data science for (physical) scientists IIa

II: physics in a probabilistic world

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

1 P(physics | 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

recap 1

Guiding principle of science practice

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

recap 2

probability

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

recap 3

statistics

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

recap 4

physics

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

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

recap

descriptive statistics:

we summarize the proprties of a distribution

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

descriptive statistics:

we summarize the proprties of a distribution

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

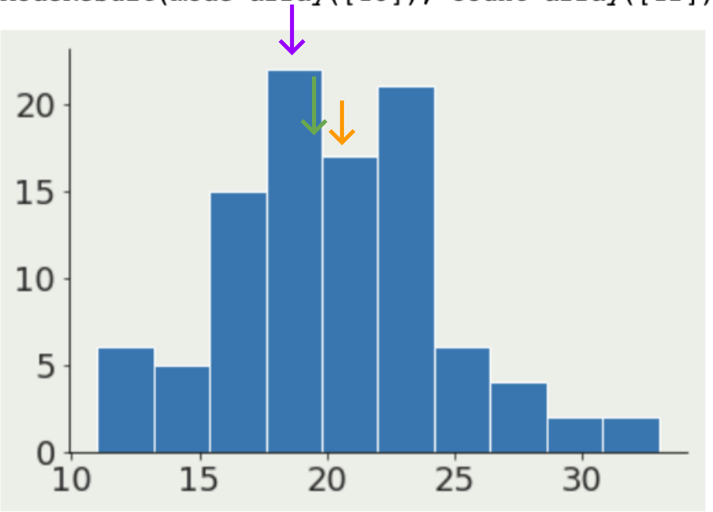
mean: n=1

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

```
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]))
```



other measures of centeral tendency:

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

mode: most popular value in the distribution

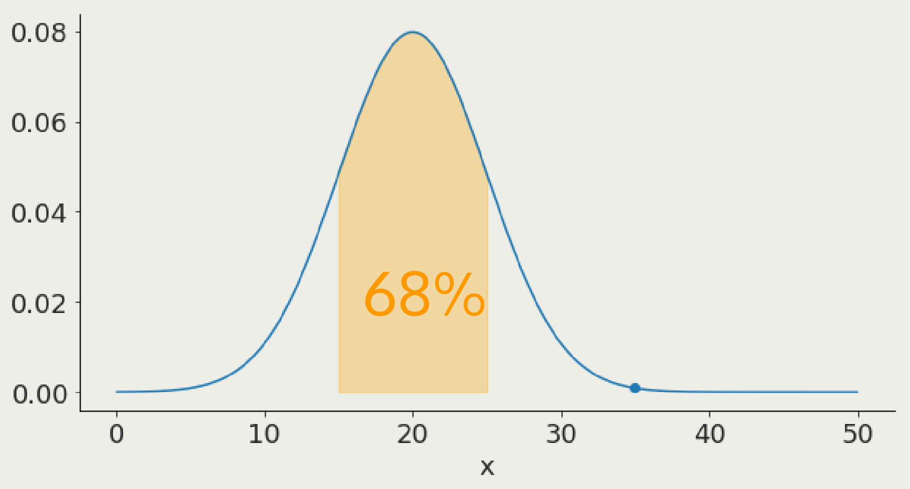
descriptive statistics:

we summarize the proprties of a distribution

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

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

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



Gaussian distribution:

1σ contains 68% of the distribution

descriptive statistics:

we summarize the proprties of a distribution

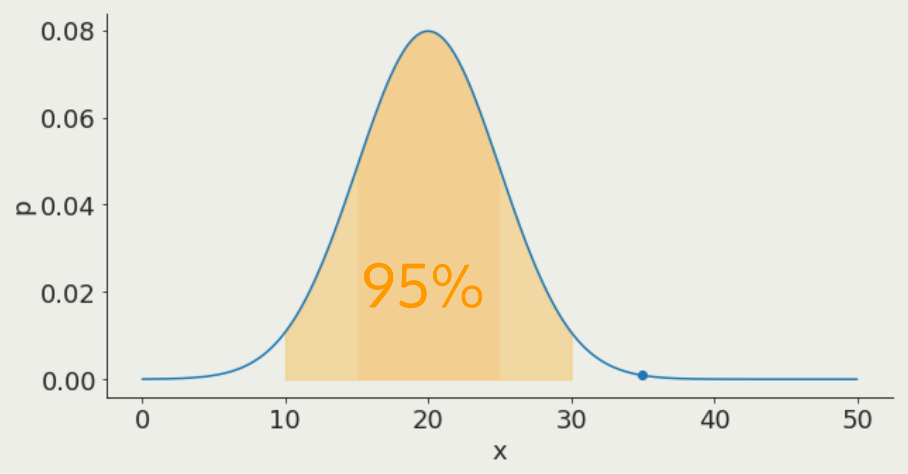
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

variance: n=2

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

standard deviation

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



Gaussian distribution:

2σ contains 95% of the distribution

descriptive statistics:

we summarize the proprties of a distribution

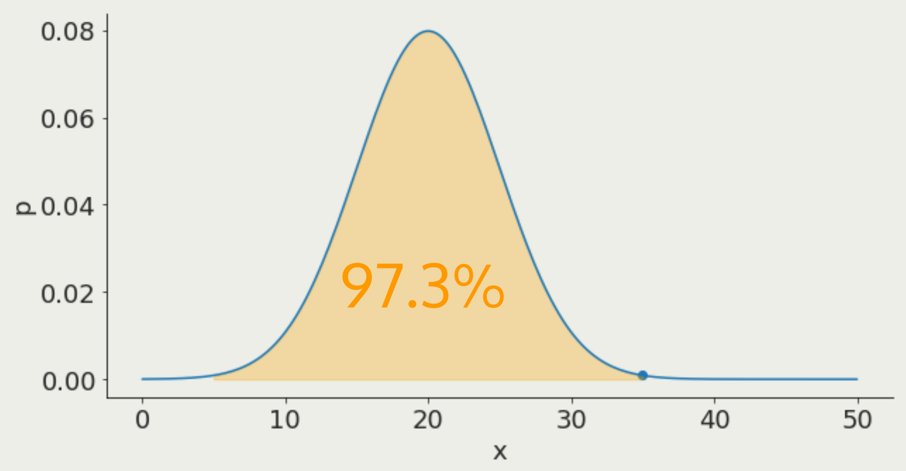
$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x.$$

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

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

Gaussian distribution:

3σ contains 97.3% of the distribution



1

the scientific method in a probabilistic context

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

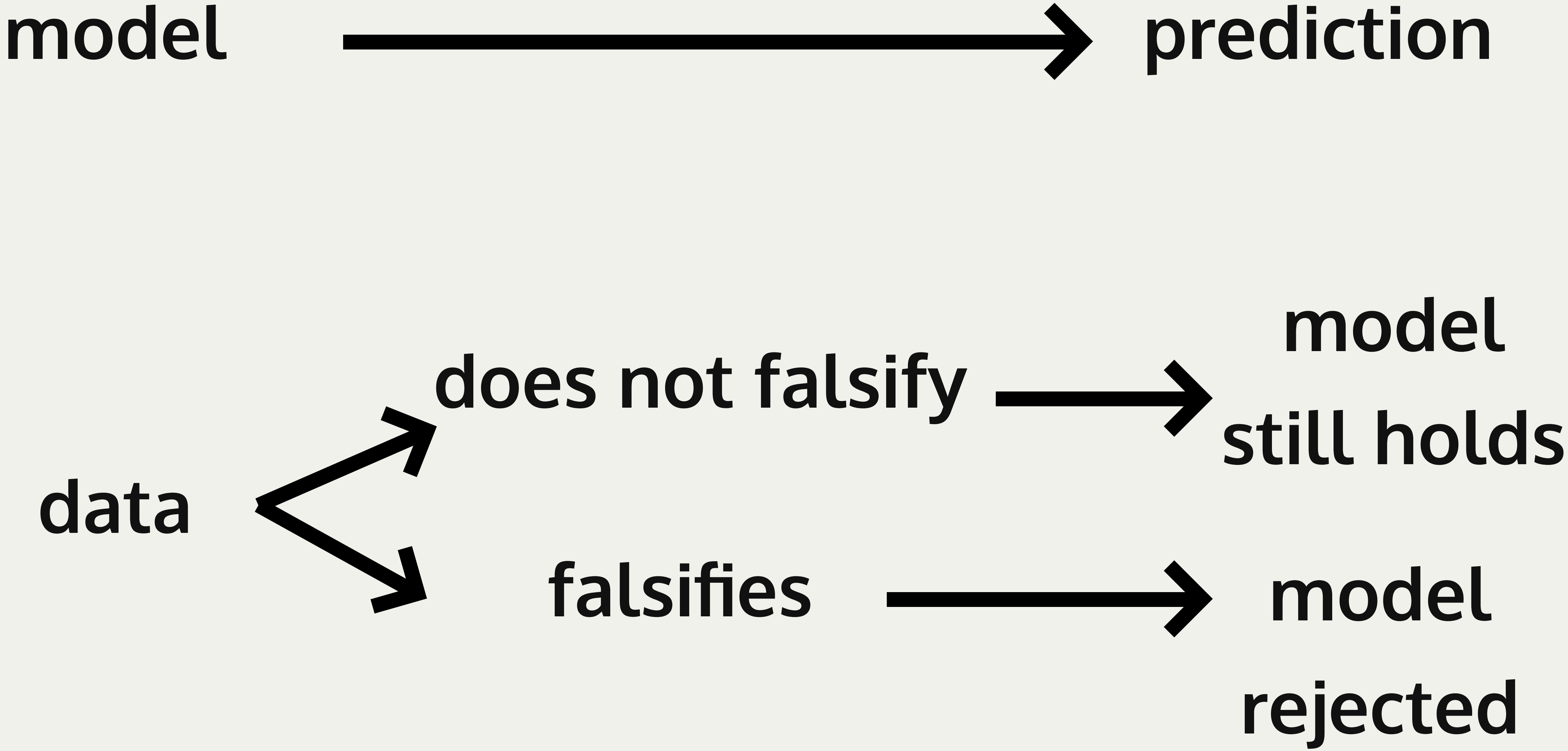
Bayesian Inference

Forward Modeling

Frequentist approach (NHRT)

p(physics | data)

