Hi!

DS501: Basic Statistics, Probability, and Linear Algebra

Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

Worcester Polytechnic Institute

WPI

Learning **objectives** for the **basic statistics** classes.

- Learn important ideas in mathematics, including:
  - Random Variables
  - Probabilities
  - Conditional Probabilities
  - Bayes Theorem
  - Basic linear algebra
  - Descriptive statistics
- Learn some Python packages, including:
  - numpy
  - pandas
  - matplotlib

Random variables: Discrete

$$X = \begin{cases} 1 & \boxed{\frac{1}{6}} \\ 2 & \frac{1}{6} \\ 3 & \frac{1}{6} \\ 4 & \frac{1}{6} \\ 5 & \frac{1}{6} \\ 6 & \frac{1}{6} \end{cases}$$

$$P(X=2) = \frac{1}{6}$$

Independent

$$P(X=2 \cup X=3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$P(x=2)\,P(y=3) = P(X=2 \cap Y=3) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

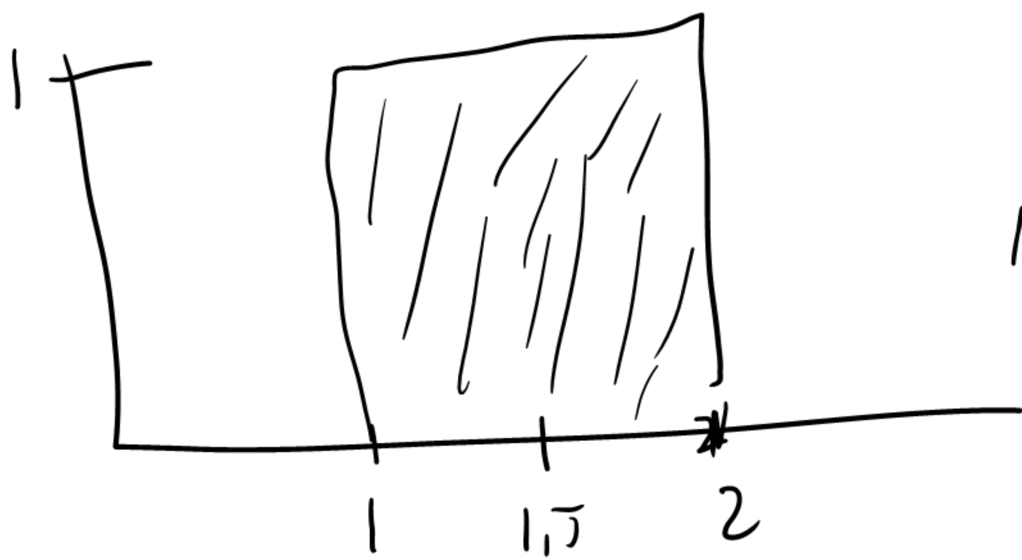
$$P(X=2 \cap X=3) = 0$$

Random variables: Continuous

Probability Distribution Function (PDF

$X = 2.7$

X



$P(X = 1.5) = 0$?

$P(1 < X < 1.5) = 0.5$

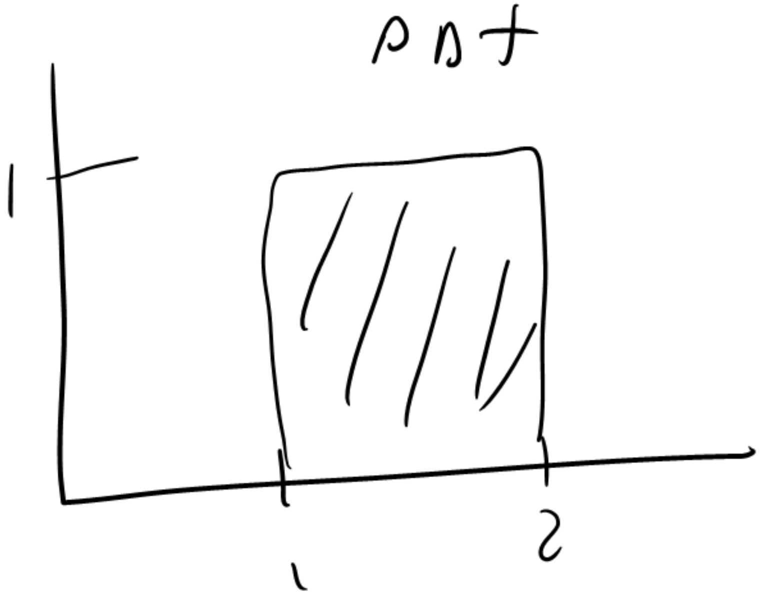
$P(1 \le X < 1.5) = 0.5$

$$P(1 < X < 1.5) = 0.5 = \int_{1}^{1.5} PDf(x)\, dx$$

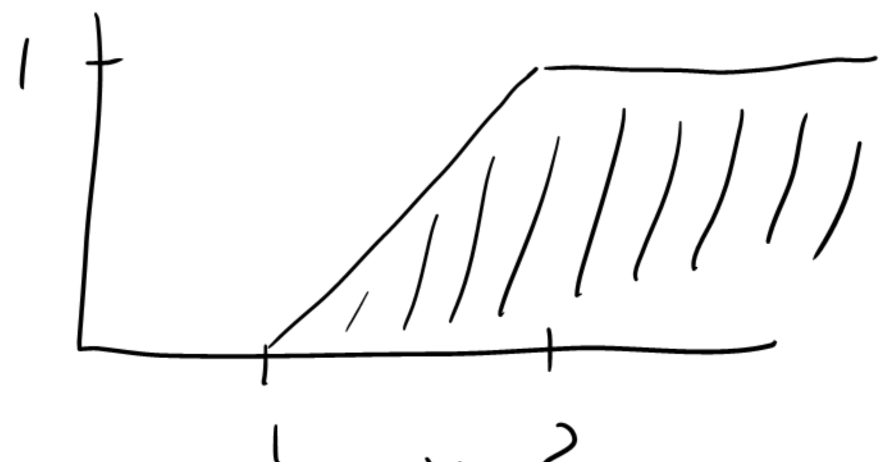
PDF

what is a CDF

Cumulative Density Function



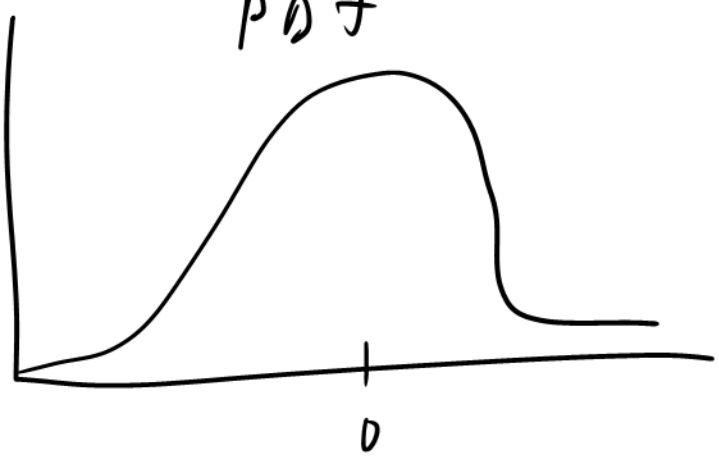
=)



