My white board writing from week 4 – 17th August

1) Example: For $\mathbf{X} = (X_1, X_2, \dots, X_n)$ i.i.d. Bernoulli with parameter θ, i.e. $f(x_i) = P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$, $x_i = 0, 1$

we claim that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

The proof is by inspecting the original definition. We have the following partitions created by T: $\mathcal{A} = (A_0, A_1, A_2, \dots, A_n)$ where $A_r = \{\mathbf{X} \mid \sum_{i=1}^{n} X_i = r\}$ for $r = 0, 1, 2, \dots, n$.

Then
$$P(\mathbf{X} = x \mid \mathbf{X} \in A_r) = \frac{P(\mathbf{X} = x \cap \mathbf{X} \in A_r)}{P(\mathbf{X} \in A_r)} \qquad (*)$$

Noting that $\mathbf{X} \in A_r$ means that $\sum_{i=1}^{n} X_i = r$ and $\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \theta)$ we have $P(\mathbf{X} \in A_r) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$

and
$$P(\mathbf{X} = x \cap \mathbf{X} \in A_r) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} x_i \neq r \\ \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} & \text{when } \sum_{i=1}^{n} x_i = r \end{cases}$$

we continue from $(*)$:
$$P(\mathbf{X} = x \mid \mathbf{X} \in A_r) = \begin{cases} 0 & \text{if } \sum x_i \neq r \\ \dfrac{\cancel{\theta^r(1-\theta)^{n-r}}}{\binom{n}{r}\cancel{\theta^r(1-\theta)^{n-r}}} = \dfrac{1}{\binom{n}{r}} & \text{if } \sum_{i=1}^{n} x_i = r \end{cases}$$

Hence this conditional probability does not depend on θ.

2) Next, I considered the proof of Neyman-Fisher's factorization criterion (for the discrete case only).

There are 2 "directions" to be shown.

$(\Leftarrow)$ Assuming that $L(X, \theta) = g(T(X), \theta)\,h(X)$ holds we need to check that T is sufficient for θ. Looking at
$$P(X = x \mid T = t) = \frac{P(X = x \cap T = t)}{P(T = t)} = \begin{cases} 0 & \text{if } T(x) \neq t \\ \dfrac{P(X = x)}{P(T = t)} & \text{if } T(x) = t \end{cases}$$

we continue in the second case as follows:
$$\frac{P(X = x)}{P(T = t)} = \frac{\cancel{g(t, \theta)}\,h(x)}{\sum_{\tilde{x} : T(\tilde{x}) = t} \cancel{g(t, \theta)}\,h(\tilde{x})} = \frac{h(x)}{\sum_{\tilde{x} : T(\tilde{x}) = t} h(\tilde{x})}$$
which does <u>not</u> involve θ.

(⇒) If on the other hand T was sufficient then

$$P_\theta(X=x) = P(X=x \cap \underbrace{T=T(x)}_{\text{since this is the sure event}}) = P(X=x \mid T=t)P_\theta(T=t) =$$

denoting $T(x)=t$

$= h(x)g(t,\theta)$ where we denoted $P(X=x \mid T=t) = h(x)$ and it is known not to depend on θ by assumption, whereas $P_\theta(T=t) = g(t,\theta)$ involves the data via the value of the statistic only, i.e. the factorization is demonstrated.

---

(3) I gave several examples of using the factorization criterion to show sufficiency:

i) For Bernoulli: $T=\sum_{i=1}^{n} X_i$ is sufficient which follows directly from: $L(X,\theta) = \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i}$, which involves the data only via the value of $\sum_{i=1}^{n} X_i = T$ so the whole RHS can be thought of as $g(t,\theta) = \theta^t(1-\theta)^{n-t}$ (and there is no need of $h(X)$ here, it can be set to the constant 1).

ii) $N(\mu,\sigma^2)$ with $\theta = \binom{\mu}{\sigma^2}$.

