

# data science for (physical) scientists Ib

I: epistemological concepts and working environment

# 1 probability

frequentist vs Bayesian interpretation

basic probability arithmetic

# 2 statistics

distributions

moments

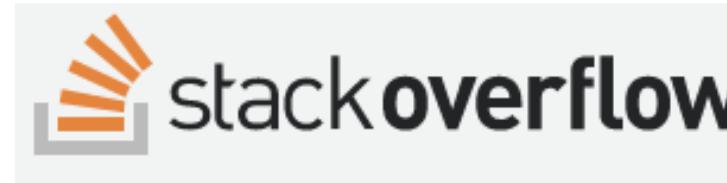
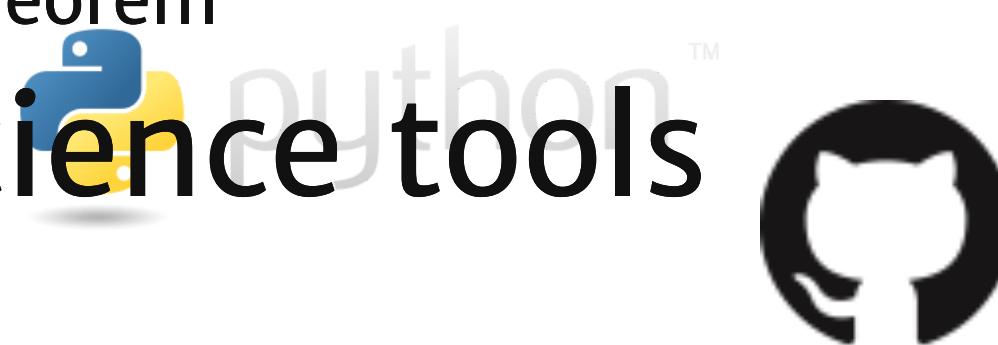
central limit theorem

# 3 data science tools

github

python

jupyter notebooks



# 3 data science tools

github

python

jupyter notebooks

google colab

stackoverflow

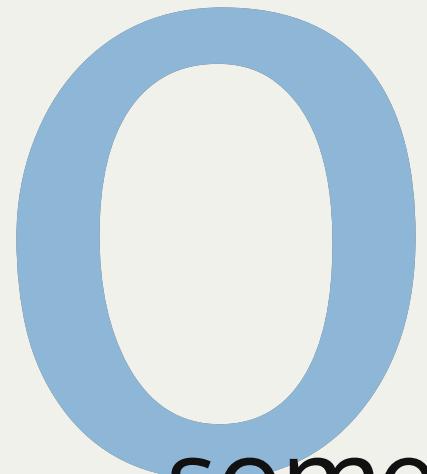


python™



this slide deck

[http://bit.ly/dsps2019\\_1](http://bit.ly/dsps2019_1)



some administrative stuff

# Syllabus

<http://bit.ly/dspssyllabus>

## Learning Outcomes

By the end of this class you should be able to formulate an appropriate analysis plan for a research question, select, gather, and prepare data for analysis, and choose and apply machine learning methods to the data.

# Syllabus

<http://bit.ly/dspssyllabus>

## Learning Outcomes

By the end of this class you should be able to formulate an appropriate analysis plan for a research question, select, gather, and prepare data for analysis, and choose and apply machine learning methods to the data.

The instructors is: Dr. **Federica Bianco** [fbianco@udel.edu](mailto:fbianco@udel.edu)

office hours: Tentatively: Tuesday 1230-2PM (or by appointment).

The Class assistants are:

Grader: **Yuqi Kong** [kongyq@udel.edu](mailto:kongyq@udel.edu)

Office Hours: Monday 5PM Smith Hall 220

Technical and coding questions

Physics instructor: Dr. **Alexandre David-Uraz** [adu@udel.edu](mailto:adu@udel.edu)

Office Hours: TBD Sharp 101B physics help center

Physics and Method-Application questions

# Syllabus

<http://bit.ly/dspssyllabus>

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

# quizz

<http://bit.ly/dspssyllabus>

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

*from beginning of class to 5 minutes past (be on time!)  
questions on previous class material and reading assignments*

# participation

<http://bit.ly/dspssyllabus>

- 10% pre-class questions ask questions
- 10% class performance and participation answer questions
- 20% homeworks get up and code
- 25% midterm extra credit assignments
- 35% final

# homework

<http://bit.ly/dspssyllabus>

- **10% pre-class questions**
- **10% class performance and participation**
- **20% homeworks**
- **25% midterm**
- **35% final**

Homework projects must be turned in as *jupyter notebooks* by checking them into your [github](#) account in a DSPS\_<firstinitialLastname> repo and the project directories HW<hw number> (unless otherwise stated).  
<finitialLastname> is e.g. fBianco

# homework

Please work in groups of up to 5 people on homework as a collaborative projects.

Individual notebooks must be returned for each homework. Different group members should lead different aspects of the work. A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example [these contributions](#)).

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

# homework

Please work in groups of up to 5 people on homework as a collaborative projects.

Individual notebooks must be returned for each homework. Different group members should lead different aspects of the work. A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example [these contributions](#)).

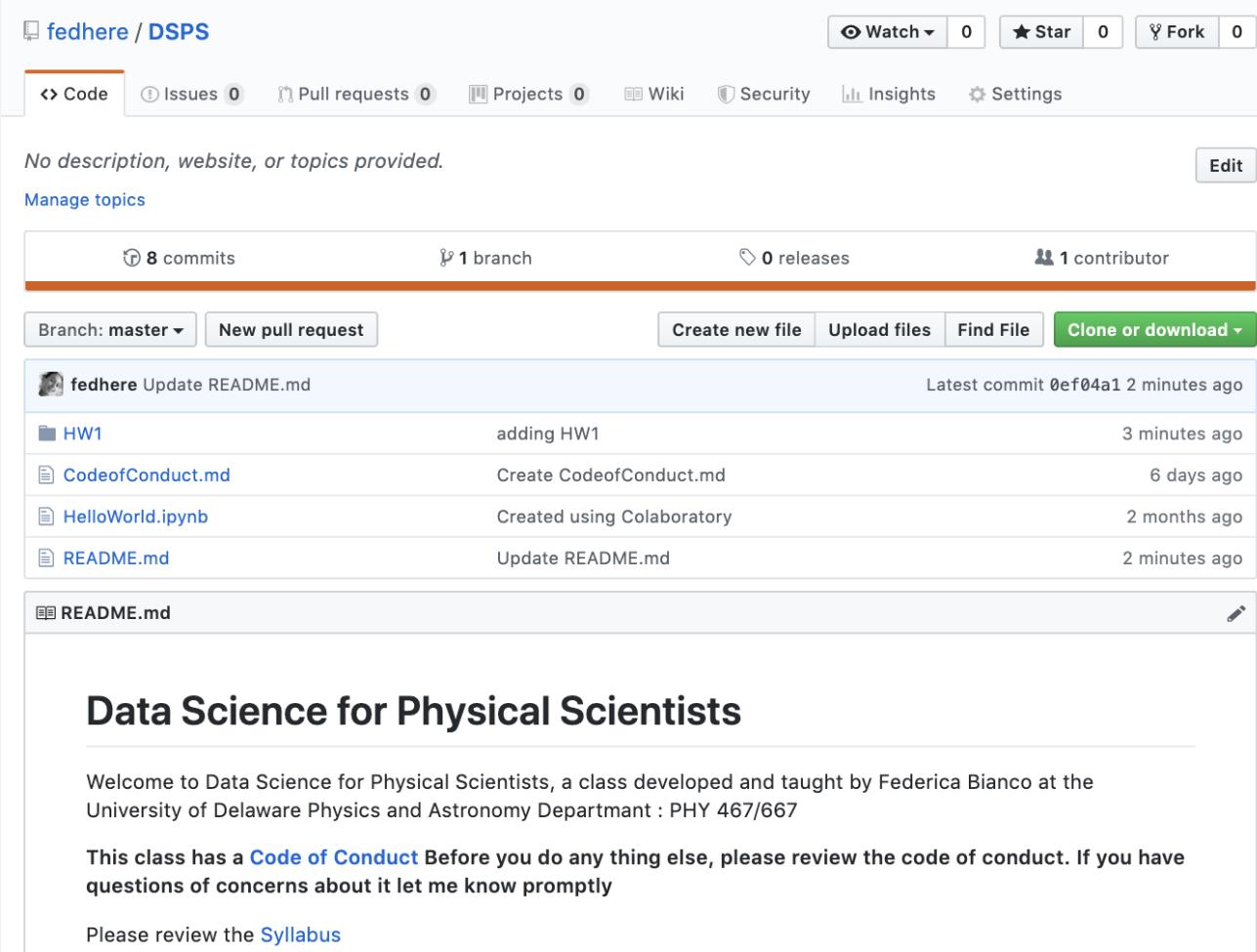
- **10% pre-class questions**
- **10% class performance and participation**
- **20% homeworks**
- **25% midterm**
- **35% final**

## Contributions

K.T. led the project and reduced the ALMA data. K.T. and D.I. wrote the manuscript. M.S.Y. reduced the Large Millimeter Telescope data and edited the final manuscript. Other authors contributed to the interpretation and commented on the ALMA proposal and the paper.

# homework

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

A screenshot of a GitHub repository page for "fedhere / DSPS". The repository has 8 commits, 1 branch, 0 releases, and 1 contributor. The README.md file contains the following content:  
**Data Science for Physical Scientists**  
Welcome to Data Science for Physical Scientists, a class developed and taught by Federica Bianco at the University of Delaware Physics and Astronomy Department : PHY 467/667  
This class has a [Code of Conduct](#) Before you do any thing else, please review the code of conduct. If you have questions of concerns about it let me know promptly  
Please review the [Syllabus](#)

instructions will be here

<https://github.com/fedhere/DSPS>

# homework

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

fedhere / DSPS

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

No description, website, or topics provided.

