MATH 10

INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduate student.

Classes

- Tuesday and Thursday, 2.25 pm – 4.15 pm, Kemeny 108.

- Tentatively : lecture, optional in-class exercise + break, lecture.
- Current plan: put slides and exercises on our webpage after every lecture.

- X-hours : Wednesday, 4.35 pm – 5.25 pm, Kemeny 108.
- X-hours are **not** used, unless I announce in class and send out an email.
- That being said, do not schedule anything in that time slot that you cannot skip.

- Office hours : MWF 3 pm -4 pm, Kemeny 212. Feel free to just email me too.

Syllabus, Homework, Grades

- Free online textbook : http://onlinestatbook.com/
- Will cover most of the textbook but will skip some sections/chapters.

- 30% weekly homework. Will give out first one next Tues, due the following Tues.
- Each homework might have different points assigned but carry the same weight.
- Your other homework : read and understand the relevant chapters in the textbook.

- 30% midterm exam : in class, week 5, Thurs, Chapter 1 to 11
- 40% final exam : in class, week 10, Tues, Chapter 1 to 19, cumulative

- Website: https://math.dartmouth.edu/~m10s18/ (all these info will be on the website)

Week 1

- **Chapter 1 – Introduction** ← **Today's lecture.**

What is Statistics? Why do you need to know Statistics?

Technical lingo and concepts: Sampling, variables, percentiles, scales, distributions, summation, linear transformations, logarithms.

- **Chapter 2 – Graphing Distributions**

Visualizing data containing qualitative and quantitative variables. Histograms etc.

- **Chapter 3 – Summarizing Distributions**

Central tendency: mean, median, mode. Variability: variance.

Chapter 1 - Introduction

- **Statistics** : the mathematics of working with data.

- How should we analyze or interpret data?
- How should we present data graphically?
- Given a set of data, how can mathematics be applied to give us more information?

- **Why study statistics?**

- Data science and big data. **cough* Facebook *cough**
- Statistics is *practical.*
- Statistical literacy is *important*.

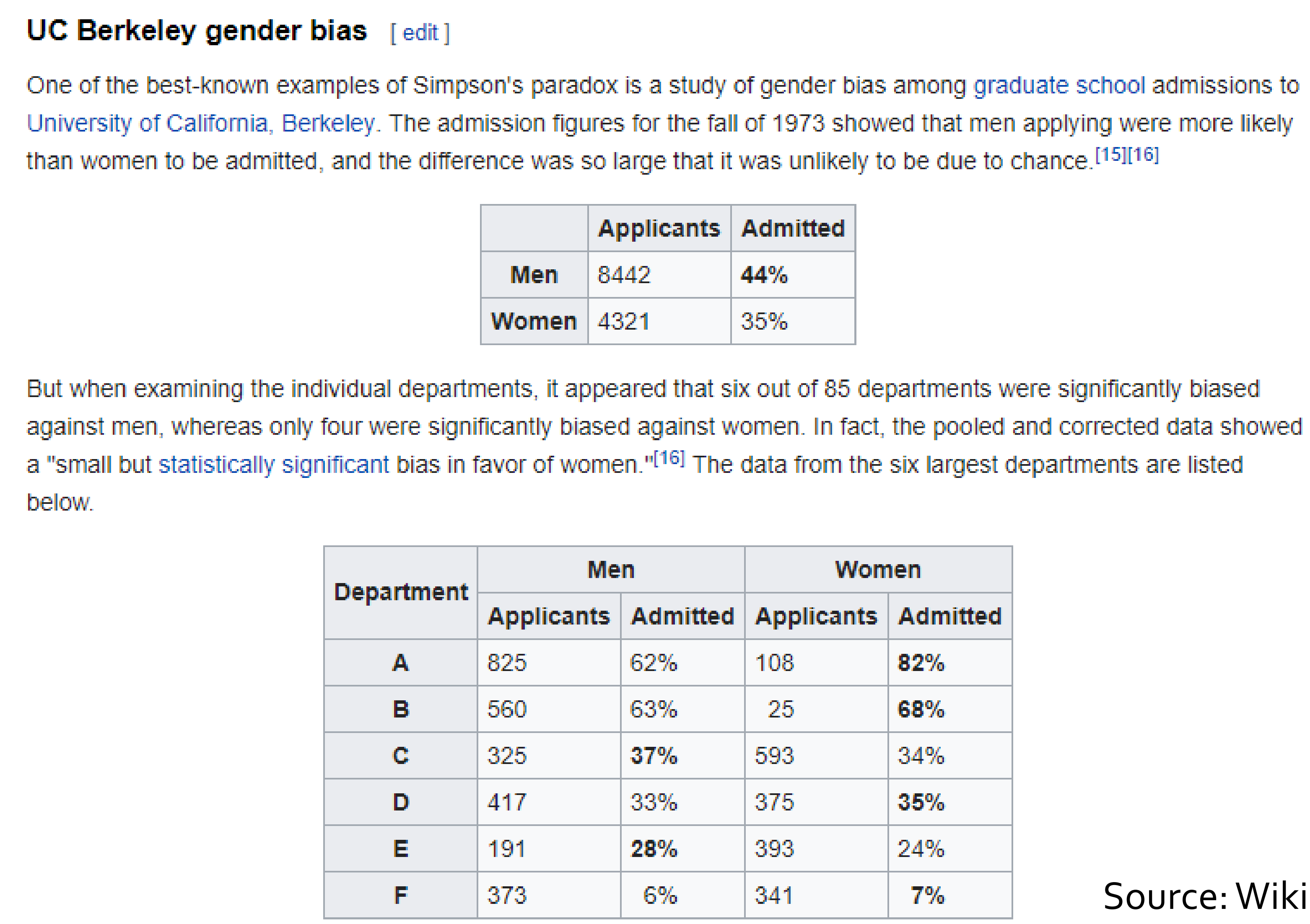
Statistics in social issues and the news: E.g. gender bias.

So…

Do we get out the pitchforks…

…or not?

¯_(ツ)_/¯



UC Berkeley gender bias [edit]

One of the best-known examples of Simpson's paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.[15][16]

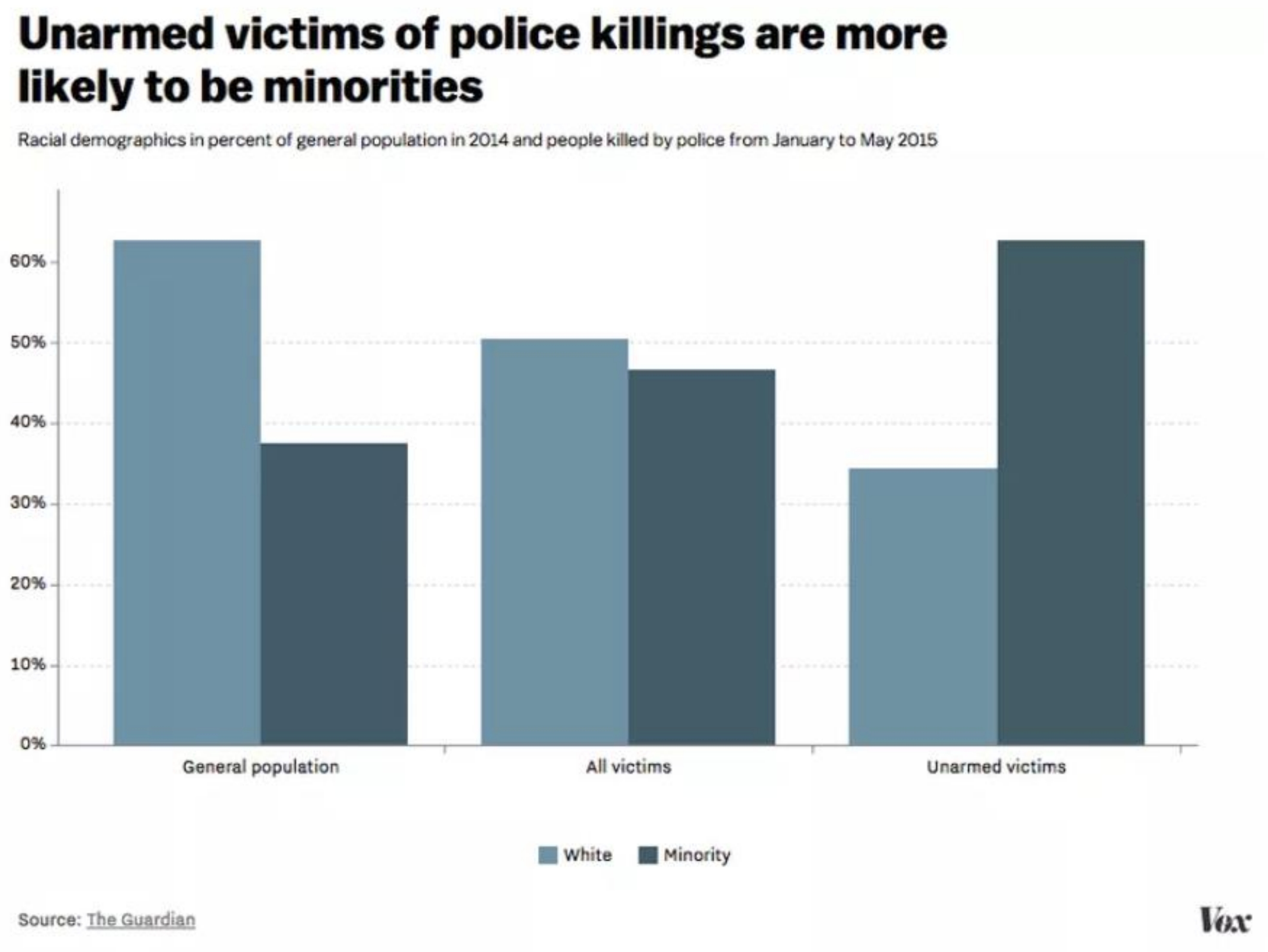
	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

But when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women."[16] The data from the six largest departments are listed below.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Source: Wikipedia

Statistics in social issues and the news: E.g. racial bias.



