Two Sample Comparison

(Hypothesis testing)

We got two samples for the same thing of interest. In stastistics, we need to collect enough evidence to support they are different.

General form of two samples comparison problem:

We are interested in unknown % of total population. (inference)

What we have got are 2 samples:

Sample 1 n_1 (100) people, Percentage 1 (50%)

Sample 2 n_2 (400) people, Percentage 2 (60%)

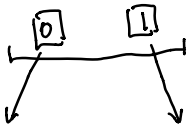
Question: Do these two samples tell us the same story about % of total population?

or Do we need to change our opinion on % of total population after we get sample 2?

Solution:

Identify the box models. Encode living on campus into 1.

Box model 1 (Sample 1)



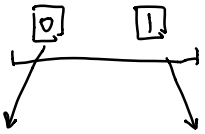
Sd of box 1

100% - Percentage 1 | Percentage 1 (50%)

(50%)

$n_1 = 100$ Sample Size

Box model 2 (Sample 2)



100% - Percentage 2 | Percentage 2 (60%)

(40%)

$n_2 = 400$ Sample Size.

This question asks about the percentage.

Why we can use some normal curve (distribution)? (1')

① Sample Sizes in both cases are large.

② The distributions (histograms) are not too skewed. (95% vs 5%
<90% vs 10%

Center:

EV_1 = Expected Value for Sample 1

= Avg of Box 1

= 0 × (100% - Percentage 1) + 1 × Percentage 1

= Percentage 1 (50%)

EV_2 = Expected Value for Sample 2

= Avg of Box 2

= Percentage 2 (60%)

$SE_1 = \frac{SD \text{ of box } 1}{\sqrt{n_1}}$ (since there're only two kinds of numbers, we can use short cut formula)

Variability.

$= \frac{(1-0) \times \sqrt{\text{Fraction of 1} \times \text{Fraction of 0}}}{\sqrt{n_1}}$

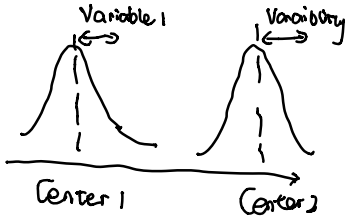
$= \frac{\sqrt{\text{Percentage 1} \times (100\% - \text{Percentage 1})}}{\sqrt{n_1}}$

$= \frac{\sqrt{50\% \times 50\%}}{\sqrt{100}} = 0.05 = 5\%$

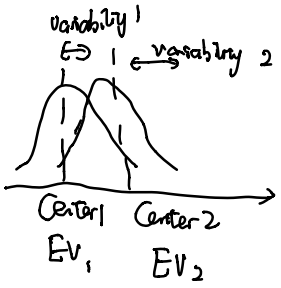
$SE_2 = \frac{\sqrt{\text{Percentage 2} \times (100\% - \text{Percentage 2})}}{\sqrt{n_2}}$

$= \frac{\sqrt{60\% \times 40\%}}{\sqrt{400}} = 2.45\%$

Don't need draw these picture:



they are different.



they are the same.

Difference:

EV for the difference $= |EV_1 - EV_2|$ ✓ $(60\% - 50\%) = 10\%$

$= |60\% - 50\%| = 10\%$

(Difference between Percentage 1 and Percentage 2)

SE for the difference $= \sqrt{SE_1^2 + SE_2^2} = \sqrt{5\%^2 + 2.45\%^2}$ ✓ ✓

$= 5.567\%$

($SE_D^2 = SE_1^2 + SE_2^2$

Variability for D = Sum of Variability)

95% Confidence Interval for the difference :

$(EV_D - 2 \times SE_D,\ EV_D + 2 \times SE_D)$

(Why use 2? From Q5. normal table, $z = 2$, middle area is 95%.)

$= (10\% - 2 \times 5.567\%,\ 10\% + 2 \times 5.567\%)$

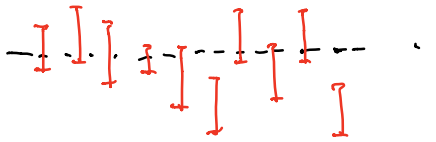
$= (-1.134\%,\ 21.134\%)$

2 explanation for 95% CI: (1')

✓ ① If we replicate the testing many many times, 95% of our confidence interval will include the true difference. ←

(5% of CI will not include true value. (we are wrong))

($148% $2.53%)
↓
New 95% CI



? ② Confidence level

if this confidence interval includes 0, we have 95% confidence to believe they are the same. ←

Since this 95% CI includes 0, we believe these two samples are the same.