Intro to Data Science

Lecture 01

Juice shop

- Assume you've opened a "juice shop", and you want to improve your business.
- If you are smart enough, you'll start recording sales data to understand your business.
- What kind of data you'll record?

Collecting Data

- Temporal data
- Text data
  - text
  - categorical
- Numerical data
  - Continuous
  - Discrete
  - Categorical

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/01/19	Sun	27	2	15	0.3	10
01/02/19	Mon	28.9	1.33	15	0.3	13
01/03/19	Tue	34.5	1.33	27	0.3	15
01/04/19	Wen	44.1	1.05	28	0.3	17
01/05/19	Thu	42.4	1	33	0.3	18
01/06/19	Fri	25.3	1.54	23	0.3	11
01/07/19	Sat	32.9	1.54	99	0.5	13
01/08/19	Sun	37.5	1.18	28	0.5	15
01/09/19	Mon	38.1	1.18	20	0.5	17
01/10/19	Tue	43.4	1.05	33	0.5	18
01/11/19	Wen	32.6	1.54	23	0.5	12
01/12/19	Thu	38.2	1.33	16	0.5	14
01/13/19	Fri	37.5	1.33	19	0.5	15
01/14/19	Sat	44.1	1.05	23	0.3	17
01/15/19	Sun	43.4	1.11	33	0.3	18
01/16/19	Mon	30.6	1.67	24	0.3	12
01/17/19	Tue	32.2	1.43	26	0.3	14
01/18/19	Wen	42.8	1.18	33	0.3	16

Sorting

- The data is sorted by date

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/01/19	Sun	27	2	15	0.3	10
01/02/19	Mon	28.9	1.33	15	0.3	13
01/03/19	Tue	34.5	1.33	27	0.3	15
01/04/19	Wen	44.1	1.05	28	0.3	17
01/05/19	Thu	42.4	1	33	0.3	18
01/06/19	Fri	25.3	1.54	23	0.3	11
01/07/19	Sat	32.9	1.54	99	0.5	13
01/08/19	Sun	37.5	1.18	28	0.5	15
01/09/19	Mon	38.1	1.18	20	0.5	17
01/10/19	Tue	43.4	1.05	33	0.5	18
01/11/19	Wen	32.6	1.54	23	0.5	12
01/12/19	Thu	38.2	1.33	16	0.5	14
01/13/19	Fri	37.5	1.33	19	0.5	15
01/14/19	Sat	44.1	1.05	23	0.3	17
01/15/19	Sun	43.4	1.11	33	0.3	18
01/16/19	Mon	30.6	1.67	24	0.3	12
01/17/19	Tue	32.2	1.43	26	0.3	14
01/18/19	Wen	42.8	1.18	33	0.3	16

Sorting

- We can sort the data by sales

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/01/19	Sun	27	2	15	0.3	10
01/06/19	Fri	25.3	1.54	23	0.3	11
01/16/19	Mon	30.6	1.67	24	0.3	12
01/11/19	Wen	32.6	1.54	23	0.5	12
01/02/19	Mon	28.9	1.33	15	0.3	13
01/07/19	Sat	32.9	1.54	99	0.5	13
01/12/19	Thu	38.2	1.33	16	0.5	14
01/17/19	Tue	32.2	1.43	26	0.3	14
01/13/19	Fri	37.5	1.33	19	0.5	15
01/08/19	Sun	37.5	1.18	28	0.5	15
01/03/19	Tue	34.5	1.33	27	0.3	15
01/18/19	Wen	42.8	1.18	33	0.3	16
01/09/19	Mon	38.1	1.18	20	0.5	17
01/14/19	Sat	44.1	1.05	23	0.3	17
01/04/19	Wen	44.1	1.05	28	0.3	17
01/15/19	Sun	43.4	1.11	33	0.3	18
01/05/19	Thu	42.4	1	33	0.3	18
01/10/19	Tue	43.4	1.05	33	0.5	18

Sorting

- Or based on Flyers

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/07/19	Sat	32.9	1.54	99	0.5	13
01/05/19	Thu	42.4	1	33	0.3	18
01/10/19	Tue	43.4	1.05	33	0.5	18
01/15/19	Sun	43.4	1.11	33	0.3	18
01/18/19	Wen	42.8	1.18	33	0.3	16
01/04/19	Wen	44.1	1.05	28	0.3	17
01/08/19	Sun	37.5	1.18	28	0.5	15
01/03/19	Tue	34.5	1.33	27	0.3	15
01/17/19	Tue	32.2	1.43	26	0.3	14
01/16/19	Mon	30.6	1.67	24	0.3	12
01/06/19	Fri	25.3	1.54	23	0.3	11
01/11/19	Wen	32.6	1.54	23	0.5	12
01/14/19	Sat	44.1	1.05	23	0.3	17
01/09/19	Mon	38.1	1.18	20	0.5	17
01/13/19	Fri	37.5	1.33	19	0.5	15
01/12/19	Thu	38.2	1.33	16	0.5	14
01/01/19	Sun	27	2	15	0.3	10
01/02/19	Mon	28.9	1.33	15	0.3	13

Sorting

- Outliers:
  - Data points far away from others
- Outliers can largely affect the analysis.
- Outliers might be mistakes or very rare.

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/07/19	Sat	32.9	1.54	99	0.5	13
01/05/19	Thu	42.4	1	33	0.3	18
01/10/19	Tue	43.4	1.05	33	0.5	18
01/15/19	Sun	43.4	1.11	33	0.3	18
01/18/19	Wen	42.8	1.18	33	0.3	16
01/04/19	Wen	44.1	1.05	28	0.3	17
01/08/19	Sun	37.5	1.18	28	0.5	15
01/03/19	Tue	34.5	1.33	27	0.3	15
01/17/19	Tue	32.2	1.43	26	0.3	14
01/16/19	Mon	30.6	1.67	24	0.3	12
01/06/19	Fri	25.3	1.54	23	0.3	11
01/11/19	Wen	32.6	1.54	23	0.5	12
01/14/19	Sat	44.1	1.05	23	0.3	17
01/09/19	Mon	38.1	1.18	20	0.5	17
01/13/19	Fri	37.5	1.33	19	0.5	15
01/12/19	Thu	38.2	1.33	16	0.5	14
01/01/19	Sun	27	2	15	0.3	10
01/02/19	Mon	28.9	1.33	15	0.3	13

Filtering

- We can filter data based on any criteria on any of the fields.
  - Day=’Sat’ or ‘Sun’

Date	Day	Temperature	Rainfall	Flyers	Price	Sales
01/07/19	Sat	32.9	1.54	99	0.5	13
01/15/19	Sun	43.4	1.11	33	0.3	18
01/08/19	Sun	37.5	1.18	28	0.5	15
01/14/19	Sat	44.1	1.05	23	0.3	17
01/01/19	Sun	27	2	15	0.3	10

  - Temperature<30

Date	Day	Temperature	Rainfall	Bills	Price	Sales
01/06/19	Fri	25.3	1.54	23	0.3	11
01/01/19	Sun	27	2	15	0.3	10
01/02/19	Mon	28.9	1.33	15	0.3	13

Drive values from existing data

- You can do any kind of calculation on any field
  - Change temperature from C° to F°
  - Adding a Month field
- Or generate a new field by combining already existing fields
  - Revenue: Sales*Price

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9

Aggregating data

We can use aggregating functions (e. g., sum) to summarize data and get feel as a whole.

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8
Total Revenue							***100.3***

Aggregating data

Count, Distinct Count, Sum, Min, Max

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8
Count	***Dcount=7***	***18***	***18***	***18***	***18***	***18***	***18***
sum		***655.5***	***23.84***	***518***	***6.8***	***265***	***100.3***
Average		***36.416666667***	***1.324444***	***28.77778***	***0.37778***	***14.7222***	***5.57222***
Min		***25.3***	***1***	***15***	***0.3***	***10***	***3***
Max		***44.1***	***2***	***99***	***0.5***	***18***	***9***

Highlighting Data

Interpreting numbers in large tables is difficult.

- We can use heatmaps to visualize the scale of values

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8

Highlighting Data

Interpreting numbers in large tables is difficult.

- We can use “data bars” to visualize the scale of values

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8

Highlighting Data

Interpreting numbers in large tables is difficult.

- We can highlight individual values that fall within sum criteria:
  - e. g., top 30% (good days) and less 30% (bad days) Revenues

Date	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/19	Sun	27	2	15	0.3	10	3
01/02/19	Mon	28.9	1.33	15	0.3	13	3.9
01/03/19	Tue	34.5	1.33	27	0.3	15	4.5
01/04/19	Wen	44.1	1.05	28	0.3	17	5.1
01/05/19	Thu	42.4	1	33	0.3	18	5.4
01/06/19	Fri	25.3	1.54	23	0.3	11	3.3
01/07/19	Sat	32.9	1.54	99	0.5	13	6.5
01/08/19	Sun	37.5	1.18	28	0.5	15	7.5
01/09/19	Mon	38.1	1.18	20	0.5	17	8.5
01/10/19	Tue	43.4	1.05	33	0.5	18	9
01/11/19	Wen	32.6	1.54	23	0.5	12	6
01/12/19	Thu	38.2	1.33	16	0.5	14	7
01/13/19	Fri	37.5	1.33	19	0.5	15	7.5
01/14/19	Sat	44.1	1.05	23	0.3	17	5.1
01/15/19	Sun	43.4	1.11	33	0.3	18	5.4
01/16/19	Mon	30.6	1.67	24	0.3	12	3.6
01/17/19	Tue	32.2	1.43	26	0.3	14	4.2
01/18/19	Wen	42.8	1.18	33	0.3	16	4.8

Grouping data

It is common to group data by categorical fields and compute subtotal values

Day	*Avg. Temp*	*Avg. Rain*	*Flyers*	*Sales*	*Revenue*
Sun	35.97	1.43	76	43	15.9
Mon	33.5	1.255	35	30	12.4
Tue	38.95	1.19	60	33	13.5
Wen	38.35	1.295	51	29	11.1
Thu	40.3	1.165	49	32	12.4
Fri	31.4	1.435	42	26	10.8
Sat	38.5	1.295	122	30	11.6
Grand Total	**36.66**	**1.304**	**435**	**223**	**87.7**

Grouping data

on more than one field

Price=0.3					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	35.2	1.555	48	28	8.4
Mon	28.9	1.33	15	13	3.9
Tue	34.5	1.33	27	15	4.5
Wen	44.1	1.05	28	17	5.1
Thu	42.4	1	33	18	5.4
Fri	25.3	1.54	23	11	3.3.
Sat	44.1	1.05	23	17	5.1
Grand Total	**36.21**	**1.301**	**197**	**119**	**35.7**

Price=0.5					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	37.5	1.18	28	15	7.5
Mon	38.1	1.18	20	17	8.5
Tue	43.4	1.05	33	18	9
Wen	32.6	1.54	23	12	6
Thu	38.2	1.33	16	14	7
Fri	37.5	1.33	19	15	7.5
Sat	32.9	1.54	99	13	6.5
Grand Total	**37.17**	**1.307**	**238**	**104**	**52**

Grouping data

on more than one field

Price=0.3					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	35.2	1.555	48	28	8.4
Mon	28.9	1.33	15	13	3.9
Tue	34.5	1.33	27	15	4.5
Wen	44.1	1.05	28	17	5.1
Thu	42.4	1	33	18	5.4
Fri	25.3	1.54	23	11	3.3.
Sat	44.1	1.05	23	17	5.1
Grand Total	**36.21**	**1.301**	**197**	**119**	**35.7**

Price=0.5					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	37.5	1.18	28	15	7.5
Mon	38.1	1.18	20	17	8.5
Tue	43.4	1.05	33	18	9
Wen	32.6	1.54	23	12	6
Thu	38.2	1.33	16	14	7
Fri	37.5	1.33	19	15	7.5
Sat	32.9	1.54	99	13	6.5
Grand Total	**37.17**	**1.307**	**238**	**104**	**52**

Grouping data

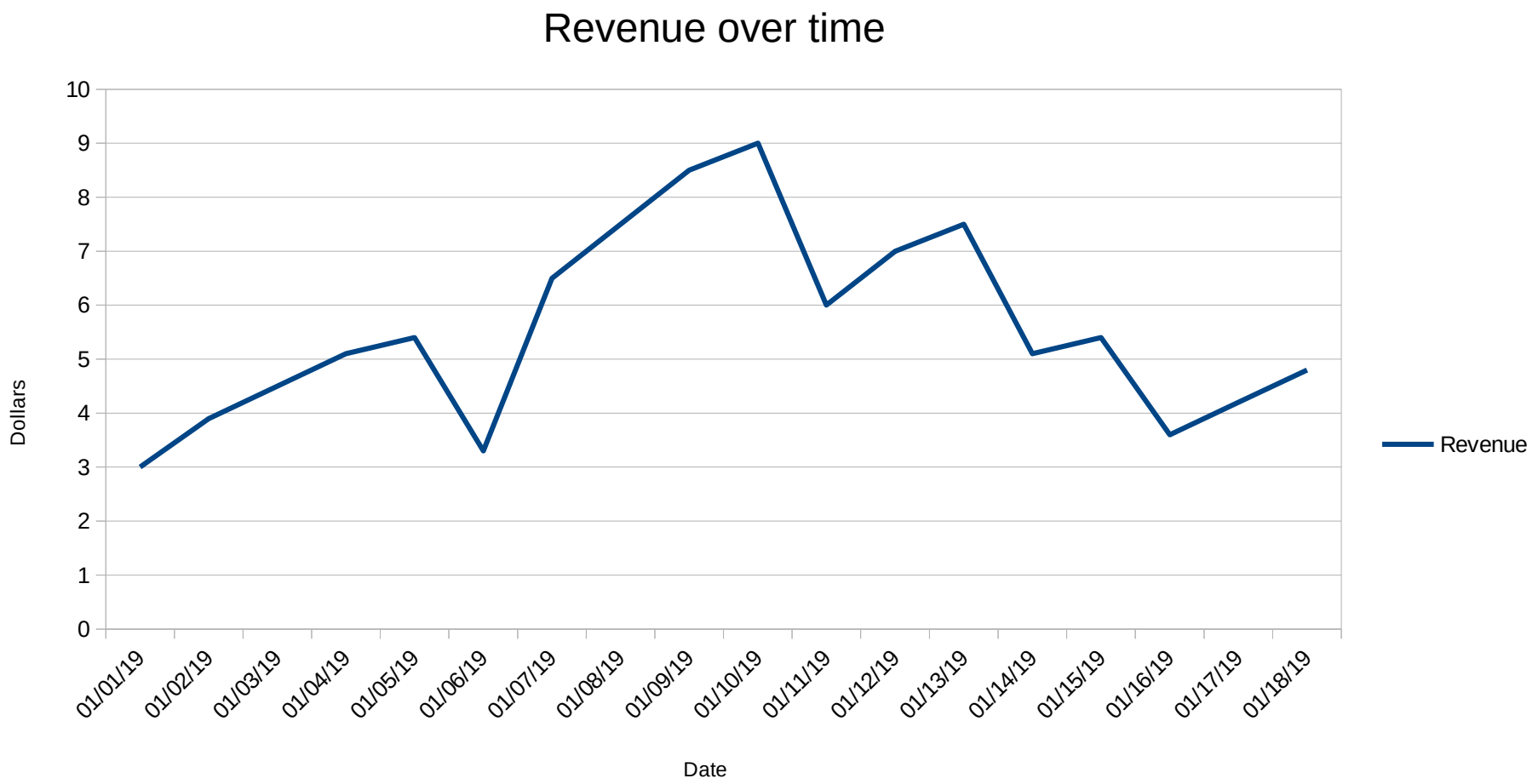
on more than one field

Price=0.3					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	35.2	1.555	48	28	8.4
Mon	28.9	1.33	15	13	3.9
Tue	34.5	1.33	27	15	4.5
Wen	44.1	1.05	28	17	5.1
Thu	42.4	1	33	18	5.4
Fri	25.3	1.54	23	11	3.3.
Sat	44.1	1.05	23	17	5.1
Grand Total	**36.21**	**1.301**	**197**	**119**	**35.7**