Unarmed victims of police killings are more likely to be minorities

Racial demographics in percent of general population in 2014 and people killed by police from January to May 2015

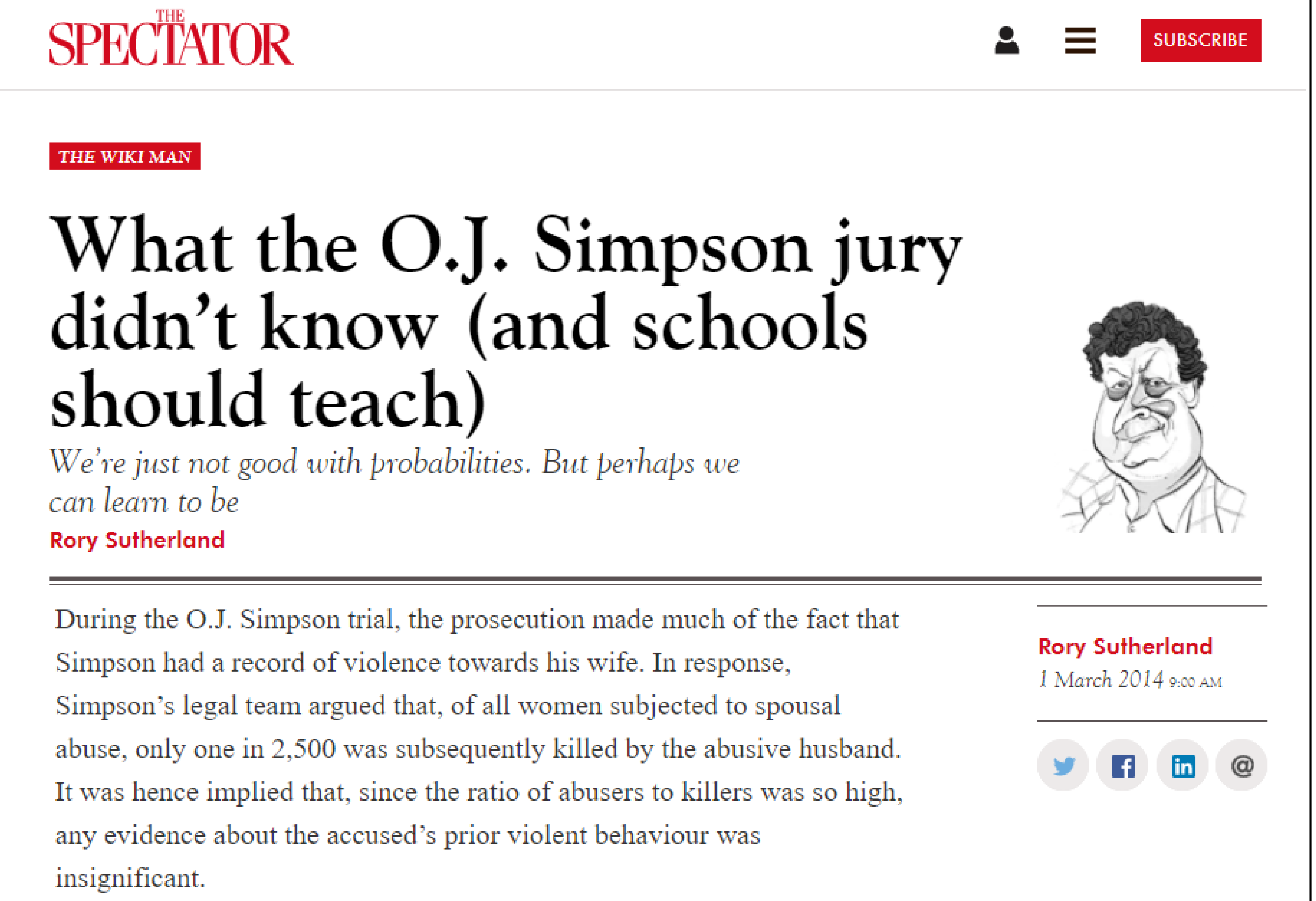
Source: The Guardian

Statistics in social issues and the news: E.g. the justice system.

"The truth is rarely pure and never simple."

— Oscar Wilde,

The Importance of Being Earnest



THE WIKI MAN

What the O.J. Simpson jury didn't know (and schools should teach)

We're just not good with probabilities. But perhaps we can learn to be

Rory Sutherland

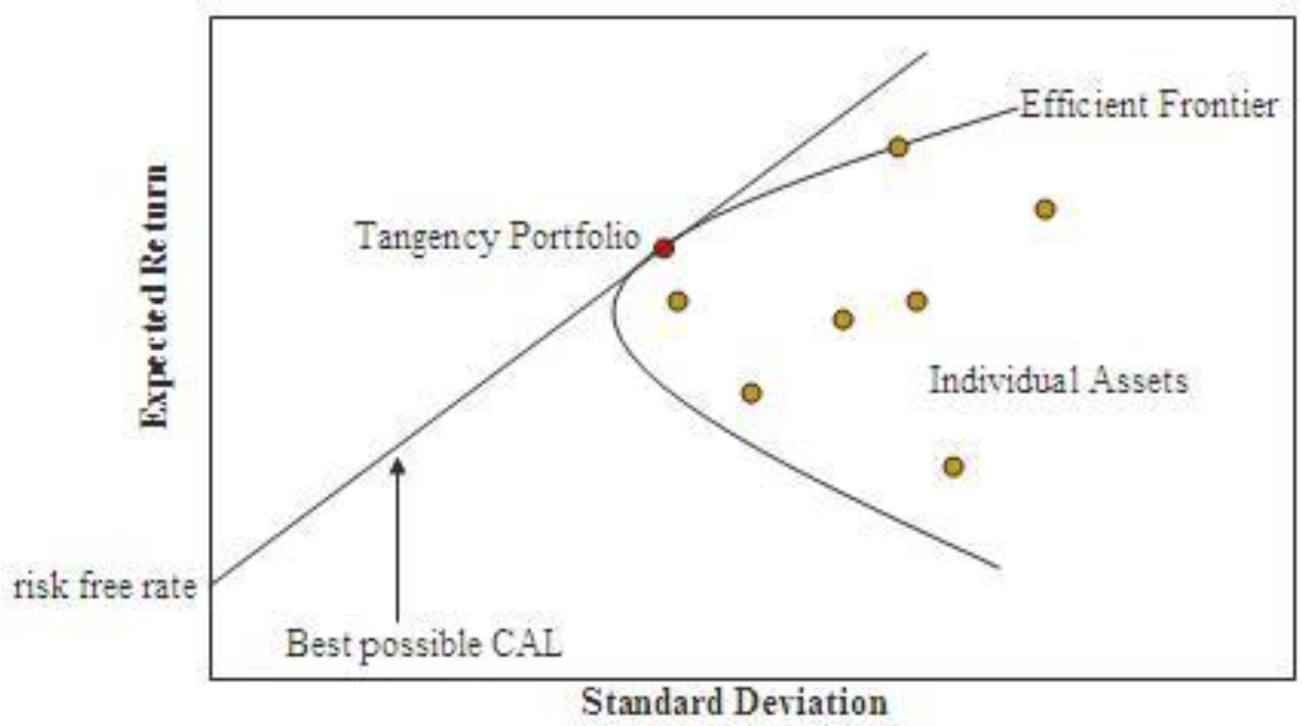
During the O.J. Simpson trial, the prosecution made much of the fact that Simpson had a record of violence towards his wife. In response, Simpson's legal team argued that, of all women subjected to spousal abuse, only one in 2,500 was subsequently killed by the abusive husband. It was hence implied that, since the ratio of abusers to killers was so high, any evidence about the accused's prior violent behaviour was insignificant.

