Introductory Statistics/ Refresher & Intro to Labs

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 1, Week1 - Module 2, September 4th, 2020

Your Github Repository

- Your Github Repository for this class
- Please create a Github repo for the Data Analytics class Labs
- Do not share the Assignment codes in your Data Analytics course repo, you share only your lab work and your individual project work/code.
- TA will collect your Github repo URLs next week.
- Example: https://github.com/tYourGithub/DataAnalytics2020_YOUR_NAME

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency
- Probability
- Significance tests
- Hypothesis (null and alternate)
- P-value
- Density and cumulative distributions

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency

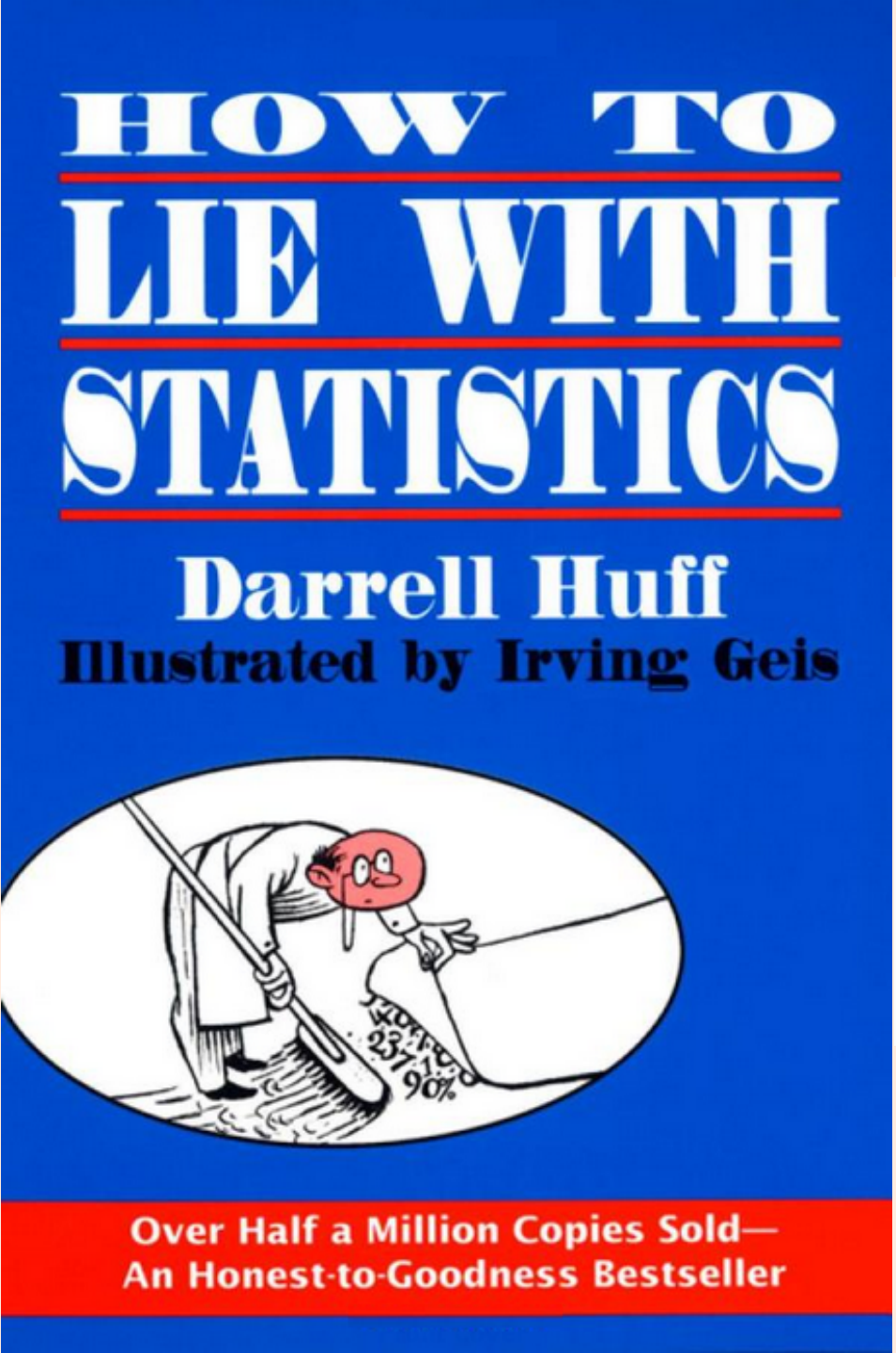
Friday’s class

- Probability
- Significance tests
- Hypothesis (null and alternate)
- P-value
- Density and cumulative distributions

Tuesday's class (next class)

Statistic and Statistics

- Statistic (not to be confused with Statistics)
  - Characteristic or measure obtained from a sample.
- Statistics
  - Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.



HOW TO LIE WITH STATISTICS
(Huff, D. 1954)

There are three kinds of lies: lies, damned lies, and statistics.
—Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
—H. G. Wells

It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so.
—Artemus Ward

Round numbers are always false. —Samuel Johnson

I have a great subject [statistics] to write upon, but feel keenly my literary incapacity to make it easily intelligible without sacrificing accuracy and thoroughness.
—Sir Francis Galton

What is "statistics"?

- The term "statistics" has two common meanings, which we want to clearly separate: **descriptive** and **inferential** statistics.
- But to understand the difference between descriptive and inferential statistics, we must first be clear on the difference between populations and samples.

Courtesy Marshall Ma (and prior sources)

Populations and samples

- A population is a set of well-defined objects
  - We must be able to say, for every object, if it is in the population or not
  - We must be able, in principle, to find every individual of the population

  A geographic example of a population is all pixels in a multi-spectral satellite image

- A sample is a subset of a population
  - We must be able to say, for every object in the population, if it is in the sample or not
  - Sampling is the process of selecting a sample from a population

  Continuing the example, a sample from this population could be a set of pixels from known ground truth points

Courtesy Marshall Ma (and prior sources)

Populations and samples

Definitions

- Population : The complete set of actual or potential elements about which inferences are made
- Sample : A subset of the population selected using some sampling methods.

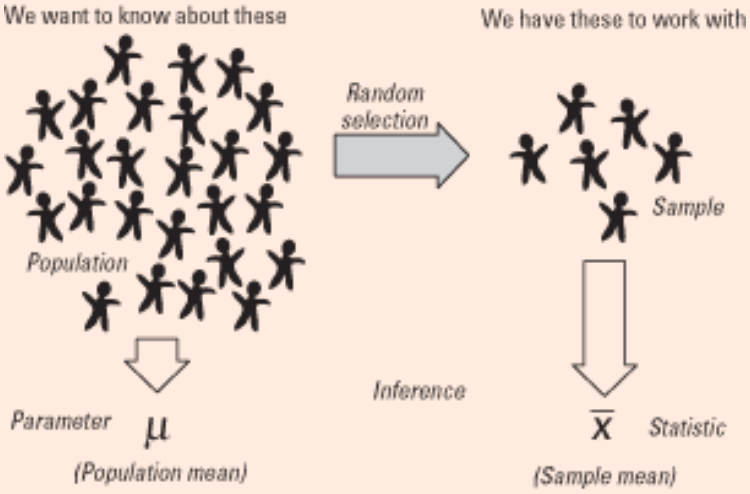
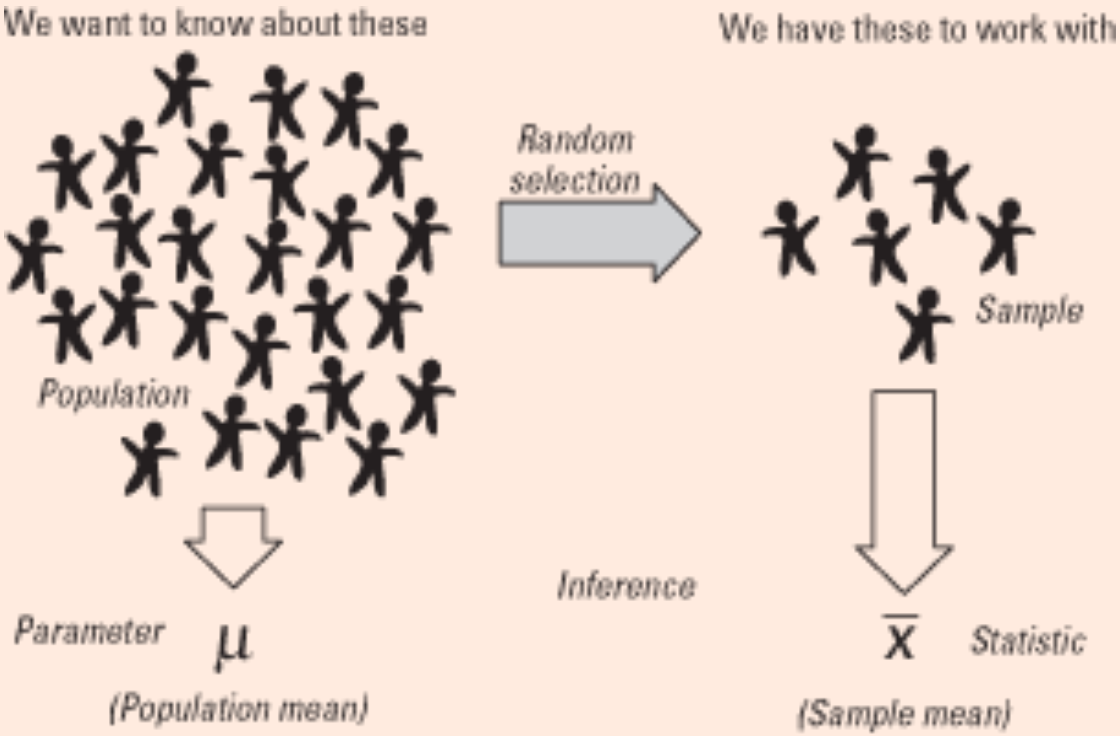


Image Credit/Reference: Quick Study Statistics

Populations and samples

- Population : The complete set of actual or potential elements about which inferences are made
- Sample : A subset of the population selected using some sampling methods.



