

# data science for (physical) scientists IIa

II: physics in a probabilistic world

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

1  $P(\text{physics} \mid \text{data})$

2 NHRT

p-values

z-test

3 comparing distributions

Z, t,  $\chi^2$ , ks-test

KL divergence

this slide deck

[http://bit.ly/dsps2019\\_2a](http://bit.ly/dsps2019_2a)



## 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)

# neap 2

## probability

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

# reap 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

# reap

## 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**

descriptive statistics:

we summarize the properties of a distribution

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

reap



descriptive statistics:

we summarize the properties of a distribution

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

mean:  $n=1$

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

other measures of central tendency:

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

50% to the right

mode: most popular value in the distribution

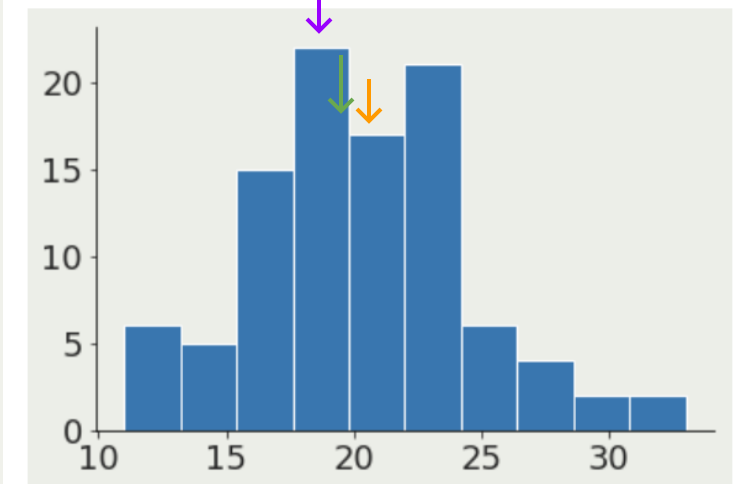
```
dist = sp.stats.poisson.rvs(size=100, mu=20)
pl.hist(dist)
print(dist.mean())
print(np.median(dist))
print(sp.stats.mode(dist))
|
```

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 properties of a distribution

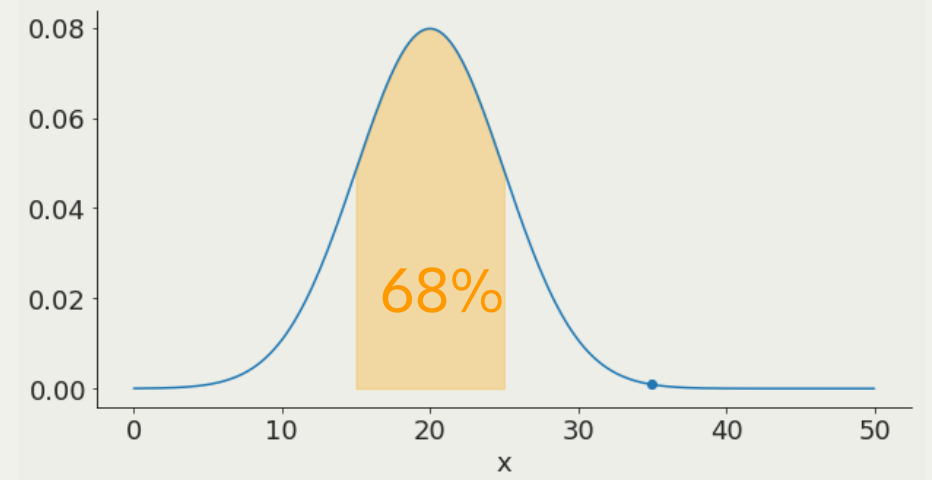
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx.$$

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

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

Gaussian distribution:

$1\sigma$  contains 68% of the distribution



descriptive statistics:

we summarize the properties of a distribution

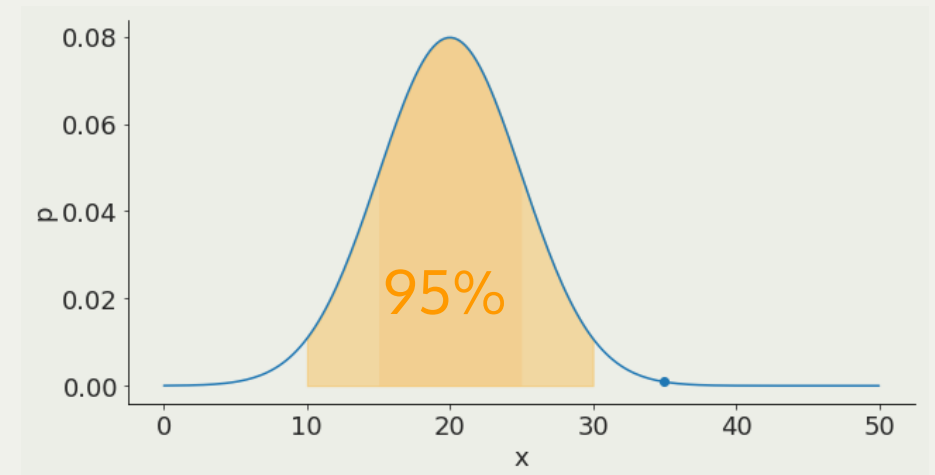
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx.$$

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

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

Gaussian distribution:

$2\sigma$  contains 95% of the distribution



descriptive statistics:

we summarize the properties of a distribution

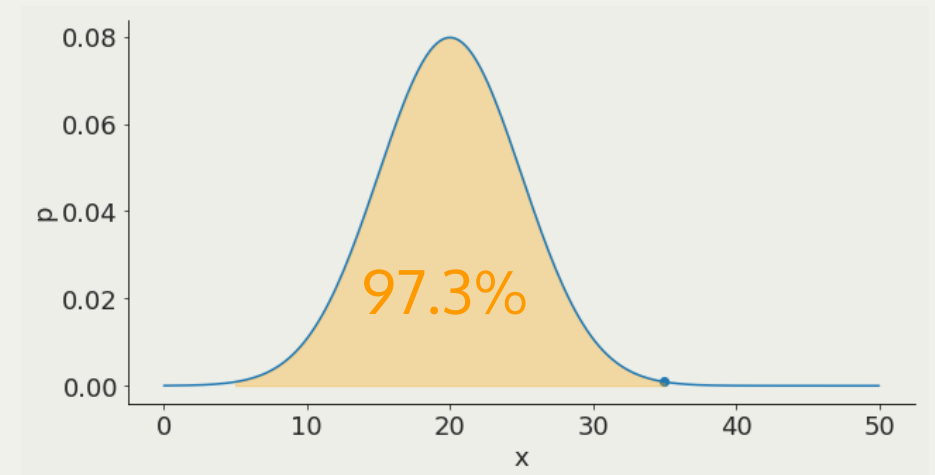
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx.$$

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

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

Gaussian distribution:

$3\sigma$  contains 97.3% of the distribution



1

the scientific method  
in a probabilistic context

**p(physics | data)**

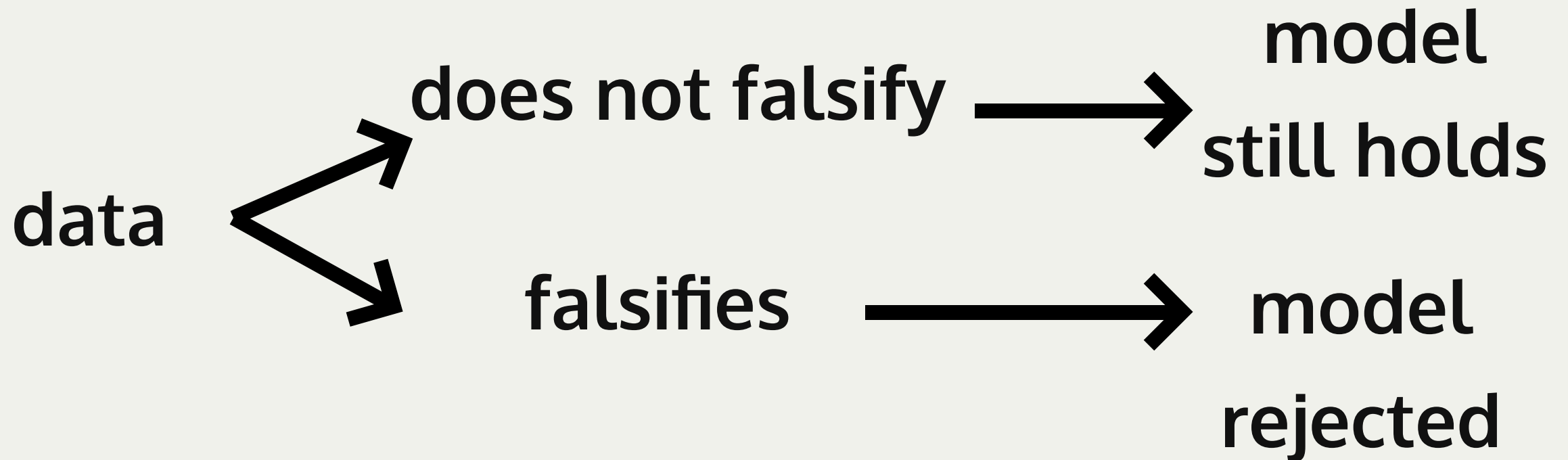
Bayesian Inference

Forward Modeling

Frequentist approach  
(NHRT)

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

**model**  **prediction**





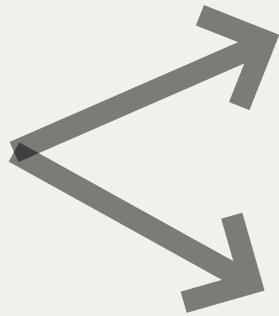
**model**



**prediction**

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

**data**



**does not falsify**



**model  
still holds**

**falsifies**



**model  
rejected**

**model**

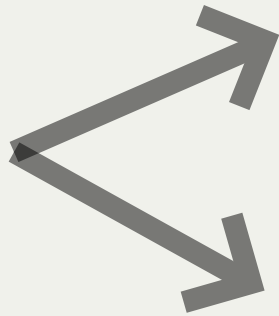


**prediction**

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

*this has a high probability  
of happening*

**data**



**does not falsify**



**model  
still holds**

**falsifies**



**model  
rejected**

**model**

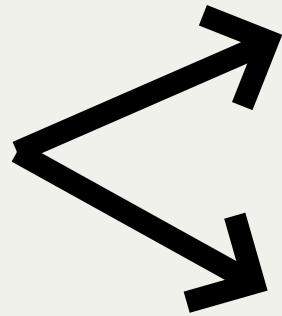


**prediction**

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

*this has a high probability  
of happening*

**data**



**does not falsify**



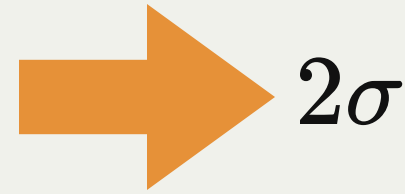
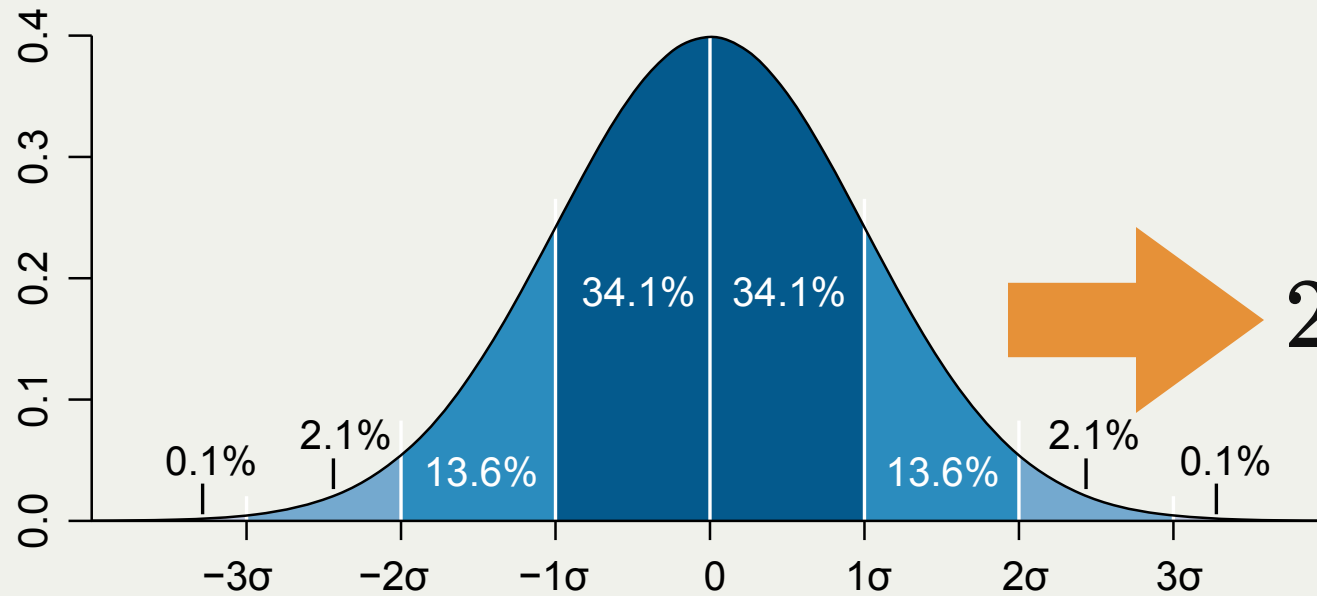
**model  
still holds**

**falsifies**

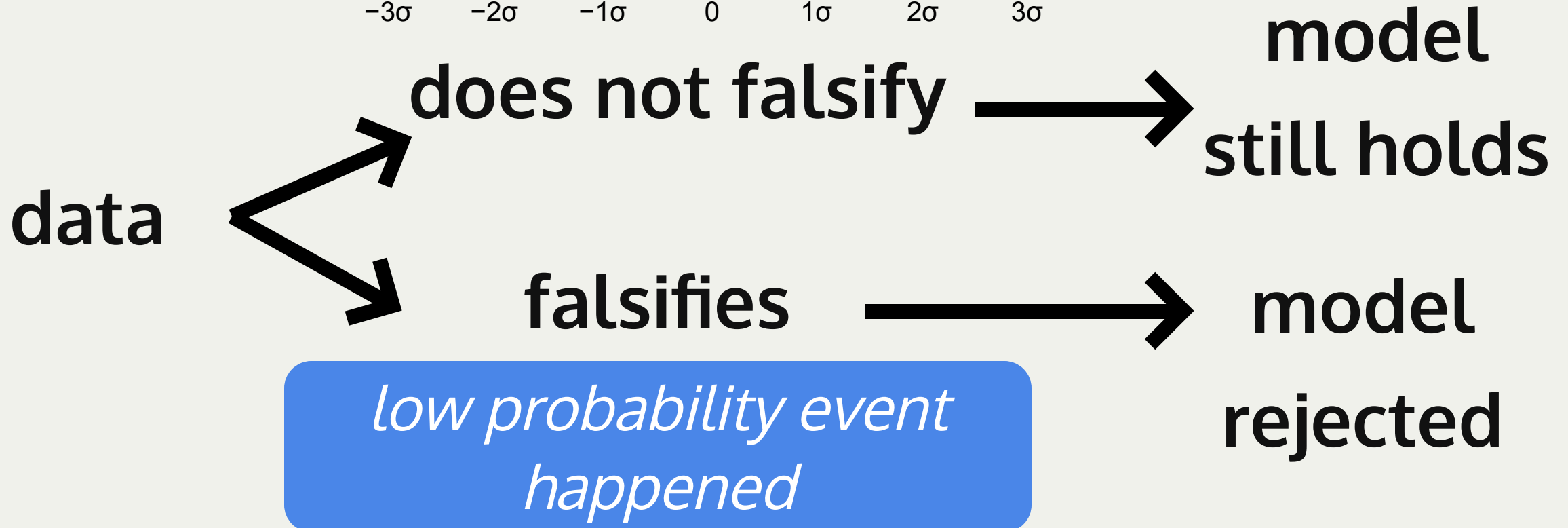


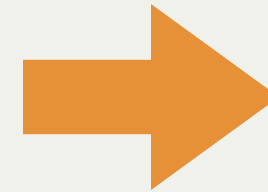
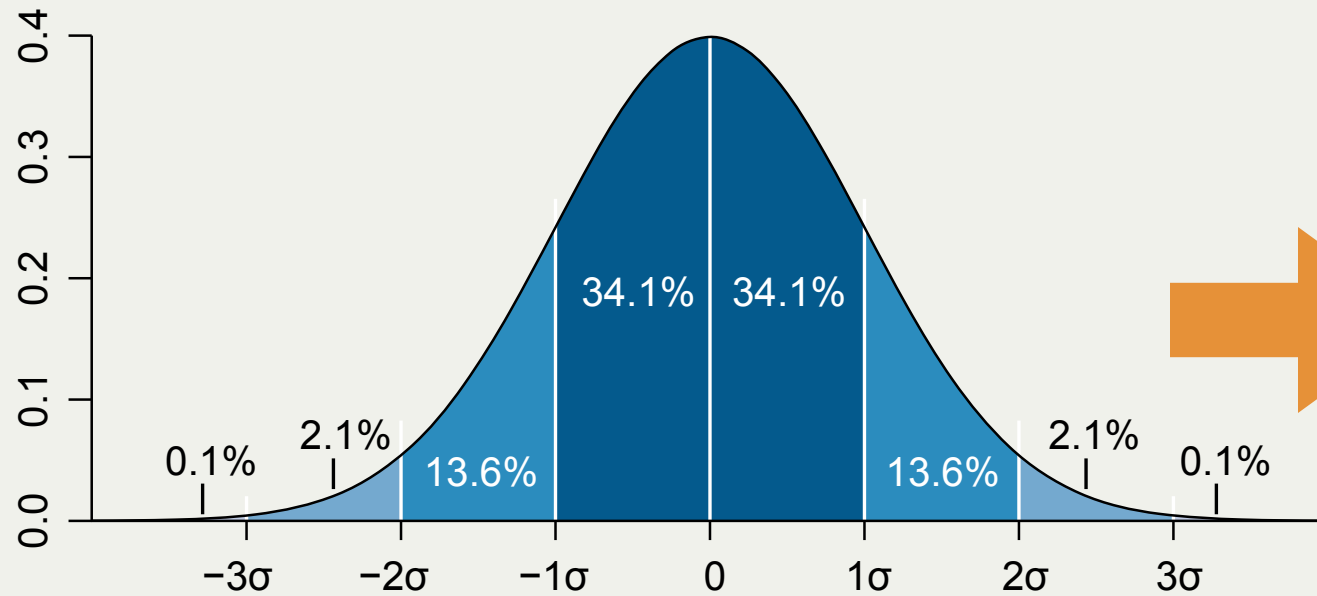
**model  
rejected**