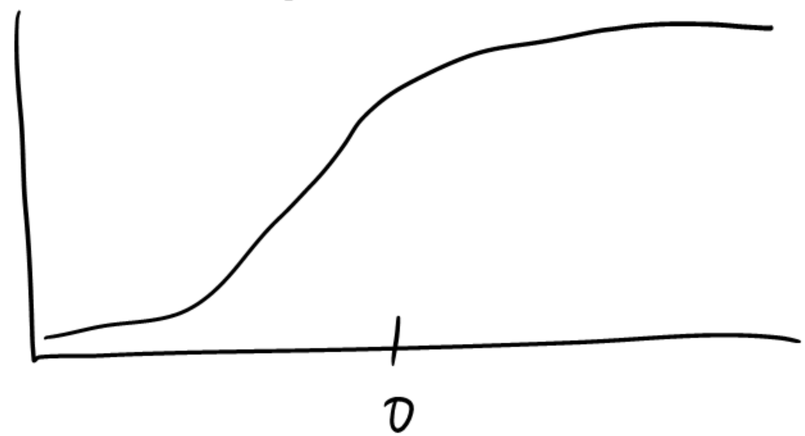
$$CDF(Y) = \int_{-\infty}^{Y} PDF(x)\,dx$$

Normal Distribution
or
Gaussian Distribution

PDF



CDF



Central limit theorem

Conditional Probabilities

$$P(X=2 \mid Y=3) = \frac{P(X=2 \wedge Y=3)}{P(Y=3)}$$

$$X = \left\{ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \quad \frac{1}{6} \right\} \qquad Y = \left\{ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \quad \frac{1}{6} \right\} \qquad Z = X + Y$$

$$P(Z=5 \wedge X=2) = P(Z=5 \mid X=2)\, P(X=2)$$

$$= \frac{1}{6} \cdot \frac{1}{6}$$

$$= \frac{1}{36}$$

			x				
		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
x	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

$$p(z=5 \cap x=2)$$

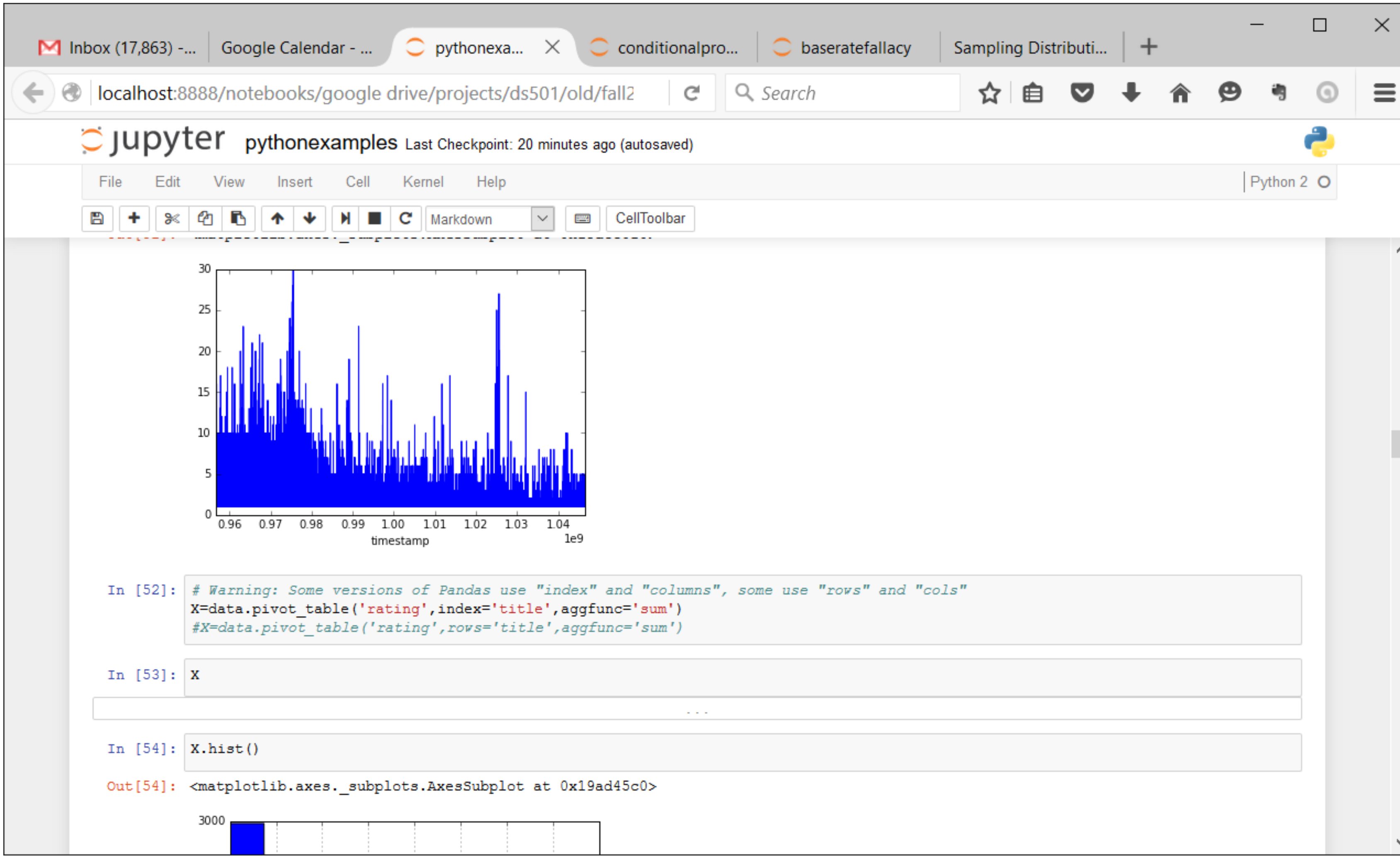
$$\frac{1}{36}$$

Central Limit Theorem

http://onlinestatbook.com/stat_sim/sampling_dist/

pythonexample
conditionalprobality

Let's see it in python



Base Rate Fallacy

- The Base Rate Fallacy is a **very** common error that people make when interpreting data.
  - It is quite easy to describe (and hopefully understand).
  - It does not require very much mathematical background.
  - It demonstrates that our intuition can lead us astray.

