

# Entropy and Mutual Information

Sofia Olhede



September 27, 2020

- 1 Entropy etc
- 2 Exponential family and Sampling Theory
- 3 Sampling Theory
- 4 Sampling distributions

# Entropy etc

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

Reminder:

$$H(M) = -E(\log f_X(x))$$

$$E(M) = \int_{x \in \mathcal{X}} x f_X(x) dx$$

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

subject to the linear constraints

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

- 
- 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



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

## Entropy etc

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

$$\begin{aligned} H(g) &= - \int_{x \in \mathcal{X}} g(x) \log\{g(x)\} dx \\ &= - \int_{x \in \mathcal{X}} g(x) \log\left\{\frac{g(x)}{f(x)} f(x)\right\} dx \end{aligned} \quad (1)$$

$$= -\text{KL}(f||g) - \int_{\mathcal{X}} g(x) \log f(x) dx \quad (2)$$

$$\leq - \int_{\mathcal{X}} g(x) \log f(x) dx, \quad (3)$$

?

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

## Entropy etc

- we have that

$$\begin{aligned}
 H(g) &\leq - \int_{\mathcal{X}} g(x) \log f(x) dx \\
 &= - \int_{\mathcal{X}} \underbrace{g(x)}_{g(x) \text{ is a pdf:}} [\log Q + \sum_{i=1}^k \lambda_i T_i(x)] dx \quad (4)
 \end{aligned}$$

side 3:  
 $f(x) = Q(\lambda_1, \dots, \lambda_k) \sup \left( \sum_i \lambda_i T_i(x) \right)$   
 $\therefore \log f(x) = \log Q + \sum_i \lambda_i T_i(x)$

$$= - \log(Q) - \int_{\mathcal{X}} f(x) \sum_{i=1}^k \lambda_i T_i(x) dx \quad (5)$$

$$= - \int_{\mathcal{X}} f(x) \log(f(x)) dx = H(f). \quad (6)$$

substituted by  $g(x)$  as it satisfies some properties

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:

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

We can also define the joint entropy

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

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

→ 
$$I(X, Y) = - \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f_{X,Y}(x, y) \log\left\{\frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)}\right\} dx dy.$$

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.

*L.i.e. if adding the knowledge of  $Y$  to  $X$  brings no new information (doesn't reduce uncertainty) then  $X$  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

$$\begin{aligned}
 I(X, Y) &= - \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f_{X,Y}(x, y) \log\left\{\frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)}\right\} dx dy \\
 &= \mathbb{E}_{X,Y}\left(\log\left\{\frac{f_X(x)f_Y(y)}{f_{X,Y}(x, y)}\right\}\right) \\
 &\geq \log \mathbb{E}_{X,Y}\left\{\frac{f_X(x)f_Y(y)}{f_{X,Y}(x, y)}\right\} \quad \text{Jensen's inequality} \\
 &= \log(1) = 0.
 \end{aligned}$$

The proof for the discrete case is very similar.

## Entropy etc

1/2: proof  $\Rightarrow$  if  $X \perp\!\!\! \perp Y \rightarrow I(X, Y) = 0$

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

$$\begin{aligned} I(X, Y) &= - \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f_{X,Y}(x, y) \log \left\{ \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)} \right\} dx dy \\ &= - \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f_{X,Y}(x, y) \underbrace{\log \left\{ \frac{f_X(x)f_Y(y)}{f_X(x)f_Y(y)} \right\}}_{\log(1)=0} dx dy = 0. \quad (7) \end{aligned}$$

independent:  
 $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$

- The discrete case follows *Mutatis mutandis.*

## Entropy etc

z/2: proof  $\Leftarrow$  if  $I(X, Y) = 0 \rightarrow X \perp\!\!\!\perp Y$

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

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

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

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

where

- (a)  $\phi = (\phi_1, \dots, \phi_k)$  is a  $k$ -dimensional parameter in  $\Phi \subseteq \mathbb{R}^k$ ;

$T_i$  is a map  $\mathcal{Y} \rightarrow \mathbb{R}$

but

- (b)  $T_i : \mathcal{Y} \rightarrow \mathbb{R}$  and  $\gamma : \mathbb{R}^k \rightarrow \mathbb{R}$  are real-valued;
- (c) The support  $\mathcal{Y}$  of  $f$  does not depend on  $\phi$ .

ex.: "on a continuous uniform distribution between  $a$  and  $b$ ,  
then the support is between  $a$  and  $b$ , so we can only have  
values between  $a$  and  $b$ , which is an obvious violation!"

# 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.
- $\phi$  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  $\theta = \eta^{-1}(\phi)$ .

# Exponential Family III

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

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

natural parameterization:      usual parameterization:  
(good for mathematics)                  (good for intuition)

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

$$\parallel \quad \phi = \eta(\theta) \quad (10)$$

$$\parallel \quad \gamma(\phi) = \gamma(\eta(\theta)) = d(\theta), \quad (11)$$

for  $d = \gamma \circ \eta$ .

- Natural parameterization: this is good for mathematical manipulation.
- Usual parameterization: this is good for intuition.

# Exponential Family IV

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

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

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  $\phi$  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

$$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \mu}{\sigma}\right)^2\right) \quad (13)$$

$$= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}. \quad (14)$$

# Exponential Family VI

- Define

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$

and also

$$T_1(y) = y, \quad T_2(y) = y^2, \quad S(y) = 0, \quad \gamma(\phi) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2} \log\left(-\frac{\pi}{\phi_2}\right).$$

- We observe that the support of  $f$  is always the entire real line.
- We in general model the observed phenomenon by a distribution  $F(y_1, \dots, Y_n; \theta)$  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  $(Y_1, \dots, Y_n)^T \in \mathcal{Y}^n$ .
- Use the sample in order to make assertions concerning the true value of  $\theta$  and we quantify the certainty of our assertions.

# Exponential Family VII

- We use sampling theory to understand how functions  $T = T(Y_1, \dots, Y_n)$  carry information about the parameter  $\theta$ .
  - 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  $T(\mathbf{Y}) = n^{-1} \sum_{i=1}^n Y_i$  is a statistic.
  - Example:  $T = T(\mathbf{Y}) = (Y_{(1)}, \dots, Y_{(n)})$  where  $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$  are the order statistics of  $\mathbf{Y}$ . Since  $T$  depends only on the values of  $\mathbf{Y}$  it is a statistic.
- T(Y) is the sample average*
- T(Y) is the ordered set of Y*

# Exponential Family VIII

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

A probability distribution is the theoretical outcome of an experiment whereas a sampling distribution is the real outcome of an experiment.

- Definition (Sampling Distribution) Let  $(Y_1, \dots, Y_n)^T \sim F(y_1, \dots, y_n; \theta)$  and let  $T$  be a  $q$ -dimensional statistic  
 $\hookrightarrow Y$  is generated from dist. function  $F$

$$T(Y_1, \dots, Y_n) = (T_1(Y_1, \dots, Y_n) \ \dots \ T_q(Y_1, \dots, Y_n)).$$

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

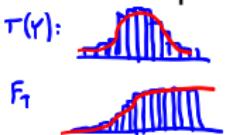
$$F_T(t_1, \dots, t_q) = \Pr(T_1(Y_1, \dots, Y_n) \leq t_1, \dots, T_q(Y_1, \dots, Y_n) \leq t_q).$$

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

# Exponential Family IX

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

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



- The sampling distribution of  $T$  depends on the unknown  $\theta$  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  $\theta$ .
- Any  $T(Y_1, \dots, Y_n)$  that is not “1-1” carries less information about  $\theta$  than the original sample. *→ obvious: sampling less info than original*
- This makes us wonder as to what are “good” and “bad” statistics.
- Definition. Ancillary statistics. A statistic  $T$  is ancillary for  $\theta$  if its distribution does not functionally depend on  $\theta$ .

# Exponential Family X



- Suppose that  $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ , identically and independently. Assume  $\sigma^2$  known. Let  $T(Y_1, \dots, Y_n) = Y_1 - Y_2$ . 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  $\theta$  then informally  $t$  carries no information about  $\theta$ .
- In order to carry any useful information about  $\theta$ , the sampling distribution  $F_T$  must explicitly depend on  $\theta$ .
- Intuitively the amount of information  $T$  carries on  $\theta$  increases as the dependence of its sampling distribution  $F_T$  on  $\theta$  increases.

