STATISTICS & ESTIMATION IN EXPLORATORY DATA ANALYSIS

w/ slides from Pierre Dragicevic

UNIVERSITY OF CALGARY

GOALS

Understand the role of statistics in **exploratory analysis**

Understand **sampling** and **confidence** intervals

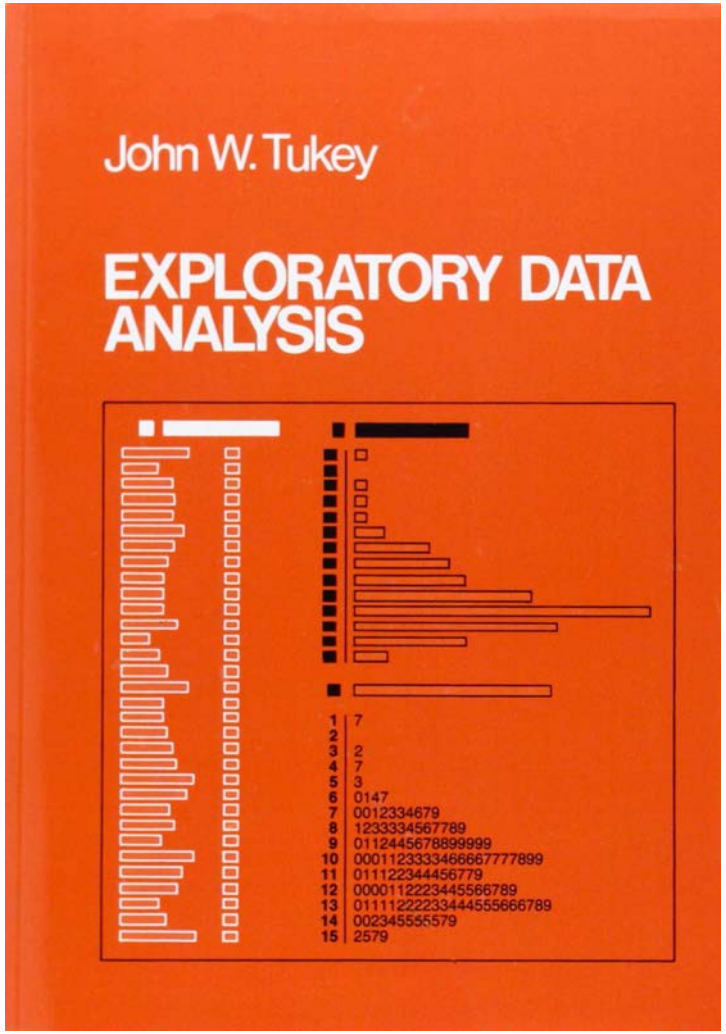
Develop a **healthy skepticism** about apparent patterns

Create visualizations that **keep variation in data visible**

STATS & VISUALIZATION

Exploratory Data Analysis

Tukey, 1977



STATISTICAL METHODS

UNDERSTANDING WHAT CONCLUSIONS AND INFERENCES WE CAN REASONABLY DRAW

STATS & VISUALIZATION

Exploratory data analysis is sometimes compared to detective work: it is the process of gathering evidence.

Confirmatory data analysis is comparable to a court trial: it is the process of evaluating evidence.

Exploratory analysis and confirmatory analysis "can–and should–proceed side by side" (Tukey; 1977).

Quoted from the SAS Institute

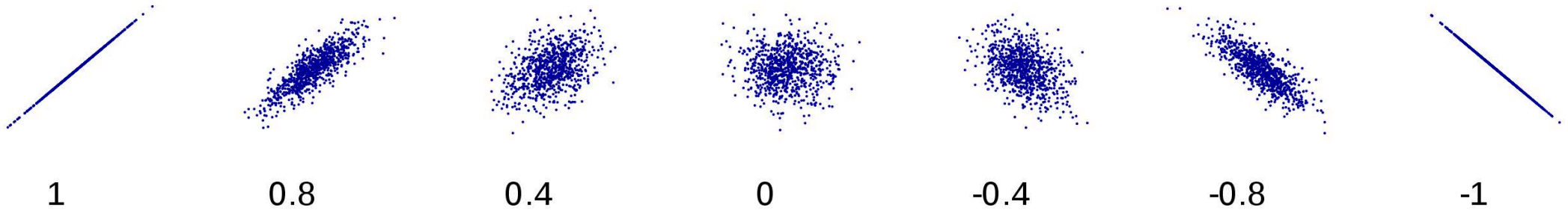
WHAT ARE STATS?

A set of tools and methods

With an old tradition:

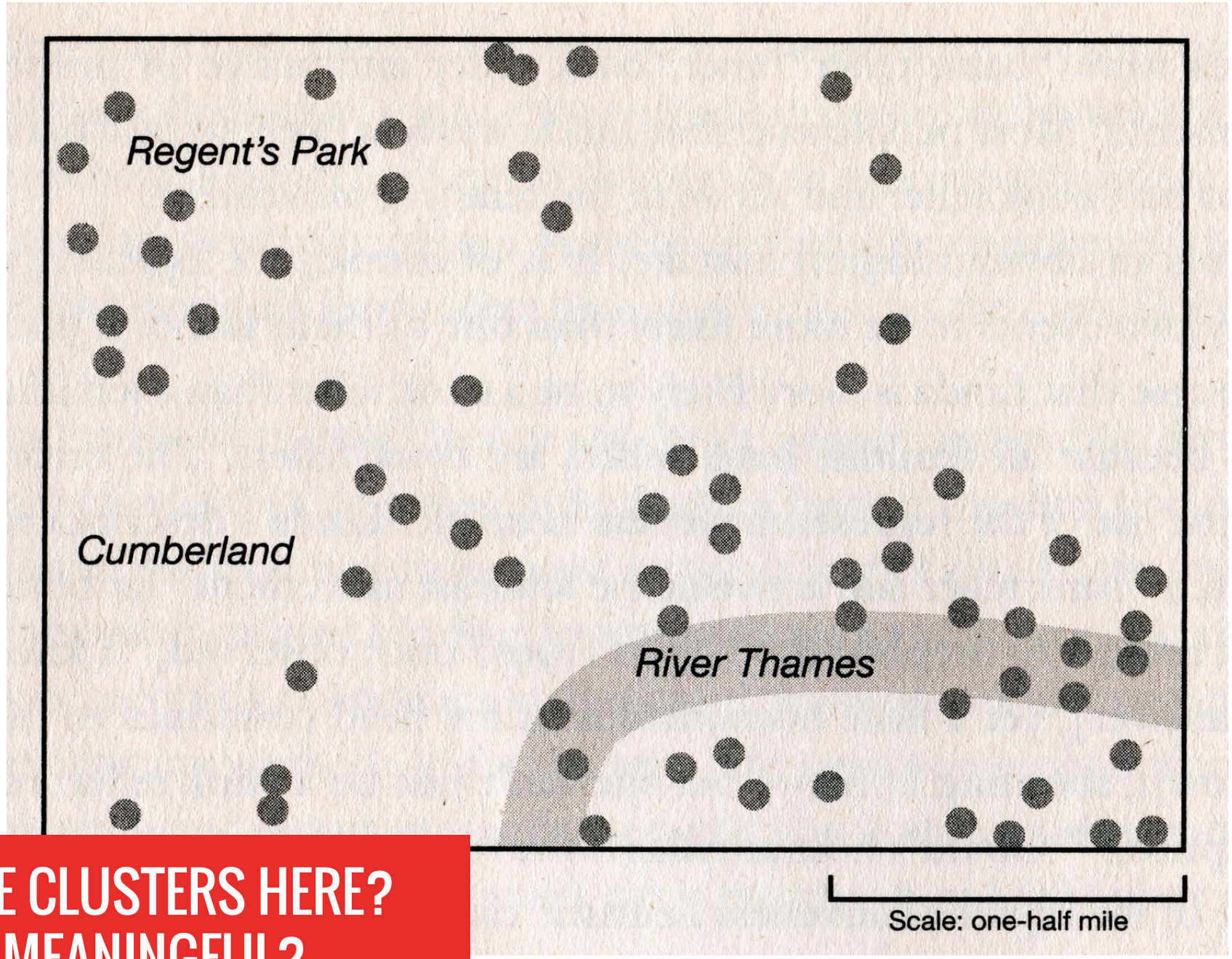
- Origins in demographics
- Anchored in mathematics & probability theory
- Visual representations play a role
- A generally strong focus on (computationally cheap) numerical calculations

VISUAL INFERENCE

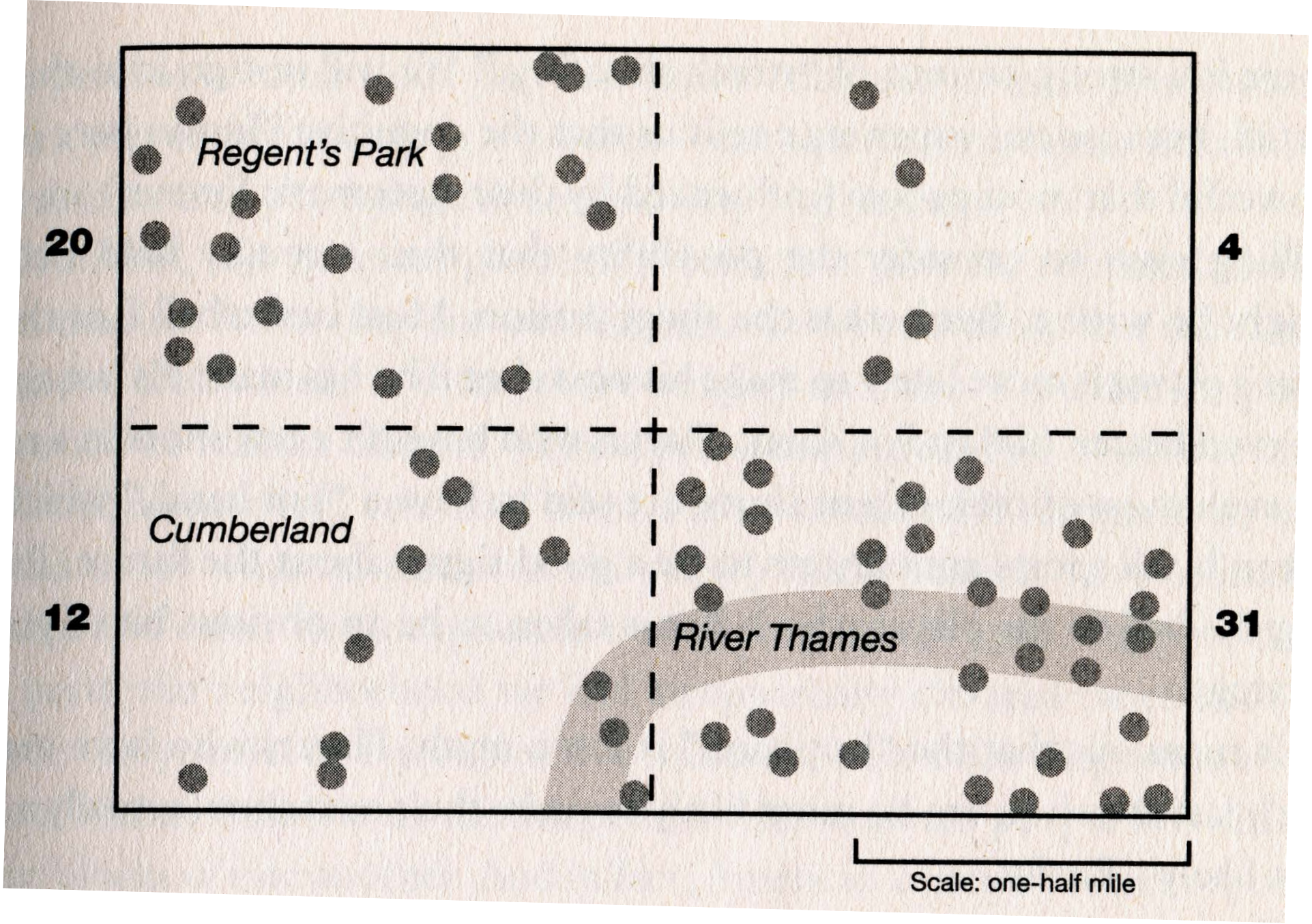


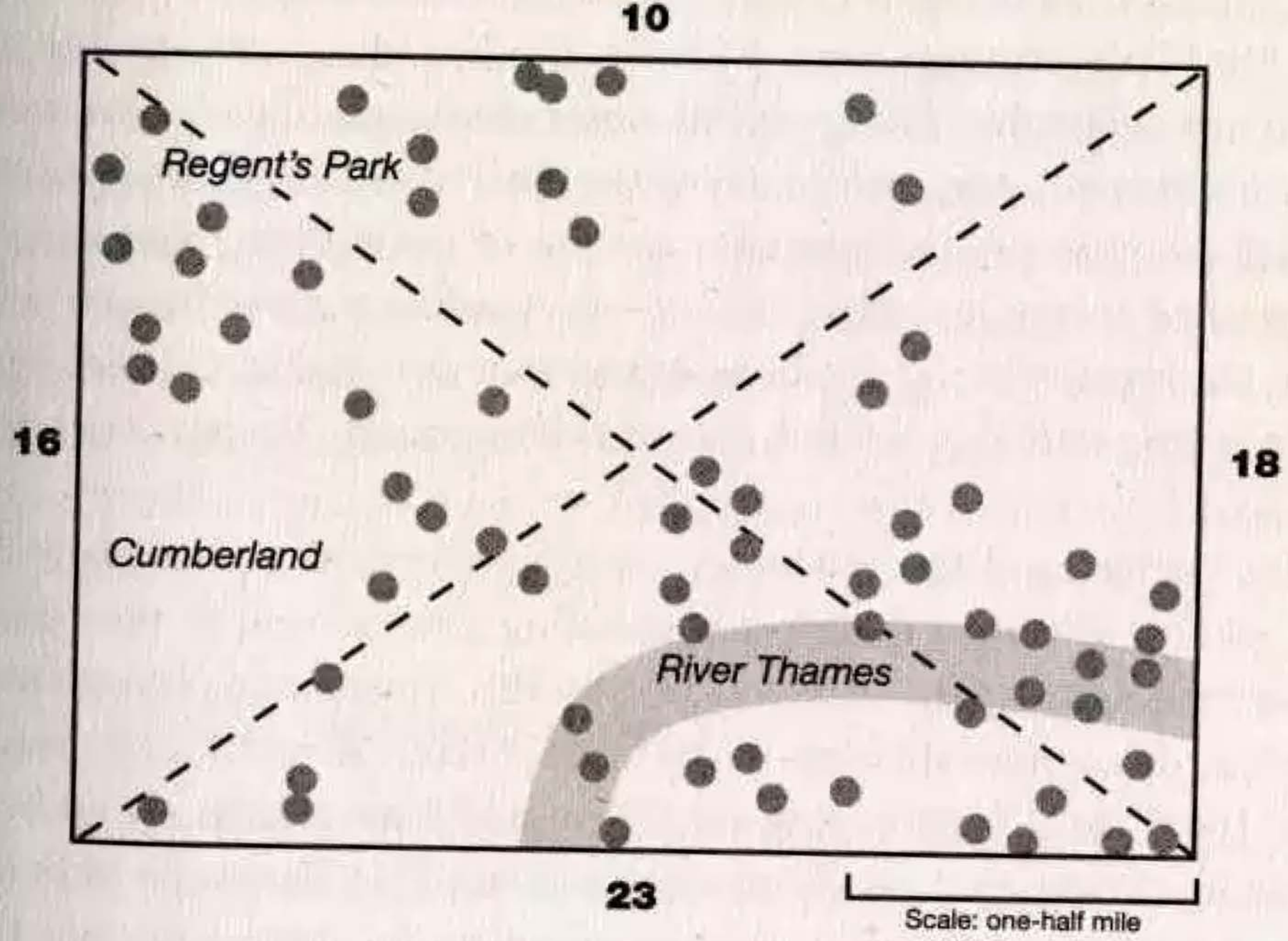
Whenever we make a judgement about a trend, correlation, difference, etc. **we are performing an implicit test.**

Extremely important to be aware of **variation**, **confidence**, and **bias.**

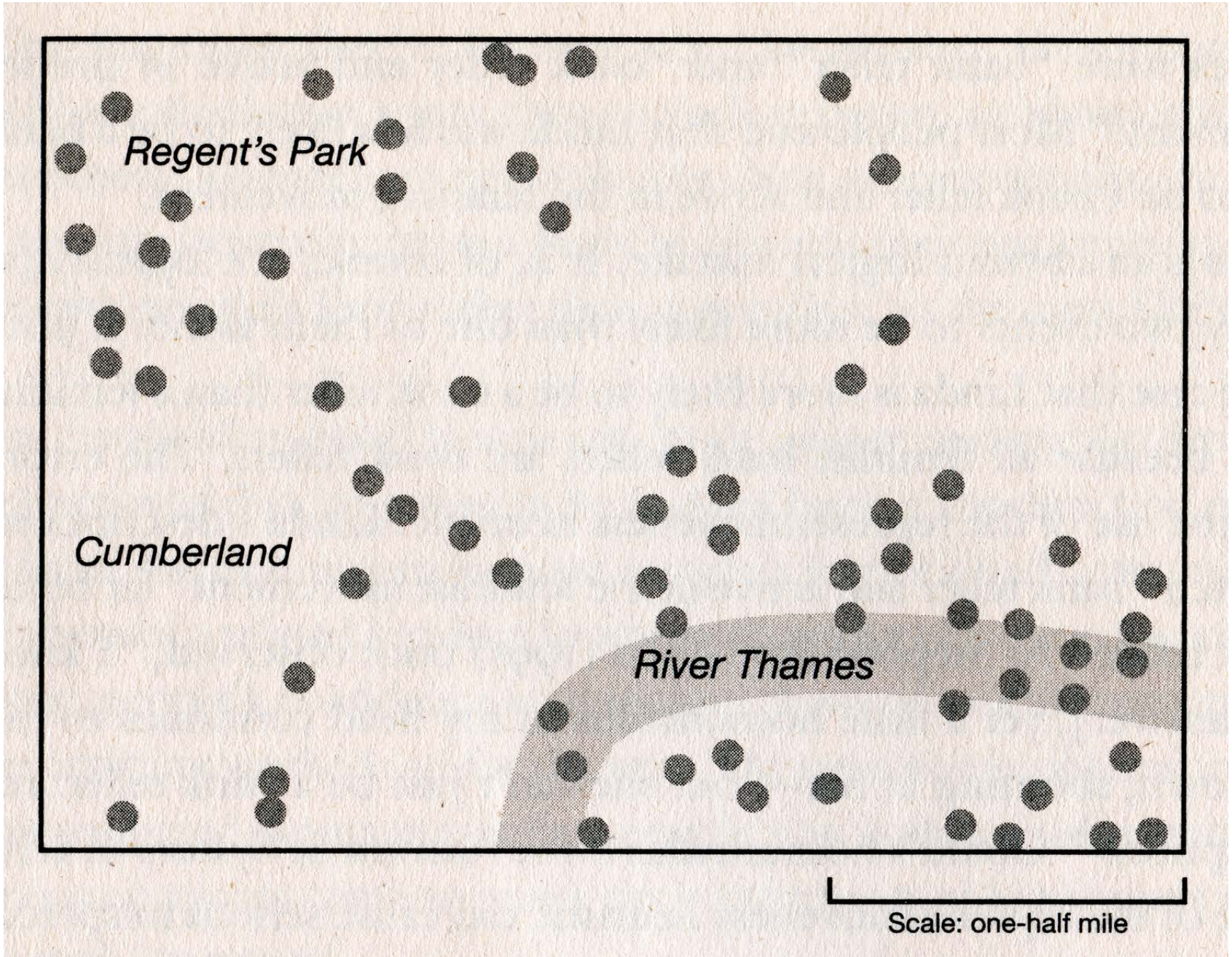


ARE THERE CLUSTERS HERE?
ARE THEY MEANINGFUL?


Regent's Park
20
4
Cumberland
12
River Thames
31
Scale: one-half mile


10
16
18
23
Regent's Park
Cumberland
River Thames
Scale: one-half mile

b


Regent's Park
Cumberland
River Thames
Scale: one-half mile

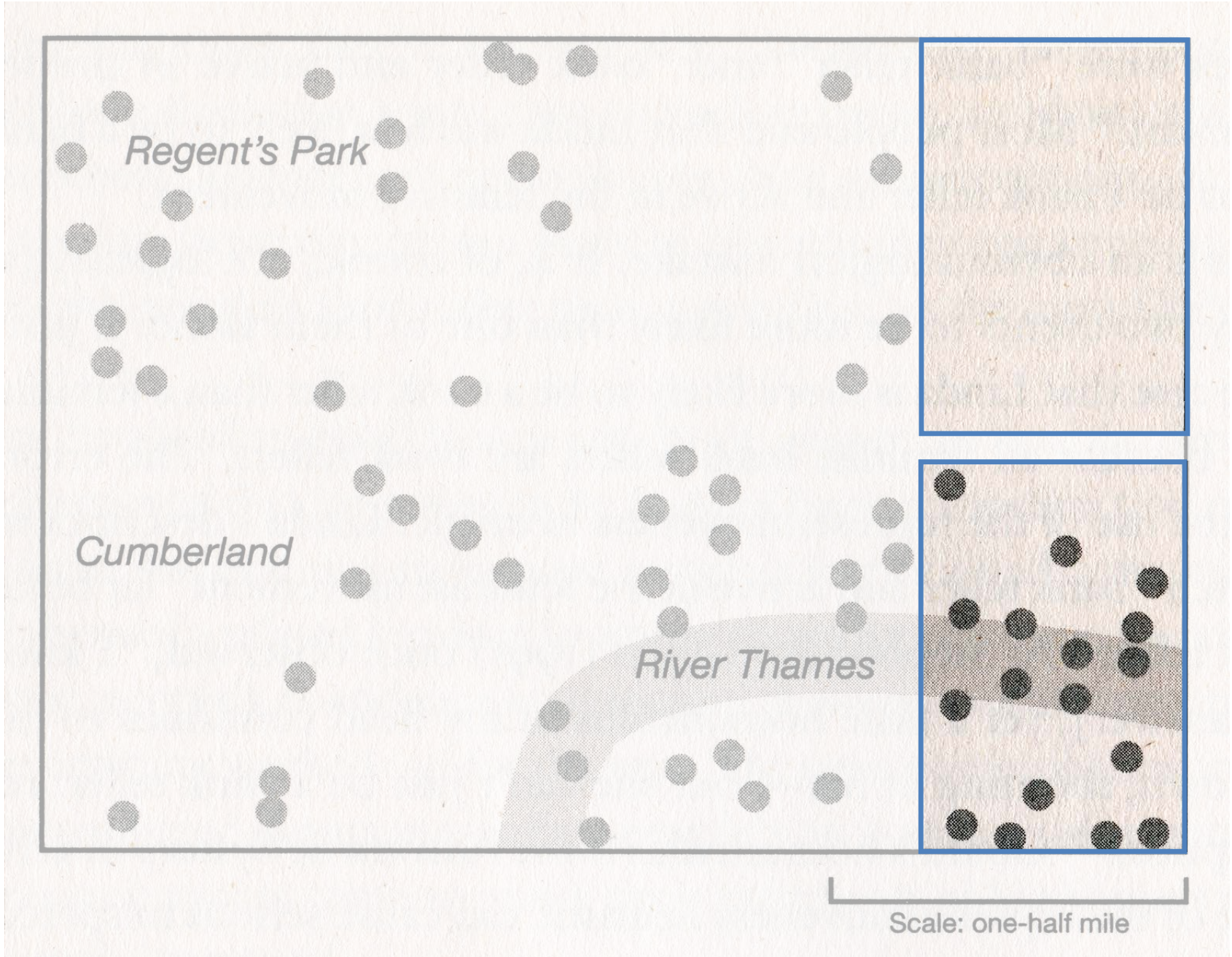
CONFIRMATORY DATA ANALYSIS

For answering questions rigorously

Strong focus on automatic procedures, computation and objectivity

Looking at data can impair objectivity:

- Cherry picking, snooping, fishing, data mining


Regent's Park
Cumberland
River Thames
Scale: one-half mile

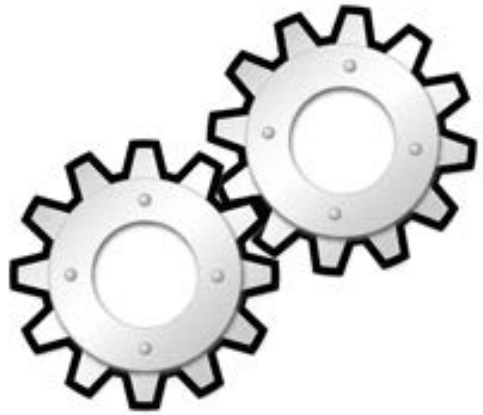
We can use statistical tools to help us understand the patterns, differences, etc. we observe in data.

...but we need to do so very carefully!

