data science for (physical) scientists 3

scaling laws

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

1 P(physics | data)

2 NHRT

p-values

z-test

3 comparing distributions

Z, t, χ2, ks-test

KL divergence

this slide deck

https://slides.com/federicabianco/dsps_3

recap 1

Guiding principle of science practice

- *Theories* should be *falsifiable* (= make predictions)
- *Analysis* should be *reproducible* (share result, share raw data, share code to get result from raw data)

recap 2

probability

- *Frequentist* interpretation: fraction of occurrence
- *Bayesian* interpretation: degree of believe that it will happen
- Basic probability algebra rules

recap 3

statistics

- links between samples (observations) and populations (general rules)
- common distributions: binomial, Poisson, Gaussian, $\chi 2$
- *Descriptive statistics:* central tendency, variance, symmetry
- Central limit theorem

recap 4

physics

- thermodynamics: the first deliberate example of application of statistics to physics

if we know the properties of the *micro* system statistically we can predict the *macro* system deterministically

recap

5

descriptive statistics:

we summarize the proprties of a distribution

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

TAXONOMY

central tendency: mean, median, mode

spread : variance, interquantile range

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,

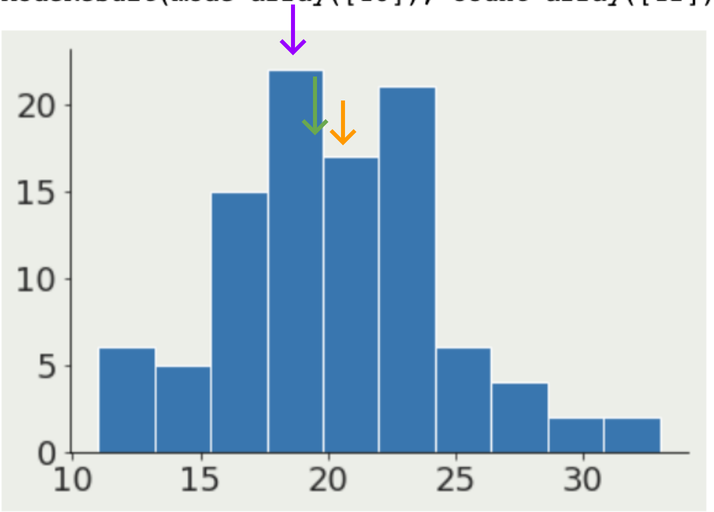
50% to the right

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

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

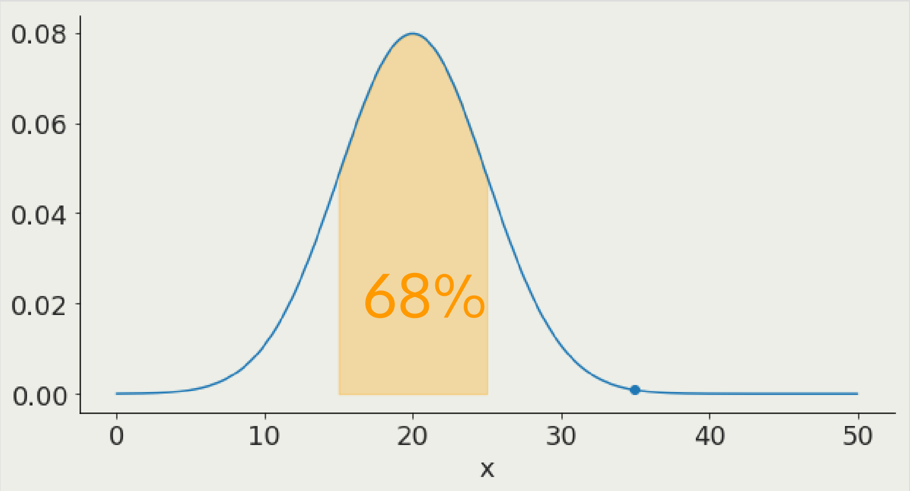
variance: n=2 $\mathrm{Var}(X) = \mathbf{E}\left[(X-\mu)^2\right].$

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

$$Var = \sigma^2$$

Gaussian distribution:

1σ contains 68% of the distribution



descriptive statistics:

we summarize the proprties of a distribution

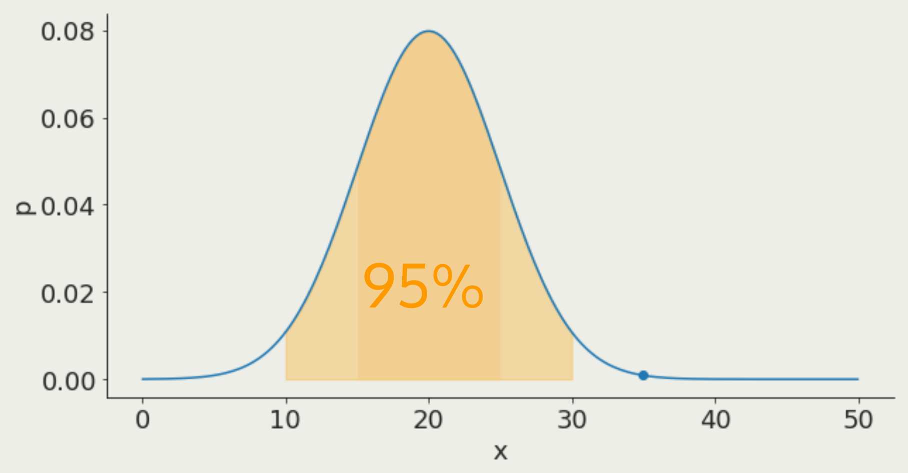
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

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

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

Gaussian distribution:

2σ contains 95% of the distribution



descriptive statistics:

we summarize the proprties of a distribution

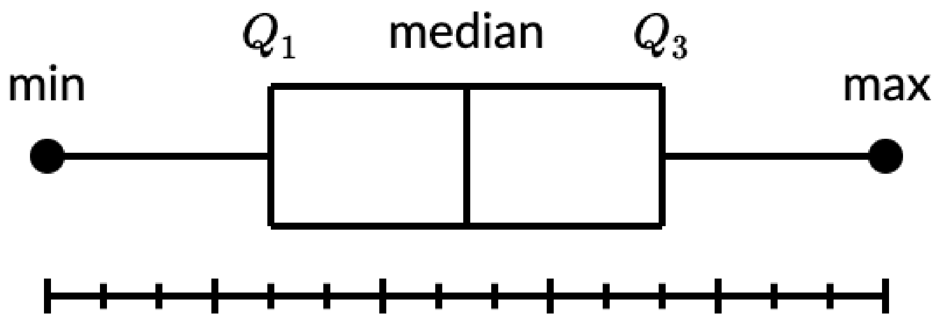
$$\mu_n = \int_{-\infty}^{\infty} (x-c)^n \, f(x) \, \mathrm{d}x.$$

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

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

We also use *interquantile range*

where are the limits within which X% of the distribution is contained



https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

descriptive statistics:

we summarize the proprties of a distribution

PROS AND CONS??

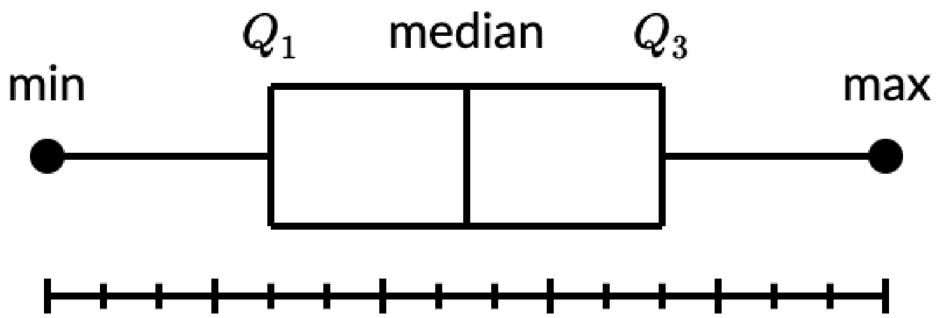
$$\mu_n = \int_{-\infty}^{\infty} (x-c)^n \, f(x)\, \mathrm{d}x.$$

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

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

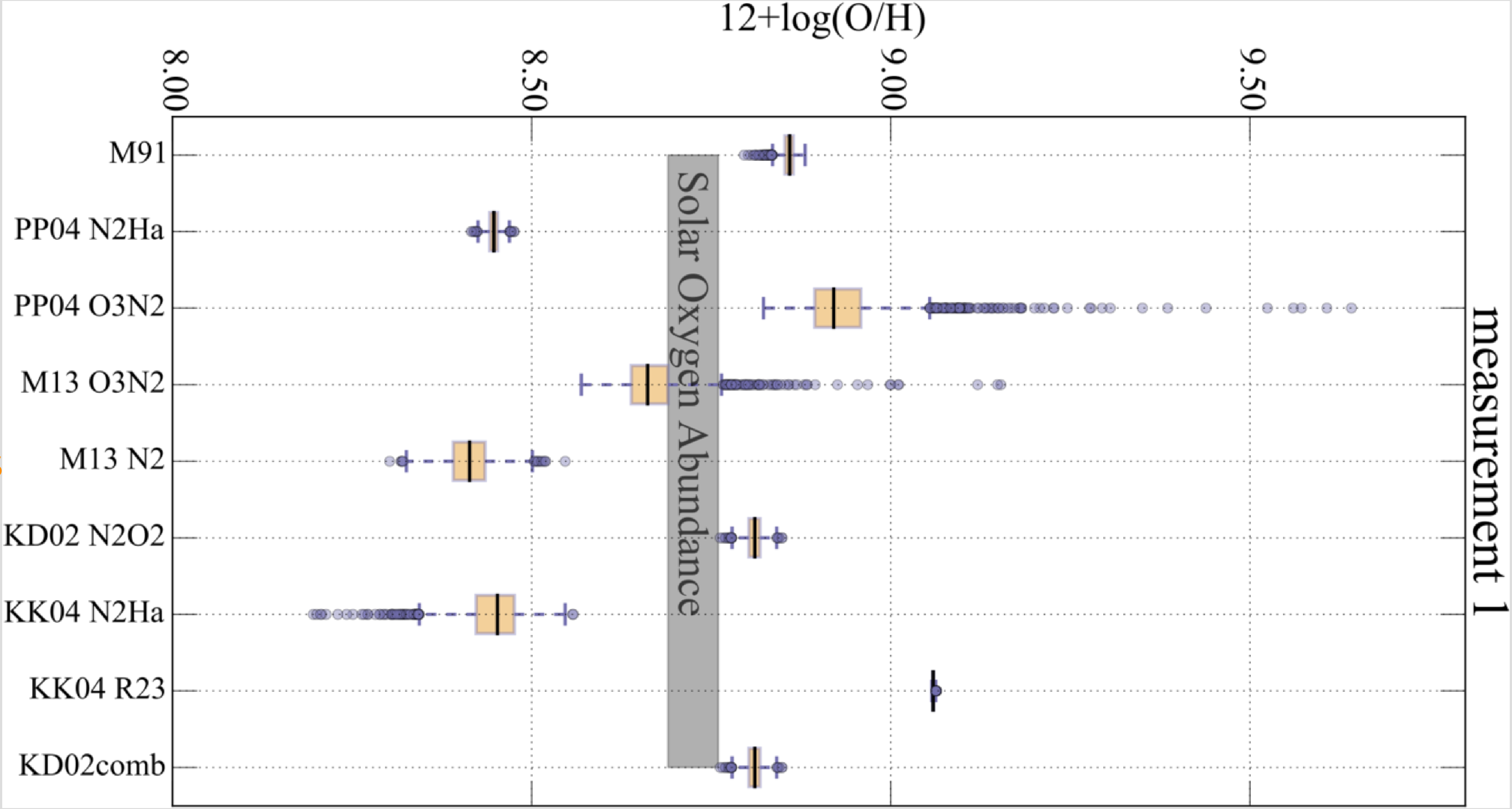
We also use *interquantile range*

where are the limits within which X% of the distribution is contained



https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

descriptive statistics:



S

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.

