Elements of statistics (MATH0487-1)

Prof. Dr. Dr. K. Van Steen

University of Liège, Belgium

September 17, 2012

Outline

Why study Statistics?

- We like to think that we have control over our lives.
- But in reality there are many things that are outside our control.
- Everyday we are confronted by our own ignorance.
- According to Albert Einstein:

God does not play dice

- The world is governed by quantum mechanics where probability reigns supreme.

Relevant Questions to Probability

If someone asks you what probability is, can you point out a key question to him/her?

Consider a day in the life of an average ULg student

- You wake up in the morning and the sunlight hits your eyes. Then suddenly without warning the world becomes an uncertain place.
- How long will you have to wait for the Number 48 bus this morning?
- When it arrives will it be full?
- Will it be out of service?
- Will it be raining while you wait?
- Will you be late for your stats lecture?

Probability is the Science of Uncertainty

- Probability originated from the study of games of chance and gambling during the 16th century
- It is derived from the verb "to probe": to "find" out what is not easily accessible or understandable
- Probability was a branch of mathematics (Blaise Pascal and Pierre Fermat)

Probability is the Science of Uncertainty

- It is used by physicists to predict the behaviour of elementary particles.
- It is used by engineers to build computers.
- It is used by economists to predict the behaviour of the economy.
- It is used by stockbrokers to make money on the stockmarket.
- It is used by psychologists to determine if you should get that job.

Probability is the Science of Uncertainty

1. Rules $\rightarrow$ data: Given the rules, describe the likelihoods of various events occurring.
2. Probability is about prediction - looking forward.
3. Probability is mathematics.

What about Statistics?

- Statistics was born in the mid 17th century.
- It originated from John Graunt reviewing a weekly church publication issued by the local parish clerk that listed the number of births, christenings and death in each parish. These so-called "Bills of Mortality" also listed the causes of death.
- The way the "data" were organized was what we call now "descriptive statistics".
- Statistics is the science of data ...

Descriptive Statistics

With **descriptive statistics** we condense a set of known numbers into a few simple values (either numerically or graphically) to simplify an understanding of those data that are available to us.

- This is analogous to writing up a summary of a lengthy book. The book summary is a tool for conveying the gist of a story to others.
- The mean and some measures of spread of a set of numbers is a tool for conveying the gist of the individual numbers (without having to specify each and every one).

Statistics is the Science of Data

- The original idea of statistics was the collection of information about and for the "state". The word statistics derives directly, not from any classical Greek or Latin roots, but from the Italian word for state.
- With the "Bills of Mortality" in mind, statistics has to borrow some concepts from sociology, such as the concept of "population".
- It has been argued that since statistics usually involves the study of human behavior, it cannot claim the precision of the physical sciences.
- Although new and ever growing diverse fields of human activities are using statistics, the field itself remains hard to access to the larger public.

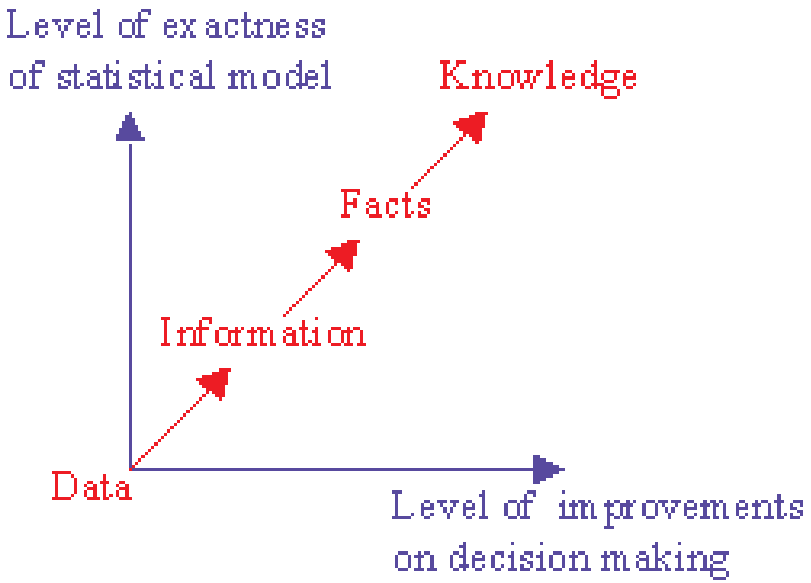
Statistics is the Science of Data

1. Rules $\leftarrow$ data: Given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess - or approximate - what that model was. We might guess wrong; we might refine our guess as we get more data.
2. Statistics is about looking backward.
3. Statistics is an art. It uses mathematical methods, but it is more than maths.
4. Once we make our best *statistical guess* about what the probability model is (what the rules are), based on looking backward, we can then use that *probability model* to predict the *future* $\rightarrow$ The purpose of statistics is to make inference about unknown quantities from samples of data

Statistics is the Science of Data

- **Sampling and experimentation**: Clarifying the question, deciding on methods of collection and analysis to produce valid information.
- **Exploring data**: Using graphical and numerical techniques to study patterns and departures from patterns (in order to interpreting data)
- **Anticipating patterns**: Exploring random phenomena using probability and simulation. Probability is our tool for anticipating distributions . . .
- **Statistical Inference**: Estimating population parameters and testing hypothesis.

Statistical Modeling under Uncertainties: From Data to Knowledge



Relevant Questions to Statistics

If someone asks you what statistics is, can you point out a key question to him/her?

Consider a day in the life of an average ULg student

- You wake up in the morning and the sunlight hits your eyes. Then suddenly without warning the world becomes an uncertain place.
- How long will you have to wait for the Number 10 bus this morning?
- When it arrives will it be full?
- Will it be out of service?
- Will it be raining while you wait?
- Will you be late for your stats lecture?

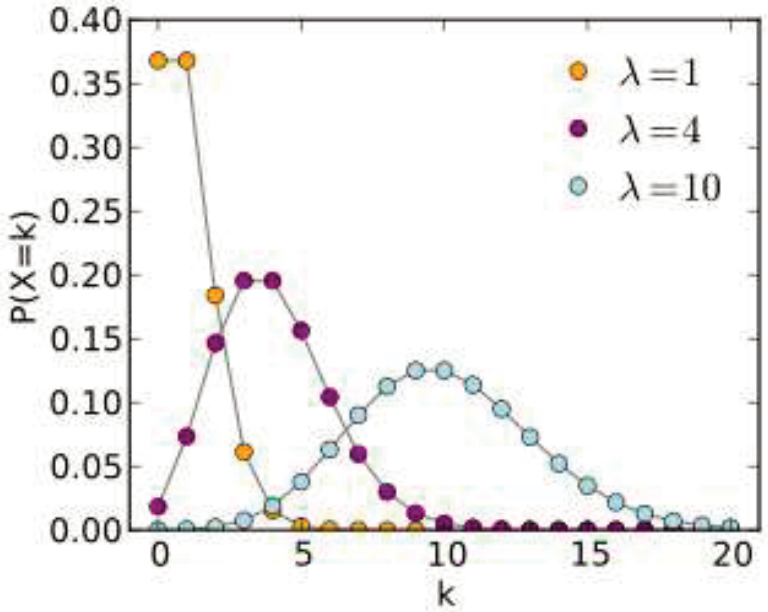
Example: Particle Emission

- X: the number of particles that *will* be emitted from a radioactive source in the next one minute period.

