

Introduction

*What is Statistics?

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.

*How do you make a decision? Prior experience, a whim, an educated guess?

[1.] Gather information and organize it---**descriptive statistics**: calculating summary numbers and drawing graphs

[2.] Use past experience to give you an idea of what to expect---**probability**: determining the likelihoods

[3.] Make your decision based on which outcome you feel is most likely---**inferential statistics**: drawing a conclusion based on the data (the facts).

*What do you want to know about?

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**.

*How do you get the information?

look at everyone (or thing) → census, this is time consuming and costly

use published information → someone else may have already made a decision about what to print

get a **sample** → some subset of the population

*So how should you get your sample? Are all samples as good?

Some samples are **biased**, they favor one side or they don't represent the entire population. To avoid this, statisticians use **randomization**. The simplest method is the **simple random sample** (see the vocabulary list).

*What does random mean? How do we randomize?

Random means not predictable or no discernable pattern, so we must use a randomization scheme rather than our (biased) judgement to get our sample. Computers, using random number generators, provide this.

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

*What is a probability?

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 then say *when* something is going to happen? or what's going to happen next?

eg., toss a coin. the probability of a head = 0.5, so will you get 1/2 of a head? will you get a head 1/2 of the time?

*Can we always predict/determine the probability?

sample vs. **population**---samples will vary (why?), 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: Say we 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 => Relative Frequency => 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.