

1 Sufficient Statistics

We want to estimate θ from an observation x (can also be an IID sample). The estimate $\hat{\theta}(x)$ can be viewed as a compression of x that provides information about θ . We would like to develop a theory of what functions of x provide information about the parameter. For example, we saw that the estimate of the covariance is a function of the empirical first and second moments.

Definition: A function $T(x)$ is called a statistic of x .

Definition: A statistic $T(x)$ is called sufficient for θ if estimators of the form $\hat{\theta}(T(x))$ are not worse than estimators $\hat{\theta}(x)$. Formally, $T(x)$ is sufficient if $p(x|T(x), \theta)$ is not dependent on θ . This is often written as:

$$p(x|T(x), \theta) = p(x|T(x)) \quad (1)$$

If the above is true, it would mean that we can generate samples of x based on $T(x)$ alone. To see this, consider the following method for generating a sample from $p(x|\theta)$:

- Draw x from $p(x|\theta)$. Calculate $T(x)$.
- Sample a new x' according to $p(x'|T(x), \theta) = p(x'|T(x))$

The distribution of x' is exactly $p(x'|\theta)$. Thus we can generate samples as if they came from the original distribution and we only use the value of $T(x)$. It's easy to see that $T(x) = x$ is a sufficient statistic.

Lets show that for Bernouli variables the sample sum is a sufficient statistic:

$$p(x_1, \dots, x_n|\theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i} \quad (2)$$

We claim that $T(x^n) = \sum_i x_i$ is a sufficient statistic.

$$\begin{aligned} p(x_1, \dots, x_n|T(x^n) = k, \theta) &= \frac{p(x_1, \dots, x_n, T(x^n) = k|\theta)}{p(T(x^n) = k|\theta)} \\ &= \frac{\delta_{k, \sum_i x_i} \theta^k (1 - \theta)^{n-k}}{\binom{n}{k} \theta^k (1 - \theta)^{n-k}} = \frac{\delta_{k, \sum_i x_i}}{\binom{n}{k}} \end{aligned}$$

The factorization lemma (Fisher-Neyman): A statistic $T(x)$ is sufficient if and only if $p(x|\theta)$ can be written as:

$$p(x|\theta) = f(\theta, T(x))g(x) \quad (3)$$

where f and g are some functions (proof is in the book).

It's easy to see that in an exponential the function $A(x)$ is sufficient for θ . More generally in a vector exponential family

$$\log p(x|\theta) = \sum_{i=1}^K A_i(x)B_i(\theta) + C(x) + D(\theta) \quad (4)$$

the (vector) statistic $T(x) = [A_1(x), \dots, A_k(x)]$ is sufficient for θ . Similarly IID samples $p(x^n|\theta)$ from an exponential family we have that $T(x) = [\sum_i A_1(x_i), \dots, \sum_i A_k(x_i)]$ is sufficient. For example in a Gaussian the first and second empirical moments are sufficient. The opposite is also true, namely if there is a set of sufficient statistics $T_1(x^n), \dots, T_s(x^n)$ for x^n then $p(x|\theta)$ is exponential as stated in the following theorem.

Pitman Koopman Darmois Theorem: The family $p(x^n|\theta)$ has a sufficient statistic whose size is independent of n if and only if $p(x|\theta)$ is exponential (this is true under conditions on $T(x)$ such as continuity etc).

The following theorem gives a nice illustration of the optimality of sufficient statistics. It tells us that for any estimator $\hat{\theta}(x)$ there is a "better" estimator that only depends on $T(x)$.

Rao-Blackwell Theorem: For every sufficient statistic $T(x)$ and any estimator $\hat{\theta}(x)$ the estimator $\hat{\theta}_T(x) = E_{p(x|T(x))}(\hat{\theta}(x))$ has lower MSE than $\hat{\theta}(x)$ for all θ values.

Note that the new estimator is only a function of $T(x)$ thus justifying our definition of the sufficient statistic.

