Data Reduction

Jacob M. Montgomery

2018

Data Reduction

Key concepts

- For traditional approaches to statistical inference, we do not want to handle our entire dataset.

Key concepts

- For traditional approaches to statistical inference, we do not want to handle our entire dataset.
- Intead, we often make parametric assumptions about the DGP that allow us to focus on specific statistics calculated from the sample.

Key concepts

- For traditional approaches to statistical inference, we do not want to handle our entire dataset.
- Intead, we often make parametric assumptions about the DGP that allow us to focus on specific statistics calculated from the sample.
- Here we focus on two conceptual quantities that we can calculate from our sample:
  - Sufficient statistics
  - The likelihood

- The basic idea is that we take our vector of data, $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{x} = (x_1, \ldots, x_n)$ and throw a bunch of data away.

- The basic idea is that we take our vector of data, $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{x} = (x_1, \ldots, x_n)$ and throw a bunch of data away.
- BUT, by making assumptions about the DGP, we can calculate statistics like $T(\mathbf{x})$ or $T(\mathbf{y})$ that include all the information we need to make statistical inference.

- The basic idea is that we take our vector of data, $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{x} = (x_1, \ldots, x_n)$ and throw a bunch of data away.
- BUT, by making assumptions about the DGP, we can calculate statistics like $T(\mathbf{x})$ or $T(\mathbf{y})$ that include all the information we need to make statistical inference.
- The value of this approach is computational efficiency. The drawback is that our inferences are only as good as our assumptions.

- The basic idea is that we take our vector of data, $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{x} = (x_1, \ldots, x_n)$ and throw a bunch of data away.
- BUT, by making assumptions about the DGP, we can calculate statistics like $T(\mathbf{x})$ or $T(\mathbf{y})$ that include all the information we need to make statistical inference.
- The value of this approach is computational efficiency. The drawback is that our inferences are only as good as our assumptions.
- In addition, we can have very different vectors $\mathbf{x}$ and $\mathbf{y}$ that give us the same values for $T(\mathbf{x})$ and $T(\mathbf{y})$.

- The basic idea is that we take our vector of data, $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{x} = (x_1, \ldots, x_n)$ and throw a bunch of data away.
- BUT, by making assumptions about the DGP, we can calculate statistics like $T(\mathbf{x})$ or $T(\mathbf{y})$ that include all the information we need to make statistical inference.
- The value of this approach is computational efficiency. The drawback is that our inferences are only as good as our assumptions.
- In addition, we can have very different vectors $\mathbf{x}$ and $\mathbf{y}$ that give us the same values for $T(\mathbf{x})$ and $T(\mathbf{y})$.

Sufficient statistics

- The core concept is that we want to determine a reduced form of the data that will tell us about the DGP.

Sufficient statistics

- The core concept is that we want to determine a reduced form of the data that will tell us about the DGP.
- First we make a paremetric assumption about the DGP, which allows us to characterize it in terms of a set of parameters θ

If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value of $T(\mathbf{X})$.

Formal definition

A statistic $T(\mathbf{X})$ is a sufficient statistic if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

Formal definition

A statistic $T(\mathbf{X})$ is a sufficient statistic if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

- In words, this means that the conditional distribution of our data does not change for any value of θ once we know $T(\mathbf{X})$

Establishing sufficiency

- Calculate $p(\mathbf{x}|\theta)$
- Choose some candidate for the sufficient statistic $T(\mathbf{X}|\theta)$
- Calculate $q(T(\mathbf{x})|\theta)$
- Calculate
$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$
- If this quantity does not depend on θ, it is suffficent.

Example 6.2.3: Binomial sufficient statistic

Let

$$X_1, \ldots, X_n$$

be iid Bernoulli random variables with parameter θ. Show that $T(\mathbf{X}) = \sum X_i$ is a sufficient statistic for θ.

Example 6.2.3: Binomial sufficient statistic

Let

$$X_1, \ldots, X_n$$

be iid Bernoulli random variables with parameter θ. Show that $T(\mathbf{X}) = \sum X_i$ is a sufficient statistic for θ. Let $t = \sum x_i$

$$\frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

Why?

