principles of Urban Science 3

NHRT

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

1. Reading in data
2. Descriptive statistics (central tendency, spread...)
3. Extracting descriptive statistics from data

1. overfitting
2. p-value inference
3. mapping in python (intro to geopandas)

this slide deck: https://slides.com/federicabianco/pus2020_3

quizz: https://forms.gle/FGKX9fy6bEXHe4A39

1 descriptive statistics

Preamble: kinds of analytical questions

- Are two measurements the same?
  - *is the amount of nitrates in Lums pond same as it was 2 years ago?*
- Are two distributions the same?
  - *is the weight of Medicare members signed up for health newsletters the same as that of members who are not signed up?*
- Can I trust that a number comes from a certain distribution? -> ***p***-value

measuring differences

are these 2 numbers the same?

clearly 1.2 ≠ 1.8

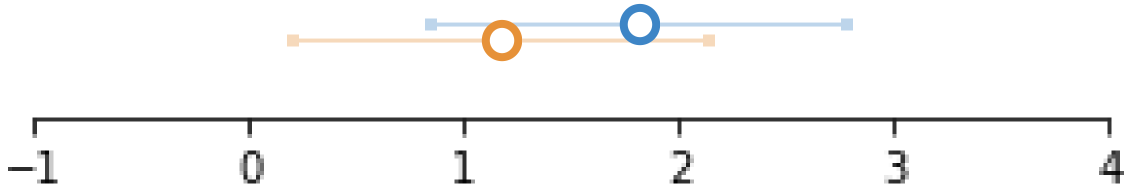
measuring differences

are these 2 numbers the same?

two numbers are never actually the same, but we understand that there are limitations in how well numbers represent reality

1.2+/- 1 = 1.8+/- 1

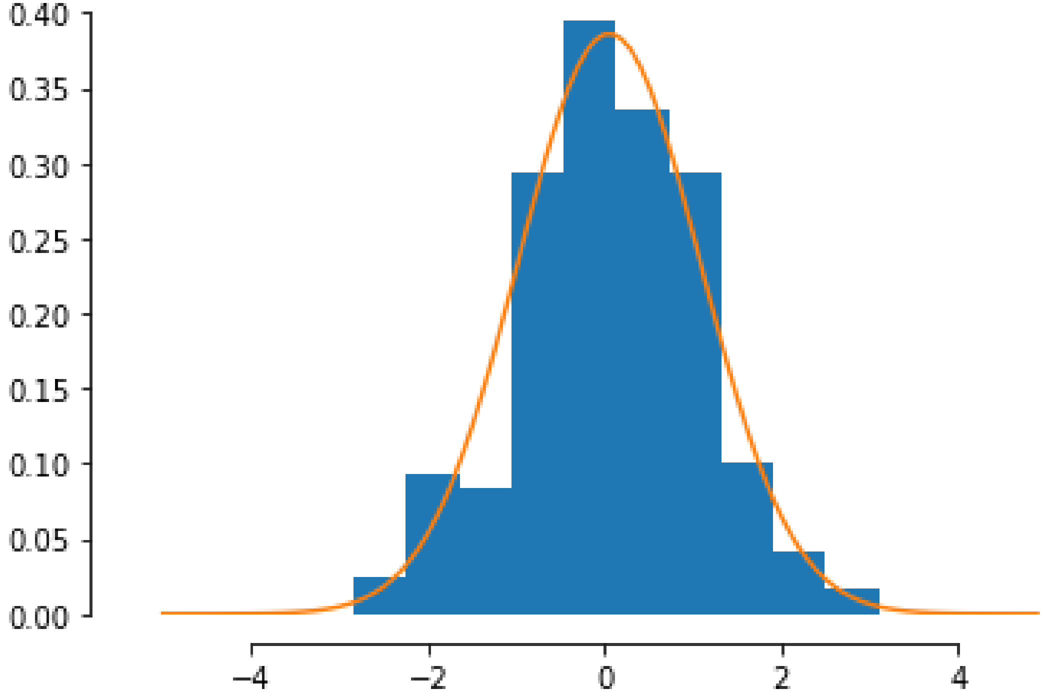
because the [0.2-2.2] interval overlaps the [0.8-2.8] interval



distribution

All data has some element of randomness either because:

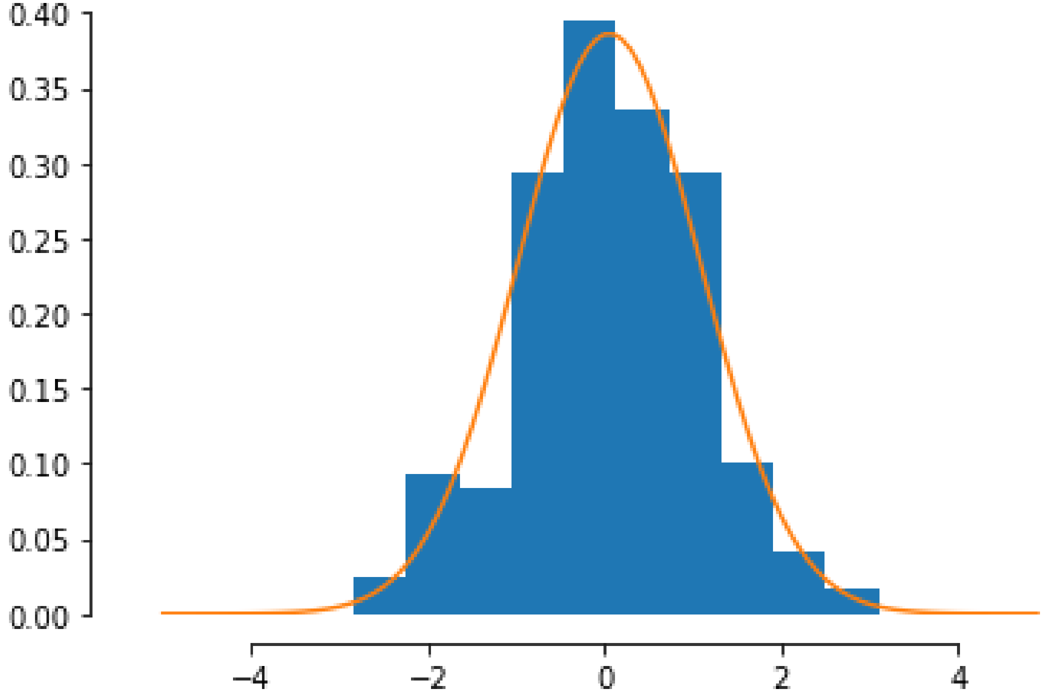
- there is randomness in the way it is generated
- there is uncertainty in the way it is measured
- both (in most cases it's both)



distributions

All data has some element of randomness either because:

- there is randomness in the way it is generated
- there is uncertainty in the way it is measured
- both (in most cases it's both)

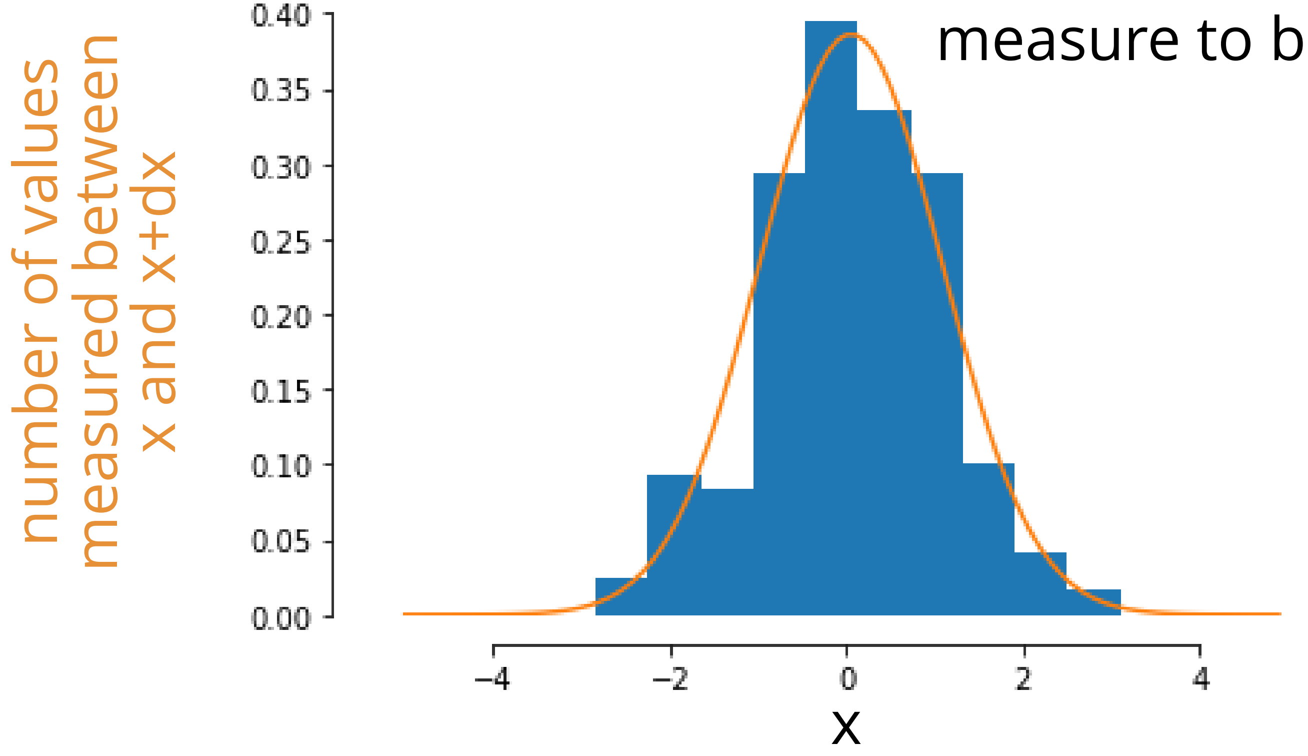


we think of data points as a number extracted from a distribution. sometimes we have expectations for that distribution, sometimes we do not.

