DATA 101

2.1.2021

Today’s Agenda

Housekeeping = [hw format, links to data, zip, office hours]

Homework review and questions

5 Min Helpers

MATH! Basic Stats & Probability

Logic! Basic logic discussion

Data Exploration - Pandas and Matplotlib.pyplot

Review of resources

https://github.com/wesm/pydata-book

https://github.com/jakevdp/PythonDataScienceHandbook/tree/master/notebooks

5 Minute Helpers

Reading Documentation

Searching for Errors

Python Nits

Simple Probability

Probability is measured in [0,1] interval with 0 being not possible and 1 being absolute.

Examples:

Suppose that you are going to throw a standard dice, and you want to know what your chances are of throwing a 6.

Now suppose that you want to know what your chances are of throwing 1 or 6.

If you throw three dice, what is the probability that you do not throw any 4s, 5s, or 6s?

*1, ⅙ ⅙ ⅙ ,   ⅙ ⅙ ⅙,   ⅙ ⅙ ⅙ = 3/6 * 3/6 * 3/6 = 27/216*

Independence vs. Dependence

Simple Probability - Dependence

To work out the probability of **both** events *(AND)*, you **multiply** the probability of one by the probability of the other.

To work out the probability of **either** event *(OR)*, you **add** the probability of one to the probability of the other.

What is the probability of drawing at least one ace from a pack of cards on two draws, if you do not replace the cards in between?

There are 52 cards in the pack, four of which are aces. **There are three possible favourable outcomes:**

You could draw two aces - Ace/Ace, Or draw one ace, either as the first or second card - Ace/Not, Not/Ace.

In AND/OR terms, these are:

- *Ace AND Ace OR*
- *Ace AND Not Ace OR*
- *Not Ace AND Ace.*

Worked Out Example

The first scenario: Ace and Ace

The probability of drawing an ace on the first card is 4/52 = 1/13.

Once you have drawn one ace, there are only 51 cards left from which to draw the second card, and only three of them are aces. The probability of drawing a second ace is therefore 3/51. You want both events, so you need to multiply them.

The probability of drawing Ace AND Ace is 1/13 x 3/51 = 1/221

The second scenario: Ace and Not Ace

The probability of drawing an ace remains 1/13. But now you have 51 cards left, all but three of which are not aces. 51−3=48.

Your chance of drawing a ‘not ace’ on the second card is therefore 48/51, and the chance of drawing Ace AND Not Ace is 1/13 x 48/51 = 16/221

The third scenario: Not Ace and Ace

he probability of drawing a ‘not ace’ on the first card is (52-4)÷52 = 48/52

The probability of drawing an ace on the second card is 4/51.