Reference: Quick Study Statistics
Image Courtesy: https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

Populations and samples

- A population = “all” of the data, if you can get it (BIG Data)
  - This is what is different about the methods you use
- A sample = “some” of the data, and you may not know how representative it is
  - This is what limits analysis but certainly the development of models

Courtesy Marshall Ma (and prior sources)

Sampling Types (basic)

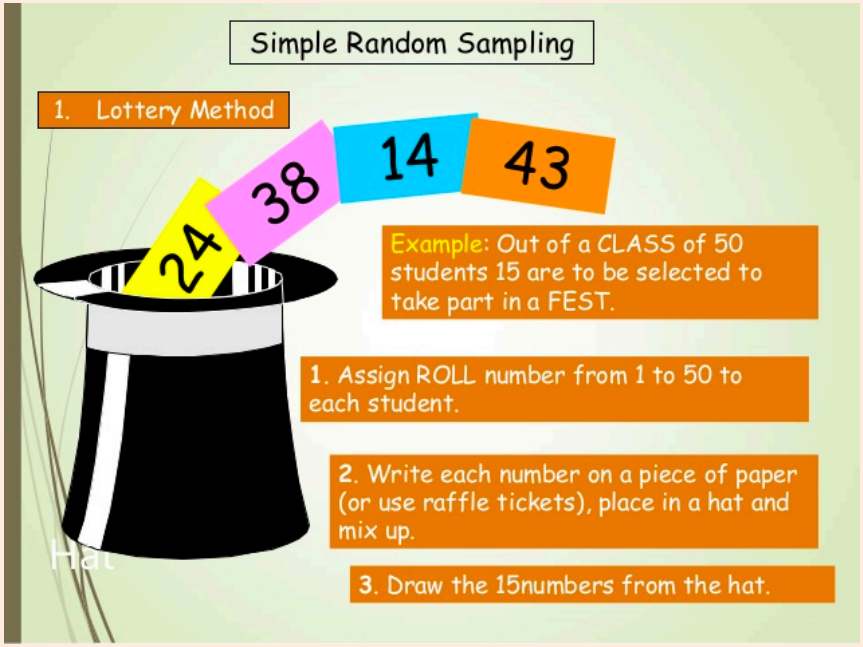
- Random Sampling
  - Sampling in which the data is collected using chance methods or random numbers.
- Systematic Sampling
  - Sampling in which data is obtained by selecting every *k*th object.
- Convenience Sampling
  - Sampling in which data is readily available is used.
- Stratified Sampling
  - Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.
- Cluster Sampling
  - Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Sampling Methods

- Simple Random Sampling
- Cluster Sampling
- Stratified Sampling

Sampling Methods

Simple Random Sampling: A sample selected so that each possible sample of the same size has an equal probability of being selected; used for most elementary inference



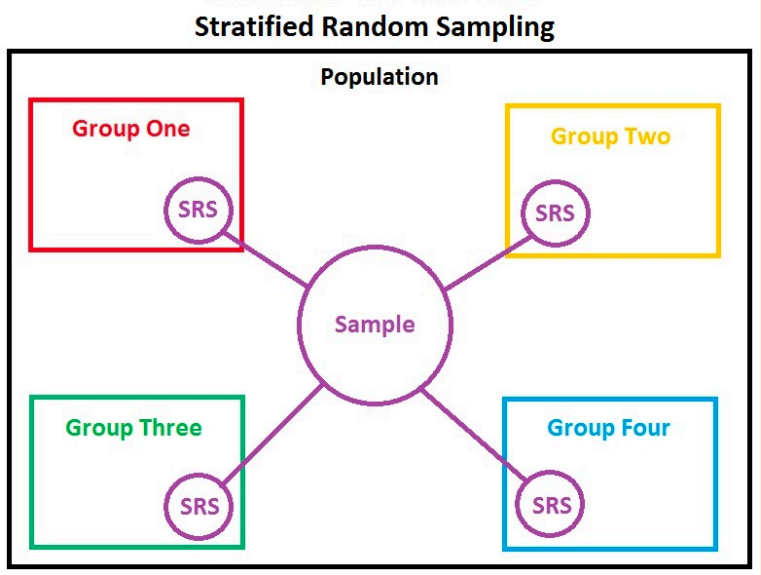
Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Reference: Quick Study Statistics

Image Courtesy: https://www.slideshare.net/mohammedzuhairy1/sampling-techniques-64917617

Sampling Methods

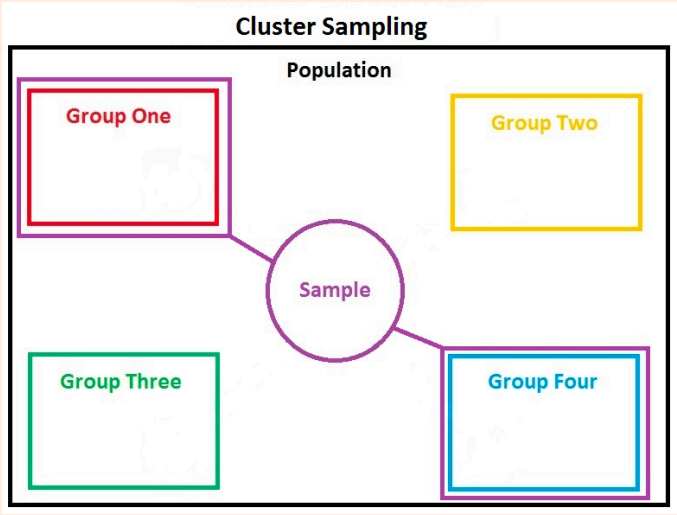
- Stratified Sampling: The population is divided into strata, and a fixed number of elements of each stratum are selected for the sample.



https://en.wikipedia.org/wiki/Stratified_sampling

Sampling Methods

- Cluster Sampling: The population is divided into groups called clusters; some clusters are randomly selected, and every member in them observed.



https://en.wikipedia.org/wiki/Cluster_sampling

Variable

- Variable: An attribute of a population or sample that can be measure.

Example: weight, height, eye-color, pulse rate are some of the many variables that can be measured for people.

Data

Data

- Types of Data:
  - Qualitative (or Categorical)
  - Quantitative (data like numeric values)

Data

Qualitative :

Qualitative (or Categorical) data are descriptive, but not numeric.

Example: your eye-color, you gender, color of a vehicle, your birthplace

Data

Quantitative :

Quantitative (data like numeric values)

- Discreate data take counting numbers (0,1,2,3..) this is used to represent things that can be counted. Example: Number of times an employee is late to work. Number of cars parking lot in the parking garage.

- Continuous data can take a range of numeric values, not just the counting numbers (fractions, decimals are included..) Example: height of a person, weight of an apple, amount of times an employee late to work.

Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Levels of Measurement

Qualitative (or Categorical) data can be measured at the:

- Nominal Level: Values are just names, without any order (example: eye-color)
- Ordinal Level: Values have some natural order, example: high school class (freshman, sophomore, ..) military rank

Quantitative (data like numeric values) can be measured at the:

- Interval Level: Numeric data with no natural zero point; intervals (differences) are meaningful but ratios are not, example: Temperature in Fahrenheit degrees 80F is not 20F hotter than 60F, but it is not 150% as hot.
- Ratio Level: Numeric data which there is true zero, both intervals and ratios are meaningful; Example: weight, length, duration

Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Types of Data

Type of data	Level of measurement	Examples
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Gender
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Parameter

- Parameter: A numeric measure that describe a population: parameters are usually not computed; but are inferred from sample statistics.
- Parameters are normally denoted using Greek symbols, whereas the corresponding statistics are denoted using Latin letters. Below you find a list of the most important parameters and the corresponding statistics.

	Parameter	Statistic
mean	μ	m
standard deviation	σ	s
correlation coefficient	ρ	r

Courtesy: Quick Study Academic – Statistics www.quickstudy.com
Resources & Image Credit: http://www.statistics4u.info/fundstat_eng/cc_parameter.html