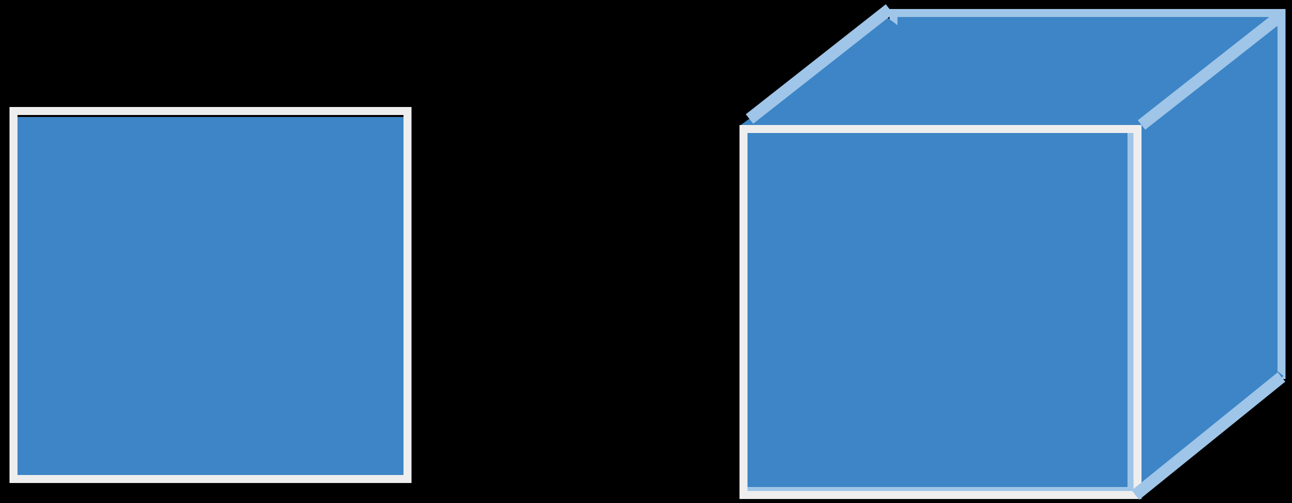
1

scaling laws

quantities that relate by powers

Example:

$$L = 1m$$

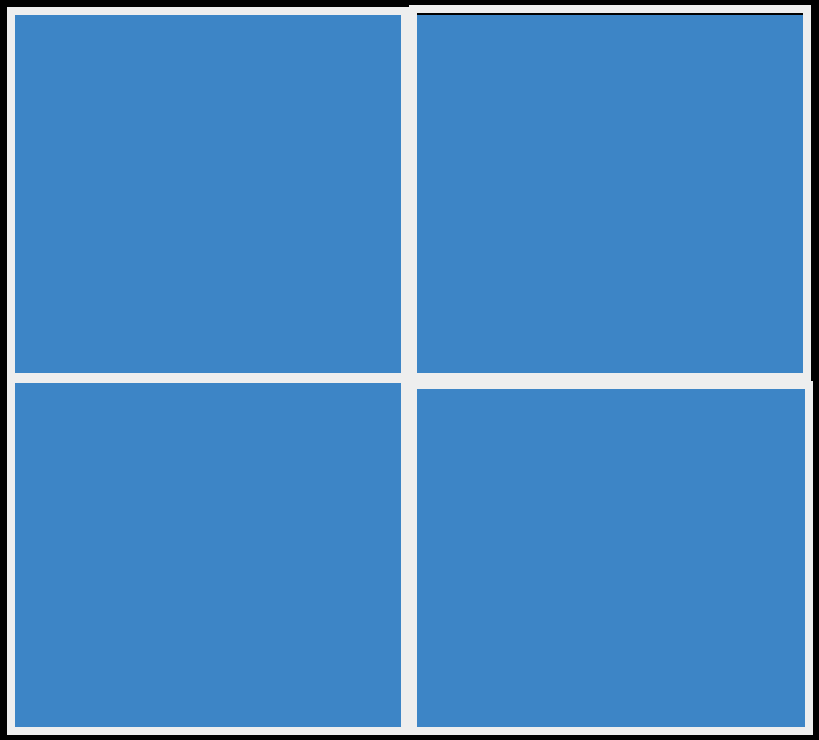


$$A = 1m^2 \qquad V = 1m^3$$

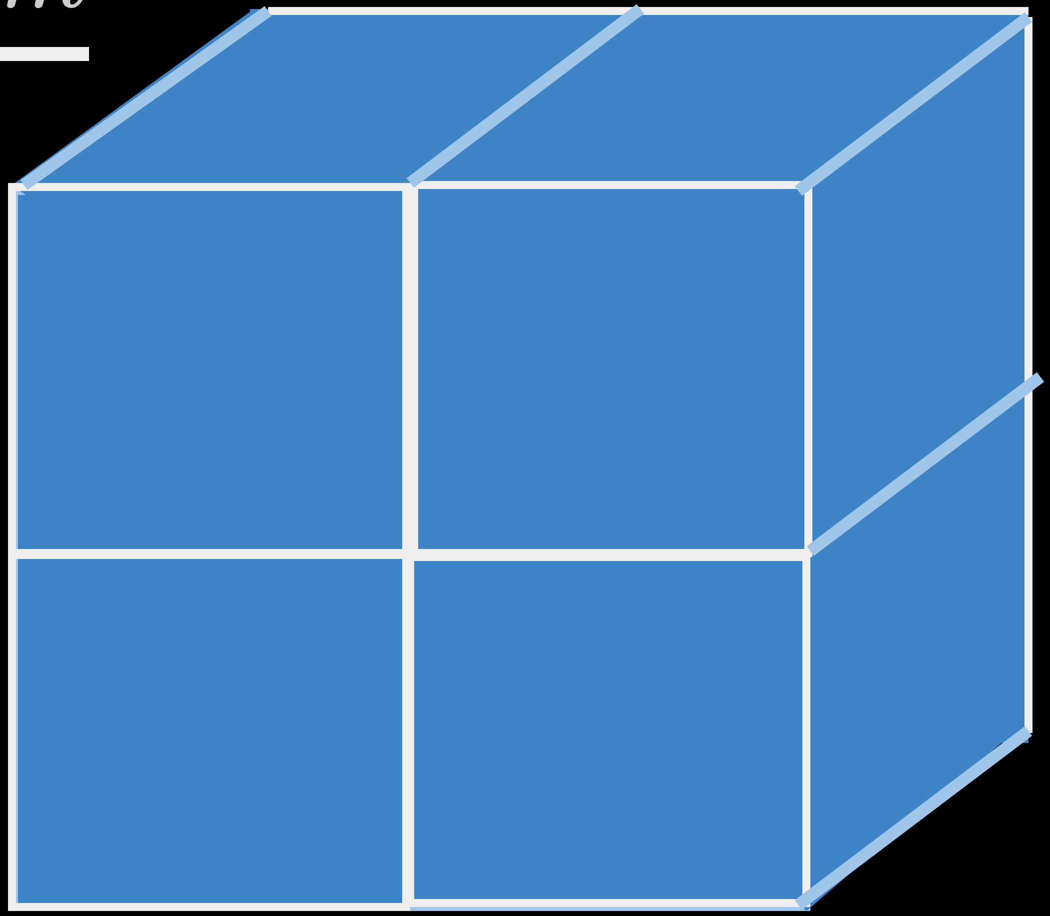
quantities that relate by powers

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

Example:



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



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

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

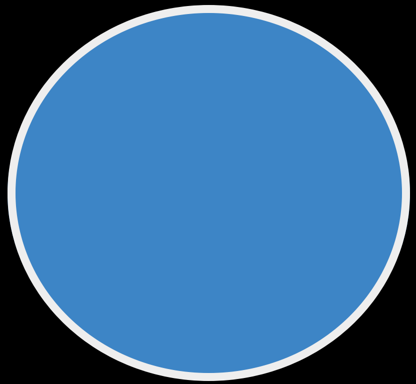
scaling law: $$(\text{ratio of volumes}) = (\text{ratio of lengths})^3$$

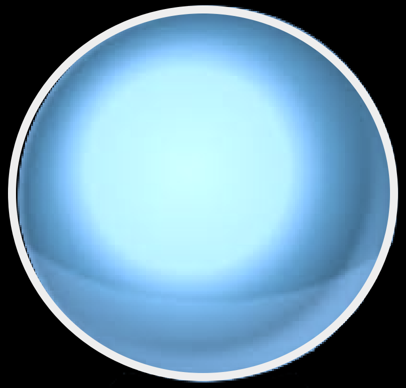
regardless of the shape!

quantities that relate by powers

Example:

$$r = 1m$$



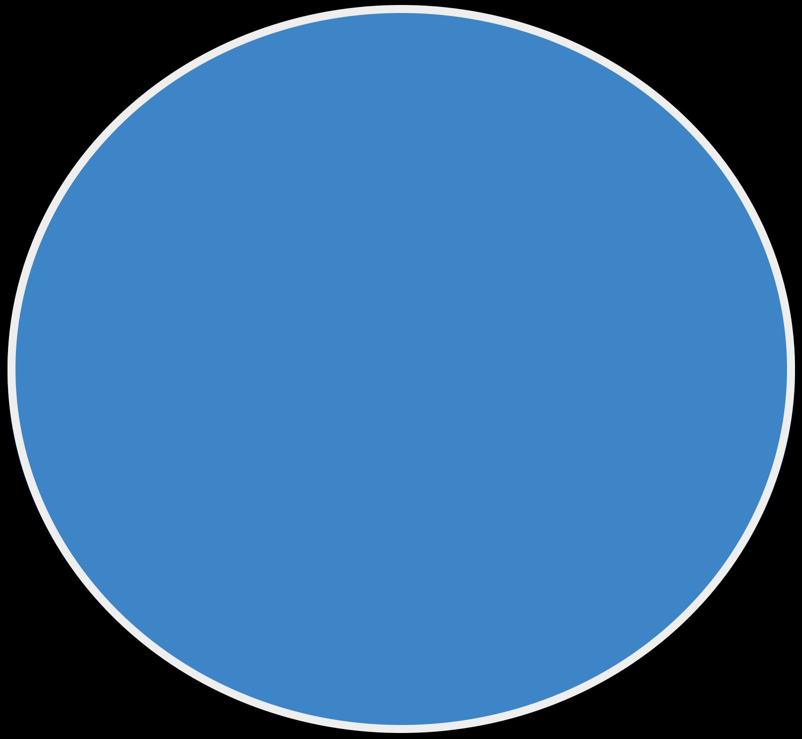


$$A = 1m^2 \qquad V = 1m^3$$

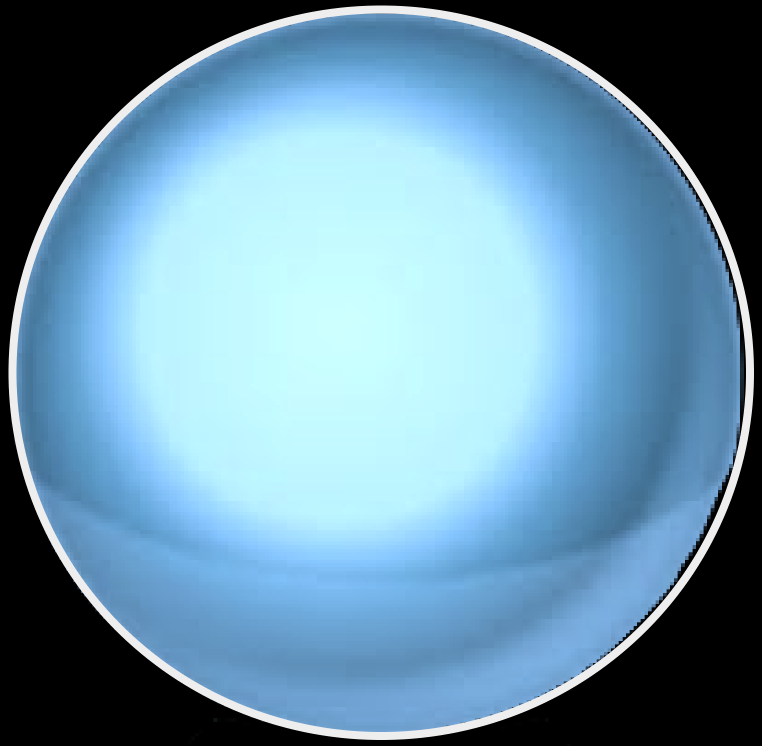
quantities that relate by powers

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?