distributions

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

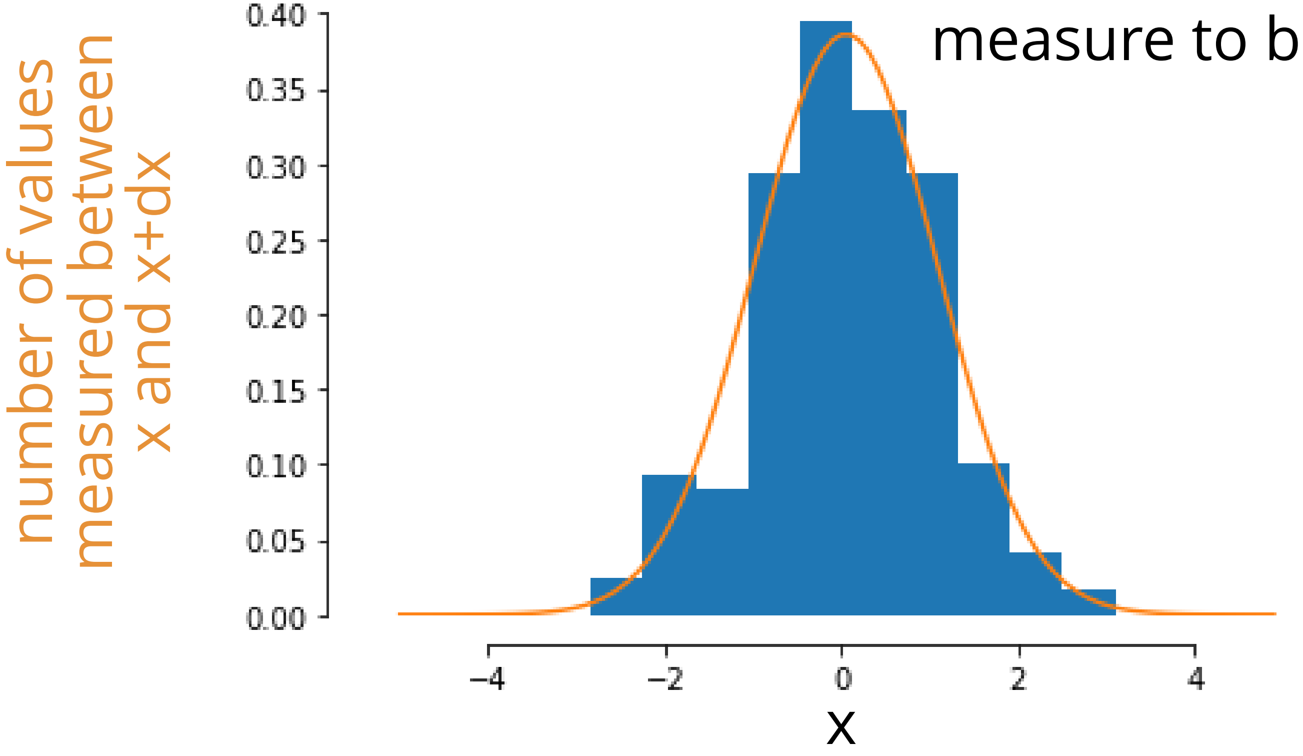


frequency ⟶ **probability**

distributions

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

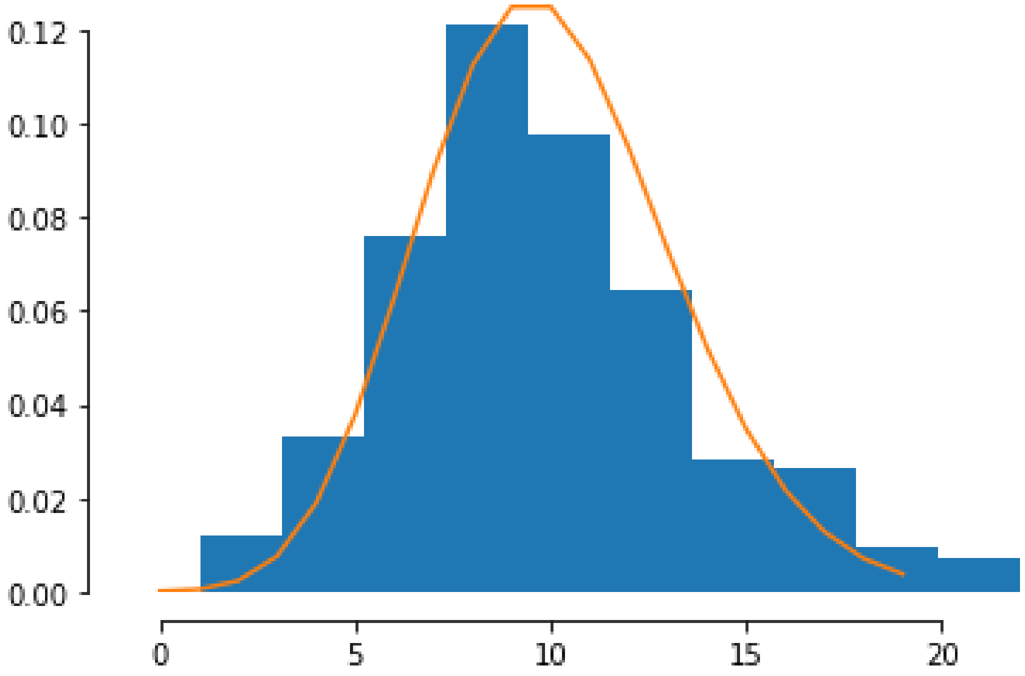
analyst approach: a distribution represent the *probability* with which a phenomenon generates a value that we measure to be ~x



frequency ⟶ **probability**

distributions

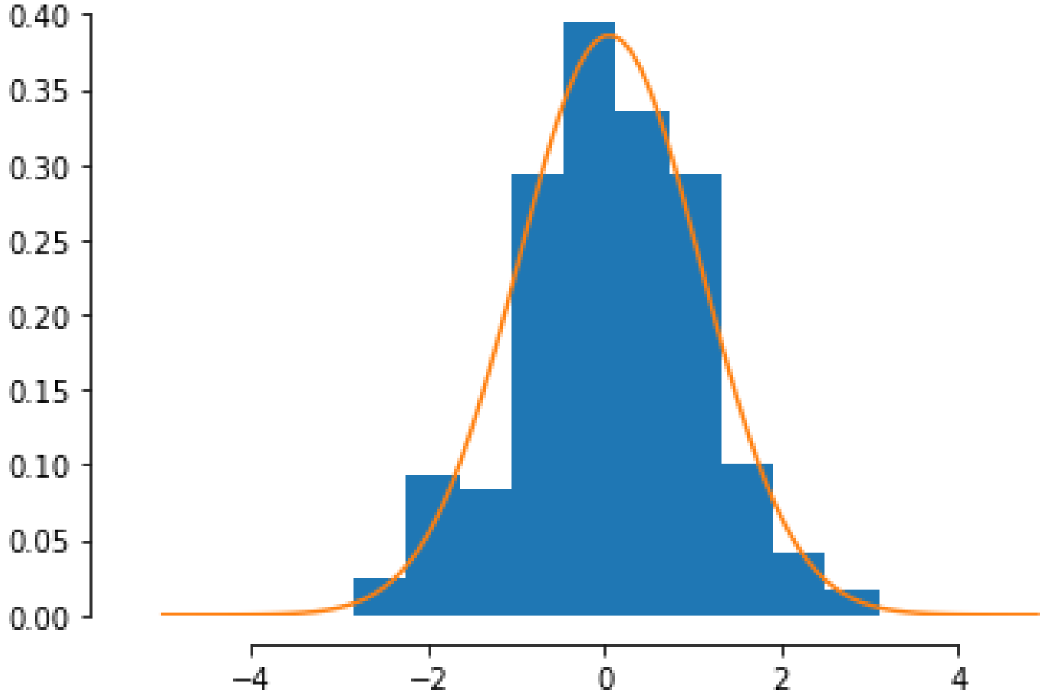
$$P(k|\lambda) \sim \frac{\lambda^k \, e^{-\lambda}}{!k}$$



Poisson

discrete support $(1, +\inf]$

$$N(r|\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}}$$



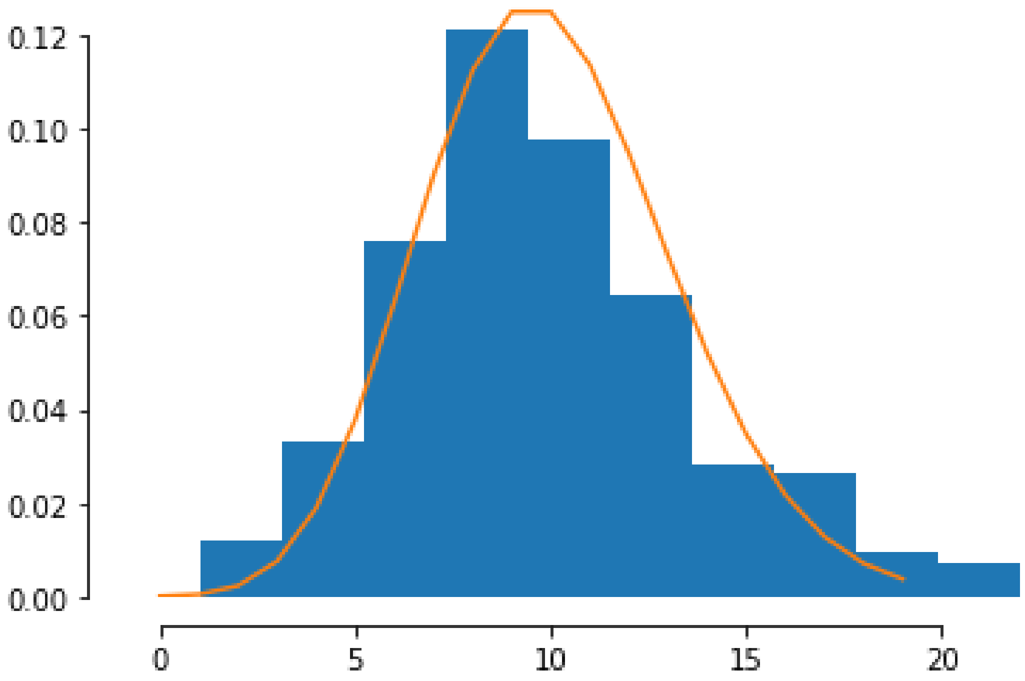
normal or Gaussian

continuous support $[-\inf, +\inf]$

distributions

parameters (*lambda*=10)

