idea on how to [illegible]

- Careful evaluation of Documented is important to know the background of the data & how to best handle it when bias

LECTURE 2: INFERENTIAL STATISTICAL ANALYSIS

BAG A or BAG B

- ~~You're to choose blindly between~~ Based on the context a bag shown, you are to decide which bag you're with.
  - Bag A or Bag B? Anyone you get, you pay for it
- Make your decision ~~based~~ using a frequency plot
  - Decision rule: Look @ the values that are possible for our chosen of our data and look for which is more likely & reject/no reject the null
    - You get $10 = More likely to be bag A → Not reject Ho
      $60 = Less likely for bag A → Reject Ho
      $30 = Hard to tell

Type I error = $\frac{2}{20}$ x100 = 10%

Type II error = $\frac{6}{20}$ x100 = 30%

α = Type I error = significance level

β = Type II error

* Based on our decision rule, our α = 0.10

- Often we pick the α option, then we look @ our data to get a p-value → How likely is it to get the results you got or more extreme
  - This is done instead of basing our decision rule off the value we got & considering that for every different bc/grad
  - Let's say we pick $40 from the washroom bag, do we reject or fail to reject? Let's look @ the p-val = $\frac{4}{20}$ = 0.2 or 20% chance of getting $40 or more. We'd have to tail to reject the null
  - Universal Decision rule → Reject Null if $p \leq \alpha$

    If the likelihood of making an error is smaller than the Type I error

TUTORIAL TO BAYESIAN

chance of getting £40 or more. We'd have to tail to reject the null

- Universal Decision rule → Reject Null if $p \leq \alpha$

  ↳ If the likelihood of making an error is smaller than the Type I error

WK 1

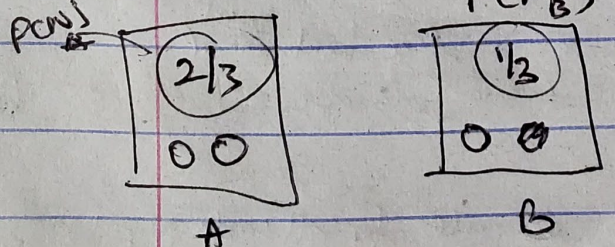
INTRODUCTION TO BAYESIAN

- There are 2 primary frameworks in stats ① Frequentist ② Bayesian stats stats
- In frequentist stats, an answer to a question can either be correct (100%) or incorrect (0%). Any uncertainty can only be given in Bayesian

Reason — In freq → probability made in world

↳ In Bayes → probability made in your head

- You can only update probability in Bayesian statistics



- 2 chances of picking W from bag A
- 1 chance of picking W from bag B

- In Bayesian ~~stats~~ you can update probabilities based on the new information you get

THE PYTHON STATISTICS LANDSCAPE

WK1 ~~This or That? Language and Notation~~

- List comprehension: my_list = [expression(i) for i in input_list]
- Your documents should explain the function use, arguments & output
- lambda function → anonymous function

  lambda arguments : expression Ex: (lambda x : x**2)(5) = 25

WK2 ESTIMATION A POPULATION PROPORTION W/ CONFIDENCE

Conf Interval = Best Estimate ± Margin of Error

~~Estimation~~ Whenever we have a statement that we want to infer or we have to come up with a Population ± (a) Parameter of interest

SE = $\frac{\sigma}{\sqrt{n}}$ or $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ → proportion

WK2 ESTIMATION A POPULATION PROPORTION W/ CONFIDENCE

Conf Interval = Best Estimate ± Margin of Error

Whenever we have a statistic that we want to infer on we have to come up with a Population ± (a) Parameter of interest

$$SE = \frac{\sigma}{n} \quad \text{or} \quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \rightarrow \text{proportion}$$

- What does CI mean? → A range of reasonable values for our param
  - We believe with 95% confidence our popular param to be with the CI

WK2 UNDERSTANDING CONFIDENCE INTERVALS

- We use our sample <statistic> proportion to construct the confidence interval but the CI is referred to a "confidence interval for a sample statistic" but it a range of values to help us estimate the POPULATION Param with a high level of confidence.
- How to best interpret & your confidence interval goes.
  - (i) State the confidence level used
  - (ii) Convey that the data is used to estimate the pop param
  - (iii) Include the rate for all such populations
  - (iv) Indicate the sample size.

