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:

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
- We denote random variables using uppercase letters, and constants using lowercase letters:
- Example 1.2:
- 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
-

Statistics

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

Parameters and Statistical Models

- Many classical probability distributions have *parameters* associated with them

- Example 1.5:

- 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:

- 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:

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 $f_0(\cdot)$.

- Example 1.8:

- 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 $f_1(\cdot)$.

- Example 1.9:

- 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 $f_{0,1}(\cdot)$.

- Example 1.10:

Poll Time!

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*.

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

- Many of your favourite distributions are included

- Example 1.11:

A Quick Review of Conditional Distributions

- $X|Y$ is a random variable, which has its own distribution called a conditional distribution
- Remember Bayes' rule:
- $X \mid Y = y$
- $X \mid X = x$
- Example 1.12:
- Example 1.13:

A Quick Review of Functions

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

- 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:

- 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:

Data Reduction: A Thought Experiment

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

- Suppose that field researchers collect a sample $\mathbf{X} = (X_1, X_2, \dots, X_n) \overset{iid}{\sim} f_\theta$, 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:

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:

Sufficiency

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

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.

The Factorization Theorem

Poll Time!

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.

- Example 1.22: 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.

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).

- 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.

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.

The Factorization Theorem: Examples

- Example 1.25: 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 $T(\mathbf{X}) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is sufficient for (μ, σ^2).

The Factorization Theorem: Examples

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

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*

- 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 $T(\mathbf{X}) = (\bar{X}, S^2)$ is not 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.
- This criterion is easier to apply than it looks
- Example 1.28:

Minimal Sufficiency: Examples

- Example 1.29: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that $T(\mathbf{X}) = (\bar{X}, S^2)$ 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.

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.

Poll Time!

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?

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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}\left([\theta, \theta + 1]\right)$, where $\theta \in \mathbb{R}$. Show that the range statistic $R(\mathbf{X}) := X_{(n)} - X_{(1)}$ 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 $F(\cdot - \mu)$, for $\mu \in \mathbb{R}$. Then the range statistic is ancillary for μ.

Proof.

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.

- 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.

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

Proof $(n = 2)$.

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

Poll Time!

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 $\mathbb{E}_\theta\left[h(U(\mathbf{X}))\right] = 0$ for all $\theta \in \Theta$ must also satisfy $\mathbb{P}_\theta\left(h(U(\mathbf{X})) = 0\right) = 1$ for all $\theta \in \Theta$.

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.

...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]

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

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.

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.

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 sufficient statistic both exist, then the complete statistic must also be minimal sufficient.

- 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 sufficient statistic exists

- 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.

The Amazingly Useful Basu's Theorem

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

Proof.

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.

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

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

Basu's Theorem: Examples

- Example 1.42: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} f_{\mu,\sigma}$ 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 $\sum_{i=1}^{n} |X_i - \mu|$.