Price=0.5					
Day	***Avg. Temp***	***Avg. Rain***	***Flyers***	***Sales***	***Revenue***
Sun	37.5	1.18	28	15	7.5
Mon	38.1	1.18	20	17	8.5
Tue	43.4	1.05	33	18	9
Wen	32.6	1.54	23	12	6
Thu	38.2	1.33	16	14	7
Fri	37.5	1.33	19	15	7.5
Sat	32.9	1.54	99	13	6.5
Grand Total	**37.17**	**1.307**	**238**	**104**	**52**

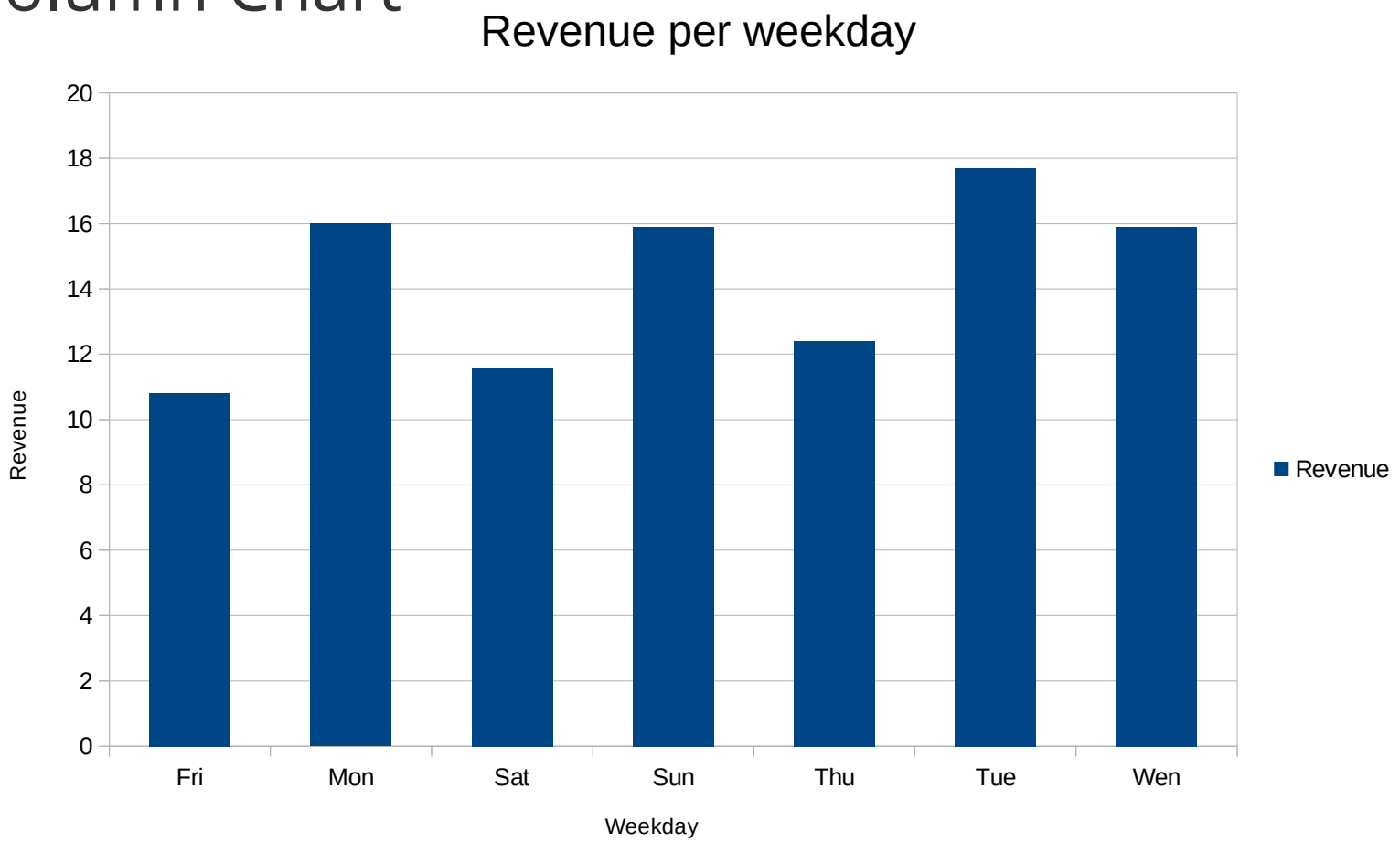
Visualization

Line Plot



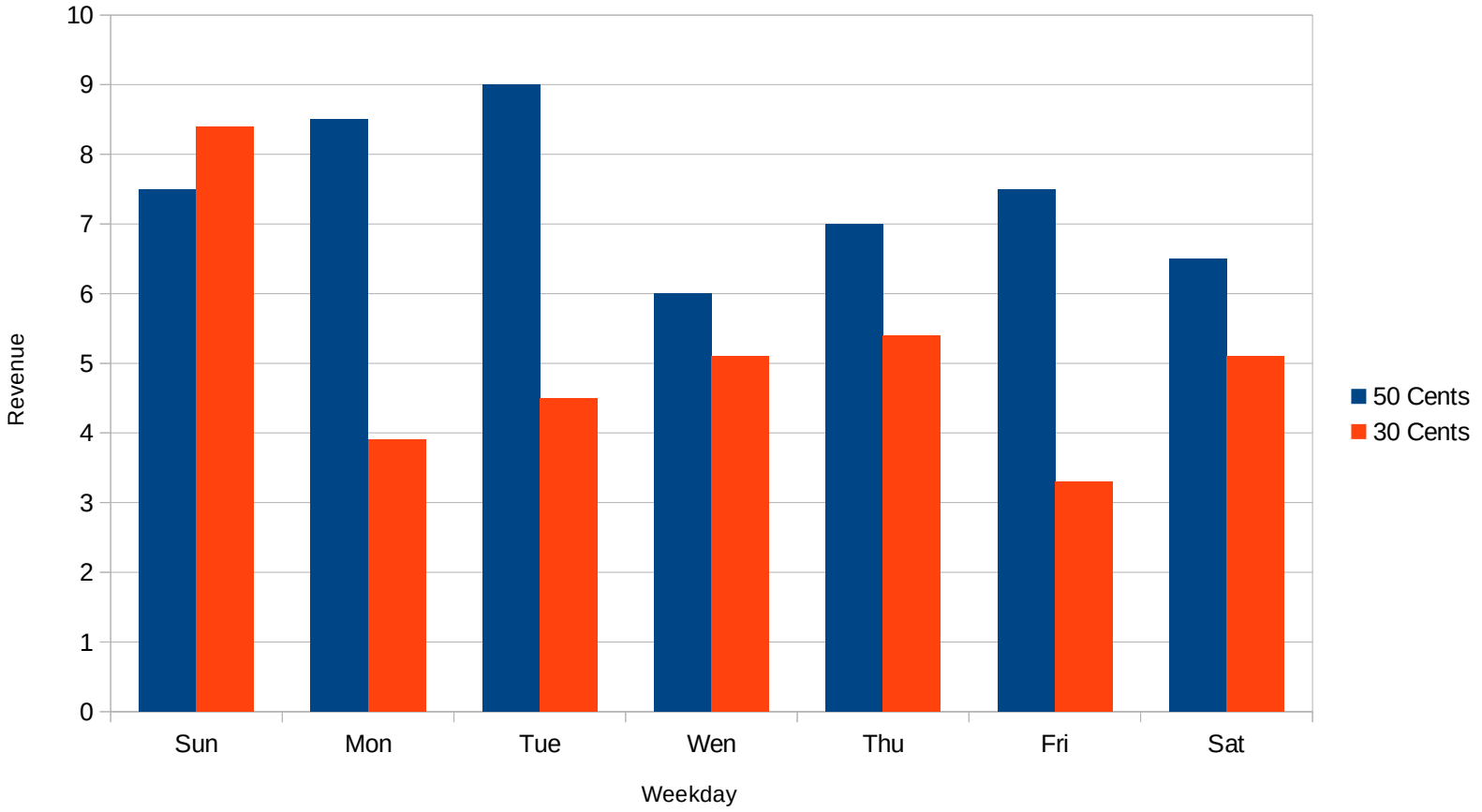
Visualization

Column Chart



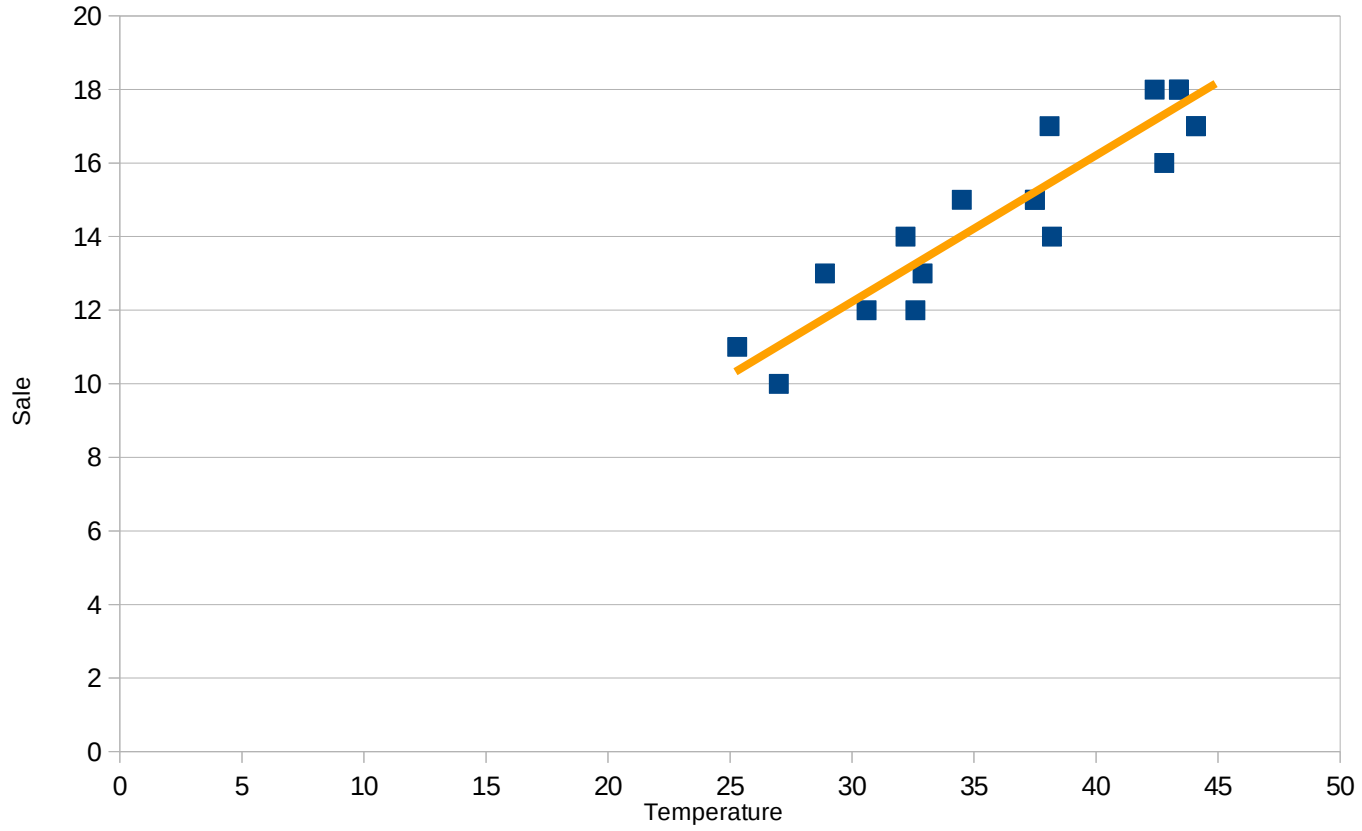
Visualization

Joint column chart



Visualization

Scater plot



Statistical analysis

- Statistics is the core of data science.

- Using Statistics
  - You can see the distribution of the data
  - How much variance there is between values
  - How changes in one feature affect values of other features

- The first point to start → descriptive statistics

Data distribution

Temperature
20
48.6
47.2
32.5
33.1
63.9
32.6
46.2
47.1
44.3
46.2
64.2
54.1
56.2
53.1

$$\bar{x} = \frac{\sum_{i-1}^{n} x_i}{n}$$

Min=20 Mean=44.62 Max=64.2

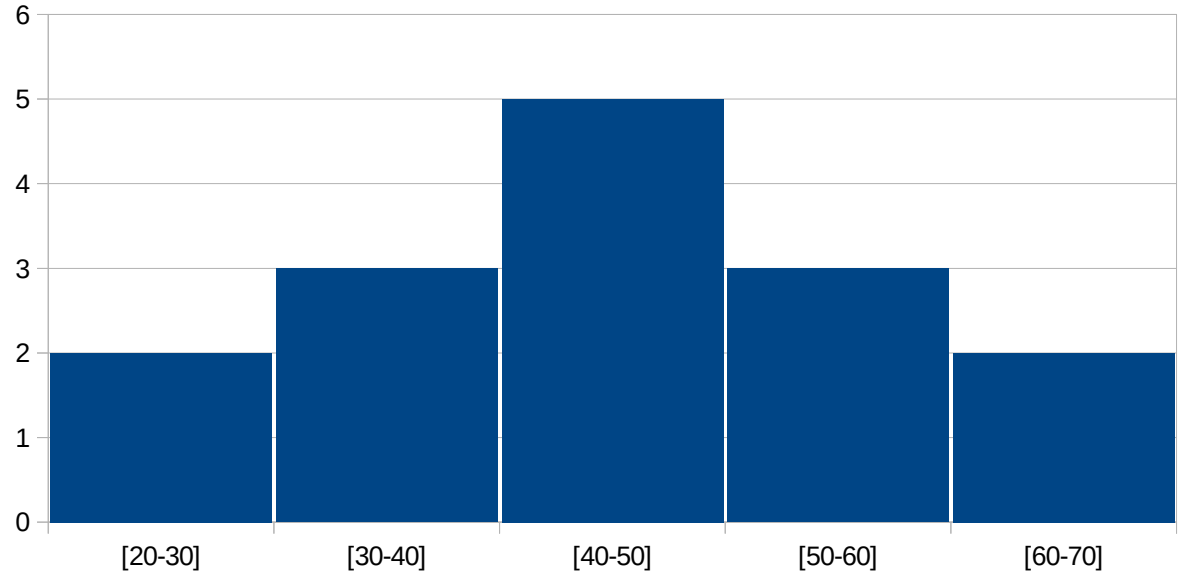
Data distribution

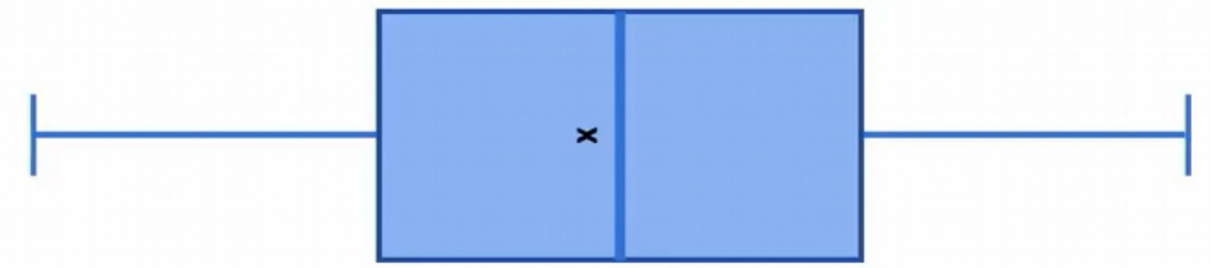
Temperature
20
32.5
32.6
33.1
44.3
46.2
46.2
47.1
47.2
48.6
53.1
54.1
56.2
63.9
64.2

