STAT 24400
Lecture 6
10/13/16

Inference

Much of physical and social science is devoted to deducing effects from causes using a model. In constructing such models, we often need to deduce causes (and hence the model) from effects (the data). This is the task of statistical inference.

Example: Polygraph Testing

so-called "Lie Detector"

Measure: Respiration Rate
Heart Rate
Blood Pressure
skin Conductivity (="sweat")

Combine into a score

Compare score to a threshold

If (score > threshold)
score as liar

Used for screening
" " investigating specific cases.

How effective is it?

Can depend on population being sampled.

The results depend on

$P(\text{"Guilty"})$ for population.

For a specific case, maybe

$P(\text{Deception}) = \frac{1}{2}$ (?)

For screening, maybe

$P(\text{Deception}) = \frac{1}{1000}$ (?)

smaller (?)

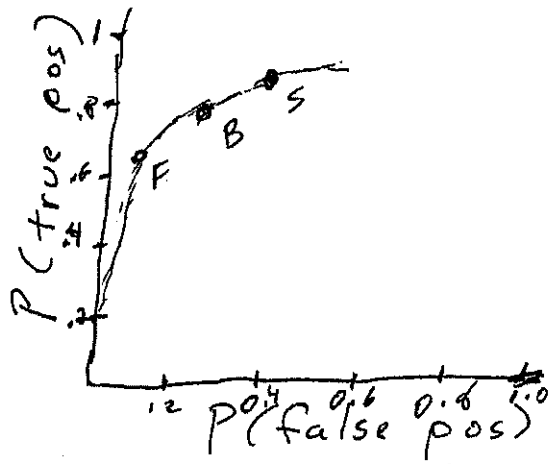
1 in 1000 are spies.

If detection threshold is set to detect the great majority (80%) of spies:

Test Result	Examinee's True Condition: Spy	Non-spy	Total
Fail test	8	1,598	1,606
Pass test	2	8,392	8,394
Total	10	9,990	10,000

If detection threshold is set to greatly reduce false positives:

Test Result	Spy	Non-spy	Total
Fail test	2	39	41
Pass test	8	9,951	9,959
Total	10	9,990	10,000



Illustrative plot of probabilities of responses for various thresholds
F = "Friendly" B = "Balanced" S = "Suspicious"

Inference - reasoning from effect to cause

Deduction - reasoning from cause to effect

Bayesian Inference

Based on far reaching consequence of a deceptively simple relationship.

Recall the definition of conditional probability:

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

of course,

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Hence

$$P(F|E)\,P(E) = P(E|F)\,P(F)$$

so

$$P(E|F) = \frac{P(F|E)\,P(E)}{P(F)}$$

Bayes's Theorem

Hypotheses: There are n possible causes $E_1, E_2, \ldots, E_n$

These are mutually exclusive, so $(E_i \cap E_j) = \phi \quad i \neq j$

They are exhaustive, so $\Omega = \bigcup_{i=1}^{n} E_i$

Given the a priori probabilities $P(E_1), P(E_2), \ldots P(E_n)$, and the observed effect F, and the conditioned probabilities $P(F/E_1), P(F/E_2), \ldots, P(F/E_n)$

(the prob. of effect given causes)

Find: $P(E_i/F)$ for $i = 1, 2, \ldots n$

a posteriori probability of cause given effect

Ex. 1: Diagnostic Testing for HIV

Patient tested for HIV.

Possible "causes"

E_1 = patient has HIV

E_2 = patient does not have HIV

If patient selected at random

$P(E_1)$ = prevalence of HIV

$P(E_2) = 1 - P(E_1)$

"effect" F is a positive reaction

suppose that (hypothetical values, correct order of mag)

$P(F|E_1) = 0.9$

$P(F|E_2) = 0.2$] — from empirical trials

Want:

$$P(E_1|F)$$

Back to:

Bayes's Theorem

We have, from the definitions,

$$P(E_1|F) = \frac{P(F|E_1)\,P(E_1)}{P(F)}$$

Need $P(F)$.

$$F = (F \cap E_1) + (F \cap E_2) + \cdots + (F \cap E_n)$$

$$\begin{aligned} P(F) &= P(F \cap E_1) + P(F \cap E_2) + \cdots + P(F \cap E_n) \\ &= P(F|E_1)P(E_1) + \cdots + P(F|E_n)P(E_n) \\ &= \sum_{i=1}^{n} P(F|E_i)P(E_i) \end{aligned}$$

Theorem:

$$P(E_1|F) = \frac{P(F|E_1)\,P(E_1)}{\sum_{i=1}^{n} P(F|E_i)P(E_i)}$$

Ex. 1: Diagnostic Test for HIV

$P(F/E_1) = .9$ true positives

$P(F/E_2) = .2$ false positives

Suppose test is positive.

$$P(E_1|F) = \frac{P(F/E_1)\,P(E_1)}{P(F/E_1)\,P(E_1) + P(F/E_2)\,P(E_2)}$$

$$= \frac{(0.9)\theta}{(0.9)\theta + (0.2)(1-\theta)} \qquad \theta = P(E_1)$$

For randomly selected patient

θ = prevalence, say $\theta = 0.001$

Then $P(E_1|F) = \frac{(0.9)(0.001)}{(0.9)(0.001) + (0.2)(0.999)} = \boxed{.0045}$

If patient seriously at risk, say $\theta = 0.5$

$$p(E_1/F) = \frac{(0.9)(0.5)}{(0.9)(0.5) + (0.1)(0.5)} = \boxed{.8182}$$

Ex. 3: (Ancient history, apocryphal, but great example)

Alan Dershowitz on "Larry King Live": "Only about $\frac{1}{10}$ of 1% of wife-batterers actually murder their wives."

Q: Is $P(OJ \text{ guilty}) = \frac{1}{10}$ of $1\% \; (0.001)$?

We condition all probabilities on husband being a batterer.

G = husband guilty of murder

M = Wife murdered

Dershowitz: $P(G) = 0.001$

Want $P(G/M)$

$P(M/G) = 0.9999 \quad (\sim 1)$

$P(M/G^c) = 0.0001$ (US statistics approx)

$$P(G/M) = \frac{P(M/G)\,P(G)}{P(M/G)\,P(G) + P(M/G^c)\,P(G^c)}$$

$$= \frac{(1.0)(0.001)}{1\,(0.001) + (0.0001)(0.999)} = \boxed{.91}$$

The Beta Distribution(s)

First, need Gamma Function

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

generalization of "factorial" to non-integers. For integers

$$\Gamma(n) = (n-1)! = (n-1)(n-2)\cdots 2\cdot 1$$

For non-integers, $\Gamma(a)$ still has the factorial property:

$$\Gamma(a) = (a-1)\Gamma(a-1).$$

Almost always we'll be considering cases where a is a positive integer, so

$$\Gamma(a) = (a-1)!$$

We'll need $\Gamma(a)$ to normalize the β distribution

The

Beta Distribution is given by

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

α and β are positive real parameters. They can be integers, but don't have to be. You could also write $f_X(x)$ or $f_X(x;\alpha,\beta)$.

If α and β are integers,

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} = \frac{(\alpha+\beta-2)!(\alpha+\beta-1)}{(\alpha-1)!(\beta-1)!}$$

$$= \binom{\alpha+\beta-2}{\alpha-1}(\alpha+\beta-1)$$

Why "Beta" Distribution?

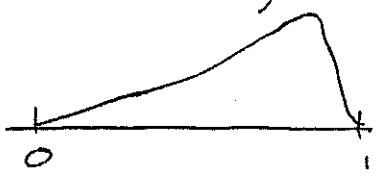
$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is called the "Beta function",

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

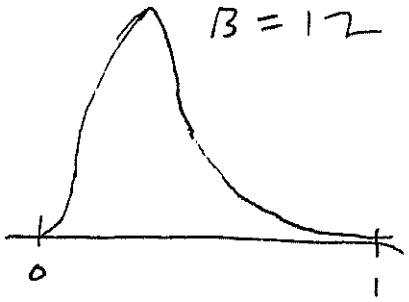
(how Beta distribution is normalized,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$\alpha = 3,\ \beta = 2$

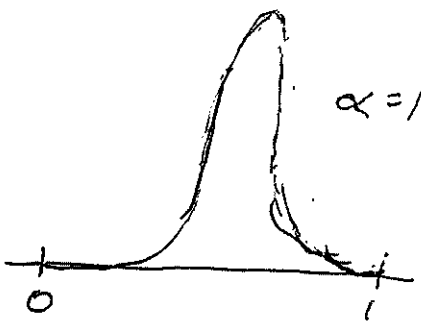


$\alpha = 5$
$\beta = 12$

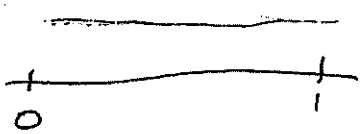


$\alpha = \beta = 3$

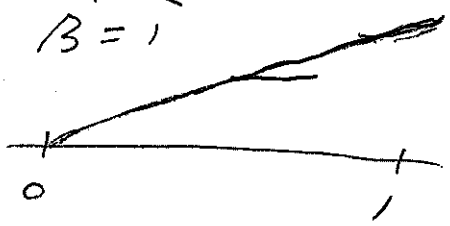
$\alpha = \beta = 10$



$\alpha = \beta = 1$



$\alpha = 2$
$\beta = 1$



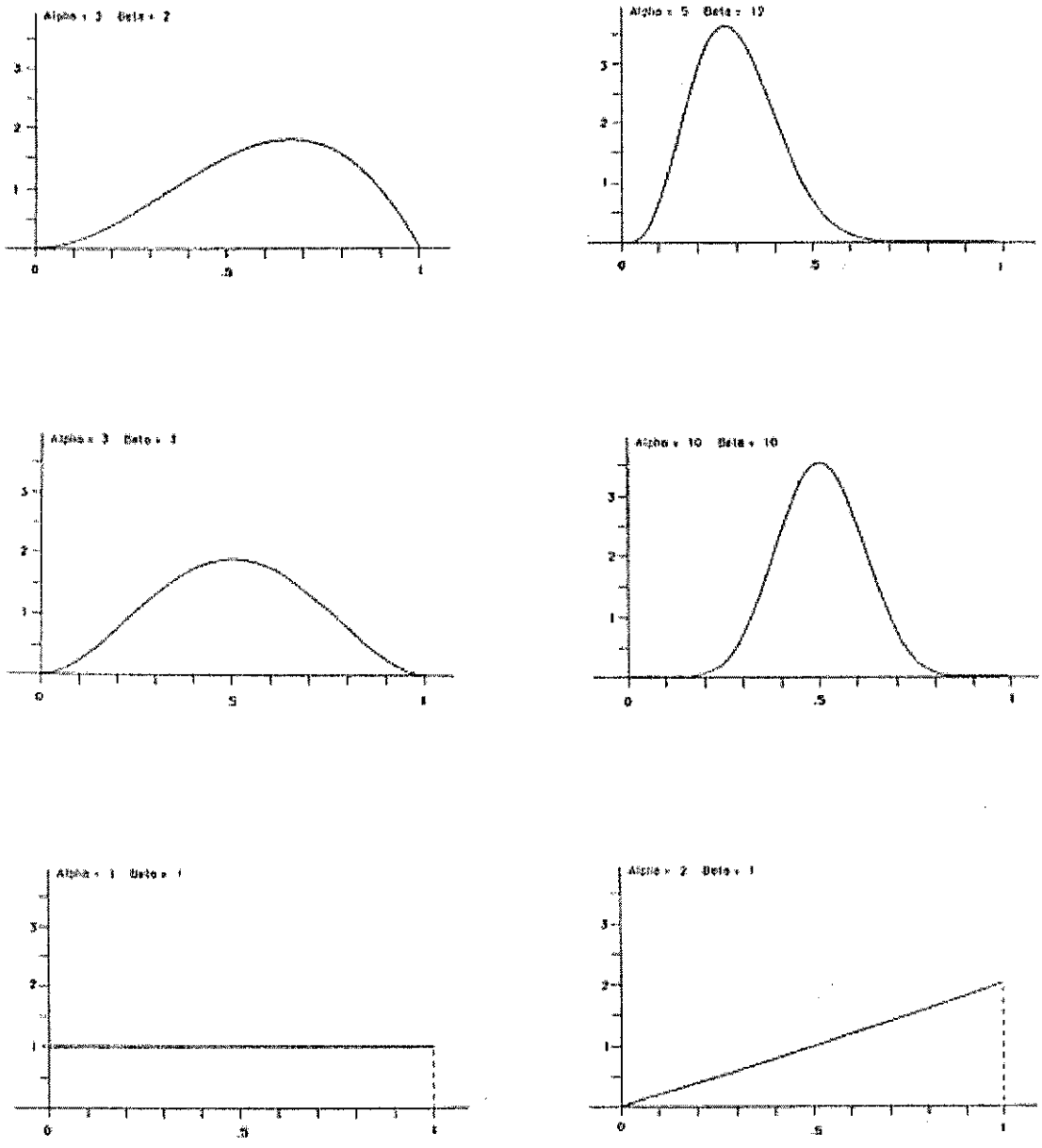


Figure 2.3. Examples of Beta densities.

12A

For the β distribution,

$$E(x) = \int_0^1 x \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

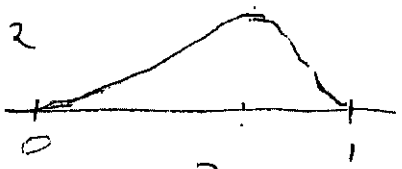
$$= \frac{\Gamma(\alpha+\beta+1)\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)\Gamma(\alpha+1)} \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}\,dx$$

$$= \underbrace{\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)\Gamma(\alpha)}}_{= \frac{\alpha}{\alpha+\beta}} \underbrace{\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}\,dx}_{= 1 \text{ Beta}(\alpha+1,\beta)}$$

$$\Rightarrow \boxed{E(x) = \frac{\alpha}{\alpha+\beta}}$$

Two of the previous examples:

$\alpha = 3$
$\beta = 2$



$$E(X) = \frac{3}{3+2} = \frac{3}{5} = 0.6$$

$$f(x) = \frac{\Gamma(3+2)}{\Gamma(3)\Gamma(2)} x^{3-1}(1-x)^{2-1}$$

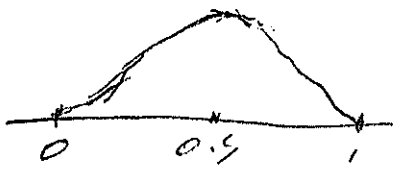
$$= \frac{4!}{2!\,1!} x^2(1-x)^1$$

$$= 12x^2(1-x)$$

max at $\frac{2}{3}$

(aka the "mode")

$\alpha = \beta = 3$



$$E(X) = \frac{3}{3+3} = \frac{1}{2}$$

(if $\alpha = \beta$, $E(X) = 0.5$)

max at $\frac{1}{2}$

We'll need the Beta distribution in a little while, but first we need to take another look at densities and at Bayes's Theorem when densities are involved.

Remarks about Densities

Normal: $\left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{x^2}{2}}$

Exponential: $(\theta)\, e^{-\theta x}$

Beta: $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) x^{\alpha-1}(1-x)^{\beta-1}$

In each case,

$f(x) = (\text{constant}) \cdot (\text{part involving } x)$

constant needed so $\int f(x)\,dx = 1$

part involving x: shape, dependence on x

To get actual probabilities, we need both parts. To identify a distribution, need only 2nd part.

Ex If we know $f(x) = \text{const} \cdot x^{17}(1-x)^{30}$ we know its Beta(18, 31)

Ex If we know $f(x) = \text{const} \cdot e^{-x^2/2}$ we know its $N(0,1)$

Recall Bayes's Theorem

E_i, $i = 1, 2, 3, \ldots, n$ "causes" "states of nature"

F "effect" or "data"

$$P(E_i|F) = \frac{P(E_i)\,P(F|E_i)}{P(F)}$$

$$P(F) = \sum_{j=1}^{n} P(E_j)\,P(F|E_j)$$

Equivalent form:

$$P(E_i|F) \propto P(E_i)\,P(F|E_i)$$

Posterior — Prior — "Liklihood"

i.e. $$P(E_i|F) = (\text{const.})\,P(E_i)\,P(F|E_i)$$

where "const" insures that

$$\sum_{i=1}^{n} P(E_i|F) = 1$$

in fact, $$\text{const} = \frac{1}{P(F)} = \frac{1}{\sum_{j=1}^{n} P(E_j)\,P(F|E_j)}$$

Bayes's Theorem in terms of densities:

Given $f(x|y)$ and $f_Y(y)$,

Find $f(y|x)$

y – "cause"
x – "effect" or "data"

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

also $f(x,y) = f(x|y)\, f_Y(y)$

so $$f(y|x) = \frac{f(x|y)\, f_Y(y)}{f_X(x)}$$

because $\int_{-\infty}^{\infty} f(y|x) = 1$

$$\int_{-\infty}^{\infty} \left(\frac{f(x|y)\, f_Y(y)}{f_X(x)} \right) dy = 1, \text{ or } \int_{-\infty}^{\infty} f(x|y)\, f_Y(y)\, dy = f_X(x)$$

so in the continuous case

$$f(y|x) = \frac{f(x|y)\, f_Y(y)}{\int_{-\infty}^{\infty} f(x|y)\, f_Y(y)\, dy}$$

in summary: Given x; so x is fixed, $f(y|x) \propto f(x|y) \cdot f_Y(y)$

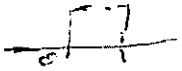
Example: Polling - a Survey

Let y = fraction of all Illinois voters for Trump

Sample n Illinois voters "at random"

Let X = # in sample who say they are for Trump

Hypotheses: $f_Y(y)$ = a priori density of y



= uniform on (0, 1), say

$$p(x|y) = \binom{n}{x} y^x (1-y)^{n-x}$$ Binomial

Observe: $n = 100$ $x = 40$

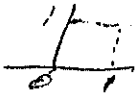
what can we infer about y?

$$f(y|x=40) \propto \underbrace{\binom{100}{40} y^{40}(1-y)^{60}}_{p(x|y)} \cdot \underbrace{1}_{f_Y(y)} \quad (\text{for } 0<y<1)$$

A Beta Density

This is a "mixed" case —

X is discrete, Y continuous



Y uniform on 0 to 1: $f_Y(y) = \begin{cases} 1 & 0<y<1 \\ 0 & \text{otherwise} \end{cases}$

Given $Y = y$,

X Binomial n trials, $\Theta = y$

$$p(x|y) = \begin{cases} \binom{n}{x} y^x (1-y)^{n-x} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

(a prob. func.)

Marginal Distribution of X?

$$p(x,y) = f_Y(y)\, p(x|y)$$

$$= \begin{cases} \binom{n}{x} y^x (1-y)^{n-x} & 0<y<1 \\ & x = 0, 1, \ldots n \\ 0 & \text{otherwise} \end{cases}$$

(prob. func. in X, density in y)

$$p_X(x) = \int_0^1 \binom{n}{x} y^x (1-y)^{n-x}\, dy$$

$$= \binom{n}{x} \int_0^1 y^x (1-y)^{n-x}\, dy$$

marg dist of X (continued)

$$p_X(x) = \binom{n}{x} \underbrace{\int_0^1 y^x (1-y)^{n-x}\, dy}$$

The is the Beta function, which we saw earlier in normalizing the Beta dist.

Explanatory Digression

[$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$

$= \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \dfrac{1}{\binom{\alpha+\beta-2}{\alpha-1}(\alpha+\beta-1)!}$

which you remember — right? ✓]

$$p_X(x) = \binom{n}{x} \cdot \frac{1}{\binom{n}{x}(n+1)}$$

$$= \frac{1}{n+1}$$

Uniform

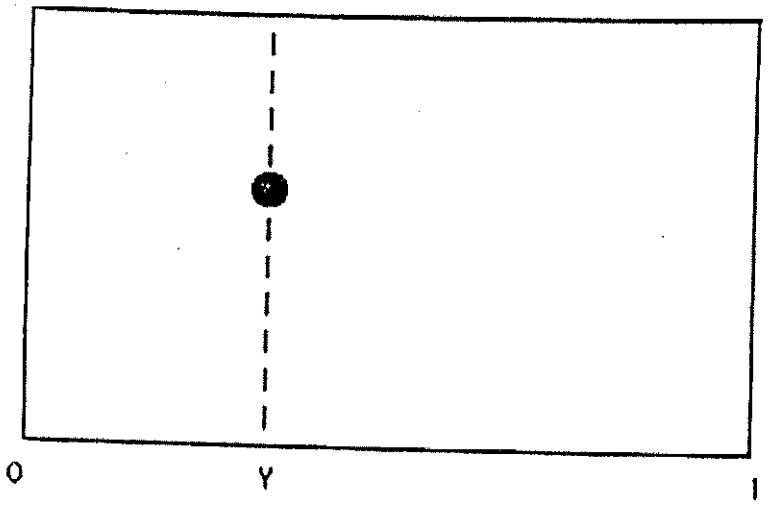


Figure 3.7. Bayes's billiard table.

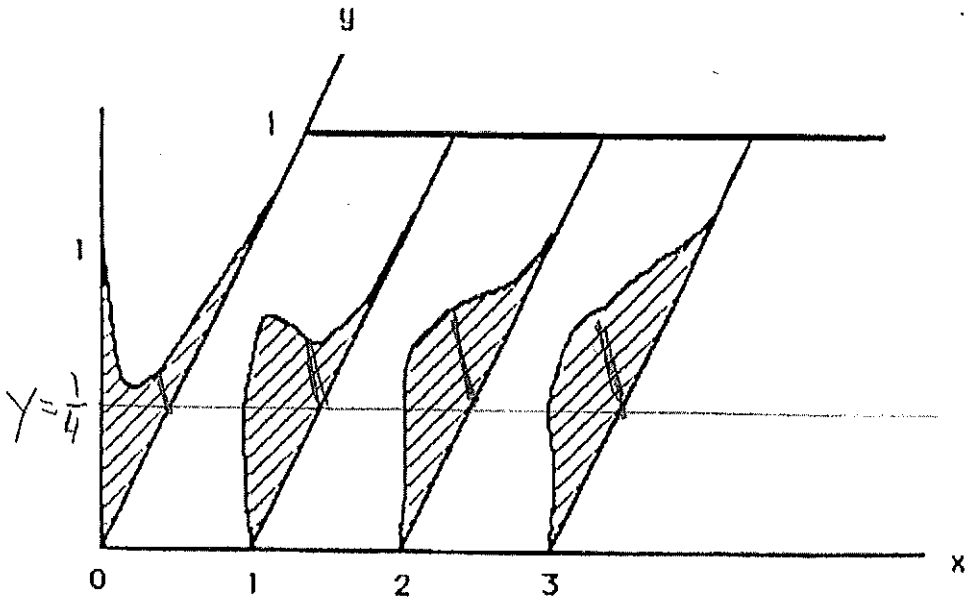
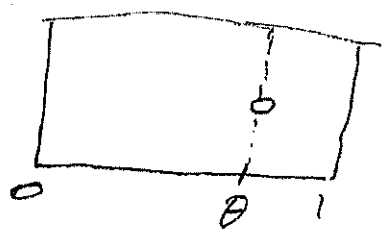


Figure 3.8

How to think about prior?

A. Billiard Table



Roll a ball. It is equally likely to stop at any point. Ball stops distance of θ from the end. Remove ball. Roll 100 more, one at a time. Count X = the number that land to the left of θ.

$$f_\theta(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x/\theta) = \binom{100}{x} \theta^x (1-\theta)^{100-x} \quad x = 0, 1, \ldots, n$$

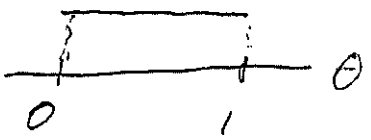
What if we are given $X = 40$?

What about θ?

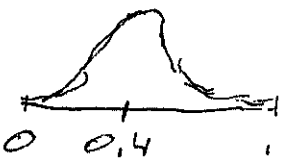
$$f(\theta \mid x) \propto \binom{100}{40} \theta^{40} (1-\theta)^{60}$$

(Beta, as previously)

Before



After



Comparison of the two examples

They are the same mathematically!

Conceptually? Is the status of $f_Y(y)$ different from $f_\Theta(\theta)$?

$f_\Theta(\theta)$ represents a tangible model

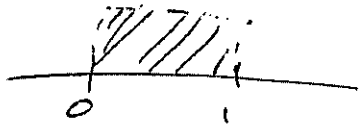
$f_Y(y)$ represents... what?
uncertainty
subjective opinion
prior knowledge

But (1) others may disagree with your assessment of f_Y

(2) Uniform f_Y may not do the job.

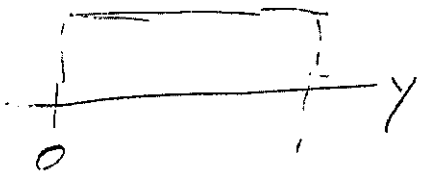
So: Hypotheses were:

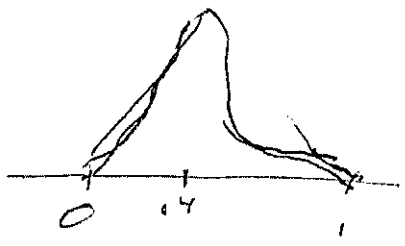
(1) y uniform on $(0,1)$ a priori



(2) Given y, X is binomial
(generated by random sampling)

Conclusion: a posteori —
after observing $x = 4/10$
y can be viewed as
having a particular Beta
distribution





BEFORE + Data = AFTER

BAYESIAN INFERENCE

Bayesian Inference, continued

STAT 244
1/26/16
Lecture 7

In the example of polling from last time, there is in fact some prior knowledge - for example, polls in other states show Clinton and Sanders close to even. We have the notion that Illinois Democrats are not that different than Democrats in other states.

How to make use of this knowledge? With a richer class of priors:

We took f_Y as Beta, so

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

(remember: the Uniform dist is a special case of the Beta dist when $\alpha = \beta = 1$)

It turns out that as α and β get bigger, and are not too different, the Beta distribution begins to resemble a Normal dist.

For $N(\mu, \sigma)$, as the

$$P(|Y - \mu| < \sigma) \approx 2/3 \approx .667$$

For Beta,

α	β	$P(\lvert Y - \mu \rvert < \sigma)$	
1	1	.577	
2	2	.626	
3	3	.644	
5	5	.659	
4	6	.661	
1	19	.812	← SKEWED!
10	10	.671	

... etc.

So suppose we expect that the vote is approximately split, with $Pr(.4 < Y < .6) \approx 2/3$.

