Intro to probability and statistical inference

The goal of statistics is to draw conclusions about a population using a sample of that population. Another way of thinking about stats is that it is a means to learn something from the data.

Samples, probability

To learn something about a population we draw samples randomly from that population. By random, we mean that all values of our variable have an equal probability of being in our sample. The range of possible outcomes in our experiment is referred to as the sample space. For instance, if we draw a sample from a deck of cards and are interested in the suit of these cards (hearts, spades, clubs, and diamonds), the sample space is the only four possible outcomes possible (hearts, spades, clubs, and diamonds).

If we repeat an experiment many times, we expect there to be some differences in the outcomes, and we can begin to assign probabilities to these occurrences. The values that are more frequent in our population should begin to occur more frequently in our sample. Probability is the foundation of statistics. For discrete characters, such as the suit of cards, the summation of these probabilities across all possible discrete outcomes will be equal one. This is the same case for continuous characters, except we define the probability (or density) of outcomes using a probability density function (PDF) and integrate over a range of values. Most of the examples and materials used in this course will be for continuous variables.

A random variable is any variable whose value depends on an unknown event. Any variable that can be tied to a PDF is a random variable. This is a bit of an over-simplification, and there are more formal, rigorous definitions of random variables, but for the sake of brevity that definition should suffice for this course. Any outcome or result that is dependant on a random variable will also be random. The example below will hopefully clarify these concepts.

Estimation and statistical inference

All probability models depend on parameters. These parameters are properties of the population ($n \rightarrow \infty$) and are unknown. We are interested in inferences on these parameters, so we randomly draw samples from the population and make inferences on these parameters. This process is called estimation. There are two ways to do estimation: via maximum likelihood (ML) or using Bayes theorem. The latter will not be covered in the course.

The goal of maximum likelihood is to estimate unknown parameters by maximizing a likelihood function. To do this we define the probability of the data in terms of the parameters (θ) and estimate θ so that it maximizes the probability of the observed data. This estimate of θ ($\hat{\theta}$) is the maximum likelihood estimate of the parameter. So what we need is a likelihood function. The probability function and the likelihood function are given by the same equation. The difference between the two is that the PDF is a function of the data when the parameters are fixed. The likelihood function is a function of the parameters when the data is fixed. We can maximize the likelihood function pretty easily if a closed form solution exists (see below), or if one does not exist we can also use numerical optimization. The latter is also not so difficult with the use of computers.

MLE of the mean

Say we're interested in finding the average height of a wheat plant in a farmers field. We assume that plant height will be normally distributed with a mean μ and variance σ^2 (i.e. $y \sim N(\mu, \sigma^2)$). We known the the PDF for a normal distribution is given by , so we can use this function to estimate the population mean using ML.

The PDF for a normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

and the likelihood function is

$$L(\mu, \sigma^2 | \mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

.

There is a closed form solution that we can use to find the MLE of the mean. What we'll do is take the partial derivative of the likelihood function, set it to zero and solve for μ. The log likelihood is easier to work with and is

$$\begin{aligned}
LL(\mu, \sigma^2|\mathbf{x}) &= \sum_{i=1}^{n} log(f(x_i)) \\
&= \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right) \\
&= \sum_{i=1}^{n}\left(log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log\left(\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right)\right) \\
&= \sum_{i=1}^{n}\left(log(\sqrt{2\pi\sigma^2}) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n}\left(-\frac{1}{2}log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{N}{2}log(2\pi\sigma^2) + \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

.

Now let's take the partial derivative to the log likelihood function with respect to μ

$$\begin{aligned}
\frac{\partial LL}{\partial\mu} &= \frac{\partial}{\partial\mu}\left(-\frac{N}{2}log(2\pi\sigma^2)\right) + \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= 0 + \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)
\end{aligned}$$

Now set this above to zero and solve for μ

$$\begin{aligned}
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\
&= \sum_{i=1}^{n}(x_i - \mu) \\
&= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu \\
&= \sum_{i=1}^{n} x_i - N\mu \\
N\mu &= \sum_{i=1}^{n} x_i \\
\mu &= \frac{\sum_{i=1}^{n} x_i}{N}
\end{aligned}$$

This seems like a lot of work to derive something that is common knowledge, but the approach outlined above can be used to derive ML estimates for other, less intuitive parameters.

