Probability and Statistical Inference

Probability is a tool for quantifying and working with uncertainty. Through probability theory, we can define the set of possible outcomes, and talk about the relative frequency of these outcomes (distribution). If the possible outcomes are numerical, we can define the expected outcome (ie the average), and the expected variation around the outcome. Summary statistics associated with the range of outcomes.

Let’s talk about some common distributions we encounter in biological research:

Normal distribution: family of distributions. Completely characterised by 2 parameters – mean and SD (variance).

Binomial distribution: family of distributions. Arises when counting # of successful outcomes out of n independent trials. Completely characterised by 2 parameters: n = number of trials, and p = probability of success.

Anyone who studies biology recognise that the living world is not deterministic. But even though we observe variation, we recognise that this variation follows the laws of probability. That is, it may be possible to ascribe a data generating process that is consistent with the data we observe.

As scientists, we observe the world. We run experiments, which we hope to learn from. Statistics is a tool that can help us learn about the world by putting it into a probabilistic framework.

By a probabilistic framework, I mean a data generating process or statistical model. Like our examples in Lecture 2, data generating processes have two components: a functional component that describes mathematical relationships, and a random component that describes the probabilistic component (expected variation or uncertainty).

The data generating process that we ascribe to the data may not be the “true” model. It’s impossible to know that true generating process. But we hope that it will be a useful model. Useful: it generates data that is consistent with our past, present and future observations.

So the challenge in statistics is to nominate a data generating process that is consistent with our observations. Here are some desirable properties:

* Our model should NOT be too complicated; or overly specific about the **ad hoc parts** of our data. It should be generalise our observations.
* Our model should not be overly simple; it should inform us about the main patterns in the data.

You can imagine that starting from data and inferring the data generating process is a task fraught with difficulties. In some sense, we never KNOW anything for certain; we can at best make educated guesses. Statistics allows us to quantify our uncertainty using probability theory. Statisticians develop methodologies and algorithms to help us make these educated guesses.

## Permutation tests

### An algorithm for assessing group differences

A field trial was planted to compare a seed lot (i.e. a batch of seeds) derived from a seed orchard (SO) with one collected from a routine plantation (P). Eight plots were planted for each seed lot, and these were thinned when the young trees were seven years of age. Tree diameters at breast height (dbh) were measured at 15 years.

Let’s import and visualise the data

```{r}

seed<- read.csv(“seed orchard data.csv”)

str(seed)

ggplot(seed, aes(x=seedlot, y=dbh, colour=seedlot) +

geom\_boxplot()

```

Let’s calculate the difference in mean dbh:

```{r}

SO <-seed$seedlot == “SO”

mean\_diff <- mean(seed$dbh[SO]) – mean(seed$dbh[!SO])

Statistical inference is a principled way of offering evidence in favour of a hypothesis. We consider two competing models. Model I: dbh is independent of the seed batch type, and Model II: dbh depends upon seed batch type. We are looking for evidence in favour of Model II, or equivalently evidence against Model 1.

If dbh is independent of seed batch, then re-assigning the seed batch type to a tree and computing the mean difference between groups should give us a mean difference similar to the observed mean difference.

Here is the algorithm:

1. compute the group mean difference from the data (observed statistic).
2. randomly re-assign seed batch labels to each tree and re-compute group mean difference.
3. Repeat Step 2 10,000 times.
4. Plot histogram of 10,000 +1 mean differences, and assess how consistent the observed statistic is relative to the distribution of the 10,000 values.

```{r}

diff2<-vector()

for (I in 1:1e4)

## Exercise. Rare bird populations

ANU researchers have been monitoring an orange bellied parrot population since 2000. Orange bellied parrots are a critically endangered parrot species that breeds in Tasmania. Population estimates are listed in obp\_pop.csv.

```{r}

obp <- read.csv(“obp\_pop.csv”)

str(obp)

ggplot(obp, aes(x=year, y= pop) ) + geom\_point()

```

What pattern do you observe in the data? Is it reasonable to conclude that the population is declining? How could you describe the population trend over time?

We may not wish to sound the alarm about the declining population unless we feel we have strong evidence in favour of this hypothesis.

There are many approached we can take. Carry out a permutation test to assess this hypothesis.