# Exponential Family XI

- Example: let  $Y_1, \dots, Y_n \stackrel{iid}{\sim} U[0, \theta]$ . 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  $\theta$ .
  - \* As  $n \rightarrow \infty$   $f_S(\cdot)$  becomes concentrated around 0.
  - \* As  $n \rightarrow \infty$   $f_T(\cdot)$  becomes concentrated around  $\theta$ .
  - \* This indicates that  $T$  provides more information about  $\theta$  than  $S$  does.
- To understand the information carried by statistics on  $\theta$  we need to understand how the statistics are related to the sample space.

# Exponential Family XII

How are statistics related to the sample space?

- Our formal relationship is understood by the following:

\*  $\mathbf{Y} = (Y_1, \dots, Y_m) \stackrel{iid}{\sim} F_\theta$  and  $T(\mathbf{Y})$  is a statistic.

\* The level sets or contours of  $T$  are the sets

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

This corresponds to all potential samples that could

$A_t = \{Y_1, Y_2\}$  have given us the value  $t$  for  $T$ .

- Any realization of  $\mathbf{Y}$  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  $\mathbf{Y}$  conditional on a given fibre  $A_t$  of  $T$ ,  $F_{Y|T=t}(y)$ .
- If  $F_{Y|T=t}(y)$  changes with  $\theta$  then we are losing information.
- If  $F_{Y|T=t}(y)$  is functionally independent of  $\theta$  then it does not matter whether we observe  $\mathbf{Y}$  or just  $T(\mathbf{Y})$ .

# Exponential Family XII

*if they're independent (?)*

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

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

does not depend on  $\theta$ .

- Example: (Coin Tossing). Let  $Y_1, \dots, Y_n \sim \text{Bernoulli}(\theta)$  independently. Let  $T(Y) = \sum_{i=1}^n Y_i$ . For  $y \in \{0, 1\}^n$  note that

$$\Pr\{Y = y | T = t\} = \frac{\Pr\{Y = y \cap T = t\}}{\Pr\{T = t\}}$$

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

*experiments where sum = t*

# Exponential Family XIII

- Thus  $T$  is sufficient for  $\theta$ .
- 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  $f(y; \theta)$ , where  $\theta \in \Theta$ . A Statistic  $T = T(Y)$  is sufficient for  $\theta$  if and only if

→ 
$$f(y; \theta) = g(T(y), \theta)h(y).$$

- **Example** Let  $Y_1, \dots, Y_n \sim \mathcal{U}[0, \theta]$  independently. This means any sample has pdf  $f(y; \theta) = \frac{\prod_{i=1}^n I(y_i \in [0, \theta])}{\theta^n}$ . Then we have that

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

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

# Exponential Family XIII

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

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

?

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

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

Now suppose that  $f(y; \theta) = g(T(y); \theta)h(y)$ . Then if  $T(y) = t$  it follows that

$$\begin{aligned} \Pr\{Y = y \mid T = t\} &= \frac{\Pr\{Y = y \cap T = t\}}{\Pr\{T = t\}} = \frac{\Pr\{Y = y\}}{\Pr\{T = t\}} I(T(y) = t) \\ &= \frac{g(T(y); \theta)h(y)}{\sum_{z, T(z)=t} g(T(z); \theta)h(z)} I(T(y) = t) = \frac{h(y)}{\sum_{z, T(z)=t} h(z)} I(T(y) = t). \end{aligned}$$

This does not depend on  $\theta$ .

