DS-UA 112
Introduction to Data Science

Lecture 3

Agenda

- Review
- Lesson
- Demo

Reminders

- Homework 1
  - Please access under Resources
  - Upload to Gradescope
- Lecture
  - Slides under Resources
  - Demos linked to Week…follow along on JupyterHub!

Review

- Average
- Frequency
- Rank

Average

- Average or Mean
  - Sum of all the elements of the collection, divided by the number of elements in the collection.

Average

- Average or Mean
  - Sum of all the elements of the collection, divided by the number of elements in the collection.
- Properties
  - Maybe not an element of collection
  - It is somewhere between the smallest and largest values in the collection.
  - It need not be halfway between the two extremes; it is not in general true that half the elements in a collection are above the mean.

Average

- Average or Mean
  - Sum of all the elements of the collection, divided by the number of elements in the collection.
- Properties
  - Maybe not an element of collection
  - It is somewhere between the smallest and largest values in the collection.
  - It need not be halfway between the two extremes; it is not in general true that half the elements in a collection are above the mean.
- Intuition
  - “Equalizes” or “Smoothens” collection
  - “Balance Point” of the collection

Average

$$
\begin{aligned}
\text{mean} &= 4.25 \\
&= \frac{2+3+3+9}{4} \\
&= 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} \\
&= 2 \cdot \frac{1}{4} + 3 \cdot \frac{2}{4} + 9 \cdot \frac{1}{4} \\
&= 2 \cdot 0.25 + 3 \cdot 0.5 + 9 \cdot 0.25
\end{aligned}
$$

Average

$$\text{Average Length of Word} = \frac{\Sigma_{\text{all words}} \text{length of word}}{\text{Total number of words}}$$

Average

$$\text{Average Length of Word} = \frac{\Sigma_{\text{all words}} \text{length of word}}{\text{Total number of words}}$$

$$\text{Frequency of Length k words} = \frac{\text{Number of words of length k}}{\text{Total number of words}}$$