Something to remember…

- Difference between “N” and “n”
- Population → N
- Sample → n

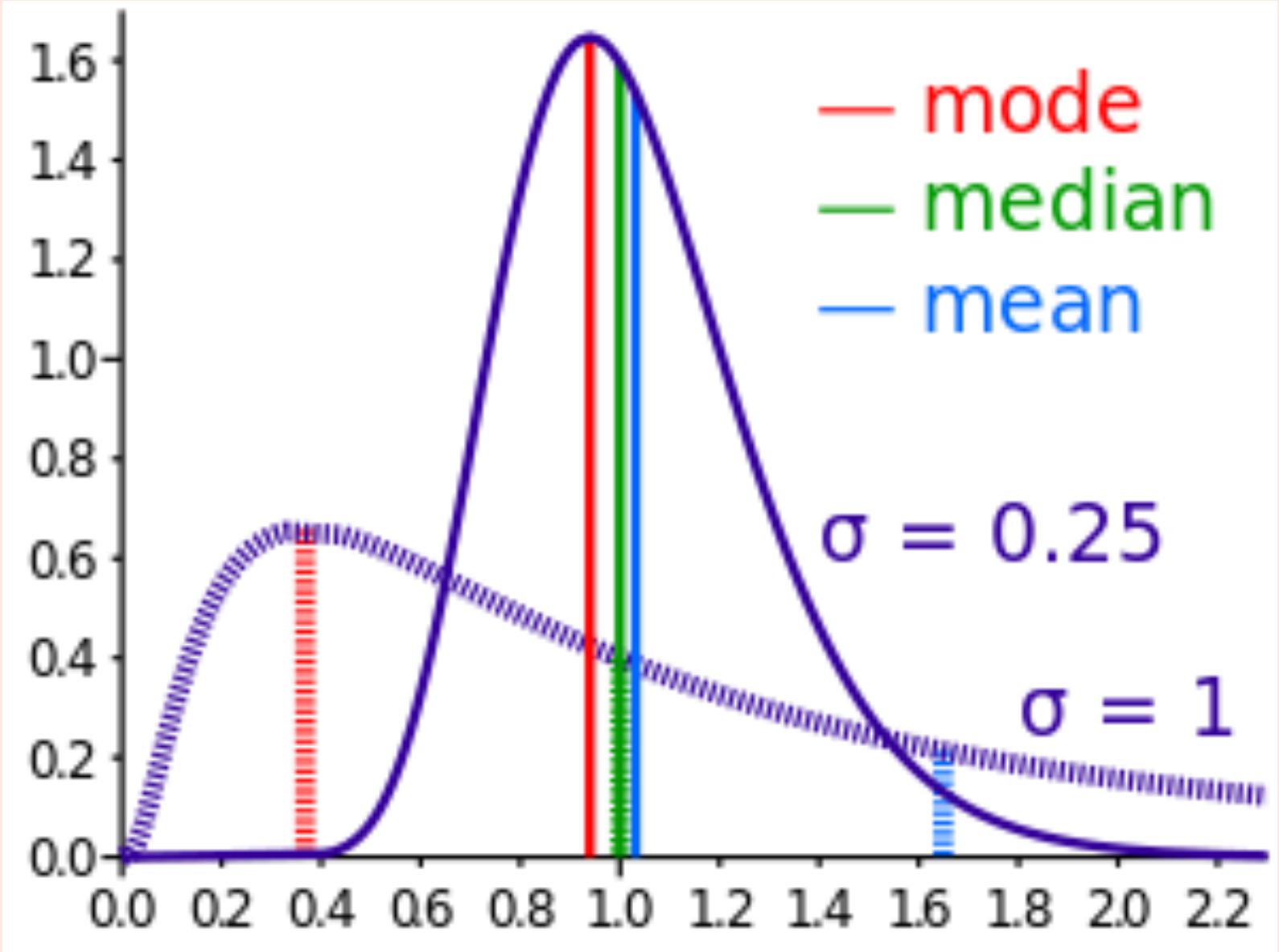
Special values in data

- Fill value
- Error value
- Missing value
- Not-a-number (NAN)
- Infinity
- Default
- Null
- Rational numbers

Outlier

- An extreme, or atypical, data value(s) in a sample.
- They should be considered carefully, before exclusion from analysis.
- For example, data values maybe recorded erroneously, and hence they may be corrected.
- However, in other cases they may just be surprisingly different, but not necessarily 'wrong'.

Central tendency – median, mean, mode



Measure of Central Tendency

- Mean: Most commonly used measure of central tendency, commonly meant by "Average", sensitive to extreme values (sensitive to outliers)
  - Population Mean
  - Sample Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

https://www.youtube.com/watch?v=k5EbijWu-Ss

Image Resource: https://www.onlinemathlearning.com/population-mean.html
Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Measure of Central Tendency

- **Median**: Value that divides the set in so the same number of observations lie on each side of it;
- ***Median is less sensitive to extreme values***
- For an even number, it is the average of middle two values

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

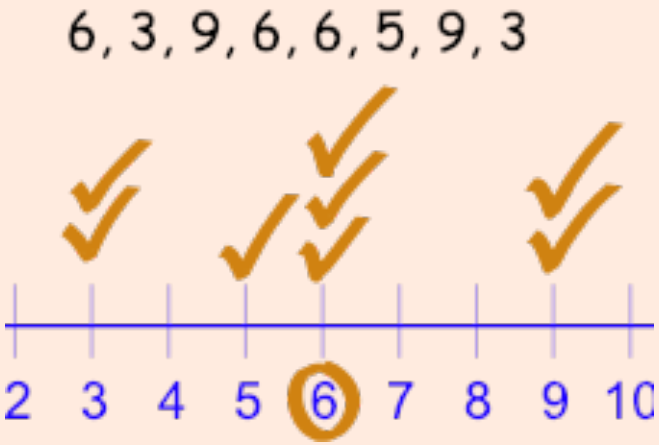
= 4.5

Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Image Resource: https://en.wikipedia.org/wiki/Median

Measure of Central Tendency

- **Mode**: Observation that occurs with the greatest frequency.



Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Image Resource: https://www.mathsisfun.com/definitions/mode.html

Mean;Median;Outlier

You Retweeted

Anna J. Egalite @annaegalite · Aug 27

In my intro stats class today, I told students the median is a "resistant" measure of a distribution's center & is often preferred to the mean in the case of salary data, etc. I jokingly referenced this meme and in the 15 mins' break they had, a student created this MASTERPIECE!

234 7.5K 35.8K

Found this on Twitter… Credit: Anna .J. Egalite

Frequencies…

- The **Absolute frequency** n_i is the number of observations belonging to a category a_i or falling into a particular class c_i. The sum of all frequencies of all categories/classes is equal to N, the total number of observations:

$$\Sigma n_i = N$$

- **Relative frequencies** f_i are obtained by normalizing the individual frequencies to a total sum of 1.0 (or 100%, respectively). This way the frequencies become independent of the sample size and will be comparable to each other.
- Frequencies are usually delineated in a **frequency table** or displayed as a **histogram**.

Example: 28 persons have been asked for their eye colors, resulting in the following frequencies:

eye color	abs. frequency	rel. frequency
brown	14	0.500 (50%)
gray	2	0.071 (7.1%)
blue	9	0.321 (32.1%)
green	3	0.107 (10.7%)

The dataset with 28 observations and one variable exihibts four categories which differ in their frequencies.

Resource & Image Credit: http://www.statistics4u.info/fundstat_eng/ee_frequency.html

Ranges: z, Percentiles, Quartiles

- The standard score is obtained by subtracting the mean and dividing the difference by the standard deviation. The symbol is z, which is why it's also called a z-score.
- **Percentiles (quantiles) (100 regions)**
  - The kth percentile is the number which has k% of the values below it. The data must be ranked.
- **Quartiles (4 regions)**
  - The quartiles divide the data into 4 equal regions.
  - Note: The 2^{nd} quartile is the same as the median. The 1^{st} quartile is the 25^{th} percentile, the 3^{rd} quartile is the 75^{th} percentile.

Getting Started : Rstudio – MASS library

install.packages("MASS") # installing the MASS package

library(MASS) # load the library MASS

attach(Boston) # attaching the dataset

?Boston # help function with "?"

head(Boston) # show the head of the dataset

dim(Boston) # dimensions of the dataset

names(Boston) # column names

str(Boston) # str function shows the structure of the dataset

nrow(Boston) # function shows the number of rows

ncol(Boston) # function shows the number of columns

summary(Boston) # summary() function shows the summary statistics

summary(Boston$crim) # summary of the "crime" column in the Boston dataset

Getting Started : Rstudio – ISLR library – Auto dataset

```
install.packages("ISLR") # installing the ISLR package
library(ISLR)
data(Auto)
head(Auto)
names(Auto)
summary(Auto)
summary(Auto$mpg)
fivenum(Auto$mpg)
boxplot(Auto$mpg)
hist(Auto$mpg)
summary(Auto$horsepower)
summary(Auto$weight)
fivenum(Auto$weight)
boxplot(Auto$weight)
mean(Auto$weight)
median((Auto$weight))
```

Time to “Play ☺” with the data

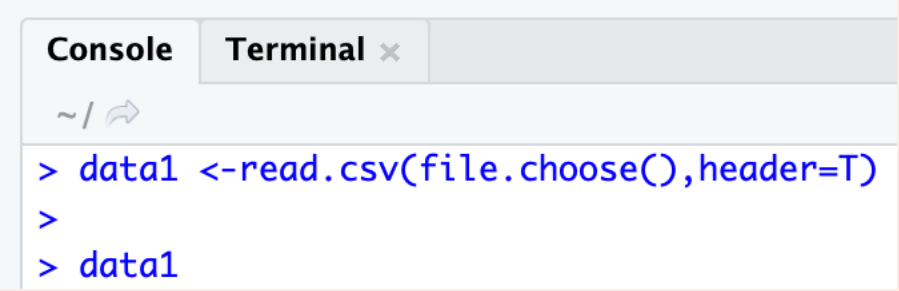
- In class work – Explore the EPI dataset –
- http://aquarius.tw.rpi.edu/html/DA/
- And some directories under this link
- – **please search before ask !!!**
- This is where the files for assignments, lab exercises are
  - data and code fragments…

Read a CSV file in R

Begin with reading a CSV file..

```
> help(read.csv)
> data1 <-read.csv(file.choose(),header=T)
> data1
```