STATISTICAL TOOLS

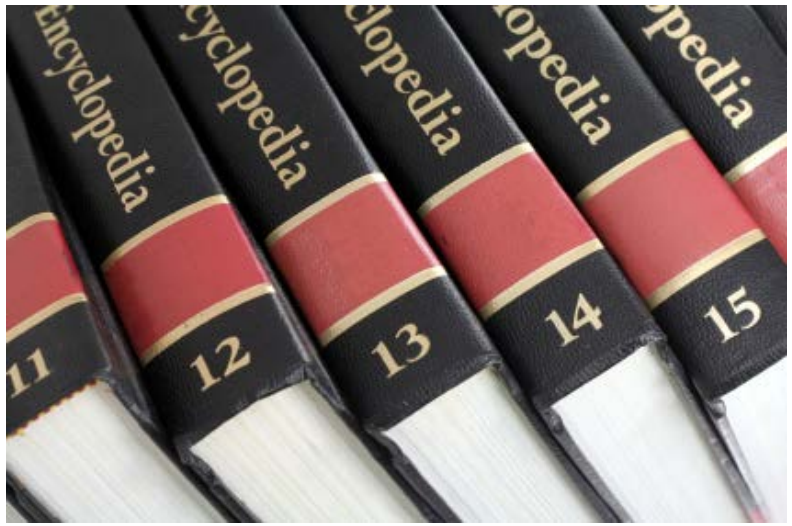
DESCRIPTIVE STATISTICS

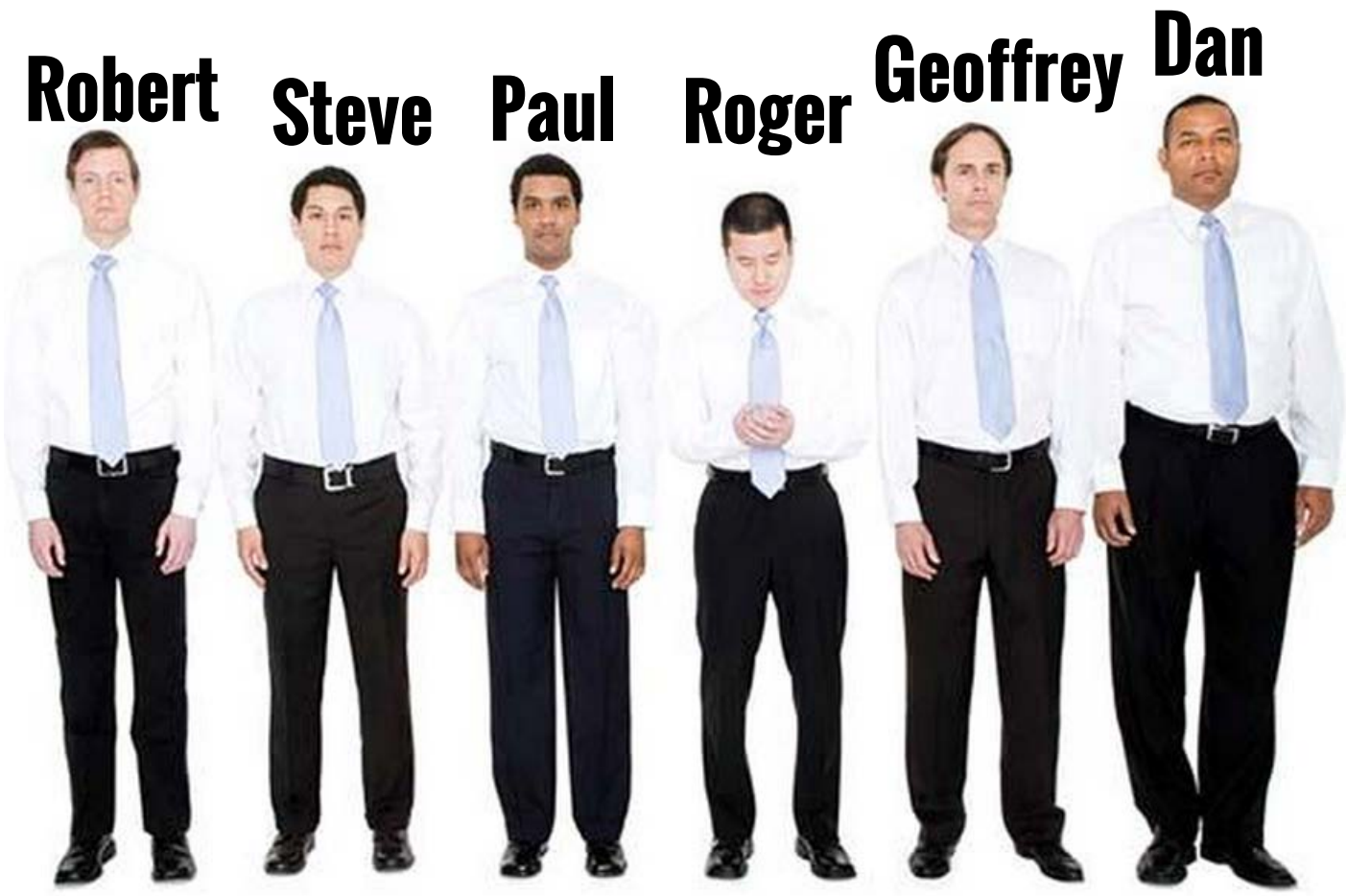
INFERENTIAL STATISTICS



AN EXAMPLE

Selling encyclopedias





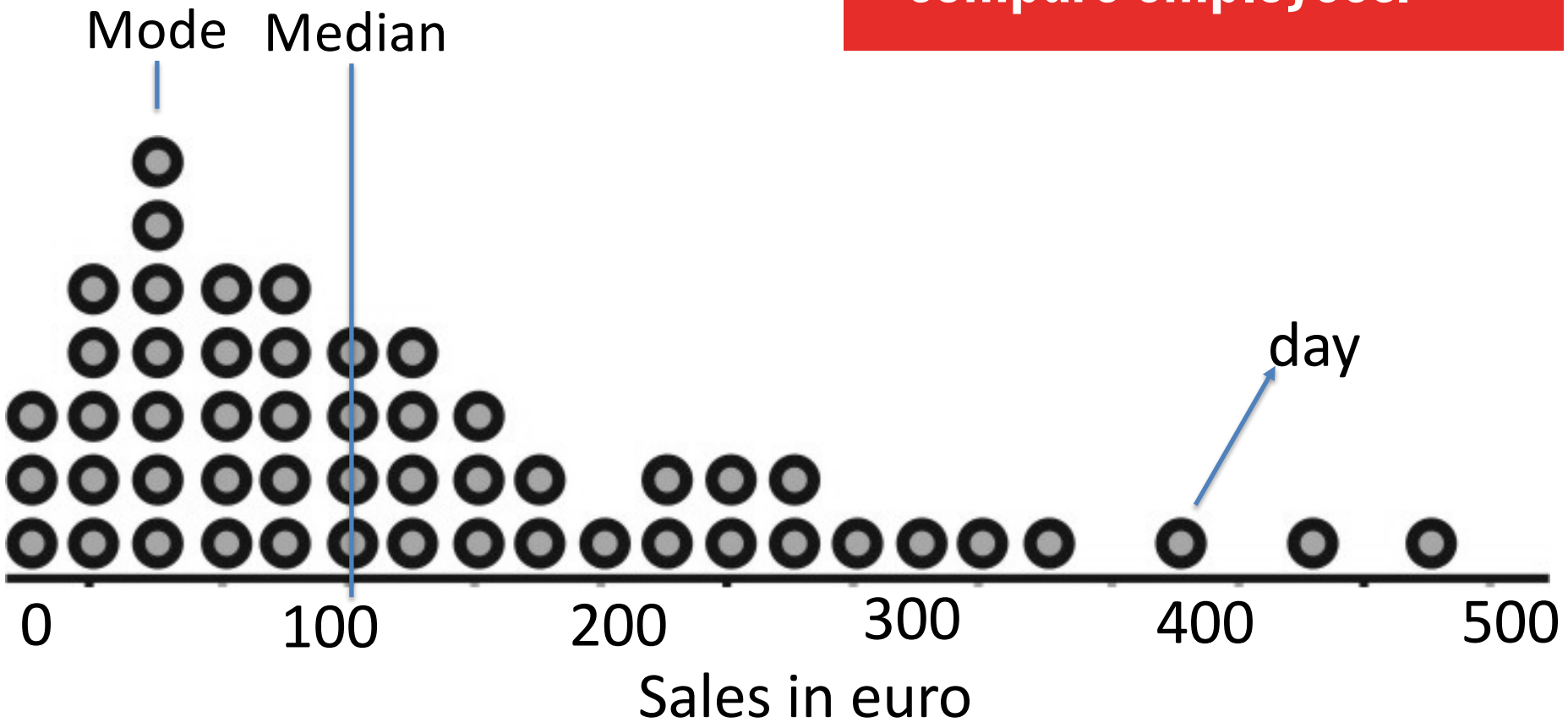
day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134
12	€47	€167	€126	€48	€93	€63
13	€34	€65	€55	€56	€333	€1,157
14	€76	€46	€89	€104	€56	€470
15	€75	€34	€184	€35	€299	€205
16	€68	€37	€275	€170	€57	€192
17	€126	€23	€114	€30	€43	€60
18	€43	€290	€89	€446	€57	€226


Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
1000
0
5
10
15
20
25
30
Can we see anything interesting?

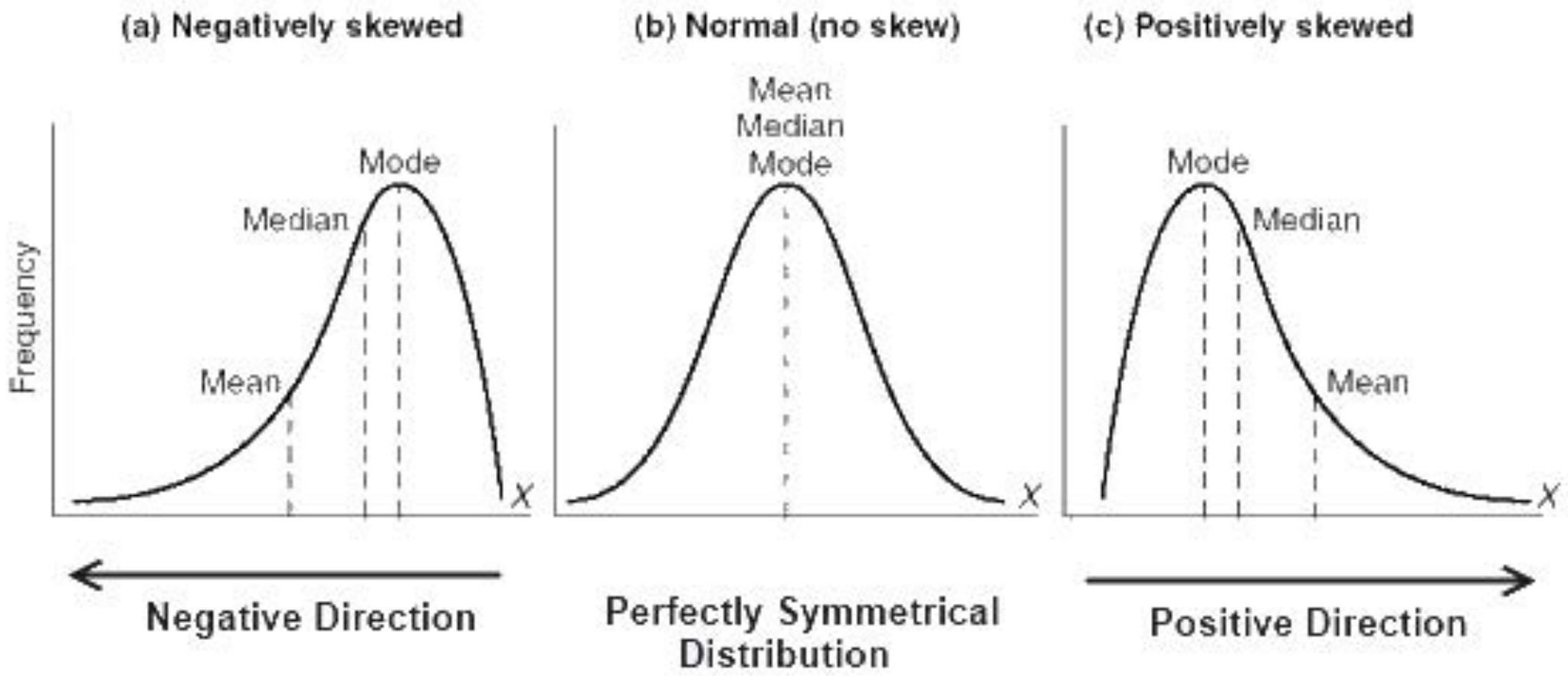
What if we compare distributions?

CENTRAL TENDENCY

One tool we might use to compare employees.



CENTRAL TENDENCY



From Shreya Sethi

DISPERSION

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

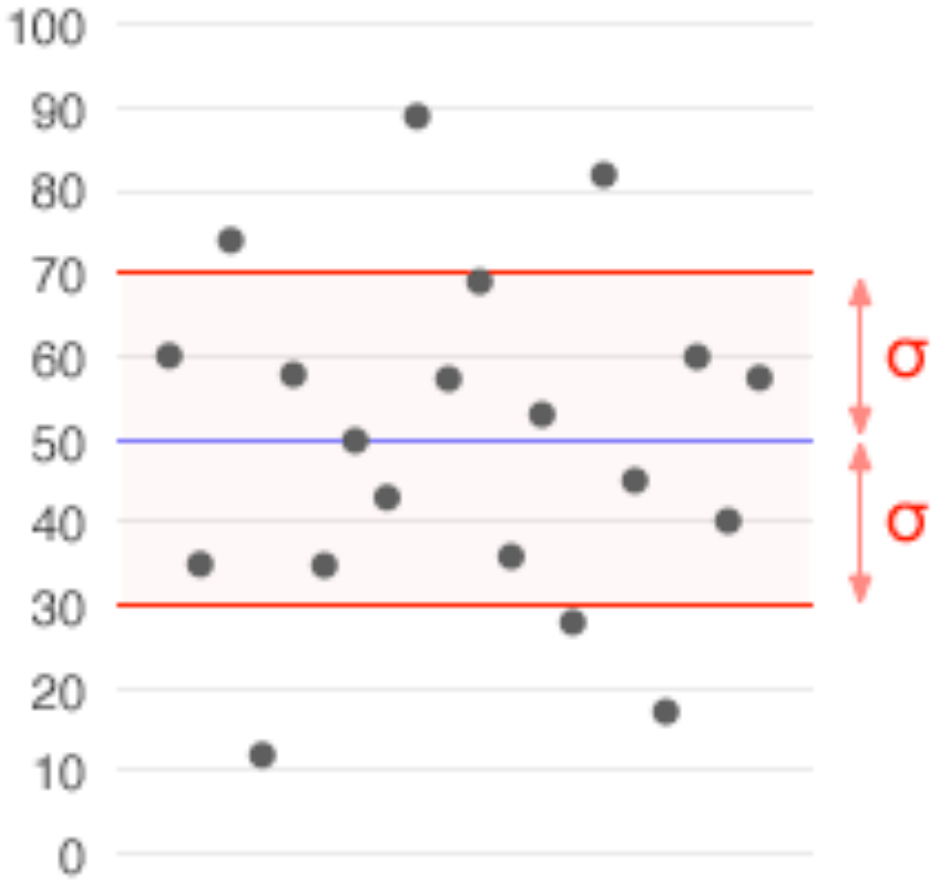
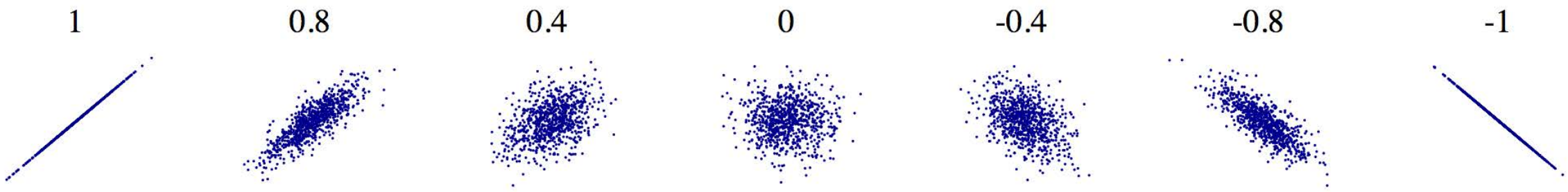


Image from Wikipedia

DEPENDENCE

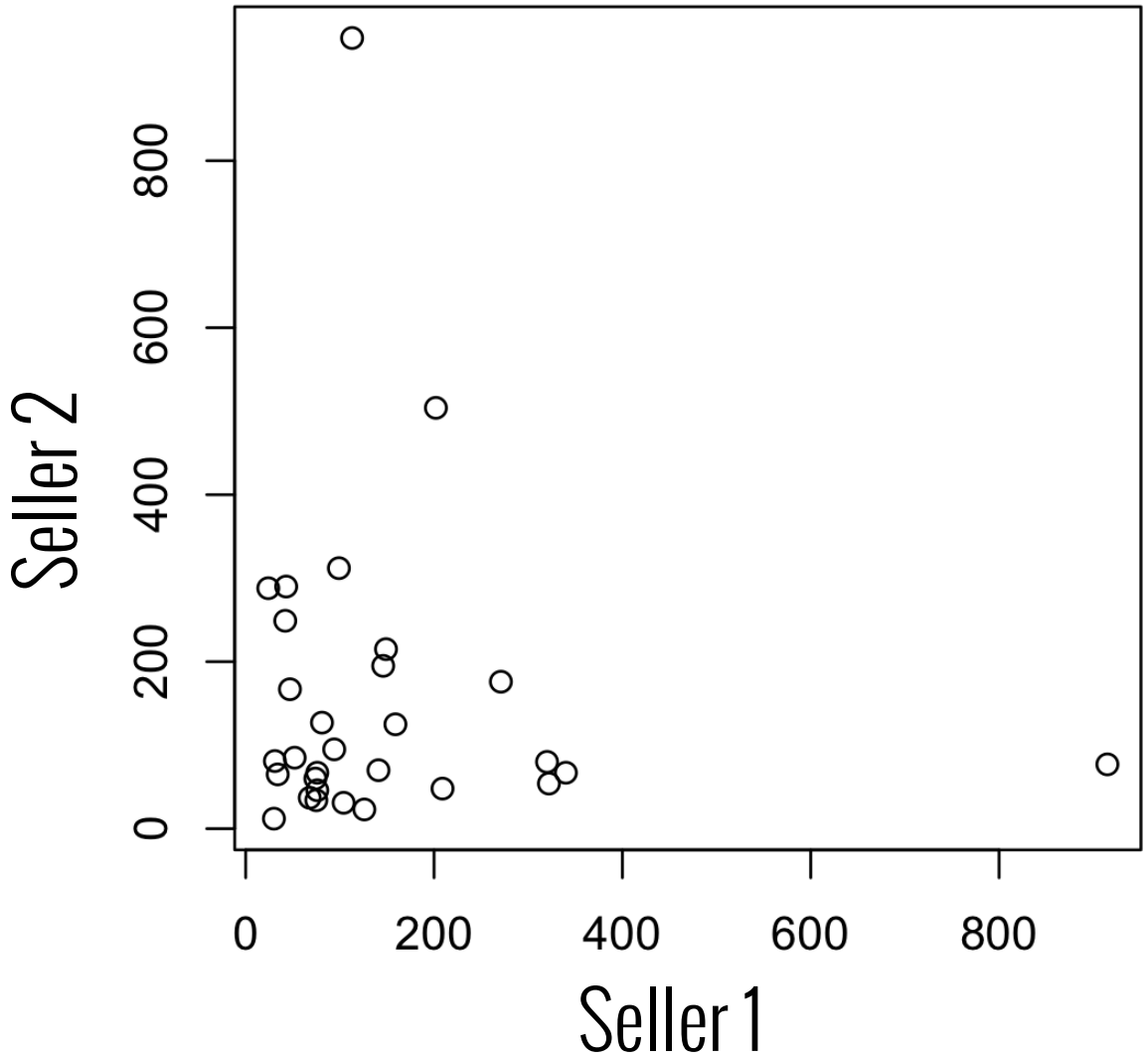
Correlation



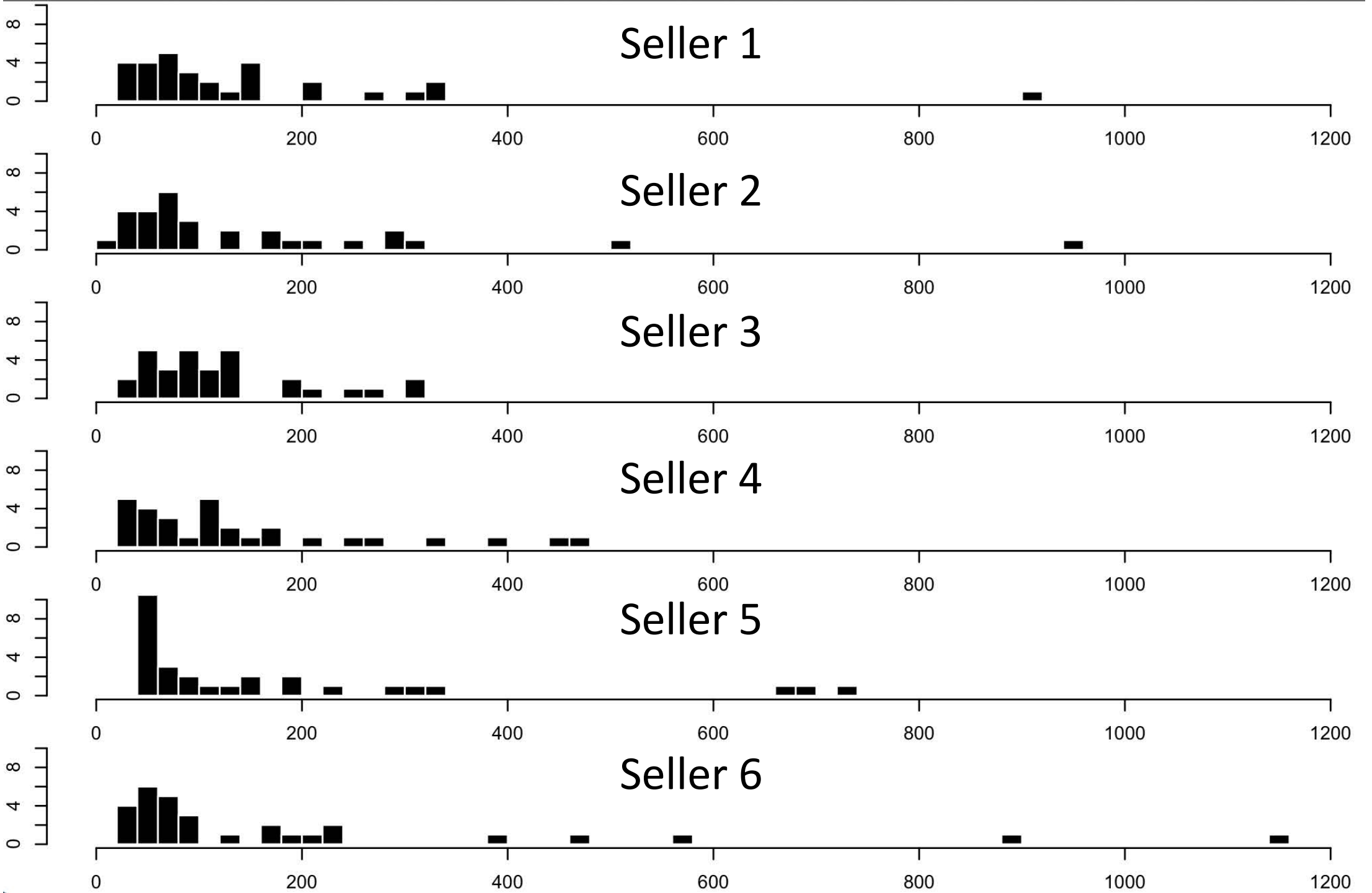
DEPENDENCE

Correlation

r = -0.08




Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
1000
0
5
10
15
20
25
30


Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
4
8
0
200
400
600
800
1000
1200

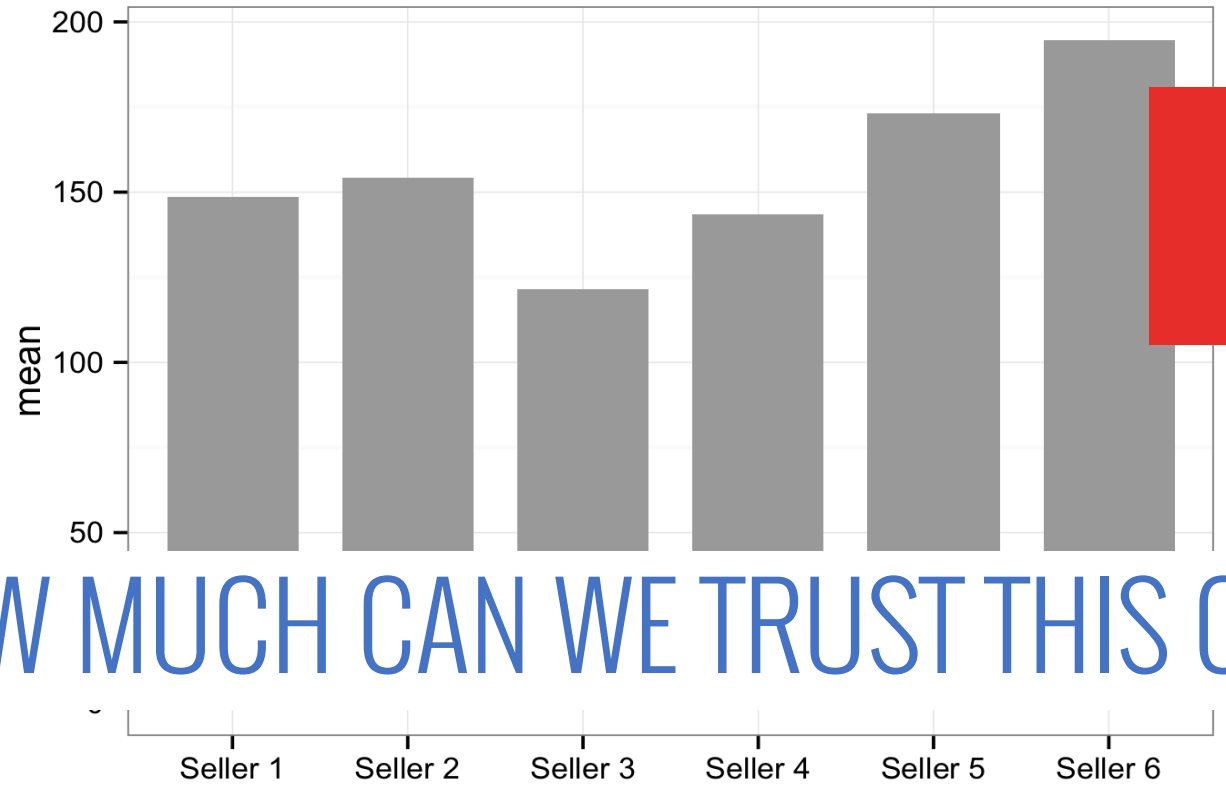
Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195

January 2018

Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195



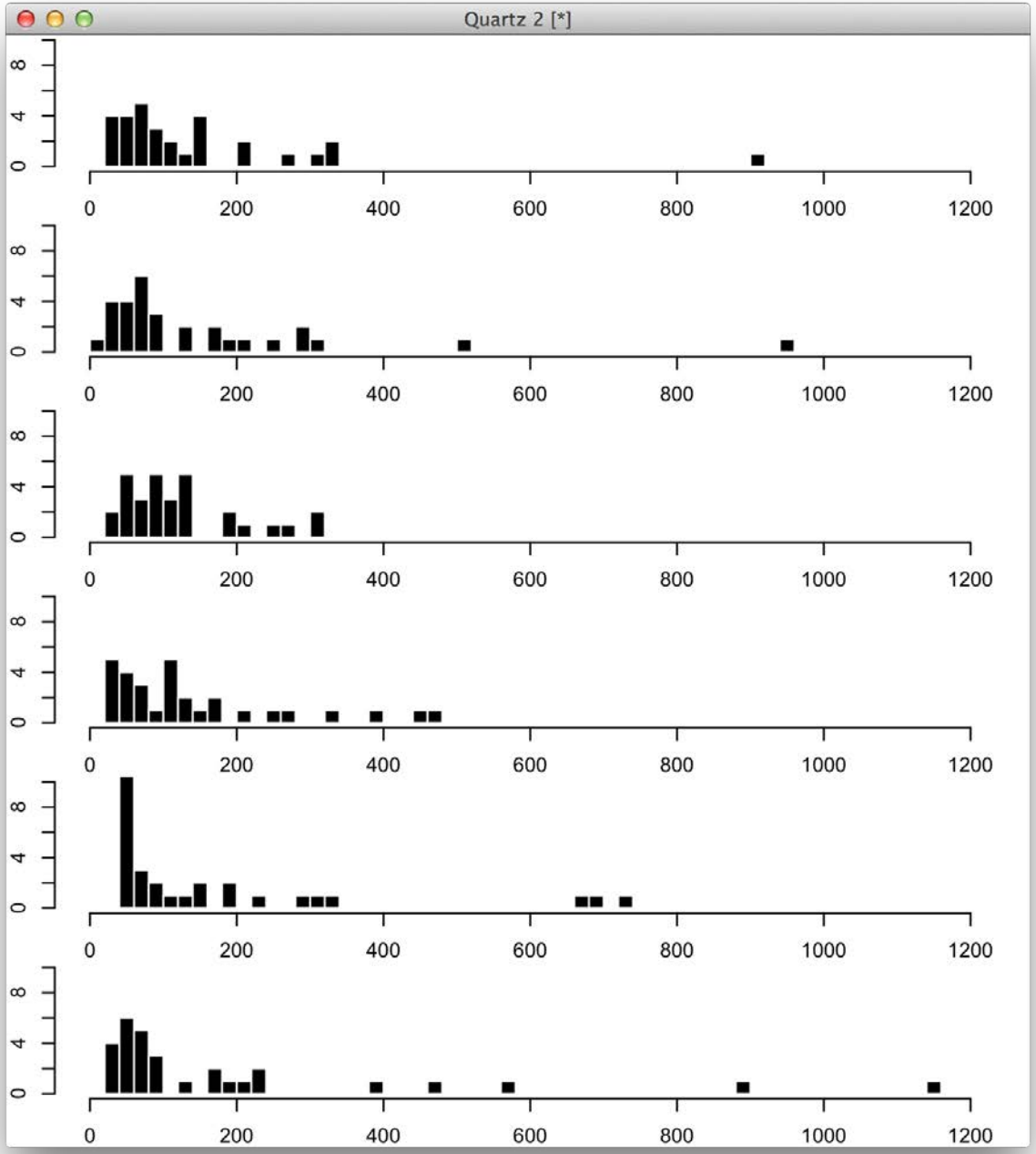
LOOKS LIKE SELLER 6 IS THE BEST!

CAN WE CALL IT A DAY THEN?

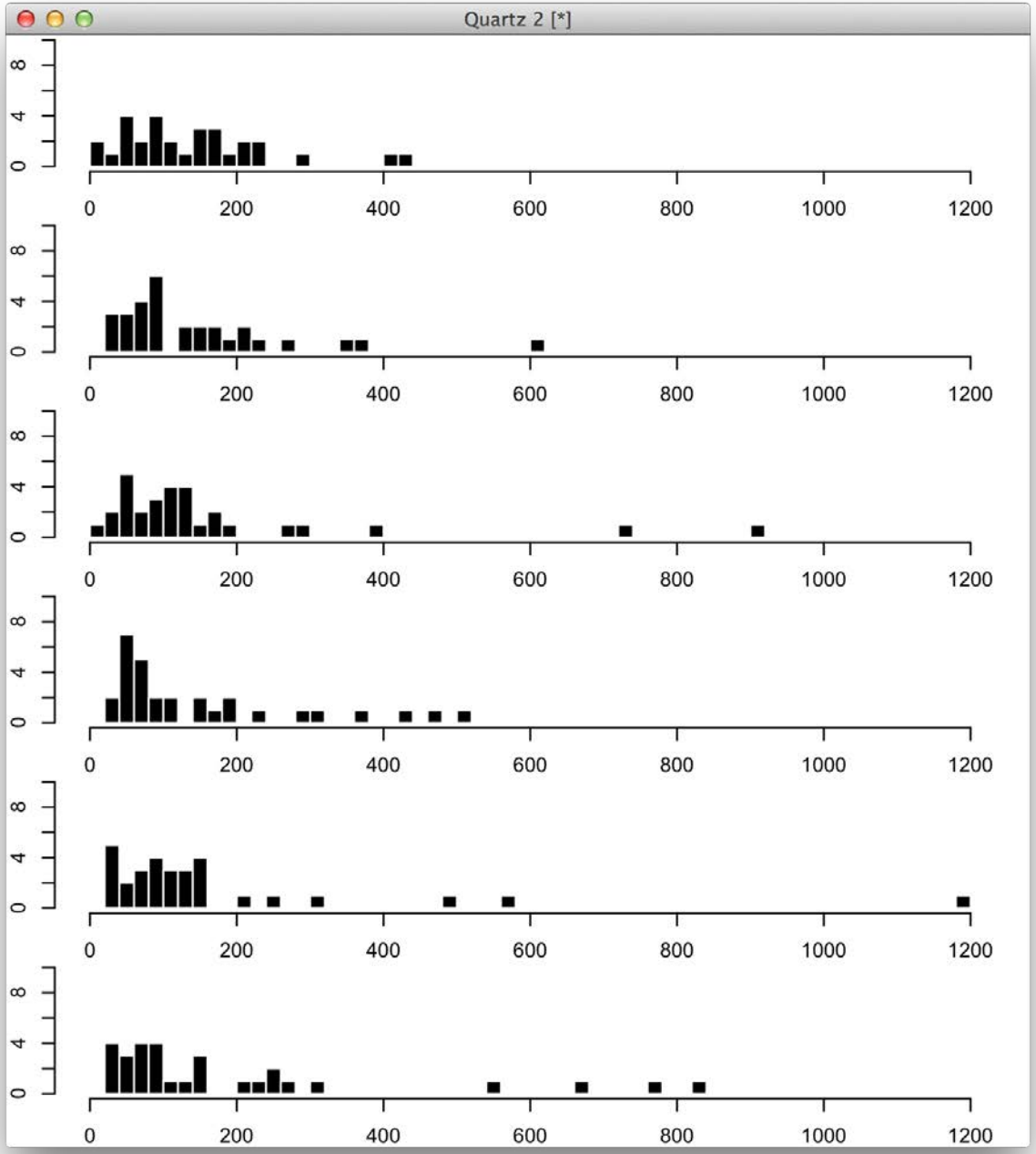
HOW MUCH CAN WE TRUST THIS CHART?

