Since the Midterm: Inference

Given a population parameter, or a deterministic function of population parameters, we can test a statement about it (hypothesis test) or give an interval estimate (confidence interval) for it.

① Hypothesis test
- (i) proportion
- (ii) difference in proportions
- (iii) mean
- (iv) difference in means
- (v) slope in a linear model

② Confidence intervals
- (i) proportion
- (ii) difference in proportions
- (iii) mean
- (iv) difference in means
- (v) slope in a linear model
- (vi) the expected value of the response

③ prediction intervals

④ connection between hypothesis test and confidence intervals

① Hypothesis test

▨ Settings

(i) $H_0: \pi \overset{\geq}{=} \pi_0$ vs $H_A: \pi \overset{<}{>} \pi_0$

(ii) $H_0: \pi_1 - \pi_2 \overset{\geq}{=} 0$ vs $H_A: \pi_1 - \pi_2 \overset{<}{\neq} 0$

(iii) $H_0: \mu \overset{\geq}{=} \mu_0$ vs $H_A: \mu \overset{<}{>} \mu_0$

(iv) $H_0: \mu_1 - \mu_2 \overset{\geq}{=} 0$ vs $H_A: \mu_1 - \mu_2 \overset{<}{\neq} 0$

(v) $H_0: \beta_1 = 0$ vs $H_A: \beta_1 \overset{<}{>} 0$

▨ Procedure

goal: to decide whether to reject H_0 or not based on a sample from the population, at significance level α

1. identifying test statistic and its null distribution
   ↑ function of the sample

desired properties
- we want it to contain useful info about the population parameter of interest
  eg. $\bar{X}$ for $H_0: \mu = \mu_0$
  $\hat{p}$ for $H_0: \pi = \pi_0$
- we want its distribution to be known under H_0
  (so that we can compute useful probabilities under H_0)

eg. $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ under H_0 (for setting (iii))

2. computing observed test statistic

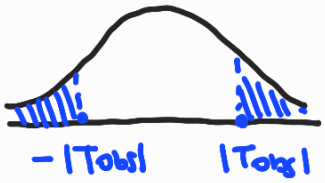
sample (random): $X_1, \ldots, X_n \rightarrow \frac{\bar{X} - \mu_0}{S/\sqrt{n}} =: T$ ← test statistic
(random)

observed sample: $x_1, \ldots, x_n \rightarrow \frac{\bar{x} - \mu_0}{s/\sqrt{n}} =: t_{obs}$
(fixed)
↑ observed test statistic

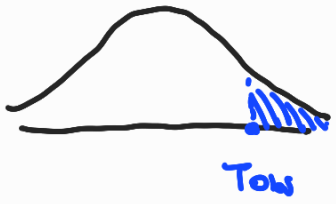
↑ e.g. for setting (iii)

3. knowing when to reject, given an observed sample
   ↳ by comparing the p-value with α
   (reject if the p-value is less than α)

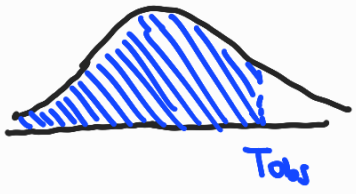
two-sided: $\mathbb{P}_{H_0}(|T| \geq |T_{obs}|)$
(e.g. $H_a: \mu \neq \mu_0$) $= \mathbb{P}_{H_0}(T \geq |T_{obs}|) + \mathbb{P}_{H_0}(T \leq -|T_{obs}|)$
$= 2\mathbb{P}_{H_0}(T \geq |T_{obs}|)$



one-sided, greater than: $\mathbb{P}_{H_0}(T \geq T_{obs})$
(e.g. $H_a: \mu > \mu_0$)



one-sided, less than: $\mathbb{P}_{H_0}(T \leq T_{obs})$
(e.g. $H_a: \mu < \mu_0$)



Summary of test statistics for different settings

	(i) $H_0: \pi = \pi_0$	(ii) $H_0: \pi_1 - \pi_2 = 0$	(iii) $H_0: \mu = \mu_0$	(iv) $H_0: \mu_1 - \mu_2 = 0$
Binomial test	$X \overset{H_0}{\sim} B(n, \pi_0)$			
z-test		$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1)$ pooled proportion if $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$, then $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$		
t-test			$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$ S: random (sample standard deviation)	(equal variance) $\frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$ S_p: pooled sd (Welch) $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{H_0}{\sim} t_\nu$ ν: complicated (paired) $\frac{\bar{D}}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$ $D_i = X_{1,i} - X_{2,i}$ S = sample standard deviation of $\{D_1, \dots, D_n\}$

	(V) $H_0: \beta_1 = \gamma$
t-test	$\frac{\hat{\beta}_1 - \gamma}{\hat{SE}[\hat{\beta}_1]} \overset{H_0}{\sim} t_{n-2}$

② Confidence intervals

- interval estimates for population parameter of interest

$P($ the parameter is in the $1-\alpha$ level (confidence interval) $) = 1-\alpha$

(confidence interval): random (before we plug in values computed from the observed sample)

- no notion of null hypothesis or alternative hypothesis

☆ We're not making any assumption on the parameter of interest

Terminology

$(1-\alpha)$ - confidence interval for γ: (γ: population parameter of interest)

$$(PE - CV \cdot \hat{SE},\ PE + CV \cdot \hat{SE})$$

$CV \cdot \hat{SE}$: MOE

PE: point estimate (estimate for γ)

CV: Critical value

↳ CV is found using confidence level $(=1-\alpha)$ and the distribution that the PE follows