Min=20 Mean=44.62 Max=64.2

Median=46.2

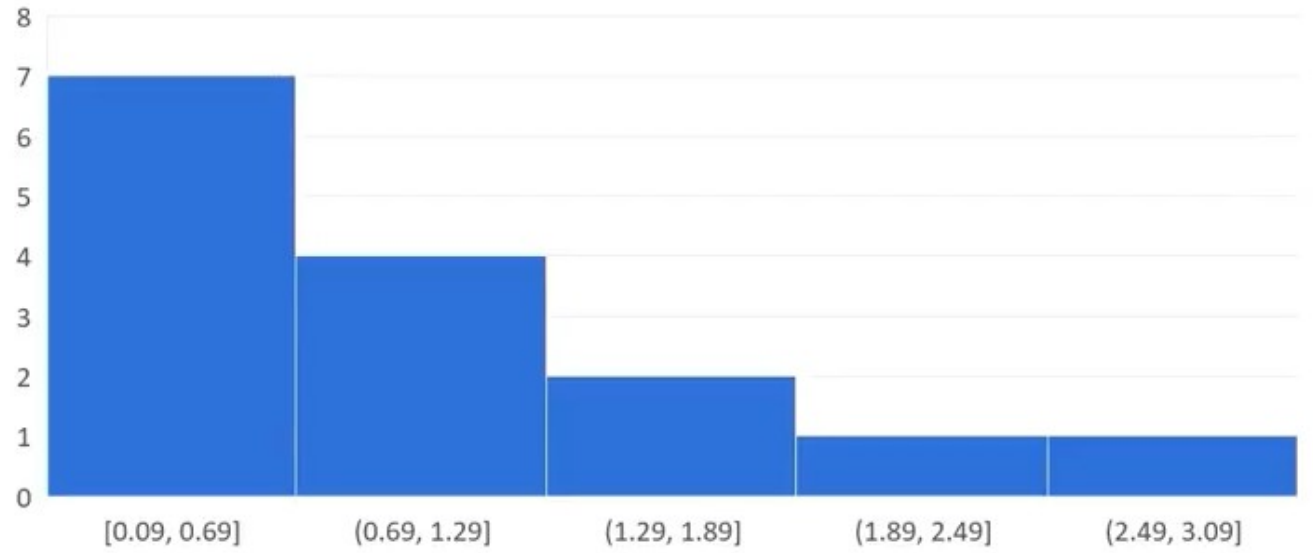
Mode=46.2





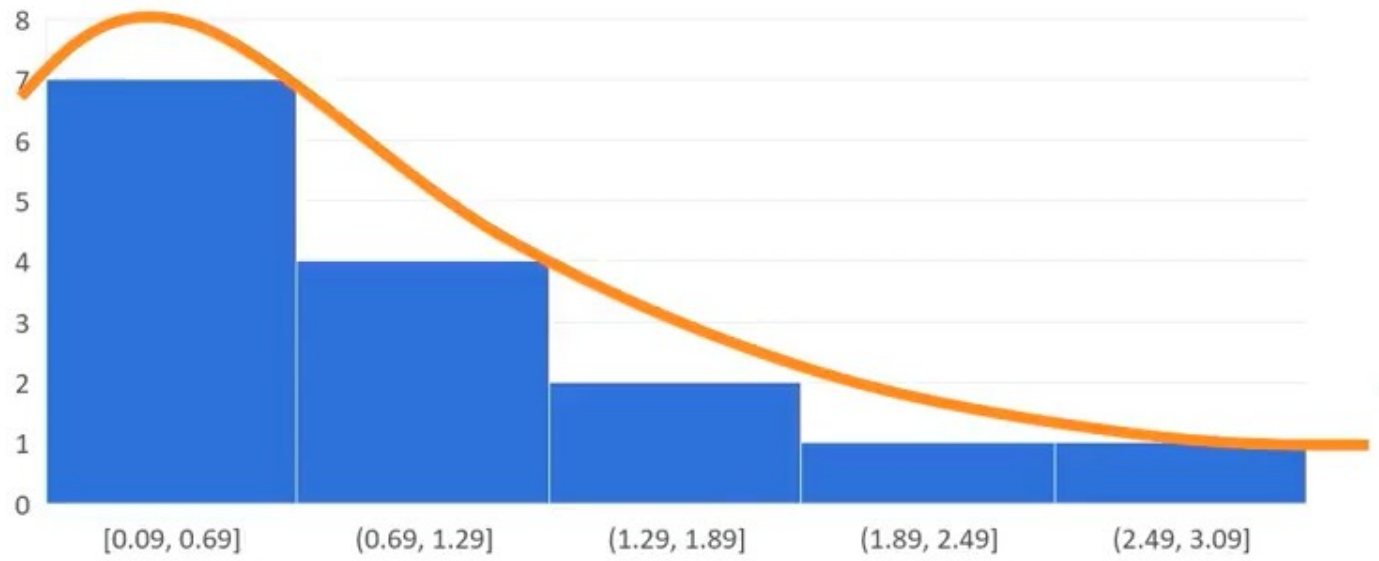
Skewed distribution

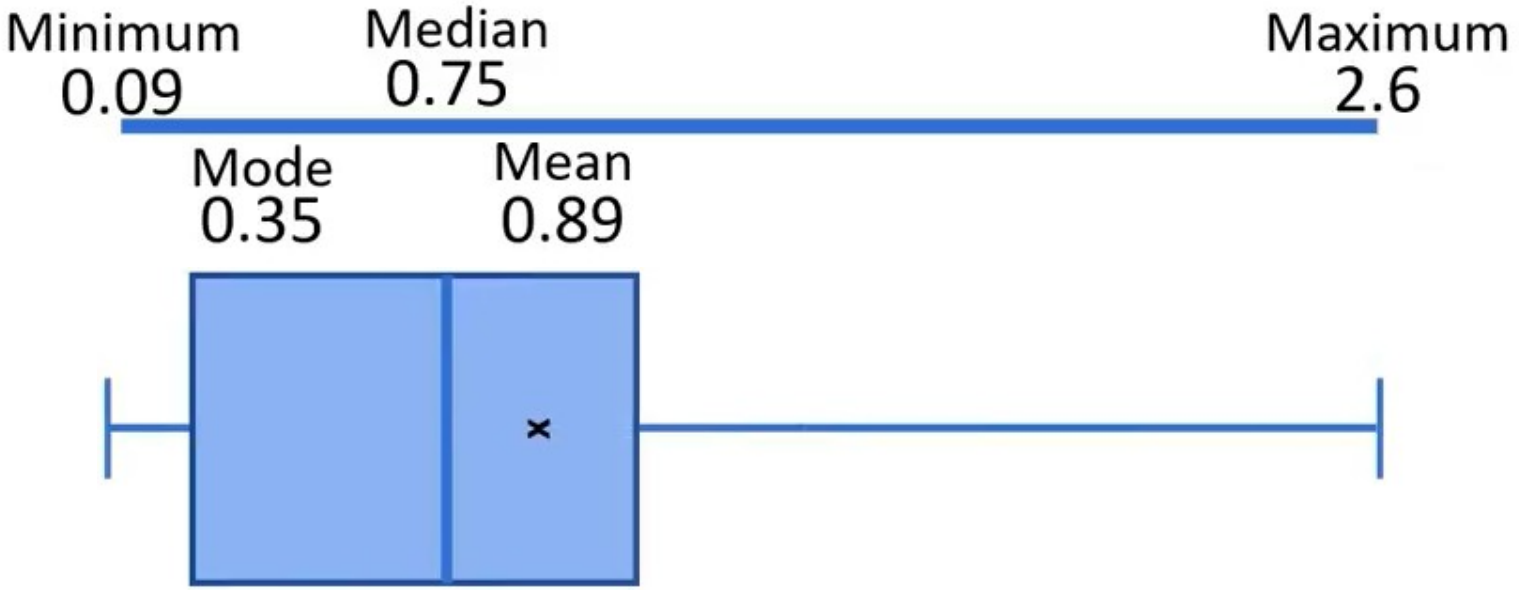
Rainfall
0.98
1.33
2.60
0.35
1.60
0.10
2.10
0.50
0.75
0.11
1.05
0.35
0.35
1.05
0.09



Skewed distribution

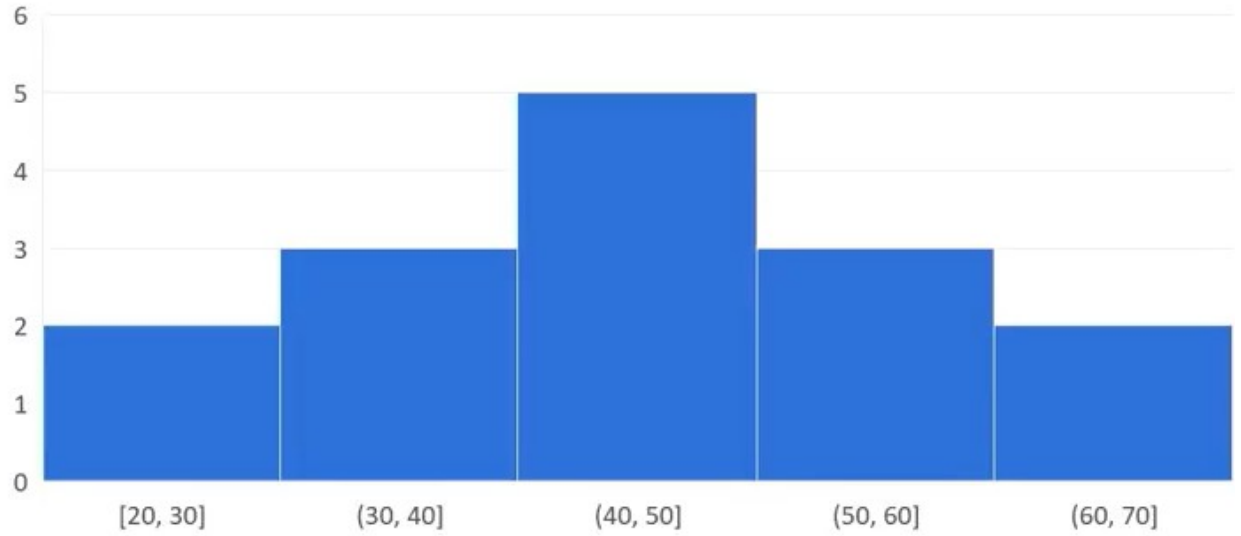
Rainfall
0.98
1.33
2.60
0.35
1.60
0.10
2.10
0.50
0.75
0.11
1.05
0.35
0.35
1.05
0.09





Variance

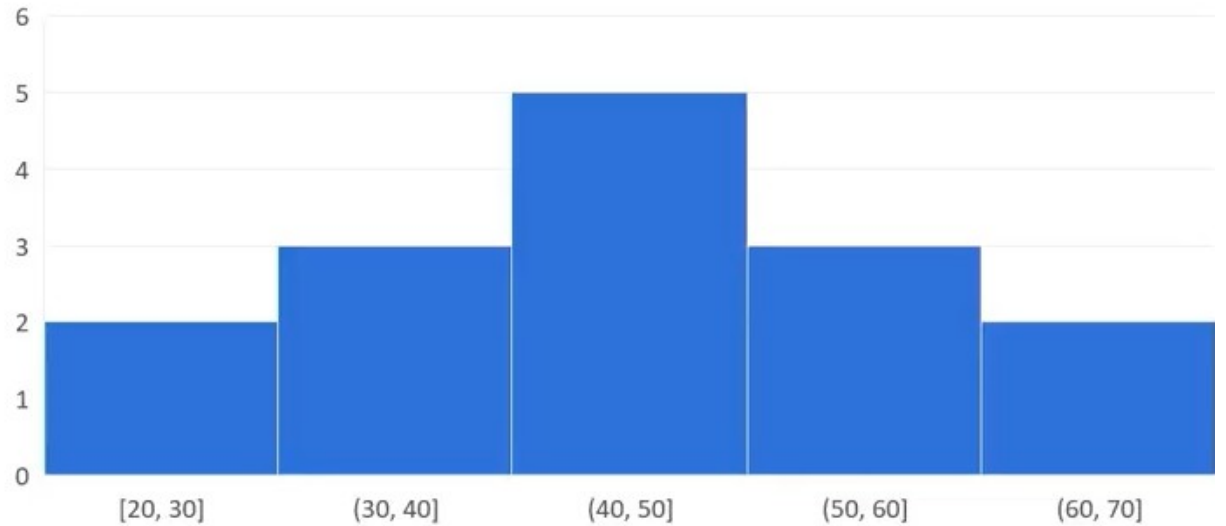
Temperature
20.0
27.2
32.5
32.6
33.1
44.3
46.2
46.2
47.1
48.6
53.1
54.1
56.2
63.9
64.2



$$\sigma^2 = \frac{\sum_{i-1}^{N} (X_i - \mu)^2}{N}$$

Standard deviation

Temperature
20.0
27.2
32.5
32.6
33.1
44.3
46.2
46.2
47.1
48.6
53.1
54.1
56.2
63.9
64.2

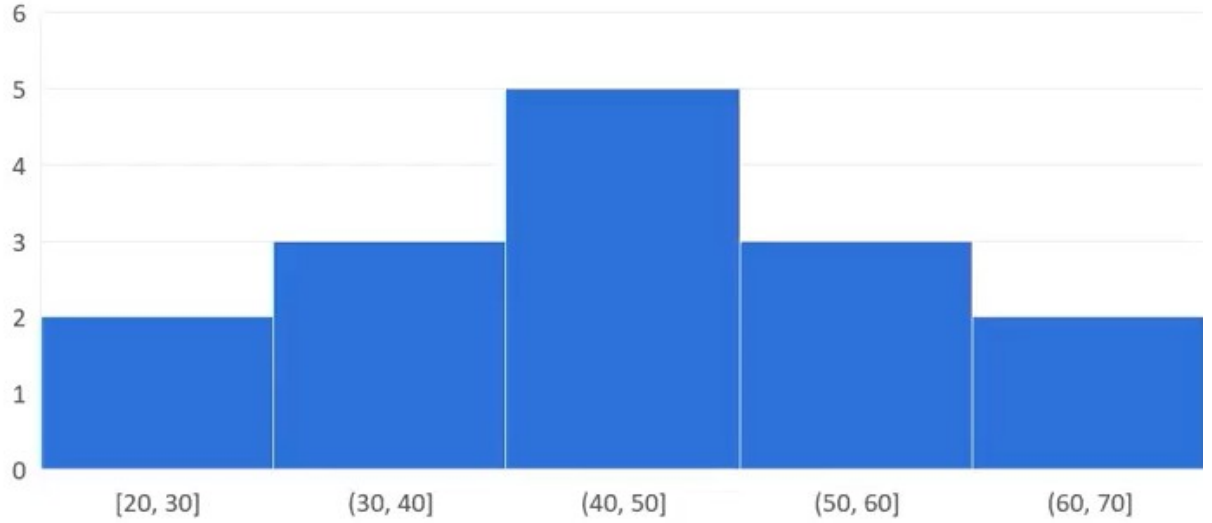


Variance: 172.27

$$\sigma = \sqrt{\frac{\sum_{i-1}^{N} (X_i - \mu)^2}{N}}$$

Standard deviation

Temperature
20.0
27.2
32.5
32.6
33.1
44.3
46.2
46.2
47.1
48.6
53.1
54.1
56.2
63.9
64.2