LET'S TRAVEL TO THE FUTURE

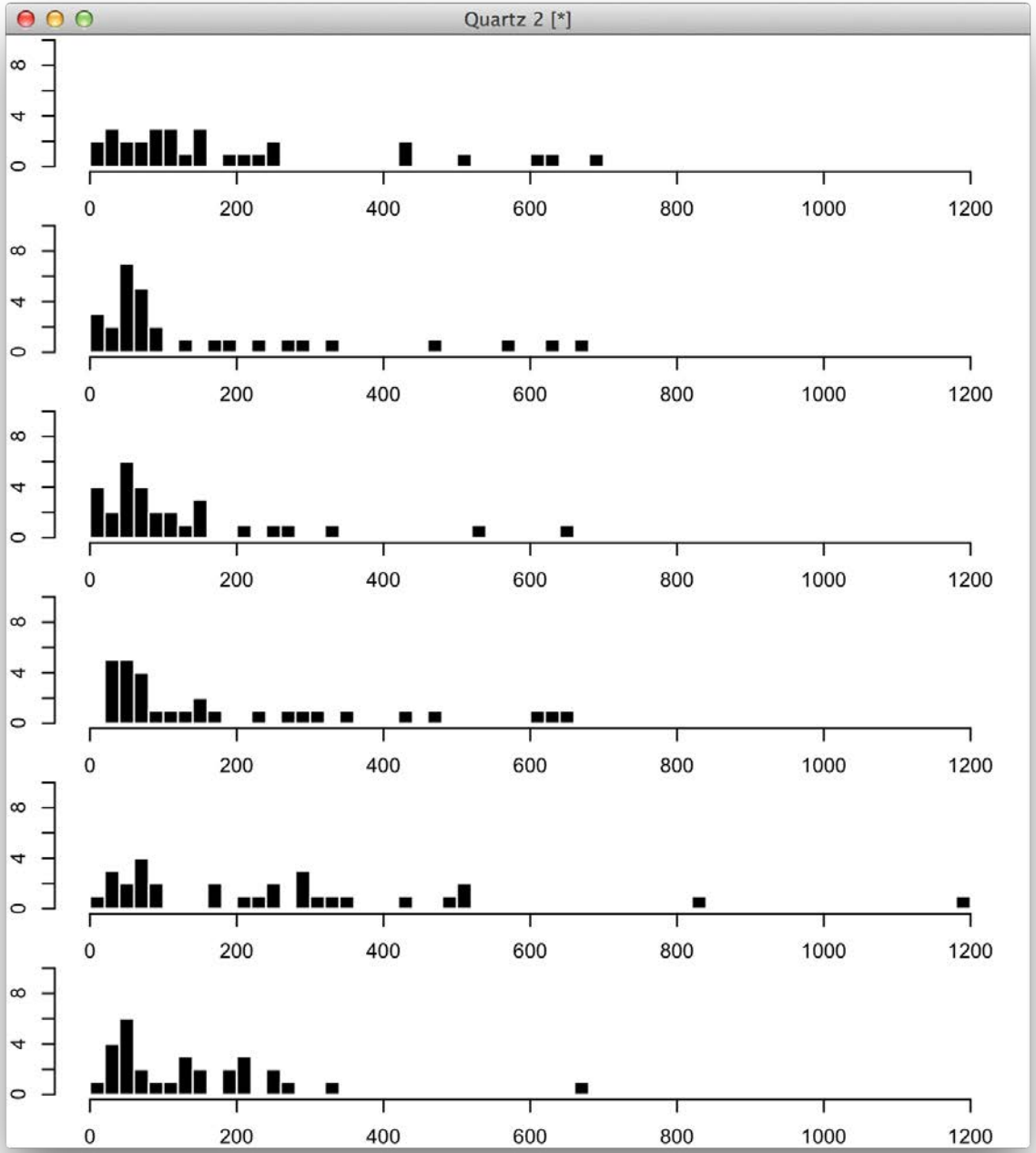
January 2020



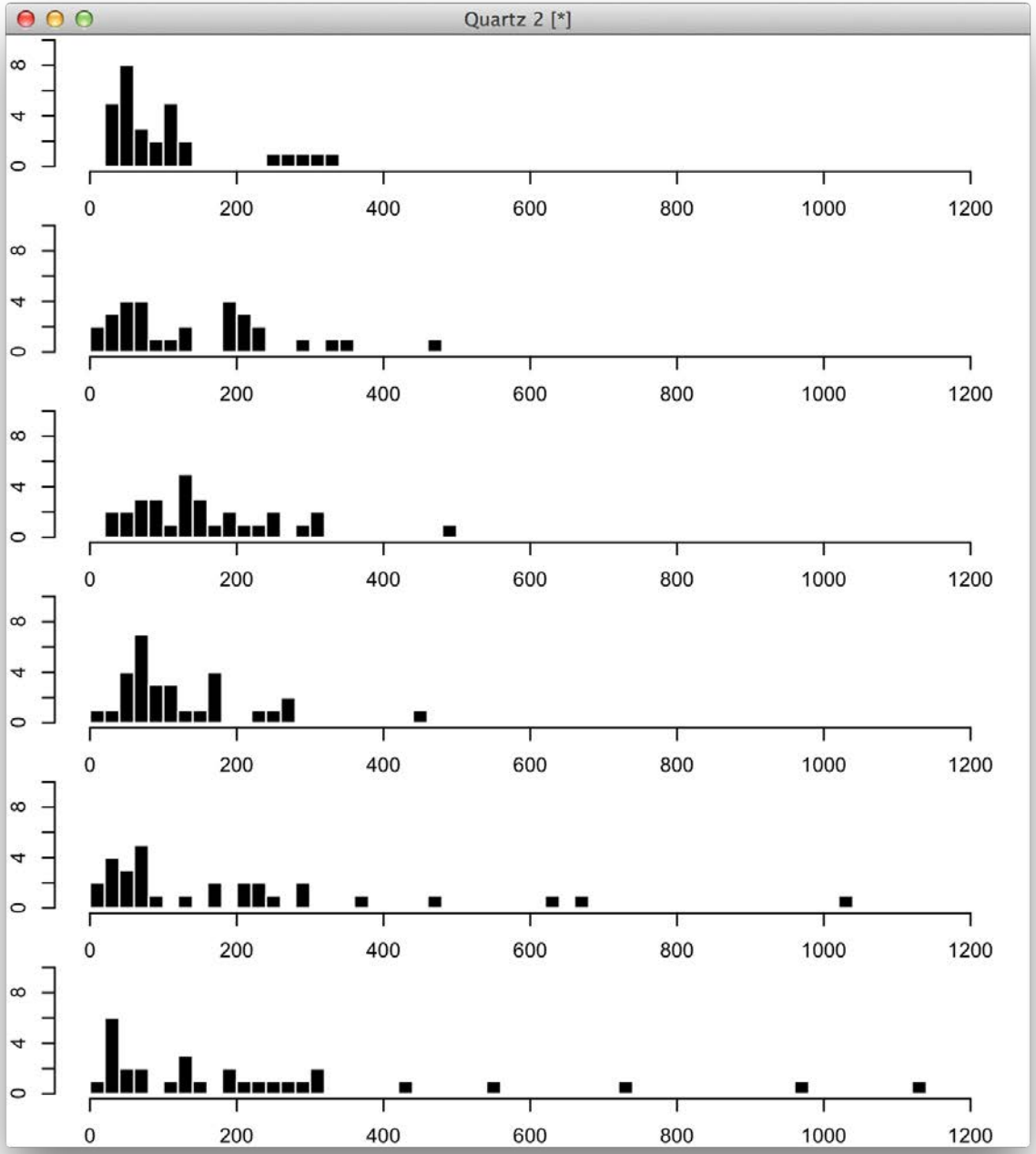
February 2020



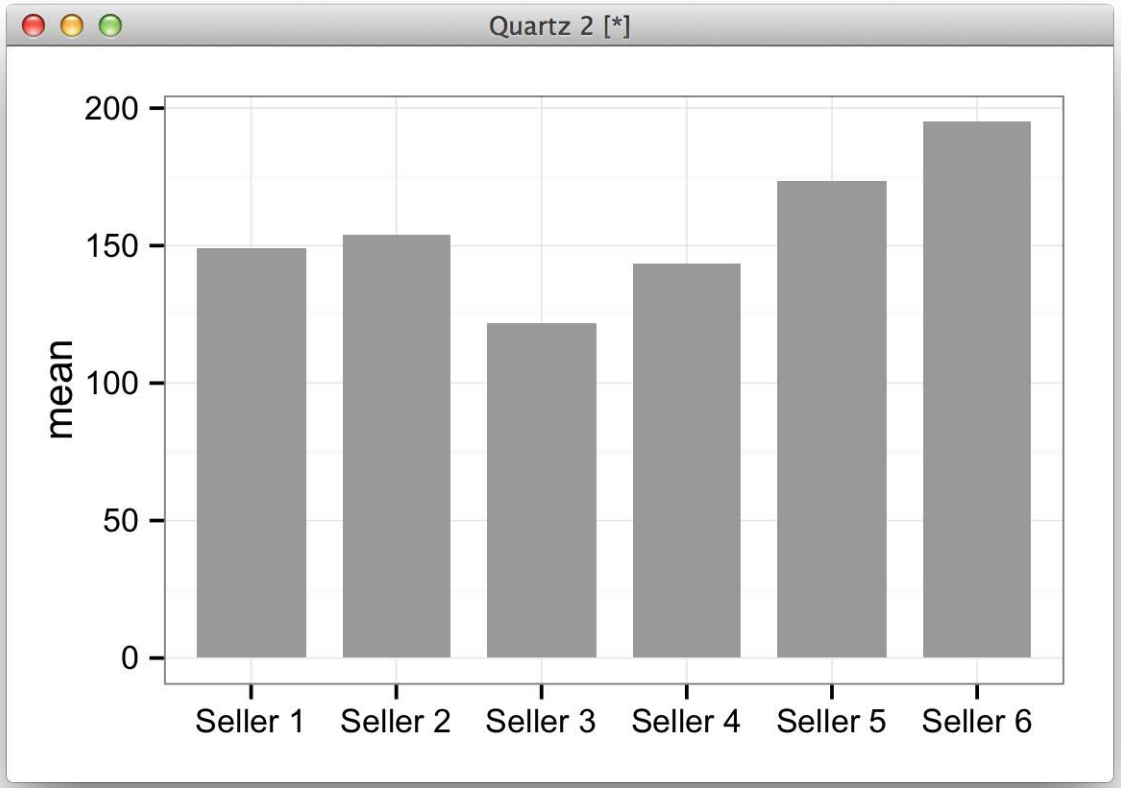
March 2020



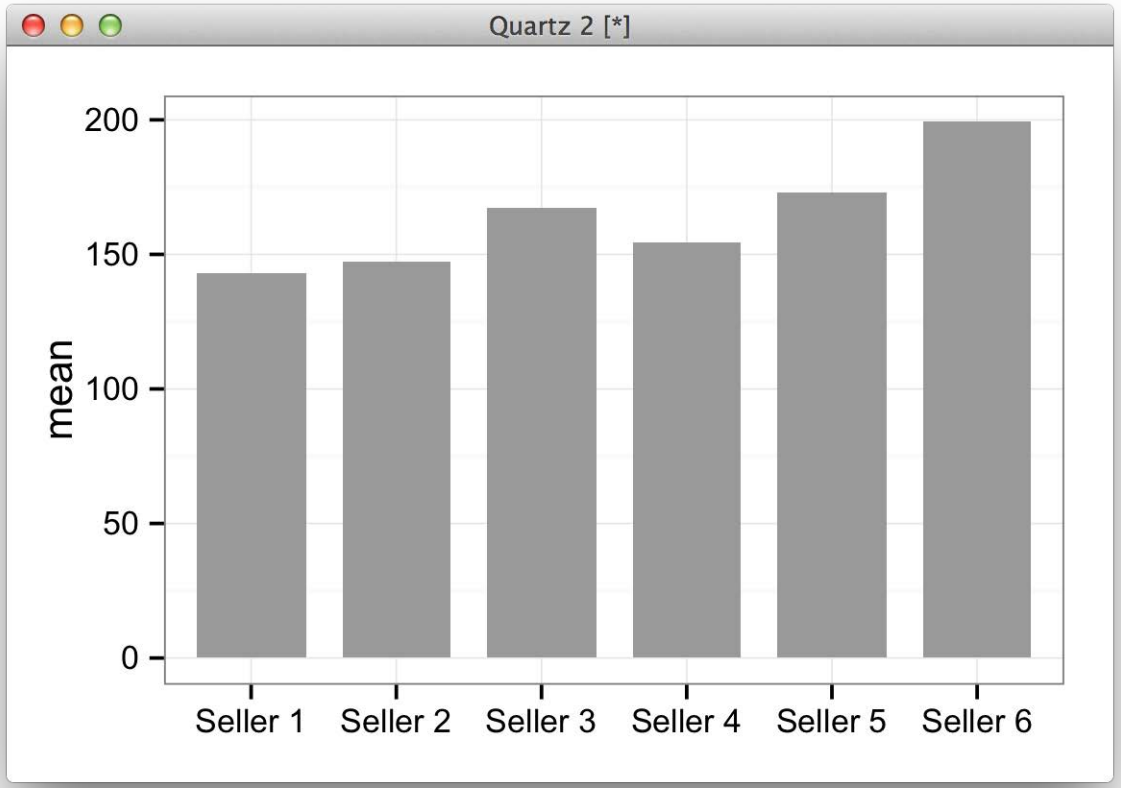
April 2020



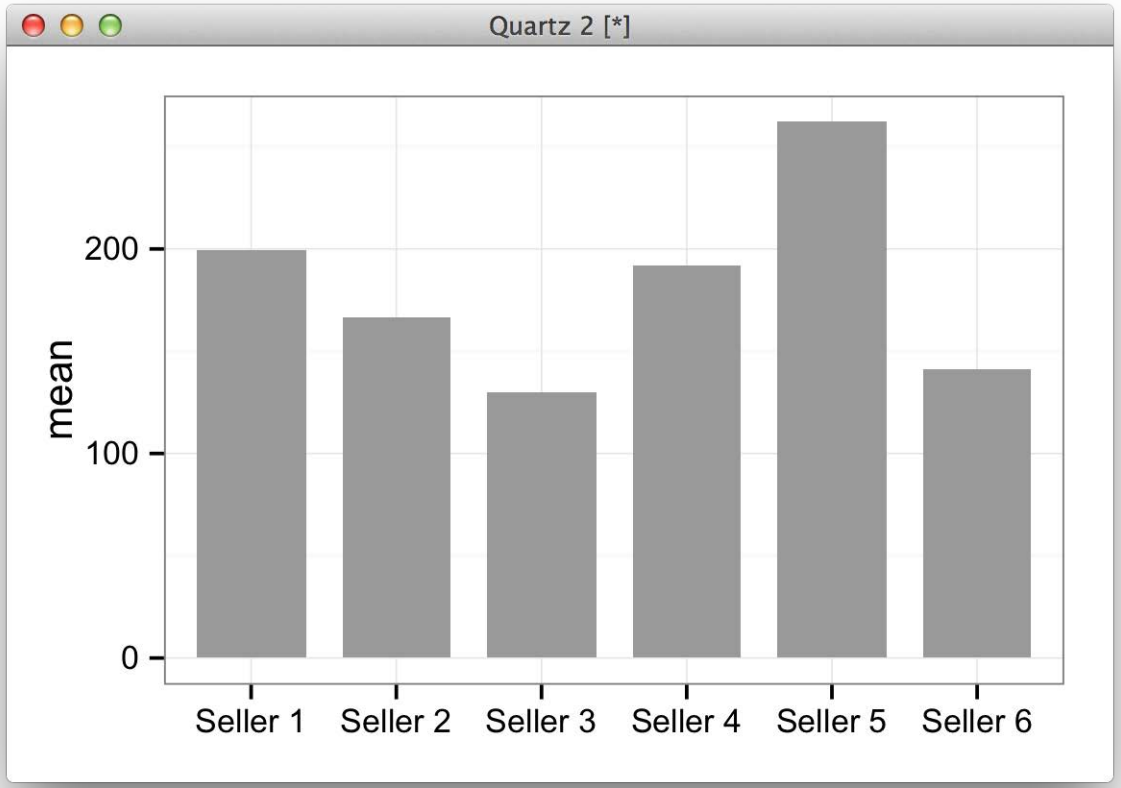
January 2020



February 2020



March 2020


Quartz 2 [*]
mean
200
100
0
Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6

April 2020

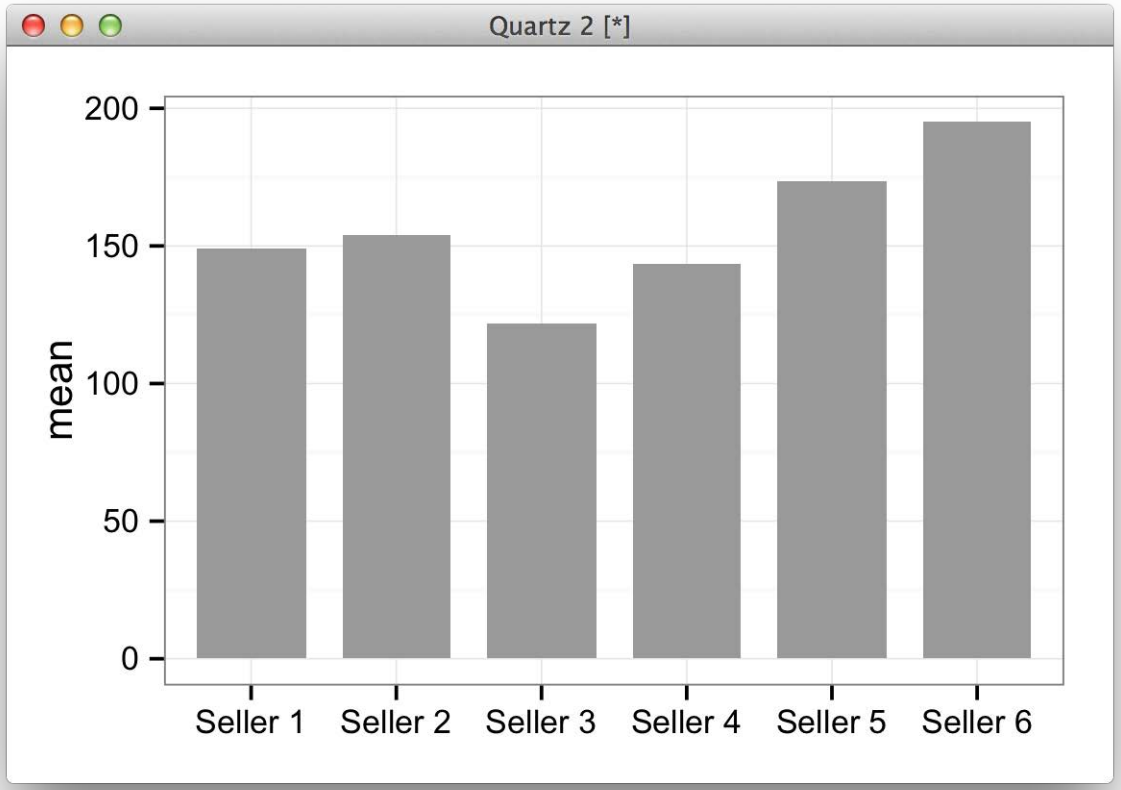

Quartz 2 [*]
250
200
150
100
50
0
mean
Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6

BACK TO THE PRESENT

January 2020

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134
12	€47	€167	€126	€48	€93	€63
13	€34	€65	€55	€56	€333	€1,157
14	€76	€46	€89	€104	€56	€470
15	€75	€34	€184	€35	€299	€205
16	€68	€37	€275	€170	€57	€192

January 2020

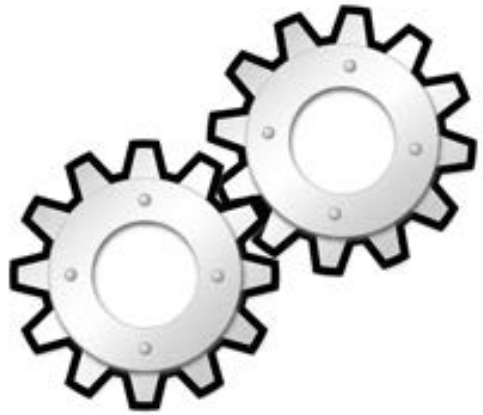


HOW MUCH CAN WE TRUST THIS CHART?

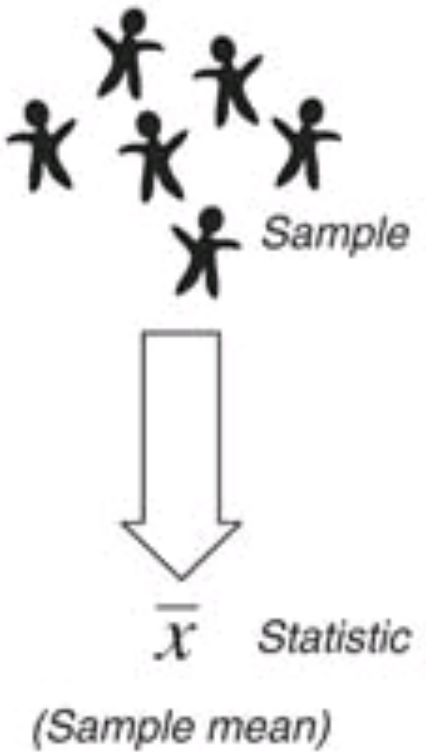
STATISTICAL TOOLS

DESCRIPTIVE STATISTICS

INFERENTIAL STATISTICS



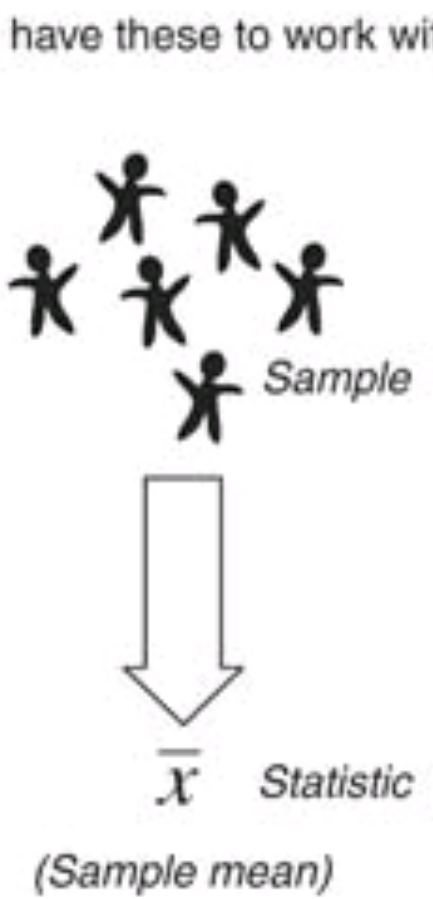
STATISTICAL INFERENCE



STATISTICAL INFERENCE

"POPULATION" IS A CONFUSING TERM

COULD BE PEOPLE, COMPANIES, SALES, TEMPERATURES, BASICALLY ANYTHING...



SAMPLE VS. POPULATION

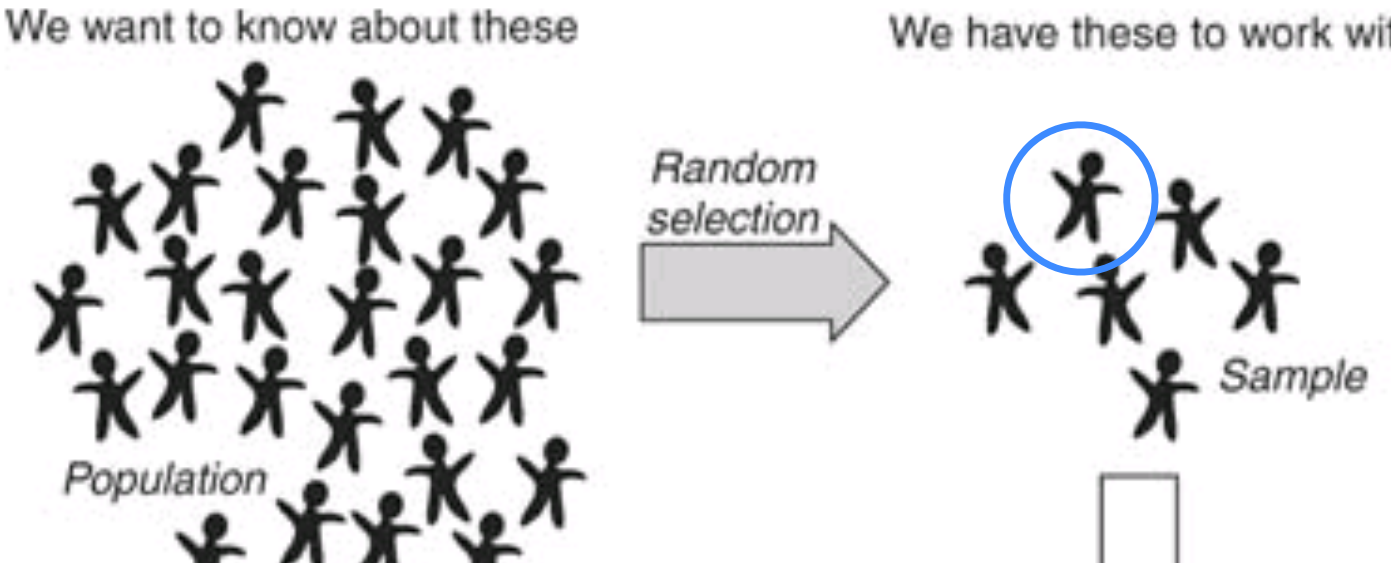
Mean, median, standard deviation, correlation, etc:

- A sample **statistic**
- A population **parameter**

STATISTICAL INFERENCE

Unit of statistical analysis

= **“the thing that I’m sampling from a larger population"**



STATISTICAL INFERENCE

Unit of statistical analysis

Q: WHAT MIGHT OUR UNIT OF ANALYSIS BE HERE?

A: IT DEPENDS!

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

STATISTICAL INFERENCE

Unit of statistical analysis

IF WE WANT TO PREDICT DAILY SALES FOR ONE SELLER, WE MIGHT USE INDIVIDUAL DAYS

day	Seller 1
1	€320
2	€74
3	€340
4	€322
5	€146
6	€24
7	€42
8	€76
9	€99
10	€915

Population =
sales from all days
(including the future)

Sample =
sales from days
in this month

STATISTICAL INFERENCE

Unit of statistical analysis

WHAT IF WE WANT TO UNDERSTAND MONTHLY SALES ACROSS MULTIPLE EMPLOYEES?

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

STATISTICAL INFERENCE

Unit of statistical analysis

Population =
sales from all months
from all employees
(including future ones)

Sample =
monthly sales from
these employees

Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195

STATISTICAL INFERENCE

Unit of statistical analysis

HOW CAN WE TELL HOW ACCURATE OUR SAMPLE STATISTIC MIGHT BE?