So

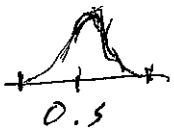
$$E(Y) \cong 0.5$$

$$Var(Y) \cong 0.1$$

For Beta distributions,

$$E(Y) = \frac{\alpha}{\alpha+\beta}, \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Set these equal to 0.5 and $(0.1)^2$, solve to get $\alpha = \beta = 12$



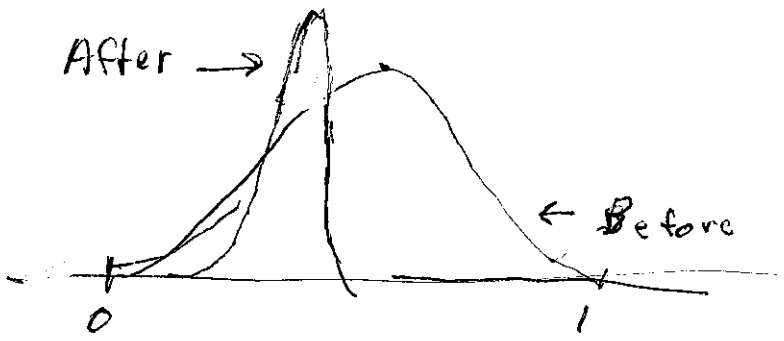
$$f_Y(y) = \frac{\Gamma(24)}{\Gamma(12)\Gamma(12)} y^{11}(1-y)^{11}$$

has $E(Y) = 0.5$, $\sqrt{Var(Y)} = 0.1$

$$f(y|x) \propto f_Y(y) f(x|y)$$

$$= (\text{constant}) y^{11}(1-y)^{11} \binom{100}{40} y^{40}(1-y)^{60}$$

$$\propto y^{51}(1-y)^{71}$$



[Beta with $\alpha = 52$, $\beta = 72$]

Before: $E(Y) = 0.5$

After: $E(Y|x=40) = \frac{52}{124} = \underline{0.42}$

For Beta distributions,

$$E(Y) = \frac{\alpha}{\alpha+\beta} = \mu_Y$$

$$Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \sigma_Y^2$$

$$\sigma_Y^2 = \left(\frac{\alpha}{(\alpha+\beta)}\right)\left(\frac{\beta}{(\alpha+\beta)}\right)\frac{1}{(\alpha+\beta+1)}$$

$$= \mu_Y$$

$$\frac{\beta}{\alpha+\beta} = \frac{\beta+\alpha-\alpha}{\alpha+\beta} = \frac{\beta+\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}$$

$$= 1 - \mu_Y$$

so

$$Var(Y) = \frac{\mu_Y(1-\mu_Y)}{\alpha+\beta-1}$$

$\mu_Y = 0.5 \quad \sigma_Y^2 = (0.1)^2$

So: $(0.1)^2 = \frac{(0.5)(1-0.5)}{\alpha+\beta+1}$

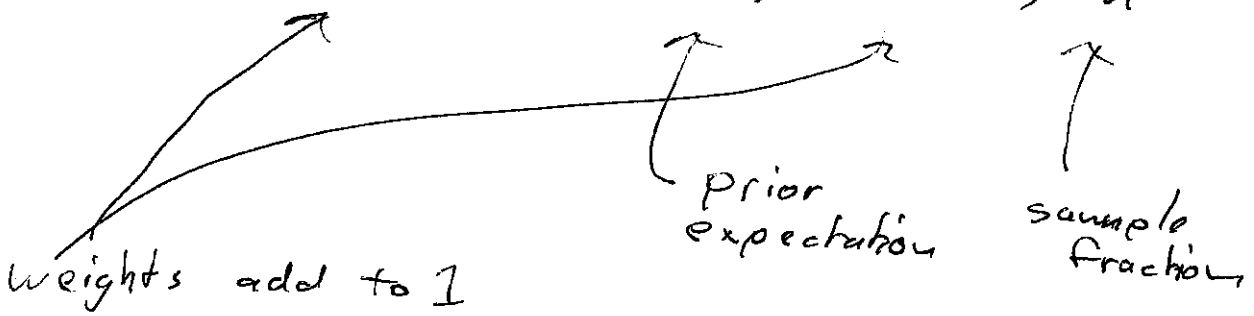
$$\alpha+\beta+1 = \frac{(0.5)^2}{(0.1)^2} = 25; \quad \alpha+\beta = 24$$

But

$$\mu_Y = \frac{\alpha}{\alpha+\beta} = 0.5 \Rightarrow \frac{\alpha}{24} = 0.5 \Rightarrow \alpha = 12, \beta = 12$$

Note: We can interpret the posterior expectation as a weighted average:

$$\frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \cdot \frac{\alpha}{\alpha + \beta} + \left(\frac{n}{\alpha + \beta + n}\right) \cdot \frac{x}{n}$$



Case 1: $\alpha + \beta$ large relative to n
("strong prior information")

Then $\frac{\alpha + \beta}{\alpha + \beta + n} \approx 1$, $\frac{n}{\alpha + \beta + n} \approx 0$

Case 2: n large relative to $\alpha + \beta$
("weak prior information")

Then $\frac{\alpha + \beta}{\alpha + \beta + n} \approx 0$, $\frac{n}{\alpha + \beta + n} \approx 1$

Case 1: $\frac{\alpha}{\alpha + \beta}$ case 2: $\frac{x}{n}$

Otherwise, a compromise!

Bayes for Normal

θ is the true value.

We take the prior $f(\theta)$

$$f(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}, \quad -\infty < \theta < \infty$$

$E(\theta) = \mu \quad Var(\theta) = \sigma^2$

X is the observed value

with error $\sim N(0, \tau^2)$ So

$X = \theta$ + error is $N(\theta, \tau^2)$

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\tau}\right)^2} \quad -\infty < x < \infty$$

likelihood

$f(\theta|x)$, the posterior will be $N(A, B^2)$

[squares completed in Stigler, chapter 4]

$$A = \frac{\tau^2\mu + \sigma^2 x}{\tau^2 + \sigma^2} \qquad B^2 = \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}$$

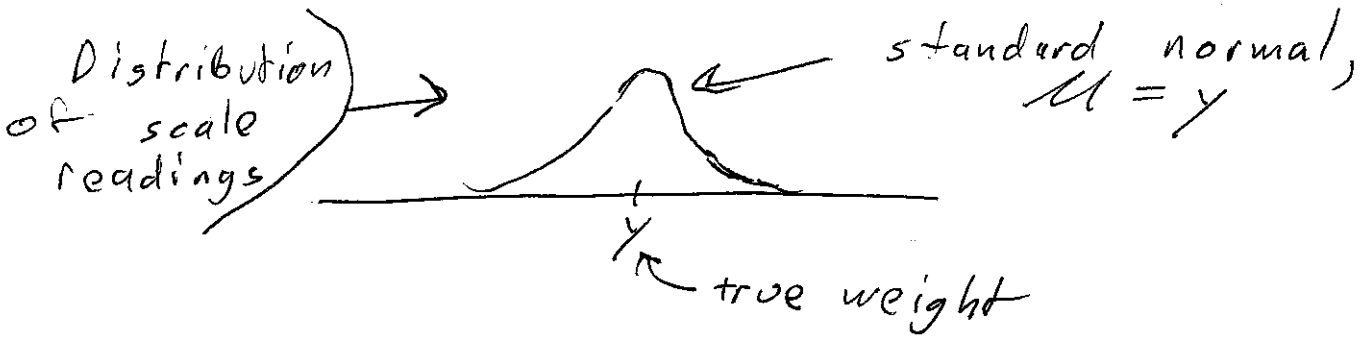
Weighted Average of μ and x

a posteriori uncertainty

Example:

Measure a weight with an imperfect scale:

Scale makes errors with std. deviation 1 kg., normally distributed:



X = recorded weight

Y = true weight

$f(x|y) \sim N(y, 1)$

$f_Y(y) \longrightarrow$??

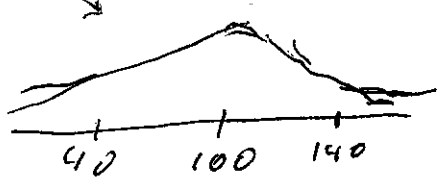
Say $N(\mu, \sigma^2)$

$\mu = 100$ kg

$\sigma^2 = (10)^2 = 100$

[Why? May be have rough idea, say from number of people needed to lift it]

$f_Y(y)$



(Aerion: $P(90 < Y \leq 110)$, $P(|Y - 100| \leq 10)$ $\approx \frac{2}{3}$

Given "data" x, want $f(y|x)$.

$$f(y|x) \propto f_Y(y) f(x|y)$$

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

$$\propto e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} - \frac{1}{2}(x-y)^2}$$

messy completion of squares not shown

$$= e^{\frac{-1}{2\sigma^2}\left[(y-\mu)^2 + \sigma^2(x-y)^2\right]}$$

$$\propto e^{-\frac{1}{2}\frac{(y-A)^2}{B}}$$ ← functional form, aka "Business Part"

$$A = \frac{x\sigma^2}{\sigma^2+1} + \frac{\mu \cdot 1}{\sigma^2+1}$$

weighted average of x & μ.

$$B = \frac{\sigma^2}{(\sigma^2+1)}$$

$f(y|x)$ is $\mathcal{N}(A, B)$

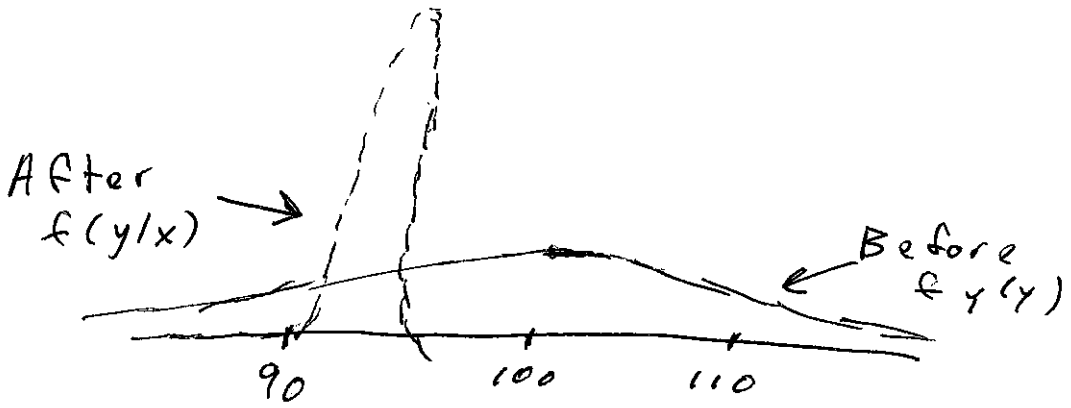
$E(Y|X) = A$ (between x and μ)

If σ^2 small (much prior info)

A near μ

If σ^2 large (little prior info)

A near X



case for $\sigma^2 = (10)^2$, $\mu = 100$

$X = 90$

That is:

$f_Y(y) \quad N(100, 10^2)$

$f(x|y) \quad N(y, 1)$

$$A = \frac{100}{101} \cdot x + \frac{1}{101} \cdot \mu = 90.9$$

$$B = \frac{100}{101}$$

$$f(y|x) \quad N\left(90.9, \frac{100}{101}\right)$$

Bayes's Theorem Processes Information!

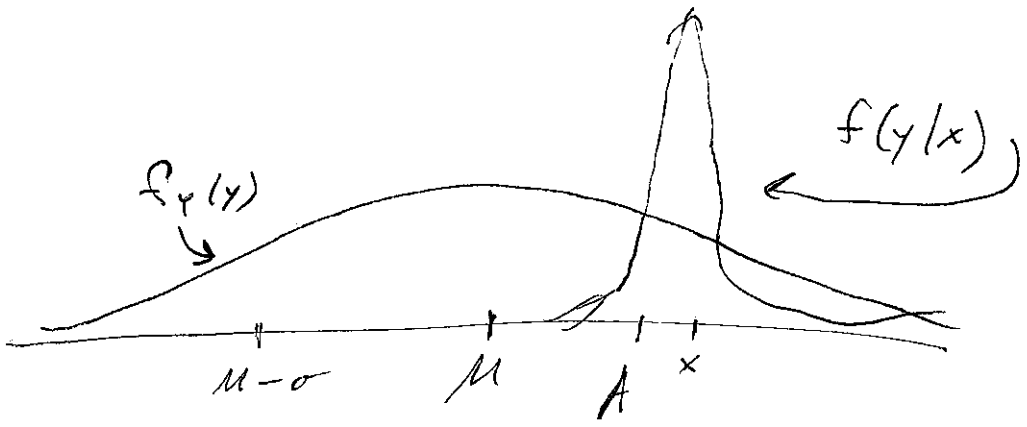
In summary, for the normal dist, we ha

$$f(y) \quad \mathcal{N}(\mu, \sigma^2)$$
$$f(x|y) \quad \mathcal{N}(y, 1)$$
$$f(y|x) \quad \mathcal{N}(A, B)$$

$$A = x \cdot \frac{\sigma^2}{\sigma^2+1} + \mu \cdot \frac{1}{\sigma^2+1}$$

$$B = \frac{\sigma^2}{\sigma^2+1}$$

(so if $\lambda = \frac{\sigma^2}{\sigma^2+1}$, $A = x \cdot \lambda + \mu(1-\lambda)$)



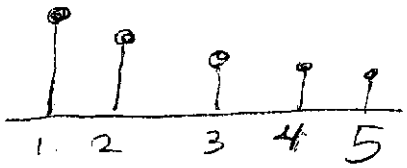
Review (for Midterm)

STAT 24400
Lecture 8
10/20/16

X a random variable

Discrete Case

$p(x) = P(X = x)$

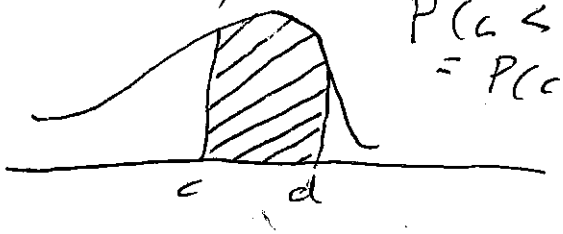


Ex Binomial
$b(x; n, \theta)$
(aka $Bin(x; n, \theta)$)
Negative Binomial
Bernoulli
Geometric
Poisson

Continuous Case

density $f(x)$

$P(c < X \le d) = P(c \le X \le d)$



Ex: Uniform
Exponential
Beta
Normal

cdf $F(x) = P(X \le x)$

Transformations

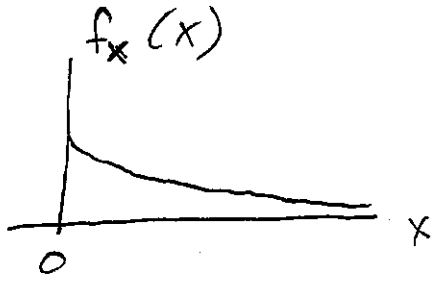
Given distribution of X ($p_X(x)$ or $f_X(x)$)

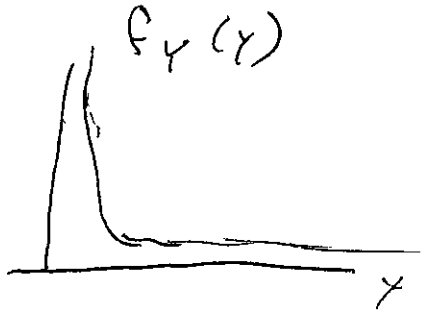
Find distribution of $Y = h(X)$ ($p_Y(y)$ or $f_Y(y)$)

Ex: $Y = X^2$ $Y = \log X$ etc

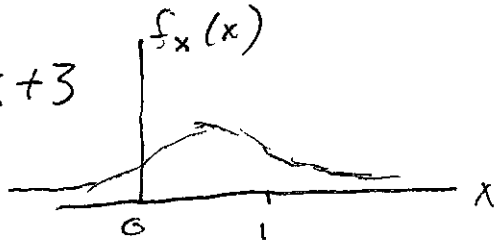
Transformation of Random Vars, continued - $Y = h(X)$

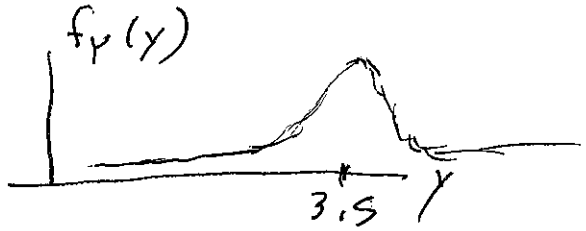
$Y = X^2$



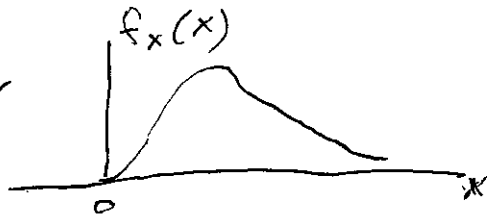


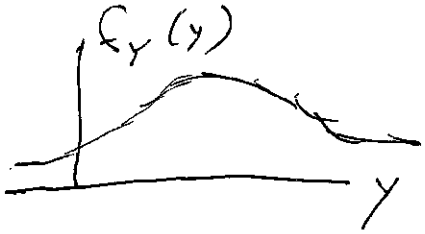
$Y = \frac{1}{2}X + 3$





$Y = \ln X$





$Y = h(X) \qquad X = g(Y)$

$$f_Y(y) = f_X(g(y)) \cdot |g'(y)|$$

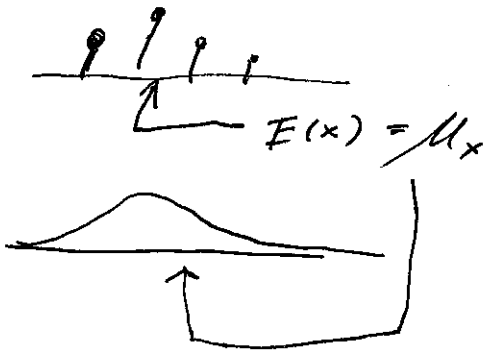
Discrete case: $p_X(x) = P_X(X = x)$

$$p_Y(y) = p_X(g(y))$$

Expectations

— "Center of Gravity"

$$E(X) = \begin{cases} \sum_{\text{all } x} x\, p(x) \\ \int_{-\infty}^{\infty} x\, f_x(x)\, dx \end{cases}$$



$$E(h(X)) = \begin{cases} \sum_{\text{all } x} h(x)\, p_x(x) \\ \int_{-\infty}^{\infty} h(x) f(x)\, dx \end{cases}$$

$E(h(X)) \neq h(E(X))$ UNLESS h linear or X constant

Variance — "spread", "dispersion"

$$Var(X) = E\left((X-\mu_x)^2\right) = E(X^2) - \left(E(X)\right)^2$$

Standard Deviation $X = \sqrt{Var(X)} = \sigma_x$

$$E(aX+b) = aE(X) + b$$

[or $\mu_{aX+b} = a\mu_x + b$]

$$Var(aX+b) = a^2 Var(X)$$

[or $\sigma^2_{aX+b} = a^2\sigma_x^2$]

Special case: $a = \frac{1}{\sigma_x}$, $b = -\frac{\mu_x}{\sigma_x}$

Then: $aX+b = \frac{X-\mu_x}{\sigma}$ — Standardized Form

Standardized form (cont'd)

So:

$E(aX+b) = 0$

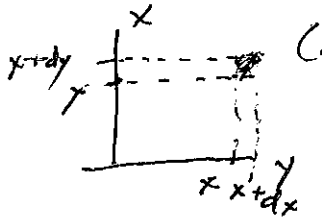
$Var(aX+b) = 1$

Multivariate Distributions

Bivariate

Discrete: $p(x,y) = P(X = x \text{ and } Y = y)$

Continuous: $f(x,y)\,dx\,dy = P(x \le X \le x+dx,\ y \le Y \le y+dy)$



Marginal Distributions ("side view")

$$p_Y(y) = \sum_{\text{all } x} p(x,y), \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Conditional Distributions ("cross sections")

$$p(y|x) = \frac{p(x,y)}{p_x(x)}, \quad f(y|x) = \frac{f(x,y)}{f_x(x)}$$

Indep Random Vars

$p(x,y) = p_x(x)\,p_y(y)$ for all x, y $\qquad f(x,y) = f(x)\,f(y)$ for all x, y

$p_x(y) = p(y|x)$ for all x, y $\qquad f_y(y) = f(y|x)$ for all x, y

---

Note that Bivariate distributions determine marginal dists, but not other way around unless X and Y are independent

but $p_x(x)$ and $p(y|x)$ do determine $p(x,y)$
$f_x(x)$ and $f(y|x)$ " " by multiplication

Moment Generating Functions

The moment generating function (mgf) $M(t)$ is $E(e^{tx})$.

discrete

$$M(t) = \sum_{\text{all } x} e^{tx} p(x)$$

continuous

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$$

The r^{th} moment of a random var is $E(X^r)$.

$$M^{(r)}(0) = E(X^r)$$

the r^{th} derivative.

Trick to find moments by differentiating instead of integrating

$$\text{mgf} \rightleftharpoons \text{pdf} \rightleftharpoons \text{cdf}$$

Any one of the above gives the other two.

Bayes's Theorem

E_i : "causes" "states of nature"

F : "effect" "data"

$$P(E_i|F) = \frac{P(E_i)\,P(F|E_i)}{P(F)}$$

$$P(F) = \sum_{j=1}^{n} P(E_j)\,P(F|E_j)$$

Equivalently:

$$\underbrace{P(E_i|F)}_{\text{posterior}} \propto \underbrace{P(E_i)}_{\text{prior}}\,\underbrace{P(F|E_i)}_{\text{likelihood}}$$

(or $P(E_i|F) = (\text{Const})\,P(E_i)\,P(F|E_i)$

$$\text{"Const"} = \frac{1}{P(F)} = \frac{1}{\sum_j P(E_j)\,P(F|E_j)}$$

Bayes's Theorem (continued)

To be clear, let's write in densities (or pmf's)

⟹ Given: $f_Y(y)$ and $f(x|y)$
⟹ Find: $f(y|x)$

These are the same!!

$$f(y|x) \propto f(x|y) f_Y(y)$$

$$f(y|x) = K f(x|y) f_Y(y) \quad (K \text{ may depend on } x)$$

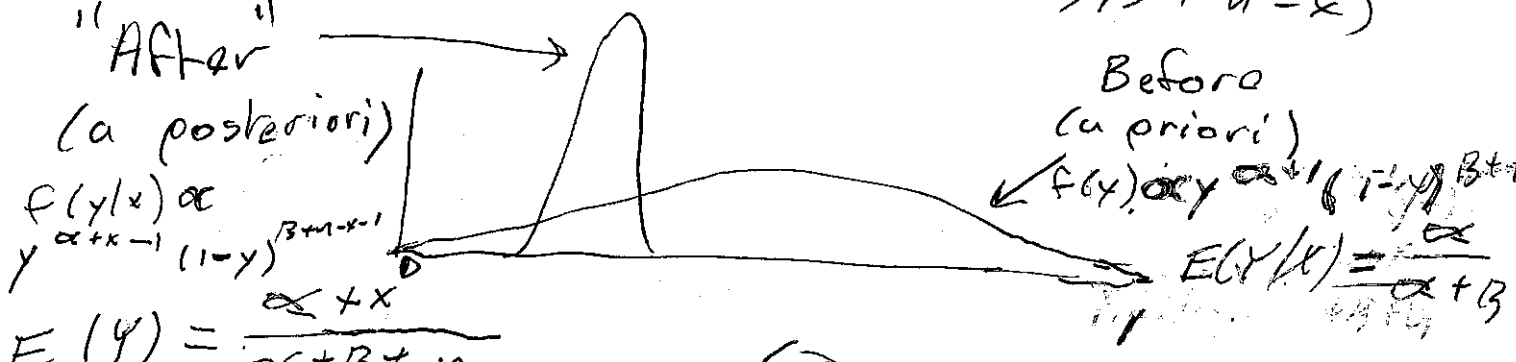
$$f(y|x) = \frac{f(x|y) f_Y(y)}{\int_{-\infty}^{\infty} f(x|u) f_Y(u) du}$$

Idea: Given X (data), make inferences about y

Ex $f(y)$ Beta (α, β)

$p(x|y)$ Binomial (n, y)

$f(y|x)$ Beta $(\alpha + x, \beta + n - x)$



Bayes for Binomial

θ = fraction, probability

$f(\theta)$ prior (Example: Beta (α, β)

$$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0<\theta<1$$

[$\alpha=\beta=1$ gives uniform]

$$\text{Expectation} = \frac{\alpha}{\alpha+\beta} = \mu_\theta$$

$$\text{Variance} = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}$$

$p(x/\theta)$ likelihood

$$p(x/\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad x=0,1,\ldots n$$

$f(\theta/x)$ posterior

$$f(\theta/x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

Beta $(x+\alpha, n-x+\beta)$

$$E(\theta/X=x) = \frac{\alpha+x}{\alpha+\beta+n}$$

$$Var(\theta/X=x) = \frac{E(\theta/x)(1-E(\theta/x))}{\alpha+\beta+n+1}$$

Bayes for Normal

$$f(y/x) \propto f_Y(y)\, f(x/y)$$

$$\propto e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \cdot e^{-\frac{1}{2}(x-y)^2}$$

$$\Rightarrow e^{-\frac{1}{2}\frac{(y-A)^2}{B}}$$

by "completing the square"

i.e.,

$$ax^2 + bx + c \Rightarrow a(x+h)^2 + k$$

This gives

$$A = \frac{x\sigma^2}{\sigma^2+1} + \frac{\mu \cdot 1}{\sigma^2+1}$$ ← weighted average

$$B = \frac{\sigma^2}{\sigma^2+1}$$

Intro to Maximum Likelihood

Today we'll go further in considering Statistical Inference.

Recall that we would like to know the "state of nature" θ. More exactly θ is a parameter that represents such a state.

We will learn about θ by considering X $[= (x_1, \ldots, x_n)]$, the data.

We need a model to describe the relation between θ and X.

Specifically, X, given θ, is a random variable with distribution $p(x|\theta)$ or $f(x|\theta)$.

Ex: θ = fraction of votes
X = # of 100 sampled
$p(x|\theta) = \binom{100}{x} \theta^x (1-\theta)^{100-x}$ $x = 1, 2, \ldots, 100$

Ex: θ = true weight
X = what scale says
$f(x|\theta)$ $N(\theta, 1)$

Ideal Goal: Find $f(\theta|x)$.

ie: After we have data ("given data"), we want to know the probability of various values of θ.

So far, we've used

Bayes's Theorem:

$$f(\theta|x) \propto \underbrace{f(\theta)}_{\text{prior}} f(x|\theta)$$

Gives what we want. But:

it requires $f(\theta)$.

$f(\theta)$ is controversial —
How to get it?
What does it mean?
Subjective bias - disagreements

OK. How about a more limited goal?

We won't use $f(\theta)$.

Instead, we will work only with $f(x|\theta)$. Then we'll

Estimate a Point, not a Distribution

We treat θ as fixed ('given') X as random. We want an estimate of $\hat{\theta} = f(x)$ that is likely to be close to θ.

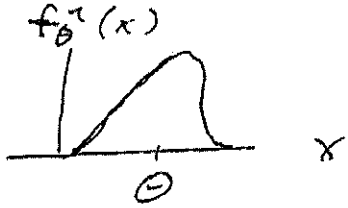
$\hat{\theta}$ depends on X

$\hat{\theta}$ is short for $\hat{\theta}(x)$ (or $\hat{\theta}(x_1, \ldots x_n)$)

X is a random variable, so $\hat{\theta}$ is a random variable

---

What does "$\hat{\theta}$ likely to be near θ" mean? From our "given θ" perspective, $\hat{\theta}$ has a distribution $f_{\hat{\theta}}(x)$ or $f_{\hat{\theta}}(x|\theta)$:



We want this distribution to be concentrated near and/or centered at θ.

Defs: $\hat{\theta}$ is <u>unbiased</u> if $E(\hat{\theta}) = \theta$, whatever θ is (i.e. $\int_{-\infty}^{\infty} x\, f_{\hat{\theta}}(x|\theta)\, dx = \theta$ for all θ)

<u>Bias</u> $= E(\hat{\theta}) - \theta$

<u>Mean Error</u> $= E(|\hat{\theta} - \theta|)$

<u>Mean Square Error</u> ("MSE") $= E[(\hat{\theta} - \theta)^2]$

It turns out the MSE has a particularly clear interpretation:

$$MSE_{\hat{\theta}}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$= E[(\hat{\theta} - E(\hat{\theta}) + (E(\hat{\theta}) - \theta)]^2$$

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\theta))(E(\theta) - \theta) + (E(\hat{\theta}) - \theta)^2\right]$$

$$= E[\hat{\theta} - E(\hat{\theta})]^2 + 2(E(\hat{\theta}) - \theta)\underline{E(\hat{\theta} - E(\hat{\theta}))} + (E(\hat{\theta}) - \theta)^2$$

$Var(\hat{\theta}|\theta)$ ← $E[\hat{\theta} - E(\hat{\theta})]^2$

$(B(\theta))^2$ ← $(E(\hat{\theta}) - \theta)^2$

$= E(\hat{\theta}) - E(\hat{\theta}) = 0$ ← $E(\hat{\theta} - E(\hat{\theta}))$

Hence

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias)^2$$

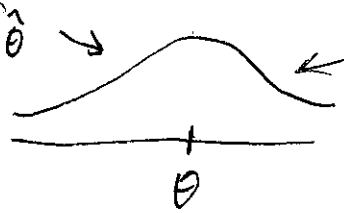
↑ "expected error"

↑ error from spread of data points

↑ error from bias

Case I

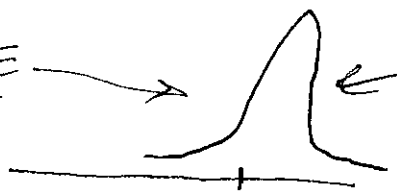
Tradeoff between bias and variance



no bias
large variance

SAME MSE

Case II



bias
small variance

Intro to Maximum Likelihood

STAT 24H
Lecture 9
10/27/16

Today we'll go further in considering Statistical Inference.

Recall that we would like to know the "state of nature" θ. More exactly θ is a parameter that represents such a state.

We will learn about θ by considering X [$= (x_1, \ldots, x_n)$], the data.

We need a model to describe the relation between θ and X.

Specifically, X, given θ, is a random variable with distribution $p(x|\theta)$ or $f(x|\theta)$.

Ex: θ = fraction of votes
X = # of 100 sampled

$$p(x|\theta) = \binom{100}{x} \theta^x (1-\theta)^{100-x} \qquad x = 1, 2, \ldots, 100$$

Ex: θ = true weight
X = what scale says

$$f(x|\theta) \quad N(\theta, 1)$$

Ideal Goal: Find $f(\theta|x)$.

ie: After we have data ("given data"), we want to know the probability of various values of θ.

So far, we've used

Bayes's Theorem:

$$f(\theta|x) \propto \underbrace{f(\theta)}_{\text{prior}} \underbrace{f(x|\theta)}_{\text{likelihood}}$$

Gives what we want. But:

it requires $f(\theta)$.

$f(\theta)$ is controversial —
- How to get it?
- What does it mean?
- Subjective bias - disagreements

OK. How about a more limited goal?

We won't use $f(\theta)$.

Instead, we will work only with $f(x|\theta)$. Then we'll

Estimate a Point, not a Distribution

We treat θ as fixed ("given") X as random. We want an estimate of $\hat{\theta} = f(x)$ that is likely to be close to θ.

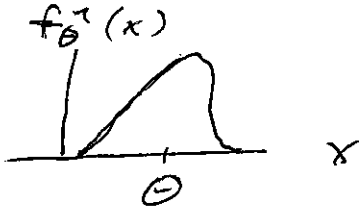
$\hat{\Theta}$ depends on X

$\hat{\theta}$ is short for $\hat{\Theta}(x)$ (or $\hat{\Theta}(x_1, \ldots x_n)$)

X is a random variable, so $\hat{\Theta}$ is a random variable

---

What does "$\hat{\Theta}$ likely to be near θ" mean? From our "given θ" perspective, $\hat{\Theta}$ has a distribution $f_{\hat{\Theta}}(x)$ or $f_{\hat{\Theta}}(x|\theta)$:



We want this distribution to be concentrated near and/or centered at θ.

Defs: $\hat{\Theta}$ is <u>unbiased</u> if $E(\hat{\Theta}) = \theta$, whatever θ is (i.e. $\int_{-\infty}^{\infty} x\, f_{\hat{\Theta}}(x|\theta)\, dx = \theta$ for all θ).

<u>Bias</u> $= E(\hat{\Theta}) - \theta$

<u>Mean Error</u> $= E(|\hat{\Theta} - \theta|)$

<u>Mean Square Error</u> ("MSE") $= E[(\hat{\Theta} - \theta)^2]$

It turns out the MSE has a particularly clear interpretation:

$$MSE_{\hat{\theta}}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$= E[(\hat{\theta} - E(\hat{\theta}) + (E(\hat{\theta}) - \theta)]^2$$

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\theta))(E(\theta) - \theta) + (E(\hat{\theta}) - \theta)^2\right]$$

$$= E[\hat{\theta} - E(\hat{\theta})]^2 + 2(E(\hat{\theta}) - \theta)\underline{E(\hat{\theta} - E(\hat{\theta}))} + (E(\hat{\theta}) - \theta)^2$$

$Var(\hat{\theta}|\theta)$ ← $E[\hat{\theta} - E(\hat{\theta})]^2$

$(B(\theta))^2$ ← $(E(\hat{\theta}) - \theta)^2$

$= E(\hat{\theta}) - E(\hat{\theta}) = 0$ ← $\underline{E(\hat{\theta} - E(\hat{\theta}))}$

Hence

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias)^2$$

↑ "expected error"

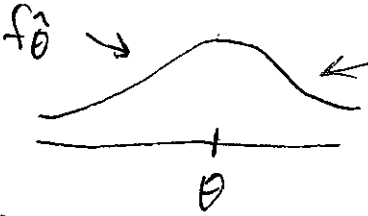
↑ error from spread of data points

↑ error from bias

Case I

$f_{\hat{\theta}}$

Tradeoff between bias and variance

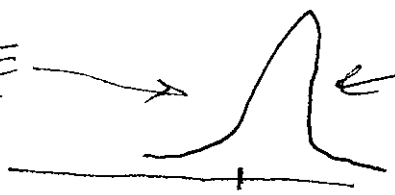


no bias
large variance

SAME MSE

Case II

$f_{\hat{\theta}}$



bias
small variance

Case I and Case II suggest that the optimal estimation scheme may vary depending on the exact details of the situation. To illustrate:

Example

Consider the polling example from the last two lectures, where the data $X \sim Bin(n, \theta)$. We want to estimate θ. Here are three possible estimators:

I. $\hat{\theta} = \frac{x}{n}$

$E(\hat{\theta}) = \frac{E(x)}{n} = \frac{(n\theta)}{n} = \theta$ unbiased

II $\theta^* = \frac{x+1}{n+2}$ where did this come from?

Well, if we're Bayesians and the prior were $\sim$ Uniform$(0,1)$. Then if $X = x$, the posterior

$\theta \sim Beta(x+1, n-x+1)$, and

$E(\theta | X = x) = \frac{x+1}{n+2}$

III. Consider a third estimator, the "stopped clock" estimator,

$\hat{\theta}^{**} = \frac{1}{2}$, It can be very accurate (or not!). It does <u>not</u> depend on data.

Note from the last lecture that we can write $\hat{\theta}^*$ as a weighted sum of the other estimators, so that

$$\hat{\theta}^* = \left(\frac{n}{n+2}\right)\hat{\theta} + \left(\frac{2}{n+2}\right)\hat{\theta}^{**}$$

Recall that $E(\hat{\theta}) = \theta$; it was unbiased.

$$E(\hat{\theta}^*) = \frac{E(x)+1}{n+2} = \frac{n\theta+1}{n+2}, \text{ Biased.}$$

$E(\hat{\theta}^{**}) = 0.5 \neq \theta$, Biased (Doh!)

What about the mean error and mean squared error?

I. $\hat{\theta}$: mean error (see Stigler for details):

$$E\left[\left|\hat{\theta}-\theta\right|\right] = 2\binom{n-1}{\lfloor n\theta \rfloor}\theta^{\lfloor n\theta\rfloor+1}(1-\theta)^{n-\lfloor n\theta\rfloor}$$

($\lfloor n\theta \rfloor$ the largest integer smaller than $n\theta$)

mean squared error:

$$MSE_{\theta}(\hat{\theta}) = E(\hat{\theta}-\theta)^2$$

$$= E\left(\frac{X}{n}-\theta\right)^2$$

$$= Var\left(\frac{X}{n}\right)$$

$$= \frac{Var(X)}{n^2}$$ ← binomial dist

$$= \frac{n\theta(1-\theta)}{n^2}$$

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

II $\hat{\theta}^*$: $$E\left[\left|\theta^*-\theta\right|\right] = \sum_{k=0}^{n}\left|\left(\frac{k+1}{n+2}\right)-\theta\right| b(k;n,\theta)$$

(evaluate numerically!)

$\hat{\theta}^*$, continued:

$$MSE_{\hat{\theta}^*} = Var(\hat{\theta}^*|\theta) + (Bias_{\hat{\theta}^*}(\theta))^2$$

$$= \frac{Var(x)}{(n+2)^2} + \frac{(1-2\theta)^2}{(n+2)^2}$$

$$= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}$$

[note: $Var(\hat{\theta}^*) < Var(\hat{\theta}) = \frac{n\theta(1-\theta)}{n^2}$

$\hookrightarrow = \frac{n\theta(1-\theta)}{(n+2)^2}$]

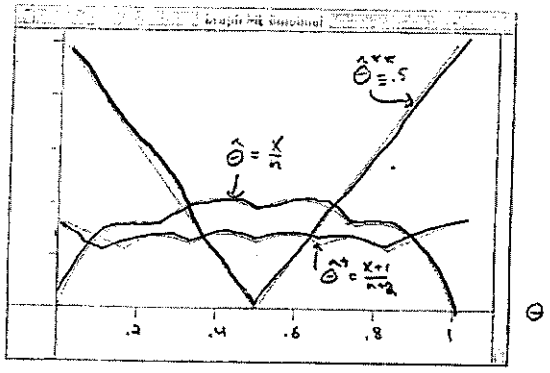
III. $\hat{\theta}^{**}$:

mean error: $E(|\hat{\theta}^{**} - \theta|) = \left|\frac{1}{2} - \theta\right|$

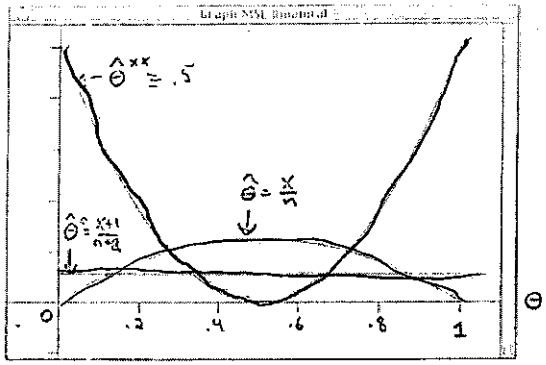
mean squared error: $MSE_{\hat{\theta}^{**}}(\theta) = \left(\frac{1}{2} - \theta\right)^2 = \frac{1}{4} - \theta(1-\theta)$

interval where $\hat{\theta}^{**}$ is better than $\hat{\theta}$

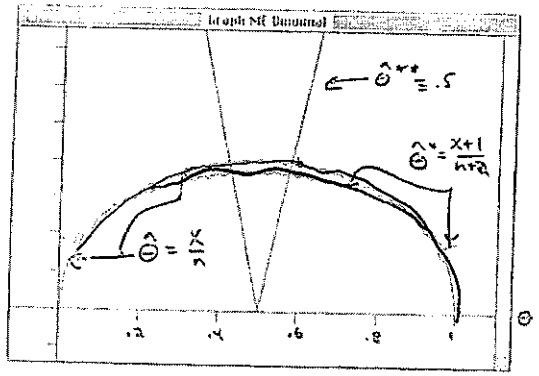
$n =$	Mean Error	Mean Squared Error
1	$.29 < \theta < .71$	$.15 < \theta < .85$
4	$.40 < \theta < .60$	$.28 < \theta < .72$
25	$.46 < \theta < .54$	$.40 < \theta < .60$
100	$.48 < \theta < .52$	$.45 < \theta < .55$



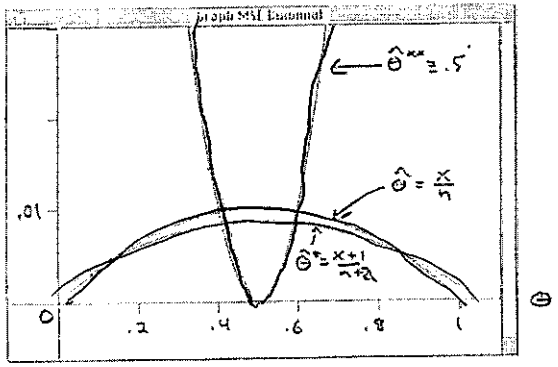
$E|\hat{\theta} - \theta|$ $n = 4$



$E(\hat{\theta} - \theta)^2$ $n = 4$



$E|\hat{\theta} - \theta|$ $n = 25$



$E(\hat{\theta} - \theta)^2$ $n = 25$

Figure 5.3

Where to get estimators?

One source is Maximum Likelihood

Recall Bayes: $f(\theta/x) \propto f(\theta) f(x/\theta)$

Might want the "most likely" θ:

θ that max's $f(\theta/x)$ (ie, max's $f(\theta)f(x/\theta)$)

But $f(\theta)$ is <u>not available</u>.

If $f(\theta)$ is flat (ie, approximately constant ≡ not much prior info) could maximize $f(x/\theta)$ instead, ie find θ to make $f(x/\theta)$ as large as possible for the given data x — call <u>that</u> $\hat{\theta}$.

<u>Definition</u>. We call $L(\theta) = f(x/\theta)$ (or $L(\theta) = f(x_1, \ldots, x_n/\theta)$), viewed as a function of θ, the <u>Likelihood function</u>. The value of θ, say $\hat{\theta}$, for which $L(\theta)$ acheives its max is called the <u>maximum likelihood estimate</u> of θ.

Interpreting $L(\theta)$

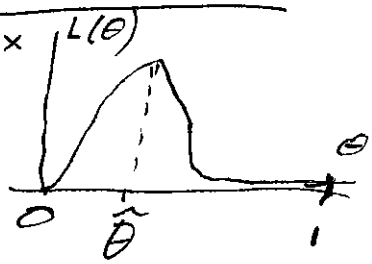
Bayesians: $L(\theta) = f(x|\theta)$ is proportional to the posterior distribution of θ given x

if $f(\theta) =$ constant.

others: $L(\theta)$ is the probability (density) that we observe the data we actually observed if the true state of nature is θ.

Hence the MLE $\hat{\theta}$ is the "state of nature" that best explains our data, for which it is most likely.

Example: $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$



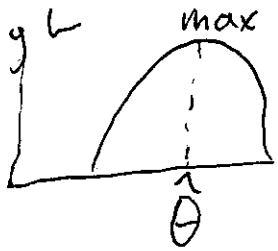
Maxing $L(\theta)$ same as

max'ing $\log(L(\theta)) = \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)$

$$\frac{d}{d\theta}\log L(\theta) = 0 + \frac{x}{\theta} - \frac{(n-x)}{1-\theta};\ \text{set} = 0 \rightarrow \boxed{\hat{\theta} = \frac{x}{n}}$$

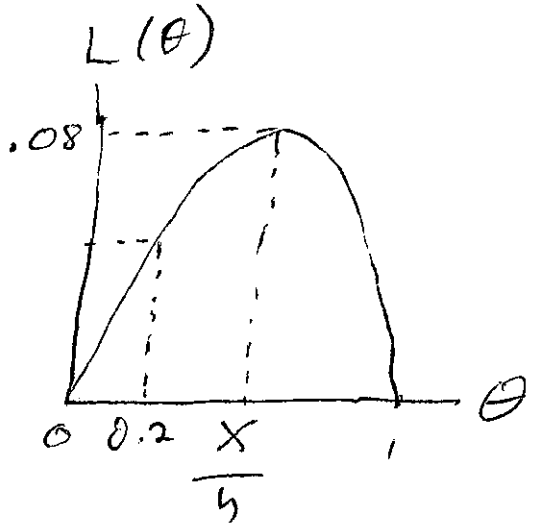
min or max ?

$$\frac{d^2}{d\theta^2}\log L(\theta) = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2} < 0$$



$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

is maximum for $\hat{\theta} = \frac{x}{n}$



$$\frac{d}{d\theta}L(\theta) = 0$$

Solve, then:

$$\frac{d^2}{d\theta^2}L(\hat{\theta}) < 0$$

check

$L(\theta)$ = Pr (observed data | state of nature is θ)

is largest for $\theta = \hat{\theta} = \frac{x}{n}$.

$\hat{\theta} = \frac{x}{n}$ "best" explains data.

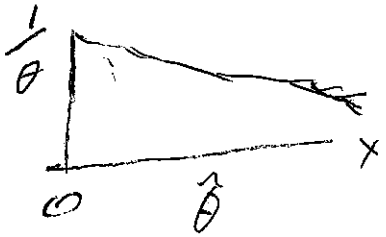
We are more likely to get 40 of 100 for Bernie if the true fraction were 0.4 than if it were any other value (.2, .5, etc). Note that $L(\hat{\theta})$ need not be large - $L(0.4) = 0.0812$ here.

Note: the MLE does not necessarily have good "sampling" properties — it does not necessarily have smallest MSE.

But: It can be proved to be often nearly best in a certain sense with large samples.

Example: Estimating Average Failure Time

Suppose a component has a constant probability of failure, so it lasts a time X with dist.



$$f(x|\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \theta > 0,\ x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E(x) = \int_0^\infty x \cdot \frac{1}{\theta} e^{-x/\theta} dx = \theta$$

Problem: The mean failure time θ is unknown. n components are tested independently, to observe $X_1, X_2, \ldots, X_n$. Estimate θ.

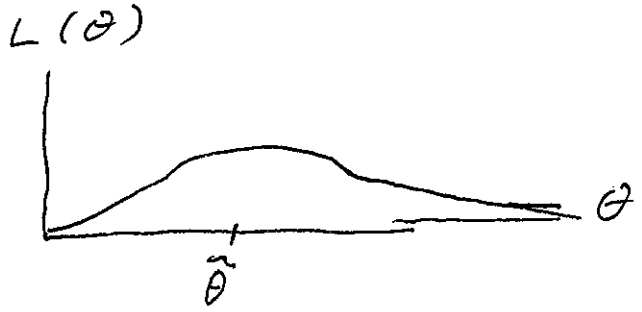
The joint density $f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) f(x_2|\theta) \ldots f(x_n|\theta)$ (by independence)

$$= \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdots$$

$$= \frac{1}{\theta^n} \cdot e^{-\frac{(x_1 + \cdots + x_n)}{\theta}}$$

iff all x's > 0 (otherwise zero)

So $L(\theta) = \frac{1}{\theta^n} e^{-\Sigma x_i/\theta}$, $\theta > 0$



You can think of $L(\theta)$ as giving "relative liklihood" of data for different values of θ

want to max $L(\theta)$. So:

$$\max \log L(\theta) = -n \log \theta - \sum X_i / \theta$$

$$\frac{d}{d\theta} \log L(\theta) = -\frac{n}{\theta} + \frac{\sum X_i}{\theta^2}$$

Set $= 0$:

$$\frac{-n}{\hat{\theta}} + \frac{\sum X_i}{\hat{\theta}^2} = 0$$

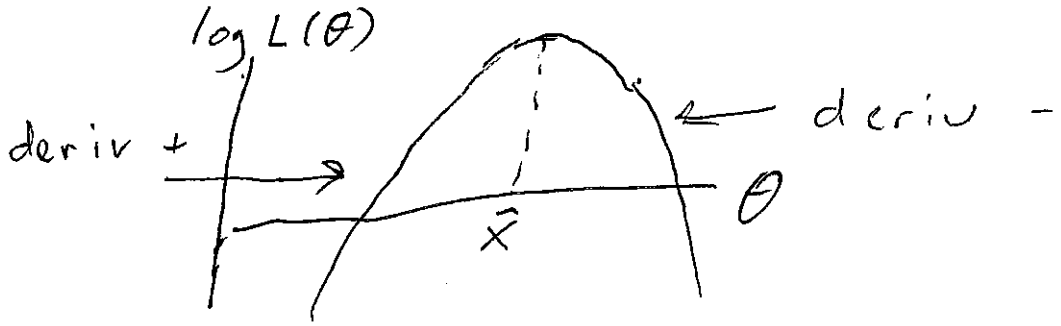
$$\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Check:

$$\frac{d}{d\theta} \log L(\theta) = \frac{n}{\theta} \left(\frac{\bar{x}}{\theta} - 1 \right)$$

For $\theta < \bar{x}$ this is > 0

For $\theta > \bar{x}$ this is < 0



Summing up: $f(x_i|\theta) = \frac{1}{\theta} e^{\frac{-x_i}{\theta}}$ $x_i > 0$

Data: $X_1, \ldots, X_n$ indep

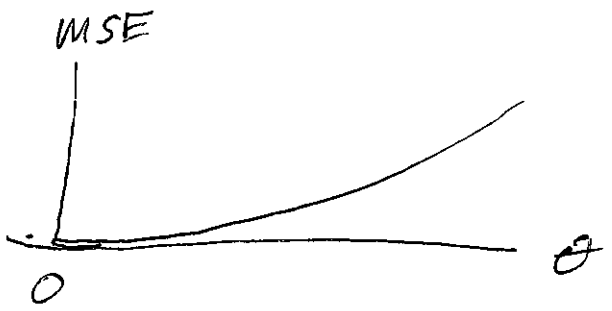
MLE $\hat{\theta} = \bar{X}$

$E(\hat{\theta}) = E(\bar{x}) = E(x_i) = \theta$

unbiased

So $MSE = Var(\hat{\theta}) = Var(\bar{x}) = \frac{Var(x_i)}{n}$

$= \boxed{\frac{\theta^2}{n}}$



So, the MSE decreases as n increases; it increases as θ increases

What about the exponential—

Say $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$

Data: $X_1, \ldots, X_n$ indep.

i.e, same model $\theta = \frac{1}{\lambda}$, $\lambda = \frac{1}{\theta}$ simply a different parametrization.

Invariance of MLE

Likelihood function for λ is $L(\frac{1}{\lambda})$, max for $\hat{\lambda} = \frac{1}{\bar{x}}$. In general, the MLE of $h(\theta)$ is $h(\hat{\theta})$.

BUT $\hat{\lambda}$ is not unbiased, because $E(\frac{1}{\bar{x}}) \neq \frac{1}{E(\bar{x})}$.

Issues:

(1) Finding MLE

→ $\frac{d}{d\theta} L(\theta) = 0$ solve

→ $\frac{d}{d\theta} \log L(\theta) = 0$ solve

→ numerical methods

→ algebraic ingenuity

(Next time:)

(2) Distribution of MLE

→ find exactly

→ Central Limit Theorem

→ Fisher's App.

(3) Properties of MLE

→ unbiased? Not usually

→ Approximate var MSE (Fisher)

→ consider exact distribution

Maximum Likelihood II

STAT 24400
Lecture 10
11/1/16

Issues:

(1) Finding MLE
- → $\frac{d}{d\theta} L(\theta) = 0$ solve
- → $\frac{d}{d\theta} \log L(\theta) = 0$ solve
- → numerical methods
- → algebraic ingenuity

(Next time:)

(2) Distribution of MLE
- → find exactly
- → Central Limit Theorem
- → Fisher's App

(3) Properties of MLE
- → unbiased? Not usually
- → Approximate Var MSE (Fisher)
- → consider exact distribution

(13)

We were here

θ "state of Nature" parameter

X (or $X_1, X_2, \ldots X_n$) data

$f(x|\theta)$ (or $f(x_1, \ldots x_n|\theta)$ model

We estimate θ by the estimator $\hat{\theta}$, a random var.

But Before we go on to the distribution of $\hat{\theta}$, need a few new results.

The Distribution of Sums

We've discussed $E(\hat{\theta})$ and $Var(\hat{\theta})$, but for detailed assessments of $\hat{\theta}$'s accuracy, we need to know its distribution.

But! $\hat{\theta} = \hat{\theta}(x) = \hat{\theta}(x_1, x_2, \ldots x_n)$ is a transformation of the data, its distribution can be VERY complicated.

Some cases are easy, though.

I. Binomial Estimators

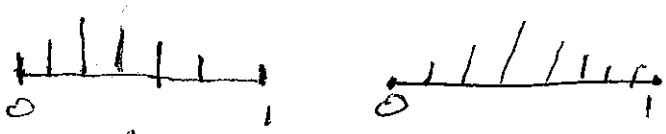
For the estimator $\hat{\theta}(x) = \frac{x}{n}$

of the parameter θ in a binomial distribution, $h(x) = \frac{x}{n}$, $h^{-1}(y) = g(y) = ny$

$$p_{\hat{\theta}}(y) = Pr(\hat{\theta} = y)$$
$$= p_x(ny)$$
$$= Bin(ny; n, \theta)$$

$\hat{\theta}^*(x) = \frac{(x+1)}{(n+2)}$, so $g(y) = (n+2)y - 1$

$$p_{\hat{\theta}^*} = Bin((n+2)y - 1; n, \theta)$$



$\hat{\theta} = \frac{x}{n}$ $\quad$ $\hat{\theta}^* = \frac{x+1}{n+2}$

← The distributions of $\hat{\theta}$ and $\hat{\theta}^*$, for $n = 6$ $\theta = 0.4$

II. $Z = X + Y$

We've already seen $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and in general sums of random variables are quite frequent. If (X, Y) are a bivariate random variable with density $f(x, y)$, then the density of

$$Z = X + Y$$

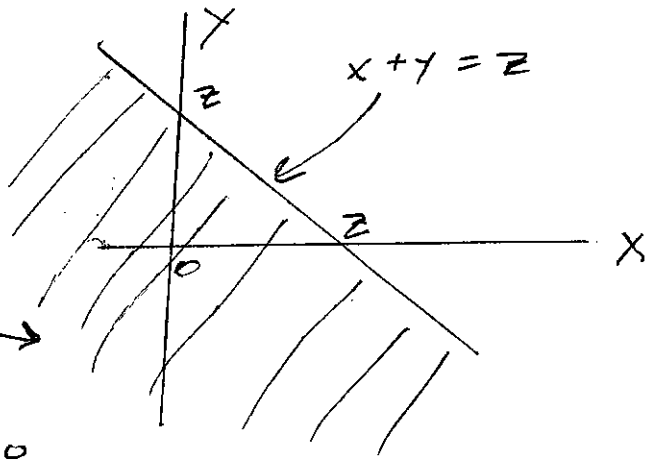
is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f(z-y, y)\,dy .$$

Pf

$$F_Z(z) = Pr(Z \le z)$$

$$= Pr(X + Y \le z)$$



the shaded region.

