BIOS6606 LectureAug28

Probability, sampling, and bias

Kathleen Torkko

Objectives

- Understand a role probability plays in statistics
- Learn some statistical terminology
- Understand potential differences in the use of statistics in human populations and in basic science studies
- Understand difference between descriptive and inferential statistics
- Learn how sampling can influence conclusions and introduce bias
- Understand what bias is and why it is important to control

Probability Basics

Probability theory uses mathematics to measure the likelihood of something happening.

Why do we care about probability theory?

Makes it possible for scientists to use statistics to quantify the extent of uncertainty inherent in their conclusions from experiments.

Probability Basics

Probability deals with calculating the likelihood of a given event occurring

i.e., What is the probability that I will get tails from a single coin toss

Measured by the ratio of # of tails possible to the total number of possible outcomes*

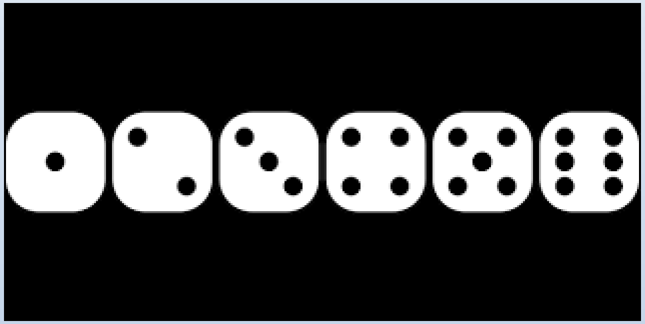
Expressed as a number between 0 and 1, a percent, or 1 in 2

*Total number of possible outcomes, AKA sample space

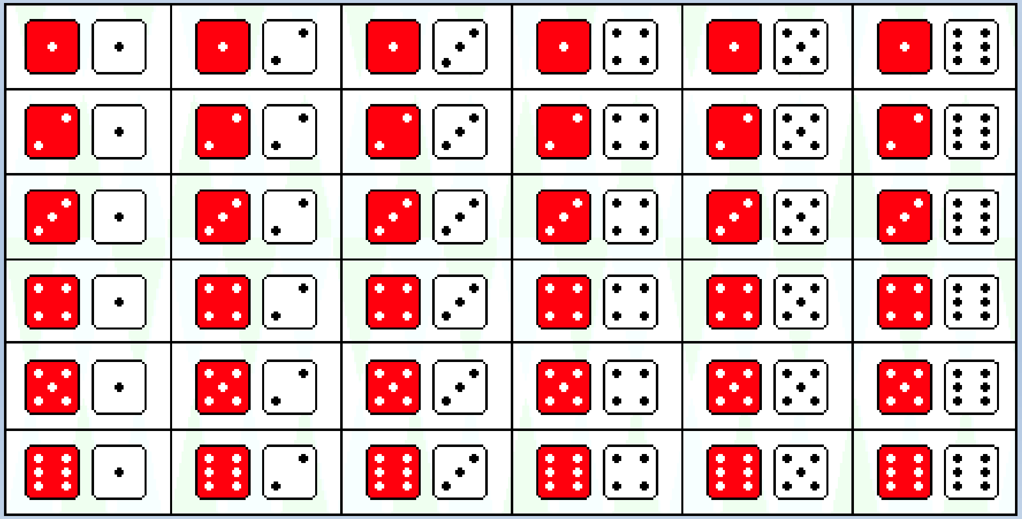
Probability Basics: Sample Space

The sample (or event) space contains all possible outcomes (events) for a given experiment

This is the sample space of rolling 1 dice.



This is the sample space of rolling 2 independent dice.

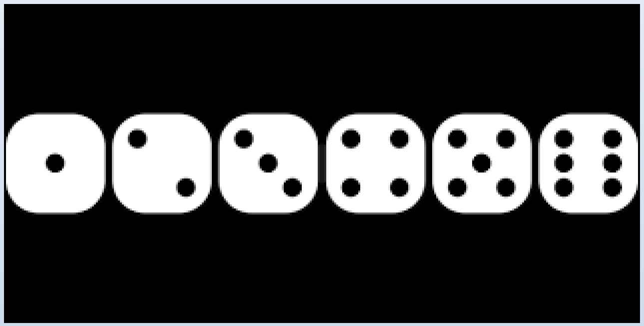


When two dice are rolled, there are 36 different and unique ways the dice can come up. This figure is arrived at by multiplying the number of ways the first die can come up (six) by the number of ways the second die can come up (six)

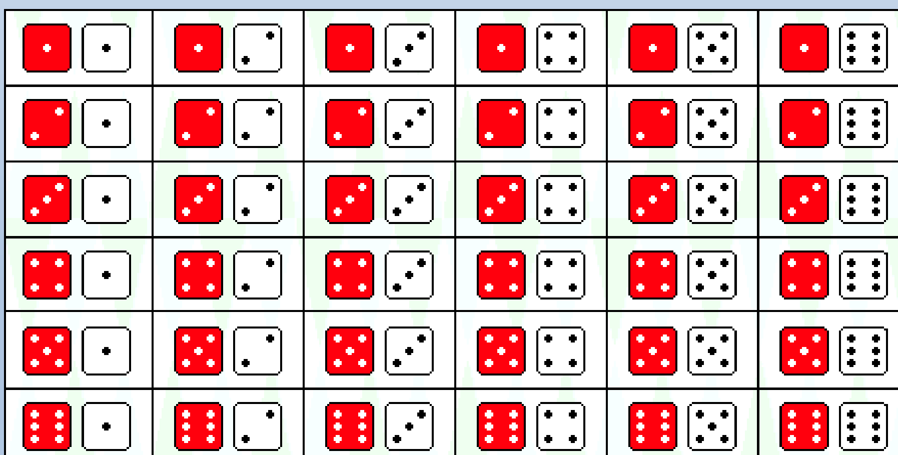
$$6 \times 6 = 36$$

Probability Basics: Probability

What is the probability of rolling a 5 from tossing 1 dice?



= 1/6 = 0.167 = 16.7% = 1 of 6



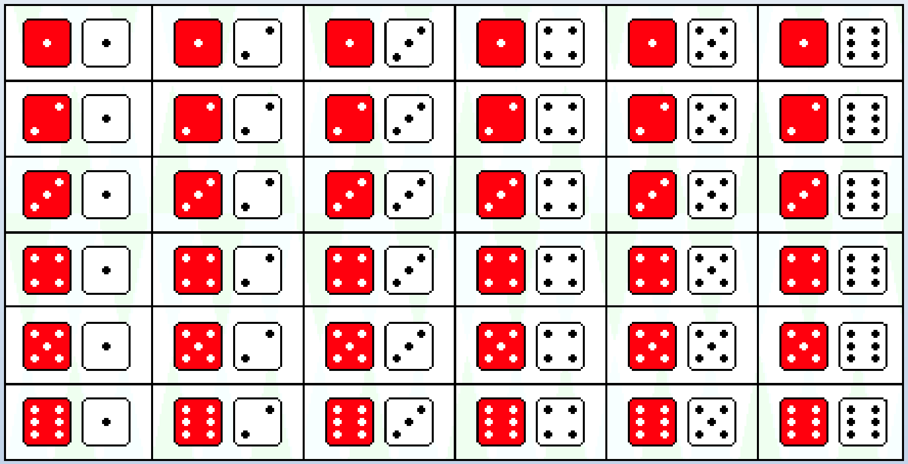
In probability speak, an outcome (event) usually designated by a capital letter

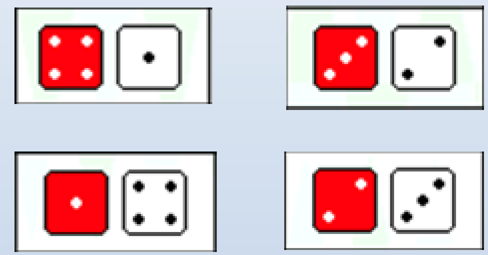
If A=rolling a 5 with 2 dice

The probability of A is written P(A)

What is P(A)?

What is P(rolling a 5 with 2 dice)?





P(A)?

P(rolling a 5 with 2 dice) = 4/36 = 1/9 = 0.111 = 11.1%

If B = not throwing a sum of five, what is P(B)?

$$P(B) = 32/36 = 8/9 = 0.889 \text{ or } 88.9\%$$

The sum of probabilities for all possible outcomes is equal to one (or 100%).

Because A and B contain all possible outcomes and are mutually exclusive,

$$P(B) = 1 – P(A) = 1 - 0.111 = 0.889$$

If an experiment can have three possible outcomes (A, B, and C), then

$$P(A) + P(B) + P(C) = 1$$

You can determine the sample space either empirically or theoretically.

Empirically: based on observation or experiment rather than theory

probability of a specific outcome = number of times the specific outcome occurs divided by the total number of outcomes observed

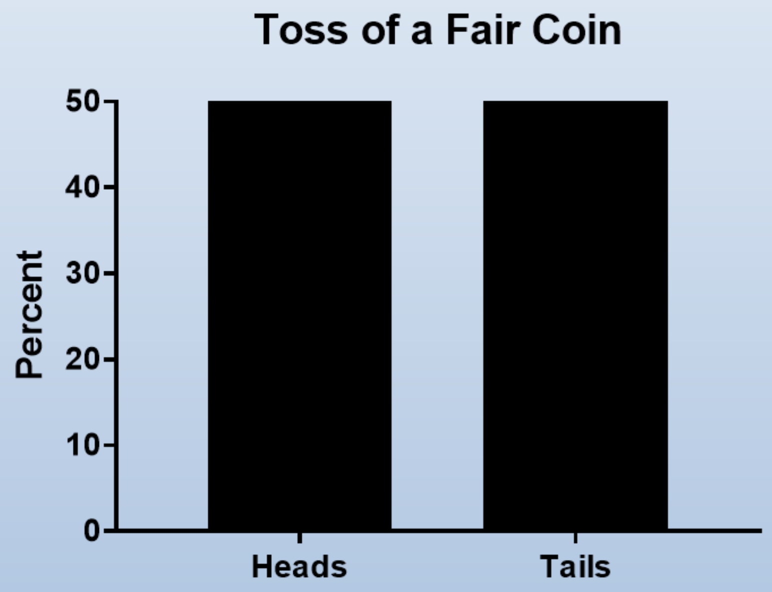
Experiment from assignment: toss 2 coins 10 times to determine the distribution of possible outcomes

Theoretically: based on theory rather than observation or experiment

Theoretical probability is given by the number of ways the specific outcome can occur divided by the total number of possible outcomes.

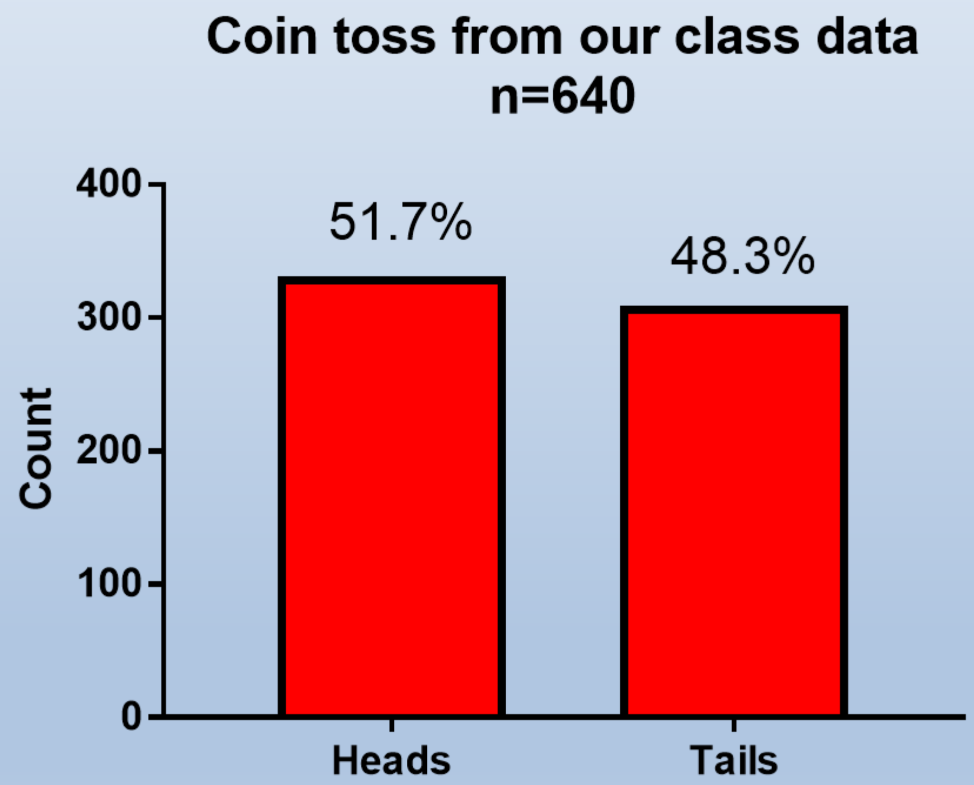
Let’s look at our random coin and dice rolling experiment

This is what we would expect theoretically



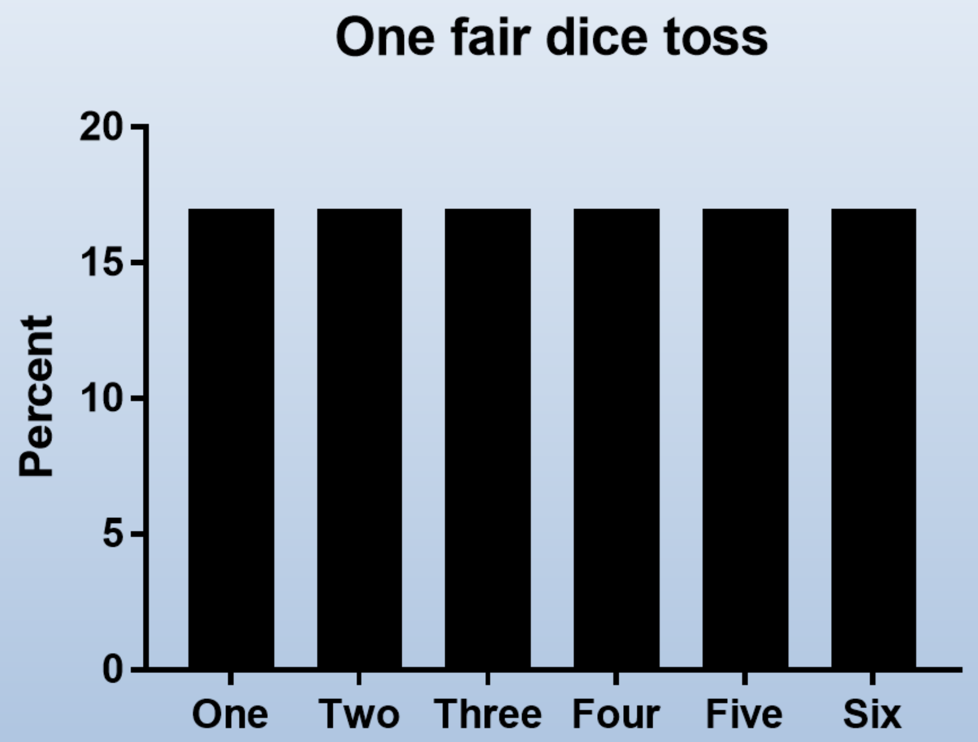
This is what we got empirically: the sample from our class experiment