$$P(k|\lambda) \sim \frac{\lambda^k \, e^{-\lambda}}{!k}$$



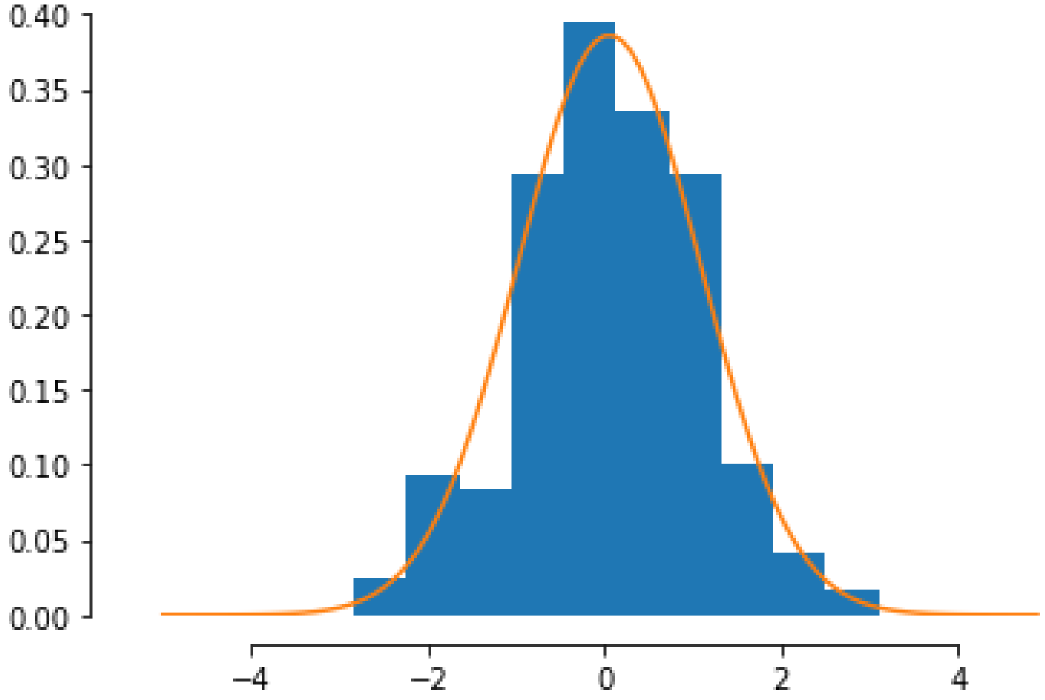
Poisson

discrete support $(1, +\inf]$

support

parameters

$$N(r|\mu, \sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}}$$



normal or Gaussian

continuous support $[-\inf, +\inf]$

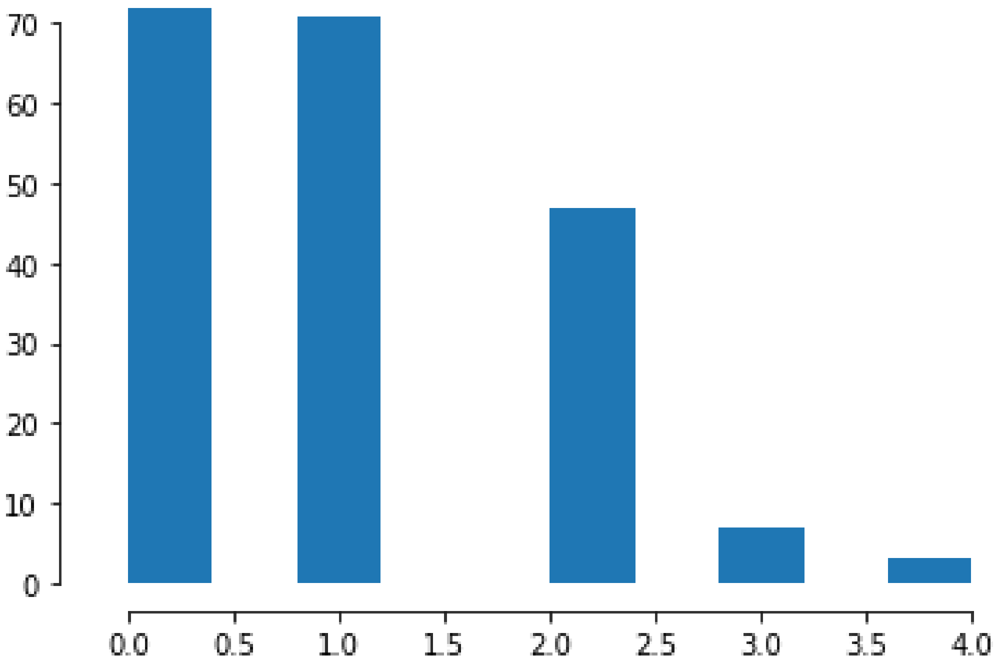
distributions

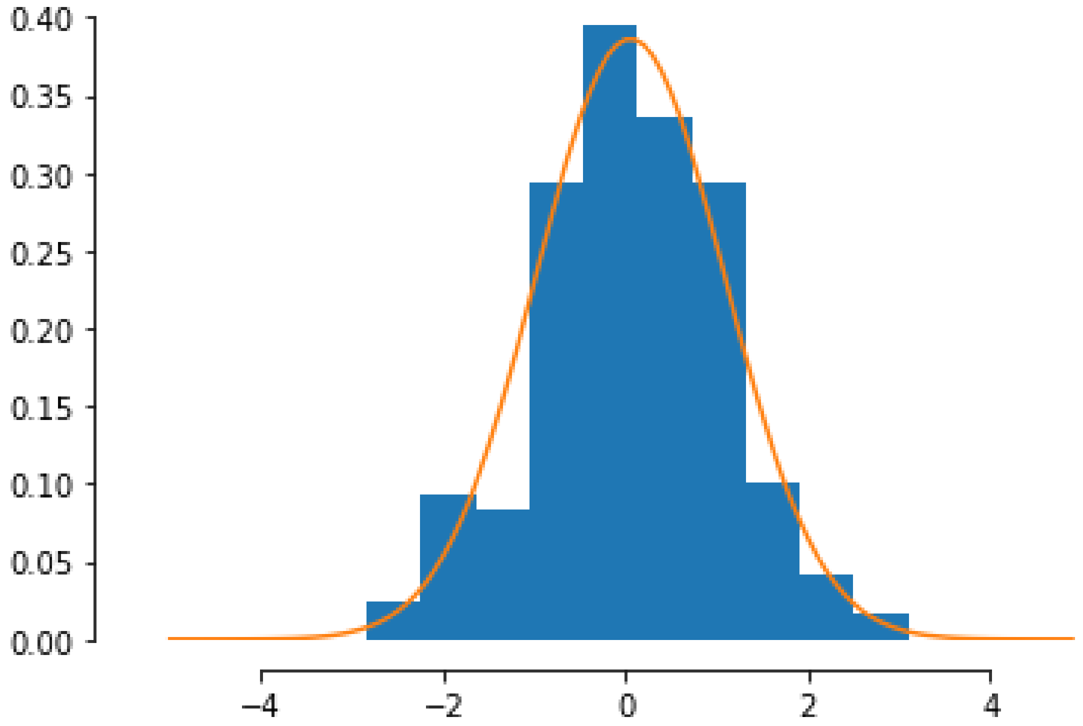
parameters (*lambda*=1)

$$P(k|\lambda) \sim \frac{\lambda^k e^{-\lambda}}{!k}$$

support parameters (-0.1, 0.9)

$$N(r|\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}}$$





Poisson

discrete support $(1, +\inf]$

normal or Gaussian

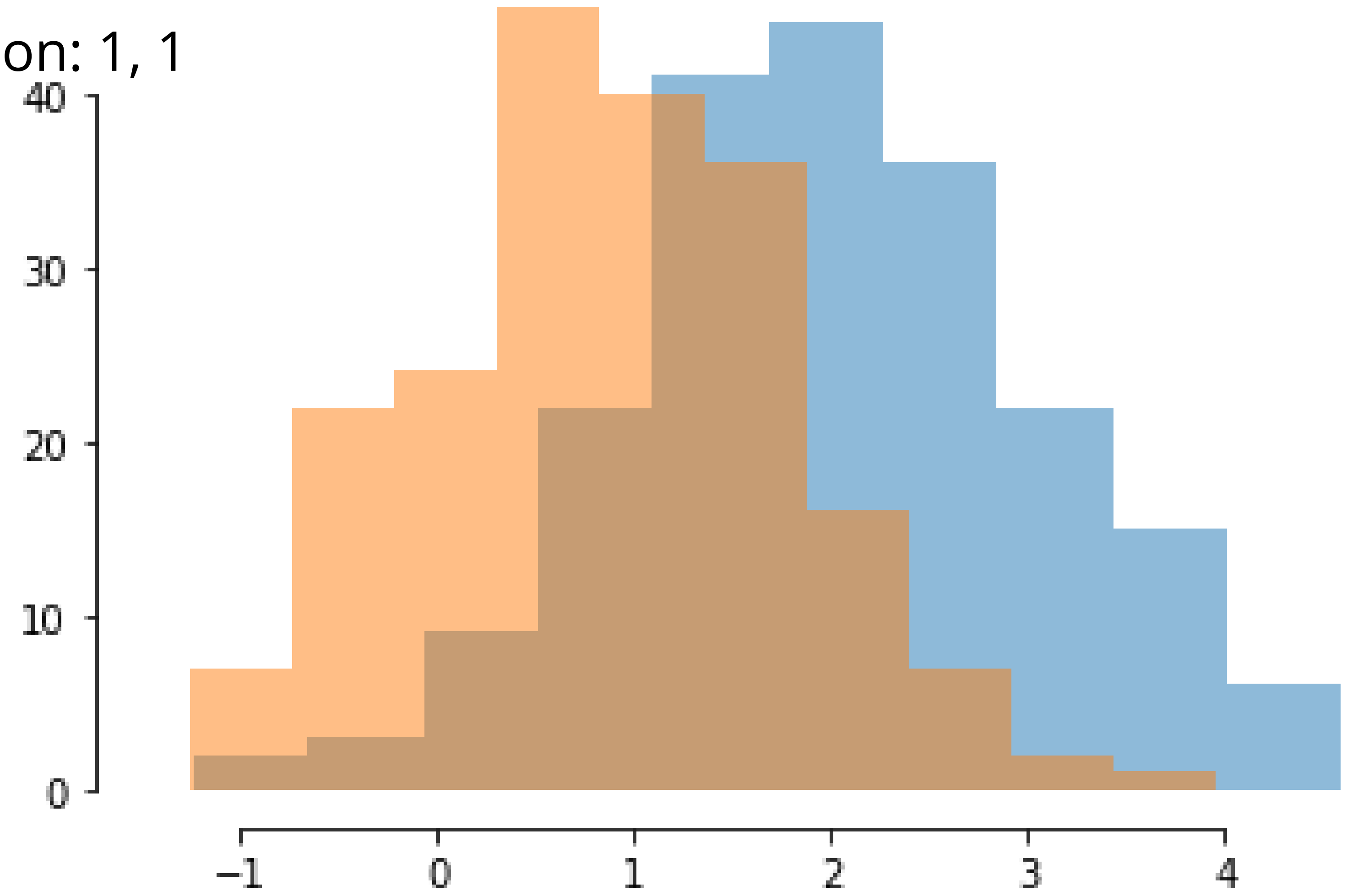
continuous support $[-\inf, +\inf]$

measuring differences between distributions

means: 1, 2

standard deviation: 1, 1

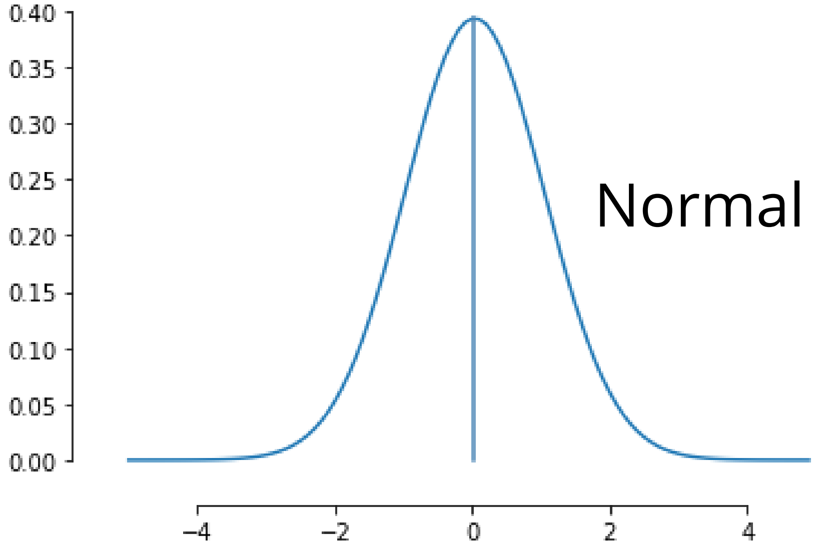
are these distributions the same?



Moments and frequentist probability

a distribution's moments summarize its properties:

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

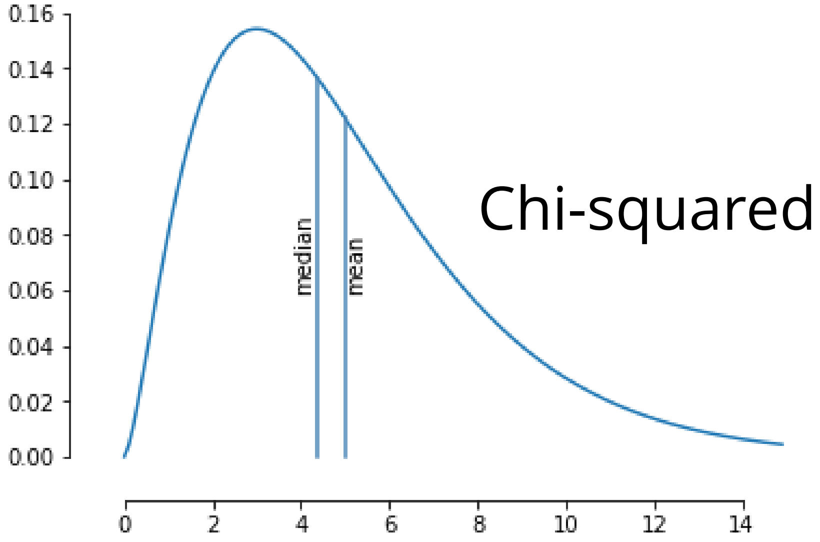


central tendency: mean (n=1), median, mode (peak)

spread: standard deviation/variance (n=2), quartiles

symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)



Moments and frequentist probability

a distribution’s moments summarize its properties:

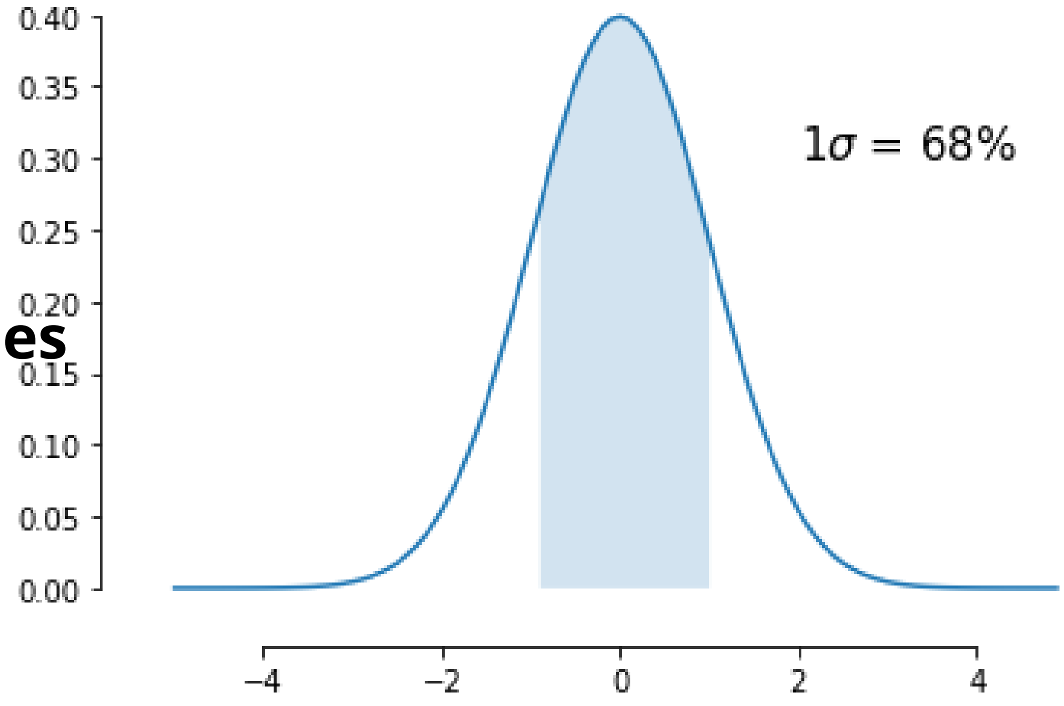
$$m_n = \int_{-\inf}^{\inf} (x-c)^n f(X)dx$$

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

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

symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)



Moments and frequentist probability

a distribution's moments summarize its properties:

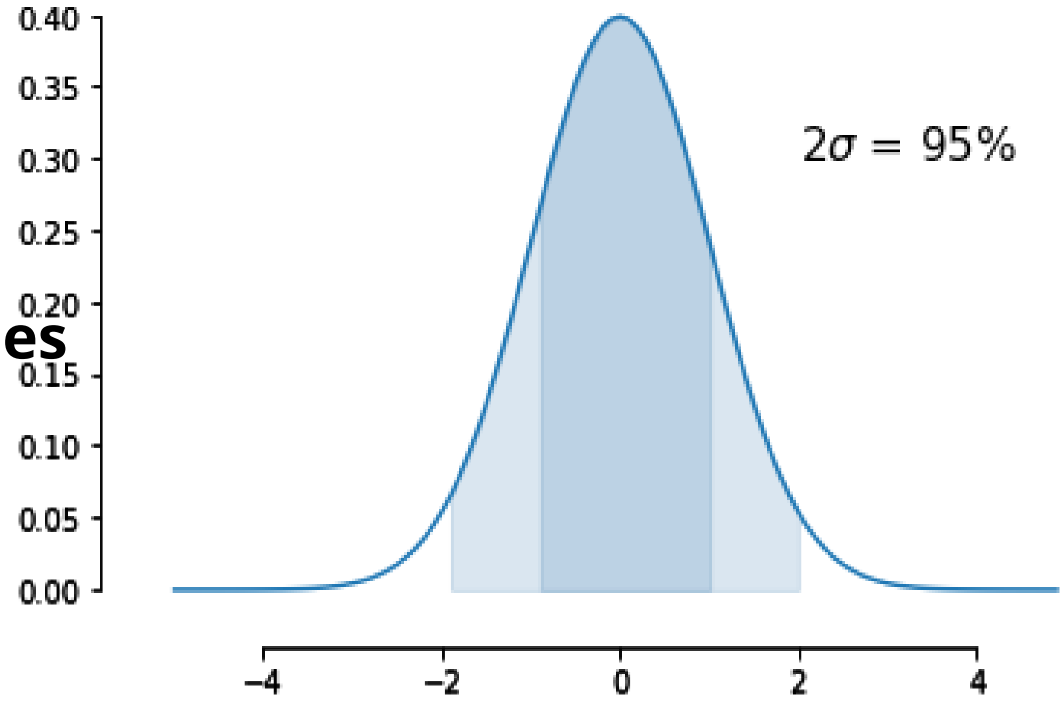
$$m_n = \int_{-\inf}^{\inf} (x-c)^n f(X) dx$$

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

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

symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)



Moments and frequentist probability

a distribution's moments summarize its properties:

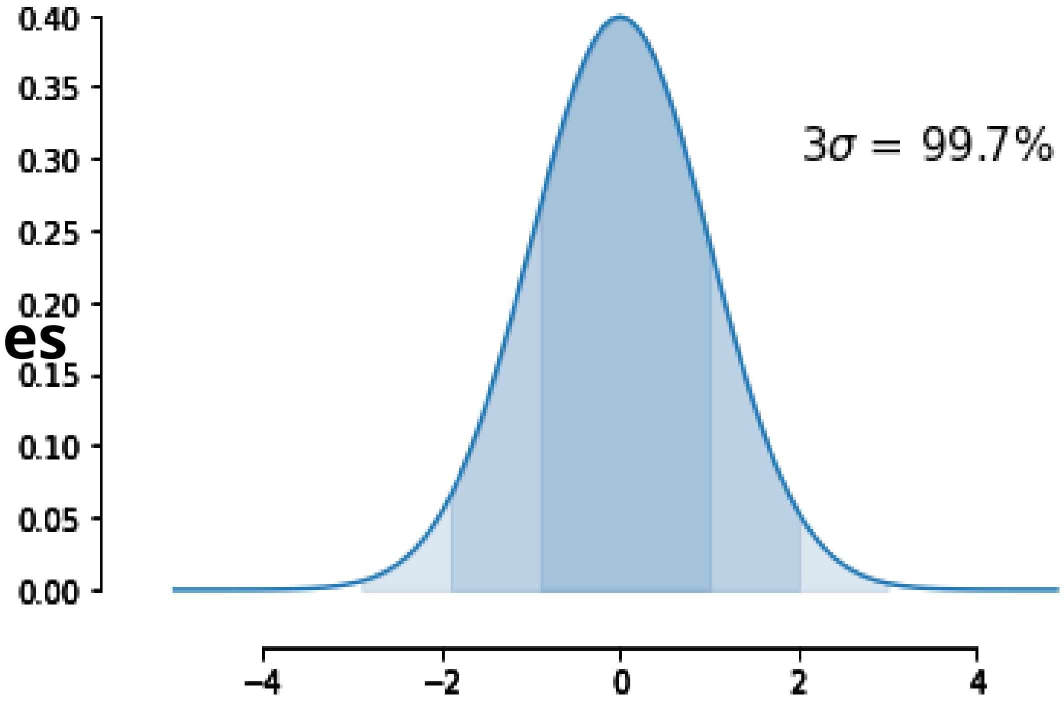
$$m_n = \int_{-\inf}^{\inf} (x - c)^n f(X) dx$$

