

Lecture_00-introduction	2
Lecture_01_02_03-descriptiveStatistics	11
Lecture_04-probability	81
Lecture_05-independence	103
Lecture_06-randomVars	126
Lecture_07-randomVars	150
Lecture_08-discreteRVs	169
Lecture_09-discreteRV	238
Lecture_10-poisson	250
Lecture_11-continuousRVs	284
Lecture_12-gaussian	296
Lecture_13-exponential	327
Lecture_14-MultipleRVs	339
Lecture_15-ParameterEstimation	362
Lecture_16-EvaluatingEstimators	382
Lecture_17-BayesianEstimates	390

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



State/UT-wise Funds Allocated under Samagra Shiksha during 2021-22

Created By: OGD Team

Created date : 6/4/2022

Data API

Data Set



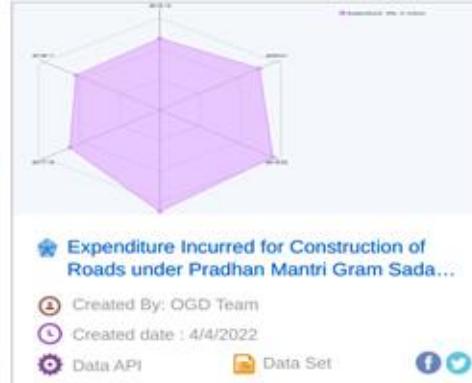
State-wise Total Allocation of Funds from State Disaster Risk Management Fund and other sources

Created By: OGD Team

Created date : 5/4/2022

Data API

Data Set



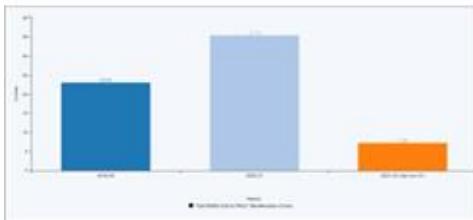
Expenditure Incurred for Construction of Roads under Pradhan Mantri Gram Sadak Yojana

Created By: OGD Team

Created date : 4/4/2022

Data API

Data Set



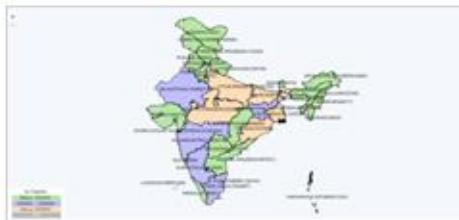
Year-wise LPG Refills Sold to PMUY Beneficiaries (14.2 kg & 5 kg) from 2019-2021

Created By: OGD Team

Created date : 4/4/2022

Data API

Data Set



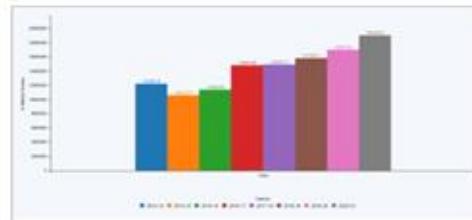
State/UT-wise Beneficiaries under National Social Assistance Programme (NSAP)

Created By: OGD Team

Created date : 1/4/2022

Data API

Data Set



Export of Major Chemicals from 2013-14 to 2020-21

Created By: OGD Team

Created date : 30/3/2022

Data API

Data Set

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.

Descriptive Statistics

Fall 2024

Instructor:

Sunita Sarawagi

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

variable / attribute

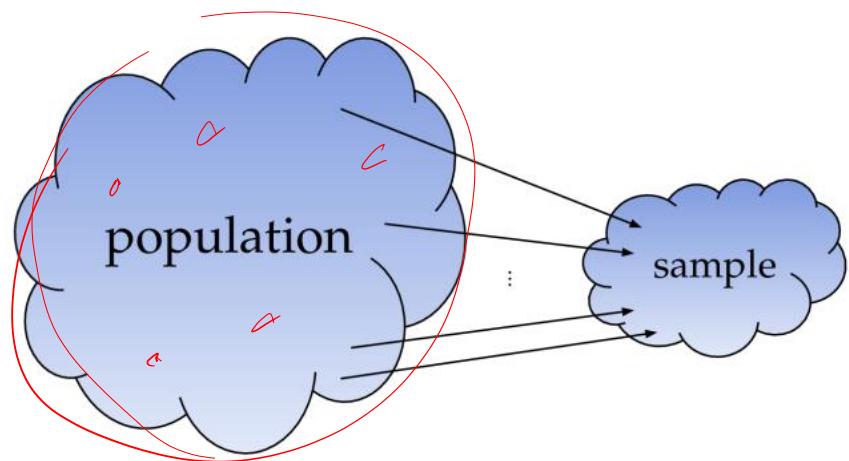
observation

Sample

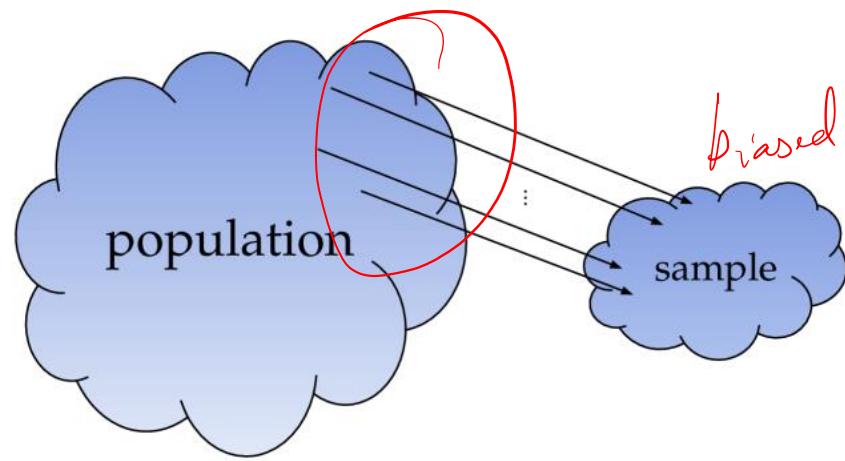
index	username	country	age	ezlvl	time	points	finished
0	mary	us	38	0	124.94	418	0
1	jane	ca	21	0	331.64	1149	1
2	emil	fr	52	1	324.61	1321	1
3	ivan	ca	50	1	39.51	226	0
4	hasan	tr	26	1	253.19	815	0
5	jordan	us	45	0	28.49	206	0
6	sanjay	ca	27	1	585.88	2344	1
7	lena	uk	23	0	408.76	1745	1
8	shuo	cn	24	1	194.77	1043	0
9	r0byn	us	59	0	255.55	1102	0
10	anna	pl	18	0	303.66	1209	1
11	joro	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

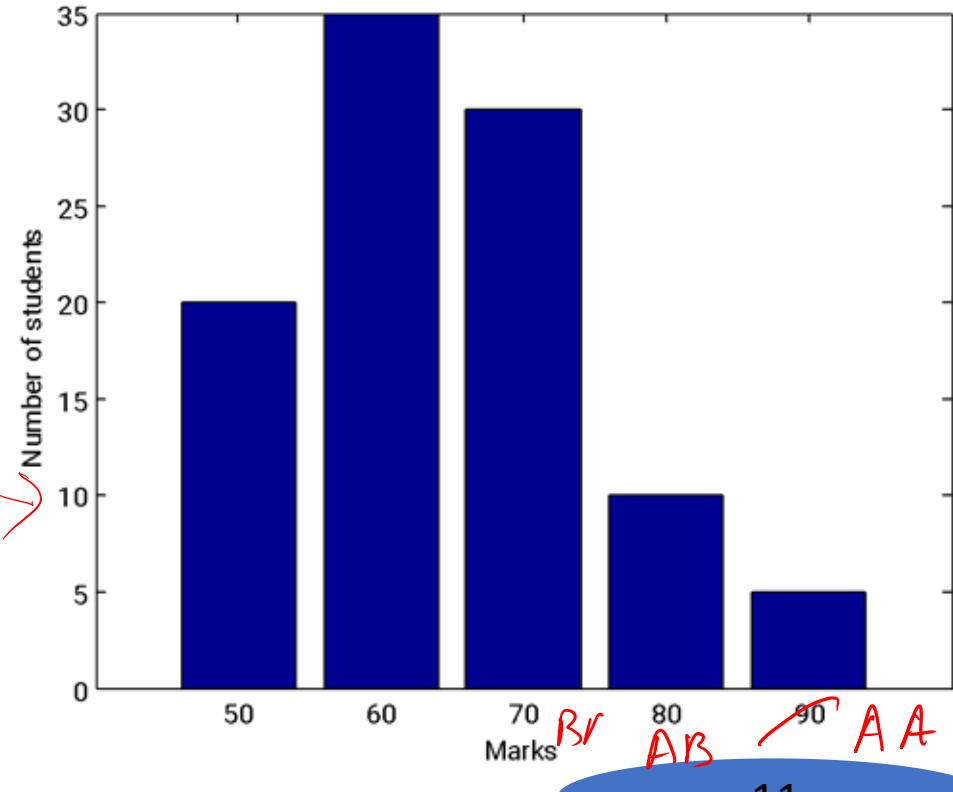
.

24 AA

Frequency Tables

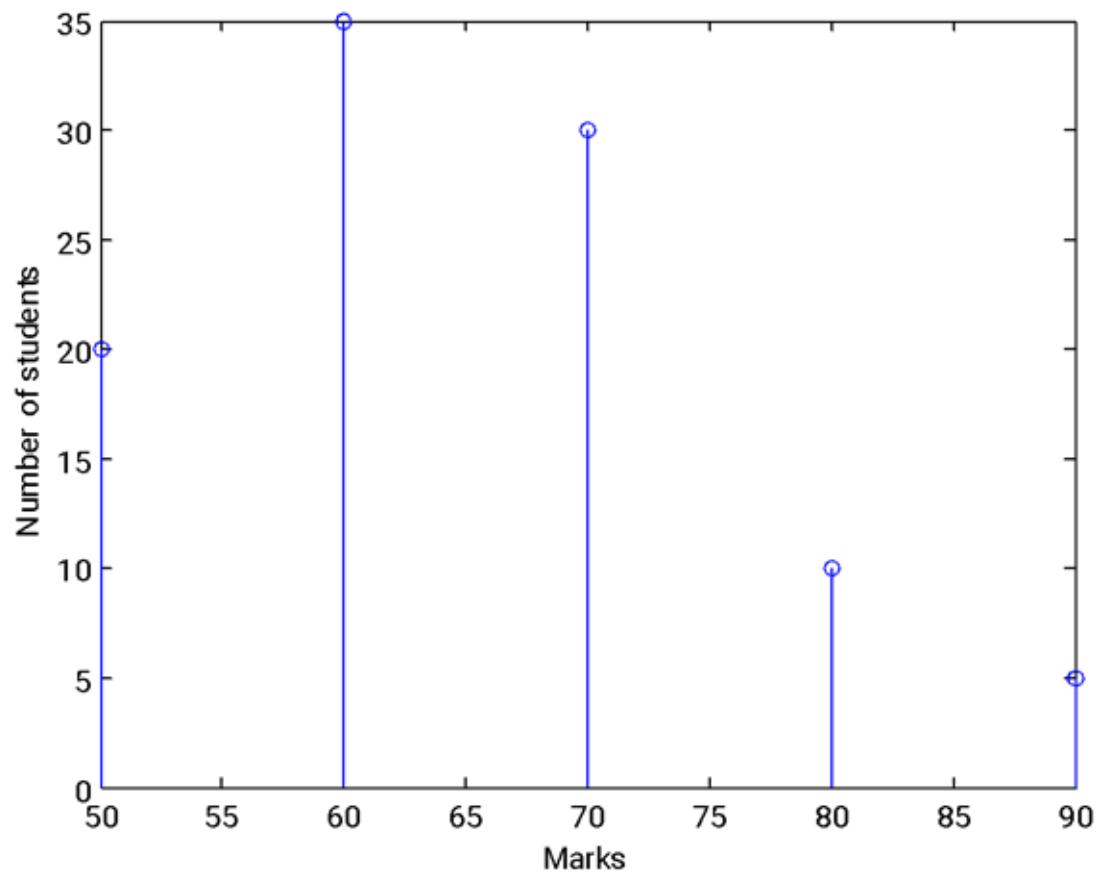
- The frequency table can be visualized using a **line graph** or a **bar graph** or a **frequency polygon**.

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



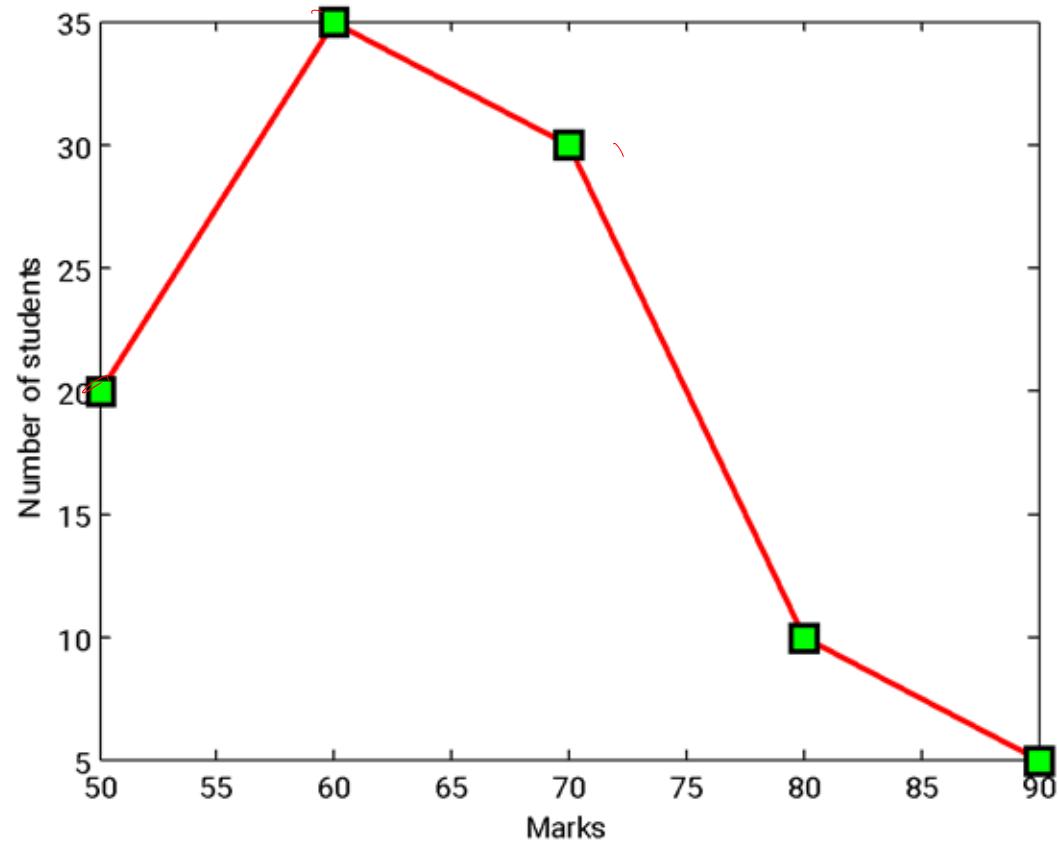
A **bar graph** plots the distinct data values on the X axis and their frequency on the Y axis by means of the height of a thick vertical bar!

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.

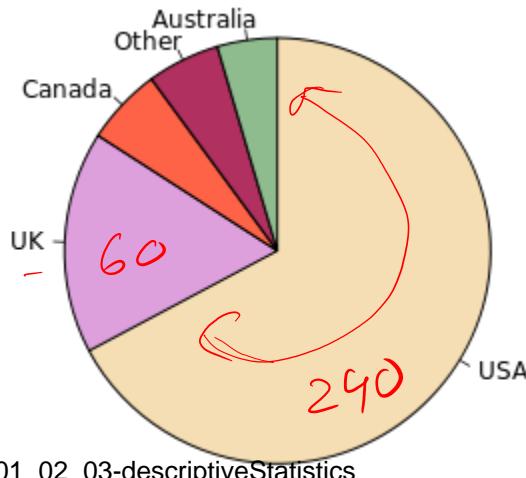
Relative frequency tables

- 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

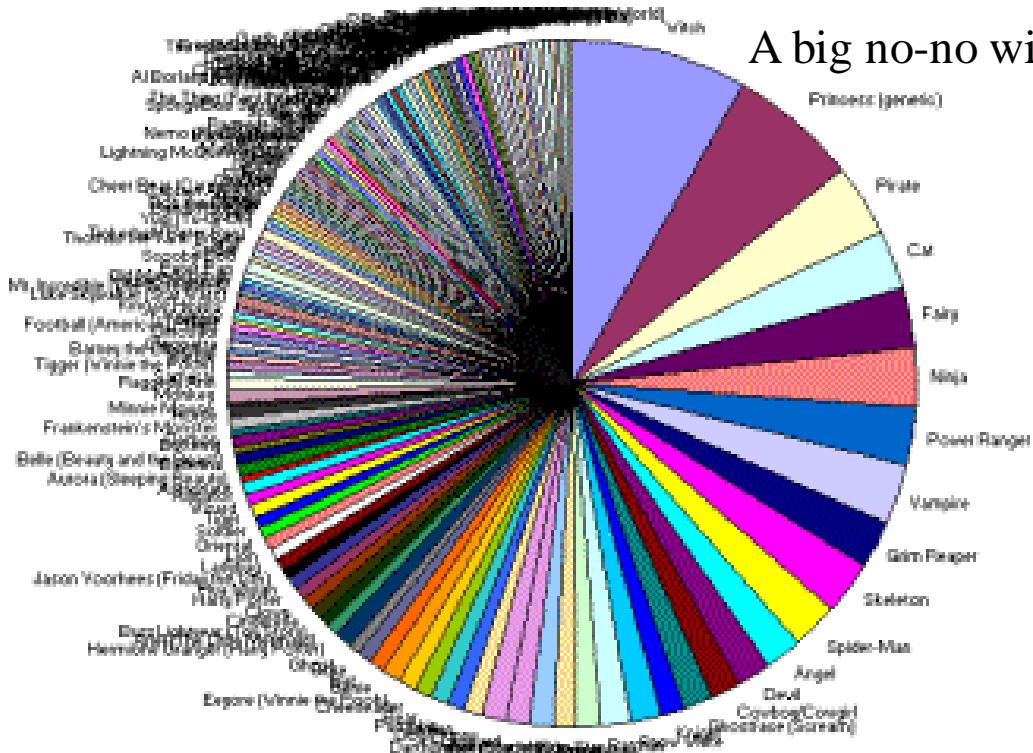
- For a small number of distinct data values which are non-numerical, one can use a **pie-chart** (it can also be used for numerical values).
- It consists of a circle divided into sectors corresponding to each data value.
- The area of each sector = relative frequency for that data value.



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

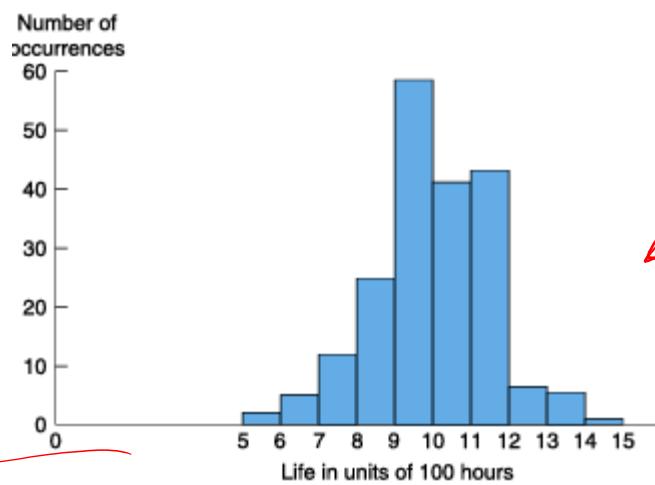


Table 2.4 A Class Frequency Table.

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	43
1200–1300	7
1300–1400	6
1400–1500	1

Dealing with continuous data

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

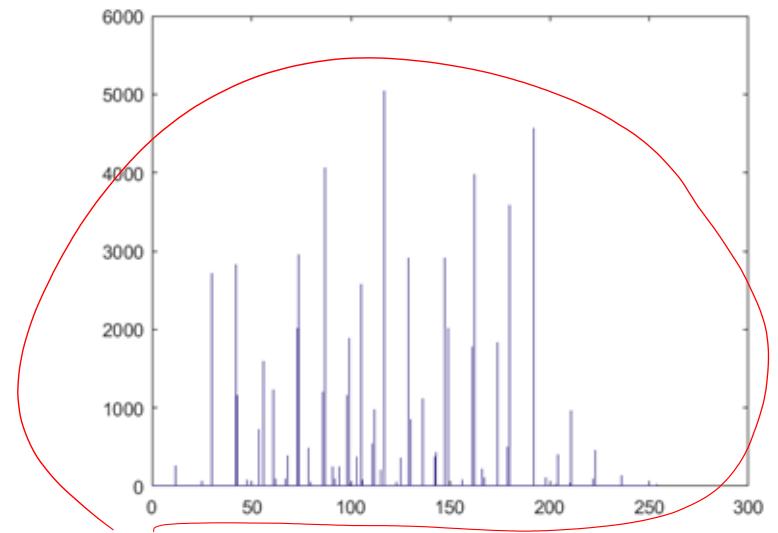
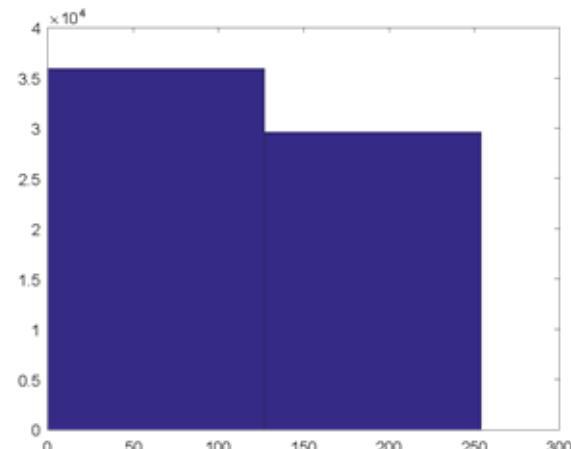
- Let the sample points be $\{x_i\}$, $1 \leq i \leq N$.
- Let there be some K ($K \ll N$) bins, where the j^{th} bin has interval $[a_j, b_j]$.
- Thus frequency f_j for the j^{th} bin is defined as follows:

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

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

The histogram binning problem

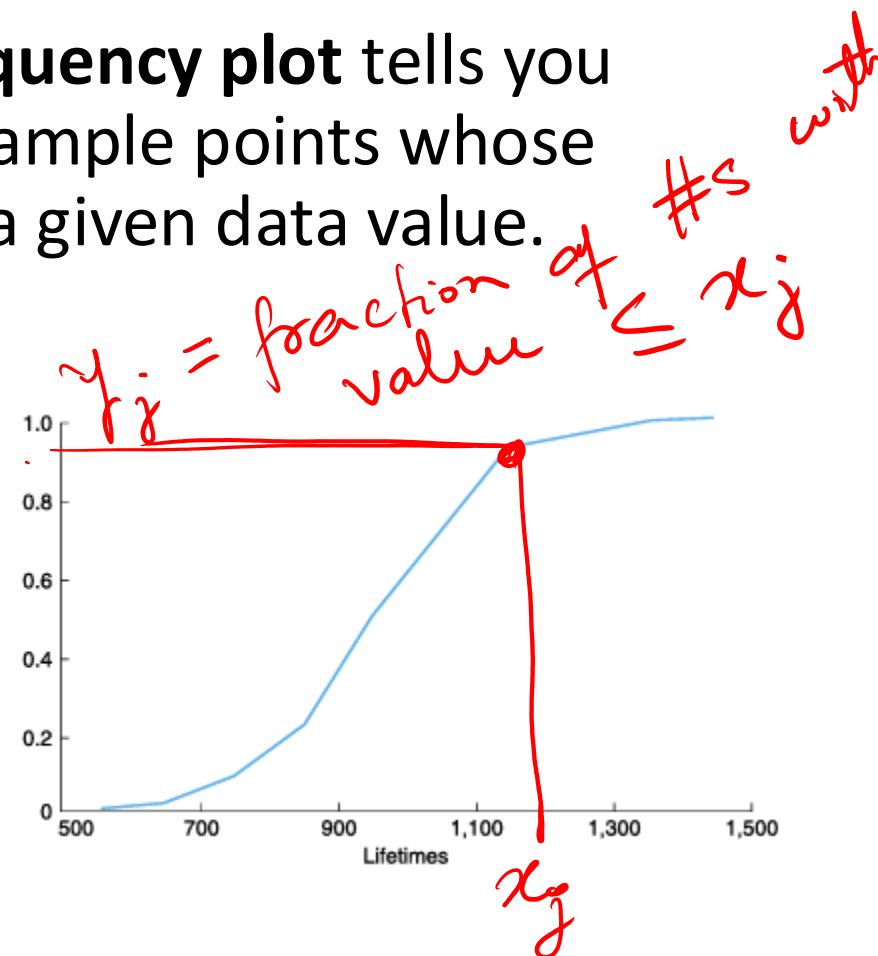
- If you have too few bins (each bin is very wide), there is very little idea you get about the data distribution from the histogram.
- Extreme: only one bin to represent whole data
- If you have many bins (all will be narrow), then there are very points falling into each bin. Again there is very little 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)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



Summarizing Data

08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	08
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	53	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	69	24	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	63	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	32	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
88	36	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	38	25	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	34	62	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	86	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	19	67	48

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

$$x_1 \leq x_2 \leq x_3 - - - \leq x_N$$

- 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.

$$x_1 \ x_2 - - - x_{\frac{N}{2}+1} \ x_{\frac{N}{2}+2} - - - x_N$$

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

$$3, 4, 10, 11, 13, 15 \quad N = 6$$

any value in-between 10 & 11.

Properties of the mean and median

→ $1, 2, 4, 5, 7 \quad q=10, b=1$

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

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

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

$$\begin{aligned}\frac{\partial_s G}{\partial y} = 0 \Rightarrow \sum_{i=1}^N \text{sign}(x_i - y) &= 0 && N \text{ is even} \\ \Rightarrow \text{equal } \# \text{ of } +1 &\neq -1 \Rightarrow y \text{ is median.} && 29\end{aligned}$$

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

N is odd.

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

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

,
:
:

⑮

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

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

$\min_y |x_{\frac{N}{2}} - y| = 0 \quad \text{if } y = x_{\frac{N}{2}}$
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.

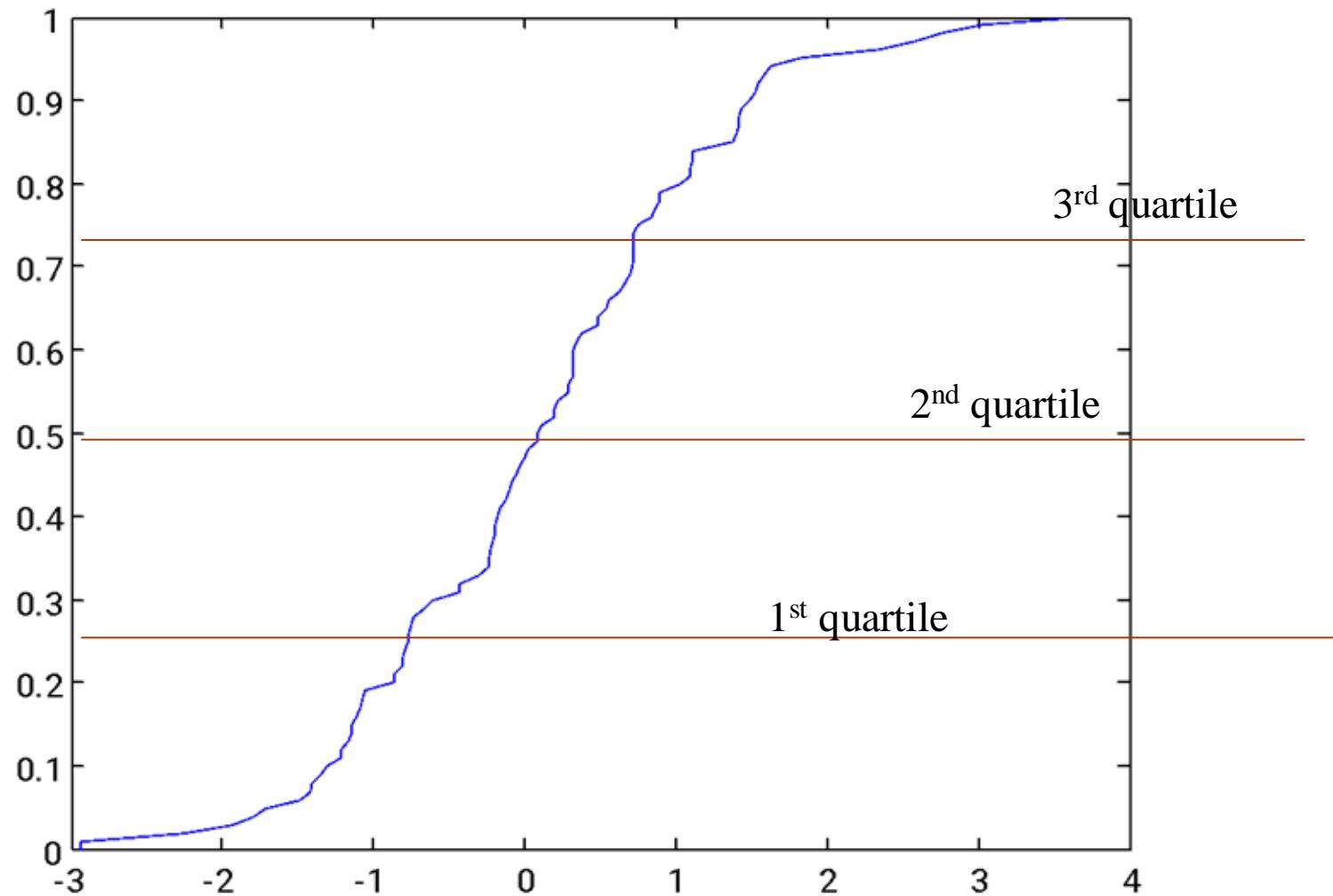
Robust statistics -

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.

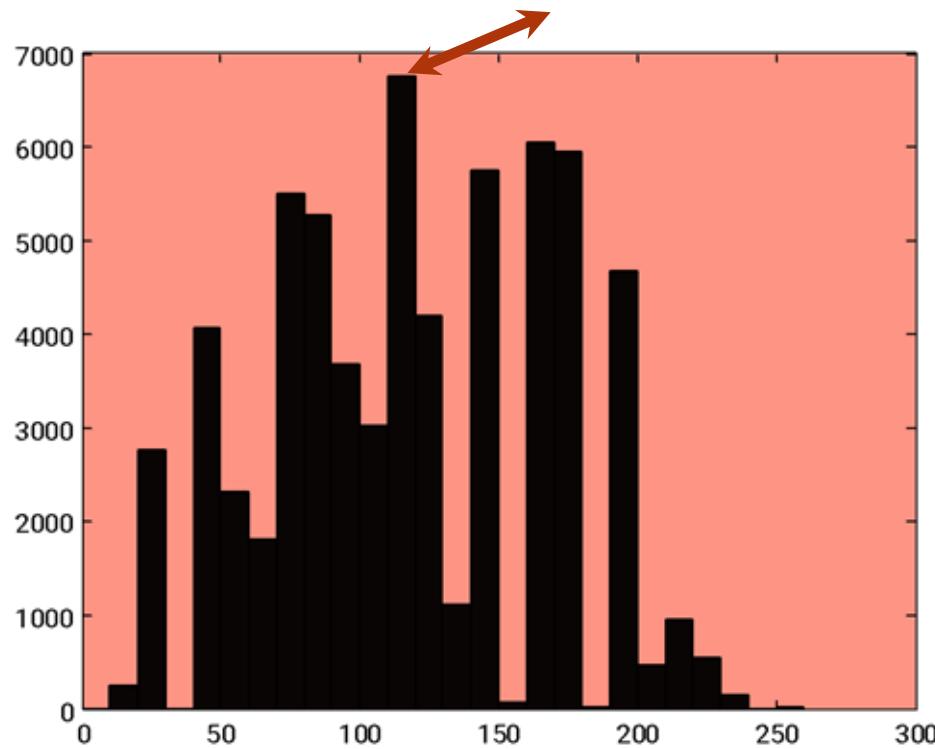
Quantiles

- 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?).



