

DAI

November 8, 2024

TOC

- 1 Lecture 00-introduction.pdf
- 2 Lecture 01 02 03-descriptiveStatistics.pdf
- 3 Lecture 04-probability.pdf
- 4 Lecture 05-independence.pdf
- 5 Lecture 06-randomVars.pdf
- 6 Lecture 07-randomVars.pdf
- 7 Lecture 08-discreteRVs.pdf
- 8 Lecture 09-discreteRV.pdf

TOC

- 9 Lecture 10-poisson.pdf
- 10 Lecture 11-continuousRVs.pdf
- 11 Lecture 12-gaussian.pdf
- 12 Lecture 13-exponential.pdf
- 13 Lecture 14-MultipleRVs.pdf
- 14 Lecture 15-ParameterEstimation.pdf
- 15 Lecture 16-EvaluatingEstimators.pdf
- 16 Lecture 17-BayesianEstimates.pdf

TOC

- 17 Lecture 18-BayesianEstimatesCont.pdf
- 18 Lecture 19-NonParametricDensity.pdf
- 19 Lecture 20-NonParametricDensity2.pdf
- 20 Lecture 21-LinearRegression.pdf
- 21 Lecture 22-MultipleLinearRegression.pdf
- 22 Lecture 23-kernelRegression.pdf
- 23 Lecture 24-TimeSeriesAnalysis1.pdf
- 24 Lecture 25-TimeSeriesAnalysis2.pdf

TOC

- 25 Lecture 26-TimeSeriesAnalysis3.pdf
- 26 Lecture 27-MVA1.pdf
- 27 Lecture 28-MVA2.pdf
- 28 Lecture 29-MVA3.pdf
- 29 Lecture 30-MVA4.pdf
- 30 Lecture 31-MVA5.pdf
- 31 Lecture 32-MVA6.pdf
- 32 Lecture 33-34-Hypothesis Testing.pdf

TOC

33 Lecture 35-NonParametricHypothesisTesting.pdf

34 Lecture 36-Robust statistics.pdf

Lecture 00-introduction.pdf

CS 215: Data Interpretation and Analysis

Sunita Sarawagi
Autumn 2024

Welcome!

What is the course about?

- Suppose you want to find reliable answers to questions:
 - Which minor should I opt for?
 - What are the types of future careers that IITB graduates favor lately?
 - How many students seats should IITB allocate to each department?
 - Which products are likely to be in high demand next month?
 - Is rainfall in Mumbai becoming more erratic lately?
 - Is inflation increasing at a faster pace in recent times?
 - How is supply of drinking water keeping pace with increasing population?
 - Is a flu vaccine useful to prevent frequent cold&fever?

How do you find the answers?

- Go by your existing vague impressions
- Ask your peer group, Ask older experienced people
- Do a websearch
- Ask ChatGPT
- ...

The data scientists approach

- Go to an authentic source that has recorded correctly the observed values over time → This is your data
 - Public data: World bank datasets, Datacommons, National Data Analytics platform, Stock prices
 - Enterprise data: Student data in universities, sales and customer interaction data in enterprises
 - Scientific data: experiments, simulations and observations in lab
- Try to find answers from the data → How?
 - This course will teach you how to get answers to top-level questions from data in a scientific way.

Several sources of public data in India



Image from data.gov.in

Some example studies on Indian Data

- Power consumption in India
- Health of Indian population
- Housing in India

Course contents

- Data analysis: gathering, summarizing, and visualizing data in intuitive ways
- Probability: Mathematical tool to represent uncertainty
- Statistical inference: Drawing probabilistic conclusions from limited data

Important pre-requisite for future courses in machine learning, image processing, computer vision, deep learning, AI, finance, etc..

Mode of running the course

- Three 55 minute slots per week:
- SAFE/Moodle/paper quizzes on the material covered in **prior** weeks
 - 20 minute duration at a pre-announced time or 55 minute quiz.
 - Grading will be done on top n-2 out of n quizzes for 20 minute quizzes.
 - No compensation for missed quizzes.
- All materials will be uploaded on Moodle, announcements via Moodle, questions on Moodle or cs215-ta@googlegroups.com
- [Course webpage](#)

Evaluation

Approximate credit structure

- 15% In-class Quizzes
- 25% Mid-semester exam
- 35% End semester exam
- 25% Programming and written homeworks: in teams (about 5 assignments)
- Attendance mandatory. Students with less than 80% may get a DX.

We will all adhere to principles of academic honesty. Penalties for violation will be severe and will be reported to DADAC. Givers and takers are equally responsible.

Lecture 01 02 03-descriptiveStatistics.pdf

Descriptive Statistics

Fall 2024

Instructor:

Sunita Sarawagi

Thanks to Prof Ajit Rajawade for the initial version of the slides

Terminology

- **Population:** The collection of all elements which we wish to study, example: data about occurrence of tuberculosis all over the world
- In this case, “population” refers to the set of people in the entire world.
- The population is often too large to examine/study.
- So we study a subset of the population – called as a **sample**.
- In an experiment, we basically collect **values** for one or more **attributes or variables** of each member of the sample.

Examples of samples

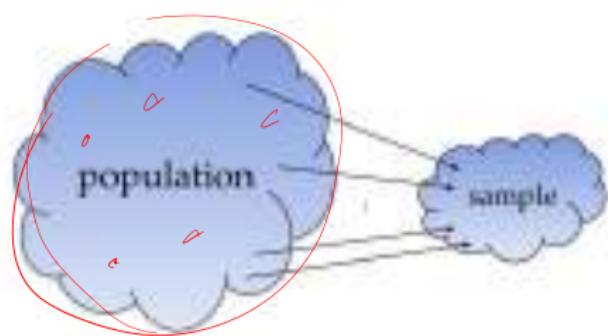
attribute

sample

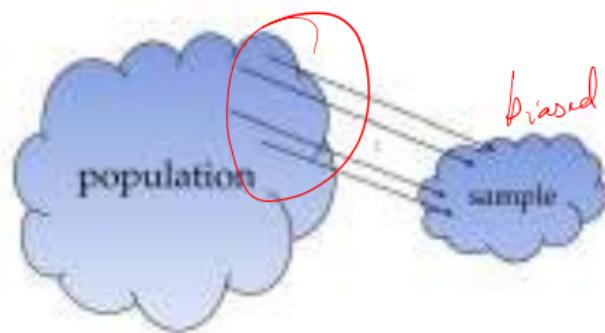
index	username	country	age	ezlvl	time	points	finished	observation
0	mary	us	38	0	124.94	418	0	
1	jane	ca	21	0	331.64	1149	1	
2	emil	fr	52	1	324.61	1321	1	
3	ivan	ca	50	1	39.51	226	0	
4	hasan	tr	26	1	253.19	815	0	
5	jordan	us	45	0	28.49	206	0	
6	sanjay	ca	27	1	585.88	2344	1	
7	lena	uk	23	0	408.76	1745	1	
8	shuo	cn	24	1	194.77	1043	0	
9	robyn	us	59	0	255.55	1102	0	
10	anna	pl	18	0	303.66	1209	1	
11	jono	bg	22	1	381.97	1491	1	

Table 1.1: A data table that contains observations of seven variables for 12 players of a computer game. Each row in this table corresponds to one player. Each column corresponds to one characteristic that was measured for all the players.

Population and Samples



(a) Representative sample selection



(b) Biased sample selection

Data Representation and Visualization

Need for data visualization

- The raw dataset or tables may be too large. Cannot make sense of the data just by inspecting raw table of numbers.
- Even if data is not too large, patterns emerge sometimes only under right type of visualization.

Outline

- Visualizing values of each variable separately
- Visualizing pairs of variables.
- Multi-dimensional data

Terminology

- **Discrete data:** Data whose values are restricted to a finite or countably infinite set. Eg: letter grades at IITB, genders, marital status (single, married, divorced), income brackets in India for tax purposes
- **Continuous data:** Data whose values belong to an uncountably infinite set (Eg: a person's height, temperature of a place, speed of a car at a time instant).

Raw data

- Example: Country of winners of any competition
- Example: Grades of students in CS 215

21 AA

25 AB

01 AP

09 BB

11 BC

02 DX

03 CC

24 AA

:

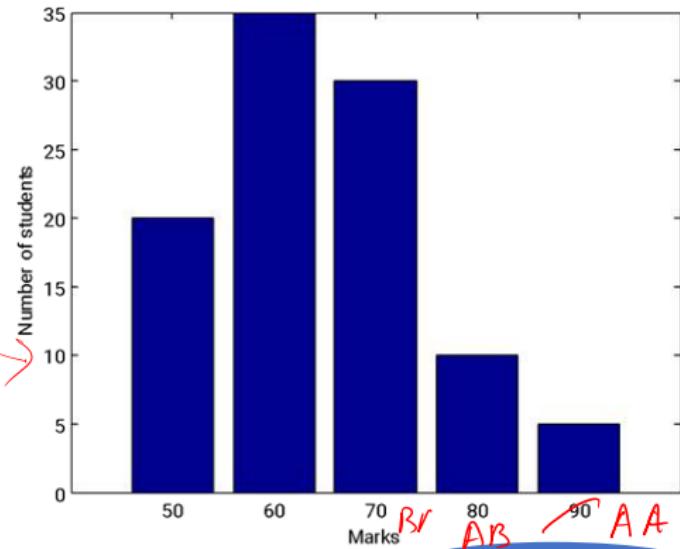
:

Frequency Tables

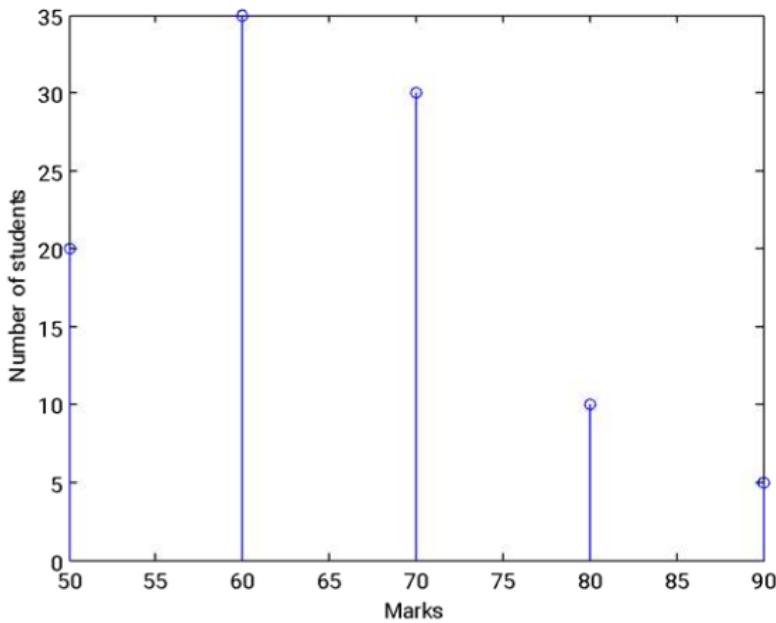
- The frequency table can be visualized using a **line graph** or a **bar graph** or a **frequency polygon**.

Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20

A **bar graph** plots the distinct data values on the X axis and their frequency on the Y axis by means of the height of a thick vertical bar!

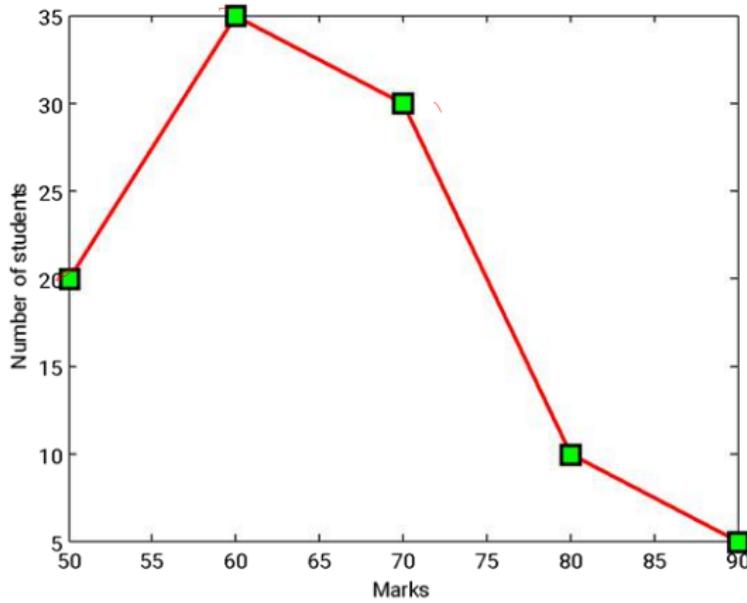


Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20



A line diagram plots the distinct data values on the X axis and their frequency on the Y axis by means of the height of a vertical line!

Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20



A **frequency polygon** plots the frequency of each data value on the Y axis, and connects consecutive plotted points by means of a line.

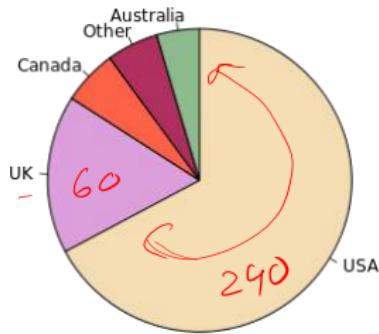
Relative frequency tables

- Sometimes the actual frequencies are not important.
- We may be interested only in the *percentage* or *fraction* of those frequencies for each data value – i.e. *relative frequencies*.

Grade	Fraction of number of students
AA	0.05
AB	0.10
BB	0.30
BC	0.35
CC	0.20

Pie charts

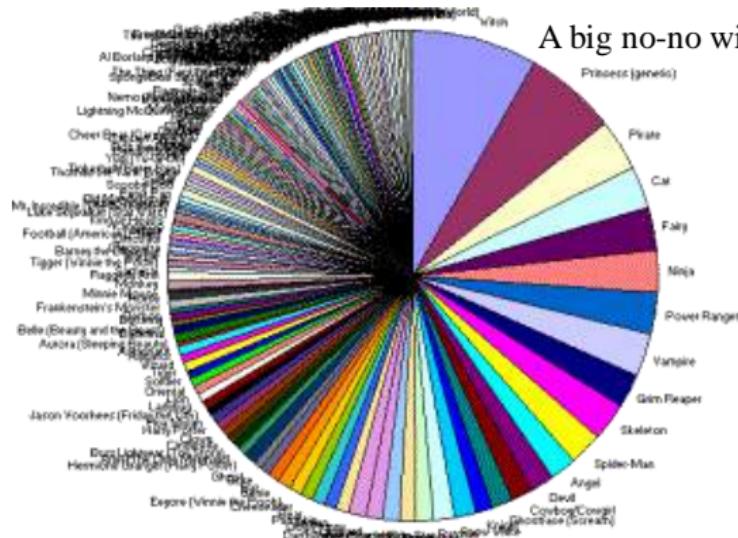
- For a small number of distinct data values which are non-numerical, one can use a **pie-chart** (it can also be used for numerical values).
- It consists of a circle divided into sectors corresponding to each data value.
- The area of each sector = relative frequency for that data value.



Population of native English speakers:
https://en.wikipedia.org/wiki/Pie_chart

$$\frac{240}{360} \approx \frac{2}{3}$$

Pie charts can be confusing



A big no-no with too many categories.

<http://stephenturbek.com/articles/2009/06/better-charts-from-simple-questions.html>

Dealing with continuous data

- Example: temperature of a place at a time instant, speed of a car at a given time instant, weight or height of an animal, etc.
- The raw data: marks in final exams.

Table 2.3 Life in Hours of 200 Incandescent Lamps.

Item Lifetimes										
1067	919	1196	785	1126	936	918	1156	920	948	
855	1092	1162	1170	929	950	905	972	1035	1045	
1157	1195	1195	1340	1122	938	970	1237	956	1102	
1022	978	832	1009	1157	1151	1009	765	958	902	
923	1333	811	1217	1085	896	958	1311	1037	702	
521	933	928	1153	946	858	1071	1069	830	1063	
930	807	954	1063	1002	909	1077	1021	1062	1157	
999	932	1035	944	1049	940	1122	1115	833	1320	
901	1324	818	1250	1203	1078	890	1303	1011	1102	
996	780	900	1106	704	621	854	1178	1138	951	
1187	1067	1118	1037	958	760	1101	949	992	966	
824	653	980	935	878	934	910	1058	730	980	
844	814	1103	1000	788	1143	935	1069	1170	1067	
1037	1151	863	990	1035	1112	931	970	932	904	
1026	1147	883	867	990	1258	1192	922	1150	1091	
1039	1083	1040	1289	699	1083	880	1029	658	912	
1023	984	856	924	801	1122	1292	1116	880	1173	
1134	932	938	1078	1180	1106	1184	954	824	529	
998	996	1133	765	775	1105	1081	1171	705	1425	
610	916	1001	895	709	860	1110	1149	972	1002	

Visualizing numerical data

- Reduce to a known problem
 - Group into bins/intervals
 - Count number in each bin.
 - Draw histogram

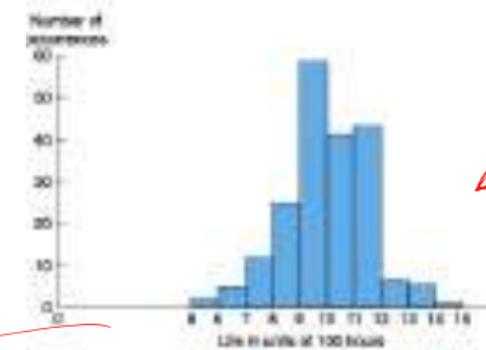


Table 2.4 A Class Frequency Table.

Class Interval	Frequency (Number of Data Values in the Interval)
500-600	2
600-700	5
700-800	12
800-900	25
900-1000	58
1000-1100	41
1100-1200	40
1200-1300	7
1300-1400	6
1400-1500	1

Dealing with continuous data

$x_1, x_2, x_3, \dots, x_N$

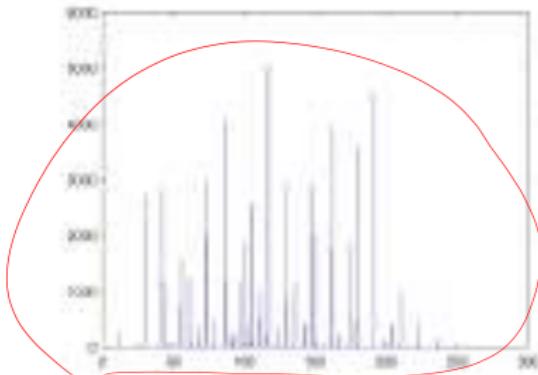
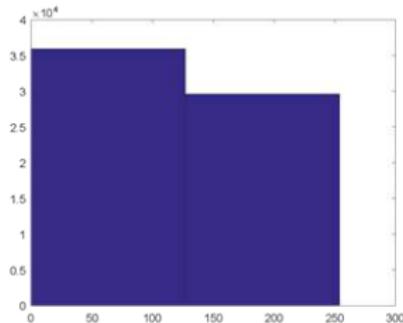
- Let the sample points be $\{x_i\}$, $1 \leq i \leq N$.
- Let there be some K ($K \ll N$) bins, where the j^{th} bin has interval $[a_j, b_j]$.
- Thus frequency f_j for the j^{th} bin is defined as follows:

$$f_j = |\{x_i : a_j \leq x_i < b_j, 1 \leq i \leq N\}|$$

- Such frequency tables are also called **histograms** and they can also be used to store relative frequency instead of frequency.

The histogram binning problem

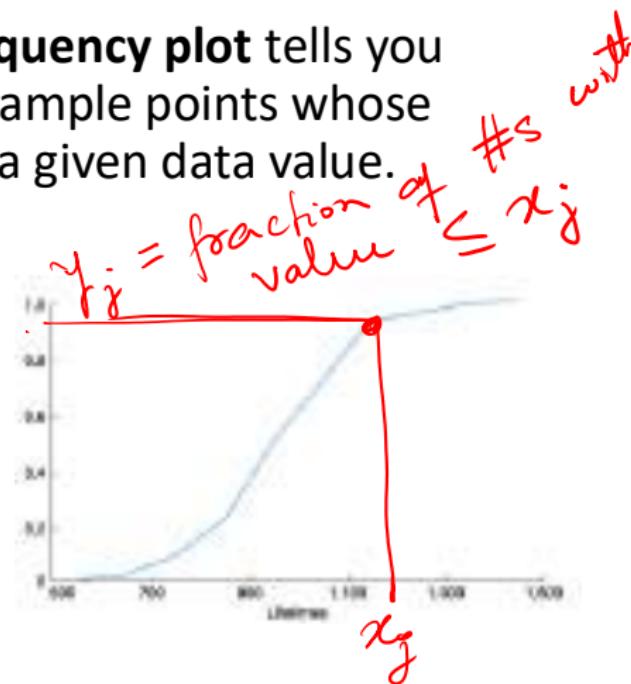
- If you have too few bins (each bin is very wide), there is very little idea you get about the data distribution from the histogram.
- Extreme: only one bin to represent whole data
- If you have many bins (all will be narrow), then there are very points falling into each bin. Again there is very little idea you get about the data distribution from the histogram.
- Extreme: One bin for each distinct value



Cumulative frequency plot

The **cumulative (relative) frequency plot** tells you the (proportion) number of sample points whose value is *less than or equal* to a given data value.

Class Interval	Frequency (Number of Data Values in the Interval)
600–600	2
600–700	5
700–800	12
800–900	25
900–1000	58 /200
1000–1100	41
1100–1200	40
1200–1300	7
1300–1400	6
1400–1500	1



Summarizing Data

98	92	21	37	56	43	20	45	48	73	58	42	27	58	51	22	38	71	31	33
98	48	89	42	27	79	29	37	66	77	47	42	48	92	48	47	79	99	42	19
52	99	31	73	66	79	54	23	89	71	40	47	52	88	20	21	94	12	34	99
50	76	85	21	19	80	21	48	74	24	68	88	10	37	88	71	57	94	98	98
44	21	48	72	32	87	94	38	84	42	48	48	32	92	42	21	36	22	44	93
79	47	52	33	36	86	45	55	89	70	89	51	78	56	88	42	28	17	21	79
52	91	32	29	94	23	37	17	48	81	42	47	58	98	75	46	18	38	81	77
47	29	21	44	12	86	54	28	48	82	89	38	32	38	42	41	48	99	44	21
58	51	31	22	38	71	88	42	77	75	38	22	38	74	58	51	29	51	21	77
24	85	23	39	78	20	76	96	28	88	88	20	40	23	87	26	32	23	38	86
78	27	58	19	28	79	93	47	38	93	93	39	91	28	13	79	95	94	58	58
18	49	28	42	88	84	32	47	48	88	88	28	32	37	49	28	88	28	88	47
44	94	60	48	98	99	97	98	94	99	99	91	91	99	91	97	49	99	37	48
28	81	61	48	10	84	87	99	28	71	81	23	28	21	77	14	29	23	49	49
29	50	58	38	97	88	98	18	97	91	72	14	26	46	73	38	21	95	49	49
48	94	49	97	57	40	29	72	28	88	87	48	88	22	32	49	38	20	49	49
54	92	58	79	36	21	49	14	24	89	71	18	18	48	42	42	74	38	28	28
23	89	28	42	72	20	20	79	28	88	99	28	81	81	82	78	19	28	28	28
23	79	41	38	78	34	85	54	74	81	44	71	48	88	81	18	23	77	21	38

Summarizing a sample-set

- There are some values that can be considered “representative” of the entire sample-set. Such values are called as a “statistic”.
- The most common statistic is the sample (arithmetic) mean:

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

- It is basically what is commonly regarded as “average value”.

Summarizing a sample-set

$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_N$$

- Another common statistic is the sample median, which is the “middle value”.

- We sort the data array \mathbf{A} from smallest to largest. If N is odd, then the median is the value at the $(N+1)/2$ position in the sorted array.

$$x_1 \leq x_2 \leq \dots \leq x_{\frac{N}{2}+1} \geq x_{\frac{N}{2}+2} \geq \dots \geq x_N$$

- If N is even, the median can take any value in the interval $(A[N/2], A[N/2+1])$ – why?

3, 4, 10, 11, 13, 15
 $N = 6$
any value in-between 10 & 11.

Properties of the mean and median

$$1, 2, 4, 5, 7 \quad a=10, b=1$$

- Consider each sample point x_i were replaced by $\underline{ax_i} + b$ for some constants a and b .

$$11, 21, 41, 51, 71$$

- What happens to the mean? What happens to the median?

$$a\bar{x} + b$$

- Consider each sample point x_i were replaced by its square.
- What happens to the mean? What happens to the median?

Properties of the mean and median

- **Question:** Consider a set of sample points x_1, x_2, \dots, x_N . For what value y , is the sum total of the **squared** difference with every sample point, the least? That is, what is:

$$\arg \min_y \sum_{i=1}^N (y - x_i)^2 ?$$

$F(y)$ Total squared deviation
(or total squared loss) $\min F(y)$ Answer: mean

$\frac{\partial F}{\partial y} = 0$

- **Question:** For what value y , is the sum total of the **absolute** difference with every sample point, the least? That is, what is:

$$\arg \min_y \sum_{i=1}^N |y - x_i| ?$$

Total absolute deviation
(or total absolute loss) Answer: median

Proof that mean minimizes square deviation

$$\min_y F(y) = \min_{y \in \mathbb{R}} \sum_{i=1}^N (x_i - y)^2$$

$$\frac{\partial F}{\partial y} = 0 ; \sum_{i=1}^N 2(x_i - y) = 0$$

$$\Rightarrow y = \frac{\sum_{i=1}^N x_i}{N} = \bar{x}$$

average or mean.

Proof that median minimize absolute deviation

$$\min_y \sum_{i=1}^N |x_i - y| \quad G(y)$$

$$\begin{aligned}\frac{\partial_s G_i}{\partial y} &= -1 \cdot \text{if } x_i - y < 0 \\ &= +1 \cdot \text{if } x_i - y \geq 0 \\ &\equiv \text{sign}(x_i - y)\end{aligned}$$

$$\begin{aligned}\frac{\partial_s G}{\partial y} = 0 \Rightarrow \sum_{i=1}^N \text{sign}(x_i - y) &= 0 && N \text{ is even} \\ \Rightarrow \text{equal } \# \text{ of } +1 \text{ & } -1 &\Rightarrow y \text{ is median}\end{aligned}$$

$x_1, x_2, x_3, \dots, x_N$ - N is odd.

① $\min_y |x_1 - y| + |x_N - y| \quad x_1 \leq y \leq x_N$

② $\min_y |x_2 - y| + |x_{N-1} - y| \quad x_2 \leq y \leq x_{N-1}$

⋮

③ $\min_y |x_{\frac{N}{2}-1} - y| + |x_{\frac{N}{2}+1} - y|$

$x_{\frac{N}{2}-1} \leq y \leq x_{\frac{N}{2}+1}$

$\min_y |x_{\frac{N}{2}} - y| = 0 \text{ if } y = x_{\frac{N}{2}}$
and all above constraints are satisfied

Properties of the mean and median

- The mean need not be a member of the original sample-set.
- The median is always a member of the original sample-set if N is odd.
- The median is not unique and will not be a member of the set if N is even.

Properties of the mean and median

- Consider a set of sample points x_1, x_2, \dots, x_N . Let us say that some of these values get grossly corrupted.
- What happens to the mean?
- What happens to the median?

Example

- Let $A = \{1, 2, 3, 4, 6\}$
- Mean (A) = 3.2, median (A) = 3
- Now consider $A = \{1, 2, 3, 4, 20\}$
- Mean (A) = 6, median(A) = 3.

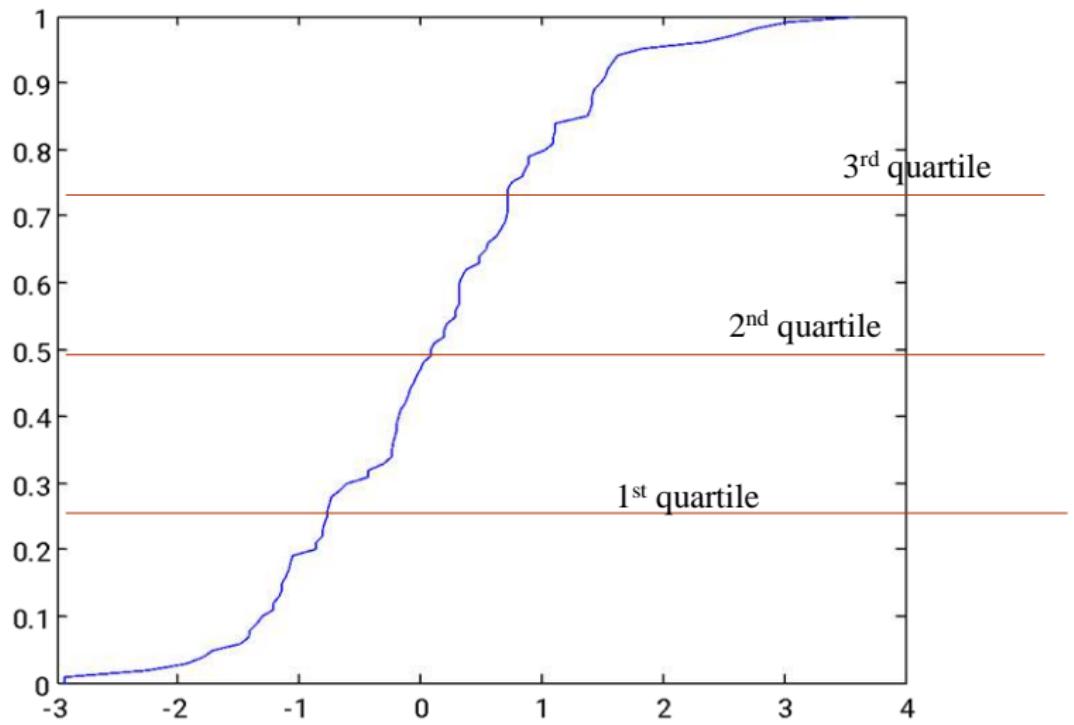
Robust statistics -

Percentiles

- The sample $100p$ percentile ($0 \leq p \leq 1$) is defined as the data value y such that $100p\%$ of the data have a value less than or equal to y , and $100(1-p)\%$ of the data have a larger value.
- For a data set with n sample points, the sample $100p$ percentile is that value such that at least np of the values are less than or equal to it. And at least $n(1-p)$ of the values are greater than it.

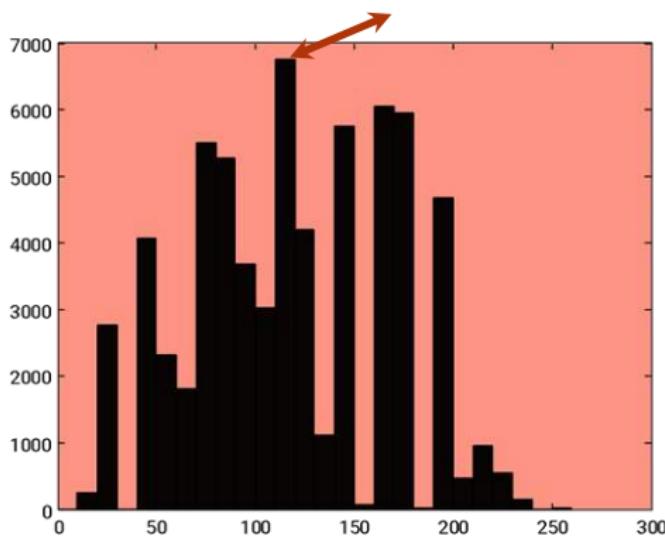
Quantiles

- The sample 25 percentile = first quartile.
- The sample 50 percentile = second quartile.
- The sample 75 percentile = third quartile.
- Quantiles can be inferred from the cumulative relative frequency plot (how?).
- Or by sorting the data values (how?).



Mode

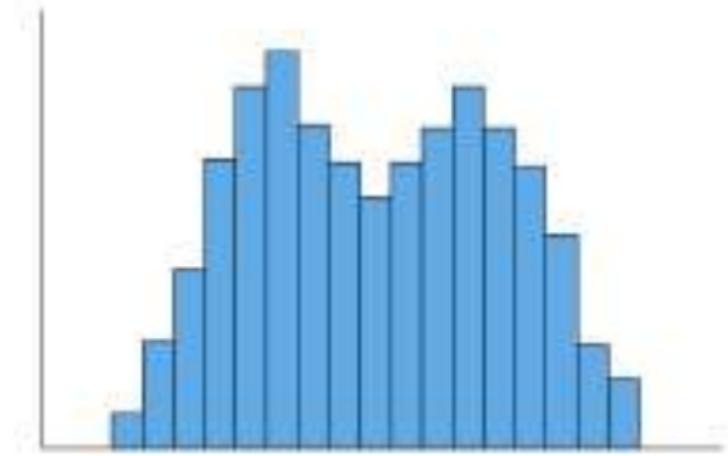
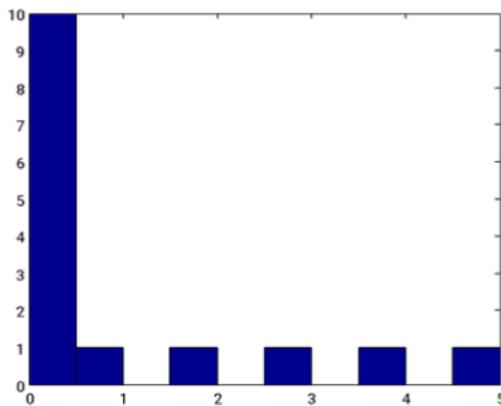
The value that occurs with the highest frequency is called the mode.



Mode

The mode may not be unique, in which case all the highest frequency values are called **modal values**.

Mode at 0



Variance and Standard deviation

- The **variance** is (approximately) the average value of the squared distance between the sample points and the sample mean. The formula is:

$$\text{variance} = s^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x} - x_i)^2$$

The division by $N-1$ instead of N is for a very technical reason which we will understand after many lectures. As such, the variance is computed usually when N is large so the numerical difference is not much.

- The variance measures the “spread of the data around the sample mean”.
- Its positive square-root is called as the **standard deviation**.

Variance and Standard deviation: Properties

Consider each sample point x_i were replaced by $\underline{ax_i + b}$ for some constants a and b . What happens to the standard deviation?

Variance is scaled by \hat{a}

Chebyshev's inequality

- Suppose you know the average marks for this course was 75 (out of 100). And that the variance of the marks was 25.
- Can you say something about how many students secured marks from 65 to 85?
- You obviously cannot predict the exact number – but you can say **something** about this number.
- That something is given by Chebyshev's inequality.

Chebyshev's inequality: and Chebyshev



https://en.wikipedia.org/wiki/Pafnuty_Chebyshev

Russian mathematician:
Stellar contributions in probability and statistics,
geometry, mechanics

Two-sided Chebyshev's inequality:

The proportion of sample points k or more than k ($k > 0$)
standard deviations away from the sample mean is less
than $1/k^2$.

Chebyshev's inequality: and Chebyshev

Two-sided Chebyshev's inequality:

The proportion of sample points k or more than k ($k > 0$) standard deviations away from the sample mean is less than or equal to $1/k^2$.

$$\begin{aligned} S_k &= \{x_i : |x_i - \bar{x}| \geq k\sigma\} \\ \frac{|S_k|}{N} &< \frac{1}{k^2} \end{aligned}$$

Chebyshev's inequality

- Applying this inequality to the previous problem, we see that the fraction of students who got less than 65 or more than 85 marks is as follows:

$$\begin{aligned} S_k &= \{x_i : |x_i - \bar{x}| \geq k\sigma\} \quad \bar{x} = 75 \\ \frac{|S_k|}{N} &\leq \frac{1}{k^2} \quad \sigma = 5 \\ \text{Circled } k &= 2 \\ \rightarrow \frac{|S_k|}{N} &\leq \frac{1}{4} \end{aligned}$$

- So the fraction of students who got from 65 to 85 is more than $1 - 0.25 = 0.75$.

Chebyshev's inequality

91

83

1	Kerala	93.91
2	Lakshadweep	92.28
3	Mizoram	91.58
4	Tripura	87.75
5	Goa	87.40
6	Daman & Diu	87.07
7	Puducherry	86.55
8	Chandigarh	86.43
9	Delhi	86.34
10	Andaman & Nicobar Islands	86.27
11	Himachal Pradesh	83.78
12	Maharashtra	82.91

Mean = 87.69

Std. dev. = 3.306

Fraction of states with literacy rate in the range

$(\mu - 1.5\sigma, \mu + 1.5\sigma)$ is $11/12 \approx 91\%$

As predicted by Chebyshev's inequality, it is **at least**

$$1 - 1/(1.5^2) \approx 0.55$$

$$1 - \frac{1}{k^2} \approx 0.55$$

The bounds predicted by this inequality are loose – but they are correct!

https://en.wikipedia.org/wiki/India_n_states_ranking_by_literacy_rate

Proof of Chebyshev's inequality

$$(N-1)\sigma^2 = \sum_{i=1}^N (x_i - \bar{x})^2 \quad S_k = \{x_i | (x_i - \bar{x}) > k\sigma\}$$

$$= \sum_{i \in S_k} (x_i - \bar{x})^2 + \sum_{i \notin S_k} (x_i - \bar{x})^2$$

$$\geq \sum_{i \in S_k} (x_i - \bar{x})^2 \quad \checkmark$$

$$\geq |S_k| k^2 \sigma^2 \quad \checkmark$$

$$\frac{|S_k|}{N} \leq \frac{(N-1)}{N k^2} \leq \frac{1}{k^2}$$

QED

One-sided Chebyshev's inequality

- Also called the Chebyshev-Cantelli inequality.

The proportion of sample points k or more than k ($k > 0$) standard deviations away from the sample mean **and greater than the sample mean** is less than or equal to $1/(1+k^2)$.



Notice: no absolute value!

$$S_k = \{x_i : x_i - \bar{x} \geq k\sigma\}$$

$$\frac{|S_k|}{N} \leq \frac{1}{1+k^2}$$

One-sided Chebyshev's inequality (Another form)

- Also called the Chebyshev-Cantelli inequality.

The proportion of sample points k or more than k ($k > 0$) standard deviations away from the sample mean **and less than the sample mean** is less than or equal to $1/(1+k^2)$.



Notice: no absolute value!

$$S_k = \{x_i : x_i - \bar{x} \leq -k\sigma\}$$

$$\frac{|S_k|}{N} \leq \frac{1}{1+k^2}$$

Hard work ✓

Constant · ← Intelligence
Luck

Perseverance ✓
Money .

Analyzing pairs of variables

Success — continuous variable [0, 1]

Correlation between different data values

- Sometimes each sample-point can have a pair of attributes.
- And it may so happen that large values of the first attribute are accompanied with large (or small) values of the second attribute for a large number of sample-points.

Correlation between different data values

- Example 1: Populations with higher levels of fat intake show higher incidence of heart disease.
- Example 2: People with higher levels of education often have higher incomes.
- Example 3: Literacy Rate in India as a function of time?

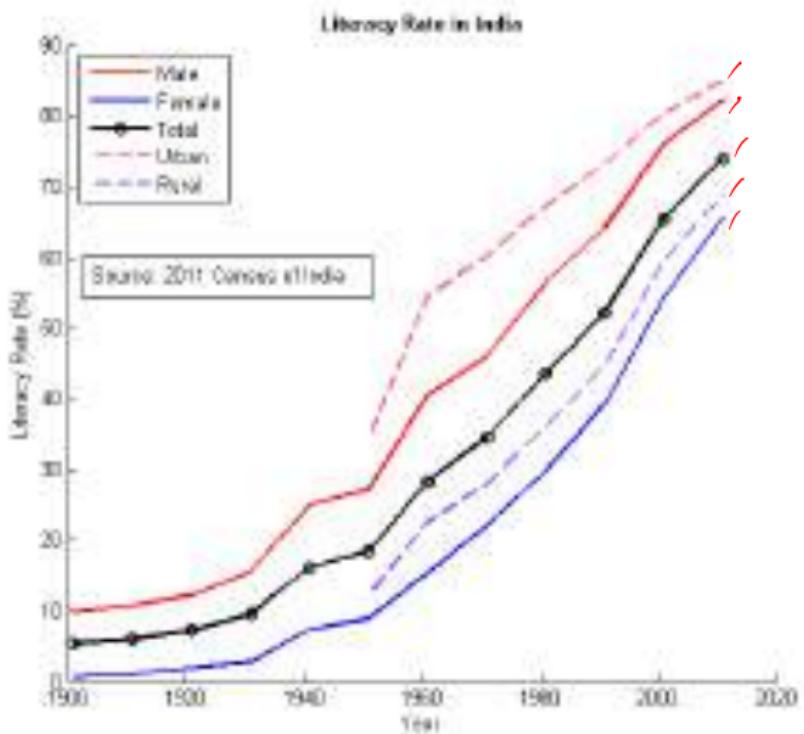


Image source

Visualizing such relationships?

- Can be done by means of a scatter plot
- X axis: values of attribute 1, Y axis: values of attribute 2
- Plot a marker at each such data point. The marker may be a small circle, a +, a *, and so on.

Table 2.8 Temperature and Defect Data

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	26.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Number
of defects

36

38

39

41

44

47

50

53

56

59

62

65

Temperature

Correlation coefficient

- Let the sample-points be given as (x_i, y_i) , $1 \leq i \leq N$.
- Let the sample standard deviations be σ_x and σ_y and the sample means be μ_x and μ_y .
- The **correlation-coefficient** is given as:

$$r(x, y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2 \sum_{i=1}^N (y_i - \mu_y)^2}} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x \sigma_y}$$

Correlation coefficient

- The correlation-coefficient is given as:

$$r(x,y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2 \sum_{i=1}^N (y_i - \mu_y)^2}} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x\sigma_y}$$

- $r > 0$ means the data are **positively correlated** (one attribute being higher implies the other is higher)
- $r < 0$ means the data are **negatively correlated** (one attribute being higher implies the other is lower)
- $r = 0$ means the data are **uncorrelated** (there is no such relationship!)
- r is **undefined** if the standard deviation of either x or y is 0.

Correlation coefficient: Properties

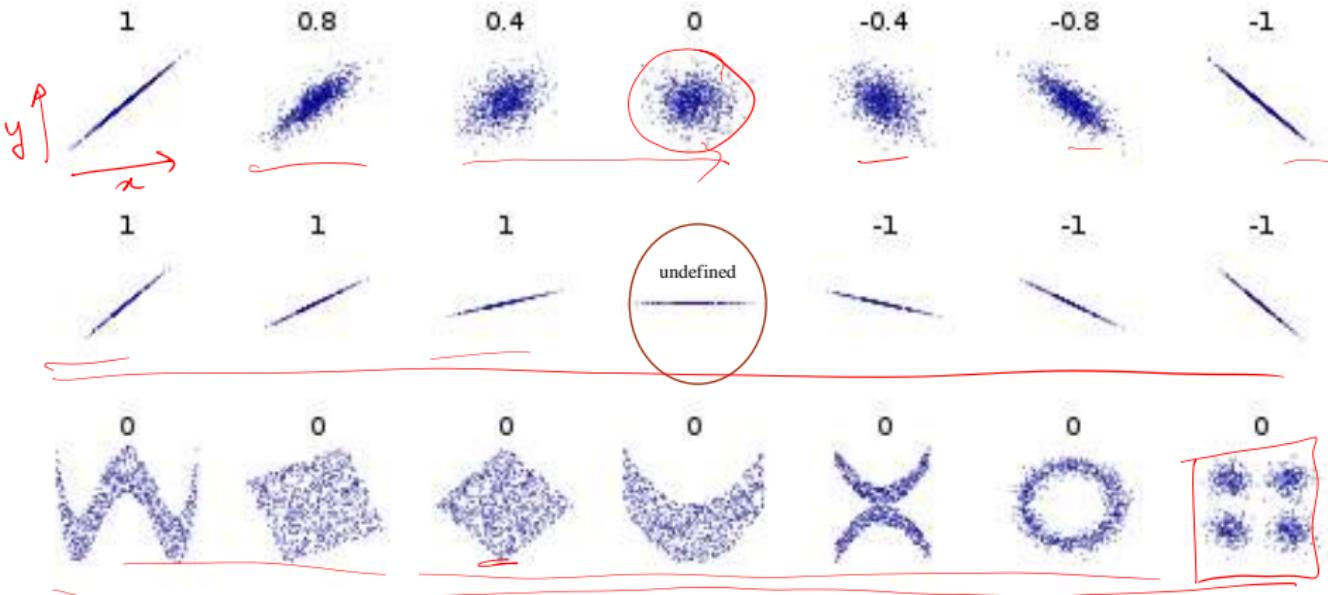
- The correlation-coefficient is given as:

$$r(x, y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2 \sum_{i=1}^N (y_i - \mu_y)^2}} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x \sigma_y}$$

- $-1 \leq r \leq 1$ always!

Prove it!

$$\begin{aligned} & i \quad (x_i - \mu_x)(y_i - \mu_y) \\ & + \quad (\overbrace{x_{i+1} - \mu_x}^{\text{!}}) (-\underbrace{y_i + 2\mu_y - \mu_y}_{y_{i+1}}) \end{aligned}$$



Correlation coefficient values for various toy datasets in 2D:
for each dataset, a scatter plot is provided

https://en.wikipedia.org/wiki/Correlation_and_dependence

Correlation coefficient: Properties

- In the following, we have a, b, c, d constant.

y_i is an affine transform of x_i

$$y_i = x_i^2$$

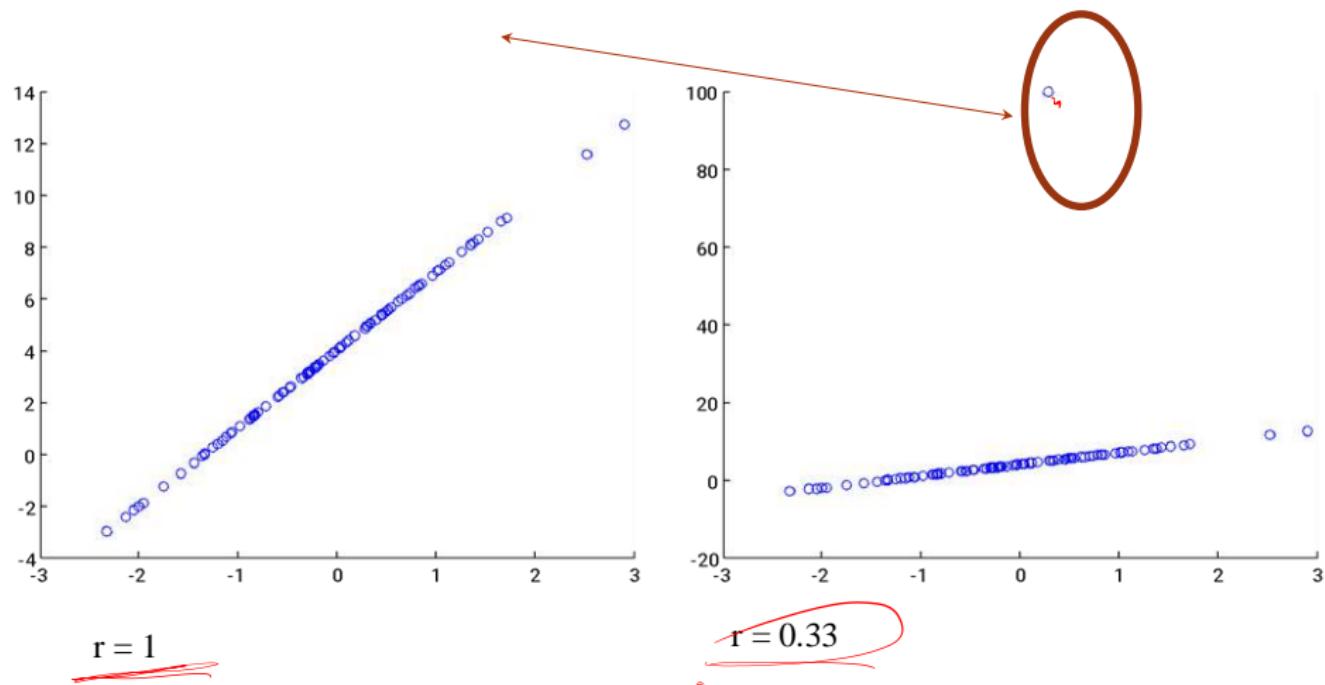
- If $y_i = a + bx_i$ where $b > 0$, then $r(x, y) = 1$.

- If $y_i = a + bx_i$ where $b < 0$, then $r(x, y) = -1$.

- If r is the correlation coefficient of data pairs as (x_i, y_i) , $1 \leq i \leq N$, then it is also the correlation coefficient of data pairs $(b+ax_i, d+cy_i)$ when a and c have the same sign.

Correlation coefficient: a word of caution

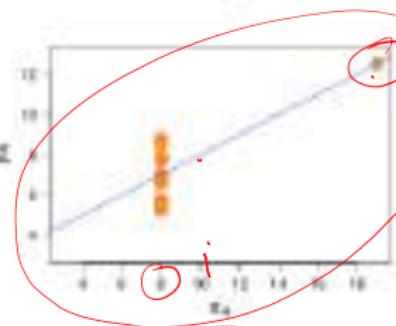
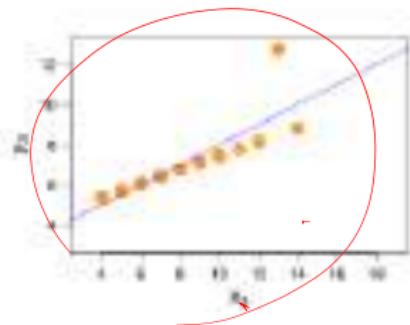
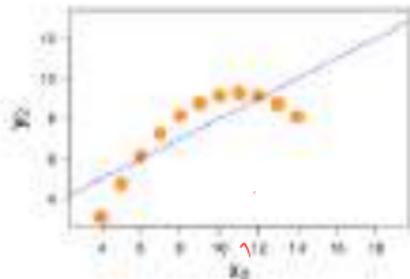
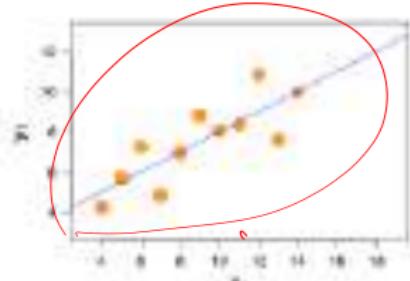
● Sensitive to outliers!



Caution with correlation: Anscombe's quartet

- The correlation coefficient can be a misleading value, and graphical examination of the data is important.
- This was illustrated beautifully by a British statistician named Frank Anscombe – by showing four examples that graphically appear very different – even though they produce identical correlation coefficients.
- These examples are famously called [Anscombe's quartet](#).

Caution with correlation: Anscombe's quartet



In each of these examples, the following quantities were the same:

- Mean and variance of x
- Mean and variance of y
- Correlation coefficient $r(x,y)$

But the data are graphically very different!

Image source

Reflective (or Uncentered) correlation coefficient

- A version of the correlation coefficient in which you do not deduct the mean values from the vectors!

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2 \sum_{i=1}^N (y_i - \mu_y)^2}} \neq r_{uncentered}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2 \sum_{i=1}^N y_i^2}}$$

- Uncentered c.c. is not “translation invariant”:

$$r(\mathbf{x}, \mathbf{y}) = r(\mathbf{x} + a, \mathbf{y} + b)$$

$$r_{uncentered}(\mathbf{x}, \mathbf{y}) \neq r_{uncentered}(\mathbf{x} + a, \mathbf{y} + b)$$

Correlation does not **necessarily** imply causation

- A high correlation between two attributes does not mean that one causes the other.
- Example 1: Fast rotating windmills are observed when the wind speed is high. Hence can one say that the windmill rotation produces speedy wind? (a **windmill** in the literal sense 😊)

Correlation does not **necessarily** imply causation

- In example 1, the cause and effect were swapped. High wind speed leads to fast rotation and not vice-versa.

- Example 2: High sale of ice-cream is correlated with larger occurrence of drowning. Hence can one say that ice-cream causes drowning?

- In this case, there is a third factor that is highly correlated with both – ice-cream sales, as well as drowning. Ice-cream sales and swimming activities are on the rise in the summer!

Correlation does not **necessarily** imply causation

- The above statement does not mean that correlation is *never* associated with causation (example: increase in age does cause increase in height in children or adolescents) – just that it is not *sufficient* to establish causation.

- Consider the argument: “High correlation between tobacco usage and lung cancer occurrence does not imply that smoking causes lung cancer.”

Correlation does not necessarily imply causation – but it **may**!

● However multiple observational studies that eliminate other possible causes do lead to the conclusion that smoking causes cancer!

- higher tobacco dosage associated with higher occurrence of cancer
- stopping smoking associated with lower occurrence of cancer
- higher duration of smoking associated with higher occurrence of cancer
- unfiltered (as opposed to filtered) cigarettes associated with higher occurrence of cancer

• See

<https://www.sciencebasedmedicine.org/evidence-in-medicine-correlation-and-causation/> and

<http://www.americanscientist.org/issues/pb/what-everyone-should-know-about-statistical-correlation> for more details.

More examples





Relationship between continuous and discrete variables

Future topics

- Multi-variate visualization
- Commercial systems for data visualization
- Visualizing special data
 - Time series
 - Text, e.g. point clouds

Lecture 04-probability.pdf

Elements of Probability

Fall 2024
Sunita Sarawagi

Thanks to Prof Ajit Rajawade for the initial version of the slides

Data interpretation from samples is uncertain

- Need a formal representation of uncertainty
- Probability provides a formal framework of expressing uncertainty when drawing conclusions from finite samples of a much larger population.

Probability in Computer Science

- Algorithm design
 - Randomized algorithms: steers around unlikely situations
 - Several hard problems that can only be solved efficiently with high probability
- Performance analysis
 - What is the probability that you will find the next accessed page in cache?
 - What is the probability that the length of the queue will be greater than 5 when a job arrives at a server?
- Network protocol design
- Machine Learning/AI: Is all about probability and statistics

Topic Overview

- Terminology: sample space, event, probability
- Composition of events; mutual exclusion and independence
- Axioms of probability
- Principles of counting
- Conditional probability and Bayes' theorem
- Some paradoxes!

Sample space

- Consider an experiment whose outcome is not known in advance.
- Example 1: A coin toss
- But we do know the complete set of possible outcomes:
Heads or tails
- The set of all possible outcomes of an experiment is called the **sample space**.

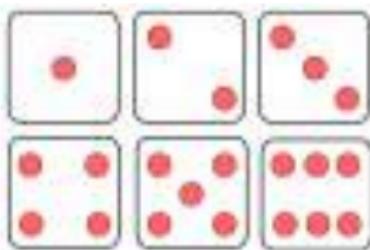


Sample space

- Example 2: Measurement of your body temperature (assume it's an integer) with a thermometer. What's the sample space?
 - Say between 30 to 40 degrees Celsius, so the sample space = $\{30, 31, \dots, 39, 40\}$
- Example 3: An experiment to randomly choose a student from the CSE 2024 batch at IITB and declare him/her the branch topper
 - Sample space = set of all students in that batch

Sample space

- Example 4: Consider a four-country ODI series between India, Pakistan, Bangladesh and Australia. What is the set of rankings?
 - Sample space = set of all $4!$ permutations of the string IPBA
- Example 5: An experiment to roll a die
 - Sample space = {1,2,3,4,5,6}



Event

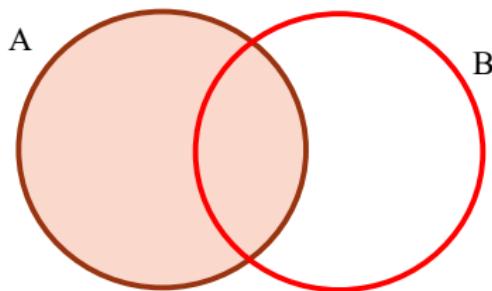
- Any subset of the sample space is called an **event**.
- If the outcome of an experiment is contained in Event E , then we say E has *occurred*.
- In example 1, if $E = \{H\}$, then E is the event that the coin produced a heads.
- In example 2, if $E = \{\text{set of temperatures from } 33 \text{ to } 37\}$, then E is the event that the temperature was “normal” (i.e. not exceeding 37 and not less than 33)

Composition of Events

- Given event E , event E^c is the event that E did not occur. E^c is called the **complement** of E .
- Given events E and F , the event G that **either** E or F (or **both**) occur is called **as the union** of E and F , and denoted as $G = E \cup F$.
- Given events E and F , the event G that **both** E and F occur is called **as the intersection** of E and F , and denoted as $G = E \cap F$ or $G = EF$

Composition of Events

- Union and intersection can be extended to handle any arbitrary number of events.
- If two events cannot occur together (for example?), then their intersection is a null set. Such events are called mutually exclusive.



This is called as a Venn diagram in set theory.

Composition of Events

- An event and its complement – are always mutually exclusive events.
- Example:
 - Let F be the event that a patient tests negative for a certain disease in a medical test.
 - Let G be the event that (s)he tests positive for the same disease in the same test.
 - Then F and G are mutually exclusive.
- Example:
 - Let E be the event that the sum of three consecutive dice throws was greater than or equal to 3.
 - Let F be the event that the sum of three consecutive dice throws was greater than or equal to 4.
 - Then E and F are NOT mutually exclusive. In fact F is a subset of E.

Probability of an event

- We conduct an experiment, whose outcomes are uncertain but come from a sample space S
- We are interested in a subset S of the sample space, which we call an event E .
- If we repeat the experiment under identical conditions very large number of times,
 - Probability of E , $P(E)$ is the fraction of times that outcome is in event E
- Example: rolling of dice.

Axioms of probability

- For an event E from sample space S, we have:

Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

Axiom 3: For mutually exclusive events E_1, E_2, \dots, E_n , we have

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i), \quad n = 1, 2, \dots, \infty$$

The notion of relative frequency of event E obeys the above axioms

Properties derivable from axioms

- Properties (can be proved by Venn diagrams):

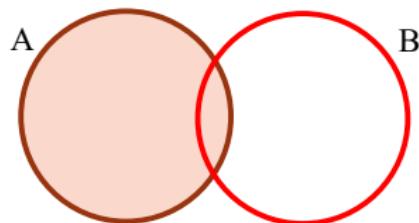
$$P(A \cup B) = P(A) + P(B) - P(AB) \leq P(A) + P(B)$$

This implies

$$(1) P(A^c) = 1 - P(A)$$

$$(2) A \subseteq B \rightarrow P(A) \leq P(B)$$

Is the converse of (2) also true?



Equally likely outcomes

- We will assume that each of the singleton outcomes in the sample space is equally likely.
- So, if the experiment is to roll a die, then all six faces will show up with equal probability.
- We will assume finite sample spaces for now.
- In such a case, the probability of an event E is given as:

$$P(E) = \frac{\text{Number of points in } E}{\text{Number of points in sample space}}$$

Principles of counting: motivating example

- Useful when solving problems on discrete probability.
- For example: Suppose a box contains 6 white and 5 black balls. If you draw two balls at random, what is the probability that one is white and the other is black?

Principles of counting: Product rule

Suppose a procedure can be broken down into a sequence of k tasks, and there are

- n_1 ways to do task 1,
- n_2 ways to do task 2, ...
- n_k ways to do task k .

Then there are $n_1 n_2 \dots n_k$ ways to do the entire procedure.

```
c = 0
for (i1 = 1 to n1)
{
    for (i2 = 1 to n2)
    {
        .
        .
        .
        for (ik = 1 to nk)
        {
            c = c + 1
        }
    }
}
```

Principles of counting: example

- For example: Suppose a box contains 6 white and 5 black balls. If you draw two balls at random, what is the probability that one is white and the other is black?
- There are two scenarios: (1) the first ball is white and second is black, or (2) vice versa.
- For (1), the probability that the white ball is picked is $6/11$, and the probability that the black ball is picked is $5/10$ (10 balls remain after the first white ball is picked). The overall probability is $30/110$ (product rule).
- For (2), the probability that the black ball is picked is $5/11$, followed by a $6/10$ probability of picking a white ball, leading to an overall probability of $30/110$ (product rule).
- The total probability is $(30+30)/110 = 6/11$ (sum rule).

Conditional Probability

- An important concept.
- Helps one quantify uncertainty of outcomes under partial knowledge or constraints.
- For example
 - What is the probability that the outcome of a dice roll is 2 given that it is even?

Let A, B be two events. Conditional probability of A, given that B has already occurred

$$P(A|B) = P(A \cap B)/P(B)$$

Bayes Formula

Example 3.7.d. In answering a question on a multiple-choice test, a student either knows the answer or she guesses. Let p be the probability that she knows the answer and $1 - p$ the probability that she guesses. Assume that a student who guesses at the answer will be correct with probability $1/m$, where m is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question given that she answered it correctly?

