ML for natural and physical scientists 2023 2

II probability and statistics

dr.federica bianco | *fbb.space* |  *fedhere* |  *fedhere*

this slide deck

https://slides.com/federicabianco/mlpns23_2

Week 1: Probability and statistics (stats for hackers)
Week 2: linear regression - uncertainties
Week 3: unsupervised learning - clustering
Week 4: kNN | CART (trees)
Week 5: Neural Networks - basics
Week 6: CNNs
Week 7: Autoencoders
Week 8: Physically motivated NN | Transformers

Tuesday: "theory"
Thursday: "hands on work"
Friday: "recap and preview"

Somewhere I will also cover:
notes on visualizations
notes on data ethics

slido.com
#2492 113

github

reproducibility

https://github.com

Reproducible research means:

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

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

allows reproducibility through code distribution

github

version control

https://github.com

the Git software

is a distributed *version control system:*

a version of the files on your local computer is made also available at a central server.

The history of the files is saved remotely so that any version (that was checked in) is retrievable.

allows version control

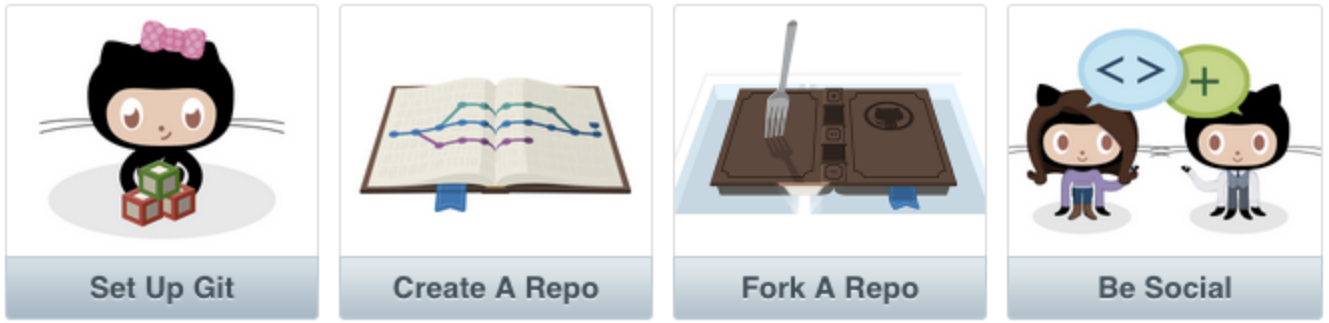
github

collaborative platform

https://github.com

collaboration tool

by fork, fork and pull request, or by working directly as a collaborator



allows effective collaboration

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 an 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
- seed your code random variables

Assignments rules

your data should be accessible e.g. reading in a URL; all data you use should be public

-> use public data with live links

-> put your data on github if not accessible

```
import pandas as pd
videos = pd.read_csv("https://github.com/fedhere/MLPNS_FBianco/blob/main/data/USyoutubes.csv?raw=true")
```

your notebook should work and reproduce exactly the results you reported in your notebook version

-> *restart kernel + run all*

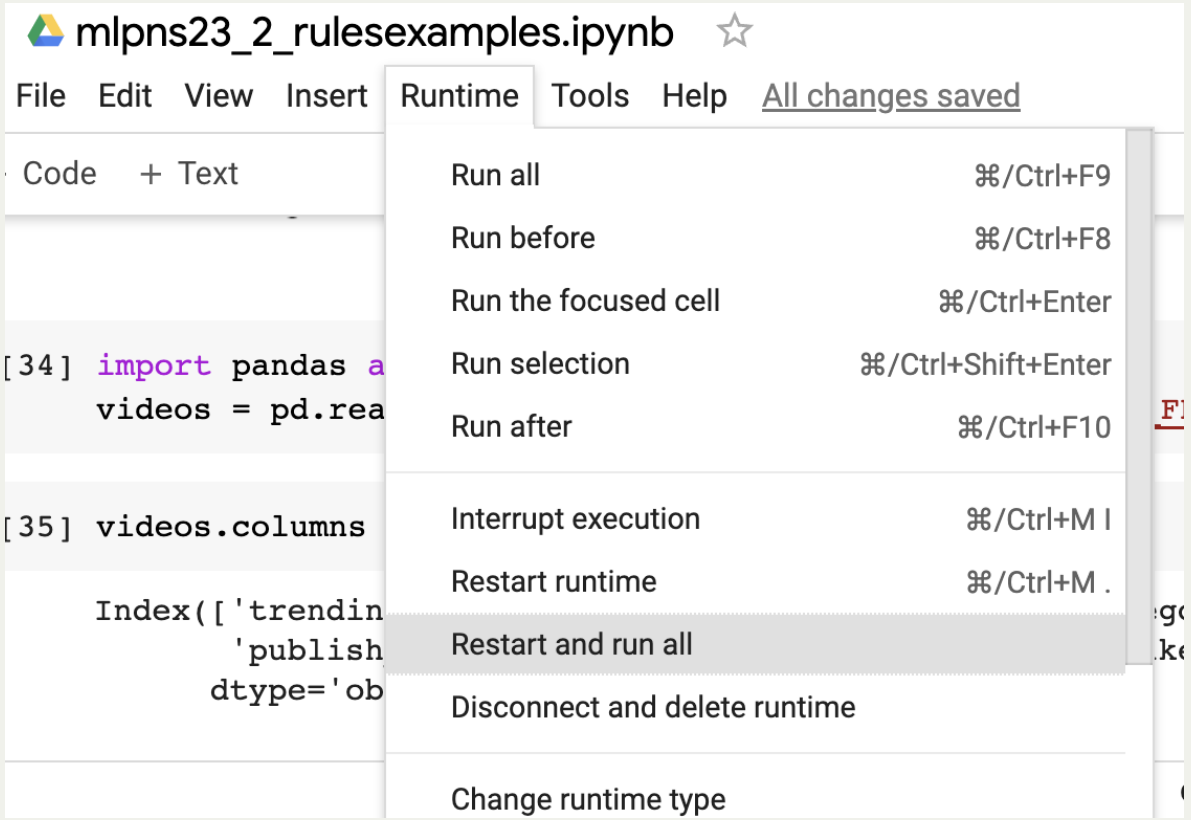
->seed your code random variables

Assignments rules

your data should be accessible e.g. reading in a URL; all data you use should be public

-> use public data with live links

-> put your data on github if not accessible



your notebook should work and reproduce exactly the results you reported in your notebook version

-> *restart kernel + run all*

->seed your code random variables

Assignments rules

your data should be accessible e.g. reading in a URL; all data you use should be public

-> use public data with live links

-> put your data on github if not accessible

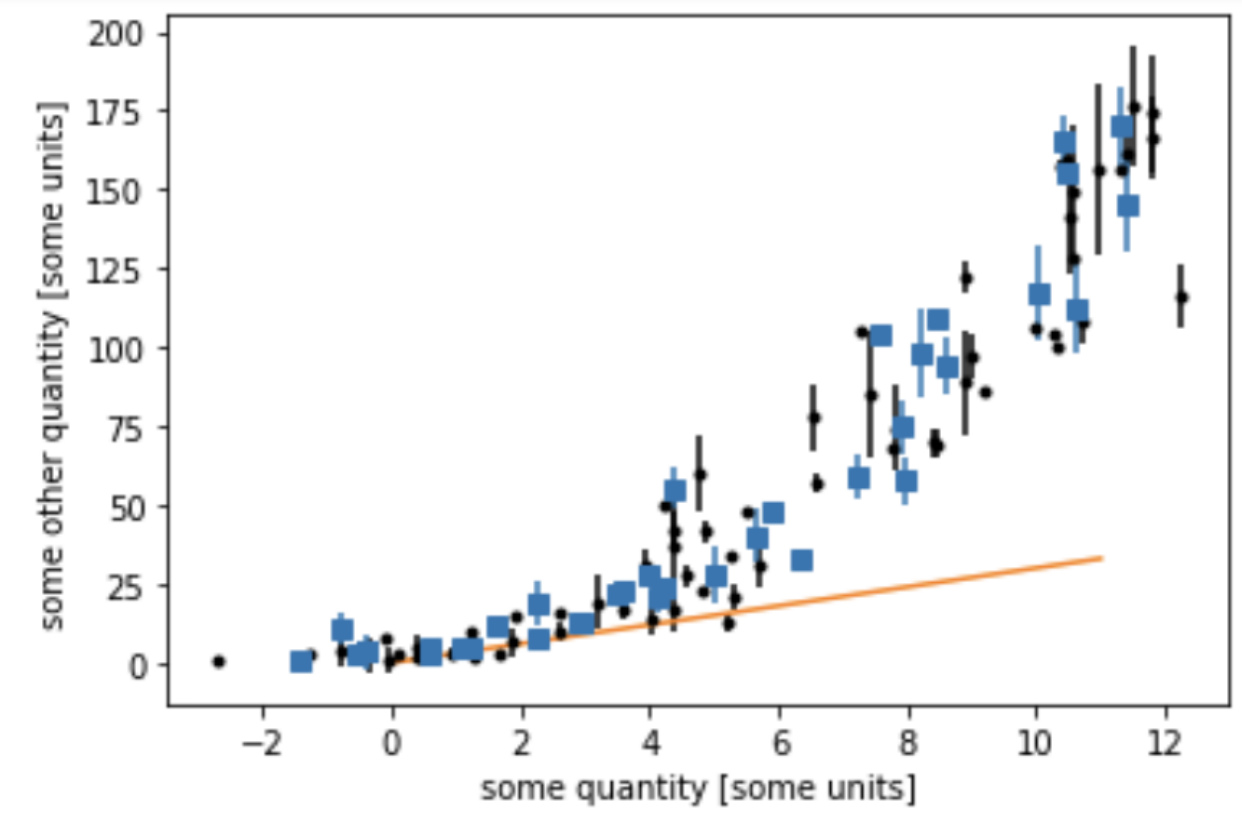
your notebook should work and reproduce exactly the results you reported in your notebook version

-> *restart kernel + run all*

->seed your code random variables

```
np.random.seed(125) # seeding random numbers i get the same sequence all the time
x = np.random.randint(0,12,100)
y = 2.0 + 0.7 * x + 1.2 * x**2
x = x + np.random.randn(x.shape[0])
scatter = np.random.randn(y.shape[0]) * np.sqrt(y)
yerr = np.random.randn(y.shape[0]) * 2 + scatter
index1 = np.random.choice(np.arange(x.shape[0]), 63, replace=False)
index2 = np.array(list(set(np.arange(x.shape[0])) - set(index1)))
index1
```

Assignments rules



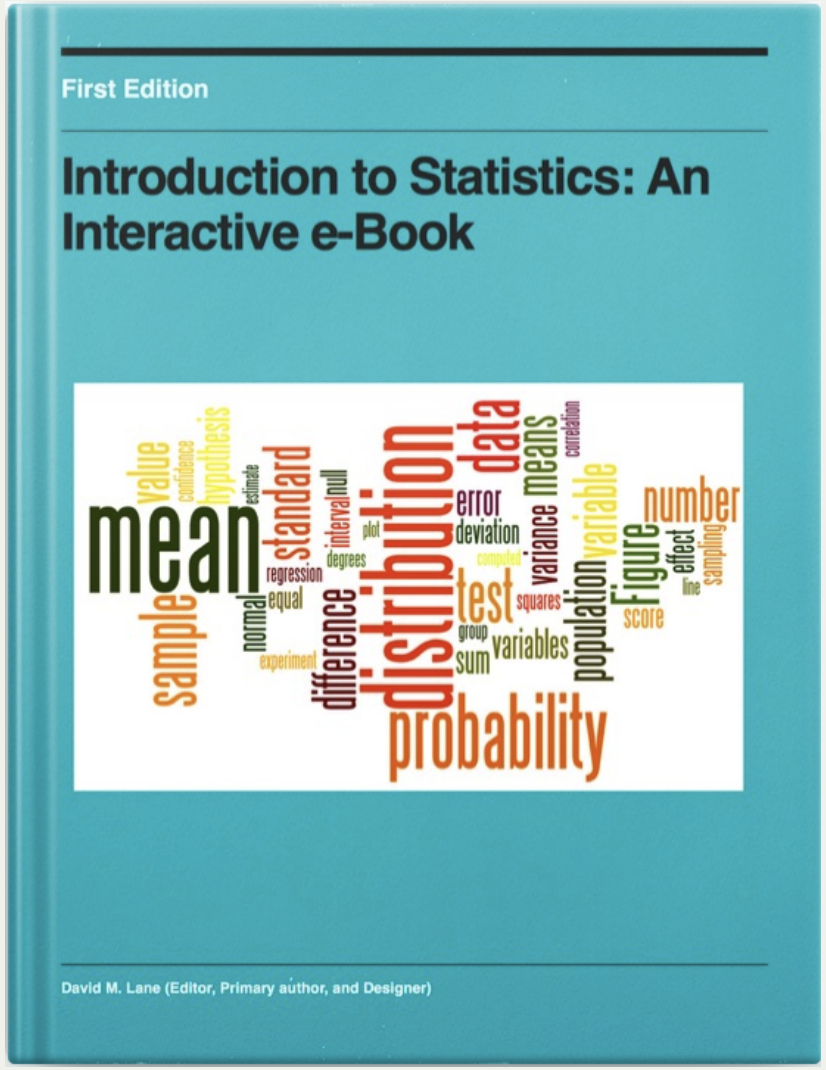
Other rules:

- Every figure should have axed labels
- Every figure should have a caption that describes :
  - what is being plotted (*this plot shows … against … black dots are … blue squares are… the line represents a linear model to…*)
  - why is interesting in the flow of your analysis (*e.g. the model (orange line) is a good fit to the data only up to about x=4, at which point the data is systematically underpredicted by the model*)
- The numbers that you print out should be explained. For example they should be embedded in a print statement (*e.g.*)

```
print("the average number of video views is {:.1f}".format(videos["views"].mean()))
```

```
the average number of video views is 2360784.6
```

what are probability and statistics?



Introduction to Statistics: An Interactive e-Book

David M. Lane

Crush Course in Statistics

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

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(coin = head) = 6/11 = 0.55

fraction of times something happens

⇕

probability of it happening

Basic Probability

Frequentist interpretation

P(E) = frequency of E

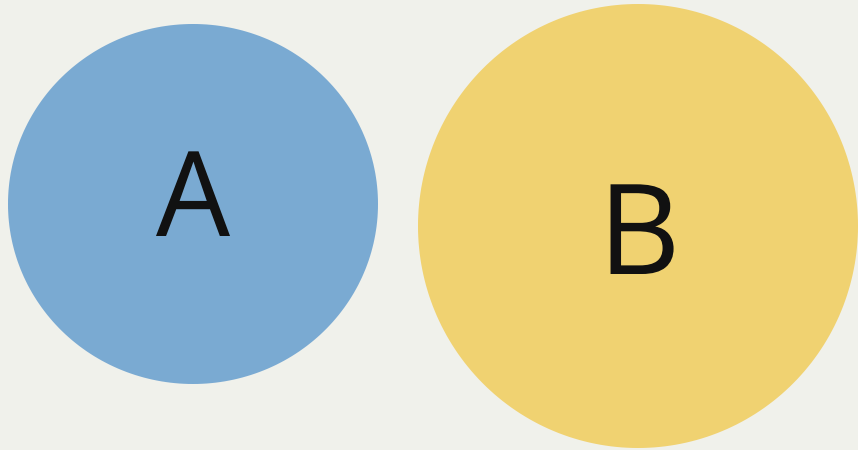
P(coin = head) = 6/11 = 0.55

P(coin = head) = 49/100 = 0.49

fraction of times something happens

probability of it happening

Basic probability arithmetics



Probability Arithmetic $0 <= P(A) <= 1$

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

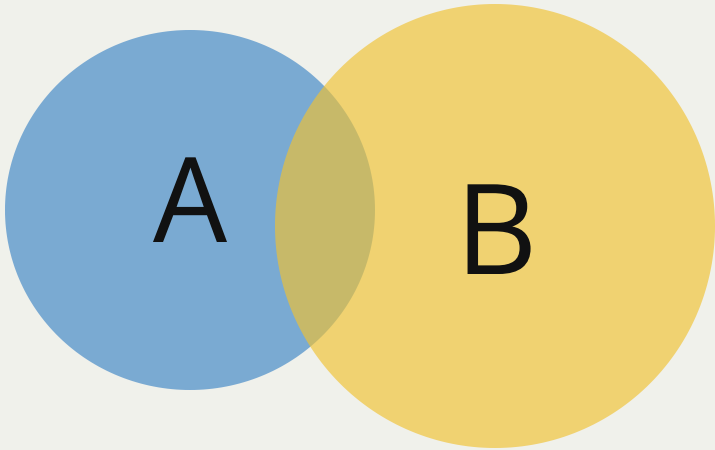
disjoint events $if\, P(A) \cap P(B) = 0 \text{ then :}$

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

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

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

Basic probability arithmetics



Probability Arithmetic

$$\text{in general :}$$

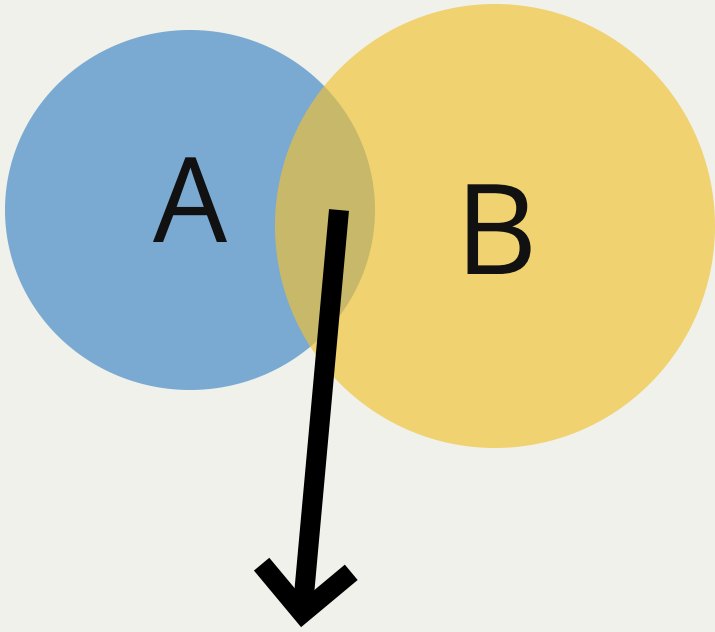
dependent events

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

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

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

Basic probability arithmetics



$P(A \cap B)$

Probability Arithmetic

$$\text{in general :}$$

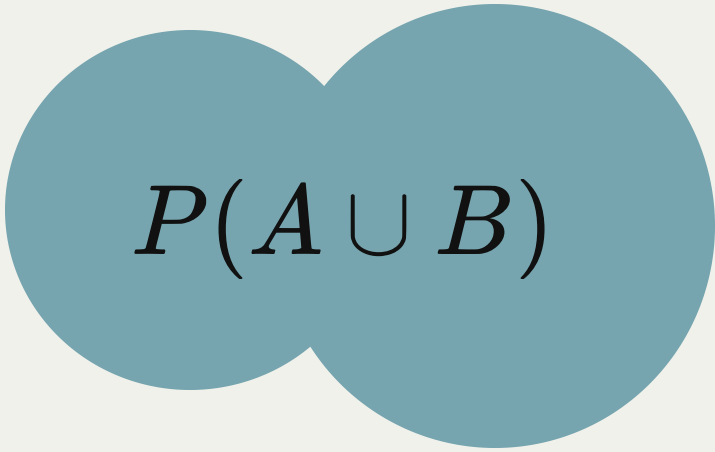
dependent events

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

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

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

Basic probability arithmetics



Probability Arithmetic

$$\text{in general :}$$

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 elemnts of a "family"

Sample: a finite subset of the population that you observe

The application of stastistics to physics

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

https://upload.wikimedia.org/wikipedia/commons/8/82/Simulation_of_gas_for_relaxation_demonstration.gif?1567607773826

https://commons.wikimedia.org/wiki/File:Simulation_of_gas_for_relaxation_demonstration.gif

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://commons.wikimedia.org/wiki/File:Simulation_of_gas_for_relaxation_demonstration.gif

Phsyics Example

Boltzmann 1872

Entropy **2015**, *17*, 1971-2009; doi:10.3390/e17041971

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76:373-435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909

Kim Sharp * and Franz Matschinsky

https://www.mdpi.com/1099-4300/17/4/1971/pdf

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,

Phsyics Example

Boltzmann 1872

Entropy **2015**, *17*, 1971-2009; doi:10.3390/e17041971

OPEN ACCESS

entropy

ISSN 1099-4300

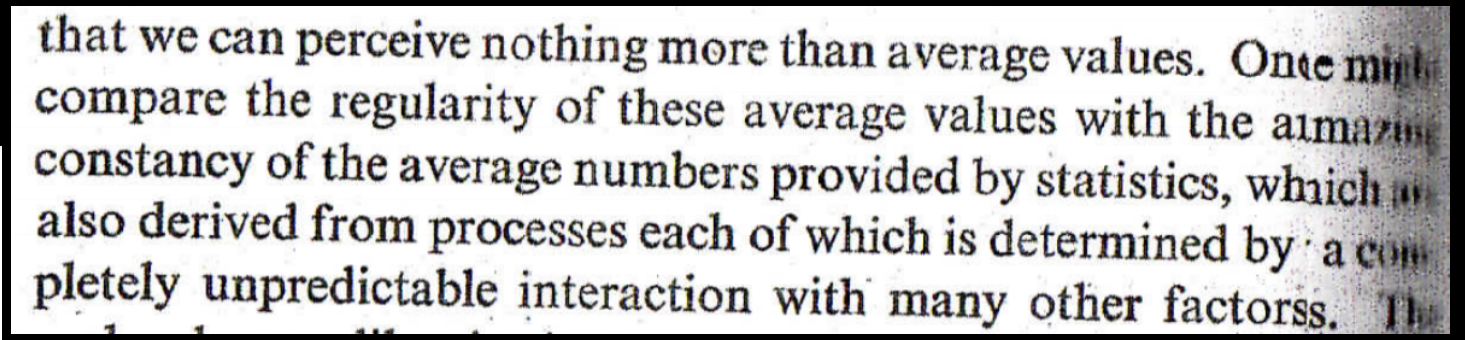
www.mdpi.com/journal/entropy

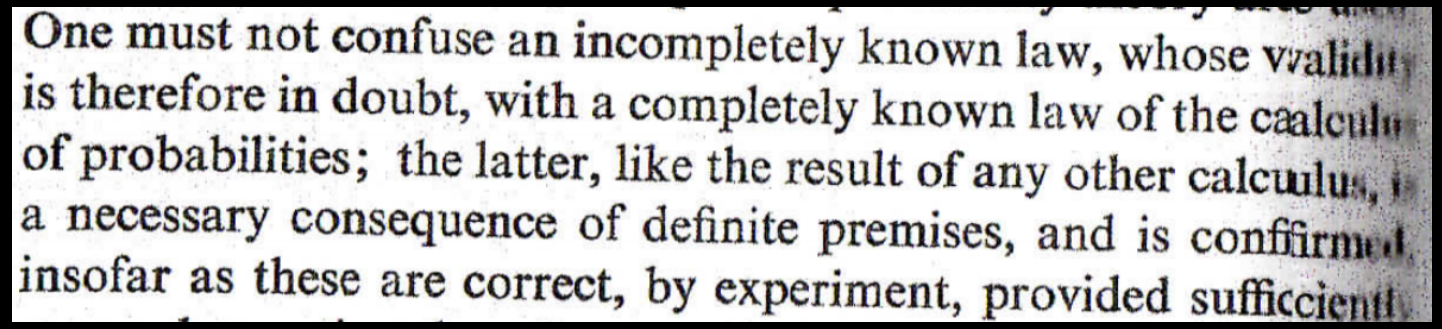
Article

Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76:373-435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909

Kim Sharp * and Franz Matschinsky

https://www.mdpi.com/1099-4300/17/4/1971/pdf


that we can perceive nothing more than average values. One mi
compare the regularity of these average values with the aima
constancy of the average numbers provided by statistics, which
also derived from processes each of which is determined by a co
pletely unpredictable interaction with many other factorss. Th


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

Phsyics Example

https://hal.archives-ouvertes.fr/hal-01662284/document

How statistics entered physics?

Olivier REY[1]

ABSTRACT: Now that statistics is a branch of mathematics, it is easy to imagine that its use in the field of human affairs is a by-product of modern science's way of looking at the world. Historical study contradicts such an idea: it is in the field of human affairs that quantitative statistics have developed, and it is only afterwards that it became a method for the natural sciences. Most physicists in the 19th century considered statistics all too human to have a place in the scientific study of nature. It took all Maxwell's authority and persuasion to make statistical analysis a new style of scientific thought in physics.