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

Base Rate Fallacy

Suppose you have taken a test for a deadly disease.

The doctor tells you that the test is quite accurate, in that, if you have the disease then the test will correctly tell you that you have the disease 100% of the time.

However, if you don't have the disease, the test will very occasionally (say 1 time in 10) mistakenly tell you that you have it.

The test comes back positive (it says you have the disease)! Are you worried!?

In particular, can you estimate the probability that you actually have the disease given that the test came back positive?

Base Rate Fallacy

- What is your estimate?
  - A) 99% probability I have the disease
  - B) 90% probability I have the disease
  - C) 50% probability I have the disease
  - D) 10% probability I have the disease
  - E) I don't know and I am mad at you for asking me!

The importance of asking the right question.

I was told the *probability* that I failed the test *given* that I have the disease.

$$Pr(\text{I fail the test}|\text{I have the disease})$$

I was told the *probability* that I failed the test *given* that I have the disease.

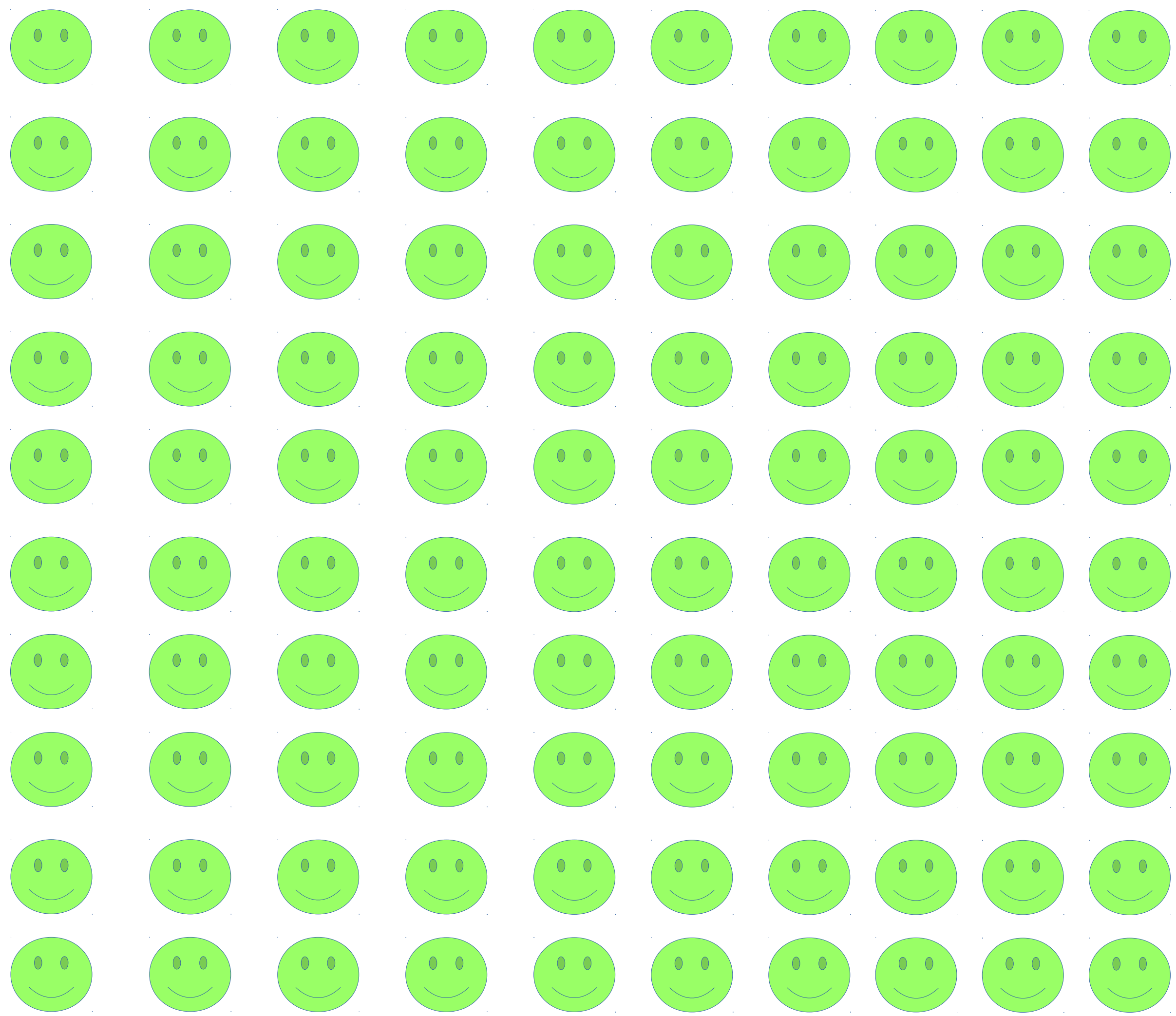
$$Pr(\text{I fail the test}|\text{I don't have the disease})$$

I want to know the *probability* that I have the disease *given* that I failed the test.

