foundations of data science for everyone

III: probability and statistics part 2

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

this slide deck

https://slides.com/federicabianco/fds_03

1 statistics

statistics

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

TAXONOMY

Population: all of the elements of a "family" or class

Sample: a finite subset of the population that you observe

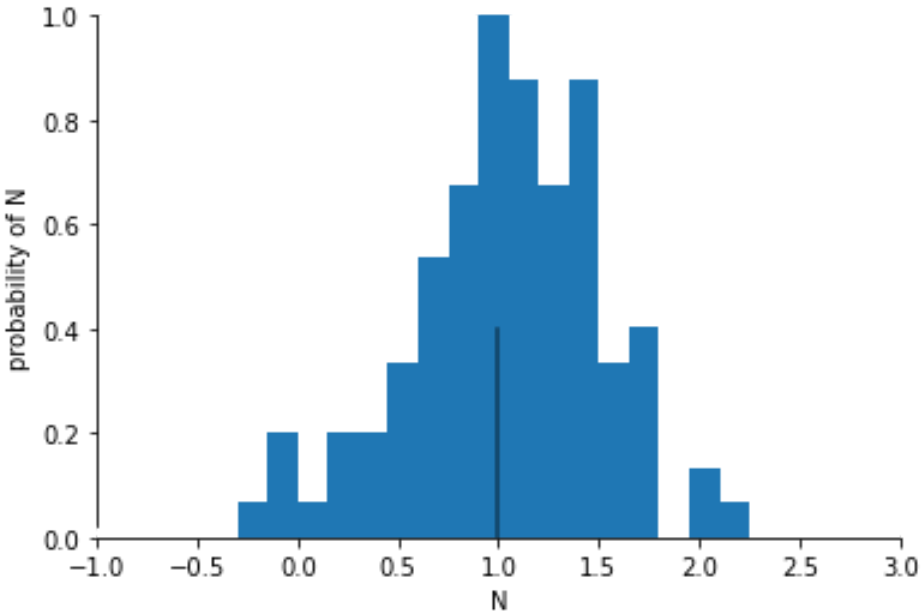
distributions: a collection of numbers with a specific shape

observational approach: a distribution represent the frequency with which we obtain a value ~x when measuring a phenomenon

HISTOGRAMS OF SAMPLES

analyst approach: a distribution represent the *probability* with which a phenomenon generates a value that we measure to be ~x

number of values measured between x and x+dx



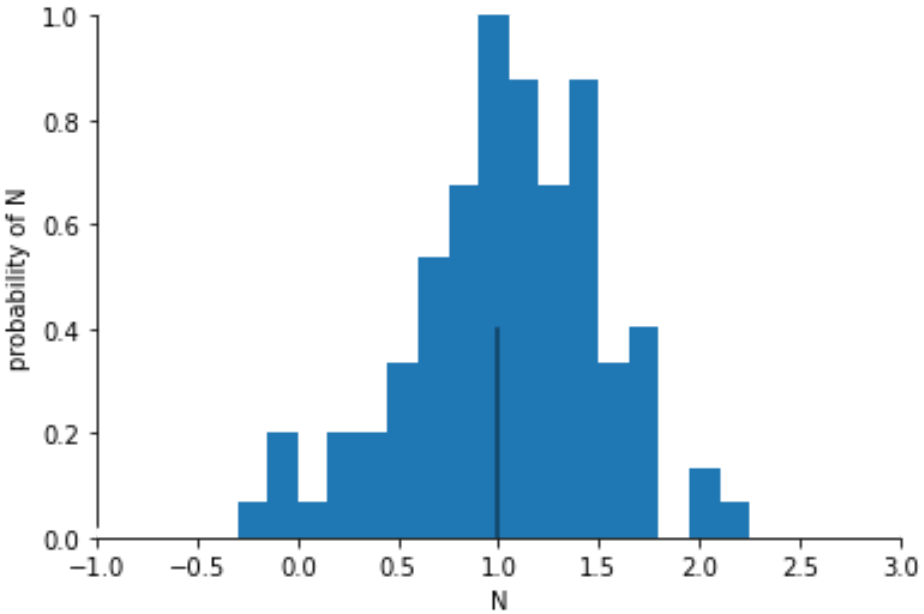
frequency ⟶ **probability**

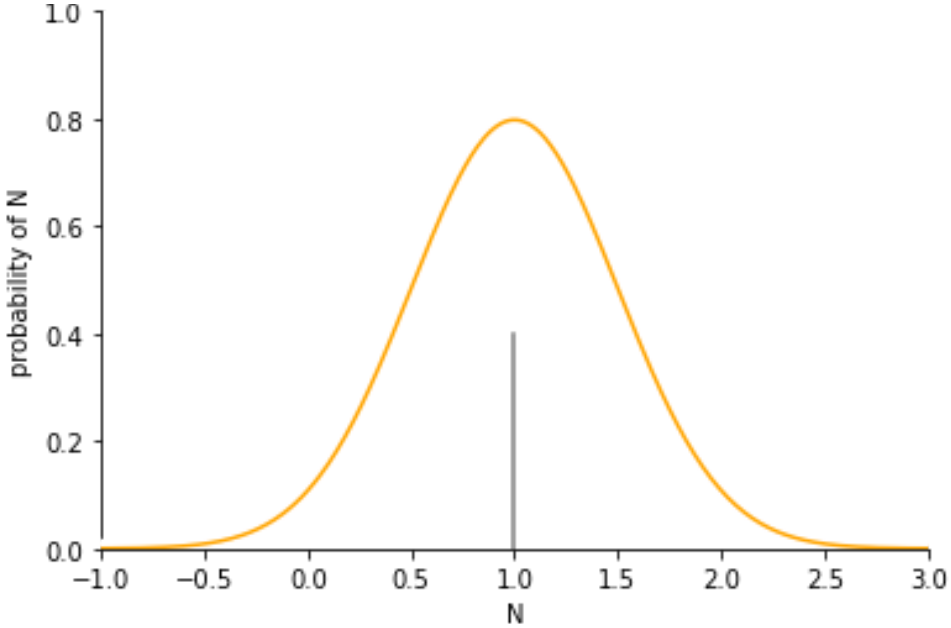
distributions: a collection of numbers with a specific shape

observational approach: a distribution represent the frequency with which we obtain a value ~x when measuring a phenomenon

analyst approach: a distribution represent the *probability* with which a phenomenon generates a value that we measure to be ~x

number of values measured between x and x+dx

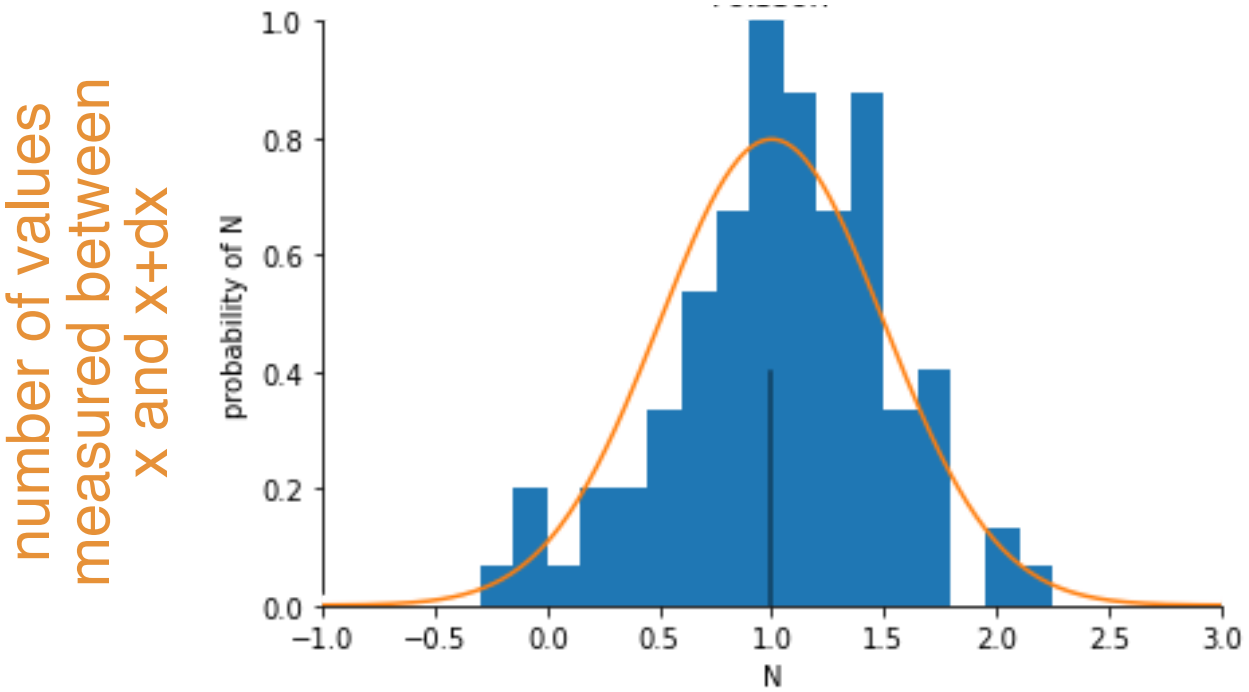




distributions: a collection of numbers with a specific shape

observational approach: a distribution represent the frequency with which we obtain a value ~x when measuring a phenomenon

analyst approach: a distribution represent the *probability* with which a phenomenon generates a value that we measure to be ~x



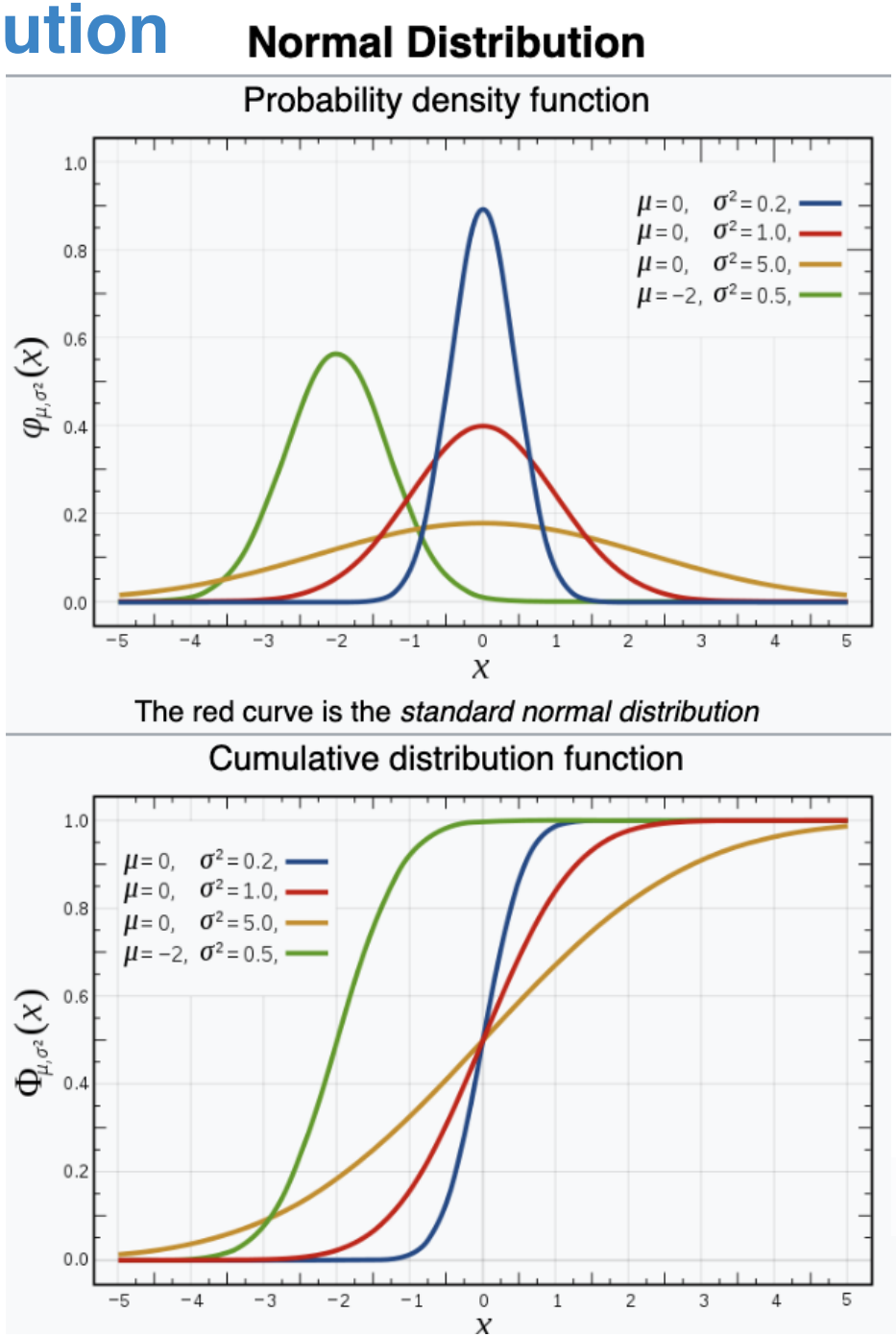
frequency → **probability**

Gaussian (or normal) distribution

most common distribution:

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

support: *the x values for which the distribution is defined*



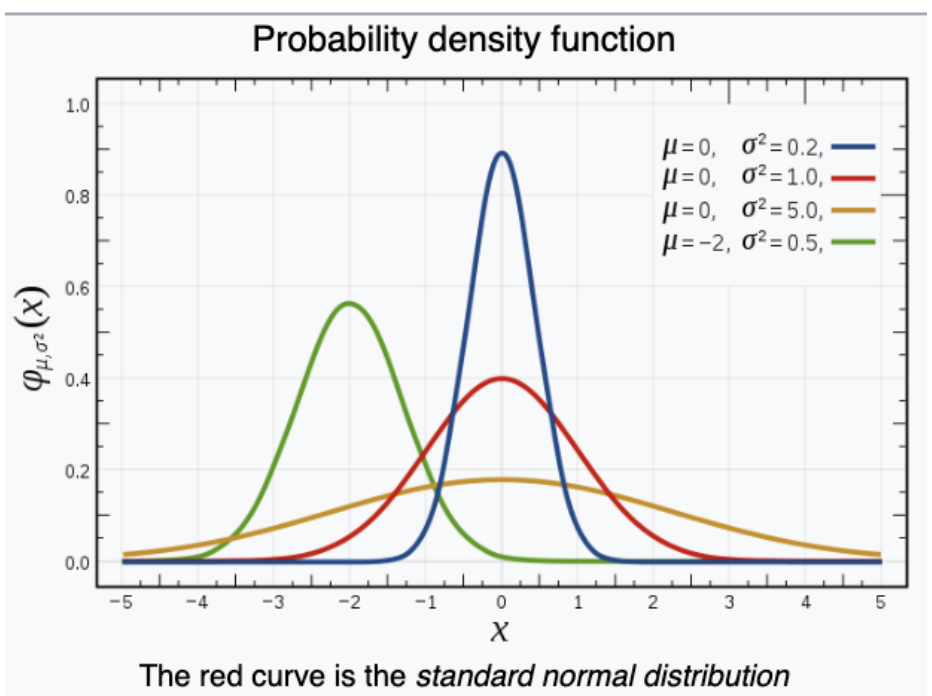
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared sc
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
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	$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

Gaussian (or normal) distribution

most common distribution:

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

support: *the x values for which the distribution is defined*



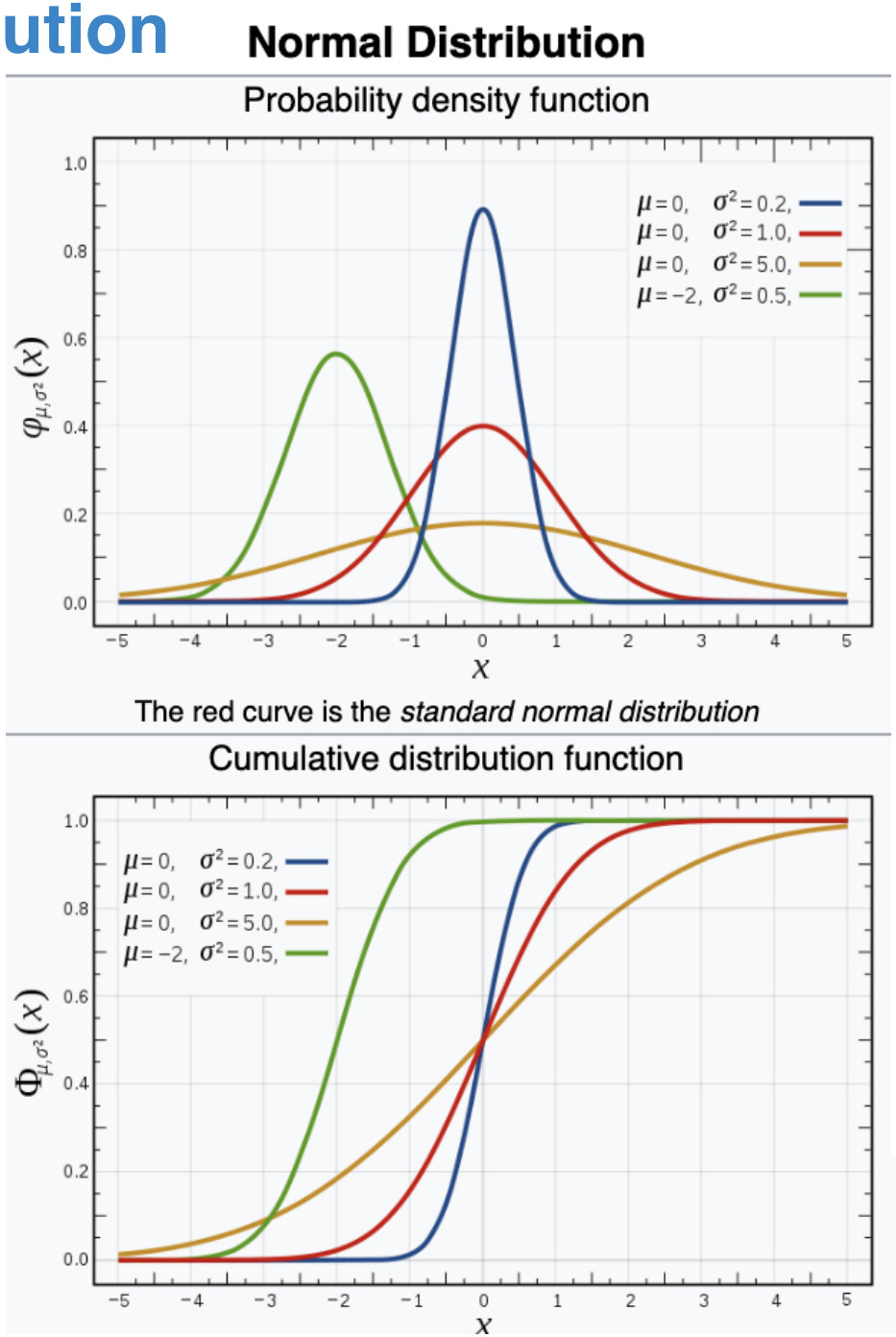
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared sc
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
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	$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

Gaussian (or normal) distribution

most common distribution:

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

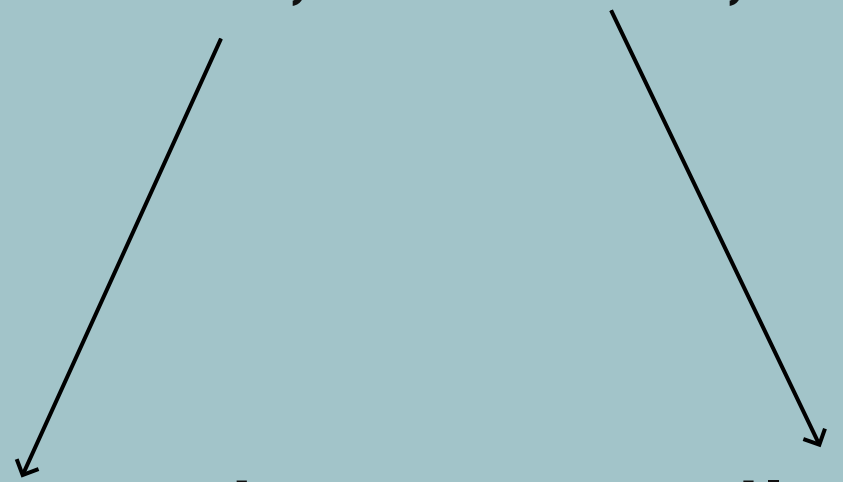
Central tendency



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	$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

TAXONOMY

central tendency: mean, median, mode



spread : variance, interquantile range

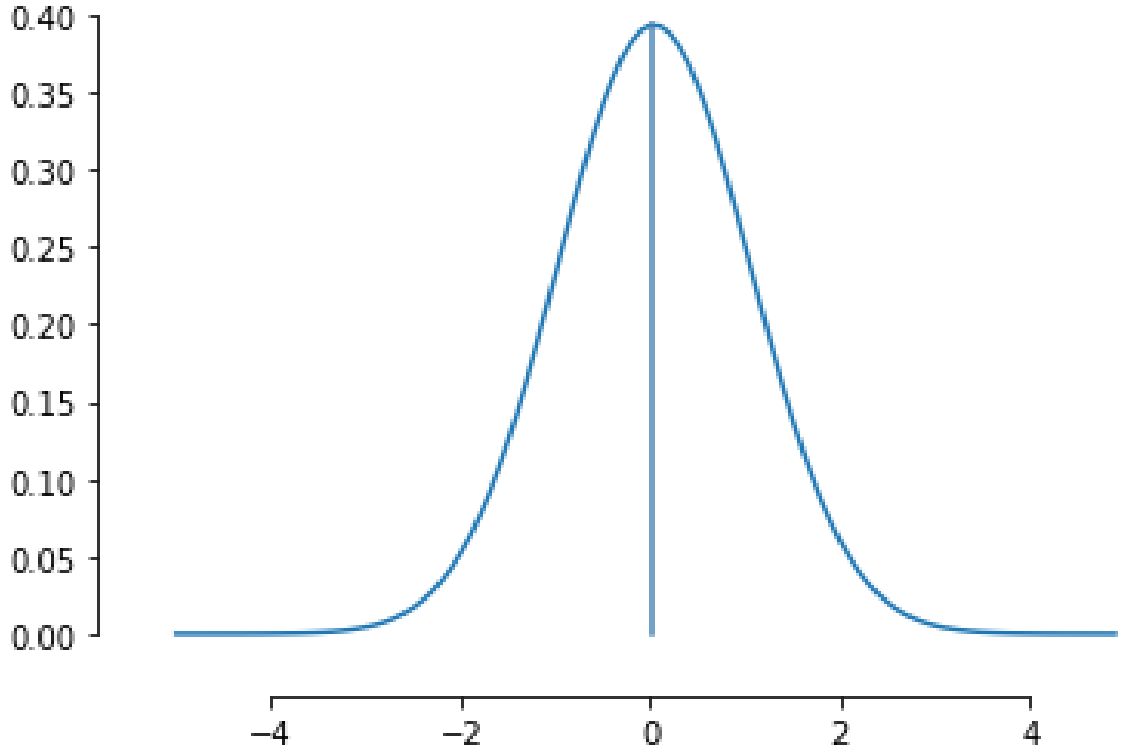
Moments and quantiles

a distribution's moments summarize its properties:

$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

central tendency: mean (n=1)

Gaussian (or Normal) distribution



Quantiles

measure what fraction of a distribution is within some x values

central tendency: median (50%)

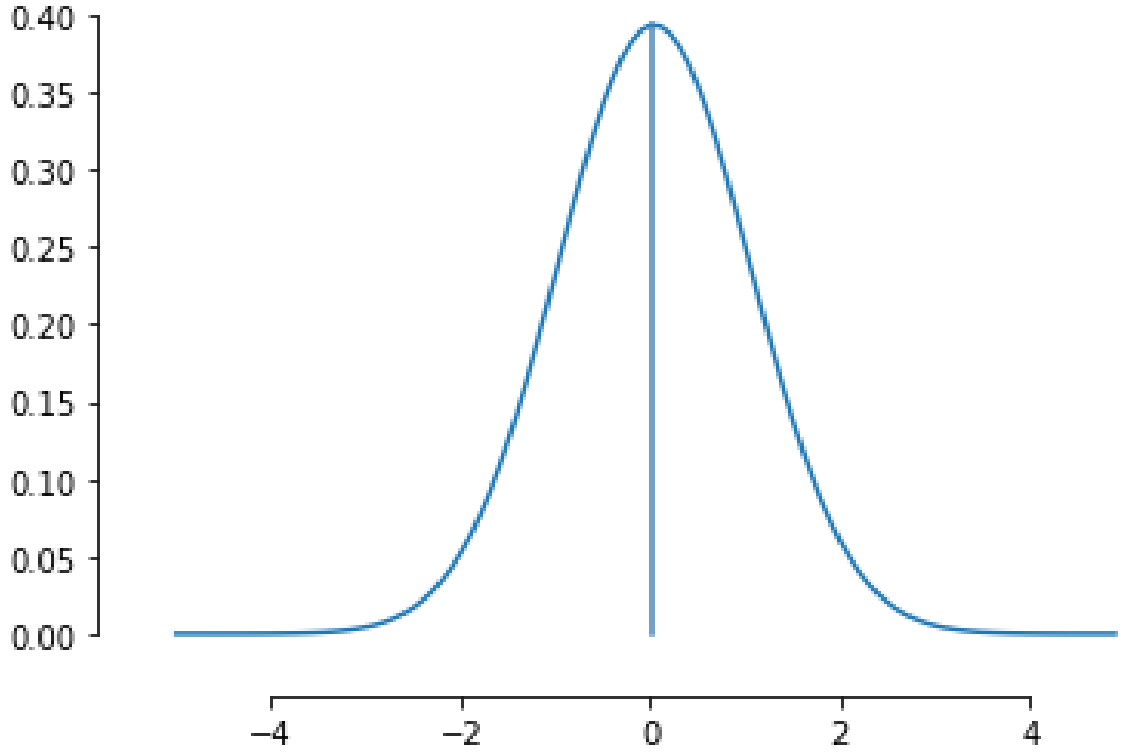
Moments

a distribution's moments summarize its properties:

$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

central tendency: mean (n=1)

Gaussian (or Normal) distribution



Quantiles

measure what fraction of a distribution is within some x values

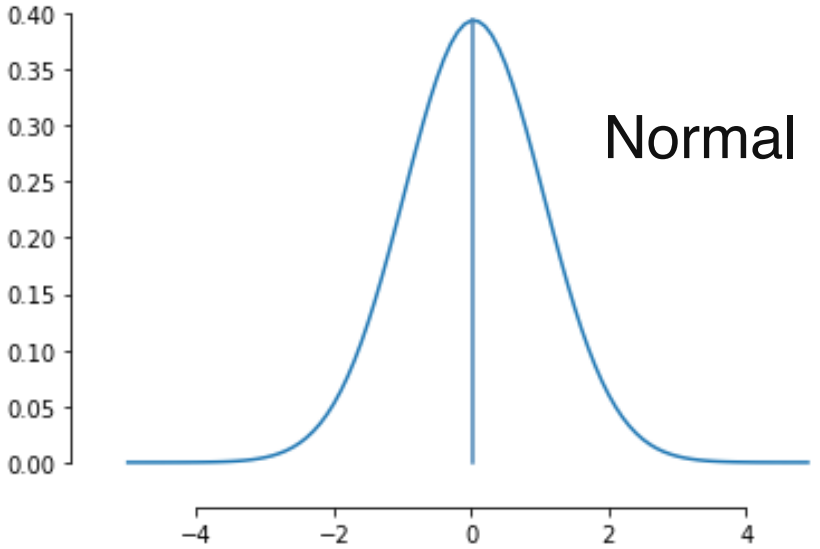
central tendency: median (50%)

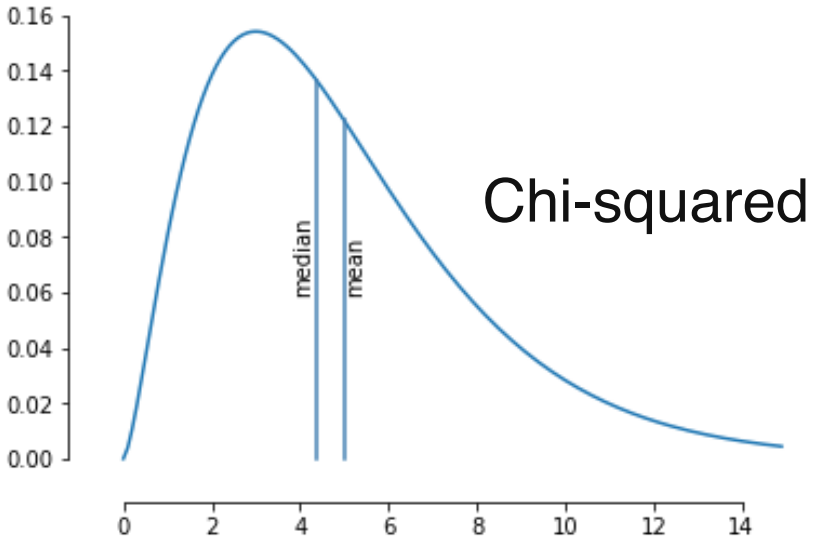
Moments

a distribution's moments summarize its properties:

symmetric distribution: mean=median

skewed distribution: mean≠median





Quantiles

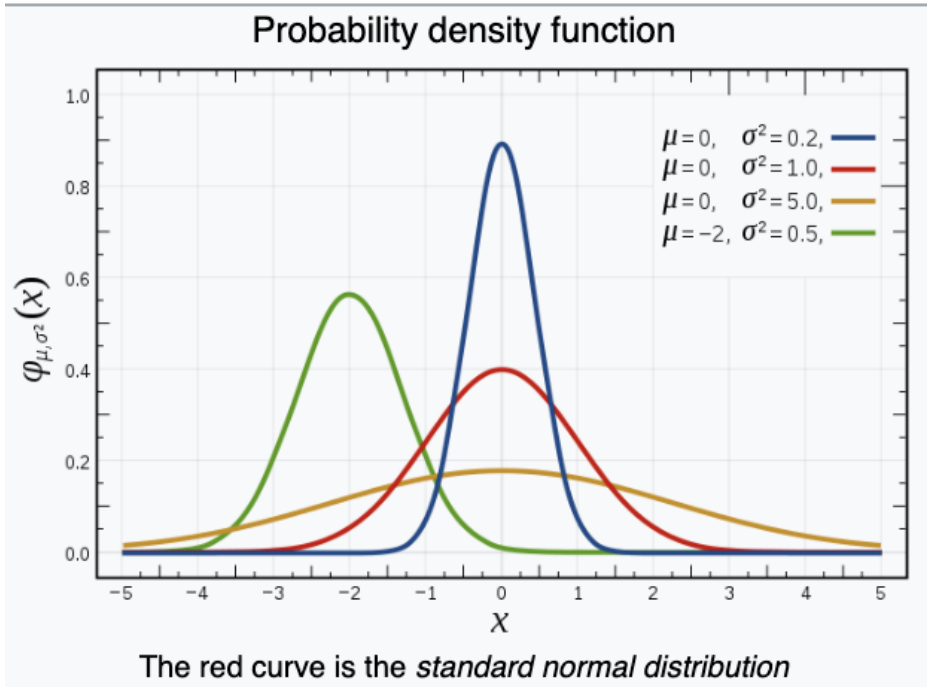
Gaussian (or normal) distribution

most common distribution:

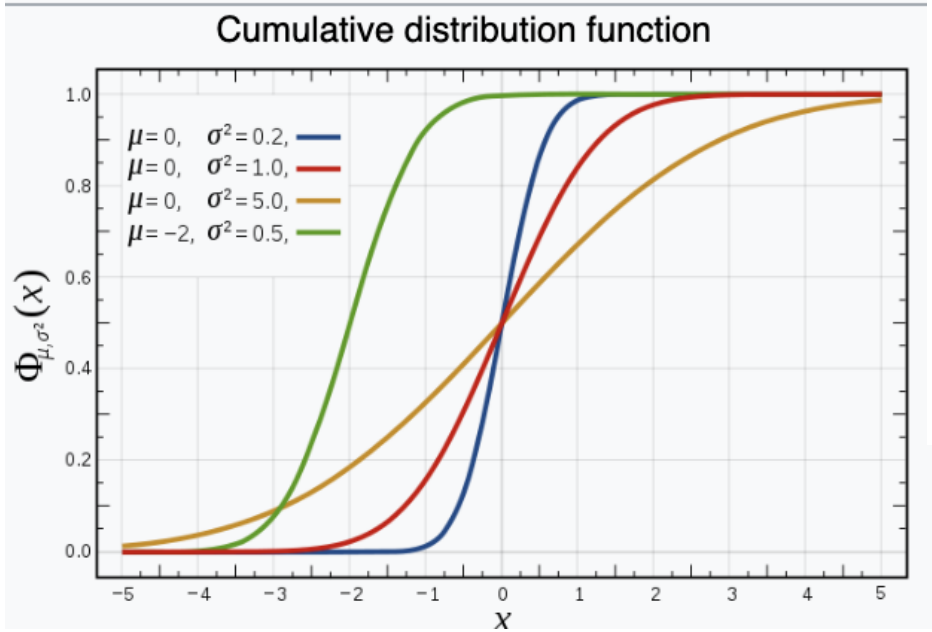
well behaved mathematically, symmetric, when we can we will assume our uncertainties or samples are Gaussian distributed

Spread

Normal Distribution



The red curve is the *standard normal distribution*



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	$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

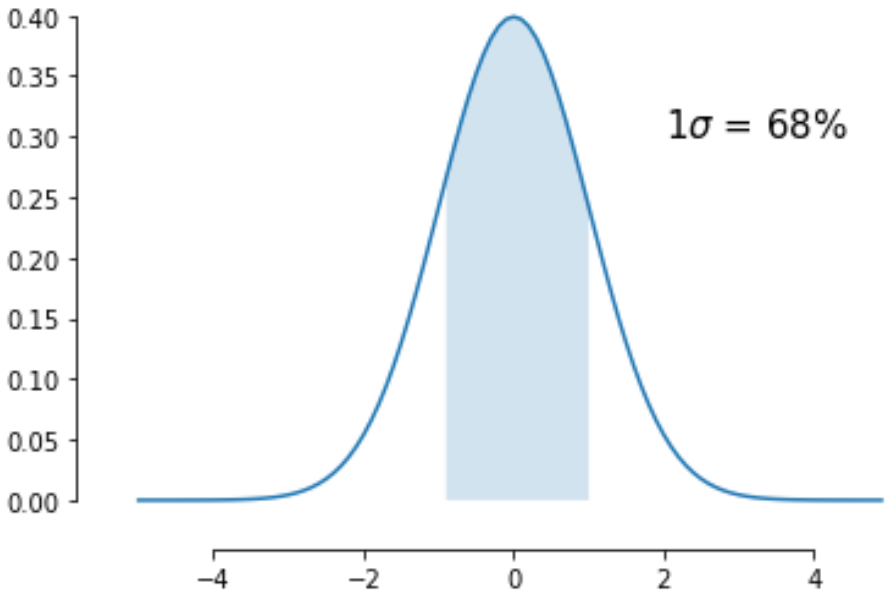
Moments

a distribution's moments summarize its properties:

$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

spread: **variance (n=2)**

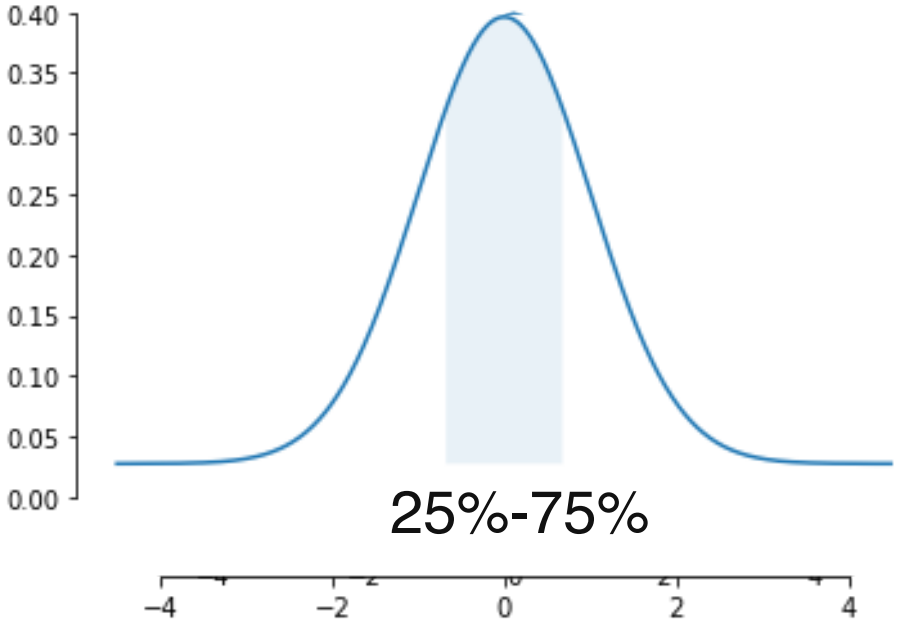
standard deviation $\sigma = \sqrt{\text{variance}}$



Quantiles

measure what fraction of a distribution is within some x values

central tendency: **quartiles (25%-75%)**



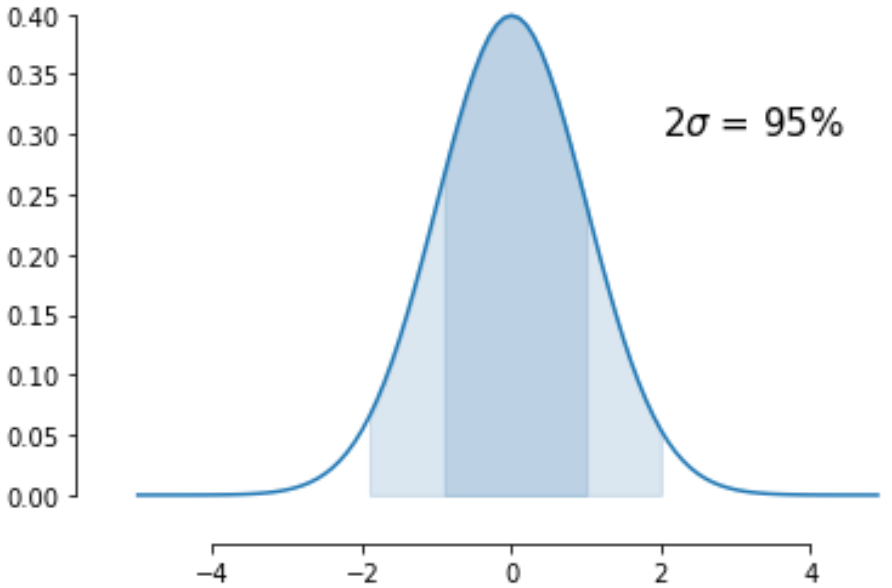
Moments and quantiles

a distribution's moments summarize its properties:

$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

spread: **variance (n=2)**

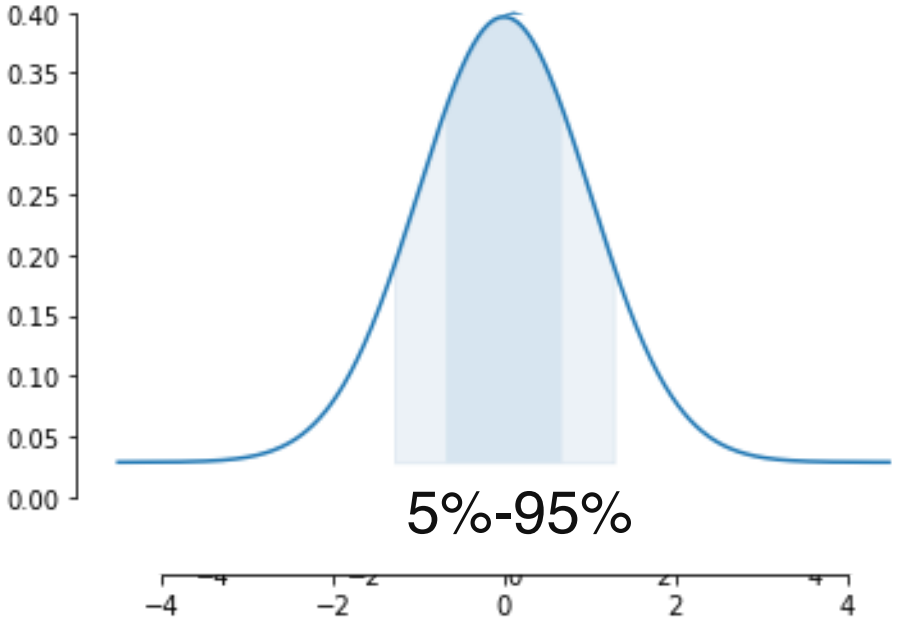
standard deviation $\sigma = \sqrt{\text{variance}}$



Quantiles

measure what fraction of a distribution is within some x values

central tendency: **quantiles (5%-95%, 1%-99%...)**



Binomial

I bet heads:

head = success

"given n tosses, each with a probability of 0.5 to get head"

Coin toss:

Parameters: p, n

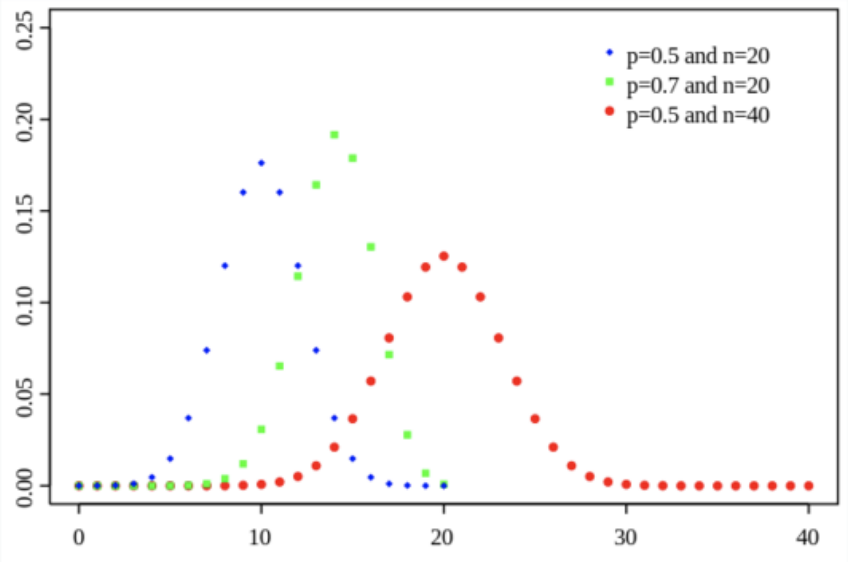
fair coin: p=0.5 n=1

Vegas coin: p=0.5 n=1

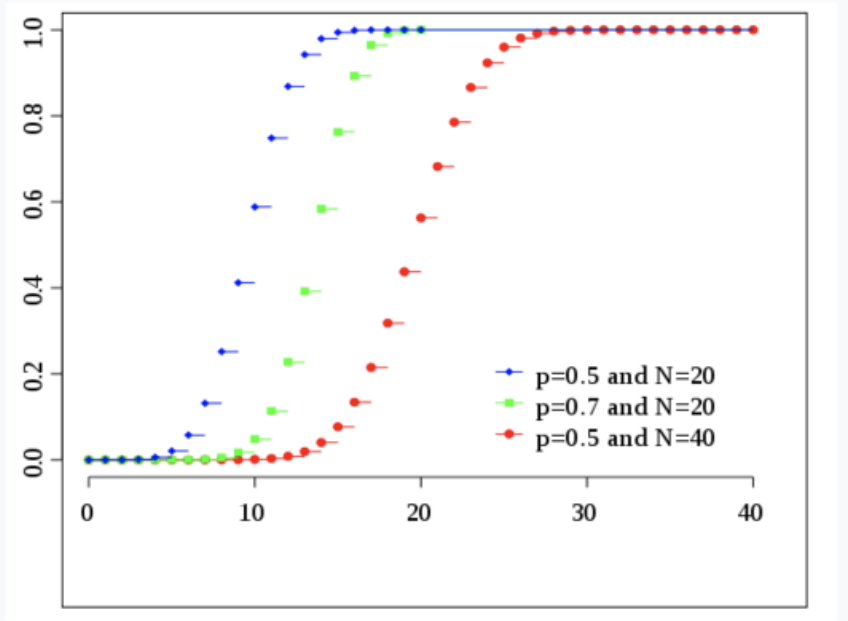
Support: integer positive
Mean: np
Variance: np*(1-p)