The probability of drawing Not Ace AND Ace is therefore 48/52 x 4/51 = 16/221

The answer is then an OR = 1/221 + 16/221 + 16/221 = 33/221

Probability practice

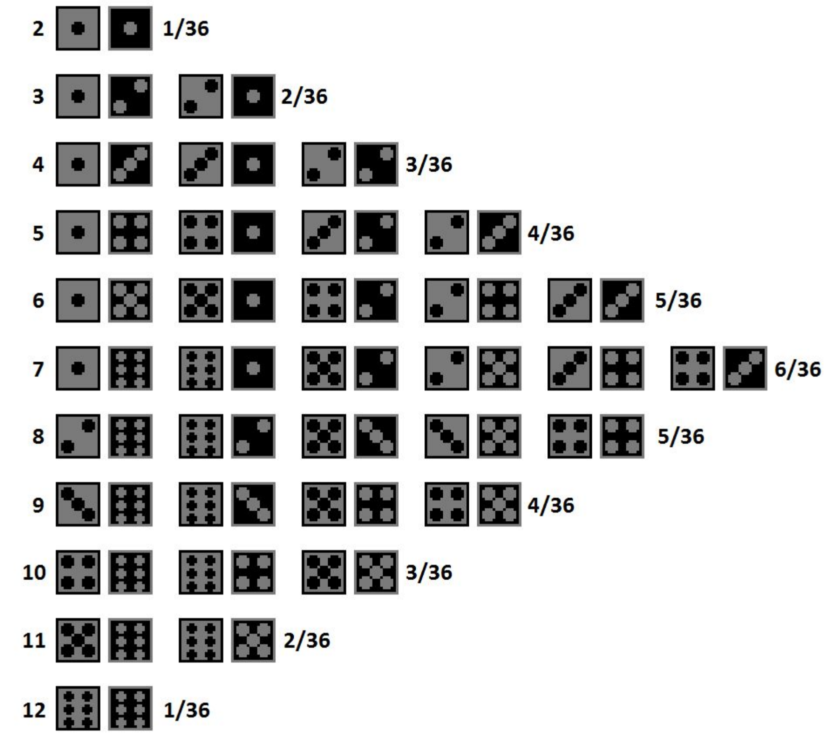
One Die

1) Odd
2) Even
3) 2 and 2

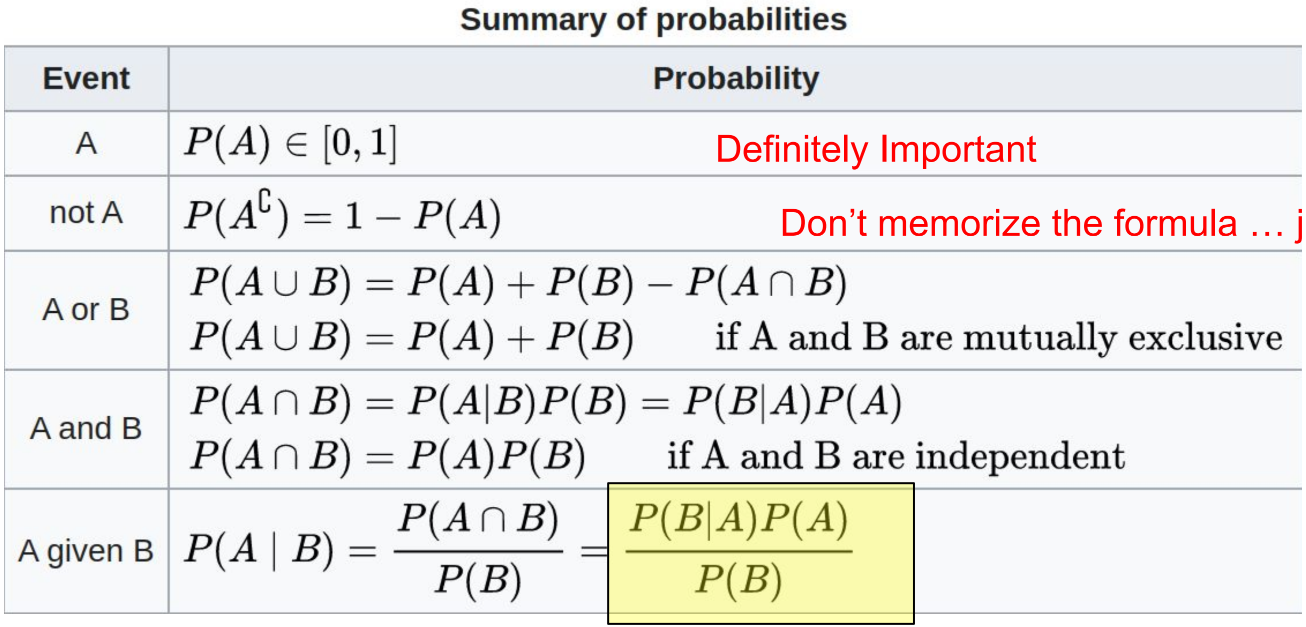
Two Dice

1) 7 and 11
2) 8, 8

Two Dice...you roll one...its a 3…
Probability of a total of 7?



Probabilities



Summary of probabilities

Event	Probability
A	$P(A) \in [0,1]$
not A	$P(A^{\complement}) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B) \quad \text{if A and B are mutually exclusive}$
A and B	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B) \quad \text{if A and B are independent}$
A given B	$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \boxed{\frac{P(B\|A)P(A)}{P(B)}}$

Definitely Important

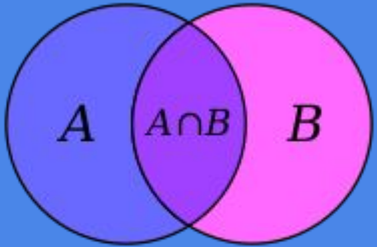
Don’t memorize the formula … just know what it means

These are results of ^^ but do not memorize formulas

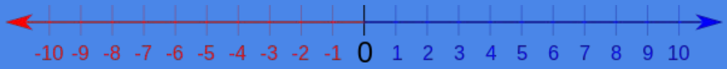
BAYES THEOREM! The most important thing you know for now...welllll in a week from now...

