INTRODUCTION TO INFERENCE

Many Ways to Describe

- Proportions, Percentages, Ratios, Percentage Change
- Central Tendency: Mode, Median, Mean
- Dispersion: Range, Variance, Standard Deviation

Tables and Graphs

- Frequency Tables, Pie Charts, Bar Charts, Histograms
- Can be very revealing, showing:
  - Dispersion
  - Proportions
  - Probabilities

Appropriate Use

- Depends on Level of Measurement
  - **Nominal:** data that are not numerical in any way (species of tree, type of pet, zip code)
  - **Ordinal:** may or may not be numerical, but values have a definite order
  - **Interval-Ratio:** data that are numerical

Descriptives Statistics

Technique	*Nominal*	*Ordinal*	*Interval-Ratio*
Proportion/%	✓	✓	✓
Percentiles		✓	✓
Ratios	✓	✓	✓
% Change	✓	✓	✓
Mode	✓	✓	✓
Median		✓	✓
Mean			✓
Range		✓	✓
Variance/SD			✓

Tables and Graphs

Technique	*Nominal*	*Ordinal*	*Interval-Ratio*
Frequency Table	✓	✓	
Pie Chart	✓	✓	
Bar Chart	✓	✓	
Histogram			✓

The Mean

- Most commonly used statistic
- **Pros:** all values are used to calculate it; it is mathematically the center of all the values (the sum of differences between the mean and each value is always 0)
- **Cons:** very sensitive to outliers; gives you no sense of the distribution

Critically Interpreting the Mean

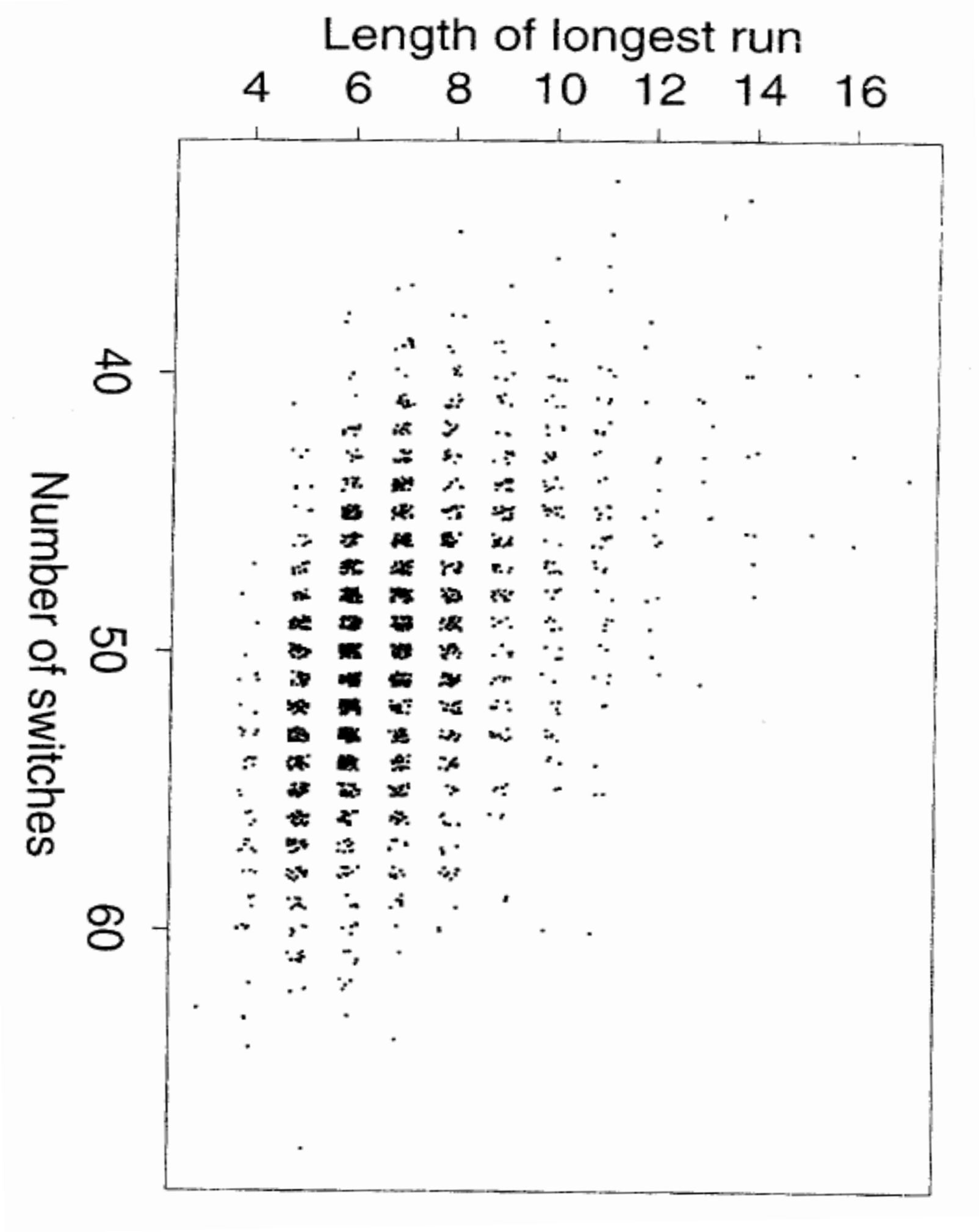
- **Median:** compared to the mean shows whether there is skew
- **Standard Deviation / Variance:** the most precise measure of dispersion
- **Histograms:** visual representation of the dispersion
- **Dispersion** is important because it shows how well or poorly the mean represents the data (do data cluster around the mean, or not?)