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)

Gaussian (or normal) distribution

most common distribution:

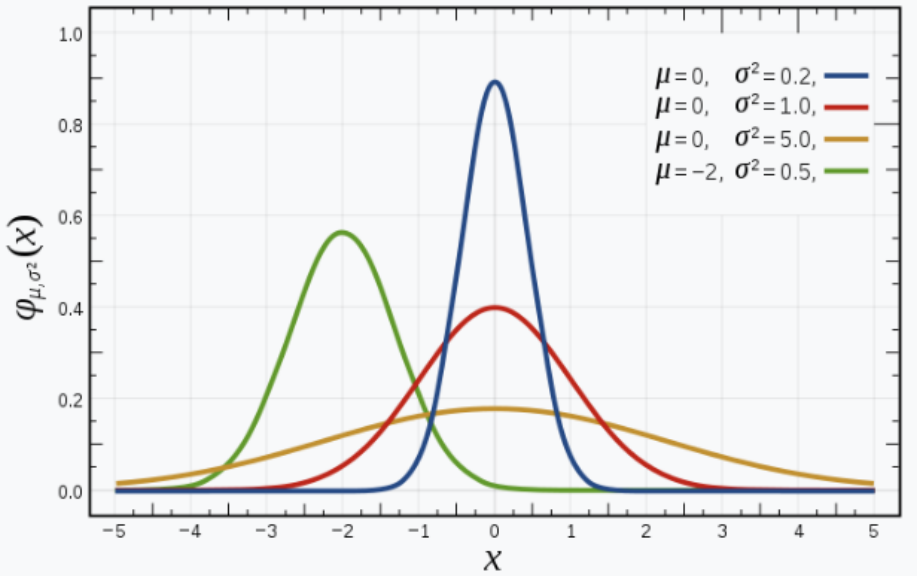
well behaved mathematically, symmetric, when we can we will assume our uncertainties or samples are Gaussian distributed

Parameters: *μ, σ*

Support: natural numbers
Mean: μ
Variance: σ

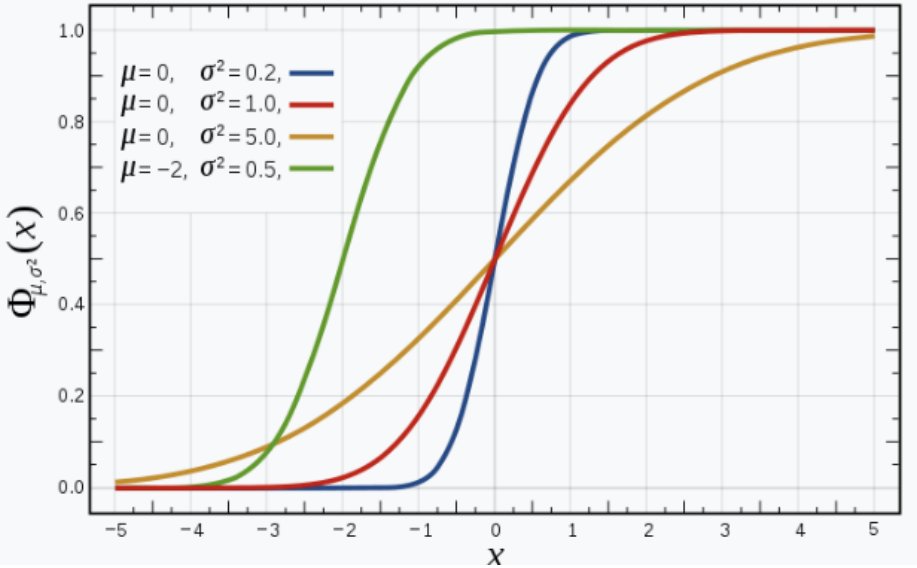
Normal Distribution

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	$D_{\mathrm{KL}}(\mathcal{N}_0 \|\mathcal{N}_1) = \frac{1}{2}\{(\sigma_0/\sigma_1)^2 + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + 2\ln\frac{\sigma_1}{\sigma_0}\}$

Poisson

Shut noise/count noise

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

Parameters: λ

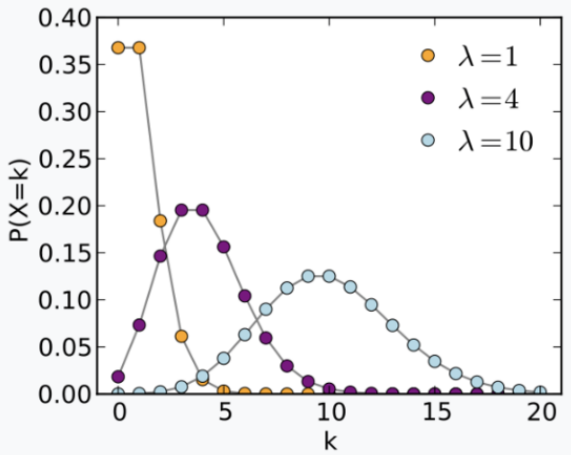
Support: natural numbers

Mean: λ

Variance: λ

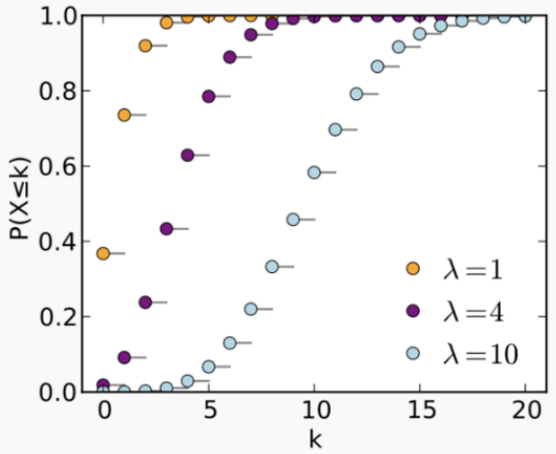
Poisson

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	$\frac{1}{\lambda}$

Moments and frequentist probability

a distribution's moments summarize its properties:

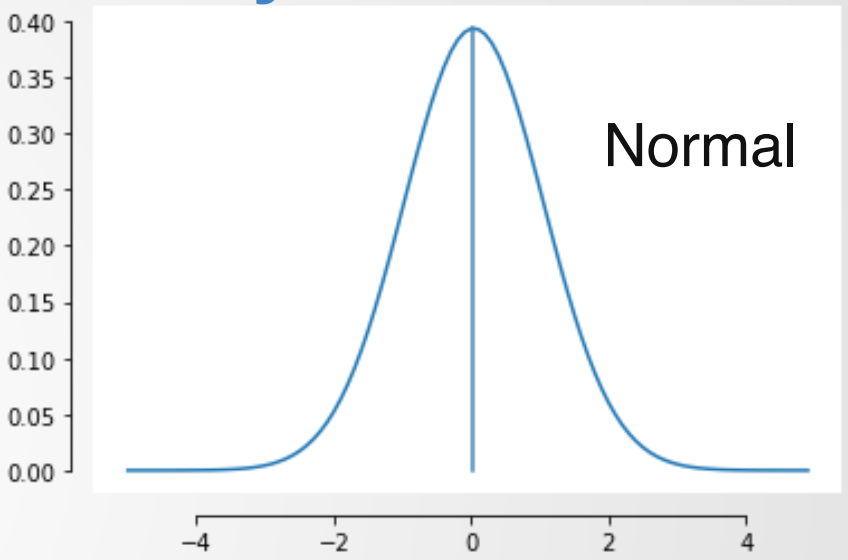
$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

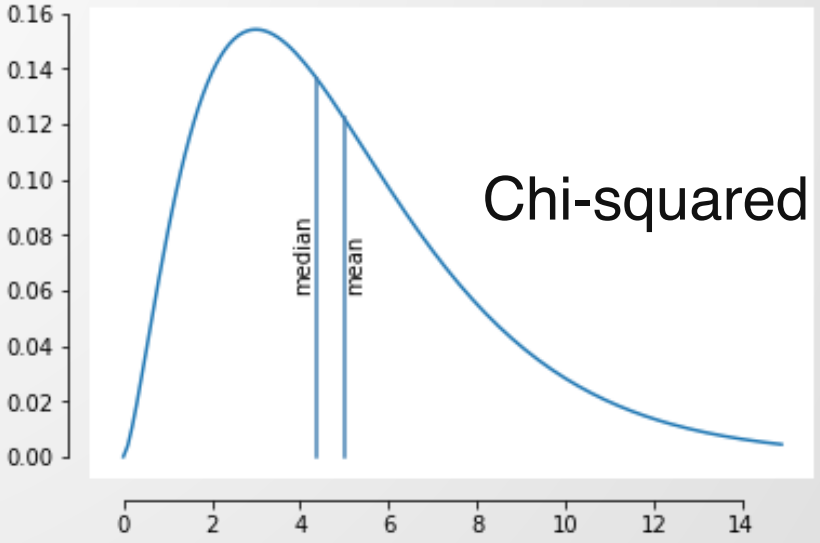
central tendency: mean (n=1), median, mode (peak)

spread: standard deviation (variance n=2), quantiles

symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)





coding time!

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

https://github.com/fedhere/FDSfE_FBianco/blob/master/statistics/distributionParametersDemo.ipynb

2 are they the same?

questions we need statistics to answer

Preamble: kinds of descriptive questions

- What is the highest/lowest value?
  - *what is the most viewed video?*
  - what is the average number of views?

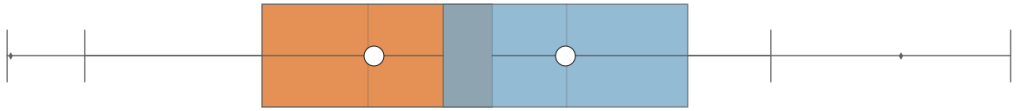
```
videos.describe()
```

	category_id	views	likes	dislikes	comment_count
count	40949.000000	4.094900e+04	4.094900e+04	4.094900e+04	4.094900e+04
mean	19.972429	2.360785e+06	7.426670e+04	3.711401e+03	8.446804e+03
std	7.568327	7.394114e+06	2.288853e+05	2.902971e+04	3.743049e+04
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	17.000000	2.423290e+05	5.424000e+03	2.020000e+02	6.140000e+02
50%	24.000000	6.818610e+05	1.809100e+04	6.310000e+02	1.856000e+03
75%	25.000000	1.823157e+06	5.541700e+04	1.938000e+03	5.755000e+03
max	43.000000	2.252119e+08	5.613827e+06	1.674420e+06	1.361580e+06

```
1 videos["views"].mean()
2 videos["views"].median()
3
4 --> 2360784.6382573447
5 --> 681861.0
6
```

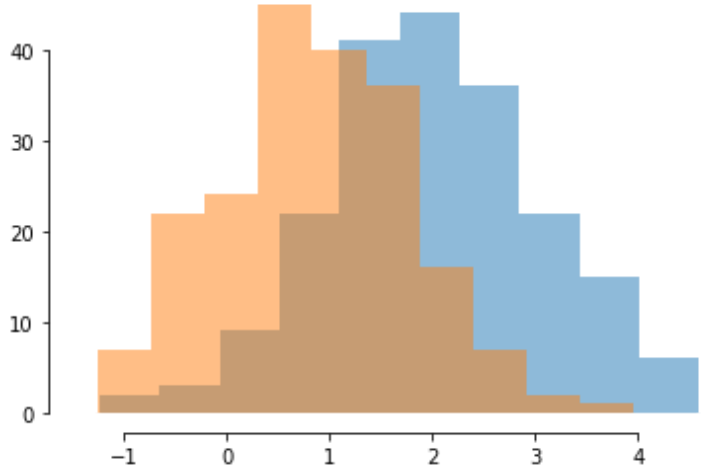
Preamble: kinds of analytical questions

- Are two measurements the same?



  - *is the amount of rain in Wilmington this year the same as last year?*

- Are two distributions the same?



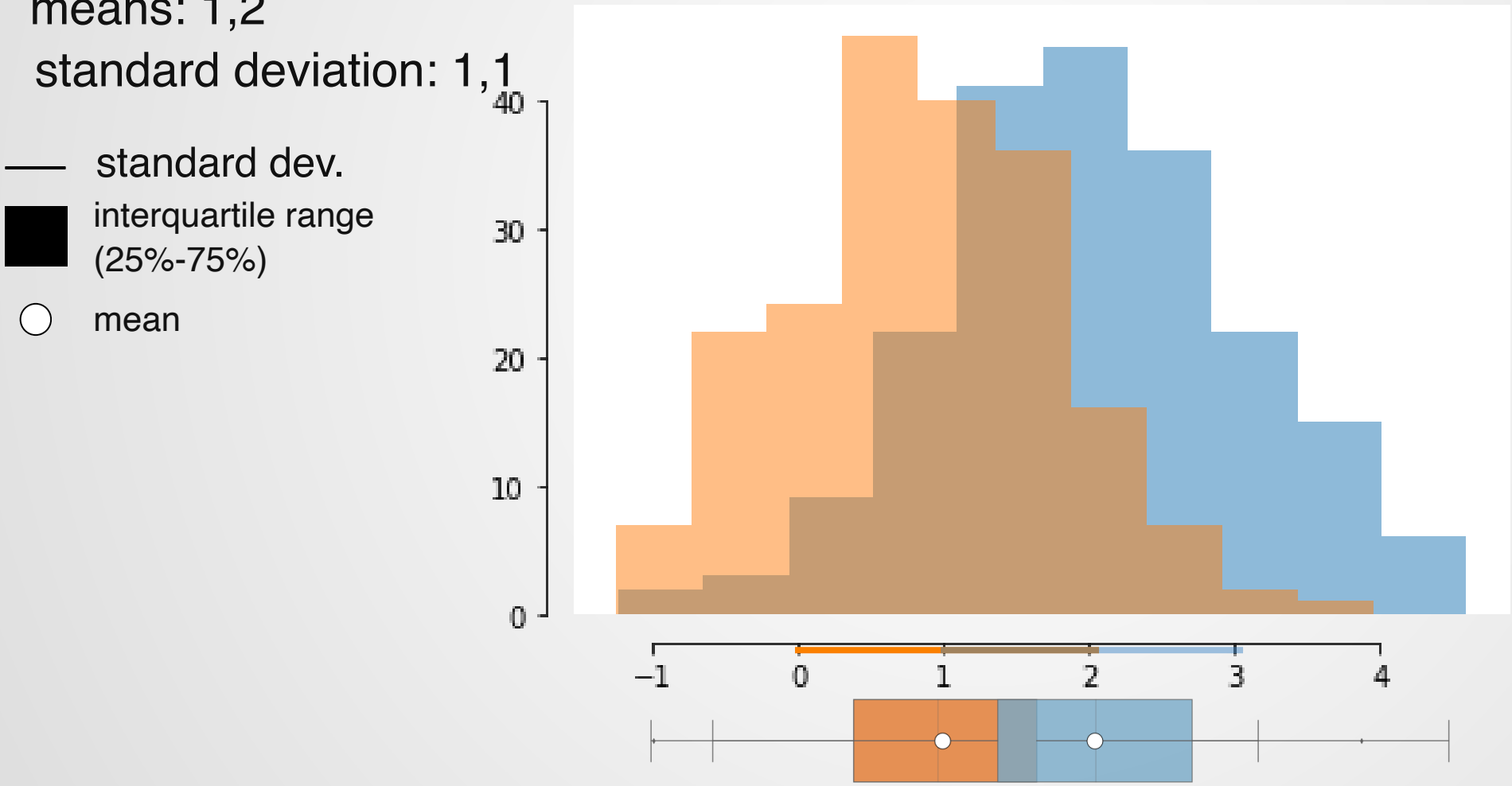
  - *is the age distribution of CityBike users the same among registered Male customers and registered Female customers?*

measuring differences between distributions

means: 1,2

standard deviation: 1,1

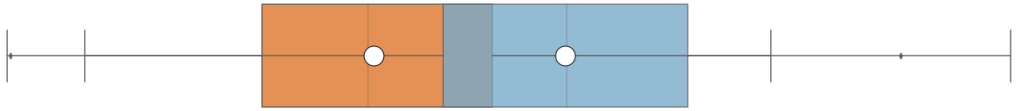
are these distributions the same?



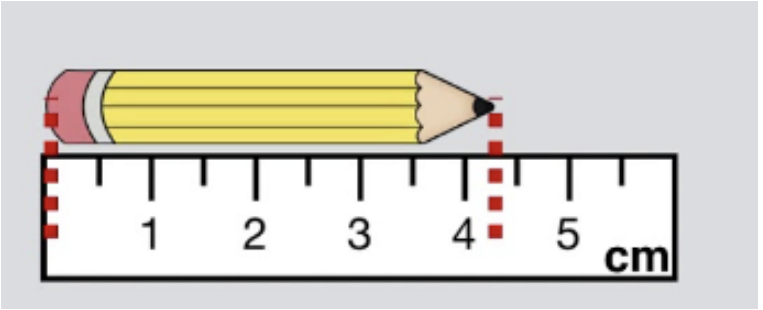
if distributions have the same measured means within 1 (or *n*) standard deviation they should be considered "the same"

21 measurement uncertainty and measurement samples

- Are two measurements the same?

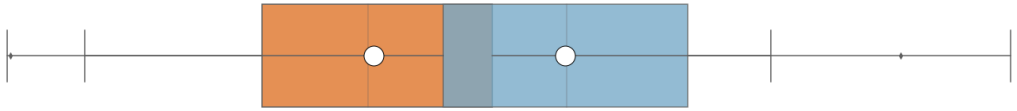


  - *is the amount of rain in Wilmington this year the same as last year?*
  - *are two pencils the same length*



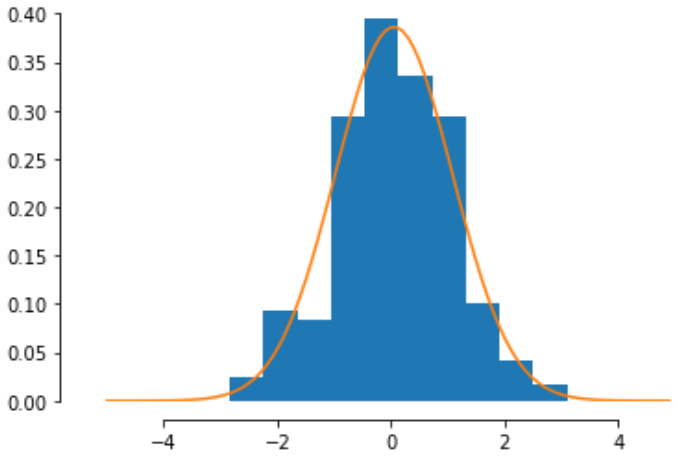
uncertainty because of the limitations of the measuring tool

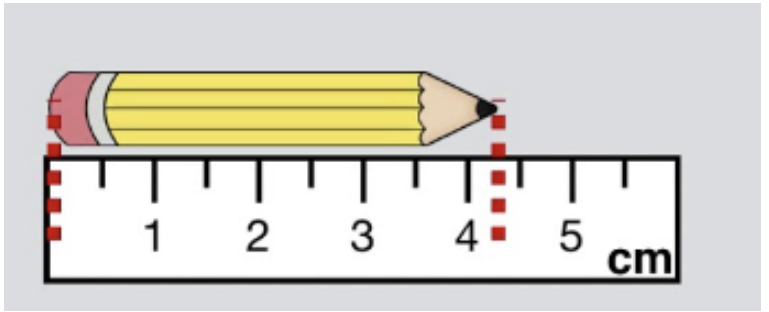
- Are two measurements the same?



  - *is the amount of rain in Wilmington this year the same as last year?*
  - *are two pencils the same length*

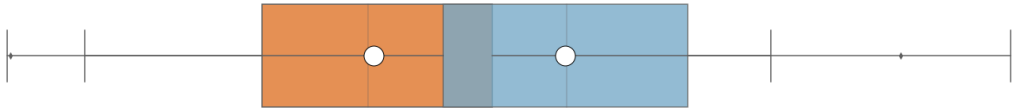
Take N measurements,
they will all be a bit different





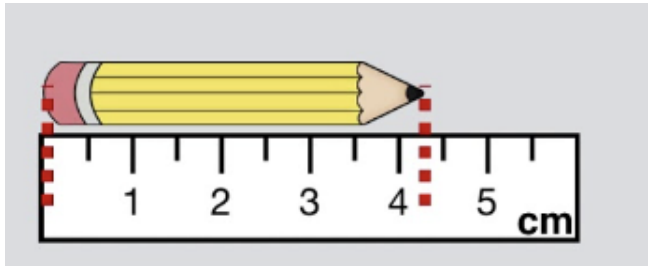
uncertainty because of the limitations of the measuring tool

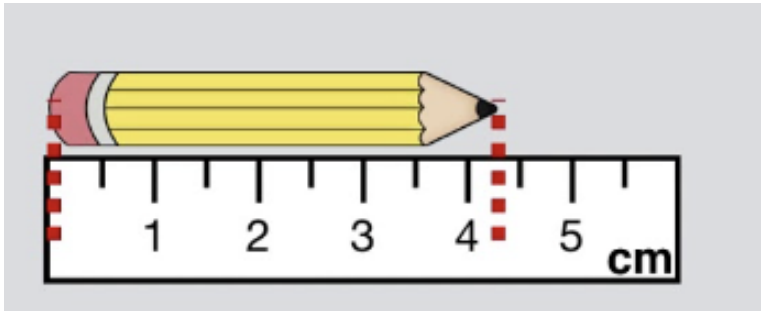
- Are two measurements the same?