Using the fundamental equality $\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$ we have:

$$L(X;\theta) = \frac{1}{(\sqrt{2\pi}\,\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2\right]\right)$$

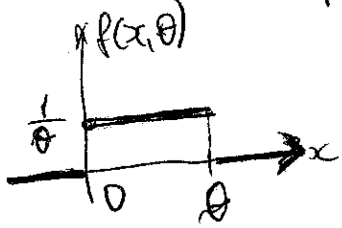
which involves the data via $T=\binom{T_1}{T_2} = \begin{pmatrix}\bar{X}\\ \sum_{i=1}^{n}(X_i-\bar{X})^2\end{pmatrix}$ only.

Hence this 2-dim vector statistic is sufficient for $\theta=\binom{\mu}{\sigma^2}$

I also noted that every 1-to-1 transformation of T is also sufficient. In particular, $\tilde{T}=\begin{pmatrix}\tilde{T}_1\\ \tilde{T}_2\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^{n} X_i\\ \sum_{i=1}^{n} X_i^2\end{pmatrix}$

is also sufficient for θ (since knowing $\begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$ we can get $\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix}$ and vice versa.

iii) $\mathbf{X} = (X_1, X_2, \dots, X_n)$ uniformly distributed in $[0, \theta)$. We claim that $T = X_{(n)}$ (the n-th order statistic, equal to the maximal of the observations) is sufficient for θ.



Note that the density (show graphically on the left) can be written using indicator function as

$$f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta)$$

Then

$$L(\mathbf{X},\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(X_i,\infty)}(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(X_i,\infty)}(\theta) =$$

$$= \boxed{\frac{1}{\theta^n} I_{(X_{(n)},\infty)}(\theta)} = g(X_{(n)},\theta) * 1$$

which represents a factorization and $T = X_{(n)}$ is sufficient according to the factorization criterion.

iv) Multivariate normal:

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be n i.i.d p-dim multivariate normal data vectors $X_i \sim N_p(\mu, \Sigma)$.

We have the p-dim version of the fundamental equality:

$$\sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)' = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' + n(\bar{X} - \mu)(\bar{X} - \mu)'$$

We also use properties of traces: $(X'AX) = tr(A(XX'))$

Then: $L(\mathbf{X}; \mu, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} tr\, \Sigma^{-1}(X_i - \mu)(X_i - \mu)'\right)$

$$\underset{\substack{\uparrow \\ \text{since } tr \text{ and } \sum_{i=1}^{n} \\ \text{are linear operators} \\ \text{and order can be exchanged}}}{=} (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\left(\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' + n(\bar{X} - \mu)(\bar{X} - \mu)'\right)\right]\right\}$$

which involves the data only via $T_1 = \bar{X}$ and $T_2 = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$

Hence T_1 and T_2 give a sufficient statistic for μ and Σ.

(4) I showed many examples about proving <u>minimal</u> sufficiency via the Lehmann-Scheffe method:

i) For Bernoulli (θ). Take 2 independent n-tuples of data $X=(X_1,X_2,\dots,X_n)$ and $Y=(Y_1,Y_2,\dots,Y_n)$ and write

$$\frac{L(Y,\theta)}{L(X,\theta)}=\frac{\theta^{\sum_{i=1}^n Y_i}(1-\theta)^{n-\sum_{i=1}^n Y_i}}{\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}}=\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n Y_i-\sum_{i=1}^n X_i}$$

For this to not depend on θ $\longrightarrow$ can only happen when $\sum_{i=1}^n X_i=\sum_{i=1}^n Y_i$ $\longrightarrow$ Hence $T=\sum_{i=1}^n X_i$ is minimal sufficient

(i.e., the sets in the minimal sufficient partition are the contours of the statistic $T(X)=\sum_{i=1}^n X_i$)

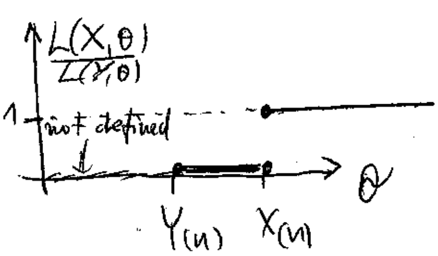
ii) $N(\mu,\sigma^2)$, $\theta=\binom{\mu}{\sigma^2}$. It can be seen easily that

