Unit 6 - CI and hypothesis tests for proportions

Confidence Intervals - Categorical

Confidence Interval means:
We are ___ % confident that the true proportion of population is from ___ to ___.
It is the specified probability of a range including parameter.

Conditions
- random
- Ind $n < .1N$
- Normal $n \cdot p \geq 10$ $n(1-p) \geq 10$

Main idea: We observe a sample. To guess the parameter, we make a range with the sample with degree of certainty because sample is likely close to the true proportion.

formula $CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

margin error = critical value × standard error

$\hat{p}$: sample proportion; n: sample size

z^*: Convert degree certainty into z^* on Calc:
- 2nd vars, invnorm
- % is area, center cause
- CI, enter.

♡ Our C.I. rather catches pop parameter or not, with repeated samples+tests, our CI would capture it % certainty times.

Hypothesis tests - categorical

Main idea: We observe a sample. If the chances of getting that sample are significantly small with the null hypothesis being true, we reject the null.

Conditions
- random
- $n \cdot p \geq 10$ $n(1-p) \geq 10$
- $n < .1N$

α - alpha
- Set alpha value: the point at which a smaller chance would be considered significant.

Hypothesis
- Set hypothesises to know what rejecting null means.
- H_0: has = sign, previous claim
- H_a: >, <, or ≠, what we are inferring

C.I. difference in proportions

$$\hat{p}_1 - \hat{p}_2 \pm z^* \underbrace{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}_{\text{standard error}}$$

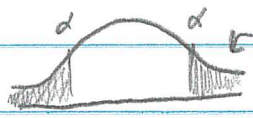
(margin of error = z^* × standard error)

If 0 is in the CI, it is not significantly different.

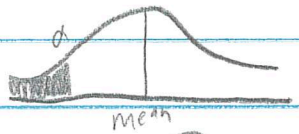
Solving P-value

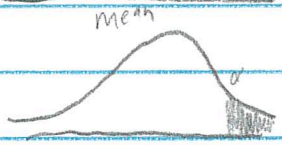
formula: $= \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z\text{-star}$

plug into 2nd vars, norm cdf, draw picture to know which side to

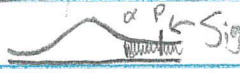
Ha: ≠  reject zone

Split % ex) 95% → 97.25% put on.

Ha: $\bar{x} <$ # 

Ha: $\bar{x} >$ # 

P-Value Interpretation

1. When $\alpha >$ p-val  Significant reject null. When null is rejected, you can accept alternate

2. When $\alpha <$ p-val 

Logic: When null is rejected, the only other option is the alternate is true. When null is not rejected, we don't have enough information to accept it but we can't confirm that it's false either.

"Don't reject null, not significant, but don't accept null or alternate."

Error types

E1: Reject a true null hypothesis

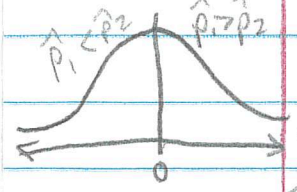
probability: α  Ex) $\alpha = .05$ means you are accepting a 5% chance you reject true

E2: fail to reject false null hypothesis.

probability: 1-power

★ Power of test is increased by sample size increasing, or the α value increasing.

Differences hypothesis test



$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c(1-\hat{p}_c) \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

→ Plug into 2nd vars, norm cdf; compare α and p-value.

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Ho: $\hat{p}_1 - \hat{p}_2 = 0$ no differ

Ha: $\hat{p}_1 - \hat{p}_2 <, >, \pm\ 0$

CI: If 0 fits in Interval, it is not significant diff

Unit 7 - CI's and hypothesis tests for numerical data

Conditions for all

random — says
Normal — $n \geq 30$ or from normal pop
Independent — $n < .1N$

- For hypothesis tests, always run Conditions, alpha, Dof, and hypotheses
- For CI run Conditions and dof.

numerical Confidence Intervals

Dof: $n-1$

$t^* = \frac{\bar{x} - M}{\left(\frac{s}{\sqrt{n}}\right)}$ or InvT conf %, dof

$CI = \bar{x} \pm t^*\left(\frac{s}{\sqrt{n}}\right)$

Context: with ___ % confidence the true parameter of ___ is from ___ to ___.