- We know that X will turn out to be equal to one of the non-negative integers but, apart from that, we know nothing about which of the possible values are more or less likely to occur.
- The quantity X is said to be a random variable. [*In fact, a random variable is a function - see later*]

Example: Particle Emission

- Suppose we are told that the random variable X has a Poisson distribution with parameter $\lambda = 2$.



Example: Particle Emission

- Then, if x is some non-negative integer, we know that the probability that the random variable X takes the value x is given by the formula

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

where $\lambda = 2$. [*It does not matter how we "denote" the parameter* ...]

- For instance, the probability that X takes the value $x = 4$ is

$$P(X = 4) = \frac{2^4 \exp(-2)}{4!} = 0.0902 \ .$$

- We have here a *probability model* for the random variable X.
- We are usually using upper case letters for random variables and lower case letters for the values taken by random variables - We shall persist with this convention throughout the course.

Example: Particle Emission

- Now suppose we are told that the random variable X has a Poisson distribution with parameter θ where θ is some unspecified (!) positive number.
- Then, if x is some non-negative integer, we know that the probability that the random variable X takes the value x is given by the formula

$$P(X = x|\theta) = \frac{\theta^x \exp(-\theta)}{x!},$$

  for $\theta \in \mathbb{R}^+$.
- However, we cannot calculate probabilities such as the probability that X takes the value $x = 4$ without knowing the value of θ.

Example: Particle Emission

- Suppose that, in order to learn something about the value of θ, we decide to measure the value of X for each of the next 5 one minute time periods.
- Let us use the notation X_1 to denote the number of particles emitted in the first period, X_2 to denote the number emitted in the second period and so forth.
- We shall end up with data consisting of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_5)$.
- Consider $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (2, 1, 0, 3, 4)$. Then $\mathbf{x}$ is a possible value for the random vector $\mathbf{X}$.

Example: Particle Emission

- We know that the probability that X_1 takes the value $x_1 = 2$ is given by the formula

$$P(X = 2|\theta) = \frac{\theta^2 \exp(-\theta)}{2!}$$

and similarly that the probability that X_2 takes the value $x_2 = 1$ is given by

$$P(X = 1|\theta) = \frac{\theta \exp(-\theta)}{1!}$$

and so on.

- However, what about the probability that $\mathbf{X}$ takes the value $\mathbf{x}$?

Example: Particle Emission

- In order for this probability to be specified we need to know something about the joint distribution of the random variables $X_1, X_2, \ldots, X_5$.
- A simple assumption to make is that the random variables $X_1, X_2, \ldots, X_5$ are *mutually independent*.
- This assumption may not be correct since X_2 may tend to be more similar to X_1 than it would be to X_5 !!!

Example: Particle Emission - an aside

- Two events are **independent** if the joint probability of both events occurring is the product of the probabilities of each event occurring:

$$P(A \cap B) = P(A) \times P(B).$$

- Independence of events is closely related to **conditional probability**: When events A and B are independent, then

$$P(A) = P(A|B) \equiv \frac{P(A \cap B)}{P(B)}, P(B) \neq 0.$$

Example: Particle Emission - an aside

- Likewise, independence of random variables is closely related to conditional distributions of random variables: The random variables $X_1, \ldots, X_k$ are **(stochastically) independent** if and only if

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i),$$

  for all $x_1, \ldots, x_k$.
- If two random variables X and Y are independent, then

$$f_{Y|X}(y|x) = f_Y(y);$$

  the conditional density of Y given x is the unconditional density of Y. Hence, to show that two random variables are NOT independent, it suffices to show that $f_{Y|X}(y|x)$ depends on x!

Example: Particle Emission

- However, with this assumption we can say that the probability that $\mathbf{X}$ takes the value $\mathbf{x}$ is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= \prod_{i=1}^{5} \frac{\theta^{x_i} \exp(-\theta)}{x_i!}, \\ &= \frac{\theta^2 \exp(-\theta)}{2!} \times \frac{\theta^1 \exp(-\theta)}{1!} \times \frac{\theta^0 \exp(-\theta)}{0!} \\ &\qquad \times \frac{\theta^3 \exp(-\theta)}{3!} \times \frac{\theta^4 \exp(-\theta)}{4!}, \\ &= \frac{\theta^{10} \exp(-5\theta)}{288}. \end{aligned}$$

Example: Particle Emission

- In general: if $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)$ is any vector of 5 non-negative integers, then the probability that $\mathbf{X}$ takes the value $\mathbf{x}$ is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= \prod_{i=1}^{5} \frac{\theta^{x_i} \exp(-\theta)}{x_i!}, \\ &= \frac{\theta^{\sum_{i=1}^{5} x_i} \exp(-5\theta)}{\prod_{i=1}^{5} x_i!}. \end{aligned}$$

- We have here a *probability model* for the random vector (!) $\mathbf{X}$.
- Our plan is to use the value $\mathbf{x}$ of $\mathbf{X}$ that we actually observe to learn something about the value of θ.
- The ways and means to accomplish this task make up a large part of this course.

What is inferential statistics?

- **Inference** Inference studies the way in which data we observe should influence our beliefs about and practices in the real world.
- **Statistical inference** Statistical inference considers how inference should proceed when the data is subject to random fluctuation.

What is inferential statistics?

- Inferential statistics is used to make claims about the populations that give rise to the data we collect.
- This requires that we go beyond the data available to us.
- Consequently, the claims we make about populations are always subject to error; hence the term "inductive inference" in the context of statistics.
- Inferential statistics encompasses a variety of procedures to ensure that the inferences are sound and rational, even though they may not always be correct.

Relation between Descriptive and Inferential Statistics

Relevant Questions for Descriptive and Inferential Statistics

Statistics (="state arithmetic")

Descriptive: describe data

- How rich are our citizens on average? → Central Tendency
- Are there many differences between rich and poor? → Variability
- Are more intelligent people richer? → Association
- How many people earn this money? → Probability distribution
- Tools: tables (all kinds of summaries), graphs (all kind of plots), distributions (joint, conditional, marginal, …), statistics (mean, variance, correlation coefficient, histogram, …)

Inferential: derive conclusions and make predictions

- Is my country so rich as my neighbors? → Inference
- To measure richness, do I have to consider EVERYONE? → Sampling
- If I don't consider everyone, how reliable is my estimate? → Confidence
- Is our economy in recession? → Prediction
- What will be the impact of an expensive oil? → Modelling
- Tools: Hypothesis testing, Confidence intervals, Parameter estimation, Experiment design, Sampling, Time models, Statistical models (ANOVA, Generalized Linear Models, …)

Problem Solving

- **Logic**: The science of correct reasoning.
- **Reasoning**: The drawing of inferences or conclusions from known or assumed facts.

