7.8.1 Statistic

**Definition 7.8.1.1: Statistic**

A statistic is any function \( T : \mathbb{R}^n \rightarrow \mathbb{R} \) of samples \( x = (x_1, \ldots, x_n) \). Examples include:

- \( T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \) (the sum)
- \( T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \) (the mean)
- \( T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\} \) (the max/largest value)
- \( T(x_1, \ldots, x_n) = x_1 \) (just take the first sample)
- \( T(x_1, \ldots, x_n) = 7 \) (ignore all samples)

7.8.2 Sufficient Statistics Intuition

Suppose we have iid samples \( x = (x_1, \ldots, x_n) \) from a known distribution with unknown parameter \( \theta \). Imagine we have two people:

- **Statistician A:** Knows the entire sample, gets \( n \) quantities: \( x = (x_1, \ldots, x_n) \).
- **Statistician B:** Knows \( T(x_1, \ldots, x_n) = t \), a single number which is a function of the samples. For example, the sum or the maximum of the samples.

Heuristically, \( T(x_1, \ldots, x_n) \) is a sufficient statistic if Statistician B can do just as good a job as Statistician A, given “less information”. For example, if the samples are from the Bernoulli distribution, knowing \( T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \) (the number of heads) is just as good as knowing all the individual outcomes, since our estimate would be something like the number of heads over the number of total trials! Hence, we don’t actually care the ORDER of the outcomes, just how many heads occurred!

7.8.3 Sufficiency

**Definition 7.8.3.1: Sufficiency**

A statistic \( T = T(X_1, \ldots, X_n) \) is a sufficient statistic if the conditional distribution of \( X_1, \ldots, X_n \) given \( T = t \) and \( \theta \) does not depend on \( \theta \).
\[ P(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) \]

This definition is hard to check, so we have a theorem next which makes it easier!

The idea is, Statistician B only knows \( T = t \), but since \( T \) is sufficient, doesn’t need \( \theta \) to generate new samples \( X_1', \ldots, X_n' \) from the distribution. Now Statistician B has \( n \) iid samples from the distribution, just like Statistician A. So no one is at any disadvantage. :)

It turns out that there is a criterion that helps us determine whether a statistic is sufficient:

### 7.8.4 Neyman-Fisher Factorization Criterion

**Definition 7.8.4.1: Neyman-Fisher Factorization Criterion**

Let \( x_1, \ldots, x_n \) be iid random samples with likelihood \( L(x_1, \ldots, x_n \mid \theta) \). A statistic \( T = T(x_1, \ldots, x_n) \) is sufficient if and only if there exist non-negative functions \( g \) and \( h \) such that:

\[ L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(T(x_1, \ldots, x_n), \theta) \]

That is, the likelihood of the data can be split into a product of two terms: the first term \( g \) can depend on the entire data, but not \( \theta \), and the second term \( h \) can depend on \( \theta \), but only on the data through the sufficient statistic \( T \). (In other words, \( T \) is the only thing that allows the data \( x_1, \ldots, x_n \) and \( \theta \) to interact!) That is, we don’t have access to the \( n \) individual quantities \( x_1, \ldots, x_n \); just the single number \( T \), the sufficient statistic.

If you are reading this for the first time, you might not think this is any better... You should be very confused right now, but let’s see some examples to make it better!

In addition to having our estimators be unbiased, consistent, and efficient, we’d like them to be sufficient (statistics) as well.

### 7.8.5 Sufficiency Examples

**Examples**

1. Let \( x_1, \ldots, x_n \) be iid random samples from \( Unif(0, \theta) \). Show that the MLE \( \hat{\theta} = T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\} \) is a sufficient statistic. (The reason this is true is because we don’t need to know each individual sample to have a good estimate for \( \theta \); we just need to know the largest!)

\[
L(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{x_i \leq \theta} = \frac{1}{\theta^n} I_{x_1 \leq \theta, \ldots, x_n \leq \theta} = \frac{1}{\theta^n} I_{x_{\text{max}} \leq \theta} = \frac{1}{\theta^n} I_{T(x_1, \ldots, x_n) \leq \theta}
\]

Choose \( g(x_1, \ldots, x_n) = 1 \)
Let $x_1, \ldots, x_n$ be iid random samples from $Poi(\theta)$. Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don’t need to know each individual sample to have a good estimate for $\theta$; we just need to know how many events happened totally!)