- S = {KC, GC, GI}
- B = Correct answer = {KC, GC}
- A = {KC, X}
- $P(A, B) = P(KC) = p$
- $\underline{P(B)} = \underline{p} + (1-p) \frac{1}{m}$

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P}{p + (1-p) \frac{1}{m}} = 0.88$$

$$m = 5$$

$$\underline{p = 0.6}$$

$$P(K|C) = \frac{0.6}{0.6 + \frac{0.4}{5}} = \frac{1}{1 + \frac{2}{3}} = 0.6$$

Example 3.7.e. A laboratory blood test is 99 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a "false positive" result for $.01$ percent of the healthy persons tested. (That is, if a healthy person is tested, then, with probability $.01$, the test result will imply he or she has the disease.) If $.5$ percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive?

$$P(T_P | D) = 0.99$$

$$P(T_P | D^c) = 0.001$$

$$P(D) = 0.005$$

$$P(D | T_P) = \frac{P(T_P | D) P(D)}{P(T_P | D) P(D) + P(T_P | D^c) P(D^c)}$$

$\underbrace{P(T_P | D) P(D)}_{P(T_P)}$

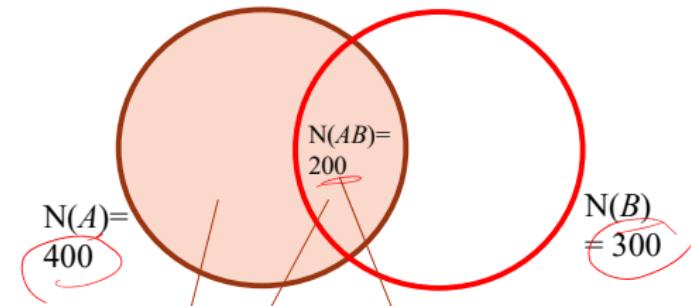
$$\therefore = \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.001 \approx (1 - 0.005)}$$

Lecture 05-independence.pdf

Joint probability

- The probability that events A and B both occur (in the same experiment) is called the **joint probability** of the events A and B . This is another word for the probability of the *intersection* of A and B .

Conditional and Joint probability: what's the difference?



conditional probability

This area divided by
the total sample space
size is the joint

$$P(A|B) = \frac{200}{300} = \frac{2}{3}$$

Joint probability of $A \cap B$

probability of A and B

$$= \frac{200}{200 + 300 - 200} = \frac{200}{500}$$

The ratio of the smaller area

to the larger area (area of the

full circle)

is the conditional

probability of B given A .

If $P(A|B) > P(AB)$? Yes

- Let the original sample space be S .
- In computing $P(B|A)$, you assume that A has already occurred.
- Therefore your new sample space S'' for computing $P(B|A)$ contains only those events which lie in A .
- For computing $P(AB)$, the sample space is the entire S .
- Now can you compute $P(A|B)$?

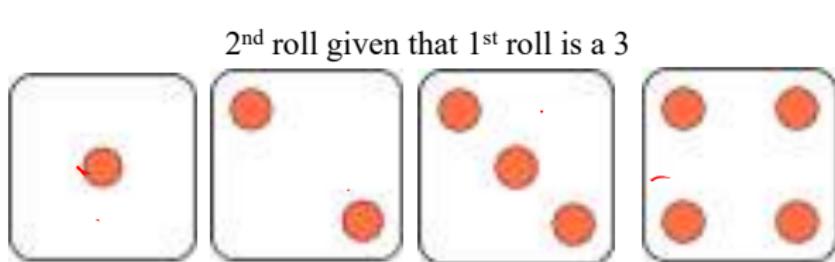
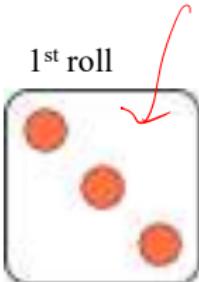
$$= \frac{200}{500}$$

$$= \frac{N(AB)}{N(A) + N(B) - N(AB)}$$

$$= \frac{2}{5}$$

Conditional and joint probability: what's the difference

- Consider two consecutive rolls of a die. Given that the first die produced a 3, what's the probability that the sum of the two throws does not exceed 7?
- Solution: $\# \text{ of combinations s.t}$
 - A = event that first throw produced a 3. $P(A) = 1/6$.
 - B = event that sum does not exceed 7. $P(B) = \frac{21}{36}$
 - Joint probability $P(AB) = \frac{4}{36}$.
 - Conditional probability: $P(B|A) = P(AB)/P(A) = (4/36)/(1/6) = 2/3$.



Example

India has a literacy rate of 74%. The state of Kerala has a literacy rate of 94% and constitutes 2.8% of India's population.



What is the probability that:

- A randomly chosen Indian person is literate $\cdot 74$ $P(L)$
- A randomly chosen Indian person is from Kerala $\cdot 028$ $P(k)$
- A randomly chosen person from Kerala is literate $\cdot 94$ $P(L|k)$
- A randomly chosen Indian person is from Kerala and is literate $P(k \cap L) = P(L|k)P(k) = 0.94 \times 0.028$
- A randomly chosen Indian person is from Kerala if you knew already that (s)he was literate

$$P(k|L) = \frac{P(L|k)P(k)}{P(L)}$$

Example

India has a literacy rate of 74%. The state of Kerala has a literacy rate of 94% and constitutes 2.8% of India's population.

What is the probability that:

- A randomly chosen Indian person is literate $P(L)=0.74$
- A randomly chosen Indian person is from Kerala $P(K)=0.028$
- A randomly chosen person from Kerala is literate $P(L|K)=0.94$
- A randomly chosen person is from Kerala and is literate
 $P(K, L)=P(L|K)P(K)=0.94*0.028$
- A randomly chosen person is from Kerala if you knew already that (s)he was literate $P(K|L)=P(K, L)/P(L)=0.94*0.028/0.74$

Independence of events

- If A and B are independent, the occurrence of A has no bearing on the probability of occurrence of B (and vice versa).
- Examples of independent events:
 - Outcomes from two dice rolls
 - Height of a person and the last digit of their mobile phone
 - Rainfall tomorrow in Mumbai and number of winning ticket lottery

- Example of dependent events

- Is it rainy in morning & is it sunny in the afternoon
- Two pulls of balls from a bag without replacement -
- Preparation for a test & marks -

Independence

Two events A and B are **independent** if:

$$P(A) = P(\underline{A|B})$$

Intuitive Definition:

Knowing that event B happened doesn't change our belief that A happens.

With independence, we can simplify the chain rule:

$$\begin{aligned} P(A \cap B) &= P(\underline{A \cap B}) = P(A|B) \cdot P(B) \\ &= P(A) \cdot P(\underline{B}) \end{aligned}$$

You can also show this \Rightarrow to prove independence

Piech & Conn, CS109, Stanford University

Independence of more than two events

- We say that $n > 2$ events are mutually independent if and only if for every subset A of $k \leq n$ events, we have:

$$P\left(\bigcap_{i=1}^k P(A_i)\right) = \prod_{i=1}^k P(A_i)$$

- Example: three events A_1, A_2, A_3 . To show that they are independent we need to show that:

$$\begin{aligned} P(A_1 A_2 A_3) &= P(A_1) P(A_2) P(A_3) \\ P(A_1 A_2) &= P(A_1) P(A_2) \\ P(A_1 A_3) &= P(A_1) P(A_3) \\ P(A_2 A_3) &= P(A_2) P(A_3) \end{aligned}$$

Example 3.8.c. Two fair dice are thrown. Let E_7 denote the event that the sum of the dice is 7. Let F denote the event that the first die equals 4 and let T be the event that the second die equals 3. Now it can be shown (see Problem 36) that E_7 is independent of F and that E_7 is also independent of T ; but clearly E_7 is not independent of FT [since $P(E_7|FT) = 1$]. ■

$$\begin{array}{l} E_7 \perp\!\!\!\perp T \\ E_7 \perp\!\!\!\perp F \end{array} \quad \left\{ \right. \not\Rightarrow E_7 \perp\!\!\!\perp TF$$

$S = \{$	[1,1]	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
	[2,1]	[2,2]	[2,3]	[2,4]	[2,5]	[2,6]
	[3,1]	[3,2]	[3,3]	[3,4]	[3,5]	[3,6]
	[4,1]	[4,2]	[4,3]	[4,4]	[4,5]	[4,6]
	[5,1]	[5,2]	[5,3]	[5,4]	[5,5]	[5,6]
	[6,1]	[6,2]	[6,3]	[6,4]	[6,5]	[6,6] $\}$

$$P(E_7) = \frac{1}{6} \quad \checkmark$$

$$P(F) = \frac{1}{6} \quad \checkmark$$

$$P(T) = \frac{1}{6}$$

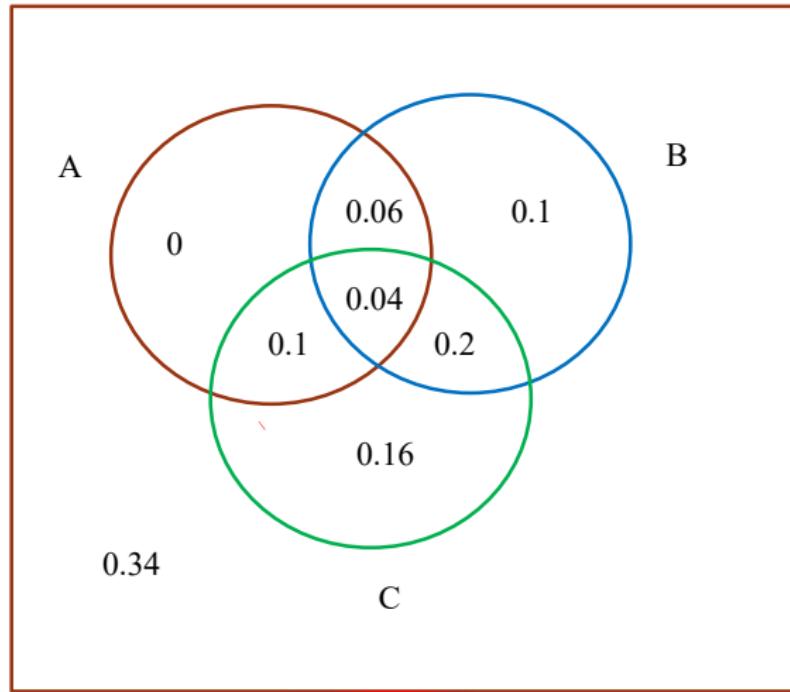
$$P(E_7 \cap F) = \frac{1}{36}$$

$$P(E_7, F) = P(E_7)P(F) \quad \checkmark \quad P(E_7 | FT) = 1 \neq P(E_7)$$

$$P(E_7, T) = P(E_7)P(T)$$

Only n-way independence does not suffice

- Note that only n-way independence of events does not imply that every pair of events are independent.
- Example: See next slide



~~$P(ABC) = 0.04 = P(A)P(B)P(C) = (0.2)(0.4)(0.5)$~~

~~$P(AB) = 0.1 \neq P(A)P(B)$~~

Independence versus Mutual Exclusion

- If A and B are mutually exclusive, then $P(AB) = 0$.
- If A and B are independent, then $\underline{P(AB)} = \underline{P(A)P(B)} \neq 0$.
- The two are usually not the same! In fact, for mutually exclusive events, the occurrence of one *does* have an effect on that of the other.

Independence and mutual exclusion

If A is independent of B, can we say that A is independent of B^c ?

Yes:

$$P(AB) = P(A) P(B) \quad \therefore A \perp\!\!\!\perp B$$

$$P(A) = P(AB) + P(AB^c) \quad [\text{Law of total probability}]$$

$$= P(A) P(B) + P(AB^c)$$

$$\Rightarrow P(AB^c) = P(A) - P(A) P(B)$$

$$= P(A) [1 - P(B)] =$$

$$= P(A) P(B^c)$$

$$\Rightarrow A \perp\!\!\!\perp B^c$$

The Core Probability Toolkit



The Law of Total Probability

$$P(B) = P(E \text{ and } B) + P(E \text{ and } B^C)$$

$$P(B) = P(E|F)P(F) + P(E|F^C)P(F^C)$$

$$S = \bigcup_{i=1}^n B_i \cap B_j^c$$

$$\begin{aligned} P(B) &= \sum_{i=1}^n P(E \text{ and } B_i) \\ &= \sum_{i=1}^n P(E|B_i)P(B_i) \end{aligned}$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|B^C)P(B^C)}$$

Definition of Conditional Probability

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

Axiom 1: $0 \leq P(B) \leq 1$

Axiom 2: $P(S) = 1$

Axiom 3: If E and F are mutually exclusive, then $P(E \text{ or } F) = P(E) + P(F)$

Otherwise, we Introduce Inclusion:

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

Chain Rule

$$\begin{aligned} P(E \text{ and } F) &= P(E|F)P(F) \\ &= P(F|E)P(E) \end{aligned}$$

$$P(B^C) = 1 - P(B)$$

De Morgan's Laws

($A \text{ and } B^C = A^C$ and if
 $(A \text{ and } B)^C = A^C \text{ or } B^C$)

Independence

$$P(E|F) = P(E)$$

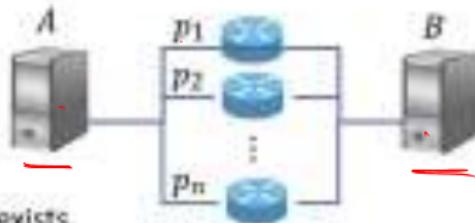
$$P(E \text{ and } F) = P(E)P(F)$$

Practice: Network Reliability

Consider the following parallel network:

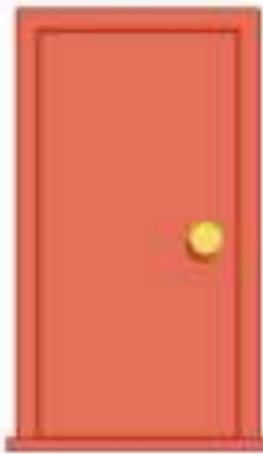
- * n independent routers, which each are working with probability p_i ($1 \leq i \leq n$)

Let E be the event that a working path from A to B exists.
What is $P(E)$?

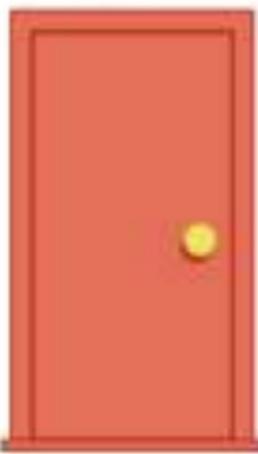


$$P(E) = 1 - \prod_{i=1}^n (1-p_i)$$

1



2



3



The Monty Hall Problem

The Monty Hall Problem

Behind one door is a prize (equally likely for each door).

Behind the other two doors are goats.

How to play:

1. We choose a door.
2. Host opens 1 of the other 2 doors, revealing a goat.
3. We are given an option to switch to the other door.



Note: If we don't switch,
 $P(\text{win}) = 1/3$

We are comparing
 $P(\text{win})$ vs. $P(\text{win}|\text{switch})$

Should we switch?

$$P(\text{win}) = P(\text{win}|\text{switch})$$

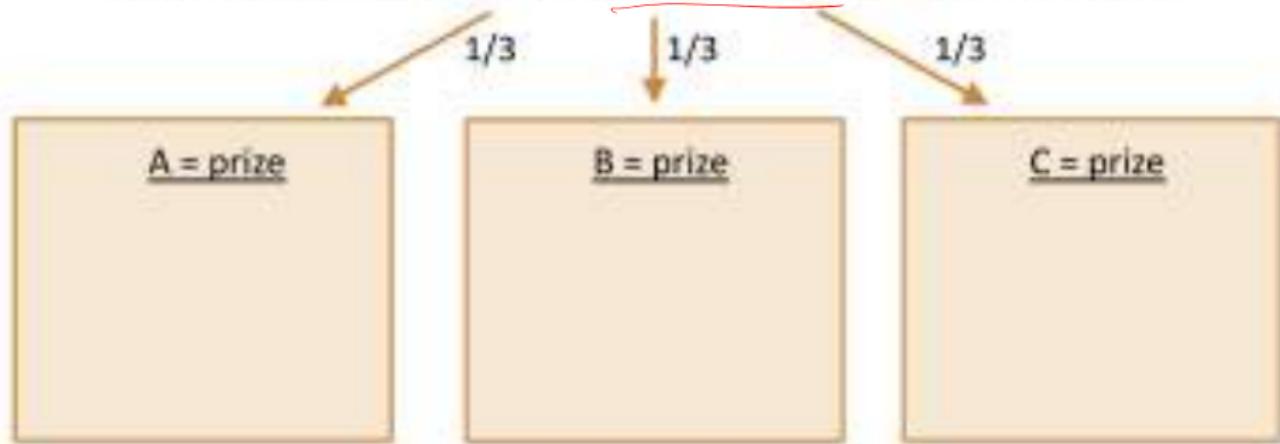
$$P(\text{prize in 1}) = P(\text{prize in 2})$$

Pitkä & Cain, C3106, Stanford University

Let's Find $P(\text{win} \mid \text{switch})$

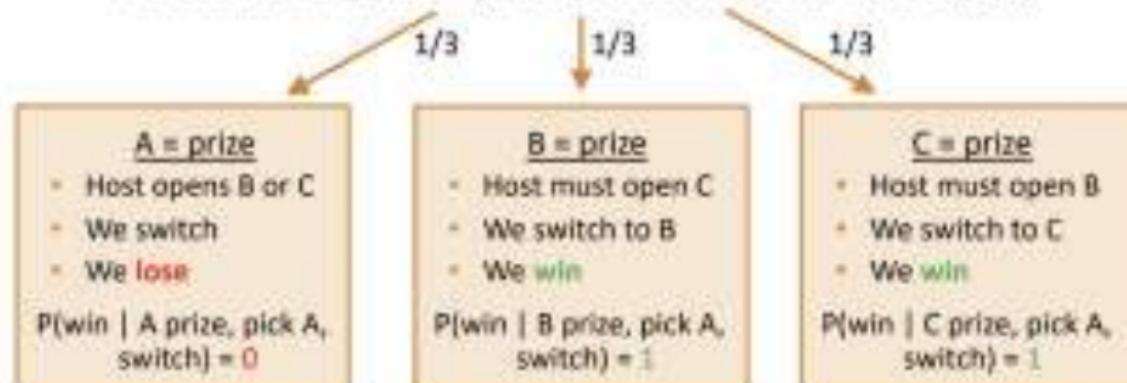
Paul Erdős

Without loss of generality, let's pick door A (out of doors A,B,C).



Let's Find $P(\text{win} \mid \text{switch})$

Without loss of generality, let's pick door A (out of doors A,B,C).



$$\begin{aligned} P(\text{win} \mid \text{pick A, switch}) &= P(\text{win} \mid \text{A prize, pick A, switch}) * P(\text{A prize}) + \\ &\quad P(\text{win} \mid \text{B prize, pick A, switch}) * P(\text{B prize}) + \\ &\quad P(\text{win} \mid \text{C prize, pick A, switch}) * P(\text{C prize}) \\ &= 1/3 * 0 + 1/3 * 1 + 1/3 * 1 = 2/3 \end{aligned}$$

You should switch!

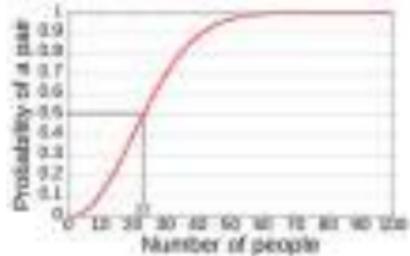
Fiech & Cain, CS509, Stanford University

The Birthday Paradox!

- Given n people in a room, what should be the least value of n such that the probability that at least 2 people in the room share the same birthday is greater than or equal to 99.9%?
- Each person can have his/her birthday on any of the 365 days. For n people, there are 365^n outcomes.
- The number of outcomes resulting in no two people sharing a birthday is $(365)(364)(363)\dots(365-n+1)$.

The Birthday Paradox!

- So required probability is
- $1 - \frac{(365)(364)(363)\dots(365-n+1)}{(365)^n} \geq 0.999$ (given)
- This is satisfied for n as small as 70.
- For $n = 20$, it is around 41%.
- For $n = 23$ it is close to 50%
- For $n = 40$, it is around 89%.
- For more information see the [wikipedia article on the birthday paradox.](#)



Conclusions

- Reasoning about probabilities is tricky
- Important to carefully analyze the sample space, and conditioning variable

Lecture 06-randomVars.pdf

Random Variables Are Variables...That Are Random

Random Variables Are Variables...That Are Random

Check out the variable **result** in the code below.

```
import random  
  
def flip_coin():  
    # returns 0 or 1 with prob. 0.5  
    return random.choice([0,1])  
  
result = flip_coin()
```

def constant():
 return 42

result = constant()
→ not random

Random Variables Are Variables...That Are Random

Check out the variable **result** in the code below.

```
import random

def flip_coin():
    # returns 0 or 1 with prob. 0.5
    return random.choice([0,1])

result = flip_coin()
```

- Do we know the value of **result** before we run the code?

Random Variables Are Variables...That Are Random

Check out the variable **result** in the code below.

```
import random

def flip_coin():
    # returns 0 or 1 with prob. 0.5
    return random.choice([0,1])

result = flip_coin()
```

- Do we know the value of **result** before we run the code? **Nope!**
- Is the value of **result** the same every time we run the code?

Random Variables Are Variables...That Are Random

Check out the variable **result** in the code below.

```
import random

def flip_coin():
    # returns 0 or 1 with prob. 0.5
    return random.choice([0,1])

result = flip_coin()
```

- Do we know the value of **result** before we run the code? **Nope!**
- Is the value of **result** the same every time we run the code? **Nope!**

Like **result**, a random variable is a variable whose value is uncertain.

Random Variables Are Variables...That Are Random

A **random variable** is a variable whose value is uncertain.

```
import random  
  
def flip_coin():  
    # returns 0 or 1 with prob. 0.5  
    return random.choice([0,1])  
  
result = flip_coin()
```



Let X be the result of flipping a coin.

$$\begin{aligned}P(X = 0) &= 0.5 \checkmark \\P(X = 1) &= 0.5\end{aligned}$$



Random Variables Are Variables...That Are Random

A **random variable** is a variable whose value is uncertain.

```
import random  
  
def flip_coin():  
    # returns 0 or 1 with prob. 0.5  
    return random.choice([0,1])  
  
result = flip_coin()
```



“Let **X** be the result of flipping a coin.”

$$P(X=0) = 0.5$$

$$P(X=1) = 0.5$$

- Random variables store the outcome of an experiment
- Random variables can be described by their possible outcomes + probabilities
 - Note: random variables can only be numbers (not “heads” or “tails”)

Random variables are an abstraction on top of events

Random variables are *not* events

Random Variables vs. Events

X

Let X be a
random variable

Random Variables vs. Events

It is an event when
 X takes on a value

$$X \quad X = 2$$

Let X be a
random variable

$$X \in \{2, 4, 6\}$$

Random Variables vs. Events

It is an event when
 X takes on a value

$$X \quad X = 2 \quad \underline{P(X = 2)}$$

Let X be a
random variable

So we can still work with
probabilities of events

Examples of Random Variables

"Let X be the result of rolling a dice."

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

Examples of Random Variables

"Let X be the result of rolling a dice."

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

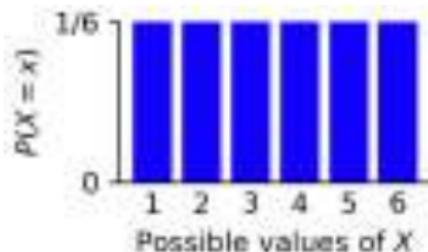
...or, $\underline{P(X=x) = 1/6 \text{ for } 1 \leq x \leq 6}$

Examples of Random Variables

"Let X be the result of rolling a dice."

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

...or, $P(X = x) = 1/6$ for $1 \leq x \leq 6$

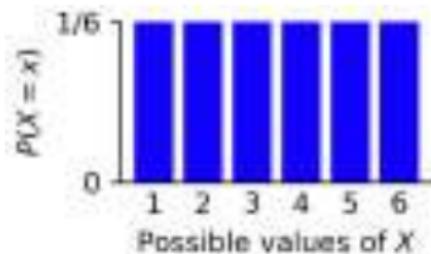


Examples of Random Variables

"Let X be the result of rolling a dice."

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

...or, $P(X = x) = 1/6$ for $1 \leq x \leq 6$



"Let \underline{Y} be the number of heads seen in 2 coin flips."

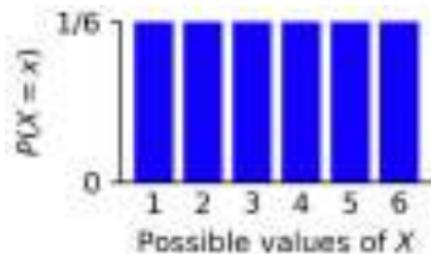
- $P(Y = 0) = 1/4$
 - $P(Y = 1) = 1/2$
 - $P(Y = 2) = 1/4$
- (T, T) (H, T), (T, H) (H, H)
- {

Examples of Random Variables

"Let X be the result of rolling a dice."

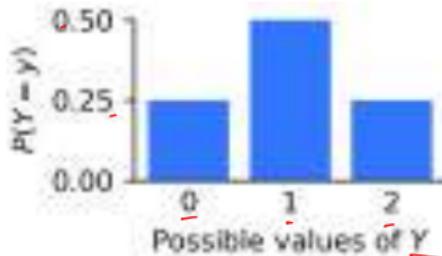
- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

...or, $P(X = x) = 1/6$ for $1 \leq x \leq 6$



"Let Y be the number of heads seen in 2 coin flips."

- $P(Y = 0) = 1/4$ (T, T)
- $P(Y = 1) = 1/2$ (H, T), (T, H)
- $P(Y = 2) = 1/4$ (H, H)



Examples of Random Variables

"Let Z be the sum of rolling two dice."

- $\underline{P(Z = 2) = 1/36}$
- $P(Z = 3) = 2/36$
- $\underline{P(Z = 4) = 3/36}$
- $P(Z = 5) = 4/36$
- $P(Z = 6) = 5/36$
- $P(Z = 7) = 6/36$
- $P(Z = 8) = 5/36$
- $P(Z = 9) = 4/36$
- $P(Z = 10) = 3/36$
- $P(Z = 11) = 2/36$
- $\underline{P(Z = 12) = 1/36}$



$$P(Z = z) = \begin{cases} \frac{z-1}{36}, & z \in \mathbb{Z}, 1 \leq z \leq 6 \\ \frac{13-z}{36}, & z \in \mathbb{Z}, 7 \leq z \leq 12 \\ 0, & \text{else} \end{cases}$$

Examples of Random Variables

"Let Z be the sum of rolling two dice."

- * $P(Z=2) = 1/36$
- * $P(Z=6) = 5/36$
- * $P(Z=10) = 3/36$
- * $P(Z=3) = 2/36$
- * $P(Z=7) = 6/36$
- * $P(Z=11) = 2/36$
- * $P(Z=4) = 3/36$
- * $P(Z=8) = 5/36$
- * $P(Z=12) = 1/36$

There's a name for what we're describing, when we list out all possible outcomes + their probabilities:

Probability Mass Function (PMF)



$$P(Z=z) = \begin{cases} \frac{1}{36}, & z \in \mathbb{Z}, 7 \leq z \leq 12 \\ 0, & \text{else} \end{cases}$$

Probability Mass Functions

Random Variables & Functions

"Let Y be the number of heads seen in 2 coin flips."

If this is a number

$$P(Y = \underline{\circled{2}})$$

Then this is a number
(between 0 and 1)

Random Variables & Functions

"Let Y be the number of heads seen in 2 coin flips."

If this is a variable

$$P(\underline{Y} = \underline{k})$$

Then this is a function

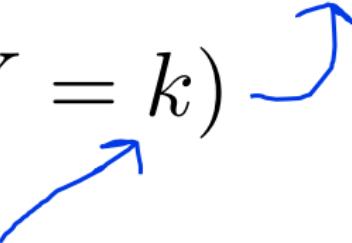
$$f_Y(k) : \{0, 1\} \rightarrow \{0, 1\}$$

Random Variables & Functions

"Let Y be the number of heads seen in 2 coin flips."

...and get out their probabilities!

0.5

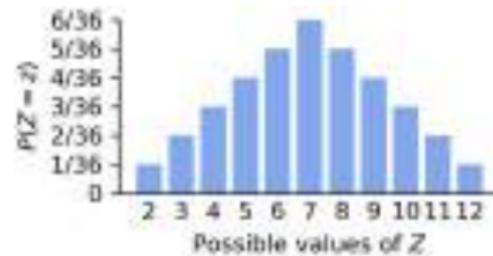
$$P(Y = k)$$


We can put in $k = 1$
different inputs...

The relationship between values a random variable can take on, and the corresponding probability, is a ***function!***

Probability Mass Function: Representations

$$P(Z = z) = \begin{cases} \frac{z-1}{36} & z \in \mathbb{Z}, 1 \leq z \leq 6 \\ \frac{13-z}{36} & z \in \mathbb{Z}, 7 \leq z \leq 12 \\ 0 & \text{else} \end{cases}$$



```
def event_probability(z):
    # probability mass function of Z
    if not z.is_integer() or z > 12 or z < 1:
        return 0

    if z < 7:
        return (z - 1) / 36
    else:
        return (13 - z) / 36
```

All of these are different ways we can represent probability mass functions!

Lecture 07-randomVars.pdf

Random Variables: Continued

Types of Random Variables

- Discrete Vs Continuous

Specifying probability of discrete R.V.

- Probability Mass Function $P(X=k)$

$X \in \{1, 2, 3, 4, 5, 6\}$ ← outcome of dice roll

$$P(X=1) = p_1$$

$$P(X=2) = p_2$$

⋮

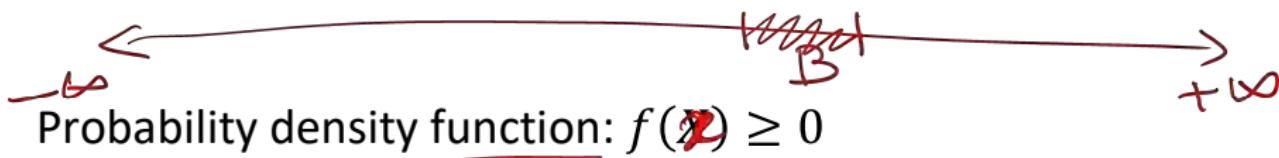
$$P(X=k) = p_k$$



enumerative representation of pmf.

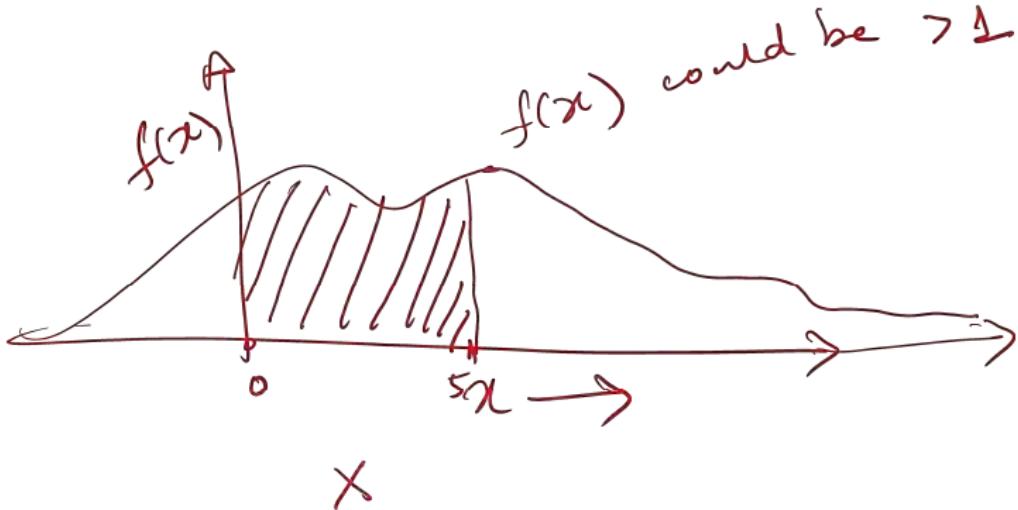
Specifying probability of continuous R.Vs

- If X is continuous, it can take an infinite number of values.
- Probability mass function: $P(X = k)$ cannot be defined.
- Instead we ask for probability that x lies in an interval B of non-zero size: $P(\underline{X} \in B)$

Probability density function: $f(x) \geq 0$

$$P(X \in B) = \int_{\underline{x} \in B} f(x) dx \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$



$$f(x) = \begin{cases} 2 & \text{if } 0 \leq x \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Cumulative distribution function (CDF)

- Assume R.V. X is ordered.
- CDF of X is a function $F(a)$ that takes a value a and return $P(X \leq a)$

- CDF of a discrete distribution. X is discrete and ordered: x_1, x_2, \dots, x_k

$$F(x) = \sum_{x_i \leq x} P(X = x_i)$$

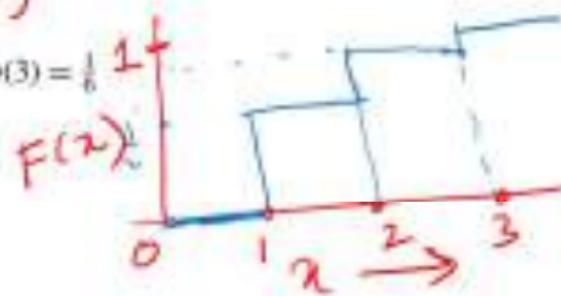
- Example: $p(1) = \frac{1}{2}, p(2) = \frac{1}{3}, p(3) = \frac{1}{6}$

$$x_1 = 1, x_2 = 2, x_3 = 3$$

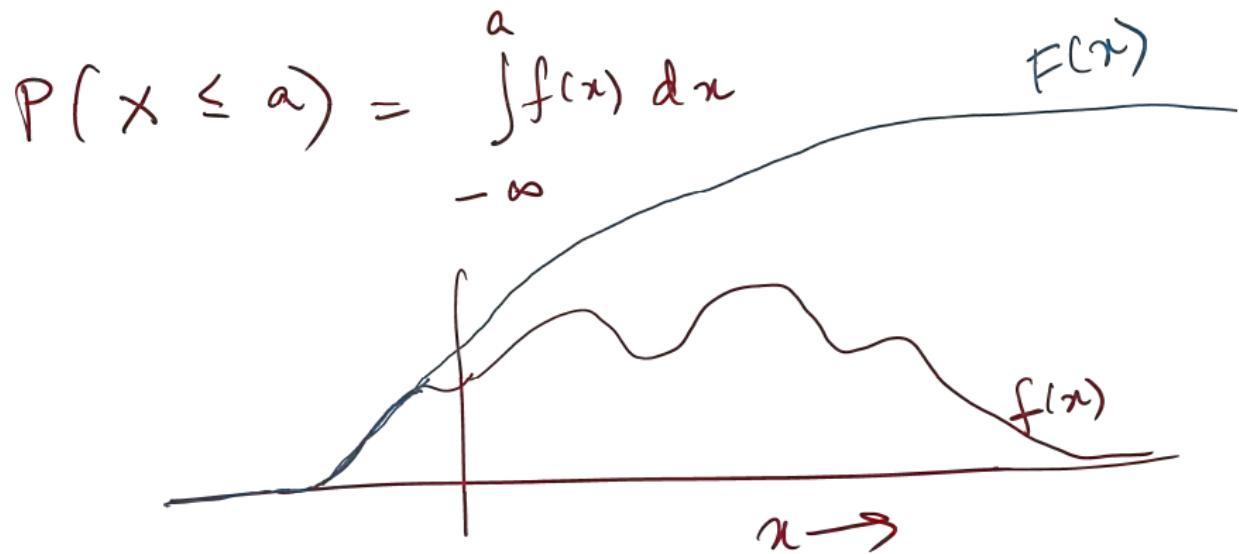
$$F(1) = \frac{1}{2}$$

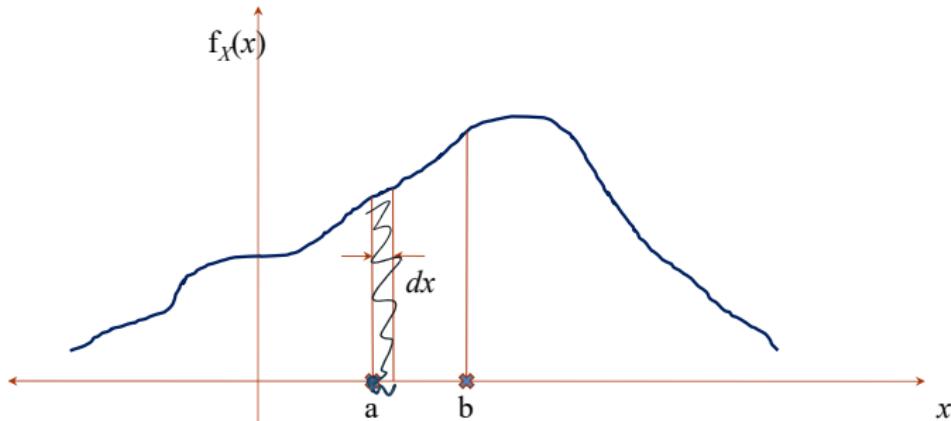
$$F(2) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

$$F(3) = \frac{5}{6} + \frac{1}{6} = 1$$



CDF of a continuous distribution





The area beneath the blue curve in between the lines $x = a$ and $x = b$ is the **cumulative interval measure** $\underline{P(a < X \leq b)} = \underline{F_X(b)} - \underline{F_X(a)}$.

$f_X(a)dx$ = probability that the random variable X takes on values between a and $a+dx$.

Random variable: continuous - example

Consider a CDF of the form:

$$F_X(x) = 0 \text{ for } x \leq 0, \text{ and}$$

$$F_X(x) = 1 - \exp(-x^2) \text{ otherwise}$$

To find: probability that X exceeds 1

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - F(1) \\ &= 1 - 1 - e^{-1^2} \\ &= e^{-1} \end{aligned}$$

Expected Value of a random variable

For a discrete random variable X , is defined as:

$$E(X) = \sum x_i P(X = x_i)$$

The expected value that shows up when you throw a die is $\frac{1}{6}(1+2+3+4+5+6) = 3.5$.

For continuous random variable X , is defined as:

$$E(X) = \int_{-\infty}^{\infty} xf_X(x)dx$$

Expected Value: examples

The game of roulette consists of a ball and wheel with 38 numbered pockets on its side. The ball rolls and settles on one of the pockets. If the number in the pocket is the same as the one you guessed, you win \$35 (probability 1/38), otherwise you lose \$1 (probability 37/38). The expected value of the amount you earn after one trial is: $(-1) \frac{37}{38} + (35) \frac{1}{38} = \underline{\underline{-0.0526}}$

A Game of Roulette



https://en.wikipedia.org/wiki/Roulette#/media/File:Roulette_casino.JPG

Expected value of a function of random variable

Consider a function $g(X)$.

The expected value of $g(X)$:

For discrete R.V. (provided the summation is well-defined):

$$E(g(X)) = \sum_i g(x_i)P(X = x_i)$$

For a continuous random variable,

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

Properties of expected value

$$\begin{aligned} E(ag(X) + b) &= \int_{-\infty}^{+\infty} (ag(x) + b) f_x(x) dx \\ &= \int_{-\infty}^{+\infty} ag(x) f_x(x) dx + \int_{-\infty}^{+\infty} bf_x(x) dx \\ &= aE(g(X)) + b \quad -- why? \end{aligned}$$

This property is called the **linearity** of the expected value. In general, a function $f(x)$ is said to be **linear** in x if $f(ax+b) = af(x)+f(b)$ where a and b are constants. In this case, the expected value is not a function but an operator (it takes a function as input). An operator E is said to be linear if $E(af(x) + b) = aE(f(x)) + E(b)$. This is equal to $aE(f(x)) + b$ for the expectation operator.

Properties of expected value

Consider a set of random variables X_1, X_2, \dots, X_n ; a set of functions g_1, g_2, \dots, g_n . Then we have:

$$E\left(\sum_{i=1}^n a_i g_i(X_i) + b_i\right) = \sum_{i=1}^n (a_i E[g_i(X_i)] + b_i)$$

a_i, b_i are scalars

$$= E(g(x_1)^2) \neq E(g(x))^2$$

- Note: for a general nonlinear function g , we have:

$$\underline{E(g(X)) \neq g(E(X))}$$

What if you have to guess the value of a R.V.?

Suppose you want to predict the value of a random variable with a known mean. On an average, what value will yield the least squared error?

$$X \sim P(X = k)$$

Goal: guess a value c s.t.

$$\min E[(X - c)^2]$$

To prove that at $c = \underbrace{E[X]}_{\mu}$ the above error is minimized.

$$\begin{aligned} E[(X - c)^2] &= E[(x - c + \mu - \mu)^2] \\ &= E[(x - \mu)^2 + (c - \mu)^2 - 2(x - \mu)(c - \mu)] \\ &= E[(x - \mu)^2] + E[(c - \mu)^2] - 2(c - \mu)E(x - \mu) \\ &= E(x - \mu)^2 + (c - \mu)^2 \end{aligned}$$

Variance

- The variance of a random variable X tells you how much its values deviate from the mean – on an average.
- The definition of variance for a continuous r.v. with mean μ is:

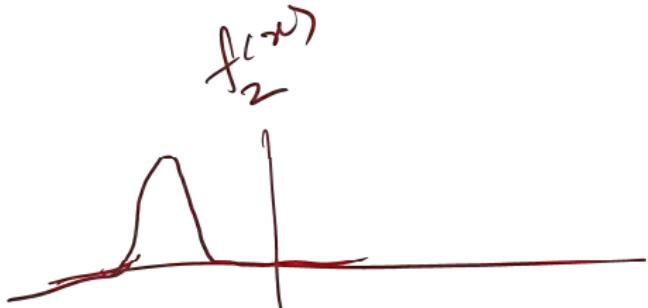
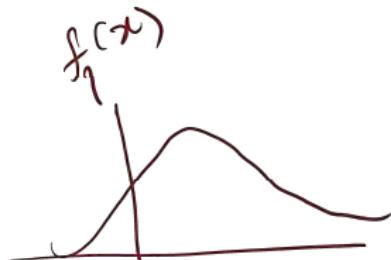
$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

- For a discrete r.v., the integration is replaced by a summation:

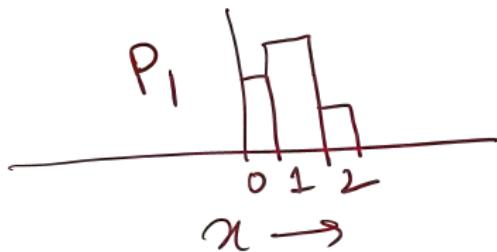
$$\text{Var}(X) = E[(X - \mu)^2] = \sum (x_i - \mu)^2 P(X = x_i)$$

- The positive square-root of the variance is called the standard deviation.
- Low-variance probability mass functions or probability densities tend to be concentrated around one point. High variance densities are spread out.

Variance: examples



$$\text{Var}(f_1) > \text{Var}(f_2)$$



$$\text{Var}(P_2) > \text{Var}(P_1)$$

The Simplest Random Variable

- Bernoulli Random Variable

$$X \in \{0, 1\}$$

$$\text{PMF of } X \quad P(X=1) = \theta$$

$$P(X=x) = \theta^x (1-\theta)^{1-x}$$

$$E[X] = 0 \cdot (1-\theta) + 1 \cdot \theta = \theta$$

$$\begin{aligned} V(X) &= (0-\theta)^2 (1-\theta) + (1-\theta)^2 \cdot \theta \\ &= \theta(1-\theta) \end{aligned}$$

Lecture 08-discreteRVs.pdf

Well-known discrete Random Variables.

Many slides from Chris Piech. CS109 in Stanford Univ

The Simplest Random Variable

- Bernoulli Random Variable Boolean R.V.

$$X \in \{0, 1\}$$

Examples:

- coin-toss
- equipment will fail or not
- whether your couch will show up or not -

PMF $P(X=1) = p$

$$P(X=0) = 1-p$$

$$E(X) = \sum_{x \in X} x \cdot P(X=x) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$\begin{aligned} V(X) &= \sum_{x \in X} (x - E(X))^2 P(X=x) = (0-p)^2(1-p) + (1-p)^2 \cdot p \\ &= p(1-p) \end{aligned}$$

Binomial Random Variable

$$X \in \{0, 1, 2, \dots, n\}$$

Imagine flipping a coin n times and counting the number of heads.

1. We will flip a coin n times: n independent trials of the same experiment
2. Each coin flip has a probability p of being heads
3. What we want to model: what is the probability of exactly k heads?

(This isn't really about flipping coins, though.)

Lots of scenarios fit the same description:

- # of 1's in randomly generated in length n bit string
- # of servers working in a large computer cluster
- # of people who vote for one of two candidates in an election
- # of jury members selected from a particular demographic

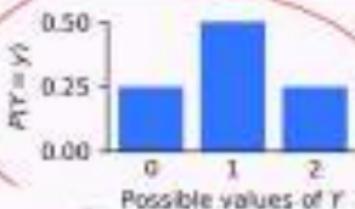
Binomial Random Variable

Imagine flipping a coin n times and counting the number of heads.

1. We will flip a coin n times: n independent trials of the same experiment
2. Each coin flip has a probability p of being heads
3. What we want to model: what is the probability of exactly k heads?

"Let Y be the # of heads in 2 coin flips."

- $P(Y = 0) = 1/4$ (T, T)
- $P(Y = 1) = 1/2$ (H, T), (T, H)
- $P(Y = 2) = 1/4$ (H, H)



This is the binomial for $n = 2$. Can we generalize from this?

Probability of Exactly k Heads in n Coin Flips

To start:

- Let's say we flip the coin 10 times. Probability of heads is p .
- For now, focus on the probability of 4 heads.

What is the probability of the outcome below?

(H, H, H, H, T, T, T, T, T, T)



Probability of Exactly k Heads in n Coin Flips

To start:

- Let's say we flip the coin 10 times. Probability of heads is p .
- For now, focus on the probability of 4 heads.

What is the probability of the outcome below?

(H, H, H, H, T, T, T, T, T, T)

$$p^4(1-p)^6$$

(H, H, H, T, H, T, T, T, T, T) ✓

Probability of Exactly k Heads in n Coin Flips

To start:

- Let's say we flip the coin 10 times. Probability of heads is p .
- For now, focus on the probability of 4 heads.

What is the probability of the outcome below?

(H, H, H, H, T, T, T, T, T, T)

$$p^4(1-p)^6$$

(H, H, H, T, H, T, T, T, T, T)

$$\underline{p^4(1-p)^6}$$

All of the outcomes with exactly 4 heads have the same probability

Probability of Exactly k Heads in n Coin Flips

H, H, H, H, T, T, T, T, T, T,
H, H, H, T, H, T, T, T, T, T,
H, H, H, T, T, H, T, T, T, T,
H, H, H, T, T, T, H, T, T, T,
H, H, H, T, T, T, T, H, T, T,
H, H, H, T, T, T, T, T, H, T,
H, H, H, T, T, T, T, T, T, H,
H, H, H, T, T, T, T, T, T, T,
H, H, H, T, H, H, T, T, T, T,
H, H, H, T, H, T, H, T, T, T,
H, H, H, T, H, T, T, H, T, T,
H, H, H, T, H, T, T, T, H, T,
H, H, H, T, H, T, T, T, T, H,
H, H, H, T, H, T, T, T, T, T,
H, H, H, T, T, H, H, T, T, T,
H, H, H, T, T, H, T, H, T, T,
H, H, H, T, T, H, T, T, H, T,
H, H, H, T, T, H, T, T, T, H,
H, H, H, T, T, H, T, T, T, T,

Then, the probability of getting k heads in any ordering is the "or" of all of these **mutually exclusive** cases

How many cases are there?

Each outcome has probability $p^k(1 - p)^{10-k}$

Probability of Exactly k Heads in n Coin Flips

HH, H, H, H, T, T, T, T, T, T
OH, H, H, T, H, T, T, T, T, T
OH, H, H, T, T, H, T, T, T, T
OH, H, H, T, T, T, H, T, T, T
OH, H, H, T, T, T, T, H, T, T
OH, H, H, T, T, T, T, T, H, T
OH, H, H, T, T, T, T, T, T, H
OH, H, H, T, T, T, T, T, T, T, H
OH, H, T, H, H, T, T, T, T, T
OH, H, T, H, T, H, T, T, T, T
OH, H, T, H, Y, T, H, T, T, T
OH, H, T, H, Y, T, T, H, T, T
OH, H, T, H, Y, T, T, T, H, T
OH, H, T, H, Y, T, T, T, T, H
OH, H, T, T, H, H, T, T, T, T
OH, H, T, T, H, T, H, T, T, T
OH, H, T, T, H, T, T, H, T, T
OH, H, T, T, H, T, T, T, H, T
OH, H, T, T, H, T, T, T, T, H

Then, the probability of getting k heads in
any ordering is the "or" of all of these
mutually exclusive cases

How many cases are there?

$$\binom{10}{k}$$

Each outcome has probability $p^k(1-p)^{10-k}$

Probability of Exactly k Heads in n Coin Flips

HH, H, H, H, T, T, T, T, T, T
OH, H, H, T, H, T, T, T, T, T
OH, H, H, T, T, H, T, T, T, T
OH, H, H, T, T, T, H, T, T, T
OH, H, H, T, T, T, T, H, T, T
OH, H, H, T, T, T, T, T, H, T
OH, H, H, T, T, T, T, T, T, H
OH, H, H, T, T, T, T, T, T, T
OH, H, T, H, H, T, T, T, T, T
OH, H, T, H, T, H, T, T, T, T
OH, H, T, H, T, H, T, T, T, T
OH, H, T, H, Y, T, H, T, T, T
OH, H, T, H, Y, T, T, H, T, T
OH, H, T, H, Y, T, T, T, H, T
OH, H, T, H, Y, T, T, T, T, H
OH, H, T, T, H, H, H, T, T, T, T
OH, H, T, T, H, T, H, T, T, T, T
OH, H, T, T, H, T, H, T, T, H, T
OH, H, T, T, H, T, T, H, T, T, T
OH, H, T, T, H, T, T, T, H, T, T
OH, H, T, T, H, T, T, T, T, H, T

Then, the probability of getting k heads in
any ordering is the "or" of all of these
mutually exclusive cases

How many cases are there?

$$\binom{10}{k}$$

Each outcome has probability $p^k(1-p)^{10-k}$

$$P(k \text{ heads}) = \binom{10}{k} p^k (1-p)^{10-k}$$

We Have Invented The Binomial



This type of random variable is so common it needs a name so that I can talk about it generally.

*I shall call it: the **Binomial** Random Variable. Huzzah.*

Jacob "James" Bernoulli (1654-1705); Swiss mathematician

One of many mathematicians in the Bernoulli family

Declaring a Random Variable to be Binomial

$$X \sim \text{Bin}(n, p)$$

Our random variable

Num trials

Probability of success on each trial

Is distributed as a

Binomial

With these parameters

The diagram illustrates the declaration of a random variable X as being distributed according to a binomial distribution with parameters n and p . The expression $X \sim \text{Bin}(n, p)$ is shown in the center. Several blue arrows point from surrounding text to specific parts of the equation:

- An arrow from "Our random variable" points to the variable X .
- An arrow from "Num trials" points to the parameter n .
- An arrow from "Probability of success on each trial" points to the parameter p .
- An arrow from "Is distributed as a" points to the word "Binomial".
- An arrow from "With these parameters" points to a red oval that encloses the parameters n and p .

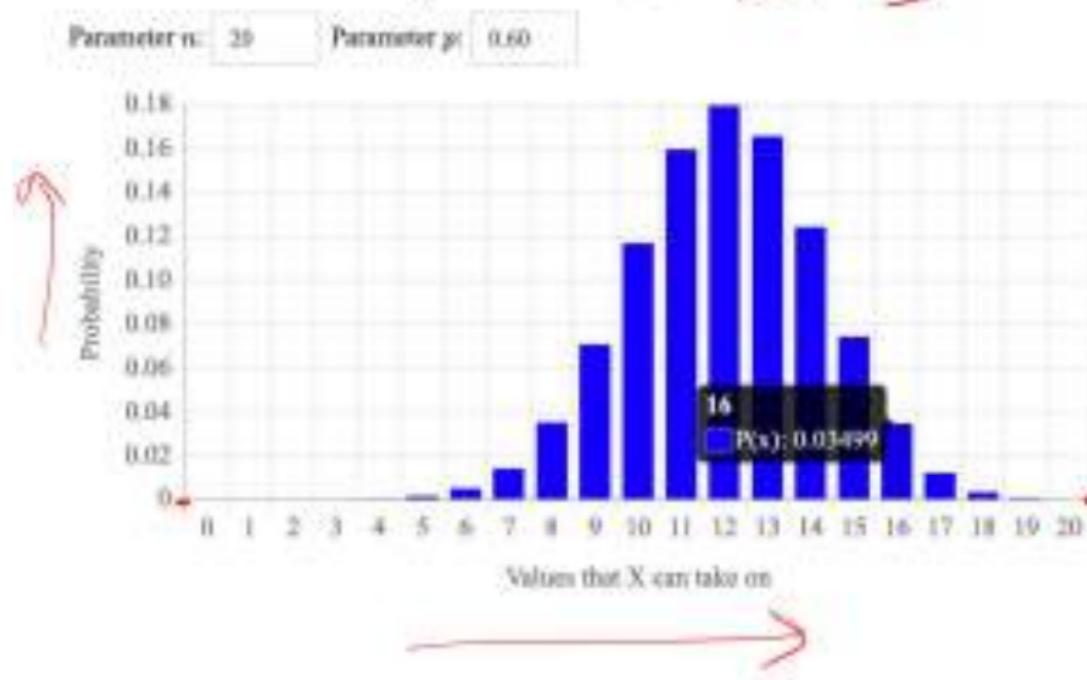
Then We Automatically Know the PMF!

Probability Mass Function
for a Binomial

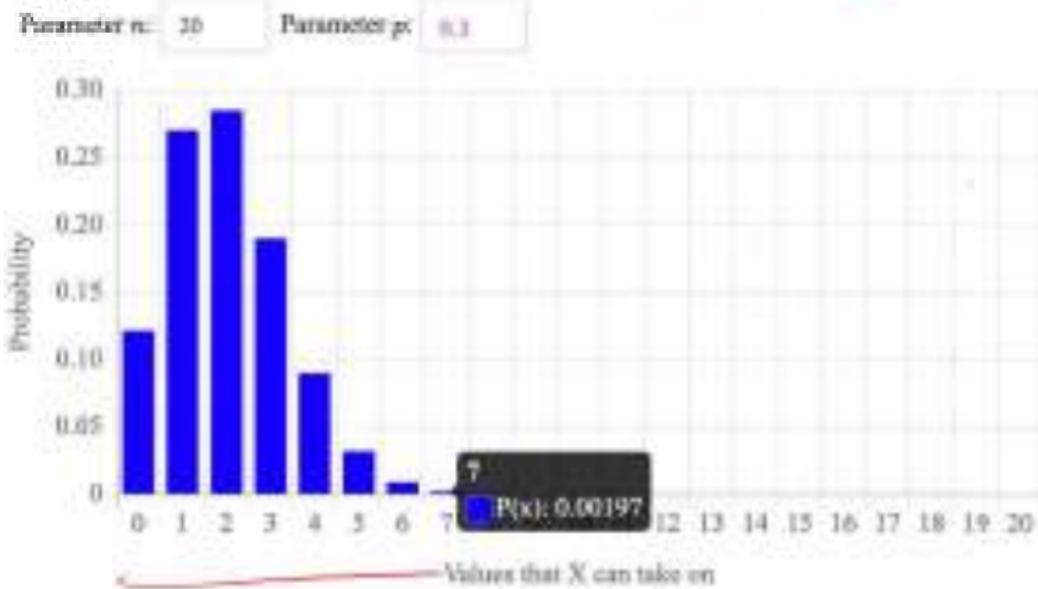
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

↑
Probability that our
variable takes on the
value k

The PMF as a Graph: $X \sim \text{Bin}(n = 20, p = 0.6)$

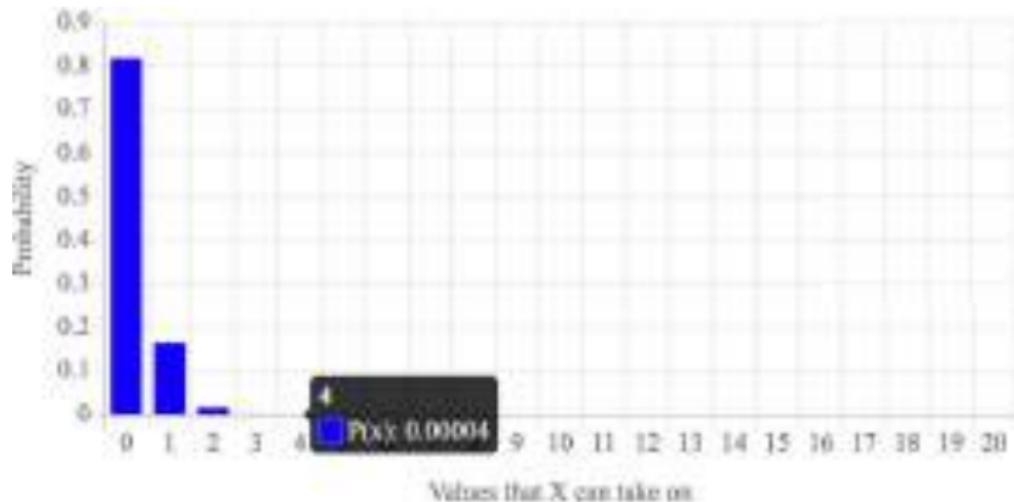


The PMF as a Graph: $X \sim \text{Bin}(n = 20, p = 0.1)$



The PMF as a Graph: $X \sim \text{Bin}(n = 20, p = 0.01)$

Parameter $n:$ Parameter $p:$



Probability of k Heads In n Flips: Now With Binomial

Three fair ($p = 0.5$ of heads) coins are flipped.

Let X be the number of heads.

$$X \sim \text{Bin}(n = 3, p = 0.5)$$

Probability of k Heads In n Flips: Now With Binomial

Three fair ($p = 0.5$ of heads) coins are flipped.

Let X be the number of heads.

$$X \sim \text{Bin}(n = 3, p = 0.5)$$

What is the probability of...

... 0 heads?

... 1 heads?

... 2 heads?

... 3 heads?

Probability of k Heads In n Flips: Now With Binomial

Three fair ($p = 0.5$ of heads) coins are flipped.

Let X be the number of heads.

$$X \sim \text{Bin}(n=3, p=0.5)$$

What is the probability of...

... 0 heads?

$$P(X=0) = \binom{3}{0} p^0 (1-p)^3 = \frac{1}{8}$$

... 1 heads?

$$P(X=1) = \binom{3}{1} p^1 (1-p)^2 = \frac{3}{8}$$

... 2 heads?

$$P(X=2) = \binom{3}{2} p^2 (1-p)^1 = \frac{3}{8}$$

... 3 heads?

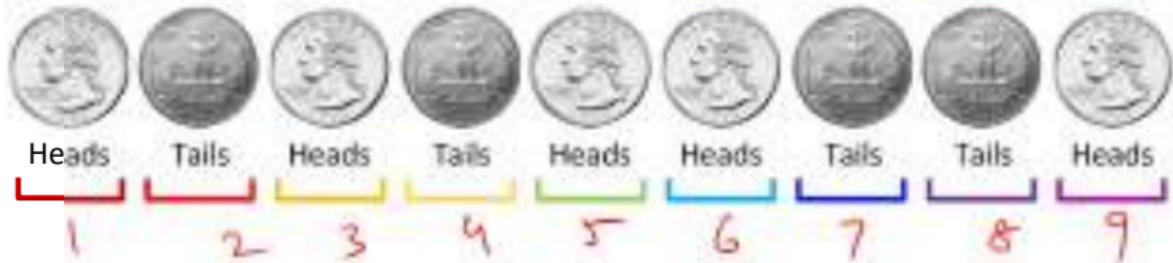
$$P(X=3) = \binom{3}{3} p^3 (1-p)^0 = \frac{1}{8}$$

Random Variable Sums

$$n=9$$

The Binomial

...is a sum of Bernoulli random variables



Random Variable Sums

The Binomial

...is a sum of Bernoulli random variables



Let $X_1 \sim \text{Bern}(p = 1/2)$ and $X_2 \sim \text{Bern}(p = 1/2)$.

$$Y \sim \text{Bin}(n = 2, p = 1/2)$$

$$Y = X_1 + X_2$$

We Can Now Calculate Expectation of Binomial

$$X \sim \text{Bin}(n, p)$$

Let Y_i be 1 if trial i was a success, otherwise 0, with i from 1 to n . $\underline{Y_i \sim \text{Bern}(p)}$.

The Binomial

...is a sum of Bernoulli random variables



We Can Now Calculate Expectation of Binomial

$$\underline{X \sim \text{Bin}(n, p)}$$

Let Y_i be 1 if trial i was a success, otherwise 0, with i from 1 to n . $\underline{Y_i \sim \text{Bern}(p)}$.

$$\underline{\mathbb{E}[X] = \mathbb{E} \left[\sum_{i=1}^n Y_i \right]} \quad \text{Since } \underline{X = \sum_{i=1}^n Y_i}$$

We Can Now Calculate Expectation of Binomial

$$X \sim \text{Bin}(n, p)$$

Let Y_i be 1 if trial i was a success, otherwise 0, with i from 1 to n . $Y_i \sim \text{Bern}(p)$.