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

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

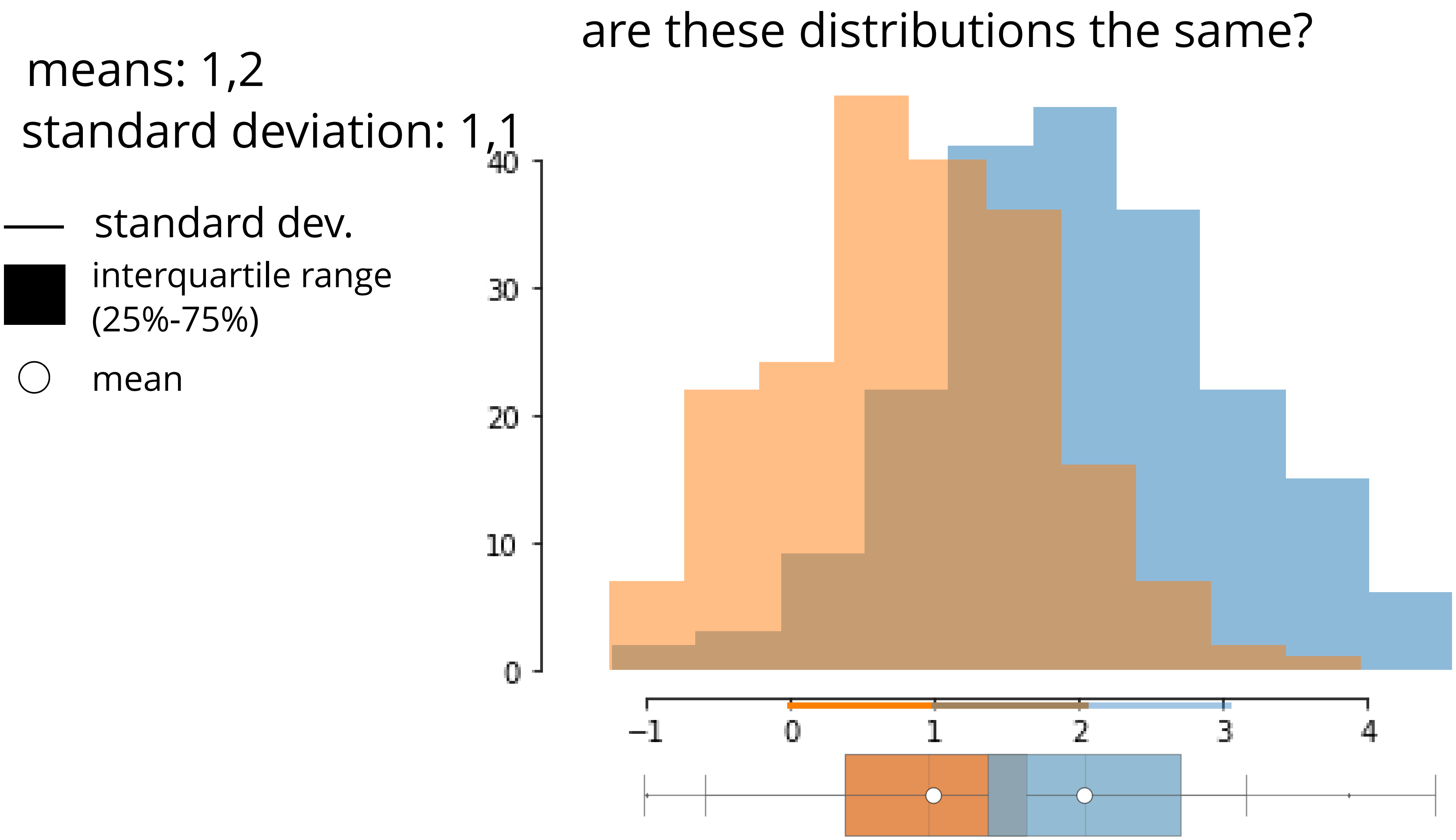
symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)



measuring differences between distributions

means: 1,2
standard deviation: 1,1



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

measuring differences between distributions

a distribution’s moments summarize its properties:

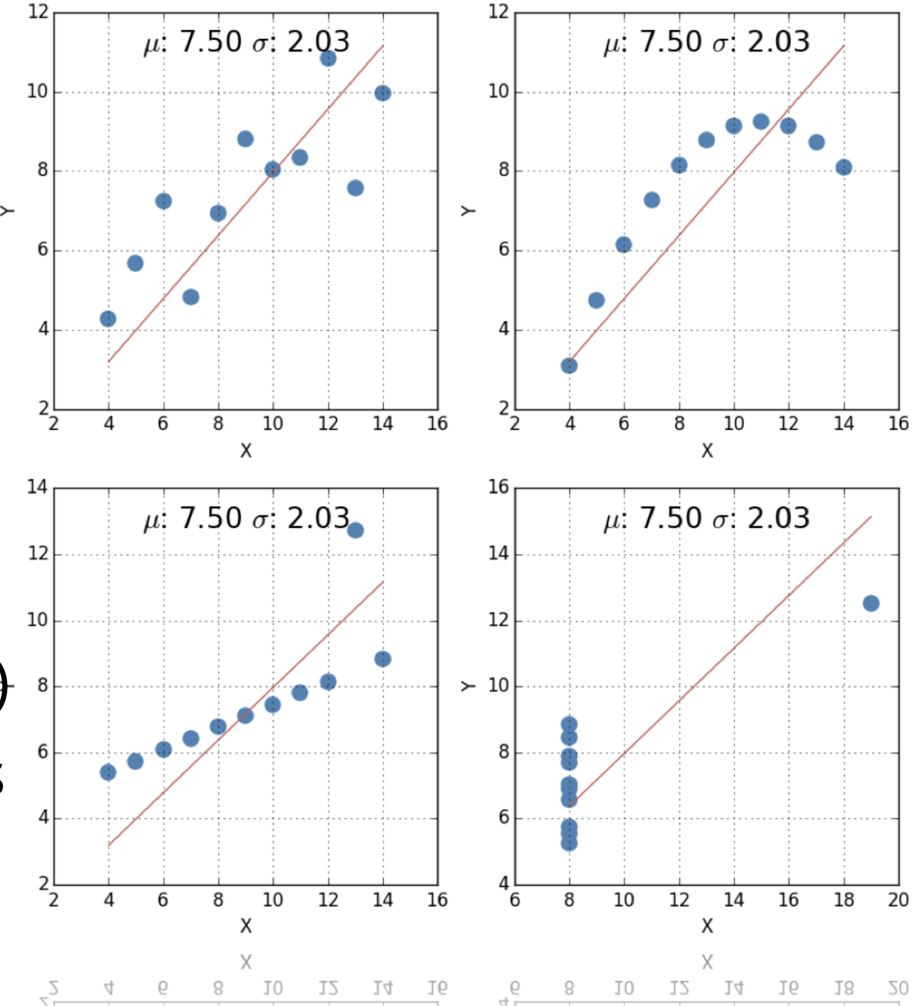
$$m_n = \int_{-\inf}^{\inf} (x-c)^n f(X) dx$$

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

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

symmetry: skewness (n=3)

cuspiness: kurtosis (n=4)



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

2

p-value hypothesis testing

Preamble: kinds of analytical questions

- Are two measurements the same?
  - *is the amount of nitrates in Lums pond same as it was 2 years ago?*
- Are two distributions the same?
  - *is the weight of Medicare members signed up for health newsletters the same as that of members who are not signed up?*
- Can I trust that a number comes from a certain distribution? -> ***p***-value

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

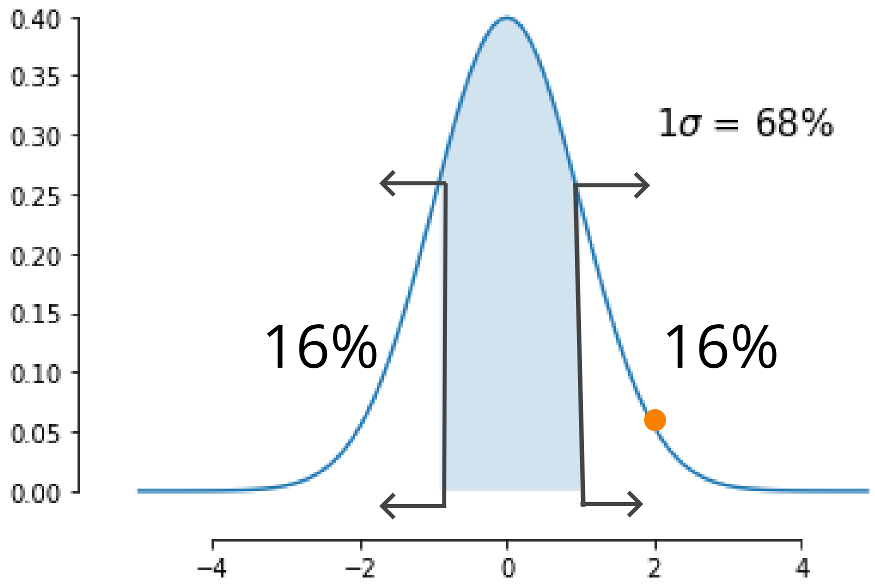
what is the probability that I would measure 1.5?

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

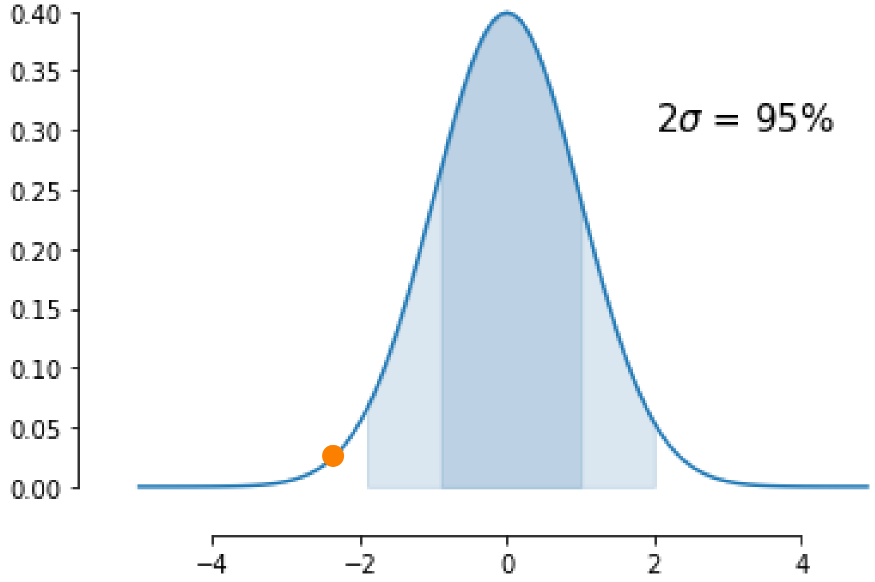
what is the probability that I would measure 1.5?

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 [-2*σ*:2*σ*] 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

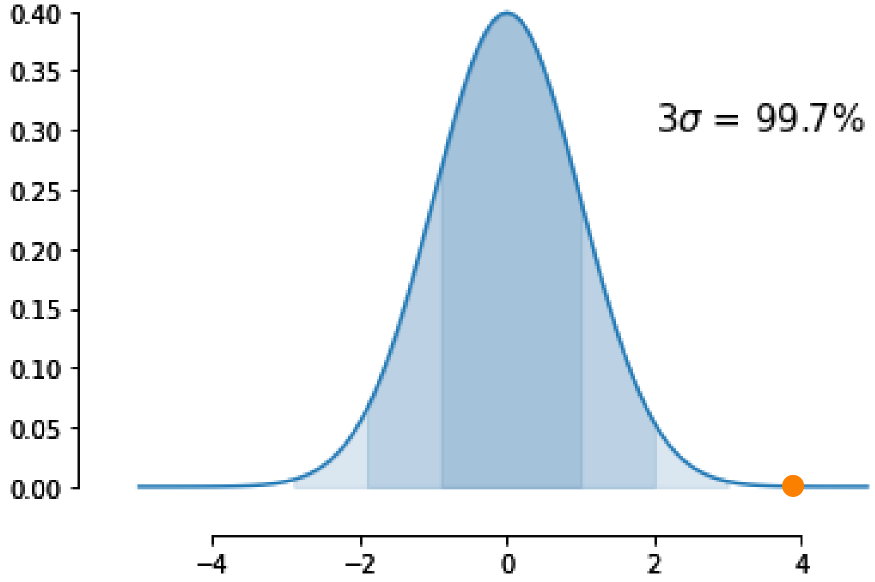
what is the probability that I would measure 1.5?