numerical Difference Confidence Intervals

Dof: $(n_1 + n_2) - 2$

$t^* = \frac{\bar{x} - M}{\left(\frac{s}{\sqrt{n}}\right)}$ or InvT %, Dof

$C.I. = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

With ___ % confidence, the true difference in means of ___ and ___ is from ___ to ___. Since 0 is included in the Interval, there is not a Significant difference between the two. OR. Since 0 isn't in our interval there's a Significant difference

numerical Hypothesis test

1. Do Conditions, Alpha, Hypothesis, and Dof.

Dof: $n-1$ α = usually .05

$t^* = \frac{obs - exp}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ → tcdf → p-val → Compare to reject / not reject

Hypothesis test numerical Differences

Dof: run these two separately, test both

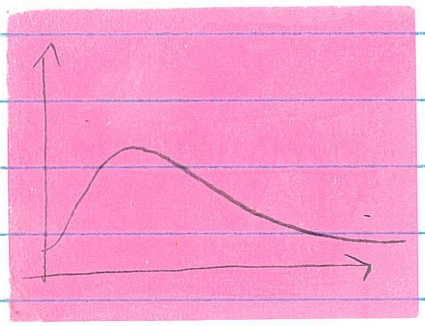
smaller sample → $n_1 - 1$ and $(n_1 + n_2) - 2$

$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - (M_1 - M_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ ← usually 0 → tcdf, compare p-value

Unit 8 - x^2 homogeneity and Independence tests

We use x^2 tests to observe if different variables are independent or dependent, and homogeneous/heterogeneous

Chi-Squared Dist:
- Almost Always right skewed
- Gets more normal as sample size goes up



♡ Always do CHIAD (conditions, hypothesis, alpha, Dof)

Conditions

- random says
- normal $n \cdot p \geq 5$ for all expected values
- Independent $n < .1N$

Dof

1 var: (# columns - 1)

2 vars: (rows-1)·(columns-1)

α - usually .05 unless says otherwise

Hypothesises

Homo: H_0: no differ dist, H_a: differ dist

Independence: H_0: no association, H_a: association

Differences of Homogeneous and Independent

Homogeneous-
- Compares 1 variable across 2 populations
- Asks if difference in the distribution of populations

Independ-
- Compares 2 variables for one population
- Asks if association between 2 variables in a population

Main order of Steps

1. Find expected values, plug in obs data L_1, etp data L_2
2. CHIAD, plug lists into formula, Sto → L3, Stats Calc vars on L3, Look at Σx. That's the test statistic → 2nd vars, x^2cdf, Compare p-value

Calculating x^2

Expected values -

- If %'s, multiply each % by the total sample size.
- If table, do $\left(\frac{\text{row total} \cdot \text{column total}}{\text{total}}\right)$

Ex)

	kayak	bike	raft	
NOC	79 observed			720
WW				
Gorge				
	123			1745

To find the expected # of kayakers at NOC, do:

$$\left(\frac{720 \cdot 123}{1745}\right)$$

- Plug obs into L_1, expected into L_2

$$x^2 = \sum \frac{(obs-exp)^2}{exp} \quad \text{So...} \quad L_3 = \frac{(L_1-L_2)^2}{L_2}$$

Sto that into L_3, then Calc L_3 and $\sum x$ is test stat.

Convert to p-value: $\sum x \rightarrow$ 2nd vars, #8 x^2cdf.

If $p > \alpha$ not reject null, results not Significant, no association between vars or no difference in dist. of populations.

If $p < \alpha$ reject null, accept alt of associated or difference in dist. of populations.

Residuals - part of unit 9

→ Sto L_4 for res

- It is possible to make residual points from $L_2 - L_3$ or (obs-ex

To get L_3: $y = \text{slope}(L_1) + y\text{-int}$. This gives expected values of LSRL. This tests if the LSRL is a good fit. Do (L_1, L_4) as residual points. If no pattern, linear is good fit.

Unit 9 - Slope applied to CI and Hypothesis tests

2-variables

How to explain slope in context:

The y-var: _____ goes up/down ____ units for every one x-var: _____ goes up.

There is variation between slope and y-int of different samples, so investigate with CI + hypo test.

C.I with slope

formula:

$b \pm t^*(SE)$

Dof = n-2

t^*: InvT, %, Dof

This C.I. gives range of true slope.

Context: we are ___% conf The true population slope of ____ is from ____ to _____.

If 0 is in our interval, it could be reasonable to conclude there is no association between the two-variables.

Computer outputs

variable	Coef	SE	t-ratio	prob
Constant	#y-int	standard error y-int		
variable	# Slope	standard error slope		

r^2 = # S = #

Variables

B = Slope pop A = y-int pop

b = slope sample a = y-int sample

Prob row: the probability we get that sample if true slope = 0. So low prob = high chance of association

THE END!