Proof:

$$\begin{aligned} \text{MSE}(\hat{\theta}_T) &= \sum_x p(x|\theta) \left(\hat{\theta}_T(x) - \theta \right)^2 \\ &= \sum_x p(x|\theta) \left(\sum_{x'} p(x'|T(x)) \hat{\theta}(x') - \theta \right)^2 \\ &= \sum_x p(x|\theta) \left(\sum_{x'} p(x'|T(x)) (\hat{\theta}(x') - \theta) \right)^2 \end{aligned}$$

The squared expression above is a square of the expected value of $\hat{\theta}(x') - \theta$. It is thus smaller than the expected value of its square

$$\begin{aligned}
 \text{MSE}(\hat{\theta}_T) &\leq \sum_x p(x|\theta) \sum_{x'} p(x'|T(x)) \left(\hat{\theta}(x') - \theta \right)^2 \\
 &= \sum_{T(x)} p(T(x)|\theta) \sum_{x'} p(x'|T(x)) \left(\hat{\theta}(x') - \theta \right)^2 \\
 &= \sum_{T(x), x'} p(x', T(x)|\theta) \left(\hat{\theta}(x') - \theta \right)^2 \\
 &= \sum_{x'} p(x'|\theta) \left(\hat{\theta}(x') - \theta \right)^2
 \end{aligned}$$

2 Testing Hypotheses about Parameters

Assume the world states Ω are partitioned into two sets Ω_0 and Ω_1 . Define two hypotheses:

- The **Null Hypothesis** H_0 which corresponds to the hypothesis that $\theta \in \Omega_0$.
- The **Alternative Hypothesis** H_1 which corresponds to the hypothesis that $\theta \in \Omega_1$.

We want to decide if the data is consistent with $\theta \in \Omega_0$.

Examples:

- We see n samples from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is unknown. We want to decide on $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.
- We see n samples from a distribution $p(x, y)$ with finite number of values for x and y and want to decide if p is independent. $H_0 : p(x, y) = p(x)p(y)$, $H_1 = p(x, y) \neq p(x)p(y)$.
- We see n_1 samples from a distribution $p(x)$ and n_2 from a distribution $p(y)$ and want to decide if they are the same. $H_0 : p(x) = q(x)$, $H_1 = p(x) \neq q(x)$.
- Test if a given distribution is Normal.
- Test if a correlation coefficient is non-zero.

Definitions:

- A statistical test $\delta(x)$ takes as input x and reject the hypothesis if $x \in D_1(\delta)$ (accepts otherwise).

- The power function $\pi(\theta)$ is the probability of rejection given that the parameter value is θ .
- A test is said to have level of significance α_0 if $\pi(\theta) \leq \alpha_0$ for $\theta \in \Omega_0$.
- The size of a test is defined as $\alpha(\delta) = \max_{\theta \in \Omega_0} \pi(\theta)$. It is the maximum probability of falsely rejecting Ω_0 . This δ is an α_0 test iff $\alpha_0 \geq \alpha(\delta)$.
- p-value. Assume we have a set of tests for a given hypothesis, with varying levels of α . For a given observation x , the *p-value* is the smallest value of α such that a size α test will reject. $p\text{-value} = \inf\{\alpha' : x \in D_1(\delta), \alpha(\delta) = \alpha'\}$. Typically the rejection region shrinks as α decreases.

The structure of a test is usually as follows: choose a statistic $T(x)$ and reject if $T(x) > c$ (i.e., set D_1 to be this region). Choose c by requiring $\max_{\theta \in \Omega_0} p(D_1|\theta) = \alpha$. This will make it a size α test.

Example: Lets x^n be a sample from a $\mathcal{N}(\mu, \sigma^2)$ with known σ^2 . We want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$ Consider a test of the of form $\bar{X} \geq c$.

The power function is (here Z is a standard normal $cN(0, 1)$).

$$\begin{aligned} \pi(\theta) = p(\bar{X} \geq c|\mu) &= p\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \geq \frac{\sqrt{n}(c - \mu)}{\sigma} \mid \mu\right) \\ &= p\left(Z \geq \frac{\sqrt{n}(c - \mu)}{\sigma} \mid \mu\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right) \end{aligned}$$

This is an increasing function in μ so that its maximum over $\mu \in H_0$ is attained at $\mu = 0$. We thus have:

$$\alpha(\delta_c) = \sup_{\mu \in H_0} \pi(\theta) = 1 - \Phi\left(\frac{\sqrt{nc}}{\sigma}\right) \quad (5)$$

So to obtain a size α test we will need $c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}$. Equivalently, we reject if $\frac{\sqrt{n}(\bar{X}-0)}{\sigma} \geq \Phi^{-1}(1 - \alpha)$ so we first turn the sample mean into a standard normal variable and then threshold it. The p-value in this case would result from solving this with equality, so $p = 1 - \Phi\left(\frac{\sqrt{n}\bar{X}}{\sigma}\right)$.

