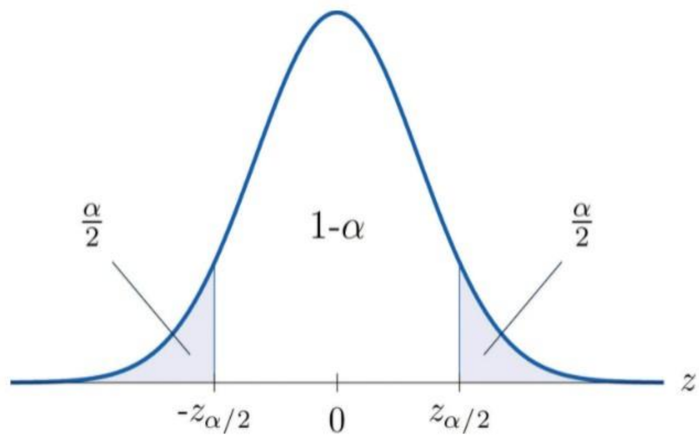
Tổng hợp lý thuyết MAS291 (Phần 3)

Chương 8:

1. Point estimate and confidence interval
    - Point estimate – điểm để ước lượng khoảng tin cậy trong 1 giả thuyết
    - Confidence interval – khoảng tin cậy: dùng để ước lượng độ tin cậy của một tham số nằm trong khoảng
    - Ví dụ về confidence interval $-z_{a/2}$ tới $z_{a/2}$

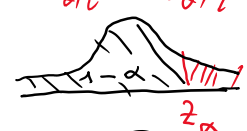


    - Critical value z_a (percentage point) là điểm giới hạn khoảng tin cậy trong giả thuyết. Trong hình trên có 2 critical value là $-z_{a/x}$ và $z_{a/2}$
2. Confidence interval for μ (Khoảng tin cậy của gái trị trung bình)
    - Khoảng tin cậy trong bài toán tìm khoảng tin cậy dành cho μ được chặn trên bởi l và chặn dưới bởi u với P(l <= μ <= u) = 1 - α (phần không tô đậm trong hình trên)
    1. Trong bài toán tìm khoảng tin cậy dành cho giá trị trung bình với phương sai cho trước (C.I. for μ if σ is known):
        i. Khi khoảng tin cậy bị giới hạn hai đầu (Two-sided confidence bound)

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

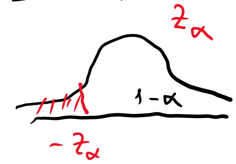
        ii. Khi khoảng tin cậy bị chặn trên (Upper confidence bound)

$$\mu \le \bar{x} + z_{\propto}\frac{\sigma}{\sqrt{n}}.$$



        iii. Khi khoảng tin cậy bị chặn dưới (Lower confidence bound)

$$\mu \ge \bar{x} - z_{\propto}\frac{\sigma}{\sqrt{n}}$$



⇨ Với $\bar{x}$ là mean của sample, $z_{\propto}$ là Critical value hay điểm giới hạn khoảng tin cậy, σ là độ lệch chuẩn (căn phương sai) đã cho biết trước và n là số lượng của sample

2. Trong bài toán tìm khoảng tin cậy dành cho giá trị trung bình với phương sai chưa cho trước (C.I. for μ if σ is unknown):
   i. Two-sided 100(1-α)% confidence interval on μ
   $$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
   ii. Upper confidence bound for μ
   $$\mu \le \bar{x} + t_{\propto;n-1}\frac{s}{\sqrt{n}}$$
   iii. Lower confidence bound for μ is
   $$\mu \ge \bar{x} - t_{\propto;n-1}\frac{s}{\sqrt{n}}.$$

   ⇨ Với $\bar{x}$ là mean của sample, $t_{\propto;n-1}$ là điểm giới hạn khoảng tin cậy trong t-distribution, σ không được biết trước nên ta thay bằng s là độ lệch chuẩn của sample và n là số lượng của sample

   ⇨ Thường các bài toán sử dụng t distribution sẽ có số lượng n > 30 và, $t_{\propto;n-1}$ cho biết trước. Việc cần làm là xác định bài toán thuộc trường hợp nào (2-sided, upper hay lower) sau đó tính point estimate $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ và số sánh t với $t_{\propto;n-1}$ hoặc $t_{\alpha/2,n-1}$ đã cho trước để đưa ra kết luận về khoảng tin cậy

3. Confidence interval for p (Bài toán kiểm định giả thuyết với xác suất)
   a. Two-sided 100(1-α)% confidence interval
   $$\hat{p} - z_{\propto/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\propto/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
   b. Upper confidence bound for μ
   $$\mathrm{p} \ge \hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
   c. Lower confidence bound for μ is
   $$\mathrm{p} \le \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   ⇨ Với $\hat{p} = \frac{x}{n}$ là ước lượng điển (point estimate) của phân phối xác suất p, z_α là điểm giới hạn khoảng tin cậy trong z-distribution và n là số lượng của sample

