Intro to Inferential Statistics with R

Workshop 4

Course: VSK1004 Applied Researcher

Workshop structure

Intro to Statistic Inference	More about inferential Statistics	Linear and Logistic regression
1. Descriptive vs Inferential statistics 2. Population, sample and sampling distribution 3. Null Hypothesis testing 4. Correlation and interpretation	1. Choosing a statistical test 2. Paired t-test 3. Anova 4. Chi-squared distribution	1. Model assumptions 2. Interpretations

Our goal in the next 40 min

In this session, we will cover some other **statistical procedures for hypothesis testing (quantitative research)**:

1. Choosing a statistical test
2. Paired t-test
3. Chi-square test for independence
4. ANOVA

1. Choosing a Statistical test for your research

Many possibilities

- Estimate Population Proportion
- Estimate Population Mean
- One sample Proportion
- Two sample Proportions
- One sample t (Mean)
- Unpaired sample t
- Paired sample t
- Correlation test
- One-Way ANOVA
- Two-Way ANOVA
- Chi-Square Test
- One Sample Variance
- Two Sample Variance
- Wilcoxon rank-sum test

Most common test for quantitative research

- Estimate Population Proportion
- Estimate Population Mean
- One sample Proportion
- Two sample Proportions
- One sample t (Mean)
- Unpaired sample t
- Paired sample t
- Correlation test
- One-Way ANOVA
- Two-Way ANOVA
- Chi-Square Test
- One Sample Variance
- Two Sample Variance
- Wilcoxon rank-sum test

Monday

- Estimate Population Proportion
- Estimate Population Mean
- One sample Proportion
- Two sample Proportions
- One sample t (Mean)
- Unpaired sample t
- Paired sample t
- Correlation test

- One-Way ANOVA
- Two-Way ANOVA
- Chi-Square Test
- One Sample Variance
- Two Sample Variance
- Wilcoxon rank-sum test

Today

- Estimate Population Proportion
- Estimate Population Mean
- One sample Proportion
- Two sample Proportions
- One sample t (Mean)
- Unpaired sample t
- Paired sample t
- Correlation test
- One-Way ANOVA
- Two-Way ANOVA
- Chi-Square Test
- One Sample Variance
- Two Sample Variance
- Wilcoxon rank-sum test

What is your purpose for research question?

- **Comparison:**
  - Is there a differences between groups?
    - e.g. females vs. males
    - e.g. control group vs. treatment groups
    - e.g. grouping individuals by color preferences (yellow, blue)
- In this different examples, we have, at least, two groups and we attempt to find the differences

- **Relationship:**
  - Is there a connection?
    - e.g. what is the equation relating height & flexibility
    - e.g. can age predict muscle mass?
    - e.g. is medication dosage linked to recovery time
- In this different examples, we are seeking out correlation or causation from one variable to the other

What is your purpose for research question?

- **Comparison**:
  - Is there a differences between groups?
    - e.g. females vs. males
    - e.g. control group vs. treatment groups
    - e.g. grouping individuals by color preferences (yellow, blue)
- In this different examples, we have, at least, two groups and we attempt to find the differences

- **Relationship:**
  - Is there a connection?
    - e.g. what is the equation relating height & flexibility
    - e.g. can age predict muscle mass?
    - e.g. is medication dosage linked to recovery time
- In this different examples, we are seeking out **correlation** or **relationship** from one variable to the other

Type of Data you are looking at:

- **Categorical:**
  - Qualitative characteristics:
    - Mortality Rate (death/survival)
    - Patient Falls Rate (fall/not fall)
    - Which gene was expressed?
- **Continuous**
  - Quantitative or numerical:
    - Heart Rate
    - Age
    - Blood pressure

3 families of statistical tests

- Chi-squared
- t-test
- correlation

Purpose

- **Comparison:**
  - Any difference?
- **Relationship:**
  - Any connection?

Type of Data

- **Categorical:**
  - No quantitative meaning
- **Continuous:**
  - Quantitative meaning

Purpose

- Comparison:
  - Any difference?
- **Relationship:**
  - Any connection?

Chi-Squared Family

Type of Data

- Categorical:
  - No quantitative meaning
- **Continuous:**
  - Quantitative meaning

Purpose

- Comparison:
  - Any difference?
- **Relationship:**
  - Any connection?

t- Test Family

Type of Data

- Categorical:
  - No quantitative meaning
- Continuous:
  - Quantitative meaning

Purpose

- **Comparison:**
  - Any difference?
- **Relationship:**
  - Any connection?

Correlation Family

Type of Data

- **Categorical:**
  - No quantitative meaning
- **Continuous:**
  - Quantitative meaning

3 families of statistical tests

- **Chi-squared:**
  - Comparison
  - Categorical *only*
- **t-Test:**
  - comparison
  - categorical *and* continuous
- **Correlation**
  - Relationship
  - continuous *only*

3 families of statistical tests

- **Chi-squared:**
  - Any number of levels/groups:
    - Chi-squared test of homogeneity
    - Chi-squared test of independence
- **t-Test:**
  - 1 level/group:
    - one-sample t-test
  - 2 levels/groups:
    - two-sample unpaired t-test
    - two-sample paired t-test
  - 3+ levels/groups:
    - one-way ANOVA
- **Correlation:**
  - 1 independent and 1 dependent variable:
    - Pearson's correlation
    - Regression

3 families of statistical tests

- **Chi-squared:**
  - Any number of levels/groups:
    - Chi-squared test of homogeneity
    - Chi-squared test of independence
- **t-Test:**
  - 1 level/group:
    - one-sample t-test
  - 2 levels/groups:
    - two-sample unpaired t-test
    - two-sample paired t-test
  - 3+ levels/groups:
    - one-way ANOVA
- **Correlation:**
  - 1 independent and 1 dependent variable:
    - Pearson’s correlation
    - Regression

2. Paired (Dependent) sample *t*-test

Paired-samples t-test: Example

A study was designed to see if XYZ drug was effective at improving their IQ. 20 patients took IQ exam and we recorded their results. The next day, the same patients received drug XYZ, took again a IQ exam and we recorded their results.

Paired-samples t-test: Paired data

As the name implies, paired data come in pairs. That is, two measurements are made on the same individual (before and after, for example) or on a linked pair of individuals (father and son, for example)

Paired-samples t-test: Research question

Is there any improvement in patient IQ score once they took the XYZ drug?

Paired-samples t-test: hypotheses

H_0: $\mu_2 = \mu_1$ (no change in their IQ)

H_a: $\mu_2 > \mu_1$, (better IQ)

Paired-samples *t*-test: Data (IQ Scores)

H_0: $\mu_2 = \mu_1$ (no change in their IQ)

H_a: $\mu_2 > \mu_1$, (better IQ)

PatientsiD	IQ_Before	IQ_After
1	101	113
2	124	127
3	89	89
4	57	70
5	135	127
6	98	104
7	69	69
8	105	127
9	114	115
10	106	99
11	97	104
12	121	120
13	93	95
14	116	129
15	102	106
16	71	71
17	88	94
18	108	112
19	144	154
20	99	96

Paired-samples *t*-test: Compute the differences between each pair

H_0: $\mu_2 = \mu_1$ (no change in their IQ)

H_a: $\mu_2 > \mu_1$, (better IQ)

PatientsiD	IQ_Before	IQ_After	Differences
1	101	113	12
2	124	127	3
3	89	89	0
4	57	70	13
5	135	127	-8
6	98	104	6
7	69	69	0
8	105	127	22
9	114	115	1
10	106	99	-7
11	97	104	7
12	121	120	-1
13	93	95	2
14	116	129	13
15	102	106	4
16	71	71	0
17	88	94	6
18	108	112	4
19	144	154	10
20	99	96	-3

Paired-samples *t*-test: T statistics formula.

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}}$$

Mean differences!