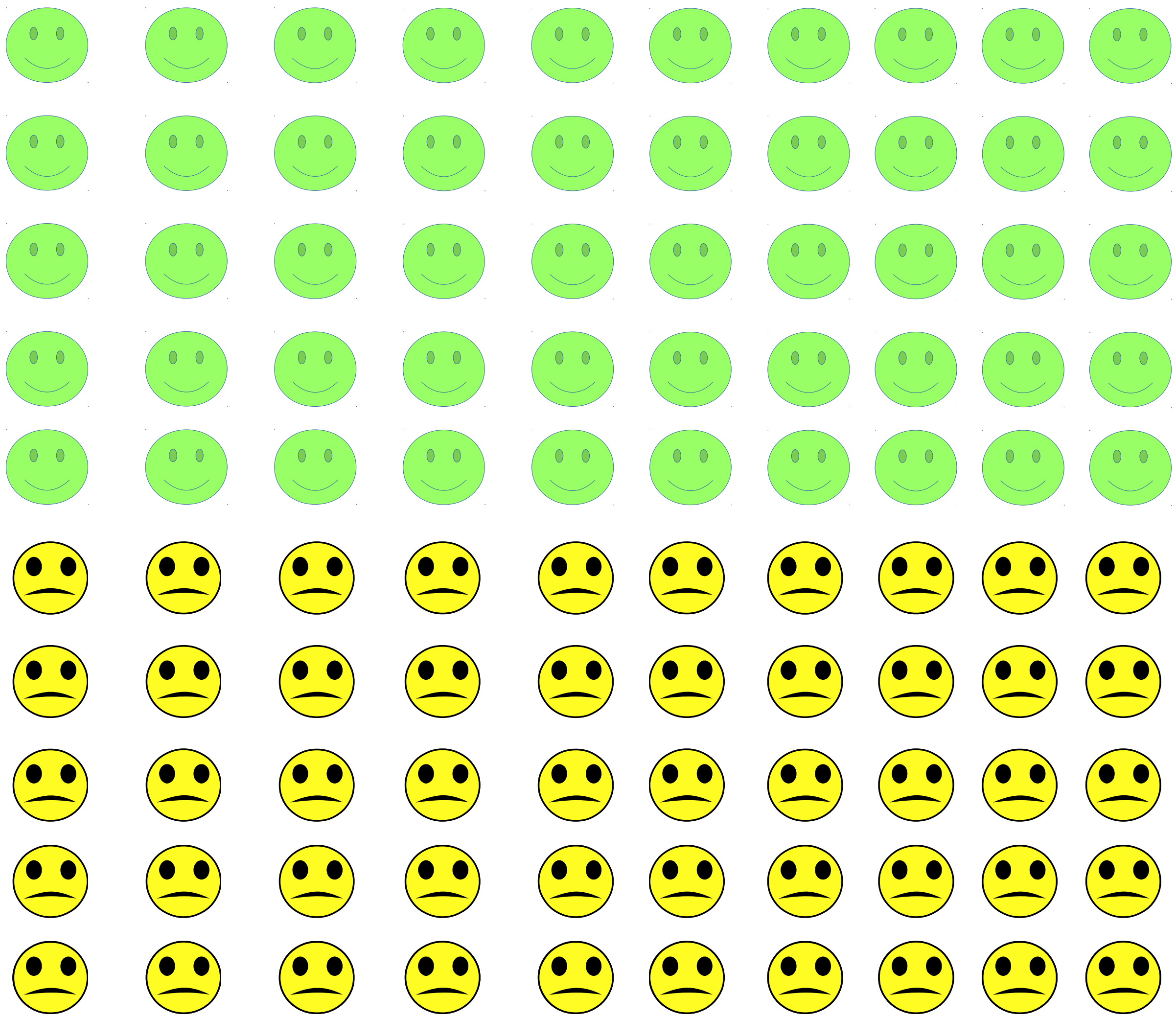
$$Pr(\text{I have the disease}|\text{I fail the test})$$