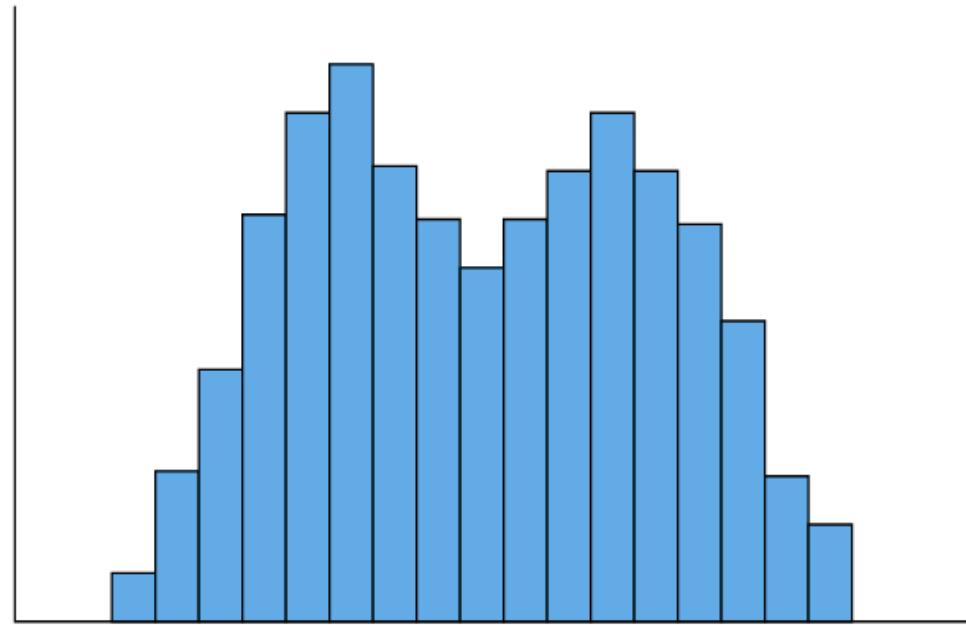
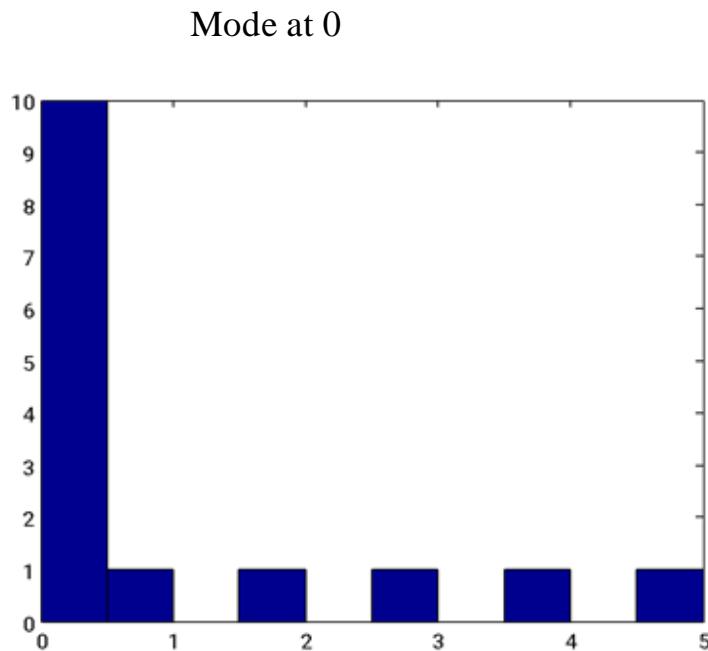
Mode

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**.



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 \tilde{a}^2

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

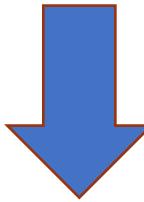
Two-sided Chebyshev's inequality:

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 $1/k^2$.


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

Annotations in red:

- A red arrow points from the symbol σ in the term $k\sigma$ to the symbol s in the denominator k^2 .
- A red arrow points from the term $x_i - \bar{x}$ in the set definition to the symbol σ in the term $k\sigma$.
- A red bracket underlines the term $|S_k|$.
- A red bracket underlines the term N .

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:

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

$\bar{x} = 75$
 $\sigma = 5$
 $k = 2$

$\rightarrow \frac{|S_k|}{N} \leq \frac{1}{4}$

- 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

$$\text{Mean} = 87.69$$

$$\text{Std. dev.} = 3.306$$

Fraction of states with literacy rate in the range

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

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

$$\begin{aligned}
 (N-1)\sigma^2 &= \sum_{i=1}^N (x_i - \bar{x})^2 & S_k = \{x_i \mid (x_i - \bar{x}) > k\} \\
 &= \sum_{i \in S_k} (x_i - \bar{x})^2 + \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
 &\geq \sum_{i \in S_k} (x_i - \bar{x})^2 \quad \checkmark \\
 &\geq |S_k| k^2 \sigma^2 \quad \checkmark
 \end{aligned}$$

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

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 $1/(1+k^2)$.

$$S_k = \{x_i : x_i - \bar{x} \leq -k\sigma\}$$
$$\frac{|S_k|}{N} \leq \frac{1}{1+k^2}$$

Notice: no absolute value!

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?

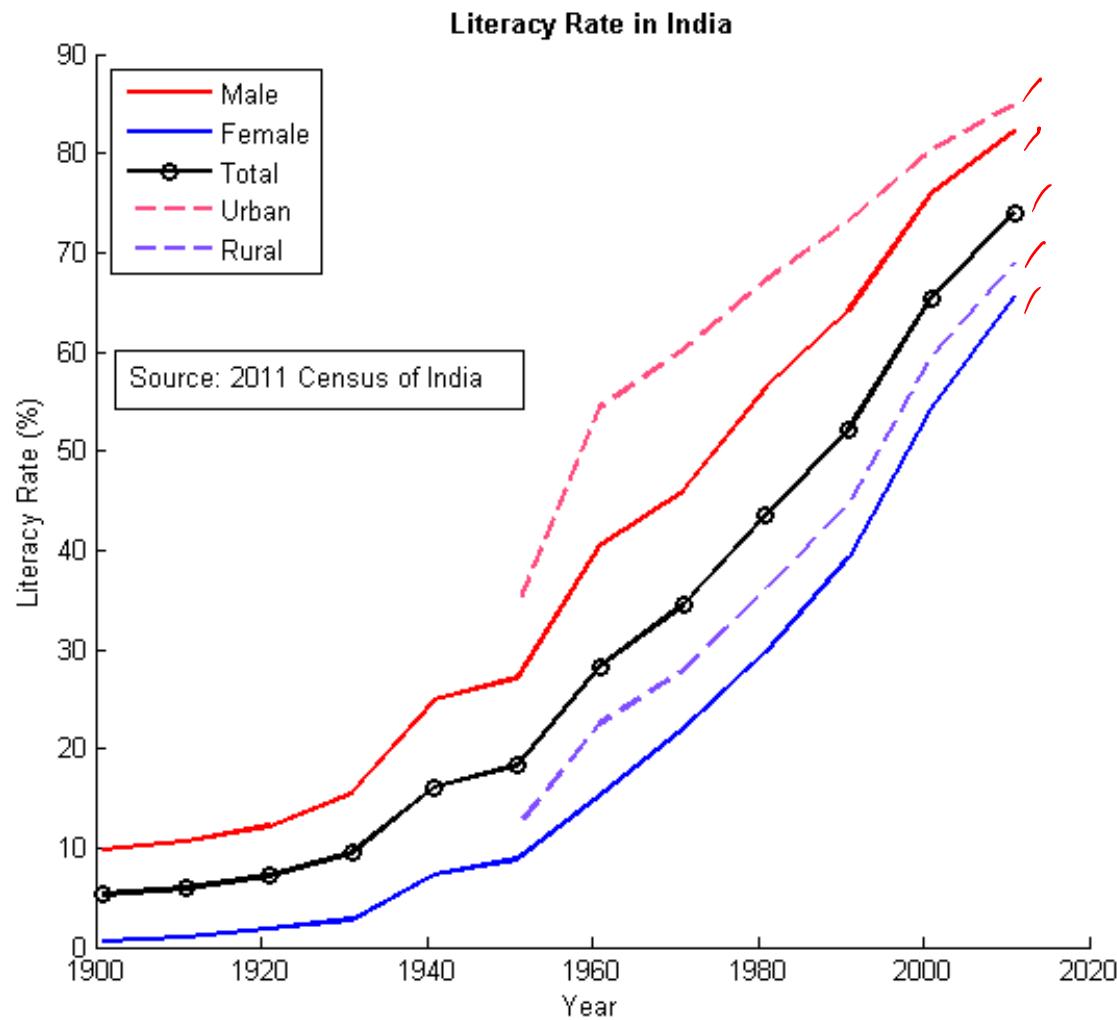


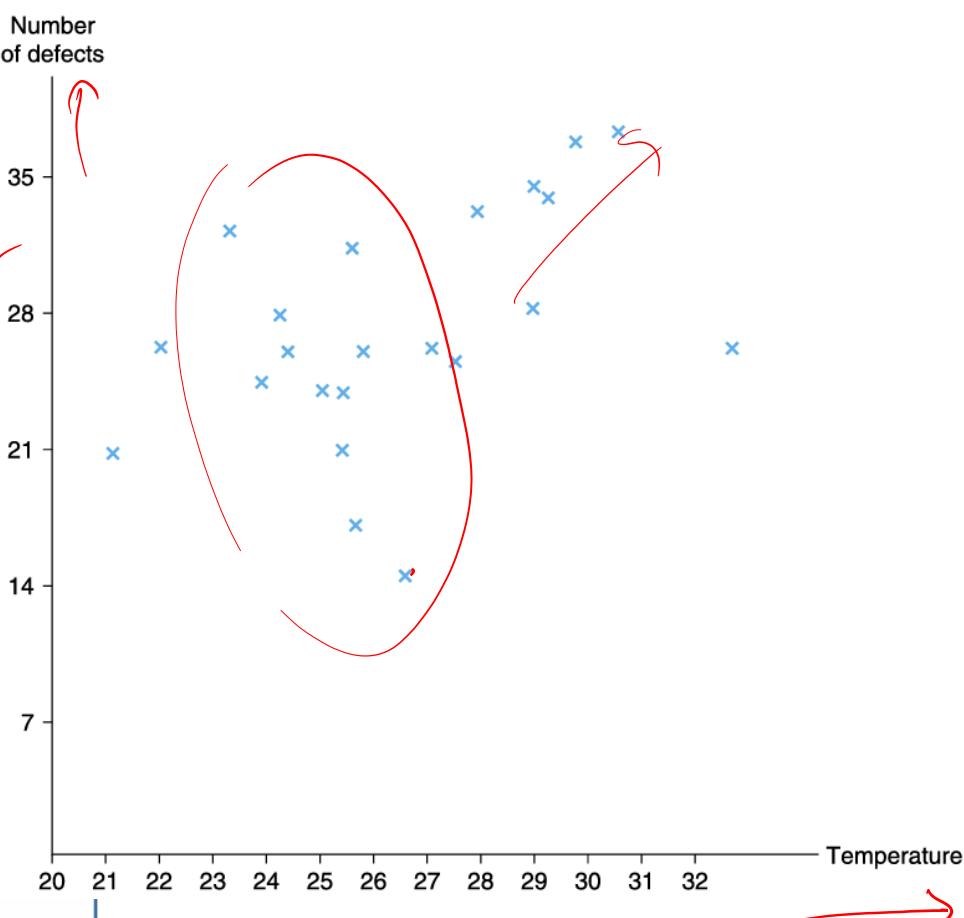
Image source

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	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24



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

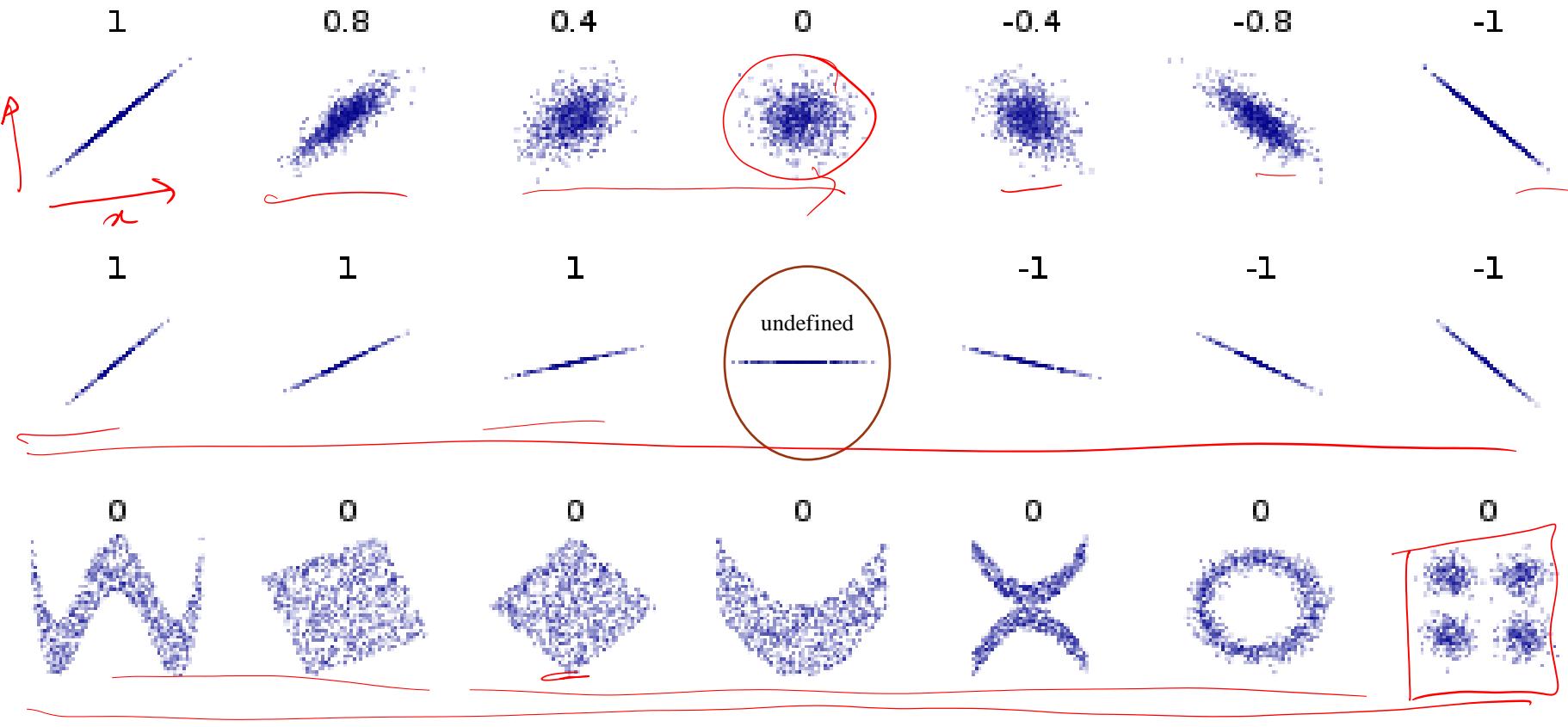
- The correlation-coefficient is given as:

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

- -1 <= r <= 1 always!

Prove it!

$$\begin{aligned} & \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \\ & + (x_{i+1} - \mu_x)(-\underbrace{y_i + 2\mu_y - \mu_y}_{y_{i+1}}) \end{aligned}$$



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 :

$$y_i = x_i^2$$

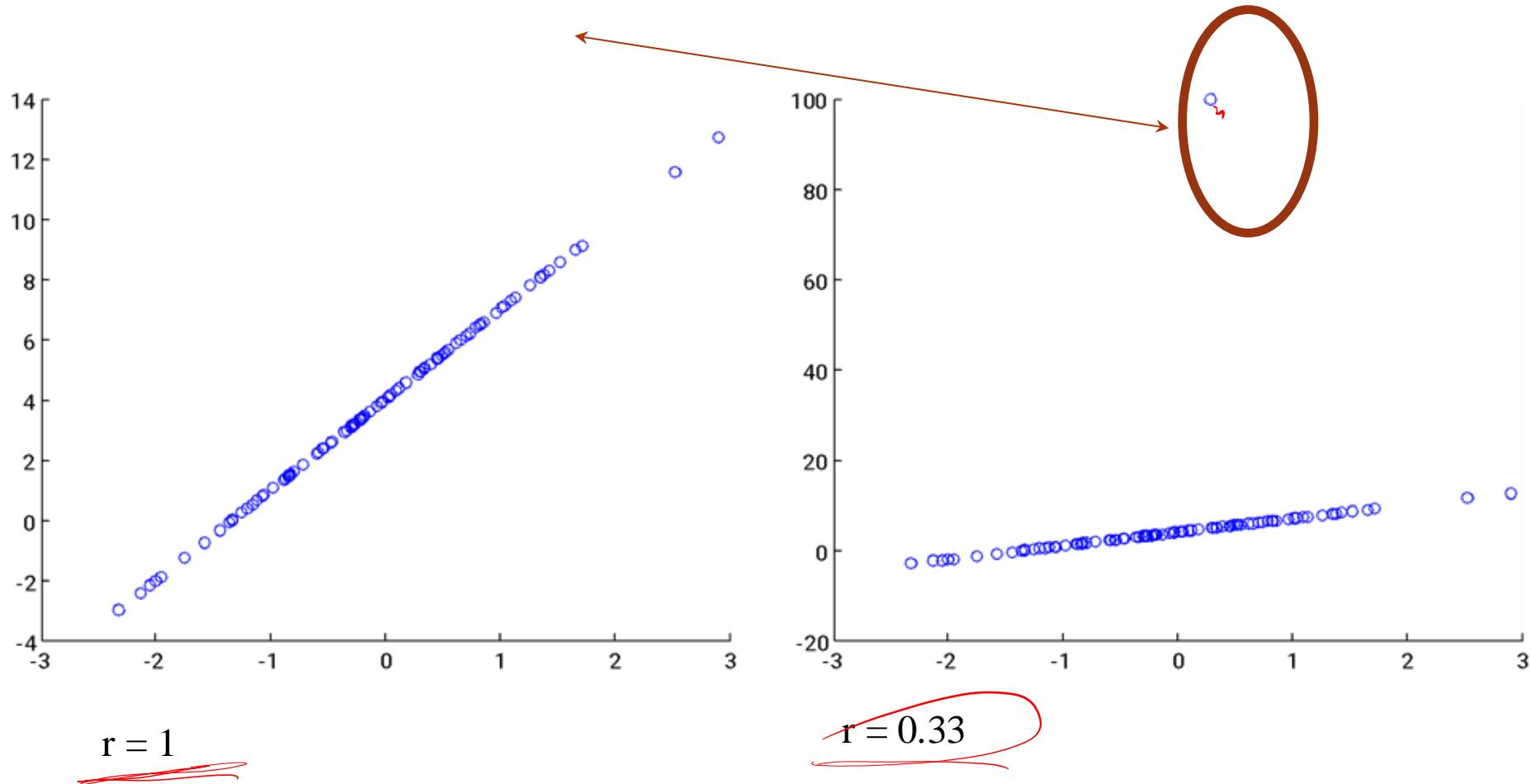
- If $y_i = a + bx_i$ where $b > 0$, then $r(x, y) = 1$.

- 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

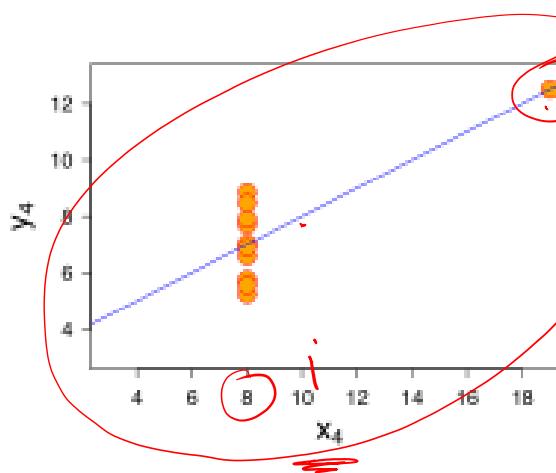
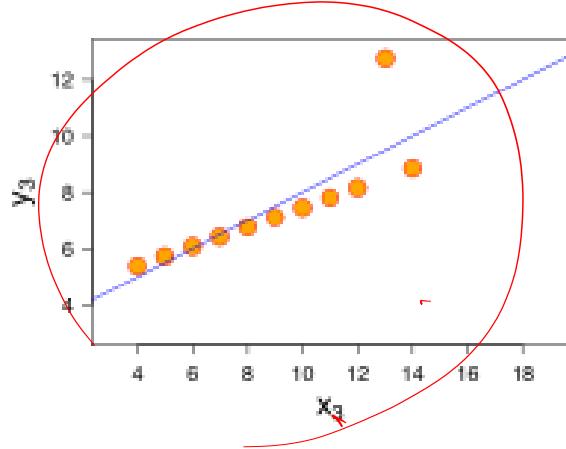
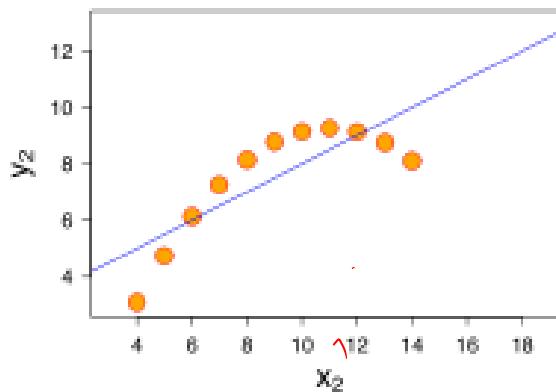
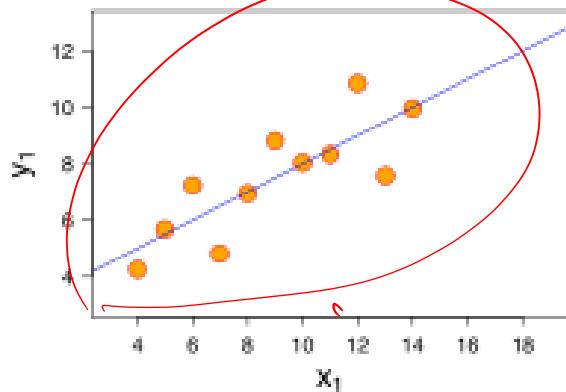
● Sensitive to outliers!



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!

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

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

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

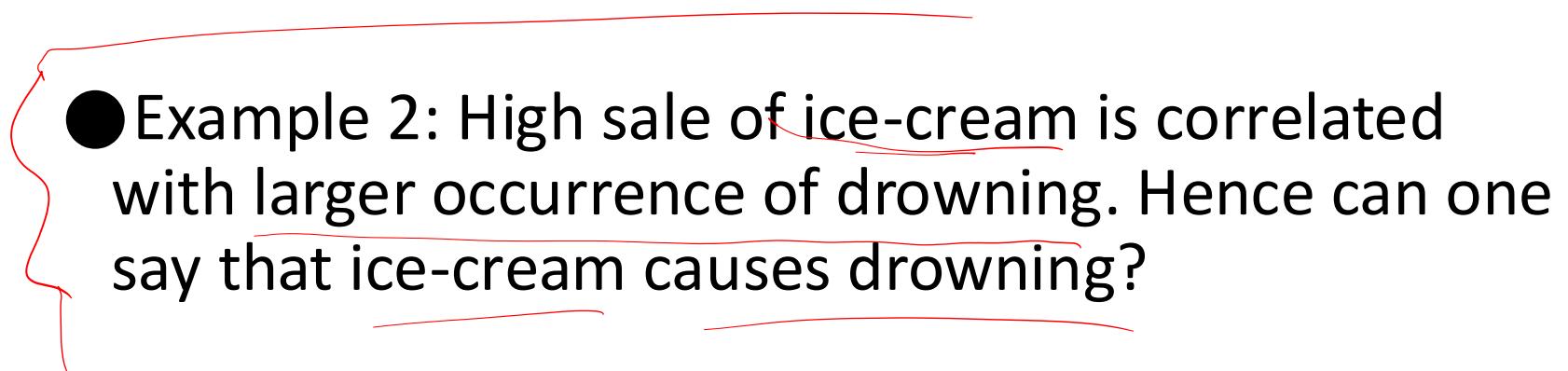
$$r_{uncentered}(\mathbf{x}, \mathbf{y}) \neq r_{uncentered}(\mathbf{x} + \mathbf{a}, \mathbf{y} + \mathbf{b})$$

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.



- 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/publish/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

Elements of Probability

Fall 2024
Sunita Sarawagi

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**.

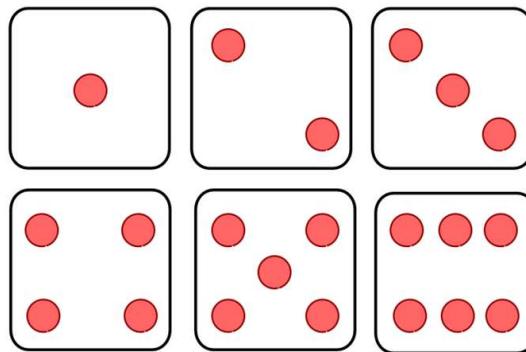


Sample space

- 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

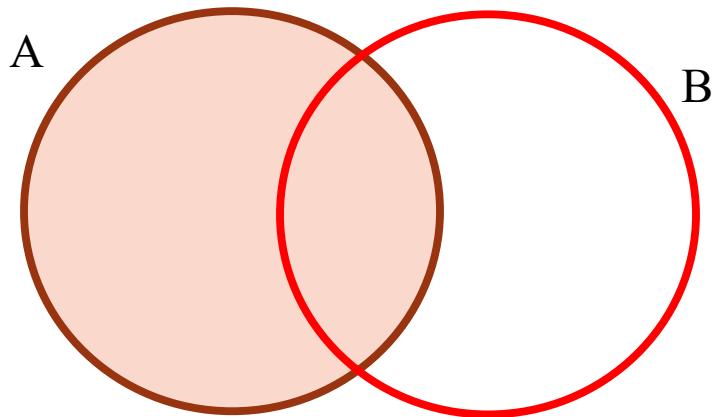
- Any subset of the sample space is called an **event**.
- If the outcome of an experiment is contained in Event E , then we say E has *occurred*.
- In example 1, if $E = \{H\}$, then E is the event that the coin produced a heads.
- In example 2, if $E = \{\text{set of temperatures from 33 to 37}\}$, 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):

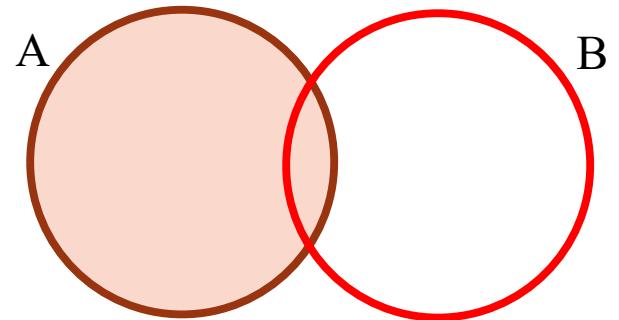
$$P(A \cup B) = P(A) + P(B) - P(AB) \leq P(A) + P(B)$$

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, GI}
- $P(A, B) = P(KC) = p$
- $P(B) = p + (1-p) \frac{1}{m}$

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

$$m = 5$$

$$p = 0.6$$

$$P(A|B) = \frac{0.6}{0.6 + \frac{0.4}{5}} = \frac{1}{1 + \frac{2}{3} \times 5} = 0.88$$

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 1% 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?

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

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

$$P(D) = 0.005$$

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

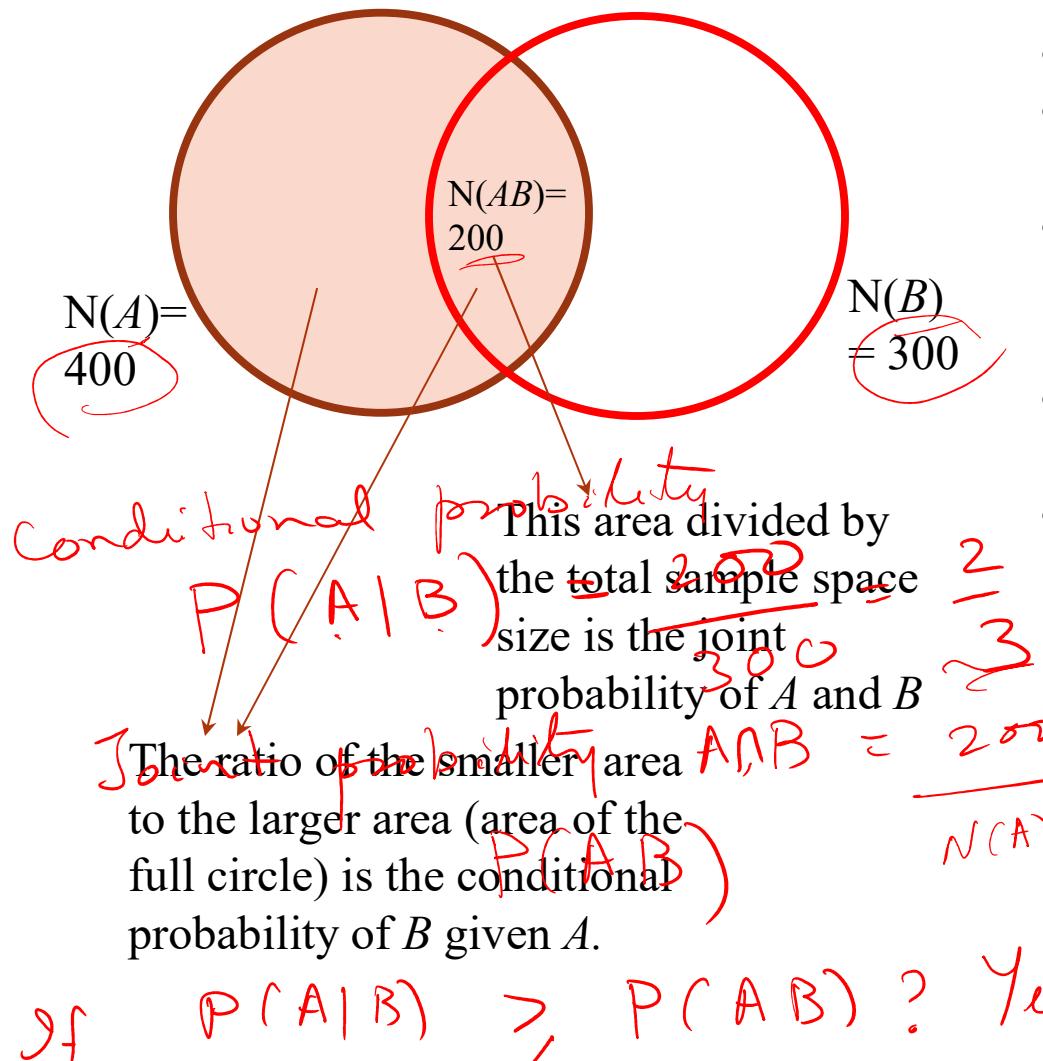
$P(T_P)$

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

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?

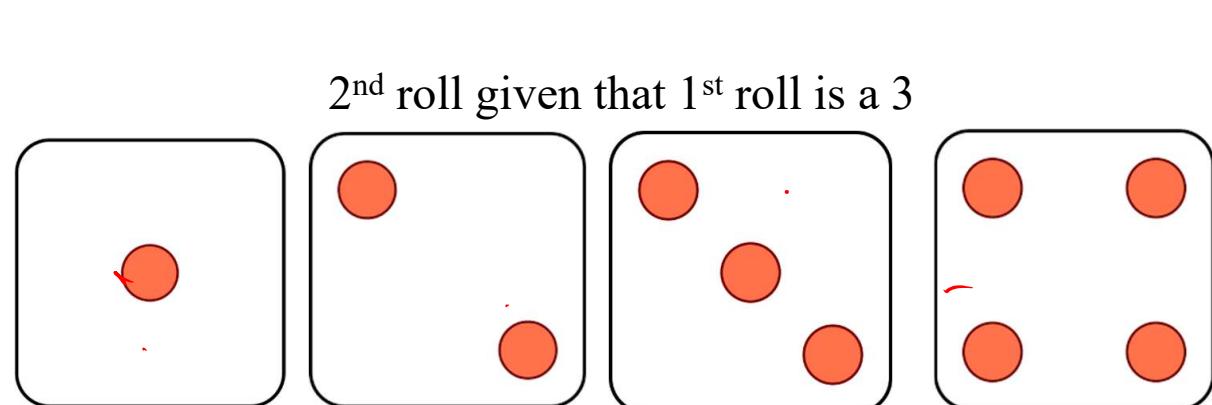
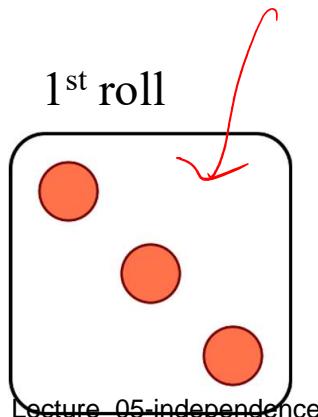


- Let the original sample space be S .
- In computing $P(B|A)$, you assume that A has already occurred.
- Therefore your new sample space S' for computing $P(B|A)$ contains only those events which lie in A .
- For computing $P(AB)$, the sample space is the entire S .
- Now can you compute $P(A|B)$?

$$\frac{N(AB)}{N(A) + N(B) - N(AB)} = \frac{200}{500} = \frac{2}{5}$$

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 • 74 P(L)
- A randomly chosen Indian person is from Kerala • 0.028 P(k)
- A randomly chosen person from Kerala is literate • 94 P(L|k)
- A randomly chosen Indian person is from Kerala and is literate $P(KL) = P(L|k)P(k) = 0.94 \times 0.028$
- 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:

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

Intuitive Definition:

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

With independence, we can simplify the chain rule:

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

You can also show this ^ to prove independence

Piech & Cain, 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:

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

$\rightarrow n=3 : A_1, A_2, A_3$

- Example: three events A, B, C. 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 $P(E_7|FT) = 1$]. ■

$$E_7 \perp\!\!\!\perp T$$

$$E_7 \perp\!\!\!\perp F$$

$$\{ \not\models E_7 \perp\!\!\!\perp TF$$

$S = \{$	[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] $\}$

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

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

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

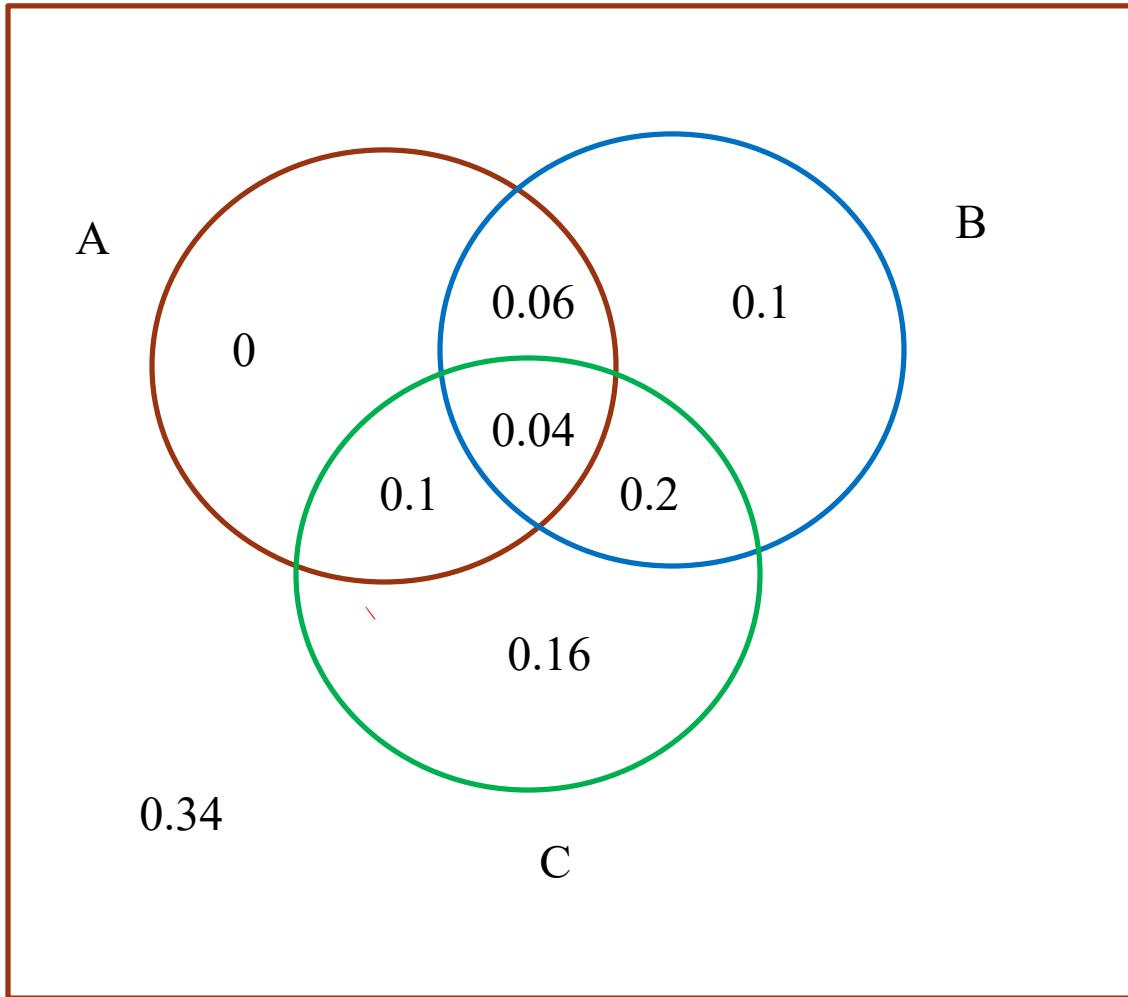
$$P(E_7 \cap F) = \frac{1}{36}$$

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

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

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



$$\cancel{P(ABC) = 0.04 = P(A)P(B)P(C) = (0.2)(0.4)(0.5)}$$

$$\cancel{P(AB) = 0.1 \neq P(A)P(B)}$$

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:

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

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

$$\begin{aligned} \Rightarrow P(A) P(B^c) &= P(A) - P(A) P(B) \\ &= P(A) [1 - P(B)] = \\ &= P(A) P(B^c) \end{aligned}$$

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

The Core Probability Toolkit



The Law of Total Probability

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

$$P(E) = \sum_{i=1}^n P(E \text{ and } B_i)$$
$$= \sum_{i=1}^n P(E|B_i)P(B_i)$$

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

Bayes' Theorem

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

Definition of Conditional Probability

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

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

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, use Inclusion-Exclusion:

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

Chain Rule

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

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

De Morgan's Laws

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

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

Independence

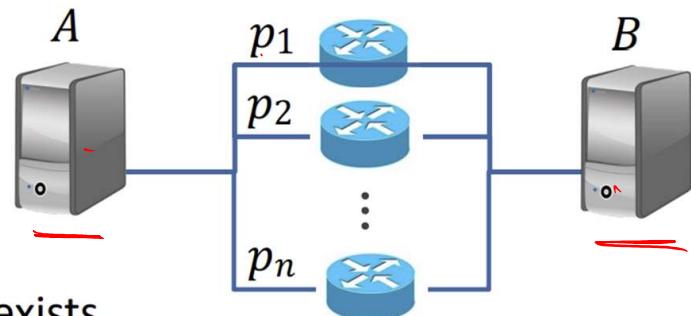
$$P(E|F) = P(E)$$
$$P(E \text{ and } F) = P(E)P(F)$$

Practice: Network Reliability

Consider the following parallel network:

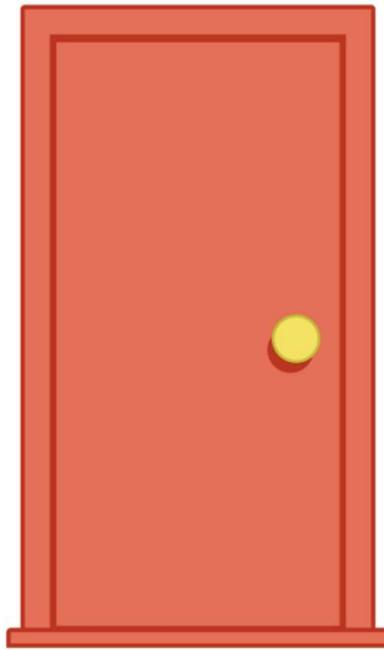
- n independent routers, which each are working with probability p_i ($1 \leq i \leq n$)

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

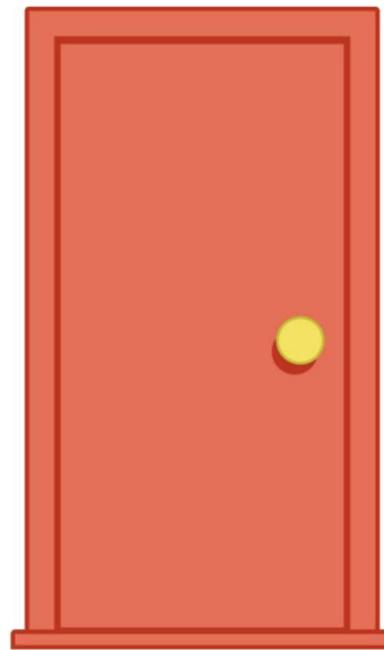


$$P(E) = 1 - \prod_{i=1}^n (1-p_i)$$

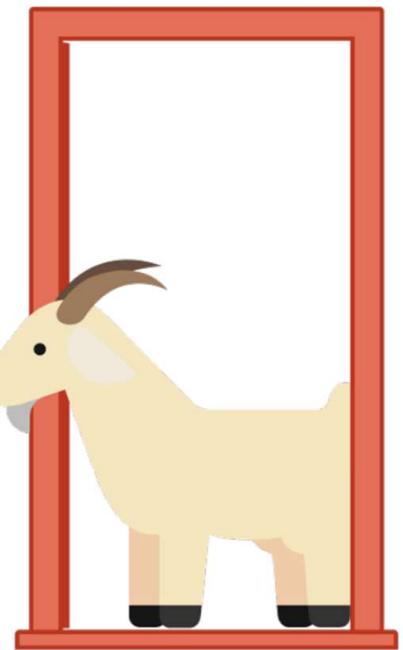
1



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?

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

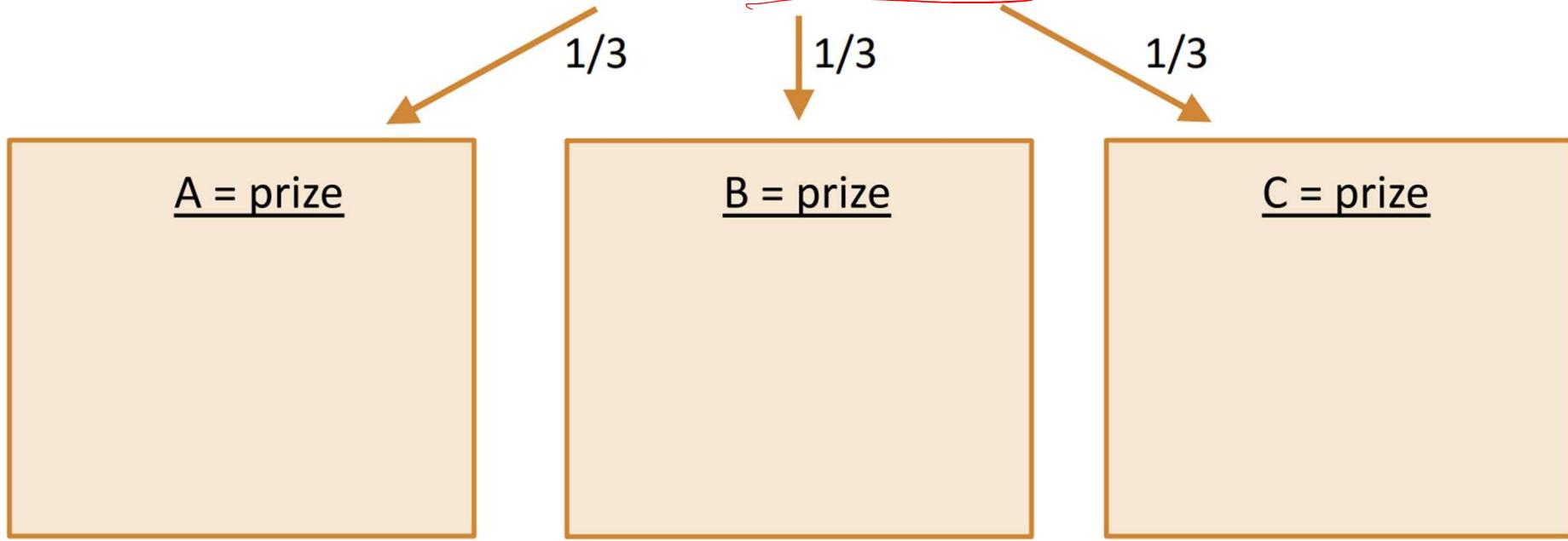
$$P(\text{prize in 1}) = P(\text{prize in 2})$$

Piech & Cain, CS109, Stanford University

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

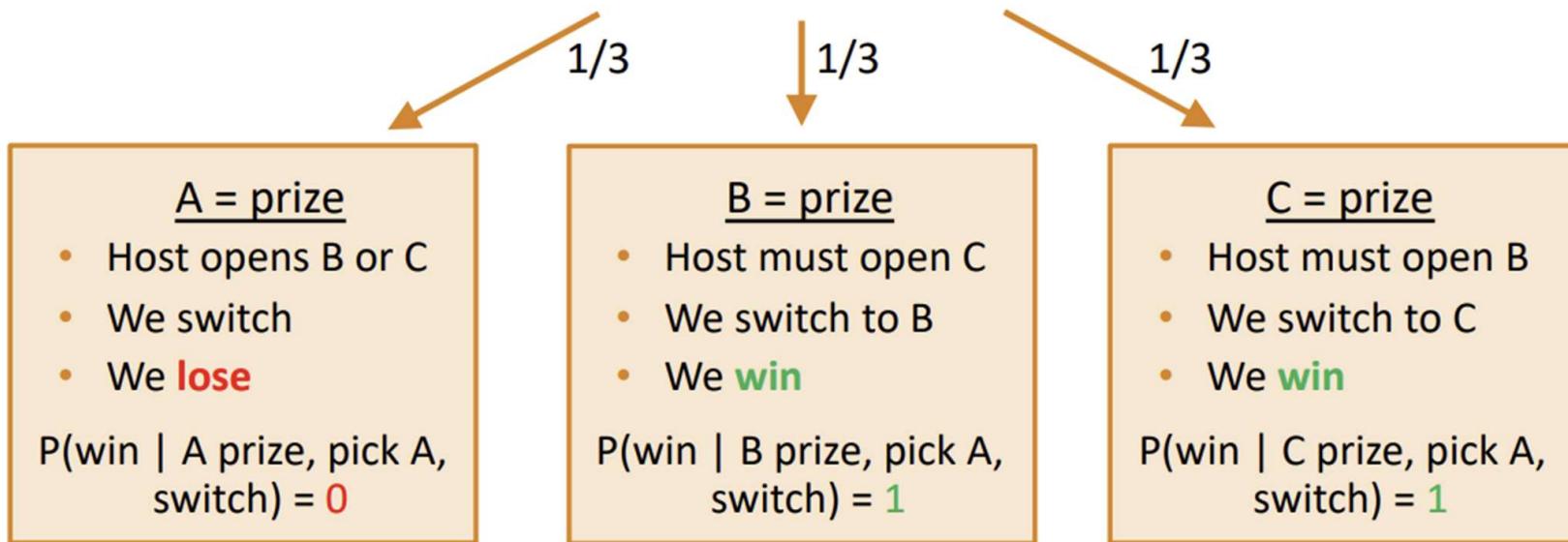
Paul Erdos

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!

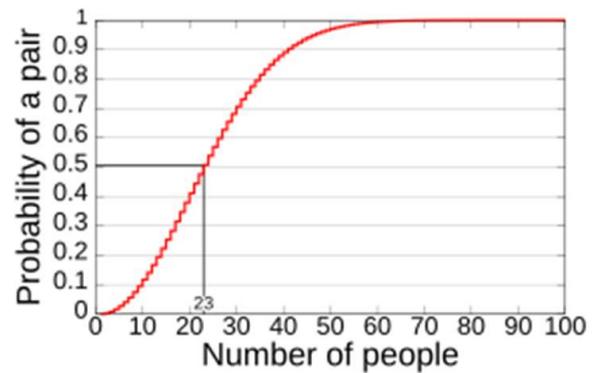
Piech & Cain, CS109, Stanford University

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

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.

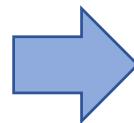
Random Variables Are Variables...That Are Random

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.”

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



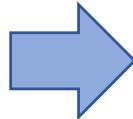
Random Variables Are Variables...That Are Random

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."

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

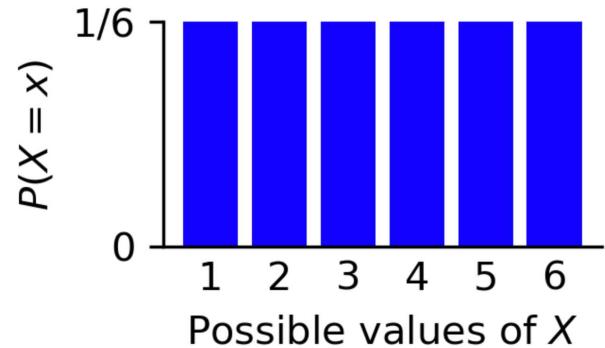
...or, $P(\underline{X} = \underline{x}) = 1/6$ for $\underline{1} \leq \underline{x} \leq \underline{6}$

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 $1 \leq x \leq 6$

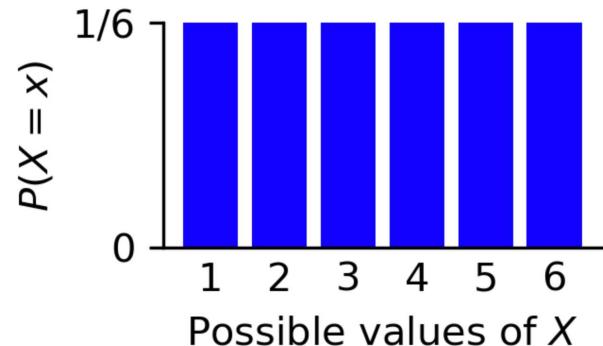


Examples of Random Variables

"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."

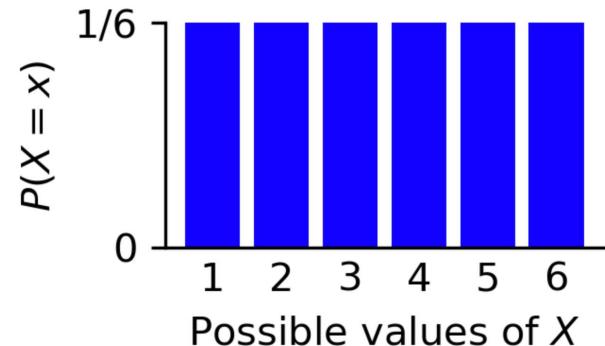
- $P(Y = 0) = 1/4$
 - $P(Y = 1) = 1/2$
 - $P(Y = 2) = 1/4$
- (T, T)
(H, T), (T, H)
(H, H)

Examples of Random Variables

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

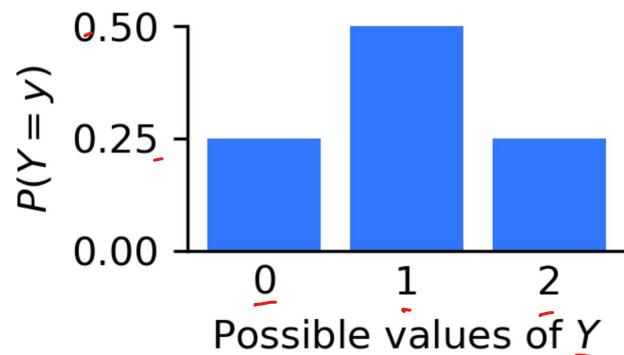
- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$
- $P(X = 4) = 1/6$
- $P(X = 5) = 1/6$
- $P(X = 6) = 1/6$

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



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

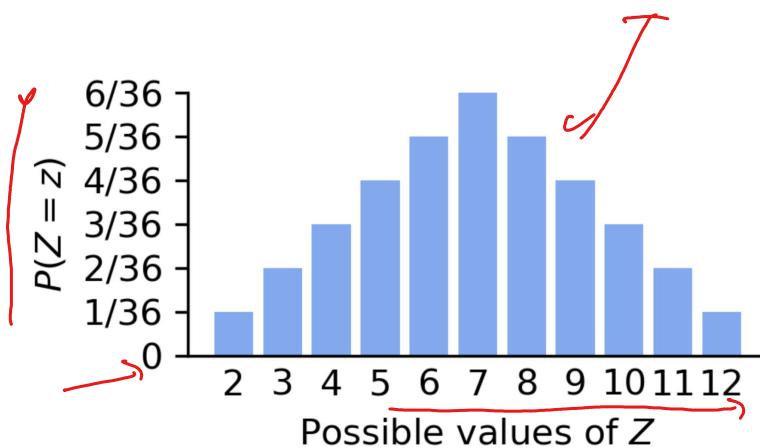
- $P(Y = 0) = 1/4$ (T, T)
- $P(Y = 1) = 1/2$ (H, T), (T, H)
- $P(Y = 2) = 1/4$ (H, H)



Examples of Random Variables

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

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



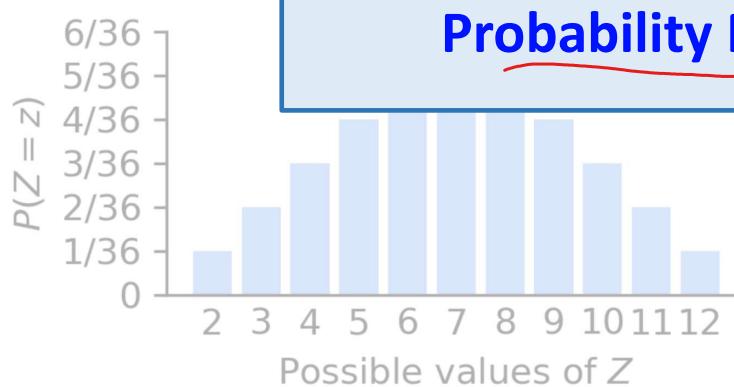
$$P(Z = z) = \begin{cases} \frac{z-1}{36}, & z \in \mathbb{Z}, 1 \leq z \leq 6 \\ \frac{13-z}{36}, & z \in \mathbb{Z}, 7 \leq z \leq 12 \\ 0, & \text{else} \end{cases}$$

Examples of Random Variables

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

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

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



Probability Mass Function (PMF)

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

$\mathbb{Z}, 1 \leq z \leq 6$

Probability Mass Functions

Random Variables & Functions

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

If this is a number

$$P(Y = 2)$$

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

$$P(Y = k)$$

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

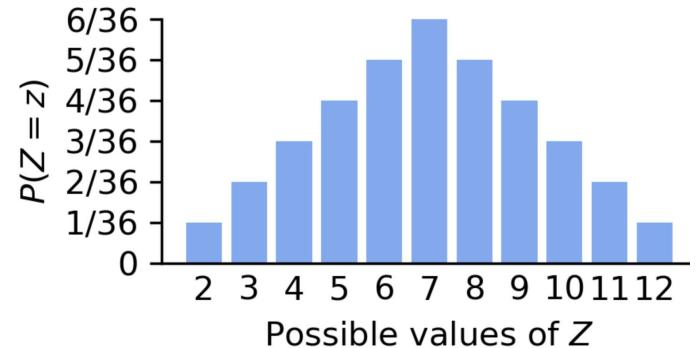
$$P(Y = k)$$

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!

Random Variables: Continued

Types of Random Variables

- Discrete Vs Continuous

Specifying probability of discrete R.V.

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

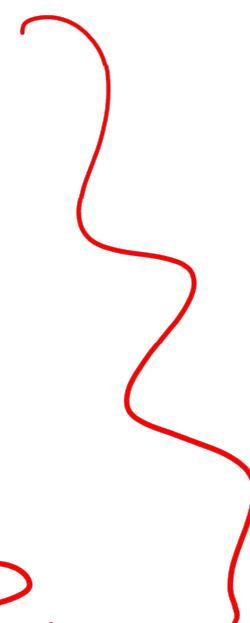
$X \in \{1, 2, 3, 4, 5, 6\}$ ← outcome of dice roll

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

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

.

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



Enumeration representation of PMF.

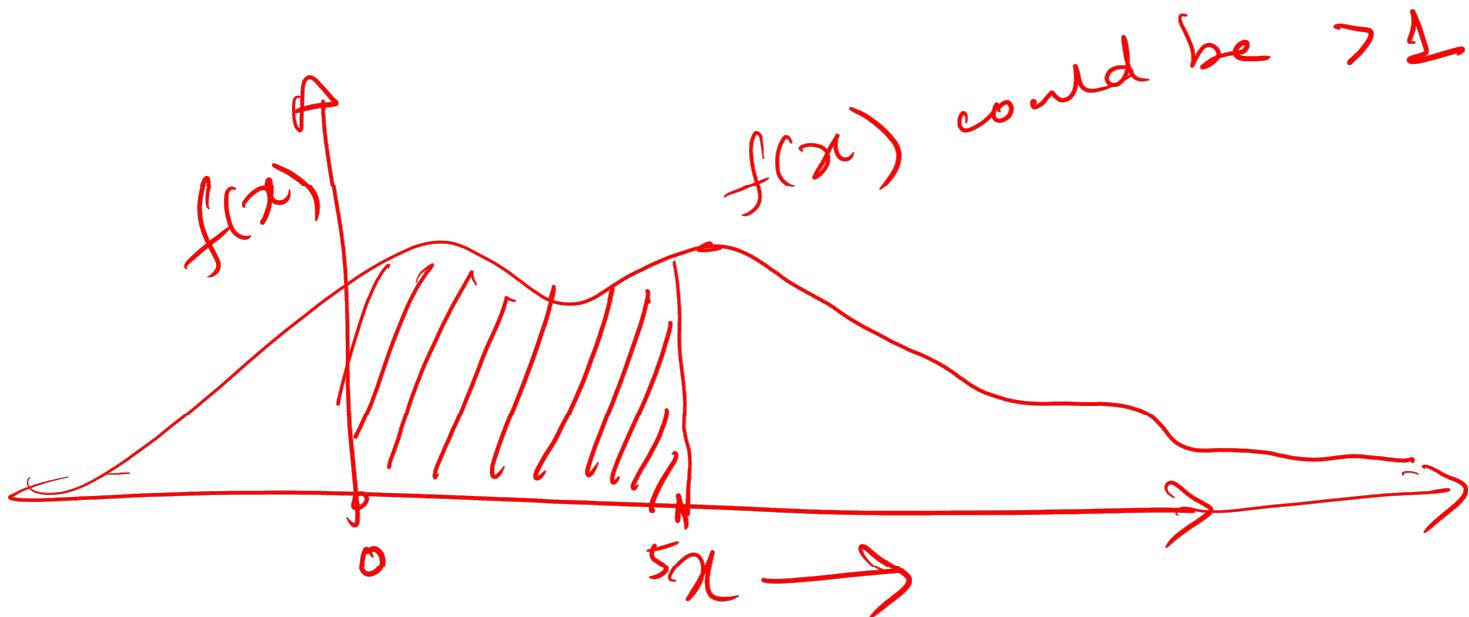
Specifying probability of continuous R.Vs

- 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: $\underline{P(X \in B)}$



$$\underline{P(X \in B)} = \int_{x \in B} f(x) dx$$

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



x

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



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_n

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

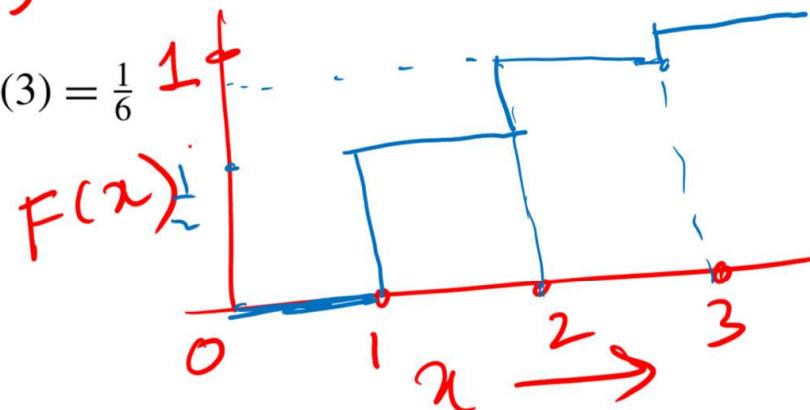
- Example: $p(1) = \frac{1}{2}, p(2) = \frac{1}{3}, p(3) = \frac{1}{6}$

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

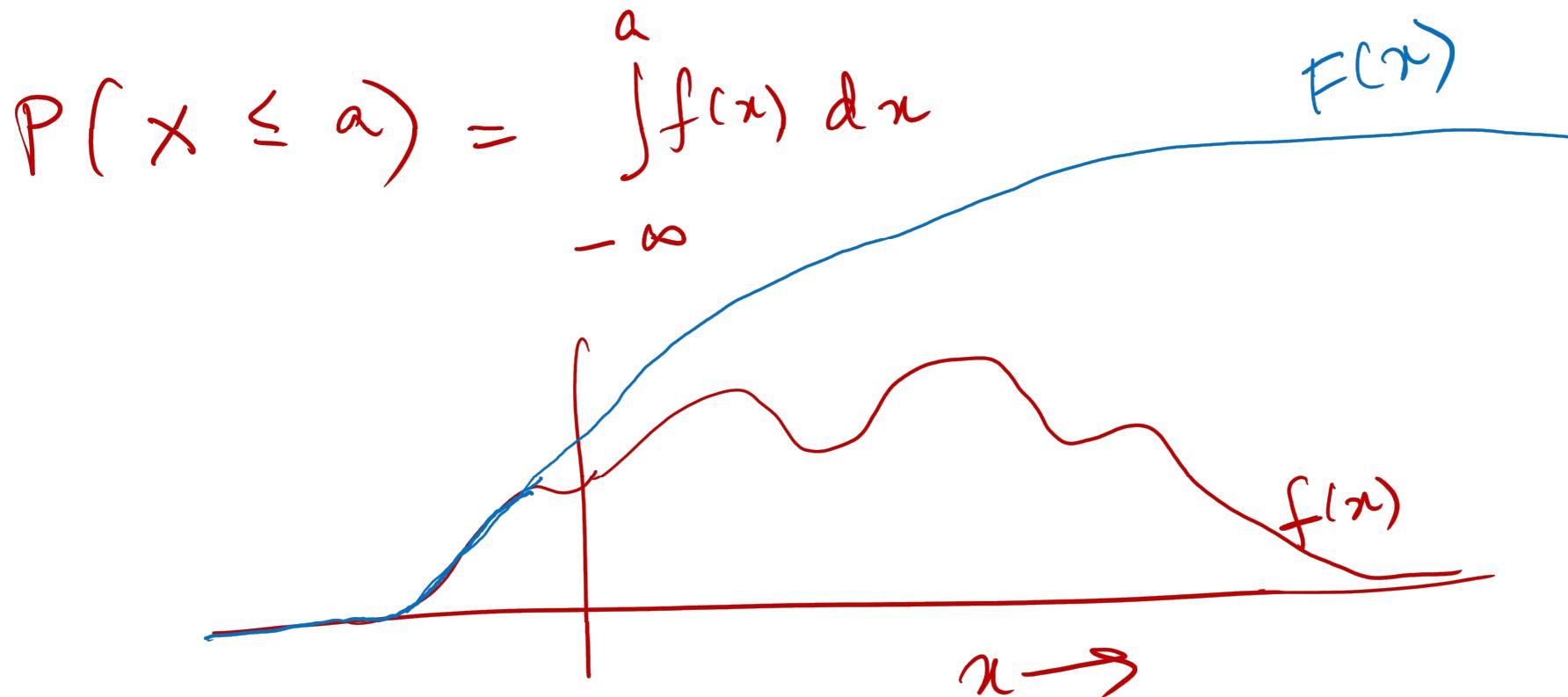
$$F(1) = \frac{1}{2}$$

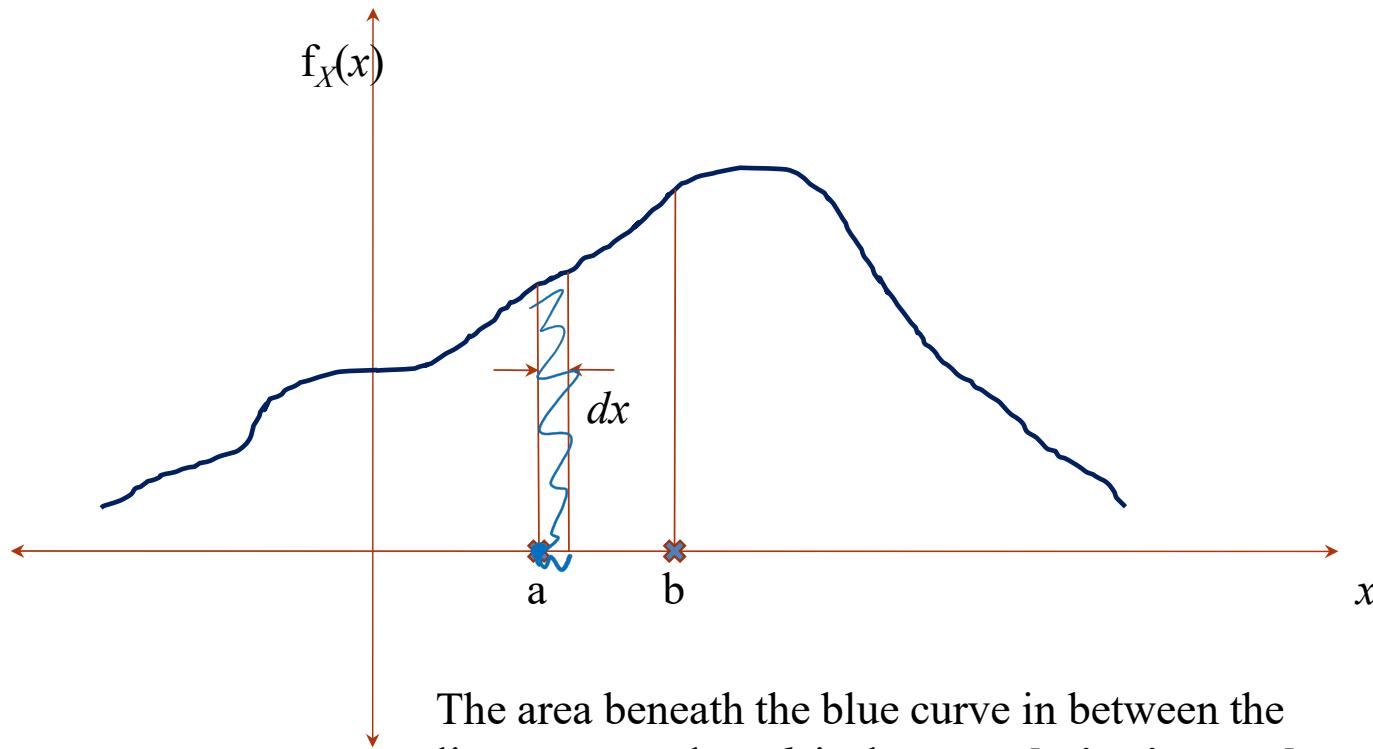
$$F(2) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

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



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 $P(a < X \leq b) = F_X(b) - F_X(a)$.

$f_X(a)dx$ = 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:

$$\text{E}(X) = \sum_i x_i P(X = x_i)$$

→ 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:

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

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} = \-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

$$\begin{aligned} E(ag(X) + b) &= \int_{-\infty}^{+\infty} (ag(x) + b) f_X(x) dx \\ &= \int_{-\infty}^{+\infty} ag(x) f_X(x) dx + \int_{-\infty}^{+\infty} bf_X(x) dx \\ &= aE(g(X)) + b \end{aligned}$$

--- why ? ---

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:

$$E(g(X)) \neq g(E(X))$$

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.

