

# INTRODUCTION TO STATISTICS

Pierre Dragicevic

# WHAT YOU WILL LEARN

Statistical  
theory

Applied  
statistics



This lecture

# GOALS

- Learn basic intuitions and terminology
- Perform basic statistical inference with R
- Focus on high-level principles
- Accent on estimation rather than null hypothesis testing ("the New Statistics")

# ORGANIZATION

- Part I - Elementary notions
- Part II - Tutorial with R
- Part III - Assignments



# A DEFINITION

- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.

Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, OUP.

# 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

# WHAT ARE STATS?

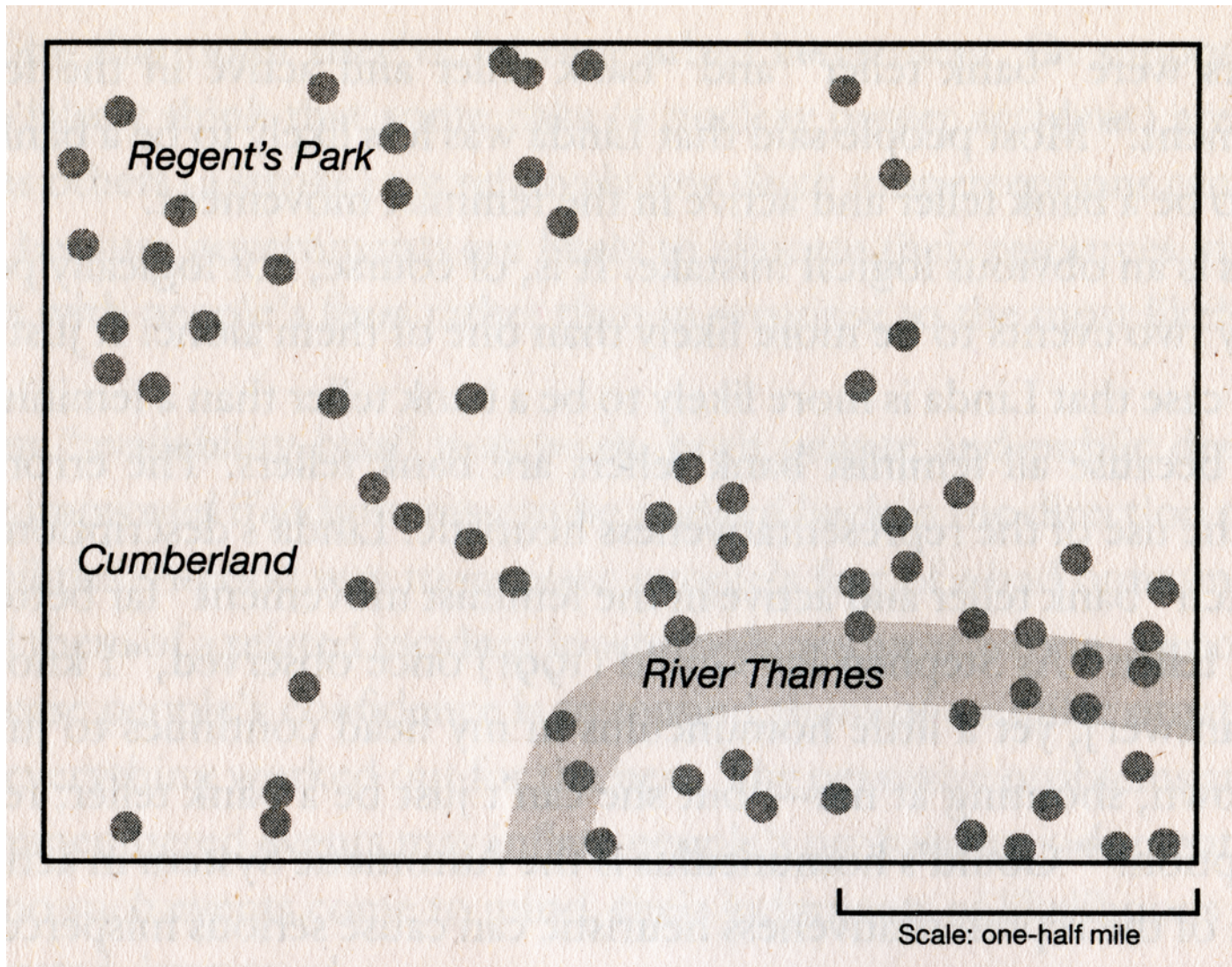
- Good for:
  - Summarizing data for presentation
  - Answering empirical questions rigorously
  - Making predictions
  - Making rational, evidence-based decisions
  - A long accumulated experience!

# STATS & VISUALIZATION

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

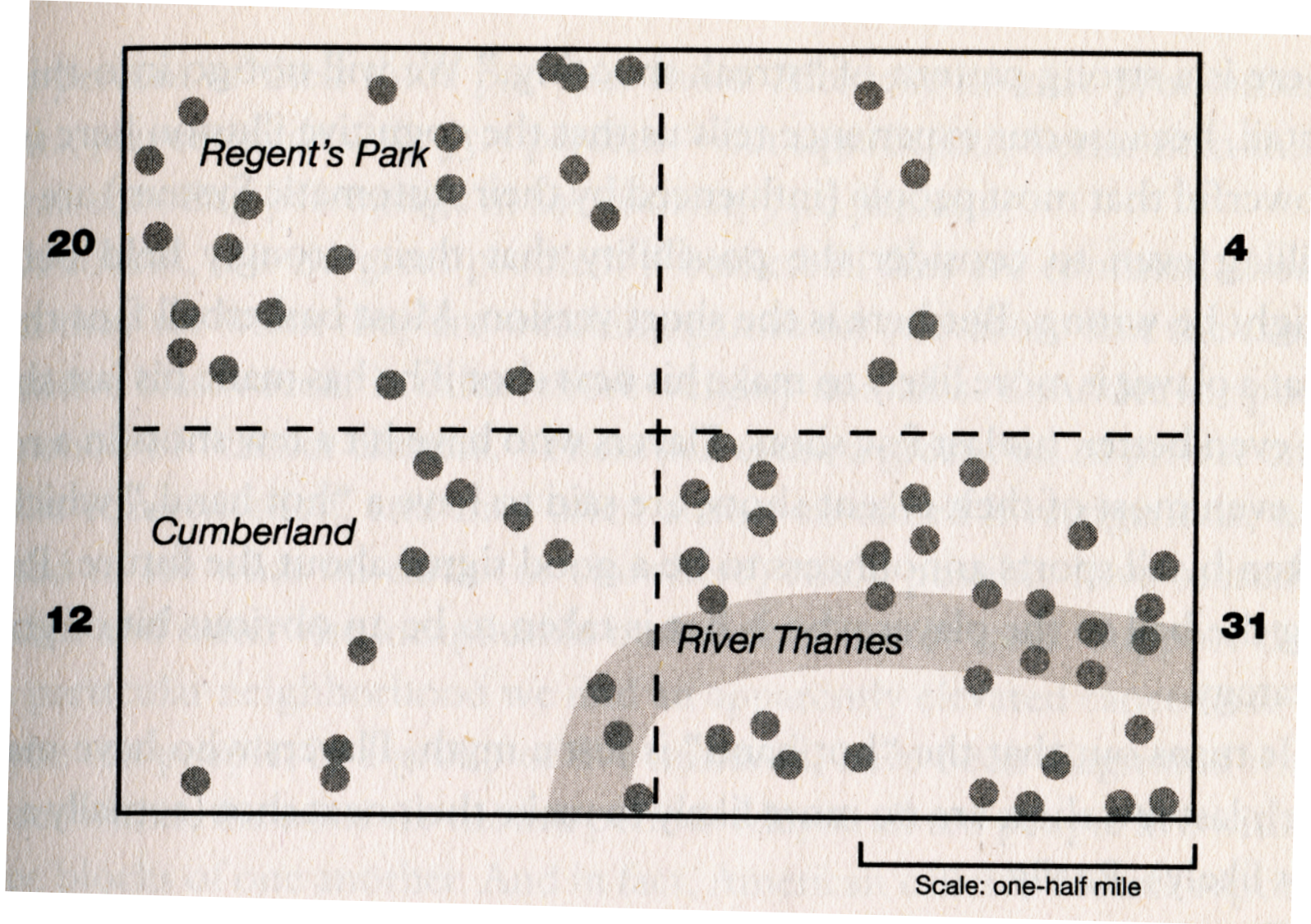
**Confirmatory data analysis** is comparable to a **court trial**: it is the process of evaluating evidence.

Exploratory analysis and confirmatory analysis “*can—and should—proceed side by side*” (Tukey; 1977).