Sets, Venn Diagrams, Number Line, Intervals

{ notation }
AND
OR
NOT

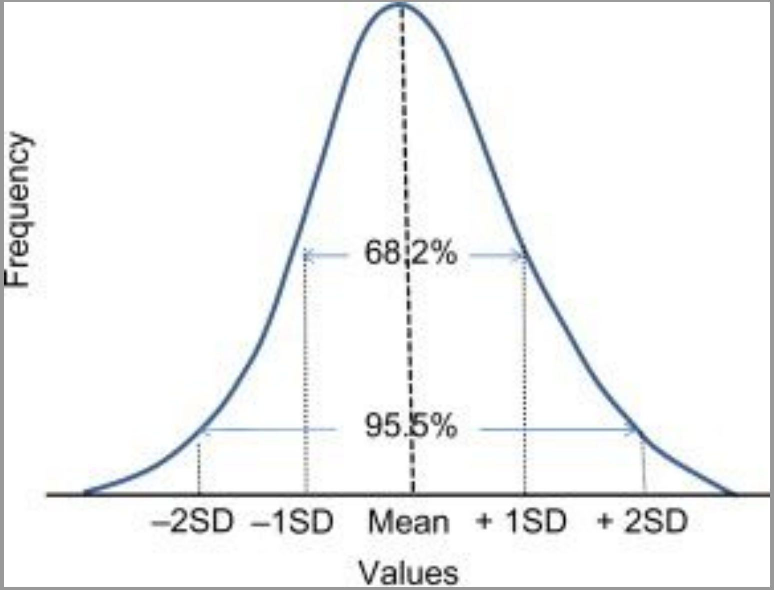


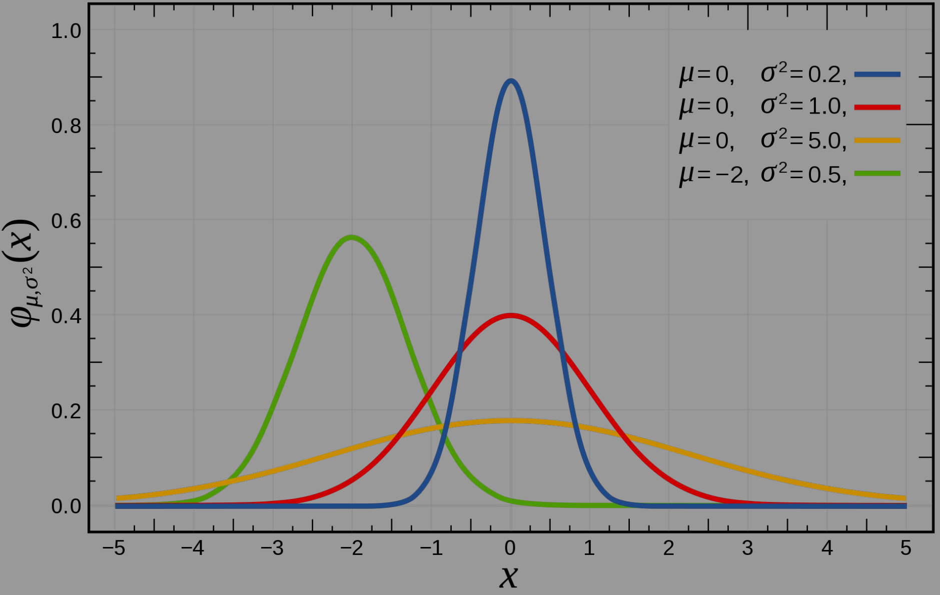
>, <, =, <=, >=
|abs|
Integer, Rationals, Irrationals, Real, Float
Intervals



Need to work on writing this out in our data to make execution easier!

