Introduction to Statistical Ideas and Methods

Sampling Distributions

Sampling Distributions

To goal of statistics is to make conclusions based on the incomplete or noisy information that we have in our data. The process of doing this is called **statistical inference**.

Here is a model for statistical inference:
We have our real world that we experience and measure and in our real world we have our observed data.
Underlying this, is a theoretical world which explains how the real world works. It includes scientific and statistical models that describe the probability distributions for random measurements, and observations of these random measurements are our data. The theoretical world could also be a population and our data could be measured on a sample from the population. In that case the randomness in the data is due to the randomness in the sampling process.

Either way we have parameters in the theoretical world which are features of the models or features of the population in the theoretical world. Typically, we do not know the values of these parameters. But they are fixed, they are features of nature or the population and therefore they are not random.

Inferential statistics is often about estimating parameters by some **statistic** calculated from our real world data.
In real life, we only get one sample or set of data. But there are many values it could be, depending on the random measurements that give us our data. Therefore statistic is a random quantity.
In order to make inferences based on one sample or set of data, we need to think about the behaviour of all of the possible sample data-sets that we could have got.

It is often a goal to understand the typical or central value of measurements, so the mean is a common parameter that we would like to make inferences about. We call our theoretical world mean μ and we will estimate it by the sample mean $\bar{x}$.
Suppose we observe random measurements from a sample of size 9:

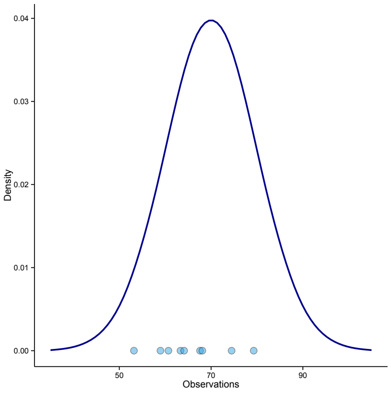

Figure 1: A sample of size 9

Then we calculate the average or sample mean of these 9 values, the purple vertical bar is where this average is located:

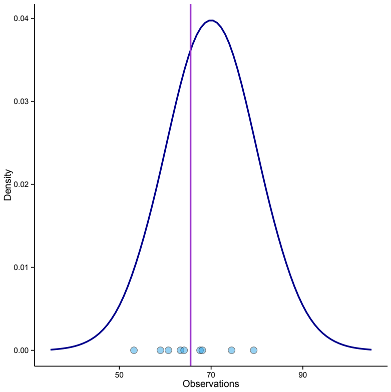

Figure 2: A sample of size 9 with average (vertical bar)

This average is a statistic which is calculated from our data. We call it $\bar{x}$ and use it to estimate μ, so $\bar{x}$ is an estimator of μ.

Of course if we had taken another random sample of data, we would get a different value for statistic. For example these are two possible samples with corresponding sample means:

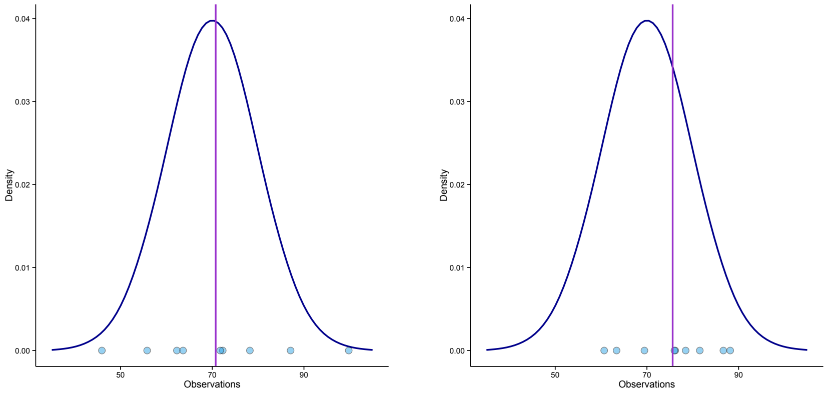

Figure 3: Two random samples of size 9 with averages (vertical bars)

The purple probability distribution shows the behaviour of all of the different sample means that it is possible to get:

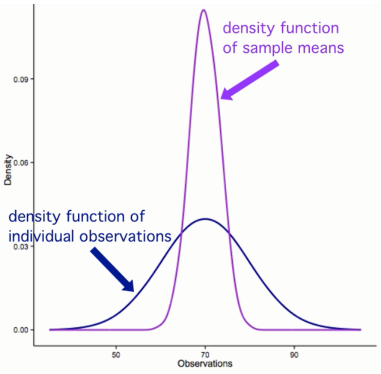

Figure 4: Purple: Distribution of sample means of size 9. Blue: Distribution of individual observations.

We have already investigated how this probability distribution for the sample means compares to the probability distribution of the individual observations. Knowing the probability distribution of the sample means is an important component of the process of statistical inference. This is called the **sampling distribution**.

Sampling distribution is the probability distribution of the possible values of an estimator.

We only observe one sample and get one sample mean, but if we make some assumptions about how the individual observations behave (if we make some assumptions about the probability distribution of the individual observations) then that tells us what the sampling distribution of the mean is.

For example if the individual observations have a Normal distribution with mean μ and standard deviation σ,

$$X \sim N(\mu, \sigma)$$

then the sample mean (sample size n) has a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

This is the sampling distribution of the mean.

The observed value of the mean varies from sample to sample of data. The variability is called the **sampling variability**.
In the case of the sampling distribution of the mean, the sampling variability is given by the standard deviation of the sampling distribution, which is $\sigma/\sqrt{n}$.

We cannot be sure that individual observations behave like observations from a Normal distribution, they could come from a uniform or skewed distribution or anything else.
But as long as our sample size is large enough, we have the Central Limit Theorem which tells us that the probability distribution of the means is approximately normal.

Note that the sampling distribution is in terms of quantities that we do not observe, in particular μ and σ. These parameters are properties of the probability distribution of the individual observations, and they exist in the theoretical world, so we do not know them. How to deal with this will be left for future work.
The most important is to understand what happens when a study is performed. We carry out the study, we get one set of data, and we calculate one sample mean. But the outcome of a single study can be considered as a random outcome from all of the possible studies that could have been carried out under the same conditions. The sampling distribution of the mean tells us how likely the various sample means that we could get in these possible studies are.

The other situation we have considered previously is a proportion.
When estimating a proportion, each individual observation is a Bernoulli random variable:

$$X \sim \text{Bernoulli}(p)$$

For the Bernoulli random variable, parameter p is the probability that it comes up a success.
We estimate p with $\hat{p}$ (proportion of times we get a success in our sample of size n).
Even though in practice we only get one sample and hence we observe only one value of $\hat{p}$, we know that the value of $\hat{p}$ would vary from sample to sample. We also know that for large n, the distribution of $\hat{p}$ is approximately normal with mean p and standard deviation equals

to $\sqrt{p(1-p)/n}$:

$$\hat{p} \mathrel{\dot{\sim}} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

This is the sampling distribution of a sample proportion.
Like we saw for the mean, the sampling distribution for $\hat{p}$ is in terms of p which we do not observe.

This type of reasoning allows us to develop statistical methods for many parameters. Starting with a presumed distribution function that describes the behaviour of the individual observations in our data, we can mathematically derive probability distribution functions for an average, or a proportion, or, other quantities, like an estimated standard deviation or, the slope of line of best fit. And knowing how these statistics behave will allow us to make inferences.