We integrate over that region, so

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x, y)\,dx\,dy$$

but

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} \frac{d}{dz} \left(\int_{-\infty}^{z-y} f(x, y)\,dx \right) dy$$

$$= \int_{-\infty}^{\infty} f(z-y, y)\,dy \quad \therefore$$

III $Z = X + Y$, X and Y normal and independent.

The "reproductive property" of normal distributions.

Say $X \sim N(\mu, \sigma^2)$, $Y \sim N(\theta, \tau^2)$.

Then
$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(z-y-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}\tau} e^{\frac{-1}{2\tau^2}(y-\theta)^2} dy$$

$$= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[(z-y-\mu)^2/\sigma^2 + (y-\theta^2)/\tau^2\right]} dy$$

The part of the exponent in brackets can be written
$$A(z-B)^2 + C(y-Dz)^2 + E$$
(trust me... complete the squares)

but now:
$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma\tau\sqrt{C}} e^{-\frac{E}{2}} \cdot e^{-\frac{A}{2}(z-B)^2} \int_{-\infty}^{\infty} \sqrt{\frac{C}{2\pi}} e^{-\frac{C}{2}(y-Dz)^2} dy$$

integral of $N(Dz, \frac{1}{C}) = 1$!

so now
$$f_Z(z) \propto e^{-\frac{A}{2}(z-B)^2}, \text{ a } N(B, \tfrac{1}{A}) \text{ density}$$

we now have

$$Z \sim N(B, 1/A)$$

$B = E(Z)$ and $1/A = Var(Z)$

but

$$E(Z) = E(X) + E(Y) = \mu + \theta$$

and since X and Y are indep,

$$Var(Z) = Var(X) + Var(Y)$$
$$= \sigma^2 + \varphi^2$$

hence $Z \sim N(\mu + \theta, \sigma^2 + \varphi^2)$.

Note: it hence follows that if $X_1, X_2, \ldots X_n$, each distributed

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

if they are all distributed $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

[Note: no limits involved here, cf CLT]

4. The Chi-Square Distribution

Way back in Lecture 3, pp. 13–14, we found the Chi-square (χ^2) distribution with one degree of freedom, namely the dist of U^2, where $U \sim N(0,1)$. It had density

$$f_{U^2}(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad \text{for } y > 0.$$

The chi-square dist for n degrees of freedom is: $\chi^2(n) = U_1^2 + U_2^2 + \cdots + U_n^2$,

U's indep and $\sim N(0,1)$.

Then

$$f_{\chi^2(n)}(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad \text{for } x > 0.$$

(0 otherwise)

Why? Well, for $n=1$, $\Gamma(\frac{n}{2}) = \sqrt{\pi}$, so for the $n=1$ case:

$$\frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} \quad \text{(same as lec 3)}$$

Let's continue by induction.

want to prove

$$f_{\chi^2(n)}(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2} \qquad x>0$$

We have the case $n=1$ taken care of. Now let's assume the above formula holds for $n = \kappa - 1$. Let

$$X = U_1^2 + \cdots + U_{\kappa-1}^2$$
$$Y = U_\kappa^2$$

$U_1, \ldots U_\kappa$ indep, $\sim N(0,1)$.
So X and Y are independent, and $\chi^2(\kappa) = X + Y$ by definition.
By hypothesis,

$$f_X(x) = \frac{1}{2^{\frac{\kappa-1}{2}}\Gamma(\frac{\kappa-1}{2})} x^{\frac{\kappa-1}{2}-1} e^{-\frac{x}{2}} \qquad x>0$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \qquad y>0$$

(both 0 otherwise.)

So,

$$f_{\chi^2(\kappa)}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy,$$

and since $f_X(z-y) = 0$ iff $y \geq z$,

$$= \int_0^z \frac{1}{2^{\frac{\kappa-1}{2}}\Gamma(\frac{\kappa-1}{2})} \cdot (z-y)^{\frac{\kappa-1}{2}-1} e^{-\frac{(z-y)}{2}} \cdot \frac{1}{\sqrt{2\pi y}} e^{-y/2}\,dy$$

Now, extract all terms not containing y...

want to prove

$$f_{\chi^2(n)}(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$ (continued..)

$$f_{\chi^2(k)}(z) = \frac{1}{2^{\frac{k-1}{2}}\Gamma(\frac{k-1}{2})\sqrt{2\pi}} e^{-\frac{z}{2}} \int_0^z (z-y)^{\frac{k-3}{2}} y^{-1/2} dy$$

The above engenders hope, but how to do the integral?

The spirit of how to procede is to note that we are proving something about $f(z)$, but the integral is with respect to y. How can we extract all z terms to the left of the integral sign?

Answer: if $y = zu$, $(z-y) \to (z - zu) = z(1-u)$

in detail: $dy = z\,du$, $(z-y)^{\frac{k-3}{2}} = z^{\frac{k-3}{2}}(1-u)^{\frac{k-3}{2}}$

$y^{-1/2} = z^{-\frac{1}{2}} u^{-\frac{1}{2}}$, so

$$\int_0^z (z-y)^{\frac{k-3}{2}} y^{-\frac{1}{2}} dy = z^{\frac{k-3}{2}} \cdot z^{-\frac{1}{2}} \cdot z \underbrace{\int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du}_{\text{const!}}$$

Hence

$$f_{\chi^2(k)}(z) = C\, z^{\frac{k}{2}-1} e^{-\frac{z}{2}} \quad \text{for } z > 0$$

... the desired functional form.

What about C? We have, at this point,

$$f_{\chi^2(k)} = \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)\sqrt{2\pi}} z^{\frac{k}{2}-1} e^{-\frac{z}{2}} \underbrace{\int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du}_{\text{a Beta function!}}$$

$$\hookrightarrow B\left(\tfrac{1}{2}, \tfrac{k-1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{k-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

$$= \frac{\sqrt{\pi}\,\Gamma\left(\frac{k-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

$$f_{\chi^2(k)} = \frac{\Gamma\left(\frac{k-1}{2}\right)\sqrt{\pi}}{2^{\frac{k-1}{2}}\cdot\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k-1}{2}\right)\sqrt{2}\sqrt{\pi}} z^{\frac{k}{2}-1} e^{-\frac{z}{2}}$$

$$= \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} z^{\frac{k}{2}-1} e^{-\frac{z}{2}}$$

$\therefore$

Because the estimator for σ^2 is $\hat{S}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ [For $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$],

$$(n-1)\hat{S}^2/\sigma^2 \sim \chi^2(n-1)$$

and hence χ^2 is the distribution of a multiple of the sample variance of a normally distributed sample.

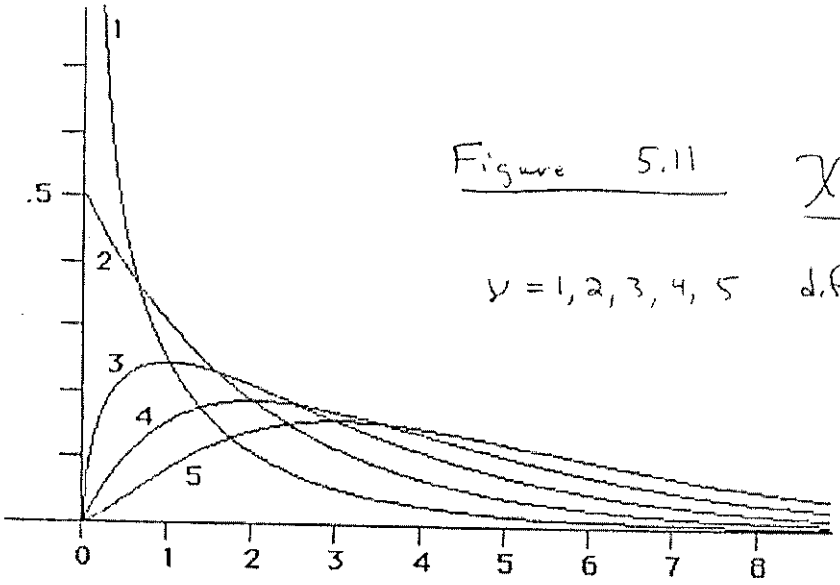
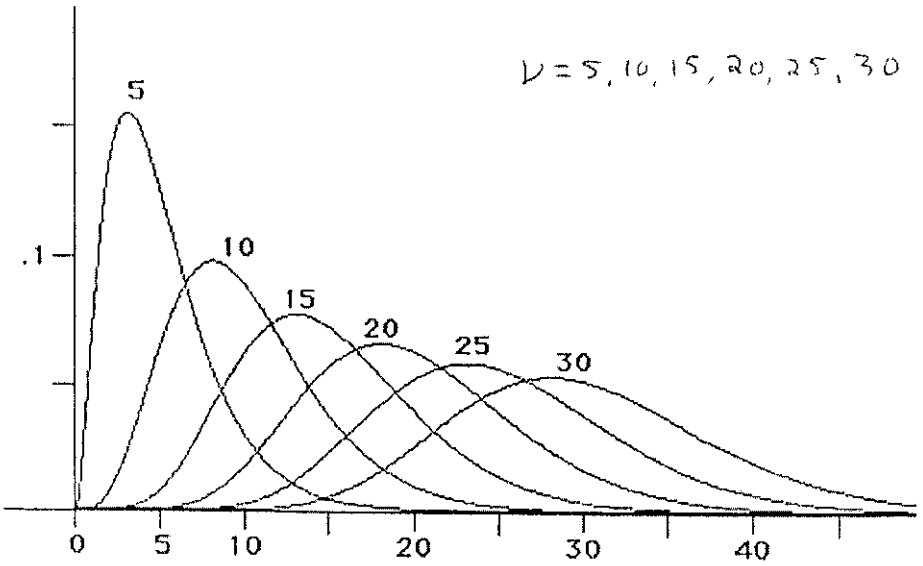


Figure 5.11 χ^2 densities

$\nu = 1, 2, 3, 4, 5$ d.f.



$\nu = 5, 10, 15, 20, 25, 30$ df.

Some MLEs are sums or averages.

Recall the Central Limit Theorem
(derived using the m.g.f. in Lecture 5)

$X_1, \ldots, X_n$ iid $E(X_i) = \mu$ $Var(X_i) = \sigma^2$

Then: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\bar{X} \overset{\text{approx}^*}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sum_{i=1}^{n} \overset{\text{approx}^*}{\sim} N(n\mu, n\sigma^2)$$

* approx gets better as n gets bigger

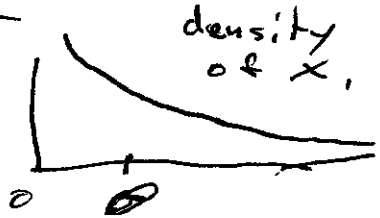
This helps explain the origin of the normal "error curve" in nature when looking at aggregates (that is, sums).

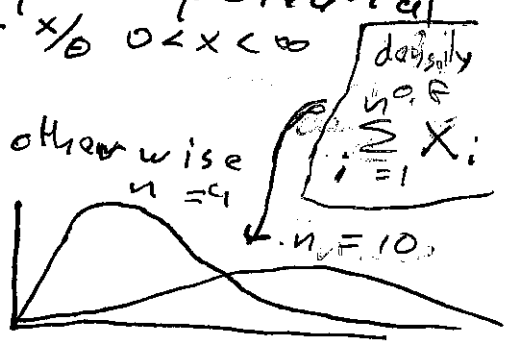
Ex: Chi-Sq, large d.f.

Ex: $X_1, X_2, \ldots, X_n$ indep exponential

Recall from last time: $\hat{\theta}(X_1, \ldots X_n) = \bar{X}$!

$$f(x_i|\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$





Now — Back to the MLE!

Recall where we are

θ "state of nature", a parameter

X (or $X_1, X_2, \ldots, X_n$) Data

$f(x|\theta)$ (or $f(x_1, \ldots, x_n|\theta)$) Model

An Estimator = a tool for estimating

$\hat{\theta} = \hat{\theta}(x)$ or $\hat{\theta}(x_1, \ldots, x_n)$

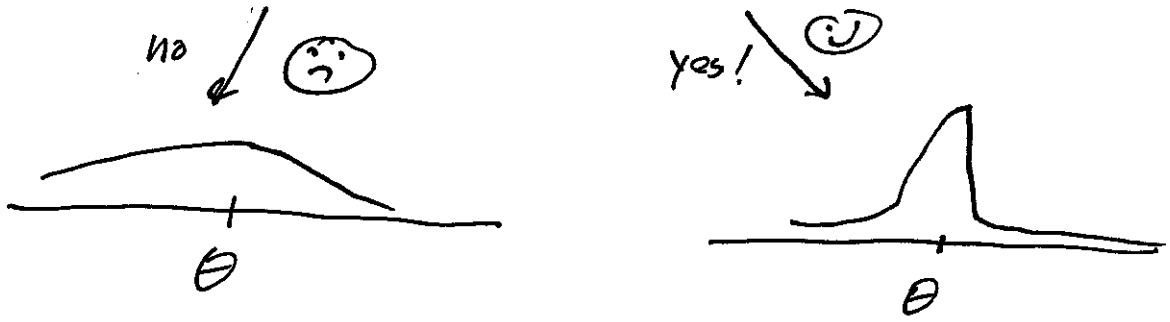
Goal: $\hat{\theta}$ near θ.

A "good" estimator tends to be near θ. Evaluate from "before data" pespective.

X random, dist $f(x|\theta)$

$\hat{\theta} = \hat{\theta}(x)$ random $f_{\hat{\theta}}(x|\theta)$

Want $f_{\hat{\theta}}(x|\theta)$ concentrated near θ



One measure:

$$MSE_\theta(\theta) = E(\hat{\theta} - \theta)^2$$

$$= Var(\hat{\theta}) + (Bias_{\hat{\theta}}(\theta))^2$$

Example: "Serial Number" Problem
(estimating tank production, WW II)

Observe serial number $= X$

Largest possible $= \theta$

$P(x=1) = P(x=2) = \ldots = P(x=\theta) = \frac{1}{\theta}$

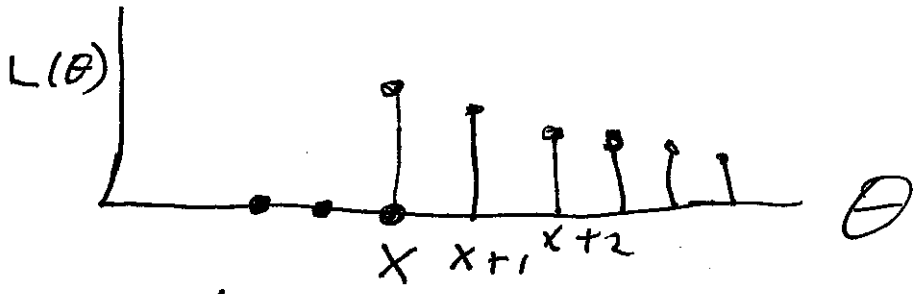
Model: $p(x|\theta) = \frac{1}{\theta}$, for $x = 1, 2, \ldots, \theta$

$= 0$ otherwise

Model: $p(x|\theta) = \frac{1}{\theta}$, for $x = 1, 2, \ldots, \theta$
0 otherwise

Maximum Likelihood:

$$L(\theta) = \frac{1}{\theta} \quad \text{for } \theta \geq x$$



MLE: $\hat{\theta}_1 = x$ (if $x = 15$, estimate 15)

Biased $E(\hat{\theta}_1) = E(x) = \frac{1+\theta}{2}$

$$\text{Bias} = \frac{1+\theta}{2} - \theta = \frac{1-\theta}{2}.$$

Unbiased Estimate: $\hat{\theta}_2 = 2x - 1$ $E(\hat{\theta}_2) = 2E(x) - 1$

$$= 2\left(\frac{1+\theta}{2}\right) - 1$$

$$= \theta$$

How do they compare in concentration?

Have Bias, need Var for MSE

Var (x):

$$E(X^2) = \frac{1}{\theta}\sum_{x=1}^{\theta} x^2 = \frac{1}{\theta}\cdot\frac{\theta(\theta+1)(2\theta+1)}{6}$$

$$= \frac{(\theta+1)(2\theta+1)}{6}$$

$$Var(x) = E(x^2) - (E(x))^2$$

$$= \frac{(\theta+1)(2\theta+1)}{6} - \left(\frac{\theta+1}{2}\right)^2$$

For MLE

$$mse\,\hat{\theta}_1 = \left[\frac{(\theta+1)(2\theta+1)}{6} - \left(\frac{\theta+1}{2}\right)^2\right] + \left[\frac{(1-\theta)}{2}\right]^2$$

$$= \frac{2\theta^2 - 3\theta + 1}{6}$$

For Unbiased:

$$mse\,\hat{\theta}_2 = 4\,Var(x) + (Bias)^2$$

$$= 4[\quad] + \theta^2$$

$$= \frac{\theta^2 - 1}{3}$$

θ	1	2	3	....	10	.....	100
$mse\,\hat{\theta}_1$	0	$\frac{1}{2}$	$\frac{5}{3}$	...	28.5		3283.5
$mse\,\hat{\theta}_2$	0	1	$\frac{8}{3}$	...	33		3333

MLE III

STAT 24400
Lecture 11
November 3, 2016

Recall where we are:

Maximum Likelihood

θ parameter (to be found)

X or $X_1, \ldots, X_n$ data

$L(\theta) = f(x|\theta)$ or $f(x_1, \ldots, x_n|\theta)$ Likelihood function (the model considered as function of θ)

MLE: $\hat{\theta} = \hat{\theta}(X) = \hat{\theta}(X_1, \ldots, X_n)$ max's $L(\theta)$ depends on data, estimates θ.

Evaluation of estimators:

Judge by performance — how concentrated is their dish around θ?

Measure by: $MSE(\hat{\theta}) = E(\hat{\theta}(x) - \theta)^2$
$= Var(\hat{\theta}(x)) + (Bias)^2$

$Bias(\hat{\theta}(x)) = E(\hat{\theta}(x)) - \theta$

So what can we say about the distribution of the random var $\hat{\theta}$?

Fisher's Approximation Theorem

If the MLE can be found from solving $\frac{d}{d\theta} L(\theta) = 0$ (or $\frac{d}{d\theta} \log L(\theta) = 0$)*†, Then $\vec{\theta}$ has an approximately* $N(\theta, \tau_n^2)$ distribution.

(* when n, the number of data points, is large).

(† How could the MLE not be found from $\frac{d}{d\theta} L(\theta) = 0$? In general, because of differentiability problems, including the max not being in the interior of the domain.

This implies that $\vec{\theta}$ is approximately unbiased

" " τ_n^2

$MSE(\vec{\theta})$ "

How to find τ_n^2?

Consider first the iid case.

For X_i iid,

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

$$\gamma_n^2 = \frac{\gamma^2}{n}, \quad \text{where}$$

$$\frac{1}{\gamma^2} = E\left[\frac{\partial}{\partial\theta} \log f(x_1|\theta)\right]^2$$

or

$$\frac{1}{\gamma^2} = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x_1|\theta)\right]$$

These are the same. In what follows, read "x" as "$x_1, \ldots, x_n$", the general case.

$$\int f(x|\theta) = 1, \text{ so}$$

$$\frac{\partial}{\partial\theta} \int f(x|\theta)\, dx = 0,$$

$$\frac{\partial}{\partial\theta} f(x|\theta) = \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)$$

so

$$0 = \frac{\partial}{\partial\theta} \int f(x|\theta)\, dx = \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)\, dx$$

daring assumption, justified with rigorous smoothness conditions not considered here.

Take a second derivative, and get

$$0 = \frac{\partial}{\partial\theta}\int\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx$$

$$= \int\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] f(x|\theta)\,dx + \int\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)\,dx$$

$$\left[\; 0 = E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] + E\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 \;\right]$$

Remark:

$$\frac{1}{\gamma_n^2} = E\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$$

is sometimes written

$$\frac{1}{\gamma_n^2} = I$$

The textbook (Rice) uses that notation and never says why.

Here is the reason.

Imagine we are considering unbiased estimators $\hat{\theta}$ so that $MSE(\hat{\theta}) = Var(\hat{\theta}) + \cancel{Bias}^{\to 0}$

and $MSE(\tilde{\theta}) = Var(\tilde{\theta}) + \cancel{Bias}^{\to 0}$

Both estimators are conditioned against the same data. We are likely to prefer $\hat{\theta}$ if it has a smaller variance than $\tilde{\theta}$, so

$$eff(\hat{\theta}, \tilde{\theta}) = \frac{Var(\tilde{\theta})}{Var(\hat{\theta})} > 1$$

and if $Var(\tilde{\theta}) = \frac{c_1}{n}$ and $Var(\hat{\theta}) = \frac{c_2}{n}$, we could use a smaller sample with $\hat{\theta}$.

There is an upper limit to efficiency: [Cramer – Rao Inequality]

Let $X_1, \ldots X_n$ be iid with density $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then

$$Var(T) \geq \frac{1}{n I(\theta)}$$

$I(\theta)$ is sometimes called "Fisher Information"

It is the most information you can squeeze out of a set of data. (Implicitly, it also reveals certain desirable math properties of the MLE).

Now, back to finding $\frac{1}{\gamma_n^2} = I$

Fisher's Approx. Thm:

If the MLE can be found from solving $\frac{d}{d\theta} \log L(\theta) = 0$, then

$\hat{\theta} \sim \mathcal{N}(\theta, \gamma_n^2)$.

1] Indep. Case $\left(L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)\right)$

so $\gamma_n^2 = \frac{\gamma^2}{n}$, where

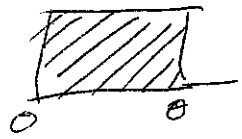
$$\frac{1}{\gamma^2} = E\left(\frac{d}{d\theta} \log f(x|\theta)\right)^2 = -E\left(\frac{d^2}{d\theta^2} \log f(x|\theta)\right)$$

2] More General Case

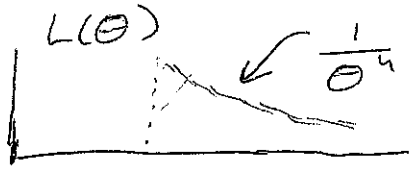
$$\frac{1}{\gamma_n^2} = E\left[\frac{d}{d\theta} \log f(x_1, \ldots x_n|\theta)\right]^2 = -E\left[\frac{d^2}{d\theta^2} \log f(x_1, \ldots, x_n|\theta)\right]$$

[if indep, $E\left[\frac{d}{d\theta} \log f(x_1, \ldots x_n|\theta)\right]^2 = nE\left[\frac{d}{d\theta} \log f(x_1|\theta)\right]^2$]

Ex: Estimate the right hand boundary of uniform dist over $0 \le x < \theta$

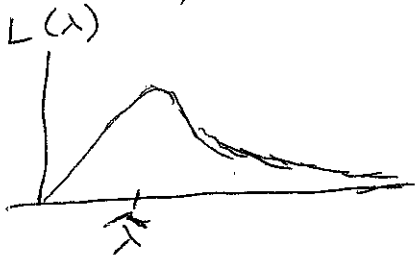


$$f(x_i|\theta) = \begin{cases} \frac{1}{\theta} & 0 < x_i < \theta \\ 0 & \text{otherwise} \end{cases}$$



$\hat{\theta} = \max(x_i)$, but $\frac{d}{d\theta} L(\theta) \neq 0$ ever!!