Rory Sutherland
1 March 2014 9:00 AM

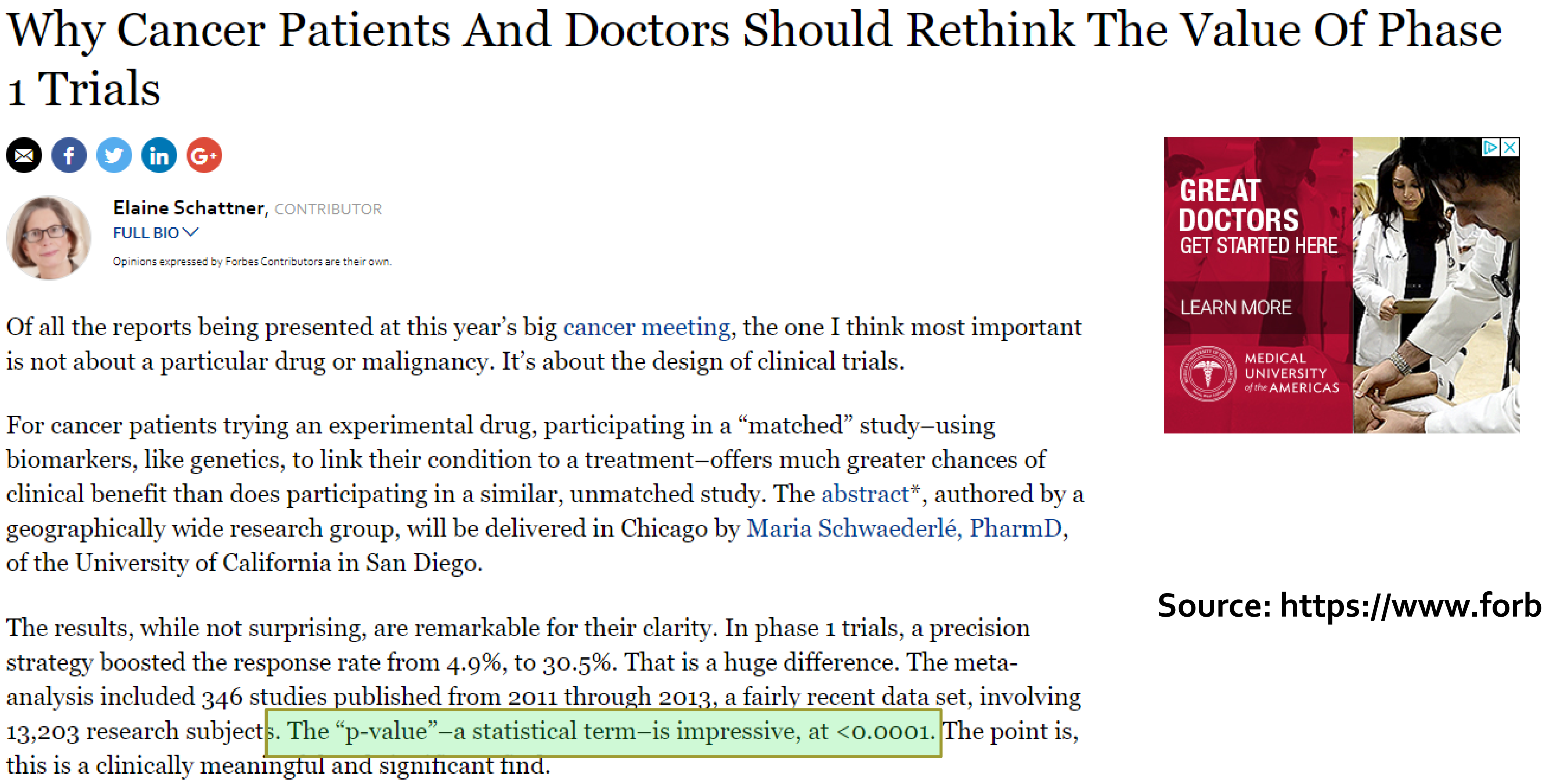
Statistics in investing and finance: E.g. risk-returns trade off

A common heuristic is that riskier investments should provide higher returns.

But how do we quantify the "returns" and "risk" of a company's stock?



Statistics in healthcare decisions: E.g. cancer treatments



Why Cancer Patients And Doctors Should Rethink The Value Of Phase 1 Trials

Elaine Schattner, CONTRIBUTOR
FULL BIO

Opinions expressed by Forbes Contributors are their own.

Of all the reports being presented at this year's big cancer meeting, the one I think most important is not about a particular drug or malignancy. It's about the design of clinical trials.

For cancer patients trying an experimental drug, participating in a "matched" study–using biomarkers, like genetics, to link their condition to a treatment–offers much greater chances of clinical benefit than does participating in a similar, unmatched study. The abstract*, authored by a geographically wide research group, will be delivered in Chicago by Maria Schwaederlé, PharmD, of the University of California in San Diego.

The results, while not surprising, are remarkable for their clarity. In phase 1 trials, a precision strategy boosted the response rate from 4.9%, to 30.5%. That is a huge difference. The meta-analysis included 346 studies published from 2011 through 2013, a fairly recent data set, involving 13,203 research subjects. The "p-value"–a statistical term–is impressive, at <0.0001. The point is, this is a clinically meaningful and significant find.

Source: https://www.forbes.com

Descriptive vs Inferential Statistics

- Chapter 1, Section 4
- **Descriptive statistics** provides a summary and description of the data.

- Chapter 1, Section 5
- **Inferential statistics**
- Collect data: sampling from a population (next slide).
- Make intelligent guesses using our data.

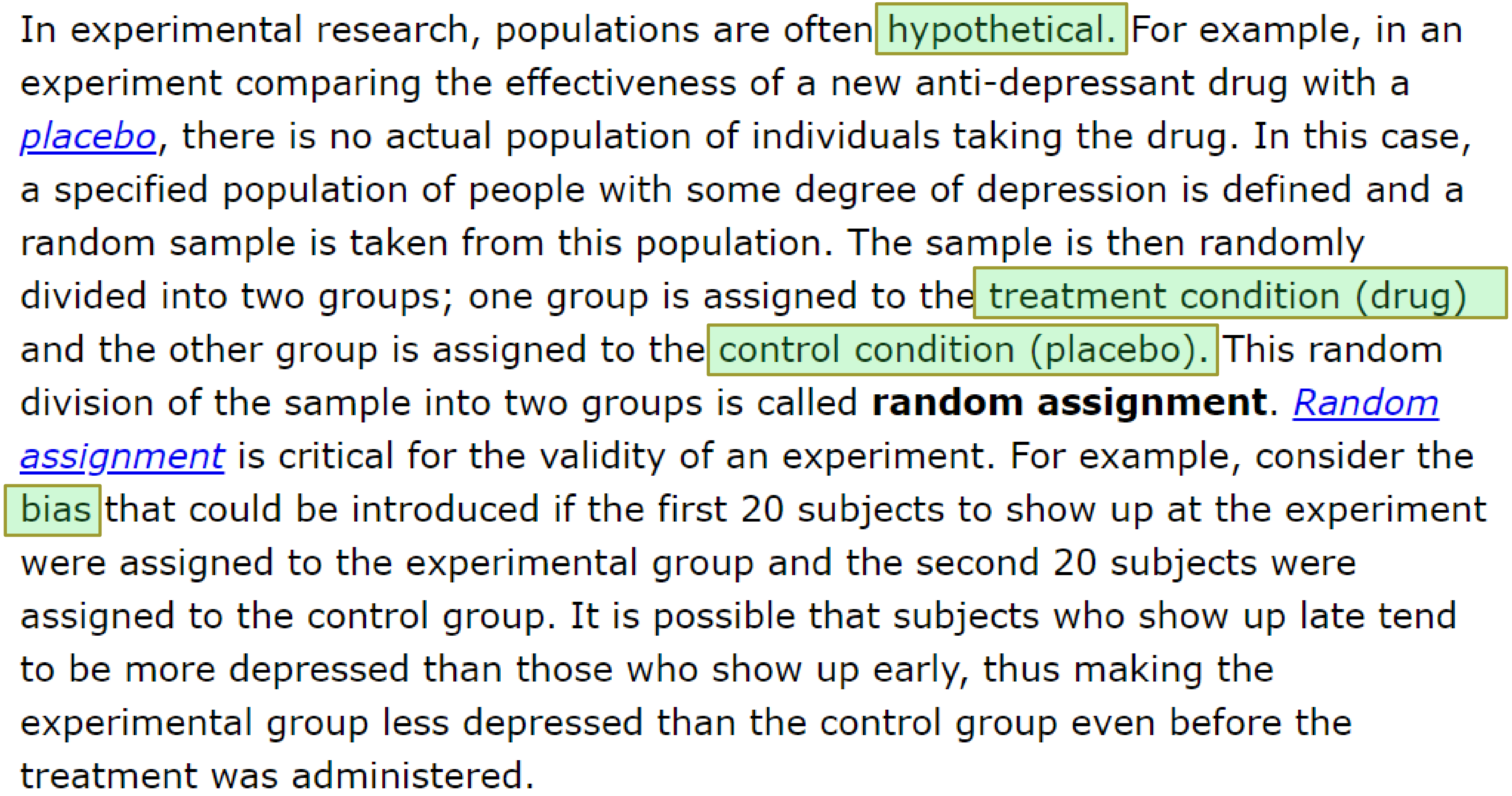
Population vs Sample

- Chapter 1, Section 5
- *"In statistics, we often rely on a sample --- that is, a small subset of a larger set of data --- to draw inferences about the larger set."*
- *"The larger set is known as the population from which the sample is drawn."*

How to sample?

- Simple random sampling.
- Random assignment – treatment and control group. → next slide
- Stratified sampling.