Probability

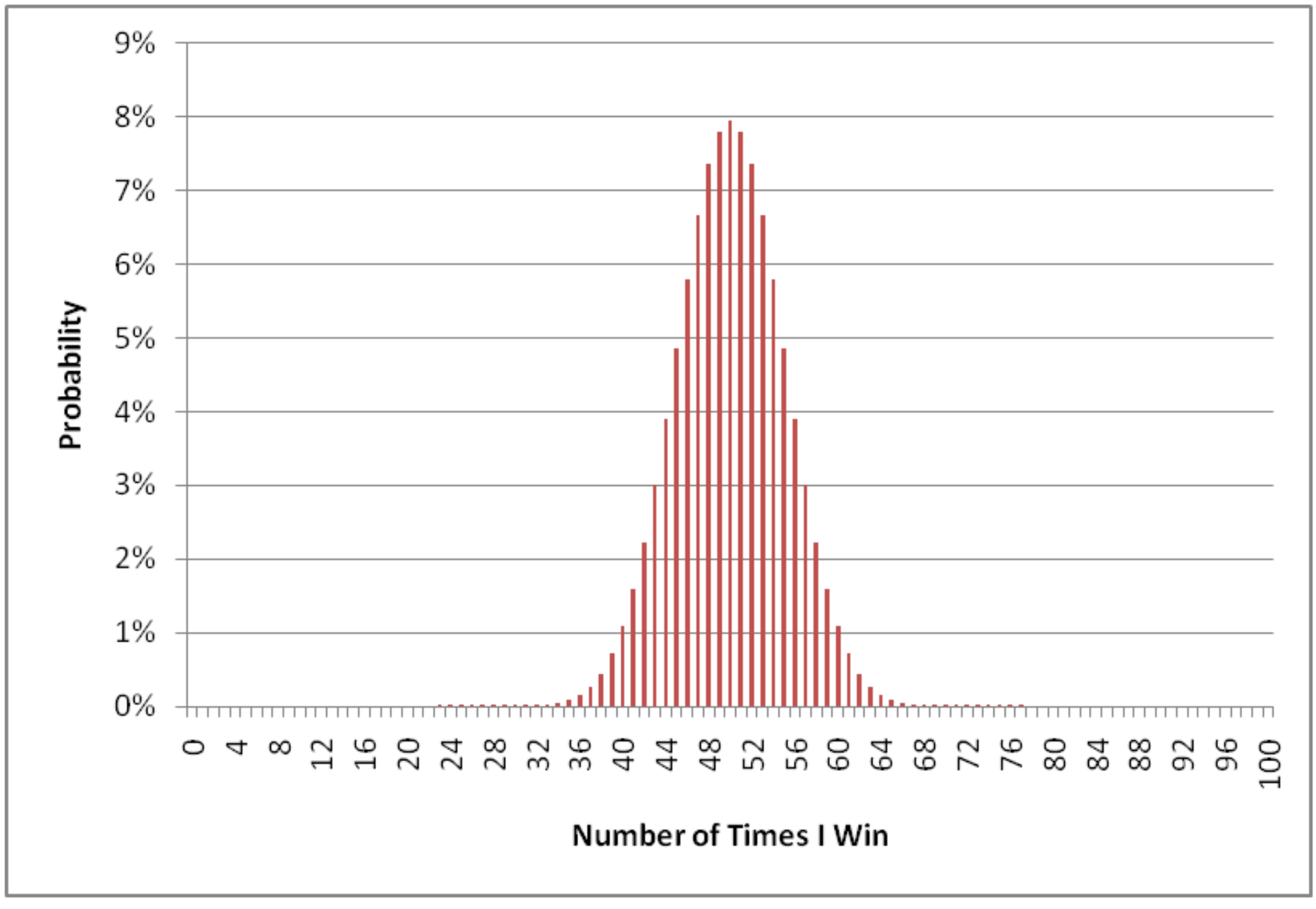
- Probability = # successes / # events
- Sometimes intuitive (coin flip), often not (Monty Hall)
- The methods for determining probability in statistics are explicit

Review: probability

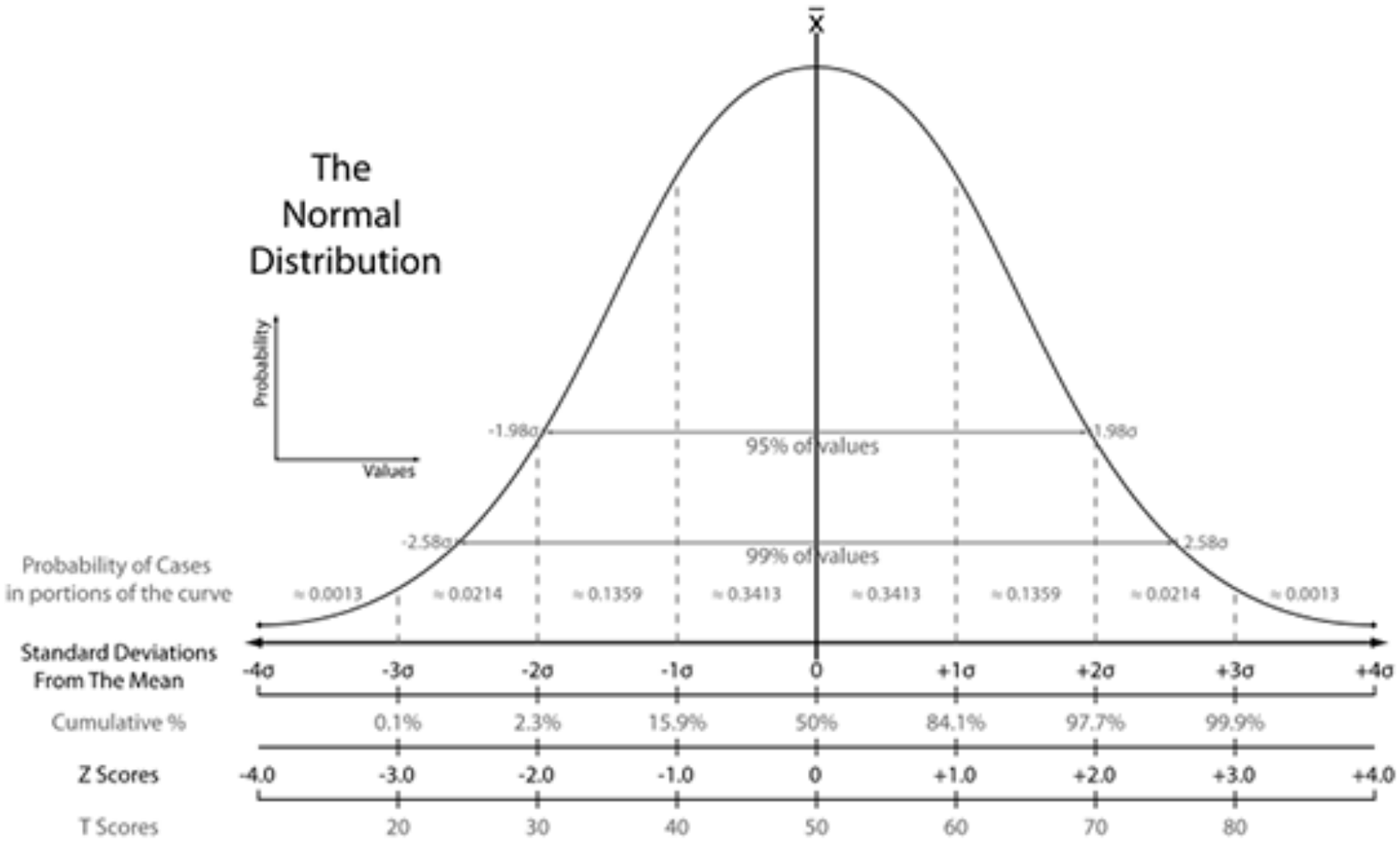
- Probability estimates the likelihood of some *random* occurrence
- In practice, we use probability *distributions*


Length of longest run
4
6
8
10
12
14
16
Number of switches
40
50
60

100 Games of Chutes & Ladders



Normal Curve



Source: http://20bits.com/wp-content/uploads/2008/10/normal-curve-small.png

Normal Curve

- Area = 1
- Mean = Median = Mode (perfectly symmetrical)
- If a variable is distributed normally, Z scores can be used to transform values and construct precise statements about proportions and probability

