Fourth Industrial Summer School

Day 1: Data Analysis Foundations - Morning

Statistical Concepts for Predictive Modelling

Session Objectives

- ✓ Predictive modeling
- ✓ Experimentation
- ✓ Hypothesis testing
- ✓ Statistical analysis
- ✓ Correlation

Predictive Modeling

Predictive Modeling

- **Predictive modeling** is a branch of advanced analytics that uses data and machine learning techniques to forecast future events or behaviors based on past data patterns and trends.
- Analyzing historical data to identify patterns and trends and then using those insights to make predictions about future outcomes.
- Applications
  - Weather forecasting
  - Email spam filtering

Predictive Modeling Steps

- **General** steps in most predictive modeling pipeline:

1. Data collection
2. Data preprocessing
3. Exploratory data analysis (EDA)
4. Model building
5. Validation and evaluation
6. Deployment

Challenges

- Data is **messy**. Vast majority of work that goes into conducting successful analyses lies in preprocessing data.

- Modeling and analysis require **multiple passes** over the data.
  - Choosing the right features, picking the right algorithms, running the right significance tests, and finding the right hyperparameters all require experimentation.

- Models become part of a production service and may need to be **rebuilt** periodically or even in real time.

Predictive Modeling Techniques

- Data scientist techniques
  - **Supervised / Unsupervised learning**
  - **Resampling Methods**
  - **Dimension Reduction**
  - **……**

Predictive Hypothesis

- Predictive hypothesis
  - **Independent** variables
  - **Dependent** variable
- Hypothesis formulation / Testing hypothesis / Statistical analysis
  - **Experimentation**

Data Types

Data Types

Categorical data

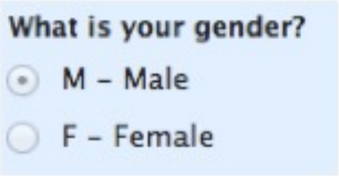
- Nominal
- Ordinal

Continuous data

- Interval
- Ratio

Nominal Scale

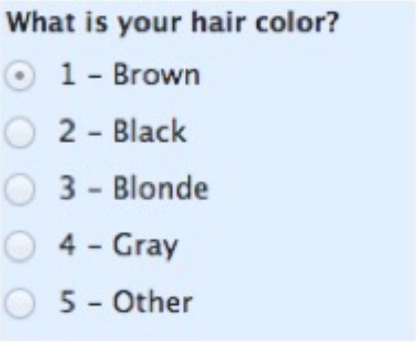
- **Nominal**
  - The nominal scale is the least powerful of the scale types. It only maps the attribute of the entity into a name or symbol.
  - The nominal scale differentiates between items or subjects based only on their names or categories and other qualitative classifications they belong to.
  - Nominal scales could simply be called “labels”
  - Examples of a nominal scale are gender, language, ethnicity etc



What is your gender?

● M – Male

○ F – Female



What is your hair color?

● 1 – Brown

○ 2 – Black

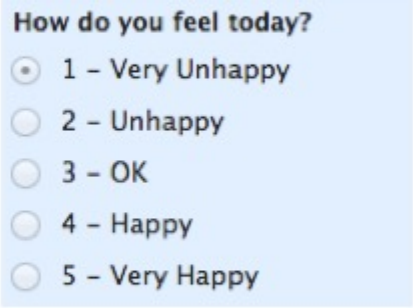
○ 3 – Blonde

○ 4 – Gray

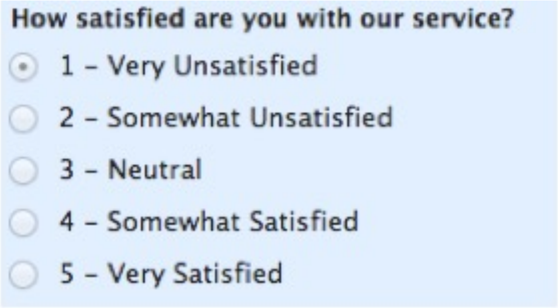
○ 5 – Other

Ordinal Scale

- **Ordinal**
  - The ordinal scale ranks the entities after an ordering criterion and is therefore more powerful than the nominal scale.
  - An ordinal variable, is one where the order matters but not the difference between values.
  - Examples of ordering criteria are “greater than”, “better than”, and “more complex”.
  - Other examples of an ordinal scale are grades and software complexity.


How do you feel today?

◉ 1 – Very Unhappy
○ 2 – Unhappy
○ 3 – OK
○ 4 – Happy
○ 5 – Very Happy


How satisfied are you with our service?

◉ 1 – Very Unsatisfied
○ 2 – Somewhat Unsatisfied
○ 3 – Neutral
○ 4 – Somewhat Satisfied
○ 5 – Very Satisfied

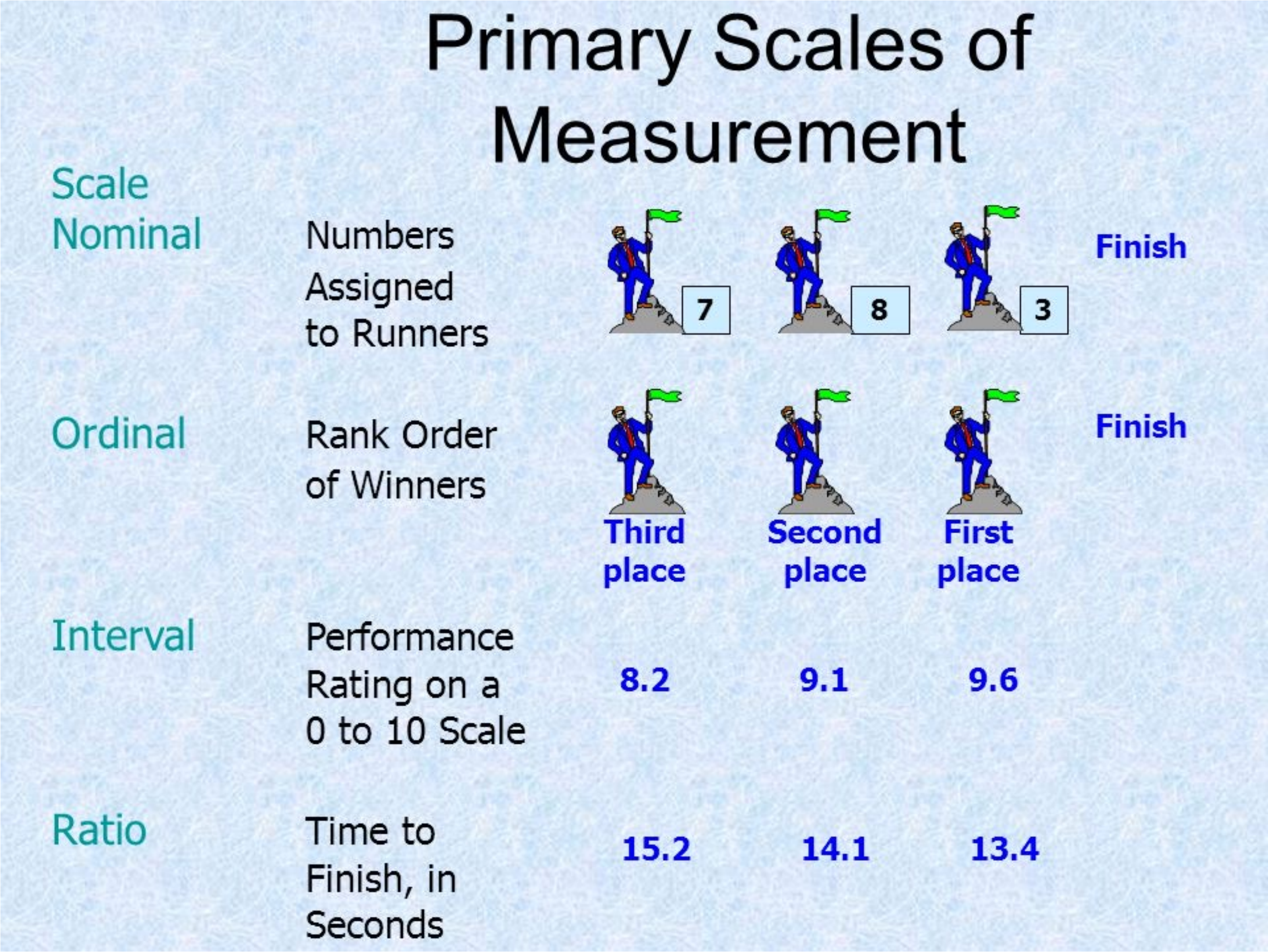
Interval Scale

- **Interval**
  - Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values.
  - This scale, orders the values in the same way as the ordinal scale but there is a notion of “relative distance” between two entities (i.e., where the difference between two values is meaningful).
  - Examples of an interval scale are temperature measured in Celsius
    - With an interval scale, you know not only whether different values are bigger or smaller, you also know *how much* bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Ratio Scale

- **Ratio**
  - If there exists a **meaningful zero value** and the **ratio between two measures is meaningful**, a ratio scale can be used.
  - It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured.
  - Example
    - Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents
    - Time is ratio since 0 time is meaningful.
    - Height and Weight

Scales Example



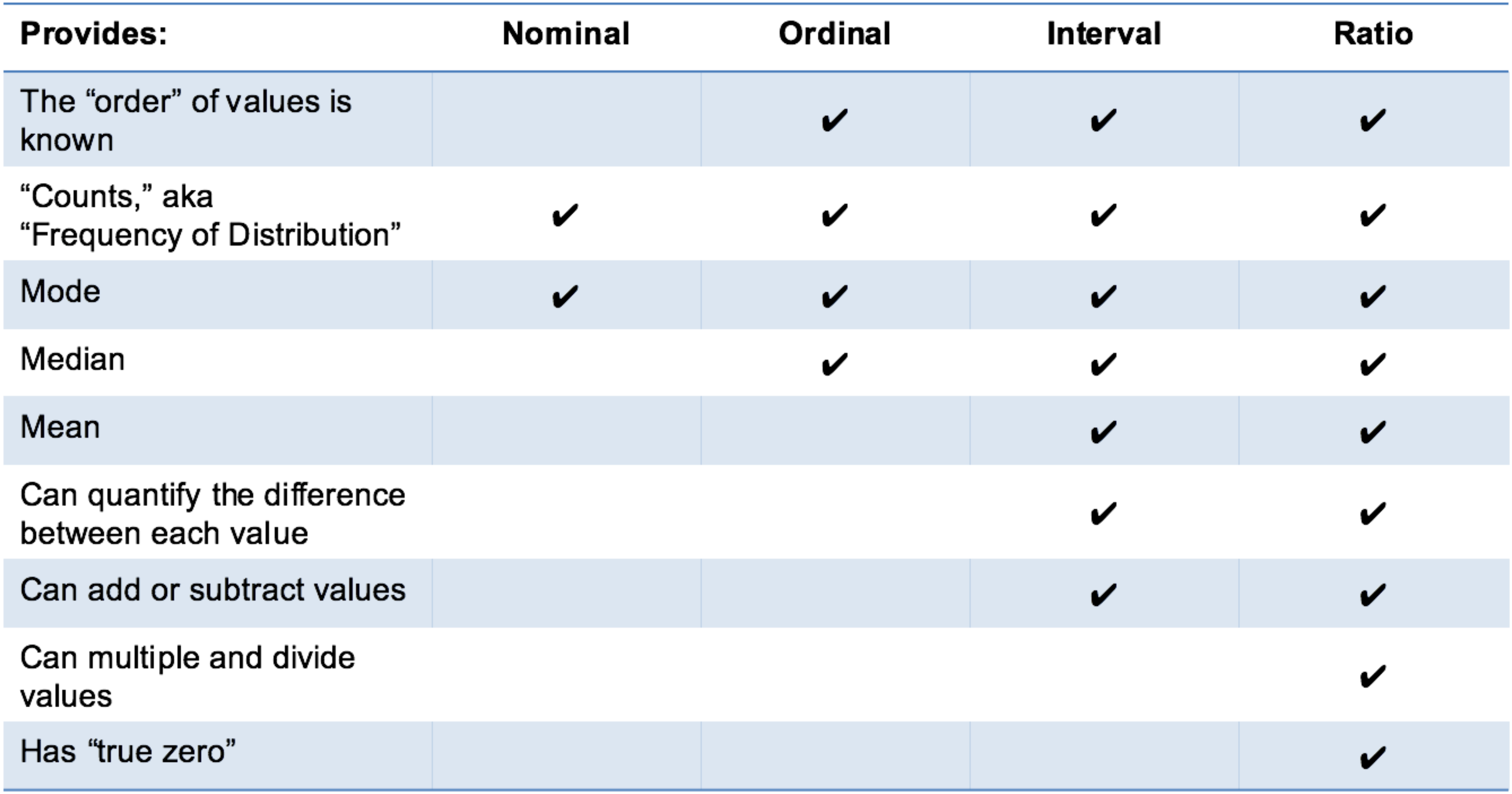
Primary Scales of Measurement

Scale					
Nominal	Numbers Assigned to Runners	7	8	3	Finish
Ordinal	Rank Order of Winners	Third place	Second place	First place	Finish
Interval	Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6	
Ratio	Time to Finish, in Seconds	15.2	14.1	13.4	

Scales Summary

Scale Type	Characterization	Example (generic)	Example (CS)
Nominal	Divides the set of objects into categories, with no particular ordering among them	Labeling, classification	Name of programming language, name of defect type
Ordinal	Divides the set of entities into categories that are ordered	Preference, ranking, difficulty	Ranking of failures (as measure of failure severity)
Interval	Comparing the differences between values is meaningful	Calendar time, temperature (Fahrenheit, Celsius)	Beginning and end date of activities (as measures of time distance)
Ratio	There is a meaningful “zero” value, and ratios between values are meaningful	Length, weight, time intervals, absolute temperature (Kelvin)	Lines of code (as measure of attribute “Program length/size”)

Scales Operations



Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Scales Examples

- For the next questions
  - think about possible answers with the data type.
- What is the type of your smartphone?
- KFUPM has a green campus?
- How many children you have?
- Assign a letter grade to a student in class?
- App star rating ★★★★☆

Experiment
(Cause-Effect)

Experiment Principles

- Experimentation is not simple; we have to **prepare**, **conduct** and **analyze** experiments properly.