Edit

Manage topics

8 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find File Clone or download

fedhere Update README.md Latest commit 0ef04a1 2 minutes ago

HW1 adding HW1 3 minutes ago

CodeofConduct.md Create CodeofConduct.md 6 days ago

HelloWorld.ipynb Created using Colaboratory 2 months ago

README.md Update README.md 2 minutes ago

README.md

Data Science for Physical Scientists

Welcome to Data Science for Physical Scientists, a class developed and taught by Federica Bianco at the University of Delaware Physics and Astronomy Department : PHY 467/667

This class has a [Code of Conduct](#) Before you do any thing else, please review the code of conduct. If you have questions of concerns about it let me know promptly

Please review the [Syllabus](#)

instructions will be here  
<https://github.com/fedhere/DSPS>

online help: <http://bit.ly/2HtkQoc>  
please sign up asap!!



# homework

The screenshot shows a course page from UD Canvas. At the top left is the UD logo. To its right is the course identifier "19F-PHYS467/PHYS667-011". On the far right is a three-line menu icon. Below the course identifier is the text "2019 Fall". A vertical sidebar on the left contains links: Account (with a person icon), Dashboard (with a clock icon), Courses (with a book icon), Calendar (with a calendar icon), and Inbox (with an envelope icon). The "Courses" link is highlighted with a blue box. The main content area has a header "Recent Activity in 19F-PHYS467/PHYS667-011". Below it is a message box with an information icon: "No Recent Messages You don't have any messages to show in your stream yet. Once you begin participating in your courses you'll see this stream fill up with messages from discussions, grading updates, private messages between you and other users, etc." To the right of the message box is a "View Course Calendar" button with a calendar icon and a "To Do" section with the text "Nothing for now".

- **10% pre-class questions**
- **10% class performance and participation**
- **20% homeworks**
- **25% midterm**
- **35% final**

of course there is also UD Canvas, which will be used to give you grades and occasionally post messages (I am still learning how to use it tho!)

HW 1 is posted already

# midterm

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential.** In the Midterm and Final YOU WILL BE EXPECTED TO WORK INDIVIDUALLY.

**Midterm... probably in class**

*issues:* stereotype thread - working under derass is not necessarily a required skill  
*advantages:* interviews for jobs are often timed

# final

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential.** In the Midterm and Final YOU WILL BE **EXPECTED TO WORK INDIVIDUALLY.**

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

Final: take home, multiple days.

# Resources

<https://github.com/fedhere/DSPS>

- SLIDES here
- HOMEWORK INSTRUCTIONS here
- RESOURCES here

If notebooks do not display

use

<https://nbviewer.jupyter.org>

No description, website, or topics provided.

Manage topics

8 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find File Clone or download

fedhere Update README.md Latest commit 0ef04a1 2 minutes ago

HW1 adding HW1 3 minutes ago

CodeofConduct.md Create CodeofConduct.md 6 days ago

HelloWorld.ipynb Created using Colaboratory 2 months ago

README.md Update README.md 2 minutes ago

**Data Science for Physical Scientists**

Welcome to Data Science for Physical Scientists, a class developed and taught by Federica Bianco at the University of Delaware Physics and Astronomy Department : PHY 467/667

This class has a [Code of Conduct](#) Before you do any thing else, please review the code of conduct. If you have questions of concerns about it let me know promptly

Please review the [Syllabus](#)

**reap**  
the scientific method

(what is science?)

# PROGRAMMING

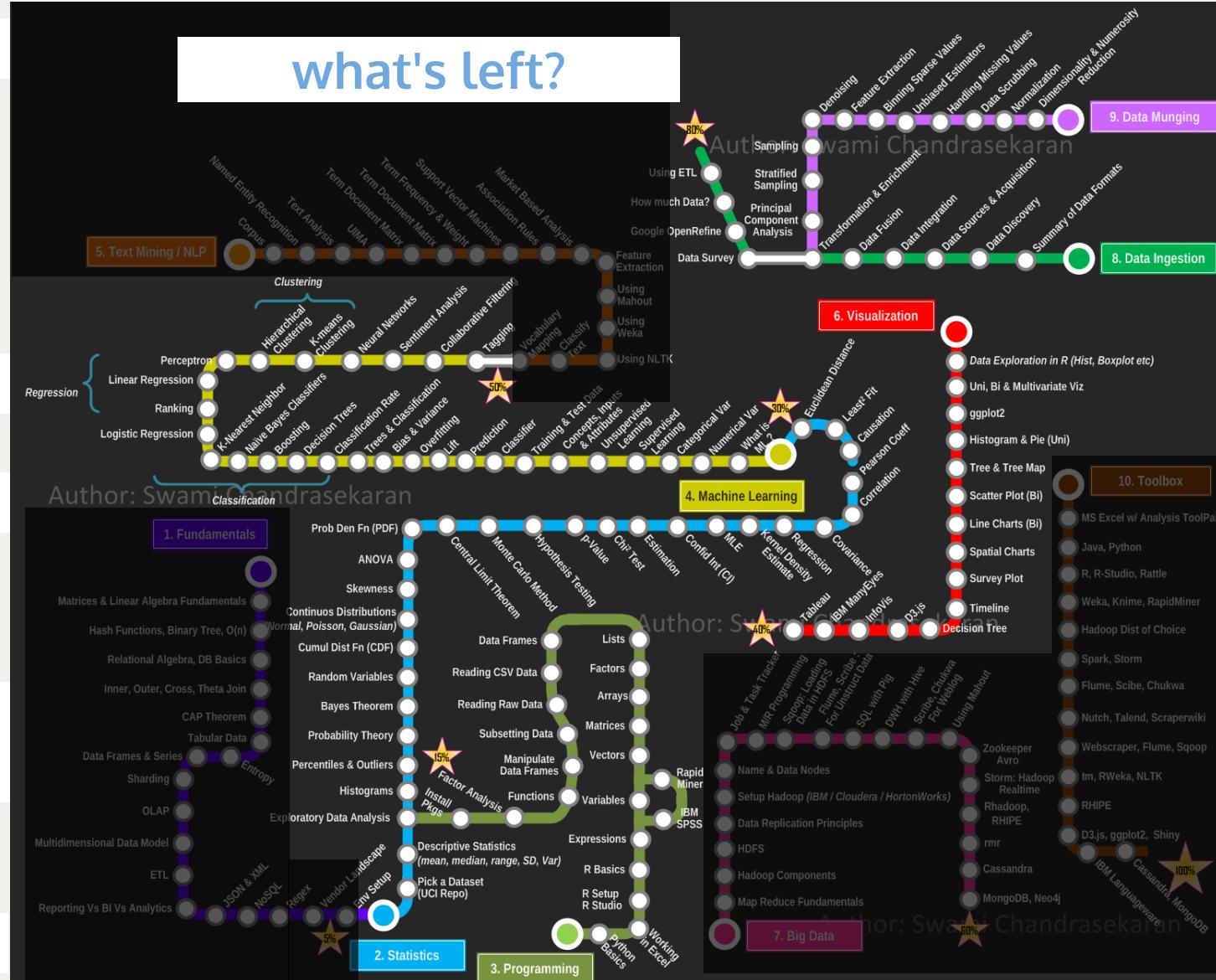
# STATISTICS

# DATA INGESTION

# DATA MUNGING

# MACHINE LEARNING

# VISUALIZATION



<http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist>

python

probability distributions

p-values

uncertainties

MCMC

regression

(linear, template)

classification

(trees, neural networks)

clustering

# *Reproducibility*

## **Reproducible research means:**

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

<https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf>

**Reproducible research in practice:**  
all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

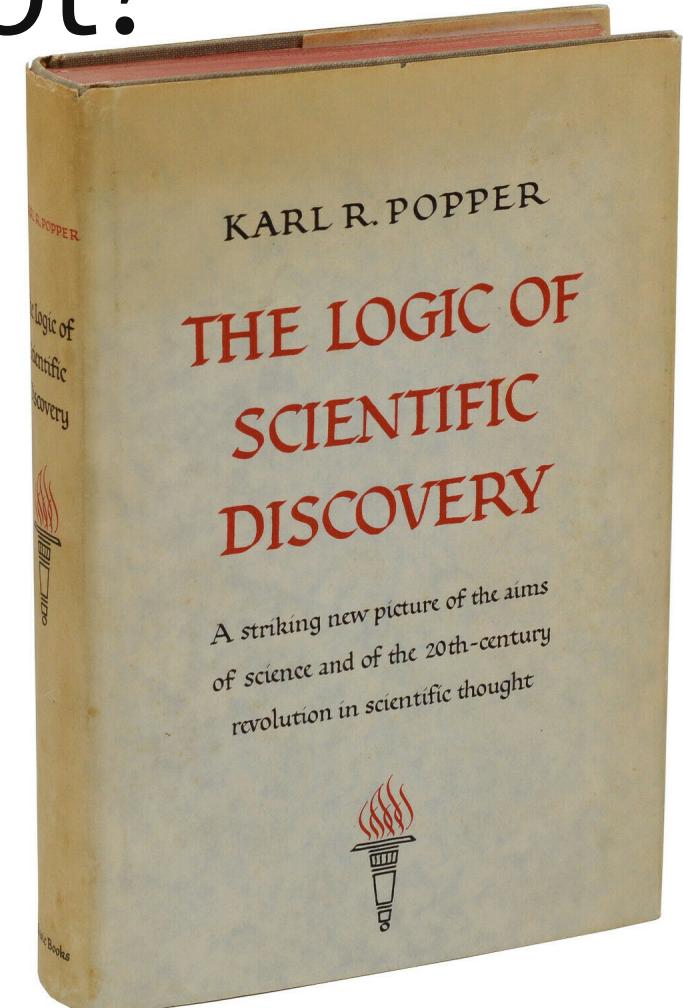
- provide raw data and code to reduce it to all stages needed to get outputs
- provide code to reproduce all figures
- provide code to reproduce all number outcomes

# the *demarcation* problem: what is science? what is not?

My proposal is based upon an *asymmetry* between **verifiability** and **falsifiability**; an asymmetry which results from the logical form of universal statements.

it is logically impossible to conclusively verify a universal proposition by reference to experience cause **not all examples could be tested**, but a single counter-instance conclusively falsifies the corresponding universal law.