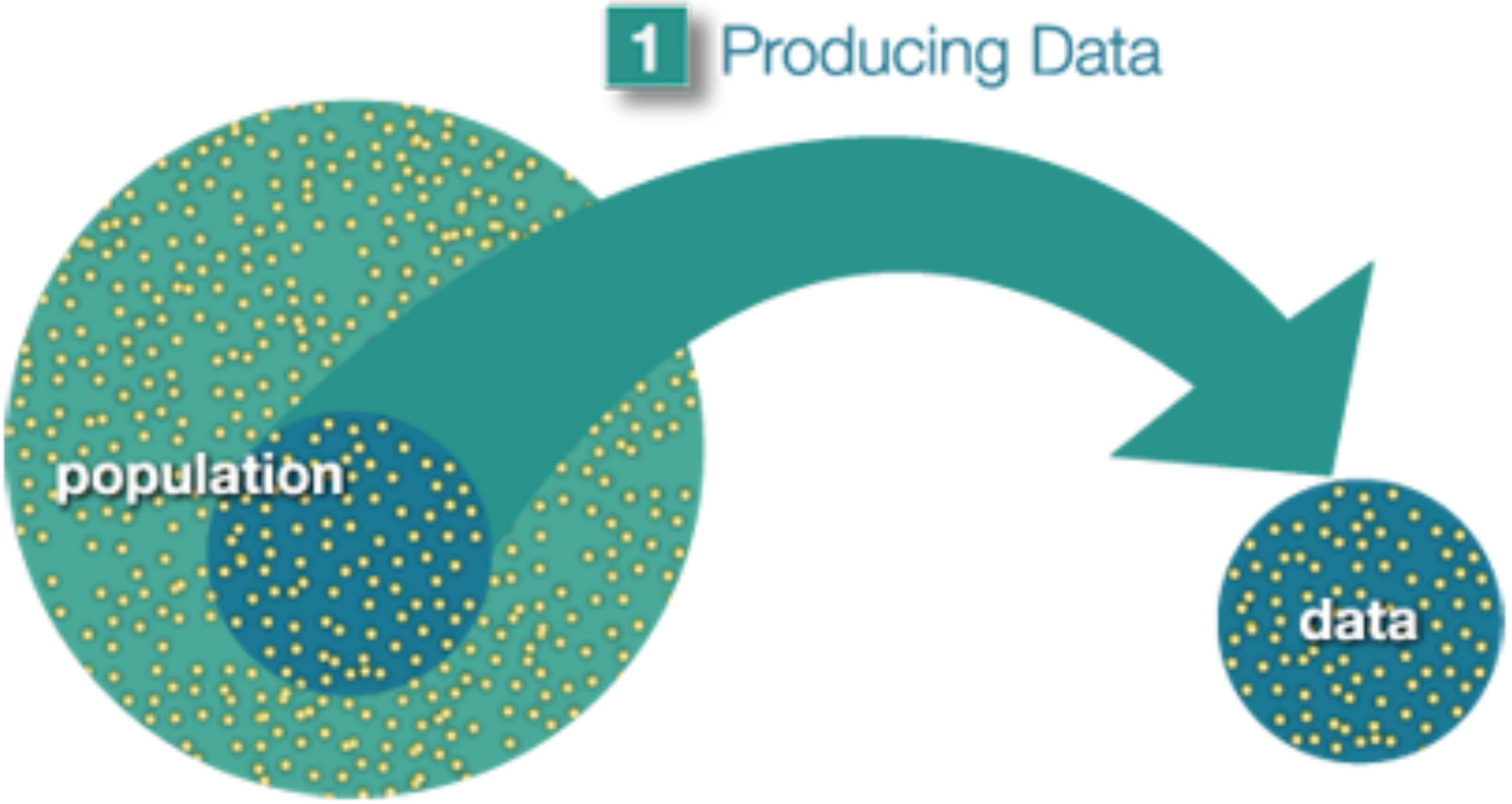
$$Z = \frac{x_i - \bar{x}}{s}$$

Practice

- You have results for 512 students that took a standardized college-entrance exam. The scores are normally distributed, with a mean score of 78 and a standard deviation of 11.
  - Are the data skewed? How do you know?
  - How many students scored above 85?
  - If 60 is the minimum passing grade, what is the pass rate?
  - If another class had test results of a mean of 84 and standard deviation of 19, did this class do better or worse than those in your sample? Why?

Practice

- A variable, age, has a mean of 33 and standard deviation of 15.
  - With this information, could you tell me the percentage of people younger than 18? Why or why not?


1 Producing Data
population
data

Sources of Data

- Population data
  - Data exists for every member of a given population
- Sample data
  - A subset of a population

Data Sources

- Existing Data for Populations
  - MIS databases (management information system)
  - Government records (TANF)
  - Organizational records (schools: students)
  - Other records/data available for entire populations
- Rare!
  - And when it exists, it’s hard to get access

Data Sources

- Existing Data for Samples
  - Public use files (US Census, General Social Survey)
  - Restricted access files

Data Sources

- Collecting new data
  - There are many ways to collect data. But, what do you plan to do with it?

Samples

- Which of these seem representative?
  - A web poll on cnn.com
  - A CBS News phone poll of 350 people about their opinion of the President of the US
  - A mail-in survey of Party priorities from the Democratic Party (sent to Democrats)
  - An email survey to all New School students about the new academic center

Samples

- What makes a sample representative:
  - Equal probability of selection (EPSEM)
  - Equal probability of response

U.S. Census: a Discussion

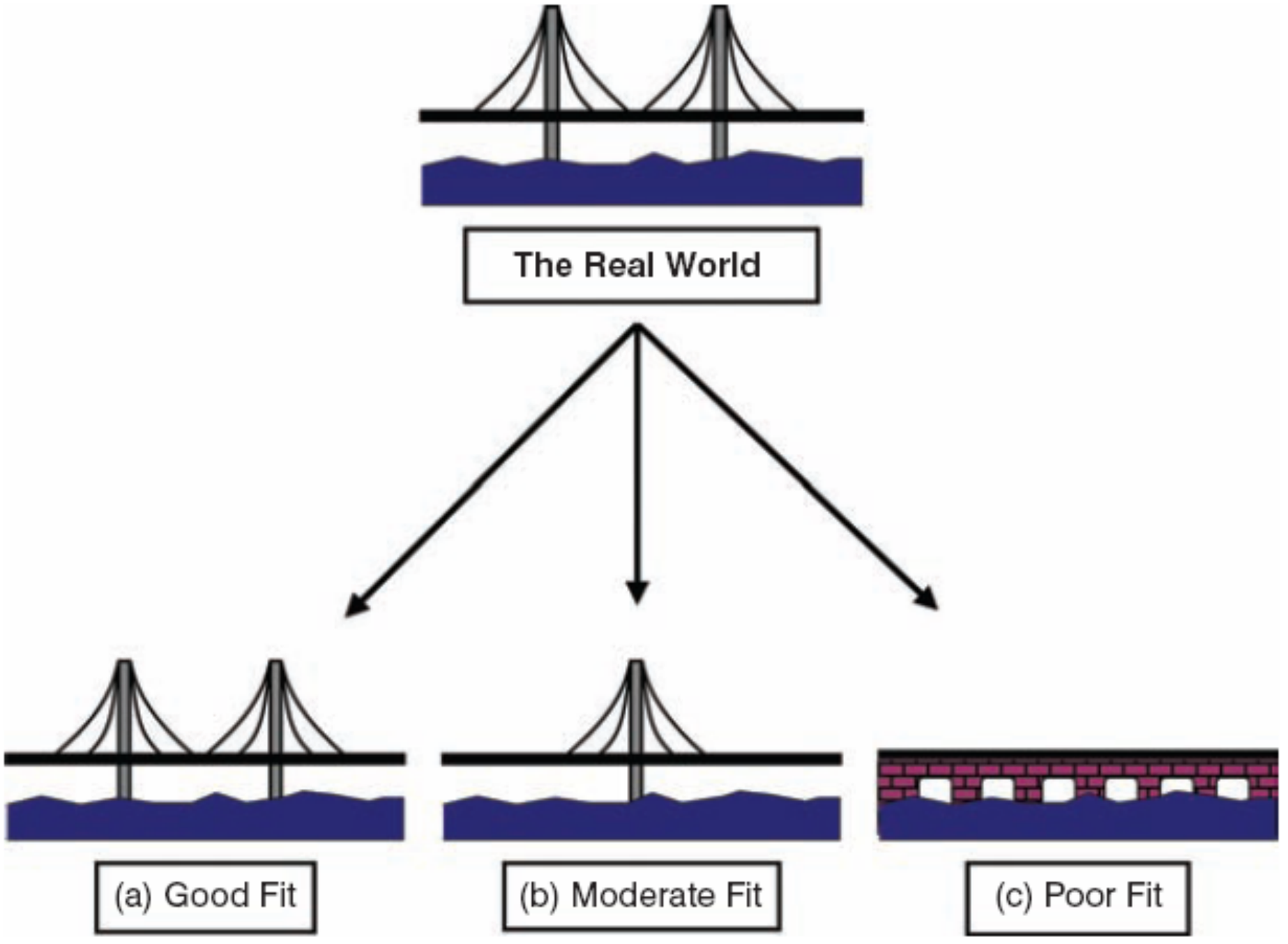
- Very large *sample*
- 72% response rate by mail (more than 225 million responses!)
- Goal: count *everyone* (not realistic)
- Reality: count as many as you can, and take extra steps to make it *representative*