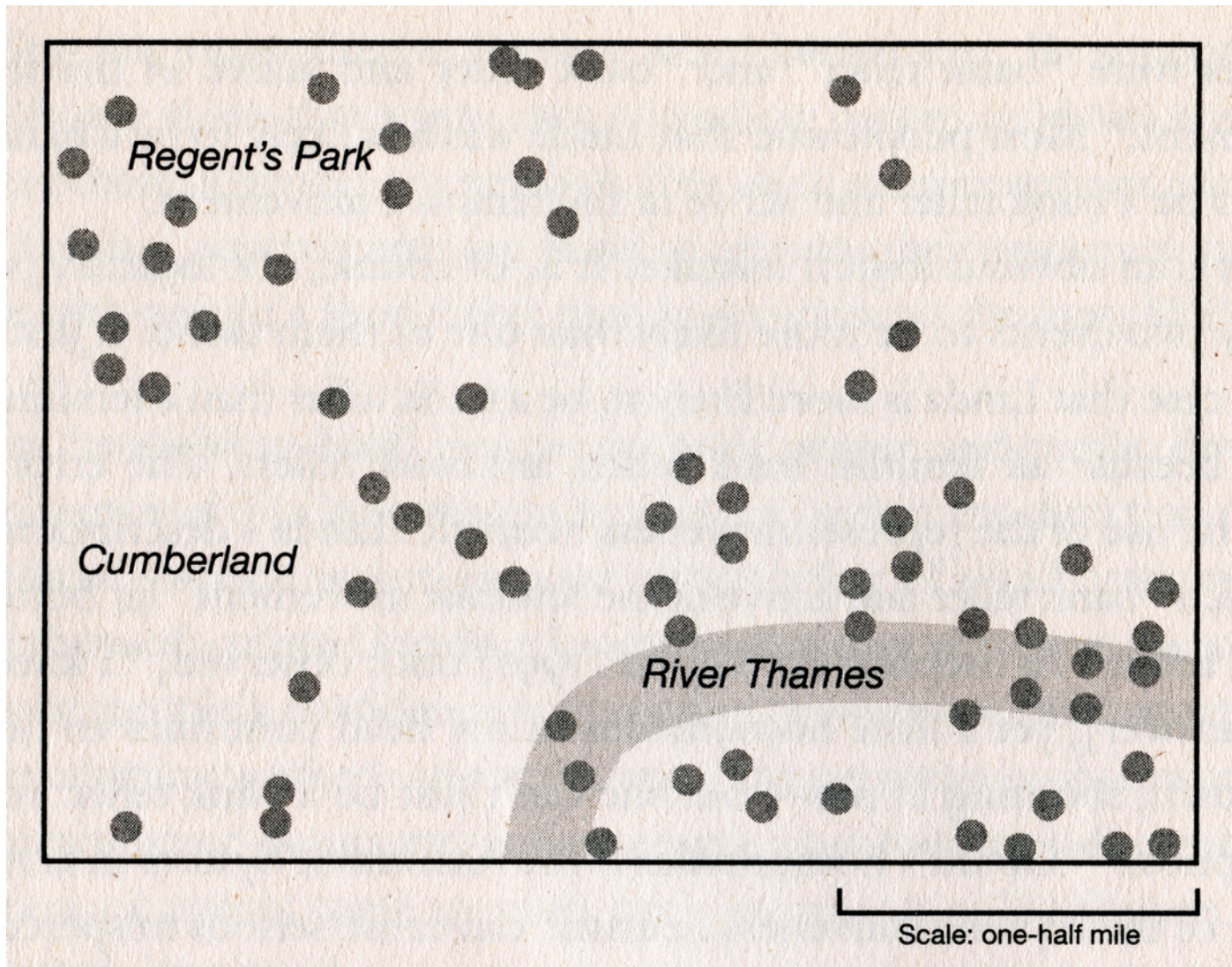
German bombings in London during WWII





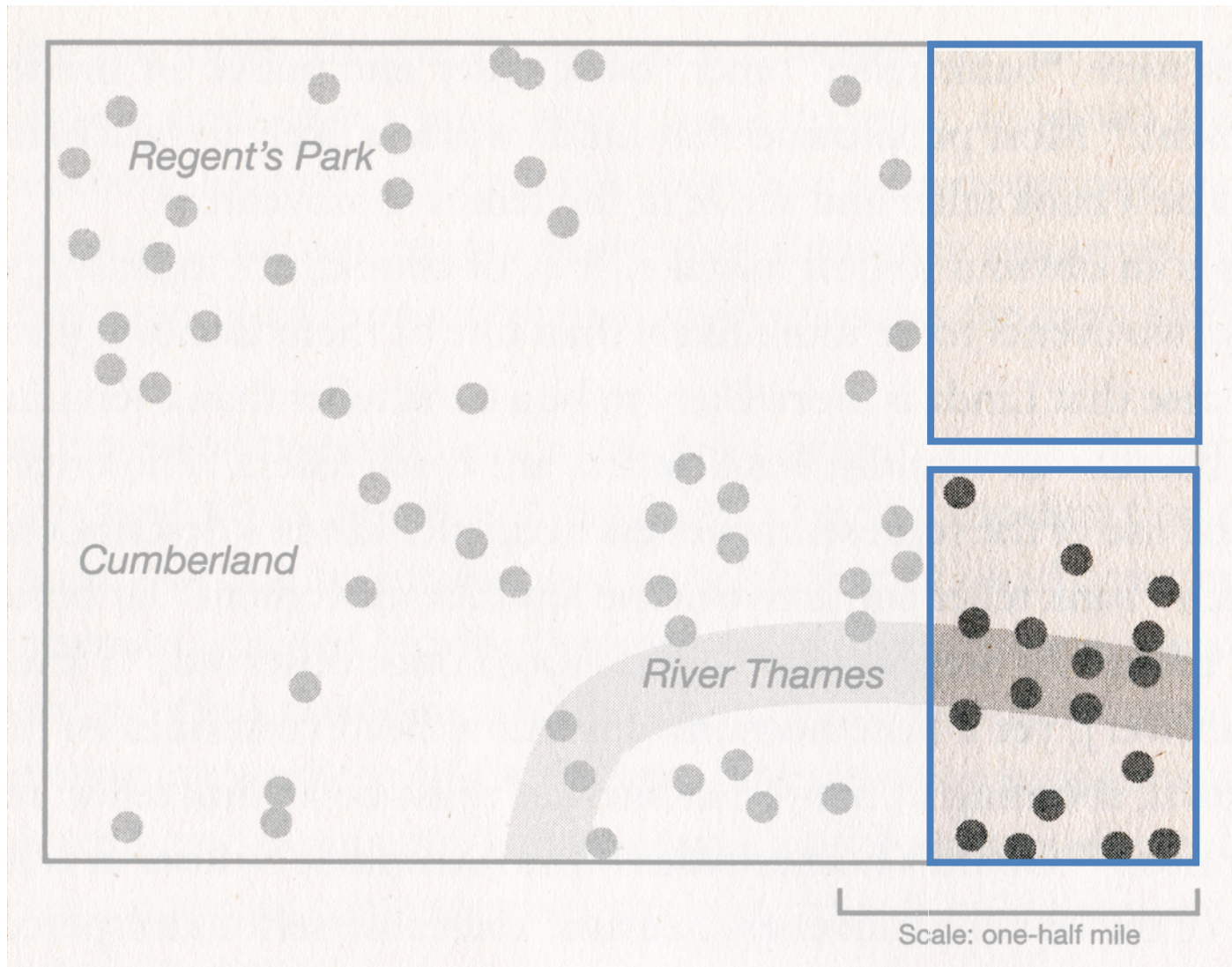
German bombings in London during WWII





German bombings in London during WWII





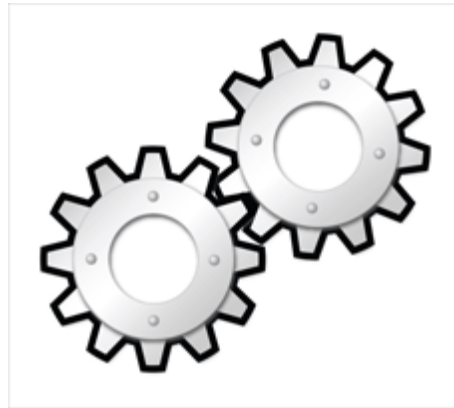
German bombings in London during WWII



# STATISTICAL TOOLS

DESCRIPTIVE STATISTICS

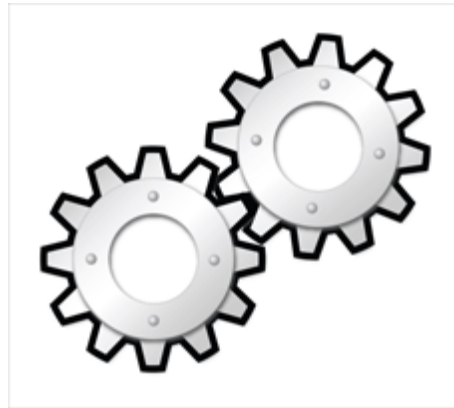
INFERENCE STATISTICS



# STATISTICAL TOOLS

DESCRIPTIVE STATISTICS

INFERENCE STATISTICS



# AN EXAMPLE

- Selling encyclopedias



Robert



Steve



Paul



Roger



Geoffrey

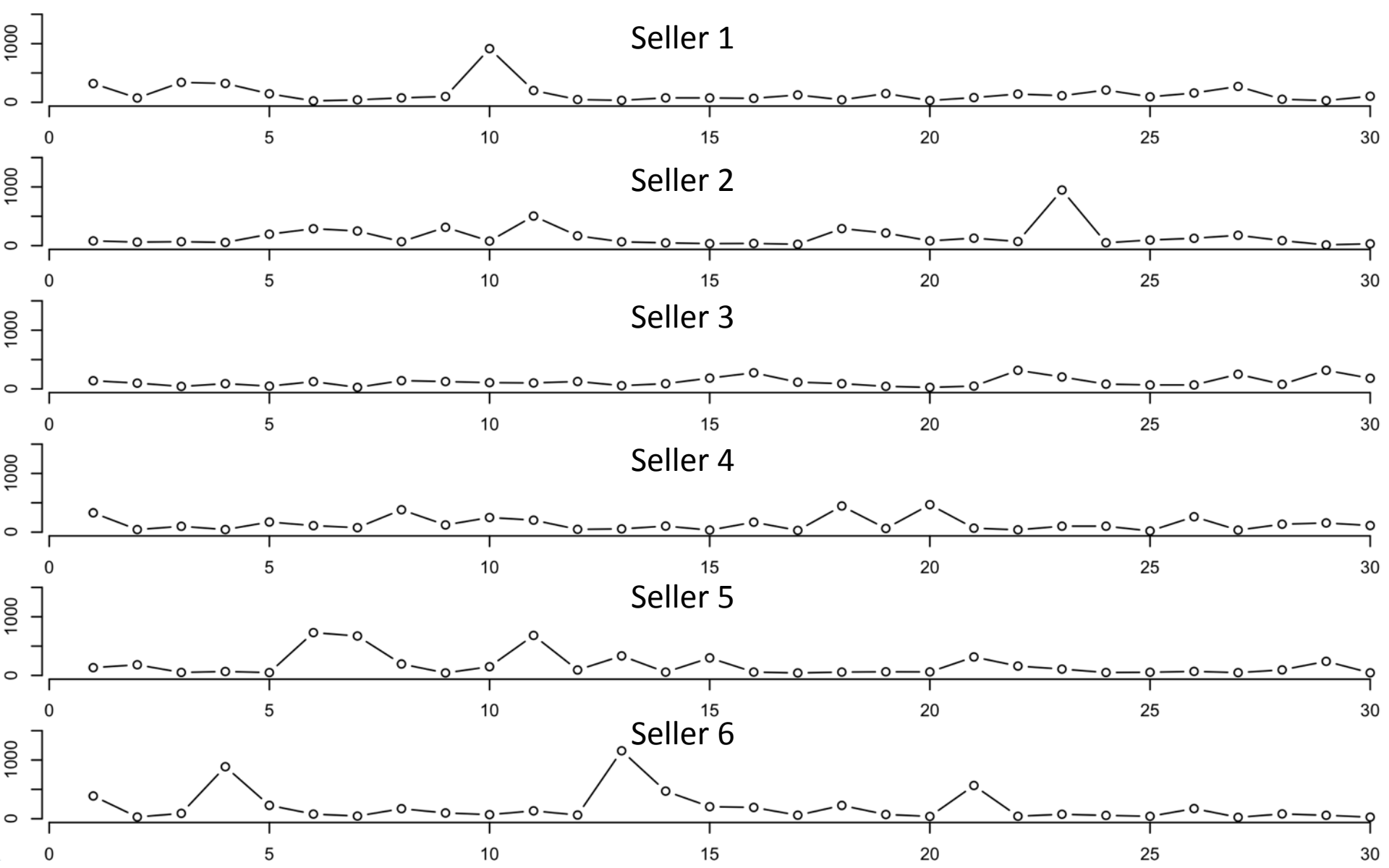


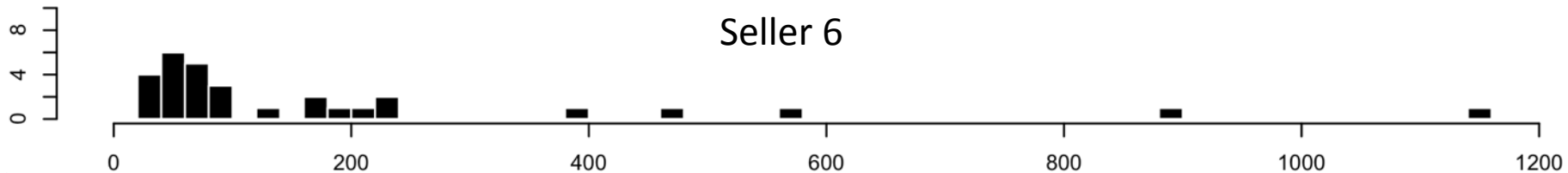
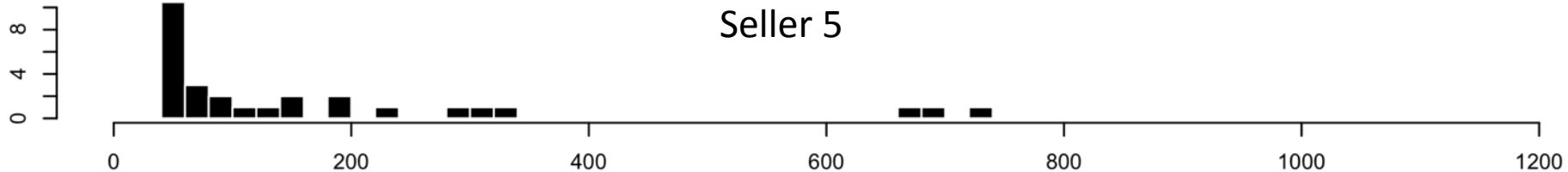
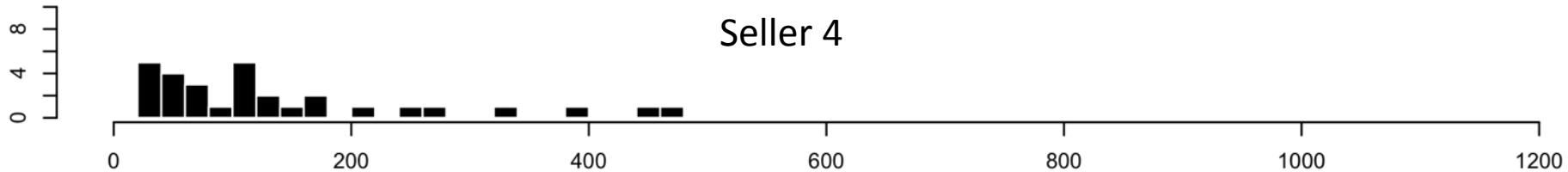
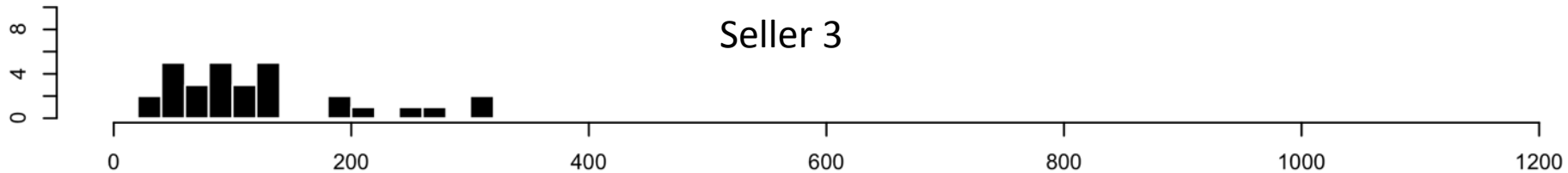
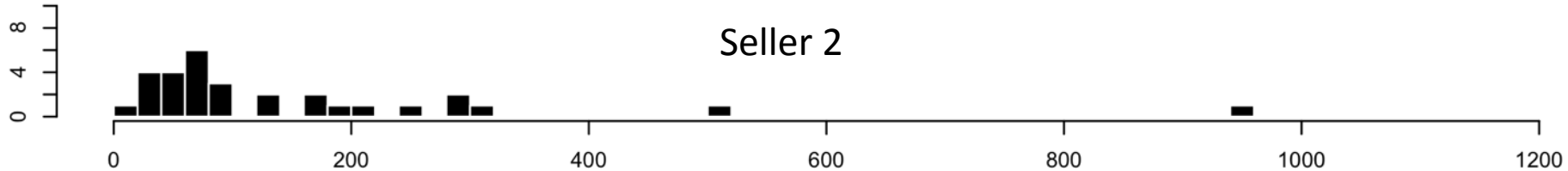
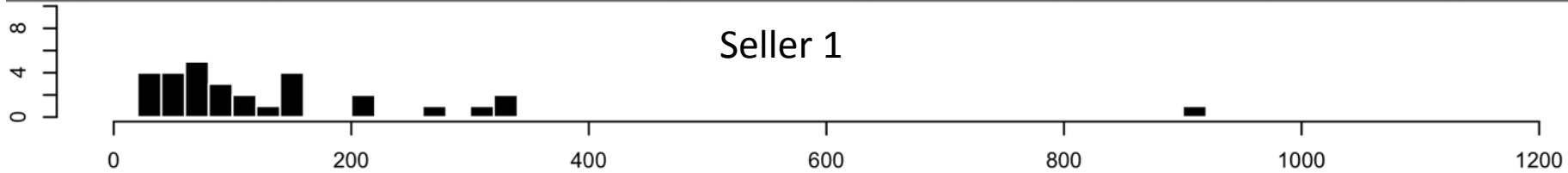
Dan



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

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
19	€149	€215	€43	€63	€62	€72
20	€31	€81	€26	€469	€60	€39
21	€81	€127	€47	€68	€315	€566
22	€141	€70	€317	€40	€160	€42
23	€113	€947	€203	€102	€108	€76
24	€209	€48	€81	€102	€50	€56
25	€94	€95	€67	€21	€54	€41
26	€159	€125	€67	€263	€69	€173
27	€271	€176	€250	€35	€48	€24
28	€52	€85	€77	€136	€95	€82
29	€30	€12	€317	€157	€240	€58
30	€104	€31	€181	€113	€45	€27



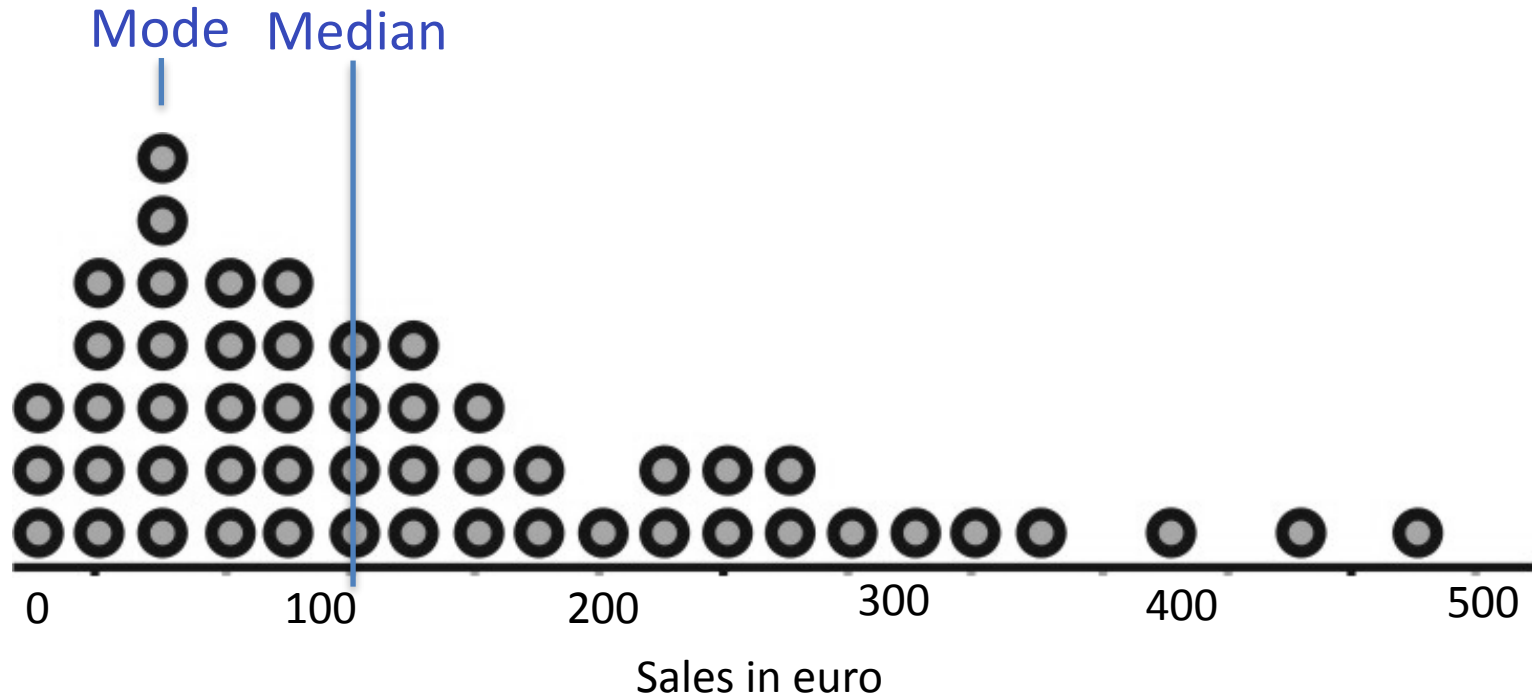


# CENTRAL TENDENCY

Name & Meaning	Formula / Example	Used for
Arithmetic Mean [average]	$\frac{\text{sum}}{\text{size}} = \frac{a+b+c}{3}$	Most situations ("average item")
Median [middle value]	Middle of sorted list (2 middles? Average 'em)	Wildly varying samples (houses, incomes)
Mode [most popular]	Most popular value	No compromises (winner takes all)
Geometric Mean [average factor]	$\sqrt[3]{abc}$	Investments, growth, area, volume
Harmonic Mean [average rate]	$\frac{3}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c}}$	Speed, production, cost

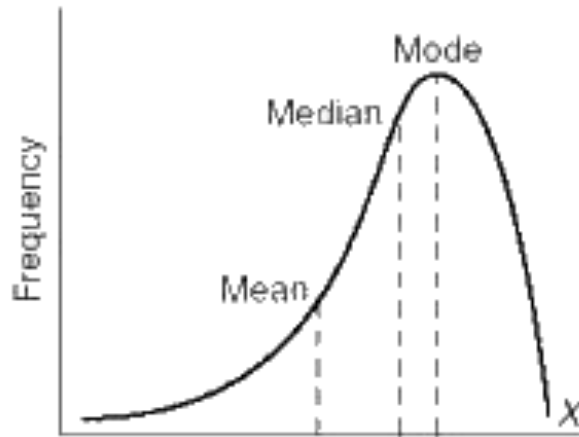


# CENTRAL TENDENCY



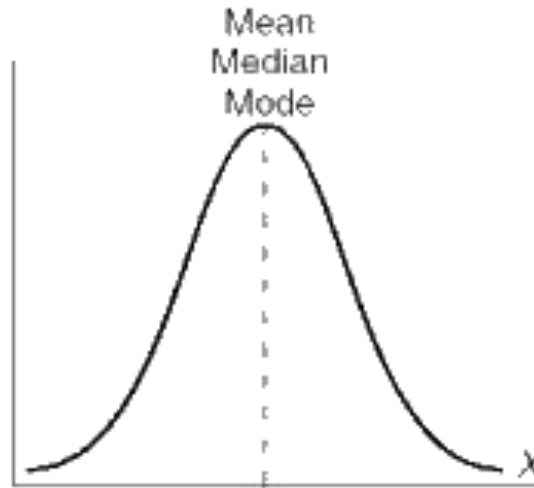
# CENTRAL TENDENCY

(a) Negatively skewed



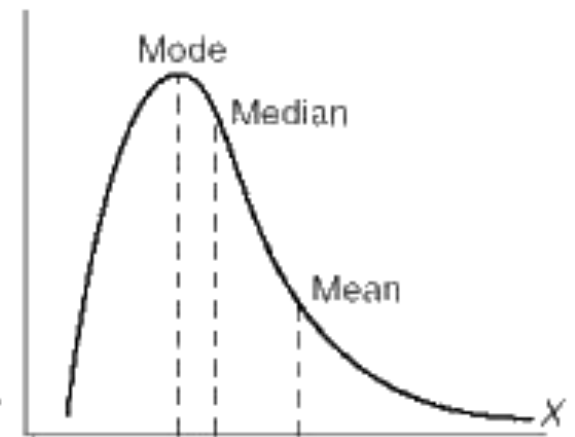
←  
Negative Direction

(b) Normal (no skew)