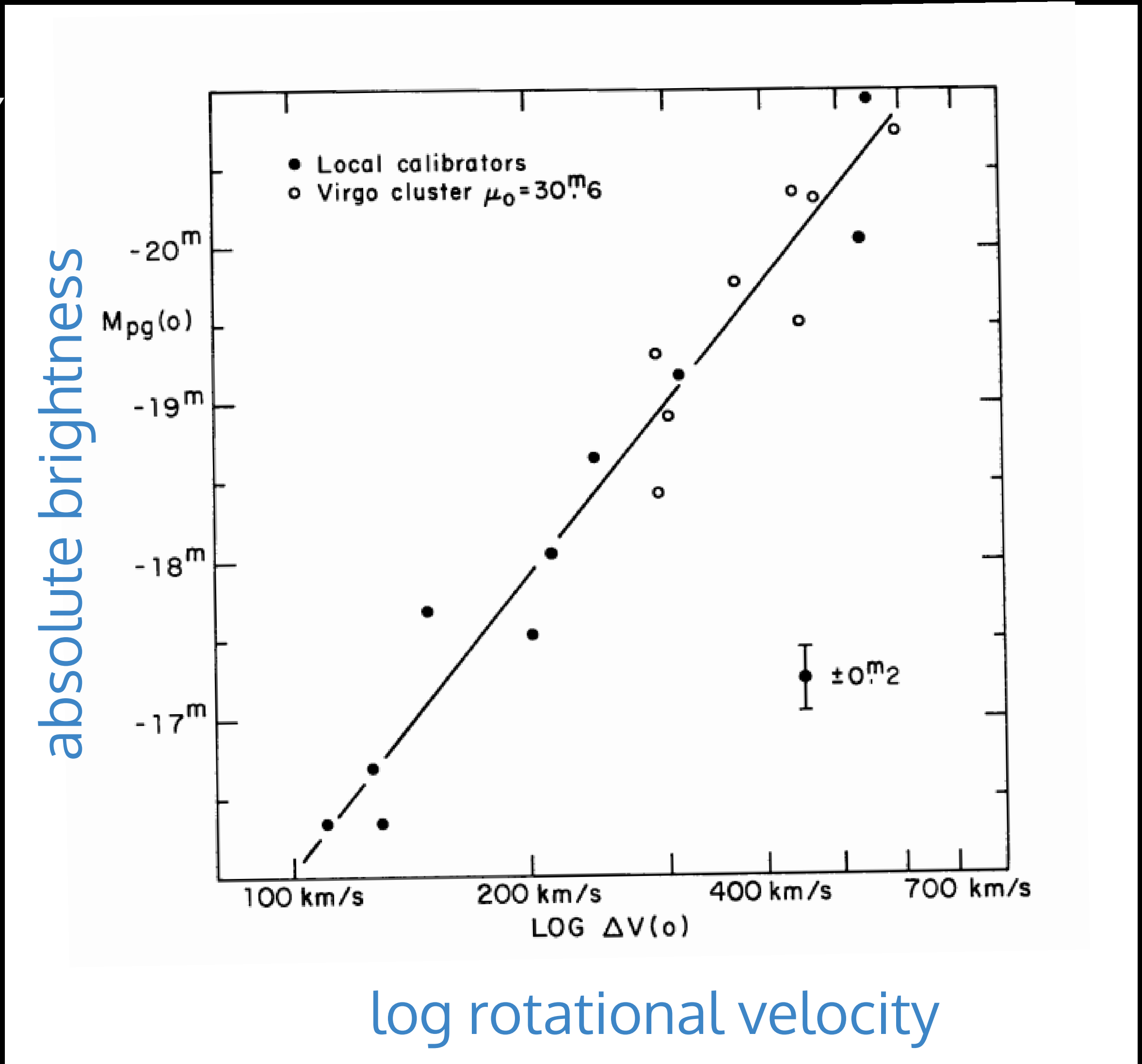
The exsitance of a **scaling** relationship between physical quantities points to an underlying driving mechanism

Astrophysics

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

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

https://www.youtube.com/embed/2KYb_l8wr0c?enablejsapi=1



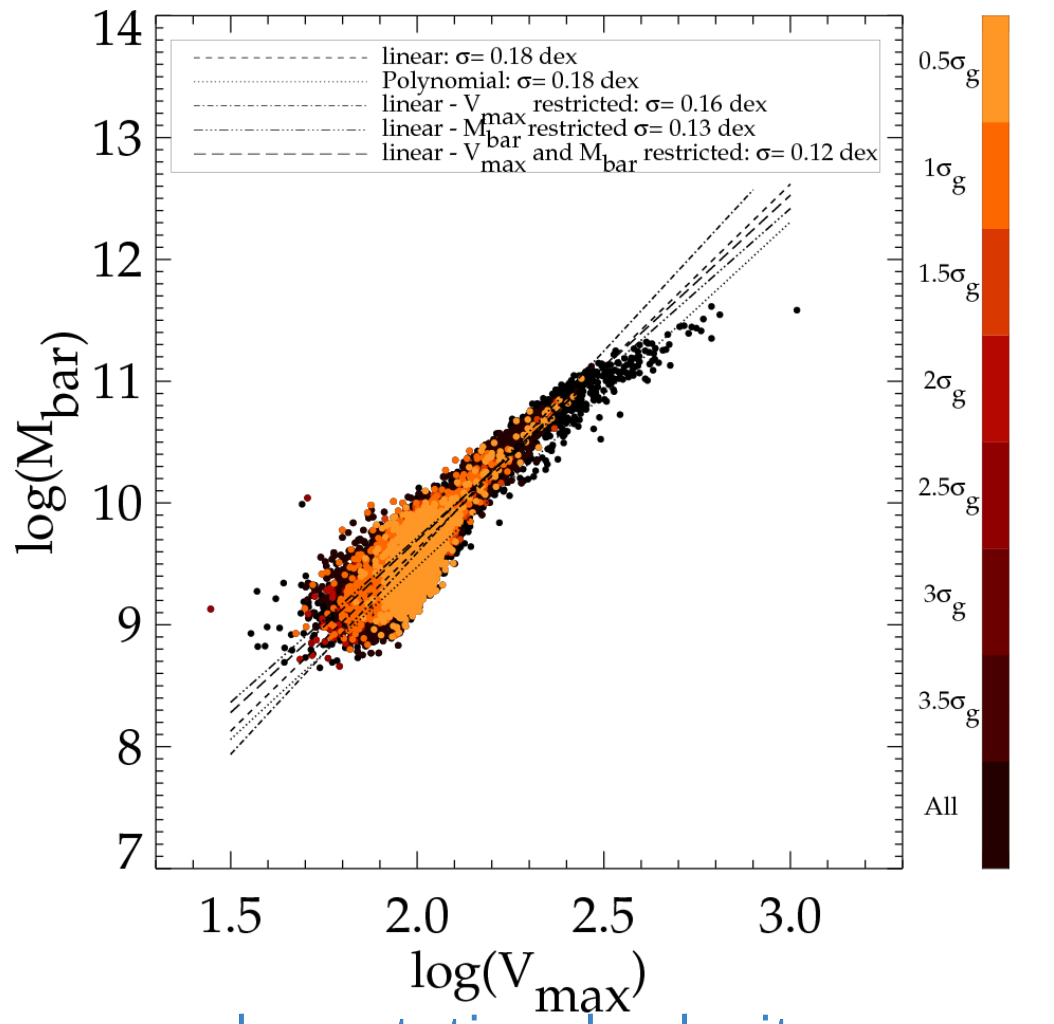
Astrophysics

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

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

GRAVITY

log Mass



log rotational velocity

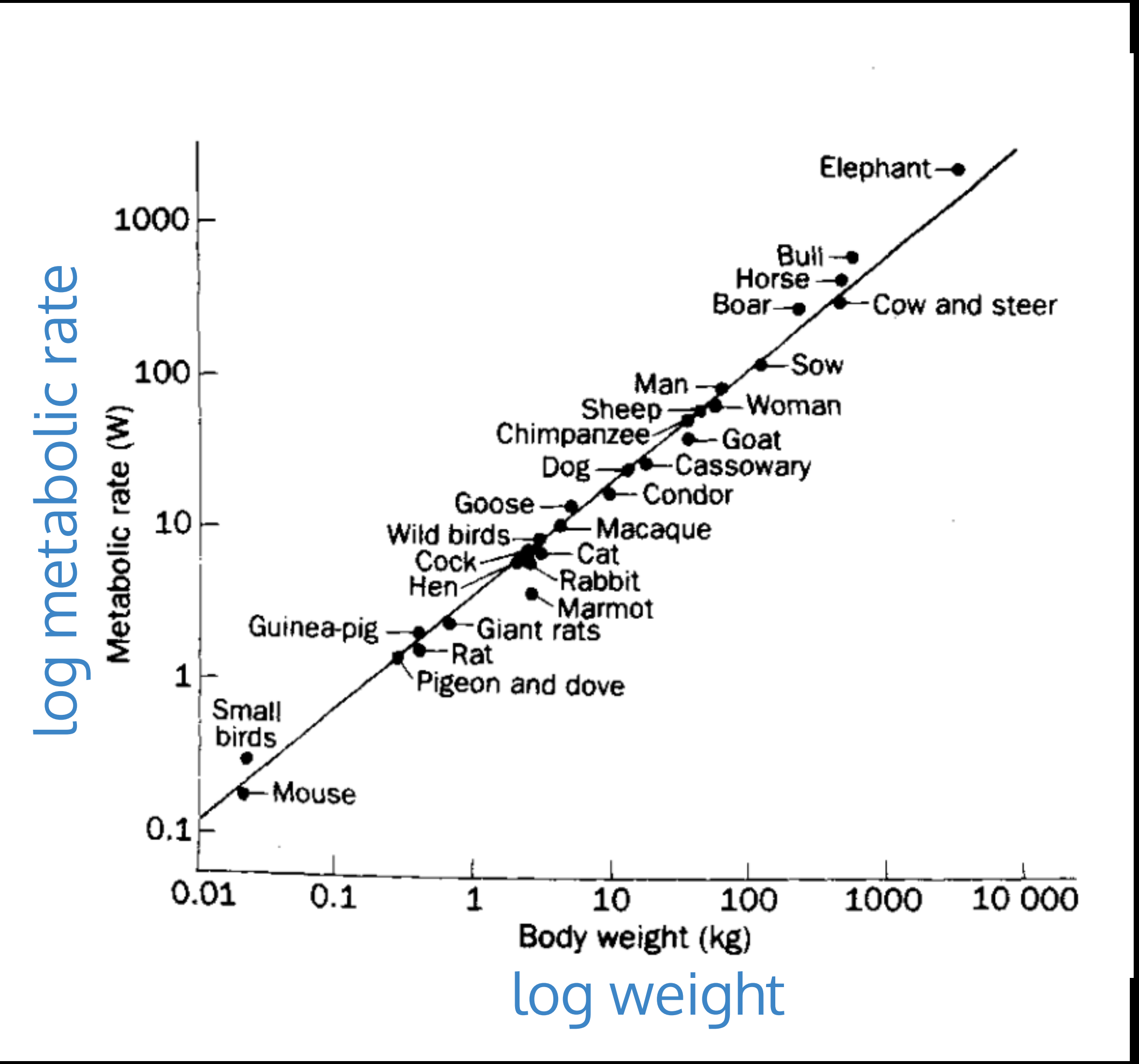
Sorce Jenny *et al.*

Biology

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

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

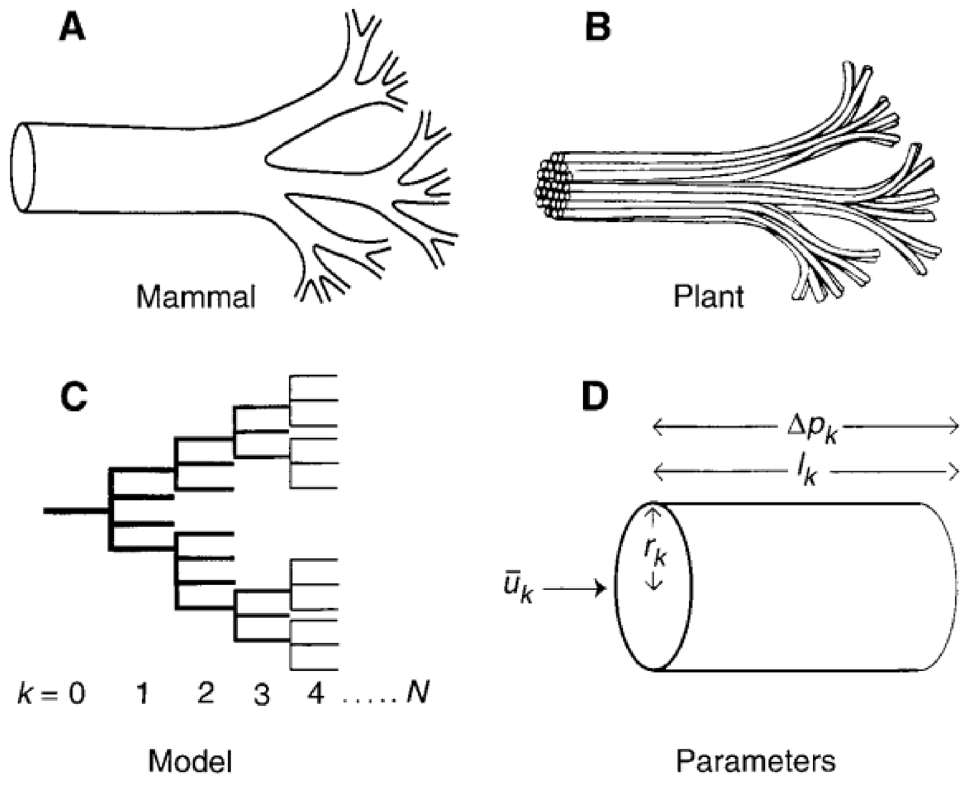
KLEIBER, M. (1932). Body size and metabolism. Hilgardia 6, 315