- *IF* the distribution of $\mathbf{X}$ does not depend on θ then
- $$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) = \frac{P(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P(T(\mathbf{X}) = T(\mathbf{x}))}$$

Why?

- *IF* the distribution of $\mathbf{X}$ does not depend on θ then
- $$P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))=\frac{P(\mathbf{X}=\mathbf{x} \text{ and } T(\mathbf{X})=T(\mathbf{x}))}{P(T(\mathbf{X})=T(\mathbf{x}))}$$
- $$P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))=\frac{P(\mathbf{X}=\mathbf{x})}{P(T(\mathbf{X})=T(\mathbf{x}))}$$

Why?

- *IF* the distribution of $\mathbf{X}$ does not depend on θ then
- $$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) = \frac{P(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P(T(\mathbf{X}) = T(\mathbf{x}))}$$
- $$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) = \frac{P(\mathbf{X} = \mathbf{x})}{P(T(\mathbf{X}) = T(\mathbf{x}))}$$
- Which can be re-written as

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$

Example 6.2.4

Let $X_1, \ldots, X_n$ by iid $N(\mu, \sigma^2)$ where σ^2 is known. Show that the sample mean is a sufficient statistic for μ.

HINT: $\sum_{i=1}^{n}(x_i - \mu)^2 = \sum(x - \bar{x})^2 + n(\bar{x} - \mu)^2$

The exponential family

- A number of very common distributions can be "factored" in such a way that they can be re-represented as having a common family form.
- This is useful because we can then prove results for this broader family without having to prove it for each individual distribution.

Defining the expontential family

Suppose $X_1, \dots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$.

Defining the expontential family

Suppose $X_1, \ldots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$. We say this is an exponential family if we can factor the distribution such that:

$$f(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t(x))$$

Note that i is indexing statistics and $w()$, so that we may have multiple parameters and statistics of interest.

Note that i is indexing statistics and $w()$, so that we may have multiple parameters and statistics of interest. This is critical for some calculations later, but the single-variable example is enough to make the point.

An equivalent way to write this is:

$$f(x|\theta) = h(x)\exp(\eta' T(x) - A(\eta))$$

Exercises

- Show that the normal distribution with known variance σ can be written as a member of the exponential family.
- Show that the poisson distribution is a member of the exponential family.

Relating back to sufficiency: Factorization theorem

Let $f(\mathbf{x}|\theta)$ *denote the joint pdf or pmf of a sample* $\mathbf{X}$*. A statistic* $T(\mathbf{X})$ *is a sufficient statistic for* θ *iff the pmf/pdf can be re-written as*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}|\theta))h(\mathbf{x})$$

Theorem 6.2.10

Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf that belongs to an exponential famlily given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

,

where theta $= (\theta_1, \theta_2, \ldots, \theta_d)$, where $d \leq k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is a sufficient statistic for θ.

Example 6.2.9: Normal sufficient statistic, both parameters unknown

Assume that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ where neither parameter is known, such that $\theta = (\mu, \sigma^2)$. Use the factorization theorem to show that $\bar{x}$ and $s^2 = \frac{\sum_{i=1}(x_i - \bar{x})}{(n-1)}$ are sufficient statistics for this distribution.

The likelihood function

- As we have seen, in some cases simply handling a sufficient statistic may be inadequate since a sufficient statistic may be the entire dataset.
- Moreover, for several types of statistical inference we will not rely on sufficient statistics at all.
- For both of these reasons, we often switch to calculate a statistic called the *likelihood*.

Defining the likelihood function

Let $f(\mathbf{x}|\theta)$ denote the joint pdf of pmf of the sample $\mathbf{X} = (X_1, \dots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the *likelihood function*.

