Lecture 4.

Data is information. About what? About the underlying distribution that the data is sampled from. We can always create data by simulation if we know the underlying distribution exactly, in which case the data has no value. In that case any lost data can be recreated and it does not make any difference. If we do not know the distribution exactly then the loss is indeed a loss and we can not recreate the data by simulation or "drawing a sample" again because we do to know which population to draw from.

A statistic is a function of the observations $t(x_1, x_2, \ldots, x_n)$. If we calculate $t = t(x_1, x_2, \ldots, x_n)$ from the sample, store it and discard the sample, we have lost some information but not all. Some times we can resurrect it.

Sufficiency. Let us go back to coin tossing. We denote the probability of a head in a toss by θ. But we have data in the form of $(x_1, x_2, \ldots, x_n)$ where each x_i is 1 for a head and 0 for a tail. The probability for a string $x_1, x_2, \ldots, x_n$ of 0's and 1's.

$$p(\theta, x_1, x_2, \ldots, x_n) = \theta^t (1-\theta)^{n-t}$$

where $t = t(x_1, x_2, \ldots, x_n) = \sum_i x_i$. The probability depends on the observations $(x_1, x_2, \ldots, x_n)$ only thorough t. What does that mean? We know the number of heads in n tosses but we have lost the information about the order in which head or tail occurred. Once we know the number of heads is t, there are $\binom{n}{t}$ possible arrangements and they all have equal probability. Arrange in a random order and nobody would notice! In other words the "order" in this case has no additional information about the parameter θ beyond what is provided by the total number t of heads.

Geometric variables. $x \geq 0$

$$p(\theta, x) = \theta(1-\theta)^x$$

If we have n of them

$$p(\theta, x_1, x_2, \ldots, x_n) = \theta^n (1-\theta)^{t(x_1, x_2, \ldots, x_n)}$$

where $t(x_1, x_2, \ldots, x_n) = \sum_i x_i$

We have a coin with unknown probability θ for a head. We keep tossing till we get n heads, the i-th head preceded by x_i tails. $t = \sum_i x_i$. If we only know t and do not know θ can we recreate $\{x_i\}$? We see that all of them have equal probability. How many are there? The number of ways of dividing t into $\{x_i\}$ is choosing t slots among $n + t - 1$ slots for the tails. Out of $n + t$ slots the last one has to be a head. Each arrangement has probability $\binom{n+t-1}{t}^{-1}$.

$$p(\theta, x_1, x_2, \ldots, x_n) = \theta^n (1-\theta)^t \binom{n+t-1}{t} \binom{n+t-1}{t}^{-1}$$

Poisson variables. $x \geq 0$

$$p(\theta, x) = \frac{e^{-\theta}\theta^x}{x!}$$

With $t = t(x_1, x_2, \ldots, x_n) = \sum_i x_i$

$$p(\theta, x_1, x_2, \ldots, x_n) = \frac{e^{-n\theta}\theta^{t(x_1,x_2,\ldots,x_n)}}{x_1!x_2!\cdots x_n!}$$

Now $p(\theta, x_1, x_2, \ldots, x_n)$ factors as

$$p(\theta, x_1, x_2, \ldots, x_n) = p_1(\theta, t) p_2(x_1, x_2, \ldots, x_n)$$

where

$$p_1(\theta, t) = \frac{e^{-n\theta} n^t \theta^t}{t!}$$

and

$$p_2(x_1, x_2, \ldots, x_n) = \frac{t!}{x_1!x_2!\ldots x_n!} \frac{1}{n^t}$$

You have monthly data on the number of accidents accidents for each month. It is known to follow a Poisson distribution, with an unknown parameter, but is known to be the same for every month. You have kept the total number of accidents for the year but have lost the information about how many each month. You can recreate it by assigning each accident randomly with probability $\frac{1}{12}$ to one of the months and you have

successfully recreated the lost data. If you can do it then you really did not need the full data, the partial information was "SUFFICIENT". The crucial step is that the conditional distribution of the full sample $x_1, \ldots, x_n$ given t should not dependent of θ, so given t we can recreate $x_1, \ldots, x_n$ with out knowing θ.

