What *is* Statistics?

A *statistical procedure* consists of three components:

1. a way to collect data,
2. a sample size, and
3. a way to compute something of interest from the data.

Choices (1) and (2) together constitute a *sampling procedure*. When the "something of interest" is an estimate of a population parameter, choices (1)-(3) together constitute an *estimation procedure.*

The fundamental concept of statistics: Any numeric result of a statistical procedure can be viewed as a random variable, and anything of interest concerning the procedure corresponds to some characteristic of the corresponding random variables.

The Language of Estimation

"My estimate of ________ (a population parameter) is ________ (a sample statistic).

"Furthermore, the estimation procedure I used had a ________ (large) chance of yielding an estimate within ________ (a small amount) of the true value."

For estimating a population mean, using simple random sampling with replacement, a sample of size n, and the sample mean as the estimate:

population parameter:	μ
sample statistic:	$\bar{x}$
confidence:	95%
margin of error:	$1.96 \cdot s/\sqrt{n}$

Some Analytical Details

1. **Properties of the sample mean**: Consider first the case of sampling with replacement.

$$E[\overline{X}] = E[\frac{X_1 + X_2 + \ldots + X_n}{n}] = \frac{1}{n} \cdot E[X_1 + X_2 + \ldots + X_n]$$

$$= \frac{1}{n} \cdot (E[X_1] + E[X_2] + \ldots + E[X_n]) = \frac{1}{n} \cdot (\mu + \mu + \ldots + \mu) = \frac{1}{n} \cdot n\mu = \mu \ .$$

$$\mathrm{Var}[\overline{X}] = \mathrm{Var}[\frac{X_1 + X_2 + \ldots + X_n}{n}] = \frac{1}{n^2} \cdot \mathrm{Var}[X_1 + X_2 + \ldots + X_n]$$

$$=^* \frac{1}{n^2} \cdot (\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n]) = \frac{1}{n^2} \cdot (\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \ .$$

(* - since, for sampling *with* replacement, X_1, X_2,..., X_n are independent.)

Otherwise (i.e., for sampling *without* replacement),

$$= \frac{1}{n^2} \cdot (\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n] + 2\,\mathrm{Cov}[X_1, X_2] + 2\,\mathrm{Cov}[X_1, X_3] + \ldots + 2\,\mathrm{Cov}[X_{n-1}, X_n])$$

$$= \frac{1}{n^2} \cdot (n\sigma^2 + n(n-1)\,c) = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right),$$

since the covariance of any two distinct observations can be shown to be $c = -\sigma^2 / (N-1)$, where N is the size of the population.

And finally, in either case, if n is at least moderately large, $\overline{X}$ is roughly normally distributed. (Of course, if the underlying population is itself normal, $\overline{X}$ is normally distributed for *any* n.)

2. **Computing the sample variance**: The sample mean $\overline{X}$ is computed by averaging the sample observations. But the sample variance s^2, an estimate of the population variance, is defined to be the sum of the squared deviations of all observations from the sample mean, divided by n-1 (instead of by n). Why?

Compare $\sum (x_i - \mu)^2 / n$ with $\sum (x_i - \overline{x})^2 / n$. The first is a legitimate estimate of σ^2, and the second will almost always (unless, by coincidence, μ is precisely equal to $\overline{x}$) be somewhat smaller. (Indeed, $\sum (x_i - t)^2$, viewed as a function of t, is minimized when $t = \overline{x}$.) To unbias the latter expression, we scale it up by a bit: It turns out that dividing by n-1 instead of n is just enough.

We frequently wish to make multiple related estimates from the same sample. When we do so, the various estimates will typically fit together a bit *too* well. In statistical lingo, each estimate "costs us a degree of freedom". We must adjust our calculations slightly to compensate for this loss.