Biology

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

West, Brown, Enquist. 1997 Science



Diagrammatic examples of segments of biological distribution networks

networks

G. West

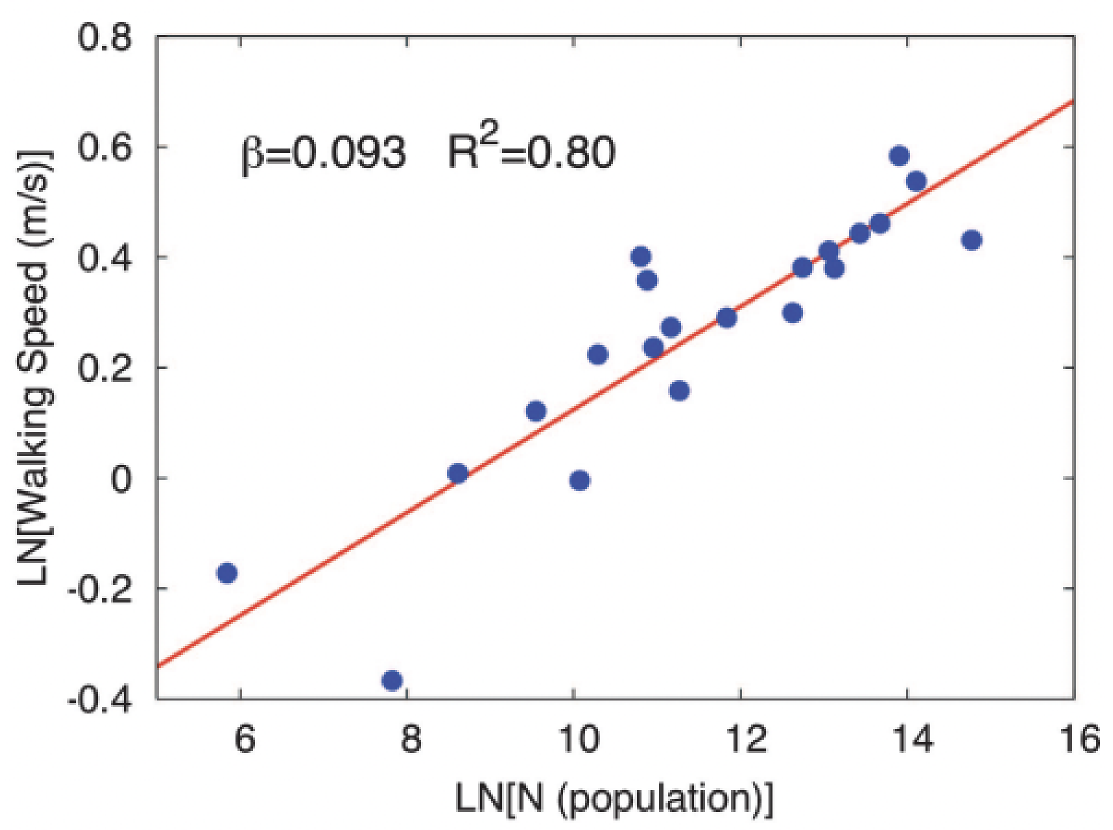
Urban Science

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



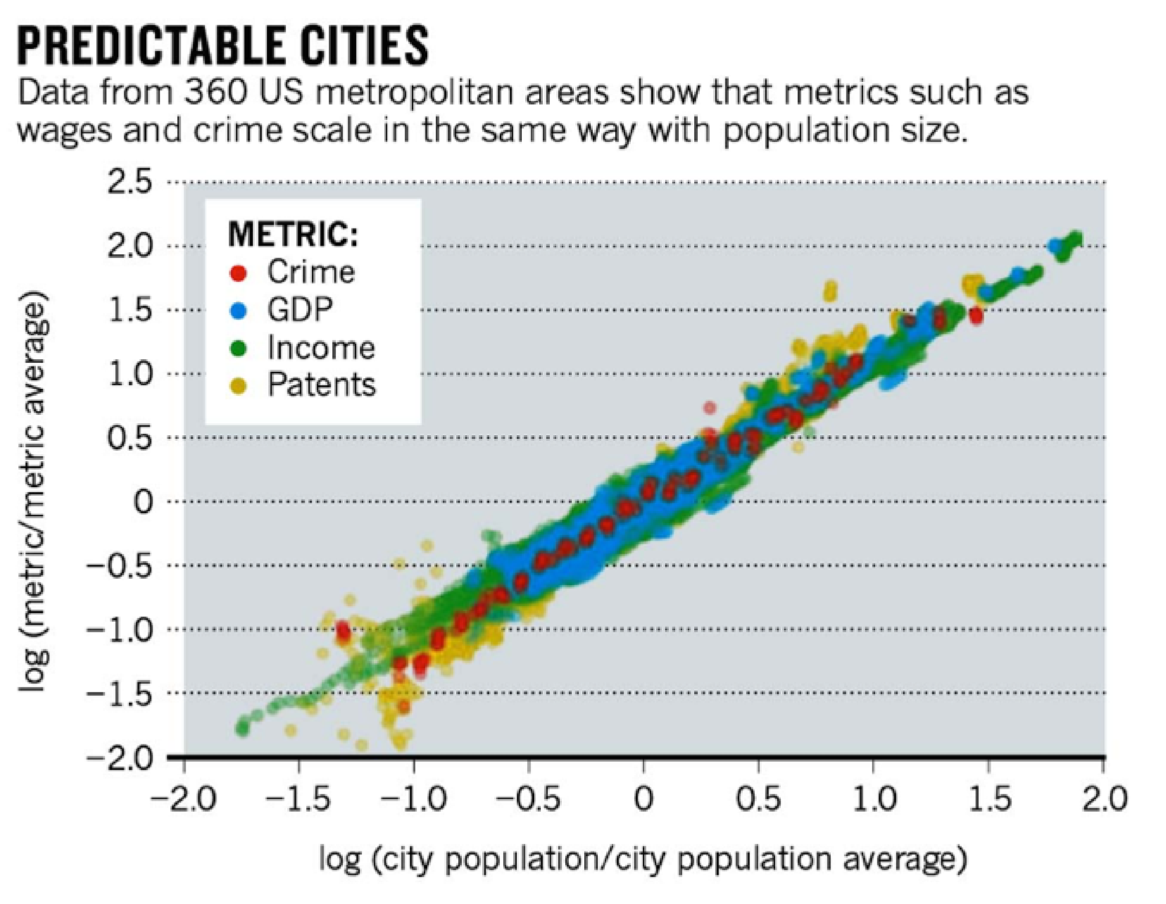
Diagrammatic examples of segments of biological distribution networks

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)

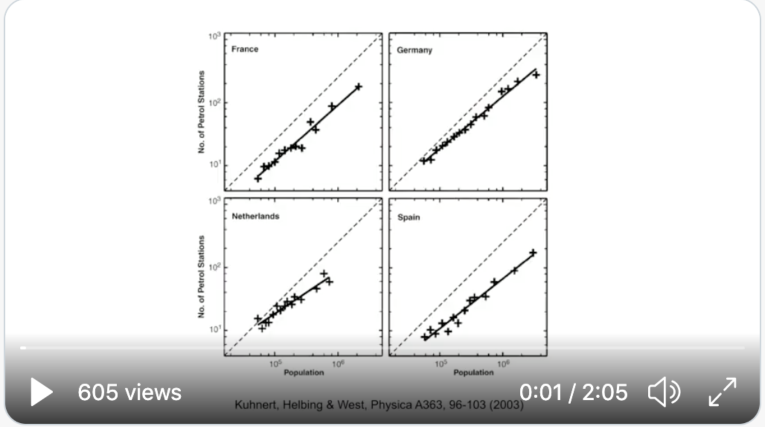
G. West



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

Patrick Sharkey @patrick_sharkey · Sep 22

Yesterday my class on urban inequality discussed Geoffrey West's ideas about "universal" laws of scaling in cities. A few questions and comments came up

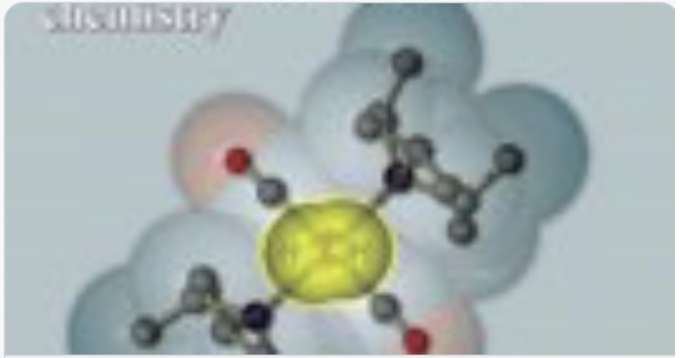


1 3 10

Patrick Sharkey @patrick_sharkey · Sep 22

Replying to @patrick_sharkey

I think the argument is fascinating. But the first q is: is his claim right? His data sources were all over the place, some pretty shaky (e.g. pace of movement!). Has there been work testing his claims?



Growth, innovation, scaling, and the pace of life in cities

Humanity has just crossed a major landmark in its history with the majority of people now living in cities. Cities have long been known to...

pnas.org

1

Patrick Sharkey @patrick_sharkey · Sep 22

For violent crime, eg, the largest cities in the US no longer have the highest levels of violence. Big cities mobilized to confront violence (thru brute force policing, mass incarceration, large-scale surveillance, private security, and *community mobilization*) - violence fell.

1

Patrick Sharkey @patrick_sharkey · Sep 22

Second, I would argue that his "laws" miss the point: growth itself, and the *distribution* of resources, people, disease, crime etc are the crucial phenomena that urbanists care about.

1

Patrick Sharkey @patrick_sharkey · Sep 22

West takes growth for granted - but why do some cities grown and others don't? Which actors play a central role? this is where power, politics, movements, and conflict drive change, where "growth machine" theory enters in, where "homevoters" become relevant.

1

Patrick Sharkey @patrick_sharkey · Sep 22

We read pieces of Urban Fortunes as well, to make the point that the very thing that West takes for granted - growth - is the thing that planners, geographers, political scientists, sociologists consider most crucial to understanding urban inequality and change.

1

Patrick Sharkey @patrick_sharkey · Sep 22

Our key questions are: Where does growth happen? What does it look like? How are resources and people distributed across communities? Which actors and institutions play a central role?

1

Patrick Sharkey @patrick_sharkey · Sep 22

My conclusion: When one tries to ignore all of this and argue for universal laws of all cities, one misses virtually everything important about how cities work and how urban inequality emerges and changes. There's nothing natural, inherent, or inevitable about urban development.

1 5

https://twitter.com/patrick_sharkey/status/1308417626574655488?s=20 https://www.ted.com/talks/geoffrey_west_the_surprising_math_of_cities_and_corporations?language=en

viewing

https://www.youtube.com/embed/XyCY6mjWOPc?enablejsapi=1

coding time!

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

we will start the homework together

2

the scientific method in a probabilistic context (already seen…)

$$p(\text{physics} \mid \text{data})$$

formulate the Null as the comprehensive opposite of your theory

model → **prediction**

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

this has a low probability of happening

data

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

falsifies alternative → **old model still holds**

low probability event happened

But instead of verifying a theory we want to falsify one

model → **prediction**

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

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 new model is false

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 new model is false

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)

3 Null hypothesis rejection testing

Null
Hypothesis
Rejection
Testing

1

formulate your prediction

Null Hypothesis

Null
Hypothesis
Rejection
Testing

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

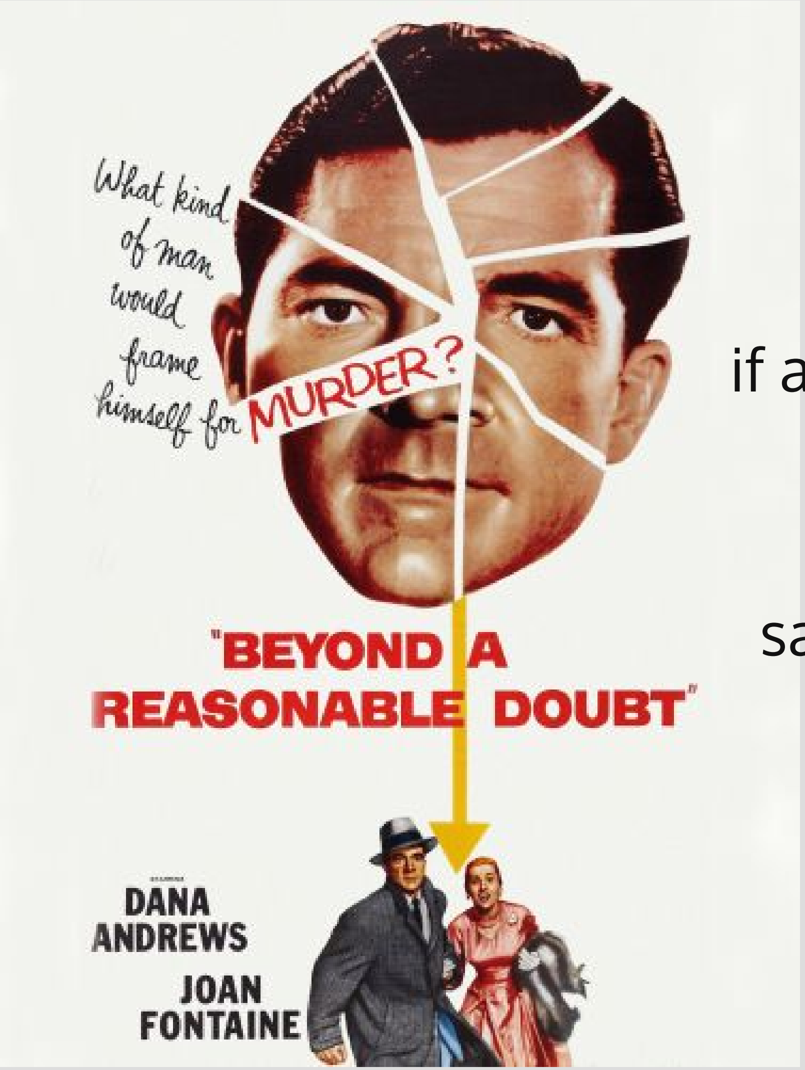
if *all alternatives* to our model are ruled out,
then our model must hold