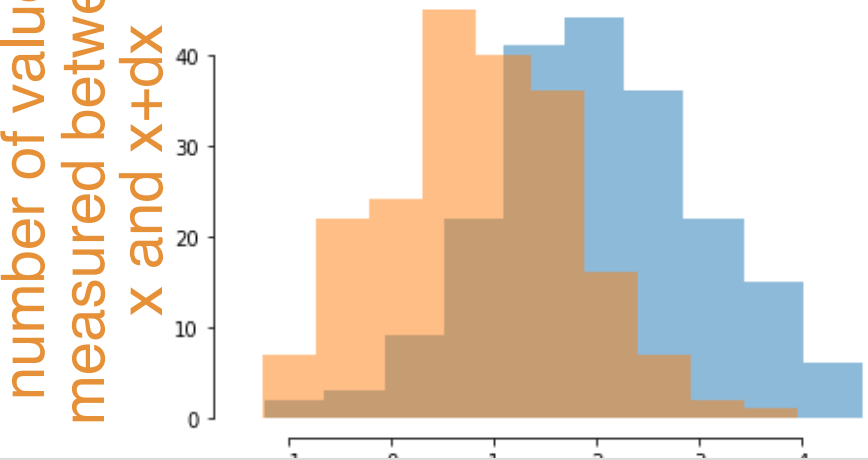


  - *is the amount of rain in Wilmington this year the same as last year?*
  - *are two pencils the same length*

Take N measurements,
they will all be a bit different

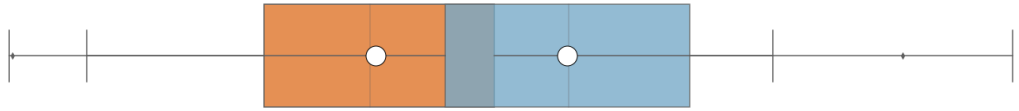






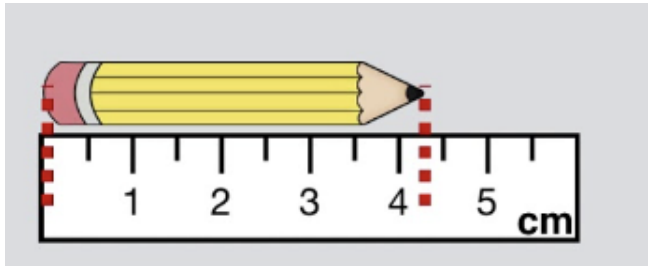
uncertainty because of the limitations of the measuring tool

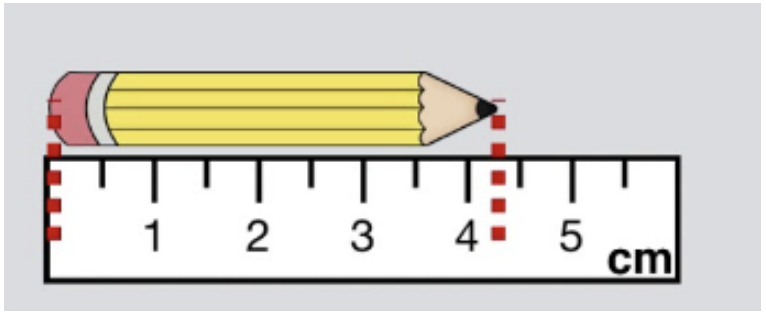
- Are two measurements the same?

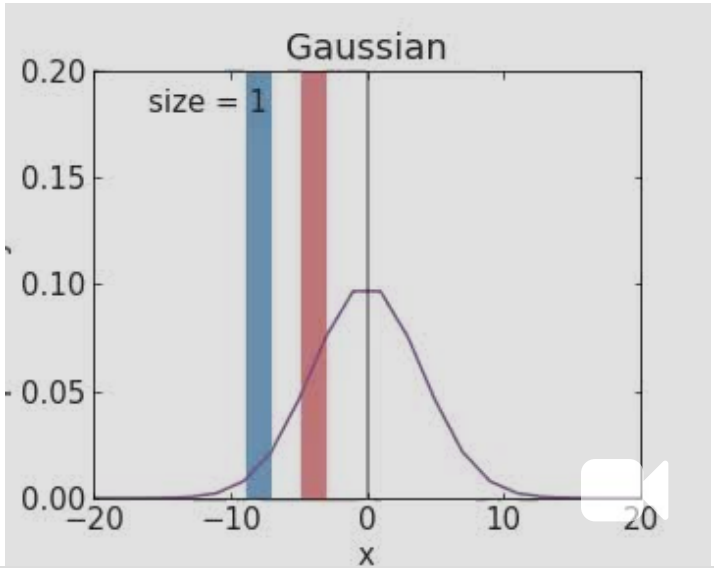


  - *is the amount of rain in Wilmington this year the same as last year?*
  - *are two pencils the same length*

Take N measurements,
they will all be a bit different







number of values measured between x and x+dx

uncertainty because of the limitations of the measuring tool

Methods and philosophy [edit]

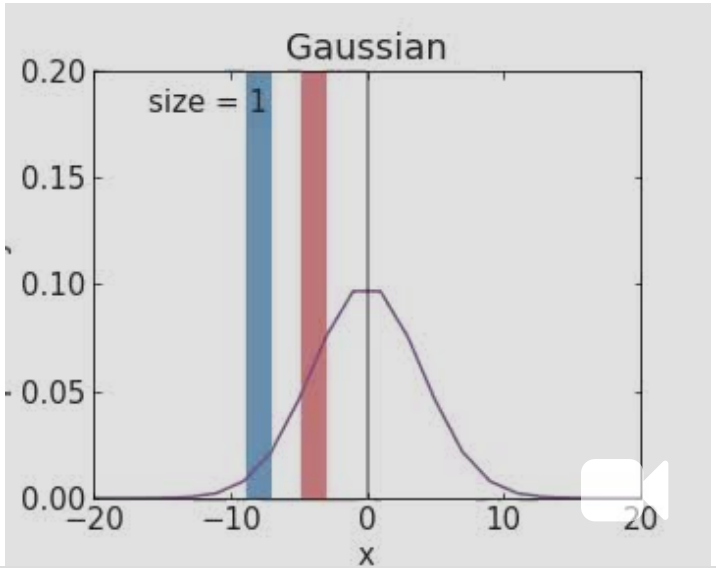
The rankings of national happiness are based on a Cantril ladder survey. Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale.[15] The

wikipedia

- Are two measurements the same?
  - *are two countries just as happy?*
  - *are two pencils the same length?*

Take N measurements,
they will all be a bit different

Country	Region	Happiness Score	Standard Error
Denmark	Western Europe	7.527	0.03328
Norway	Western Europe	7.522	0.03880



number of values measured between x and x+dx

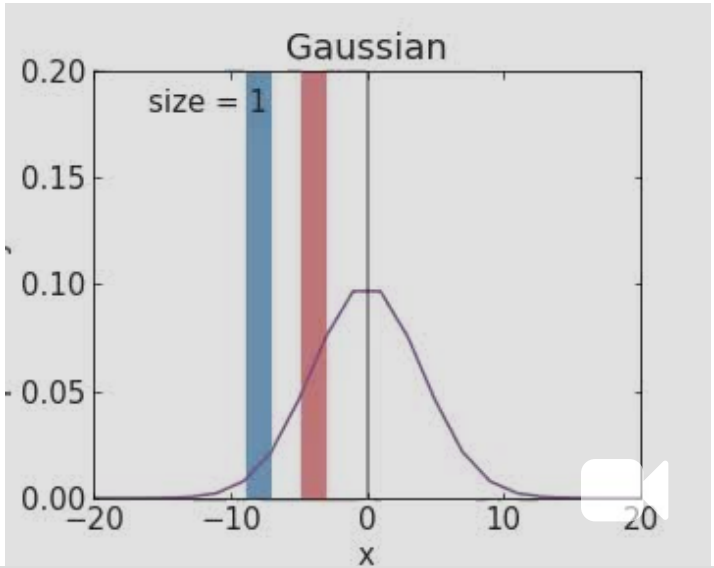
intrinsic variance in the phenomenon

KEY CONCEPT:

the larger the number of samples from the distribution the more similar the distribution of our sample is to the actual "generative process": i.e. the histogram will look more and more like the actual distribution curve

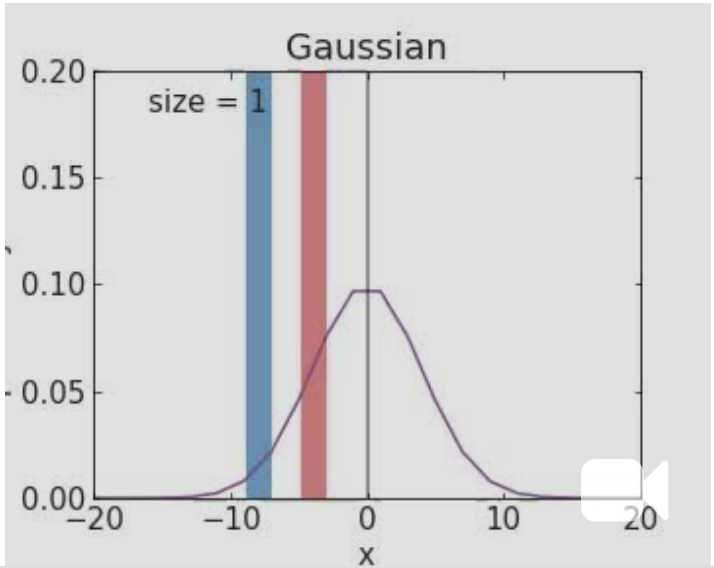
Take N measurements, they will all be a bit different

number of values measured between x and x+dx



Take N measurements,
they will all be a bit different

number of values measured between x and x+dx



KEY CONCEPT:

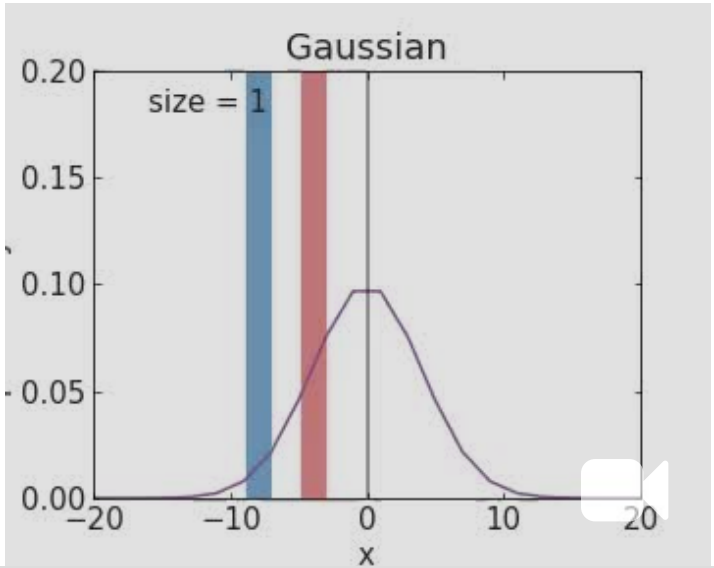
the larger the number of samples from the distribution the more similar the distribution of our sample is to the actual "generative process": i.e. the histogram will look more and more like the actual distribution curve

THEREFORE

It is easier to tell if two distributions are the same when the samples are large

Take N measurements,
they will all be a bit different

number of values measured between x and x+dx



KEY CONCEPT:

the larger the number of samples from the distribution the more similar the distribution of our sample is to the actual "generative process": i.e. the histogram will look more and more like the actual distribution curve

THEREFORE

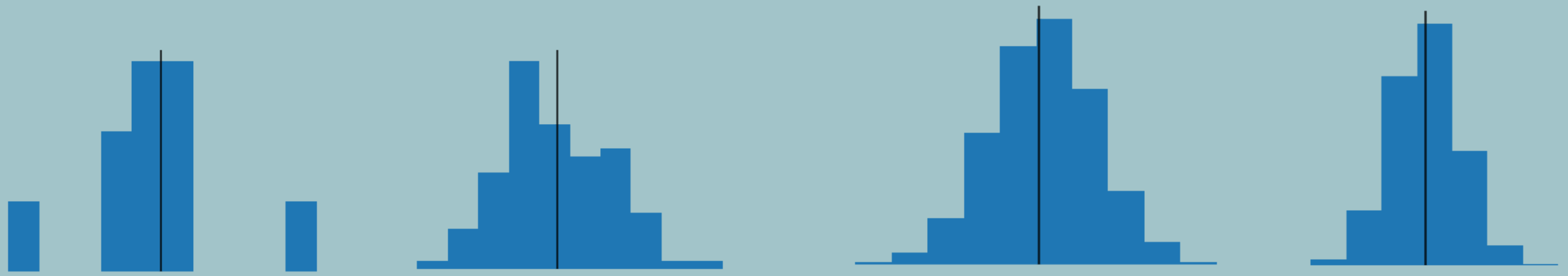
It is easier to tell if two distributions are the same when the samples are large

THEREFORE

I need statistical tests that acknowledge the size of the sample when I compare distributions

Law of Large Numbers

As the size of a _____ tends to infinity the mean of the sample tends to the mean of the _____



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

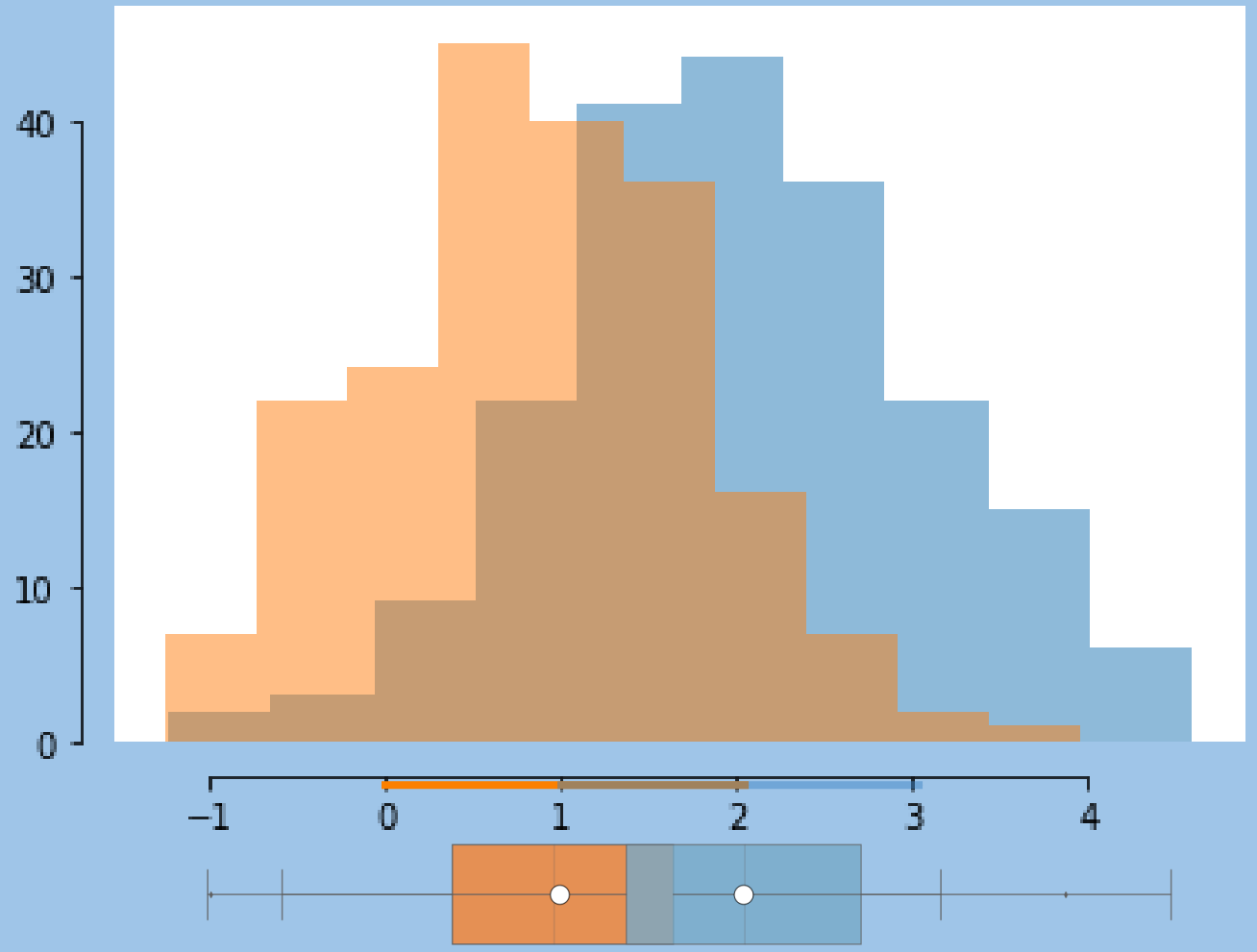
the sample mean $\bar{x}$ approaches a Normal (Gaussian) distribution with mean μ and standard deviation σ

regardless of the distribution of X

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

Easy way to assess if two number are the same:

Are the mean farther than the standard deviations?



standard deviation: 1,1

standard dev.

interquartile range (25%-75%)

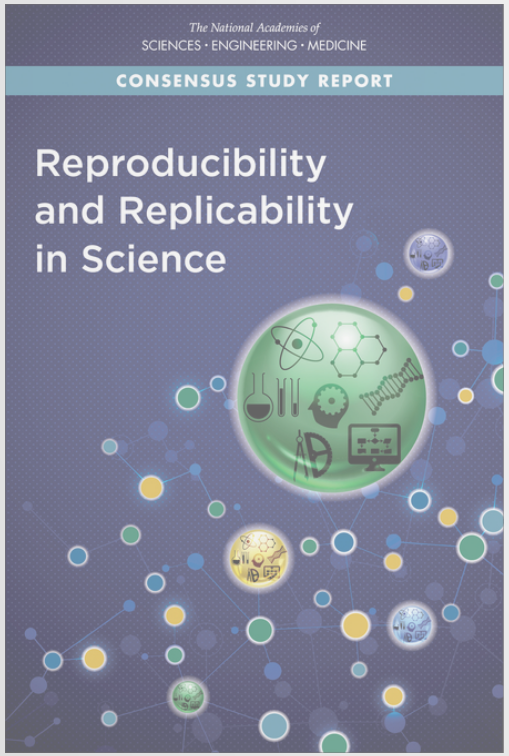
mean

3

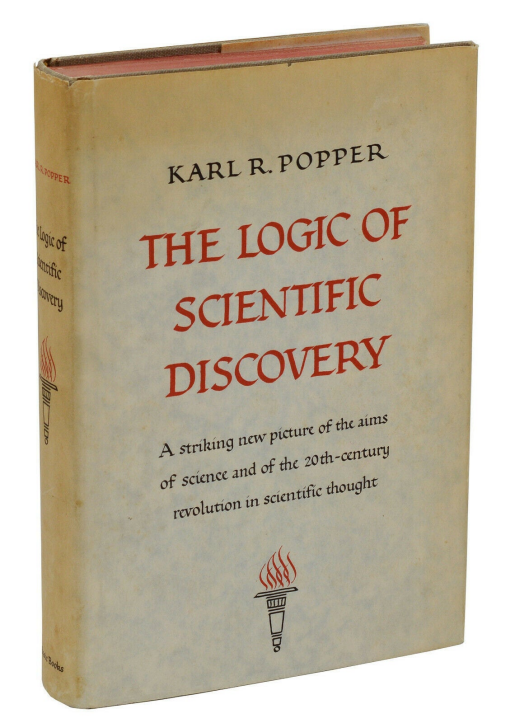
the principle of Falsifiability

3 General principles of "good" science

Reproducibility



Falisifiability



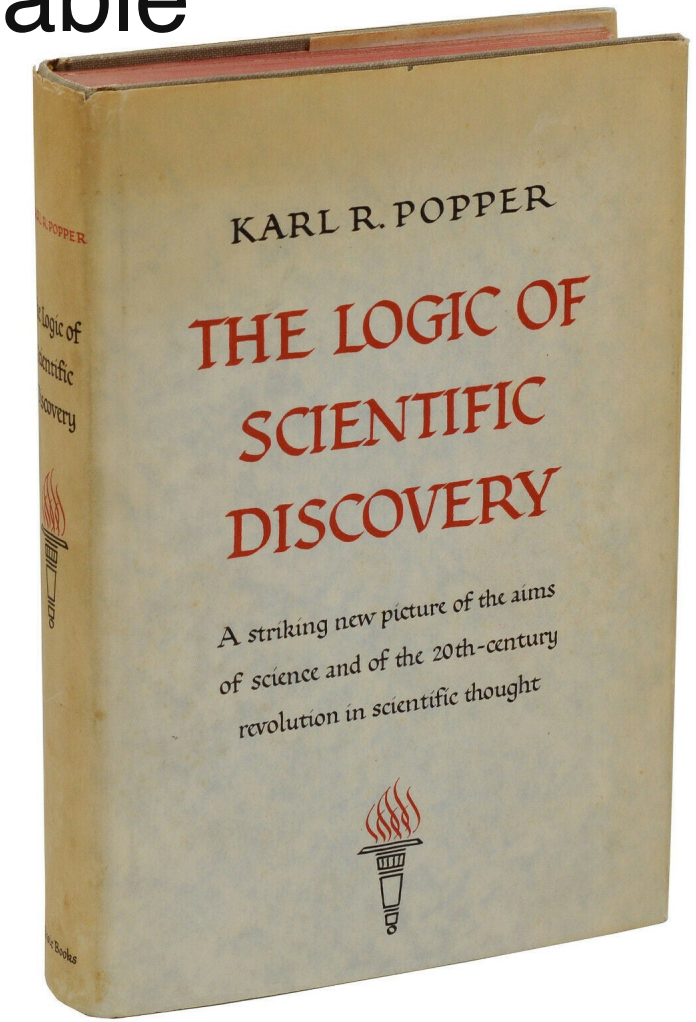
Parsimony

the *demarcation* problem:

science hypotheses needs to be falsifiable

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*



the *demarcation* problem:

science hypotheses needs to be falsifiable

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.

