

Class 01 - Introduction to Statistics

JIN, ICK HOON

2009. 07. 06.

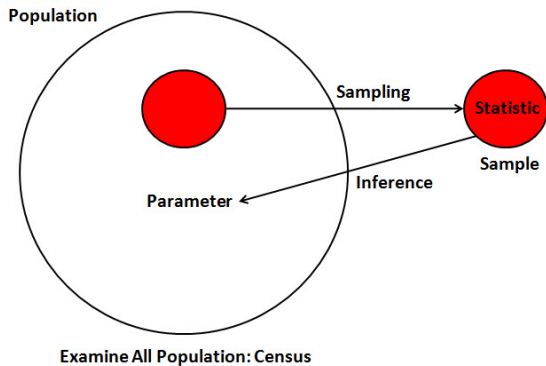
- Most people think statistics deals strictly with probabilities:
 - How likely am I to win the lottery?
 - Should I bet on this 'game'?
 - Will I get cancer?
- Statistics is actually divided into three areas: descriptive statistics, probability and inferential statistics(decision making). Statistics helps us make decisions in a way that will be cost effective, both in time and money.

Sort of Statistics

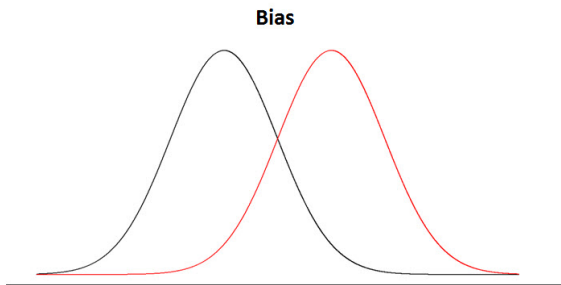
- 1 **Descriptive Statistics:** Gather information and organize it. Calculating summary numbers and drawing graphs.
- 2 **Probability:** Use past experience to give you an idea of what to expect. Determining the likelihoods.
- 3 **Inferential Statistics:** Make your decision based on which outcome you feel is most likely. Drawing a conclusion based on the data.

The group of interest is called the **population**. The characteristic of the population that we are interested in is called the population **parameter**. The same characteristic in a sample is called a **statistic**.

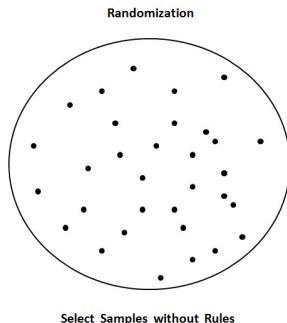
- Look at everyone (or thing) \Rightarrow **Census**, this is time consuming and costly.
- Use published information \Rightarrow Someone else may have already made a decision about what to print.
- Get a sample \Rightarrow Some subset of the population.



Some samples can be **biased**, they favor one side or they do not represent the entire population. To avoid this, statisticians use **randomization**. The simplest method is the **simple random sample**.



Random means not predictable or no discernable pattern, so we must use a randomization scheme rather than our (biased) judgement to get our sample.



So how can we use random samples to help us predict/determine what is going on in the population? This is where probability comes in.

- Probability** = How likely something is to occur?
- = The chance something will occur over time.
 - = The proportion of times something will occur in the long run.

But can we say when something is going to happen? or what is going to happen next?

- **Sample vs. Population:** Samples will vary, where the population is the 'whole truth'. The variability of the samples, looking at multiple samples, helps us determine how often one particular sample will occur. By studying sampling variability, we get a better idea of what the population looks like.
- (eg.) The coin toss: Toss the coin 10 times and count the number of heads. Each of these 'experiments' will give us a different number of heads, but most will be around 5. Rarely would we get 0 or 10. If we looked at the **distribution** of all of the experiments' outcomes, it would be centered at 5, the true population value. So, we can use samples to determine or estimate probabilities when the true (population) probabilities are unknown.

- Probability \Rightarrow Relative Frequency \Rightarrow Simulation.
- Often we don't know the population distribution, so we must estimate the probabilities using sample proportions. We use the 'relative frequency' approach to probability which says "the long-run proportion is the probability". Rather than perform an experiment numerous times, we use computer simulations which can run many experiments almost instantaneously. Once we have the 'true' probability distribution, we call it the population distribution. From this, we can calculate the population parameters.

- **Statistics:** A collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty.
- **Data:** A plural word referring to numbers or non-numerical labels collected from a set of entities.
- **Population:** The entire collection of individuals or measurements about which information is desired.
- **Sample:** A subset of the population selected for study. Takes less time and money, but we no longer know everything exactly, we have some error.
- **Data set:** A collection of observations on one or more variables.

- **Distribution:** A pattern of variation of a variable. Can be mathematical or graphical.
- **Variable:** A characteristic which changes when from observation to observation in a population.
- **Categorical (Qualitative) Variable:** A variable whose values cannot be interpreted as numbers. With categorical data, we can replace numbers with letters, symbols, etc.

Note: When analyzing categorical data, we will typically work with counts or percentages of objects.

- **Numerical (Quantitative) Variable:** Measurements or counts, values which have meanings as numbers.
 - **Discrete:** If the set of all possible values, when pictured on the number line, consists only of isolated points. The most common type of discrete variable we will encounter is a counting variable.
 - **Continuous:** If the set of all values, when pictured on the number line, consists of intervals.
- **Census:** Look at every object or individual in the population. We know everything exactly, but it takes a lot of time and money, may even be impossible to do if measurements are destructive.

- **Sampling Variability:** The differences between samples from the same population.
- **Random Sample:** A subset of the population selected so that every individual has a specified probability of being part of the sample.
- **Simple Random Sample:** A random subset of size n from a population where the subset is chosen in such a way that every possible subset of size n has the same chance of being selected as any other.
- **Sample Survey:** A method of gathering opinions and information from each individual included in the sample.

Possible problems: nonresponse bias and volunteer sample, both cause a nonrepresentative sample.

- **Observational Study:** Characteristics of the participants in the sample are observed.
- **Case-Control Study:** A study in which 'cases' having a particular condition are compared to 'controls' who do not. The idea is to compare them to see how they differ on the variable of interest.
- **Experiment:** A planned study in which the researcher collects data under different experimental conditions. The researcher controls one or more variables and measures the effect of changes in that variable on some outcome of interest.
- **Randomized Experiment:** A study in which the treatments are randomly assigned to participants. This may allow the researcher to determine cause and effect.

	Population - Parameter	Sample - Statistic
Sample Size	N	n
Mean	μ	\bar{x}
Median		M or \tilde{x}
Proportion	π or p	p or \hat{p}
Variance	σ^2	$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$
Standard Deviation	σ	s
Correlation Coefficient	ρ	r
Intercept	β_0	b_0
Slope	β_1	b_1

- **Standard Normal Distribution Notation:** $Z \sim N(0, 1^2)$
⇒ Z , a random variable, is distributed normally with mean, $\mu = 0$, variance, $\sigma^2 = 1^2$
- **Non-Standard Normal Distribution:** $X \sim N(\mu_x, \sigma_x^2)$
⇒ X , a random variable, is distributed normally with mean, μ_x , variance, σ_x^2 .
- **Sampling Distribution of the Sample Mean:** $\bar{X}_n \sim N(\mu_x, \sigma_x^2/n)$
- **Sampling Distribution of the Sample Proportion:** $p_n \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$