- STATS 412: 4 P.M. SKB 2500
  - ~~HT/CIs~~ for population proportion
  - ~~HT/CIs~~ for diff of population means
  - ~~HT/CIs~~ for mean of population diffs (same as 1-pop)
  - ~~HT/CIs~~ for diff between population proportions
  - ~~"General Inference Topics"~~ ??
    →
  - Correlation Coeff + LSRL
  - Uncertainties in data
  - Check assumptions + transform data

HYPOTHESIS TESTS FOR POP. PROPORTION:

- Pop. proportion ≡ pop. mean for 0s, 1s population
  - → $X \sim N(\mu = np, \ \sigma^2 = np(1-p))$ [Binomial]
  - → $\hat{p} = \frac{X}{n} \sim N\left(\mu = p, \ \sigma^2 = \frac{p(1-p)}{n}\right)$
  - → USE Z-TEST
  - → Assumptions:
    - RANDOM SAMPLE (no interdependence)
    - Approx Normal
      - → # successes, # failures ≥ 10
      - → $np \geq 10$; $n(1-p) \geq 10$
  - $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ ← stdev of original

① Let p be the population proportion of all Gen Zers who would say that " . . . " is a goo[d] thing for our society.

$H_0: \hat{p}_z \leq 0.52$ ← population proportion

$H_1: \hat{p}_z > 0.52$

② We would like a RS from Bernoulli population and distribution of $\hat{p}$ can be approximated w/ Normal dist.

① R.S.: Not stated. We hope that any one individual sampled had no effect on any other individual's answer. Peer pressure may be a concern.

- 1178 is 5% of 23560. Reasonable that pop. size of Gen Z-ers is > 23560

② Is Normal Approx Good?: $np = 1178(0.52) = 612.6 \geq 10$; $1178(1-0.52) = 565.4 \geq 10$ } thus, $\hat{p}$ ~ approx Normal. So, can use z-procedures

③ Step 3: Calculate test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{730/1178 - 0.52}{\sqrt{(0.52)(0.48)/1178}} = 6.85$$ ← very large

$$\text{P-value} = P[\hat{p} \geq 0.52] = P[z \geq 6.85] \approx 0$$ ↳ reject H_0

Step 4: As we rejected H_0, there is strong ($p \approx 0$) evidence to suggest that the population proportion of all Gen-Zers who would say that increased diversity is higher than 52%, which is the corresponding popul. prop of Gen X-ers.

C.I.s for POP. PROPORTION:

- AGRESTI-COULL ADJUSTMENT
  - → $\tilde{n} = n + 4$; $\tilde{p} = \dfrac{X+2}{n+4}$ (add 4 observations
- CI: $\tilde{p} \pm z_{\alpha/2} \sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{n}}$

  for 95%, this would be 2.5%

- Assumptions:
  - ONLY RANDOM SAMPLE
  - no need to check normality

Hypothesis Tests for Diff of Pop Means $(\bar{X} - \bar{Y})$ — not paired

- Assumptions:
  - 2 indepent R.S. from normally dist. pops
  - ① R.S. (for both)
  - ② independent samples
  - ③ normality

  $\Big\}$ "sufficiently large"

- Use t-test unless both pop. stdevs known (rare)

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}$$

(df complicated, given on exam)

- CIs:

$$(\bar{x} - \bar{y}) \pm t_{\nu, \alpha/2} \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

Population Mean of Differences $(\overline{X - Y})$ (paired)

- Assumptions:
  - → RS from Normal pop (same as single pop)
- $T = \dfrac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$
- $\bar{d} \pm t_{n-1, \alpha/2} \dfrac{S_D}{\sqrt{}}$

Difference Between Two Proportions:

- Assumptions:
  - Independence between samples
  - R.S.
  - ~Normal (10 succ / fail in each sample)