When solving a problem, one must

- understand the question,
- gather all pertinent facts,
- analyze the problem i.e. compare with previous problems (note similarities and differences),
- perhaps use pictures or formulas to solve the problem

Deductive Reasoning

- **Deductive Reasoning**: A type of logic in which one goes from a general statement to a specific instance.
- The classic example:
  - All men are mortal. (major premise)
  - Socrates is a man. (minor premise)
  - Therefore, Socrates is mortal. (conclusion)

Deductive Reasoning

- The example on the previous slide is an example of a syllogism.
- **Syllogism**: An argument composed of two statements or premises (the major and minor premises), followed by a conclusion.
- For any given set of premises, if the conclusion is guaranteed, the argument is said to be valid.
- If the conclusion is not guaranteed (at least one instance in which the conclusion does not follow), the argument is said to be invalid.
- Be careful, do not confuse "truth" with "validity" ...

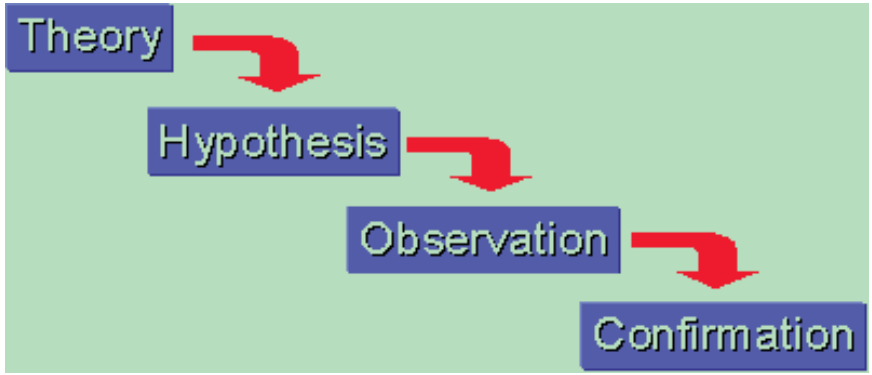
Inductive Reasoning

- **Inductive Reasoning** involves going from a series of specific cases to a general statement.
- The conclusion in an inductive argument is never guaranteed.
- Example:

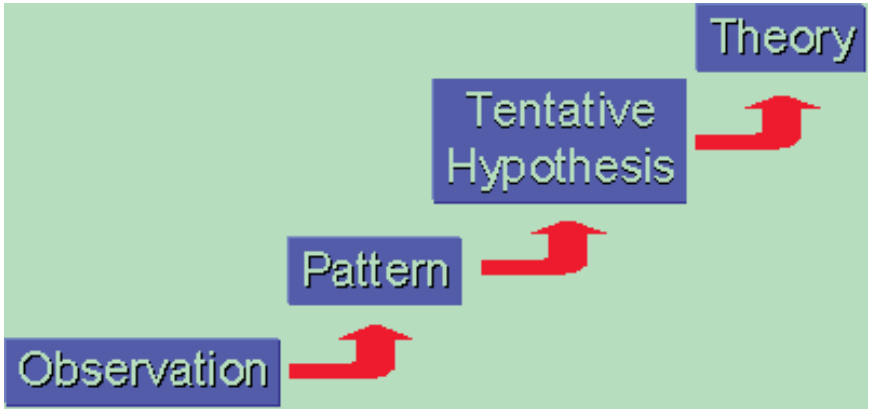
  Suppose we have a storage bin that contains 10 million flower seeds which we know will each produce either white or red flowers. The information which we want is: How many of these 10 million seeds will produce whit flowers? The only way in which we can be absolutely sre that this question is answered correctly is to plant every seed and observe the number producing white flowers . . .

Types of Statistical Inference

Deductive Inference



Inductive Inference



Basic Probability Revisited: Events

- A **sample space** Ω: the totality of possible outcomes of a conceptual experiment of interest ("universe")
- An **event space** $\mathcal{A}$: a set of subsets of Ω.

  The event space $\mathcal{A}$ is assumed to be a (Boolean) algebra (explaining the use of the symbol $\mathcal{A}$), meaning that the collection of events $\mathcal{A}$ satisfies the following properties:
  - The *universum* $\Omega \in \mathcal{A}$
  - If $A \in \mathcal{A}$ then $\Omega - A = \overline{A} \in \mathcal{A}$
  - If A_1 and $A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$
- An **event** is any collection (subset) of outcomes contained in the sample space. An event is *simple* if it consists of exactly one outcome and *compound* if it consists of more than one outcome

Probability

- Probability is a measure of uncertainty about the occurrence of events
- Two definitions of probability:
  - Classical definition (also referred to as "a priori probability")
  - Relative frequency definition

Probability

- **Classical definition** If a random experiment can result in n mutually exclusive and equally likely outcomes and if n_A of these outcomes have an attribute A, then the probability of A is the fraction n_A/n.
- Limitations:
  - The definition of probability must be modified somehow when the total number of possible outcomes is infinite [*e.g. draw from positive integers*]
  - Suppose that we toss a coin known to be biased in favor of heads [*What is the probability of a head?*]
  - Suppose notions of symmetry and equally likely do not apply? [*What is the probability that a male will die before the age of 60?*]

Probability

- **Relative frequency definition** Assuming that a random experiment is performed a large number of times, say n, then for any event A let n_A be the number of occurrences of A in the n trials and define the ratio n_A/n as the relative frequency of A. The limiting value of the relative frequency is a probability measure of A.
- Intuitive interpretation:
  - The probability of A is the limit of the relative frequency of A, as the number of experiments (see later: sample size) n goes to infinity.
  - "Long run relative frequency"

Probability Function

The **probability function** $P(.)$ is a set function having domain $\mathcal{A}$ and counterdomain the interval $[0, 1]$. Probability functions allow to compute the probability of certain "events" and satisfy the defining properties or axioms:

- $P(A) \geq 0$ for all $A \in \mathcal{A}$
- $P(\Omega) = 1$
- If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events in $\mathcal{A}$ (i.e., $A_i \cap A_j = \phi$ for $i \neq j; i, j = 1, 2, \ldots$) and if $A_1 \cup A_2 \cup \ldots = \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Probability Function

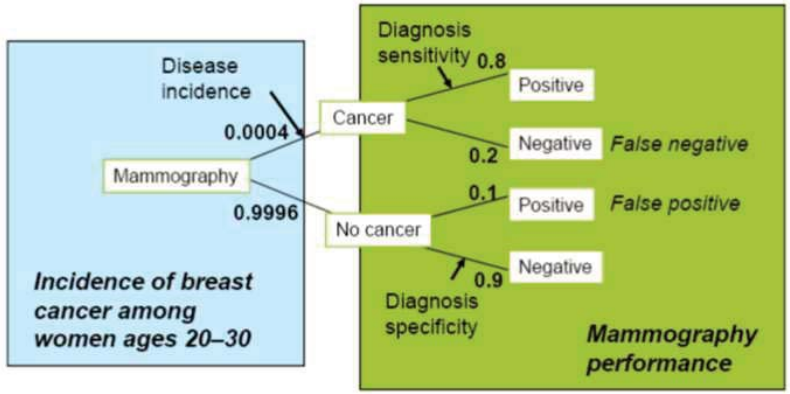
Some remarks:

- The previous definition of a probability function is mathematical one, motivated by the defintions of classical and relative frequency probabilities.
- It tells us which set functions can be called probability functions.
- It does not tell us what value the probability function $P(.)$ assigns to a given event A.
- We will have to model our random experiment in some way in order to obtain values for the probability of events.