*low probability event  
happened*

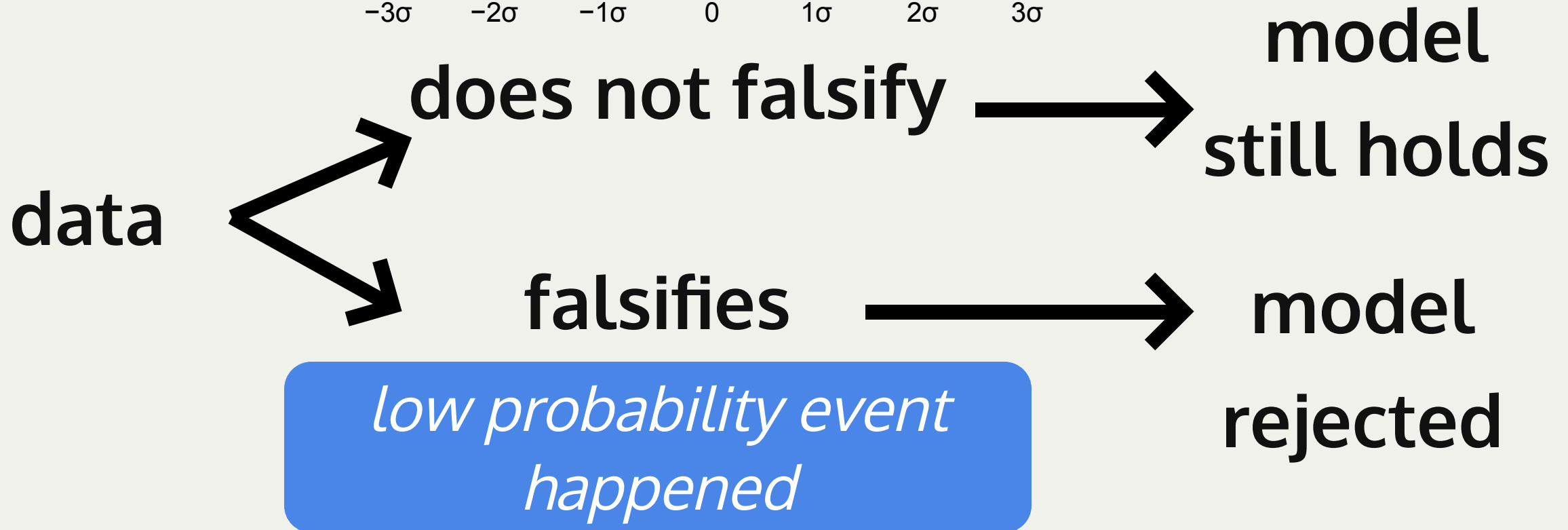


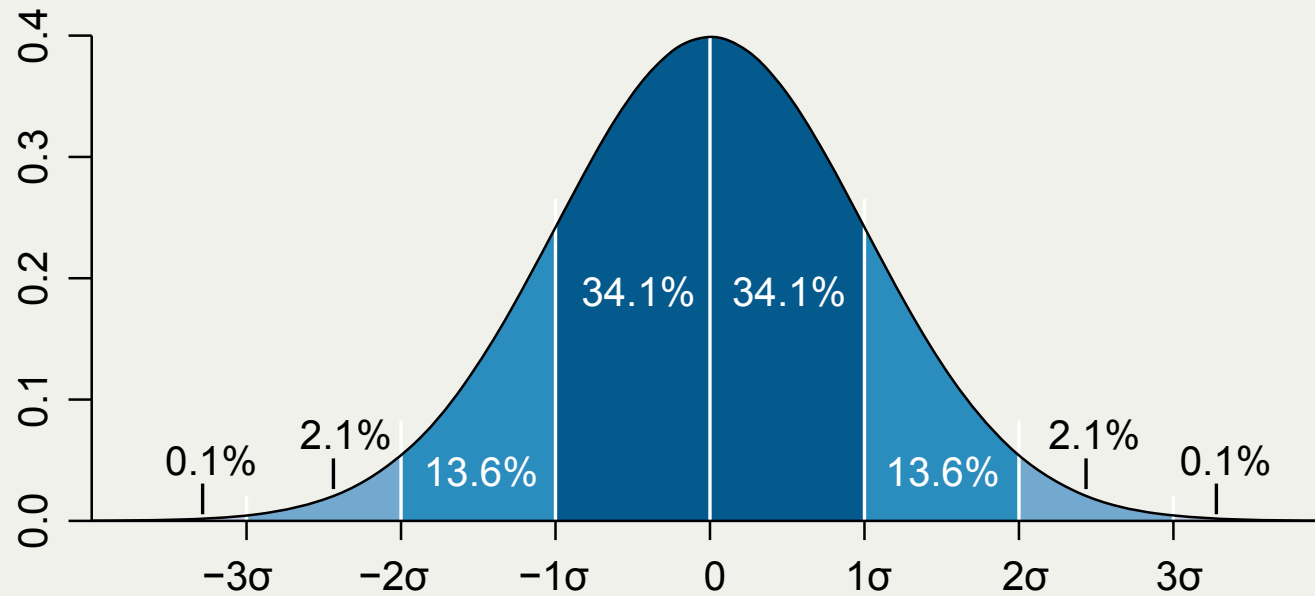
rejected at 95%  
0.05 p-value  
5% confidence





rejected at 99.7%  
 $3\sigma$  0.003 p-value  
0.3% confidence



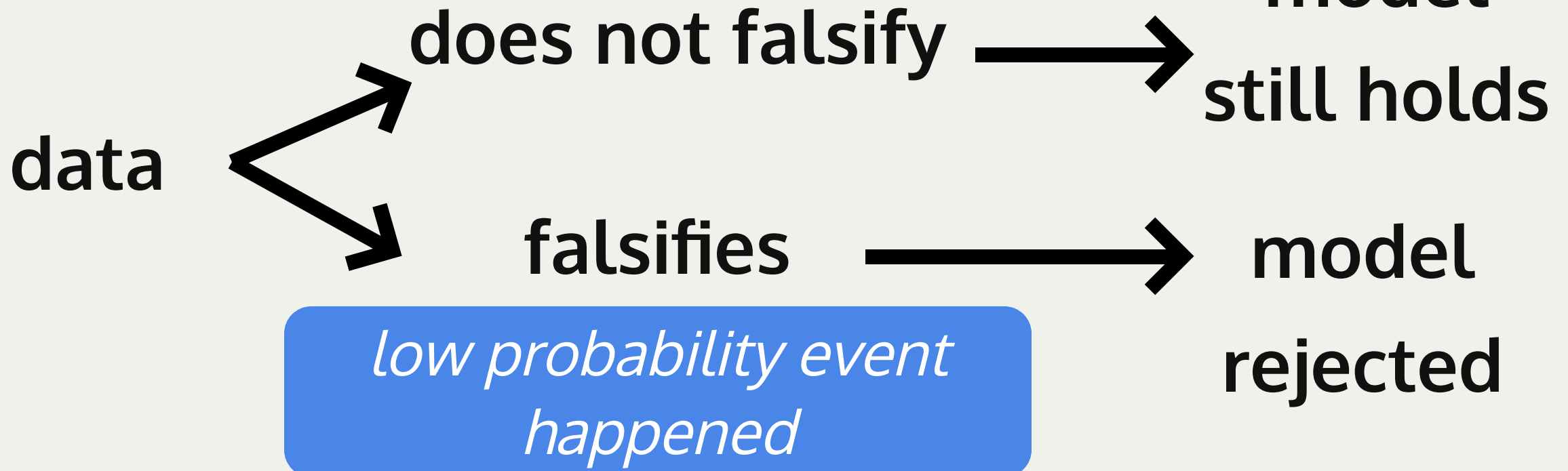


rejected at 99.99...%

3-e7 p-value

3-e5% confidence

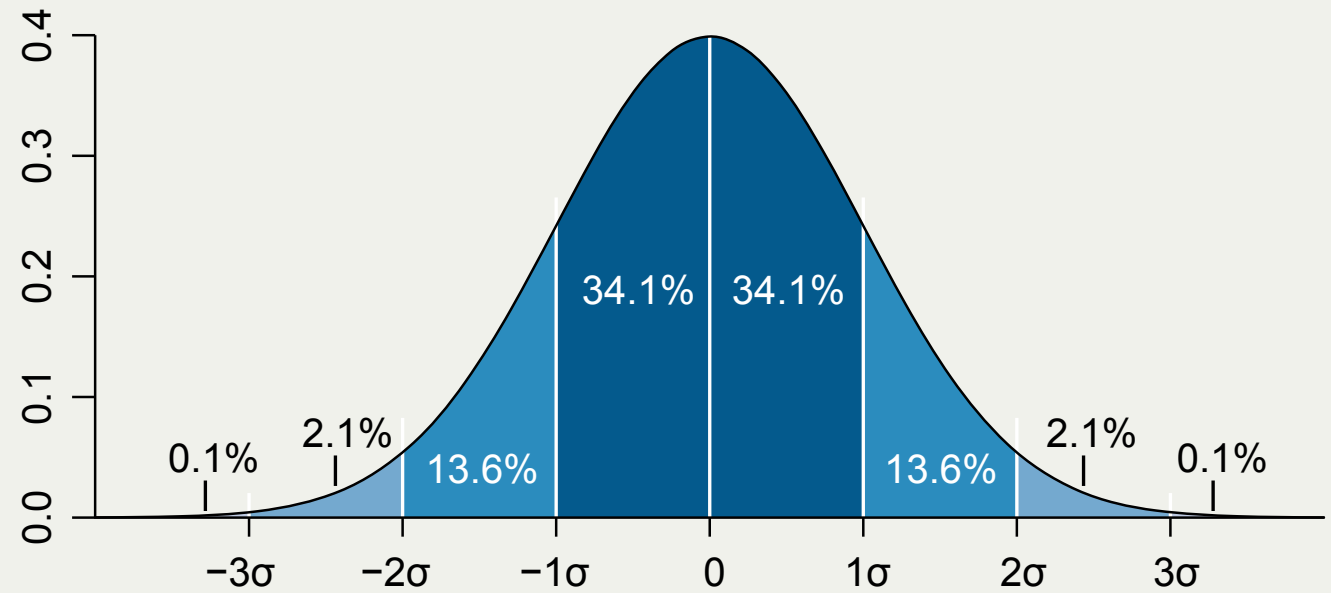
→  $5\sigma$



2

Null hypothesis rejection testing

Null  
Hypothesis  
Rejection  
Testing



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



Null  
Hypothesis  
Rejection  
Testing

1

formulate your prediction

Null Hypothesis

Null

Hypothesis

Rejection

Testing

2

identify all alternative  
outcomes

Alternative Hypothesis

Null

Hypothesis

Rejection

Testing



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

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

2  
identify all alternative  
outcomes

Alternative Hypothesis

But instead of verifying a theory we want to falsify one  
**model**  **prediction**