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

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:

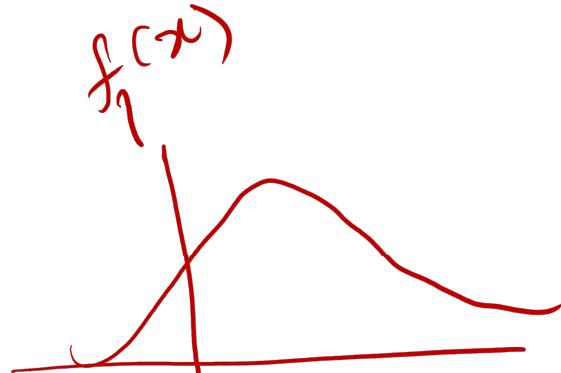
$$Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

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

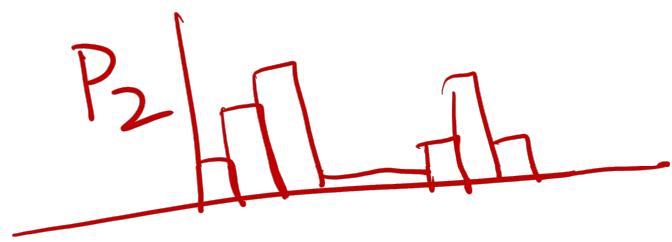
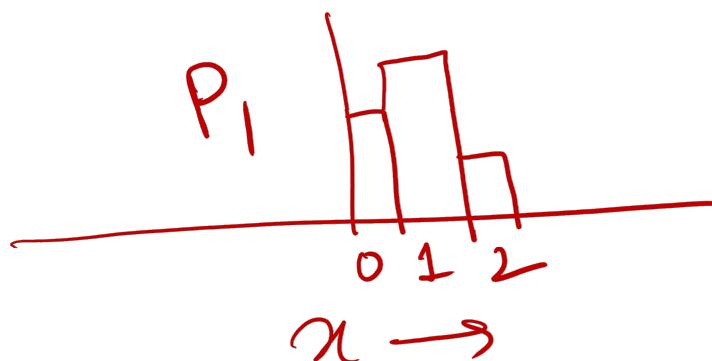
$$Var(X) = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 P(X = x_i)$$

- 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

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

PMF of X $P(X=1) = \theta$

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

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

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

Well-known discrete Random Variables.

The Simplest Random Variable

- Bernoulli Random Variable Boolean R.V.

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

Examples:

- coin - toss
- equipment will fail or not
- whether your crush 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

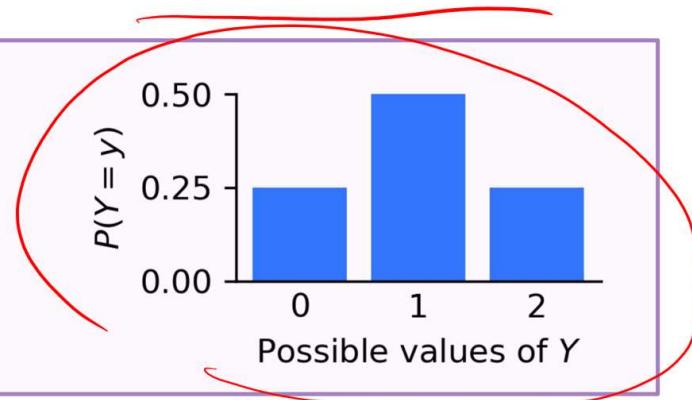
Binomial Random Variable

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)
- $P(Y = 1) = 1/2$ $(H, T), (T, H)$
- $\underline{P(Y = 2) = 1/4}$ (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?

$$\underbrace{(\text{H}, \text{ H}, \text{ H}, \text{ H}, \text{ T}, \text{ T}, \text{ T}, \text{ T}, \text{ T}, \text{ T})}_{\text{4 heads}} \quad p^4(1-p)^6$$
$$\underbrace{(\text{H}, \text{ H}, \text{ H}, \text{ T}, \text{ H}, \text{ T}, \text{ T}, \text{ T}, \text{ T}, \text{ T})}_{\text{3 heads}} \quad \checkmark$$

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

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

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

(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, T, H, H, T, T, T, T, T)
(H, H, T, H, T, H, T, T, T, T)
(H, H, T, H, T, H, T, T, T, T)
(H, H, T, H, T, T, H, T, T, T)
(H, H, T, H, T, T, T, H, T, T)
(H, H, T, H, T, T, T, T, H, T)
(H, H, T, H, T, T, T, T, T, H)
(H, H, T, T, H, H, T, T, T, T)
(H, H, T, T, H, T, H, T, T, T)
(H, H, T, T, H, T, T, H, T, T)
(H, H, T, T, H, T, T, T, H, T)
(H, 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 $p^k(1 - p)^{10-k}$

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

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

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

Our random variable →

Num trials →

Probability of success on each trial →

Is distributed as a →

Binomial ↑

With these parameters

Then We Automatically Know the PMF!

Probability Mass Function
for a Binomial


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

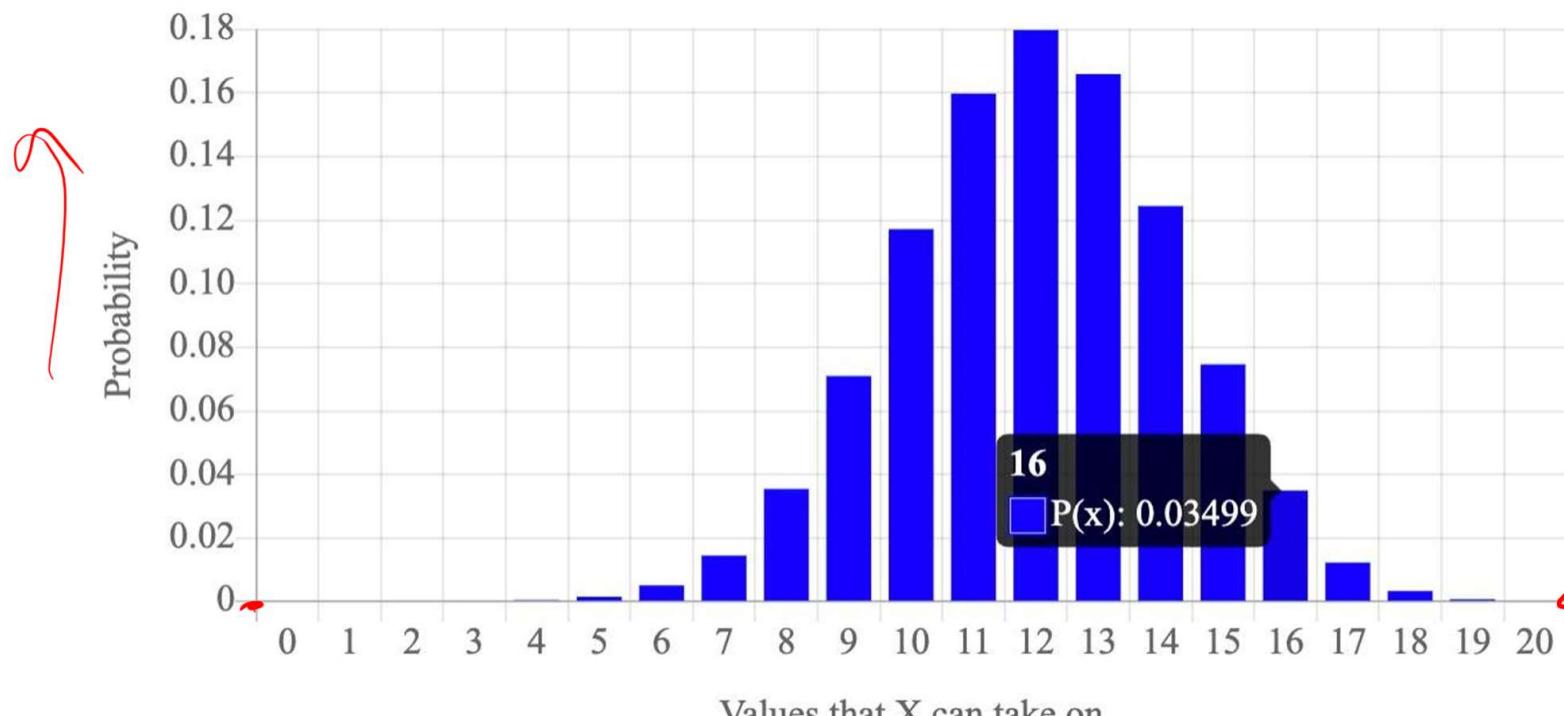
\uparrow

Probability that our variable takes on the value k

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

Parameter n : 20

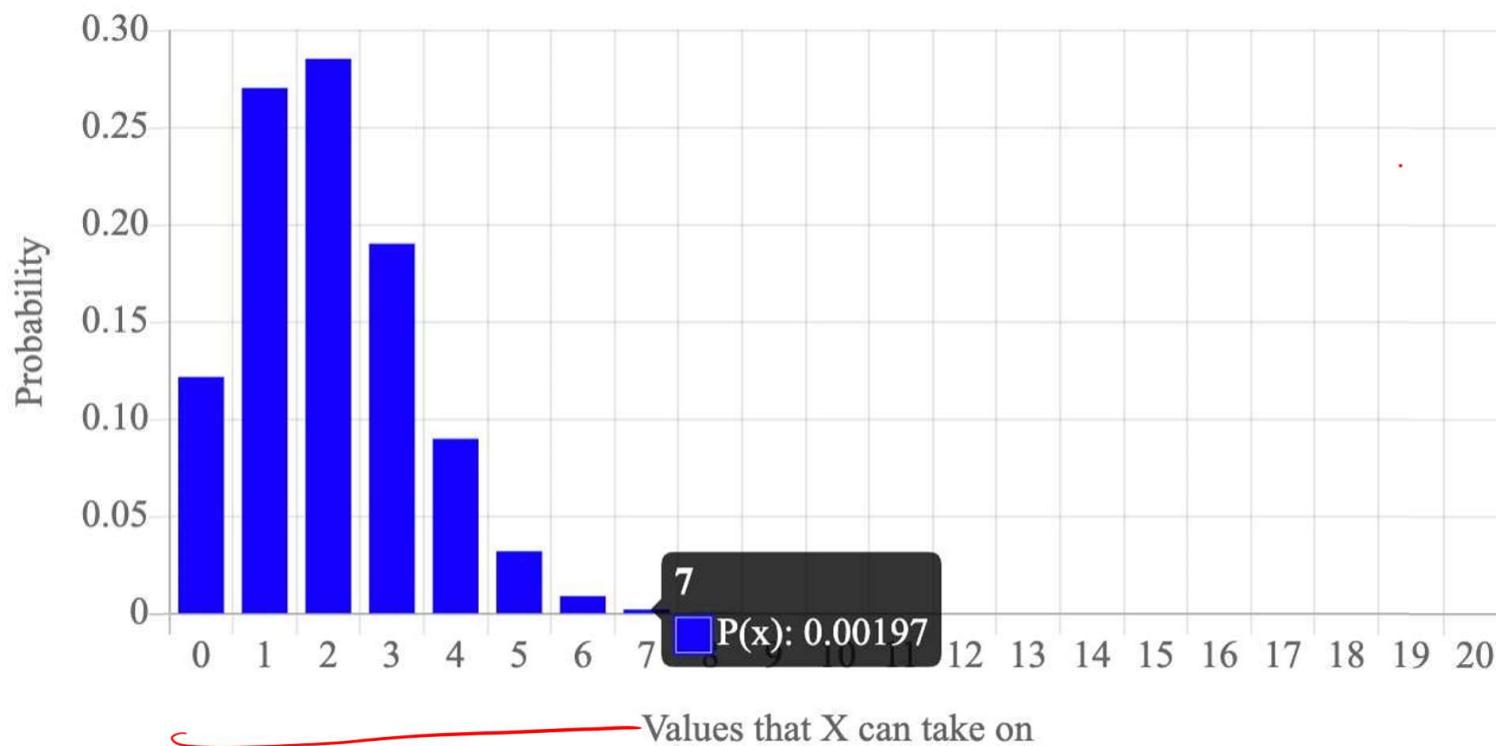
Parameter p : 0.60



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

Parameter n : 20

Parameter p : 0.



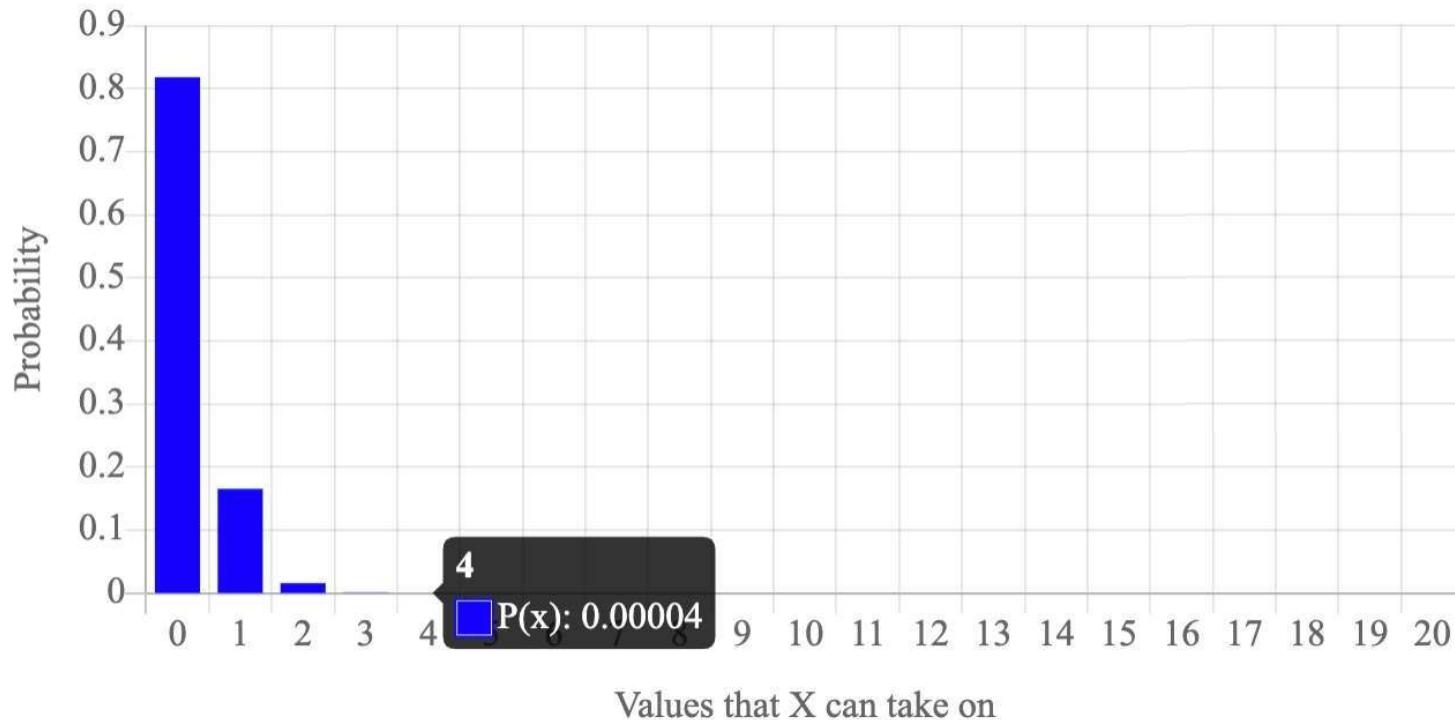
The PMF as a Graph: $X \sim \text{Bin}(n = 20, p = 0.01)$

Parameter n :

20

Parameter p :

0.01



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?

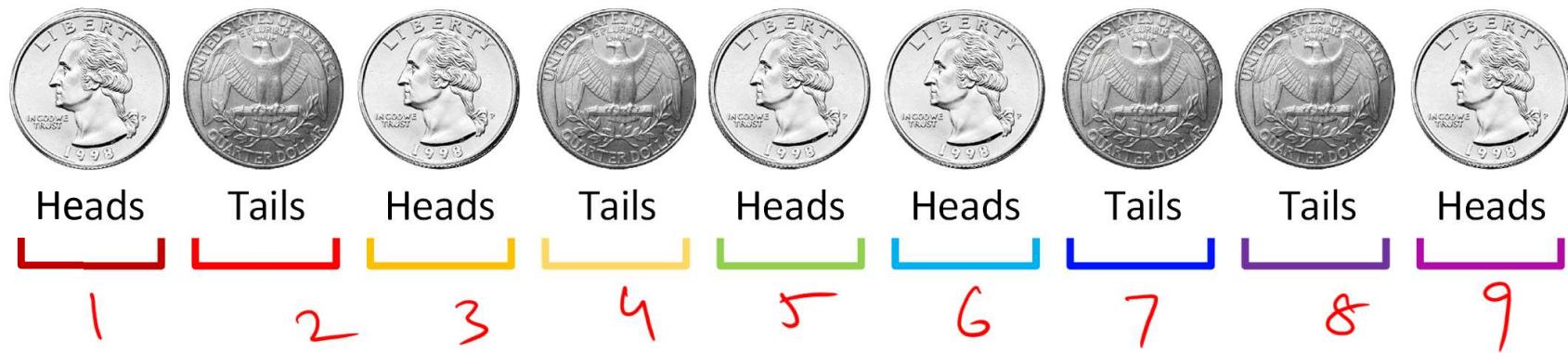
$$P(X = 3) = \binom{3}{3} p^3 (1-p)^0 = \frac{1}{8}$$

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 $\underline{X}_1 \sim \text{Bern}(p = 1/2)$ and $\underline{X}_2 \sim \text{Bern}(p = 1/2)$.

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

$$\underline{Y} = \underline{X}_1 + \underline{X}_2$$

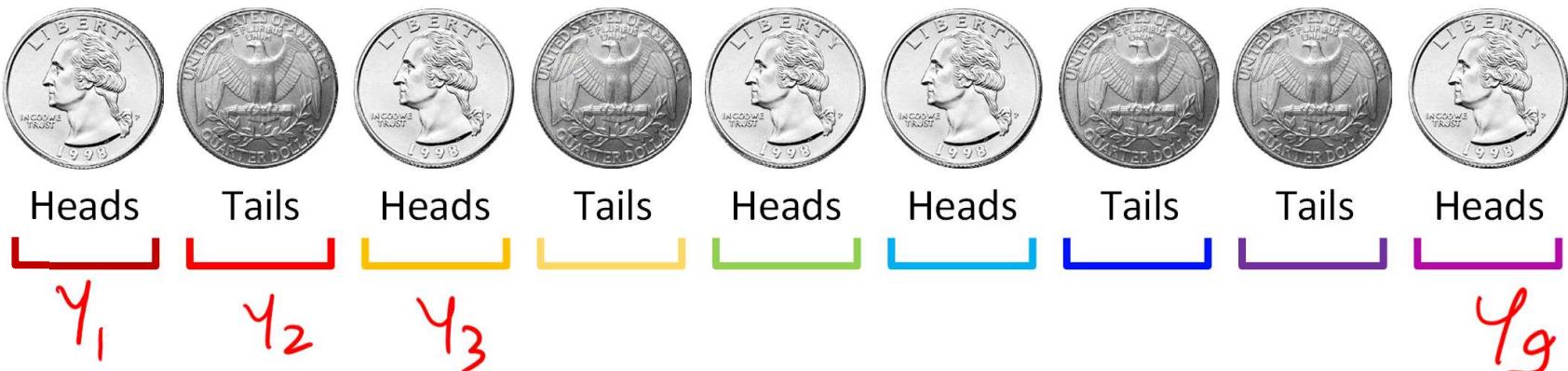
We Can Now Calculate Expectation of Binomial

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

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

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 . $Y_i \sim \text{Bern}(p)$.

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

$$= \sum_{i=1}^n \mathbb{E}[Y_i] \quad \text{Expectation of sum}$$

$$= \sum_{i=1}^n p \quad \text{Expectation of Bernoulli}$$

$$= n \cdot p \quad \text{Sum } n \text{ times}$$

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 . $Y_i \sim \text{Bern}(p)$.

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

True for every binomial ever

Variance of Binomial R.V.s $X = \sum_{i=1}^n Y_i$ $Y_i \sim \text{Bern}(p)$

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

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

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

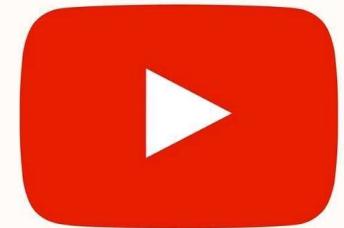
$$\text{Var}(X) = np(1-p)$$

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

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

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

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

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).

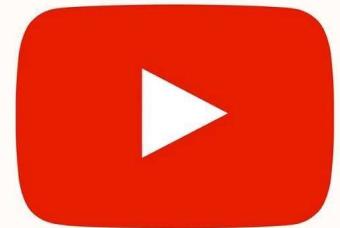
What is the probability of this ad getting **20** clicks?

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

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

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

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?