   ⇨ Cách làm: Tính z = $\frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}}$ là standard normal z- distribution. So sánh z với $z_{\propto/2}$ (với bài toán 2 phía) hoặc z_α với bài toán 1 phía rồi đưa ra kết luận về khoảng tin cậy

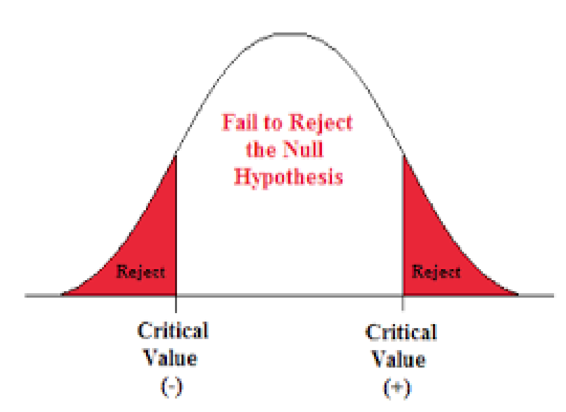
Chương 9: Test of hypotheses for a single sample (Kiểm định giả thuyết 1 mẫu)

1. Một số khái niệm trong bài toán kiểm định giả thuyết
    - H_0: null hypothesis (phát biểu đang được mặc định)
    - H_1: alternative hypothesis (phát biểu người thu thập dữ liệu muốn c/m)
    - Purpose: Nếu điểm z (point estimate) nằm trong khoảng tin cậy ($-z_{\frac{\alpha}{2}} \leq z \leq z_{\frac{\alpha}{2}}$) hoặc $z \leq z_\alpha$ hoặc $z \geq -z_\alpha$ (khoảng màu trắng của hình dưới đây) -> fail to reject H_0 -> fail to reject hoặc reject claim tuỳ từng trường hợp cụ thể
    - Một số lỗi thường gặp trong bài toán kiểm định giả thuyết
        - Type 1 error: H_0 true but reject H_0
        - Type 2 error: H_0 false but fail to reject H_0

Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	no error	type II error
Reject H_0	type I error	no error

    - p - value: smallest α that would lead to rejection of H_0



2. Kiểm định giả thuyết với giá trị trung bình và xác suất
    - Xem cách làm và công thức trong slides

Chương 10: Test of hypotheses for a two sample (Kiểm định giả thuyết 2 mẫu)

Vẫn tiếp tục tham khảo slides

Chương 11: Simple linear regression and Correlation (Hồi qui tuyến tính đơn giản và Hệ số tương quan)

1. Simple linear regression – Khái niệm về hồi qui tuyến tính
    - Phương trình tổng quát: $Y = \beta_0 + \beta_1 x$
    - Trong một mẫu có n điểm dữ liệu (x_i, y_i):
        - Xấp xỉ điểm (point estimates) của β_0, β_1, σ^2 được ký hiệu là $\hat{\beta}_0$, $\hat{\beta}_1, \hat{\sigma}^2$ (giống với xấp xỉ điểm của xác suất p được ký hiệu là $\hat{p}$)
        - Đường hồi qui tuyền tính có dạng $\hat{y} = \hat{\beta}_0$+$\hat{\beta}_1$x
2. Một số công thức tính xấp xỉ điểm trong hồi qui tuyến tính
    - $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
    - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
    - Với $S_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i{}^2 - \frac{(\sum x_i)^2}{n}$ và
    - $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

ANOVA

	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

    - SS Regression – $SS_R = \sum(\hat{y}_i - \overline{y})^2 = \hat{\beta}_1 S_{xy}$
    - SS Residual – $SS_E = \sum(y_i - \hat{y}_i)^2 = SS_T - SS_R$
    - SS Total - $SS_T = \sum(y_i - \overline{y})^2 = \sum y_i{}^2 - \frac{(\sum y_i)^2}{n}$
    - Unbiased estimator: $\hat{\sigma}^2 = \frac{SS_E}{n-2}$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

    - Coeficients slope β_1 is $\hat{\beta}_1$ (0.10977)
    - Coeficients intercept β_0 is $\hat{\beta}_0$ (98.24833)
    - Estimated standard error of the slope is $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ (0.03297)
    - Estimated standard error of the intercept is (58.03348)

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$$

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

    - Sample correlation coefficient R= $\frac{S_{xy}}{\sqrt{S_{xx} SS_T}}$