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

*this has a **low** probability  
of happening*



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 model is true*

*this has a **low** probability  
of happening*

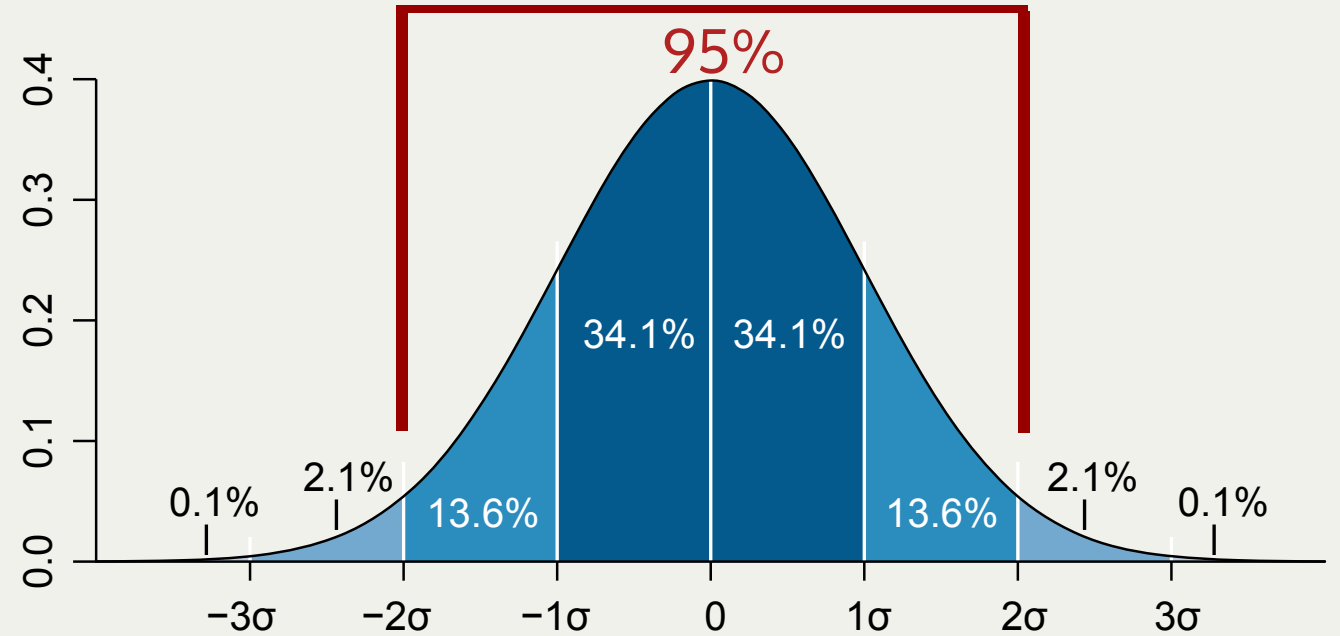


**Earth is flat is Null**

**Earth is round is Alternative:**

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

# Null Hypothesis Rejection Testing



3 set confidence threshold

$2\sigma$  confidence level

0.05 p-value

95%  $\alpha$  threshold

Null

Hypothesis

Rejection

Testing

*pivotal quantities*

4

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

# *pivotal quantities*

quantities that under the Null Hypothesis follow a known distribution

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?



## *pivotal quantities*

quantities that under the Null  
Hypothesis follow a known distribution

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

Null

Hypothesis

Rejection

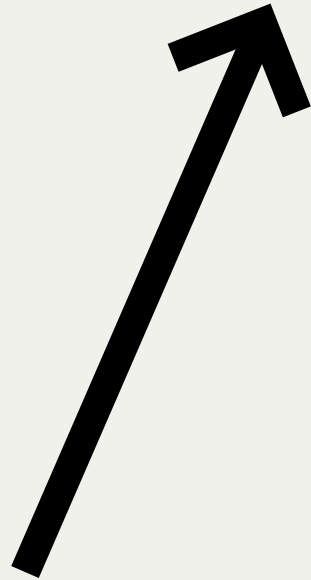
Testing

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

prediction is unlikely

Null rejected

Alternative holds



test data against

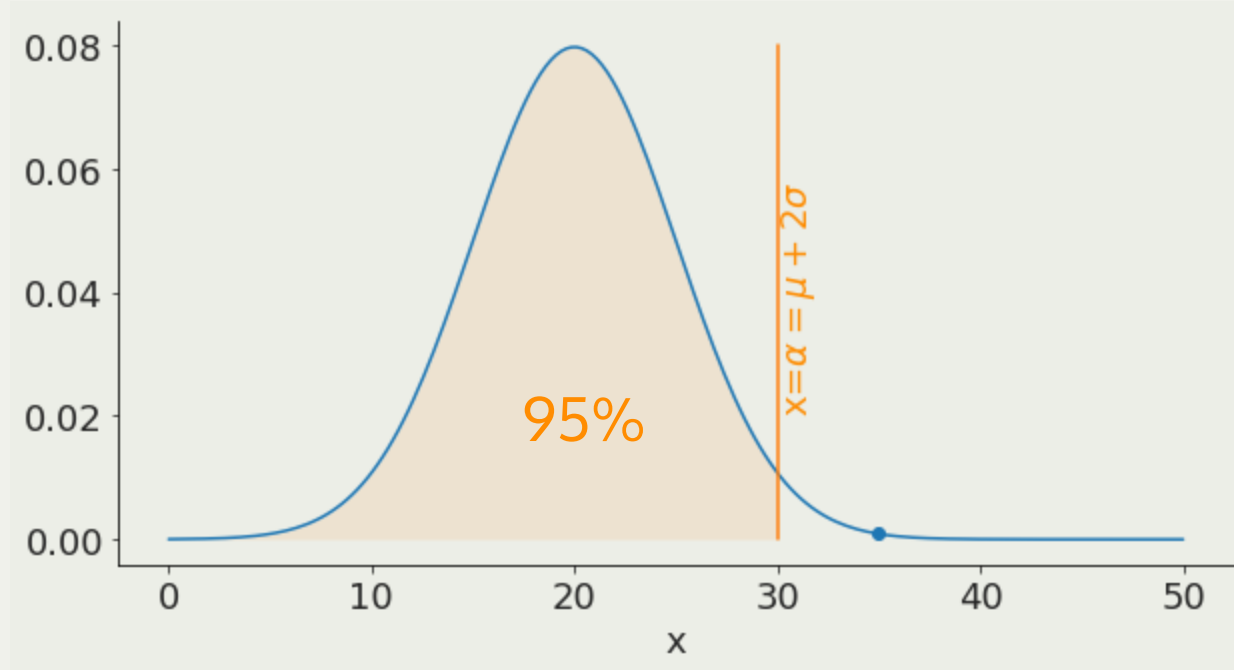
*alternative* outcomes



# Null Hypothesis Rejection Testing

## what is $\alpha$ ?

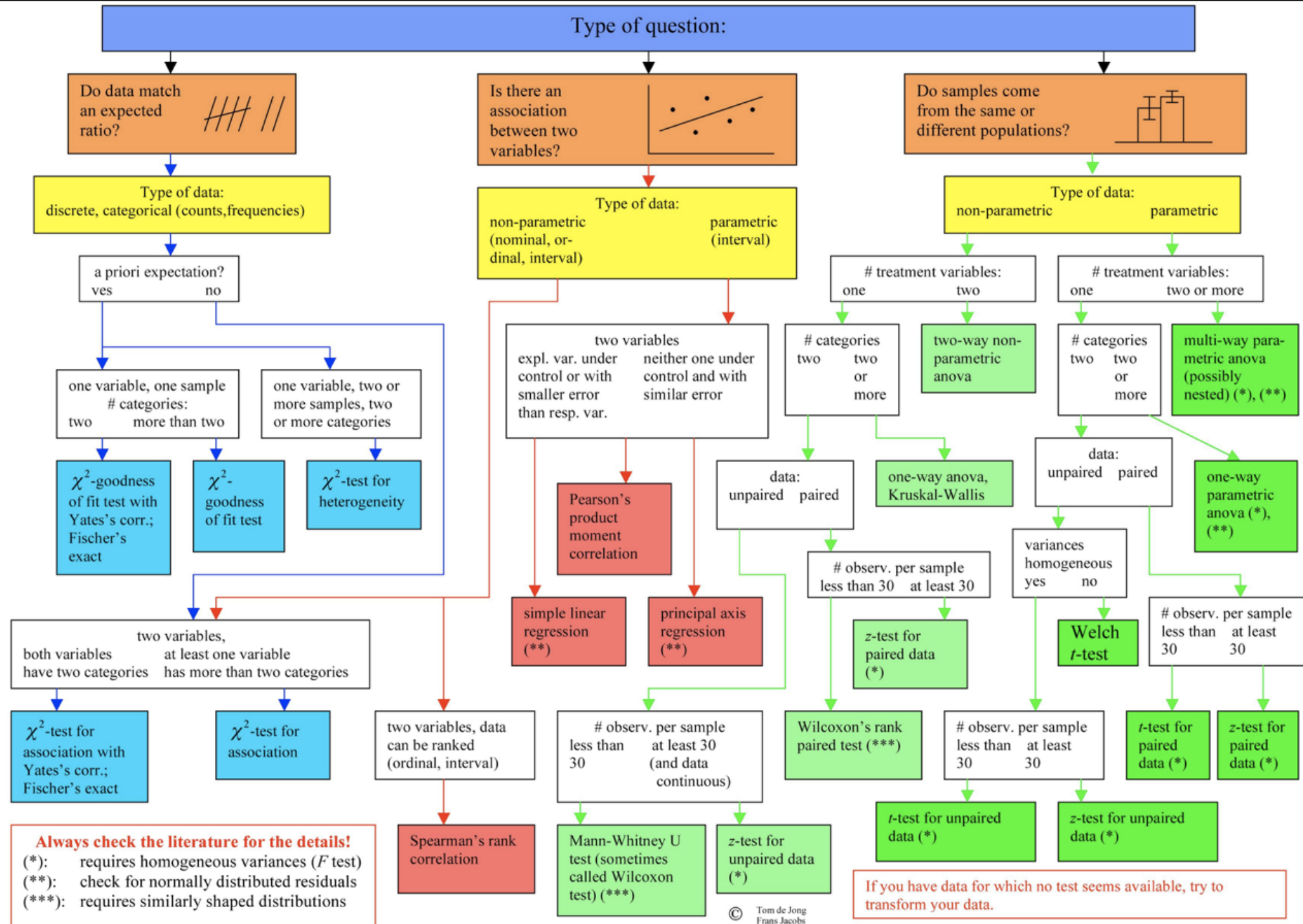
$\alpha$  is the x value corresponding to a chosen threshold



test data against  
*alternative* outcomes

3

common tests and pivotal quantities



# *pivotal quantities*

quantities that under the Null Hypothesis follow a known distribution

also called "statistics"

e.g.:  $\chi^2$  statistics: difference between expectation and reality squared

$Z$  statistics: difference between means

$K-S$  statistics: maximum distance of cumulative distributions.

# Z-test

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

*pivotal quantity*

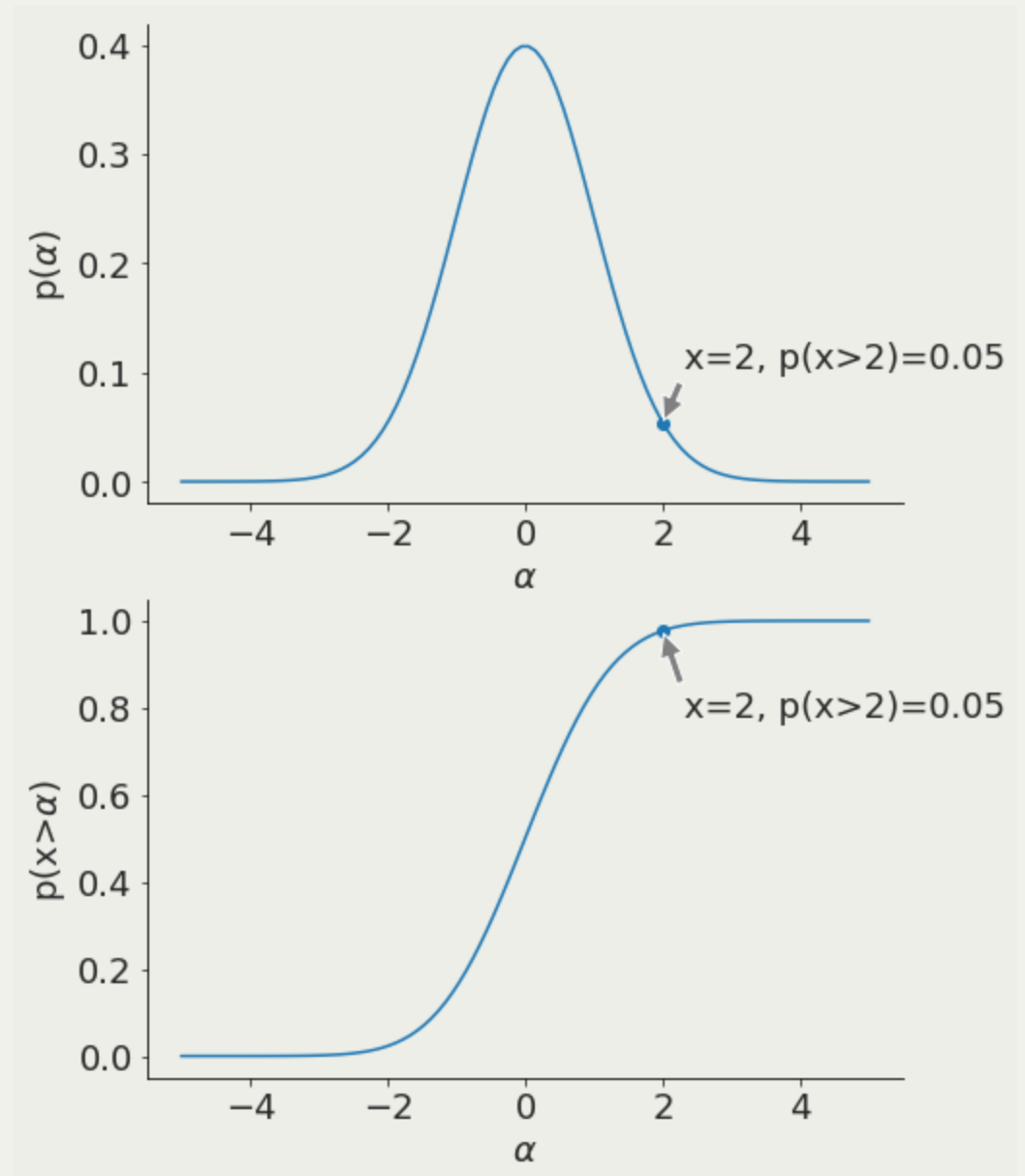
$$Z = (\bar{X} - \mu_0) / s$$

sample  
mean

population  
mean

sample  
variance

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



# Z-test

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

*pivotal quantity*

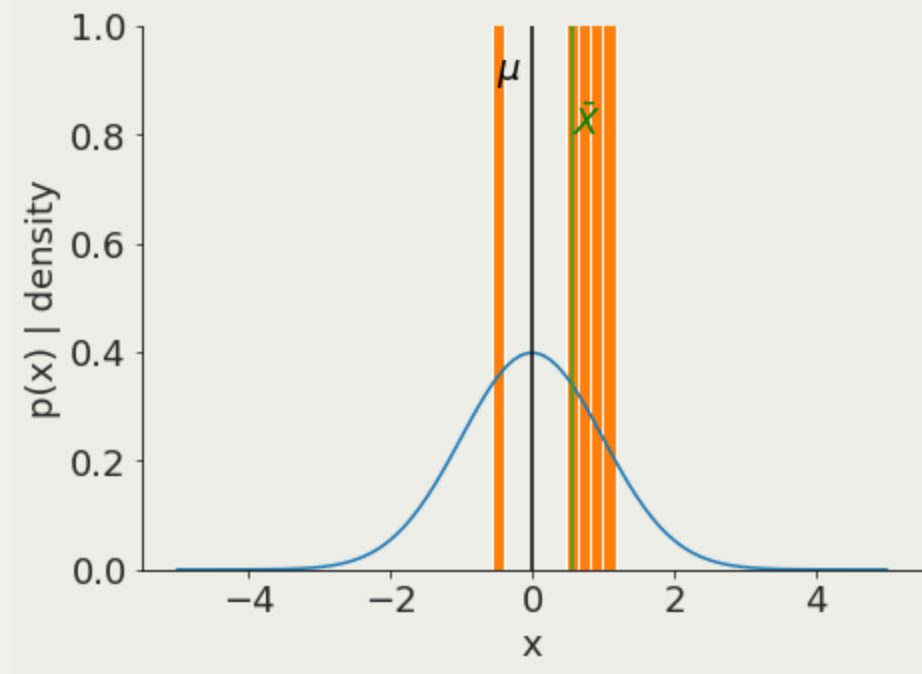
$$Z = (\bar{X} - \mu_0) / s$$

sample  
mean

population  
mean

sample  
variance

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



why do we need a test? why not just measuring the means and seeing if 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*