2018



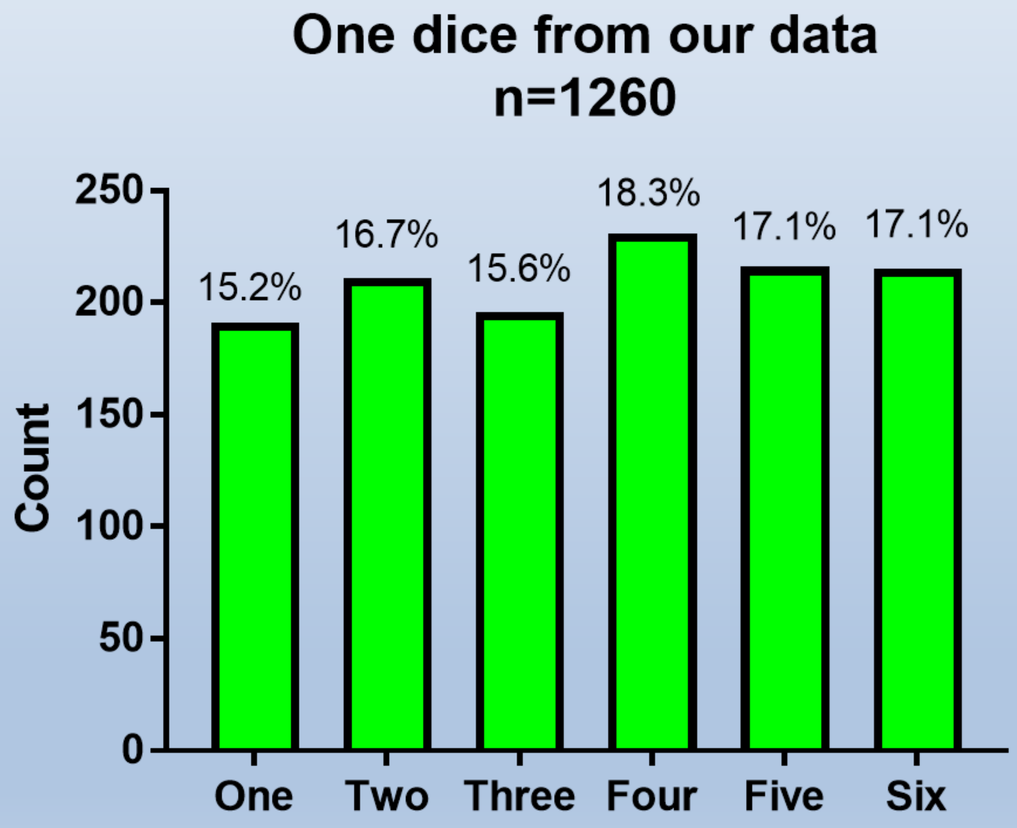
What distribution would you expect from our 2 coin toss experiment?

This is what we would expect theoretically

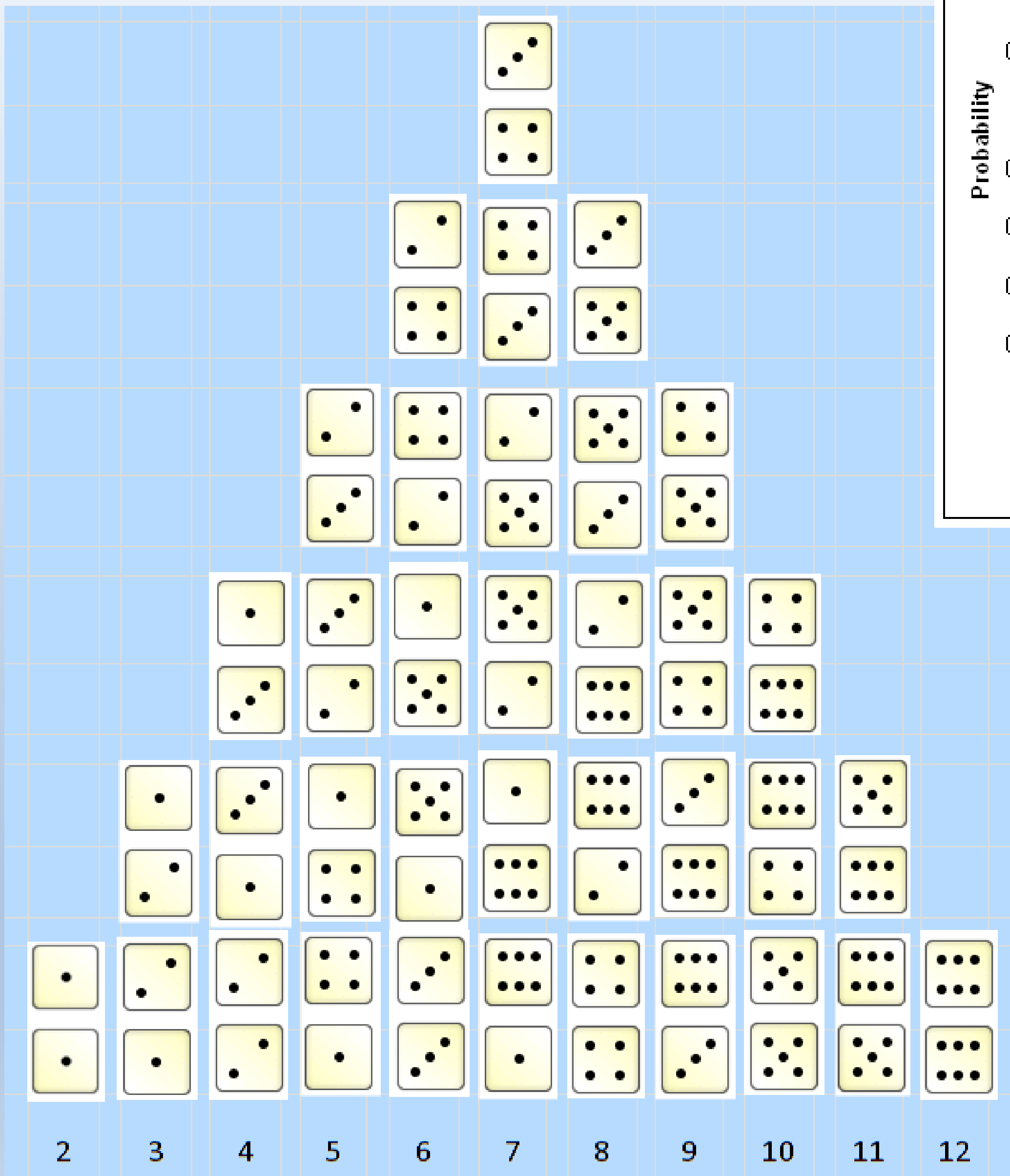


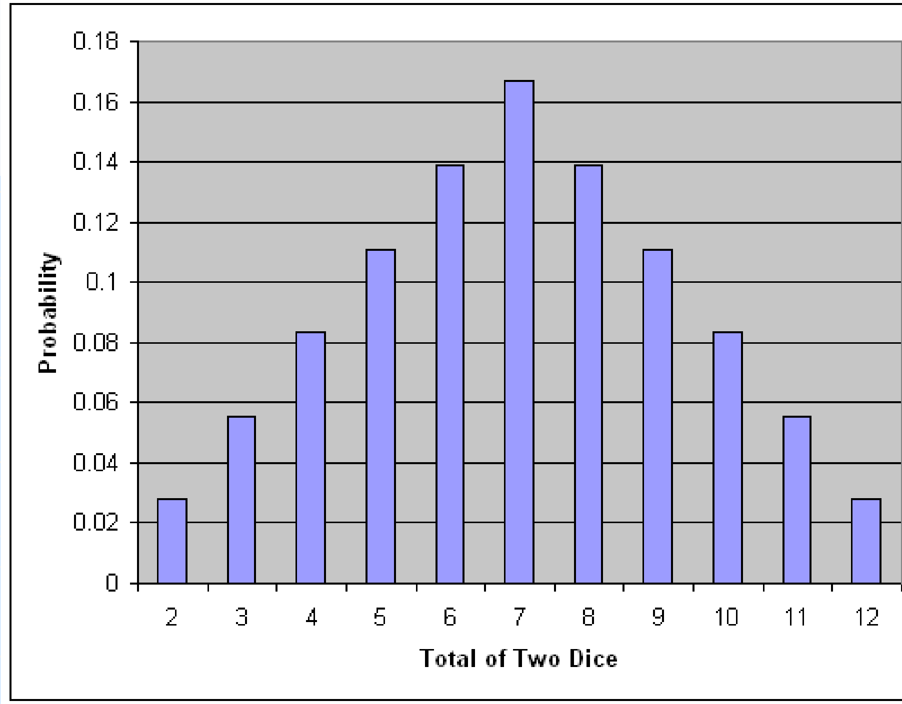
This is what we got empirically: the sample from our experiment

2018



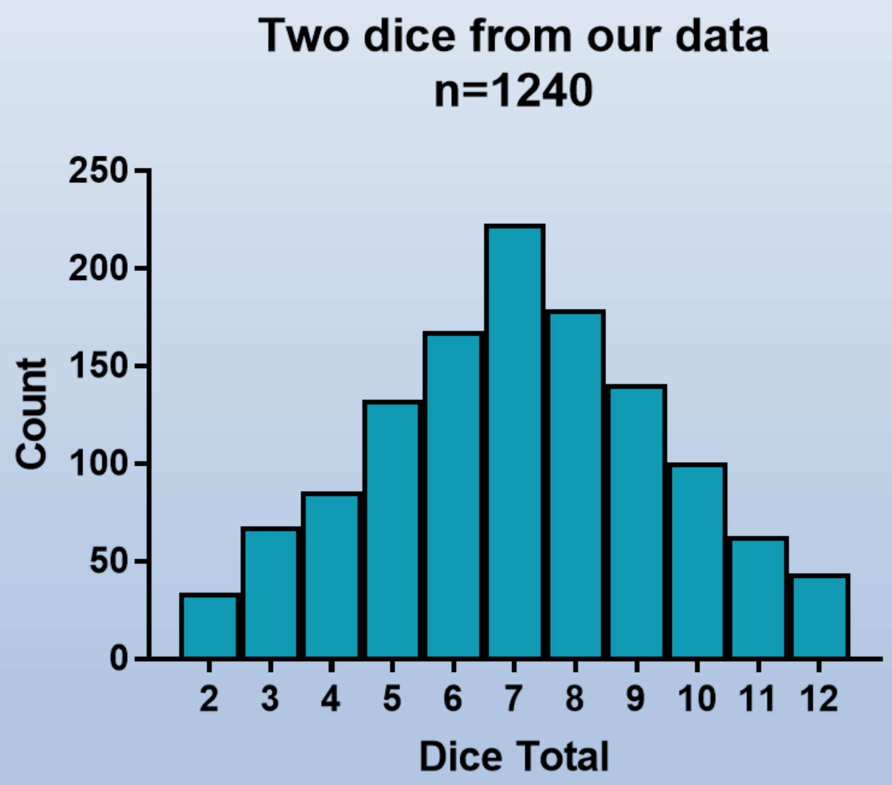
Theoretical distribution of outcomes for 2 dice





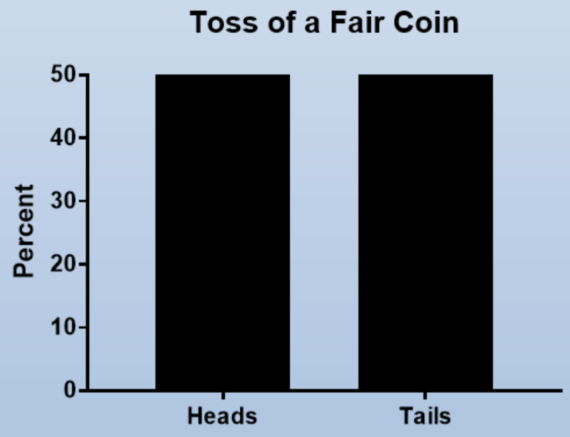
This is what we got empirically: the sample from our class experiment

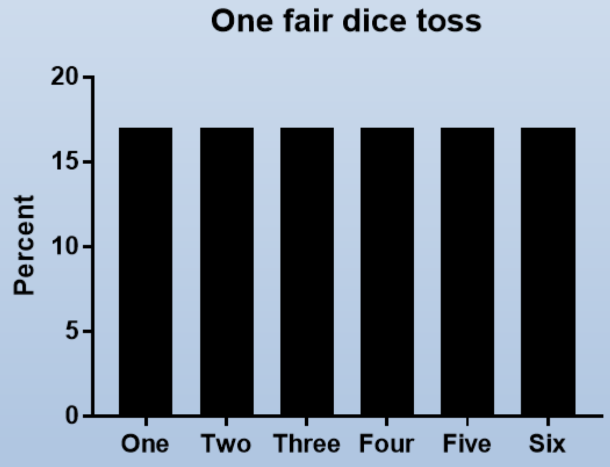
2018

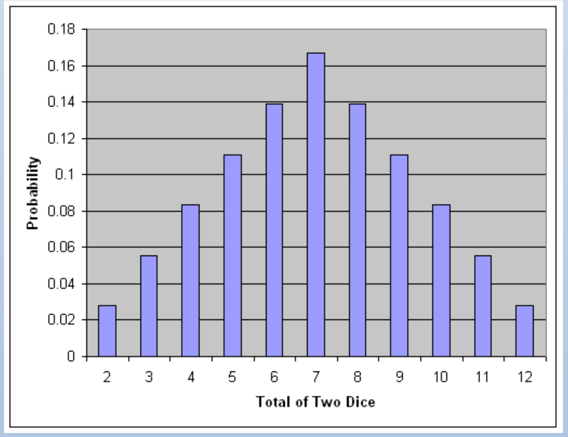


What is a Probability Distribution?

A probability distribution is a table or a frequency distribution or an equation that links each possible value that an outcome can assume with its probability of occurrence.

