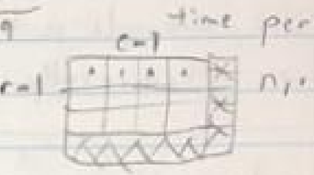
Lecture 16 11/2/20

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

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

$\Rightarrow \hat{\phi} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j})^2}{\underbrace{n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}}_{E_{ij}}} \xrightarrow{d} \chi^2_{\underbrace{\underbrace{(r-1)}_{3}\underbrace{(c-1)}_{3}}_{9}}$ a fact that could be proven in 368 if time permitted.

"Chi-squared test of independence"



$\hat{\hat{\phi}} = 41.28$ $\alpha = 5\%$ $F_{\chi^2_9}(16.92) = 95\%$

$\notin$ RET $\Downarrow$

$\Rightarrow$ Reject H_0 RET $= [0, 16.92]$

we conclude that hair color and eye color are dependent.

This class has focused mainly on the 3 goals of inference: estimation, testing and confidence sets. We will continue to study these 3 goals but... we will also study some tangential "meta" concepts that are classic.

Here's one such classic "meta concept." Usually you are

a) given a dataset $x_1, \ldots, x_n$, then you

b) assume a DGP, then you

c) define one (or many) inferential target parameters θ, then

d) compute $\hat{\hat{\theta}} = w(x_1, \ldots, x_n)$ and then you

e) make a CI / run a test at size α

Let's examine (b). How do you just "assume a DGP"? Sometimes you really do know the DGP e.g. a coin flipped repeatedly is iid Bern(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 DGP's 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 statisticians do. This is called "model fitting." DGP = model. Models you kinda make up and hopefully they're useful for whatever you're doing. Why don't we proceed as follows: let's guess M candidate models / DGPs $m = 1, 2, \ldots, M$ 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 two celestial bodies with masses m_1, m_2 at a distance r from each other (i.e. "gravity").

Consider the following models:

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

Mod 2: $F = G_1 \frac{m_1 m_2}{r^2} + G_2 \frac{m_1 m_2}{r^3}$ — Newton's Extension

Mod 3: $F = G_1 \frac{m_1 m_2}{r^2} e^{-G_2 r}$ — Laplace's Extension

which model is the best? We know all these are wrong because Einstein came and disproved them with general relativity.

Let's talk about model selection techniques. Our data $x_1, \ldots, x_n$ comes from an unknown DGP. Here are M candidate models:

$$\text{Model } 1: f_1(x_1, \ldots, x_n; \theta_{11}, \ldots, \theta_{1K_1}) = \mathcal{L}_1(\theta_{11}, \ldots, \theta_{1K_1}; x_1, \ldots, x_n)$$

$$\text{Model } 2: f_2(x_1, \ldots, x_n; \theta_{21}, \ldots, \theta_{2K_2}) = \mathcal{L}_2(\theta_{21}, \ldots, \theta_{2K_2}; x_1, \ldots, x_n)$$

$$\vdots$$

$$\text{Model } m: f_m(x_1, \ldots, x_n; \theta_{M1}, \ldots, \theta_{MK_M}) = \mathcal{L}_M(\theta_{M1}, \ldots, \theta_{MK_M}; x_1, \ldots x_n)$$

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

$$m_* := \underset{m}{\text{argmax}} \left\{ \mathcal{L}_m (\theta_{m_1}, \ldots, \theta_{mK_m} ; x_1, \ldots, x_n) \right\}$$

$$= \underset{m}{\text{argmax}} \left\{ \ell_m (\theta_{m_1}, \ldots, \theta_{mK_m} ; x_1, \ldots, x_n) \right\}$$

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

So let's nicheardize $K_1 + K_2 + \ldots + K_M$ times! We'll estimate each of the parameters using MLE's and plug them all in and then...

$$m_* = \text{argmax} \left\{ \ell \left(\hat{\theta}_{m_1}^{MLE}, \ldots, \hat{\theta}_{mK_m}^{MLE} ; x_1, \ldots, x_n \right) \right\}$$

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

$\ell \left(\hat{\theta}_{m_1}^{MLE}, \ldots, \hat{\theta}_{mK_m}^{MLE} ; x_1, \ldots, x_n \right)$ is an estimator for $\ell (\theta_{m_1}, \ldots, \theta_{mK_m} ; x_1, \ldots, x_n)$

and it's biased... With many assumptions, you can prove that asymptotically ... meaning as n gets large...

$\text{Bias} \left[\ell \left(\hat{\theta}_{m_1}^{MLE}, \ldots, \hat{\theta}_{mK_m}^{MLE} ; x_1, \ldots, x_n \right) \right] = K_m > 0$

There is positive bias (meaning that the log-likelihood would appear higher on average) and this bias increases you use more parameters in your candidate models. This was figured out by H. Akaike, a Japanese statistician, and he published it in 1974. Once you have the bias, you just use it to correct your estimate:

$$\ell (\theta_{m_1}, \ldots, \theta_{mK_m}) \approx \ell \left(\hat{\hat{\theta}}_{m_1}^{MLE}, \ldots, \hat{\hat{\theta}}_{mK_m}^{MLE} ; x_1, \ldots, x_n \right) - K_m$$

Recall that log likelihood is always negative for discrete DGP's and almost always negative for continuous DGP's. So let's flip it's sign and multiply by two:

$$\underbrace{AIC_m}_{\text{(Akaike's Information Criterion)}} := -2\ell \left(\hat{\hat{\theta}}_{m_1}^{MLE}, \ldots, \hat{\hat{\theta}}_{mK_m}^{MLE} ; x_1, \ldots, x_n \right) + \overbrace{2K_m}^{\text{complexity penalty}}$$

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.