Ex: Failure time; parametrize by λ (not $\bar{x}$!)



$$f(x_i|\lambda) = \begin{cases} \lambda e^{-\lambda x_i} & x_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\hat{\lambda} = \frac{1}{\bar{x}}$, NOT unbiased. (but found from $\frac{d}{d\lambda} \log L(\lambda) = 0$)

Finding γ^2

$$\frac{d}{d\lambda} \log f(x_i|\lambda) = \frac{d}{d\lambda} (\log \lambda - \lambda x_i) = \frac{1}{\lambda} - x_i$$

$$\frac{1}{\gamma^2} = E\left(\frac{1}{\lambda} - x_i\right)^2 = \mathrm{Var}(x_i) = \frac{1}{\lambda^2}$$

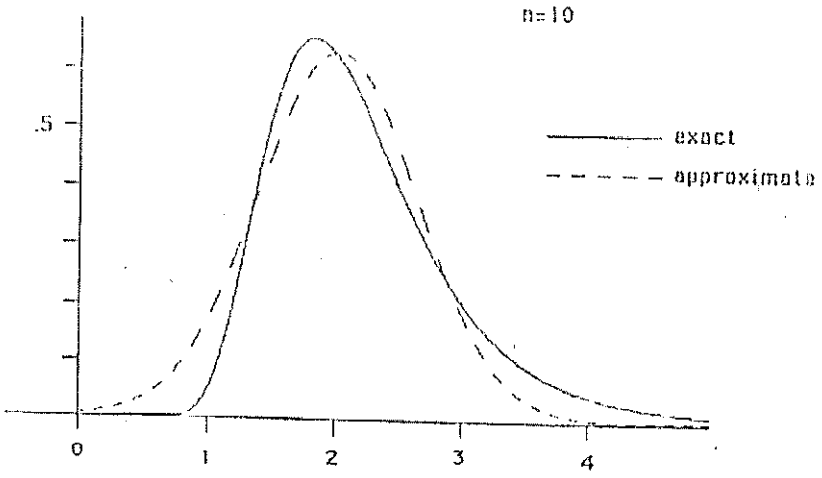
$\rightarrow \hat{\lambda}$ is approximately dist. $\sim \left(\lambda, \frac{\lambda^2}{n}\right)$

or take the 2nd deriv of $\log f(x_i|\lambda)$:

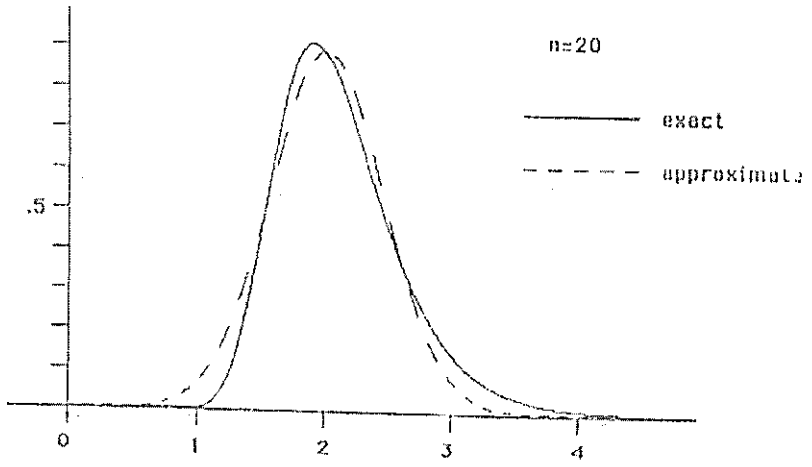
$$\frac{d^2}{d\lambda^2}(\quad) = \frac{d}{d\lambda}\left(\frac{1}{\lambda} - x_i\right) = \frac{-1}{\lambda^2} \qquad E\left(-\frac{1}{\lambda^2}\right) = \frac{-1}{\lambda^2}$$

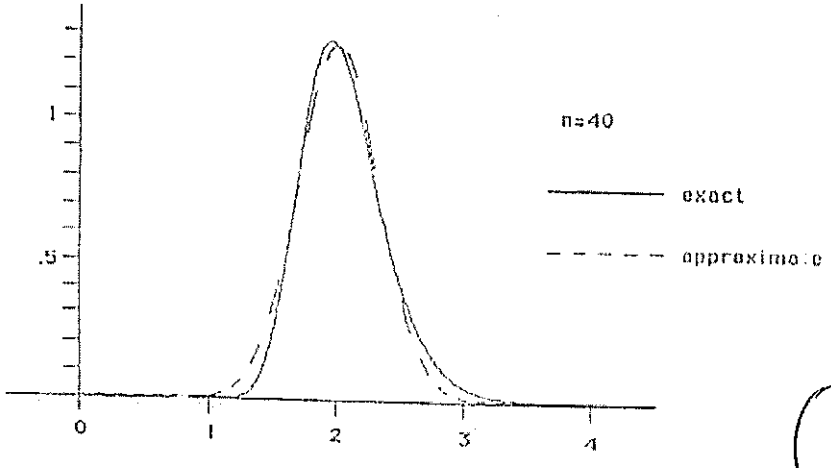
Figure 5.13

Distributions for $\hat{\lambda} = \frac{1}{\bar{x}}$



$\lambda = 2$





Ex: Binomial "Independent case"

$X_1, \dots, X_n \quad P(X_i = 1 \mid \theta) = 1 - P(X_i = 0 \mid \theta) = \theta$

$$p(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

$$\frac{d}{d\theta} \log p(x_i \mid \theta) = \frac{d}{d\theta} \left(x_i \log \theta + (1 - x_i) \log (1-\theta) \right)$$

$$= \frac{x_i}{\theta} - \frac{(1-x_i)}{1-\theta}$$

$$= \frac{x_i (1-\theta) - \theta (1-x_i)}{\theta (1-\theta)}$$

$$= \frac{x_i - \theta}{\theta (1-\theta)}$$

$$E\left(\frac{d}{d\theta} \log p(x_i \mid \theta) \right)^2 = \frac{E(x_i - \theta)^2}{(\theta (1-\theta))^2} \rightarrow \mathrm{Var}(x_i) = \theta (1-\theta)$$

$$= \frac{1}{\theta (1-\theta)}$$

$$\Rightarrow \gamma_n^2 = \frac{\theta (1-\theta)}{n}$$

or

$$\frac{d^2}{d\theta^2} \log p(x_i \mid \theta) = -\frac{x_i}{\theta^2} - \frac{(1-x_i)}{(1-\theta)^2}$$

$$E\left[\frac{d^2}{d\theta^2} \log p(x_i \mid \theta) \right] = -\left[\frac{x_i (1-\theta)^2 + (1-x_i)\theta^2}{\theta^2 (1-\theta)^2} \right]$$

$$= \frac{\theta - 2\theta^2 + \theta^2}{\theta^2 (1-\theta)^2}$$

$$= -\left[\frac{x_i - 2\theta x_i + \theta^2 x_i - \theta^2 x_i + \theta^2}{\theta^2 (1-\theta)^2} \right]$$

$$= \frac{1}{\theta (1-\theta)}$$

$$= -\left[\frac{x_i - 2\theta x_i + \theta^2}{\theta^2 (1-\theta)^2} \right]$$

Ex: Binomial "General Case"

X = # Successes n trials

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = L(\theta)$$

$$\frac{d}{d\theta}\log L(\theta) = \frac{d}{d\theta}\left[\log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)\right]$$

$$= \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x-n\theta}{\theta(1-\theta)}$$

$$E\left[\frac{d}{d\theta}\log L(\theta)\right]^2 = \frac{E(x-n\theta)^2}{(\theta(1-\theta))^2} \leftarrow \text{Var}(x)$$

$$= \frac{n\theta(1-\theta)}{(\theta(1-\theta))^2} = \frac{n}{\theta(1-\theta)}$$

$$\Rightarrow \quad \gamma_n^2 = \frac{\theta(1-\theta)}{n}$$

Although we are demonstrating the use of Fisher's approximation, for the Binomial dist we also know the exact answer, so in this case, the variance (here also the MSE) of the MLE

$\frac{X}{n}$ is exactly $\frac{\theta(1-\theta)}{n}$!

Ex. Genetic Linkage in Corn

$n = 3839$ seedlings in 4 classes

	Green	White
Starchy	1997	906
Sugary	904	32

$= \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

Probs:

	Green	White	
starchy	$(2+\theta)/4$	$(1-\theta)/4$	$3/4$
sugary	$(1-\theta)/4$	$\theta/4$	$1/4$
	$3/4$	$1/4$	

θ = a coeff of linkage, unknown.

n indep trials, each with these probabilities ($\theta = \frac{1}{4}$ if no linkage)

Liklihood:

$$L(\theta) = \left(\frac{2+\theta}{4}\right)^a \left(\frac{1-\theta}{4}\right)^b \left(\frac{1-\theta}{4}\right)^c \left(\frac{\theta}{4}\right)^d$$

$$= (2+\theta)^a (1-\theta)^{b+c} \theta^d / 4^n$$

$$\log L(\theta) = a \log(2+\theta) + (b+c)\log(1-\theta) + d \log \theta$$

$$\frac{d}{d\theta} \log L(\theta) = \frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta}.$$

Set $= 0$: $\frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta} = 0$

$\Rightarrow$ quadratic in θ

$\hat{\theta} = 0.0357$ (only positive root)

Find γ_n:

(use general case)

Find $\frac{d^2}{d\theta^2} \log L(\theta)$

$$\frac{d}{d\theta} \log L(\theta) = \frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta}$$

$$\frac{d^2}{d\theta^2} \log L(\theta) = \frac{-a}{(2+\theta)^2} - \frac{b+c}{(1-\theta)^2} - \frac{d}{\theta^2}$$

Take expectations: $E(a) = \frac{2+\theta}{4} \cdot n$

$$E(b) = E(c) = \frac{1-\theta}{4} \cdot n \qquad E(d) = \frac{\theta}{4} \cdot n$$

$$-E\left(\frac{d^2}{d\theta^2} \log L(\theta)\right) = \frac{n}{4}\left[\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta}\right]$$

$$\text{So, } \gamma_n^2 = \frac{4}{n}\left(\frac{1}{\frac{1}{(2+\theta)} + \frac{2}{(1-\theta)} + \frac{1}{\theta}}\right)$$

$$= \frac{1}{n} \cdot \frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)}$$

Plugging in $\hat{\theta}$, we get

$$\approx \frac{.13}{n}$$

Now let's do it for the indep

case (1): $\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} x_{i1} & x_{i2} \\ x_{i3} & x_{i4} \end{pmatrix}$

Poss X_i	$f(x_i\|\theta)$	$\log f(x_i\|\theta)$	$\frac{d}{d\theta}\log f$	$\frac{d^2}{d\theta^2}\log f$
$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\frac{(2+\theta)}{4}$	$\log(2+\theta)+c$	$\frac{1}{(2+\theta)}$	$\frac{-1}{(2+\theta)^2}$
$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\frac{(1-\theta)}{4}$	$\log(1-\theta)+c$	$\frac{-1}{(1-\theta)}$	$\frac{-1}{(1-\theta)^2}$
$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\frac{(1-\theta)}{4}$	$\log(1-\theta)+c$	$\frac{-1}{(1-\theta)}$	$\frac{-1}{(1-\theta)^2}$
$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\frac{\theta}{4}$	$\log(\theta)+c$	$+\frac{1}{\theta}$	$-\frac{1}{\theta^2}$

$$E\left(\frac{d^2}{d\theta^2}\log f\right) = \frac{2+\theta}{4}\cdot\left(\frac{-1}{(2+\theta)^2}\right) + \frac{1-\theta}{4}\left(\frac{-1}{(1-\theta)^2}\right)$$

$$+ \frac{(1-\theta)}{4}\left(\frac{-1}{(1-\theta)^2}\right) + \frac{\theta}{4}\left(\frac{-1}{\theta^2}\right)$$

$$= \frac{1}{4(2+\theta)} - \frac{1}{2(1-\theta)} - \frac{1}{4\theta} = \frac{-1}{\gamma^2}$$

$$\rightarrow \gamma^2 = \frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)}$$

Another estimator:

$$\hat{\theta}^* = \frac{a - b - c + d}{n}$$

$$E(\hat{\theta}^*) = \frac{E(a) - E(b) - E(c) + E(d)}{n}$$

$$= \frac{n\left(\frac{2+\theta}{4}\right) - n\left(\frac{1-\theta}{4}\right) - n\left(\frac{1-\theta}{4}\right) + n\left(\frac{\theta}{4}\right)}{n}$$

$$= \theta \qquad \text{unbiased}$$

It can be shown (and will, shortly) that

$$Var(\hat{\theta}^*) = \frac{1-\theta^2}{n} = \frac{(1-\theta)(1+\theta)}{n}$$

Then $$\frac{Var(MLE\,\hat{\theta})}{Var(\hat{\theta}^*)} \geq \frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)(1-\theta)(1+\theta)}$$

$$= \frac{2\theta^2 + 4\theta}{2\theta^2 + 3\theta + 1}$$

< 1 since $\theta < 1$, so $4\theta < 3\theta + 1$

Cramer - Rao in action...

STAT 24400
Lecture 12
November 8, 2016

MLE IV: Multivarido Case and more on the Fisher Approximation Theorem

As you now know by heart,

θ is the parameter, a state of nature.

random vars $X_1, X_2, \ldots X_n$ data

$L(\theta) = f(x|\theta)$ or $f(x_1, \ldots x_n|\theta)$

the likelihood function (model as a function of θ)

We want $\hat{\theta} = \hat{\theta}(x_1, \ldots x_n)$, a random variable, to estimate θ. $\hat{\theta}$ maximizes $L(\theta)$.

In real life, of course, we frequently have multidimensional θ and x.

Ex. Normal Case

n measurements $X_1, \ldots, X_n$

Model: X_i's indep, each $\mathcal{N}(\mu, \sigma^2)$

Data: $X = (X_1, \ldots, X_n)$

Parameter: $\theta = \vec{\theta} = (\mu, \sigma^2)$

$$L(\theta) = f(x|\theta) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2)$$

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

$$= (2\pi\phi)^{-\frac{n}{2}} e^{-\frac{1}{2\phi}\sum(x_i-\mu)^2} \quad (\phi = \sigma^2)$$

$$\log(L(\theta)) = -\frac{n}{2}\log(2\pi\phi) - \frac{1}{2\phi}\sum(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}\log(L(\theta)) = \frac{-1}{\phi}\sum(x_i-\mu) = \frac{n}{\phi}(\bar{x}-\mu)$$

$$\frac{\partial}{\partial\phi}\log(L(\theta)) = \frac{-n}{2\phi} + \frac{1}{2\phi^2}\sum(x_i-\mu)^2$$

Set $= 0$:

$$\hat{\mu} = \bar{X}, \quad \hat{\phi} = \frac{1}{n}\sum(x_i-\hat{\mu})^2$$

$$= \frac{1}{n}\sum(x_i-\bar{X})^2$$

To show this is a max, need 2nd derivs:

Let $\ell_{11}(\theta) = \frac{\partial^2}{\partial \mu^2} \log L(\theta)$

$\ell_{22}(\theta) = \frac{\partial^2}{\partial \phi^2} \log L(\theta)$

$\ell_{12}(\theta) = \frac{\partial^2}{\partial \mu \partial \theta} \log L(\theta)$

Need to show that the eigenvalues of the Hessian matrix, $\left[\frac{\partial^2}{\partial x_i \partial x_j}\right]$, are all negative to ensure a max.

This turns out to be easy!

$\ell_{12}(\theta) = \frac{-n}{\phi^2}(\bar{x} - \mu)$, so $\ell_{12}(\hat{\theta}) = 0$.

Hence the Hessian is diagonal, with eigenvalues given by the diagonal elements. These are:

$\ell_{11}(\theta) = \frac{-n}{\phi}$, so $\ell_{11}(\hat{\theta}) < 0$

$\ell_{22}(\theta) = \frac{n}{2\phi^2} - \frac{1}{\phi^3}\sum (x_i - \mu)^2$

$\ell_{22}(\hat{\theta}) = \frac{n}{2\hat{\phi}^2} - \frac{1}{\hat{\phi}^3}\sum (x_i - \mu)^2$ ← this is $\hat{\phi}$!

$= \frac{n}{2\hat{\phi}^2} - \frac{\hat{\phi}}{\hat{\phi}^3} = -\frac{n}{2\hat{\phi}^2} < 0$

both eigenvalues < 0 ⇒ a max!!

OK. MLE's are $\hat{\mu} = \bar{X}$ and

$$\hat{\sigma}^2 = \hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Bias: $E(\hat{\mu}) = E(\bar{X}) = \mu$ (unbiased)

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$ (after some algebra)

Use $E(W^2) = Var(W) + [E(W)]^2$

so:

$$E(\bar{X}^2) = \frac{E(\Sigma X_i)^2}{n^2} = \frac{Var(\Sigma X_i) + [E(\Sigma X_i)]^2}{n^2}$$

$$= \frac{n\sigma^2 + [n\mu]^2}{n^2} = \frac{\sigma^2}{n} + \mu^2$$

$$E\left(\frac{1}{n}\Sigma X_i^2\right) = \frac{1}{n}\Sigma(X_i^2) = \frac{1}{n}\Sigma(\sigma^2 + \mu^2)$$

$$= \sigma^2 + \mu^2$$

so

$$E(\hat{\sigma}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \left(\frac{n-1}{n}\right)\sigma^2$$

(biased!)

Common to use $s^2 = \frac{n}{n-1}\hat{\sigma}^2$

$$= \frac{1}{n-1}\Sigma(X_i - \bar{X})^2$$

$$E(s^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2 \quad \text{(unbiased)}$$

Note: As long as X_i's are independent with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$,

$\bar{X}$ unbiased for μ

S^2 unbiased for σ^2

('normal' not used for this)

Note 2: MLE of σ is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2},$$

but both $\hat{\sigma}$ and s are biased for σ a little bit, because

$$E(\sqrt{S^2}) \neq \sqrt{E(S^2)} = \sqrt{\sigma^2} = \sigma$$

In fact $E(S) = b(n)\sigma$, a bias correction factor. It gets close to 1 as n grows:

n	4	10	100
$b(n)$	.778	.923	.992

$$b(n) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$ (see Stigler S-14)

MSE:

$$MSE(S^2) = Var(S^2) = \frac{2\sigma^4}{(n-1)}$$ ← move next quarter!

$$Bias(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

$$Var(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)^2 Var(S^2) = \frac{2(n-1)}{n^2}\sigma^4$$

$$MSE(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4$$

Since $\frac{2n-1}{n^2} < \frac{2}{n-1}$ for all $n \geq 2$

$$MSE(\hat{\sigma}^2) < MSE(S^2)$$

$$\left(\text{But } \frac{MSE(\hat{\sigma}^2)}{MSE(S^2)} = 1 - \left(\frac{3n+1}{2n^2}\right) \approx 1\right)$$

How distributed? Next Quarter

We will show that it is exactly true that:

if X_i's are normal themselves

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \text{ dist. } \chi^2_{n-1} \text{ d.f.}$$

Fisher's Theorem

Multidimensional Parameter

$$L(\vec{\theta}) = f(\vec{x} \mid \vec{\theta})$$

If $\hat{\vec{\theta}}$ is found by setting derivs equal to 0, then if n large, $\hat{\vec{\theta}}$ has approx a multiple dim. normal dist. $N(\vec{\theta}, \vec{\tau}^2)$ where $\vec{\tau}^2$ is the inverse of the matrix

$$\left[-E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\vec{\theta}) \right) \right]$$

Idea for a Proof of Fisher's Theorem

Up to now, we found MLE's by solving

$$\frac{d}{d\theta} \log L(\theta) = 0$$

exactly. This isn't always possible. Sometimes (including numerical problems) it is useful to use an approximate method, such as Newton-Raphson.

Let

$$g(\theta) = \frac{d}{d\theta} \log L(\theta)$$

$$g'(\theta) = \frac{d^2}{d\theta^2} \log L(\theta)$$

We want to find $\hat{\theta}$ s.t. $g(\hat{\theta}) = 0$.

Suppose $\hat{\theta}$ is near θ. Then the mean value theorem says that

$$g(\hat{\theta}) - g(\theta) \simeq (\hat{\theta} - \theta) g'(\theta)$$

But we supposed $g(\hat{\theta}) = 0$, so

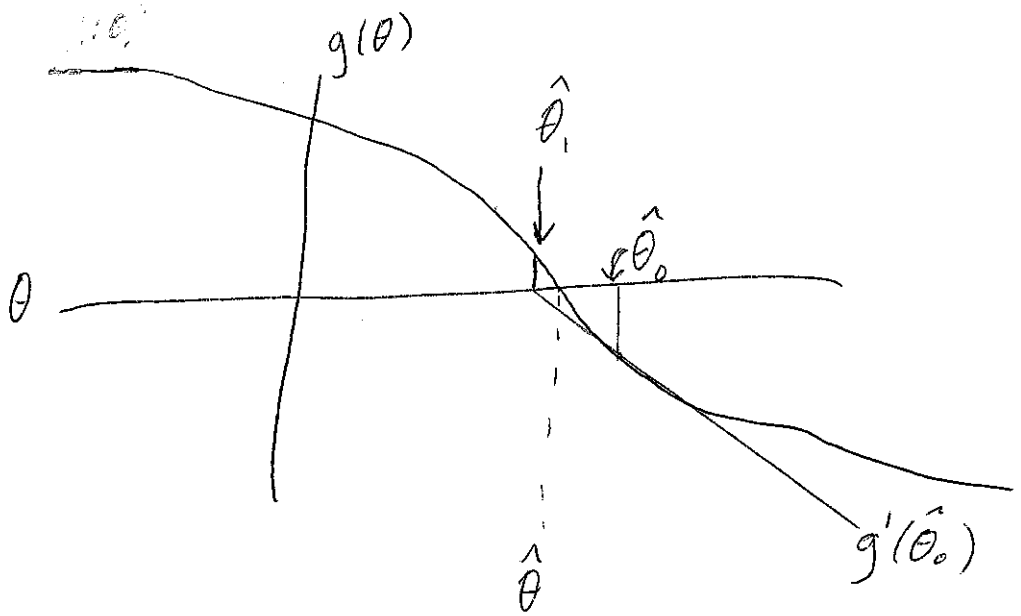
$$-g(\theta) \simeq (\hat{\theta} - \theta) g'(\theta)$$

$$\hat{\theta} - \theta = -\frac{g(\theta)}{g'(\theta)}$$

$$\hat{\theta} = \theta - \frac{g(\theta)}{g'(\theta)}$$

an approximation for $\hat{\theta}$. For numerical work, we can let this approximation be our next guess for $\hat{\theta}$, so if the first guessed θ is $\hat{\theta}_0$, we have just found $\hat{\theta}_1$. Then we can continue by taking

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{g(\hat{\theta}_n)}{g'(\hat{\theta}_n)}$$



This is also the basis of a proof of Fisher's Theorem.

Let's make the X_i i.i.d.

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$\log L(\theta) = \sum \log f(x_i|\theta)$$

Let

$$g(\theta) = \frac{d}{d\theta} \log L(\theta), \text{ and}$$

define

$$Z_i(\theta) = \frac{d}{d\theta} \log f(x_i|\theta)$$

$$= \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)} \quad \text{(we'll need this below)}$$

so

$$g(\theta) = \sum_{i=1}^{n} Z_i(\theta)$$

is a sum of indep. random vars.

Let's calculate $E(Z_i(\theta))$.

$$E(z_i(\theta)) = \int_{-\infty}^{\infty} z_i(\theta)\, f(x/\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} f(x/\theta)}{f(x/\theta)} f(x/\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{d}{d\theta} f(x/\theta)\, dx$$

$$= \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x/\theta)\, dx$$

$$= \frac{d}{d\theta} \cdot (\text{const} = 1)$$

$\int_{-\infty}^{\infty} f(x/\theta)\, d\theta$ is a density

$$= 0.$$

$$Var(z_i(\theta)) = E[(z_i(\theta))^2] - E[z_i(\theta)]^2$$

$$= E\left[\left(\frac{d}{d\theta} \log f(x_i/\theta)\right)^2\right]$$

But we know that

$$E\left[\left(\frac{d}{d\theta} \log f(x_i/\theta)\right)^2\right] = -E\left[\frac{d^2}{d\theta^2} \log f(x_i/\theta)\right] = \underline{\underline{\frac{1}{I^2(\theta)}}}$$

But remember:

$$g(\theta) = \sum_{i=1}^{n} Z_i(\theta).$$

$E(Z) = 0$ and $Var(Z) = \frac{1}{\mathcal{I}}$, so the CLT says that

$$\frac{g(\theta)}{\sqrt{n}} \text{ is distributed approx } N\left(0, \frac{1}{\mathcal{I}^2(\theta)}\right)$$

Also,

$$g'(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} Z_i(\theta) = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

is a sum of indep random vars, so the Law of Large Numbers says that as $n \to \infty$,

$$\frac{1}{n} g'(\theta) \xrightarrow{P} E\left(\frac{d}{d\theta} Z_i(\theta)\right) = E\left(\frac{d^2}{d\theta^2} \log f(x_i|\theta)\right)$$

$$= -\frac{1}{\mathcal{I}^2(\theta)}$$

So now,

$$\sqrt{n}\left(\frac{-g(\theta)}{g'(\theta)}\right) = \frac{g(\theta)/\sqrt{n}}{(g'(\theta)/n)}$$

$$\approx \frac{g(\theta)/\sqrt{n}}{1/\gamma^2(\theta)}$$

$$= \gamma^2(\theta) \cdot \frac{g(\theta)}{\sqrt{n}}$$

will have an approximate distribution of $N\left(0, (\gamma^2(\theta))^2 \cdot \frac{1}{\gamma(\theta)}\right)$, or

$$N(0, \gamma^2(\theta)).$$

Hence $\frac{-g(\theta)}{g'(\theta)}$ is approximately distributed $N(0, \gamma^2(\theta))$.

But we have approximated $\log L(\theta)$ by $\frac{-g(\theta)}{g'(\theta)}$ near $\hat{\theta}$, which is what we sought to ~~prove~~ argue for.

STHT 24400
Lecture 13
11/10/16

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 MSE) from a given set of data. Now let's ask a complementary question, which is how to write an estimator (or "statistic")* that has <u>all</u> 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 $X_1, \ldots X_n$, with parameter θ. Obviously $\vec{T} = (X_1, \ldots, X_n)$ has all the information about θ that can be obtained from these observations. What about

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

$$P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$$

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

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

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

Definition

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

Factorization Theorem (Neyman):

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

$$f(x_1, \ldots, x_n | \theta) = g[T(x_1, \ldots, x_n), \theta]\, h(x_1, \ldots, x_n).$$

Pf For discrete case.

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

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

1st, 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:

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

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

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

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

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

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 statistic not an estimator. 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,

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

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

To prove this, we need a small lemma (Rice, 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 $E(Y) = E[E(Y|X)]$, (Rice, p. 119)

we can write

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

so:

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

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

Pf

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

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

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

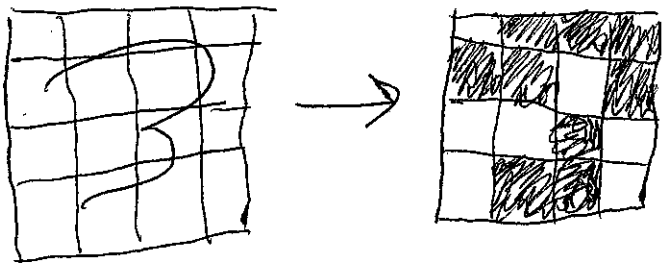
$$Var(\theta^*) = Var(\tilde{\theta}) + E[Var(\theta^*/T)]$$

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

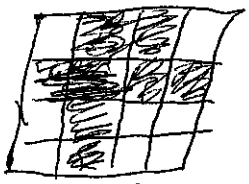
Topic II:
Hypothesis Testing

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.

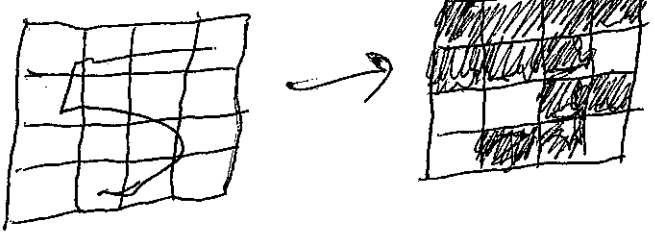
Example: Pattern recognition



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

Husband \ Wife	T	M	S	
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 $\quad \theta_{TT} = \theta_{HT} \cdot \theta_{WT}$

Testing 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 $\frac{f(x|\theta_1)}{f(x|\theta_0)} > k$.

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

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

$\pi = 1 - \beta =$ power of the test.

The LR $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ orders x values

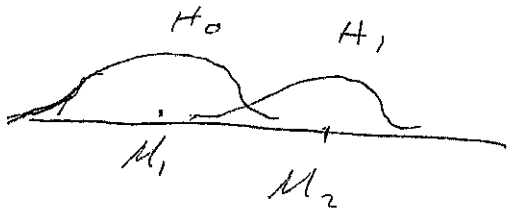
high LR is stronger evidence for H_1
low LR is " " " H_0

K draws the line

Example $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$

$H_1: \mu = \mu_1$



($\sigma^2 = \sigma_0^2$ given)

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

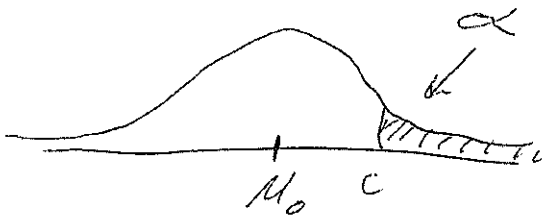
large when $(\mu_1 - \mu_0)\sum x_i$ is large

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

So we obtain the test,

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

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



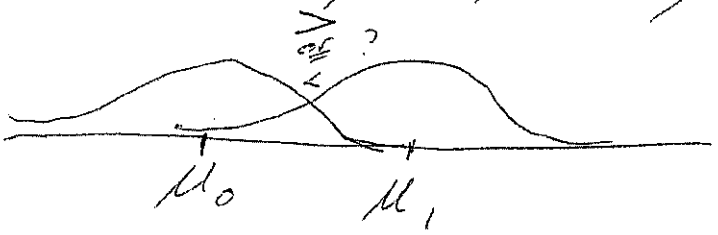
$$c = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

Restricted solution may 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 $\bar{X} > C = \mu_0 + Z_\alpha \frac{\sigma_0}{\sqrt{n}}$

Note

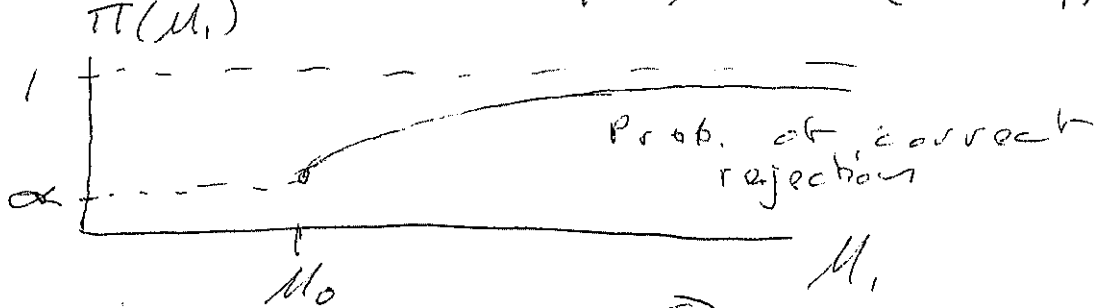
Same test for any $\mu_1 > \mu_0$

But the power depends on μ_1

The test is uniformly most powerful

Describe performance with power function

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



STAT 2.4400
Lecture 14
11/15/16

Testing 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 (1933): Best test to use is Likelihood ratio (LR) test.