PatientsiD	IQ_Before	IQ_After	Differences
1	101	113	12
2	124	127	3
3	89	89	0
4	57	70	13
5	135	127	-8
6	98	104	6
7	69	69	0
8	105	127	22
9	114	115	1
10	106	99	-7
11	97	104	7
12	121	120	-1
13	93	95	2
14	116	129	13
15	102	106	4
16	71	71	0
17	88	94	6
18	108	112	4
19	144	154	10
20	99	96	-3

Paired-samples *t*-test: Compute the mean (*m*) and (sd) of the column *differences*.

```
# Table of the mean and sdv of the differences
summarise(IQStudy,
    count = n(),
    mean = mean(Differences, na.rm = TRUE),
    sd = sd(Differences, na.rm = TRUE)
  )
```

count <int>	mean <dbl>	sd <dbl>
20	4.2	7.266361

$$t = \frac{\bar{x} - 0}{\frac{\boxed{s}}{\sqrt{n}}}$$

PatientsiD	IQ_Before	IQ_After	Differences
1	101	113	12
2	124	127	3
3	89	89	0
4	57	70	13
5	135	127	-8
6	98	104	6
7	69	69	0
8	105	127	22
9	114	115	1
10	106	99	-7
11	97	104	7
12	121	120	-1
13	93	95	2
14	116	129	13
15	102	106	4
16	71	71	0
17	88	94	6
18	108	112	4
19	144	154	10
20	99	96	-3

Paired-samples *t*-test: Compute the t statistic value

```
# calculate the t
n = mean(IQStudy$Differences)-0 # numerator
d = sd(IQStudy$Differences)/sqrt(20) # denominator
t = n/d
t
# Table of the mean and sdv of the differences
summarise(IQStudy,
    count = n(),
    mean = mean(Differences, na.rm = TRUE),
    sd = sd(Differences, na.rm = TRUE)
  )

```
```

```
[1] 2.584921
```

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$
```

Paired-samples *t*-test: Compute the t statistic value

```r
# calculate the t
n = mean(IQStudy$Differences)-0 # numerator
d = sd(IQStudy$Differences)/sqrt(20) # denominator
t = n/d
t
# Table of the mean and sdv of the differences
summarise(IQStudy,
    count = n(),
    mean = mean(Differences, na.rm = TRUE),
    sd = sd(Differences, na.rm = TRUE)
  )
```
```

[1] 2.584921
```

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$

***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Paired-samples *t*-test: Compute the t statistic value

Degrees of freedom = n(number of patients)-1 = 19
Level of significance = .05 (Interval confidence 95%)

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

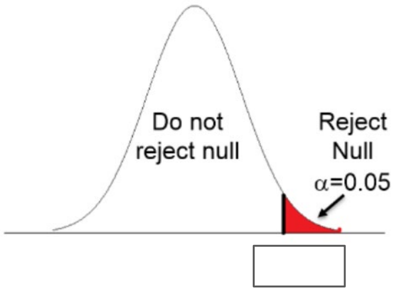
Paired-samples *t*-test: Compute the t statistic value

Degrees of freedom = n(number of patients)-1 = 19
Level of significance = .05 (Interval confidence 95%)
One-tailed paired t-test

H_0: $\mu_b = \mu_a$ (m = 0), (no change in their IQ)

H_a: $\mu_b > \mu_a$ (**better IQ**)

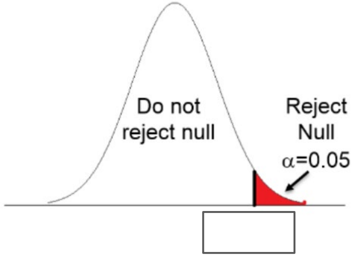
$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$



Paired-samples *t*-test: Compute the t statistic value

Degrees of freedom = n(number of patients)-1 = 19
Level of significance = .05 (Interval confidence 95%)
One-tailed paired t-test

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539

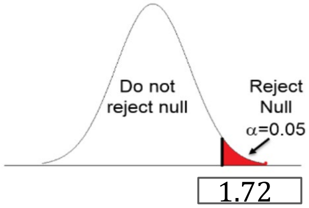
Paired-samples *t*-test: Compute the t Statistic