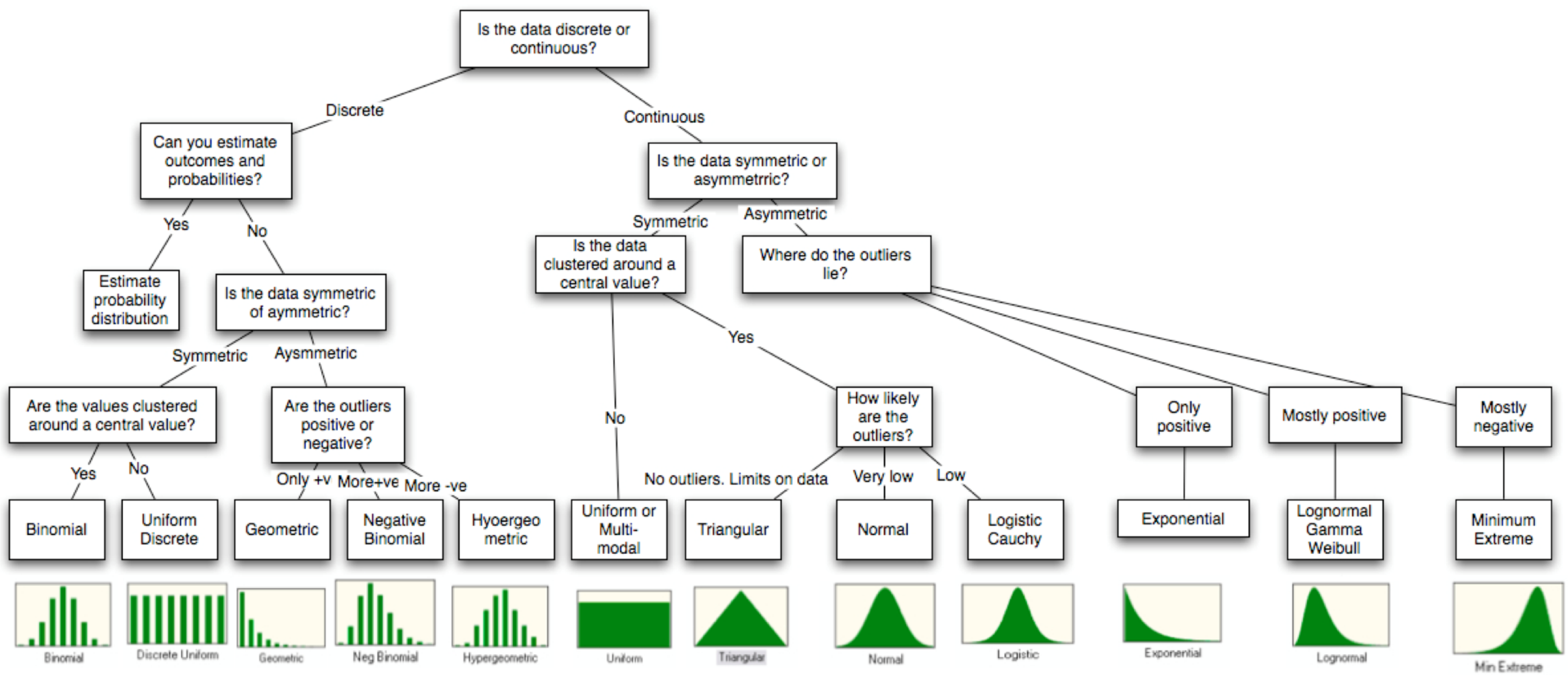




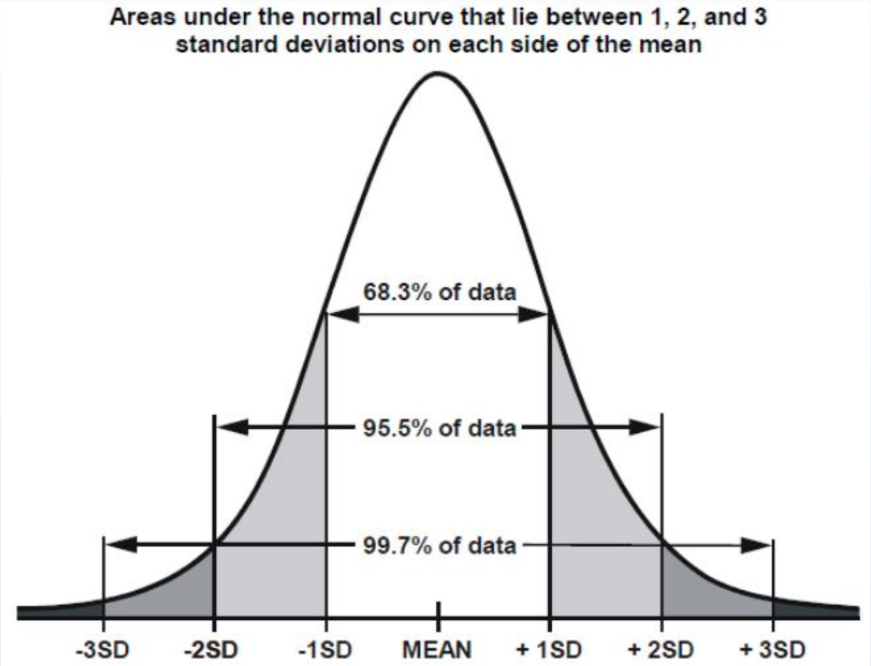


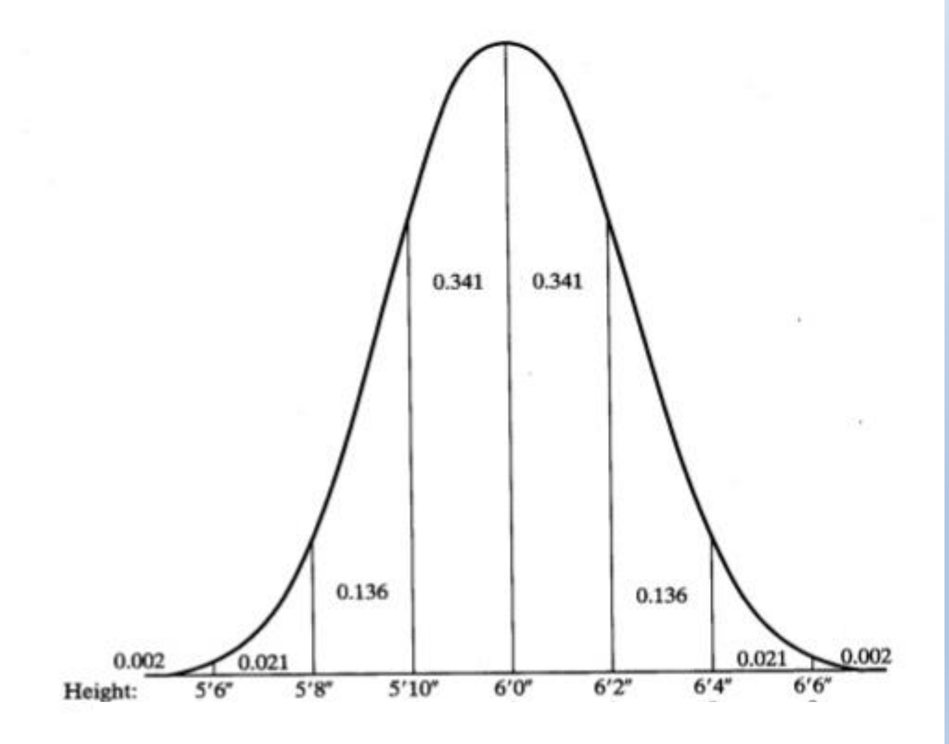
In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

Figure 6A.15: Distributional Choices



http://people.stern.nyu.edu/adamodar/pdfiles/papers/probabilistic.pdf


Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean
68.3% of data
95.5% of data
99.7% of data
-3SD
-2SD
-1SD
MEAN
+ 1SD
+ 2SD
+ 3SD


0.341
0.341
0.136
0.136
0.002
0.021
0.021
0.002
Height:
5'6"
5'8"
5'10"
6'0"
6'2"
6'4"
6'6"

my.ilstu.edu

Statistical hypothesis testing uses different types of probability distributions to determine whether the results are statistically significant ($p \leq 0.05$).

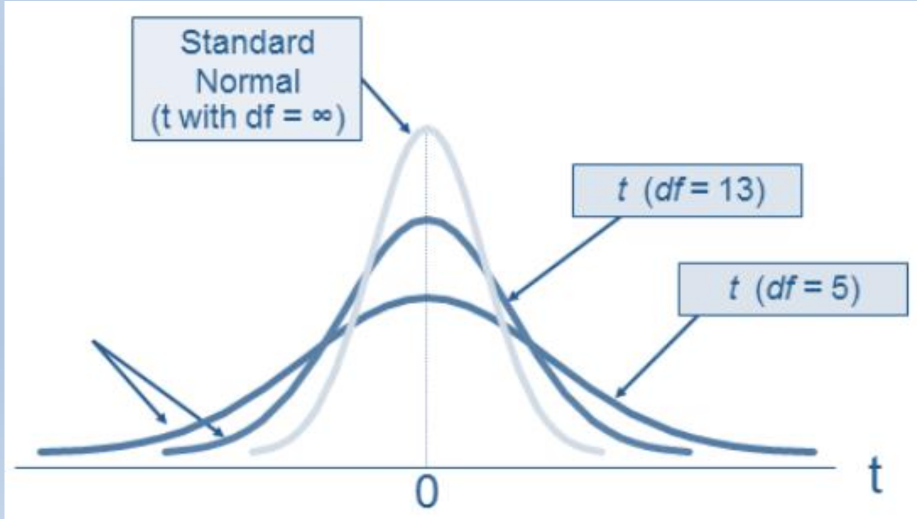
Each type of statistical test calculates a test statistic

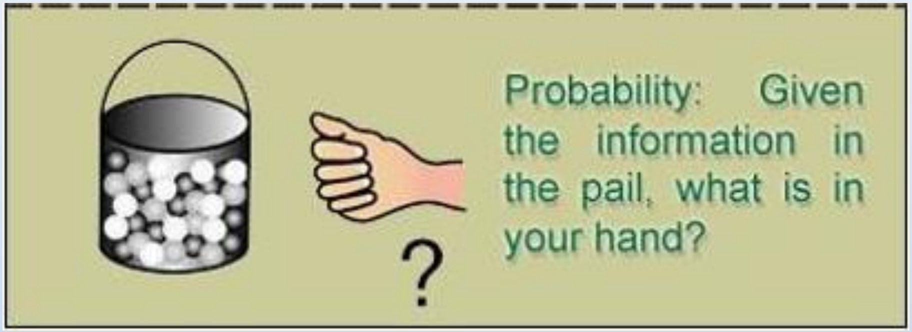
t-tests equations calculate t statistics

ANOVA equations calculate F statistics

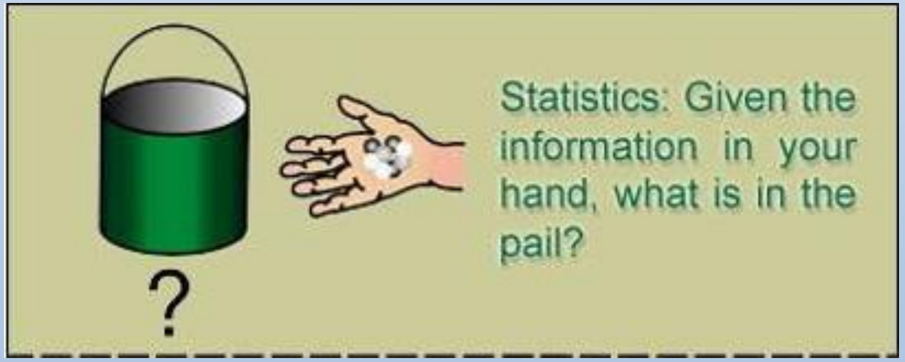
Chi-square test equations calculate chi-square statistics

Statistical tests use the probability distributions of these test statistics to calculate p-values (*probability*-values) for each test



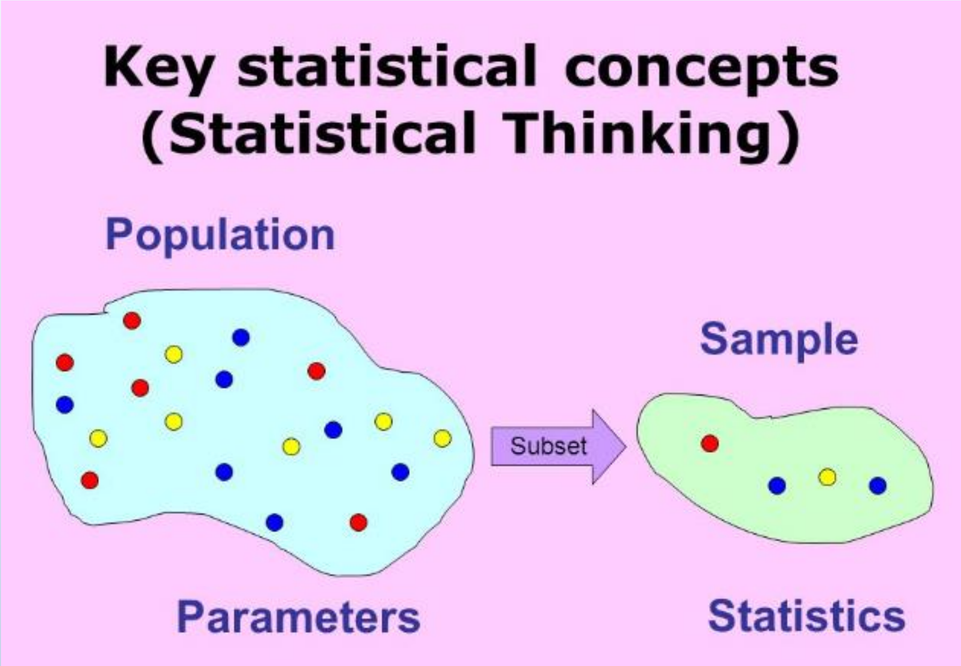


You know the sample space (contents of the pail). We know the frequency distributions



But we often don’t know the sample space or the frequency distribution, so you guess it from a sample you took from the pail

So we sample



The entire collection of people, animals, cell lines or things from which we collect data

The entire group we are interested in, which we wish to describe or draw conclusions about (e.g., people with breast cancer)

A set of subjects/samples drawn from the population

Preferably a random sample (hopefully representative of the population)