`file.choose()` Choose the CSV file



EPI data set

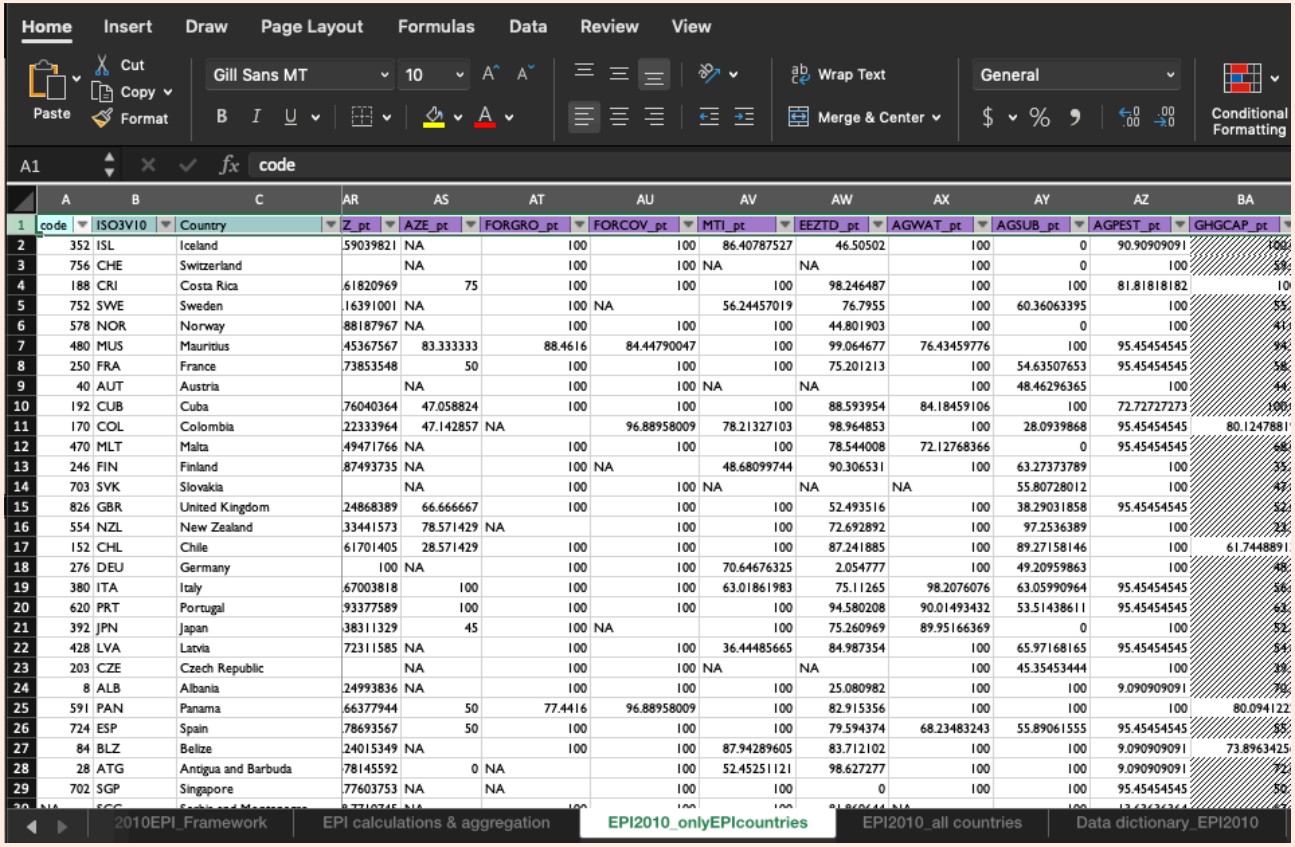
Index of /html/DA/EPI

Name	Last modified	Size	Description
Parent Directory		-	
2010EPI_data.csv	05-Feb-2016 00:28	10M	
2010EPI_data.xls	05-Feb-2016 00:35	11M	
2016 EPI Wastewater Data Appendix.xls	19-Jan-2018 16:01	907K	
2016EPI_Backcasted_Scores.xls	19-Jan-2018 16:01	1.3M	
2016EPI_Full_Report_opt.pdf	19-Jan-2018 16:02	15M	
2016EPI_Raw_Data.xls	19-Jan-2018 16:02	1.5M	
2016_epi_framework_indicator_scores_friendly.xls	19-Jan-2018 16:02	740K	
2016epi_weightings_0.xls	19-Jan-2018 16:02	660K	
EPI_data.csv	05-Feb-2016 00:28	232K	
EPI_data.xls	05-Feb-2016 00:36	11M	
Fisheries_Penalties.xls	19-Jan-2018 16:02	120K	
OnlyEPI_data.csv	05-Feb-2016 00:29	10M	
OnlyEPI_data.xls	05-Feb-2016 00:37	11M	
filters_materiality_for_2016epi.xls	19-Jan-2018 16:02	64K	

Apache/2.2.14 (Ubuntu) Server at aquarius.tw.rpi.edu Port 443

https://aquarius.tw.rpi.edu/html/DA/EPI/

2010EPI_data.xls



Home Insert Draw Page Layout Formulas Data Review View

Cut Copy Format Paste — Gill Sans MT — 10 — Wrap Text — Merge & Center — General — Conditional Formatting

A1 — code

	A	B	C	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA
1	code	ISO3V10	Country	Z_pt	AZE_pt	FORGRO_pt	FORCOV_pt	MTI_pt	EEZTD_pt	AGWAT_pt	AGSUB_pt	AGPEST_pt	GHGCAP_pt
2	352	ISL	Iceland	59039821	NA	100	100	86.40787527	46.50502	100	0	90.90909091	100
3	756	CHE	Switzerland		NA	100	100	NA	NA	100	0	100	59
4	188	CRI	Costa Rica	61820969	75	100	100	100	98.246487	100	100	81.81818182	10
5	752	SWE	Sweden	16391001	NA	100	NA	56.24457019	76.7955	100	60.36063395	100	55
6	578	NOR	Norway	88187967	NA	100	100	100	44.801903	100	0	100	41
7	480	MUS	Mauritius	45367567	83.333333	88.4616	84.44790047	100	99.064677	76.43459776	100	95.45454545	94
8	250	FRA	France	73853548	50	100	100	100	75.201213	100	54.63507653	95.45454545	58
9	40	AUT	Austria		NA	100	100	NA	NA	100	48.46296365	100	44
10	192	CUB	Cuba	76040364	47.058824	100	100	100	88.593954	84.18459106	100	72.72727273	100
11	170	COL	Colombia	22333964	47.142857	NA	96.88958009	78.21327103	98.964853	100	28.0939868	95.45454545	80.1247881
12	470	MLT	Malta	49471766	NA	100	100	100	78.544008	72.12768366	0	95.45454545	68
13	246	FIN	Finland	87493735	NA	100	NA	48.68099744	90.306531	100	63.27373789	100	35
14	703	SVK	Slovakia		NA	100	100	NA	NA	NA	55.80728012	100	47
15	826	GBR	United Kingdom	24868389	66.666667	100	100	100	52.493516	100	38.29031858	95.45454545	52
16	554	NZL	New Zealand	33441573	78.571429	NA	100	100	72.692892	100	97.2536389	100	23
17	152	CHL	Chile	61701405	28.571429	100	100	100	87.241885	100	89.27158146	100	61.7448891
18	276	DEU	Germany	100	NA	100	100	70.64676325	2.054777	100	49.20959863	100	48
19	380	ITA	Italy	67003818	100	100	100	63.01861983	75.11265	98.2076076	63.05990964	95.45454545	56
20	620	PRT	Portugal	93377589	100	100	100	100	94.580208	90.01493432	53.51438611	95.45454545	63
21	392	JPN	Japan	38311329	45	100	NA	100	75.260969	89.95166369	0	100	52
22	428	LVA	Latvia	72311585	NA	100	100	36.44485665	84.987354	100	65.97168165	95.45454545	54
23	203	CZE	Czech Republic		NA	100	100	NA	NA	100	45.35453444	100	39
24	8	ALB	Albania	24993836	NA	100	100	100	25.080982	100	100	9.090909091	70
25	591	PAN	Panama	66377944	50	77.4416	96.88958009	100	82.915356	100	100	100	80.094122
26	724	ESP	Spain	78693567	50	100	100	100	79.594374	68.23483243	55.89061555	95.45454545	55
27	84	BLZ	Belize	24015349	NA	100	100	87.94289605	83.712102	100	100	9.090909091	73.8963425
28	28	ATG	Antigua and Barbuda	78145592	0	NA	100	52.45251121	98.627277	100	100	9.090909091	72
29	702	SGP	Singapore	77603753	NA	NA	100	100	0	100	100	95.45454545	50

2010EPI_Framework | EPI calculations & aggregation | EPI2010_onlyEPIcountries | EPI2010_all countries | Data dictionary_EPI2010

Data Prepared for Analysis = Munging

- Missing values, null values, etc.
- E.g. in the EPI_data – they use "--"
  - Most data applications provide built ins for these higher-order functions – in R "NA" is used and functions such as is.na(var), etc. provide powerful filtering options (we'll cover these on next Friday)
- Of course, different variables often are missing "different" values
- In R – higher-order functions such as: Reduce, Filter, Map, Find, Position and Negate will become your enemies and then your friends: http://www.johnmyleswhite.com/notebook/2010/09/23/higher-order-functions-in-r/

Explore the “Missing values” -- NA

ISL	Iceland	.59039821	NA	100
CHE	Switzerland		NA	100
CRI	Costa Rica	.61820969	75	100
SWE	Sweden	.16391001	NA	100
NOR	Norway	.88187967	NA	100
MUS	Mauritius	.45367567	83.333333	88.4616
FRA	France	.73853548	50	100
AUT	Austria		NA	100
CUB	Cuba	.76040364	47.058824	100
COL	Colombia	.22333964	47.142857	NA
MLT	Malta	.49471766	NA	100
FIN	Finland	.87493735	NA	100
SVK	Slovakia		NA	100
GBR	United Kingdom	.24868389	66.666667	100
NZL	New Zealand	.33441573	78.571429	NA
CHL	Chile	.61701405	28.571429	100
DEU	Germany	100	NA	100
ITA	Italy	.67003818	100	100
PRT	Portugal	.93377589	100	100
JPN	Japan	.38311329	45	100
LVA	Latvia	.72311585	NA	100
CZE	Czech Republic		NA	100
ALB	Albania	.24993836	NA	100
PAN	Panama	.66377944	50	77.4416

