STA261 - Module 1

Statistics

Rob Zimmerman

University of Toronto

July 5-7, 2022

Data and samples

- *Data* is factual information collected for the purposes of inference (Merriam-Webster)

- *Inference* is the act of passing from statistical sample 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:

  – height of typical UofT student $\overset{\text{assume}}{\sim} N(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$, $\sigma^2 > 0$

  – # of courses STA261 students are taking $\overset{?}{\sim}$ Poisson(λ)? for some $\lambda > 0$.

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 $Z \sim N(0,1) \Rightarrow P(Z>0) = 1/2$, $z \in \mathbb{R} \Rightarrow P(z>0) \in \{0,1\}$, $z = 261 \Rightarrow P(z>0) = 1$.

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

- Example 1.2: $\vec{X}_n = (X_1, X_2, \ldots, X_n)$ random sample; $\vec{x}_n = (x_1, x_n, \ldots, x_n) \in \mathbb{R}^n$ observed data

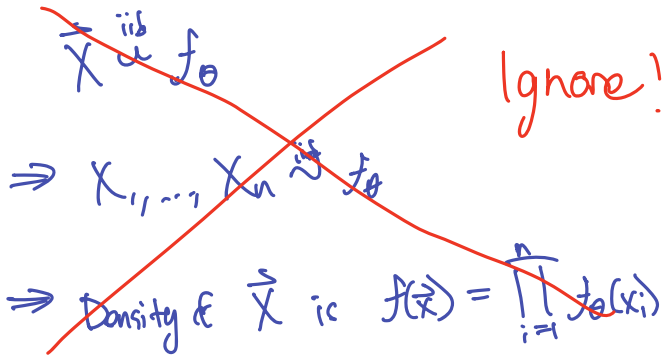
- It is **very** important to clearly distinguish between the two quantities. But why?

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 "Let $X_1, X_2, \ldots, X_n$ be a random sample from same distribution with pdf/pmf f_θ" (same thing!)

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

  Ex: "Let $Z_1, Z_2 \ldots, Z_n \overset{iid}{\sim} N(0,1)$"



Ignore! Not worth the confusion.

A slight abuse of notation: we'll usually write $f_\theta(\vec{x})$ for the joint pdf/pmf of $\vec{X} = (X_1, \ldots, X_n)$ and $f_\theta(x_i)$ for the pdf/pmf of any X_i, in the iid case.

So, for example, if $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, then

$$f_\lambda(\vec{x}) = \prod_{i=1}^{n} f_\lambda(x_i) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

Statistics

$\vec{X}$

- Definition 1.1: A **statistic** is a function of the (random) data sample which is free of any unknown constants
- Example 1.3:

$T(\vec{X}) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$T(\vec{X}) = 24 \cdot X_3^2$

$T(\vec{X}) = 261$

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

Say $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$. Let $T(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu$.

NOT A STATISTIC!! Cuz μ is unknown.

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

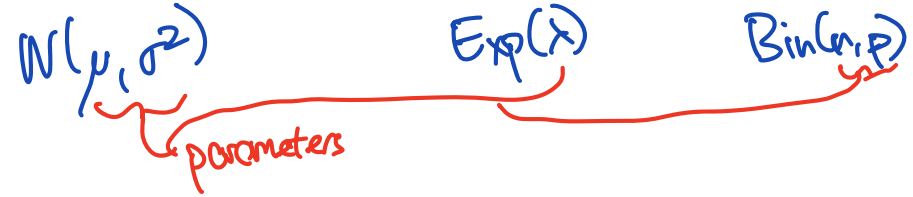
Intuitively, $T(\vec{X}) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is useful if we're interested in μ.

- Ex: $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \in (0,1)$.

Then $T(\vec{X}) = \bar{X}_n$ is also useful for understanding p.

Parameters and Statistical Models

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



- 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: $\left\{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) : \mu \in \mathbb{R}\right\} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$

$\{\lambda e^{-\lambda x} : \lambda > 0\} = \{Exp(\lambda) : \lambda > 0\}$

capital theta

$\Theta = \{\theta_1, \theta_2\}$

$\Rightarrow \{f_\theta : \theta \in \Theta\} = \{f_{\theta_1}, f_{\theta_2}\}$

- 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:

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

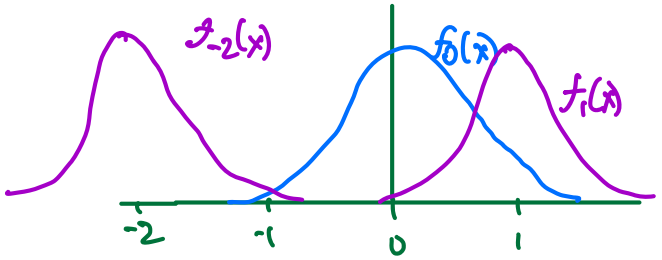
$$= \{N(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