Problem Solving Steps in Probability

- **Step 1**: Find the sample space

  [When the sample space is not too large, it is feasible to use tree diagrams, as in the breast cancer example below, to capture the sample space]



- **Step 2**: Define events of interest
- **Step 3**: Assign outcome probabilities
- **Step 4**: Compute event probabilities

Basic Probability Rules

- $P(A^c) = 1 - P(A)$
- **Addition rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Mutually exclusive** events are events which cannot occur at the same time: $P(AB) = P(A \cap B) = 0$
- Probability of A occuring given that B has occurred, $P(A|B) = P(AB)/P(B)$
- **Multiplicative rule**:

$$\begin{aligned} P(AB) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

Independent Events

- A and B are **independent events** if the occurrence of one event does not affect the probability of the other event.
- Relying on $P(A|B) = P(AB)/P(B)$, if A and B are independent then:
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
  - $P(AB) = P(A)P(B)$

Example 1: Positive Test for Disease (I)

- Suppose one in every 10,000 people in Ireland suffer from AIDS
- There is a test for HIV/AIDS which is 95% accurate.
- You are not feeling well and you go to hospital where your Physician tests you.
- He says you are positive for AIDS and tells you that you have 18 months to live.

How should you react?

Example 1: Positive Test for Disease (II)

- Let D be the event that you have AIDS
- Let T be the event that you test positive for AIDS
- $P(D) = 0.0001$
- $P(T|D) = 0.95$
- $P(D|T) =?$

Example 1: Positive Test for Disease (III)

$$
\begin{aligned}
P(D|T) &= \frac{P(D \cap T)}{P(T)} \\
&= \frac{P(T|D)P(D)}{P(\{T \cap D\} \cup \{T \cap D^c\})} \text{ (denominator: theorem of total probabilities)} \\
&= \frac{P(T|D)P(D)}{P(T \cap D) + P(T \cap D^c)} \\
&= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\
&= \frac{(0.95)(0.0001)}{(0.95)(0.0001) + (0.05)(0.9999)} \\
&= 0.001897
\end{aligned}
$$

Bayesian Odds

- On occasion when there are two events, say A and B, whose comparative **posterior probabilities** are of interest, it may be more advantageous to consider the ratios, i.e.:

$$\frac{p(A|C)}{p(B|C)} = \frac{p(C|A)}{p(C|B)} \cdot \frac{p(A)}{p(B)}.$$

- Ward Edwards gives a simple example where this comes in handy: There are two bags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue chips. Flip a fair coin to determine which one of the bags to use. Chips are drawn with replacement. In 12 samples, 8 red and 4 blue chips showed up.

What is the probability that it was the predominantly red bag?

Bayesian Odds

- Let A be the event of selecting the first bag. Let B be the event of selecting the second bag. Finally, let C be the result of the experiment, i.e., drawing 8 red and 4 blue chips from the selected bag. Clearly,

$$p(C|A) = \left(\frac{7}{10}\right)^8 \left(\frac{3}{10}\right)^4 \tag{1}$$

$$p(C|B) = \left(\frac{7}{10}\right)^4 \left(\frac{3}{10}\right)^8 \tag{2}$$

  so that $\frac{p(C|A)}{p(C|B)} = \left(\frac{7}{3}\right)^4 \approx 29.642$.
- Now, $p(A) = p(B) = 0.5$, implying that $\frac{p(A|C)}{p(B|C)} = \frac{p(C|A)}{p(C|B)} \times 1 = 29.642$.
- From $p(A|C) + p(B|C) = 1$, it then follows that $\frac{p(A|C)}{1-p(A|C)} = 29.642$ (this is **an odds** ...) and

$$p(A|C) \approx \frac{29.642}{1 + 29.642} = \frac{29.642}{30.642} \approx 0.967$$

Odds

- **Odds** are just an alternative way of expressing the likelihood of an event such as catching the flu. Probability is the expected number of flu patients divided by the total number of patients. Odds would be the expected number of flu patients divided by the expected number of non-flu patients.
- During the flu season, a medical doctor might see ten patients in a day. One would have the flu and the other nine would have something else.
  - So the probability of the flu in your patient pool would be one out of ten.
  - The odds would be one to nine.
- It is easy to convert a probability into an odds, and vice versa.

$$odds = probability \; / \; (1\text{-}probability)$$

$$probability = odds \; / \; (1 + odds)$$

Example 2: Monty Hall Problem (I)

- Game Show
- 3 doors
- 1 Car & 2 Goats
- You pick a door - e.g. #1
- Host knows what's behind all the doors and he opens another door, say #3, and shows you a goat
- He then asks if you want to stick with your original choice #1, or change to door #2?

Would you change doors?

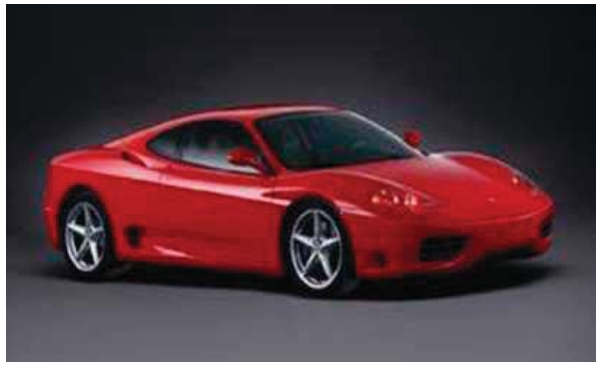
Example 2: Monty Hall Problem (II)

- At the beginning the sample space (set of all outcomes) is $\{CGG, GCG, GGC\}$
- Pick a door, e.g., #1
- 1 in 3 chance of winning
- The host then shows you a goat.
- So now $\{CGG, GCG, GGC\}$
- Hence, switching is the best option! The second door has a 2/3 of winning.