- One of the main advantages of an experiment is the **control** of, for example, subjects (e.g., people), objects (e.g., tools) and instrumentation (e.g., questions).

- Other advantages include ability to perform **statistical analysis** using **hypothesis testing** methods and opportunities for **replication**.

- To ensure that we make use of the advantages, we need a process supporting us in our objectives in doing experiments

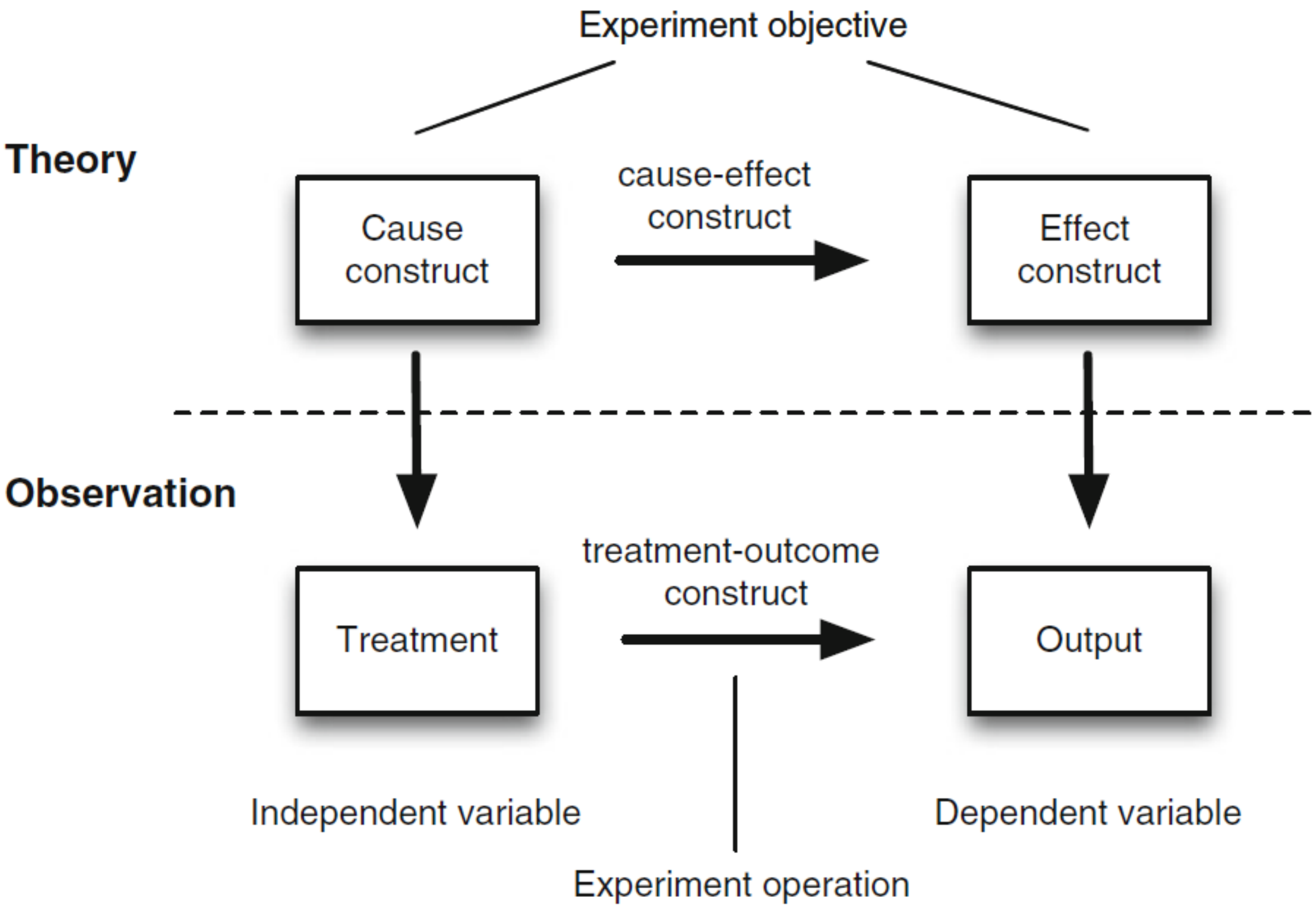
Experiment Principles – cont’d

- The starting point is that we have an idea of a **cause** and **effect** relationship

- In order to test hypothesis, we may use an **experiment**.

- In the design of the experiment, we have a number of treatments (values that the studied variable can take) over which we have control.

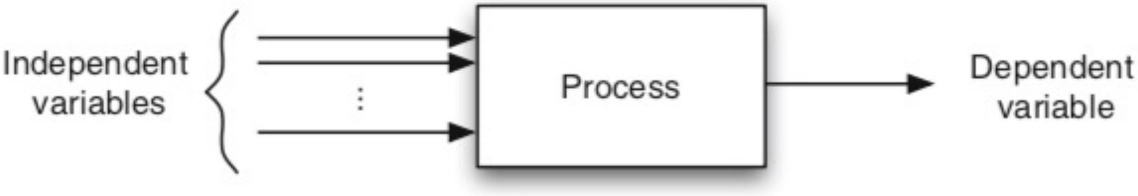
Experiment Principles – cont’d



Experiment Principles – cont’d

- When conducting a formal experiment, we want to study the outcome when we vary some of the input variables to a process.

- There are two kinds of variables in an experiment.
  - All variables in a process that are manipulated and controlled are called ***independent*** *variables.*
  - Those variables that we want to study to see the effect of the changes in the independent variables are called ***dependent*** *variables*



Experiment Principles – cont’d

- An experiment studies the effect of changing one or more independent variables.
  - Those variables are called **factors**.
  - A **treatment** is one particular value of a factor.
  - The other independent variables are controlled at a fixed level during the experiment, or else we cannot say if the factor or another variable causes the effect.

Experiment Principles – cont’d

- The experiment is performed and we are able to test the relationship between the **treatment** and the **outcome**.
  - If the experiment is properly set up, we should be able to draw conclusions about the relationship between the cause and the effect for which we stated a hypothesis.

Example

- The effect of aerobic training on Percentage of Body Fat

  - The dependent variable
    - percentage of body fat
  - The independent variable
    - aerobic training intensity (controlled)

- In the above example identify
  - Factor: ?
  - Treatment: ?

Hypothesis Formulation

Example

- We want to study the effect of a new development method on the productivity of the personnel. We may have chosen to introduce an object-oriented design method instead of a function-oriented approach.
  - The **dependent** variable in the experiment is the productivity.
  - **Independent** variables may be the development method, the experience of the personnel, tool support, and the environment.

Hypothesis formulation 1/3

- The basis for the **statistical analysis** of an experiment is **hypothesis testing**.
- A hypothesis is stated formally and the data collected during the course of the experiment is used to, if possible, reject the hypothesis.
- If the hypothesis can be **rejected** then conclusions can be drawn, based on the hypothesis testing under given risks.
- In the planning phase, the experiment definition is formalized into hypotheses

Hypothesis formulation 2/3

- A **null hypothesis**, H_o
  - It states that there are no real underlying trends or patterns in the experiment setting; the only reasons for differences in our observations are coincidental (by chance).
  - This is the hypothesis that the experimenter wants to reject with as high significance as possible.
  - An example hypothesis is that a new object-oriented design method offers similar productivity as the function-oriented method.

$$H_0 : \mu_{N\,old} = \mu_{N\,new}$$

  - Where μ denotes the **average** and N is the productivity.

Hypothesis formulation 3/3

- An **alternative hypothesis**, H_a; H_1, etc.
  - It is the hypothesis in favor of which the null hypothesis is rejected.
  - An example hypothesis is that an object-oriented design on average more productive than function-oriented method.

$$H_1 : \mu_{N\,old} < \mu_{N\,new}$$

  - Where μ denotes the **average** and N is the productivity

Question [Hypothesis formulation]

- To examine the relationship between texting and driving skill in students, a researcher uses orange traffic cones to set up a driving circuit in a parking lot. A group of students is then tested on the circuit, once while receiving and sending text messages, and once without texting. For each student, the researcher records the number of cones hit while driving each circuit.
  - What are experiment **variables**? What is the appropriate scale measurement for each variable?
  - Write a **hypothesis** for this problem?

Hypothesis testing

SciPy

- **SciPy** is a python library used for scientific computing and technical computing.
- SciPy contains modules for optimization, linear algebra, integration, signal and image processing and other tasks common in science and engineering.
- SciPy library is one of the core packages that make up the SciPy stack.
- built on NumPy

Link: https://www.scipy.org/scipylib/

Hypothesis Testing

- The objective of hypothesis testing is to see if it is possible to **reject** a certain **null hypothesis**, H_0, based on a sample from some statistical distribution.
  - The null hypothesis describes some properties of the distribution from which the sample is drawn, and the experimenter wants to reject that these properties are true with a given significance.

Type of Tests

- **Parametric tests**
  - Based on a model that involves a specific distribution.
  - In most cases, it is assumed that some of the parameters, involved in a parametric test, are normally distributed.
  - The data should represent an interval or ratio scale of measurement
  - Requires a fewer sample size

- **Non-Parametric tests**
  - Makes no assumption about the distribution of the variable in the population, that is, the shape of the distribution
  - More general than parametric tests
  - Can be used instead of parametric tests, but parametric tests cannot generally be used when non-parametric tests can be used.
  - Used when the data represent a nominal or ordinal scale
  - Usually requires a larger sample size

Test selection

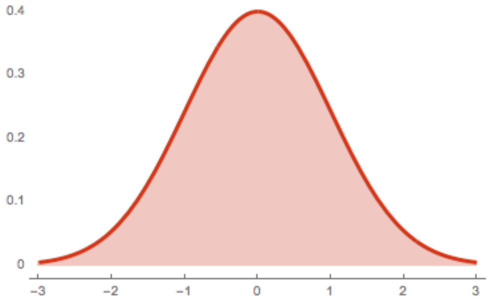
- Applicability
  - What are the assumptions made by the different tests? It is important that assumptions regarding distributions of parameters and assumptions concerning scales are realistic.

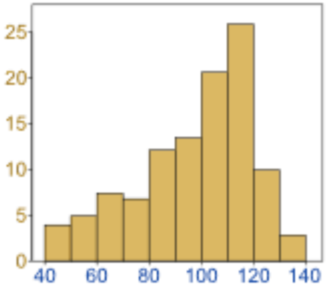
- Power
  - The power of parametric tests is generally higher than for nonparametric tests. Therefore, parametric tests require fewer data points, and therefore smaller experiments, than non-parametric test if the assumptions are true.

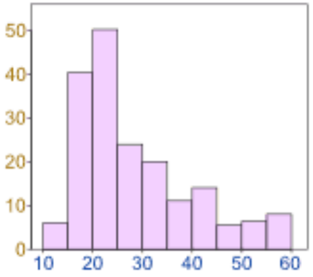
- Even if it is a risk using parametric methods when the required conditions are not fulfilled, it is in some cases worth taking that risk.