Step	Result
Null (H_0)	No change in IQ
Alternative (H_a)	Better IQ
Level significance (α)	0.05 level
Critical values	[1.7249]
Test statistic	2.5849
p-value	0.00908
Decision	Reject Ho

```r
# Compute paired t test
t.test(IQStudy$IQ_After, # after sample
       IQStudy$IQ_Before, # before sample
       alternative = 'greater',
       paired = TRUE)
```

```
        Paired t-test

data:  IQStudy$IQ_After and IQStudy$IQ_Before
t = 2.5849, df = 19, p-value = 0.00908
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.724718      Inf
sample estimates:
mean of the differences
                    4.2
```

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}} \qquad t = \frac{4.2 - 0}{\frac{7.266361}{\sqrt{20}}} = 2.585$$



Paired-samples t-test: Decision p-value approach

Since the p-value is less than alpha(α), we reject the H_0.

There is enough evidence to suggest that treatment (XYZ drug) has achieved better change (i.e. patients after treatment scores got higher than before the treatment).

4. ANOVA: one-way

ANOVA: Analysis of the Variance

$$V(X) = \frac{\sum(X - \bar{X})^2}{n - 1}$$

ANOVA: Analysis Of Sum of Squares

$$SST = \sum(X - \bar{X})^2$$

ANOVA: Analysis Of Sum of Squares

$$SST = \sum(X - \bar{X})^2$$

Example:
Find the total SS for the following two samples

A: {2,2,3,5}

```
> mean_A
[1] 3
```

ANSWER:

$SST_A = (-1)^2 + (-1)^2 + (0)^2 + (2)^2 = 6$

B:{4,10,13}

```
> mean_B
[1] 9
```

ANSWER:

$SST_B = (-5)^2 + (1)^2 + (4)^2 = 42$

One-way ANOVA: Example 1

Scores from a stats test (9 students):
{1,3,4,5,5,5,6,7,9}

STT = 42

Stream I
{1,5,9}

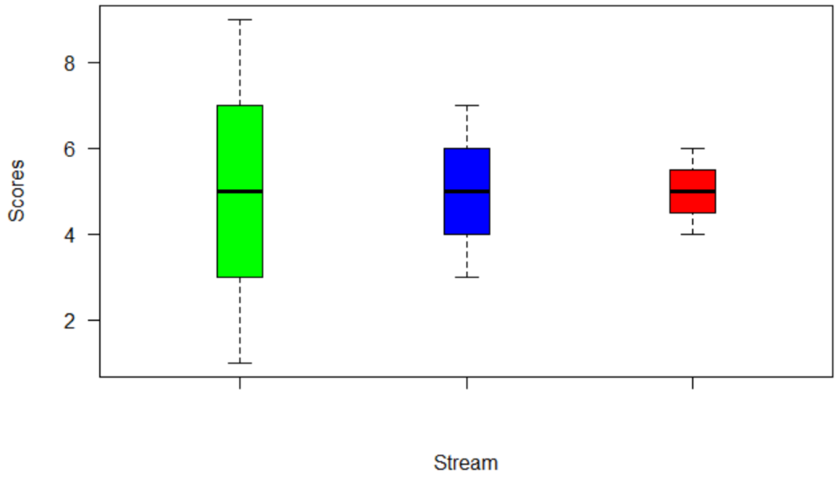
Stream II
{4,5,6}

Stream III
{3,5,7}

One-way ANOVA: Example 1

STT = 42

Stream I	Stream II	Stream III
{1,5,9}	{4,5,6}	{3,5,7}



One-way ANOVA: Example 1

STT = 42

Stream I	Stream II	Stream III
{1,5,9}	{4,5,6}	{3,5,7}

$\bar{x}_{I} = 5$ $\bar{x}_{II} = 5$ $\bar{x}_{III} = 5$

SSW = Sum of squares within groups = $\sum(X - \bar{X}_i)^2$

SSB = Sum of squares between groups = $\sum(\bar{X}_i - \bar{\bar{X}})^2$

$= n(\bar{X}_i - \bar{\bar{X}})^2$

One-way ANOVA: Example 1

STT = 42

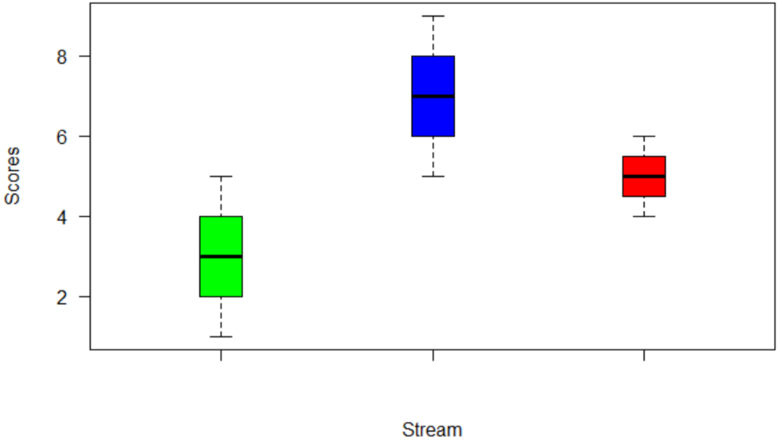
Stream I	Stream II	Stream III
{1,5,9}	{3,5,7}	{4,5,6}
$\bar{x}_I = 5$	$\bar{x}_{II} = 5$	$\bar{x}_{III} = 5$

	Stream I	Stream II	Stream III	
SSW =	$(-4)^2+0^2+4^2$	$(-2)^2+0^2+2^2$	$(-1)^2+0^2+1^2$	
	32	8	2	$= 42$
SSB =	$3(0)^2$	$3(0)^2$	$3(0)^2$	
	0	0	0	$= 0$

One-way ANOVA: Example 2

STT = 42

Stream I	Stream II	Stream III
{1,3,5}	{5,7,9}	{4,5,6}



One-way ANOVA: Example 2

STT = 42

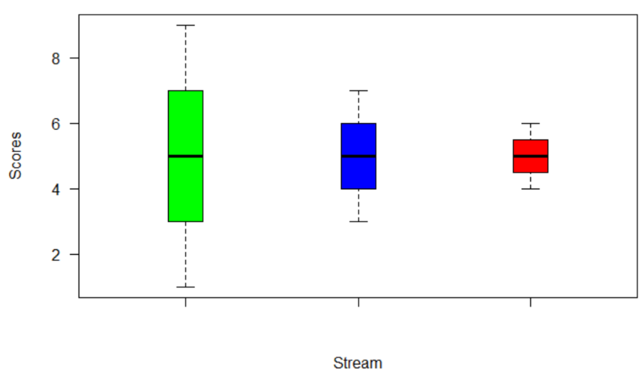
	Stream I	Stream II	Stream III	
	{1,3,5}	{5,7,9}	{4,5,6}	
	$\bar{x}_I = 3$	$\bar{x}_{II} = 7$	$\bar{x}_{III} = 5$	
SSW =	$(-2)^2+0^2+2^2$	$(-2)^2+0^2+2^2$	$(-1)^2+0^2+1^2$	$\boxed{= 18}$
	8	8	2	
SSB =	$3(3-5)^2$	$3(7-5)^2$	$3(5-5)^2$	$\boxed{= 24}$
	12	12	0	

One-way ANOVA: Example 2

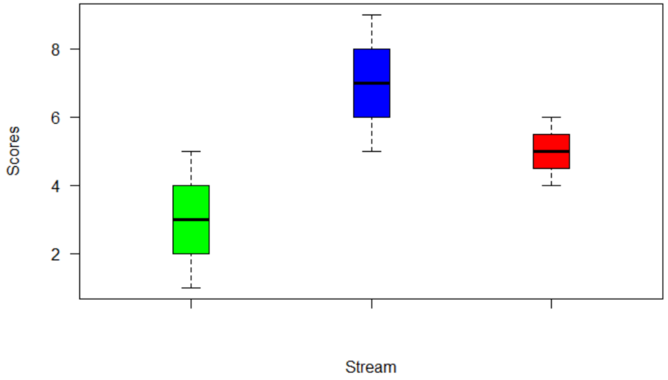
SST = SSW + SSB

$$F = \frac{MSB}{MSW} = \frac{SSB/(c-1)}{SSW/(n-c)}$$

One-way ANOVA: Example 1 vs Example 2



n= 9, c = 3



SSW = 42
SSB = 0

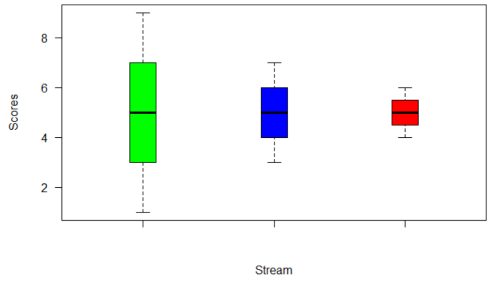
F = 0

$$F = \frac{MSB}{MSW} = \frac{SSB/(c-1)}{SSW/(n-c)}$$

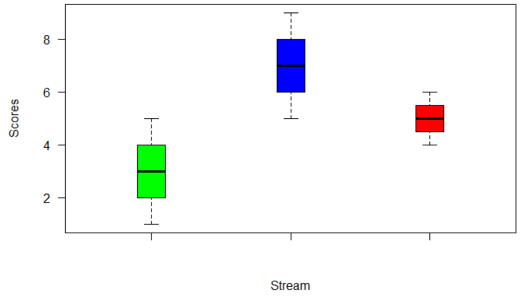
SSW = 18
SSB = 24

F = 4.0

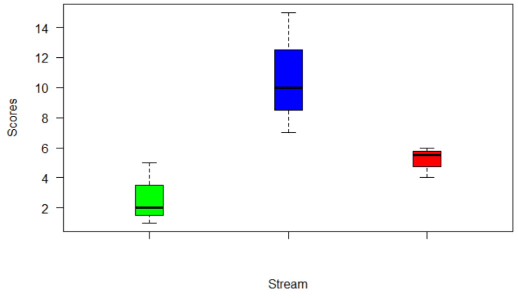
One-way ANOVA: Example 1 vs Example 2 vs Example 3



F = 0



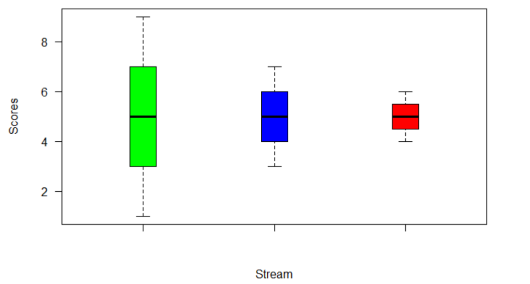
F = 4.0

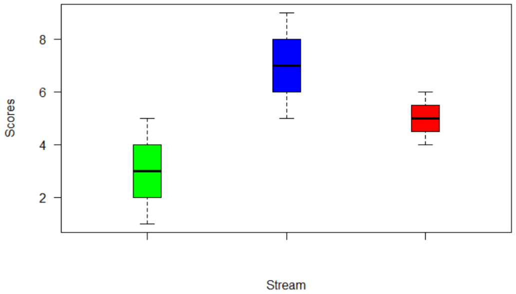


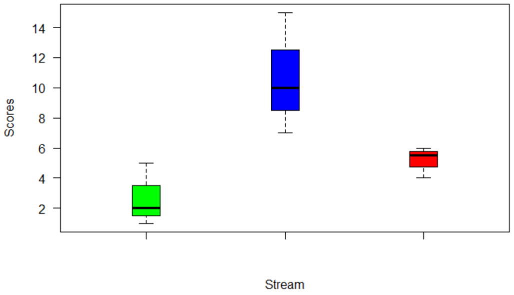
F = 46.2

One-way ANOVA: Example 1 vs Example 2 vs Example 3

$$H_0: \mu_I = \mu_{II} = \mu_{III}$$







F = 0
(p =1.000)