United States Census 2010

IT'S IN OUR HANDS

Problems with Samples

- Sampling design should be decided with the ultimate goal: *that the sample be representative of its population*
- Not always an easy task: the U.S. Census, for example

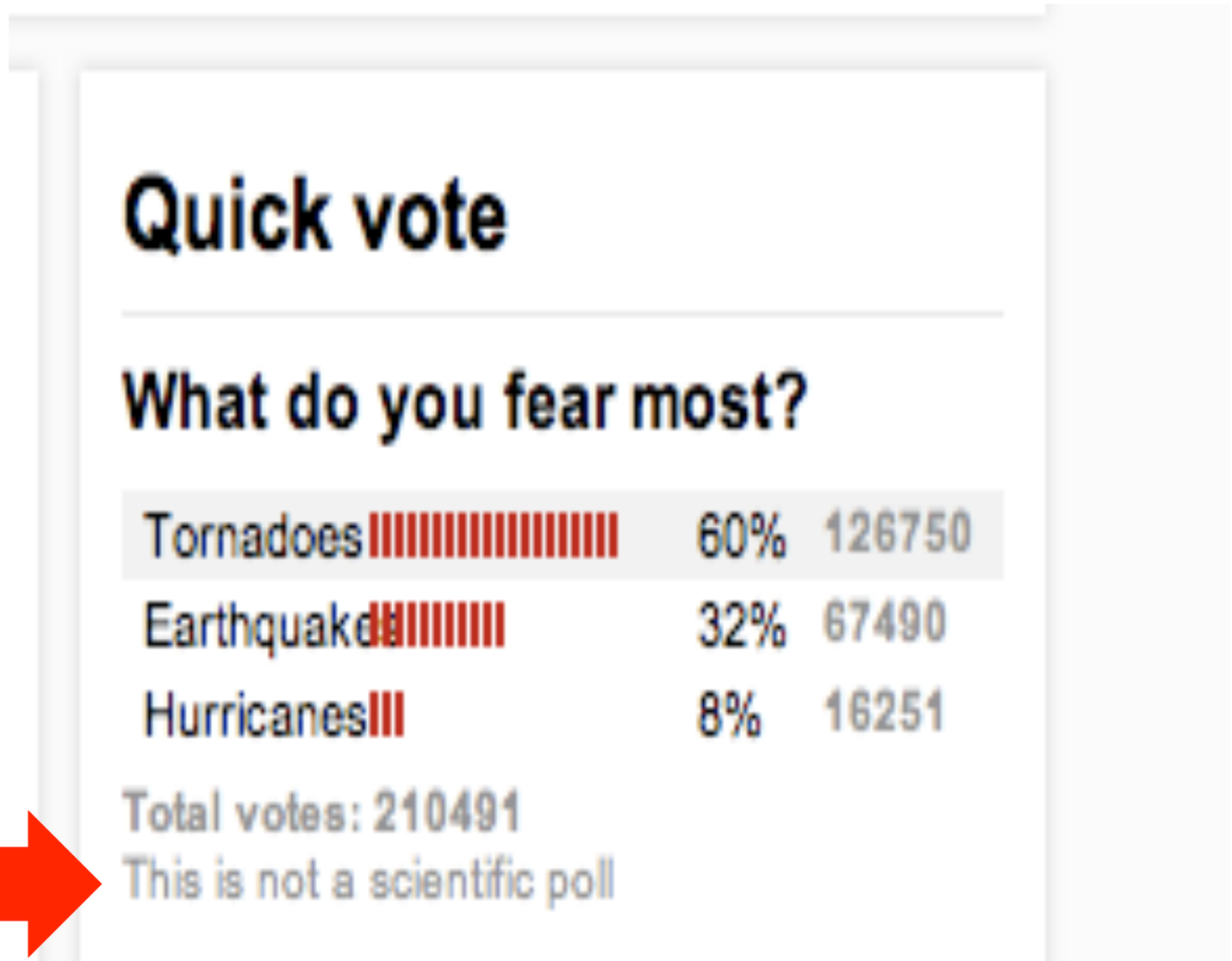

The Real World
(a) Good Fit
(b) Moderate Fit
(c) Poor Fit

Types of Samples

- Probability Sampling
  - Used for statistical inference
- Nonprobability Sampling
  - Convenience sampling
  - Snowball sampling

Nonprobability Sampling

- 8% of Americans fear hurricanes!



Probability Sampling

- A “random sample”

  Everyone in the target population has an equal chance of being selected into the survey

Common Probability Samples

- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

Simple Random Sample

- Randomly select from the entire population
  - Example: draw names out of a hat

Systematic Sampling

- Select sample members from the entire population based on some system
  - Example: select every fifth name in the telephone directory

Cluster Sample

- Randomly select “clusters” and include everyone within those clusters
  - Example: randomly select schools, and include all students within those schools

Stratified Sample

- Randomly sample within each important stratum
  - Example: draw a random sample within each of New York City's five boroughs

Exercise!

The Problem

A farmer has planted corn in a field with 100 plots: 10 plots wide and 10 plots long. On the easternmost side of the field runs a river. The farmer is ready to harvest, but the corn doesn't look so great and she is worried that it won't be worth the money. Consequently, she will harvest only samples, and will use the results to determine whether the entire field should be harvested. Using the sampling methods we discussed, you will construct four estimates of the field yield.