Such a point is of fundamental importance. Indeed, because of their initial anchoring in the human and social sphere, statistics suffered a long time from a bias among scientists and, in particular, among physicists. Making them acknowledge that statistics could be, and should be used in physics, was not a small undertaking, and that's this story I would like to sketch.

descriptive statistics:

we summarize the proprties of a distribution

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

3 descriptive statistics:

```
# myarray is a numpy array
myarray.mean() # the mean of all numbers in the array
myarray.mean(axis=1) # the mean along the second axis (one number per array row)
```

descriptive statistics:

we summarize the proprties of a distribution

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x)\,\mathrm{d}x.$$

mean: n=1 $$\mu = \frac{1}{N}\sum_1^N x_i$$

other measures of centeral tendency:

median: 50% of the distribution is to the left,

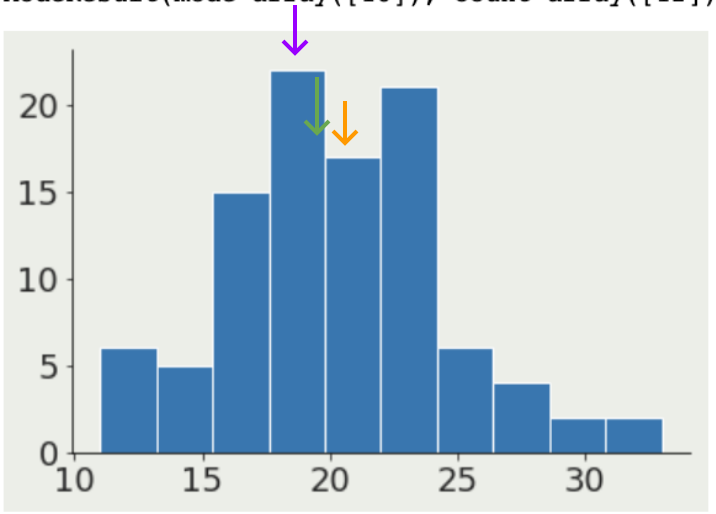
50% to the right

mode: most popular value in the distribution

```
dist = sp.stats.poisson.rvs(size=100, mu=20)
pl.hist(dist)
print(dist.mean())
print(np.median(dist))
print(sp.stats.mode(dist))
```

executed in 125ms, finished 15:01:20 2019-09-09

```
20.06
20.0
ModeResult(mode=array([18]), count=array([12]))
```



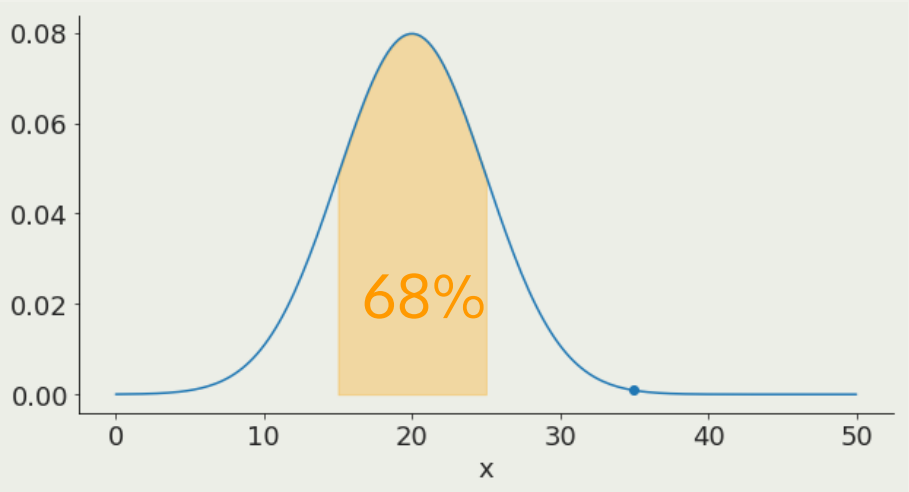
descriptive statistics:

we summarize the proprties of a distribution

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x)\, \mathrm{d}x.$$

variance: n=2 $$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

standard deviation $$\sigma(X) = \mathrm{E}\left[(X - \mu)\right].$$



Gaussian distribution:

1σ contains 68% of the distribution

```
# myarray is a numpy array
myarray.std() # the standard dev of all numbers in the array
myarray.std(axis=1) # the standard dev  along the second axis
```

descriptive statistics:

we summarize the proprties of a distribution

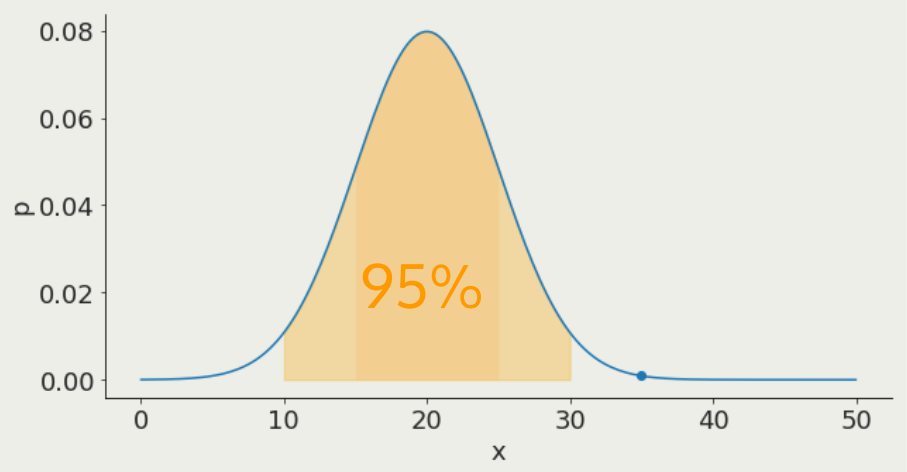
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

variance: n=2 $$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

standard deviation $$\sigma(X) = \mathrm{E}\left[(X - \mu)\right].$$

Gaussian distribution:

2σ contains 95% of the distribution



descriptive statistics:

we summarize the proprties of a distribution

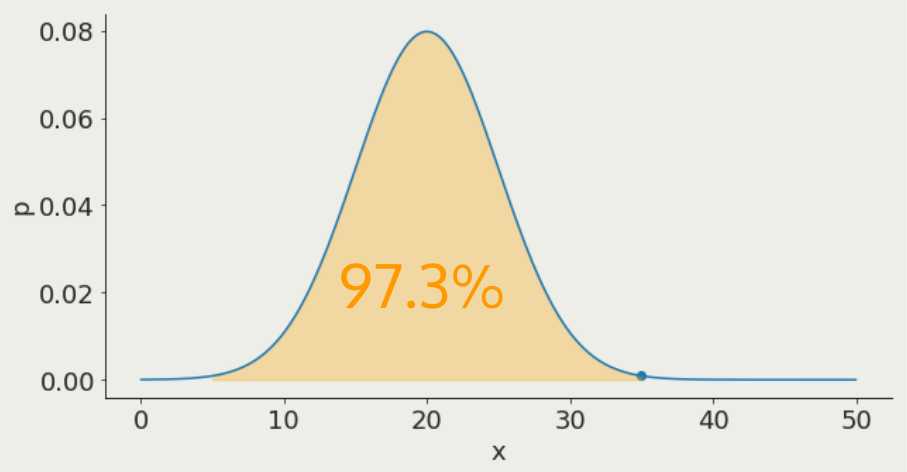
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

variance: n=2 $$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

standard deviation $$\sigma(X) = \mathrm{E}\left[(X - \mu)\right].$$

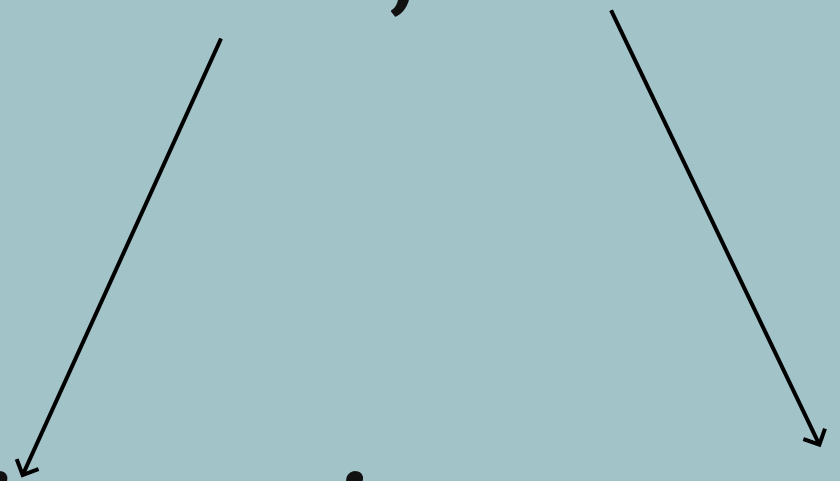
Gaussian distribution:

3σ contains 97.3% of the distribution



TAXONOMY

central tendency: mean, median, mode



spread : variance, interquantile range

Probability distributions

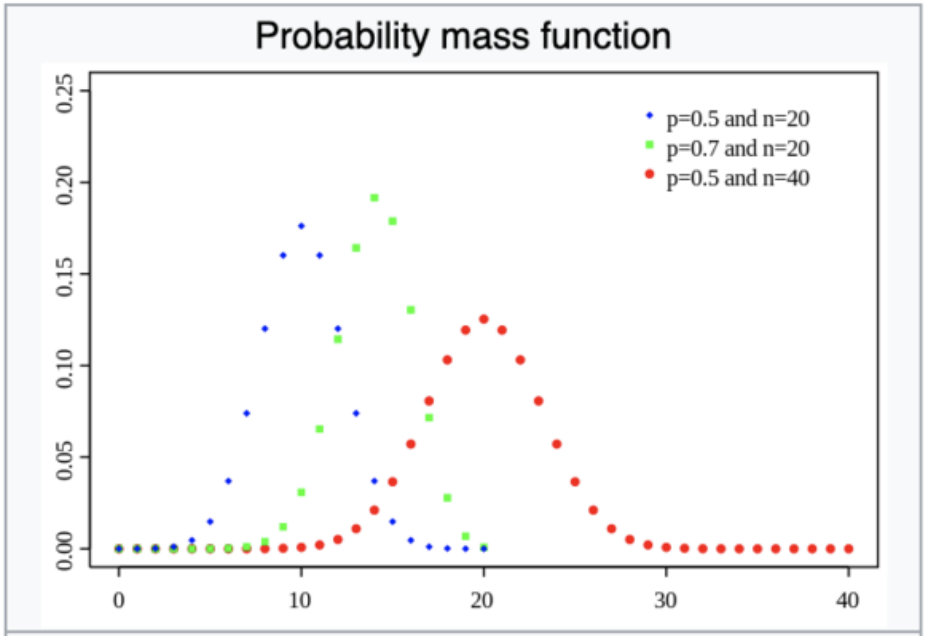
Binomial

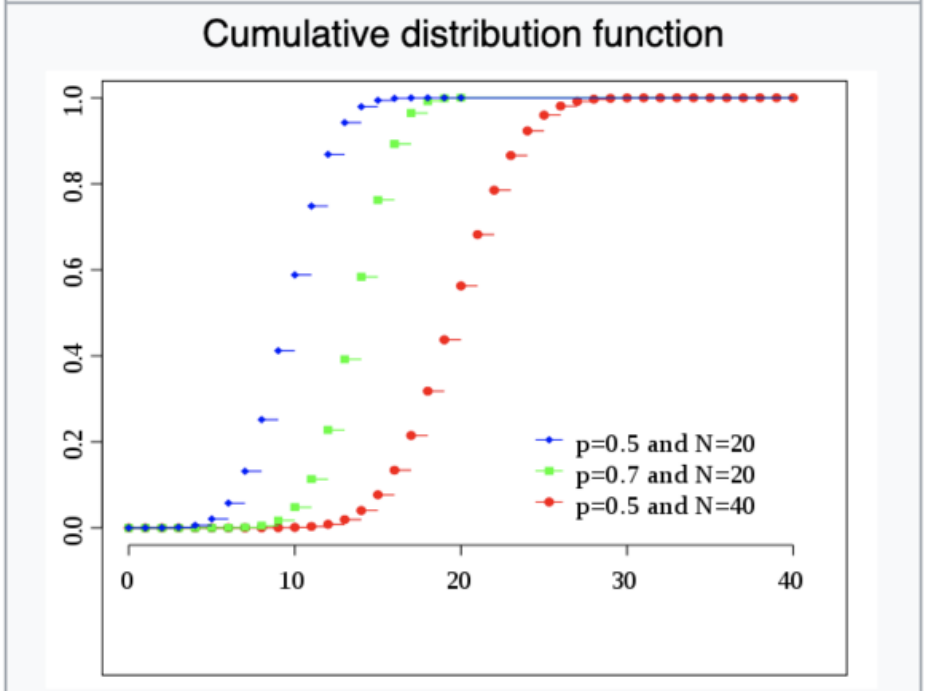
Coin toss:

fair coin: p=0.5 n=1

Vegas coin: p≠0.5 n=1

Binomial distribution





Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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

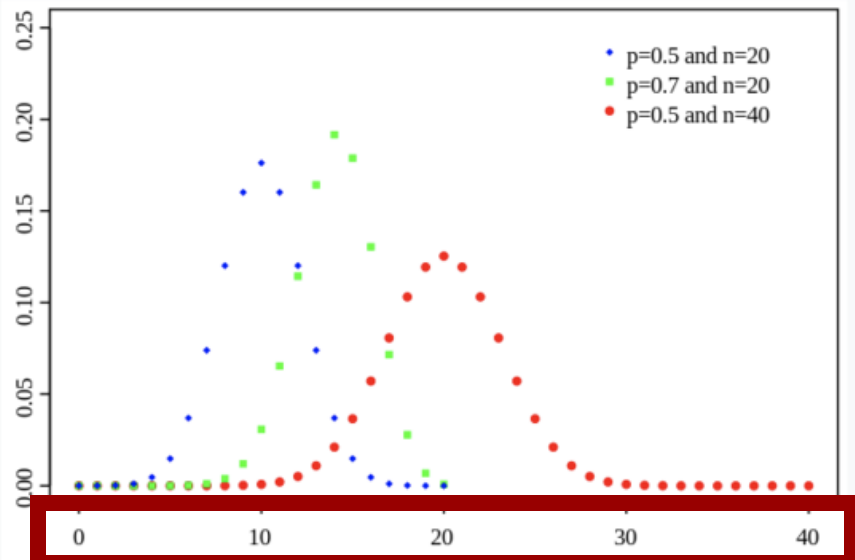
Coin toss:

fair coin: p=0.5 n=1

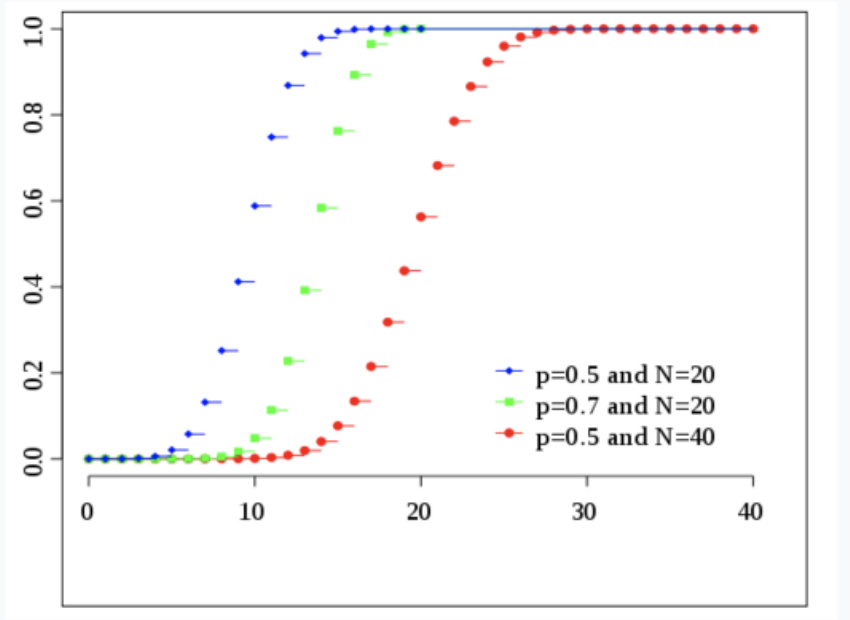
Vegas coin: p≠0.5 n=1

Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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

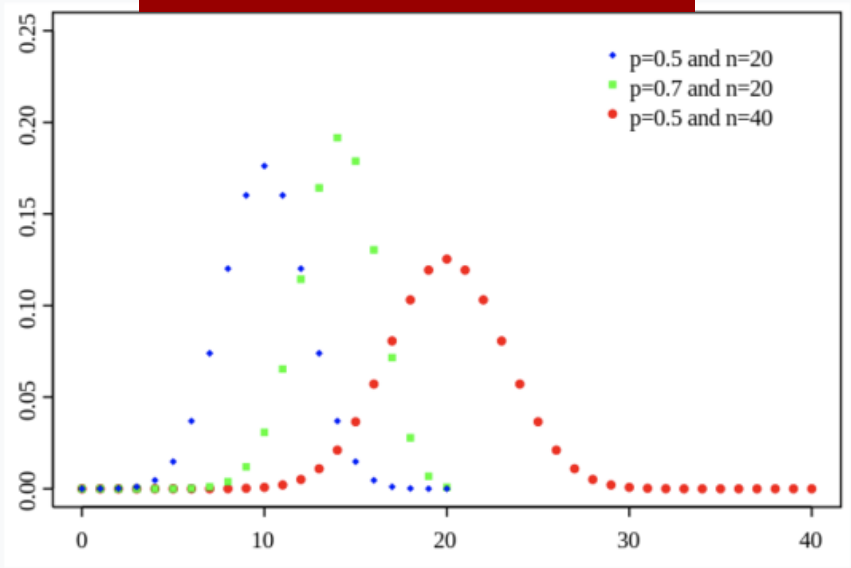
Coin toss:

fair coin: p=0.5 n=1

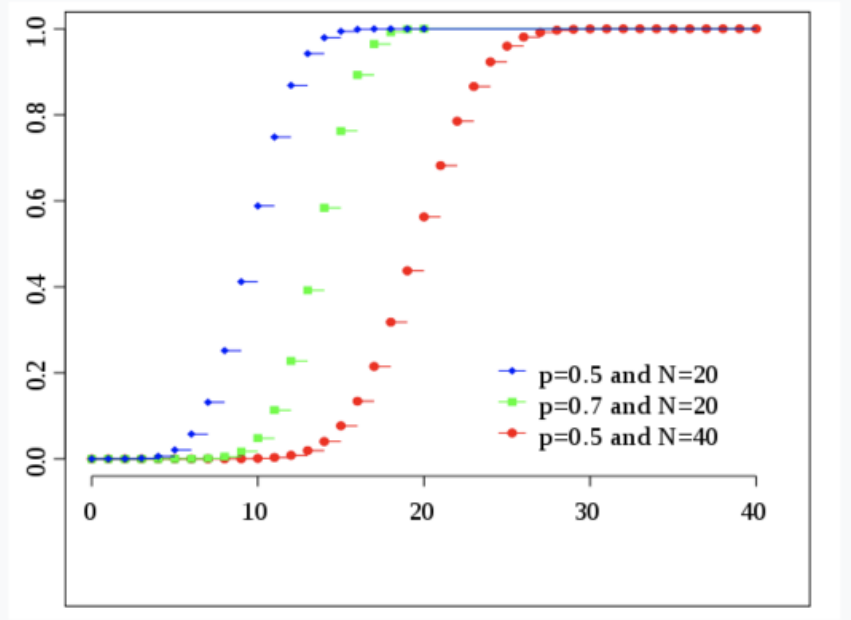
Vegas coin: p≠0.5 n=1

Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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

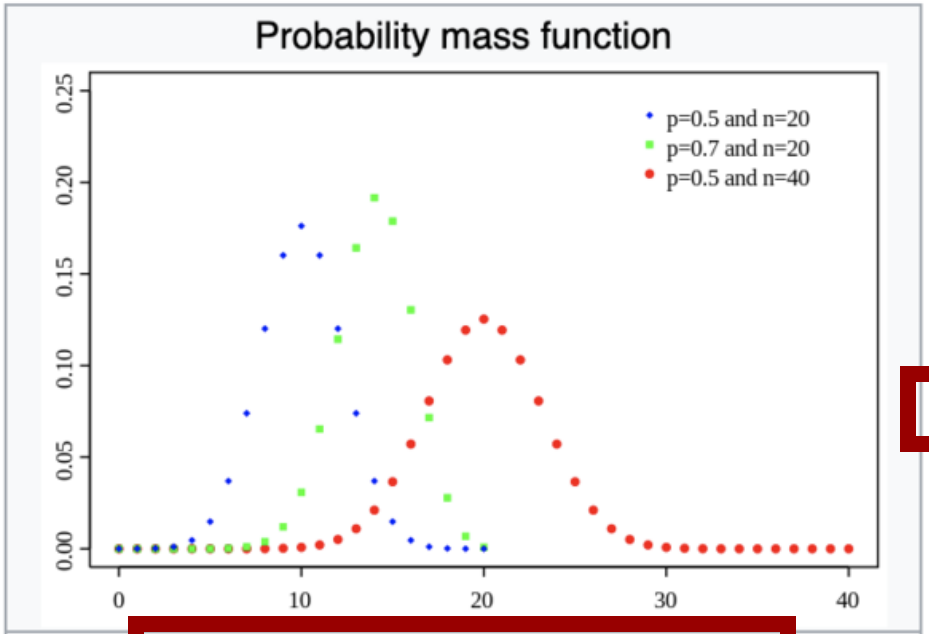
Binomial

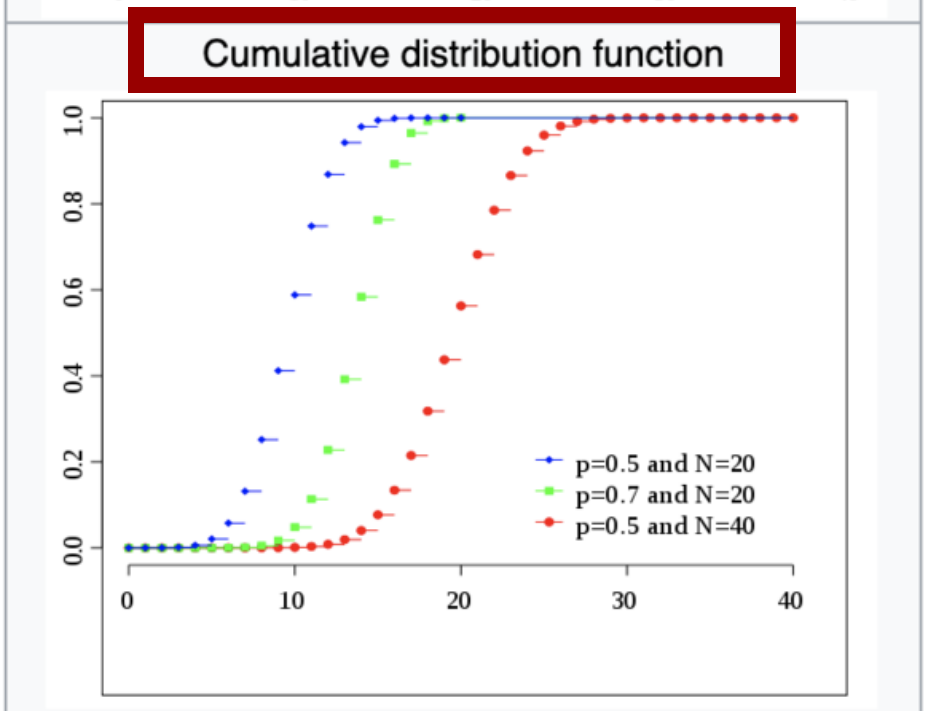
Coin toss:

fair coin: p=0.5 n=1

Vegas coin: p≠0.5 n=1

Binomial distribution





Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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

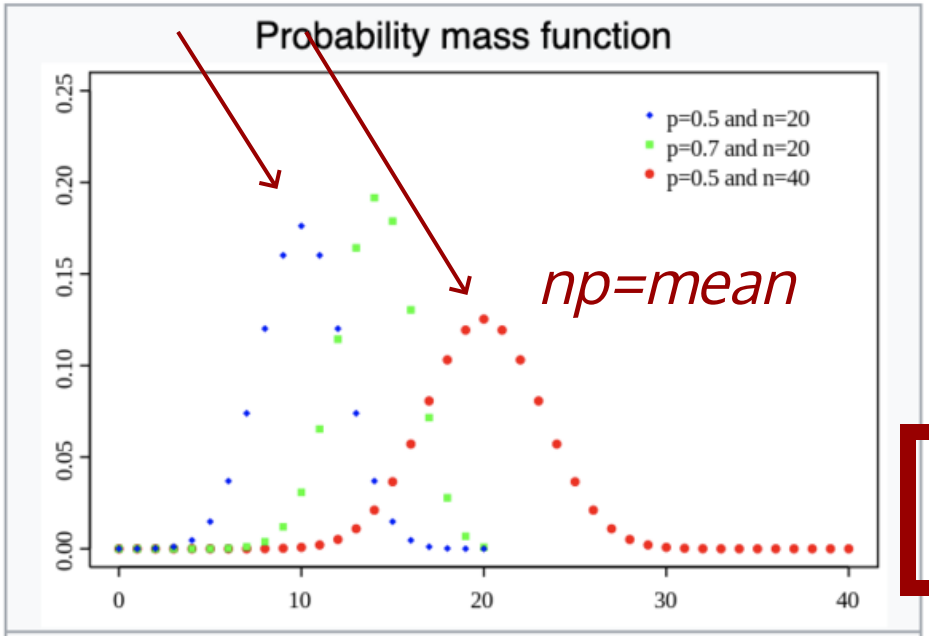
Binomial

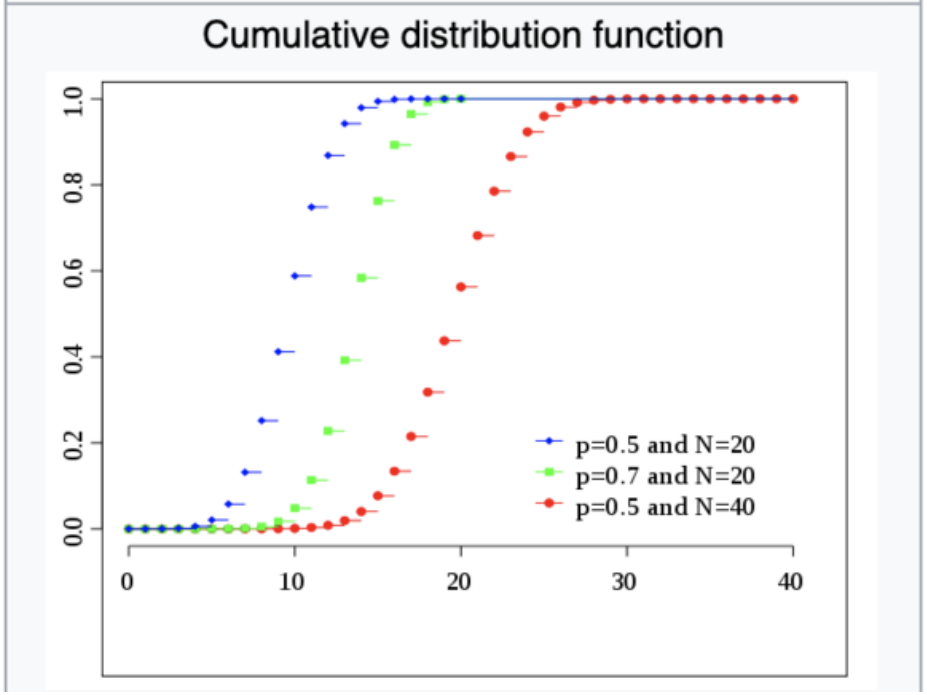
Coin toss:

fair coin: p=0.5 n=1

Vegas coin: p≠0.5 n=1

Binomial distribution





central tendency

Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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

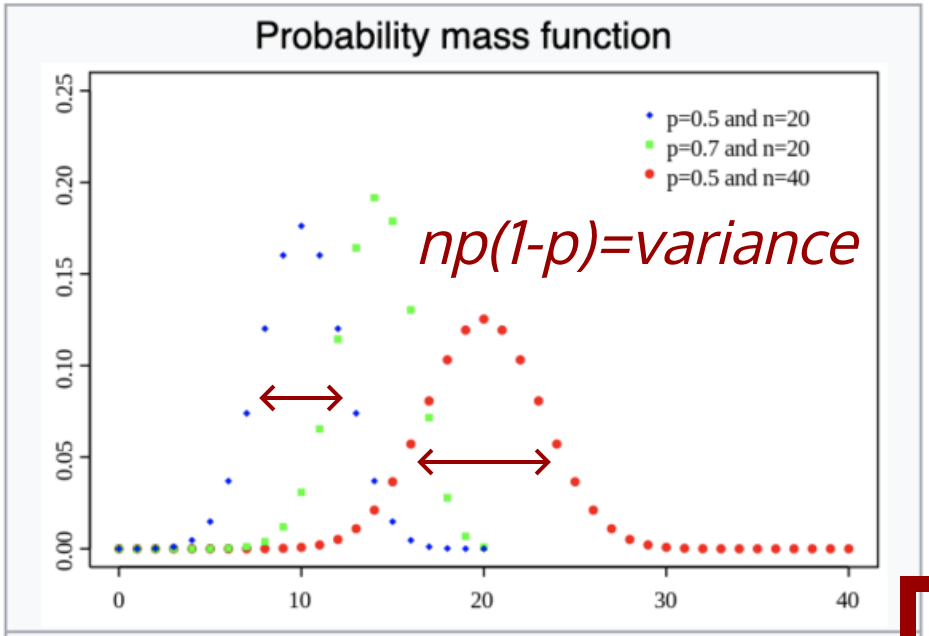
Binomial

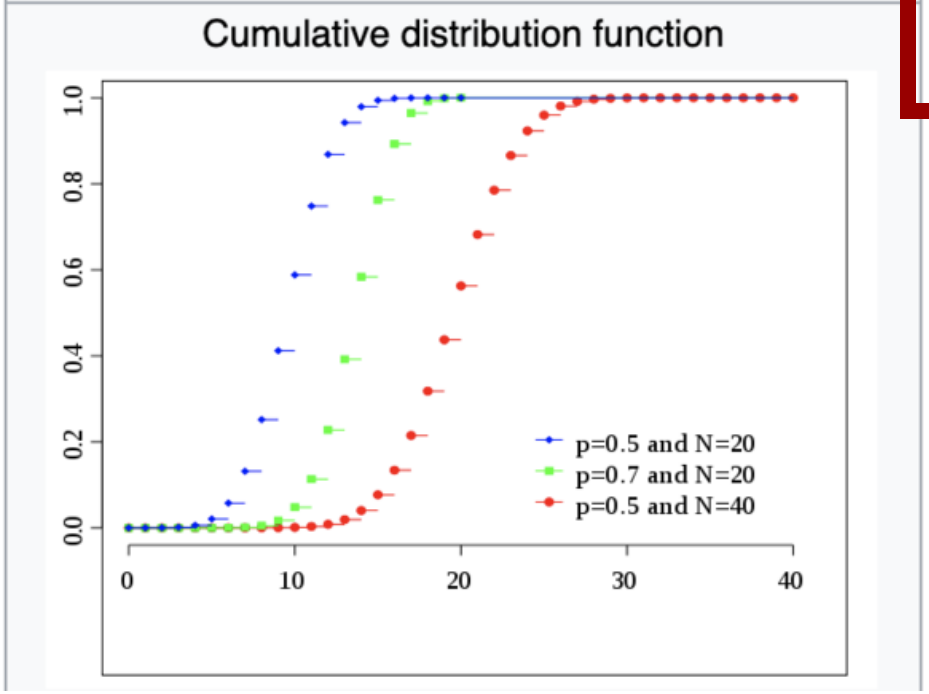
Coin toss:

fair coin: p=0.5 n=1

Vegas coin: p≠0.5 n=1

Binomial distribution





Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial
Support	$k \in \{0,1,\ldots,n\}$ – 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 shannons. For nats, 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)

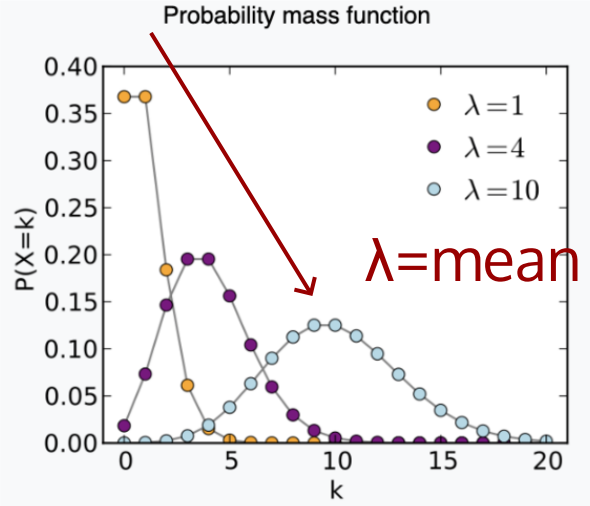
Probability distributions

Poisson

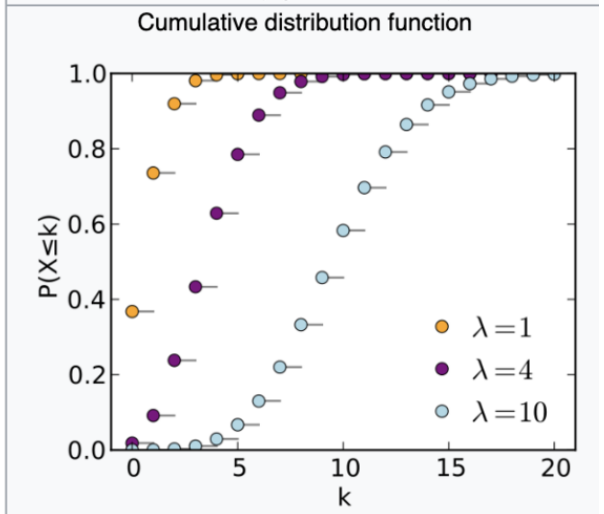
Shut noise/count noise

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

Poisson



The horizontal axis is the index *k*, the number of occurrences. *λ* is the expected number of occurrences, which need not be an integer. The vertical axis is the probability of *k* occurrences given *λ*. 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.

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

Probability distributions

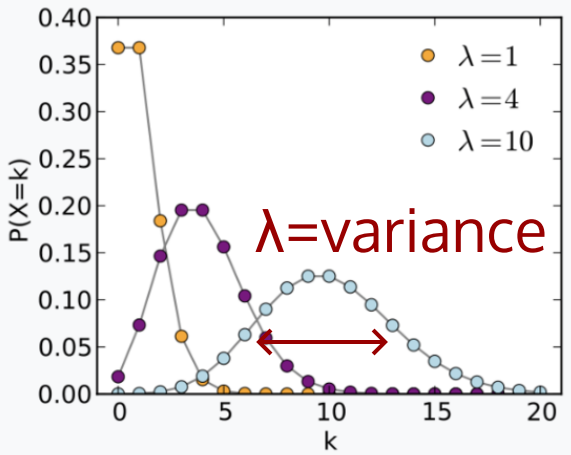
Poisson

Shut noise/count noise

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

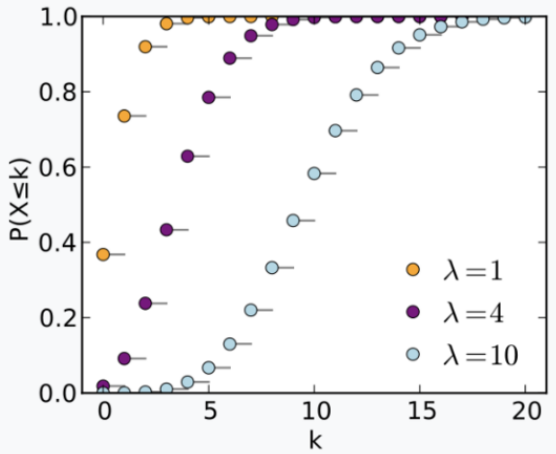
Poisson

Probability mass function



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

Cumulative distribution function



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.

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

Probability distributions

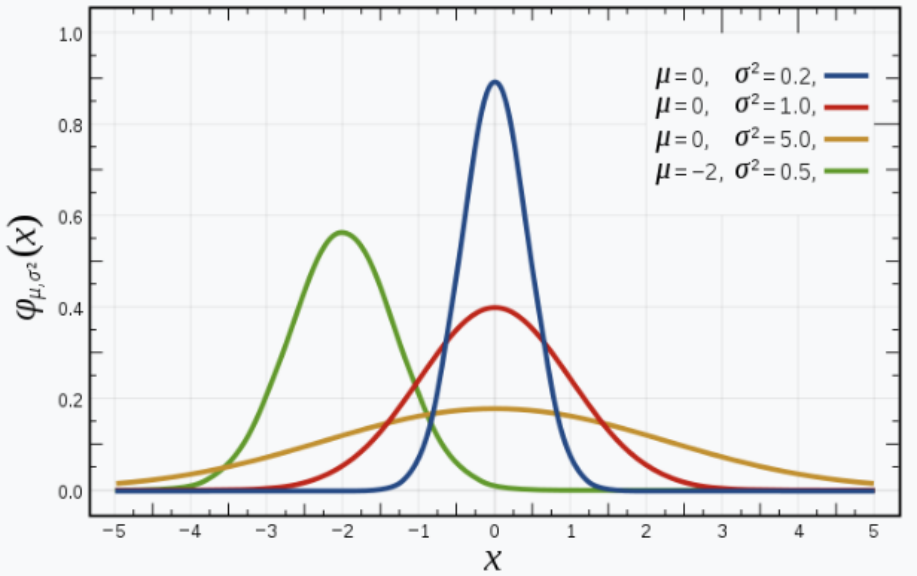
Gaussian

most common noise:

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

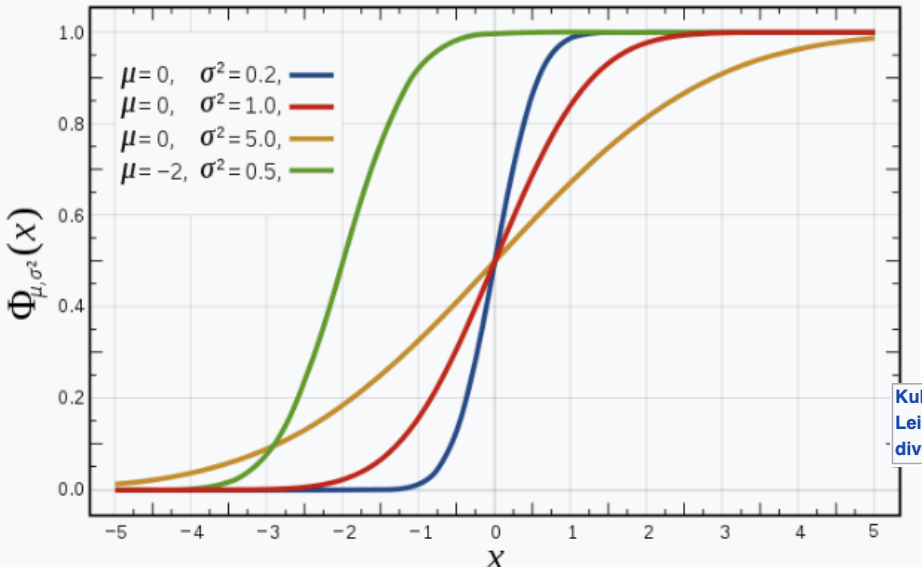
Normal Distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared sc
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	μ
Median	μ
Mode	μ
Variance	σ^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	$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

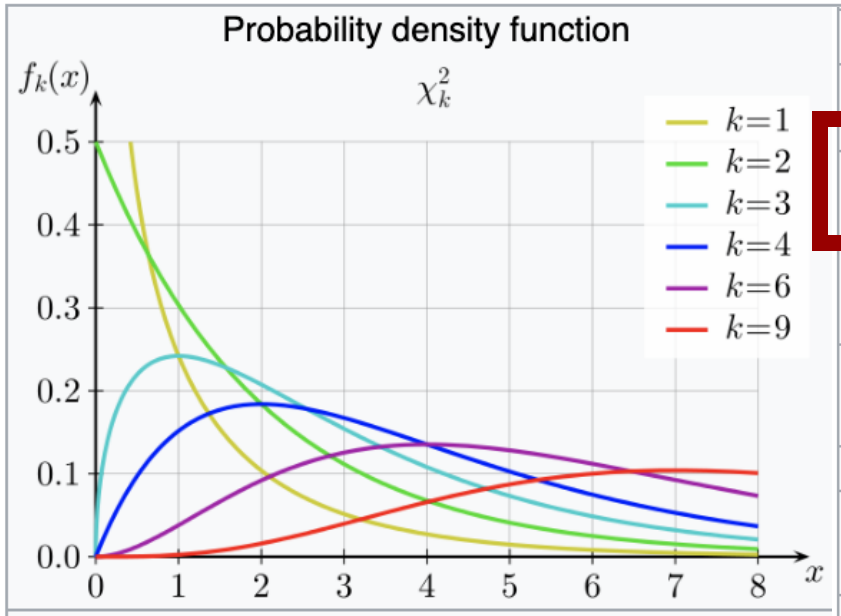
Probability distributions

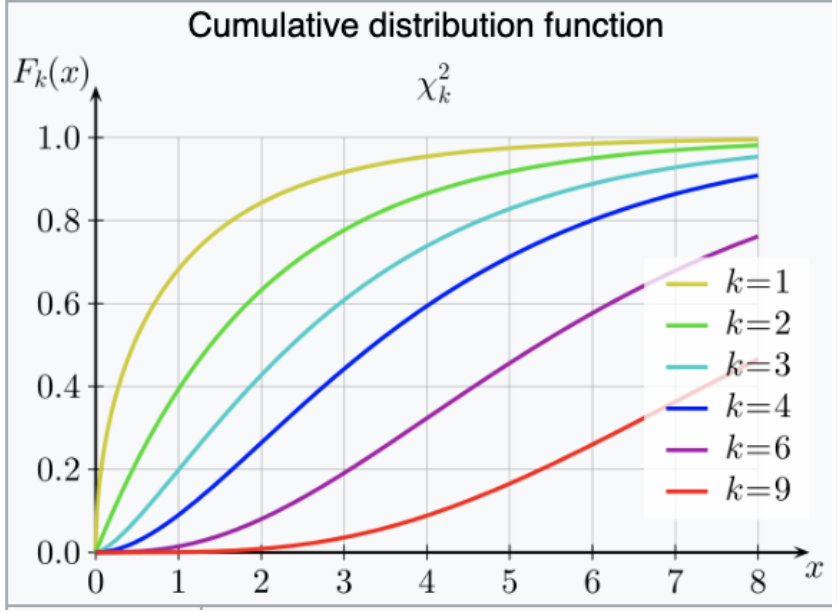
Chi-square (χ2)

turns out its extremely common

many pivotal quantities follow this distribution and thus many tests are based on this

chi-square





Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$
Skewness	$\sqrt{8/k}$
Ex. kurtosis	$\frac{12}{k}$
Entropy	$\frac{k}{2} + \ln(2\Gamma(\frac{k}{2})) + (1 - \frac{k}{2})\psi(\frac{k}{2})$
MGF	$(1-2t)^{-k/2}$ for $t < \frac{1}{2}$
CF	$(1-2it)^{-k/2}$ [1]
PGF	$(1-2\ln t)^{-k/2}$ for $0 < t < \sqrt{e}$

Law of large numbers

Suppose X1, X2, . . . , Xn are independent and identically-distributed, or i.i.d (=independent
random variables with the same underlying distribution).

=>Xi all have the same mean **μ** and standard deviation **σ**.

Let *X* be the mean of Xi $X = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Note that *X* is itself a random variable.

As N grows, the probability that X is close to μ goes to 1.

In the limit of N -> infinity

the mean of a sample of size N approaches the mean of the population **μ**

Central Limit Theorem

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

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

mean **μ** and standard deviation **σ**

In the limit of N -> infinity

the sample mean $\bar{x}$ approaches a Normal (Gaussian) distribution with mean μ and standard deviation σ

regardless of the distribution of X

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

coding time!

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

https://github.com/fedhere/MLPNS_FBianco/blob/main/statistics/distributionParametersDemo.ipynb

key concepts

interpretation of probability

distributions

central limit theorem

homework

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

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

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

David M. Lane et al.

Introduction to Statistics (XVIII)

http://onlinestatbook.com/Online_Statistics_Education.epub

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

resources

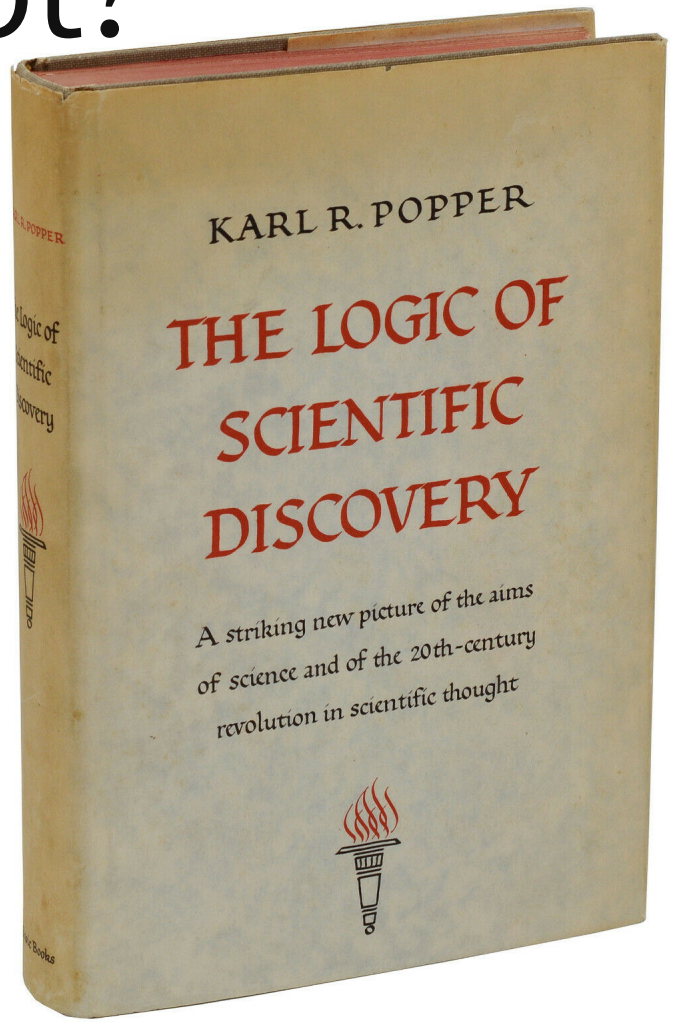
the *demarcation* problem:

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*



4

the scientific method in a probabilistic context

p(physics | data)

https://speakerdeck.com/dfm/emcee-odi

Bayesian Inference

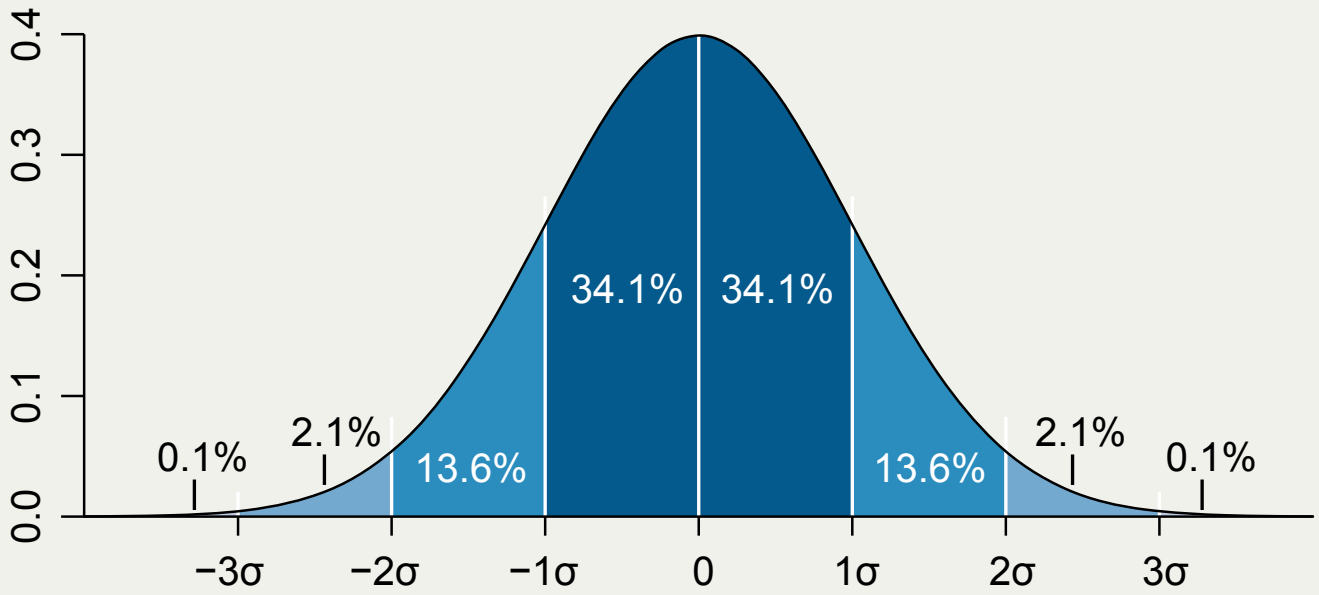
Forward Modeling

Frequentist approach
(NHRT)

p(physics | data)