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

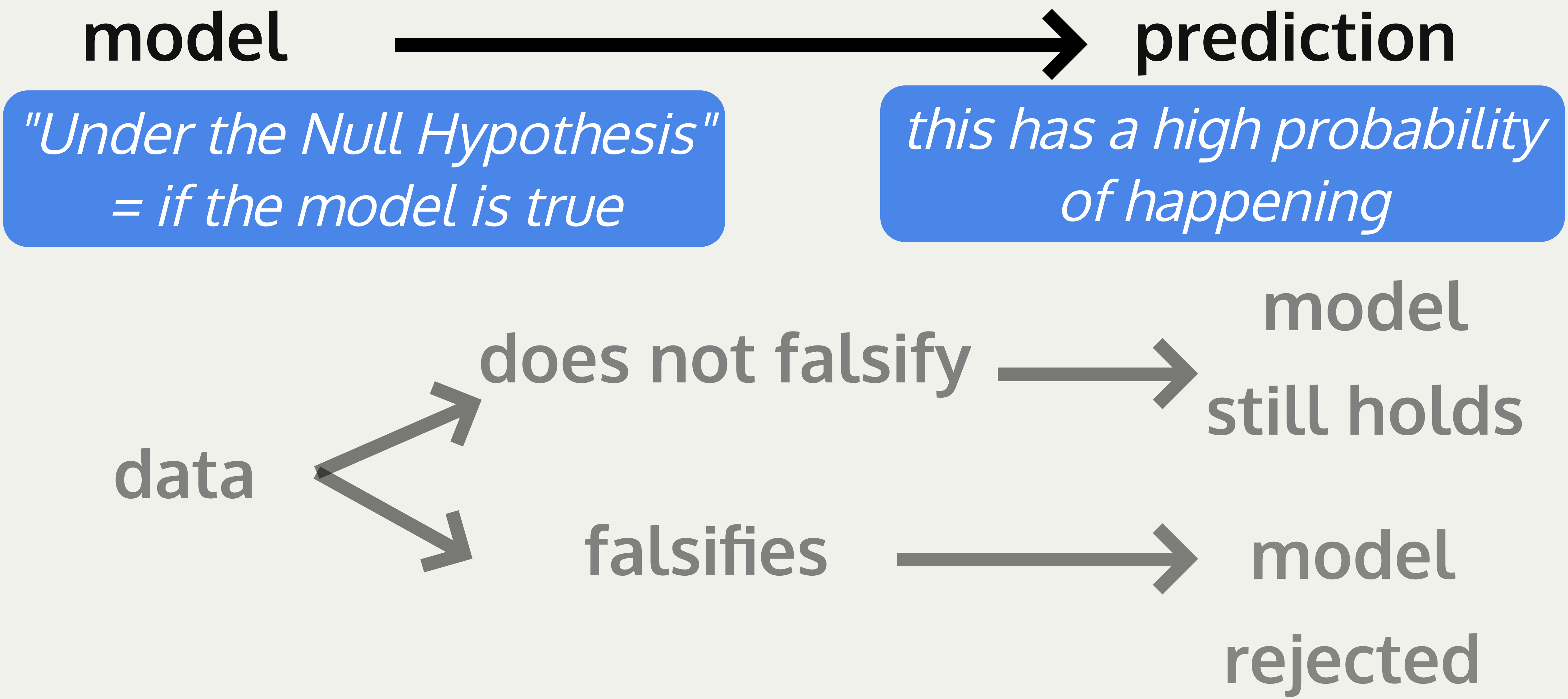

model
prediction
data
does not falsify
model
still holds
falsifies
model
rejected

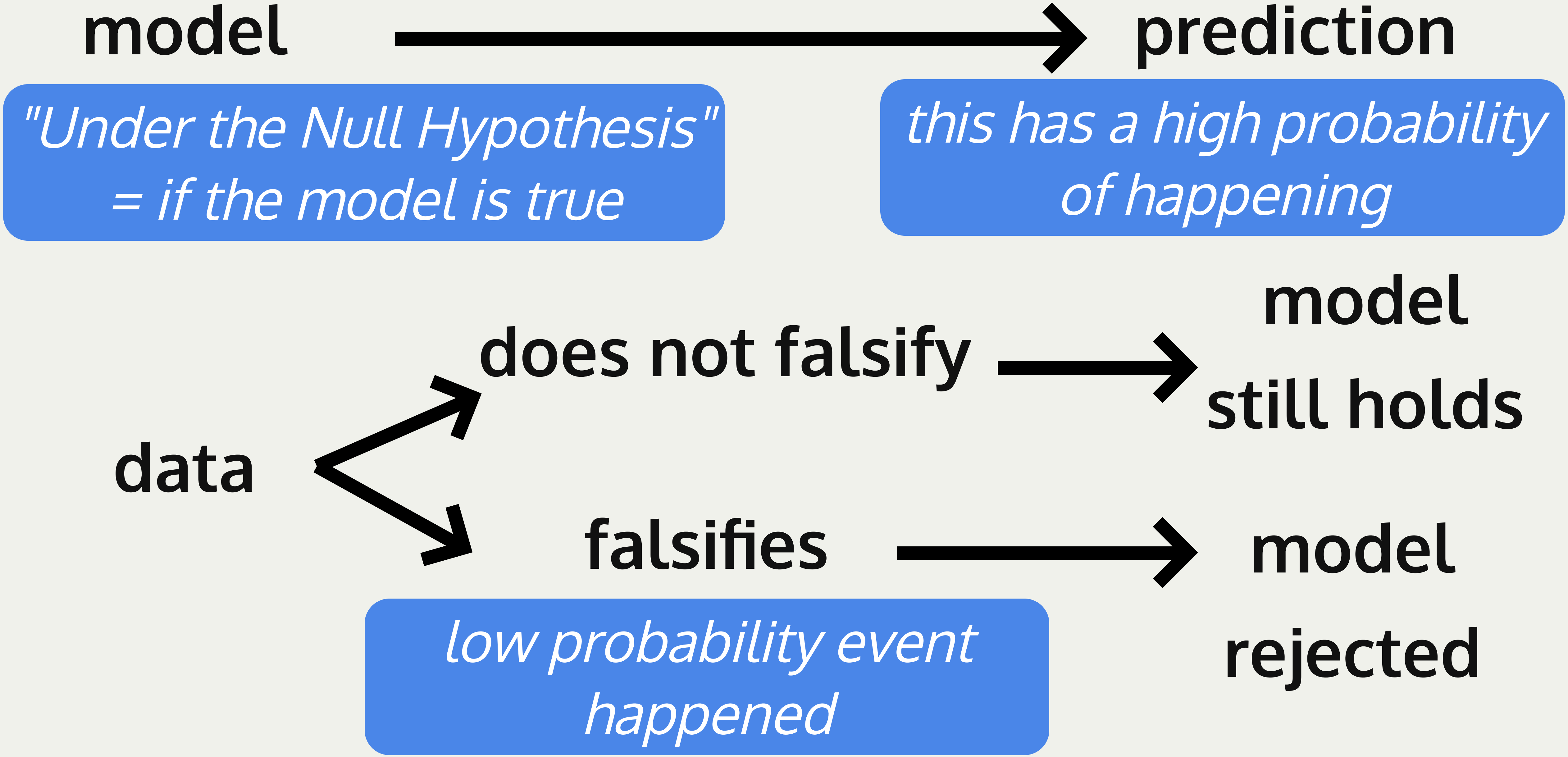
model → prediction

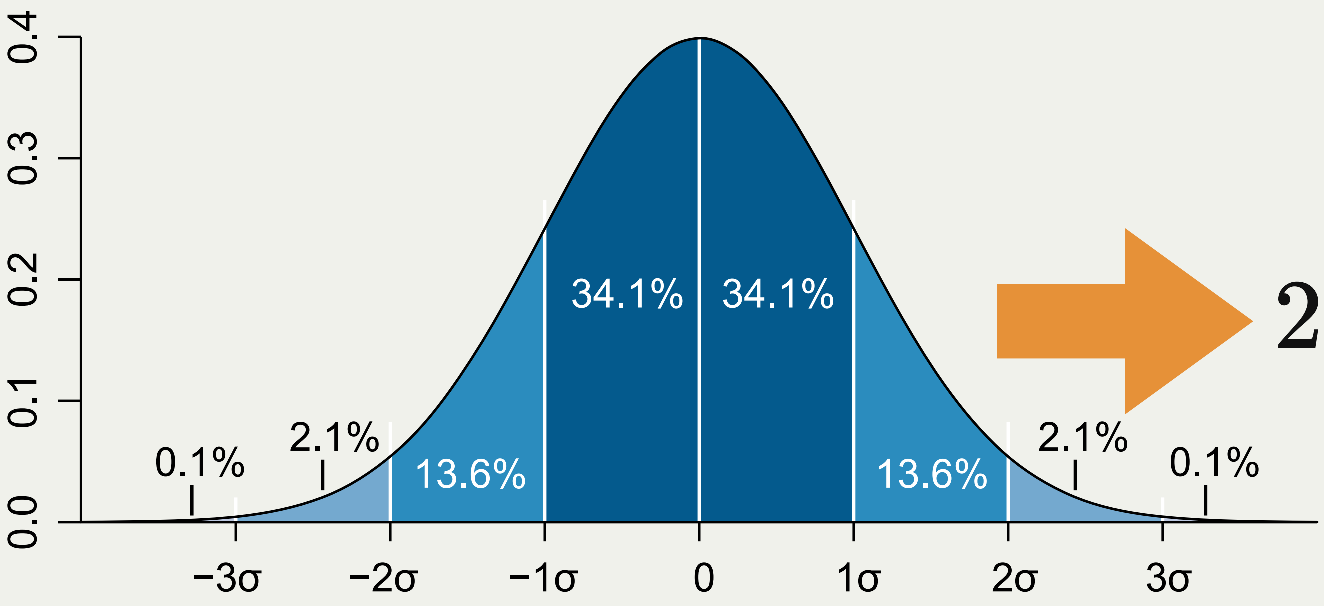
"Under the Null Hypothesis"
= if the model is true

