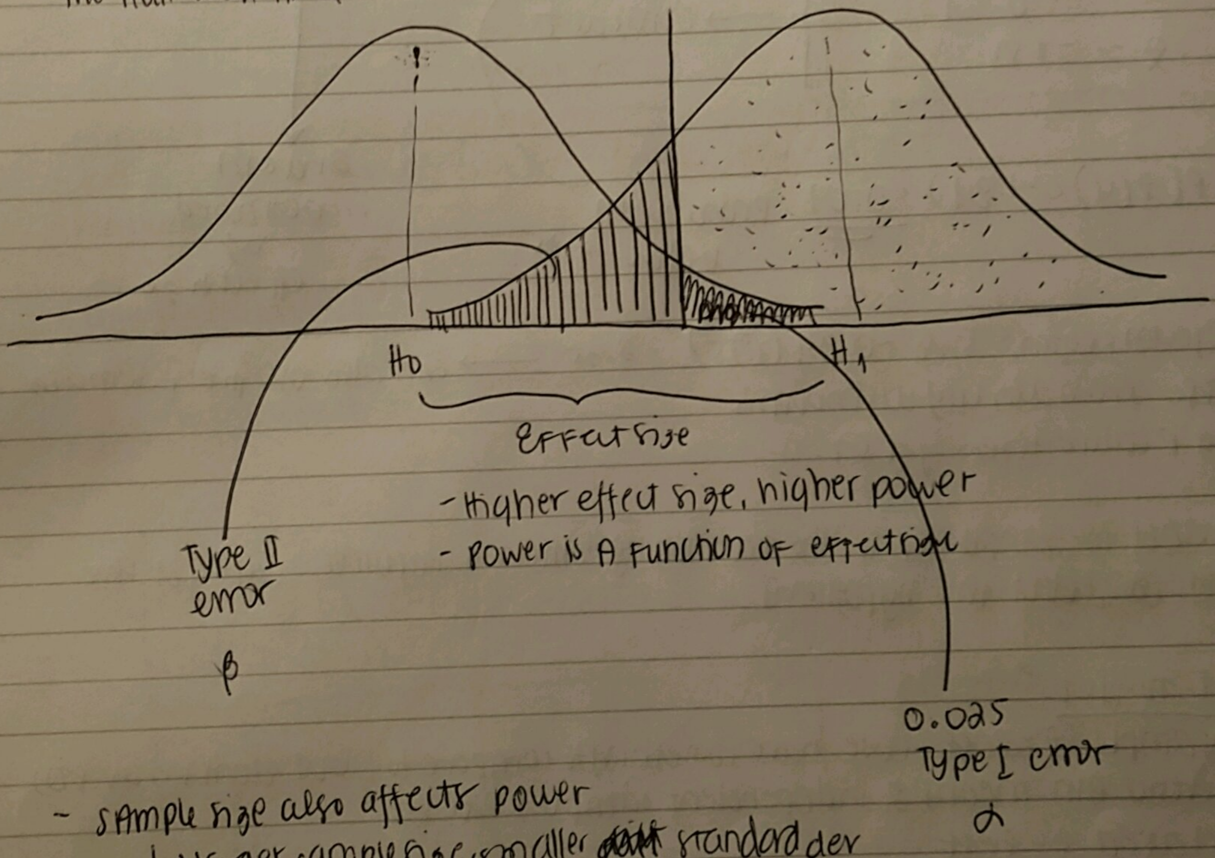
Lecture #4

POWER, REGRESSION & ANOVA — 03/27/19

- Decision Theory vs inferential statistics

Power

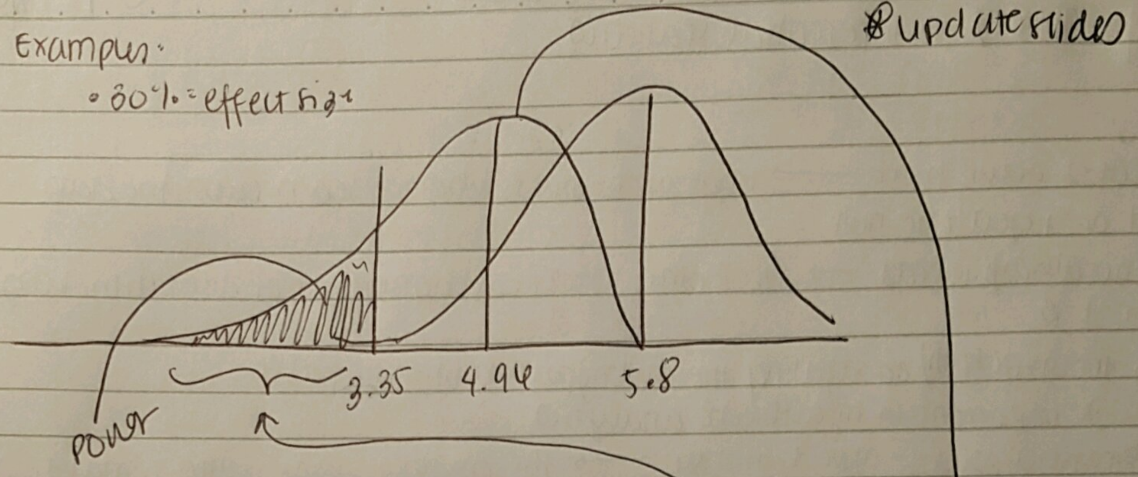
- True states occur here ⟶ what we know is what we learn from the data
- Accept or reject the null
- Null is not always correct ⟶ Type II error - practical significance or decision making
- power = $1-\beta$
  - ↳ the probability of rejecting the null hypothesis when it's false
    - * important in healthcare analytics
- complement to power - Type 1 mistake ⟶ the mistake we're willing to make, rejecting the null when it's true



