd  min   Q1 median   Q3   max
## 1 America 28 53.6 26.7 17.1 34.5   48.8 62.8 124.3
## 2  Europe 33 59.6 29.9 10.1 40.9   52.6 78.7 132.6
```

```
ggplot(countries, aes(x =  Public.Debt.Perc.of.GDP)) +
  geom_histogram(bins = 9)+
  facet_grid(Region ~.)
```


10.0
7.5
5.0
2.5
0.0
count
America
Europe
50
100
150
Public.Debt.Perc.of.GDP

Two-samples Mann-Whitney (or Wilcoxon–Mann–Whitney) Test: Two Tailed Test

```r
wilcox.test(Public.Debt.Perc.of.GDP ~ Region, data = countries,
            alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Public.Debt.Perc.of.GDP by Region
## W = 401, p-value = 0.3812
## alternative hypothesis: true location shift is not equal to 0
```

P-value $= 0.38$, which means that Null hypothesis is NOT rejected, so there is No enough evidence to claim, that the median value of Debt to GDP ratio is not the same in two regions.

Two-samples Mann-Whitney (or Wilcoxon–Mann–Whitney) Test: Right Tailed Test

H_0 : $median_{America} \leq median_{Europe}$

H_1 : $median_{America} > median_{Europe}$

```
wilcox.test(Public.Debt.Perc.of.GDP ~ Region, data = countries,
            alternative = "greater")
```

```
## Warning in wilcox.test.default(x = c(56.5, 65.7, 103, 76.3, 39.7, 73.7, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Public.Debt.Perc.of.GDP by Region
## W = 401, p-value = 0.8133
## alternative hypothesis: true location shift is greater than 0
```

P-value > alpha, Do Not Reject the Null hypothesis

Two-samples Mann-Whitney (or Wilcoxon–Mann–Whitney) Test: Left Tailed Test

H_0 : $median_{America} \geq median_{Europe}$

H_1 : $median_{America} < median_{Europe}$

```
wilcox.test(Public.Debt.Perc.of.GDP ~ Region, data = countries,
            alternative = "less")
```

```
## Warning in wilcox.test.default(x = c(56.5, 65.7, 103, 76.3, 39.7, 73.7, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Public.Debt.Perc.of.GDP by Region
## W = 401, p-value = 0.1906
## alternative hypothesis: true location shift is less than 0
```

P-value > alpha, Do Not Reject the Null hypothesis

Wilcoxon Signed Rank Test for two related population

H_0 : Difference between the related populations follows a symmetric distribution around zero

H_1 : Difference between the related populations does not follow a symmetric distribution around zero

```
summary(countries$Labor.Freedom_2015)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.50   48.80   60.20   59.06   70.90   91.00
```

```
summary(countries$Labor.Freedom_2016)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.76   52.81   62.55   62.54   72.96   97.34
```

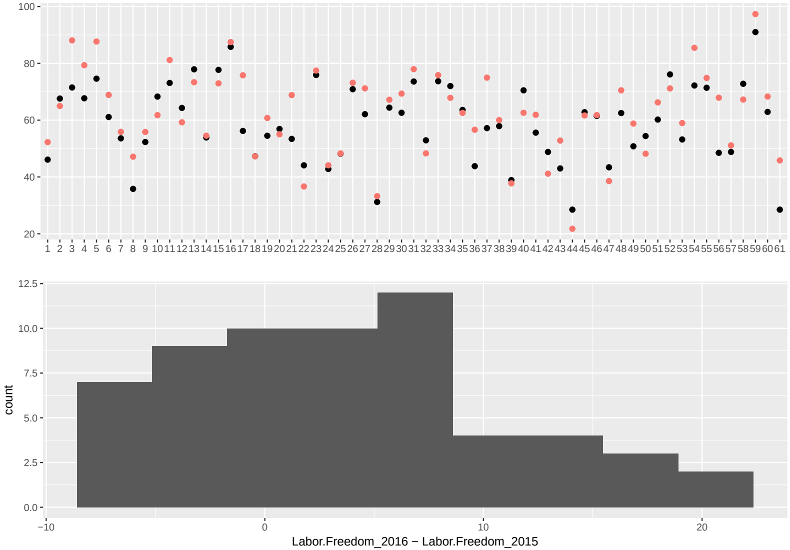
Wilcoxon Signed Rank Test for two related population