—Karl Popper, *The Logic of Scientific Discovery*

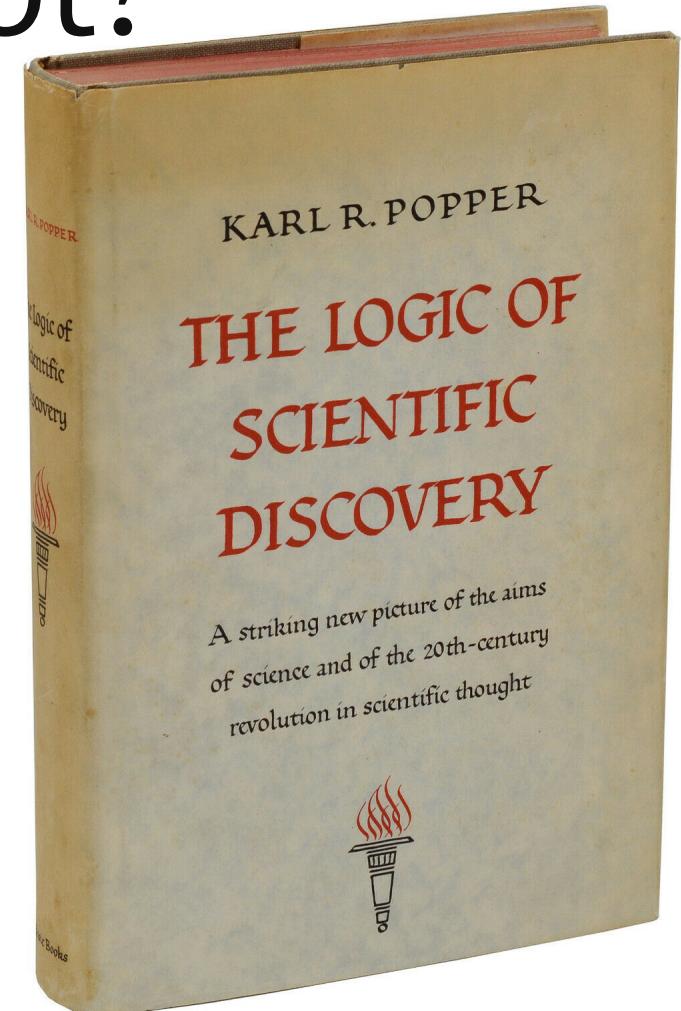


# the *demarcation* problem: what is science? what is not?

My proposal is based upon an *asymmetry* between **verifiability** and **falsifiability**; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

—Karl Popper, *The Logic of Scientific Discovery*

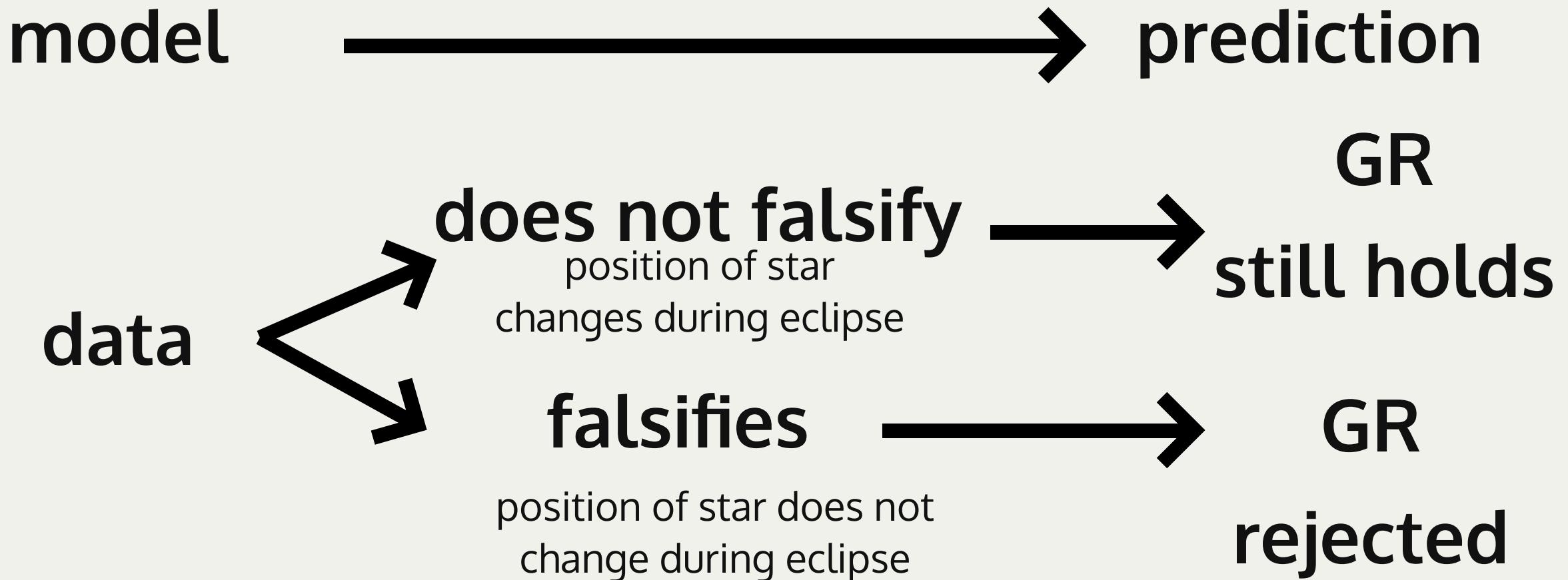
a scientific theory must be  
*falsifiable*



# the *demarcation* problem

model —————→ prediction

# the *demarcation* problem



the *demarcation* problem

is psychology a science?

DISCUSS!

# the *demarcation* problem

things can get more complicated though:

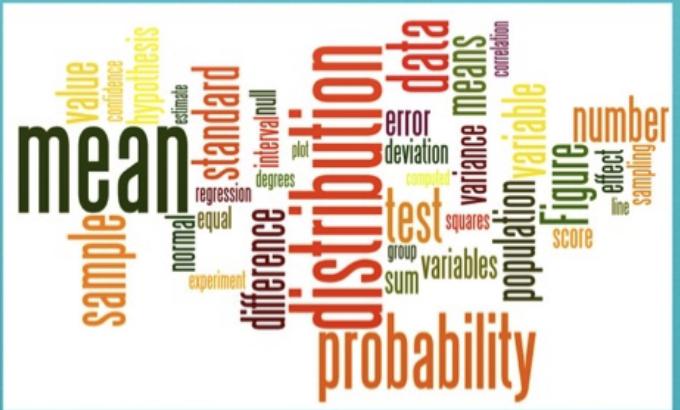
most scientific theories are actually based largely on *probabilistic induction* and modern *inductive inference* (Solomonoff, frequentist vs Bayesian methods...)

# 1

probability and statistics

First Edition

## Introduction to Statistics: An Interactive e-Book



David M. Lane (Editor, Primary author, and Designer)

# Introduction to Statistics: An Interactive e-Book

David M. Lane

## Crush Course in Statistics

freee statistics book: <http://onlinestatbook.com/>

what are probability and  
statistics?

# 1

## probability

# Basic Probability

## *Frequentist* interpretation

fraction of times something happens



probability of it happening



# Basic Probability

## Bayesian interpretation

represents a level of certainty relating to a potential outcome or idea:

*if I believe the coin is unfair (tricked)  
then even if I get a head and a tail I  
will still believe I am more likely to  
get heads than tails*

# Basic Probability

## *Frequentist* interpretation

$P(E)$  = frequency of E

$P(\text{coin} = \text{head}) = 6/11 = 0.55$

fraction of times something happens



probability of it happening



# Basic Probability

## *Frequentist* interpretation

fraction of times something happens



probability of it happening

$P(E)$  = frequency of E

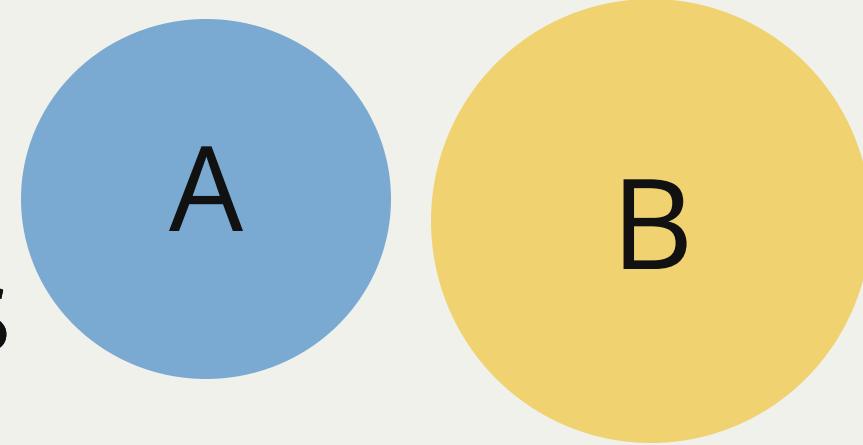
$P(\text{coin} = \text{head}) = 6/11 = 0.55$

$P(\text{coin} = \text{head}) = 51/100 = 0.51$





# Basic probability arithmetics



Probability Arithmetic

$$0 \leq P(A) \leq 1$$

$$P(A) + P(\bar{A}) = 1$$

disjoint events

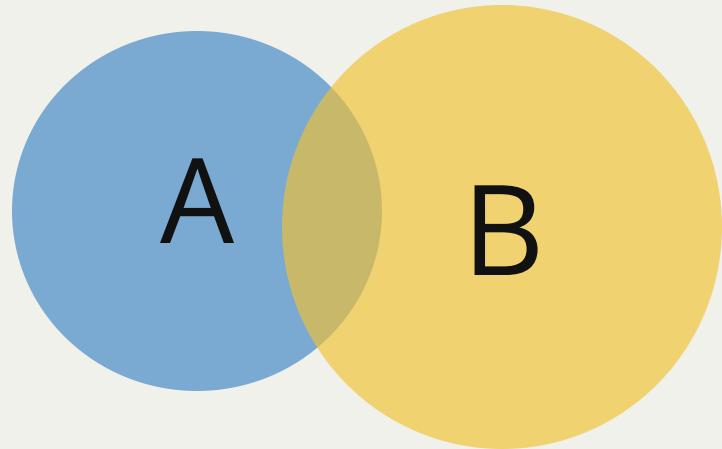
if  $P(A) \cap P(B) = 0$  then :

$$P(A \text{or } B) = P(A) + P(B)$$

$$P(A \text{and } B) = P(A) * P(B)$$

$$P(A|B) = P(A)$$

# Basic probability arithmetics



## Probability Arithmetic

in general :