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

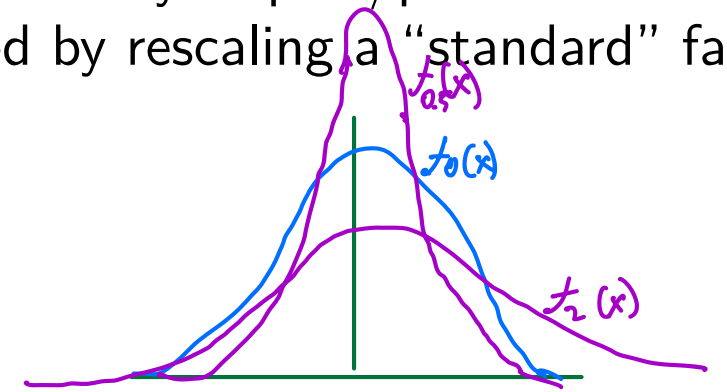
if we know in advance that $\mu > 0$.

Important Parametric Families: Location-Scale Families

- Definition 1.3: A **location family** is a family of pdfs/pmfs $\{f_\mu(\cdot) = f_0(\cdot - \mu) : \mu \in \mathbb{R}\}$ formed by translating a "standard" family member $f_0(\cdot)$.
- Example 1.8: $\{N(\mu, 1) : \mu \in \mathbb{R}\}$

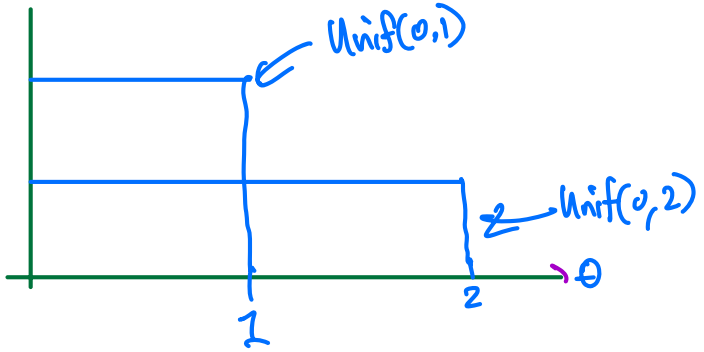


- Definition 1.4: A **scale family** is a family of pdfs/pmfs $\{f_\sigma(\cdot) = f_1(\cdot/\sigma)/\sigma : \sigma > 0\}$ formed by rescaling a "standard" family member $f_1(\cdot)$.
- Example 1.9: $\{N(0, \sigma^2) : \sigma^2 > 0\}$

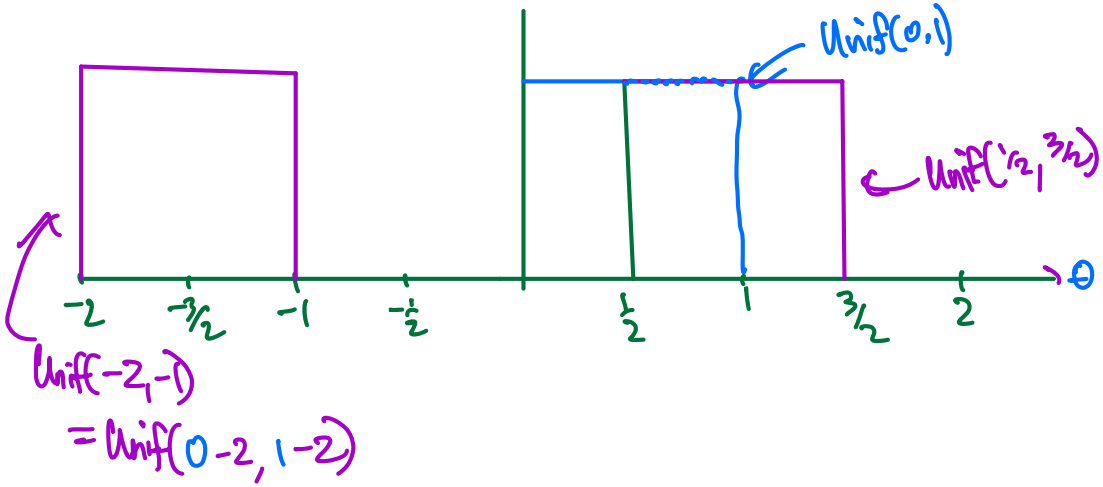


- Definition 1.5: A **location-scale family** is a family of pdfs/pmfs $\{f_{\mu,\sigma}(\cdot) = f_{0,1}\left(\frac{\cdot - \mu}{\sigma}\right)/\sigma : \mu \in \mathbb{R}, \sigma > 0\}$ formed by translating and rescaling a "standard" family member $f_{0,1}(\cdot)$.
- Example 1.10: $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$

Poll Time!