- Pooled Proportion:

$$\hat{p} = \frac{X + Y}{n_x + n_y}$$

$$Z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

$$\hat{p}_x = \frac{X}{n_x}$$

$$\hat{p}_y = \frac{Y}{n_y}$$

$$\hat{p} = \frac{X + Y}{n_x + n_y}$$

- CONFIDENCE INTERVALS
  → No need to check sample sizes / normality
  - Agresti - Coull
    → Add 4 samples still
      → 2 to each, 1 S / 1 F
      → $\tilde{n}_x = n_x + 2$ $\quad \tilde{p}_x = \frac{X+1}{n_x + 2}$
      $\tilde{n}_y = n_y + 2$ $\quad \tilde{p}_y = \frac{Y+1}{n_y+2}$

  - $(\tilde{p}_x - \tilde{p}_y) \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_x(1-\tilde{p}_x)}{n_x} + \frac{\tilde{p}_y(1-\tilde{p}_y)}{n_y}}$

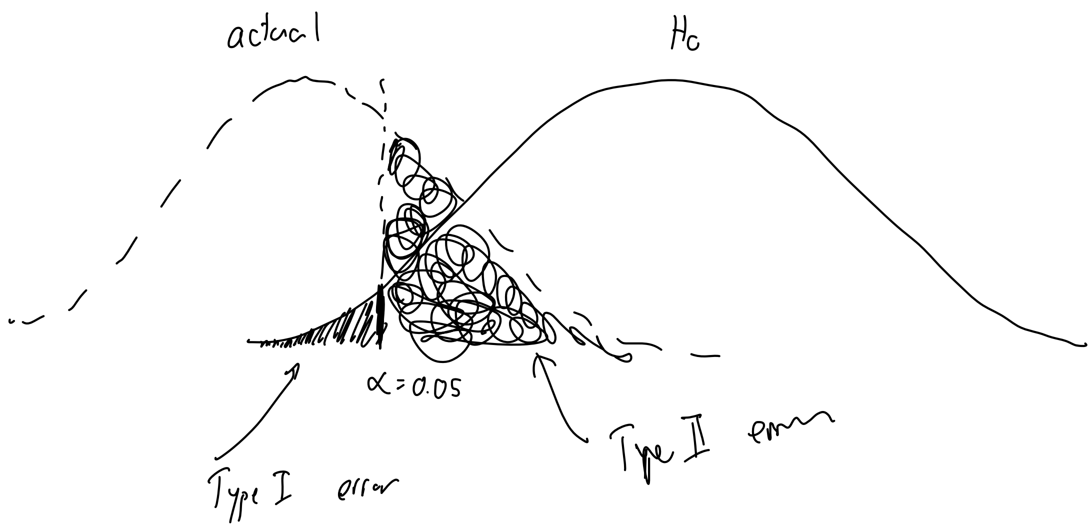
- Fixed-level testing: decide α before data
- Type I / Type II errors:
  - → If H_0 is true and you reject → TYPE I
  - → If H_0 is false and you accept → Type II

I: H_0 true

II: H_1 true

→ $P(\text{type I}) = \alpha$ (for single-sided test)

traditionally, we focus on minimizing Type I



As n increases, $P(\text{error}) \downarrow$

INTRO TO LINEAR REGRESSION:

1. Form
2. Direction
3. Strength
4. Outliers / Unusual

$\left.\right\}$ describing scatterplots

$\mu_y = B_0 + B_1 x$

$\hat{y} = B_0 + B_1 x$

Residual: e = Actual − Pred

$y_i - \hat{y}_i$ (negative → overestimate)

- $b_1 = r \cdot \frac{S_y}{S_x}$
- $b_0 = \bar{y} - b_1 \bar{x}$
- Use Multiple R^2 from R output
- Do NOT extrapolate
- R^2: (%) amount of variation explained by regression line

$$\frac{S_y^2 - S_{resid}^2}{S_y^2} = R^2 = 0.879$$

- TYPES OF OUTLIERS
  - Non-leverage: weird y, but in x-range
  - Leverage: Out of x-range
    - Influential: changes line (y weird)
    - Non-influential: does not chg line (y fine for that x)

- Checking Conditions

→ L: Linearity

→ I: Independent Samples (no time series / time-pattern)

→ N: Nearly-Normal Residuals (no outliers, look @ Q-Q plot)

→ E: Equal Variability; variability around LSRL ~ same
(Look @ residuals vs. fitted plot)
No fanning