It is a little more complicated in continuous variables. Let for $\theta > 0$, on $[0, \infty)$ the density be given by

$$f(\theta, x) = \theta e^{-\theta x}$$

Then with $t = \sum_i x_i$

$$f(\theta, x_1, x_2, \ldots, x_n) = \theta^n e^{-\theta t}$$

Given the sum t, $x_1, x_2, \ldots, x_n$ is uniformly distributed over the hyperplane $x_i \geq 0$, $\sum_i x_i = t$ no matter what the value of t is. So t is sufficient.

What if

$$f(\theta, x) = \theta^2 e^{-\theta x} x$$

We have again with $t = \sum_i x_i$

$$f(\theta, x_1, x_2, \ldots, x_n) = \theta^{2n} e^{-\theta t} x_1 x_2 \cdots x_n$$

Now given t the distribution on $\sum_i x_i = t$ is not uniform but has a weight $x_1 x_2 \cdots x_n$ that is still independent of θ. It will have some normalization $c(t)$ and we will factor

$$f(\theta, x_1, x_2, \ldots, x_n) = \theta^{2n} e^{-\theta t} c(t) \times [c(t)]^{-1} x_1 x_2 \cdots x_n$$

$$c(t) = \frac{t^{2n-1}}{\Gamma(2n)}$$

Some times we need more than one "t", especially if there are many parameters.

Normal Family

$$f(\mu, \theta, x) = c(\theta) \exp[-\frac{(x-\mu)^2}{2\theta}] = c(\theta, \mu) \exp[\frac{\mu}{\theta} x - \frac{x^2}{2\theta}]$$

We need both $t_1 = \frac{1}{n}\sum_i x_i$ and $t_2 = \sum_i x_i^2$. Given t_1 and t_2, we have uniform distribution on the intersection of the sphere $\sum_i x_i^2 = t_2$ with the hyperplane $\sum_i x_i = t_1$. Do not need to know the values of μ, θ for this. t_1, t_2 are sufficient for μ, θ. Some times we may need more than one statistic for a single parameter. If we know that $\theta = \mu^2$ and $\mu \neq 0$ is the only parameter still with

$$f(\mu, \theta, x) = c(\theta) \exp[-\frac{(x-\mu)^2}{2\mu^2}] = c(\mu) \exp[\frac{1}{\mu}x - \frac{x^2}{2\mu^2}]$$

we still need $\sum_i x_i$ and $\sum_i x_i^2$.

Do sufficient statistics exist. The entire sample is sufficient. Order is not needed. Use symmetric functions. No need to use information you do not need!

Other examples. Uniform distribution on $[0, \theta]$.

$$f(\theta, x_1, x_2, \ldots, x_n) = \theta^{-n}; \quad 0 \leq x_i \leq \theta \quad \forall i$$

$t = \max_i x_i$ is sufficient. Given t one of the $\{x_i\}$ has to be t and the remaining are uniform over $[0, t]$.

$$f(\theta, x_1, x_2, \ldots, x_n) = n\theta^{-n} t^{n-1} \frac{1}{nt^{n-1}}$$

and $f(\theta, t) = n\theta^{-n} t^{n-1}$ gives the density of t on $[0, \theta]$.