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

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

for some $k \in \mathbb{N}$, where all functions of x and θ are *known*.

Usually $k=1$ when $\theta \in \mathbb{R}$, whence $f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp(w(\theta) \cdot T(x))$

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

- Many of your favourite distributions are included

eg
- $\text{Bin}(n,p)$, n known
- $N(\mu, \sigma^2)$
- $\text{Gamma}(\alpha, \beta)$
- $\text{Beta}(\alpha, \beta)$
- $\text{Exp}(\lambda)$

$\text{Poisson}(\lambda)$

$\text{Multinomial}(n; p_1, \ldots, p_k)$, n known

$\chi^2_{(n)}$

- Example 1.11:

$X \sim \text{Exp}(\lambda)$

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

$= 1 \cdot \lambda \cdot \exp(-\lambda \cdot x)$

$h(x)$, $g(\lambda)$, $w(\lambda)$, $T(x)$

$x = \exp(\log(x)) \quad \forall x > 0$

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

$f_\theta(x) = \theta^x (1-\theta)^{1-x}$

$= (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x$

$= (1-\theta) \cdot \exp\left(\log\left(\left(\frac{\theta}{1-\theta}\right)^x\right)\right)$

$= 1 \cdot (1-\theta) \cdot \exp\left(x \cdot \log\left(\frac{\theta}{1-\theta}\right)\right)$

$h(x)$, $g(\theta)$, $T(x)$, $w(\theta)$

A Quick Review of Conditional Distributions

- $X|Y$ is a random variable, which has its own distribution called a conditional distribution
- Remember Bayes' rule: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, $\mathbb{P}(B) \neq 0$
- $X \mid Y = y$ $\quad f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

- $X \mid X = x$ degenerate/constant at x
- Example 1.12: $\mathbb{E}[X|Y]$ is a random variable (a function of Y)
- Example 1.13: $\mathbb{E}[X|X] = X$. If $X \perp\!\!\!\perp Y$, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$

"Tower property": $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$

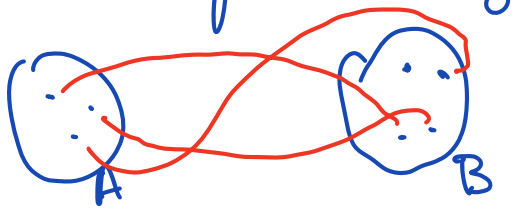
"Conditional variance": $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$

EXERCISE: Prove these, if you haven't already!

A Quick Review of Functions

- Let $f : A \to B$ be a function
- If f is one-to-one, then "injective"

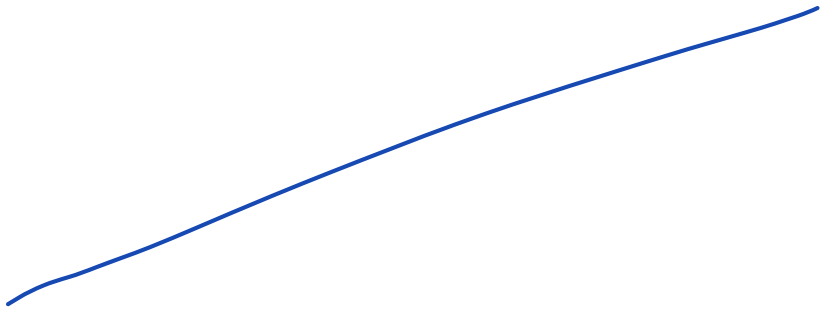
$f(x) = f(y) \Rightarrow x = y$



- If f is onto, then "surjective"

$\forall b \in B, \exists a \in A$ s.t. $f(a) = b$

- If f is a bijection, then f is one-to-one AND onto.
- 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 $g: \Theta \times \mathbb{R} \to \mathbb{R}$ s.t. $f_\theta(x) = g(\theta, x)$

- 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

  $f_\theta(x) = \theta x^2$ is not free of θ

  Say $X \sim N(\mu, 1)$. Then $P_\mu(X - \mu \le x) = \Phi(x)$ 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 $X \sim \text{Exp}(\lambda)$, then the distribution of λX is free of λ

  CHECK THIS!

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: $X_1, \ldots, X_n$ iid $N(\mu, 1)$, $\mu \in \mathbb{R}$.