Example 2: Monty Hall Problem (III)

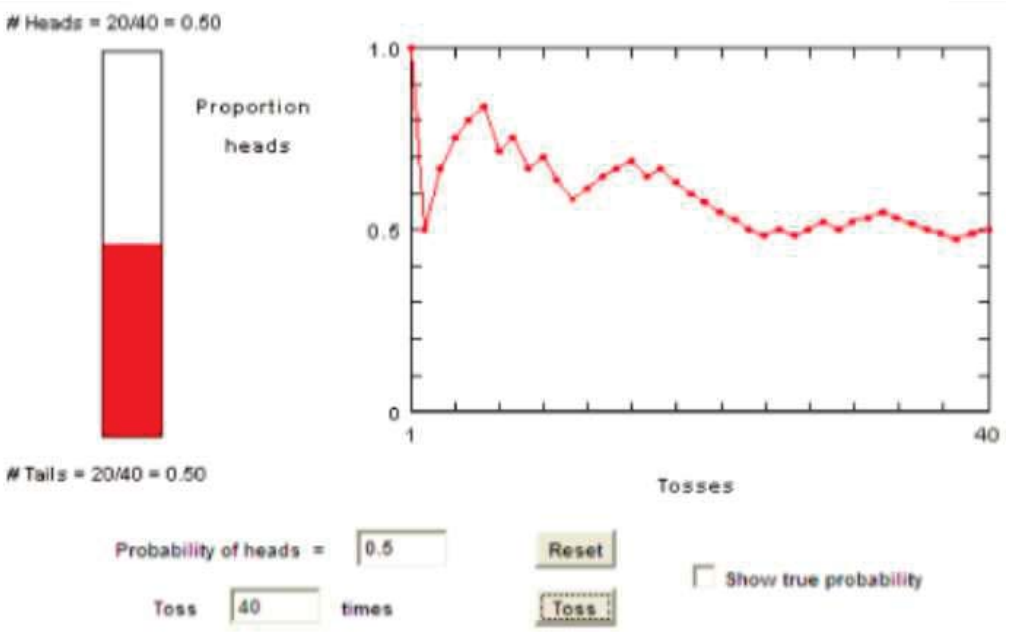
If your are not convinced, image a game with 100 doors . . .

- 1 Ferrari, 99 goats
- Pick a door
- Host opens 98 of the 99 other doors
- Do you stick with your original choice? (1/100 probability?)
- Or do you switch to the unopened door? (99/100 probability?)



Example 3: Tossing a Fair Coin (I)

Oh no .... not again !!!!



Example 2: Tossing a Fair Coin (II)

Misconception 1:

- "There is no reason to assume at any point that a change of luck is warranted based on prior trials (flips), because every outcome observed will always have been as likely as the other outcomes that were not observed for that particular trial, given a fair coin. "
- Assume a fair 16-sided die, where a win is defined as rolling a 1. Assume a player is given 16 rolls to obtain at least one win [1-P(rolling no ones)].
- The probability of having at least 1 win in the 16 rolls is: $1 - (\frac{15}{16})^{16} = 64.39\%$
- Assume that the first roll was a loss, then the probability of having at least 1 win is $1 - (\frac{15}{16})^{15} = 62.02\%$
- The previous losses in no way contribute to the results of the remaining attempts, but there are fewer remaining attempts to gain a win, which results in a lower probability of obtaining it.

Example 3: Tossing a Fair Coin (III)

Misconception 2:

Quote by William Feller (1957):

"...It is usual to read into the law of large numbers things which it definitely does not imply. If Peter and Paul toss a perfect coin 10,000 times, it is customary to expect that Peter will be in the lead roughly half the time. This is not true. In a large number of different coin-tossing games it is reasonable to expect that any fixed moment heads will be in the lead in roughly half of all cases. But it is quiet likely that the player who ends at the winning side has been in the lead for practically the whole duration of the game. "

Example 3: Tossing a Fair Coin (IV)

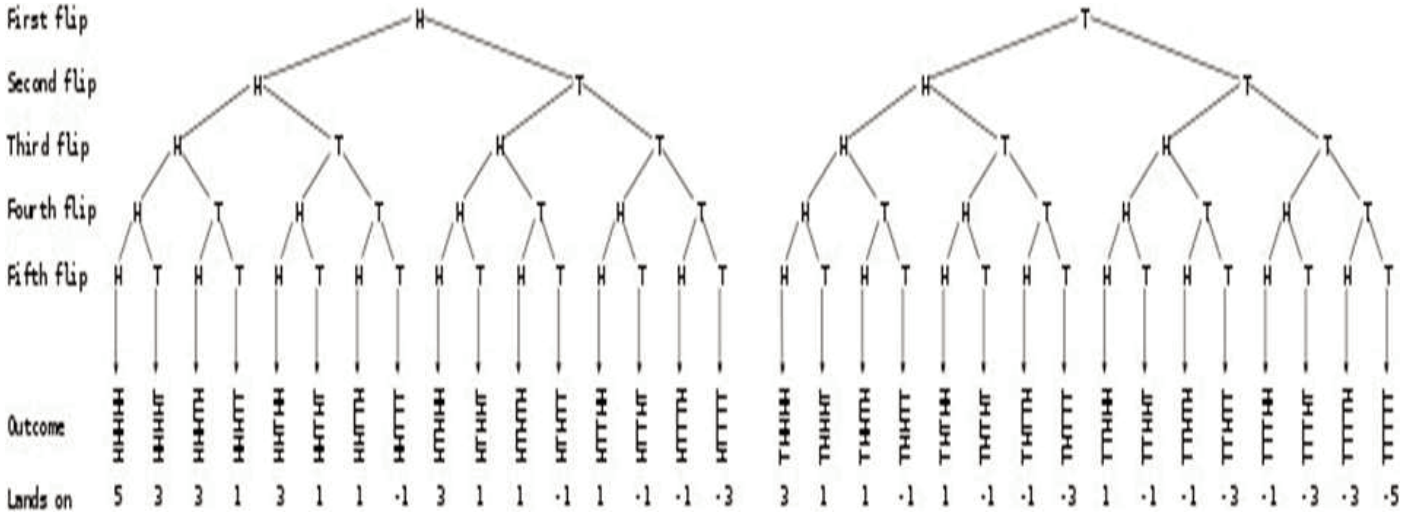
- When Peter and Paul toss a coin 10,000 times each, $N = 10,000$ and $2N$ coins are tossed in total.
- As N increases, the chances that there are equal numbers of heads and tails among the $2N$ tosses increases.
- So by the Law of Large Numbers:

$$\lim_{n \to \infty} P(\#H = \#T) = 1$$

- In the limit, as N tends to infinity, the probability of matching numbers of heads and tails approaches 1.

Example 3: Tossing a Fair Coin (V)

- One of the key ideas in probability is to study not just events but *processes*, which evolve in time and are driven by forces with a random element. [*Probability* - *Statistics* - *Stochastic Processess*]
- A *one-dimensional random walk* can be illustrated as follows: A marker is placed at zero on the number line and a fair coin is flipped. If it lands on heads, the marker is moved one unit to the right. If it lands on tails, the marker is moved one unit to the left.