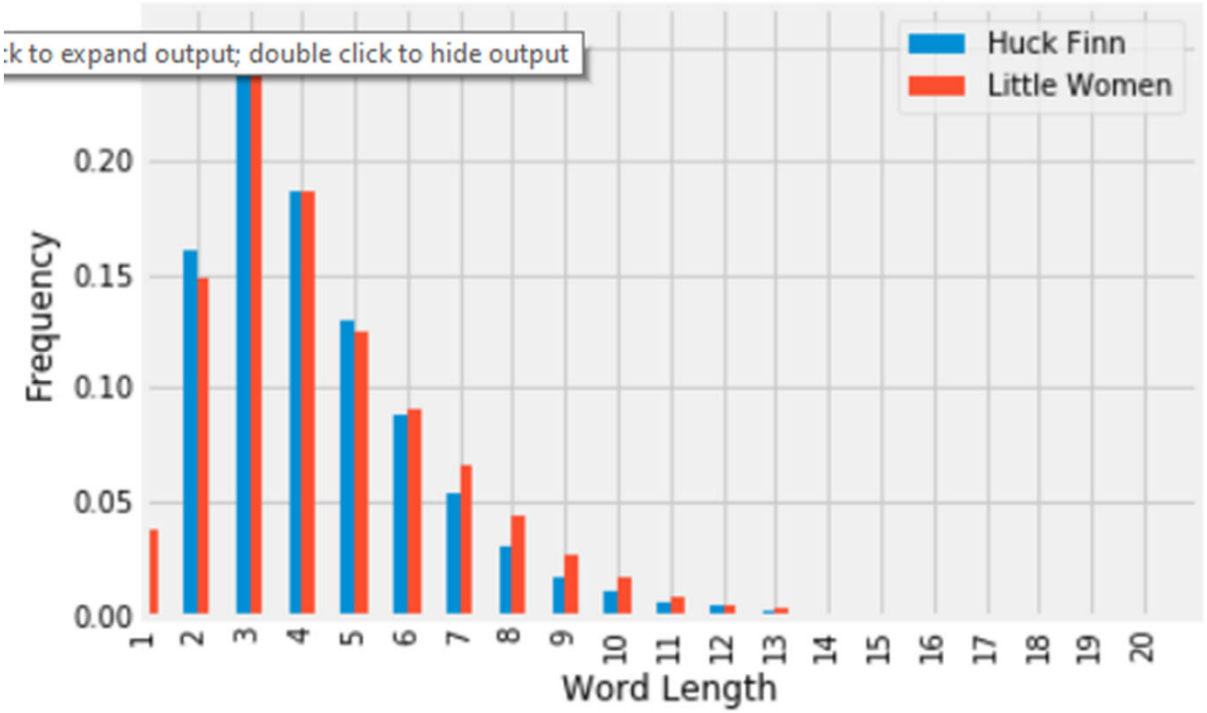
Average

$$\text{Average Length of Word} = \frac{\Sigma_{\text{all words}} \text{length of word}}{\text{Total number of words}}$$

$$\text{Frequency of Length k words} = \frac{\text{Number of words of length k}}{\text{Total number of words}}$$

$$\text{Average Length of Word} = \Sigma_{\text{all words}}\, k \times \text{Frequency of Length k Words}$$

Frequency



	Huck Finn
1	6000
2	18154
3	28809
4	21145
5	14628
6	10003
7	6105
8	3350
9	1820
10	1167
11	666
12	481
13	255
14	125
15	65
16	57
17	29
18	14
19	8
20	7

	Little Women
1	7139
2	28049
3	45459
4	35207
5	23651
6	17187
7	12420
8	8271
9	4981
10	3044
11	1579
12	921
13	491
14	185
15	86
16	46
17	20
18	7
19	3
20	5

Rank

$$\text{Frequency} \approx \frac{1}{\text{Rank}^{0.85}}$$

	Huck Finn
3	28809
4	21145
2	18154
5	14628
6	10003
7	6105
1	6000
8	3350
9	1820
10	1167
11	666
12	481
13	255
14	125
15	65
16	57
17	29
18	14
19	8
20	7

	Little Women
3	45459
4	35207
2	28049
5	23651
6	17187
7	12420
8	8271
1	7139
9	4981
10	3044
11	1579
12	921
13	491
14	185
15	86
16	46
17	20
18	7
20	5
19	3

Rank

$$\text{Frequency} \propto \frac{1}{\text{Rank}^{0.85}}$$

	Huck Finn
3	28809
4	21145
2	18154
5	14628
6	10003
7	6105
1	6000
8	3350
9	1820
10	1167
11	666
12	481
13	255
14	125
15	65
16	57
17	29
18	14
19	8
20	7

	Little Women
3	45459
4	35207
2	28049
5	23651
6	17187
7	12420
8	8271
1	7139
9	4981
10	3044
11	1579
12	921
13	491
14	185
15	86
16	46
17	20
18	7
20	5
19	3

