Information contained in an experiment

It is of interest to know how informative is an experiment about the unknown parameters. Binomial and negative binomial sampling provide different amount of information depending on how large or small p is. In *Information Theory*, Shannon information is mostly used, which is a measure of entropy or randomness, but in statistics different measures are used. The notion that is described and used here is based on 'the difference that we see when we change the model continuously from one to another'.

Information number (Fisher). Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions satisfying the following mathematical regularity conditions.

(A) $A = \{x : f(x|\theta) > 0\}$ does not depend on θ. For all $x \in A$ and $\theta \in \Theta$, the score function,

$$S(x) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}$$

exists and is finite.

$S(x)$ measures the relative rate at which $f(x|\theta)$ changes at x. Since x varies (due to X being random) this needs averaging.

$$\begin{aligned} I(\theta) &= E_\theta \left[\frac{\partial}{\partial\theta} \log f(X|\theta) \right]^2 \\ &= \int \left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right)^2 f(x|\theta)\, dx \end{aligned}$$

is called the Fisher Information number of θ contained in $f(.|\theta)$ or P_θ. Clearly, $0 \leq I(\theta) \leq \infty$. To get a feeling for $I(\theta)$, consider an extreme case where $f(x|\theta)$ is free of θ. Clearly, in this case there can be no information about θ in $\mathbf{X}$. On the other hand, if $I(\theta)$ is large, then on an average a small change in θ leads to a big change in $\log f(x|\theta)$, i.e., f depends strongly on θ and one expects there is a lot that can be learned about θ.

Example. $X \sim$ Bernoulli(p). i.e., how much information is there in a single

toss of a coin on its success probability, p?

$$\begin{aligned} f(x|p) &= p^x(1-p)^{1-x},\ x=0,1 \\ \log f(x|p) &= x\log p + (1-x)\log(1-p) \\ \frac{\partial}{\partial p}\log f(x|p) &= \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-xp-p+xp}{p(1-p)} = \frac{x-p}{p(1-p)},\ \text{so} \\ I(p) &= E_p\left[\frac{\partial}{\partial p}\log f(X|p)\right]^2 = E_p\left[\frac{(X-p)^2}{p^2(1-p)^2}\right] \\ &= \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)} = \frac{1}{Var_p(X)}. \end{aligned}$$

In other words, Information is inversely proportional to the variance. Intuitively, if the variance is large, or if p is far away from 0 or 1, then one will need a large number of observations to get a reliable estimate of p. If p is close to 0 or 1, the observations will be mostly 0, or mostly 1, so that estimation is easy. On the other hand, if p is close to 1/2, there will be a lot of fluctuation, and much variability. Then it will be difficult to distinguish between models.

Theorem. If (A) holds, and
(B) the derivative with respect to θ of $\int f(x|\theta)\,dx$ can be obtained by differentiating under the integral sign, then
(i) $E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = 0$, and
(ii) $I(\theta) = Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]$.
In addition, if
(C) the second derivative (w.r.t. θ) of $\log f(x|\theta)$ exists for all x and θ, and the second derivative of $\int f(x|\theta)\,dx$ can be obtained by differentiating twice under the integral sign, then
(iii) $I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$.

(A)-(C) are called Cramer-Rao (C-R) regularity conditions.

Proof. (i). Since $\int_{-\infty}^{\infty} f(x|\theta)\,dx = 1$, we have,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x|\theta)\,dx = \int\frac{\partial}{\partial\theta}f(x|\theta)\,dx = \int\left\{\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right\}f(x|\theta)\,dx \\ &= \int\left\{\frac{\partial}{\partial\theta}\log f(x|\theta)\right\}f(x|\theta)\,dx = E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}. \end{aligned}$$

(ii). Now, using this, we get,

$$\begin{aligned} I(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta) - E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}\right]^2 \\ &= Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]. \end{aligned}$$

(iii). To obtain this alternative formula, note that,

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}\log f(x|\theta) &= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] \\ &= \frac{\frac{\partial^2}{\partial^2\theta}f(x|\theta)}{f(x|\theta)} - \frac{\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)^2}{\left[f(x|\theta)\right]^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} E_\theta\left[\frac{\partial^2}{\partial^2\theta}\log f(X|\theta)\right] &= E_\theta\left\{\frac{\frac{\partial^2}{\partial^2\theta}f(X|\theta)}{f(X|\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right)^2\right\} \\ &= \int \frac{\frac{\partial^2}{\partial^2\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx - E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 \\ &= \int \frac{\partial^2}{\partial^2\theta} f(x|\theta)\,dx - E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 \\ &= 0 - I(\theta), \end{aligned}$$

since

$$0 = \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty} f(x|\theta)\,dx = \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)\,dx.$$

The condition (A) that the support of $f(\cdot|\theta)$ is free of θ is essential. The exponential families satisfy these regularity conditions. Location-scale families may or may not satisfy, usually the critical assumption is that relating to the support of f. Thus the Cauchy location-scale family satisfies these conditions but not the uniform or the exponential density

$$f(x|\mu,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right),\ x > \mu.$$