- The population param is unknown hence the need for the study to estimate its value

NOTE: If u incorrect to ~~the~~ interpret the confidn level as probability or chance

↳ It's not a 95% chance that our pop param is in the interval

this is because that ~~interval~~ interval is fixed & our populatn param is fixed but unknown

The confidence interval refers to our confidence in the statistical procedure used to make the confidence interval

- We are confident that with our method used, 95% of the time, it'll contain our pop. parameter

⇒ we only did this once & we used an unbiased estimate with a representative study & with a 95% conf interval, we are sure that if 95% of the time, the interval gotten will contain the population param.

VVL2 ASSUMPTIONS for a SINGLE POPULATION PROPORTION CONF. INTERVAL

① Assumpt. 1: Our sample is a SRS

study & with a 95% conf interval, we are sure that if 95% of the time, the interval gotten will contain the population param.

WK2 ASSUMPTIONS for a SINGLE POPULATION PROPORTION CONF. INTERV

(i) Asmpt. 1: Our samp is a SRS

(ii) Asmpt -2: We need a large enough sample size so that we can approx our sample dist with a normal curve

How do we check (i) → Analyze how the sample was collected to make sure its representative. If complex use techniq mentioned in last cour weighting, etc.

(ii) Make sure you have at least 10 of each response outcom

WK2 CONSERVATIVE APPROACH & SAMPLE SIZE CONSIDERATN

- Conservativ approach is used when we "doubt" our point estimate
- Margin of Error is dependent on → confidence level
  → sample size
- Usually we pick a Conservatively; you use your desired MoE & conf. level to find out a sample size that will give you your desired MoE

"What sample size would we need to have a 95% conservativ confidence interval with a Margin of Error of only 3% (0.03)"

* Split screenshot question by topic e.g. ↓ where ... through

For best estimate → $\hat{p} = 0.5$

WK 2 PRACTICE QUIZ

conf int = 95%, MoE = ~~5~~%

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}}$

Question 6

confidence MoE $= z^* \frac{1}{2\sqrt{n}} = \frac{1.96}{2\sqrt{n}}$

MoE $= \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{n}}(0.98)$

$0.03 = \frac{1}{\sqrt{n}} \rightarrow 0.03^2 = \frac{1}{n} \times 0.98^2$

$n \geq 1068 \leftarrow n \geq 1067.11$

WK 2 QUIZ

Q3 → MoE $= 2 * \frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{232}}$

Q6

MoE $= z \times \frac{1}{2\sqrt{n}}$

$0.04 = \frac{1.96}{2\sqrt{n}} \rightarrow (\sqrt{n})^2 = \left(\frac{1.96}{2(0.04)}\right)^2$

$n = 600.25 \rightarrow n \geq 601$

Q7 98% MoE $= \frac{z}{2\sqrt{n}} = \frac{2.326}{2\sqrt{n}}$

$n = \left(\frac{2.326}{2 \times 0.03}\right)^2 = 1502.87$

$n \geq 1503$

WK 2 ESTIMATING A DIFF. IN POP. PROPS W/ CONFIDENCE

$\frac{\sigma}{\sqrt{n}}$ $\quad 0.04 = \frac{1.96}{2\sqrt{n}} \rightarrow \sqrt{n} = \frac{1.96}{2(0.04)}$

(Q7) 95% $\quad MoE = \frac{z}{2\sqrt{n}} = \frac{2.326}{2\sqrt{n}}$ $\quad n = 600.25 \rightarrow n \geq 601$

$$n = \left(\frac{2.326}{2 \times 0.03}\right)^2 = 1502.87 \quad n \geq 1503$$

WK-2 **ESTIMATING A DIFF. IN POP. PROPS W/ CONFIDENCE**

- Best Estimate $\pm$ Margin of Error

$$\hat{p}_1 - \hat{p}_2 \pm z_{crit} \cdot SE_{(\hat{p}_1 - \hat{p}_2)}$$

From table, $\hat{p}_1 = 0.55$ ... 55% of parents in our sample had kids sig lessons

$p_2 = 0.37$ ... 37% of parents of black kids have had swim lessons

$$\hat{p}_1 - \hat{p}_2 \pm z_{crit} \times \sqrt{\frac{\hat{p}_1(\hat{p}_1 - 1)}{n_1} + \frac{\hat{p}_2(\hat{p}_2 - 1)}{n_2}}$$

what 95% conf

$\rightarrow = (0.1123, 0.2477)$

Interpretation: With 95% confidence, the population proportion of parents w/ white kids who have taken swimming lessons is 11.23% to 24.77% higher than the population proportion of parents with black children who have taken swimming lessons

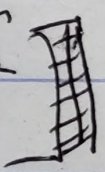
WK2

INTERPRETATIONS & ASSUMPTIONS for TWO ~~CONFIDENCE~~ POPULATION PROPORTION INTERVALS

Note if 0 is in between our confidence interval for the difference in effects, then there is a ~~prob~~ possibility that there might be no difference.

