Data Reduction

Any statistic \( T(X) \) defines a form of data reduction.

**Motivation:** If you use \( T(X) \) rather than \( X \), you'll treat \( X \) and \( Y \) as equal if \( T(X) = T(Y) \).

**Sufficient:** \( X_1, \ldots, X_n \sim f(x \mid \theta) \)

A sufficient statistic \( T(X) \) for a parameter \( \theta \) captures all the information about \( \theta \).

**Formally:** \( T(X) \) is sufficient for \( \theta \) (or for the family \( f = \{f(x \mid \theta) : \theta \in \Theta \} \)) if the conditional distribution of \( X_1, \ldots, X_n \) given \( T(X) \) is free of \( \theta \).

\[
\frac{f(x \mid T(X) = T(x))}{f_T(T(X) = T(x))} = \frac{f(x, T(X) = T(x))}{f_T(T(X) = T(x) \mid \theta)} = \frac{f(x \mid \theta)}{f_T(T(X) = T(x) \mid \theta)}
\]

We need:
1. Joint pmf/pdf of the sample
2. Distribution of the statistic

If \( T(X) \) is suff., \( \rightarrow \) LHS is free of \( \theta \), say \( h(x) \), then

\[
h(x) = \frac{f(x \mid \theta)}{f_T(T(X) = T(x) \mid \theta)} \rightarrow \frac{f(x \mid \theta)}{f_T(T(X) = T(x) \mid \theta) h(x)}
\]

\( \rightarrow \) **Fisher-Neyman Factorization Thm.** \( T(X) \) is suff. for \( \theta \) iff

\[
f(x \mid \theta) = g(T(X) \mid \theta) \cdot h(x)
\]

(\( f(x \mid \theta) \) can be expressed in terms of functions \( g \) and \( h \)).