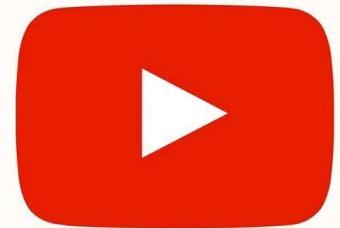
Let X be the number of ad clicks. $X \sim \text{Bin}(n = 1000, p = 0.01)$.

```
[>>> 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?

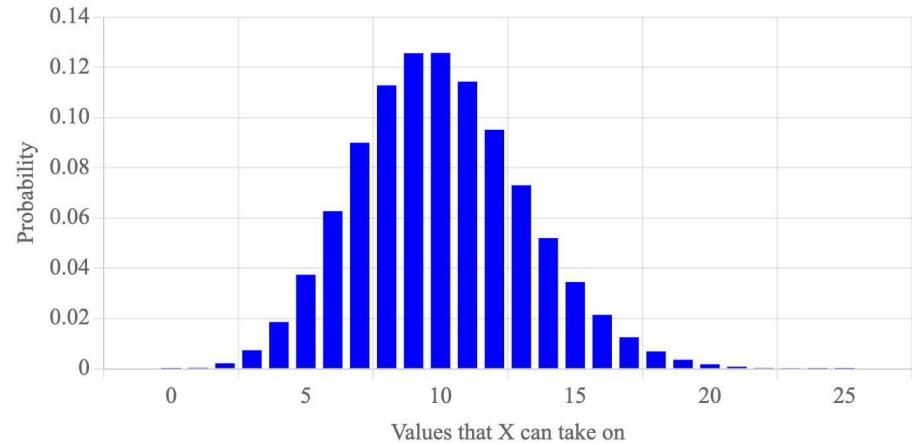
Let X be the number of ad clicks.

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

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 \mathbf{X} be the number of servers alive.

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

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

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

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

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

$$\underline{P(X < 2) = P(X = 0) + P(X = 1)}$$

Server Redundancy



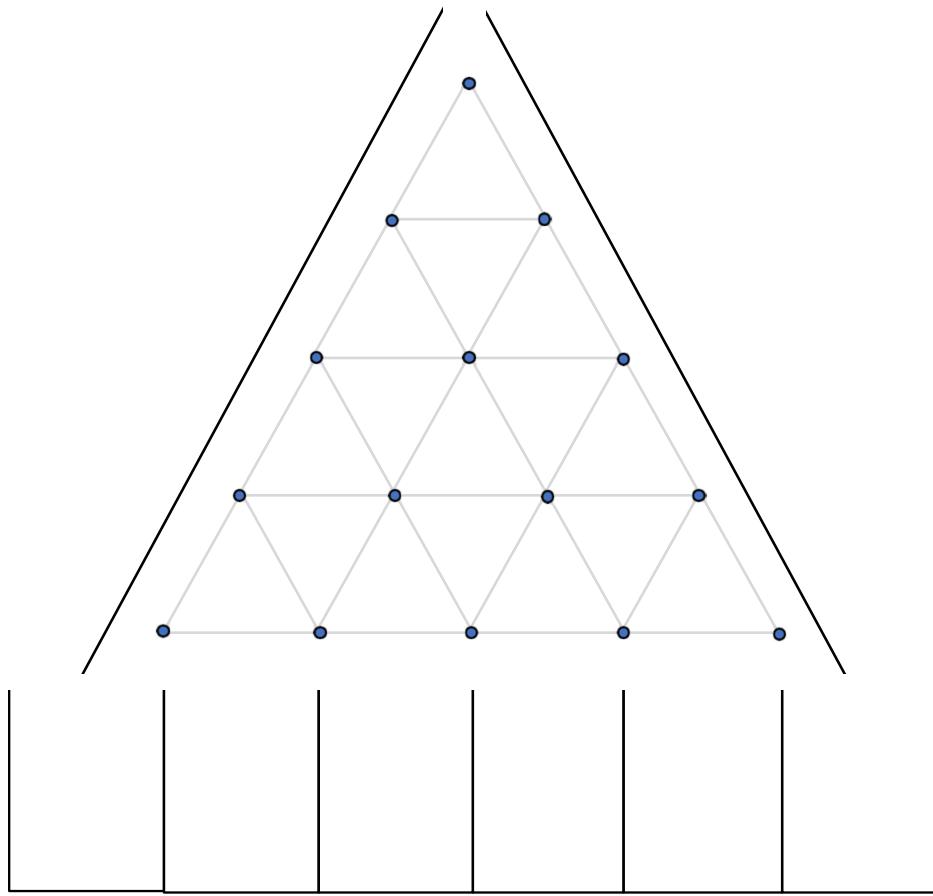
A network can remain functional as long as at least 2 out of 7 servers are alive.
The probability of any server working is 0.8.
What is the probability that less than 2 servers are alive?

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

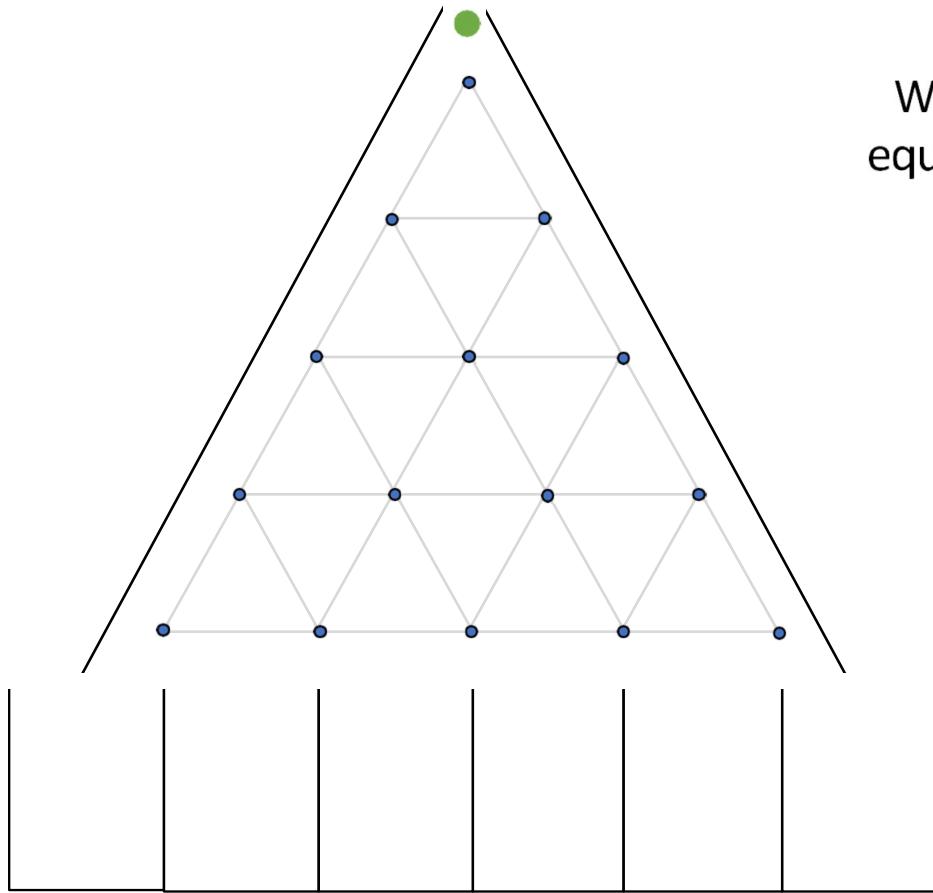
$$P(X = k) = \binom{7}{k} (0.8)^k (0.2)^{7-k}$$

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

Galton Board Fun

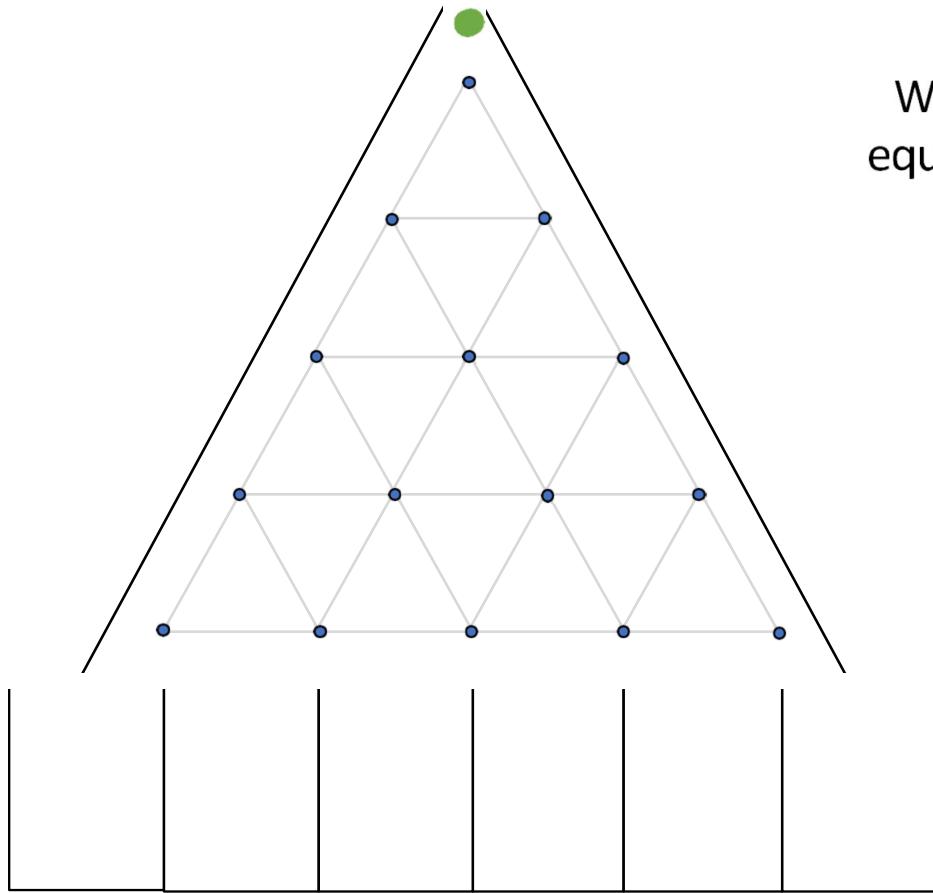


Galton Board Fun



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

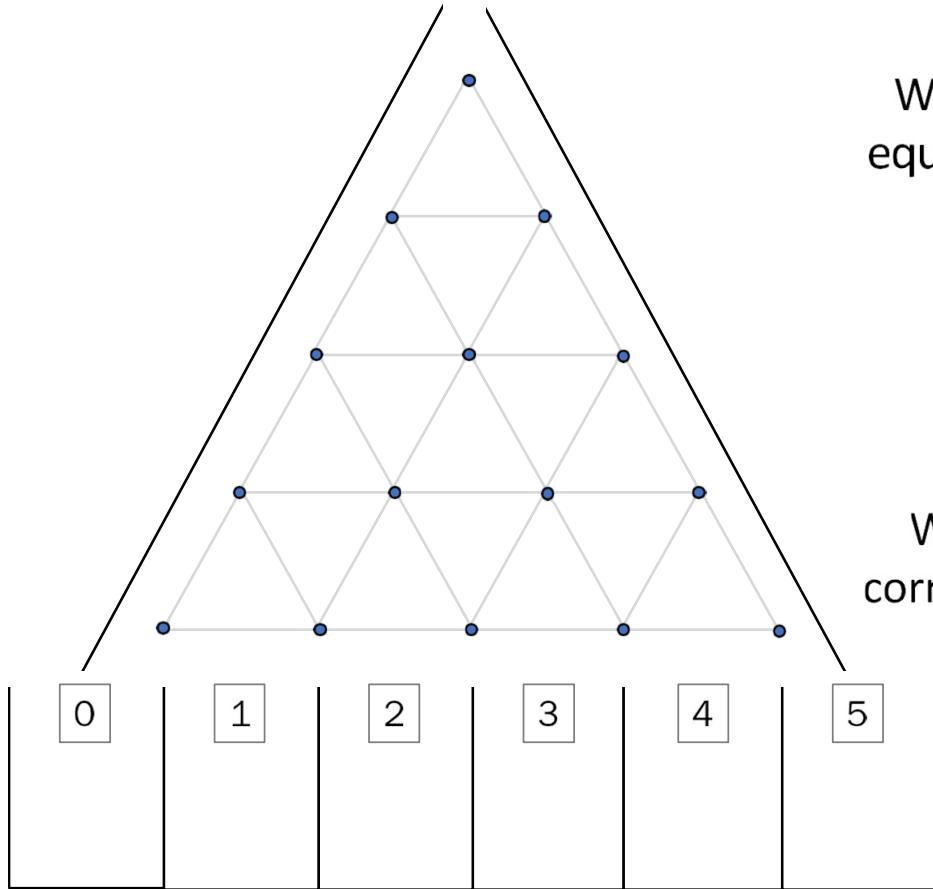
Galton Board Fun



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

Each pin represents an independent event.

Galton Board Fun

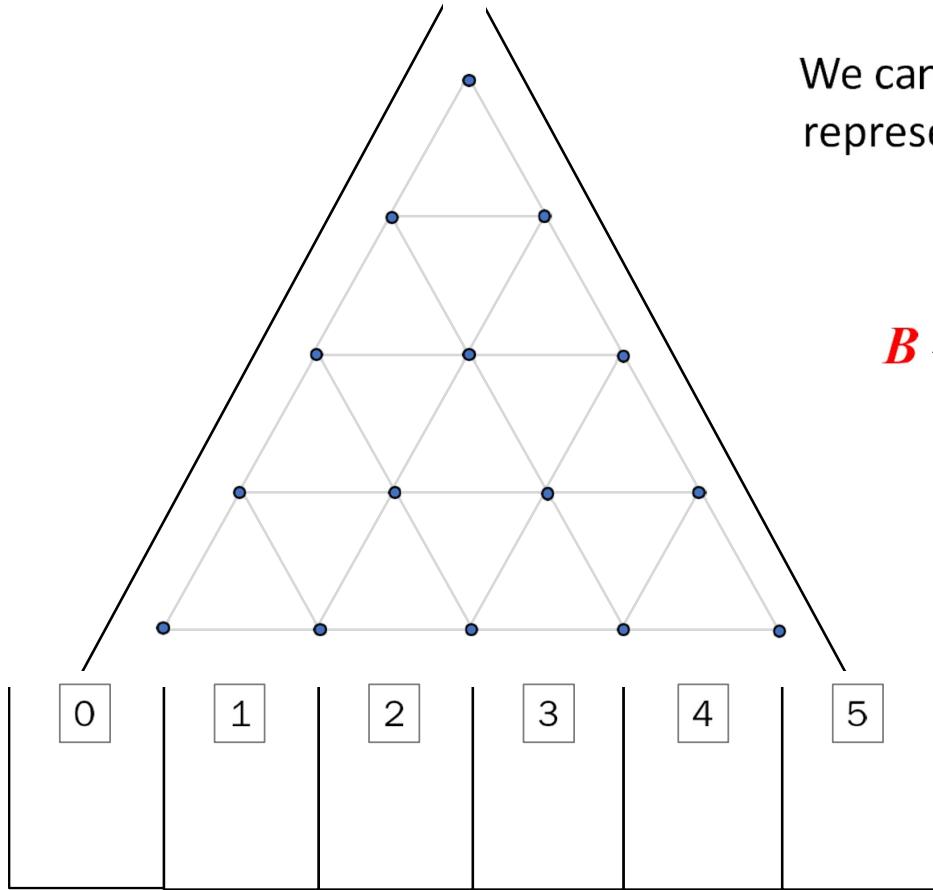


When a marble hits a pin, it has equal chance of going left or right.

Each pin represents an independent event.

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

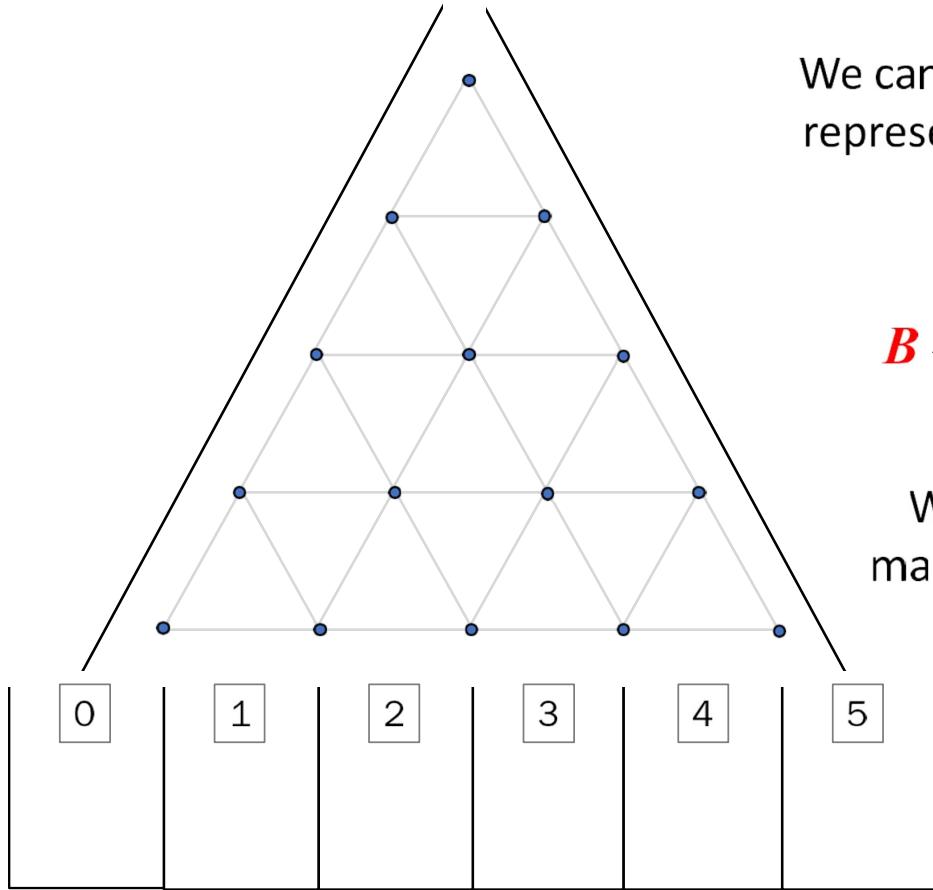
Galton Board Fun



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

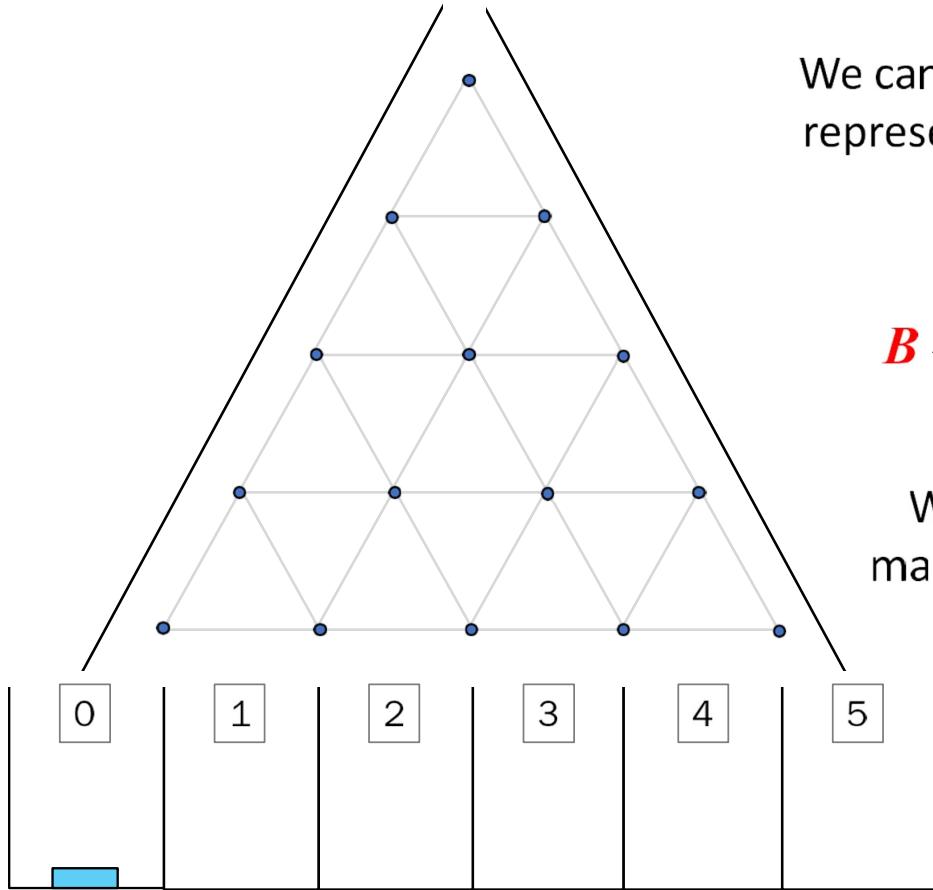


We can define a random variable (B) representing which bucket a marble lands in.

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

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

Galton Board Fun



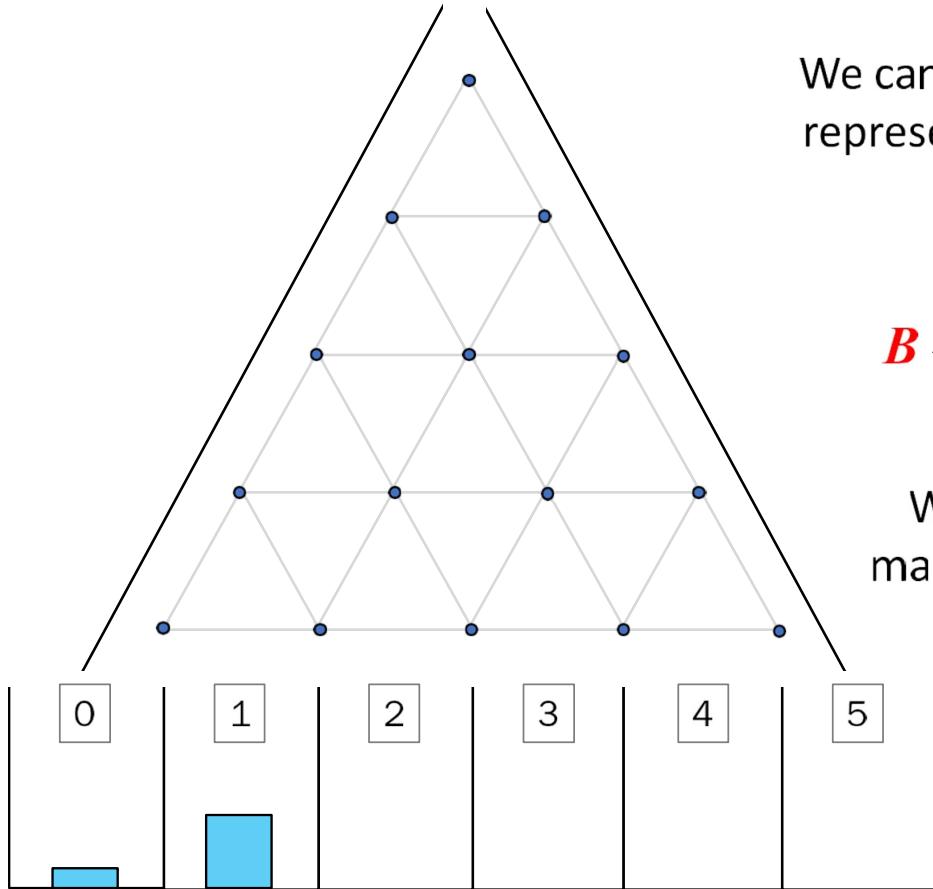
We can define a random variable (B) representing which bucket a marble lands in.

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

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.

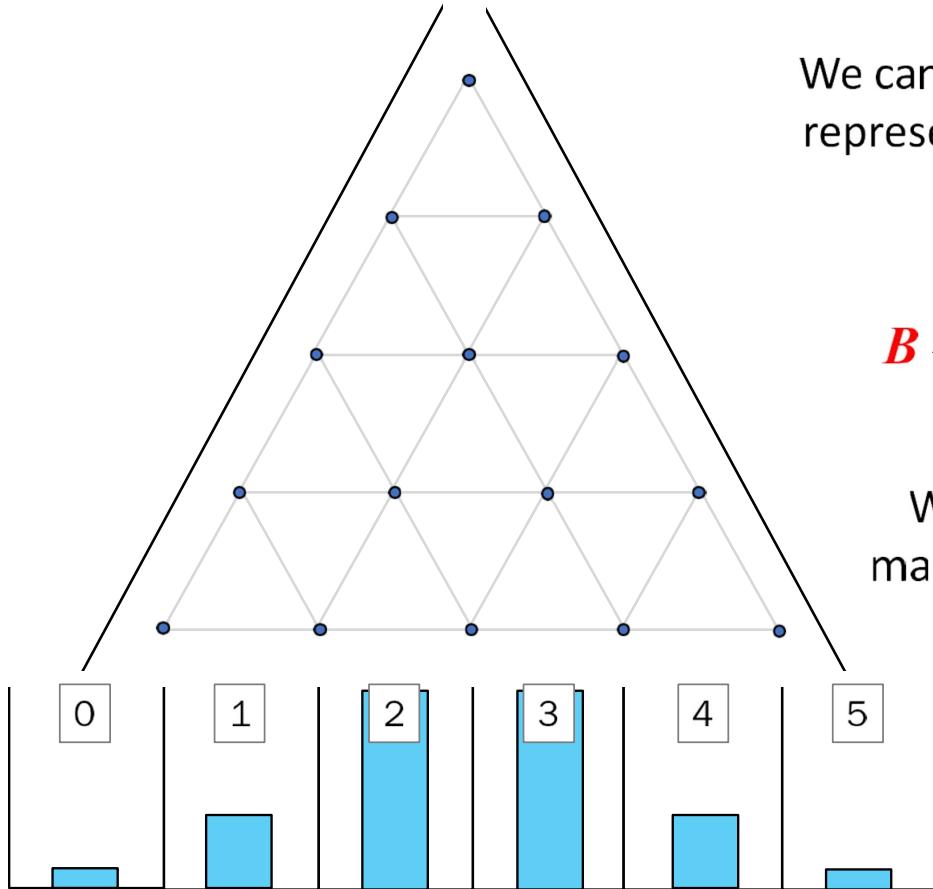
$$B \sim \text{Bin}(n = \text{levels}, p = 0.5)$$

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?

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

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

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:

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

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

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

/ / !
...

/ / !

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:

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

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

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

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

...

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?

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

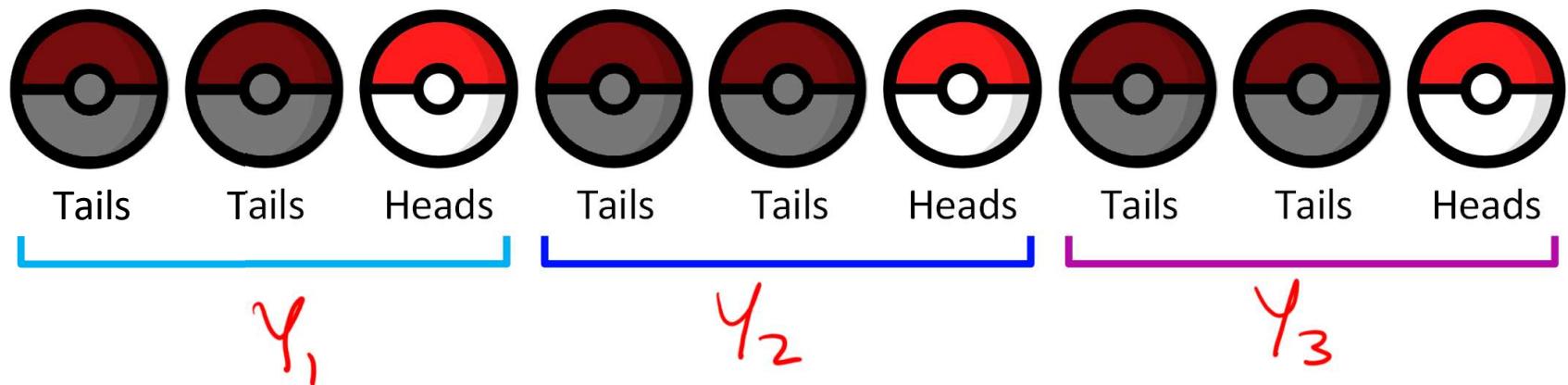
$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Random Variable Sums

The Negative Binomial



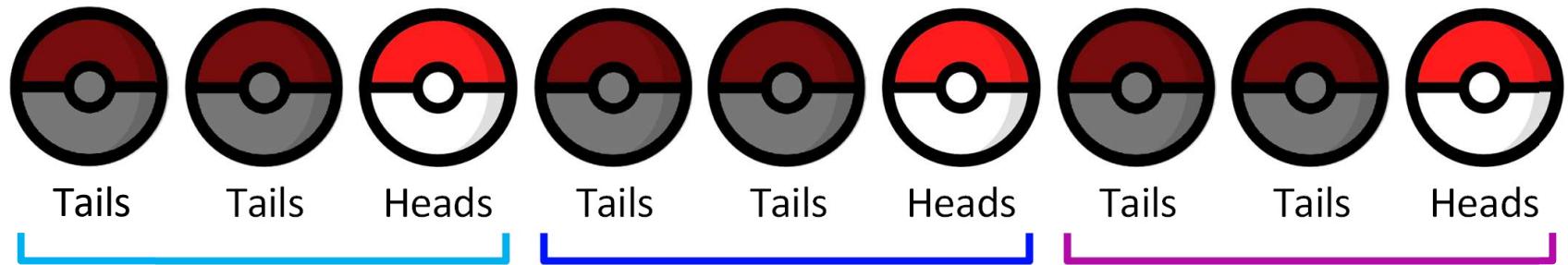
$\sigma = 3$



$$X = Y_1 + Y_2 + Y_3$$

Random Variable Sums

The Negative Binomial

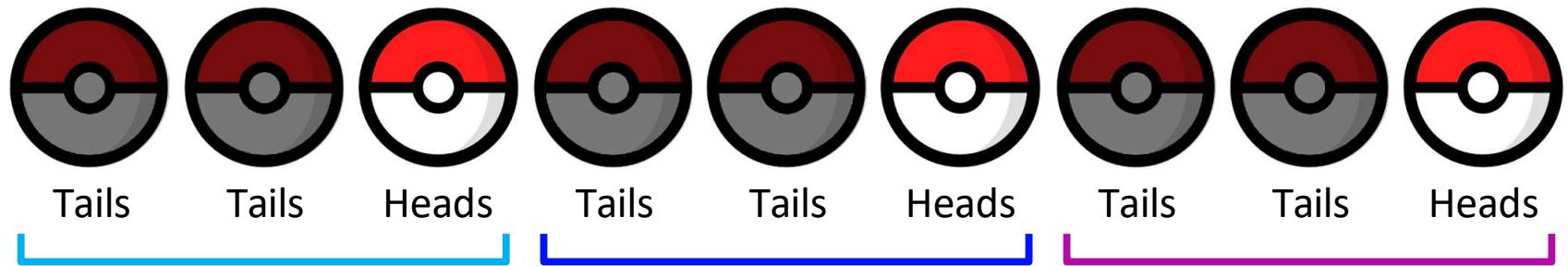


...is a sum of Geometric random variables

Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables

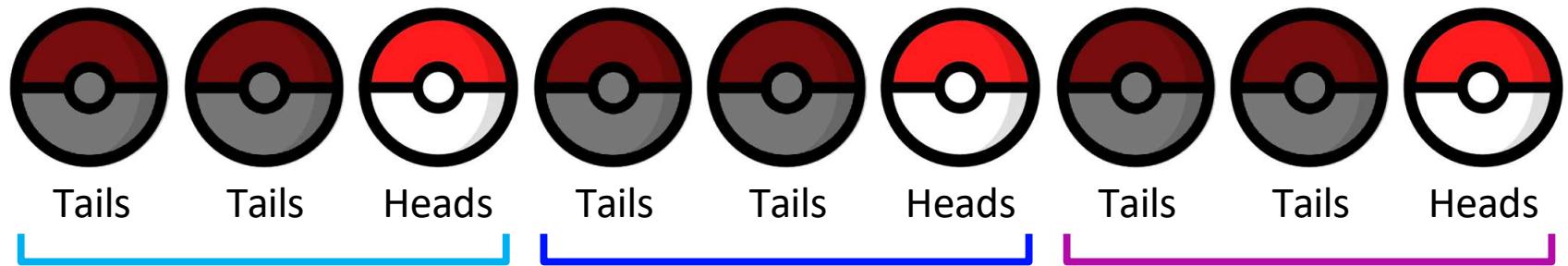


Let $X_1 \sim \text{Geo}(p = 1/3)$, $X_2 \sim \text{Geo}(p = 1/3)$, and $X_3 \sim \text{Geo}(p = 1/3)$.

Random Variable Sums

The Negative Binomial

...is a sum of Geometric random variables



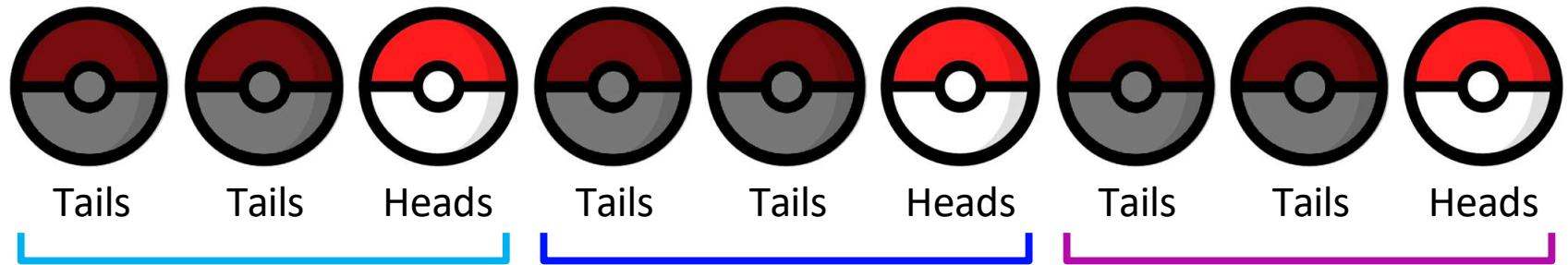
Let $X_1 \sim \text{Geo}(p = 1/3)$, $X_2 \sim \text{Geo}(p = 1/3)$, and $X_3 \sim \text{Geo}(p = 1/3)$.

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

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 $X_3 \sim \text{Geo}(p = 1/3)$.

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

$$Y = X_1 + X_2 + X_3$$

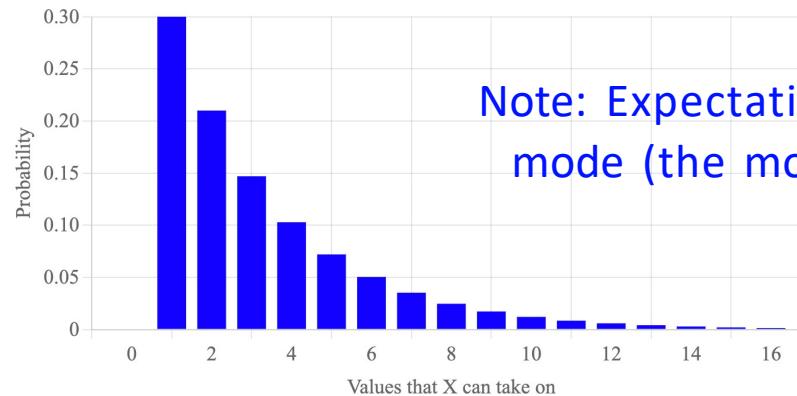
Expected Value of The Geometric

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

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

Expected Value of The Negative Binomial

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

The Negative Binomial



Tails



Tails



Heads



Tails



Tails



Heads



Tails



Tails



Heads

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

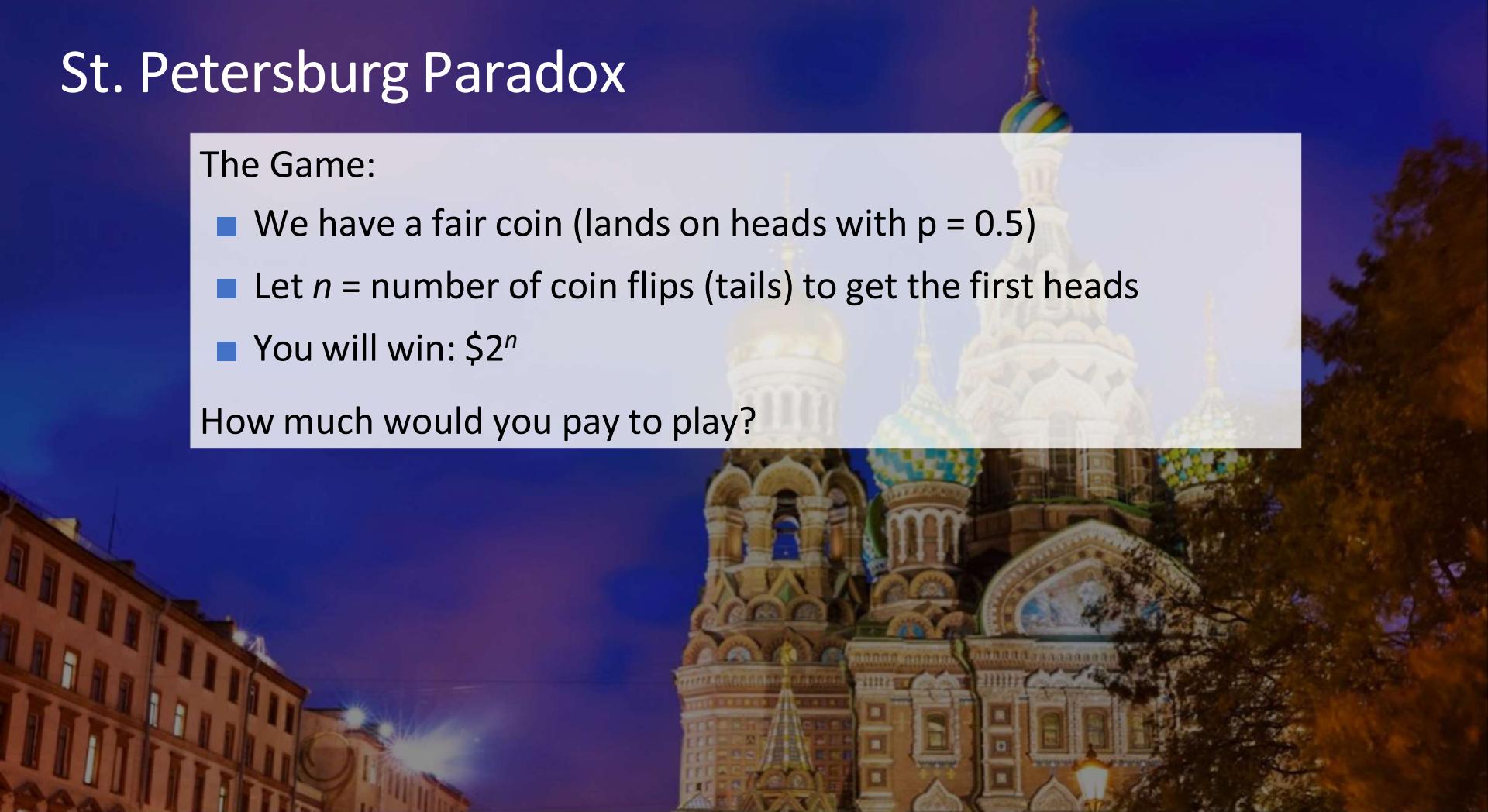
$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \\ &= \sum_{i=1}^r \frac{1}{p} = \frac{r}{p} \end{aligned}$$

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

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

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

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

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

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

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

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

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

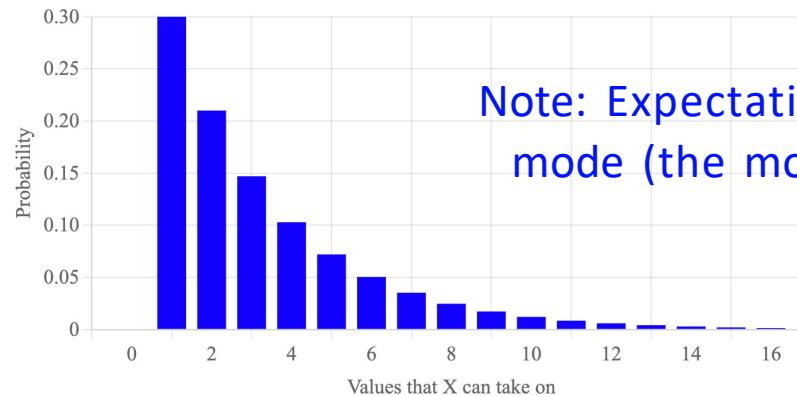
Expected Value of The Geometric

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)

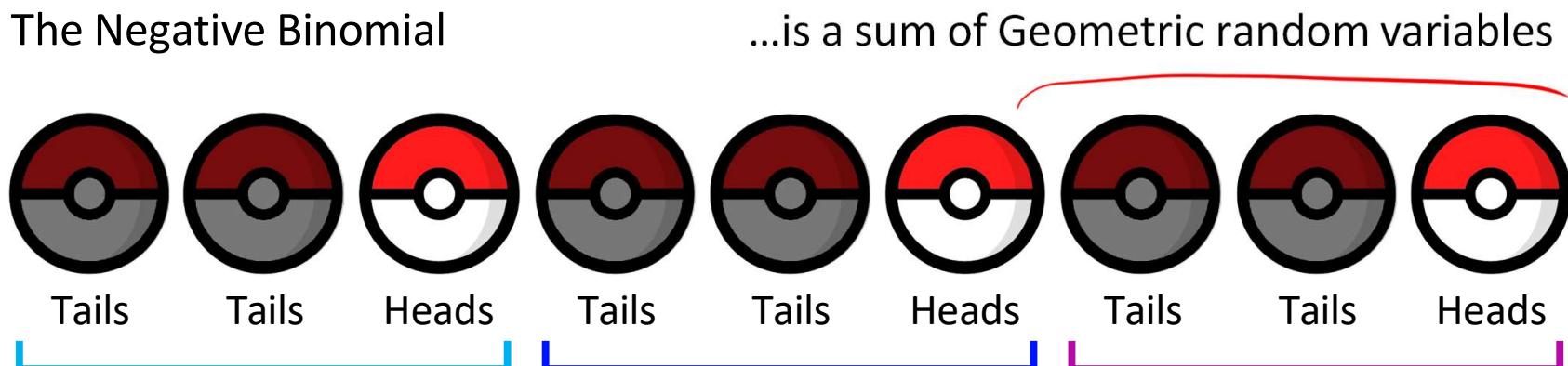
Expected Value of The Geometric

$$\begin{aligned} E[Y] &= \sum_{\substack{n=1 \\ n \neq 1}}^{\infty} n \cdot (1-p)^{n-1} \cdot p = \frac{1}{p} \\ &= 1 \cdot p + 2(1-p)p + 3(1-p)^2 p + 4(1-p)^3 p \\ &= p \left(1 + \overset{+}{2(1-p)} + \overset{-}{3(1-p)^2} + \overset{-}{\dots} \right) = \underset{\Sigma}{Sp} \\ &= p \left((-p) + \overset{+}{2(1-p)^2} + \overset{+}{3(1-p)^3} + \overset{+}{\dots} \right) = \underset{\Sigma}{Sp} \\ &= \end{aligned}$$

Recall SEE math --

Expected Value of The Negative Binomial

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



Expected Value of The Negative Binomial

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

Let $\underline{X_i \sim \text{Geo}(p)}$, 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 .

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

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

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

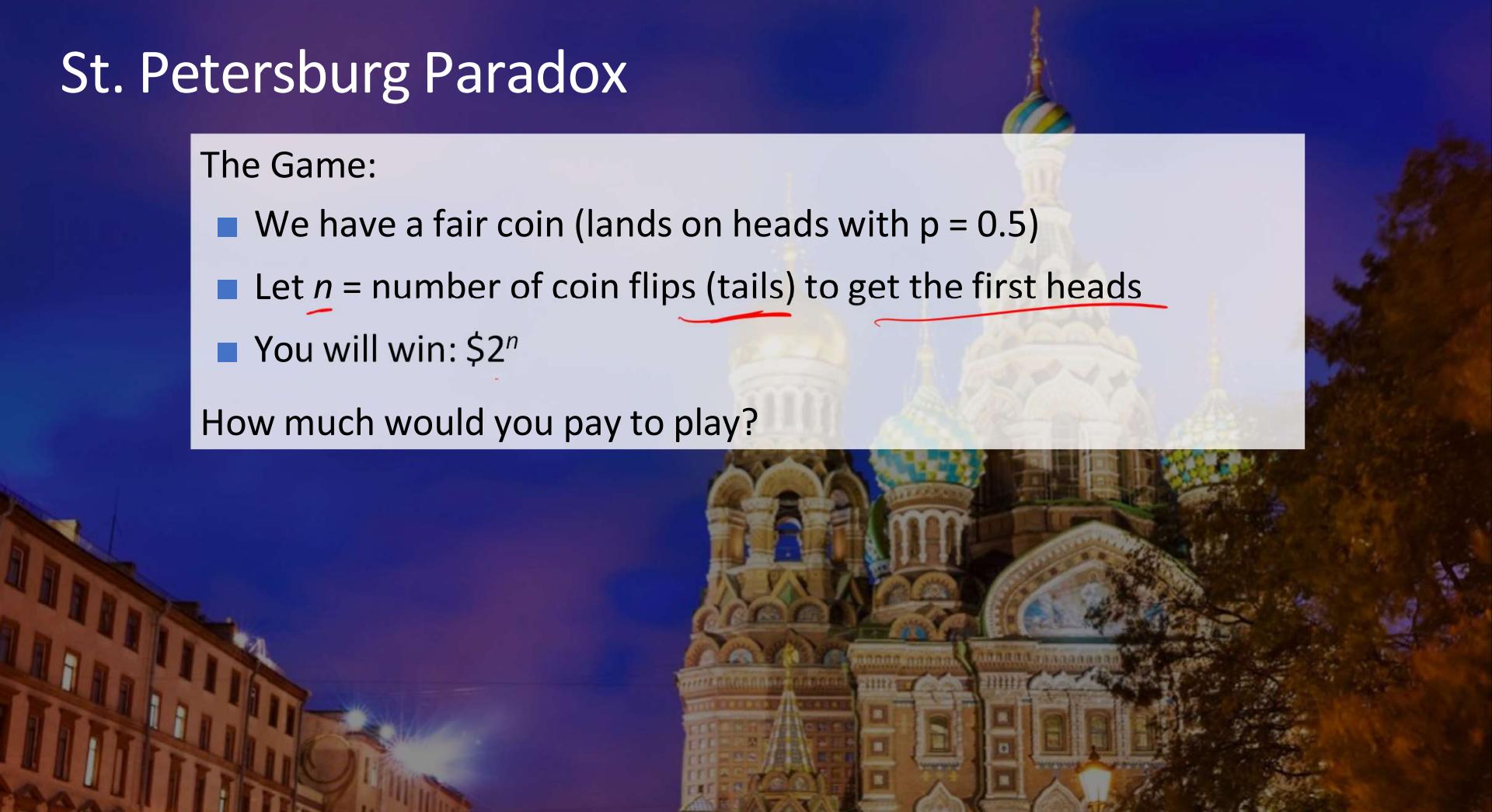
$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^r X_i\right] \\ &= \sum_{i=1}^r E[X_i] \\ &= \sum_{i=1}^r \frac{1}{p} = \frac{r}{p} \end{aligned}$$

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?

$$E(g(x))$$

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

Let X be your winnings.

$$g(x) = 2^x$$

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

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

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

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

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

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

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

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

$$E[X] = p$$

$$Y \in \{\tau, \tau+1, \dots, \infty\}$$

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

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

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

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

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

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

$$\overline{E[Y]} = n \cdot p$$

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

Variance of Classic Random Variables

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

Homework

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

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

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

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

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

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

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

Poisson Random Variable

Expected # of auto's in an hour = 10

Time interval = 5 minutes.

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

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



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

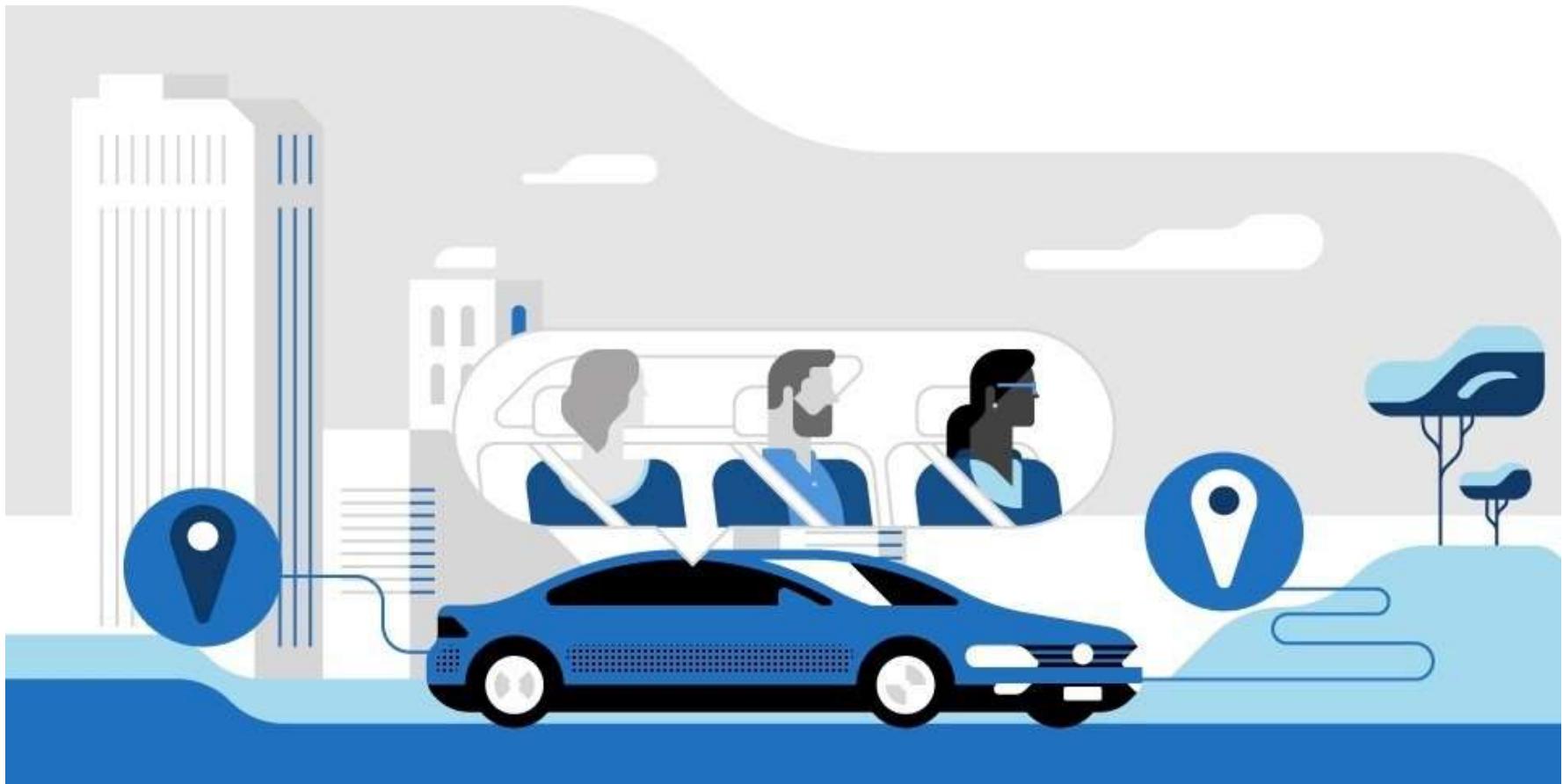
$$n = 300$$

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

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

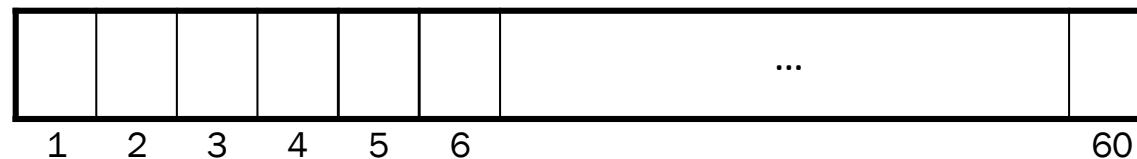


Probability of k Requests From This Area Next Minute



Probability of k Requests From This Area Next Minute

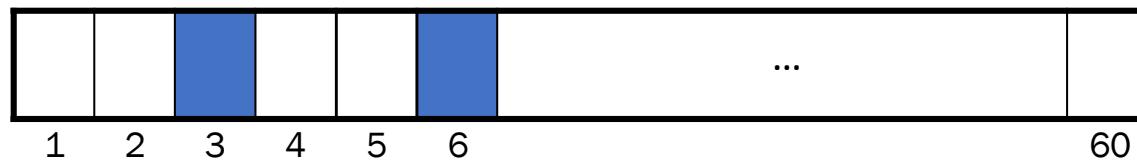
Idea: we can break a minute down into 60 seconds...



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...

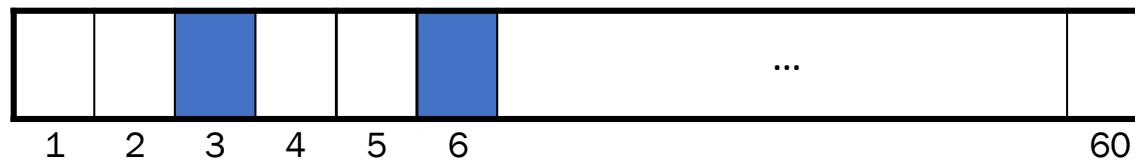


At each second, you either get a request or don't.

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...



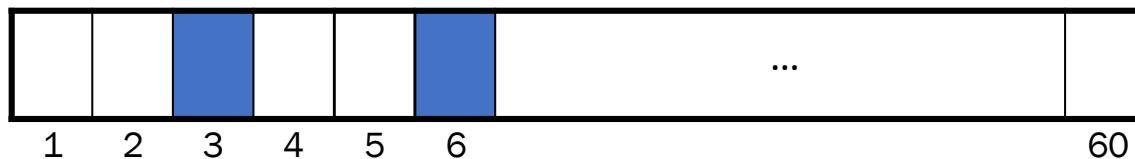
At each second, you either get a request or don't.
Let X be the number of requests in a minute.

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

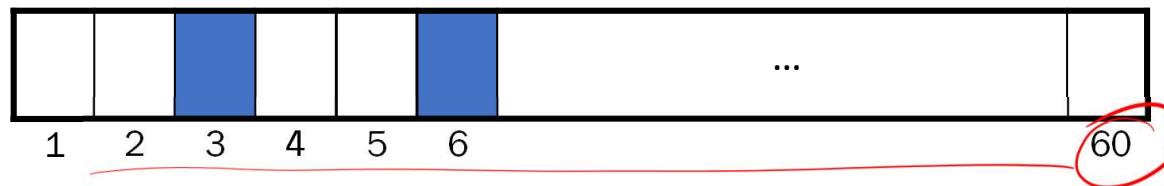
$$X \leftarrow \text{Bin}(n = 60, p = 5/60)$$

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

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

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

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

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

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

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

$$X \leftarrow \text{Bin}(n = 60000, p = \frac{\lambda}{n})$$
$$p = \frac{\lambda}{n} = \frac{5}{60000}$$
$$P(X = k) = \binom{60000}{k} \left(\frac{5}{60000}\right)^k \left(1 - \frac{5}{60000}\right)^{60000-k}$$

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

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

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

Probability of k Requests From This Area Each Minute

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

$\boxed{\frac{e^{-\lambda} \lambda^k}{k!}}$

The Poisson Random Variable

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

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

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.

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

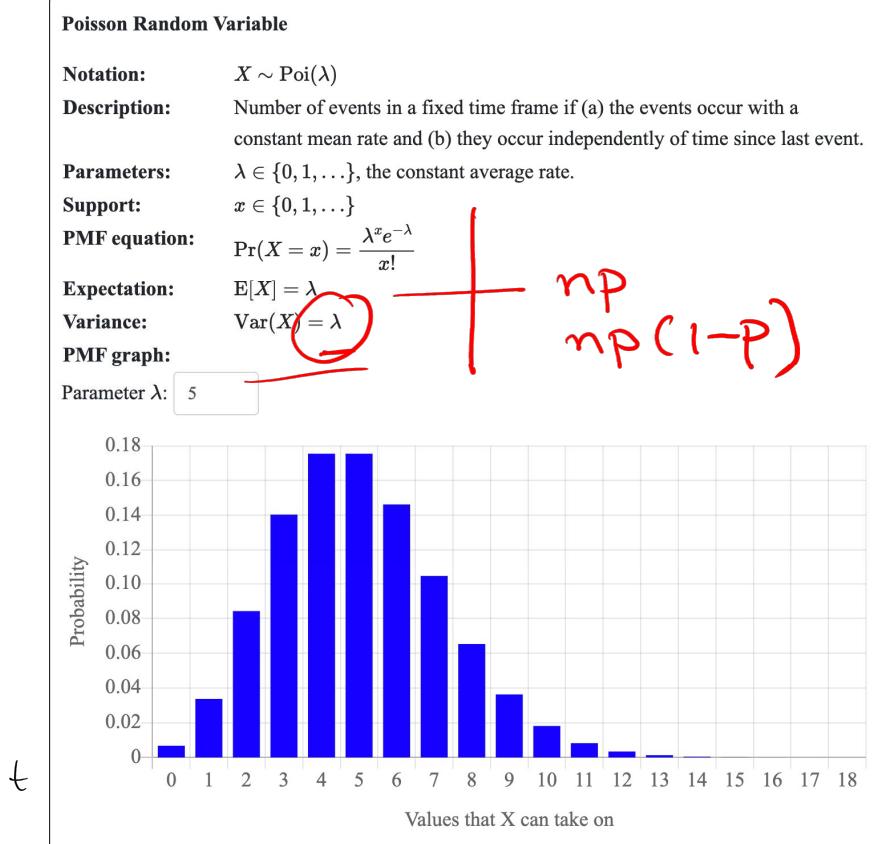
Is Lambda All You Need? Yes

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

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

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

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

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

Moment Generating Function

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

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

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

$$\frac{\partial \phi(t)}{\partial t} = \frac{\partial}{\partial t} \sum_x e^{tx} p(x) = \sum_x x e^{tx} p(x) \quad \left. \begin{array}{l} | \\ t=0 \end{array} \right. = \sum_x x p(x) \\ = E(x)$$

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

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

MGF of Poisson distribution

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

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

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

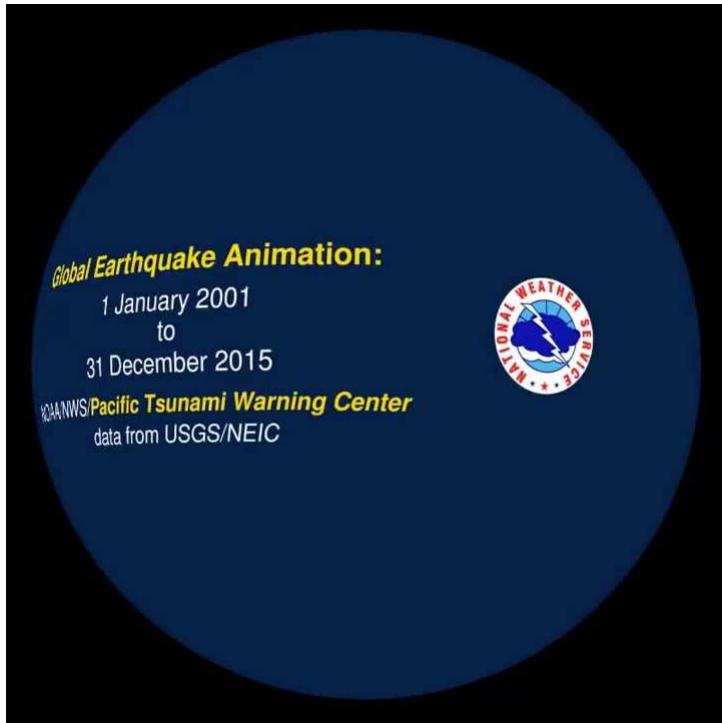