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

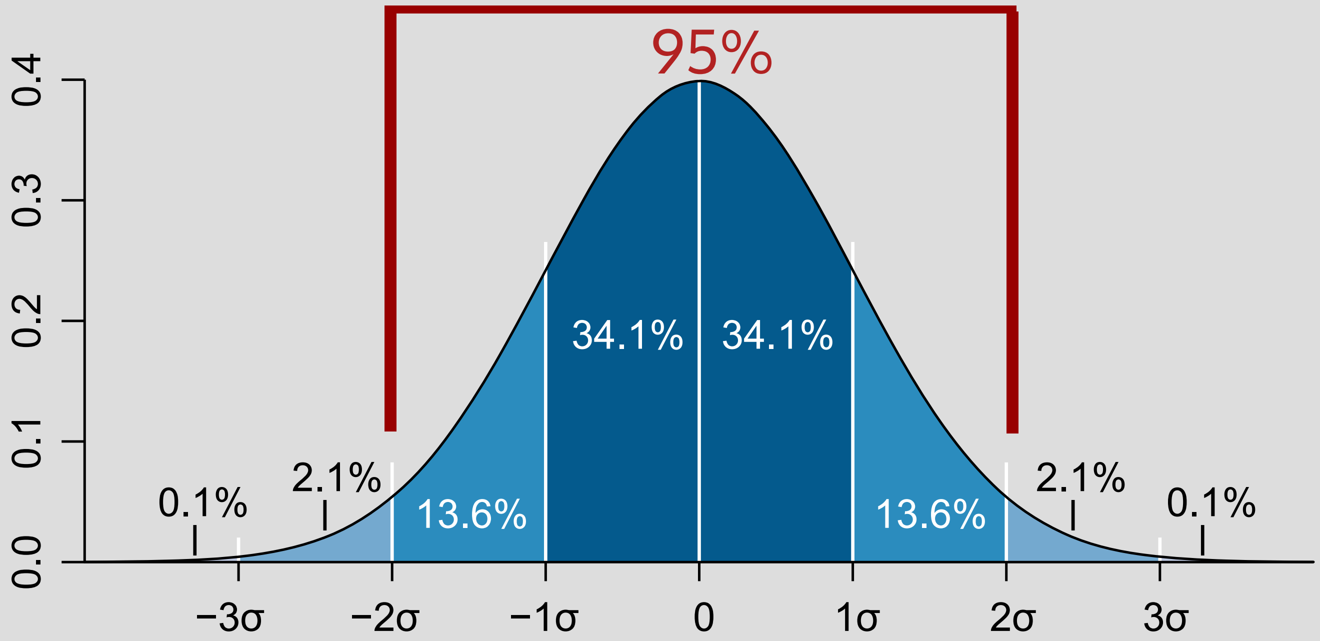
guilty beyond reasonable doubt

2

identify all alternative outcomes

Alternative Hypothesis

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

Z-test

The distribution of sample means for (independent) samples extracted from a population

with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ is

Normally distributed

$$\bar{X} \sim N(\mu = 0, \sigma = 1)$$

Z-test

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

pivotal quantity

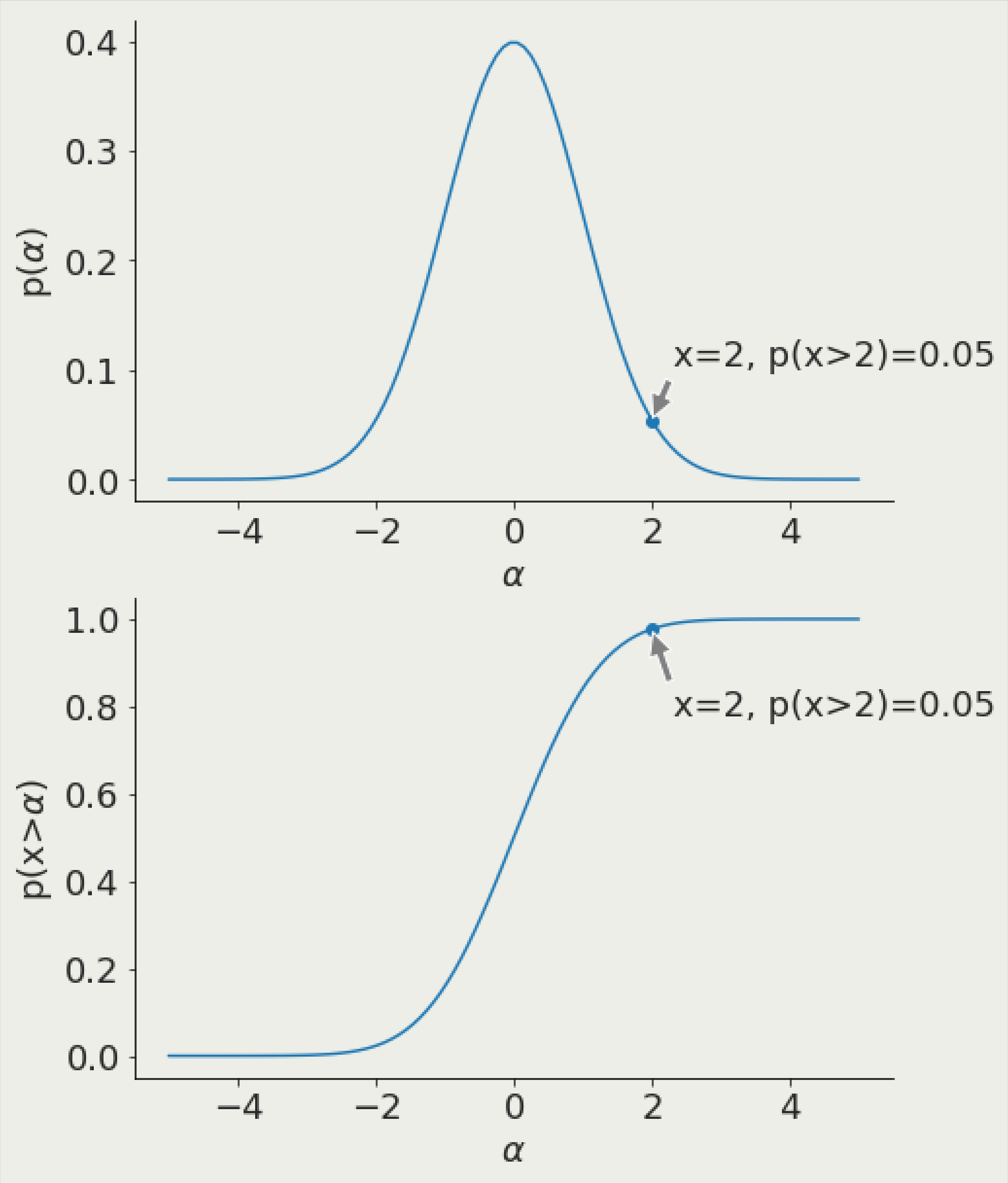
$$Z = (\bar{X} - \mu_0)/s$$

sample mean

population mean

sample variance = $\sigma_0/\sqrt{n}$

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



coding time!

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

we will start the homework together

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 = $\sigma_0 / \sqrt{n}$

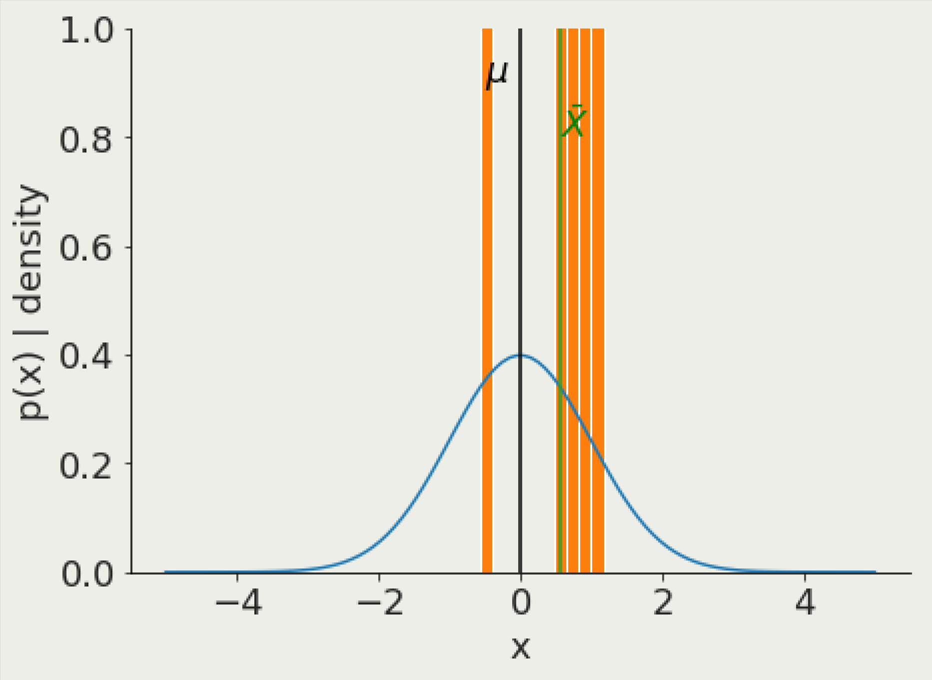
$$Z \sim N(\mu = 0,\ \sigma = 1)$$

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)