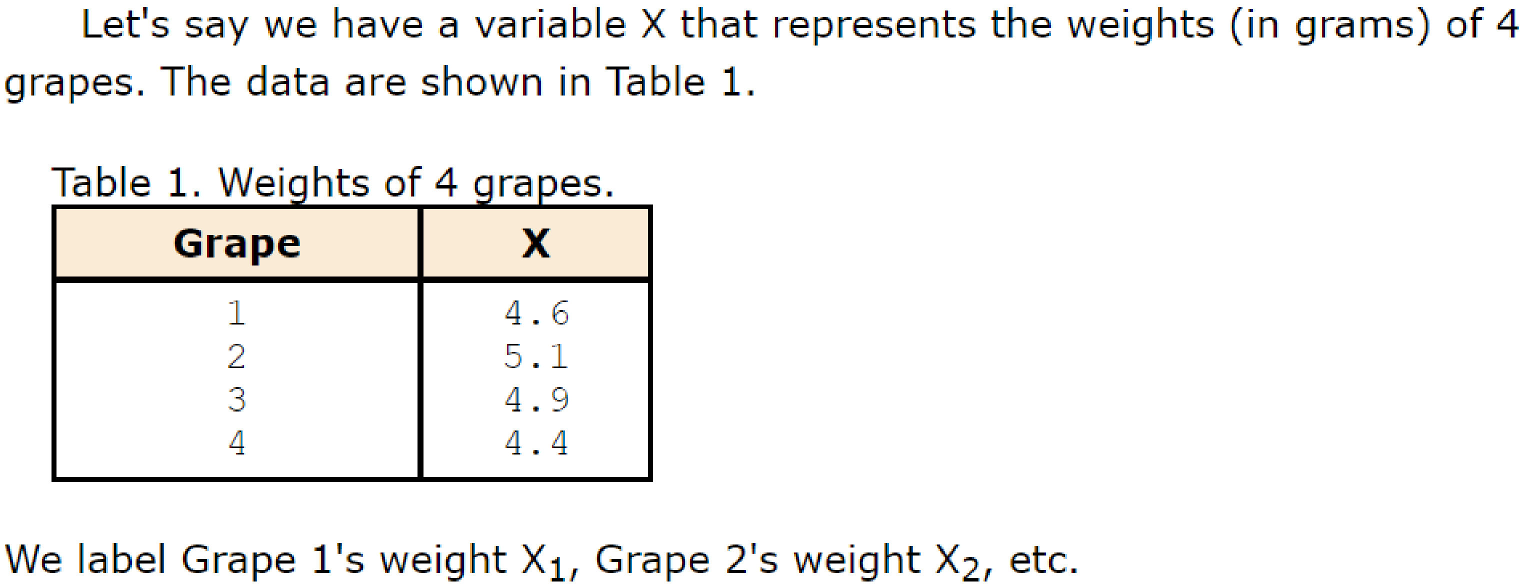
Random assignment of treatment and control groups



In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a *placebo*, there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (drug) and the other group is assigned to the control condition (placebo). This random division of the sample into two groups is called **random assignment**. *Random assignment* is critical for the validity of an experiment. For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

Variables

- Chapter 1, Section 7
- **Numerical variables**



Let's say we have a variable X that represents the weights (in grams) of 4 grapes. The data are shown in Table 1.

Table 1. Weights of 4 grapes.

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

We label Grape 1's weight X_1, Grape 2's weight X_2, etc.

- Be comfortable with using variables in place of numbers.

- Sometimes we might play with (random) variables that does not have numerical values attached to them yet.

E.g. X_1, X_2, X_3, X_4

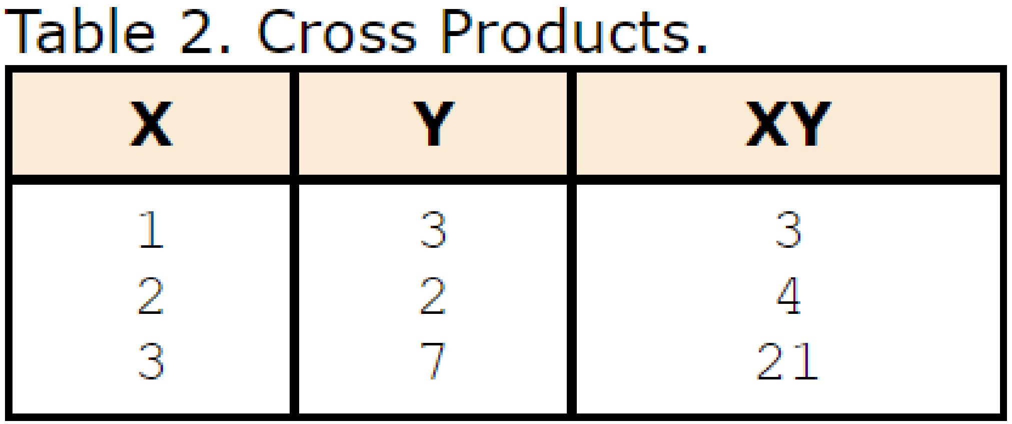


Table 2. Cross Products.

X	Y	XY
1	3	3
2	2	4
3	7	21

Things we can do to variables…

- Miscellaneous Topics in Chapter 1 – Section 12, 13, 14

- **Summation Notation.**
- **Linear Transformation.**
- **Logarithms.**

True story : on an exam, a student was asked "what does this mathematical operation ***A*** *do to the variable* ***x****?" The student answered "****A*** *does terrible things to* ***x*** *when no one is looking".*

Things we can do to variables…

- Chapter 1 – Section 12

- **Summation Notation.**

$$\sum_{i=1}^{n} X_i \;, \qquad \sum_{i=1}^{n} X_i^2 \;, \qquad \left(\sum_{i=1}^{n} X_i\right)^2$$

And etc…

Things we can do to variables…

- Chapter 1 – Section 12
- **Linear Transformation.**

Well, this is not exactly *linear*, but we'll follow the textbook's lingo.

Our use of the word linear and how it differs from general mathematic usage is mentioned in Chapter 3, section 18.

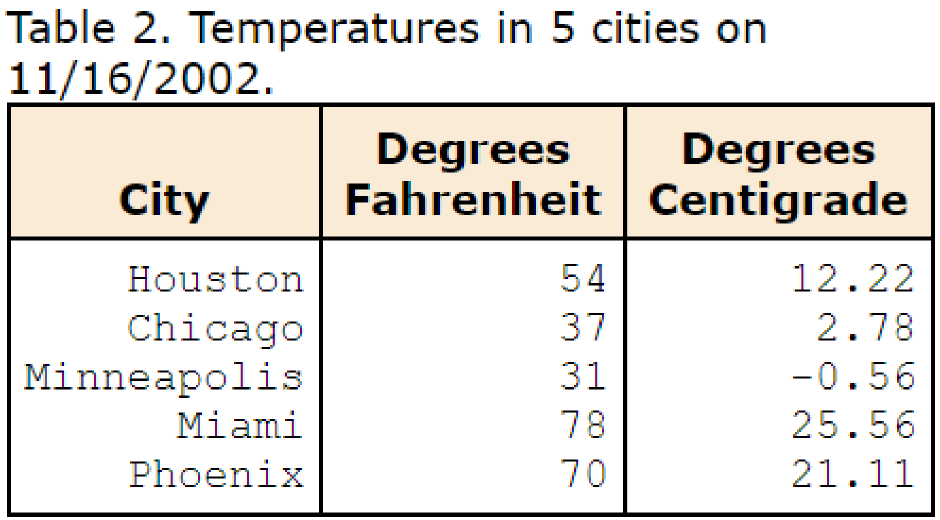
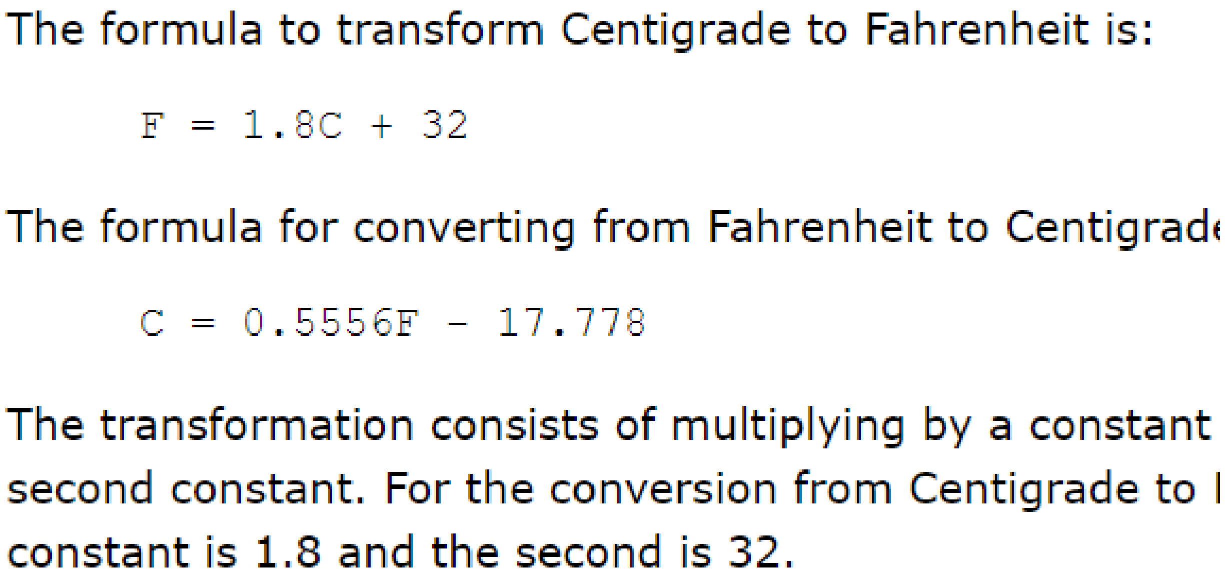


Table 2. Temperatures in 5 cities on 11/16/2002.