Reject H_0 if $\frac{f(x|\theta_1)}{f(x|\theta_0)} > K$.

$\alpha = P(\text{Rej. } H_0 \mid H_0 \text{ true})$ "Type 1" "False Positive"

$\beta = P(\text{Acc. } H_0 \mid H_1 \text{ true})$ "Type 2" "False negative"

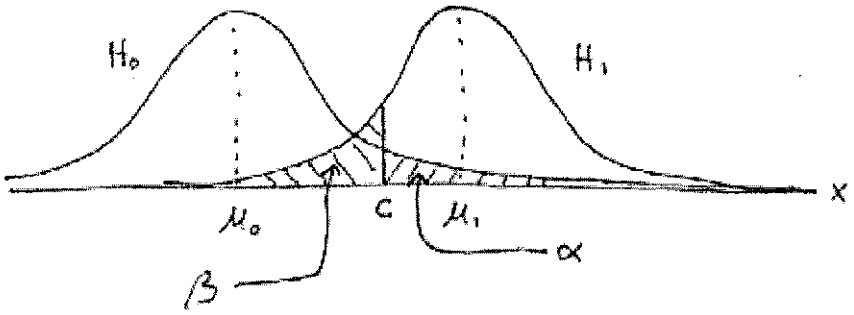
$\pi = 1 - \beta =$ power of the test.

The LR $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ orders x values

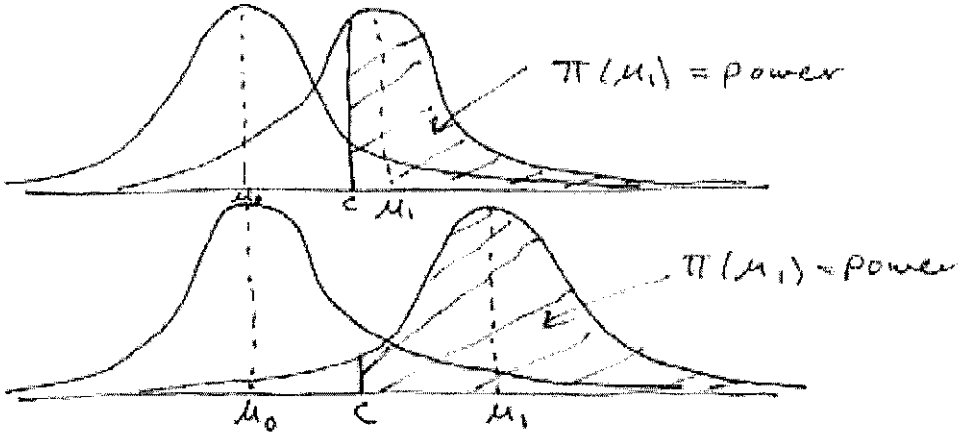
High LR is stronger evidence for H_1
low LR is " " " H_0

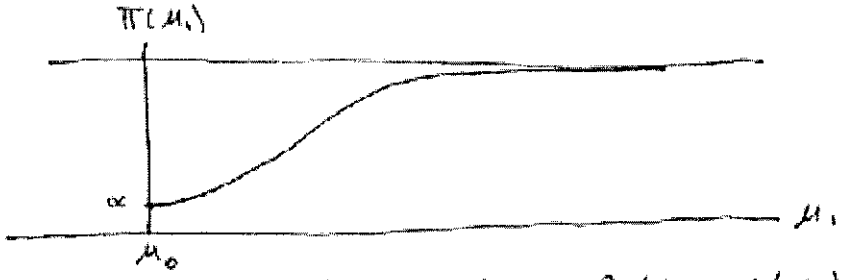
K draws the line

Figure 6.3



(a) Testing a simple hypothesis vs. a simple alternative.





(b) The power function $\pi(\mu_1) = P_r(\text{Reject}(\mu_1))$, as a function of the alternative μ_1

[From Stigler, Chap 6]

The Neyman - Pearson Lemma

Given α, no test with the same or lower α has a lower β than the likelihood ratio with the given α.

Proof

The LR test rejects if $\lambda = x$ for any x satisfying

$$f(x|\theta_1) > k f(x|\theta_0).$$

Define an "indicator function"

$$I_{NP} = \begin{cases} 1 & \text{iff } f(x|\theta_1) > k f(x|\theta_0) \\ 0 & \text{otherwise} \end{cases}$$

I_{NP} is a Bernoulli random variable.

$$E[I_{NP}] = 0 \cdot Pr(I_{NP}=0) + 1 \cdot Pr(I_{NP}=1) = Pr(I_{NP}=1)$$

Let α_{NP} be the probability of a type I error for the NP test. Then

$$\alpha_{NP} = E[I_{NP}(x)|\theta_0] \quad (\text{reject } H_0 \mid H_0 \text{ true})$$

$$1 - \beta_{NP} = E[I_{NP}(x)|\theta_1] \quad (\text{reject } H_0 \mid H_1 \text{ true}, \text{ "power"})$$

Let T be any other test with $\alpha_T \leq \alpha_{NP}$, and let $I_T = \begin{cases} 1 & \text{iff } T \text{ rejects } H_0 \text{ data } X = x \\ 0 & \text{otherwise} \end{cases}$

Then $\alpha_T = E[I_T(X) | \theta_0]$ ("reject θ_0 H_0 true")

$1 - \beta_T = E[I_T(x) | \theta_1]$ ("accept θ_1, H_1 true")

Claim: for all x,

$$I_{NP}(x)\left[f(x|\theta_1) - K f(x|\theta_0)\right] \geq I_T(x)\left[f(x|\theta_1) - K f(x|\theta_0)\right]$$

Why? The part in [ ] is the same on both sides. If $I_{NP}(x) = 1$, then $[\,] \geq 0$, and since $I_{NP}(x) = 1 \geq I_T(x)$, the inequality is true. If $I_{NP}(x) = 0$, $[\,] \leq 0$ and the inequality holds because

$$I_T(x) \geq 0 = I_{NP}(x).$$

Multiply out the inequality to get

$$I_{NP}(x) f(x|\theta_1) - K I_{NP}(x) f(x|\theta_0) \geq I_T(x) f(x|\theta_1) - K I_T(x) f(x|\theta_0)$$

Now, sum or integrate or multiply integrate over x to give expectations.

$$E[I_{NP}(x)|\theta_1] - kE[I_{NP}(x)|\theta_0] \geq E[I_T(x)|\theta_1] - kE[I_T(x)|\theta_0]$$

Now let's change notation back to α's and β's:

$$1 - \beta_{NP} - k\alpha_{NP} \geq 1 - \beta_T - k\alpha_T$$

$$1 - \beta_{NP} \geq 1 - \beta_T + k(\alpha_{NP} - \alpha_T)$$

but: $\alpha_{NP} - \alpha_T \geq 0$, $k \geq 0$, so

$$1 - \beta_{NP} \geq 1 - \beta_T$$

$$\beta_T \geq \beta_{NP}$$

∴

So: the LR test is the <u>most</u> <u>powerful</u> for a particular θ_1, given α.

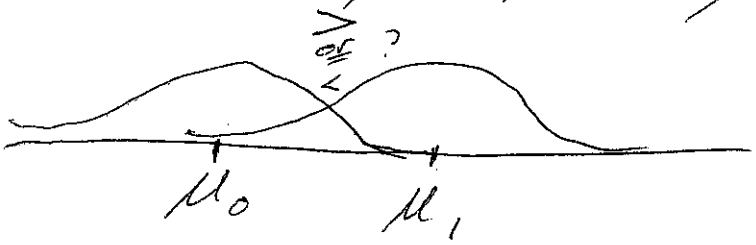
What about composite tests?

Restricted solution may 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 $\bar{X} > C = \mu_0 + Z_\alpha \frac{\sigma_0}{\sqrt{n}}$

Note

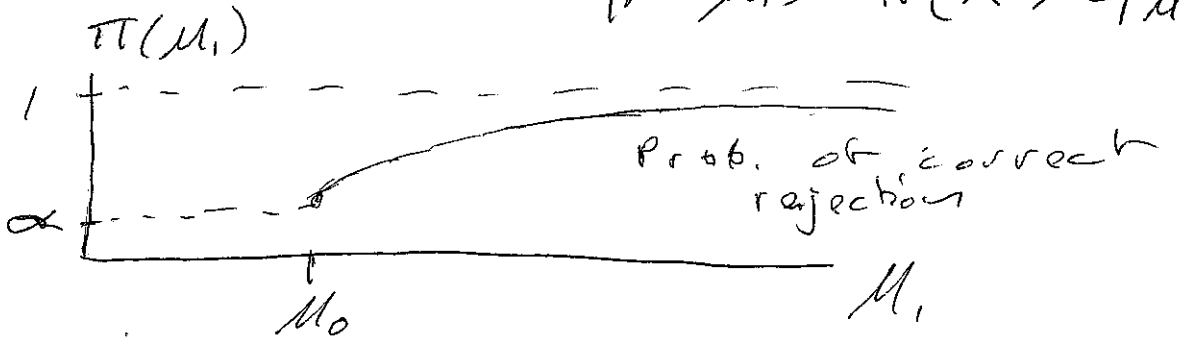
Same test for any $\mu_1 > \mu_0$

But the power depends on μ_1.

The test is uniformly most powerful!

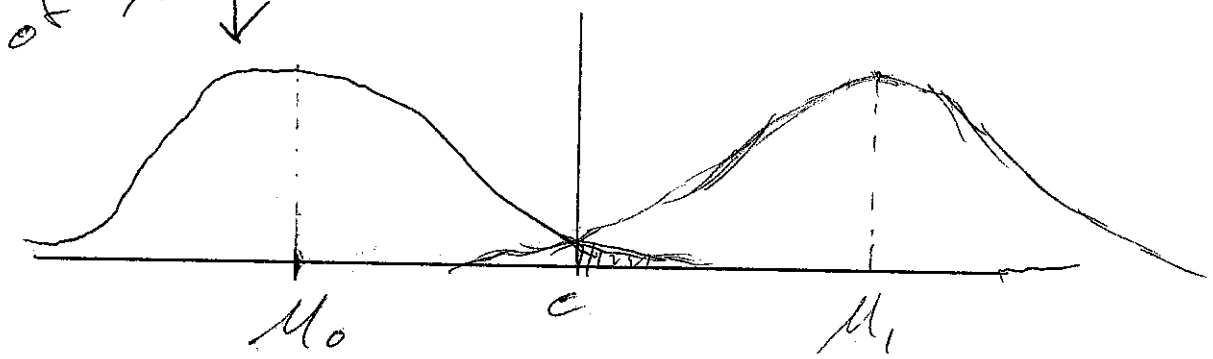
Describe performance with power function

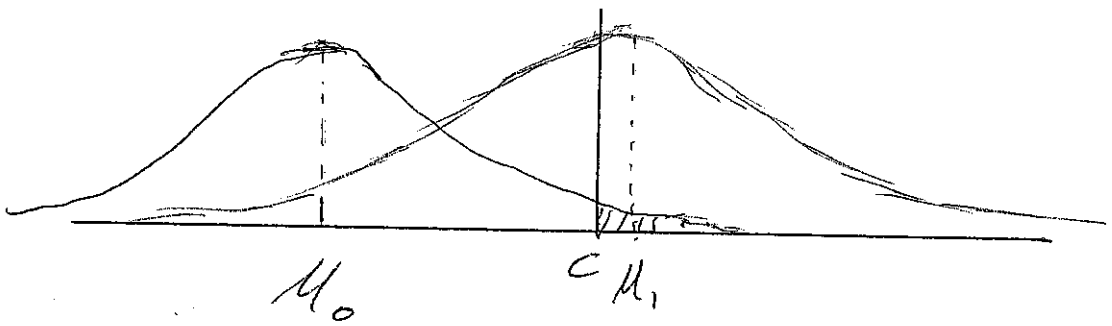
$\pi(\mu_1) = Pr(\text{Reject } H_0 | \mu = \mu_1) = Pr(\bar{X} > C | \mu = \mu_1)$

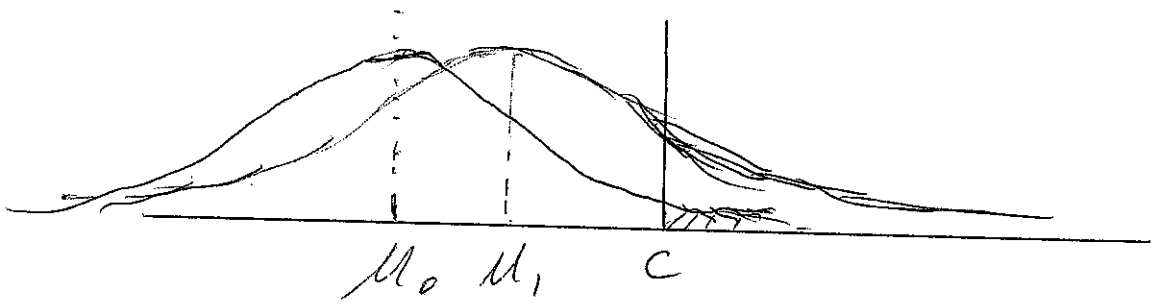


$$C = \mu_0 + Z_\alpha \frac{\sigma_0}{\sqrt{n}}$$

density of $\bar{X}$ under H_0







Example: X binomial (n, θ)

$H_0: \theta = \theta_0 \left(= \frac{1}{2}\text{? "fair coin"}\right)$

$H_1: \theta = \theta_1 > \theta_0$

Likelihood Ratio: $\frac{p(x|\theta_1)}{p(x|\theta_0)} = \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^n$

$\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right) > 1$ — (because: $\frac{\theta_1}{\theta_0} > 1$, $\frac{1-\theta_0}{1-\theta_1} > 1$)

large when x large

Test: Reject H_0 if $x > c$

want $R(x > c \mid \theta = \theta_0) = \alpha$

(not possible exactly for all α)

Ex: $n = 5$ $\theta_0 = \frac{1}{2}$

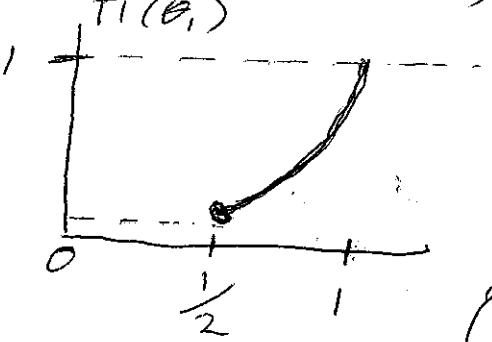
x	0	1	2	3	4	5
$p(x\|\theta_0)$	.03	.16	.31	.31	.16	.03

$\alpha = .03 \rightarrow$ Reject H_0 if $x > 4$ (≥ 5)

$\alpha = .19 \rightarrow$ Reject H_0 if $x > 3$

Power function? For $\alpha = .03$, $c = 4$

$\pi(\theta_1) = P(X > 4 \mid \theta = \theta_1) = P(x = 5 \mid \theta = \theta_1) = \theta_1^5$



UMP

But: In General, when testing composite* hypotheses [*ie more than one distribution in H_0 and/or H_1]

there is no UMP test.

Ex: $X_1, \ldots, X_n \quad \mathcal{N}(\mu, \sigma_0^2)$ ← Known

$H_0: \mu = \mu_0$

$H_1: \mu = \mu_1 \neq \mu_0$ [composite]

Possible Tests:

(1) Reject if $\bar{X} > C$ ——————

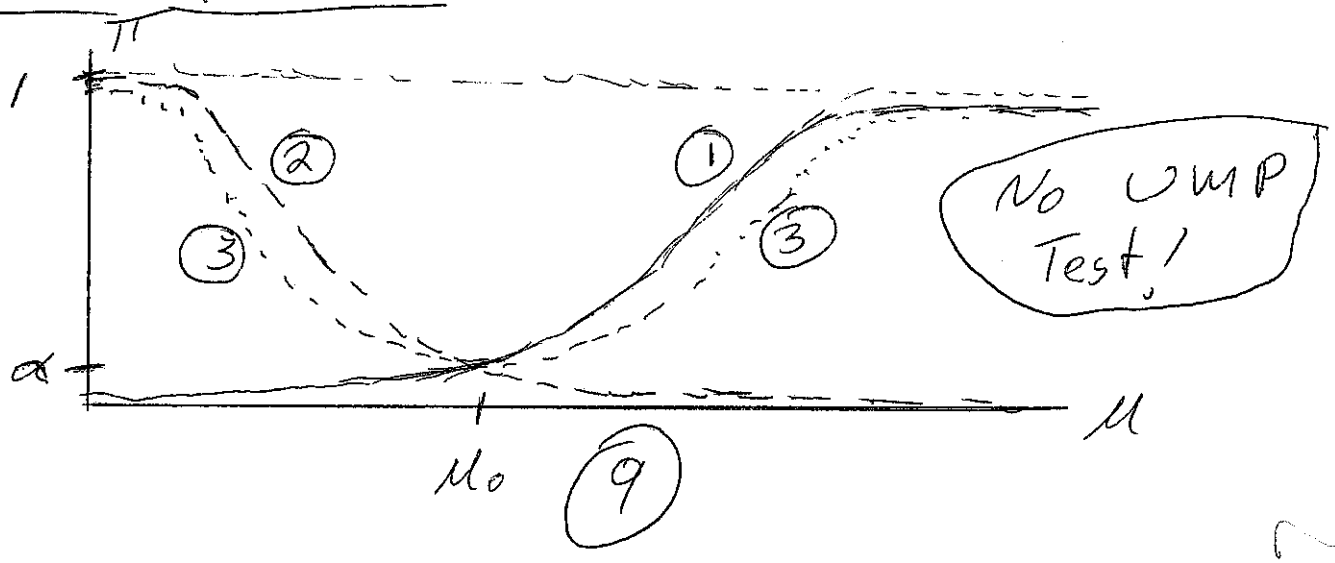
(Best vs $\mu_1 > \mu_0$

(2) Reject if $\bar{X} < C'$ — — — — —

(Best vs $\mu_1 < \mu_0$)

(3) Reject if $|\bar{X} - \mu_0| > C'''$ .....

Power Functions:



Likelihood Ratio Tests – General Case

Test H_0: Group of θ_0's

H_1: Group of θ_1's

Idea: Compare "champ" of H_0 to "champ" of H_1.

Could compare $\max_{\theta_0\text{'s}} L(\theta)$ to $\max_{\theta_1\text{'s}} L(\theta)$

Instead, compare $\max_{\theta_0\text{'s}} L(\theta)$ to $\max_{\text{all } \theta} L(\theta)$

(max at MLE!)

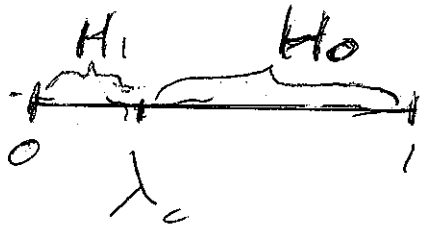
Let $\lambda = \dfrac{\max_{\theta_0\text{'s}} L(\hat{\theta}_0)}{\max_{\text{all } \theta\text{'s}} L(\hat{\theta}_1)}$

If H_0 clearly best, $\lambda = 1$

If H_1 clearly best, $\lambda \ll 1$

So: Reject H_0 if $\lambda < \lambda_c$, with $P(\lambda < \lambda_c \mid H_0) \leq \alpha$

"Likelihood Ratio Test"



Examples of Likelihood Ratio Tests

1) Neyman-Pearson Tests are a special case

2) Student's t-tests, tests of 1 or 2 means with unknown variances.

3) ANOVA - "Analysis of Variance"

4) Regression tests

5) Variance comparisons

6) Chi-Square tests

2, 3, 4, and 5: 245!

Chi-Square Tests

Contingency Tables

Tests of Fit

But First:

Multinomial Distributions

Multinomial: Generalization of Binomial

n independent trials

For each trial:

(mutually exclusive) Outcomes $A_1, A_2, \ldots, A_K$

Probabilities $\theta_1, \theta_2, \ldots, \theta_K$

$(\theta_1 + \theta_2 + \cdots + \theta_K = 1)$

Counts $X_1, X_2, \ldots, X_K$

$(X_1 + X_2 + \cdots + X_K = n)$

Ex: $K=2$ $A_1 =$ "H", $A_2 =$ "T"

$X_1 = X$, $X_2 = n - X$, X Binomial

Ex: $K = 11$ $A_i =$ "pair of dice total $i+1$"

$(i = 1, 2, \ldots, 11)$

Roll pair of dice n times

$X_1 =$ #2's, $\ldots$, $X_{11} =$ #12's, $\sum X_i = n$

Ex: $K = 38$. Roulette wheel, n spins

$X_1 =$ #1's, $\ldots$, $X_{36} =$ #36's, $X_{37} =$ #0's, $X_{38} =$ #00's

n trials, K outcomes each

X_i = count of # A_i's

Note: For any A_i (say A_3) we can regroup: A_3 vs. all others. Then X_3 = # A_3's can be seen to have a binomial marginal distribution:

$$P(X_3 = j) = \binom{n}{j} \theta_3^j (1-\theta_3)^{n-j}$$

$$E(X_3) = n\theta_3, \quad Var(X_3) = n\theta_3(1-\theta_3)$$

Same for any single X_i;

Multinomial Distribution:

$(X_1, X_2, \ldots, X_k)$ are dependent, multivariate

$$p(x_1, x_2, \ldots, x_n | \theta's) = Pr(X_1 = x_1, \ldots, X_n = x_n | \theta's)$$

$$= \frac{n!}{x_1! x_2! \ldots x_n!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}$$

if $x_1 + x_2 + \cdots + x_k = n$

$= 0$ otherwise

Estimation

$$L(\theta_1, \ldots, \theta_k) = \frac{n!}{x_1! x_2! \cdots x_n!} \theta_1^{x_1} \cdots \theta_k^{x_k}$$

$$\theta_1 + \cdots + \theta_k = 1$$

MLE's $\hat{\theta}_i = \frac{x_i}{n}$ – sample fractions

[can show this from $\frac{\partial}{\partial \theta_i} L(\vec{\theta})$

Testing: $H_0: \theta_1 = a_1, \ldots, \theta_k = a_k$

H_1: "otherwise"

L. R. Test: $$\lambda = \frac{\max_{H_0 \theta's} L(\theta_1, \ldots, \theta_n)}{\max_{\text{all } \theta's} L(\theta_1, \ldots, \theta_n)}$$

Reject if $\lambda < \lambda_c$. Let $E(x_i | H_0) = m_i = n a_i$

Then $$\lambda = \frac{L(a_1, \ldots, a_k)}{L(\hat{\theta}_1, \ldots, \hat{\theta}_n)} = \left(\frac{a_1}{\hat{\theta}_1}\right)^{x_1} \cdots \left(\frac{a_k}{\hat{\theta}_k}\right)^{x_k}$$

$$= \left(\frac{m_1}{x_1}\right)^{x_1} \cdots \left(\frac{m_k}{x_k}\right)^{x_k}$$

λ small $\rightarrow$ $-\log \lambda$ large, reject if

$$-\log \lambda > K$$

Multinomial LR test, cont:

$$\lambda = \left(\frac{m_1}{x_1}\right)^{x_1} \cdots \left(\frac{m_k}{x_k}\right)^{x_k}$$

Let's invoke Taylor's Thm:

$$f(x-m) = f(m) + f'(m)(x-m) + \frac{1}{2} f''(m)(x-m)^2 + \text{Rem}$$

Let $f(x) = x \log\left(\frac{x}{m}\right)$, so $f(m) = 0$, $f'(m) = 1$, $f''(m) = \frac{1}{m}$

Then $f(x) = (x-m) + \frac{1}{2}\frac{(x-m)^2}{m} + \cdots$

$$-\log \lambda = -\sum x_i \log\left(\frac{m_i}{x_i}\right)$$

$$= \sum x_i \log\left(\frac{x_i}{m_i}\right)$$

$$= \sum (x_i - m_i) + \frac{1}{2}\sum \frac{(x_i - m_i)^2}{m_i} + \text{Rem}$$

$\sum (x_i - m_i) = \sum x_i - \sum m_i = n - n = 0$

$$= 0 + \frac{1}{2}\sum \frac{(x_i - m_i)^2}{m_i} + \text{Rem}$$

$$= \frac{1}{2}\chi^2 + \text{Rem}$$

$$\boxed{-2 \log \lambda \stackrel{\sim}{=} \chi^2} \quad \text{(if Rem small)}$$

So.. "reject if $\chi^2 > c$" is almost the L. R. test..

more next time.

Hypothesis Testing

STAT 244
Lecture 15
11/17/16

Recap:

(1) "Simple" Hypotheses (test one distinct distribution for the data X vs another)

$H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$

NP: use likelihood ratio to test:

density or pmf → $\frac{P(\text{data}|\theta_1)}{P(\text{data}|\theta_0)}$ Reject H_0 if too large

Procedure: (a) Find form of test from Likelihood Ratio (e.g. "Rej. if $X > c$")

(b) Fix $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$

(2) "Composite" hypotheses (test one distribution vs a set of distributions)

e.g. $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$

Sometimes the test from (1) is the same for all $\theta_1 > \theta_0$; then the test is UMP ("uniformly most powerful")

LR test: $\lambda = \frac{\max_{H_0 \theta\text{'s}} L(\theta_1, \ldots \theta_n)}{\max_{\text{all } \theta\text{'s}} L(\theta_1, \ldots, \theta_n)}$

Reject if $\lambda < \lambda_c$

Composite tests like $H_1: \mu > \mu_0$ are great when possible, but the most intellectually revealing examples involve multiple distinct outcomes.

Classic Example: Weldon's Dice

Weldon and assistants rolled 12 dice 26,306 times and counted the number showing 5 or 6 up.

Results:

ARE THE DICE FAIR??

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	203	−18
1	1149	1217	−68
2	3265	3345	−80
3	5475	5576	−101
4	6114	6273	−159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10	14	13	1
11	4	1	3
12	0	0	0
Total	26,306	26,306	0

"Theory" column assumes $X \sim Bin(12, \frac{1}{3})$,

so $26{,}306 \cdot \binom{12}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{10} = 3{,}345.366.$

$\theta = \binom{12}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{10} = 0.128$

$\sqrt{n\theta(1-\theta)} = 54$, so

$X = 2$ column only 1.48 std. dev

Weldon said agreement good, dice fair.

