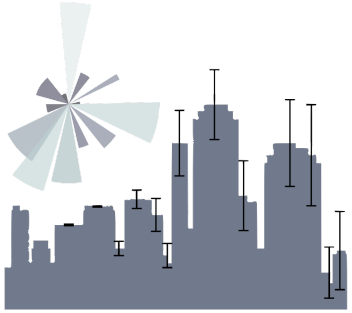
principles of Urban Science 4



choosing NHRT tests - geopandas

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

1. ~~Reading in data~~
2. ~~Descriptive statistics (central tendency, spread...)~~
3. ~~Extracting descriptive statistics from data~~
4. ~~p-value inference~~
5. Choosing a statistical test

mapping in python (intro to geopandas)

fitting lines to data

fitting and overfitting

this slide deck: https://slides.com/federicabianco/pus2020_4

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.

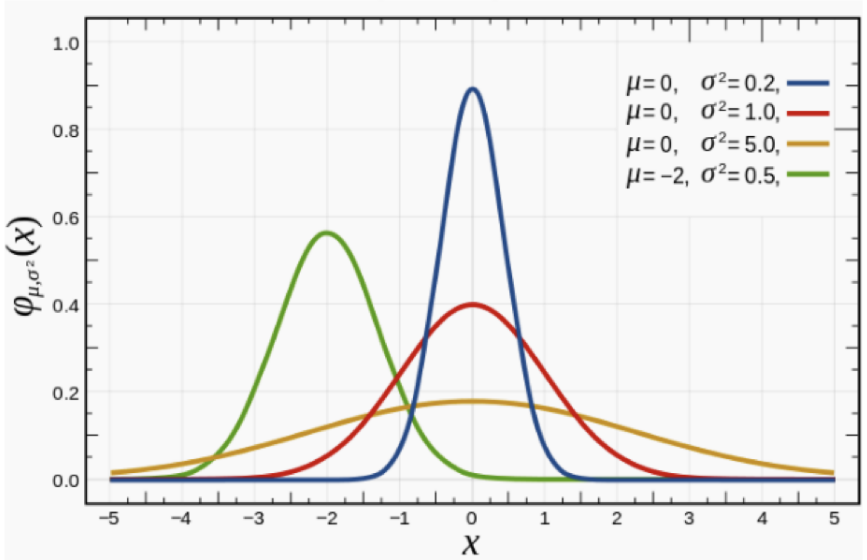
Has travel time changed?

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

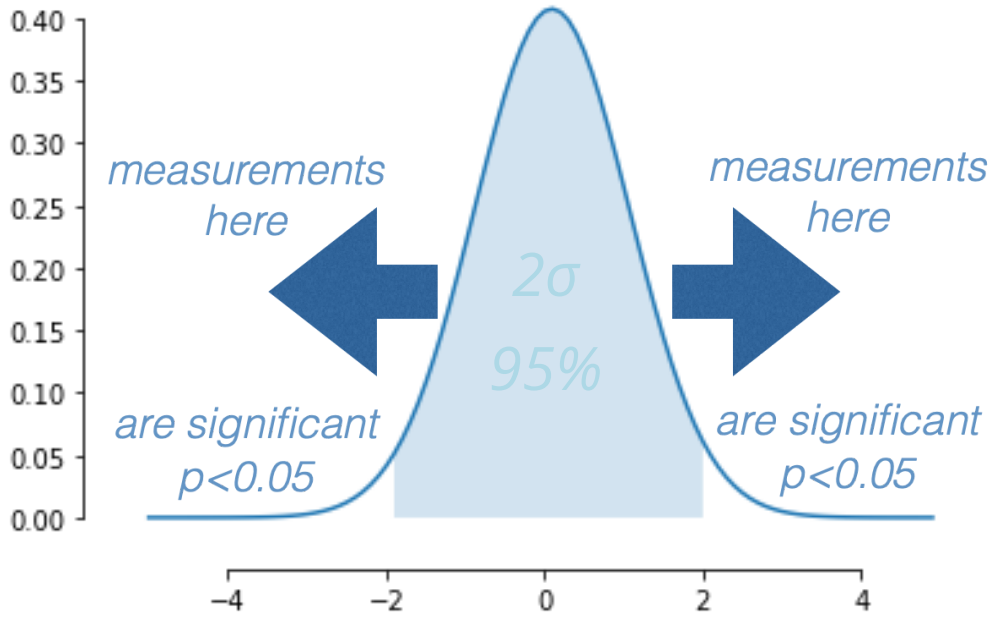
In absence of effect (i.e. under the Null)

