Statistical Inference

Week 1

Probability

Assigns a number between 0 and 1 to a event to give a sense of the chance of the event.
Probability as become out default model for apparently random phenomena.
Our eventual goal is to use probability models, or formal mechanism for correlating our data to a population.

- Rules that probability must follow
  - The probability that nothing occurs is 0
  - the probability that something occurs is 1
  - The probability of something is 1 minus the probability that opposite occurs
  - The probability of at least one of two (or more) things that can simultaneously (mutually exclusive) is the sum of the respective probability.
  - If the event A implies the occurrence of the event b, then the probability of A occurring is less than the probability than B occurs
  - For any two events the probability that at least on occurs is the sum of theirs probability minus the intersection.

Probability Mass Function (PMF)

A **probability mass function**(pmf) evaluated a value correspond to the probability that a random variable takes a value. To be a valid pmf a function(p) must satisfy:

- it always be larger than or equal to 0
- the sum of the possible values that the random variable can take has to add up to one

Probability density function (PDF)

A **probability density function**(pdf) is associated with a continuous random variable. To be a valid pdf a function(p) must satisfy:

- Is must be larger than or equal to zero everywhere
- the total under it must be one

Areas under the pdfs correspond to probability for that random variable.

Note: A pdf is a statement about the population of "matter in study" in the case. Not a statement about the data itself. We use that to evaluate statements about the population probability. Probability is for a population quantity

Cumulative distribution function(CDF) and Survival Function

The **cumulative distribution function**(cdf) of a random variable x return the probability that a random variable is less that or equal to the value x

$$F(x) = P(X \leq \mathbf{x})$$

The **survival function** of a variable x is defined as the probability that the random variable is greater than value x

$$F(x) = P(X \geq \mathbf{x})$$

Quantiles

If you where the 95th on a ex am, you know that 95% of the people scored worse that you and 5% scored better.

Definition: The : α^{st} quantile of a distribution function F is the point x_α so that

$$F(x) = \alpha$$

Summary

- You might wandering at this point `'I've heard of a median before it where I had a sample quantity ->  its a estimator. In this course we gonna build up not only estimators but also the target of a estimation or the estimand.'` didn't require integration where is the data.
- We're referring to the population quantities. therefore , the median being discussed is the population median
- a probability model connect the data to the population using assumption
- therefore the median we're discussing is the estimand, for the sample will be the estimation
- the formal process of statistical inference linking your sample to a population

Conditional Probability

Let B an event so that $P(B > 0)$

$P(A|B)$ be an event so that $P(B) > 0$

$P(A|B) = P(A \cap B)/P(B)$ -> if the event A and B are independent this will be to the probability of A, B doesn't affect anything.

Bayes Rule

Bayes rules allows us to reverse the role of the conditioning set and the set that we want the probability of.

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A|B) * P(B) + P(A|B^c) * P(B^c)}$$

Example Diagnostic Tests

Let + and - be the events that result of a diagnostic test is positive and negative respectively. Let D and D^c be the event that subject of the test has or does not have the disease respectively.

Sensitive: $P(+|D)$

Specificity: $P(-|D^c)$

Positive Predictive Value = $P(D|+)$

Negative Predictive Value = $P(D^c|-)$

Prevalence of the disease = $P(D)$

Specificity of a 98.5% | Sensitivity of 99.7% | population with a 0.001 ($P(D)$)

$$P(D|+) = \frac{P(+|D) * P(D)}{P(+|D) * P(D) + P(+|D^c) * P(D^c)} = \frac{P(+|D) * P(D)}{P(+|D) * P(D) + (1 - P(-|D^C)) * (1 - P(D))} = \frac{0.997 * 0.001}{0.997 * 0.001 + 0.15 * 0.999} = 0.062(6\%)$$

Independence

IID random variables

Random variables are *iid* if they are **independent** and **identically distributed**:

- Independent: statistically irrelated from one to another
- Identically distributed: all having been draw from the same population

Expected Values

The process of making conclusions about populations from noisy data that was draw from it

- the mean (expect value) is a characterization of its center
- Variance and standard deviation are characterization about of about spread it is.

Summarize

- Expected values are properties of distributions
- The population mean is the center of mass of population
- The sample mean is the center mass of the observed data
- The sample mean is an estimate of the population mean
- The sample mean is unbiased
    - The population mean of its distributions is the mean that is trying to estimate
- The more data that goes into the sample mean the more concentrate its density/mass function is around the population mean.