Assumptions: (i) We have two indepent random samples → in the sampling screen

(ii) We need large enough sample [We need @ least 10 yes's & @ least 10 no's for each sample



WK2

ESTIMATION A POPULATION MEAN W/ CONFIDENCE

- Now our response is quantitative instead of categorical <proportion>
<mean>

Graphical Summary

- Histogram: Shows that CWD were for below 70 as to above 110
  ↳ we see a right skew in our dist ↑
- QQ plot = Also shows us with the tail @ the lower end off.

Numerical summary → min, max, mean, SD, n

wk 2 **ESTIMATION A POPULATION MEAN W/ CONFIDENCE**

- Now our response is quantitative instead of categorical <proportion>
  <mean>

Graphical Summary

- Histogram: Shows that CWD ~~were~~ for below 70 as to above 110
  ↳ we see a right skew in our dist
- QQ plot = Also shows us ↑ with the tail @ the lower end off.

Numerical summary → min, max, mean, SD, n

↳ The measures varied from the mean of X with an average of (SD) in .

- Our best estimate is the sample mean but it'll vary ~~of~~ with repeated sampling

Q ‖ Does the response variable need to be normal for the sample mean dist to be normal?

∴ Standard Error of the Sample Mean = $\frac{\sigma}{\sqrt{n}}$ depends on how much variab & how much data

(we see a right skew in our dist ↑

- QQ plot = Also shows us with the tail @ the lower end off.

Numerical summary → min, max, mean, SD, n

↳ The measures varied from the mean of X with an average of (SD) in .

- Our best estimate is the sample mean but it'll vary if # repeated with sampling

Q Does the response variable need to be normal for the sampling mean dist to be normal?

∴ Standard Error of the Sample Mean = $\frac{\sigma}{\sqrt{n}}$ depends on how much variab & how much data was sampled

Drawback ∴ We can never calculate it because we don't know the population SD (σ). ∴ We need to come up with an estimate.

▷ Estimate Standard Error of the Mean = $\frac{(s)}{\sqrt{n}}$ → sample SD

95% Confidence Interval

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

- **Why t statistic?** — Here you don't need a large sample size to conduct a confidence interval for a population mean unlike the requirement for that of proportion.
  - And since we're using (s); we'd have to adjust for multiplier which determines the distance from the mean and t comes from a t distribution with n-1 degrees of freedom
  - The t statistic varies based on the degree of freedom which is dependent on the sample size.

E.g. for 95%

$n = 25 \rightarrow t^* = 2.064$	
$n = 1000 \rightarrow t^* = 1.962$	

s differs depending on the sample size. It get closer to the true σ as n increases so t^* accounts for that

Interpreting the Confidence level:

- It's giving us a range of reasonable vals for our parameter

dependent on the sample size.

E.g. For 95% | $n = 25 \rightarrow t^* = 2.064$

$n = 1000 \rightarrow t^* = 1.962$

on the sample size. It get closer to the true 0 as n increases so t^* account for that

Interpreting the Confidence level

- It's giving us a range of reasonable vals for our parameter

e.g. With 95% confidence, the population mean cartwheel Dist for all adults is ESTIMATED to be between 76.26 inches & 88.70 in.

↳ We have context, what param we're estimating & Conf level of use

Our confidence is in the procedure we're using

- we would expect 95% of those resulting intervals to contain the population mean cartwheel distance.

- Assumptions for Confidence Interval for Population Mean
  - data considered from a random sample
  - population of responses is normal (also a large sample size helps)

WK2

ESTIMATING A MEAN DIFFERENCE FOR PAIRED DATA

- With paired data we treat the simultaneously because they're matched in some way
  - e.g.: How someone writes w/ both hands, pre & post treatment of an individual, matched wolves data etc. Twin data, due to blood relation, by age or gender or weight.

variable: difference of measurements btw pairs.