== 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

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

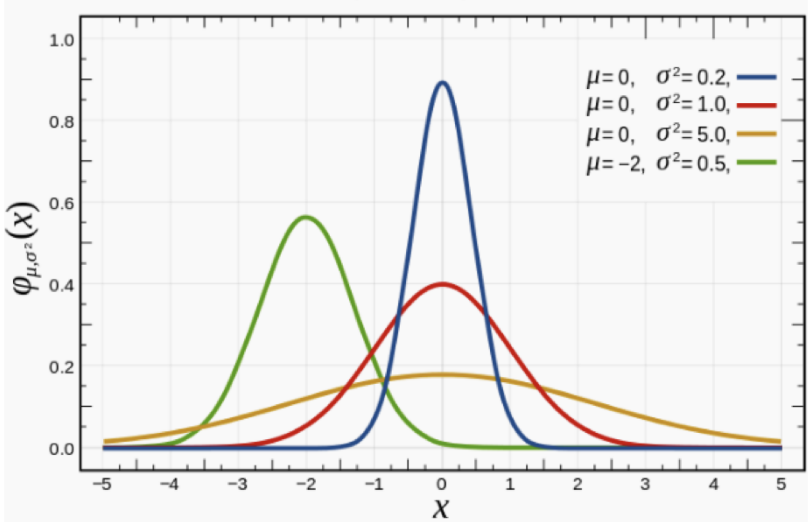
$$SE = \sqrt{p(1-p)\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}$$

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

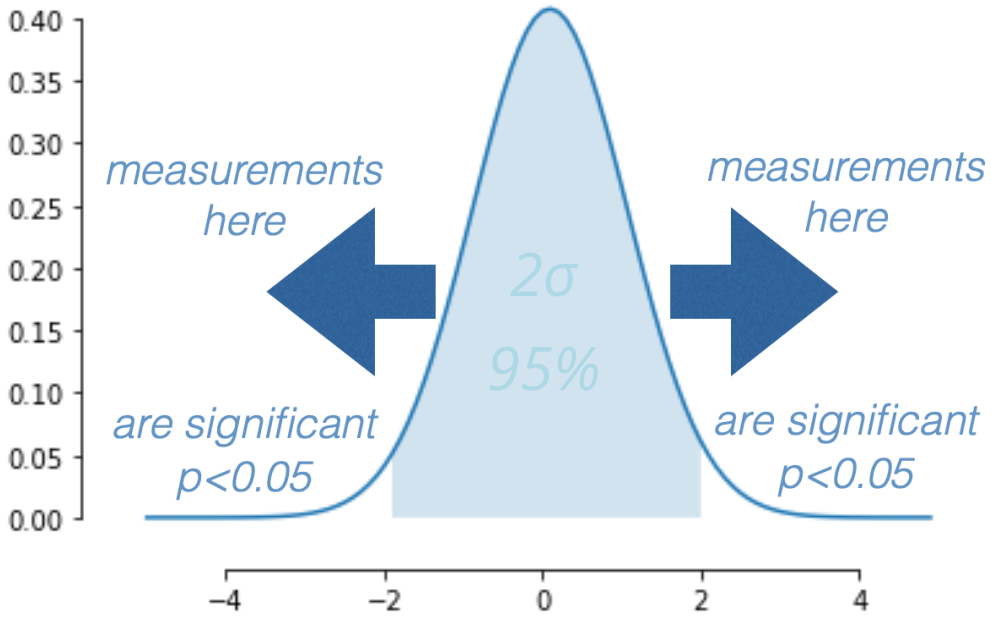
In absence of effect (i.e. under the Null)

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



Statistics and tests

Z statistics Gaussian

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

Student's *t*

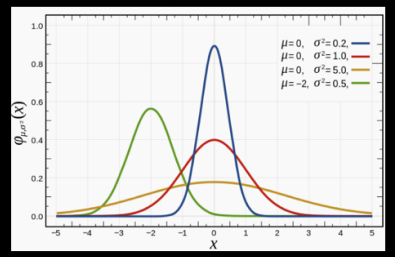
$$t = \frac{\mu - \overline{x}}{s/\sqrt{n}}$$