City	Degrees Fahrenheit	Degrees Centigrade
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11



The formula to transform Centigrade to Fahrenheit is:

$$F = 1.8C + 32$$

The formula for converting from Fahrenheit to Centigrad

$$C = 0.5556F - 17.778$$

The transformation consists of multiplying by a constant second constant. For the conversion from Centigrade to constant is 1.8 and the second is 32.

Things we can do to variables…

- Chapter 1 – Section 14
- Logarithms : An example of a non-linear transformation.

- $Log_b\ x =$ how many b we need to multiple together to get x
- Or, the power b needs to be raised to, to get x

- E.g. $Log_{10}\ 100 = 2$, $Log_{10}\ 1000 = 3$, $Log_2\ 8 = 3$.

- $Log_b\ (xy) = Log_b\ x$ + $Log_b\ y$
- $Log_b\ \left(\frac{x}{y}\right) = Log_b\ x$ - $Log_b\ y$

BREAK TIME!   \o/

Variables - Chapter 1, Section 7

- **Independent and dependent variables.**

Identify them in your model/conjecture.

Many examples in your textbook. :3

- **Qualitative and quantitative.**

Qualitative/categorical : hair color, movie preferences etc.

Quantitative : numbers!

- **Discrete and continuous.**

Percentiles

- No standard universal definition.
- E.g. standardized tests (SAT, GRE) percentiles.
- Rough intuition : “if the 65th percentile score is 50, then 65% of the people scored lower than 50”.

- We will follow the textbook’s preferred definition.

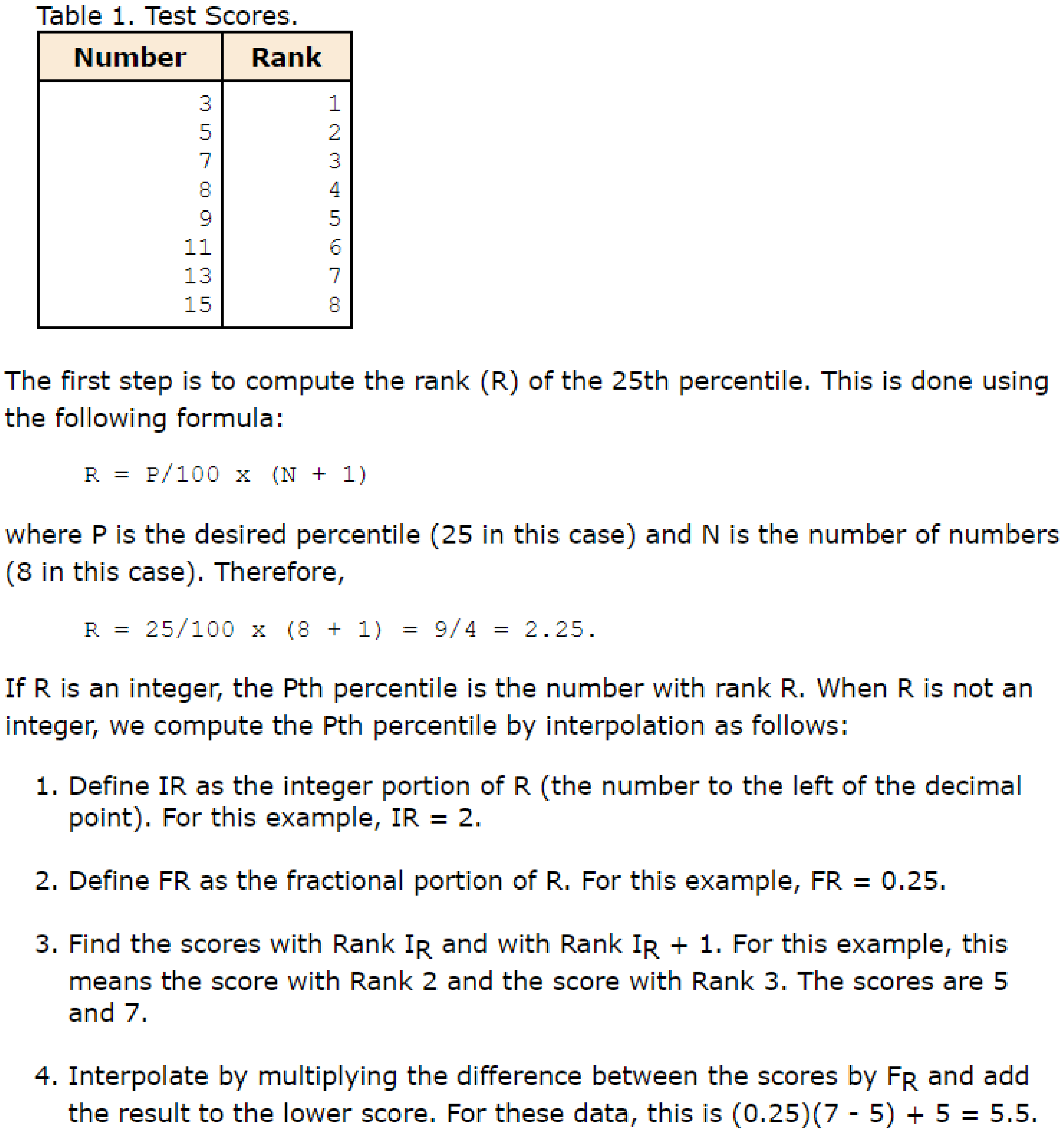


Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = P/100 \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8 + 1) = 9/4 = 2.25.$$

If R is an integer, the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, IR = 2.

2. Define FR as the fractional portion of R. For this example, FR = 0.25.

3. Find the scores with Rank I_R and with Rank $I_R + 1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.

4. Interpolate by multiplying the difference between the scores by F_R and add the result to the lower score. For these data, this is (0.25)(7 - 5) + 5 = 5.5.

Distributions

- How much data is in each type/category.
- We often want a visual representation of this.
- Symmetric vs skewed distribution. Chapter 3, section 11.

