Recitation Two

PART ONE: Testing

Basics of z-tests

We have a τ-stat that we believe is approximately normal under the true θ_0. For now, I will assume

if $\tau = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} \xrightarrow{d} N(0,1)$ if $\theta_0 = \mathbb{E}[\hat{\theta}]$ truly.

More specifically, say $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\tau = \frac{\bar{X} - \mu_0}{SE(\bar{X})} \xrightarrow{d} N(0,1)$$

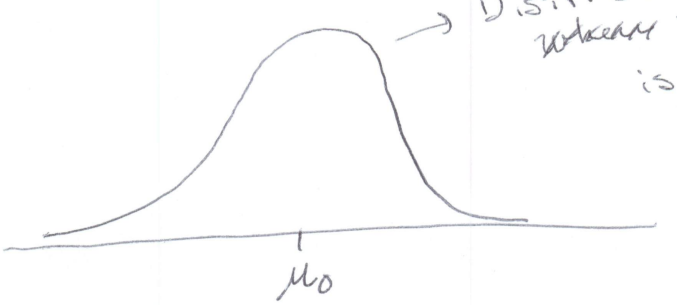
OR

$$\frac{\bar{X} - \mu_0}{SE(X_i)/\sqrt{n}} \sim N(0,1) \quad \text{if } n \text{ large.}$$

Then under the assumption that

$\mathbb{E}[X_i] = \mu_0$, $\dfrac{\bar{X} - \mu_0}{SE(X_i)/\sqrt{n}} \sim N(0,1)$

or $\bar{X} \sim N\left(\mu_0, \dfrac{VAR(X_i)}{n}\right)$



Distribution of $\bar{X}$ assuming null is true.

Variance of this distribution depends on $VAR(X_i)$ and n.

Our question is:

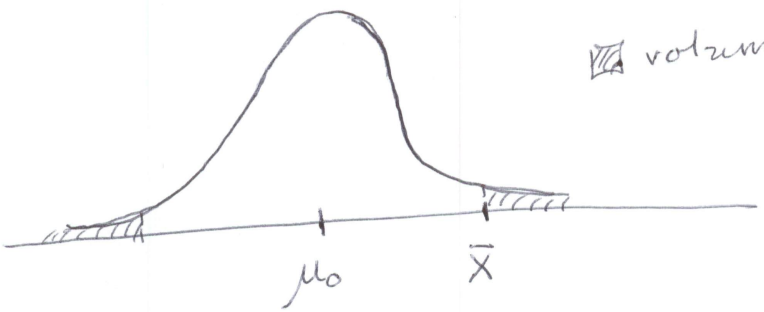
Is the null we assumed reasonable?

Another way of asking this is:

Assuming the null is true, how likely would our data we observe truly be?

How do we answer this? p-value!

p-value is the probability of seeing something more extreme than we saw with the data assuming the null is correct.



volume = p-value.

So, we think: if the prob. of seeing a more extreme value is low, then the null is not very reasonable.

The critical value is simply a value that gives us a specified p-value. This specified p-value is called a significance level)