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!} = \left( \prod_{i=1}^n e^{-\theta} \right) \left( \prod_{i=1}^n \theta^{x_i} \right) \left( \prod_{i=1}^n \frac{1}{x_i!} \right) = \frac{1}{\prod_{i=1}^n x_i!} e^{-n\theta} \theta^{\sum_{i=1}^n x_i}$$

Choose

$$g(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$$

and

$$h(T(x_1, \ldots, x_n), \theta) = e^{-n\theta} \theta^{T(x_1, \ldots, x_n)}$$

By the Neyman-Fisher Factorization Criterion, $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is sufficient. The mean $\frac{\sum_{i=1}^n x_i}{n} = \frac{T(x_1, \ldots, x_n)}{n}$ is as well, since knowing the total number of events and the average number of events is equivalent (since we know $n$)! This is a good property of an estimator! Notice that here the only interaction between the data and parameter $\theta$ happens through the sufficient statistic (the sum/mean of all the values). We don’t actually need to know each individual $x_i$.

3. Let $x_1, \ldots, x_n$ be iid random samples from $Ber(\theta)$. Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don’t need to know each individual sample to have a good estimate for $\theta$; we just need to know how many heads happened totally!)

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \frac{\theta^{x_i} (1 - \theta)^{1-x_i}}{x_i!} = \left( \prod_{i=1}^n \theta^{x_i} \right) \left( \prod_{i=1}^n (1 - \theta)^{1-x_i} \right)$$

Choose

$$g(x_1, \ldots, x_n) = 1$$

and

$$h(T(x_1, \ldots, x_n), \theta) = \theta^{T(x_1, \ldots, x_n)} (1 - \theta)^{n-T(x_1, \ldots, x_n)}$$
By the Neyman-Fisher Factorization Criterion, \( T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \) is sufficient. The mean \( \hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{T(x_1, \ldots, x_n)}{n} \) is as well, since knowing the total number of heads and the sample proportion of heads is equivalent (since we know \( n \)!)! This is a good property of an estimator! Notice that here the only interaction between the data and parameter \( \theta \) happens through the sufficient statistic (the sum/mean of all the values). We don’t actually need to know each individual \( x_i \).

7.8.6 Properties of Estimators Summary

**Fact 7.8.6.1: Bias**

Let \( \hat{\theta} \) be an estimator for \( \theta \). The bias of \( \hat{\theta} \) as an estimator for \( \theta \) is

\[
\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta
\]

As estimator is unbiased if \( \text{Bias}(\hat{\theta}, \theta) = 0 \) or equivalently, \( E[\hat{\theta}] = \theta \).

**Fact 7.8.6.2: Mean Squared Error (MSE)**

The mean squared error of an estimator \( \hat{\theta} \) of \( \theta \) measures the expected squared error from the true value \( \theta \), and decomposes into a bias term and variance term. This term results in the phrase ”Bias-Variance Tradeoff” - usually these are opposing forces and minimizing MSE is a result of choosing the right balance.

\[
MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias^2(\hat{\theta}, \theta)
\]

If \( \hat{\theta} \) is an unbiased estimator of \( \theta \), then the MSE reduces to just: \( MSE(\hat{\theta}, \theta) = Var(\hat{\theta}) \).

**Fact 7.8.6.3: Consistency**

An estimator \( \hat{\theta}_n \) (depending on \( n \) iid samples) of \( \theta \) is consistent if it converges (in probability) to \( \theta \). That is, for any \( \epsilon > 0 \),

\[
\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0
\]

**Fact 7.8.6.4: Efficiency**

An unbiased estimator \( \hat{\theta} \) is efficient if it achieves the Cramer-Rao Lower Bound, meaning it has the lowest variance possible.

\[
e(\hat{\theta}) = \frac{I(\theta)^{-1}}{Var(\hat{\theta})} = 1 \iff Var(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{-E \left[ \frac{\partial^2 \ln L(x|\theta)}{\partial \theta^2} \right]}
\]
Fact 7.8.6.5: Sufficiency

An estimator $\hat{\theta} = T(x_1, \ldots, x_n)$ is **sufficient** if it satisfies the **Neyman-Fisher Factorization Criterion**. That is, there exist non-negative functions $g$ and $h$ such that:

$$L(x_1, \ldots, x_n | \theta) = g(x_1, \ldots, x_n) \cdot h(\hat{\theta}, \theta)$$