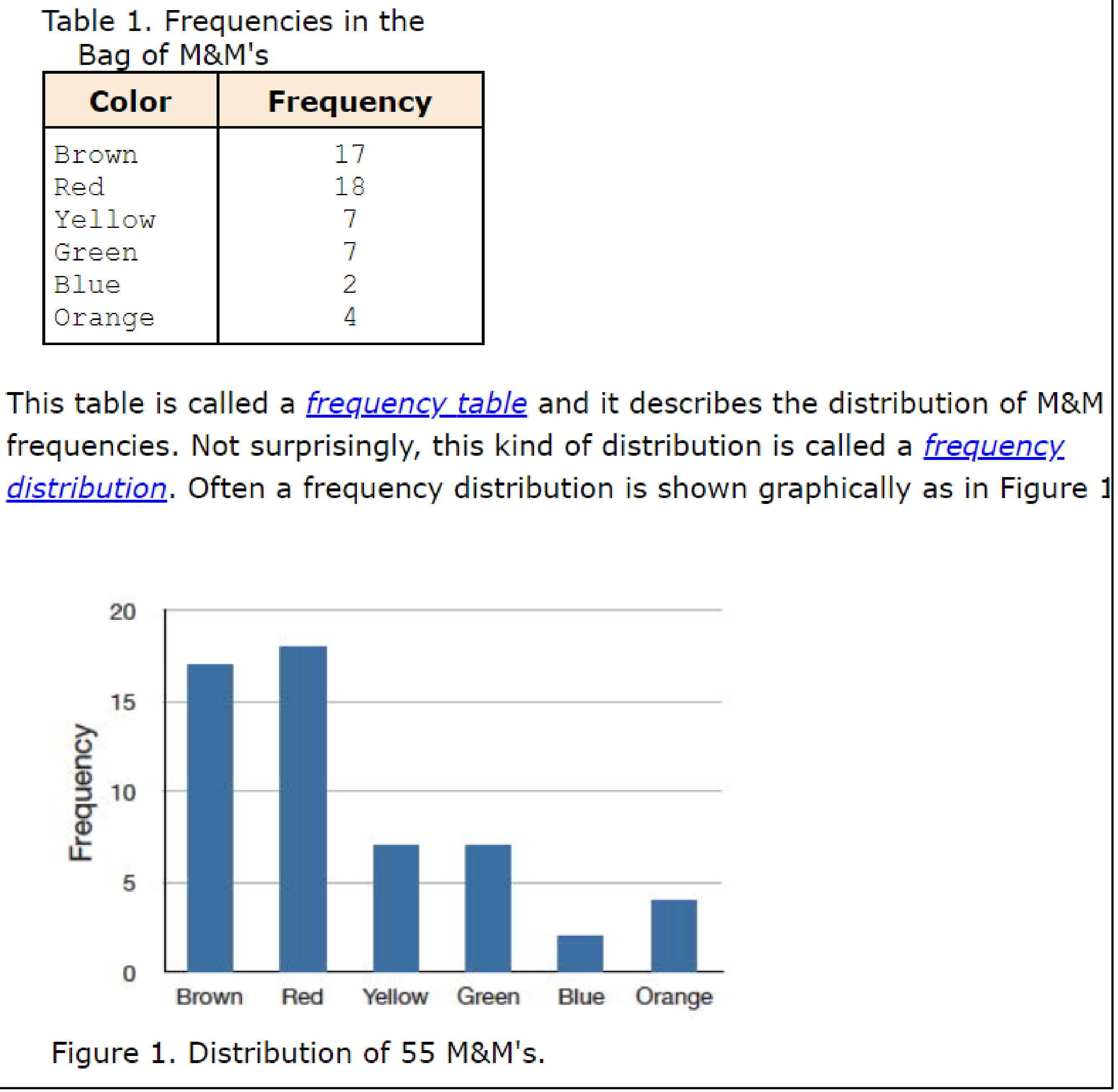


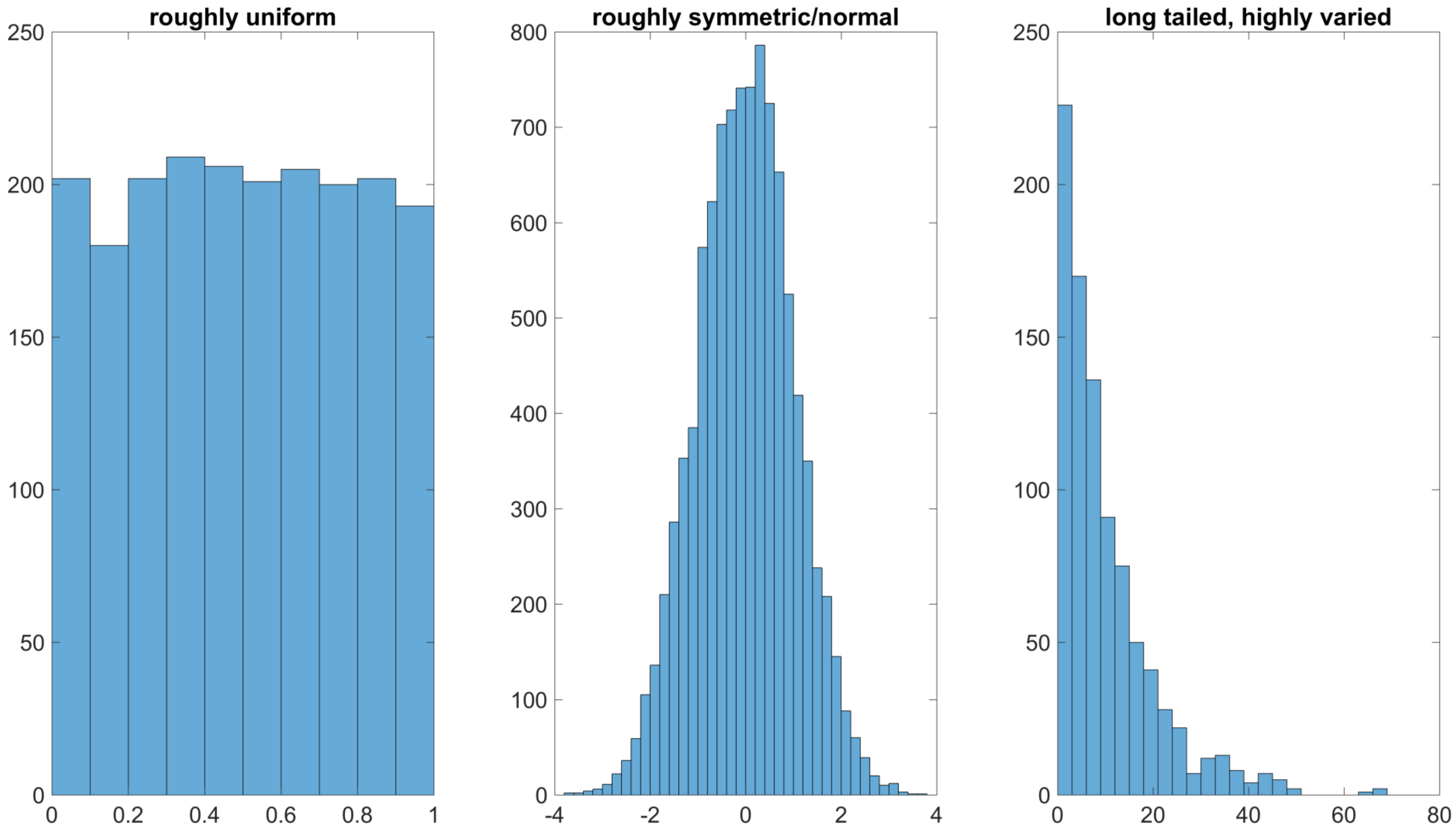
Table 1. Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a *frequency table* and it describes the distribution of M&M frequencies. Not surprisingly, this kind of distribution is called a *frequency distribution*. Often a frequency distribution is shown graphically as in Figure 1

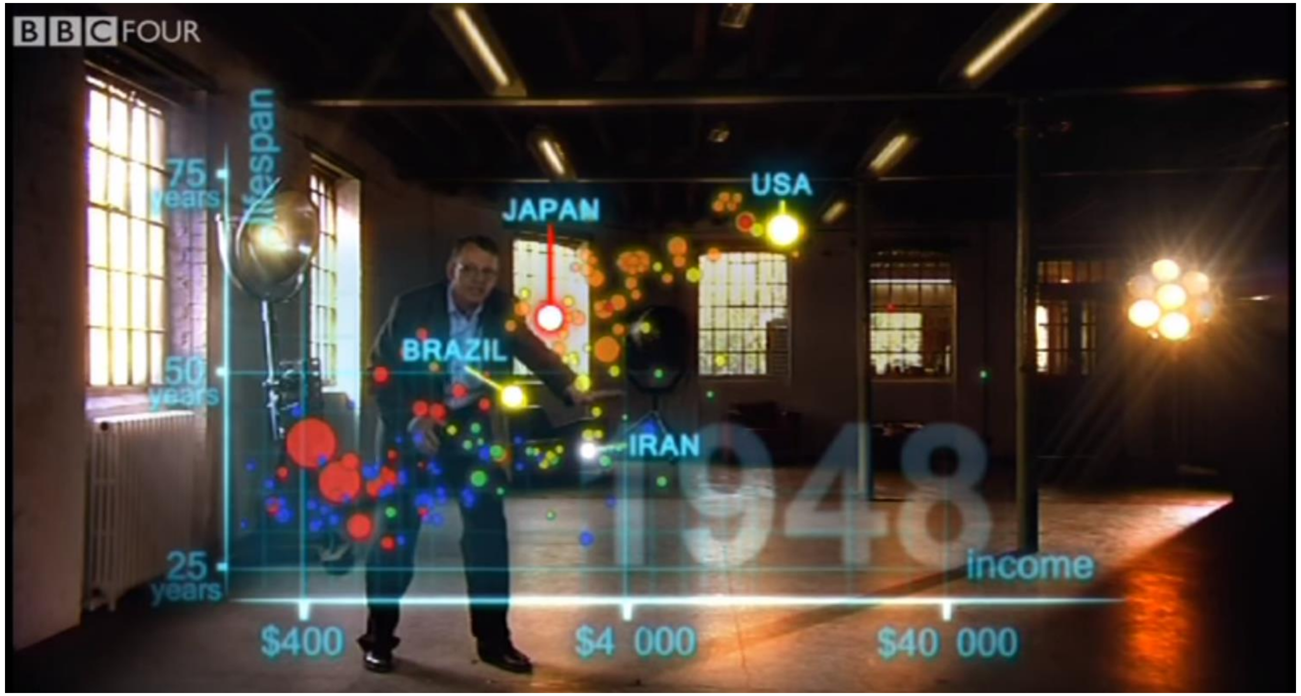
Figure 1. Distribution of 55 M&M's.