Sampling distributions

The sample average will vary depending on the set of observations sampled from the population. This, if we repeatedly sample from the population and compute the mean, we can describe the sampling distribution of this statistic. The sampling distribution is a theoretical probability distribution of the values of some sample statistic that would occur if we were to draw all possible samples of a given size from a population. This can give us insight into how common or rare it is to observe a given value for that statistic.

The sampling distributing of the mean has several interesting properties. If we sample from a normally distributed population, the distribution of the sample means ($\bar{x}$) is also normal. If we draw a large enough sample from a non-normal population the sampling distribution of the mean is also normal. The mean of the sampling distribution is the true mean. The sampling distribution for the mean is $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ where n is the sample size. Thus, the sampling distribution of the mean is different for different sample sizes.

Central limit theorem

There are several variants of the central limit theorem (CLT), but we will not discuss these variants.The general idea of the CLT is that the normalized sum (or average) of independent random variables will tend to follow a normal distribution. This will be the case even if each random variable is not normally distributed.

x_i is the ith draw of a random variable from a uniform distribution (non-normal distribution; note that X_i is still a random variable). We create another random variable (S_n) that represents this average over the draws $S_n = \frac{x_1+X_2+x_3+\cdots+x_i}{n}$. The CLT states that for a large enough n, S_n will begin to follow a normal distribution. The R code below demonstrates this. It is important to note that the average computed below is not normalized.

This code will average over n draws where $n = 1, 2, 15, 30$ and will repeat the process 10,000 times.

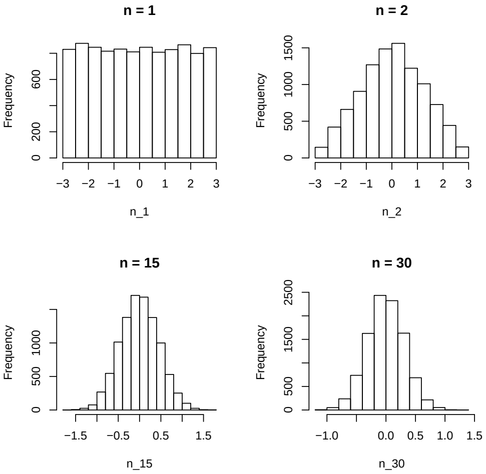
```r
n_1 <- NULL
for(i in 1:10000){
  n_1 <- c(n_1, mean(runif(1, min = -3, max = 3)))
}

n_2 <- NULL
for(i in 1:10000){
  n_2 <- c(n_2, mean(runif(2, min = -3, max = 3)))
}

n_15 <- NULL
for(i in 1:10000){
  n_15 <- c(n_15, mean(runif(15, min = -3, max = 3)))
}

n_30 <- NULL
for(i in 1:10000){
  n_30 <- c(n_30, mean(runif(30, min = -3, max = 3)))
}
```

The histogram of the random variables is shown below.



This behavior is also the case for random variables that are independent, but do not have identical distributions (see the Lyapunov CLT; this will not be explained here). There are many reasons why the CLT is important. In linear regression (or variants of linear regression) we often assume the error term is normally distributed. The reason for this is that the CLT states that the sum of a bunch of random variables will be approximately normally distributed. So returning to our plant height example, we can ask why there is a random deviation from the "true" value for plant height. If you ask someone with a good background in biology they will give you many reasons why plant height might deviate from this true value. Each one of these potential causes is a random variable, and because we cannot possible measure or account for all of these causes the CLT says that we can just throw them into the error term and assume they come from a normal distribution.