Population = all identical twins

Param of interest = μ_d (diff = older - younger)

goal: Construct 95% interval for the mean diff of...

CI global = (0.0025, 0.1652) yrs

Interpreting the CI: w/ 95% the pop mean diff of the older twin's (vs the younger twin's) self reported education is estimated to be btwn 0.0025 yrs and 0.1652 yrs. with 95% confi

variable: difference of measurements btw pairs.

Population = all identical twins

parameter of interest = μ_d (diff = older − younger)

goal: Construct 95% interval for the mean diff of...

CI given = (0.0025, 0.1652) yrs

Interpreting the CI: W/ 95% the pop mean diff of the older twin (vs the younger twin) self reported education is estimated to be btw 0.0025 yrs and 0.1652 yrs with 95% confidence.

Means that both ends are in the +ve end, we expect the older twin has more education. With values being close to 0, that education jump might not be that large.

Assumptions — random sample of identical twins

— Population of diff is normal (or large enough sample size)

WK 2 ESTIMATING A DIFFERENCE IN POPULATION MEANS W/ CONFIDENCE

(INDEPENDENT GROUPS)

(0.0025, 0.1652) yrs

Interpreting the CI: W/ 95% the pop mean diff of the older twin's (vs the younger twin's) self reported educ or educ to be btw 0.0025 yrs and 0.1652 yrs.

Means that both ends are on the +ve end, we expect with 95% confidence the older twin has more educ. With values being close to 0, that educ jump might not be that large.

Assumptions — (i) random sample of identical twin pairs

— (ii) Population of diff is normal (or large enough sample size)

WK-2

ESTIMATING A DIFFERENCE IN POPULATION MEANS W/ CONFIDENCE

<FOR INDEPENDENT GROUPS>

sampled based on gender

- Population: Mexican-American adults (18-29 yrs)
- Parameter of interest ($\mu_1 - \mu_2$): BMI (kg/m^2) → A quantitative variable

2 Approaches for discovering the MoE

(i) Pooled approach: If variances are close to each other, we can assume them as equal

(ii) Unpooled approach: Assumption of equal var. dropped

- for df for finding t^* use the smaller of $n_1 - 1$ and $n_2 - 1$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

→ est. standard er $= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- df $= n_1 + n_2 - 2$
- This is only used if pop. variance are equal

QQ plots compare sample quantiles w/ quantiles of a normal dist.

- We compare IQRs (boxplots) & sds of both groups to find out if we use pooled / unpooled variance

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

→ est. standard er $= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- df $= n_1 + n_2 - 2$
- This is only used if pop. variance are equal

QQ plots compare sample quantiles w/ quantiles of a normal data

- We compare IQRs (box plots) & sds of both groups to find out if we use pooled / unpooled variance

Based on CI $(-0.385, 1.565)$ with 95% confidence, there is a possibility of there being no difference in population means.

- Proportions have a special property, that is the variance is completely determined by the mean.

WK 2

NORMAL DISTRIBUTION: AN INTRODUCTORY GUIDE TO PDF

& CDF → <copy & past from website>

QQ plots compare sample quantiles w/ quantiles of a normal dist

- We compare IQRs <box plots> & sds of both groups to find out if we use pooled / unpooled variance

Based on CI (-0.385, 1.565) with 95% confidence, there is a possibility of there being no difference in population means.

- Proportions have a special property, that is the variance is completely determined by the mean.

?? ??

WK 2

NORMAL DISTRIBUTION: AN INTRODUCTORY GUIDE TO PDF & CDF → <copy & paste from website>

- A normal distribution (Gaussian dist) is a continuous probability distribution for real-valued variables. For a normal dist, other values other than the mean are less probable.
- CLT → Laplace observed that, even if a population doesn't follow a normal distribution, as the number of the samples taken increases, the distribution of the sample means tends to be a normal distribution

- The probability density function (pdf) & cumulative distribution function (cdf) helps us determine the probabilities & range of probabilities when data follows a normal distribution.

- The CDF is the integrate, from left to right, of the PDF

pdf

- The pdf is a statistical expression that defines a probability distribut (the likelihood of an outcome) for a discrete random variable as opposed to a continuous random variable. When pdf is graphically portrayed, the area under the curve ≡ the interval in which the variable will fall

- This The bias of an estimator is the difference btw the estimator's expected value and the true value of the parameter being estimated.

- Why we (n-1) instead of (n) when calculate sample variance?