```
grid.arrange(

ggplot(countries, aes(x = factor(1:nrow(countries)), y = Labor.Freedom_2015)) +
  geom_point(size = 2) +
  geom_point(aes(y = Labor.Freedom_2016, col="red"), size = 2) +
  xlab("") + ylab("")+
  theme(legend.position = "none"),

ggplot(countries, aes(x = Labor.Freedom_2016 - Labor.Freedom_2015)) +
  geom_histogram(bins = 9),

ncol = 1)
```


count
Labor.Freedom_2016 − Labor.Freedom_2015

Wilcoxon Signed Rank Test for two related population

H_0 : Difference between the related populations follows a symmetric distribution around zero

H_1 : Difference between the related populations does not follow a symmetric distribution around zero

```
wilcox.test(countries$Labor.Freedom_2016, countries$Labor.Freedom_2015,
            paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  countries$Labor.Freedom_2016 and countries$Labor.Freedom_2015
## V = 1387, p-value = 0.001537
## alternative hypothesis: true location shift is not equal to 0
```

Kruskal Wallis Test: One-Way Anova by Ranks

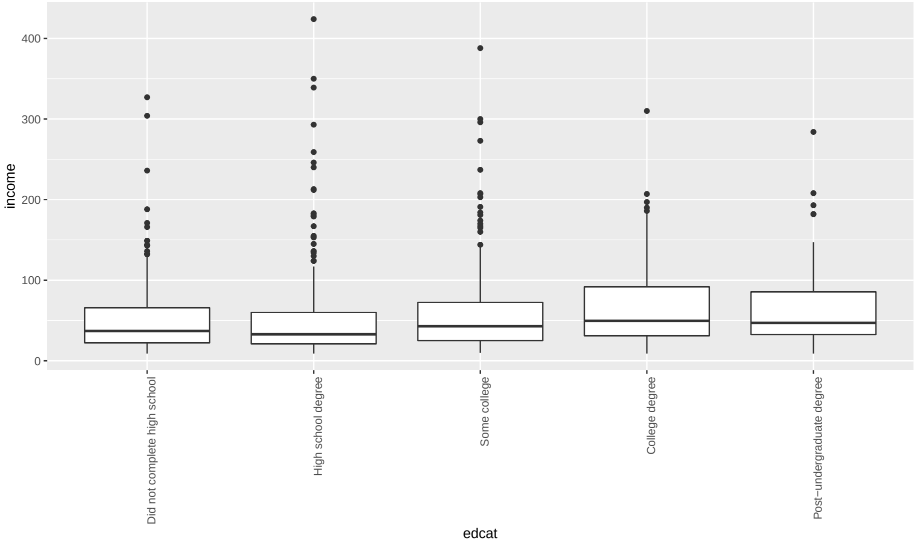
H_0 : The medians of all groups are equal

H_1 : At least one population median is different from at least one other population median

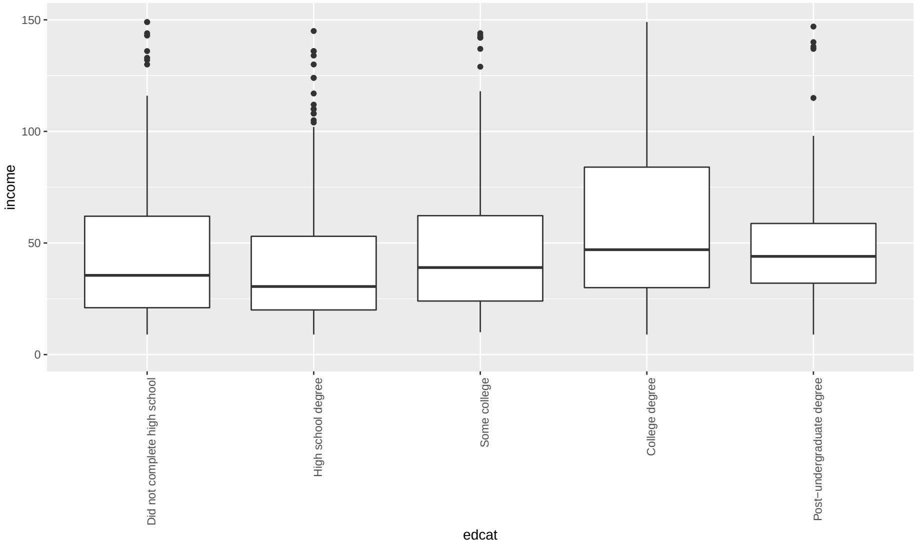
```
Summary <- Summarize(data = credit, income ~ edcat)
Summary[, -1] <- round(Summary[, -1])
Summary
```