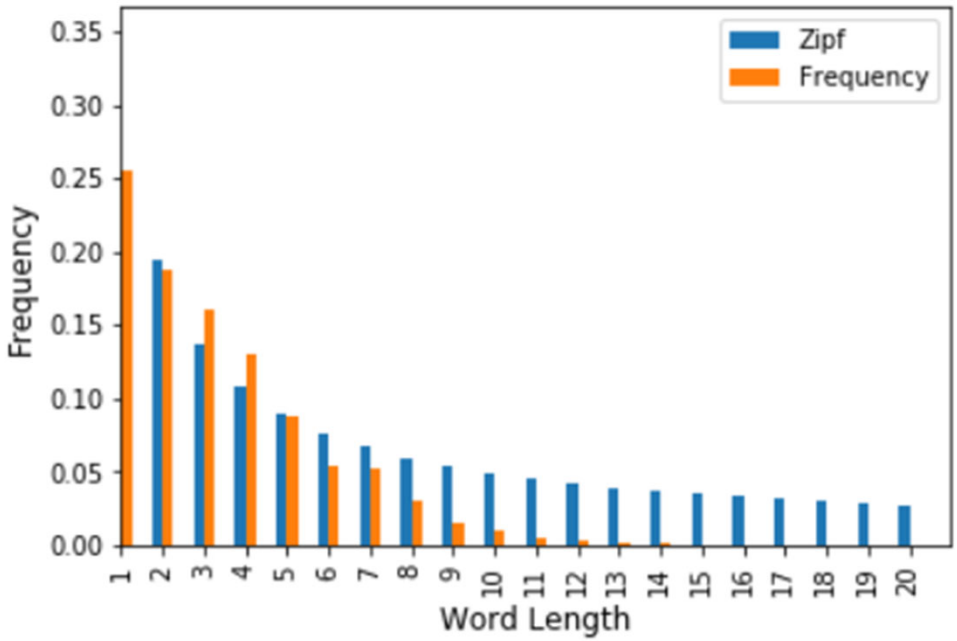
Rank

$$\text{Frequency} = \frac{0.17}{\text{Rank}^{0.85}}$$

	Huck Finn
3	28809
4	21145
2	18154
5	14628
6	10003
7	6105
1	6000
8	3350
9	1820
10	1167
11	666
12	481
13	255
14	125
15	65
16	57
17	29
18	14
19	8
20	7

	Little Women
3	45459
4	35207
2	28049
5	23651
6	17187
7	12420
8	8271
1	7139
9	4981
10	3044
11	1579
12	921
13	491
14	185
15	86
16	46
17	20
18	7
20	5
19	3

Rank



	Length	Huck Finn	Zipf	Frequency
1	3	28809	0.170000	0.255200
2	4	21145	0.094313	0.187310
3	2	18154	0.066818	0.160814
4	5	14628	0.052324	0.129580
5	6	10003	0.043284	0.088610
6	7	6105	0.037070	0.054080
7	1	6000	0.032517	0.053150
8	8	3350	0.029028	0.029675
9	9	1820	0.026263	0.016122
10	10	1167	0.024013	0.010338
11	11	666	0.022144	0.005900
12	12	481	0.020566	0.004261
13	13	255	0.019213	0.002259
14	14	125	0.018040	0.001107
15	15	65	0.017013	0.000576
16	16	57	0.016104	0.000505
17	17	29	0.015296	0.000257
18	18	14	0.014570	0.000124
19	19	8	0.013916	0.000071
20	20	7	0.013322	0.000062

Lesson

- Data Science requires answers to questions with partial information
  - Missing
  - Biased
  - Noisy
- Study chances with frequencies
  - Outcome Space
  - Event
  - Probability

When an Event Happens?

$$0 \leq P(\text{an event happens}) \leq 1$$

$$P(\text{an event doesn't happen}) = 1 - P(\text{an event happens})$$

When Events Have the Same Chances?

$$\frac{\text{Number of Even Faces}}{\text{Total Number of Faces}} = \frac{\#\{2,4,6\}}{\#\{1,2,3,4,5,6\}} = \frac{1}{2}$$

When Events Have the Same Chances?

$$P(\text{event}) = \frac{\#\text{ of outcomes in event}}{\#\text{ of all possible outcomes}}$$
$$= \text{proportion of outcomes in event}$$

When Two Events Occur in Order?

$$P(\text{green first, then red}) = \frac{\#\{\text{GR}\}}{\#\{\text{RB, BR, RG, GR, BG, GB}\}} = \frac{1}{6}$$

When Two Events Occur In Order?

$$P(\text{two events both happen}) = P(\text{one event happens}) \times$$

$$P(\text{the other event happens given that the first one happened})$$

When Two Events Occur In Order?

$$P(\text{two events both happen}) = P(\text{one event happens}) \times$$

$$P(\text{the other event happens} \mid \text{the first one happened})$$

When Two Events Occur In Order?

$$P(\text{the other event happens} \mid \text{the first one happened}) = \frac{P(\text{two events both happen})}{P(\text{the first one happened})}$$

Independence

$$P(\text{the other event happens} \mid \text{the first one happend}) = P(\text{the other event happens})$$

When Multiple Events Occur in Order?