Do not reject Ho (at 5%)

F = 4.0
(p = 0.07087)

Do not reject Ho (at 5%)

F = 46.2
(p = 0.0002)

Reject Ho (at 5%)

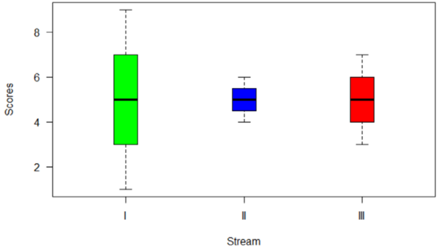
Compute one-way ANOVA in R

````
```{r}
# create vectors
streams <- c('I','II','III') # streams categorical variable
scores <- c(1,4,3,5,5,5,9,6,7) # scores numerical variable

# create data frame as combination both vectors (= columns)
data.scores <- data.frame(streams,
                          scores,
                          stringsAsFactors = TRUE) # convert to factor

# check the data frame characteristics
str(data.scores) # check structure of dataframe
levels(data.scores$streams) # check the levels of the variable streams

# compute descriptive stats
library(dplyr) # import the library dplyr
summarise(group_by(data.scores,streams,
                   count = n(),
                   mean = mean(scores, na.rm = TRUE), # if NA
                   sd = sd(scores, na.rm = TRUE)))
# box plot score by group
boxplot(scores~streams, data.scores,
        col = c("green","blue", "red"),
        ylab = "Scores", xlab="Stream",
        las = 1,
        boxwex = 0.2)

# compute one-way anova
score.aov <- aov(scores ~ streams, data.scores)
summary(score.aov) # summarize the analysis of variance model.
```
````

streams	scores
I	1
I	5
I	9
II	4
II	5
II	6
III	3
III	5
III	7

streams <fctr>	count <int>	mean <dbl>	sd <dbl>
I	9	5	2.291288
II	9	5	2.291288
III	9	5	2.291288

3 rows




```
          Df Sum Sq Mean Sq F value Pr(>F)
streams    2      0       0       0      1
Residuals  6     42       7
```

3. Chi Square test for Independence

Chi Square test for Independence:

- The Chi-Square Test for Independence evaluates the relationship between two variables
- It is a nonparametric test that is performed on categorical(nominal) data.
- Null Hyothesis is No relationship or No Differences

Example:

We conduct a survey with 500 Data Science graduate students (boys and girls) and we asked which is their favourite course: statistics, computer science, or Ethics & Responsibility. We would like to know if there is any relationship between gender and favourite course. We use a significant level of 5%.

Source: https://www.youtube.com/watch?v=LE3AIyY_cn8

Data Collected: Contingency Table

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

Chi-square test for independence (Steps):

1. Define Null and Alternative Hypotheses
2. Looking for critical value:
   a) State Alpha
   b) Calculate degrees of freedom
   c) Look at chi square table
3. State Decision Rule
4. Calculate chi square statistic
5. State Results and Conclusion

Step 1: Define Null and Alternative hypotheses:

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

Ho: Gender and favourite course are not related (no relationship)

Ha: Gender and favorite course are related

Step 2: a) State alpha: 0.05

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

How confident should you be in your test result?

Level of significance, commonly accepted 5%, then alpha = 0.05.

Step 2: b) Calculate the Degrees of Freedom

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

$$df = (rows - 1)(columns - 1)$$

$$df = (2 - 1)(3 - 1)$$

$$df = (1)(2) = 2$$

Step 2: c) Look at chi-square table

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	$\boxed{5.99}$	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81

18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49

Step 3: State Decision Rule

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

Critical value approach:

If χ^2 is greater than 5.99 then, reject H_0

Step 3: State Decision Rule

P-value value approach?