The Language of Statistics

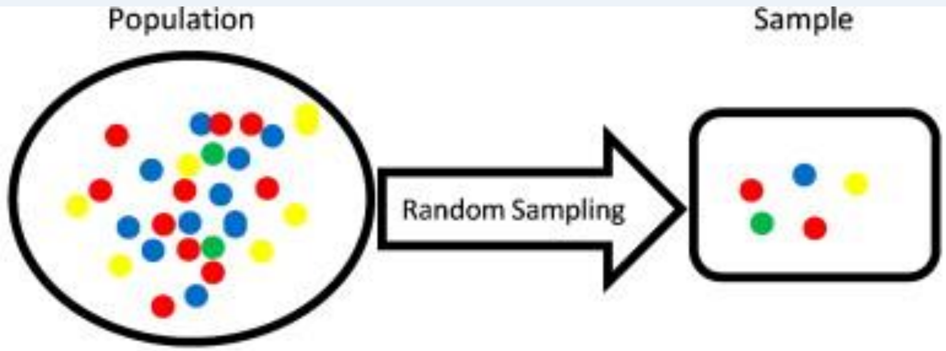
<u>Parameter</u>

- A value used to represent a *population* characteristic
- Usually unknown, and therefore has to be estimated (using sample data)
- Usually represented by a Greek letter
  Greek letter mu (μ) = the population mean
  Greek letter sigma (σ) = the population SD

<u>Statistic</u>

- A quantity that is calculated from a *sample* of data
- Used to estimate the unknown value of the parent population parameter
- Usually represented with a Roman letter
  Sample mean ($\overline{x}$) used to estimate the mean of the population from which that sample was drawn (μ)
  Sample standard deviation (*s*) used to estimate the corresponding population SD(σ)

Notations for Population vs. Sample

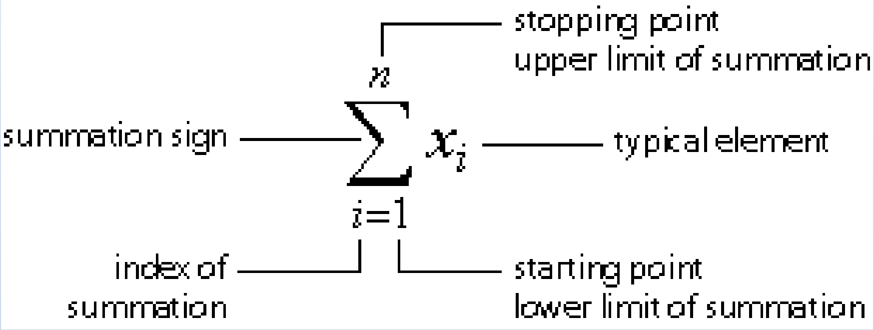


	Parameter	Statistic
Mean	$\mu = \frac{\sum X}{n}$	$\overline{X} = \frac{\sum X}{n}$
Variance	$\sigma^2 = \frac{\sum(\mu - X)^2}{n}$	$s^2 = \frac{\sum(\overline{X} - X)^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Random selection of a sample from the population *should* provide a representative estimate of the true population parameters.

Marino MJ. Biochemical Pharmacology, Volume 87, Issue 1, 2014, 78–92

Notation for Summation



i	x_i
1	10
2	20
3	30
4	40

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 10 + 20 + 30 + 40 = 100$$

SAMPLING

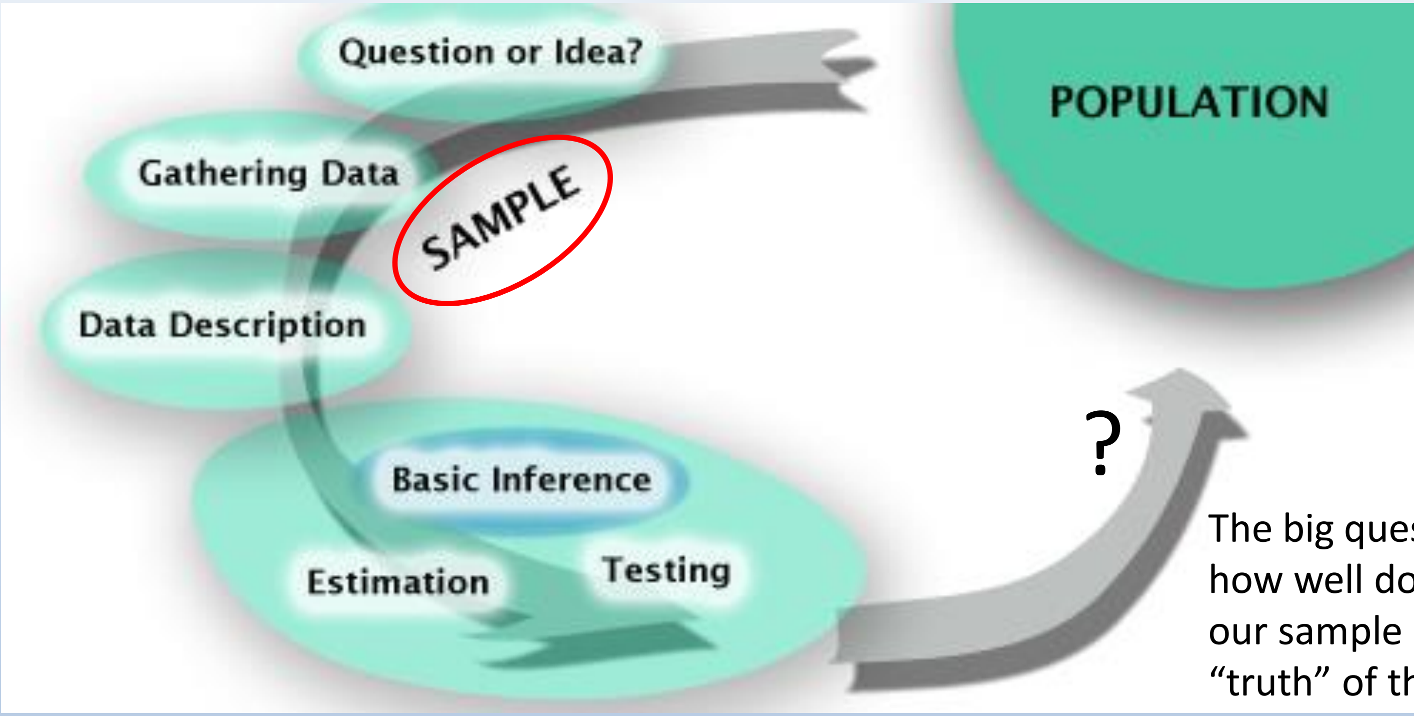
Why sample?

Costs less than measuring the entire population

It is impossible to study the whole population

It just must be done properly…

Classic Population Based Research:
Samples taken from a population of people



The big question is:
how well do the results from our sample represent the "truth" of the population?

Start with a question (hypothesis) about a population
Gather data from a subset of that population, known as a sample.
Describe that data using the various methods.
Use the data from the sample to make inferences about the larger population
Determine a measure of effect (estimation) and test a statistical hypothesis

Classic Research Design: Case-Control Study

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded increased from 612 to 9,287, or roughly fifteenfold. This remarkable increase is, of course, out of all proportion to the increase of population—both in total and, particularly, in its older age groups. Stocks (1947), using standardized death rates to allow for whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

Two main causes have from time to time been put forward: (1) a general atmospheric pollution from the exhaust fumes of cars, from the surface dust of tarred roads, and from gas-works, industrial plants, and coal fires; and (2) the smoking of tobacco. Some characteristics of the

They found reported cases of lung cancer from several hospitals

They matched each case to people from the same hospital who did not have lung cancer

They asked each person about their smoking habits

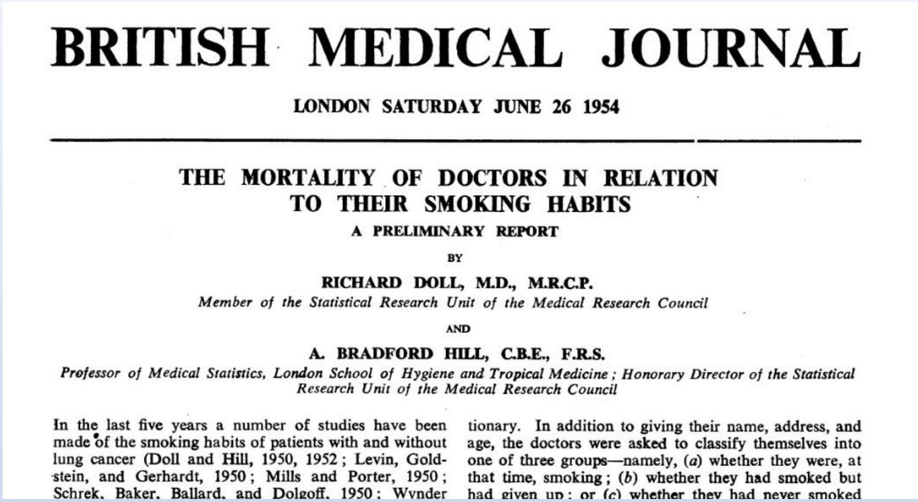
TABLE IV.—*Proportion of Smokers and Non-smokers in Lung-carcinoma Patients and in Control Patients with Diseases Other Than Cancer*

Disease Group	No. of Non-smokers	No. of Smokers	Probability Test
Males: Lung-carcinoma patients (649)	2 (0·3%)	647	P (exact method) = 0·00000064
Control patients with diseases other than cancer (649) ..	27 (4·2%)	622	
Females: Lung-carcinoma patients (60)	19 (31·7%)	41	$\chi^2 = 5{\cdot}76$; n = 1 $0{\cdot}01 < P < 0{\cdot}02$
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	

This is a case-control study.

They are sampling from the larger population of people with lung cancer and those without

Classic Research Design: Cohort Study



BRITISH MEDICAL JOURNAL

LONDON SATURDAY JUNE 26 1954

THE MORTALITY OF DOCTORS IN RELATION TO THEIR SMOKING HABITS

A PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, C.B.E., F.R.S.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

In the last five years a number of studies have been made of the smoking habits of patients with and without lung cancer (Doll and Hill, 1950, 1952; Levin, Goldstein, and Gerhardt, 1950; Mills and Porter, 1950; Schrek, Baker, Ballard, and Dolgoff, 1950; Wynder

tionary. In addition to giving their name, address, and age, the doctors were asked to classify themselves into one of three groups—namely, (*a*) whether they were, at that time, smoking; (*b*) whether they had smoked but had given up; or (*c*) whether they had never smoked

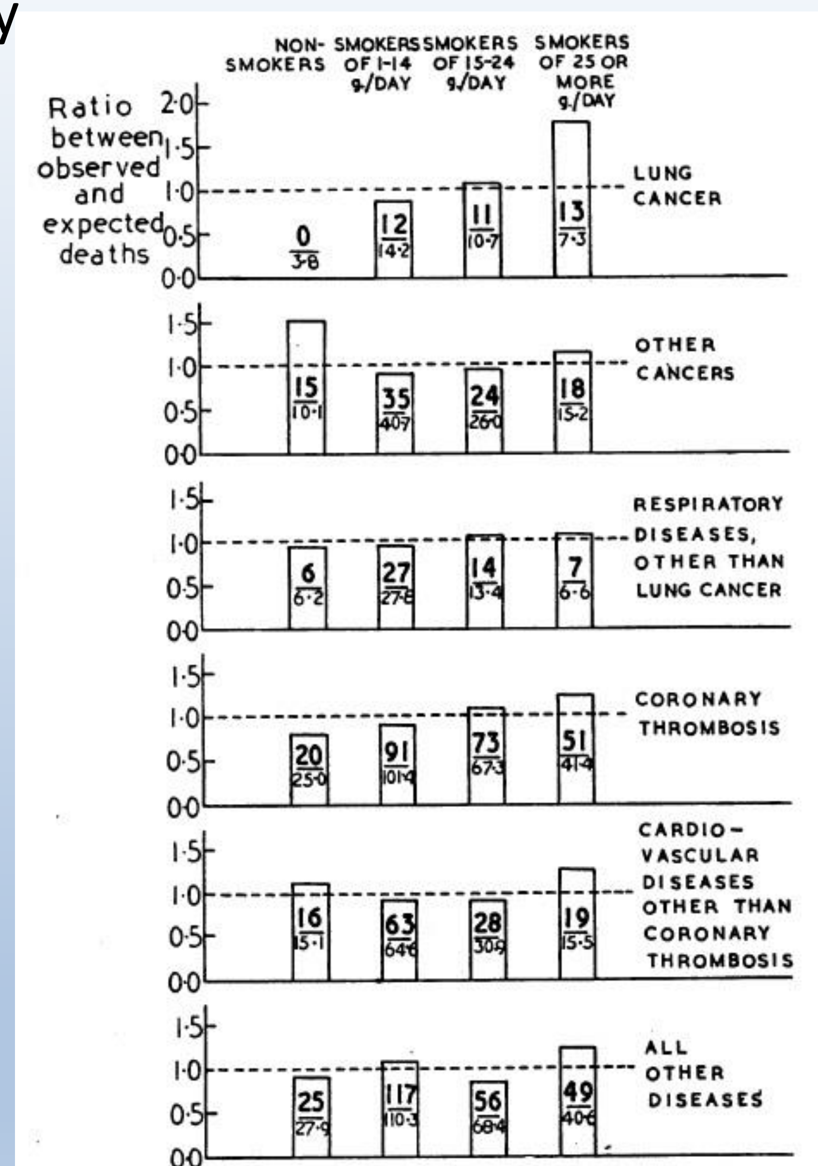


Chart showing variation in mortality with amount smoked. The ordinate shows the ratio between the number of deaths observed and the number expected (as entered in each column).

They surveyed all licensed physicians in the UK
- Asked about smoking habits
- Followed over time and recorded who died from what

This is a cohort (AKA prospective) study

The sample is drawn from the population of physicians (representative of the general population?)