Distributions often seen in the wild…



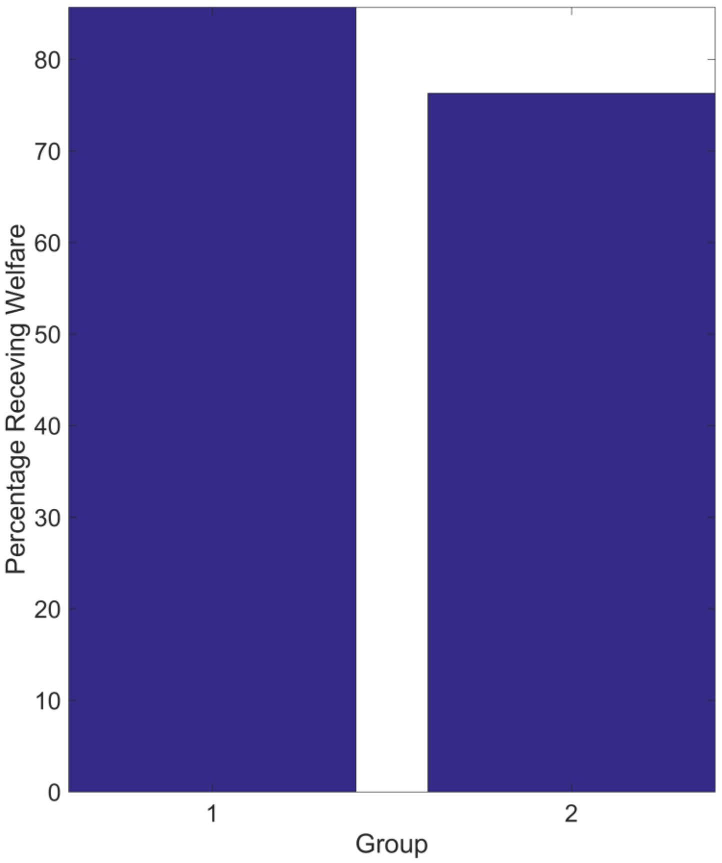
Chapter 2 – Graphing Distributions

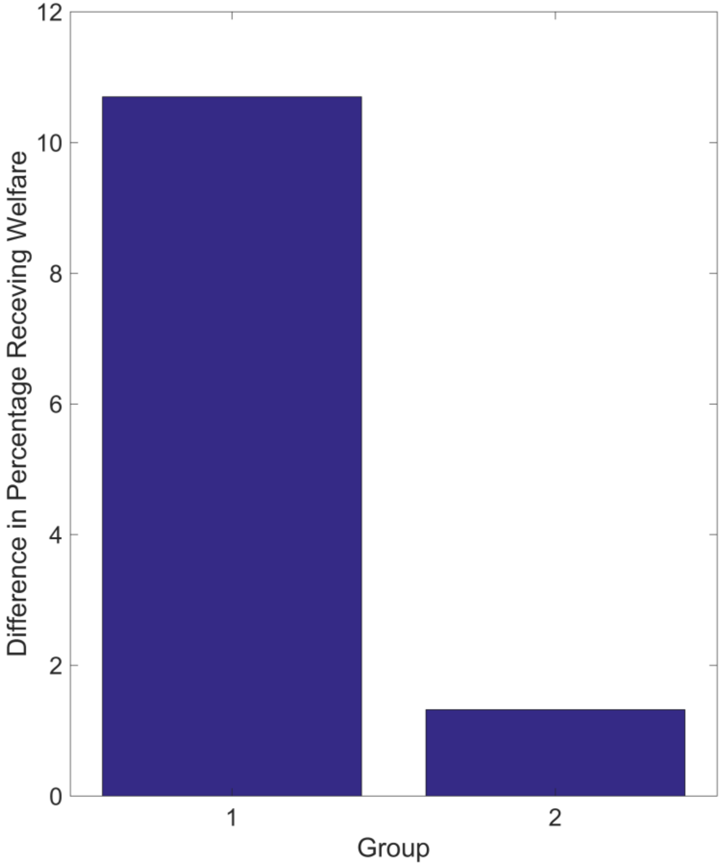
- Visualizing data helps us see patterns, support our conjectures, and can also help *sell* our ideas.
- https://www.youtube.com/watch?v=jbkSRLYSojo
- Hans Rosling, Statistician, on the BBC YouTube channel.



Chapter 2 – Graphing Distributions

- Unfortunately, sometimes visualization sell our ideas a little bit *too well*.





erratum: the "difference" in the 2nd figure is the difference of group 1 and 2 from some other percentage (70% in this case).

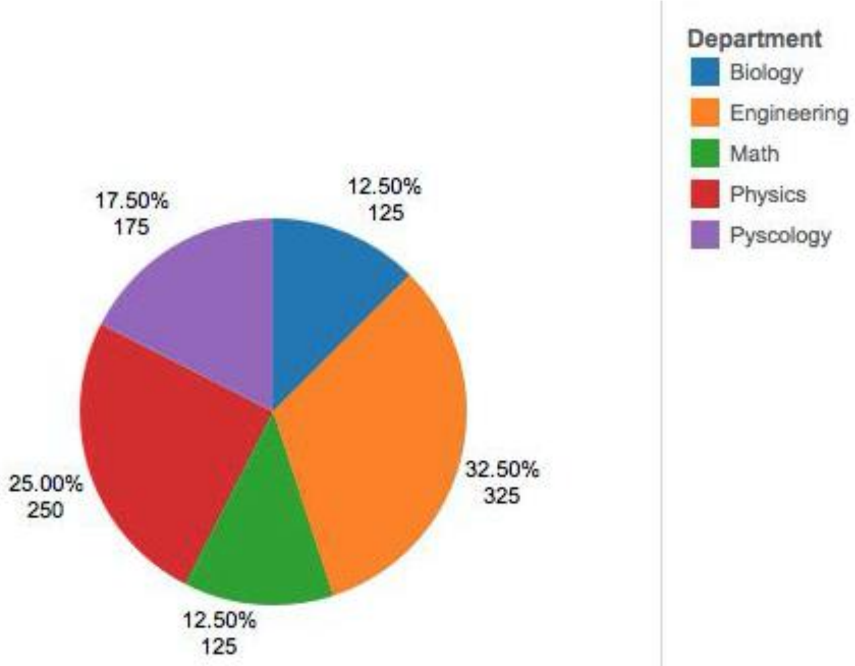
E.g. difference from the national average.

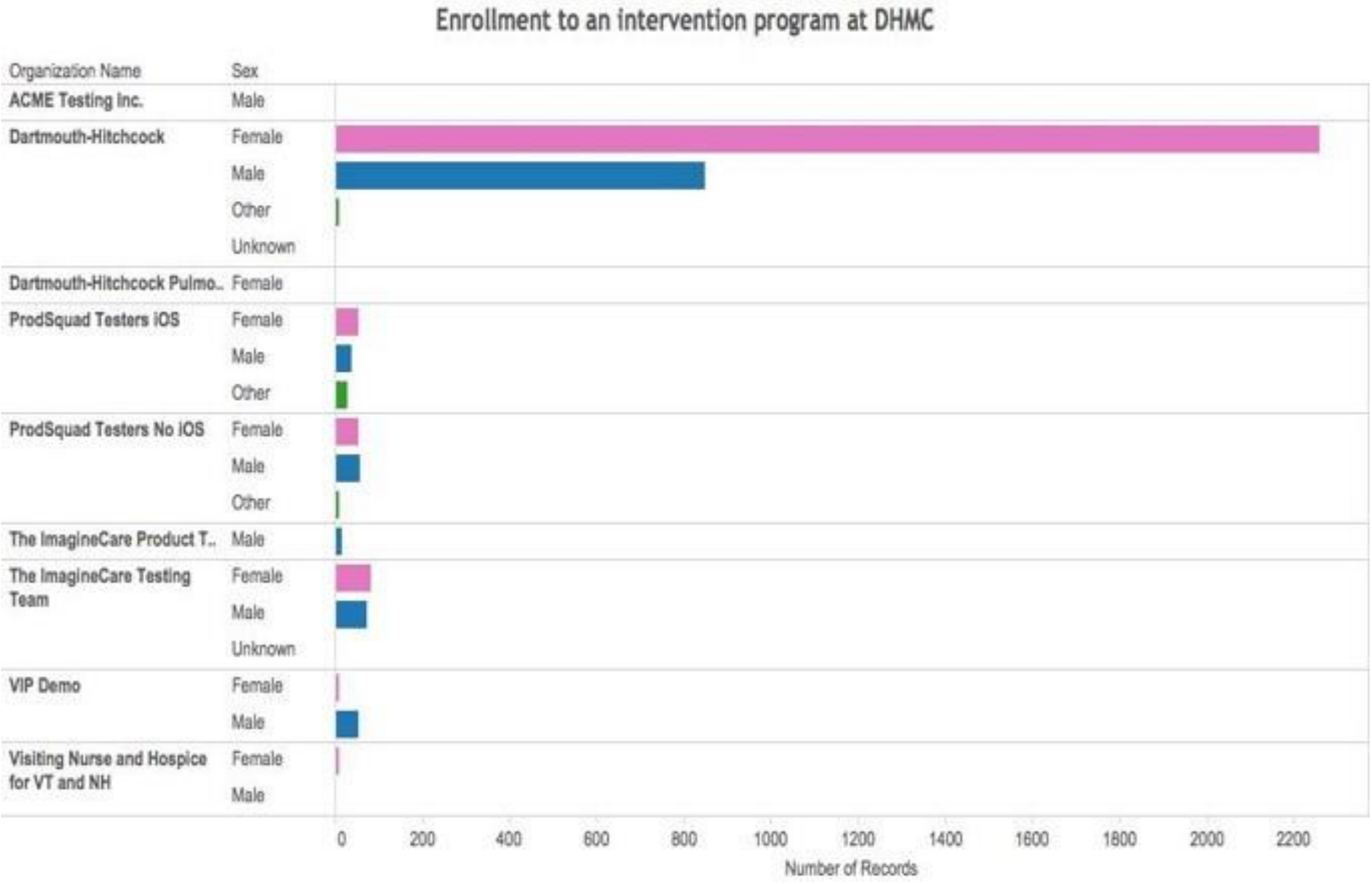
Chapter 2

- **Qualitative variables**
- Not numerical. Usually categories. E.g. hair color, favorite movie etc.
- Use frequency tables, pie charts, bar charts to visualize.