- Higher effect size, higher power
- Power is A function of effect size

- sample size also affects power
  - ↳ larger sample size, smaller ~~dat~~ standard dev
- power calculator
  - effect size
  - standard deviation = prior research

- Example:
  - 80% = effect size



$t = \frac{3.35 - 4.96}{0.89}$ } → critical t

$t = -1.81$

$P(t(v) < -1.81) = \underline{\underline{0.07}}$ probability below 3.35

- Rejection Region: $5.8 - 0.89(2.75) = 3.35$ → can also use the formula
- inverse probability distribution
- Find P value then get SE.

- increase the sample size if the research hypothesis is inaccurate, can't get the power to detect this hypothesis

<u>WHAT TEST TO USE?</u>

1. one sample t-test (we have something to compare to, but don't know SD)
2. Difference b/n means & independent sample t-test
3. correlated t-test
4. z-scores — take out impact of day to day i eliminate impact of circumstances you don't care about
   - t-test work w/ sample mean. vs z test work w/ 2 populations.
5. one sample t-test, no SD but can compare
6. 2-sample t-test (since 2 diff graduates)
7. correlated t test
8. t-test (population information) → diff t-test 2 sample and $\underline{\underline{z}}$ test
   ↳ z
   ↳ differences

⊗ CLASS ON SAT AT 9:00 AM

REGRESSION & ANOVA

- predict GPA on the basis of test scores
  - selection (Bias)? — Regression
  - prediction problem
- years of education & income
  - correlation
- ~~[illegible]~~ ANOVA
- ANOVA

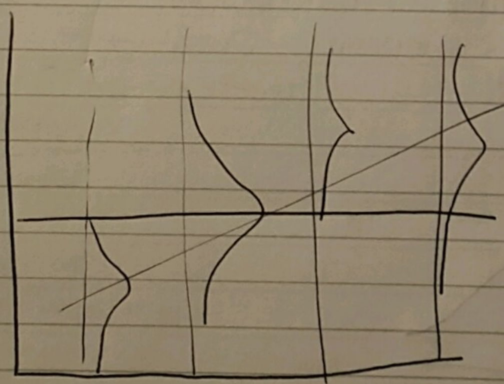
diff b/n corr e Reg?
corr = semantic rel
Reg = conditional analysis
$P(y|x) \neq P(x|y)$

Does Age of clock predict its price?

- error term not needed if you don't have the actual one? prediction term
  - * simple regression — 1 indpt var
- simple Regression if just Age ⟶ it depends on the client
- ~~[illegible] your constraints from the staff~~
- Realtime calculation

- Plotting Data
- Average Price = Best to guess ⟶ "Affine PREDICTION"
  - minimize error

$\bar{y}$



⟶ multiple predictions

-perfect relationships

calculate the data:
more error - lower corr, more regression to the mean we obtain
⟶ mean

- converge to the memory
- standardize both of them

- model deterioration
- premium

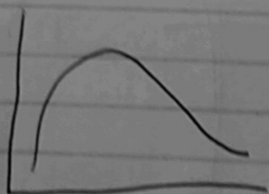
$X = 0$

$y = \alpha + \beta(x) + \varepsilon$

- Testing hypothesis here: $H_0: \beta = 0$
  $H_1: \beta \neq 0$
  - naive prediction
- semantic variables
  - SSTO = sum of squares of model + SSE → varing due to indert model
    ↳ variammy
    ↳ var that can't be explained by model
  - $SSTO = \sum_{i=1}^{n} (y - \bar{y})^2$
    - Also the same as sum of squares
    - Quantitative method
  - $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - $SSM = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$ → var in y due to x
  - $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- Decomposing the variance = most of statistics
  - df = 30
  - (n-1) - df } depends on polynomial
- mean squared error

$F = \frac{MSM}{MSE} \sim 2$



F equals or > 1

F (df for MSM, and df for MSE)

ing $\boxed{F_{(1, 30)} = \frac{MSM}{MSE} \sim 2}$

- working w/ variances

$\boxed{-R^2}$ - F.DIST & - Excel

Significance Test

$t = \frac{b}{SE_b}$ $SE_b = \sqrt{\frac{MSE}{SS_X}}$

$p = 0.0000$

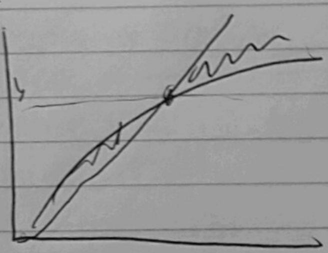
p - test ~~[illegible]~~ experiment

β - generate [illegible]

~~[illegible]~~

~~[illegible]~~ Assumption and ~~[illegible]~~ [illegible]

- [illegible]
- [illegible]



[illegible]

multi [illegible] → [illegible]

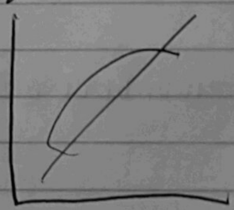
t

- Are we approximating [illegible]
- [illegible]

[illegible] transform [illegible]

transform X

[illegible]



Transformations?

- [illegible]

$y = \alpha + \beta_1 (x) + \beta_2 (x)$

- Approximating Linearity

- standardized resid

Homoscedasticity!

[illegible] plot

multiple regression