Rao-Blackwell theorem. If there is a sufficient statistic use it! We have a parametric family $f(x, \theta)$. We may want to estimate θ or a given function $h(\theta)$ of the unknown parameter θ from a random sample from the population. We have an estimator $u(x_1, \ldots, x_n)$ and a sufficient statistic $t(x_1, x_2, \ldots, x_n)$. We can replace $u(x_1, \ldots, x_n)$ by $\hat{u}(t) = E[u(x_1, \ldots, x_n)|t]$ which will not depend on θ. Why is $\hat{u}$ better than u? From the properties

of conditional expectation

$$\begin{aligned}
E[u(x_1,\ldots,x_n)] &= \sum_{x_1,\ldots,x_n} u(x_1,\ldots,x_n)p(\theta,x_1,\ldots,x_n) \\
&= \sum_{x_1,\ldots,x_n} u(x_1,\ldots,x_n)p_1(\theta,t)p_2(t,x_1,\ldots,x_n) \\
&= \sum_t p_1(\theta,t) \sum_{\substack{x_1,\ldots,x_n \\ t(x_1,\ldots,x_n)=t}} u(x_1,\ldots,x_n)p_2(t,x_1,\ldots,x_n) \\
&= \sum_t \hat{u}(t)p_1(\theta,t) \\
&= E[\hat{u}(t)]
\end{aligned}$$

$$E[(u-\theta)^2] = E\big[[(u-\hat{u})+(\hat{u}-\theta)]^2\big]$$

The cross term

$$E[(u-\hat{u})(\hat{u}-\theta)] = E\big[E[(u-\hat{u})|t](\hat{u}-\theta)\big] = 0$$

Therefore

$$E[(u-\theta)^2] \geq E[(\hat{u}-\theta)^2]$$

Works for $E[|u-\theta|]$ or any convex loss function. $E[L(u,\theta)]$ It is better to replace u by $\hat{u}$.

Jensen's inequality. For convex ϕ.

$$E[\phi(X)] \geq \phi(E[X])$$

Proof:

$$\phi(x) = \sup_{a,b\in S}[ax+b]$$

$$E[\max(f,g)] \geq \max([E[f],E[g]])$$

Conditional expectation is still an expectation.

Cramér-Rao inequality. Let u be an unbiased estimator of θ.

$$\sum p(\theta,x_1,x_2,\ldots,x_n) \equiv 1$$

$$\sum [u(x_1, x_2, \ldots, x_n) p(\theta, x_1, x_2, \ldots, x_n)] \equiv \theta$$

Differentiate

$$\sum [\frac{dp(\theta, x_1, x_2, \ldots, x_n)}{d\theta}] \equiv 0$$

$$\sum [\frac{d \log p(\theta, x_1, x_2, \ldots, x_n)}{d\theta} p(\theta, x_1, x_2, \ldots, x_n)] \equiv 0$$

$$\sum [u(x_1, x_2, \ldots, x_n) \frac{dp(\theta, x_1, x_2, \ldots, x_n)}{d\theta}] \equiv 1$$

$$\sum [u(x_1, x_2, \ldots, x_n) \frac{d \log p(\theta, x_1, x_2, \ldots, x_n)}{d\theta} p(\theta, x_1, x_2, \ldots, x_n)] \equiv 1$$

Set $X = u, Y = \frac{d \log p(\theta, x_1, x_2, \ldots, x_n)}{d\theta}$. Then

$E[Y] = 0, E[XY] = 1; E[(X - E(X))(Y - E(Y)] = E[XY] - E[X]E[Y] = 1$

By Schwarz's inequality

$$E[(X - E[X])^2] E[Y^2] \geq 1$$

On the other hand

$$Y = \sum_i \frac{d \log p(\theta, x_i)}{d\theta}$$

and

$$E[Y^2] = nE[(\frac{d \log p(\theta, x)}{d\theta})^2] = n\, I(\theta)$$

You can not have an unbiased estimator with variance smaller than $\frac{1}{nI(\theta)}$.

Examples. Coin Tossing.

$$p(\theta, 1) = \theta, p(\theta, 0) = 1 - \theta$$

$$I(\theta) = \frac{1}{\theta} + \frac{1}{1 - \theta} = \frac{1}{\theta(1 - \theta)}$$

Lower bound is $\frac{1}{n}\theta(1 - \theta)$. can not improve $\frac{r}{n}$ Minimum variance unbiased estimator.