```
##                          edcat   n mean sd min Q1 median Q3 max
## 1               College degree 178   66 48   9 31     50 92 310
## 2 Did not complete high school 206   51 46   9 22     37 66 327
## 3           High school degree 299   52 55   9 21     33 60 424
## 4    Post-undergraduate degree  43   71 63   9 32     47 86 284
## 5                 Some college 182   62 60  10 25     43 72 388
```

```
credit$edcat <- factor(credit$edcat,
                       levels = c("Did not complete high school",
                                  "High school degree",  "Some college",
                                  "College degree",
                                  "Post-undergraduate degree"))
ggplot(credit, aes(x = edcat, y = income)) + geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```r
ggplot(credit, aes(x = edcat, y = income)) + geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylim(c(0, 150))
```


income
150
100
50
0
Did not complete high school
High school degree
Some college
College degree
Post−undergraduate degree
edcat

Kruskal Wallis Test: One Way Anova by Ranks

H_0 : The medians of all groups are equal

H_1 : At least one population median is different from at least one other population median

```
kruskal.test(data = credit, income ~ edcat)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  income by edcat
## Kruskal-Wallis chi-squared = 32.318, df = 4, p-value = 1.647e-06
```

p-value $<$ alpha, reject the null hypothesis: there is enough evidence to claim that the median income is different for at least one education category

The proportion of defaults is the same across males and females (alpha = 0.01)

$H_0 : \pi_{male} = \pi_{female}$
$H_1 : \pi_{male} \neq \pi_{female}$

```
cross <- table(credit$gender, credit$default)
cross # Observed frequencies
```

```
##
##           No Yes
##   Female 380  97
##   Male   306 125
```

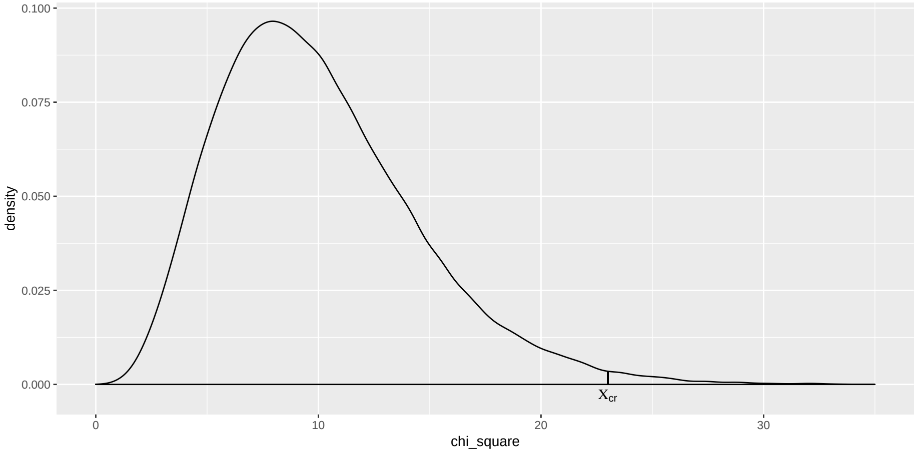
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o = observed frequency in a particular cell
f_e = expected frequency in a particular cell if H_0 is true
$df = (r-1)(c-1)$

Pooled Proportion

$\bar{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{X}{n}$