- **Qualitative variables**

Department	Enrollment
Physics	250
Math	125
Engineering	325
Biology	125
Phycology	175





Chapter 2

- **Qualitative variables**
- Not numerical. Usually categories. E.g. hair color, favorite movie etc.
- Use frequency tables, pie charts, bar charts to visualize.

- **Quantitative variables**
- Numbers! → we will see a lot of these in this course.
- Use stem and leaf, histograms, frequency polygons, box plots, bar charts, line graphs, dot plots.
- We will talk about histograms first then go on to Chapter 3 before giving a quick tour of the rest.

Quantitative variables – Histogram Part 1

- Very useful for visualizing the shape of a distribution when number of observations is large.
- **Textbook's example** : 642 students, scores ranged from 46 to 167. A simple frequency table will contain over 100 rows.
- Sort the N obervations into bins or classes intervals. For this course, we'll stick to the same width for each class.
- How many classes or bins? Trial and error → a.k.a. "the eyeball method".
- **Sturges' Rule** : as close to $(1 + log_2(N))$ classes as possible.
- **Rice Rule** : $2\ \sqrt[3]{N}$ classes. → can differ greatly from Sturge's Rule.
- These good to know, but don't worry, we won't ask you to memorize and state these in the exam.

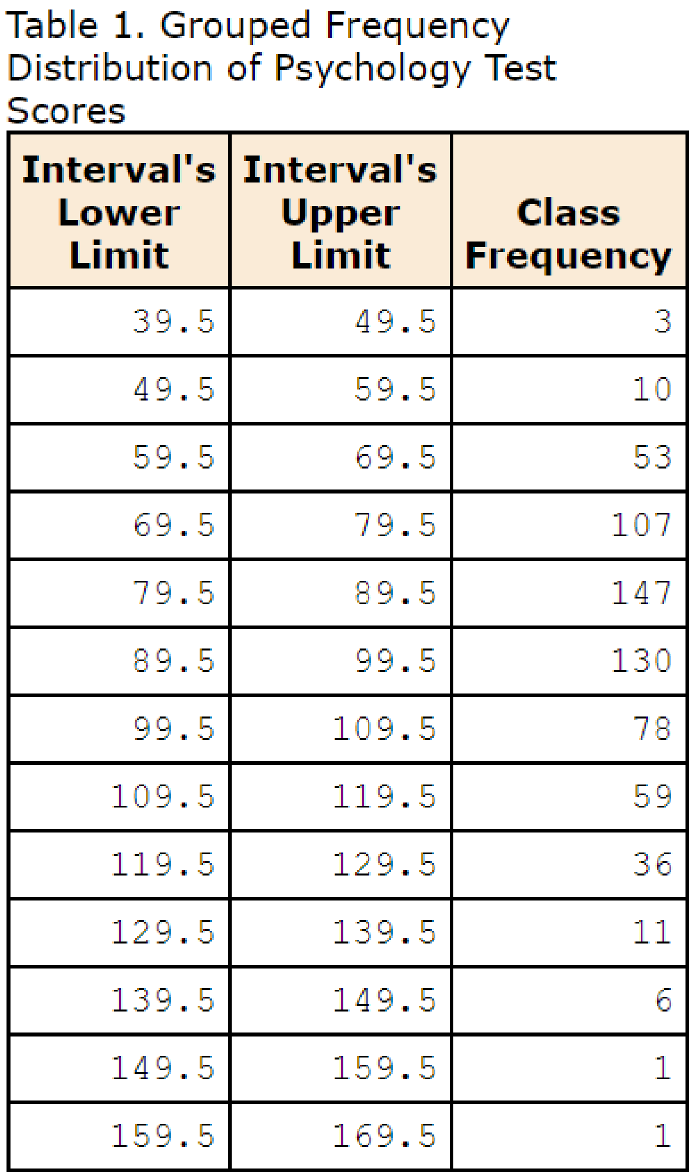


Table 1. Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Quantitative variables – Histogram Part 2

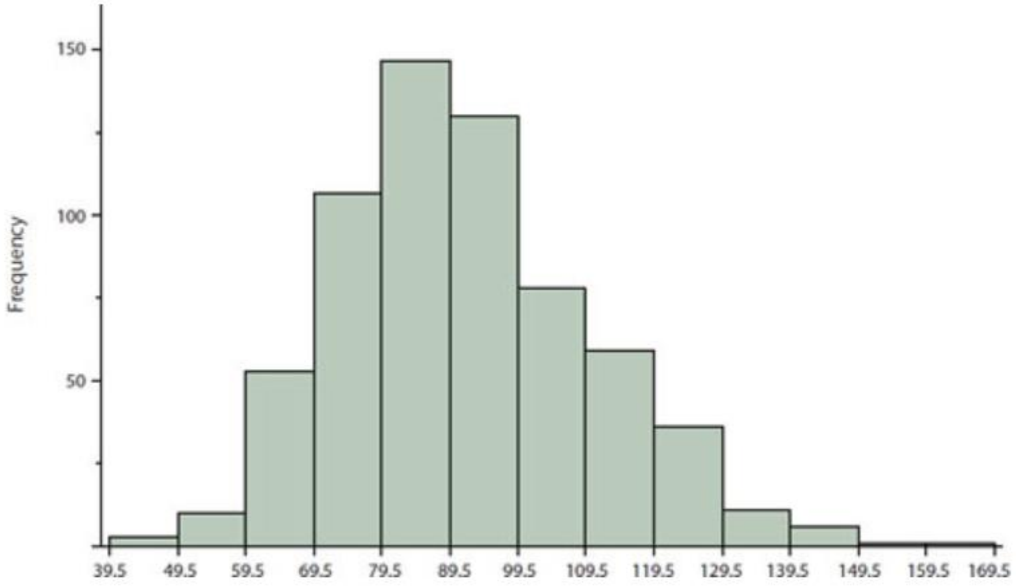
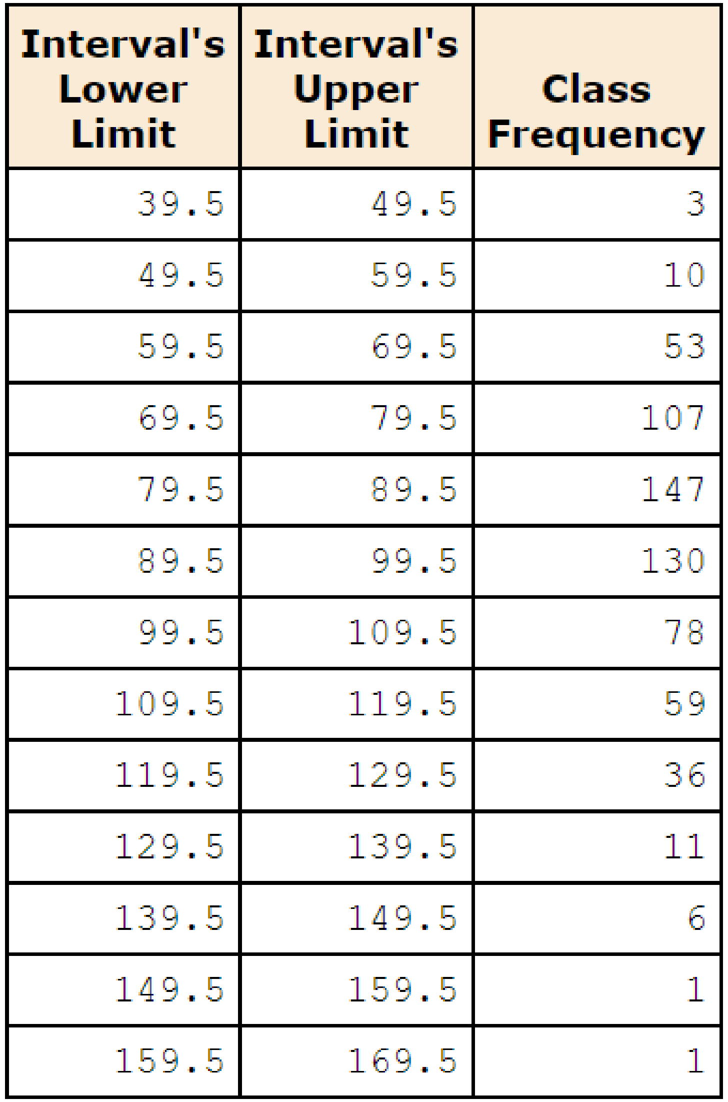


Figure 1. Histogram of scores on a psychology test.

- Vertical axis is the frequency or count for each class/bin.
- We can divide frequency by total number of observations, to get relative frequencies or proportions instead.
- E.g. Relative frequency or proportion of scoring between 69.5 and 79.5 is $107/642 = 0.1667$.

Table 1. Grouped Frequency Distribution of Psychology Test Scores



Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1