Entropy and Mutual Information

Sofia Olhede

EPFL

September 27, 2020

Entropy etc

- The following variational problem is interesting: Determine the probability distribution f supported on $\mathcal{X}$ with maximum entropy

$$H(f) = -\int_{x\in\mathcal{X}} f_X(x)\log\{f_X(x)\}dx,$$

subject to the linear constraints

$$\int_{\mathcal{X}} T_i(x)f(x)\,dx = \alpha_i, \quad i = 1, 2, 3, \ldots, k.$$

- This is in the setting of choosing a probability model that gives the highest uncertainty (max entropy).
- Proposition: When a solution to the above constrained optimisation problem exists, it is unique and has the form

$$f(x) = Q(\lambda_1, \ldots, \lambda_k)\exp(\sum_{i=1}^{k} \lambda_i T_i(x)).$$

Entropy etc

- Proof: Let $g(x)$ be a density also satisfying the constraints (also means that we assume $f(x)$ does). Then

$$
\begin{aligned}
H(g) &= -\int_{x\in\mathcal{X}} g(x)\log\{g(x)\}dx \\
&= -\int_{x\in\mathcal{X}} g(x)\log\{\frac{g(x)}{f(x)}f(x)\}dx && (1)\\
&= -\mathrm{KL}(f||g) - \int_{\mathcal{X}} g(x)\log f(x)dx && (2)\\
&\leq -\int_{\mathcal{X}} g(x)\log f(x)dx, && (3)
\end{aligned}
$$

as $-\mathrm{KL}(f||g)\leq 0$. We have assumed $g(x)$ satisfies the constraints and so:

Entropy etc

- we have that

$$
\begin{aligned}
H(g) &\leq -\int_{\mathcal{X}} g(x) \log f(x) dx \\
&= -\int_{\mathcal{X}} g(x)[\log Q + \sum_{i=1}^{k} \lambda_i T_i(x)] dx && (4) \\
&= -\log(Q) - \int_{\mathcal{X}} f(x) \sum_{i=1}^{k} \lambda_i T_i(x) dx && (5) \\
&= -\int_{\mathcal{X}} f(x) \log(f(x))\, dx = H(f). && (6)
\end{aligned}
$$

Uniqueness in turn follows from when the divergence can be zero.

Entropy etc

- We also define the conditional entropy $H(f_Y|f_X)$ as the entropy of the conditional distribution averaged over the domain of X. Let Y have distribution f_Y and X in turn f_X, and as usual let the conditional distribution be $f_{Y|X}(y|x)$. We will now swap notation from densities to random variables:

$$H(Y|X) = -\int_{x\in\mathcal{X}} f_X(x) \int_{\mathcal{Y}} f_{Y|X}(y|x) \log\{f_{Y|X}(y|x)\} dydx.$$

  We can also define the joint entropy

$$H(X,Y) = -\int_{x\in\mathcal{X},y\in\mathcal{Y}} f_{X,Y}(x,y) \log\{f_{X,Y}(x,y)\} dxdy.$$

  By introducing this notation we can state the entropy *chain rule* of

$$H(X,Y) = H(X) + H(Y|X).$$

Entropy etc

- One of the measures of dependence between X and Y is to use the mutual information $I(X,Y)$. as

$$I(X,Y) = -\int_{x\in\mathcal{X},y\in\mathcal{Y}} f_{X,Y}(x,y)\log\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\}dxdy.$$

  It transpires that

$$I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y).$$

  $I(X,Y)$ measures the reduction in uncertainty of X due to knowledge of Y and is symmetric in X and Y.

- Proposition: $I(X,Y) \geq 0$. Furthermore $I(X,Y) = 0 \Leftrightarrow X$ and Y are independent.

Entropy etc

- First consider the continuous case. $-\log(x)$ is a convex function on $x \geq 0$. Therefore using Jensen's inequality (lecture 2) we know

$$
\begin{aligned}
I(X, Y) &= -\int_{x\in\mathcal{X}, y\in\mathcal{Y}} f_{X,Y}(x,y) \log\{\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\} dxdy \\
&= \mathbb{E}_{X,Y}\left(\log\{\frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)}\}\right) \\
&\geq \log \mathbb{E}_{X,Y}\{\frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)}\} \\
&= \log(1) = 0.
\end{aligned}
$$

The proof for the discrete case is very similar.

Entropy etc

- $\Rightarrow$. If X and Y are independent then their pdfs factorize. Then

$$\begin{aligned} I(X,Y) &= -\int_{x\in\mathcal{X},y\in\mathcal{Y}} f_{X,Y}(x,y)\log\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\}dxdy \\ &= -\int_{x\in\mathcal{X},y\in\mathcal{Y}} f_{X,Y}(x,y)\log\{\frac{f_X(x)f_Y(y)}{f_X(x)f_Y(y)}\}dxdy = 0. \end{aligned} \quad (7)$$

- The discrete case follows *Mutatis mutandis*.

Entropy etc

- $\Leftarrow$. Now if $I(X,Y)=0$ then

$$0=-\int_{x\in\mathcal{X},y\in\mathcal{Y}} f_{X,Y}(x,y)\log\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\}dxdy.$$

  This is equivalent to saying $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ everywhere, as it is multiplied by a non-negative function, which is the definition of independence.
- The discrete case follows *Mutatis mutandis*.

Exponential Family

- A probability distribution is said to be a member of a k-parameter exponential family, if its density (or frequency), admits the representation

$$f(y) = \exp\left\{ \sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi_1, \ldots, \phi_k) + S(y) \right\} \tag{8}$$

where

(a) $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ is a k-dimensional parameter in $\Phi \subseteq \mathbb{R}^k$;
(b) $T_i : \mathcal{Y} \to \mathbb{R}$ and $\gamma : \mathbb{R}^k \to \mathbb{R}$ are real-valued;
(c) The support $\mathcal{Y}$ of f does not depend on $\boldsymbol{\phi}$.

Exponential Family II

- A very rich class of models. (Sometimes requiring fixing some parameters to satisfy last condition): Binomial, Negative Binomial, Poisson, Gamma, Gaussian, Pareto, Weibull, Laplace, logNormal, inverse Gaussian, inverse Gamma, Normal-Gamma, Beta, Multinomial . . .
- Basis for Generalised Linear Models (GLM).
- We will gradually appreciate the tractable properties of such models.
- ϕ is called the natural parameter.
- We can transform this parameter to write the family in other ways.
- The word "natural" here comes from a mathematics point of view. The usual parameter that is used is $\boldsymbol{\theta} = \eta^{-1}(\phi)$.

Exponential Family III

- Thus we may write and equate natural and usual parameterisations:

$$\exp\left\{\sum_{i=1}^{k}\phi_i T_i(y)-\gamma(\boldsymbol{\phi})+S(y)\right\}=\exp\left\{\sum_{i=1}^{k}\eta_i(\boldsymbol{\theta})T_i(y)-d(\boldsymbol{\theta})+S(y)\right\}. \tag{9}$$

- Here $\eta: \mathbb{R}^k \to \mathbb{R}^k$ is a C^2 map such that

$$\boldsymbol{\phi}=\eta(\boldsymbol{\theta}) \tag{10}$$

$$\gamma(\boldsymbol{\phi})=\gamma(\eta(\boldsymbol{\theta}))=d(\boldsymbol{\theta}), \tag{11}$$

  for $d=\gamma\circ\eta$.
- Natural paramaterization: this is good for mathematical manipulation.
- Usual parameterization: this is good for intuition.

Exponential Family IV

- Example: binomial exponential family. Let $Y \sim \mathrm{Binom}(n, p)$. Observe that

$$\binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{ y \cdot \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\left(\binom{n}{y}\right)\right\} \tag{12}$$

Define the new parameterisation

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(y) = y,$$

and additionally define

$$S(y) = \log\left(\binom{n}{y}\right), \quad \gamma(\phi) = n \log(1 + e^{\phi}) = -n \log(1-p).$$

Exponential Family V

- Keeping n fixed and allowing only p to vary, the support of f does not depend on ϕ and we get a 1-parameter family. Note that:

$$p = \frac{e^{\phi}}{1+e^{\phi}}, \quad \phi = \log\left(\frac{p}{1-p}\right).$$

- Thus the usual parameter is $p \in (0,1)$ but the natural one is $\phi \in \mathbb{R}$.
- Example, the Gaussian distribution.
- Let $Y \sim N(\mu, \sigma^2)$. We shall write

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) && (13)\\ &= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}. && (14) \end{aligned}$$

Exponential Family VI

- Define
$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$
and also
$$T_1(y) = y,\ T_2(y) = y^2,\ S(y) = 0,\ \gamma(\phi) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2}\log(-\frac{\pi}{\phi_2}).$$
- We observe that the support of f is always the entire real line.
- We in general model the observed phenomenon by a distribution $F(y_1, \ldots, Y_n; \theta)$ on $\mathcal{Y}^n$ for some $n \geq 1$.
- The distributional form is assumed known but $\theta \in \Theta$ is assumed to be unknown.
- We observe a realisation (or a sample) of $(Y_1, \ldots, Y_n)^T \in \mathcal{Y}^n$.
- Use the sample in order to make assertions concerning the true value of θ and we quantify the certainty of our assertions.

Exponential Family VII

- We use sampling theory to understand how functions $T = T(Y_1, \ldots, Y_n)$ carry information about the parameter θ.
- We determine the probability distribution of T and determine how that relates to the distribution of the sample.
- Definition (Statistic). A statistic is any function T of the data whose domain is the sample space $\mathcal{Y}^n$ but which does not depend on any unknown parameters.
- Intuitively any function that can be evaluated from the sample is a statistic.
- Any statistic is a random variable with its own distribution.
- Example: with n known $T(\boldsymbol{Y}) = n^{-1}\sum_{i=1}^{n} Y_i$ is a statistic.
- Example: $T = T(\boldsymbol{Y}) = (Y_{(1)}, \ldots, Y_{(n)}$ where $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are the order statistics of $\boldsymbol{Y}$. Since T depends only on the values of $\boldsymbol{Y}$ it is a statistic.

Exponential Family VIII

- Definition (Sampling Distribution) Let $(Y_1 \ldots, Y_n)^T \sim F(y_1, \ldots, y_n; \theta)$ and let T be a q-dimensional statistic

$$T(Y_1, \ldots, Y_n) = \begin{pmatrix} T_1(Y_1 \ldots, Y_n) & \ldots & T_q(Y_1 \ldots, Y_n) \end{pmatrix}.$$

The sampling distribution of T under $F(y_1, \ldots, Y_n; \theta)$ is the distribution:

$$F_T(t_1, \ldots t_q) = \Pr(T_1(Y_1 \ldots, Y_n) \leq t_1, \ldots, T_q(Y_1 \ldots, Y_n) \leq t_q).$$

- Will normally write just T rather than $T = T(Y_1, \ldots, Y_n)$.
- Very often $T : \mathcal{Y}^n \to \mathbb{R}$ in which case this notation can be simplified.

Exponential Family IX

- Very often $T : \mathcal{Y}^n \to \mathbb{R}$ in which case this notation can be simplified.
- In this case we have

$$F_T(t) = \Pr\{T(\mathrm{Y}) \leq t\}, \quad t \in \mathbb{R}.$$

- The sampling distribution of T depends on the unknown θ but it can be computed from the data alone.
- The extent and form of this dependence is crucial for inference.
- Evident from previous examples, some statistics are more informative and others are less informative regarding the true value of θ.
- Any $T(Y_1, \ldots, Y_n)$ that is not “1-1” carries less information about θ than the original sample.
- This makes us wonder as to what are “good” and “bad” statistics.
- Definition. Ancillary statistics. A statistic T is ancillary fo θ if its distribution does not functionally depend on θ.