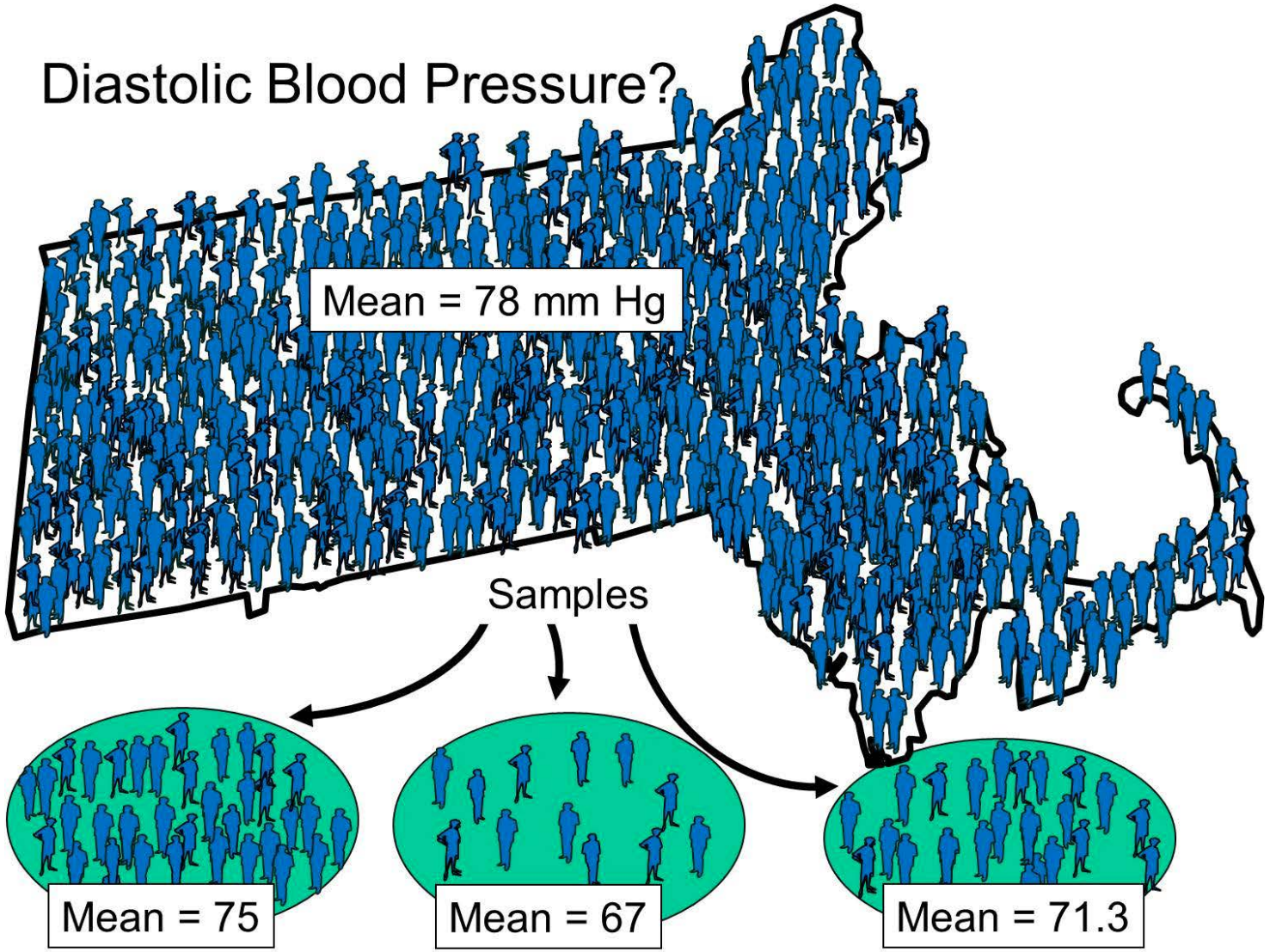
day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

IT HELPS TO THINK ABOUT HOW THE **SAMPLING DISTRIBUTION** RELATES TO THE **POPULATION**

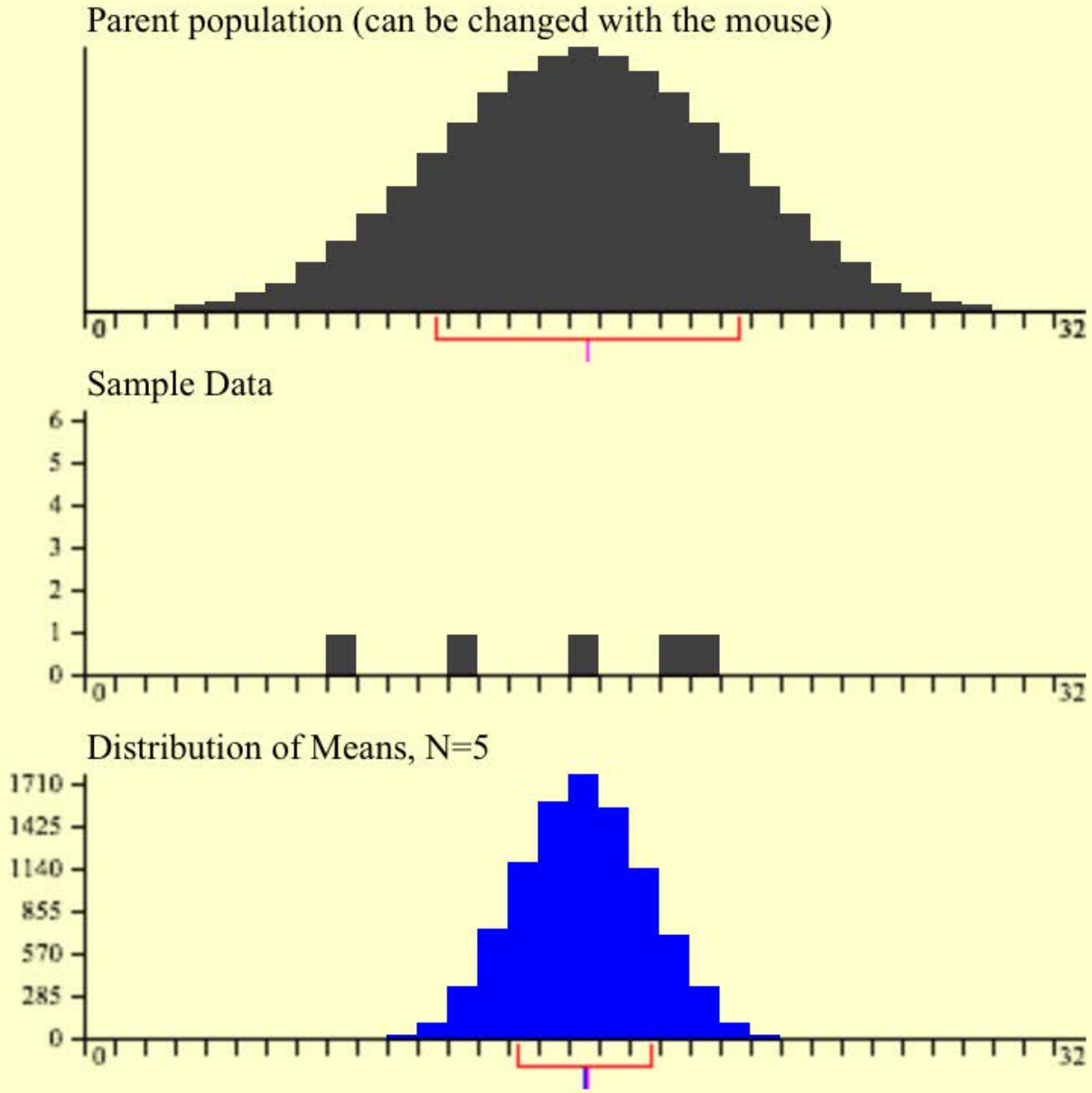
"The **sampling distribution** of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n."

"It may be considered as the distribution of the statistic for all possible samples from the same population of a given size"

SAMPLING DISTRIBUTION



From Lisa Sullivan



A DEMO

https://goo.gl/tjnI9P

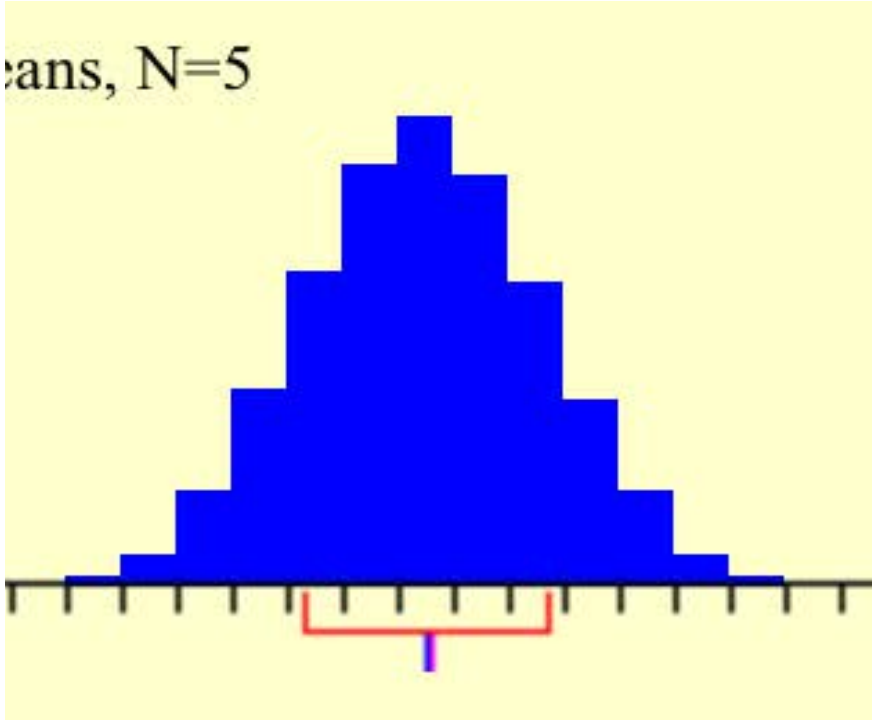
A COUPLE OF LESSONS

1. Statistics computed using **larger samples** will tend to more accurately reflect the real population parameters.

2. We can use the properties of the sampling distribution to reason about **how accurate our statistics might be!**

SAMPLING DISTRIBUTION

TENDS TO BE NORMALLY DISTRIBUTED...

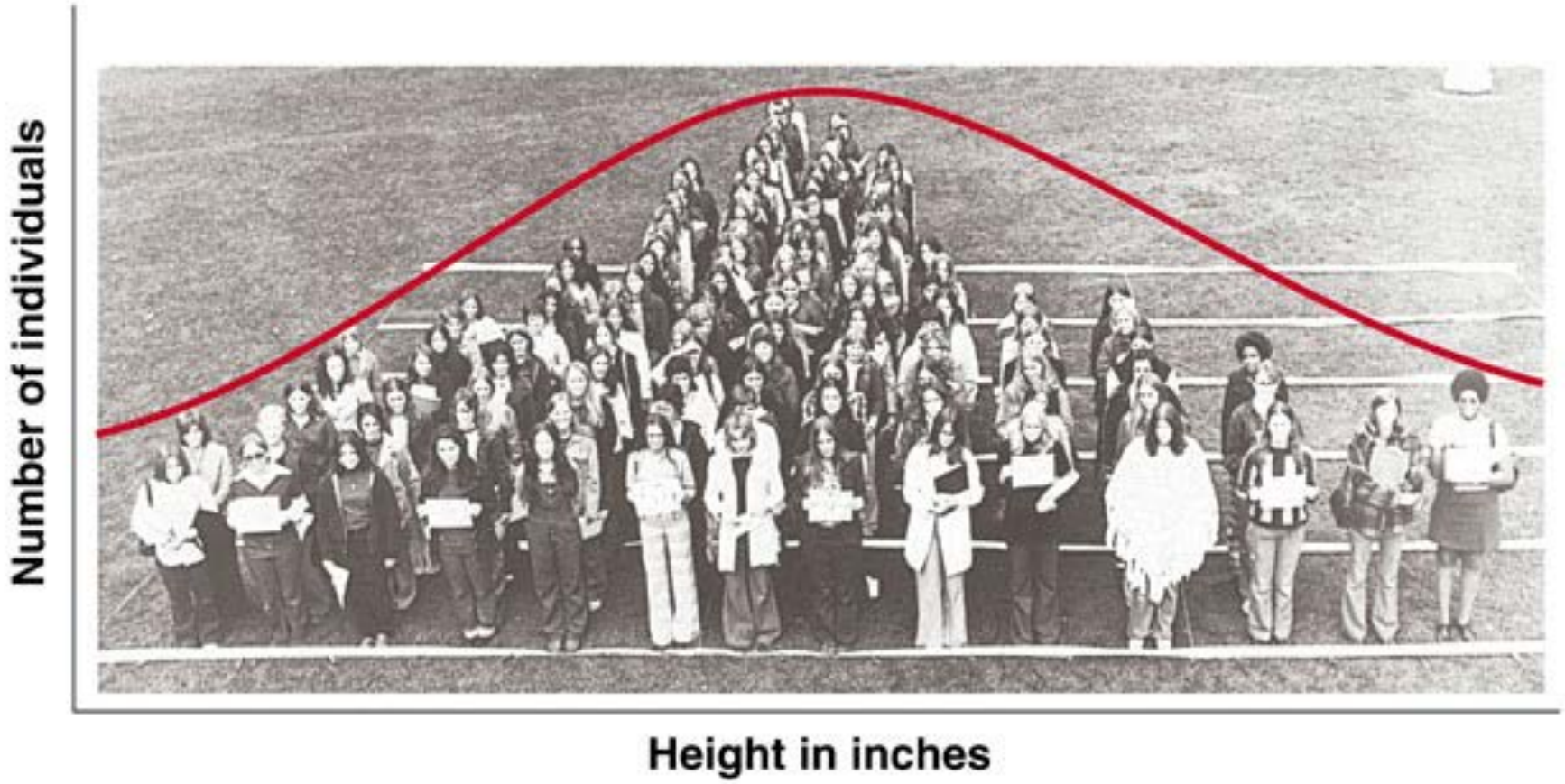


WHY?

Central Limit Theorem

Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed.

NATURAL PHENOMENA TEND TO BE NORMALLY DISTRIBUTED

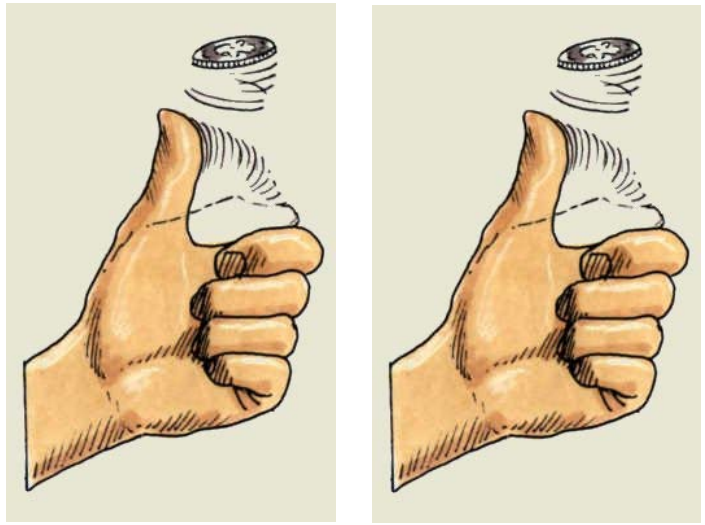


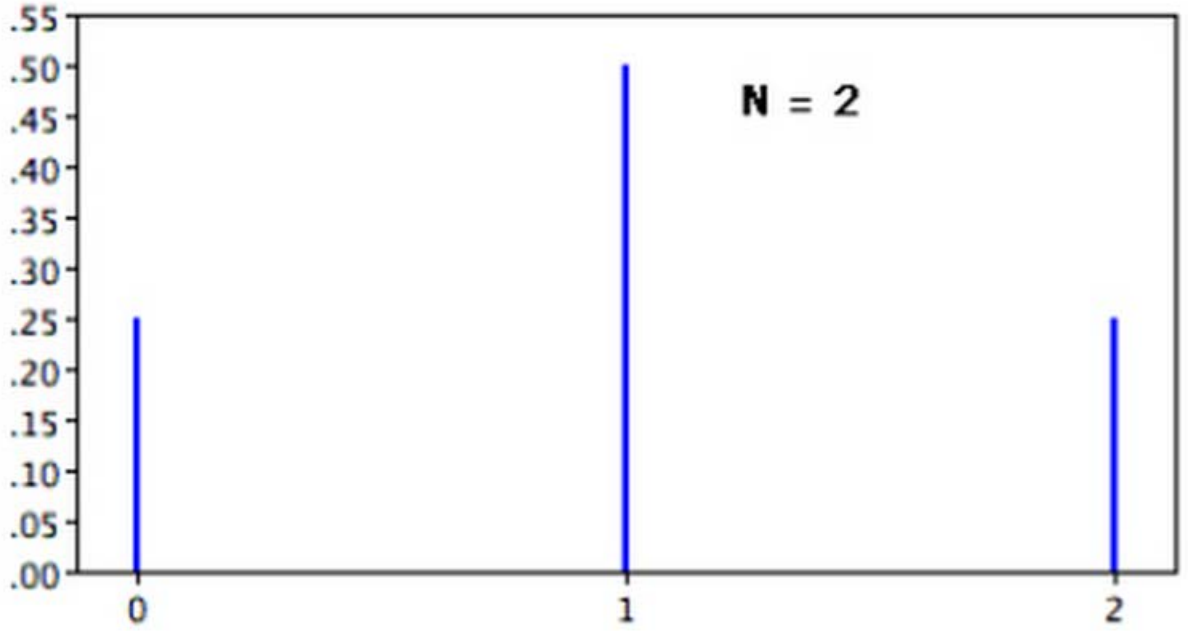
AND SAMPLING FROM A POPULATION TENDS TO ALSO HAVE THIS CHARACTERISTIC (ASSUMING INDEPENDENCE OF SAMPLES)

Abraham De Moivre
1667 - 1754

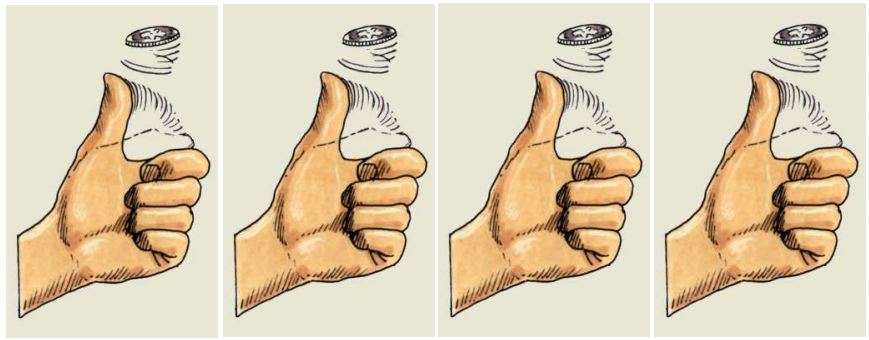
Abraham De Moivre
1667 - 1754

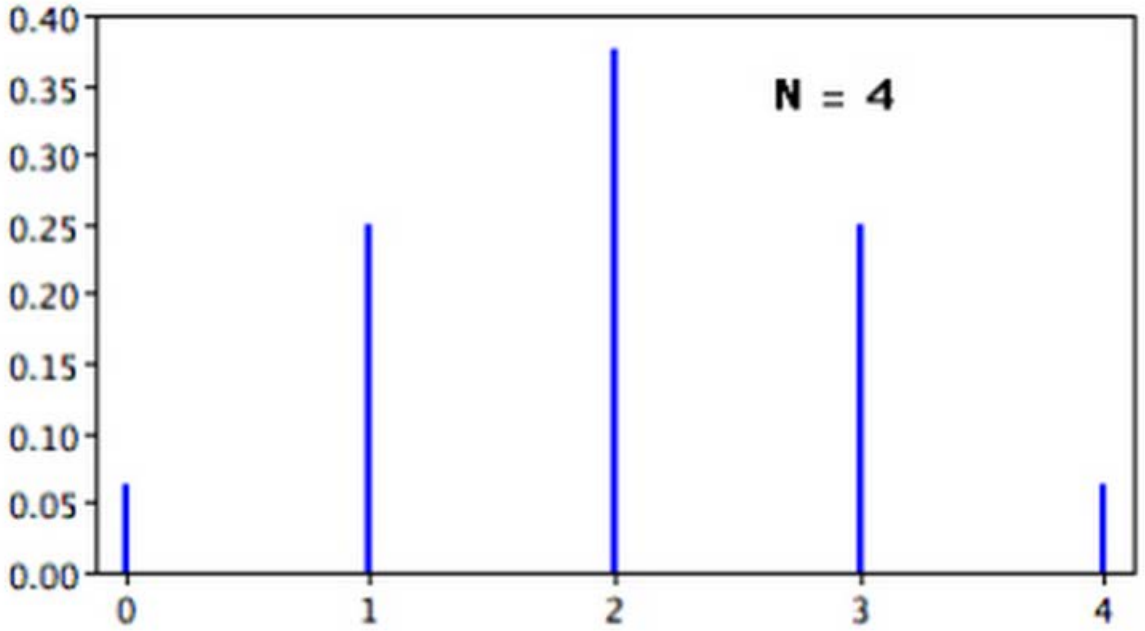
Abraham De Moivre
1667 - 1754



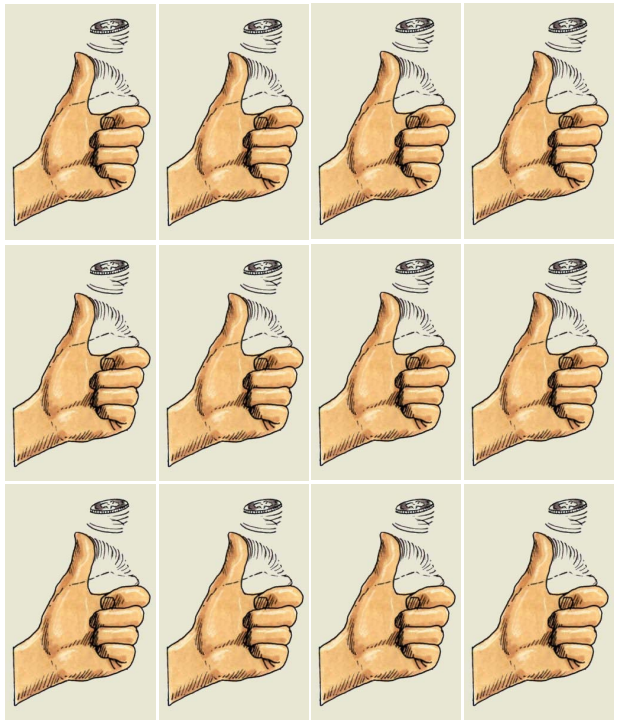


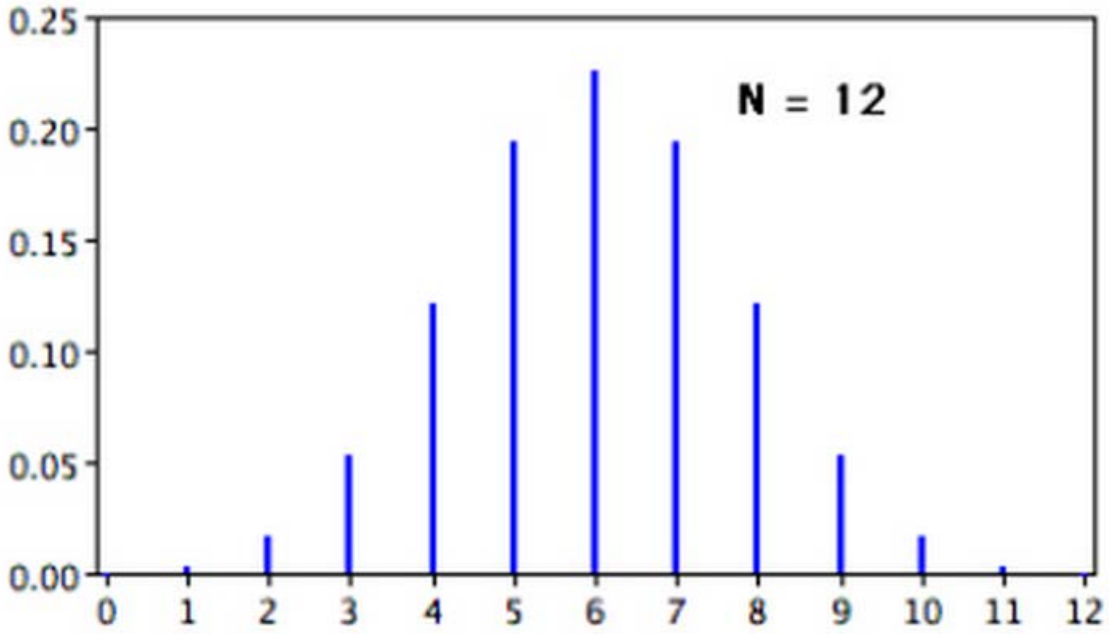
Abraham De Moivre
1667 - 1754





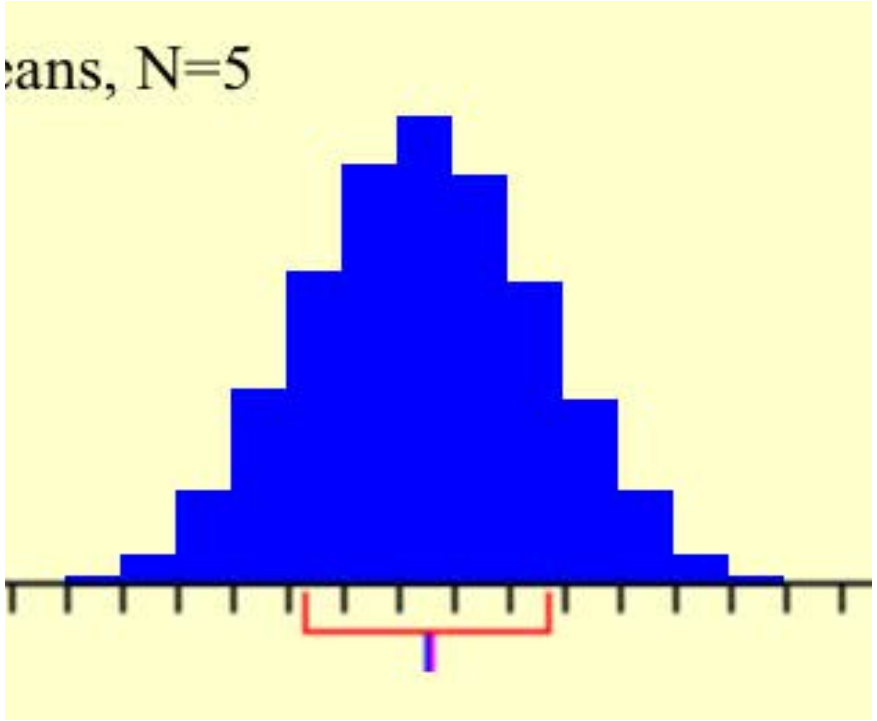
Abraham De Moivre
1667 - 1754





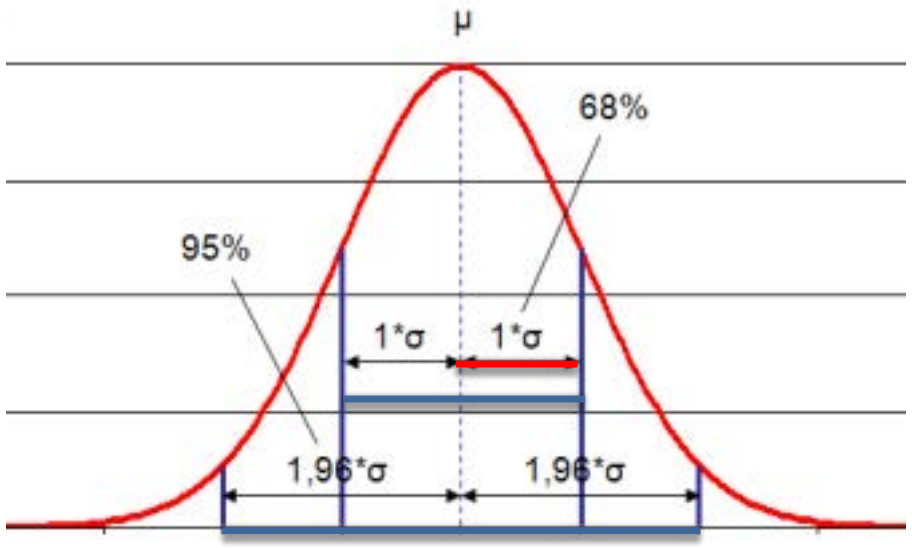
SAMPLING DISTRIBUTION

TENDS TO BE NORMALLY DISTRIBUTED...