data → does not falsify → model still holds

data → 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



2σ

rejected at 95%

0.05 p-value

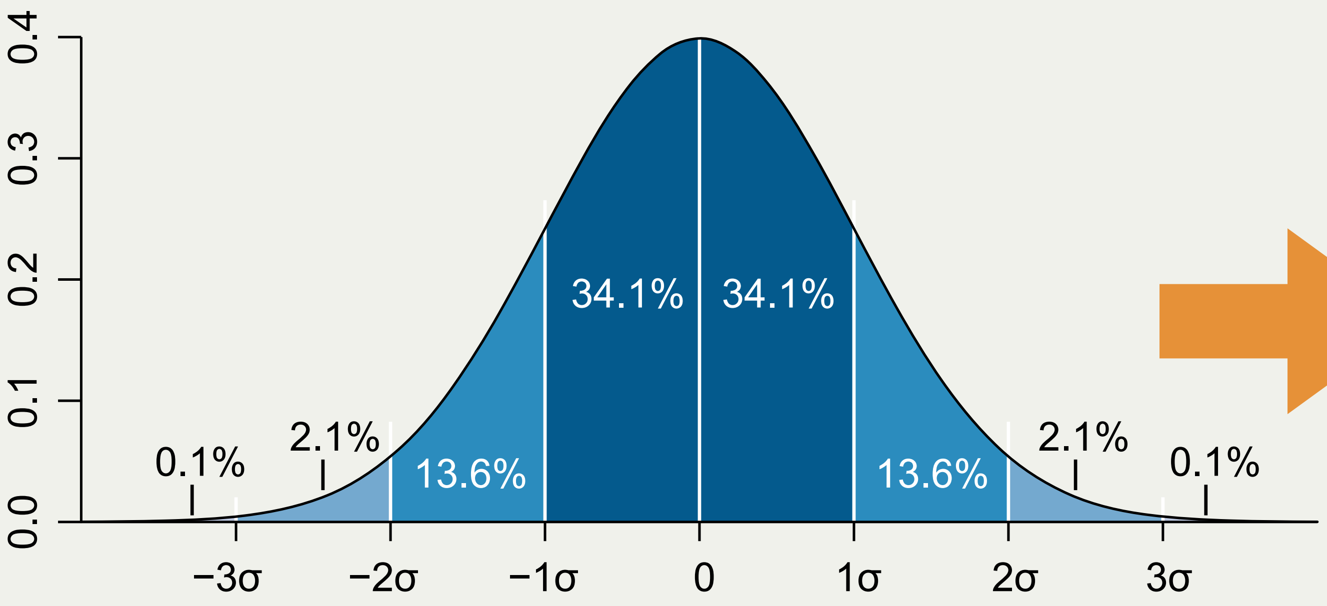
5% confidence

data

does not falsify → model still holds

falsifies → model rejected

low probability event happened



rejected at 99.7%

3σ 0.003 p-value

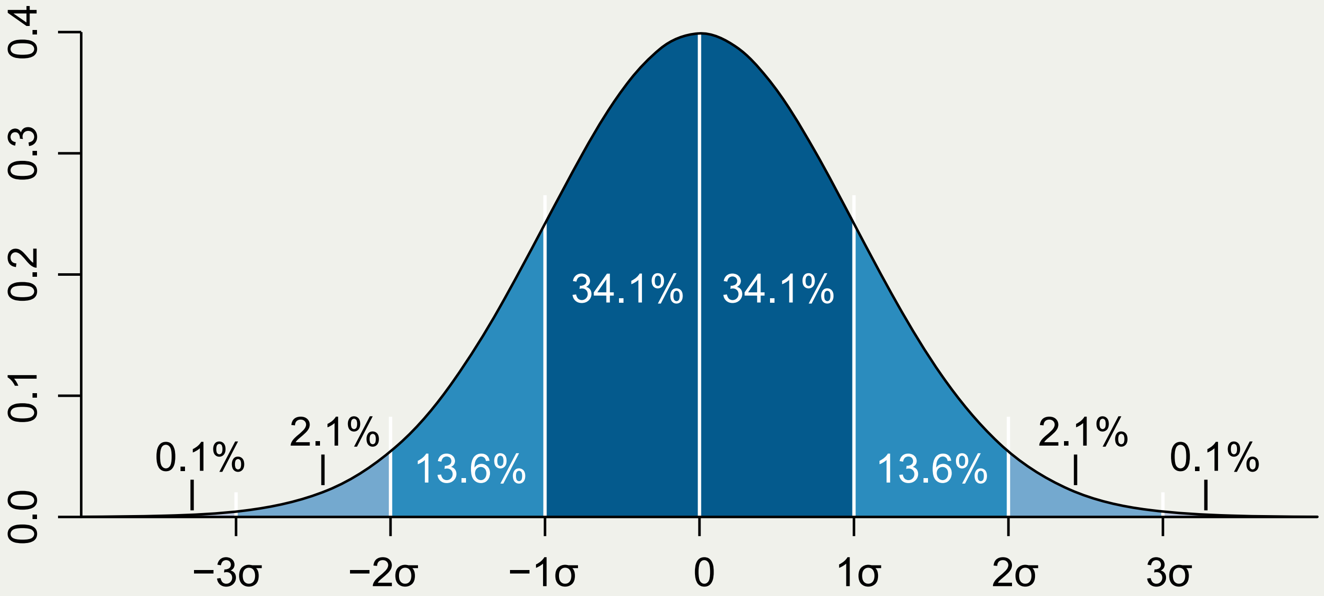
0.3% confidence

data

does not falsify → **model still holds**

falsifies → **model rejected**

low probability event happened



rejected at 99.99...%

3-e7 p-value

3-e5% confidence

→ 5σ

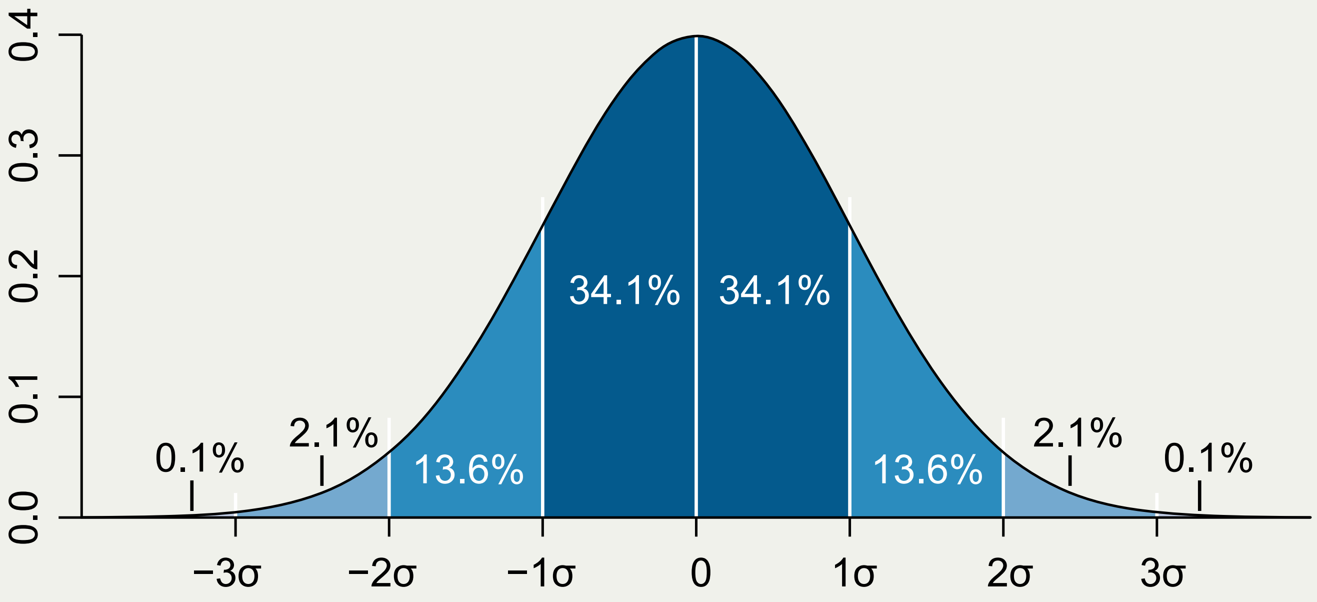
data → does not falsify → model still holds