Exponential Family X

- Suppose that $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, identically and independently. Assume σ^2 known. Let $T(Y_1, \ldots, Y_n) = Y_1 - Y_2$. Then T has a normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for $\theta = (\mu, \sigma^2)$.
- If T is ancillary for θ then informally t carries no information about θ.
- In order to carry any useful information about θ, the sampling distribution F_T must explicitly depend on θ.
- Intuitively the amount of information T carries on θ increases as the dependence of its sampling distribution F_T on θ increases.

Exponential Family XI

- Example: let $Y_1, \dots, Y_n \overset{iid}{\sim} U[0, \theta]$. Take $S = \min(Y_1, \dots, Y_n)$ and take $T = \max(Y_1, \dots, Y_n)$. Note that

  * Note that $f_S(y; \theta) = \frac{n}{\theta}\left(1 - \frac{y}{\theta}\right)^{n-1}$ for $0 \leq y \leq \theta$ and $f_T(y; \theta) = \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1}$ for $0 \leq y \leq \theta$.
  * In this case neither S nor T are ancillary for θ.
  * As $n \to \infty$ $f_S(\cdot)$ becomes concentrated around 0.
  * As $n \to \infty$ $f_T(\cdot)$ becomes concentrated around θ.
  * This indicates that T provides more information about θ then S does.

- To understand the information carried by statistics on θ we need to understand how the statistics are related to the sample space.

Exponential Family XII

- Our formal relationship is understood by the following:

  * $\boldsymbol{Y} = (Y_1, \ldots, Y_m) \overset{iid}{\sim} F_{\boldsymbol{\theta}}$ and $T(\boldsymbol{Y})$ is a statistic.
  * The *level sets* or *contours* of T are the sets

  $$A_t = \{\boldsymbol{y} \in \mathcal{Y}^n : \; T(\boldsymbol{y}) = t\}.$$

  This corresponds to all potential samples that could have given us the value t for T.
- Any realization of $\boldsymbol{Y}$ that falls in a given level set is equivalent as far as T is concerned, as T reduces all these values to the same output.
- Any inference drawn through T will be the same within a given level set.
- Therefore it makes sense to consider the distribution of Y conditional on a given fibre A_t of T, $F_{\mathrm{Y}|T=t}(\mathrm{y})$.
- If $F_{\mathrm{Y}|T=t}(\mathrm{y})$ changes with $\boldsymbol{\theta}$ then we are losing information.
- If $F_{\mathrm{Y}|T=t}(\mathrm{y})$ is functionally independent of $\boldsymbol{\theta}$ then it does not matter whether whether we observe Y or just $T(\mathrm{Y})$.

Exponential Family XII

- Knowing the exact value of Y in addition to knowing $T(\mathrm{Y})$ does not give us any additional information. In a sense Y is irrelevant if we know $T(\mathrm{Y})$.
- Sufficient Statistic: A Statistic $T = T(\mathrm{Y})$ is said to be sufficient for the parameter θ if the conditional probability distribution of the sample given the statistic

$$F_{\mathrm{Y}\,|\,T=y}(y_1,\ldots,y_n) = \Pr\{\,Y_1 \leq y_1,\ldots,Y_n \leq y_n \,|\, T = t\},$$

does not depend on θ.
- Example: (Coin Tossing). Let $Y_1,\ldots,Y_n \sim \mathrm{Bernoulli}(\theta)$ independently. Let $T(\mathrm{Y}) = \sum_{i=1}^{n} Y_i$. For $\mathrm{y} \in \{0,1\}^n$ note that

$$\begin{aligned}\Pr\{\mathrm{Y}=\mathrm{y}\,|\,T=t\} &= \frac{\Pr\{\mathrm{Y}=\mathrm{y}\,\cap\,T=t\}}{\Pr\{T=t\}}\\ &= \frac{\Pr\{\mathrm{Y}=\mathrm{y}\}}{\Pr\{T=t\}}\mathrm{I}\left(\sum_{i=1}^{n} y_i = t\right)\end{aligned}$$