Five-number summary

The five-number summary is a set of descriptive statistics that provide information about a dataset. It consists of the five most important sample percentiles:

1. the sample minimum (smallest observation)
2. the lower quartile or first quartile
3. the median (the middle value)
4. the upper quartile or third quartile
5. the sample maximum (largest observation)

Read: https://en.wikipedia.org/wiki/Five-number_summary

Getting started – look at the data

- Visually
  - What is the improvement in the understanding of the data as compared to the situation without visualization?
  - Which visualization techniques are suitable for one's data?
    - Scatter plot diagrams
    - Box plots (min, 1st quartile, median, 3rd quartile, max)
    - Stem and leaf plots
    - Frequency plots
    - Group Frequency Distributions plot
    - Cumulative Frequency plots
    - Distribution plots

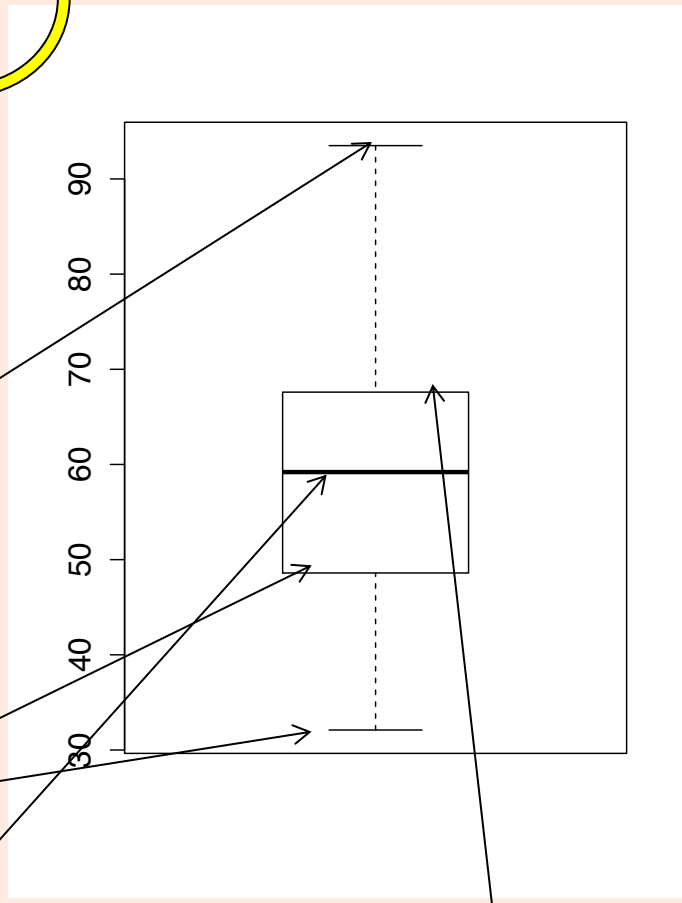
Why visualization?

- Reducing amount of data
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products

Exploring the distribution

```
> summary(EPI)       # stats
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  32.10   48.60   59.20   58.37   67.60   93.50      68
```

> boxplot(EPI)

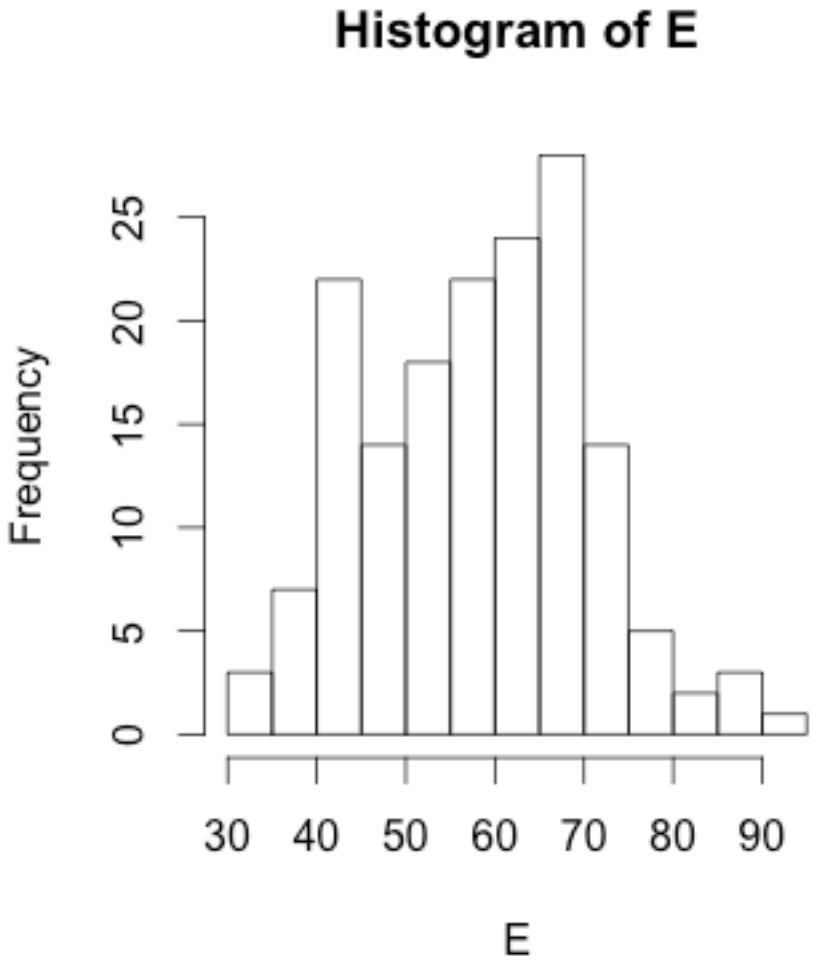



```
> fivenum(EPI,na.rm=TRUE)
[1] 32.1 48.6 59.2 67.6 93.5
```

Tukey: min, lower hinge, median, upper hinge, max

Grouped Frequency Distribution aka binning

> hist(EPI) #defaults



Distributions

- Shape
- Character
- Parameter(s)

- Which one fits?

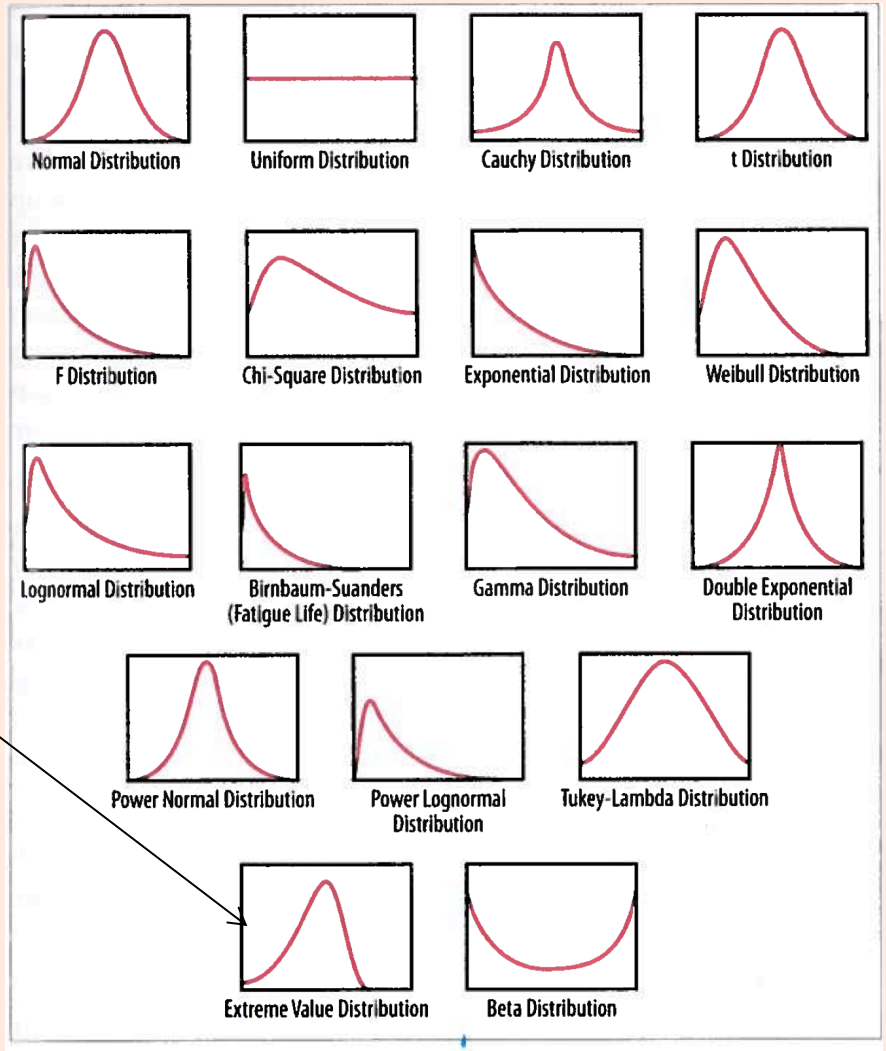


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Distributions

- http://www.quantitativeskills.com/sisa/rojo/alldist.zip
- Shape
- Character
- Parameter(s)
  - Mean
  - Standard deviation
  - Skewness
  - Etc.

