Real Estate in Jeff Bezos's Neighborhood

A lesson in robustness in statistics

CHIDATA

Recap: Significance

Data 1

Data 2

Is the difference between two datasets just random chance?

Recap: Two-Sample Z-Test

Compares the means of two samples!

Sample 1

$$\bar{\mu}_1, \bar{\sigma}_1^2, K_1$$

Calculate mean, variance,
and size of sample

Sample 2

$$\bar{\mu}_2, \bar{\sigma}_2^2, K_2$$

Calculate mean, variance,
and size of sample

$$z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}}$$

Look up p-value via stats library

Recap: Two-Sample Z-Test

Compares the means of two samples!

Sample 1

$$\bar{\mu}_1, \bar{\sigma}_1^2, K_1$$

Calculate mean, variance,
and size of sample

Sample 2

$$\bar{\mu}_2, \bar{\sigma}_2^2, K_2$$

Calculate mean, variance,
and size of sample

Null Hypothesis: no difference in means only error due to sampling

p-value: how likely is the observed difference assuming the null

Recap: Two-Sample Z-Test

Compares the means of two samples!

Sample 1

$$\bar{\mu}_1, \bar{\sigma}_1^2, K_1$$

Calculate mean, variance, and size of sample

Sample 2

$$\bar{\mu}_2, \bar{\sigma}_2^2, K_2$$

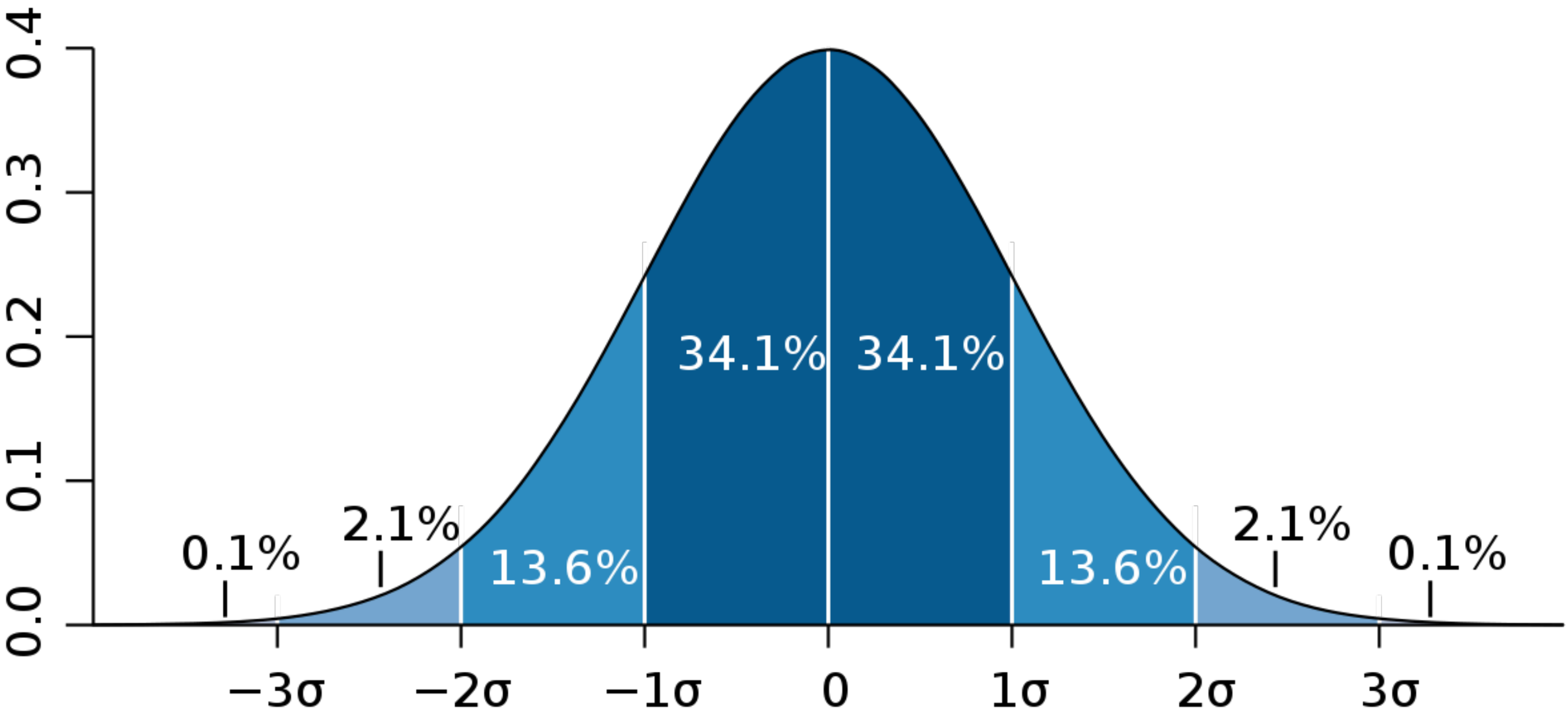
Calculate mean, variance, and size of sample