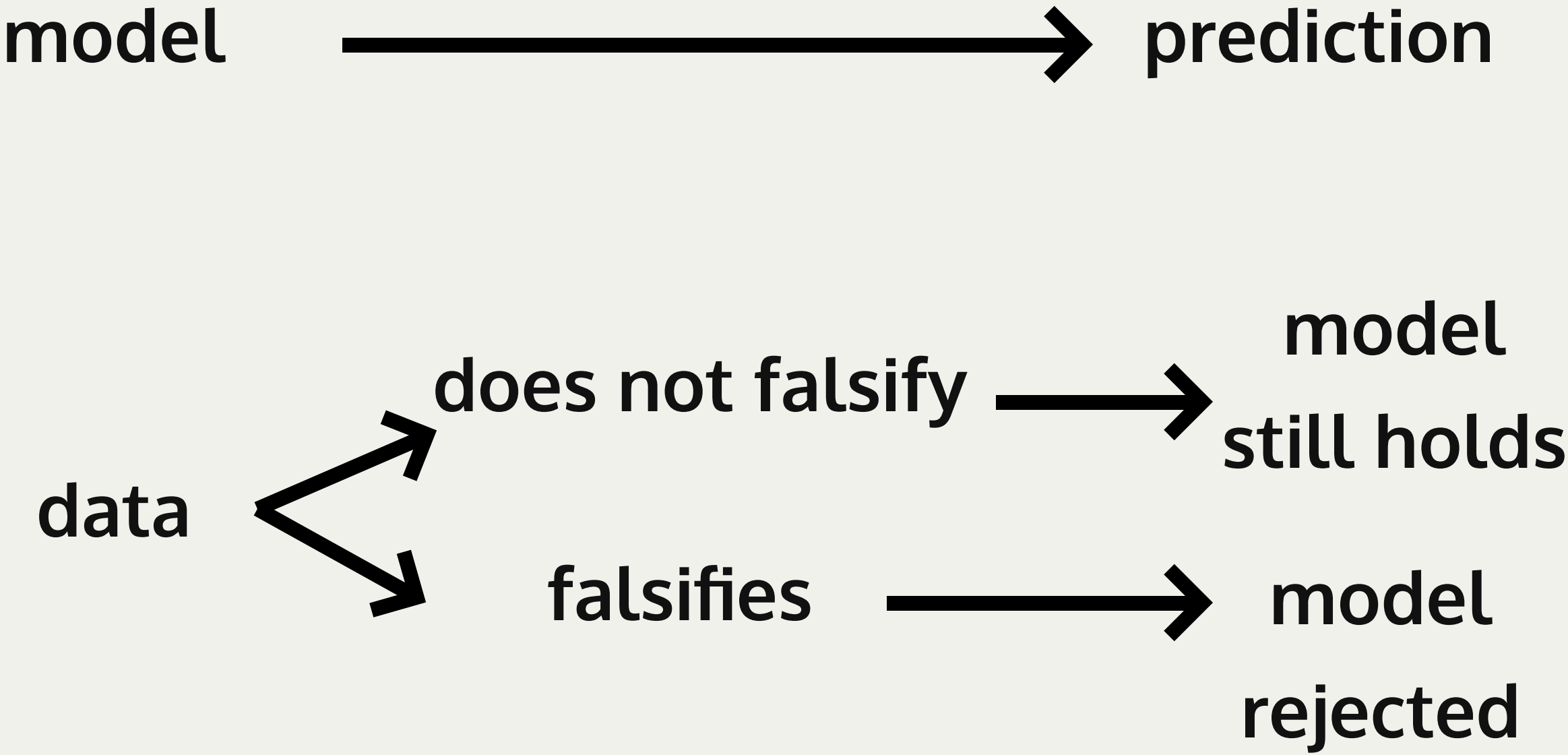
https://speakerdeck.com/dfm/emcee-odi

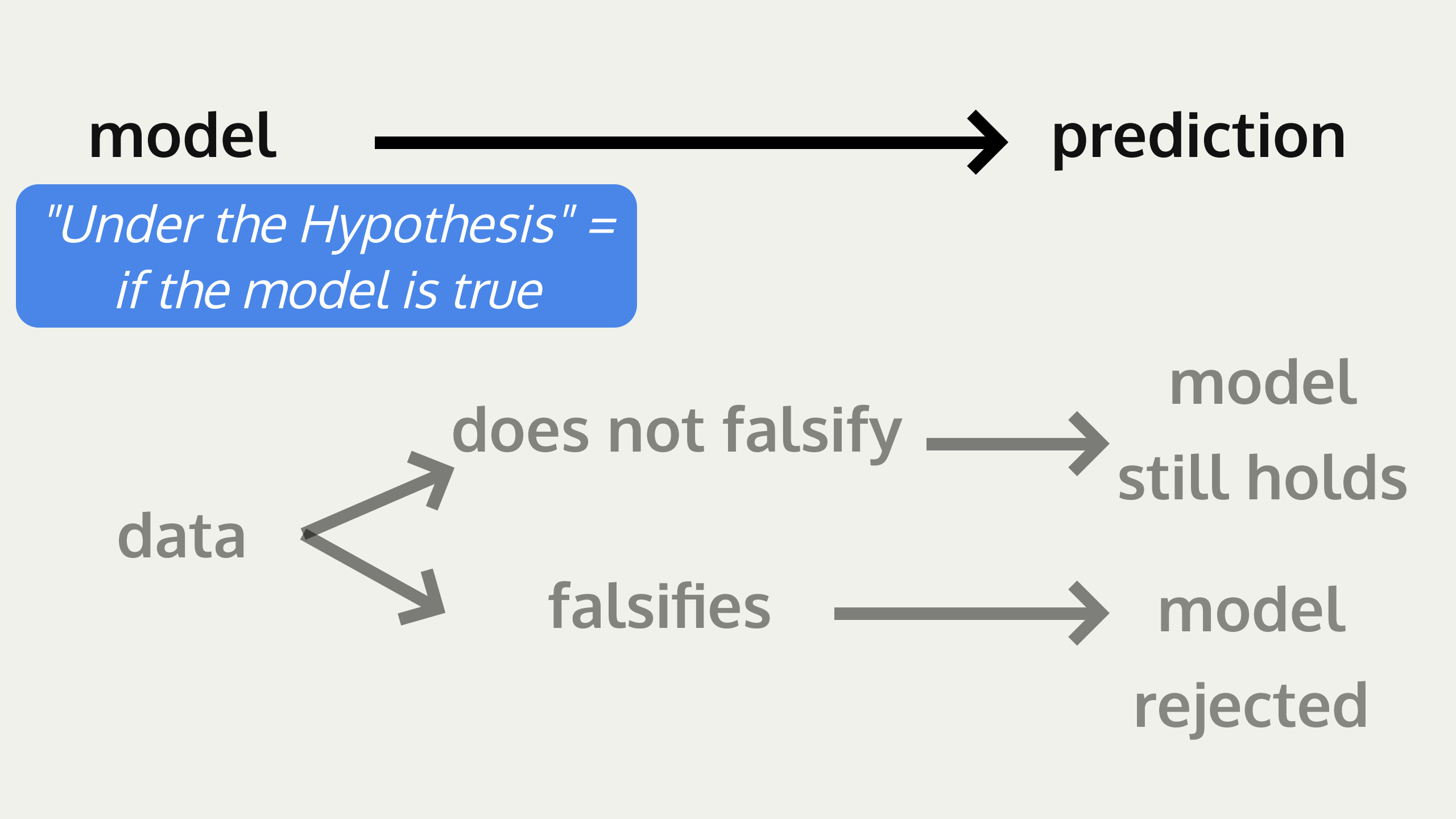
Null
Hypothesis
Rejection
Testing

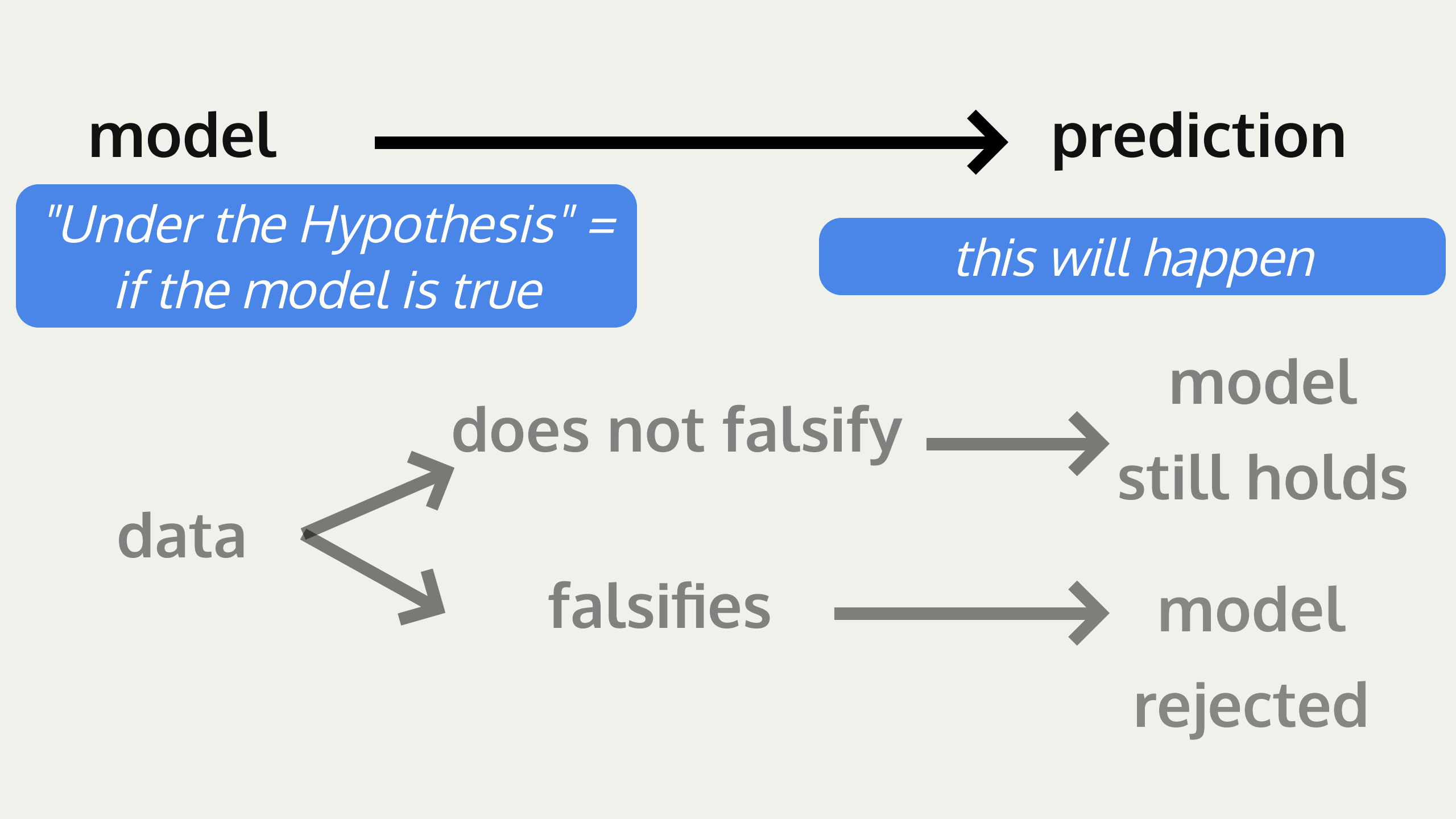


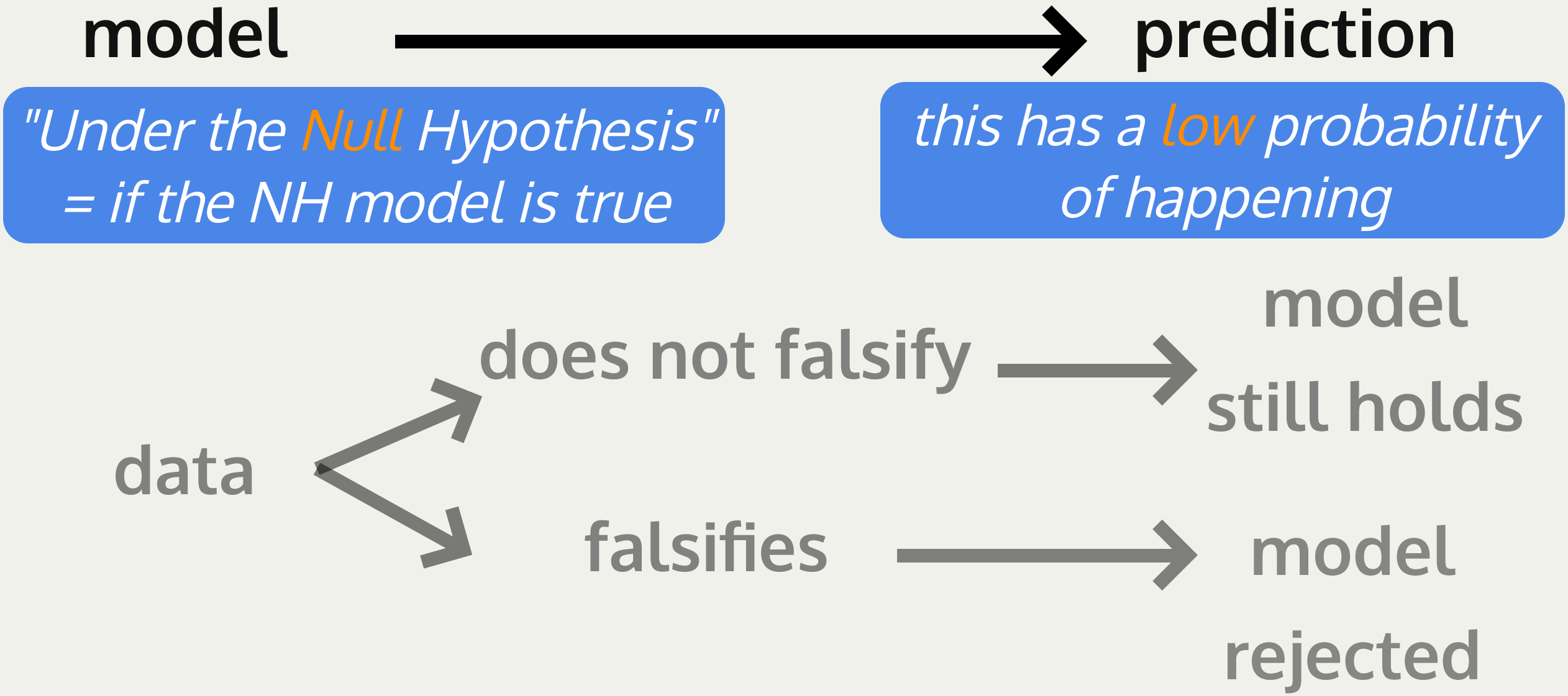
p(physics | data)

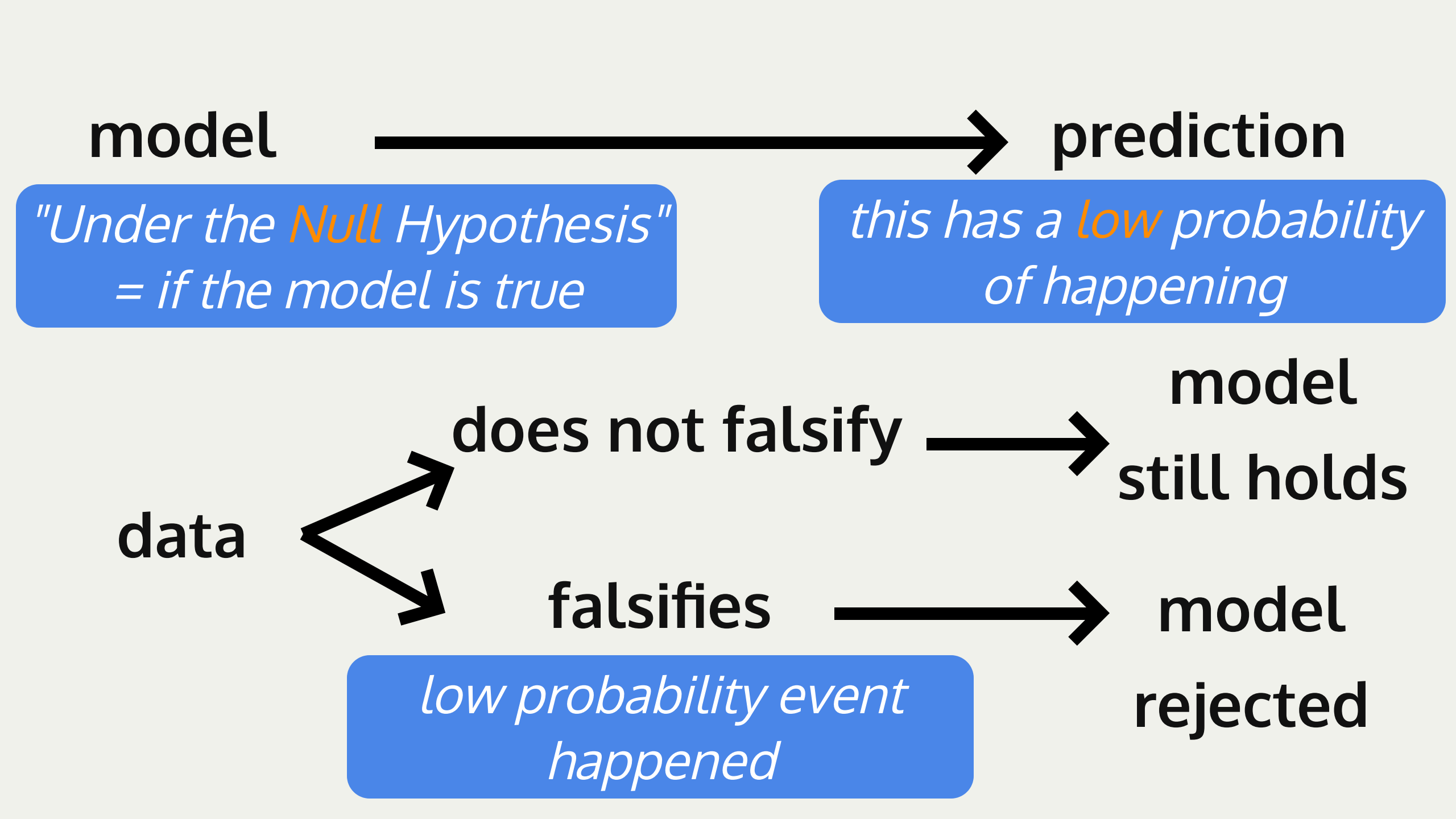
https://speakerdeck.com/dfm/emcee-odi

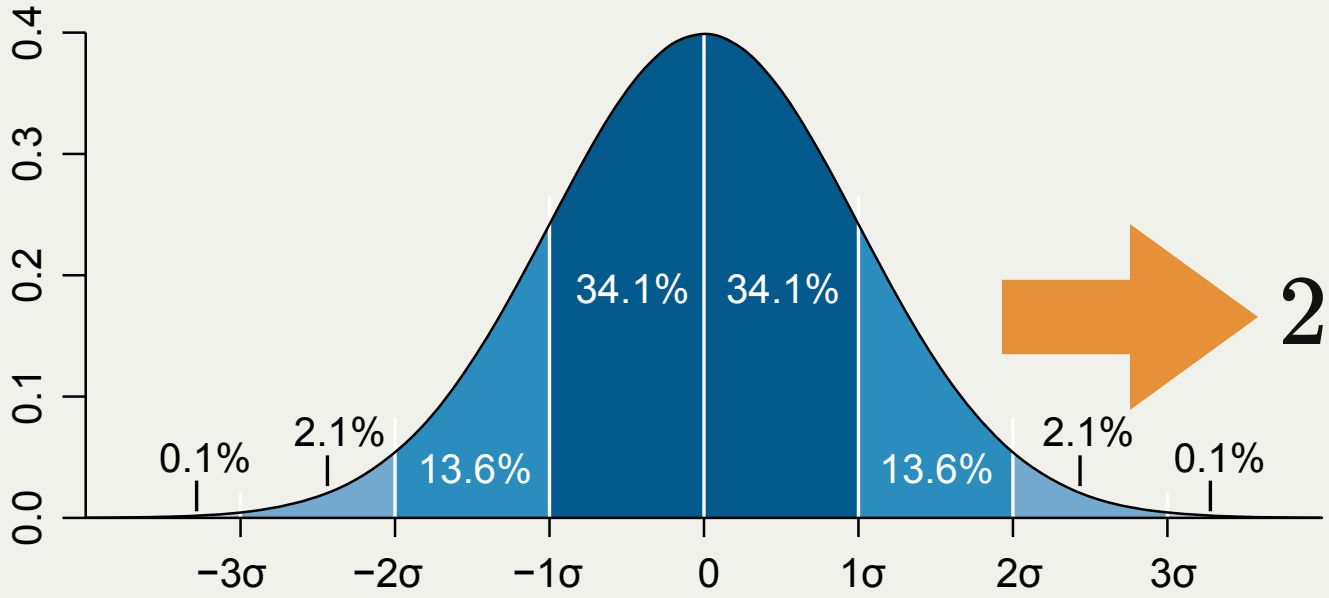

model
prediction
data
does not falsify
model
still holds
falsifies
model
rejected


model
prediction
"Under the Hypothesis" =
if the model is true
data
does not falsify
model
still holds
falsifies
model
rejected


model
prediction
"Under the Hypothesis" =
if the model is true
this will happen
data
does not falsify
model
still holds
falsifies
model
rejected


model
prediction
"Under the Null Hypothesis"
= if the NH model is true
this has a low probability
of happening
data
does not falsify
model
still holds
falsifies
model
rejected


model
prediction
"Under the Null Hypothesis"
= if the model is true
this has a low probability
of happening
data
does not falsify
model
still holds
falsifies
model
rejected
low probability event
happened