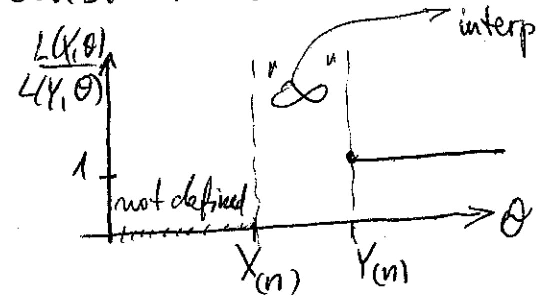
$$\frac{L(Y,\theta)}{L(X,\theta)}=\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n Y_i^2-\sum_{i=1}^n X_i^2-2\mu\left(\sum_{i=1}^n Y_i-\sum_{i=1}^n X_i\right)\right)\right)$$

and for this to not depend on θ we need both $\begin{pmatrix}\sum_{i=1}^n X_i=\sum_{i=1}^n Y_i\\ \sum_{i=1}^n X_i^2=\sum_{i=1}^n Y_i^2\end{pmatrix}$ to hold. Hence $T=\begin{pmatrix}\sum_{i=1}^n X_i\\ \sum_{i=1}^n X_i^2\end{pmatrix}$ is minimal sufficient for $\theta=\binom{\mu}{\sigma^2}$ (as is also any 1-1 transformation of T)

iii) <u>Uniform in $[0,\theta)$</u>. For $\frac{L(X,\theta)}{L(Y,\theta)}$ we get $\frac{I_{(X_{(n)},\infty)}(\theta)}{I_{(Y_{(n)},\infty)}(\theta)}$

This is independent of θ if and only if $X_{(n)}=Y_{(n)}$ which implies that $T=X_{(n)}$ is minimal sufficient. (Indeed, if $X_{(n)}\neq Y_{(n)}$ we can consider two cases: interpreted $\left[\frac{1}{0}\right]\overset{\text{Def}}{=}\infty$





In both cases when $X_{(n)} \neq Y_{(n)}$, the ratio's value (where defined), depends on the position of θ, i.e., it is not independent of θ. To have it not depending on θ, we need $X_{(n)} = Y_{(n)}$ to hold.)

iv) One more example to show that not necessarily is the dimension of the minimal sufficient statistic equal to the dimension of the parameter (as it was in the previous 3 examples):

If $X_1, X_2, \dots, X_n$ are i.i.d. Cauchy(θ) (i.e., with density $f(x,\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, $-\infty < x < \infty$) then

$$\frac{L(Y,\theta)}{L(X,\theta)} = \frac{\prod_{i=1}^{n} (1+(X_i-\theta)^2)}{\prod_{i=1}^{n} (1+(Y_i-\theta)^2)}$$ and we see that

unless $\begin{pmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{pmatrix} = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \\ \vdots \\ Y_{(n)} \end{pmatrix}$, the ratio will depend on θ. Hence $T = \begin{pmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{pmatrix}$ is the minimal sufficient statistic in this case (although its dimension is equal to the sample size, so virtually no dimension reduction is possible in this case)

5) Next, I showed several examples of one-parameter exponential families and explained the related minimal sufficient statistics for them:

i) $f(x,\theta) = \theta \exp(-\theta x)$: $a(\theta) = \theta$, $b(x) = 1$, $c(\theta) = -\theta$, $d(x) = x$
$\longrightarrow T = \sum_{i=1}^{n} X_i$ is minimal sufficient

ii) $f(x,\theta) = \frac{e^{-\theta}\theta^{x}}{x!} = e^{-\theta}\frac{1}{x!}e^{x\ln\theta} \longrightarrow \begin{cases} a(\theta) = e^{-\theta} \\ b(x) = \frac{1}{x!} \\ c(\theta) = \ln\theta \\ d(x) = X \end{cases}$

Hence $T = \sum_{i=1}^{n} X_i$ is minimal sufficient

iii); etc. → for your own exercise

iv) $N(0,\theta^2) \longrightarrow f(x,\theta) = \frac{1}{\sqrt{2\pi}\,\theta} \cdot e^{-\frac{1}{2\theta^2}X^2}$

Hence $a(\theta) = \frac{1}{\sqrt{2\pi}\,\theta}$, $b(x) = 1$, $c(\theta) = -\frac{1}{2\theta^2}$, $d(x) = X^2$

Hence $T = \sum_{i=1}^{n} X_i^2$ is minimal sufficient

The generalization for K-parameter exponential families:

Example: $N(\mu,\sigma^2)$:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}X^2 + X\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$$

is 2 par. exponential and you can choose $d_1 = X$, $d_2 = X^2$ to end up with $\begin{pmatrix}\sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2\end{pmatrix}$ as a minimal sufficient.

