Hyunseung Kang, Stat 111

Procedure for Hypothesis Testing:

Step 1: Write out your two competing hypotheses

(Null) H_0: "Status quo/What is already believed" hypothesis (e.g. popu. Avg = 50 inches)

(Altr.)H_a: "What you want to prove/claim" hypothesis (e.g. popu. Avg > 50 inches)

Tip 1: Figure out H_a first and basically write H_0 as the opposite/equality version

Tip 2: μ_0 is the numerical boundary point for H_0. In the example, $\mu_0 = 50$

Step 2: Choose the correct hypothesis test (see boxes below)

Tip 1: When you compute the values, the direction of your H_a doesn't matter.

Step 3: Find the p-value

a. Mark the value you got from Step 2 (Z or t) on either the Normal Dist. or the t-dist.

b. Figure out which area is of interest. To do this, look at H_a (see examples below)

c. To find p-value on t-table, (i) find the degrees of freedom to the right (ii) find the corresponding t-value on that row, and(iii) read off the percentage on the top column

Example 1:

For H_a: popu avg > 3, look at the right side of your z or t value

Calculator:

Z:2nd+DISTR →normalcdf(z,100)

t:2nd+DISTR →tcdf(t,100,deg of fred)

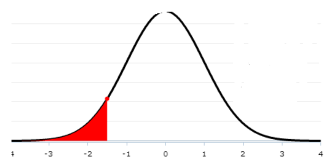
Example 2:

For H_a: popu avg < 3, look at the left side of your z or t value

Calculator:

Z:2nd+DISTR→normalcdf(-100,z)

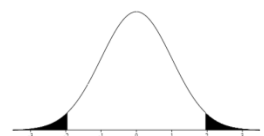
t:2nd+DISTR→tcdf(-100,t,deg fre)



Example 3:

For H_a: popu avg ≠ 3, look at the right and left side of the z or t value (basically, the two tails of the curve)

Calculator:Get numbers from example 1 and 2. Add these two numbers together!



Procedure for Confidence Intervals:

Step 1: Check to make sure that the assumptions of CLT are met

a. SRS sample from the population whose pop avg. is of interest for CI

b. We have enough draws from the population

c. The quantity of interest is a function of the sum of tickets (e.g. average or %)

Tip 1: If you can't construct the CI, it's most likely due to violation of (a)

Step 2: *For __% Confidence Interval for Population Average*

[sample mean - Z*SE_{avg}, sample mean + Z*SE_{avg}],

$SE_{avg}=\frac{Box\ SD}{\sqrt{N}}$ → (if Box SD unknown, bootstrap)→ $SE_{avg}=\frac{Sample\ SD}{\sqrt{N}}$

(e.g. 95% CI: [sample mean - 2*SE_{avg}, sample mean + 2*SE_{avg}])

For __% Confidence Interval for Population %

[sample % - Z*$SE_{\%}$, sample % + Z*$SE_{\%}$],

$SE_{\%}=\frac{Box\ SD}{\sqrt{N}}$→ bootstrap→$SE_{\%}=\frac{\sqrt{sample\% * (1-sample\%)}}{\sqrt{N}}$

(e.g. 95% CI: [sample % - 2*$SE_{\%}$, sample %+ 2*$SE_{\%}$])

Step 3: Things to consider

Point 1: Margin of Error = Z*SE

Point 2: If the sample is a considerable proportion of the population, correct the SE and use the corrected SE in CI formula

$SE_{corrected}$=CorrFactor*$SE_{uncorrected}$, CorrFactor=$\sqrt{\frac{NPopu-NSample}{NPopu.-1}}$

T-Statistic for Population Average:

When to Use:

1) You are testing one population average AND
2) The box/population model follows a Normal distribution AND (e.g. Gauss measurement model where errors are Normal)
3) SD of Box (i.e. SD of **Population**) is **unknown**

Tip: This is generally used when sample size is small (say less than 10)