2σ

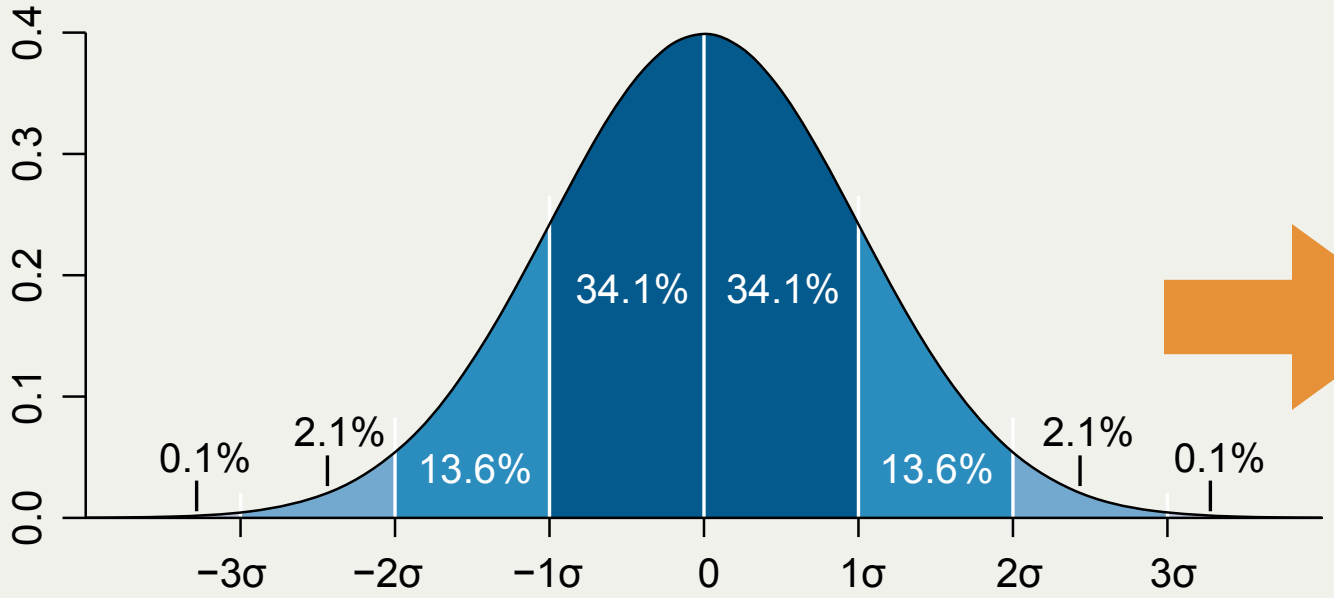
rejected at 95%
0.05 p-value
5% confidence

data

does not falsify → **model still holds**

falsifies → **model rejected**

low probability event happened



rejected at 99.7%

3σ 0.003 p-value

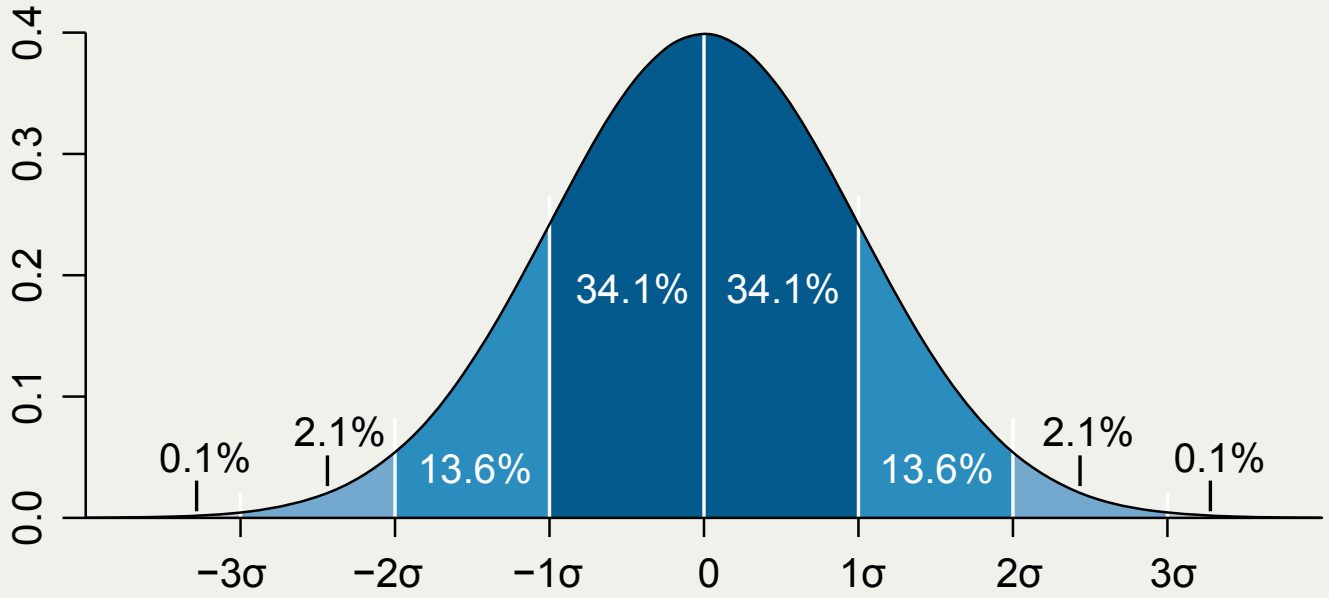
0.3% confidence

data

does not falsify → **model still holds**

falsifies → **model rejected**

low probability event happened



rejected at 99.99...%

3-e7 p-value

3-e5% confidence

5σ

data

does not falsify → **model still holds**

falsifies → **model rejected**

low probability event happened

formulate the Null as the comprehensive opposite of your theory

model → **prediction**

"Under the Null Hypothesis" = if the proposed model is false

this has a low probability of happening

data → **does not falsify alternative** → **model rejected**

data → **falsifies alternative** → **model still holds**

low probability event happened

5 Null hyposthesis rejection testing

Null

Hypothesis

Rejection

Testing

1 formulate your prediction

Null Hypothesis

Null

Hypothesis

Rejection

Testing

2

identify all alternative outcomes

Alternative Hypothesis

Null

Hypothesis

Rejection

Testing

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

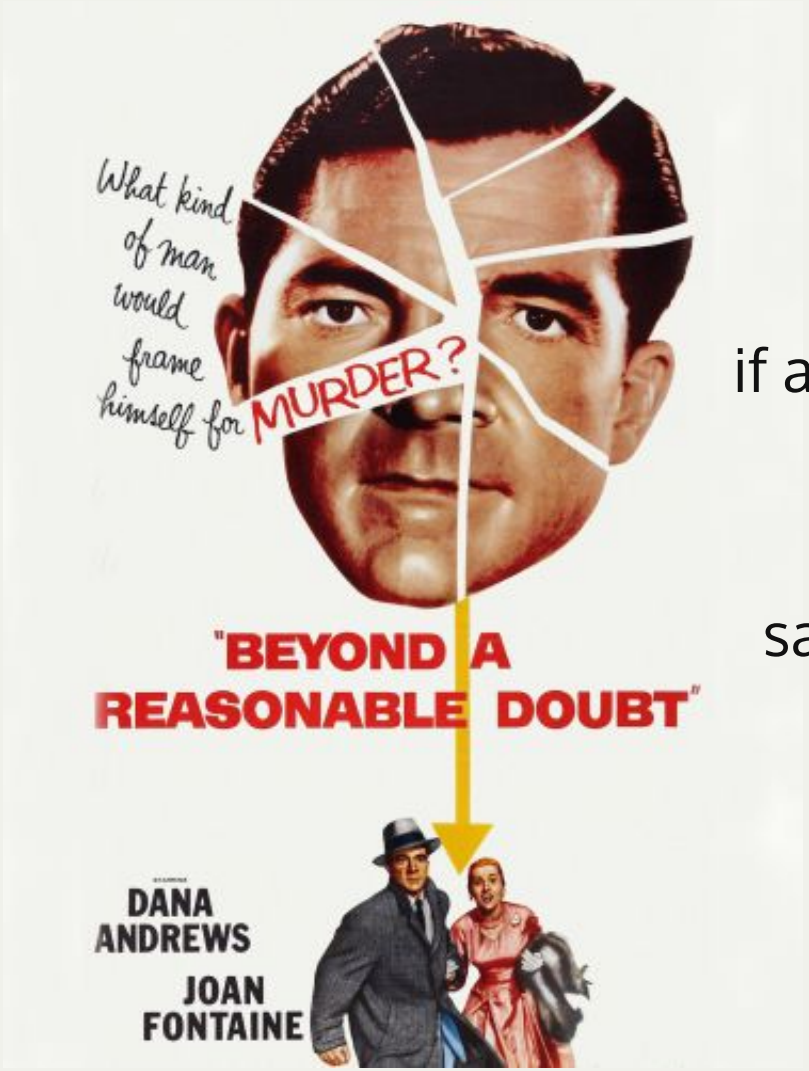
if *all alternatives* to our model are ruled out, then our model must hold

2

identify all alternative outcomes

Alternative Hypothesis

Null
Hypothesis
Rejection
Testing



if all alternatives to our model are ruled out, then our model must hold

same concept guides prosecutorial justice

guilty beyond reasonable doubt

2

identify all alternative outcomes

Alternative Hypothesis

But instead of verifying a theory we want to falsify one

model → **prediction**

"Under the Null Hypothesis"
= if the old model is true

this has a low probability of happening

data → does not falsify → Null still holds

data → falsifies → Null rejected

generally, our model about how the world works is the *Alternative* and we try to reject the non-innovative thinking as the *Null*

But instead of verifying a theory we want to falsify one

model → prediction

"Under the Null Hypothesis"
= if the old model is true

this has a low probability of happening

data → does not falsify → Null still holds

data → falsifies → Null rejected

Earth is flat is Null

Earth is round is Alternative:

we reject the Null hypothesis that the Earth is flat (p=0.05)

But instead of verifying a theory we want to falsify one

model → **prediction**

"Under the Null Hypothesis"
= if the old model is true

this has a low probability of happening

data → does not falsify → Null still holds

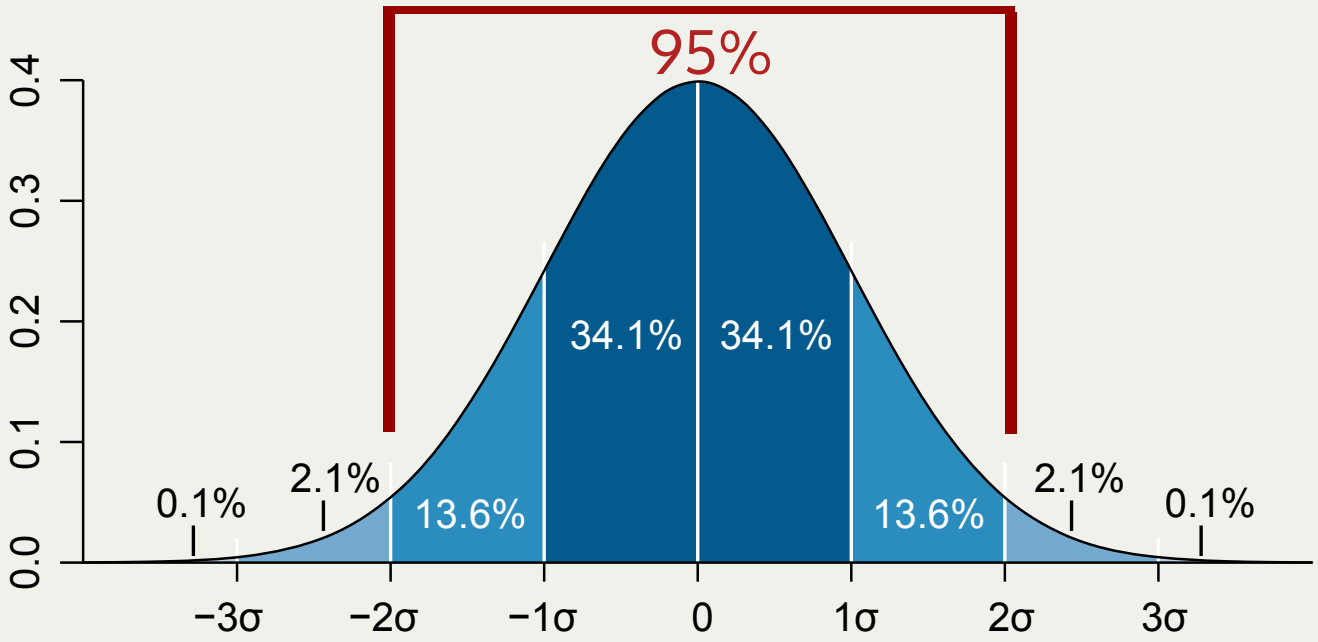
data → falsifies → Null rejected

Earth is flat is Null

Earth is ~~round~~ not flat is Alternative:

we reject the Null hypothesis that the Earth is flat (p=0.05)

Null
Hypothesis
Rejection
Testing



3

set confidence threshold

2σ confidence level

0.05 p-value

95% α threshold

Null

Hypothesis

Rejection

Testing

pivotal quantities

4

find a measurable quantity which under the Null has a known distribution

Null

Hypothesis

Rejection

Testing

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

4

if a quantity follows a known distribution, once I measure its value I can what the probability of getting that value actually is! was it a likely or an unlikely draw?

Null

Hypothesis

Rejection

Testing

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

also called "statistics"

e.g.: *χ2 statistics:* difference between expetation and reality squared

Z statistics: difference between means

K-S statistics: maximum distance of cumulative distributions.

Null
Hypothesis
Rejection
Testing

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

$$p(\text{pivotal quantity}|NH) \sim p(NH|D)$$

Null

Hypothesis

Rejection

Testing

pivotal quantities

5

calculate it!

Null

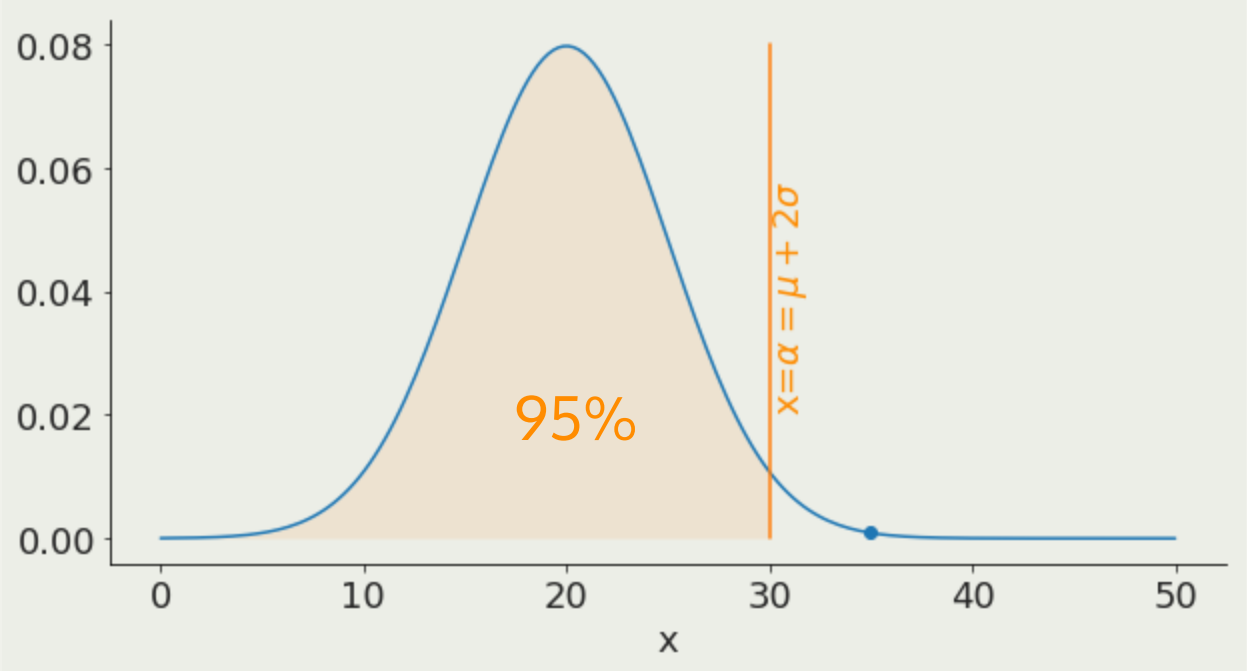
Hypothesis

Rejection

Testing

what is α?

α is the x value corresponding to a chosen threshold



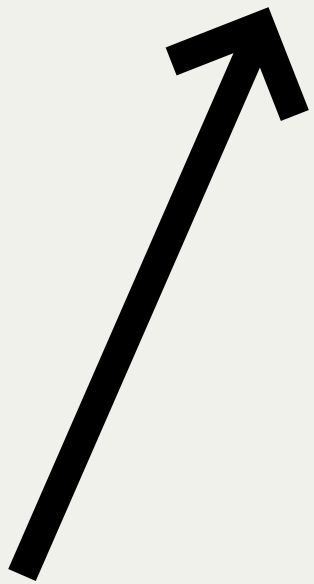
6

test data against
alternative outcomes

Null
Hypothesis
Rejection
Testing

$$p(NH|D) < \alpha$$

prediction is unlikely
Null rejected
Alternative holds



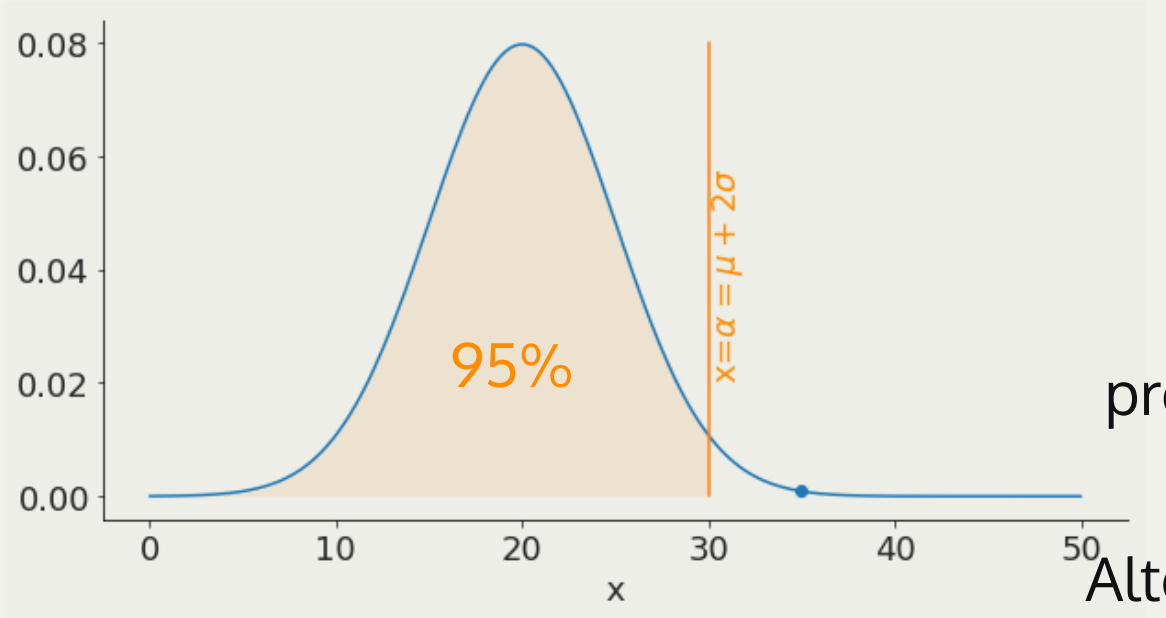
6

test data against
alternative outcomes

Null
Hypothesis
Rejection
Testing

$$p(NH|D) < \alpha$$

prediction is unlikely
Null rejected
Alternative holds



prediction is likely
Null holds
Alternative rejected

6

test data against
alternative outcomes

$$p(NH|D) \geq \alpha$$

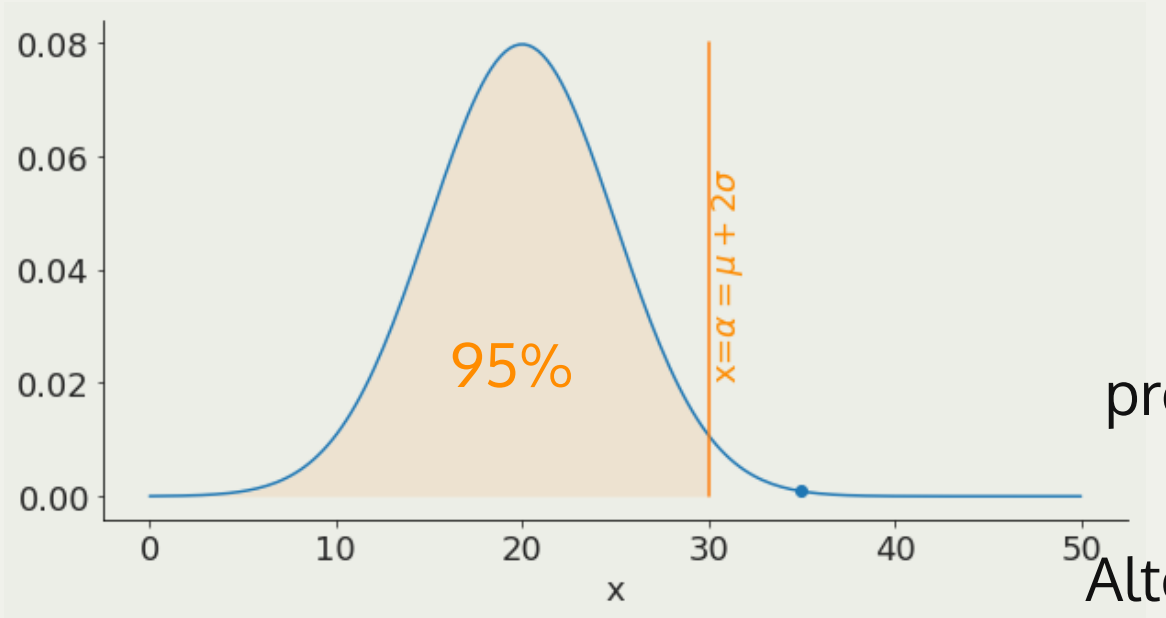
Null
Hypothesis
Rejection
Testing

6

test data against
alternative outcomes

$$p(NH|D) < \alpha$$

prediction is unlikely
Null rejected
Alternative holds



$$p(NH|D) \geq \alpha$$

prediction is likely
Null holds
Alternative rejected

formulate the Null as the comprehensive opposite of your theory

model → **prediction**

"Under the Null Hypothesis" = if the model is false

this has a low probability of happening

data

does not falsify alternative → **everything but model is rejected**

falsifies alternative → **model holds**

Key Slide

low probability event happened

1 formulate your prediction (NH)

2 identify all alternative outcomes (AH)

3 set confidence threshold (*p*-value)

4 find a measurable quantity which under the Null has a known distribution (pivotal quantity)

5 calculate the pivotal quantity

6 calculate probability of value obtained for the pivotal quantity under the Null

Key Slide

if probability < *p*-value : reject Null

Jacob Cohen, 1994

The earth is round (p=0.05)

http://fbb.space/dsps/Cohen1994_TheEarthIsRound_AmPsych.pdf

the original link:

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

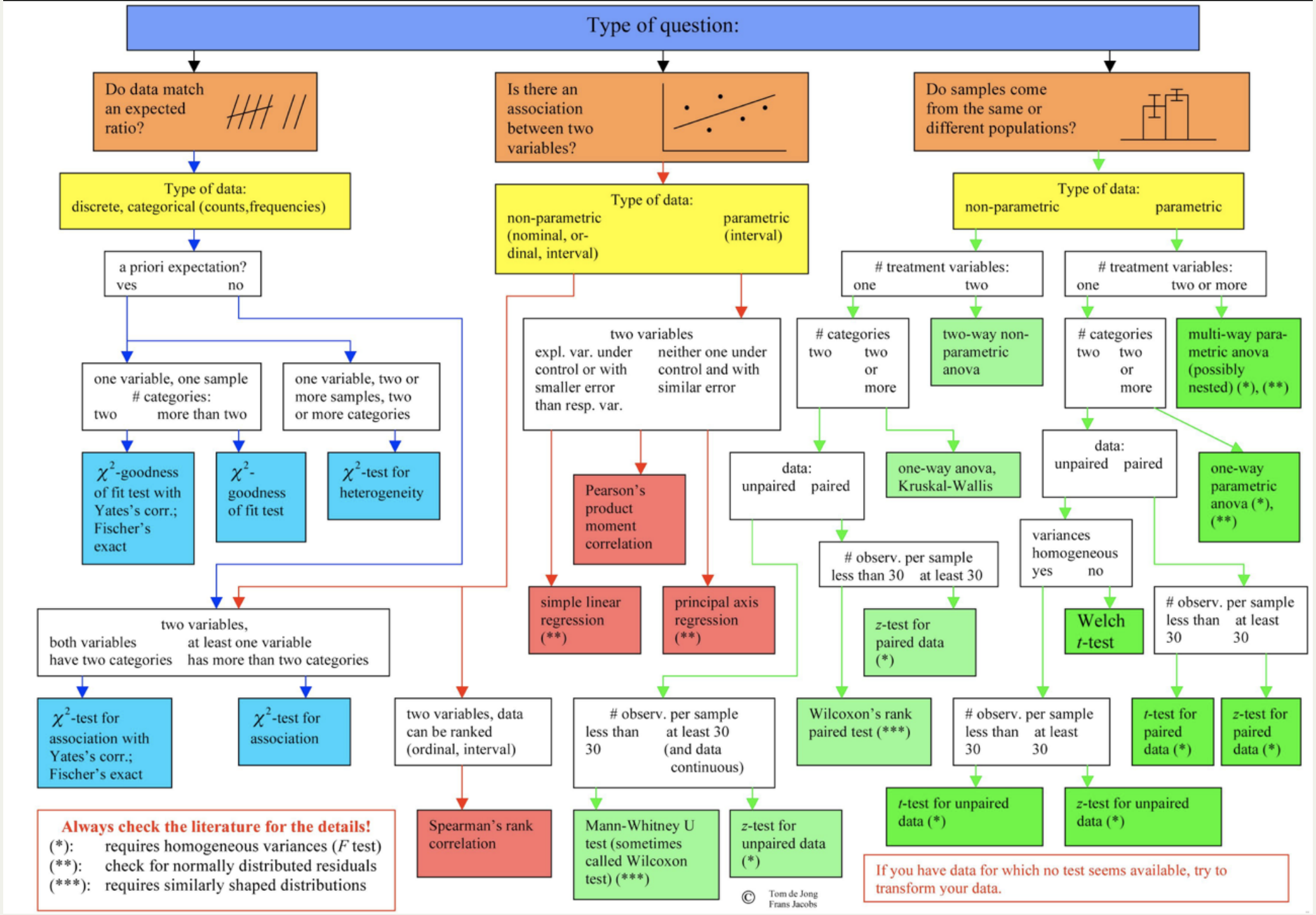
(this link nees access to science magazine, but ou can use the link above which is the same file)

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.*

6 common tests and pivotal quantities


Type of question:
Do data match an expected ratio?
Type of data: discrete, categorical (counts,frequencies)
a priori expectation? yes no
one variable, one sample # categories: two more than two
one variable, two or more samples, two or more categories
χ^2-goodness of fit test with Yates's corr.; Fischer's exact
χ^2-goodness of fit test
χ^2-test for heterogeneity
two variables, both variables have two categories at least one variable has more than two categories
χ^2-test for association with Yates's corr.; Fischer's exact
χ^2-test for association
Is there an association between two variables?
Type of data: non-parametric (nominal, ordinal, interval) parametric (interval)
two variables expl. var. under control or with smaller error than resp. var. neither one under control and with similar error
Pearson's product moment correlation
simple linear regression (**)
principal axis regression (**)
two variables, data can be ranked (ordinal, interval)
Spearman's rank correlation
observ. per sample less than 30 at least 30 (and data continuous)
Mann-Whitney U test (sometimes called Wilcoxon test) (***)
z-test for unpaired data (*)
Do samples come from the same or different populations?
Type of data: non-parametric parametric
treatment variables: one two
categories two two or more
two-way non-parametric anova
data: unpaired paired
one-way anova, Kruskal-Wallis
observ. per sample less than 30 at least 30
z-test for paired data (*)
Wilcoxon's rank paired test (***)
observ. per sample less than 30 at least 30
t-test for unpaired data (*)
z-test for unpaired data (*)
treatment variables: one two or more
categories two two or more
multi-way parametric anova (possibly nested) (*), (**)
data: unpaired paired
one-way parametric anova (*), (**)
variances homogeneous yes no
Welch t-test
observ. per sample less than 30 at least 30
t-test for paired data (*)
z-test for paired data (*)
Always check the literature for the details!
(*): requires homogeneous variances (F test)
(**): check for normally distributed residuals
(***): requires similarly shaped distributions
If you have data for which no test seems available, try to transform your data.
© Tom de Jong Frans Jacobs