Exponential Family XIII

- Thus T is sufficient for θ.
- In general the definition of sufficiency is hard to verify.
- **Theorem (Fisher–Neyman factorization theorem)**: suppose that Y has a joint density or frequency function $f(\mathrm{y};\theta)$, where $\theta \in \mathbf{\Theta}$. A Statistic $T = T(\mathrm{Y})$ is sufficient for θ if and only if

$$f(\mathrm{y};\theta) = g(T(\mathrm{y}),\theta)h(\mathrm{y}).$$

- **Example** Let $Y_1,\dots,Y_n \sim \mathcal{U}[0,\theta]$ independently. This means any sample has pdf $f(y;\theta) = \frac{\mathrm{I}(y\in[0,\theta])}{\theta}$. Then we have that

$$f(\mathrm{y};\theta) = \frac{\mathrm{I}(\max_i y_i \leq \theta)\mathrm{I}(\min_i y_i \geq 0)}{\theta^n}.$$

- From this equation we may deduce that $T(\mathrm{y}) = \max_i y_i$ is sufficient for θ.

Exponential Family XIII

- **Proof of the Fisher–Neyman factorization theorem.** Suppose that T is sufficient. Then

$$f(y;\theta) = \Pr(\mathrm{Y} = \mathrm{y}) = \sum_t \Pr\{\mathrm{Y} = \mathrm{y}, T = t\}$$

$$= \Pr\{\mathrm{Y} = \mathrm{y}, T = t(\mathrm{y})\} = \Pr\{T = t(\mathrm{y})\} \Pr\{\mathrm{Y} = \mathrm{y} \mid T = t(\mathrm{y})\}.$$

Since T is sufficient, it follows that $\Pr\{\mathrm{Y} = \mathrm{y} \mid T = t(\mathrm{y})\}$ is independent of θ and so $f(\mathrm{y};\theta) = g(T(\mathrm{y});\theta)h(\mathrm{y})$.
Now suppose that $f(\mathrm{y};\theta) = g(T(\mathrm{y});\theta)h(\mathrm{y})$. Then if $T(\mathrm{y}) = t$ it follows that

$$\Pr\{\mathrm{Y} = \mathrm{y} \mid T = t\} = \frac{\Pr\{\mathrm{Y} = \mathrm{y} \cap \ T = t\}}{\Pr\{T = t\}} = \frac{\Pr\{\mathrm{Y} = \mathrm{y}\}}{\Pr\{T = t\}} \mathrm{I}(T(\mathrm{y}) = t)$$

$$= \frac{g(T(\mathrm{y});\theta)h(\mathrm{y})}{\sum_{\mathrm{z}, T(\mathrm{z}) = t} g(T(\mathrm{z});\theta)h(\mathrm{z})} \mathrm{I}(T(\mathrm{y}) = t) = \frac{h(\mathrm{y})}{\sum_{\mathrm{z}, T(\mathrm{z}) = t} h(\mathrm{z})} \mathrm{I}(T(\mathrm{y}) = t).$$

This does not depend on θ.

Sampling Theory I

- Example: (sufficient statistics for iid normal samples). Let $Y_1, \dots Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Recall that we may write

$$f(y; \mu, \sigma^2) = \frac{e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}} = \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}.$$

  It follows from that expression that we have

$$f(\mathrm{y}; \mu, \sigma^2) = \exp\left\{-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\mu \sum_i y_i}{\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

- Consequently Fisher-Neyman factorization implies that the vector-valued statistic

$$S(\mathrm{Y}) = \begin{pmatrix} S_1(\mathrm{Y}) & S_2(\mathrm{Y}) \end{pmatrix}^T = \begin{pmatrix} \sum_i Y_i & \sum_i Y_i^2 \end{pmatrix}^T = \begin{pmatrix} \overline{Y} & \sum_i Y_i^2 \end{pmatrix}^T,$$

  is sufficient for the parameter (μ, σ^2) and so is the statistic

$$T(\mathrm{Y}) = \begin{pmatrix} T_1(\mathrm{Y}) & T_2(\mathrm{Y}) \end{pmatrix}^T = \begin{pmatrix} \overline{Y} & \frac{1}{n}\sum_i (Y_i - \overline{Y})^2 \end{pmatrix}^T,$$