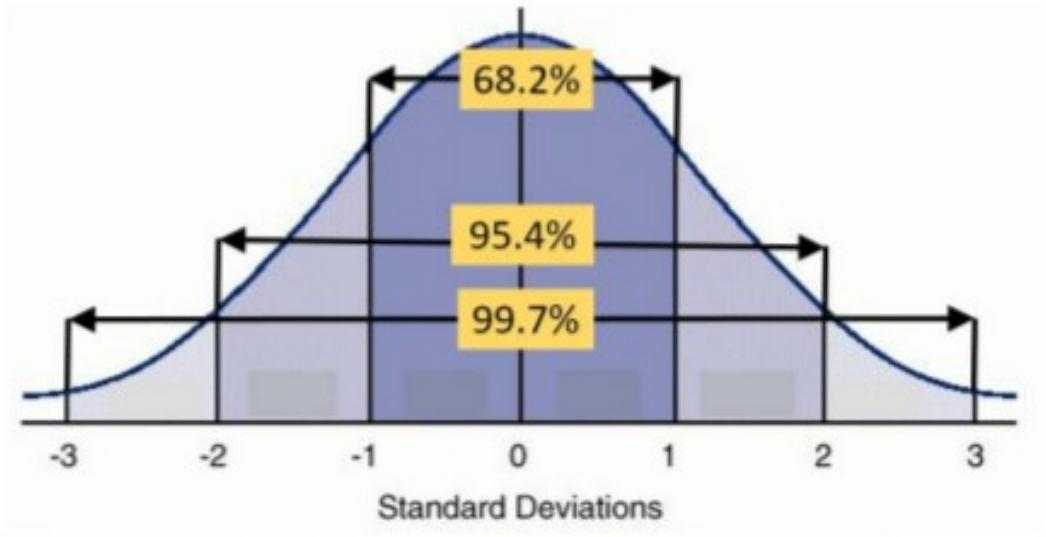


Variance: 172.27

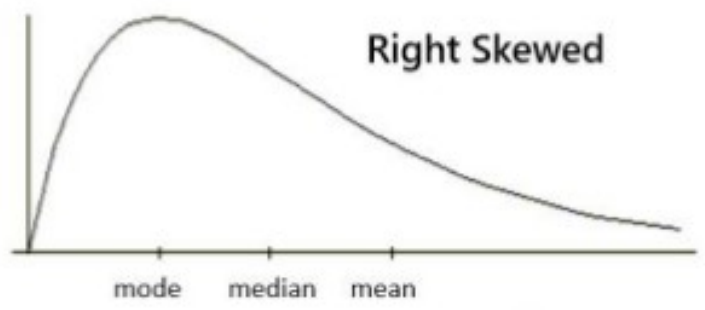
Standard Deviation: 13.13

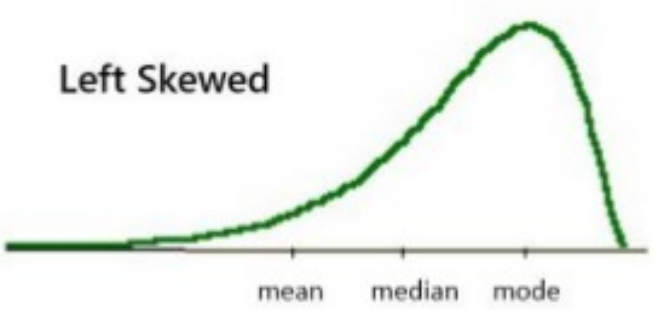
Normal distribution

Normal Distribution:



Skewed Distributions:

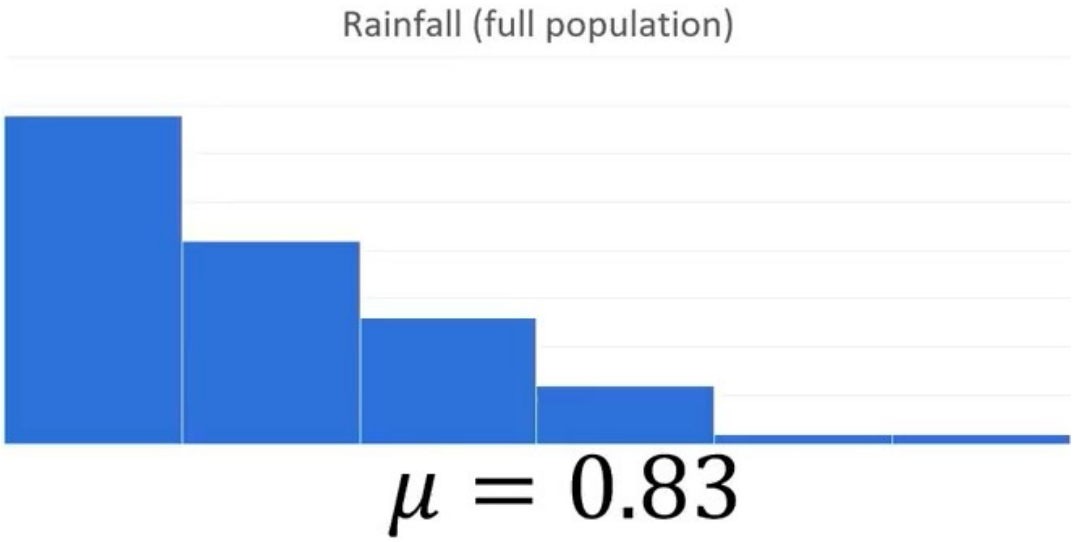




Full and sampled data

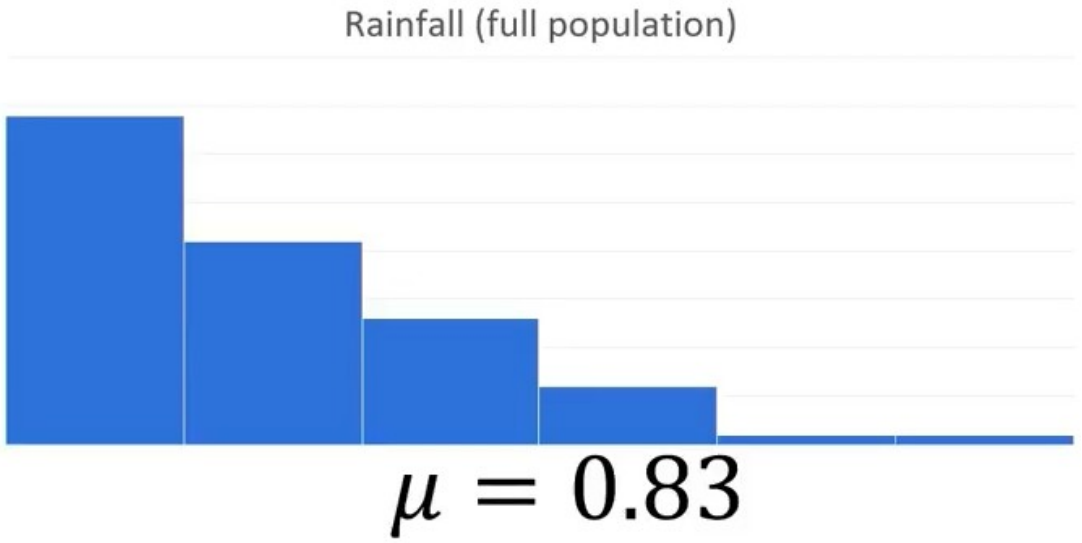
- It is not always possible to access the full data and we just have a sample data.
- Even if we have the full data, it is so large and is not easy to handle.
- The best case is to have a sample of the data that represents the whole of the data.
- This way we can use sample data statistics to approximate the whole data statistics.

Full and sampled data



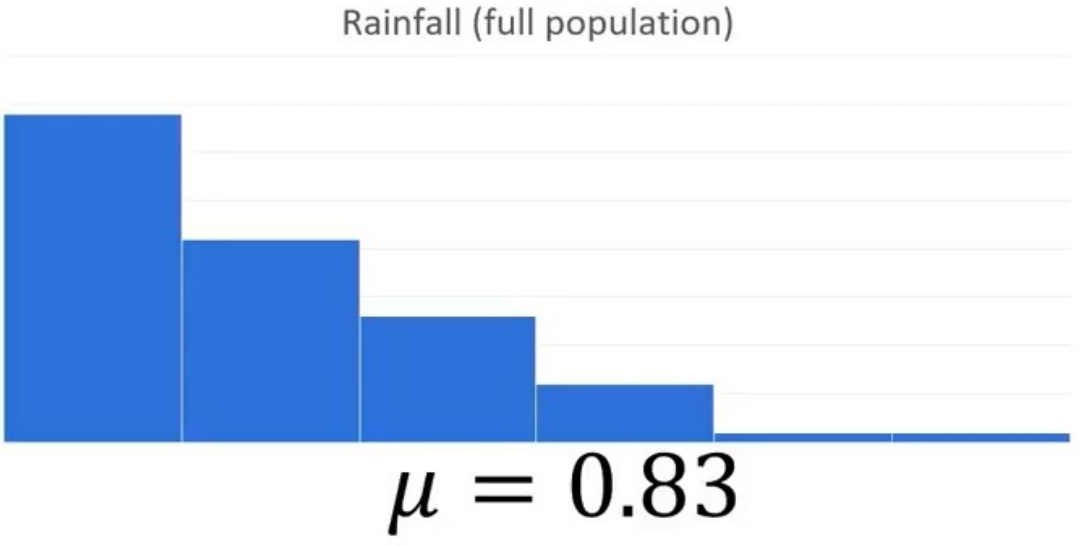
Full and sampled data

Sample
0.98
1.33
2.60
0.35
1.60
0.10
2.10
0.50
0.75
0.11
1.05
0.35
0.35
1.05
0.09



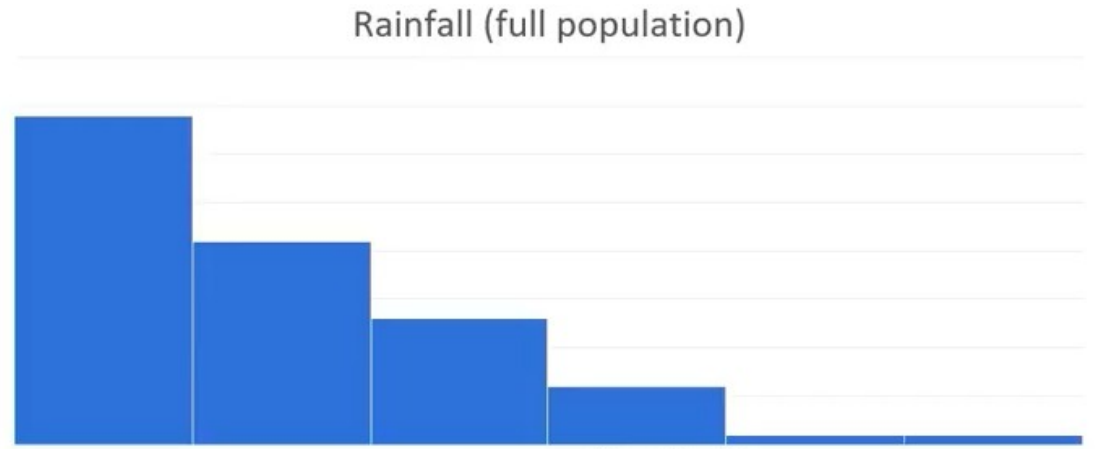
Full and sampled data

Sample
0.98
1.33
2.60
0.35
1.60
0.10
2.10
0.50
0.75
0.11
1.05
0.35
0.35
1.05
0.09

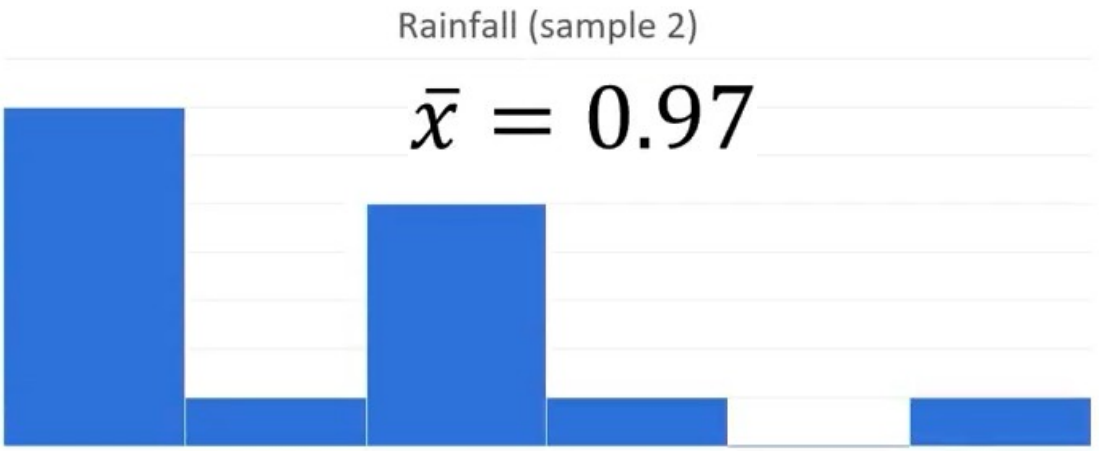


Full and sampled data

Sample
1.33
1.05
1.05
1.00
1.00
0.87
0.83
1.11
0.95
0.87
0.87
1.00
0.91
0.91
0.83



