Hypothesis testing with CLT (cont.)

Difference in means

10/31/24

Housekeeping

- No office hours tomorrow
- Daylight savings this weekend
- Data collection proposal due Monday 11/4 midnight!

Recap

- Test for difference in two proportions
  - Learned about $\hat{p}_{pooled}$
- Test for a single mean
  - z-test: we know σ, use standard Normal distribution
  - t-test: we don't know σ, use t distribution

Hypothesis test for mean paired difference

Paired data (recap)

- Recall paired data: we have two set of data $\boldsymbol{x}$ and $\boldsymbol{y}$ where each x_i has a corresponding to one y_i
  - Can obtain differences $d_i = y_i - x_i$
  - We are interested in the true mean difference μ_d
- Recall: if observational units are independent and the differences are approximately Normal, then CLT gives us:

$$\bar{d} \mathrel{\dot\sim} N\left(\mu_d, \frac{\sigma_d}{\sqrt{n}}\right)$$

- We don't typically know σ_d, so replace with sample s_d (and then use t distribution)

Hypothesis test

- Hypotheses: $H_0 : \mu_d = \mu_0$ versus $H_A : \mu_d \neq \mu_0$ (or $>$ or $<$)
- Obtain summary statistics $\bar{d}_{obs}$ and s_d
- Check if CLT holds. If so, what is our **null distribution**?

$$\bar{d} \dot{\sim} N\left(\mu_0, \frac{\sigma_d}{\sqrt{n}}\right)$$

- Because we don't know σ_d, our **test statistic** here is:

$$t = \frac{\bar{d}_{obs} - \mu_0}{\frac{s_d}{\sqrt{n}}} \sim t_{df}$$

where $df = n - 1$

Example: zinc (revisited)

Data consist of measured zinc concentrations in bottom water and surface water at 10 randomly sampled wells:

Do the data suggest that the true average concentration in the bottom water is greater than that of surface water? Let's now answer this using a hypothesis test at the 0.05 level.

- Define parameters and hypotheses
  - Let μ_d be the true mean difference between zinc concentrations (bottom-surface)
  - $H_0 : \mu_d = 0$ versus $H_A : \mu_d > 0$
- Last week, we saw conditions for CLT were satisfied

Example: zinc (cont.)

```r
zinc <- zinc |>
  mutate(d = bottom - surface)
d_bar <- mean(zinc$d)
d_bar
```

```
[1] 0.0804
```

```r
s_d <- sd(zinc$d)
s_d
```

```
[1] 0.05227321
```

Find the test-statistic

Example: zinc (cont.)

$$t = \frac{\bar{d}_{obs} - \mu_0}{s_d/\sqrt{n}} = \frac{0.0804 - 0}{0.052/\sqrt{10}} = 4.889 \sim t_9$$

- So our p-value is $\Pr(T \geq t) = \Pr(T \geq 4.889) = 1 - \texttt{pnorm}(4.889,\ 9) = 0$
- We reject H_0! The data provide convincing evidence that zinc concentrations of bottom well water is greater than those of surface water.

Hypothesis test for difference in means

Sampling distribution for difference in means

- Two populations, interest in $\mu_1 - \mu_2$ (or other order)
- Samples of size n_1 and n_2
- If CLT holds, we learned sampling distribution of difference in sample means is:

$$\bar{X}_1 - \bar{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

- When we don't know the population standard deviations, we replace the σ with s and use a t distribution
- Same thing will happen for hypothesis test!
  - Same conditions for inference: independence (extended) and approximate normality/large sample size (extended)

Hypothesis test

Hypotheses $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$ (or > or <)

- If CLT holds, our **null distribution** for the difference in sample means is:

$$\bar{X}_1 - \bar{X}_2 \dot{\sim} N\left(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

- In practice, use s_1 and s_2. So our **test-statistic** is…

$$t = \frac{\text{point est} - \text{null value}}{\widehat{\text{SE}}_0} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

where $df = \min(n_1 - 1, n_2 - 1)$

Activity

Munchkins!