Sampling Distributions

The Distribution of a Sample Mean: Part 1

Imagine that we observe the value of a random measurement and suppose the probability distribution that describes the behaviour of the possible values of the measurement is a normal distribution. Then, although the value of our measurement is a random quantity, we know that it is more likely to be a value close to the mean, than from the tails of its normal distribution.

Now, suppose we observe 10 random measurements where the probability distribution of each measurement is normal. Then we take these measurements and average them together. *What values are we likely to see for this average? Compared to a single value, is it more likely to be close to the mean of the normal distribution, or further from the mean?*

In this document we explore how the probability distribution of an average of a number of measurements compares to the probability distribution of a single measurement in the case of the normal distribution.

Here is a Normal density function with a mean of 70 and a standard deviation of 10.

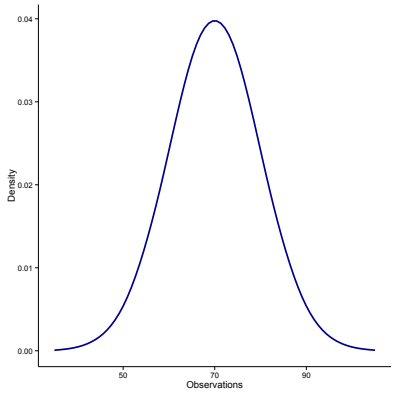

Figure 1: Normal probability density: mean=70, SD=10

Suppose this is the probability distribution that describes the behaviour of a measurement of some quantity. For example it can be the probability distribution of the marks on a standardized exam.

Next we observe a randomly generated observation from this distribution:

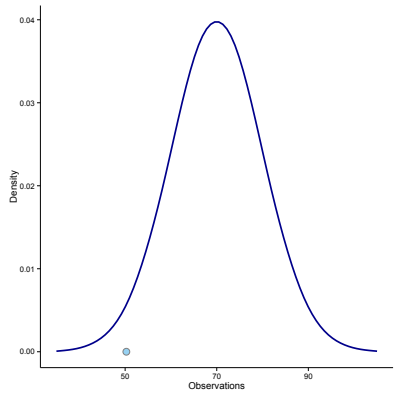

Figure 2: Randomly generated observation

It has value a little above 50. So this randomly chosen exam writer got a mark just over 50 on the exam. We can observe another random measurement from this distribution:

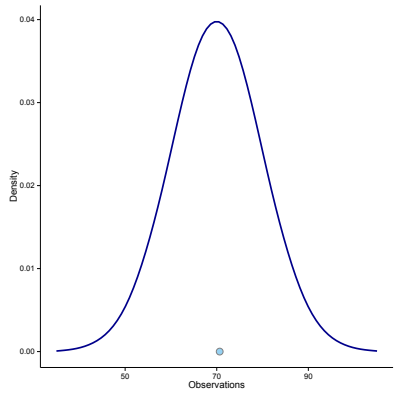

Figure 3: Randomly generated observation

This writer got above the mean of 70.
According to this probability distribution, about 95% of writers will get an exam mark between 50 and 90, with values closer to the mean more likely, and marks greater than 90 or less than 50 possible, but unlikely.

Now suppose we have 9 randomly chosen measurements following this probability distribution. Here is one random sample from all possible samples of 9 measurements:

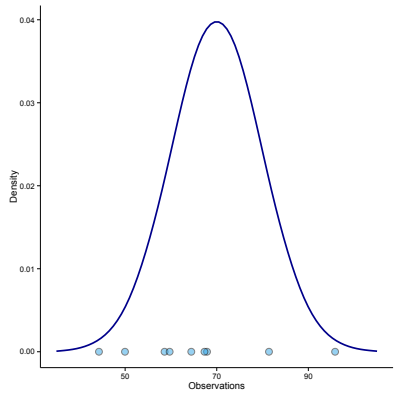

Figure 4: One random sample of 9 measurements

The measurements in our sample are distributed across the probability distribution, with more values close to the distribution mean of 70, although by chance we got a few more values below the mean than above the mean. Suppose we are interested in the average of these 9 measurements. It is 65.5. We can call this average the **sample mean**, since it is a mean, or average, from a sample of values from the probability distribution.
If we had another random sample of 9 measurements we would get a different 9 values, and consequently a different average.
If we repeat this process 10 times and calculate the sample mean (of 9 measurements) each time, we get the following random sample means: $65.5, 70.7, 66.4, 65.6, 73.1, 74.8, 74.3, 63.5, 66.8, 66.9$. Although the individual observations cover the range of possible values from our probability distribution (from 40 to 100), the sample means are always quite close to the mean of the normal probability distribution.