Introduction to Gaussian Distribution





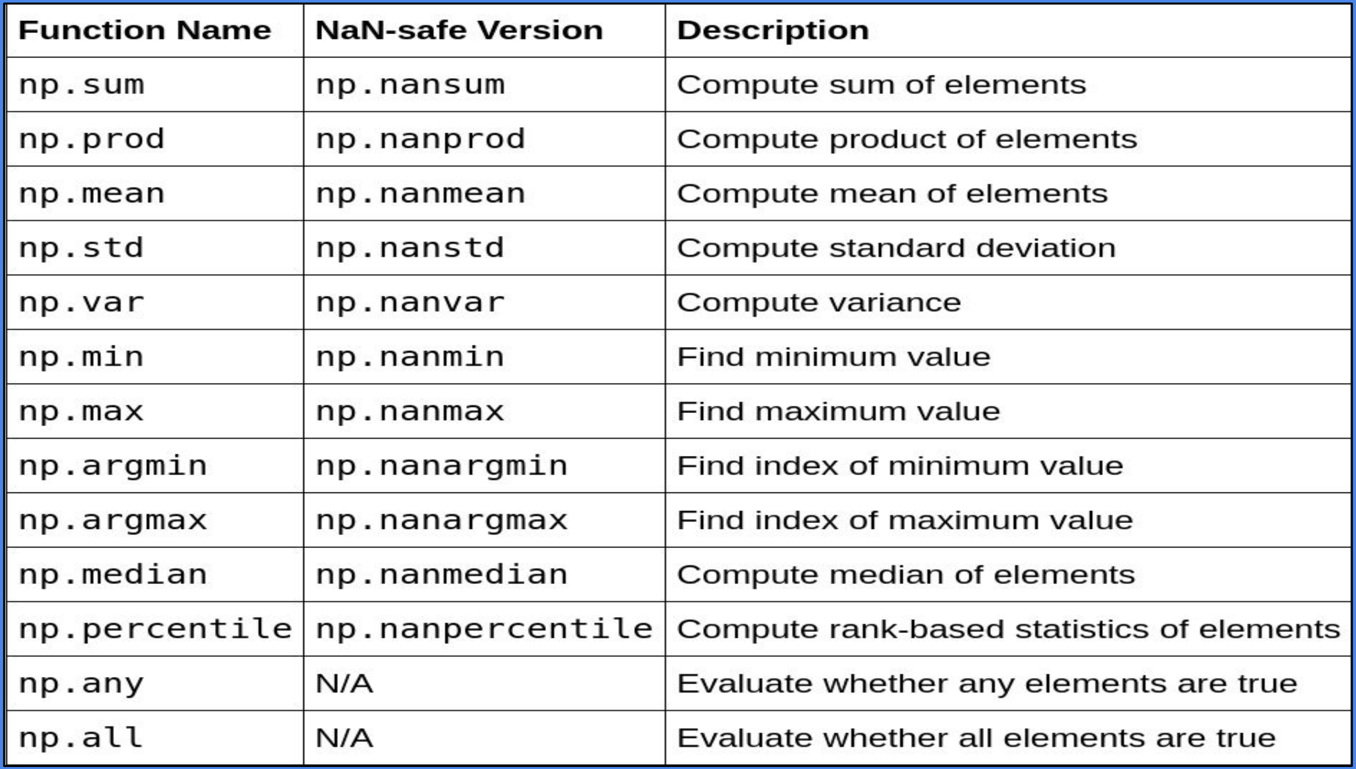
	Population	Sample
# of subjects	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$

Note: S^2 is the formula for unbiased sample variance, since we're dividing by $n - 1$.

Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Note: Finding S by taking $\sqrt{S^2}$ reintroduces bias.

Mean, Median, Variance, Standard Deviation



Function Name	NaN-safe Version	Description
`np.sum`	`np.nansum`	Compute sum of elements
`np.prod`	`np.nanprod`	Compute product of elements
`np.mean`	`np.nanmean`	Compute mean of elements
`np.std`	`np.nanstd`	Compute standard deviation
`np.var`	`np.nanvar`	Compute variance
`np.min`	`np.nanmin`	Find minimum value
`np.max`	`np.nanmax`	Find maximum value
`np.argmin`	`np.nanargmin`	Find index of minimum value
`np.argmax`	`np.nanargmax`	Find index of maximum value
`np.median`	`np.nanmedian`	Compute median of elements
`np.percentile`	`np.nanpercentile`	Compute rank-based statistics of elements
`np.any`	N/A	Evaluate whether any elements are true
`np.all`	N/A	Evaluate whether all elements are true