- if $PE \sim N(\ ,\ )$ then $CV = z_{\frac{\alpha}{2}}$

- if $PE \sim t_v$ then $CV = t_{v,\frac{\alpha}{2}}$

$\hat{SE}$: estimate of the standard error of PE

↳ standard error of PE is often unknown, and thus we use an estimate of it

MOE: margin of error, $CV \cdot \hat{SE}$

Summary of CI for different settings

	(i) π	(ii) $\pi_1 - \pi_2$	(iii) μ	(iv) $\mu_1 - \mu_2$
z-CI (Wald CI)				
PE	$\hat{p}$	$\hat{p}_1 - \hat{p}_2$		
CV	$z_{\frac{\alpha}{2}}$	$z_{\frac{\alpha}{2}}$		
$\widehat{SE}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$		
	Let $\hat{p} = \frac{x}{n}$	Let $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$		
Agresti-Coull CI				
PE	$\tilde{p} := \frac{x+2}{n+4}$	$\tilde{p}_1 - \tilde{p}_2$, where $\tilde{p}_1 = \frac{x_1+1}{n_1+2}$	and $\tilde{p}_2 = \frac{x_2+1}{n_2+2}$	
CV	$z_{\frac{\alpha}{2}}$	$z_{\frac{\alpha}{2}}$		
$\widehat{SE}$	$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$	$\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$		$D_i = x_{1,i} - x_{2,i}$ S = sample standard deviation of $\{D_1, \dots, D_n\}$
				(paired)
t-CI			PE $\bar{x}$	PE $\bar{D}$
			CV $t_{n-1,\frac{\alpha}{2}}$	CV $t_{n-1,\frac{\alpha}{2}}$
			$\widehat{SE}$ $\frac{s}{\sqrt{n}}$	$\widehat{SE}$ $\frac{s}{\sqrt{n}}$
				(Welch)
				PE $\bar{x}_1 - \bar{x}_2$
				CV $t_{(v),\frac{\alpha}{2}}$ ← complicated
				$\widehat{SE}$ $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
				(equal variance)
				PE $\bar{x}_1 - \bar{x}_2$
				CV $t_{n_1+n_2-2,\frac{\alpha}{2}}$
				$\widehat{SE}$ $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ ← pooled sd

	(v) β_1		(vi) $E[Y_* \mid X_*]$
	PE $\hat{\beta}_1$	PE	$\hat{Y}_*$ ← predicted response for X_*
t-CI	CV $t_{n-2,\frac{\alpha}{2}}$	CV	$t_{n-2,\frac{\alpha}{2}}$
	SE $\widehat{SE}[\hat{\beta}_1]$	SE	$\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_*-\bar{X})^2}{\sum_{i=1}^n (X_i-\bar{X})^2}}$

$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (Y_i-\hat{Y}_i)^2}$

※ $E[Y_* \mid X_*]$ is a deterministic function of parameters

↳ $E[Y_* \mid X_*] = E[\beta_0 + X_*\beta_1 + \varepsilon_* \mid X_*]$ ($\varepsilon_* \sim N(0,\sigma^2)$)

$= \beta_0 + X_*\beta_1$

β_0: parameter; β_1: parameter; X_*: considered to be a fixed constant

③ Prediction intervals for Y_*

$$\left[\hat{Y}_* - t_{n-2,\frac{\alpha}{2}}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_*-\bar{X})^2}{\sum_{i=1}^n (X_i-\bar{X})^2}},\ \hat{Y}_* + t_{n-2,\frac{\alpha}{2}}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_*-\bar{X})^2}{\sum_{i=1}^n (X_i-\bar{X})^2}}\right]$$

Wider than the CI for $E[Y_* \mid X_*]$ due to the randomness in ε_*

$Y_* = \beta_0 + X_*\beta_1 + \varepsilon_*$ is a random variable, not a deterministic function of parameters. This is why it's called a prediction interval instead of a confidence interval

④ connection between hypothesis test and confidence interval

Given an observed sample, we reject H_0 in a two-sided hypothesis test for a parameter at significance level α

$\Leftrightarrow$ $(1-\alpha)$ level observed CI does not contain the value of the parameter specified in H_0.

Other remarks

[Linear regression]

- linear model

  Given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$, we assume

  $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{for } i \in [1, \dots, n],$$

  where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- Estimates of coefficients

  $\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$, where $r \in [-1, 1]$ is the sample correlation coefficient

  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

- Predicted response for X_i:

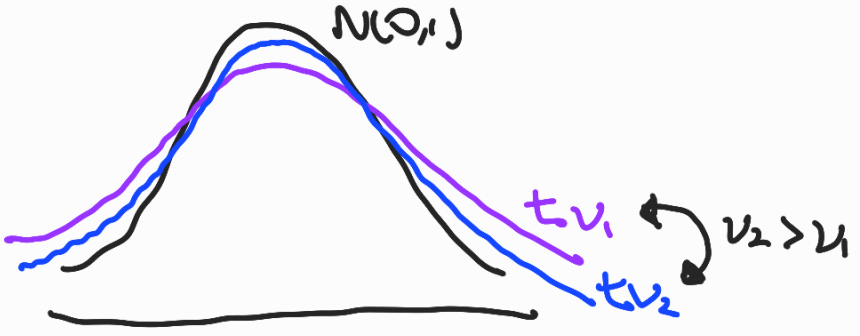
  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- Residual for i^{th} observation

  $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

- Knowing how to read the R output is important, e.g. where to find observed $\hat{\beta}_1$ and observed $\widehat{SE}[\hat{\beta}_1]$.

[normal vs t-distribution]



As the degrees of freedom of the t-distribution increases, the t-distribution approaches the normal distribution.

[Binomial distribution]

X is a binomial random variable if X can be thought of as the number of successes in a ① fixed number of trials. where the trials are ② independent and have the ③ same success probability.