$$z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}}$$

Mean/Variance key parameters for modeling

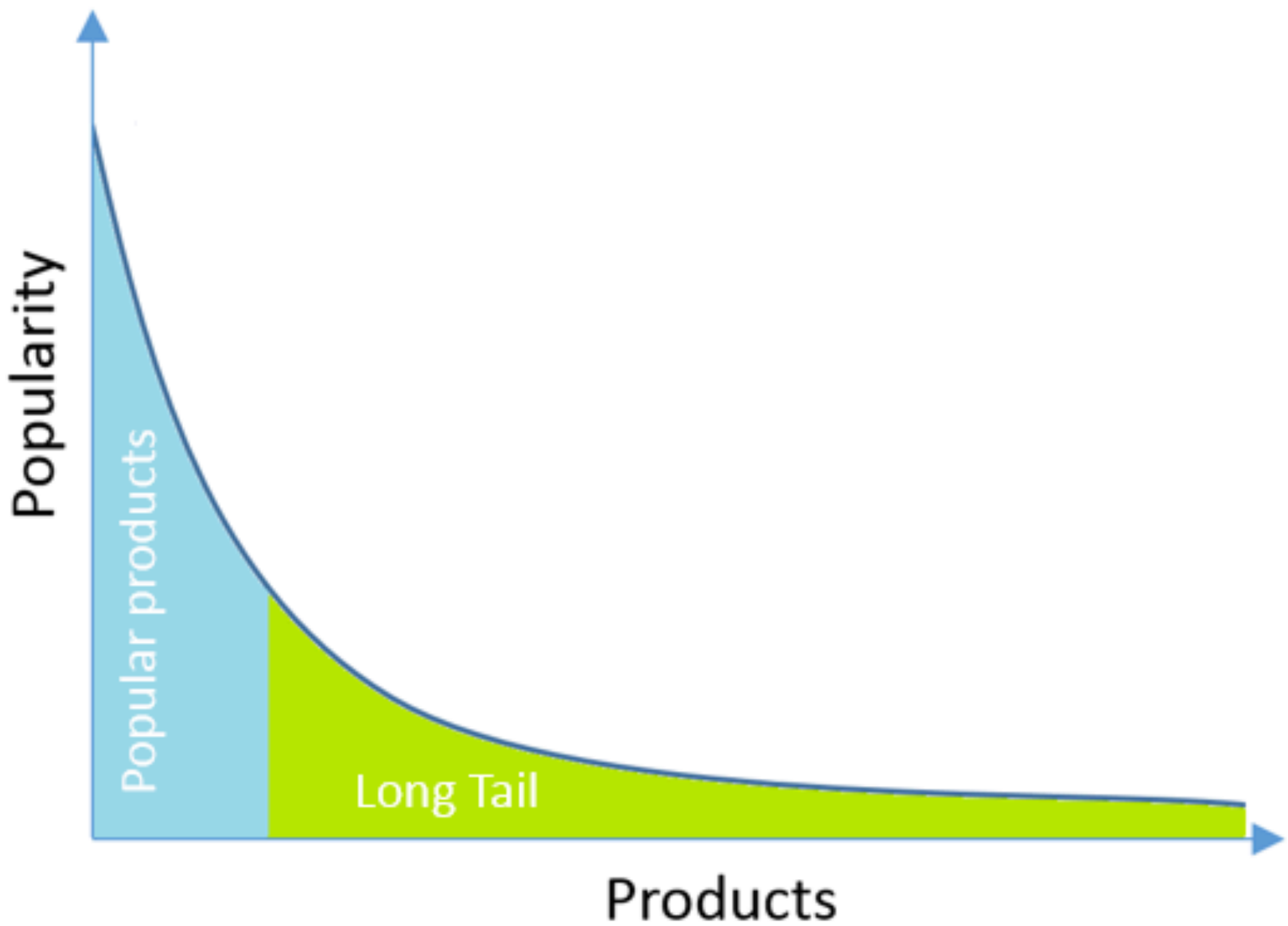
2 Standard Deviations From The Mean

Normal Distribution: ~95%



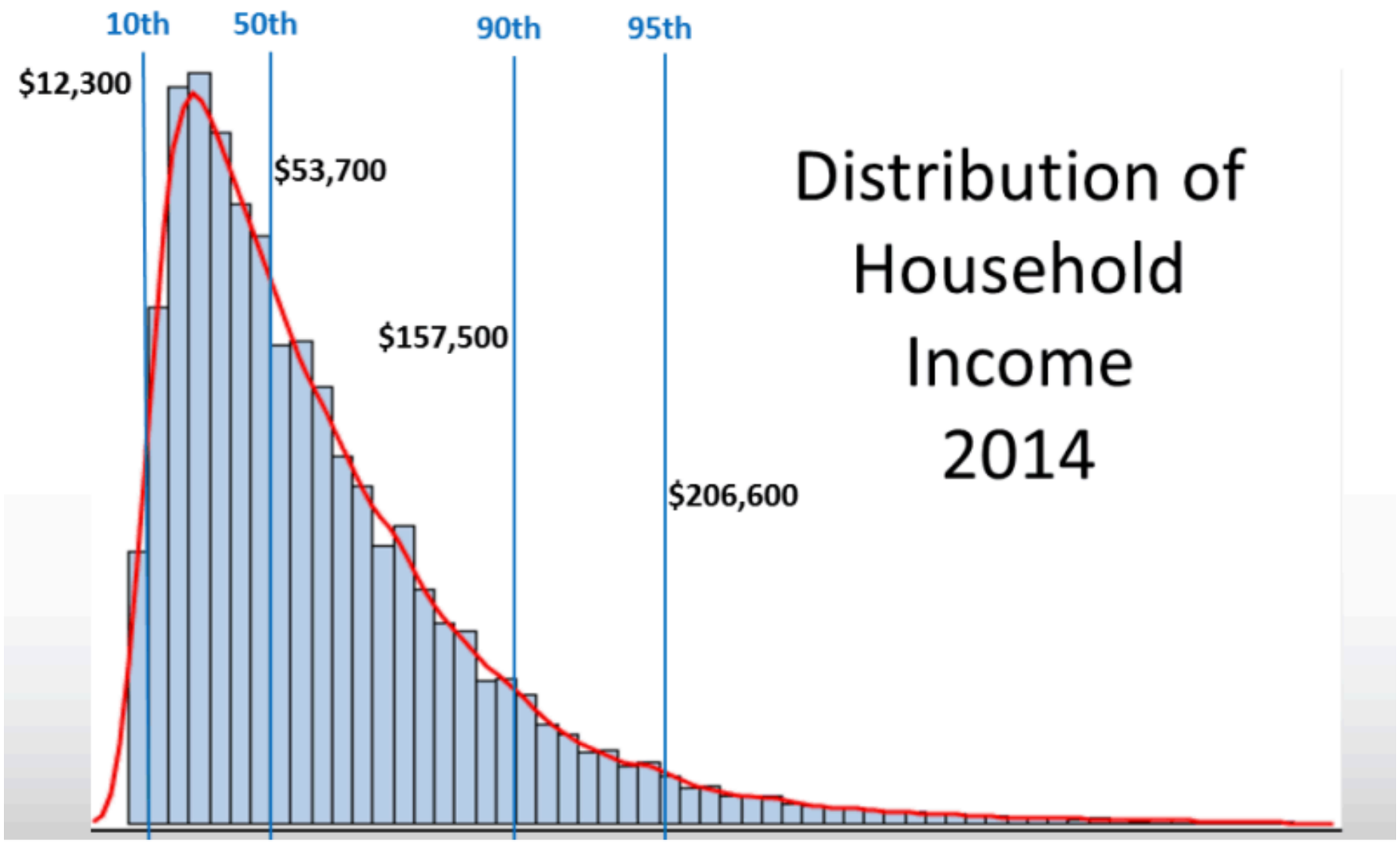
Long "Tailed" Distributions

Not all distributions are so "even"



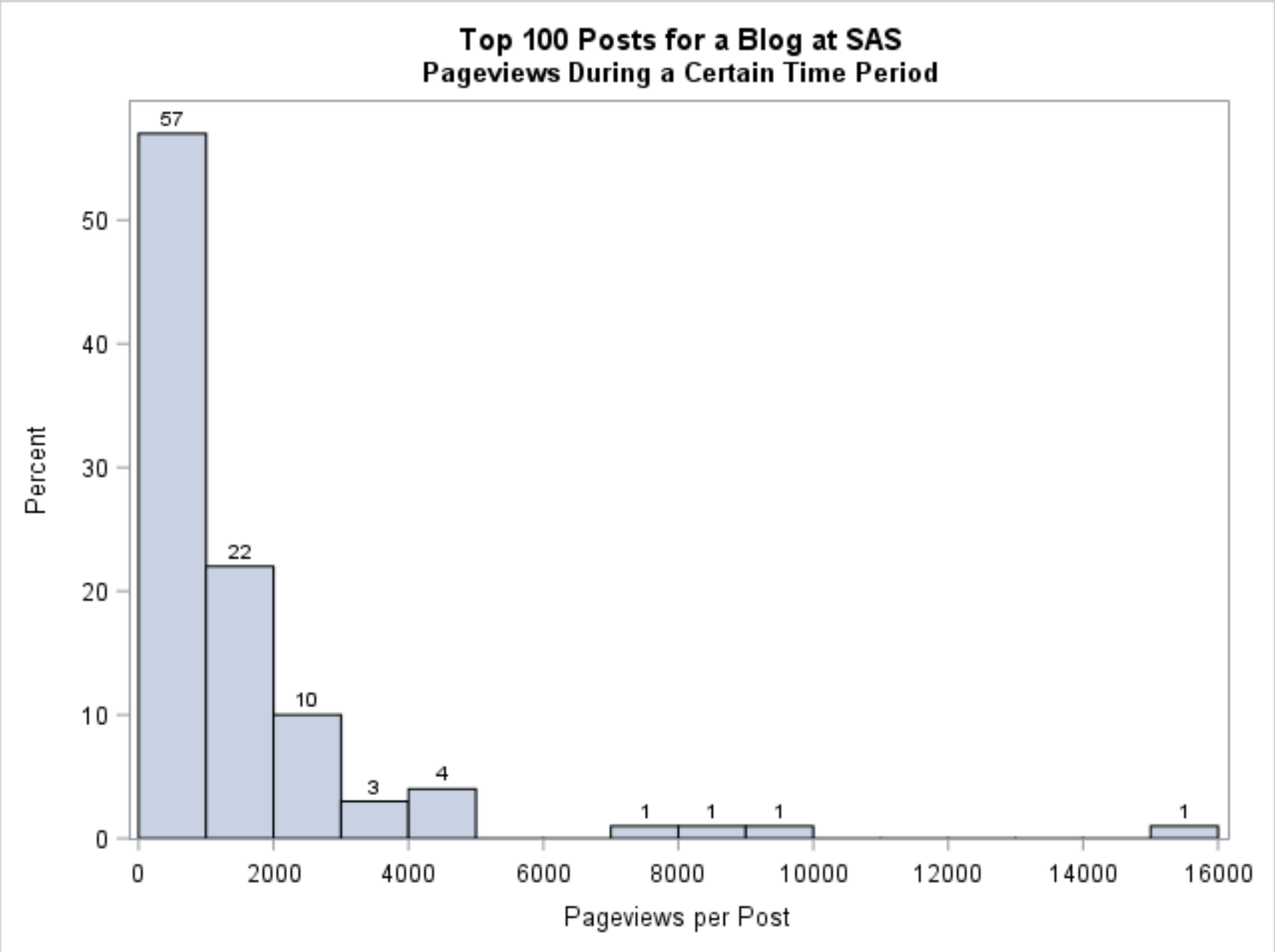
Long "Tailed" Distributions

Not all distributions are so "even"