F statistics

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

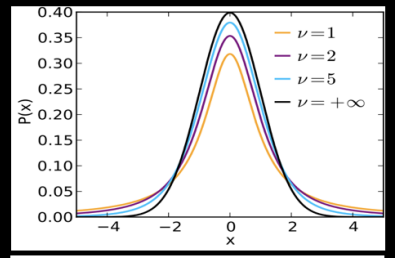
Pearson's χ^2

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



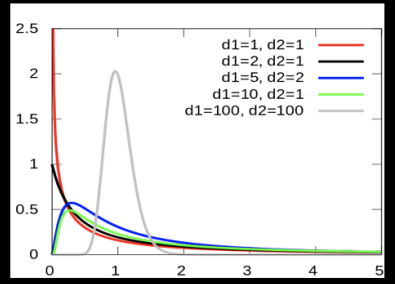
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

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



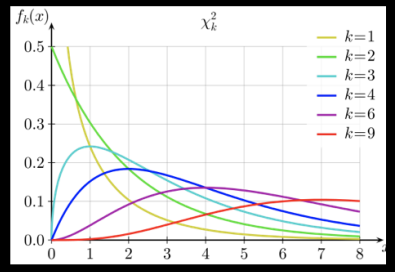
Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},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



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	$\frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)}$ 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 $d_2 > 6$



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	$\sqrt{8/k}$

see
Statistics in a Nutshell

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

3

data types 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 these ____ we infer the _____ has 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:

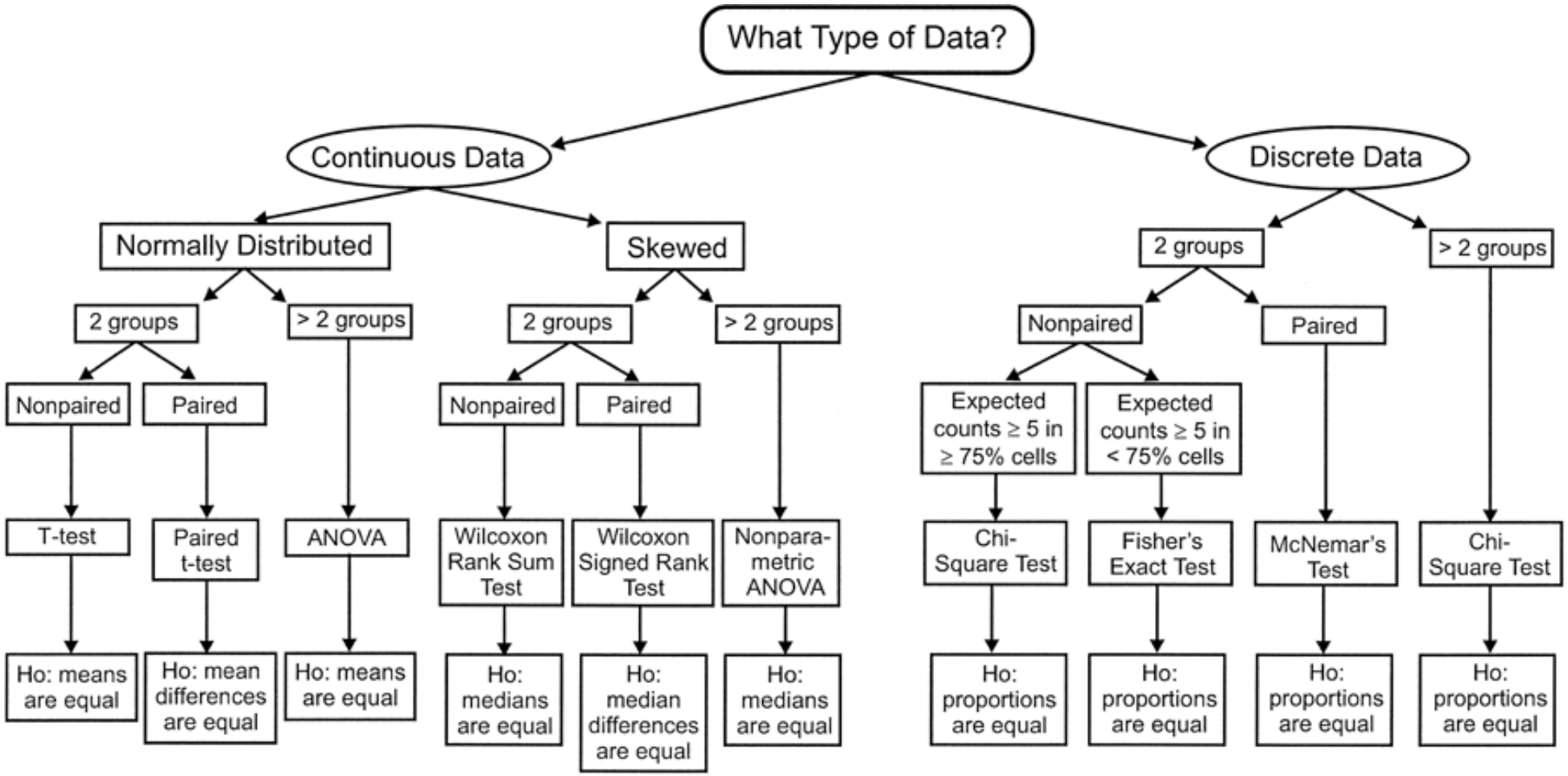
Car speed 0 is naturally defined

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

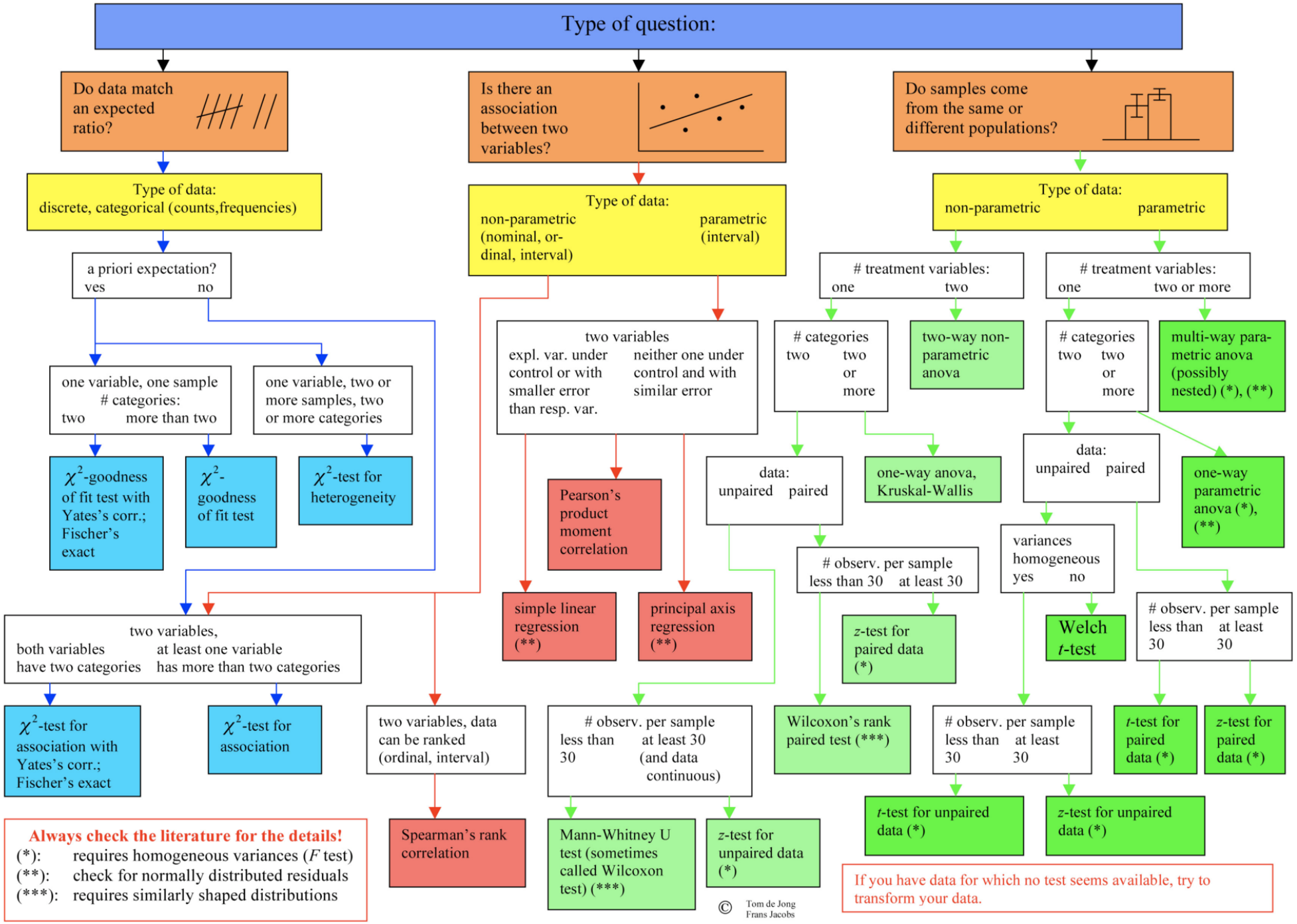
Censored: age>90

4

which is the right test for me?