PART ONE

	A	B	C	D	E	F	G	H	I	J
1	6	17	20	38	47	55	69	76	82	97
2	7	14	23	34	43	56	63	75	81	92
3	2	14	28	30	50	50	62	80	85	96
4	9	15	27	34	43	51	65	72	88	91
5	4	15	28	32	44	50	64	76	82	97
6	5	16	27	31	48	59	69	72	86	99
7	5	18	28	34	50	60	62	75	90	90
8	8	15	20	38	40	54	62	77	88	93
9	7	17	29	39	44	53	61	77	80	90
10	7	19	22	33	49	53	67	76	86	97

PART ONE

	Group 1	Group 2	Group 3	Group 4
Sampling Method	**Mean Yield**	**Mean Yield**	**Mean Yield**	**Mean Yield**
Convenience Sample				
Simple Random Sample				
Stratified Sample				
Cluster Sample				
Actual	**50.04**	**50.04**	**50.04**	**50.04**

PART TWO

	A	B	C	D	E	F	G	H	I	J
1	79	81	95	69	65	59	88	65	66	91
2	80	75	88	80	82	66	76	99	62	61
3	97	50	92	92	91	84	75	85	63	89
4	99	71	55	75	65	66	66	86	96	50
5	57	95	51	79	98	71	70	86	89	76
6	57	53	90	71	50	76	56	91	85	64
7	69	95	98	90	93	97	79	95	73	90
8	58	99	75	51	67	81	55	63	89	74
9	98	62	73	54	50	76	91	50	90	55
10	91	59	69	59	71	72	85	85	86	97

PART TWO

	Group 1	Group 2	Group 3	Group 4
Sampling Method	**Mean Yield**	**Mean Yield**	**Mean Yield**	**Mean Yield**
Convenience Sample				
Simple Random Sample				
Stratified Sample				
Cluster Sample				
Actual	**76.03**	**76.03**	**76.03**	**76.03**

Probability Samples

- To be able to infer meaning from a sample it should be ***representative***
- A sample is representative when everyone in the *population* has an equal chance at being selected into the *sample*

Other Sampling Considerations

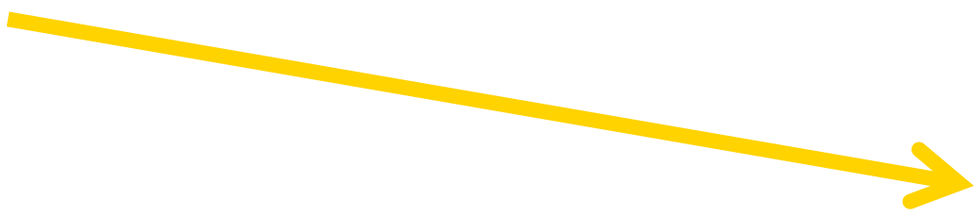
- Under-represented groups
  - Over sampling
- Survey samples have additional sources of error
  - Sampling frame
  - Measurement

Response Bias

- Even if you have a perfectly random sample… if those who respond differ from those who didn't, your sample will be subject to response bias.

Which sampling technique should I use?

- When there are clear patterns or groups in the data



- Convenience sample
- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

Which sampling technique should I use?

- When there is no discernable pattern in the population

- Convenience sample
- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

Which sampling technique should I use?

- When it would be logistically taxing to sample from an entire population, but they gather in large groups

- Convenience sample
- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

Which sampling technique should I use?

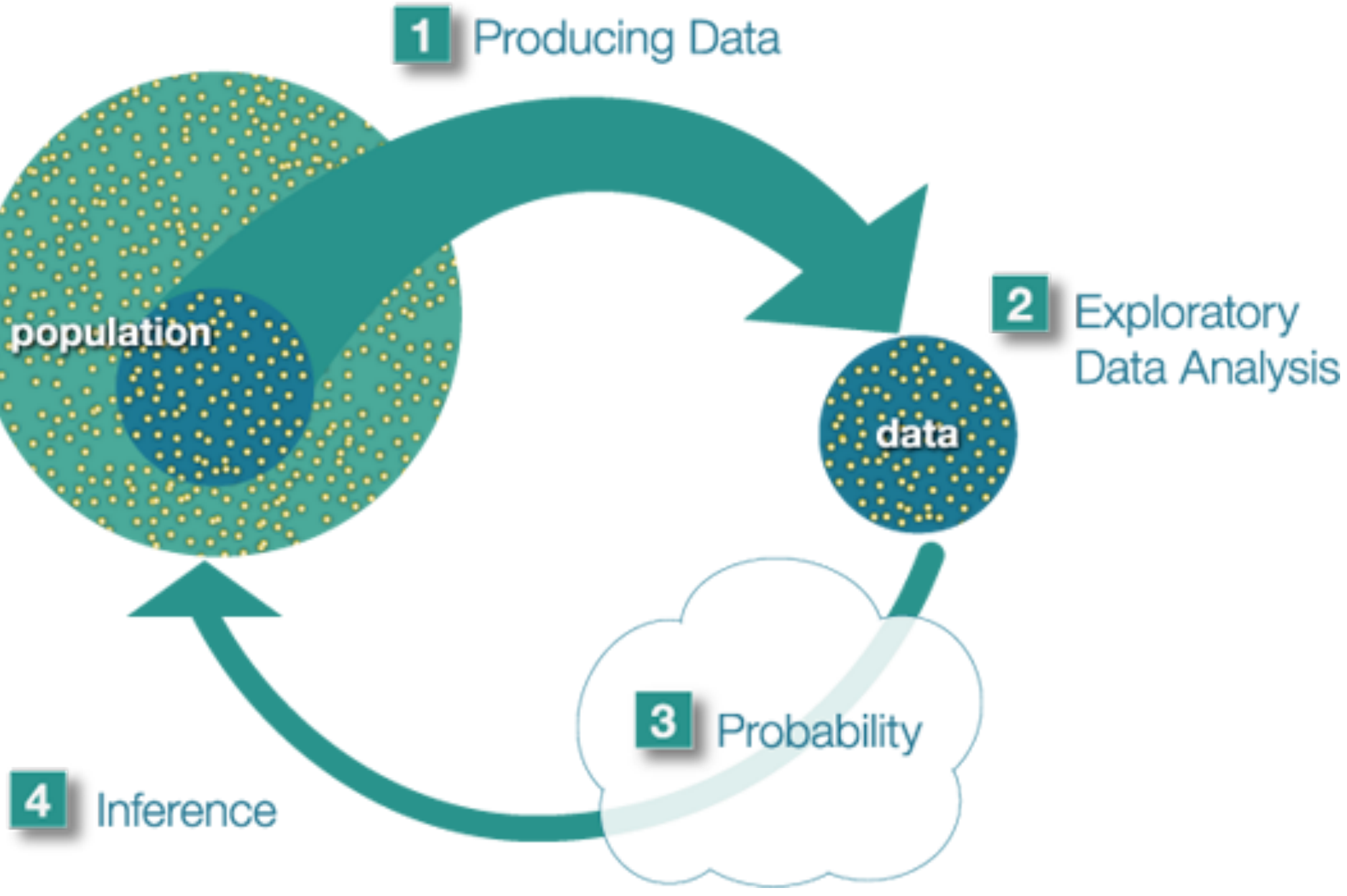
- When I don't want anyone to believe the results

- Convenience sample
- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

Which sampling technique should I use?