data → falsifies → model rejected

low probability event happened

2 Null hypothesis rejection testing

Null
Hypothesis
Rejection
Testing



p(physics | 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

data

does not falsify → Null still holds

falsifies → Null rejected

generally, out 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

data → does not falsify → Null still holds

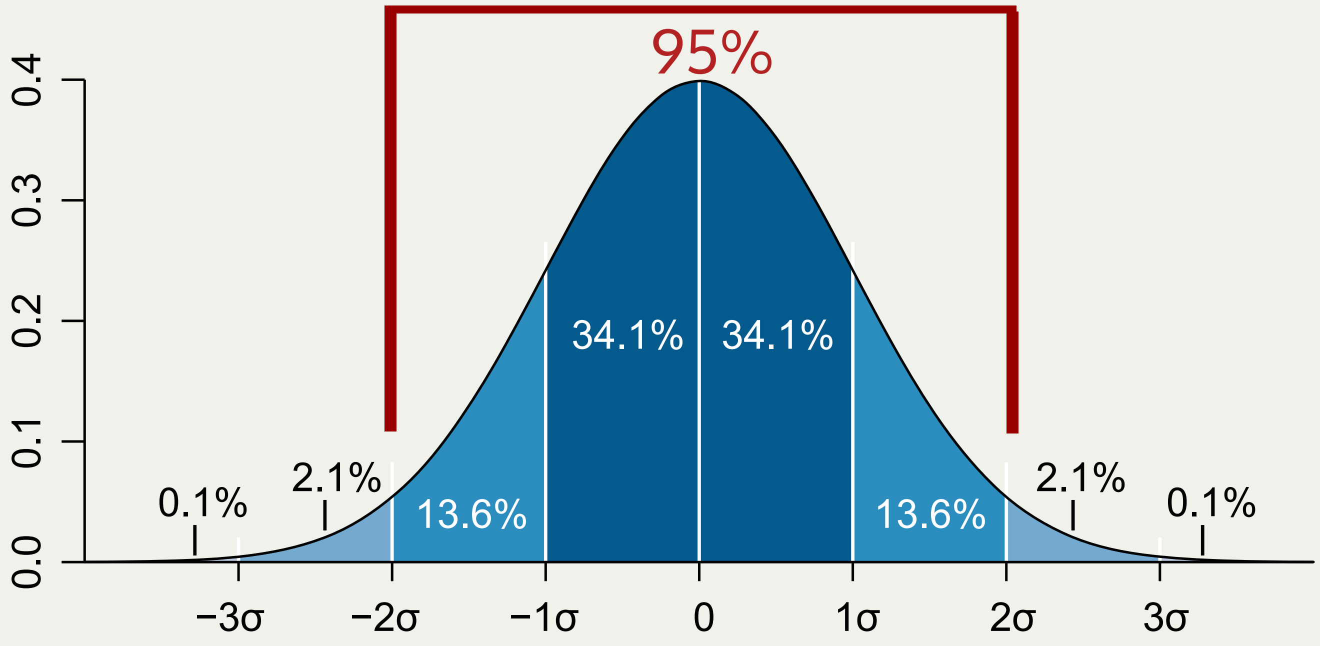
data → falsifies → Null rejected

Earth is flat is Null

Earth is round is Alternative:

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

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

pivotal quantities

quantities that under the Null Hypotheis 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 Hypotheis 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

☺

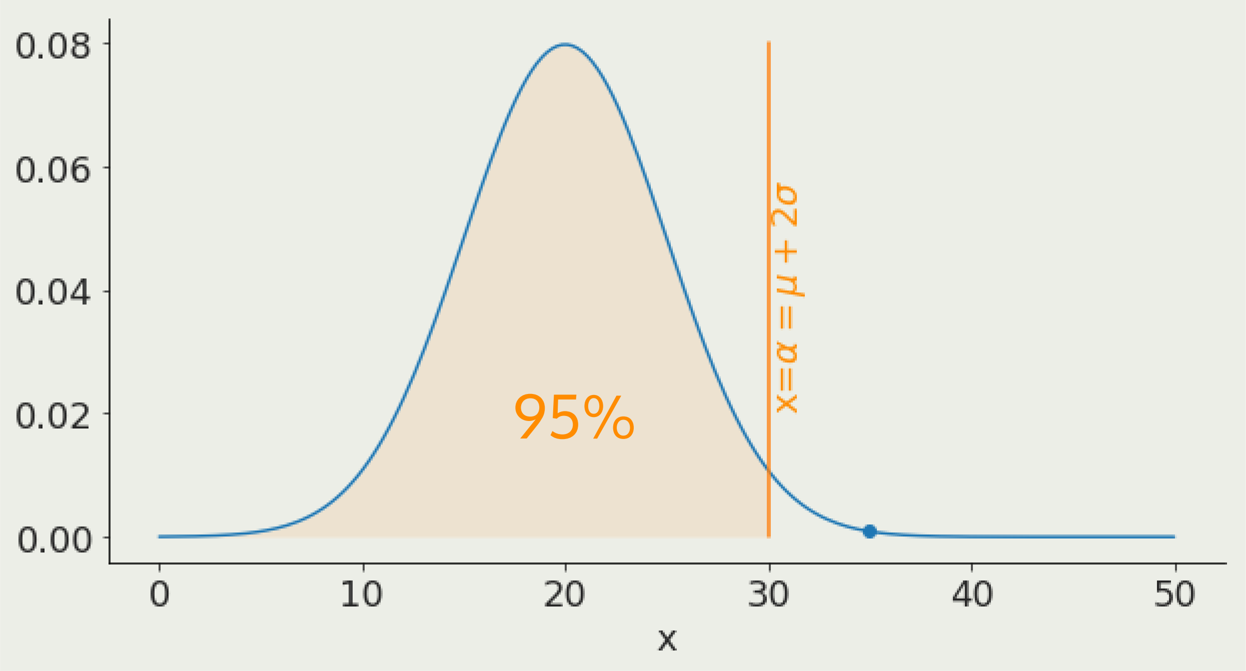
4

test data against
alternative outcomes

Null
Hypothesis
Rejection
Testing

what is α?

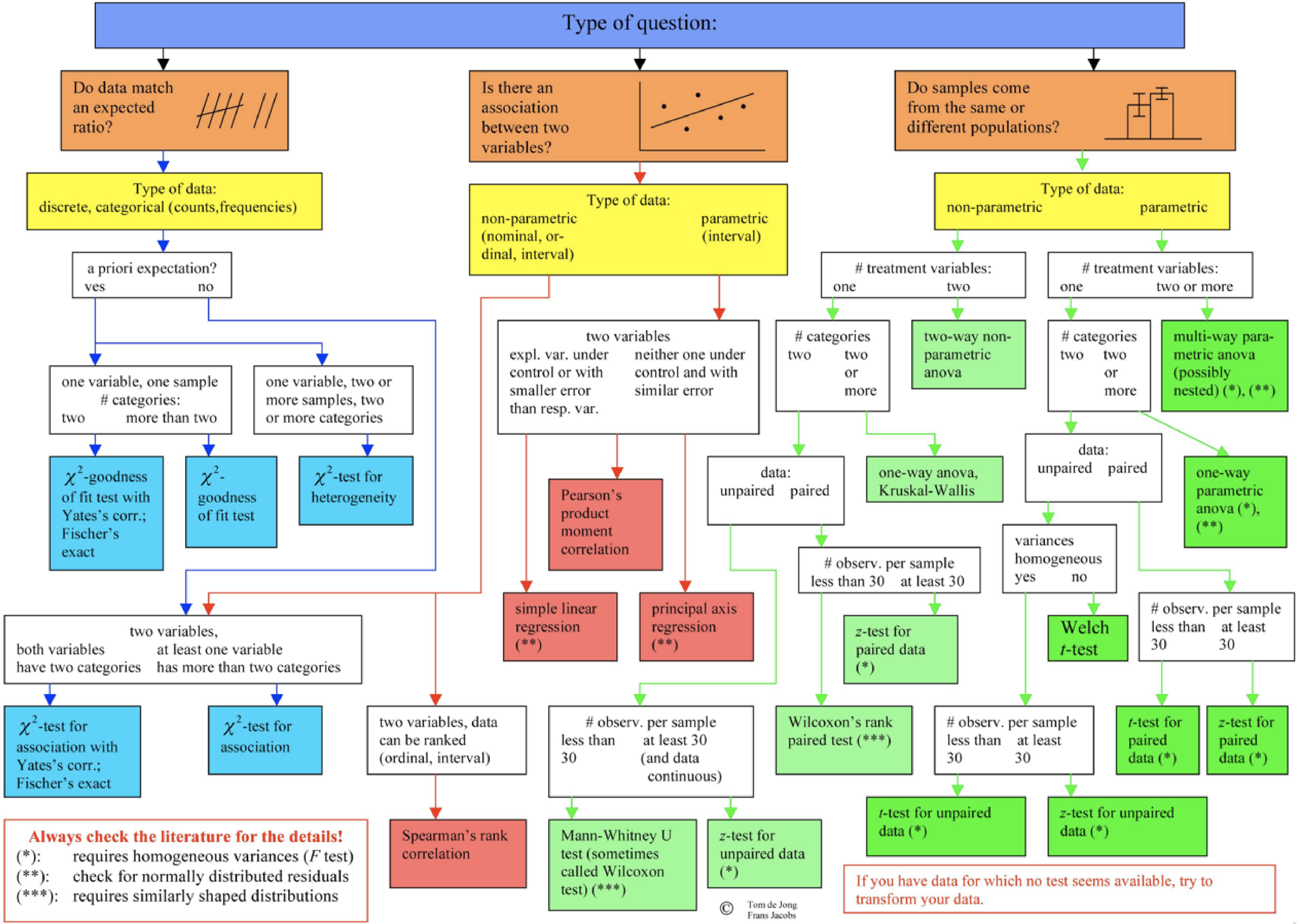
α is the x value corresponding to a chosen threshold



4

test data against
alternative outcomes

3 common tests and pivotal quantities


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

pivotal quantities

quantities that under the Null Hypotheis follow a known distribution

also called "statistics"

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

Z statistics: difference between means

K-S statistics: maximum distance of cumulative distributions.

Z–test

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

pivotal quantity

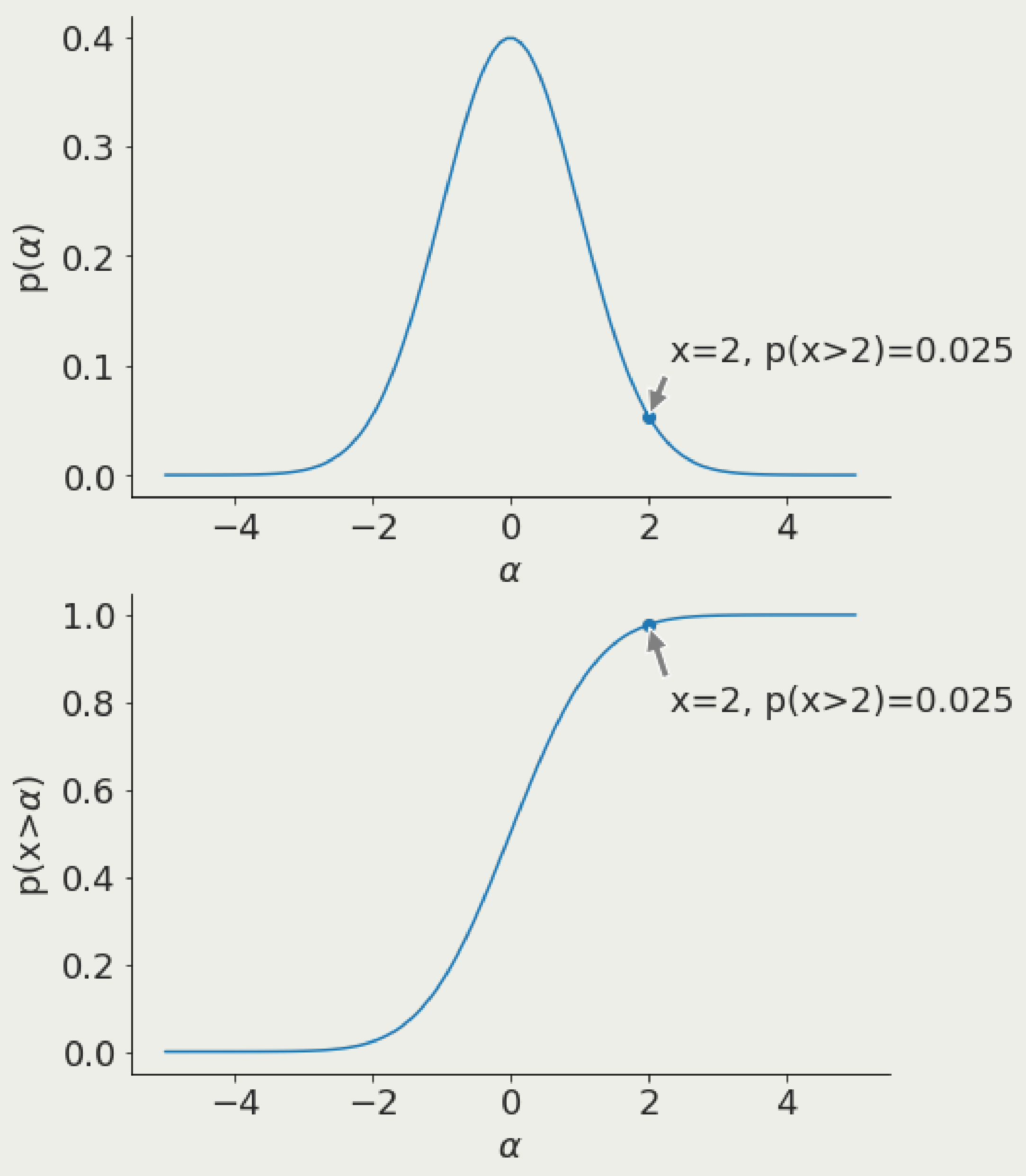
$$Z = (\bar{X} - \mu 0)/s$$

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

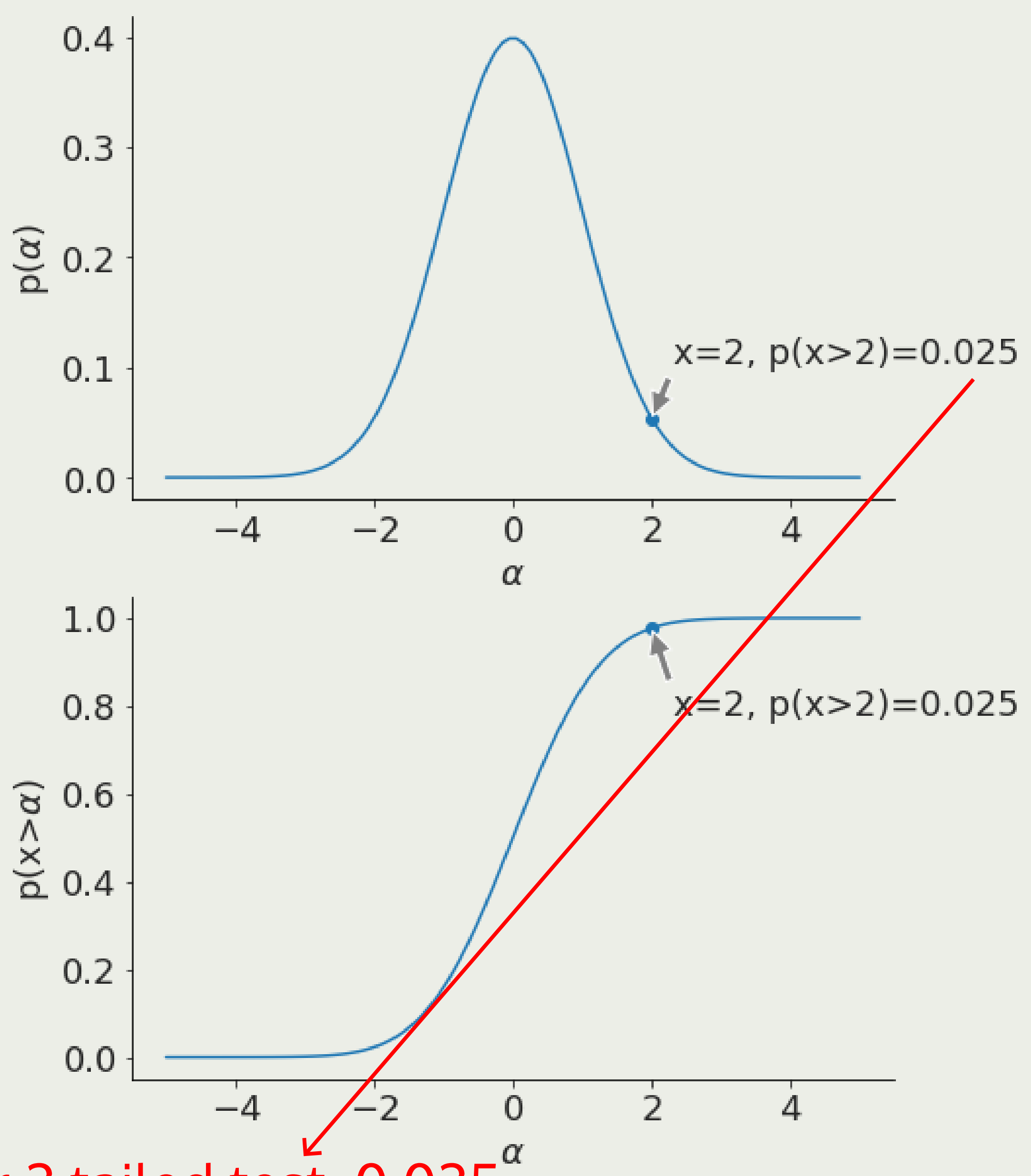
$$Z = (\bar{X} - \mu 0)/s$$

sample mean

population mean

sample variance

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



NOTE: 2-σ is p=0.05 for 2 tailed test, 0.025 for 1 tailed test

Z–test

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

pivotal quantity

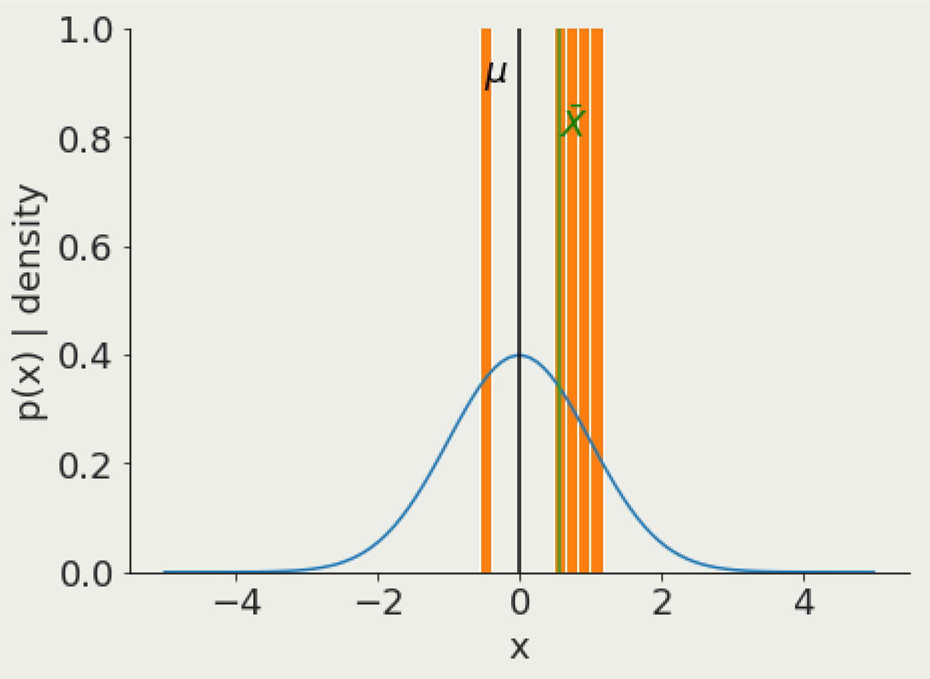
$$Z = (\bar{X} - \mu 0) / s$$

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 it they are the same?