  Instead of sending us $x_1, \ldots, x_n$, what if we only got $\bar{x}_n = \frac{1}{n}\sum x_i$?

  Can we still make inferences 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

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

  is really free of θ, then the information about θ in $\mathbf{X}$ and the information about θ in $T(\mathbf{X})$ are "equal"

- Example 1.18: $T(\vec{X}) = \vec{X}$ is sufficient:

$$= \frac{P_\theta(\vec{X} = \vec{x})}{P_\theta(T(\vec{X}) = T(\vec{x}))}$$

$$P_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = \vec{x}) = P_\theta(\vec{X} = \vec{x} \mid \vec{X} = \vec{x}) = 1$$

But no data reduction!

Sufficiency

- Example 1.19: 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.

$T(\vec{X}) \sim \text{Bin}(n, \theta) \Rightarrow P(T(\vec{X}) = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}$

Then $P_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = t)$

$= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, \sum X_i = t)$

$= P_\theta(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} x_i = t)$

$= P_\theta(X_1 = x_1, \ldots, X_n = t - \sum_{i=1}^{n-1} x_i)$

$= P_\theta(X_1 = x_1) \cdot \ldots \cdot P_\theta(X_n = t - \sum_{i=1}^{n-1} x_i)$ by independence

$= \theta^{x_1}(1-\theta)^{1-x_1} \ldots \theta^{t - \sum_{i=1}^{n-1} x} (1-\theta)^{1 - t + \sum_{i=1}^{n-1}}$

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

Therefore, $P_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = t) = \dfrac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \dfrac{1}{\binom{n}{t}}$ which 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.

$T(\vec{X}) \sim N(\mu, \sigma^2/n)$ so has density $f_T(t) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(\frac{-n(t-\mu)^2}{2\sigma^2}\right)$.

For the numerator, observe that for $t = \frac{1}{n}\sum_i x_i$

$$\begin{aligned}\sum(x_i-\mu)^2 &= \sum(x_i - t + t - \mu)^2 \\ &= \sum\left[(x_i-t)^2 + 2(x_i-t)(t-\mu) + (t-\mu)^2\right] \\ &= \sum(x_i-t)^2 + n(t-\mu)^2\end{aligned}$$

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

$$\begin{aligned}&= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{-\sum(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{-\sum(x_i-t)^2}{2\sigma^2} - \frac{n(t-\mu)^2}{2\sigma^2}\right) \quad \text{from above} \\ &= f_{\vec{X},T}(\vec{x}, t)\end{aligned}$$

Therefore,

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

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

which is free of μ.

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

The Factorization Theorem

- 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

Proof. (Discrete case). Want to show $\dfrac{P_\theta(\vec{X}=\vec{x} \wedge T(\vec{X})=t)}{P_\theta(T(\vec{X})=t)}$ is free of θ

iff $P_\theta(\vec{X}=\vec{x}) = h(\vec{x}) \cdot g_\theta(t)$.

Continuous case requires measure theory

The Factorization Theorem

($\Rightarrow$) Assume T is sufficient for Θ.

Then $P_\theta(\vec{X}=\vec{x}) = P_\theta(\vec{X}=\vec{x} \wedge T(\vec{X})=t)$

$= \underbrace{P_\theta(\vec{X}=\vec{x} \mid T(\vec{X})=t)}_{=: h(\vec{x})} \cdot \underbrace{P_\theta(T(\vec{X})=t)}_{=: g_\theta(t)} = h(\vec{x}) \cdot g_\theta(t).$

because T is sufficient

($\Leftarrow$) Assume $P_\theta(\vec{X}=\vec{x}) = h(\vec{x}) \cdot g_\theta(t)$ for some h, g_θ.

Then the pmf of T is $P(T(\vec{X})=t) = \sum_{\vec{x}: T(\vec{x})=t} P_\theta(\vec{X}=\vec{x}, T(\vec{X})=t)$

$= \sum_{\vec{x}: T(\vec{x})=t} P_\theta(\vec{X}=\vec{x})$

$= \sum_{\vec{x}: T(\vec{x})=t} h(\vec{x}) \cdot g_\theta(t)$ by ass.

$= \left(\sum_{\vec{x}: T(\vec{x})=t} h(\vec{x})\right) g_\theta(t)$

Therefore, $P_\theta(\vec{X}=\vec{x} \mid T(\vec{X})=t)$

$= \frac{P_\theta(\vec{X}=\vec{x})}{P_\theta(T(\vec{X})=t)} = \frac{h(\vec{x}) \cdot g_\theta(t)}{\left(\sum_{\vec{x}: T(\vec{x})=t} h(\vec{x})\right) g_\theta(t)} = \frac{h(\vec{x})}{\sum_{\vec{x}: T(\vec{x})=t} h(\vec{x})}$ which is free of Θ. $\therefore T(\vec{X})$ is sufficient for Θ. $\square$

Poll Time!

From the proof of the FT,

$$f_\theta(x) = P_\theta(X = x) = \underbrace{P_\theta(X = x \mid T(X) = T(x))}_{h(x)} \cdot \underbrace{P_\theta(T(X) = T(x))}_{g_\theta(x)}$$

But $h(x) = P(X = x \mid T(X) = T(x))$

$= P(X = x \mid X = x) = 1.$

So $h(x)$ must be constant wrt x, but it can be *any* constant $c \neq 0$ as long as $g_\theta(x) = \frac{1}{c} \cdot f_\theta(x)$.

The Factorization Theorem: Examples

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

Let $t = \sum x_i$

$$f_\theta(\vec{x}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$
$$= 1 \cdot \theta^t (1-\theta)^{n-t}$$

$h(x)$ $\quad$ $g_\theta(t)$

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.

Let $t = \frac{1}{n}\sum_i x_i$.

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

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum(x_i-t)^2 - n(t-\mu)^2}{2\sigma^2}\right) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum(x_i-t)^2}{2\sigma^2}\right)}_{=: h(\vec{x})} \cdot \underbrace{\exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right)}_{=: g_\mu(t)}$$

By the FT, $T(\vec{X})$ is sufficient for μ.

The Factorization Theorem: Examples

- Example 1.23: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The **sample variance** is the statistic $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Show that $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ is sufficient for (μ, σ^2).

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

$$f_{\mu,\sigma^2}(\vec{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum(x_i-\bar{x})^2 - n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

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

$h(\vec{x}) = 1$

By the FT, $T(\vec{x})$ is sufficient for (μ, σ^2).

- 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.

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

$$= \theta^{-n} \cdot \mathbb{1}_{0 \le x_1 \le \theta, \ldots, 0 \le x_n \le \theta}$$

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

By the FT, $T(\vec{x}) = x_{(n)}$ is sufficient for θ.

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

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

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 $T(\mathbf{X}) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is sufficient for $(\mu, \sigma^2) = \theta$

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

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

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

$h(x)$ $\quad$ $g(\theta)$ $\quad$ $w_1(\theta)$ $\quad$ $w_2(\theta)$ $\quad$ $T_1(x)$ $\quad$ $T_2(x)$

By Theorem 1.2, $T(\vec{x}) = \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right)$ is sufficient for $\theta = (\mu, \sigma^2)$.

The Factorization Theorem: Examples

- Example 1.26: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(\{1, 2, \ldots, \theta\})$, where $\theta \in \mathbb{N}$. Show that $T(\mathbf{X}) = X_{(n)}$ is sufficient for θ.

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

$$= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{x_i \in \{1,2,\ldots,\theta\}}$$

$$= \theta^{-n} \cdot \mathbb{1}_{x_i \in \{1,2,\ldots,\theta\}\ \forall i}$$

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

By the FT, $T(\vec{X}) = X_{(n)}$ is sufficient for θ.

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. If $T(\vec{X})$ is sufficient for Θ, then by the FT

$f_\theta(\vec{x}) = h(\vec{x}) \cdot g_\theta(T(\vec{x}))$ for some $h(\cdot)$, $g_\theta(\cdot)$.

$= h(\vec{x}) \cdot g_\theta(r^{-1}(r(T(\vec{x}))))$

$= h(\vec{x}) \cdot \tilde{g}_\theta(r(T(\vec{x})))$ where $\tilde{g}_\theta(t) = g_\theta(r^{-1}(t))$.

By the FT, $r(T(\vec{X}))$ is sufficient for Θ. $\square$

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*

$N(\mu, 1)$: $\vec{X}$ is sufficient for μ.
So is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ which is a function of $\vec{X}$ (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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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!

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: $X_1, \dots, X_n$ iid Bernoulli(θ), $\theta \in (0,1)$. Then $T(\vec{X}) = \sum_i X_i$ is minimal sufficient for θ. Let $\vec{x}, \vec{y} \in \mathcal{X}^n$. Then

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

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

$\therefore T(\vec{X})$ 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 $T(\mathbf{X}) = (\bar{X}, S^2)$ is minimal sufficient for $(\mu, \sigma^2) = \theta$.

Let $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

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

$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(\frac{-(n-1)s_x^2 - n(\bar{x}-\mu)^2}{2\sigma^2}\right)$.

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

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

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

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

By Theorem 1.4, $T(\vec{x}) = (\bar{x}, s^2)$ is minimal sufficient for $\theta = (\mu, \sigma^2)$.

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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}([\theta, \theta+1])$, where $\theta \in \mathbb{R}$. Show that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.

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

$$= \mathbb{1}_{\theta \le x_1, x_2, \ldots, x_n \le \theta+1}$$

$$= \mathbb{1}_{\theta \le x_{(1)} \wedge x_{(n)} \le \theta+1}$$

$$= \mathbb{1}_{x_{(n)}-1 \le \theta \le x_{(1)}}$$

Let $\vec{x}, \vec{y} \in \mathcal{X}^n$. Then $\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})} = \frac{\mathbb{1}_{x_{(n)}-1 \le \theta \le x_{(1)}}}{\mathbb{1}_{y_{(n)}-1 \le \theta \le y_{(1)}}}$ is free of θ iff $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$.

Hence, by Theorem 1.4, $T(\vec{x}) = (x_{(1)}, x_{(n)})$ is minimal sufficient for θ.

Poll Time!

(Q: how do you feel about sufficiency?)

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?
- 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

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

$T(\vec{X}) = (X_1 - \frac{1}{2}\sum X_i, X_2 - \frac{1}{2}\sum X_i)$

might not depend on μ

Ancillarity

- Definition 1.10: 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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}([\theta, \theta+1])$, 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 $Y_i \overset{iid}{\sim} \text{Unif}(0,1)$ and $Y_{(j)} = X_{(j)} - \theta$ is also free of θ.

Then $P_\theta(R(\vec{X}) \le r)$

$= P_\theta(X_{(n)} - X_{(1)} \le r)$

$= P_\theta([X_{(n)} - \theta] - [X_{(1)} - \theta] \le r)$

$= P_\theta(Y_{(n)} - Y_{(1)} \le r)$

$= P_\theta(\text{Beta}(n,1) - \text{Beta}(1,n) \le r)$ does not depend on θ.

Hence $R(\vec{X})$ is ancillary for θ.

Ancillarity: Examples

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

- Theorem 1.5: Let $X_1, \ldots, X_n$ 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 $Y_i = X_i - \theta \sim F(\cdot)$.

Then $P_\theta(R(\vec{X}) \le r)$

$= P_\theta(X_{(n)} - X_{(1)} \le r)$

$= P_\theta([X_{(n)} - \theta) - (X_{(1)} - \theta) \le r)$

$= P_\theta(Y_{(n)} - Y_{(1)} \le r)$, which is free of θ since the distribution of $Y_{(1)}, Y_{(n)}$ is free of θ.

□

Ancillarity: Examples

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

Let $Z_i = \frac{X_i}{\sigma} \overset{iid}{\sim} N(0,1)$.

Then $P_{\sigma^2}(D(\vec{X}) \le x)$

$= P_{\sigma^2}\left(\frac{X_1}{X_n} + \frac{X_2}{X_n} + \cdots + \frac{X_{n-1}}{X_n} \le x\right)$

$= P_{\sigma^2}\left(\frac{X_1/\sigma}{X_n/\sigma} + \frac{X_n/\sigma}{X_2/\sigma} + \cdots + \frac{X_{n-1}/\sigma}{X_n/\sigma} \le x\right)$

$= P_{\sigma^2}\left(\frac{Z_1}{Z_n} + \cdots + \frac{Z_{n-1}}{Z_n} \le x\right)$ is free of σ^2.

Hence $D(\vec{X})$ 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 $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{(n-1)}$.

Proof $(n = 2)$.

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

$(n-1)S^2 = \sum_{i=1}^{2} (X_i - \bar{X})^2 = \left(X_1 - \frac{1}{2}(X_1+X_2)\right)^2 + \left(X_2 - \frac{1}{2}(X_1+X_2)\right)^2$
$= \left(\frac{1}{2}X_1 - \frac{1}{2}X_2\right)^2 + \left(\frac{1}{2}X_2 - \frac{1}{2}X_1\right)^2$
$= \frac{1}{2}(X_1 - X_2)^2$
$\overset{d}{=} \frac{1}{2} \cdot 2\sigma^2 \cdot \chi^2_{(1)}$
$\overset{d}{=} \sigma^2 \cdot \chi^2_{(1)}$

$\Rightarrow \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{(1)}$. □

(i.e., $(n-1)S^2 \overset{d}{=} \sigma^2 \cdot Z$ where $Z \sim \chi^2_{(1)}$)

General n: exercise (use induction).

- 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^2 is ancillary for μ.

We showed that $S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{(n-1)}$ which is free of μ. So S^2 is ancillary for μ.

Poll Time!

$Z \sim f_Z$ has no info about the unknown θ in our sample $X_1, \dots, X_n$.

$X_1 - X_n$ and X_1 / X_n may be ancillary in some models, but not in general!

Support of a function: $\text{Supp}(f) := \{x : f(x) \neq 0\}$

If $f : \mathbb{R} \to \mathbb{R}$ is not supported on $\mathbb{R}$, multiply by an indicator for $\text{Supp}(f)$ to be safe.

eg. if $X \sim \text{Unif}(0,1)$, then $f_X(x) = 1$ is FALSE!

TRUE: $f_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & \text{otherwise} \end{cases} = \mathbb{1}_{x \in [0,1]}$.

Indicators are useful! $x \in A \cap B \iff \mathbb{1}_{A \cap B}(x) = 1 \iff \mathbb{1}_A(x) \cdot \mathbb{1}_B(x) = 1$

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 1.11: A statistic $U(\mathbf{X})$ is **complete** if *any* function $h(\cdot)$ which satisfies $\mathbb{E}_\theta\left[h(U(\mathbf{X}))\right] = 0$ for all $\theta \in \Theta$ must also satisfy $\mathbb{P}_\theta\left(h(U(\mathbf{X})) = 0\right) = 1$ for all $\theta \in \Theta$.

continuous: $\int h(u) \cdot f_\theta(u)\, du = 0 \quad \forall \theta \in \Theta$ where $f_\theta(\cdot)$ is true pdf $U(\vec{X})$

discrete: $\sum_{u \in \mathcal{U}} h(u) \cdot P_\theta(U(\vec{X}) = u) = 0 \quad \forall \theta \in \Theta$.

$U(\vec{X})$ complete means this implication is T: $\left(E_\theta[h(U(\vec{X}))] = 0 \ \forall\theta \Rightarrow P_\theta(h(U(\vec{X})) = 0) = 1 \ \forall\theta \right)$

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
- If $U(\mathbf{X})$ is complete and supported on $\{u_1, \ldots, u_n\}$, then $\sum_{i=1}^n h(u_i) \cdot \mathbb{P}_\theta \left(U(\mathbf{X}) = u_i\right) = 0$ implies $h(u_i) = 0$ 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.

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

Suppose $h(\cdot)$ is some function s.t. $E_\theta[h(U(\vec{X}))] = 0 \quad \forall \theta \in (0,1)$

$\Rightarrow E_\theta[h(\sum_{i=1}^{n} X_i)] = 0 \quad \forall \theta \in (0,1)$

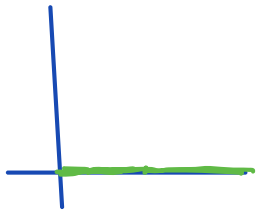
Then $0 = \sum_{j=0}^{n} h(j) \cdot \binom{n}{j} \theta^j (1-\theta)^{n-j}$

$= (1-\theta)^n \sum_{j=0}^{n} h(j) \cdot \left[\binom{n}{j} \left(\frac{\theta}{1-\theta}\right)^j\right]$ $\quad (1-\theta)^n \neq 0$ so we can divide through by it

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

$= \sum_{j=0}^{n} \hat{h}(j) \cdot r^j$ $\quad$ where $\hat{h}(j) = \binom{n}{j} h(j)$

is a polynomial in r which is 0 for all $r \in (0, \infty)$ (i.e., constant at 0).



Therefore $\hat{h}(j)$ is 0 $\forall j$.

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

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

$\Rightarrow h(\cdot) = 0$ on $\{0, 1, \ldots, n\}$.

Hence, $P_\theta(h(U(\vec{X})) = 0) = 1 \quad \forall \theta \in \Theta$, so $U(\vec{X})$ 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

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

Then $T(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i) \right)$ is a complete statistic, as long as each component of Θ contains an open interval in $\mathbb{R}$.[1]

g: $\Theta = \mathbb{R} \supseteq (0,1)$ ✓
$\Theta = \mathbb{R} \times (0,\infty) \supseteq (0,1) \times (0,1)$ ✓
$\Theta = \{1,2,3,\ldots,\theta\}$ ✗ DOES NOT CONTAIN AN OPEN INTERVAL

- 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.

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

$$= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)}_{h(x)} \cdot \underbrace{\exp\left(\frac{-\mu^2}{2\sigma^2}\right)}_{g(\theta)} \cdot \exp\left(\underbrace{\frac{\mu}{\sigma^2}}_{w(\theta)} \cdot \overbrace{x}^{T(x)}\right)$$

Since $\Theta = \mathbb{R}$ clearly contains an open interval,

$T(\vec{X}) = \bar{X}_n$ is complete for μ 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.