Modeling Breast Cancer Intertumor and Intratumor Heterogeneity Using Xenografts

ALEJANDRA BRUNA,[1] OSCAR M. RUEDA,[1] AND CARLOS CALDAS[1,2]
[1]*Department of Oncology and Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge CB2 0RE, United Kingdom*
[2]*Breast Cancer Programme, Cambridge Cancer Centre, Cambridge CB2 2QQ, United Kingdom*
Correspondence: carlos.caldas@cruk.cam.ac.uk

Cold Spring Harb Symp Quant Biol 2016.
81: 227-230

From one woman's tumor directly

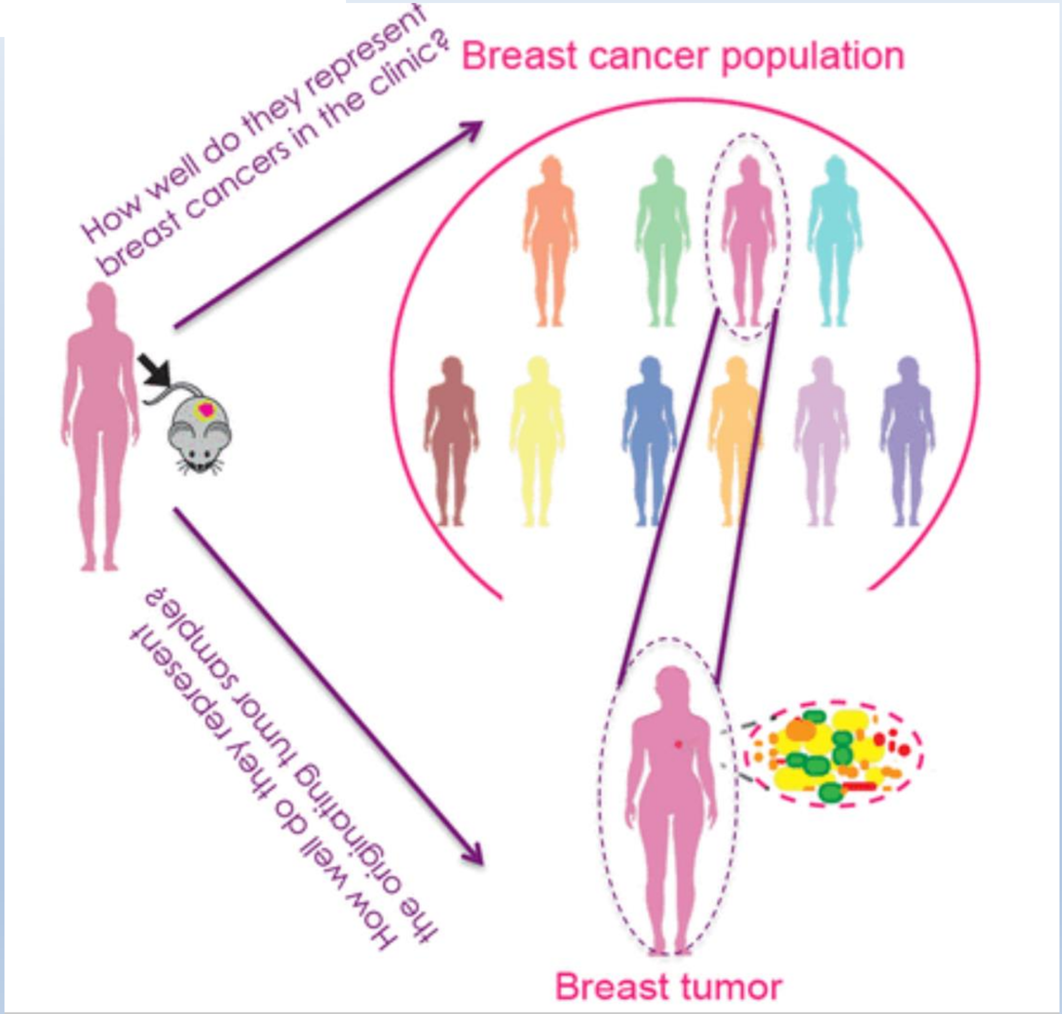
or from

a human breast cancer cell Line

MCF-7

Invasive ductal carcinoma derived from metastatic site: Pleural effusion

Sex of cell:	Female
Age at sampling:	69Y
Category:	Cancer cell line



web.expasy.org/cellosaurus

Review

Molecular heterogeneity in breast cancer: State of the science and implications for patient care

Seminars in Cell & Developmental Biology 64 (2017) 65–72

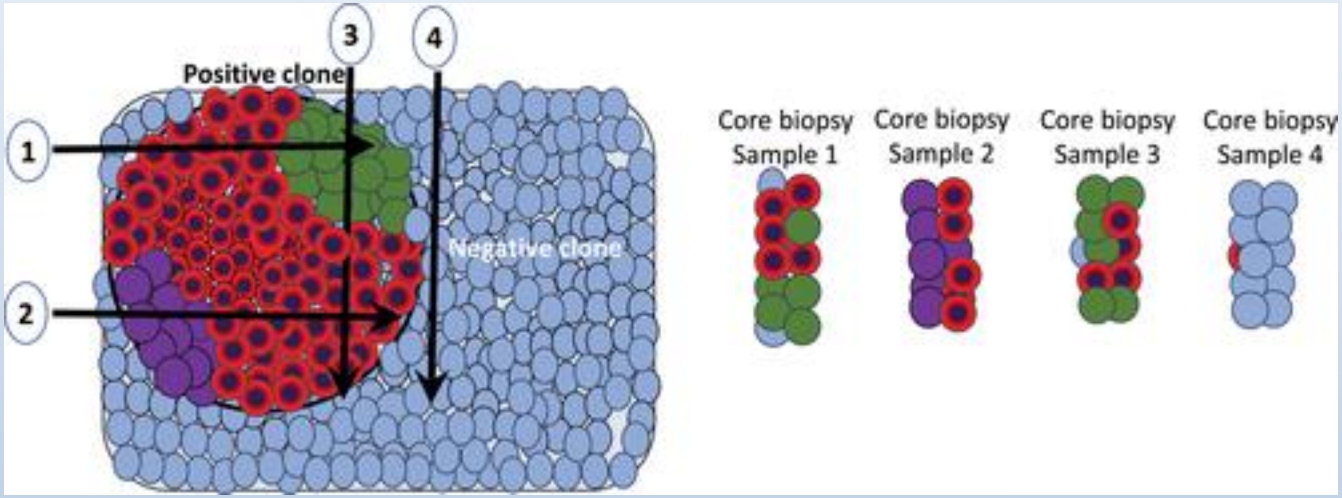
Rachel E. Ellsworth[a], Heather L. Blackburn[b], Craig D. Shriver[c], Patrick Soon-Shiong[d], Darrell L. Ellsworth[b,*]

Breast cancer cell lines: Extensive cell to cell heterogeneity within a single cell line.

Single-cell-derived sub-clones from the HCC38 breast cancer cell line have extensive copy number heterogeneity among individual cells that could be attributed to novel DNA changes during cell division in genomically unstable cancer cells.

Sequencing of single cells from an ER-/HER2+ cell line revealed additional mutations not detected in bulk genomic DNA from the same cell line and showed that no two cells had identical mutation profiles.

Schematic representation of breast cancer intratumor heterogeneity and **sampling** in needle core biopsy

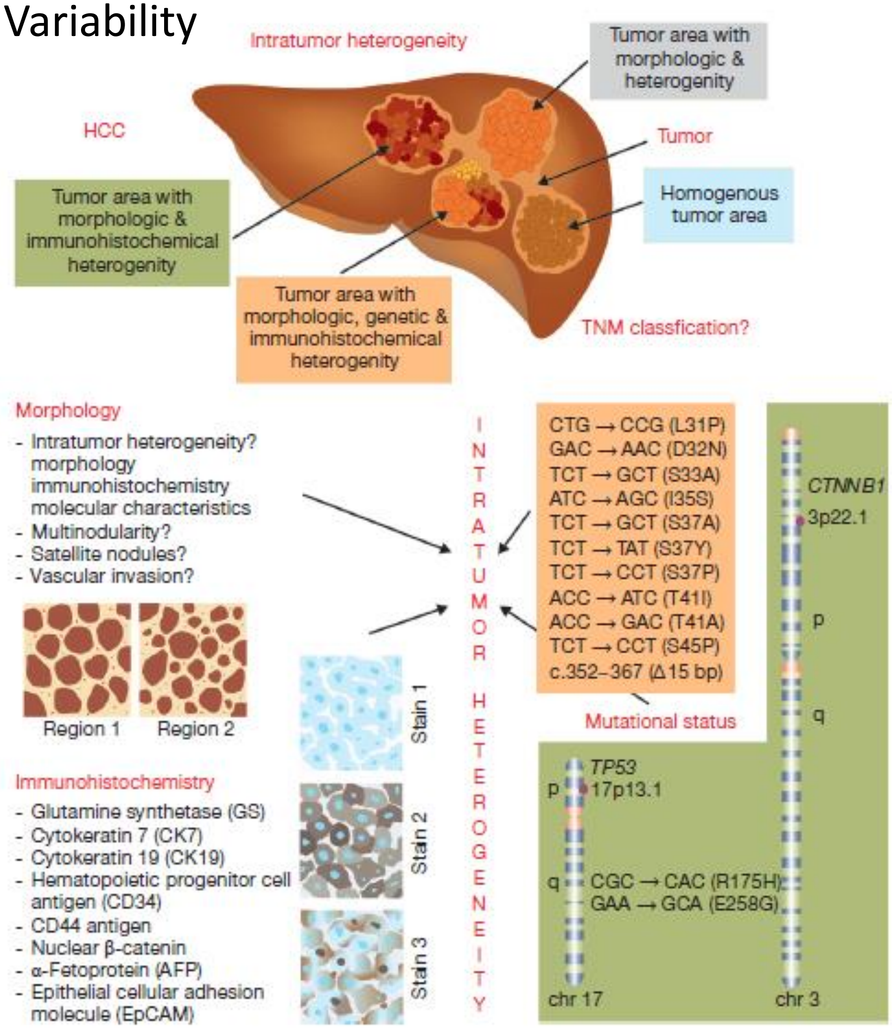


Intratumor heterogeneity shows clones that are positive and negative for specific features in the same tumor tissue. **1, 2, 3** and **4** show the incomplete representation of the different clones in core biopsies of the tumor tissue, because of intratumor heterogeneity. Blue denotes a negative clone. Red, green and purple denote different clones in the tumor.

Histopathology, Volume: 73, Issue: 5, Pages: 717-731, First published: 02 May 2018, DOI: (10.1111/his.13642)

HepatoBiliary Surgery and Nutrition, Vol 5, No 2 April 2016

Intratumor Variability

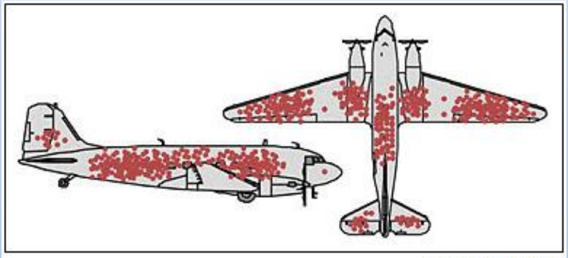


23 patients with HCC

only 3 had tumors that were homogeneous based on morphology and immunohistochemistry

A Historical Example of Why Sampling Error Matters

Abraham Wald, an Austrian statistician, contributed to the U.S. effort in World War II as a member of the Statistical Research Group. One of his projects was to estimate the survival rates of aircraft that had been shot. The data showed that returning aircraft were riddled with bullet holes to the wings but the engines had emerged relatively unscathed. Some people concluded that the wings needed more reinforcement.



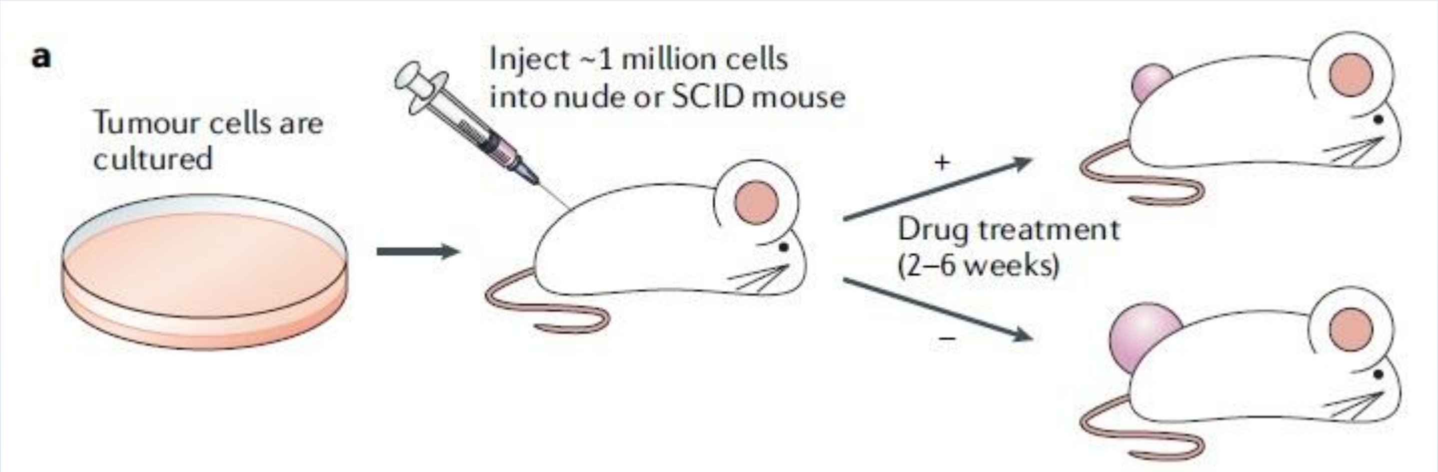
Credit: Cameron Moll

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Wald realized that the exact opposite was the case. The reason that the returning planes had few bullet holes in the engine was that the planes that *had* been hit there had crashed. In other words, there was a strong *bias* in how the planes entered their study. The data did not come from a random sample of all the airplanes that had flown, but only the planes that had survived. Those planes by definition had less serious, non-fatal damage. Thus, he argued, the most important place to reinforce the hull was the engine.

What is the reference population?

Classic Basic Research Study: Xenograft Study



Example: Cultured cells are generally injected subcutaneously into immunodeficient mice, which are then treated with the compound of interest for 2–6 weeks during which time subcutaneous tumors develop. The hypothesis is the compound of interest reduces the rate of growth of tumors

- What is the population to which we will apply the results of the study?
- What population did the sample come from?

Differences Between Research in Human Populations and in Basic Science

- The reference population is less clearly defined in basic sciences
- Independence between samples is often not easy to define in basic sciences
  - An important assumption for correct use of statistical tests (more on this later)
- Sample sizes are often small in basic science experiments ($n<10$)
- Statistics is usually taught from the prospective of classic population studies
  - Exception: big data and the "-omics" disciplines have their own special statistical rules
- The issues special to the basic sciences are often ignored

Will discuss more of this in a future lecture….For now let's proceed with classical statistics

Descriptive vs. Inferential Statistics

Descriptive statistics uses data to provide descriptions of the population or sample, either through numerical calculations (i.e., means, standard deviations) or graphs or tables.

