STHT 24400
Lecture 13
2/18/16

Sufficiency

We have been comparing ways in which to estimate the state of nature, θ, from an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, where X_i is a random var. denoting an observation.
We have compared estimators by considering which give the most concentrated estimates (minimum SSE) from a given set of data. Now let's ask a complementary question, which is how to write an estimator (or "statistic")* that has **all** the available information about θ present in the observations.

---

* A "statistic" means a function of the observations x_i: $T(x_1, \ldots x_n)$

Ex Consider n iid Bernoulli trials $X_1, \dots X_n$, with parameter θ. Obviously $\tilde{T} = (X_1, \dots, X_n)$ has all the information about θ that can be obtained from these observations. What about

$$T = \sum_{i=1}^{n} X_i \,?$$

$$P(X_1 = x_1, \dots, X_n = x_n / T = t)$$

$$= \frac{P(X_1 = x_1, \dots X_n = x_n, T = t)}{P(T = t)}$$

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

That is: if you know $T = t$, the likelihood is <u>independent</u> <u>of</u> θ, so $T = \sum_{i=1}^{n} X_i$ contains all available information about θ that is in the observations.

Definition

A statistic $T(X_1, \dots X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \dots, X_n$ given $T = t$ does not depend on θ for any value of t.

Factorization Theorem (Neyman):

A necessary and sufficient condition for $T(X_1, \dots, X_n)$ to be sufficient for a parameter θ is that the joint {density OR pmf} factors as follows:

$$f(x_1, \dots, x_n \mid \theta) = g\left[T(x_1, \dots, x_n), \theta\right] h(x_1, \dots, x_n).$$

Pf For discrete case.

We'll write $\vec{X} = (X_1, \ldots, X_n)$

$\vec{x} = (x_1, \ldots, x_n)$

1^{st}, show that factorization $\rightarrow$ sufficiency:

we'll need: $$P(T = t) = \sum_{T(x)=t} P(\vec{X} = x)$$

$$= g(t, \theta) \sum_{T(\vec{x})=t} h(\vec{x})$$

so then

$$P(\vec{X} = \vec{x} \mid T = t) = \frac{P(\vec{X} = \vec{x}, T = t)}{P(T = t)}$$

$$= \frac{h(\vec{x})}{\sum_{T(\vec{x})=t} h(\vec{x})}$$ doesn't depend on θ!

sufficiency $\rightarrow$ factorization:

Suppose $p(\vec{X} \mid T)$ is indep of θ. Let

$$g(t, \theta) = P(T = t \mid \theta)$$

$$h(\vec{x}) = P(\vec{X} = \vec{x} \mid T = t)$$

Then:

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

$$= g(t, \theta)\, h(\vec{x})$$

Ex: Same as before: Bernoulli iid rv's $X_1, \ldots, X_n$. Let's factor:

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

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

$$= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}(1-\theta)^n$$

$$\boxed{t = \sum_{i=1}^{n}}$$

$$= \underbrace{\left(\frac{\theta}{1-\theta}\right)^{t}(1-\theta)^n}_{g(T(\vec{x}),\theta)} \cdot \underbrace{1}_{h(x)}$$

Ex: Let's revisit the Normal example from last time from a sufficiency perspective:
We sample $(X_1, \ldots X_n)$ from a Normal distribution. Want to find μ and σ:

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(x_i-\mu)^2}$$

$$f(x_1,\dots,x_n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)}$$

Let $\sum_{i=1}^{n} x_i^2 = t_1$; $\sum_{i=1}^{n} x_i = t_2$

(We have a 2-dimensional estimation problem here, but the factorization thm. still holds].

We now have

$$f(x_1,\dots,x_n|\mu,\sigma) = \underbrace{\frac{1}{\sigma^n (2\pi)^{n/2}} e^{\frac{-1}{2\sigma^2}(t_1 - 2\mu t_2 + n\mu^2)}}_{g(T(\vec{x},\theta))} \cdot \underbrace{1}_{h(x)}$$

Note that we call $T(\vec{x})$ a <u>statistic</u> not an <u>estimator</u>. So far, it is just a package to put $\vec{x}$ into.

What is the relationship between sufficient statistics and estimators?

Answer:

If T is sufficient for θ, then the MLE $\hat{\theta}$ is a function of T.

Pf $f(\vec{x}|\theta) = L(\theta) = g(T,\theta)h(x)$,

so to maximize $L(\theta)$, we need only to maximize $g(T,\theta)$.

In fact, it is possible to do much better:

Thm (Rao-Blackwell)

Let θ^* be an estimator of θ with $E(\theta^{*2}) < \infty$ for all θ. Suppose T is sufficient for θ, and set $\tilde{\theta} = E(\theta^*|T)$. Then for all θ,

$$E[(\tilde{\theta}-\theta)^2] \leq E[(\theta^*-\theta)^2]$$

(inequality is strict unless $\theta^* = \tilde{\theta}$)

To prove this, we need a small lemma (Rico, p. 151)

$$\boxed{Var(Y) = Var[E(Y/x)] + E[Var(Y/x)]}$$

pf $Var(Y/x) = [E(Y^2/x)] - [E(Y/x)]^2$

$$E[Var(Y/X)] = E[E(Y^2/X)] - E[[E(Y/X)]^2]$$

furthermore,

$$Var[E(Y/X)] = E[[E(Y/X)]^2] - [E[E(Y/X)]]^2$$

and because $E(Y) = E[E(Y/X)]$, (Rico, p. 119) we can write

$$Var(Y) = E(Y^2) - [E(Y)]^2$$
$$= E[E(Y^2/x)] - [E[E(Y/x)]]^2$$

so:

$$Var(Y) = E[E(Y^2/X)] - [E[E(Y/X)]]^2$$
$$= E[E(Y^2/X)] - E[[E(Y/X)]^2] + E[[E(Y/X)]^2] - [E[E(Y/X)]]^2$$
$$= E[Var(Y/X)] + Var[E(Y/X)]$$

$$E[(\tilde{\theta}-\theta)^2] \le E[(\theta^*-\theta)^2], \quad \tilde{\theta} = E(\theta^*/T)$$

Pf

$$E(\tilde{\theta}) = E[E(\theta^*/T)] = E(\theta^*)$$

Hence to compare the SSE's, we need only compare the variances.

$$Var(\theta^*) = Var[E(\theta^*/T)] + E[Var(\theta^*/T)]$$

$$Var(\theta^*) = Var(\tilde{\theta}) + E[Var(\theta^*/T)]$$

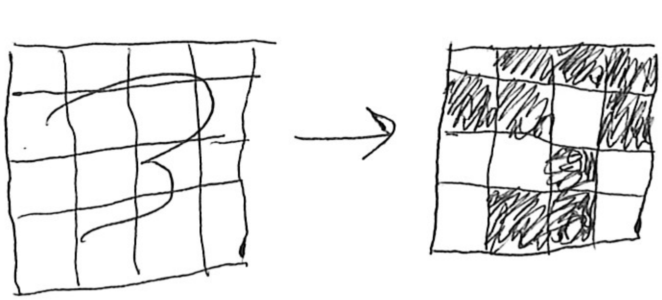
so $Var(\theta^*) > Var(\tilde{\theta})$ unless $Var(\hat{\theta}/T) = 0$, which is only true if θ^* is a function of T, implying $\theta^* = \tilde{\theta}$

STAT 24400
Lecture 12
[illegible]

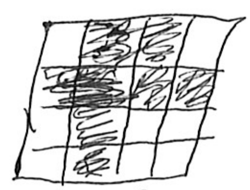
Topic IV:
Hypothesis Testing

Sometimes settling for less information gives us very useful techniques. By giving up the full picture of the posterior distribution, we got the MLE and Fisher's Thm. Now, instead of looking for parameter values ($\hat{\theta}$), let's use the parameters to make decisions.

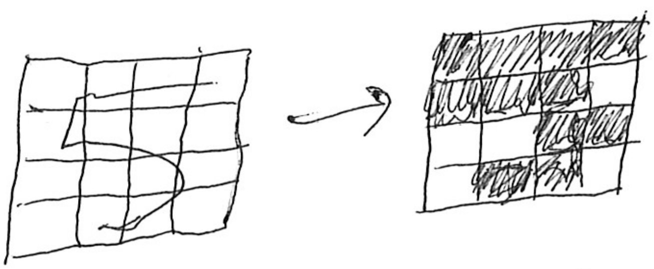
Example: Pattern recognition



θ = "3" or "5"

X = 

a set of pixels.



observe X, decide which θ, knowing $p(x|\theta)$ from trials.

Ex. Acceptance Sampling

X_i exponential λ, $E(X_i) = \frac{1}{\lambda}$

$i = 1, \ldots, n$ inspected. Is $\lambda \leq 0.1$

(lot is "good") or $\lambda > 0.1$ ("bad")?