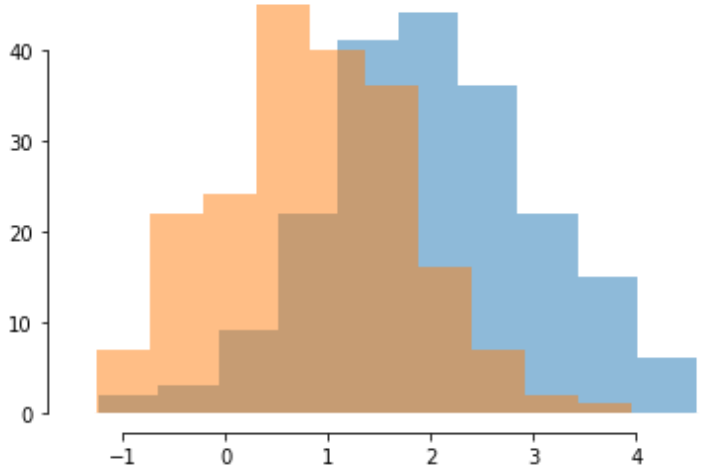
—Karl Popper, *The Logic of Scientific Discovery*

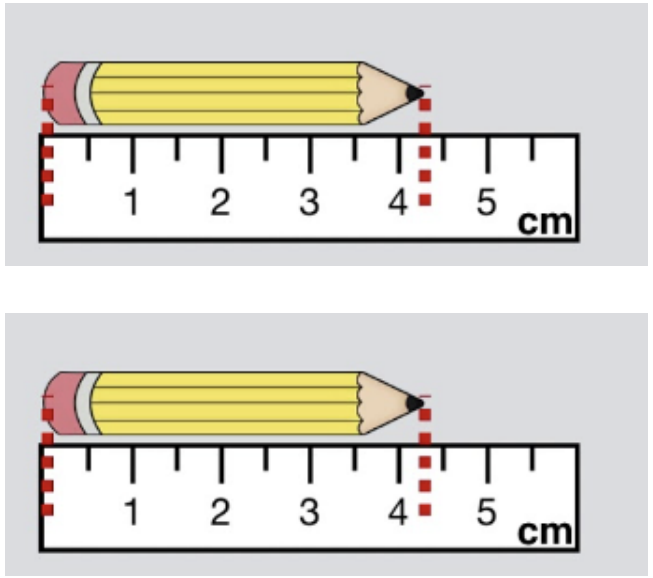
I need to see only 1 black swan to tell that the statement that all swans are white is not true. But even if I dont see a black one it does not mean all swans are white

the *demarcation* problem:

science hypotheses needs to be falsifiable

But what happens when I have distributions of measurements?





Beyond any reasonable doubt

same concept guides prosecutorial justice

guilty beyond reasonable doubt

in a probabilistic sense, all hypotheses we make are possible

We will reject a hypothesis if its probability is lower than a predefined threshold

4 hyposthesis testing

4 hyposthesis testing

We do not "prove our hypothesis"

we falsify the opposite of our hypothesis

Null
Hypothesis
Rejection
Testing

The NULL hypothesis is typically what I want to reject: its the way I think the world does NOT work

the pencils are the NOT same length (tho i think they are)

the earth is NOT round (spoiler alert... it is!!)

1 formulate your prediction

Null Hypothesis

Null
Hypothesis
Rejection
Testing

The ALTERNATIVE hypothesis is the complete opposite of the NULL

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

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

2 identify all alternative outcomes

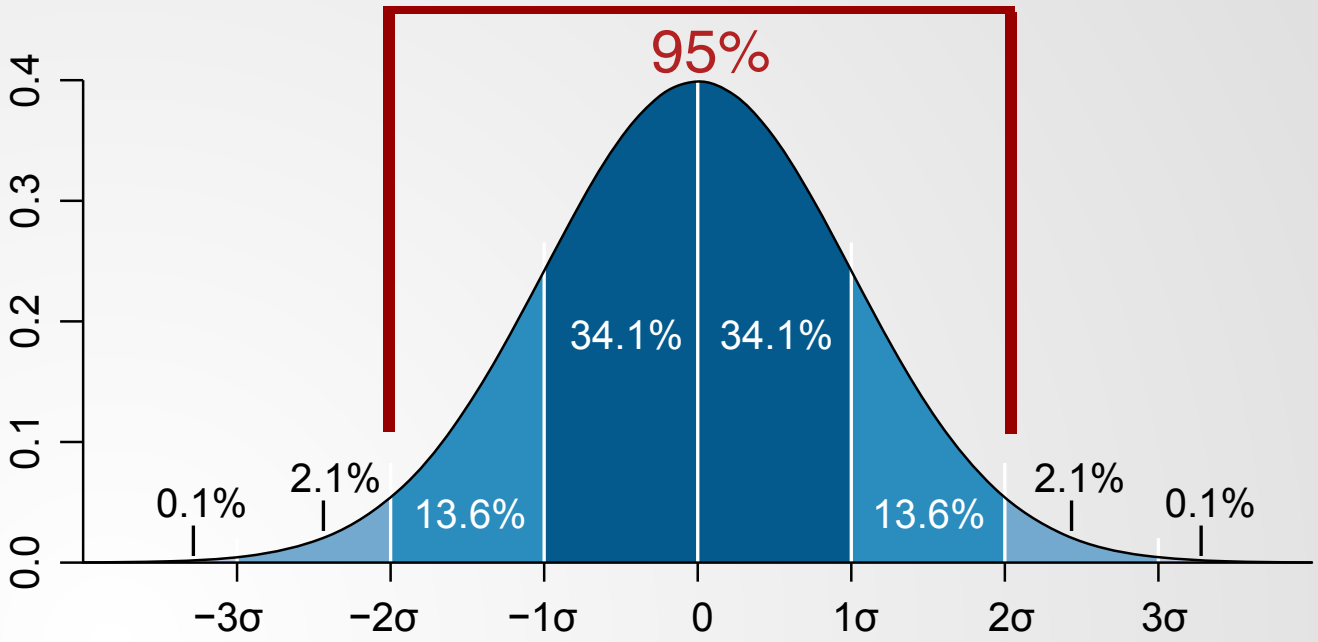
Alternative Hypothesis

Null

Hypothesis

Rejection

Testing



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

If I measure the mean of two samples

(the samples of pencil measurements)

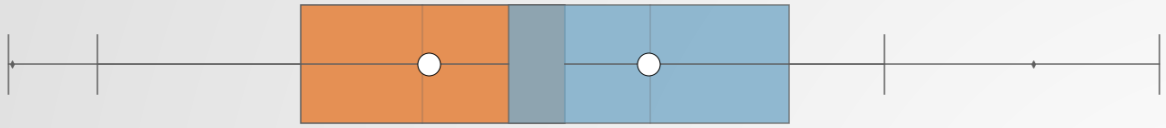
I expect the difference to be a number drawn from a standard normal:

Highest prob. is 0

Prob that the number is within 1σ of the mean is 68%

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

Z-test



If the means are within 1-sigma of each other that means that I cannot rule out that the distributions are the same to a 1-sigma level (p-value 0.32)

If the means are not 1-sigma of each other that means that I can rule out that the distributions are the same to a 1-sigma level (p-value 0.32)

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

Z-test

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

pivotal quantity

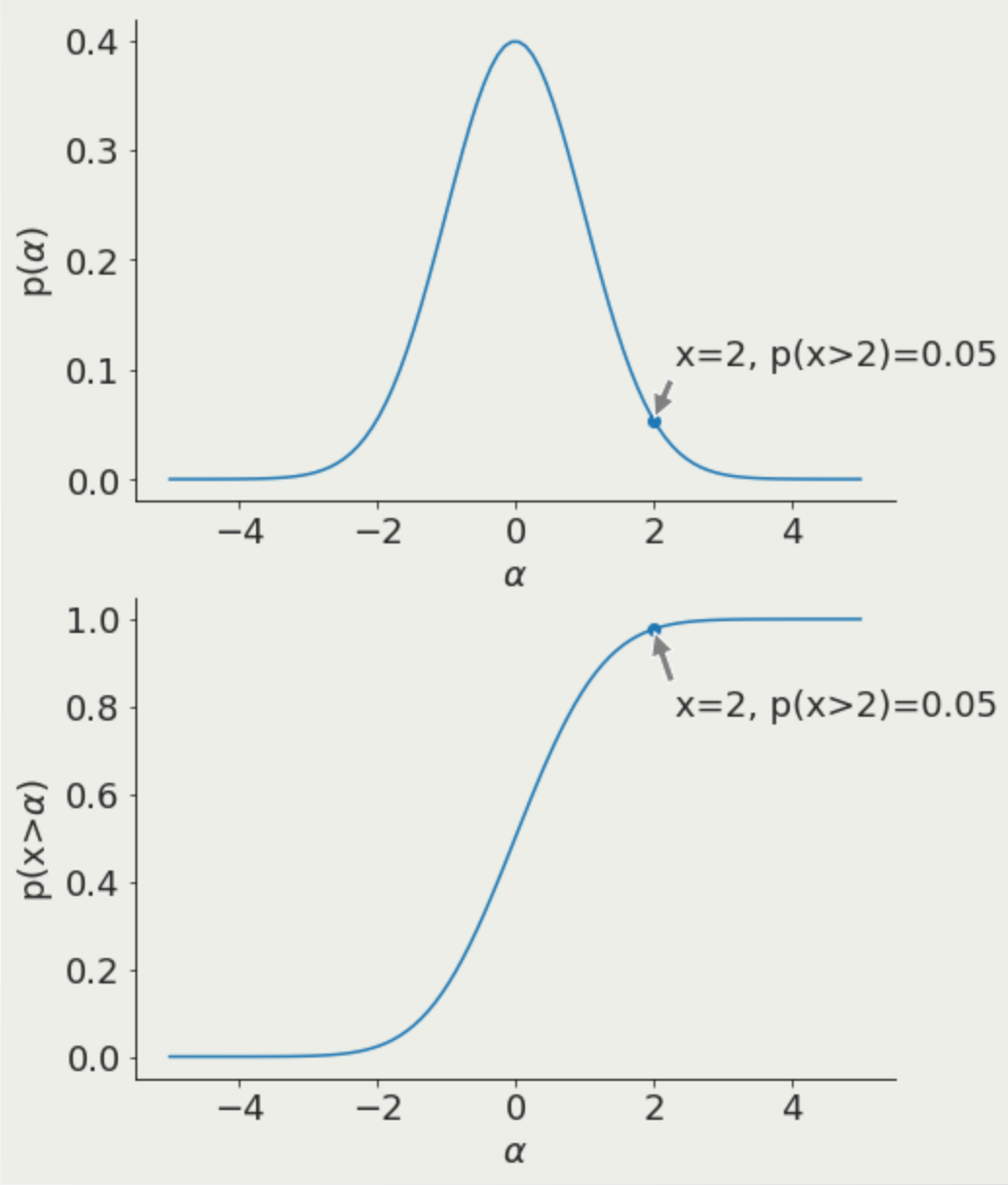
$$Z = (\bar{X} - \mu_0)/s$$

sample mean

population mean

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

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



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

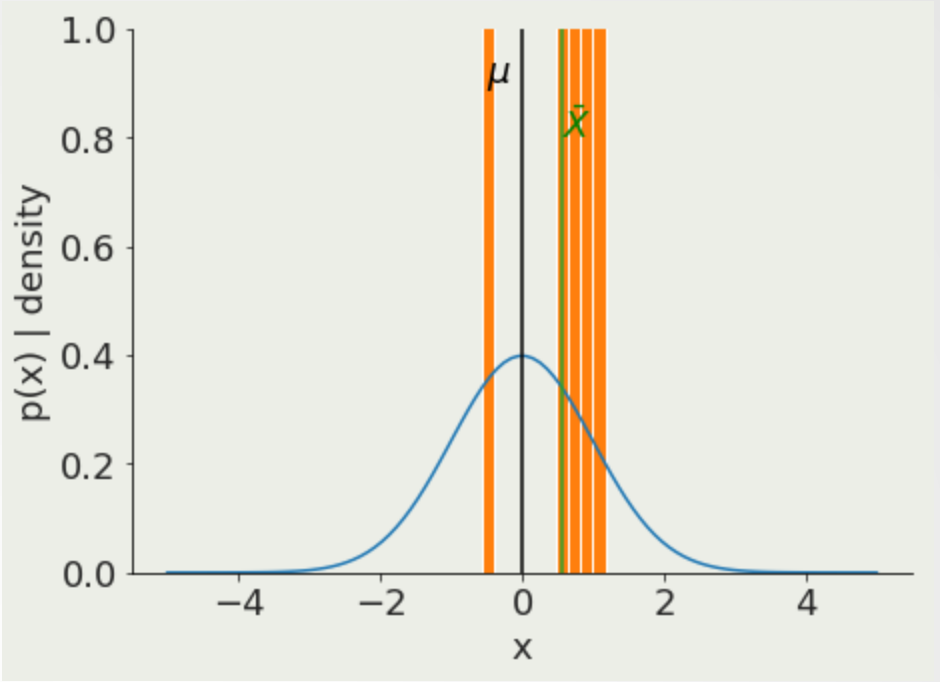
$$Z = (\bar{X} - \mu_0)/s$$

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?

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

Key Slide

if probability < *p*-value : reject Null

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 prediction 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

4

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

Null

Hypothesis

Rejection

Testing

pivotal quantities

5

calculate it!

Null

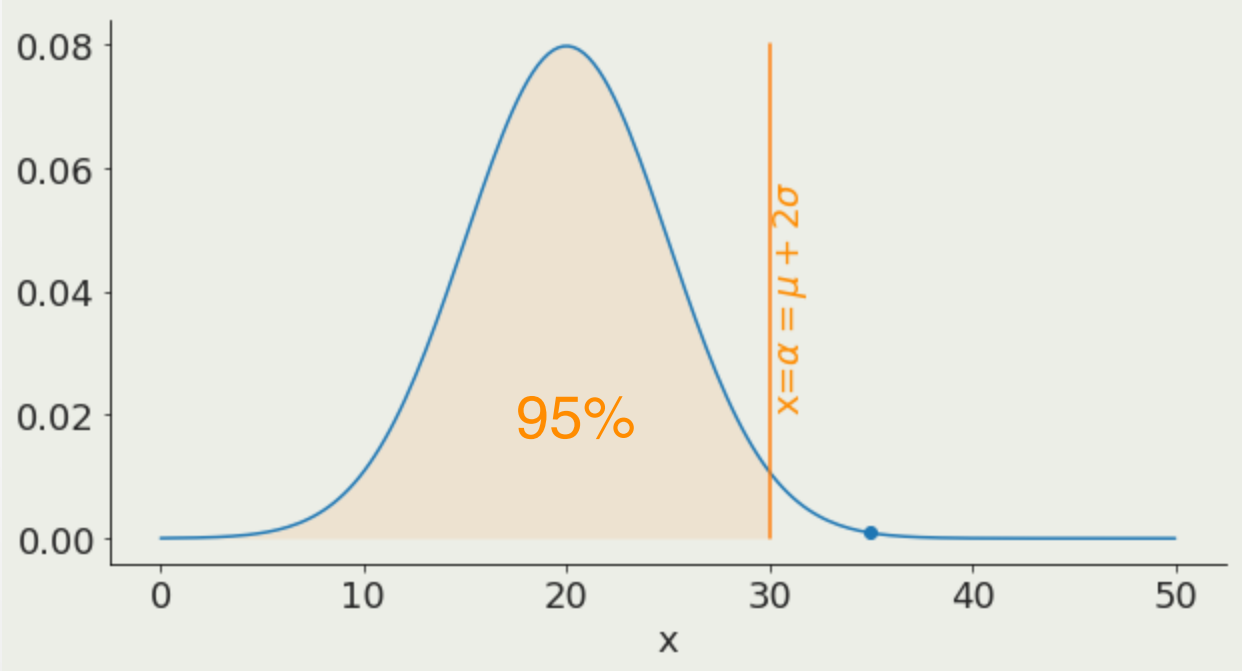
Hypothesis

Rejection

Testing

what is α?

