STA261 - Module 1

Statistics

Rob Zimmerman

University of Toronto

July 2-4, 2024

Data and samples

- *Data* is factual information collected for the purposes of inference (Merriam-Webster)
- *Inference* is the act of passing from statistical sample of data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty (also Merriam-Webster)
- We collect a *sample* of data from a *population* associated with some probability distribution, and we would like to infer unknown properties of that distribution
- Example 1.1: Maybe...

- Height of STA261 student is well-approximated by $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$

- Indicator (0 or 1) that an Ontario high school student goes on to university is Bernoulli(p), $p \in (0,1)$

- Number of defective parts produced by a factory is Poisson(λ), $\lambda > 0$

Random variables versus observed data (this is really important)

- Our data sample goes through two phases of life: first as a *random sample*, and then as *observed data*

- A random sample is a set of *random variables*; observed data is a set of *constants*; the same goes for functions thereof If $Z \sim N(0,1)$, then $P(Z>0) = 1/2$.
  If we observe $Z = z$, then $P(z>0) = \mathbb{1}_{z>0} \in \{0,1\}$.

- We denote random variables using uppercase letters, and constants using lowercase letters:

- Example 1.2: $\vec{X}_n = (X_1, \ldots, X_n)$ = vector of heights of STA261 students before measuring (random vector!)

  $\vdots$ (we measure)

  $\vec{x}_n = (x_1, \ldots, x_n)$ = vector of measured heights (constant!)

- It is **very** important to clearly distinguish between the two quantities. But why?

iid-ness

- "iid" stands for "**independent and identically distributed**"
- This term is used everywhere in statistics, because it saves a lot of time
- Instead of writing "Let $(X_1, \ldots, X_n)$ be a vector of independent random variables (ie, a <u>random sample</u>) each distributed according to the same pdf/pmf f_θ"

we write "Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$"

Or sometimes F_θ

- Eg: let $Z_1, \ldots, Z_n \overset{iid}{\sim} N(\mu, 1)$ $\underbrace{\text{for } \mu \in \mathbb{R}}_{\text{i.e., } \mu \text{ is unknown}}$.

Statistics

- Definition 1.1: A **statistic** $T(\mathbf{X})$ is a function of the random data sample $\mathbf{X}$ which is free of any unknown constants. If we observe $\mathbf{X} = \mathbf{x}$, then $T(\mathbf{x})$ is the **observed value of** T.

- Example 1.3: $T(\vec{X}) = \bar{X}_n = \frac{1}{n}\sum_i X_i$

  $T(\vec{X}) = 24 \cdot X_2^3$

  $T(\vec{X}) = X_{(n)}$

  $T(\vec{X}) = (\bar{X}_n, S_n^2)$

  Say $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$. Then $T(\vec{X}) = \bar{X}_n - \mu$ is **NOT** a statistic (because we don't know μ)

- A statistic is useful when it allows us to summarize the data sample in ways that helps us with inference

- Different statistics are useful for different models

- Example 1.4: Say $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

  Intuitively, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ could help us "understand" μ.

  For example, $\mathbb{E}[\bar{X}_n] = \mu$. Also, $\bar{X}_n \xrightarrow{P} \mu$ by the WLLN.

Similarly, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ could help us with σ^2. $\mathbb{E}[S_n^2] = \sigma^2$, etc..

Slight abuse of notation: for iid samples, we'll usually write $f_\theta(\vec{x})$ for the joint pdf/pmf of $\vec{X} = (X_1, \dots, X_n)$ and $f_\theta(x)$ for the pdf/pmf of each X_i.

For example, if $X_1, \dots, X_n \overset{iid}{\sim}$ Poisson(λ), $\lambda > 0$,

then $f_\lambda(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

and $f_\lambda(\vec{x}) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} \dfrac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \left(\dfrac{1}{\prod_{i=1}^{n} x_i!}\right) \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$

$$\left(\begin{aligned} f_\theta &: \mathcal{X} \to [0, \infty) \\ f_\theta &: \mathcal{X}^n \to [0, \infty) \end{aligned} \right)$$

Parameters and Statistical Models

All parameters

- Many classical probability distributions have *parameters* associated with them
- Example 1.5: $N(\mu, \sigma^2)$, $\text{Bin}(n, p)$, $\text{Poisson}(\lambda)$

Capital Θ

- Definition 1.2: A **statistical model** is a set of pdfs/pmfs $\{f_\theta(\cdot) : \theta \in \Theta\}$ defined on the same sample space, where each θ is a fixed **parameter** in a known **parameter space** Θ. When $\Theta \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}$, the set is also called a **parametric model** (or **parametric family**).
- Example 1.6: $\left\{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) : \mu \in \mathbb{R}\right\}$

  $\{\lambda e^{-\lambda x} : \lambda > 0\}$

  Say f_θ is a pdf for any θ.

  If we know $\Theta = \{\theta_1, \theta_2\}$, then $\{f_\theta : \theta \in \{\theta_1, \theta_2\}\}$

- Statistical inference is classically concerned with figuring out which one of those distributions generated the data, based on the data sample we have available
- This amounts to inferring the particular parameter θ

Parameters and Statistical Models: More Examples

- Example 1.7: We generally write θ for the unknown parameter of interest (which may be a vector!)

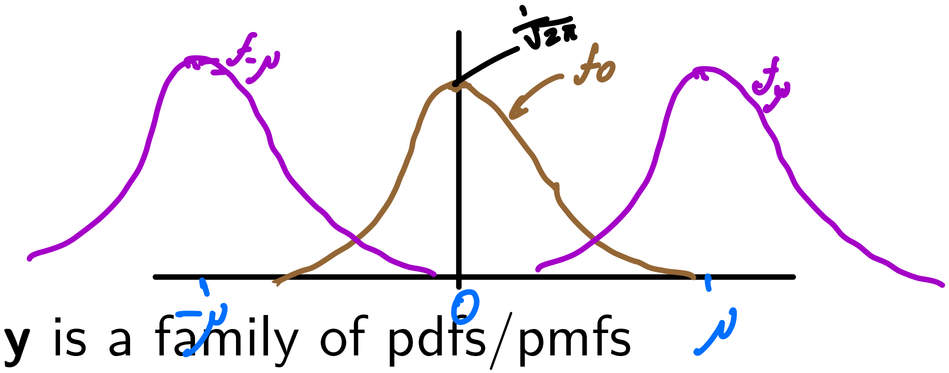
eg: $\{N(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\} = \{N(\mu,\sigma^2) : \theta = (\mu,\sigma^2) \in \underbrace{\mathbb{R}\times(0,\infty)}_{\Theta}\}$

Maybe we know in advance that $\mu > 0$. Then maybe our parametric family is

$$\{N(\mu,\sigma^2) : (\mu,\sigma^2) \in (0,\infty)^2\}$$

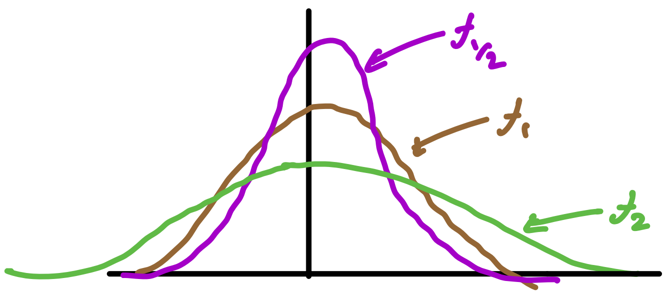
Important Parametric Families: Location-Scale Families

- Definition 1.3: A **location family** is a family of pdfs/pmfs $\{f_\mu(\cdot) = f(\cdot - \mu) : \mu \in \mathbb{R}\}$ formed by translating a "standard" family member $f(\cdot) := f_0(\cdot)$.



- Example 1.8: $\{N(\mu, 1) : \mu \in \mathbb{R}\}$

- Definition 1.4: A **scale family** is a family of pdfs/pmfs $\{f_\sigma(\cdot) = f(\cdot/\sigma)/\sigma : \sigma > 0\}$ formed by rescaling a "standard" family member $f(\cdot) := f_1(\cdot)$.



- Example 1.9: $\{N(0, \sigma^2) : \sigma^2 > 0\}$

- Definition 1.5: A **location-scale family** is a family of pdfs/pmfs $\{f_{\mu,\sigma}(\cdot) = f\left(\frac{\cdot - \mu}{\sigma}\right)/\sigma : \mu \in \mathbb{R}, \sigma > 0\}$ formed by translating and rescaling a "standard" family member $f(\cdot) := f_{0,1}(\cdot)$.

- Example 1.10: $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$

Poll Time!

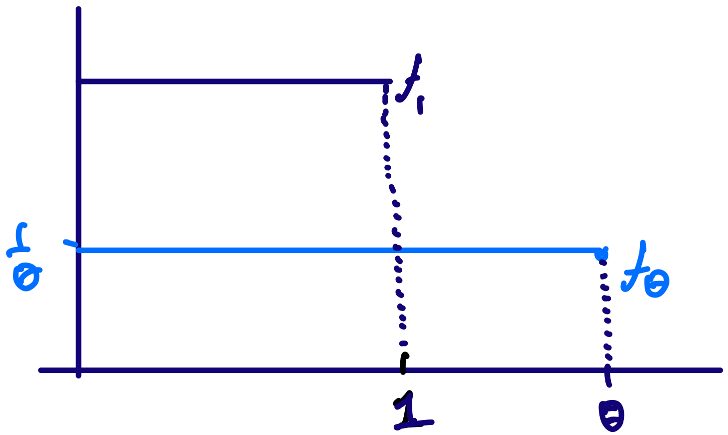
$X \sim \text{Cauchy}(\mu, \sigma^2) \iff f_{\mu,\sigma^2}(x) = \frac{1}{\pi}\left(\frac{\sigma}{(x-\mu)^2+\sigma^2}\right)$. Check: location-scale family.

but $\mathbb{E}[X]$ is undefined (EXERCISE!)

On Quercus: Module 1 - Poll 1

$\text{Unif}(0, \theta)$.

$f_\theta(x) = \frac{1}{\theta} \cdot \mathbb{1}_{x \in (0,\theta)}$



f_1 and f_θ are not shifts of one another!

So not a location family...

Important Parametric Families: Exponential Families

- Definition 1.6: An **exponential family** is a parametric family of pdfs/pmfs of the form

$$f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \cdot T_j(x)\right),$$

Usually $k=1$ when $\Theta = \mathbb{R}$, whence $f_\theta(x) = h(x) \cdot g(\theta) \cdot e^{\eta(\theta) \cdot T(x)}$

for some $k \in \mathbb{N}$, where all functions of x and θ are *known* and the support of f_θ does not depend on θ.

- Lots of theory simplifies considerably if we assume our random sample comes from an exponential family

- Many of your favourite distributions are included

$\text{Bin}(n,p)$, n known; $N(\mu, \sigma^2)$; $\text{Gamma}(a,b)$; $\text{Beta}(\alpha,\beta)$; $\text{Exp}(\lambda)$; $\chi^2_{(n)}$; $\text{Multinomial}(n; p_1, \ldots, p_k)$, n known

EXERCISE: Show!

- Example 1.11:

$X \sim \text{Exp}(\lambda)$, $\lambda > 0$.

$$f_\lambda(x) = \lambda e^{-\lambda x} = \underbrace{1}_{h(x)} \cdot \underbrace{\lambda}_{g(\lambda)} \cdot \exp(\underbrace{-\lambda}_{\eta(\lambda)} \cdot \underbrace{x}_{T(x)})$$

$X \sim \text{Bernoulli}(\theta)$, $\theta \in (0,1)$.

$$f_\theta(x) = \theta^x (1-\theta)^{1-x}$$
$$= 1 \cdot (1-\theta) \cdot \left(\frac{\theta}{1-\theta}\right)^x = \underbrace{1}_{h(x)} \cdot \underbrace{(1-\theta)}_{g(\theta)} \cdot \exp\left(\underbrace{x}_{T(x)} \cdot \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\eta(\theta)}\right)$$

Note: the $g, h, \eta_1, \dots, \eta_k, T_1, \dots, T_k$ are <u>not</u> unique, in general!

For example, if $f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp\left(T(x) \cdot \eta(\theta)\right)$

then for any $c \neq 0$, $\tilde{T}(x) = \frac{1}{c} T(x)$

and $\tilde{\eta}(\theta) = c \cdot \eta(\theta)$

give us the <u>same</u> family!

<u>FYI</u>: if $\eta(\theta) = \theta$, the family is said to be in <u>canonical form</u>.

used in STA303

If $\eta(\theta) = \theta$ <u>and</u> $T(x) = x$, the family is said to be a

<u>natural exponential family</u>.

$x = \exp(\log(x)), \ x > 0$

A Quick Review of Conditional Distributions

- Remember Bayes' rule: $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Conditional distributions and expectations $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}, \quad f_Y(y) > 0$
- For any fixed y, $\mathbb{E}[X|Y = y]$ is a constant
- But $\mathbb{E}[X|Y]$ is a *random variable*

$X \perp\!\!\!\perp Y$ means "X and Y are independant"

- Example 1.12: $\mathbb{E}[X|X] = X$. $\mathbb{E}[X|X=x] = x$.
  Say $X \perp\!\!\!\perp Y$. $\mathbb{E}[X|Y] = \mathbb{E}[X] = \mathbb{E}[X|Y=y]$
- Example 1.13:

"Tower property"/"Law of total expectation": $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

"Law of total variance": $\mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}(X|Y)) = \mathrm{Var}(Y)$

EXERCISE: prove these if you haven't seen them!

A Quick Review of Functions

domain codomain

- Let $f : A \to B$ be a function

- If f is one-to-one, then $f(a) = f(b) \iff a = b$

  "injective"

- If f is onto, then $\forall b \in B, \exists a \in A$ s.t. $b = f(a)$

  "surjective"

- If f is a bijection, then f is one-to-one <u>and</u> onto (and hence admits an inverse $f^{-1}: B \to A$ which is also a bijection)

- Example 1.14:

Freedom From θ

- Most of the functions $f_\theta(x)$ we will deal with have parameters involved in addition to the "independent variable"

- If the parameter θ can vary too, then $f_\theta(x)$ is really a function of both x and θ
  i.e., there exists $g : \Theta \times \mathbb{R} \to [0, \infty)$ such that $g(\theta, x) = f_\theta(x)$ $\forall \theta \in \Theta, x \in \mathcal{X}$

- If $f_\theta(x)$ is actually *not* a function of θ (i.e., it's constant with respect to θ), we might also say that it's "free of θ" or that it "does not depend on θ"

- Example 1.15:
  $f_\theta(x) = x^2$ is free of Θ.
  $f_\theta(x) = \Theta e^{-\Theta x}$ is not free of Θ.
  Say $X \sim N(\mu, 1)$. Then
  $P_\mu(X - \mu \leq x) = \Phi(x)$ is free of μ.

- So if we say that the distribution of X is free of θ, we mean that the cdf of X (and hence the pdf/pmf) is the same for all $\theta \in \Theta$

- Example 1.16: If $X \sim Exp(\lambda)$, then the distribution of λX is free of λ.

  EXERCISE!

Data Reduction: A Thought Experiment

- Is there a such thing as "more data than necessary"?

- Suppose that field researchers collect a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n) \overset{iid}{\sim} f_\theta$, where n is astronomically large; they want us statisticians to do inference on θ, but sending us $\mathbf{X}$ would take weeks

- Wouldn't it be great if we didn't need the entire sample $\mathbf{X}$ to make inferences about θ, but rather a much smaller statistic $T(\mathbf{X})$ – perhaps just a single number – that still contained as much information about θ as $\mathbf{X}$ itself did?

- The researchers observe $\mathbf{X} = \mathbf{x}$, calculate $T(\mathbf{x}) = t$ on their end, and then text t over to us

- Example 1.17: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$. Instead of the experimenters sending us $\vec{x}_n = (x_1, \ldots, x_n) \in \mathbb{R}^n$, what if they just sent us $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i \in \mathbb{R}$?

  Can we still learn something about μ?

Sufficiency

- How do we "encode" this idea?
- If we know that $T(\mathbf{X}) = t$, then there should be nothing else to glean from the data about θ
- Definition 1.7: A statistic $T(\mathbf{X})$ is a **sufficient statistic** for a parameter θ if the conditional distribution of $\mathbf{X} \mid T(\mathbf{X}) = t$ does not depend on θ.
- An interpretation: if the conditional distribution

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) = \frac{\overbrace{\mathbb{P}_\theta\left(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})\right)}^{= \mathbb{P}_\theta(\vec{X} = \vec{x})}}{\mathbb{P}_\theta\left(T(\mathbf{X}) = T(\mathbf{x})\right)}$$

is really free of θ, then the information about θ in $\mathbf{X}$ and the information about θ in $T(\mathbf{X})$ and "cancel each other out" (heavy quotes here)

- Example 1.18: $T(\vec{X}) = \vec{X}$ is <u>always</u> sufficient for whatever parameter,

Why? Because $\mathbb{P}_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = T(\vec{x})) = \mathbb{P}_\theta(\vec{X} = \vec{x} \mid \vec{X} = \vec{x}) = 1$, which is free of θ

Sufficiency

- Example 1.19: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Bernoulli (θ), where $\theta \in (0,1)$. Show that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

$T(\vec{X}) \sim Bin(n, \theta)$.

$\mathbb{P}_\theta(T(\vec{X}) = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$, $t \in \{0, \dots, n\}$.

Let $t = T(\vec{x})$. Then

$\mathbb{P}_\theta(\vec{X} = \vec{x} \wedge T(\vec{X}) = t)$

$= \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^{n} X_i = t)$

$= \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^{n} x_i = t)$

$= \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = t - \sum_{i=1}^{n-1} x_i)$

$= \mathbb{P}_\theta(X_1 = x_1) \cdot \dots \cdot \mathbb{P}_\theta(X_n = t - \sum_{i=1}^{n-1} x_i)$

$= \theta^{x_1}(1-\theta)^{1-x_1} \cdot \dots \cdot \theta^{t - \sum_{i=1}^{n-1} x_i}(1-\theta)^{1-t+\sum_{i=1}^{n-1} x_i}$

$= \theta^t (1-\theta)^{n-t}$

So $\mathbb{P}_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = t)$

$= \dfrac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$

$= \dfrac{1}{\binom{n}{t}} = \dfrac{1}{\binom{n}{\sum_{i=1}^{n} x_i}}$ is free of θ.

$\therefore T(\vec{X})$ is sufficient for θ.

Sufficiency

- Example 1.20: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Show that the sample mean $T(\mathbf{X}) = \bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient for μ. $T(\vec{X}) \sim N(\mu, \sigma^2/n)$ has pdf $f_T(t) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right)$.

Let $t = \frac{1}{n}\sum x_i = \bar{x}_n$. Then

$$\sum_{i=1}^n (x_i-\mu)^2 = \sum_i (x_i - t + t - \mu)^2$$
$$= \sum_i \left[(x_i-t)^2 - 2(x_i-t)\cdot(t-\mu) + (t-\mu)^2 \right]$$
$$= \sum_i (x_i-t)^2 + n\cdot(t-\mu)^2$$

So $f_{\vec{X}}(\vec{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum_i (x_i-t)^2}{2\sigma^2} - \frac{n(t-\mu)^2}{2\sigma^2}\right)$$
$$= f_{\vec{X},T}(\vec{x}, t)$$

So

$$f_{\vec{X}|T}(\vec{x}|t) = \frac{(2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum_i (x_i-t)^2}{2\sigma^2} - \cancel{\frac{n(t-\mu)^2}{2\sigma^2}}\right)}{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\cancel{\frac{n(t-\mu)^2}{2\sigma^2}}\right)}$$

$$= \frac{1}{\sqrt{n}(2\pi\sigma^2)^{\frac{n-1}{2}}} \cdot \exp\left(-\frac{\sum (x_i-t)^2}{2\sigma^2}\right)$$

is free of μ.

$\therefore T(\vec{X})$ is sufficient for μ.

The Factorization Theorem

Fisher-Neyman

- Theorem 1.1 (**Factorization theorem**): Let $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta(\mathbf{x})$, where $f_\theta(\mathbf{x})$ is a joint pdf/pmf. A statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g_\theta(t)$ and $h(\mathbf{x})$ such that

$$f_\theta(\mathbf{x}) = h(\mathbf{x}) \cdot g_\theta(T(\mathbf{x})) \quad \text{for all } \theta \in \Theta,$$

where $h(\mathbf{x})$ is free of θ and $g_\theta(T(\mathbf{x}))$ only depends on $\mathbf{x}$ through $T(\mathbf{x})$.

- In other words, $T(\mathbf{X})$ is sufficient whenever the "part" of $f_\theta(\mathbf{x})$ that actually depends on θ is a function of $T(\mathbf{x})$, rather than $\mathbf{x}$ itself

Continuous case needs measure theory

Proof. (Discrete case)

Let $t = T(\vec{x})$.

We want to show $\dfrac{\mathbb{P}_\theta(\vec{X} = \vec{x} \wedge T(\vec{X}) = t)}{\mathbb{P}_\theta(T(\vec{X}) = t)}$ is free of θ iff

$\mathbb{P}_\theta(\vec{X} = \vec{x}) = h(\vec{x}) \cdot g_\theta(t)$,

for some $h(\cdot)$, $g_\theta(\cdot)$.

The Factorization Theorem

($\Longrightarrow$) Assume T is sufficient for θ. If $t = T(\vec{x})$, then

$$\mathbb{P}_\theta(\vec{X} = \vec{x}) = \mathbb{P}_\theta(\vec{X} = \vec{x} \wedge T(\vec{X}) = t)$$

$$= \underbrace{\mathbb{P}_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = t)}_{=: h(\vec{x}) \text{ which is free of } \theta \text{ b/c } T \text{ is sufficient}} \cdot \underbrace{\mathbb{P}_\theta(T(\vec{X}) = t)}_{=: g_\theta(t)} = h(\vec{x}) \cdot g_\theta(t).$$

($\Longleftarrow$): Assume $\mathbb{P}_\theta(\vec{X} = \vec{x}) = h(\vec{x}) \cdot g_\theta(t)$ for some $h(\cdot)$, $g_\theta(\cdot)$. Then, if $A_t = \{\vec{x} : T(\vec{x}) = t\}$,

$$\mathbb{P}_\theta(T(\vec{X}) = t) = \sum_{\vec{x} \in A_t} \mathbb{P}(\vec{X} = \vec{x} \wedge T(\vec{X}) = t) \text{ by the law of total probability}$$

$$= \sum_{\vec{x} \in A_t} \mathbb{P}(\vec{X} = \vec{x})$$

$$= \sum_{\vec{x} \in A_t} h(\vec{x}) \cdot g_\theta(t) \text{ by assumption}$$

$$= \left(\sum_{\vec{x} \in A_t} h(\vec{x})\right) \cdot g_\theta(t)$$

Then

$$\mathbb{P}_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = t) = \frac{h(\vec{x}) \cdot g_\theta(t)}{\left(\sum_{\vec{x} \in A_t} h(\vec{x})\right) \cdot g_\theta(t)} = \frac{h(\vec{x})}{\sum_{\vec{x} \in A_t} h(\vec{x})} \text{ is free of } \theta.$$

$\square$

Poll Time!

On Quercus: Module 1 - Poll 2

$$\frac{f_{\theta_1}(\vec{x})}{f_{\theta_2}(\vec{x})} = \frac{h(\vec{x}) \cdot g_{\theta_1}(T(\vec{x}))}{h(\vec{x}) \cdot g_{\theta_2}(T(\vec{x}))} = \frac{g_{\theta_1}(T(\vec{x}))}{g_{\theta_2}(T(\vec{x}))}$$ depends on $\vec{x}$ only through $T(\vec{x})$

The Factorization Theorem: Examples

- Example 1.21: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ), where $\theta \in (0,1)$. Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is sufficient for θ.

Let $t = \sum_{i=1}^n x_i$. Then $f_\theta(\vec{x}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$

$$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$= \underbrace{1}_{=:h(\vec{x})} \cdot \underbrace{\theta^t(1-\theta)^{n-t}}_{g_\theta(t)}$$

By the factorization theorem, $T(\vec{X})$ is sufficient for θ.

- Example 1.22: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Show that the sample mean $T(\mathbf{X}) = \bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient for μ.

$f_\mu(\vec{x}) = \prod_{i=1}^n f_\mu(x_i)$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum(x_i-t)^2}{2\sigma^2} - \frac{n(t-\mu)^2}{2\sigma^2}\right)$$

$$= \underbrace{(2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left(-\frac{\sum(x_i-t)^2}{2\sigma^2}\right)}_{=:h(\vec{x})} \cdot \underbrace{\exp\left(\frac{-n(t-\mu)^2}{2\sigma^2}\right)}_{=:g_\mu(t)}.$$

By the factorization theorem, $T(\vec{x})$ is sufficient for μ.

The Factorization Theorem: Examples

- Example 1.23: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ is sufficient for $(\mu, \sigma^2) = \theta$.

Let $t_1 = \bar{x}_n$ and $t_2 = \frac{1}{n-1} \sum_i (x_i - t_1)^2$. Then

$$f_\theta(\vec{x}) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum(x_i - t_1)^2}{2\sigma^2} - \frac{n(t_1-\mu)^2}{2\sigma^2}\right)$$

$$= \underbrace{1}_{h(\vec{x})} \cdot \underbrace{(2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{-(n-1)\cdot t_2}{2\sigma^2} - \frac{n(t_1-\mu)^2}{2\sigma^2}\right)}_{g_\theta(t_1, t_2)}.$$

By the factorization theorem, $T(\vec{x}) = (T_1(\vec{x}), T_2(\vec{x}))$ is sufficient for θ.

- Example 1.24: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$ where $\theta > 0$. Show that $\bar{X}_n$ is *not* sufficient for θ, and find a statistic that is.

$$f_\theta(\vec{x}) = \prod_{i=1}^n \frac{1}{\theta} \cdot \mathbb{1}_{0 \le x_i \le \theta}$$

$$= \theta^{-n} \cdot \mathbb{1}_{0 \le x_i \le \theta \ \forall i}$$

$$= \theta^{-n} \cdot \mathbb{1}_{x_{(1)} \ge 0 \wedge x_{(n)} \le \theta}$$

$$= \underbrace{\mathbb{1}_{x_{(1)} \ge 0}}_{=: h(\vec{x})} \cdot \underbrace{\theta^{-n} \cdot \mathbb{1}_{x_{(n)} \le \theta}}_{g_\theta(x_{(n)})}.$$

By the factorization theorem, $T(\vec{x}) = X_{(n)}$ is sufficient for θ (and $\bar{X}_n$ is <u>not</u>).

The Factorization Theorem: Examples

- Theorem 1.2: Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$ be a random sample from an exponential family, where

$$f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \cdot T_j(x)\right).$$

Then $T(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right)$ is sufficient for θ.

Proof. Important EXERCISE!

The Factorization Theorem: Examples

- Example 1.25: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that $T(\mathbf{X}) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is sufficient for $(\mu, \sigma^2) = \theta$.

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

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

$$= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \cdot \underbrace{\frac{1}{\sigma} \exp\left(\frac{-\mu^2}{2\sigma^2}\right)}_{g(\mu,\sigma^2)} \cdot \exp\left(\underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\mu,\sigma^2)} \cdot \underbrace{x^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{\eta_2(\mu,\sigma^2)} \cdot \underbrace{x}_{T_2(x)}\right)$$

By Theorem 1.2, $T(\vec{X}) = \left(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i)\right)$

$= \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is sufficient for (μ, σ^2).

The Factorization Theorem: Examples

- Example 1.26: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Unif$(\{1, 2, \ldots, \theta\})$, where $\theta \in \mathbb{N}$. Show that $T(\mathbf{X}) = X_{(n)}$ is sufficient for θ. Not an exponential family!

$$f_\theta(\vec{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} \cdot \mathbb{1}_{x_i \in \{1,\ldots,\theta\}}$$

$$= \theta^{-n} \cdot \mathbb{1}_{x_i \in \{1,\ldots,\theta\}\ \forall i}$$

$$= \underbrace{\mathbb{1}_{x_{(1)} \in \mathbb{N}}}_{=: h(\vec{x})} \cdot \underbrace{\theta^{-n} \cdot \mathbb{1}_{x_{(n)} \le \theta}}_{=: g_\theta(x_{(n)})}.$$

By the factorization theorem, $T(\vec{X})$ is sufficient for θ.

Indicator arithmetic: let P_1 and P_2 be two propositions (ie, either T or F). Let A_1, A_2 be sets.

$\mathbb{1}_P := \begin{cases} 0, & P \text{ is false} \\ 1, & P \text{ is true.} \end{cases}$

$\mathbb{1}_A(x) := \begin{cases} 0, & x \notin A \\ 1, & x \in A. \end{cases}$

Then...

$\mathbb{1}_{P_1 \text{ AND } P_2} = \mathbb{1}_{P_1} \cdot \mathbb{1}_{P_2}$ $\quad \mathbb{1}_{P_1 \text{ OR } P_2} = \mathbb{1}_{P_1} + \mathbb{1}_{P_2} - \mathbb{1}_{P_1} \cdot \mathbb{1}_{P_2}$

$\mathbb{1}_{A \cap B}(x) = \mathbb{1}_A(x) \cdot \mathbb{1}_B(x)$, $\quad \mathbb{1}_{A \cup B}(x) = \mathbb{1}_A(x) + \mathbb{1}_B(x) - \mathbb{1}_A(x) \cdot \mathbb{1}_B(x)$,

$\mathbb{1}_{\text{NOT } P} = 1 - \mathbb{1}_P$

$\mathbb{1}_{A^c}(x) = 1 - \mathbb{1}_A(x)$ $\quad$ etc, etc.

If There's One, There's More...

- If we have some sufficient statistic, we can always come up with (infinitely) many others...

- Theorem 1.3: Let $T(\mathbf{X})$ be sufficient for θ and suppose that $r(\cdot)$ is a bijection. Then $r(T(\mathbf{X}))$ is also sufficient for θ.

Proof. Say that $T(\vec{x})$ is sufficient for θ.

By the factorization theorem, $f_\theta(\vec{x}) = h(\vec{x}) \cdot g_\theta(T(\vec{x}))$ for some functions $h(\cdot)$ and $g_\theta(\cdot)$.

$$
\begin{aligned}
f_\theta(\vec{x}) &= h(\vec{x}) \cdot g_\theta(T(\vec{x})) \\
&= h(\vec{x}) \cdot g_\theta(r^{-1}(r(T(\vec{x})))) \\
&= h(\vec{x}) \cdot \tilde{g}_\theta(r(T(\vec{x}))), \text{ where } \tilde{g}_\theta = g_\theta \circ r^{-1} \ (\text{i.e., } \tilde{g}_\theta(t) = g_\theta(r^{-1}(t))).
\end{aligned}
$$

By the factorization theorem, $r(T(\vec{x}))$ is sufficient for θ. $\square$

in the other direction!

Too Many Sufficient Statistics

- So there are lots of sufficient statistics out there
- We saw that $T(\mathbf{X}) = \mathbf{X}$ is always sufficient – it's also pretty useless as far as data reduction goes
- There are usually "better" ones out there – how do we get the best bang for our buck?
- Another issue: the factorization theorem makes it easy to show that a statistic is sufficient (if it actually is), but less so to show that a statistic is *not* sufficient
- We will develop theory that takes care of both of these issues at once

Minimal Sufficiency

- Definition 1.8: A sufficient statistic $T(\mathbf{X})$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $U(\mathbf{X})$, there exists a function h such that $T(\mathbf{X}) = h(U(\mathbf{X}))$.

- In other words, a minimal sufficient statistic is some function of *any other sufficient statistic*

Eg: $X_1, \ldots, X_n \stackrel{iid}{\sim} N(\mu, 1)$. We know that $T_1(\vec{X}) = \vec{X}$ is sufficient for μ.
So is $T_2(\vec{X}) = \bar{X}_n$. T_2 is a function of T_1,
but not the other way around!

- A minimal sufficient statistic achieves the greatest reduction of data possible (while still maintaining sufficiency)

- Example 1.27: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Show that $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ is not minimal sufficient for μ.

We saw that $\bar{X}_n$ is sufficient for μ. But $T(\vec{X})$ is not a function of $\bar{X}_n$.

So it can't be minimal sufficient for μ.

Poll Time!

On Quercus: Module 1 - Poll 3

If $X_1 + \cdots + X_{n-1} + X_n$ is minimal sufficient

$\Rightarrow X_1 + \cdots + X_{n-1}$ NOT minimal sufficient

A Criterion For Minimal Sufficiency

- It's usually not that hard to show that a statistic is not minimal sufficient

- But how can we possibly show that a statistic *is* minimal?

- Theorem 1.4: Let $f_\theta(\mathbf{x})$ be the pdf/pmf of a sample $\mathbf{X}$. Suppose there exists a function $T(\cdot)$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$, $T(\mathbf{x}) = T(\mathbf{y})$ if and only if the ratio $f_\theta(\mathbf{x})/f_\theta(\mathbf{y})$ is free of θ. Then $T(\mathbf{X})$ is minimal sufficient for θ.

No proof...

- This criterion is easier to apply than it looks

- Example 1.28: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ), where $\theta \in (0,1)$. Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is minimal sufficient for θ.

Let $\vec{x}, \vec{y} \in \mathcal{X}^n = \{0,1\}^n$. Then...

$$\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}} = \theta^{\sum x_i - \sum y_i} \cdot (1-\theta)^{\sum y_i - \sum x_i},$$

which is free of θ iff $\sum_i x_i = \sum_i y_i$ (i.e., $T(\vec{x}) = T(\vec{y})$).

By Theorem 1.4, $T(\vec{X})$ is minimal sufficient for θ.

Minimal Sufficiency: Examples

- Example 1.29: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ is minimal sufficient for $(\mu, \sigma^2) = \theta$.

Let $s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$.

From Ex 1.23, $f_\theta(\vec{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-(n-1)s_x^2 - n(\bar{x}-\mu)^2}{2\sigma^2} \right)$. Then

$$\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})} = \frac{\cancel{(2\pi\sigma^2)^{-n/2}} \cdot \exp\left(\frac{-(n-1)s_x^2 - n(\bar{x}-\mu)^2}{2\sigma^2} \right)}{\cancel{(2\pi\sigma^2)^{-n/2}} \cdot \exp\left(\frac{-(n-1)s_y^2 - n(\bar{y}-\mu)^2}{2\sigma^2} \right)}$$

$$= \exp\left(\frac{-(n-1)(s_x^2 - s_y^2) - n(\bar{x}^2 - \bar{y}^2 - 2\mu(\bar{x}-\bar{y}))}{2\sigma^2} \right)$$

... is free of (μ, σ^2) iff $\bar{x} = \bar{y}$ AND $s_x^2 = s_y^2$.

By Theorem 1.4, $T(\vec{x})$ is minimal sufficient for (μ, σ^2).

Minimal Sufficiency: Examples

- Example 1.30: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Find a minimal sufficient statistic for λ.

EXERCISE!

Minimal Sufficiency: Examples

- A minimal sufficient statistic isn't always as minimal as you would expect...
- Example 1.31: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Unif$([\theta, \theta+1])$, where $\theta \in \mathbb{R}$. Show that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.

$$f_\theta(\vec{x}) = \prod_{i=1}^{n} f_\theta(x_i)$$
$$= \prod_{i=1}^{n} \mathbb{1}_{\theta \le x_i \le \theta+1}$$
$$= \mathbb{1}_{\theta \le x_1, \dots, x_n \le \theta+1}$$
$$= \mathbb{1}_{\theta \le x_{(1)} \wedge x_{(n)} \le \theta+1}$$
$$= \mathbb{1}_{x_{(n)}-1 \le \theta \le x_{(1)}}.$$

Let $\vec{x}, \vec{y} \in \mathcal{X}^n = [\theta, \theta+1]^n$. Then...

$$\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})} = \frac{\mathbb{1}_{x_{(n)}-1 \le \theta \le x_{(1)}}}{\mathbb{1}_{y_{(n)}-1 \le \theta \le y_{(1)}}} \text{ is free of } \theta$$

iff $x_{(n)} = y_{(n)}$ and $x_{(1)} = y_{(1)}$.

By Theorem 1.4, $T(\vec{x})$ is minimal sufficient for θ.

Poll Time!

On Quercus: Module 1 - Poll 4

The "Opposite" of Sufficiency?

- We know that a sufficient statistic contains all the information about θ that the original sample has
- What about a statistic that contains *no* information about θ?
- Why would such a thing be useful?

Eg: $X_1, X_2 \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$.

$T(\vec{X}) = (X_1 - \overline{X}_2, X_2 - \overline{X}_2)$ <u>might</u> not depend on μ.

Ancillarity

- Definition 1.9: A statistic $D(\mathbf{X})$ is an **ancillary statistic** for a parameter θ if the distribution of $D(\mathbf{X})$ does not depend on θ

- Example 1.32: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Unif$([\theta, \theta+1])$, where $\theta \in \mathbb{R}$. Show that the range statistic $R(\mathbf{X}) := X_{(n)} - X_{(1)}$ is ancillary for θ.

Let $Y_i = X_i - \theta$. Then $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}(0,1)$ and the distributions of $Y_{(1)}$ and $Y_{(n)}$ are free of θ.

Then $\mathbb{P}_\theta(R(\vec{X}) \leq r)$

$$= \mathbb{P}_\theta(X_{(n)} - X_{(1)} \leq r)$$

$$= \mathbb{P}_\theta((X_{(n)} - \theta) - (X_{(1)} - \theta) \leq r)$$

$$= \mathbb{P}_\theta(Y_{(n)} - Y_{(1)} \leq r)$$

$$= \mathbb{P}(\text{Beta}(n,1) - \text{Beta}(1,n) \leq r)$$

From Assignment 0

does not depend on θ. $\therefore R(\vec{X})$ is ancillary for θ.

Ancillarity: Examples

- Did we actually use the uniform distribution anywhere in the previous example?

- Theorem 1.5: Let $X_1, \dots, X_n$ be a random sample from a location family with cdf $F(\cdot - \theta)$, for $\theta \in \mathbb{R}$. Then the range statistic is ancillary for θ.

Proof. Let $Y_i = X_i - \theta \sim F(\cdot)$. ← Check!

Then $\mathbb{P}_\theta(R(\vec{X}) \le r)$

$= \mathbb{P}_\theta(X_{(n)} - X_{(1)} \le r)$

$= \mathbb{P}_\theta(Y_{(n)} - Y_{(1)} \le r)$, which is free of θ because the distributions of $Y_{(1)}$ and $Y_{(n)}$ are free of θ. $\square$

Ancillarity: Examples

$\sigma^2 > 0.$

- Example 1.33: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Show that $D(\mathbf{X}) = \frac{X_1 + \cdots + X_{n-1}}{X_n}$ is ancillary for σ^2.

Let $Z_i = X_i/\sigma$. Then $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$.

Then $\mathbb{P}_{\sigma^2}(D(\vec{X}) \leq x)$

$$= \mathbb{P}_{\sigma^2}\left(\frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} \leq x\right)$$

$$= \mathbb{P}_{\sigma^2}\left(\frac{X_1/\sigma}{X_n/\sigma} + \cdots + \frac{X_{n-1}/\sigma}{X_n/\sigma} \leq x\right)$$

$$= \mathbb{P}_{\sigma^2}\left(\frac{Z_1}{Z_n} + \cdots + \frac{Z_{n-1}}{Z_n} \leq x\right)$$ does not depend on σ^2.

So $D(\vec{X})$ is ancillary for σ^2.

- Theorem 1.6: Let $X_1, \ldots, X_n$ be a random sample from a scale family with cdf $F(\cdot/\sigma)$, for $\sigma > 0$. Then any statistic which is a function of the ratios $X_1/X_n, \ldots, X_{n-1}/X_n$ is ancillary for σ.

EXERCISE!

Ancillarity: Examples

- Recall that if $Z_1, \ldots, Z_n \overset{iid}{\sim} \mathcal{N}(0,1)$, then the distribution of $Y = \sum_{i=1}^n Z_i^2$ is called a **chi-squared distribution with n degrees of freedom**, which we write as $Y \sim \chi^2_{(n)}$.

- Theorem 1.7: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{(n-1)}$.

Proof ($n = 2$).

$X_1 - X_2 \sim N(0, 2\sigma^2)$

$\overset{d}{=} \sqrt{2}\sigma \cdot N(0,1)$.

$\Rightarrow (X_1 - X_2)^2 \overset{d}{=} (\sqrt{2}\sigma \cdot N(0,1))^2$

$= 2\sigma^2 \cdot \chi^2_{(1)}$

$(2-1) S_2^2 = \sum_{i=1}^{2} (X_i - \bar{X}_n)^2 = (X_1 - \frac{1}{2}(X_1 + X_2))^2 + (X_2 - \frac{1}{2}(X_1 - X_2))^2$

$= (\frac{1}{2}X_1 - \frac{1}{2}X_2)^2 + (\frac{1}{2}X_2 - \frac{1}{2}X_1)^2$

$= \frac{1}{2}(X_1 - X_2)^2$

$\overset{d}{=} \frac{1}{2} \cdot 2\sigma^2 \cdot \chi^2_{(1)}$

$= \sigma^2 \cdot \chi^2_{(1)}$

EXERCISE: General case. Use induction!

$\Rightarrow \frac{(2-1)}{\sigma^2} S_2^2 \overset{d}{=} \chi^2_{(2-1)}$. $\square$

- Example 1.34: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that the sample variance S_n^2 is ancillary for μ.

From above, $S_n^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{(n-1)}$, so its distribution is free of μ.

Poll Time!

On Quercus: Module 1 - Poll 5

Completeness: An Abstract Definition

- Everything so far has been about ways to reduce the amount of data we need while still retaining all information about θ
- We've seen that ancillary statistics are bad at it, sufficient statistics are good at it, and minimal sufficient statistics are very good at it
- We will study one more kind of statistic, but the definition isn't pretty
- Definition 1.10: A statistic $U(\mathbf{X})$ is **complete** if *any* function $h(\cdot)$ which satisfies $\mathbb{E}_\theta[h(U(\mathbf{X}))] = 0$ for all $\theta \in \Theta$ must also satisfy $\mathbb{P}_\theta(h(U(\mathbf{X})) = 0) = 1$ for all $\theta \in \Theta$.

$\mathcal{U}$ = range of $U(\cdot)$

continuous: $\int_{\mathcal{U}} h(u) \cdot f_\theta(u)\, du = 0$ where $U(\vec{X}) \sim f_\theta$

discrete: $\sum_{u \in \mathcal{U}} h(u) \cdot \mathbb{P}_\theta(U(\vec{X}) = u) = 0$

$\forall \theta \in \Theta$

If $U(\vec{X})$ is complete, then $\left(\mathbb{E}_\theta[h(U(\vec{X}))] = 0 \Rightarrow \mathbb{P}_\theta(h(U(\vec{X})) = 0) = 1 \ \forall \theta \right)$ is TRUE

Completeness: An Abstract Definition

- The concept of completeness is notoriously unintuitive – probably the most abstract one in our course – but it will pay off later

- For now, you can think about the finite case a bit like a finite-dimensional basis from linear algebra

- If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ span $\mathbb{R}^n$, then $\sum_{i=1}^{n} a_i \mathbf{v}_i = \mathbf{0}$ implies $a_i = 0$ for all i

  all of these span $\mathbb{R}^n$ coefficient

- If $U(\mathbf{X})$ is complete and supported on $\{u_1, \ldots, u_n\}$, then $\sum_{i=1}^{n} h(u_i) \cdot \underbrace{\mathbb{P}_\theta\left(U(\mathbf{X}) = u_i\right)}_{\text{all of these add up to 1}} = 0$ implies $\underbrace{h(u_i)}_{\text{"coefficient"}} = 0$ for all i

- The meaning will become clearer at the end of Module 2

- So why bring it up now?

Showing Completeness is Very Difficult In General...

- Example 1.35: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ) with $\theta \in (0,1)$. Show that $U(\mathbf{X}) = \sum_{i=1}^n X_i$ is complete. $U \sim Bin(n,\theta)$.

Suppose that $h(\cdot)$ is some function such that $\mathbb{E}_\theta[h(U(\vec{X}))] = 0 \ \forall \theta \in (0,1)$.

Then $0 = \sum_{j=0}^{n} h(j) \cdot \mathbb{P}_\theta(U=j)$

$= \sum_{j=0}^{n} h(j) \cdot \binom{n}{j} \theta^j (1-\theta)^{n-j}$

$= (1-\theta)^n \cdot \sum_{j=0}^{n} h(j) \cdot \binom{n}{j} \left(\frac{\theta}{1-\theta}\right)^j$

$\Rightarrow 0 = \sum_{j=0}^{n} h(j) \cdot \binom{n}{j} \cdot \left(\frac{\theta}{1-\theta}\right)^j$.

Let $r = \frac{\theta}{1-\theta} \in (0,\infty)$ and $\tilde{h}_j = h(j) \cdot \binom{n}{j}$

$= \sum_{j=0}^{n} \tilde{h}_j \cdot r^j$ $\quad \forall \theta \in (0,1)$ $\Leftrightarrow \forall r \in (0,\infty)$.

So we have a polynomial in r which is 0 for all r.

So that polynomial is the zero polynomial $\Rightarrow \tilde{h}_j = 0 \ \forall j$.

$\Rightarrow h(j) \cdot \binom{n}{j} = 0 \ \forall j$

$\Rightarrow h(j) = 0 \ \forall j$.

$\therefore \mathbb{P}_\theta(h(U(\vec{X})) = 0)$

$= \mathbb{P}_\theta(0=0)$

$= 1. \quad \forall \theta \in \Theta$.

$\therefore U(\vec{X})$ is complete.

...But for Exponential Families, There's Nothing To It

- Theorem 1.8: Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$ be a random sample from an exponential family, where

$$f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \cdot T_j(x)\right),$$

where each $\eta_j(\cdot)$ is continuous on Θ and each component of Θ contains an open interval in $\mathbb{R}$.[1] Then $T(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right)$ is a complete statistic. No proof.

- Recall from Theorem 1.2 that in this case, $T(\mathbf{X})$ is also sufficient for θ

- So it's really easy to find complete sufficient statistics for exponential families

[1] More generally, Θ must contain an open set in $\mathbb{R}^k$ – this requirement is sometimes called the "open set condition".

Completeness: Examples

- Example 1.36: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Show that $\bar{X}_n$ is complete for μ.

$$f_\mu(x) = (2\pi\sigma^2)^{-1/2} \cdot \exp\left(\frac{-(x^2 - 2\mu x + \mu^2)}{2\sigma^2}\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-x^2}{2\sigma^2}\right)}_{h(x)} \cdot \underbrace{\exp\left(\frac{-\mu^2}{2\sigma^2}\right)}_{g(\mu)} \cdot \exp\left(\overset{\eta(\mu)}{n\frac{\mu}{\sigma^2}} \cdot \underset{T(x)}{\frac{1}{n}x}\right)$$

$\eta(\mu) = n\mu/\sigma^2$ is clearly continuous. Also $\Theta = \mathbb{R}$ contains an open interval,

$T(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is complete by Theorem 1.8.

Completeness: Examples

- Example 1.37: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Show that $\bar{X}_n$ is complete for λ.

$\Theta = (0, \infty)$

$$f_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!} \cdot e^{-\lambda} \cdot \exp(x \cdot \log(\lambda))$$

$$= \underbrace{\frac{1}{x!}}_{h(x)} \cdot \underbrace{e^{-\lambda}}_{g(\lambda)} \cdot \exp\Big(\underbrace{\frac{1}{n}x}_{T(x)} \cdot \underbrace{n \cdot \log(\lambda)}_{\eta(\lambda).}\Big)$$

$\eta(\lambda) = n \cdot \log(\lambda)$ is continuous on $\Theta = (0, \infty)$. Also Θ contains an open interval.

By Theorem 1.8, $T(\vec{x}) = \bar{X}_n$ is complete.

Completeness: Examples

- Example 1.38: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_{\mu,\sigma}$ where

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right), \quad x \in \mathbb{R},$$

where $\sigma > 0$ and μ is known. Find a complete statistic for σ.

EXERCISE!

Complete Statistics Are Minimal Sufficient!

- There is nothing resembling sufficiency in the definition of completeness; the two concepts seem completely unrelated
- And yet, Theorem 1.8 says that for exponential families, certain complete statistics are sufficient
- What about in general? The answer might surprise you...
- Theorem 1.9 (**Bahadur's theorem**): A complete sufficient statistic is a minimal sufficient statistic. No proof...
- That's *not* the same as saying that all minimal sufficient statistics are complete (which is unfortunately not true)

Minimal Sufficient Statistics Are Not Always Complete

- Bahadur implies that if a minimal sufficient statistic exists and it's not complete, then no complete sufficient statistic exists

  Why? See the note following Slide 54

- This is probably the simplest example of a minimal sufficient statistic that is not complete

- Example 1.39: Let $X_1 \sim \text{Unif}(\theta, \theta+1)$, where $\theta \in \mathbb{R}$. Show that $T(X_1) = X_1$ is minimal sufficient for θ, but not complete.

$f_\theta(x) = \mathbb{1}_{\theta \le x \le \theta+1}$.

Let $x, y \in \mathcal{X}$. Then

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\mathbb{1}_{\theta \le x \le \theta+1}}{\mathbb{1}_{\theta \le y \le \theta+1}}$$

does not depend on θ iff $x = y$. By Theorem 1.4, $T(X_1)$ is minimal sufficient.

However, consider $h(x) = \sin(2\pi x)$. Then for all $\theta \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}_\theta[h(T(X_1))] &= \mathbb{E}_\theta[h(X_1)] \\
&= \mathbb{E}_\theta[\sin(2\pi X_1)] \\
&= \int_\theta^{\theta+1} \sin(2\pi x)\,dx \\
&= \frac{-\cos(2\pi x)}{2\pi}\Big|_\theta^{\theta+1} = \frac{-\cos(2\pi(\theta+1)) + \cos(2\pi\theta)}{2\pi} = 0.
\end{aligned}$$

But h is not identically 0 on $(\theta, \theta+1)$, so $P_\theta(h(X_1) = 0) \neq 1$

$\implies T$ is ~~not~~ not complete.

The Amazingly Useful Basu's Theorem

- Theorem 1.10 (**Basu's theorem**): Complete sufficient statistics are independent of *all* ancillary statistics.

u

Proof. (Discrete case). Let $T = T(\vec{X})$ be a complete sufficient statistic. Let $S = S(\vec{X})$ be an ancillary statistic for θ. It suffices to show $\mathbb{P}_\theta(S=s \mid T=t) = \mathbb{P}(S=s)$.

By the law of total probability,

$$\mathbb{P}(S=s) = \sum_{t \in \mathcal{T}} \mathbb{P}_\theta(S=s \mid T=t) \cdot \mathbb{P}_\theta(T=t) \qquad (1)$$

Also, $1 = \sum_{t \in \mathcal{T}} \mathbb{P}_\theta(T=t)$, so $\mathbb{P}(S=s) = \left(\sum_{t \in \mathcal{T}} \mathbb{P}_\theta(T=t)\right) \cdot \mathbb{P}(S=s)$. (2)

So $0 = (1) - (2) = \sum_{t \in \mathcal{T}} \Big[\underbrace{\mathbb{P}_\theta(S=s \mid T=t) - \mathbb{P}(S=s)}_{=:\, h(t)} \Big] \cdot \mathbb{P}_\theta(T=t)$

$= \sum_{t \in \mathcal{T}} h(t) \cdot \mathbb{P}_\theta(T=t)$

$= \mathbb{E}_\theta[h(T)] \quad \forall \theta \in \Theta.$

Since T is complete, we <u>must</u> have $\mathbb{P}_\theta(S=s \mid T=t) = \mathbb{P}(S=s)$.

$\square$

Poll Time!

On Quercus: Module 1 - Poll 6

Basu's Theorem: Examples

- Example 1.40: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that the sample mean $\bar{X}_n$ is independent of the sample variance S_n^2.

By Exemple 1.36, we know $\bar{X}_n$ is a complete sufficient statistic for μ.

By Example 1.34, S_n^2 is ancillary for μ.

By Basu's theorem, $\bar{X}_n \perp\!\!\!\perp S_n^2$.

FYI:

- This is actually a characterizing property of the Normal distribution: $\bar{X}_n \perp S_n^2$ if and only if $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

Basu's Theorem: Examples

- Example 1.41: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Exp(θ), where $\theta > 0$. Use Basu's theorem to find $\mathbb{E}_\theta\left[\frac{X_1}{X_1+\cdots+X_n}\right]$.

$\{\text{Exp}(\theta) : \theta > 0\}$ is a scale family $\Rightarrow$ $\frac{X_1}{X_1+\cdots+X_n}$ is ancillary for θ by Theorem 1.6.

Also, it's in an exponential family with $T(x) = x \Rightarrow T(\vec{X}) = X_1 + \cdots + X_n$ is a complete sufficient statistic for θ. By Basu's theorem, $\frac{X_1}{X_1+\cdots+X_n} \perp\!\!\!\perp X_1 + \cdots + X_n$.

$$\mathbb{E}\left[\frac{X_1}{X_1+\cdots+X_n}\cdot(X_1+\cdots+X_n)\right] = \mathbb{E}\left[\frac{X_1}{X_1+\cdots+X_n}\right]\cdot\mathbb{E}[X_1+\cdots+X_n]$$

$$\Rightarrow \mathbb{E}[X_1] = \mathbb{E}\left[\frac{X_1}{X_1+\cdots+X_n}\right]\cdot\mathbb{E}[X_1+\cdots+X_n]$$

$$\Rightarrow \frac{1}{\theta} = \mathbb{E}\left[\frac{X_1}{X_1+\cdots+X_n}\right]\cdot\frac{n}{\theta}$$

$$\Rightarrow \mathbb{E}\left[\frac{X_1}{X_1+\cdots+X_n}\right] = \frac{1}{n}.$$

Basu's Theorem: Examples

- Example 1.42: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_{\mu,\sigma}$ where

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right), \quad x \in \mathbb{R},$$

where $\sigma > 0$ and μ is known. Show that X_1/X_n is independent of $\sum_{i=1}^n |X_i - \mu|$.

$\{f_\sigma : \sigma > 0\}$ is a scale family.

$$\sigma \cdot f_\sigma(x\sigma) = \sigma \cdot \frac{1}{2\sigma} \cdot \exp\left(-\frac{|x\sigma - \mu|}{\sigma}\right)$$
$$= \frac{1}{2} \cdot \exp\left(-|x - \mu/\sigma|\right)$$

So we're in a scale family.

By Theorem 1.6, X_1/X_n is ancillary for σ.

$$f_\sigma(x) = \underbrace{\tfrac{1}{2}}_{h(x)} \cdot \underbrace{\tfrac{1}{\sigma}}_{g(\sigma)} \cdot \exp\Big(-\underbrace{|x-\mu|}_{T(x)} \cdot \underbrace{\tfrac{1}{\sigma}}_{\eta(\sigma)}\Big)$$

$\eta(\sigma) = 1/\sigma$ is continuous on $(0,\infty)$. Also $\Theta = (0,\infty)$ contains an open interval. By Theorem 1.8, $T(\vec{X}) = \sum_{i=1}^n |X_i - \mu|$ is complete sufficient for σ.

By Basu's theorem,

$\frac{X_1}{X_n}$ is independent of $\sum_{i=1}^n |X_i - \mu|$

From Slide 49:

- Bahadur implies that if a $\underbrace{\text{minimal sufficient statistic}}^{\text{say } T(\vec{x})}$ exists and it's not complete, then no complete sufficient statistic exists

Why? Suppose that a complete sufficient statistic $U(\vec{x})$ did exist. By Bahadur, $U(\vec{x})$ must be minimal sufficient. But then $U(\vec{x})$ and $T(\vec{x})$ must be one-to-one functions of each other, since they're both minimal sufficient. But then $T(\vec{x})$ is a one-to-one function of a complete statistic, and hence itself complete (Assignment 1). Contradiction! So $U(\vec{x})$ cannot exist after all...