```
##
##        No       Yes
## 0.7555066 0.2444934
```


0.100
0.075
0.050
0.025
0.000
density
0
10
20
30
chi_square
X_{cr}

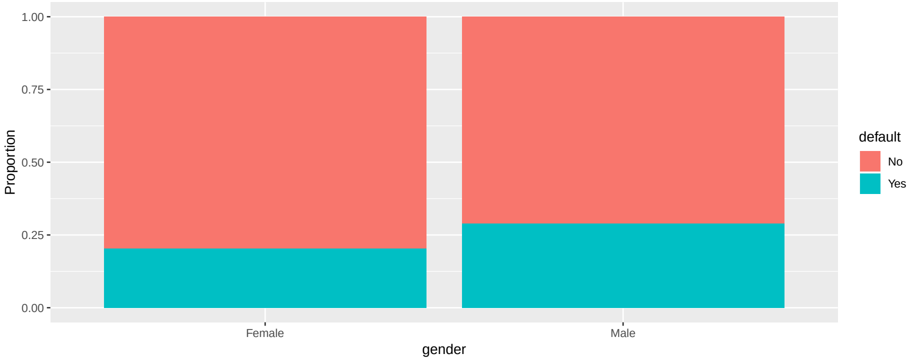
Chi Square test: Compare two populations proportions

The proportion of defaults is the same across males and females (alpha $= 0.01$)

```
prop.table(cross, 1)
```

```
##
##                  No       Yes
##   Female 0.7966457 0.2033543
##   Male   0.7099768 0.2900232
```

```
cross_df <- data.frame(prop.table(cross, 1))
colnames(cross_df) <- c("gender", "default", "Proportion")
ggplot(cross_df, aes(x = gender, y = Proportion, fill = default)) +
  geom_bar(stat = "identity")
```




```r
chisq.test(cross)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cross
## X-squared = 8.7441, df = 1, p-value = 0.003106
```

p-value $<$ alpha, reject the null hypothesis at the 0.01 level of significance.

The proportion of defaults is significantly different across males and females

The proportion of defaults is the same across married and not married customers (alpha 0.05)

$H_0 : \pi_{married} = \pi_{unmarried}$
$H_1 : \pi_{married} \neq \pi_{unmarried}$

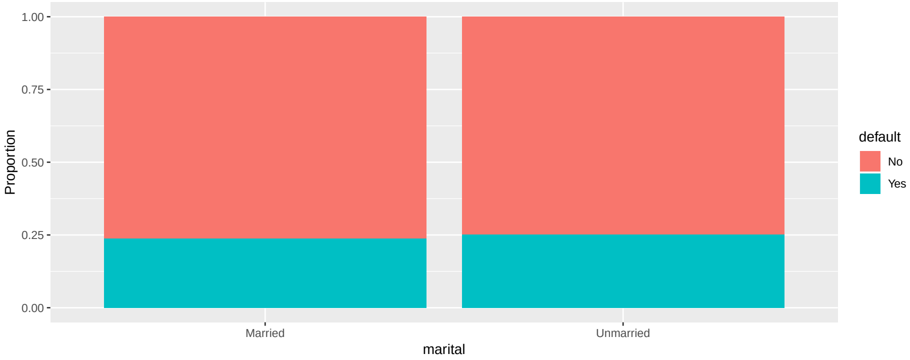
```
cross <- table(credit$marital, credit$default)
cross
```

