Class 1 and 2: Frequentist statiscal Inference

Frequentist Statistical Inference

We want to learn about a population. A population could be the inhabitants of a country, bacteria in an environment or a keyboard produced by a company. We want to know a specific characteristic of that population; for example, the average height in a country, the median size of the bacteria, or, average the number of key presses it takes for the keyboard to fail.

It would be unfeasible to calculate this measure directly from the population. For example, to pinpoit the exact average time for a product to fail, we would have repeatidly press every keyboard in production to find the number. This is why we want to take a sample from the population and use the data in the best way so we can learn about the specific characteristic.

For example, we want to know the average height of the population(μ). We will use a sample to ***guess*** the value of μ (i.e. the parameter of interest).

Caveat: We can *never* know if the guess we get from our sample is close to our true parameter. All we can do is to *hope* that our guess gives us a fair estimate.

But that *hope* is backed by a whole theory behind it. That's why we use inference everyday. We can be *wrong* but we had made the best *guess* we could.

Population and Sampling Distributions

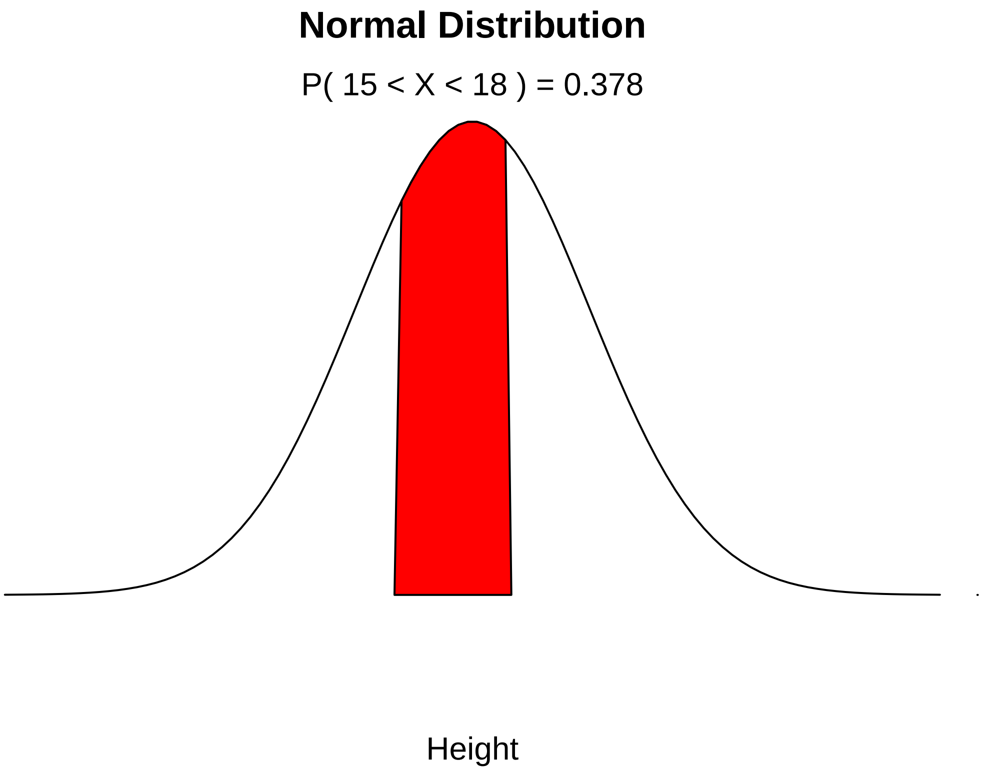
Our population has some distribution. This distribution tells us what can happen when I extract one subject from the population. For example, suppose that our population height (in dm = 10cm) is distributed

$$X \sim \mathcal{N}(17, \sigma^2 = 9)$$

.

Then suppose that I want to know what is the probaility that one individual will be between 15dm and 19 dm $ P(15= lb & x <= ub lines(x, hx) polygon(c(lb,x[i],ub), c(0,hx[i],0), col="red")

area <- pnorm(ub, mean, sd) - pnorm(lb, mean, sd) result <- paste("P(",lb,"< X <",ub,") =", signif(area, digits=3)) mtext(result,3) axis(1, at=seq(40, 160, 20), pos=0) "'



Statistics and their properties

Consider that we take a sample $(x_1, x_2, ..., x_n)$ from the population X.

Definition: A *statistic* $T = T(x_1, x_2, ..., x_n)$ is a function of the random sample $(x_1, x_2, ..., x_n)$. As such, it is itself a random variable with a distribution PDF $f_T(t)$ or CDF $F_T(t)$.

This statistic will give be helpful to unconver an estimation of the parameter (θ, or in the case of the sample mean μ) of the population that we are interested in.

We expect that the statistic has several properties:

1. Unbiasedness: Expected value of the statistic should be the parameter in question.

$$\mathbb{E}(T) = \theta$$

For example, in the case of $T(\bar{x}) = \sum_{i \in (1,..n)} x_i$, we get that $E(T) = \mu$.

2, Consistency: As n becomes larger (we sample more subjects), the statistic will converge in probability to θ, i.e. If we consider a series of estimators indexed by n, we get

$$\{T_n\} \xrightarrow{p} \theta$$

What is convergence in probability? Definition:

$$\{T_n\} \xrightarrow{p} \theta$$

For example, if we go back to the avearage case, we can index the statistic by n and get:

$$T_j(\bar{x}) = \sum_{i \in (1,..n_j)} x_i$$

Definition: 3. Variance get smaller

$$\forall \epsilon > 0, \forall \theta \in \Theta, lim_{n \to \infty} P(|T_n - \theta| \geq \epsilon) = 0 \iff lim_{n \to \infty} P(|T_n - \theta| \leq \epsilon) = 1$$