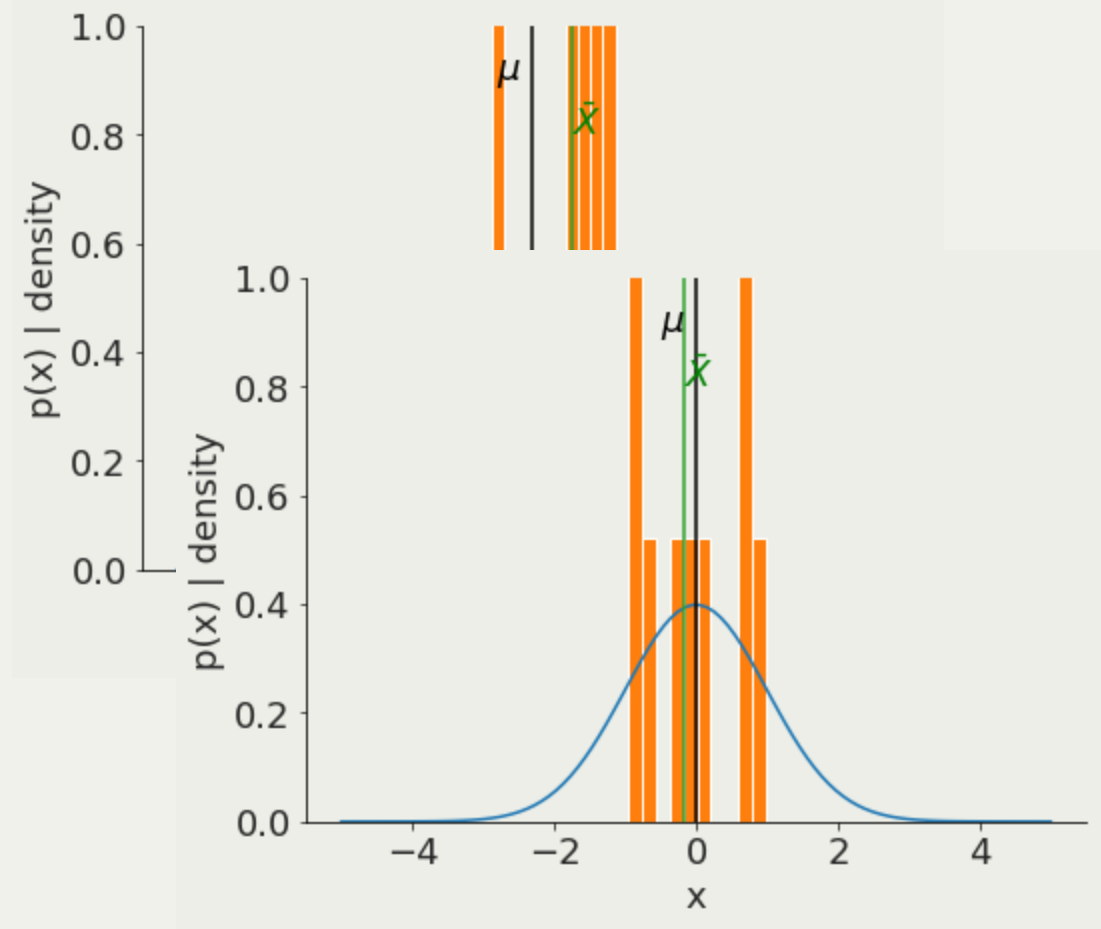
$$Z = (\bar{X} - \mu_0) / s$$

sample  
mean

population  
mean

sample  
variance

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



# Z-test

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

*pivotal quantity*

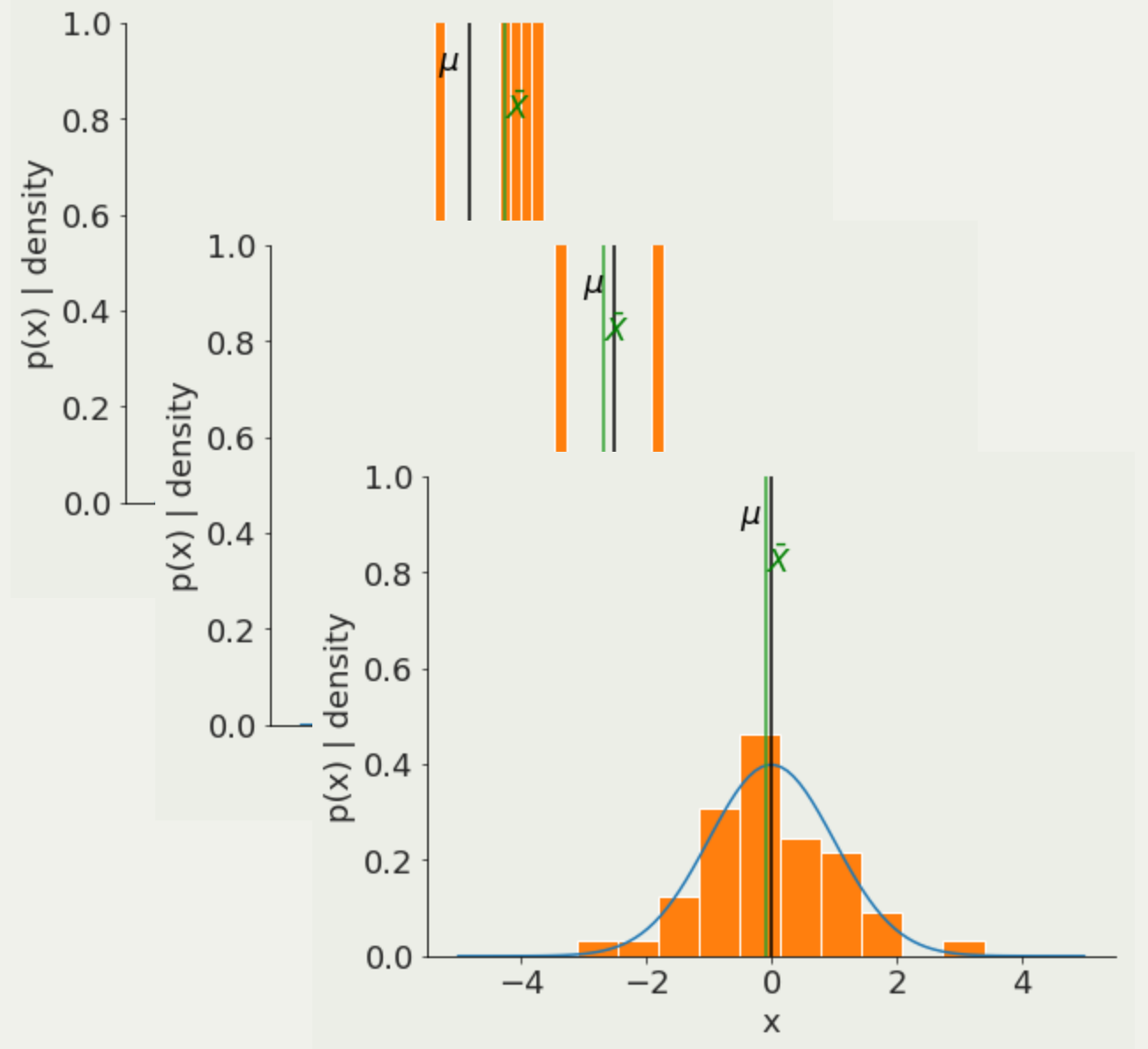
$$Z = (\bar{X} - \mu_0) / s$$

sample  
mean

population  
mean

sample  
variance

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



# Z-test

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

*pivotal quantity*

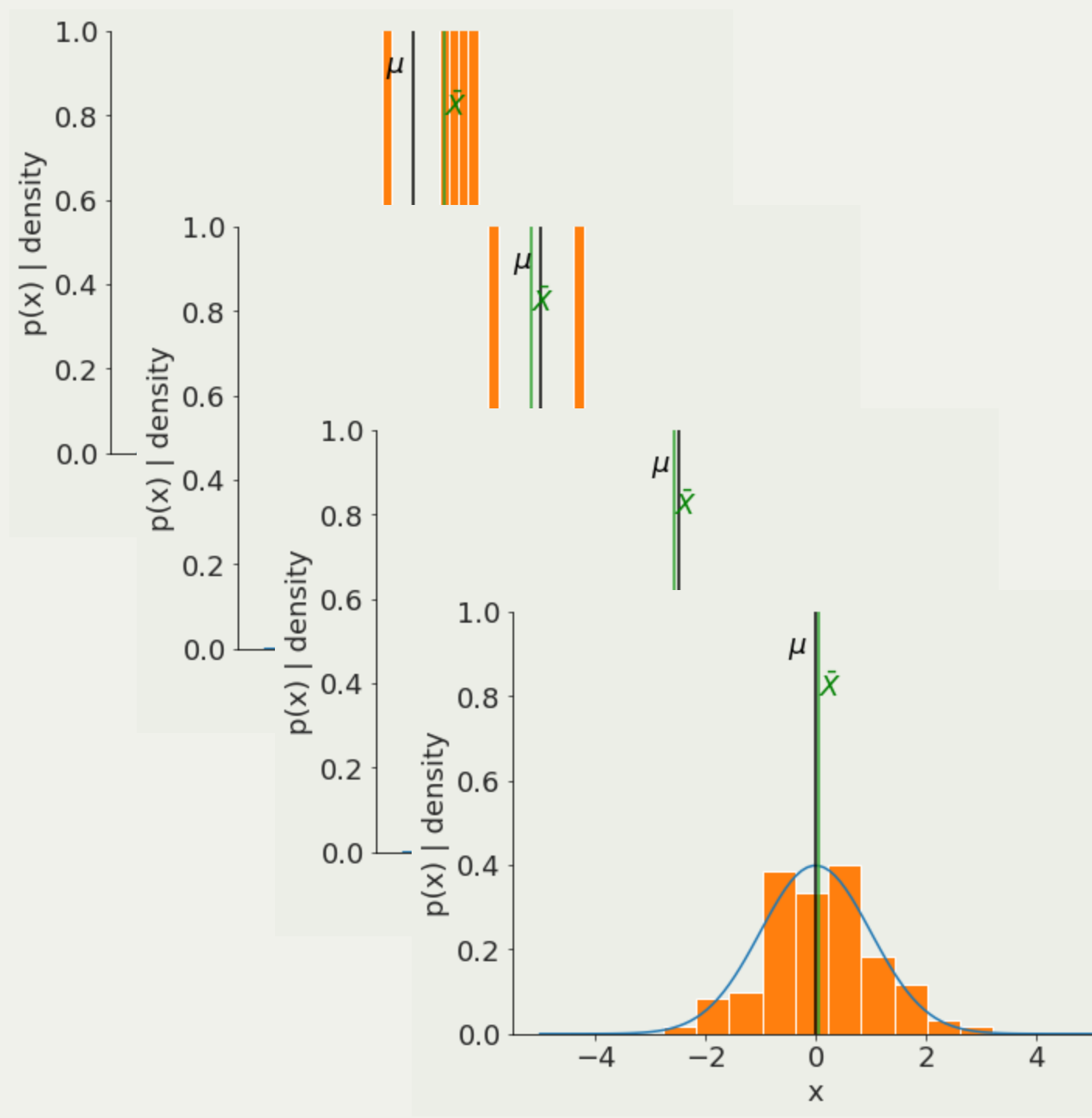
$$Z = (\bar{X} - \mu_0) / s$$

sample  
mean

population  
mean

sample  
variance

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



# t- test

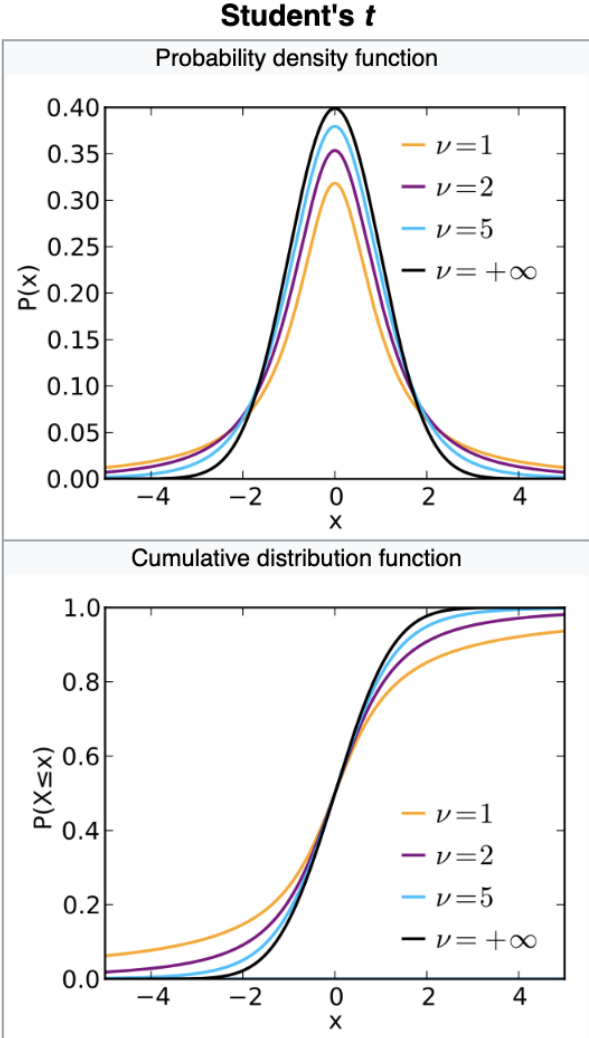
Are the means of 2 samples significantly different?