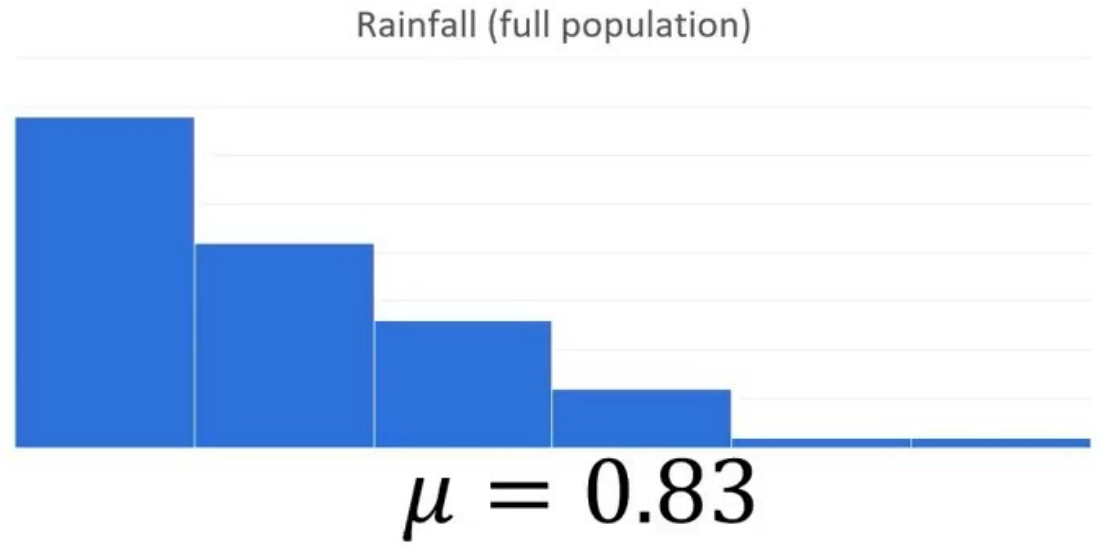
$$\mu = 0.83$$



$$\bar{x} = 0.97$$

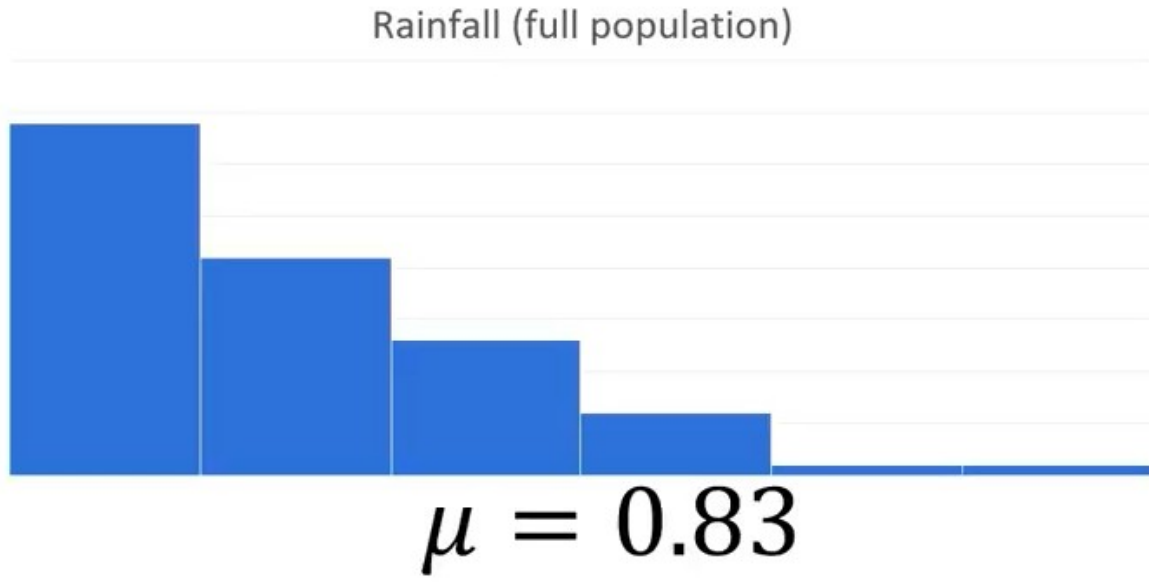
Full and sampled data

Sample
0.61
0.67
0.69
0.67
0.63
0.69
0.69
0.67
0.63
0.65
0.65
0.67
0.65
0.65
0.59

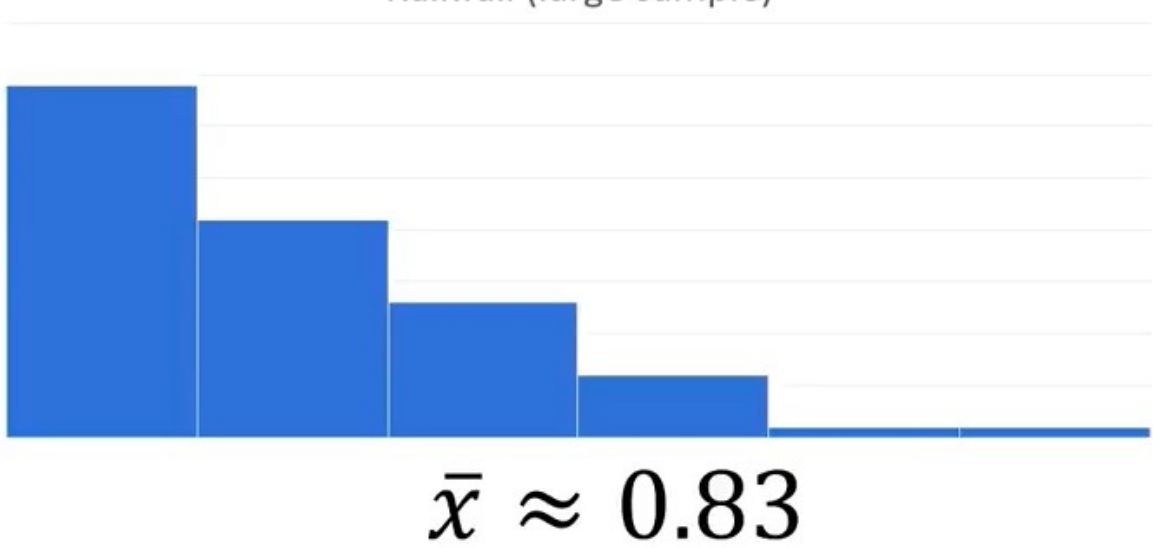


Large sample
better represents the full population

Sample
102.6
75.3
75.9
86.5
89.4
102.9
93.4
81.5
84.2
73.6
91.7
82.5
83.2
77.9
98
83.5
80.2
78.9
92
82.5
79.2
80.9
99.3
83.8
86.5
76.9
99.6



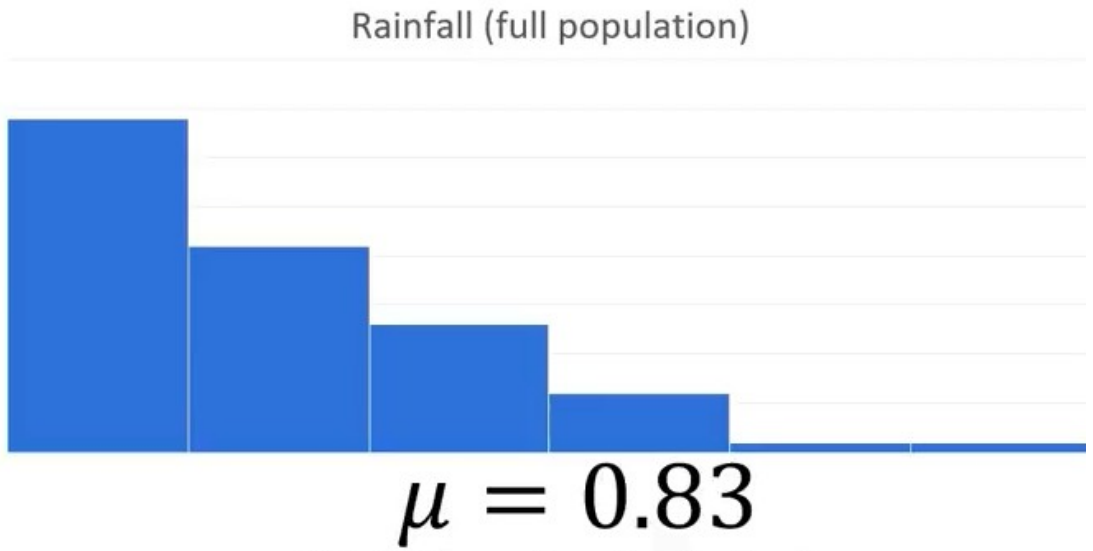
$$\mu = 0.83$$



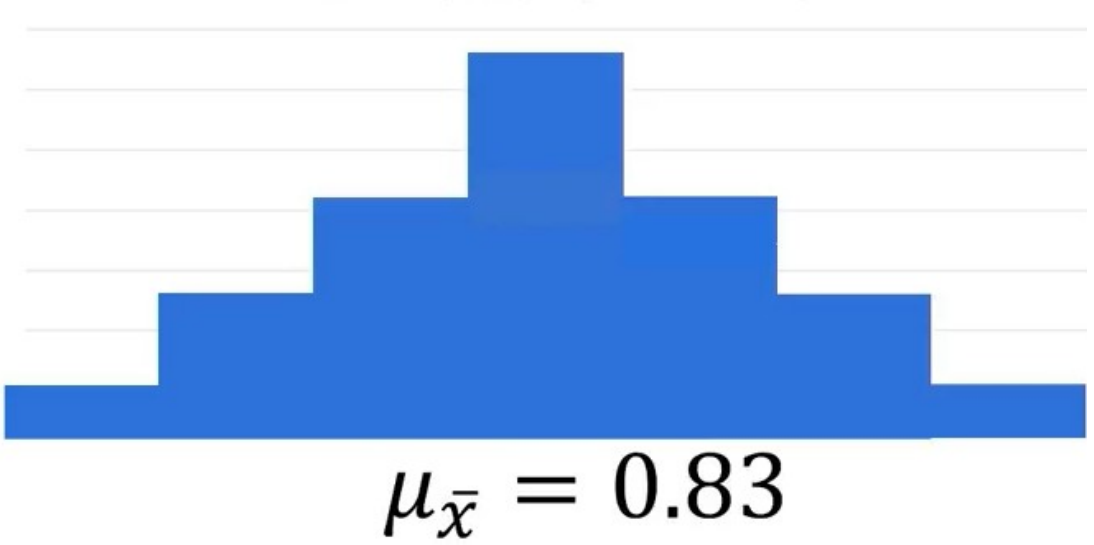
$$\bar{x} \approx 0.83$$

Mean of the sample means

would be the same as population mean



$$\mu = 0.83$$



$$\mu_{\bar{x}} = 0.83$$

$$E(\bar{x}) = \bar{x},$$

$$Var(\bar{x}) = \frac{Var(X)}{n}$$

Covariance & Correlation

The covariance of two variables in full population:

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).$$

The covariance of two variables in a sample:

$$\sigma_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

The covariance of two variables in full population:

$$Cor(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$$

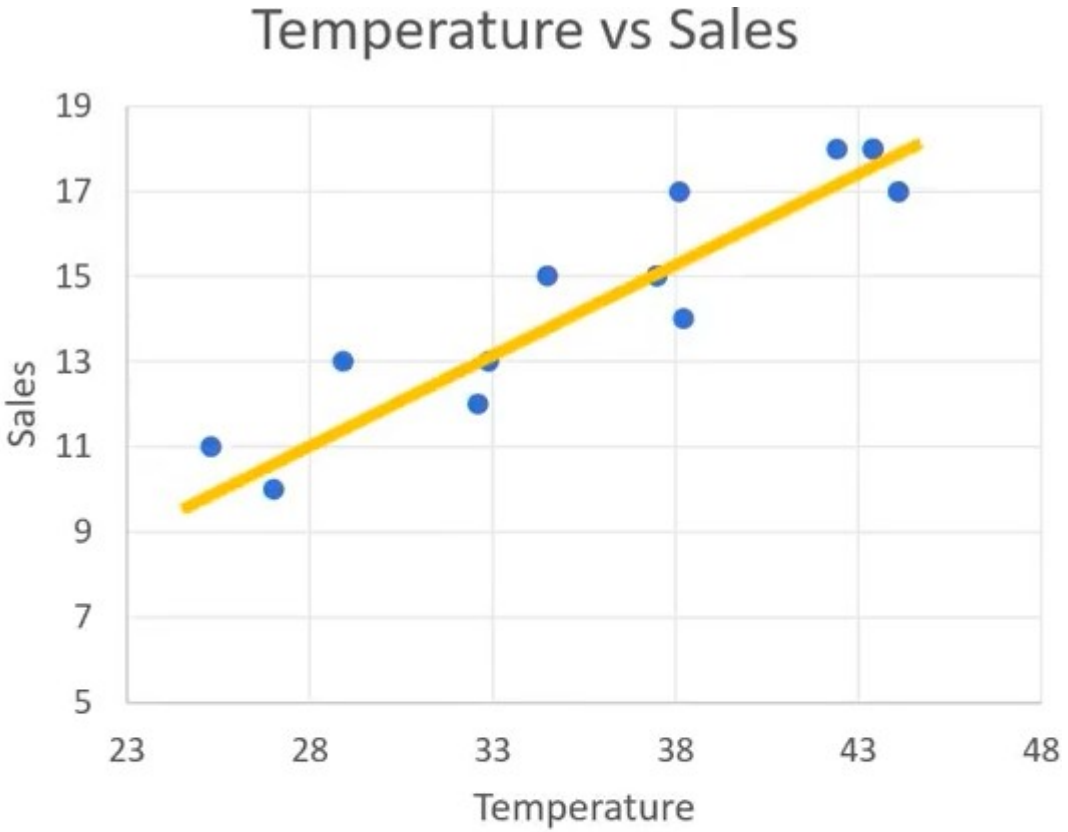
The covariance of two variables in a sample:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation

Finding relation between different fields of data

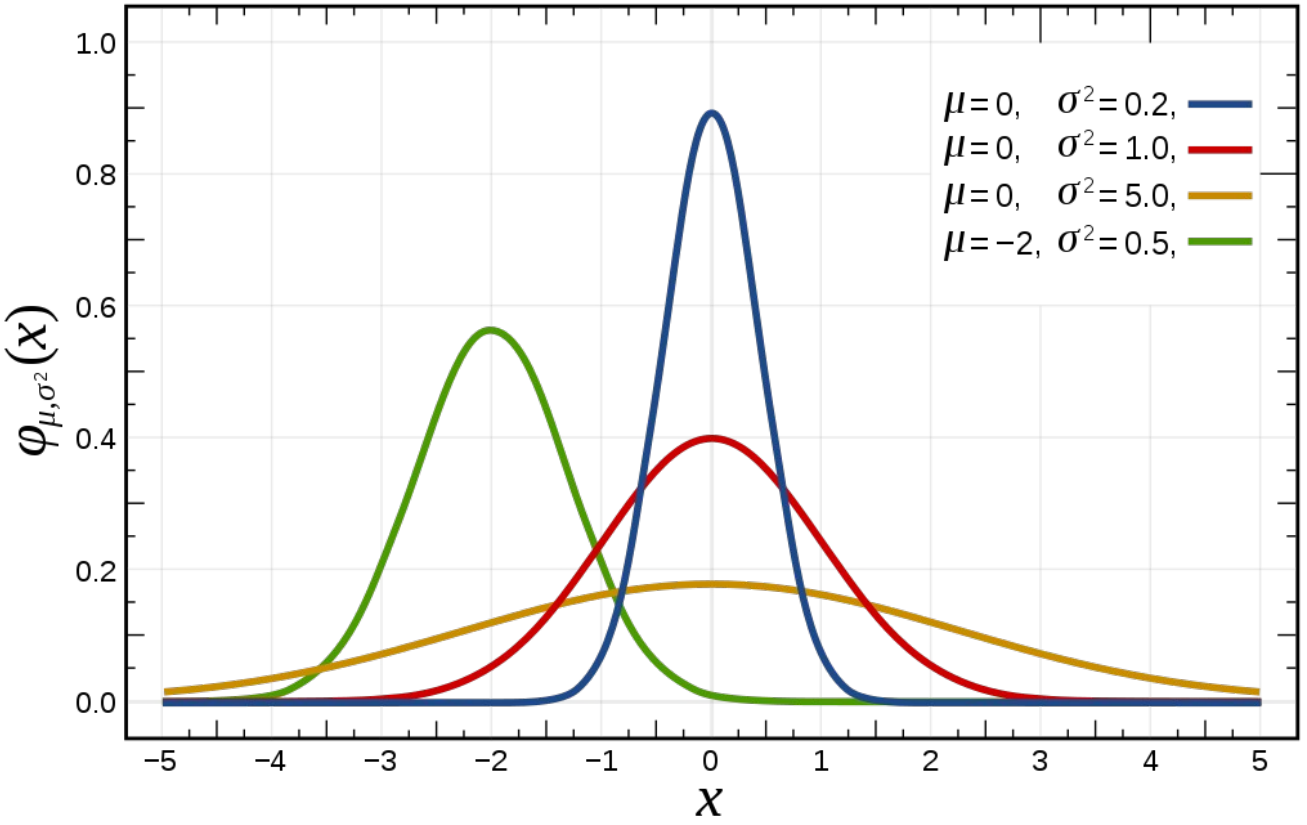
Temperature	Sales
27.0	10
28.9	13
34.5	15
44.1	17
42.4	18
25.3	11
32.9	13
37.5	15
38.1	17
43.4	18
32.6	12
38.2	14
37.5	15
44.1	17
43.4	18



Here we see a linear relation between temperature and the number of sales.

Normal Distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Hypothesis testing

Combining sample distribution with probability

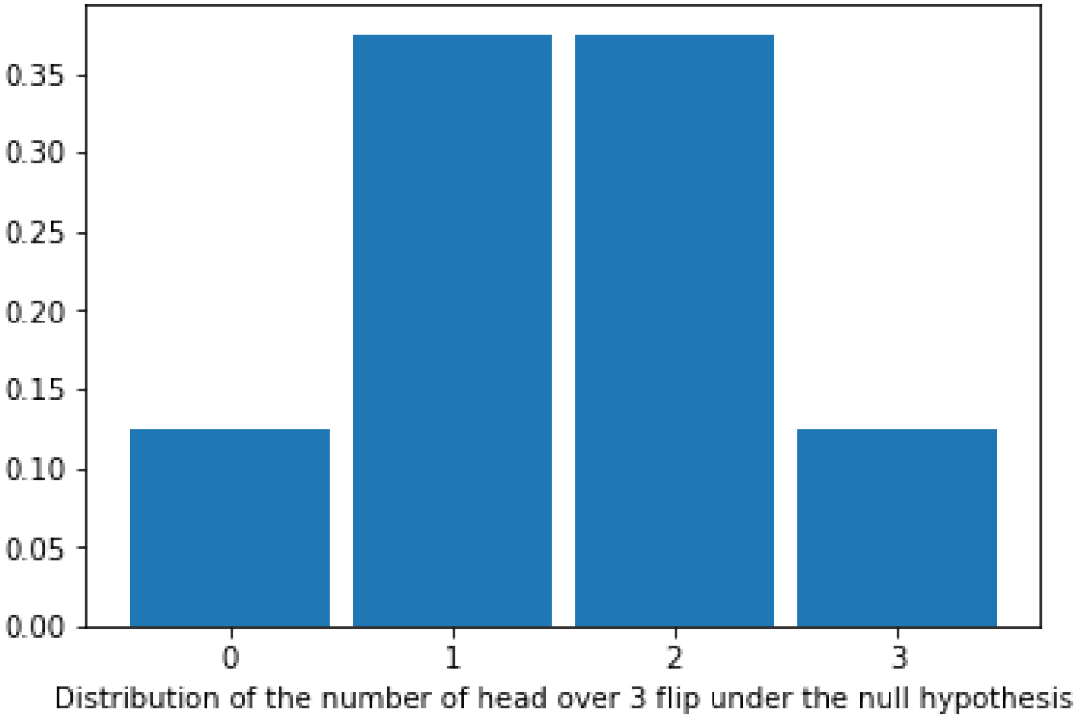
- Examples
  - Test a proportion:
    - Biased coin? 200 heads have been found over 300 flips, is it coins biased?
  - Test the association between two variables.
    - height and sex: In a sample of 25 individuals (15 females, 10 males), is female height is different from male height?
    - age and arterial hypertension: In a sample of 25 individuals is age height correlated with arterial hypertension?

Biased coin ? 2 heads have been found over 3 flips, is it biased coin?

1	2	3	count #heads
			0
H			1
	H		1
		H	1
H	H		2
H		H	2
	H	H	2
H	H	H	3

$P(x = 0) = 1/8 \quad P(x = 2) = 3/8$

$P(x = 1) = 3/8 \quad P(x = 3) = 1/8$



the probability (p-value) to observe a value larger or equal that 2

$$P(x \geq 2|H_0) = P(x = 2) + P(x = 3) = 3/8 + 1/8 = 4/8 = 1/2$$

Hypothesis testing

Hypothetically we sell more juice on hot days

Temperature	Sales
27.0	10
28.9	13
34.5	15
44.1	17
42.4	18
25.3	11
32.9	13
37.5	15
38.1	17
43.4	18
32.6	12
38.2	14
37.5	15
44.1	17
43.4	18

$\mu = 25$

True population mean

Temperature	Sales
44.1	17
42.4	18
37.5	15
38.1	17
43.4	18
38.2	14
37.5	15
44.1	17
43.4	18

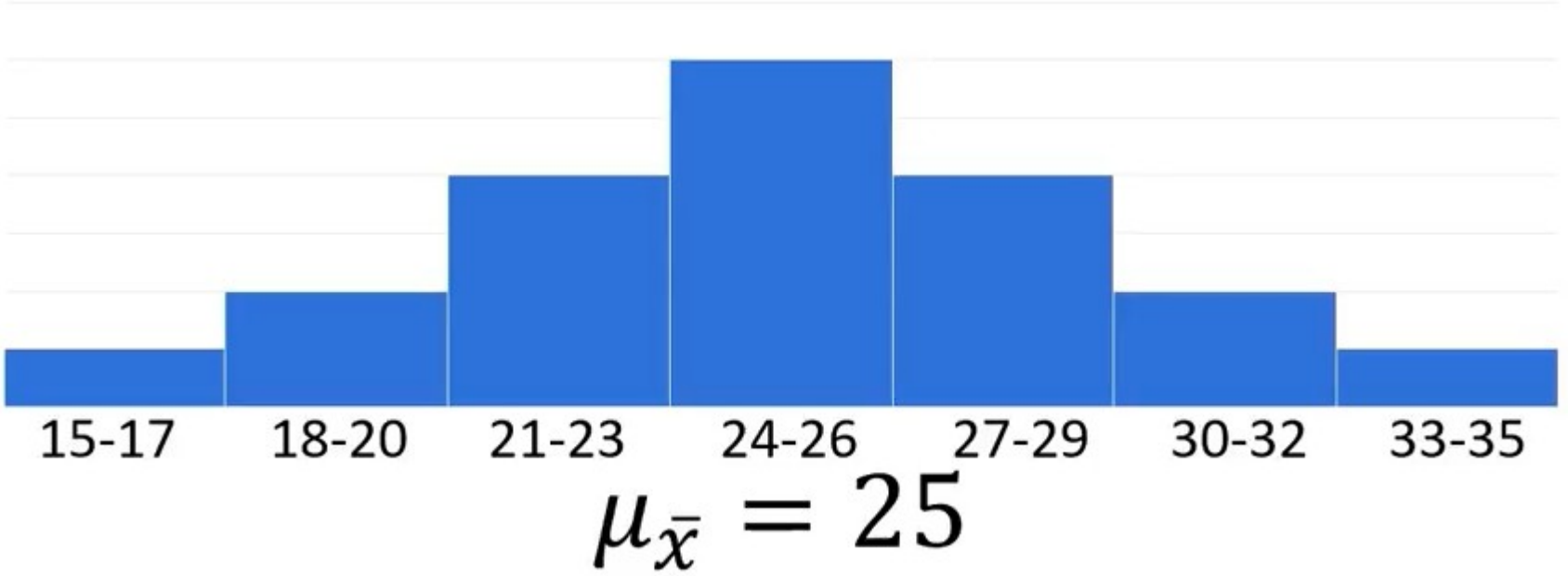
$\bar{x} = 33$

A 30 hot-days sample mean

It seems that our hypothesis is TRUE

Hypothesis testing

Consider a series of completely random 30-days samples:

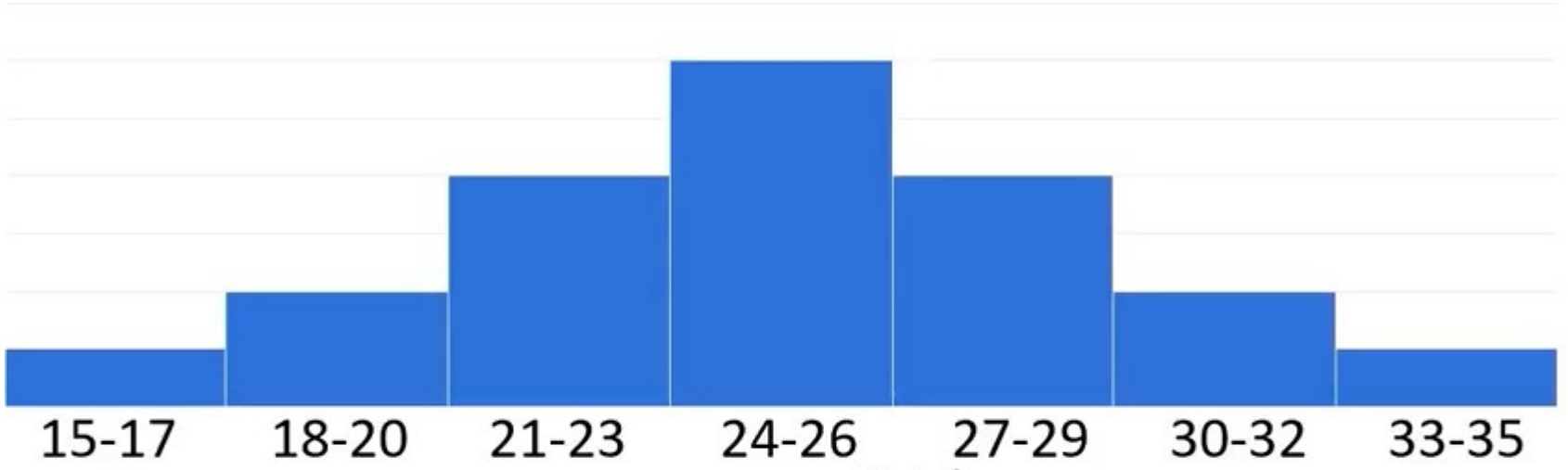


As seen it is probable to have a completely random sample with mean 33.

So how to get sure that our hypothesis is true for any hot-days sample?

We calculate how much it is probable to get
a random sample with the mean of 33

Hypothesis testing

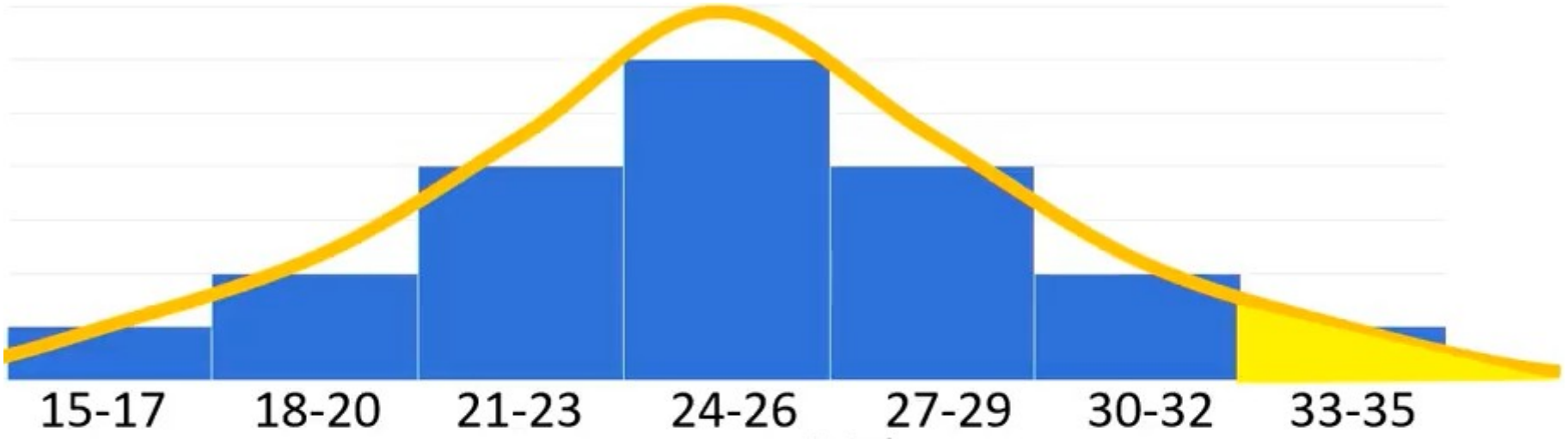


H_0: mean sales for hot days = mean sales for population

H_1: mean sales for hot days > mean sales for population

α: 0.05

Hypothesis testing



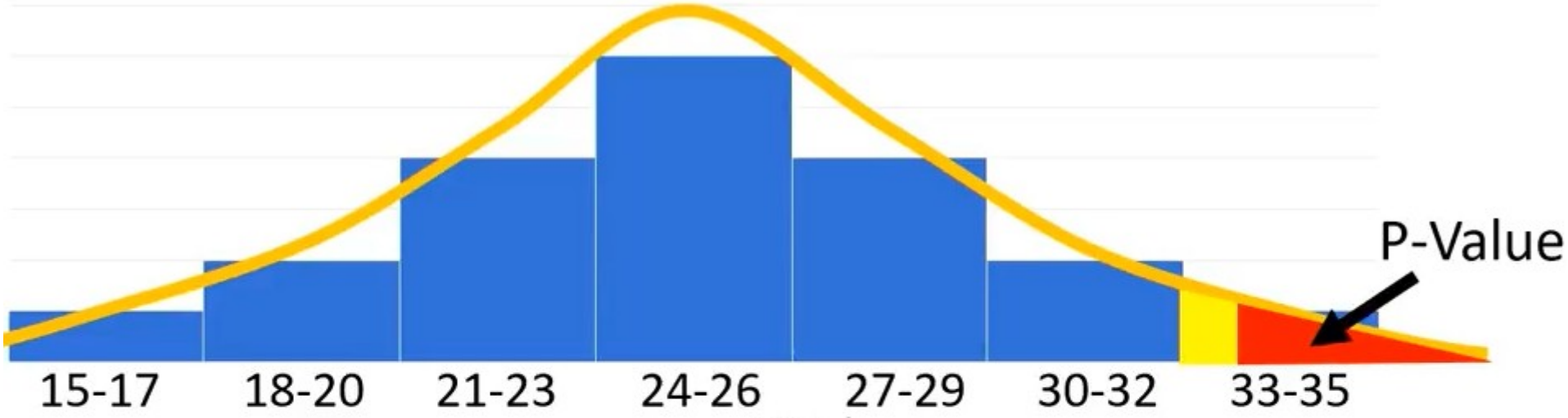
$$\mu_{\bar{x}} = 25$$

H_0: mean sales for hot days = mean sales for population

H_1: mean sales for hot days > mean sales for population

α: 0.05

Z-test (if the population variance is known) t_test (with large sample size~ z-test)



H_0: mean sales for hot days = mean sales for population

H_1: mean sales for hot days > mean sales for population

α: 0.05

t-test

Assume that height is normally distributed: $X \sim \mathcal{N}(\mu, \sigma)$, ie:

$$\text{height}_i = \text{average height over the population} + \text{error}_i$$

$$x_i = \bar{x} + \varepsilon_i$$

The ε_i are called the residuals

2 Fit: estimate the model parameters

$\bar{x}, s_x$ are the estimators of μ, σ.

