M378K: November 18th, 2024.

Sufficient Statistics.

Consider 2 election candidates: (A) & (B) ←

Goal: To "predict" if A wins.

Let the random sample be $Y_1, \dots, Y_n$

Scenarios to book-keeping:

- $T_1 = (Y_1, \dots, Y_n)$
- $T_2 = Y_1 + Y_2 + \dots + Y_n$
- $T_3 = 1_{\{\bar{Y} > 0.5\}}$

Formally,
- the conditional dist'n of $T_1 = (Y_1, \dots, Y_n)$ given $T_2 = Y_1 + \dots + Y_n$ does not depend on p ←
- the conditional dist'n of $T_1 = (Y_1, \dots, Y_n)$ given $T_3 = 1_{\{\bar{Y} > 0.5\}}$ DOES DEPEND on p.

In general, consider two r.v.s Y and T,
the "conditional dist'n" of Y given T
are these probabilities $\mathbb{P}[Y = y \mid T = t]$ ←

In the discrete case, take $Y = (Y_1, \dots, Y_n)$, we write

$$p_{Y_1,\dots,Y_n,T}(y_1, \dots, y_n, t) = \mathbb{P}[Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, T = t]$$

Def'n. The conditional joint pmf for t such that $\mathbb{P}[T = t] > 0$ is

$$p_{Y_1,\dots,Y_n|T}(y_1, \dots, y_n \mid t) = \mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n \mid T = t]$$

$$= \frac{\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n, T = t]}{\mathbb{P}[T = t]}$$

Analogously, in the continuous case:

$$f_{Y_1,\dots,Y_n|T}(y_1,\dots,y_n|t) = \frac{f_{Y_1,\dots,Y_n,T}(y_1,\dots,y_n,t)}{f_T(t)}$$

Def'n. A statistic T of a random sample $(Y_1,\dots,Y_n)$ is said to be sufficient for an unknown parameter Θ if the conditional dist'n of the sample $(Y_1,\dots,Y_n)$ given T does not depend on Θ.

Example. Claim T_2 is sufficient for T_1.

$$p_{Y_1,\dots,Y_n|T}(y_1,\dots,y_n|t) = \frac{\mathbb{P}[Y_1=y_1,\dots,Y_n=y_n,T=t]}{\mathbb{P}[T=t]} \quad \checkmark$$

1st: $y_1+\dots+y_n \neq t$ we get 0

2nd: $y_1+\dots+y_n = t$

$$\mathbb{P}[Y_1=y_1,\dots,Y_n=y_n,T=t] = \mathbb{P}[Y_1=y_1,\dots,Y_n=y_n]$$
$$= p^{y_1}(1-p)^{1-y_1}\,p^{y_2}(1-p)^{1-y_2}\cdots p^{y_n}(1-p)^{1-y_n}$$
$$= p^{\sum y_i}(1-p)^{n-\sum y_i}$$
$$= p^t(1-p)^{n-t}$$

$$\mathbb{P}[T=t] = \binom{n}{t}p^t(1-p)^{n-t}$$

$$p_{Y_1,\dots,Y_n|T}(y_1,\dots,y_n|t) = \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Theorem. The Fisher-Neyman Factorization Criterion.

Let $Y_1,\dots,Y_n$ be a random sample w/ the likelihood f'tion $L(\Theta;y_1,\dots,y_n)$

The statistic T is sufficient for Θ if and only if

L can be expressed as

$$L(\theta; y_1, \dots, y_n) = g(\theta, T(y_1, \dots, y_n)) \cdot h(y_1, \dots, y_n)$$

Example. **Bernoulli.**

$$L(p; y_1, \dots, y_n) = p^{\sum y_i}(1-p)^{n-\sum y_i} = p^t(1-p)^{n-t}$$

$$g(p,t) = p^t(1-p)^{n-t} \quad \text{and} \quad h \equiv 1$$

Example. **Normal w/ a known σ.**

$$L(\mu; y_1, \dots, y_n) = \frac{}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$