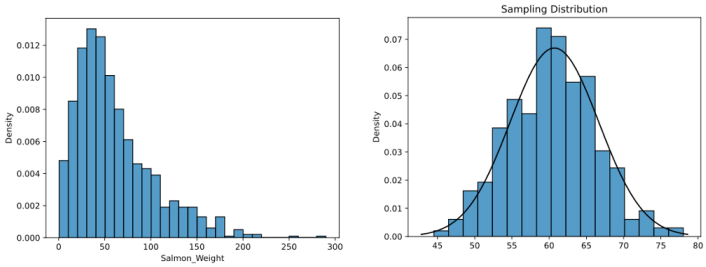
Sampling for Data Science

Central Limit Theorem

According to the Central Limit Theorem, the sampling distribution of the mean:

- is normally distributed
- has a mean equal to the population mean
- has standard deviation (also called standard error) equal to the population standard deviation divided by the square root of the sample size

In the plots provided, the left plot shows the population distribution of salmon weights, and the right plot shows the sampling distribution of the mean salmon weights.



Standard Error & Sample Size

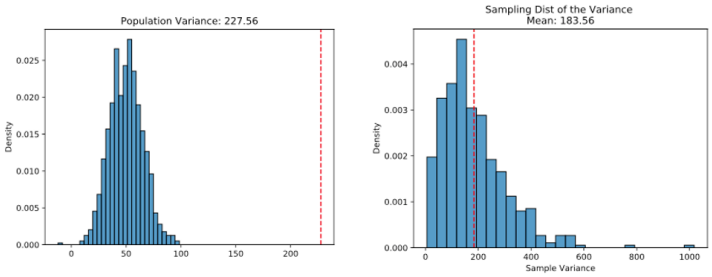
When you increase the sample size, the standard error of the mean decreases. This can be seen from the formula:

$$Standard\ Error = \frac{\text{Population Stand}}{\sqrt{\text{Sampl}}}$$

As sample size increases, the denominator increases while the numerator remains constant.

Biased Estimators

A *biased estimator* is a statistic such that the mean of that statistic's sampling distribution is not equal to the value of that statistic for the population.
Minimum is an example of a biased estimator because any particular sample minimum is likely to be larger than the population minimum. Variance is another example of a biased estimator, and this is shown in the provided plot.



CLT & CDF

If we want to know the probability that a sample from a population will have a mean in some specific range, we can:

1. Use the CLT to determine the mean and standard deviation of the sampling distribution of the mean
2. Use the cumulative density function of a normal distribution with that mean and standard deviation to calculate the probability

The code block given shows how to do this using Python.

```
# calculate standard error using population standard deviation and sample size
standard_error = std_dev / (samp_size**.5)
# use the cdf scipy method to calculate the probability of observing some value x or lower
stats.norm.cdf(x,mean,standard_error)
```

Central Limit Theorem Assumptions

The CLT holds true if:

- the population is normally distributed. OR
- if the population is skewed or otherwise not normally distributed, the sample size must be sufficiently large (n>30).

Since we often don't know the distribution of the population, it is safer to always make sure to have a sufficiently large sample size.

Standard Error

The standard deviation of a sampling distribution is also known as the *standard error of the estimate of a mean*. The standard error for a sample mean can be calculated with the following formula:

$$Standard\ Error = \frac{\text{Population Stand}}{\sqrt{\text{Sampl}}}$$