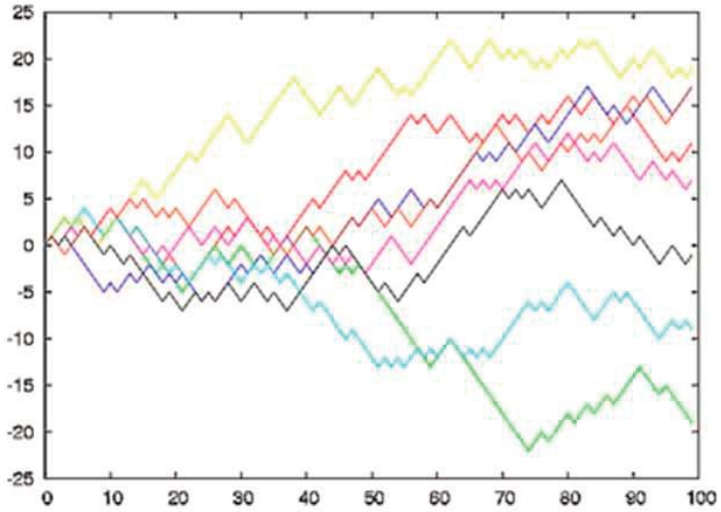
Example 3: Tossing a Fair Coin (VI)

- Take independent random variables Z_i, $i = 1, \ldots, n$ where each variable is either 1 or -1 with a 50% probability for either value, and set $S_0 = 0$ and $S_n = \sum_{i=1}^{n} Z_i$. The series is called the simple random walk on the integer numbers.
- This series of 1's and -1's gives the distance walked, if each part of the walk is of length 1.
- The expectation S_n is 0. That is, the mean of all coin flips approaches zero as the number of flips increase. This also follows by the finite additivity property of expectations:

$$\mathrm{E}(S_n) = \sum_{i=1}^{n} \mathrm{E}(Z_i) = 0.$$

Example 3: Tossing a Fair Coin (VII)

How many times will a random walk cross the zero line?



Example 3: Tossing a Fair Coin (VIII)

- The following, perhaps surprising, theorem is the answer: for any random walk in one dimension, every point in the domain will almost surely be crossed an infinite number of times. [In two dimensions, this is equivalent to the statement that any line will be crossed an infinite number of times.] This problem has many names: the "level-crossing problem", the "recurrence problem" or the "gambler's ruin" problem.

Problem Solving Steps in Statistics

- **Step 1**: Observation

  The first step of the scientific method is to make an observation regarding some event or characteristic of the world. This observation should lead to a question regarding the event or characteristic.

- **Step 2**: Ask a question

  The scientific method starts when you ask a question about something that you observe:

  How ? What ? When ? Who ? Which ? Why ? Where ?

  In order for the scientific method to answer the question it must be about something that you can measure, preferably with a number.

- **Step 3**: Construct a statistical hypothesis

  A hypothesis is an educated guess about how things work:

  If [I do this], then [this] will happen.

Problem Solving Steps in Statistics

- **Step 3**: Construct a statistical hypothesis

  . . . You must state your hypothesis in a way that you can easily measure and so that you are able to answer your original question.

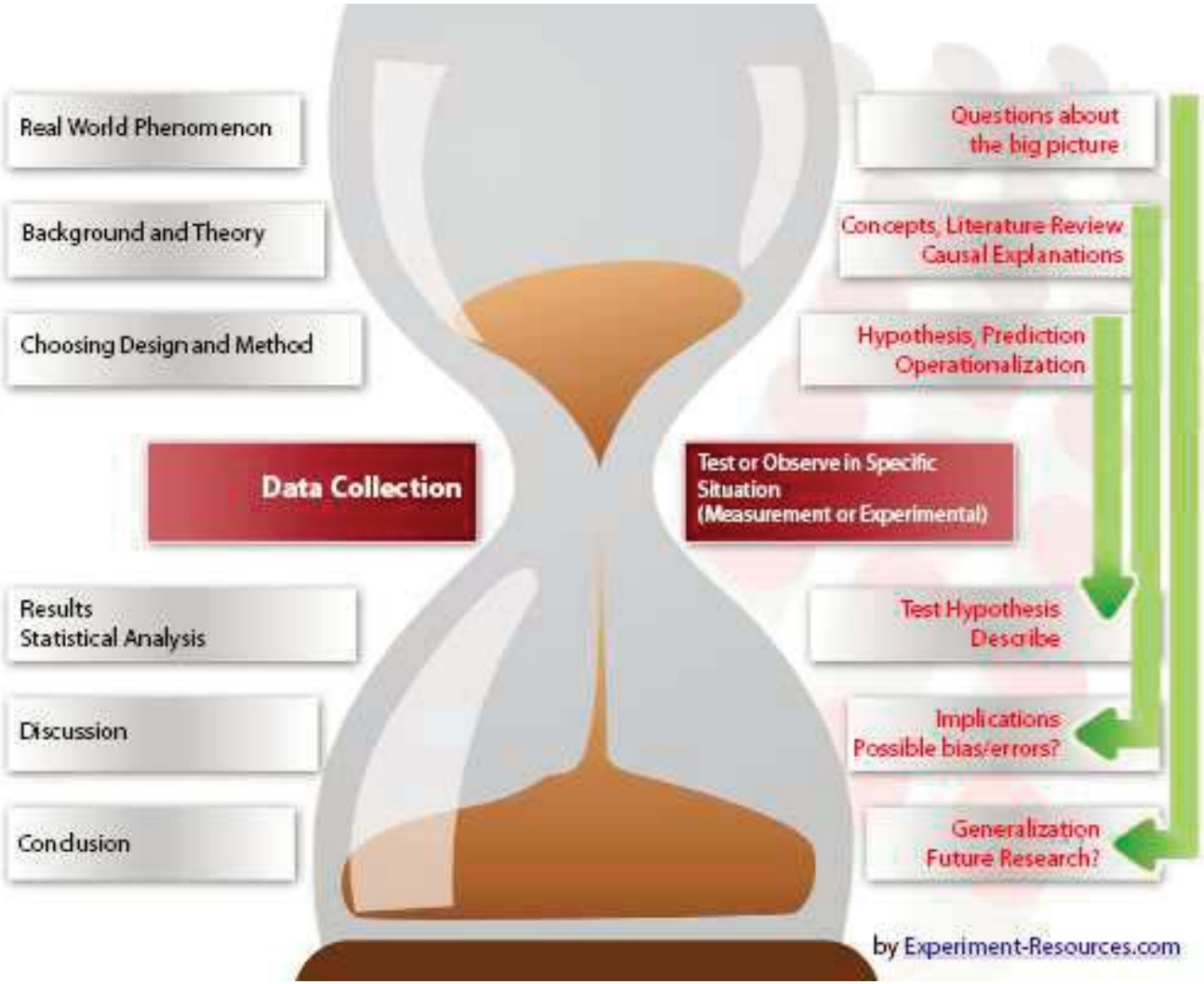
- **Step 4**: Test your hypothesis with data

  Collect an appropriate sample, check the quality of your data and carry out the test. It must be a fair test. All conditions are preferentially the same for each factor in your experiment except the one factor you are testing.

- **Step 5**: Draw conclusions

  These may be directly based on the formal hypothesis testing in Step 4, but may also involve making "predictions", with an assessment of how "reliable" these predictions are.