Perfectly Symmetrical  
Distribution

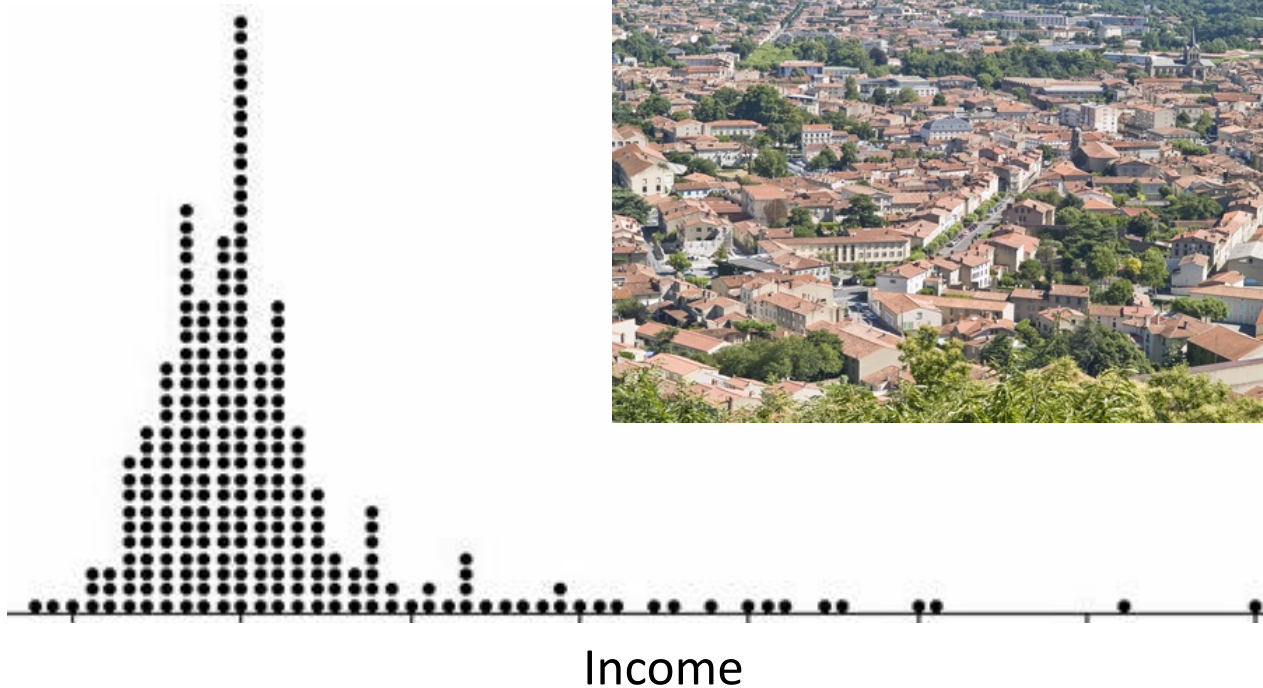
(c) Positively skewed



→  
Positive Direction

# CENTRAL TENDENCY

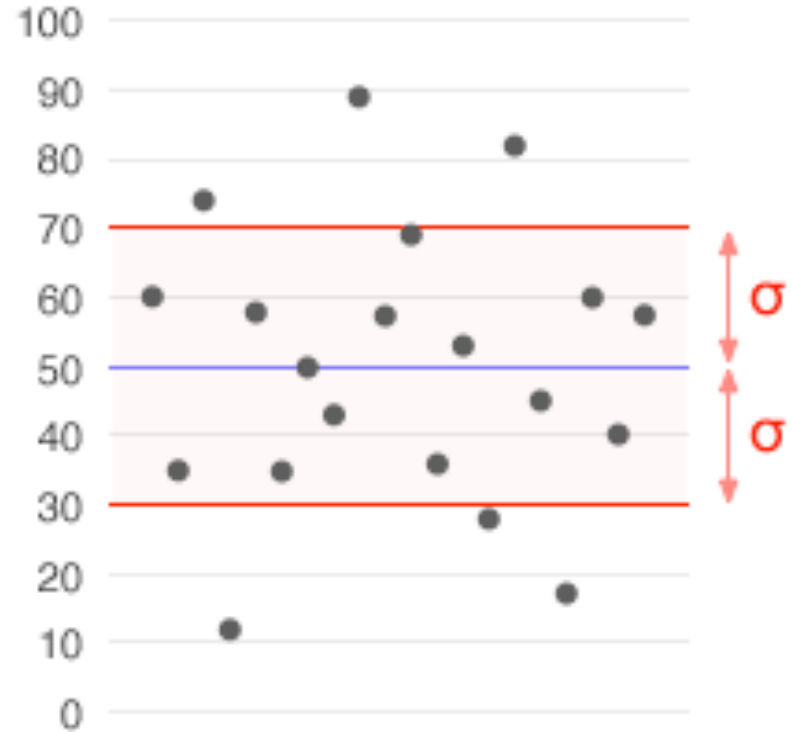
What is the best measure of central tendency?



# DISPERSION

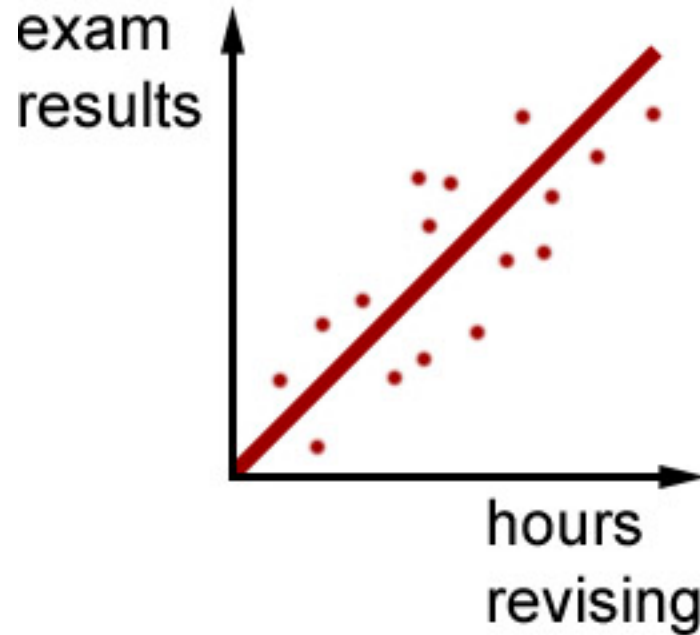
## Standard Deviation

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



# DEPENDENCE

- Correlation

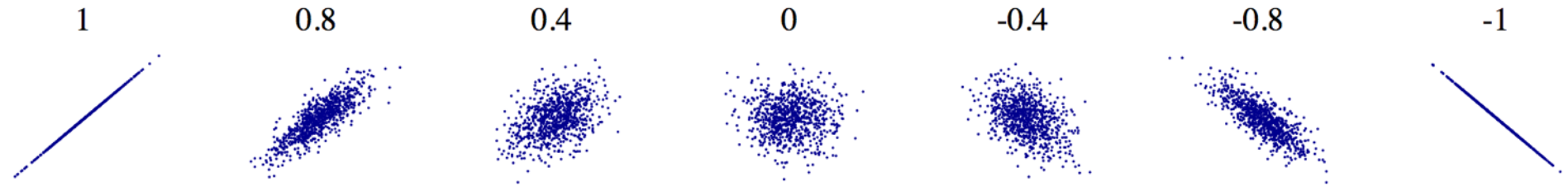


POSITIVE CORRELATION

- people who do more revision get higher exam results.

# DEPENDENCE

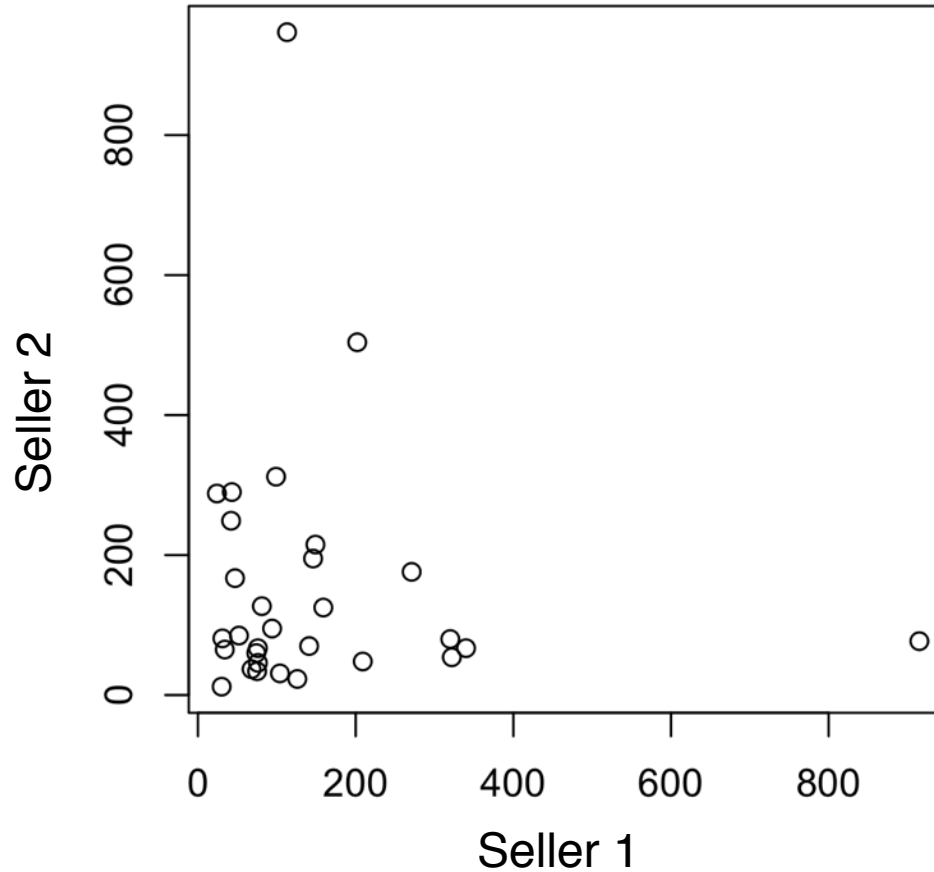
- Correlation

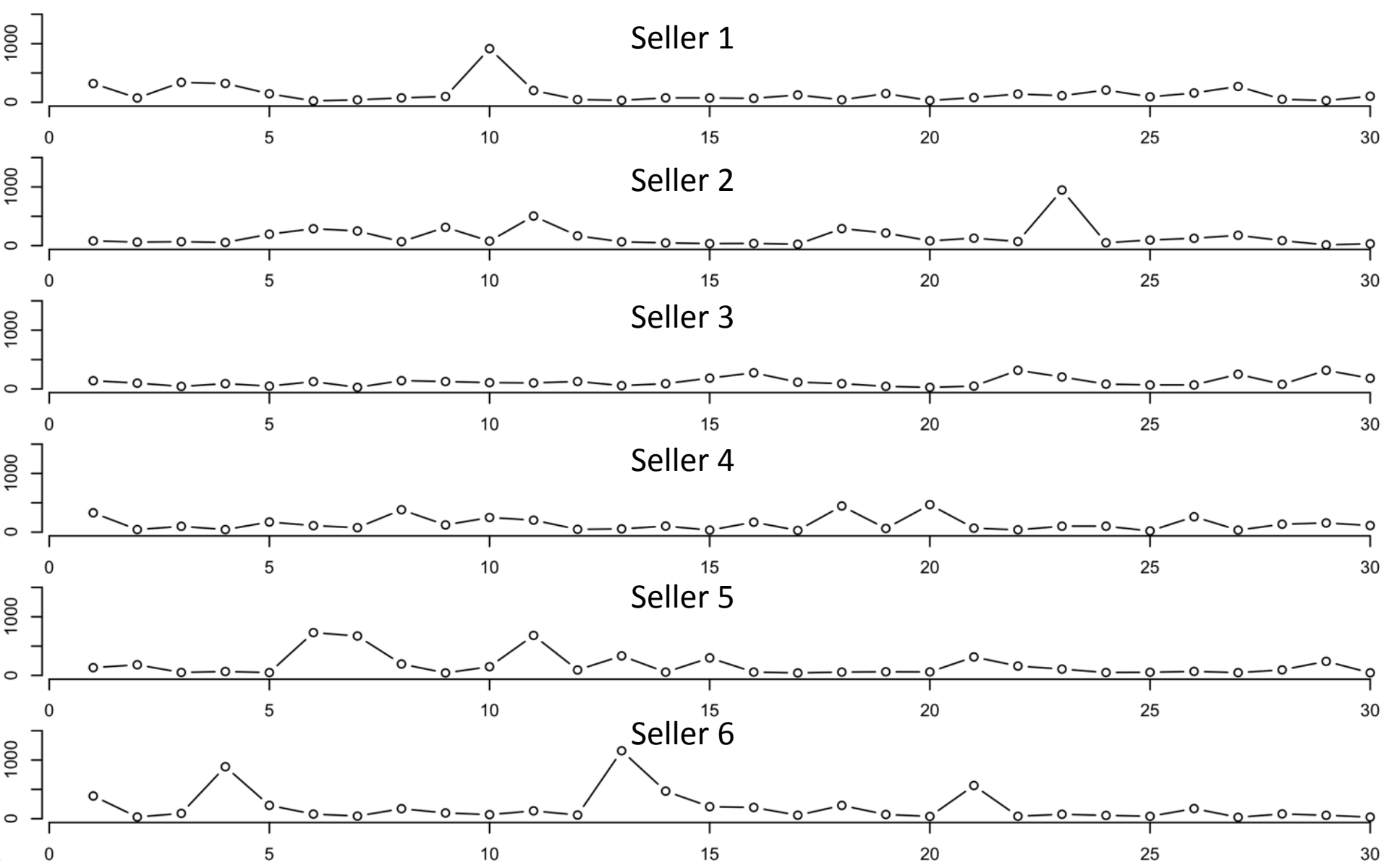


# DEPENDENCE

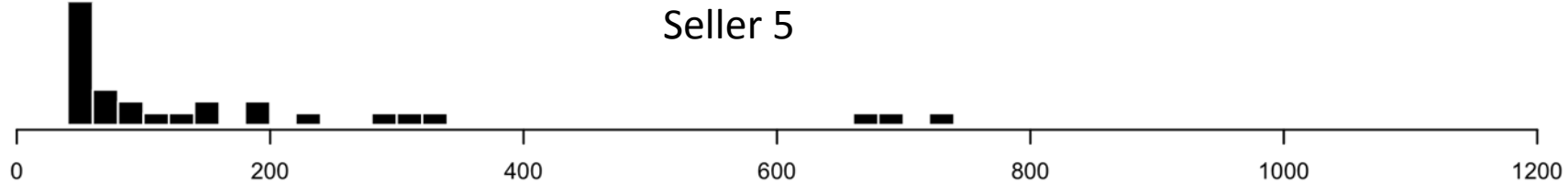
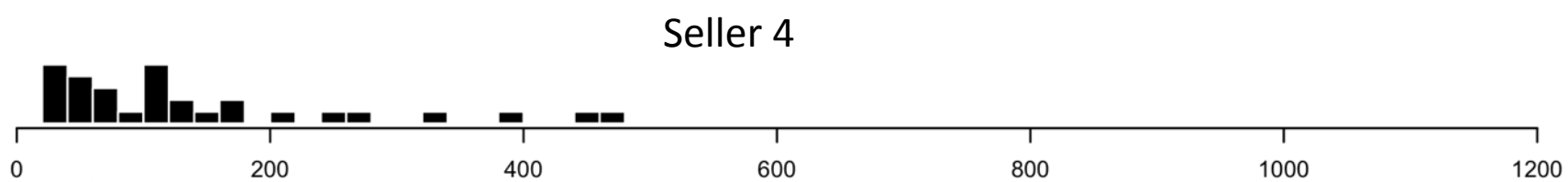
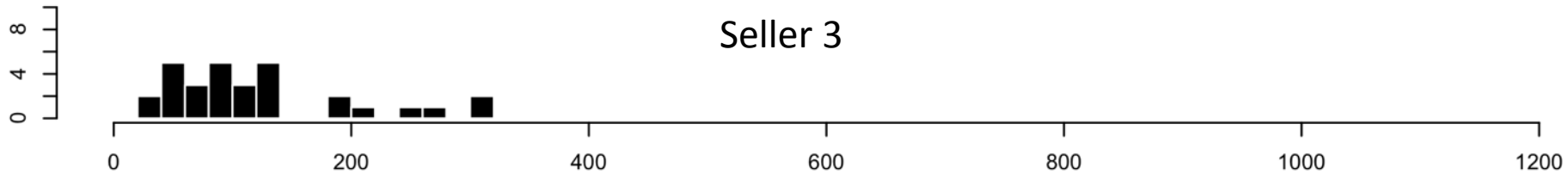
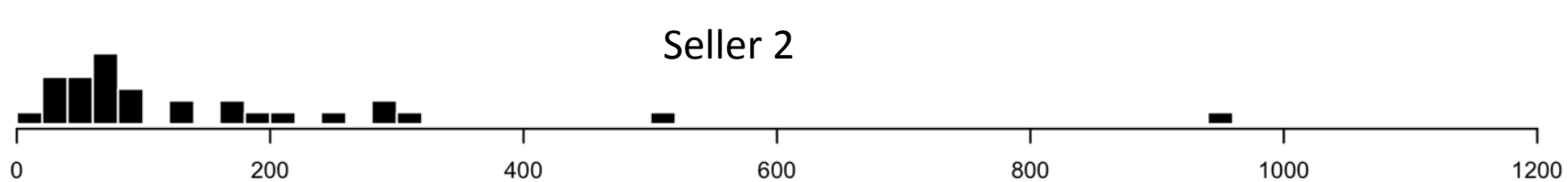
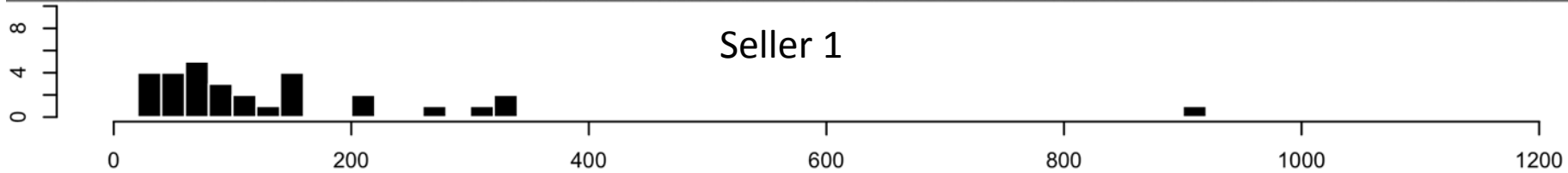
- Correlation

$$r = -0.08$$







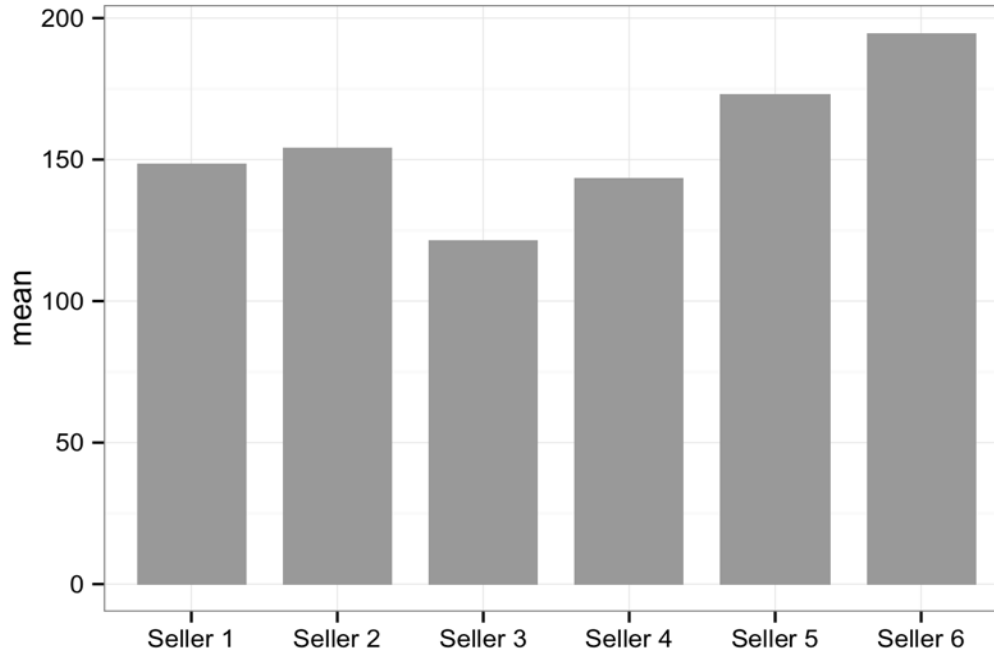


## Average Sales

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

## Average Sales

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

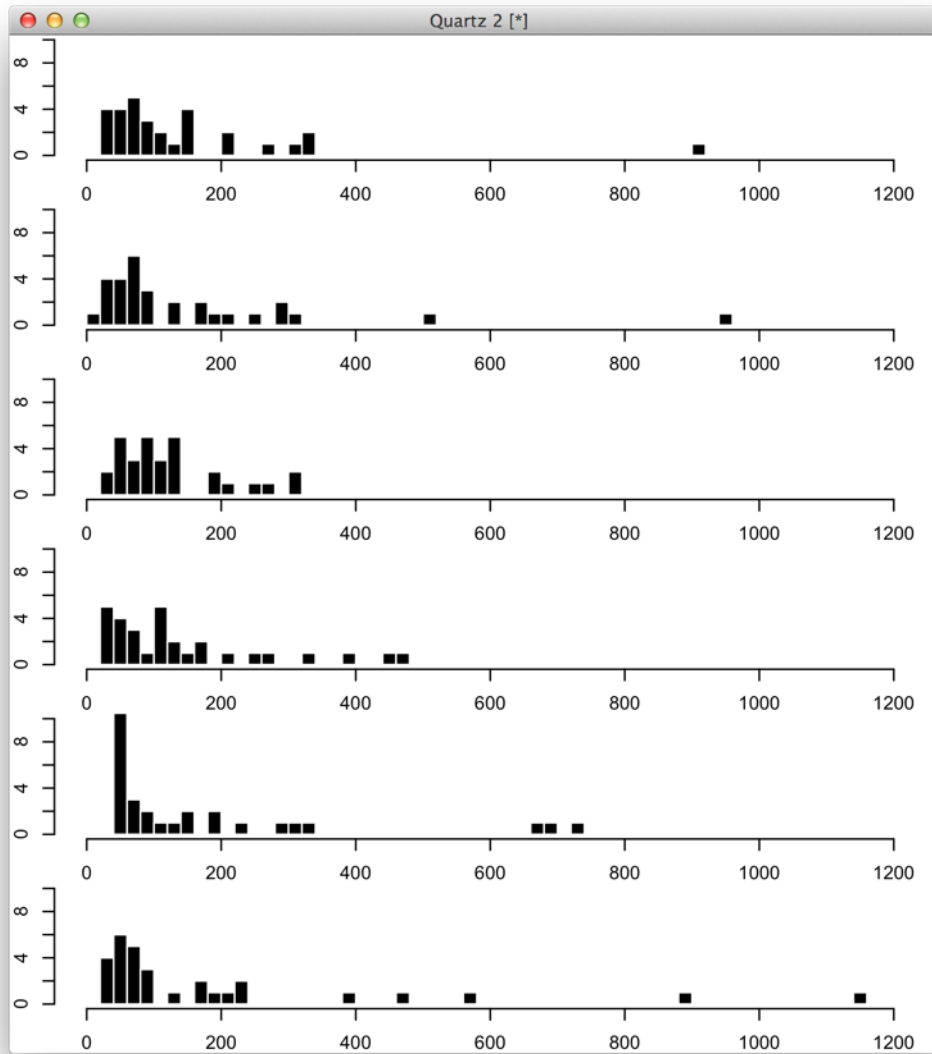




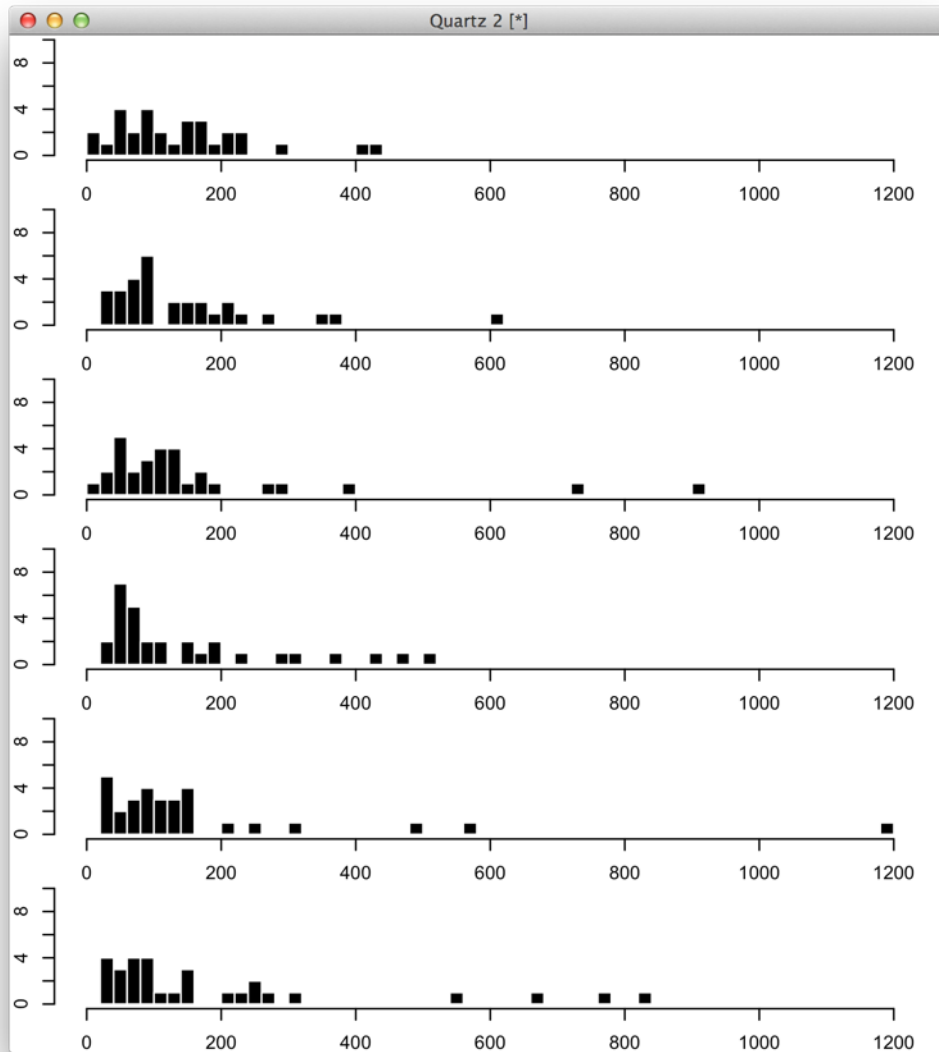
How much can we trust this chart?

LET US TRAVEL TO THE FUTURE

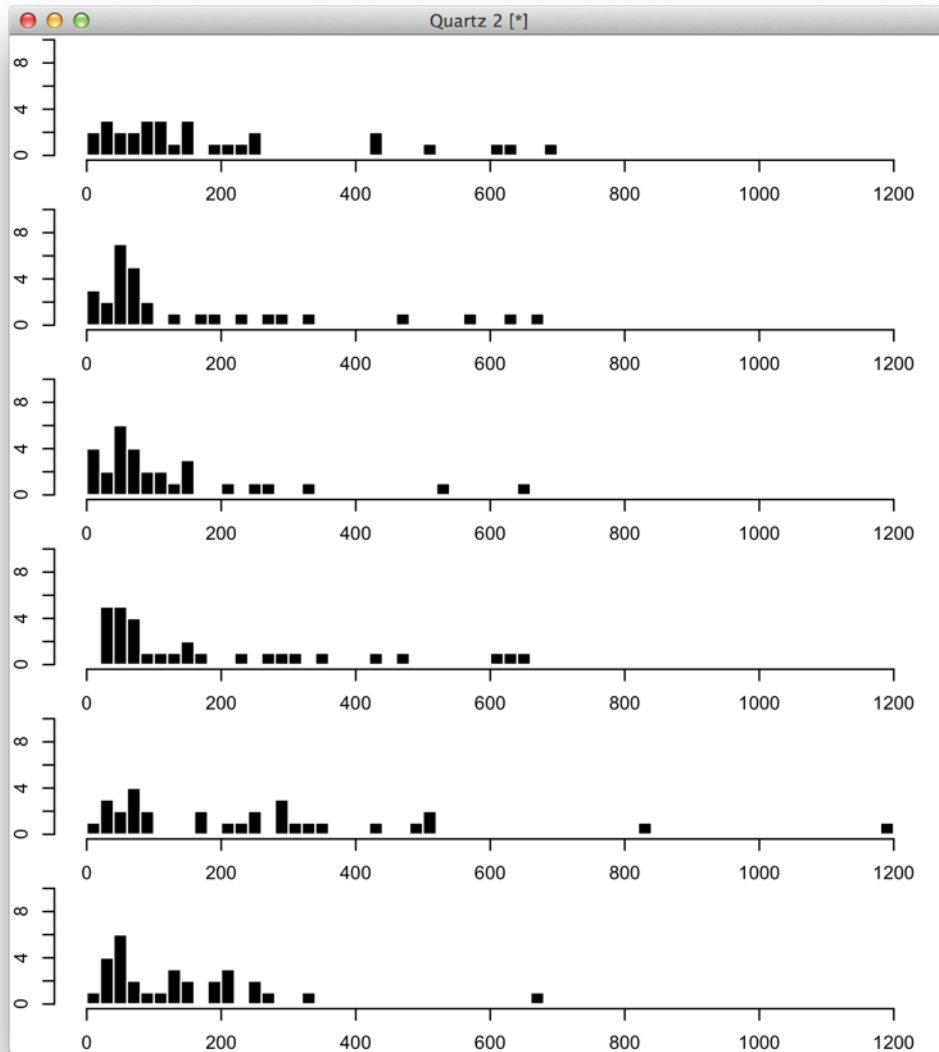
September 2014



October 2014

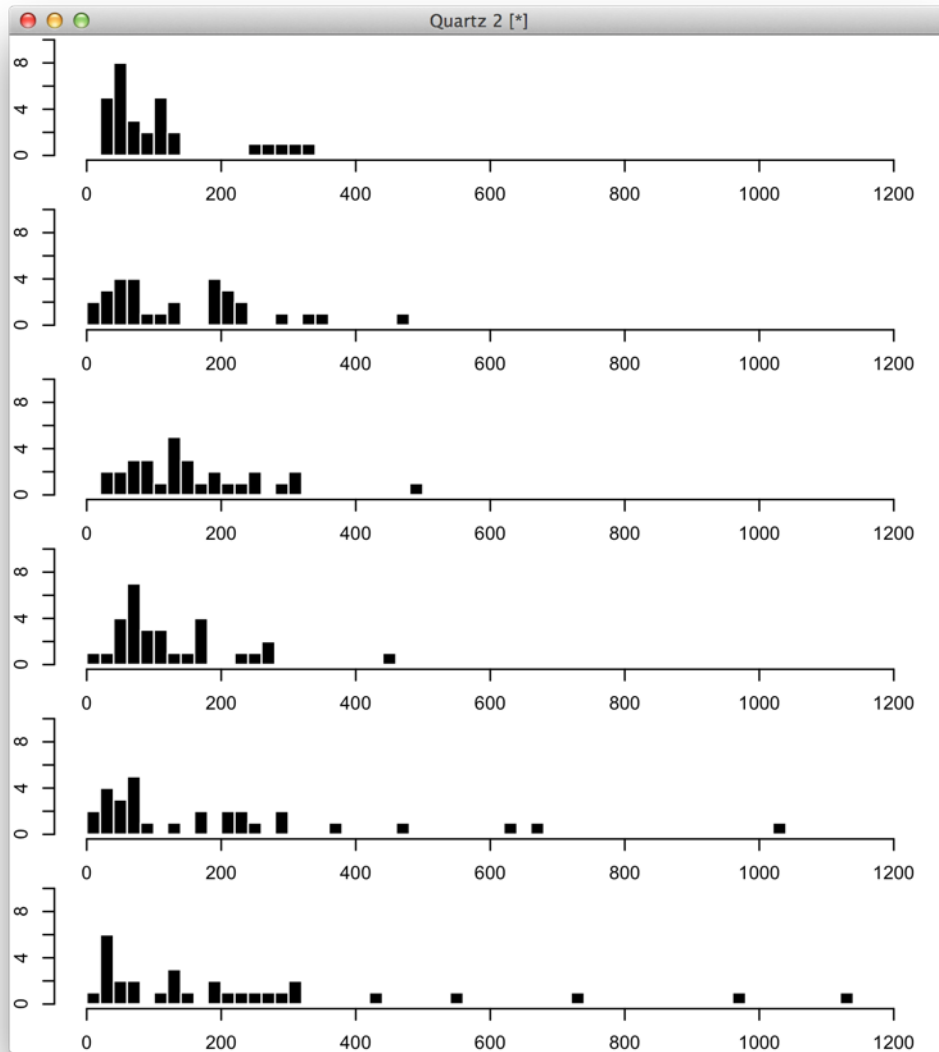


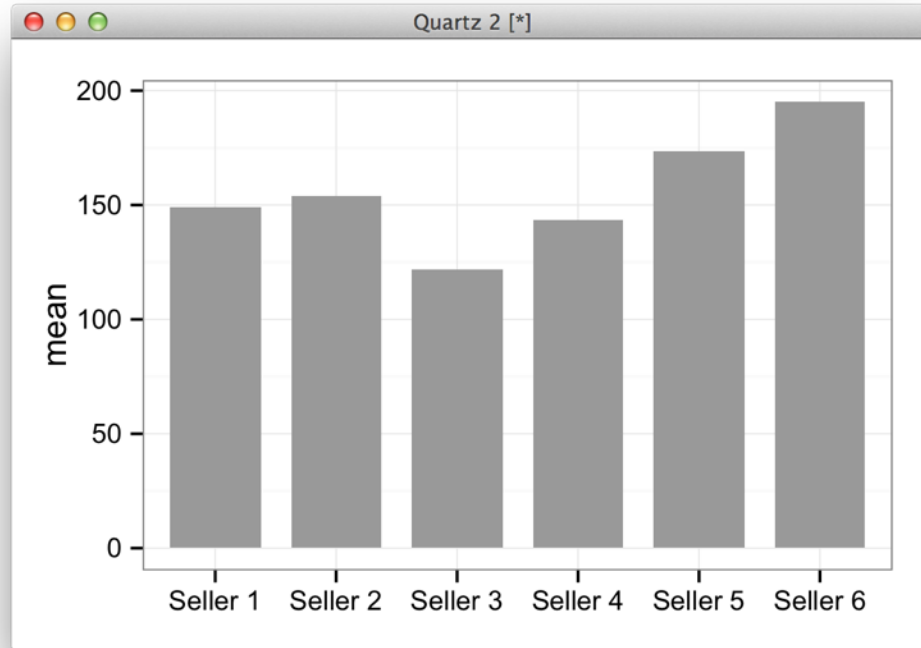
November 2014

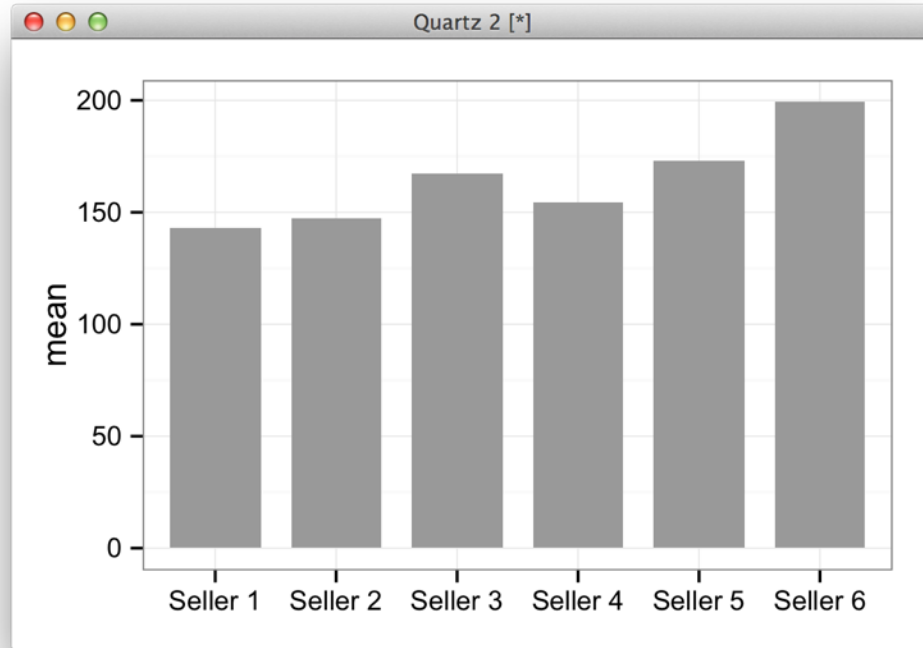




December 2014











BACK TO THE PRESENT

September 2014

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



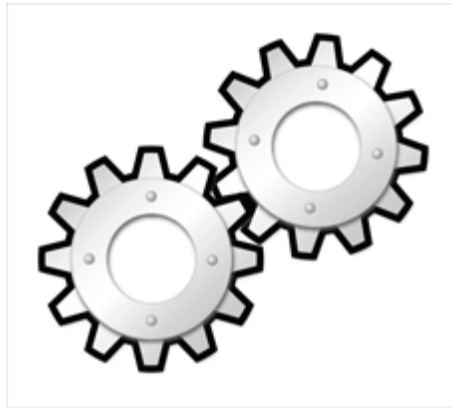
How much can we trust this chart?



# STATISTICAL TOOLS

DESCRIPTIVE STATISTICS

INFERENCE STATISTICS



# STATISTICAL INFERENCE

We want to know about these



Parameter  $\mu$

(Population mean)

We have these to work with



$\bar{x}$  Statistic

(Sample mean)

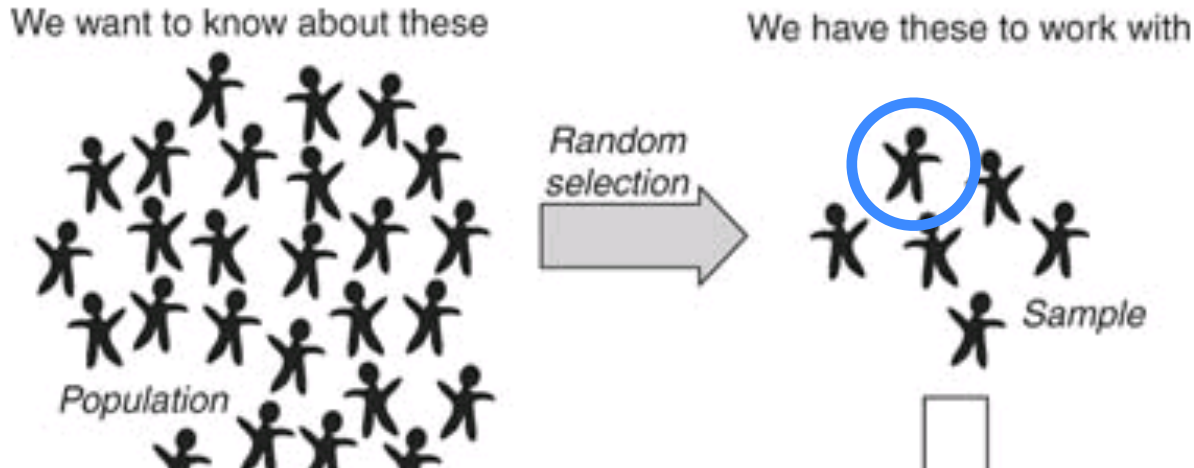


# STATISTICAL INFERENCE

- Terminology:
  - **Sample vs. population**
  - Mean, median, standard deviation, correlation, etc:
    - A sample **statistic** (e.g.,  $M$ )
    - A population **parameter** (e.g.,  $\mu$ )

# STATISTICAL INFERENCE

- Unit of statistical analysis

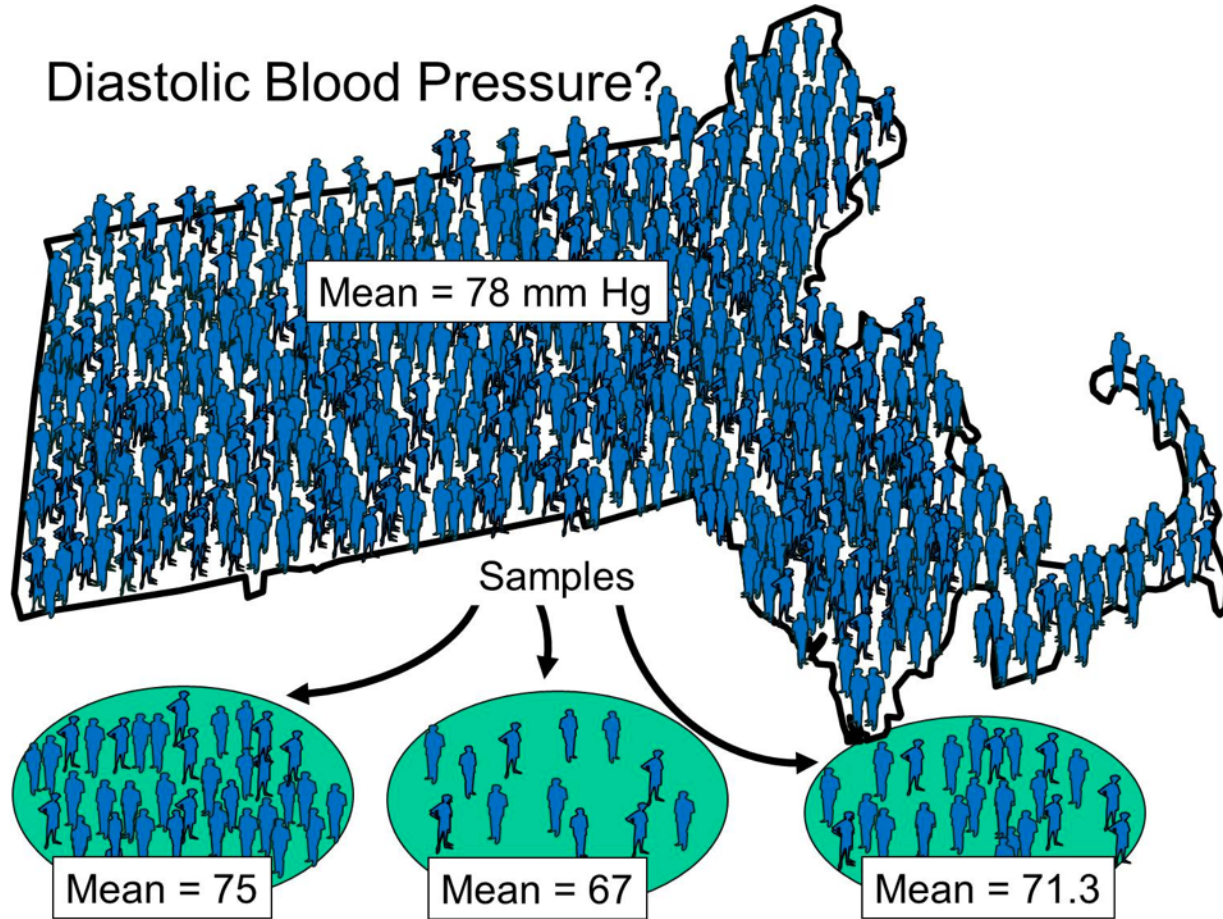


= “the thing that I’m sampling from a larger population”



What is the unit of statistical analysis?

# SAMPLING ERROR

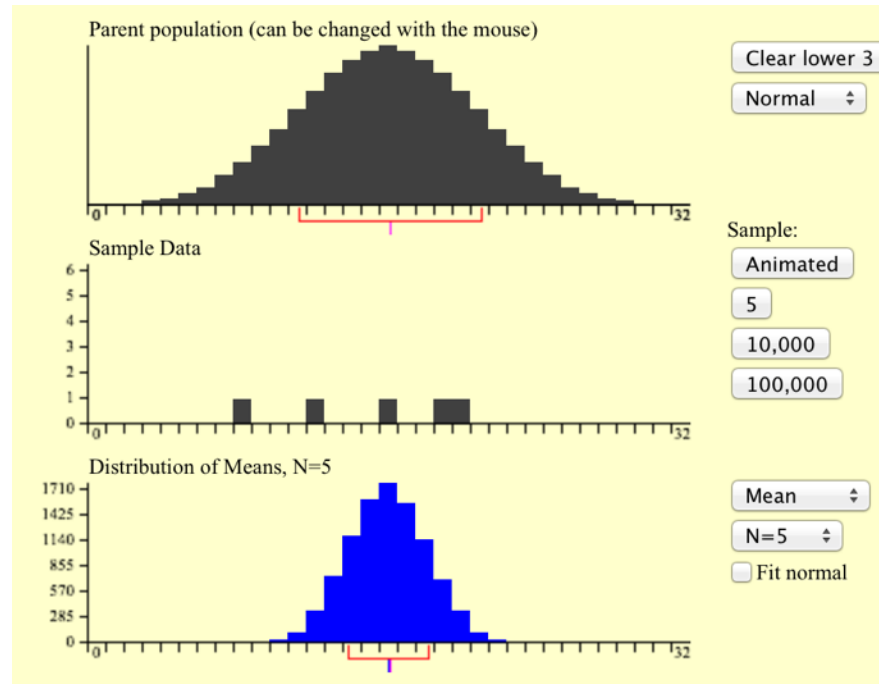


# SAMPLING DISTRIBUTION

- *"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"*

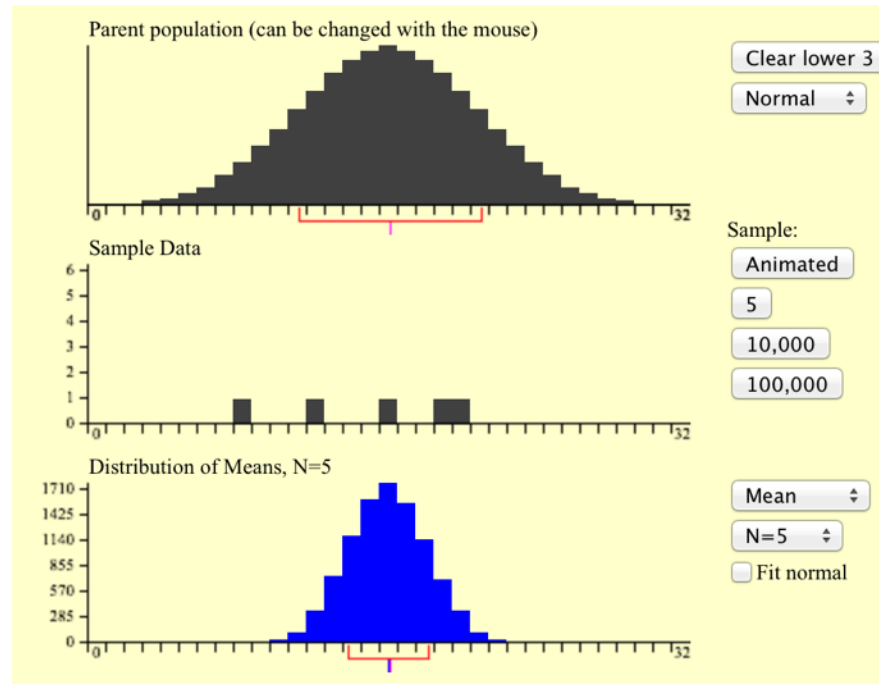
# SAMPLING DISTRIBUTION

- Demo [http://onlinestatbook.com/stat\\_sim/sampling\\_dist/](http://onlinestatbook.com/stat_sim/sampling_dist/)