SAMPLING DISTRIBUTION

... AND WE CAN APPOXIMATE IT



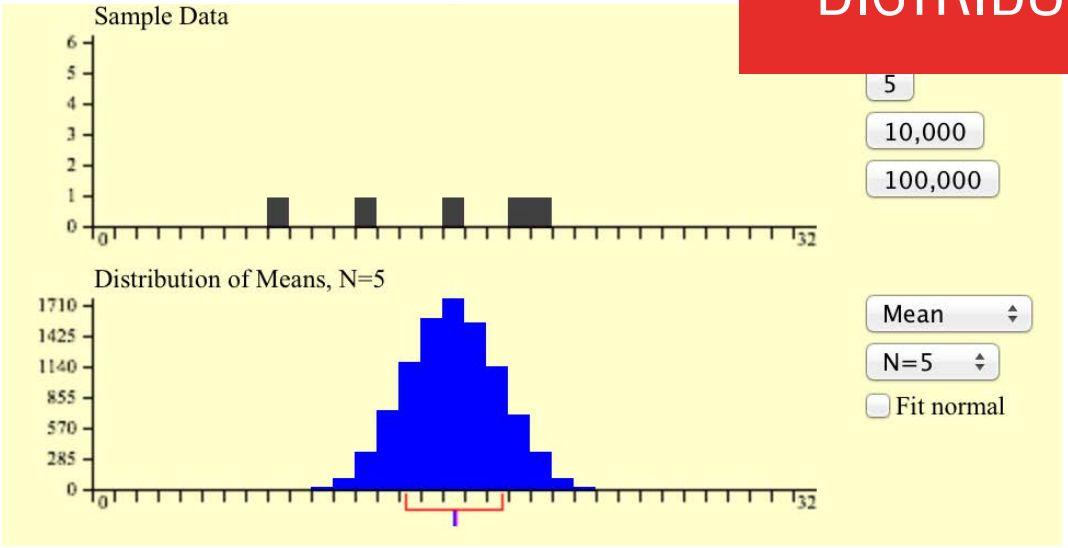
Standard error

95% confidence interval

SAMPLING DISTRIBUTION

BUT WE DON'T HAVE THE ORIGINAL POPULATION

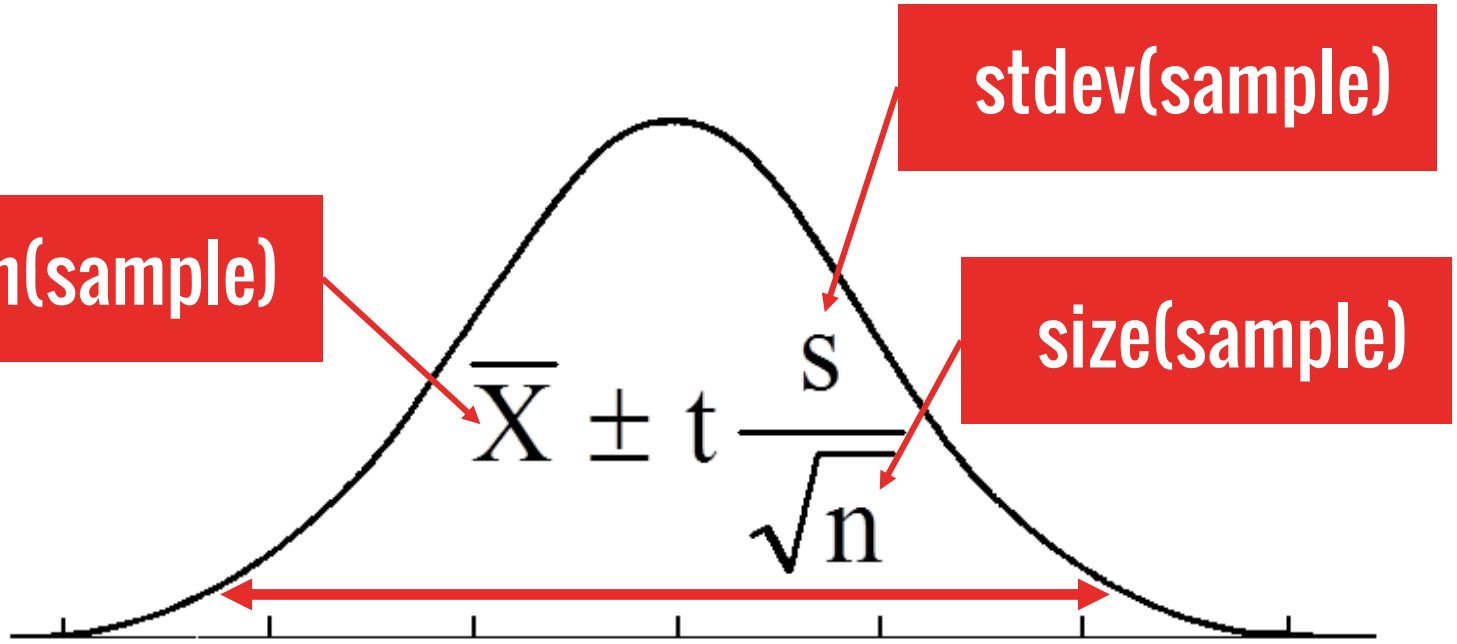
(IN FACT WE DON'T REALLY EVEN KNOW THE SAMPLING DISTRIBUTION)



BUT...

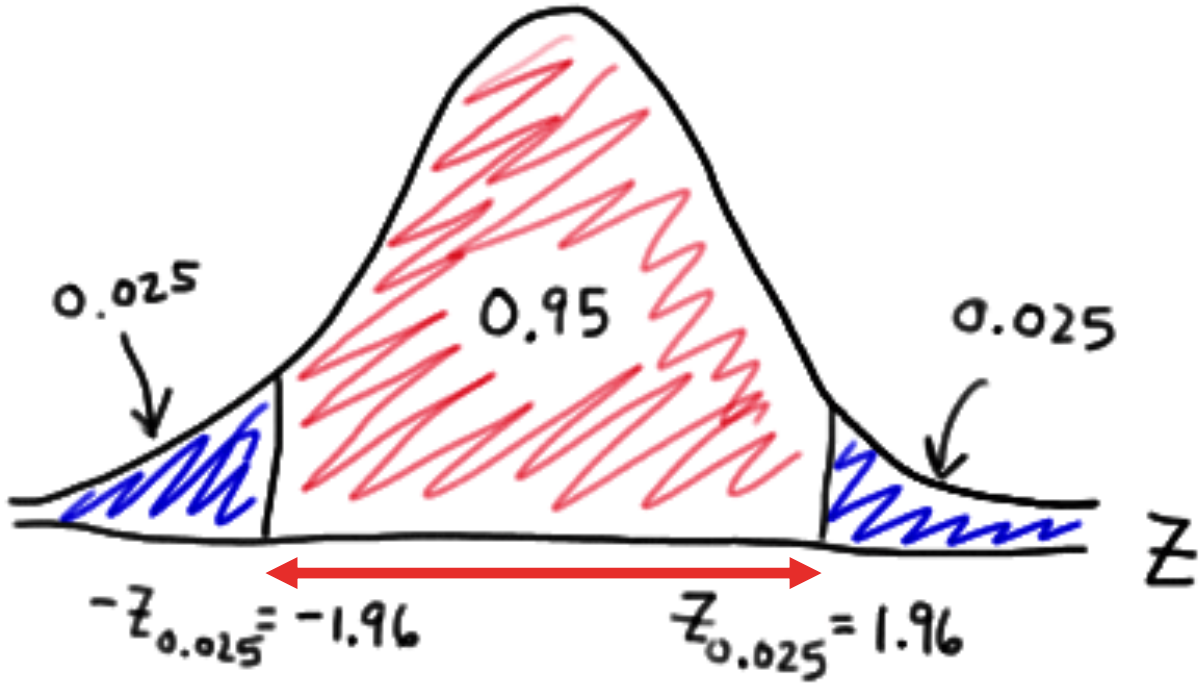
We can make some reasonable assumptions about the sampling distribution, based on the properties of the sample!

(For example,
computing a 95% CI)



t ~ 1.96 for large samples

CONFIDENCE IN OUR STATISTIC



95% of the time, the **population parameter** (the real value) should vary by less than this much.

A "**confidence interval**" for our statistic.

WHAT ARE CONFIDENCE INTERVALS?

"a range of plausible values for x. Values outside the CI are relatively implausible."

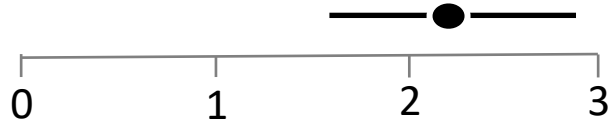
– Cumming and Finch, 2005

Examples of presentation formats:

2.2m, 95% CI [1.6m, 2.8m]

2.2m +/- 0.6m

from 1.6m to 2.8m

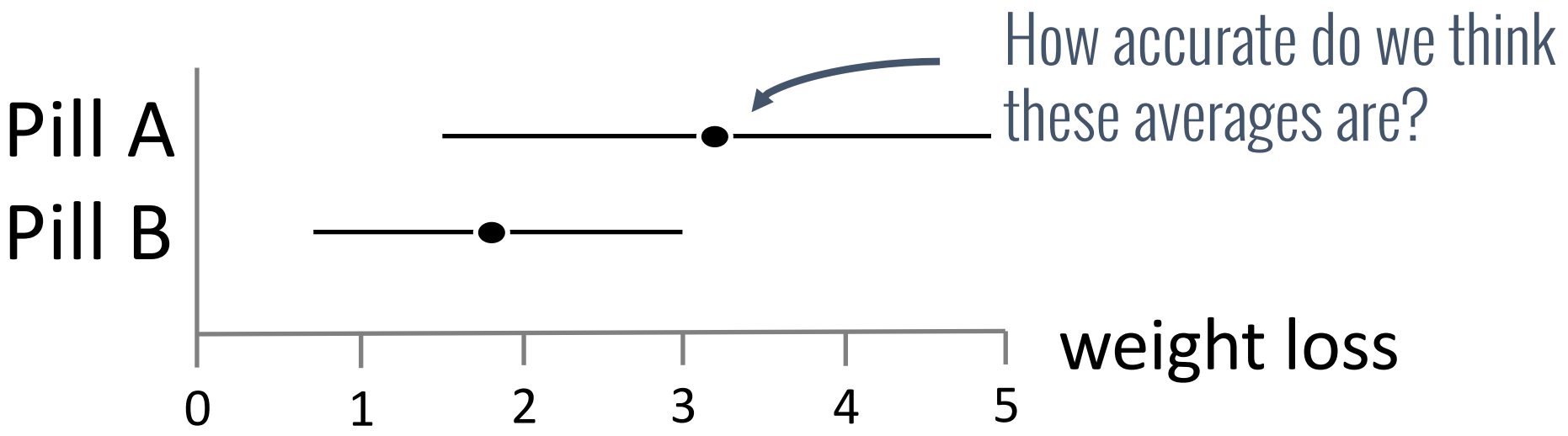


HIGHLIGHTS BOTH SIZE OF AN EFFECT AND HOW PLAUSIBLE IT IS.

INTERPRETING CONFIDENCE INTERVALS

"A range of plausible values for x. Values outside the CI are relatively implausible."

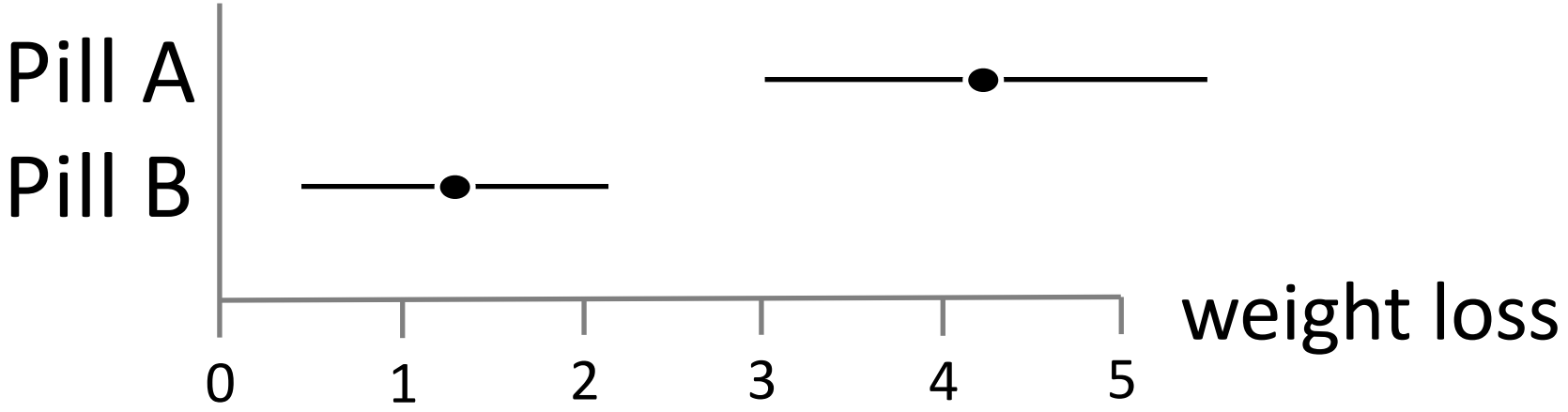
– Cumming and Finch, 2005



INTERPRETING CONFIDENCE INTERVALS

"A range of plausible values for x. Values outside the CI are relatively implausible."

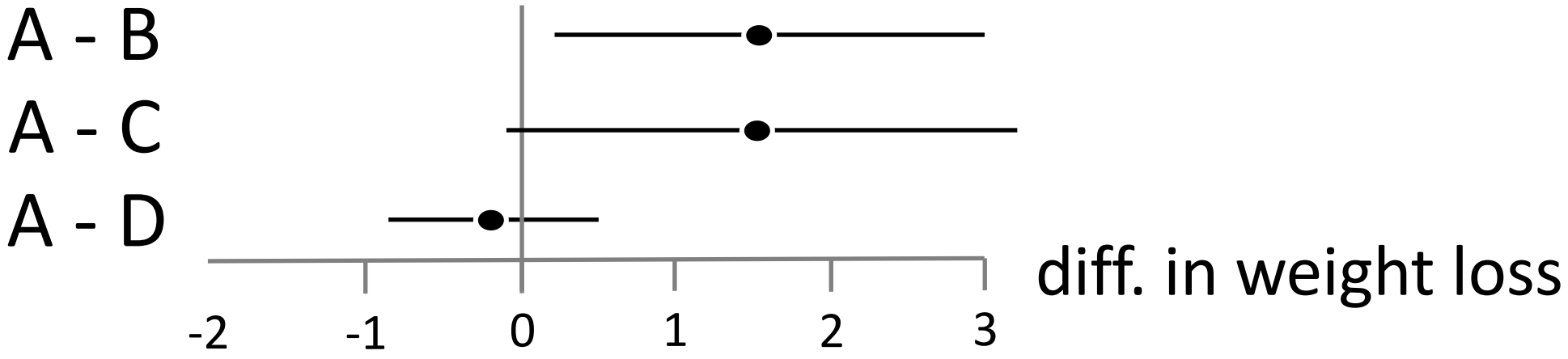
– Cumming and Finch, 2005



INTERPRETING CONFIDENCE INTERVALS

"A range of plausible values for x. Values outside the CI are relatively implausible."

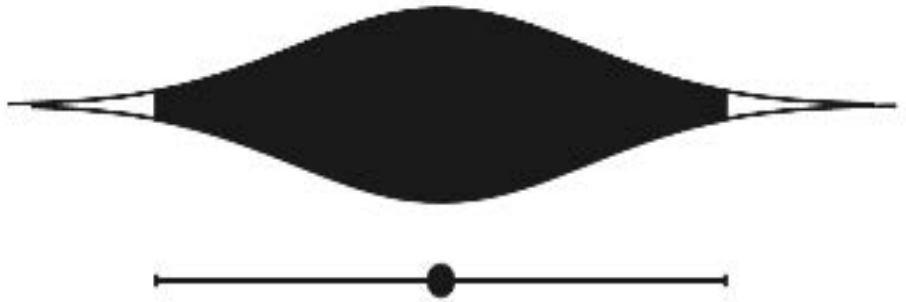
– Cumming and Finch, 2005



INTERPRETING CONFIDENCE INTERVALS

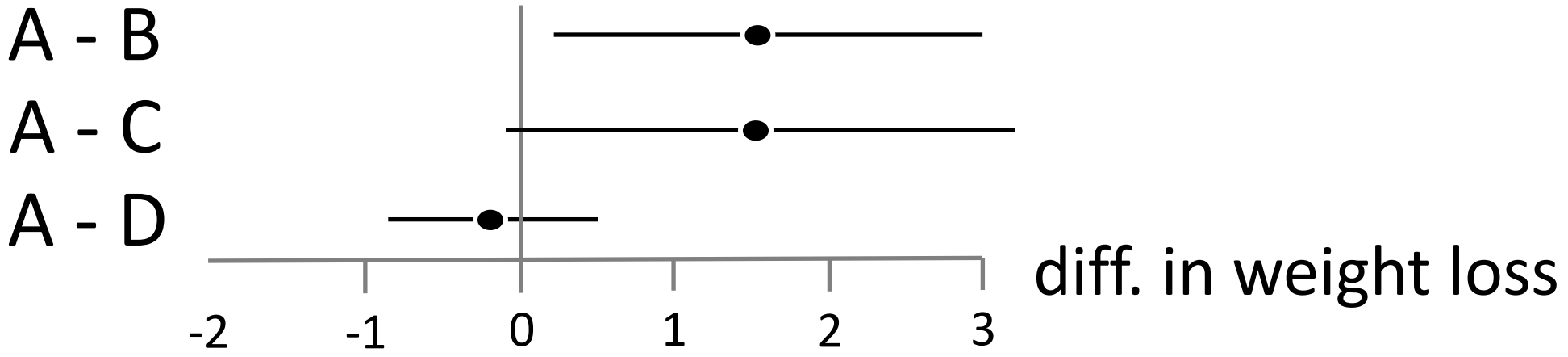
"values close to our M are the best bet for X, and values closer to the limits of our CI are successively less good bets."

–Cumming, 2013

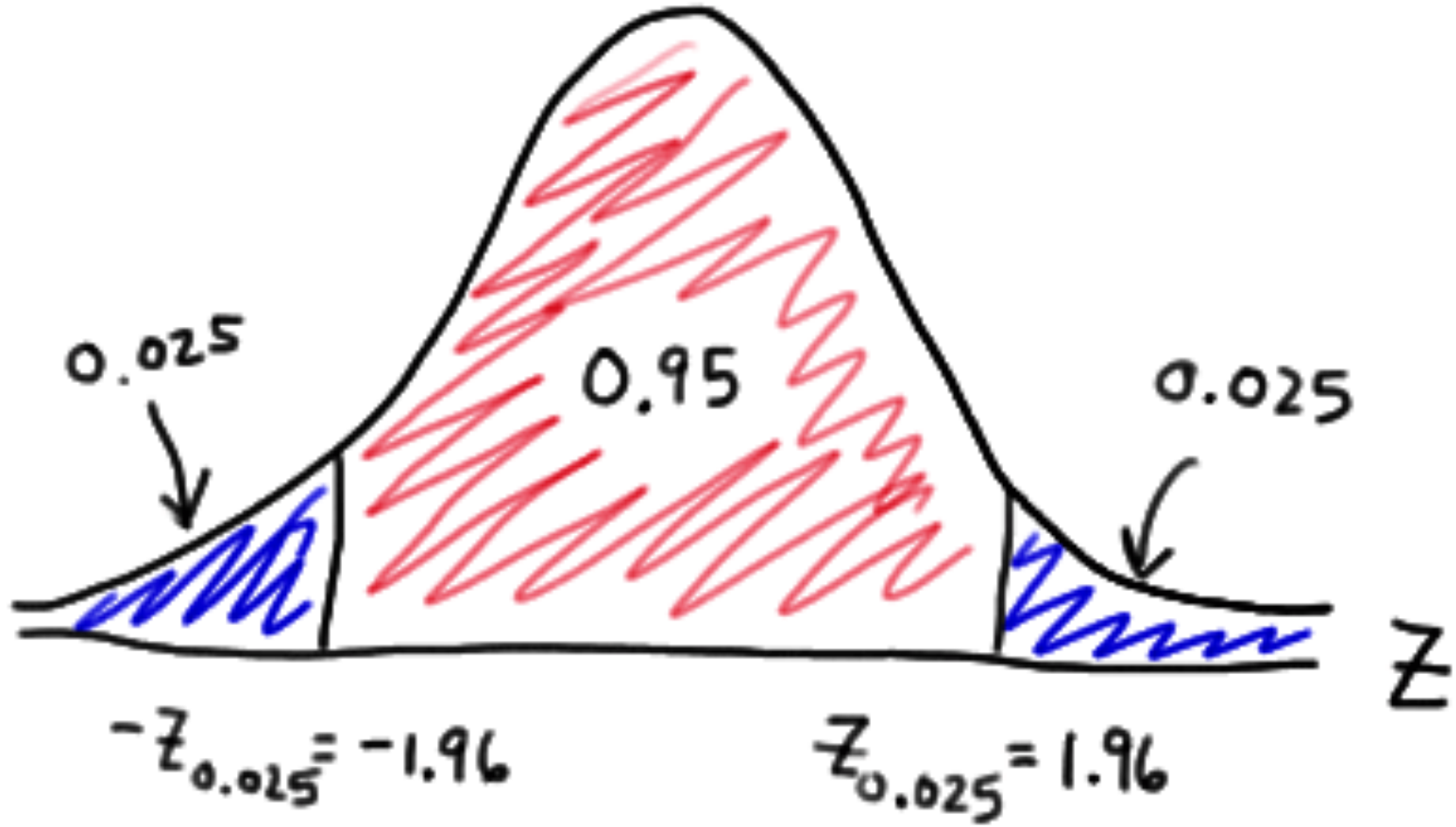


INTERPRETING CONFIDENCE INTERVALS

BE CAREFUL ABOUT THE CONCLUSIONS YOU DRAW!



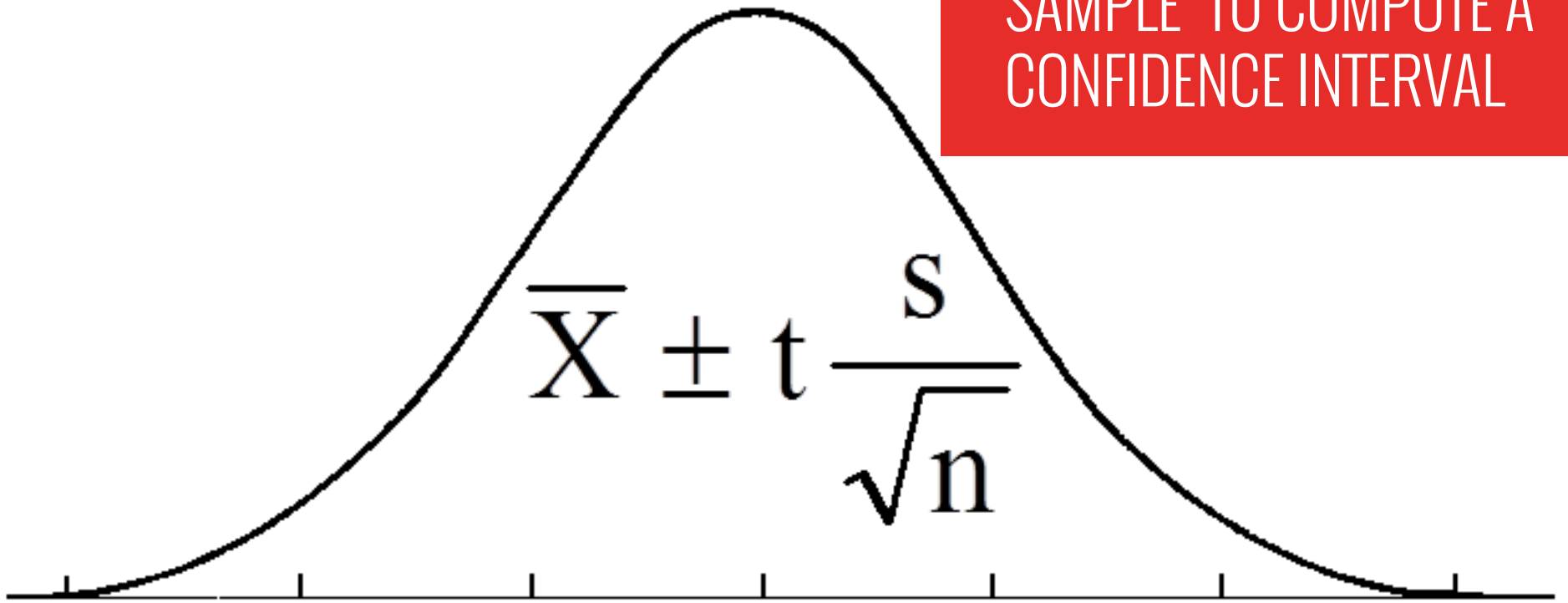
"the figure provides good evidence that B outperforms A, whereas C and A seem very similar, and results are largely inconclusive concerning the difference between D and A."


0.025
0.95
0.025
Z
$-Z_{0.025} = -1.96$
$Z_{0.025} = 1.96$

CI METHOD #1

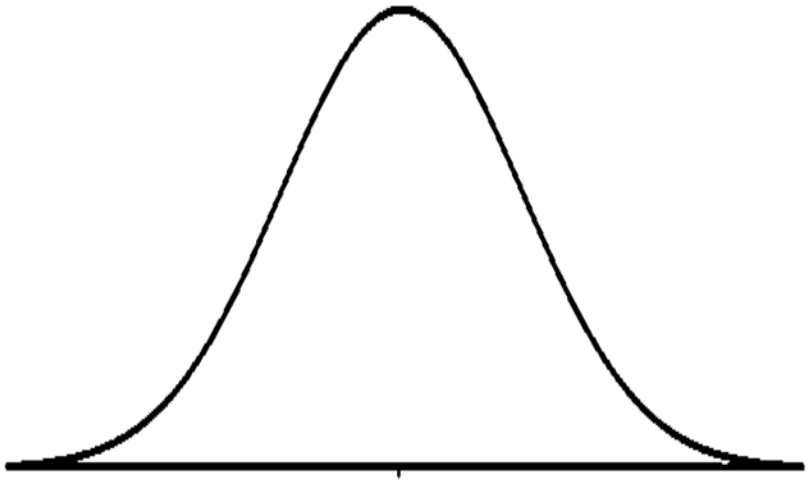
"Exact" Confidence Intervals

USE THE **MEAN**, **STANDARD DEVIATION**, & **SIZE** OF THE SAMPLE TO COMPUTE A CONFIDENCE INTERVAL



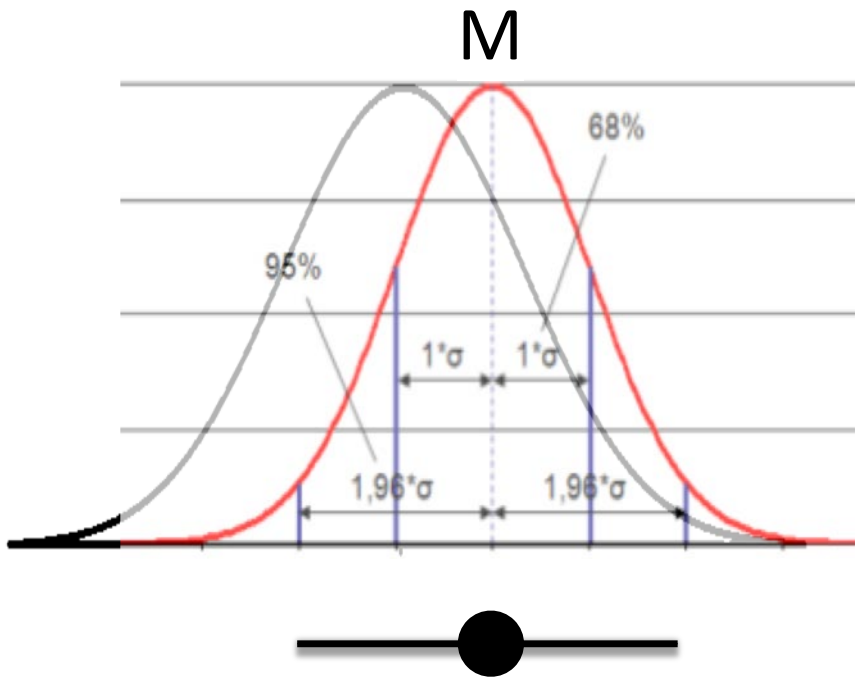
t ~ 1.96 for large samples

CONFIDENCE INTERVALS



CONFIDENCE INTERVALS

TECHNICALLY, WE CAN ONLY TALK ABOUT THE **ACCURACY** OF THE CI, NOT THE **ACCURACY** OF THE MEAN

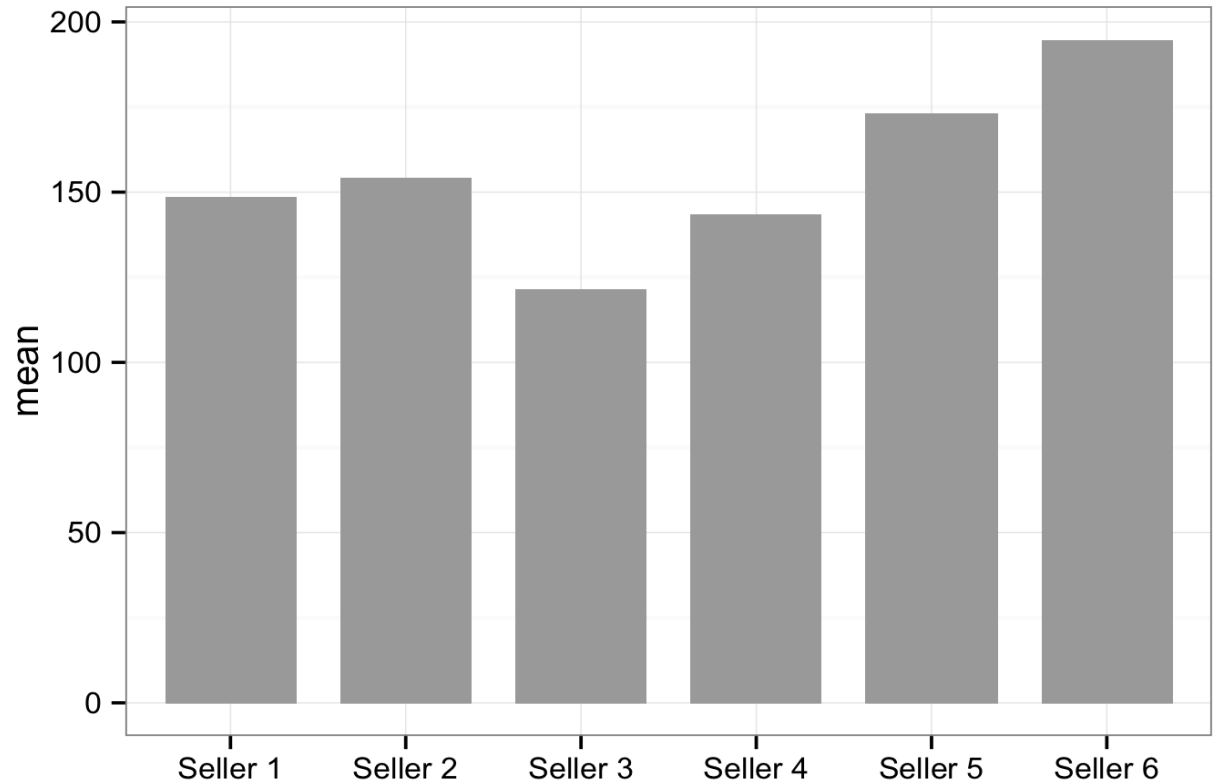


95% confidence interval

BACK TO OUR FIRST EXAMPLE

Average Sales

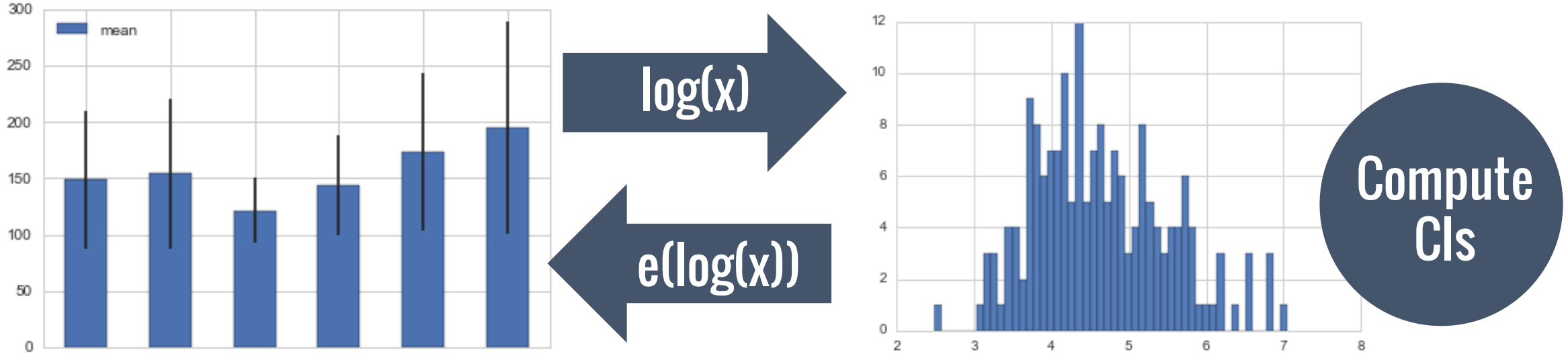
Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195



LET'S TRY IT IN TABLEAU

CI METHOD #1.5

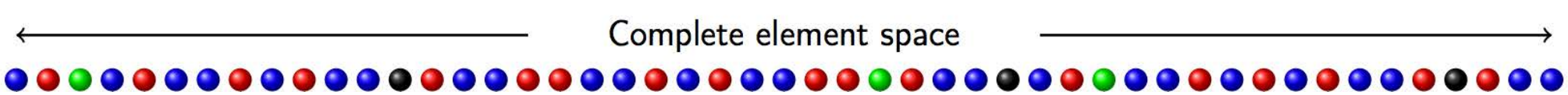
If our data is **not normally distributed**, we can sometimes transform it, compute CIs and then transform back to improve results.