Thinking about the likelihood function

- We seem to be defining the likelihood the same as the pdf/pmf.
- The only difference is how we will think about θ and $\mathbf{x}$.
  - For $f(\mathbf{x}|\theta)$ we consider $\mathbf{x}$ as the variable and θ to be fixed.
  - For $L(\theta|\mathbf{x})$ we consider $\mathbf{x}$ to be the observed sample and θ to be varying over all possible parameter values.
- Bayesian thinking will consider θ as a variable. Other approaches tend to think of θ as a fixed but unknown parameter.

Poisson likelihood.

Let

$$X_1, \ldots, X_n$$

be iid Poisson random variables with parameter θ. Assume that the observed values of $\mathbf{X}$ are $\mathbf{x} = (4, 17, 4)$.

- Find $L(\theta|\mathbf{x})$.
- Write out the generic version for any (non-empty) observed data $\mathbf{x}$

Binomial Likelihood.

Let

$$X_1, \ldots, X_n$$

be iid Bernoulli random variables with parameter θ. Find $L(\theta|\mathbf{x})$.

Normal likelihood.

Let $X_1, \ldots, X_n$ by iid $N(\mu, \sigma^2)$ where σ^2 is known.

- Find $L(\theta|\mathbf{x})$.
- Can it be represented in terms of the sufficient statistic $T(\mathbf{x})$?

The likelihood principal

If x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is, there exists a constant $C(x, y)$ such that

$$L(\theta|x) = C(x, y)L(\theta|y)\forall\theta,$$

then the conclusion drawn from x and y should be identical.

Thinking about the likelihood principal

- If $L(\theta_2|x) = 2L(\theta_1|x)$, this means that θ_2 is twice as likely.

Thinking about the likelihood principal

- If $L(\theta_2|x) = 2L(\theta_1|x)$, this means that θ_2 is twice as likely.
- If we instead observe $L(\theta_2|y)$ and $L(\theta_1|y)$, then θ_2 should still be twice as likely such that $L(\theta_2|y) = 2L(\theta_1|y)$.

Thinking about the likelihood principal

- If $L(\theta_2|x) = 2L(\theta_1|x)$, this means that θ_2 is twice as likely.
- If we instead observe $L(\theta_2|y)$ and $L(\theta_1|y)$, then θ_2 should still be twice as likely such that $L(\theta_2|y) = 2L(\theta_1|y)$.
- So long as the underlying DGP does not change, no one realization from the data should change our conclusions about which values of θ are *relatively* more likely *so long as the likelihood of the two points is proportional*.

Thinking about the likelihood principal

- If $L(\theta_2|x) = 2L(\theta_1|x)$, this means that θ_2 is twice as likely.
- If we instead observe $L(\theta_2|y)$ and $L(\theta_1|y)$, then θ_2 should still be twice as likely such that $L(\theta_2|y) = 2L(\theta_1|y)$.
- So long as the underlying DGP does not change, no one realization from the data should change our conclusions about which values of θ are *relatively* more likely *so long as the likelihood of the two points is proportional*.
- Imagine if we knew that $L(\theta_1|x) = 4L(\theta_1|y)$ and $L(\theta_2|x) = 4L(\theta_2|y)$ but somehow concluded $L(\theta_1|x) > L(\theta_2|x)$ and $L(\theta_1|y) > L(\theta_2|y)$

Thinking about the likelihood principal

- If $L(\theta_2|x) = 2L(\theta_1|x)$, this means that θ_2 is twice as likely.
- If we instead observe $L(\theta_2|y)$ and $L(\theta_1|y)$, then θ_2 should still be twice as likely such that $L(\theta_2|y) = 2L(\theta_1|y)$.
- So long as the underlying DGP does not change, no one realization from the data should change our conclusions about which values of θ are *relatively* more likely *so long as the likelihood of the two points is proportional*.
- Imagine if we knew that $L(\theta_1|x) = 4L(\theta_1|y)$ and $L(\theta_2|x) = 4L(\theta_2|y)$ but somehow concluded $L(\theta_1|x) > L(\theta_2|x)$ and $L(\theta_1|y) > L(\theta_2|y)$
- This seems almost tautologically true, but we shall see that frequentist approaches to inference actually break this rule.