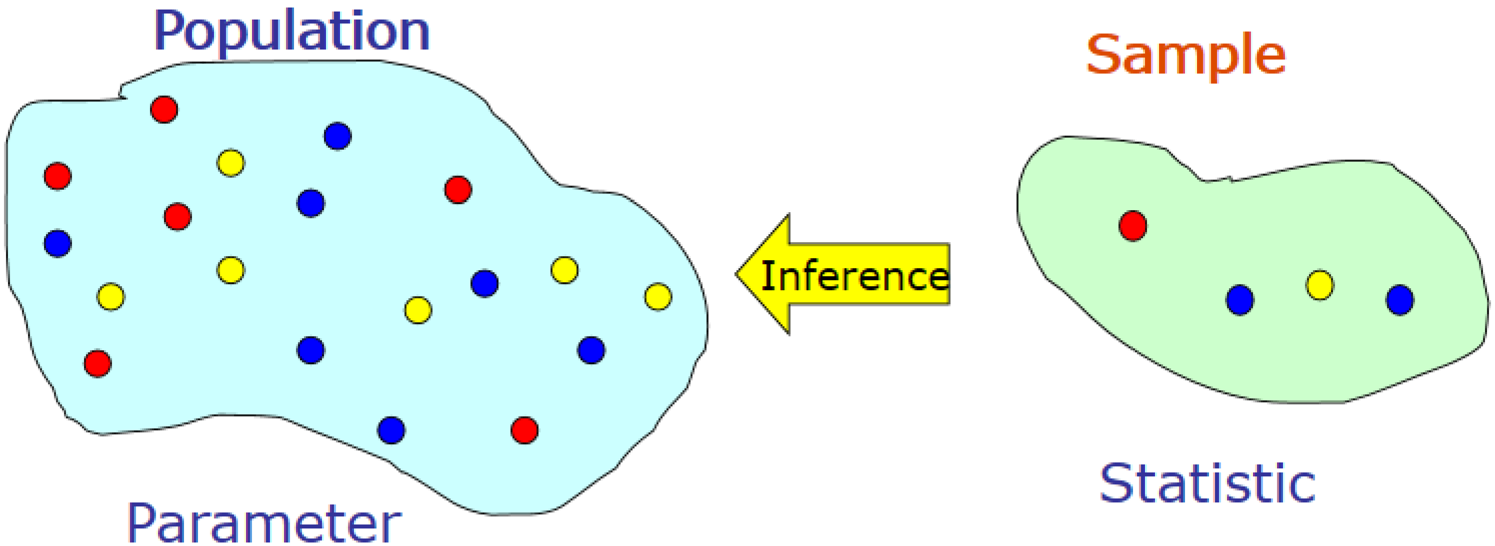
Used to summarize the data (population or sample)

Inferential statistics use a sample taken from a population to describe and make inferences about the population.

Testing a statistical hypothesis and drawing conclusions about a population, based on a sample using statistical tests like the ANOVA, t-Test, Chi-Square , or calculating confidence intervals, etc.

Statistical inference (Decisions)

- *Statistical inference* is the *process* of making an estimate, prediction, or decision about a population based on a sample.



What can we **infer** about a *population's parameters* based on a *sample's statistics*?

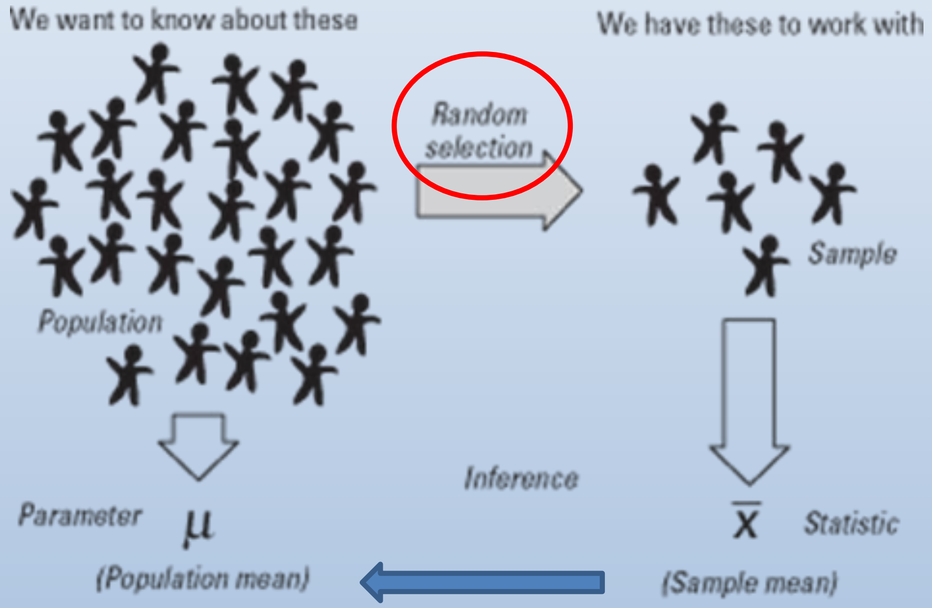
14

If your "sample" is the entire population, only descriptive statistics are needed

Image: slideplayer.com/slide/7536181/

Populations and samples

Taking a sample from a population



But do sample data ‘represent’ the population?
Random sampling helps.

Characteristics of a “Good” Sample

Random

A random *sample* is one in which each item in the *sample* has an equal chance of being selected

Representative of the parent population

Independent

Sufficient Size (power)

What is Random Sampling?

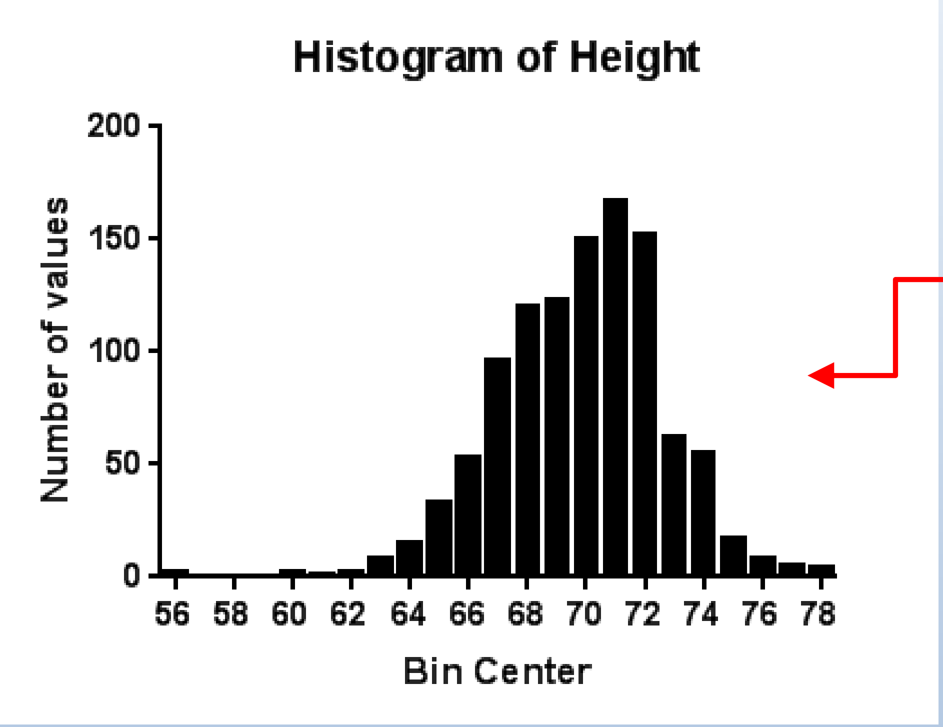
Random sampling is important because it results in samples that are more likely to be *representative* of the population

Random sampling allows us to apply the laws of probability to sample data, and to make valid inferences about the corresponding populations

Non-probability samples

- Convenience samples (ease of access)
  - Case-control sample of Doll and Hill paper
- Probability of being chosen is unknown
- Could be cheaper- but unable to generalize
- Potential for bias

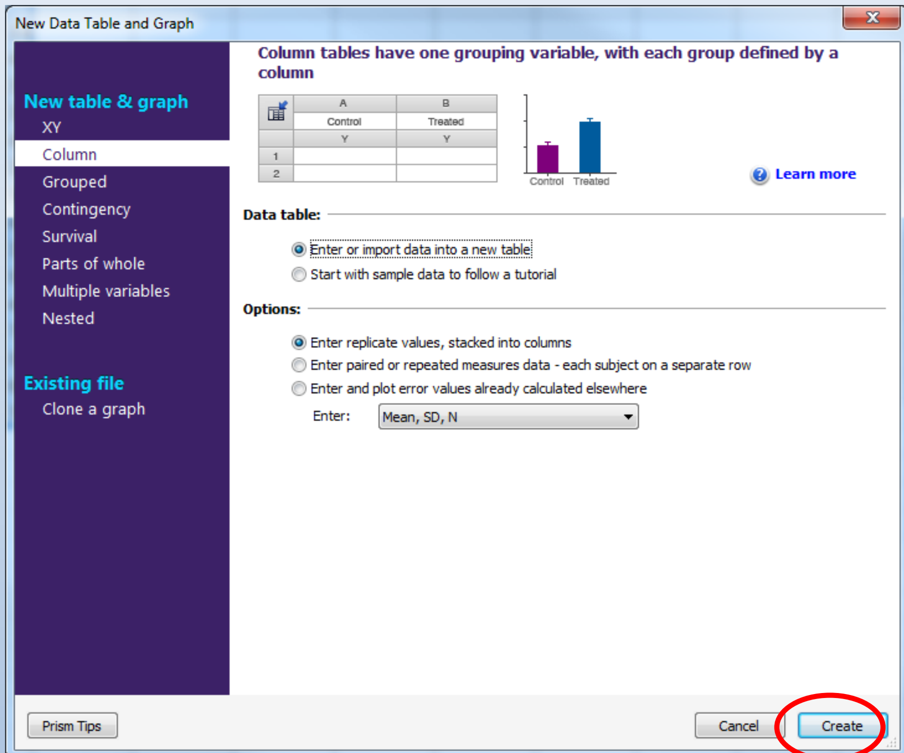
Example of Simple Random Sampling
Distribution of Heights (Inches) in a Cohort of 1,075 Men