- When I can't conduct a simple random sample but could systematically construct a near-random sample

- Convenience sample
- Simple random sample
- Systematic sampling
- Stratified random sample
- Cluster sampling

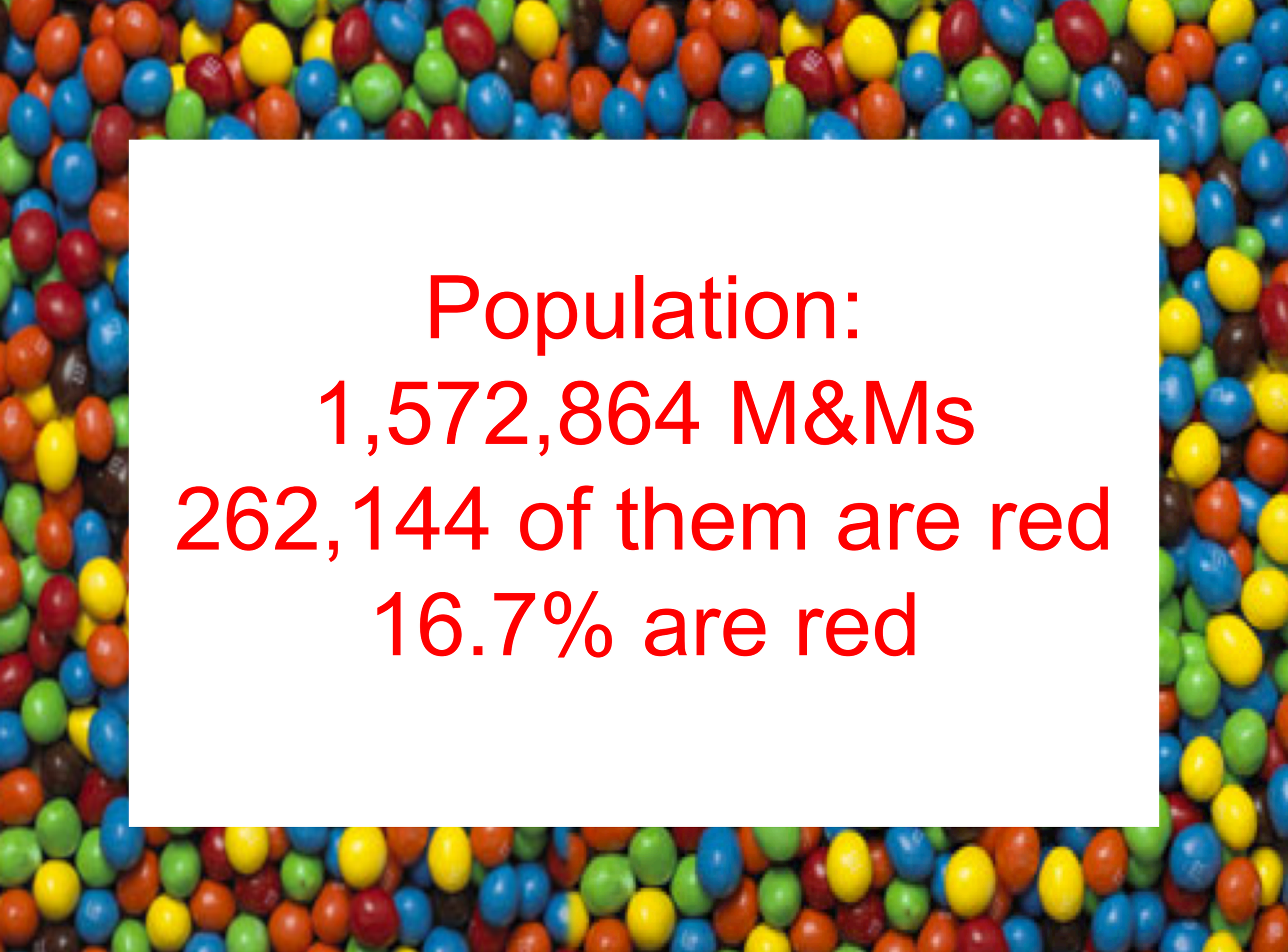

1 Producing Data
population
data
2 Exploratory Data Analysis
3 Probability
4 Inference

Inference = Generalization

- To generalize your sample to its population it should follow EPSEM standards.
- The link from the sample to the population is the **Sampling Distribution**

Sampling Distribution

- Links the sample with the population
- Its shape is normal (if the sample is EPSEM)
- If you took thousands of samples drawn from a single population, the distribution of the sample means would be normal
- The mean of the sample means would be equal to the mean of the population



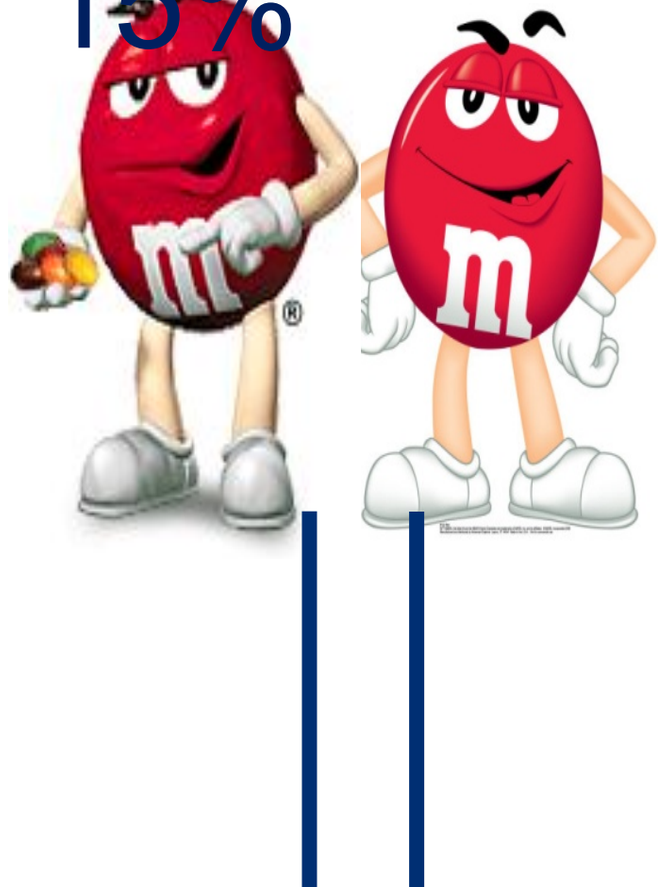
Population:
1,572,864 M&Ms
262,144 of them are red
16.7% are red