dependent events

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

$$P(A|B) < P(A)$$

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

# Basic probability arithmetics

Probability Arithmetic

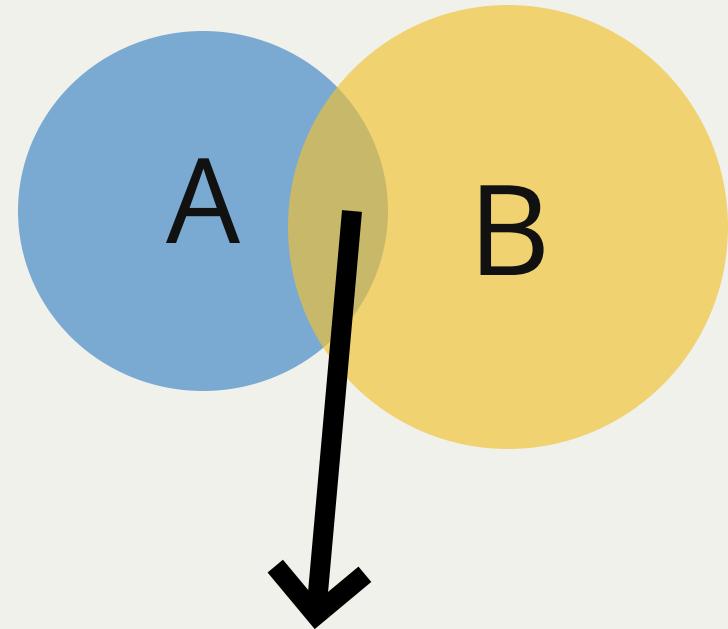
in general :

dependent events

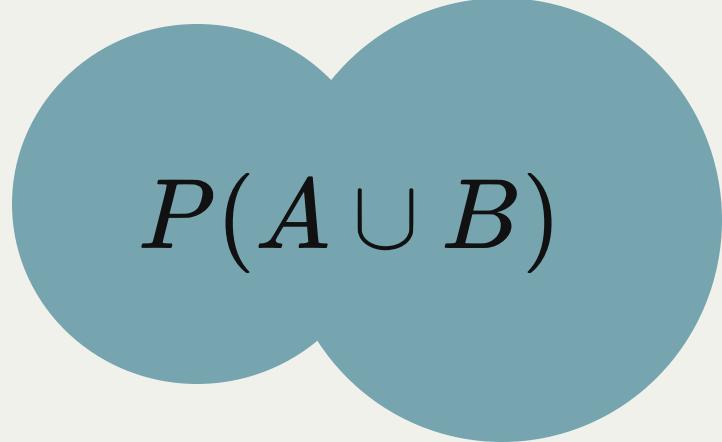
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) < P(A)$$

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



$$P(A \cap B)$$


$$P(A \cup B)$$

# Basic probability arithmetics

## Probability Arithmetic

in general :

dependent events

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

$$P(A|B) < P(A)$$

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

2

statistics

# statistics

takes us from observing a limited  
number of samples to infer on the  
population

# TAXONOMY

**Distribution:** a formula (a model)

**Population:** all of the elements of a "family"

**Sample:** a finite subset of the population that you observe

**HW**

# Phsyics Example

describe properties of the Population while  
the population is too large to be observed.

Statistical Mechanics:  
explains the properties of the macroscopic  
system by statiscal knowledge of the  
microscopic system, even the the state of  
each element of the system cannot be  
known exactly

example: Maxwell Boltzman distribution of  
velocity of molecules in an ideal gas

[https://upload.wikimedia.org/wikipedia/commons/8/82/Simulation\\_of\\_gas\\_for\\_relaxation\\_demonstration.gif?1567607773826](https://upload.wikimedia.org/wikipedia/commons/8/82/Simulation_of_gas_for_relaxation_demonstration.gif?1567607773826)

# Phsyics Example

Boltzmann 1872

The mechanical theory of heat assumes that the molecules of a gas are not at rest, but rather are in the liveliest motion. Hence, even though the body does not change its state, its individual molecules are always changing their states of motion, and the various molecules take up many different positions with respect to each other. The fact that we nevertheless observe completely definite laws of behaviour of warm bodies is to be attributed to the circumstance that the most random events, when they occur in the same proportions, give the same average value. For the molecules of the body are indeed so numerous, and their motion is so rapid,

**HW**

# Phsyics Example

Boltzmann 1872

that we can perceive nothing more than average values. One must compare the regularity of these average values with the amazing constancy of the average numbers provided by statistics, which are also derived from processes each of which is determined by a completely unpredictable interaction with many other factors.

One must not confuse an incompletely known law, whose validity is therefore in doubt, with a completely known law of the calculus of probabilities; the latter, like the result of any other calculus, is a necessary consequence of definite premises, and is confirmed insofar as these are correct, by experiment, provided sufficient

# Probability distributions

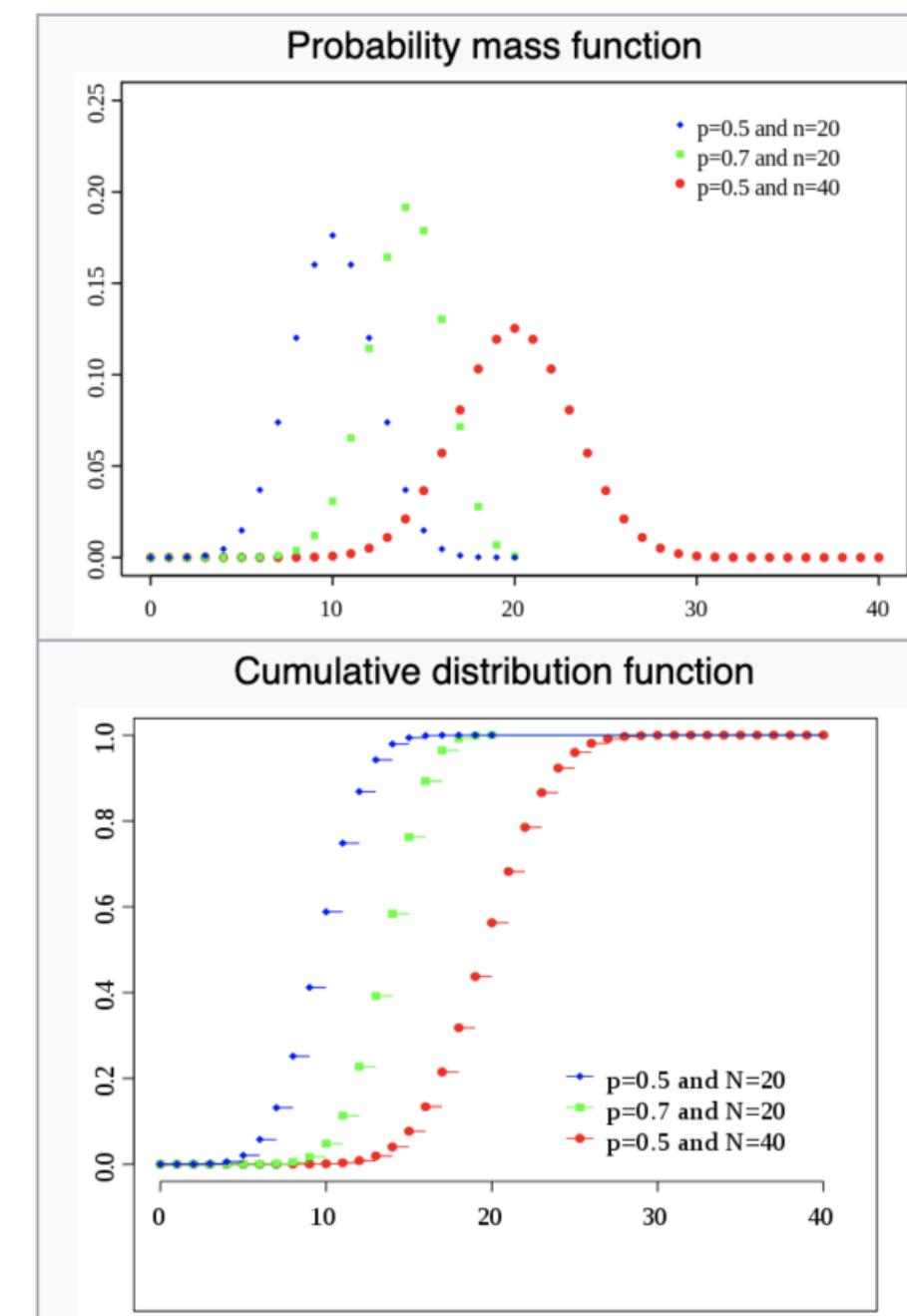
## Binomial

Coin toss:

fair coin:  $p=0.5$   $n=1$

Vegas coin:  $p \neq 0.5$   $n=1$

### Binomial distribution



<b>Notation</b>	$B(n, p)$
<b>Parameters</b>	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
<b>Support</b>	$k \in \{0, 1, \dots, n\}$ – number of successes
<b>pmf</b>	$\binom{n}{k} p^k (1-p)^{n-k}$
<b>CDF</b>	$I_{1-p}(n - k, 1 + k)$
<b>Mean</b>	$np$
<b>Median</b>	$\lfloor np \rfloor$ or $\lceil np \rceil$
<b>Mode</b>	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
<b>Variance</b>	$np(1-p)$
<b>Skewness</b>	$\frac{1-2p}{\sqrt{np(1-p)}}$
<b>Ex. kurtosis</b>	$\frac{1-6p(1-p)}{np(1-p)}$
<b>Entropy</b>	$\frac{1}{2} \log_2(2\pi enp(1-p)) + O\left(\frac{1}{n}\right)$ in <a href="#">shannons</a> . For <a href="#">nats</a> , use the natural log in the log.
<b>MGF</b>	$(1-p + pe^t)^n$
<b>CF</b>	$(1-p + pe^{it})^n$
<b>PGF</b>	$G(z) = [(1-p) + pz]^n$
<b>Fisher information</b>	$g_n(p) = \frac{n}{p(1-p)}$ (for fixed $n$ )