α is the x value corresponding to a chosen threshold



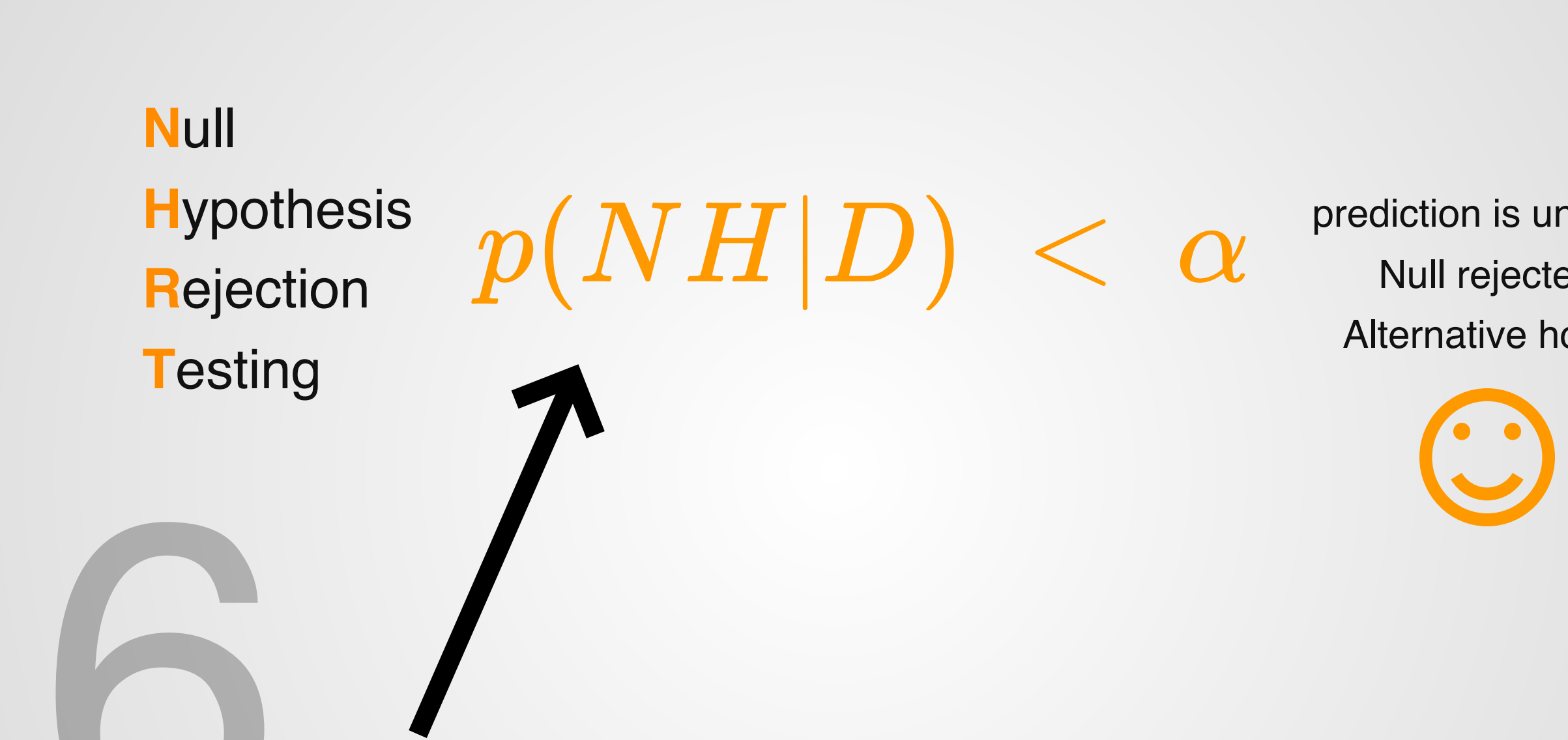
6

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

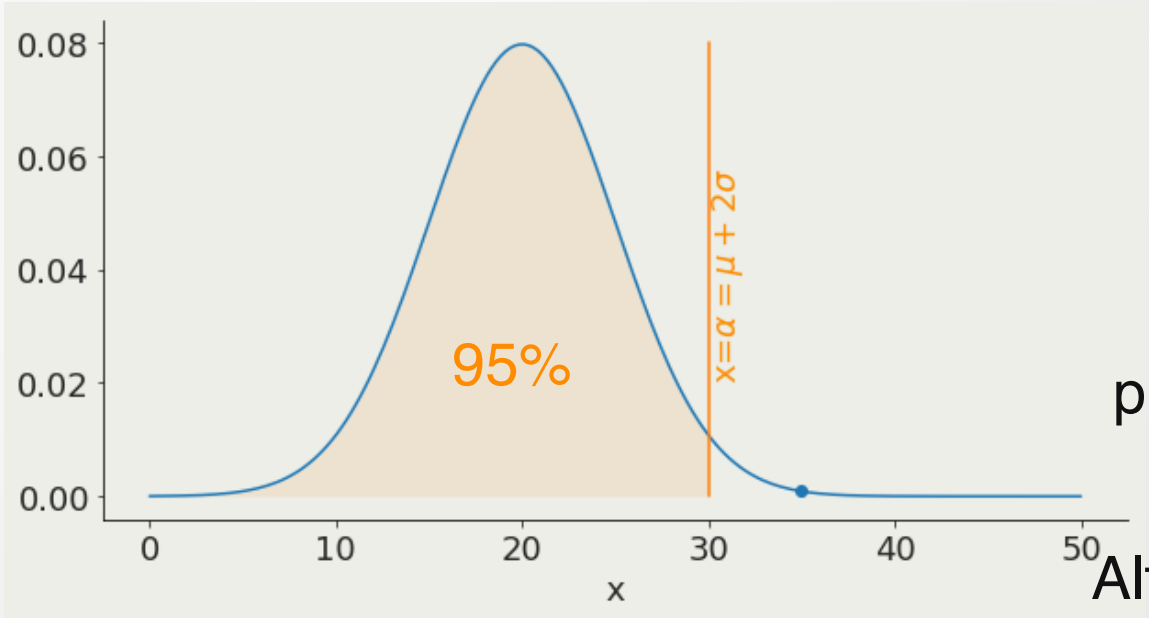
Null
Hypothesis
Rejection
Testing

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

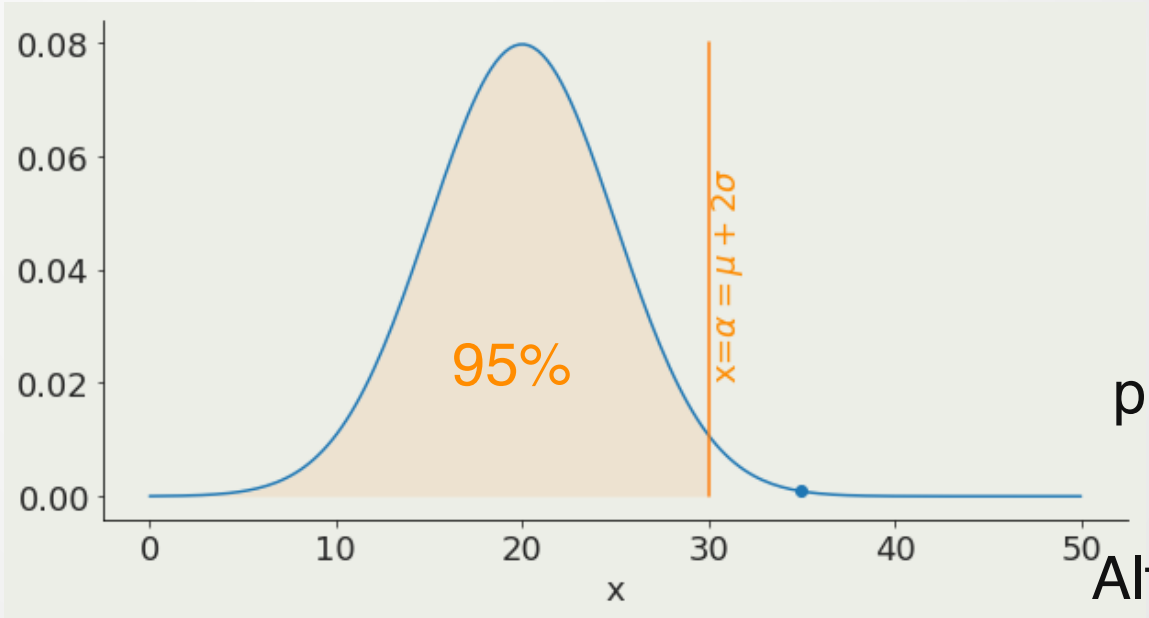
Null
Hypothesis
Rejection
Testing

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

K-S test

Kolmogorof-Smirnoff :

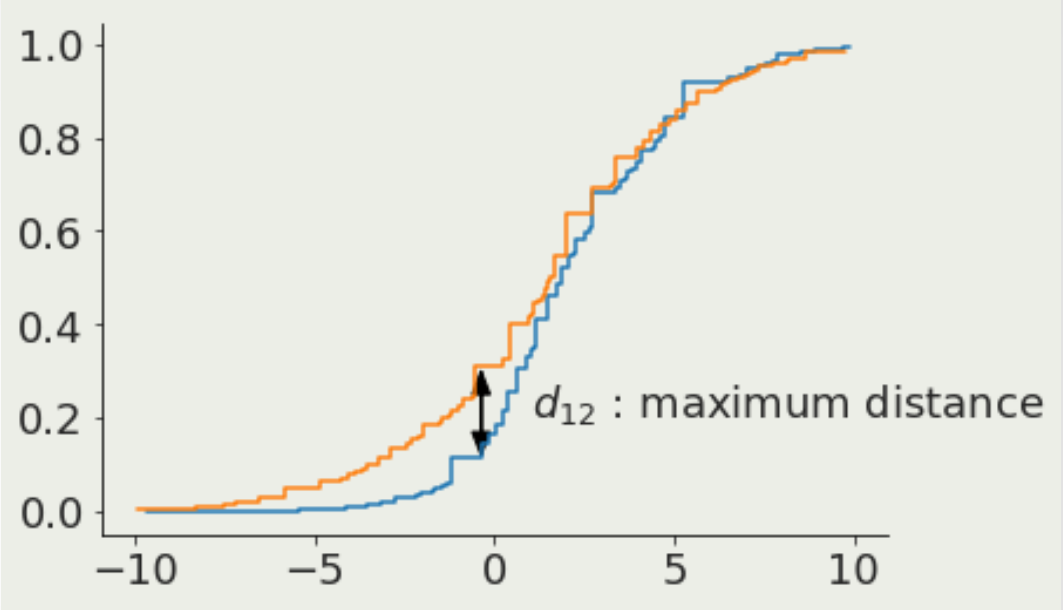
do two samples come from the same parent distribution?

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 :

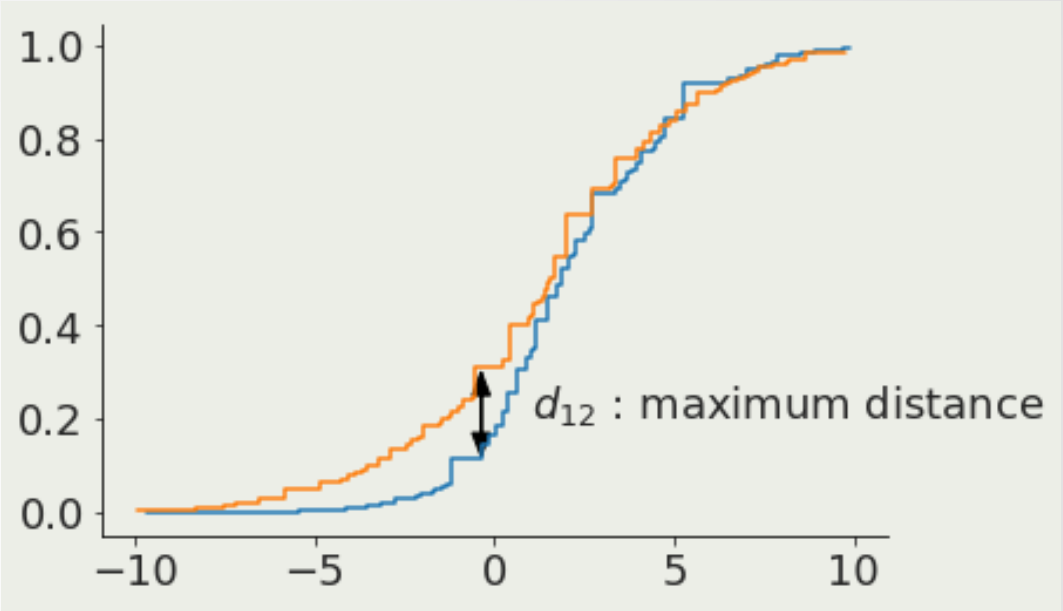
do two samples come from the same parent distribution?