Z-test

Is the mean of a sample *with known variance* the same as that of a known population?

pivotal quantity

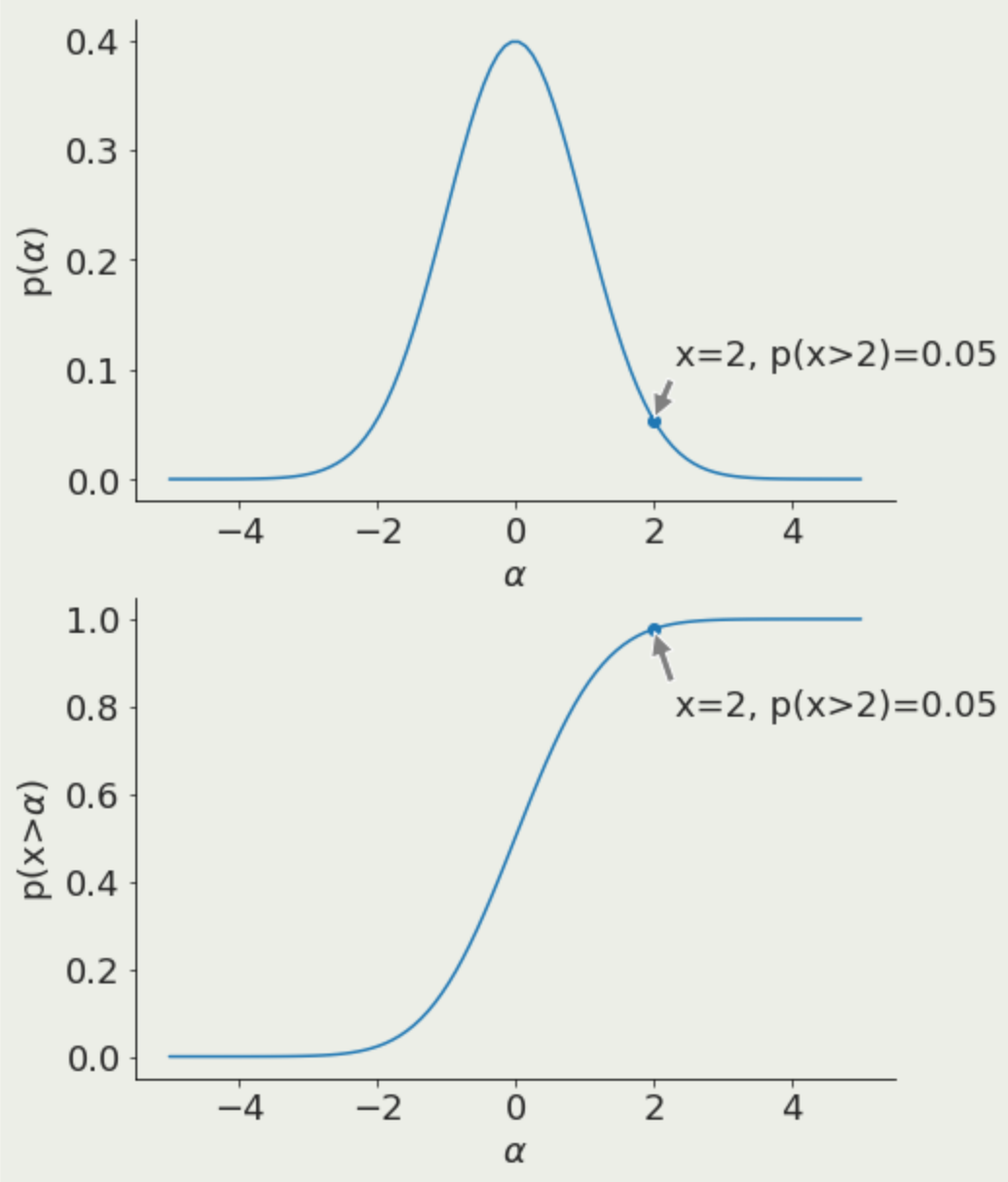
$$Z = (\bar{X} - \mu 0)/s$$

sample mean

population mean

sample variance = $\sigma/\sqrt{n}$

$$Z \sim N(\mu = 0,\ \sigma = 1)$$



Z-test

Is the mean of a sample *with known variance* the same as that of a known population?

pivotal quantity

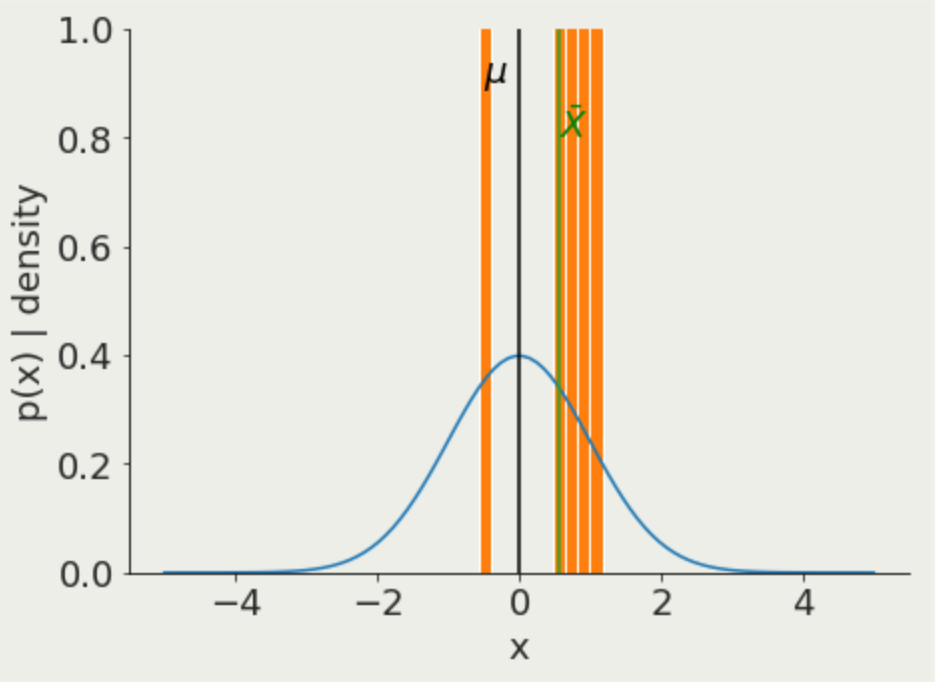
$$Z = (\bar{X} - \mu 0)/s$$

sample mean

population mean

sample variance = $\sigma/\sqrt{n}$

$$Z \sim N(\mu = 0,\ \sigma = 1)$$



why do we need a test? why not just measuring the means and seeing it they are the same?

Z-test

Is the mean of a sample *with known variance* the same as that of a known population?

pivotal quantity

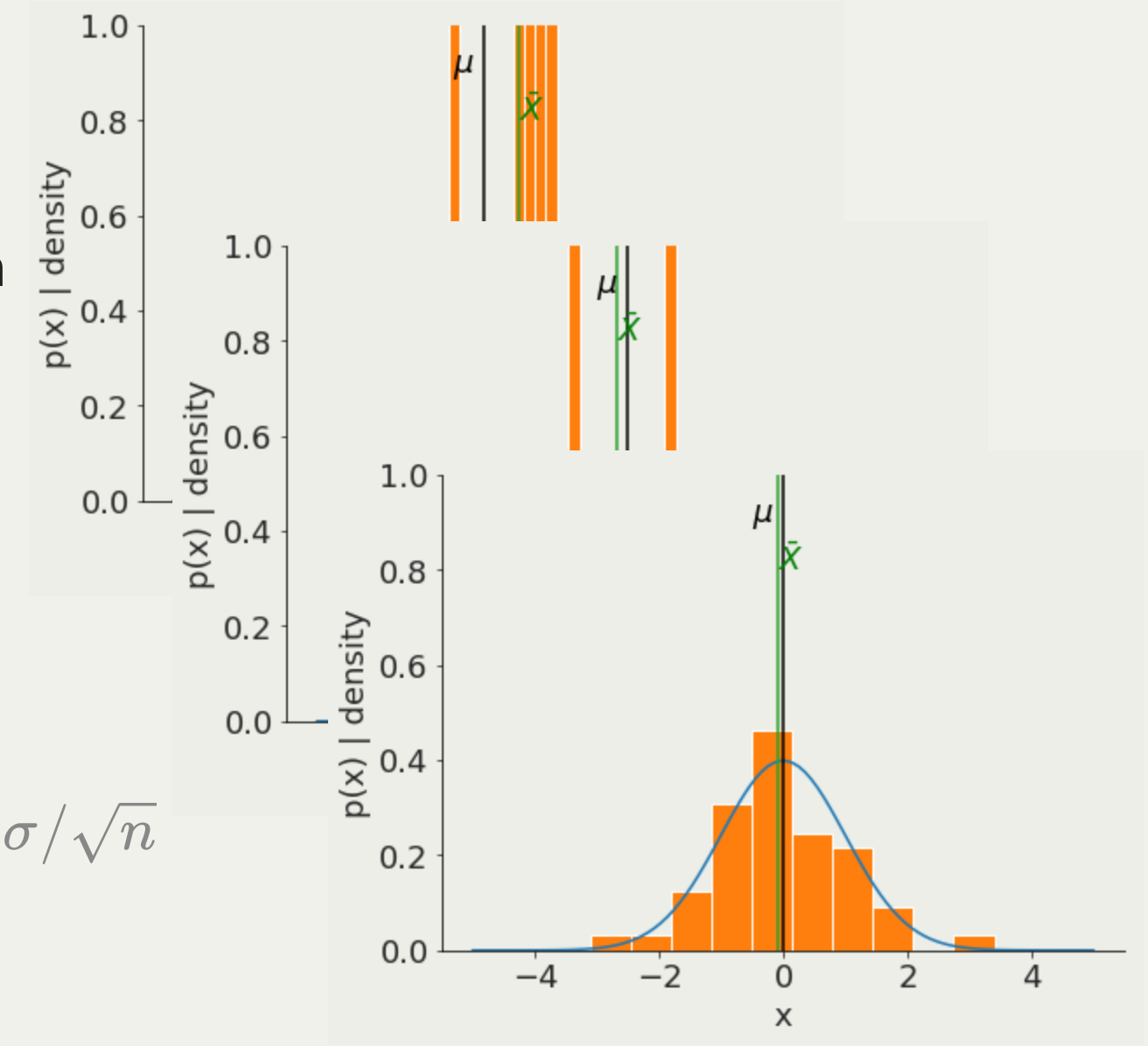
$$Z = (\bar{X} - \mu 0)/s$$

sample mean

population mean

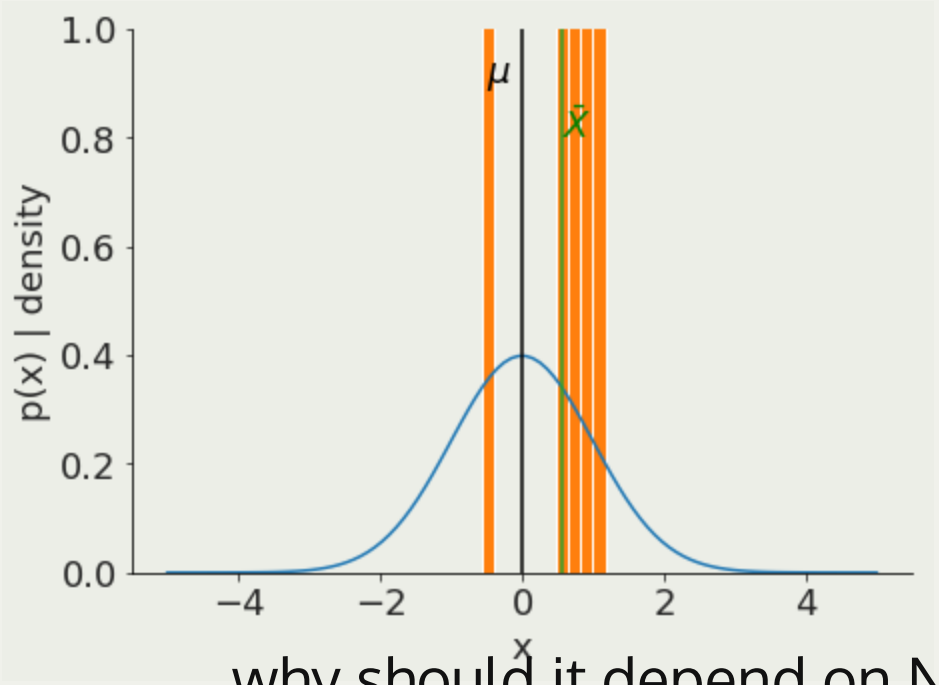
sample variance = $\sigma/\sqrt{n}$

$$Z \sim N(\mu = 0,\ \sigma = 1)$$



Z-test

Is the mean of a sample *with known variance* the same as that of a known population?



pivotal quantity

$$Z = (\bar{X} - \mu 0)/s$$

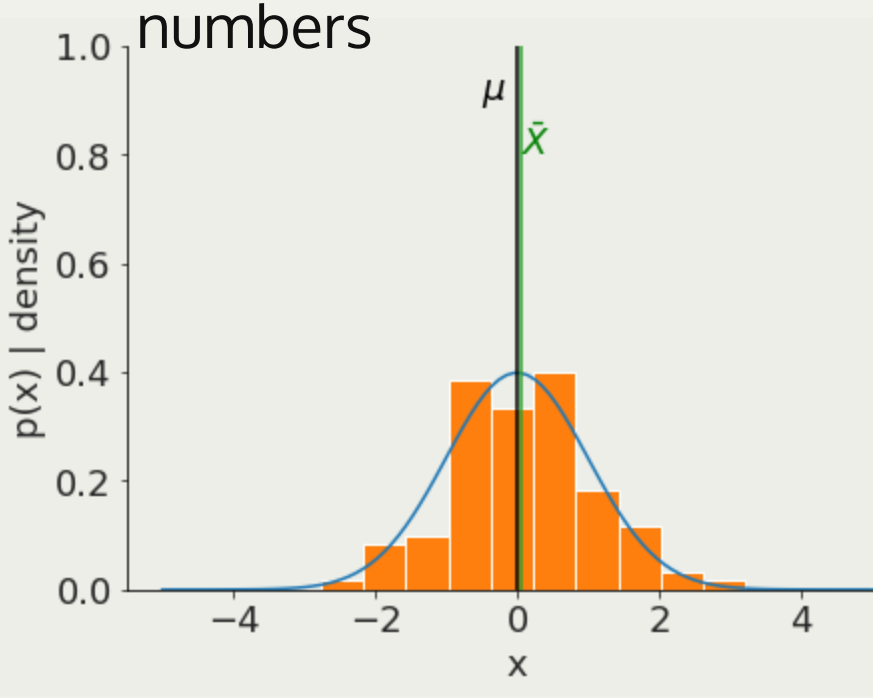
sample mean

population mean

sample variance = $\sigma/\sqrt{n}$

$$Z \sim N(\mu = 0,\ \sigma = 1)$$

why should it depend on N? Law of large numbers



Z-test

Is the mean of a sample *with known variance* the same as that of a known population?

pivotal quantity

$$Z = (\bar{X} - \mu 0)/s$$

sample mean

population mean

sample variance = $\sigma/\sqrt{n}$

$$Z \sim N(\mu = 0,\ \sigma = 1)$$

The Z test provides a trivial interpretation of the measured quantity: the Z value is exactly the distance for the mean of the standard distribution of possible outcomes *in units of standard deviation*

so a result of 0.13 means we are 0.13 standard deviations to the mean (p>0.05)

t– test

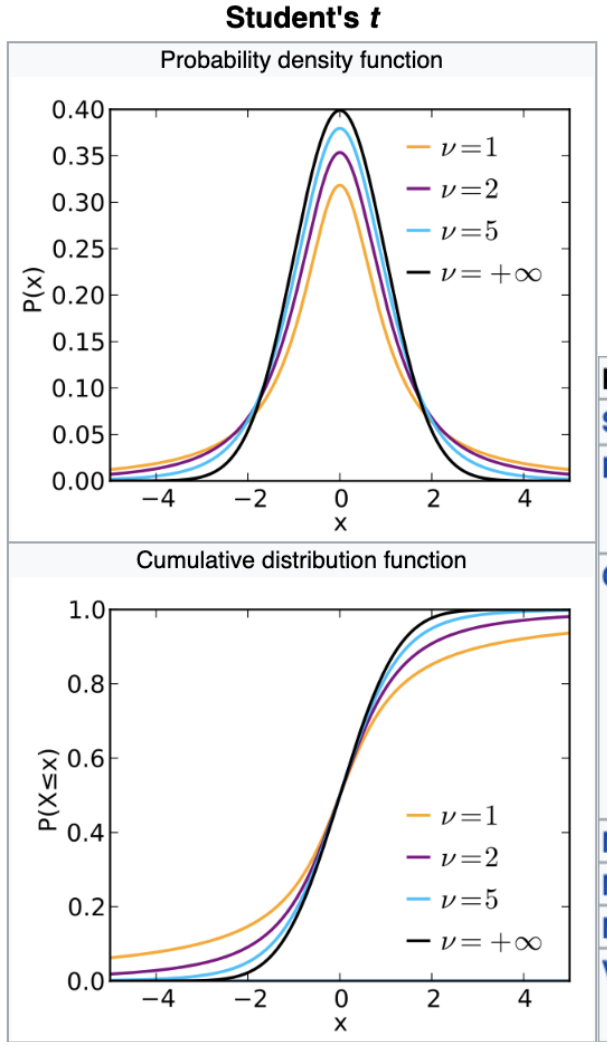
Are the means of 2 samples significantly different?

pivotal quantity

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample



WIKIPEDIA
The Free Encyclopedia

Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty, \infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric functio
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$, otherwise undefined

$$t \sim \text{Student's } t \left(\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right)$$

t– test

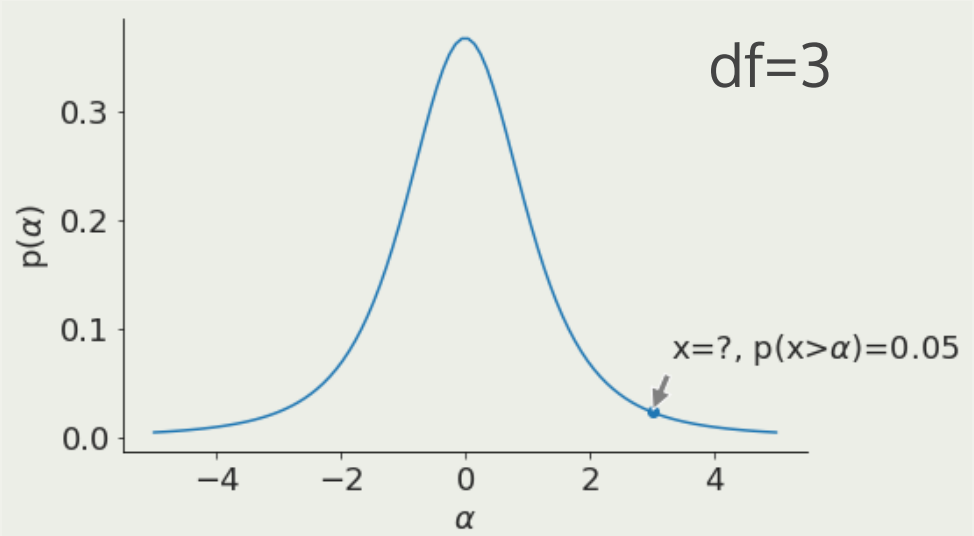
Are the means of 2 samples significantly different?

pivotal quantity

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample



$$t \sim \text{Student's } t \left(\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right)$$

t– test

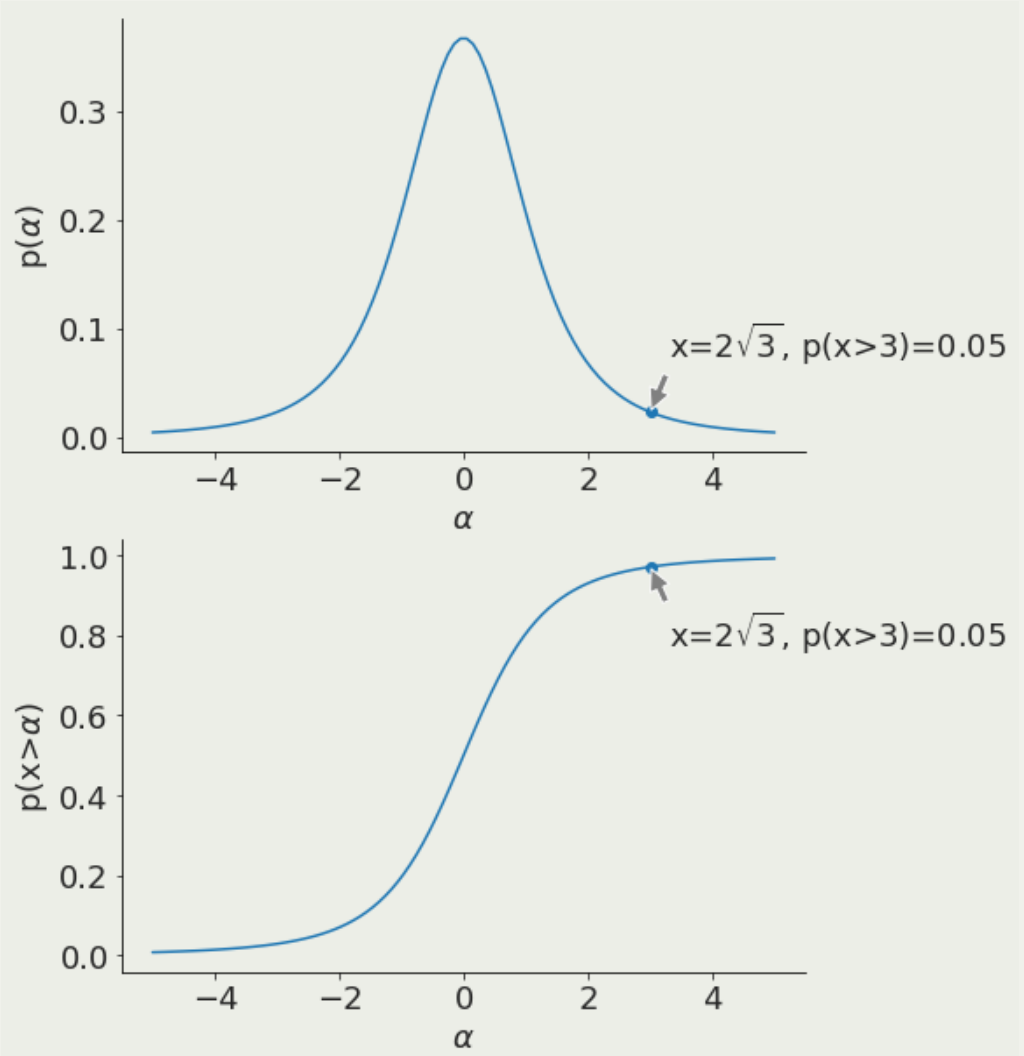
Are the means of 2 samples significantly different?

pivotal quantity

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample



$$t \sim \text{Student's } t \left(\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right)$$

t– test

Are the means of 2 samples significantly different?