Formula: (work in decimals throughout your calculation)

$$t = \frac{sample\ mean - \mu_0}{\frac{SD^+}{\sqrt{N}}}, \qquad SD^+ = Sample\ SD * \sqrt{\frac{N}{N-1}},$$

Degrees of freedom = N-1

Example: SRS sample of 10. Sample average is 3.9 and sample SD is 0.15. From step 1, we have the hypotheses

H_0: population avg = 4 *(here, $\mu_0 = 4$)*

H_a: population avg < 4

Then SD^+= 0.15 * $\sqrt{\frac{10}{9}}$= 0.158, $t = \frac{3.9-4}{\frac{0.158}{\sqrt{10}}}$=-2.001, and

degrees of freedom=N-1=9

Z-Statistic for Population Percentage:

When to Use:

1) You are testing one population percentage AND
2) CLT assumptions are met

Formula: (work in decimals throughout your calculation)

$$Z = \frac{sample\ \% - \mu_0}{\frac{\sqrt{\mu_0 * (1-\mu_0)}}{\sqrt{N}}}$$

Example: SRS sample of 1000. We observe 70% in our sample. From step 1, we have the hypotheses

H_0: population % = 2/3, *(here μ_0=2/3)*

H_a: population % > 2/3

Then $Z = \frac{0.7-\frac{2}{3}}{\frac{\sqrt{\frac{2}{3}*(1-\frac{2}{3})}}{\sqrt{1000}}}$=2.24

Tip 1: Make sure μ_0 is in decimals/fractions!

Z-Statistic for Population Average

When to Use:

1) You are testing one population average AND
2) SD of Box (i.e. **SD of Population**) **is known** AND
3) CLT assumptions are met **OR** the box/population is Normal SRS samples

Formula:

$$Z = \frac{sample\ average - \mu_0}{\frac{SD\ of\ Box}{\sqrt{N}}}$$

Example: SRS sample of 50. We observe a sample average of 4. SD of population is 2. From step 1, we have the hypotheses

H_0: population avg = 5, *(here $\mu_0 = 5$)*

H_a: population avg < 5.

Then $Z = \frac{4-5}{\frac{2}{\sqrt{50}}} = -3.54$

Two-Sample Z-Statistic for Difference in Population Percentage:

When to Use:
1) You are testing two population percentages' difference AND
2) The two samples **are independent** (i.e. not paired) AND
2) CLT assumptions are met for each box/population

Formula: (work in decimals throughout your calculation)

$$Z = \frac{\Delta - \mu_0}{SE_{\%,diff}}$$

Δ= % from one group (say group A) – % from another group (say group B)

$SE_{\%,diff} = \sqrt{SE^2_{\%,A} + SE^2_{\%,B}}$,

$SE_{\%,A}$ =SE of % from group A = $\frac{\sqrt{(Sample\ \%\ of\ A*(1-Sample\ \%\ of\ A)}}{\sqrt{N_A}}$

Tip: when calculating the SE of %, you're "boostrapping the SD" like in the CI formula

Example: SRS sample of 1000 men and 800 women. In the sample, 47% of men like Coke over Pepsi and 46% of women like Coke over Pepsi. Do males prefer Coke over Pepsi more than females?

H_0: popu. % of males like Coke – popu. % of females like Coke = 0, (*here* $\mu_0 = 0$)

H_a: popu. % of males like Coke – popu. % of females like Coke > 0

Then, $SE_{\%,males} = \frac{\sqrt{0.47*0.53}}{\sqrt{1000}}$=0.016, $SE_{\%,females} = \frac{\sqrt{0.46*0.54}}{\sqrt{800}}$=0.018,

$SE_{\%,diff} = \sqrt{0.016^2 + 0.018^2}$ =0.024. Thus, $Z = \frac{0.47-0.46}{0.024}$=0.417

Two-Sample Z-Statistic for Difference in Population Averages:

When to Use:
1) You are testing two population averages' difference AND
2) The two samples **are independent** (i.e. not paired) AND
2) CLT assumptions are met for each box/population