WPI

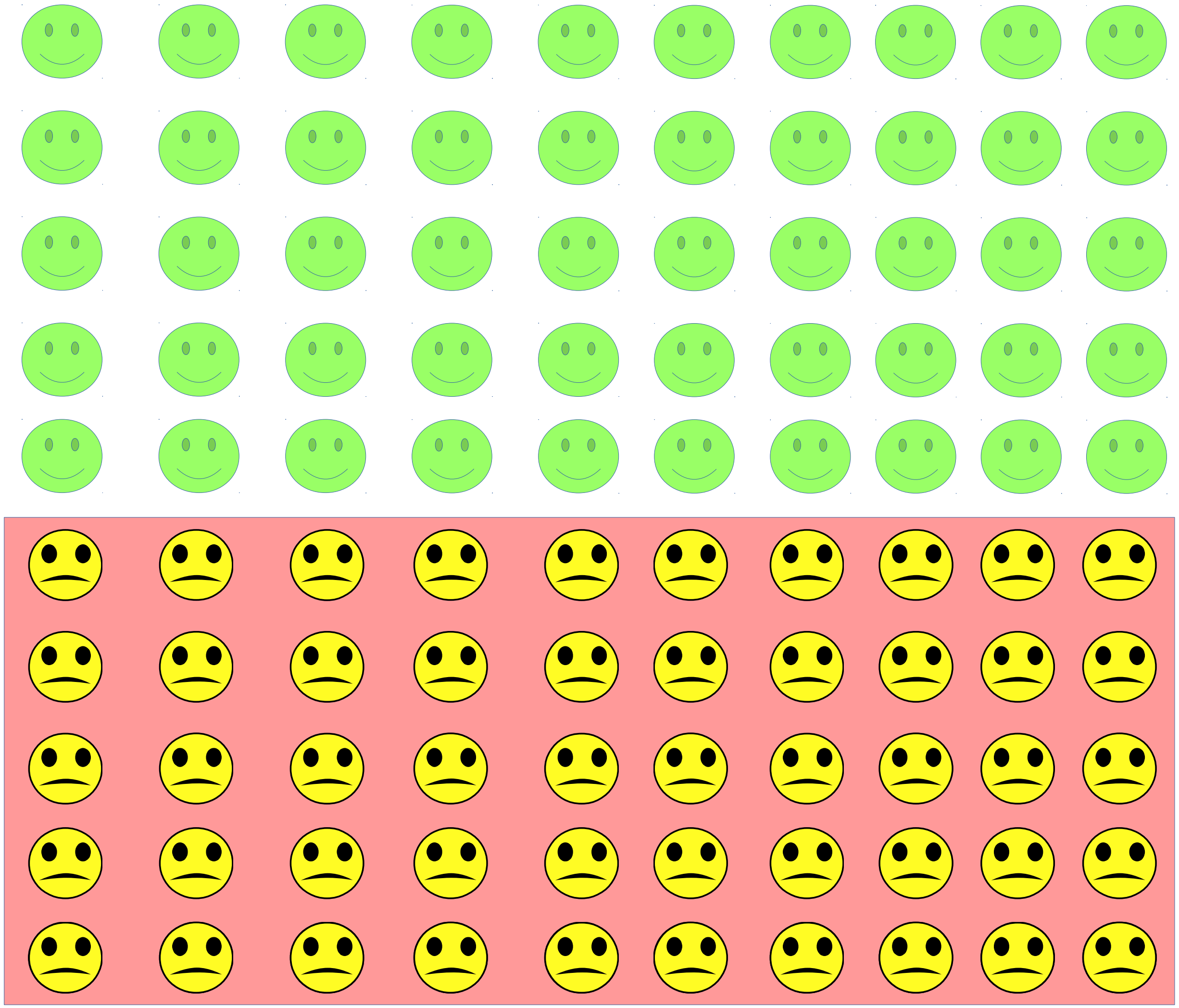
Base Rate Fallacy



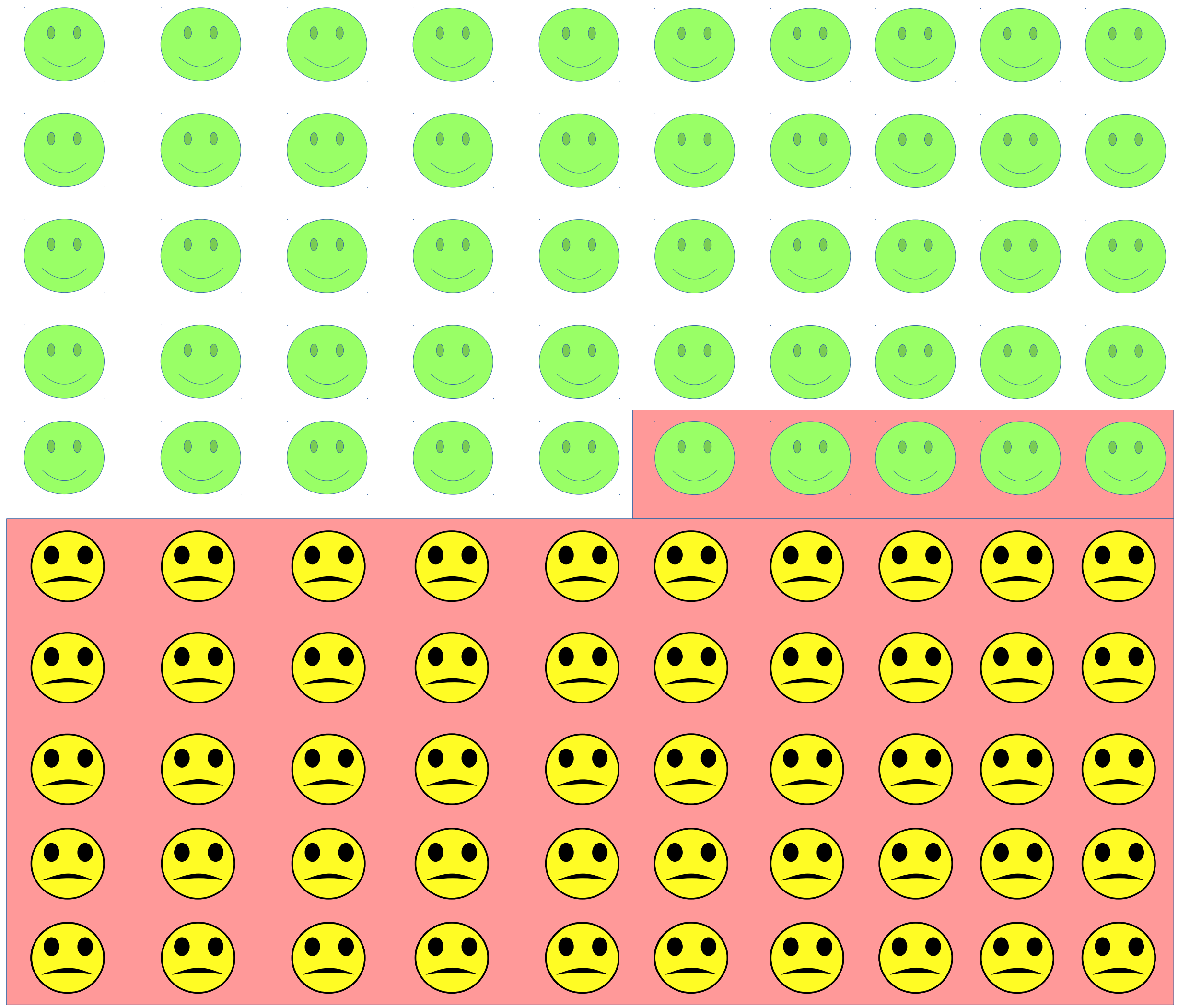
Base Rate Fallacy



Base Rate Fallacy

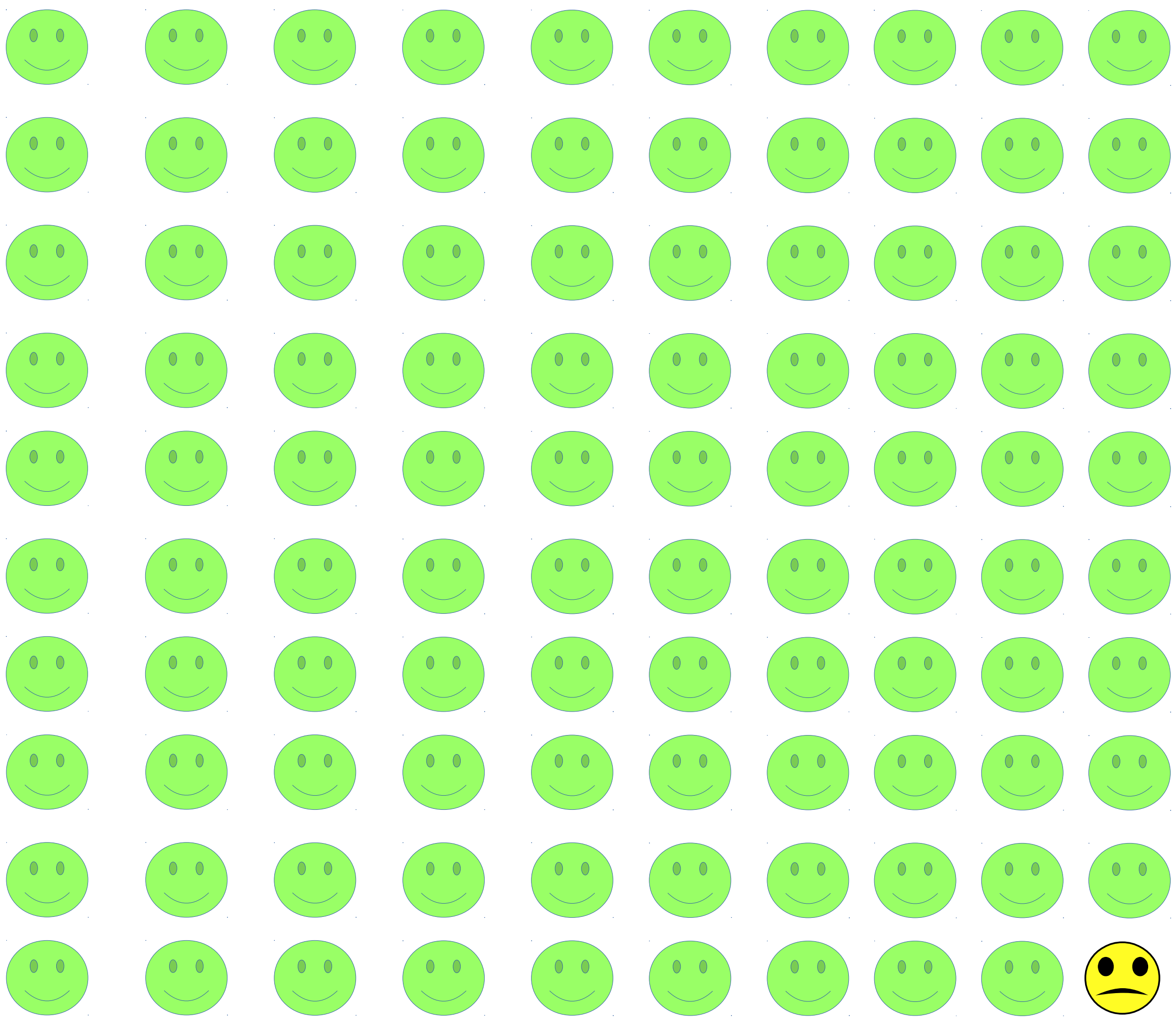


Base Rate Fallacy



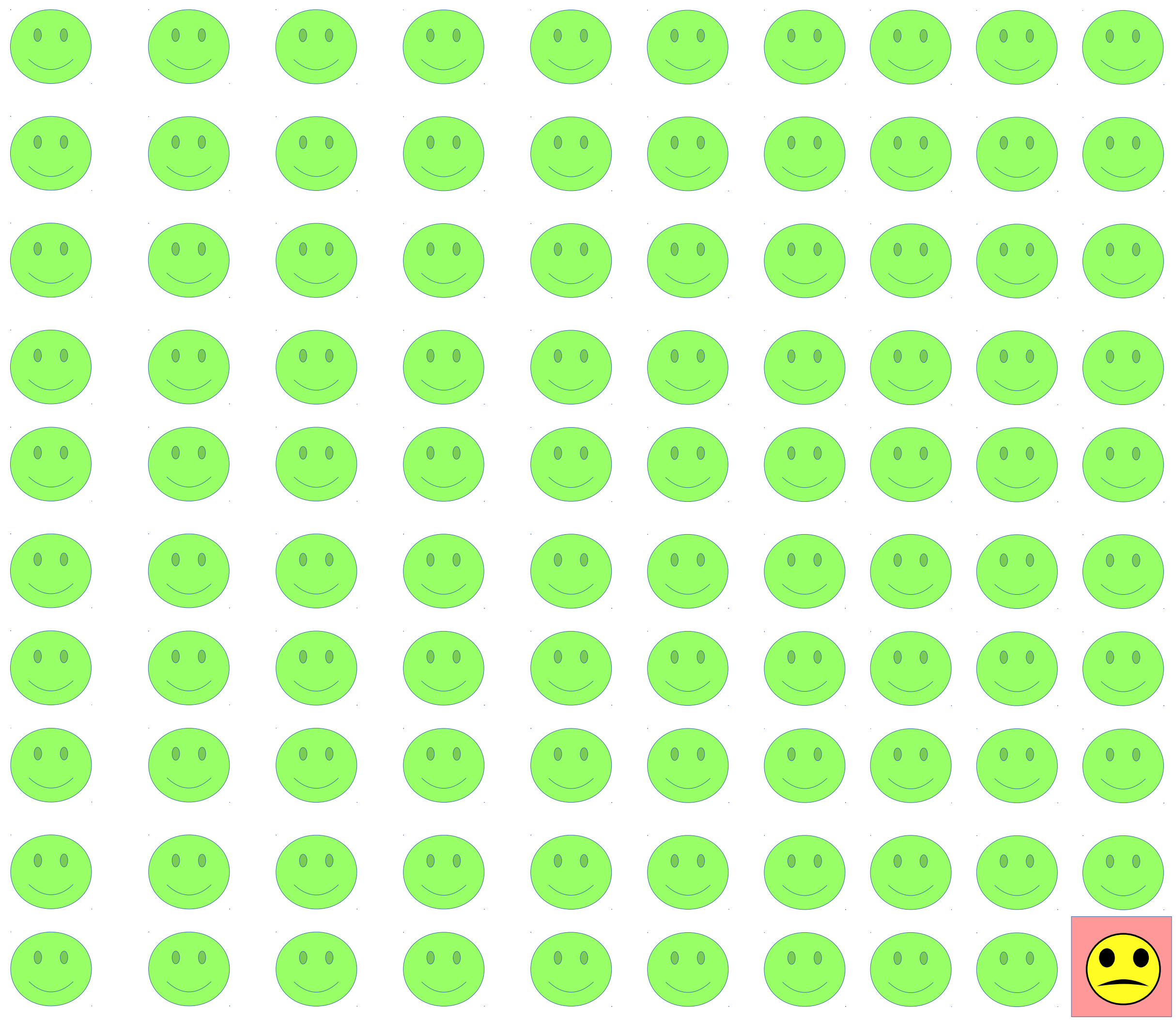
$$\frac{50}{55} = 90\%$$

Base Rate Fallacy

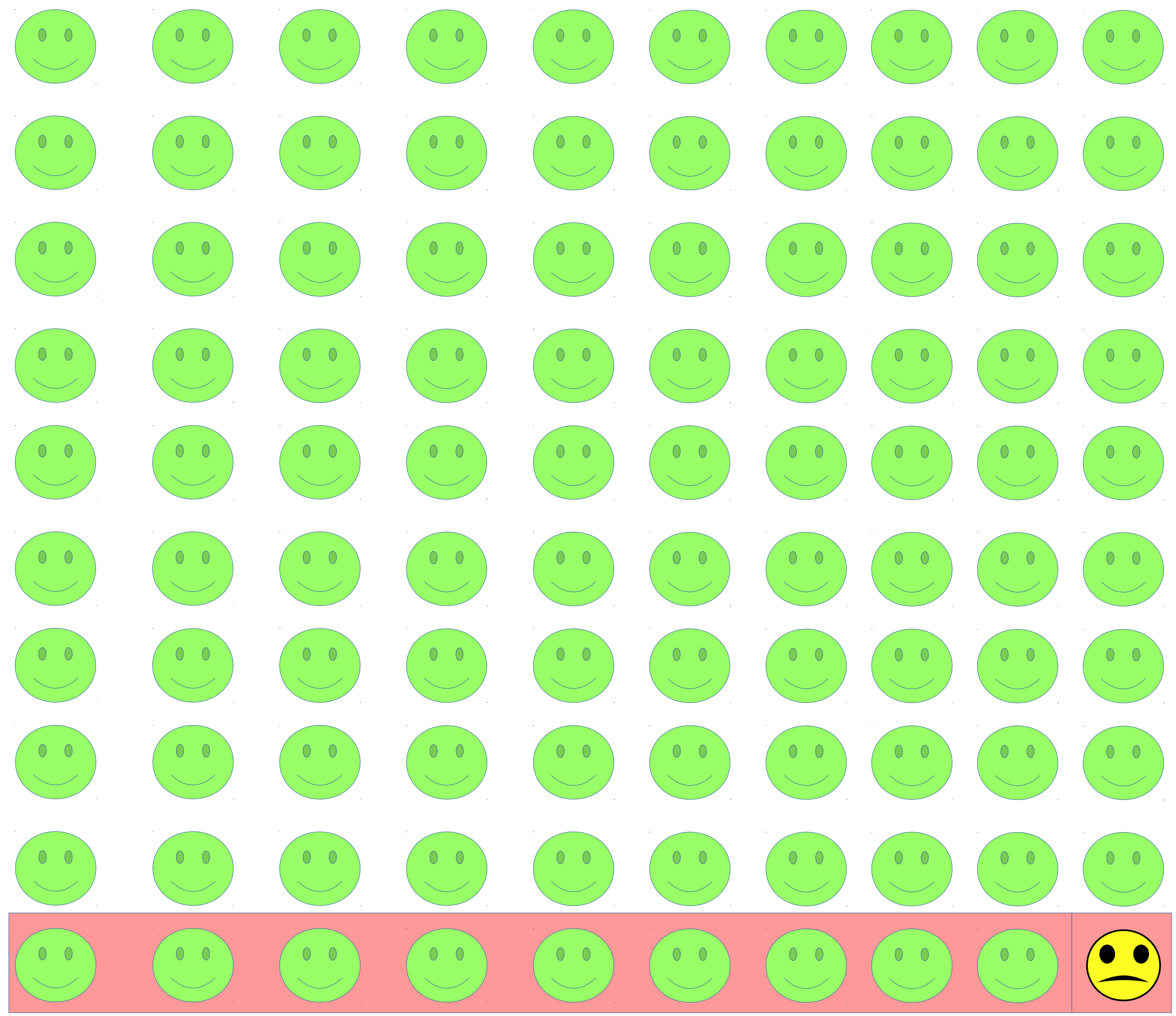


WPI

Base Rate Fallacy



Base Rate Fallacy



$\frac{1}{10}$

10%

The importance of asking the right question.

I want to know the *probability* that I have the disease *given* that I failed the test.

$$Pr(\text{I have the disease}|\text{I fail the test})$$

I do need to know the *probability* that I failed the test *given* that I have the disease.

$$Pr(\text{I fail the test}|\text{I have the disease})$$

I also need to know the *probability* that I have the disease.

$$Pr(\text{I have the disease})$$

I also need to know the *probability* that I failed the test.

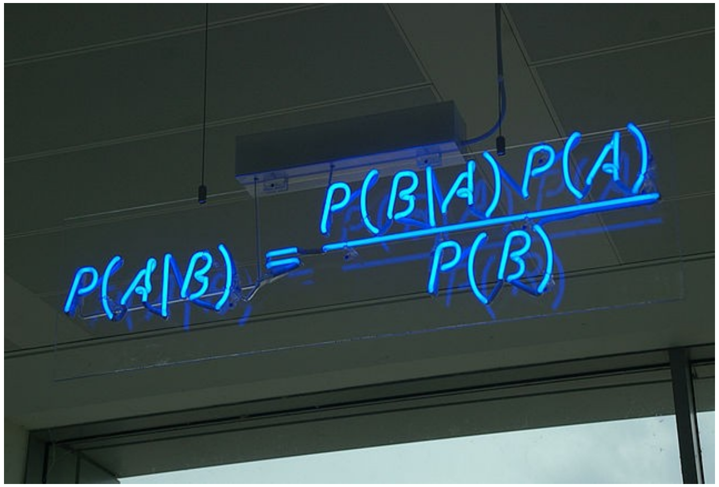
$$Pr(\text{I fail the test})$$

WPI

Bayes Theorem

$$Pr(\text{I have the disease}|\text{I fail the test}) = \frac{Pr(\text{I fail the test}|\text{I have the disease})Pr(\text{I have the disease})}{Pr(\text{I fail the test})}$$

"Thomas Bayes" by unknown - [2][3]. Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Thomas_Bayes.gif#/media/File:Thomas_Bayes.gif



"Bayes' Theorem MMB 01" by mattbuck (category) - Own work by mattbuck.. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Bayes%27_Theorem_MMB_01.jpg#/media/File:Bayes%27_Theorem_MMB_01.jpg

Even T-shirts!

https://www.google.com/search?site=&tbm=isch&source=hp&biw=1241&bih=518&q=bayes+theorem+t-shirt&oq=bayes+theorem+t-shirt&gs_l=img.3...371.4856.0.4955.21.7.0.9.9.0.231.555.0j1j2.3.0....0...1ac.1.64.img..15.6.580.yrkdHV_w79w

Detailed example

X	T=Y	T=N
have virus	1	0
not have virus	.01	.99

$\frac{1}{1000}$ have the disease

$$P(X='yes' \mid Y='yes') = \frac{P(Y='yes' \mid X='yes')\, P(X='yes')}{P(Y='yes')}$$