→ Variance is the average of the sum of squares of the diff. of the obsv. from the mean. So, when we use the sample mean as an approx of the pop mean for calcut. the sample variance, the num (i.e $\sum$ squared diff. from the mean) can be small@

(the likelihood of an outcome) for a discrete random variable as opposed to a continuous random variable. When pdf is graphically portrayed, the area under the curve ≡ the interval in which the variable will fall

• Bias The bias of an estimator is the difference btw the estimator's expected value and the true value of the parameter being estimated.

• Why we (n-1) instead of (n) when calculating sample variance?

→ Variance is the average of the sum of squares of the diff. of the obsv. from the mean. So, when we use the sample mean as an approx of the pop mean for calcul. the sample variance, the num (i.e Σ squared dist. from the mean) can be small@ times. In those cases, we will get smaller sample variances. Hence when we divide the sample var. by n, we underestimate (i.e. get a biased value) the population var. In order to compensate for this, we make the denominator of the sample var (n-1) to obtain a larger value. This reduces the bias of the sample variance as an est. of the pop. var.

pdf: $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$

the sample mean as an approx of the pop mean for calcul. the sample variance, the num (i.e $\sum$ squared dist. from the mean) can be small@ times. In those cases, we will get smaller sample variances. Hence when we divide the sample var. by n, we underestimate (i.e. get a biased value) the population var. In order to compensate for this, we make the denominator of the sample var (n-1) to obtain a larger value. This reduces the bias of the sample variance as an est. of the pop. var.

Standard Normal pdf: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\cdot\sigma^2}} \cdot e^{-1/2\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$

- A standard normal distribution has a pop. mean of 0 & standard dev = 1.
  - It is gotten by subtracting the mean from each value in the sample & dividing the difference by the standard deviation

Reasons why we would do this? (i) The data is widely spread out

(ii) To put all our variables on the same scale because most times our data might be measured in diff scales (units and might affect our interpretation (bias)

- Pdf of a standard Do; $f(z) = \frac{1}{2\pi} \exp\left(\frac{-z^2}{2}\right)$

- CDF → It is a function derived from the probability density function for a continuous random variable. It gives the prob. of finding the random variable @ a variable less than or equal to a given cutoff i.e., $P(X \leq x)$

Note: Chap in "How to build a 95% CI, what is t^* in" from week2_segment into notes.

WK3 SECTION 08 A TEST FOR A POPULATION PROPORTION ; Test of one pop. proportion

Why carry out hypothesis tests? → we have a question about the pop. Can use data to support the claim concerning the question

- You make your hypothesis first before data collection to avoid influencing the beliefs after data collection

$H_0 : p = 0.52$; $H_a : p > 0.52$

- State what p is ← NOTE ; $\alpha = 0.05$] → cut off pt at which we found smth. to be significant

Note: Chap in "How to build a 95% CI, what a [illegible] ment into notes.

WK3 SECTION 08 A TEST FOR A POPULATION PROPORTION & DIFF. BTW POP. PROPS

Testing one

Why carry out hypothesis tests? ⇒ we have a question about the pop

Can use data to support the claim concerning the question

- You make your hypothesis first before data collection to avoid influencing the beliefs after data collection

$H_0: p = 0.52$; $H_a: p > 0.52$

- State what p is ← NOTE ; $\alpha = 0.05$] → cut off pt at which we found result to be significant

- Next survey collection; $n = 1018$; $\hat{p} = 56\%$

[Assumption check ⇒ (i) Rand sample (ii) Large sample size to ensure normality
$n \cdot p_0 \geq 10$
$n(1-p_0) \geq 10$

- Run hypothesis test

test statistic = $\frac{\text{best estimate} - \text{hypothesized est}}{\text{SE of estimate}} = \frac{\hat{p} - p_0}{SE}$

beliefs after data collection

$H_0 : p = 0.52$; $H_a : p > 0.52$

- State what p is ← NTCE ; $\alpha = 0.05$] → cut off pt at which we found smth. to be significant

- Next survey collected: $n = 1018$; $\hat{p} = 56\%$

[Assumption check = (i) Rand sample (ii) Large sample size to ensure normality
$n \cdot p_0 \geq 10$
$n(1-p_0) \geq 10$