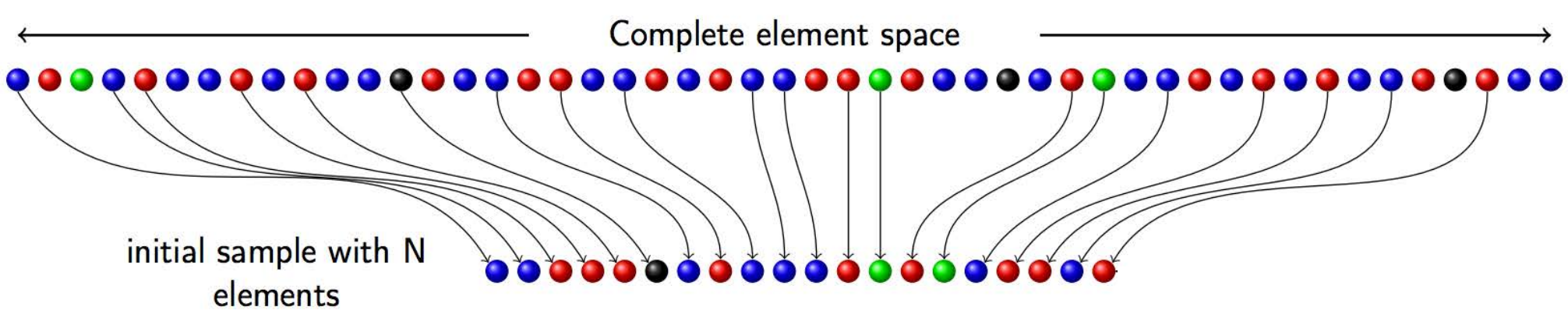
CI METHOD #2

“Bootstrapping”

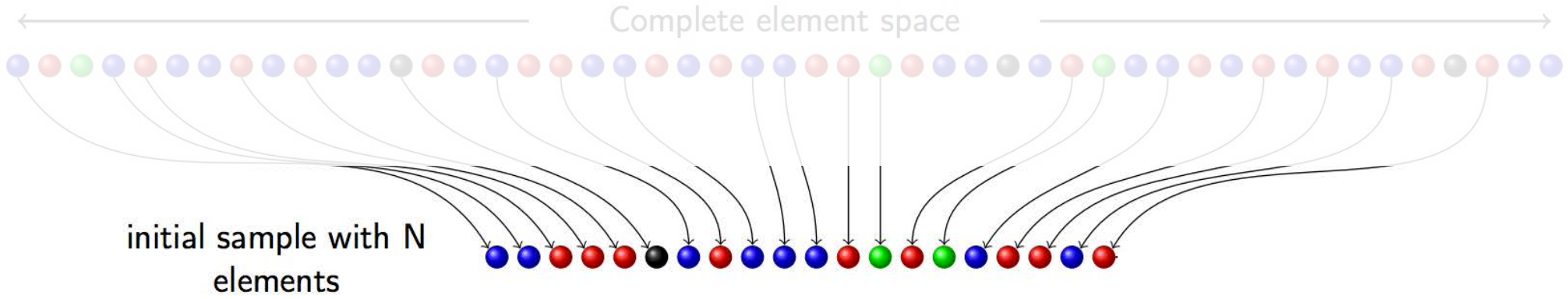
Re-sample repeatedly from our existing sample to simulate repeated samples from the population.



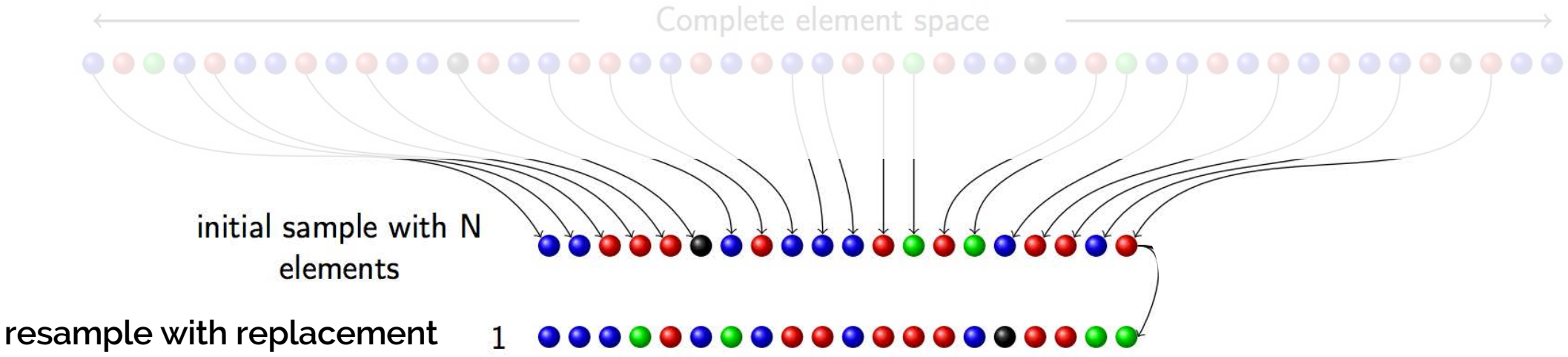
From Germain Salvato-Vallverdu



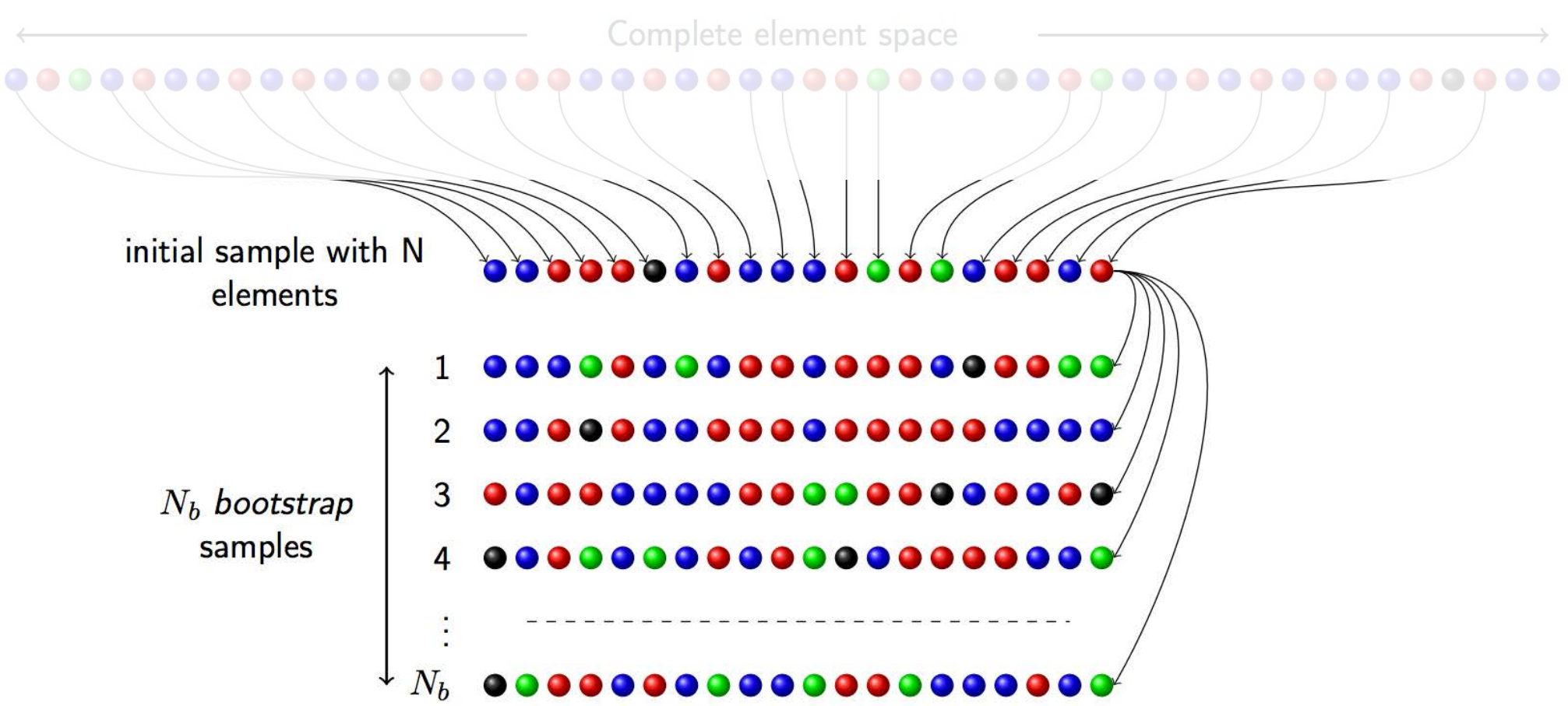
From Germain Salvato-Vallverdu



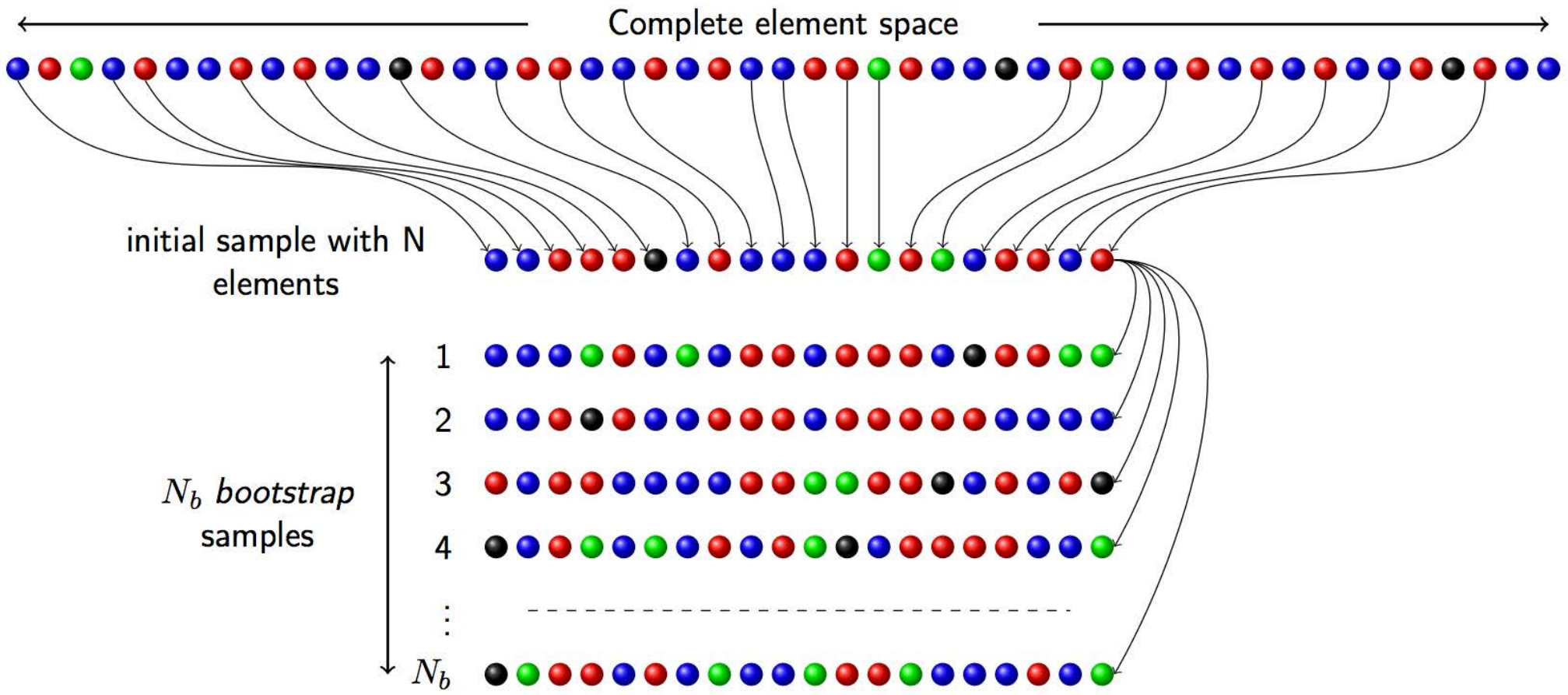
From Germain Salvato-Vallverdu



From Germain Salvato-Vallverdu



From Germain Salvato-Vallverdu

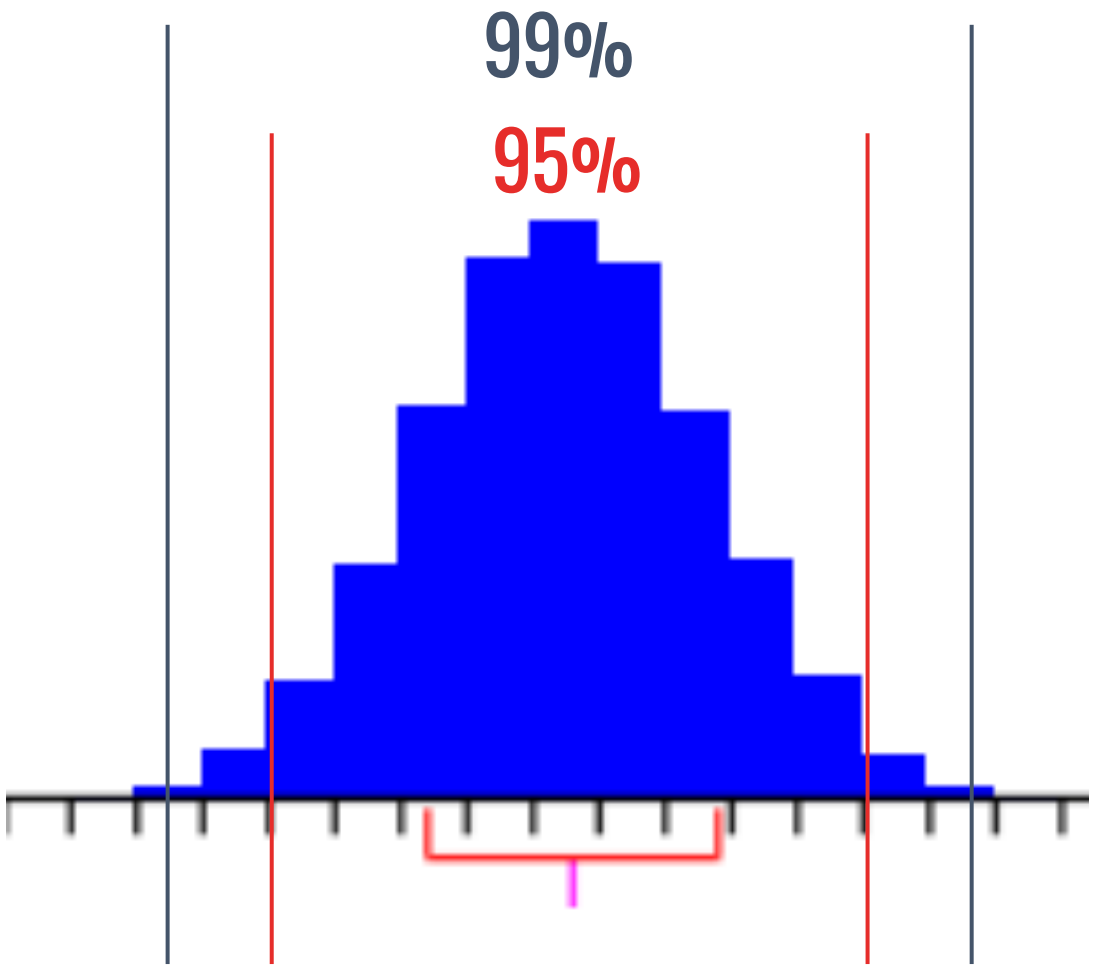


Theorem (B. Efron, Ann. Statist. 1979)

When N tend to infinity, the distribution of average values computed from bootstrap samples is equal to the distribution of average values obtained from ALL samples with N elements which can be constructed from the complete space. Thus the width of the distribution gives an evaluation of the sample quality.

From Germain Salvato-Vallverdu

Simulate the sampling distribution, then just pick the cutoffs we want!



Calculating Confidence Intervals

Informal

Traditional Normal-based

Bootstrapping

CIs IN JUPYTER

BOOTSTRAPPING ALTERNATIVES IN PYTHON

Seaborn – Great for plotting.
Hard to get actual CI values (need to use seaborn.utils.ci)

Bootstrapping Libraries – Varying levels of polish / Pandas compatibility

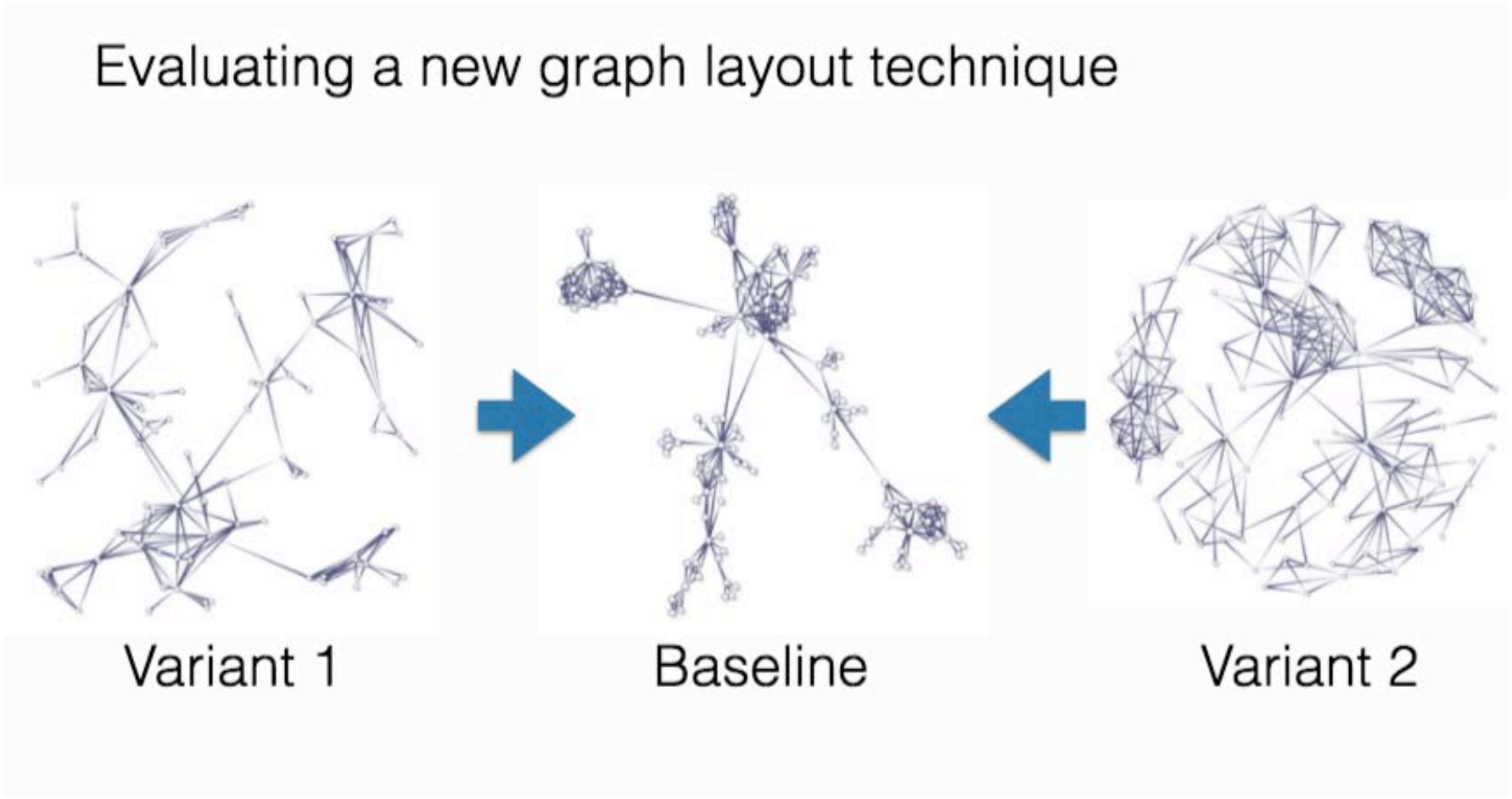
- scikits.bootstrap
- bootstrapped
- statsmodels

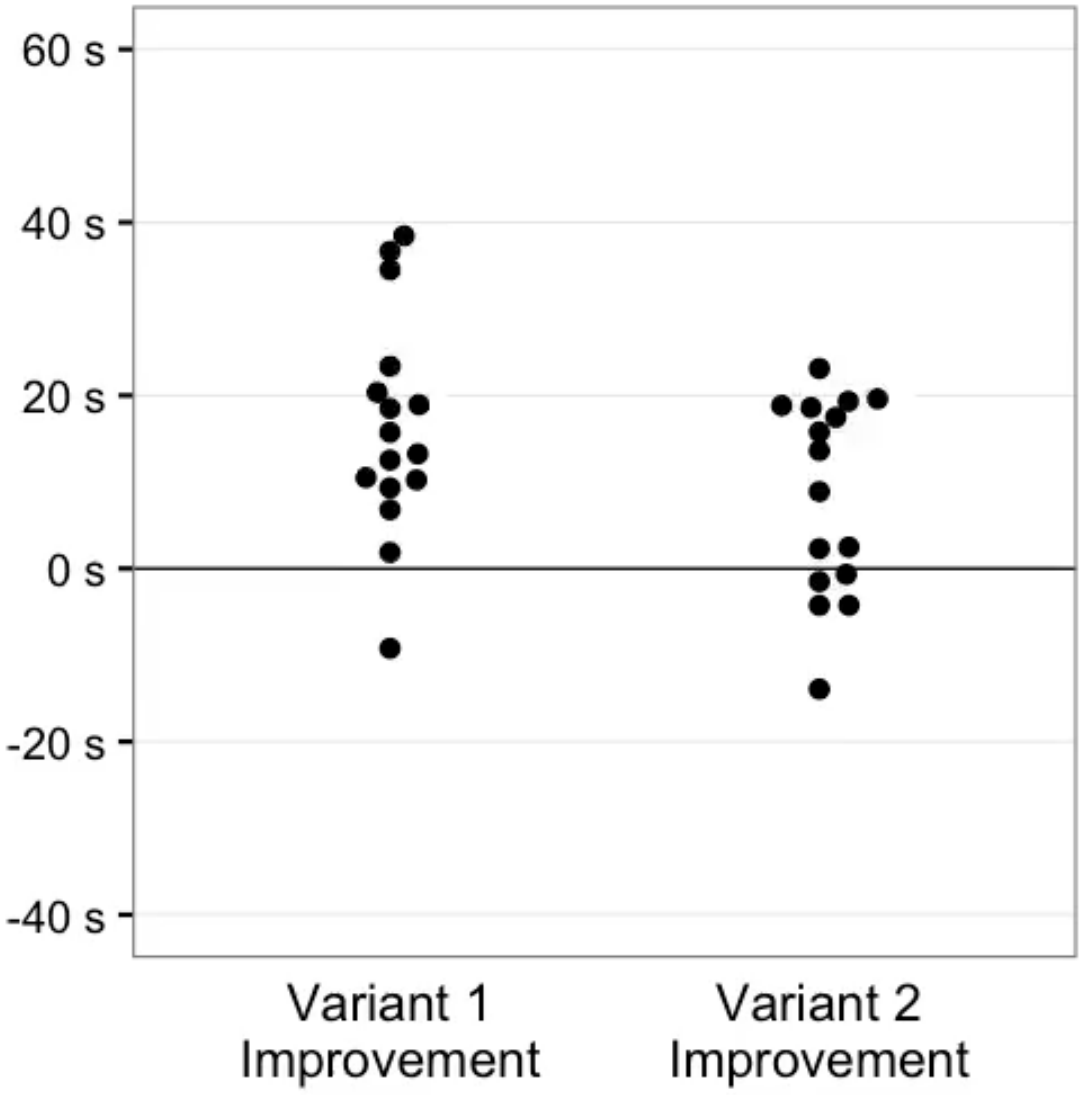
D-I-Y – Bootstrapping is conceptually simple. You can roll your own!