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 [-3*σ*:3*σ*] 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

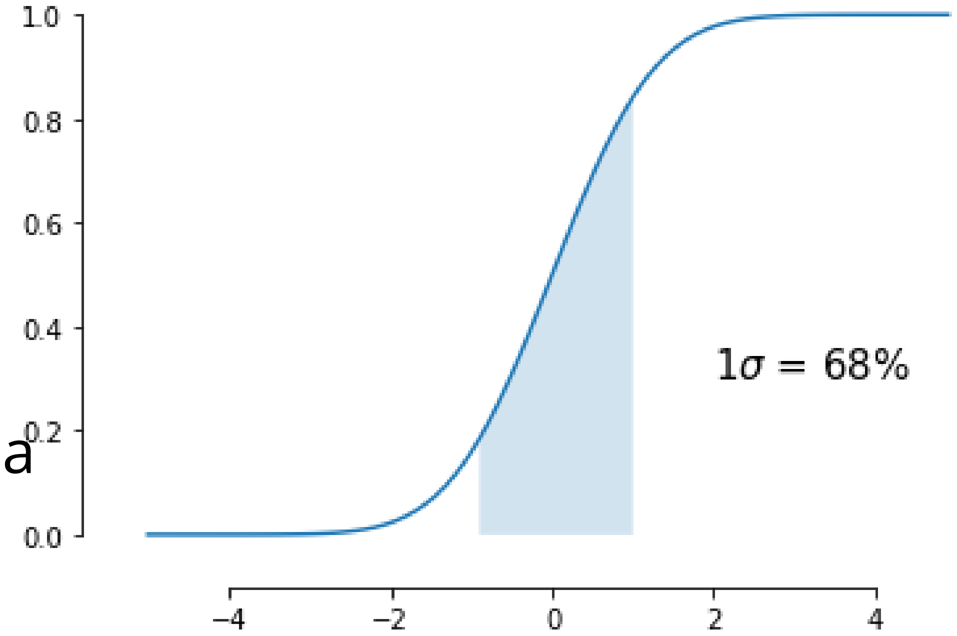
what is the probability that I would measure 1.5?

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%.

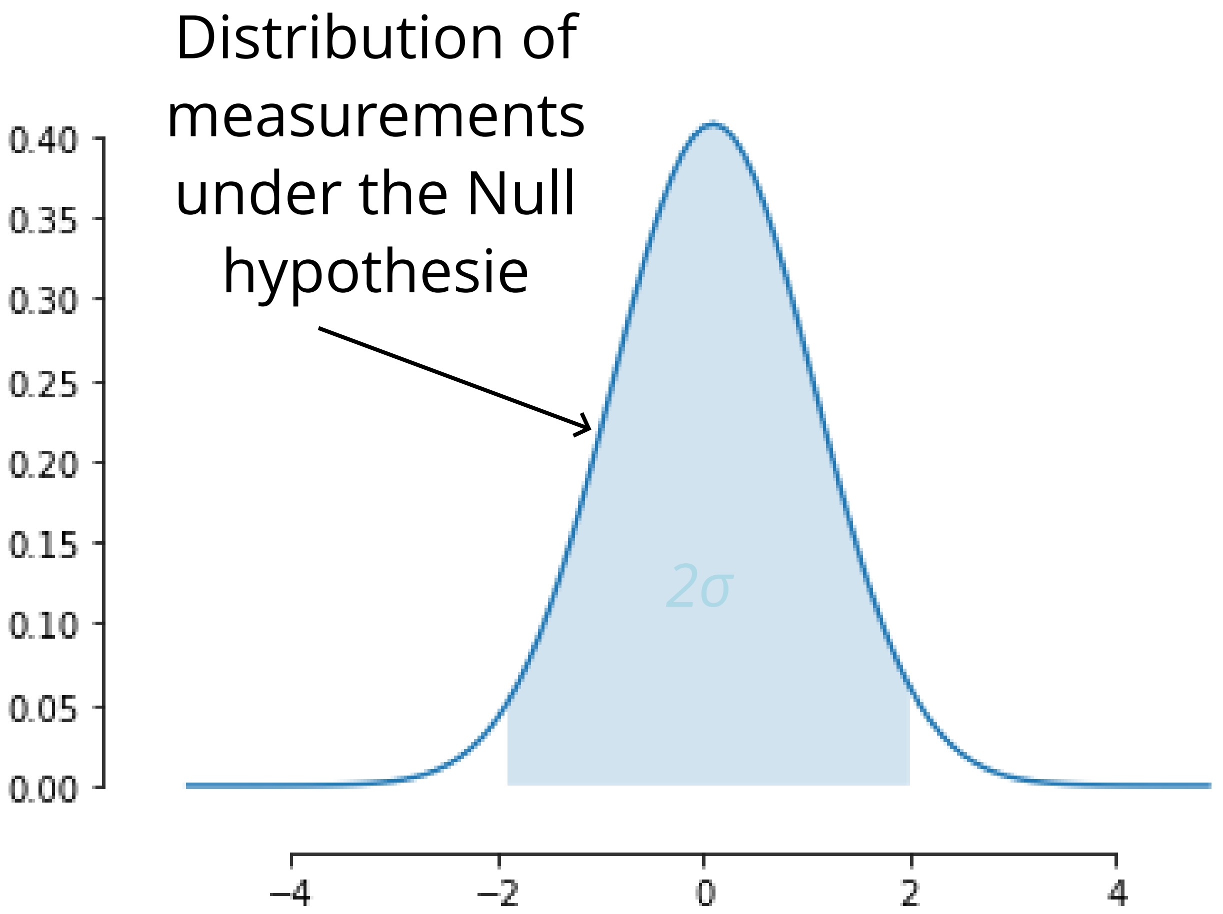
So the probability of measuring something outside is 100-99.7 = 0.3%.

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



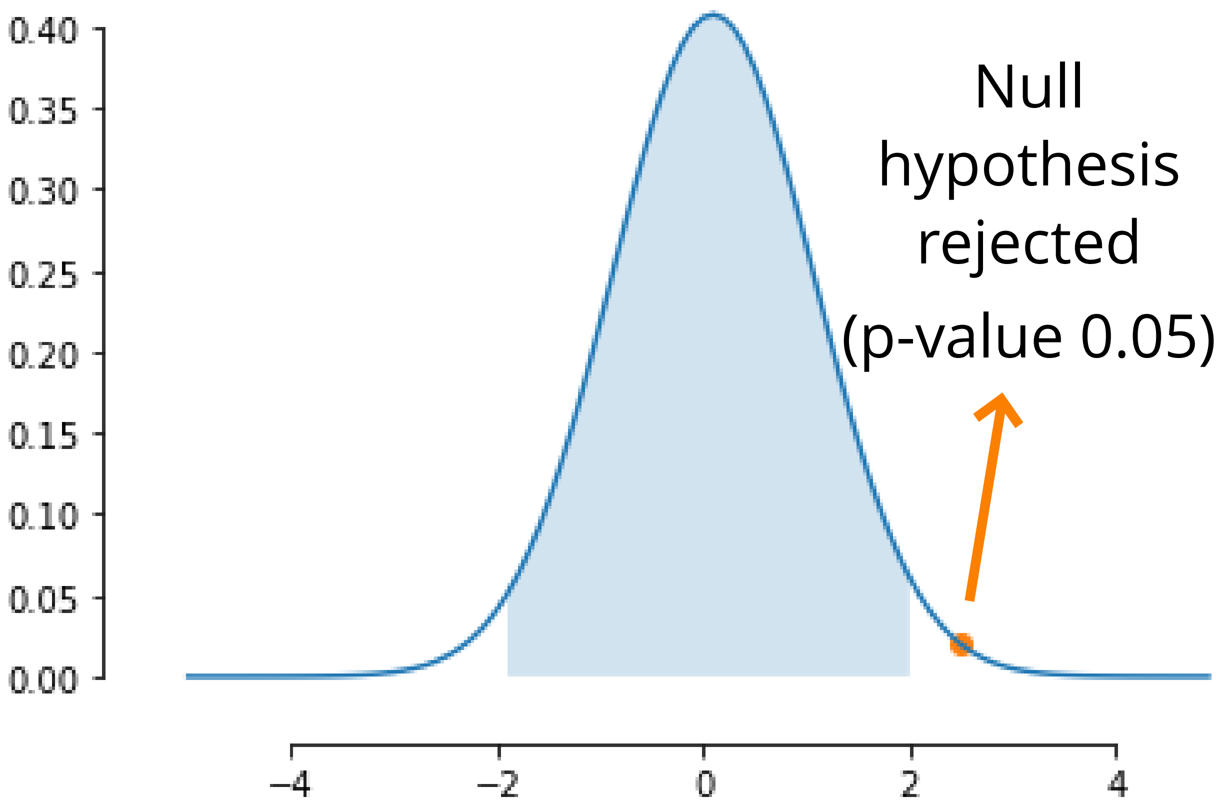
Moments and frequentist probability in the *falsification* framework: *p*-value

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.



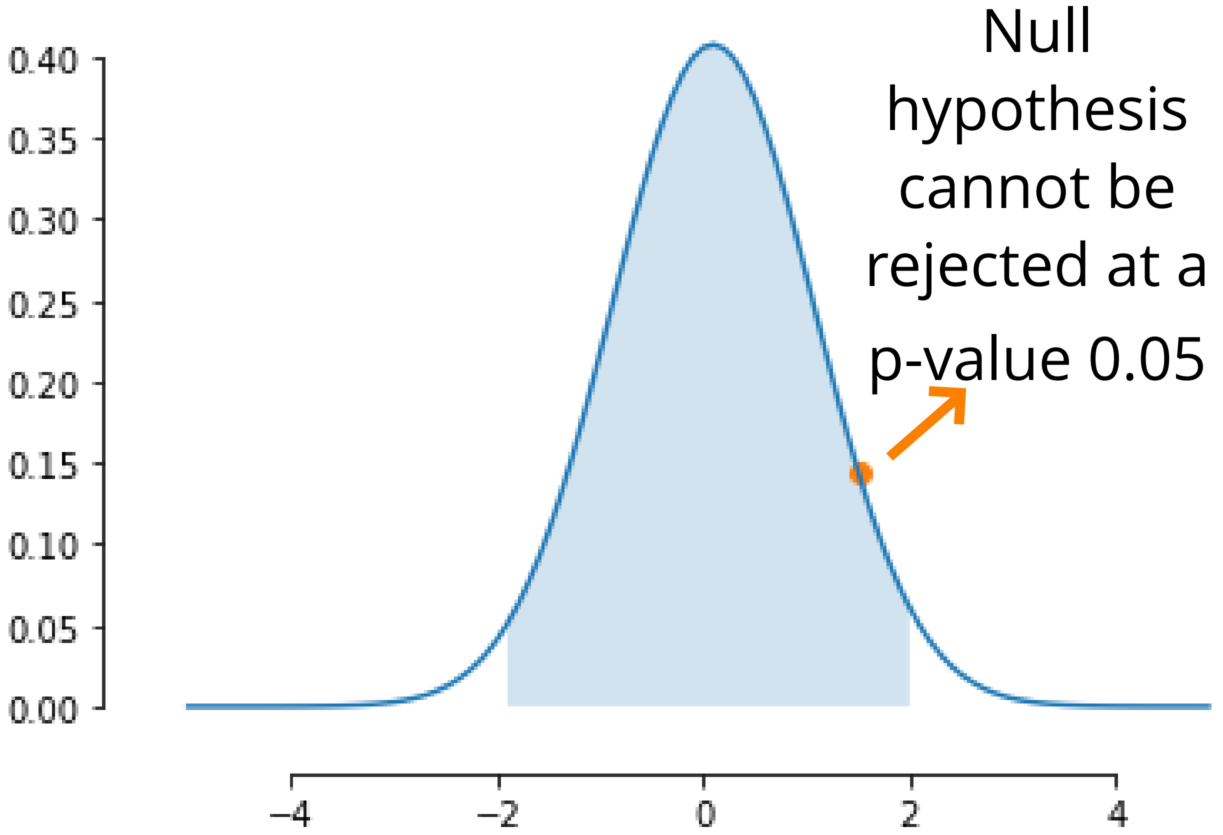
Moments and frequentist probability in the *falsification* framework: *p*-value

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.