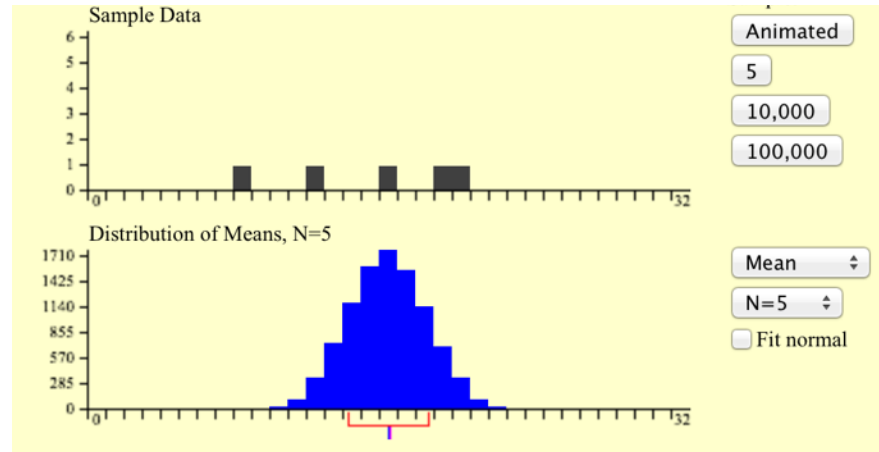


# SAMPLING DISTRIBUTION



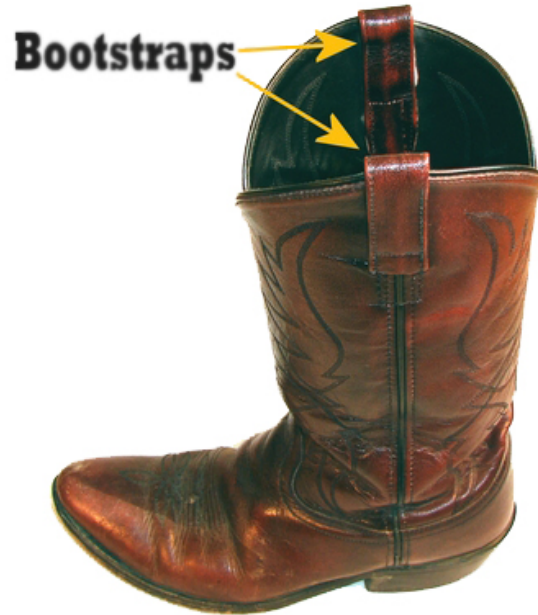
# SAMPLING DISTRIBUTION

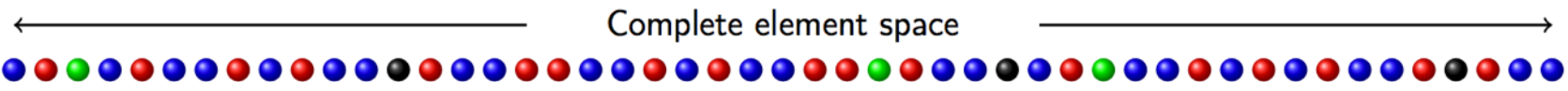
- But we don't know the population distribution!

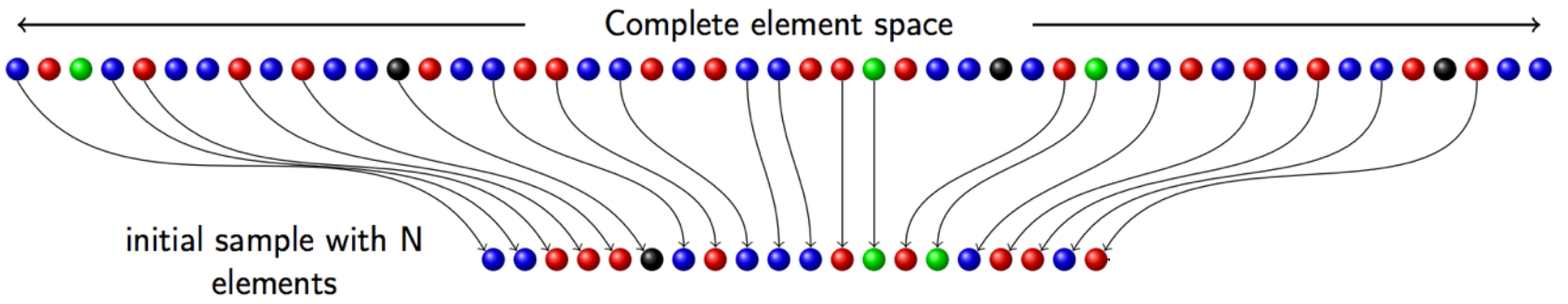


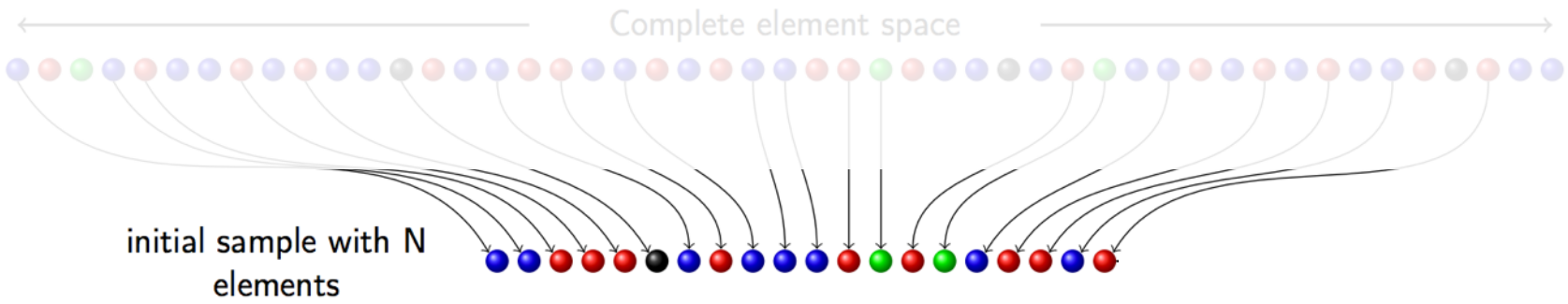
# SAMPLING DISTRIBUTION

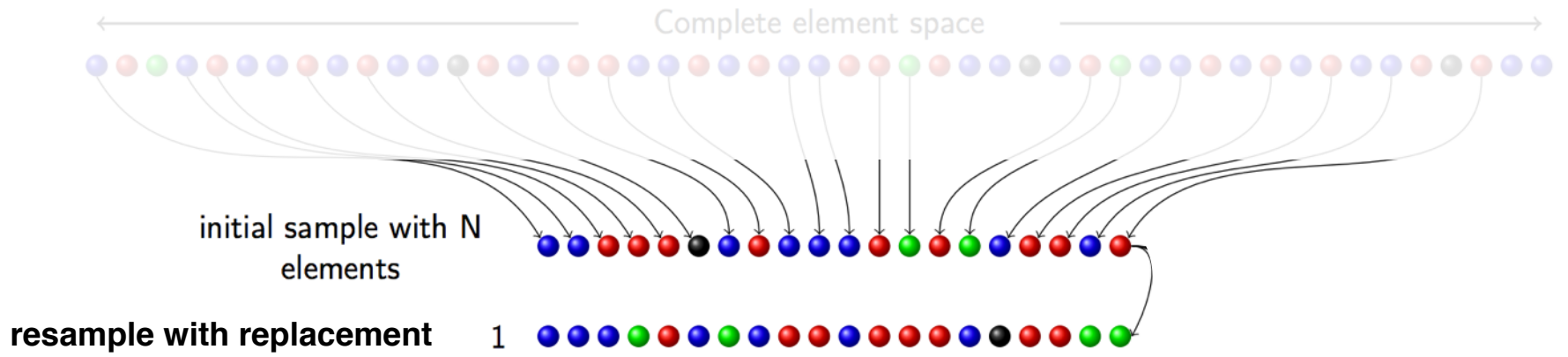
- Resampling techniques
  - Bootstrapping