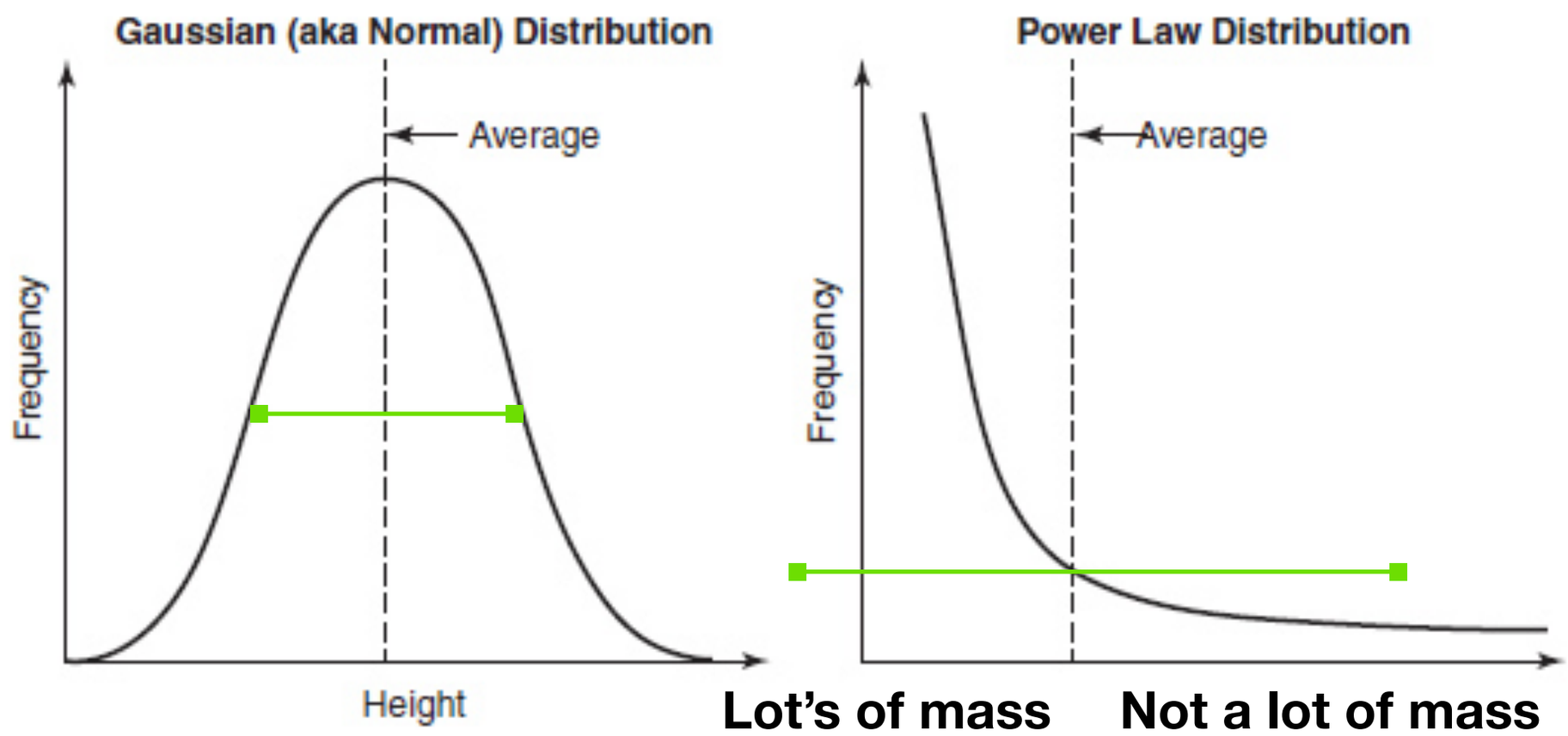
Long "Tailed" Distributions

Not all distributions are so "even"



Means/Standard Deviations

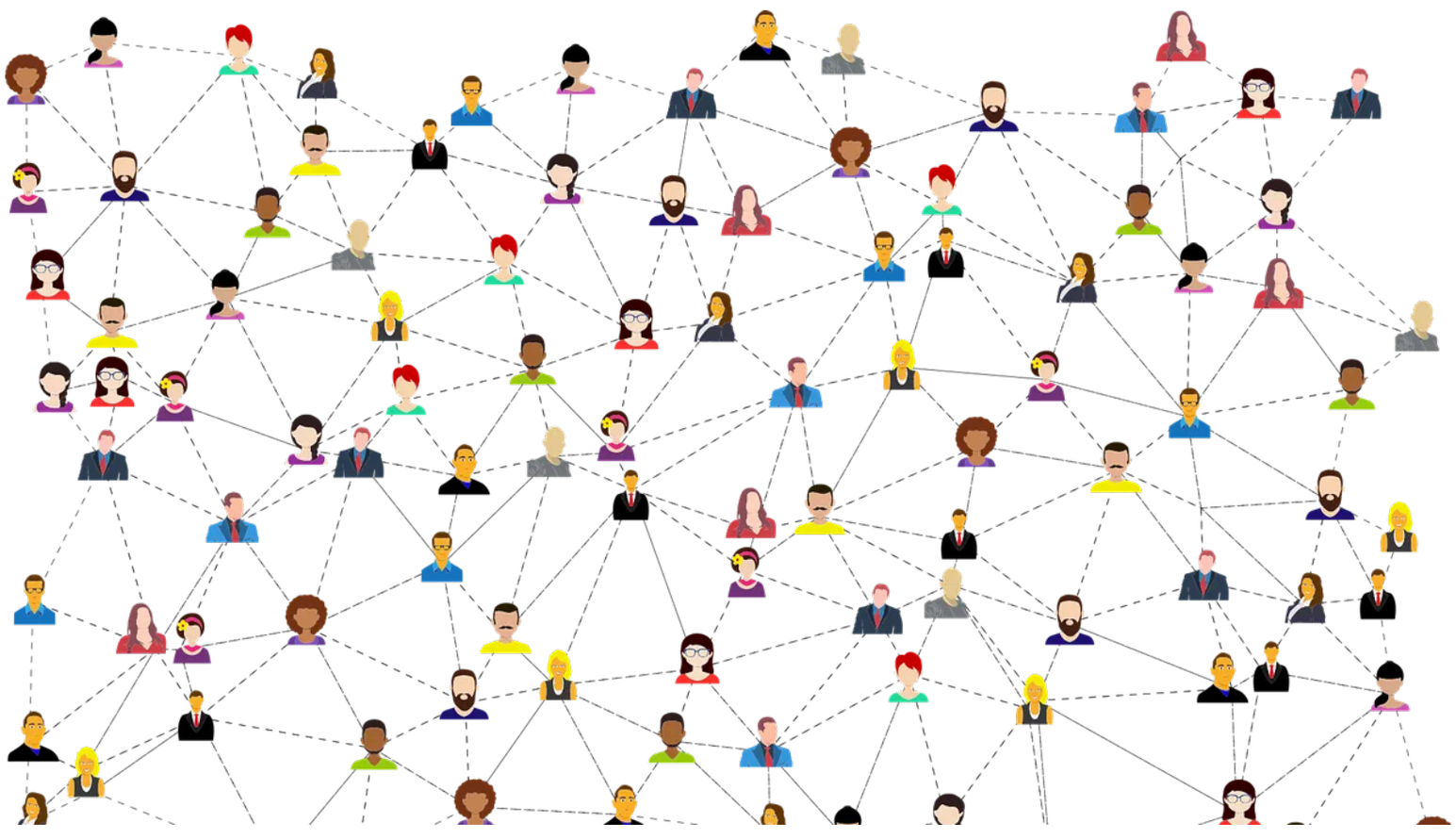
Mean and Standard Deviations Are Misleading For Long-Tailed Distributions.



