

Probability - Introduction

Chapter 3, part 1

Mary Lindstrom

(Adapted from notes provided by Professor Bret Larget)

January 27, 2004

Why Learn Probability?

- Some biological processes seem to be directly affected by chance outcomes. Examples include
 - formation of gametes
 - occurrence of genetic mutations
 - number of colonies that grow out on a plate
 - location of electrons in a atom
 - radioactive decay
- Statistical analysis of biological data assumes that variation not explained by measured variables is caused by chance.

Why Learn Probability?

- Chance might be used in the design of an experiment, such as the random allocation of treatments or random sampling of individuals.
- Probability is the language with which we express and interpret assessment of uncertainty in a statistical analysis.
- Probability comes up in everyday life — predicting the weather, lotteries or sports betting, strategies for card games, understanding risks of passing genetic diseases to children, assessing your own risks of diseases associated in part with genetic causes.
- Statistical analysis depends on modeling observed data as the realization of a random process.

Random Sampling

Most of the methods of statistical analysis we will use in this class are based on the assumption:

the individual units in the **sample** are sampled **at random** from the **population of interest**.

Random Sampling

- Taking a **simple random sample of size n** is equivalent to:
 1. Represent each individual in the population with a single ticket.
 2. Put the tickets into a box
 3. Mix the tickets **completely**
 4. Draw out n tickets **randomly** (without replacement)
- We will ignore for the present that sampling at random is almost impossible.
- How would you generate a random sample of size 1 ($n = 1$) from the population of numbers from 1 to 10?

Simple Random Sampling

What sampling is not simple random sampling?

- **Stratified random sampling** - Sampling is done in parts, e.g., a random sample taken in each of a number of geographical locations.
- **cluster sampling** - each first level experimental unit is sampled multiple times creating a second level of experimental units, e.g. measuring the same individual over time.

These are examples of random sampling processes that are not simple. Data analysis for these types of sampling strategies go beyond the scope of this course.

Simple Random Sampling

The defining characteristic of the process of simple random sampling is that every possible sample of size n has the same chance of being selected.

In particular, this means that

- (a) every individual has the same chance of being included in the sample.
- (b) members of the sample are chosen independently of each other.

Simple Random Sampling

Note that point (a) above is insufficient to define a simple random sample.

Every individual has the same chance of being included in the sample.

Consider sampling one couple at random from a set of ten couples. Each person would have a one in ten chance of being in the sample, but the sampling is not independent.

Possible samples of two people from the population who are not in a couple have no chance of being sampled while each couple has a one in ten chance of being sampled.

Using R to Take a Random Sample

Suppose that you have a set of individuals, numbered from 1 to 98, and that you wanted to sample ten of these. Here is some R code that will do just that.

```
> sample(1:98, 10)
[1] 17 79 37 32 57 94 12 27 52 93
```

In the `sample()` function, the first argument is the set from which to sample (in this case the integers from 1 to 98) and the second argument is the sample size.

In the output, the `[1]` is R's way of saying that that row of output begins with the first element.

The `random()` function samples **without** replacement by default.

Using R to Take a Random Sample

The same code executed again will result in a different random sample.

```
> sample(1:98, 10)
[1] 51 49 52 53 82 78 11 65 81 25
> sample(1:98, 10)
[1] 23  2 13  9 98 74 56 83 51 68
> sample(1:98, 10)
[1] 38 37 17 44 25 32 82 19 53 91
```

Samples and Populations

Statistical inference involves making statements about populations on the basis of analysis of sampled data.

Population: The group of individuals we are interested in

Sample: The group of individuals we actually measure.

Samples and Populations

The **Simple Random Sampling** model is useful because it allows precise description of the discrepancy between:

Statistical Estimates: numbers calculated from the sample, e.g. the sample mean.

Population Parameters: properties of the population, e.g. the population mean.

If we do the sampling correctly we can quantify this error.

Samples and Populations

When using Random Sampling it is important to ask:

- What is the **population**?
- How was the **sample** generated?

We must know these things to determine whether the sample is a **simple random sample** of the population of interest.

If we don't have a simple random sample, and we use statistical methods that assume we do, then our conclusions will be incorrect.

This type of error is called **sampling bias**. Sampling bias leads to incorrect conclusions because the sample is unrepresentative of the population in important ways.

Probability

- **Probability** is a numerical measure of the likelihood of an event.
- Probabilities are always **between 0 and 1**, inclusive.
- **Notation:** The probability of an event E is written $\Pr\{E\}$.

Probability - Examples

If a fair coin is tossed, the probability of a head is

$$\Pr\{\text{Heads}\} = 0.5 = 50\%$$

If bucket contains 23 white balls and 77 red balls and a ball is drawn at random, the probability that the drawn ball is white is

$$\Pr\{\text{white}\} = 23/100 = 0.23 = 23\%$$

If we throw 1 die the probability of a 5 or greater is

$$\Pr\{5 \text{ or } 6\} = 2/6 \doteq 0.33 = 33\%$$

Note on percents: Percent means into “per 100” or “divided by 100”. So $33\% = 0.33$

Probabilities in the real world

Coin-tossing

- It is reasonable to consider tossing a coin many times where each coin toss can be thought of as a repetition of the same basic chance operation.
- The probability of heads can be thought of as the long-run relative frequency of heads.

Super Bowl

- The outcome Sunday's Super Bowl is uncertain.
- Currently the odds of the Panthers winning are listed as approximately .65 .
- [What does this mean?](#)
- We can't use the "repeat many, many times argument" here since the odds would change if they played each other once a week over a long periods of time.

Probabilities in the real world

Evolution

- We can make the statement “all living individuals classified as mollusks have a common ancestor that is not an ancestor of any non-mollusks” .
- It is uncertain whether or not this statement is true.
- Does it make sense to place a probability on the truth of the statement?

Interpretations of Probability

There are two main approaches for connecting probability to the real world: **Frequentist** and **Bayesian**.

- The **Frequentist (or frequency) interpretation** of probability defines the probability of an event E as the proportion of times event E would occur if we repeated the experiment many, many times.

The textbook follows a frequency interpretation of probability.

- The **Bayesian (or subjective) interpretation** of probability defines probability as an individual's or group's degree of belief in the likelihood of an outcome given the existing data.

This may seem odd and unscientific but consider the following:

Interpreting Probabilities

Pr(heads)

- **Frequentist:** Use the frequency argument (out of many, many coin flips what proportion will be heads)?
- **Bayesian:** Use our belief about how likely the coin is to turn up heads.

Interpreting Probabilities

Pr(Panthers win Super Bowl on Sunday)

- **Frequentist:** Use the “many possible worlds” argument which goes something like:
 - No you can’t really repeat the Super Bowl many, many times.
 - But, you can imagine many copies of our world where many Super Bowls were being played and they would have different outcomes.
 - The probability of the Panthers winning is the proportions of the “worlds” where the Panthers win.
- **Bayesian:** Use our belief about how likely the Panthers are to win given the data we have (exactly what the bookies are doing).

Interpreting Probabilities

Pr(mollusks have a common ancestor that is not an ancestor of any non-mollusks)

- **Frequentist:**
 - There is no probability involved. Either the statement is true or it is not.
 - Reformulate problem: Ask what is the likelihood of the data we have observed (archaeological, biological) assuming the statement is true?
- **Bayesian:** Use our belief about how likely the statement is to be true based on the information we have about the history of mollusks.