Normal Distribution

- The **Normal Distribution** has:
  - Same value of mean, median, and mode
  - Symmetric
  - 50% of values less than the mean and 50% greater than the mean





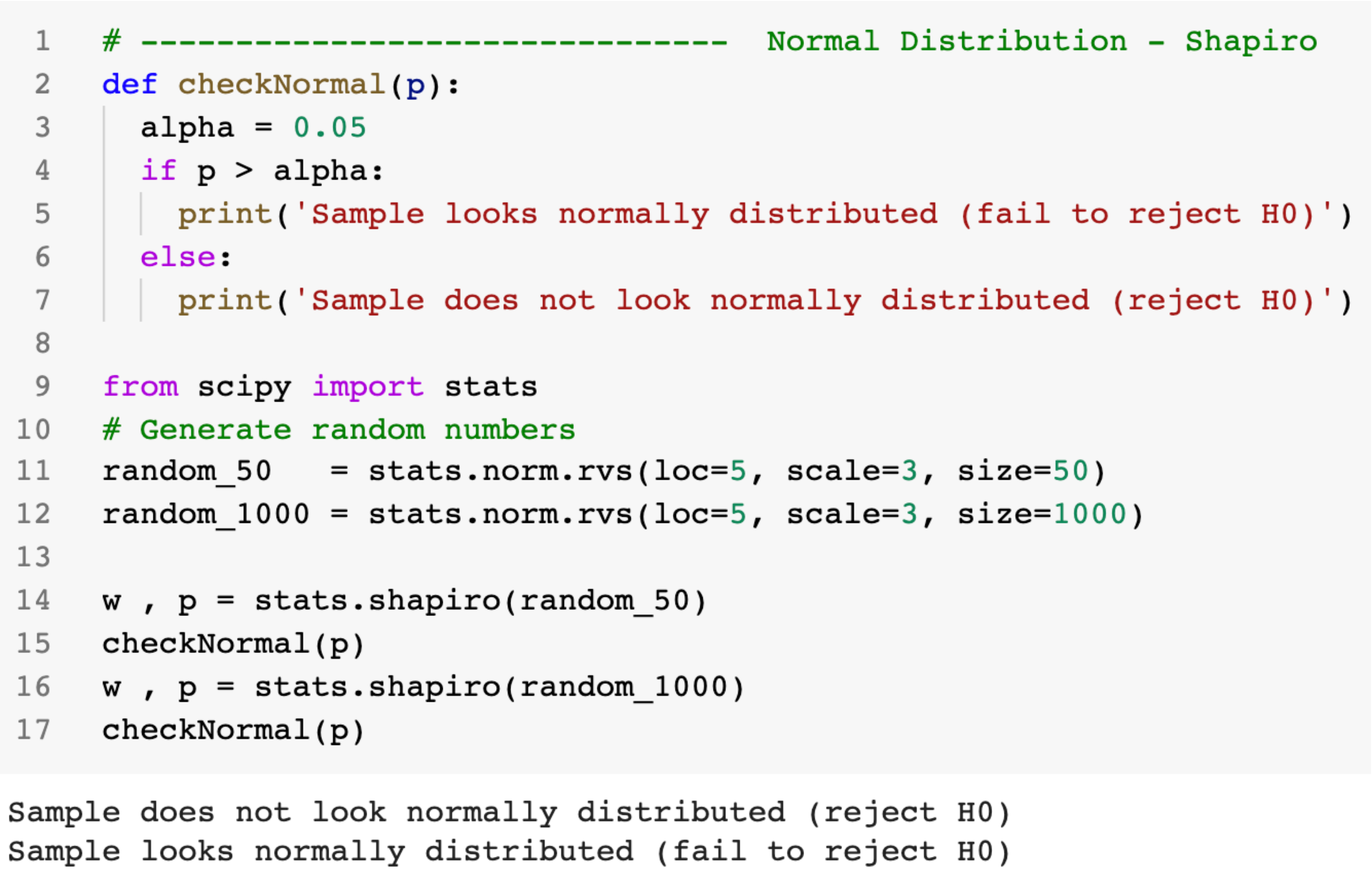


https://www.mathsisfun.com/data/standard-normal-distribution.html

Normal Distribution – cont’d

- Statistical tests
  - Shapiro–Wilk test
    - If the p value returned is less than (.05), then the null hypothesis is rejected and there is evidence that the data is not from a normally distributed population.
  - Kolmogorov-Smirnov (KS) test
  - Anderson test

Normal Distribution – cont’d



```python
# -------------------------------  Normal Distribution - Shapiro
def checkNormal(p):
    alpha = 0.05
    if p > alpha:
        print('Sample looks normally distributed (fail to reject H0)')
    else:
        print('Sample does not look normally distributed (reject H0)')

from scipy import stats
# Generate random numbers
random_50   = stats.norm.rvs(loc=5, scale=3, size=50)
random_1000 = stats.norm.rvs(loc=5, scale=3, size=1000)

w , p = stats.shapiro(random_50)
checkNormal(p)
w , p = stats.shapiro(random_1000)
checkNormal(p)
```

```
Sample does not look normally distributed (reject H0)
Sample looks normally distributed (fail to reject H0)
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html?highlight=rvs

Central Limit Theorem

- Regardless of the population distribution model, as the sample size increases, the sample mean tends to be normally distributed around the population mean, and its standard deviation shrinks as n increases.
  - The **Central Limit Theorem** is what the shape of the distribution of means will be when we draw repeated samples from a given population.

- Conditions:
  - The samples must be independent
  - The sample size must be large

Statistical Significance

- Researchers use a **test of significance** to determine whether to reject or fail to reject the null hypothesis
- **Test statistical significance** could be determined using P value or confidence intervals
- Confidence intervals involve pre-selecting a level of probability, “α” (e.g., α = .05) that serves as the criterion to determine whether to reject or fail to reject the null hypothesis
  - To say that a result has statistical significance at the 0.05 level means that a difference as large as or larger than what we observed between the groups could have occurred by chance only 5 times or less out of 100.
  - In other words, we are 95% confident that observed difference between the groups of samples **was not** occurred by chance.

Statistical Significance

- The confidence level is equivalent to 1 – the α level. So, if your significance level is 0.05, the corresponding confidence level is 95%.
  - If the P value is less than your significance (α) level, the hypothesis test is statistically significant.
  - If the confidence interval does not contain the null hypothesis value, the results are statistically significant.
  - If the P value is less than α, the confidence interval will not contain the null hypothesis value.

- For example, the P value (0.031) is less than the significance level (0.05), which indicates that our results are statistically significant.

P value

- P values evaluates how well the sample data support the null hypothesis is true.
  - Defines the distance the sample mean must be from the null hypothesis to be considered statistically significant.
  - It measures how compatible your data are with the null hypothesis.
  - How likely is the effect observed in your sample data if the null hypothesis is true?

- High P values: your data are likely with a true null.
- Low P values: your data are unlikely with a true null

- P value and significance level:
  - If the P value is less than your significance (α) level, the hypothesis test is statistically significant.
    - Reject H_0