$$P(\text{heads first and tails second}) = P(\text{heads first}) \times$$

$$P(\text{tails second} \mid \text{heads first})$$

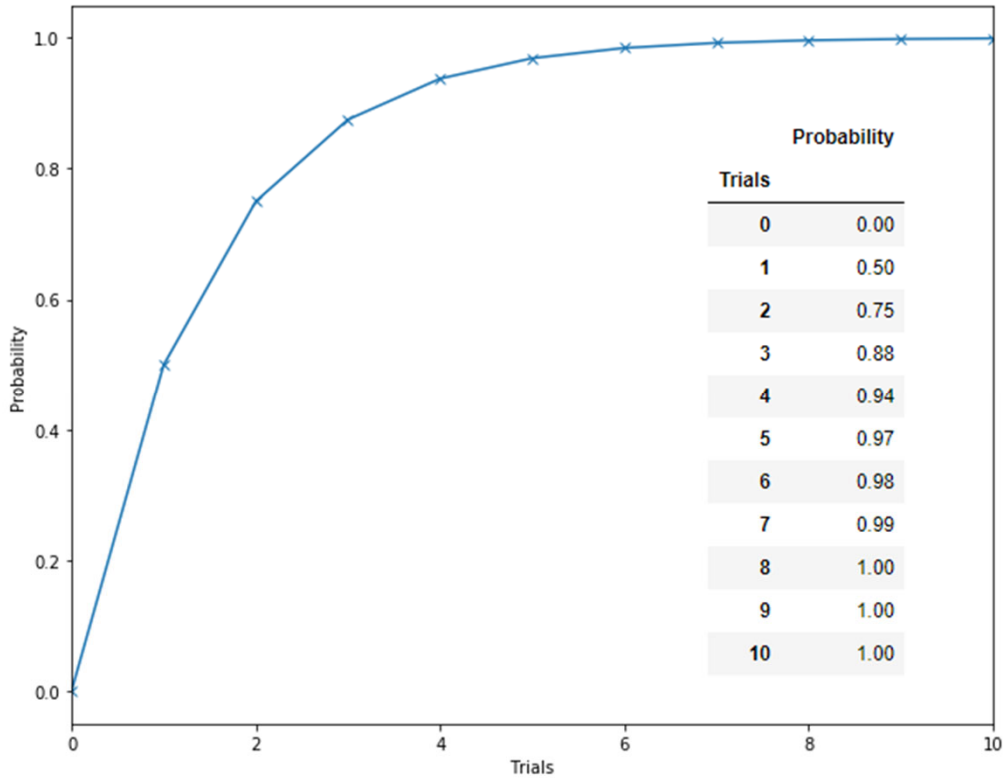
$$= P(\text{heads first}) \times$$

$$P(\text{tails second})$$

When Multiple Events Occur in Order?

$$P(\text{at least one head in 17 tosses}) = 1 - P(\text{all 17 are tails}) = 1 - \left(\frac{1}{2}\right)^{17}$$

When Multiple Events Occur in Order?



Trials	Probability
0	0.00
1	0.50
2	0.75
3	0.88
4	0.94
5	0.97
6	0.98
7	0.99
8	1.00
9	1.00
10	1.00

When Two Events Occur in Any Order?

$$P(\text{one green and one red}) = P(\text{GR}) + P(\text{RG}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

When Two Events Occur in Any Order?

$$P(\text{an event happens}) = P(\text{first way it can happen}) + P(\text{second way it can happen})$$

When Two Events Occur in Any Order?

$$1 = P(\text{one green or one red}) \neq$$

$$P(\text{one green}) + P(\text{one red}) = \frac{2}{3} + \frac{2}{3}$$

When Two Events Occur in Any Order?

$$1 = P(\text{one green or one red}) =$$

$$P(\text{one green}) + P(\text{one red}) - P(\text{one red and one green}) = \frac{2}{3} + \frac{2}{3} - \frac{1}{3}$$

When Two Events Occur in Any Order?

$$P(\text{one or the other event happens}) =$$

$$P(\text{one event happens}) + P(\text{other event happens}) - P(\text{both happen})$$

Joint Probability Table

	FIRST is GREEN	FIRST is RED	FIRST is BLUE
SECOND is GREEN	0	1/6	1/6
SECOND is RED	1/6	0	1/6
SECOND is BLUE	1/6	1/6	0

Demo

- Goals
  - How can frequencies indicate the chance of events?
  - How can probability allow us to incorporate our understanding of the problem?

How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{P(n)}$$

How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{P((n \text{ and } 1888) \text{ or } (n \text{ and } 1889) \ldots)}$$

How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{\Sigma_y P(n \text{ and } y)}$$

How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{\Sigma_y P(n \mid y) \cdot P(y)}$$