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.



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**

set up threshold α → identify how you expect your measurement to be distributed *under the Null hypothesis*

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...)

set up threshold α → identify how you expect your measurement to be distributed *under the Null hypothesis* → measure outcome from data: x_0

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.

set up threshold α → identify how you expect your measurement to be distributed *under the Null hypothesis* → measure outcome from data: x_0

$$p(|x| > |x_0|) < \alpha \implies \boxed{\textbf{falsify } \textbf{\textit{H0}}}$$

$$p(|x| > |x_0|) \geq \alpha \implies \boxed{\textbf{\textit{H0 cannot be falsified}}}$$

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.

this quantity is called a "statistics"

set up threshold α → identify how you expect your measurement to be distributed *under the Null hypothesis* → measure outcome from data: x_0

$$p(|x| > |x_0|) < \alpha \implies$$ **falsify *H0***

$$p(|x| > |x_0|) \geq \alpha \implies$$ ***H0 cannot be falsified***

2.1

statistics

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

$Z \sim N(0,1)$

Does a sample come from a known population? Z -test

Example: new bus route implementation.

https://github.com/fedhere/PUS2020_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.

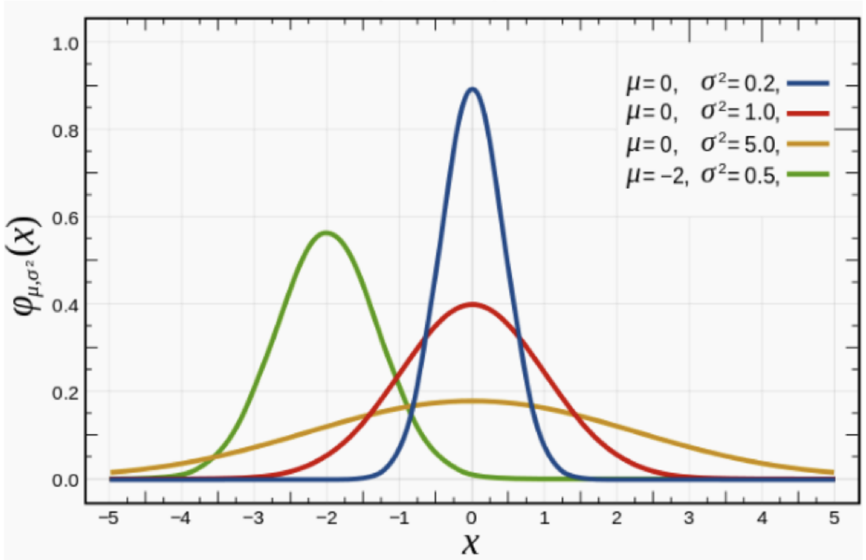
Has travel time changed?

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

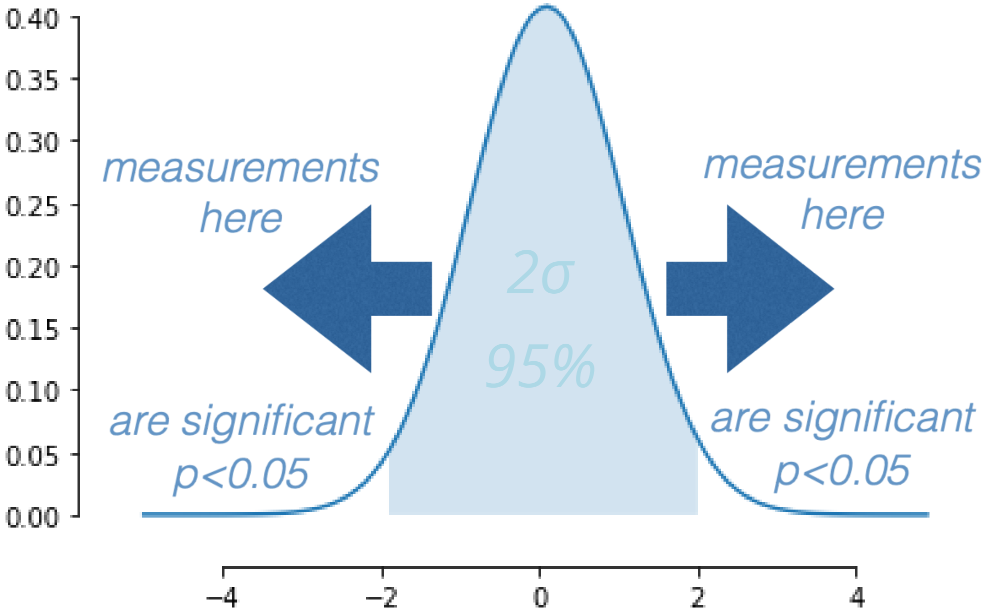
In absence of effect (i.e. under the Null)

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

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



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	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$



Are 2 proportions (fractions) the same? Z -test

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

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

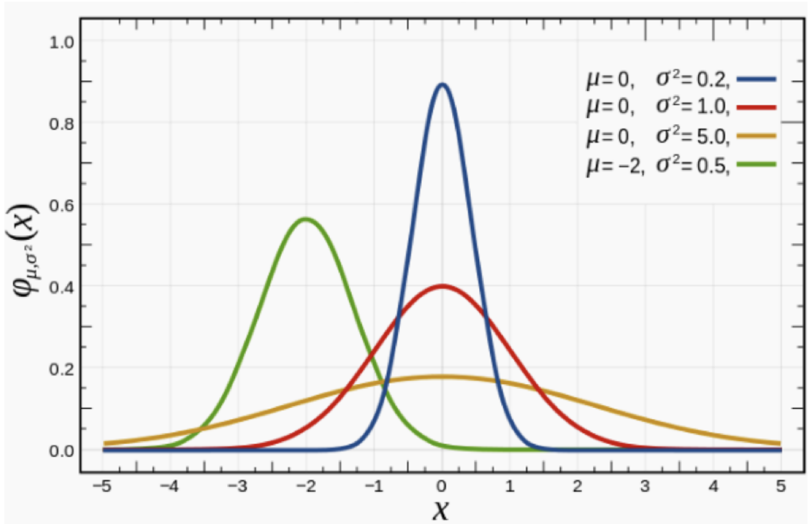
$$SE = \sqrt{p(1-p)\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}$$

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

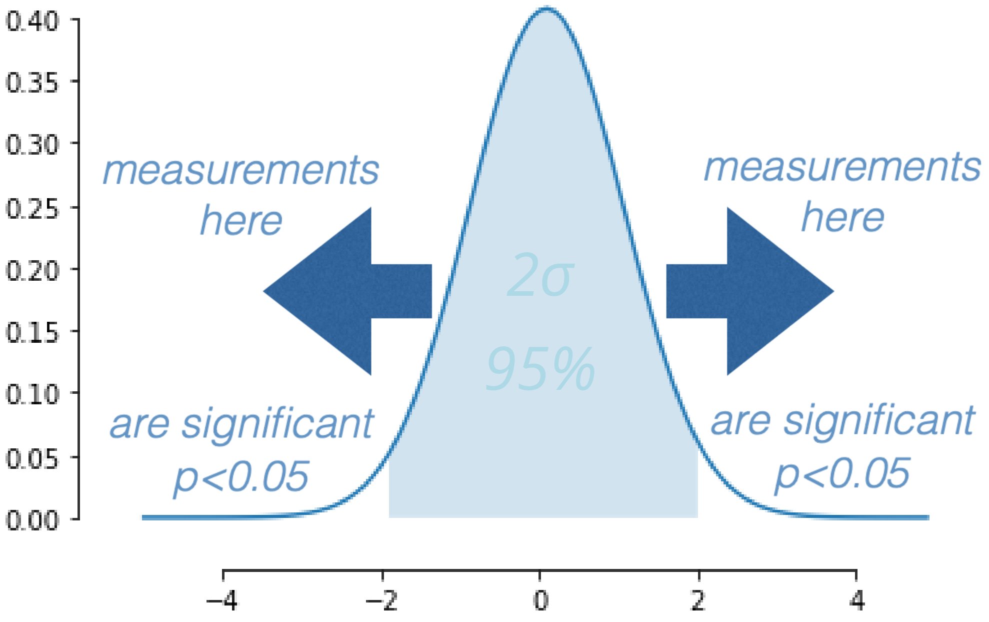
In absence of effect (i.e. under the Null)

== the proportions of men and women are the same

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



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	$\frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$



Are 2 proportions (fractions) the same? Z -test

Example: citibike women usage patterns

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

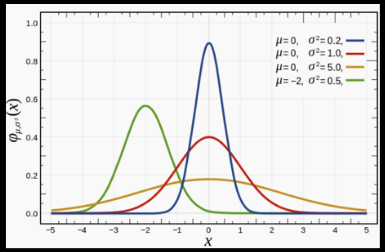
You want to know if women are less likely than man to use citibike to commute.

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

Statistics and tests

Z statistics Gaussian

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

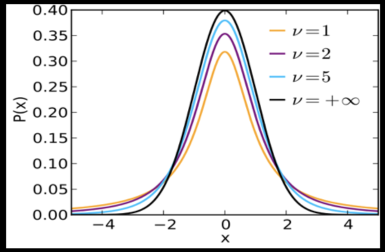


Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

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

Student's *t*

$$t = \frac{\mu - \overline{x}}{s/\sqrt{n}}$$

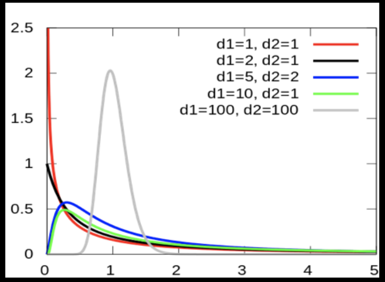


Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu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

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

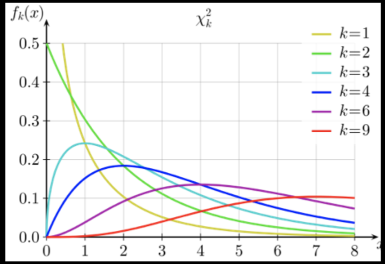


Parameters	$d_1, d_2 > 0$ deg. of freedom
Support	$x \in [0, +\infty)$
PDF	$\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$
CDF	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$

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

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



Notation	$\chi^2(k)$ or χ_k^2
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 $\chi_F^2 = \sum_i \frac{(m_i - x_i)^2}{e_i}$

see
Statistics in a Nutshell

3

data kinds and nomenclature

Types of Data:

Data Definitions

Data: observations that have been collected

Population: the complete body of subjects we want to infer about

Sample: the subset of the population about which data is collected/available

Census: collection of data from the *entire population*

Parameter: the subset of the population we actually studied collection of data from the entire population

Statistics: numerical value describing an attribute of the *population* numerical value describing an attribute of the *sample*

Data Definitions

The analysis of our ______

showed that for our 10 ______ the mean income is $60k.