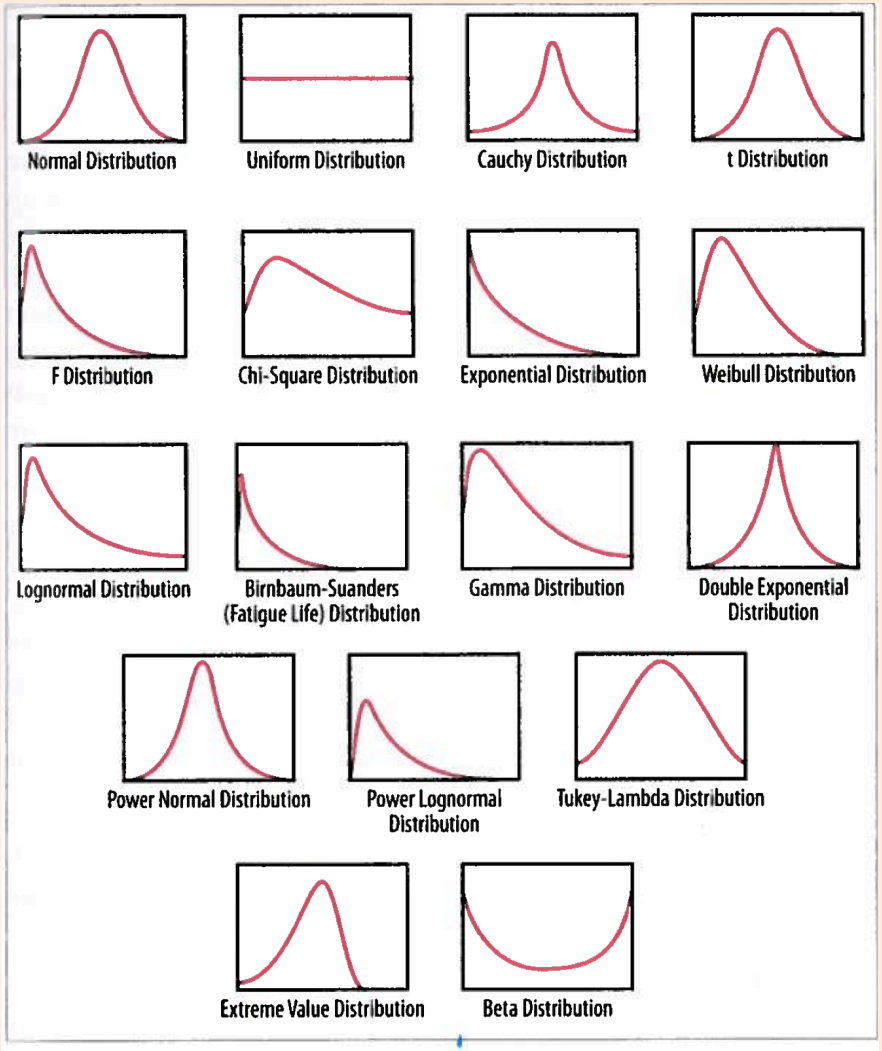


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Plotting these distributions

- Histograms and binning
- Getting used to log scales
- Going beyond 2-D
- More of this in lab 1 and module 3

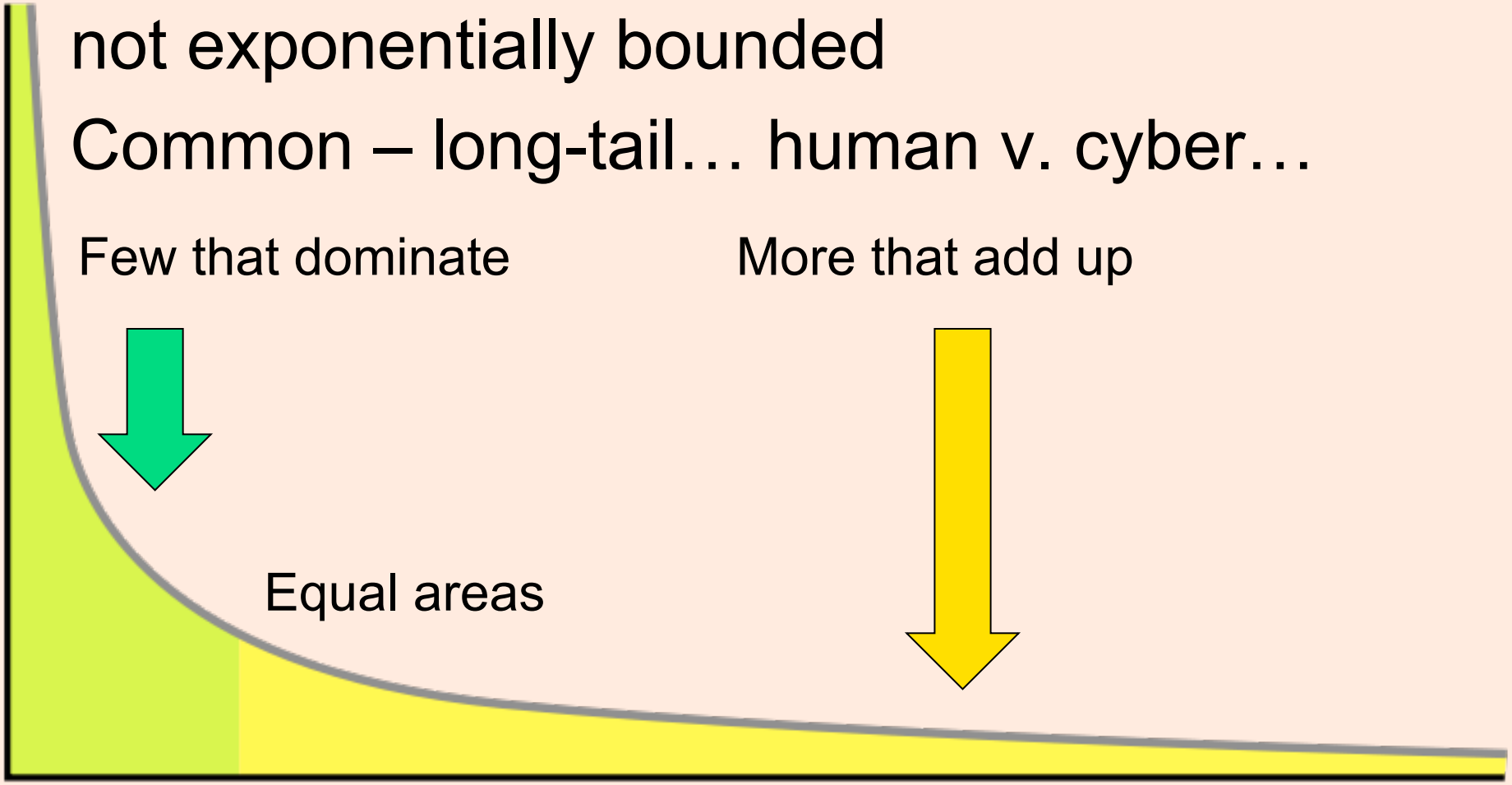
In applications

- Scipy: http://docs.scipy.org/doc/scipy/reference/stats.html
- R: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/Distributions.html
- Matlab: http://www.mathworks.com/help/stats/_brn2irf.html
- Excel: see http://aquarius.tw.rpi.edu/html/DA/distribution

Heavy-tail distributions

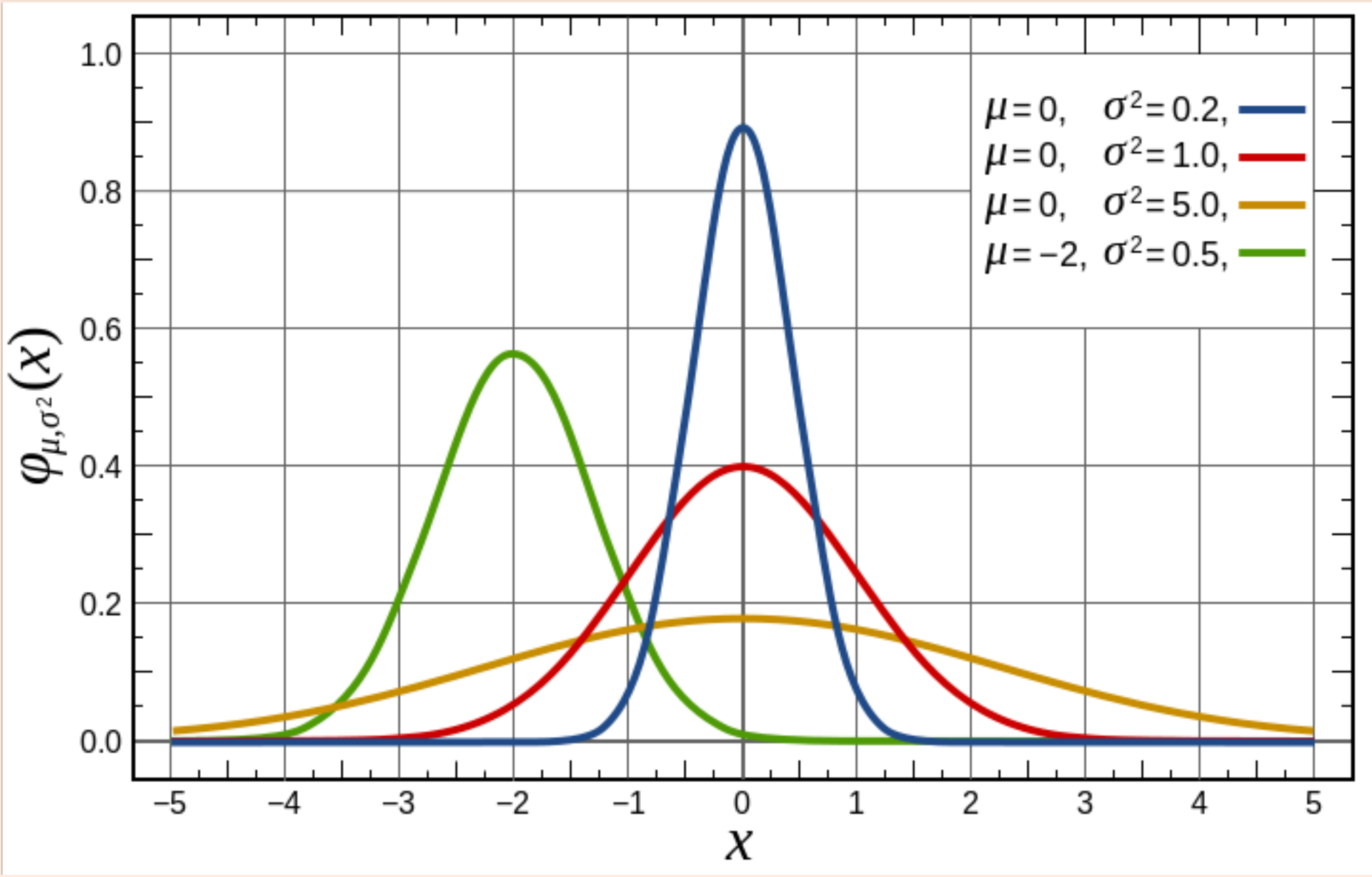
- are probability distributions whose tails are not exponentially bounded