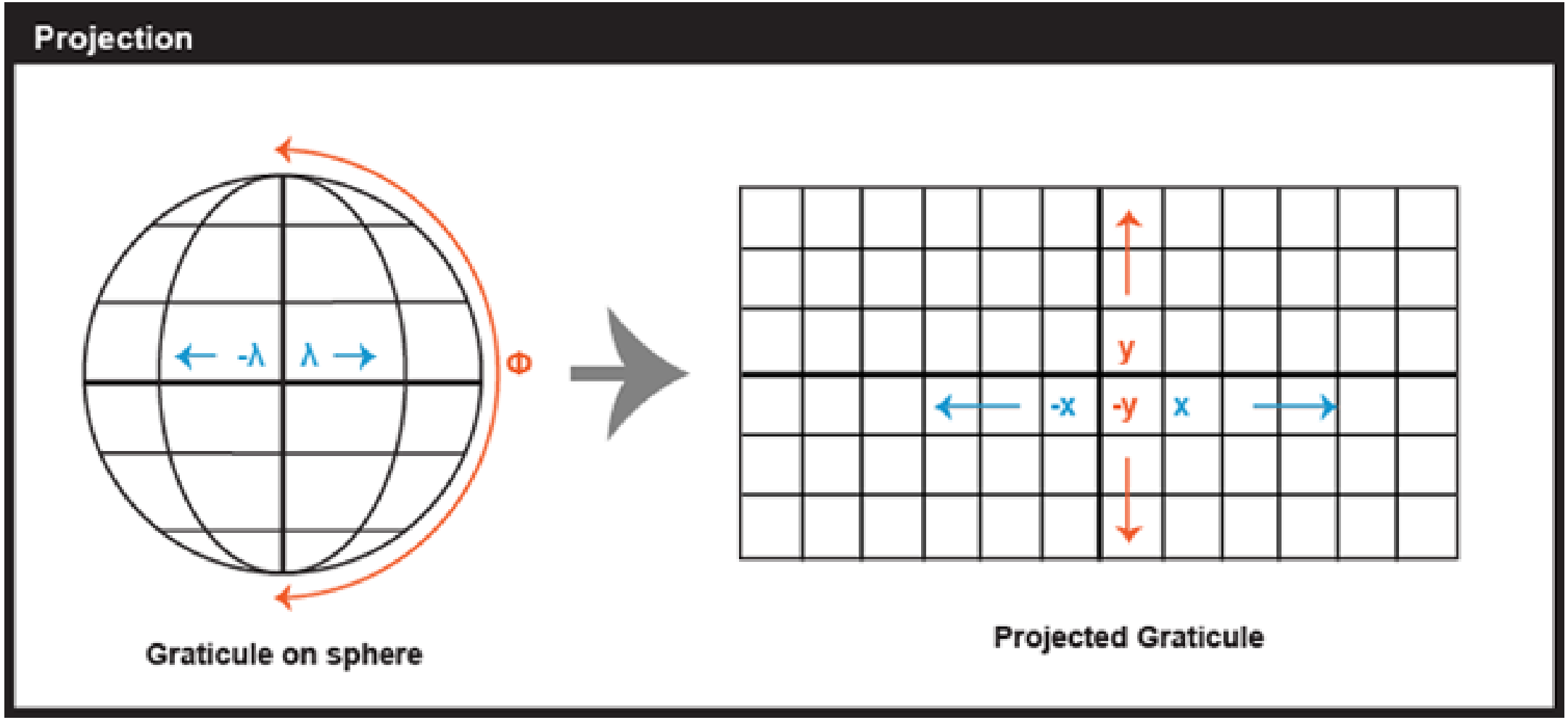
Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com