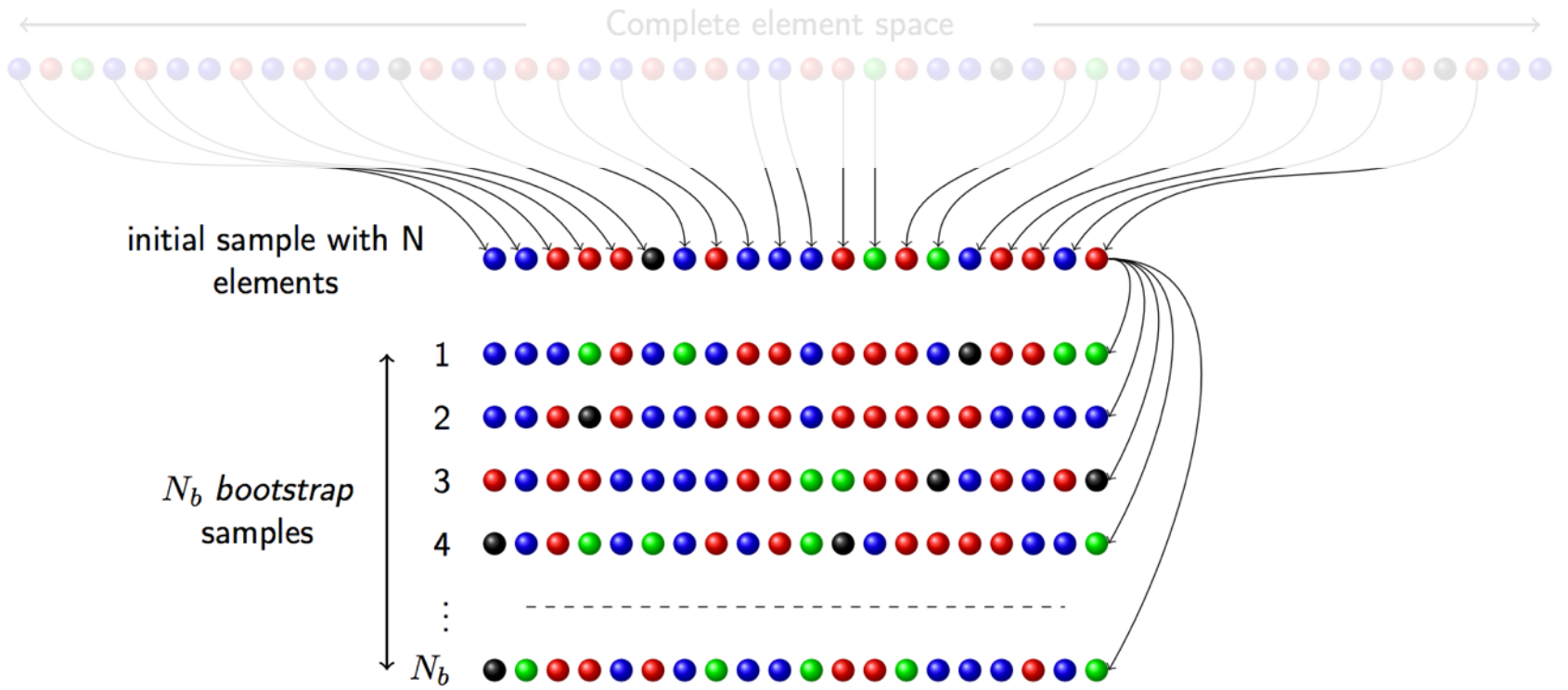




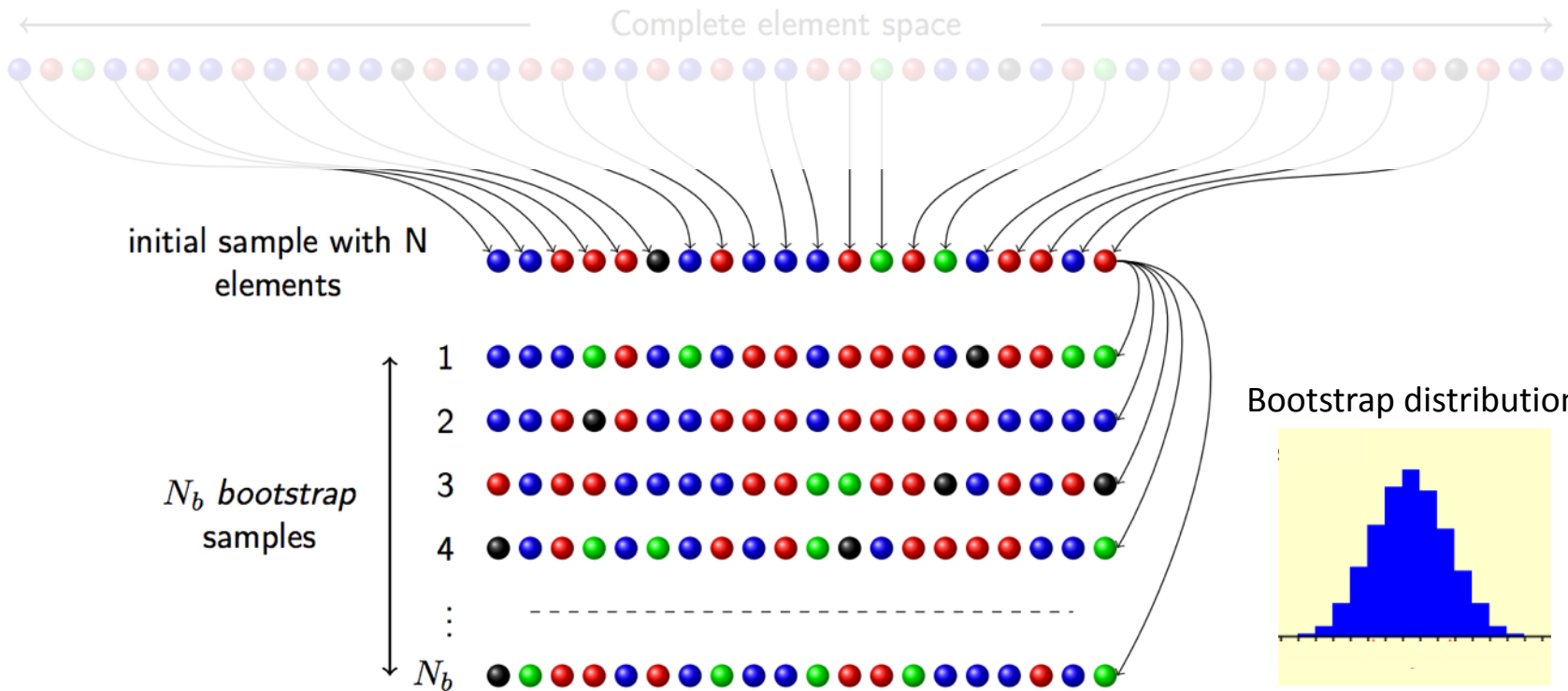


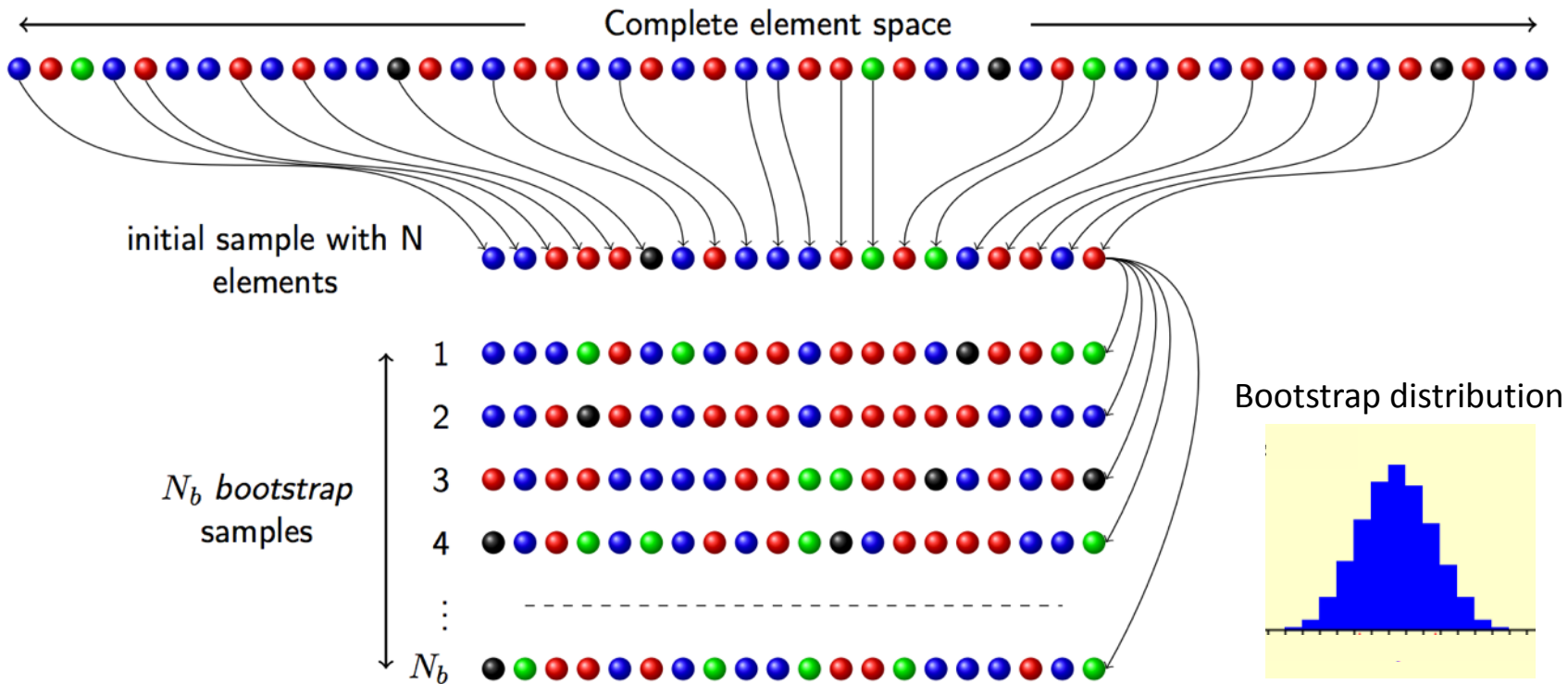










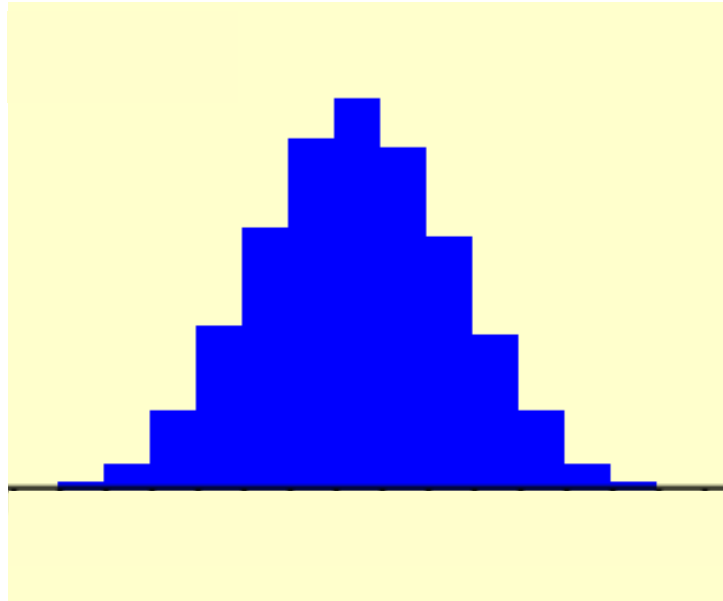


**Theorem (B. Efron, Ann. Statist. 1979)**

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.

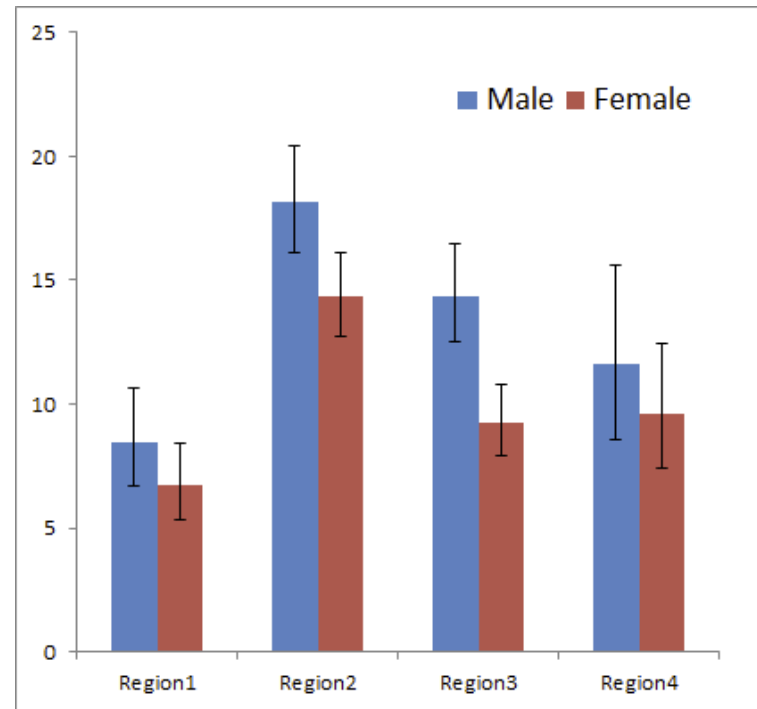
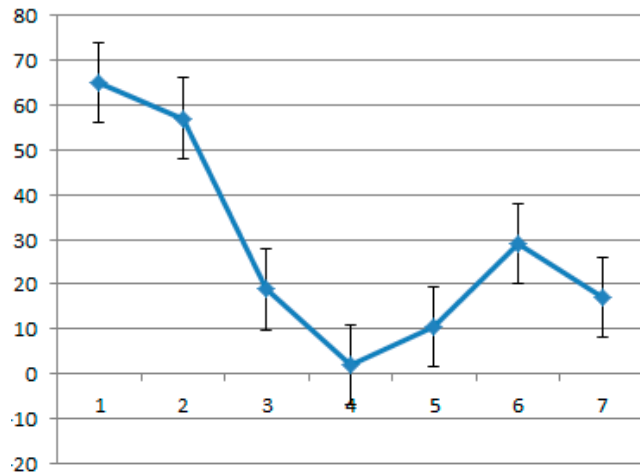
# SAMPLING DISTRIBUTION

- How to summarize a sampling distribution?

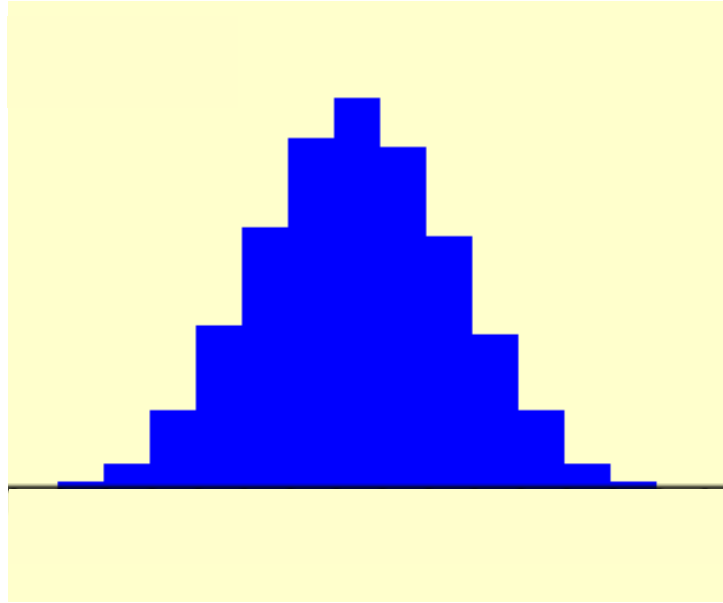


# SAMPLING DISTRIBUTION

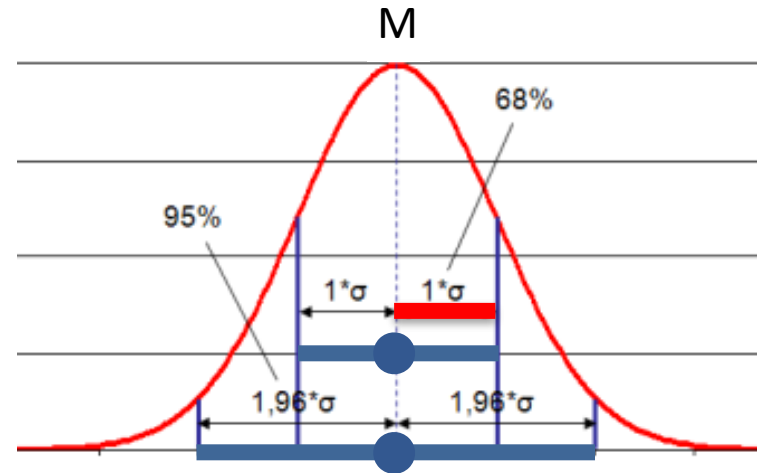
- How to summarize a sampling distribution?
- With an error bar



# SAMPLING DISTRIBUTION



# SAMPLING DISTRIBUTION

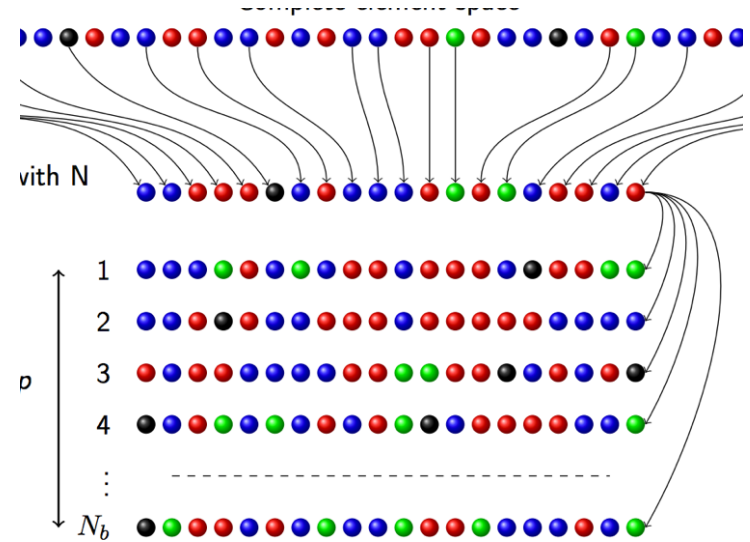
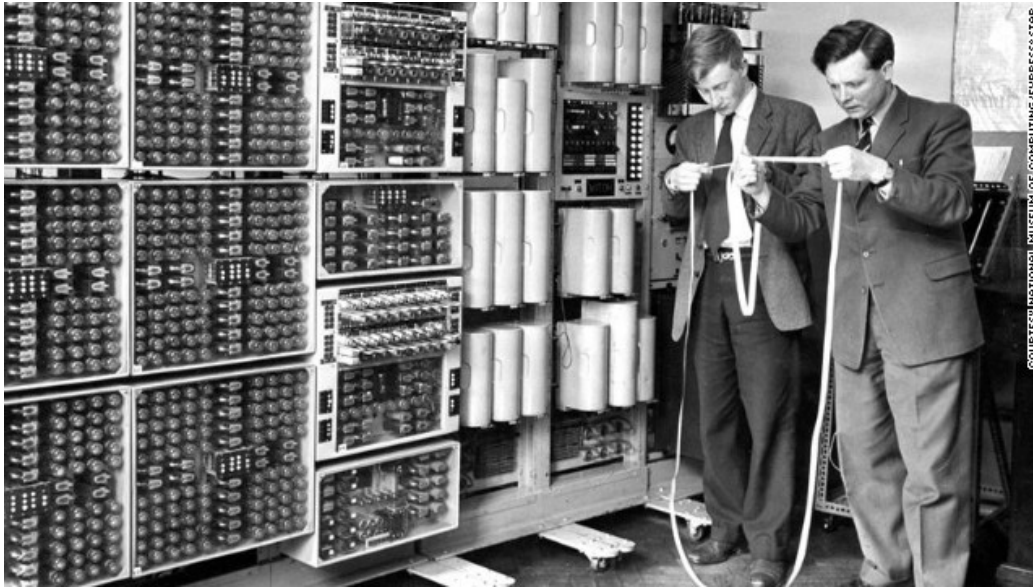


Standard error

95% confidence interval

# SAMPLING DISTRIBUTION

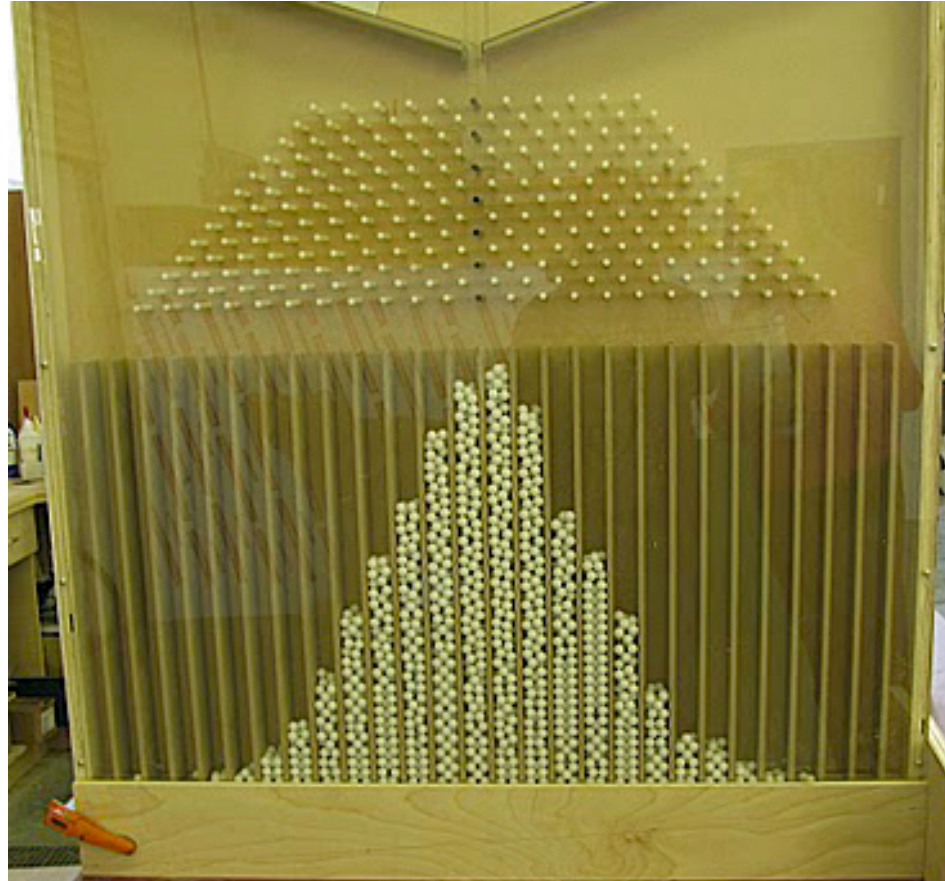
- How did people do before computers?