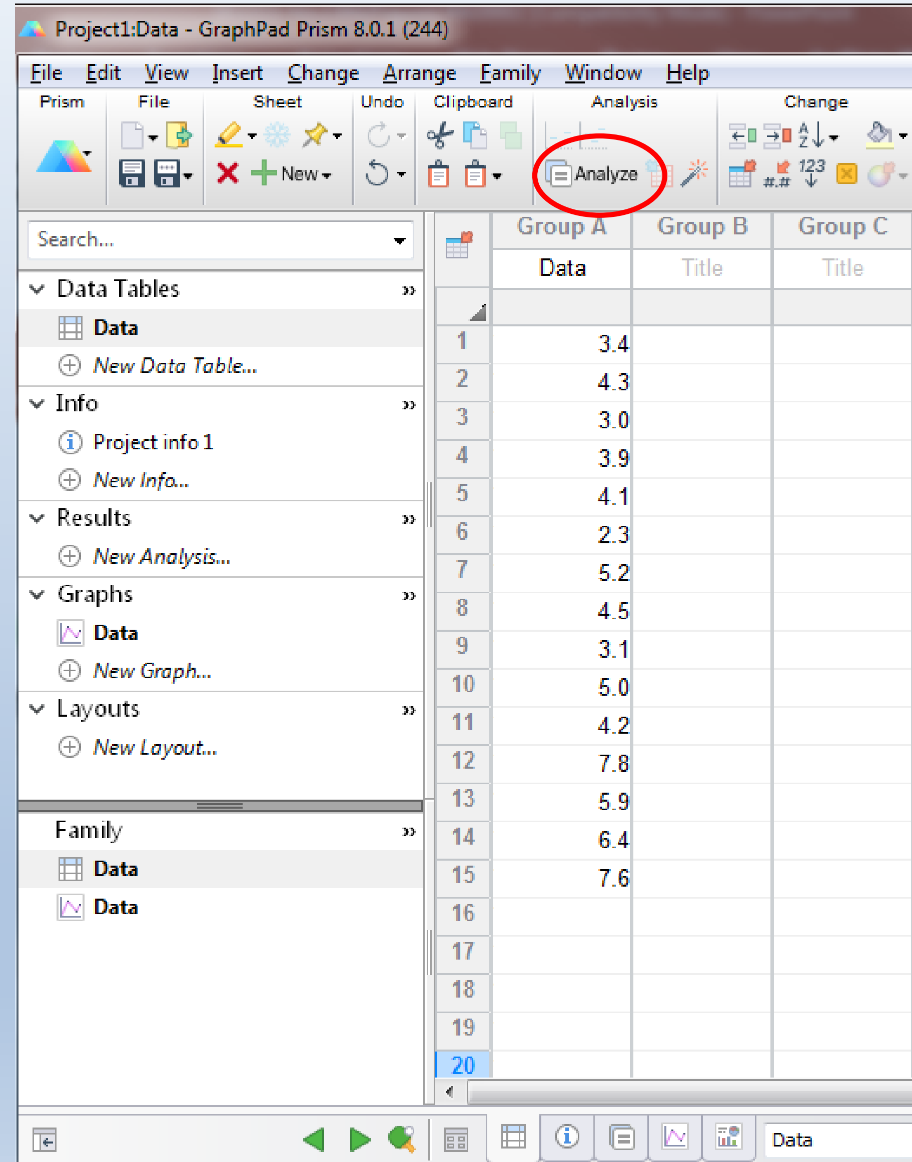


Frequency distribution graph

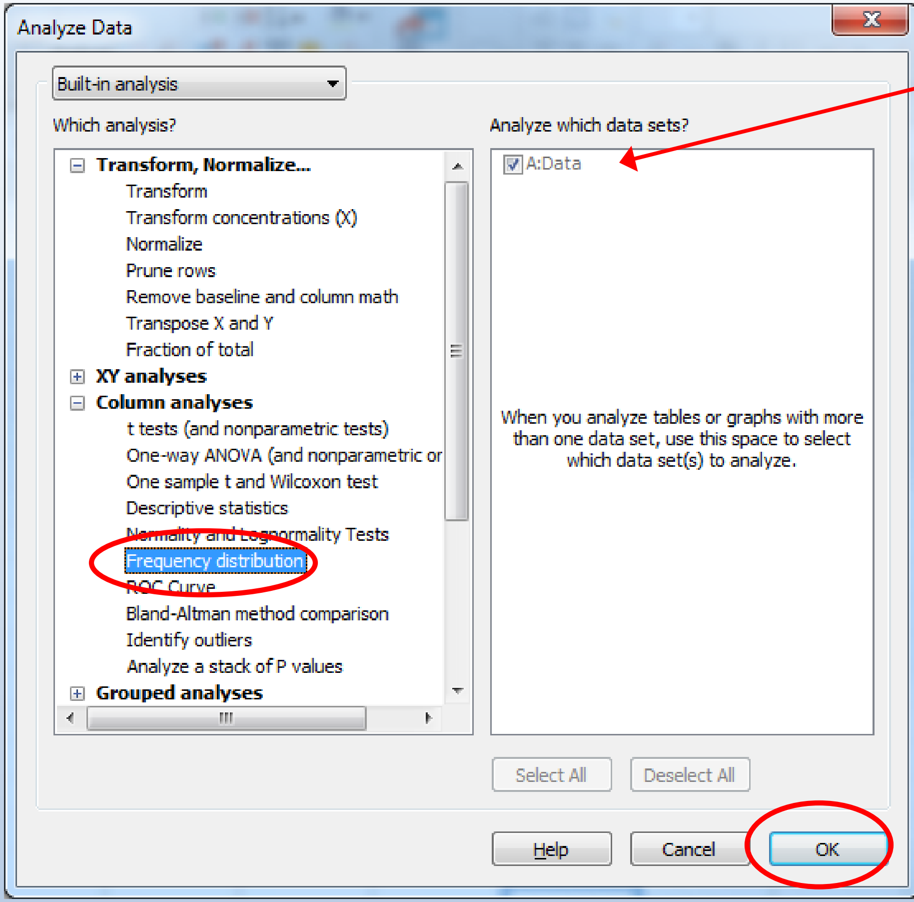
Mean = 69.8 inches
Range: 56-78

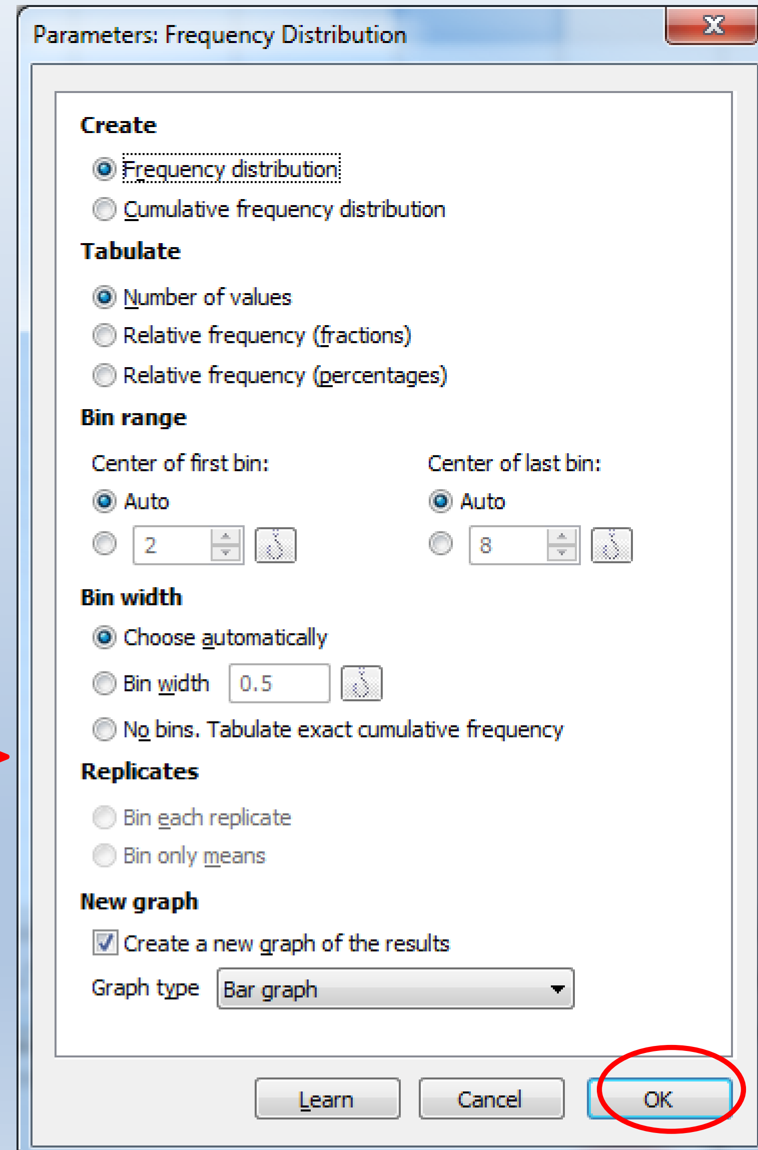
Creating Frequency Distribution Graphs in Prism



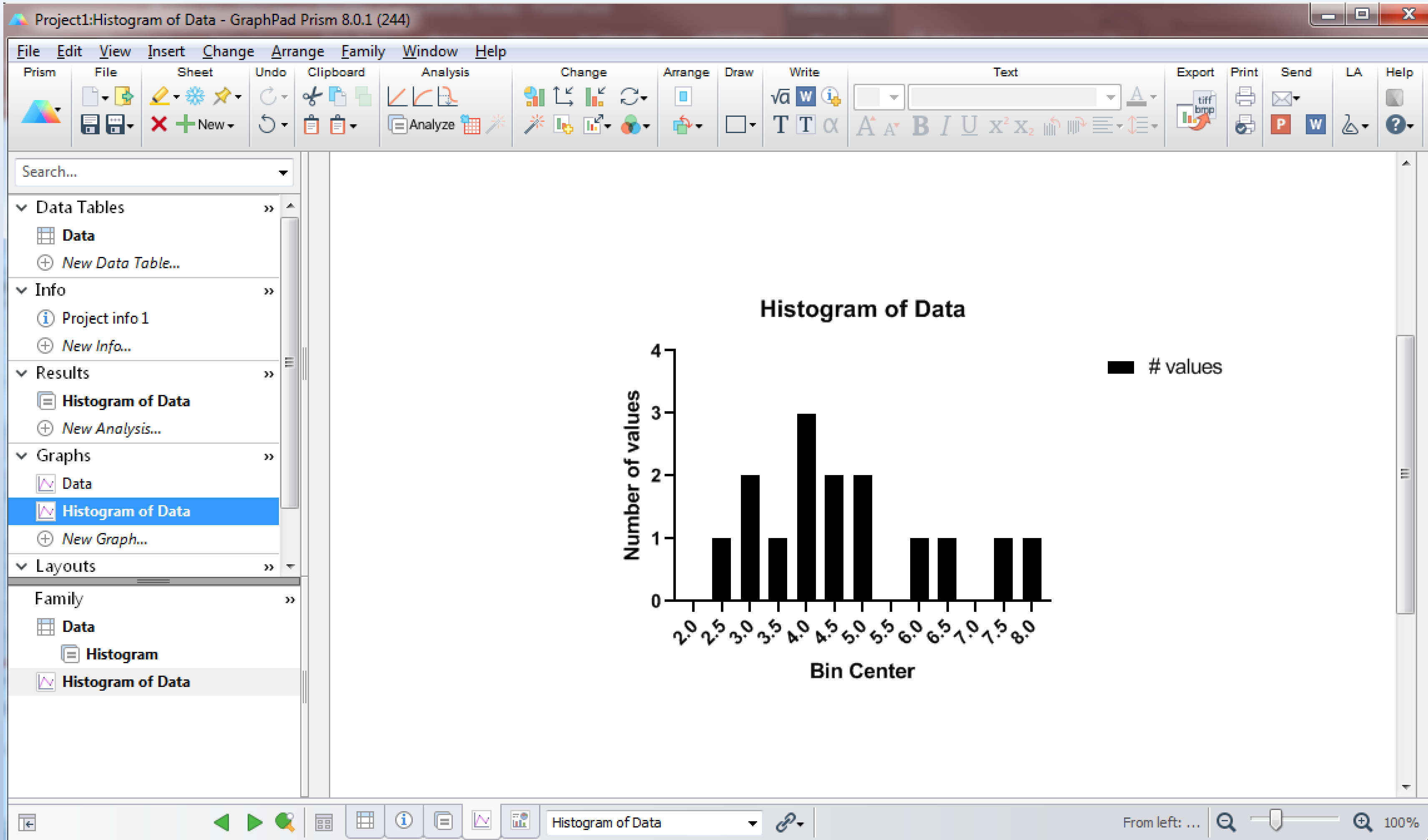


If you have multiple columns, check the one you want for the frequency distribution





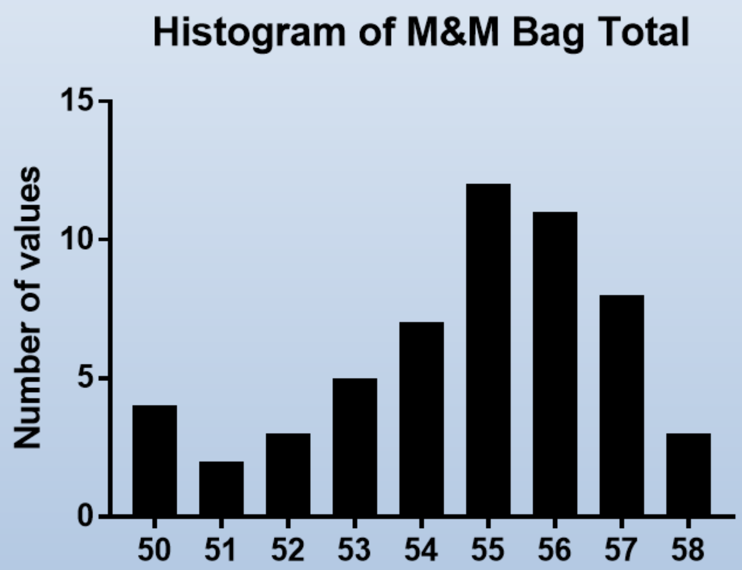
Voilà, a frequency distribution

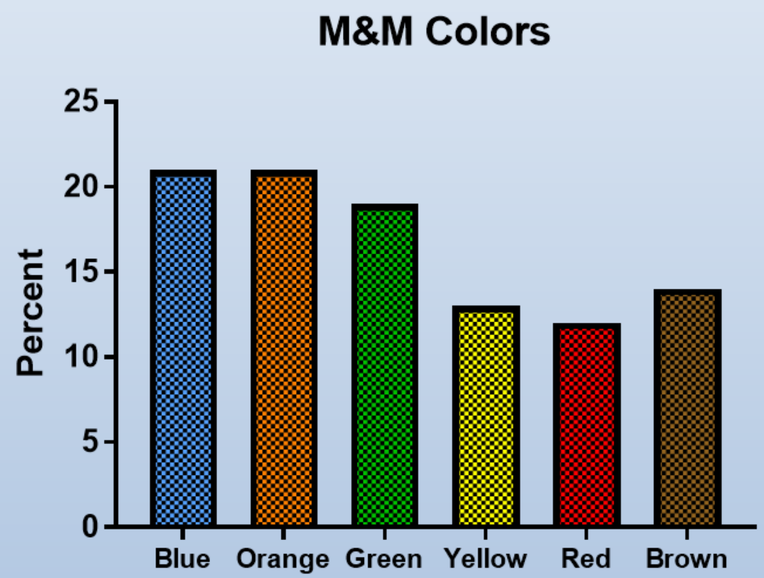


You can edit the figure and then copy it into Word or Excel.

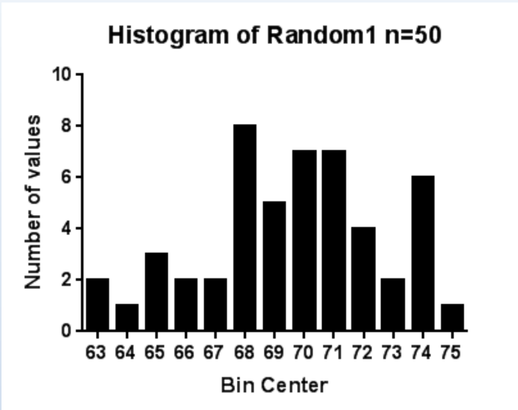
Our Data from Week01Assignment: M&Ms

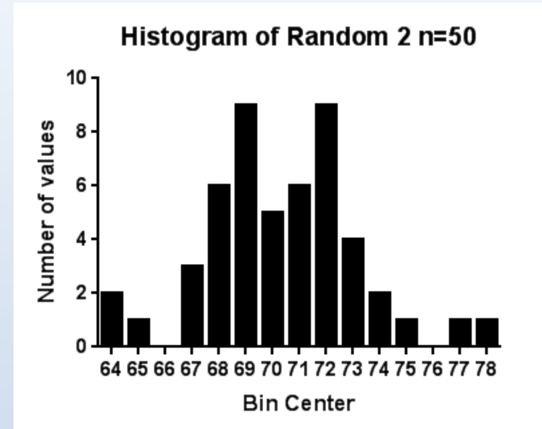
2018



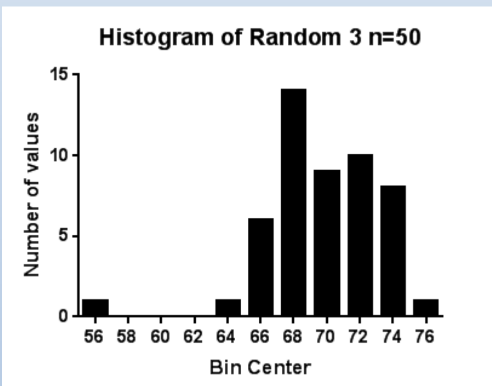


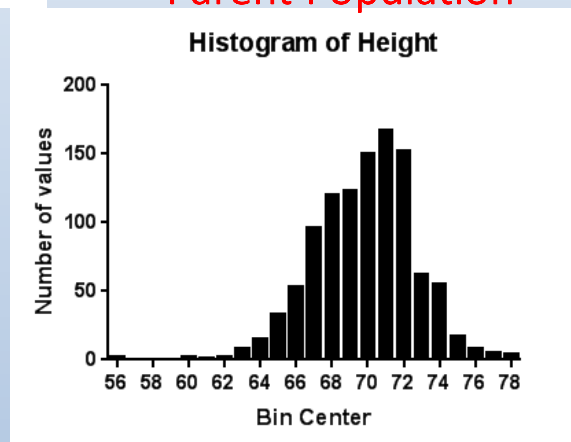
How Does Sampling Estimate the Truth: Examples of Random Samples of 50

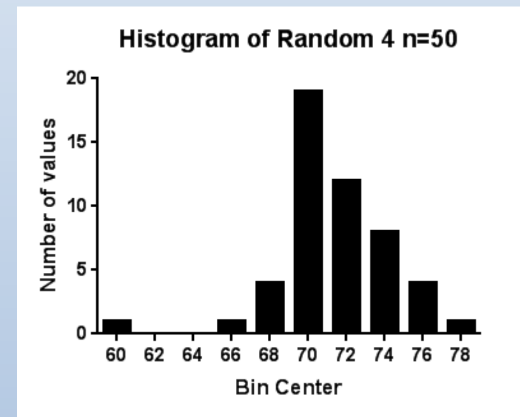


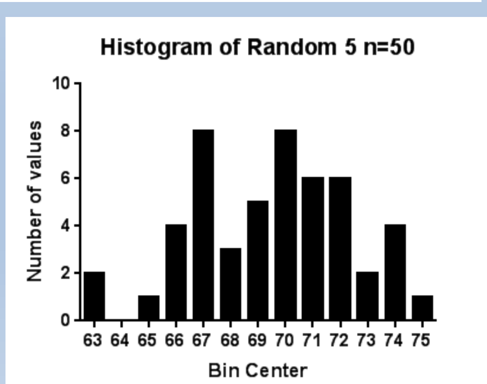


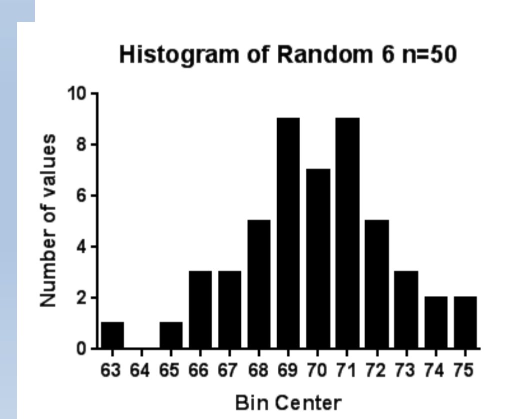
Parent Population





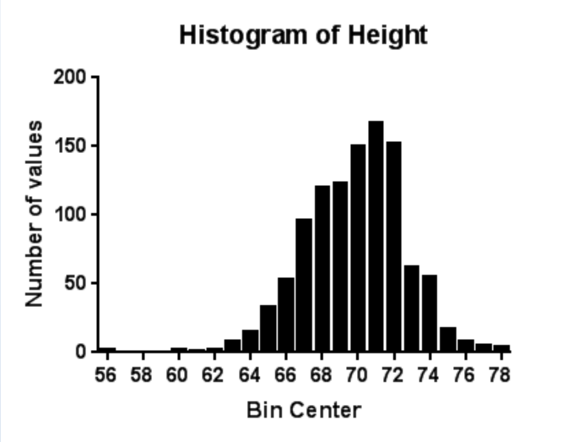






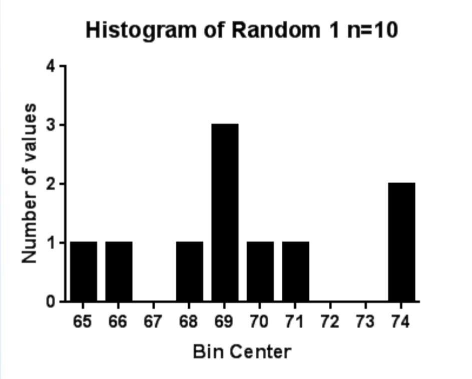
Means for Height from the Random Samples of 50
How Do They Compare to the Truth (the mean from the parent population)?