$$\phi'(t) = e^{-\lambda(1-e^t)} \cdot \lambda e^t \Big|_{t=0}$$

$$\begin{aligned}\phi''(t) &= \frac{d}{dt} e^{-\lambda + \lambda e^t + t} = \lambda e^{-\lambda + \lambda e^t + t} (\lambda e^t + 1) \Big|_{t=0} \\ &= \lambda + \lambda^2 = E(X^2)\end{aligned}$$

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

$$= \lambda$$

Example: Earthquakes



Bulletin of the Seismological Society of America

Vol. 64

October 1974

No. 5

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

BY J. K. GARDNER and L. KNOPOFF

ABSTRACT

Yes.

Earthquakes

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

Here's the PMF for \underline{X} :

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

\underline{X} is a Poisson!
What is $E[X]$ (1)?

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

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

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 $P(X < 5)$?



$$\lambda = \frac{120}{60} = 2 \text{ average # of requests per second}$$

$$P(X < 5) = \sum_{x=0}^4 e^{-\lambda} \frac{\lambda^x}{x!}$$

/



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 $P(X < 5)$?

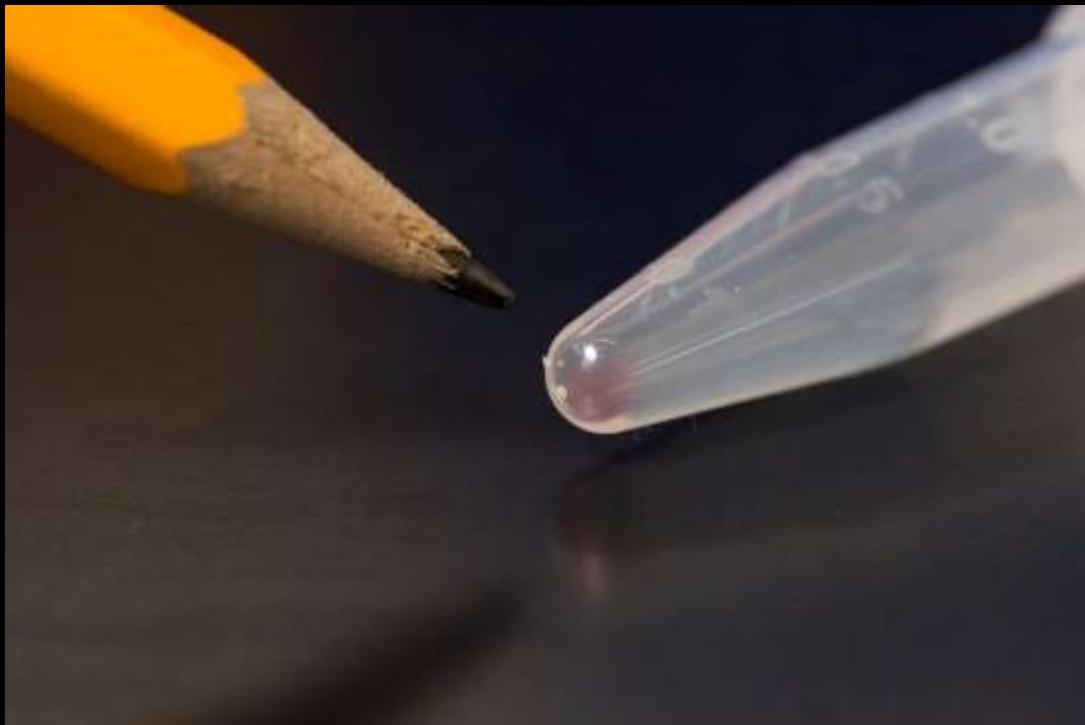


$$X \leftarrow \text{Poi}(\lambda = 2)$$



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.

$$\sum_{x=0}^{10^8} \binom{n}{x} p^x (1-p)^{n-x} = X \sim \text{Bin}(10^8, 10^{-6})$$

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.

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

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

Approximating Binomial With Poisson: General Rule

The Poisson approximates the Binomial well when:

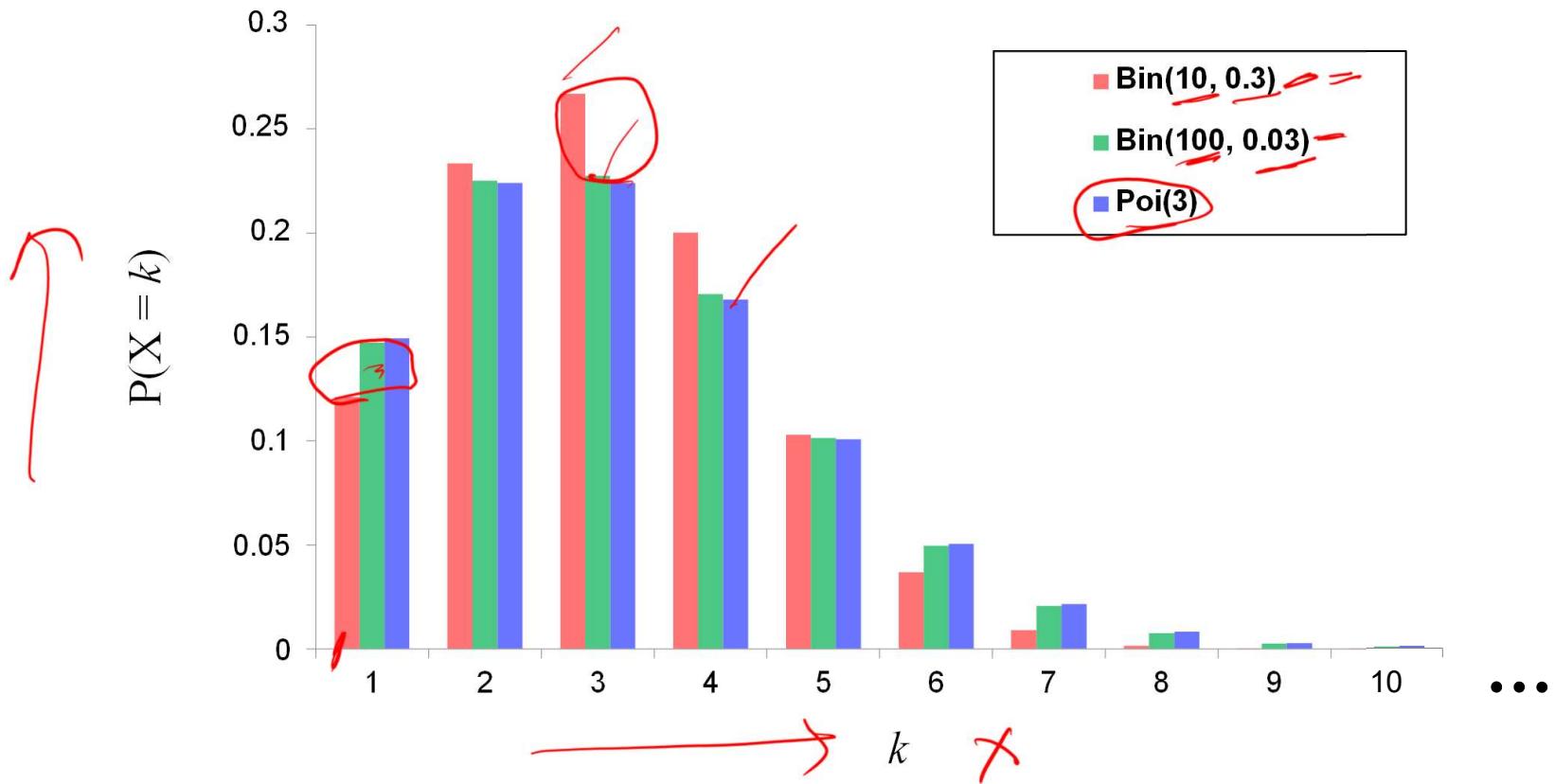
1. n is large ✓
2. p is small ✓
3. Therefore, $\lambda = np$ is "moderate" ↗

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$
- $n > 100$ 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 ?



Special Continuous Random Variables

Uniform Random Variable

- X is uniformly distributed between α and β

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

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

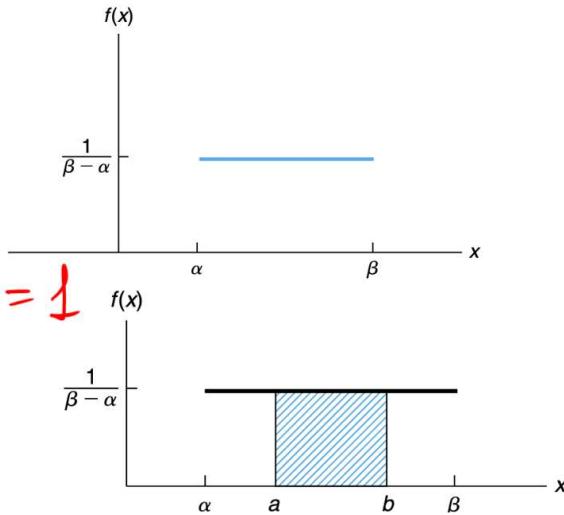
$$\int_{x=\alpha}^{\beta} p(x) dx = 1$$

$$= \int_{x=\alpha}^{\beta} p(x) dx = \int_{\alpha}^{\beta} \frac{1}{\beta - \alpha} dx = 1$$

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

- $E[X] = \frac{b - a}{\beta - \alpha}$

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



Variance of uniform random variable

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

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

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 $[R, x_i]$

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

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

- Initialize R with first k elements.
- Foreach 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_i to R_i after seeing i elements is $\frac{k}{i}$

By induction

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

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

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

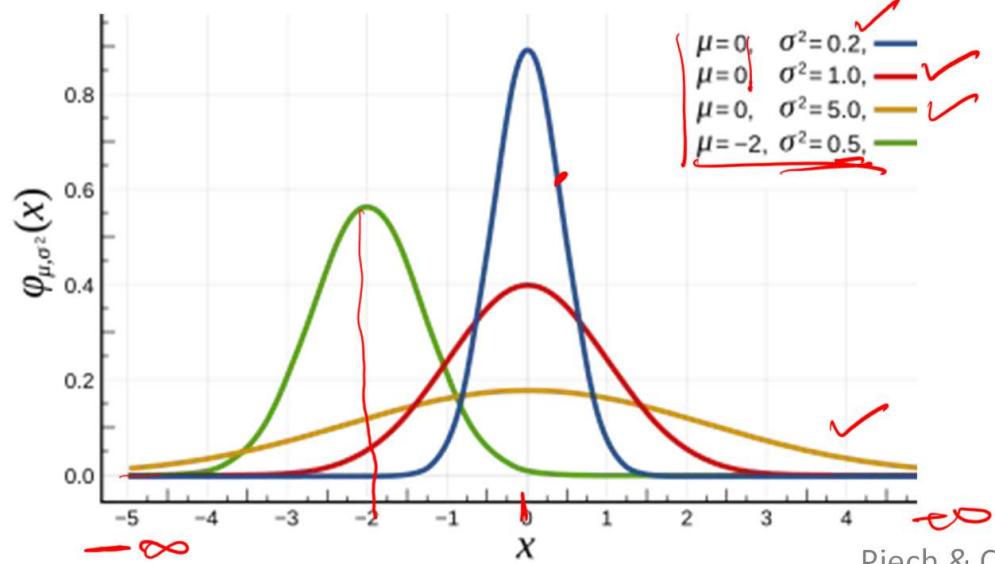
Normal (Gaussian) Random Variable

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

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

mean
variance

density



Piech & Cain, CS109, Stanford University

Normal (Gaussian) Random Variable

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

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

mean
↓
variance
↓

PDF:

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

Piech & Cain, CS109, Stanford University

Anatomy of a The Normal PDF

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

distance to the mean
(makes the PDF symmetric around the mean)

a constant:
makes the integral over all possible outcomes sum to 1

...normalized by the variance

Piech & Cain, CS109, Stanford University

Expected value of a normal distribution

Verify that μ is the expected value of
 $X \sim N(\mu, \sigma^2)$

$$\overline{E[(X-\mu)]} = E(X) - \mu$$
$$\int_{-\infty}^{+\infty} (x-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \left. \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma(\sigma^2)} \right|_{-\infty}^{+\infty} = 0$$

$$E[(X-\mu)] = 0 \Rightarrow E(X) = \mu$$

Variance

$$\begin{aligned} E((X - \mu)^2) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-(y^2/2)} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (y)(ye^{-(y^2/2)}) dy \quad \begin{matrix} u \\ v \end{matrix} \quad \begin{matrix} \nearrow u \\ \searrow v \end{matrix} \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left(-ye^{-y^2/2} \right)_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-y^2/2} dy \right] \quad \begin{matrix} \int udv = uv - \int vdu \\ \int ye^{-y^2/2} dy = -e^{-y^2/2} \end{matrix} \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2 \end{aligned}$$

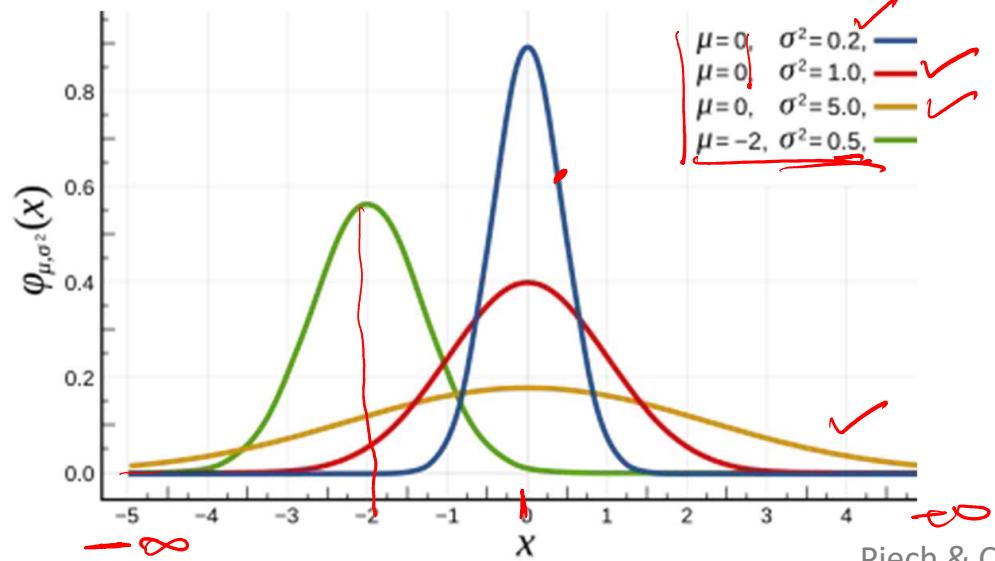
Normal (Gaussian) Random Variable

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

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

mean
variance

density



Piech & Cain, CS109, Stanford University

Normal (Gaussian) Random Variable

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

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

mean
↓
variance
↓

PDF:

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

Piech & Cain, CS109, Stanford University

Anatomy of a The Normal PDF

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

distance to the mean
(makes the PDF symmetric
around the mean)

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
 $x \sim N(\mu, \sigma^2)$

$$\overline{E((x-\mu))} = E(x) - \mu$$
$$\int_{-\infty}^{\infty} (x-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \left. \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma(\sigma^2)} \right|_{-\infty}^{+\infty} = 0$$

$$E((x-\mu)) = 0 \Rightarrow E(x) = \mu$$

Variance

$$\cancel{E((X - \mu)^2)} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx \quad \checkmark$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-(y^2/2)} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\cancel{y})(ye^{-(y^2/2)}) dy$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-y^2/2} dy \right]$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \underline{\sigma^2}$$

$$\int u dv = uv - \int v du$$

$$\int ye^{-y^2/2} dy = -e^{-y^2/2}$$

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

$$F_Y = P(Y \leq y) \quad f_Y = \frac{d}{dy} F_Y(y)$$

$$F_X = P(X \leq x) \quad f_X = \frac{d}{dx} F_X(x)$$

Let $a > 0$

$$P(Y \leq y) = P(ax + b \leq y)$$

$$\left[P(Y \leq y) = P\left(X \leq \frac{y-b}{a}\right) \right] \Rightarrow \begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) \\ f_Y(y) &= \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) \frac{d}{dx} F_X\left(\frac{y-b}{a}\right) \\ &= f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} \end{aligned}$$

$$f_x\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\mu a+b))^2}{2\sigma^2 \cdot a^2}}$$

$$f_y(y) = f_x\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} = \frac{1}{\sqrt{2\pi}\sigma^2 a} e^{-\frac{(y-(\tilde{\mu} a+b))^2}{2a^2\sigma^2}}$$

$\Rightarrow Y \sim N(\mu a + b; \tilde{\sigma}^2)$ if $a > 0$

$$\begin{aligned} a < 0 \\ F_Y(y) &= P(Y \leq y) = P(ax + b \leq y) = P(X \geq \frac{y-b}{a}) \\ &= 1 - F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

$$Y \sim N(a\mu + b; \sigma^2 a^2)$$

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)

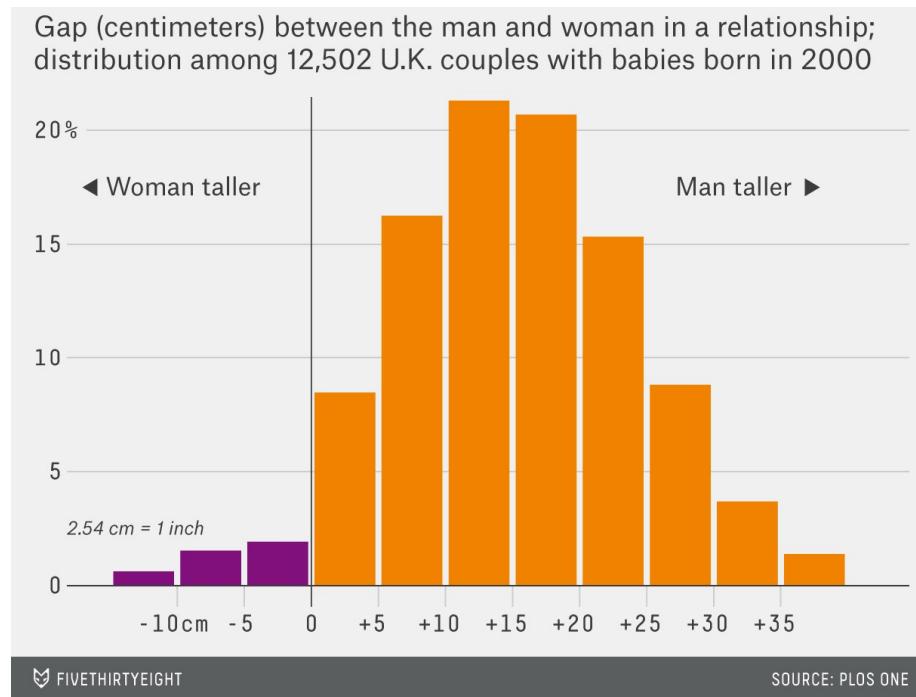
- German mathematician
- Sort-of invented the normal distribution
- Also astronomer, geologist, physicist
- Super influential in a lot of fields



Piech & Cain, CS109, Stanford University

Why the Normal?

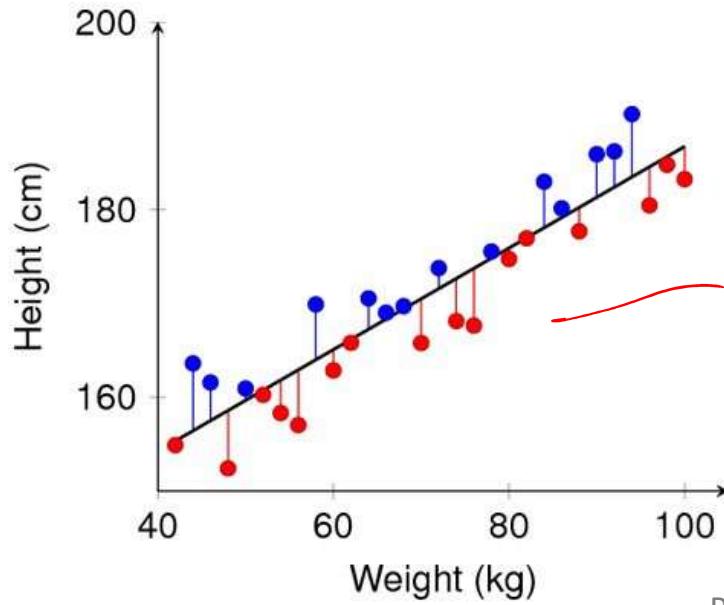
- Common for natural phenomena: human height, weight, shoe sizes, etc.



Piech & Cain, CS109, Stanford University

Why the Normal?

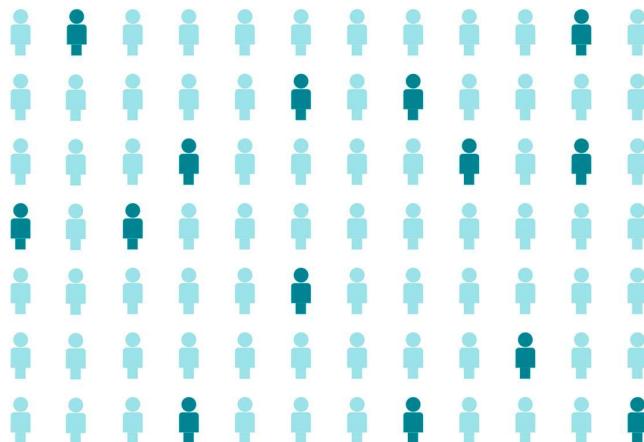
- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression



Piech & Cain, CS109, Stanford University

Why the Normal?

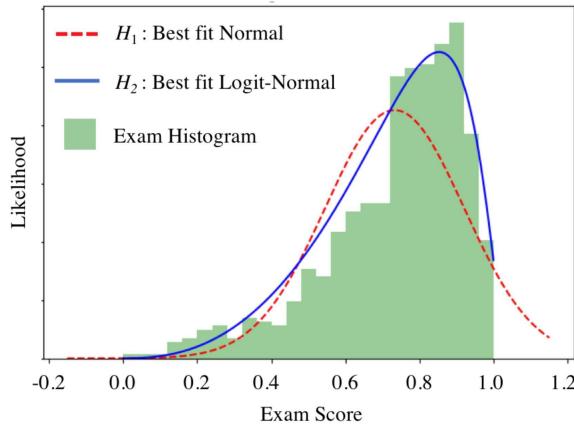
- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for 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

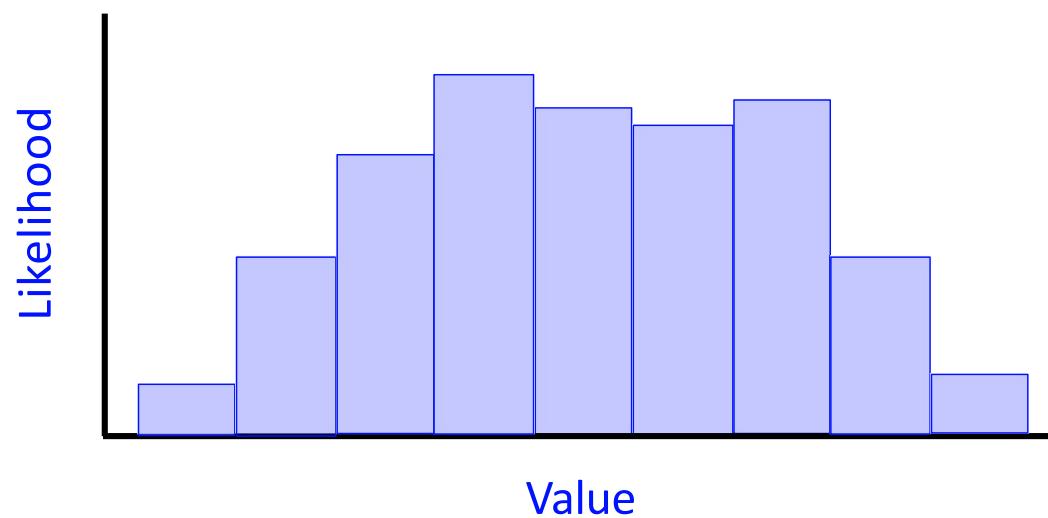
People also just assume things are normally distributed a lot.

- They can do this in part because the Normal is so common
- But there's a deeper reason to it...

Piech & Cain, CS109, Stanford University

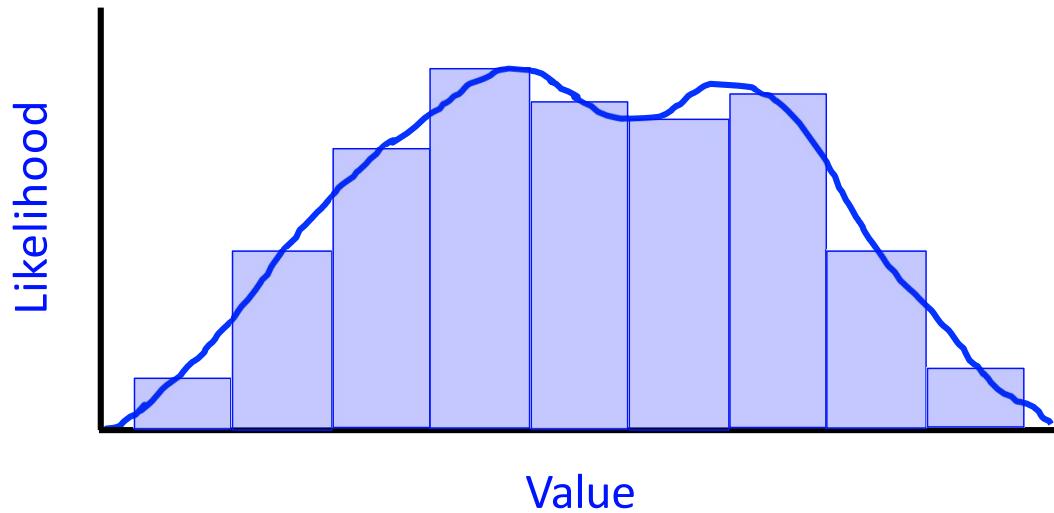


When We Fit Models To Data, We Try To Keep It Simple



Piech & Cain, CS109, Stanford University

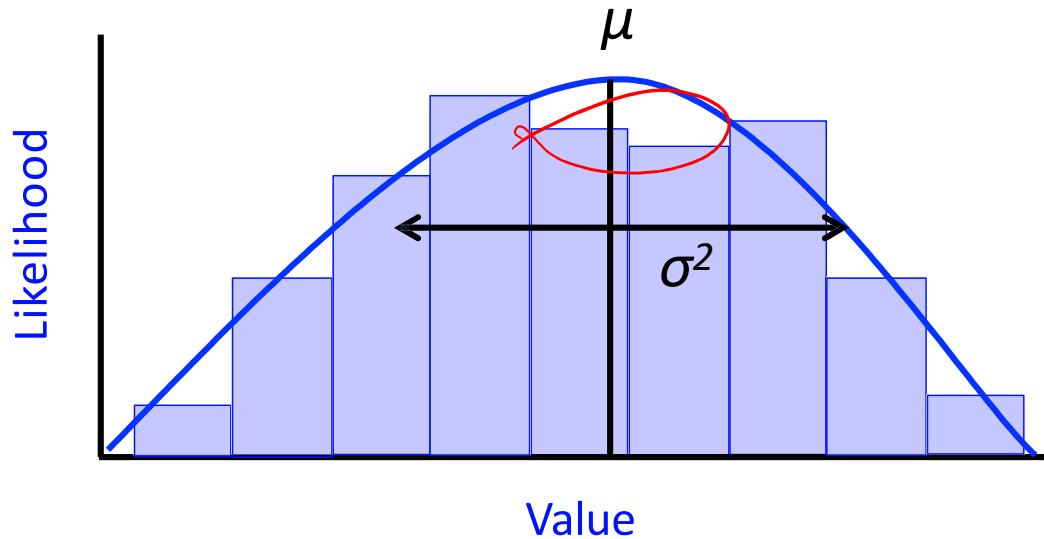
When We Fit Models To Data, We Try To Keep It Simple



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

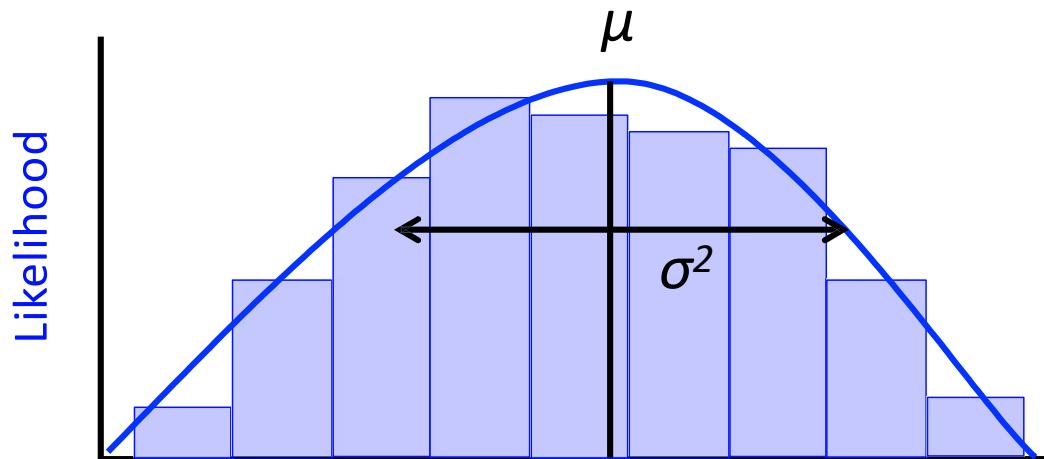
When We Fit Models To Data, We Try To Keep It Simple



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: $X, f(x)$

$$\text{Entropy}(X) = - \int_X f(x) \log f(x) dx$$

$\text{Ent}(X) \rightarrow \text{discrete } E(X) \leq 0 \quad E(X) > 0$

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 $p(x_1) \cdot p(x_k) - \sum_{x_i} p(x_i) \log p(x_i)$
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 logs) 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.

$$\text{Ent}_\sigma(x) = - \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} f(x) \right] \log e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx - \log \sqrt{2\pi}\sigma$$
$$= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} [f(x)] dx + \log \sqrt{2\pi}\sigma \int f(x) dx$$
$$= \frac{\sigma^2}{2\sigma^2}$$
$$= \frac{1}{2} + \log \sqrt{2\pi}\sigma$$

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*

$$P(X \leq x) = \Phi(x) = F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-(z)^2/2} dz$$

$F(x)$

- 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 $X = \frac{Y - \mu}{\sigma}$ $X \sim N(0, 1)$

- $F_Y(y) = F_X\left(\frac{y - \mu}{\sigma}\right) = \Phi\left(\frac{y - \mu}{\sigma}\right)$

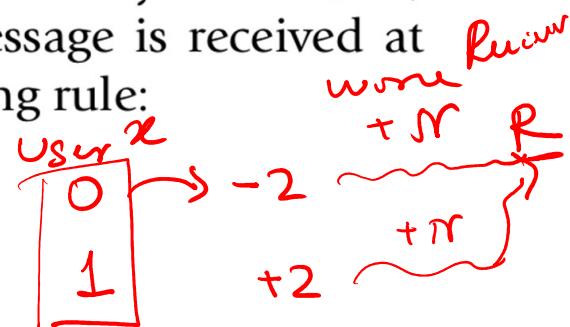
$$P(Y \leq y) = P(\sigma X + \mu \leq y) = P\left(X \leq \frac{y - \mu}{\sigma}\right) = F_X\left(\frac{y - \mu}{\sigma}\right)$$

from a library $\Rightarrow \Phi\left(\frac{y - \mu}{\sigma}\right)$

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.

$$P(y=0 | x=1) = P(N < -1.5) \quad N \sim \mathcal{N}(0, 1)$$

$$= \phi(-1.5) = 1 - \Phi(1.5) = 0.0668$$

$$P(y=1 | x=0) = P(N > 2.5)$$

$$= 1 - P(N \leq 2.5)$$

$$= 1 - \Phi(2.5) = 0.0062$$

$$P\{\text{error} | \text{message is "1"}\} = P\{N < -1.5\}$$

$$= 1 - \Phi(1.5) = .0668$$

and

$$P\{\text{error} | \text{message is "0"}\} = P\{N > 2.5\}$$

$$= 1 - \Phi(2.5) = .0062$$

Properties

MGF of $\underline{Z} \sim N(0,1)$

$$\begin{aligned} E(e^{t\underline{z}}) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-z^2/2 + tz} dz \end{aligned}$$

MGF of $X \sim N(\mu, \sigma^2)$

Properties

MGF of $Z \sim N(0,1)$

$$\begin{aligned} E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2 - 2tx)/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= e^{t^2/2} \end{aligned}$$

$$X = \sigma Z + \mu$$

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

$$\begin{aligned} E[e^{t\underline{X}}] &= E[e^{t\mu + t\sigma Z}] \\ &= E[e^{t\mu} e^{t\sigma Z}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{t\mu} e^{(\sigma t)^2/2} \\ &= \boxed{e^{\mu t + \sigma^2 t^2/2}} \end{aligned}$$

Sum of Gaussian Random Variables

- Let $\underline{Y} = \underline{X_1 + X_2 + \dots + X_n}$
 - Where each $\underline{X_i} \sim N(\mu_i, \sigma_i^2)$
 - What is the distribution of \underline{Y} ?
-