To generalize this, we simulated 2000 random samples of size 9 (from normal distribution with mean 70 and standard deviation of 10), found the average for each, and plotted the 2000 sample means in the following histogram:

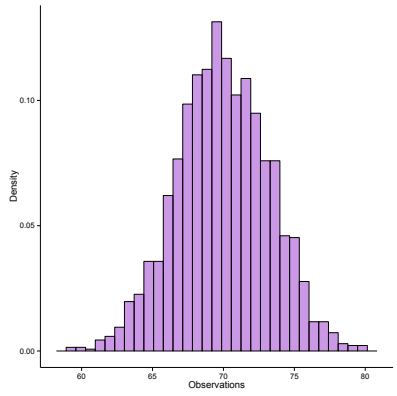

Figure 5: Density histogram of 2000 random sample means

The y-axis has been scaled so that the total area under the histogram is 1, like a probability density function.
There are 3 things to note about the values of the sample mean that we observe in this histogram:

1. The most common values are close to the mean of the probability distribution of the individual values (in our case 70).
2. The sample means range from about 60 to 80, with most of them between about 65 and 75. Compared to a range about 40 to 100 with most values between 50 and 90 for our individual values. So there is less variability in the sample means than there is in our individual values.
3. The histogram is symmetric, and bell-shaped, like a normal distribution.

We can smooth this histogram and get an estimate of the density function of the means:

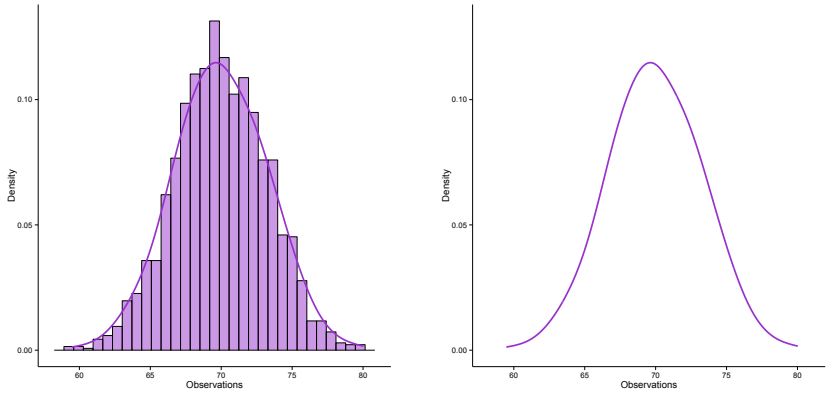

Figure 6: Left: Density histogram with smooth. Right: Estimated density curve

For a better comparison of the probability distribution of the individual measurements versus the probability distribution of the sample means of random samples of size 9, we plot both on the same scale:

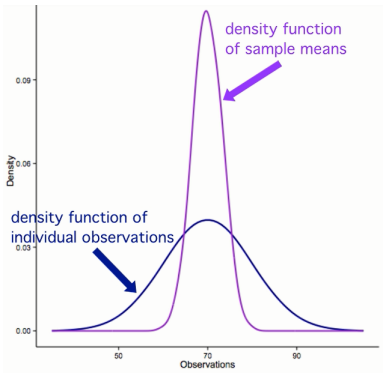

Figure 7: Density of individual measurement (blue) and density of sample means (purple)

Both probability distributions are normal, both normal distributions have the same mean, but the purple probability density function has less spread. That is, the standard deviation of the probability distribution of the sample means is smaller than the standard deviation of the probability distribution of the individual observations. This is consistent with our intuition, an average of several observations gives a better estimate than a single observation, where accuracy is captured by the spread in the distribution of values it could be.

Let's now introduce some notation.
As is usual for the normal distribution, we call the center of the distribution (or expected value) the Greek letter 'mu' (μ) and we call the standard deviation the Greek letter 'sigma' (σ). μ and σ are properties of the probability distribution and they are fixed (not random).

We call X_1 a random variable that follows this distribution.
We had 9 observations of such a random variable in our samples, so we have $X_1, X_2, , X_9$.
The notation for the sample mean (the average of these 9 observations) is X with a bar over it:

$$\bar{X} = \frac{X_1 + \cdots + X_9}{9}$$

Note that $\bar{X}$ is a random quantity. It varies, randomly, with each random sample that we might observe.

Previously it was shown that if we roll a die 100 times and calculate the average of those 100 rolls then:

$$E(\bar{X}) = \text{expected value of a single roll}$$

$$Var(\bar{X}) = \frac{\text{variance of a single roll}}{100}$$

We can take the square root of the variance to get the standard deviation:

$$SD(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\text{standard deviation of a single roll}}{\sqrt{100}}$$

So the average has the same mean as the probability distribution of a single measurement, but the variance decreases by dividing it by the number of measurements in our random sample.

Generally: for $X_1, , X_n$ independent random variables (n is the sample size) with

$$E(X_i) = \mu, \quad SD(X_i) = \sigma$$

and $\bar{X}$ is the average of the n observations. Then

$$E(\bar{X}) = \mu, \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Averages are often used to estimate the expected value (μ) of the underlying probability distribution. The fact that $E(\bar{X})$ equals to μ is a nice property of the $\bar{X}$. It tell us that an average is an **unbiased** estimator of the μ. So in the long run, with a large sample size, $\bar{X}$ will give us a value very close to what we want to estimate.

Summary:

1. When measurements are random values that follow a normal distribution, the probability distribution of sample means (the average of the data) is also a normal distribution.
2. The mean of the normal probability distribution of the sample means is the same as the mean of the probability distribution of the individual measurements.
3. The standard deviation of the probability distribution of the sample means is smaller than the standard deviation of the probability distribution of the individual observations.
   If there are n values in the random sample and σ is the standard deviation of the probability distribution of the individual observations, the standard deviation of the probability distribution of the sample means is $\sigma/\sqrt{n}$.