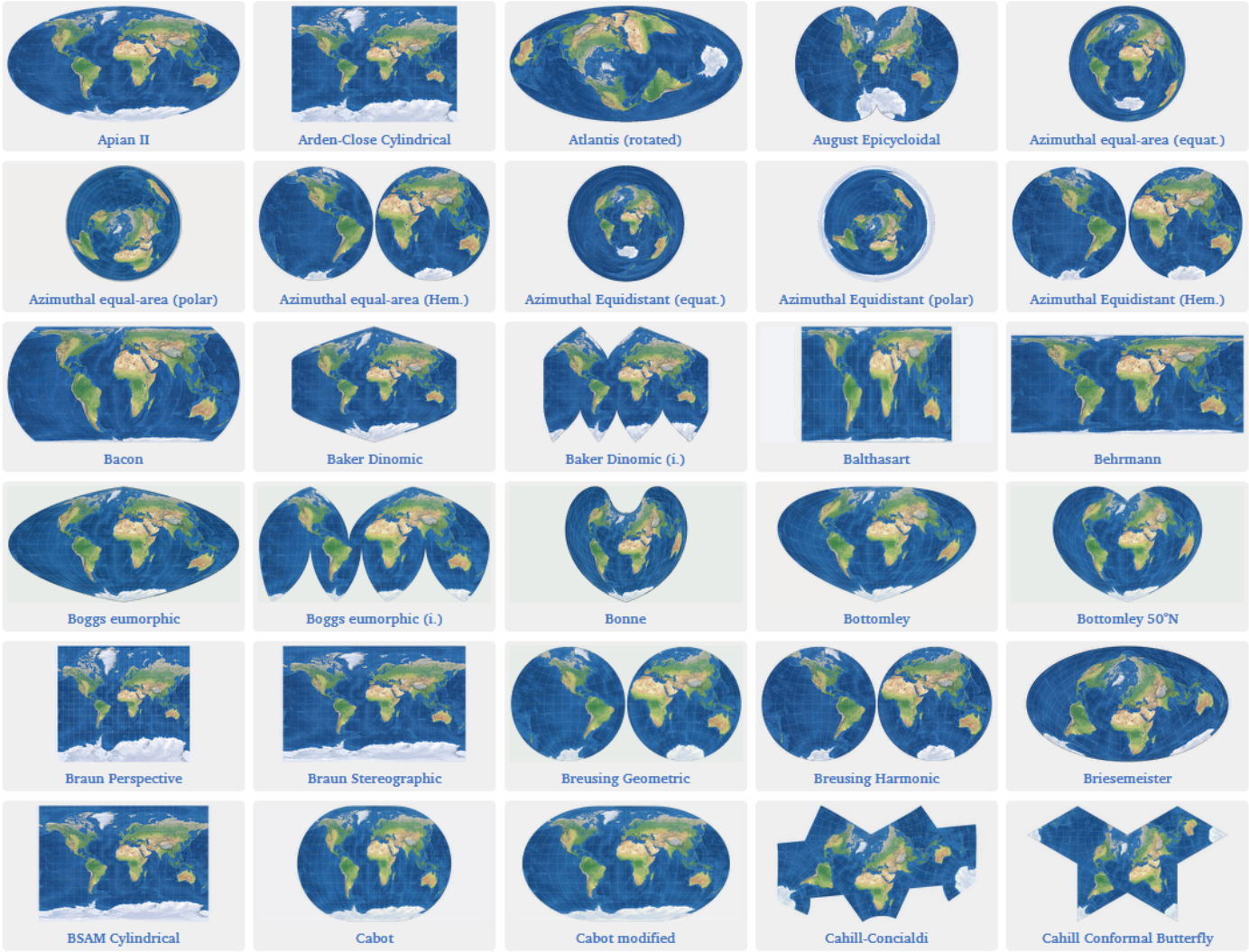

Type of question:
Do data match an expected ratio?
Is there an association between two variables?
Do samples come from the same or different populations?
Type of data: discrete, categorical (counts,frequencies)
Type of data: non-parametric (nominal, ordinal, interval) parametric (interval)
Type of data: non-parametric parametric
a priori expectation? yes no
one variable, one sample # categories: two more than two
one variable, two or more samples, two or more categories
χ²-goodness of fit test with Yates's corr.; Fischer's exact
χ²-goodness of fit test
χ²-test for heterogeneity
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 (**)
treatment variables: one two
treatment variables: one two or more
categories two two or more
two-way non-parametric anova
categories two two or more
multi-way parametric anova (possibly nested) (*), (**)
data: unpaired paired
one-way anova, Kruskal-Wallis
data: unpaired paired
one-way parametric anova (*), (**)
observ. per sample less than 30 at least 30
z-test for paired data (*)
variances homogeneous yes no
Welch t-test
observ. per sample less than 30 at least 30
two variables, both variables have two categories at least one variable has more than two categories
χ²-test for association with Yates's corr.; Fischer's exact
χ²-test for association
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 (*)
Wilcoxon's rank paired test (***)
observ. per sample less than 30 at least 30
t-test for unpaired data (*)
z-test for unpaired data (*)
t-test for paired data (*)
z-test for paired 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