When we don't know the variance σ^2 we need to estimate it from samples and then divide by it. The resulting statistic follows a t distribution.

3 Confidence Intervals

Thus far we considered estimators whose goal was to give the exact values of θ . When θ is continuous this is not very realistic and it would be better

to have an interval in which θ is likely to lie. Such an interval is called a confidence interval.

Definition: We say that $C(x^n) = [a(x^n), b(x^n)]$ is a confidence interval for θ with confidence level $1 - \alpha$ if for all θ

$$\mathbb{P}_\theta [\theta \notin C(x^n)] \leq \alpha \quad (6)$$

We can treat $C(x^n)$ as a prediction of the location of θ . Either it succeeds or fails given θ and a sample x^n from it. What the confidence interval says is that this prediction has a probability less than α of being wrong (i.e., there will some x^n samples for which θ will be outside the interval, but the probability of seeing those is low).

To find such intervals we need to use inequalities that tell us how far empirical averages are from their means.

Theorem (Special case of Hoeffding's inequality: Given n independent observations X_1, \dots, X_n from a Bernouli distribution with expected value p , for every $\epsilon > 0$ it holds that

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2} \quad (7)$$

Lets use this to derive a $1 - \alpha$ confidence interval. We'd like to find an ϵ such that

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2} = \alpha \quad (8)$$

So

$$\epsilon_\alpha = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \quad (9)$$

So $C(x^n) = [\bar{X} - \epsilon_\alpha, \bar{X} + \epsilon_\alpha]$ is a confidence interval for p .

4 Non-Parametric Methods

So far we talked about estimating a set of parameters $\theta_1, \dots, \theta_d$ of a distribution $p(x|\theta)$ give a sample x^n . Now suppose we don't want to assume anything about the shape of $p(x)$, and we don't know that it depends on a finite set of parameters. For example we know our data comes from some distribution $p(x)$ but we don't know its shape. It seems like if we have more and more data we should be able to get a more precise estimate of $p(x)$. There is also no reason to limit the number of parameters to be independent (and smaller) than n . Methods whose description of $p(x)$ uses a number of parameters that is $O(n)$ are called non-parametric methods.

4.1 Histogram Methods

Assume for now that x is continuous and we want to estimate its pdf. One approach would be to partition x space into a set of intervals B_1, \dots, B_k where the width of each interval is h , a decreasing function of n . We then estimate the relevant frequency of points within B_i so that

$$\hat{p}(x) = \frac{v_{i(x)}}{nh} \quad (10)$$

where $i(x)$ is the interval in which x lies, and v_j is the number of sample points in interval j . It's easy to see that this integrates to one.

There are two difficulties here. One is choosing h as a function of n . There are several ways of choosing it, but its clear that there is a bias variance tradeoff. One popular method is to use cross-validation to estimate how good a given choice of h is.

The other difficulty is that there is an un-natural boundary between the B_i . Seems like it would make more sense to generate a smoother curve.

4.2 Kernel Methods

The problem with histograms is that they strongly rely on the exact location of x points. What if we viewed each x point as possibly arising from points near x .

Assume we are given a function $K(x)$ such that $\int K(x)dx = 1$, $K(x) \geq 0$ and $\int K(x)x dx = 0$ (the latter is not strictly necessary for defining the method). The kernel density estimator is given by:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \quad (11)$$

It's easy to see that this integrates to one. As $h \rightarrow 0$ the distribution becomes more peaked at the sample points, and as h grows it becomes smoother. Again, there is the matter of the choice of h . Parzen windows are an example of this method.

5 Non parametric regression

In regression we have a variable Y that depends on a variable X and we want to determine $f(x) = E(Y|X = x)$. We do this from a sample $(x_1, y_1), \dots, (x_n, y_n)$. One approach is to use linear regression (or polynomial fitting). There is also a non-parametric option. Say we construct a kernel estimate of $p(x, y)$:

$$\hat{p}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right) \quad (12)$$

With this estimate we have

$$\begin{aligned} E_p(Y|X = x) &\approx E_{\hat{p}}(Y|X = x) = \frac{\sum_y \hat{p}(x, y)y}{\hat{p}(x)} \\ &= \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \end{aligned}$$

This is called the Nadaraya-Watson estimator.