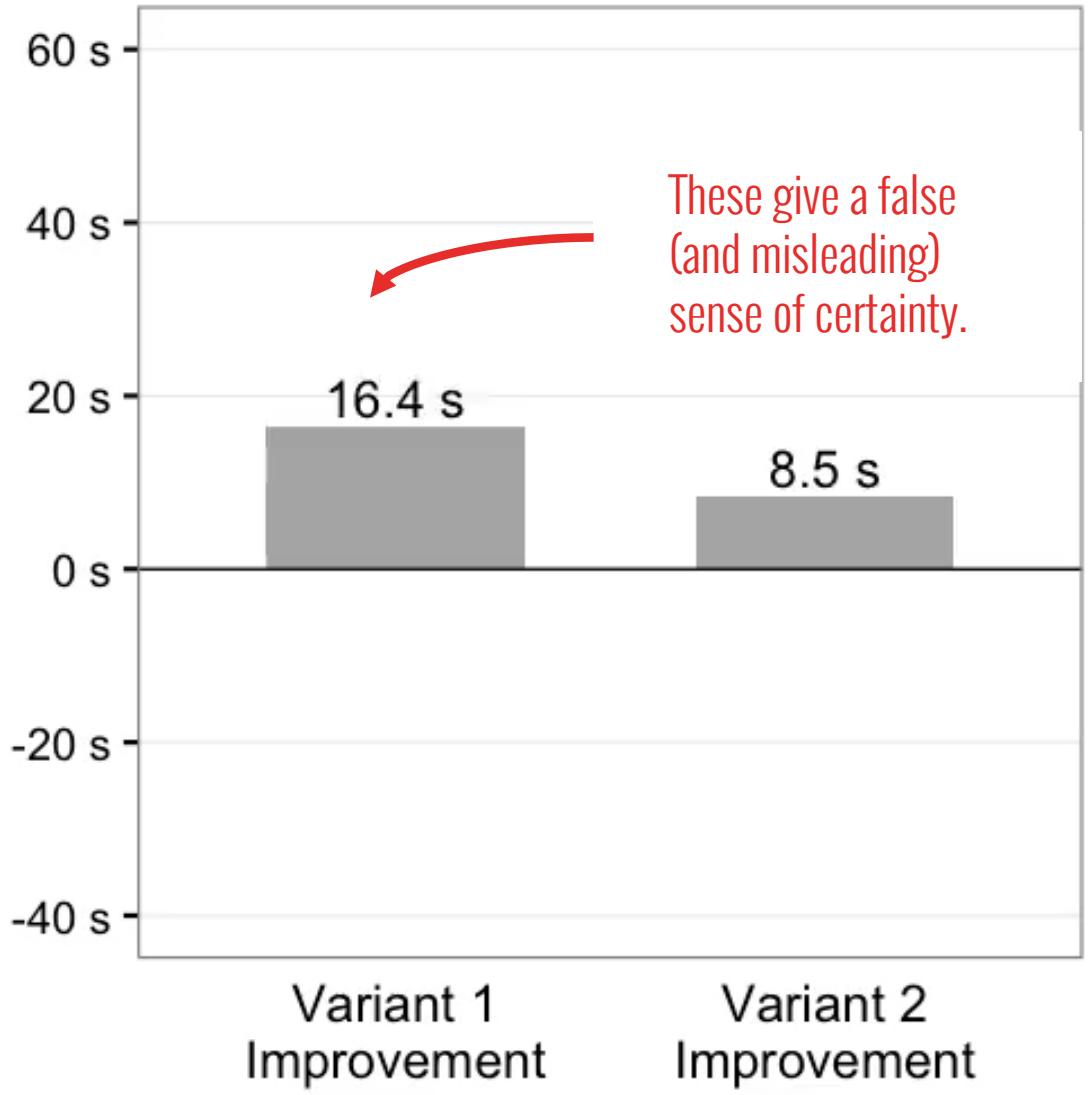
ONE MORE THING

Interpreting confidence intervals and other statistical tests.

IMAGINE YOU'RE EVALUATING A NEW SYSTEM...

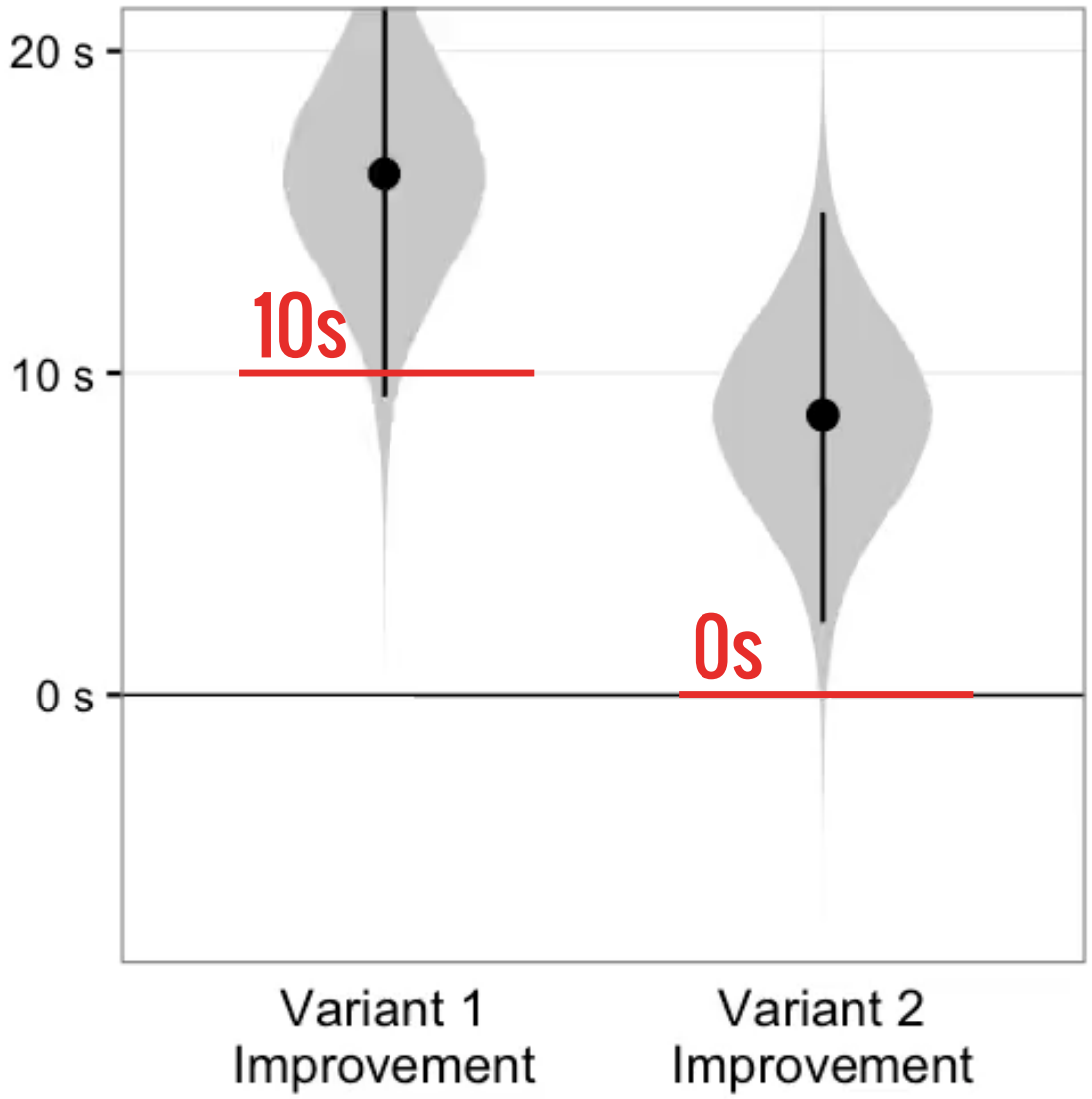






DON'T BLINDLY TRUST SAMPLE STATISTICS!

So what if we use **confidence intervals**?



CONFIDENCE INTERVALS AREN'T FOOLPROOF (BUT THEY ARE BETTER)

STATISTICAL SIGNIFICANCE

effect of METHOD ($F_{4,44} = 10.1, p < 0.0001$
$1\ F_{3,33} = 49.1, p < 0.0001$) for both datasets
4) and a significant effect of SCALE for the data
ıt not for SCALE≥ 4 ($F_{2,22} = 2.7, p = 0.0885$).
$= 0.1116$ and $F_{1,11} = 3.9, p = 0.0718$).
ractions of METHOD $\times$ W ($F_{12,132} = 6.1, p <$
$> < 0.0001$ and $F_{6,66} = 10.6, p < 0.0001$) for
SCALE $= 1$ in particular, we have a higher error
his difference vanishes as W increases. The
maining

HOW MANY OF YOU REMEMBER / HAVE USED SIGNIFICANCE TESTS?

Student's T-test, Mann-Whitney U-test, Chi-squared test, etc.

"NULL-HYPOTHESIS SIGNFICANCE TESTING"

Dichotomizes the distinction of whether or not a difference is due to chance.

Results are either "significant" or not.

AN EXAMPLE HYPOTHESIS (H_1)

"Variant 1 will perform better than the baseline"

THE NULL HYPOTHESIS (H_0)

"There is no effect (no difference)."

```
import scipy.stats as stats

stats.ttest_ind(a= variant_one,
                b= baseline)
```

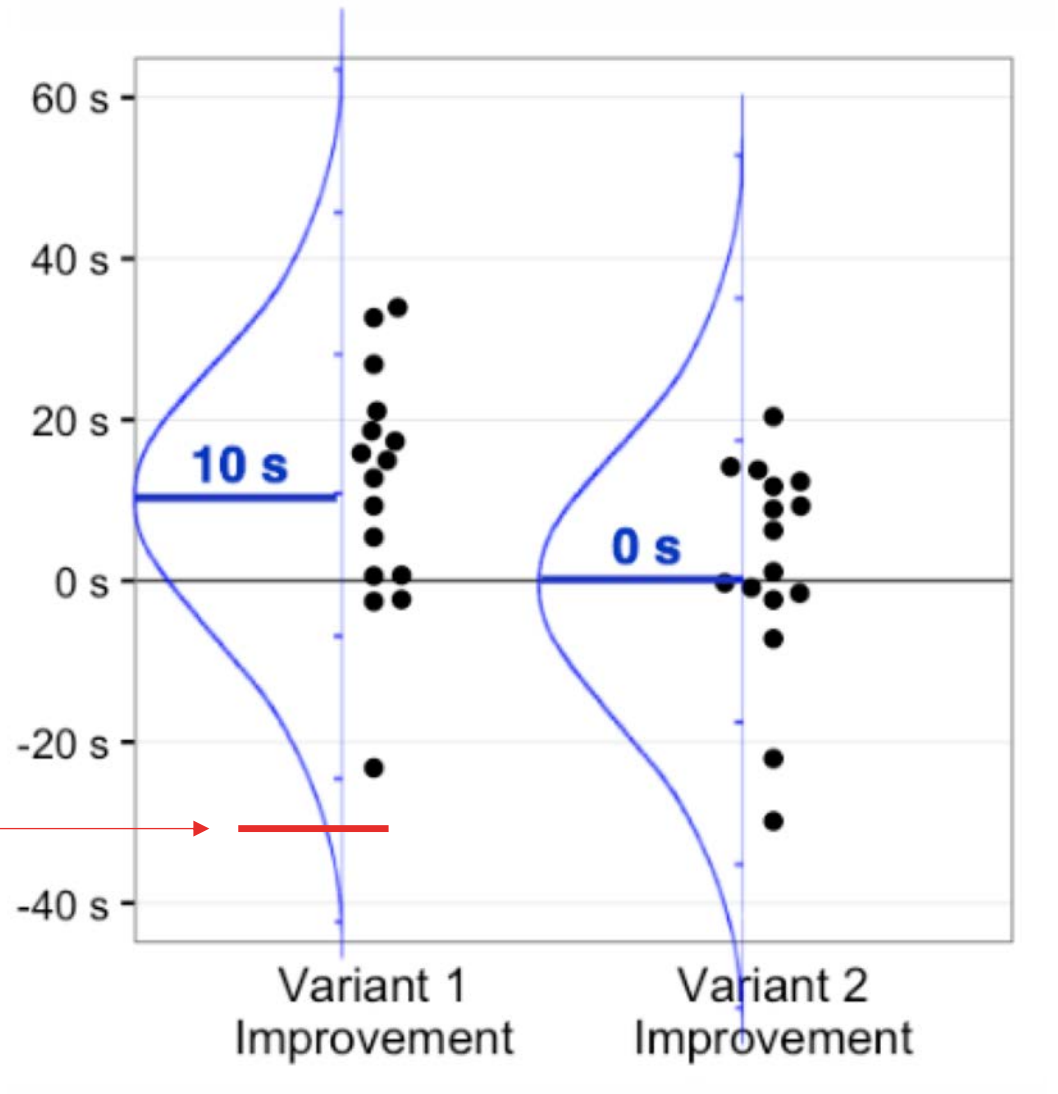
```
Ttest_indResult(statistic==9.7892346273,
                pvalue=0.090731043439577483)
```

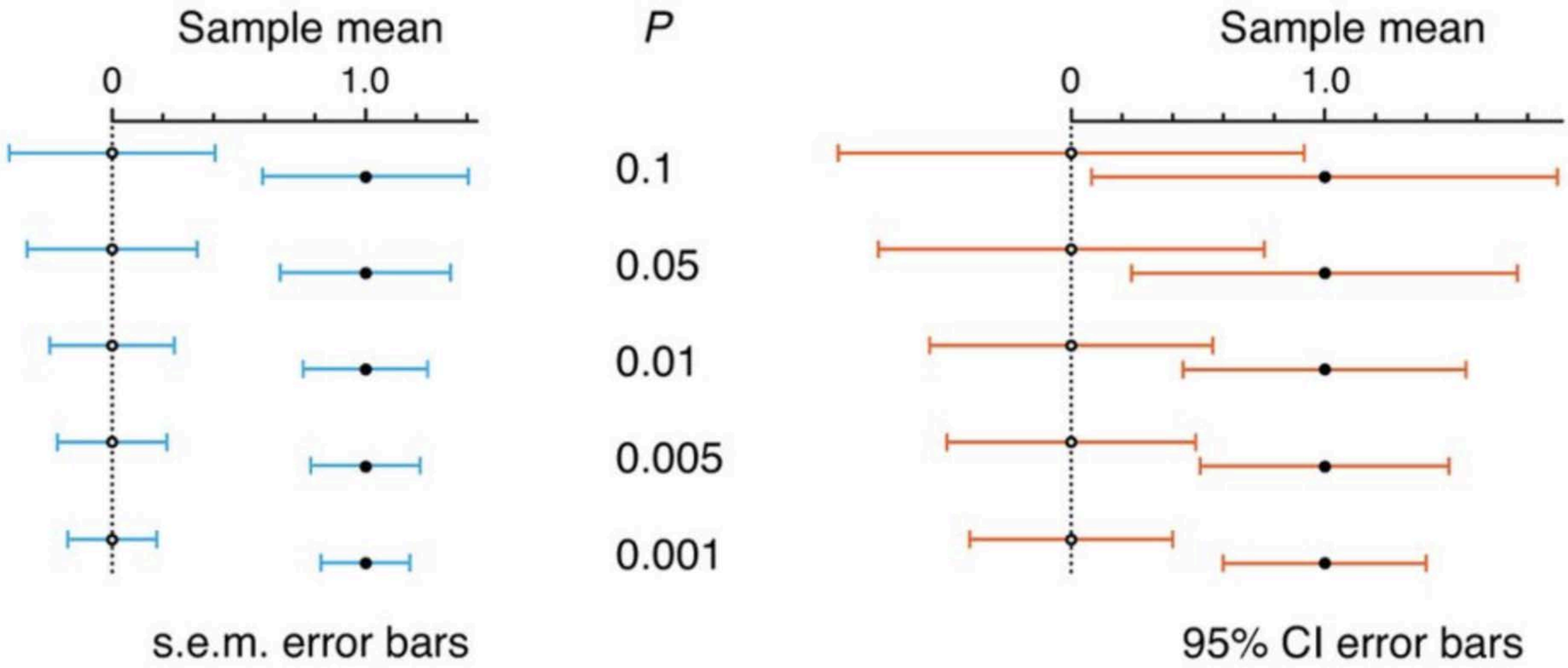
Test statistic – how different the variant one is from the baseline.

P-value – likelihood the difference is due to chance.

Usually we would reject H_0 any time $P<0.05$

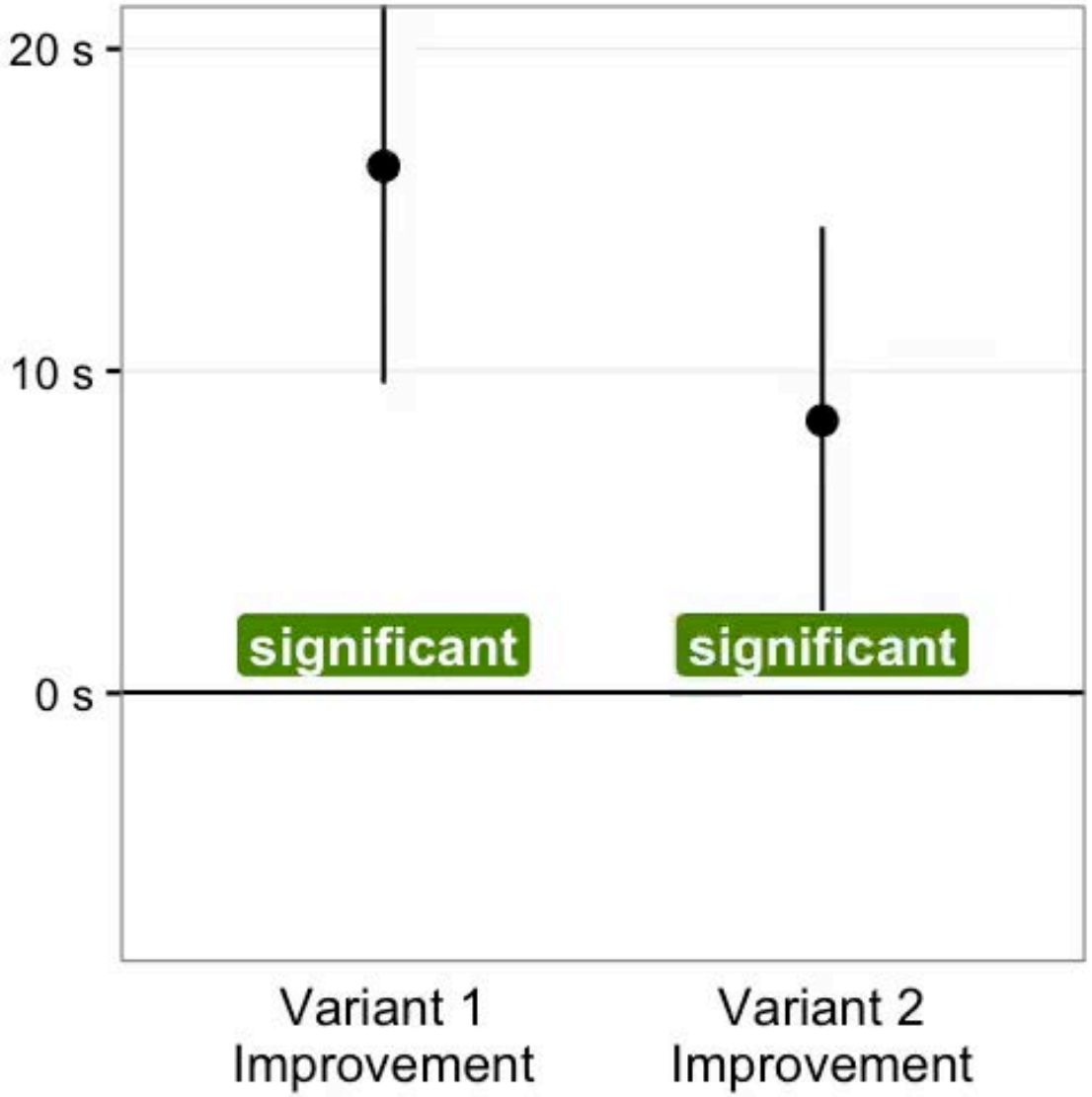
p=0.05

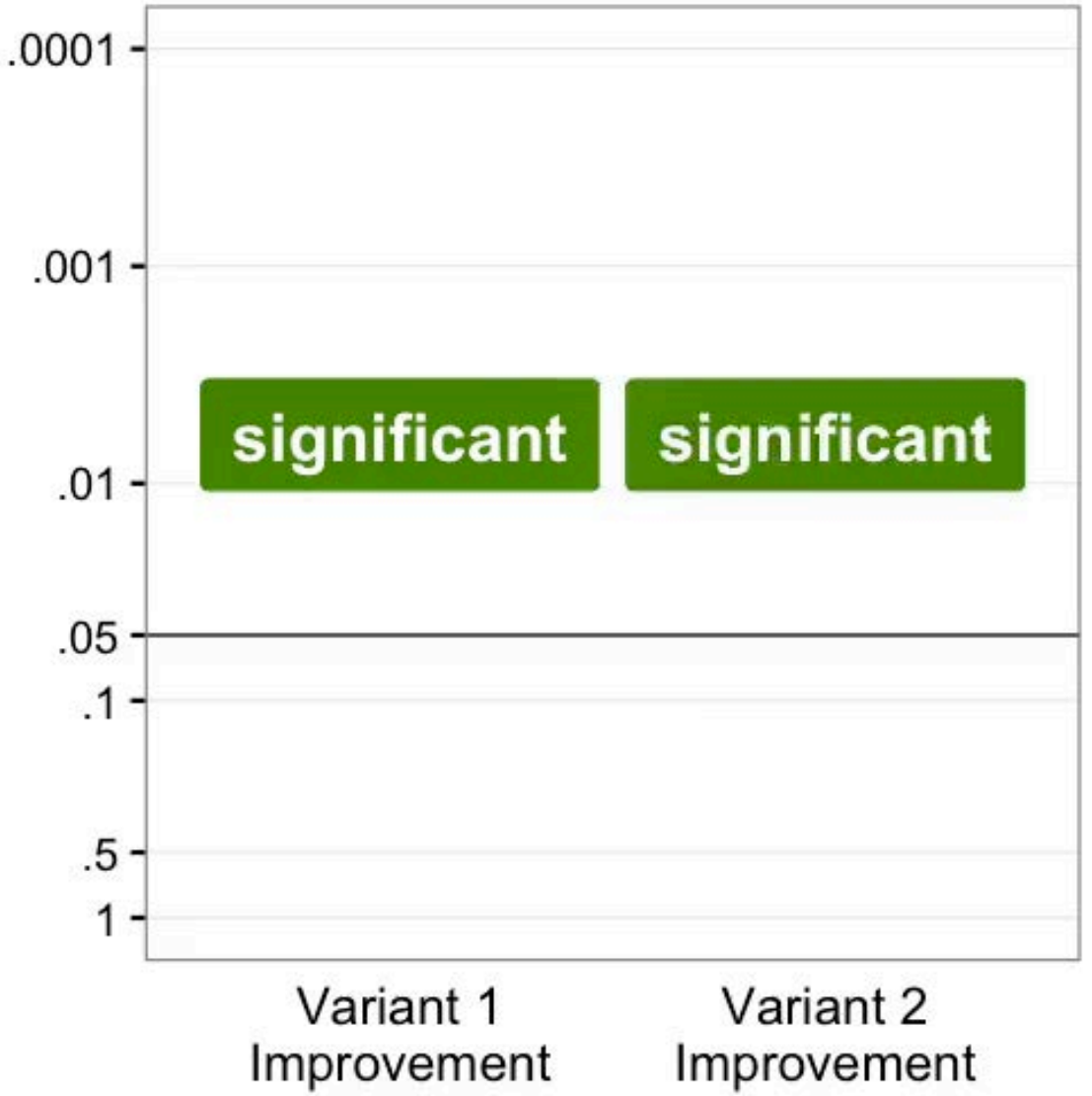

60 s
40 s
20 s
0 s
-20 s
-40 s
10 s
0 s
Variant 1
Improvement
Variant 2
Improvement



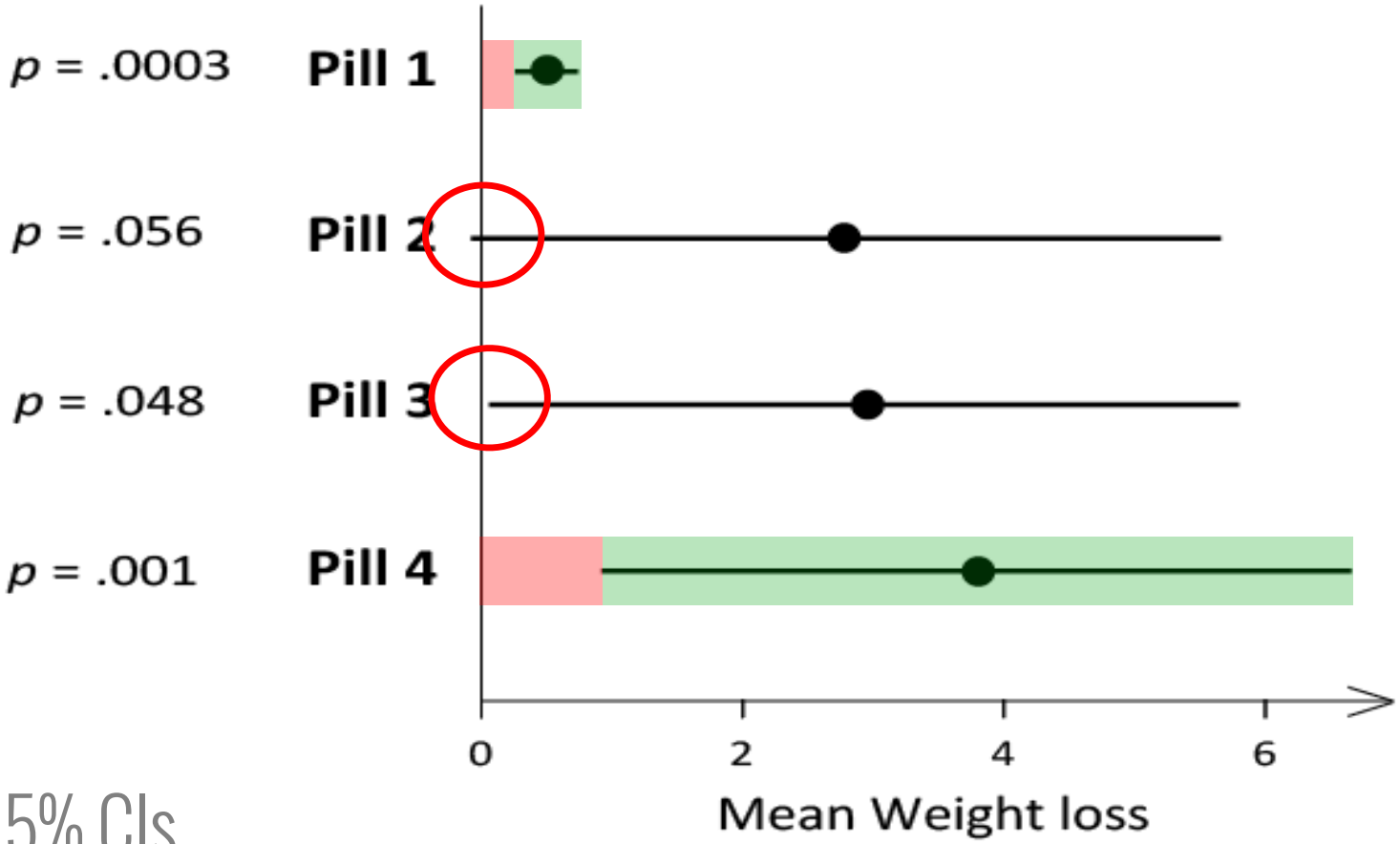
Martin Krzywinski & Naomi Altman (2013)

Taking a more nuanced representation that shows **effect size and confidence** and reducing it to a dichotomous yes/no.





STATISTICAL SIGNIFICANCE



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

STATISTICAL SIGNIFICANCE

"[Null-Hypothesis Significance Testing] is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research."

—Rozeboom (1960)

“P-HACKING”

It’s easier to reach false conclusions
when you’re fixated on one number. You might:

- **Stop collecting data** when p<.05
- Analyze many measures or conditions, but **report only those** with p<.05
- **Exclude participants** to get p<.05
- **Transform the data** to get p<.05
- Etc.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV.

Ioannidis, 2005
PLoS Medicine

MORE USEFUL DISCUSSION OF THIS PHENOMENON

https://www.youtube.com/watch?v=42QuXLucH3Q

SO WHAT SHOULD WE DO?

(When doing visual analysis.)

Be conscious about what you're trying to claim.

Are you making **a descriptive statement about the sample**?
Or **an inference about the population**?

If you're trying to make an inferential claim
– be **cautious and conservative in your interpretations.**

SO WHAT SHOULD WE DO?

DON'T MAKE CLAIMS ABOUT:

- POPULATION-LEVEL DIFFERENCES
- "SIGNIFICANCE"
- CAUSALITY

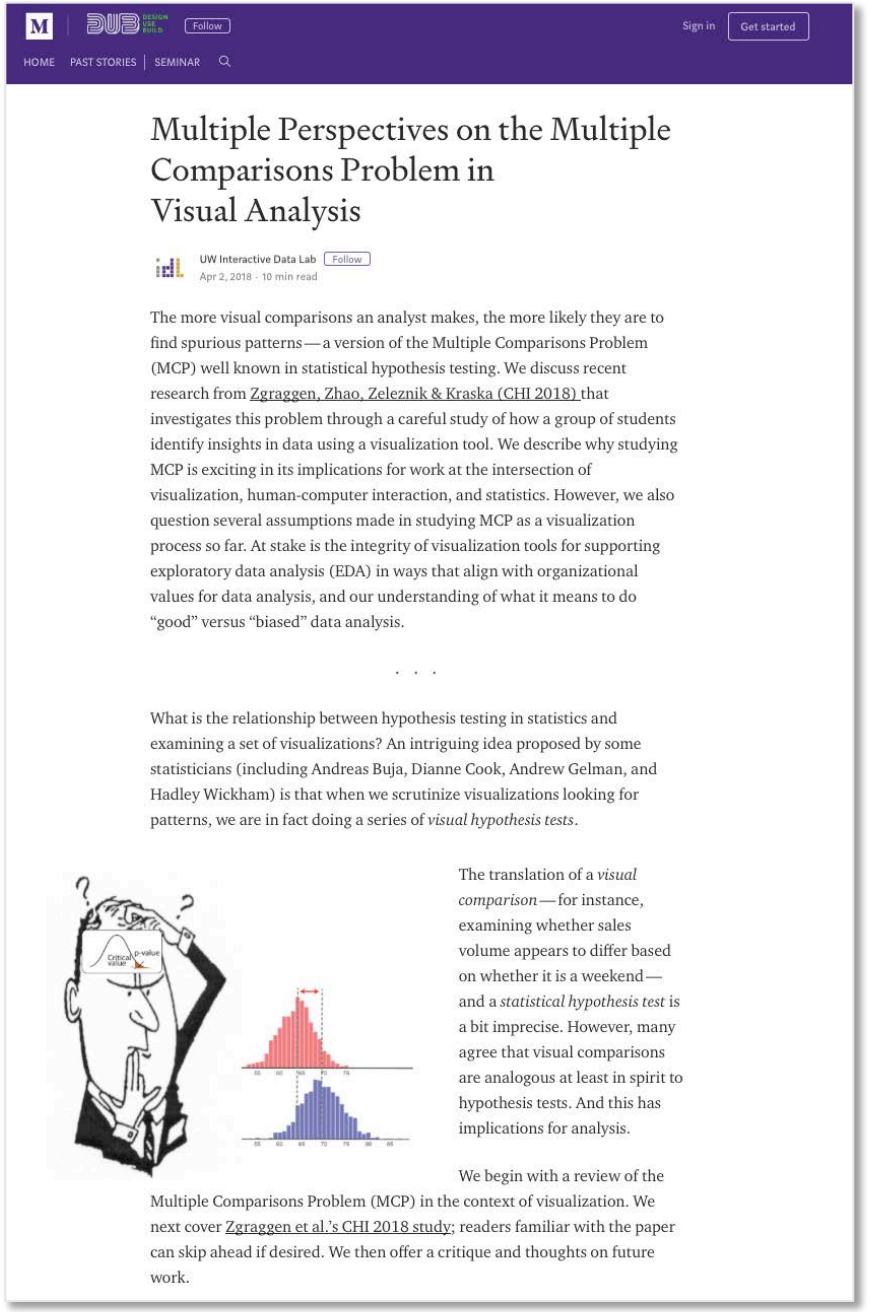
...UNLESS THE RESULTS ARE **VERY** ROBUST.