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^n Y_i\right] && \text{Since } X = \sum_{i=1}^n Y_i \\ &= \sum_{i=1}^n E[Y_i] && \text{Expectation of sum} \end{aligned}$$

Expectation of a sum is the sum of expectations: $E[X + Y] = E[X] + E[Y]$

We Can Now Calculate Expectation of Binomial

$$X \sim \text{Bin}(n, p)$$

Let Y_i be 1 if trial i was a success, otherwise 0, with i from 1 to n . $Y_i \sim \text{Bern}(p)$.

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^n Y_i\right] && \text{Since } X = \sum_{i=1}^n Y_i \\ &= \sum_{i=1}^n E[Y_i] && \text{Expectation of sum} \\ &= \sum_{i=1}^n p && \text{Expectation of Bernoulli} \\ &= n \cdot p && \text{Sum } n \text{ times} \end{aligned}$$

We Can Now Calculate Expectation of Binomial

$$X \sim \text{Bin}(n, p)$$

Let Y_i be 1 if trial i was a success, otherwise 0, with i from 1 to n . $Y_i \sim \text{Bern}(p)$.

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^n Y_i\right] && \text{Since } X = \sum_{i=1}^n Y_i \\ &= \sum_{i=1}^n E[Y_i] && \text{Expectation of sum} \\ &= \sum_{i=1}^n p && \text{Expectation of Bernoulli} \\ &= n \cdot p && \text{Sum } n \text{ times} \end{aligned}$$

True for every binomial ever

Variance of Binomial R.V.s

$$X = \sum_{i=1}^n Y_i \quad Y_i \sim \text{Bern}(p)$$

$$\text{Var}(Y_i) = p(1-p)$$

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \text{Var}(Y_i)$$

If $Y_i \perp\!\!\! \perp Y_j \quad \forall i, j$

$$\text{Var}(X) = np(1-p)$$

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting 10 clicks?

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting 10 clicks?

Let X be the number of ad clicks.

$X \sim \text{Bin}(n = 1000, p = 0.01)$.

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting 10 clicks?

Let X be the number of ad clicks. $X \sim \text{Bin}(n = 1000, p = 0.01)$.

$$P(X = k) = \binom{1000}{k} (0.01)^k (0.99)^{1000-k}$$

$$P(X = 10) = \binom{1000}{10} (0.01)^{10} (0.99)^{990} \approx 0.125$$

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting 20 clicks?

Let X be the number of ad clicks. $X \sim \text{Bin}(n = 1000, p = 0.01)$.

$$P(X = k) = \binom{1000}{k} (0.01)^k (0.99)^{1000-k}$$

$$P(X = 20) = \binom{1000}{20} (0.01)^{20} (0.99)^{980} \approx 0.0018$$

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting **20** clicks?

Let X be the number of ad clicks. $X \sim \text{Bin}(n = 1000, p = 0.01)$.

```
>>> from scipy import stats  
>>> stats.binom.pmf(10, 1000, 0.01)  
0.1257402111262075  
>>> stats.binom.pmf(20, 1000, 0.01)  
0.0017918782400182195
```

k ↑ n ↑ p ↗

Practice: Ad Clicks



Every day, YouTube shows a particular ad 1000 times.

Each ad served is clicked with $p = 0.01$ (otherwise it's ignored).

What is the probability of this ad getting **20** clicks?

Let X be the number of ad clicks.

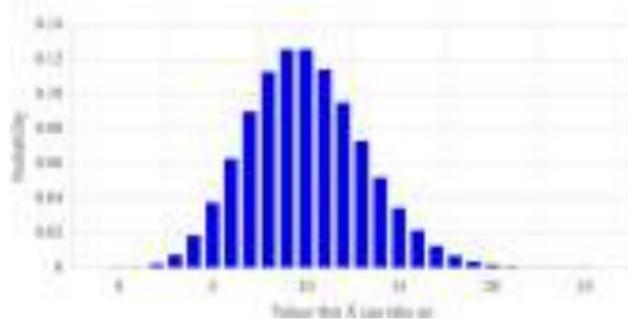
$X \sim \text{Bin}(n = 1000, p = 0.01)$.

PMF graph:

Parameter n: 1000

Parameter p:

0.01 ±



Server Redundancy



A network can remain functional as long as at least 2 out of 7 servers are alive.

The probability of any server working is 0.8.

What is the probability that less than 2 servers are alive?

Server Redundancy



A network can remain functional as long as at least 2 out of 7 servers are alive.

The probability of any server working is 0.8.

What is the probability that less than 2 servers are alive?

Let X be the number of servers alive.

$X \sim \text{Bin}(n = 7, p = 0.8)$.

Server Redundancy



A network can remain functional as long as at least 2 out of 7 servers are alive.

The probability of any server working is 0.8.

What is the probability that less than 2 servers are alive?

Let X be the number of servers alive. $X \sim \text{Bin}(n = 7, p = 0.8)$.

$$P(X = k) = \binom{7}{k} (0.8)^k (0.2)^{7-k}$$

Server Redundancy



A network can remain functional as long as at least 2 out of 7 servers are alive.

The probability of any server working is 0.8.

What is the probability that less than 2 servers are alive?

Let X be the number of servers alive. $X \sim \text{Bin}(n = 7, p = 0.8)$.

$$P(X = k) = \binom{7}{k} (0.8)^k (0.2)^{7-k}$$

$P(X < 2) = P(X = 0) + P(X = 1)$

Server Redundancy



A network can remain functional as long as at least 2 out of 7 servers are alive.

The probability of any server working is 0.8.

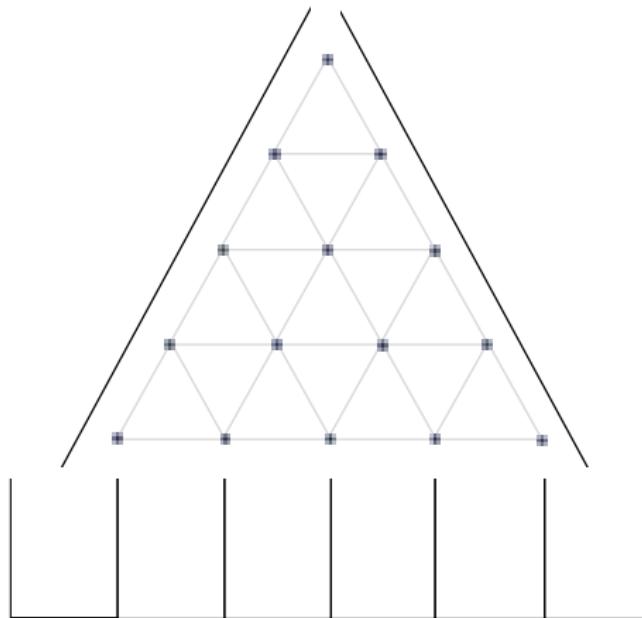
What is the probability that less than 2 servers are alive?

Let X be the number of servers alive. $X \sim \text{Bin}(n = 7, p = 0.8)$.

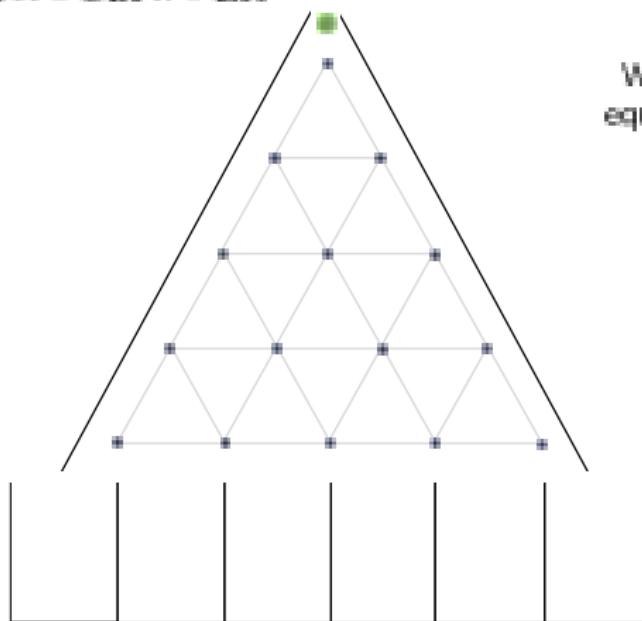
$$P(X = k) = \binom{7}{k} (0.8)^k (0.2)^{7-k}$$

$$P(X < 2) = P(X = 0) + P(X = 1) = \binom{7}{0} (0.8)^0 (0.2)^{7-0} + \binom{7}{1} (0.8)^1 (0.2)^{7-1} \approx 0.0004$$

Galton Board Fun

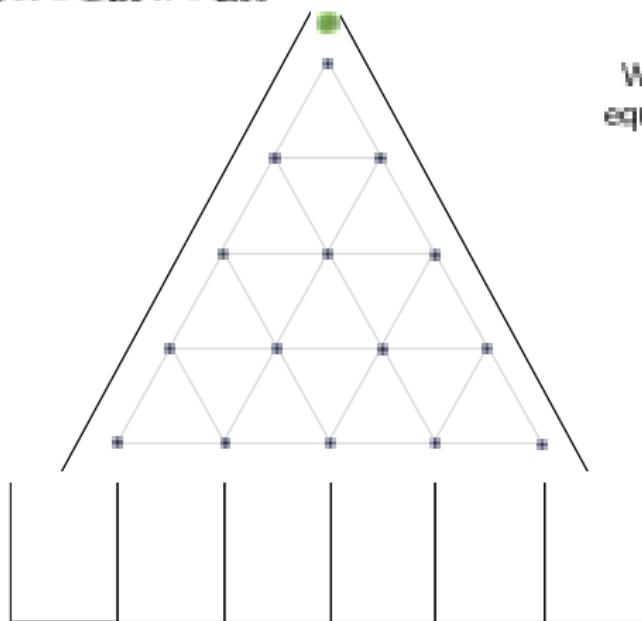


Galton Board Fun



When a marble hits a pin, it has equal chance of going left or right.

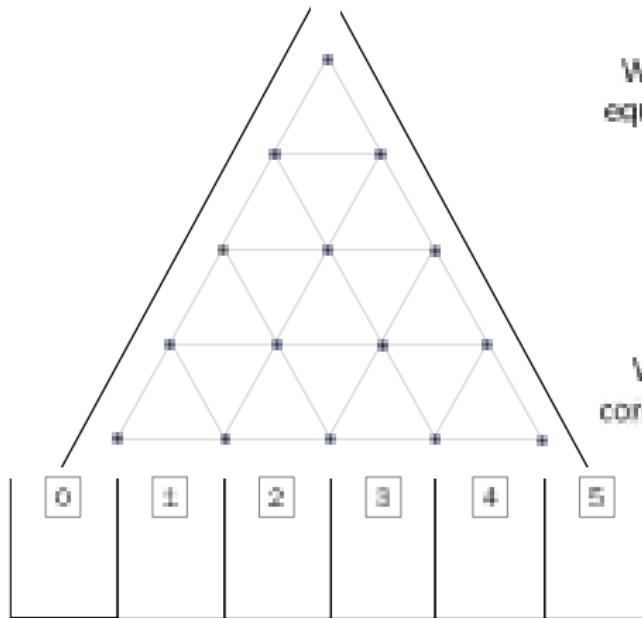
Galton Board Fun



When a marble hits a pin, it has equal chance of going left or right.

Each pin represents an independent event.

Galton Board Fun

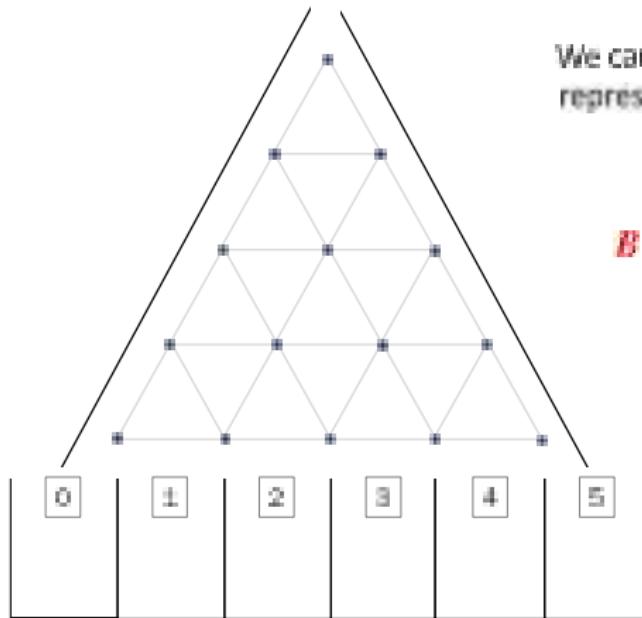


When a marble hits a pin, it has equal chance of going left or right.

Each pin represents an independent event.

Which bucket a marble lands in corresponds to the number of times the marble went right.

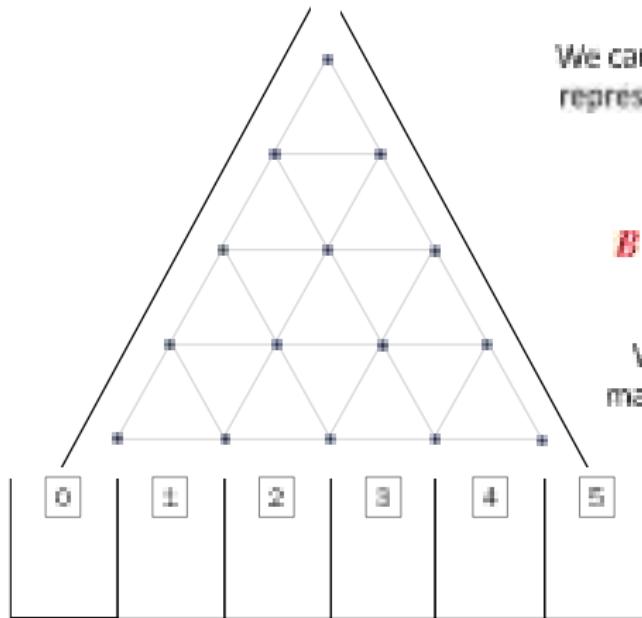
Galton Board Fun



We can define a random variable (B) representing which bucket a marble lands in.

$$B \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

Galton Board Fun

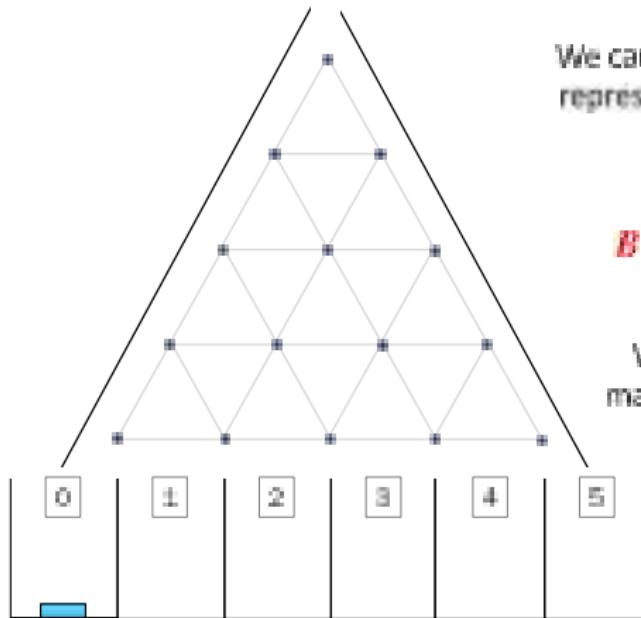


We can define a random variable (R) representing which bucket a marble lands in.

$$R \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

What is the probability of a marble landing in each bucket?

Galton Board Fun



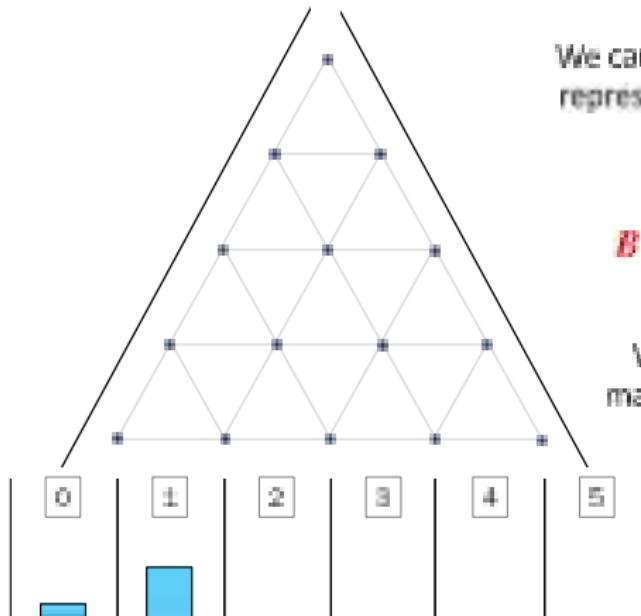
We can define a random variable (B) representing which bucket a marble lands in.

$$B \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

What is the probability of a marble landing in each bucket?

$$P(B=0) = \binom{5}{0} \frac{1}{2}^5 \approx 0.03$$

Galton Board Fun



We can define a random variable (B) representing which bucket a marble lands in.

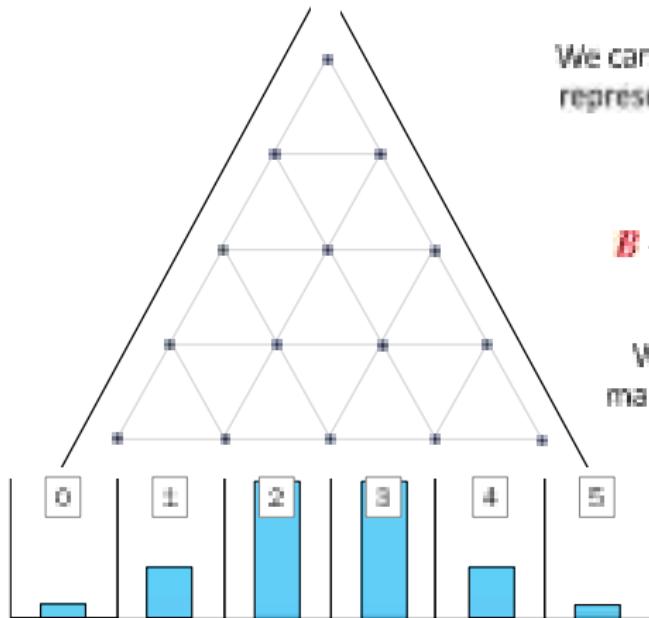
$$B \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

What is the probability of a marble landing in each bucket?

$$P(B=0) = \binom{5}{0} \frac{1}{2}^5 \approx 0.03$$

$$P(B=1) = \binom{5}{1} \frac{1}{2}^5 \approx 0.16$$

Galton Board Fun



We can define a random variable (B) representing which bucket a marble lands in.

$$B \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

What is the probability of a marble landing in each bucket?

This is the PMF of the binomial

The Geometric Random Variable

Imagine flipping a coin until you see your first heads. ~

Each coin flip is an independent trial, with probability p of getting heads.

Want to model: how many coin flips until the first heads?

$$\underline{X} \sim \underline{\text{Geo}}(\underline{p})$$

$$X \in \{1, 2, 3, \dots, \dots, \dots, \infty\}$$

The Geometric Random Variable

Imagine flipping a coin *until you see your first heads.*

Each coin flip is an independent trial, with probability p of getting heads.

Want to model: how many coin flips until the first heads?

$$X \sim \text{Geo}(p)$$

Deriving the PMF:

$$P(\text{heads on first flip}) = \underline{p}$$

$$P(\text{tails, then heads}) = \underline{(1-p) * p}$$

$$P(\text{tails, tails, heads}) = \underline{\underline{(1-p)^2 * p}}$$

/

/

! ...

The Geometric Random Variable

Imagine flipping a coin *until you see your first heads.*

Each coin flip is an independent trial, with probability p of getting heads.

Want to model: how many coin flips until the first heads?

$$X \sim \text{Geo}(p)$$

Deriving the PMF:

$$P(\text{heads on first flip}) = p$$

$$P(\text{tails, then heads}) = (1 - p) * p$$

$$P(\text{tails, tails, heads}) = (1 - p)^2 * p$$

$$P(X = n) = \underline{(1 - p)^{n-1} p}$$

...

The Negative Binomial Random Variable

Imagine flipping a coin *until you see r heads.*

Each coin flip is an independent trial, with probability p of getting heads.

Want to model: how many coin flips until r heads?

$$X \in \{r, r+1, \dots, \infty\}$$
$$P(X=n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

The Negative Binomial Random Variable

Imagine flipping a coin *until you see r heads.*

Each coin flip is an independent trial, with probability p of getting heads.

Want to model: how many coin flips until r heads?

$$X \sim \text{NegBin}(r, p)$$

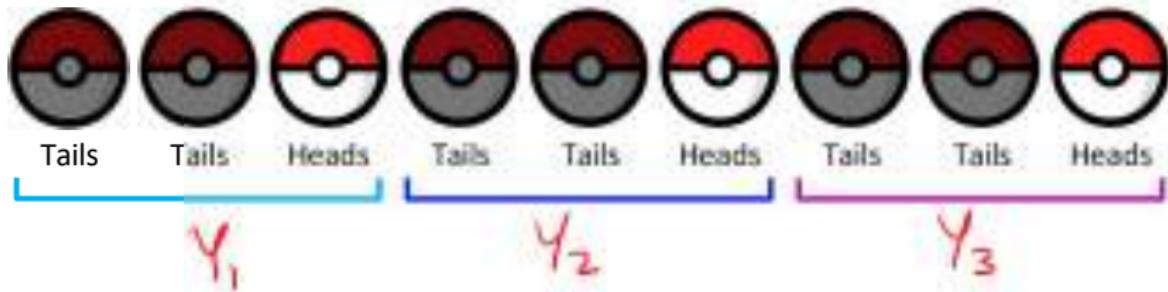
$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Random Variable Sums

The Negative Binomial



$\gamma = 3$

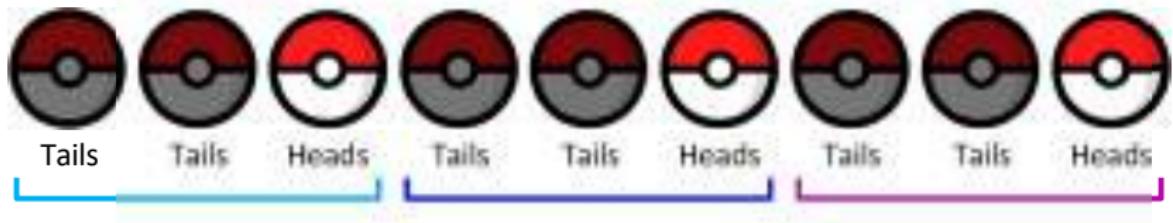


$$X = Y_1 + Y_2 + Y_3$$

Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables



Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables



Let $X_1 \sim \text{Geo}(p = 1/3)$, $X_2 \sim \text{Geo}(p = 1/3)$, and $X_3 \sim \text{Geo}(p = 1/3)$.

Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables



Let $X_1 \sim \text{Geo}(p = 1/3)$, $X_2 \sim \text{Geo}(p = 1/3)$, and $X_3 \sim \text{Geo}(p = 1/3)$.

$$Y \sim \text{NegBin}(r = 3, p = 1/3)$$

Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables



Let $X_1 \sim \text{Geo}(p = 1/3)$, $X_2 \sim \text{Geo}(p = 1/3)$, and $X_3 \sim \text{Geo}(p = 1/3)$.

$$Y \sim \text{NegBin}(r = 3, p = 1/3)$$

$$Y = X_1 + X_2 + X_3$$

Expected Value of The Geometric

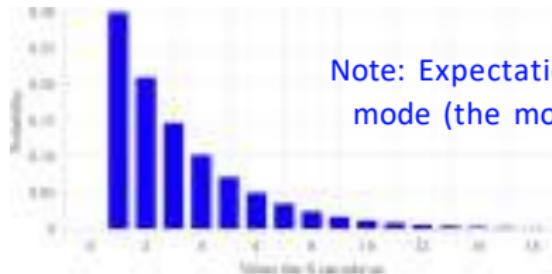
If $X \sim \text{Geo}(p)$, then

$$E[X] = \frac{1}{p}$$



This definition has intuition built in:

- If a coin has probability $\frac{1}{2}$ of a head, then on average, it will take him two tosses to get a head. $E[X] = (1/2)^{-1} = 2$.



Note: Expectation is often **not** the mode (the most likely outcome)

Expected Value of The Geometric

$$E[Y] = \sum_{i=1}^{\infty} n \cdot (\underbrace{(1-p)}_{n-1} \cdot p) = \frac{1}{p}$$

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

The Negative Binomial

...is a sum of Geometric random variables



Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[Y] = E \left[\sum_{i=1}^r X_i \right]$$

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \end{aligned}$$

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \\ &= \sum_{i=1}^r \frac{1}{p} = \frac{r}{p} \end{aligned}$$

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

Let X be your winnings.

$$E[X] = \left(\frac{1}{2}\right)^1 2^1 + \left(\frac{1}{2}\right)^2 2^2 + \left(\frac{1}{2}\right)^3 2^3 + \dots = \sum_{i=0}^{\infty} 1 = \infty$$

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

Let X be your winnings.

$$E[X] = \left(\frac{1}{2}\right)^1 2^1 + \left(\frac{1}{2}\right)^2 2^2 + \left(\frac{1}{2}\right)^3 2^3 + \dots = \sum_{i=0}^{\infty} 1 = \infty$$

What if you could play this game for only \$1000...but just once?

Expectations of Classic Random Variables

$X \sim \text{Geo}(p)$

$$E[X] = \frac{1}{p}$$

$X \sim \text{Bern}(p)$

$$E[X] = p$$

$Y \sim \text{NegBin}(r, p)$

$$E[Y] = \frac{r}{p}$$

$Y \sim \text{Bin}(n, p)$

$$E[Y] = n \cdot p$$

Variance of Classic Random Variables

$X \sim \text{Geo}(p)$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

$X \sim \text{Bern}(p)$

$$\text{Var}(X) = p(1-p)$$

$Y \sim \text{NegBin}(r, p)$

$$\text{Var}(X) = \frac{r \cdot (1-p)}{p^2}$$

$Y \sim \text{Bin}(n, p)$

$$\text{Var}(Y) = n \cdot p(1-p)$$

Lecture 09-discreteRV.pdf

Expected Value of The Geometric

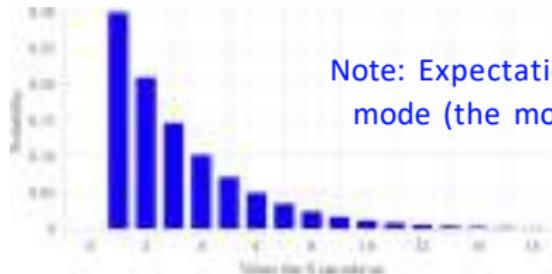
If $X \sim \text{Geo}(p)$, then

$$E[X] = \frac{1}{p}$$



This definition has intuition built in:

- If a coin has probability $\frac{1}{2}$ of a head, then on average, it will take him two tosses to get a head. $E[X] = (1/2)^{-1} = 2$.



Note: Expectation is often **not** the mode (the most likely outcome)

Expected Value of The Geometric

$$E[Y] = \sum_{\substack{n=1 \\ n \neq 1}}^{\infty} n \cdot \underbrace{(1-p)^{n-1} \cdot p}_{\text{Probability of success}} = \frac{1}{p}$$

$$= 1 \cdot p + 2(1-p)p + 3(1-p)^2 p + 4(1-p)^3 p$$

$$= p \left(1 + \overset{+}{2(1-p)} + \overset{-}{3(1-p)^2} + \overset{-}{\dots} \right) = \cancel{Sp}$$

$$= p \left(\cancel{(1-p)} + 2(1-p)^2 + 3(1-p)^3 + \dots \right) =$$

=

Recall SEE math --

Expected Value of The Negative Binomial

We can derive using the sum of expectations property, similar to binomials.

The Negative Binomial

...is a sum of Geometric random variables



Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$\underline{E[X_i]} = \frac{1}{\underline{p}}$$

Let $\underline{Y} \sim \text{NegBin}(r, p)$.

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

$$E[Y] = E\left[\sum_{i=1}^r X_i\right]$$

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \end{aligned}$$

Expected Value of The Negative Binomial

We can derive using the **sum of expectations** property, similar to binomials.

Let $X_i \sim \text{Geo}(p)$, for each i from 1 to r .

$$E[X_i] = \frac{1}{p}$$

Let $Y \sim \text{NegBin}(r, p)$.

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \\ &= \sum_{i=1}^r \frac{1}{p} = \frac{r}{p} \end{aligned}$$

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

$$E[g(X)] = \sum_{i=0}^{\infty} g(i) \cdot P(X=i)$$

Let X be your winnings. $g(x) = 2^x$

$$E[g(X)] = \left(\frac{1}{2}\right)^1 2^1 + \left(\frac{1}{2}\right)^2 2^2 + \left(\frac{1}{2}\right)^3 2^3 + \dots = \sum_{i=0}^{\infty} 1 = \infty$$

St. Petersburg Paradox

The Game:

- We have a fair coin (lands on heads with $p = 0.5$)
- Let n = number of coin flips (tails) to get the first heads
- You will win: $\$2^n$

How much would you pay to play?

Let X be your winnings.

$$E[X] = \left(\frac{1}{2}\right)^1 2^1 + \left(\frac{1}{2}\right)^2 2^2 + \left(\frac{1}{2}\right)^3 2^3 + \dots = \sum_{i=0}^{\infty} 1 = \infty$$

What if you could play this game for only \$1000...but just once?

Expectations of Classic Random Variables

$X \in \{1, 2, \dots, \infty\}$

$$X \sim \text{Geo}(p)$$
$$P(X=n) = (1-p)^{n-1} p$$
$$E[X] = \frac{1}{p}$$

$Y \in \{0, 1, 2, \dots\}$

$$Y \sim \text{NegBin}(r, p)$$
$$P(Y=k) = \binom{r+k-1}{k} (1-p)^k p^r$$
$$E[Y] = \frac{r}{p}$$

$$Y = \sum_{i=1}^r X_i \quad X_i \sim \text{Geo}(p)$$

$X \in \{0, 1\}$

$$X \sim \text{Bern}(p)$$
$$P(X=x) = p^x (1-p)^{1-x}$$
$$E[X] = p$$

$Y \sim \text{Bin}(n, p) \quad Y \in \{0, 1, \dots, n\}$

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k}$$
$$E[Y] = n \cdot p$$

$$Y = \sum_{i=1}^n X_i \quad X_i \sim \text{Bern}(p)$$

Variance of Classic Random Variables

$X \sim \text{Geo}(p)$

$$\underline{\text{Var}}(X) = \frac{1-p}{p^2}$$

$X \sim \text{Bern}(p)$

$$\underline{\text{Var}}(X) = p(1-p)$$

$Y \sim \underline{\text{NegBin}}(r, p)$

$$\underline{\text{Var}}(X) = \frac{r \cdot (1-p)}{p^2}$$

$Y \sim \underline{\text{Bin}}(n, p)$

$$\underline{\text{Var}}(Y) = n \cdot p(1-p)$$

Lecture 10-poisson.pdf

Poisson Random Variable

Expected # of autos in an hour = 10

Time interval = 5 minutes.

Probability that one auto will come
in the next 5 minutes.
improve approximation

$$\approx 1 - \frac{\left(\frac{5}{10}\right)^5}{\binom{60}{10}}$$

$$1 - \frac{\left(\frac{3600 - 5 \times 60}{10}\right)^5}{\binom{3600}{10}}$$



3600



5 min ≈ 300

$$P = \frac{10}{3600} \times \# \text{ of contours in } \rightarrow$$

$$n = 300$$

$$\begin{aligned} P(X \geq k) &= 1 - P(X = 0) \\ &= 1 - \binom{300}{k} (1-P)^{300-k} P^k \\ k=0 &= 1 - (1-P)^{300} = 1 - \left(1 - \frac{10}{3600}\right)^{300} \approx 0.57. \end{aligned}$$

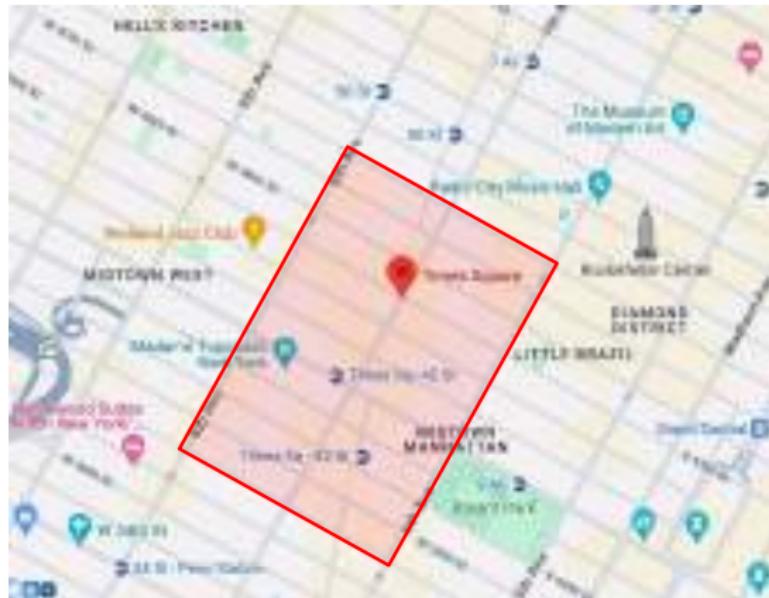
Situations from Poisson R.V is useful

- In all four discrete R.V.s so far (Bernoulli, Binomial, Geometric, Negative binomial), we were counting some outcome from a set of possible discrete options.
 - Multiple dice rolls
 - Servers in operation
 - View of ads on YouTube.
- In many real-life applications, the substrate is continuous, example time.
 - We are counting outcomes of interest in this continuous space.

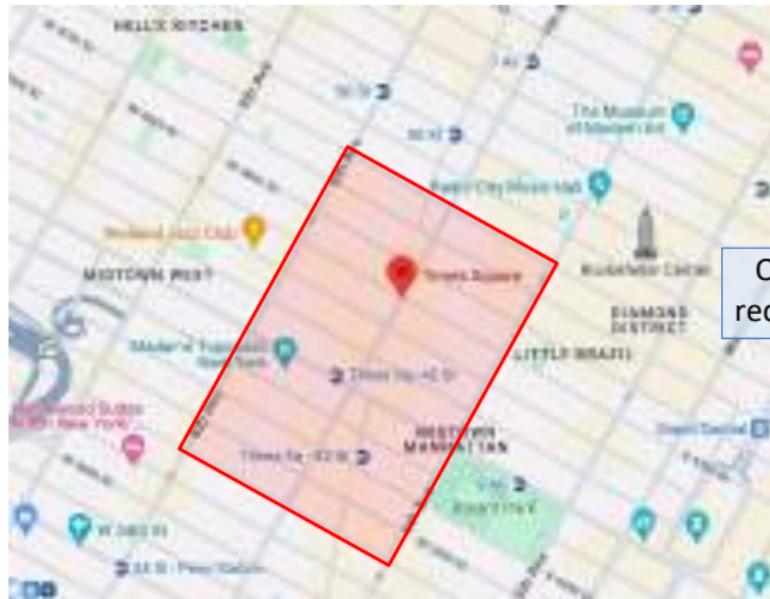
Case Study: Ride Sharing Apps



Probability of k Requests From This Area Next Minute



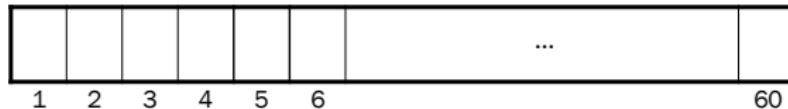
Probability of ***k*** Requests From This Area Next Minute



On average, $\lambda = 5$
requests per minute

Probability of ***k*** Requests From This Area Next Minute

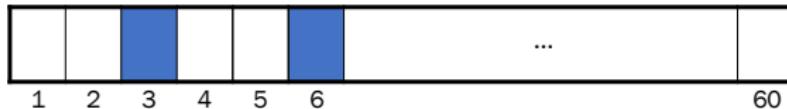
Idea: we can break a minute down into 60 seconds...



On average, $\lambda = 5$
requests per minute

Probability of ***k*** Requests From This Area Next Minute

Idea: we can break a minute down into 60 seconds...



At each second, you either get a request or don't.

On average, $\lambda = 5$
requests per minute

Probability of ***k*** Requests From This Area Each Minute

Idea: we can break a minute down into 60 seconds...



At each second, you either get a request or don't.

Let X be the number of requests in a minute.

On average, $\lambda = 5$
requests per minute

$$X \sim \text{Bin}(n = 60, p = ?)$$

Probability of ***k*** Requests From This Area Each Minute

Idea: we can break a minute down into 60 seconds...



At each second, you either get a request or don't.

Let X be the number of requests in a minute.

On average, $\lambda = 5$
requests per minute

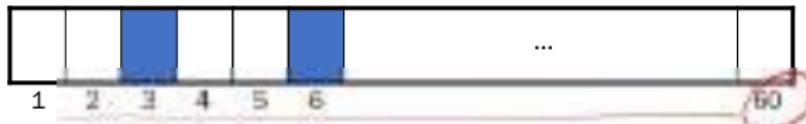
$$X \sim \text{Bin}(n = 60, p = 5/60)$$

$$p = \frac{\lambda}{n}$$

$$P(X = 3) = \binom{60}{3} (5/60)^3 (1 - 5/60)^{57}$$

Probability of k Requests From This Area Each Minute

Idea: we can break a minute down into 60 seconds...



At each second, you either get a request or don't.

Let X be the number of requests in a minute.

On average, $\lambda = 5$ requests per minute

$$X \sim \text{Bin}(n = 60, p = 5/60)$$

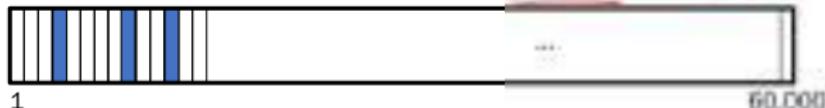
$$p = \frac{\lambda}{n}$$

$$P(X = 3) = \binom{60}{3} (5/60)^3 (1 - 5/60)^{57}$$

But what if there are two requests in the same second?

Probability of k Requests From This Area Each Minute

Idea: we can break a minute down into 60,000 milliseconds...

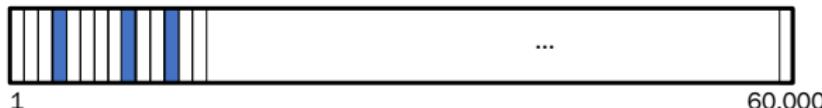


At each ms, you either get a request or don't.
Let X be the number of requests in a minute.

On average, $\lambda = 5$
requests per minute

Probability of k Requests From This Area Each Minute

Idea: we can break a minute down into 60,000 milliseconds...



At each ms, you either get a request or don't.
Let X be the number of requests in a minute.

On average, $\lambda = 5$
requests per minute

$$X \sim \text{Bin}(n = 60000, p = \lambda/n)$$
$$P(X = k) = \binom{60000}{k} \left(\frac{\lambda}{60000}\right)^k \left(1 - \frac{\lambda}{60000}\right)^{n-k}$$
$$p = \frac{\lambda}{n}$$

Can we do even better?

Probability of k Requests From This Area Each Minute

Idea: we can break a minute down into *infinitely small* buckets

too small to draw ®

1

In each bucket, you either get a request or don't.
Let X be the number of requests in a minute.

On average, $\lambda = 5$
requests per minute

$$X \sim \text{Bin}(n = \infty, p = \lambda/n)$$

$$p = \frac{\lambda}{n}$$

$$\lim_{n \rightarrow \infty} P(X = k) = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Probability of **k Requests** From This Area Each Minute

$$\begin{aligned} P(X = k) &= \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \cancel{\lambda^k} \cancel{\frac{1}{k!}} \lim_{n \rightarrow \infty} \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \\ &= \frac{\lambda^k}{k!} \underbrace{\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n}_{e^{-\lambda}} \\ &= \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \boxed{\frac{e^{-\lambda} \lambda^k}{k!}} \end{aligned}$$

The Poisson Random Variable

A **Poisson** random variable models the number of occurrences that happen in a fixed interval of time.

$$X \sim \text{Poi}(\lambda)$$

PMF:

$$P(X = k) = e^{-\lambda} \lambda^k / k!$$

X takes on values 0, 1, 2...up to infinity.

Simeon-Denis Poisson

Prolific French mathematician (1781-1840)

He published his first paper at 18?

Became a professor at 21???

And published over 300 papers in his life?????



He reportedly said, *“Life is good for only two things: discovering mathematics and teaching mathematics.”*

Problem Solving with The Poisson

Say you want to model events occurring over a given time interval.

- Earthquakes, radioactive decay, queries to a web server, etc.

The events you're modeling must follow a **Poisson Process**:

- 1. Events happen *independently* of one another
- 2. Events arrive at a fixed rate: λ events per interval of time

If those conditions are met:

Let X be the number of events that happen in the time interval.

$$X \sim \text{Poi}(\lambda)$$

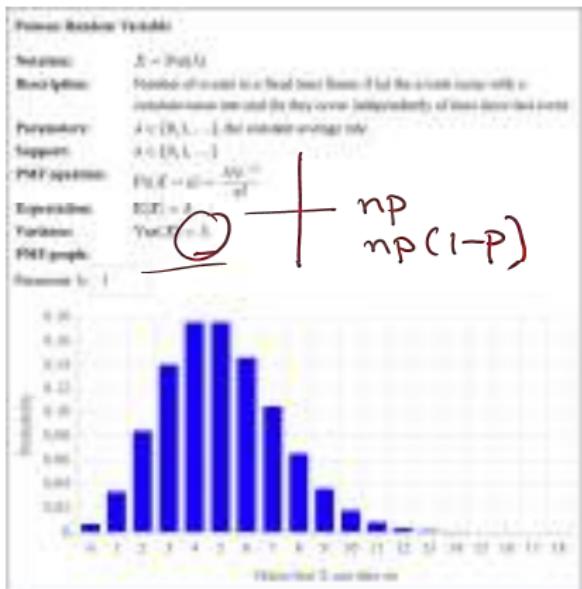
Is Lambda All You Need? Yes

Let X be the number of Uber requests from Powai each minute.

$$X \sim \text{Poi}(\lambda = 5)$$

Calculate $E[X]$, $\text{Var}(X)$

First calculate moment generating function of X .



Moment Generating Function

Expected value of a special function that will make it easy to calculate mean and variance of several random variables.

Let X be a random variable, and $P(X)$ be its pmf or density function.

Recall $E[g(X)] = \sum_{x \in X} g(x) P(x)$

Let $g(X) = e^{tX}$, $\phi(t) = E[e^{tX}] = \sum_{x \in X} e^{tx} p(x)$

For many special random variables, $\phi(t)$ can be calculated in closed form.

Moment Generating Function

$$\underline{\phi(t)}$$

$$E_p(x) = \sum_x x p(x)$$

$$\phi(t) = \sum_x e^{tx} p(x)$$

$$\frac{\partial \phi(t)}{\partial t} = \frac{\partial}{\partial t} \sum_x e^{tx} p(x) = \sum_x x e^{tx} p(x) \Big|_{t=0} = \sum_x x p(x)$$

$$\phi'(t) \Big|_{t=0} = E(x)$$

$$\boxed{\phi''(t) \Big|_{t=0} = E(x^2)}$$

MGF of Poisson distribution

$$\phi(t) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$
$$= \lambda$$

$$= e^{-\lambda} \left[\sum_{x=0}^{\infty} \frac{(e^t)^x}{x!} \right] = e^{-\lambda} e^{t\lambda}$$

$$\phi(t) = e^{-\lambda(1-e^t)}$$

$$\phi'(t) = e^{-\lambda(1-e^t)} \cdot \lambda e^t \Big|_{t=0}$$

$$\phi''(t) = \frac{\partial^2}{\partial t^2} \left[e^{-\lambda} e^{\lambda + \lambda e^t + t} \right] = \lambda e^{-\lambda + \lambda e^t + t} (\lambda e^t + 1) \Big|_{t=0}$$
$$= \lambda + \lambda^2 = E(x^2)$$

Example: Earthquakes



Bulletin of the Seismological Society of America

Vol. 64

October 1974

No. 3

IN THE SEQUENCE OF EARTHQUAKES IN SOUTHERN CALIFORNIA,
WITH AFTERSHOCKS REMOVED. POISSONIAN?

By J. R. Gephart and L. Keefer

Abstract

None

Earthquakes

Let X be the number of earthquakes that happen in California every year.

Here's the PMF for X :

$$P(X = x) = \frac{69^x e^{-69}}{x!}$$

X is a Poisson!
What is $E[X]$ (11)

What is the probability that there are 60 earthquakes in California next year?

$$P(X = 60) = \frac{69^{60} e^{-69}}{60!} \approx 0.028$$

Just plug numbers into the PMF!

Practice: Web Server Load

Historically, a particular web server averages 120 requests each minute.

Let X be the number of hits this server receives in a second. What is $P[X < 5]$?



$$\lambda = \frac{120}{60} = 2 \text{ average # of requests per second}$$
$$P(X < 5) = \sum_{x=0}^4 \frac{\lambda^x e^{-\lambda}}{x!}$$



Practice: Web Server Load

Historically, a particular web server averages 120 requests each **minute**.

Let X be the number of hits this server receives in a **second**. What is $P(X < 5)$?

$$X \leftarrow \text{Poi}(\lambda = 2)$$



The Poisson approximates the Binomial when n is large

Storing Data in DNA: Super Promising Technology



The amount of data contained
in ~ 600 smartphones
(10,000 gigabytes) can be
stored in just the faint pink
smear of DNA at the end of
this test tube.

https://en.wikipedia.org/wiki/DNA_digital_data_storage#:~:text=DNA%20digital%20data%20storage%20is,slow%20read%20and%20write%20times.

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^8$ positions.

What's the probability that $< 1\%$ of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$\sum_{k=0}^{10^8} \binom{n}{k} p^k (1-p)^{n-k}$$

But the PMF for this would be unwieldy to compute :/

There are lots of cases where extreme n and p values arise:

- Errors sending streams of bits over an imperfect network
- Server crashes per day in giant data center

Approximating with Poisson

Let X be the number of corrupted positions.

$$X \sim \text{Poi}(\lambda = 10^8 * 10^{-6} = 100)$$

$$P(X < 0.01 \cdot 10^8) = P(X < 10^6) = \sum_{k=0}^{10^6-1} P(X = k) = \sum_{k=0}^{10^6-1} \frac{100^k \cdot e^{-100}}{k!}$$

Approximating Binomial With Poisson: General Rule

The Poisson approximates the Binomial well when:

1. n is large ✓
2. p is small ✓
3. Therefore, $\lambda = np$ is "moderate" ↗

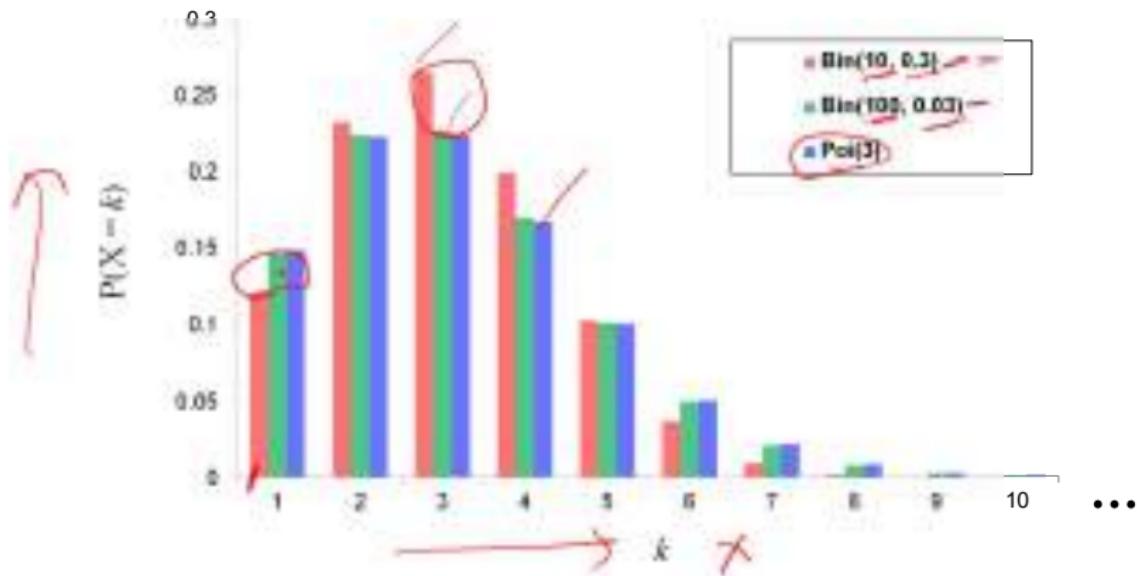
Different interpretations of "moderate":

$$n > 20 \text{ and } p < 0.05$$

$$n > 100 \text{ and } p < 0.1$$

Really, Poisson is Binomial as
 $n \rightarrow \infty$ and $p \rightarrow 0$, where $np = 1$

How Similar Are The Shapes, With Different n and p ?



Lecture 11-continuousRVs.pdf

Special Continuous Random Variables

Uniform Random Variable

- X is uniformly distributed between α and β

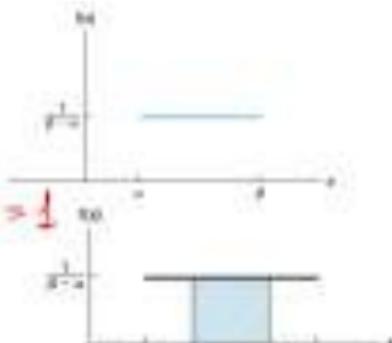
- $X \sim U(\alpha, \beta)$

- $P(X) = \frac{1}{\beta - \alpha}$

- $P(X \in [a, b])$

- $E[X] = \frac{\alpha + \beta}{2}$

$$\int_a^b x p(x) dx = \int_{\alpha}^{\beta} x \frac{1}{\beta - \alpha} dx = \left[\frac{x^2}{2} \right]_{\alpha}^{\beta} = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2}$$



Variance of uniform random variable

$$E(x^2) = \int_{\alpha}^{\beta} x^2 \left(\frac{1}{\beta - \alpha} \right) dx = \frac{1}{\beta - \alpha} \left(\frac{x^3}{3} \right) \Big|_{\alpha}^{\beta} = \frac{\beta^3 - \alpha^3}{(\beta - \alpha) 3}$$

$$= \frac{\beta^2 + \alpha^2 + \alpha \beta}{3}$$

$$\begin{aligned}\text{Var}(x) &= E[x^2] - E(x)^2 \\ &= \frac{\beta^2 + \alpha^2 + \alpha \beta}{3} - \left(\frac{\alpha + \beta}{2} \right)^2 = \frac{(\alpha - \beta)^2}{12}\end{aligned}$$

An example application of uniform R.V.s

- Given a set n elements x_1, x_2, \dots, x_n . You need to write an algorithm for selecting a random subset k of the n elements given access to a uniform random number generator $U(0,1)$

- $R = \emptyset$

for $i=0$ to $n-1$

$$u_i \sim U(0,1)$$

$$\tau_i = |R|$$

$$p_i = \frac{k-\tau_i}{n-i}$$

if $(u_i < p_i)$ add x_i

stop if $|R| = k$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = N$$

All possible subsets of size k

$$S_0, S_1, \dots, S_{N-1}$$

$$u \sim U(0,1)$$

if $u \in \left[\frac{j}{N}, \frac{j+1}{N} \right]$ then choose set S_j



Selecting a random subset in a streaming setting

Say you are hosting a webserver. You want to track the interaction of a random subset k of customers that arrive at the webserver. But you do not know the number of customers that will arrive in advance.

You have limited memory k and cannot store all possible customers data that arrive and then select a subset.

You can generate uniform random numbers between 0 and 1.

$R \leftarrow \{x_1, x_2, \dots, x_k\}$
for $i = k+1$ to ∞
choose a random # between $1 - \frac{k+1}{|R|}$
reject customer from $[R \setminus x_i]$

Reservoir sampling: n is unknown, data arrives in a stream

$$R_k \leftarrow \{x_1, x_2, \dots, x_k\}$$

- Initialize R_k with first k elements.

- For each subsequent x_i

- Sample a uniform integer s from $1, 2, \dots, i$.

- If $s \leq k$, $R[s] = x_i$

Let R_i denote the state of reservoir R after seeing x_i

Prove that the probability with which we add element x_j to R_i after seeing i elements is $\frac{k}{i}$

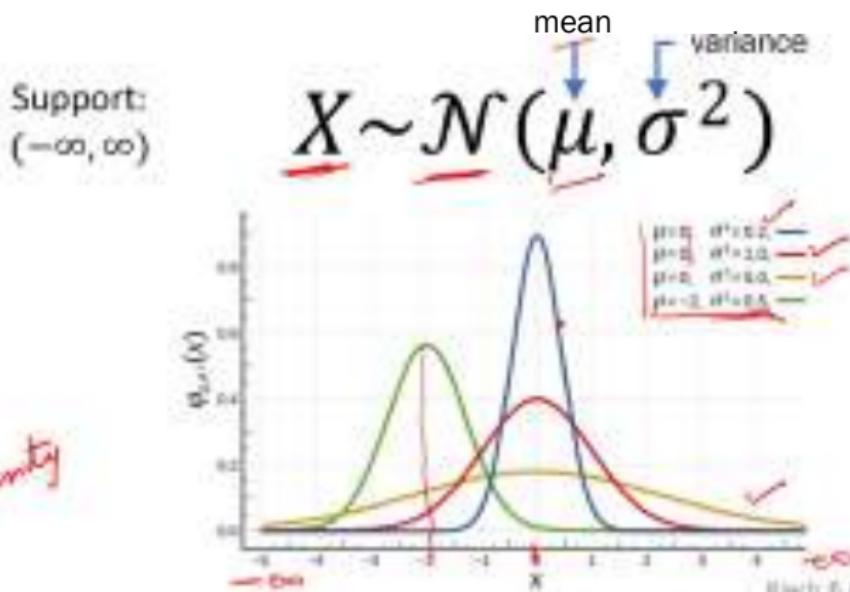
By induction

- Base case $i = k$ holds
- Assume that at $i-1$ $P(x_j \in R_{i-1}) = \frac{k}{i-1}$ $j \in [1, i-1]$
- $P(x_j \in R_i) = P(x_j \in R_{i-1}) \cdot (1 - \frac{k}{i} \cdot \frac{1}{k})$

$$= \frac{k}{i-1} \left(1 - \frac{1}{i}\right)$$

$$= \frac{k}{i}$$

Normal (Gaussian) Random Variable



Fiech & Cain, CS109, Stanford University

Normal (Gaussian) Random Variable

Support:
 $(-\infty, \infty)$

$$X \sim \mathcal{N}(\underline{\mu}, \sigma^2)$$

mean
↓
variance

PDF:

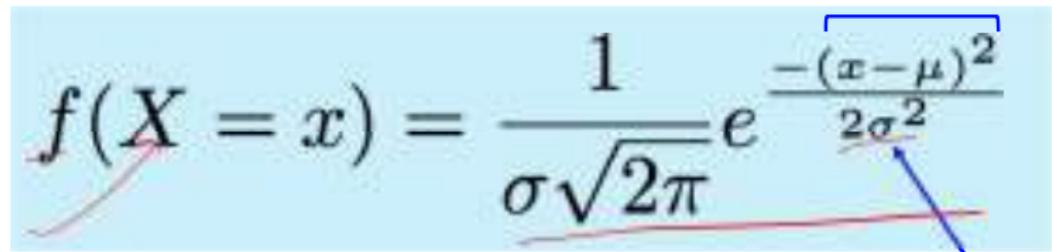
$$\underline{f}(\underline{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Piech & Cain, CS109, Stanford University

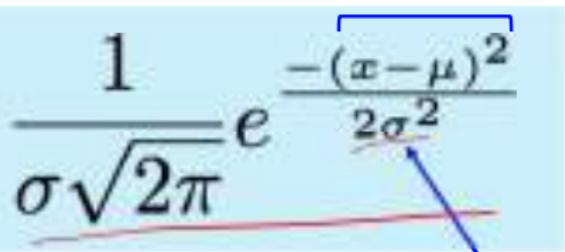
Anatomy of a The Normal PDF

distance to the mean
(makes the PDF symmetric
around the mean)

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



a constant:
makes the integral
over all possible
outcomes sum to 1


...normalized by
the variance

Expected value of a normal distribution

Verify that μ is the expected value of
 $x \sim N(\mu, \sigma^2)$

$$\underline{E((x-\mu))} = E(x) - \mu$$

$$\int_{-\infty}^{+\infty} (x-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \left. \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma (\sigma^2)} \right|_{-\infty}^{+\infty} = 0$$

$$E((x-\mu)) = 0 \Rightarrow E(x) = \mu$$

Variance

$$\begin{aligned}E((X-\mu)^2) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx \\&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-(y^2/2)} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\underbrace{y}_v)(\underbrace{ye^{-(y^2/2)}}_u) du \\&= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-y^2/2} dy \right] & \int udv = uv - \int vdu \\&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2 & \int ye^{-y^2/2} dy = -e^{-y^2/2}\end{aligned}$$

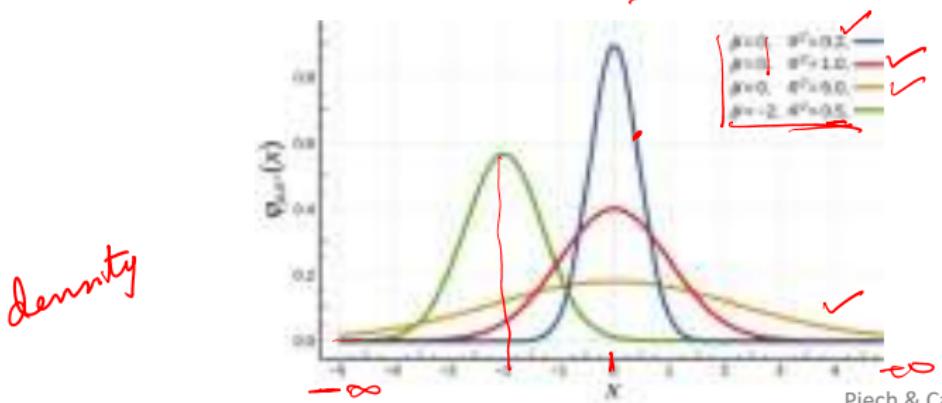
Lecture 12-gaussian.pdf

Normal (Gaussian) Random Variable

Support:
 $(-\infty, \infty)$

$$\underline{X} \sim \mathcal{N}(\mu, \sigma^2)$$

mean
variance



Piech & Cain, CS109, Stanford University

Normal (Gaussian) Random Variable

Support:
 $(-\infty, \infty)$

$$X \sim \mathcal{N}(\underline{\mu}, \underline{\sigma^2})$$

mean
↓
variance
↓

PDF:

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

Piech & Cain, CS109, Stanford University

Anatomy of a The Normal PDF

distance to the mean
(makes the PDF symmetric
around the mean)

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

a constant:
makes the integral
over all possible
outcomes sum to 1

...normalized by
the variance

Expected value of a normal distribution

Verify that μ is the expected value of
 $x \sim N(\mu, \sigma^2)$

$$E(x - \mu) = E(x) - \mu$$

$$\int_{-\infty}^{+\infty} (x - \mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \left. \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma (\sigma^2)} \right|_{-\infty}^{+\infty} = 0$$

$$E[(x - \mu)] = 0 \Rightarrow E(x) = \mu$$

Variance

$$\underline{\underline{E((X-\mu)^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx}}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-(y^2/2)} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (y)(\cancel{ye^{-(y^2/2)}}) dy$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-y^2/2} dy \right]$$

$$\int u dv = uv - \int v du$$

$$\int ye^{-y^2/2} dy = -e^{-y^2/2}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \underline{\underline{\sigma^2}}$$

Properties

If $X \sim N(\mu, \sigma^2)$ and if $Y = aX + b$, then $a + b$ are scalars.