# NORMAL DISTRIBUTION

- Sir Francis Galton  
1822 – 1911

Bean Machine  
or Galton Board:





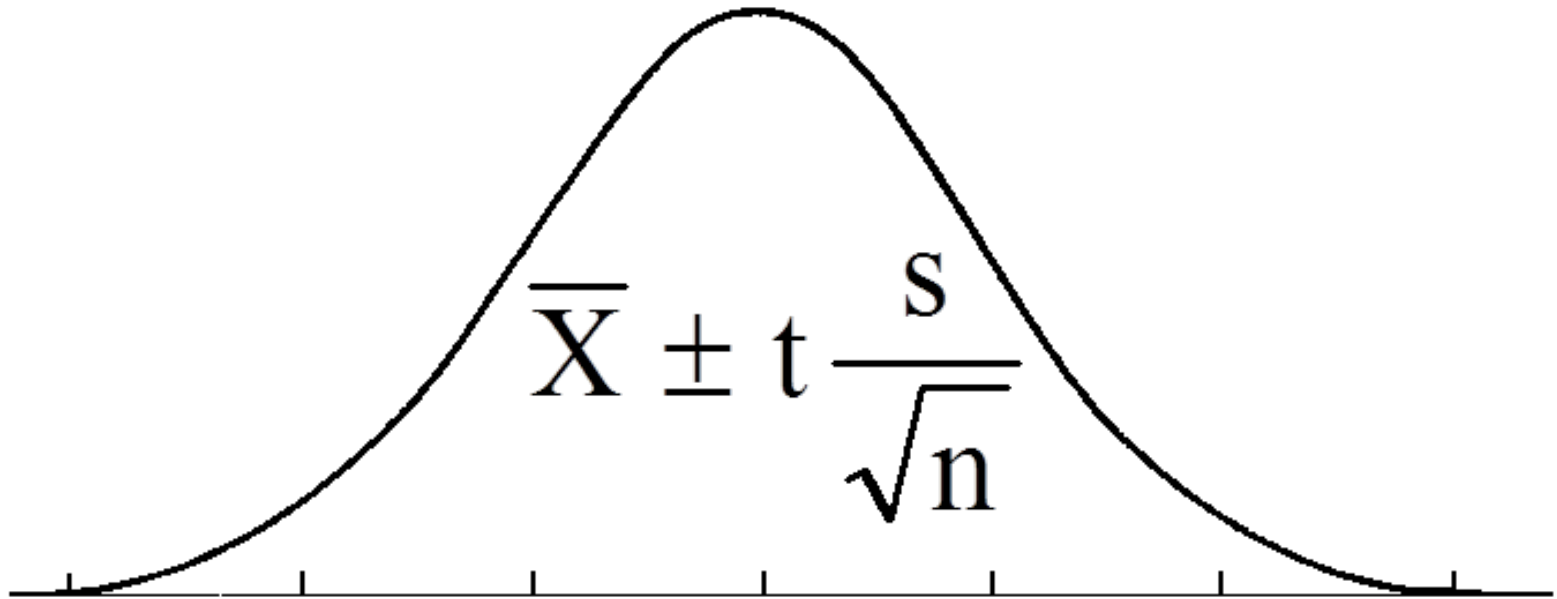
# NORMAL DISTRIBUTION

## 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

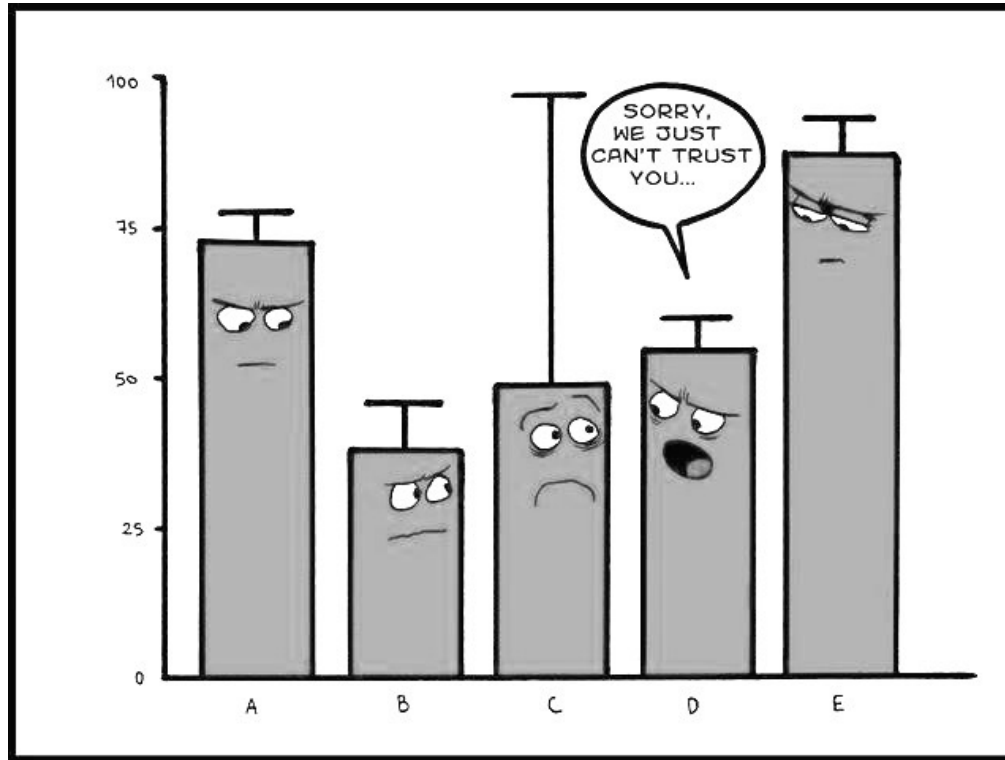
# NORMAL DISTRIBUTION

“Exact” t-based confidence intervals

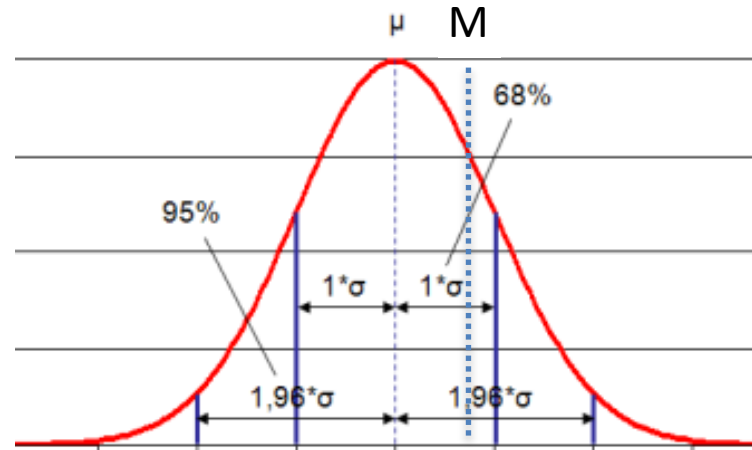


$t \sim 1.96$  for large samples

# CONFIDENCE INTERVALS

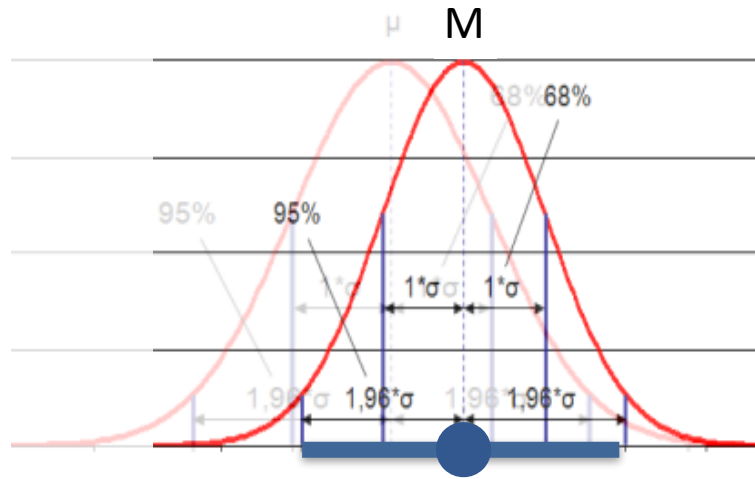


# CONFIDENCE INTERVALS



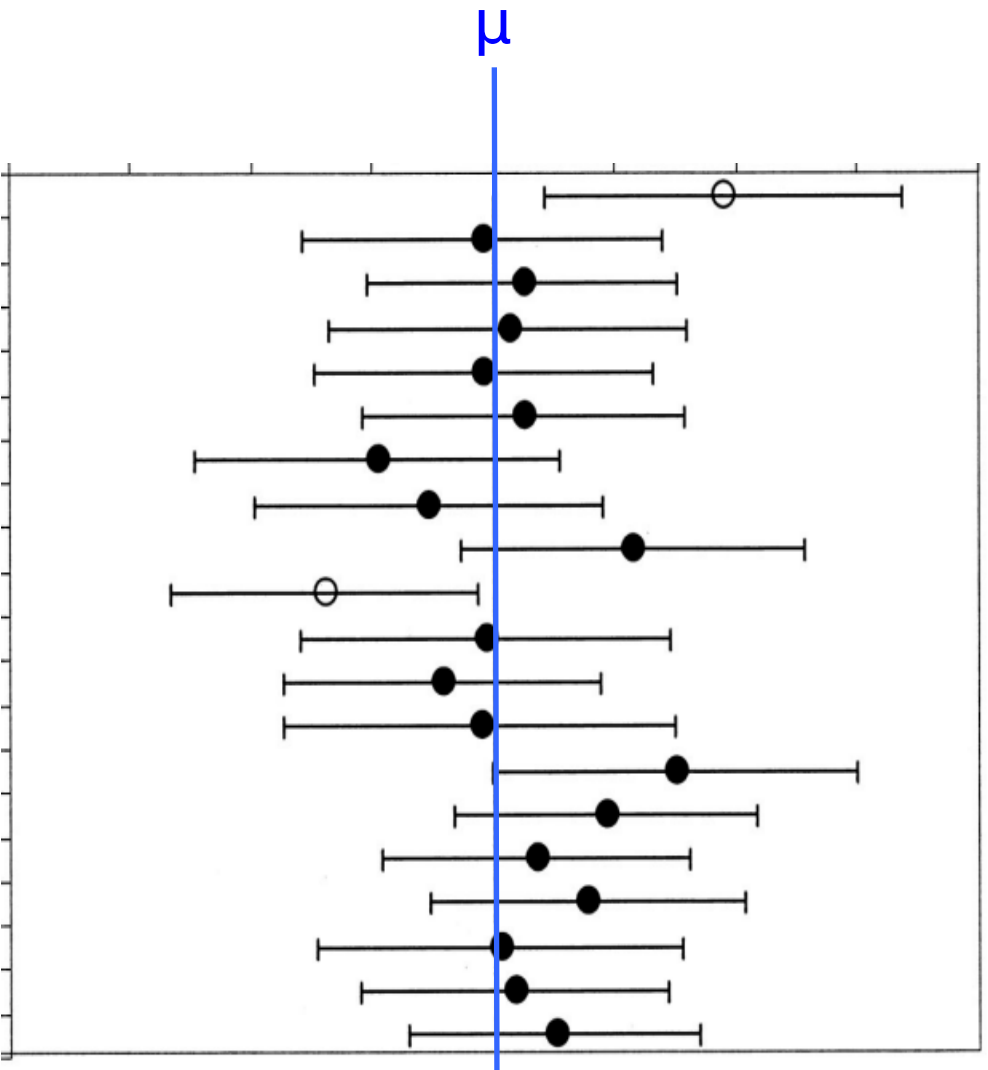
True sampling distribution

# CONFIDENCE INTERVALS



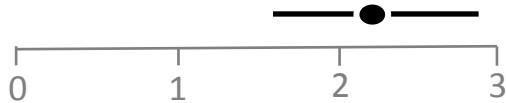
95% confidence interval

Different random samples



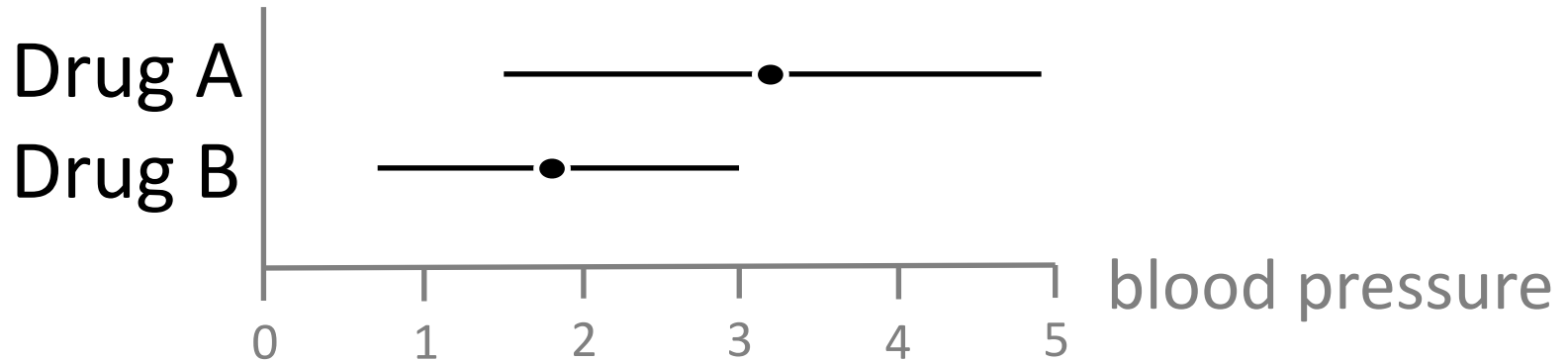
# CONFIDENCE INTERVALS

- Several interpretations
- « *a range of plausible values for  $\mu$ . 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



# CONFIDENCE INTERVALS

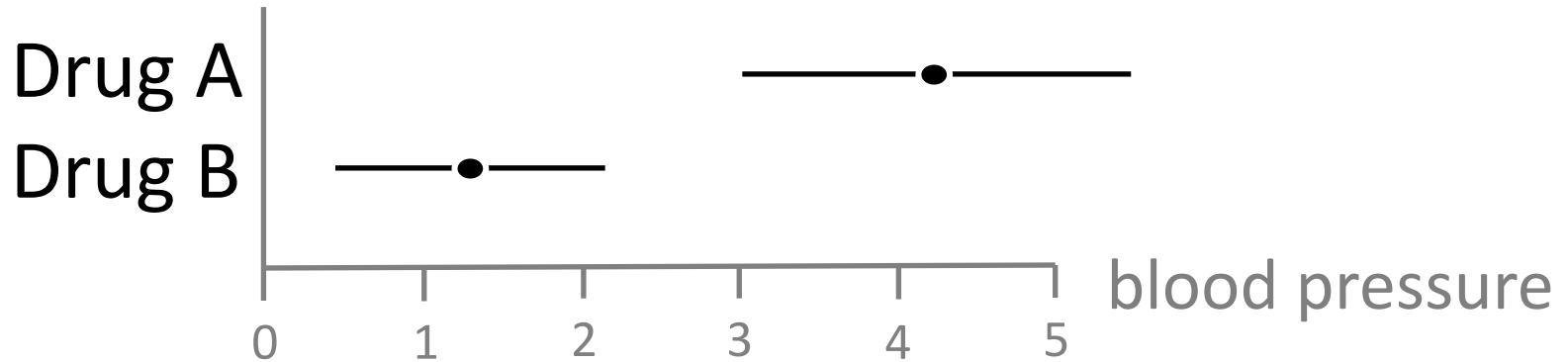
- « *a range of plausible values for  $\mu$ . Values outside the CI are relatively implausible.* »  
(Cumming and Finch, 2005)





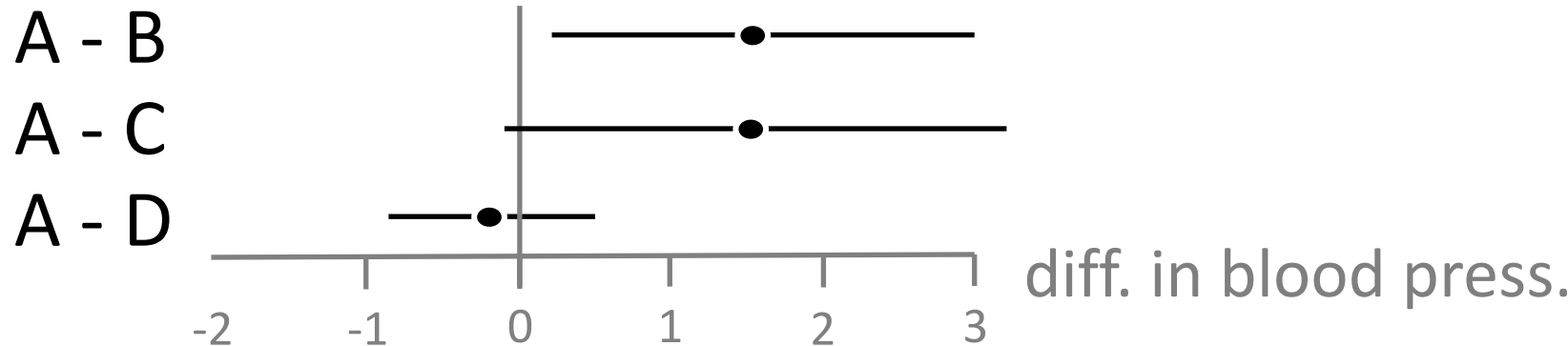
# CONFIDENCE INTERVALS

- « *a range of plausible values for  $\mu$ . Values outside the CI are relatively implausible.* »  
(Cumming and Finch, 2005)



# CONFIDENCE INTERVALS

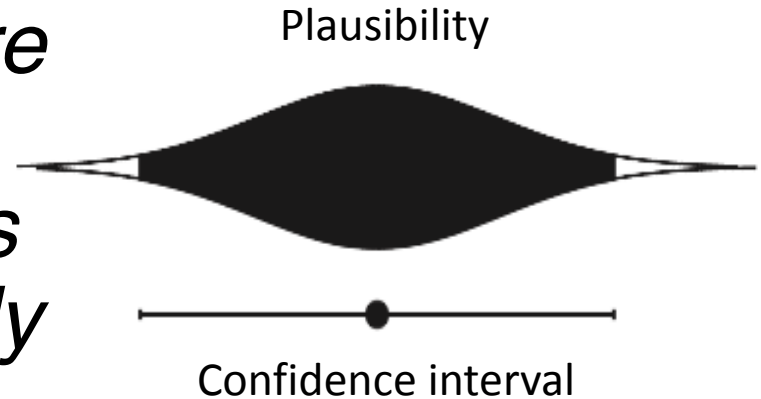
- « *a range of plausible values for  $\mu$ . Values outside the CI are relatively implausible.* »  
(Cumming and Finch, 2005)



# CONFIDENCE INTERVALS

- *“values close to our  $M$  are the best bet for  $\mu$ , and values closer to the limits of our CI are successively less good bets.”*

(Cumming, 2013)



# BACK TO OUR EXAMPLE

- Selling encyclopedias



## Average Sales

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