SO WHAT SHOULD WE DO?

BE WARY OF THE **MULTIPLE-COMPARISON PROBLEM**

BE SUSPICIOUS OF **AUTOMATED CORRELATION DETECTION**

https://bit.ly/2PW65MO



Follow

Sign in | Get started

HOME | PAST STORIES | SEMINAR

Multiple Perspectives on the Multiple Comparisons Problem in Visual Analysis

UW Interactive Data Lab Follow

Apr 2, 2018 · 10 min read

The more visual comparisons an analyst makes, the more likely they are to find spurious patterns—a version of the Multiple Comparisons Problem (MCP) well known in statistical hypothesis testing. We discuss recent research from Zgraggen, Zhao, Zeleznik & Kraska (CHI 2018) that investigates this problem through a careful study of how a group of students identify insights in data using a visualization tool. We describe why studying MCP is exciting in its implications for work at the intersection of visualization, human-computer interaction, and statistics. However, we also question several assumptions made in studying MCP as a visualization process so far. At stake is the integrity of visualization tools for supporting exploratory data analysis (EDA) in ways that align with organizational values for data analysis, and our understanding of what it means to do "good" versus "biased" data analysis.

. . .

What is the relationship between hypothesis testing in statistics and examining a set of visualizations? An intriguing idea proposed by some statisticians (including Andreas Buja, Dianne Cook, Andrew Gelman, and Hadley Wickham) is that when we scrutinize visualizations looking for patterns, we are in fact doing a series of *visual hypothesis tests*.

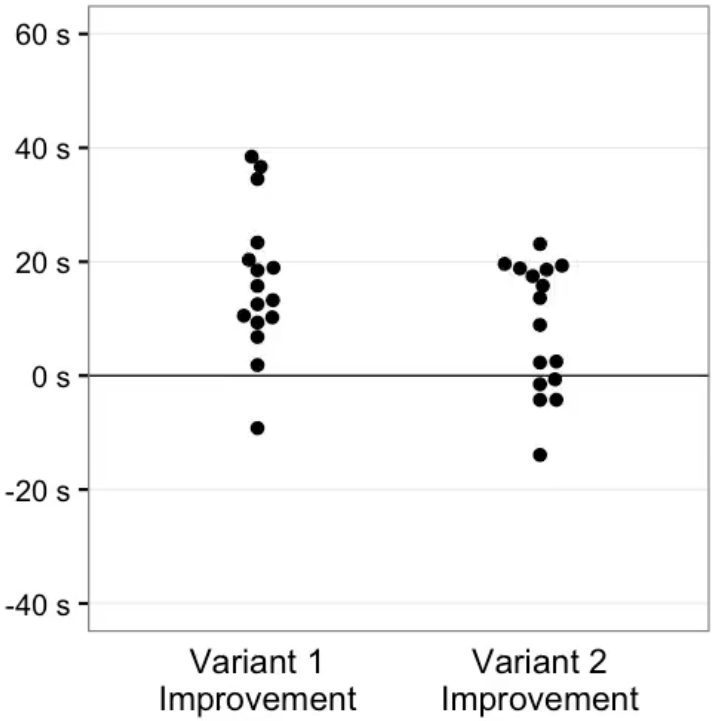
The translation of a *visual comparison*—for instance, examining whether sales volume appears to differ based on whether it is a weekend—and a *statistical hypothesis test* is a bit imprecise. However, many agree that visual comparisons are analogous at least in spirit to hypothesis tests. And this has implications for analysis.

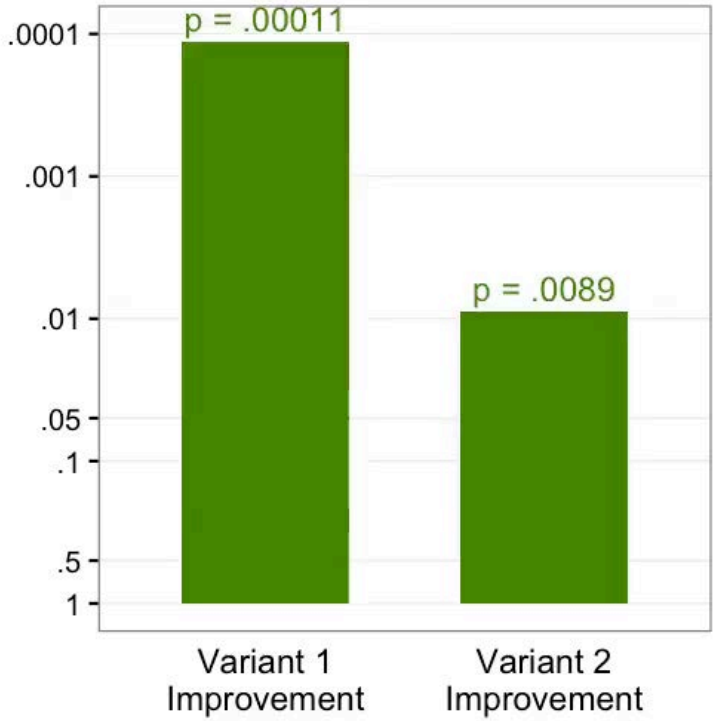
We begin with a review of the Multiple Comparisons Problem (MCP) in the context of visualization. We next cover Zgraggen et al.'s CHI 2018 study; readers familiar with the paper can skip ahead if desired. We then offer a critique and thoughts on future work.

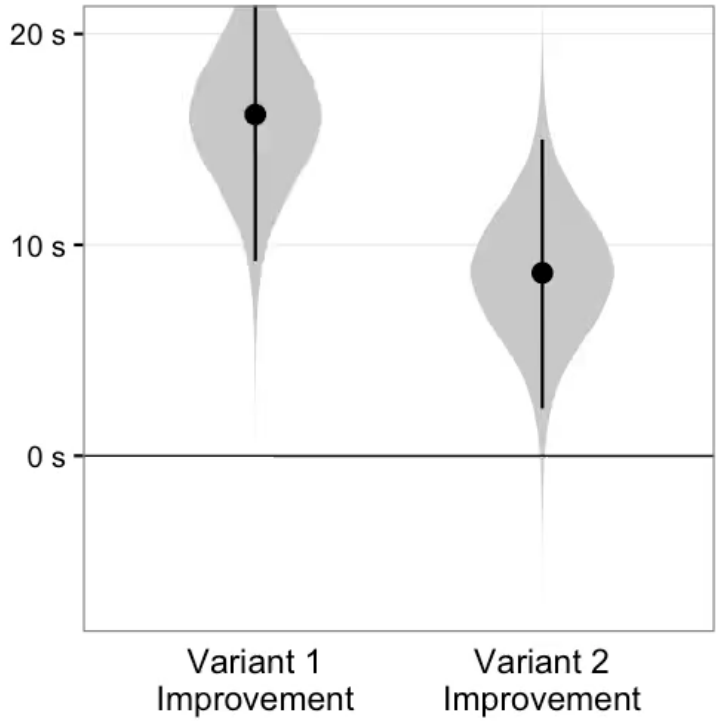
SO WHAT SHOULD WE DO?

Use Larger Sample Sizes!

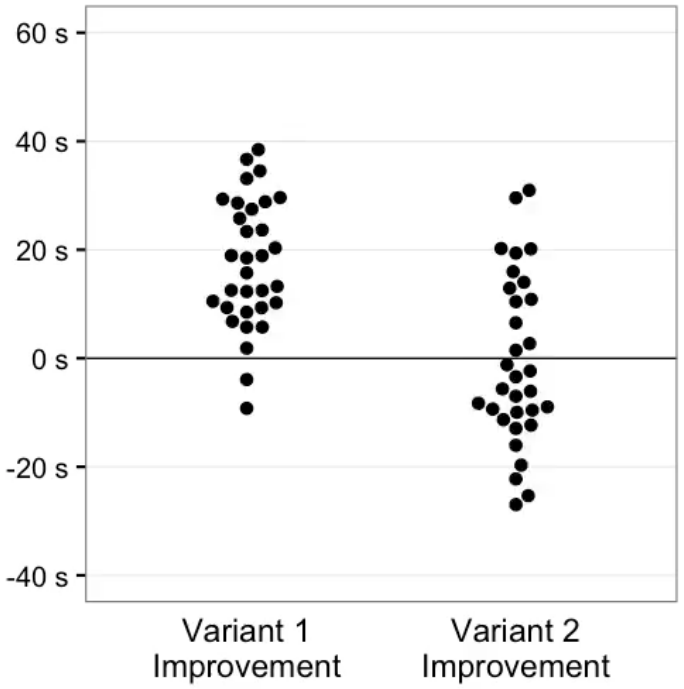
N=16

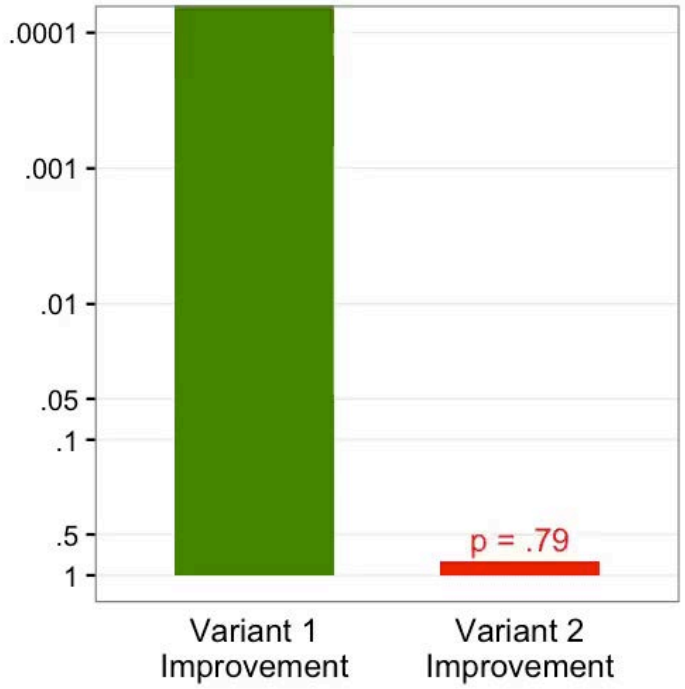


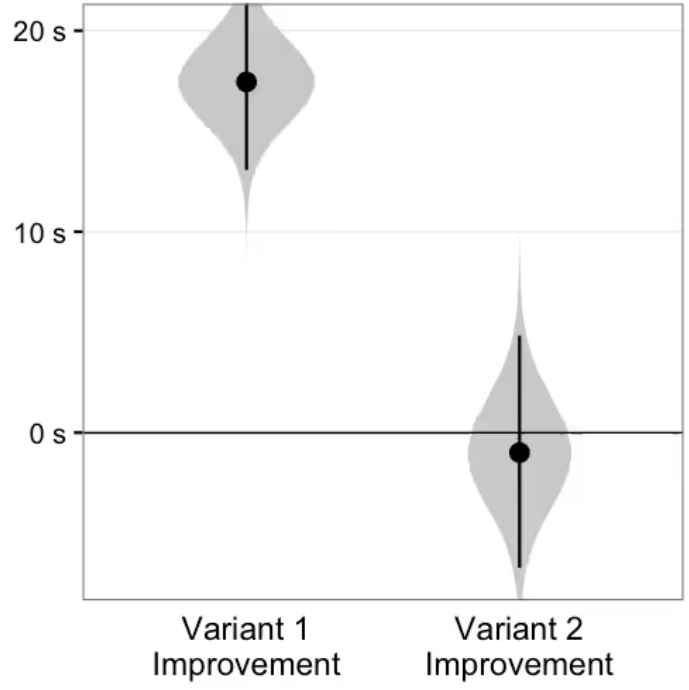




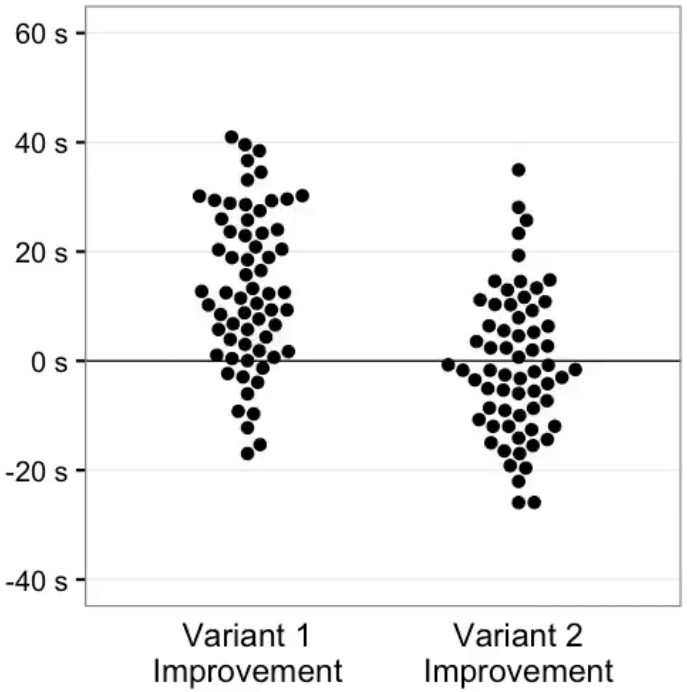
N=32

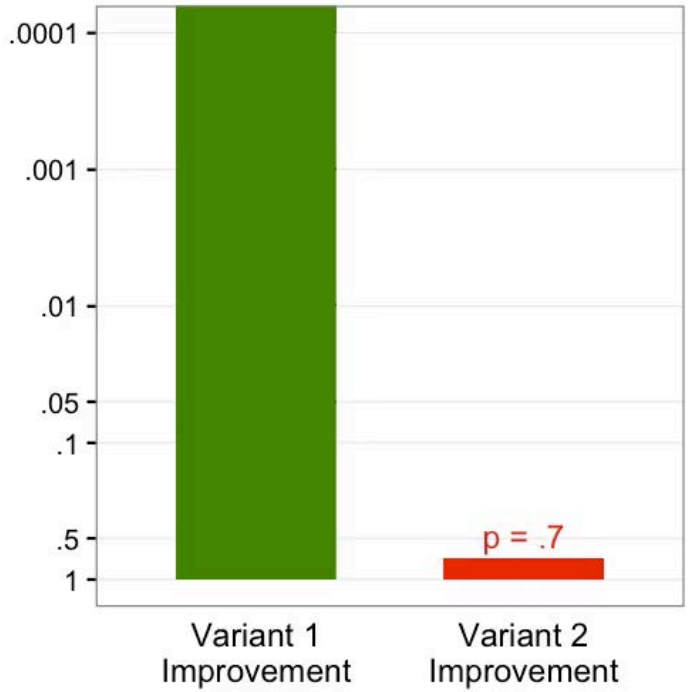

60 s
40 s
20 s
0 s
-20 s
-40 s
Variant 1 Improvement
Variant 2 Improvement

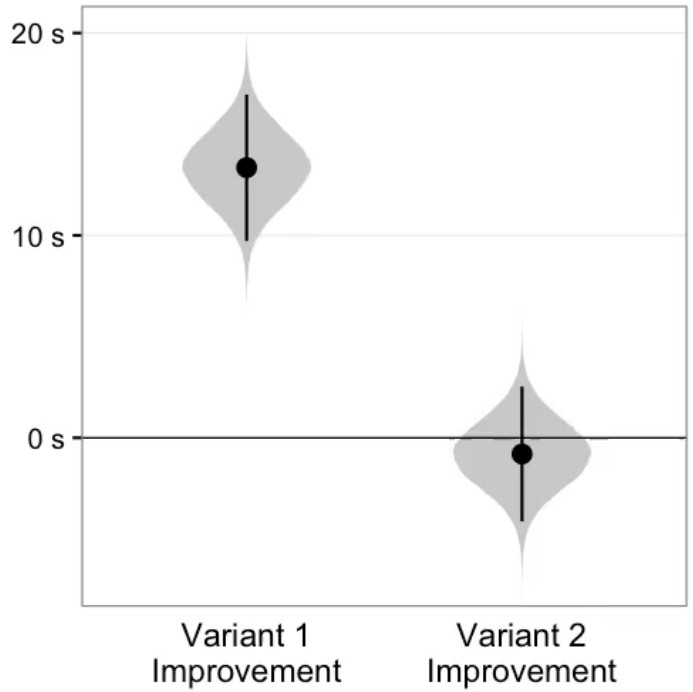

.0001
.001
.01
.05
.1
.5
1
p = .79
Variant 1 Improvement
Variant 2 Improvement


20 s
10 s
0 s
Variant 1 Improvement
Variant 2 Improvement

N=64


60 s
40 s
20 s
0 s
-20 s
-40 s
Variant 1 Improvement
Variant 2 Improvement

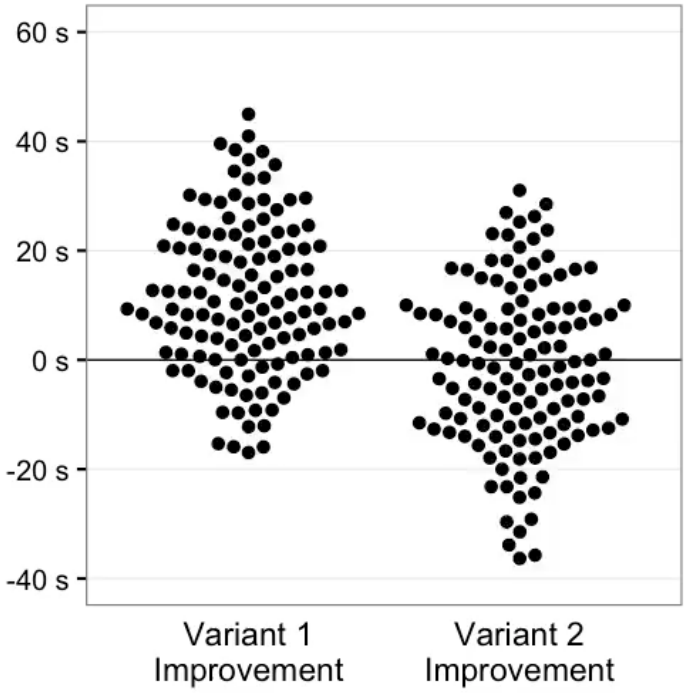

.0001
.001
.01
.05
.1
.5
1
p = .7
Variant 1 Improvement
Variant 2 Improvement

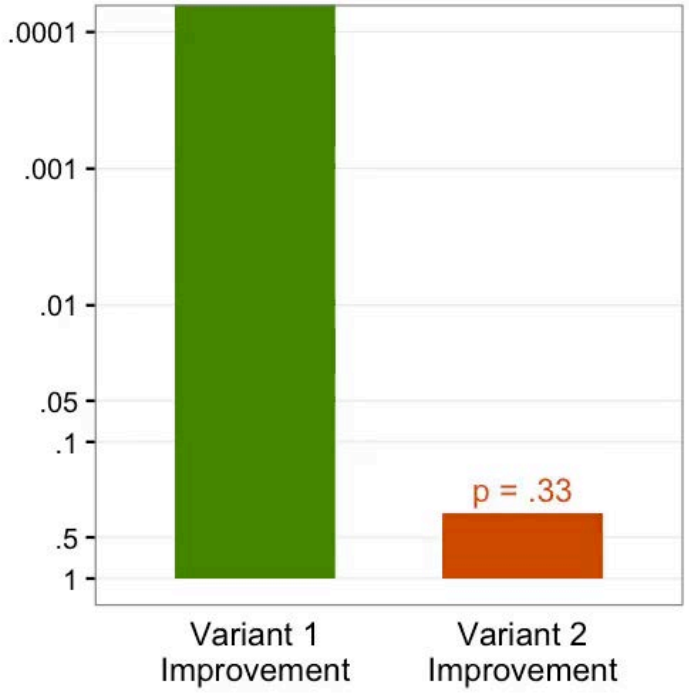

20 s
10 s
0 s
Variant 1 Improvement
Variant 2 Improvement

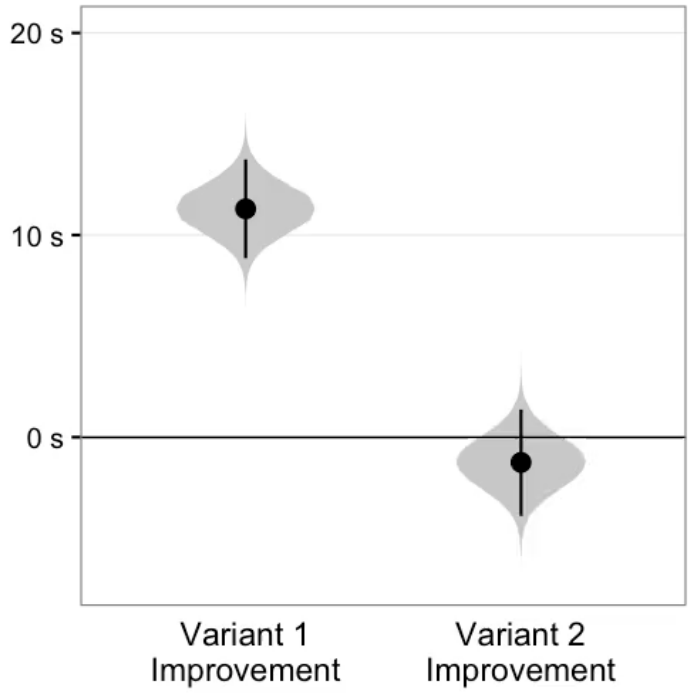
N=128

But, you don't always have this option...

- Often collecting data is hard/expensive.
- Maybe you didn't collect it at all!







A MORE HOLISTIC UNDERSTANDING OF THE DATA

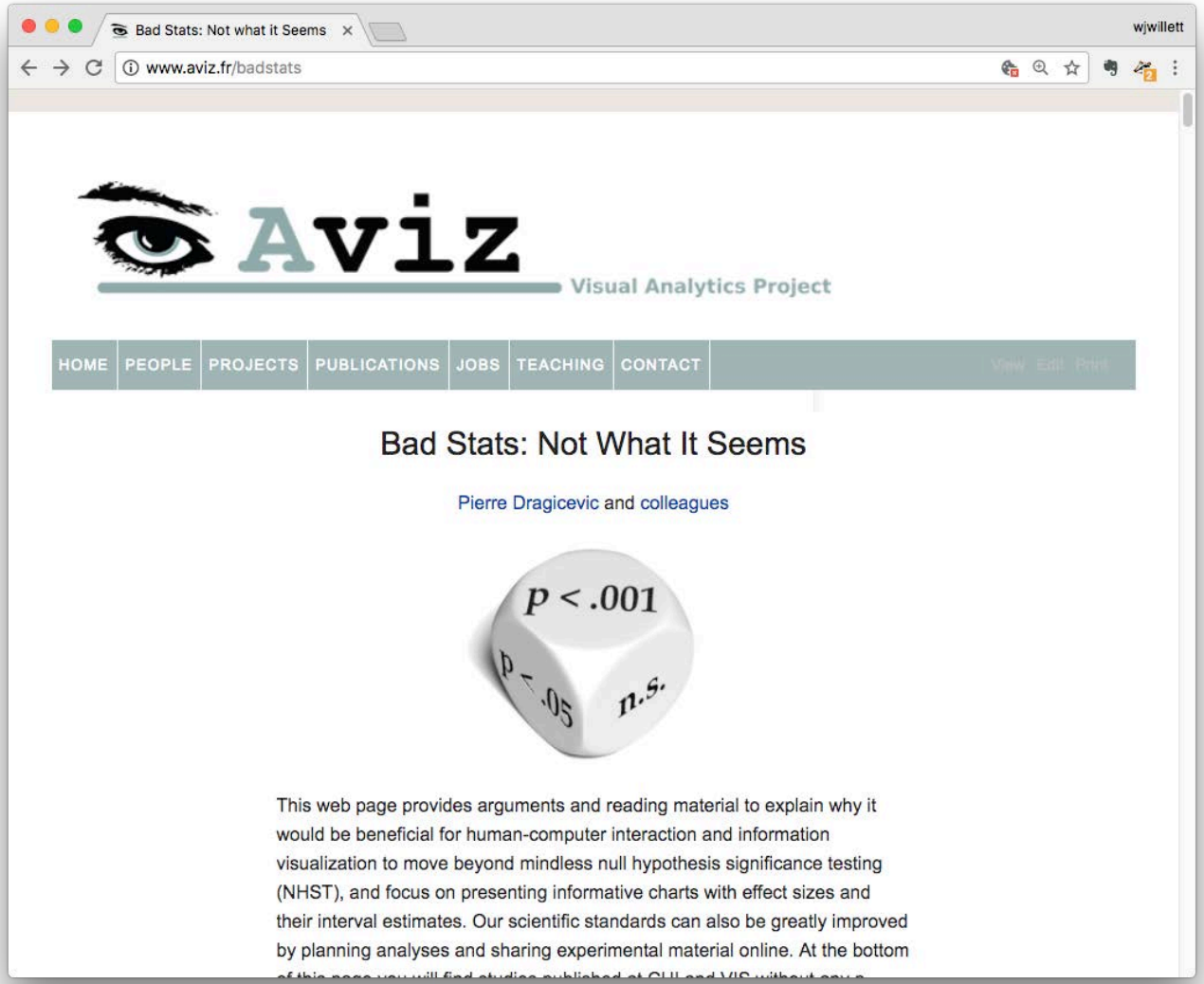
You can make these **same mistakes** using confidence intervals, significance tests, Baysean statistics, etc...

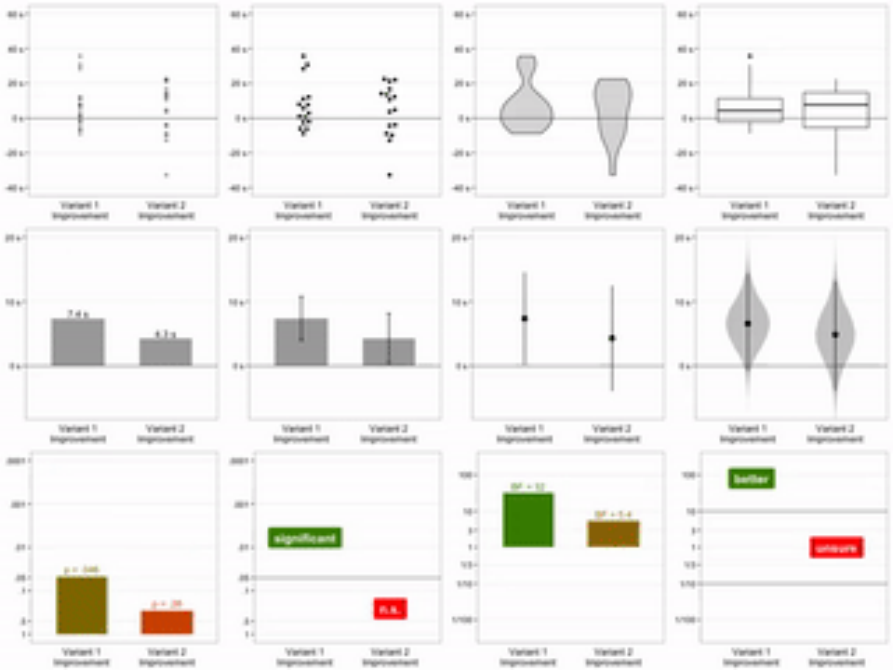
What's important is:

- understanding the **size of the effects**,
- paying attention to the **amount of (un)certainty**,
- being aware of your **potential biases**,
- and **interpreting the data in context**.

Use all the tools you have, and **don't overstate your results.**

For more: www.aviz.fr/badstats





PIERRE DRAGICEVIC