$AIC_1, AIC_2, \ldots, AIC_M$. Let $AIC_* = \min\{AIC_m\}$ (best one)

let $w_m = \dfrac{e^{-(AIC_m - AIC_*)/2}}{\sum_{j=1}^{M} e^{-(AIC_j - AIC_*)/2}}$ Akaike weights (they sum to 1)

$$e^{-(AIC_m - AIC_*)/2} = e^{\frac{-(-2\ell_m + 2K_m) - (-2\ell_* + 2K_*)}{2}}$$

$$= e^{(\ell_m - \ell_*) + (K_m - K_*)} = \frac{e^{\ell_m}}{e^{\ell_*}} \frac{e^{K_m}}{e^{K_*}} = \frac{\mathcal{L}_m}{\mathcal{L}_*} \frac{e^{K_m}}{e^{K_*}}$$

kind of like a ratio of probabilities

If the "true model / DGP" is one of the candidate models, then w_m is the probability that model m is the true model.

Beyond the scope of the class, people use Akaike weights to create a model which is an average over the candidates:

$$f = \sum_{m=1}^{M} w_m f(x_1, \ldots x_n; \hat{\theta}_{m1}^{MLE}, \ldots \hat{\theta}_{mK_m}^{MLE})$$ mixture model

It turns out that in low n situations, the bias is very incorrect, so there's a correction term that is employed to fix the bias and make the AIC more accurate, it's called "AIC-Corrected" or AICC:

$$AICC_m := -2\ell_m(\hat{\theta}_{m1}^{MLE}, \ldots \hat{\theta}_{mK_m}^{MLE}; x_1, \ldots x_n) + 2K_m \left(\frac{n}{n - K_m - 1}\right)$$

inflates the penalty

Midterm II ↑

Final ↓

The concept of "practical significance" (or "clinical significance" if you happen to be in a medical / health context). Let's say you're testing

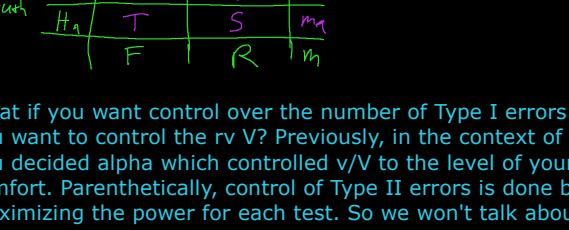
$H_a: \theta \neq \theta_0$ vs $H_0: \theta = \theta_0$

but the true value of the parameter is $\theta_0 + \varepsilon$ where ε is a small number. So H_0 is technically false. Now you won't be able to reject H_0 unless your n is very high because power to find small effects is low. But... given enough n, you always reject for any epsilon and any alpha.

Proof: assume thetahat is asymptotically normal and epsilon is positive (for HW you'll prove it for negative). This means:

$$\hat{\theta} | H_0 \sim N(\theta_0, SE[\hat{\theta}](\theta_0)^2) = N(\theta_0, (\tfrac{\sigma_0/\sqrt{n}}{\sqrt{n}})^2)$$

$$\hat{\theta} \sim N(\theta_0 + \varepsilon, SE[\hat{\theta}](\theta_0+\varepsilon)^2) = N(\theta_0 + \varepsilon, (\tfrac{\sigma/\sqrt{n}}{\sqrt{n}})^2)$$



$$\text{Pow} = P(\text{Reject } H_0) = P\left(\hat{\theta} > \theta_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right) = P\left(\frac{\hat{\theta} - (\theta_0 + \varepsilon)}{\frac{\sigma}{\sqrt{n}}} > \frac{\theta_0 + z\frac{\sigma_0}{\sqrt{n}} - (\theta_0 + \varepsilon)}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(Z > \frac{-\varepsilon + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z > \underbrace{-\frac{\varepsilon}{\sigma}\sqrt{n}}_{a} + \underbrace{z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sigma}}_{b}\right)$$

$$\lim_{n\to\infty} \text{Pow} = \lim_{n\to\infty} = P(Z > -\infty) = 1 \Rightarrow$$ your estimate is "statistically significant"

$p_{val} < \alpha$

So one can argue that in the real world, theta is never exactly some value you propose. Take the case of a coin. You want to prove it's unfair. H_0: theta = 50% exactly. But take a look at this coin... the real theta is probably 49.9999% so epsilon = 0.0001%. But.. is a coin with Prob(heads) = 49.9999% actually "unfair"? NO. Practically this coin is fair (for all practical uses).

Thus, if you flip enough times, you will get a "statistically signicant" estimate that has "no practical significance".

No amount of math can tell you what "practical signficance" means. You have to define it yourself based on your own context and your own objective.

As an example. Let's say you're testing a weight loss pill so you randomly give n_T subjects the pill (T = treatment) and randomly give n_C subjects the placebo (C = control) and then you measure

$\bar{X}_T - \bar{X}_C$ and run the a test H_0: theta_T = theta_C (no mean difference between pill group and control group). Where theta_T is mean weight loss in the pill group and theta_C is mean weight loss in control group.

You get a pvalue of 0.001 < 5% => reject and you have "statistical significance. But.... xbar_T - xbar_C = 0.1 pounds. This is not "clinically significant" (that's our feeling).

Next "meta concept" that I will cover is called the "multiple testing problem" or "multiple comparisons problem". Recall the possible decisions / outcomes from a hypothesis test:

Truth \ Decision	Retain H_0	Reject H_0
H_0	Justified retainment	Type I error
H_a	Type II error	Justified rejection

we "control" the probability of Type I errors by setting it to be at most...

$P(\text{Type I error}) \leq \alpha$

Let's say you're doing m hypothesis tests (many of them) each with alpha (controlled Type I error probability). This collection of tests is called a "family of tests". Among these tests, you reject r of them and retain f of them so that r + f = m. But... unbeknownst to you, you could've made some Type I or Type II errors. Here's a contingency table with the number of each possibility:

Truth \ Decision	Retain H_0	Reject H_0	
H_0	u	v	m_0
H_a	t	s	m_a
	f	r	m

observe / constants

v is the number of Type I errors AKA "false rejections" AKA "false discoveries".

you don't know any of these values

Furthermore, which quantities are random? And the randomness is due to the DGP. The ones with capital letters below:

Truth \ Decision	Retain H_0	Reject H_0	
H_0	U	V	m_0
H_a	T	S	m_a
	F	R	m

What if you want control over the number of Type I errors v i.e. you want to control the rv V? Previously, in the context of m=1, you decided alpha which controlled v/V to the level of your comfort. Parenthetically, control of Type II errors is done by maximizing the power for each test. So we won't talk about it.

Why do we care about controlling the Type I errors? Here's an example that should get you scared. Let's say you have m independent hypothesis tests each with size alpha = 5%. Also, let m = m_0 = 30 i.e. all H_0's are true. By chance alone,

R ~ Bin(m, alpha)

I'm interested in the chance I make at least one Type I error. By the simple 241 calculations,

P(R > 0) = 1 - P(R = 0) = 1 - P(all retain) = 1 - (1 - alpha)^m

$$= 1 - (1 - 5\%)^{30} \approx 76\%.$$

So there's a huge chance you make at least one "false discovery". Maybe this 76% probability is too high for you.