Z–test

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

pivotal quantity

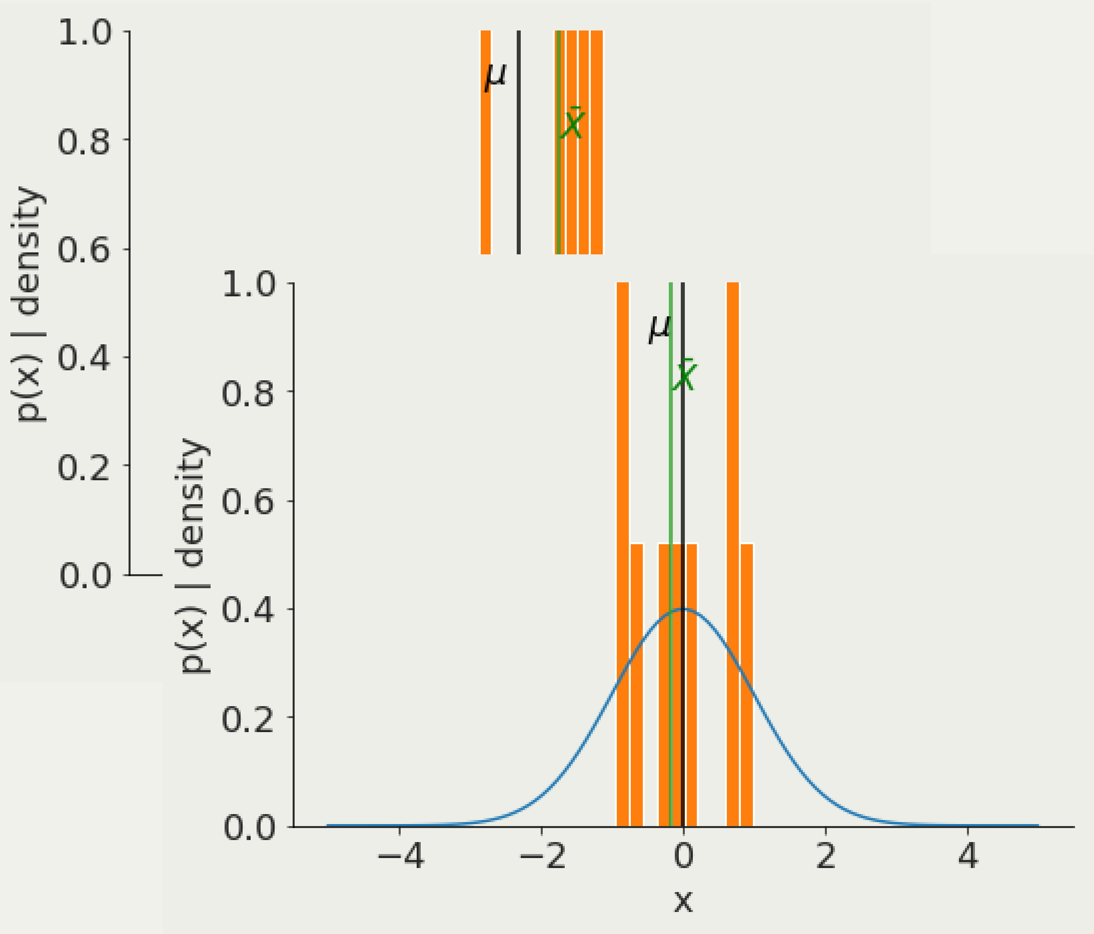
$$Z = (\bar{X} - \mu 0) / s$$

sample mean

population mean

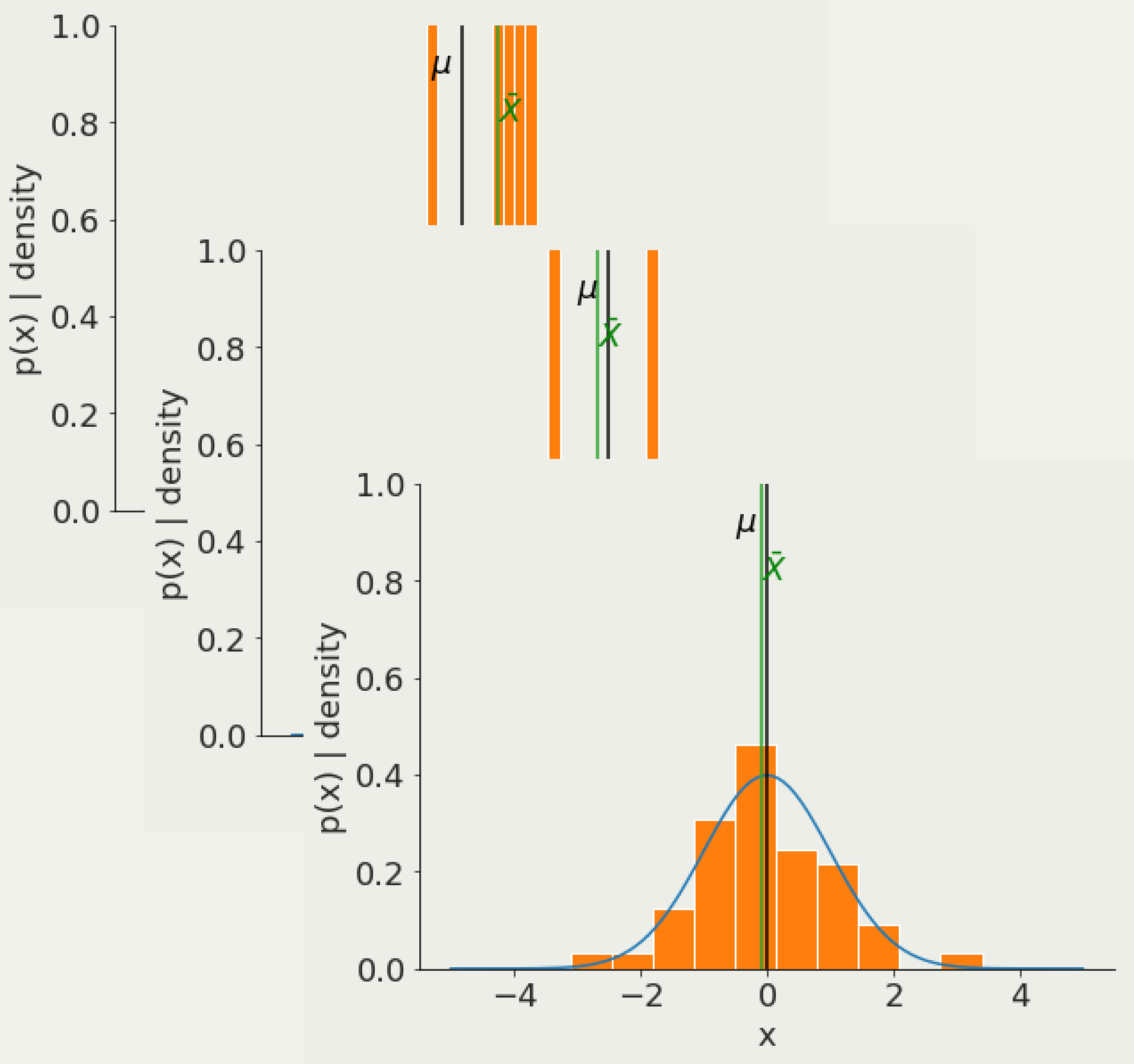
sample variance

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



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

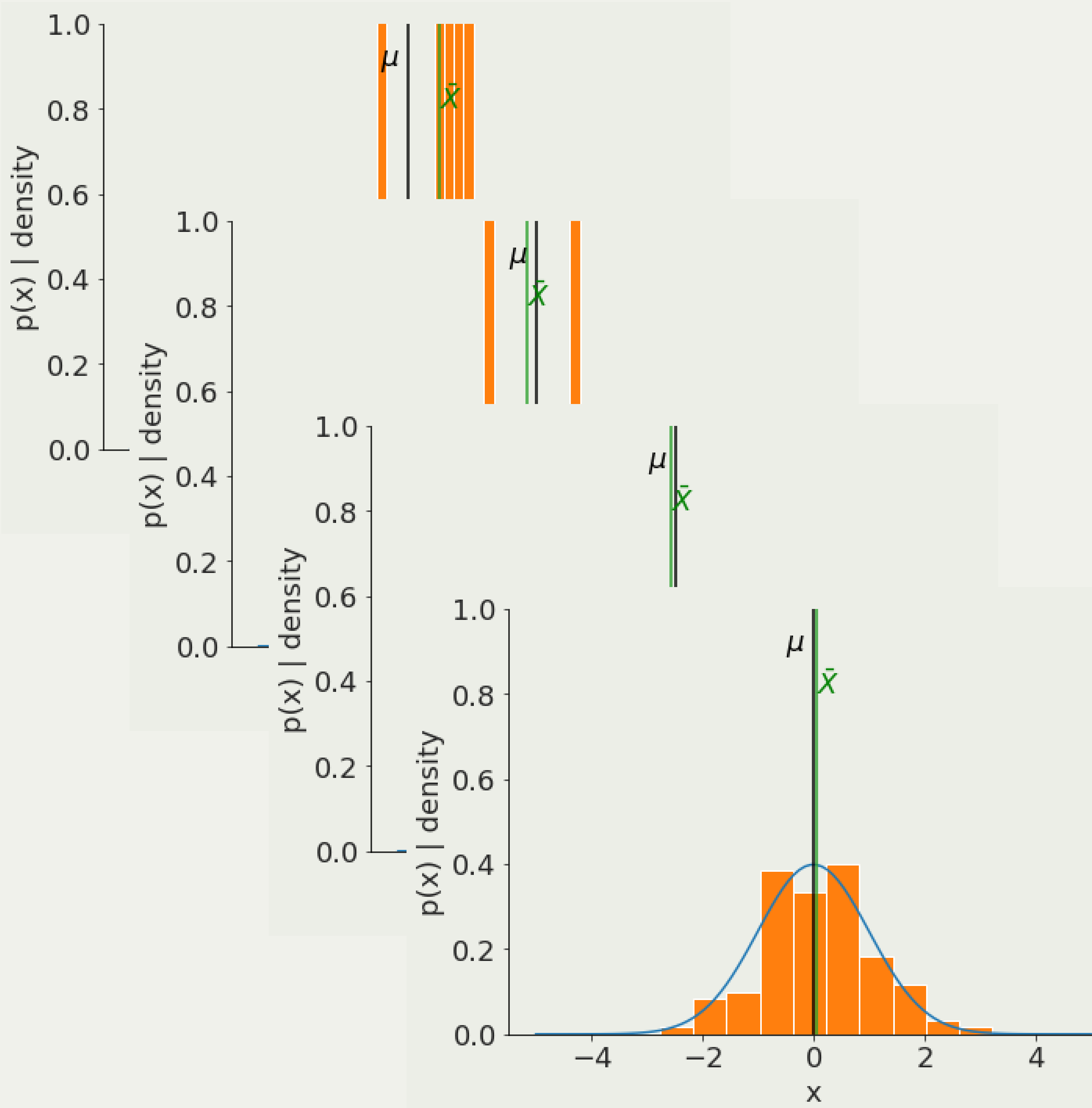
$$Z = (\bar{X} - \mu 0)/s$$

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

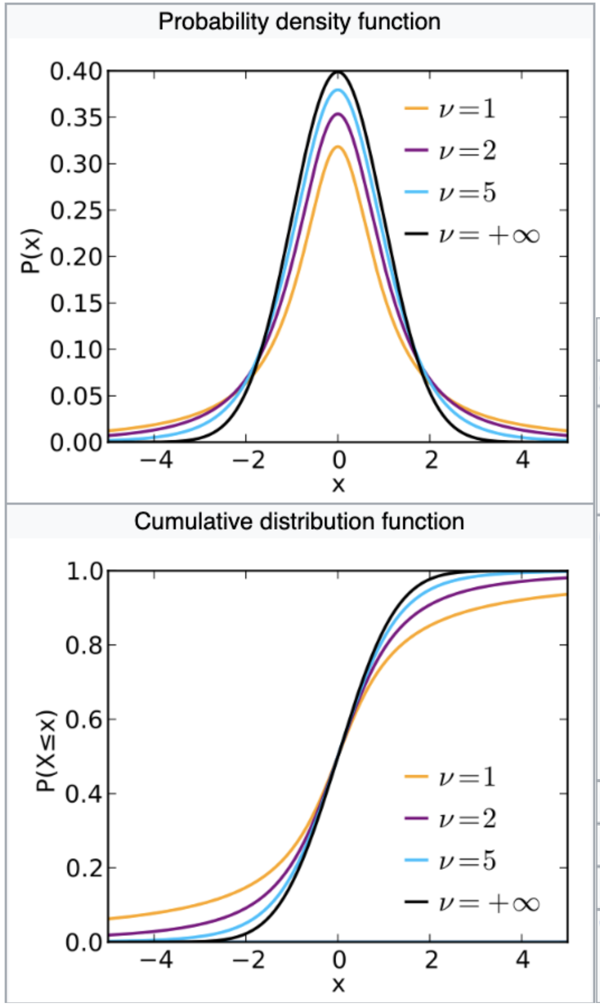
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unbias variance estimator

size of sample

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

Student's t



WIKIPEDIA