Sampling Theory II

- since T and S are 1-1 functions of each other.
- Example: Sufficient statistics for k-parameter exponential families. More generally, consider a k-parameter exponential family with density

$$f(y) = \exp\left\{ \sum_{j=1}^{k} \phi_j T_j(y) - \gamma(\phi_1, \cdots, \phi_k) + S(y) \right\}, \quad y \in \mathcal{Y}.$$

Then an iid sample $(Y_1, \ldots, Y_n)^T$ has a joint distribution of

$$f(\mathrm{y}) = \exp\left\{ \sum_{j=1}^{k} \phi_j \tau_j(y_1, \ldots, y_n) - n\gamma(\phi_1, \cdots, \phi_k) + \sum_{i=1}^{n} S(y_i) \right\},$$

where

$$\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i).$$

Sampling Theory III

- So the statistic

$$\tau(Y_1,\dots,Y_n) = \left(\tau_1(y_1,\dots,y_n),\dots,\tau_k(y_1,\dots,y_n)\right)^T$$

  which is sufficient for $(\phi_1,\dots,\phi_k)$ by Fisher-Neyman factorization.
- This, and other examples, show that sufficient statistics compress the data without information loss on the parameter of interest.
- How much information can be thrown away?
- Definition (Minimally sufficient statistic). A statistic $T = T(\mathrm{Y})$ is said to be minimally sufficient for the parameter θ if it is sufficient for θ and for any other sufficient statistic $S = S(\mathrm{Y})$ there exists a function $g(\cdot)$ with

$$T(\mathrm{Y}) = G(S(\mathrm{Y})).$$

- Lemma: If T and S are minimally sufficient statistics for a parameter θ, then there exists injective functions g and h such that $S = g(T)$ and $T = h(S)$.

Sampling Theory IV

- Theorem: Let $\mathrm{Y} = \left(Y_1, \ldots, Y_n\right)$ have joint density or frequency function $f(\mathrm{y}; \theta)$ and let $T = T(\mathrm{Y})$ be a statistic. Suppose that $f(\mathrm{y}; \theta)/f(\mathrm{z}; \theta)$ is independent of θ if and only if $T(\mathrm{y}) = T(\mathrm{z})$. Then T is minimally sufficient for θ.
- Proof: Assume for simplicity that $f(\mathrm{y}; \theta) > 0$ for all $\mathrm{y} \in \mathbb{R}^n$ and $\theta \in \Theta$. Let $\mathcal{T} = \{T(\mathrm{u}) : \mathrm{u} \in \mathbb{R}^n\}$ be the image of $\mathbb{R}^n$ under T and let A_t be the level sets of T. For each t, choose a representative element $\boldsymbol{w}_t \in A_t$. Notice that for any y $\boldsymbol{w}_{T(y)}$ is in the same level set as y so that

$$f(\mathrm{y}; \theta)/f(\boldsymbol{w}_{T(y)}; \theta),$$

does not depend on θ by assumption. Let $g(t, \theta) \equiv f(\boldsymbol{w}_t; \theta)$ and notice that

$$f(\mathrm{y}; \theta) = \frac{f(\boldsymbol{w}_{T(y)}; \theta) f(\mathrm{y}; \theta)}{f(\boldsymbol{w}_{T(y)}; \theta)} = g(T(y), \theta) h(y).$$

Sufficiency follows from the Fisher-Neyman factorization theorem.