Step 3: State Decision Rule

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

p-value value approach:

If p-value is smaller than level of significance, then reject H_0

i.e. the relationship is significant (we are unlikely to have got that by chance)

Step 5: Calculate Chi square statistic

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

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

where

$f_o = Observed\ frequencies$

$f_e = Expected\ frequencies$

Step 5: Calculate Chi square statistic

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

$$\chi^2 = \Sigma \frac{(f_O - f_e)^2}{f_e}$$

where

f_o = *Observed frequencies*

f_e = *Expected frequencies*

Step 5: Calculate Chi square statistic

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

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

$$f_e = \frac{(f_c f_r)}{n}$$

where

f_o = *Observed frequencies*

f_e = *Expected frequencies*

Step 5: Calculate Chi square statistic

	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	N = 500

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

$$f_e = \frac{(f_c f_r)}{n}$$

where

f_o = *Observed frequencies*

f_e = *Expected frequencies*

Step 5: Calculate Chi square statistic (*fe*)

Observed table	Statistics	Computer Science	Ethics and Responsibility	TOTAL
Boys	100	150	20	270
Girls	20	30	180	230
TOTAL	120	180	200	n = 500

Expected table	Statistics			**TOTAL**
Boys $f_e = \frac{(f_c f_r)}{n}$	(120*270)/500 = **64.8**			**270**
TOTAL	**120**	**180**	**200**	**n = 500**

Step 5: Calculate Chi square statistic (*fe* and *fo*)

Observed table (fo)	Statistics	Computer Science	Ethics and Responsibility	**TOTAL**
Boys	100	150	20	**270**
Girls	20	30	180	**230**
TOTAL	**120**	**180**	**200**	**n = 500**

Expected table (fe)	Statistics	Computer Science	Ethics and Responsibility	**TOTAL**
Boys	64.8	97.2	108	**270**
Girls	55.2	82.8	92	**230**
TOTAL	**120**	**180**	**200**	**n = 500**

Step 5: Calculate Chi square statistic (*fe* and *fo*)

Observed (Expected)	Statistics	Computer Science	Ethics and Responsibility	**TOTAL**
Boys	100 (64.8)	150 (97.2)	20(108)	**270**
Girls	20 (55.2)	30 (82.8)	180 (92)	**230**
TOTAL	**120**	**180**	**200**	**n = 500**

$$\chi^2 = \frac{(f_O - f_e)^2}{f_e} = \frac{(100 - 64.8)^2}{64.8} + \frac{(20 - 55.2)^2}{55.2} + \frac{(150 - 97.2)^2}{97.2} + \cdots + \frac{(180 - 92)^2}{92}$$

$$\chi^2 = 259.8$$

Step 5: State the results

Step	Result
Null (H_0)	Gender and favourite color are not related
Alternative (H_a)	Gender and favourite color are related
Level significance (α)	0.05 level
Degrees of freedom *(df)*	2
Chi-square	259.8
p-value	.00000000000000022
Decision	Reject Ho

```r
# Create contingency table
ResponsableDS <-  as.table(rbind(c(100, 150, 20), c(20, 30, 180)))
dimnames(ResponsableDS) <- list(gender = c("Boys", "Girls"),       # row names
                     course= c("Statistics", "Computer Science","Ethics & Responsability"))
```

```r
# use chisq() function
(Xsq <- chisq.test(ResponsableDS))  # Prints test summary
Xsq$observed
Xsq$expected
```

```
      course
gender  Statistics Computer Science Ethics & Responsability
  Boys         100              150                      20
  Girls         20               30                     180
```

```
      course
gender  Statistics Computer Science Ethics & Responsability
  Boys        64.8             97.2                     108
  Girls       55.2             82.8                      92
```

```
	Pearson's Chi-squared test

data:  ResponsableDS
X-squared = 259.8, df = 2, p-value < 2.2e-16
```

Step 6: State the results

"A chi-square test of independence was performed to examine the relation between gender and the favorite course within Data Science Graduate Program. As the p-value is smaller than the .05 significance level, we do reject the null hypothesis that the gender and favorite course are not related and therefore, we can conclude that there is a statistically significant relationship between them".