- $\underline{Y} \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$
 - Proof via MGF.

/ .

MGF of sum of Gaussians

$$\begin{aligned}
 \bullet E(e^{tY}) &= E_Y\left(e^{t(x_1 + x_2 + \dots + x_n)}\right) \\
 &= E_Y\left(e^{tx_1} \cdot e^{tx_2} \cdots e^{tx_n}\right) = \\
 &= E_{x_1}(e^{tx_1}) \cdots E_{x_n}(e^{tx_n}) \\
 &= \prod_{i=1}^n E_{x_i}(e^{tx_i}) = \prod_{i=1}^n e^{tu_i + t\sigma_i^2/2} \\
 &= e^{\sum_{i=1}^n tu_i + t\sum_i \sigma_i^2/2} \\
 &= e^{\boxed{tu + t^2\sigma^2}} \quad \text{where } \mu = \sum_i u_i \\
 &\rightarrow \text{MGF of } \mathcal{N}(\mu, \sigma^2) \quad \left[\begin{array}{l} \text{MGF} \& \text{density fn.} \\ \text{done on } i-1 \text{ corresponds} \end{array} \right]
 \end{aligned}$$

Exponential Random Variable

For any Poisson Process, the **Exponential** RV models *time until an event*:

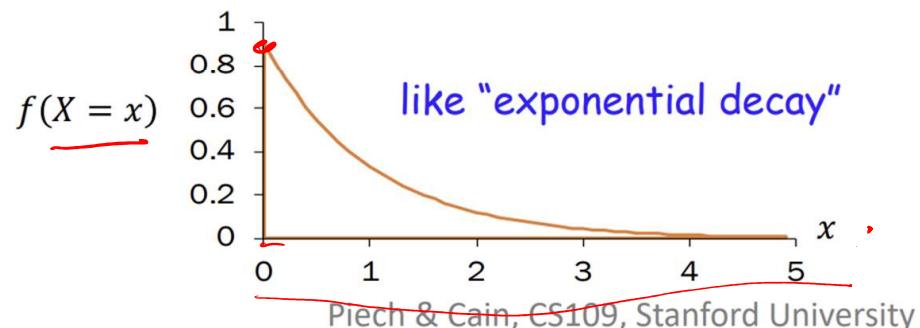
$$\underline{X \sim \text{Exp}(\lambda)}$$

PDF:

$$\underline{f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}}$$

Examples:

- Time until next earthquake
- Time until a ping reaches a web server
- Time until a Uranium atom decays

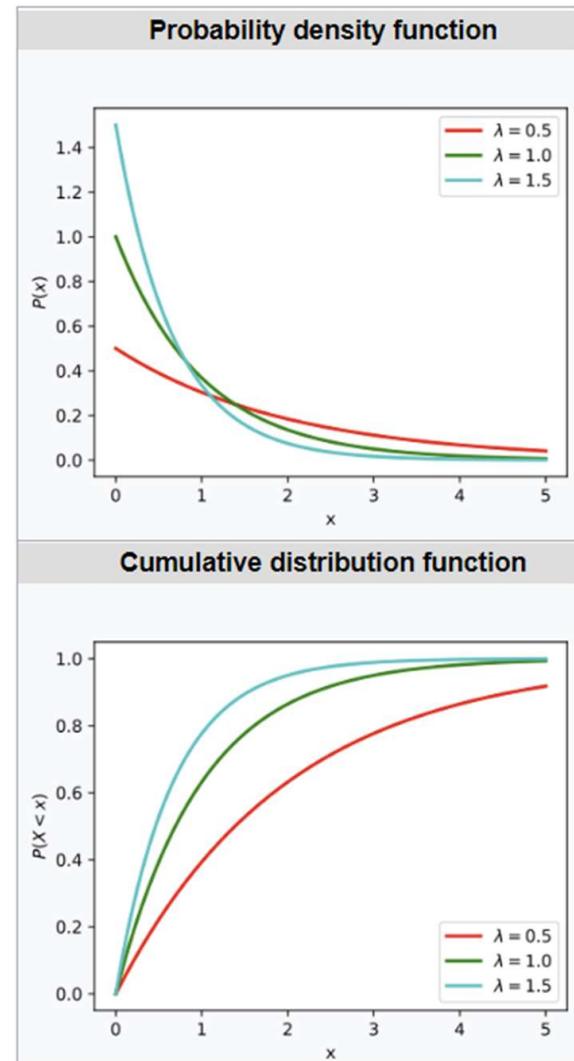


Cumulative Distribution function

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

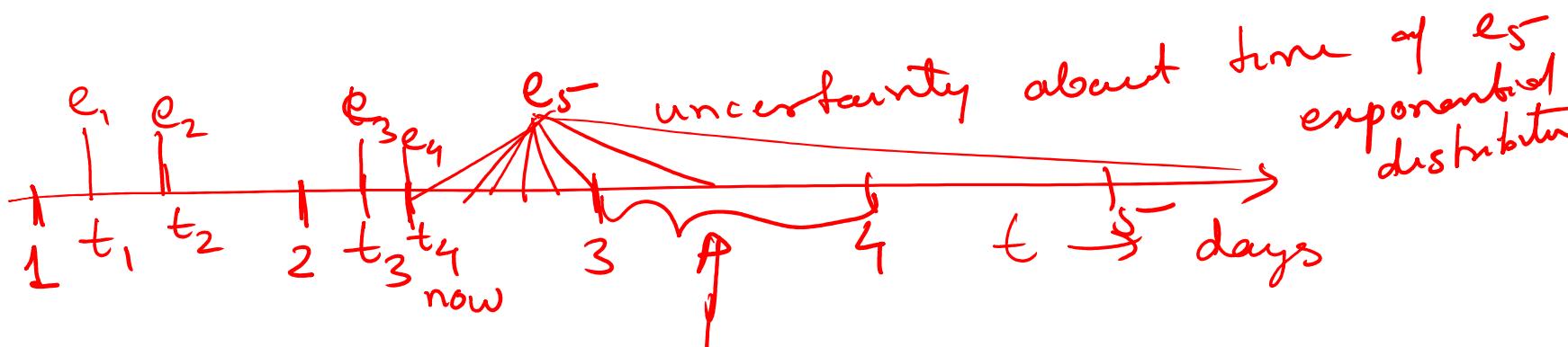
$$= \int_0^x \lambda e^{-\lambda y} dy$$

$$= 1 - e^{-\lambda x}, \quad x \geq 0$$



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

$$\phi(t) = \underline{E[e^{tX}]} \quad X \sim \text{exp}(\lambda)$$

$$= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx$$

$$= \lambda \int_0^\infty e^{-(\lambda-t)x} dx$$

$$= \frac{\lambda}{\lambda - t}, \quad t < \lambda$$

Differentiation yields

$$\underline{\phi'(t)} = \frac{\lambda}{(\lambda - t)^2}$$

$$\underline{\phi''(t)} = \frac{2\lambda}{(\lambda - t)^3} \cdot \frac{2\lambda}{\lambda^3}$$

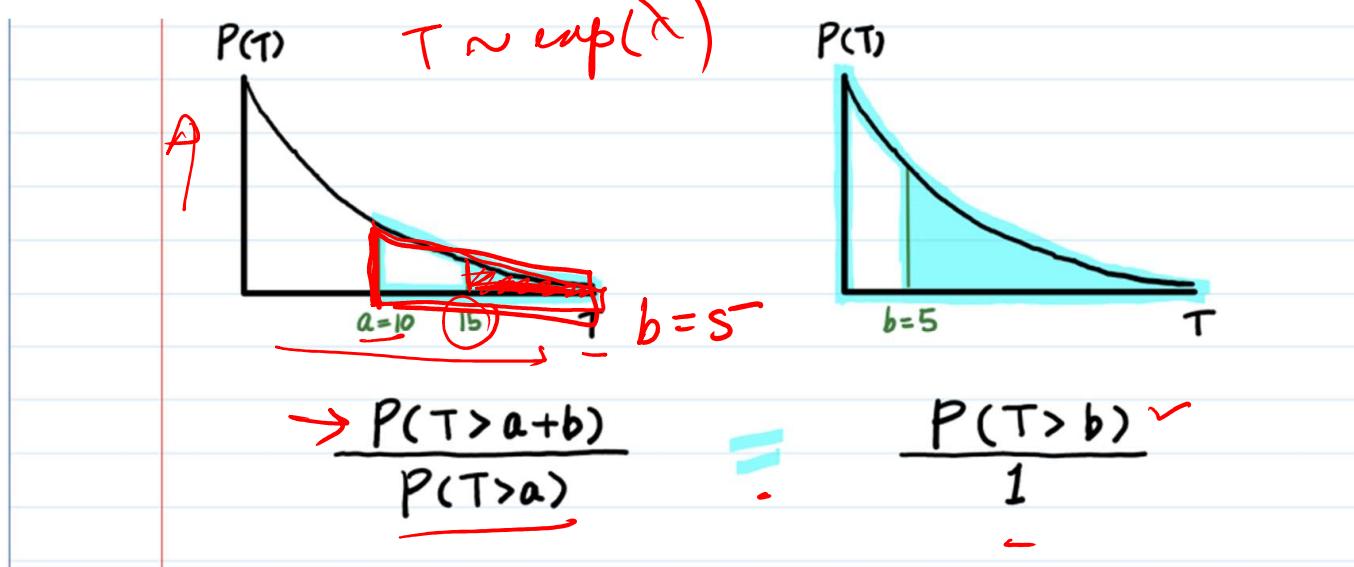
$$E[X] = \underline{\phi'(0)} = 1/\lambda$$

$$\begin{aligned} \underline{\text{Var}(X)} &= \underline{\phi''(0)} - (E[X])^2 \\ &= \underline{2/\lambda^2} - \underline{1/\lambda^2} \\ &= \underline{1/\lambda^2} \end{aligned}$$

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-7bdd08590e2a>

Proof of the memory-less property

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

$$X \sim \exp(\lambda)$$

$$\text{CDF}(x) = \frac{1 - e^{-\lambda x}}{P(X \leq x)} \quad t > 0$$

$$\begin{aligned} P(X > s+t | X > s) &= \frac{P(X > s+t, X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)} \\ &= \frac{1 - \text{CDF}_X(s+t)}{1 - \text{CDF}_X(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= 1 - \text{CDF}_X(t) \\ &= P(X > t) \end{aligned}$$

$\Rightarrow X$ 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 $G(2t) = G(t)^2$, $G(3t) = G(t)^3$, ..., $G(kt) = G(t)^k$.

Similarly, from the above we see that $G(\frac{t}{2}) = G(t)^{\frac{1}{2}}$, ..., $G(\frac{t}{k}) = G(t)^{\frac{1}{k}}$.

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.

$$X \sim \exp(\lambda) \quad \lambda = \frac{1}{10000} \quad E(X) = \frac{1}{\lambda}$$

$$P(X > 5000) = e^{-\lambda \cdot 5000} = e^{-\frac{1}{10000} \cdot 5000}$$

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 $\sum_{i=1}^n \lambda_i$.

$$Y = \min(X_1, X_2, \dots, X_n) \quad X_i \sim \text{exp}(\lambda_i)$$

$$P(Y > b) = P(X_1 > b, X_2 > b, \dots, X_n > b)$$

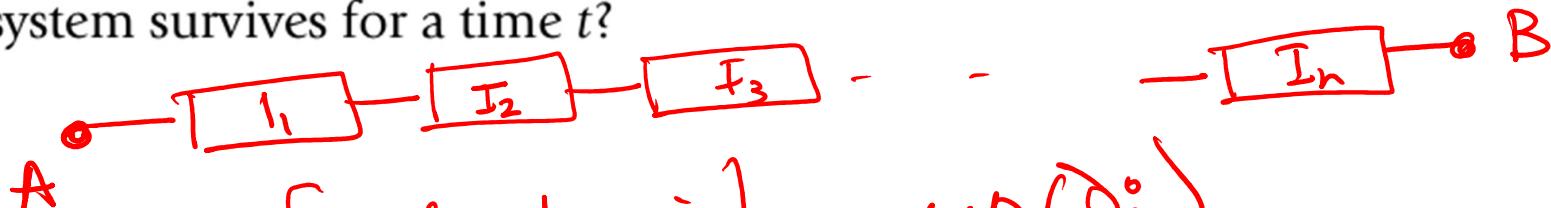
$$\begin{aligned} P(\min(X_1, \dots, X_n) > b) &= \prod_{i=1}^n P(X_i > b) \\ &= \prod_{i=1}^n e^{-\lambda_i b} = e^{-b \left(\sum_{i=1}^n \lambda_i \right)} \end{aligned}$$

$$\Rightarrow Y \sim \text{exp}\left(\sum_{i=1}^n \lambda_i\right)$$

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 ?



$$[I_j \text{ functioning}] \sim \exp(-\lambda_j t)$$

$Y = A \text{ to } B \text{ connection is functioning}$

$$P(Y \geq r) = P(\min(I_1, I_2, \dots, I_n) > r) = e^{-r \left[\sum_{i=1}^n \lambda_i t \right]}$$

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 variate X which is greater than or equal to zero and for which $E[X]$ is fixed.^[2]

Multiple Random Variables

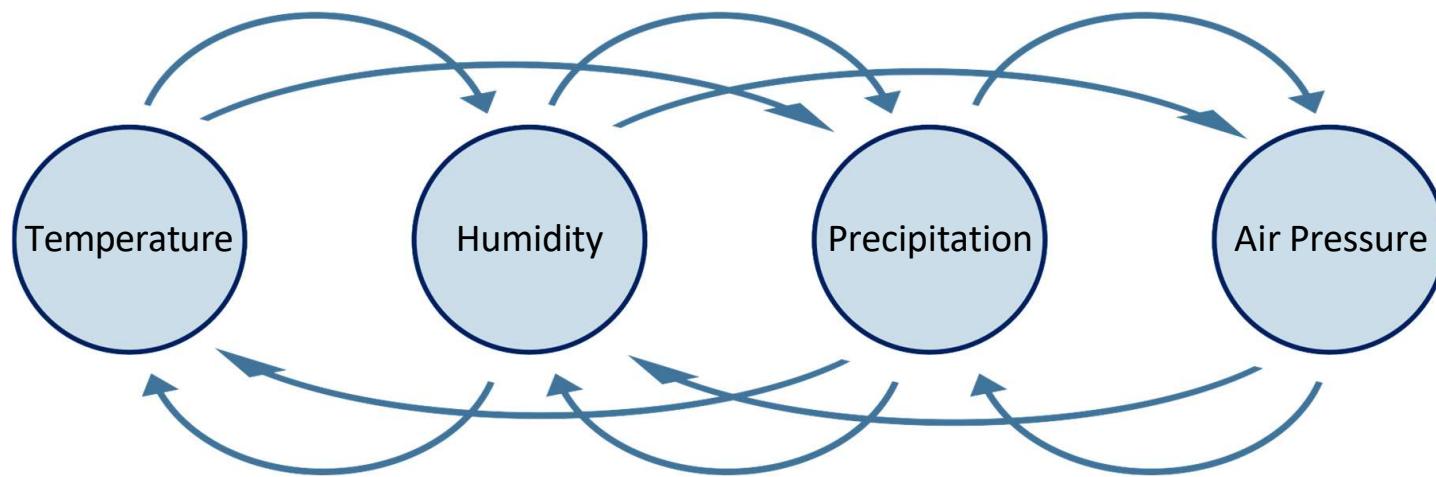
What Are We Missing?



The world is full of interesting probability problems...
...and many of them involve *multiple* random variables, being random *together*

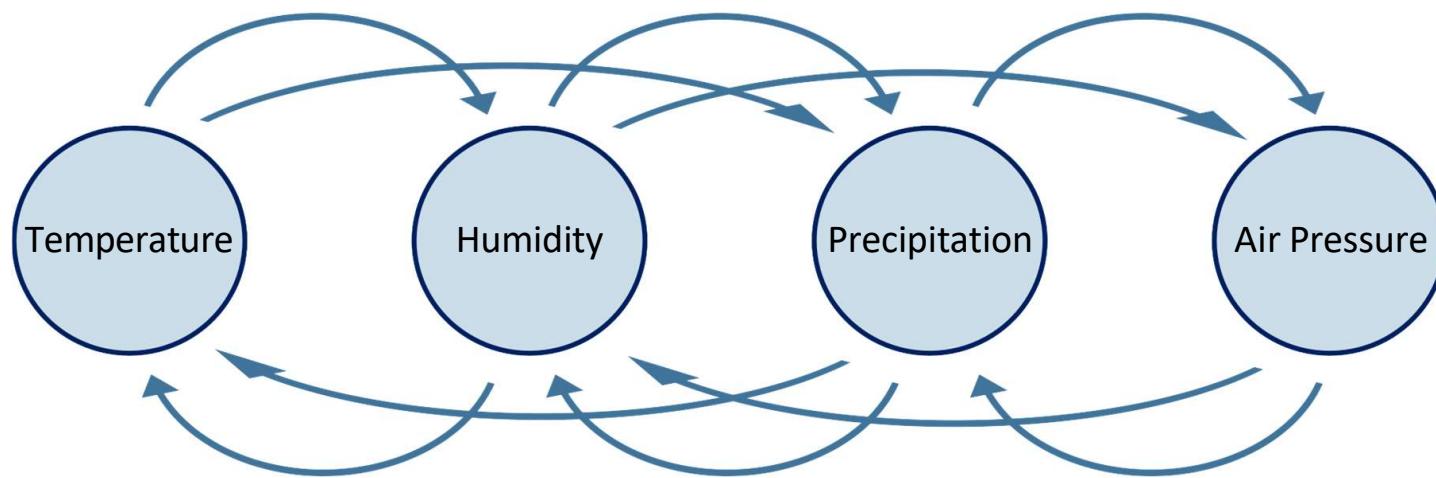
How Do We Model Multiple Random Variables Together?

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**:

$$P(X = x, Y = y) \quad \text{for } X = 2, Y = 4$$

$P(X = \text{male}, Y = \text{5'9 ft})$
gender height

0.5134 ...

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:

$$\underline{\underline{f(X = x, Y = y)}}$$

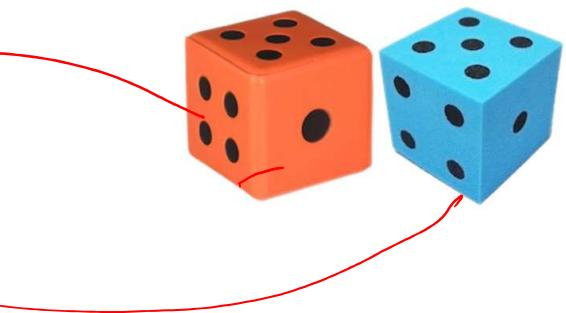
$$P\{(X, Y) \in C\} = \iint_{(x, y) \in C} f(x, y) dx dy$$

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

X
random variable

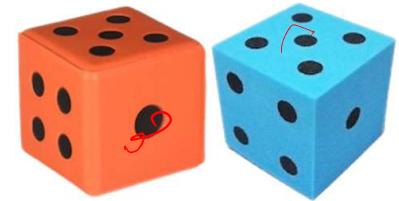
$P(X = 1)$
probability of
an event



$P(X = k)$
probability mass function

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

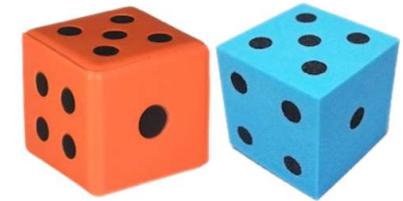
$\underline{P(X = 1, Y = 6)}$
probability of the intersection
of two events

$P(X = x, Y = y)$
joint probability mass function

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

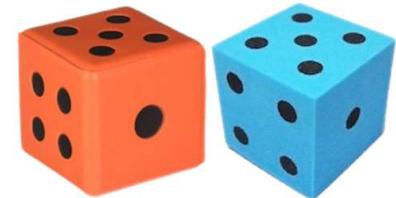
What is $P(X = x, Y = y)$?



Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?



$$P(\underbrace{X = x}_{}, \underbrace{Y = y}_{}) = \frac{1}{\underbrace{36}_{}}$$

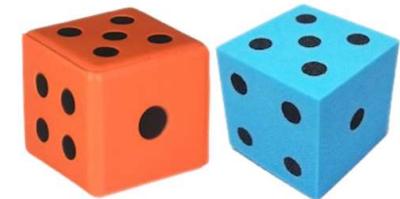
$$(x, y) \in \{(1,1), \dots, (6,6)\}$$

Example Joint PMF: Two Dice

Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?

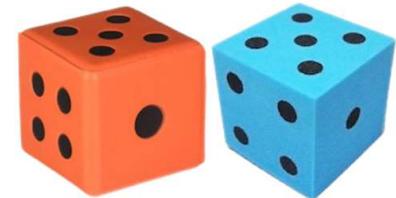
	1/36
1
2
3
4
5
6	1/36



$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \dots, (6,6)\}$$

Example Joint PMF: Two Dice



Roll two 6-sided dice, yielding values X and Y .

What is $P(X = x, Y = y)$?

	X					
	1	2	3	4	5	6
1	1/36	1/36
2
3
4
5
6	1/36	1/36

$P(X = 4, Y = 3)$

$$P(X = x, Y = y) = \frac{1}{36}$$

$(x, y) \in \{(1,1), \dots, (6,6)\}$

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

$$p(i, j) = \frac{\binom{3}{i} \binom{4}{j} \binom{5}{3-i-j}}{\binom{12}{3}}$$

$$p(0, 0) = \binom{5}{3} / \binom{12}{3} = 10/220$$

$$p(0, 1) = \binom{4}{1} \binom{5}{2} / \binom{12}{3} = 40/220$$

$$p(0, 2) = \binom{4}{2} \binom{5}{1} / \binom{12}{3} = 30/220$$

Table 4.1 $P\{X = i, Y = j\}$

		0	1	2	3	Row Sum	
		\sim # of new batteries	\sim # of used batteries			$= P\{X = i\}$	
		0	1	2	3	defective	
i	j	0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
	0	0	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
	1	0	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
	2	0	0	0	0	0	$\frac{1}{220}$
	3	0	0	0	0	0	0
		Column Sums =	$P\{Y = j\}$				
		3	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

Example with continuous density

$$C \equiv \{X < Y\}$$

Example 4.3.c. The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

$$C \equiv \{X > 1, Y < 1\}$$

$$\begin{aligned} P\{X > 1, Y < 1\} &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y} dx dy \\ &= \int_0^1 2e^{-2y}(-e^{-x}|_1^\infty) dy \end{aligned}$$

$$P(X < Y) = \iint f(x, y) dy$$

y x

$$\begin{aligned} P\{X < Y\} &= \iint_{(x,y):x < y} 2e^{-x}e^{-2y} dx dy \\ &= \int_0^\infty \int_0^y 2e^{-x}e^{-2y} dx dy \\ &= \int_0^\infty 2e^{-2y}(1 - e^{-y}) dy \\ &= \int_0^\infty 2e^{-2y} dy - \int_0^\infty 2e^{-3y} dy \\ &= 1 - \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

Marginals

$$\underline{P\{X < a\}} = \int_0^a \int_0^\infty 2e^{-2y} e^{-x} dy dx$$

~~$x=0$~~ ~~$y=0$~~

$$= \int_0^a e^{-x} dx$$
$$= 1 - e^{-a}$$

■

$$P(X < a) \sim \exp(\gamma = 1)$$

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 $P(X, Y)$

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

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

$$\underbrace{P\{X \in A, Y \in B\}}_{\text{Product rule}} = \underbrace{P\{X \in A\}}_{\text{Marginal}} \underbrace{P\{Y \in B\}}_{\text{Marginal}} \quad (4.3.7)$$

This also implies that

$$\underbrace{P(X \leq a, Y \leq b)}_{\text{Joint distribution}} = \underbrace{P(X \leq a)}_{\text{Marginal}} \underbrace{P(Y \leq b)}_{\text{Marginal}}$$

Or $F_{X,Y}(a, b) = F_X(a)F_Y(b)$

In the jointly continuous case, the condition of independence is equivalent to

$$\underbrace{f(x, y)}_{\text{Joint density}} = \underbrace{f_X(x)f_Y(y)}_{\text{Marginal densities}} \quad \text{for all } x, y$$

Example 4.3.d. Suppose that X and Y are independent random variables having the common density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(y) = \begin{cases} e^{-y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} z &= x+y \\ \frac{\partial z}{\partial x} &= 1 \\ \frac{\partial z}{\partial y} &= 1 \end{aligned}$$

Find the density function of the random variable X/Y .

$$f(z = \frac{x}{y})$$

$$\begin{aligned} P(z \leq a) &= P\left(\frac{x}{y} \leq a\right) = P(x \leq a y) = \int_0^\infty \int_0^{ay} f(x, y) dx dy \\ &= \int_0^\infty e^{-y} \int_0^{ay} e^{-x} dx dy = F_z(a) = \frac{a}{a+1} \end{aligned}$$

$$\begin{aligned} f(z) &= \frac{\partial}{\partial a} F_z(a) & z = T(x) & f_x(x) \\ &= 0 & f_z(z) &= f_x(T^{-1}(z)) \frac{\partial T^{-1}}{\partial z} \end{aligned}$$

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 $f(x, y)$:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f(y)}$$

$$\begin{aligned} f_{X|Y}(x|y) dx &= \frac{\int f(x, y) dx dy}{\int f(y) dy} \\ &\approx \frac{P\{x \leq X \leq x + dx, y \leq Y \leq y + dy\}}{P\{y \leq Y \leq y + dy\}} \\ &= P\{x \leq X \leq x + dy | y \leq Y \leq y + dy\} \end{aligned}$$

Example 4.3.h. 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 $0 < \underline{y} < 1$.

Solution. For $0 < x < 1$, $0 < y < 1$, we have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \quad | \\ &= \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) dx} \quad | \\ &= \frac{x(2-x-y)}{\int_0^1 x(2-x-y) dx} \quad | \\ &= \frac{x(2-x-y)}{\frac{2}{3} - y/2} \quad | \\ &= \frac{6x(2-x-y)}{4-3y} \quad | \blacksquare \end{aligned}$$

$$f_Y(y) = \int_{x=0}^1 f(x, y) dx$$

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

$$p(x_1, x_2, \dots, x_n) = P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$$

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

$$\begin{aligned} P\{X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n\} \\ = \int_{A_n} \int_{A_{n-1}} \dots \int_{A_1} f(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n \end{aligned}$$

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

$$P\{\text{daily change is } i\} = P(X_i)$$

-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

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. ■

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)
- Uni(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”

Given model, **parameters** yield actual distribution

- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

Non parametric example - model-histogram

Stanford University

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
 $X = \# \text{ of failures in a day}$ $X \sim \text{Bin}(C)$

2. Gaussian?

$X = \text{change from one day to the next}$

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 scr 1	day 1 scr 2	\bar{x}_3	\bar{x}_4	\bar{x}_5	\bar{x}_6	x_7	x_8	day 3 scr 3	\bar{x}_{12}
	x_1 0	x_2 0	1	0	0	0	0	1	---	0

2. Stock price change over a 10 days

change	1 → 2	2 → 3	-	-	-	-	-	9-10
	x_1 1%	x_2 -2%						

3. Number of clicks on the ad over the last 20 hour

Hour	1	2	3	<u>...</u>	20	x_{20}
	10 x_1	15 x_2	5 x_3	7		

- Step 4: Estimate the unknown parameters using the training sample

The overall setup in parameter estimation

density of
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_N are independent
