Covariance and Correlation

We often want to understand the interrelationship between two random variables. One measure of the linear dependence between two random variables is their **covariance**: $cov(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)]$, which basically measures how well the random variables agree in their deviations about their means. It is difficult to compare covariances because of scale differences, so we usually also calculate their correlation, which is a scaled version of their covariance: $\rho = cov(y_1, y_2)/(\sigma_1 \sigma_2)$.

Estimation and Confidence Intervals

We use **sample statistics** as **estimators** of **population parameters**. We would like these estimators to be **unbiased** and have small **variance**. If we are comparing biased estimators, we can compare their **mean squared error** (**MSE**). When n, N, and N-n are all large, then the sample mean will tend to have a **normal distribution**.

We generally want to make statements like $P(| \theta\text{-HAT} - \theta | <= B) = 1 - \alpha$, to quantify the amount of error of estimation. This leads to a **confidence interval** (θ-HAT - B, θ-HAT + B) with **confidence coefficient** $1 - \alpha$. Generally B is set to be 2(STD- θ-HAT) (2 times the standard error of the estimator), in which case Tchebysheff's theorem states that we achieve at least 75% confidence. If θ-HAT is normal, we have 95% confidence.

Simple Random Sampling

If a sample of size n is drawn from a population of size N such that every possible sample of size n is equally likely, the sampling procedure is called **simple random sampling**.

How to draw a simple random sample:

Estimation of the population mean and total: