Lecture 16:

$\underbrace{\hat{\hat{\theta}}_{1.}}_{HB} = \frac{n_{1.}}{n}$ proportion of people with black hair.

$\Rightarrow \hat{\hat{\theta}}_{i.} = \frac{n_{i.}}{n}$, $\hat{\hat{\theta}}_{.j} = \frac{n_{.j}}{n}$

$$\Rightarrow \hat{\phi} = \sum_{i=1}^{r} \sum_{J=1}^{c} \frac{\left(\theta_{i.j} - n\,\hat{\theta}_{i.}\,\hat{\theta}_{.j}\right)}{\underbrace{n\,\hat{\theta}_{i.}\,\hat{\theta}_{.j}}_{E_{ij}}} \xrightarrow{d} \chi_{(r-1)(c-1)}$$

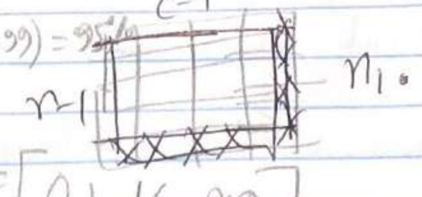
a fact that could be proved in 368 if time permitted

" Chi-squared test of independence "

$\hat{\phi} = 41.20$ $\alpha = 5\%$. $F_{\chi^2_9} = (16.99) = 95\%$

$\Downarrow$

RET$= [0, 16.99]$



$\Rightarrow$ Reject H_0

We conclude that haire color and eye colore are dependent.

Thie class has focused mainly on three goals of inference : estimation, testing and confidence sets. we will continue to study these three goals deut, ... we will also study some tangential "meta" concept that are classic:

Here is one such classic "meta concept" . usually you are

(a) given a dataset $x_1, .. x_n$ then you

(b) assume a DGP, then you

(c) define one or (many) inferential target parameters theta, the

(d) Compute $\hat{\theta} = \omega_1 (x_1 .. x_n)$ and then you

(e) make a CI / run a test at size α.

let examine (b). How do you just "assume a DGP"? Sometimes you really do know the DGP e.g a coin flipped repeatedly is iid Bernoulli (θ), a die rolled repeatedly is iid uniform discrete. But what about "daily wind speeds at JFK airport" or "rat survival times" (like on midterm) or "daily percentage returns of the S&P 500"? The DGPS for the last three are very complicated and they're unknown.

What if we wanted to "guess" the DGP's? This is actually a really big Part of what

Statiscians do. Thisis called "model fitting". DGP = model. Model you kinda make up and hopefully they are useful fore whateverr you are doing. why don't we proceed as follows:

let's guess M candidate models/ DGPs $m = 1, 2, .., M$ and then
(1) pick the "best" model out of my M guesses and maybe

(2) provide a weighting score to each of the M guesses (low scores indicate bad guesses and high scores indicate good guesses). Goal (1) is famous and called "Model selection". In 342, we do a

little of this atheoretically. Here we'll do it more theoretically.

Model Selection is more fundamental than You realize. It's actually the entire problem of all of science. For example, let's say you have some astronomical data on movement of different planets, Stars, etc. You want to fit a model (guess a DGP) for the force on celestial bodies with masses m_1 and m_2 at a distance r from eachother (i.e "gravity"). Consider the following models,

Model 1: $F = G\frac{m_1 m_2}{r^2}$ Newton's Law

Model 2: $F = G_1\frac{m_1 m_2}{r^2} + G_2\frac{m_1 m_2}{r^2}$ Newton's extension

Model 3: $F = G_1 \frac{m_1 m_2}{r^2} e^{-G_2 r}$ Laplace extension

Which model is best? We know all these are wrong because Einstein Came and disprove them with general relativity.

---

lets talk about Model Selection techniques. Our data $x_1 \dots x_n$ Comes from an Unknown DGP. Here are M candidate models:

Model 1 : $f_1(X_1 \dots X_n ; \theta_{11}, \theta_{1K}) = \mathcal{L}(\theta_{11}, \dots, \theta_{1K_1}, X_1 \dots X_n)$

model 2: $f_2(X_1 \dots X_n ; \theta_{21} \dots \theta_{2K_2}) = \mathcal{L}(\theta_{21} \dots, \theta_{2K_2}, X_1 \dots X_n)$

$\vdots$

Model M: $f_m(X_1 \dots X_n ; \theta_{m1} \dots \theta_{mK_m}) = \mathcal{L}(\theta_{m1} \dots \theta_{mK_m}, X_1 \dots X_n)$

K_1 is the # of Parameters in model1,

K_2 is the num of parameter in model 2, K_m is the num of parameter into model M. Each K_m could be different.

Why don't we just select the model that has the highest likelihood?

$$m_\ell := \underset{m}{\arg\max} \left\{ \mathcal{L}_M(\theta_{M_1} \quad \theta_{M_{K_m}} ; x_1 \ldots x_n) \right\}$$

$$= \underset{m}{\arg\max} \left\{ \ell_m(\theta_{M_1} \ldots \theta_{M K_m} ; x_1 \quad x_n) \right\}$$

The problem withe this is we don't know the values of θ for any of the models!

So let's do rechardize $K_1 + K_2 +$ K_m times!

We'll estimate each of the parameters

using MLE's and plug them all in and then

$$m_\alpha = \arg\max \left\{ l\left(\hat{\theta}^{MLE}_{M_1}, \dots \hat{\theta}^{MLE}_{mk_m}\ x_1 .. x_n\right) \right\}$$

You could do this. But... it will not give you the best model. why?

$l(\hat{\theta}^{MLE}_{m_1}, \dots \hat{\theta}^{MLE}_{mk_m}; x_1 .. x_n)$ is an estimator for $l(\theta_{m_1}, \dots \theta_{mk_m}; x_1 .. x_n)$

and it's biased... with many assumptions, you can prove that

$$\text{Bias}\left[l(\hat{\theta}^{MLE}_{m}, \dots \hat{\theta}^{MLE}_{mk_m}; x_1 .. x_n) \right] = K_m > 0$$

There is positive Bias (meaning the log-likelihood would appear higher on average) and this bias increases You use more parameteres in your Candidate models. This was figured out by H. Akaike, a Japanese Statistian and he published it in 1974.

Once you have the bias, you Just use it to Correct your estimate:

$$\ell(\theta_{m_1}, \dots \theta_{mk_m}) \approx \ell(\hat{\theta}^{MLE}_{m_1}, \dots \hat{\theta}^{MLE}_{mk_m}; x_1 \dots x_n) - K_m$$

Recall that log-likelihood is always negative for discrete DGP'S and alomost always negative for continuous DGP'S So let's flip it's sign and multiply by 2:

(Alkaime

$$AIC_m = -2\ell(\hat{\theta}^{MLE}_{m_1,\dots} \hat{\theta}^{MLE}_{mk_m}, x_1 \dots x_n) + \overbrace{2K_m}^{\text{Complexity plunalty}}$$

(Akaike's Information Criterion)

The "best" log likelihood is the largest i.e closest to zero

So once negated, the "best" negative log likelihood is the smallest i.e closest to zero.

$$m_\alpha = \text{argmax}($$