Karl Pearson said: "No way - dice loaded" need 13 tests!?

Last time we introduced the

Multinomial Distribution,

a generalization of the Binomial Distribution for K distinct outcomes.

We found that for this dist,

$$L(\theta_1, \ldots, \theta_K) = \frac{n!}{x_1!\, x_2! \cdots x_K!} \theta_1^{x_1} \cdots \theta_K^{x_K}$$

(K outcomes, n trials)

$$\hat{\theta}_K = \frac{x_i}{n}$$

$H_0: \theta_1 = a_1, \ldots, \theta_K = a_K$

H_1: "otherwise", one equality in H_0 does not hold.

$$\lambda = \frac{L(a_1, \ldots, a_K)}{L(\hat{\theta}_1, \ldots, \hat{\theta}_K)} \qquad m_i = n a_i = E(X_i | H_0)$$

$$= \frac{L(a_1, \ldots, a_K)}{L\left(\frac{x_1}{n}, \ldots \frac{x_K}{n}\right)} = \left(\frac{m_1}{x_1}\right)^{x_1} \cdots \left(\frac{m_K}{x_K}\right)^{x_K}$$

$$-\log \lambda = \sum_{i=1}^{K} \log \left[\frac{m_i}{x_i}\right]^{x_i} = \sum_{i=1}^{K} x_i \log\left(\frac{x_i}{m_i}\right) > c.$$

We further showed that

$$-\log \lambda \overset{\sim}{=} \frac{1}{2} \sum \frac{(x_i - m_i)^2}{m_i} = \frac{1}{2} \chi^2$$

For "large" n, tests equivalent

("large" means all $m_i \geq 3.5$)

$$\chi^2 = \sum_{\substack{\text{all} \\ \text{categories}}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

How is χ^2 distributed? We'll do the $k=2$ case: ($X_2 = n - X_1$; $a_2 = 1 - a_1$, $m_2 = n - m_1$)

$$\chi^2 = \sum_{i=1}^{2} \frac{(X_i - m_i)^2}{m_i}$$

$$= \frac{(X_1 - m_1)^2}{m_1} + \frac{(X_2 - m_2)^2}{m_2}$$

$$= \frac{(X_1 - na_1)^2}{na_1} + \frac{((n - X_1) - n(1-a_1))^2}{n(1-a_1)}$$

$$= \frac{(X_1 - na_1)^2}{na_1} + \frac{(X_1 - na_1)^2}{n(1-a_1)}$$

$$= \frac{(X_1 - na_1)^2}{na_1(1-a_1)}$$

$$= \left(\frac{X_1 - na_1}{\sqrt{na_1(1-a_1)}} \right)^2$$

X_1 is the result of adding n trials, $E(X) = na_1$, $Var(X) = na_1(1-a_1)$ SO WE INVOKE THE CLT which says that approximately,

$$\left(\frac{X_1 - na_1}{\sqrt{na_1(1-a_1)}} \right) \sim N(0,1)$$

Hence χ^2 is distributed as X^2, where $X \sim N(0,1)$. Hence χ^2 is chi-square distributed with 1 degree of freedom for $k=2$.

In general, for K outcomes we have $K-1$ degrees of freedom.

Moreover:

$$E(\chi^2) = E\sum_{i=1}^{K} \frac{(x_i - na_i)^2}{na_i}$$

$$= \sum_{i=1}^{K} \frac{E(x_i - na_i)^2}{na_i}$$

$$= \sum_{i=1}^{K} \frac{Var(x_i)}{na_i}$$

$$= \sum_{i=1}^{K} \frac{na_i(1-a_i)}{na_i}$$

$$= \sum_{i=1}^{K} (1-a_i) = K - \sum_{i=1}^{K} a_i$$

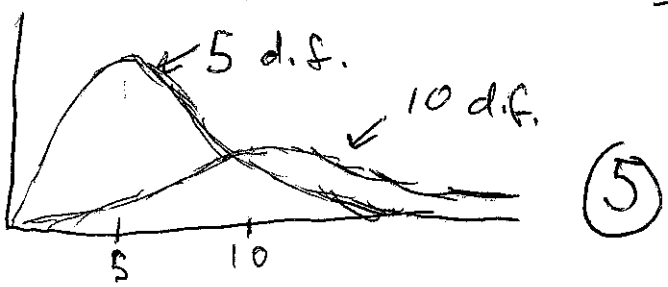
$$= K-1$$

$$\chi^2 = \sum \frac{(obs - Exp)^2}{Exp} = \sum_{i=1}^{n} \frac{(x_i - na_i)}{na_i}$$

Reject H_0 if $\chi^2 > C$

C from χ^2 dist $k-1$ d.f.

[essentially an LR test]



Multinomial Dist.

Outcome: A_i

Probability: θ_i

Expected count: $n\theta_i$

H_0 Prob: a_i

H_0 Expected counts na_i

observed counts: x_i

counts total n

Probs total 1

Weldon's data have too few counts in $X = 11$ or $X = 12$ to use the χ^2 approximation, so let's group 10, 11, and 12's.

Table 7.2. Weldon's dice data with the last three categories grouped together.

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	203	–18
1	1149	1217	–68
2	3265	3345	–80
3	5475	5576	–101
4	6114	6273	–159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10 – 12	18	14	4
Total	26,306	26,306	0

$$\chi^2 = \frac{(-18)^2}{203} + \cdots + \frac{(4)^2}{14} = 35.5 \quad \text{(no roundoff)}$$

$K = 11$, so d.f $= K - 1 = 10$

Table in Rice says $\chi^2_{10} = 25.19$ for $Pr(H_0) = 0.005$

In fact, $\chi^2_{10} = 35.5 \Rightarrow Pr(H_0) \sim 10^{-4}$

Karl Pearson: H_0 is "intrinsically incredible" ... correct!

What if we need to find H_0 from the data?

Instead of $H_0: X \sim Bin(12, \frac{1}{3})$, what if we want to consider

$H_0: X \sim Bin(12, \theta)$ for some θ?

General Problem: k cells, $\theta_i = p(i)$

$H_0: \theta = a_i(\theta)$, $i = 1, \ldots, k$

H_1: "otherwise"

Ex: $a_i(\theta) = \binom{12}{i} \theta^i (1-\theta)^{12-i}$

$i = 0, 1, \ldots, 12$ (13 cells)

H_0: Some Binomial

H_1: "other"

etc etc ...

We procede as before with Weldon's data, but with 2 changes:

(1) a_i replaced by $a_i(\hat{\theta})$ and m_i replaced by $n a_i(\hat{\theta})$, where $\hat{\theta}$ = MLE of θ, assuming H_0 holds.

(2) d. f. $= k - 1 - 1 = k - 2$

← price of estimating one parameter

Results:

$$\hat{\theta} = \frac{\#\,5\text{'s or } 6\text{'s}}{\#\,\text{trials}} = \frac{(0)(185) + (1)(1149) + (2)(3265) + \cdots}{(12)(26,306)}$$

$$= 0.33769862 \quad (\text{from ungrouped data})$$

recompute the table:

Table 7.2. Weldon's Dice Data. The Theory column has been recomputed using the maximum likelihood estimate of the probability of a 5 or 6, namely 0.33769862.

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	187.4	-2.4
1	1149	1146.5	2.5
2	3265	3215.2	49.8
3	5475	5464.7	10.3
4	6114	6269.3	-155.3
5	5194	5114.7	79.3
6	3067	3042.5	24.5
7	1331	1329.7	1.3
8	403	423.8	-20.8
9	105	96.0	9.0
10	14	14.7	-0.7
11	4	1.4	2.6
12	0	0.1	-0.1
Total	26,306	26,306	0.0

group these, as before

$$\chi^2 = \frac{(185-187.4)^2}{187.4} + \frac{(1149-1146.5)^2}{1146.5} + \cdots$$

$$= 8.2, \text{ with } 11-1-1 = 9 \text{ d.f.}$$

close to expectation (and median) of χ^2, H_0 strongly supported.

We <u>can</u> compare these hypotheses, as long as degrees of freedom included. Should have (8) grouped same way, but effect small.

Other types of data can sometimes be treated by the multinomial distr, and hence χ^2. (other dists that give rise to contingency tables can also be treated by χ^2 techniques, since we derived χ^2 by taking a Taylor expansion around a likelihood maximum)

Contingency Tables

(a cross classification of data into 2 (in general N) categories)

Ex. Galton's Data

Husband \ Wife	T	M	S	
T	18	28	14	60
M	20	51	28	99
S	12	25	9	46
	50	104	51	205

3 × 3 Table

$n = 205$ counts

$K = 3 \times 3 = 9$ cells

In General

Cell counts X_{ij} ← rows, $i = 1, \ldots, r$; columns, $j = 1, \ldots, c$

$r \times c$ table: notation

Factor A \ Factor B	B_1	B_2	...	B_c	
A_1	X_{11}	X_{12}	---	X_{1c}	X_{1+}
A_2	X_{21}				
⋮	⋮		X_{ij}		
A_r	X_{r1}			X_{rc}	
	X_{+1}				

$X_{1+} = \sum_{j=1}^{c} X_{1j}$

row i, column j

$X_{+1} = \sum_{i=1}^{r} X_{i1}$

(multinomial contingency tables, cont'd)

$r \times c$ cells.

$\theta_{ij} = P(\text{Trial gives } A_i \cap B_j)$, $\sum_i \sum_j \theta_{ij} = 1$

$X_{ij} = \#$ trials with $A_i \cap B_j$, $\sum_i \sum_j X_{ij} = n$

X_{ij}'s multinomial n trials, probs θ_{ij}

$\theta_{ij} = P(A_i \cap B_j)$

H_0: Factors independent

or $P(A_i \cap B_j) = P(A_i) P(B_j)$

or $\theta_{ij} = (\theta_{i+})(\theta_{+j})$ $(\theta_{i+} = \sum_j \theta_{ij})$

(e.g. $P(\text{Hus T} \cap \text{Wf T}) = P(\text{Hus T}) P(\text{Wf T})$)

H_1: "otherwise". in detail:

H_1 is <u>Composite hypothesis</u>: $(a_{ij}\ (\theta_{i+}, \theta_{+j}))$ ← comma

depends on $\begin{cases} \theta_{1+}\ \theta_{2+} \cdots \theta_{r+} \\ \theta_{+1}\ \theta_{+2} \cdots \theta_{+c} \end{cases}$

$(r-1) + (c-1)$ parameters

MLE's $\hat{\theta}_{i+} = \frac{X_{i+}}{n}$, $\hat{\theta}_{+j} = \frac{X_{+j}}{n}$

Ex. Galton Data

$$\frac{X_{1+}}{n} = \frac{18+28+14}{205} \quad \text{estimates } P(\text{Hus } T)$$

So, under H_0, MLE of

θ_{ij} is $\hat{\theta}_{i+} \cdot \hat{\theta}_{+j} = \dfrac{X_{i+} X_{+j}}{n \cdot n}$

and the MLE of $m_{ij} = n\theta_{ij}$ is

$$n \cdot \frac{X_{i+} X_{+j}}{n \cdot n} = \frac{X_{i+} X_{+j}}{n}$$

$$= \frac{(\text{row total})(\text{col total})}{(\text{total})}$$

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(x_{ij} - \frac{X_{i+} X_{+j}}{n}\right)^2}{\left(\frac{X_{i+} X_{+j}}{n}\right)}$$

for the Galton Data:

$$\chi^2 = \frac{\left(18 - \frac{60 \cdot 50}{205}\right)^2}{\frac{60 \cdot 50}{205}} + \cdots \quad (9 \text{ terms})$$

$$= 2.91$$

$$df = r \cdot c - 1 - [(r-1) + (c-1)]$$

"k" − 1 − "m"

$$= (r-1)(c-1) = 2 \cdot 2 = 4$$

χ^2 even less than expected value!

STAT 249
11/22/2016
Lecture 16

Review

Testing Composite Hypotheses

H_0: Group of θ_0's — Mutually Exclusive
H_1: Group of θ_1's

Likelihood Ratio Test

Reject H_0 if $\lambda < \lambda_c$ $(0 \le \lambda \le 1$ always$)$

$$\lambda = \frac{\max_{\theta_0\text{'s}} L(\theta)}{\max_{\text{all } \theta\text{'s}} L(\theta)}$$

← Restricted to H_0
← No Restrictions

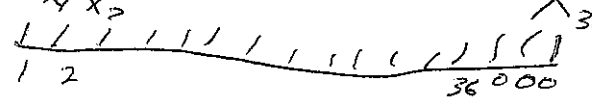
[Alternative, equivalent version:
Reject H_0 if $-2 \log \lambda > K$]

Multinomial Distrib. k dimensional

$X = (X_1, X_2, \ldots, X_k)$, $X_1 + \cdots + X_k = n$

(n indep trials, $A_1, \ldots, A_n$ outcomes,
$X_i = \#A_i$'s, $\theta_i = P(X_i)$, $\sum \theta_i = 1$)

Ex: Roulette
n spins $k = 38$



Marginal Dist. X_i is Binomial(n, θ_i)
X_i's NOT indep!

Testing Roulette Wheel

χ^2 test is approx LR test.

k cells

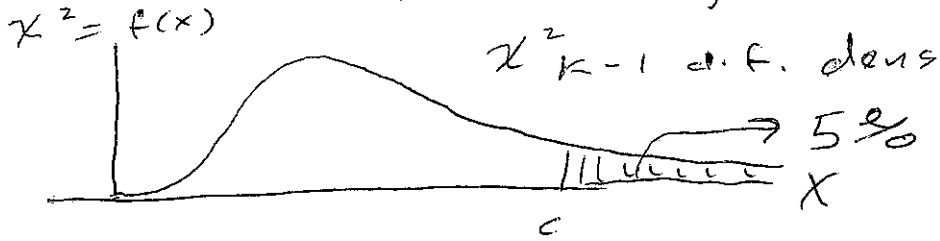
$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{obs} - \text{Exp})^2}{\text{Exp}} = \sum_{i=1}^{k} \frac{(x_i - n a_i)^2}{n a_i}$$

$a_i = H_0$ prob of cell i

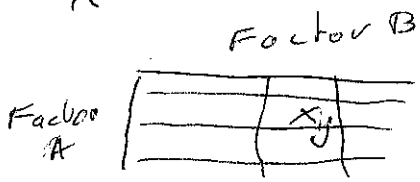
$n a_i = u_i$ = expected counts under H_0

Reject if $\chi^2 > C$

C from χ^2 table, $k-1$ d.f.



Contingency Tables



X_{ij} multinomial n trials probs θ_{ij}

$$\theta_{ij} = P(A_i \cap B_j)$$

H_0: Factors indep

$$\theta_{ij} = (\theta_{i+})(\theta_{+j})$$

H_1: "otherwise"

We can use the multinomial dist because we can regard the $r \times c$ cells as k outcomes. What if margins are fixed?

$r \times c$ table

$n = \sum_i \sum_j$ trials

$i = 1, \ldots, r$

$j = 1, \ldots, c$

cell counts X_{ij}

$$X_{i+} = \sum_{j=1}^{c} X_{ij}$$

$$X_{+j} = \sum_{i=1}^{r} X_{ij}$$

Tests of Homogeneity – Draft Lottery

Table 7.4.

Drawing numbers	Months												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Totals
1-122	9	7	5	8	9	11	12	13	10	9	12	17	122
123-244	12	12	10	8	7	7	7	7	15	15	12	10	122
245-366	10	10	16	14	15	12	12	11	5	7	6	4	122
Totals	31	29	31	30	31	30	31	31	30	31	30	31	366

row totals fixed

Example 7.F (Right-Handedness). To what degree is the propensity to be right-handed socially determined? Is it the same in different cultures? In different historical epochs? Two psychologists addressed this question by examining works of art that portrayed activities that could be judged as being done right- or left-handedly. (Stanley Coren and Clare Porac, "Fifty Centuries of Right-Handedness: The Historical Record" Science (1977), Vol. 198, pp. 631-632.) The following tables summarize their findings, looking at the data in two different ways.

Table 7.5. Counts of 1180 art works showing activity that can be categorized as left- or right-handed, (a) by geographical area, and (b) by historical epoch.

(a)	Right	Left	Total	% Right
Central Europe	312	23	335	93%
Medit. Europe	300	17	317	95%
Middle East	85	4	89	96%
Africa	105	12	117	90%
Central Asia	93	8	101	92%
Far East	126	13	139	91%
Americas	72	10	82	88%
Total	1093	87	1180	92.6%

Left and Right-Handedness by Historical Epoch or Location

(b)				
Pre 3000 BC	35	4	39	90%
2000 BC	44	7	51	86%
1000 BC	89	10	99	90%
500 BC	134	8	142	94%
~0 BC	130	4	134	97%
AD 500	39	3	42	93%
AD 1000	57	7	64	89%
AD 1200	40	1	41	98%
AD 1400	44	6	50	88%
AD 1500	63	5	68	93%
AD 1600	68	4	72	94%
AD 1700	66	5	71	93%
AD 1800	95	6	101	94%
AD 1850	38	1	39	97%
AD 1900	71	6	77	92%
AD 1950	80	10	90	89%
Total	1093	87	1180	92.6%

column totals fixed

Ex Adaptation in Evolution (The "MK" Test)

Nature (1991) 351:652

Adaptive protein evolution at the *Adh* locus in *Drosophila*

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

PROTEINS often differ in amino-acid sequence across species. This difference has evolved by the accumulation of neutral mutations by random drift, the fixation of adaptive mutations by selection, or a mixture of the two. Here we propose a simple statistical test of the neutral protein evolution hypothesis based on a comparison of the number of amino-acid replacement substitutions to synonymous substitutions in the coding region of a locus. If the observed substitutions are neutral, the ratio of replacement to synonymous fixed differences between species should be the same as the ratio of replacement to synonymous polymorphisms within species. DNA sequence data on the *Adh* locus (encoding alcohol dehydrogenase, EC 1.1.1.1) in three species in the *Drosophila melanogaster* species subgroup do not fit this expectation; instead, there are more fixed replacement differences between species than expected. We suggest that these excess replacement substitutions result from adaptive fixation of selectively advantageous mutations.

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

	Fixed (DNA changes between species)	Polymorphic (DNA changes within species)	Total
replacement (DNA changes, protein changes)	7	2	$9 = X_{1+}$
synonymous (DNA changes, protein stays the same)	17	42	$59 = X_{2+}$

row totals

Fixed

Fixed Margins:

Product-Multinomial, or Testing Homogeneity of proportions.

Ex 1st Word Usage - James Mill and John Stuart Mill

	I	II	
James	X_{11}	X_{12}	$X_{1+} = 1075$
J. S.	X_{21}	X_{22}	$X_{2+} = 451$
			$n = 1626$

X_{1+}, X_{2+} fixed, given

X_{ij} multinomial <u>within rows</u>.

Ex: 200 applic. to UC (Grad)
100 male, 100 female

	Adm	Deny	Adm no aid	
M	X_{11}	X_{12}	X_{13}	$X_{1+} = 100$
F	X_{21}	X_{22}	X_{23}	$X_{2+} = 100$

H_0: $P(I \mid James) = P(I \mid J.S.)$

or $P(Adm \mid M) = P(Adm \mid F)$

or $P(\text{synonymous} \mid \text{fixed}) = P(\text{synonymous} \mid \text{polymorphic})$

OR $P(B_j \mid A_i) = P(B_j)$ all i, j

Test of Homogeneity

H_0: $(X_{i1}, \ldots, X_{ic})$ Multinomial

X_{i+} trials

probs $(b_1, \ldots, b_c)$ the same for all rows i

Under H_0:

$$L(b_1, b_2, \ldots, b_c) = \prod_{i=1}^{r} \left(\frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!} b_1^{X_{i1}} \cdots b_c^{X_{ic}} \right)$$

$$= \binom{\text{involves}}{X\text{'s}} b_1^{X_{+1}} \cdots b_c^{X_{+c}}$$

$$\hat{b}_j = X_{+j}/n$$

Under H_1:

$$\max L = \binom{\text{involves}}{X\text{'s}} \prod_{i=1}^{r} \left(\frac{X_{i1}}{X_{i+}}\right)^{X_{i1}} \cdots \left(\frac{X_{ic}}{X_{i+}}\right)^{X_{ic}}$$

$$\lambda = \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{m_{ij}}{X_{ij}}\right)^{X_{ij}}, \quad m_{ij} = \frac{X_{i+} X_{+j}}{n}$$

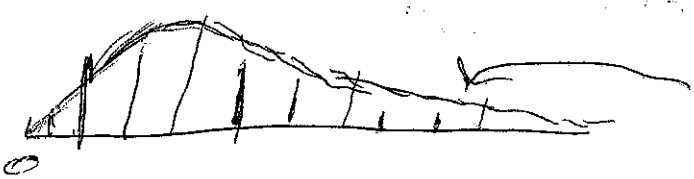
Same test as in full multinomial situation! Same χ^2 statistic, same d.f.

Key Points

χ^2 is for large samples —

χ^2 distribution is an approx to the distribution given Ho (the actual distribution is discrete)

("Expected" ≥ 3.9)



approx. dist. of χ^2

χ^2's were originally designed to avoid being misled by deviation selected as "large"

Ex: Roulette

Ex: Authorship

Ex: Evolutionary Adaptation

$$\chi^2 = \sum_{\substack{\text{all} \\ \text{categories}}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$H_0 = E(\text{count} | H_0)$ or, when this is incompletely specified, the MLE of $E(\text{count} | H_0)$

Uses	d.f.
(1) Test fit, H_0 completely specified [Ex: Roulette] [Ex: Weldon's dice, $P(5 \cup 6) = 1/3$]	$k-1$
(2) Test fit, H_0 not completely specified [Ex: Weldon's dice as test of Binomial dist, $P(5 \cup 6) = \hat{\theta}$]	$k-1-(\#\text{params estimated})$
(3) Test independence [Ex: Galton, tall/short spouse selection]	$(r-1)(c-1)$
(4) Test homogeneity (fixed margin) [Ex. MK test for selection draft lottery]	$(r-1)(c-1)$

22 d.f.
$\chi^2 = 32.16$

Table 7.4.

Drawing numbers	Months												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Totals
1-122	9	7	5	8	9	11	12	13	10	9	12	17	122
123-244	12	12	10	8	7	7	7	7	15	15	12	10	122
245-366	10	10	16	14	15	12	12	11	5	7	6	4	122
Totals	31	29	31	30	31	30	31	31	30	31	30	31	366

Example 7.F (Right-Handedness). To what degree is the propensity to be right-handed socially determined? Is it the same in different cultures? In different historical epochs? Two psychologists addressed this question by examining works of art that portrayed activities that could be judged as being done right- or left-handedly. (Stanley Coren and Clare Porac, "Fifty Centuries of Right-Handedness: The Historical Record" Science (1977), Vol. 198, pp. 631-632.) The following tables summarize their findings, looking at the data in two different ways.

Table 7.5. Counts of 1180 art works showing activity that can be categorized as left- or right-handed, (a) by geographical area, and (b) by historical epoch.

(a)

	Right	Left	Total	% Right
Central Europe	312	23	335	93%
Medit. Europe	300	17	317	95%
Middle East	85	4	89	96%
Africa	105	12	117	90%
Central Asia	93	8	101	92%
Far East	126	13	139	91%
Americas	72	10	82	88%
Total	1093	87	1180	92.6%

6 d.f.
$\chi^2 = 8.14$

(b)

Pre 3000 BC	35	4	39	90%
2000 BC	44	7	51	86%
1000 BC	89	10	99	90%
500 BC	134	8	142	94%
~0 BC	130	4	134	97%
AD 500	39	3	42	93%
AD 1000	57	7	64	89%
AD 1200	40	1	41	98%
AD 1400	44	6	50	88%
AD 1500	63	5	68	93%
AD 1600	68	4	72	94%
AD 1700	66	5	71	93%
AD 1800	95	6	101	94%
AD 1850	38	1	39	97%
AD 1900	71	6	77	92%
AD 1950	80	10	90	89%
Total	1093	87	1180	92.6%

15 d.f.
$\chi^2 = 17.4$

But how to interpret?

P-values

Classical Testing:

- fix level of α (say, .05)
- choose test (preferably high power)
- look at data, accept or reject α

[works well with tables, e.g. in Rice]

Ex.: H_0: Factors in 3×4 table indep.

Test: Reject if $X^2 > 12.59$ ($2 \times 3 = 6$ d.f.)

Observe $X^2 = 12.50$ accept indep.

Observe $X^2 = 12.60$ reject indep.

Are these two situations really different??

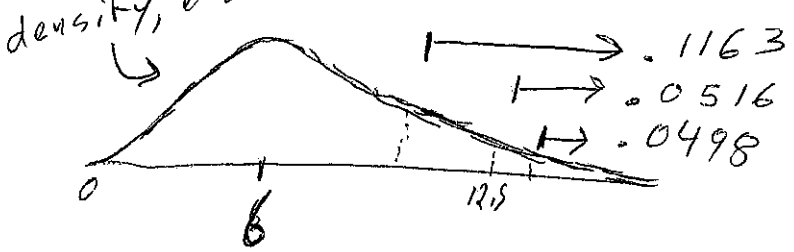
P-value: The smallest level α at which you would reject H_0

Ex: (as above)

$X^2 = 12.50 \longrightarrow P = .0516$

$X^2 = 12.60 \longrightarrow P = .0498$

$X^2 = 10.20 \longrightarrow P = .1163$



Can report P, or if only tables are available, interval: $P > .1$ or $.01 < P < .05$ or $P < .01$ or $P \ll .01$

d.f. = 22
$\chi^2 = 32.16$
p-value 2%

Table 7.4.

Drawing numbers	Months												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Totals
1-122	9	7	5	8	9	11	12	13	10	9	12	17	122
123-244	12	12	10	8	7	7	7	7	15	15	12	10	122
245-366	10	10	16	14	15	12	12	11	5	7	6	4	122
Totals	31	29	31	30	31	30	31	31	30	31	30	31	366

Example 7.F (Right-Handedness). To what degree is the propensity to be right-handed socially determined? Is it the same in different cultures? In different historical epochs? Two psychologists addressed this question by examining works of art that portrayed activities that could be judged as being done right- or left-handedly. (Stanley Coren and Clare Porac, "Fifty Centuries of Right-Handedness: The Historical Record" Science (1977), Vol. 198, pp. 631-632.) The following tables summarize their findings, looking at the data in two different ways.

Table 7.5. Counts of 1180 art works showing activity that can be categorized as left- or right-handed, (a) by geographical area, and (b) by historical epoch.