Rich Get Richer Effects

Preferential Attachment

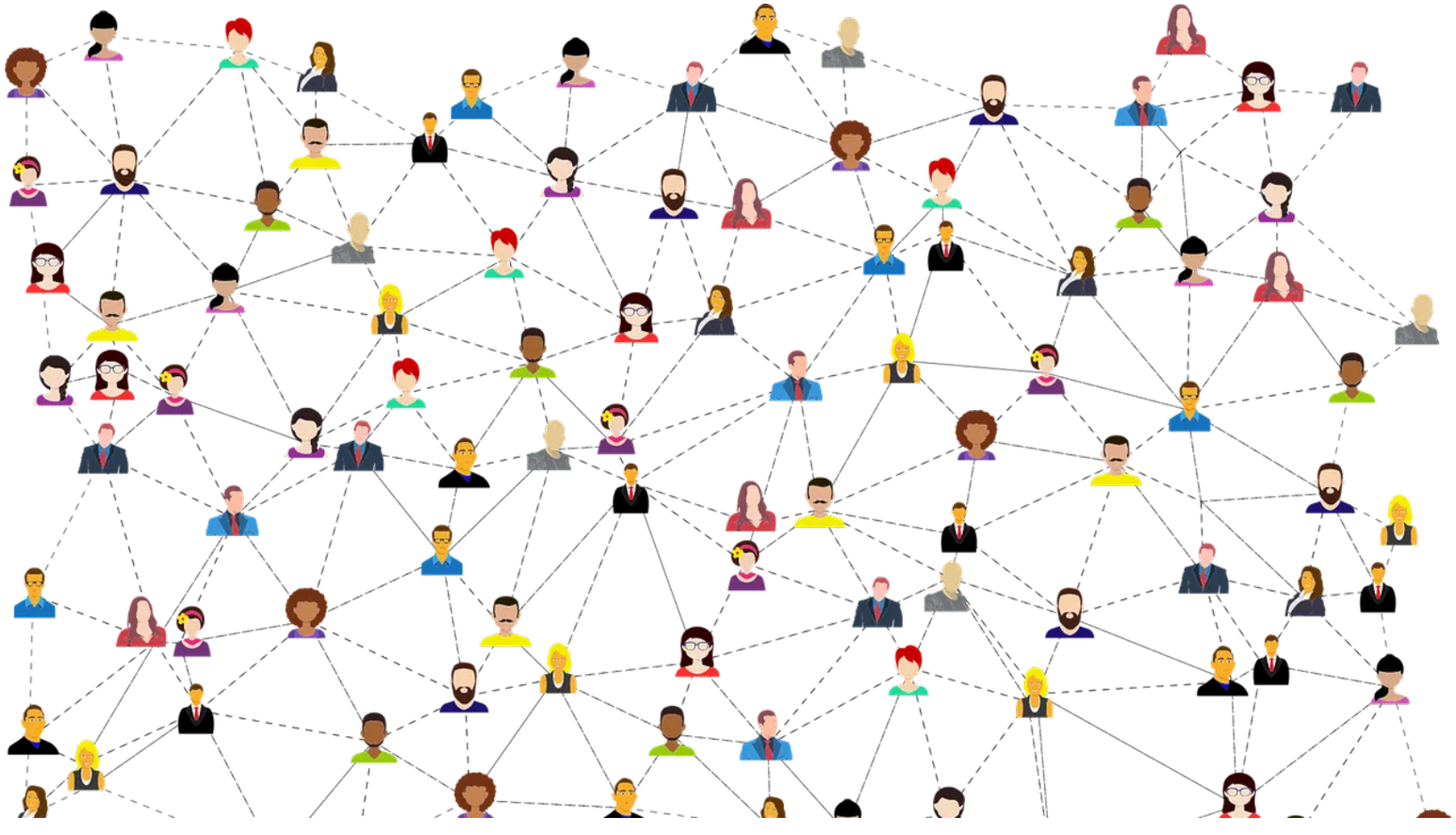
New users in a social network attach to already popular existing users.



Rich Get Richer Effects

Preferential Attachment

People with wealth have a greater ability to grow their wealth



Why Do These Happen?

Normal Distribution

Central Limit Theorem! Average of independent factors

- Sampling error
- Sensor/electrical noise
- Human height

Long-Tailed Distributions

Preferential attachment

- Wealth/income
- Social Networks/Internet
- Word usage

What do we do?

How to compare the samples?

Sample 1

Long-tailed distribution

Sample 2

Long-tailed distribution

Wilcoxon Rank-Sum Test

Compare "average rank" rather than the mean value

Sample 1

7.6	6
3.1	3
4.5	5

Avg-rank: 14/3

Sample 2

0.6	1
3.9	4
2.4	2

Avg-rank: 7/3

Wilcoxon Rank-Sum Test

Null Hypothesis: Assume that data is evenly spread out to both samples.

Expected rank of a randomly chosen sample:

$$\frac{K_1 + K_2 + 1}{2}$$

How far is actual "average rank" to this expectation?

Wilcoxon Rank-Sum Test

Compare "average rank" rather than the mean value

Sample 1

7.6	6
3.1	3
4.5	5

Avg-rank: 14/3

Null-rank: 7/2

Sample 2

0.6	1
3.9	4
2.4	2

Avg-rank: 7/3

Null-rank: 7/2

Wilcoxon Rank-Sum Test

Won't show in class but…you can derive a standard deviation as well.

More importantly

```
>>> from scipy.stats import ranksums
>>> sample1 = np.random.uniform(-1, 1, 200)
>>> sample2 = np.random.uniform(-0.5, 1.5, 300) # a shifted distribution
>>> ranksums(sample1, sample2)
RanksumsResult(statistic=-7.887059, pvalue=3.09390448e-15)  # may vary
```

Wilcoxon Rank Sum Test

Average rank is effectively comparing “Medians”

Sample 1

Long-tailed distribution

Sample 2

Long-tailed distribution

Compare two samples by their medians rather than the means.

Sanjay's $0.02

Comparing Normal'ish Populations

Two-sample z-test

- More intuitive results

Comparing All Other Populations

Wilcoxon Rank Sum Test

- More robust comparison

P.S I wouldn't fault you for always using a rank-sum test!

Same Trick Works For Correlation

Var1		Var2	
7.6	3	**7.1**	3
3.1	1	**2.6**	1
4.5	2	**5.3**	2
102.0	4	**11.7**	4

Corr(Var1, Var2) Low!

Corr(Ranked-Var1, Ranked-Var2) High!