Let F_Y be the cumulative density of Y

$$F_Y = P(Y \leq y) \quad f_Y = \frac{\partial}{\partial y} F_Y(y)$$

$$F_X = P(X \leq x) \quad f_X = \frac{\partial}{\partial x} F_X(x)$$

Let $a > 0 \Rightarrow P(ax+b \leq y) = P(X \leq \frac{y-b}{a})$

$$\begin{aligned} P(Y \leq y) &= P(X \leq \frac{y-b}{a}) \Rightarrow \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} F_X\left(\frac{y-b}{a}\right) \\ f_X\left(\frac{y-b}{a}\right) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-b-\mu)^2}{2\sigma^2}} \end{aligned}$$
$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_X\left(\frac{y-b}{a}\right) \frac{\partial}{\partial y} \left[\frac{y-b}{a}\right] \\ &= f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} \end{aligned}$$

$$f_x\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\mu a+b))^2}{2\sigma^2 \cdot a^2}}$$

$$f_y(y) = f_x\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} = \frac{1}{\sqrt{2\pi}\sigma a} e^{-\frac{(y-(\mu a+b))^2}{2a^2\sigma^2}}$$

$$\Rightarrow Y \sim N(\mu a + b; \frac{\sigma^2}{a^2}) \text{ if } a > 0$$

$$\begin{aligned} a < 0 \\ F_Y(y) &= P(Y \leq y) = P(ax + b \leq y) = P(X \geq \frac{y-b}{a}) \\ &= 1 - F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

$$Y \sim N(\mu a + b; \sigma^2 a^2)$$

Properties

- Median = mean (why?)
- Because of symmetry of the pdf about the mean
- Mode = mean – can be checked by setting the first derivative of the pdf to 0 and solving, and checking the sign of the second derivative.

Carl Friedrich Gauss (1777-1855)

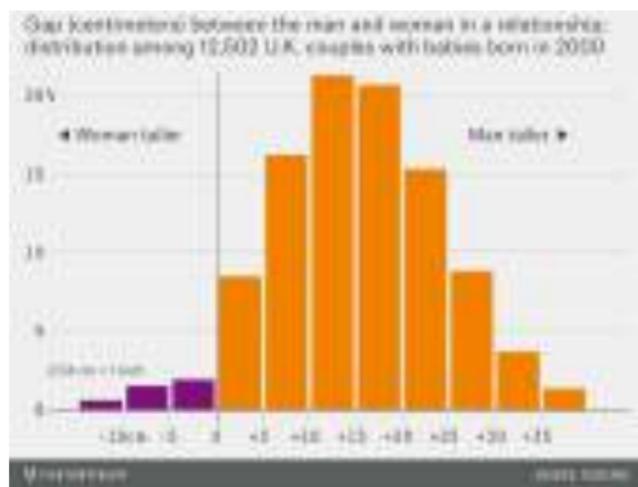
- German mathematician
- Sort-of invented the normal distribution
- Also astronomer, geologist, physicist
- Super influential in a lot of fields



Piech & Cain, CS109, Stanford University

Why the Normal?

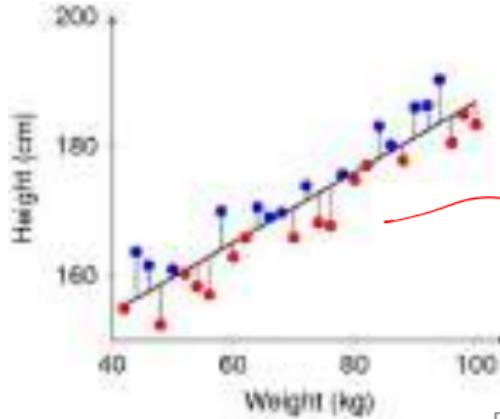
- Common for natural phenomena: human height, weight, shoe sizes, etc.



Piech & Cain, CS109, Stanford University

Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression



Piech & Cain, CS109, Stanford University

Why the Normal?

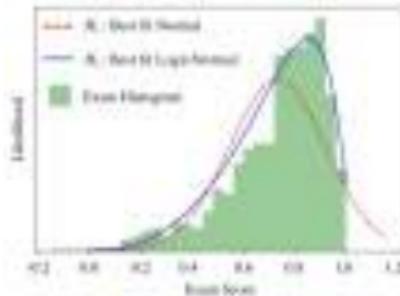
- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics



Piech & Cain, CS109, Stanford University

Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics
- Even things that aren't Normal might fit a normal-related distribution



Piech & Cain, CS109, Stanford University

Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics
- Even things that aren't Normal might fit a normal-related distribution

People also just assume things are normally distributed a lot.

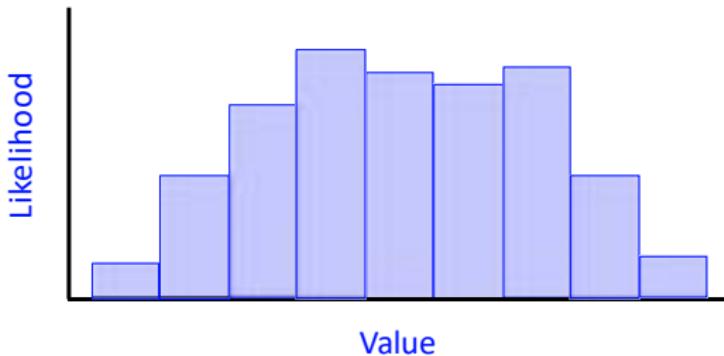
- They can do this in part because the Normal is so common
- But there's a deeper reason to it...

Ockham's razor

Shaving your hypothesis since 14th Century

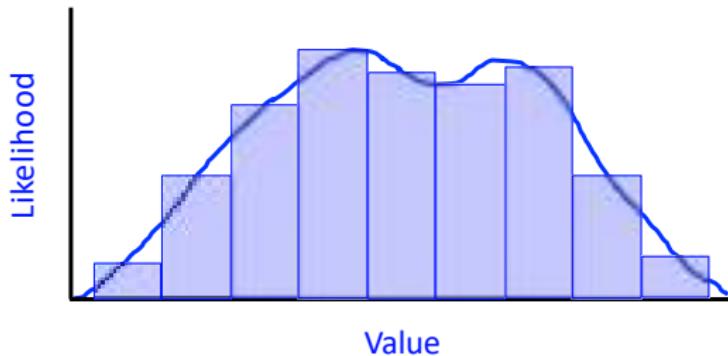


When We Fit Models To Data, We Try To Keep It Simple



Piech & Cain, CS109, Stanford University

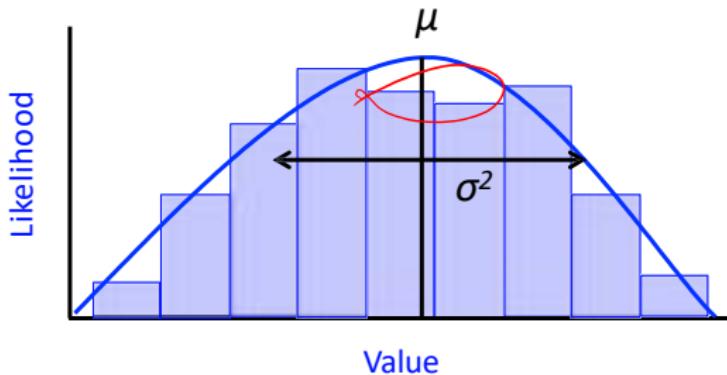
When We Fit Models To Data, We Try To Keep It Simple



This curve fits the data well, but does it really represent the distribution?
Or is it “overfit”, so that the curve captures too much of the noise?

Piech & Cain, CS109, Stanford University

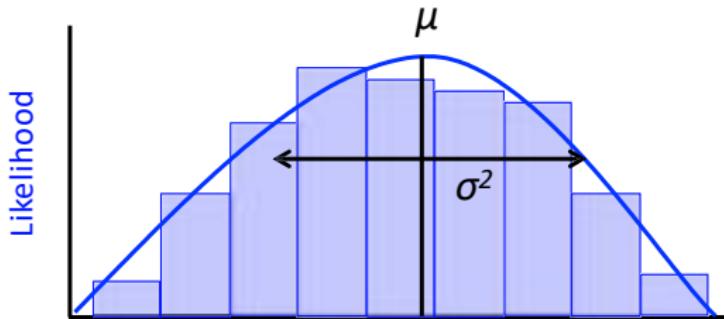
When We Fit Models To Data, We Try To Keep It Simple



This curve fits the data about as well, but appears to overfit less.
We could say that this simpler distribution makes fewer assumptions.
The formal concept for this idea is entropy

Piech & Cain, CS109, Stanford University

When We Fit Models To Data, We Try To Keep It Simple



For a fixed mean and variance, the unique distribution that maximizes the entropy is the normal distribution.

Entropy

- Measures the amount of uncertainty associated with a distribution.
- High entropy → high uncertainty or chaos.
- Formula of entropy:

x is continuous.

Entropy: $x, f(x)$

$$\text{Entropy}(x) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx$$

$\text{Ent}(x) \rightarrow$ discrete $E(x) \leq 0 \quad E(x) \geq 0$

Minimum entropy $p(x_i) = 1$ for any $i=1$

Maximum entropy: $p(x_i) = \frac{1}{k}$ for all i

Discrete $x, \text{ PMF } p(x)$
 $x \in \{x_1, x_2, \dots, x_k\}$

$$\text{Entropy}(x) = - \sum p(x_i) \log p(x_i)$$

Goal: find $p(x_i)$ s.t.

$$\max_{p(x_1), p(x_2), \dots, p(x_k)} - \sum_{x_i} p(x_i) \log p(x_i)$$

s.t. $p(x_i) \geq 0$

$$\sum_{i=1}^k p(x_i) = 1$$

Question in class

Optional information: Not in syllabus

- Example of a distribution with negative entropy: $X \sim U(0,1/2)$
- What is the interpretation of entropy for continuous R.V.
 - Entropy for continuous R.V is more precisely referred to as Differential entropy

The differential entropy describes the equivalent side length (in log₂) of the set that contains most of the probability of the distribution.

This is nicely illustrated and explained in Theorem 8.2.3 in *Elements of Information Theory* by Thomas M. Cover, Joy A. Thomas.

https://poincare.matf.bg.ac.rs/nastavno/viktor/Differential_Entropy.pdf

<https://stats.stackexchange.com/questions/256203/how-to-interpret-differential-entropy>

Entropy of Gaussian distribution.

$$\begin{aligned} \text{Ent}_{\sigma}(x) &= - \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} f(x) \right] \log e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \log \sqrt{2\pi}\sigma \int_{-\infty}^{\infty} f(x) dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} [f(x)] dx + \log \sqrt{2\pi}\sigma \int_{-\infty}^{\infty} f(x) dx \\ &= \frac{\sigma^2}{2\sigma^2} + \log (\sqrt{2\pi}\sigma) \\ &= \frac{1}{2} + \log \sqrt{2\pi}\sigma \end{aligned}$$

Proof that Gaussian distribution maximizes entropy given fixed mean and variance.

- Not in syllabus...
- For the interested, check out.

https://en.wikipedia.org/wiki/Differential_entropy

<https://medium.com/mathematical-musings/how-gaussian-distribution-maximizes-entropy-the-proof-7f7dcb2caf4d>

<https://statproofbook.github.io/P/norm-maxent.html>

Why is the Gaussian density defined so?

Optional topic: Not in syllabus.

- One student asked after class: how did the Gaussian density end up with such a non-intuitive form?
- It is possible to derive the Gaussian density function just starting from the desire to maximize entropy while matching a given mean μ , and variance σ^2
- Proof here: https://en.wikipedia.org/wiki/Differential_entropy

And here:

- [How Gaussian Distribution Maximizes Entropy — The Proof | by Freedom Preetham | Mathematical Musings | Medium](#)

CDF of a Gaussian distribution

- $X \sim N(0,1)$ *Standard normal distribution.*

$$P(X \leq x) = \Phi(x) = F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz$$

~~F(x)~~

- Not easy to compute in closed form: You can use libraries to access pre-computed values.

- CDF $F_Y(y)$ of a general $Y \sim N(\mu, \sigma^2)$

- Convert Y to standard form $X = \frac{Y-\mu}{\sigma}$

$$X \sim N(0, 1)$$

- $F_Y(y) = F_X\left(\frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right)$

$$P(Y \leq y) = P(\sigma X + \mu \leq y) = P(X \leq \frac{y-\mu}{\sigma}) = F_X\left(\frac{y-\mu}{\sigma}\right)$$

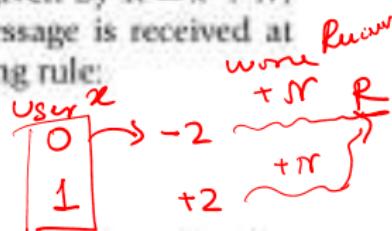
$$= \Phi\left(\frac{y-\mu}{\sigma}\right)$$

from a library

Example 5.5.b. Suppose that a binary message — either "0" or "1" — must be transmitted by wire from location A to location B. However, the data sent over the wire are subject to a channel noise disturbance and so to reduce the possibility of error, the value 2 is sent over the wire when the message is "1" and the value -2 is sent when the message is "0." If x , $x = \pm 2$, is the value sent at location A then R , the value received at location B, is given by $R = x + N$, where N is the channel noise disturbance. When the message is received at location B, the receiver decodes it according to the following rule:

if $R \geq .5$, then "1" is concluded

if $R < .5$, then "0" is concluded



Because the channel noise is often normally distributed, we will determine the error probabilities when N is a standard normal random variable.

$$N \sim \mathcal{N}(0, 1)$$

Let y denote the final 0/1 decoded value at the receiver.

$$\begin{aligned} P(y=0 | x=1) &= P(N < -1.5) \quad N \sim \mathcal{N}(0, 1) \\ &= \Phi(-1.5) = 1 - \Phi(1.5) = 0.0668 \end{aligned}$$

$$\begin{aligned} P(y=1 | x=0) &= P(N > 2.5) \\ &= 1 - P(N \leq 2.5) \\ &= 1 - \Phi(2.5) = 0.0062 \end{aligned}$$

$$\begin{aligned} P[\text{error message is "1"}] &= P[N < -1.5] \\ &= 1 - \Phi(1.5) = .0668 \end{aligned}$$

and

$$\begin{aligned} P[\text{error message is "0"}] &= P[N > 2.5] \\ &= 1 - \Phi(2.5) = .0062 \end{aligned}$$

Properties

MGF of $\underline{Z} \sim N(0,1)$

$$E(e^{t\underline{Z}}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2} dz$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-z^2/2 + tz} dz$$

MGF of $X \sim N(\mu, \sigma^2)$

Properties

MGF of $Z \sim N(0,1)$

$$\begin{aligned} E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2 - 2tx)/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= e^{t^2/2} \end{aligned}$$

MGF of $X \sim N(\mu, \sigma^2)$

$$\begin{aligned} E[e^{tX}] &= E[e^{t\mu + t\sigma Z}] \\ &= E[e^{t\mu} e^{t\sigma Z}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{t\mu} e^{(\sigma t)^2/2} \\ &= \boxed{e^{\mu t + \sigma^2 t^2/2}} \end{aligned}$$

Sum of Gaussian Random Variables

- Let $Y = \underbrace{X_1 + X_2 + \cdots + X_n}$
 - Where each $X_i \sim N(\mu_i, \sigma_i^2)$
 - What is the distribution of Y?
-
- $Y \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$
 - Proof via MGF.

/ .

MGF of sum of Gaussians

$$\begin{aligned} \bullet E(e^{tY}) &= E_Y \left(e^{t(x_1 + x_2 + \dots + x_n)} \right) \\ &= E_Y \left(e^{tx_1} \cdot e^{tx_2} \cdots e^{tx_n} \right) = \\ &= E_{x_1} \left(e^{tx_1} \right) \cdots E_{x_n} \left(e^{tx_n} \right) \\ &= \prod_{i=1}^n E_{x_i} \left(e^{tx_i} \right) = \prod_{i=1}^n e^{tu_i + t\sigma_i^2/2} \\ &= e^{\sum u_i t + t^2 \sum \sigma_i^2 / 2} \\ &= \boxed{e^{tu + t^2 \sigma^2}} \quad \text{where } \mu = \sum u_i \\ &\rightarrow \text{MGF of } N(\mu, \sigma^2) \quad \left[\because \text{MGF \& density fn. have a 1-1 correspondence} \right] \end{aligned}$$

Lecture 13-exponential.pdf

Exponential Random Variable

For any Poisson Process, the Exponential RV models time until an event:

$$X \sim \text{Exp}(\lambda)$$

PDF:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

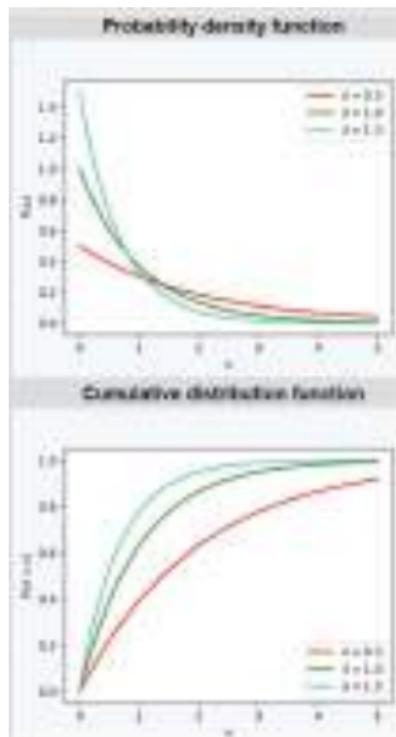
Examples:

- Time until next earthquake
- Time until a ping reaches a web server
- Time until a Uranium atom decays



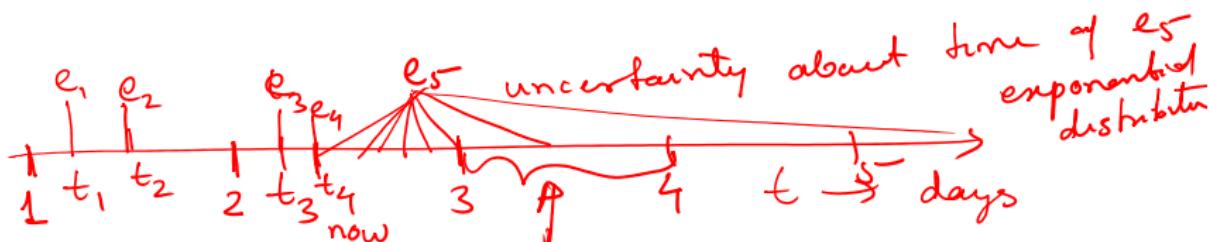
Cumulative Distribution function

$$\begin{aligned}F(x) &= P\{X \leq x\} \\&= \int_0^x \lambda e^{-\lambda y} dy \\&= 1 - e^{-\lambda x}, \quad x \geq 0\end{aligned}$$



Relationship to Poisson distribution

- Both are applicable when events occur continuously and independently at a constant average rate λ



- Poisson R.V is discrete over the number of events in a given time
- Exponential R.V is continuous and is the distance between two events.

Moment Generating Function , Mean, Variance

$$\phi(t) = E[e^{tX}] \quad X \sim \text{exp}(\lambda)$$

$$= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx$$

$$= \lambda \int_0^\infty e^{-(\lambda-t)x} dx$$

$$= \frac{\lambda}{\lambda - t}, \quad t < \lambda$$

$$E[X] = \phi'(0) = 1/\lambda$$

Differentiation yields

$$\phi'(t) = \frac{\lambda}{(\lambda - t)^2}$$

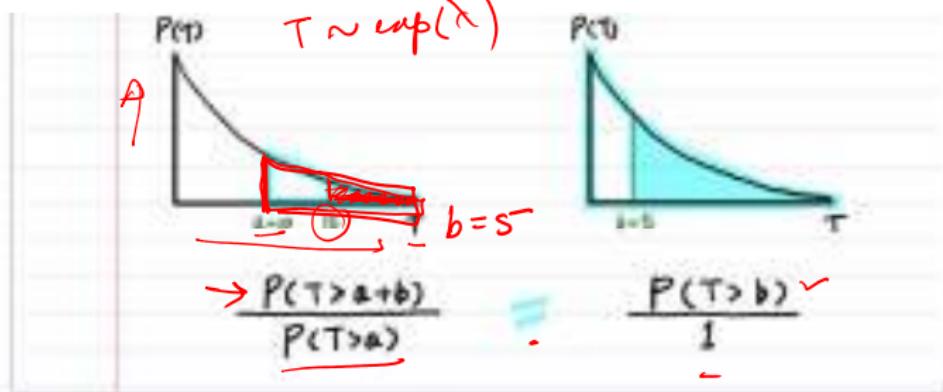
$$\phi''(t) = \frac{2\lambda}{(\lambda - t)^3} \cdot \frac{2\lambda}{\lambda^3}$$

$$\begin{aligned} \text{Var}(X) &= \phi''(0) - (E[X])^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned}$$

Memoryless property of exponential distribution

$$P(X > s + t | X > s) = P(X > t)$$

Example: lifetime T of a lamp if exponentially distributed, then remaining lifetime does not depend on how long lamp has been in use!



<https://towardsdatascience.com/what-is-exponential-distribution-79cd28f40e2a>

Proof of the memory-less property $P(A|B) = \frac{P(A, B)}{P(B)}$

$$X \sim \exp(\lambda)$$

$$\text{CDF}(x) = \frac{1 - e^{-\lambda x}}{e} \quad P(X \leq x) \quad t > 0$$

$$\begin{aligned} P(X > s+t | X > s) &= \frac{P(X > s+t, X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)} \\ &= \frac{1 - \text{CDF}_X(s+t)}{1 - \text{CDF}_X(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= 1 - \text{CDF}_X(t) \\ &= P(X > t) \end{aligned}$$

$\Rightarrow X$ is memory less

Memoryless property is unique to exponential!

- If X is a continuous random variable where $P(X>s+t|X>s)=P(X>t)$ then $P(X)$ is an exponential distribution. [Proof not part of the syllabus]

Proof:

Let F be the CDF of X , and let $G(x) := P(X > x) = 1 - F(x)$. The memoryless property says $G(s+t) = G(s)G(t)$, we want to show that only the exponential will satisfy this.

Try $s = t$, this gives us $G(2t) = G(t)^2, G(3t) = G(t)^3, \dots, G(nt) = G(t)^n$.

Similarly, from the above we see that $G(\frac{1}{2}t) = G(t)^{\frac{1}{2}}, \dots, G(\frac{1}{k}t) = G(t)^{\frac{1}{k}}$.

Combining the two, we get $G(\frac{m}{n}t) = G(t)^{\frac{m}{n}}$ where $\frac{m}{n}$ is a rational number.

Now, if we take the limit of rational numbers, we get real numbers. Thus, $G(xt) = G(t)^x$ for all real $x > 0$.

If we let $t = 1$, we see that $G(x) = G(1)^x$ and this looks like the exponential. Thus,

$G(1)^x = e^{x\ln(G(1))}$, and since $0 < G(1) \leq 1$, we can let $\ln(G(1)) = -\lambda$.

Therefore $e^{x\ln(G(1))} = e^{-\lambda x}$ and only exponential can be memoryless.

<https://math.stackexchange.com/questions/1801830/on-the-proof-that-every-positive-continuous-random-variable-with-the-memoryless>

Example

- Suppose the number of kms that a car can run before the battery wears down is exponentially distributed with average distance as 10000. If the person takes a 5000 km trip, what is the probability that the battery will not run down.

$$x \sim \text{exp}(\lambda) \quad \lambda = \frac{1}{10000} \quad E(x) = \frac{1}{\lambda}$$
$$P(x > 5000) = e^{-\lambda \cdot 5000} = e^{-\frac{1}{2}}$$

Another interesting property of exponential distribution

Proposition 5.6.1. If X_1, X_2, \dots, X_n are independent exponential random variables having respective parameters $\lambda_1, \lambda_2, \dots, \lambda_n$, then $\min(X_1, X_2, \dots, X_n)$ is exponential with parameter $\sum_{i=1}^n \lambda_i$.

$$Y = \min(X_1, X_2, \dots, X_n) \quad X_i \sim \exp(\lambda_i)$$

$$P(Y > b) = P(X_1 > b, X_2 > b, \dots, X_n > b)$$

$$P(\min(X_1, X_2, \dots, X_n) > b) = \prod_{i=1}^n P(X_i > b)$$

$$= \prod_{i=1}^n e^{-\lambda_i b} = e^{-b \left(\sum_{i=1}^n \lambda_i \right)}$$

$$\Rightarrow Y \sim \exp\left(\sum_{i=1}^n \lambda_i\right)$$

X_1, \dots, X_n are independent

Example

Example 5.6.c. A series system is one that needs all of its components to function in order for the system itself to be functional. For an n -component series system in which the component lifetimes are independent exponential random variables with respective parameters $\lambda_1, \lambda_2, \dots, \lambda_n$, what is the probability that the system survives for a time t ?



$$A \quad [I_j \text{ functioning}] \sim \exp(\lambda_j)$$

$$Y = A \text{ to } B \text{ connection is functioning} \quad \rightarrow \left[\sum_{i=1}^n \lambda_i \right]$$
$$P(Y \geq r) = P(\min(I_1, I_2, \dots, I_n) > r) = e^{-r \left[\sum_{i=1}^n \lambda_i \right]}$$

Another fun property of exponential distribution

Maximum entropy distribution

Among all continuous probability distributions with support $[0, \infty)$ and mean μ , the exponential distribution with $\lambda = 1/\mu$ has the largest differential entropy. In other words, it is the maximum entropy probability distribution for a random variable X which is greater than or equal to zero and for which $E[X]$ is fixed.^[2]

Lecture 14-MultipleRVs.pdf

Multiple Random Variables

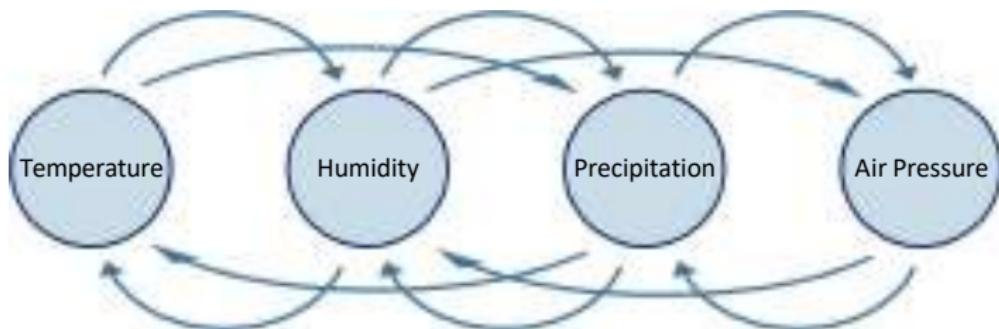
What Are We Missing?



The world is full of interesting probability problems...
...and many of them involve *multiple* random variables, being random *together*

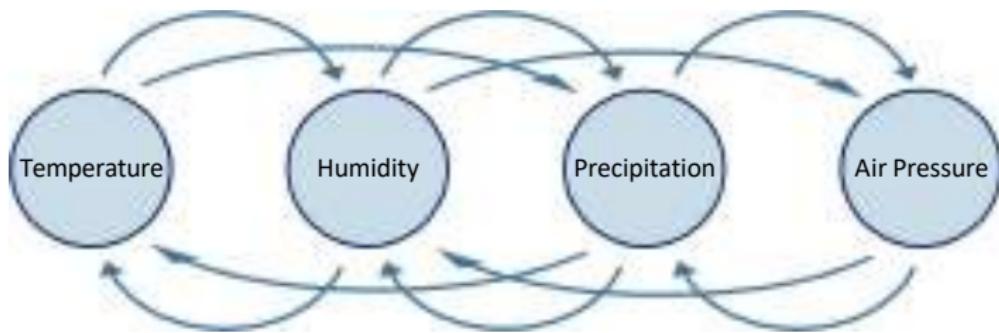
How Do We Model Multiple Random Variables Together?

Often, all the random variables involved are not independent of each other.



How Do We Model Multiple Random Variables Together?

Often, all the random variables involved are not independent of each other.



So we can't just have a single distribution for each random variable — we need a way to talk about all the random variables at the same time.

The “Joint” Distribution of Multiple Random Variables

For *discrete* random variables X and Y , we have a **joint probability mass function**:

$$P(X = x, Y = y)$$

The joint is the “and” between an assignment to X , and an assignment to Y

The same as $P(A \text{ and } B)$ for events A and B !

The “Joint” Distribution of Multiple Random Variables

For *discrete* random variables X and Y , we have a **joint probability mass function**:

$$P(X = x, Y = y) \quad \text{0.5134 ...}$$

$X = 2, Y = 4$ $P(\underset{\text{gender}}{X = \text{male}}, \underset{\text{height}}{Y = 5.9 \text{ feet}})$

The joint is the “and” between an assignment to X , and an assignment to Y

The same as $P(A \text{ and } B)$ for events A and B !

The “Joint” Distribution of Multiple Random Variables

For discrete random variables X and Y , we have a **joint probability mass function**:

$$\underline{P(X = x, Y = y)}$$

For continuous random variables, we have a **joint probability density function**:

$$\underline{f(X = x, Y = y)}$$

$$P(\underline{(X, Y) \in C}) = \iint_{\substack{(x,y) \in C}} \underline{f(x, y)} dx dy$$

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values \underline{X} and \underline{Y} .



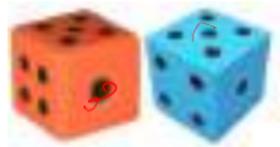
X
random variable

$P(\underline{X} = 1)$
probability of
an event

$P(X = k)$
probability mass function

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .



X
random variable

$P(X = 1)$
probability of
an event

$P(X = k)$
probability mass function

X, Y
random variables

$P(X = 1, Y = 6)$
probability of the intersection
of two events

$P(X = x, Y = y)$
joint probability mass function

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $\underline{P(X = x, Y = y)}$?



Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?



$$P(\underline{X = x}, \underline{Y = y}) = \frac{1}{\underline{36}}$$

$$(x, y) \in \{(1,1), \dots, (6,6)\}$$

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?



$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \dots, (6,6)\}$$

		X					
		1	2	3	4	5	6
Y	1	1/36	1/36
	2
	3
	4
	5
	6	1/36	1/36

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?



$$P(X = x, Y = y) = \frac{1}{36}$$

$(x, y) \in \{(1,1), \dots, (6,6)\}$

	X					
	1	2	3	4	5	6
1	1/36	1/36
2
3
4
5
6	1/36	1/36

Q1

$P(X = 4, Y = 3)$

This is a **joint probability table**: it contains the probabilities of all possible outcomes for a set of discrete random variables

Another Example

Example 4.3.a. Suppose that 3 batteries are randomly chosen from a group of 3 new, 4 used but still working, and 5 defective batteries. If we let X and Y denote, respectively, the number of new and used but still working batteries that are chosen, then the joint probability mass function of X and Y , $p(i, j) = P(X = i, Y = j)$, is given by

$$p(i, j) = \frac{\binom{3}{i} \binom{4}{j} \binom{5}{3-i-j}}{\binom{12}{3}}$$

$$p(0, 0) = \binom{3}{0} / \binom{12}{3} = 10/220$$

$$p(0, 1) = \binom{3}{1} \binom{4}{1} / \binom{12}{3} = 40/220$$

$$p(0, 2) = \binom{4}{2} \binom{3}{1} / \binom{12}{3} = 30/220$$

Table 6.1 $P(X = i, Y = j)$

i	j	0	1	2	3	
		0	1	2	3	
0	0	10	0	0	0	10
1	1	0	10	0	0	10
2	2	0	0	10	0	10
3	3	0	0	0	10	10
	Column Sum	10	10	10	10	40
	Row Sum	10	10	10	10	40
	Total	40	40	40	40	160
		new	used	defective		sd

Annotations:

- $X \sim \# \text{ of new batteries}$
- $Y \sim \# \text{ of used batteries}$
- (3) circled in red at the bottom right corner of the table.

Example with continuous density

Example 4.3.c. The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

$$C = \{X > 1, Y < 1\}$$

$$\begin{aligned} P(X > 1, Y < 1) &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y} dx dy \\ &= \int_0^1 2e^{-2y}(-e^{-x}) \Big|_1^\infty dy \end{aligned}$$

$$P(X < Y) = \iint_{y < x} f(x, y) dy$$

$$C = \{X < Y\}$$

$$\begin{aligned} P\{X < Y\} &= \iint_{\{x < y\}} 2e^{-x}e^{-2y} dx dy \\ &= \int_0^\infty \int_0^y 2e^{-x}e^{-2y} dx dy \\ &= \int_0^\infty 2e^{-2y}(1 - e^{-y}) dy \\ &= \int_0^\infty 2e^{-2y} dy - \int_0^\infty 2e^{-3y} dy \\ &= 1 - \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

Marginals

$$\begin{aligned} P(X < a) &= \int_{-\infty}^a \int_0^\infty 2e^{-2x} e^{-y} dy dx \\ &= \int_0^a e^{-y} dy \\ &= 1 - e^{-a} \quad \blacksquare \end{aligned}$$

$$P(X < a) \sim \exp(\lambda = 1)$$

Law of total probability

Joint table expresses the complete information about the random variables

$$\underline{P(X = x)} = \sum_y P(X = x, Y = y)$$

$P(X = x)$ is called the marginal of the joint distribution $P(X, Y)$

$$f(x) = \int_y f(x, y) dy$$

Independent Random Variables

The random variables X and Y are said to be independent if for any two sets of real numbers A and B

$$P[X \in A, Y \in B] = P[X \in A]P[Y \in B] \quad (4.3.7)$$

This also implies that

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

Or $F_{X,Y}(a, b) = F_X(a)F_Y(b)$

In the jointly continuous case, the condition of independence is equivalent to

$$f(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y$$

Example 4.3.d. Suppose that X and Y are independent random variables having the common density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(y) = \begin{cases} e^{-y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{array}{c} (x+y) \\ \text{---} \\ \text{C} \end{array}$$

Find the density function of the random variable $Z = X/Y$.

$$f(z = \frac{x}{y})$$

$$\begin{aligned} P(Z \leq a) &= P\left(\frac{X}{Y} \leq a\right) = P(X \leq aY) = \int_0^{\infty} \int_0^{\infty} f(x,y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{ay} e^{-x} e^{-y} dx dy = F_z(a) = \frac{a}{a+1} \end{aligned}$$

$$f(z) = \frac{\partial}{\partial a} F_z(a)$$

$$z = T(x) \quad f_z(z) = f_x(T^{-1}(z)) \frac{\partial T^{-1}}{\partial z}$$

Conditional Probability

Given two discrete random variables X, Y. The conditional probability of X given a specific value of Y is given as:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

For continuous variables with joint density of X, Y as f(x, y):

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f(y)}$$

$$\begin{aligned} f_{X|Y}(x|y) dx &= \frac{f(x, y) dx dy}{f(y) dy} \\ &\approx \frac{P\{x \leq X \leq x + dx, y \leq Y \leq y + dy\}}{P\{y \leq Y \leq y + dy\}} \\ &= P\{x \leq X \leq x + dy | y \leq Y \leq y + dy\} \end{aligned}$$

Example 4.3.b. The joint density of X and Y is given by

$$f(x, y) = \begin{cases} \frac{12}{5}x(2-x-y) & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute the conditional density of X , given that $\underline{Y = y}$, where $0 < y < 1$.

Solution. For $0 < x < 1, 0 < y < 1$, we have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \quad | \\ &= \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) dx} \quad | \\ &= \frac{x(2-x-y)}{\int_0^1 x(2-x-y) dx} \quad | \\ &= \frac{x(2-x-y)}{\frac{2}{3} - y/2} \quad | \\ &= \frac{6x(2-x-y)}{4-3y} \quad | \blacksquare \end{aligned}$$

$$f_Y(y) = \int_{x=0}^2 f(x, y) dx$$

Joint distribution of n random variables

If X_1, X_2, \dots, X_n are n random variables. Their joint distribution is defined for the discrete case as

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$$

Further, the n random variables are said to be jointly continuous if there exists a function $f(x_1, x_2, \dots, x_n)$, called the joint probability density function, such that for any set C in n -space

$$P\{X_1, X_2, \dots, X_n \in C\} = \int \int_{x_1, \dots, x_n \in C} \dots \int f(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n$$

In particular, for any n sets of real numbers A_1, A_2, \dots, A_n

$$P[X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n]$$

$$= \int_{A_n} \int_{A_{n-1}} \dots \int_{A_1} f(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n$$

Example 4.3.e. Suppose that the successive daily changes of the price of a given stock are assumed to be independent and identically distributed random variables with probability mass function given by

x_i	
-3	with probability .05
-2	with probability .10
-1	with probability .20
0	with probability .30
1	with probability .20
2	with probability .10
3	with probability .05

$$P[\text{daily change is } i] = P(x_i)$$

Then the probability that the stock's price will increase successively by 1, 2, and 0 points in the next three days is

$$P[X_1 = 1, X_2 = 2, X_3 = 0] = (.20)(.10)(.30) = .006$$

where we have let X_i denote the change on the i th day. ■

Lecture 15-ParameterEstimation.pdf

Parameter Estimation

Sunita Sarawagi

CS 215. Fall 2024

So far..

- Computing probabilities of outcomes given a fixed distribution.
- Distributions were given to us as a function..
- Functions had parameters with fixed values

What are Parameters?

Consider some probability distributions:

- Ber(p)
- Poi(λ)
- Unj(α, β)
- Normal(μ, σ^2)
- $Y = mX + b$ $X \sim N(0, 1)$
- etc...

$$\begin{aligned}\theta &= p \\ \theta &= \lambda \\ \theta &= \{\alpha, \beta\} \\ \theta &= \{\mu, \sigma^2\} \\ \theta &= \{m, b\}\end{aligned}$$

Call these "parametric models"

Non parametric model example - histograms

Given model, **parameters** yield actual distribution

- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

Today's class

How to determine the values of the parameters.

Parameters differ based on the task and application. These are not fixed like the speed of light.

The setup for parameter estimation in real-life

- Step 1: A real-life problem:

1. Estimating the probability that at least two out of four servers will be alive next day ✓
2. The probability that stock price will rise by 10% in the next week
3. The expected number of clicks on an advertisement in the next 3 hours

- Step 2: Model the problem: Choose a functional form of the uncertainty.

1. Binomial?

Assume that servers fail independently
 $X = \# \text{ of failures in a day}$ $X \sim \text{Bin}(n)$

2. Gaussian?

$X = \text{change from } t \text{ day to the next}$

3. Poisson?

$X = \# \text{ of clicks on the ad per hour}$

The setup for parameter estimation in real-life

Step 3: Collect a training sample by observing over several days.

1. Sample server failure data observed over 3 days

	day 1 ser 1	day 1 ser 2	\bar{x}_3	\bar{x}_4	\bar{x}_5	\bar{x}_6	x_7	x_8	day 3 ser 3
	x_1	x_2							x_{12}
	0	0	1	0	0	0	0	1	0

2. Stock price change over a 10 days

change $1 \rightarrow 2$ $2 \rightarrow 3$ - - - - - $9-10$

1% -2% 3%
 x_1 x_2 x_9

3. Number of clicks on the ad over the last 20 hour

Hour	1	2	3	<u>4</u>	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	x_1	x_2	x_3							x_4										x_{20}

- Step 4: Estimate the unknown parameters using the training sample



The overall setup in parameter estimation

density or pmf

- Given: a density or distribution function with parameters $f(x, \theta)$
 - Given: sample: $D = \{x_1, x_2, \dots, x_N\}$
 - The i-th sample is a random variable X_i assumed to be independently identically distributed as per the unknown $f(x, \theta)$
 - Find θ .
-
- Since D is a finite sample, we cannot really know the actual θ . Best we can do is obtain an estimate of θ
 - We will denote the estimate as $\hat{\theta}$
 - Goodness of estimate will be discussed later.

Types of estimators

- Maximum likelihood: sample D is all you got.
- Bayesian estimation: in addition to sample, we got prior beliefs.

$\hat{\theta}$ point estimator

Maximum Likelihood Estimation

- If θ were known we could have calculated the probability of getting the N outcomes in $D = \{x_1, x_2, \dots, x_N\}$ from the distribution as
 x_1, x_2, \dots, x_N are independent
for both continuous & discrete
- $P(D|\theta) = P(x_1, \dots, x_N|\theta) = \prod_i P(x_i|\theta) = \boxed{\prod_i f(x_i; \theta)}$
- Likelihood refers to the above function. Often denoted as $L(\theta)$
- Maximum likelihood estimator:
 - Choose the parameter θ for which the above likelihood is maximized

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N f(x_i; \theta) \rightarrow$$

Finding θ that maximizes likelihood

- Use log-likelihood instead of likelihood to convert products into sums

$$\bullet LL(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log f(x_i, \theta)$$

\nwarrow

$\max_{\theta} LL(\theta)$

↑
sum over observations

- Maximum likelihood estimator

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log f(x_i | \theta)$$

Solved using numerical optimization methods applying calculus.

MLE for Bernoulli

$$\underline{X \sim \text{Bern}(p)} \quad x \in \{0, 1\}$$

$$\underline{f(x; p) = p^x (1-p)^{1-x}}$$

Data sample: D

	x_1	x_2	x_3	x_4	\dots	x_N	x_6	x_7	x_8	x_9	x_{10}
	0	1	1	0	0	0	0	0	0	1	1

$$\begin{aligned} \max_p \underline{LL}(\theta = [p]) &= LL_D(p) = \max_p \sum_{i=1}^N \log P(x_i) (1-P)^{1-x_i} \\ &= \max_p \sum_{i=1}^N x_i \underline{\log P} + \left(N - \sum_{i=1}^N x_i\right) \underline{\log(1-p)} \end{aligned}$$

$$\text{Let} - \sum_{i=1}^N x_i = N_1$$

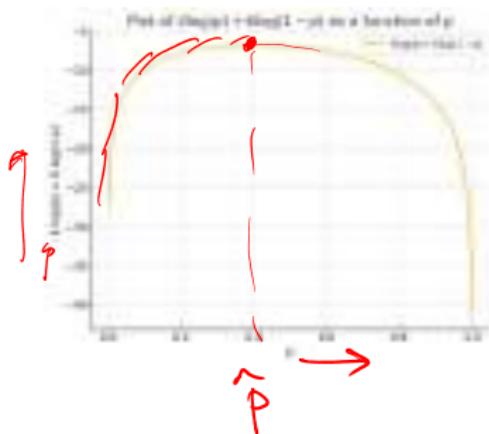
$$\underset{P}{\text{max}} \quad N_1 \log P + (N - N_1) \log (1 - P)$$

$\underbrace{\hspace{10em}}$
 $LL(P)$

$$\frac{\partial LL}{\partial p} = \frac{N_1}{\hat{p}} - \frac{N-N_1}{1-\hat{p}} = 0$$

$$\Rightarrow \hat{P} = \frac{N}{Z}$$

\leftarrow concave in P
 \Rightarrow unique maxima at the P where $\frac{\partial L}{\partial P} = 0$



Examples: MLE for Poisson

$$f(k, \lambda) = \frac{\bar{e}^{\lambda} \lambda^k}{k!}$$

$$x \sim \exp(\lambda)$$
$$f(x, \lambda) = \frac{\bar{e}^{\lambda} \lambda^x}{x!}$$

$$D = \{x_1, x_2, \dots, x_N\}$$

$$LL(\lambda) = \sum_{i=1}^N \log \frac{\bar{e}^{\lambda} \lambda^{x_i}}{x_i!} = \left(\sum_{i=1}^N x_i \right) \log \lambda - \sum_{i=1}^N \log x_i!$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{arg\max}} \left(\sum_{i=1}^N x_i \right) \log \lambda - \lambda N$$
$$\frac{\partial LL}{\partial \lambda} = \sum_{i=1}^N x_i - N \quad : \quad \hat{\lambda} = \frac{\sum x_i}{N} \leftarrow \begin{matrix} \text{sample} \\ \text{mean} \end{matrix}$$

MLE for Gaussian

Homework

$$x \sim N(\mu, \sigma^2)$$
$$f(x, \theta = [\mu, \sigma^2]) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{\partial LL}{\partial \mu} = 0$$

at μ calculated above

$$\frac{\partial LL}{\partial \sigma} = 0$$

$$f(x_1, \dots, x_n | \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right]$$

The logarithm of the likelihood is thus given by

$$\log f(x_1, \dots, x_n | \mu, \sigma) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

In order to find the value of μ and σ maximizing the log-likelihood, we compute

$$\frac{\partial}{\partial \mu} \log f(x_1, \dots, x_n | \mu, \sigma) = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma} \log f(x_1, \dots, x_n | \mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3}$$

Equating these equations to zero yields that

$$\hat{\mu} = \sum_{i=1}^n x_i / n$$

and

$$\hat{\sigma} = \left[\sum_{i=1}^n (x_i - \hat{\mu})^2 / n \right]^{1/2}$$

Example 7.2.d. The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

4, 0, 6, 5, 2, 1, 2, 0, 4, 3

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

Homework

• Most difficult question: what distribution to use to model accidents in a city?

- Binomial? Will need to know total number of drivers
- Gaussian?
- Poisson?

Solution in notebook

MLE for a new distribution: Gamma distribution

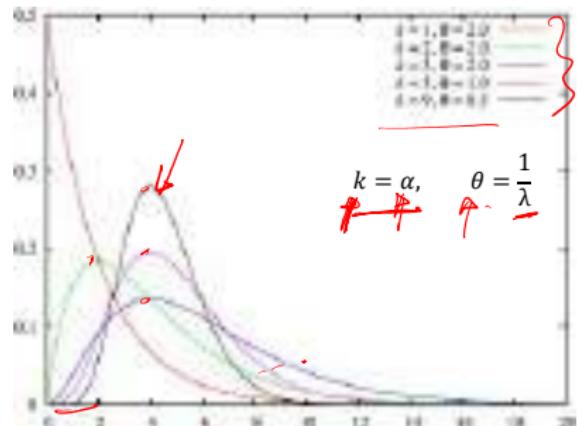
A random variable is said to have a gamma distribution with parameters (α, λ) , $\lambda > 0$, $\alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for what α is $f(x) \sim \exp(\lambda)$
 $= \lambda e^{-\lambda x}$

$$\alpha = 1$$

- Can look like Gaussian for positive random variables.
- Reduces to exponential when $\alpha = 1$
- More flexible than exponential since mode is not at 0.
- Useful to model one-sided long tails e.g. blue curve here.



What is $\Gamma(\alpha)$? Gamma function

$$\begin{aligned}\underline{\Gamma(\alpha)} &= \underline{\int_0^\infty \lambda e^{-\lambda x} (\lambda x)^{\alpha-1} dx} \\ &= \underline{\int_0^\infty e^{-y} y^{\alpha-1} dy} \quad (\text{by letting } y = \lambda x)\end{aligned}$$

The integration by parts formula $\int u dv = uv - \int v du$ yields, with $u = y^{\alpha-1}$, $dv = e^{-y} dy$, $v = -e^{-y}$, that for $\alpha > 1$,

$$\begin{aligned}\int_0^\infty e^{-y} y^{\alpha-1} dy &= -e^{-y} y^{\alpha-1} \Big|_{y=0}^{y=\infty} + \int_0^\infty e^{-y} (\alpha-1) y^{\alpha-2} dy \\ &= (\alpha-1) \int_0^\infty e^{-y} y^{\alpha-2} dy\end{aligned}$$

or

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

$$\Gamma(1) = 1$$

(5.7.1)

If α is + integer then

$$\Gamma(\alpha) = (\alpha-1)!$$

Estimate MLE of parameter λ of Gamma distribution -

$$D = \{x_1, x_2, \dots, x_N\}$$

$$\hat{\theta} = \hat{\lambda} = \underset{\lambda}{\operatorname{arg\,max}} \sum_{i=1}^N \log \frac{x_i^{\lambda-1} e^{-(\lambda x_i)}}{\Gamma(\lambda)}$$

$$= \underset{\lambda}{\operatorname{arg\,max}} \frac{\sum_{i=1}^N [-\lambda x_i + (\lambda-1) \log \lambda + \log \Gamma(\lambda)]}{F(\lambda)}$$

$$\frac{\partial F}{\partial \lambda} = -\sum_{i=1}^N x_i + \frac{(\lambda-1)N}{\lambda} + \frac{N}{\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{\alpha N}{\sum_{i=1}^N x_i}$$

MLE for α

$$\underset{\alpha}{\operatorname{argmax}} \quad (\alpha - 1) \left[N \log x + \sum_{i=1}^N \log x_i \right] - N \log \Gamma(\alpha)$$

$$\frac{\partial \ell(\alpha)}{\partial \alpha} = \sum_{i=1}^N \log x_i - \frac{N}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha} + N \log \Gamma(\alpha) = 0$$

Not easy to solve in closed form.
But can be estimated numerically.

Lecture 16-EvaluatingEstimators.pdf

Evaluating a point estimator (Chapter 7.7)

- Given sample $D = \{X_1, X_2, \dots, X_N\}$
- Given density/PMF: $f(x, \theta)$
- Let $\hat{\theta}_D$ be any estimated value of θ , example maximum likelihood estimate.
- How do we measure quality of the estimate?
 - Square difference from actual parameter.
 - $Error(\hat{\theta}_D) = (\hat{\theta}_D - \theta)^2$

This error is a function of a specific data sample D.

Often, we want the expected square error where expectation is over all possible Ds.

Expected square error of the mean estimate

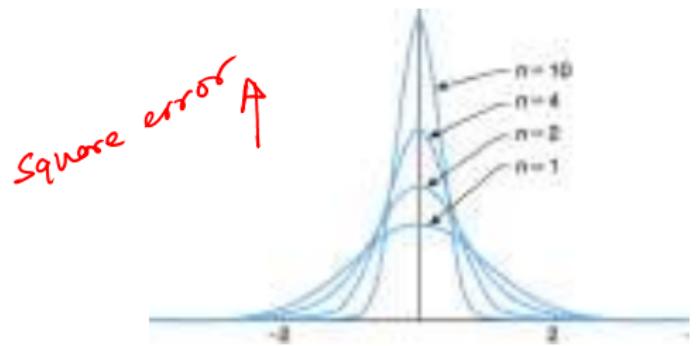
A common estimated parameter is the mean of the distribution.

$$\underline{\theta} = \mu = E_f(X), \quad \hat{\theta} = \underline{(X_1 + X_2 + \dots + X_N)/N} \leftarrow \text{sample mean}$$

- Expected square error of the above estimate $E_f\left(\sum_i \frac{x_i}{N} - \underline{\theta}\right)^2 = \sigma^2/N$

$$\text{where } \sigma^2 = E_f(X - \mu)^2$$

$$\begin{aligned} & E_f\left(\frac{\sum_i x_i - N\theta}{N}\right)^2 \\ &= \frac{1}{N^2} E_f\left(\sum_{i=1}^N (x_i - \theta)^2 + 2 \sum_{i \neq j} E(x_i - \theta)(x_j - \theta)\right) \\ &= \frac{1}{N^2} \sum_{i=1}^N E(x_i - \theta)^2 + 2 \underbrace{\sum_i E(x_i - \theta)}_{j \neq i} \sum_j E(x_j - \theta) \\ &= \frac{N\sigma^2}{N^2} + 2 \cdot 0 \cdot \sigma^2 / N \quad [\because E(x_i - \theta) = 0] \end{aligned}$$



Biased and Unbiased estimator

- The estimated parameter $\hat{\theta}_D$ is a random variable since it depends on D which is a random sample.
 - For example: $|D| = 3$. $\theta \equiv \gamma$ of an exponential distribution.
 - Two different samples and means.
 $D_1 = \{1, 1.5, 0.5\}$ $D_2 = \{1.2, 0.8, 1.8\}$
 $\bar{x}_{D_1} = \frac{3}{3} = 1$ $\bar{x}_{D_2} = \frac{3.8}{3} = 1.26$
 D_1, D_2, \dots
- An interesting question: what is the expected value $E_D(\hat{\theta}_D)$ over different random samples D ? How does that compare with true θ ?
- Unbiased: $E_D(\hat{\theta}_D) = \theta$
 - Biased: $E_D(\hat{\theta}_D) \neq \theta$

Example: two unbiased estimator

- Parameter $\theta = \mu$ of Gaussian distribution.
 - Two different estimators:
 - Lame estimator: just take first element: $\hat{\theta}_D = X_1$
 - MLE: $\hat{\theta}_D = \frac{X_1 + X_2 + \dots + X_N}{N}$
- $$E_D[\hat{\theta}_D] = E_f[X_1] = \mu$$
- $$E_f\left[\frac{X_1 + X_2 + \dots + X_N}{N}\right] = \frac{N \cdot \mu}{N} = \mu$$

Example: a biased estimator

- A constant estimator.

$$\hat{\theta}_D = 5.7 \text{ foot.}$$

- MLE of Variance parameter of Gaussian: $\hat{\sigma}_D^2 = \frac{\sum_{i=1}^N (x_i - \hat{\mu}_D)^2}{N}$
- Proof in $E_D[\hat{\sigma}_D^2] \neq \sigma^2$ $N=1 \Rightarrow \hat{\sigma}_{|D|=1}^2 = 0$

[https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample variance](https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance)

An unbiased estimator of variance of Gaussian

$$S^2 = \frac{\sum_{i=1}^N (x_i - \bar{x}_D)^2}{N-1}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

where $\bar{x} = \sum_{i=1}^n x_i/n$. It follows from this identity that

$$(n-1)S^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Taking expectation of both sides of the preceding equality using the fact that for any random variable W , $E[W^2] = \text{Var}(W) + (E[W])^2$,

$$\begin{aligned} (n-1)E[S^2] &= E\left[\left(\sum_{i=1}^n x_i^2\right) - n\bar{x}^2\right] \\ &= nE[X_1^2] - nE[\bar{X}^2] \\ &= n\text{Var}(X_1) + nE[X_1]^2 - n\text{Var}(\bar{X}) - nE[\bar{X}]^2 \\ &= n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2 \\ &= n\sigma^2 \end{aligned}$$

or

$$E[S^2] = \sigma^2$$

Consistent estimator

- An estimator is consistent if the estimation error goes to zero as N (size of D) goes to infinity.

$$\hat{\theta}_D \rightarrow \theta \text{ as } |D| \rightarrow \infty$$

Example of an unbiased estimator that is not consistent.

- Parameter $\theta = \mu$ of Gaussian distribution, Lame estimator: just take first element: $\hat{\theta}_D = X_1$

Example of an unbiased, consistent estimator:

- Parameter $\theta = \text{mean of a distribution. } \hat{\theta} = (X_1 + X_2 + \dots + X_N)/N$ ✓

Example of a biased, consistent estimator:

- Parameter $\theta = \sigma$ of Gaussian distribution, $\hat{\sigma}$ is sample variance.

Lecture 17-BayesianEstimates.pdf

Limitation of MLE

- Over-reliance on data sample D. If data is limited, estimates can be very wrong.
 - Example, Bernoulli p could be zero if no 1s in 10 trials.
- No indication on the uncertainty of the estimated parameters.
 - Example, for a Bernoulli parameters whether estimation is made from two with 50% heads or 1000 examples with 50% heads, the estimated parameter is the same.
- No mechanism to specify human's prior knowledge of the parameters.

$$\begin{array}{lll} \text{--- } D_1 & |D_1| = 2 & n_1(D_1) = 1, \quad N - n = 1 \\ \cancel{\text{--- } D_2} & |D_2| = 1000 & n_1(D_2) = 500 \end{array} \quad \begin{array}{l} \hat{p}_1 = 0.5^- \\ \hat{p}_2 = 0.5^- \end{array}$$

Example of limitations of MLE

- Suppose you toss a coin 10 times and get

H, H, H, H, H, H, H, H, H, H

Estimate p ? MLE: $\hat{p} = \# \text{ ones} / N = 1$

What is your guess on the probability p of head?

- Suppose you want to form a music band, and you are looking for bass guitarist. You ask 7 random batchmates: "Can you play the bass guitar" and you get answers

N, N, N, N, N, N, N \hat{D}

What fraction of batchmates play bass guitar?

MLE: 0

Do you have a different guess? 0.01

$p \approx 0$

Bayesian estimation

- Treat the parameters as a random variable which has a distribution.
- Step 1: Humans specify their prior knowledge of the values of the parameters as a distribution $f_{\text{pr}}(\theta)$
 - Example: $f_{\text{pr}}(\theta) \sim U(0,1)$ where θ denotes the parameter p of a Bernoulli
 - Example for Gaussian:

Temperature of CPU on your laptop $T \sim \mathcal{G}(0, 15^2)$

$f_{\text{pr}}(\theta) \sim N(30, 10)$

Also called prior probability

Bayesian estimation

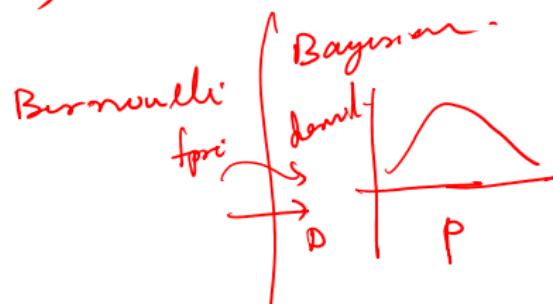
- Calculate the posterior distribution of parameters after observing data D following Bayes rule

$$\cancel{f(D|\theta) = f(\theta)f(D|\theta)/\int_{\theta} f(\theta)f(D|\theta)}$$

$$f_{\text{post}}(\theta|D) = \frac{f(D|\theta)f_{\text{pri}}(\theta)}{\int f(D|\theta')f_{\text{pri}}(\theta')} \quad \checkmark$$

Posterior probability

$$D \rightarrow \begin{array}{c} \text{MLE} \\ \downarrow \\ \hat{p} = 0.6 \end{array}$$



Using Bayesian estimates

$$f(\theta|D) \equiv f_{\text{pos}}(\theta|D)$$

- Exact Bayesian probability computation:

- Given a new x , calculate $f(X|D)$

$$f(x|D) = \int_{\theta} f(x|\theta) f_{\text{pos}}(\theta|D) d\theta$$

Binomial e.g.