Problem Solving Steps in Statistics



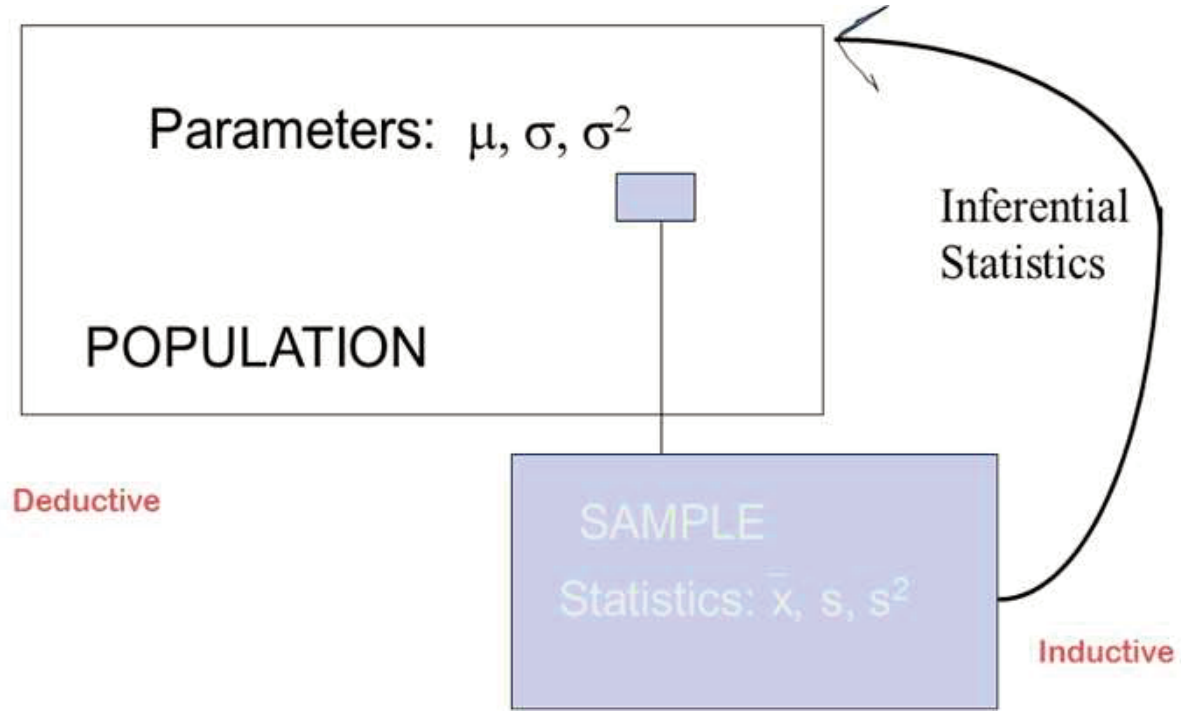
Samples and populations: Trying to Understand the True State of Affairs

- The world just happens to be a certain way, regardless of how we view it.
- The phrase "true state of affairs" refers to the real nature of any phenomenon of interest.
- In statistics, the **true state of affairs** refers to some quantitative property of a "population". Numeric properties of populations (such as their means, standard deviations, and sizes) are called "parameters". Parameters of a population (say, its mean and standard deviation) are based on each and every element in that population.
- Thus, for the scientist who uses inferential statistics, "population parameters" represent the true state of affairs.

Trying to Understand the True State of Affairs

- We seldom know the true state of affairs. The process of inferential statistics consists of making use of the data we do have (observed data) to make inferences about population parameters.
- Unfortunately, the true state of affairs is also dependent on all of the data we don't have (unobserved data). Nevertheless, an important aspect of "sample data" is that they are actual elements from an underlying population. In this way, sample data are "representatives" of the population that gave rise to them. This implies that sample data can be used to estimate population parameters.
- Therefore, as we have seen before, inferential statistics (both estimating and testing components) involve inductive reasoning: "from specific towards more general"

Parameters and statistics



Samples and Populations

- Since sample data are only representatives, they are not expected to be perfect estimators. Consider that we necessarily lose information about a book when we only read a book review. Similarly, we lack information about a population when we only have access to a subset of that population.
- It would be useful to have some measure of how "reliable" (or representative) our sample data really are. What is the probability of making an error?
- Obviously, in order to get a better handle on how representative our data are, we must first consider the sampling process itself: we must first study "how to generate samples from populations", before we can learn to generalize from samples to populations
- It is in this context that the importance of random and independent sampling begins to emerge.

True state of affairs + Chance = Sample data

- Some elements (say, "heights") in a population are more frequent than others. These more frequent elements are thus over-represented in the population compared to less common elements (e.g., the heights of very short and very tall individuals).
- The laws of chance tell us that it is always possible to randomly select any element in a population, no matter how rare (or under-represented) that element may be in the population. If the element exists, then it can be sampled, plain and simple.
- However, the laws of probability tell us that rare elements are not expected to be sampled often, given that there are more numerous elements in that same population. It is the more numerous (or more frequent) elements that tend to be sampled each time a random and independent sample is obtained from the population.

Some Definitions

- **Target population** The totality of elements which are under discussion and about which information is desired will be called the target population.
- **Random sample** Let the random variables $X_1, X_2, \ldots, X_n$ have a joint density $f_{X_1,X_2,\ldots,X_n}$ that factors as

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n),$$

where $f(.)$ is the (common) density of each X_i. Then $X_1, X_2, \ldots, X_n$ is defined to be a random sample of size n from a population with density $f(.)$

Some Definitions

- **Random variable** For a given probability space $(\Omega, \mathcal{A}, P(.))$, a random variable, denoted by X or $X(.)$, is a function with domain Ω and counterdomain the real line. The function X must be such that the set defined by $\{\omega : X(\omega) \leq r\}$ belongs to $\mathcal{A}$ for every real number r.
  - Ω: the sample space, this is the totality of possible outcomes of a conceptual experiment of interest
  - $\mathcal{A}$ is a set of subsets of Ω, called the event space.
- **Sampled population** Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density $f(.)$, then this population is called the sampled population.

Some Definitions - an aside

- **Cumulative distribution function** Any function $F(.)$ with domain the real line and counterdomain [0,1] satisfying the following 3 properties is defined to be a cumulative distribution function:
  - $F(-\infty) \equiv \lim_{x \to -\infty} F(x) = 0$ and $F(\infty) \equiv \lim_{x \to \infty} F(x) = 1$
  - $F(.)$ is a monotone, nondecreasing function [$F(a) \leq F(b)$ for any $a < b$]
  - $F(.)$ is continuous from the right; that is $\lim_{0 < h \to 0} F(x + h) = F(x)$