The Free Encyclopedia

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

$t-$ test

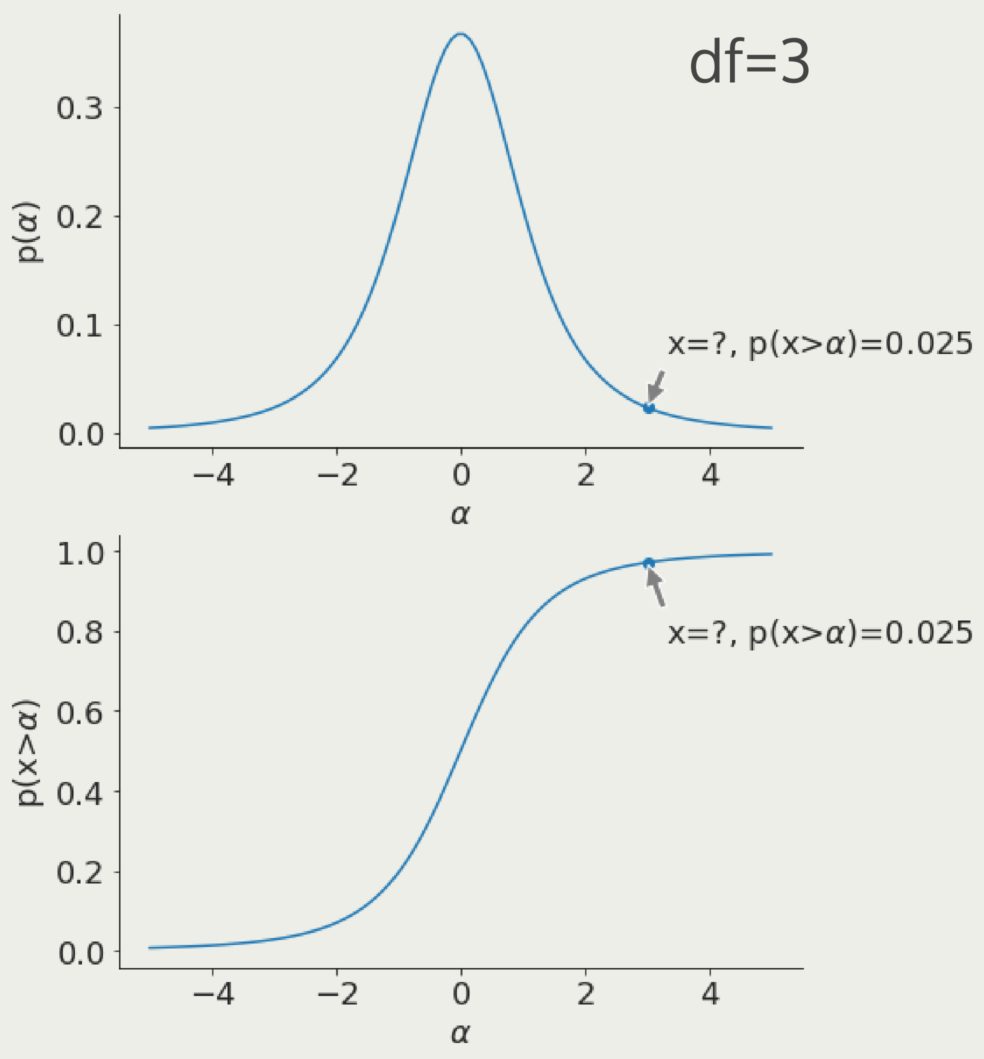
Are the means of 2 samples significantly different?

pivotal quantity

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

unbias variance estimator

size of sample



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

t– test

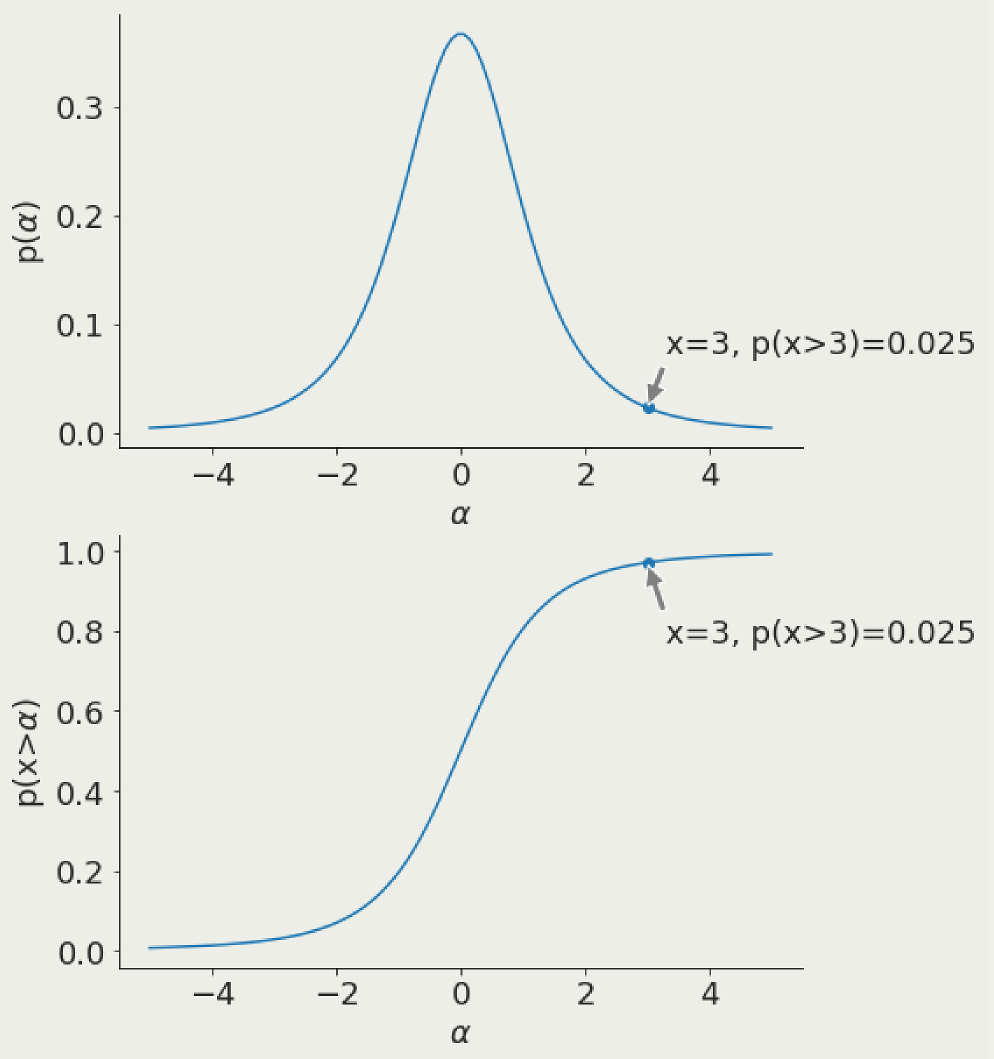
Are the means of 2 samples significantly different?

pivotal quantity

unbias variance estimator

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

size of sample

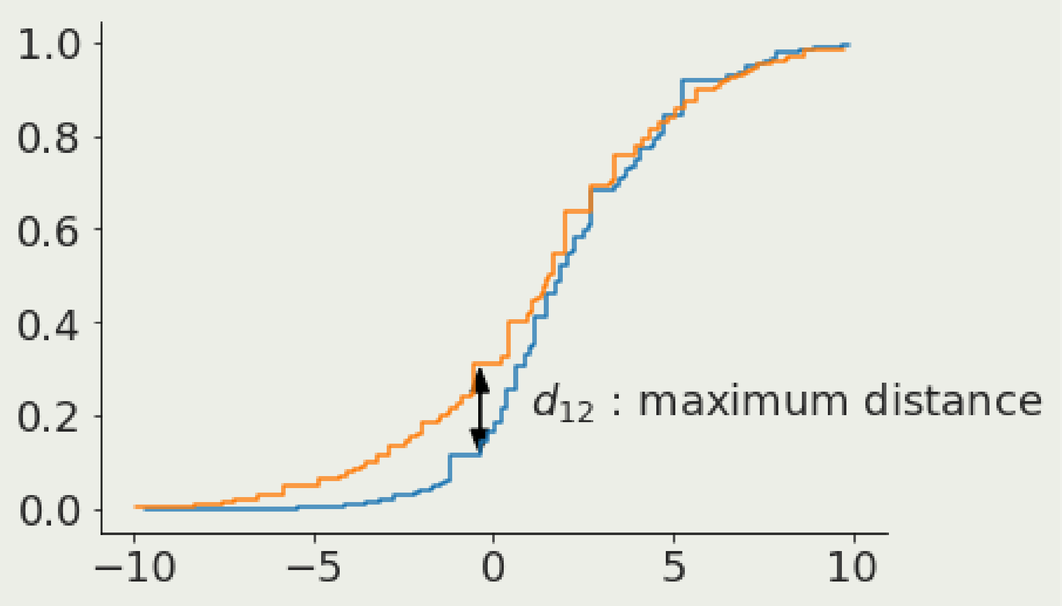


$$t \sim \text{Student's } t\left(\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}\right)$$