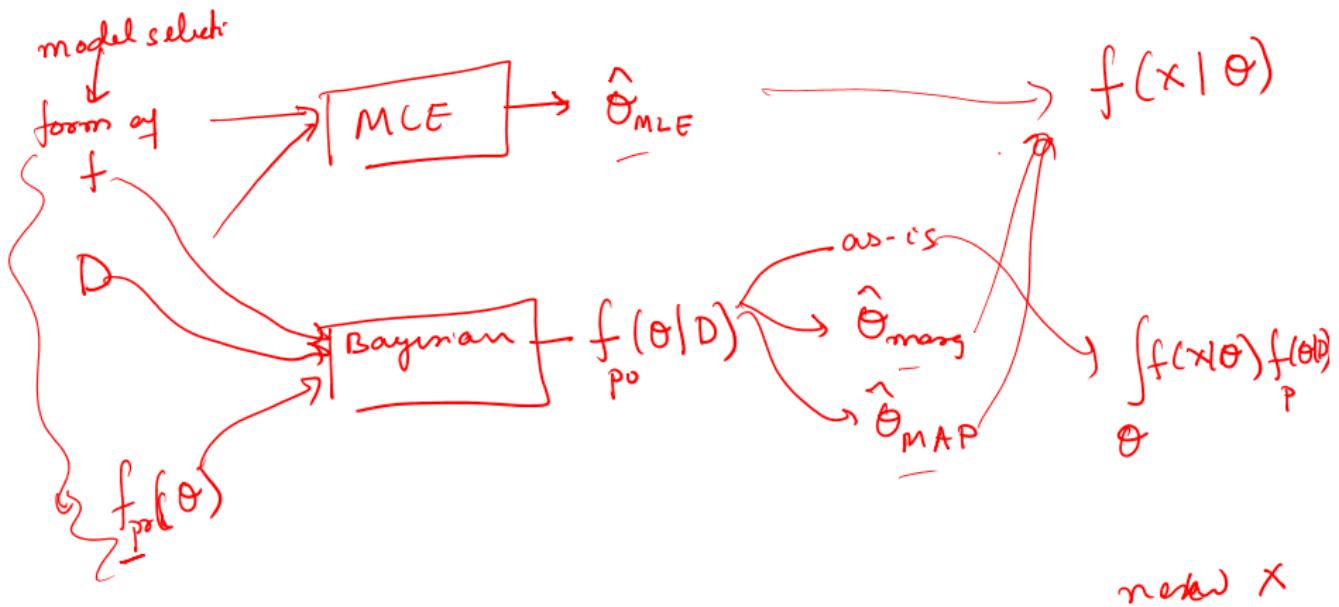
- Expected value of parameters: calculate expected value of $f(\theta|D)$

$$\hat{\theta}_{\text{marg}} = E[\theta] = \int_{\theta} \theta \cdot f_{\text{pos}}(\theta|D) d\theta$$

- MAP estimate: use $\max_{\theta} f(\theta|D)$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{arg\,max}} f(\theta|D)$$

Overall pipeline for MLE Vs Bayesian



Example:
Bayesian estimation of
Bernoulli/Binomial parameter p

Bayesian estimation of Bernoulli parameter

$$\theta \in [0 \dots 1]$$

- Choose a prior distribution over parameter $\underline{\theta}$ or \underline{p} of Bernoulli

$$f_{\text{pr}}(\theta) \sim U(0,1)$$

- Data \underline{D} has n are ones and remaining $N-n = m$ are 0s.

$$D = \{n, m\} \quad D = \{x_1, x_2, \dots, x_N\} \quad \text{eg: } \{0, 1, 1, \dots, 0 \dots 0\}$$
$$f(D|\theta) = \prod_{i=1}^n f(x_i|\theta) = \theta^n (1-\theta)^{m-n} = m$$

- Posterior distribution is:

$$f(\theta|D) = \frac{f(\theta) f(D|\theta)}{\int_{\theta'} f(\theta') f(D|\theta')} = \frac{1 \cdot \theta^n (1-\theta)^m}{\int_{\theta'} \theta'^n (1-\theta')^m}$$

$$f_{p\theta}(\theta|D) = \frac{(1-\theta)^m (\theta)^n}{Z} \quad 0 \leq \theta \leq 1$$

~~$f_{p\theta}$~~ $Z \leftarrow \text{normalizer}$.

mode of $f_{p\theta}(\theta|D)$

$$\begin{aligned} & \max (1-\theta)^m \theta^n \\ & \stackrel{\theta}{=} \frac{d}{d\theta} [(1-\theta)^m \theta^n] = -m(1-\theta)^{m-1} \theta^n + n\theta^{n-1}(1-\theta)^m \\ & \Rightarrow -m\theta + n(1-\theta) = 0 \\ & \theta(n+m) = n \Rightarrow \theta = \frac{n}{m+n} \end{aligned}$$

Beta Random Variable (Generic defn. of Beta)

X is a Beta Random Variable: $X \sim \text{Beta}(a, b)$

* Probability Density Function (PDF): (where $a, b > 0$)

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

$$B(z_1, z_2) = \frac{\Gamma(z_1) \Gamma(z_2)}{\Gamma(z_1 + z_2)}$$

$z_1, z_2 \in \text{integers}$
 $(z_1-1)! (z_2-1)!$
 $\frac{(z_1+z_2-1)!}{(z_1+z_2)!}$

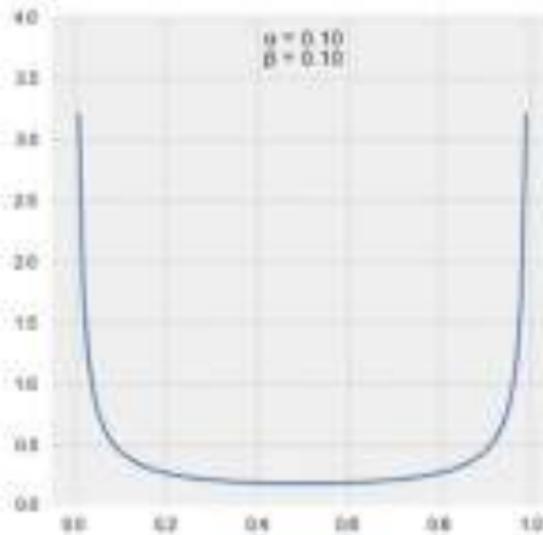
* Symmetric when $a = b$

$$E[X] = \frac{a}{a+b}$$

$$\text{Var}(X) = \frac{ab}{(a+b)^2 (a+b+1)}$$

$$\text{Mode} : \frac{a-1}{a+b-1}$$

The shapes of the Beta distribution



Beta distribution is the
distribution of probabilities

More properties of Beta distributions

- Uniform distribution $U(0,1) = B(1,1)$

$$f(\theta | a, b) = \frac{\theta^{a-1} (1-\theta)^{b-1}}{B(a, b)} = \frac{1 \cdot 1}{B(a, b)} = 1$$

- Relationship between Beta and Gamma distribution

- Let $Y = G(a, 1)$ and $W = G(b, 1)$

$$f(y | a, 1) = e^{-y} y^{a-1} / \Gamma(a)$$

$$f(x | \alpha, \gamma) = \frac{\gamma^x}{\Gamma(\alpha)} e^{-\gamma x} (1-x)^{\alpha-1}$$
$$f(w | b, 1) = \frac{e^{-w}}{\Gamma(b)} w^{b-1}$$

- The $X = Y/(Y+W)$ follows a Beta distribution $B(a, b)$

$$X = \frac{Y}{Y+W} \text{ Then } X \sim B(a, b)$$

Expected value of the posterior of Binomial

$$f_{P^0}(\theta | D) \equiv \frac{\theta^n (1-\theta)^m}{Z} = B(a=n+1, b=m+1)$$

$$D = \{n, m\}$$

p p

#1s #0s

$$\hat{\theta}_{\text{marg}} = \frac{n+1}{n+m+2}$$

Laplace smoothing..

Bass guitar example:

$$\hat{\theta}_{\text{marg}} = \frac{1}{q}$$

Contrast with MLE

$$\hat{\theta}_{\text{MLE}} = \frac{n}{m+n}$$

$$\hat{\theta}_{\text{MLE}} = \frac{0}{7}$$

Lecture 18-Bayesian Estimates Cont.pdf

Overview of parameter estimation

- We have a density function $f(X; \theta)$ whose parameters θ are unknown.

- We have a dataset D of n independent observations from f

- D is random variable denoting X_1, X_2, \dots, X_n where each $X_i \sim f(X; \theta)$

- We use any method to get an estimate $\hat{\theta} = A(D)$ as some function of D . Thus $\hat{\theta}$ is also a R.V.

- Alternative notations $\hat{\theta}_n, \hat{\theta}_D$ to stress that estimate depends on D .

Example: $f(x; \theta) \equiv N(x; \theta \equiv (\mu, \sigma^2))$ $x \equiv$ height of people.

$$D \equiv \{X_1, X_2, \dots, X_{10}\}$$

$$\hat{\theta}_{10} \equiv A(D) = \frac{v_1 + v_2 + \dots + v_{10}}{\text{Sample-mean } 10}$$

$$\underline{f(x; \theta)} \quad \text{hidden}$$

$$D \quad X_1, X_2, \dots, X_n$$

$$X_1 = v_1, X_2 = v_2, \dots, X_n = v_n$$

$$\hat{\theta} \equiv \underline{A(D)}$$

$$\hat{\theta}_D \quad \hat{\theta}_n$$

$x \equiv$ height of people.

Risk of an estimate — theoretical exercise

- Let $\theta, \hat{\theta}_n$ be actual and estimated quantities.

- Risk = Expected square error

$$\underline{E_D [(\hat{\theta}_n - \theta)^2]}$$

$$\begin{aligned}\hat{\theta}_n^1 &= 5.7 \text{ ft} \\ \hat{\theta}_n^2 &= 5.9 \text{ ft}\end{aligned}$$

$$\int (\hat{\theta}_n - \theta)^2 p(D) dD \quad D = \{x_1, \dots, x_n\}$$

$$\int_{x_1} \int_{x_2} \dots \int_{x_n} \underbrace{A(x_1, x_2, \dots, x_n)}_{\hat{\theta}_n} - \theta)^2 f(x_1) f(x_2) \dots f(x_n) dx_1 \dots dx_n \quad \theta^{\infty} = 6.0 \text{ ft}$$

Bias and Variance

$\hat{\theta}_n$ is a R.V

- Expected value of $E_D[\hat{\theta}_n] = \int_D A(D)f(D) = \int_{X_1} \dots \int_{X_n} A(X_1, \dots, X_n) \prod_i f(X_i) dX_1 dX_2 \dots dX_n$

Bias $E_p[\hat{\theta}_n] - \theta$

Variance $E_D[(\hat{\theta}_n - E_p[\hat{\theta}_n])^2] = E_p[\hat{\theta}_n^2] - (E_p[\hat{\theta}_n])^2$

$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$

$$\xrightarrow{\text{To show}} E_D[(\hat{\theta}_n - \theta)^2] = (E_D[\hat{\theta}_n] - \theta)^2 + \text{Var}(\hat{\theta}_n)$$

Proof:

$$\begin{aligned} E_D[(\hat{\theta}_n - \theta)^2] &= E_D[(\hat{\theta}_n - E_D[\hat{\theta}_n] + E_D[\hat{\theta}_n] - \theta)^2] \\ &= E_D[(\hat{\theta}_n - E_D[\hat{\theta}_n])^2] + E_D[(E_D[\hat{\theta}_n] - \theta)^2] \\ &\quad - 2 E_D[(\hat{\theta}_n - E_D[\hat{\theta}_n])(E_D[\hat{\theta}_n] - \theta)] \\ &= \text{Var}(\hat{\theta}_n) + (E_D[\hat{\theta}_n] - \theta)^2 - [2(E_D[\hat{\theta}_n] - E_D[\hat{\theta}_n])] \\ &\quad = 0 \end{aligned}$$

Estimating CDF of any scalar random variable

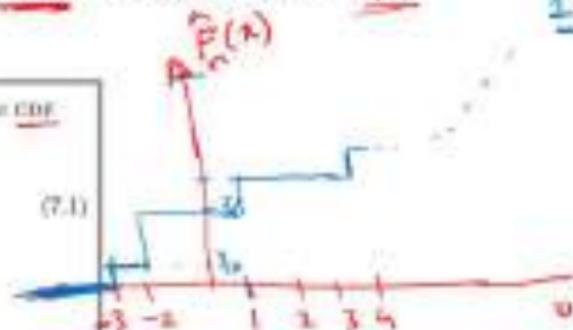
- Given sample: $D = \{X_1, X_2, \dots, X_n\}$ sampled i.i.d from an unknown $f(X)$
- Goal: estimate CDF function using D

7.1 Definition. The empirical distribution function \hat{F}_n is the CDF that puts mass $1/n$ of each data point X_i . Formally,

$$\hat{P}(X \leq x) = \hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}$$

where:

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$



Example: $n = 10, D = \{-3, -2, -1, 1, 2, 2, 4, 6, 8, 10\}$

Analyzing Bias, Variance, and Risk of empirical CDF

Bias: at a x

$E_D[\hat{F}_n(x)]$

$\hat{F}_n(x)$ is unbiased.

$$\begin{aligned} &= E_D \left[\frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \right] \\ &= \frac{1}{n} \sum_{i=1}^n E_D[I(X_i \leq x)] = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^x f(x_i) dx_i \\ &= \frac{1}{n} \sum_{i=1}^n \int_{x_i}^x I(X_i \leq x) f(x_i) dx_i \underbrace{\int f(x_1) dx_1}_{x_1=1} \underbrace{\int f(x_2) dx_2}_{x_2=1} \cdots \underbrace{\int f(x_n) dx_n}_{x_n=1} \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{\int I(X_i \leq x) f(x_i) dx_i}_{x_i = -\infty} \underbrace{\text{except } i}_{x_i = 1} = \frac{n}{n} F(x) = \underline{F(x)} \end{aligned}$$

Variance $\hat{F}_n(x)$

$$\begin{aligned} \text{Var}_D\left(\left[\sum_{i=1}^n I(X_i \leq x)\right]\right) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}_{X_i}\left(\left[I(X_i \leq x)\right]\right) \\ &= \frac{1}{n^2} \cdot n \text{Var}_{X_1}\left(I(X_1 \leq x)\right) \\ &\leftarrow = \frac{1}{n} \left(\left[E_{X_1}\left(I(X_1 \leq x)^2\right) \right] - F(x)^2 \right) \\ &= \frac{1}{n} \left(1 \cdot F(x) - F(x)^2 \right) \\ &= \frac{1}{n} \left(F(x) - F(x)^2 \right) \\ &= \frac{1}{n} F(x) (1 - F(x)) \end{aligned}$$

Non-parametric density estimation

Reading material: https://faculty.washington.edu/yenchi/18W_425/Lec6_hist_KDE.pdf

Motivation

- $D = \{X_1, X_2, \dots, X_n\}$ sampled i.i.d from an unknown $f(X)$
- Estimate $\hat{f}(x)$ without committing on a specific parametric form of $f(X)$
- Why not use empirical CDF?

- Too inefficient to maintain. Need to store entire data
 - Too jerky. Non-zero density at observed points, zero elsewhere.
-
- Density estimation: assume some form of smoothness of $f(X)$

Lecture 19-NonParametricDensity.pdf

Non-parametric density estimation

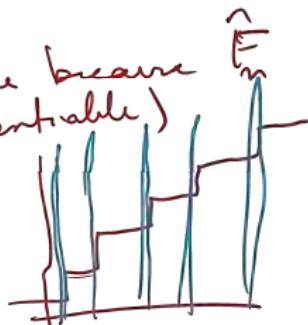
Reading material: https://faculty.washington.edu/yenchi/18W_425/Lec6_hist_KDE.pdf

Motivation

- $D = \{X_1, X_2, \dots, X_n\}$ sampled i.i.d from an unknown $f(X)$
- Estimate $\hat{f}(x)$ without committing on a specific parametric form of $f(X)$
- Why not use empirical CDF?

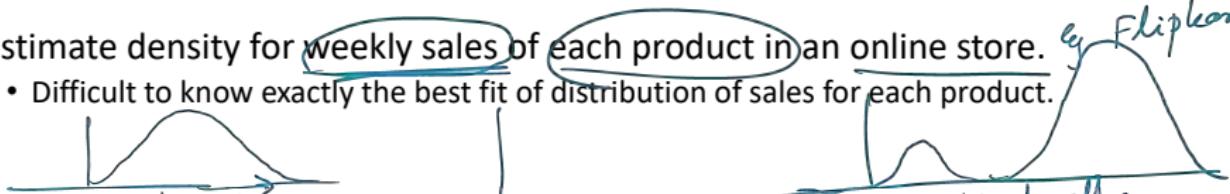
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$$

$$\hat{f}_n(x) = \frac{\partial}{\partial x} \hat{F}_n(x) \quad (\text{Not very usable because is not differentiable})$$



- Too inefficient to maintain. Need to store entire data
- Too jerky. Non-zero density at observed points, zero elsewhere.
- Density estimation: assume some form of smoothness of $f(X)$

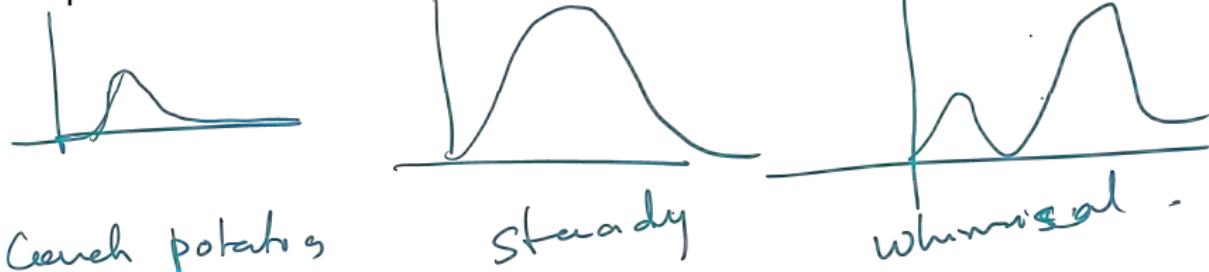
Real-life examples where parametric form is hard.

- Estimate density for weekly sales of each product in an online store.
 - Difficult to know exactly the best fit of distribution of sales for each product.

Lamp

Umbrella

Flipkart
- Estimate distribution of number of steps taken by any arbitrary individual on a phone.

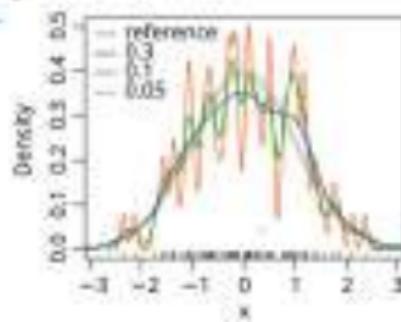


Density estimation methods

- Histogram



- Kernel density



Histogram

- Let $X \in [0,1]$.
- To estimate $\hat{f}_n(x)$ using $D: \{x_1, x_2, \dots, x_n\}$
- Partition the $[0,1]$ range into M equal sized bins.

$$B_1 = \left[0, \frac{1}{M}\right], B_2 = \left[\frac{1}{M}, \frac{2}{M}\right], \dots, B_{M-1} = \left[\frac{M-2}{M}, \frac{M-1}{M}\right], B_M = \left[\frac{M-1}{M}, 1\right]$$

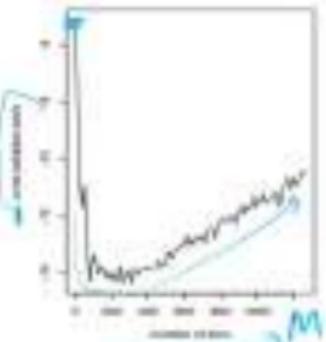
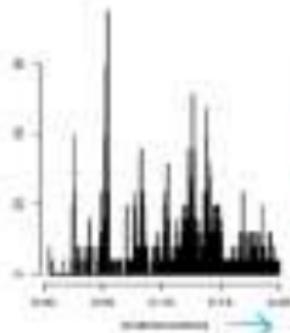
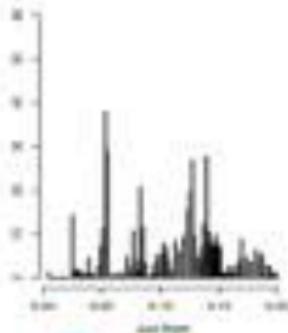
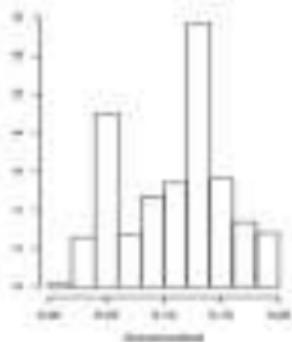


- Estimated density

$$\begin{aligned}\hat{f}_n(x) &= (\text{Fraction of instances in } D \text{ in } B_j) / \text{Width of bin} \\ &= \frac{\sum_{i=1}^n I(x_i \in B_j)}{n} \cdot \frac{N_j}{n} \cdot \frac{1}{M}\end{aligned}$$

$$N_j = \#\{x_i \in D : x_i \in B_j\}$$

20.2 Example. Figure 20.3 shows three different histograms based on $n = 1,266$ data points from an astronomical sky survey. Each data point represents the distance from us to a galaxy. The galaxies lie on a “pencilbeam” pointing directly from the Earth out into space. Because of the finite speed of light, looking at galaxies farther and farther away corresponds to looking back in time. Choosing the right number of bins involves finding a good tradeoff between bias and variance. We shall see later that the top left histogram has too few bins resulting in oversmoothing and too much bias. The bottom left histogram has too many bins resulting in undersmoothing and too few bins. The top right histogram is just right. The histogram reveals the presence of clusters of galaxies. Seeing how the size and number of galaxy clusters varies with time, helps cosmologists understand the evolution of the universe. ■



$$n = 1266$$

Time distance of an observed galaxy from center

Bias, Variance, Risk

$$E[f(x)] - f(x) \equiv \text{Bias}$$

$$E[\hat{f}_n(x)] = E\left[\frac{1}{n} \sum_{i=1}^n I(x_i \in B_j)\right] \text{ if } x \in B_j$$

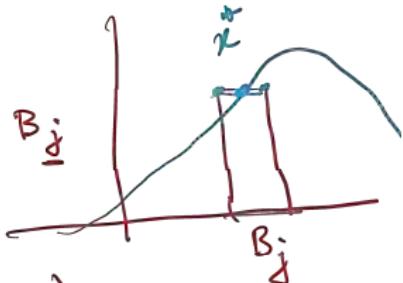
$$= \underline{M} \cdot n \mathbb{E}_x(I(x \in B_\delta))$$

$$= \frac{M}{n} \cdot n \int_{x \in B_i} f(x) dx = \overbrace{M P(x \in B_i)}^{\text{Probability}} = \overbrace{M \left[F\left(\frac{j}{m}\right) - F\left(\frac{(j-1)}{m}\right) \right]}^{\text{Width of subinterval}}$$

$$= F\left(\frac{x}{m}\right) - \overline{F\left(\frac{x-1}{m}\right)} = f(x^*) = f(x)$$

$$\left(\frac{j}{m} - \frac{j-1}{m} \right)$$

$x^* \in B_j$
 $f(x)$ is differentiable



$$\underline{P(X \in B_g)} = \frac{f(x^*)}{M}$$

Variance of $\hat{f}_n(x)$

$$\begin{aligned}\text{Var}(\hat{f}_n(x)) &= \frac{M^2}{n^2} \text{Var}\left(\sum_{i=1}^M I(x_i \in B_j)\right) \quad \text{if } x \in B_j \\ &= \frac{M^2}{n} \text{Var}_x(I(x \in B_j)) \\ &= \frac{M^2}{n} P(x \in B_j)(1 - P(x \in B_j)) \\ &= \frac{M^2}{n} \left[\frac{f(x^*)}{M} \right] \left[1 - \frac{f(x^*)}{M} \right] \\ &= \frac{M f(x^*)}{n} - \frac{f(x^*)^2}{n}\end{aligned}$$

Variance increases as M is increased
 $M = \# \text{ of bins:}$

$$\begin{aligned}
 \text{Risk}(\hat{f}_n(x)) &= \text{Bias}^2 + \text{Var} \\
 &= \underbrace{[f(\bar{x}^*) - f(\bar{x})]^2}_{\text{Bias}} + \frac{Mf(\bar{x}^*)}{n} - \frac{\underline{f(x^*)}}{n}^2 \quad \bar{x}^* \in B_i \\
 &= f'(\tilde{x})(\bar{x}^* - \bar{x}) + \frac{Mf(\bar{x}^*)}{n} - \frac{\underline{f(x^*)}}{n}^2 \\
 &\stackrel{\text{MV thm:}}{\leq} \underbrace{|f'(\tilde{x})| \left\{ \frac{1}{M} \right\}}_{\text{UB on bias}} + \underbrace{\frac{M}{n} \underline{f(x^*)} - \frac{\underline{f(x^*)}}{n}}_{\text{variance}}
 \end{aligned}$$

Bias drops with increasing M vs Variance
which increases.

Kernel Density Estimation

- One of the most convenient and accurate ways to estimate any density.
- Given data sample
 $D = \{x_1, x_2, \dots, x_n\}$
- Assume a special function called a kernel that puts a density mass around each training point.

→ 1. $K(x)$ is symmetric.

2. $\int_x K(x)dx = 1.$

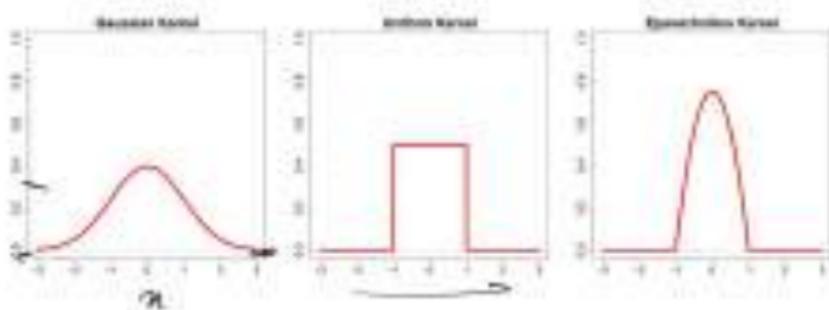
3. $\lim_{x \rightarrow -\infty} K(x) = \lim_{x \rightarrow +\infty} K(x) = 0.$

- Examples of kernels

Gaussian $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$

Uniform $K(x) = \frac{1}{2} I(-1 \leq x \leq 1),$

Epanechnikov $K(x) = \frac{3}{4} \cdot \max\{1 - x^2, 0\}.$

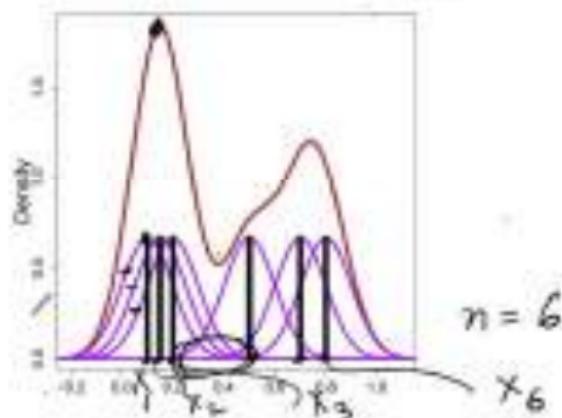


Kernel density estimator

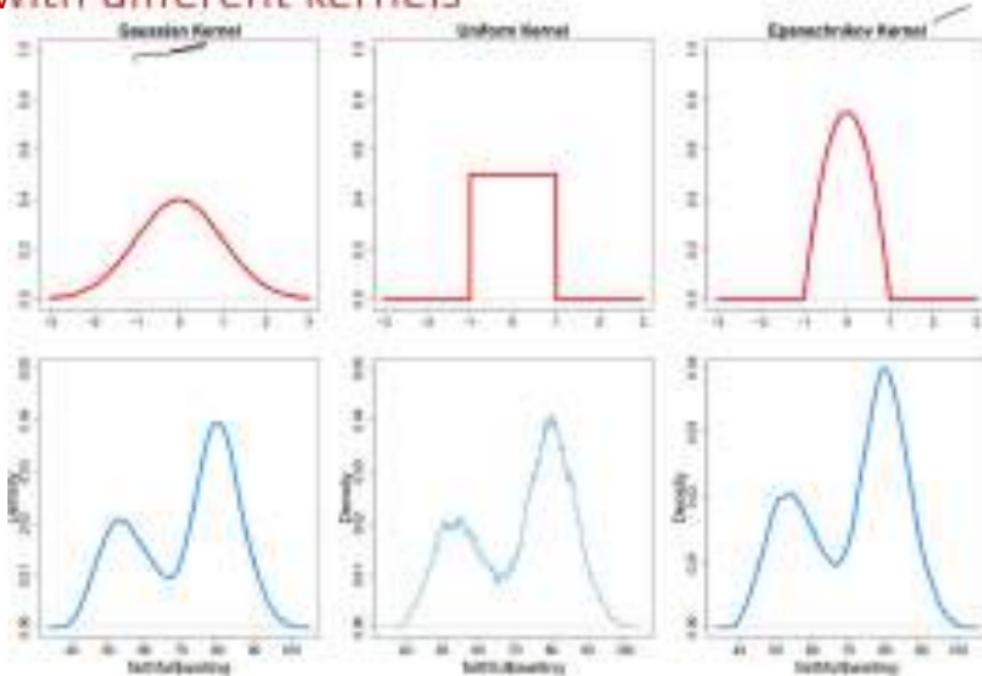
$$h = \frac{l}{M}$$

20.12 Definition. Given a kernel K and a positive number h called the bandwidth, the kernel density estimator is defined to be

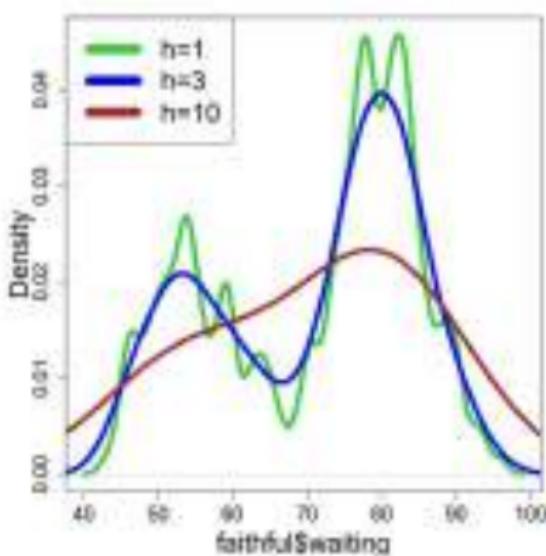
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \quad (20.21)$$



KDE with different kernels



Effect of kernel width



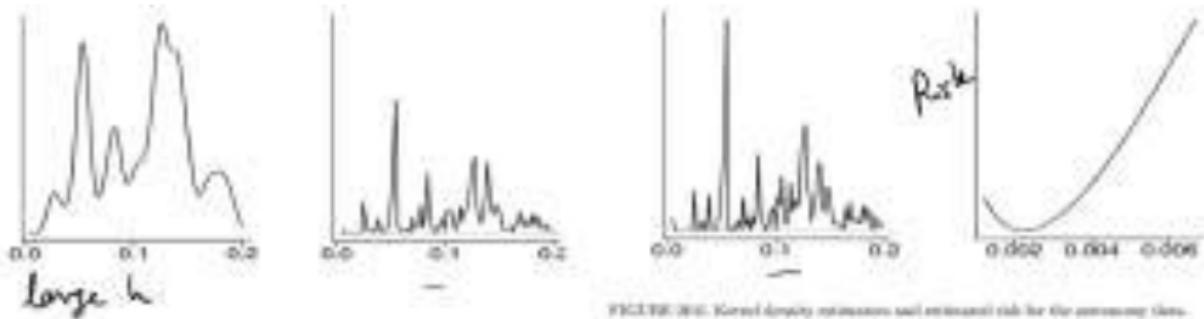


FIGURE 20. Kernel density estimation (solid line) and true (dashed line) rate for the anonymous times. Top left: communication. Top right: just right (handwritten digits by cross-validation). Bottom left: uncommunicated. Bottom right: cross-validation curve as a function of λ . The bandwidth was chosen to fit the value of λ where the curve is a parabola.

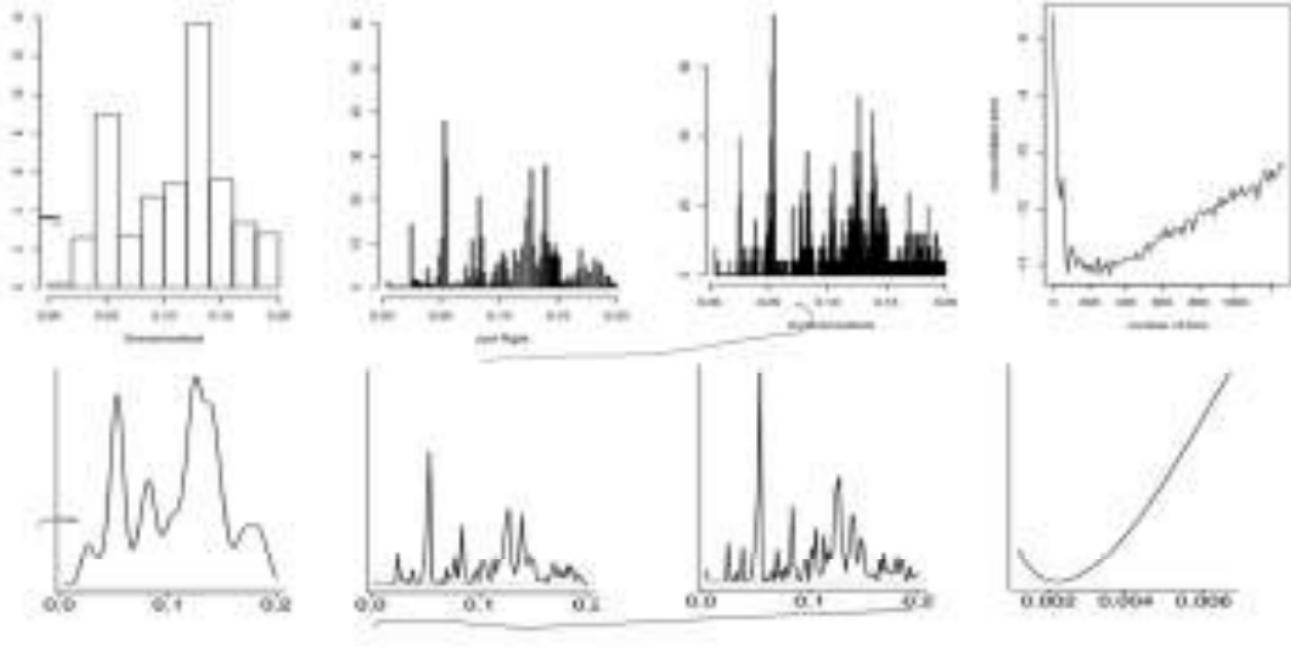


FIGURE 21.2 Recent trends in inflation rates and estimated risk that the economy stays “too hot” (oversimplified). Top right: past eight 10-month periods chosen by error-minimization. Bottom left: undersimplified. Bottom right: same solidification curve as a function of Δu_0 . The undersimplification was chosen to be three times as large as the oversimplification.

Demo

[https://colab.research.google.com/github/fbeilstein/machine_learning/
blob/master/lecture_15_kernel_density_estimation.ipynb#scrollTo=pGU9KIkxe-FY](https://colab.research.google.com/github/fbeilstein/machine_learning/blob/master/lecture_15_kernel_density_estimation.ipynb#scrollTo=pGU9KIkxe-FY)

Lecture 20-NonParametricDensity2.pdf

Analyzing the bias and variance of KDE

Bias:

$$E[\hat{f}_n(x)]$$

$$\mathbb{E}[K_h(x, X)] - f(x) \approx \frac{1}{2} \sigma_h^2 h^2 f''(x).$$

Variance:

$$\mathbb{V}[\hat{f}_n(x)] \approx \frac{f(x) \int K^2(x) dx}{n h_n},$$

Risk:

39.14 Theorem. Under mild assumptions on f and K ,

$$R(f, \hat{f}_n) = \frac{1}{4} \sigma_h^2 h^4 \int (f'(x))^2 + \frac{\int K^2(x) dx}{nh} \quad (39.21)$$

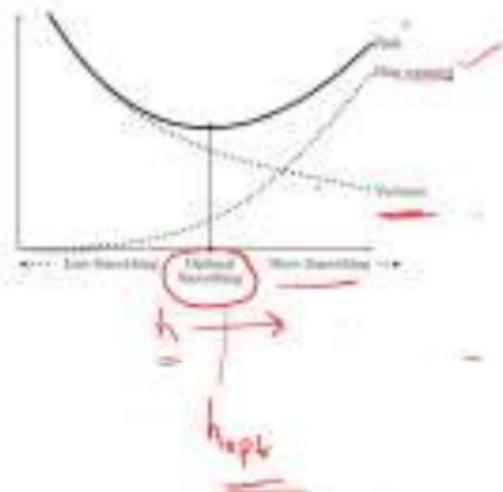
where $\sigma_h^2 = \int x^2 K(x) dx$. The optimal bandwidth is

$$h^* = \frac{\sqrt{c_1 c_2 c_3 n^{-1/6}}}{\sqrt{c_0}} \quad (39.22)$$

where $c_1 = \int x^2 K(x) dx$, $c_0 = \int K(x)^2 dx$ and $c_2 = \int (f''(x))^2 dx$. With this choice of bandwidth,

$$R(f, \hat{f}_n) = \frac{c_2}{c_0^{1/3}}$$

(for some constant $c_3 > 0$)



Proof for Bias $E[\hat{f}_n(x)] - f(x)$

$$\begin{aligned}
 E[\hat{f}_n(x_0)] &= \int_{x_1}^{x_n} \left(\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) f(x_i) - f(x_n) \right) dx_1 \dots dx_n \\
 &= \frac{n}{n} \int_x K\left(\frac{x_0 - x}{h}\right) f(x) \frac{dx}{h} \quad \text{Let } u = \frac{x - x_0}{h} \quad \begin{matrix} \text{small} \\ h \downarrow \end{matrix} \\
 &= \int_u K(u) f(x_0 + uh) du \\
 &= \int_k(u) \left[f(x_0) + uh f'(x_0) + \frac{u^2 h^2}{2} f''(x_0) + O(h^3) \right] du \\
 &= f(x_0) \underbrace{\int_u k(u) du}_{\substack{u=0 \\ u=1}} + h f'(x_0) \underbrace{\int_u k(u) du}_{u=0} + \frac{h^2}{2} f''(x_0) \underbrace{\int_u u^2 k(u) du}_{\substack{u=0 \\ u=1}} + O(h^3) \\
 &\quad \boxed{M_K}
 \end{aligned}$$

$$\begin{aligned}
 f(x_0) + \frac{h^2 f''(x_0)}{2} M_K \\
 \text{where} \\
 M_K = \int_u^2 k(u) du
 \end{aligned}$$

Proof for Variance

$$\begin{aligned}\text{Variance}(\hat{f}_n(x_0)) &= \text{Var}\left(\frac{1}{nh} \sum_i k\left(\frac{x_i - x_0}{h}\right)\right) \\ &= \frac{1}{nh^2} \text{Var}_x\left(k\left(\frac{x - x_0}{h}\right)\right) = E[k^2] - \underbrace{E[k]}_{}^2 \\ &\leq \frac{1}{nh^2} \int_{-\infty}^{\infty} k\left(\frac{x - x_0}{h}\right)^2 f(x) dx \\ &\quad \vdots \\ &\quad \vdots \\ &\quad \vdots \\ &\quad \vdots \\ \underbrace{\frac{1}{nh} \hat{f}_n(x_0) \sigma_k^2}_{\cdot} + O(h^2) \text{ where } \sigma_k^2 = \int_u k^2(u) du\end{aligned}$$

$$\text{Risk}(\hat{f}_n(x_0), f(x)) = \text{Bias}(\hat{f}_n(x_0))^2 + \text{Var}(\hat{f}_n(x_0))$$

Optimal h

$$\min_h \left\{ \frac{1}{4} h^4 f''(x_0)^2 u_n^2 + \frac{1}{nh} f(x_0) \sigma_u^2 \right\}$$

\tilde{R}

$$\frac{\partial \tilde{R}}{\partial h} = h^3 f''(x_0)^2 u_n^2 - \frac{1}{nh^2} f(x_0) \sigma_u^2$$

$$h_{\text{opt}} \stackrel{\circ}{=} h_{\text{opt}}^3 f''(x_0)^2 u_n^2 - \frac{1}{nh^2} f(x_0) \sigma_u^2 = 0$$

$$\Rightarrow h_{\text{opt}}(x_0) \stackrel{\circ}{=} \left[\frac{f(x_0) \sigma_u^2}{n f''(x_0) u_n^2} \right]^{1/2}$$

Evaluate \tilde{R} at $h_{opt} = \frac{C}{n^{4/5}} = O(\underline{\overline{n^{-4/5}}})$

fn of gold f(x) which is unknown

Contrast with same error of
maximum likelihood estimate of mean parameter
eg: μ of $N(\mu, \sigma^2)$

$$\text{Risk}(\hat{\mu}_n, \mu) = \frac{\sigma^2}{n} = O(n^{-1})$$

Convergence analysis of histogram density estimator

$$\text{Bias} [\hat{f}_n(x_0)] \leq |f'(x)| \left(\frac{1}{M}\right) = |f'(x)|h \quad h = \frac{1}{M}$$

$$\text{Var} [\hat{f}_n(x)] = \frac{f(x)}{nh} - \frac{f(x)^2}{n}$$

$$L = \max_x |f'(x)|$$

$$\tilde{R} = h^2 L^2 + \frac{f(x)}{nh} - \frac{f(x)^2}{n}$$

Find h_{opt}

$$\frac{\partial \tilde{R}}{\partial h} = 2h L^2 - \frac{f(x)}{nh^2} \Rightarrow h_{\text{opt}} = \left(\frac{f(x)}{2n L^2} \right)^{1/3}$$

Evaluate \tilde{R} at h_{opt} :

$$\tilde{R}_{\text{opt}} = O(n^{-2/3})$$

Risk reduces at the rate $n^{2/3}$.

Comparing Histogram and KDE

- Risk of histogram reduces at rate $O(n^{-2/3})$
- Risk of KDE reduces at rate $O(n^{-4/5})$

Summary of estimation methods

D

1) Parametric methods:

Assumed or found functional form of density
 $f(x; \theta)$

Estimated θ

Maximum likelihood
estimation

Bayesian estimation.

2) Empirical CDF.

Histograms

3) Non-parametric density

KDE.

Lecture 21-LinearRegression.pdf

Regression

Sunita Sarawagi

CS 215. Fall 2024

Reading: Chapter 9 in Ross Textbook

Problem definition

- So far, we have been estimating the density of single random variables $f(X)$
- In many applications, we need to reason about the distribution of values of a continuous random variable Y but as a function of some other variables $\mathbf{x} = (x_1, \dots, x_k)$
 - Y is called output or response or dependent variable
 - (x_1, \dots, x_k) are called input or covariate or independent variables.
- We want to estimate the conditional density: $f(Y | x_1, \dots, x_k)$
- We are given data samples D of the form:
- $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} = \{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$

Motivating examples

- How does errors in assembled circuits (Y) depend on temperature (x₁), percentage of copper (x₂), humidity (x₃)
 $f(\text{yield} | \text{temperature}, \% \text{ copper}, \text{humidity})$
- Predict stock price tomorrow (Y) as a function of stock price in the last 7 days (x₁, x₂, ..., x₇)
- Express CPU temperature (Y) as a function of workload (x₁), ambient temperature (x₂), fan-speed (x₃), chip model (x₄), etc.

How to represent conditional density?

- Many different forms of $f(Y|x)$ will be discussed later.
- A simple form:

$$f(Y|x_1, \dots, x_k) \sim N(\mu_x, \sigma^2), \quad \text{where } \mu_x = \beta_1 x_1 + \dots + \beta_k x_k + \alpha$$

- Parameters of the above model are

$$\beta_1, \beta_2, \dots, \beta_k, \alpha, \sigma^2$$

- Alternate way to view the above:

$$Y = \underbrace{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \alpha}_{\text{deterministic}} + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2)$

random error

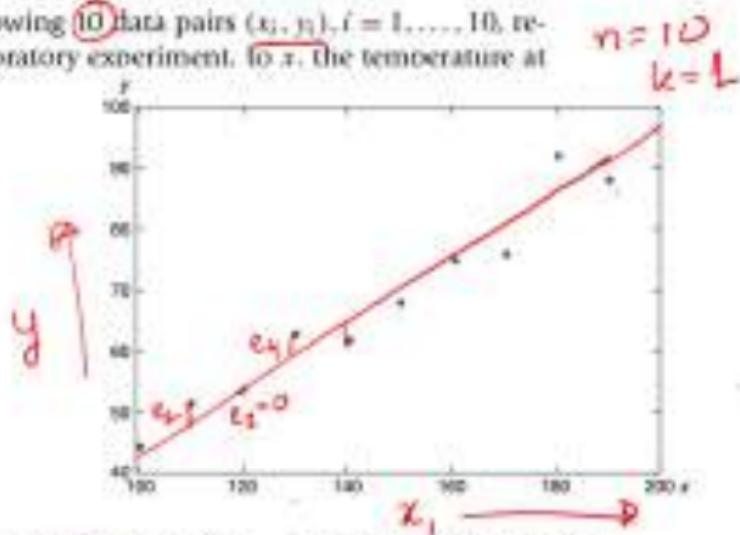
Special case $k = 1$

$$f(y|x_1) \sim N(\beta_0 + \alpha_1 x_1; \sigma^2)$$

$$y = \beta_0 + \alpha_1 x_1 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

Example 9.1.a. Consider the following 10 data pairs (x_i, y_i) , $i = 1, \dots, 10$, relating y , the percent yield of a laboratory experiment, to x , the temperature at which the experiment was run.

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88



Our goal is to use the data to learn α, β, σ so that we can write

$$y_i = \beta x_i + \alpha + e_i \quad e_i \sim N(0, \sigma^2)$$

Notation
Actual unknown parameter
 β, α, σ^2

Estimated parameter using n samples -
 $\hat{\beta}_n, \hat{\alpha}_n, \hat{\sigma}_n^2 \}$ previous notation

Instead following the Ross textbook notation -

$$\begin{matrix} \beta, \alpha, \sigma^2 \\ \uparrow \quad \uparrow \quad \uparrow \\ \beta \quad \alpha \quad \sigma \end{matrix}$$

Estimating parameters from data using Maximum Likelihood

- Given data $D = \{(x_1, y_1), \dots, (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^n$
- Theorem: The parameters β, α that maximize the likelihood of D are:

$$B = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad A = \bar{y} - B \bar{x} \quad [\text{Ignore estimation of } \sigma^2]$$

Proof:

$$\begin{aligned} \text{Log likelihood of the data:} & \quad \log L(B, A) = \sum_{i=1}^n \log P(y_i | x_i) = \sum_{i=1}^n \log \frac{e^{-\frac{(y_i - Bx_i - A)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \\ \max_{A, B} LL(D) &= \sum_{i=1}^n (y_i - Bx_i - A)^2 \end{aligned}$$

$$\frac{\partial LL}{\partial A} = \sum_{i=1}^n (y_i - Bx_i - A)(-1) = 0$$

$$\Rightarrow nA = \sum_{i=1}^n y_i - B \sum_{i=1}^n x_i$$

$$\Rightarrow A = \bar{y} - B\bar{x} \quad \text{where } \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

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

$$\frac{\partial LL}{\partial B} = \sum_{i=1}^n (y_i - Bx_i - A)(-x_i) = 0$$

$$\sum_{i=1}^n (y_i - Bx_i - \bar{y} + B\bar{x})x_i = 0$$

$$\Rightarrow B \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = \sum x_i y_i - \sum x_i \bar{y}$$

-

$$B = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

Analyzing risk of the linear regression parameters

We will show the following results about the mean and variance of the maximum likelihood estimators \hat{A} , \hat{B} of α , β respectively.

- $E[\hat{B}] = \beta$, $E[\hat{A}] = \alpha$: Both are unbiased estimators.

- $Var[\hat{B}] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ $Var(\hat{A}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}$

More generally we can show that both parameters follow a Gaussian distribution with above mean and variance.

$$\hat{B} \sim N\left(\beta, \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}\right), \quad \hat{A} = N(\alpha, Var(\hat{A}))$$

$$B = \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n \bar{x}^2$$

$$B = \sum_{i=1}^n Y_i \left[\frac{x_i - \bar{x}}{SS} \right]$$

$$\Rightarrow B = \sum_{i=1}^n w_i Y_i \Rightarrow B \sim N\left(\sum_{i=1}^n w_i E[Y_i]; \frac{\sum_{i=1}^n w_i^2}{\text{var}(Y_i)}\right)$$

B is normally distributed.

$$E[B] = \sum_{i=1}^n \frac{(x_i - \bar{x})(\beta x_i + \alpha)}{SS} = \beta -$$

sample manipulations.

$$\frac{\sum_i (x_i - \bar{x})(\beta x_i + \alpha)}{\sum_i x_i^2 - n \bar{x}^2} = \beta \left(\sum_{i=1}^n (x_i^2 - \bar{x} x_i) \right) + \alpha \left(\sum_i (x_i - \bar{x}) \right)$$

$$\alpha \sum_i x_i - \alpha n \bar{x} = \\ = \alpha n \bar{x} - \alpha n \bar{x} = 0$$

$$\frac{\beta \left[\sum_{i=1}^n x_i^2 - \bar{x} n \bar{x} \right]}{\sum_i x_i^2 - n \bar{x}^2} = \beta -$$

$$\text{Var}[\beta] = \sum_i \text{Var}(Y_i) \left[\frac{(x_i - \bar{x})^2}{(\sum_i x_i^2 - n \bar{x}^2)^2} \right] = \sigma^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i x_i^2 - n \bar{x}^2}$$

$$\sum_i x_i^2 + n \bar{x}^2 - 2 \bar{x} \sum_i x_i = \sum_i x_i^2 - n \bar{x}^2$$

$$\text{Var}(B) = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\underline{B} \sim N(B; \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}) = N(B; \frac{\sigma^2}{n\sigma_x^2}) \text{ where } \sigma_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$

Distribution of A estimate:

$$A = \bar{Y} - B\bar{X}$$

$$E[A] = E[\bar{Y}] - E[B\bar{X}] \quad \checkmark$$

$$\text{Var}[A] = \text{Var}(\bar{Y}) + \text{Var}(B\bar{X}) \leftarrow \text{erroneous because } \bar{Y} \text{ is not independent of } B$$

\bar{Y} is not independent of B .

$$A = \sum_{i=1}^n Y_i \left(\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{S_{xx}} \right)$$

$A = \bar{Y} - B \bar{x}$

$A = \sum_{i=1}^n Y_i \sigma_i$ *constants.* $\Rightarrow A \sim N\left(\sum_i \sigma_i (\beta x_i + \alpha); \sigma^2 \sum_i \sigma_i^2\right)$

$$\begin{aligned} E[A] &= E[\bar{Y}] - E[B \bar{x}] : \frac{1}{n} E\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n E[Y_i]/n \\ &= [\beta \bar{x} + \alpha] - \bar{x} \beta = \alpha \\ &= \underbrace{\sum_{i=1}^n \beta x_i + \alpha}_{n} \end{aligned}$$

$\text{Var}[A] = \text{homework.}$ $= \beta \bar{x} + \alpha$

Estimating σ^2

- Calculate sum of square of residuals

- Residuals = difference between actual y_i and predicted value $Bx_i + A$

$$SS_E = \sum_{i=1}^n (Y_i - A - Bx_i)^2$$

- The MLE estimate would be:

- The above biased like for normal Gaussian parameters.
- We will use a different method:

The Chi-Square distribution (Section 5.8 of textbook)

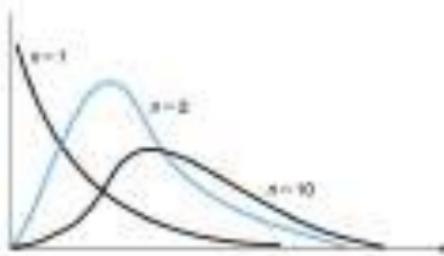
Definition. If Z_1, Z_2, \dots, Z_n are independent standard normal random variables, then X , defined by

$$X = Z_1^2 + Z_2^2 + \dots + Z_n^2 \quad (5.8.1)$$

is said to have a *chi-square distribution with n degrees of freedom*. We will use the notation

$$X \sim \chi_n^2$$

to signify that X has a chi-square distribution with n degrees of freedom.



Deriving the density of χ_n^2 distribution

- Use MGF.
- Consider n=1 first.

$$E[e^{tY}] = E[e^{tZ^2}] \text{ where } Z \sim N(0, 1)$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} e^{tx^2} f_Z(x) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2(1-2t)/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\theta^2} dx \quad \text{where } \theta^2 = (1-2t)^{-1} \\ &= (1-2t)^{-1/2} \frac{1}{\sqrt{2\pi}\theta} \int_{-\infty}^{\infty} e^{-x^2/2\theta^2} dx \\ &= (1-2t)^{-1/2} \end{aligned}$$

General n

- $E_X[e^{\{tX\}}] = E \left[e^{t \sum_i Z_i^2} \right] = \prod_i E \left[e^{t Z_i^2} \right] = (1 - 2t)^{-n/2}$
- The above is MGF of gamma distribution with parameters $(n/2, 1/2)$.

A random variable is said to have a gamma distribution with parameters (α, λ) , $\lambda > 0, \alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Thus, density of χ^2 distribution is

- $$f(x) = \frac{\frac{1}{2} e^{-x/2} \left(\frac{x}{2}\right)^{(n/2)-1}}{\Gamma\left(\frac{n}{2}\right)}, \quad x > 0$$

Expected value of χ_n^2 distribution

- $E[\chi_n^2] = n$ [Can be derived from the MGF]
- $\text{Var}[\chi_n^2] = 2n$

Estimating σ^2

- Calculate sum of square of residuals where

- Residuals = difference between actual y_i and predicted value $Bx_i + A$

$$SS_R = \sum_{i=1}^n (Y_i - A - Bx_i)^2$$

- It can be show that $\frac{SS_R}{\sigma^2}$ follows a Chi-square distribution with $n-2$ degrees of freedom
 - Book has a kind of intuitive proof...

Estimating σ^2

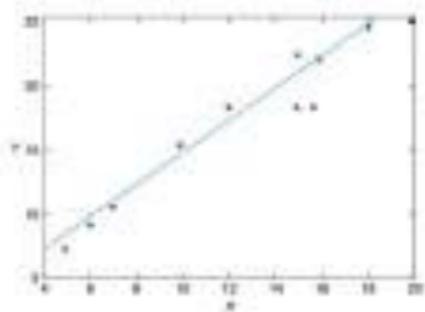
Let estimate of σ^2 be called S.

$$S = \frac{SS_R}{n - 2}$$

S is an unbiased estimate of σ^2 . It is easy to see that $E[S] = \sigma^2$

Example 9.3.a. The following data relate x , the moisture of a wet mix of a certain product, to Y , the density of the finished product.

x_i	y_i
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
20	24.8



Compute the least square fit. Estimate A, B, S

$$y = 2.463 + 1.206x$$

Multi-variable linear regression

Reading material: Section 9.10 of Ross Textbook

General case: $k > 1$

$$f(Y | x_1, \dots, x_k) \sim N(\mu_x, \sigma^2), \quad \text{where } \mu_x = \beta_1 x_1 + \dots + \beta_r x_r + \beta_0$$

Or

$$Y = \beta_1 x_1 + \dots + \beta_r x_r + \beta_0 + e \quad \text{where } e \sim N(0, \sigma^2)$$

Training data D will be denoted as

$$\{(x_{i1}, x_{i2}, \dots, x_{ir}, y_i) : i = 1 \dots n\}$$

MLE estimates of parameters:

Solving the MLE

Solving the MLE

$$\sum_{i=1}^n Y_i = \alpha \beta_0 + \beta_1 \sum_{i=1}^n x_{i1} + \beta_2 \sum_{i=1}^n x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{ik} \quad (9.10.1)$$

$$\sum_{i=1}^n x_{i1} Y_i = \beta_0 \sum_{i=1}^n x_{i1} + \beta_1 \sum_{i=1}^n x_{i1}^2 + \beta_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{i1} x_{ik}$$

⋮

$$\sum_{i=1}^n x_{ik} Y_i = \beta_0 \sum_{i=1}^n x_{ik} + \beta_1 \sum_{i=1}^n x_{ik} x_{i1} + \beta_2 \sum_{i=1}^n x_{ik} x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{ik}^2$$

Matrix notation for k-dimensional covariates.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

then \mathbf{Y} is an $n \times 1$, \mathbf{X} an $n \times p$, $\boldsymbol{\beta}$ a $p \times 1$, and $\boldsymbol{\epsilon}$ an $n \times 1$ matrix where $p = k + 1$.

The regression problem now becomes: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$\begin{aligned}
 X'X &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 0 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
 &= \begin{bmatrix} n & \sum_j x_{1j} & \sum_j x_{1j} & \cdots & \sum_j x_{1j} \\ \sum_i x_{i1} & \sum_j x_{1j}^2 & \sum_i x_{i1} x_{1j} & \cdots & \sum_i x_{i1} x_{1j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{ik} & \sum_j x_{ik} x_{ij} & \sum_i x_{ik} x_{ij} & \cdots & \sum_j x_{ik}^2 \end{bmatrix}
 \end{aligned}$$

and

$$X'Y = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_{1i} Y_i \\ \vdots \\ \sum_i x_{ni} Y_i \end{bmatrix}$$

Solving the MLE

$$\sum_{i=1}^n Y_i = \beta_0 + \beta_1 \sum_{i=1}^n x_{i1} + \beta_2 \sum_{i=1}^n x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{ik} \quad (9.10.1)$$

$$\sum_{i=1}^n x_{i1} Y_i = \beta_0 \sum_{i=1}^n x_{i1} + \beta_1 \sum_{i=1}^n x_{i1}^2 + \beta_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{i1} x_{ik}$$

⋮

$$\sum_{i=1}^n x_{ik} Y_i = \beta_0 \sum_{i=1}^n x_{ik} + \beta_1 \sum_{i=1}^n x_{ik} x_{i1} + \beta_2 \sum_{i=1}^n x_{ik} x_{i2} + \cdots + \beta_k \sum_{i=1}^n x_{ik}^2$$



$$X'XB = X'Y$$



$$B = (X'X)^{-1}X'Y$$

where X' is the transpose of X .

Example

Get least square estimate on this data with last column as y.

Table F.2					
Age (years)	Diameter (1000 ft)	Root- thickness	Specific Gravity	Diameter at Breast Height (inches)	
1	10	1.5	.85	30	10.5
2	15	2.0	.84	30	10.8
3	20	2.5	.83	30	11.2
4	25	3.0	.82	30	11.6
5	30	3.5	.80	30	12.0
6	35	4.0	.79	30	12.4
7	40	4.5	.78	30	12.8
8	45	5.0	.77	30	13.2
9	50	5.5	.76	30	13.6
10	55	6.0	.75	30	14.0
11	60	6.5	.74	30	14.4
12	65	7.0	.73	30	14.8
13	70	7.5	.72	30	15.2
14	75	8.0	.71	30	15.6
15	80	8.5	.70	30	16.0

$$y = 11.54873 + 0.05728x_1 + 0.08712x_2 + 7.3323(x_3)$$

Non-parametric regression

Lecture 22-MultipleLinearRegression.pdf

Estimating σ^2

- Calculate sum of square of residuals

- Residuals = difference between actual y_i and predicted value $Bx_i + A$

$$SS_R = \sum_{i=1}^n (Y_i - A - Bx_i)^2$$

- The MLE estimate would be:

$$\hat{\sigma}_{MLE}^2 = \frac{\sum_{i=1}^n (y_i - (\hat{A} + \hat{B}x_i))^2}{n}$$

Example $\sigma = 2$

$$\hat{\sigma}_{MLE} = 0$$

- The above is biased like for normal Gaussian parameters.

- We will use a different method:

The Chi-Square distribution (Section 5.8 of textbook)

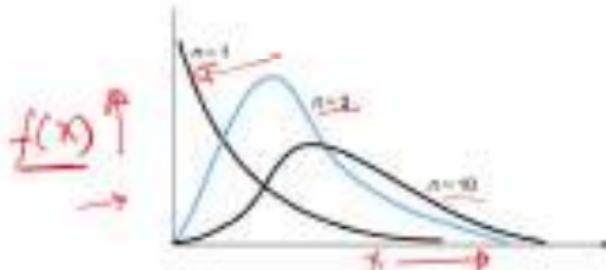
Definition. If Z_1, Z_2, \dots, Z_n are independent standard normal random variables, then X , defined by

$$\underline{X} = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \quad (5.8.1)$$

is said to have a chi-square distribution with n degrees of freedom. We will use the notation

$$\underline{X} \sim \chi_n^2$$

to signify that X has a chi-square distribution with n degrees of freedom.



Deriving the density of χ_n^2 distribution

- Use MGF.

- Consider $n=1$ first.

MGF + $\rightarrow E[e^{tX}] = E[e^{tZ^2}]$ where $Z \sim N(0, 1)$

$$\begin{aligned} &= \int_{-\infty}^{\infty} e^{tx^2} f_Z(x) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{tx^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2t)x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/\hat{\sigma}^2} dx \quad \text{where } \hat{\sigma}^2 = (1-2t)^{-1} \\ &= (1-2t)^{-1/2} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \int_{-\infty}^{\infty} e^{-x^2/\hat{\sigma}^2} dx \\ &= (1-2t)^{-1/2} \end{aligned}$$

General n

$$\bullet E_X[e^{\{tX\}}] = E \left[e^{t \sum_i Z_i^2} \right] = \prod_i E \left[e^{t Z_i^2} \right] = (1 - 2t)^{-n/2}$$

- The above is MGF of gamma distribution with parameters (n/2, 1/2).

A random variable is said to have a gamma distribution with parameters α, λ , $\lambda > 0, \alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha \alpha!}{x^{\alpha-1} \Gamma(\alpha)} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- Thus, density of χ^2 distribution is

$$\bullet f(x) = \frac{1}{2} e^{-x/2} \left(\frac{x}{2} \right)^{(n/2)-1} \frac{1}{\Gamma(n/2)}, \quad x > 0$$

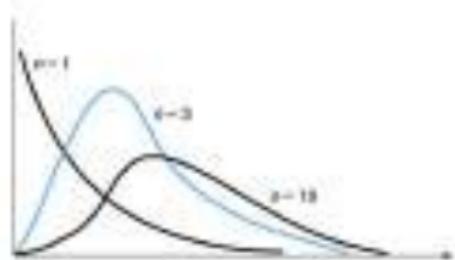
Expected value of χ_n^2 distribution [Homework]

- $E[\chi_n^2] = n$ [Can be derived from the MGF]

$$\frac{d}{dt} MGF(\chi_n^2)$$

- $\text{Var}[\chi_n^2] = 2n$

- Mode = $\max\{n-2, 0\}$



Estimating σ^2

- Calculate sum of square of residuals where

- Residuals = difference between actual y_i and predicted value $Bx_i + A$

$$SS_R = \sum_{i=1}^n (Y_i - A - Bx_i)^2$$

$$\sum_{i=1}^n \frac{(Y_i - Bx_i - \alpha)^2}{\sigma^2} \sim \chi_n^2$$

$Y_i \sim N(\mu_{Y_i}, \sigma^2)$
 $Z_i = \frac{Y_i - (\alpha + Bx_i)}{\sigma}$
 $Z_i \sim N(0, 1)$

- It can be show that $\frac{SS_R}{\sigma^2}$ follows a Chi-square distribution with $n-2$ degrees of freedom

- Book has a kind of intuitive proof...

A, B are functions of Y_i

∴ each of the n terms in SS_R are not independent of each other

Estimating σ^2

Let estimate of σ^2 be called S.