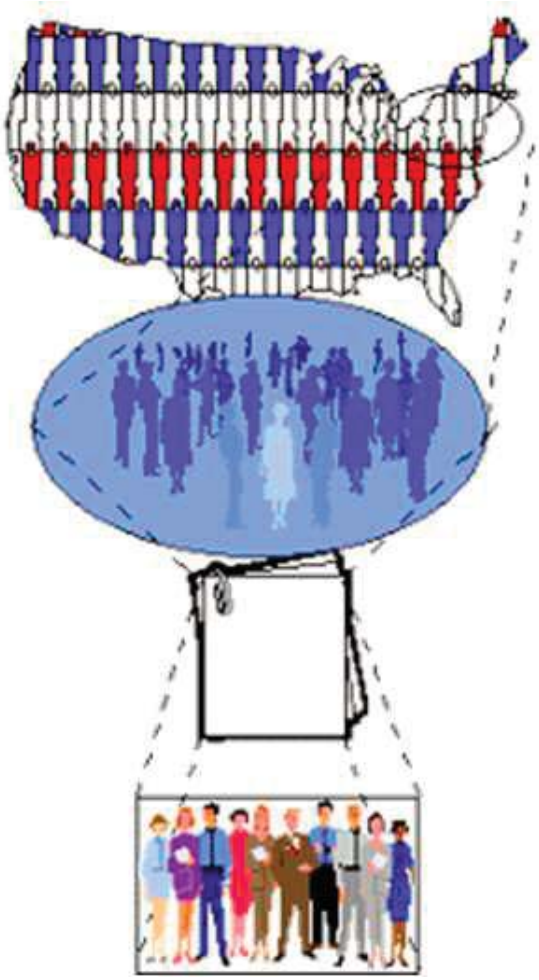
Example: 10 Million Flowers

- In the example of the 10 million flower seeds (syllabus), each seed is an element of the target population we wish to sample and will produce a white or red flower.
- Strictly speaking, there is not a numerical value associated with each element of the population. When we associate for instance number 1 with white and number 0 with red, then there is a numerical value associated with each element of the population, and we can discuss whether a particular sample is random or not.
- The random variable X_i is then 1 or 0 depending on whether the i-th seed sampled produces a white or red flower, $i = 1, \ldots, n$.
- If the sampling is performed in such a way that the random variables $X_1, X_2, \ldots, X_n$ are independent and have the same density (cfr i.i.d.), then, according to the previous definition of a random sample, the sample is random.
- We will see later in this course how we can look for signs against "randomness" or "independent" obvervations.

Sampling Frame

Who do you want to generalize to?

The Theoretical Population



What population can you get access to?

The Study Population

How can you get access to them?

The Sampling Frame

Who is in your study?

The Sample

Who are Those Angry Women?

In 1987, Shere Hite published a best-selling book called "Women and Love: A Cultural Revolution in Progress". This 7-year research project produced a controversial 922-page publication that summarized the results from a survey that was designed to examine how American women felt about their relationships with men. Hite mailed out 100,000 fifteen-page questionnaires to women who were members of a wide variety of organizations across the U.S. Questionnaires were actually sent to the leader of each organization. The leader was asked to distribute questionnaires to all members. Each questionnaire contained 127 open-ended questions with many parts and follow-ups. Part of Hite's directions read as follows: "Feel free to skip around and answer only those questions you choose." Approximately 4,500 questionnaires were returned . . .

- The population: all American women.
- The sample: the 4,500 women who responded.
- The sampling frame ?

Who are Those Angry Women?

- It is also easy to identify that the sampling unit was an American woman. So, the key question is "What is the sampling frame?"
- Most people think the sampling frame was the 100,000 women who received the questionnaires.

Is this answer correct?

Who are Those Angry Women?

- This answer is not correct because the sampling frame was the list from which the 100,000 who were sent the survey was obtained. In this instance, the sampling frame included all American women who had some affiliation with an organization.
- There is no statistical term to attach to the 100,000 women who received the questionnaire. However, if the response rate had been 100%, the sample would have been the 100,000 women who responded to the survey. You should also remember that ideally the sampling frame should include the entire population. If this is not possible, the sampling frame should appropriately represent the desired population. In this case, the sampling frame of all American women who were "affiliated with some organization" did not appropriately represent the population of all American women. This problem is called "selection bias".

Who are Those Angry Women?

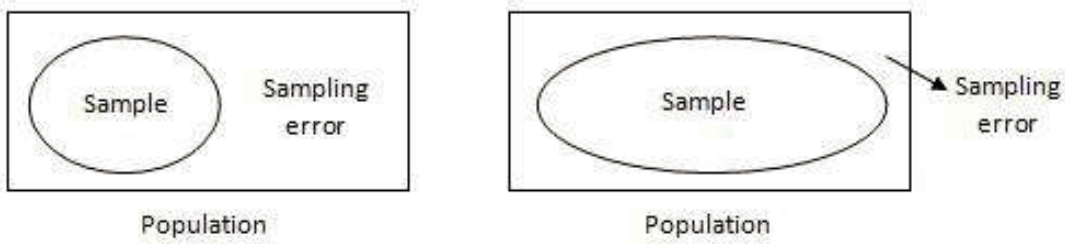
Three difficulties that are possible when samples are obtained for (for example) surveys:

1. Using the wrong sampling frame. This problem is also called **selection bias**.
2. Not reaching the individuals selected. Because the questionnaire was sent to leaders of organizations, there is no guarantee that these questionnaires actually reached the women who were supposed to be in the sample.
3. Getting "no response" or a "volunteer response." This problem can also be called **nonresponse bias**

Issues for Review

We already know the differences between theoretical / study population and sampling frame / sample, and that several difficulties may arise when samples are obtained.

- Does sampling work?
- What is sampling error and what can be done to reduce it?
- How much sampling error can be tolerated?



Issues for Review

- Is the sample size sufficient for "extrapolations"?
- How much precision is there in the "extrapolation"?
- How much larger should the sample be if more precision is desired?
- What is the most "optimal" sampling scheme?

Problem Setting

- Why do we sample?
  - Size of the population
  - Cost of obtaining elements
  - Convenience and accessibility of elements
- It is clear that in most practical applications, we need to generate a sample. However, at this point, it is less clear how we "best" select from an infinite number of observations we could possibly make ...

How do we Decide What to Observe?

- This decision should be a matter of deliberate choice rather than chance.
- Representativeness: a small sample of individuals from a population must contain essentially the same variations that exist in the population
- Limit to those characteristics that are relevant to the substantive interests of the study, not ALL aggregate characteristics

How do we Decide Who to Choose?

Basically two sampling strategies available:

- **Probability sampling** each member of the population has a certain probability to be selected into the sample
- **Non-probability sampling** members selected not according to logic of probability (or mathematical rules), but by other means (e.g. convenience, or access)

Non-probability Sampling

- Sometimes it is not possible to get the kind of information about populations that is required for probability sampling
- When the sampling frame is not known
- Complicates and limits statistical analyses: Non-probability sampling is well suited for exploratory research intended to generate new ideas that will be systematically tested later. However, if the goal is to learn about a large population, it is imperative to avoid judgment of non-probabalistic samples.
- Often well-suited for so-called **qualitative research** (i.e., a form of systematic empirical inquiry into meaning - Shank 2002), where distribution of characteristics is not important

Convenience Sampling

- Rely on available respondents
- Most convenient method
- Risky: exercise caution !!!

Purposive Sampling

- Rely on those subjects that fit a specific purpose.
- Select the sample on the basis of knowledge of the population: use your own knowledge, or use expert judges to identify candidates to select
- Typically used for very rare populations, such as deviant cases, or in market research.

Purposive Sampling

Expert

Heterogeneity

Quota

Snowball

Major Types of Probability Sampling