15%

17%

4%

32%



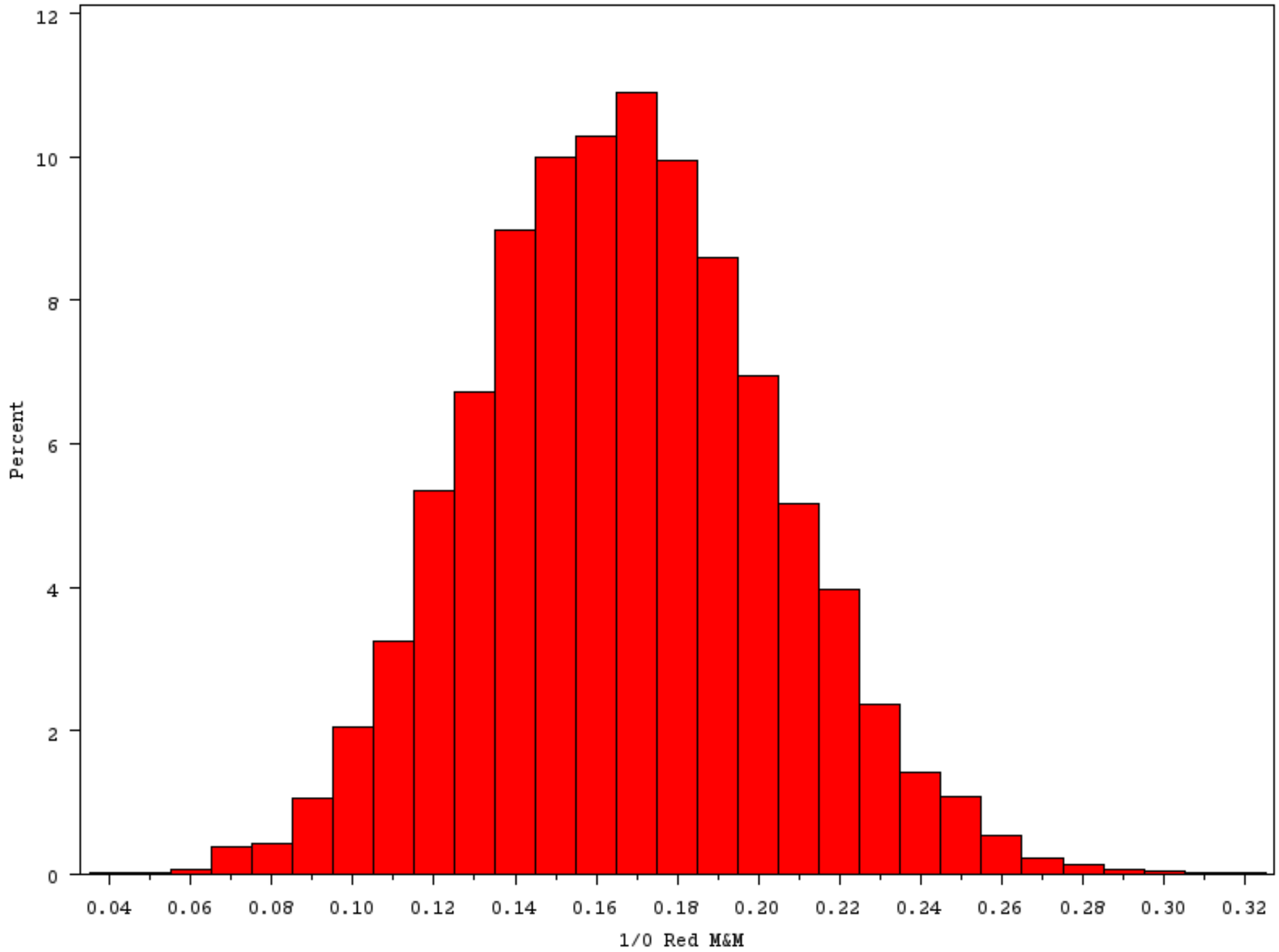
0%

+0%

16.7%

12%

Variable: red M&Ms (%)



Distribution of red M&M estimates from 10,000 random samples*

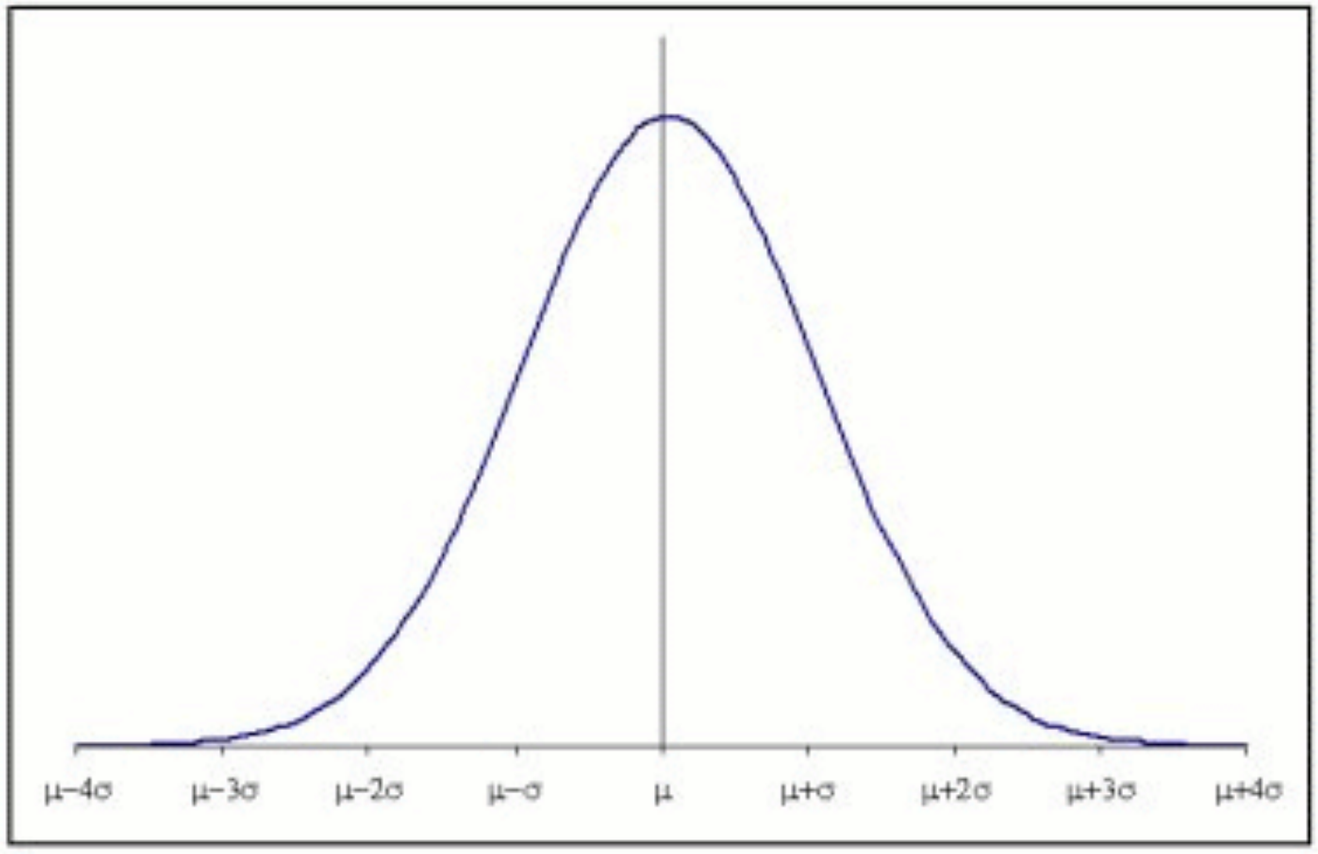
*N=100

Central Limit Theorem

- If a sample is EPSEM and N is 100 or greater, the sampling distribution takes a normal shape.
- Thus, the mean of any given sample can be related to all the other (theoretical) sample means--as well as the population mean--through the properties of the normal curve.