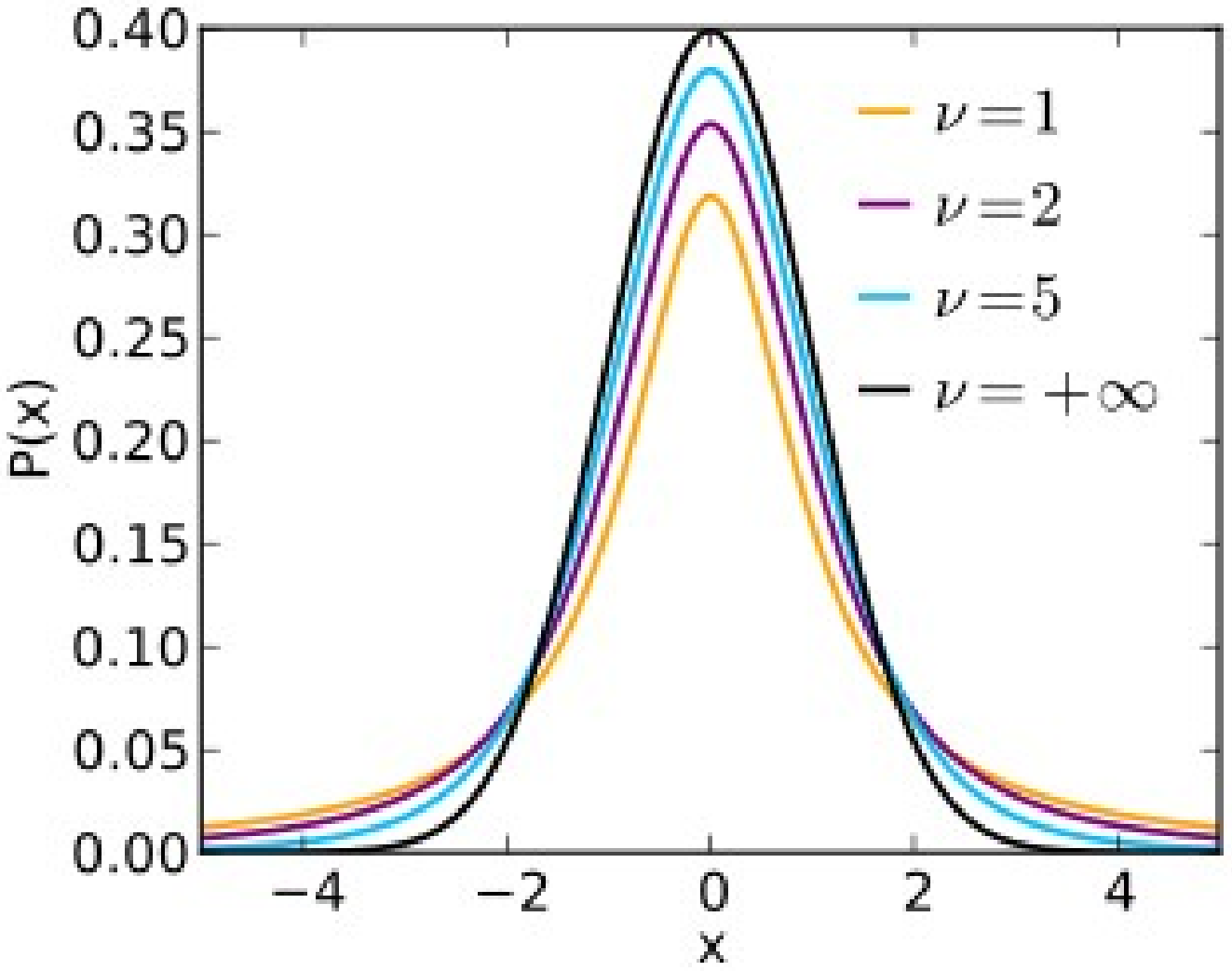
3 Compute a test statistic

In testing the null hypothesis that the population mean is equal to a specified value $\mu_0 = 1.75$

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

T-ditribution

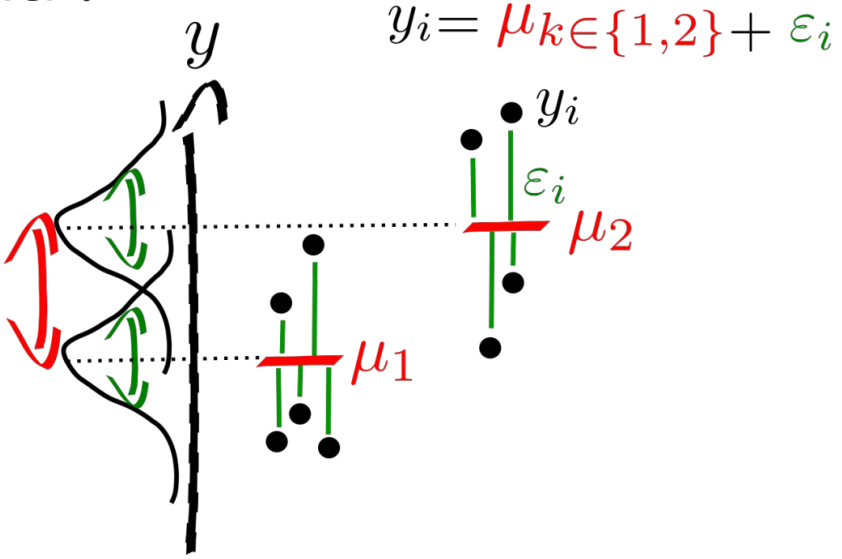
probability density function



Two sample t-test

- The two-sample t-test is used to determine if two population means are equal.

$$y_1 \sim \mathcal{N}(\mu_1, \sigma_1), \quad y_2 \sim \mathcal{N}(\mu_2, \sigma_2).$$



Since y_1 and y_2 are independant:

$$t = \frac{\text{difference of means}}{\text{its standard error}} = \frac{\bar{y_1} - \bar{y_2}}{s_{\bar{y_1}-\bar{y_2}}}$$

$$s^2_{\bar{y_1}-\bar{y_2}} = s^2_{\bar{y_1}} + s^2_{\bar{y_2}} = \frac{s^2_{y_1}}{n_1} + \frac{s^2_{y_2}}{n_2}$$

thus

$$s_{\bar{y_1}-\bar{y_2}} = \sqrt{\frac{s^2_{y_1}}{n_1} + \frac{s^2_{y_2}}{n_2}}$$

Hypothesis testing

Combining sample distribution with probability

- Steps

  1. Model the data

  2. Fit: estimate the model parameters (frequency, mean, correlation, regression coeficient)

  3. Compute a test statistic from model the parameters.

  4. Formulate the null hypothesis: What would be the (distribution of the) test statistic if the observations are the result of pure chance.

  - 5. Compute the probability (p-value) to obtain a larger value for the test statistic by chance (under the null hypothesis)

Chi-squared distribution

The chi-square or χ^2_n distribution with n degrees of freedom (df) is the distribution of a sum of the squares of n independent standard normal random variables $\mathcal{N}(0,1)$

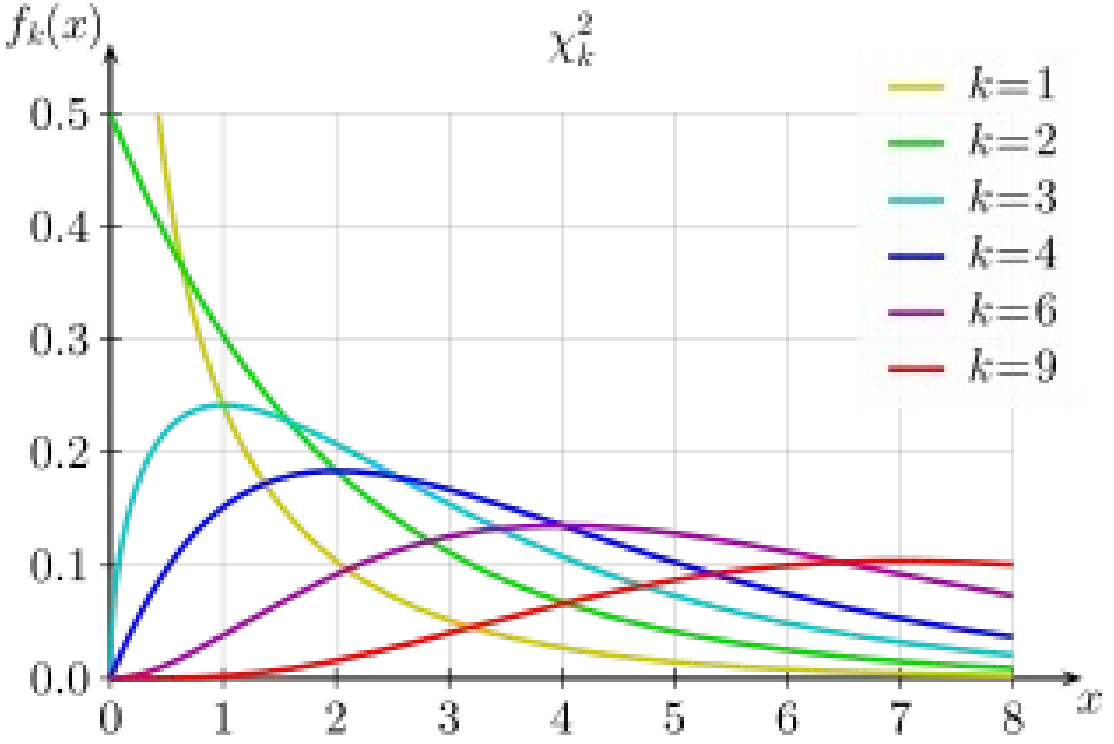
Let $X \sim \mathcal{N}(\mu, \sigma^2)$ then, $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$, then:

- The squared standard $Z^2 \sim \chi^2_1$ (one df).
- **The distribution of sum of squares** of n normal random variables: $\sum_i^n Z_i^2 \sim \chi^2_n$

The χ^2-distribution is used to model the distribution of the sample **variance**.

Chi-squared distribution

probability density function



Fisher distribution

The F-distribution, $F_{n,p}$, with n and p degrees of freedom is the ratio of two independent χ^2 variables. Let $X \sim \chi^2_n$ and $Y \sim \chi^2_p$ then:

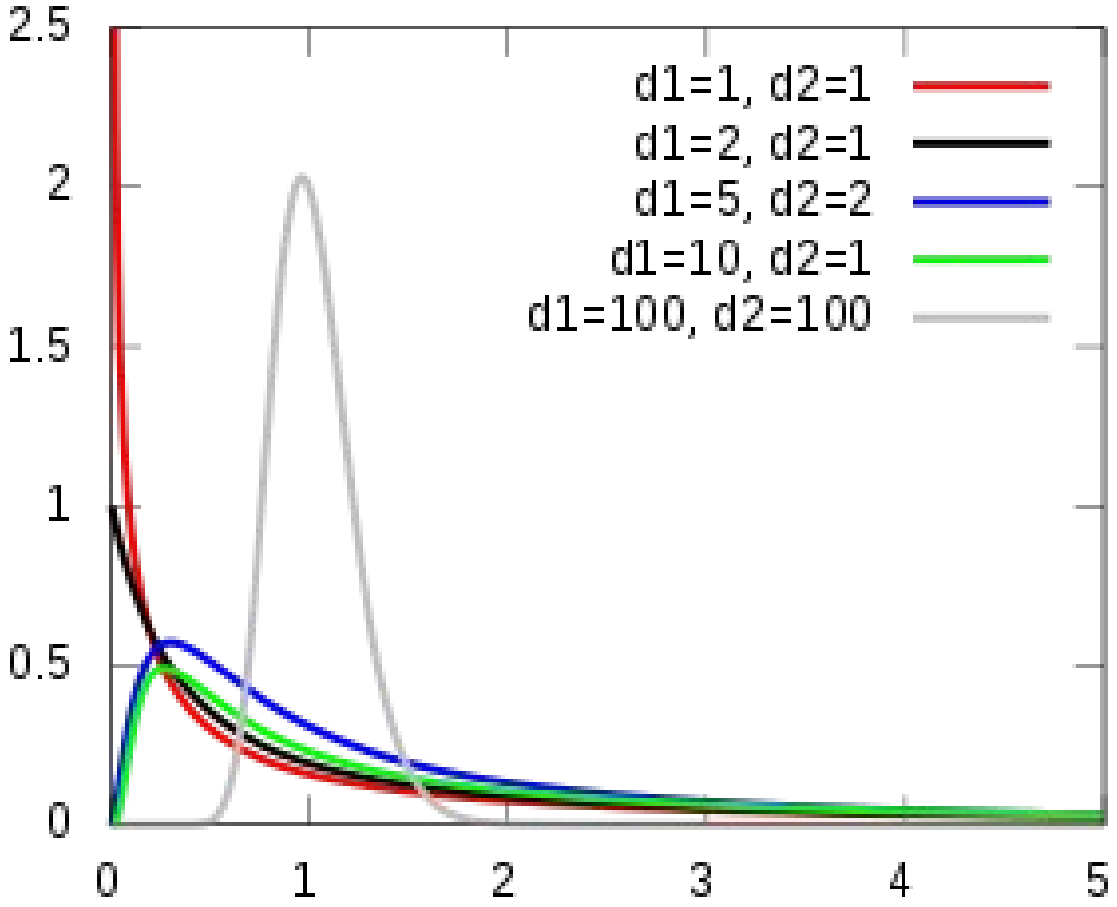
$$F_{n,p} = \frac{X/n}{Y/p}$$

Are two variances equals?

is the ratio or two errors significantly large ?

Fisher distribution

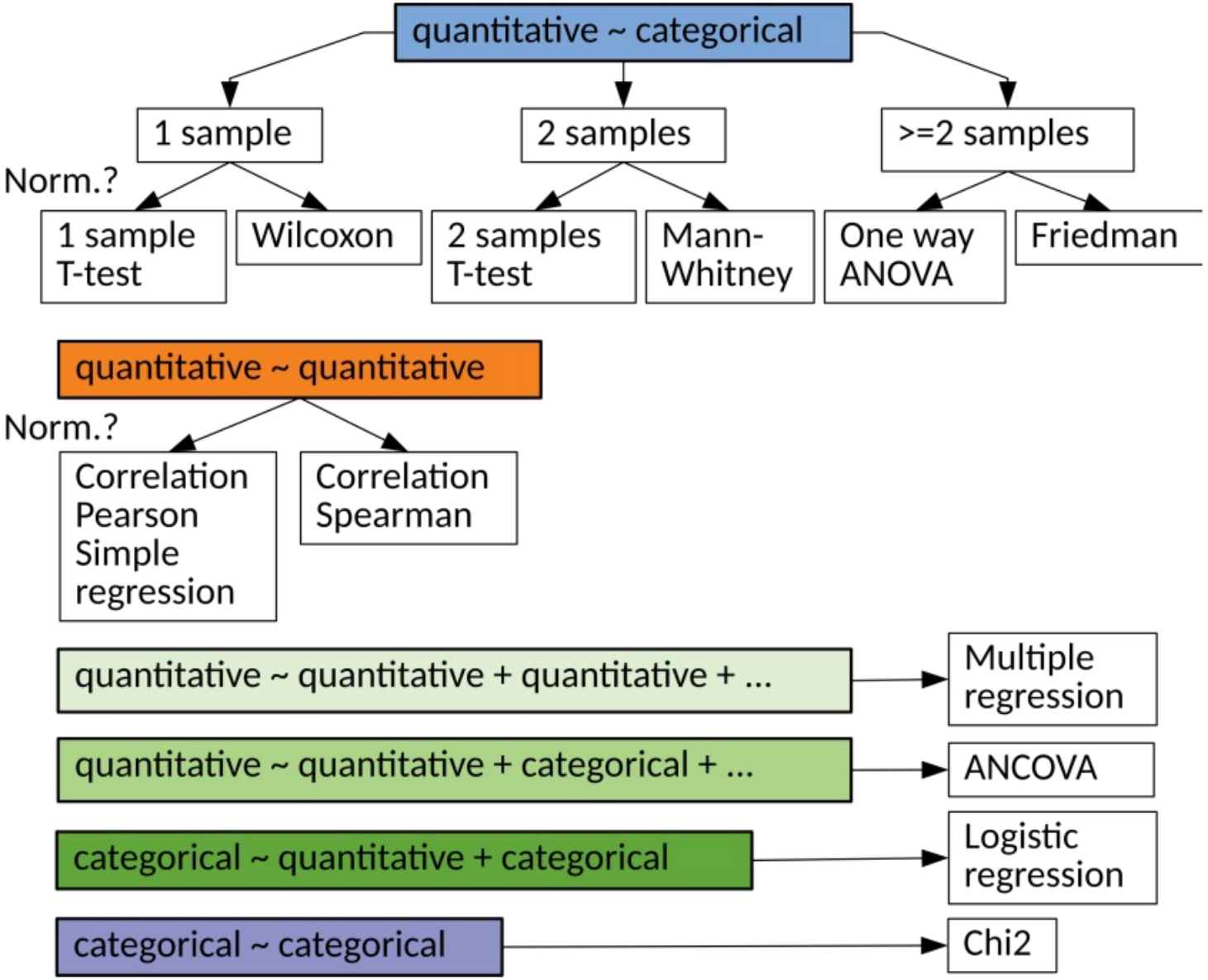
probability density function



Testing pairwise associations

- In statistics, a **categorical variable** or factor is a variable that can take on one of a limited, and usually fixed, number of possible values.
- An **ordinal variable** is a categorical variable with a clear ordering of the levels. For example: drinks per day (none, small, medium and high).
- A **continuous** or **quantitative** variable $x \in R$ is one that can take any value in a range of possible values, possibly infinite. E.g.: salary, experience in years, weight.