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)
```

5

p-value hypothesis testing

Moments and frequentist probability

Imagine that I take a measurements of a quantity that is expected to be normally distributed with mean 0 and stdev 1

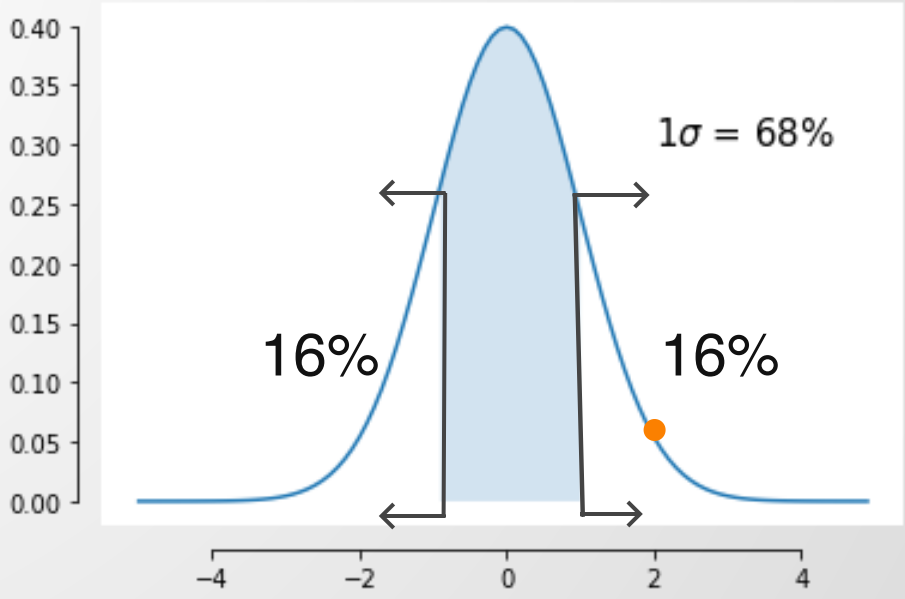
what is the probability that I would measure 1.5?

The probability of measuring any one value is mathematically *0*... however I can say that

the probability of measuring something between *-1σ* and *1σ* (within 1-sigma) is 68%.

So the probability of measuring something outside is 100-68 = 32%.

So if I measure something outside of [-*1σ*:*1σ*] that had a probability <32% of being measured.



Moments and frequentist probability

Imagine that I take a measurements of a quantity that is expected to be normally distributed with mean 0 and stdev 1

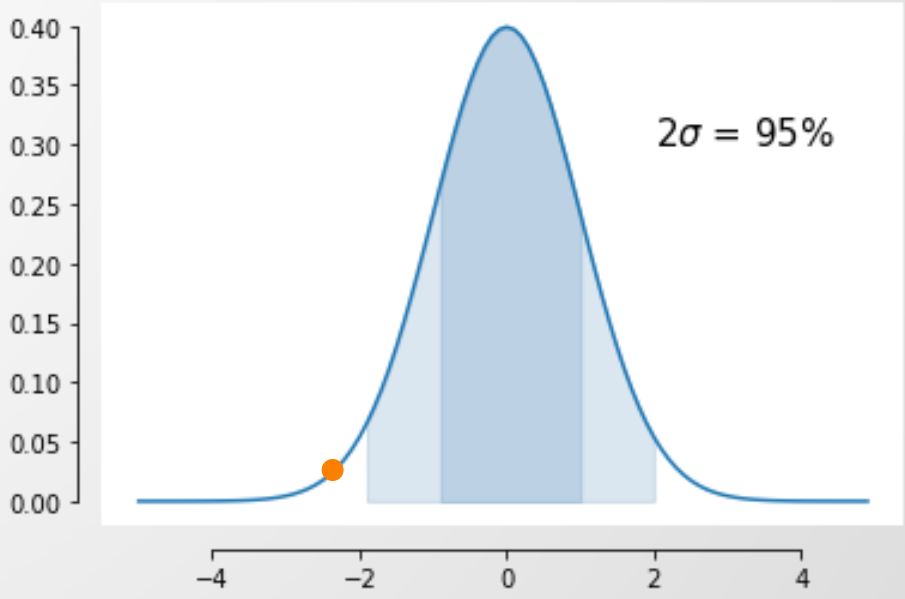
what is the probability that I would measure 1.5?

The probability of measuring any one value is mathematically *0*... however I can say that

the probability of measuring something between *-2σ* and *2σ* (within 2-sigma) is 95%.

So the probability of measuring something outside is 100-95 = 5%.

So if I measure something outside of [-2σ:2σ] that had a probability <5% of being measured.



Moments and frequentist probability

Imagine that I take a measurements of a quantity that is expected to be normally distributed with mean 0 and stdev 1

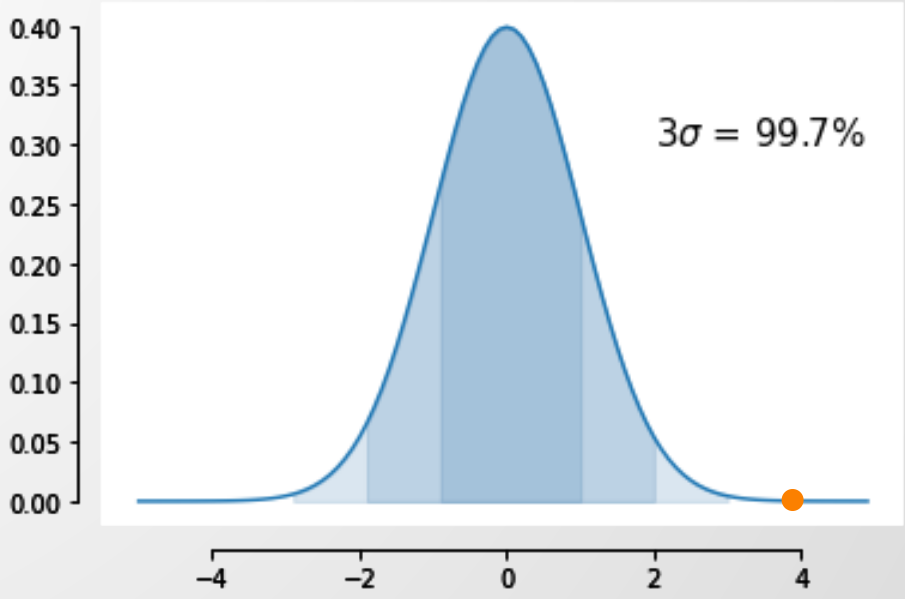
what is the probability that I would measure 1.5?

The probability of measuring any one value is mathematically *0*... however I can say that

the probability of measuring something between *-3σ* and *3σ* (within 3-sigma) is 99.7%.

So the probability of measuring something outside is 100-99.7 = 0.3%.

So if I measure something outside of [-3σ:3σ] that had a probability <0.3% of being measured.



Moments and frequentist probability

Imagine that I take a measurements of a quantity that is expected to be normally distributed with mean 0 and stdev 1

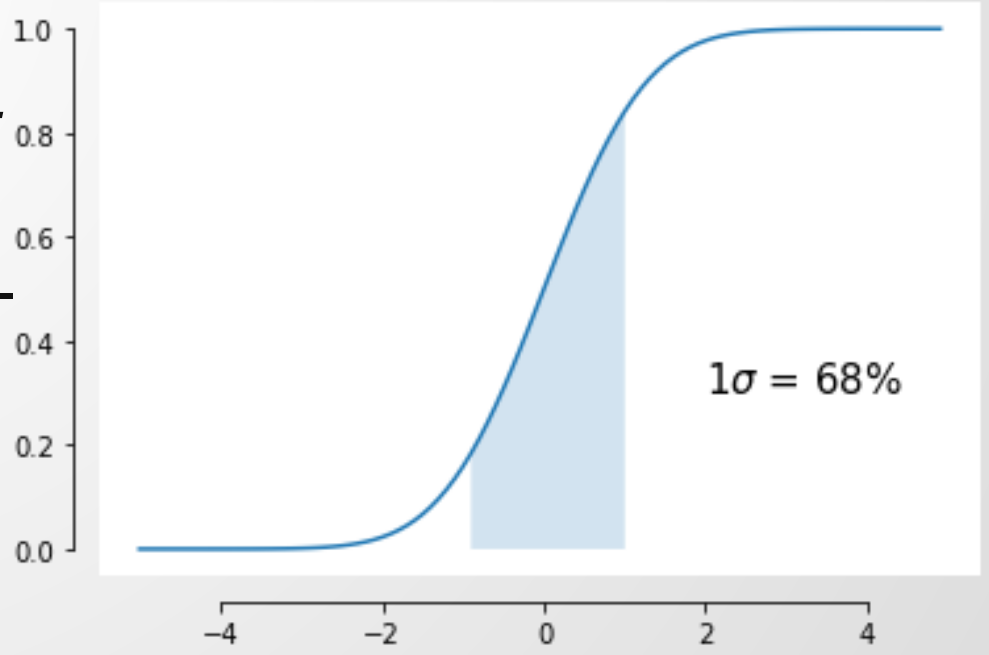
what is the probability that I would measure 1.5?

it might be easier to think about it as cumulative distributions if you are comfortable with integrals

the probability of measuring something between *-3σ* and *3σ* (within 3-sigma) is 99.7%.

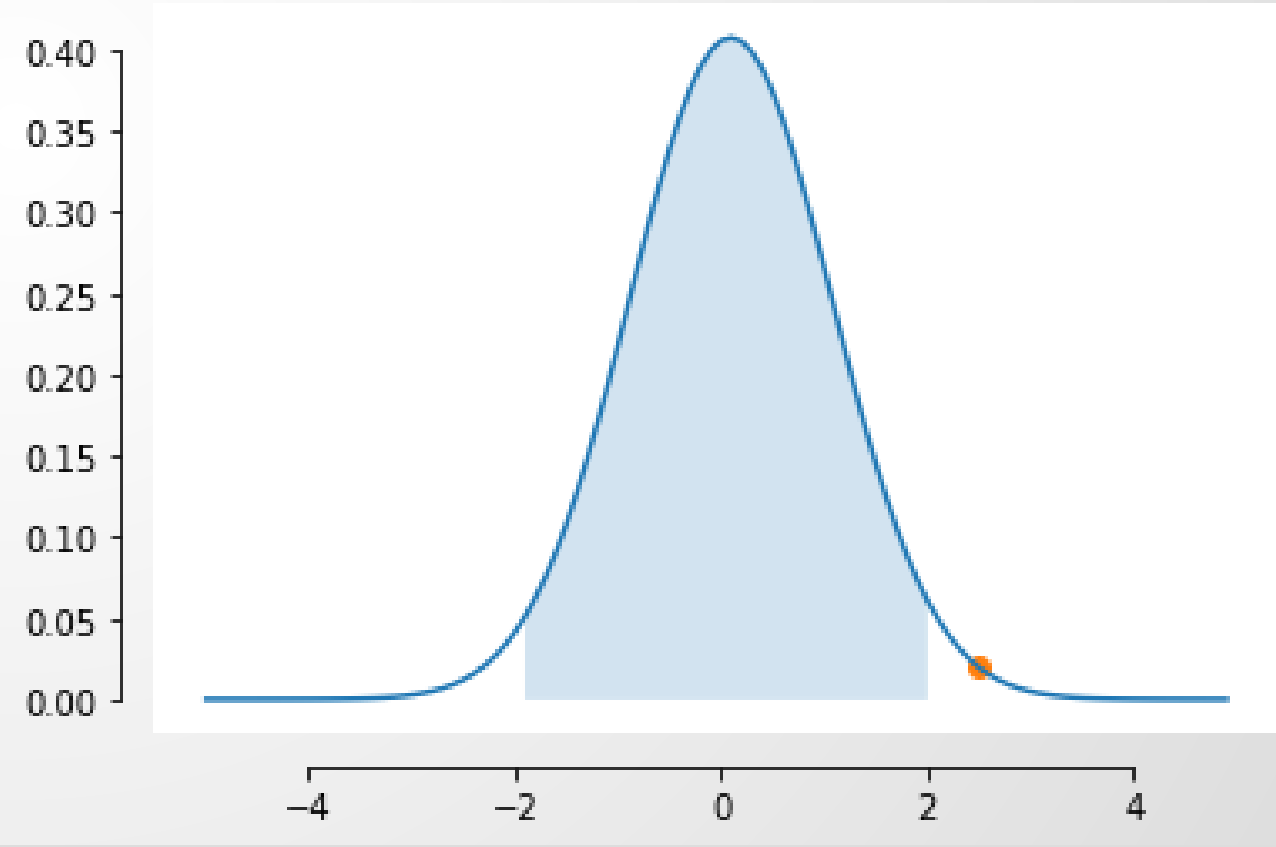
So the probability of measuring something outside is 100-99.7 = 0.3%.

So if I measure something outside of [-3σ:3σ] that had a probability <0.3% of being measured.



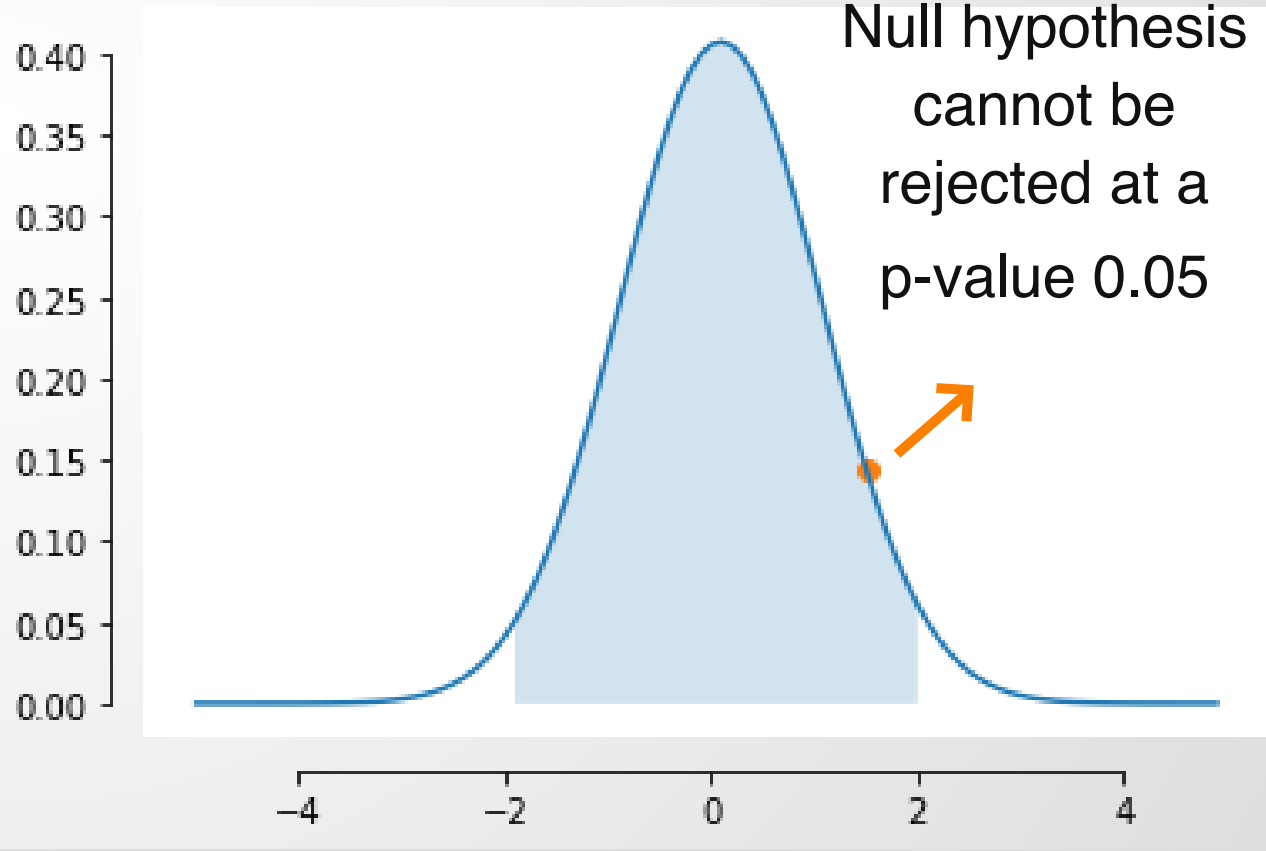
Moments and frequentist probability in the *falsification* framework: *p*-value

1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**
3. Measure your outcome from the data ***x***
4. If ***x*** is outside of the area of "reasonable doubt" under the Null hypothesis => the null hypothesis is rejected at ***p-value = α***, otherwise the Null cannot be rejected.



Moments and frequentist probability in the *falsification* framework: *p*-value

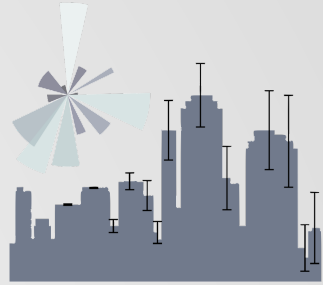
1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**
3. Measure your outcome from the data ***x***
4. If ***x*** is outside of the area of "reasonable doubt" under the Null hypothesis => the null hypothesis is rejected at ***p-value = α***, otherwise the Null cannot be rejected.



6

p-value hypothesis testing step by step

NHRT: *p*-value
Null Hypothesis Rejection Testing

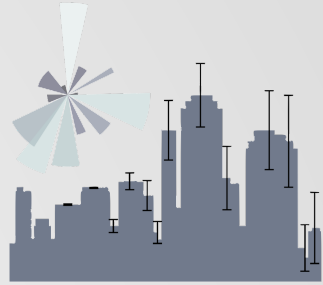


1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05

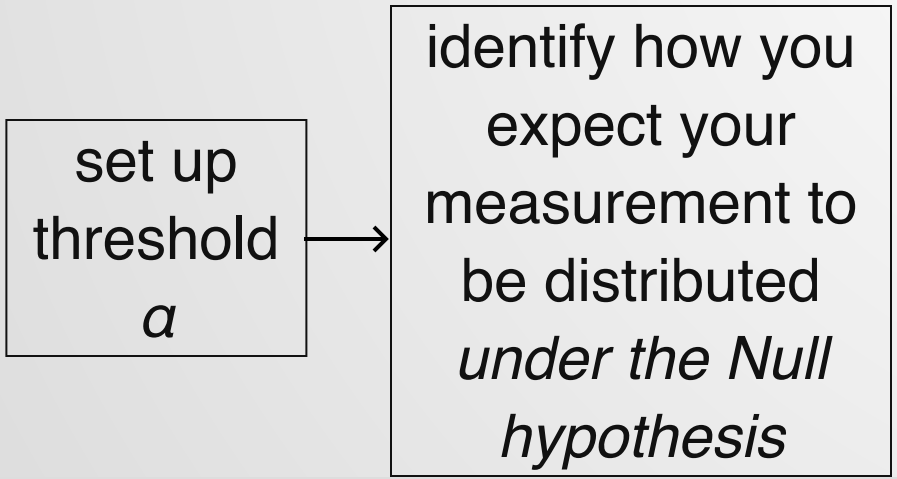
its important to do this first. If we do not we may be tempted to choose a threshold that fits our result, thus always reporting rejection of null hypothesis

set up threshold α

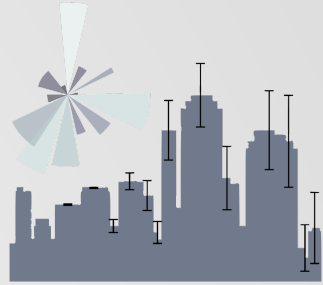
NHRT: *p*-value
Null Hypothesis Rejection Testing



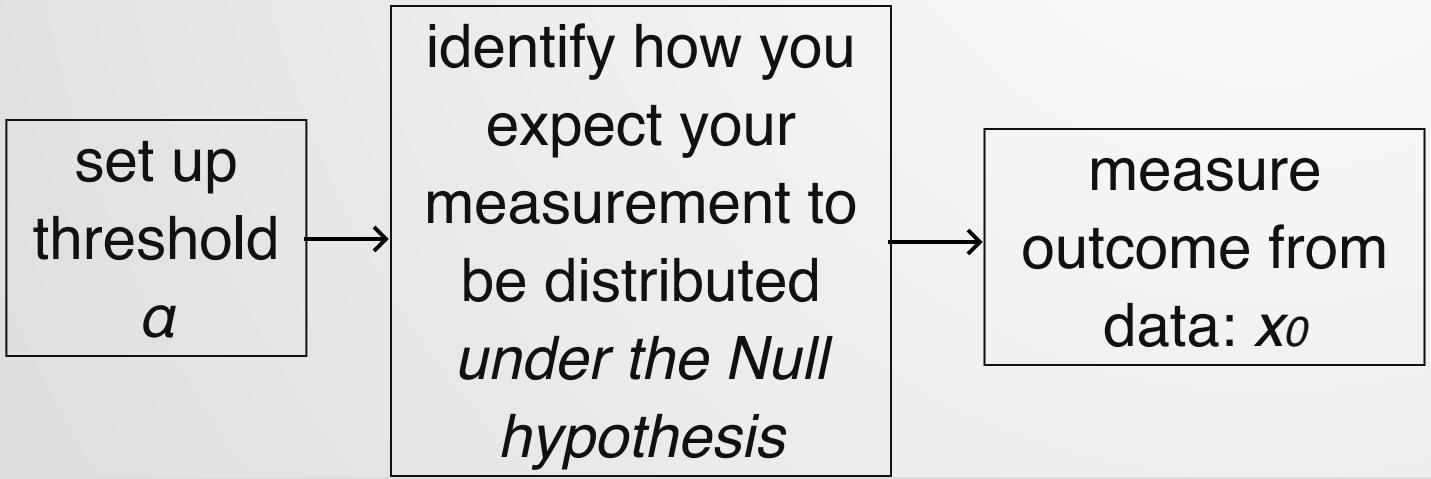
1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**



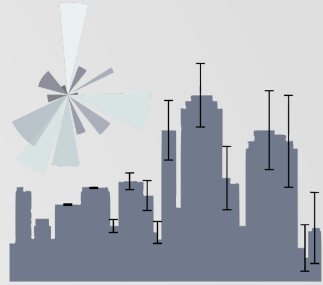
NHRT: *p*-value
Null Hypothesis Rejection Testing



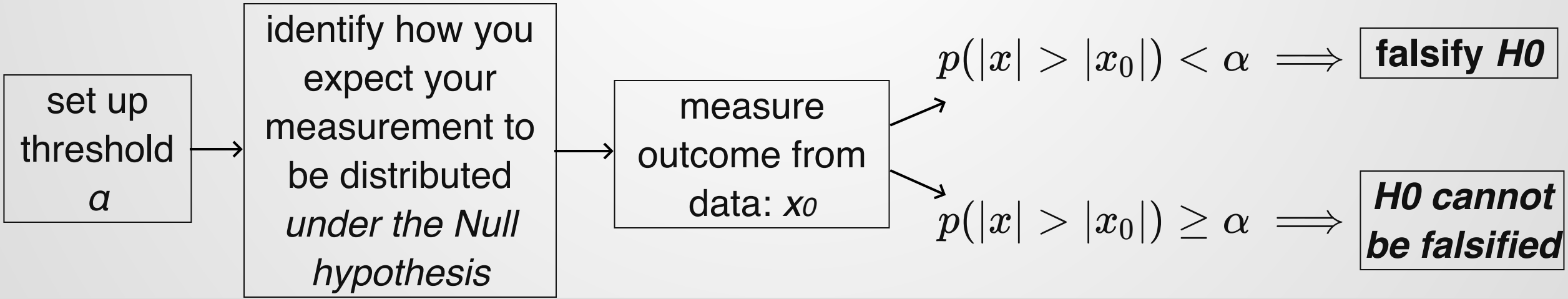
1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**
3. Measure your outcome from the data ***x***; extract the appropriate statistics from a set of data (e.g. mean, median...)



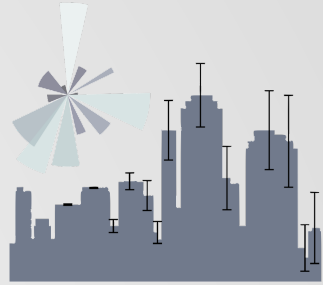
NHRT: *p*-value
Null Hypothesis Rejection Testing



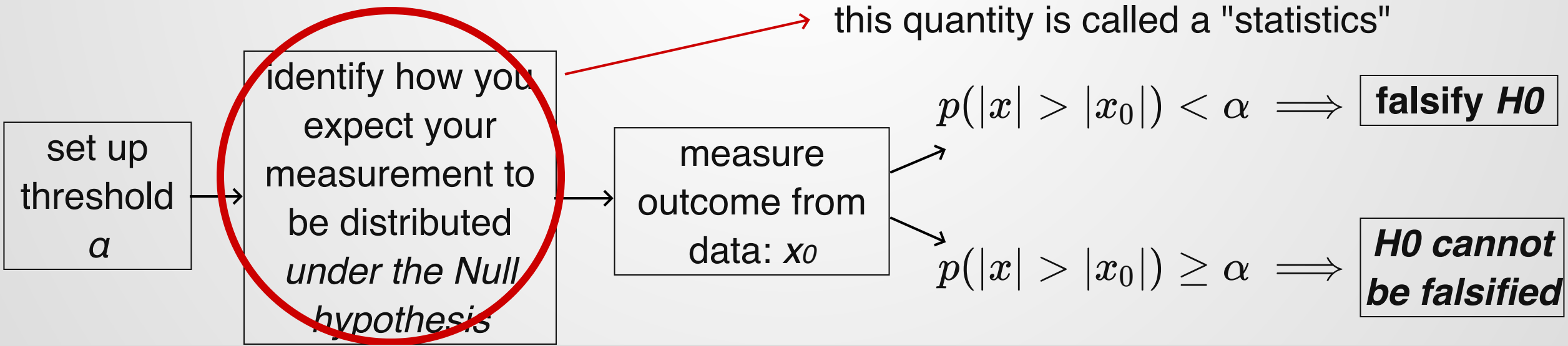
1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**
3. Measure your outcome from the data ***x***; extract the appropriate statistics from a set of data (e.g. mean, median...)
4. If ***x*** is outside of the area of "reasonable doubt" under the Null hypothesis the null hypothesis is rejected at ***p-value = α***, otherwise the Null cannot be rejected.



NHRT: *p*-value
Null Hypothesis Rejection Testing



1. Set a threshold you believe corresponds to "reasonable doubt" 95% => α=0.05
2. Identify what you expect your measurement's distribution to be **if the Null hypothesis holds**
3. Measure your outcome from the data ***x***; extract the appropriate statistics from a set of data (e.g. mean, median...)
4. If ***x*** is outside of the area of "reasonable doubt" under the Null hypothesis the null hypothesis is rejected at ***p-value = α***, otherwise the Null cannot be rejected.

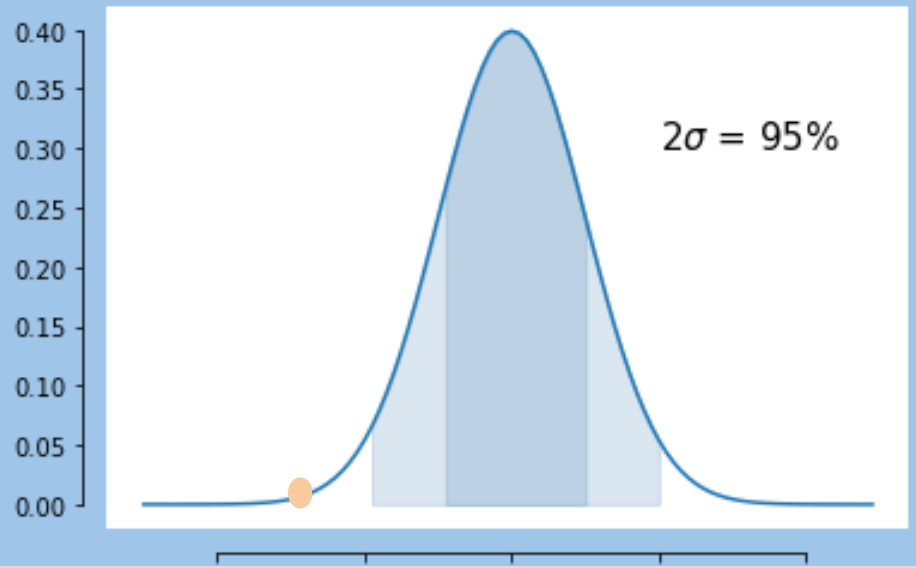


7

p-value hypothesis testing common tests

Statistical way to measure differences:

In NHRT a statistics is a quantity that relates to the data which has a known distribution under the Null Hypothesis



In NHRT a statistics is a quantity that relates to the data which has a known distribution under the Null Hypothesis

e.g.: Z statistics is Normally distributed

$Z \sim N(0,1)$

Statistics that follow a Standard Normal distribution

Z -test

Does a sample come from a known population?

$$Z = \frac{\mu - \bar{x}}{\sigma / \sqrt{N}}$$

In absence of effect (i.e. under the Null)

== the sample mean is the same as the population mean

Z is distributed according to a Gaussian N(μ=0, σ=1)

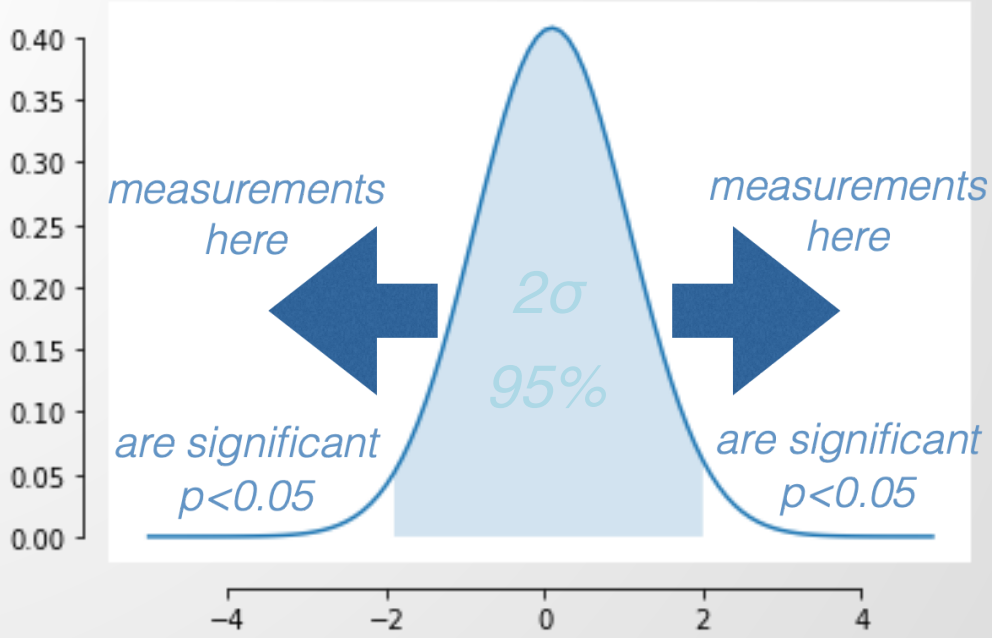
Example: new bus route implementation.

https://github.com/fedhere/PUS2022_FBianco/blob/master/classdemo/ZtestBustime.ipynb

You know the mean and standard deviation of a but travel route: that is the population

You measure the new travel time between two stops 10 times: that is your sample.

Has travel time changed?



Statistics that follow a Standard Normal distribution

Z -test

Does a sample come from a known population?

How to interpret the number you get?

$$Z = \frac{\mu - \bar{x}}{\sigma / \sqrt{N}}$$

In absence of effect (i.e. under the Null)

== the sample mean is the same as the population mean

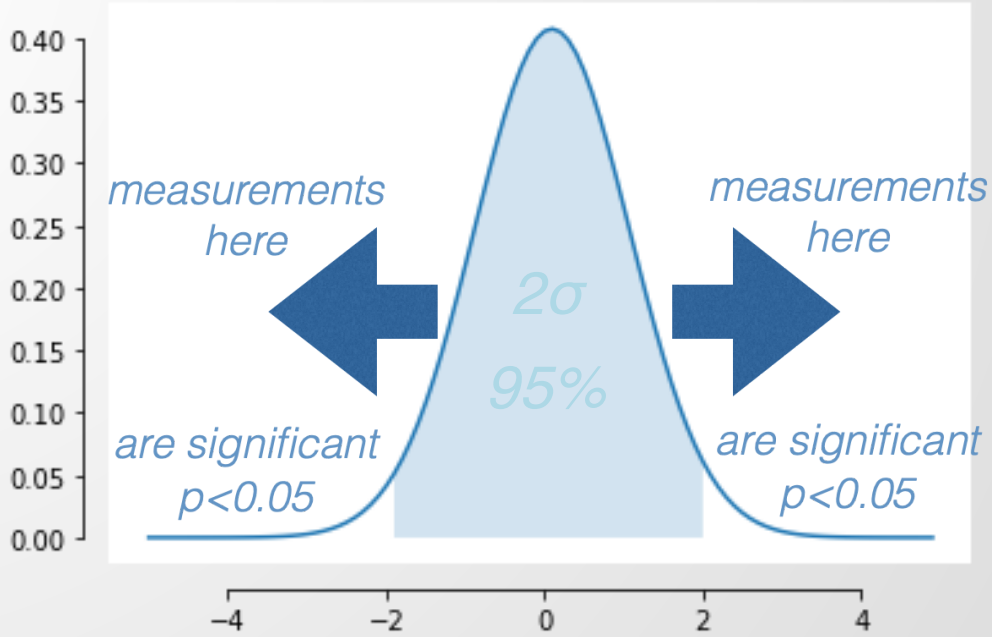
Z is distributed according to a Gaussian N(μ=0, σ=1)

The expectation is that Z will be distributed following a standard normal: a Gaussian with mean 0 std 1.

Values away from 0 are increasingly less probable.

68% probability to get a number b/w -1 and +1

95% probability to get a number b/w -2 and +2



Statistics that follow a Standard Normal distribution

Z -test

Does a sample come from a known population?

How to interpret the number you get?

$$Z = \frac{\mu - \bar{x}}{\sigma / \sqrt{N}}$$

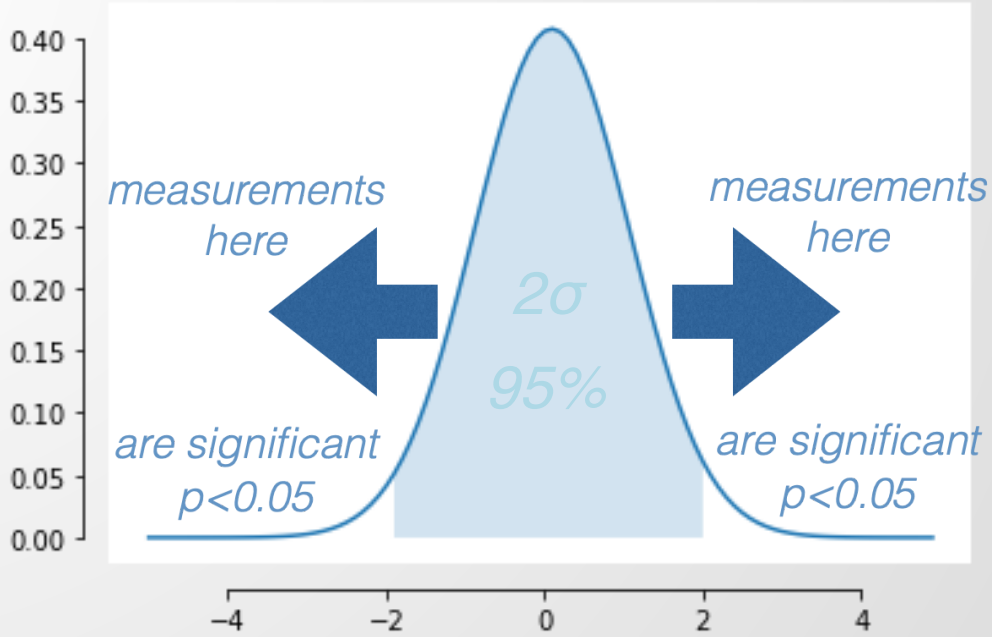
In absence of effect (i.e. under the Null)

== the sample mean is the same as the population mean

Z is distributed according to a Gaussian N(μ=0, σ=1)

IF OUR p-value THRSHOLD IS 1-sigma that means the 68% region is between -1 and +1

=> We have less 68% probability of getting a number <-1 or > 1



Statistics that follow a Standard Normal distribution

Z -test

Are 2 proportions (fractions) the same?

$$Z = \frac{(p_0 - p_1)}{SE}$$

In absence of effect (i.e. under the Null)

== the proportions of men and women are the same

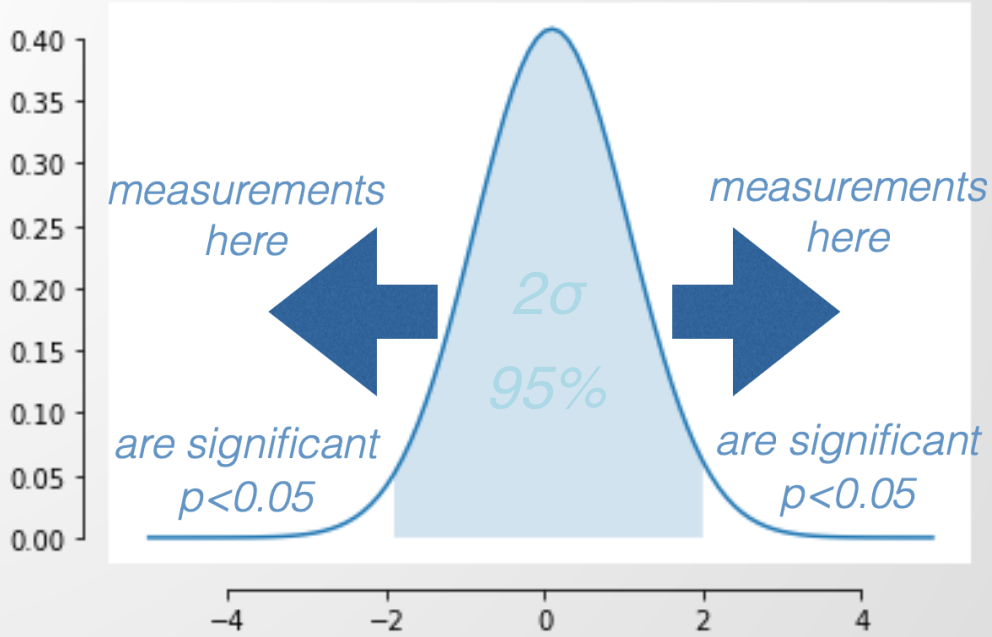
Z is distributed according to a Gaussian N(μ=0, σ=1)

Example: citibike women usage patterns

https://github.com/fedhere/PUS2020_FBianco/blob/master/classdemo/citibikes_gender.ipynb

You want to know if women are less likely than man to use citibike to commute.

You know the fraction of rides women (men) take during the week



Statistics that follow a Standard Normal distribution

Z -test

Are 2 proportions (fractions) the same?

$$Z = \frac{(p_0 - p_1)}{SE}$$

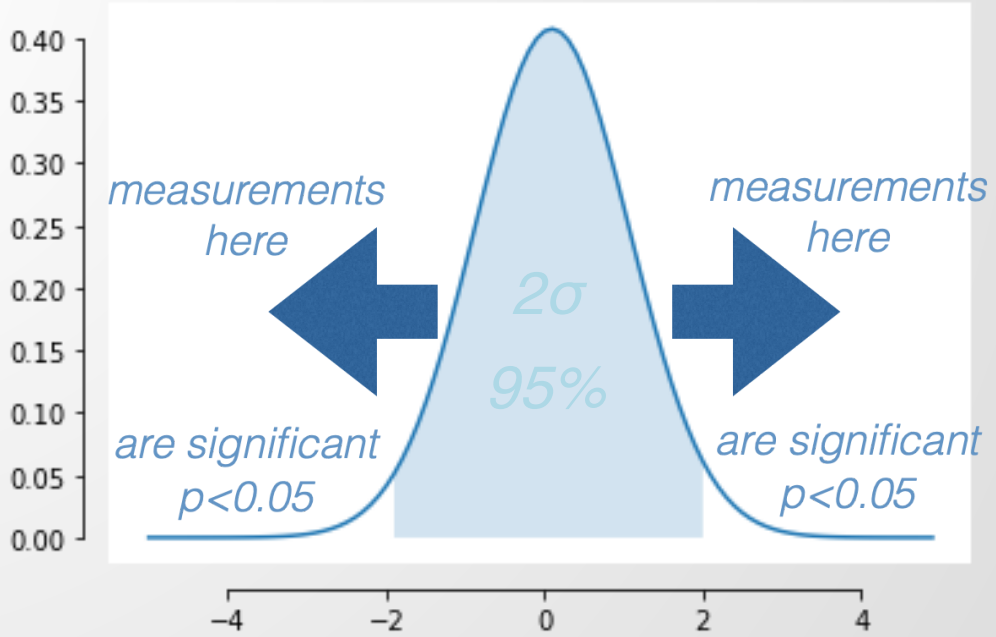
In absence of effect (i.e. under the Null)

== the proportions of men and women are the same

Z is distributed according to a Gaussian N(μ=0, σ=1)

$$p = \frac{p_0 n_0 + p_1 n_1}{n_0 + n_1}$$

$$SE = \sqrt{p(1-p)\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}$$



Statistics that ***do not*** follow a Standard Normal distribution

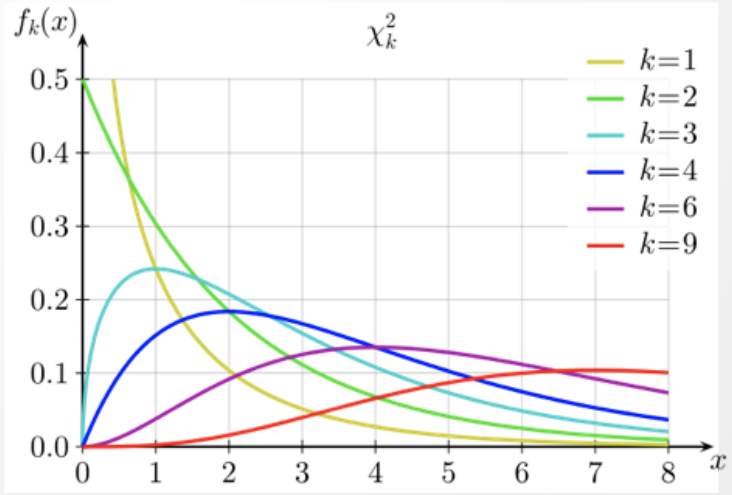
KS-test

Are 2 samples the same?

In absence of effect (i.e. under the Null)

== the samples are drawn from the same population

The KS test is chi-square distributed



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in [0, +\infty)$
PDF	$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
CDF	$\frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$

Statistics that ***do not*** follow a Standard Normal distribution

KS-test

Are 2 samples the same?

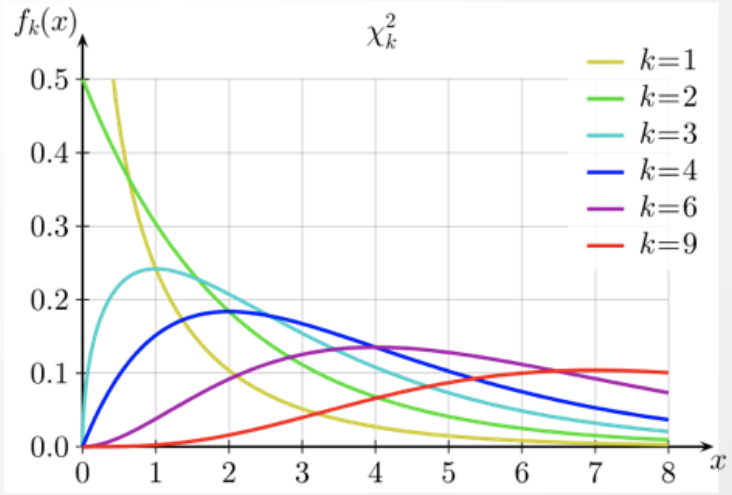
pivotal quantity

In absence of effect (i.e. under the Null)
== the samples are drawn from the same population
The KS test is chi-square distributed

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

Cumulative distribution 1

Cumulative distribution 2



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in [0, +\infty)$
PDF	$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
CDF	$\frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$

$$P(d > observed) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}\sqrt{\frac{N_1 N_2}{N_1+N_2}}d_{12}$$

Statistics that ***do not*** follow a Standard Normal distribution

KS-test

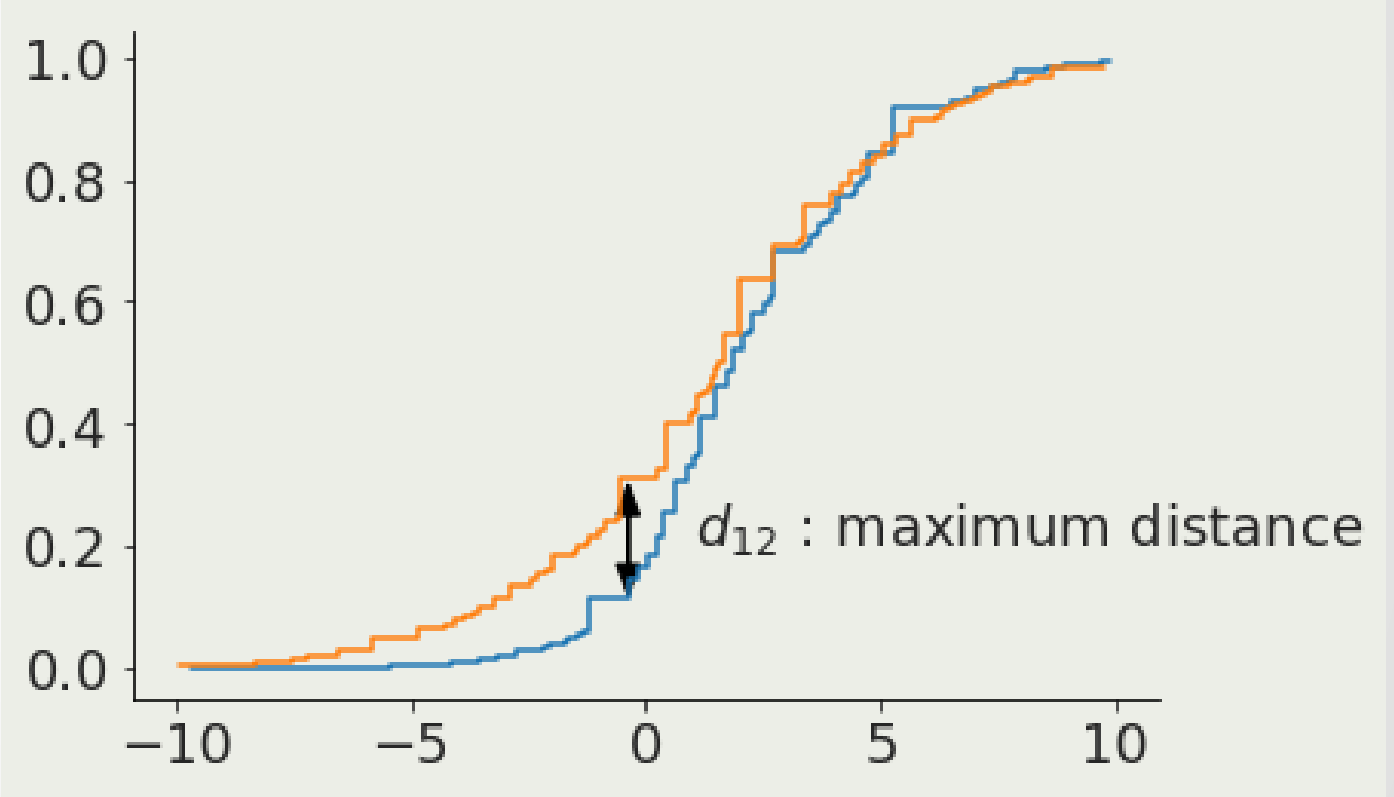
Are 2 samples the same?

pivotal quantity

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

Cumulative distribution 1

Cumulative distribution 2



$$P(d > observed) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}\sqrt{\frac{N_1 N_2}{N_1+N_2}}d_{12}$$

Statistics that ***do not*** follow a Standard Normal distribution

KS-test

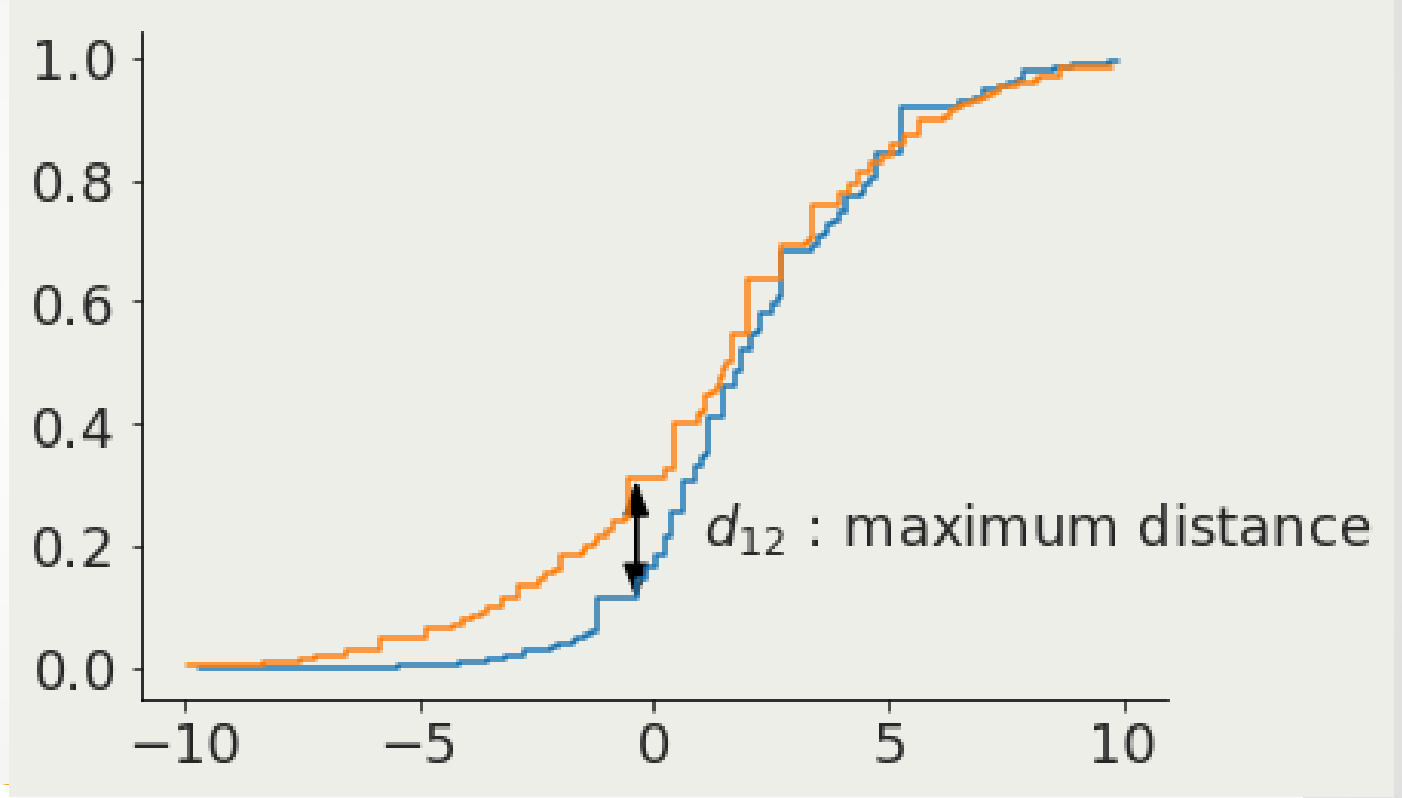
Are 2 samples the same?

pivotal quantity

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

Cumulative distribution 1

Cumulative distribution 2



$$P(d > observed) =$$

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

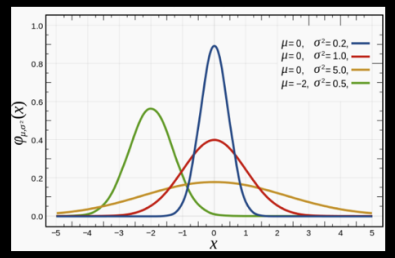
executed in 7ms, finished 14:45:10 2019-09-09