# Probability distributions

## Binomial

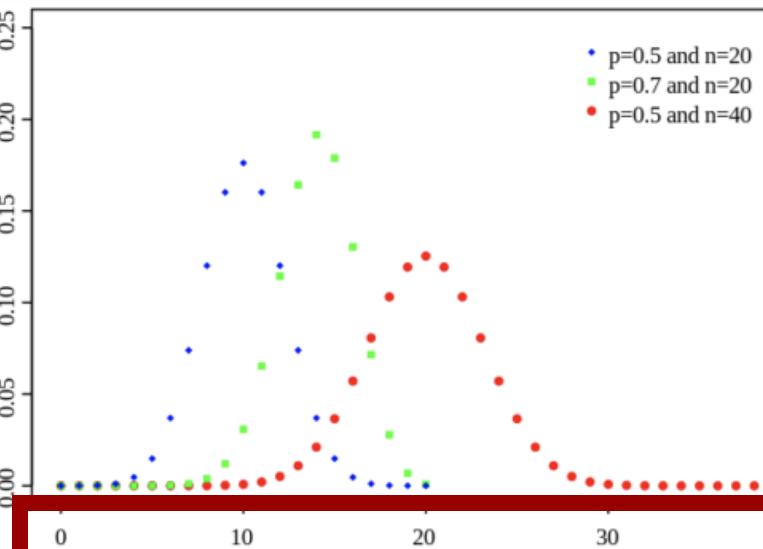
Coin toss:

fair coin:  $p=0.5$   $n=1$

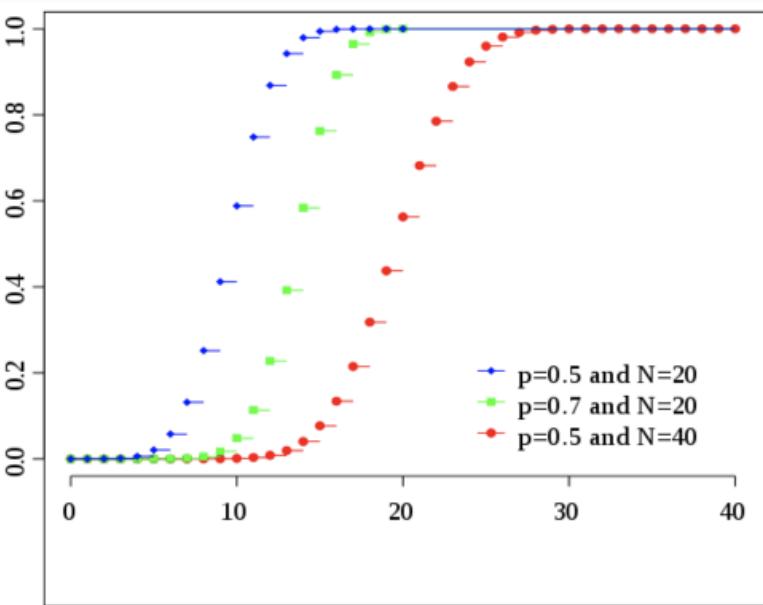
Vegas coin:  $p \neq 0.5$   $n=1$

### Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
Support	$k \in \{0, 1, \dots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n - k, 1 + k)$
Mean	$np$
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Ex. kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2} \log_2(2\pi enp(1-p)) + O\left(\frac{1}{n}\right)$ in <a href="#">shannons</a> . For <a href="#">nats</a> , use the natural log in the log.
MGF	$(1-p + pe^t)^n$
CF	$(1-p + pe^{it})^n$
PGF	$G(z) = [(1-p) + pz]^n$
Fisher information	$g_n(p) = \frac{n}{p(1-p)}$ (for fixed $n$ )

# Probability distributions

## Binomial

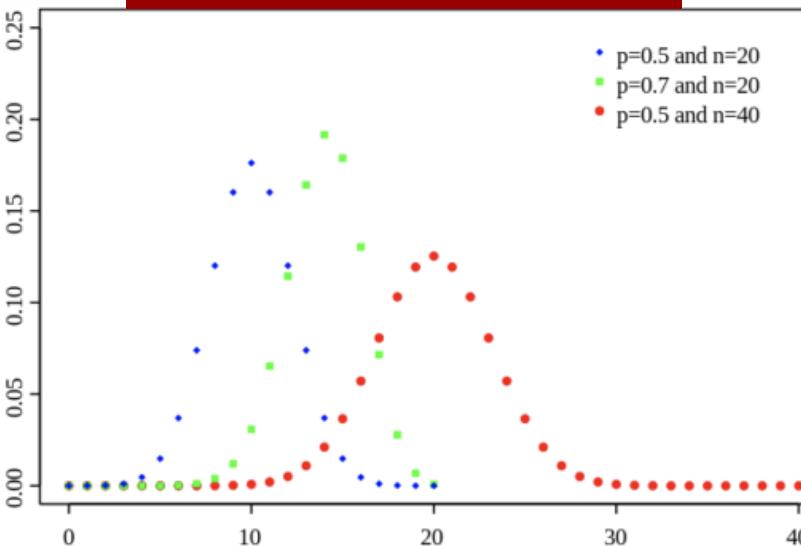
Coin toss:

fair coin:  $p=0.5$   $n=1$

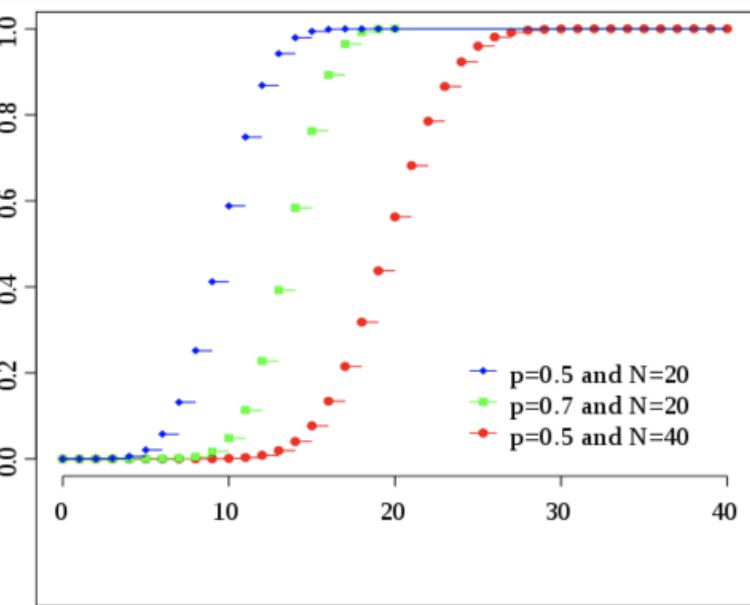
Vegas coin:  $p \neq 0.5$   $n=1$

### Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
Support	$k \in \{0, 1, \dots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$F_{1-p}(n - k, 1 + k)$
Mean	$np$
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Ex. kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2} \log_2(2\pi enp(1-p)) + O\left(\frac{1}{n}\right)$ in <a href="#">shannons</a> . For <a href="#">nats</a> , use the natural log in the log.
MGF	$(1-p + pe^t)^n$
CF	$(1-p + pe^{it})^n$
PGF	$G(z) = [(1-p) + pz]^n$
Fisher information	$g_n(p) = \frac{n}{p(1-p)}$ (for fixed $n$ )

# Probability distributions

## Binomial

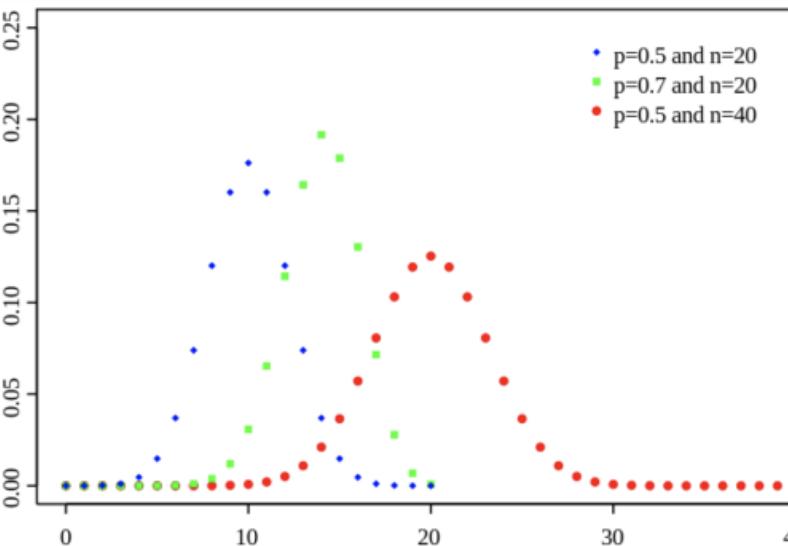
Coin toss:

fair coin:  $p=0.5$   $n=1$

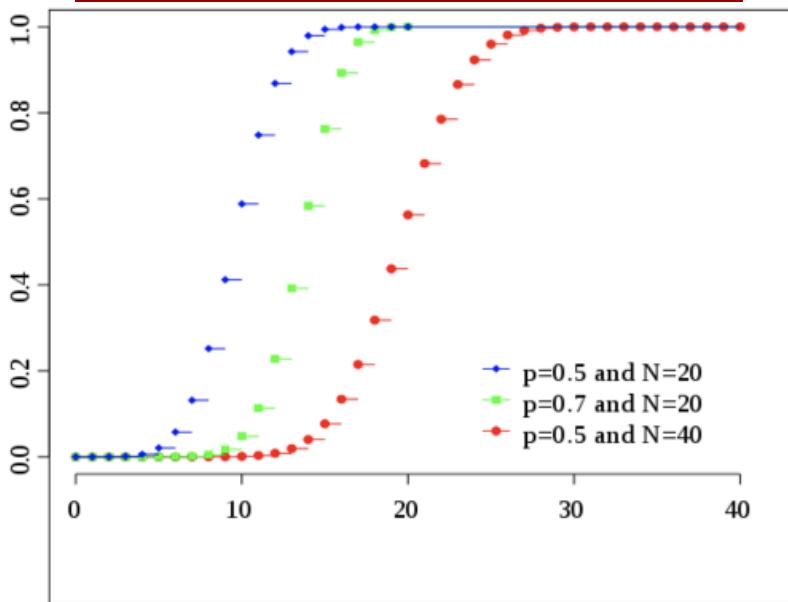
Vegas coin:  $p \neq 0.5$   $n=1$

### Binomial distribution

Probability mass function



Cumulative distribution function



<b>Notation</b>	$B(n, p)$
<b>Parameters</b>	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
<b>Support</b>	$k \in \{0, 1, \dots, n\}$ – number of successes
<b>pmf</b>	$\binom{n}{k} p^k (1 - p)^{n-k}$
<b>CDF</b>	$I_{1-p}(n - k, 1 + k)$
<b>Mean</b>	$np$
<b>Median</b>	$\lfloor np \rfloor$ or $\lceil np \rceil$
<b>Mode</b>	$\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
<b>Variance</b>	$np(1 - p)$
<b>Skewness</b>	$\frac{1 - 2p}{\sqrt{np(1 - p)}}$
<b>Ex. kurtosis</b>	$\frac{1 - 6p(1 - p)}{np(1 - p)}$
<b>Entropy</b>	$\frac{1}{2} \log_2(2\pi enp(1 - p)) + O\left(\frac{1}{n}\right)$ in <a href="#">shannons</a> . For <a href="#">nats</a> , use the natural log in the log.
<b>MGF</b>	$(1 - p + pe^t)^n$
<b>CF</b>	$(1 - p + pe^{it})^n$
<b>PGF</b>	$G(z) = [(1 - p) + pz]^n$
<b>Fisher information</b>	$g_n(p) = \frac{n}{p(1 - p)}$ (for fixed $n$ )