Examples. Normal.

$$p(\theta, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

$$\frac{d \log p}{d\theta} = (x - \theta)$$

$$E[(x-\theta)^2] = 1,\ Var(u) \geq \frac{1}{n}$$

$\bar{x}$ is a minimum variance unbiased estimator. What if we do not demand unbiasedness. But allow a bias, $b(\theta)$. The lower bound is $\frac{(1+b'(\theta))^2}{nI(\theta)}$.

Estimation from parametric families. We have a family of probability distributions $\{p(\theta, x)\}$ or densities $\{f(\theta, x)\}$, indexed by one are more parameters. We have a sample $\{X_1, X_2, \ldots, X_n\}$ form one of these, we do not know which. Guess ?. $\theta = t(X_1, X_2, \ldots, X_n)$ is an estimate. It is purely a function of $\{X_1, X_2, \ldots, X_n\}$ and does not depend on θ. Not all choices of $t(\cdot)$ are good choices. What makes a choice good and how do find good choices ?

Examples:

1. Tossing a coin n times with unknown probability θ of head. If $\frac{r}{n}$ is the proportion of heads in n tosses then we saw

$$E_\theta[\frac{r}{n}] = \theta, E_\theta[(\frac{r}{n} - \theta)^2] = \frac{\theta(1-\theta)}{n}$$

2. How about densities $f(\theta, x) = \theta x^{\theta-1}$ on $[0, 1]$?

$$E[X] = \int \theta x^\theta dx = \frac{\theta}{\theta + 1}$$

If

$$t_n = t_n(X_1, X_2, \ldots, X_n) = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

then

$$E_\theta[t_n] = \frac{\theta}{\theta + 1}$$

Variance of X is

$$\int_0^1 \theta x^{\theta+1} dx - (\frac{\theta}{\theta+1})^2 = \frac{\theta(p\theta+1)^2 - \theta^2(\theta+2)}{(\theta+2)(\theta+1)^2} = \frac{\theta}{(\theta+2)\theta+1)^2}$$

Variance of t_n is therefore $\frac{\theta}{n(\theta+2)(\theta+1)^2}$.

If $t_n \simeq \frac{\theta}{\theta+1}$ then is $u_n = \frac{t_n}{1-t_n} \simeq \theta$? Can we do better?

How about

$$E[-\log X]? = -\int_0^1 \theta \log x x^{\theta-1} dx = \int x^{\theta-1} dx = \frac{1}{\theta}$$

Is the estimate $v_n = \frac{n}{-\sum \log X_i}$ better? How do we compare them?

$$\begin{aligned} E[(u_n - \theta)^2] &= E\left[\left[\frac{t_n}{1-t_n} - \frac{\frac{\theta}{\theta+1}}{1-\frac{\theta}{\theta+1}}\right]^2\right] \\ &\simeq E[f(t_n) - f(E(t_n))]^2 \\ &\simeq [f'(E(t_n))]^2 E[(t_n - E(t_n))^2] \\ &\simeq [f'(E(t_n))]^2 Var(t_n) \end{aligned}$$

$f(t) = \frac{t}{1-t}, f'(t) = (\frac{1}{1-t})^2 = (\frac{1}{1-\frac{\theta}{\theta+1}})^2 = (\theta+1)^2$.

Variance of u_n is roughly $\frac{\theta(\theta+1)^2}{n(\theta+2)}$. On the other hand

$$\theta \int (\log x)^2 x^{\theta-1} dx = -2 \int \log x \, x^{\theta-1} dx = \frac{2}{\theta^2}$$

Variance of $\log X$ is $\frac{1}{\theta^2}$. $f(t) = \frac{1}{t}$. $f'(t) = \frac{-1}{t^2}$

Variance of v_n is roughly $\frac{\theta^4}{n\theta^2} = \frac{\theta^2}{n}$.

One can verify that

$$\theta^2 < \frac{\theta(\theta+1)^2}{(\theta+2)}$$