- Run hypothesis test

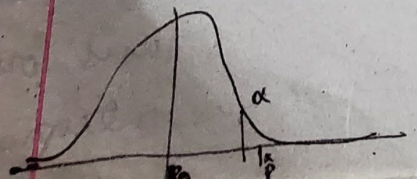
test statistic = $\frac{\text{best estimate} - \text{hypothesized est.}}{\text{SE of estimate}} = \frac{\hat{p} - p_0}{SE}$

$S.E(\hat{p}) = \sqrt{\frac{p_0 \cdot (1-p_0)}{n}}$

↳ null standard error [under the H_0]

$Z = \frac{0.56 - 0.52}{0.0157} = 2.555$

* Z test statistic because we're under a normal dist. When using proportions, we use a norm dist. as long as n is large enough

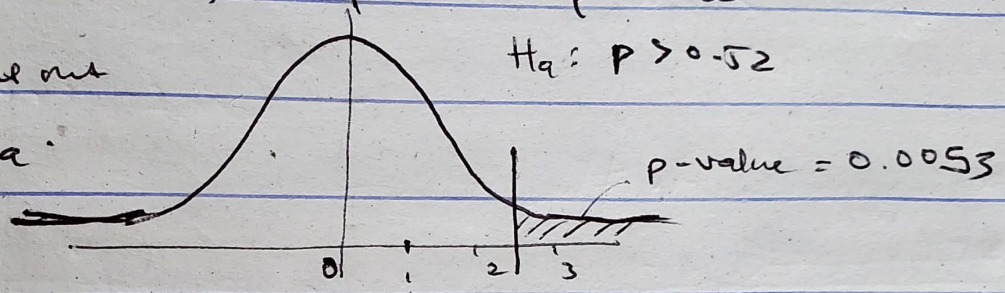


Interpret the Z-test statistic: Our observed sample proportion is 2.555 null standard errors above our hypothesized population proportion.

~~A Z-test statistic~~

- With the Z test statistic, we find the p-value

Note that p-value is not the same as p used as H_a. That our null proportion



since p-value < α ; we reject H_0

- Conclusion statement: There is sufficient evidence to conclude that the popular proportion of parents ...

WK3 SETTING UP A TEST OF DIFFERENCE IN POPULATION PROP.

- population = All parents of black kids <6-8 yrs> & all parent of hispanic children age <6-8>
- Parameter of Interest -> $P_1 - P_2$ <1- black kids, 2- Hispanic kids>
- significance level = 10% ; $\alpha = 0.10$

0 1 2 3

Since p-value $< \alpha$; we reject H_0

- Conclusion statement: There is sufficient evidence to conclude that the popular proportion of parents ...

WK3 SECTION 10 A TEST OF DIFFERENCE IN POPULATION PROP.

- population = All parents of black kids <6-18 yrs> & all parent of hispanic children <6-18> age
- Parameter of interest → $P_1 - P_2$ <1- black kids, 2- Hispanic kids>
- significance level = 10% ; $\alpha = 0.10$

$H_0 : P_1 - P_2 = 0$; $H_a : P_1 - P_2 \neq 0$

- survey results $n_1 = 247$; $\hat{P}_1 = 91/247$

$n_2 = 308$; $\hat{P}_2 = 120/308$

- Check our assumptions
- Calculate best estimate of the parameter ; $\hat{P}_1 = 0.37$, $\hat{P}_2 = 0.39$

Children <6-8>

- Parameter of Interest → $p_1 - p_2$ <1- black kids, 2- Hispanic kids>
- significance level = 10% ; $\alpha = 0.10$

$H_0 : p_1 - p_2 = 0$; $H_a : p_1 - p_2 \neq 0$

- survey results $n_1 = 247$; $\hat{p}_1 = 91/247$

$n_2 = 308$; $\hat{p}_2 = 120/308$

- Check our assumptions
- Calculate best estimate of the parameter ; $\hat{p}_1 = 0.37$, $\hat{p}_2 = 0.39$

$\hat{p}_1 - \hat{p}_2 = -0.02$

MK2

TESTING A DIFF. IN POPULATION PROPORTIONS

test statistic $= \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}$

$SE_{(\hat{p})} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$= 0.0414$

z test stat $= -0.4831$

- Check mean of Z stats is not 0
- p-value (z) = 0.63
  - ↳ two tailed
- since p-value > α ; we fail to reject null hypothesis

- Alternative approaches to this question
  - ① Chi square test but the hypothesis phrasing will be different
  - ② Fisher's Exact test = for smaller sample size