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 = $\sigma_0/\sqrt{n}$

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

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

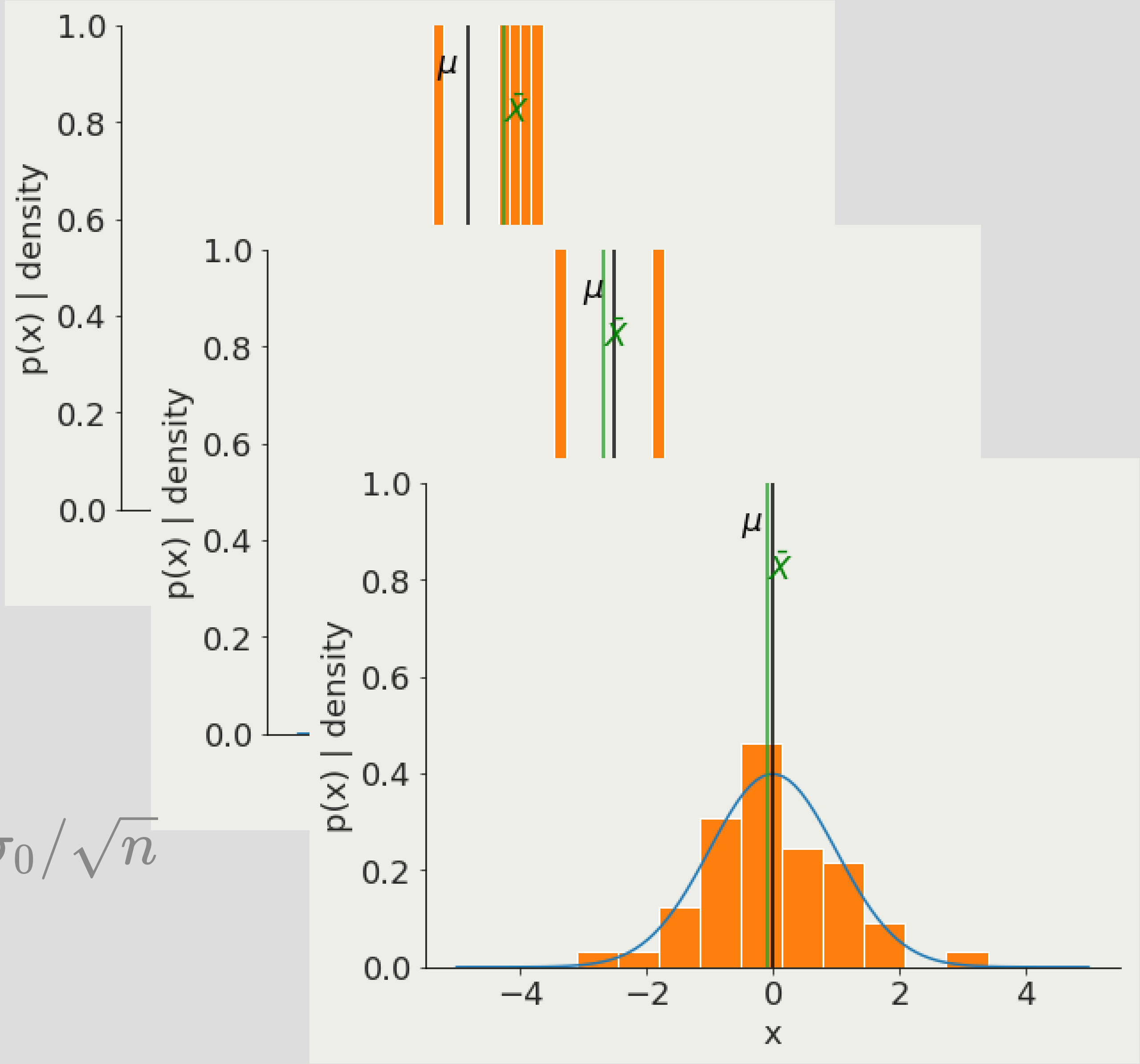
$$Z = (\bar{X} - \mu_0)/s$$

sample mean

population mean

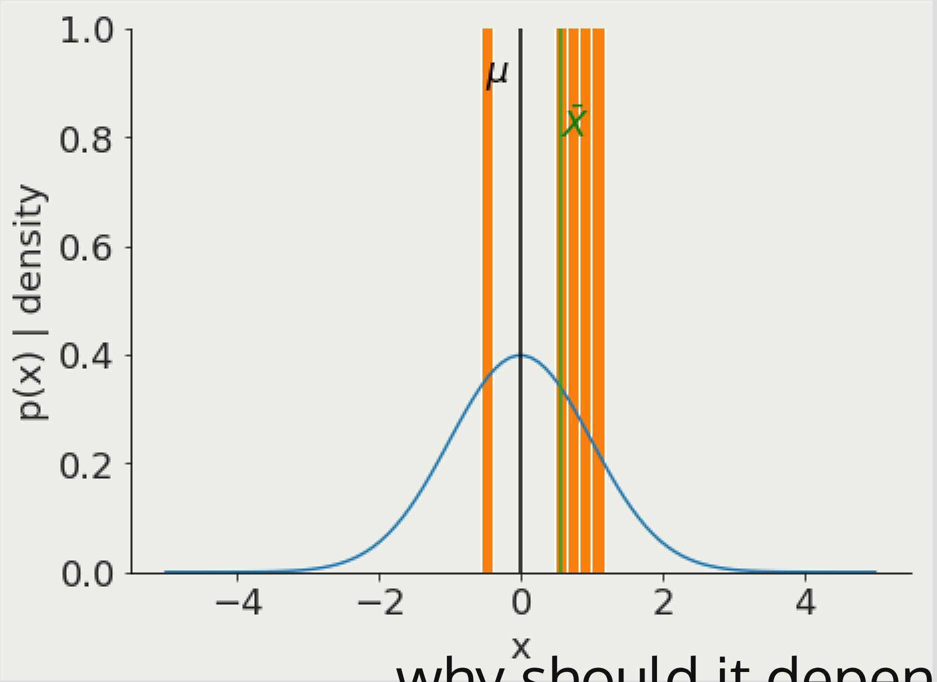
sample variance = $\sigma_0/\sqrt{n}$

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



Z-test

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



pivotal quantity

$$Z = (\bar{X} - \mu_0)/s$$

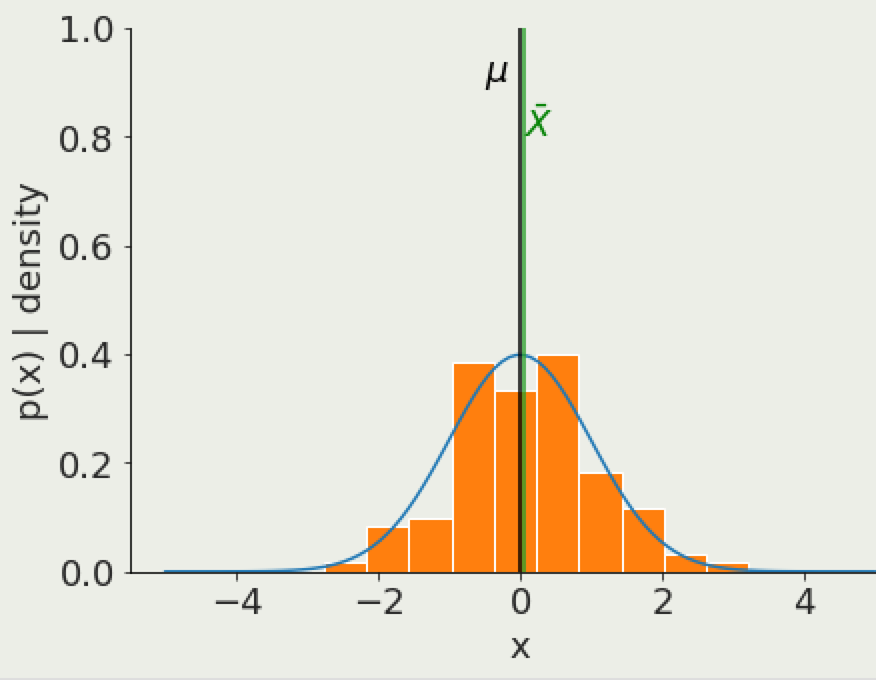
sample mean

population mean

sample variance = $\sigma_0/\sqrt{n}$

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

why should it depend on N?



Z-test

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

pivotal quantity

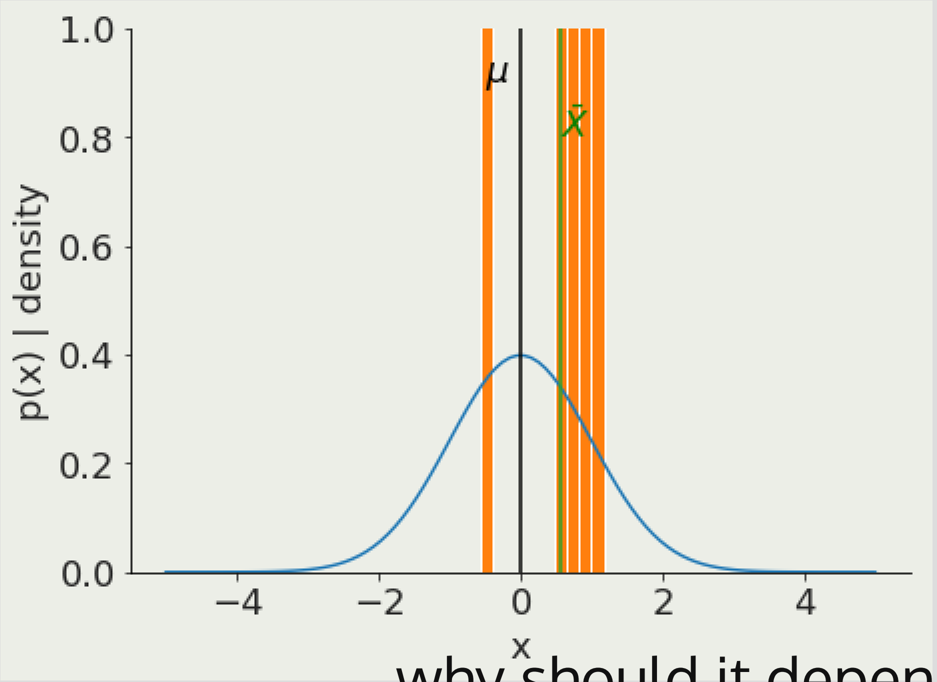
$$Z = (\bar{X} - \mu_0)/s$$

sample mean

population mean

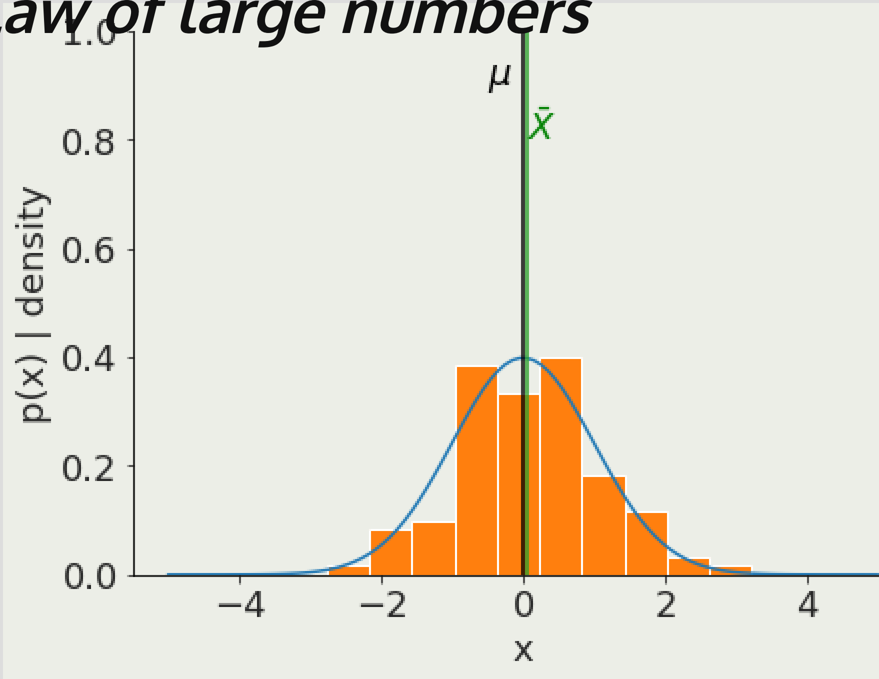
sample variance = $\sigma_0/\sqrt{n}$

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



why should it depend on N?

Law of large numbers



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.: *$\chi 2$ statistics:* difference between prediction and reality squared

Z statistics: difference between means

K-S statistics: maximum distance of cumulative distributions.

4

Null
Hypothesis
Rejection
Testing

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

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

4

Null
Hypothesis
Rejection
Testing

pivotal quantities

5

calculate it!

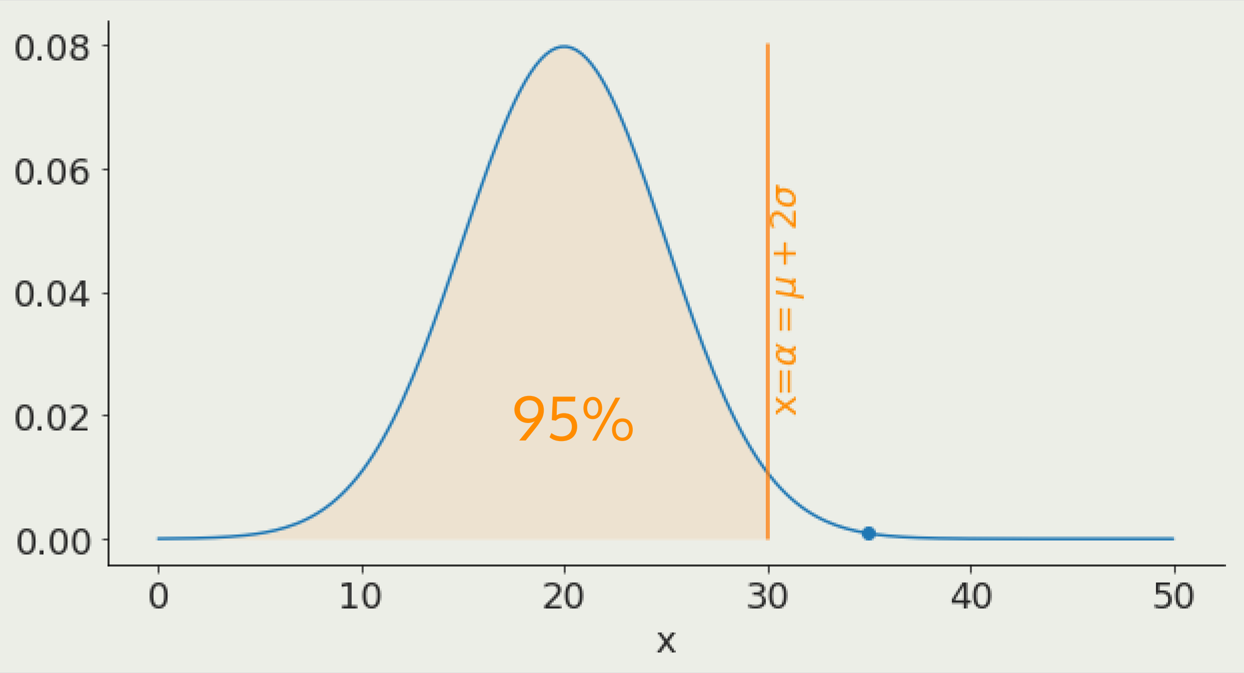
Null
Hypothesis
Rejection
Testing

6

test data against
alternative outcomes

what is α?