“Spearman” Correlation

Measures the correlation in ranking between variables

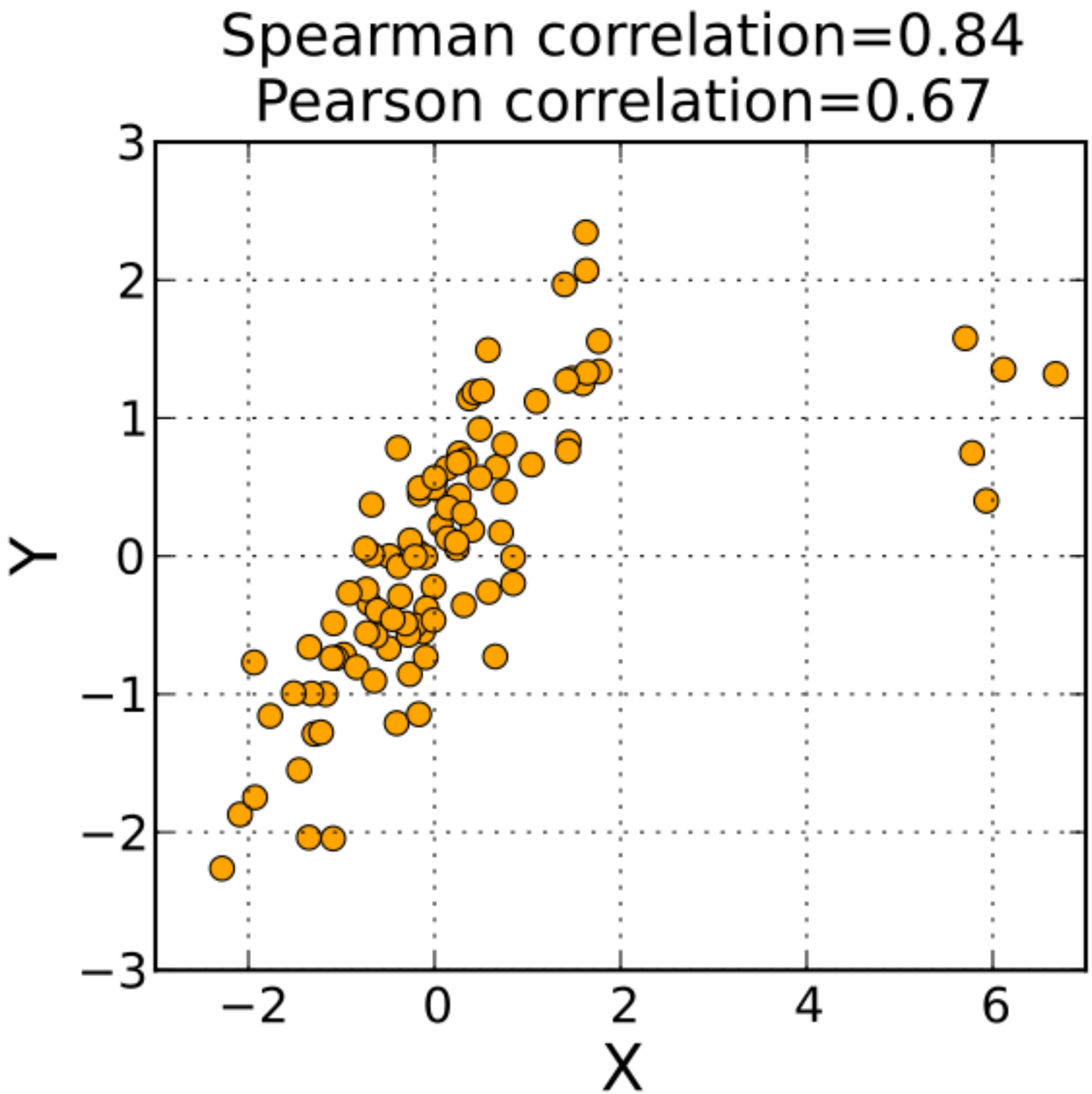
Var1		Var2	
7.6	3	**7.1**	3
3.1	1	**2.6**	1
4.5	2	**5.3**	2
102.0	4	**11.7**	4

“Spearman” Correlation

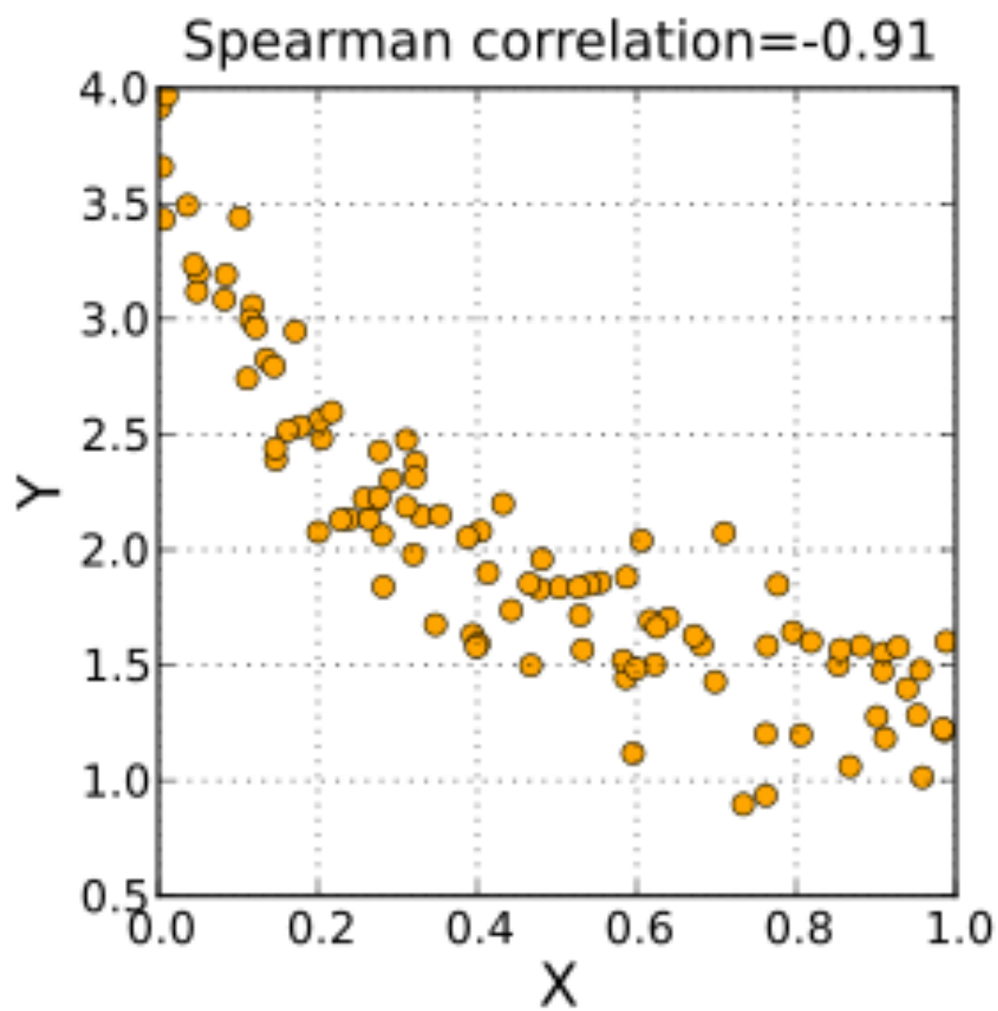
```
from scipy.stats import spearmanr
```

```
coef, p = spearmanr(data1, data2)
```

Spearman Correlation



Spearman Correlation



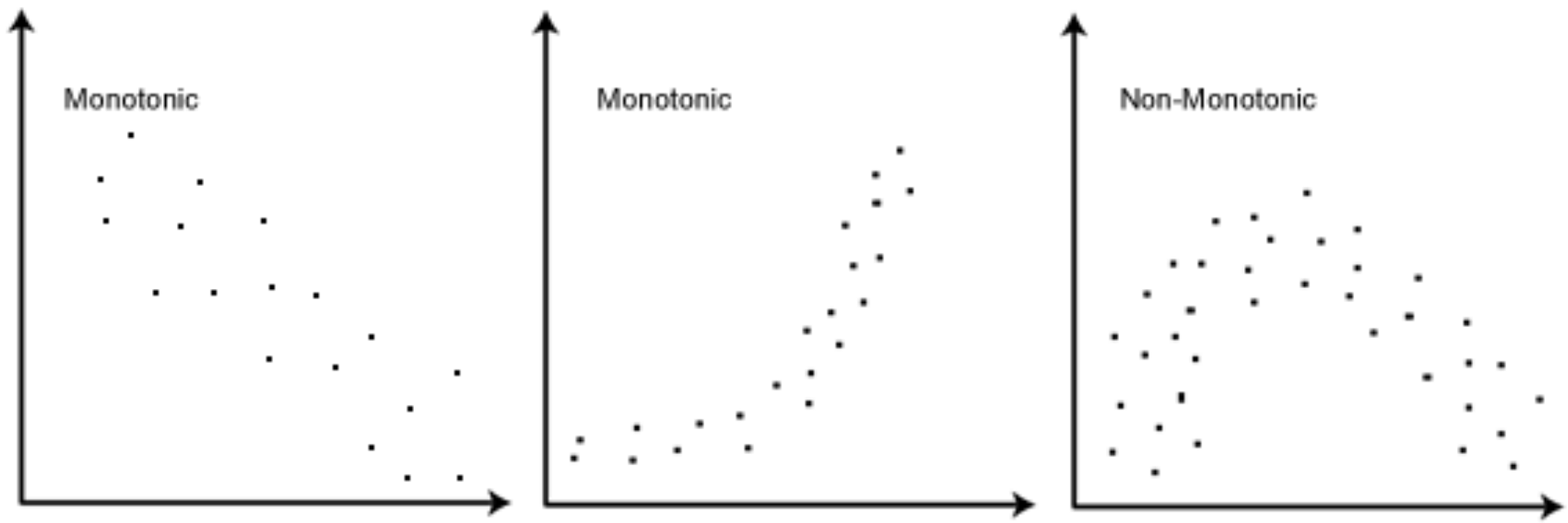
Spearman Correlation

Spearman: Strength of monotonic relationship between two variables

Pearson: Strength of linear relationship between two variables

Note: Linear relationships are monotonic

Spearman Correlation



Ex. Non-Monotonic Relationships

Non-monotonic: Age v.s Mortality Rate

High infant mortality, High elder mortality, relatively safe middle age.

Vehicle maintenance

Company/Security Ratings

The “Rank Trick”

Mean and Standard Deviations Are Misleading For Long-Tailed Distributions.

Measures majority-minority relationships (but lose scale!).

“Robust Statistics”

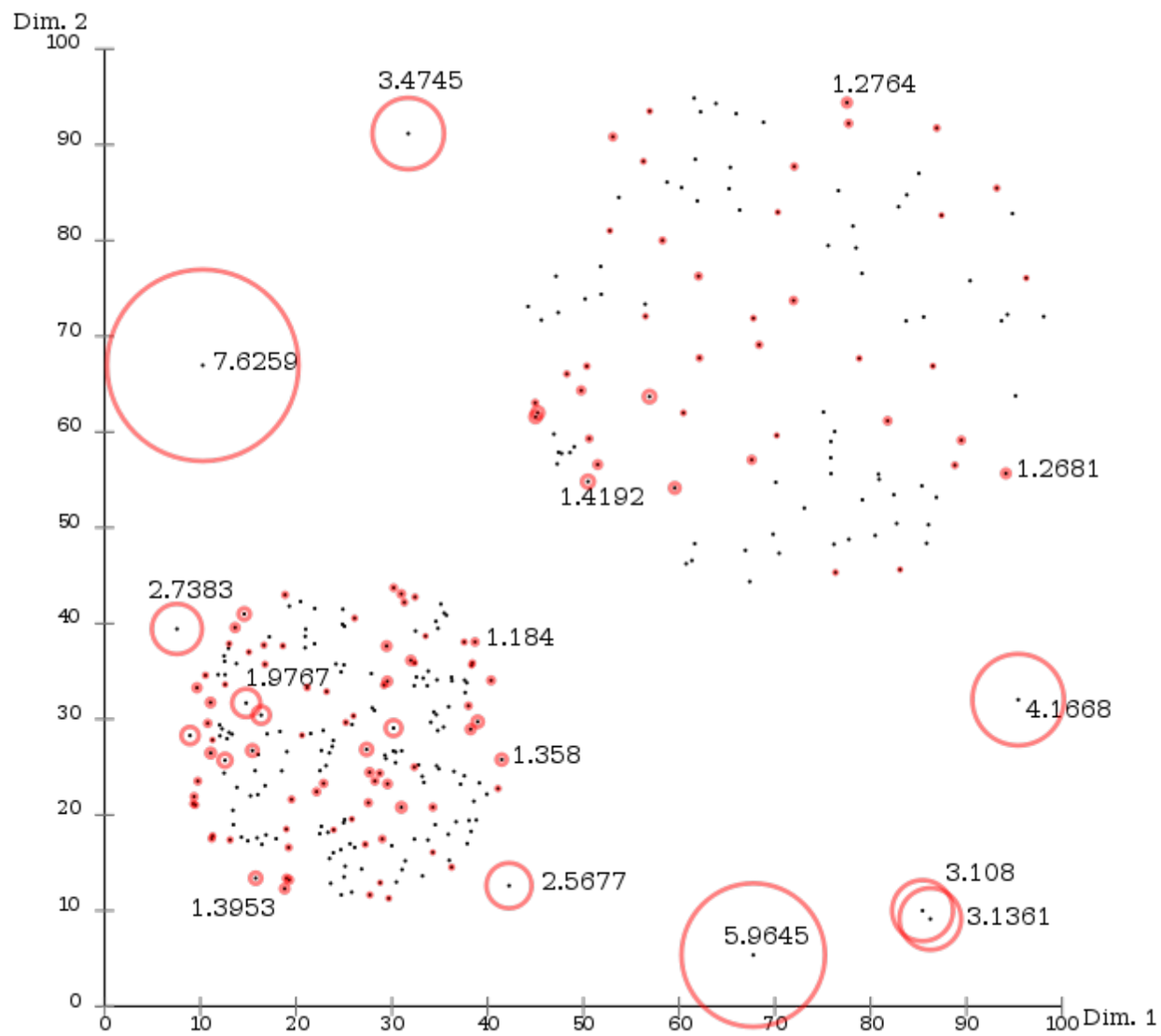
Robustness and “Outliers”