The standard deviation of the ______ means is $12k.

From this ____ we infer for the ________ a mean income

__________ $60k +/- $12k

data

sample

statistics

population

parameter

At the root is the fact that a sample drawn from a parent distribution will look increasingly more like the parent distribution as the size of the sample increases.

More formally: The distribution of the means of N samples generated from the same parent distribution will

I. be normally distributed (i.e. will be a Gaussian)

II. have *mean* equal to the *mean of the parent distribution*, and

III. have *standard deviation* equal to the *parent population standard deviation divided by the square root of the sample size*

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

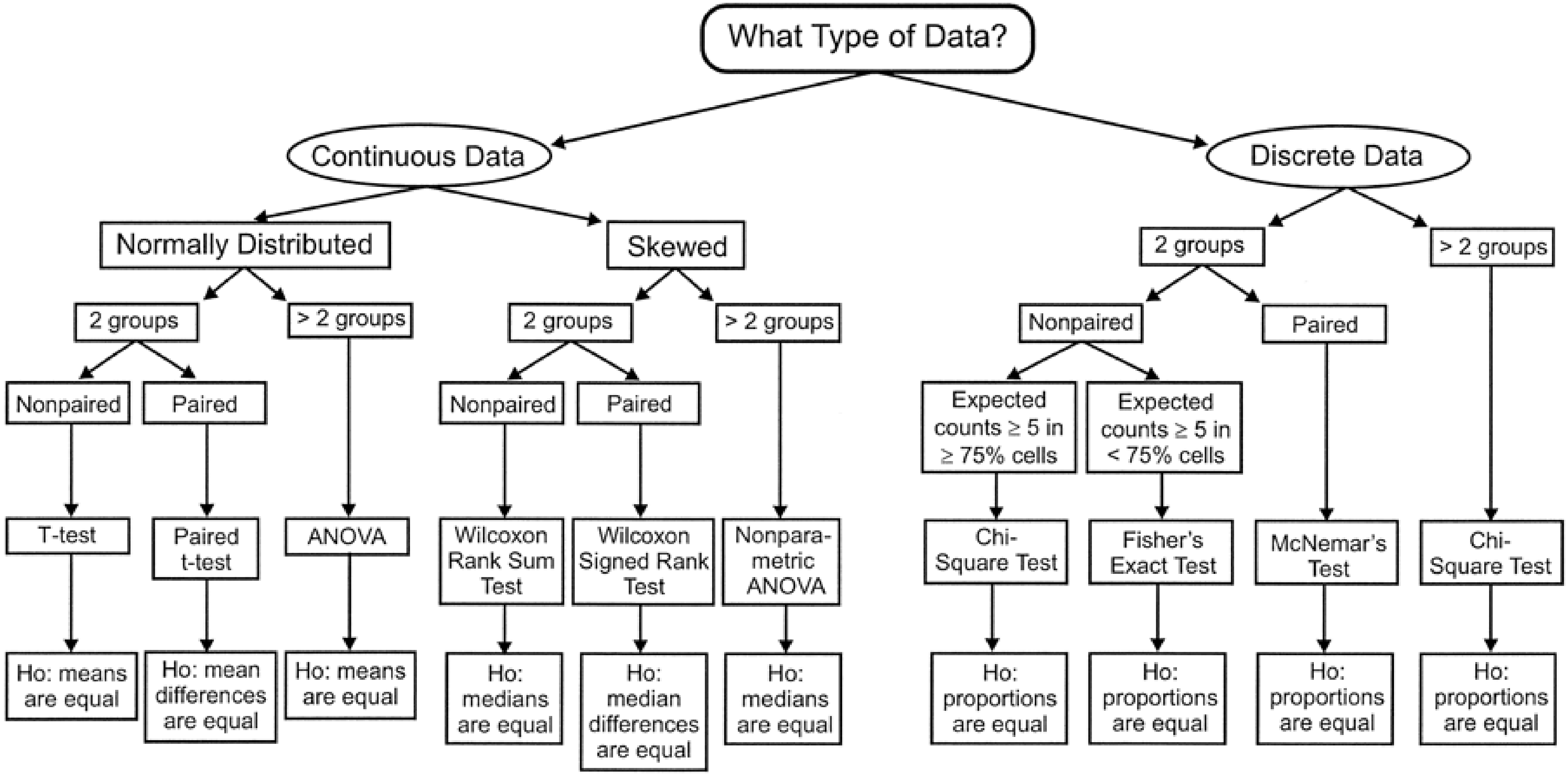
Ratio:
Car speed 0 is naturally defined

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

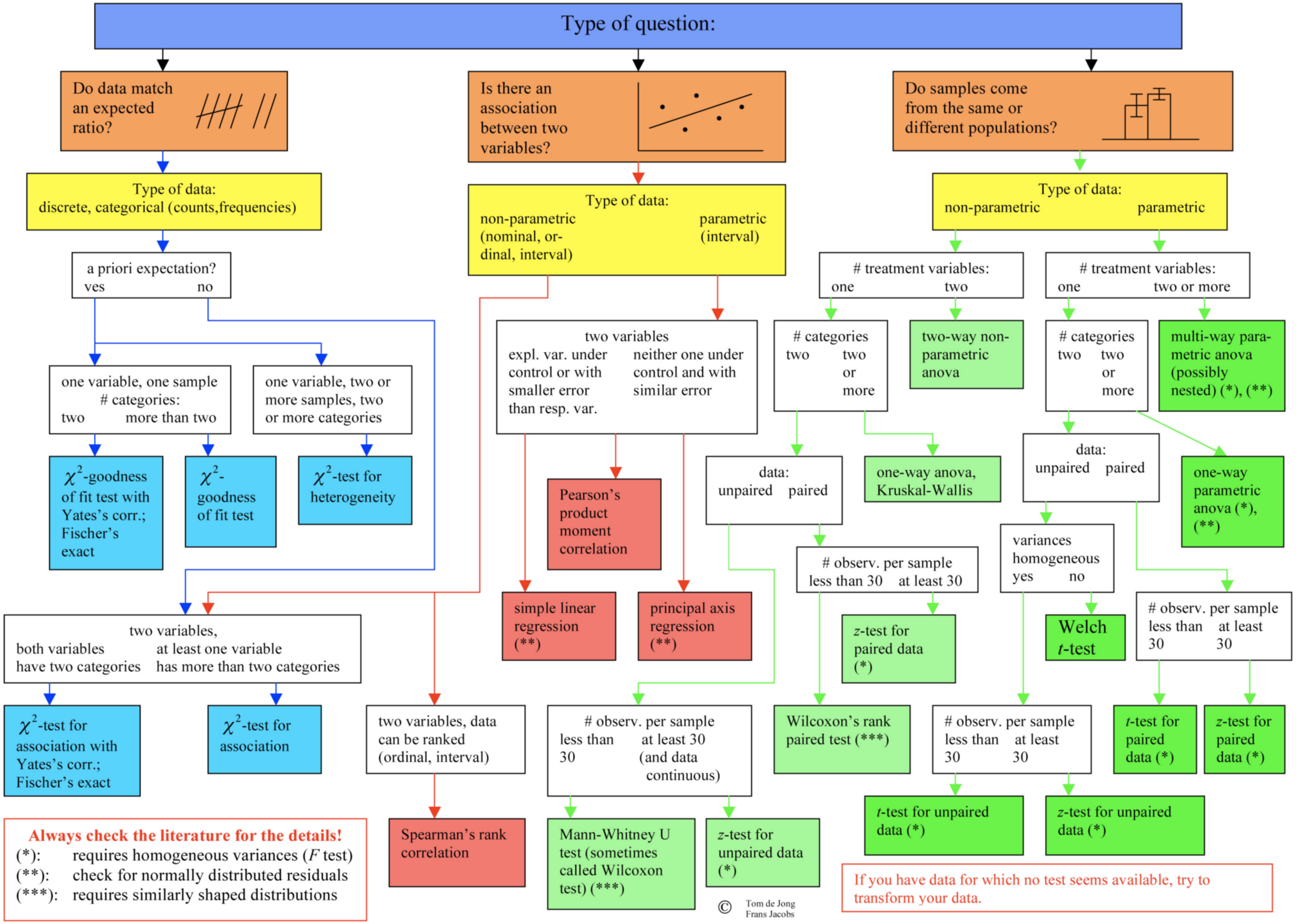
Censored: age>90

4

which is the right test for me?



Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com



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

key concepts

Distributions: frequency and probability interpretations

Descriptive statistics: mean, median, standard deviation, interquartile range

NHRT Null Hypothesis Rejection Testing and *p*-values

Definition: Types of data

Definition: Parameters, features, variables

Statistical tests:

how to use it (statistics value compared to the distribution under the null)

how to choose it : what kind of data? what kind of question?

key concepts

From idea to hypothesis

idea → measurable quantity → statistics → *H0* Null hypothesis / *Ha* Alternative hypothesis → **falsify *H0***

set up threshold α → identify how you expect your measurement to be distributed *under the Null hypothesis* → measure outcome from data: x_0

$$p(|x| > |x_0|) < \alpha \implies$$ **falsify *H0***

$$p(|x| > |x_0|) \geq \alpha \implies$$ ***H0 cannot be falsified***

key concepts

NHRT setup:

set up threshold α $\longrightarrow$ identify how you expect your measurement to be distributed *under the Null hypothesis* $\longrightarrow$ measure outcome from data: *x_0*

$$p(|x| > |x_0|) < \alpha \implies$$ **falsify *H0***

$$p(|x| > |x_0|) \geq \alpha \implies$$ ***H0 cannot be falsified***

https://pdfs.semanticscholar.org/fa63/cbf9b514a9bc4991a0ef48542b689e2fa08d.pdf

The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p *as the probability that* H_0 *is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects* H_0 *one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.*

reading

optional follow up:

The Earth is Flat (p<0.05): significance thresholds and the crisis of unreplicable research

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5502092/

Homework

follow the notebook, create your own question about citibike, write the idea, the null and alternative hypothesis and the relative formulate. The question should be measurable by a test of proportions. Follow the example of the gender and usage of bikes for commuting to perform the tests and interpret the results.

Measure the effect size and follow the wikipedia entry to evaluate if the effect size is large or small based on Cohen's criterion

homework

https://github.com/fedhere/PUS2020_FBianco/tree/master/HW3