  - If the P value is more than α, the hypothesis test is not statistically significant.
    - Accept H_0

Types of errors

Hypothesis testing involves two main types of risks:

- Type-1-error (false positive)
  - the risk of rejecting a true null hypothesis

$$P\,(\text{type-I-error}) = P\,(\text{reject } H_0 \mid H_0 \text{ true})$$

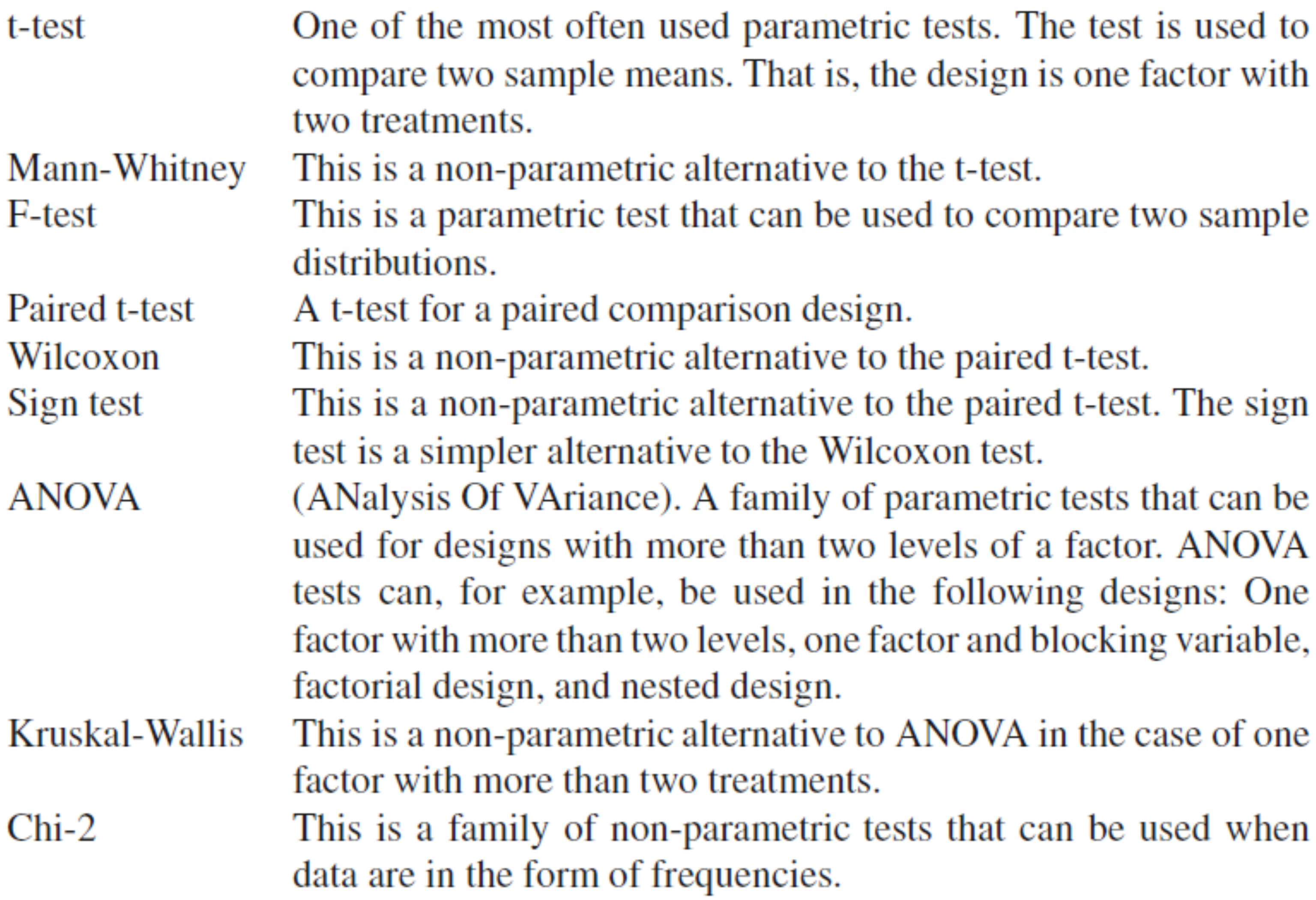
- Type-2-error (false negative)
  - the risk of not rejecting a false null hypothesis

$$P\,(\text{type-II-error}) = P\,(\text{not reject } H_0 \mid H_0 \text{ false})$$

Types of errors

	Null hypothesis is TRUE	**Null hypothesis is FALSE**
Reject null hypothesis	Type I Error (False positive)	Correct outcome ! (True positive)
Fail to reject null hypothesis	Correct outcome ! (True negative)	Type II Error (False negative)

Statistical Tests



t-test	One of the most often used parametric tests. The test is used to compare two sample means. That is, the design is one factor with two treatments.
Mann-Whitney	This is a non-parametric alternative to the t-test.
F-test	This is a parametric test that can be used to compare two sample distributions.
Paired t-test	A t-test for a paired comparison design.
Wilcoxon	This is a non-parametric alternative to the paired t-test.
Sign test	This is a non-parametric alternative to the paired t-test. The sign test is a simpler alternative to the Wilcoxon test.
ANOVA	(ANalysis Of VAriance). A family of parametric tests that can be used for designs with more than two levels of a factor. ANOVA tests can, for example, be used in the following designs: One factor with more than two levels, one factor and blocking variable, factorial design, and nested design.
Kruskal-Wallis	This is a non-parametric alternative to ANOVA in the case of one factor with more than two treatments.
Chi-2	This is a family of non-parametric tests that can be used when data are in the form of frequencies.

Statistical Tests

Design and Statistical Tests

Design	Parametric	Non-parametric
One factor, one treatment		Chi-Square, Binomial test
One factor, two treatments, independent sample	t-test	Mann-Whitney, Chi-Square
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-Square
More than one factor	ANOVA	

Chi-Square test

- **Chi-Square test** is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

- All Chi-Square tests are, however, based on that data is in the form of frequencies from a single population.

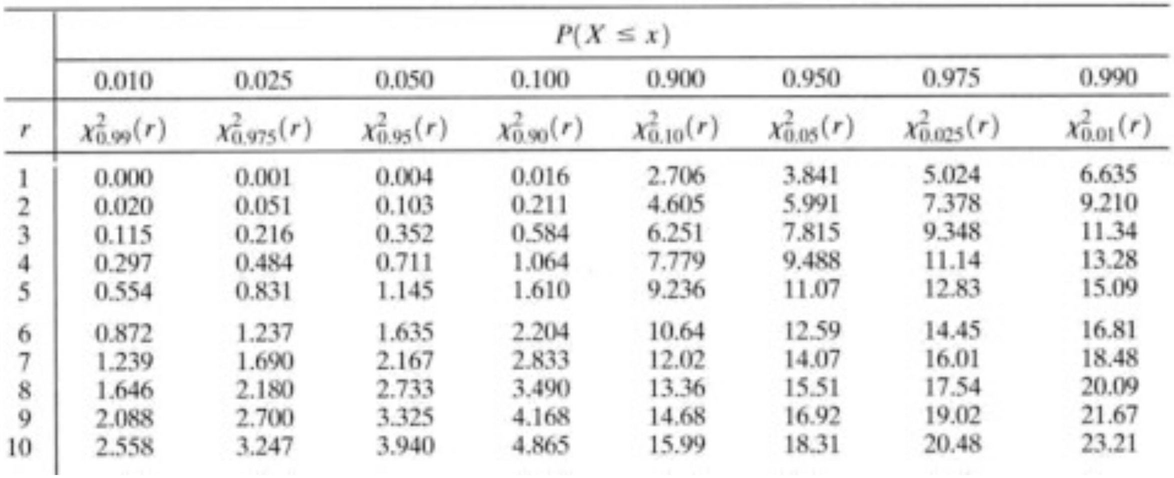
$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_1)^2}{E_i}$$

  - Observed frequency = data which you collect
  - Expected frequency = theoretical value

Chi-Square test

- Often use null hypothesis, i.e. there is no significant difference between the observed and expected frequency.
  - Significant value also called p value is a measure of the strength of the evidence against the null hypothesis
  - If p value < 0.05 the null hypothesis will be rejected

- Degrees of freedom = n – 1
  - **n** is the number of categories (variables)
  - Chi square distribution table



	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

Chi-Square test example

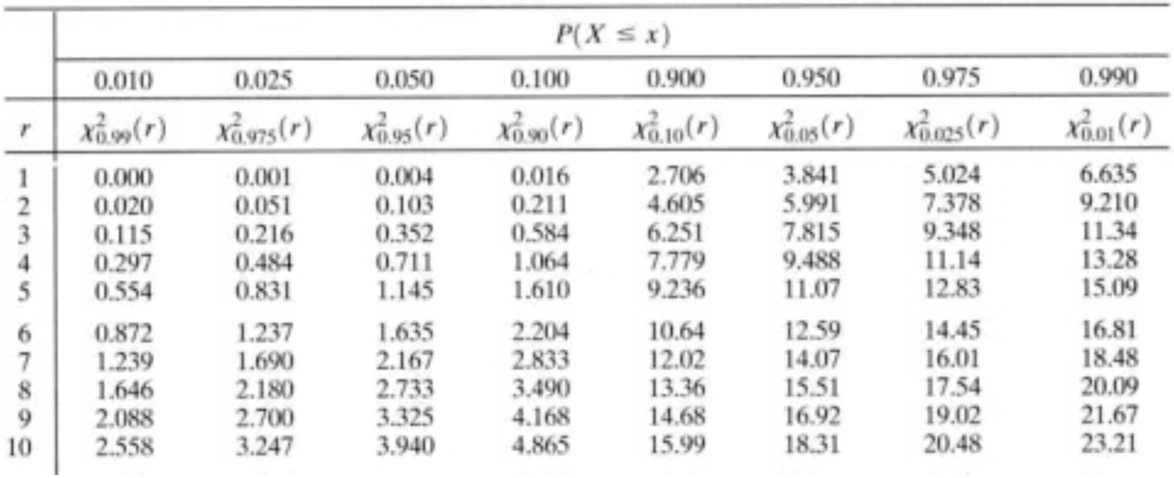
- Consider coin with two options, i.e. head and tail.

- Think of hypothesis?

- Lets flip 50 coins
  - Expected frequency = 25 Head and 25 Tails
  - Observed frequency = 28 Head and 22 Tails
  - Head = $(28-25)^2$ / 25 = 9/25
  - Tail = $(22-25)^2$ / 25 = 9/25
  - X^2 = 9/25 + 9/25 = 18/25 = 0.72

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_1)^2}{E_i}$$

Chi-Square test example

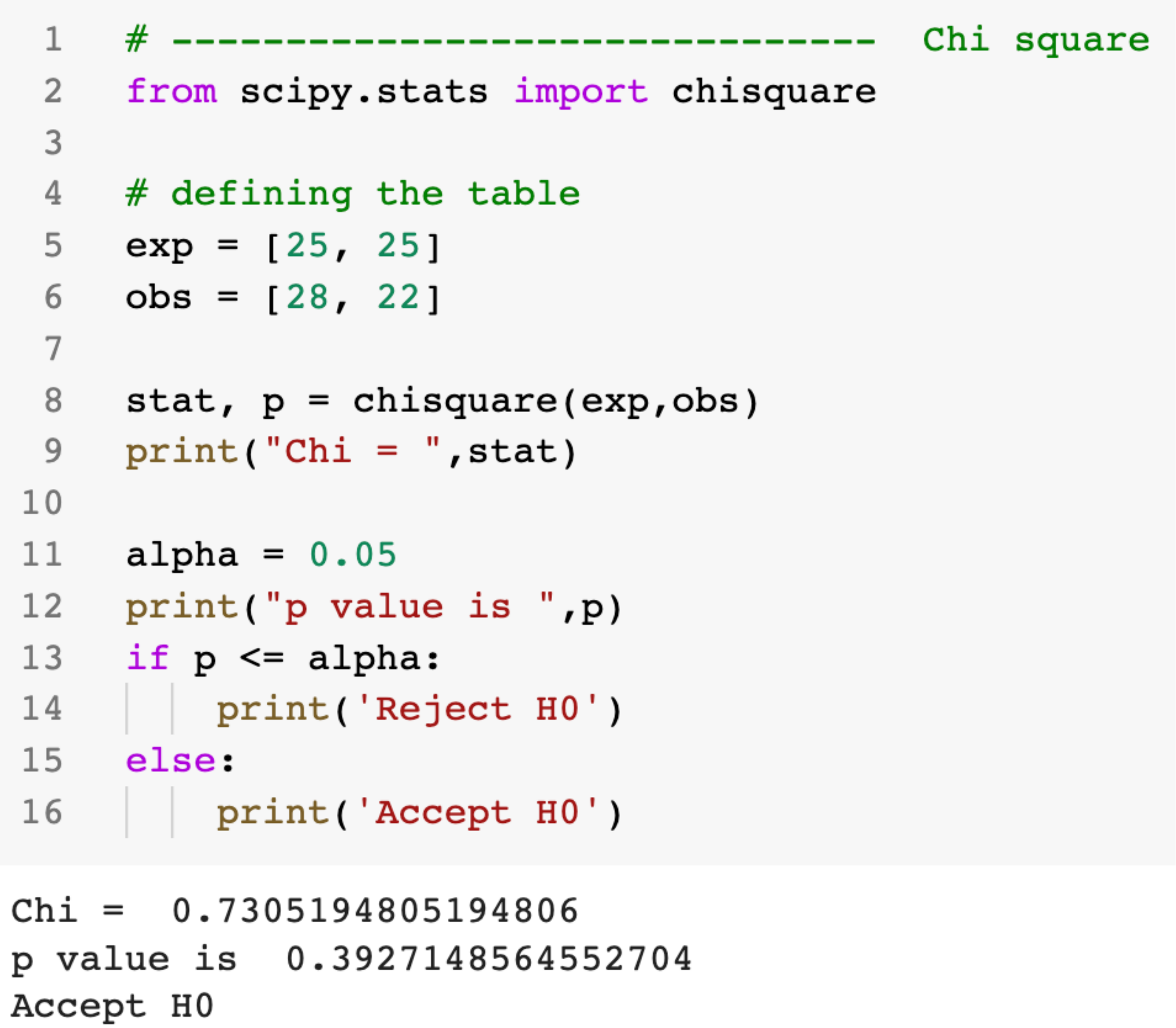


	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

- Degrees of freedom = n-1
- $X^2 = 0.72$
- Critical value = 3.841
- If X^2 > Critical value, Then Reject H_0
- Else if X^2 < Critical value, Then **Accept H_0**

No statistical difference between what we observe and what we expect to see.

Example - Python code