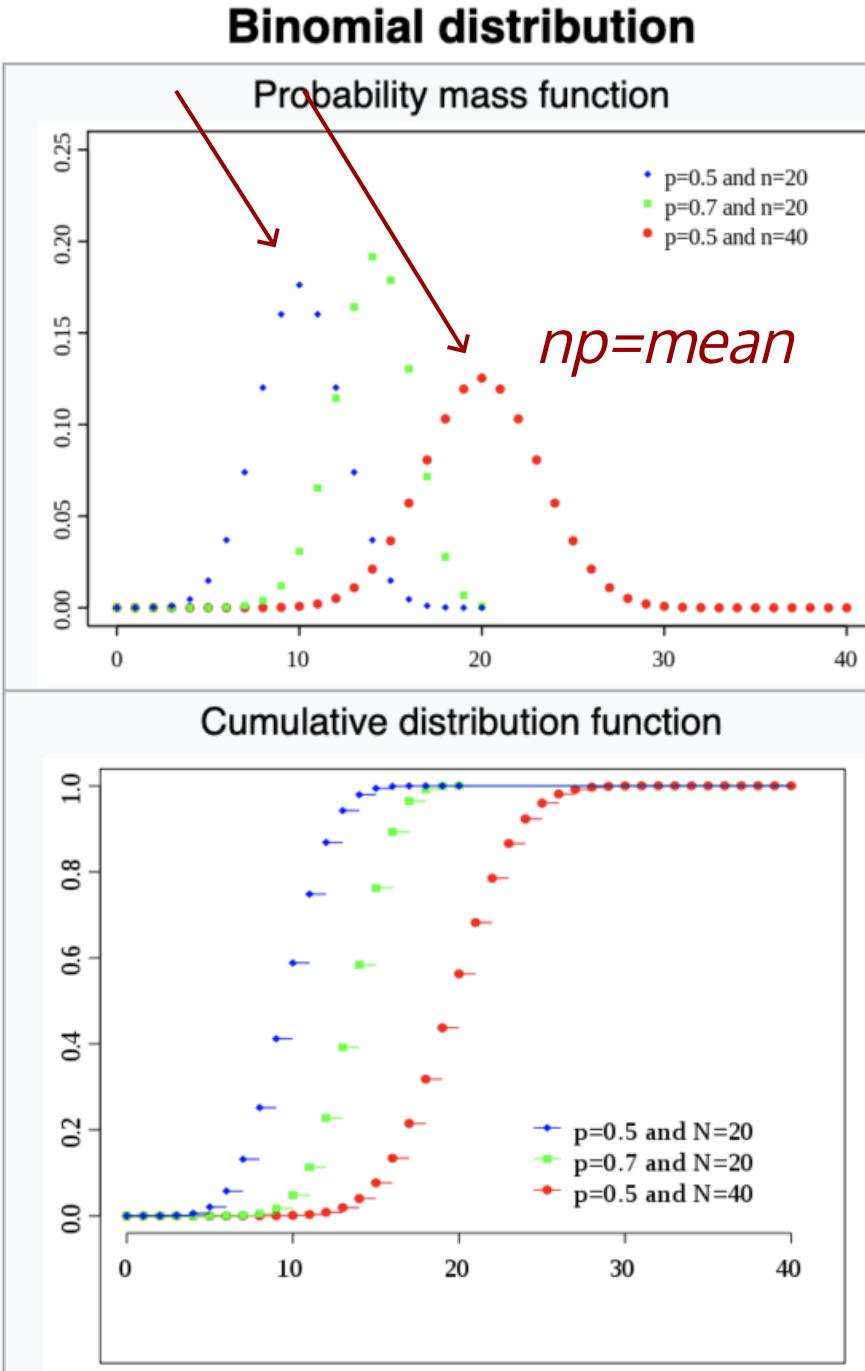
# Probability distributions

## Binomial

Coin toss:

fair coin:  $p=0.5$   $n=1$

Vegas coin:  $p \neq 0.5$   $n=1$



<b>Notation</b>	$B(n, p)$
<b>Parameters</b>	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
<b>Support</b>	$k \in \{0, 1, \dots, n\}$ – number of successes
<b>pmf</b>	$\binom{n}{k} p^k (1-p)^{n-k}$
<b>CDF</b>	$F(k) = \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i}$ <i>central tendency</i>
<b>Mean</b>	$np$
<b>Median</b>	$\lfloor np \rfloor$ or $\lceil np \rceil$
<b>Mode</b>	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
<b>Variance</b>	$np(1-p)$
<b>Skewness</b>	$\frac{1-2p}{\sqrt{np(1-p)}}$
<b>Ex. kurtosis</b>	$\frac{1-6p(1-p)}{np(1-p)}$
<b>Entropy</b>	$\frac{1}{2} \log_2(2\pi e n p(1-p)) + O\left(\frac{1}{n}\right)$ in <b>shannons</b> . For <b>nats</b> , use the natural log in the log.
<b>MGF</b>	$(1-p+pe^t)^n$
<b>CF</b>	$(1-p+pe^{it})^n$
<b>PGF</b>	$G(z) = [(1-p)+pz]^n$
<b>Fisher information</b>	$g_n(p) = \frac{n}{p(1-p)}$ (for fixed $n$ )

# Probability distributions

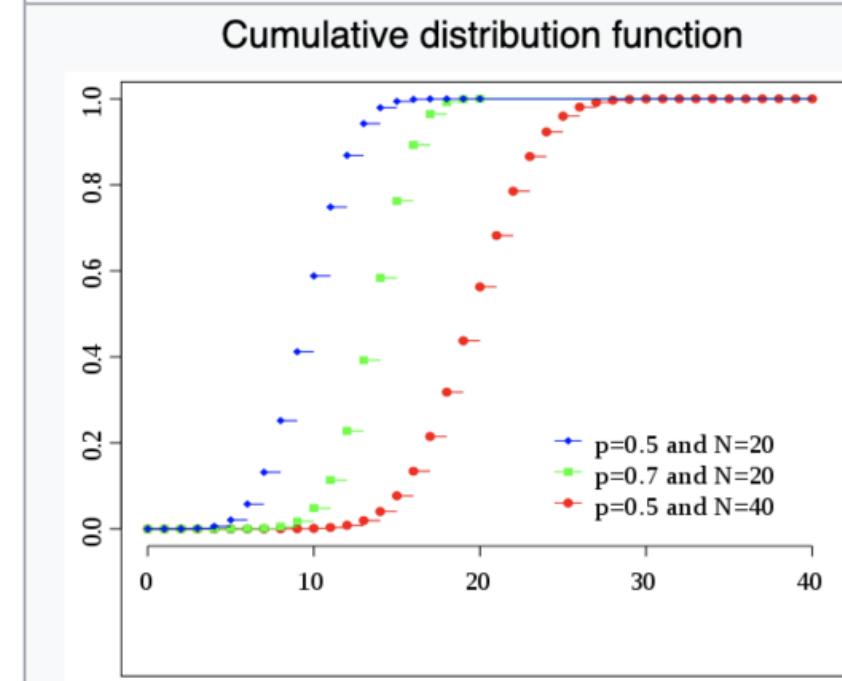
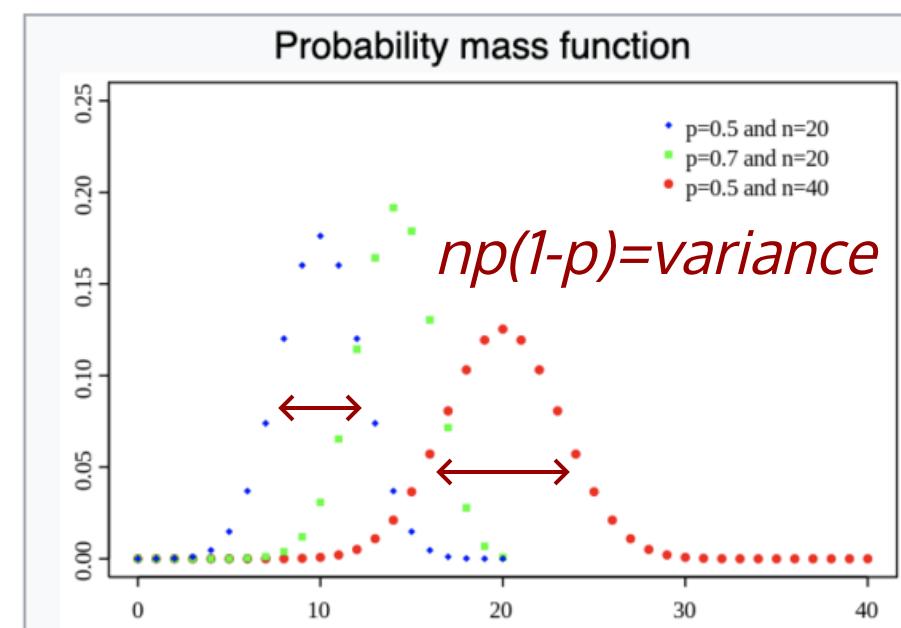
## Binomial