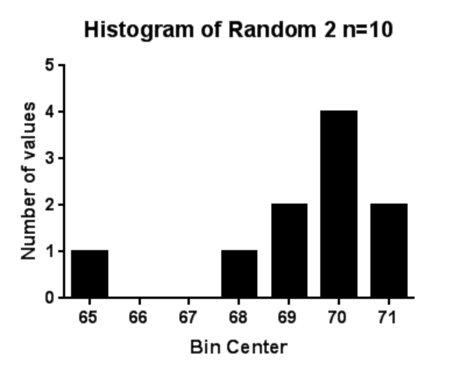
Group	Mean	Difference
Truth	69.8	---
Random1	69.6	-0.2
Random2	70.3	0.5
Random3	69.2	-0.6
Random4	70.9	1.1
Random5	69.5	-0.3
Random6	69.9	0.1
Random7	69.5	-0.3
Random8	70.1	0.3
Random9	70.4	0.6
Random10	70.1	0.3
Mean	69.9	0.43*
Range	69.2-70.9	-0.6 to 1.1



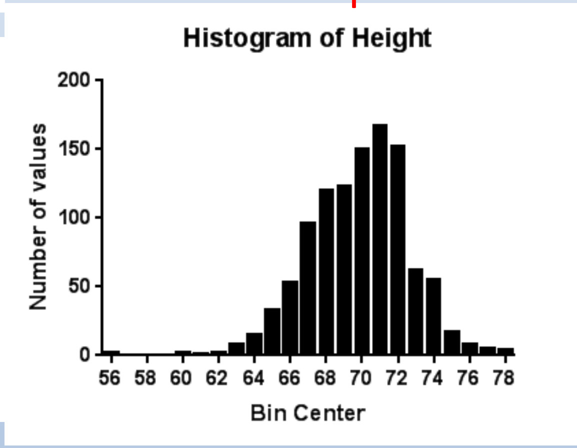
*Calculated from absolute values of the difference

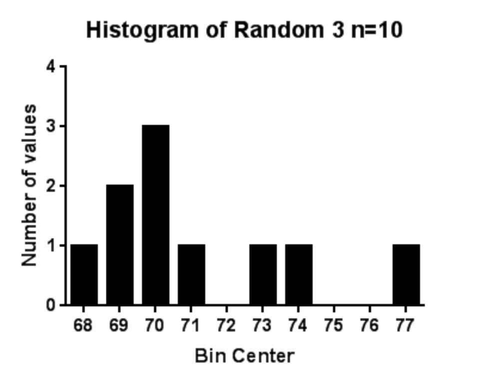
Examples of Random Samples of 10

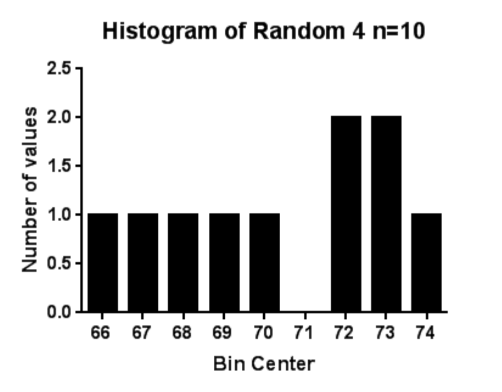


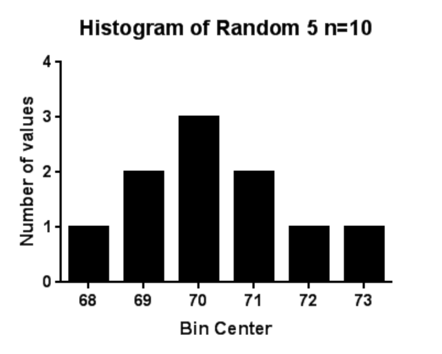


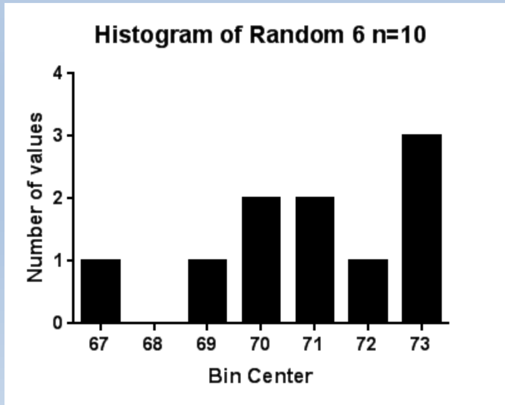
Parent Population







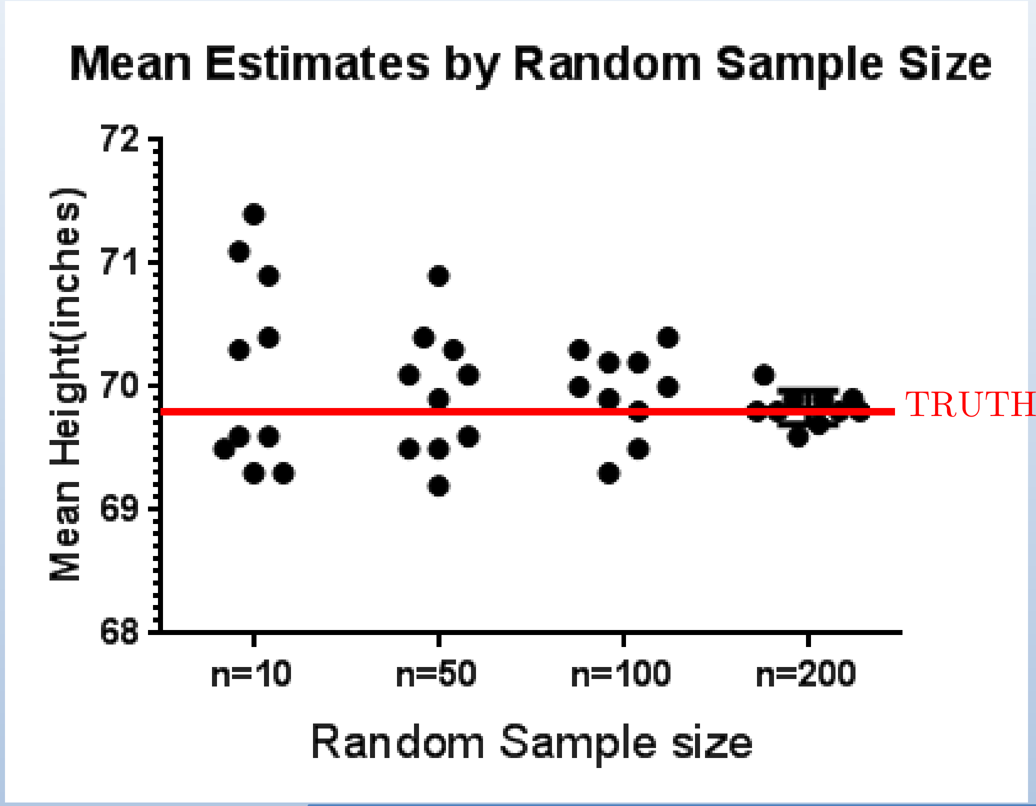


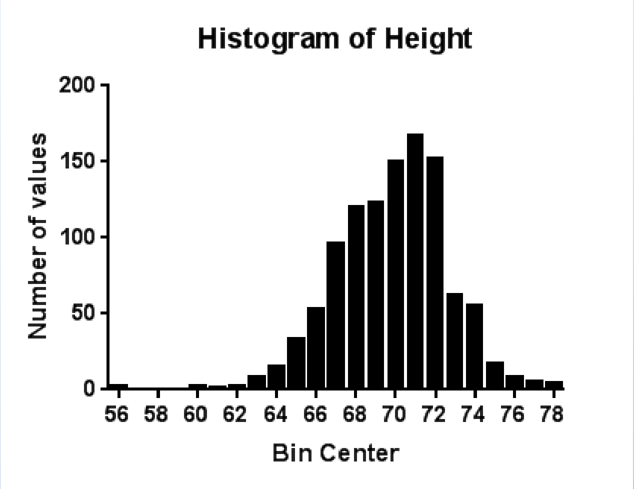


Means from the Random Samples of 50 and 10
How Do They Compare to the Truth?

Group	Mean	Difference	Group	Mean	Difference
Truth	69.8	---	Truth	69.8	---
n=50			n=10		
Random1	69.6	-0.2	Random1	69.5	-0.3
Random2	70.3	0.5	Random2	69.3	-0.5
Random3	69.2	-0.6	Random3	71.1	1.3
Random4	70.9	1.1	Random4	70.4	0.6
Random5	69.5	-0.3	Random5	70.3	0.5
Random6	69.9	0.1	Random6	70.9	1.1
Random7	69.5	-0.3	Random7	69.3	-0.5
Random8	70.1	0.3	Random8	71.4	1.6
Random9	70.4	0.6	Random9	69.6	-0.2
Random10	70.1	0.3	Random10	69.6	-0.2
Mean	69.9	0.43*		70.1	0.68*
Range	69.2–70.9	-0.6 to 1.1		69.3 – 71.4	-0.5 to 1.6

*Calculated from absolute values of the difference





	n=10	n=50	n=100	n=200
% of total sample	0.9%	4.7%	9.3%	18.6%
Mean difference from TRUTH for 10 random samples	0.68	0.43	0.32	0.10

Conclusions from Our Little Sampling Experiment

- The larger your sample size (the larger proportion of the parent population you sample), the more closely you approximate the truth
  - Sampling only a small proportion of a population can lead to an inaccurate estimate of the truth from individual samples

- Individual random samples can have estimates that are very different from the truth

  \- This is called **sampling error**

  i.e., our sample estimate for the mean height does not represent the truth

  \- Sampling error can also be caused by a non-random, non-representative sample from the parent population

Sampling error is everywhere – it is part of the sampling process

What is Bias?
A brief introduction

Systematic error in design or conduct or analysis of a study

Something we want to avoid.
It lurks everywhere…

http://skepchick.org/wp-content/uploads/2014/04/experimenter-bias-vuokud.jpg

BIAS

Systematic, non-random deviation of results and inferences from the truth, or processes leading to such deviation. Any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions which are systematically different from the truth.

Dictionary of Epidemiology, 3rd ed.

Sources of bias in *basic* research

Table 1. Sources and "Locations" of Bias in Marker Research

Source of Bias	Location of Bias: Before or After Specimens Are Received in the Laboratory		Example
	Before	After	
Features of subjects, determined in selection: Age Sex Comorbid conditions Medications	X		Cancer subjects are male, whereas control subjects are mainly female. Bias: Assay results may depend on sex.
Specimen collection	X		Cancer specimens come from one clinic, whereas controls come from a different clinic. Bias: Assay results may depend on conditions that differ between clinics.
Specimen storage and handling	X	X	Cancer specimens are stored for 10 years because it takes longer to collect them, whereas control specimens are collected and stored over 1 year. Bias: Assay results may vary with duration of storage, or with different numbers of thaw-freeze cycles.
Specimen analysis		X	Cancer specimens are run on one day, whereas control specimens are run on a different day. Bias: Assay results may depend on day of analysis in a machine that "wanders" over time.

NOTE. The table shows examples of different sources of bias and the location of the bias before or after specimens are received in the laboratory. The list is not exhaustive; other biases may be important, and the biases listed may or may not be important in any given research study, depending on details of biology and technology (ie, what is being measured and how it might be influenced).

Bias vs. Sampling Error

Bias	Sampling error
Is due to *mistakes* which can be avoided Cannot be precisely measured Control and prevention requires careful attention	Is unavoidable if sampling < 100% of population Can be controlled by selecting appropriate sample size and sampling method

- Bias is the difference between sample value and population value due to error in measurement, selection of non-representative sample or other factors
- A large sample size cannot guarantee absence of bias
- The best way to reduce bias is to plan, plan, plan.

Plan Ahead, Really Ahead, especially when substantial effort will be involved in collecting data

There are many questions

What should my *n* be?
How do I collect data to answer my research questions?
What would be the ideal outcome of my experiment?
What kind of statistical tests might I need?
How do I collect data to fit the appropriate tests?
How would you interpret the results?

Find and control for factors that might introduce systematic errors (bias).

Asking questions at the design stage can save headaches at the analysis stage:

One of the Ten Simple Rules for Effective Statistical Practice:

Talk to Your Statistician Before You Start Collecting Data

Statistical experts can help their scientific collaborators with ways that data might answer their research questions and what kinds of studies might be most useful

Identify potential sources of variability and bias
Develop analytic goals and strategies.

The collaborative process works best when initiated early in an investigation

Kass RE et al (2016) Ten Simple Rules for Effective Statistical Practice. PLoS Comput Biol 12(6): e1004961.

Careful data collection can greatly simplify analysis and make it more rigorous. Or, as Sir Ronald Fisher put it:

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of”

Assignment Week01
Already done

"Numbers don't lie. That's where we come in."