What statistical test should I use?



Pearson correlation test

Pearson correlation test tests association between two quantitative variables.

Let x and y two quantitative variables, where n samples were obeserved. The linear correlation coeficient is defined as :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Under H_0, the test statistic $t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$ follow Student distribution with $n-2$ degrees of freedom.

This t statistic has a Student's t-distribution in the null case (correlation r is zero)

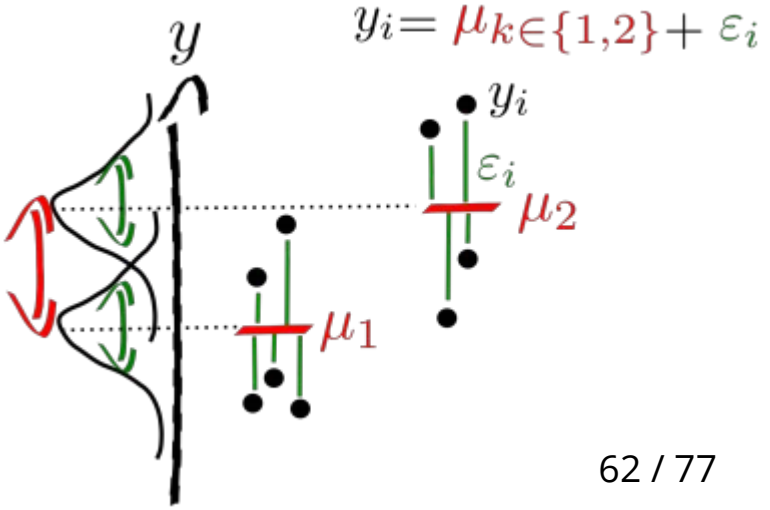
t-test

- **One sample t-test**:
  - The one-sample t-test is used to determine whether a sample comes from a population with a specific mean.

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

- **Two sample (Student) t-test:**
  - compares two means

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_{y1}^2}{n_1} + \frac{s_{y2}^2}{n_2}}}.$$



- **Two sample (Student) t-test:**
  - Equal variances(Equal or unequal sample sizes)

$$t = \frac{\bar{y_1} - \bar{y_2}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

  - Equal sample sizes, equal variances

$$t = \frac{\bar{y_1} - \bar{y_2}}{s\sqrt{2}} \cdot \sqrt{n}$$

Analysis of variance (ANOVA)

- ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups.
  - Test if any group is on average superior, or inferior, to the others versus the null hypothesis that all four strategies yield the same mean response
  - Detect any of several possible differences.
  - The advantage of the ANOVA F-test is that we do not need to pre-specify which strategies are to be compared, and we do not need to adjust for making multiple comparisons.
  - The disadvantage of the ANOVA F-test is that if we reject the null hypothesis, we do not know which strategies can be said to be significantly different from the others.

ANOVA F-test

- Lets assume

$$Y_1 \sim N(\mu_1, \sigma_1), Y_2 \sim N(\mu_2, \sigma_2) \text{ and } Y_3 \sim N(\mu_3, \sigma_3).$$

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}}$$

$$= \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{s_B^2}{s_W^2}.$$

$$s_B^2 = \sum_i n_i(\bar{y}_{i\cdot} - \bar{y})^2/(K-1),$$

$$s_W^2 = \sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2/(N-K),$$

where y_{ij} is the jth observation in the ith out of K groups and N is the overall sample size. This F-statistic follows the F-distribution with $K-1$ and $N-K$ degrees of freedom under the null hypothesis.

Chi-square $\chi 2$ (categorial ~ categorial)

- Used to check the relation between two categorical variables
- Computes the chi-square, $\chi 2$, statistic and p-value for the hypothesis test of **independence of frequencies in the observed contingency table** (cross-table)
- The observed frequencies are tested against an **expected contingency** table obtained by computing expected frequencies based on the marginal sums under the assumption of independence.

Example:

  - 20 participants: 10 exposed to some chemical product and 10 non exposed (exposed= 1 or 0). Among the 20 participants 10 had cancer 10 not (cancer = 1 or 0). $\chi 2$ tests the association between those two variables.

Chi-square test

```
import numpy as np
import pandas as pd
import scipy.stats as stats

# Dataset:
# 15 samples:
# 10 first exposed
exposed = np.array([1] * 10 + [0] * 10)
# 8 first with cancer, 10 without, the last two with.
cancer = np.array([1] * 8 + [0] * 10 + [1] * 2)

crosstab = pd.crosstab(exposed, cancer, rownames=['exposed'],
                       colnames=['cancer'])
print("Observed table:")
print("---------------")
print(crosstab)

chi2, pval, dof, expected = stats.chi2_contingency(crosstab)
print("Statistics:")
print("-----------")
print("Chi2 = %f, pval = %f" % (chi2, pval))
print("Expected table:")
print("---------------")
print(expected)
```

Chi-square test

```
Observed table:
---------------
cancer   0  1
exposed
0        8  2
1        2  8
Statistics:
-----------
Chi2 = 5.000000, pval = 0.025347
Expected table:
---------------
[[5. 5.]
 [5. 5.]]
```

Are the levels of the row variable differentially distributed over levels of the column variables?

Significance in this hypothesis test means that interpretation of the cell frequencies is warranted.

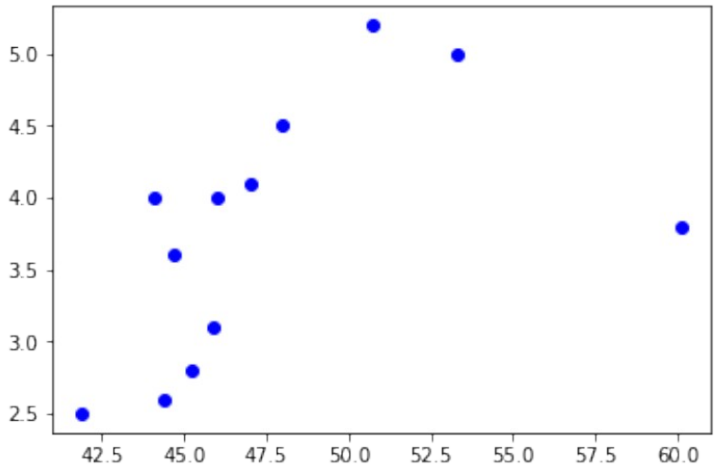
Spearman correlation

Non-parametric test of pairwise associations

- The Spearman correlation is a non-parametric measure of the monotonicity of the relationship between two datasets.
- When to use it? Observe the data distribution: - presence of outliers - the distribution of the residuals is not Gaussian.
- Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation.

```python
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

x = np.array([44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 46, 47, 48, 60.1])
y = np.array([2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8, 4, 4.1, 4.5, 3.8])

plt.plot(x, y, "bo")

# Non-Parametric Spearman
cor, pval = stats.spearmanr(x, y)
print("Non-Parametric Spearman cor test, cor: %.4f, pval: %.4f" % (cor, pval))

# "Parametric Pearson cor test
cor, pval = stats.pearsonr(x, y)
print("Parametric Pearson cor test: cor: %.4f, pval: %.4f" % (cor, pval))
```

```
Non-Parametric Spearman cor test, cor: 0.7110, pval: 0.0095
Parametric Pearson cor test: cor: 0.5263, pval: 0.0788
```



Dependent t-test for paired samples

- This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired".

$$t = \frac{\bar{X}_D - \mu_0}{\frac{s_D}{\sqrt{n}}}.$$

- The average (X_D) and standard deviation (s_D) of those differences are used in the equation.
- The constant μ_0 is zero if we want to test whether the average of the difference is significantly different.
- The degree of freedom used is n − 1, where n represents the number of pairs.

Wilcoxon signed-rank

- The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two **related samples**, **matched samples**, or r**epeated measurements on a single sample** to assess whether their population mean ranks differ (i.e. it is a paired difference test).
- It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed.
- When to use it? Observe the data distribution: the distribution of the data is not Gaussian

Wilcoxon signed-rank

Null hypothesis H_0: difference between the pairs follows a symmetric distribution around zero.

```
import scipy.stats as stats
n = 20
# Buisness Volume time 0
bv0 = np.random.normal(loc=3, scale=.1, size=n)
# Buisness Volume time 1
bv1 = bv0 + 0.1 + np.random.normal(loc=0, scale=.1, size=n)

# create an outlier
bv1[0] -= 10

# Paired t-test
print(stats.ttest_rel(bv0, bv1))

# Wilcoxon
print(stats.wilcoxon(bv0, bv1))
```

```
Ttest_relResult(statistic=0.7821450892478711, pvalue=0.4437681541620575)
WilcoxonResult(statistic=35.0, pvalue=0.008967599455194583)
```

Linear model

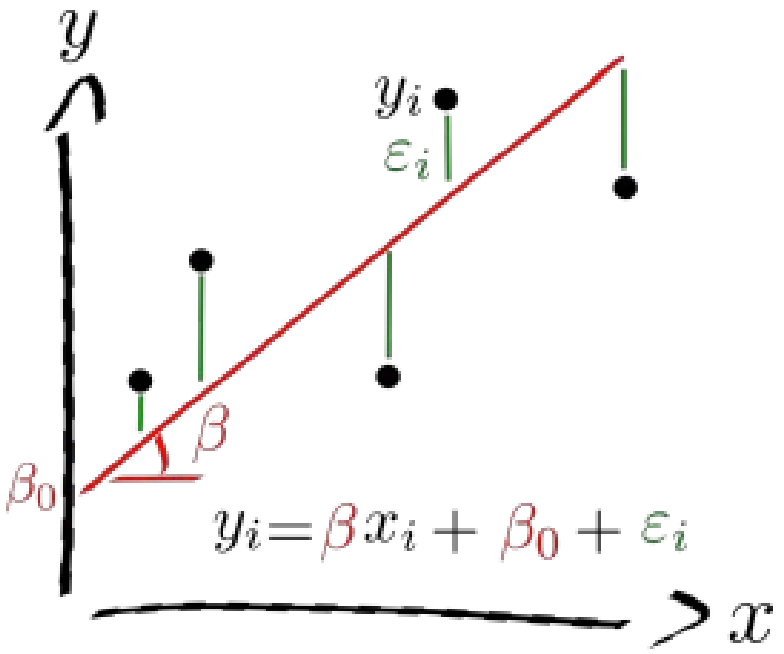
Given n random samples $(y_i, x_{1i}, \ldots, x_{pi})$, $i = 1, \ldots, n$, the linear regression models the relation between the observations y_i and the independent variables x_i^p is formulated as

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \qquad i = 1, \ldots, n$$

- The β's are the model parameters, ie, the regression coeficients.
- β_0 is the intercept or the bias.
- ε_i are the **residuals**.

An independent variable (IV). It is a variable that stands alone and isn't changed by the other variables you are trying to measure.

A dependent variable. It is something that depends on other factors.



Fit the model

The goal it so estimate β, β_0 and σ^2.

Minimizes the **mean squared error (MSE)** or the **Sum squared error (SSE)**. The so-called **Ordinary Least Squares (OLS)** finds β, β_0 that minimizes the $SSE = \sum_i \epsilon_i^2$

$$SSE = \sum_i (y_i - \beta\, x_i - \beta_0)^2$$

Recall from calculus that an extreme point can be found by computing where the derivative is zero, i.e. to find the intercept, we perform the steps:

$$\frac{\partial SSE}{\partial \beta_0} = \sum_i (y_i - \beta\, x_i - \beta_0) = 0$$

$$\sum_i y_i = \beta \sum_i x_i + n\, \beta_0$$

$$n\, \bar{y} = n\, \beta\, \bar{x} + n\, \beta_0$$

$$\beta_0 = \bar{y} - \beta\, \bar{x}$$

To find the regression coefficient, we perform the steps:

$$\frac{\partial SSE}{\partial \beta} = \sum_i x_i(y_i - \beta\, x_i - \beta_0) = 0$$

Plug in β_0:

$$\sum_i x_i(y_i - \beta\, x_i - \bar{y} + \beta\bar{x}) = 0$$

$$\sum_i x_i y_i - \bar{y}\sum_i x_i = \beta \sum_i (x_i - \bar{x})$$

Divide both sides by n:

$$\frac{1}{n}\sum_i x_i y_i - \bar{y}\bar{x} = \frac{1}{n}\beta \sum_i (x_i - \bar{x})$$

$$\beta = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_i (x_i - \bar{x})} = \frac{Cov(x,y)}{Var(x)}.$$

Test

The total sum of squares is the total squared sum of deviations from the mean of y, i.e.

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

The regression sum of squares, also called the explained sum of squares:

$$SS_{\text{reg}} = \sum_i (\hat{y}_i - \bar{y})^2,$$

where $\hat{y}_i = \beta x_i + \beta_0$ is the estimated value of salary $\hat{y}_i$ given a value of experience x_i.

The sum of squares of the residuals, also called the residual sum of squares (RSS) is:

$$SS_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2.$$

R^2 is the explained sum of squares of errors. It is the variance explain by the regression divided by the total variance, i.e.

$$R^2 = \frac{\text{explained SS}}{\text{total SS}} = \frac{SS_{\text{reg}}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}.$$

Test

Let $\hat{\sigma}^2 = SS_{\text{res}}/(n-2)$ be an estimator of the variance of ϵ. The 2 in the denominator stems from the 2 estimated parameters: intercept and coefficient.

- **Unexplained variance**: $\frac{SS_{\text{res}}}{\hat{\sigma}^2} \sim \chi^2_{n-2}$
- **Explained variance**: $\frac{SS_{\text{reg}}}{\hat{\sigma}^2} \sim \chi^2_1$. The single degree of freedom comes from the difference between $\frac{SS_{\text{tot}}}{\hat{\sigma}^2} (\sim \chi^2_{n-1})$ and $\frac{SS_{\text{res}}}{\hat{\sigma}^2} (\sim \chi^2_{n-2})$, i.e. $(n-1)-(n-2)$ degree of freedom.

The Fisher statistics of the ratio of two variances:

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}} = \frac{SS_{\text{reg}}/1}{SS_{\text{res}}/(n-2)} \sim F(1, n-2)$$