I also left it for you to convince yourself that

$$f(X;\theta_1,\theta_2) = \frac{1}{B(\theta_1,\theta_2)}X^{\theta_1-1}(1-X)^{\theta_2-1}, \quad X\in(0,1), \quad \theta_1,\theta_2 > 0$$

(the Beta density) belongs to a 2 parameter exponential family and a minimal sufficient statistic for $\cancel{\theta} = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}$ is $\begin{pmatrix}\sum_{i=1}^{n}\ln X_i \\ \sum_{i=1}^{n}\ln(1-X_i)\end{pmatrix}$.

6) Then I moved over to the ancillarity principle in inference. I abstain from reproducing the discussion here since it is thorough enough in the notes.

I just summarize my discussion about the Pitman estimator $\hat{\theta}_p$.

The claim is that if you consider equivariant estimators $\hat{\theta}$ (i.e. satisfying $\hat{\theta}(X_1+c, X_2+c, \dots, X_n+c) = \hat{\theta}(X_1, X_2, \dots, X_n) + c$) when dealing with location parameter estimation then you can find <u>the best</u> equivariant estimator with respect to mean-squared error (that is in the class of these estimators, there is one particular one (namely $\hat{\theta}_p$) which minimizes $E_\theta(\theta - \hat{\theta})^2$ for all θ! (i.e. has uniformly smallest risk with respect to quadratic loss). What is interesting is that in its construction you utilise the ancillary statistic $\tilde{T}_2 = (X_2 - X_1, X_3 - X_1, \dots, X_n - X_1)$. Starting with an <u>arbitrary</u> equivariant estimator $\tilde{\theta}$, you construct $\hat{\theta}_p$ as $\hat{\theta}_p = \tilde{\theta} - E_0(\tilde{\theta} / \tilde{T}_2)$.

When the loss is quadratic, you end up with $\hat{\theta}_p$ posessing the above optimality. It turns out that

$$\hat{\theta}_p = \frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^n f(X_i - \theta)\, d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^n f(X_i - \theta)\, d\theta}$$

which you can obviously interpret as a Bayes estimator w.r. quadratic loss and w.r. <u>improper</u> prior $\pi(\theta) \equiv 1$ on R^1

Of course, this prior is improper because it is not a density over $(-\infty, \infty)$ but you can exploit the analogy.

Finally, I also convinced you that if the location family $f_\theta(x) = f(x-\theta)$ we were dealing with was the $N(\theta, 1)$ family then $\hat{\theta}_p$, as discussed above, is the familiar $\overline{X}$ (the arithmetic mean).

Indeed: $$\hat{\theta}_p = \frac{\int_{-\infty}^{\infty} \theta e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} d\theta}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} d\theta}$$

Within the integrals, we can multiply with any factor on top and bottom as long as it does not involve the θ. Hence we can write:

$$\hat{\theta}_p = \frac{\int_{-\infty}^{\infty} \theta e^{-\frac{n}{2}\theta^2 + \theta\sum_{i=1}^{n} x_i} d\theta}{\int_{-\infty}^{\infty} e^{-\frac{n}{2}\theta^2 + \theta\sum_{i=1}^{n} x_i} d\theta}$$

multiply by $e^{-\frac{n}{2}(\overline{X})^2}$ to complete the square

$$= \frac{\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \theta e^{-\frac{n}{2}(\theta-\overline{X})^2} d\theta}{\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{n}{2}(\theta-\overline{X})^2} d\theta}$$

$\Leftarrow$ If we interprete θ as a random variable with $\theta \sim N(\overline{X}, \frac{1}{n})$

then on top we have written the expected value of this variable (which is $\overline{X}$) and on bottom we have integrated out the density of this random variable (which gives us 1). Hence the ratio is equal to $\frac{\overline{X}}{1} = \overline{X}$.

I also mentioned that Pitman is a famous AUSTRALIAN Statistician