To interpret the outcome of a t-test I have to figure out the probability of a give p

pivotal quantity

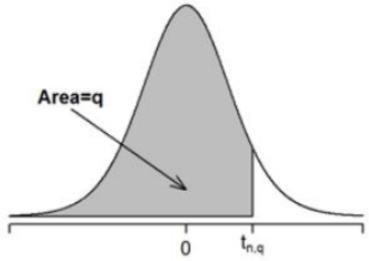
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample

Quartiles of the t Distribution

The table gives the value if $t_{n;q}$ - the qth quantile of the t distribution for n degrees of freedom



	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
n = 1	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781

$$t \sim \text{Student's } t \left(\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right)$$

K-S test

Kolmogorof-Smirnoff :

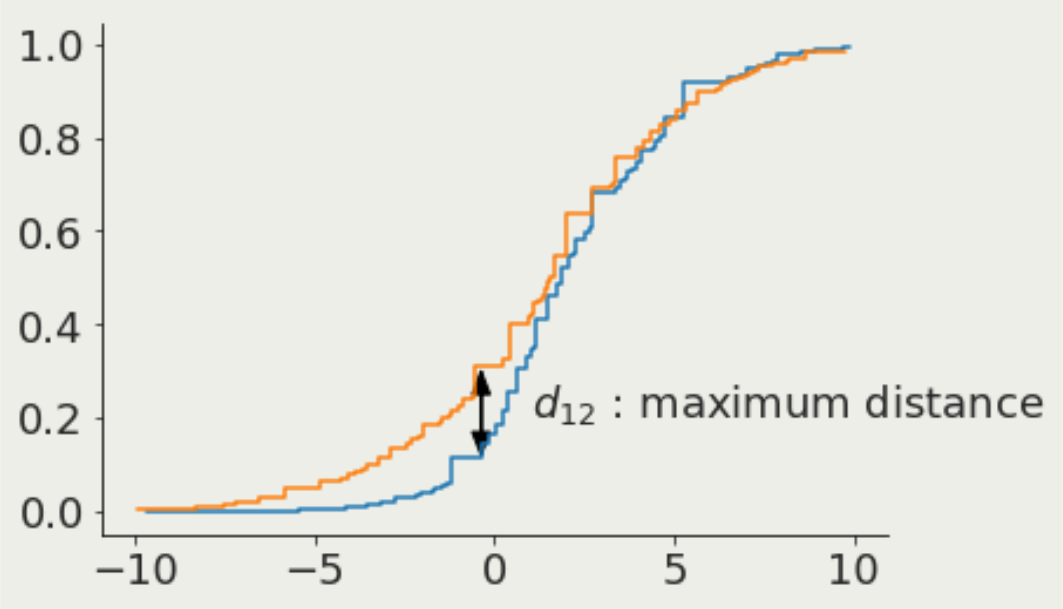
do two samples come from the same parent distribution?

pivotal quantity

$$d_{12} \equiv max_x \left| C_1(x) - C_2(x) \right|$$

Cumulative distribution 1

Cumulative distribution 2



$$P(d > observed) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}\sqrt{\frac{N_1 N_2}{N_1+N_2}D}$$

K-S test

Kolmogorof-Smirnoff :

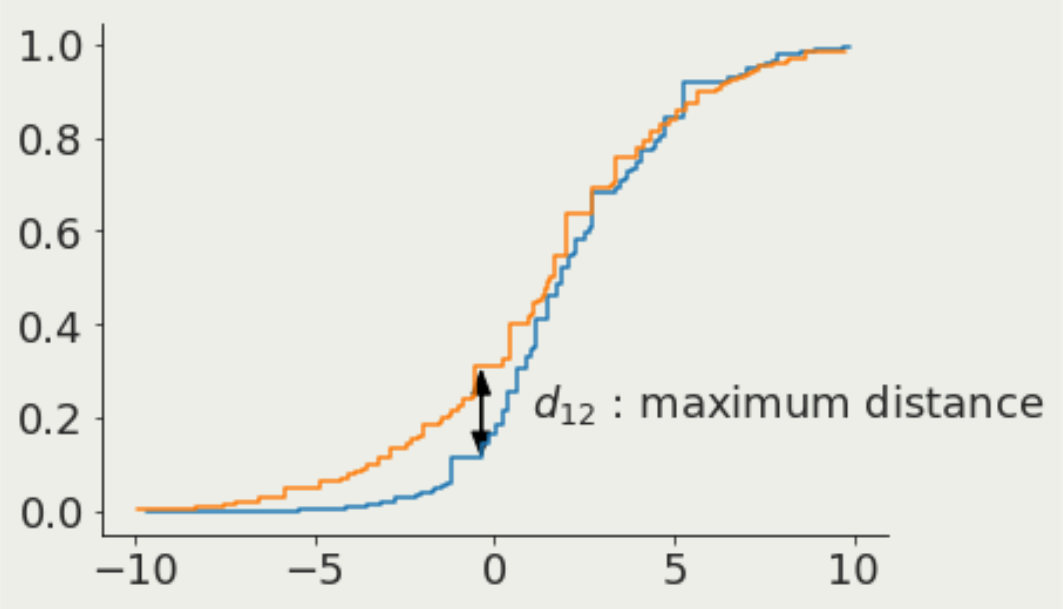
do two samples come from the same parent distribution?

pivotal quantity

$$d_{12} \equiv max_x \left| C_1(x) - C_2(x) \right|$$

Cumulative distribution 1

Cumulative distribution 2



$$P(d > observed) =$$

```
sp.stats.ks_2samp(x, y)
```

executed in 7ms, finished 14:45:10 2019-09-09

```
Ks_2sampResult(statistic=0.4, pvalue=0.312852676016955 8)
```

$\chi 2$ test

are the data what is expected from the model (if likelihood is Gaussian... we'll see this later) - ther are a few χ2 tests. The one here is the "Pearson's χ2 tests"

pivotal quantity

$$\chi^2 \equiv \sum_i \frac{(f(x_i) - y_i)^2}{\sigma_i^2}$$

model

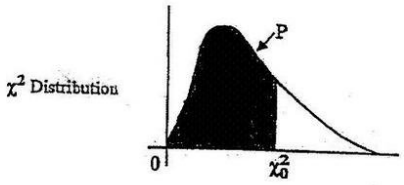
uncertainty

observation

$$\chi^2 \sim \chi^2(df = n - 1)$$

number of observation

this should actually be the number of params in the model



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] = P$ for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

$\chi 2$ test

are the data what is expected from the model (if likelihood is Gaussian... we'll see this later) - ther are a few χ2 tests. The one here is the "Pearson's χ2 tests"

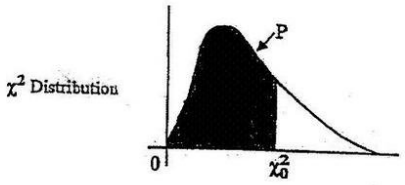
pivotal quantity

$$\chi^2 \equiv \sum_i \frac{(f(x_i) - y_i)^2}{\sigma_i^2}$$

model — observation — uncertainty

$$\frac{\chi^2}{n-1} \sim \chi^2(df = 1)$$

number of observation



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] = P$ for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
								535	20.090	21.955
								023	21.666	23.589
								483	23.209	25.188
								920	24.725	26.757
								337	26.217	28.300
								736	27.688	29.819
								119	29.141	31.319
								488	30.578	32.801
								345	32.000	34.267
								191	33.409	35.718
								526	34.805	37.156
								852	36.191	38.582
									37.566	39.997

Parameters	$k \in \mathbb{N}^*$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k

the *demarcation* problem in *Bayesian* context

7

beyond frequentism and NHRT

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.

- *Thomas Bayes* *Essay towards solving a Problem in the Doctrine of Chances* (1763)

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.

- *Thomas Bayes* *Essay towards solving a Problem in the Doctrine of Chances* (1763)

$$p(M|D) =$$

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.[1]

- *Thomas Bayes Essay towards solving a Problem in the Doctrine of Chances* (1763)

$p(M|D) = P(M)...$ *"prior"*

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.

- *Thomas Bayes* *Essay towards solving a Problem in the Doctrine of Chances* (1763)

$$p(M|D) = P(M)\ P(D|M)...$$

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.

- *Thomas Bayes* *Essay towards solving a Problem in the Doctrine of Chances* (1763)

$$p(M|D) = \frac{P(M)\ P(D|M)}{P(D)}$$

"evidence"

Null
Hypothesis
Rejection
Testing

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

$$p(\text{pivotal quantity}|NH) \sim p(NH|D)$$

Bayes theorem

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

Bayes theorem

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

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

Bayes theorem

$$P(\text{model}|\text{data})P(\text{data}) = P(\text{data}|\text{model})P(\text{model})$$

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}$$

Bayes theorem

$$\underbrace{P(\theta|\mathrm{D})}_{\text{posterior}} = \frac{\overbrace{P(\mathrm{D}|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\mathrm{D})}_{\text{evidence}}}$$

θ model parameters

D data

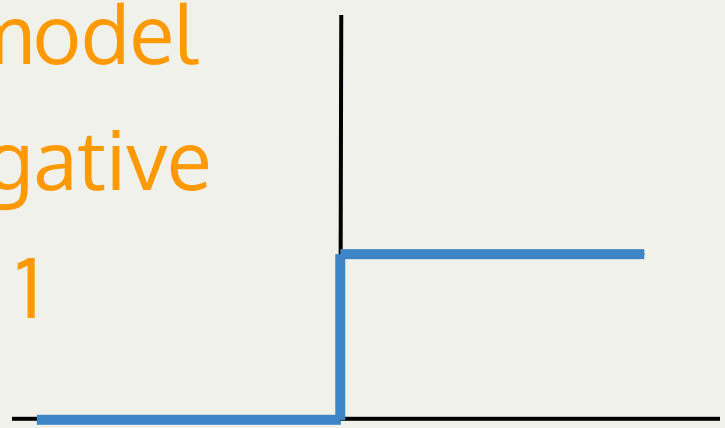
Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

prior: constraints on the model

e.g. flux is never negative

P(f<0) = 0 P(f>=0) = 1



θ model parameters

D data

Bayes theorem

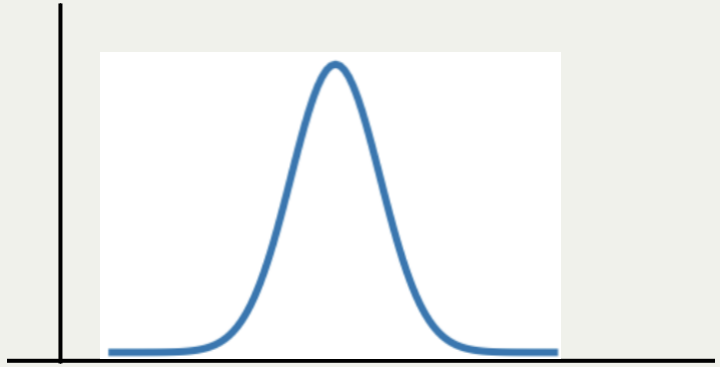
$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

prior: constraints on the model

people's weight <1000lb

& people's weight >0lb

P(w) ~ $\mathcal{N}$(105lb, 90lb)



θ model parameters

D data

Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

θ model parameters

D data

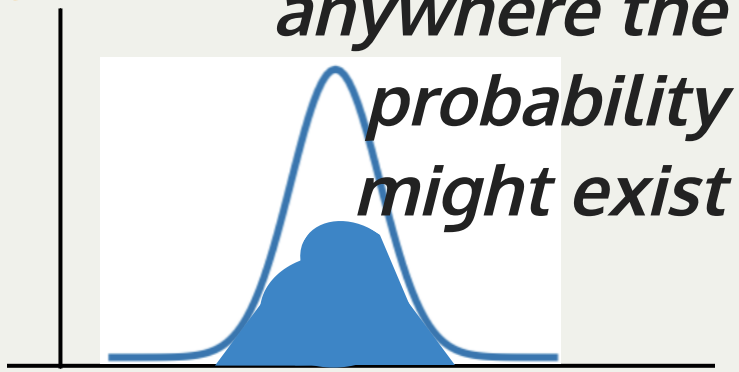
prior: constraints on the model

people's weight <1000lb

& people's weight >0lb

P(w) ~ $\mathcal{N}$(105lb, 90lb)

the prior should not be 0 anywhere the probability might exist



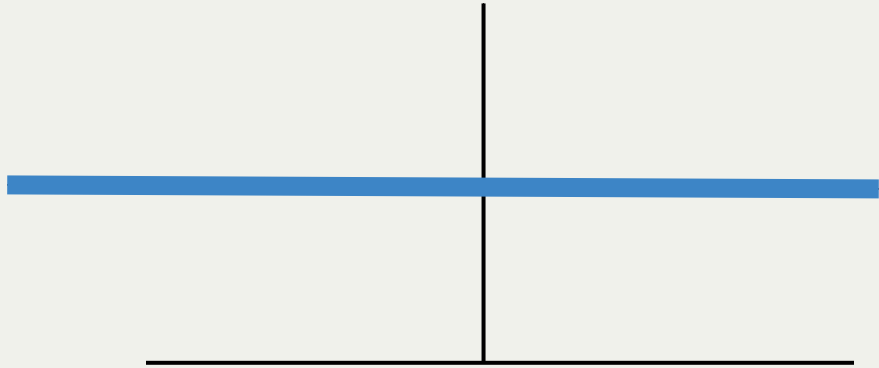
Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

prior: "uninformative prior"

θ model parameters

D data



Bayes theorem

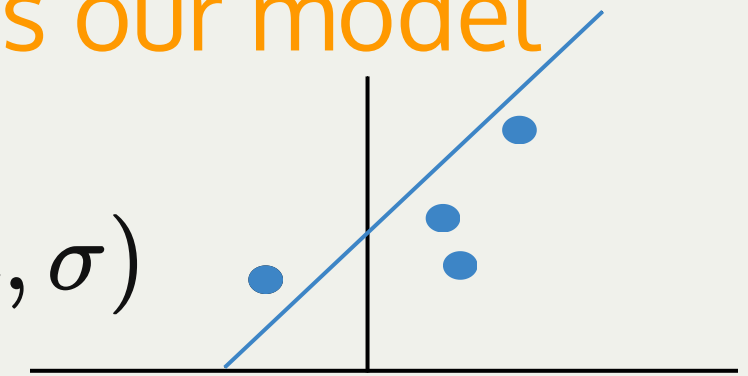
$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

likelihood: this is our model

θ model parameters

D data

$$P(D|\theta) = ax + b + \epsilon; \epsilon \sim N(\mu, \sigma)$$



Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

evidence ????

θ model parameters

D data

Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

evidence ????

θ model parameters

D data

it does not matter if I want to use this for model comparison

Bayes theorem

$$P(\theta_1|\mathrm{D}) = \frac{P(\mathrm{D}|\theta_1)P(\theta_1)}{\cancel{P(\mathrm{D})}} \qquad P(\theta_2|\mathrm{D}) = \frac{P(\mathrm{D}|\theta_2)P(\theta_2)}{\cancel{P(\mathrm{D})}}$$

$$P(\theta|\mathrm{D}) \propto P(\mathrm{D}|\theta)P(\theta)$$

which has the highest posterior probability?

θ model parameters

D data

it does not matter if I want to use this for model comparison

Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

posterior: joint probability distributin of a parameter set (θ, e.g. (m, b)) condition upon some data D and a model hypothesys f

prior: "intellectual" knowledge about the model parameters condition on a model hypothesys *f. This should come from domain knowledge or knowledge of data that is not the dataset under examination*

evidence: marginal likelihood of data under the model

$$P(D|f) = \int P(D|\theta, f)P(\theta|f)d\theta$$

Bayes theorem

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

posterior: joint probability distributin of a parameter set (θ, e.g. (m, b)) condition upon some data D and a model hypothesys f

prior: "intellectual" knowledge about the model parameters condition on a model hypothesys *f. This should come from domain knowledge or knowledge of data that is not the dataset under examination*

evidence: marginal likelihood of data under the model

in reality all of these quantities are conditioned on the shape of the model: this is a model fitting, not a model selection methodology $P(D|f) = \int_{-\infty}^{\infty} P(D|\theta, f)P(\theta|f)d\theta$

the *demarcation* problem in *Bayesian* context

The probability that a belief is true given **new evidence** equals the probability that the belief is true **regardless of that evidence** times the **probability that the evidence is true given that the belief is true** divided by the **probability that the evidence is true regardless** of whether the belief is true.

- Thomas Bayes Essay towards solving a Problem in the Doctrine of Chances (1763)

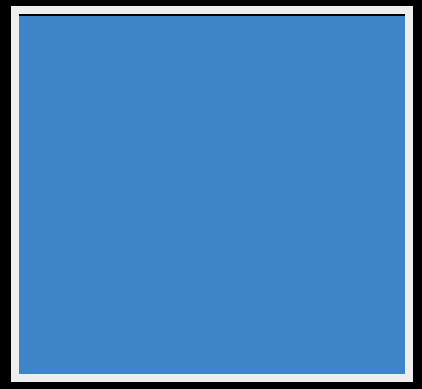
$$p(M|D) = \frac{P(M)\ P(D|M)}{P(D)}$$

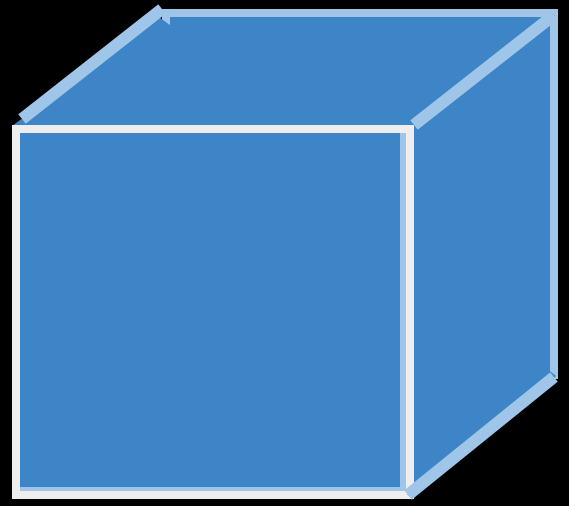
8 scaling laws

quantities that relate by powers

Example:

$L = 1m$





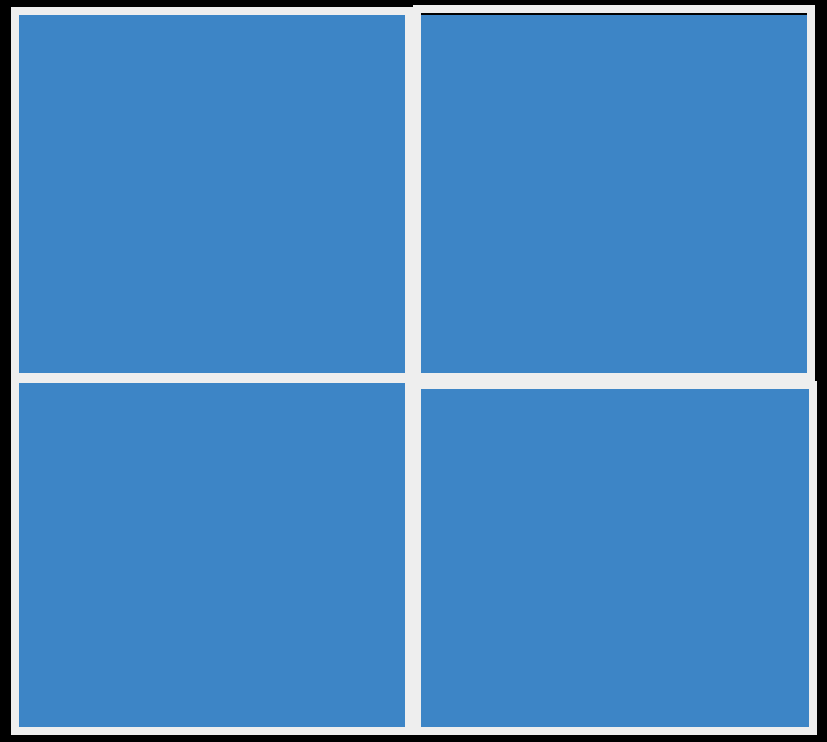
$A = 1m^2$

$V = 1m^3$

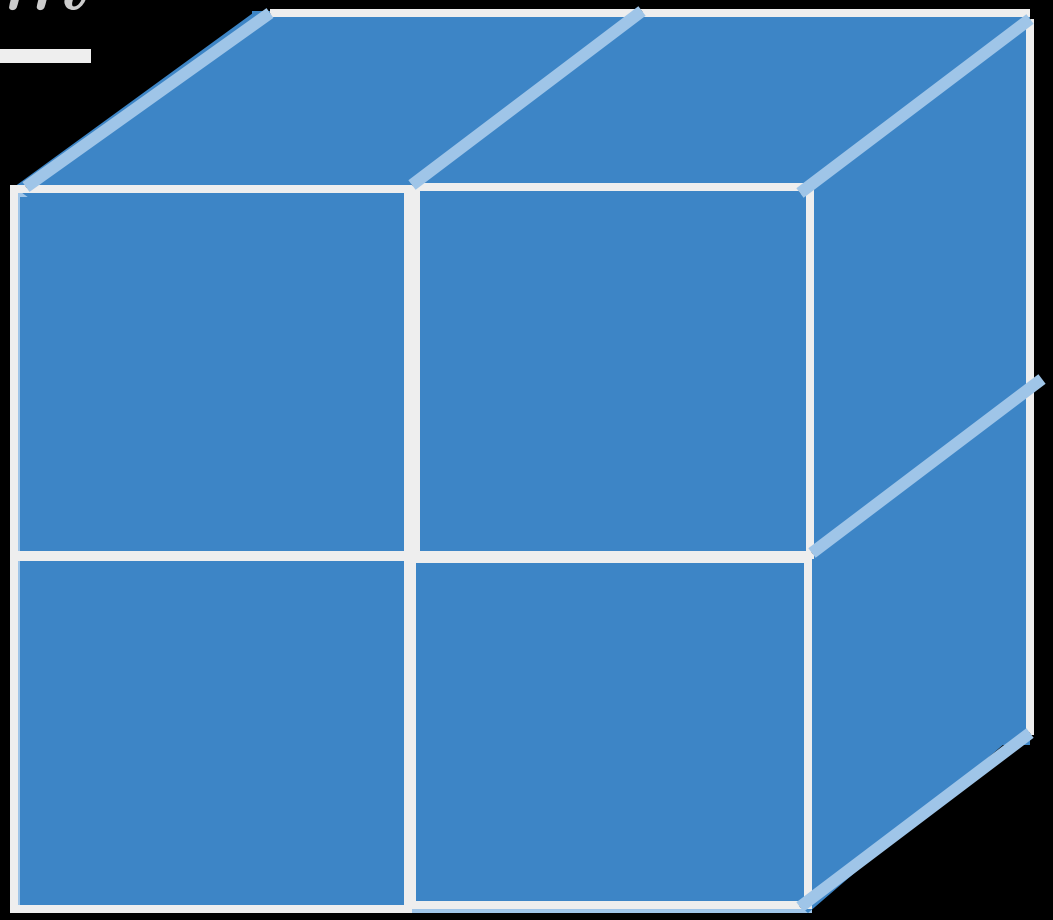
quantities that relate by powers

Example:

$$L = 2\text{x} = 2m$$



$$A = 4\text{x} = 4m^2$$



$$V = 8\text{x} = 8m^3$$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

scaling law: $(\text{ratio of volumes}) = (\text{ratio of lengths})^3$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

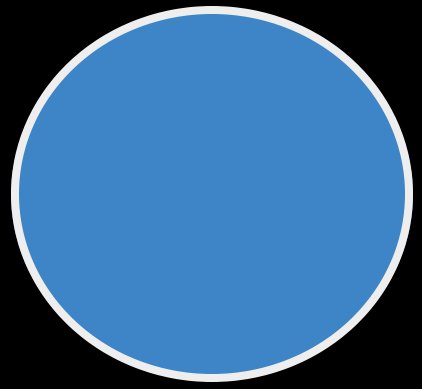
scaling law: $(\text{ratio of volumes}) = (\text{ratio of lengths})^3$

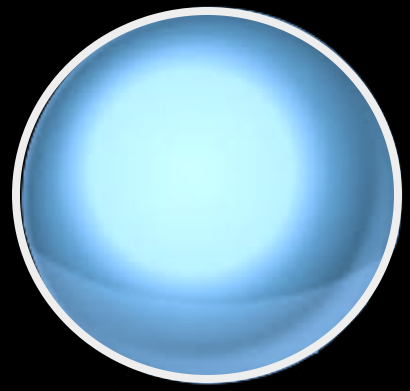
regardless of the shape!

quantities that relate by powers

Example:

$r = 1m$





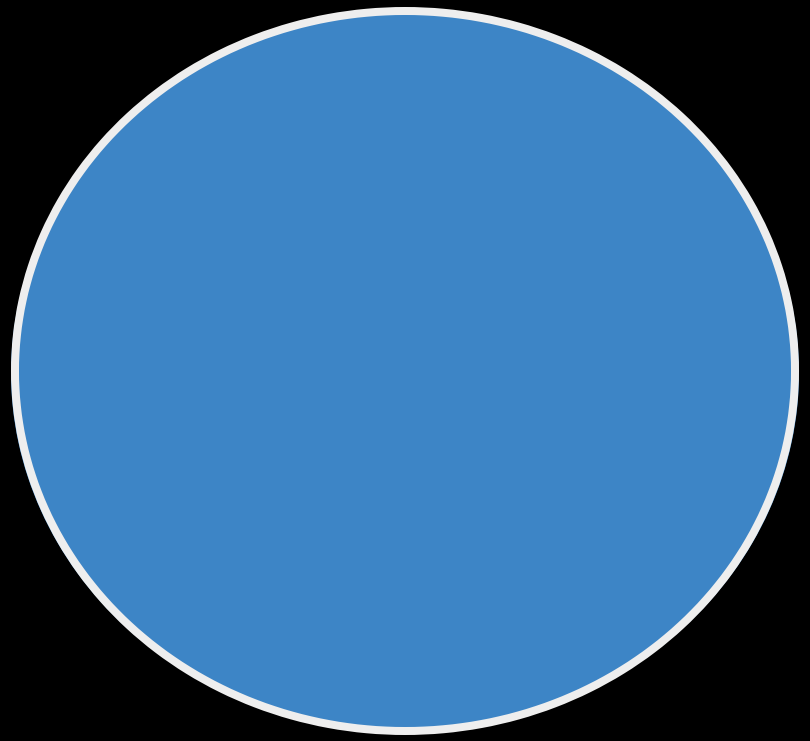
$A = 1m^2$

$V = 1m^3$

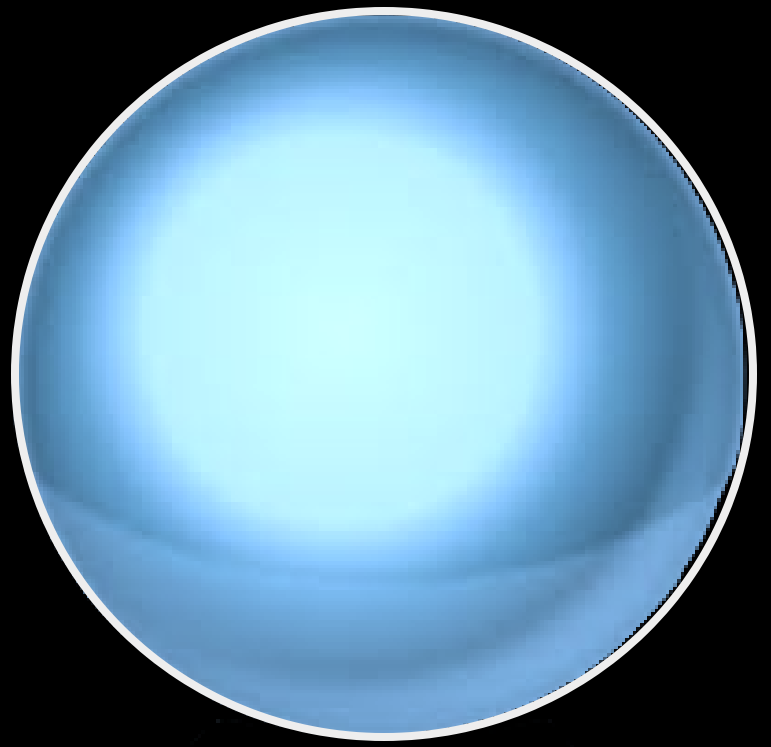
quantities that relate by powers

Example:

$r = 1m$



$V \sim 4\text{x}, V = \text{const}\ r^2$



$V \sim 8\text{x}, V = \text{const}\ r^3$

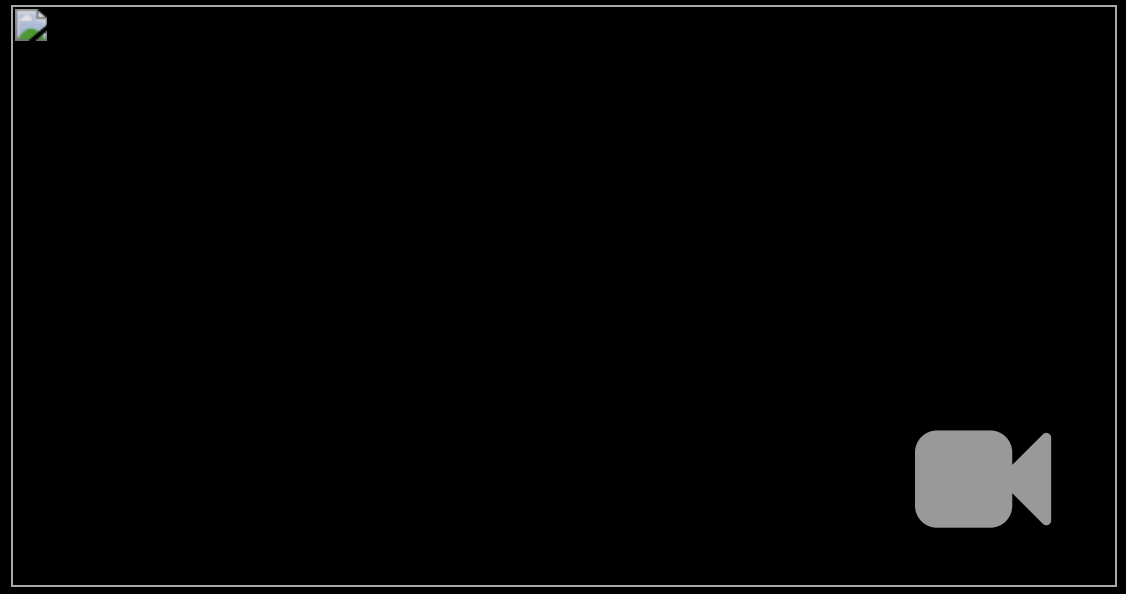
why is this important?

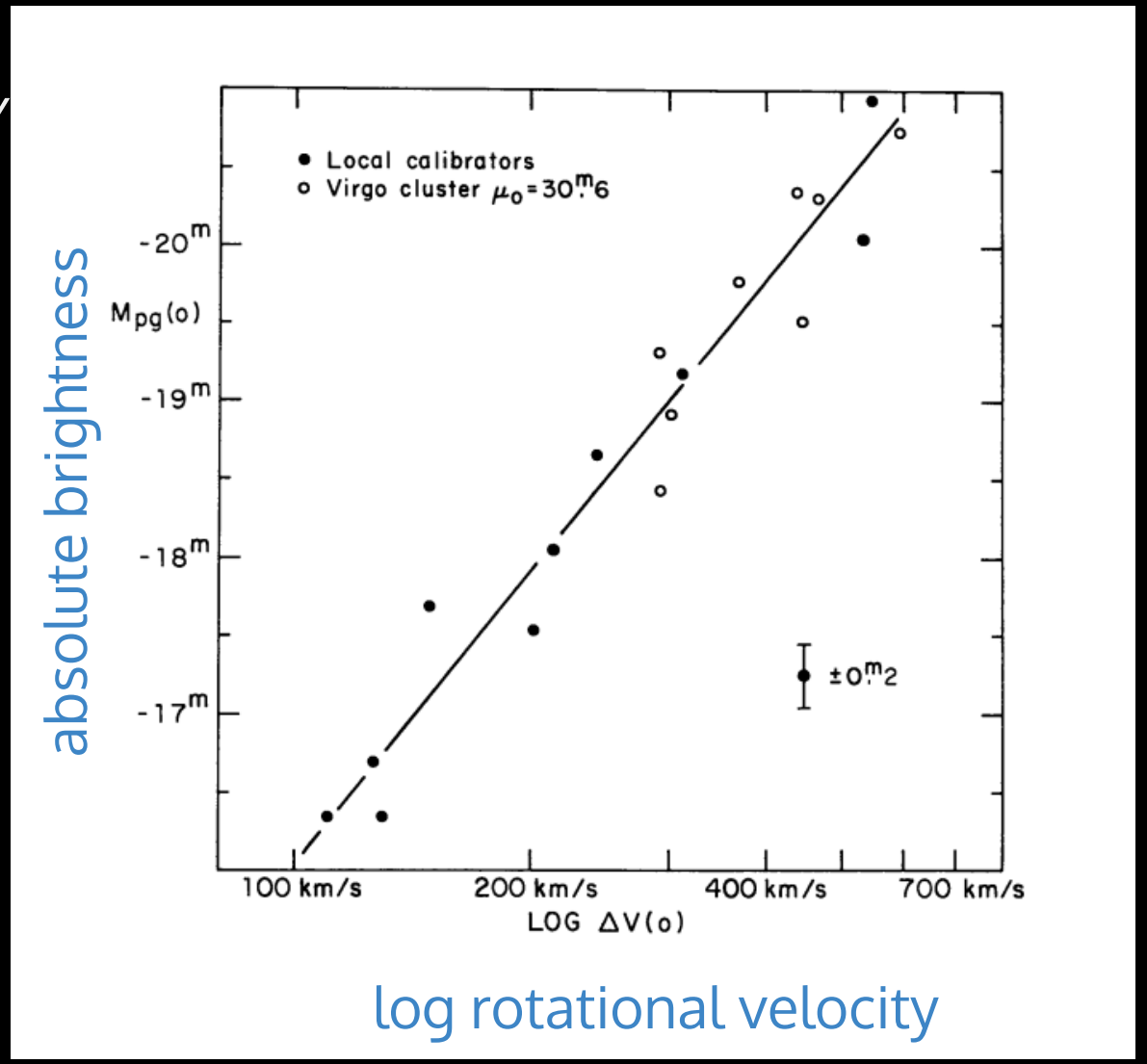
The exsitance of a **scaling** relationship between physical quantities reveals an underlying driving mechanism

Astrophysics

The **Tully–Fisher relation** is an *empirical relationship between the intrinsic luminosity of a spiral galaxy and its torational velocity*

R. Brent **Tully** and J. Richard **Fisher, 1977**
Astronomy and Astrophysics, 54, 661



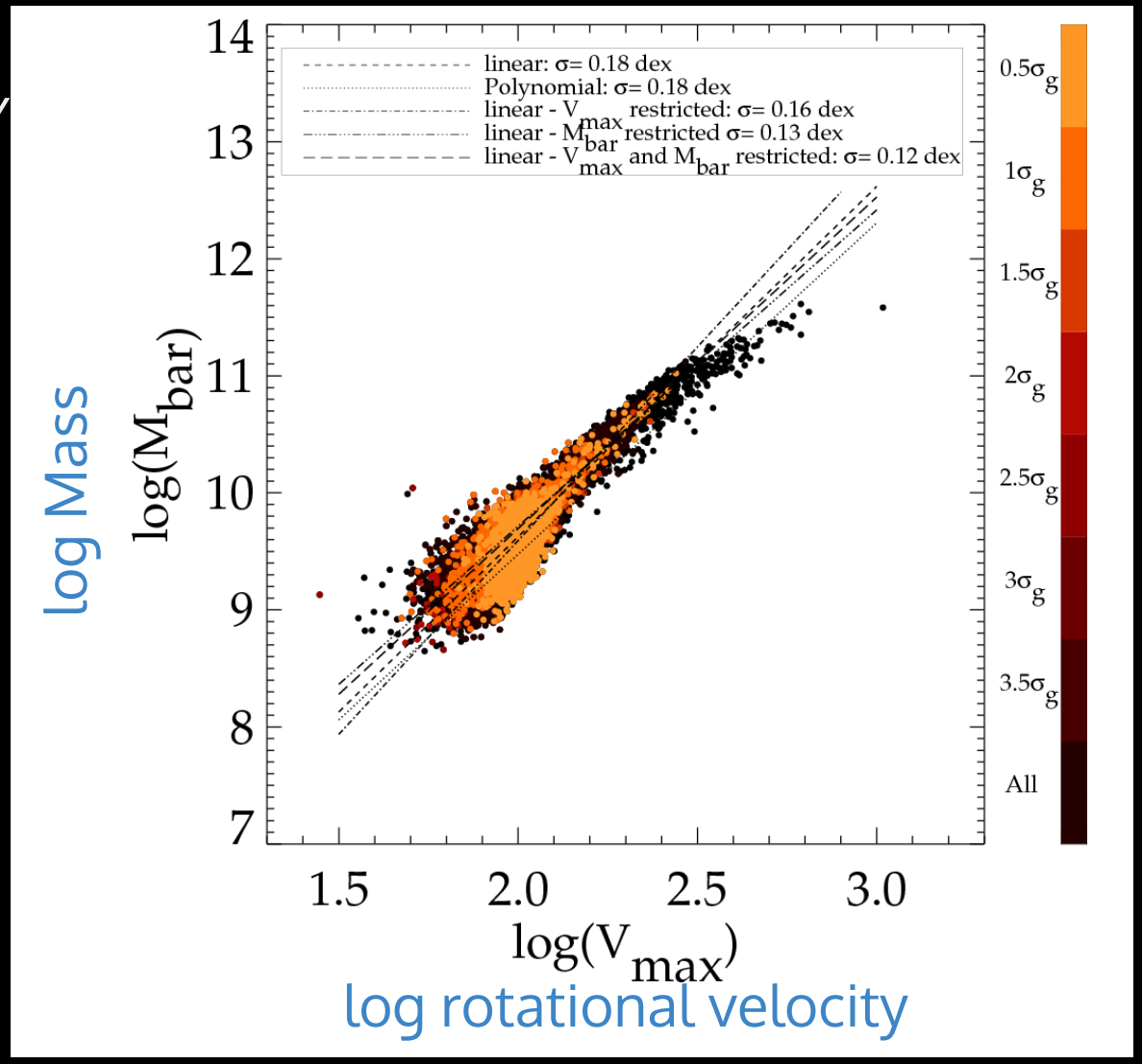


Astrophysics

The **Tully–Fisher relation** is an *empirical relationship between the intrinsic luminosity of a spiral galaxy and its torational velocity*

R. Brent **Tully** and J. Richard **Fisher, 1977**

GRAVITY



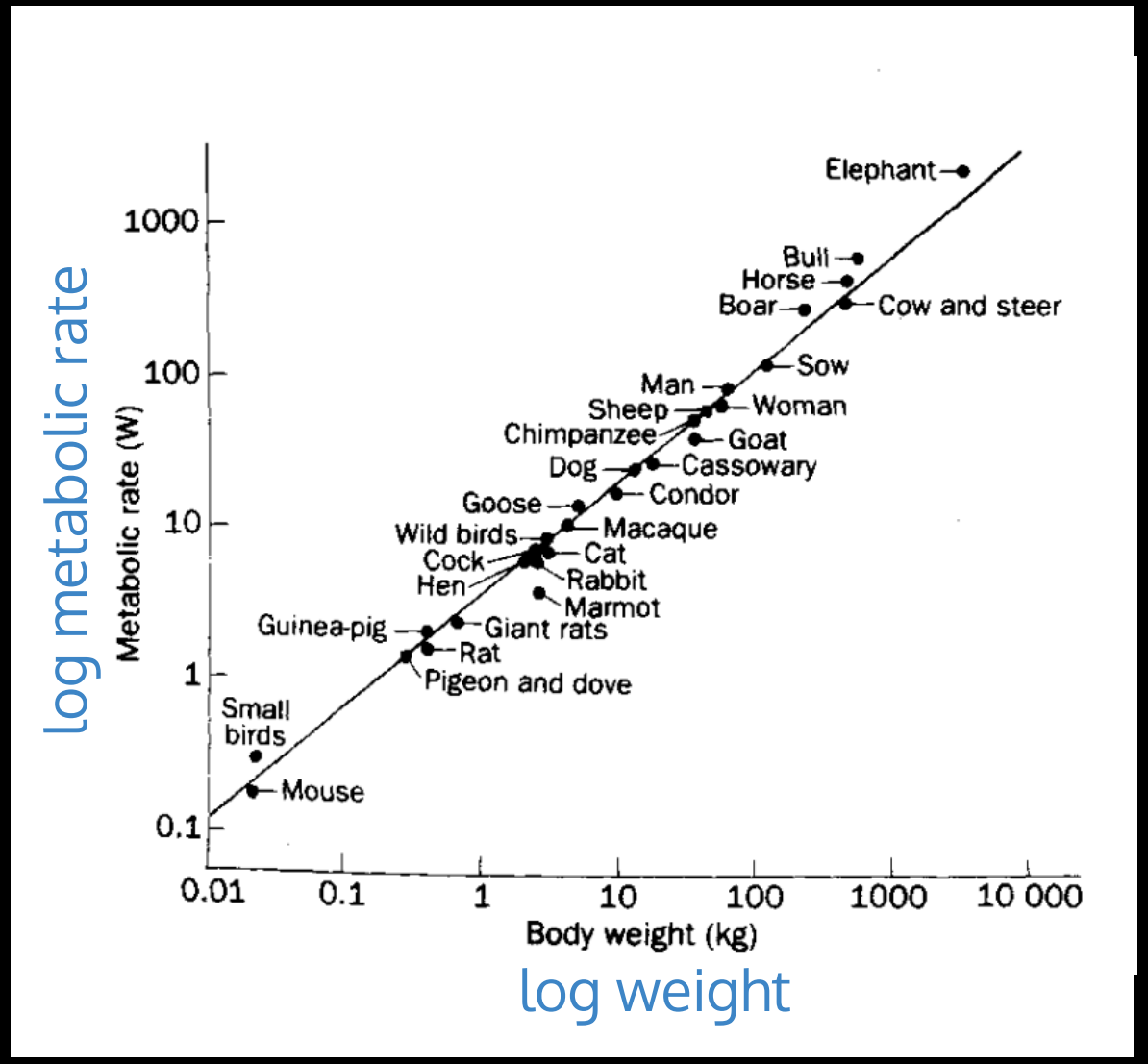
Sorce Jenny *et al.*

Biology

Basal metabolism of mammals (that is, the minimum rate of energy generation of an organism) has long been known to scale empirically as

$$B \propto M^{3/4}$$

KLEIBER, M. (1932). Body size and metabolism. Hilgardia 6, 315



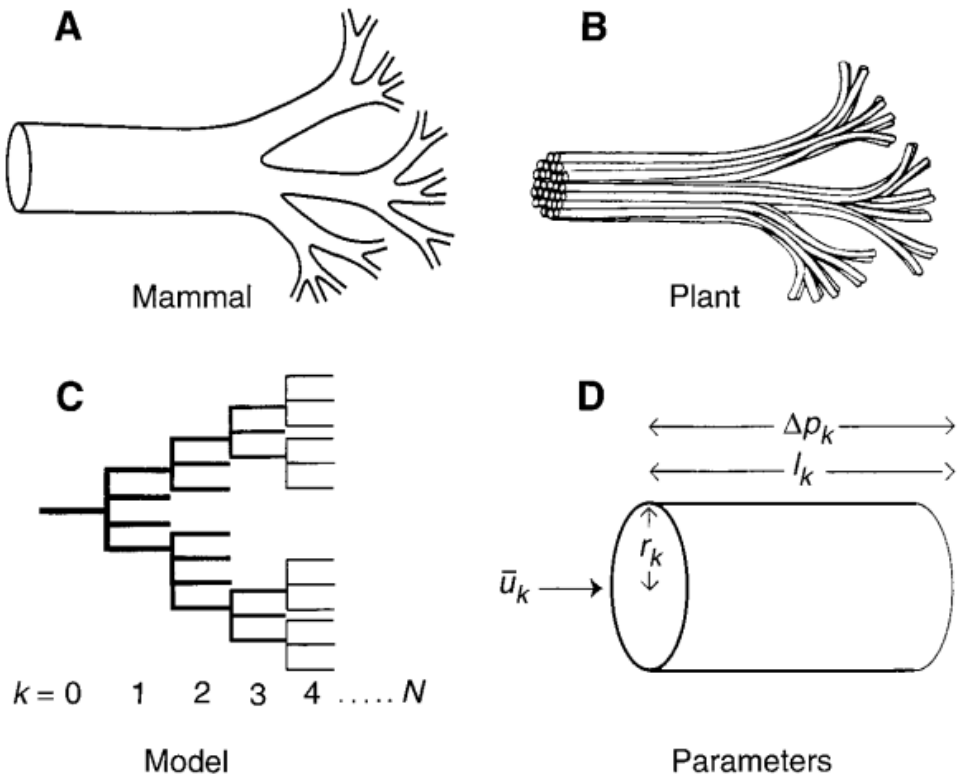
Biology

A general model that describes how essential materials are transported through space-filling fractal networks of branching tubes.

West, Brown, Enquist. 1997 Science

networks

G. West

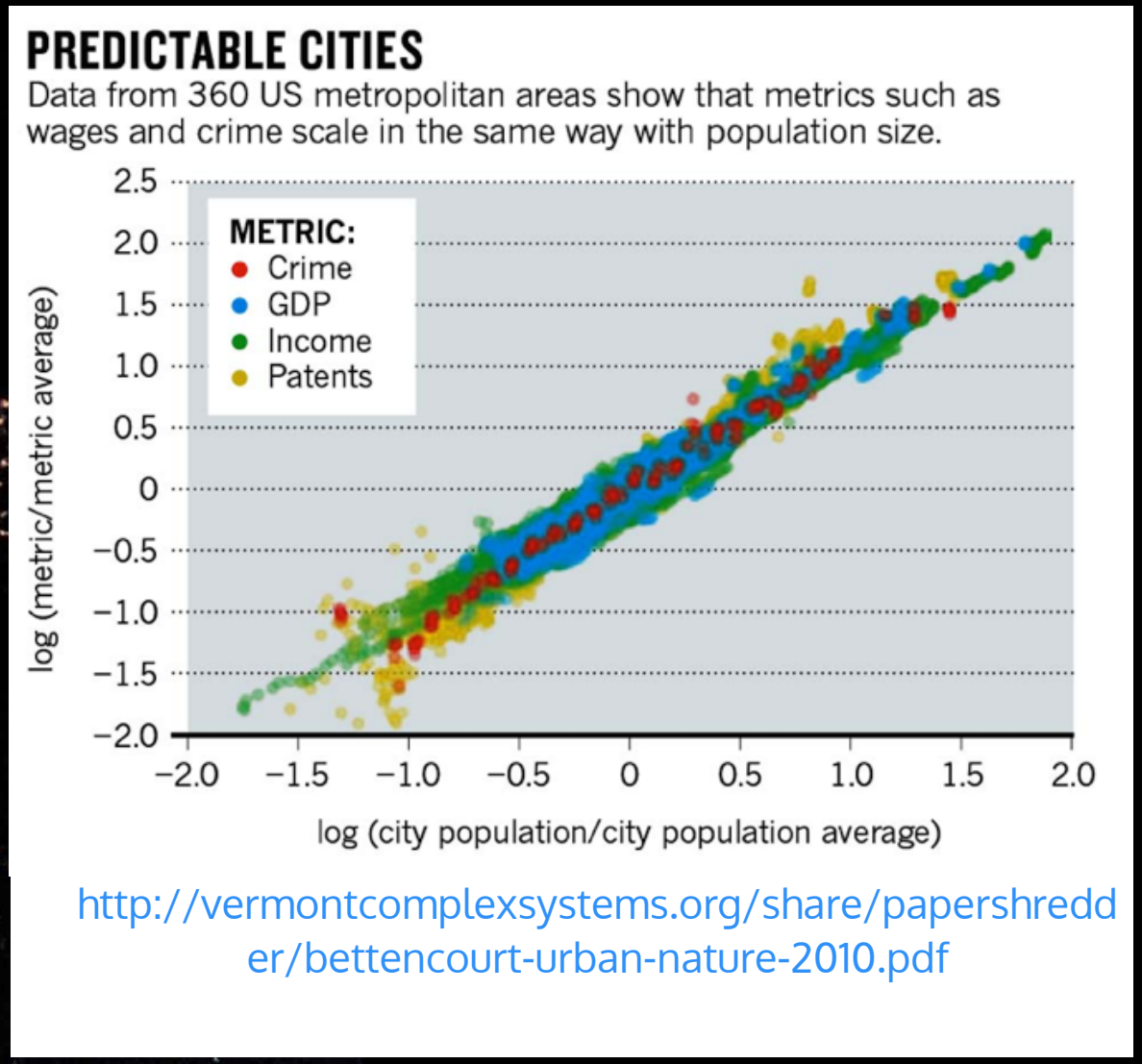


Diagrammatic examples of segments of biological distribution networks

Urban Science

Cities are networks too! And they obey scaling laws on a ridiculus number of parameters!

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Proc. Natl Acad. Sci. USA 104, 7301–7306 (2007)



http://vermontcomplexsystems.org/share/papershredder/bettencourt-urban-nature-2010.pdf

key concepts

descriptive statistics

null hypothesis rejection testing setup

pivotal quantities

Z, t, $\chi 2$, K-S tests

the importance of scaling laws

homework

HW1 : earthquakes and KS test:

reproduce the work of Carrell 2018 using a KS-test to demonstrate the existence of s scaling law in the frequency of earthquakes

https://arxiv.org/pdf/0910.0055.pdf

International Journal of Modern Physics B

https://arxiv.org/pdf/0910.0055.pdf

STATISTICAL TESTS FOR SCALING IN THE INTER-EVENT TIMES OF EARTHQUAKES IN CALIFORNIA

ÁLVARO CORRAL

Centre de Recerca Matemàtica, Edifici Cc, Campus UAB, E-08193 Bellaterra, Barcelona, Spain
ACorral at crm dot es

Received Day Month Year
Revised Day Month Year

We explore in depth the validity of a recently proposed scaling law for earthquake inter-event time distributions in the case of the Southern California, using the waveform cross-correlation catalog of Shearer *et al.* Two statistical tests are used: on the one hand, the standard two-sample Kolmogorov-Smirnov test is in agreement with the scaling of the distributions. On the other hand, the one-sample Kolmogorov-Smirnov statistic complemented with Monte Carlo simulation of the inter-event times, as done by Clauset *et al.*, supports the validity of the gamma distribution as a simple model of the scaling function appearing on the scaling law, for rescaled inter-event times above 0.01, except for the largest data set (magnitude greater than 2). A discussion of these results is provided.

Keywords: Statistical seismology; scaling; goodness-of-fit tests; complex systems.

https://www.ted.com/talks/geoffrey_west_the_surprising_math_of_cities_and_corporations?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

watching

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

David M. Lane et al.

Introduction to Statistics (XVIII)

http://onlinestatbook.com/Online_Statistics_Education.epub

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

Bernard J. T. Jones, Vicent J. Martínez, Enn Saar, and Virginia Trimble

Scaling laws in physics

https://ned.ipac.caltech.edu/level5/March04/Jones/Jones1_3.html

Bettencourt , Strumsky, West

Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013541

resources