1,000,000 people

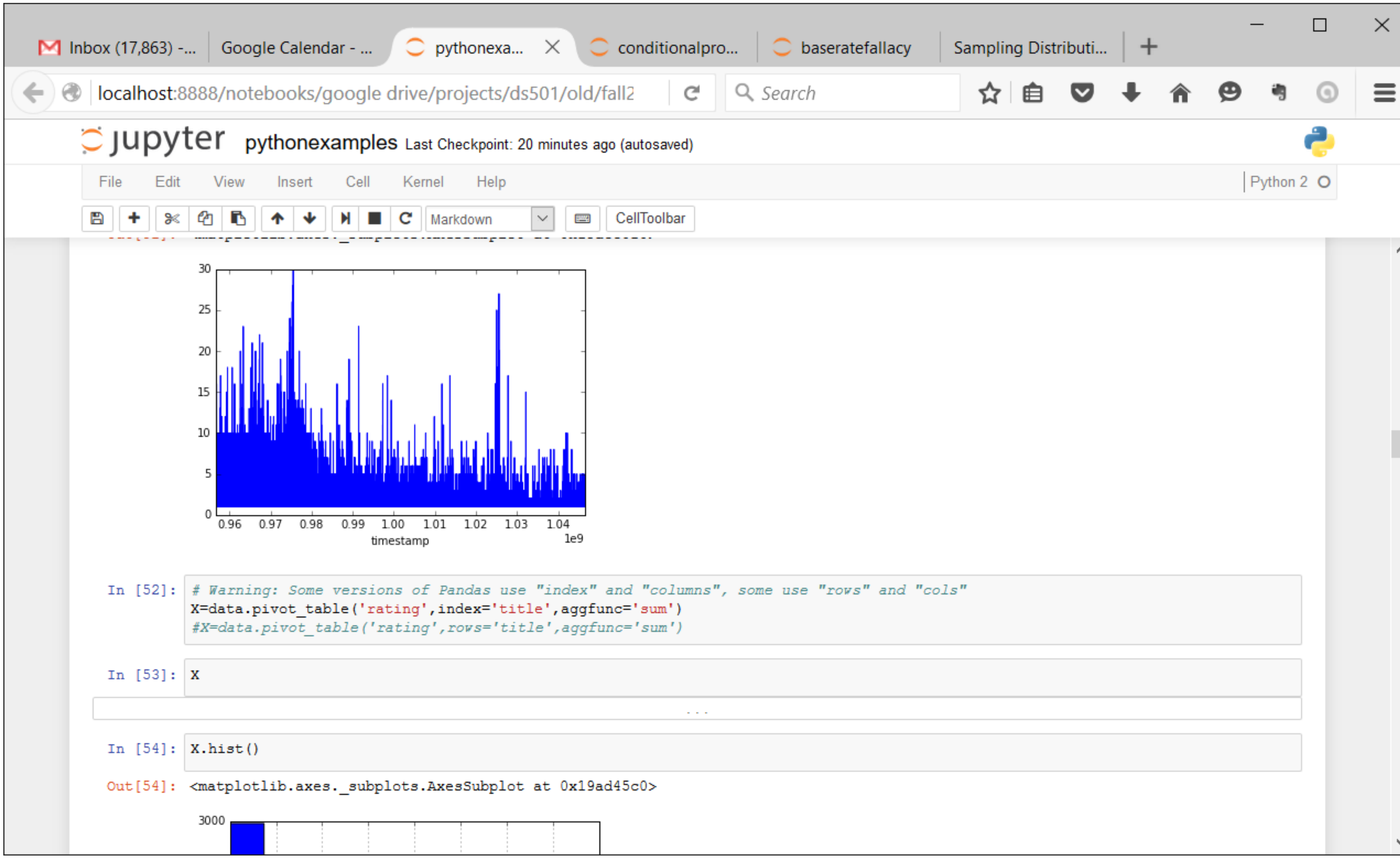
1,000 "yes" sick

~ 10,000 "yes" test went wrong

$$= \frac{1 \cdot 0.001}{\frac{11,000}{1,000,000}} \approx .1$$

WPI

Let's see it in python

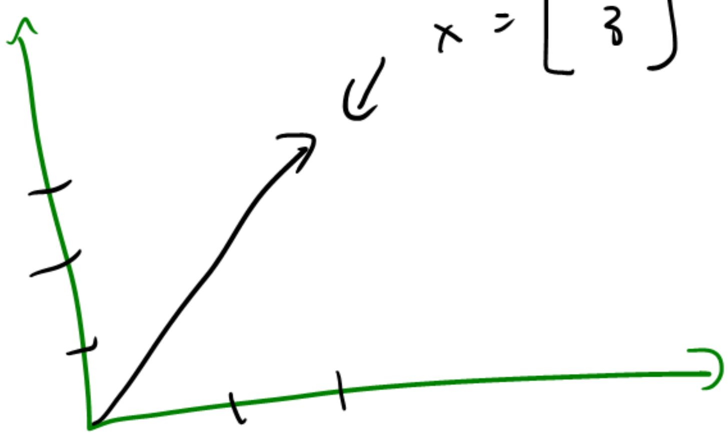


Linear Algebra: Vectors

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \qquad X \in \mathbb{R}^{n \times 1}$$

$\mathbb{C}$ complex $\quad \mathbb{I}$ integers

$$x = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$



Linear Algebra: Dot products

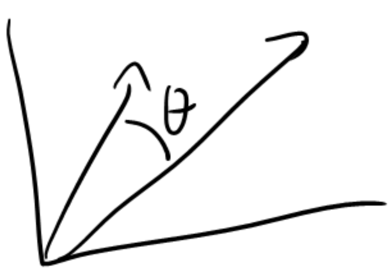
$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Transpose

$$X^T = [1, 2, 3] \qquad X^T Y = [1, 2, 3] \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

$$= 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 32$$

$$X^T Y = \|X\| \, \|Y\| \cos\theta$$



$$X^T X = [1, 2, 3] \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$= 1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3 = 14$$

Linear Algebra: Matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad A \in \mathbb{R}^{3 \times 3} \qquad y \in \mathbb{R}^{3 \times 1}$$

$$\underset{3\times3}{A}\ \underset{3\times1}{x} = \underset{3\times1}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1^2 + 2^2 + 3^2 \\ 4\cdot1 + 5\cdot2 + 6\cdot3 \\ 7\cdot1 + 8\cdot2 + 9\cdot3 \end{bmatrix}$$

Transformation

<u>Linear</u> transformation

Linear transformation!

Claim f is linear

1) $f(x+y) = f(x) + f(y)$

2) $f(\alpha x) = \alpha f(x)$

scalar

Linear Algebra: Identities and Inverses

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad Ix = x$$

eigenvalue

$$BA = I \qquad A^{-1}A = I \qquad Ax = \lambda x$$

when does A^{-1} ~~e~~exist

if $\exists x$ s.t.

$$Ax = 0, \quad x \neq 0$$

then A^{-1} does not exist $\Leftarrow$ singular

Let's see it in python