α is the x value corresponding to a chosen threshold



Null

Hypothesis

Rejection

Testing

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

prediction is unlikely

Null rejected

Alternative holds

☺

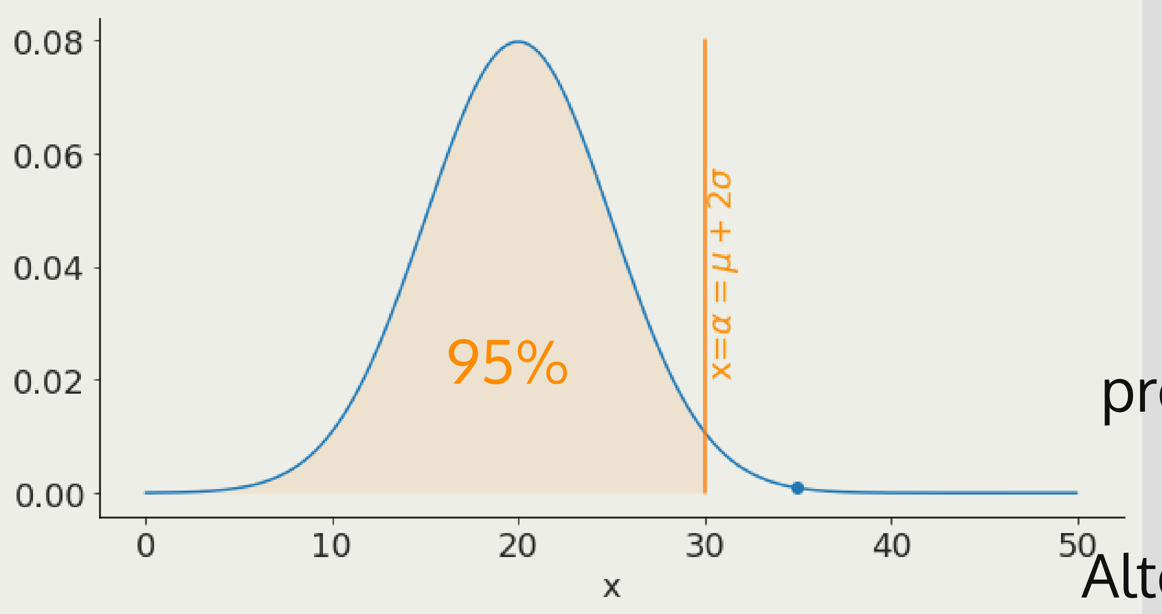
6

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

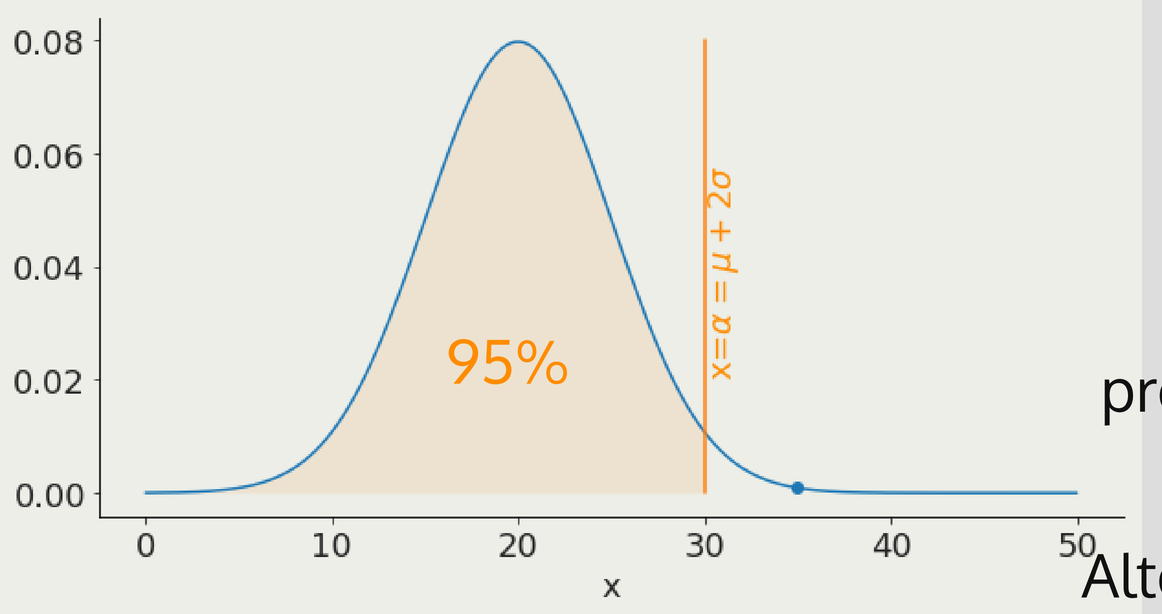
test data against
alternative outcomes

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

Null
Hypothesis
Rejection
Testing

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

prediction is unlikely
Null rejected
Alternative holds



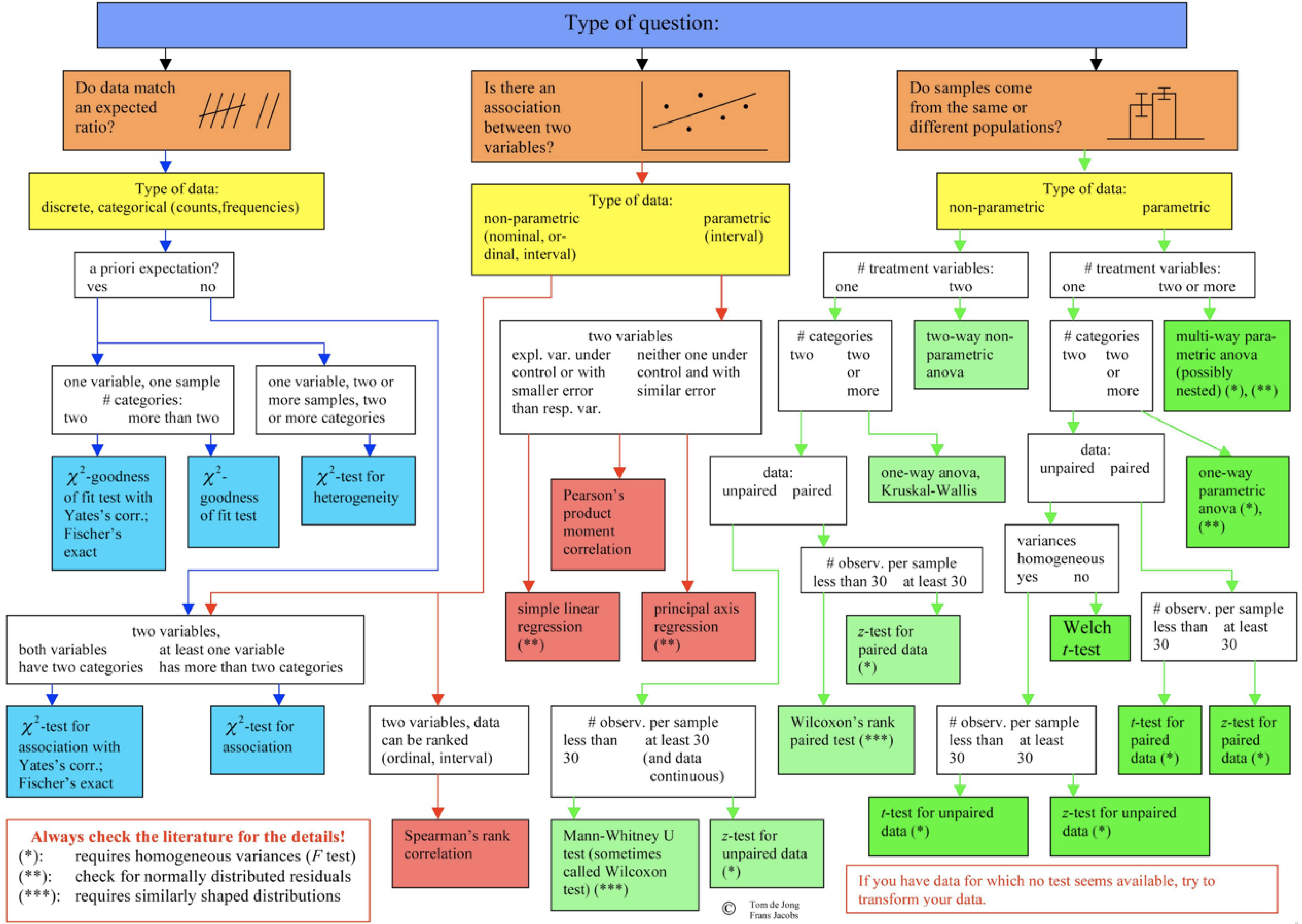
6

test data against
alternative outcomes

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

prediction is likely
Null holds
Alternative rejected

4 common tests and pivotal quantities and tests



Type of question:
Do data match an expected ratio?
Is there an association between two variables?
Do samples come from the same or different populations?
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
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
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 (and data continuous)
Mann-Whitney U test (sometimes called Wilcoxon test) (***)
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 (*)
observ. per sample less than 30 at least 30
t-test for unpaired data (*)
z-test for unpaired 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

$$Z = (\bar{X} - \mu_0)/s$$

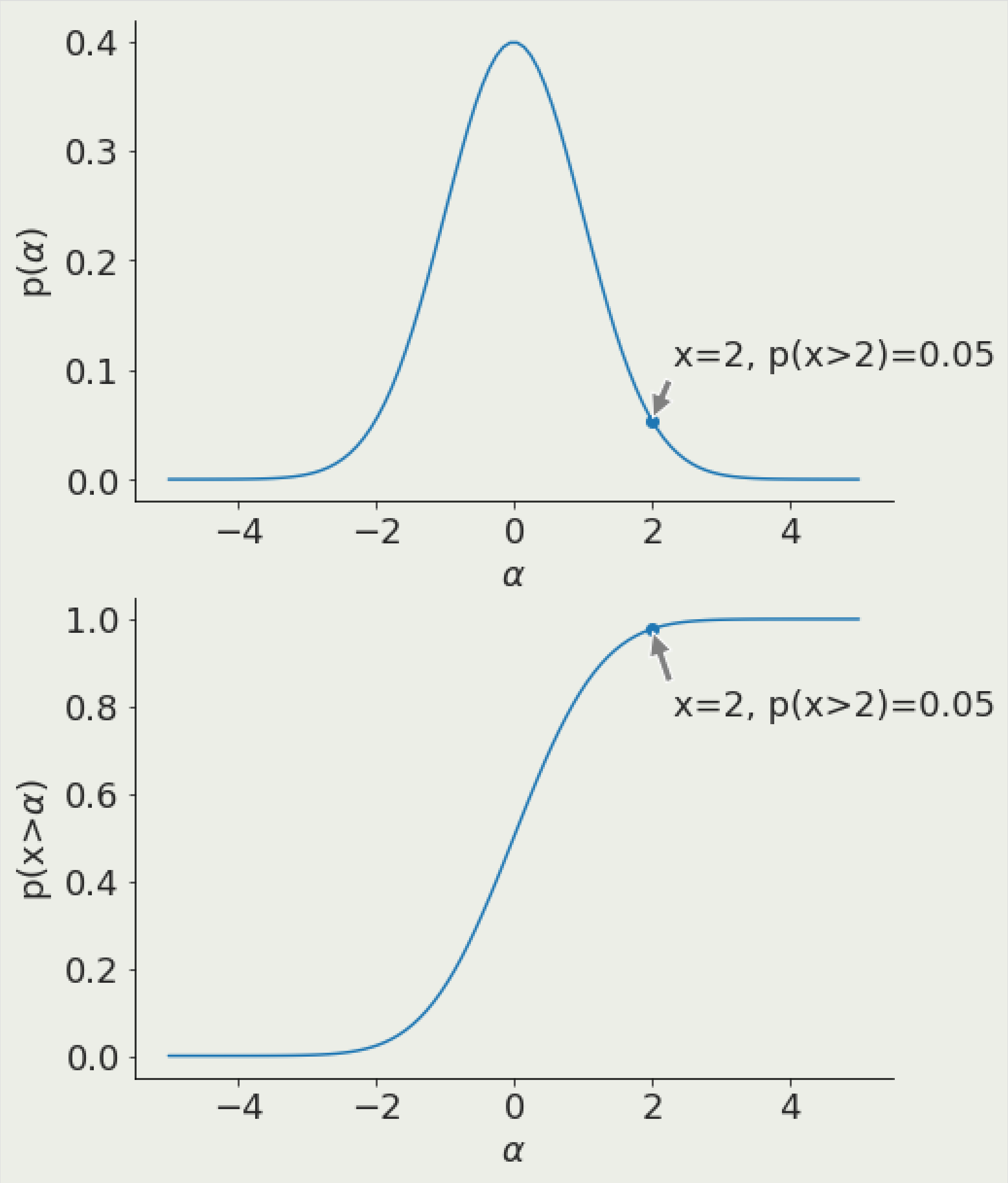
sample mean

population mean

sample variance = $\sigma_0/\sqrt{n}$

$$Z \sim N(\mu_z = 0,\ \sigma_z = 1)$$

need to know the population μ and σ



$t-$ test

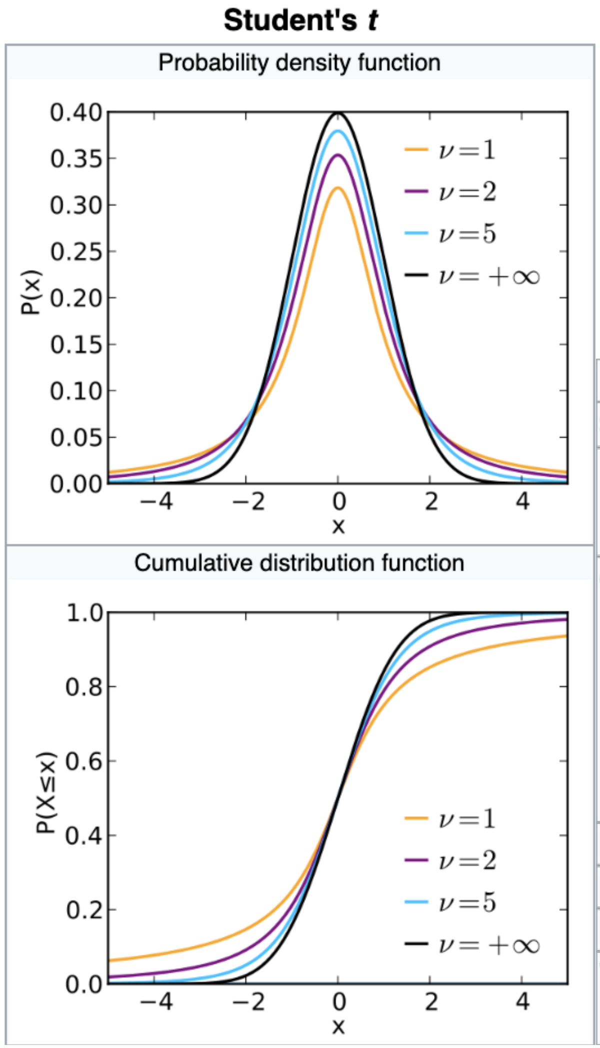
Are the means of 2 samples significantly different?