*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

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



WIKIPEDIA  
The Free Encyclopedia

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

# $t$ - test

Are the means of 2 samples significantly different?

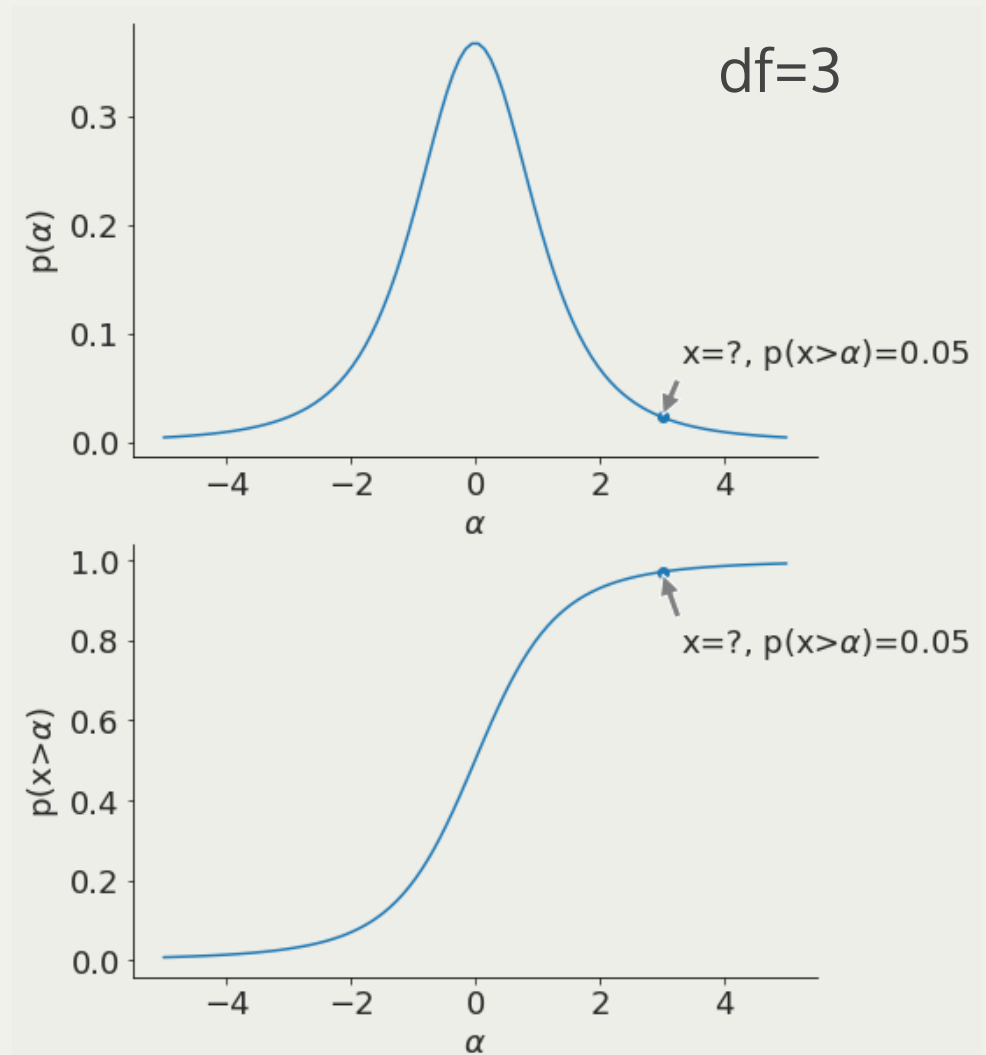
*pivotal quantity*

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

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



# t- test

Are the means of 2 samples significantly different?

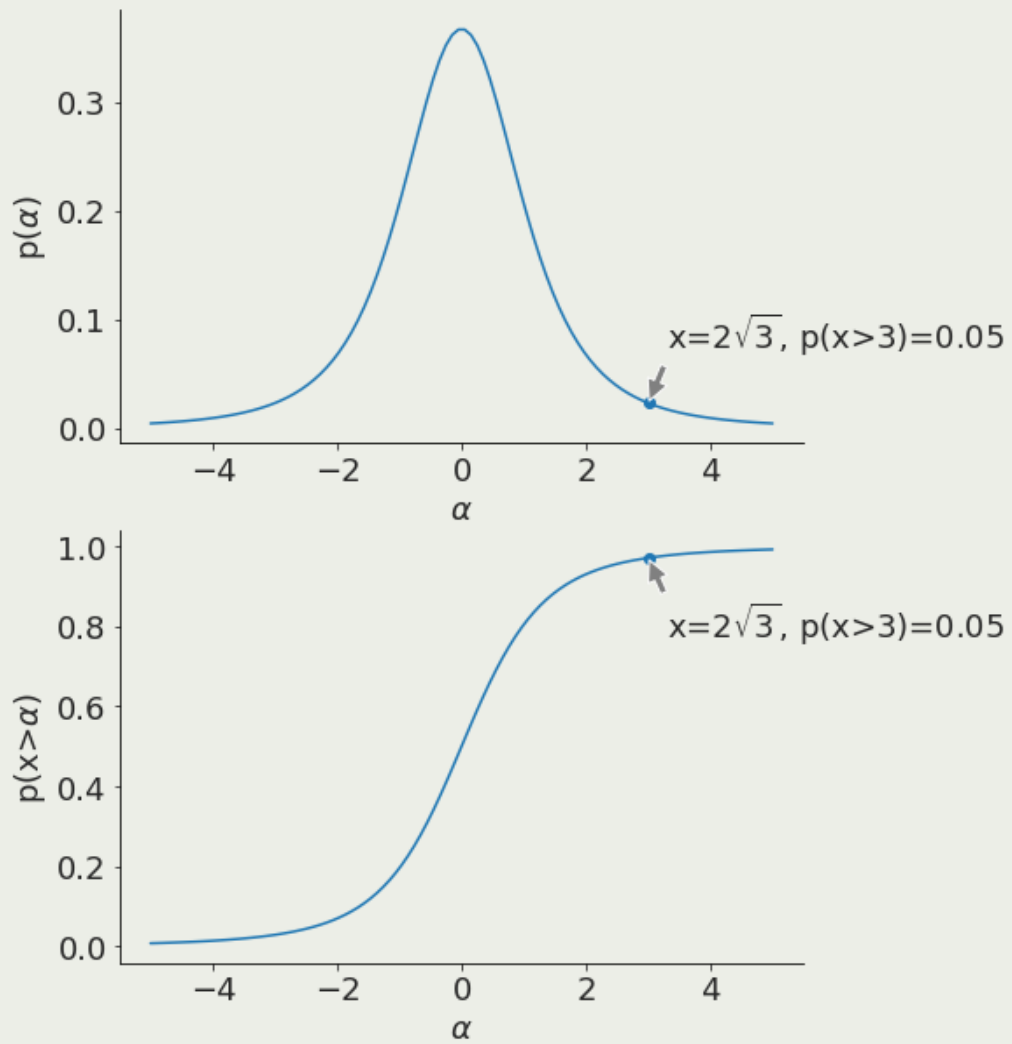
*pivotal quantity*

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

unbias variance estimator

size of sample

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



# K-S test

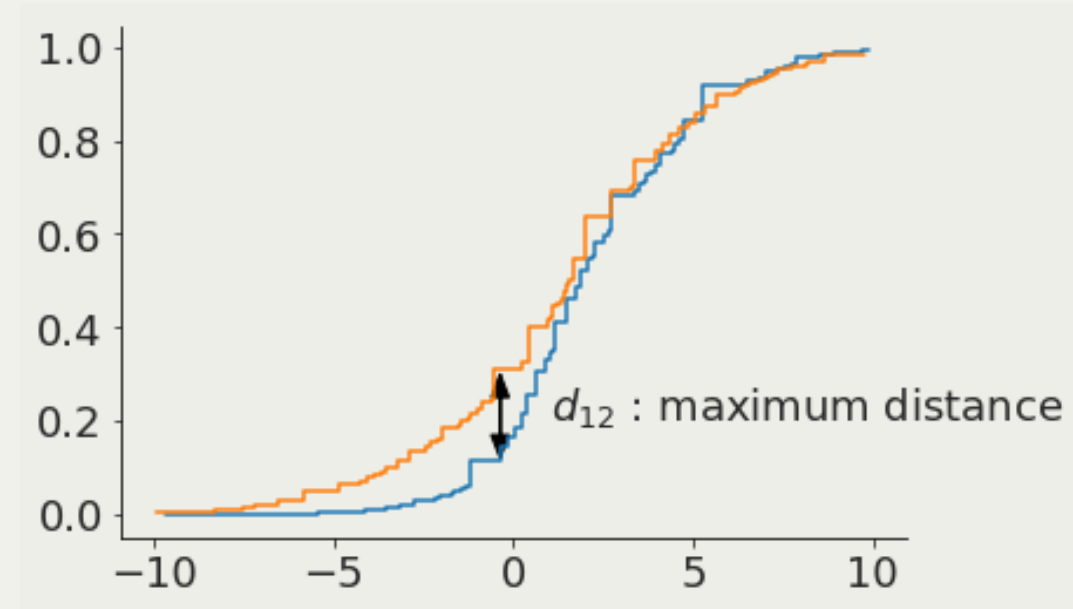
Kolmogorof-Smirnoff :

do two samples come from the same  
parent distribution?