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

$h(x)$ $g(\lambda)$ $T(x)$ $w(\lambda)$

I can multiply/divide by any constant I want!

$\Omega = (0, \infty)$ contains an open interval, so $T(\vec{X}) = \bar{X}_n$ is complete for λ by Theorem 1.8.

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.

$T(\vec{X}) = \sum_{i=1}^{n} |X_i - \mu|$

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**): If a minimal sufficient statistic and a complete statistic both exist, then the complete statistic must also be minimal sufficient.

sufficient

sufficient

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

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

  ↑ sufficient

- 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.

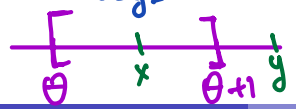
$f_\theta(x) = \mathbb{1}_{\theta \le x \le \theta+1}$.

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

$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\mathbb{1}_{\theta \le x \le \theta+1}}{\mathbb{1}_{\theta \le y \le \theta+1}}$ is free of θ iff $x = y$.

By Theorem 1.4, $T(X) = X$ is minimal sufficient for θ.

If $x < y$, then we could have $x \le \theta+1 \le y$, so $\mathbb{1}_{\theta \le x \le \theta+1} \neq \mathbb{1}_{\theta \le y \le \theta+1}$ $\forall \theta$:



However, consider $h(x) = \sin(2\pi x)$.

We have $\mathbb{E}_\theta[h(T(X))]$

$= \mathbb{E}_\theta[h(X)]$

$= \int_\theta^{\theta+1} \sin(2\pi x)\, dx$

$= 0 \quad \forall \theta \in \mathbb{R}$.

But clearly, $h(\cdot)$ is not identically 0.

So $P_\theta(h(X) = 0) \neq 1 \quad \forall \theta \in \mathbb{R}$.

So $T(X) = X$ is not complete.

The Amazingly Useful Basu's Theorem

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

Proof. (Discrete)

Let $T = T(\vec{X})$ be a complete sufficient statistic.
Let $S = S(\vec{X})$ be ancillary for θ.
It suffices to show $P_\theta(S=s \mid T=t) = P(S=s)$.
By the Law of Total Probability, $P(S=s) = \sum_{t \in T} P(S=s \mid T=t) \cdot P_\theta(T=t)$ ①
Moreover $1 = \sum_{t \in T} P(T=t)$ so $P(S=s) = \left(\sum_{t \in T} P(T=t)\right) \cdot P(S=s)$ ②
Therefore, $0 = ① - ② = \sum_{t \in T} \underbrace{\left[P(S=s \mid T=t) - P(S=s)\right]}_{=: h(t)} \cdot P_\theta(T=t)$

$$= \sum_{t \in T} h(t) \cdot P_\theta(T=t)$$

$$= E_\theta[h(T)] \quad \forall \theta \in \Theta.$$

Since T is complete, $h(t) = 0 \ \forall t \Rightarrow P(S=s \mid T=t) = P(S=s) \Rightarrow S \perp T$. □

Poll Time!

Basu's Theorem: Examples

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

Example 1.36 $\Rightarrow$ $\bar{X}_n$ is a complete sufficient statistic for μ.

Example 1.34 $\Rightarrow$ S^2 is ancillary for μ.

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

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

(allowing for degenerate Normals with $\sigma^2 = 0$, since constants are independent of everything)

Basu's Theorem: Examples

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

$f_\theta(x) = \theta e^{-\theta x}$

$\Rightarrow \mathbb{E}[X] = \int_0^\infty x \cdot \theta e^{-\theta x} dx = \theta \cdot \frac{1}{\theta^2} = \frac{1}{\theta}$.

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

$\{\text{Exp}(\theta): \theta > 0\}$ is an exponential family with $T(x) = x \Rightarrow X_1+\cdots+X_n$ is complete sufficient for θ.

By Basu's theorem, $\frac{X_1}{X_1+\cdots+X_n} \perp\!\!\!\perp X_1+\cdots+X_n$.

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

$= \mathbb{E}_\theta\left[\frac{X_1}{X_1+\cdots+X_n}\right] \cdot \mathbb{E}_\theta[X_1+\cdots+X_n]$ by independence

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

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

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 $\sum_{i=1}^n |X_i - \mu|$.

This is a scale family, so $\frac{X_1}{X_n}$ is ancillary for σ by Theorem 1.6.

Moreover, $\sum_{i=1}^n |X_i - \mu|$ is a complete sufficient statistic by Example 1.38.

By Basu's theorem, $\sum_{i=1}^n |X_i - \mu| \perp\!\!\!\perp \frac{X_1}{X_n}$.