```python
# ------------------------------  Chi square
from scipy.stats import chisquare

# defining the table
exp = [25, 25]
obs = [28, 22]

stat, p = chisquare(exp,obs)
print("Chi = ",stat)

alpha = 0.05
print("p value is ",p)
if p <= alpha:
    print('Reject H0')
else:
    print('Accept H0')
```

```
Chi =  0.7305194805194806
p value is  0.3927148564552704
Accept H0
```

Design and Statistical Tests

Design	Parametric	Non-parametric
One factor, one treatment		Chi-Square, Binomial test
One factor, two treatments, independent sample	t-test	Mann-Whitney, Chi-Square
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-Square
More than one factor	ANOVA	

Mann–Whitney *U* test

- It is a **non-parametric** statistical hypothesis test to compare differences between two independent groups when the dependent variable is either ordinal or continuous **but not normally distributed**

- The *U* test is useful in the same situations as the **independent samples *t*-test**

Mann–Whitney *U* test

- *U* remains the logical choice when the data are ordinal and/or distribution is not normal

- We are interested in the median rather than the mean.

- The hypothesis test of interest is
  - H_0: $\text{median}_a = \text{median}_b$
  - H_A: $\text{median}_a \mathrel{!=} \text{median}_b$

Example 1 - Mann–Whitney *U* test

- In our company, there are managers with a **Bsc** and **Msc**, and we would like to see if there is a **significant difference** between their salaries.

Bsc	Msc
30221	44330
39907	52404
40324	41034
42198	

Example 1 - Mann–Whitney *U* test

- Rank values / Sum of ranks

Salary	Rank	Degree
30221	1	Bsc
39907	2	Bsc
40324	3	Bsc
41034	4	Msc
42198	5	Bsc
44330	6	Msc
52404	7	Msc

Degree	Mean Rank	Sum of Ranks
Bsc	2.75	11
Msc	5.67	17

Example 1 - Mann–Whitney *U* test

- Identify group with smaller sum of ranks
  - Bsc

- Calculate U_{stat}

Salary	Rank	Degree	U
30221	1	Bsc	0
39907	2	Bsc	0
40324	3	Bsc	0
41034	4	Msc	
42198	5	Bsc	1
44330	6	Msc	
52404	7	Msc	

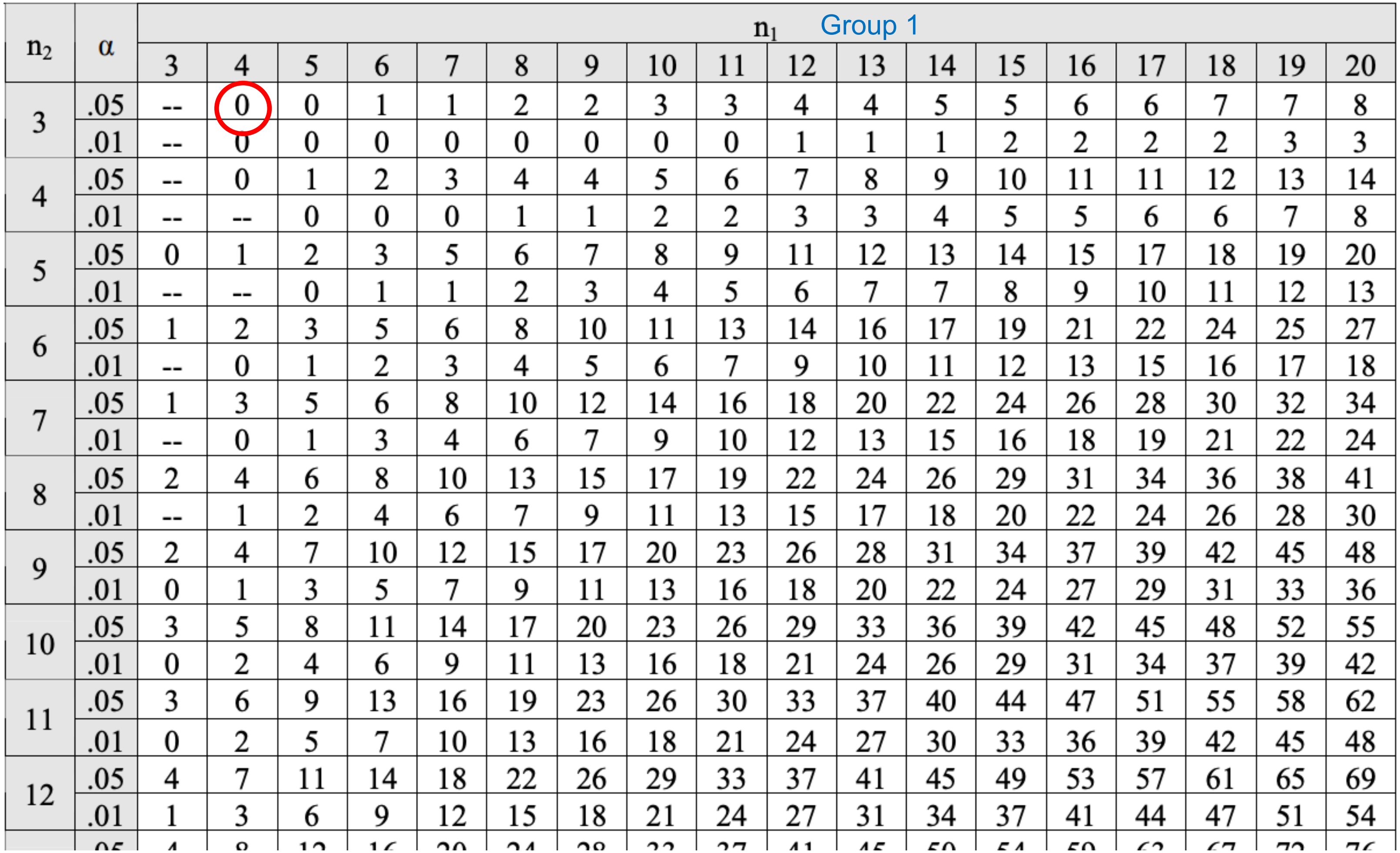
- Then, $U_{stat} = 1$

Distribution table

- Compare (U_{stat}) to U_{crit} from distribution table

Group 2

Group 1



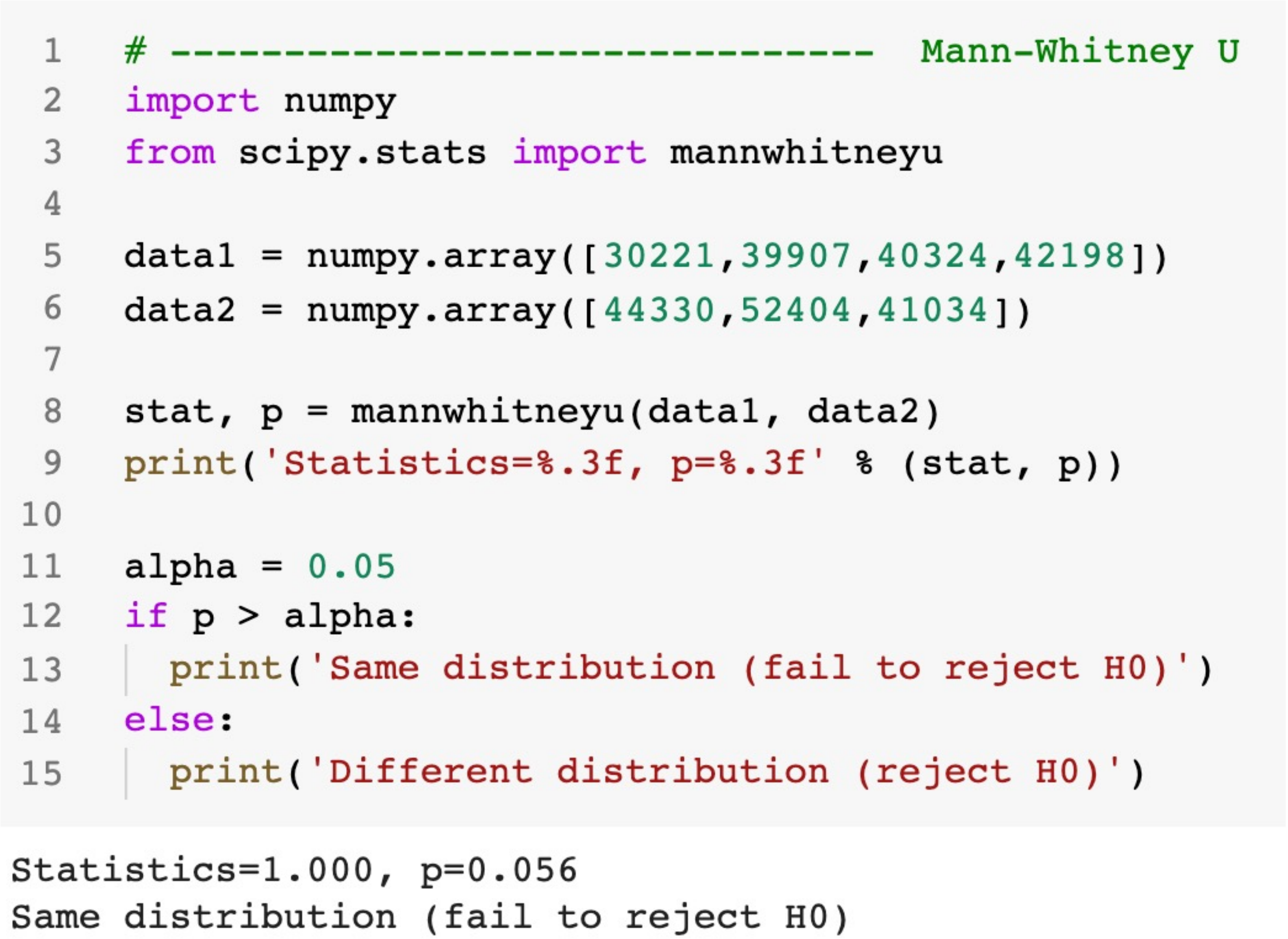
n2	α	n1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76

Example 1 - Mann–Whitney *U* test

- If our calculated $\mathbf{U_{stat} < U_{crit}}$ from distribution table
  - Then, reject null hypothesis

- In our example: $U_{stat = 1}$ and $U_{crit = 0}$
  - Then, we cannot reject the null hypothesis

Example 1 - Python code



```python
# ------------------------------  Mann-Whitney U
import numpy
from scipy.stats import mannwhitneyu

data1 = numpy.array([30221,39907,40324,42198])
data2 = numpy.array([44330,52404,41034])

stat, p = mannwhitneyu(data1, data2)
print('Statistics=%.3f, p=%.3f' % (stat, p))

alpha = 0.05
if p > alpha:
  print('Same distribution (fail to reject H0)')
else:
  print('Different distribution (reject H0)')
```

```
Statistics=1.000, p=0.056
Same distribution (fail to reject H0)
```

Example 2 - Mann–Whitney *U* test

- The surgeons are interested in the economics of the **two types of surgery**.

- One of the costs of interest is the **anesthesia cost**. The cost (in dollars) for several of the patients in each of the two groups is given here.

- Hypotheses
  - Null: $\text{median}_{\text{surgery1}} = \text{median}_{\text{surgery2}}$
  - Alternative: $\text{median}_{\text{surgery1}} \neq \text{median}_{\text{surgery2}}$

Surgery 1	**Surgery 2**
1011.07	496.44
1066.82	541.76
610.80	1562.01
1111.44	2515.12
955.68	1133.99
1203.84	300.33
1600.32	482.55
555.90	503.22
1302.95	2744.23
182.34	1232.22
1233.20	
1402.09	

Example 2 - Mann–Whitney *U* test

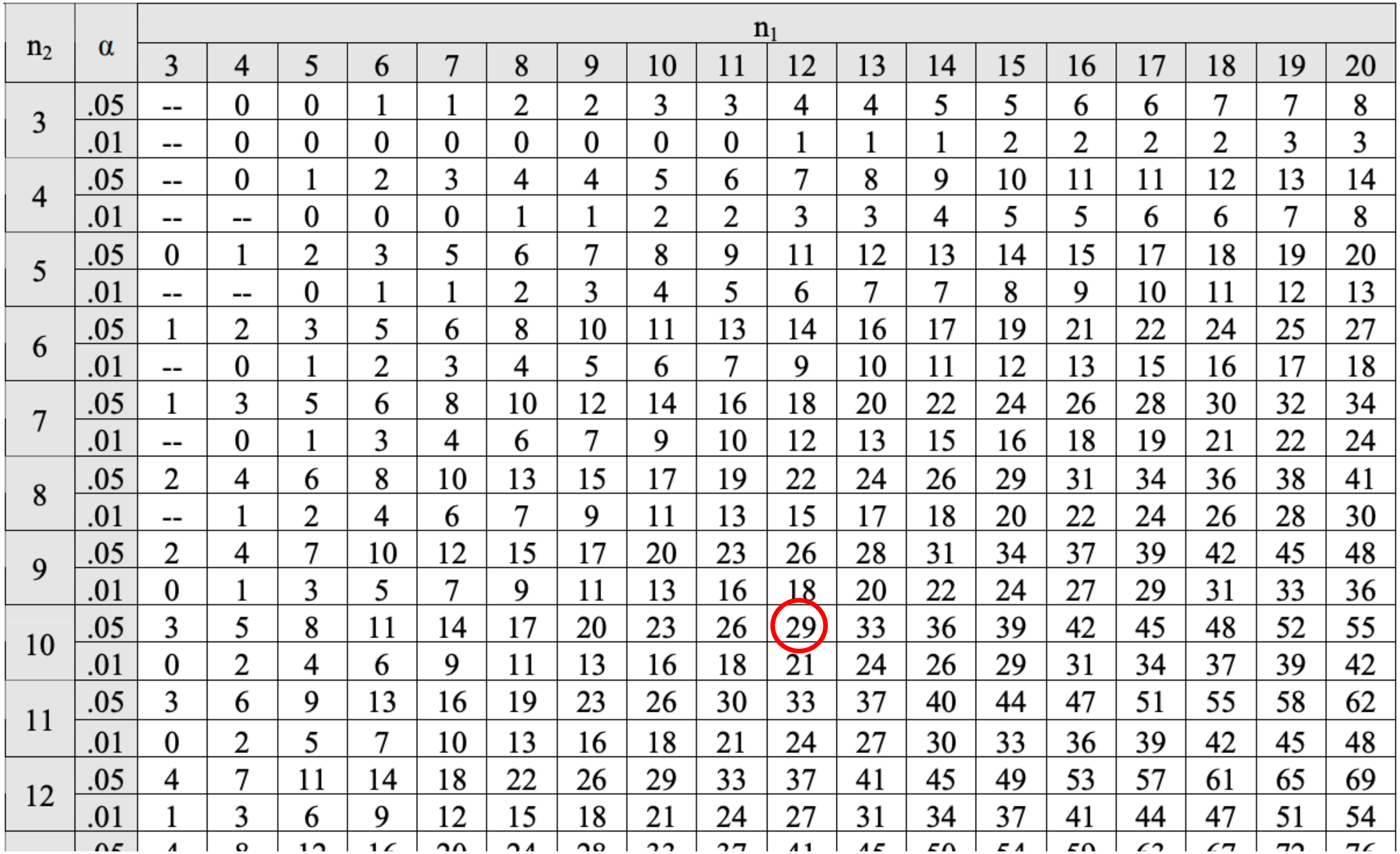
Surgery 1	Rank	Surgery 2	Rank
182.34	1	300.33	2
555.9	7	482.55	3
610.8	8	496.44	4
955.68	9	503.22	5
1011.07	10	541.76	6
1066.82	11	1133.99	13
1111.44	12	1232.22	15
1203.84	14	1562.01	19
1233.2	16	2515.12	21
1302.95	17	2744.23	22
1402.09	18		
1600.32	20		
sum of ranks	143		110
Average	11.92		11

Surgery 1	
Surgery 2	1
Surgery 2	1
Surgery 2	1
Surgery 2	1
Surgery 2	1
Surgery 1	
Surgery 1	
Surgery 1	
Surgery 1	
Surgery 1	
Surgery 1	
Surgery 2	7
Surgery 1	
Surgery 2	8
Surgery 1	
Surgery 1	
Surgery 1	
Surgery 2	11
Surgery 1	
Surgery 2	12
Surgery 2	12
U	**55**

- $U_{stat} = 55$

Or…. U_{stat} = RankSum – (n*(n+1))/2

Mann–Whitney *U* distribution table

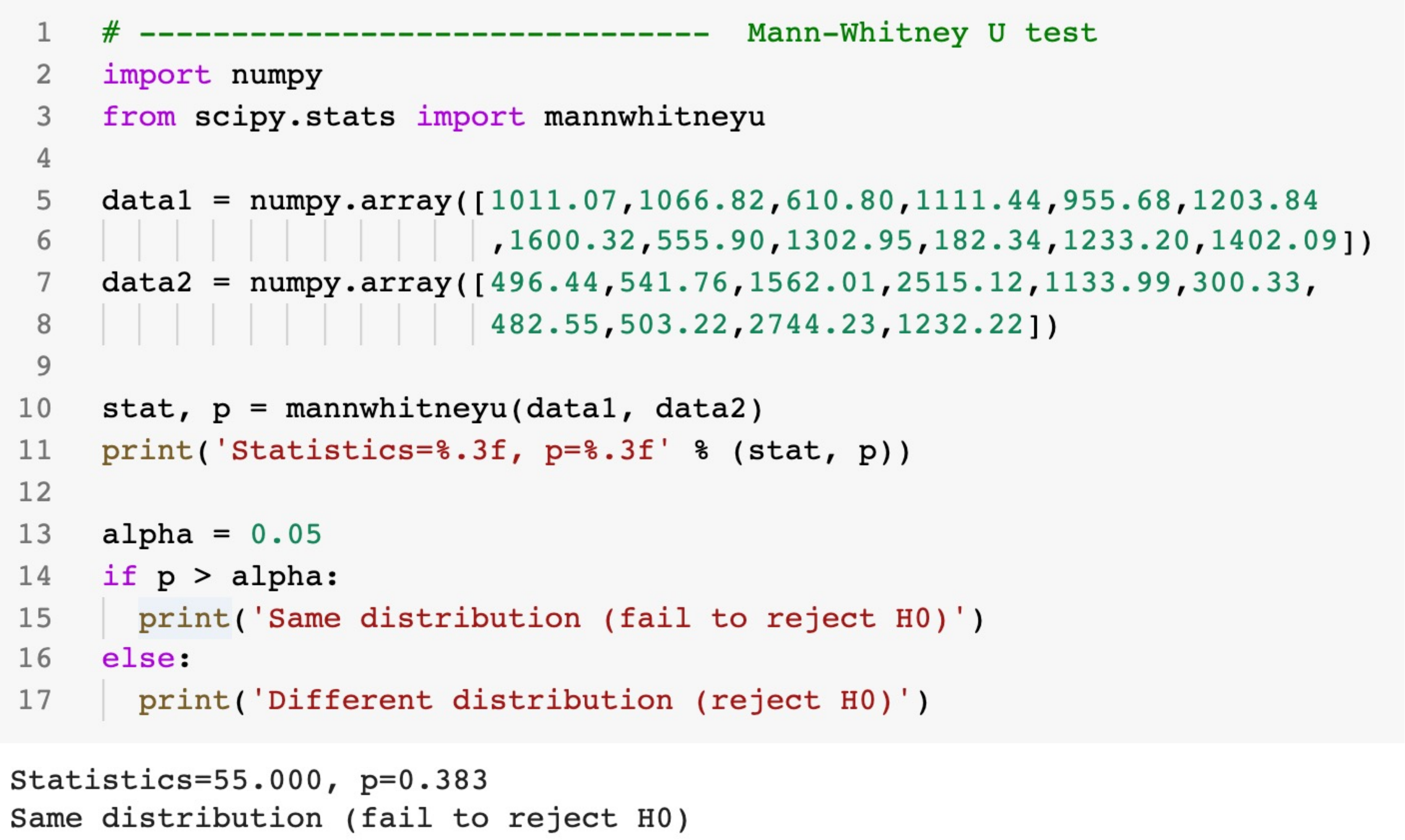


n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76

Example 2 - Mann–Whitney *U* test

- U_{crit}= 29, alpha=0.05
- U_{stat} = 55
- P-value = 0.77
- $U_{stat} > U_{crit}$ from distribution table
  - Fail to reject null hypothesis
- Conclusion: There is no evidence of a difference between the cost in the surgery 1 patients and the surgery 2 patients.

Example 2 - Python code