P-VALUE, P-HACKING & MORE

- P-value - level of surprise. Tells about the surprise of the data. Low = higher exp of surprise,
- P-value doesn't reveal the level of the truth of the data.
- Having a low value could mean statistical significance but not necessarily practical significance
  - You're almost guaranteed to get low p-val w/ large datas
- P-hacking - whenever we can no longer interpret the p-value by its mathematical definition. You need to know all done on the back

higher exp of surprise,

- p-value doesn't reveal the level of the truth of the data.
- Having a low value could mean statistical significance but not necessarily practical significance
  - You're almost guaranteed to get ~~small~~ low p-value w/ large datas
- p-hacking — whenever we can no longer interprete the p-value by its mathematical definition. You need to know all done on the side
  - ways of hacking: many multiple tests on a data
  - consequences: leads to uninformative studies
- p-value hacking also affects all the other estimates
- For pure p-value, pre-register, & write down all the choices, methodology that you want to follow. This helps you follow a more confirmatory test than an exploratory test (which leads to more p-hacking)
  - Include all the steps taken in the study, not just what lead to the p-value

WK3 ONE MEAN: TESTING ABOUT A POPULATION MEAN W/ CONFIDENCE

- p-hacking — whenever we can no longer interpret the p-value by its mathematical definition. You need to know all done on the data
  - ways of hacking: using multiple tests on a data
  - consequences: leads to uninformative studies
- p-value hacking also affects all the other estimates
- For pure p-value, pre-register, & write down all the choices, methodology that you want to follow. This helps you follow a more confirmatory test than an exploratory test (which leads to more p-hacking)
  - Include all the steps taken in the study, not just what lead to the p-value

WK3

OVERVIEW: TESTING ABOUT A POPULATIONS MEAN W/ CONFIDENCE

Populat: All adults

Param of interest: Population mean CWB, μ

- Define hypothesis → $H_0: \mu = 80$
  $H_1: \mu > 80$, $\alpha = 0.05$

significance level: → the results should be so unusual that we would see them no more than 5%

$n = 25$, min = 63, max = 115, $\bar{x} = 82.48$, $s = 15.06$ in

- Example assumptn —— (i) Random sample
  └ (ii) Normal distribution for CW Distance
  - histograms show right skew w/ 25 obs we can use the CLT.

- Is the sample mean of 82.48 in eg. greater than hypothesized mean of 80 in. SE of the sample mean = $\frac{\sigma}{\sqrt{n}}$ — we don't have pop std

- estimated SE of the mean = $\frac{s}{\sqrt{n}}$

- t-test statistic = $\frac{\bar{x} - \mu}{SE} = 0.82$

our sample mean is only 0.82 estimated SE above the null value of 80 in

$n = 25$, $\min = 63$, $\max = 115$, $\bar{x} = 82.48$, $s = 15.06$ in

- Example assumptn —— (i) Random sample
  - (ii) Normal distribution for CW Distances
- Histograms show right skew but w/ 25 obs we can use the CLT.

- Is the sample mean of 82.48 in eg. greater than hypothesized mean of 80 in.

SE of the sample mean $= \frac{\sigma}{\sqrt{n}}$ — we don't have pop std

- estimated SE of the mean $= \frac{s}{\sqrt{n}}$

- t-test statistic $= \frac{\bar{x} - \mu}{SE} = 0.82$ | Our sample mean is only 0.82 estimated SE above the null value of 80 in

- Determine p-value | using t-test dist table ; p-value = 0.21

since p-value $> \alpha$ $(0.21 > 0.05)$; we fail to reject the null hypothesis

- Based on our sample mean (82.48 in), we . . .

of 80 in.

SE of the sample mean = $\frac{\sigma}{\sqrt{n}}$ — we don't have pop std

- estimated SE of the mean = $\frac{s}{\sqrt{n}}$

- t-test statistic = $\frac{\bar{x} - \mu}{SE} = 0.82$ | our sample mean is only 0.82 estimated SE above the null value of 80 in

- Determine p-value | using t-test dist table; p-value = 0.21

since p-value > α (0.21 > 0.05), we fail to reject the null hypothesis

- Based on our est. sample mean (82.48 in), we . . .

- Looking @ 90% CI estimate

WK3 TESTING A POPULATION MEAN DIFFERENCE

- In what way is our data paired? How was it paired?