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

$$P(d > observed) \; = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \sqrt{\frac{N_1N_2}{N_1+N_2}D}$$

K-S test

Kolmogorof-Smirnoff :

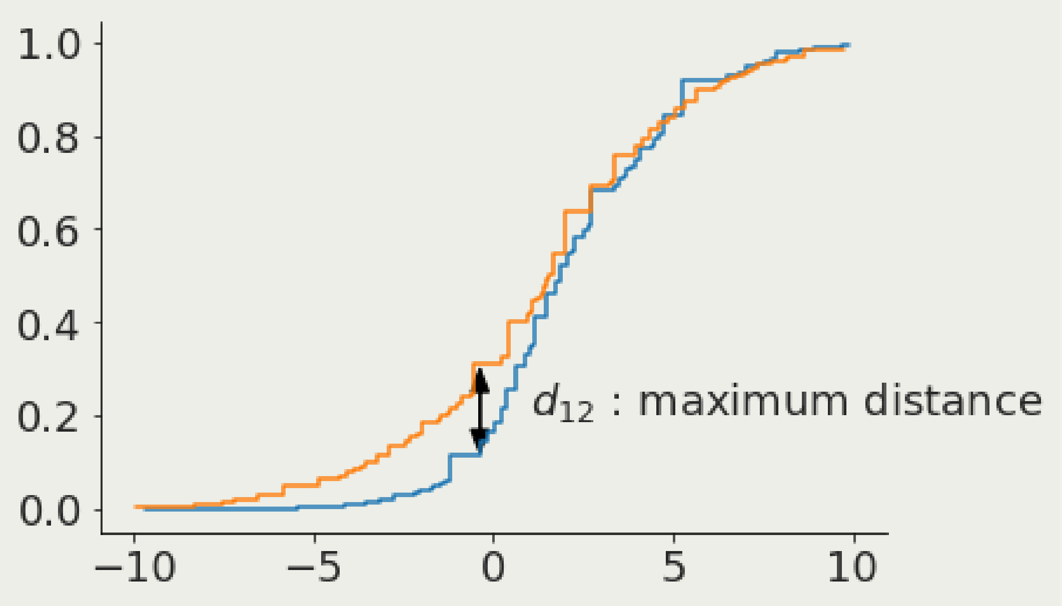
do two samples come from the same parent distribution?

pivotal quantity

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

Cumulative distribution 1

Cumulative distribution 2



$$P(d > observed) =$$

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

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

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

$\chi 2$ test

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

pivotal quantity

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

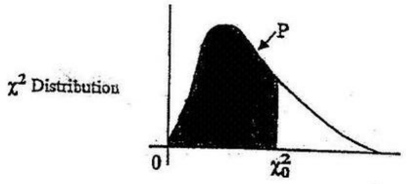
model

observation

uncertainty

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

number of observation



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

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

$\chi 2$ test

are the data what is expected from the model (if likelihood is Gaussian... we'll see this later) - ther are a few $\chi 2$ 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

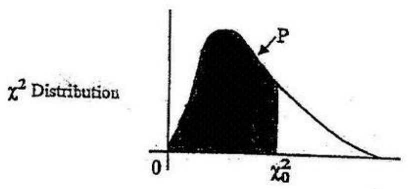
uncertainty

observation

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

number of observation

χ^2 Distribution



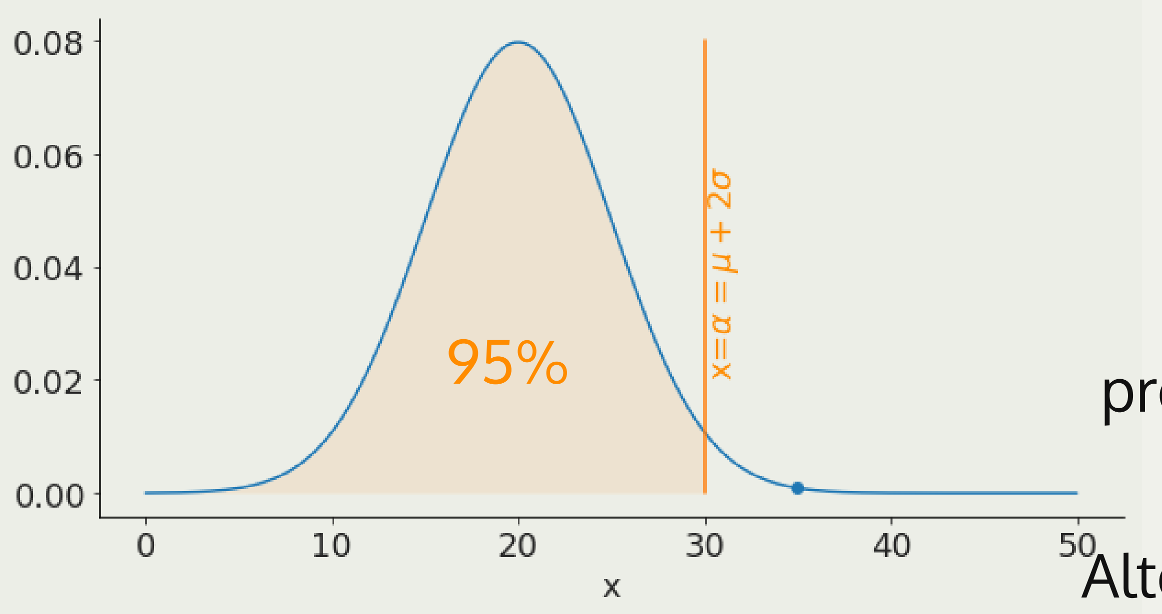
The table below gives the value x_0^2 for which $P[x^2 < x_0^2] = P$ for a given number of degrees of freedom and a given value of P.

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

Null
Hypothesis
Rejection
Testing

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

prediction is unlikely
Null rejected
Alternative holds



prediction is likely
Null holds
Alternative rejected

4

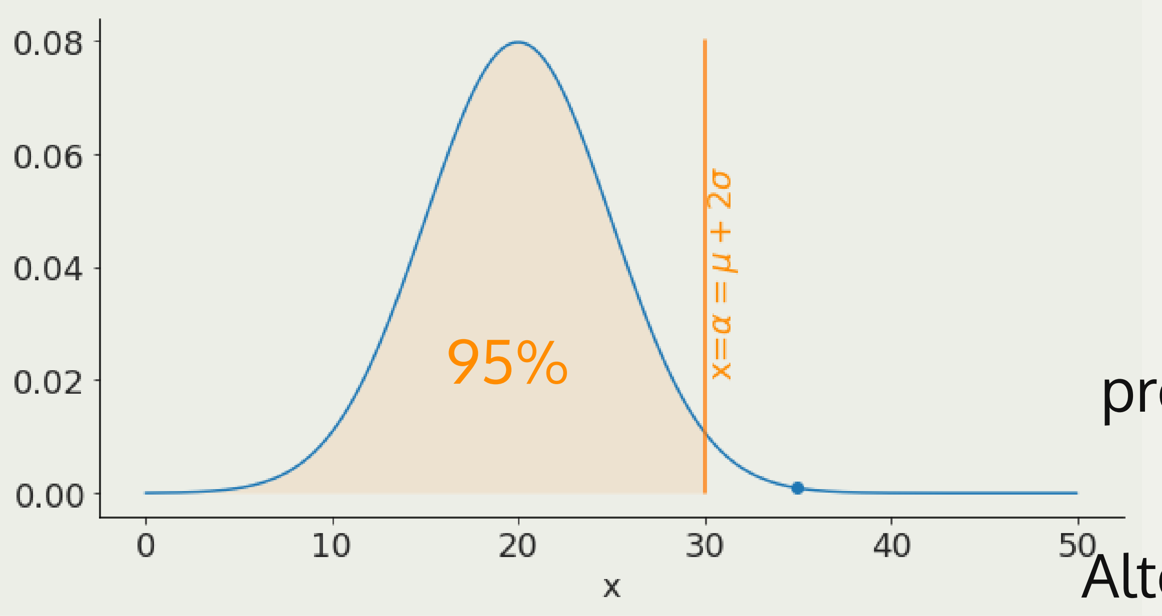
test data against
alternative outcomes

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

Null
Hypothesis
Rejection
Testing

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

prediction is unlikely
Null rejected
Alternative holds



prediction is likely
Null holds
Alternative rejected

4

test data against
alternative outcomes

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

formulate the Null as the comprehensive opposite of your theory

model → prediction

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

this has a low probability of happening

data

does not falsify alternative → everything but model is rejected

falsifies alternative → model holds

Key Slide

low probability event happened

the *demarcation* problem in *Bayesian* context

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

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

the *demarcation* problem in *Bayesian* context

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

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

$$p(M|D) =$$

the *demarcation* problem in *Bayesian* context

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

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

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

"prior"

the *demarcation* problem in *Bayesian* context

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

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

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

the *demarcation* problem in *Bayesian* context

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

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

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

"evidence"

the *demarcation* problem in *Bayesian* context

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

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

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

key concepts

descriptive statistics

null hypothesis rejection testing setup

pivotal quantities

Z, t, χ2, K-S tests

homework

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

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008

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

https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C

David M. Lane et al.

Introduction to Statistics (XVIII)

http://onlinestatbook.com/Online_Statistics_Education.epub

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

resources