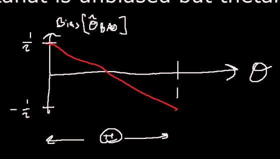
$X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$. You seek point estimation for θ.

$\hat{\theta} = \frac{1}{n} \sum X_i, \quad \hat{\theta}_{BAD} = \frac{1}{2}$

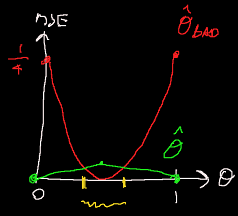
Common sense should tell you guessing one single value all the time independent of your data is a *bad* idea. Why is thethat better?
* Thetahat is unbiased but thetahatbad is biased:



What about risk (expected loss)? Under squared error loss, risk is the same thing as mean squared error (MSE). So let's look at MSE:

$$MSE[\hat{\theta}] \overset{\text{lec 2}}{=} \cancel{\text{Bias}[\hat{\theta}]^2}^{\,0} + \text{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n} \qquad n = 10$$

$$MSE[\hat{\theta}_{BAD}] = \text{Bias}[\hat{\theta}_{BAD}]^2 + \cancel{\text{Var}[\hat{\theta}_{BAD}]}^{\,0} = \left(\frac{1}{2} - \theta\right)^2$$



* As n increases, MSE decreases for thetahat everywhere but as n increases, the MSE of thetahatbad does not decrease.

There is a small region of theta where thethatbad is more accurate than thetahat. This region decreases with n. Estimators cannot be expected to be the best everywhere (for all theta). There are compromises.

How do I compare two MSE curves? There are many ways. One way is called "supremum risk" (max risk).

$$\sup_{\theta \in \Theta} \{ MSE[\hat{\theta}] \} = \frac{1}{4n} \quad \text{vs.} \quad \sup_{\theta \in \Theta} \{ MSE[\hat{\theta}_{BAD}] \} = \frac{1}{4}$$

* For any n, thetahat beats thetahatbad in max risk.

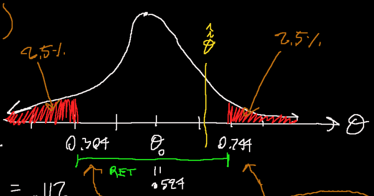
Let's return to hypothesis testing for θ in the iid Bern(θ) setting. Using the CLT, we have the following approximate distribution:

$$\hat{\theta} \mid H_0 \,\dot{\sim}\, N\left(\theta_0, \left(\sqrt{\frac{\theta_0(1-\theta_0)}{n}}\right)^2\right) = N(.524, .112^2)$$

let $\alpha = 5\%$ ($n = 20$, lec 1)

$\rightarrow H_0: \theta = \theta_0 = .524$

$H_a: \theta \neq \theta_0 = .524$



$$SE[\hat{\theta} \mid H_0] = \sqrt{\frac{.524(1-.524)}{20}} = .112$$

How do we construct RET / REJ regions? We need

$$F^{-1}_{\hat{\theta}|H_0}\left(\frac{\alpha}{2}\right) \qquad F^{-1}_{\hat{\theta}|H_0}\left(1 - \frac{\alpha}{2}\right) = +1.96\,SE + \theta_0 = .744$$

$$\underset{2.5\%}{\parallel} \qquad\qquad \underset{97.5\%}{\parallel}$$

$$= \hat{\hat{\theta}} \text{ s.t. } P(\hat{\theta} \mid H_0 < \hat{\hat{\theta}}) = 2.5\% \qquad X \sim N(\mu, \sigma^2)$$

$$= P\left(\underbrace{\frac{\hat{\theta} \mid H_0 - \theta_0}{SE[\hat{\theta} \mid H_0]}}_{} < \frac{\hat{\hat{\theta}} - \theta_0}{SE[\hat{\theta} \mid H_0]}\right) = 2.5\% \qquad Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

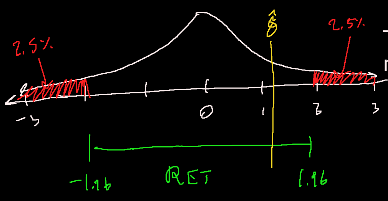
standard normal i.e. N(0,1)

$$= P(Z < z) = 2.5\% \overset{\text{calc.}}{\Longrightarrow} z = -1.96$$

$$\Rightarrow -1.96 = \frac{\hat{\hat{\theta}} - \theta_0}{SE} \Rightarrow \hat{\hat{\theta}} = -1.96\,SE + \theta_0 = -1.96 \cdot .112 + .524 = .304$$

If our data estimate thetahatatahat ∈ RET = [.304, .774] => Retain H0
∈ REJ = [0, .304) ∪ (.774, 1] => Reject H0

Traditionally, this test is done on the Z scale (not the thetahat scale) because Z and thetahat are 1:1. For example:



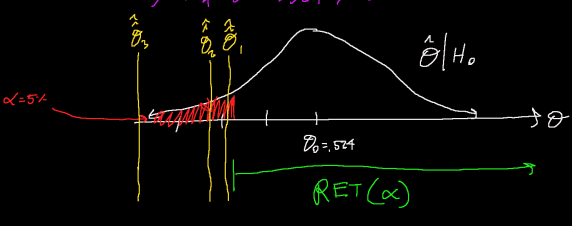
This thetahat under the null standardized i.e. N(0,1)

$$z = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta} \mid H_0]} = \frac{.65 - .524}{.112} = 1.125 \in RET \Rightarrow \text{Retain } H_0.$$

Fisher's p-values.

To illustrate this concept, we'll use the iid Bern(θ) DGP and

$H_0: \theta \geq .524, \quad H_a: \theta < .524, \quad \alpha = 5\%$

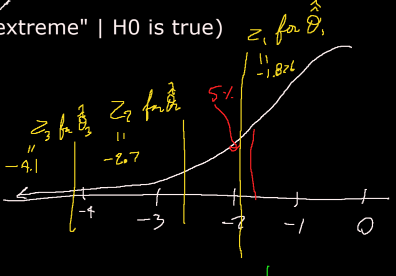


$z = -1.645$

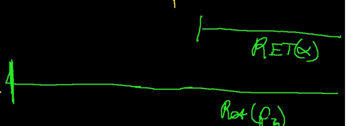
As thetahathat decreases, you still "only" reject. But shouldn't that rejection become a "stronger rejection"? Fisher was bothered by the same thing so he defined his "p-value" as:

$$p_{val} := \max\{\alpha : \hat{\hat{\theta}} \in RET\}$$

= P(thetahathat is "more extreme" | H0 is true)



The p_1 (pval for thetahathat1) is



$p_1 = P(Z < z_1) = 3.03\%$ via calc.

$p_2 = P(Z < z_2) = 0.34\%$ via calc

$p_3 = P(Z < z_3) = 2 \times 10^{-5}$ via calc

⇒ the p-value allows you to compare rejections. The third estimate provides the strongest evidence against the null since it's the smallest p-val.

Note: if the p-val is less than alpha => Reject. This wasn't the interesting innovation of Fisher's p-value.

Type II errors and statistical power

Assume our iid Bern(θ) DGP and consider these two hypotheses just as a hypothetical. This isn't a test you would ever really run.

$H_0: \theta = .524 = \theta_0$

$H_a: \theta = .716 = \theta_a$

$H_0 \cup H_1 \neq \Theta$ Weird!

$\alpha = 5\%$, $n = 20$