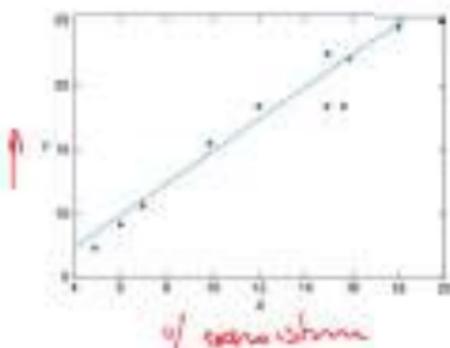
$$\hat{\sigma}^2 = S = \frac{SS_R}{n - 2}$$

S is an unbiased estimate of σ^2 . It is easy to see that $E[S] = \sigma^2$

Example 9.3.a. The following data relate x , the moisture of a wet mix of a certain product, to Y , the density of the finished product.

x_i	y_i	$x_i y_i$	x_i^2	$S S R$
5	7.4	37	25	
6	9.3			
7	10.6			
10	15.4			
12	18.1			
15	22.2			
18	24.1			
20	24.8			

Compute the least square fit. Estimate A, B, S.



$$B = \frac{\sum x_i y_i - \bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$y = 2.463 + 1.206x$$

Multi-variable linear regression

Reading material: Section 9.10 of Ross Textbook



General case: $k > 1$

$$f(Y | \underbrace{x_1, \dots, x_k}_{\alpha}) \sim N(\mu_x, \sigma^2), \text{ where } \mu_x = \underbrace{\beta_1 x_1 + \dots + \beta_k x_k + \beta_0}_{\alpha}$$

Or

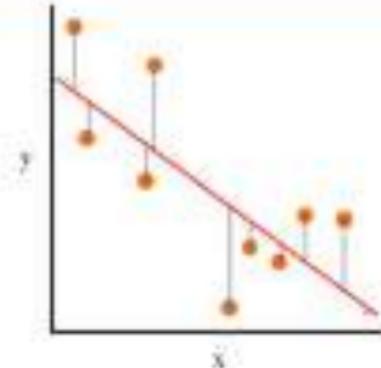
$$Y = \underbrace{\beta_1 x_1 + \dots + \beta_k x_k + \beta_0}_{\alpha} + e \quad \text{where } e \sim N(0, \sigma^2)$$

Training data D will be denoted as

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1 \dots n\}$$

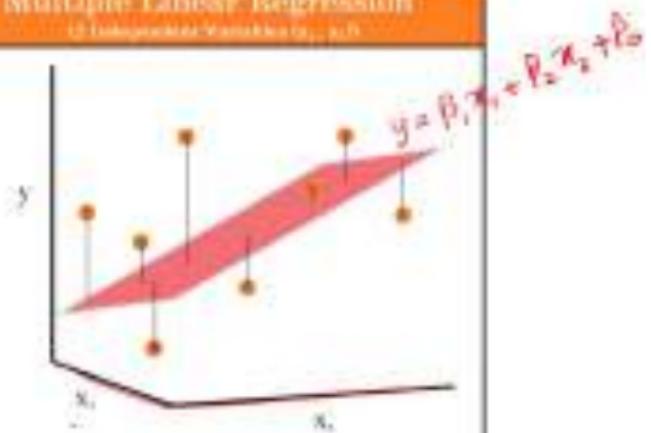
$$\{(x_i, y_i) : i = 1 \dots n\}$$

Simple Linear Regression



Multiple Linear Regression

(\hat{y} = Interpreted Variable (y_1, \dots, y_n))



Parameter estimation using MLE

$$LL(D) = \sum_{i=1}^n \log e^{-\frac{(Y_i - (B_0 + \dots + B_k x_k))}{\sigma^2}} - n \log(2\pi\sigma)$$

$$\frac{\partial LL}{\partial B_0} = 0 \quad \sum_{i=1}^n (Y_i - (B_0 + \dots + B_k x_k))(-1) = 0$$

$$\frac{\partial LL}{\partial B_1} = 0 \quad \sum_{i=1}^n (Y_i - (B_0 + \dots + B_k x_k))x_1 = 0$$

⋮

$$\frac{\partial LL}{\partial B_k} = 0 \quad \sum_{i=1}^n (Y_i - (B_0 + \dots + B_k x_k))x_k = 0$$

Solving the MLE

$$\sum_{i=1}^n Y_i = \alpha B_0 + B_1 \sum_{i=1}^n x_{i1} + B_2 \sum_{i=1}^n x_{i2} + \dots + B_k \sum_{i=1}^n x_{ik} \quad (9.10.1)$$

$$\sum_{i=1}^n x_{ij} Y_i = B_0 \sum_{i=1}^n x_{i1} + B_1 \sum_{i=1}^n x_{i1}^2 + B_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + B_k \sum_{i=1}^n x_{i1} x_{ik}$$

$$\sum_{i=2}^k x_{ij} Y_i = B_0 \sum_{i=1}^n x_{i2} + B_1 \sum_{i=1}^n x_{i2} x_{i1} + B_2 \sum_{i=1}^n x_{i2} x_{i2} + \dots + B_k \sum_{i=1}^n x_{i2}^2$$

Matrix notation for k-dimensional covariates.

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underline{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \xrightarrow{(1 \times n)} \underline{X} \in n \times (k+1)$$
$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \xrightarrow{(k+1) \times 1}$$

then \underline{Y} is an $n \times 1$, \underline{X} an $n \times p$, $\underline{\beta}$ a $p \times 1$, and \underline{e} an $n \times 1$ matrix where $p = k + 1$.

The regression equation on this data becomes:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$$

$$\underline{\underline{X'X}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1k} & X_{2k} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}$$

X^t = transpose of X

$$= \begin{bmatrix} n & \sum X_{11} & \sum X_{12} & \cdots & \sum X_{1k} \\ \sum x_{11} & \sum x_{11}^2 & \sum x_{11}x_{12} & \cdots & \sum x_{11}x_{1k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{n1} & \sum x_{n1}x_{11} & \sum x_{n1}x_{12} & \cdots & \sum x_{n1}^2 \end{bmatrix}$$

(i, i)

$$\sum_{i=1}^n x_{ii} x_{ii}$$

Covariance matrix

if $\bar{Y} = 0$

and

$$\underline{\underline{X'Y}} = \begin{bmatrix} \sum Y_i \\ \sum x_{11}Y_1 \\ \vdots \\ \sum x_{kk}Y_k \end{bmatrix}$$

Solving the MLE

$$\sum_{i=1}^n Y_i = \beta_0 S_0 + \beta_1 \sum_{i=1}^n x_{i1} + \beta_2 \sum_{i=1}^n x_{i2} + \dots + \beta_k \sum_{i=1}^n x_{ik} \quad (9.10.1)$$

$$\sum_{i=1}^n x_{ij} Y_i = \beta_0 \sum_{i=1}^n x_{i1} + \beta_1 \sum_{i=1}^n x_{i1}^2 + \beta_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + \beta_k \sum_{i=1}^n x_{i1} x_{ik}$$

⋮

$$\sum_{i=1}^n x_{ik} Y_i = \beta_0 \sum_{i=1}^n x_{i0} + \beta_1 \sum_{i=1}^n x_{i0} x_{i1} + \beta_2 \sum_{i=1}^n x_{i0} x_{i2} + \dots + \beta_k \sum_{i=1}^n x_{i0}^2$$

⋮

$$\underline{\underline{X'XB = X'Y}}$$



$$\underline{\underline{B = (X'X)^{-1} X'Y}}$$

where X' is the transpose of X .

Example

Get least square estimate on this data with last column as y.

Age (years)	Diameter (mm)	Root length (mm)	Specific Gravity	Diameter at Breast Height (inches)
1	10	1.5	.95	10.5
2	15	2.0	1.0	10.8
3	20	2.5	1.0	11.0
4	25	3.0	1.0	11.2
5	30	3.5	1.0	11.5
6	35	4.0	1.0	11.8
7	40	4.5	1.0	12.0
8	45	5.0	1.0	12.2
9	50	5.5	1.0	12.5
10	55	6.0	1.0	12.8
11	60	6.5	1.0	13.0
12	65	7.0	1.0	13.2
13	70	7.5	1.0	13.5
14	75	8.0	1.0	13.8
15	80	8.5	1.0	14.0

$$k = 4$$

$$y = 11.54873 + 0.05728x_1 + 0.08712x_2 + 7.33231x_3$$

Evaluating goodness of a fit: The coefficient of determination

- Measure amount of variation in the data:

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- Measure the amount of variation in the residual after a model is fit

$$SS_R = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \hat{Y}_i = Bx_i$$

- The coefficient of determination

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}}$$
$$= 1 - \frac{SS_R}{S_{YY}}$$

$\left| \begin{array}{l} \text{if } \forall i \quad \hat{Y}_i = \bar{Y} \text{ what is } R^2 = 0 \\ \text{if } \forall i \quad \hat{Y}_i = Y_i \quad R^2 = 1 \end{array} \right.$
 $0 \leq R^2 \leq 1$

Demo

[https://colab.research.google.com/github/rafiag/DTI2020/blob/main/02a_Multi_Linear_Regression_\(EN\).ipynb](https://colab.research.google.com/github/rafiag/DTI2020/blob/main/02a_Multi_Linear_Regression_(EN).ipynb)

Lecture 23-kernelRegression.pdf

Non-parametric regression

Motivation

- Linear regression fits a linear line, which might be a poor fit for general datasets.
- Need a powerful estimator of $E(Y|X)$ without making any assumption about the functional form.
- $E(Y|x_1, \dots, x_k) = m(x)$ where $x = (x_1, \dots, x_k)$
- The function $m(x)$ we will derive under the assumption that $f(X,Y)$ and $f(X)$ are both estimated using kernels.

$$m(x) = E(Y|x) = \int_y y \underbrace{f(y|x)}_{y} dy$$

x is also a random variable

$$= \int_y y \frac{f(x,y)}{f(x)} dy$$

$$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

$\hat{f}(x|y)$ → estimate using KDE on D

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$$

$$\begin{aligned} \hat{f}(x,y) &= K_{xy}\left[(x_i, y_i), (x, y)\right] = K\left(\frac{x_i - x}{h}, K\left(\frac{y_i - y}{h}\right)\right) \\ &= K\left(\frac{x_i - x}{h}\right) K\left(\frac{y_i - y}{h}\right) \end{aligned}$$

$$\begin{aligned}
 E(Y|x) &= \int y \frac{\sum_{i=1}^n k\left(\frac{x_i-x}{h}\right) k\left(\frac{y_i-y}{h}\right) + \hat{f}(x)y}{\sum_{i=1}^n k\left(\frac{x_i-x}{h}\right)} dy \\
 &= \frac{\sum_{i=1}^n k\left(\frac{x_i-x}{h}\right) Y_i}{\sum_{i=1}^n k\left(\frac{x_i-x}{h}\right)} = \hat{Y} \\
 m(x) &= \frac{\int y k\left(\frac{y_i-y}{h}\right) dy}{\sum_{i=1}^n k\left(\frac{x_i-x}{h}\right)} \quad \text{change of variable} \\
 &= Y_i \\
 \text{Test } (\underline{x}) \rightarrow \hat{Y} &= \sum_{i=1}^n w_i(x) Y_i \quad w_i(x) = \frac{k\left(\frac{x_i-x}{h}\right)}{\sum_{j=1}^n k\left(\frac{x_j-x}{h}\right)}
 \end{aligned}$$

Non-parametric estimate of $E[Y|x]$

Starting with the definition of conditional expectation.

$$E(Y|X=x) = \int y f(y|x) dy = \int y \frac{f(x,y)}{f(x)} dy$$

we estimate the joint distribution $f(x,y)$ and $f(x)$ using kernel density estimation with a kernel K :

$$\hat{f}(x,y) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) K_h(y - y_i),$$

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i),$$

We get:

$$\begin{aligned} E(Y|X=x) &= \int y \frac{\hat{f}(x,y)}{\hat{f}(x)} dy \\ &= \int y \frac{\sum_{i=1}^n K_h(x - x_i) K_h(y - y_i)}{\sum_{i=1}^n K_h(x - x_i)} dy \\ &= \frac{\sum_{i=1}^n K_h(x - x_i) / \int y K_h(y - y_i) dy}{\sum_{i=1}^n K_h(x - x_i)} \\ &= \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)}, \end{aligned}$$

which is the Nadaraya-Watson estimator.

Nadaraya–Watson kernel regression

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n K_h(x - x_i)y_i}{\sum_{i=1}^n K_h(x - x_i)}$$

Demo: <https://colab.research.google.com/github/tufts-ml-courses/cs135-23f-assignments/blob/main/labs/day20-KernelRegression.ipynb>

Choosing bin-width is again a problem

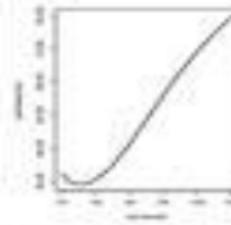
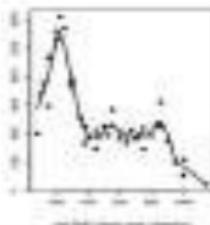
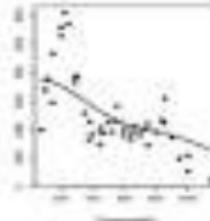
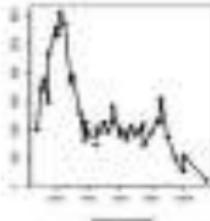
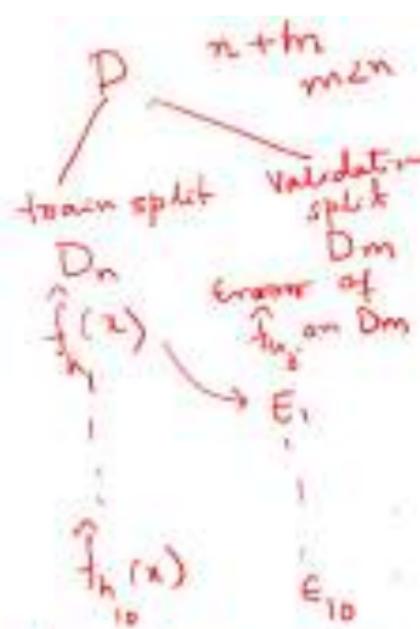


FIGURE 3.6: Frequency analysis of the 1000 values. The left is a recommended bin width is too narrow, and the right is based on overestimation. The left panel shows the estimated value versus the bandwidth of the histogram. The axes are from 0 to 1000000. Source: and Figueira



choose h with minimum error

Multidimensional extension.

$$k\left(\frac{x_i - x}{n}\right) = k\left(\frac{\|x_i - x\|^2}{n}\right)$$

$$\begin{aligned}x &\in \mathbb{R}^k \\x &= [x_1, \dots, x_k]\end{aligned}$$

$$\begin{aligned}L_p &= \|x_i - x\|_p \\&= \left(\sum_{j=1}^k |x_{ij} - x_j|^p \right)^{1/p}\end{aligned}$$

Summary of regression

- Linear regression (1-D data)
 - MLE estimates of slope and intercept
 - Unbiased chi-squared distribution based estimate of σ^2
- Distribution of parameters
- Linear regression (Arbitrary k)
 - Just the derivation of MLE estimate
- Kernel regression
 - Just the final estimate.

Lecture 24-TimeSeriesAnalysis1.pdf

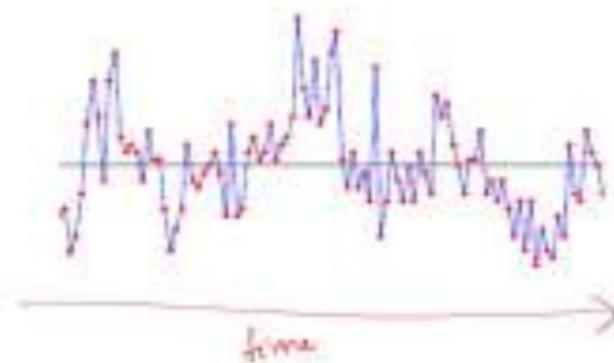
Time Series Analysis

Reading material

<https://online.stat.psu.edu/stat510/lesson/1>

What is a time-series

- Sequence of values recorded at regular time intervals
 - Time interval: E.g. Weekly, monthly, daily, hourly, annually, etc.
 - Values recorded:
 - Scalar: single value like sales
 - Vector of values



Motivation

- Daily traffic on individual webpages from different regions in Wikipedia
- Hourly load on various servers of different services in a Data center
- Monthly demand for products from different regions in Flipkart
- Stock price of various companies
- Rice production in Maharashtra each year
- Consumer price index of various food items.

Objective of time series analysis

Identify characteristics of the time-series

- Provide a model of the data (e.g. test scientific hypothesis)
- Forecasting
 - Predict future values as a function of past values,



- Finding outliers and filling in missing values
- Provide a compact description of the data (data compression)

Important characteristics of time-series

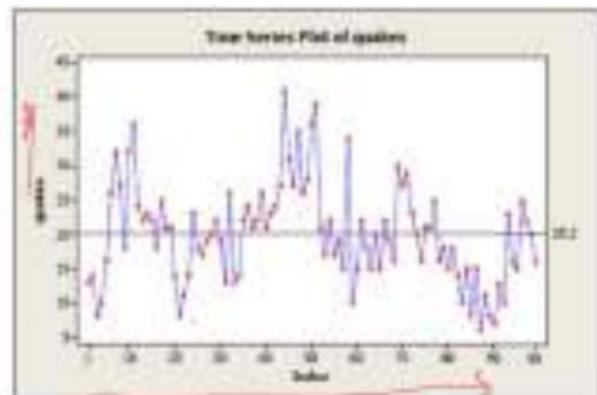
- Is there a **trend**, meaning that, on average, the measurements tend to increase (or decrease) over time?
- Is there **seasonality**, meaning that there is a regularly repeating pattern of highs and lows related to calendar time such as seasons, quarters, months, days of the week, and so on?
- Are there **outliers**? In regression, outliers are far away from your line. With time series data, your outliers are far away from your other data.
- Is there a **cycle**: data rises and falls but without a fixed frequency.
- Is there **constant variance** over time, or is the variance non-constant?
- Are there any **abrupt changes** to either the level of the series or the variance?

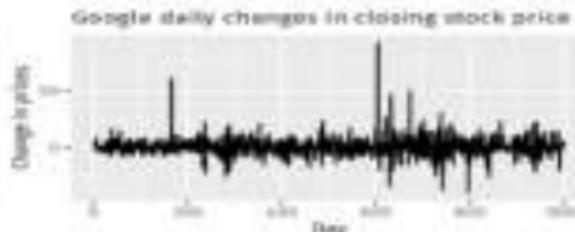
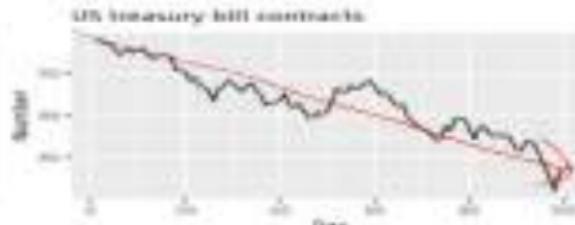
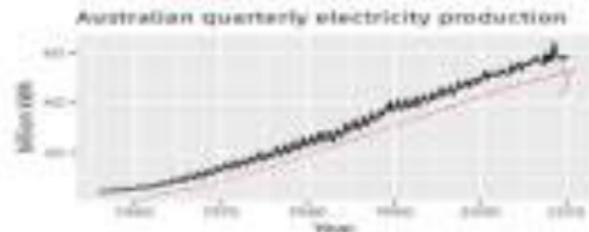
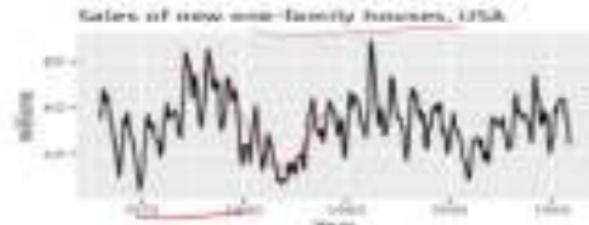
Example 1-1 (from book)

Annual number of earthquakes in the world
with seismic magnitude over 7.0, for 99
consecutive years

Characteristics:

- No consistent trend. Values on both sides of mean along time
- No seasonality
- No outliers





- The monthly housing sales (top left) show strong seasonality within each year, as well as some strong cyclic behaviour with a period of about 6–10 years. There is no apparent trend in the data over this period.
- The US treasury bill contracts (top right) show results from the Chicago market for six-month期的 trading days in 1991. Here there is no seasonality, but an obvious downward trend. Possibly, if we had a much longer series, we would see that this downward trend is actually part of a long cycle, but when viewed over only 310 days it appears to be a trend.
- The Australian quarterly electricity production (bottom left) shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behaviour here.
- The daily change in the Google closing stock price (bottom right) has no trend, seasonality or cyclic behaviour. There are random fluctuations which do not appear to be very predictable, and no strong patterns that would help with developing a forecasting model.

Example 1-2

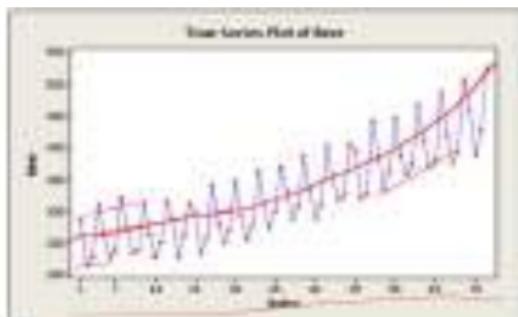
Quarterly production of beer in Australia

For the past 18 years.

- Length = $4 \times 18 = 72$

Characteristics

- There is an upward trend, possibly a curved one.
- There is seasonality – a regularly repeating pattern of highs and lows related to quarters of the year.
- There are no obvious outliers.
- There might be increasing variation as we move across time, although that's uncertain.



Fitting simple regression models for quantifying pattern in time-series

Suppose that the observed series is x_t , for $t = 1, 2, \dots, n$.



- For a linear trend, use t (the time index) as a predictor variable in a regression.
- For a quadratic trend, we might consider using both t and t^2 .
- For quarterly data, with possible seasonal (quarterly) effects, we can define indicator variables such as $S_j = 1$ if the observation is in quarter j of a year and 0 otherwise. There are 4 such indicators.

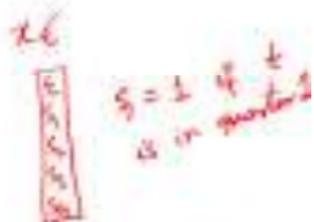
Let $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. A model with additive components for linear trend and seasonal (quarterly) effects might be written:

$$x_t = \beta_1 t + \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \epsilon_t$$

To add a quadratic trend, which may be the case in our example, the model is

$$x_t = \beta_1 t + \beta_2 t^2 + \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \epsilon_t$$

parameters

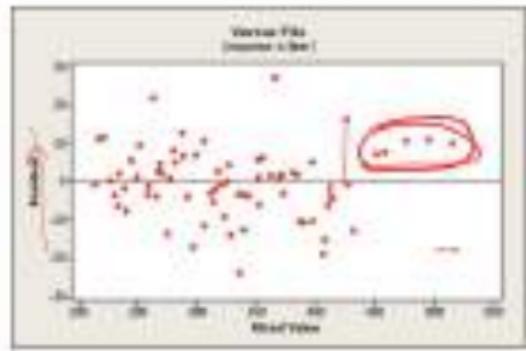


Fitting the dataset of 72 values using least square regression

Predictor	Coeff	SE Coef
No constant		
Time	0.5881 β_1	0.2193
tsqrd	0.031214 β_2	0.002911
quarter_1	251.430 α_1	3.937
quarter_2	212.165 α_2	3.968
quarter_3	228.415 α_3	3.994
quarter_4	310.880 α_4	4.018

Residual analysis

- Ideal residuals: mean 0, normally distributed.
- For time-series: correlation between residues separated by a fixed time-span should be zero.



Limitation of regression models

- Does not account for strong temporal correlation among values.
- Predicts each value independent of its neighbors.
- Need a model that can directly exploits correlations within a series.

Sample Autocorrelation Function (ACF)

- Correlation between a series value x_t and lagged values for different values of lags h .
- Lag h auto-correlation function: correlation between x_t, x_{t-h} for all t .

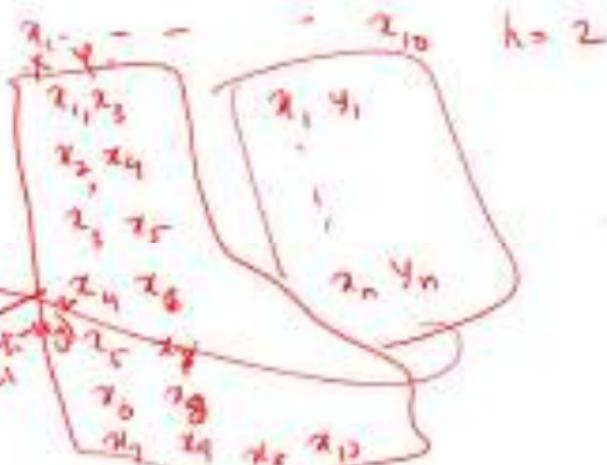
→ Covariance(x_t, x_{t-h})

Std.Dev.(x_t) Std.Dev.(x_{t-h})

Correlation (X, Y)

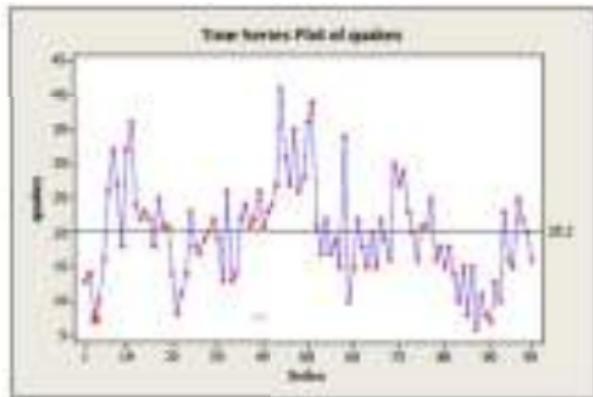
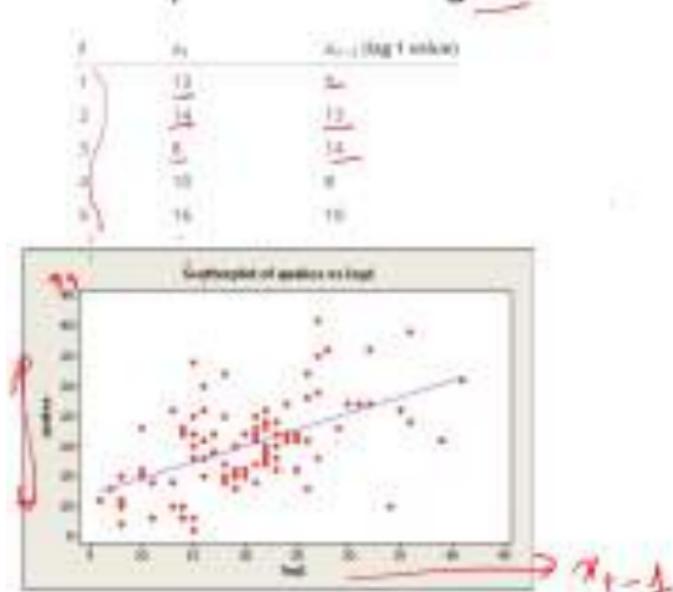
= Covariance (X, Y) / Std. Dev. (X)

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



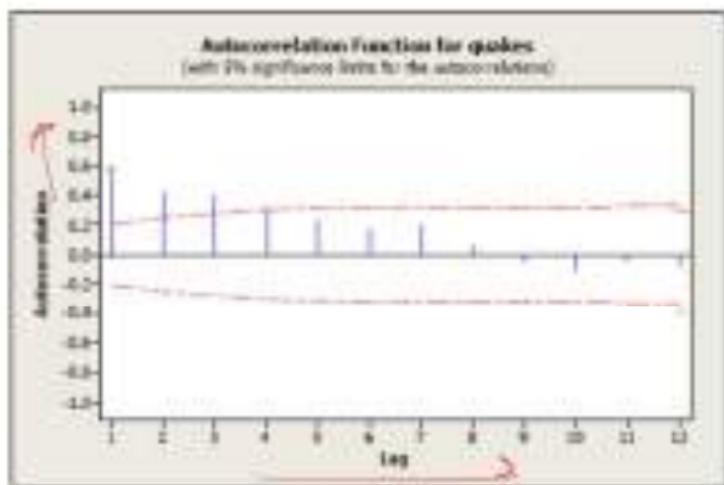
Example 1-1 (from book)

- Earthquake data with lag $h=1$



Example ACR for lag > 1

Lag	ACF
1.	0.541733
2.	0.418884
3.	0.397955
4.	0.324047
5.	0.237164
6.	0.171794
7.	0.190228
8.	0.061202
9.	-0.048505



Stationary series

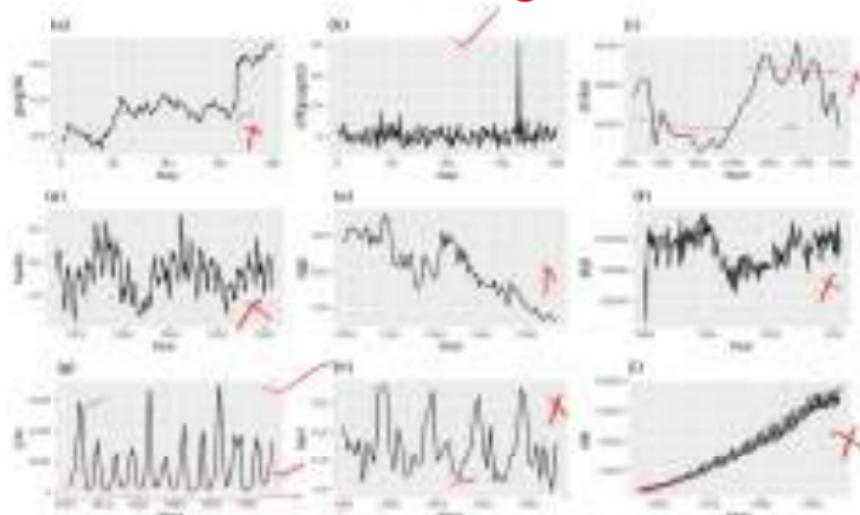
A series x_t is said to be **(weakly) stationary** if it satisfies the following properties:

- The mean $E(x_t)$ is the same for all t .
- The variance of x_t is the same for all t .
- The covariance (and also correlation) between x_t and x_{t-h} is the same for all t at each lag $h = 1, 2, 3$, etc.

Autocorrelation function for stationary series:

$$ACF(h) = \frac{\text{Covariance}(x_t, x_{t-h})}{\text{Std.Dev.}(x_t)\text{Std.Dev.}(x_{t-h})} = \frac{\text{Covariance}(x_t, x_{t-h})}{\text{Variance}(x_t)}$$

Which of the following series are stationary?



seasonality rules out series (d), (h) and (i). Trends and changing levels rules out series (a), (c), (e), (f) and (i). Increasing variance also rules out (i). That leaves only (b) and (g) as stationary series.

Figure 6.8 Which of these series are stationary? (a) Google stock price for 200 consecutive days; (b) Daily change in the Google stock price for 200 consecutive days; (c) Annual number of visitors to the Whitechapel gallery of fine and decorative arts; (d) 3rd Annual price of a kilogramme of beef (Exmouth market); (e) Monthly total of pigs slaughtered in Victoria, Australia; (f) Annual total of pigs totalled in the Parry Sound district of North-West Ontario, ON, March pig Australian beer production; (g) Monthly Australian beer production; (h) Monthly Australian electricity production.

<https://otexts.com/fpp2/stationarity.html>

Models for time-series

- Auto-regressive models AR(p):
- Moving average models MA
- ARIMA models (Combines the above two)
- SARIMA: Generalization of ARIMA to handle seasonality.

Auto-regressive models

- AR(p) model: The value of x at time t is a linear function of the value of x at time $t-1, t-2, \dots, t-p$.

$$AR(p): x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- AR(1)

$$x_t = \delta + \phi_1 x_{t-1} + w_t$$

parameter word

$$x_t = x_{t-1} + w_t$$

$\phi_1 = 1$

- $w_t \stackrel{iid}{\sim} N(0, \sigma_w^2)$, meaning that the errors are independently distributed with a normal distribution that has mean 0 and constant variance.
- Properties of the errors w_t are independent of x_t .
- The series x_1, x_2, \dots is (weakly) stationary. A requirement for a stationary AR(1) is that $|\phi_1| < 1$. We'll see why below.

Lecture 25-TimeSeriesAnalysis2.pdf

Models for time-series

- Auto-regressive models AR(p):
- Moving average models MA
- ARIMA models (Combines the above two)
- SARIMA: Generalization of ARIMA to handle seasonality.

Auto-regressive models

- AR(p) model: The value of x at time t is a linear function of the value of x at time $t-1, t-2, \dots, t-p$.

$$AR(p): x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- AR(1)

$$x_t = \delta + \phi_1 x_{t-1} + w_t$$

parameter word

$$x_t = x_{t-1} + w_t$$

$\phi_1 = 1$

- $w_t \stackrel{iid}{\sim} N(0, \sigma_w^2)$, meaning that the errors are independently distributed with a normal distribution that has mean 0 and constant variance.
- Properties of the errors w_t are independent of x_t .
- The series x_1, x_2, \dots is (weakly) stationary. A requirement for a stationary AR(1) is that $|\phi_1| < 1$. We'll see why below.

Finding the best values of parameters: $\delta, \phi_1, \dots, \phi_p$

- Standard least square fit. $D: \{x_1, x_2, \dots, x_n\}$
- Each t defines a training instance

Our goal is to find $\phi_1, \phi_2, \dots, \phi_p, n, \sigma_w^2$ s.t.
we have a good fit on each x_t

$$x_t = \underline{\phi_1} x_{t-1} + \underline{\phi_2} x_{t-2} + \dots + \underline{\phi_p} x_{t-p} + w_t$$
$$w_t \sim N(0, \sigma_w^2)$$

Example 37.1

- The number of disk access for 50 database queries were measured to be: 73, 67, 83, 53, 78, 88, 57, 1, 29, 14, 80, 77, 19, 14, 41, 55, 74, 98, 84, 88, 78, 15, 66, 99, 80, 75, 124, 103, 57, 49, 70, 112, 107, 123, 79, 92, 89, 116, 71, 68, 59, 84, 39, 33, 71, 83, 77, 37, 27, 30.

- For this data: $\sum_{t=1}^{50} x_t = 3313 \quad \sum_{t=1}^{50} x_{t-1} = 3356$

$$\sum_{t=2}^{50} x_t x_{t-1} = 248147 \quad \sum_{t=2}^{50} x_{t-1}^2 = 272102 \quad n = 49$$

$$\begin{aligned}\text{Cov}_x &= \frac{\sum x_t \sum x_{t-1}^2 - \sum x_{t-1} \sum x_t x_{t-1}}{n \sum x_{t-1}^2 - (\sum x_{t-1})^2} \\ &= \frac{3313 \times 272102 - 3356 \times 248147}{49 \times 272102 - 3356^2} = 33.181\end{aligned}$$

Example 37.1 (Cont)

$$\begin{aligned} \rho_1 &= \frac{n \sum x_t x_{t-1} - \sum x_t \sum x_{t-1}}{n \sum x_{t-1}^2 - (\sum x_{t-1})^2} \\ &= \frac{49 \times 248147 - 3313 \times 3356}{49 \times 272102 - 3356^2} = 0.503 \end{aligned}$$

- The AR(1) model for the series is:

$$x_t = 33.181 + 0.503 x_{t-1} + \epsilon_t$$

- The predicted value of x_2 given x_1 is:

$$\hat{x}_2 = a_0 + a_1 x_1 = 33.181 + 0.503 \times 73 = 69.880$$

- The actual observed value of x_2 is 67. Therefore, the prediction error is:

$$\epsilon_2 = x_2 - \hat{x}_2 = 67 - 69.880 = -2.880$$

- Sum of squared errors SSE = 32995.57

Properties of a time-series following AR(1) model

- The (theoretical) mean of x_t is

$$\underline{E(x_t)} = \mu = \frac{\cancel{\alpha} + \eta}{1 - \phi_1}$$

- The variance of x_t is

$$\text{Var}(x_t) = \frac{\sigma_u^2}{1 - \phi_1^2}$$

- The correlation between observations h time periods apart is

$$\underline{\rho_h} = \underline{\phi_1^h}$$

Proofs.

- Easy proofs [here](#)

$$x_t = \eta + \varphi_1 x_{t-1} + w_t$$

$$\begin{aligned} E[x_t] &= E(\eta) + E(\varphi_1 x_{t-1}) + E(w_t) \\ &= \eta + \varphi E(x_{t-1}) + 0 \end{aligned}$$

Assume series is stationary $E[x_t] = E(x_{t-1})$

$$E(x_t) = \frac{\eta}{1 - \varphi_1}$$

$$V(x_t) = E$$

.. .

$A \subset F(1)$

$$\frac{E[(x_t - \mu_x)(x_{t-1} - \mu_x)]}{\text{var}(x_t)}$$

$$= E[(\varphi x_{t-1} + \eta + w_t - \mu_x)(x_{t-1} - \mu_x)]$$

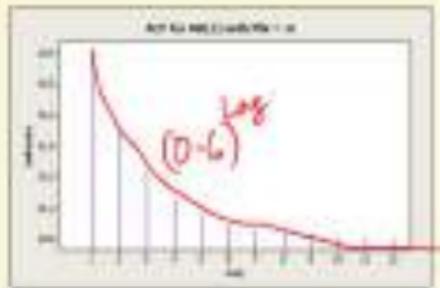
$$= \varphi E[x_{t-1}^2] + n E[x_{t-1}] + E[x_{t-1} w_t]$$

Shape of ACF of a series following AR(1) model

Following is the ACF of an AR(1) with $\phi_1 = 0.6$, for the first 12 lags:

Note!

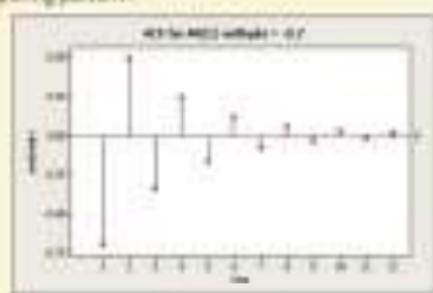
The tapering pattern.



The ACF of an AR(1) with $\phi_1 = -0.7$ follows:

Note!

The alternating and tapering pattern.



Choosing the p for which AR(p) provides a good fit

Partial Auto Correlation Function (PACF), also called conditional ACF.

- Auto-correlation after adjusting for the intervening values
- Conditional correlation: correlation between x_t and x_{t-h} under known values of x -s in-between them.
- ACF might show that x_t and x_{t-h} are correlated but that might be because they are both correlated with the values in-between. PACF corrects for that.
 x_t is correlated with $\{x_{t-1}, x_{t-2}\}$ & x_{t-1} is correlated with $\{x_{t-2}, x_{t-3}\}$
 x_{t-2} is correlated with $\{x_{t-3}, x_{t-4}\}$
- The 1st order partial autocorrelation will be defined to equal the 1st order autocorrelation,
- The 2nd order (lag) partial autocorrelation is

$$\frac{\text{Covariance}(x_t, x_{t-2}|x_{t-1})}{\sqrt{\text{Variance}(x_t|x_{t-1})\text{Variance}(x_{t-2}|x_{t-1})}}$$

Computing Partial Auto Correlation Function (PACF)

- To compute PACF between x_t and x_{t-h} fit an AR(h) model. The coefficient ϕ_h is the measure of PACF

$$x_t = \eta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_h x_{t-h} + w_t$$

$$\phi_h \equiv \text{PACF}(x_t, x_{t-h})$$

But $\phi_2 \neq \text{PACF}(x_t, x_{t-2})$ for $h > 2$

ARMA models

- Extending auto-regressive models with smoother noise.

In AR model for each t , we associate an independent noise w_t

Rice production in Maharashtra
AR(1)

$$x_t = \eta + \phi_1 x_{t-1} + w_t + \theta_1 w_{t-1}$$
$$x_t = \eta + \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \boxed{\theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_p w_{t-p}} + w_t$$

MA terms

Need smoother handling of noise.

A moving average (MA) model provides that.

- ARMA models: AR models + MA models

Moving average models (MA models)

- A value x_t in a time-series sometimes cannot be explained just in terms of its past values.
- External (unknown) variables might be influencing the values
 - Example: Total wheat export of India in 2023 can be determined by wheat export in 2022, but also other external factors like weather patterns, war, exchange rates, etc.
- External variables are also time-varying → errors at each position cannot be independent.
- Moving average models capture dependency on such external unknowns.

Properties of a series following MA(1) model $p=0, q=1$

$$x_t = \eta + \theta_1 w_{t-1} + w_t$$

- Mean is $E(x_t) = \mu$ $E(x_t) =$
- Variance is $Var(x_t) = \sigma_w^2(1 + \theta_1^2)$
- Autocorrelation function (ACF) is: $E(x_t x_{t-h})$

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}, \text{ and } \rho_h = 0 \text{ for } h \geq 2$$

Papers here: <http://online.stat.psu.edu/stat510/lesson/2/2.18/paragraph-264>

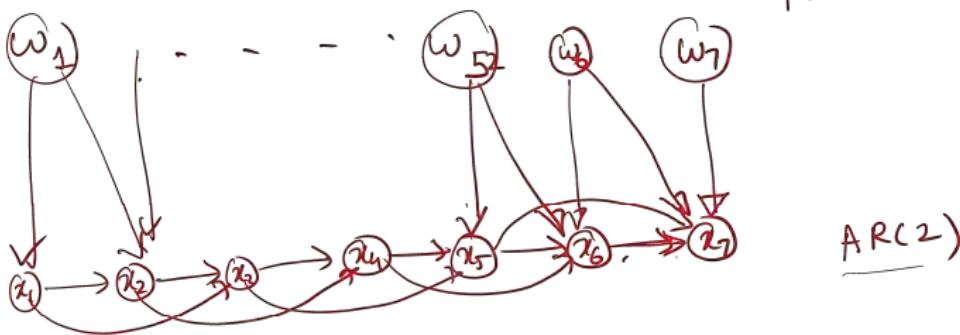
Pictorial representation of dependency.

$$E(x_t | x_{t-1}) = E((n + \theta_1 w_{t-1} + w_t) x_{t-1})$$

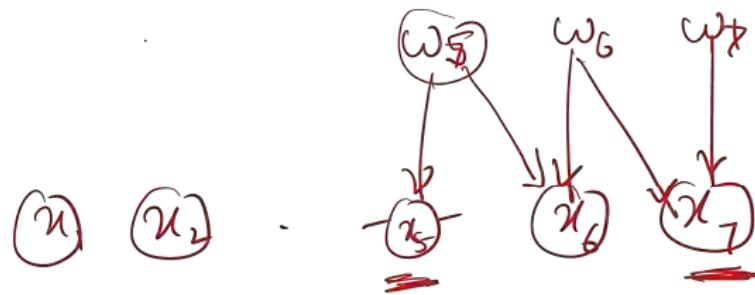
$$= n E(x_{t-1}) + \theta_1 E(w_{t-1} \cdot x_{t-1}) + E(w_t x_{t-1})$$

=

MA(1)



$$p = 0, \quad q = 1$$



ARMA (p,q) model

Each x_t depends on p previous x-values, and q-previous error values

$$x_t = \eta + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

Estimating all the parameters of this model is not as straightforward as least-square regression since the w_t values are not observed (Not covered)

Lecture 26-TimeSeriesAnalysis3.pdf

ARMA models

- Extending auto-regressive models with smoother noise.

In AR model for each t , we associate an independent noise w_t

Rice production in Maharashtra
AR(1)

$$x_t = \eta + \varphi_1 x_{t-1} + w_t + \theta_1 w_{t-1}$$

$$x_t = \eta + \varphi_1 x_{t-1} + \dots + \varphi_p x_{t-p} + \boxed{\theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_p w_{t-p}} + w_t$$

Need smoother handling of noise.

A moving average (MA) model provides that.

- ARMA models: AR models + MA models

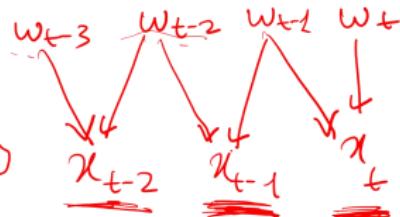
Moving average models (MA models)

- A value x_t in a time-series sometimes cannot be explained just in terms of its past values.
- External (unknown) variables might be influencing the values
 - Example: Total wheat export of India in 2023 can be determined by wheat export in 2022, but also other external factors like weather patterns, war, exchange rates, etc.
- External variables are also time-varying → errors at each position cannot be independent.
- Moving average models capture dependency on such external unknowns.

Properties of a series following MA(1) model $p=0, q=1$

$$x_t = \eta + \theta_1 w_{t-1} + w_t$$
$$x_{t-1} = \eta + \theta_1 w_{t-2} + w_{t-1}$$
$$x_{t-2} = \eta - \theta_1 w_{t-3} + w_{t-2}$$

- Mean is $E(x_t) = \mu$
- Variance is $Var(x_t) = \sigma_w^2(1 + \theta_1^2)$
- Autocorrelation function (ACF) is:



$$E(x_t x_{t-1})$$

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}, \text{ and } \rho_h = 0 \text{ for } h \geq 2$$

Proofs here: <https://online.stat.psu.edu/stat510/lesson/2/2.1#paragraph--264>

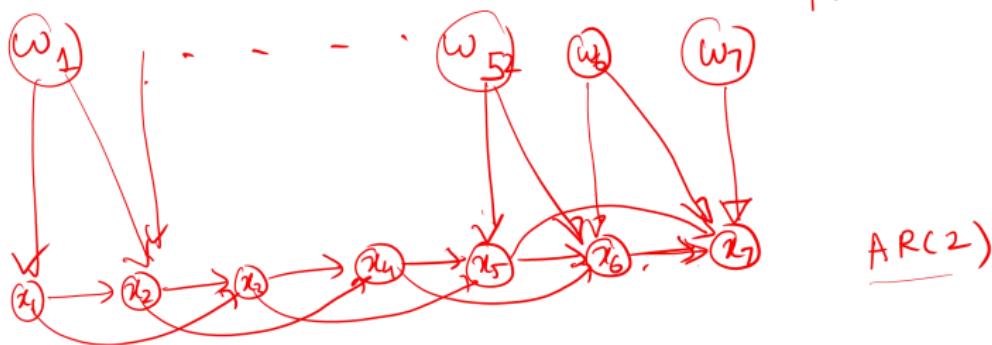
Pictorial representation of dependency.

$$E(x_t x_{t-1}) = E((n + \theta_1 w_{t-1} + w_t) x_{t-1})$$

$$= n E(x_{t-1}) + \theta_1 E(w_{t-1} \cdot x_{t-1}) + E(w_t x_{t-1})$$

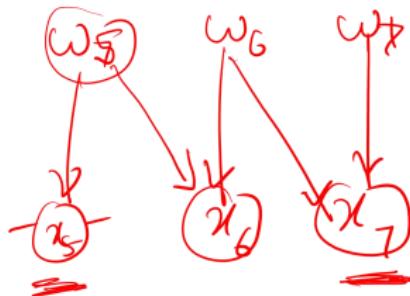
=

MAC(1)



ARC(2)

$$p = 0, \quad q = 1$$



Original:

$$\rightarrow x_t = \theta_1 \underline{w_{t-1}} + w_t + \eta \quad : \text{To determine PACF}(x_t, x_{t-2})$$

$$\leftarrow x_t = \underline{\varphi_1 x_{t-1}} + \underline{\varphi_2 x_{t-2}} + \tilde{\eta}$$

$= \varphi_2$

ARMA (p,q) model

Each x_t depends on p previous x-values, and q-previous error values

$$x_t = \eta + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

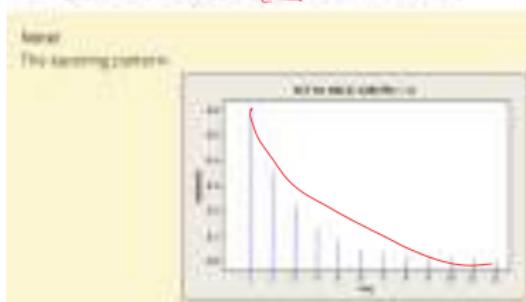
Estimating all the parameters of this model is not as straightforward as least-square regression since the w_t values are not observed (Not covered)

Comparing AR(1) and MA(1) on ACF and PACF

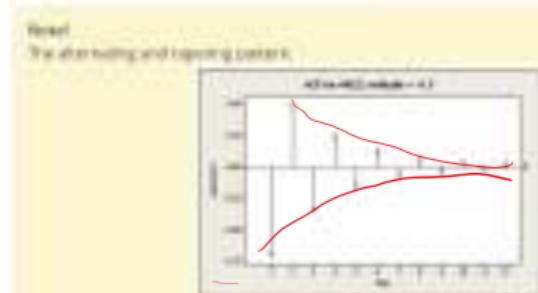
- ACF=plain correlation
- PACF(x_t, x_{t-2})=conditional correlation or what extra contribution you get from x_{t-2} after you x_{t-1}

Shape of ACF and PACF of a series following AR(1) model

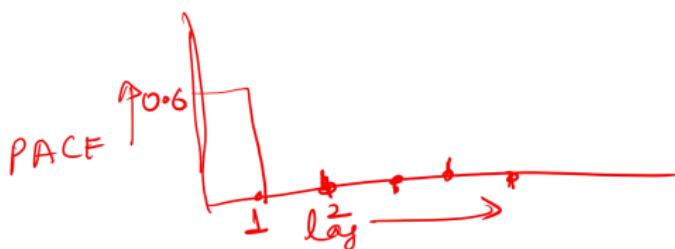
Following is the ACF of an AR(1) model with $\rho = 0.6$, for the first 12 lags:



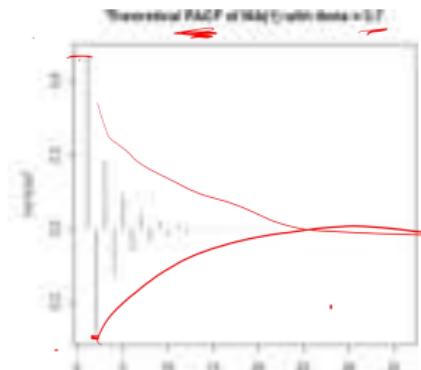
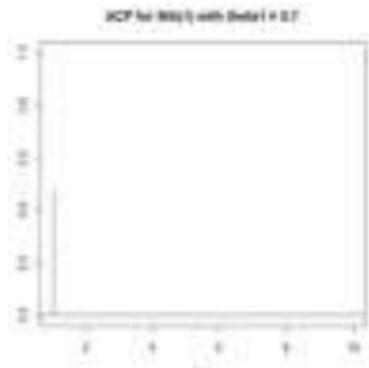
The ACF of an AR(1) model with $\rho = -0.6$ is as follows:



$$\text{PACF}(x_t, x_{t-1}) = \text{ACF}(x_t, x_{t-1}) ; \quad \text{PACF}(x_t, x_{t-2}) = 0$$



Shape of ACF and PACF of a series following MA(1) model



$$\begin{aligned}
 x_t &= \theta_1 w_{t-1} + w_t & ; \quad w_t = x_t - \theta_1 w_{t-1} \\
 &= \theta_1 (x_{t-1} - \theta_1 w_{t-2}) + w_t \\
 \downarrow &= \theta_1 (x_{t-1} - \theta_1 (x_{t-2} - \theta_1 w_{t-3})) + w_t \\
 &= \left\{ \theta_1 x_{t-1} - \theta_1^2 x_{t-2} + \theta_1^3 w_{t-3} + w_t \right\}
 \end{aligned}$$

Choosing p,q

- Data may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced data show the following patterns:
 - the ACF is exponentially decaying or sinusoidal;
 - there is a significant spike at lag p in the PACF, but none beyond lag p.
- The data may follow an ARIMA(0,d,q) model if the ACF and PACF plots of the differenced data show the following patterns:
 - ||| • The PACF is exponentially decaying or sinusoidal;
 - There is a significant spike at lag q in the ACF, but none beyond lag q.

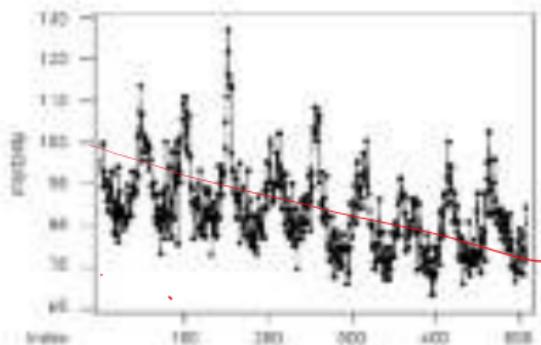
. [Tips given in this link](#)

Handling trend in time-series.

$$\underline{x}_t = \underline{\alpha}t + \varphi x_{t-1} + \theta, w_t + w_t$$

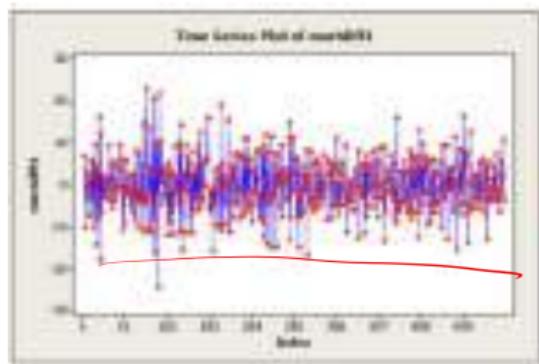
- If a time-series has a linear trend, then replace each value x_t with difference of consecutive x-values

$$\underline{y}_t = x_t - x_{t-1}$$



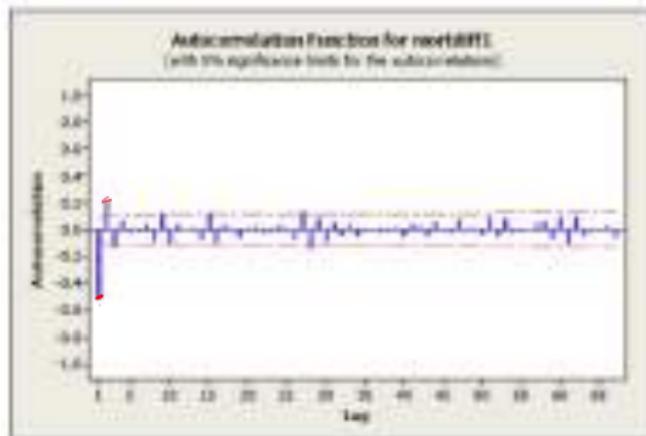
Daily cardiovascular mortality rate in Los Angeles County, 1970-1979.

Clear downward trend.



Plot of first differences

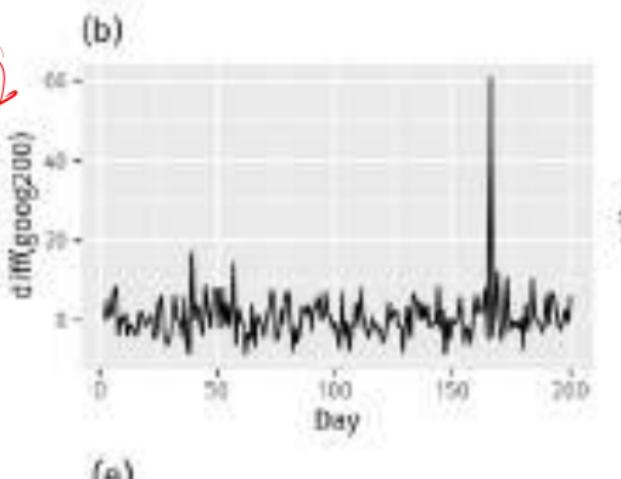
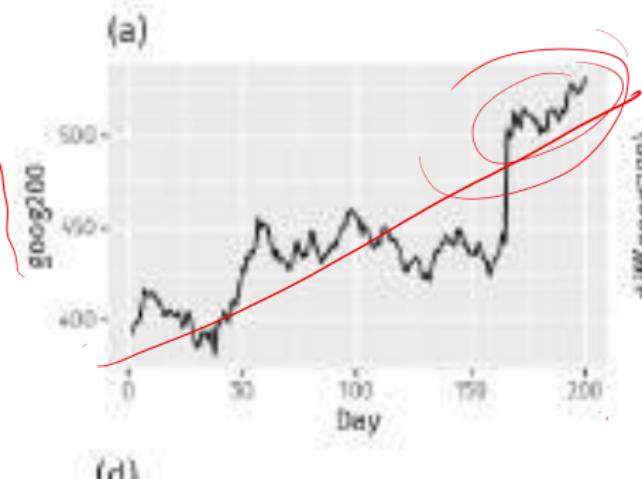
ACF of first differences.



Lag.	ACF
1.	-0.506029
2.	0.205100
3.	-0.126110
4.	0.062476
5.	0.015190

$$\hat{y}_t = -0.04627 - 0.50636y_{t-1}$$

Another example



ARIMA(p,d,q) models

- p is the order of the autoregressive part,
- d is the degree of first difference involved,
- q is the order of the moving average part.

Example: ARIMA(2,1,1) model

Incorporating seasonality.

- Seasonality in a time series is a regular pattern that repeats over S time periods.
 - Example: monthly seasonality repeats over S=12 (months of the year)
 - Example: quarter seasonality repeats over S=4 period
- Extending ARIMA to handle seasonality. One or more of the above might work
 - Introduce a AR term x_{t-S} in the model for every period S.
 - Introduce MA term w_{t-S} in the model for every period S.
 - Create seasonal differences $y_t = x_t - x_{t-S}$

Demo

- https://colab.research.google.com/drive/1Z4zNI_bVXoFQBsCHUtxBDCBno6yhXceB?usp=sharing#scrollTo=deWKK_D1mNlr

Lecture 27-MVA1.pdf

Multivariate Analysis

CS 215 Fall 2024

Multivariate data

- Data reduction or summarization
 - Studying data as simply as possible without sacrificing useful information
- Sorting and grouping
 - Clustering similar object together
- Investigating dependence among variables
 - Mutual independence, conditional independence etc. (already done)
- Prediction
 - Regression (already done)
- Hypothesis testing
 - Validate assumptions

Multivariate data organization.

Consequently, n measurements on p variables can be displayed as follows:

	Variable 1	Variable 2	...	Variable k	...	Variable p
Item 1:	x_{11}	x_{12}	...	x_{1k}	...	x_{1p}
Item 2:	x_{21}	x_{22}	...	x_{2k}	...	x_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Item j :	x_{j1}	x_{j2}	...	x_{jk}	...	x_{jp}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Item n :	x_{n1}	x_{n2}	...	x_{nk}	...	x_{np}

Or we can display these data as a rectangular array, called \mathbf{X} , of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

x_{ij} ↗ item-id ↘ variable-id

The array \mathbf{X} , then, contains the data consisting of all of the observations on all of the variables.

Multivariate descriptive statistics

Sample means

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^n x_{jk} \quad k = 1, 2, \dots, p$$

X

Sample variances
and covariances

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

$$s_{ik} = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Sample correlations

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}} \sqrt{s_{kk}}}$$

Unbiased sample covariance

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \right\}$$

Example

Example 1.1 (A data array) A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales): 42 52 48 58

Variable 2 (number of books): 4 5 4 3

Using the notation just introduced, we have

$$x_{11} = 42 \quad x_{21} = 52 \quad x_{31} = 48 \quad x_{41} = 58$$

$$x_{12} = 4 \quad x_{22} = 5 \quad x_{32} = 4 \quad x_{42} = 3$$

and the data array \mathbf{X} is

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

with four rows and two columns.

The sample means are

$$\bar{x}_1 = \frac{1}{4} \sum_{j=1}^4 x_{j1} = \frac{1}{4}(42 + 52 + 48 + 58) = 50$$

$$\bar{x}_2 = \frac{1}{4} \sum_{j=1}^4 x_{j2} = \frac{1}{4}(4 + 5 + 4 + 3) = 4$$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

The sample variances and covariances are

$$s_{11} = \frac{1}{3} \sum_{j=1}^4 (x_{j1} - \bar{x}_1)^2 \\ = \frac{1}{3}((42 - 50)^2 + (52 - 50)^2 + (48 - 50)^2 + (58 - 50)^2) = 34$$

$$s_{22} = \frac{1}{3} \sum_{j=1}^4 (x_{j2} - \bar{x}_2)^2 \\ = \frac{1}{3}((4 - 4)^2 + (5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2) = 2$$

$$s_{12} = \frac{1}{3} \sum_{j=1}^4 (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2) \\ = \frac{1}{3}((42 - 50)(4 - 4) + (52 - 50)(5 - 4) \\ + (48 - 50)(4 - 4) + (58 - 50)(3 - 4)) = -1.5$$

$$s_{11} = s_{22}$$

$$\mathbf{S}_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 2 \end{bmatrix}$$

The sample correlation is

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}} \sqrt{s_{22}}} = \frac{-1.5}{\sqrt{34} \sqrt{2}} = -.36$$

$$r_{11} = r_{22}$$

$$\mathbf{R} = \begin{bmatrix} 1 & -.36 \\ -.36 & 1 \end{bmatrix}$$

Descriptive statistics in matrix notation

$x_{\text{variable-id}, \text{item-id}}$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{y}'_1 \mathbf{1}}{n} \\ \frac{\mathbf{y}'_2 \mathbf{1}}{n} \\ \vdots \\ \frac{\mathbf{y}'_p \mathbf{1}}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_p \end{bmatrix} \mathbf{1}^T$$

$\mathbf{X}' \quad \mathbf{1} \quad \mathbf{X}_{n \times p}$

Diagram illustrating the calculation of descriptive statistics:

- The matrix \mathbf{X}' represents the transpose of the data matrix $\mathbf{X}_{n \times p}$.
- The vector $\mathbf{1}$ is a column vector of ones.
- The product $\mathbf{X}' \mathbf{1}$ results in a column vector of length n , where each element is the sum of all elements in a row of \mathbf{X} .
- The mean vector $\bar{\mathbf{x}}$ is obtained by dividing each element of the resulting vector by n .