```python
# -------------------------------  Mann-Whitney U test
import numpy
from scipy.stats import mannwhitneyu

data1 = numpy.array([1011.07,1066.82,610.80,1111.44,955.68,1203.84
                     ,1600.32,555.90,1302.95,182.34,1233.20,1402.09])
data2 = numpy.array([496.44,541.76,1562.01,2515.12,1133.99,300.33,
                     482.55,503.22,2744.23,1232.22])

stat, p = mannwhitneyu(data1, data2)
print('Statistics=%.3f, p=%.3f' % (stat, p))

alpha = 0.05
if p > alpha:
    print('Same distribution (fail to reject H0)')
else:
    print('Different distribution (reject H0)')
```

```
Statistics=55.000, p=0.383
Same distribution (fail to reject H0)
```

Design and Statistical Tests

Design	Parametric	Non-parametric
One factor, one treatment		Chi-Square, Binomial test
One factor, two treatments, independent sample	t-test	Mann-Whitney, Chi-Square
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-Square
More than one factor	ANOVA	

Wilcoxon signed rank

- The Wilcoxon signed-rank test is a **non-parametric** statistical hypothesis test used when comparing two **related/paired samples** to assess whether their population mean **ranks** differ.

- It can be used as an alternative to the paired t-test when
  - the population cannot be assumed to be normally distributed and/or
  - the data is on the ordinal scale

- The null and alternative hypotheses are
  - H_0: median diff = 0
  - H_A: median diff ≠ 0

Example - Wilcoxon signed rank

- Is there a **significant difference** in the pre and post surgery platelet count?
- We have **paired observations** on each of the patients
- We are interested in the difference between the two measurements

Patient	Pre	Post
1	492	375
2	297	382
3	272	325
4	367	585
5	206	181
6	284	237
7	338	273
8	212	243
9	161	147
10	384	326
11	224	214
12	251	292
13	224	263

Example - Wilcoxon signed rank

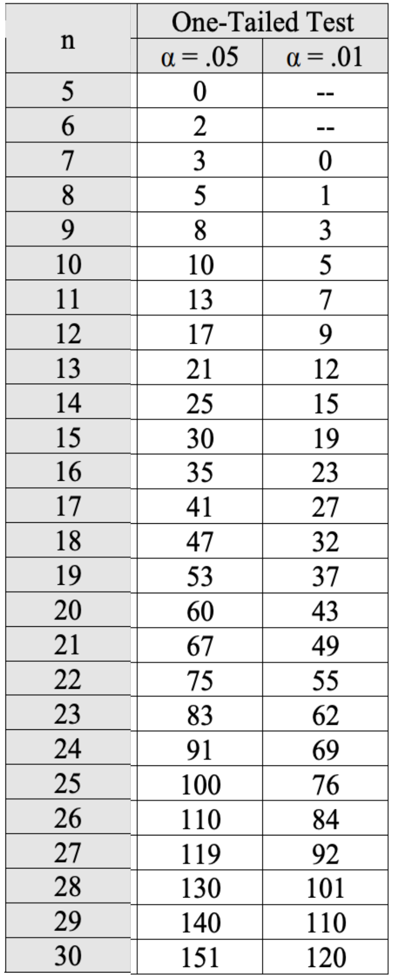
Patient	Pre	Post	Diff
1	492	375	117
2	297	382	-85
3	272	325	-53
4	367	585	-218
5	206	181	25
6	284	237	47
7	338	273	65
8	212	243	-31
9	161	147	14
10	384	326	58
11	224	214	10
12	251	292	-41
13	224	263	-39

Diff	Rank	sum of + ranks	sum of - ranks
117	12	12	
85	11		11
53	8		8
218	13		13
25	3	3	
47	7	7	
65	10	10	
31	4		4
14	2	2	
58	9	9	
10	1	1	
41	6		6
39	5		5
		44	47

- W_{stat}= 44

Wilcoxon critical values table

- Sample size (n) = 13
- W_{crit} = 21



n	One-Tailed Test	
	α = .05	α = .01
5	0	--
6	2	--
7	3	0
8	5	1
9	8	3
10	10	5
11	13	7
12	17	9
13	21	12
14	25	15
15	30	19
16	35	23
17	41	27
18	47	32
19	53	37
20	60	43
21	67	49
22	75	55
23	83	62
24	91	69
25	100	76
26	110	84
27	119	92
28	130	101
29	140	110
30	151	120

Example - Wilcoxon signed rank

- Paired data, wilcoxon test, alpha=0.05

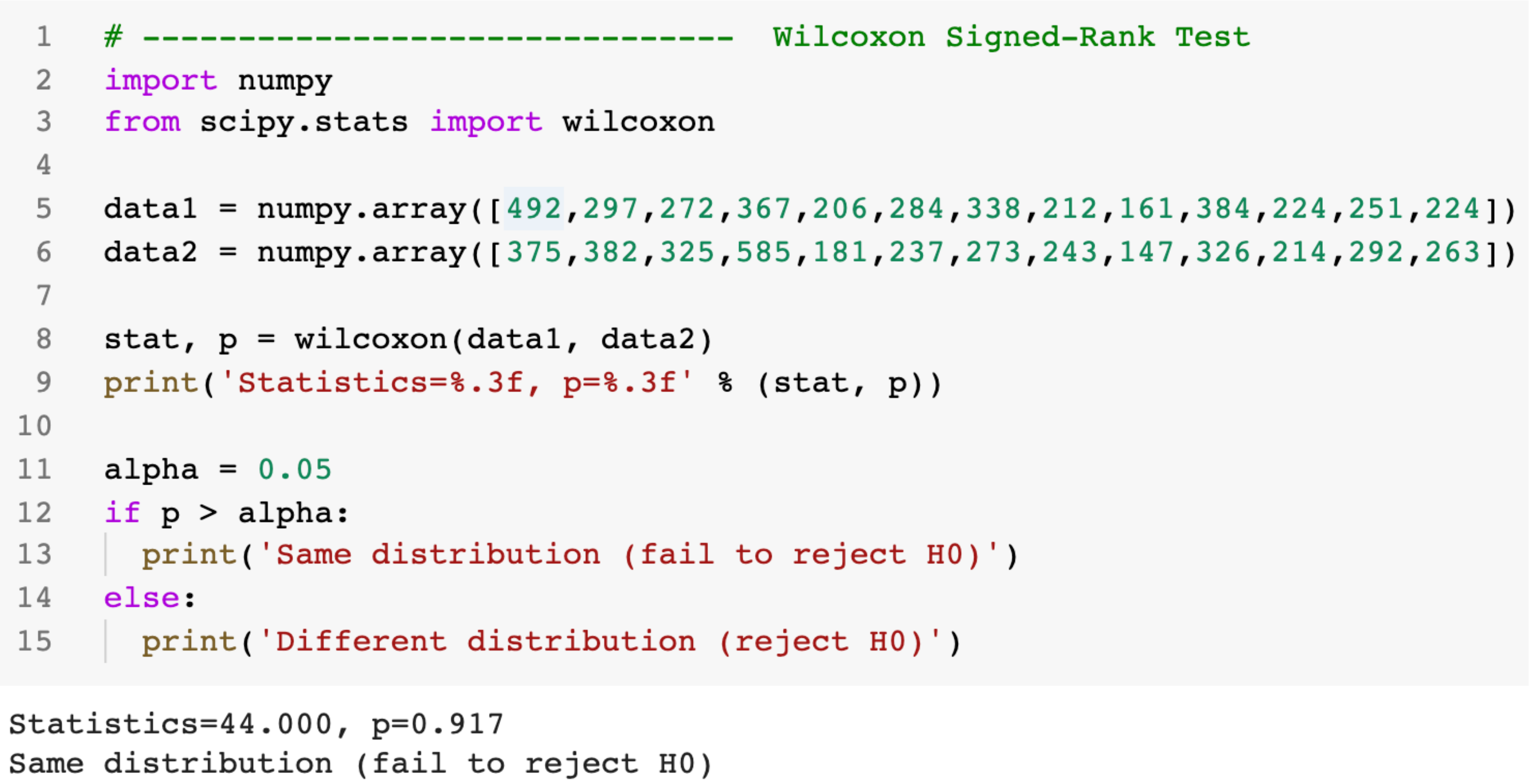
- Hypotheses
  - Null: median difference = 0
  - Alternative: median difference ≠ 0

- p-value = 0.917,

- Fail to reject null hypothesis
  - Calculated $W_{stat} > W_{crit}$
  - p-value > alpha

- Conclusion: There is **no evidence of a difference** between the pre and post platelet counts for patients.

Example - Python code



```python
# -------------------------------  Wilcoxon Signed-Rank Test
import numpy
from scipy.stats import wilcoxon

data1 = numpy.array([492,297,272,367,206,284,338,212,161,384,224,251,224])
data2 = numpy.array([375,382,325,585,181,237,273,243,147,326,214,292,263])

stat, p = wilcoxon(data1, data2)
print('Statistics=%.3f, p=%.3f' % (stat, p))

alpha = 0.05
if p > alpha:
  print('Same distribution (fail to reject H0)')
else:
  print('Different distribution (reject H0)')
```