Common – long-tail… human v. cyber…

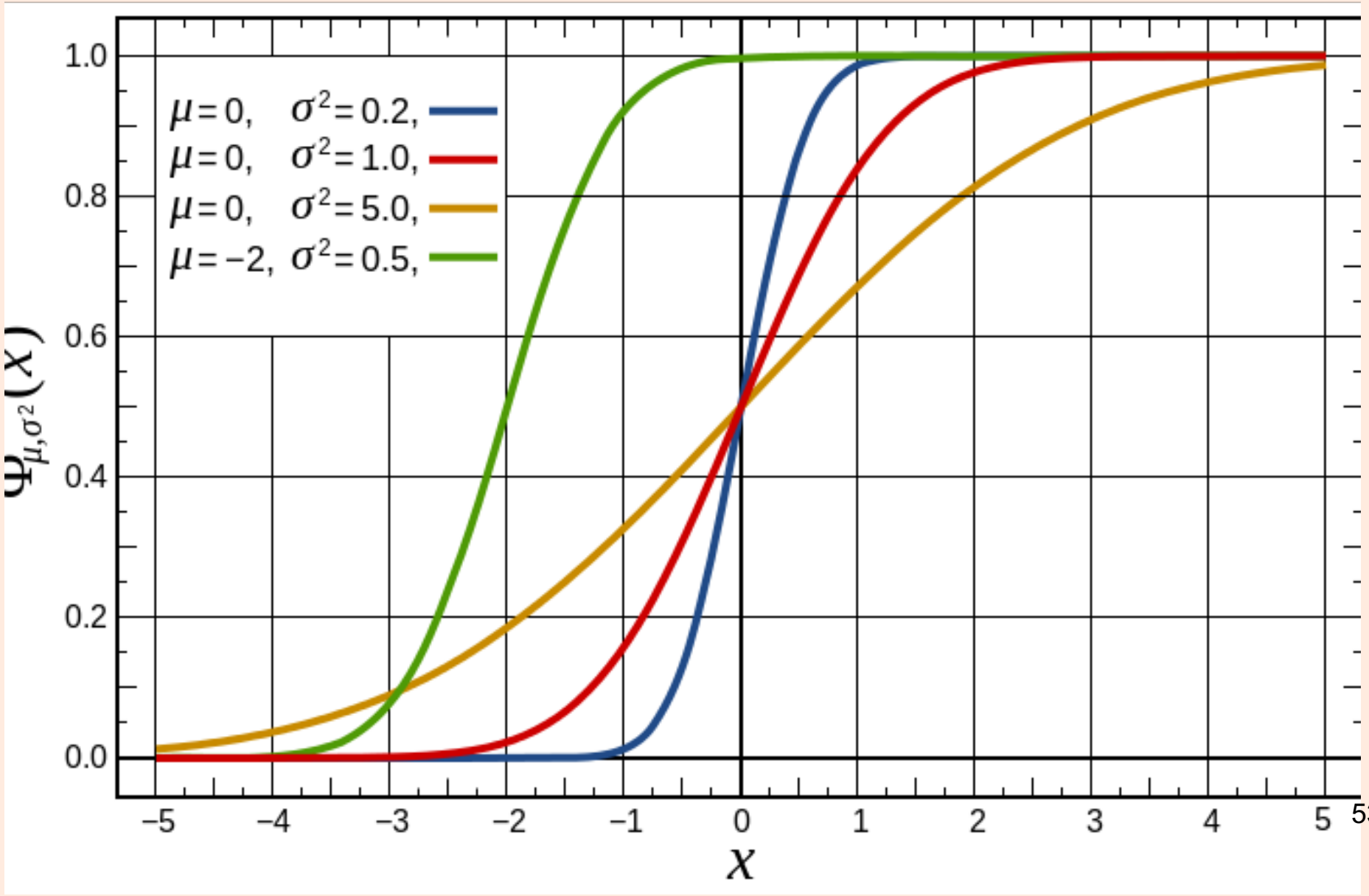


http://en.wikipedia.org/wiki/Heavy-tailed_distribution

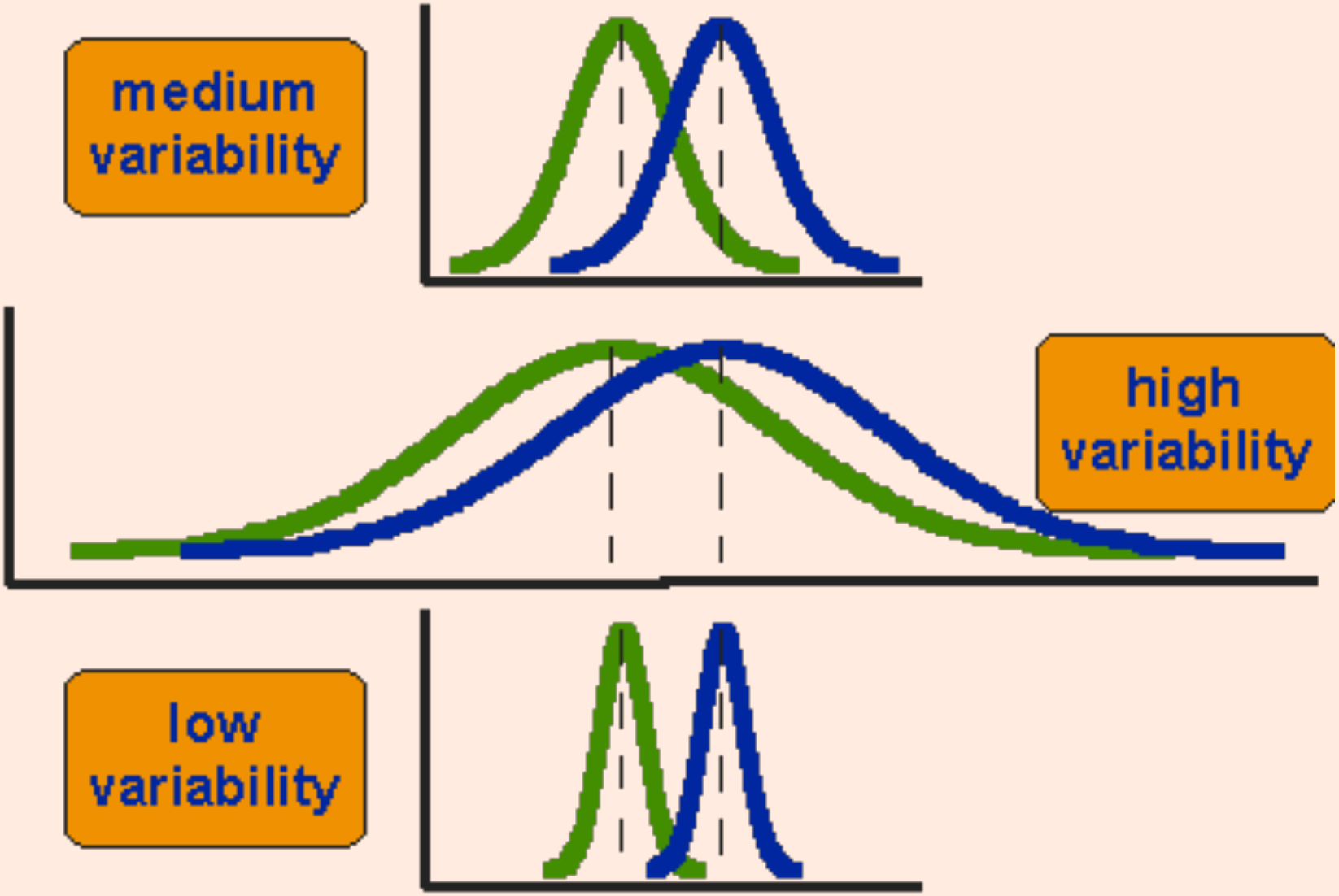
Probability Density



Cumulative…


$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$
$\Psi_{\mu,\sigma^2}(x)$
x
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5