*pivotal quantity*

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

↓                      ↓  
Cumulative      Cumulative  
distribution 1    distribution 2



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

# K-S test

Kolmogorof-Smirnoff :

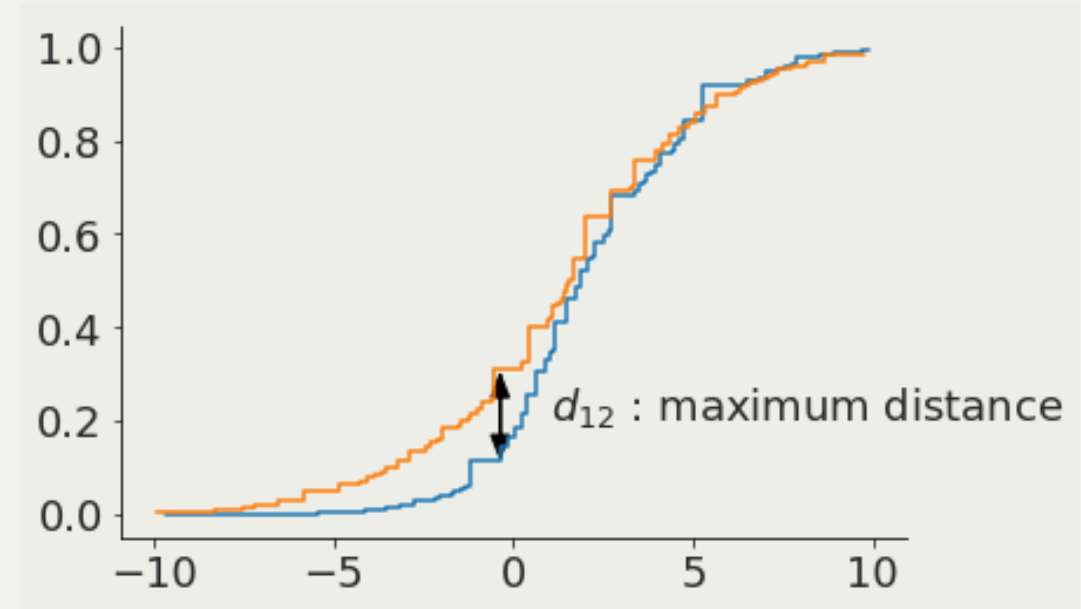
do two samples come from the same  
parent distribution?

*pivotal quantity*

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

↓  
Cumulative  
distribution 1

↓  
Cumulative  
distribution 2



*$P(d > \text{observed}) =$*

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

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

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



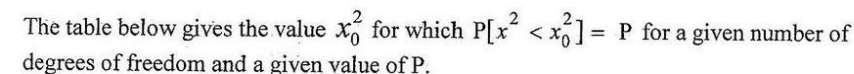
are the data what is expected from the model (if likelihood is Gaussian... we'll see this later) - there are a few  $\chi^2$  tests. The one here is the "Pearson's  $\chi^2$  tests"

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

model                  uncertainty                  observation

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

↓  
number of  
observation



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) - there are a few  $\chi^2$  tests. The one here is the "Pearson's  $\chi^2$  tests"

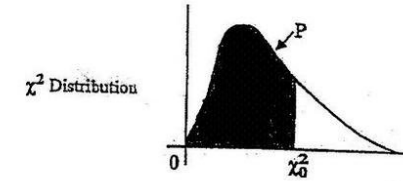
*pivotal quantity*

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

model      observation  
            uncertainty

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

number of  
observation



The table below gives the value  $\chi_0^2$  for which  $P[\chi^2 < \chi_0^2] = P$  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

Null

Hypothesis

Rejection

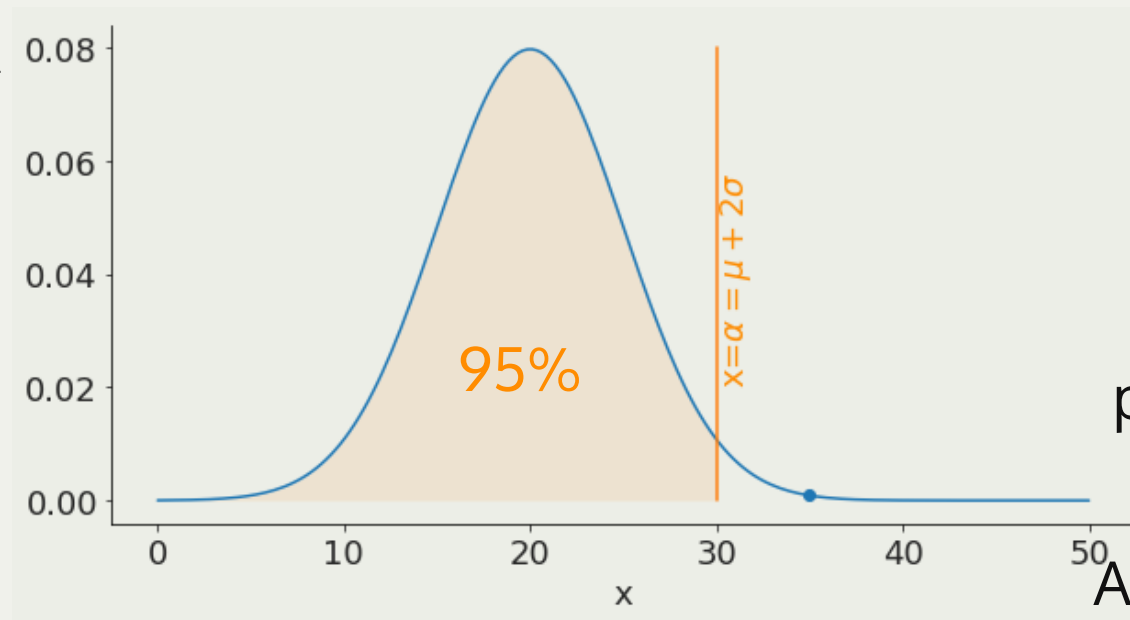
Testing

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

prediction is unlikely

Null rejected

Alternative holds



prediction is likely

Null holds

Alternative rejected



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

test data against  
*alternative* outcomes

Null

Hypothesis

Rejection

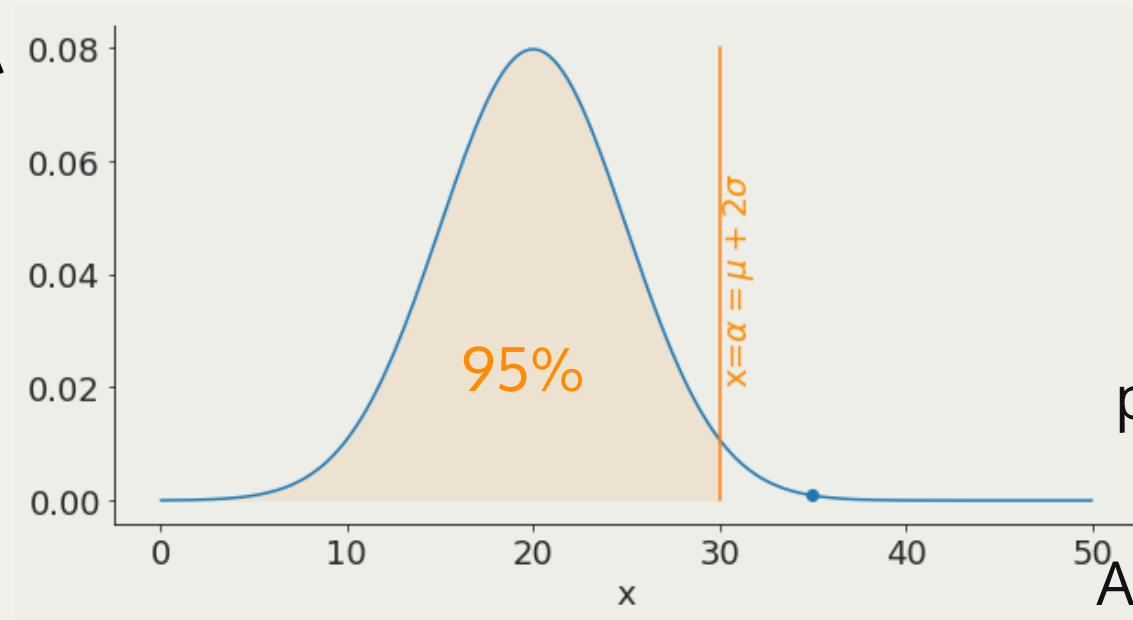
Testing

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

prediction is unlikely

Null rejected

Alternative holds



prediction is likely

Null holds

Alternative rejected



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

test data against  
*alternative* outcomes

formulate the Null as the comprehensive opposite of your theory

**model**



**prediction**

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

*this has a low  
probability of happening*

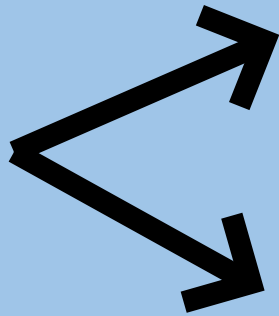
**everything**

**but model  
is rejected**

**does not falsify  
alternative**



**data**



**falsifies  
alternative**



**model  
holds**

**Key Slide**

*low probability event happened*

# key concepts

descriptive statistics

null hypothesis rejection  
testing setup

pivotal quantities

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

- HW1 : earthquakes and KS test  
(<https://arxiv.org/pdf/0910.0055.pdf>)
- ....

# homework

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008

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

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

David M. Lane et al.

**Introduction to Statistics (XVIII)**

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

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

# resources