Reject null if p-value < sig. level

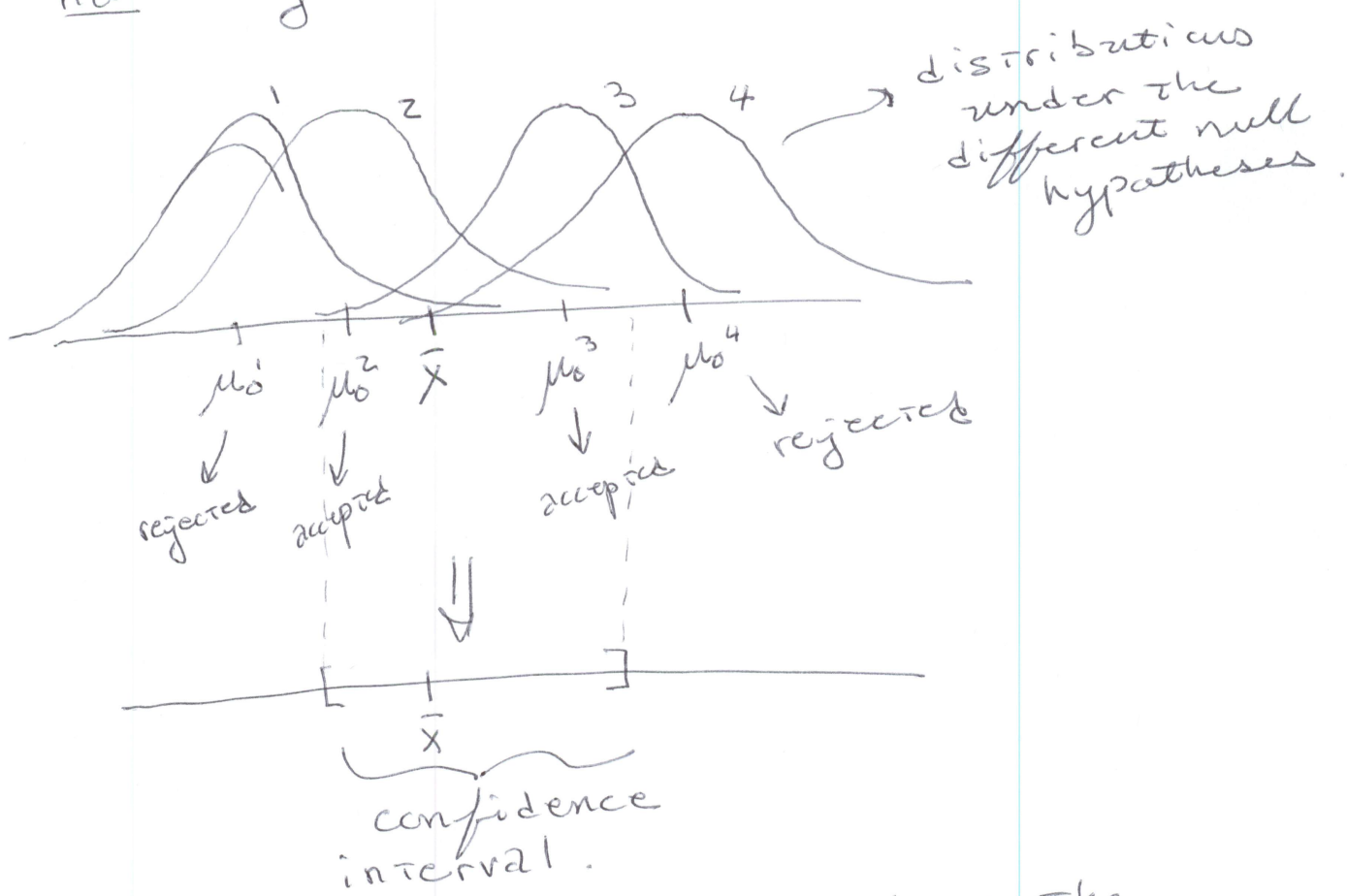
OR $|\bar{X}| <$ crit. value.

These are the same thing.

Finally, CONFIDENCE INTERVALS

A CONFIDENCE INTERVAL is an interval that we are $(1-\alpha)\%$ confident that the true parameter value is in that interval. (α is the confidence level - you choose this - usually $\alpha = .05$)