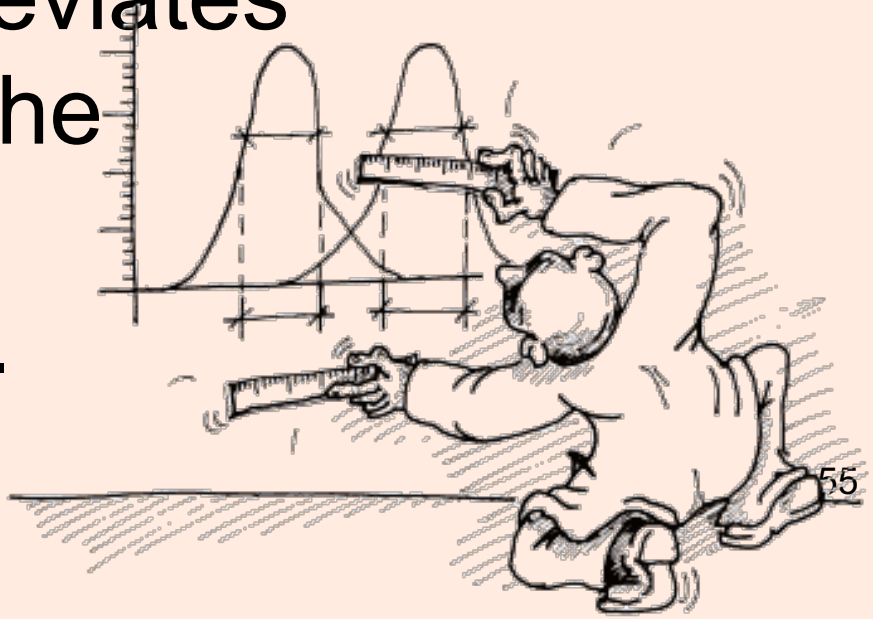
Variability in normal distributions



F-test

$F = S_1^2 / S_2^2$
where S_1 and S_2 are the sample variances.

The more this ratio deviates from 1, the stronger the evidence for unequal population variances.



T-test

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$

Note on Standard Error

- Versus standard deviation = SD (i.e. from the mean)
- SE ~ SD/sample size

- So, as size increases SE << SD !! Big data

Frequencies v. Probabilities

- Actual rate of occurrence in a sample or population – frequency
- Expected or estimate likelihood of a value or outcome – probability

- Coin toss – two outcomes (binomial)

  p= 0.5 (of “heads”)
- Male/Female
- Which US State you live in

Hypothesis

1. Write the original claim and identify whether it is the null hypothesis or the alternative hypothesis.
2. Write the null and alternative hypothesis. Use the alternative hypothesis to identify the type of test.
3. Write down all information from the problem.
4. Find the critical value using the tables
5. Compute the test statistic
6. Make a decision to reject or fail to reject the null hypothesis. A picture showing the critical value and test statistic may be useful.
7. Write the conclusion.

Hypothesis

- What are you exploring?
- Regular data analytics features ~ well defined hypotheses
  - Big Data messes that up - discuss

- E.g. Stock market performance / trends versus unusual events (crash/ boom):
  - Populations versus samples – which is which?
  - Why?
- E.g. Election results are predictable from exit polls

Null and Alternate Hypotheses

- H0 - null
- H1 – alternate
- If a given claim contains equality, or a statement of no change from the given or accepted condition, then it is the null hypothesis, otherwise, if it represents change, it is the alternative hypothesis.

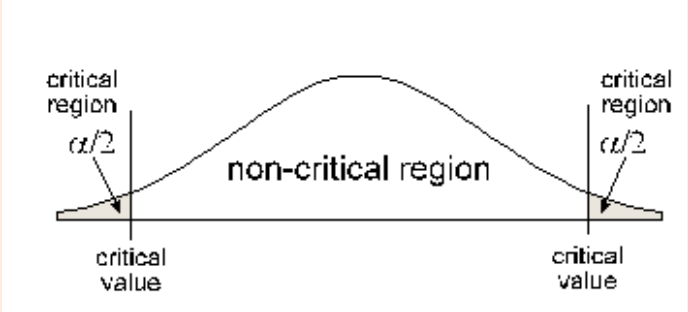
- It never snows in Troy in January
- Students will attend their scheduled classes

P-value

- One common way to evaluate significance, especially in R output
  - approaches hypothesis testing from a different manner. Instead of comparing z-scores or t-scores as in the classical approach, you're comparing probabilities, or areas.
- The level of significance (alpha) is the area in the critical region. That is, the area in the tails to the right or left of the critical values.

P-value

- The p-value is the area to the right or left of the test statistic.
  - If it is a two tail test, then look up the probability in one tail and double it.
- If the test statistic is in the critical region, then the p-value will be less than the level of significance.
  - It does not matter whether it is a left tail, right tail, or two tail test. This rule always holds.



Accept or Reject?

- **Reject the null hypothesis if the p-value is less than the level of significance.**

- **You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance. In English – you accept that there is no relation!**

- **Typical significance 0.05 (!)**

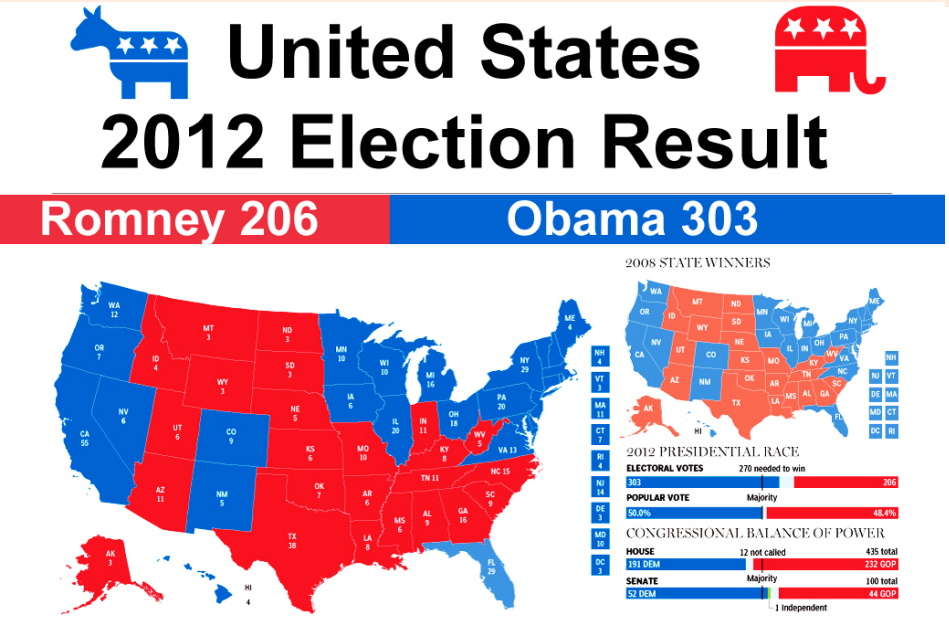
E.g. Election prediction

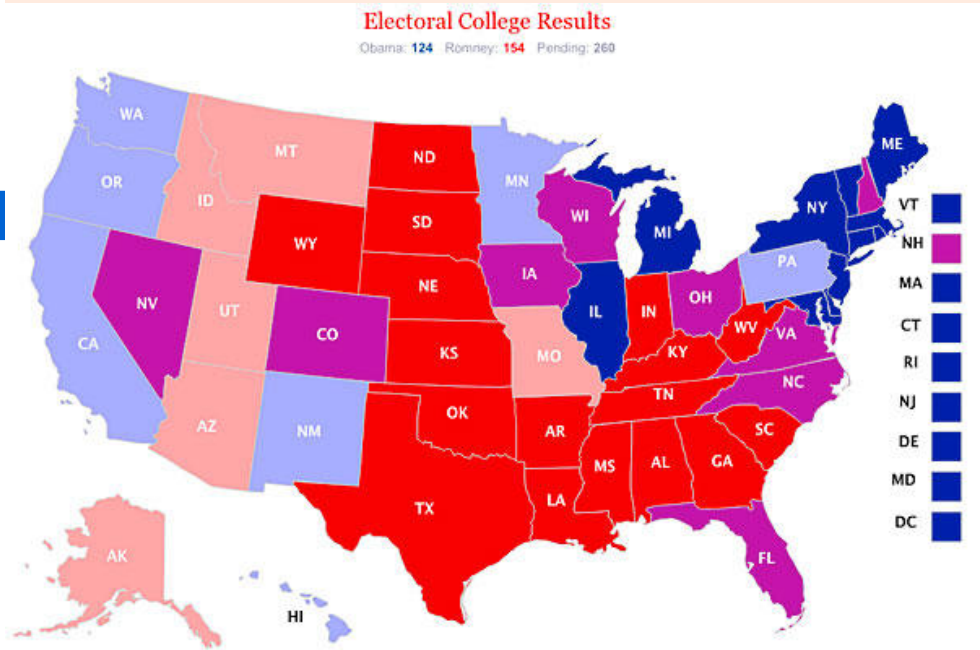
- Exit polls versus election results
  - Human versus cyber
- How is the “population” defined here?

- What is the sample, how is it chosen?
  - What is described and how is that used to predict?
  - Are results categorized? (where from, M/F, age)
- What is the uncertainty?
  - It is reflected in the “sample distribution”
  - And controlled/ constraints by “sampling theory”

Bias difference: between cyber and human data

- 2012 (!not 2016) election results and exit polls
  - What are examples of bias in election results?
  - In exit polls?





Random Numbers

- Can a computer generate a random number?
- Can you?
- Origin – to reduce selection bias!
- In R – many ways – see help on Random {base} and get familiar with set.seed()

We will talk about this more later…