M378K: April 11^{th}, 2025.

Sufficient Statistics.

Consider two election candidates: A & B

Goal: To "predict" if A wins.

Let the random sample be: $Y_1, Y_2, \dots, Y_n$

Scenarios for book-keeping:

$$\left.\begin{array}{l} \bullet\ T_1 = (Y_1, \dots, Y_n) \\ \bullet\ T_2 = Y_1 + Y_2 + \dots + Y_n \\ \bullet\ T_3 = \mathbb{1}_{\{\bar{Y} > 0.5\}} \end{array}\right\}$$

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 distn of $T_1 = (Y_1, \dots, Y_n)$ given $T_3 = \mathbb{1}_{\{\bar{Y} > 0.5\}}$ 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

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

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

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

Analogously, in the continuous case:

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

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. $\boxed{\text{Claim:}}$ T_2 is sufficient for p

$$p_{Y_1,\dots,Y_n|T_2}(y_1,\dots,y_n|t) = \frac{\mathbb{P}[Y_1=y_1,\dots,Y_n=y_n, T_2=t]}{\mathbb{P}[T_2=t]} \checkmark$$

Cases:

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

2nd $\quad \boxed{y_1+\dots+y_n=t}$ ✓ $\quad \{T_2=t\} \supseteq \{Y_1=y_1,\dots,Y_n=y_n\}$

$$\mathbb{P}[Y_1=y_1,\dots,Y_n=y_n, \boxed{T_2=t}] = \mathbb{P}[Y_1=y_1,\dots,Y_n=y_n]$$

$$= \mathbb{P}[Y_1=y_1]\cdot\mathbb{P}[Y_2=y_2]\cdots\mathbb{P}[Y_n=y_n]$$

$$= p^{y_1}\cdot(1-p)^{1-y_1}\cdot p^{y_2}(1-p)^{1-y_2}\cdots p^{y_n}(1-p)^{1-y_n}$$

$$= p^{\Sigma y_i}(1-p)^{n-\Sigma y_i}$$

$$= p^t(1-p)^{n-t}$$

$$\mathbb{P}[T_2=t] = \binom{n}{t}p^t(1-p)^{n-t}$$

$$p_{Y_1,\dots,Y_n|T_2}(y_1,\dots,y_n|t) = \frac{\cancel{p^t(1-p)^{n-t}}}{\binom{n}{t}\cancel{p^t(1-p)^{n-t}}} = \frac{1}{\binom{n}{t}}$$

Theorem. **The Fisher-Neyman Factorization Criterion**

Let $Y_1, \dots, Y_n$ be a random sample w/ the likelihood function

$$L(\theta; y_1, \dots, y_n).$$

The statistic T is sufficient for θ if and only if L can be expressed as

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