- Population = all rows | param of interest = Pop mean diff of Celsius quads

↳ μ_d

$\alpha = 0.05$

$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$

$\sqrt{n}$

- t-test statistic $= \frac{\bar{x} - \mu}{SE} = 0.82$ | our sample mean is only 0.82 estimate SE above the null value of 80 in

- Determine p-value | using t-test dist table ; p-value = 0.21

since p-value > α (0.21 > 0.05) ; we fail to reject the null hypothesis

- Based on our sample mean (82.48 cm), we . . .

- Looking @ 90% CI estimate

WK3 **TESTING A POPULATION MEAN DIFFERENCE**

- In what way is our data paired ? How it paired ?

- Population : all rows | param of interest = Pop mean diff (μ_d) Glucose

$\alpha = 0.05$

$H_0 : \mu_d = 0$ $H_1 : \mu_d \neq 0$

- Assumptions — Random sample of diff
  - normal dist. of (sample size of 25+) | we have a rough bell shape

$n = 20$, $\bar{x}_d = \$17.30$, $s_d = \$28.49$

t-test statistic $= \dfrac{\bar{x}_d - \mu_d}{\frac{s}{\sqrt{n}}} = \dfrac{17.30 - 0}{\frac{28.49}{\sqrt{20}}} = 2.72$ ← SE_d

p-value = 0.014 ; p-value < 0.05 — reject H_0

95% CI : ($3.97 ; $30.63)

Wk3 TESTING for ~~DIFFERENCE~~ a DIFFERENCE IN POPULATION MEANS (FOR INDEPENDENT GROUPS)

- Population : Men-American adults in NY
- Parameter $(\mu_1 - \mu_2)$: Body Mass Index
- Define hypotheses ; $H_0: \mu_1 - \mu_2 = 0$; $H_a: \mu_1 - \mu_2 \neq 0$

$\bar{x}_1 = 23.57$, $\bar{x}_2 = 22.83$; $n_1 = 258$, $n_2 = 259$

$s_1 = 6.24$ $s_2 = 6.43$

- Check Assumptions — (i) Simple random sam
  - (ii) iid
  - (iii) Normality or n is large
- Calc: test statistic

* t-test statistic $= \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{}$

$s_1 = 6.24$ $s_2 = 6.43$

— Check Assumptions — (i) Simple random sam
— (ii) iid — (iii) Normality or n is large

— Calc. test statistic

* test statistic $= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE_d}$

• Best estimate: $\bar{x}_1 - \bar{x}_2 = 0.74$

SE using pooled approach $= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $t = 1.30$ //

SE_d using pooled approach (interpret the IQR of boxplot) is $s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

" In terms of actual calculation, the CLT doesn't differ much from assuming that the data is normal. The key difference is that with the CLT, you do not need to assume that the data is norm. dist. Instead, you can rely on the fact that as the sample size increases, the samp. dist. of the sample mean becomes incr. normal, regardless of the shape of the original dist.

∘ Relevant p-value: p-value = 0.19 | we fail to reject Null ; p > 0.05

∘ Clip in the 6th paragraph under "Hypothesis Testing: Other Cons" > Namche of them

↳ This gives me more insight as to why t-dist is used over Z for normal dist.

~~WK3~~ ~~Python~~ ~~PRACTICAL NOTES~~

WK 4 → THE IMPORTANCE of GOOD RESEARCH QUESTIONS for SOUND INFERENCE

∘ ~~Data~~ is everywhere and easy to be gotten but still we need a well formulated research quest that informs our study

Key aspects of GOOD

→ (i) Target population of interest

→ (ii) Is research quest descriptive or analytic

→ (iii) Has the quest been asked b4? What will the new study add?

→ (iv) Are the variables ready available & measurable & they actual capture the concepts we want

Quest → doesn't check our 4 key aspects?

Key Aspects of Qns

- Target population of interest
- Is research question descriptive or analytic
- Has the question been asked b4? What will the new study add?
- Are the variables ready available & measurable & they actual capture the concepts we want

Bad Questions → doesn't check on 4 key aspects ?

~~(measure? type, attendance, assign ; * mean student)~~ income

Good questions

WK4 ~~DESCRIPTIVE~~ INF. EXAMPLES for SINGLE VARIABLES USING CI

~~Remember~~ Remember that the confidence interval estimates for an actual population parameter

WK4 DESCRIPTIVE INF EXAMPLES for SINGLE VARIABLES USING HYP TEST

WK4 ~~COMPARING MEANS for~~ TWO ~~INDEPENDENT~~ SAMPLES : AN EXAM