Ex Contingency Tables

$n = 205$ couples

		Wife T	Wife M	Wife S	
Husband	T	18	28	14	
	M	20	51	28	
	S	12	25	9	
					205

Is there 'selection based on height' or are heights independent?

Test if

$$P(T \cap T) = P(T) \cdot P(T)$$

etc $$\theta_{TT} = \theta_{HT} \cdot \theta_{WT}$$

Testing $\underline{\text{Simple}}^*$ Hypotheses

$*$ means "Distribution of data is completely specified, with no parameters to estimate".

X data

$f(x|\theta)$ model

$H_0: \theta = \theta_0$, or dist. of X is $f(x|\theta_0)$

$H_1: \theta = \theta_1$, or dist of X is $f(x|\theta_1)$

Neyman - Pearson Lemma: Best test to use is Likelihood ratio (LR) test.

Reject H_0 if $\dfrac{f(x|\theta_1)}{f(x|\theta_2)} > K$.

$\alpha = P(\text{Rej. } H_0 \mid H_0 \text{ true})$

$\beta = P(\text{Acc. } H_0 \mid H_1 \text{ true})$

$\pi = 1 - \beta = \underline{\text{power}}$ of the test.

The LR $\dfrac{f(x|\theta_1)}{f(x|\theta_2)}$ orders x values

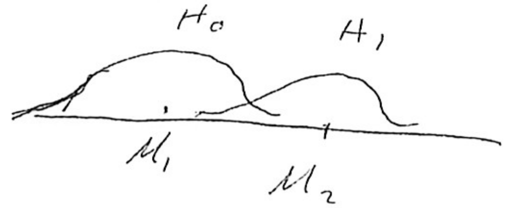
high LR is stronger evidence for H_1
low LR is " " " H_0

K draws the line

Example $\quad X_1, \ldots, X_n \quad N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$

$H_1: \mu = \mu_1$



$(\sigma^2 = \sigma_0^2 \text{ given})$

$$\frac{f(x_1, \ldots, x_n \mid \mu_1)}{f(x_1, \ldots, x_n \mid \mu_0)} = e^{\frac{1}{\sigma_0^2}\left[(\mu_1 - \mu_0)\sum x_i\right]} e^{-\frac{n}{2\sigma_0^2}\left[\mu_1^2 - \mu_0^2\right]}$$

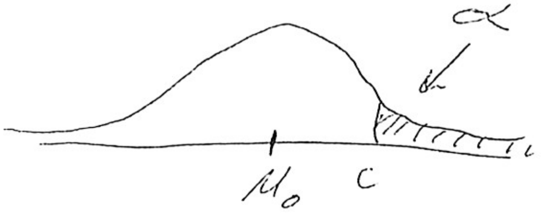
large when $(\mu_1 - \mu_0)\sum x_i$ is large

$\hookrightarrow = (\mu_1 - \mu_0) \cdot n\bar{x}$

So we obtain the test,

Reject H_0 if $\bar{x} > c$,

where $P(\bar{x} > c \mid \mu = \mu_0) = \alpha$



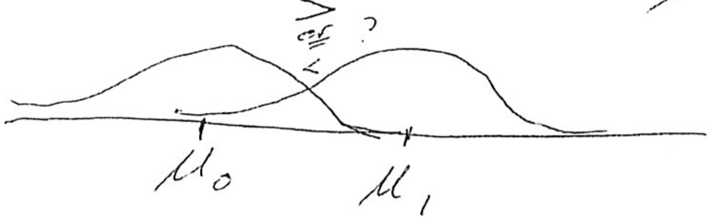
$$c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

Restricted solution <u>may</u> solve more general problem

Ex $X_1, \ldots, X_n$ $N(\mu, \sigma_0^2)$, σ_0^2 known

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1 > \mu_0$$



Test: Reject H_0 if $\bar{X} > C = \mu_0 + Z_\alpha \frac{\sigma_0}{\sqrt{n}}$

<u>Note</u>

Same test for <u>any</u> $\mu_1 > \mu_0$

But the <u>power</u> depends on μ_1

The test is <u>uniformly</u> <u>most</u> <u>powerful</u>

Describe peformance with <u>power</u> <u>function</u>

$$\pi(\mu_1) = Pr(\text{Reject } H_0 \mid \mu = \mu_1) = Pr(\bar{X} > c \mid \mu = \mu_1)$$