Lecture 29.

Recall: If we are sampling a population with proportion p in some subclass, then sampling with size n, $X = Bin(n, p)$, $\hat{p} = \frac{X}{n}$ is an unbiased estimator of p.

Suppose our test is:

$H_0: p = p_0$

H_1: (I) $p \neq p_0$, (II) $p > p_0$ (III) $p < p_0$.

significance level α.

- If H_0 is true, then $X \approx N(np_0, np_0(1-p_0))$ and so $\hat{p} = \frac{X}{n} \approx N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

- We choose our test statistic

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Suppose, for some sample, $Z_0 = z_0$. Then:

(I) $H_1: p \neq p_0$ Reject H_0 if $|z_0| \geq z_{\alpha/2}$.

(II) $H_1: p > p_0$ Reject H_0 iff $z_0 \geq z_\alpha$.

(III) $H_1: p < p_0$ Reject H_0 iff $z_0 \leq -z_\alpha$.

Example: In 2018, McMaster claimed that 27% of all incoming engineering students were female. Of a random sample of 25 engineering students, 4 are female. At a significance level of $\alpha = 0.01$, does the evidence support McMaster's claim?

Solution: $\hat{p} = \frac{4}{25} = 0.16$

$H_0: p = 0.27$.

$H_1: p < 0.27$

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.16 - 0.27}{\sqrt{\frac{(0.27)(0.73)}{25}}} = -1.2389$$

For $\alpha = 0.01$, $z_\alpha = 2.33$

so $-z_\alpha = -2.33$

$$-1.2389 \geq -2.33$$

So don't reject H_0.

Prob. of a Type II error & Sample Size

- When testing proportion of a population, we can obtain a formula for $\beta = \text{Prob(Type II error)}$

① $H_1: p \neq p_0$:

$$\beta = \Phi\left(\frac{p_0 - p + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{p_0 - p - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right)$$

(II) $H_2: p < p_0$:

$$\beta = 1 - \Phi\left(\frac{p_0 - p - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right)$$

(III) $H_2: p > p_0$: $\beta = \Phi\left(\frac{p_0 - p_0 + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right)$.

We can solve for n to get a formula for approximately the sample size based on β:

(I) $$n \approx \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0}\right)^2$$

(II) and (III) $$n \approx \left(\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0}\right)^2$$

Ex: How many people should we sample to test

$H_0: p = 0.9 (= p_0)$

$H_1: p < 0.9$

$\alpha = 0.05$

if the true proportion is $p = 85\%$ and we want $\beta = 0.1$?

Solution: From tables: $z_\alpha = z_{0.05} = 1.64$, $z_\beta = z_{0.1} = 1.28$

One-sided test, so

$$n \approx \left[\frac{1.64\sqrt{(0.9)(0.1)} + 1.28\sqrt{(0.85)(0.15)}}{0.9 - 0.8}\right]^2$$

$\approx 360.279 \longrightarrow n = 361.$

Skip a bunch of sections.....

10.2) Tests & Confidence Intervals For the difference in the means of two normal populations, variances unknown.

- We consider the situation where we sample the means μ_1, μ_2 of two normal distributions $N(\mu_i, \sigma_i^2)$, $i=1,2$ where the variances are unknown.

- We assume the first population is sampled with sample size n_1 and the second is sampled with size n_2.

Test:

$H_0: \mu_1 - \mu_2 = \Delta_0$ some #.

H_1: I $\mu_1 - \mu_2 \neq \Delta_0$ II $\mu_1 - \mu_2 > \Delta_0$ III $\mu_1 - \mu_2 < \Delta_0$.

(Choose some α sig. level).

- Let $\overline{X}_i$ be the sample mean, $i=1,2$.

Then $\overline{X}_i = N(\mu_i, \sigma_i^2/n)$ $i=1,2$.

So $\overline{X}_1 - \overline{X}_2$ is normal, $\overline{X}_1 - \overline{X}_2 = N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

- We consider two scenarios:

1) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (still unknown, but equal).

2) $\sigma_1^2 \neq \sigma_2^2$.

Case 1). $\bar{X}_1 - \bar{X}_2 = N\left(\mu_1 - \mu_2,\ \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$.

- we form an estimator for σ^2 from the sample variances, S_1^2 and S_2^2. This is the pooled estimator of σ^2:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

$$= \underbrace{\frac{(n_1-1)}{(n_1+n_2-2)}}_{w} S_1^2 + \underbrace{\frac{(n_2-1)}{(n_1+n_2-2)}}_{1-w} S_2^2$$

- Note that S_p^2 is an unbiased estimator for σ^2:

$$E[S_p^2] = w E[S_1^2] + (1-w)E[S_2^2] = w\sigma^2 + (1-w)\sigma^2 = \sigma^2.$$

- So we take our test stat to be

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

($\Delta_0 = \mu_1 - \mu_0$ assuming H_0)

which has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, when H_0 is true!!! so use ~~t-test~~ from before as appropriate.

Case 2): $\sigma_1^2 \neq \sigma_2^2$

- Cannot pool s_1^2, s_2^2.
- Take:

$$T_0^* = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- Assuming H_0, T_0^* is again t-distributed, but the has degrees of freedom ν given by

$$\nu = \left\lfloor \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1^2)}{n_1 - 1} + \frac{(S_2^2/n_2^2)}{n_2 - 1}} \right\rfloor$$

floor: round down to the nearest integer.

So from here one can use the t-tests from before.

Note: In practice the t-test for pooled variances is very sensitive to the assumption that $\sigma_1^2 = \sigma_2^2$, so if you are not totally, completely sure, its best to assume $\sigma_1^2 \neq \sigma_2^2$.