Coin toss:

fair coin:  $p=0.5$   $n=1$

Vegas coin:  $p \neq 0.5$   $n=1$

### Binomial distribution



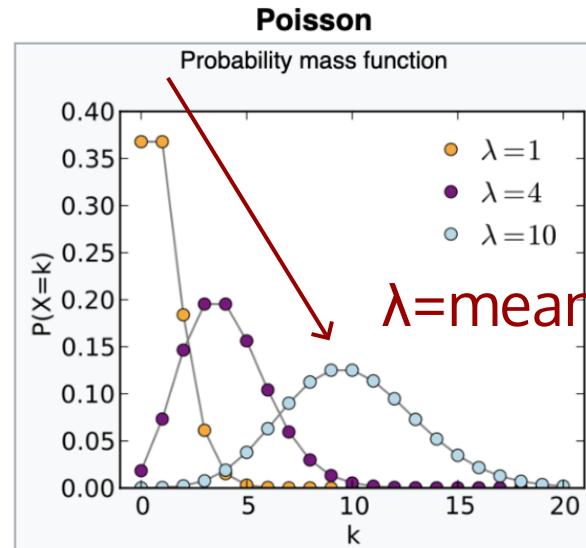
Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
Support	$k \in \{0, 1, \dots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k (1 - p)^{n-k}$
CDF	$I_{1-p}(n - k, 1 + k)$
Mean	$np$
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
Variance	$np(1 - p)$
Skewness	$\frac{1 - 2p}{\sqrt{np(1 - p)}}$
Ex. kurtosis	$\frac{1 - 6p(1 - p)}{np(1 - p)}$
Entropy	$\frac{1}{2} \log_2(2\pi enp(1 - p)) + O\left(\frac{1}{n}\right)$ in <a href="#">shannons</a> . For <a href="#">nats</a> , use the natural log in the log.
MGF	$(1 - p + pe^t)^n$
CF	$(1 - p + pe^{it})^n$
PGF	$G(z) = [(1 - p) + pz]^n$
Fisher information	$g_n(p) = \frac{n}{p(1 - p)}$ (for fixed $n$ )

# Probability distributions

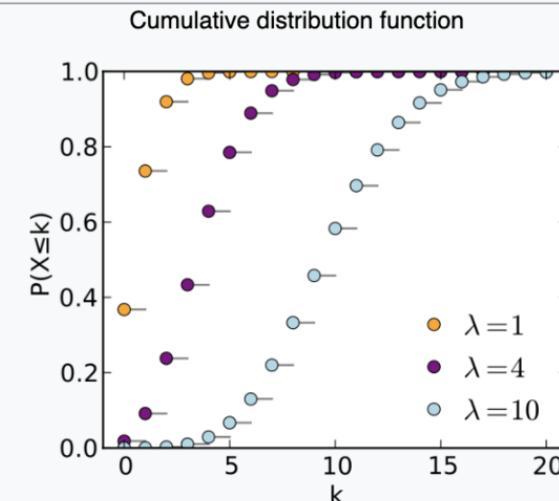
## Poisson

Shut noise/count noise

The innate noise in natural steady state processes (star flux, rain drops...)



The horizontal axis is the index  $k$ , the number of occurrences.  $\lambda$  is the expected number of occurrences, which need not be an integer. The vertical axis is the probability of  $k$  occurrences given  $\lambda$ . The function is defined only at integer values of  $k$ . The connecting lines are only guides for the eye.



The horizontal axis is the index  $k$ , the number of occurrences. The CDF is discontinuous at the integers of  $k$  and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

<b>Notation</b>	$\text{Pois}(\lambda)$
<b>Parameters</b>	$\lambda > 0$ , (real) — rate
<b>Support</b>	$k \in \{0, 1, 2, \dots\}$
<b>pmf</b>	$\frac{\lambda^k e^{-\lambda}}{k!}$
<b>CDF</b>	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor!}, \text{ or } e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}, \text{ or } Q(\lfloor k+1 \rfloor, \lambda) \text{ (for } k \geq 0 \text{, where } \Gamma(x, y) \text{ is the upper incomplete gamma function, } \lfloor k \rfloor \text{ is the floor function, and Q is the regularized gamma function)}$
<b>Mean</b>	$\lambda$
<b>Median</b>	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
<b>Mode</b>	$\lceil \lambda \rceil - 1, \lceil \lambda \rceil$
<b>Variance</b>	$\lambda$
<b>Skewness</b>	$\lambda^{-1/2}$
<b>Ex. kurtosis</b>	$\lambda^{-1}$
<b>Entropy</b>	$\lambda[1 - \log(\lambda)] + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \log(k!)}{k!}$ (for large $\lambda$ ) $\frac{1}{2} \log(2\pi e \lambda) - \frac{1}{12\lambda} - \frac{1}{24\lambda^2} - \frac{19}{360\lambda^3} + O\left(\frac{1}{\lambda^4}\right)$
<b>MGF</b>	$\exp(\lambda(e^t - 1))$
<b>CF</b>	$\exp(\lambda(e^{it} - 1))$
<b>PGF</b>	$\exp(\lambda(z - 1))$
<b>Fisher information</b>	$\frac{1}{\lambda}$

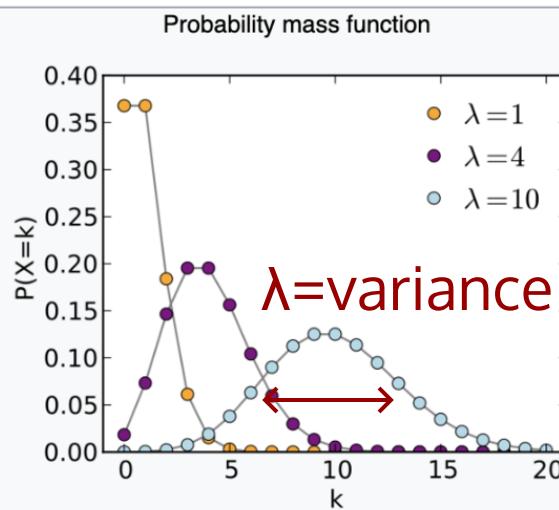
# Probability distributions

## Poisson

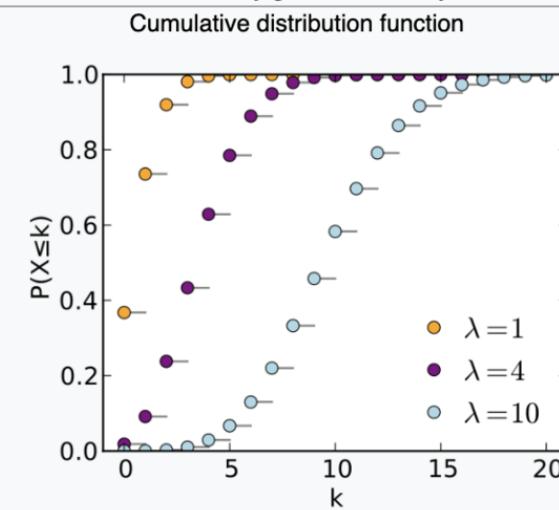
Shut noise/count noise

The innate noise in natural steady state processes (star flux, rain drops...)

Poisson	
<b>Notation</b>	$\text{Pois}(\lambda)$
<b>Parameters</b>	$\lambda > 0$ , (real) — rate
<b>Support</b>	$k \in \{0, 1, 2, \dots\}$
<b>pmf</b>	$\frac{\lambda^k e^{-\lambda}}{k!}$
<b>CDF</b>	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor!}, \text{ or } e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}, \text{ or } Q(\lfloor k+1 \rfloor, \lambda) \text{ (for } k \geq 0 \text{, where } \Gamma(x, y) \text{ is the upper incomplete gamma function, } \lfloor k \rfloor \text{ is the floor function, and } Q \text{ is the regularized gamma function)}$
<b>Mean</b>	$\lambda$
<b>Median</b>	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
<b>Mode</b>	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
<b>Variance</b>	$\lambda$
<b>Skewness</b>	$\lambda^{-1/2}$
<b>Ex. kurtosis</b>	$\lambda^{-1}$
<b>Entropy</b>	$\lambda[1 - \log(\lambda)] + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \log(k!)}{k!}$ (for large $\lambda$ ) $\frac{1}{2} \log(2\pi e \lambda) - \frac{1}{12\lambda} - \frac{1}{24\lambda^2} - \frac{19}{360\lambda^3} + O\left(\frac{1}{\lambda^4}\right)$
<b>MGF</b>	$\exp(\lambda(e^t - 1))$
<b>CF</b>	$\exp(\lambda(e^{it} - 1))$
<b>PGF</b>	$\exp(\lambda(z - 1))$
<b>Fisher information</b>	$\frac{1}{\lambda}$



The horizontal axis is the index  $k$ , the number of occurrences.  $\lambda$  is the expected number of occurrences, which need not be an integer. The vertical axis is the probability of  $k$  occurrences given  $\lambda$ . The function is defined only at integer values of  $k$ . The connecting lines are only guides for the eye.