What does it mean? Comparisons are still meaningful if data are not normally distributed.

What does it not mean? Tolerate data corruption or noise

Outliers

Not a precise term….means something different to everyone.



Data Corruption/Error

Sometimes the data in our dataset is simply "wrong"

Integrated from different sources

Data entry error/Human error

Software bug

CHIDATA

Sanjay's $0.02

DO NOT TRY TO HANDLE SYSTEMATIC DATA CORRUPTION WITH ROBUST STATISTICS!!!

What does it not mean? Tolerate data corruption or noise

Paradox of Structured Data

Always certain, seldom right

- Programming tools for data (SQL, Python-Pandas, etc.) are grounded in formal logic. They "assume" the data that is processed is correct.
- A *database* holds facts and allows users to make principled inferences from these facts

Integrity Constraints

- Datasets not only should include attribute types but also constraints on the values they can take.

Person(First: String, Last: String, SSN: String)

Syntactic Constraints

- First name: must be alphabetical (upto apostrophes and dashes)
- Last name: must be alphabetical (upto apostrophes and dashes)
- SSN: must follow a format XXX-XX-XXXX

Semantic Constrains

- SSNs must be unique

Integrity Constraints

Syntactic Constraints: Formatting of a single cell of data

Semantic Constrains: Relationships between data

Customers

CustomerId	FirstName	LastName	DateCreated	Cli
1	Homer	Simpson	13/06/2014 3:33:37 PM	
2	Peter	Griffin	13/06/2014 9:09:56 PM	
3	Stewie	Griffin	13/06/2014 9:16:07 PM	
4	Brian	Griffin	13/06/2014 9:16:36 PM	
5	Cosmo	Kramer	13/06/2014 9:16:41 PM	
6	Philip	Fry	13/06/2014 9:17:02 PM	
7	Amy	Wong	13/06/2014 9:22:05 PM	
8	Hubert J.	Farnsworth	13/06/2014 9:22:19 PM	
9	Marge	Simpson	13/06/2014 9:22:37 PM	
10	Bender	Rodríguez	13/06/2014 9:22:52 PM	
11	Turanga	Leela	13/06/2014 9:23:37 PM	
(New)			15/06/2014 9:00:01 PM	

Integrity Constraints

Syntactic Constraints: Formatting of a single cell of data

No cell can be empty

Customers

CustomerId	FirstName	LastName	DateCreated	Cli
1	Homer	Simpson	13/06/2014 3:33:37 PM	
2	Peter	Griffin	13/06/2014 9:09:56 PM	
3	Stewie	Griffin	13/06/2014 9:16:07 PM	
4	Brian	Griffin	13/06/2014 9:16:36 PM	
5	Cosmo	Kramer	13/06/2014 9:16:41 PM	
6	Philip	Fry	13/06/2014 9:17:02 PM	
7	Amy	Wong	13/06/2014 9:22:05 PM	
8	Hubert J.	Farnsworth	13/06/2014 9:22:19 PM	
9	Marge	Simpson	13/06/2014 9:22:37 PM	
10	Bender	Rodríguez	13/06/2014 9:22:52 PM	
11	Turanga	Leela	13/06/2014 9:23:37 PM	
(New)			15/06/2014 9:00:01 PM	

Integrity Constraints

Semantic Constraints: Relationships between data

All firstname, lastname pairs are unique

Customers

CustomerId	FirstName	LastName	DateCreated	Cli
1	Homer	Simpson	13/06/2014 3:33:37 PM	
2	Peter	Griffin	13/06/2014 9:09:56 PM	
3	Stewie	Griffin	13/06/2014 9:16:07 PM	
4	Brian	Griffin	13/06/2014 9:16:36 PM	
5	Cosmo	Kramer	13/06/2014 9:16:41 PM	
6	Philip	Fry	13/06/2014 9:17:02 PM	
7	Amy	Wong	13/06/2014 9:22:05 PM	
8	Hubert J.	Farnsworth	13/06/2014 9:22:19 PM	
9	Marge	Simpson	13/06/2014 9:22:37 PM	
10	Bender	Rodríguez	13/06/2014 9:22:52 PM	
11	Turanga	Leela	13/06/2014 9:23:37 PM	
(New)			15/06/2014 9:00:01 PM	

Unit Tests for Data

- What can a programmer assume is true about the data!

Person(First: String, Last: String, SSN: String)

Syntactic Constraints

- First name: must be alphabetical (upto apostrophes and dashes)
- Last name: must be alphabetical (upto apostrophes and dashes)
- SSN: must follow a format XXX-XX-XXXX

Semantic Constrains

- SSNs must be unique

Sanjay's $0.02

DO NOT TRY TO HANDLE SYSTEMATIC DATA CORRUPTION WITH ROBUST STATISTICS!!!

Similar to handling all exceptions with try-catch statements.

A Recipe For Success

Step 1. Build a logical data model

Describe dataset attributes, data types, and integrity constraints.

Use software or manual effort to fix violations of the data model

Step 2. Build a statistical data model

Describe dataset attributes in terms of their distributions

Choose the appropriate robust/non-robust tests/statistics to work with those distributions.