# Sampling Theory I

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

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

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

$$S(Y) = (S_1(Y) \quad S_2(Y))^T = (\sum_i Y_i \quad \sum_i Y_i^2)^T = (\bar{Y} \quad \sum_i Y_i^2)^T,$$

is sufficient for the parameter  $(\mu, \sigma^2)$  and so is the statistic

$$T(Y) = (T_1(Y) \quad T_2(Y))^T = (\bar{Y} \quad \frac{1}{n} \sum_i (Y_i - \bar{Y})^2)^T,$$

# 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

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

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

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

where

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

# Sampling Theory III

- So the statistic

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

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  $T = T(Y)$  is said to be minimally sufficient for the parameter  $\theta$  if it is sufficient for  $\theta$  and for any other sufficient statistic  $S = S(Y)$  there exists a function  $g(\cdot)$  with

$$T(Y) = G(S(Y)).$$

*→ ie any other "minimally suff."  $S(Y)$  can be transformed into  $T(Y)$*

- Lemma: If  $T$  and  $S$  are minimally sufficient statistics for a parameter  $\theta$ , then there exists injective functions  $g$  and  $h$  such that  $S = g(T)$  and  $T = h(S)$ . *→ ie  $T$  is a transf. of  $S$  and vice-versa*

# Sampling Theory IV

- Theorem: Let  $Y = (Y_1, \dots, Y_n)$  have joint density or frequency function  $f(y; \theta)$  and let  $T = T(Y)$  be a statistic. Suppose that  $f(y; \theta)/f(z; \theta)$  is independent of  $\theta$  if and only if  $T(y) = T(z)$ . Then  $T$  is minimally sufficient for  $\theta$ . ?
- Proof: Assume for simplicity that  $f(y; \theta) > 0$  for all  $y \in \mathbb{R}^n$  and  $\theta \in \Theta$ . Let  $\mathcal{T} = \{T(u) : u \in \mathbb{R}^n\}$  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  $w_t \in A_t$ . Notice that for any  $y$   $w_{T(y)}$  is in the same level set as  $y$  so that

$$f(y; \theta)/f(w_{T(y)}; \theta),$$

does not depend on  $\theta$  by assumption. Let  $g(t, \theta) \equiv f(w_t; \theta)$  and notice that

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

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  $\exists g', h' : f(y; \theta) = g'(T'(y); \theta)h'(y)$ .

Let  $y$  and  $z$  be such that  $T'(y) = T'(z)$ . Then

$$\frac{f(y; \theta)}{f(z; \theta)} = \frac{g'(T'(y); \theta)h'(y)}{g'(T'(z); \theta)h'(z)} = \frac{h'(y)}{h'(z)}.$$

Since the ratio does not depend on  $\theta$ , we have by assumption  $T(y) = T(z)$ . Hence  $T$  is a function of  $T'$  so is minimal by an arbitrary choice of  $T'$ .

- Example (Bernoulli trials). Let  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$ . Consider  $z$  and  $y$ , two possible distinct outcomes. Then we may note that

$$\frac{f(z; \theta)}{f(y; \theta)} = \frac{\theta^{\sum z_i} (1 - \theta)^{n - \sum z_i}}{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}}.$$

This ratio is constant only if  $\sum z_i = \sum y_i = T(y)$ . We may note that  $T$  is minimally sufficient.

# Sampling Theory VI

- Example: minimal sufficiency for the  $k$ -parameter exponential family.

An iid sample  $(Y_1, \dots, Y_n)^T$  from an exponential family has joint distribution

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

where

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

?

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

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



# 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  $\phi$  and this implies minimal sufficiency of  $\tau$ .

# Sampling Distributions

- By studying sampling distributions we aim to determine what different information do different forms of  $T$  carry about  $\theta$ .
- Theorem: (Sampling Distributions of Gaussian Sufficient Statistics).

Let  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$  and define

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

sample mean      sample variance       $\xrightarrow{\text{population mean/variance}}$

- The pair  $(\bar{Y}, S^2)$  are minimally sufficient for  $(\mu, \sigma^2)$  and
  - The sample mean has distribution  $\bar{Y} \sim N(\mu, \sigma^2/n)$ ,
  - The random variables  $\bar{Y}$  and  $S^2$  are independent,
  - The random variable  $S^2$  satisfies  $\frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2$ .
- Corollary: (Moments of Sufficient Statistics).

If  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$  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  $(Z_1, \dots, Z_k)$  be iid  $N(0, 1)$  random variables. Then

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

↗ chi-square with  
 k degrees of freedom

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

$$E(X) = k, \quad \text{Var}(X) = 2k, \quad M(t) = (1 - 2t)^{-k/2}, \quad t < 1/2.$$

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

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