```
Ks_2sampResult(statistic=0.4, pvalue=0.3128526760169558)
```

Statistics and tests

Z statistics Gaussian

$$Z = \frac{\mu - \overline{x}}{\sigma/\sqrt{n}}$$

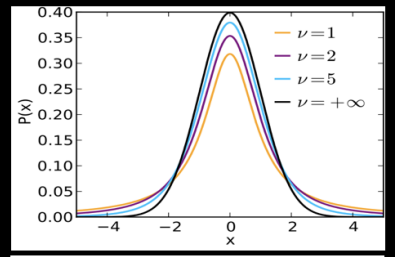


Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

Quantile	$\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2

Student's t

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{S_{AB}\sqrt{\frac{1}{N_A} + \frac{1}{N_B}}}$$

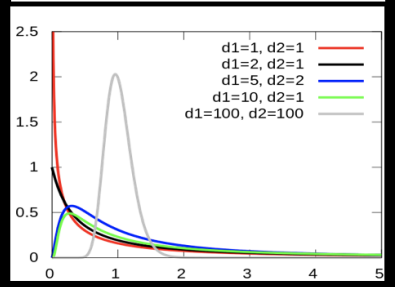


Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function

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

F statistics

$$F = \frac{\sum_i n_i(\overline{x}_i - \overline{x})^2/(K-1)}{\sum_{ij}(x_{ij} - \overline{x}_i)^2/(N-K)}$$

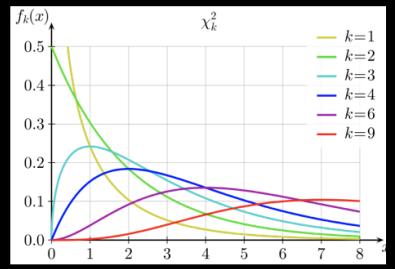


Parameters	$d_1, d_2 > 0$ deg. of freedom
Support	$x \in [0, +\infty)$
PDF	$\frac{\sqrt{\frac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$
CDF	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$

Mean	$\frac{d_2}{d_2 - 2}$ for $d_2 > 2$
Mode	$\frac{d_1 - 2}{d_1}\,\frac{d_2}{d_2 + 2}$ for $d_1 > 2$
Variance	$\frac{2\,d_2^2\,(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)}$ for $d_2 > 4$
Skewness	$\frac{(2d_1 + d_2 - 2)\sqrt{8(d_2-4)}}{(d_2-6)\sqrt{d_1(d_1+d_2-2)}}$ for $d_2 > 6$

Pearson's χ^2

$$\chi^2_P = \sum_i \frac{(O_i - E_i)^2}{E_i}$$



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in [0, +\infty)$
PDF	$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
CDF	$\frac{1}{\Gamma\left(\frac{k}{2}\right)}\,\gamma\left(\frac{k}{2}, \frac{x}{2}\right)$

Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max\{k-2, 0\}$
Variance	$2k$
Skewness	$\sqrt{8/k}$

goodness of fit χ^2 $\chi^2_F = \sum_i \frac{(m_i - x_i)^2}{e_i}$

see
Statistics in a Nutshell

5

data kinds and nomenclature

Types of Data:

Qualitative variables

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

Types of Data:

Qualitative variables

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

Types of Data:

Qualitative variables

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

Quantitative variables

Ordering is meaningful

Time, Distance, Age, Length, Intensity, Satisfaction, Number of

Types of Data:

Qualitative variables

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

Quantitative variables

Ordering is meaningful

Time, Distance, Age, Length, Intensity, Satisfaction, Number of

discrete

Counts:

number of people in a county

Ordinal:

survey response Good/Fair/Poor

Types of Data:

Qualitative variables

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

Quantitative variables

Ordering is meaningful

Time, Distance, Age, Length, Intensity, Satisfaction, Number of

discrete ● ● ● ● ● ● ● ● ● ● ● ● ● ● **continuous**

Counts:

number of people in a county

Ordinal:

survey response Good/Fair/Poor

Continuous Ordinal:

Earthquakes (notlinear scale)

Interval:

F temperature interval size preserved

Ratio:

Car speed

0 is naturally defined

Types of Data:

***Qualitative* variables**

No ordering

UrbanScience e.g. precinct, state, gender, Also called *Nominal, Categorical*

***Quantitative* variables**

Ordering is meaningful

Time, Distance, Age, Length, Intensity, Satisfaction, Number of

discrete **continuous**

Counts:

number of people in a county

Ordinal:

survey response Good/Fair/Poor

Continuous Ordinal:

Earthquakes (notlinear scale)

Interval:

F temperature interval size preserved

Ratio:

Car speed

0 is naturally defined

Missing: “Prefer not to answer” (NA / NaN)

Censored: age>90

key concepts

descriptive statistics

null hypothesis rejection testing setup

pivotal quantities

Z, K-S tests

https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291

Understanding Descriptive Statistics

READING

Descriptive Statistics [Image 1] (Image courtesy: My Photoshopped Collection)

Statistics is a branch of mathematics that deals with collecting, interpreting, organization, and interpretation of data.