```
Statistics=44.000, p=0.917
Same distribution (fail to reject H0)
```

Design and Statistical Tests

Design	Parametric	Non-parametric
One factor, one treatment		Chi-Square, Binomial test
One factor, two treatments, independent sample	t-test	Mann-Whitney, Chi-Square
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-Square
More than one factor	ANOVA	

ANOVA Analysis

- **AN**alysis **O**f **VA**riance

- ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups.

- One-way ANOVA is used to test for differences among three or more independent groups (means)

What does ANOVA do?

- At its simplest ANOVA tests the following hypotheses:
  - H_0: The means of all the groups are equal.
  - H_a: Not all the means are equal

ANOVA

- Calculations steps
  - Compute correction for the mean (CM)

$$\text{CM} = \frac{\text{Total of all observations}^2}{n_{total}}$$

  - Compute Sum of Squares (Between groups) and (Within groups)

$$\text{Sum Square}_{\text{Between}} = \sum \frac{X^2}{n} - CM$$

$$\text{Sum Square}_{\text{Within}} = \sum X^2 - CM$$

  - Compute degrees of freedom (Between groups) and (Within groups)

$\text{Degrees of Freedom}_{\text{Between}}$: DF = k − 1

where k is the number of groups

$\text{Degrees of Freedom}_{\text{Within}}$ = N − k

where N is the total number of subjects

https://goodcalculators.com/one-way-anova-calculator/

ANOVA

- Calculations steps
  - Compute total Sum of Squares

$$\text{Sum of Square}_{\text{Total}} = \text{Sum of Square}_{\text{Between}} + \text{Sum of Square}_{\text{Within}}$$

  - Compute Mean Square Between Groups

$$\text{Mean Square}_{\text{Between}} = SS_B \,/\, (k - 1)$$

$$\text{Mean Square}_{\text{Within}} = SS_W \,/\, (N - k)$$

  - Calculate **F_{stat}**

$$F_{stat} = \text{Mean Square}_{\text{Between}} \,/\, \text{Mean Square}_{\text{Within}}$$

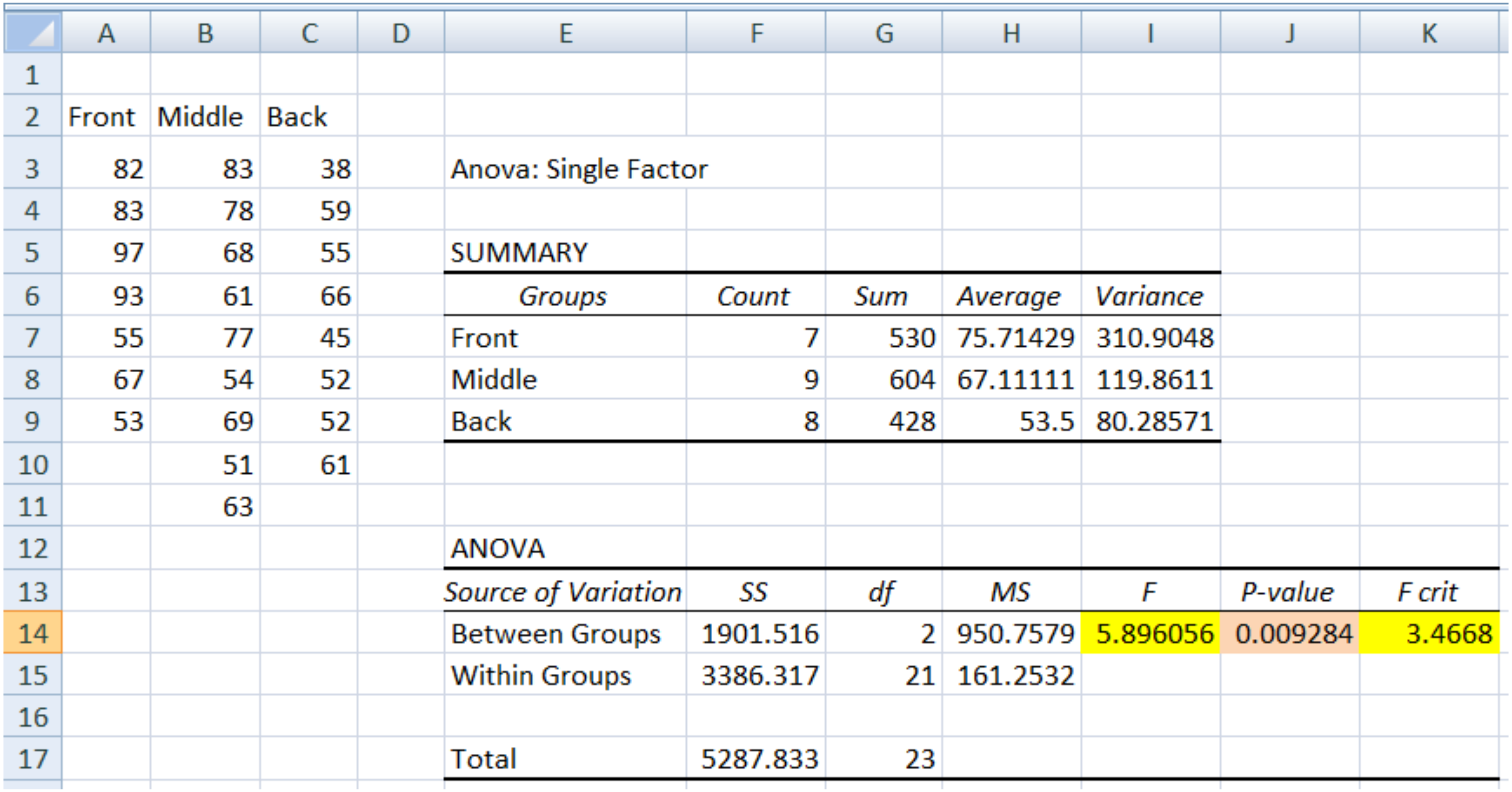
Example - ANOVA

- A random sample of the students in each row was taken.

- The score for those students on the second exam was recorded

Front:	82, 83, 97, 93, 55, 67, 53
Middle:	83, 78, 68, 61, 77, 54, 69, 51, 63
Back:	38, 59, 55, 66, 45, 52, 52, 61

Example - ANOVA

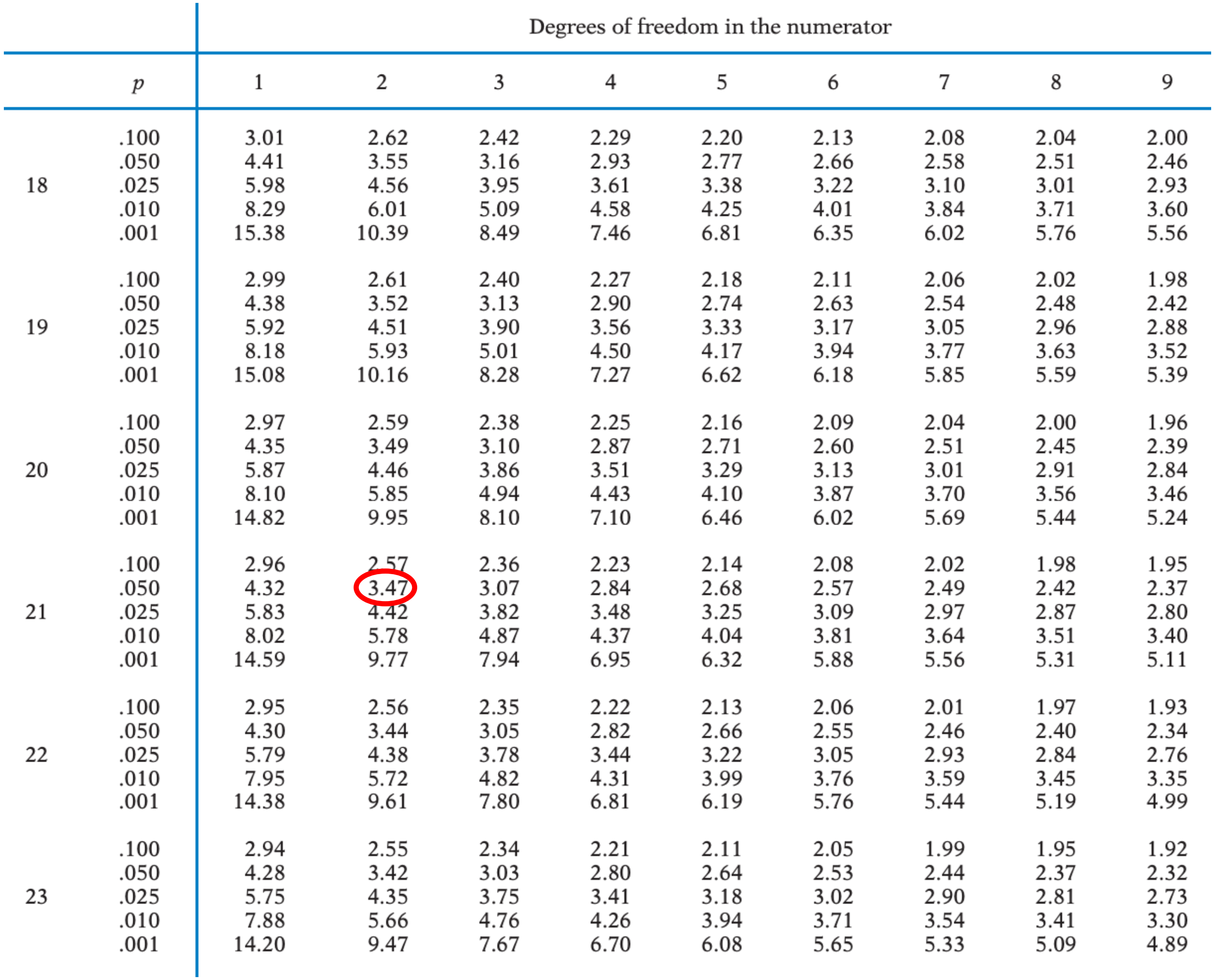


	A	B	C	D	E	F	G	H	I	J	K
1											
2	Front	Middle	Back								
3	82	83	38		Anova: Single Factor						
4	83	78	59								
5	97	68	55		SUMMARY						
6	93	61	66		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
7	55	77	45		Front	7	530	75.71429	310.9048		
8	67	54	52		Middle	9	604	67.11111	119.8611		
9	53	69	52		Back	8	428	53.5	80.28571		
10		51	61								
11		63									
12					ANOVA						
13					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
14					Between Groups	1901.516	2	950.7579	5.896056	0.009284	3.4668
15					Within Groups	3386.317	21	161.2532			
16											
17					Total	5287.833	23				

https://www.socscistatistics.com/pvalues/fdistribution.aspx

F Distribution Table

- $F_{crit} = 3.47$



		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89

Example - ANOVA

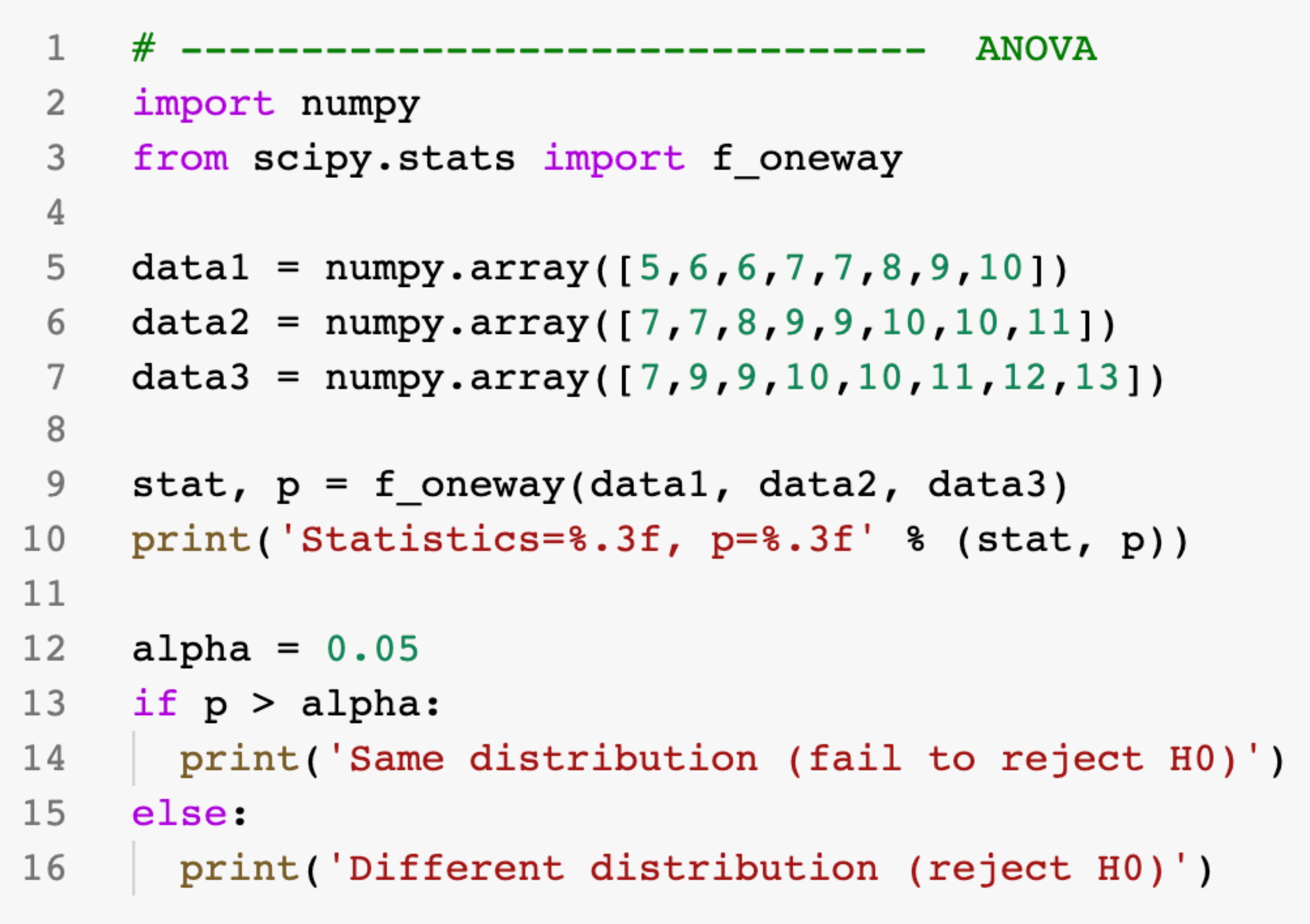
- Fail to reject null hypothesis
  - Calculated $F_{stat} < F_{crit}$
  - p-value > alpha

- The p-value is **0.009**, which is less than the significance level of 0.05, so we reject the null hypothesis.

- The null hypothesis is that the means of the three rows in class were the same, but we **reject that**, so at least one row has a different mean.

- There is enough evidence to support the claim that there is a difference in the mean scores of the front, middle, and back rows in class.

Example - Python Code

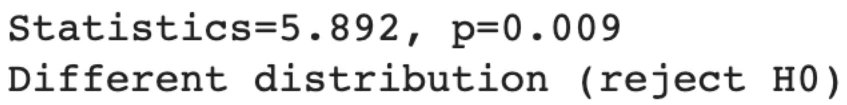