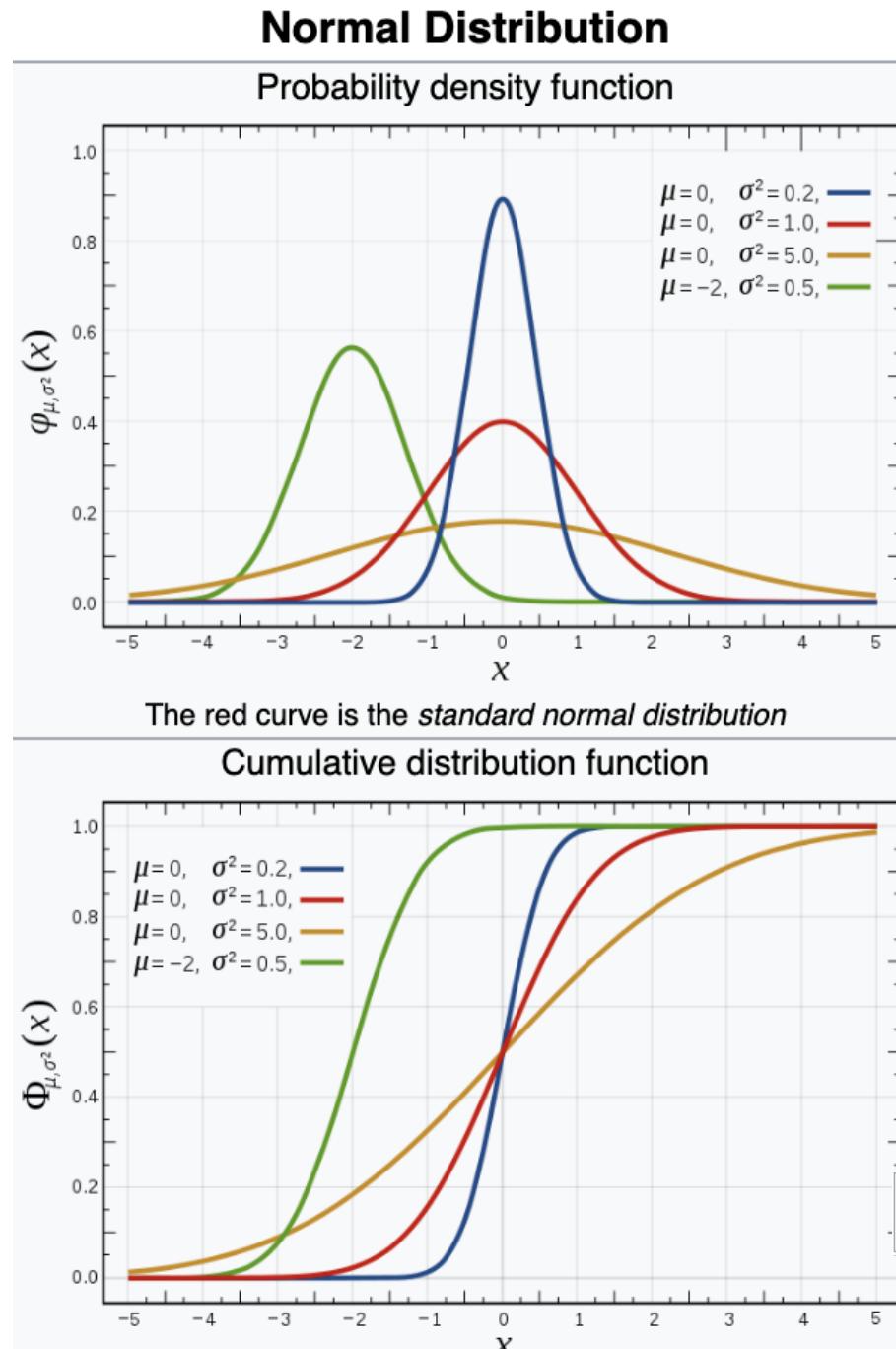
The horizontal axis is the index  $k$ , the number of occurrences. The CDF is discontinuous at the integers of  $k$  and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

# Probability distributions

## Gaussian

most common noise:

well behaved mathematically,  
symmetric, when we can we will  
assume our uncertainties are  
Gaussian distributed

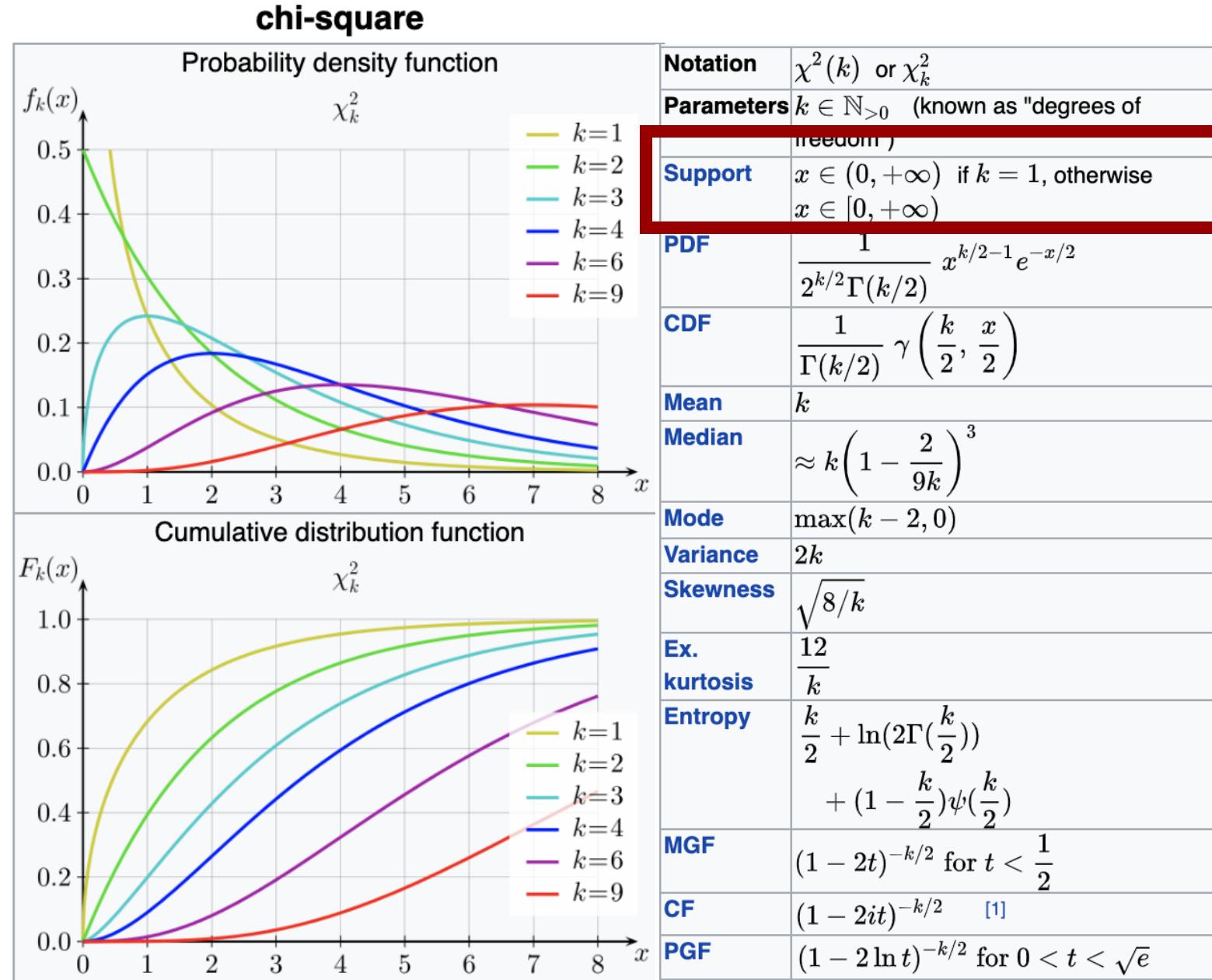


Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared standard deviation)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2} \left[ 1 + \operatorname{erf}\left( \frac{x-\mu}{\sigma\sqrt{2}} \right) \right]$
Quantile	$\mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2F - 1)$
Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Variance	$\sigma^2$
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2} \log(2\pi e \sigma^2)$
MGF	$\exp(\mu t + \sigma^2 t^2 / 2)$
CF	$\exp(i\mu t - \sigma^2 t^2 / 2)$
Fisher information	$\mathcal{I}(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$
Kullback-Leibler divergence	$D_{\text{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2} \left\{ (\sigma_0/\sigma_1)^2 + \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2} - 1 + 2 \ln \frac{\sigma_1}{\sigma_0} \right\}$

# Probability distributions

## Chi-square ( $\chi^2$ )

turns out its extremely common  
many pivotal quantities follow this  
distribution and thus many tests are  
based on this



# coding time!



<https://colab.research.google.com/>

<https://github.com/fedhere/DSPS/blob/master/statistics/distributionParametersDemo.ipynb>

# Central Limit Theorem

Laplace (1700s) but also: Poisson, Bessel, Dirichlet, Cauchy, Ellis

Let  $x_1 \dots x_N$  be an  $N$ -elements sample from a population  
whose distribution has

mean  $\mu$  and standard deviation  $\sigma$

In the limit of  $N \rightarrow \infty$

the sample mean  $\bar{x}$  approaches a Normal (Gaussian)  
distribution with mean  $\mu$  and standard deviation  $\sigma$   
regardless of the distribution of  $X$

$$\bar{x} \sim N\left(\mu, \sigma/\sqrt{N}\right)$$

interpretation of  
probability

distributions

central limit theorem

null hypothesis rejection  
testing setup

# key concepts

# homework

- HW1 : explore the Maxwell Boltzmann distribution
- HW2: graphic demonstration of the Central Limit Theorem

Jacob Cohen, 1994

## The earth is round ( $p=0.05$ )

[http://fbb.space/dsps/Cohen1994\\_TheEarthIsRound\\_AmPsych.pdf](http://fbb.space/dsps/Cohen1994_TheEarthIsRound_AmPsych.pdf)

read

the original link:

<http://psycnet.apa.org/fulltext/1995-12080-001.html>

(this link needs access to science magazine, but you can use the link above which is the same file)

## The Earth Is Round ( $p < .05$ )

Jacob Cohen

*After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of  $p$  as the probability that  $H_0$  is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects  $H_0$  one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.*

# Foundations of Statistical Mechanics 1845—1915

Stephen G. Brush

Archive for History of Exact Sciences Vol. 4, No. 3 (4.10.1967), pp. 145-183

University of Delaware Library  
DELCAT Discovery



Course Reserves

Chat with a librarian

My Items (0)

Search University of Delaware Library and beyond.

additional reading

Foundations of Statistical Mechanics 1845—1915



[Advanced Search](#)

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008

**Statistics in a Nutshell (Chapters 3,4,5)**

[https://books.google.com/books/about/Statistics\\_in\\_a\\_Nutshell.html?id=ZnhgO65Pyl4C](https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C)

David M. Lane et al.

**Introduction to Statistics (XVIII)**

[http://onlinestatbook.com/Online\\_Statistics\\_Education.epub](http://onlinestatbook.com/Online_Statistics_Education.epub)

<http://onlinestatbook.com/2/index.html>

resources