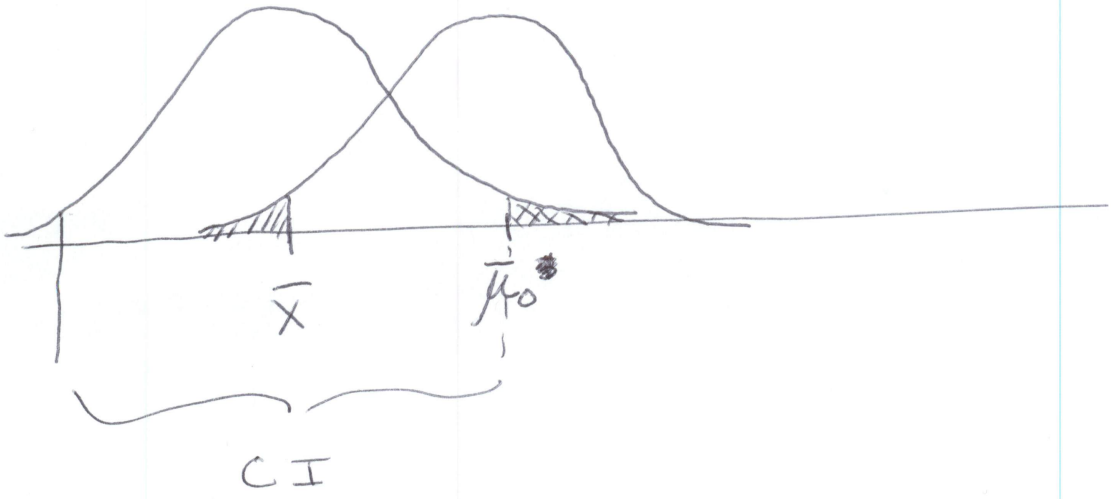
For a ~~t test~~ sample mean test, one such confidence interval is the set of null hypothesis values μ_0 such that $H_0: \mu = \mu_0$ is not rejected.



It just so happens that, since the normal distribution is symmetric, this set can be calculated as

$$\bar{X} \pm SE(\bar{X}) Z_{1-\alpha/2}$$

Why?



All $\mu_0 > \bar{\mu}_0$ are rejected
if and only if
All $\mu_0 > \bar{\mu}_0$ are greater than
$\bar{X} + SE(\bar{X})\, z_{1-\alpha/2}$

If $\bar{\mu}_0$ is the upper bound on the confidence interval, both the area of ▨ and ▩ are equal to $\alpha/2$.
If they got further apart, both would be less than $\alpha/2$.

Now, going back to the beginning, this can all be re-analyzed in the exact same fashion, but instead of setting $\hat{\theta} = \bar{X}$, we set $\hat{\theta} = \hat{\beta}_1$ (from OLS) and instead of $\hat{\theta} = \bar{X}$. And we set $SE(\hat{\theta}) = SE(\hat{\beta}_1)$

instead of $SE(\hat{\beta}) = SE(\bar{x})$

One of the major points of the past few weeks of lectures is that:

1) $\dfrac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} \xrightarrow{d} N(0,1)$

that is, the can use the analysis talked about before

2) $SE(\hat{\beta}_1) = \sqrt{\dfrac{1}{n} \dfrac{var[(X_i - \mu_x) u_i]}{(\sigma_x^2)^2}}$

where μ_x is the TRUE mean of X_i

and σ_x^2 is the TRUE var of X_i

(when I say of X_i, I mean the distribution of X_i)

Topic Two: Regression

OLS Assumptions: Without these, we cannot ensure $\hat{\beta}_1$ is unbiased, consistent, asymptotically normally distributed, or anything else we would like. In short, without these assumptions it is very hard to interpret $\hat{\beta}_1$.

1) $E[u_i \mid X_i = x] = 0$

Means u_i and X_i do not correlate. X_i cannot "explain" u_i at all.

2) (X_i, Y_i) are iid.

That is person i is chosen randomly from the population. e.g. i and another draw are not siblings because we have family data.

3) Large outliers are rare:

$$E[X_i^4] < \infty, \quad E[Y_i^4] < \infty$$

R^2: $\widehat{Var}(Y_i) = \widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i)$

$\hat{u}_i \equiv Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$$= \widehat{Var}(\hat{Y}_i) + \widehat{Var}(\hat{u}_i) + 2\underbrace{\widehat{cov}(\hat{Y}_i, \hat{u}_i)}_{=0}$$

$TSS = ESS + RSS$

so, $R^2 = \frac{ESS}{TSS}$, the proportion of variance in Y_i "explained" by X_i

Final Note: $SER = \frac{RSS}{n-k-1}$ = estimate of $var(u_i)$ ($k=1$ for us, so far)