Annotations in red:

- Red boxes highlight the first row of \mathbf{X}' ($x_{11}, x_{12}, \dots, x_{1n}$) and the first column of \mathbf{X}' ($x_{11}, x_{21}, \dots, x_{p1}$).
- Red arrows point from the labels y'_1, y'_2, \dots, y'_p to the corresponding elements in the vector $\frac{\mathbf{y}' \mathbf{1}}{n}$.
- A red bracket underlines the term $\frac{1}{n} \mathbf{y}' \mathbf{1}$.
- A red arrow points from the label $\mathbf{1}^T$ to the rightmost column of the matrix \mathbf{X}' .

Sample covariance

$$(n-1) \underset{(p \times p)}{\mathbf{S}} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}) (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$$

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \underbrace{[\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]}_{n \times 1}$$

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_1 & \cdots & x_{n1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & x_{22} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} - \bar{x}_p & x_{2p} - \bar{x}_p & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\mathbf{1} \bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} n \times p$$

Lecture 28-MVA2.pdf

Distance between two points

Two points: $P = [x_1, x_2, \dots, x_p]', Q = [y_1, y_2, \dots, y_p]'$

Euclidean distance between them:

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2}$$

- Gives equal importance to all dimensions.
- Does not make sense always.
 - Example mobiles phone statistics
 - x_1, x_2 refers to screen size in cms, and cost in rupees
 - A 2 point difference in screen size is more significant than a 2 rupee difference in cost.
 - Example: states with production of various grains
 - x_1, x_2 refers to production of rice Vs mustard in kilo tons.

	Screen size	Price	Weight	Memory
Stnd 1	5"	-	-	-
Stnd 2	-	-	-	-
Your.	-	-	-	-
70	-	-	-	-
	Rice	Wheat	Rai	Jeera
MA	-	-	-	-
UD	-	-	-	-
Bihar	-	-	-	-

Mahalonobis distance

Prasanta Chandra Mahalanobis

卷之三

ISSN 1063-271X • 10000 • 10000

Please indicate if this belief pertains to

Prasanta Chandra Mahalanobis (1893–1972) was an Indian statistician and biometrist. He is best remembered for the Mahalanobis distance, a statistical measure, and for being one of the members of the first Planning Commission of free India, the much-prominent statistician subsequently to India. He founded the Indian Statistical Institute, and contributed to the design of large-scale sample surveys.^{[1][2][3]} For his contributions, Mahalanobis has been considered the Father of statistics in India.^[4]

Early life



Final Manuscript

Matsaonka was born on 28 June 1922, in Calcutta, Bengal Presidency (now West Bengal). Matsaonka belonged to a poor and itinerant Brahmin family of tanned-gentry in Bihar. Chaitanya, Bengal Presidency (now in Bangladesh) [1821]. His grandfather Gourcharan [1813–1878] moved to Calcutta in 1854 and built up a business, starting a chemist shop in 1863. Gourcharan was influenced by Dwarakanath Tagore [1817–1895], leader of the Hindu Pra-advaitic party. Nalakumar Tagore, his father, was



10

29 June 2022
EBC-221-BWYU
EBC-221-BWYU

34

28-June-2023 page 701

300 words

University of California (2003
Maggie College, Cambridge
page 2)

Mahalanobis/Statistical distance between two points

Two points: $P = [x_1, x_2, \dots, x_p]', Q = [y_1, y_2, \dots, y_p]'$

- If points are more spread out in one dimension, then we expect distance between any two random points to be larger in that dimension.
- A new distance function: standard deviation scaled.

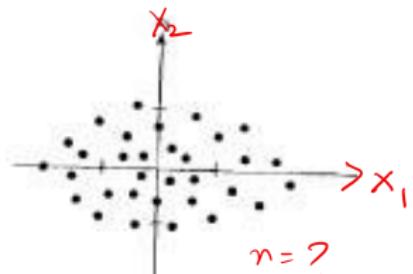
$$d(P, Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}^2} + \frac{(x_2 - y_2)^2}{s_{22}^2}}$$

variance

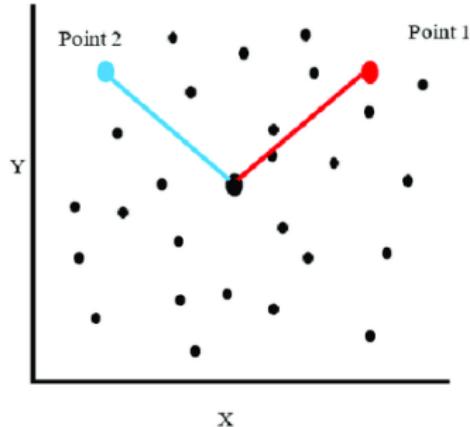
s₁₁ *s₂₂*

↗ —————— ↗

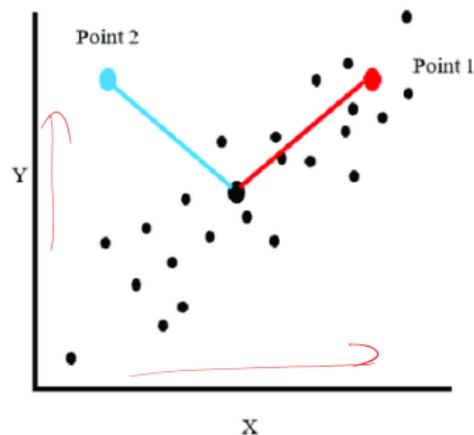
variance along dimension 1



X and Y are not correlated



X and Y are correlated



When X and Y are not correlated, the Euclidean distance from the Centroid can be useful to infer if a point is member of the distribution

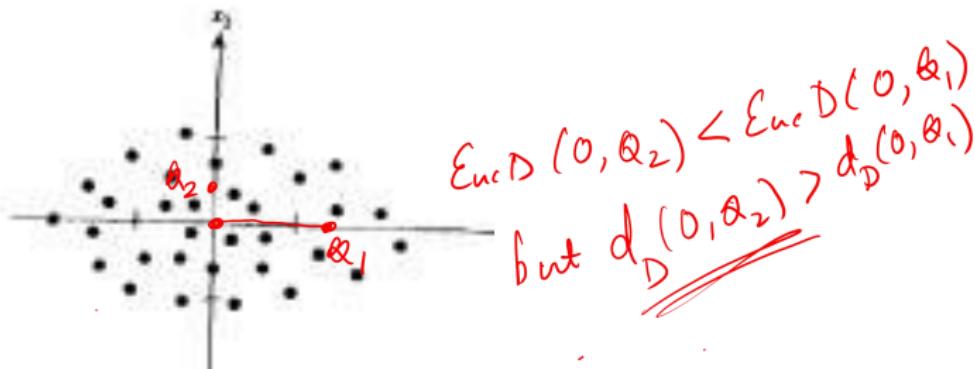
Point one and two have the same Euclidean Distance from Centroid but only point one is a member of the distribution. To detect point two as outlier, $\text{dist.}(\text{point two}, \text{centroid})$ should be much higher than $\text{dist.}(\text{point one}, \text{Centroid})$. Mahalanobis distance can be used here instead.

Variance scaled distance:

- Distance to origin

$$d(p=0; \alpha) = \sqrt{\frac{y_1^2}{s_{11}} + \frac{y_2^2}{s_{22}} + \dots + \frac{y_p^2}{s_{pp}}}$$

- Points closer by Euclidean distance might get further by variance-scaled distance



Contours of equal distance from the origin

Set of all points at the same distance from the mean (= origin here)

$$P=2 ;$$

$$\frac{x_1^2}{S_{11}} + \frac{x_2^2}{S_{22}} = C$$

Contour

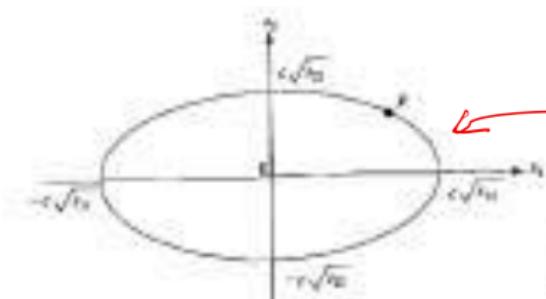


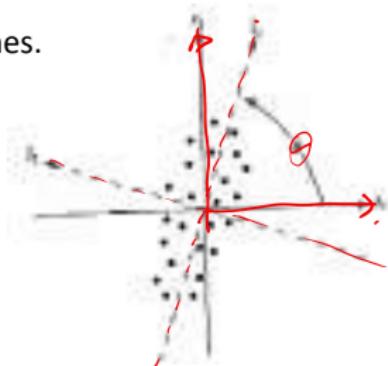
Figure 7.21 The ellipse of constant statistical distance:
 $d^2(O, P) = x_1^2/S_{11} + x_2^2/S_{22} = r^2$.

Distance when data has correlation among variables

Define a new co-ordinate system where the correlation vanishes.
We will see how to do that in general.

$$\begin{aligned} \bar{x}_1 &= x_1 \cos(\theta) + x_2 \sin(\theta) \\ \bar{x}_2 &= -x_1 \sin(\theta) + x_2 \cos(\theta) \end{aligned}$$

$$d(O, P) = \sqrt{\frac{\bar{x}_1^2}{s_{11}} + \frac{\bar{x}_2^2}{s_{22}}}$$



$$d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

a_{11}, a_{12}, a_{22} are functions of s_{11}, s_{12}, s_{22}
Exact form will be provided later.

$$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

Generalization to multiple dimensions:

$$d^2(P, Q) = \sum_{i=1}^P \sum_{j=1}^P a_{ij} (x_i - y_i)(x_j - y_j)$$

$$A_{P \times P} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1P} \\ \vdots & \ddots & \ddots & \vdots \\ a_{P1} & & a_{PP} & \end{bmatrix} = S^{-1}$$

Statistical distance between points in matrix notation

- Square of distance between two points

$$\begin{aligned} & (x-y)^T A (x-y) \quad \text{eg: } p=2 \\ & [x_1 - y_1 \quad x_2 - y_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}^T \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix} \\ & = (x_1 - y_1)^2 a_{11} + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2 \end{aligned}$$

- Distance to origin $d(O, P) \quad P = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$

$$x^T A x$$

Valid A are those where distance is always non-negative.

$$x^T A x \geq 0$$

Quadratic forms

- Given a vector x of size p , and a square matrix A of size $p \times p$, the quadratic form of x is

$$Q_A(x) = x^T A x$$

$$Q(x) = [x_1 \ x_2] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2x_1x_2 + x_2^2$$

$$Q(x) = [x_1 \ x_2 \ x_3] \begin{bmatrix} 1 & 3 & 0 \\ 3 & -1 & -2 \\ 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + 6x_1x_2 - x_2^2 - 4x_2x_3 + 2x_3^2$$

Positive definite and semi-definite matrices

- A square matrix A is positive definite if for all vectors y , the value of the quadratic form is > 0 .
 $y^T A y > 0 \quad \forall y$
 - Positive semi-definite if quadratic form is ≥ 0 .

$$y^T A y \geq 0 \quad \forall y.$$

Covariance matrix is positive semi-definite

$$S = \left(\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \bar{\mathbf{x}} \right)^T \left(\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \bar{\mathbf{x}} \right)$$
$$= \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

$\mathbf{X}_{n \times p}$

$$\forall \mathbf{y} \quad \mathbf{y}^T S \mathbf{y} \geq 0$$

$$\mathbf{y}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{y} = \underbrace{(\tilde{\mathbf{X}} \mathbf{y})^T}_{\mathbf{U}} \tilde{\mathbf{X}} \mathbf{y} = \mathbf{U}^T \mathbf{U} \geq 0$$

- Inverse of a positive semi-definite matrix is positive semi-definite [Proof in terms of spectral decomposition will be shown later.]

Population mean and covariance

- Let $f(X_1, \dots, X_p)$ be joint distribution over p variable.

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i \\ \sum_{\text{all } x_i} x_i p_i(x_i) \end{cases}$$

$$\sigma_{ik} = E[(X_i - \mu_i)(X_k - \mu_k)]$$
$$= \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k) dx_i dx_k \\ \sum_{\text{all } x_i, x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) \end{cases}$$

In matrix notation

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu$$

Warning! The X here refers to a random vector of length p . This should not be confused with the X used to denote the data matrix in earlier slides which is of size $n \times p$

Covariance

S_D

$$\boxed{S} = \text{Cov}(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$S = E(X - \mu)(X - \mu)^T$$

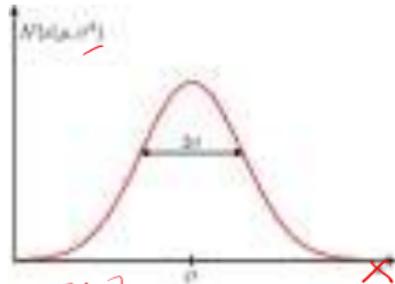
$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} (X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p)^T\right)$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_1 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & E(X_1 - \mu_1)(X_p - \mu_p) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \cdots & E(X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p - \mu_p)(X_1 - \mu_1) & E(X_p - \mu_p)(X_2 - \mu_2) & \cdots & E(X_p - \mu_p)^2 \end{bmatrix}$$

Multidimensional Gaussian Distribution

The Gaussian Distribution

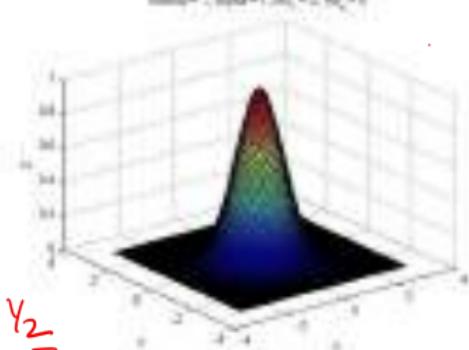


$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \cdot \\ \vdots & \ddots & \vdots \end{bmatrix} \quad |\boldsymbol{\Sigma}| = \sigma_2$$



Bi-variate Gaussian density

$$\underline{x} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}} \sqrt{\sigma_{22}}}$$

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}$$

Introducing the correlation coefficient ρ_{12} by writing $\rho_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$, and the squared distance becomes:

$$(x - \mu)^T \Sigma^{-1} (x - \mu)$$

$$= [x_1 - \mu_1, x_2 - \mu_2] \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}$$

$$\begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

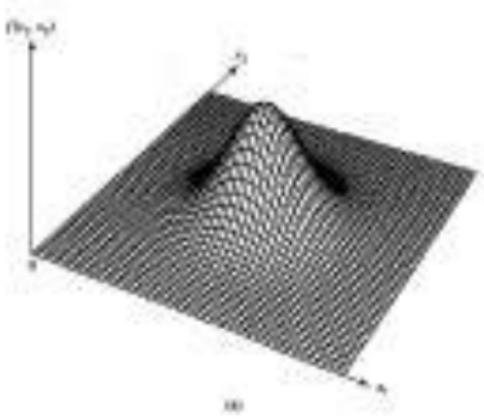
$$= \frac{\sigma_{22}(x_1 - \mu_1)^2 + \sigma_{11}(x_2 - \mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}$$

$$= \frac{1}{1 - \rho_{12}^2} \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right] \quad (4.5)$$

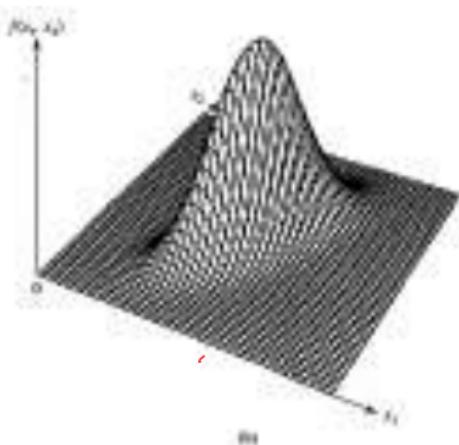
$$|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$$

$$\begin{aligned}
 f(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}} \\
 &\times \exp\left\{-\frac{1}{2(1 - \rho_{12}^2)}\left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2\right.\right. \\
 &\quad \left.\left.- 2\rho_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}
 \end{aligned} \tag{4-6}$$

Visualization



(a)



(b)

Figure 4.2: Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$.
(b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .35$.

Constant density or contour plots

Constant probability density contour = {all \mathbf{x} such that $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$ }

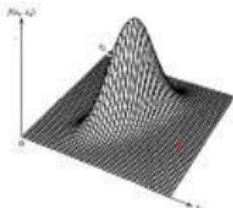
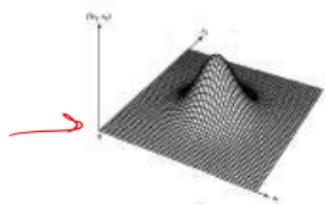


Figure 4.2 Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$.
(b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

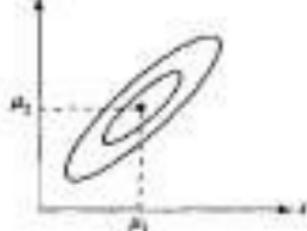
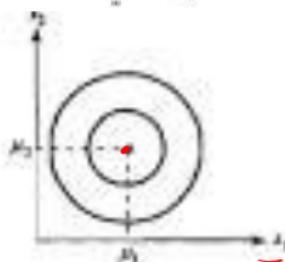
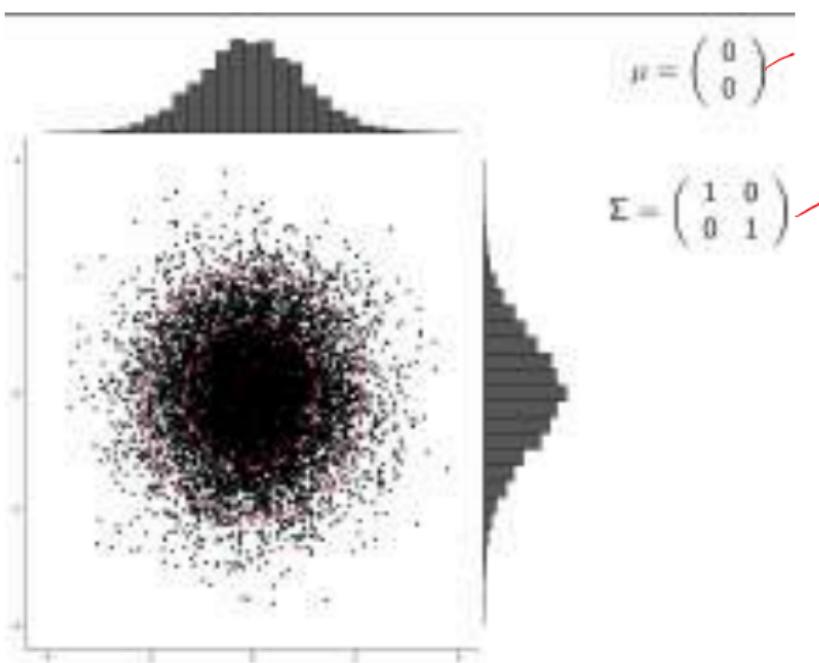
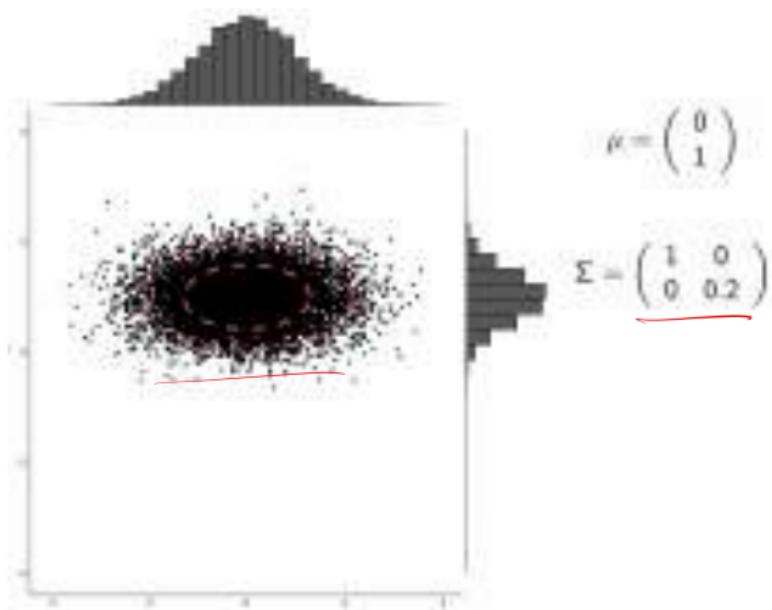


Figure 4.4 The 90% and 90% contours for the bivariate normal distributions in Figure 4.2.

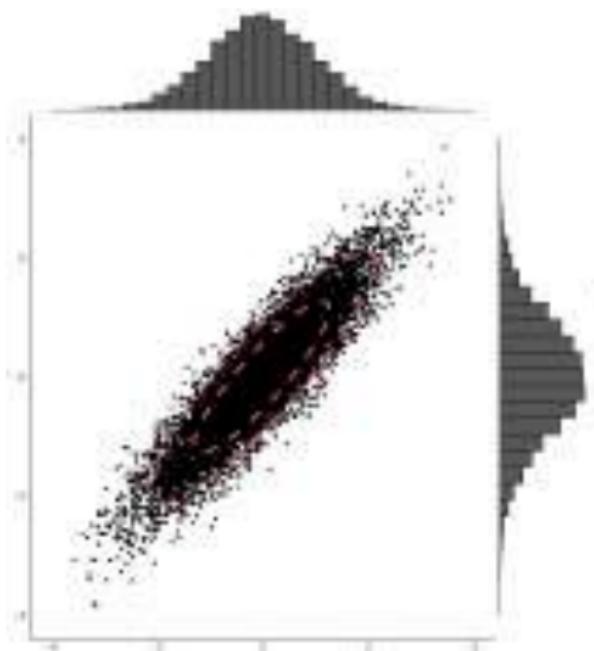
Visualizing in 2-D via contours



Different variance along each dimension



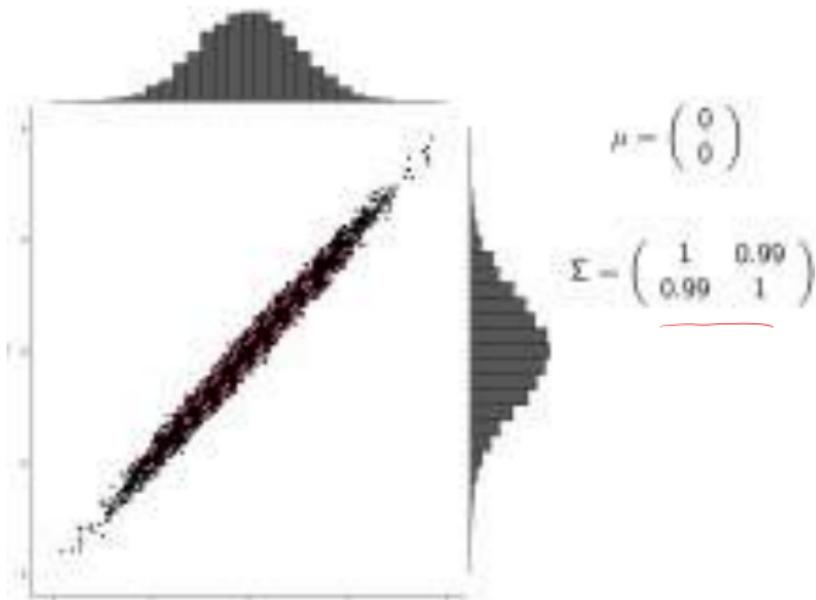
Correlated variables



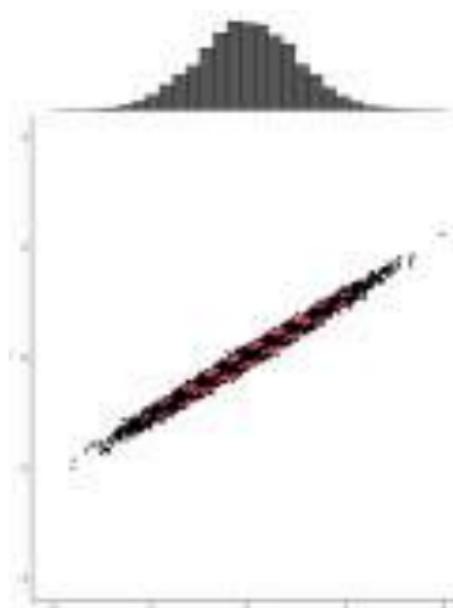
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

Highly correlated variables



Correlation with different variance.



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.54 \\ 0.54 & 0.3 \end{pmatrix}$$

$$\text{Cor}(Y_1, Y_2) = \\ 0.54 / \sqrt{0.3} = \\ 0.99$$

Lecture 29-MVA3.pdf

Moments of the Multivariate Gaussian (1)

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \mathbf{x} d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z} \right\} (\mathbf{z} + \boldsymbol{\mu}) d\mathbf{z}\end{aligned}$$

thanks to symmetry of \mathbf{z}

$$\int e^{-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z}} d\mathbf{z}_1 \dots d\mathbf{z}_p = 1$$

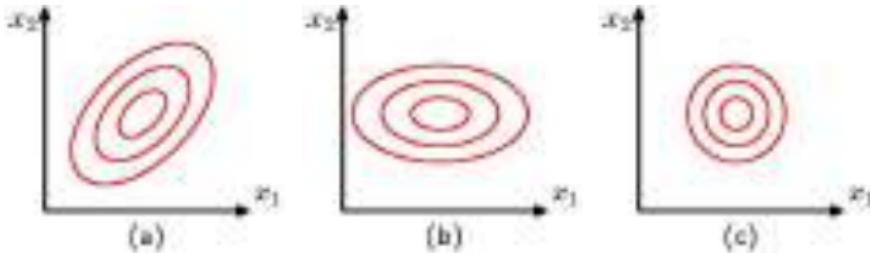
since a valid density

Moments of the Multivariate Gaussian (2)

$$\begin{bmatrix} \text{cov}(x_i, x_j) \\ \vdots \end{bmatrix} = \text{cov}(x)$$

$$\mathbb{E}[xx^T] = \mu\mu^T + \Sigma$$

$$\text{cov}[x] = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T] = \underline{\Sigma}$$



Properties of Gaussian Distribution

- Let X be a p -dimensional random vector following a Normal distribution

$$\underline{\underline{X}} \sim N(\underline{\mu}_{p \times 1}, \underline{\Sigma}_{p \times p})$$
$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Linear combination of components of \underline{X} are also normal.

- Let c_1, c_2, \dots, c_p be arbitrary constants

$$\underline{\underline{Y}} = c_1 \underline{X}_1 + c_2 \underline{X}_2 + \dots + c_p \underline{X}_p = \sum_{j=1}^p c_j \underline{X}_j = \underline{c}' \underline{X}$$

$$c = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix}$$

- Then Y also follows a normal distribution

$$\underline{\underline{Y}} \sim N(\underline{c}' \underline{\mu}, \underline{c}' \underline{\Sigma} \underline{c})$$

$$Y \in \mathbb{R}$$
$$\mu(Y) = c^\top \mu$$
$$\text{var}(Y) = c^\top \Sigma c$$

Reading material: Lesson 2: Linear Combinations of Random Variables | STAT 505 (psu.edu)

Example 2.2: Monthly Employment Data

Another example where we might be interested in linear combinations is in the Monthly Employment Data. Here we have observations on 6 variables:

- X_1 : Number people laid off or fired
- X_2 : Number of people resigning
- X_3 : Number of people retiring
- X_4 : Number of jobs created
- X_5 : Number of people hired
- X_6 : Number of people entering the workforce

Net employment change *increase*

Looking at the net job increase, which is equal to the number of jobs created, minus the number of jobs lost.

$$Y = X_4 - X_1 - X_2 - X_3$$

In this case, we have the number of jobs created, (X_4), minus the number of people laid off or fired, (X_1), minus the number of people resigning, (X_2), minus the number of people retired, (X_3). These are all of the people that have left their jobs for whatever reason.

In this case

$$C^T X = Y$$

$$\underline{c_1 = c_2 = c_3 = -1 \text{ and } c_4 = 1}$$

Mean and variance of Y

The population mean of a linear combination is equal to the same linear combination of the population means of the component variables. i.e.

thus

$$Y = c_1X_1 + c_2X_2 + \dots + c_pX_p = \sum_{j=1}^p c_j X_j = c'X$$

$$E(Y) = c_1\mu_1 + c_2\mu_2 + \dots + c_p\mu_p = \sum_{j=1}^p c_j\mu_j = c'\mu$$

$$\text{var}(Y) = \sum_{j=1}^p \sum_{k=1}^p c_j c_k \sigma_{jk} = c' \Sigma c$$

Applies for any set of p R.V.s irrespective of whether they are jointly Gaussian

$$f(x_1 \dots x_p)$$
$$\text{var}(Y) = E[(Y - E(Y))^2] -$$

$$\int_y (y - E(y)) f(y) dy = \int_{x_1 \dots x_p} \left(\left(\sum_{j=1}^p c_j x_j - \sum_{j=1}^p c_j \mu_j \right)^2 f(x_1 \dots x_p) \right)^2 dx_1 \dots x_p$$

Proof

$$f(x_1 \dots x_p) = f(x_j, x_k) f(\dots | x_j, x_k)$$

$$\int \left(\sum_{j=1}^p c_j (x_j - \mu_j) \right)^2 f(x_1 \dots x_p) dx_1 \dots x_p$$

$$\begin{aligned} &= \sum_{j=1}^p \sum_{k=1}^p \int_{x_1 \dots x_p} c_j c_k (x_j - \mu_j)(x_k - \mu_k) \underbrace{f(x_1 \dots x_p)}_{\int_{x-\{x_j, x_k\}} f(x-x_i, x_j, x_k)} dx_1 \dots x_p \\ &= \sum_{j=1}^p \sum_{k=1}^p c_j c_k \cdot \int_{\substack{x_j, x_k \\ x_j, x_k}} (x_j - \mu_j)(x_k - \mu_k) f(x_j, x_k) dx_j dx_k \end{aligned}$$

$$\sum_{j=1}^p \sum_{k=1}^p c_j c_k \sigma_{jk} \quad \sigma_{jk} \uparrow \quad \text{cov}(x_j, x_k)$$

Example 2-1: Women's Health Survey (Linear Combinations)

The Women's Health Survey data contains observations for the following variables:

- X_1 : calcium (mg)
- X_2 : iron (mg)
- X_3 : potassium (mg)
- X_4 : vitamin A (ug)
- X_5 : vitamin C (mg)

In addition to addressing questions about the individual nutritional components, we may wish to address questions about certain combinations of these components. For instance, we might want to ask what is the total intake of vitamins A and C (6 mg). We note that in this case, Vitamin A is measured in micrograms while Vitamin C is measured in milligrams. There are a thousand micrograms per milligram so the total intake of the two vitamins, V , can be expressed as the following:

$$V = 0.001X_4 + X_5 \quad Y \text{ is in milligrams}$$

In this case, our coefficients c_1 , c_2 and c_3 are all equal to 0 since the variables X_1 , X_2 and X_3 do not appear in this expression. In addition, c_4 is equal to 0.001 since one microgram of vitamin A is equal to 0.001 milligrams of vitamin A. In summary, we have

$$c_1 = c_2 = c_3 = 0, c_4 = 0.001, c_5 = 1$$

1985, the USDA commissioned a study of women's nutrition. Nutrient intake was measured for a random sample of 737 women aged 25-50 years.

Example 2.3: Women's Health Survey (Population Mean)

The following table shows the sample means for each of the five nutritional components that are computed in the previous lesson.

Variable	Mean
Carbohydrate	624.0 mg
Fibre	11.3 mg
Protein	45.3 g
Vitamin A	899.0 µg
Vitamin C	75.3 mg

As previously, we define \bar{Y} to be the total intake of vitamins A and C (in mg) as:

$$\bar{Y} = 0.001\bar{X}_A + \bar{X}_C$$

Then we can work out the estimated mean intake of the two vitamins as follows:

$$\bar{Y} = 0.001\bar{X}_A + \bar{X}_C = 0.001 \times 899.0 + 75.3 = 0.899 + 75.3 = 76.200 \text{ mg}$$

Example 2-4: Women's Health Survey (Population Variance)

Looking at the Women's Nutrition survey data we obtained the following variance/covariance matrix as shown below from the previous lesson:

$$S = \begin{pmatrix} 157820.4 & 981.1 & 9075.8 & 102411.1 & 6701.6 \\ 981.1 & 35.8 & 114.1 & 2383.2 & 137.7 \\ 9075.8 & 114.1 & 814.9 & 7330.1 & 477.2 \\ 102411.1 & 2383.2 & 7330.1 & 2068452.4 & 13943.3 \\ 6701.6 & 137.7 & 477.2 & 22960.3 & 5410.2 \end{pmatrix}$$

If we wanted to take a look at the total intake of vitamins A and C (in mg) (remember we defined this earlier as:

$$Y = 0.001X_4 + X_5$$

Therefore the sample variance of Y is equal to $(0.001)^2$ times the variance for X_4 , plus the variance for X_5 , plus 2 times 0.001 times the covariance between X_4 and X_5 . The next four lines carry out the mathematical calculations using these values.

$$\begin{aligned} s_Y^2 &= 0.001^2 s_{X_4}^2 + s_{X_5}^2 + 2 \times 0.001 s_{XY} \\ &= 0.000001 \times 2068452.4 + 13943.3 + 0.002 \times 22960.3 \\ &\approx 2.7 + 3418.3 + 44.1 \\ &= 3485.1 \end{aligned}$$

More examples

- [4.1 - Comparing Distribution Types | STAT 505 \(psu.edu\)](#)

Lecture 30-MVA4.pdf

Other properties

- Every single variable has a univariate normal distribution.

$$x_j \sim N(\mu_j, \sigma_{jj}^2) \quad \text{and} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N(\boldsymbol{\mu}, \Sigma)$$

- Any subset of the variables also has a multivariate normal distribution.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_3 \end{bmatrix}; \Sigma\right)$$

- Zero covariance terms or a diagonal covariance matrix implies that the variables are independent of each other.

$$\Sigma_{3 \times 3} = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \quad x_1 \perp\!\!\!\perp x_2 \quad x_3 \perp\!\!\!\perp x_1 \\ x_2 \perp\!\!\!\perp x_3$$

- Any conditional distribution for a subset of the variables conditional on known values for another subset of variables is a multivariate distribution.

Partitioned Gaussian Distributions

$$p(\underline{\mathbf{x}}) = \mathcal{N}(\mathbf{x} | \underline{\mu}, \underline{\Sigma})$$

$$\underline{x}_a \cap \underline{x}_b = \emptyset$$

$$\underline{\mathbf{x}} = \begin{pmatrix} \underline{\mathbf{x}}_a \\ \underline{\mathbf{x}}_b \end{pmatrix} \quad \underline{\mu} = \begin{pmatrix} \underline{\mu}_a \\ \underline{\mu}_b \end{pmatrix} \quad \underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_{aa} & \underline{\Sigma}_{ab} \\ \underline{\Sigma}_{ba} & \underline{\Sigma}_{bb} \end{pmatrix}$$

$$\underline{\Lambda} \equiv \underline{\Sigma}^{-1} \quad \underline{\Lambda} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

$$\underline{x}_a \sim \mathcal{N}(\underline{\mu}_a; \overline{\underline{\Sigma}_{aa}})$$

Partitioned Conditionals and Marginals

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a | \underline{\boldsymbol{\mu}_{a|b}}, \underline{\Sigma_{a|b}})$$

$$\underline{\Sigma_{a|b}} = \underline{\Lambda_{aa}^{-1}} = \underline{\Sigma_{aa}} - \underline{\Sigma_{ab}} \underline{\Sigma_{bb}^{-1}} \underline{\Sigma_{ba}}$$

$$\underline{\boldsymbol{\mu}_{a|b}} = \underline{\Sigma_{a|b}} \{ \underline{\Lambda_{aa}} \underline{\boldsymbol{\mu}_a} - \underline{\Lambda_{ab}} (\mathbf{x}_b - \underline{\boldsymbol{\mu}_b}) \}$$

$$= \underline{\boldsymbol{\mu}_a} - \underline{\Lambda_{aa}^{-1}} \underline{\Lambda_{ab}} (\mathbf{x}_b - \underline{\boldsymbol{\mu}_b})$$

$$= \underline{\boldsymbol{\mu}_a} + \underline{\Sigma_{ab}} \underline{\Sigma_{bb}^{-1}} (\mathbf{x}_b - \underline{\boldsymbol{\mu}_b})$$

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$

$$= \mathcal{N}(\mathbf{x}_a | \underline{\boldsymbol{\mu}_a}, \underline{\Sigma_{aa}})$$

$$a = \{1\}; \quad b = \{2\}$$

$$P(x_1 | x_2) = \mathcal{N}(x_1 | \underline{\boldsymbol{\mu}_{1|2}}, \underline{\Sigma_{1|2}})$$

$P(\underline{x}_1 | \underline{x}_2)$) Derive for bi-variate case.

$$\underline{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix} -$$

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \underline{\boldsymbol{\mu}_1} \\ \underline{\boldsymbol{\mu}_2} \end{bmatrix}$$

$$\sigma_{1|2} = \frac{\sigma_{11}^2 - \sigma_{12}^2}{\sigma_{22}^2} \quad \checkmark$$

$$\underline{\boldsymbol{\mu}_{1|2}} = \underline{\boldsymbol{\mu}_1} + \frac{\sigma_{12}}{\sigma_{22}} (\underline{\boldsymbol{\mu}_2} - \underline{\boldsymbol{\mu}_1})$$

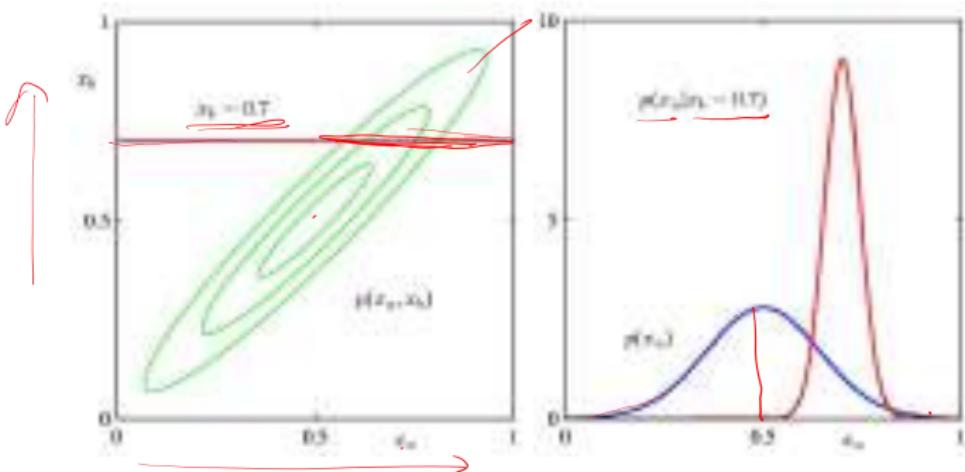
$$\therefore \therefore \therefore$$

Conditional distribution for bivariate case

$$\text{Mean} = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$$

$$\text{Variance} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$


Partitioned Conditionals and Marginals



Demos: https://colab.research.google.com/github/goodboychan/goodboychan.github.io/blob/main/_notebooks/2021-08-11-Multivariate-distribution.ipynb

Example 6-1: Conditional Distribution of Weight Given Height for College Men

Suppose that the weights (lbs) and heights (inches) of undergraduate college men have a multivariate normal distribution with mean vector $\mu = \begin{pmatrix} 175 \\ 71 \end{pmatrix}$ and covariance matrix $\Sigma = \begin{pmatrix} 250 & 48 \\ 48 & 8 \end{pmatrix}$. $\bar{\mu}_{x|a} = \begin{pmatrix} 80 \\ 71 \end{pmatrix}$ — $\Sigma_{x|a} = \begin{pmatrix} 550/4 & 20 \\ 20 & 8 \end{pmatrix}$

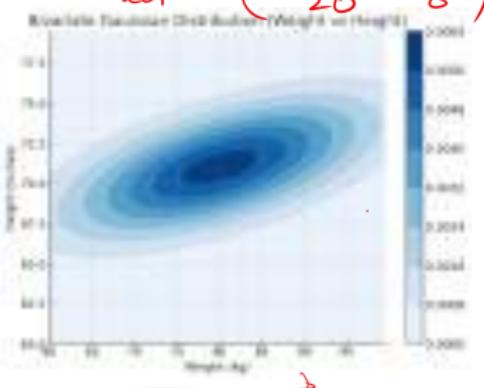
The conditional distribution of X_1 , weight given $x_2 = \text{height}$ is a normal distribution with:

$$\begin{aligned}\text{Mean: } \bar{\mu}_{x|a} &= \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2) \\ &= 175 + \frac{48}{8}(71 - 71) \\ &= 175 + 0\end{aligned}$$

$$\begin{aligned}\text{Variance: } \sigma_{x|a}^2 &= \frac{\sigma_{11}^2}{\sigma_{22}^2} \\ &= 180 - \frac{48^2}{8} \\ &= 180 - 576 \\ &= 250\end{aligned}$$

For instance, for men with height = 70, weights are normally distributed with mean = $175 + 5(70) = 170$ pounds and variance = $250/4 = 62.5$ pounds squared.

Notice that we have generated a simple linear regression model that relates weight to height.



Geometry of the Multivariate Normal Distribution

- Can we characterize the shape and orientation of the ellipse that defines that contours of equal density?

Constant probability density contour = (all \mathbf{x} such that $(\mathbf{x} - \mu)' \Sigma^{-1}(\mathbf{x} - \mu) = c^2$)

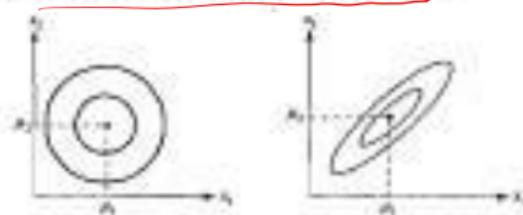


Figure 4.4 The 50% and 90% contours for the bivariate normal distribution in Figure 4.2.

- We will see that these can be characterized using eigen vectors and values of the covariance matrix.

Eigen values and Eigen vectors

- A square matrix A has a eigen value, eigen vector pair $\lambda, e \neq 0$ if $Ae = \lambda e$ where norm of e is 1

Let A be a $k \times k$ square symmetric matrix. Then A has k pairs of eigenvalues and eigenvectors namely,

$$\lambda_1, e_1 \quad \lambda_2, e_2 \quad \dots \quad \lambda_k, e_k \quad e_i^T e_j = e_i^T e_i - \langle e_i, e_j \rangle \quad (2-15)$$

The eigenvectors can be chosen to satisfy $1 = e_1^T e_1 = \dots = e_k^T e_k$ and be mutually perpendicular. The eigenvectors are unique unless two or more eigenvalues are equal.

$$e_j^T e_k = 0 \quad \forall j \neq k$$

Spectral decomposition of A

$$\underbrace{A}_{(k \times k)} = \lambda_1 \underbrace{e_1}_{(k \times 1)(1 \times k)} e_1^T + \lambda_2 \underbrace{e_2}_{(k \times 1)(1 \times k)} e_2^T + \dots + \lambda_k \underbrace{e_k}_{(k \times 1)(1 \times k)} e_k^T$$

Spectral decomposition of a positive semi-definite matrix

- If A is positive-definite than all eigen-values ≥ 0
- Example:

$$\underline{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

$y^T A y \geq 0$
choose $y = e_j$ to show
that $\lambda_j \geq 0$

- First find Eigen values and vectors.

- [4.5 - Eigenvalues and Eigenvectors | STAT 505 \(psu.edu\)](#) [HW]

$$e_1 = \begin{pmatrix} \frac{\sqrt{1+\rho}}{\sqrt{2}} \\ \frac{\sqrt{1-\rho}}{\sqrt{2}} \end{pmatrix} \text{ for } \lambda_1 = 1 + \rho \text{ and } e_2 = \begin{pmatrix} \frac{\sqrt{1-\rho}}{\sqrt{2}} \\ \frac{\sqrt{1+\rho}}{\sqrt{2}} \end{pmatrix} \text{ for } \lambda_2 = 1 - \rho$$

$$e_1 \cdot e_2 = 0$$

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = (1+\rho) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} + (1-\rho) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

$$\text{if } \Sigma = \lambda_1 e_1 e_1^T + \dots + \lambda_p e_p e_p^T$$

then

$$\Sigma^{-1} = \frac{1}{\lambda_1} e_1 e_1^T + \dots + \frac{1}{\lambda_p} e_p e_p^T$$

Geometry of the Multivariate Gaussian

$$C^2 = \underline{\Delta^2} = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

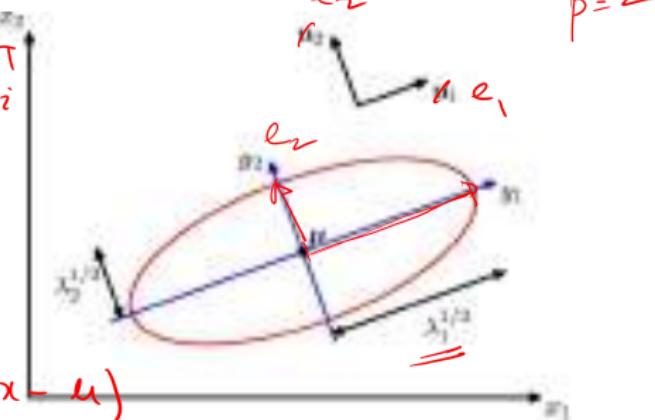
$$\underline{\Sigma^{-1}} = \sum_{i=1}^P \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

$$C^2 = \Delta^2 = \sum_{i=1}^P \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$$

$$(\mathbf{x} - \boldsymbol{\mu})^T \left[\sum_{i=1}^P \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^T \right] (\mathbf{x} - \boldsymbol{\mu})$$

$$\underline{\Delta^2} = \sum_{i=1}^P \frac{1}{\lambda_i} (e_i^T (\mathbf{x} - \boldsymbol{\mu}))^2$$

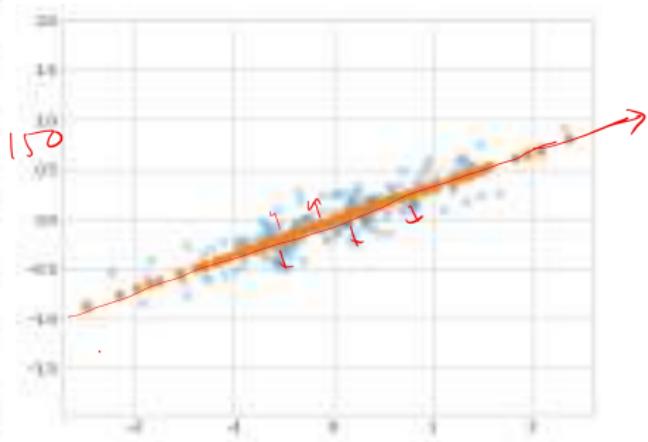
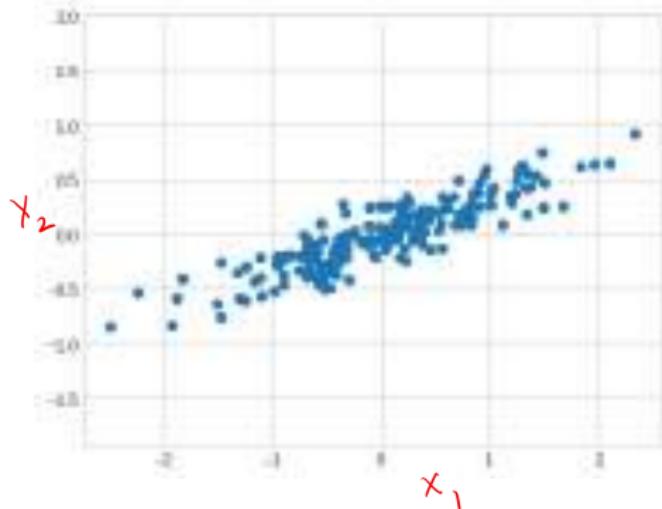


Principal component analysis

Projecting high-dimensional data

- When multivariate dataset has a large number of variables, analysis and interpretation of the data may be hard.
- Too many variables pairs, so pairwise correlation may be hard to grasp.
- For convenient visualization and interpretation
 - Reduce the number of variables.
- How to reduce number of variables while capturing most of the information in the data
 - Information == variance

Example



What is the best way to summarize this two dimensional data into a single dimension without losing much of the dispersion?

How to reduce number of variables: many methods

- Principal component analysis ~~is~~
- Factor analysis is
- Other embedding methods
 - Random projection
 - T-SNE ~~is~~

Principal component analysis

- Let original set of p variables be $\underline{X_1, X_2, \dots, X_p}$
- Define a smaller set of new variables that are linear combinations of existing variables.

$$\begin{array}{lcl} Y_1 & = & e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 & = & e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ & \vdots & \\ Y_p & = & e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{array}$$

Variance and Co-variance of the new variables.

Let

$$\text{var}(\underline{\underline{X}}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Then:

$$\text{var}(Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{ik} e_{il} \sigma_{kl} = \underline{\underline{e}_i' \Sigma e_i}$$

$$Y_i = \begin{bmatrix} e_{i1} \\ \vdots \\ e_{ip} \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

$$\text{cov}(Y_i, Y_j) = \sum_{k=1}^p \sum_{l=1}^p e_{ik} e_{jl} \sigma_{kl} = \underline{\underline{e}_i' \Sigma e_j}$$

Principal components

- First principal component Y_1 is chosen to maximize the variance among all possible linear combinations such that the norm of coefficients is 1.

More formally, select $e_{11}, e_{12}, \dots, e_{1p}$ that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^p \sum_{l=1}^p e_{1k} e_{1l} \sigma_{kl} = \mathbf{e}'_1 \Sigma \mathbf{e}_1$$

subject to the constraint that

$$\mathbf{e}'_1 \mathbf{e}_1 = \sum_{j=1}^p e_{1j}^2 = 1$$

Second principal component

Select $e_{21}, e_{22}, \dots, e_{2p}$ that maximizes the variance of this new component...

$$\text{var}(Y_2) = \sum_{k=1}^p \sum_{l=1}^p e_{2k} e_{2l} \sigma_{kl} = \underline{\mathbf{e}_2' \Sigma \mathbf{e}_2}$$

subject to the constraint that the sums of squared coefficients add up to one,

$$\underline{\mathbf{e}_2' \mathbf{e}_2} = \sum_{j=1}^p e_{2j}^2 = 1$$

along with the additional constraint that these two components are uncorrelated,

$$\text{cov}(Y_1, Y_2) = \sum_{k=1}^p \sum_{l=1}^p e_{1k} e_{2l} \sigma_{kl} = \underline{\mathbf{e}_1' \Sigma \mathbf{e}_2} = 0$$

i^{th} Principal Component (PCA); Y_i

We select e_1, e_2, \dots, e_N to maximize

$$\text{var}(Y_i) = \sum_{k=1}^K \sum_{l=1}^L e_{ik} e_{il} \sigma_{kl} = e'_i \Sigma e_i$$

subject to the constraint that the sum of squared coefficients add up to one... along with the additional constraint that this new component is uncorrelated with all the previously defined components.

$$e'_i e_i = \sum_{j=1}^K e_{ij}^2 = 1$$

$$\text{cov}(Y_1, Y_i) = \sum_{k=1}^K \sum_{l=1}^L e_{1k} e_{il} \sigma_{kl} = e'_1 \Sigma e_i = 0$$

$$\text{cov}(Y_2, Y_i) = \sum_{k=1}^K \sum_{l=1}^L e_{2k} e_{il} \sigma_{kl} = e'_2 \Sigma e_i = 0$$

⋮

$$\text{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^K \sum_{l=1}^L e_{i-1,k} e_{il} \sigma_{kl} = e'_{i-1} \Sigma e_i = 0$$

For what Y_1 is $\text{Variance}(Y_1)$ maximized?

- The coefficient of the first principal component correspond to the Eigen vector with the maximum Eigen value.

More generally

- The i -th principal component corresponds to the i -th largest eigen vector.

The variance for the i -th principal component is equal to the i -th eigenvalue.

$$\text{var}(Y_i) = \text{var}(c_{i1}X_1 + c_{i2}X_2 + \dots + c_{ip}X_p) = \lambda_i$$

$$\text{cov}(Y_i, Y_j) = 0$$

The proportion of variance explained

- The total variance of X
- We can show that sum of p Eigen values equals the total variance

- The fraction of variance explained by the i -th Eigen value

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$$

Reducing number of dimensions

- Variance explained by first k Eigen values

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$$

11.3 - Example: Places Rated

Example 11.2: Places Rated

We will use the Places Rated Almanac data (Boyal and Saenger) which rates 829 communities according to nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

[11.3 - Example: Places Rated | STAT 505 \(psu.edu\)](#)

Notes

- The data for many of the variables are strongly skewed to the right.
- The log transformation was used to normalize the data.

More demos

- <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb>

Lecture 31-MVA5.pdf

For what $\underline{Y_1}$ is $\text{Variance}(Y_1)$ maximized?

- The coefficient of the first principal component correspond to the Eigen vector with the maximum Eigen value.

$$\begin{cases} \max_{e_1} e_1^\top \Sigma e \\ \text{s.t. } e_1^\top e = 1 \end{cases}$$

$$\xrightarrow{\text{Lagrangean multiplier based}} \max_{e_1} e_1^\top \Sigma e + \lambda [e_1^\top e - 1]$$

$$F(e_1) =$$

Lagrangean
multiplier based
reverts to
original constraint
objective.

$$\nabla_{e_1} F = 0$$

$$\Rightarrow \Sigma e_1 + \lambda e_1 = 0$$

$$\Rightarrow \Sigma e_1 = -\lambda e_1 \Rightarrow \boxed{\Sigma e_1 = \lambda e_1}$$

$$\max_{e_i} \vec{e}_i^T \Sigma \vec{e}_i = \max_{e_i, \lambda} \vec{e}_i^T \lambda e = \max_{\lambda} \lambda$$

s.t e_i is a eigen vector.

choose the e_i corresponding to
the largest eigen value.

More generally

- The i -th principal component corresponds to the i -th largest eigen vector.

The variance for the i -th principal component is equal to the i -th eigenvalue.

$$\text{var}(Y_i) = \text{var}(c_{i1}X_1 + c_{i2}X_2 + \dots + c_{ip}X_p) = \lambda_i$$

$$\text{cov}(Y_i, Y_j) = 0$$

The proportion of variance explained

- The total variance of X

$$\sum_{j=1}^p \sigma_j^2 = \sum_{j=1}^p \text{Var}(x_j) \quad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots \\ \sigma_{12}^2 & \sigma_{22}^2 & \dots \\ \vdots & \vdots & \ddots \\ \sigma_{pp}^2 & \dots & \dots \end{bmatrix}$$

- We can show that sum of p Eigen values equals the total variance

Show

$$\sum_{j=1}^p \sigma_j^2 = \sum_{j=1}^p \lambda_j$$

$$\Sigma = \sum_{j=1}^p \lambda_j e_j e_j^\top$$

$$e_j^\top e_j = 1 = \sum_{r=1}^p e_{jr}^2$$

$$\begin{aligned} \Sigma &= \sum_{j=1}^p \lambda_j e_j e_j^\top \\ &= \left[\begin{array}{c} \lambda_1 e_1 e_1^\top \\ \vdots \\ \lambda_p e_p e_p^\top \end{array} \right] = \lambda_j \begin{bmatrix} e_{j1}^2 & e_{j2}^2 & \dots & e_{jp}^2 \end{bmatrix} \\ &\quad \text{Trace}(\lambda_j e_j e_j^\top) = \lambda_j \end{aligned}$$

Reducing number of dimensions

- Variance explained by first k Eigen values

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_n} \quad ||$$

11.3 - Example: Places Rated

Example 11.2: Places Rated

We will use the Places Rated Almanac data (Loyola and Saenger) which rates 829 communities according to nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

[11.3 - Example: Places Rated | STAT 505 \(psu.edu\)](#)

Notes

- The data for many of the variables are strongly skewed to the right.
- The log transformation was used to normalize the data.

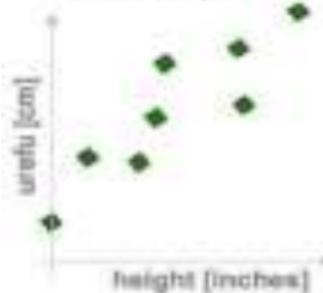
More demos

- <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb>

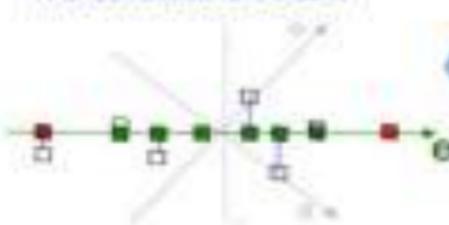
PCA in a nutshell

1. correlated hi-d data

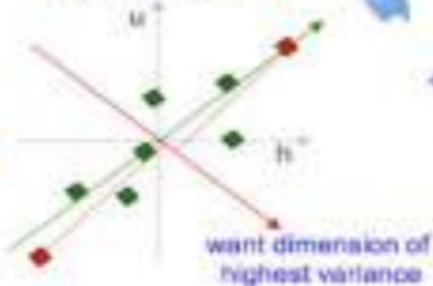
FDA? (mean "height" is 60in)



7. uncorrelated low-d data

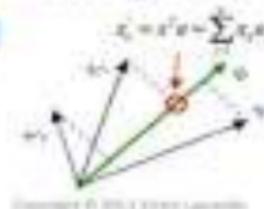


2. center the points



want dimension of highest variance

6. project data points to those eigenvectors



Copyright © 2010, Yann LeCun et al.

3. compute covariance matrix

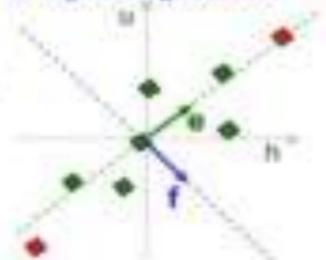
$$\frac{1}{n} \begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.8 \end{bmatrix} = cov(\mathbf{x}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$$

4. eigenvectors + eigenvalues

$$\begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.8 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$
$$\begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.8 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \lambda_2 \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

`np.linalg.eig(cov(data))`

5. pick m-d eigenvectors w. highest eigenvalues



T-SNE: T-distributed stochastic neighbourhood embedding

Reading material:

<https://www.dailydoseofds.com/formulating-and-implementing-the-t-sne-algorithm-from-scratch/>

T-SNE

- Another data projection method, specifically designed for visualizing high dimensional data in two dimensions.
- Preserves local similarities and clusters better than PCA
- Creates non-linear projection

Lecture 32-MVA6.pdf

T-SNE: T-distributed stochastic neighbourhood embedding

Reading material:

<https://www.dailydoseofds.com/formulating-and-implementing-the-t-sne-algorithm-from-scratch/>

Kevin Murphy's book chapter: Section 20.4.10 in [Probabilistic ML book](#).

Demos:

- <https://projector.tensorflow.org/>
- <https://distill.pub/2016/misread-tsne/>

T-SNE

- Another data projection method, specifically designed for visualizing high dimensional data in two dimensions.
- Preserves local similarities and clusters better than PCA
- Creates non-linear projection

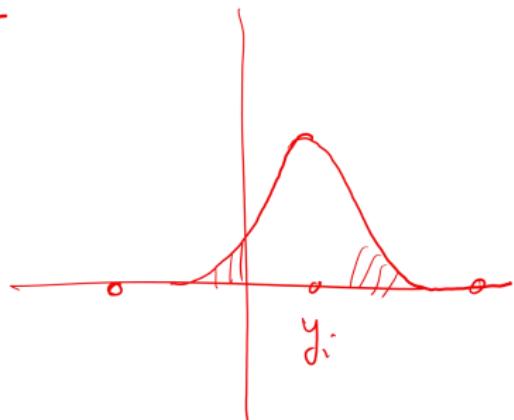
First, SNE

- Described here.
- <https://www.dailydoseofds.com/formulating-and-implementing-the-t-sne-algorithm-from-scratch>

Limitations of SNE

- A fundamental problem with SNE and many other embedding techniques is that they tend to squeeze points that are relatively far away in the high dimensional space close together in the low dimensional (usually 2d) embedding space; this is called the the crowding problem, and arises due to the use of squared errors (or Gaussian probabilities).

$$q(j|i) \propto e^{-\frac{\|\gamma_i - \gamma_j\|^2}{2}}$$



T-SNE

- Use a probability distribution in latent space that has heavier tails, which eliminates the unwanted attractive forces between points that are relatively far in the high dimensional space.

Student t distribution - with one degree of freedom:

$$q(j|i) = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}$$

20.4.10.4 Choosing the length scale

An important parameter in t-SNE is the local bandwidth σ_i^2 . This is usually chosen so that P_{ij} has a perplexity chosen by the user.⁷ This can be interpreted as a smooth measure of the effective number of neighbors.

Unfortunately, the results of t-SNE can be quite sensitive to the perplexity parameter, so it is wise to run the algorithm with many different values. This is illustrated in Figure 20.42. The input data is 2d, so there is no distortion generating by mapping to a 2d latent space. If the perplexity is too small, the method tends to find structure within each cluster which is not truly present. At perplexity 30 (the default for scikit-learn), the clusters seem equi-distant in embedding space, even though some are closer than others in the data space. Many other caveats in interpreting t-SNE plots can be found in [WVJ16].

7. The perplexity is defined to be $2^{\text{H}(P_i)}$, where $\text{H}(P_i) = -\sum_j p_{ij} \log_2 p_{ij}$ is the entropy; see Section 6.1.5 for details. A big radius around each point (large value of σ_i) will result in a high entropy, and thus high perplexity.

Lecture 33-34-Hypothesis Testing.pdf

Hypothesis Testing

Chapters 8

Introduction

- Hypothesis testing is key to scientific inquiry
- We just have some hypothesis H
- To check it we collect data D
- We check if H is consistent with D
 - Consistency is probabilistic, and there is no 0/1 answer.
- Example:
 - Say you got a shipment of cables, and average breaking strength is claimed to be at least 7000 pounds per square inch (PSI).
 - D = You test 10 random cables and record PSI of each.
 - Hypothesis: $\text{PSI} \geq 7000$.
 - Hypothesis testing: Is the hypothesis consistent with the observed data?

More real-world examples of hypothesis testing

- Number of clicks on the video is at least 100
- Average order value has increased since last financial year
- Investing in A brings a higher return than investing in B
- The new user interface converts more users into customers than the expected 30%

Types of hypothesis testing

- Parametric Vs Non-parametric
 - Parametric
 - Assume that the underlying data distribution has a known parametric form. E.g. Normal or exponential
 - Non-parametric
 - Unknown functional form of the distribution.
- One-sample Vs Two-sample
 - One-sample: D = sample from one distribution
 - Two-sample: D1, D2 = samples from two different distributions. Hypothesis is comparing them.
- Paired Vs Unpaired:

Parametric hypothesis testing for a single population

8.3.2 of Ross Text-book

Parametric Hypothesis testing

- Let $F_\theta(X)$ be a distribution on X with unknown parameters $\underline{\theta}$
- We want to test some property of θ E.g.
 - i) $\theta = \theta_0$
 - ii) $\theta \geq \theta_0$
 - iii) $\theta \leq \theta_0$

Simple tests: all parameters fully specified

$$\underline{\theta} = \theta_0 \quad \text{eg: } X \sim F_0(x) = \text{Exp}(\lambda); \quad \lambda = 2$$

Complex tests

$$\theta \geq \theta_0 \quad \text{or} \quad \theta \leq \theta_0 \quad \text{or} \quad \theta \in (\theta_1, \theta_2]; \quad \theta_1 = \theta_{1,0}$$

$N(\mu, \sigma^2)$

θ_2 is unconstraint
 $\mu = 2$
 σ^2 is unspecified

Step 1: Collect data (Evidence)

$$D = \{x_1, x_2, x_3, \dots, x_n\}$$

Why not simple likelihood tests?

For simple hypothesis where all parameters are specified by the user: $\theta = \theta_0$

A default method

Measure log likelihood of D = $\sum_{i=1}^N \log F_{\theta_0}(x_i)$ ✓

If $LL(D|\theta_0)$ is high enough $LL(D|\theta_0)$
then accept hypothesis

Shortcomings:

- 1) Threshold is not specified.
- 2) Only applicable for simple hypothesis

Step 2: Formulate the question using a pair of hypothesis

- Null hypothesis H_0 that tests for equality (Reason will be clear later)

$$\theta = \theta_0 \quad (H_0)$$

- Alternative hypothesis: Alternative values of the parameters

a) User is asserting that $\theta = \theta_0$, then

$$H_1: \theta \neq \theta_0$$

b) User wants to test if $\theta \leq \theta_0$, then

$$H_1: \theta \leq \theta_0$$

c) User wants to test if $\theta \geq \theta_0$, then

$$H_1: \theta \geq \theta_0$$

Examples

- Is the average IQ of students in this class greater than 120?

$$H_0: \mu = \mu_0 = 120 \quad ; \quad H_1: \mu > 120$$

- Is the rise in temperature over the last ten years less than 2 degrees?

$$H_0: \mu = 2 \text{ deg} \quad ; \quad H_1: \mu < 2$$

Step 3: Compute if D is extreme given hypothesis

- Step 3.1: Compute a test statistic T from the data --- some summary of the data
D (Design step)
- Step 3.2: Identify the probability of T under the null hypothesis $P_{\theta_0}(T)$
 - For some F_{θ}, T it may be possible to show this in closed parametric form
 - For others, simulations may be required.
- Step 3.3: User specifies a significance level α , the error tolerance of rejecting the null hypothesis even if it is true

Step 3.4: Compute if \hat{T} is extreme under $P_{\theta_0}(T)$

Two related ways:

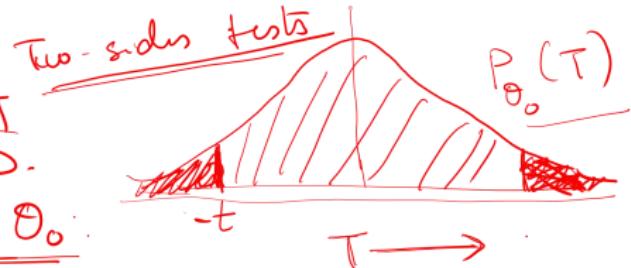
- Define a critical region C such that D being in that region is unlikely if H_0 is true, and where H_1 is more likely, Or
- Compute a p-value: the probability — assuming the null hypothesis was true — of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed.

t = observed value of T
from given data D .

$$H_0: \underline{\theta = \theta_0} ; H_1: \underline{\theta \neq \theta_0}$$

Assume $\underline{t} > 0$

$$P_{\theta_0}(T < -t \text{ or } T > t) \rightarrow \underline{\text{p-value}} \quad \theta = \theta_0$$



$H_0: \theta = \theta_0$; $H_1: \theta \leq \theta_0$ - Left-sided test

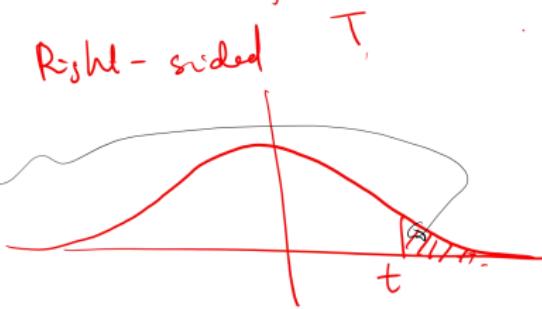
Compute

$$P_{\theta_0}(T \leq t) \text{ :: p-value}$$



$H_0: \theta = \theta_0$; $H_1: \theta > \theta_0$ Right-sided test

$$P(T \geq t) \text{ :: p-value}$$



Step 3.5: Accept/Reject decisions based on p-values

- Step 3.5: Compare p-value with given significance-level α

if p-value of test is $> \bar{\alpha}$ then
then accept the null hypothesis H_0
else
accept the alternative H_1

Step 3.4 using critical regions method instead of p-values

- Define a Critical or reject region C such that the probability of T being in C is at most α , and if t lies in C then H_1 is more likely to be true than H_0
- Defining C for different hypothesis types

i) $H_0: \theta = \theta_0$; $H_1: \theta \neq \theta_0$

Define t_{left} & t_{right} such that if

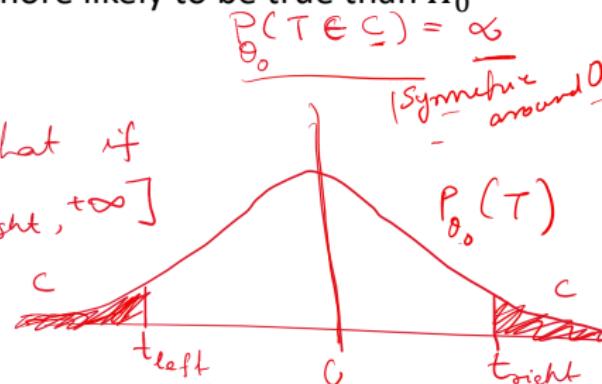
$$C = [-\infty, t_{\text{left}}] \cup [t_{\text{right}}, +\infty]$$

$$P(T \in C) = \alpha$$

$$\text{or } P(T \leq t_{\text{left}} \text{ or } T \geq t_{\text{right}})$$

Often $P_{\theta_0}(T)$ will be symmetric around 0.

so, find $t_{\alpha/2}$ s.t. $P(T \leq -t_{\alpha/2}) = \frac{\alpha}{2}$ & $P(T \geq t_{\alpha/2}) = \frac{\alpha}{2}$



r

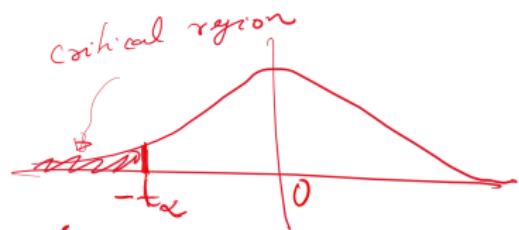
Defining C for other hypothesis

2) $H_0: \theta = \theta_0 ; H_1: \theta \leq \theta_0$

$C = [-\infty, t_\alpha]$ s.t

$$P_{\theta_0}(T \in C) = P_{\theta_0}(T \leq -t_\alpha) = \alpha$$

t_α = the upper percentile of $P_{\theta_0}(T)$



3) $H_0: \theta = \theta_0 ; H_1: \theta > \theta_0$

$C = [t_\alpha, +\infty]$ s.t

$$P_{\theta_0}(T \in C) = P_{\theta_0}(\theta \geq t_\alpha) = \alpha$$

Step 3.5: Accept/Reject decisions based on critical regions

If $\underline{t} \in C$ then reject null hypothesis
else accept " "

1) If $H_0: \theta = \theta_0 ; H_1: \theta \neq \theta_1$
 $\underline{t} < -t_{\alpha/2}$ or $\underline{t} > t_{\alpha/2}$

2)

3)

Example: Hypothesis test on mean of a normal distribution with unknown variance

Suppose that X_1, \dots, X_n is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0: \mu = \mu_0$$

against the alternative hypothesis

$$H_1: \mu \neq \mu_0$$

where μ_0 is some specified constant.

$$D = \{x_1, x_2, \dots, x_n\}$$

Possible test statistic: $T = \left| \frac{\sum_{i=1}^n x_i}{n} - \mu_0 \right|$
variance is not known so cannot define $f_{\theta_0}(T)$

A better choice of test statistic

$$T = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}}$$

$$P_{\theta_0}(T) ??$$

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

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$P(\bar{x}) \sim N(\mu_0; \frac{\sigma^2}{n})$$

- A good test-statistic is one where the distribution of T can be easily computed, and in that distribution the null hypothesis region is well separated from the alternative hypothesis.

Distribution of test statistic

- Property of sample mean and sample variance of a normal distribution

Theorem 6.5.1. If X_1, \dots, X_n is a sample from a normal population having mean μ and variance σ^2 , then \bar{X} and S^2 are independent random variables, with \bar{X} being normal with mean μ and variance σ^2/n and $(n - 1)S^2/\sigma^2$ being chi-square with $n - 1$ degrees of freedom.

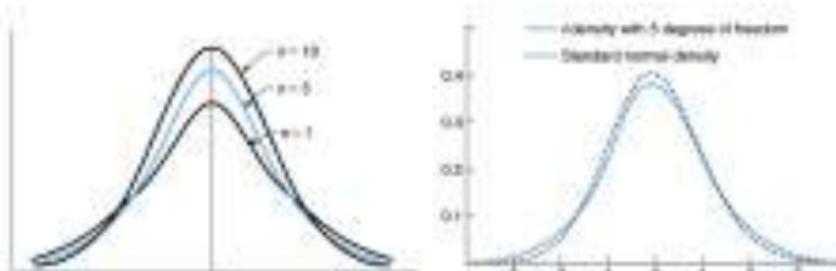
$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right]^2$$

Distribution of test statistic (t-distribution)

If Z and χ^2_n are independent random variables, with Z having a standard normal distribution and χ^2_n having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{\chi^2_n/n}}$$

is said to have a t-distribution with n degrees of freedom. A graph of the density function of T_n is given in Figure 5.13 for $n = 1, 5$, and 10 .

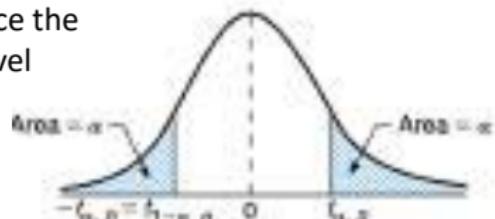


$$E[T_n] = 0, \quad n > 1$$
$$\text{Var}(T_n) = \frac{n}{n-2}, \quad n > 2$$

Distribution of test statistic

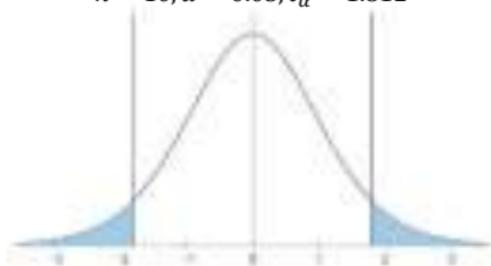
Since T depends only on n , we can compute in advance the cutoffs at which the area is less than a significance level

$$P\{T_n \geq t_{\alpha,n}\} = \alpha$$



$$P\{T_n \geq -t_{\alpha,n}\} = 1 - \alpha$$

$$n = 10, \alpha = 0.05, t_\alpha = 1.812$$



$$\text{Defining critical region} \quad T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s}, \text{ or with } n^{-1/2} \text{ df.}$$

Define

$$P_{\mu_0} \left\{ -t_{\alpha/2, n-1} \leq \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \leq t_{\alpha/2, n-1} \right\} = 1 - \alpha$$

$\frac{T}{S}$ is $\frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$

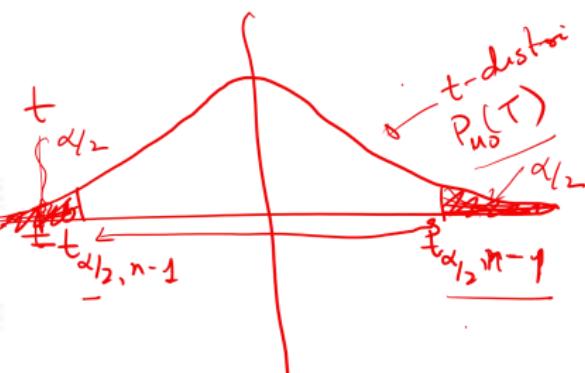
$t_{\alpha/2, n-1}$ is upper $\alpha/2$ percentile of t_{n-1}

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

is, when σ^2 is unknown, to

accept H_0 if $\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| \leq t_{\alpha/2, n-1}$

reject H_0 if $\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| > t_{\alpha/2, n-1}$



Example 8.3.g. Among a clinic's patients having blood cholesterol levels ranging in the medium to high range (at least 220 milliliters per deciliter of serum), volunteers were recruited to test a new drug designed to reduce blood cholesterol. A group of 50 volunteers was given the drug for 1 month and the changes in their blood cholesterol levels were noted. If the average change was a reduction of 14.8 with a sample standard deviation of 6.4, what conclusions can be drawn?

Solution. Let us start by testing the hypothesis that the change could be due solely to chance — that is, that the 50 changes constitute a normal sample with mean 0. Because the value of the t -statistic used to test the hypothesis that a normal mean is equal to 0 is

$$H_0: \mu = 0 ; H_1: \mu > 0$$

$$n = 50$$

$$t = \sqrt{n} \bar{X}/S = \sqrt{50} 14.8/6.4 = 16.352$$

P-value $P(T > 16.352) = t$ -distribution with 49 degrees of freedom

$$P_{H_0}(T > 16.352) = 5.6 \times 10^{-22} \leftarrow \text{very small}$$

Reject null hypothesis. Drug was effective.

Example 8.3.b. A public health official claims that the mean home water use is 350 gallons a day. To verify this claim, a study of 20 randomly selected homes was instigated with the result that the average daily water uses of these 20 homes were as follows:

340	344	362	375
356	386	354	364
332	402	340	355
362	322	372	324
318	360	338	370

Do the data contradict the official's claim?

Solution. To determine if the data contradict the official's claim, we need to test

$$H_0: \mu = 350 \quad \text{versus} \quad H_1: \mu \neq 350$$

This can be accomplished by noting first that the sample mean and sample standard deviation of the preceding data set are

$$\bar{Y} = 353.8, \quad S = 21.8478$$

Thus, the value of the test statistic is

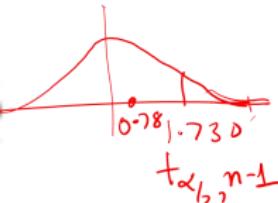
$$T = \frac{\sqrt{20}(3.8)}{21.8478} = .7778$$

$$T = \frac{\sqrt{20}(3.8)}{21.8478} = .7778$$

Because this is less than $t_{0.95,19} = 1.730$, the null hypothesis is accepted at the 10 percent level of significance. Indeed, the p -value of the test data is

$$\underline{p\text{-value}} = P[|T_{19}| > .7778] = 2P[T_{19} > .7778] = .4462$$

indicating that the null hypothesis would be accepted at any reasonable significance level, and thus that the data are not inconsistent with the claim of the health official. ■



Example 8.3.1. The manufacturer of a new fiberglass tire claims that its average life will be at least 40,000 miles. To verify this claim a sample of 12 tires is tested, with their lifetimes (in 1000s of miles) being as follows:

Tire	1	2	3	4	5	6	7	8	9	10	11	12
Life	36.1	40.2	33.8	38.5	42	35.8	37	40	36.0	37.2	33	36

Test the manufacturer's claim at the 5 percent level of significance.

Solution. To determine whether the foregoing data are consistent with the hypothesis that the mean life is at least 40,000 miles, we will test

$$H_0: \mu \geq 40,000 \quad \text{versus} \quad H_1: \mu < 40,000$$

A computation gives that

$$\bar{Y} = 37.2833, \quad S = 2.7319$$

and so the value of the test statistic is

$$T = \frac{\sqrt{12}(37.2833 - 40)}{2.7319} = -3.4448$$

Since this is less than $-t_{0.05,11} = -1.796$, the null hypothesis is rejected at the 5 percent level of significance. Indeed, the *p*-value of the test data is

$$p\text{-value} = P[T_{11} < -3.4448] = P[T_{11} > 3.4448] = .0027$$

indicating that the manufacturer's claim would be rejected at any significance level greater than .005. ■

Practice question

Parametric hypothesis testing for two populations

Sections: 8.4.2 and 8.4.4 of Ross Text-book

Two-sample t-test

- We have two distributions $F_{\theta_x}(X)$ and $F_{\theta_y}(Y)$ with unknown parameters
- Hypothesis:
 - Whether the two parameters are equal.
 - Whether mean of one is greater than another.
- Data D: n samples of X, and m samples of Y

$$\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n \quad \bar{y}_1, \bar{y}_2, \dots, \bar{y}_m$$

Special case: normal distributions with equal but unknown variance

$$E_{\theta_x}(X) \sim N(\mu_x, \sigma^2) \quad E_{\theta_y}(Y) \sim N(\mu_y, \sigma^2)$$

σ^2 is unknown.

- Hypothesis: are their means equal?

$$H_0: \mu_x = \mu_y \quad \text{versus} \quad H_1: \mu_x \neq \mu_y$$

$$\Leftrightarrow H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$$

- T-statistic:

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

$$S_y^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1}$$

Pooled sample variance

$$\Rightarrow S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$$

Distribution of test statistic T

- We can show that T follows a t-distribution with $n+m-2$ degrees of freedom.

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2(1/n + 1/m)}} \sim t_{n+m-2}$$

- Proof: [Section 7.4 of text book]

Example:

Twenty-two volunteers at a cold research institute caught a cold after having been exposed to various cold viruses. A random selection of 10 of these volunteers was given tablets containing 1 gram of vitamin C. These tablets were taken four times a day. The control group consisting of the other 12 volunteers was given placebo tablets that looked and tasted exactly the same as the vitamin C tablets. This was continued for each volunteer until a doctor, who did not know if the volunteer was receiving the vitamin C or the placebo tablets, decided that the volunteer was no longer suffering from the cold. The length of time the cold lasted was then recorded. At the end of this experiment, the following data resulted.

At the end of this experiment, the following data resulted.

Treated with Vitamin C	Treated with Placebo
8.0	8.5
8.0	8.0
7.0	8.0
8.0	7.0
7.0	8.5
8.0	8.0
7.0	7.0
8.5	8.0
7.0	7.0
8.0	8.0
8.0	8.0
7.0	7.0

10 12
vit c placebo

Do the data listed prove that taking 4 grams daily of vitamin C reduces the mean length of time a cold lasts? At what level of significance?

Solution

- $H_0: \mu_C - \mu_P = 0, H_1: \mu_C - \mu_P < 0$
- $t = -1.8987, df = 20, p\text{-value} = 0.03606$
- Accept the hypothesis that vitamin-C reduces duration of cold at 5 percent significance level

Paired t-test

Paired t-test

- For the same instance i , we have values before a treatment X_i , and after a treatment Y_i
- Example:
 - Suppose we are interested in determining whether the installation of a certain antipollution device will affect a car's mileage. To test this, a collection of n cars that do not have this device are gathered. Each car's mileage per gallon is then determined both before and after the device is installed.
 - How can we test the hypothesis that the antipollution control has no effect on gas consumption? The data can be described by the n pairs $(X_i, Y_i), i = 1, \dots, n$, where X_i is the gas consumption of the i th car before installation of the pollution control device, and Y_i of the same car after installation.

T-statistic for this test

- Assume that $W_i = \underline{X_i} - \underline{Y_i}$ is Gaussian with unknown mean and variance.
- Hypothesis to test:

$$H_0: \mu_w = 0 \quad \text{versus} \quad H_1: \mu_w \neq 0$$

- Test statistic for Gaussian with unknown variance as discussed earlier is

$$T = \frac{\sqrt{n} \bar{W}}{S_w} \quad S_w^2 = \frac{1}{n} \left(\sum_{i=1}^n (X_i - Y_i) - \bar{W} \right)^2$$

- Accept/reject decision using critical regions

accepting H_0 if $-t_{\alpha/2, n-1} < \frac{\sqrt{n} \bar{W}}{S_w} < t_{\alpha/2, n-1}$

rejecting H_0 otherwise

W

Example 8.4.c. An industrial safety program was recently instituted in the computer chip industry. The average weekly loss (averaged over 1 month) in labor-hours due to accidents in 10 similar plants both before and after the program are as follows:

Test statistic with difference array:

Plant	Before	After	$\Delta = \text{Before} - \text{After}$
1	30.5	23	-7.5
2	39.0	31	-8
3	24.5	27	-2.5
4	32	38.5	-6.5
5	36	14.5	-21.5
6	30	15.5	-14.5
7	33.5	34.5	1
8	25.5	31	-5.5
9	39	23.5	-15.5
10	38	16.5	-21.5

> $d = c(-7.5, 2.5, -2.5, -3.5, -1.5, 5, 1, -4.5, -4.5, -1.5)$

> $v = \text{sqrt}(10/\text{var}(d)) * \text{mean}(d)$

> v

[1] -2.265949 = T

> $pT(v, 9)$

[1] 0.02484552

Thus, $v = -2.265949$, with resulting

$p\text{-value} = P(T_9 \leq -2.265949) = 0.02484552$

Hypothesis test in Bernoulli population (Section 8.6 of Ross Book)

- The distribution changes to Bernoulli. $p(x) = p^x (1-p)^{1-x} \quad x \in \{0, 1\}$
- Hypothesis: $H_0: p = p_0$ versus $H_1: p > p_0$
- Data samples of size n will consist of 1s and 0s. x_1, \dots, x_n
- T-statistic: number of 1s in D.
 $T = \sum_{i=1}^n x_i$; Let t denote observed constant.
- Distribution of T-statistic under a parameter p_0 will be Binomial(n, p_0)
 $P_{p_0}(T) \sim \text{Binomial}(n, p_0)$
- Calculate p-value of this distribution

$$P\{\text{Binomial}(n, p_0) \geq t\} = \sum_{i=t}^n \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Example 8.6.a. A computer chip manufacturer claims that no more than 2 percent of the chips it sends out are defective. An electronics company, impressed with this claim, has purchased a large quantity of such chips. To determine if the manufacturer's claim can be taken literally, the company has decided to test a sample of 300 of these chips. If 10 of these 300 chips are found to be defective, should the manufacturer's claim be rejected?

Solution. Let us test the claim at the 5 percent level of significance. To see if rejection is called for, we need to compute the probability that the sample of size 300 would have resulted in 10 or more defectives when p is equal to .02. (That is, we compute the p -value.) If this probability is less than or equal to .05, then the manufacturer's claim should be rejected. Now

$$\begin{aligned}P_{0.05}(t \geq 10) &= 1 - P_{0.05}(t \leq 9) \\&= 1 - \text{pbinom}(9, 300, .02) \\&= 0.08183807\end{aligned}$$

practise questions

and so the manufacturer's claim cannot be rejected at the 5 percent level of significance. ■

Example 8.6.b. In an attempt to show that proofreader A is superior to proofreader B, both proofreaders were given the same manuscript to read. If proofreader A found 28 errors, and proofreader B found 18, with 10 of these errors being found by both, can we conclude that A is the superior proofreader?

Solution. To begin note that A found 18 errors that B missed, and that B found 8 that A missed. Hence, a total of 26 errors were found by just a single proofreader. Now, if A and B were equally incompetent then they would be equally likely to be the sole-finder of an error found by just one of them. Consequently, if A and B were equally competent then each of the 26 singly found errors would have been found by A with probability 1/2. Hence, to establish that A is the superior proofreader the result of 18 successes in 26 trials must be strong enough to reject the null hypothesis when testing

$$H_0: p \leq 1/2 \text{ versus } H_1: p > 1/2$$

where p is a Bernoulli probability that a trial is a success. Because the resultant p -value for the data cited is

$$p\text{-value} = P(\text{Bin}(26, .5) \geq 18) = 0.03776$$

the null hypothesis would be rejected at the 5 percent level of significance, thus enabling one to conclude (at that level of significance) that A is the superior proofreader. ■

practise questions

Lecture 35-NonParametricHypothesisTesting.pdf

Non-parametric tests

12.2, 12.4 of Ross Textbook

Non-parametric tests

- We make no assumptions of the form of the distribution function unlike previous cases where we assume Normal or Binomial.
- Generically denote distribution as $F(X)$. Note form of $F(X)$ is not known
- Possible hypothesis that can be tested in such cases
 - What is the median of $F(X)$?
 - Sign test
 - Is the distribution around the median similar
 - Sign rank test
 - Given samples of two distributions: Are they likely to be from the same or different distributions?
 - Two sample test

Sign test

Let X_1, \dots, X_n denote a sample from a continuous distribution F and suppose that we are interested in testing the hypothesis that the median of F , call it m , is equal to a specified value m_0 . That is, consider a test of

$$H_0 : m = m_0 \quad \text{versus} \quad H_1 : m \neq m_0$$

where m is such that $F(m) = .5$.

$$T = \frac{\#\text{ of } X_i\text{'s less than } m_0}{n} \quad \hat{m} = \text{median of } (X_1, \dots, X_n)$$
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$$

$$P_{H_0}(T) \sim \text{Binomial}(n, \frac{1}{2}) \quad T = \hat{F}(m_0)$$

T-statistic

- $T = \text{sum of the sign of } m_0 - X_i$

$$T = \sum_i I_i$$

$$I_i = \begin{cases} 1 & \text{if } X_i < m_0 \\ 0 & \text{if } X_i \geq m_0 \end{cases}$$

$$P(I_i) \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

- What is the distribution of T under the null hypothesis?

$$P_{H_0}(T) \sim \text{Binomial}(n, \frac{1}{2})$$

$$T = \sum_{j=1}^{n/2} \left(|x_j - m_0| + |x_{n-j} - m_0| \right) \quad P(T)$$


Example 12.2.b. A financial institution has decided to open an office in a certain community if it can be established that the median annual income of families in the community is greater than \$90,000. To obtain information, a random sample of 80 families was chosen, and the family incomes determined. If 28 of these families had annual incomes below and 52 had annual incomes above \$90,000, is this significant enough to establish, say, at the 5 percent level of significance, that the median annual income in the community is greater than \$90,000?

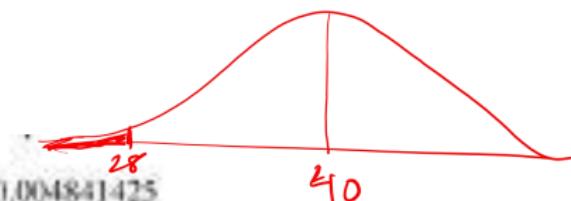
$$m_0 = 90k$$

$$n = 80$$

$$T = 28$$

Solution. We need to see if the data are sufficient to enable us to reject the null hypothesis when testing

$$H_0 : m \leq 90 \quad \text{versus} \quad H_1 : m > 90$$



$$p\text{-value} = P(\text{Bin}(80, 1/2) \leq 28) = \underline{\text{pbinnom}}(28, 80, 1/2) = \underline{0.004841425}$$

and so the null hypothesis that the median income is less than or equal to \$90,000 is rejected. ■

Signed rank test

Given n sample X_1, \dots, X_n from unknown distribution F , we are interested in the hypothesis that F is symmetric about a given median m_0 , that is,

- $H_0: P(X > m_0 + a) = P(X < m_0 - a)$, for all a

Let $Y_i = |X_i - m_0|$, $i = 1, \dots, n$ and rank (that is, order) the absolute values $|Y_1|, |Y_2|, \dots, |Y_n|$. Set, for $j = 1, \dots, n$,

$$I_j = \begin{cases} 1 & \text{if the } j\text{th smallest value comes from a data value that is smaller} \\ & \text{than } m_0 \\ 0 & \text{otherwise} \end{cases}$$

test statistic

$$T = \sum_{j=1}^n j I_j$$

Example 12.3.a. If $n = 4$, $m_0 = 2$, and the data values are $X_1 = 4.2$, $X_2 = 1.8$, $X_3 = 5.3$, $X_4 = 1.7$, then the rankings of $|X_i - 2|$ are $.2, .3, 2.2, 3.3$. Since the first of these values — namely, $.2$ — comes from the data point X_2 , which is less than 2, it follows that $I_1 = 1$. Similarly, $I_2 = 1$, and I_3 and I_4 equal 0. Hence, the value of the test statistic is $T = 1 + 2 = 3$. ■

X_i	4.2	1.8	5.3	1.7		$P_{H_0}(T \leq 3)$
	-2	-2	-2	-2		
	<u>2.2</u>	<u>-0.2</u>	<u>3.3</u>	<u>-0.3</u>		
y_i	2.2	-0.2	3.3	-0.3		
$ Y_i $	6.2	0.3	2.2	3.3		
	1	1	0	0		
	=					

Distribution of test statistic $P_{H_0}(T)$ under the null hypothesis?

Expected value and variance of T under H_0
Probability that the j^{th} absolute difference is from an
 smallest

$$P(I_j = 1) = \frac{1}{2} = P(I_j = 0), \quad j = 1, \dots, n \quad E[I_j] = \frac{1}{2}, \quad \text{Var}(I_j) = \frac{1}{4} \quad x_k < m_0.$$

Hence, we can conclude that under H_0 ,

$$\begin{aligned} E[T] &= E\left[\sum_{j=1}^n j I_j\right] \\ &= \sum_{j=1}^n \frac{j}{2} = \frac{n(n+1)}{4} \end{aligned}$$

$$\begin{aligned} \text{Var}(T) &= \text{Var}\left(\sum_{j=1}^n j I_j\right) \\ &= \sum_{j=1}^n j^2 \text{Var}(I_j) \\ &= \sum_{j=1}^n \frac{j^2}{4} = \frac{n(n+1)(2n+1)}{24} \end{aligned}$$

$P_{H_0}(T)$ = approximately normal for large n with mean and variance as above. But we can do better..

An exact computation of probability $P_{H_0}(T)$ recursively

$$\begin{aligned} P_k(i) &= P_{H_0} \left[\sum_{j=1}^k I_{I_j \leq i} \right] \quad P_{H_0}(T \leq i) = P_{H_0}(\\ &= P_{H_0} \left[\sum_{j=1}^k I_{I_j \leq i} | I_k = 1 \right] P_{H_0}(I_k = 1) \\ &\quad + P_{H_0} \left[\sum_{j=1}^k I_{I_j \leq i} | I_k = 0 \right] P_{H_0}(I_k = 0) \\ &= P_{H_0} \left[\sum_{j=1}^{k-1} I_{I_j \leq i-k} | I_k = 1 \right] P_{H_0}(I_k = 1) \\ &\quad + P_{H_0} \left[\sum_{j=1}^{k-1} I_{I_j \leq i} | I_k = 0 \right] P_{H_0}(I_k = 0) \\ &= P_{H_0} \left[\sum_{j=1}^{k-1} I_{I_j \leq i-k} \right] P_{H_0}(I_k = 1) + P_{H_0} \left[\sum_{j=1}^{k-1} I_{I_j \leq i} \right] P_{H_0}(I_k = 0) \end{aligned}$$

Continued..

$$\underline{P_{H_0}(I_k = 1)} = P_{H_0}(I_k = 0) = \frac{1}{2}$$

$$P_k(i) = P_{k-1}(i-k) P\{I_k = 1\} + P_{k-1}(i) P\{I_k = 0\}$$

we see that

$$\underline{P_k(i) = \frac{1}{2} P_{k-1}(i-k) + \frac{1}{2} P_{k-1}(i)}$$

Base Case:

$$P_1(i) = \begin{cases} 0 & i < 0 \\ \frac{1}{2} & i = 0 \\ 1 & i \geq 1 \end{cases}$$

$$P(I_1 < 0) = 0$$

$$P_1(i) = P_{H_0}(T \leq i)$$

$$P(I_1 \leq 0) = \frac{1}{2}$$

$$P(I_1 \leq 1) = 1$$

/

Example

$$\begin{aligned} \text{Compute: } P_4(3) &= \frac{1}{2} P_3(-1) + \frac{1}{2} P_3(3) \\ \left(= P_{H_0} \left(\sum_{j=1}^4 j I_j \leq 3 \right) \right) &= 0 + \frac{1}{2} [P_2(0) + P_2(3)] \\ &\quad + \frac{1}{2} \left[P_1(-2) + P_1(0) + P_1(1) + P_1(3) \right] \end{aligned}$$

HW

- How to extend paired-t-test to the non-parametric case?

Are two distributions equal?

- Let F and G be two continuous distributions of unknown form
- Given
 - n samples X_1, \dots, X_n from F
 - m samples Y_1, \dots, Y_m from G
- Null hypothesis: $H_0: F = G$
- Test is called: Rank-sum test, Mann-Whitney test, Wilcoxon test

Rank order the $n+m$ items.

R_i = rank of the data value X_i

Test statistic:

$$T = \sum_{i=1}^n R_i$$

Example 12.4.a. An experiment designed to compare two treatments against corrosion yielded the following data in pieces of wire subjected to the two treatments.

Treatment 1 65.2, 67.1, 69.4*, 78.2, 74, 80.3

Treatment 2 59.4, 72.1, 68, 66.2, 58.5

(The data represent the maximum depth of pits in units of one thousandth of an inch.) The ordered values are 58.5, 59.4, 65.2*, 66.2, 67.1*, 68, 69.4*, 72.1, 74*, 78.2*, 80.3* with an asterisk noting that the data value was from sample 1. Hence, the value of the test statistic is $T = 3 + 5 + 7 + 9 + 10 + 11 = 45$. ■

Distribution of test-statistic under the null hypothesis $P_{H_0}(T)$

- Again we will compute recursively.
- Let $P(n, m, t) = P_{H_0}(T \leq t)$

Self-study

Either the last item in the rank is one of the $N X_i$ s, or it is one of the $M Y_j$ s. Under the null hypothesis, this probability:

$$P(N, M, K) = \frac{N}{N+M} P(N-1, M, K-N+M) + \frac{M}{N+M} P(N, M-1, K)$$

Starting with the boundary condition

$$P(1, 0, K) = \begin{cases} 0 & K \leq 0 \\ 1 & K > 0 \end{cases}, \quad P(0, 1, K) = \begin{cases} 0 & K < 0 \\ 1 & K \geq 0 \end{cases}$$

Example 12.4.b. Suppose we wanted to determine $P(2, 1, 3)$. We use Equation (12.4.3) as follows:

$$P(2, 1, 3) = \frac{2}{3}P(1, 1, 0) + \frac{1}{3}P(2, 0, 3)$$

and

$$P(1, 1, 0) = \frac{1}{2}P(0, 1, -2) + \frac{1}{2}P(1, 0, 0) = 0$$

$$\begin{aligned} P(2, 0, 3) &= P(1, 0, 1) \\ &= P(0, 0, 0) = 1 \end{aligned}$$

Example 12.4.a. An experiment designed to compare two treatments against corrosion yielded the following data in pieces of wire subjected to the two treatments.

Treatment 1 65.2, 67.1, 69.4, 78.2, 74, 80.3

Treatment 2 59.4, 72.1, 68, 66.2, 58.5

(The data represent the maximum depth of pits in units of one thousandth of an inch.) The ordered values are 58.5, 59.4, 65.2*, 66.2, 67.1*, 68, 69.4*, 72.1, 74*, 78.2*, 80.3* with an asterisk noting that the data value was from sample 1. Hence, the value of the test statistic is $T = 3 + 5 + 7 + 9 + 10 + 11 = 45$. ■

$$P(6,5,45) = \frac{6}{11} P(5,5,34) + \frac{5}{11} P(6,4,45) = \dots$$

Wilcoxon rank sum test

data: x and y

$W = 24$, p-value = 0.1255

Errors in Hypothesis testing

- Type-I error: Rejecting H_0 even when H_0 is true.
 - The probability with which it happens is called significant level α
- Type-II error: Accepting H_0 when it is false

Summary of hypothesis testing

- Follow this framework:

- Formulate null and alternative hypothesis
- Collect data
- Decide on test statistic
- Identify distribution of test statistic under null hypothesis
- Apply p-value or critical region test to accept or reject null hypothesis

- We applied this framework on

- Mean of Gaussian with unknown variance is μ_0
- Are means of two normal distributions with shared unknown variance same?
- Difference in means of two normal with unknown variance from paired observations

Summary..

- Parameter p of Bernoulli is p_0
- Non-parametric tests
 - Median is a given value
 - Distribution is symmetric around a median
 - Are two distributions equal

Topics not covered.

- Goodness of fit tests
- Test on sequences

Lecture 36-Robust statistics.pdf

Robust statistics

Reference material

- Hampel,F.R., Ronchetti,E.M., Rousseeuw, P.J., Stahel, W.A. Robust Statistics: the Approach based on Influence Functions.Wiley Series in Probability and Mathematical Statistics.,1986.

Motivation

- Real world data often contains outliers or extreme values
- Most methods discussed so far on inferring models or parameter values from data can be adversely affected by outliers
 - Example: estimates of mean and variance from data
 - Estimation of linear regression parameters
- Robust statistics attempts to fit models that are largely unaffected by outliers, and fit based on “majority” of normal data.
- Robust fits enable better detection of outliers, as values that deviate from the fitted model

Assumptions

- We assume that the majority of the observations satisfy a parametric model and we want to estimate the parameters of this model.

E.g. $x_i \sim N(\mu, \sigma^2)$

$\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$

- Moreover, we assume that some of the observations might not satisfy this model.
- We do NOT model the outlier generating process.
- We do NOT know the proportion of outliers in advance.

Example

The classical methods for estimating the parameters of the model may be affected by outliers.

Example. Location-scale model: $x_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$,

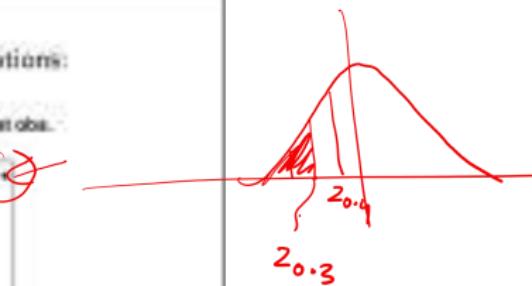
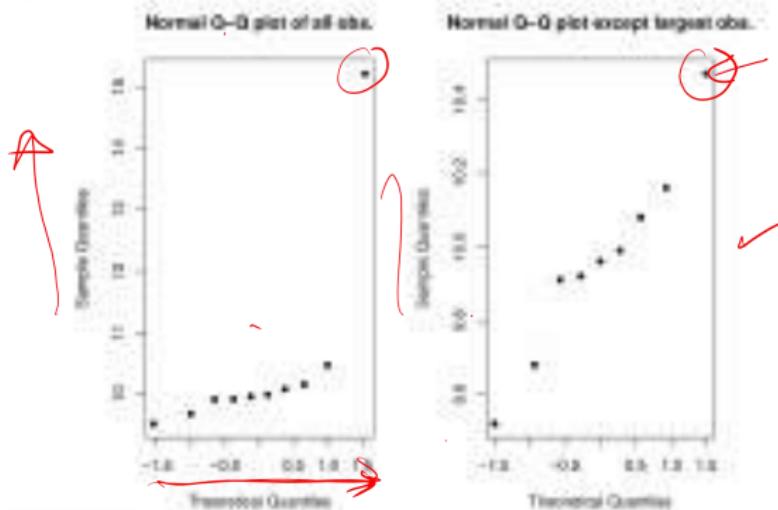
Data: $X_n = \{x_1, \dots, x_{10}\}$ are the natural logarithms of the annual incomes (in US dollars) of 10 people.

9.52	9.68	10.16	9.96	10.08
9.99	10.47	9.91	9.92	15.21

Example

The income of person 30 is much larger than the other values.

Normality cannot be rejected for the remaining ('regular') observations:



Classical versus robust estimators

Location:

Classical estimator: arithmetic mean

$$\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

Robust estimator: sample median

$$\hat{\mu} = \text{med}(X_n) = \begin{cases} x_{\lceil n/2 \rceil + 1} & \text{if } n \text{ is odd} \\ \frac{1}{2} (x_{\lfloor n/2 \rfloor} + x_{\lfloor n/2 \rfloor + 1}) & \text{if } n \text{ is even} \end{cases}$$

with $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ the ordered observations.

Classical versus robust estimators

Score

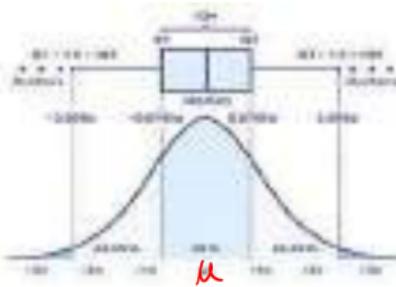
Classical estimator: sample standard deviation

$$\text{Sd} = \text{Sd}_{\text{dev}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Роль ван-актимов: интервью с Е. Гончар

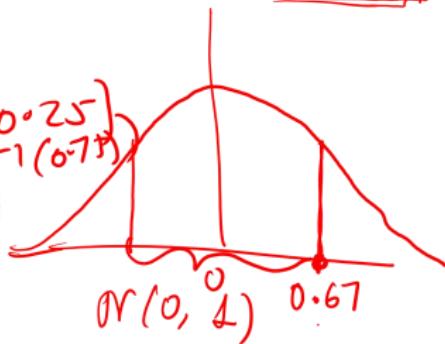
$$z = \text{IQF}(X_n) = \frac{1}{3\Phi^{-1}(0.75)} (x_{(n-(n/4)+1)} - x_{(n/4)})$$

$$\Phi^{-1}(0.75) - \Phi^{-1}(0.25) = \Phi^{-1}(0.75) - (-\Phi^{-1}(0.25))$$



For Gaussian $N(\mu, \sigma^2)$:

IQR can be shown to be $2\Phi^{-1}(0.75)$



Classical versus robust estimators

For the data of the example we obtain:

	the 9 regular observations	all 10 observations
\bar{x}_n	9.97	10.49
med	9.96	9.98
Stdsv _n	0.27	1.68
IQRN	0.13	0.17

- The classical estimators are highly influenced by the outlier
- The robust estimators are less influenced by the outlier
- The robust estimate computed from all observations is comparable with the classical estimate applied to the non-outlying data.

Classical versus robust estimators

→ Robustness: being less influenced by outliers

→ Efficiency: being precise at uncontaminated data

Robust estimators aim to combine high robustness with high efficiency

Outliers

An outlier is an observation that deviates from the fit suggested by the majority of the observations.

The usual standardized values (z-scores, standardized residuals) are:

$$r_i = \frac{x_i - \bar{x}_n}{\text{Stdev}_{ii}} \quad r_i \sim N(0, 1)$$

Classical rule: if $|r_i| > 3$, then observation x_i is flagged as an outlier.

Here: $|r_{16}| = 2.8 \rightarrow ?$

Outlier detection based on robust estimates:

$$r_i = \frac{x_i - \text{med}(X_n)}{\text{IQRN}(X_n)}$$

Here: $|r_{16}| = 31.0 \rightarrow$ very pronounced outlier!

MASKING is when actual outliers are not detected.

SWAMPING is when regular observations are flagged as outliers.

Characterizing robustness

- Breakdown value
- Sensitivity curve
- Influence function



Breakdown value

Breakdown value (breakdown point) of a location estimator

A data set with n observations is given. If the estimator stays in a fixed bounded set even if we replace any $m - 1$ of the observations by any outliers, and this is no longer true for replacing any m observations by outliers, then we say that:

the breakdown value of the estimator at that data set is m/n

Notation:

$$\varepsilon_n^*(T_n, X_n) = \underline{m/n}$$

Typically the breakdown value does not depend much on the data set. Often it is a fixed constant as long as the original data set satisfies some weak condition, such as the absence of ties.

Breakdown value

Example: $X_n = \{x_1, \dots, x_n\}$ univariate data, $T_n(X_n) = \text{med}(X_n)$.

Assume n odd, then $T_n = x_{((n+1)/2)}$.

- Replace $\frac{n-1}{2}$ observations by any value, yielding a set X_n^*
 $\Rightarrow T_n(X_n^*)$ always belongs to $[x_{(1)}, x_{(n)}]$, hence $T_n(X_n^*)$ is bounded.
- Replace $\frac{n+1}{2}$ observations by $+\infty$, then $T_n(X_n^*) = +\infty$.
- More precisely, if we replace $\frac{n+1}{2}$ observations by $x_{(n)} + n$,
 where n is any positive real number, then $T_n(X_n^*) = x_{(n)} + n$.
 Since we can choose n arbitrarily large, $T_n(X_n^*)$ cannot be bounded.

For n odd or even, the (finite-sample) breakdown value ε_n^* of T_n is

$$\varepsilon_n^*(T_n, X_n) = \frac{1}{n} \left[\frac{n+1}{2} \right] \approx 50\% .$$

Note that for $n \rightarrow \infty$ the finite-sample breakdown value tends to $\varepsilon^* = 50\%$
 (which we call the asymptotic breakdown value).

For instance, the arithmetic mean satisfies $\varepsilon_n^*(T_n, X_n) = \frac{1}{n} \rightarrow \varepsilon^* = 0\%$.

Sensitivity curve

The sensitivity curve measures the effect of a single outlier on the estimator.

Assume we have $n - 1$ fixed observations $X_{n-1} = \{x_1, x_2, \dots, x_{n-1}\}$.

Now let us see what happens if we add an additional observation equal to x , where x can be any real number.

Sensitivity curve

$$SC(x, T_n, X_{n-1}) = \frac{T_n(x_1, \dots, x_{n-1}, x) - T_{n-1}(x_1, \dots, x_{n-1})}{1/n}$$

Example: for the arithmetic mean $T_n = \bar{X}_n$ we find $SC(x, T_n, X_{n-1}) = x - \bar{x}_{n-1}$.

Note that the sensitivity curve depends strongly on the data set X_{n-1} .

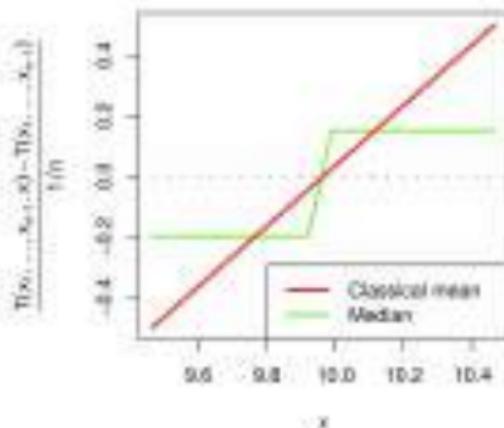
Sensitivity curve: example

Annual income data: let X_5 consist of the 9 "regular" observations.

9.52 9.66 9.81 9.92 9.96 9.99 10.00 10.16 10.47

mean: 9.97
med: 9.96

Sensitivity curve



Influence function

- The influence function is the asymptotic version of the sensitivity curve. It is computed for an estimator T at a certain distribution F , and does not depend on a specific data set.
- For this purpose, the estimator should be written as a function of a distribution F . For example, $T(F) = E_F[X]$ is the functional version of the sample mean, and $T(F) = F^{-1}(0.5)$ is the functional version of the sample median.
- The influence function measures how $T(F)$ changes when contamination is added in x . The contaminated distribution is written as

$$F_{\varepsilon, x} = (1 - \varepsilon)F + \varepsilon \Delta_x$$

for $\varepsilon > 0$, where Δ_x is the distribution that puts all its mass in x .

$$\Delta_x(x) = \begin{cases} 1 & \text{if } x=x \\ 0 & \text{otherwise} \end{cases}$$

Influence function

Influence function

$$\text{IF}(x, T, F) = \lim_{\varepsilon \rightarrow 0} \frac{T(F_{\varepsilon x}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon x})|_{\varepsilon=0}$$

Example: for the arithmetic mean $T(F) = E_F[X]$ at a distribution F with finite first moment:

$$\begin{aligned}\text{IF}(x, T, F) &= \frac{\partial}{\partial \varepsilon} E[(1-\varepsilon)F + \varepsilon \Delta_x]|_{\varepsilon=0} \\ &= \frac{\partial}{\partial \varepsilon} [\varepsilon x + (1-\varepsilon)T(F)]|_{\varepsilon=0} = x - T(F)\end{aligned}$$

At the standard normal distribution $F = \Phi$ we find $\text{IF}(x, T, \Phi) = x$.

We prefer estimators that have a bounded influence function.

Other robust estimates of location

- Median

- **Trimmed mean:** ignore the m smallest and the $n - m$ largest observations and just take the average of the observations in between:

$$\hat{\mu}_{TM} = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)}$$

with $m = \lceil (n-1)/4 \rceil$ and $0 \leq \alpha < 0.5$.

For $\alpha = 0$ this is the mean, and for $\alpha \rightarrow 0.5$ this becomes the median.

- **Winsorized mean:** replace the m smallest observations by $x_{(m+1)}$ and the m largest observations by $x_{(n-m)}$. Then take the average

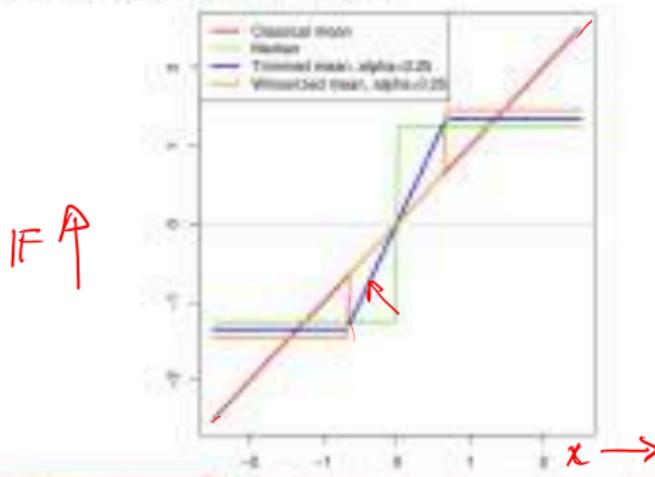
$$\hat{\mu}_{WM} = \frac{1}{n} \left(mx_{(m+1)} + \sum_{i=m+1}^{n-m} x_{(i)} + mx_{(n-m)} \right)$$

Robustness properties

Breakdown value: $\varepsilon_n^*(\text{med}) \rightarrow 0.5$; $\varepsilon_n^*(\hat{\mu}_{TM}) = \varepsilon_n^*(\hat{\mu}_{WM}) = (m+1)/n \rightarrow \alpha$.

Maxbias: For any ε , the median achieves the smallest maxbias among all location equivariant estimators.

Influence function at the normal model:



The pure scale model

The scale model assumes that the data are i.i.d. according to:

$$F_\sigma(x) = F\left(\frac{x}{\sigma}\right)$$

where $\sigma > 0$ is the unknown scale parameter. As before F is a continuous distribution with density f , but now

$$f_\sigma(x) = F'_\sigma(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right),$$

We say that a scale estimator S is Fisher-consistent at this model iff

$$S(F_\sigma) = \sigma \quad \text{for all } \sigma > 0.$$

ignore

Robust estimates of scale

Some explicit scale estimators:

- Standard deviation (Stdev) Not robust.
- Interquartile range

$$\text{IQR}(X_n) = \underline{x_{(n-1)n/4+1}} - \underline{x_{(n/4)}}$$

However, at $F_\sigma = N(0, \sigma^2)$ it holds that $\text{IQR}(F_\sigma) = 2\Phi^{-1}(0.75)\sigma \neq \sigma$.

Normalized IQR:

$$\text{IQRN}(X_n) = \frac{1}{\underline{2\Phi^{-1}(0.75)}} \text{IQR}(X_n)$$

The constant $1/2\Phi^{-1}(0.75) = 0.7413$ is a consistency factor.

When using software, it should be checked whether the consistency factor is included or not!

Explicit scale estimators

Estimators with 50% breakdown value:

• Median absolute deviation

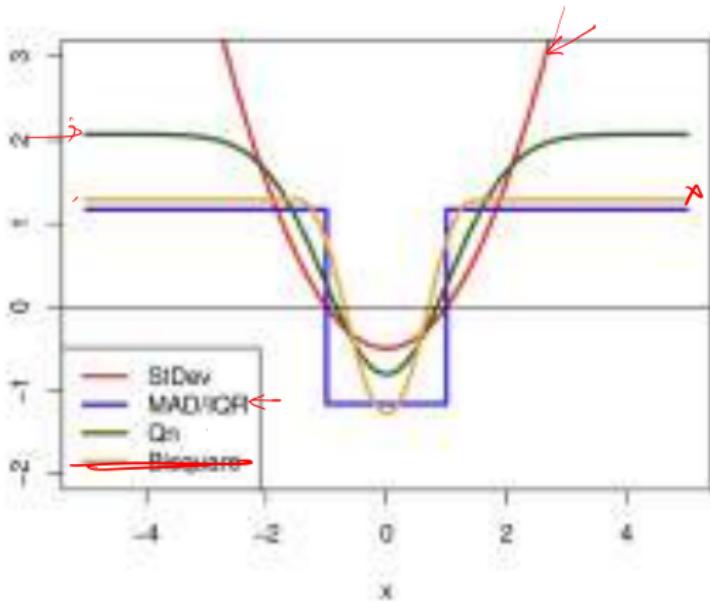
$$\text{MAD}(X_n) = \text{med}_i(|x_i - \text{med}(X_n)|)$$

At any symmetric sample it holds that IQR = 2 MAD.

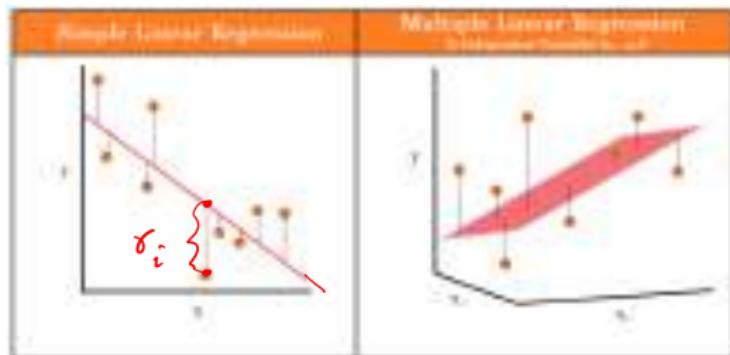
At the normal model we use the normalized version:

$$\text{MADN}(X_n) = \frac{1}{\Phi^{-1}(0.75)} \text{MAD}(X_n) \quad \text{MAD}(X_n) = 1.4826 \text{MAD}(X_n)$$

Influence function of various scale estimators



Robust regression



Reading material

- Primary
 - https://en.wikipedia.org/wiki/Robust_regression
- Additional
 - [T.1.1 - Robust Regression Methods | STAT 501](#)

Ordinary least square regression

$$f(Y | \underline{x_1}, \dots, \underline{x_k}) \sim N(\mu_x, \sigma^2), \quad \text{where } \mu_x = \beta_1 x_1 + \dots + \beta_k x_k + \beta_0$$

Training data D denoted as

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1 \dots n\}$$

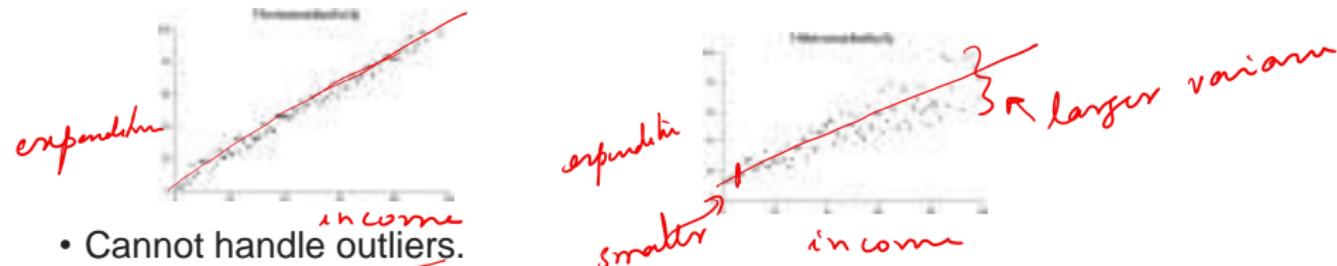
MLE training objective

$$\min_{\{\beta\}} \sum_i \frac{(y_i - \beta^\top x_i)^2}{\sigma^2}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$
$$x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$$

Limitations

- Assumes same variance for all examples (**homoscedastic**). Real-life data often does not satisfy this assumption (**heteroscedasticity**)
 - For example, the variance of expenditure is often larger for individuals with higher income than for individuals with lower incomes.



- Cannot handle outliers.
 - The least squares predictions are dragged towards the outliers
 - The variance of the estimates is artificially inflated, causing outliers to be masked.
 - In many situations, including some areas of geostatistics and medical statistics, it is precisely the outliers that are of interest.

Robust regression: modify the loss function

- Replace square loss by least absolute deviation

- $\min_{\{\beta\}} \sum_i |y_i - \beta x_i|$

- Solvable using a linear program.

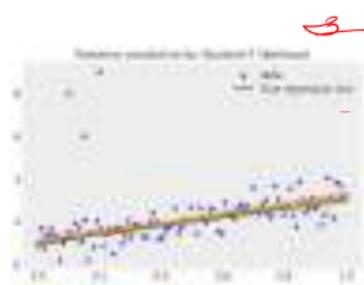
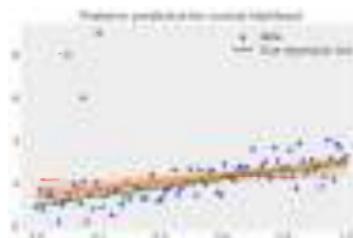
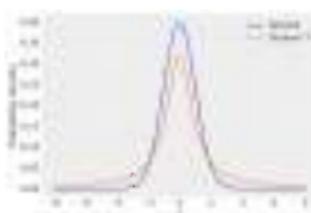
- Least trimmed square: Ignore $n-k$ largest residues during minimization

$$\min_{\beta} \sum_{\text{Bottom- } k} (|y_1 - \beta x_1| - \dots - |y_n - \beta x_n|)$$

Robust regression: parametric alternative

Replace the normal distribution:

- Choose a heavy-tailed distribution. A t-distribution with 4–6 degrees of freedom has been reported to be a good choice in various practical situations.



https://colab.research.google.com/github/pymc-devs/pymc-examples/blob/main/examples/generalized_linear_models/GLM-robust.ipynb#scrollTo=285a756b

Robust regression: continued...

- Choose a mixture of normal and outlier distribution --- majority of observations are from a specified normal distribution, but a small proportion are from a normal distribution with much higher variance.

$$e_i \sim (1 - \varepsilon)N(0, \sigma^2) + \varepsilon N(0, c\sigma^2).$$

Usefulness of robust regression

- If number of data points is large, the actual fitted model may not be different with robust methods, but robust estimation of the variance, can lead to better outlier detection.