(a)

	Right	Left	Total	% Right
Central Europe	312	23	335	93%
Medit. Europe	300	17	317	95%
Middle East	85	4	89	96%
Africa	105	12	117	90%
Central Asia	93	8	101	92%
Far East	126	13	139	91%
Americas	72	10	82	88%
Total	1093	87	1180	92.6%

6 d.f.
$\chi^2 = 8.14$
$P > 10\%$

(b)

Pre 3000 BC	35	4	39	90%
2000 BC	44	7	51	86%
1000 BC	89	10	99	90%
500 BC	134	8	142	94%
~0 BC	130	4	134	97%
AD 500	39	3	42	93%
AD 1000	57	7	64	89%
AD 1200	40	1	41	98%
AD 1400	44	6	50	88%
AD 1500	63	5	68	93%
AD 1600	68	4	72	94%
AD 1700	66	5	71	93%
AD 1800	95	6	101	94%
AD 1850	38	1	39	97%
AD 1900	71	6	77	92%
AD 1950	80	10	90	89%
Total	1093	87	1180	92.6%

15 d.f.
$\chi^2 = 17.4$
$P > 10\%$

MK Test

Expected:

	Fixed	Polymorphic
Replacement	3.16	5.81
Synonymous	20.84	38.19

Observed:

	Fixed	Polymorphic	row marginals
	7	2	9
	17	42	59
	24	44	68

column marginals

$\chi^2 = 8.27$ 1 dif.

P - value 0.0042

The authors actually used a "G-test" to get a P-value of 0.006. The G-test uses an exact value of $-\log \lambda = -\sum x_i \log\left(\frac{m_i}{x_i}\right)$ rather than $-\log \lambda \approx \frac{1}{2}\chi^2$. Unfortunately the G-test still has to approximate the distribution of $-\log \lambda$ by the distribution of χ^2, so either way there is an approximation.

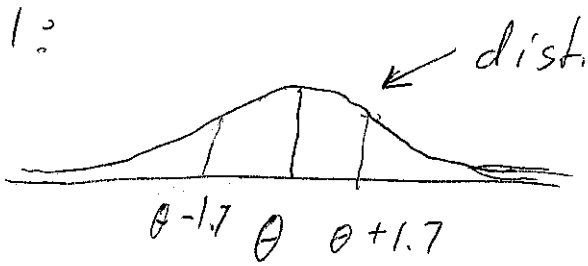
Confidence Intervals

Point estimation

$\hat{\theta}$ estimate (may be from MLE)
(accompanied by SE = "standard error"
= estimated standard deviation)

Ex: $\hat{\theta} = 14.3$, $SE = 1.7$

If normal:



Confidence intervals (C.I.)

"$[\hat{L}, \hat{U}]$, a 95% C.I."
→ an interval estimate

Ex: "The interval (11.0, 17.6) is an approx 95% CI for θ"

Questions: How to find?
What does the statement mean?
(What is "Confidence")

Finding CI's

A pivotal quantity is a function of observations and unobservable parameters whose distribution does not depend on the parameters:
eg $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Ex: (by Pivotal Method)

Data: $X_1, \ldots, X_n$

Model: X_i's indep. $N(\theta, 25)$

So MLE $\bar{X}$ is $N(\theta, \frac{25}{n})$

So $P\left(-1.96 \cdot \frac{5}{\sqrt{n}} \leq \bar{X} - \theta \leq 1.96 \cdot \frac{5}{\sqrt{n}}\right) = .95$

$P\left(-\bar{X} - 1.96\frac{5}{\sqrt{n}} \leq -\theta \leq -\bar{X} + 1.96\frac{5}{\sqrt{n}}\right) = .95$

$P\left(\bar{X} - 1.96 \cdot \frac{5}{\sqrt{n}} \leq \theta \leq \bar{X} + 1.96\frac{5}{\sqrt{n}}\right) = .95$

The interval $[\hat{L}, \hat{U}]$ is a 95% CI

$\hat{L} = \bar{X} - 1.96 \cdot \frac{5}{\sqrt{n}}$

$\hat{U} = \bar{X} + 1.96\frac{5}{\sqrt{n}}$

Interpretation: This is a random interval that includes θ with prob. .95

Interval is random

θ is not random

adopt a Bayesian approach as in Chapter 4. The virtue of confidence intervals is that they combine an assessment of accuracy with the estimate; their drawback is the propensity of the statement to be misinterpreted as a Bayesian statement when it is in fact somewhat weaker than that.

Confidence Interval Example. Eighty samples of size n = 25 were taken from an N(350, 25) distribution and the eighty 95% confidence intervals for the mean θ = 350 were computed. In this example, 5 of the 80 missed the target, about as expected.

$\bar{X}$ →

X-bar	Lower	Upper	Covers?
348.17	346.21	350.13	
351.21	349.25	353.17	
350.15	348.19	352.11	
350.69	348.73	352.65	
348.51	346.55	350.47	
350.69	348.73	352.65	
352.94	350.98	354.90	No
350.35	348.39	352.31	
349.11	347.15	351.07	
348.77	346.81	350.73	
349.88	347.92	351.84	
349.40	347.44	351.36	
349.60	347.64	351.56	
349.39	347.43	351.35	
350.82	348.86	352.78	
350.38	348.42	352.34	
349.62	347.66	351.58	
349.77	347.81	351.73	
350.02	348.06	351.98	
349.81	347.85	351.77	
349.14	347.18	351.10	
349.10	347.14	351.06	
348.47	346.51	350.43	
349.73	347.77	351.69	
348.79	346.83	350.75	
350.43	348.47	352.39	
350.65	348.69	352.61	
349.29	347.33	351.25	
349.17	347.21	351.13	
350.00	348.04	351.96	
349.97	348.01	351.93	
349.60	347.64	351.56	
351.41	349.45	353.37	
350.86	348.90	352.82	
351.28	349.32	353.24	
351.14	349.18	353.10	
349.54	347.58	351.50	
350.59	348.63	352.55	
351.58	349.62	353.54	
350.93	348.97	352.89	

X-bar	Lower	Upper	Covers?
350.44	348.48	352.40	
349.52	347.56	351.48	
347.75	345.79	349.71	No
349.10	347.14	351.06	
349.44	347.48	351.40	
348.47	346.51	350.43	
348.60	346.64	350.56	
349.37	347.41	351.33	
351.37	349.41	353.33	
350.10	348.14	352.06	
349.15	347.19	351.11	
350.97	349.01	352.93	
350.46	348.50	352.42	
350.16	348.20	352.12	
351.29	349.33	353.25	
350.37	348.41	352.33	
348.92	346.96	350.88	
349.25	347.29	351.21	
349.31	347.35	351.27	
351.23	349.27	353.19	
349.99	348.03	351.95	
350.29	348.33	352.25	
350.88	348.92	352.84	
347.41	345.45	349.37	No
349.91	347.95	351.87	
348.53	346.57	350.49	
350.03	348.07	351.99	
352.13	350.17	354.09	No
349.99	348.03	351.95	
350.81	348.85	352.77	
350.14	348.18	352.10	
350.39	348.43	352.35	
349.50	347.54	351.46	
351.29	349.33	353.25	
349.74	347.78	351.70	
351.14	349.18	353.10	
349.89	347.93	351.85	
350.80	348.84	352.76	
347.98	346.02	349.94	No
349.07	347.11	351.03	

(15)

$X_1, \ldots, X_{25}$ indep $N(350, 25)$

$\bar{X} = \frac{1}{25} \sum X_i$ $N(350, 1)$

Lower $= \bar{X} - 1.96$ Upper $= \bar{X} + 1.96$

More on the meaning of CIs

Suppose we calculate $\bar{X}$ from data and find

$$\hat{L} = 11.0 \quad \hat{U} = 17.6.$$

Does this mean

$$P(11.0 \leq \theta \leq 17.6) = .95?$$

No → only the interval is random. Can't make prob. statements about θ without a prior distribution for θ. Our confidence is in the procedure being used - it is 95% reliable, but in a given application that's all we can say.

Could have

$$P(11.0 \leq \theta \leq 17.6 \mid \bar{x}) > .95$$

or $< .95$

" $= 0$!

even

∴ Cif prior prob $P(\theta > 10) = 0$

Ex

Consider $X_1, \ldots, X_n$ iid observations from a normal distribution having unknown mean μ and known variance σ^2.

Let $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

Suppose we have a test at false positive level α that rejects if

$$|\bar{X} - \mu_0| > x_0$$

where we picked x_0 so that

if H_0 is true, $P(|\bar{X} - \mu_0| > x_0) = \alpha$

What is x_0? denote the standard deviation of $\bar{X}$ by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

This is a two sided test, so we want $x_0 = \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$, $z = \frac{\bar{X} - \mu}{\sigma}$

so the test will accept when

$$|\bar{X} - \mu_0| < \sigma_{\bar{x}} \left(\frac{\alpha}{2}\right)$$

Getting rid of the absolute values by writing positive and negative cases explicitly, we now have

$$-\sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right) < \bar{x} - \mu_0 < \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$$

$$\bar{x} - \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right) < \mu_0 < \bar{x} + \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$$

This is almost exactly the same expression we got at the end of lec 16. That was a confidence interval for the estimation of μ. Here we see that if the hypothesis test accepts H_0 at level α, the $100(1-\alpha)\%$ confidence interval for μ_0 is

$$\left[\bar{x} - \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right), \bar{x} + \sigma_{\bar{x}}\left(\frac{\alpha}{2}\right)\right]$$

So the confidence interval is precisely those values of μ_0 for which $H_0: \mu = \mu_0$ is accepted.

This is true in general.

Thm Suppose for every value θ_0 in Θ there is a test at level α of the hypothesis $H_0: \theta = \theta_0$. Denote the acceptance region of the test by $A(\theta_0)$. Then the set

$$C(\vec{x}) = \{\theta : \vec{x} \text{ in } A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

Pf $P[\vec{x} \text{ in } A(\theta_0) \mid \theta = \theta_0] = 1 - \alpha$ by definition of $A(\theta)$.

$P[\theta_0 \text{ in } C(\vec{x}) \mid \theta = \theta_0] = P[x \text{ in } A(\theta_0) \mid \theta = \theta_0]$ by definition of $C(\vec{x})$.

So, a $100(1-\alpha)\%$ confidence region for θ consists of all those values of θ_0 for which $H_0: \theta = \theta_0$ will <u>not</u> be rejected at level α.

True the other way also:

Thm

Suppose $C(\vec{x})$ is a $100(1-\alpha)\%$ confidence region for θ; in other words for every θ_0,

$$P[\theta_0 \in C(\vec{x}) \mid \theta = \theta_0] = 1 - \alpha$$

Then the acceptance region for a test at level α of the hypothesis $H_0: \theta = \theta_0$ is

$$A(\theta_0) = \{\vec{x} \mid \theta_0 \in C(\vec{x})\}$$

Pf The test has level α because

$$P(\vec{x} \in A(\theta_0) \mid \theta = \theta_0) = P(\theta_0 \text{ in } C(\vec{x}) \mid \theta = \theta_0)$$
$$= 1 - \alpha$$

That is,

$H_0: \theta = \theta_0$ is accepted if θ_0 lies in the confidence region.

STAT 249

Confidence Intervals

Lecture 17

11/29/16

recap:

Confidence Interval:

a random interval $[\hat{L}, \hat{U}]$ that includes the "state of nature" θ with some probability $1-\alpha$ (in the example last time, 95%).

recap: **Testing Composite Hypotheses**

H_0 : Group of θ_0's
H_1 : Group of θ_1's
} mutually exclusive

Likelihood Ratio Test

Reject H_0 if $\lambda < \lambda_c$ $(0 \leq \lambda \leq 1)$

$$\lambda = \frac{\max_{\theta_0\text{'s}} L(\theta)}{\max_{\text{all } \theta\text{'s}} L(\theta)}$$

$\alpha = P(\text{Reject } H_0 / H_0 \text{ true})$ "Type 1 error" "False negative"

$\beta = P(\text{Accept } H_0 / H_1 \text{ true})$ "Type 2 error" "False positive"

$\pi = 1 - \beta =$ power of the test

EX

Consider $X_1, \ldots, X_n$ iid observations from a normal distribution having unknown mean μ and known variance σ^2.

Let $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

Suppose we have a test at false positive level α that rejects if

$$|\bar{X} - \mu_0| > x_0$$

where we picked x_0 so that if H_0 is true, $P(|\bar{X} - \mu_0| > x_0) = \alpha$

What is x_0? denote the standard deviation of $\bar{X}$ by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

This is a two sided test, so we want $x_0 = \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$, $z = \frac{\bar{X} - \mu}{\sigma}$

so the test will accept when

$$|\bar{X} - \mu_0| < \sigma_{\bar{x}} \left(\frac{\alpha}{2}\right)$$

Getting rid of the absolute values by writing positive and negative cases explicitly, we now have

$$-\sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right) < \bar{X} - \mu_0 < \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$$

$$\bar{X} - \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right) < \mu_0 < \bar{X} + \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right)$$

This is almost exactly the same expression we got at the end of lec 16. That was a confidence interval for the estimation of μ. Here we see that if the hypothesis test accepts H_0 at level α, the $100(1-\alpha)\%$ confidence interval for μ_0 is

$$\left[\bar{X} - \sigma_{\bar{x}} z\left(\frac{\alpha}{2}\right), \bar{X} + \sigma_{\bar{x}}\left(\frac{\alpha}{2}\right)\right]$$

So the confidence interval is precisely those values of μ_0 for which $H_0 : \mu = \mu_0$ is accepted.

This is true in general.

Thm Suppose for every value θ_0 in Θ there is a test at level α of the hypothesis $H_0: \theta = \theta_0$. Denote the acceptance region of the test by $A(\theta_0)$. Then the set

$$C(\vec{x}) = \{\theta : \vec{x} \text{ in } A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

Pf $P[\vec{x} \text{ in } A(\theta_0) \mid \theta = \theta_0] = 1 - \alpha$ by definition of $A(\theta)$.

$P[\theta_0 \text{ in } C(\vec{x}) \mid \theta = \theta_0] = P[x \text{ in } A(\theta_0) \mid \theta = \theta_0]$ by definition of $C(\vec{x})$.

So, a $100(1-\alpha)\%$ confidence region for θ consists of all those values of θ_0 for which $H_0: \theta = \theta_0$ will <u>not</u> be rejected at level α.

True the other way also:

Thm

Suppose $C(\vec{x})$ is a $100(1-\alpha)\%$ confidence region for θ; in other words for every θ_0,

$$P[\theta_0 \in C(\vec{x}) \mid \theta = \theta_0] = 1 - \alpha$$

Then the acceptance region for a test at level α of the hypothesis $H_0: \theta = \theta_0$ is

$$A(\theta_0) = \{\vec{x} \mid \theta_0 \in C(\vec{x})\}$$

Pf The test has level α because

$$P(\vec{x} \in A(\theta_0) \mid \theta = \theta_0) = P(\theta_0 \text{ in } C(\vec{x}) \mid \theta = \theta_0)$$
$$= 1 - \alpha$$

That is,

$H_0: \theta = \theta_0$ is accepted if θ_0 lies in the confidence region.

Multinomial and Poisson

Consider a set of indep. random vars

$(X_1, \ldots, X_n)$ $\quad X \sim \text{Poisson}(\lambda)$

H_0: $\lambda_i = \lambda$ for all i (rates the same)

H_1: $\lambda_i \neq \lambda_k$ $i \neq k$ (rates different)

For an indep Poisson r.v.,

$$f(x|\theta) = f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Denote the MLE for H_0 by $\hat{\lambda} = \bar{X}$

Denote the n MLEs for H_1 by $\tilde{\lambda}_i = \tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$. $\tilde{\lambda}_1 = x_1, \ldots, \tilde{\lambda}_n = x_n$

note: we are denoting MC Poisson 'state of nature' by λ, not θ

The likelihood ratio is then (write Λ to avoid confusion):

$$\Lambda = \frac{\max_{\lambda} L(\lambda)}{\max_{\text{all } \lambda_i\text{'s}} L(\lambda)} = \frac{\prod_{i=1}^{n} \hat{\lambda}^{x_i} \frac{e^{-\hat{\lambda}}}{x_i!}}{\prod_{i=1}^{n} \tilde{\lambda}_i^{x_i} \frac{e^{-\tilde{\lambda}_i}}{x_i!}}$$

$$= \prod_{i=1}^{n} \left(\frac{\hat{\lambda}}{\tilde{\lambda}_i}\right)^{x_i} e^{\lambda_i - \lambda} = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}}$$

We had the LR

$$\Lambda = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}}$$

$$\log \Lambda = \sum_{i=1}^{n} \left(x_i \log\left(\frac{\bar{x}}{x_i}\right) + (x_i - \bar{x}) \right)$$

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

But wait! We've seen this before, Lec 14, pp 14-15. The L.R. for the multinomial dist was (write LR as Λ)

$$\log \Lambda = \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{m_i}\right)$$

x_i observed

m_i expected from multinomial dist

there, we found that

$$-2 \log \Lambda \cong \sum_{i=1}^{n} \frac{(x_i - m_i)^2}{m_i} = \chi^2$$

with Poisson,

$$-2 \log \Lambda \cong \frac{1}{\bar{x}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \chi^2$$

under H_0, there is one parameter $\hat{\lambda}$, so d.f. $= n - 1$, just like multinomial.

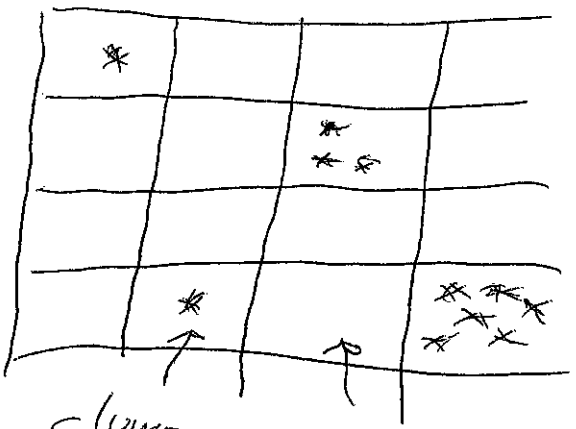
Now, the estimated variance

$$s^2 = \frac{1}{(n-1)} \sum (x_i - \bar{x})^2,$$

so $$-2 \log \Lambda \approx \frac{1}{\bar{x}} \sum_{i=1}^{n} (x_i - \bar{x})^2 \propto \frac{\text{var}}{\text{mean}}$$

For Poisson r.v.'s, $\sigma^2 = \mu$, so deviations from this ratio are being tested.

Ex clumps of bacteria
0.01 ml milk spread out on slide with grid:



There aren't very many bacteria in the milk (fortunately!), and one often is told that Poisson statistics are useful for "rare events" (e.g. death by Horsekick).

So let us consider some actual data (Bliss and Fisher, Biometrics 9: 174-200)

The clump data: (from 400 squares)

# bacteria/sq.	0	1	2	3	4	5	6	7	8	9	10	19
Frequency	56	104	80	62	42	27	9	9	5	3	2	1

$$\hat{\lambda} = \bar{x} = \frac{0(56) + 1(104) + 2(80) + \cdots + 19(1)}{400}$$

$$= 2.44$$

Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
χ^2 contrib	12.8	4.2	5.5	5.9	1.8	.14	.14	45.0

$\chi^2_6 = 75.4$, d.f. $= 8 - 1 - 1 = 6$

(8 cells; 1 for Poisson or multinomial; 1 parameter estimated)

6 d.f

p-value $\ll .005$ ... reject!

In general, χ^2 is a one-sided test - reject H_0 if $\chi^2 > c$

But small χ^2 is sometimes informative, too.

Mendel's Peas (test for indep, multinomial dist)

We expect:

	Smooth	Wrinkled
yellow	$\frac{9}{16}$	$\frac{3}{16}$
green	$\frac{3}{16}$	$\frac{1}{16}$

For a particular experiment, Mendel reported (compare to observed, calculated from above table)

Type	Observed	Expected
smooth yellow	315	312.75
smooth green	108	104.25
wrinkled yellow	102	104.25
wrinkled green	31	34.75

mmm... looks like good agreement!

d.f. = 3. $X^2 = 0.604$ (G test gives 0.618)

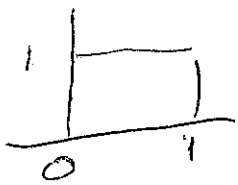
p-value close to 0.9. Can't reject. But would expect agreement worse than this 90% of the time... means one should look at Mendel's other experiments.

To do this, we need a way of combining results of many experiments ("meta analysis"), and we'll need to (somehow), combine their p-values.

We'll need (also see Lec. 3, p. 15)

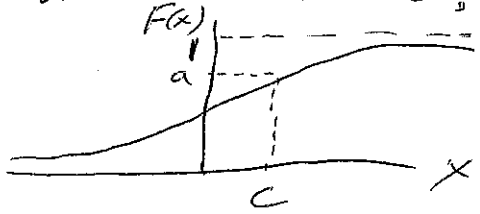
Thm X a continuous rand. var. with c.d.f. $F(x)$.

$Y = F(X)$ (transform of X).



Then $Y \sim$ Uniform $[0,1]$

Pf $F(x)$ continuous and monotone:



$$P(Y \leq y) = P(F(X) \leq y)$$
$$= P(X \leq F^{-1}(y))$$
$$= F(F^{-1}(y)) = y$$

$a = F(c)$, $c = F^{-1}(a)$

hence density of Y is $= \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Corollary: Test H_0 with test statistic T, where T has a cont. dist., with c.d.f. F_0 under H_0, and we reject for small T. Then if $T = t$ is observed, $F_0(t) =$ P-value, smallest α for which you reject H_0

Then $P = F_0(T)$ is the P-value and has a uniform (0,1) dist under H_0.

[Family of tests: Reject if $T \le c$, small c, small α. Smallest = T]

Combining Indep Tests

k indep tests of some H_0

P-values $P_1, P_2, \ldots, P_k$

Do these, "combined", provide evidence to reject H_0? What is the "combined" P-value?

One idea: $P = P_1 P_2 P_3 \ldots P_k$ (product)

not right - way too small ($\to 0$ as $k \to \infty$) and indep tests of the same hypothesis not completely indep.

But look at $-2 \log P$. Claim:

This has a χ^2_{2k} dist under H_0 (if dists continuous).

Remember that χ^2_k has pdf

$$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

$$-2 \log P = \sum_1^n -2 \log P_i. \quad \text{Let } Y_i = -2 \log P_i$$

Under H_0, $P(Y \le y) = P(-2 \log P_i \le y)$

$$= P(P_i \ge e^{-y/2})$$

$$= 1 - e^{-y/2}$$

So under H_0, Y is exponential density $\frac{1}{2} e^{-y/2}$

↳ see above χ^2_2 is exponential

!!

So

$$-2 \log P = \chi^2_2 + \chi^2_2 + \cdots + \chi^2_2$$

$$= (u_1^2 + u_2^2) + (u_3^2 + u_4^2) + \cdots$$

dist χ^2_{2k}

P small, $-2 \log P$ large

Examples:

① P_i's: 0.3, 0.4, .5, .6, .7

$P = \prod_{i=1}^{5} P_i = 0.025$

$-2 \log P = 7.36$

$k = 5$

$2k = 10$ d.f.

(P-value = .69)

② P_i's: 0.1, 0.15, 0.08, 0.2, 0.07

$P = .000017$

$-2 \log P = 21.99$

"Fisher's Method of Combination"

(meta analysis)

Done on Mendel's experiments, Fisher got P-value of 0.99996. (!?!)

Mendel's laws are correct, but Mendel's variance is awfully small. Probably he kept the best results, not knowing principles of statistics, not yet invented when he did his work.

Topics

249 AvQ Topic Summary

1-2: Permutations and combinations
Conditional Probability
Random Variables + Distributions (discrete)
Bernoulli
Binomial
Negative Binomial
Poisson

3. Continuous Random Variables
Functions of Random Variables
Discrete
Continuous
$Y = X^2$, $X \sim N, \rightarrow Y \sim \chi^2_1$
$Y = aX + b \rightarrow$ std normal

4. Describing Distributions

pmf/cdf
median
Expectations
expectation of function of r.v.
mean
linear transformations thereof
variance
linear transformations thereof

Multivariate Distributions
Discrete
continuous
Marginal Dists
Conditional Dists

5. Expectations of Joint Distributions
   Expectations of Marginal Dists
   Covariance
   Law of Large Numbers
   Moments and Moment Generating Functions
   CLT derived (not proved) with MGFs
6. INFERENCE
   Bayesian Inference
   Bayes' Thm
      Discrete Dists
      Continuous Dists
      Mixed
   Beta Distribution
      Gamma Function
   Meaning of the Prior
7. Expectation of Beta Dist
   Bayesian Inference on Conjugate Priors
      Beta
      Normal
      Weighted average of prior + posterior
8. Likelihood
   Point estimation: $\hat{\theta}(x)$ from $f(x/\theta)$
   $MSE = Var + (Bias)^2$
   Maximum Likelihood $L(\theta) = f(x/\theta)$
      Find by setting $\frac{d}{d\theta} \log L(\theta) = 0$
      check " verifying $\frac{d^2}{d\theta^2} \log L(\theta) < 0$
9. Distributions of Sums
   χ^2 density function derivation
10. Fisher's Thm
    If MLE found by setting $\frac{d}{d\theta} \log L(\theta) = 0$,
    $\hat{\theta}$ normally dist.

$$\frac{1}{\tau^2} = E\left[\frac{d}{d\theta} \log f(x/\theta)\right]^2 = -E\left[\frac{d^2}{d\theta^2} \log f(x/\theta)\right]$$

10. (cont) Fisher's Thm (cont)

Cramer-Rao inequality

Case where $L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$, $\gamma_u = \frac{\gamma}{n}$

Fisher doesn't apply if max at edge of domain

11. Fisher's Thm for multivariate case

Although MLE's are often biased,

$\bar{X}$ unbiased for μ, s^2 for σ^2

Non-rigorous proof of Fisher's Thm

11A. Sufficient Statistics and Neyman Factorization Thm; Rao-Blackwell Thm

12. Hypothesis Testing

Simple Hypotheses: α, β, $\pi = 1 - \beta$

Likelihood Ratios, which are best tests

Neyman-Pearson Lemma

Power Functions

Uniformly most Powerful Tests

13. Proof of Neyman-Pearson

LR

$$\lambda = \frac{\max_{\theta_0's} L(\hat{\theta}_0)}{\max_{\text{all } \theta_i's} L(\hat{\theta}_i)}$$

Reject H_0 if $\lambda < \lambda_c$

$P(\lambda < \lambda_c | H_0) \leq \alpha$

The Multinomial Distribution

14. Composite Hypotheses

Reject H_0 if $\lambda < \lambda_c \Rightarrow -\log \lambda > C$

χ^2

$-\log \lambda \cong \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \chi^2$, dist chi-square

Multinomial, k outcomes, χ^2_{k-1}
(test 'fairness')

Multinomial, k outcomes, p $\hat{\theta}$'s by MLE, χ^2_{k-p-1}
(test form of distribution)

Contingency Tables - $r-1$, $c-1$

Depend on row and column totals,

Multinomial $(r-1)(c-1)$ d.f.

H_0: cell probs product of marginal probs

15. Tests of Homogeneity

Multinomial X_{ij} trials

H_0: probs the same all rows

Same χ^2, same d.f. as other multinomial tests

χ^2 an approx.

all expected > 3.5 (?)
80% " > 5 (?)
etc.
No expected $= 0$!

} opinion differs on exact borders

P-values

smallest α to reject H_0

Confidence Intervals

random interval with $1-\alpha$ prob of containing θ

16. dual to hypothesis tests

ASI HOMEWORK

questions about these topics easier

More χ^2:

Multinomial χ^2 for Poisson

Small χ^2 and Mendel's peas

Meta-analysis by Fisher's method of combination of statistics

[illegible] test, [illegible] function.