Sampling Theory V

- To obtain minimality we suppose that T' is another sufficient statistic. By the factorization theorem $\exists g',\ h' : \quad f(\mathrm{y};\theta) = g'(T'(\mathrm{y});\theta)h'(\mathrm{y})$. Let y and z be such that $T'(\mathrm{y}) = T'(\mathrm{z})$. Then
$$\frac{f(\mathrm{y};\theta)}{f(\mathrm{z};\theta)} = \frac{g'(T'(\mathrm{y});\theta)h'(\mathrm{y})}{g'(T'(\mathrm{z});\theta)h'(\mathrm{z})} = \frac{h'(\mathrm{y})}{h'(\mathrm{z})}.$$
Since the ratio does not depend on θ, we have by assumption $T(\mathrm{y}) = T(\mathrm{z})$. Hence T is a function T' so is minimal by an arbitrary choice of T'.
- Example (Bernoulli trials). Let $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathrm{Ber}(\theta)$. Consider z and y, two possible distinct outcomes. Then we may note that
$$\frac{f(\mathrm{z};\theta)}{f(\mathrm{y};\theta)} = \frac{\theta^{\sum z_i}(1-\theta)^{n-\sum z_i}}{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}}.$$
This ratio is constant only if $\sum z_i = \sum y_i = T(\mathrm{y})$. We may note that T is minimially sufficient.

Sampling Theory VI

- Example: minimal sufficiency for the k parameter exponential family. An iid sample $(Y_1, \ldots, Y_n)^T$ from an exponential family has joint distribution

$$f(\mathrm{y}) = \exp\left\{\sum_{j=1}^{k} \phi_j \tau_j(y_1, \ldots, y_n) - n\gamma(\phi_1, \cdots, \phi_k) + \sum_{i=1}^{n} S(y_i)\right\},$$

where

$$\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i).$$

- If the summary statistics $\{T_j\}_{j=1}^{k}$ are non-trivial then $f(\boldsymbol{y})/f(\boldsymbol{z})$ will be constant with respect to the collection $\{\phi_j\}$ if and only if as $\{\phi_j\}$ varies the following quantity is constant;

$$\sum_{j=1}^{k} \phi_j[\tau_j(y_1, \ldots, y_n) - \tau_j(z_1, \ldots, z_n)].$$

Sampling Theory VI

- If $(\phi_1, \dots, \phi_k)$ range over an open parameter space of dimension k this implies we must require

$$\tau_j(y_1, \dots, y_n) = \tau_j(z_1, \dots, z_n).$$

Conversely when this equation holds for all j then the density ratio does not depend on $\boldsymbol{\phi}$ and this implies minimal sufficiency of $\boldsymbol{\tau}$.

Sampling Distributions

- By studying sampling distributions we aim to determine what different information do different forms of T carry about θ.
- Theorem: (Sampling Distributions of Gaussian Sufficient Statistics). Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and define

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2.$$

- The pair $(\bar{Y}, S^2)$ are minimally sufficient for (μ, σ^2) and
  - (a) The sample mean has distribution $\bar{Y} \sim N(\mu, \sigma^2/n)$,
  - (b) The random variables $\bar{Y}$ and S^2 are independent,
  - (c) The random variable S^2 satisfies $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.
- Corollary: (Moments of Sufficient Statistics). If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then

$$\mathbb{E}(\bar{Y}) = \mu, \quad \mathbb{V}\text{ar}\{\bar{Y}\} = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2, \quad \mathbb{V}\text{ar}\{S^2\} = \frac{2\sigma^4}{n-1}.$$

Sampling Distributions II

- Theorem (Sum of Gaussian Squares) Let $(Z_1, \ldots, Z_k)$ be iid $N(0,1)$ random variables. Then

$$Z_1^2 + \cdots + Z_k^2 \sim \chi_k^2.$$

- Recall that $\chi_k^2 \equiv \mathrm{Ga}(k/2, 1/2)$. The pdf, mean, variance and moment generating functions of this distribution is

$$E(X) = k, \ , \mathbb{V}\mathrm{ar}(X) = 2k, \ M(t) = (1-2t)^{-k/2}, \quad t < 1/2.$$

- Theorem: (Student's statistic and its sampling distribution). Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. Then

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$