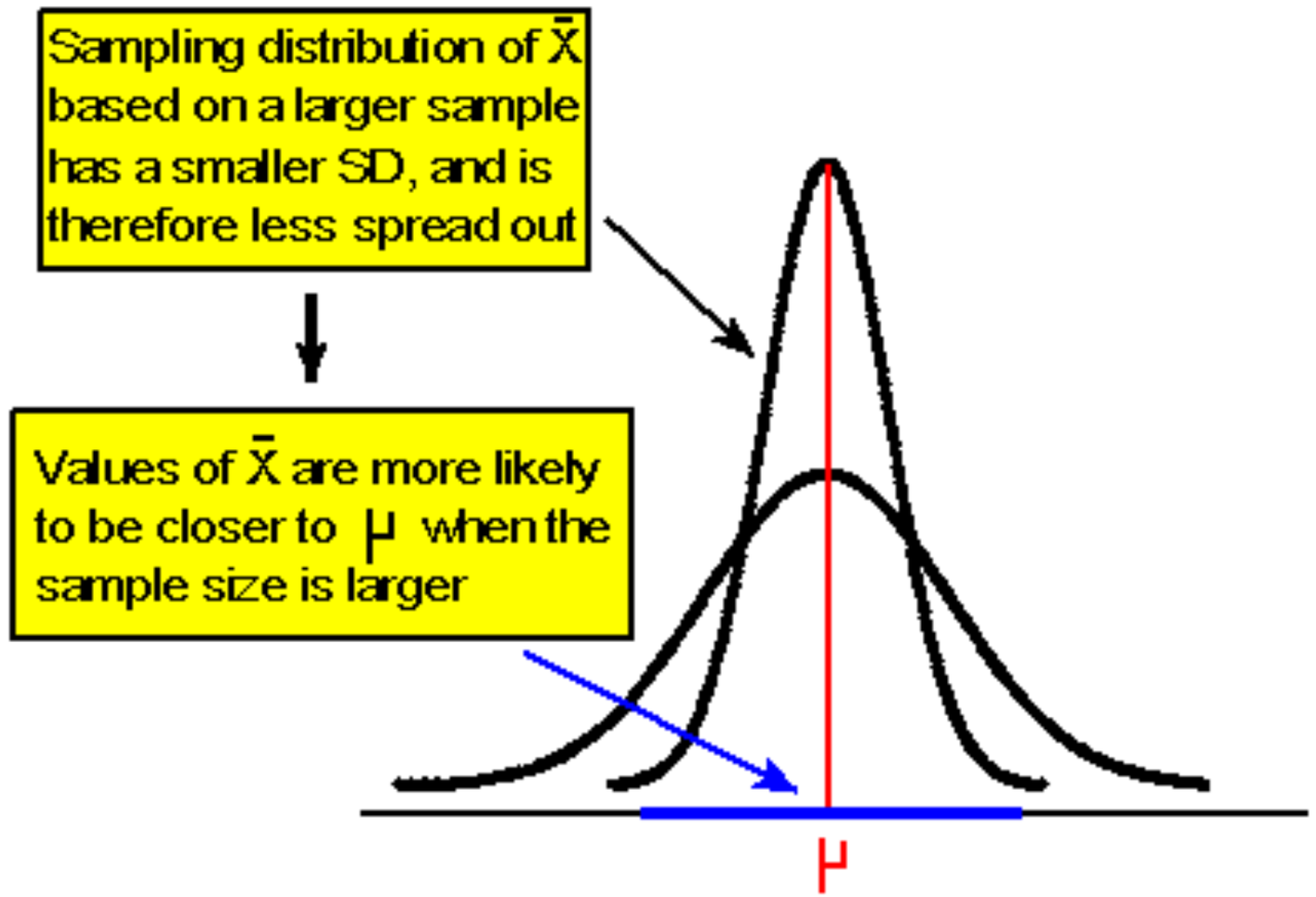
Central Limit Theorem

Fig 1: The Normal distribution function



μ is the mean and σ the standard deviation of the distribution

Central Limit Theorem



Standard Error

- Is (theoretically) the standard deviation *between* sample means
- Is approximated by:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Important Symbols

	Mean	SD	Prop.
Samples	$\bar{X}$	s	P_s
Populations	μ	σ	P_u

Scenario #1

- Post a music-lovers' survey on a university bulletin board, asking students to vote for their favorite type of music
- Type: Volunteer sample. Individuals self-select
- Very biased. Why?
- Cannot be used to represent the New School student population.

Scenario #2

- Stand outside Union Square, across from the 16th Street building, and ask students passing by to respond to your question about musical preference
- Type: Convenience sample - individuals are selected based on their proximity to the researcher at the time she has decided to collect data
- Less biased than volunteer sample, but still very biased. Why?
- Not very credible for representing a population.

Scenario #3

- Ask your professors for email rosters of all the students in your classes. Randomly sample some addresses, and email those students with your question about musical preference.
- Type: Convenience sample
- The potential for bias is lessening, but still very likely to be biased. Why?
- Not very credible for representing a population.

Scenario #4

- Obtain a student directory with email addresses of all the university's students, and send the music poll to every 50th name on the list.
- Type: Systematic sampling
- Lessens the chance for bias, but not fail-proof. Why?
- Can be used to represent a population.

Scenario #5

- Obtain a student directory with email addresses of all the university's students, and send your music poll to a simple random sample of students.
- Type: Simple random sample.
- Students are equally likely to be selected. This removes the potential for selection bias. Random sampling is the strongest method for selecting a sample

Sources of bias

- Bias is a serious problem that is impossible to eliminate, but as good researchers we work to remove as much as possible. Potential biases can be troubling for your research:
  - Only music lovers respond
  - You pick out only the people that you find attractive
  - You interview only people in your degree program
  - You send it out the night that Project Runway is on and Parsons students disproportionately respond at much lower levels than other departments

ESTIMATION PROCEDURES

How much risk of error is acceptable?

- For election polls?
- For opinion surveys?
- For new medicines being developed by pharmaceutical companies?

Chance for Error?

- Yes, but the risk is managed
- The amount of acceptable risk of error is set by the researcher
- This is called “Alpha”
  - Alpha: the probability of error (Type I)
  - Common alphas: .10, .05, .01, .001

Confidence Level

- A counterpart to Alpha
- Another way of expressing probability of error
- If Alpha = .05, then Confidence Level = 95%
  - (Probability of error = .05, so probability of non-error = .95; .95 * 100 = 95%
- If Alpha = .01 then Confidence Level = ?

Confidence Interval

- NOT to be confused with Confidence Level
- Point Estimates are precise, but misleading
- Confidence Interval is the range of the expected population value (which includes the point estimate)