intermission

geopandas

The earth is rounds… my monitor is flat





incomplete list of map projections

5

reproducible analysis

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

1

assures a result is grounded in evidence

#openscience

#opendata

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

facilitates scientific progress by avoiding the need to duplicate unoriginal research

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

3

facilitate collaboration and teamwork

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

- provide raw data and code to reduce it to all stages needed to get outputs
- provide code to reproduce all figures
- provide code to reproduce all number outcomes

github

reproducibility

https://github.com

Reproducible research means:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

allows reproducibility through code distribution

github

version control

https://github.com

the Git software

is a distributed *version control system:*

a version of the files on your local computer is made also available at a central server.

The history of the files is saved remotely so that any version (that was checked in) is retrievable.

allows version control

github

collaborative platform

https://github.com

collaboration tool

by fork, fork and pull request, or by working directly as a collaborator

allows effective collaboration

key concepts

Statistical tests:

how to use it

choose measure and compare a pivotal quantity

how to choose it :

first ask yourself what kind of data? what kind of question?

key concepts

NHRT setup:

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

From idea to hypothesis

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

Gaussian and Poisson distribution

Moments of a distribution

key concepts

Reproducible research

why do we want research to be reproducible?

1) more trustworthy, 2) faster progress

what constitute reproducible research?

2) your plots can be remade and your numbers rederived

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

reading

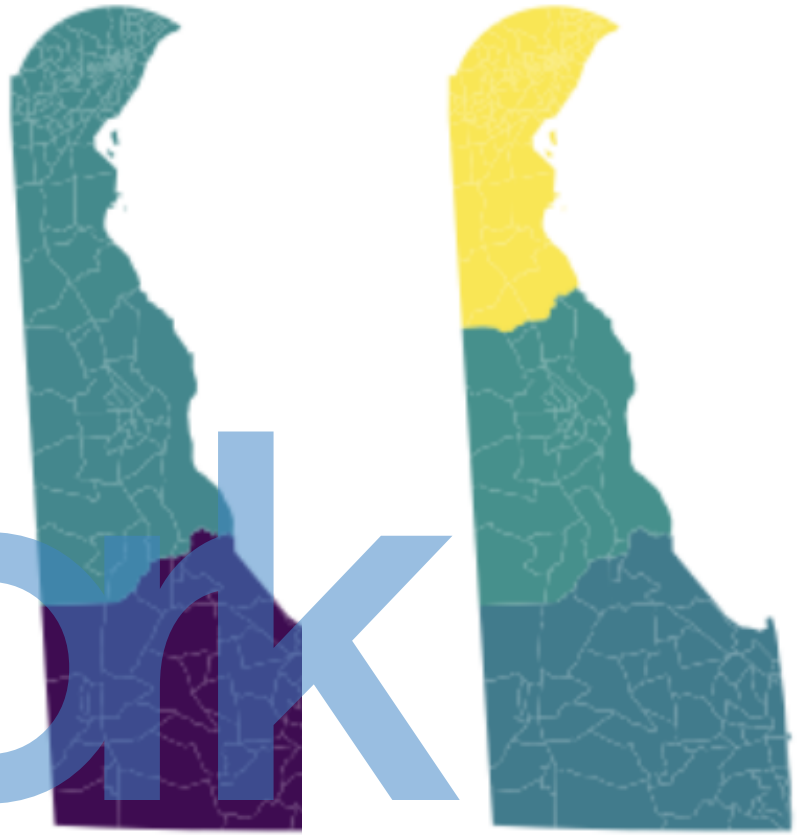
Plot Delaware mean age by gender and country:

Get a CENSUS API key

Extract the relevant census data

plot it on maps of DE

homework



https://github.com/fedhere/PUS2020_FBianco/HW4