```
##
##              No Yes
##   Married   344 107
##   Unmarried 342 115
```

```
prop.table(cross, 1)
```

```
##
##                    No       Yes
##   Married   0.7627494 0.2372506
##   Unmarried 0.7483589 0.2516411
```

```
cross_df <- data.frame(prop.table(cross, 1))
colnames(cross_df) <- c("marital", "default", "Proportion")
ggplot(cross_df, aes(x = marital, y = Proportion, fill = default)) +
  geom_bar(stat = "identity")
```




```r
chisq.test(cross)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cross
## X-squared = 0.18254, df = 1, p-value = 0.6692
```

p-value > alpha, do not reject the null hypothesis at the 0.05 level of significance.

The proportion of defaults is not significantly different across married and unmarried customers

Chi Square test: Independence of two categorical variables

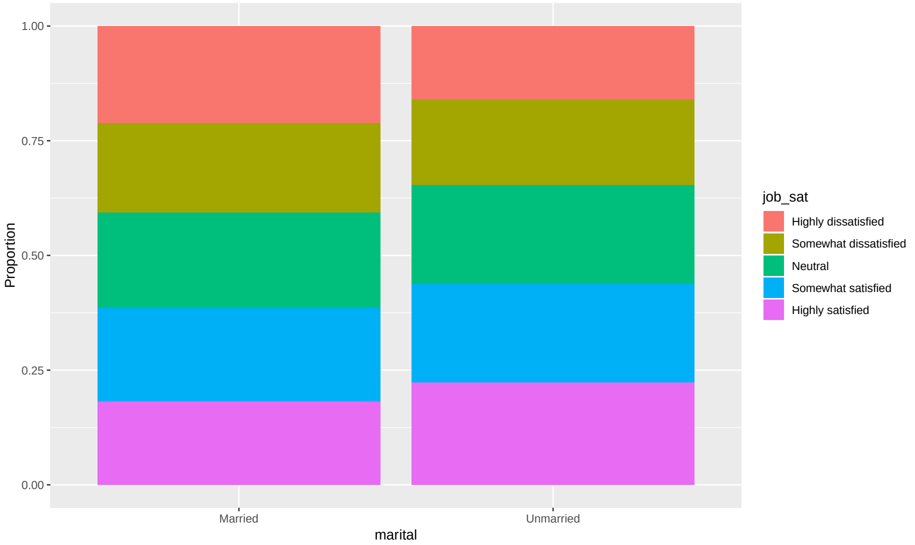
Job satisfaction and marital status are related (alpha 0.05)

H_0 : Job satisfaction and marital status are not related
H_1 : Job satisfaction and marital status are related

```r
credit$jobsat <- factor(credit$jobsat,
                        levels = c ("Highly dissatisfied",
                                    "Somewhat dissatisfied",
                                    "Neutral",
                                    "Somewhat satisfied",
                                    "Highly satisfied"))
cross <- table(credit$jobsat, credit$marital)
round(prop.table(cross, 2), 3)
```

```
##
##                         Married Unmarried
##   Highly dissatisfied     0.211     0.160
##   Somewhat dissatisfied   0.195     0.186
##   Neutral                 0.208     0.217
##   Somewhat satisfied      0.204     0.214
##   Highly satisfied        0.182     0.223
```

```r
cross_df <- data.frame(prop.table(cross, 2))
colnames(cross_df) <- c("job_sat", "marital", "Proportion")
ggplot(cross_df, aes(x = marital, y = Proportion, fill = job_sat)) +
  geom_bar(stat = "identity")
```


1.00
0.75
0.50
0.25
0.00
Proportion
Married
Unmarried
marital
job_sat
Highly dissatisfied
Somewhat dissatisfied
Neutral
Somewhat satisfied
Highly satisfied

Chi Square test: Independence of two categorical variables

```r
chisq.test(cross)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cross
## X-squared = 5.3865, df = 4, p-value = 0.2499
```

p-value > alpha, Do not reject the null hypothesis at the 0.05 level of significance.

There is no significant evidance to claim that Job satisfaction and marital status are related

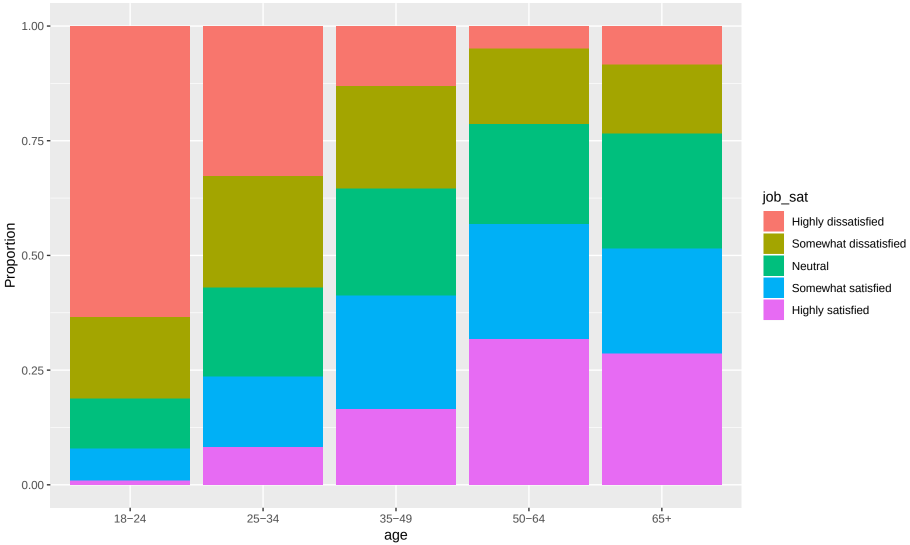
Job satisfaction and age are related (alpha 0.05)

H_0 : Job satisfaction and age are not related
H_1 : Job satisfaction and age are related

```
cross <- table(credit$jobsat, credit$agecat)
round(prop.table(cross, 2),3)
```

```
##
##                          18-24 25-34 35-49 50-64   65+
##   Highly dissatisfied   0.634 0.326 0.130 0.048 0.083
##   Somewhat dissatisfied 0.178 0.243 0.224 0.165 0.151
##   Neutral               0.109 0.194 0.233 0.218 0.250
##   Somewhat satisfied    0.069 0.153 0.247 0.250 0.229
##   Highly satisfied      0.010 0.083 0.166 0.319 0.286
```

```
cross_df <- data.frame(prop.table(cross, 2))
colnames(cross_df) <- c("job_sat", "age", "Proportion")
ggplot(cross_df, aes(x = age, y = Proportion, fill = job_sat)) +
  geom_bar(stat = "identity")
```


1.00
0.75
0.50
0.25
0.00
Proportion
18−24
25−34
35−49
50−64
65+
age
job_sat
Highly dissatisfied
Somewhat dissatisfied
Neutral
Somewhat satisfied
Highly satisfied

Chi Square test: Independence of two categorical variables

```r
chisq.test(cross)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cross
## X-squared = 246.47, df = 16, p-value < 2.2e-16
```

p-value $<$ alpha, Reject the null hypothesis at the 0.05 level of significance.

There is significant evidance to claim that Job satisfaction and age are related

All you need to know for testing hypotheses in R

Descriptive Analysis

- Summarize(), summary()
- table(), prop.table()
- ggplot()

Parametric Hypothesis testing

- t.test()
- aov(), oneway.test()
- TukeyHSD()
- prop.test()
- var.test(), leveneTest()

Non-Parametric Hypothesis testing

- wilcox.test()
- kruskal.test()
- chisq.test()

Decision Rule

- p-value < alpha, Reject the Null Hypothesis

Thank You!

Questions?