- $P(D|\theta) = P(x_1, \dots, x_N|\theta) = \prod_i P(x_i|\theta) = \prod_i f(x_i; \theta)$ * for both continuous & discrete
- 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

$$\bullet LL(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log f(x_i, \theta)$$

*↑
sum over observations*

$\max_{\theta} LL(\theta)$

- 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

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

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

$$\underline{f(x; p) = p^x (1-p)^{1-x}}$$

Data sample: D

x_1	x_2	x_3	x_4	\dots	x_6	x_7	x_8	x_9	$x_{10} \leftarrow$
0	1	1	0	0	0	0	0	1	1

$$\begin{aligned} \max_p \underline{LL_D(p)} &= LL_D(p) = \max_p \sum_{i=1}^N \log p^{x_i} (1-p)^{1-x_i} \\ &= \max_p \sum_{i=1}^N x_i \log p + \left(N - \sum_{i=1}^N x_i \right) \log (1-p) \end{aligned}$$

$$\text{Let} - \sum_{i=1}^n x_i = N_1$$

max $N_1 \log p + (N-N_1) \log(1-p)$

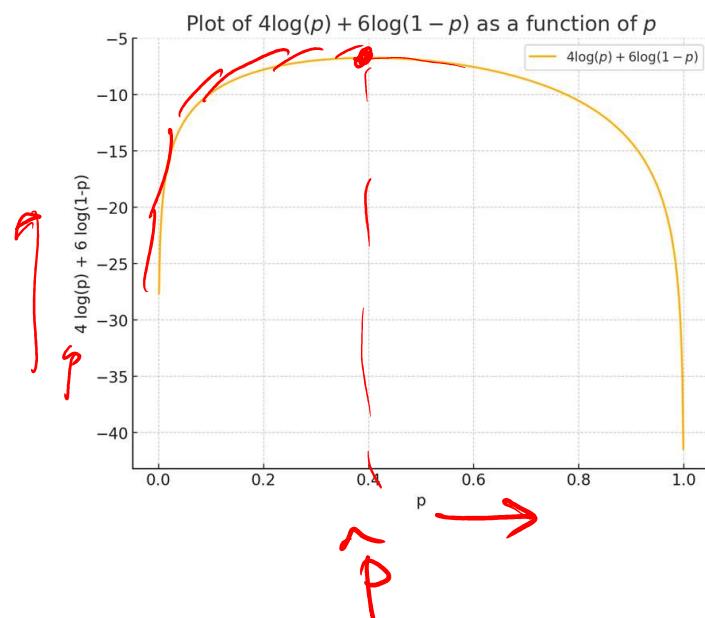
p

$LL(p)$

$$\frac{\partial LL}{\partial p} = \frac{N_1}{p} - \frac{N-N_1}{1-p} = 0$$

$$\Rightarrow \hat{p} = \frac{N_1}{N}$$

concave in p
 unique maxima at the p where
 $\frac{\partial LL}{\partial p} = 0$



Examples: MLE for Poisson

$$f(k, \lambda) = \frac{\bar{e}^\lambda \lambda^k}{k!}$$

$$x \sim \exp(\lambda)$$

$$f(x, \lambda) = \frac{\bar{e}^\lambda \lambda^x}{x!}$$

$$D = \{x_1, x_2, \dots, x_N\}$$

$$LL(\lambda) = \sum_{i=1}^N \log \frac{\bar{e}^\lambda \lambda^{x_i}}{x_i!} = \left(\sum_{i=1}^N x_i \right) \log \lambda - \sum_{i=1}^N \log x_i!$$

$$\hat{\lambda} = \arg \max_{\lambda} \left(\sum_{i=1}^N x_i \right) \log \lambda - \lambda N$$

$$\frac{\partial LL}{\partial \lambda} = \sum_{i=1}^N x_i - N \quad \therefore \hat{\lambda} = \frac{\sum x_i}{N}$$

$\sum x_i \leftarrow \text{sample mean}$

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

$$\frac{\partial LL}{\partial \sigma} = 0$$

$$\begin{aligned}
 f(x_1, \dots, x_n | \mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right] \\
 &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right]
 \end{aligned}$$

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 foregoing, we compute

$$\begin{aligned}
 \frac{\partial}{\partial \mu} \log f(x_1, \dots, x_n | \mu, \sigma) &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \\
 \frac{\partial}{\partial \sigma} \log f(x_1, \dots, x_n | \mu, \sigma) &= -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3}
 \end{aligned}$$

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

Homework

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

• 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 textbook

MLE for a new distribution: Gamma distribution

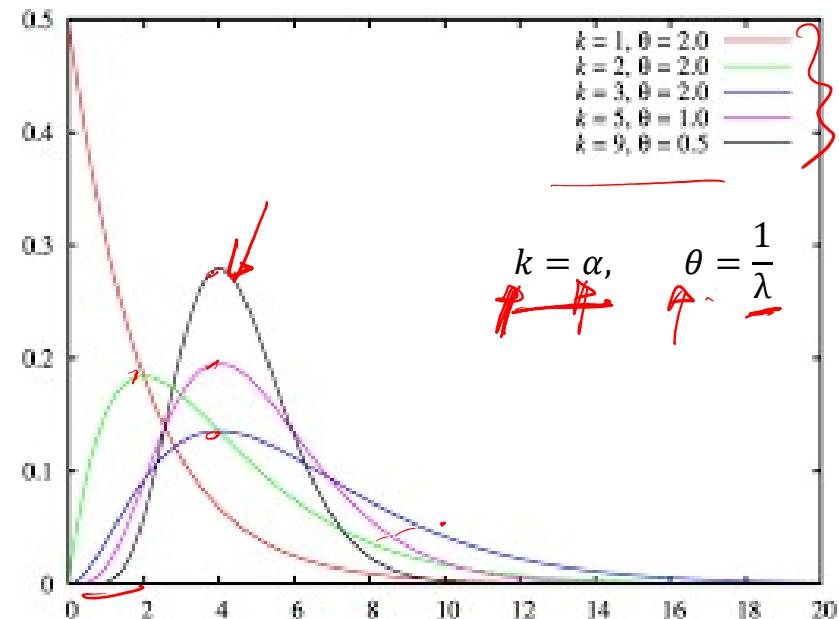
A random variable is said to have a gamma distribution with parameters (α, λ) , $\lambda > 0$, $\alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for what α is $f(x) \sim \exp(-\lambda) \lambda^x$
 $= \lambda e^{-\lambda} \lambda^x$

$$\alpha = 1$$

- 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

$$\begin{aligned}\underline{\Gamma(\alpha)} &= \int_0^\infty \lambda e^{-\lambda x} (\lambda x)^{\alpha-1} dx \\ &= \int_0^\infty e^{-y} y^{\alpha-1} dy \quad (\text{by letting } y = \underline{\lambda x})\end{aligned}$$

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

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

$$\Gamma(1) = 1 \quad (5.7.1)$$

If α is $+ \text{integer}$ then

$$\Gamma(\alpha) = (\alpha-1)!$$

Estimate MLE of parameter λ of gamma distribution -

$$D = \{x_1, x_2, \dots, x_N\}$$

$$\hat{\theta} = \hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \sum_{i=1}^N \log \frac{x_i^{\lambda-1} e^{(\lambda x_i) \lambda}}{\Gamma(\lambda)}$$

$$= \underset{\lambda}{\operatorname{argmax}} \sum_{i=1}^N \left[-\lambda x_i + (\lambda-1) \log \lambda + \log \Gamma(\lambda) \right]$$

$$\frac{\partial F}{\partial \lambda} = -\sum_{i=1}^N x_i + \frac{(\lambda-1)N}{\lambda} + \frac{N}{\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{\alpha N}{\sum_{i=1}^N x_i}$$

MLE for α

$$\underset{\alpha}{\operatorname{argmax}} \quad (\alpha - 1) \left[N \log \alpha + \sum_{i=1}^N \log x_i \right] - N \log \Gamma(\alpha)$$

$$\frac{\partial \ell(\alpha)}{\partial \alpha} = \sum_{i=1}^N \log x_i - \frac{N}{\Gamma(\hat{\alpha})} \frac{\partial}{\partial \alpha} \left. \Gamma(\alpha) \right|_{\hat{\alpha}} + N \log \Gamma(\hat{\alpha}) = 0$$

Not easy to solve in closed form.
But can be estimated numerically.

Evaluating a point estimator (Chapter 7.7)

- Given sample $\underline{D} = \{\underline{X}_1, \underline{X}_2, \dots, \underline{X}_N\}$
- Given density/PMF: $f(\underline{x}, \theta)$
- Let $\hat{\theta}_{\underline{D}}$ be any estimated value of θ , example maximum likelihood estimate.
- How do we measure quality of the estimate?
 - Square difference from actual parameter.
 - $\text{Error}(\hat{\theta}_{\underline{D}}) = (\hat{\theta}_{\underline{D}} - \theta)^2$

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

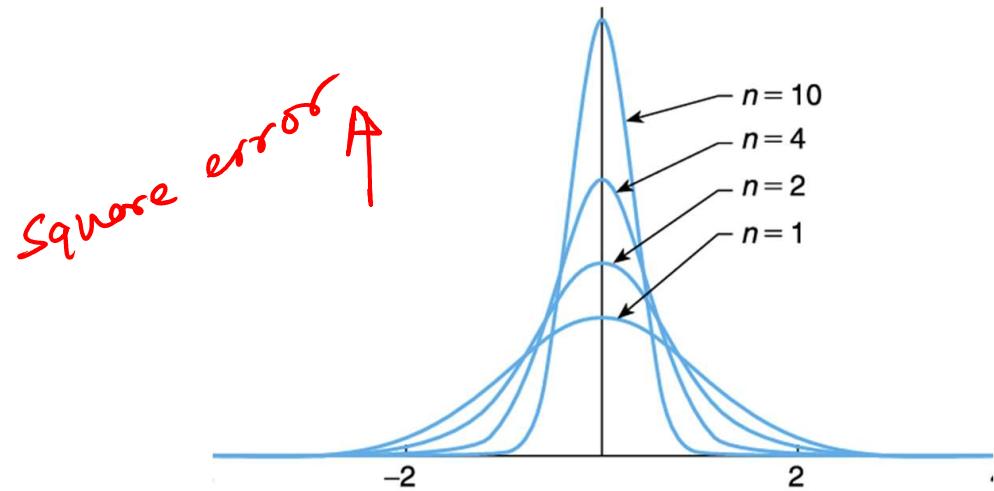
A common estimated parameter is the mean of the distribution.

$$\underline{\theta} = \mu = \underline{E_f(X)}, \quad \hat{\theta} = \underline{(X_1 + X_2 + \dots + X_N)/N} \quad \leftarrow \text{sample mean}$$

- Expected square error of the above estimate $E_f\left(\sum_i \frac{X_i}{N} - \hat{\theta}\right)^2 = \sigma^2/N$
where $\sigma^2 = E_f(X - \mu)^2$

$$\begin{aligned}
 & E_f\left(\frac{\sum_i x_i - N\hat{\theta}}{N}\right)^2 \\
 &= \frac{1}{N^2} E_f\left[\sum_{i=1}^N (x_i - \hat{\theta})^2 + 2 \sum_{i \neq j} E(x_i - \hat{\theta})(x_j - \hat{\theta})\right] \\
 &= \frac{1}{N^2} \sum_{i=1}^N E(x_i - \hat{\theta})^2 + 2 \sum_i E(x_i - \hat{\theta}) \sum_{j \neq i} E(x_j - \hat{\theta}) \\
 &= \frac{N\sigma^2}{N^2} + 2 \cdot 0 \cdot \dots \cdot [\because E(x_i - \hat{\theta}) = 0]
 \end{aligned}$$

$\hat{\theta}$ $x_i + x_j$



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.
 $D_1 = \{1, 1.5, 0.5\}$ $D_2 = \{1.2, 0.8, 1.8\}$
 $\bar{D}_1 = \frac{3}{3} = 1$ $\bar{D}_2 = \frac{3.8}{3} = 1.26$
- 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: $\hat{\theta}_D = X_1$

$$E_D[\hat{\theta}_D] = E_f[X_1] = \mu$$

- MLE: $\hat{\theta}_D = \frac{x_1 + x_2 + \dots + x_N}{N}$

$$E_f\left[\frac{x_1 + x_2 + \dots + x_N}{N}\right] = \frac{N \cdot \mu}{N} = \mu$$

Example: a biased estimator

- A constant estimator.

$$\hat{\theta}_D = 5.7 \text{ foot.}$$

- MLE of Variance parameter of Gaussian:

$$E_D[\hat{\sigma}_D^2] \neq \sigma^2$$

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

- Proof in

[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

$$S^2 = \frac{\sum_{i=1}^N (x_i - \hat{\mu}_0)^2}{N-1}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

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 expectations of both sides of the preceding yields, upon using the fact that for any random variable W , $E[W^2] = \text{Var}(W) + (E[W])^2$,

$$\begin{aligned} (n-1)E[S^2] &= E\left[\sum_{i=1}^n X_i^2\right] - nE[\bar{X}^2] \\ &= nE[X_1^2] - nE[\bar{X}^2] \\ &= n\text{Var}(X_1) + n(E[X_1])^2 - n\text{Var}(\bar{X}) - n(E[\bar{X}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

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.

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.

$$\begin{array}{lll} \text{D}_1 & |D_1|=2 & n_1(D_1) = 1, \quad N-n = 1 \\ \cancel{\text{D}_2} & |D_1|=1000 & n_1(D_2) = 500 \end{array} \quad \begin{array}{l} \hat{p}_1 = 0.5^- \\ \hat{p}_2 = 0.5^- \end{array}$$

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 D

What fraction of batchmates play bass guitar?

MLE: 0

Do you have a different guess? 0.01

P ~ 0.01

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 $T \sim \mathcal{G}(0, 5^2)$

$f_{\text{pr}}(\theta) \sim \mathcal{N}(30, 10)$

Also called prior probability

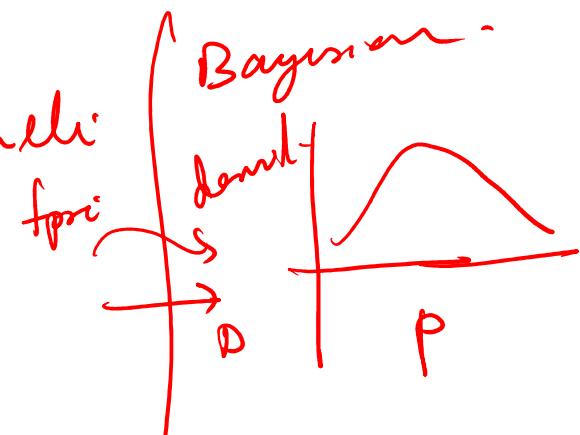
Bayesian estimation

- Calculate the posterior distribution of parameters after observing data D following Bayes rule

$$f(D|\theta) = f(\theta)f(D|\theta) / \int_{\theta} f(\theta)f(D|\theta)$$
$$f_{\text{post}}(\theta|D) = \frac{f(D|\theta)f_{\text{pri}}(\theta)}{\int_{\theta'} f(D|\theta')f_{\text{pri}}(\theta')} \quad \checkmark$$

Posterior probability

$$\begin{array}{c} \text{MLE} \\ \downarrow \\ D \rightarrow \hat{p} = 0.6 \end{array}$$



Using Bayesian estimates

$$f(\theta|D) \equiv f_{\frac{P}{P_0}}(\theta|D)$$

- Exact Bayesian probability computation:

- Given a new x , calculate $f(x|D)$

$$f(x|D) = \int_{\theta} f(x|\theta) \underbrace{f_{P_0}(\theta|D)}_{\text{Binomial e.g.}} d\theta$$

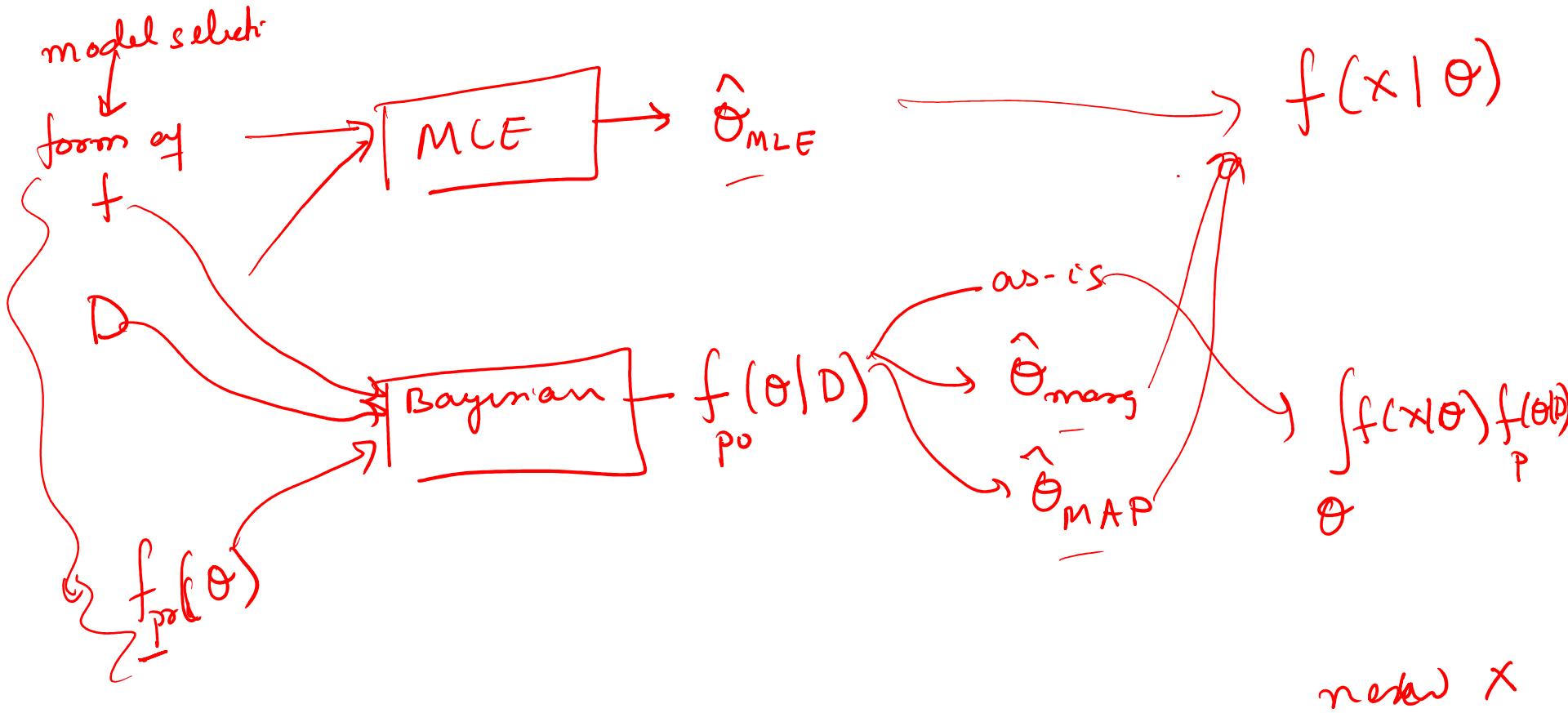
- Expected value of parameters: calculate expected value of $f(\theta|D)$

$$\hat{\theta}_{\text{mary}} = E[\theta] = \int_{\theta} \theta \cdot \underbrace{f_{P_0}(\theta|D)}_{\text{pos}} d\theta$$

- 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

$$p \in [0 \dots 1]$$

- Choose a prior distribution over parameter $\underline{\theta}$ or \underline{p} of Bernoulli
 - $f_{\text{pr}}(\theta) \sim U(0,1)$

- Data D has n are ones and remaining $N-n = m$ are 0s.

$$D = \{n, m\}$$

$$D = \{x_1, x_2, \dots, x_N\} \text{ e.g. } \{0, 1, 1, \dots, 0, 0\}$$

$$f(D|\theta) = \prod_{i=1}^n f(x_i|\theta) = \theta^n (1-\theta)^{N-n} = m$$

\uparrow #1s \uparrow #0s

- Posterior distribution is:

$$\frac{f(\theta|D)}{P_0} = \frac{f(\theta) f(D|\theta)}{\int f(\theta') f(D|\theta')} = \frac{1 \cdot \theta^n (1-\theta)^m}{\int \theta'^n (1-\theta')^m}$$

\uparrow

$$f_{p\theta}(\theta|D) = \frac{(1-\theta)^m (\theta)^n}{Z} \quad 0 \leq \theta \leq 1$$

~~$f_{p\theta}$~~ $Z \leftarrow \text{normalizer.}$

mode of $f_{p\theta}(\theta|D)$

$$\begin{aligned} & \max (1-\theta)^m \theta^n \\ &= \frac{\partial}{\partial \theta} \left[(1-\theta)^m \theta^n \right] = -m(1-\theta)^{m-1} \theta^n + n\theta^{n-1} (1-\theta)^m = 0 \\ & \Rightarrow -m\theta + n(1-\theta) = 0 \\ & \theta(n+m) = n \Rightarrow \theta = \frac{n}{m+n} \end{aligned}$$

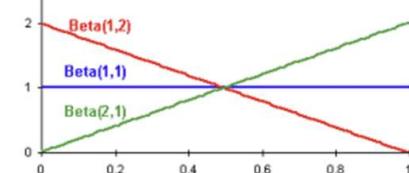
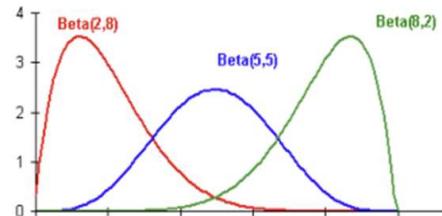
Beta Random Variable

(General defn. of Beta)

X is a Beta Random Variable: $X \sim \text{Beta}(a, b)$

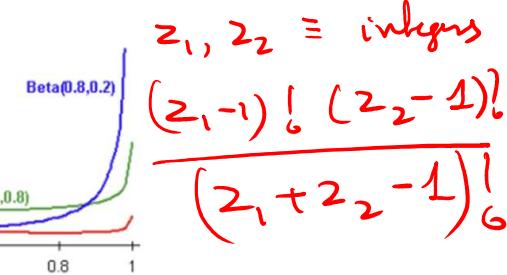
- Probability Density Function (PDF): (where $a, b > 0$)

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



$$B(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

$$B(z_1, z_2) = \frac{\Gamma(z_1) \Gamma(z_2)}{\Gamma(z_1 + z_2)}$$



$$z_1, z_2 \equiv \text{integers}$$

$$\frac{(z_1-1)! (z_2-1)!}{(z_1+z_2-1)!}$$

- Symmetric when $a = b$

$$E[X] = \frac{a}{a+b}$$

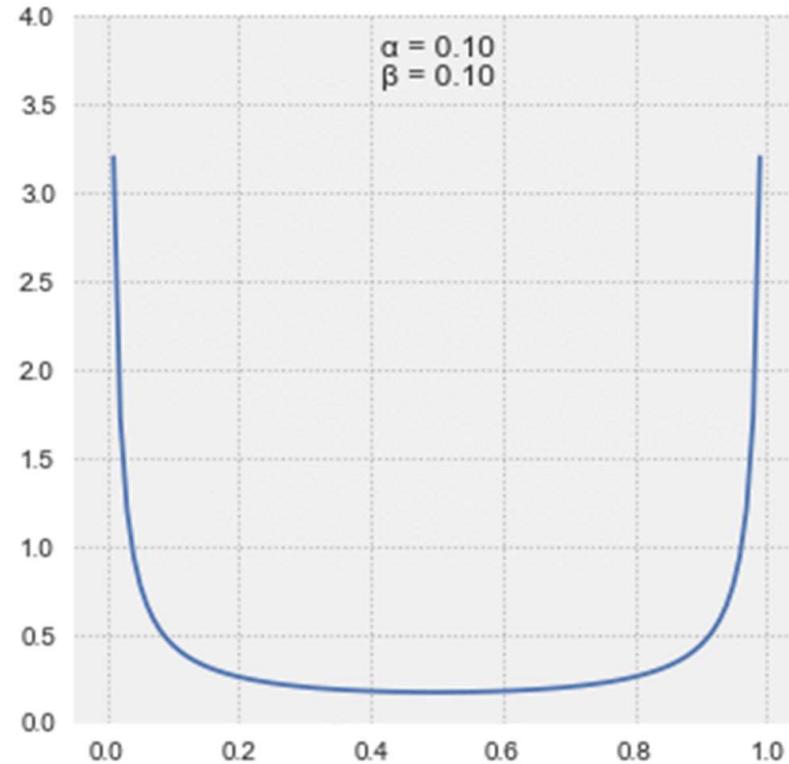
$$\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Mode : $\frac{a-1}{a+b-1}$

Chris Piech, CS109, 2021

Stanford University

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 $\underline{Y} = G(a, 1)$ and $\underline{W} = G(b, 1)$

$$f(\underline{y} | a, 1) = \frac{e^{-y} y^{a-1}}{\Gamma(a)}$$

$$\begin{aligned} f(x | \alpha, \gamma) &= \frac{\gamma^x e^{-\gamma x} (\gamma x)^{\alpha-1}}{\Gamma(\alpha)} \\ f(w | b, 1) &= \frac{e^{-w} w^{b-1}}{\Gamma(b)} \end{aligned}$$

- The $\underline{X} = \underline{Y}/(\underline{Y}+\underline{W})$ follows a Beta distribution $B(a, b)$

$$X = \frac{Y}{Y+W} \quad \text{then} \quad X \sim B(a, b)$$

Expected value of the posterior of Binomial

$$f_{P_0}(\theta | D) \equiv \frac{\theta^n (1-\theta)^m}{Z} \equiv B(a = n+1, b = m+1)$$

$$D = \begin{cases} n, m \\ p, p \end{cases}$$

#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

$$\hat{\theta}_{\text{MLE}} = \frac{n}{m+n}$$

$$\hat{\theta}_{\text{MLE}} = \frac{0}{7}$$