Formula:

$$Z = \frac{\Delta - \mu_0}{SE_{avg,diff}}$$

Δ= average from one group (say group A) – average from another group (say B)

$SE_{avg,diff} = \sqrt{SE^2_{avg,A} + SE^2_{avg,B}}$,

$SE_{avg,A}$=SE of Averages from group A = $\frac{Sample\ SD\ of\ GroupA}{\sqrt{N_A}}$

Tip: when calculating the SE of average, you're "boostrapping the SD" like in the CI formula

Example: SRS sample of 50 men and 60 women. Average height for male in sample is 76 inches with SD = 3 inches and average height for females is 75 inches with SD 1 inches. Are males, on average, taller than females?

H_0: population avg of males – population avg of females = 0, (*here* $\mu_0 = 0$*)*

H_a: population avg of males – population avg of females > 0

Then, $SE_{avg,males} = \frac{3}{\sqrt{50}}$=0.424, $SE_{avg,females} = \frac{1}{\sqrt{60}}$=0.129, $SE_{diff} = \sqrt{0.424^2 + 0.129^2}$ =0.44

$Z = \frac{76-75}{0.44}$=2.27

Two-Sample, BUT Paired, Test for Differences in Population Average:

When to Use:
1) You are testing two population averages AND
2) The two samples are **dependent** (e.g. paired, before-after experiments)
2) CLT assumptions are met

Tip 1: You can only use this test if the SD of differences is given to you. SD for each group will not be enough for this test to work!

Tip 2:You can only use this test if the two populations have the identical sample size!

Formula:

$$Z = \frac{\Delta - \mu_0}{\frac{SD\ of\ Difference}{\sqrt{N}}}$$

Δ= average from one group (say group A) – average from another group (say group B)
= average difference between groups

SD of Difference= should be given to you

N=number of paired samples (i..e how many **paired** observations do you have?)

Example: SRS sample of 200 twins. The average difference in their height is 1 inches and the SD of the difference is 4. I think there is a difference in twin's heights

H_0: avg height for one twin – avg height for another twin = 0, *(here* μ_0*=0)*

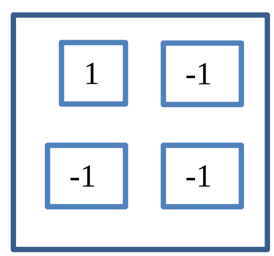
H_a: avg height for one twin – avg height for another twin ≠0

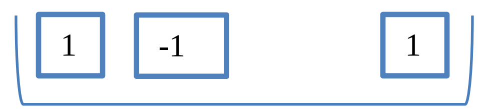
Then $Z = \frac{1}{\frac{4}{\sqrt{200}}} = 3.54$

Note: It's possible, although rare in the class, that the sample size is small and the difference is assumed to be Normally distributed. In that case, use the one-sample t-test for averages.

Box Model, Survey Sampling and Probability:

Box model Example 1. Suppose you win a dollar with 25% and lose a dollar with 75%. You play the game 80 times.





We take 80 draws, with replacement, from the box
We are expected to lose Box Average*N=-0.5*80=-40
The SE of sums is Box SD*sqrt(N)=0.866*sqrt(80)= 7.75
95% of our net winnings will be within 2*SE of -40

Box Average = (1+-1+-1+-1)/4=-0.5

Box SD = (Big –Little)*sqrt(FracBig*FracSmall)=(1-(-1))*sqrt(1/4*3/4)=0.866

Central Limit Theorem: If
(i) we draw with replacement (or SRS if sample is small in comparison to popu.) from the box
(ii) we draw enough tickets
(iii) we take the sum of these tickets,
then the sum of these tickets is Normally distributed with mean=expected sum=Box Average*N and SD=SE of sums=Box SD*sqrt(N)

Sample Average and CLT: If CLT conditions are met, average of tickets is Normally distributed with mean=Box Average and SD=SE of avg=Box SD/sqrt(N)