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



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 $1 < \nu \le 2$, 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

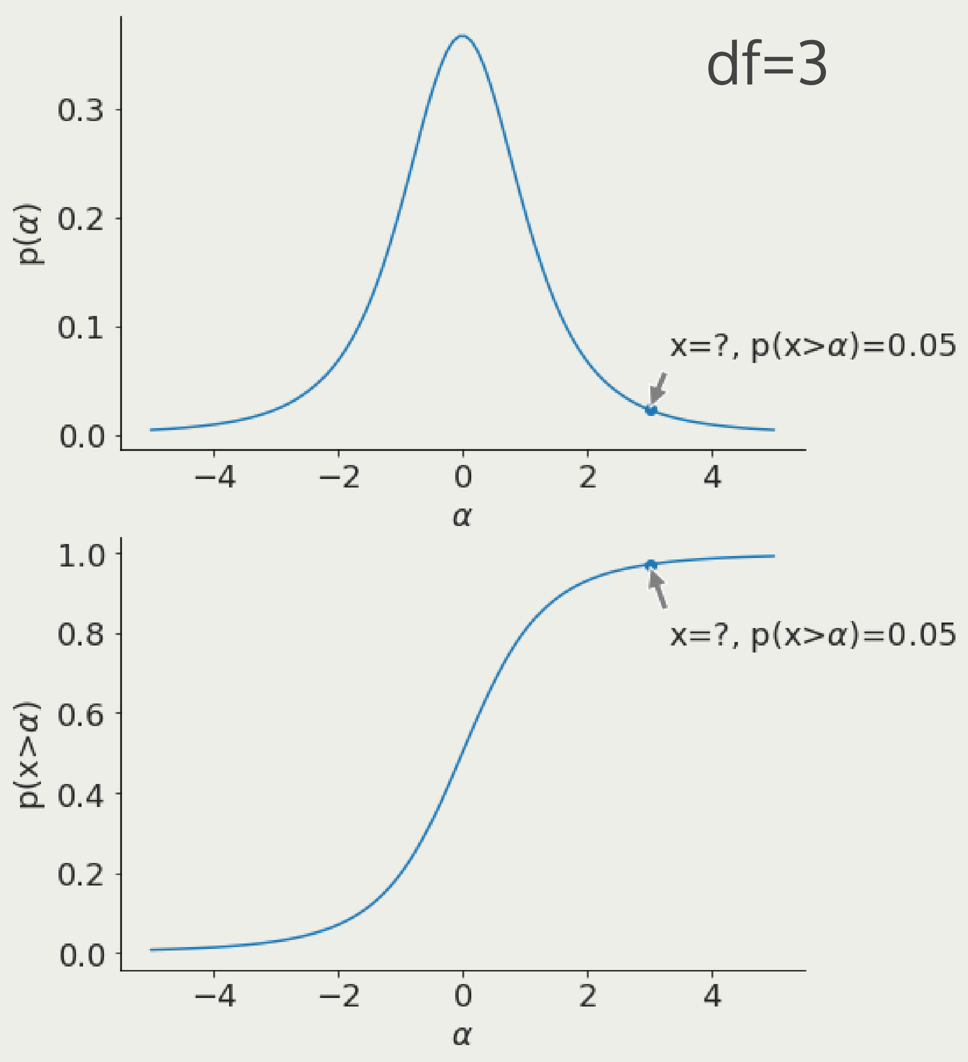
Are the means of 2 samples significantly different?

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

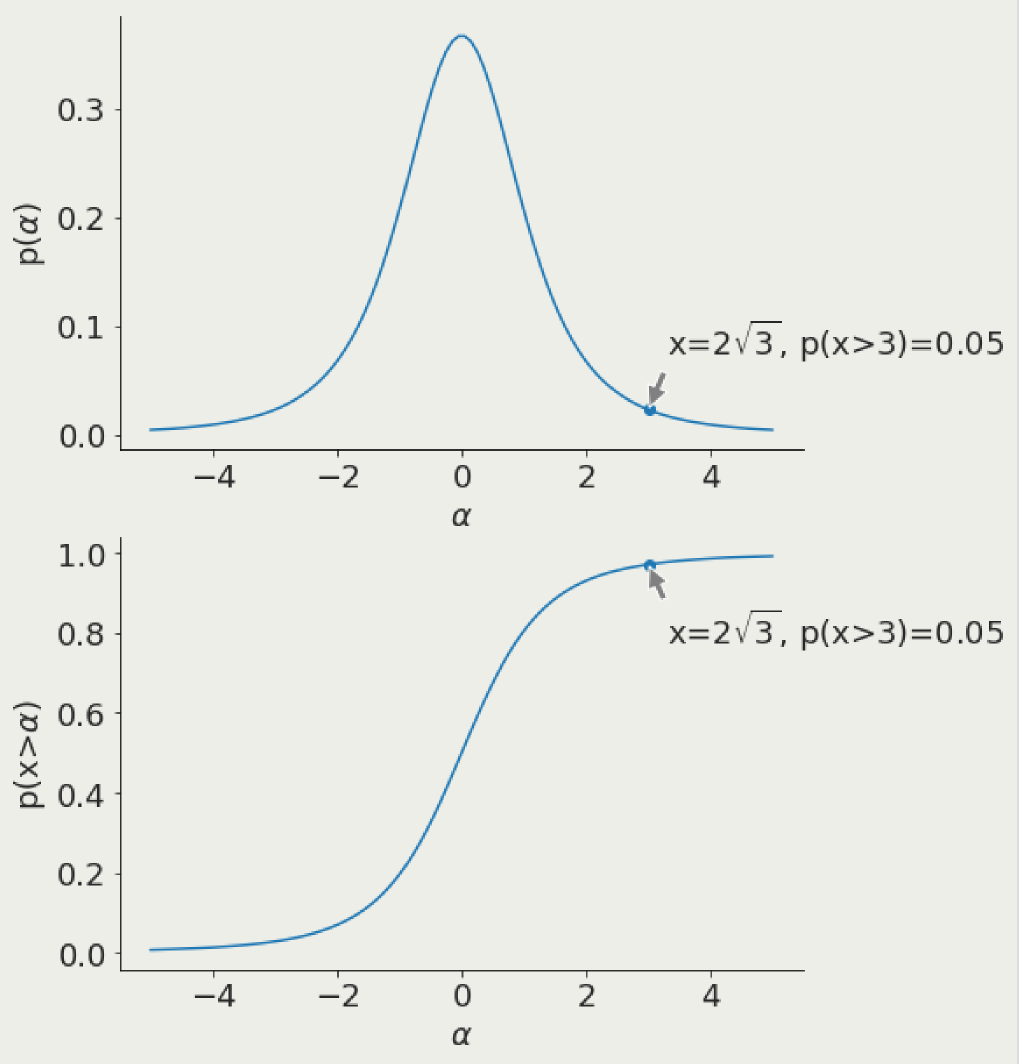
Are the means of 2 samples significantly different?

pivotal quantity

unbias variance estimator

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

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

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

Are the means of 2 samples significantly different?

pivotal quantity

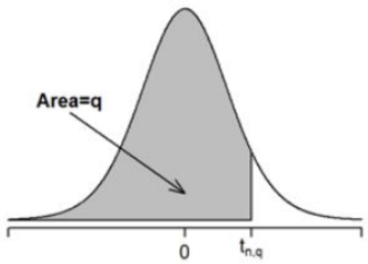
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample

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 :

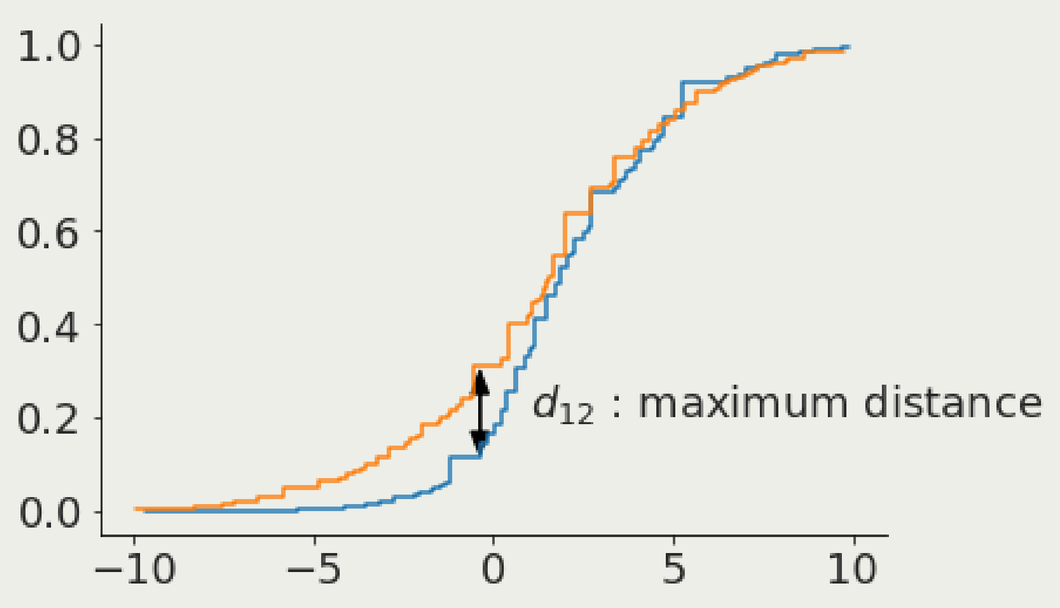
do two samples come from the same parent distribution?

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 :

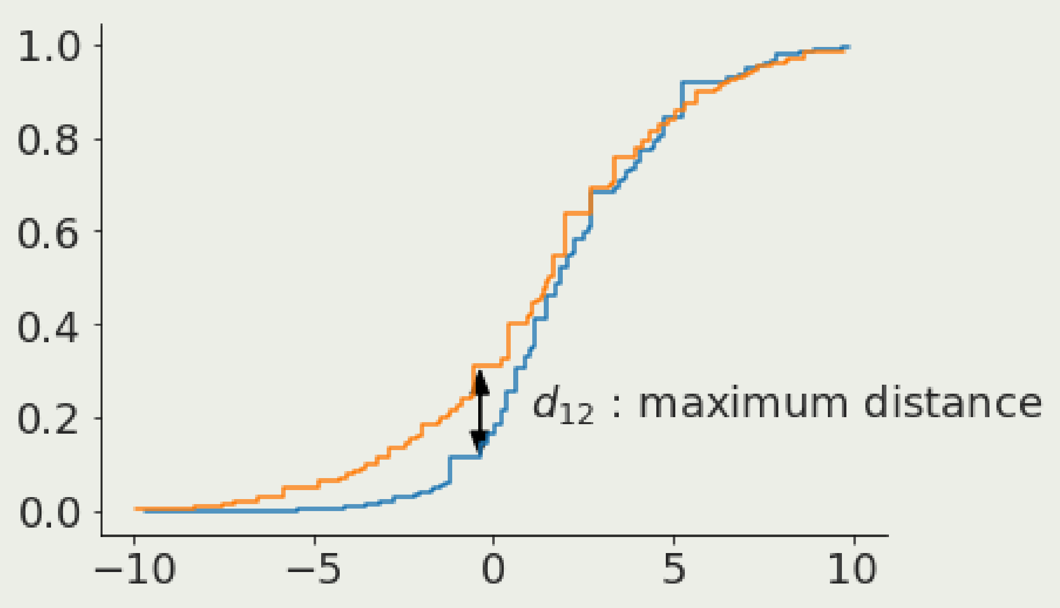
do two samples come from the same parent distribution?

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)
```

χ2 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

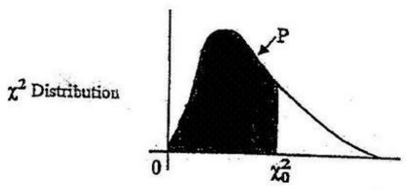
uncertainty

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

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

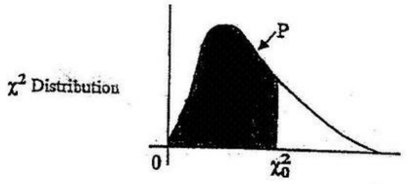
model

uncertainty

observation

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

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
								352	36.191	38.582
								[illegible]	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

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 → **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

if probability < p-value : reject Null

Key Slide

key concepts

descriptive statistics

null hypothesis rejection testing setup

pivotal quantities

Z, t, $\chi2$, K-S tests

the importance of scaling laws

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

homework

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

reading

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

https://embed.ted.com/talks/lang/en/geoffrey_west_the_surprising_math_of_cities_and_corporations

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