```python
# -------------------------------  ANOVA
import numpy
from scipy.stats import f_oneway

data1 = numpy.array([5,6,6,7,7,8,9,10])
data2 = numpy.array([7,7,8,9,9,10,10,11])
data3 = numpy.array([7,9,9,10,10,11,12,13])

stat, p = f_oneway(data1, data2, data3)
print('Statistics=%.3f, p=%.3f' % (stat, p))

alpha = 0.05
if p > alpha:
  print('Same distribution (fail to reject H0)')
else:
  print('Different distribution (reject H0)')
```



```
Statistics=5.892, p=0.009
Different distribution (reject H0)
```

Correlation

Correlation

- Determines whether and to what degree a relationship exists between two or more quantifiable variables

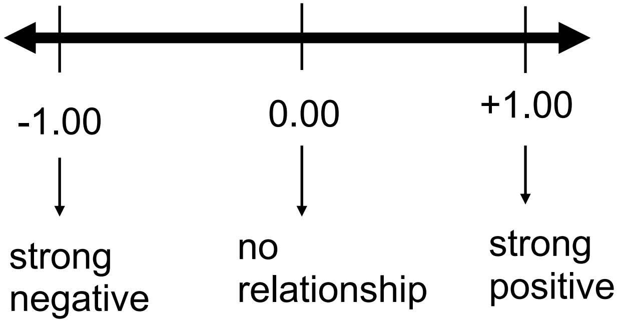
- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
  - For example, height and weight are related;

- The degree of the relationship is expressed as a **coefficient of correlation**

- Correlation describes the three types of relationship positive, negative and non-correlated. It also describe the magnitude of correlation from 0 to 1 and from -1 to 0.

Correlation

- Magnitude
- Direction

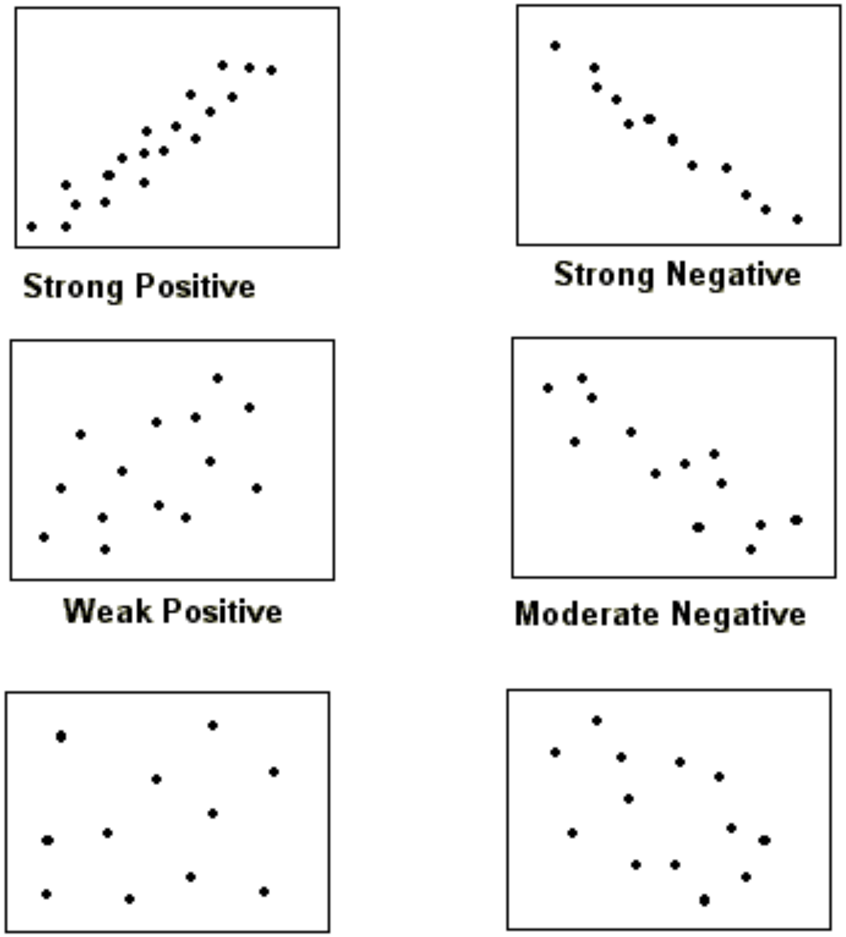


-1.00 → strong negative

0.00 → no relationship

+1.00 → strong positive

Degree of Correlation



Strong Positive

Strong Negative

Weak Positive

Moderate Negative

None

Weak Negative

Causation

- Correlation is of great importance in statistical analysis, as it helps explain the data and sometimes highlights predictive relationships between variables.

- Correlation does not establish a **cause-effect** relationship.

- A cause-effect relationship means that one variable is causing a change in another variable.

Pearson Correlation

- In statistics, the **Pearson correlation coefficient** or **Pearson's *r*** is a measure of the linear correlation (dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive

- Where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation.

- It is widely used in the sciences as a measure of the degree of linear dependence between two variables

Spearman Rank Correlation

- The Spearman's **rank-order** correlation is the **non-parametric** version of the Pearson product-moment correlation.

- Spearman's correlation coefficient, (p also signified by r_s) measures the strength of association between two **ranked** variables.

- Nonparametric statistics uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking.

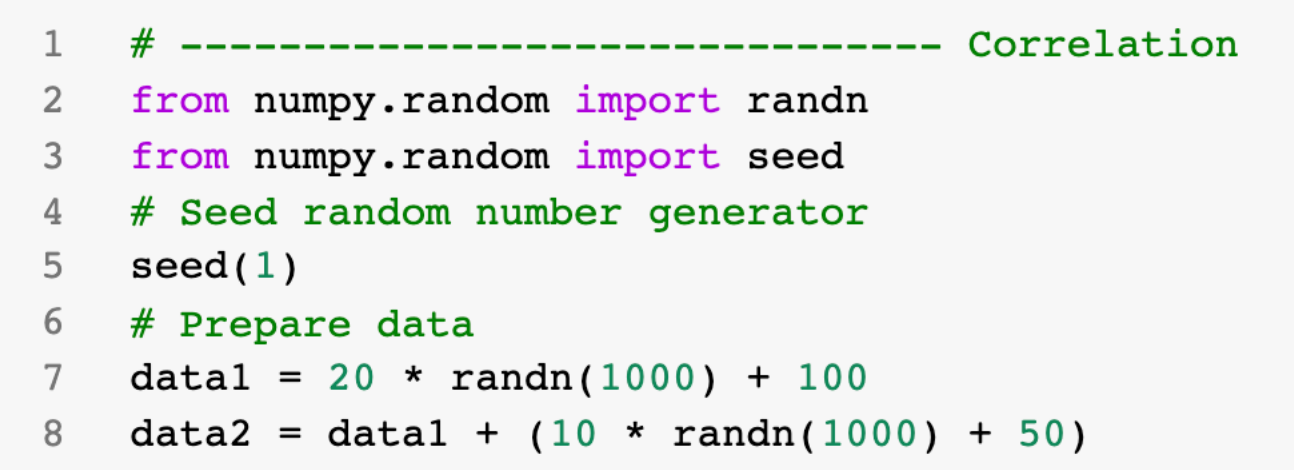
Spearman is computed on ranks while Pearson is on true values.

Pearson or Spearman

- If you want to explore your data it is best to compute both, since the relation between the Spearman and Pearson correlations will give some information.
- Briefly, Spearman is computed on ranks while Pearson is on true values.

Correlation example

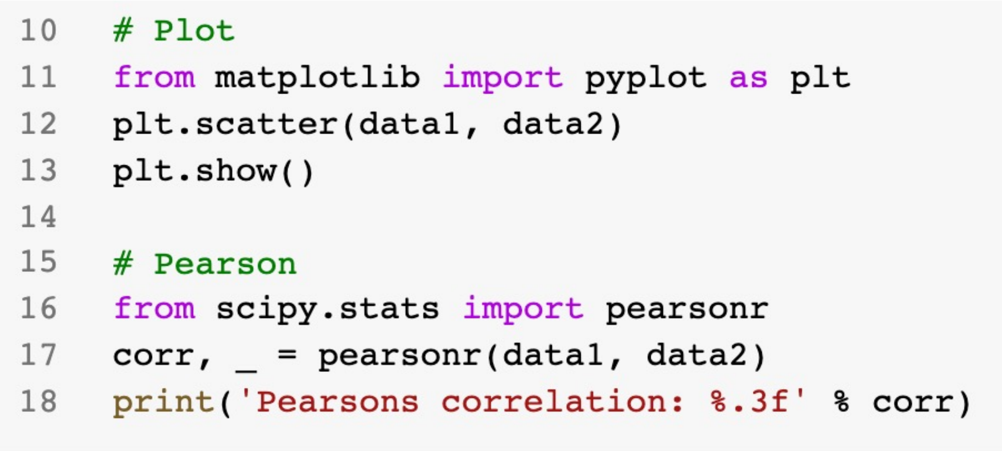
- Generate random numbers with positive correlation



```
# ------------------------------- Correlation
from numpy.random import randn
from numpy.random import seed
# Seed random number generator
seed(1)
# Prepare data
data1 = 20 * randn(1000) + 100
data2 = data1 + (10 * randn(1000) + 50)
```

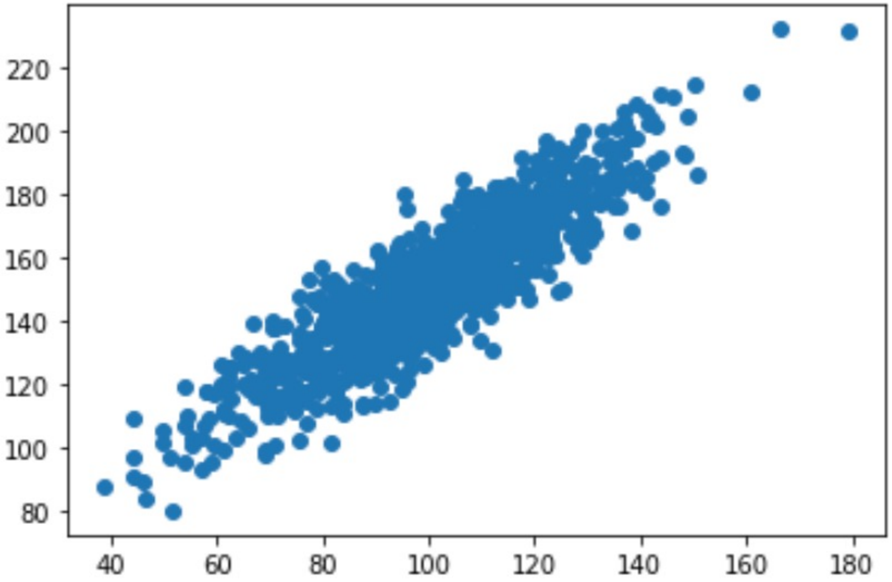
Correlation example

- Plot scatter plot



```python
# Plot
from matplotlib import pyplot as plt
plt.scatter(data1, data2)
plt.show()

# Pearson
from scipy.stats import pearsonr
corr, _ = pearsonr(data1, data2)
print('Pearsons correlation: %.3f' % corr)
```



```
Pearsons correlation: 0.888
```

Hands on session

Problem Solving

Reference

Claes Wohlin · Per Runeson
Martin Höst · Magnus C. Ohlsson
Björn Regnell · Anders Wesslén

Experimentation in Software Engineering

Springer

Chapter 6