1. Тест для пропорций

Если $\hat{p}$ - это наблюдаемая пропорция, то используются следующие тестовые статистики :

Для проверки гипотезы $H_0 : p = p_0$, используем

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1).$$

Шаги проверки гипотез:

1. **Определить H_0** *и альтернативную гипотезы, и задать уровень доверия, например* $\alpha = 0.05$.
2. **Собрать данные из выборки для тестовой статистики**
3. *Предполагая справедливость нуль-гипотезы найти p-value*

P-value это вероятность наблюдения выборки в предположении справедливости нуль-гипотезы

4. *Если* p-value ≥ α, не отвергаем H_o , если p-value < α, *отвергаем* H_o

2. Сравнение двух пропорций

Кто пьет больше, мужчины или женщины? Для мужчин пропорция p_1 для женщин p_2.

Нулевая и альтернативная гипотезы

$$H_0 : p_1 = p_2, \; H_1 : p_1 \neq p_2$$

Используя нормальную аппроксимацию для выборочных пропорций имеем

$\hat{p}_1 \sim N(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}})$ и $\hat{p}_2 \sim N(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}})$

Если мы рассматриваем разницу двух независимых нормальных распределений, то

Если $Z_1 \sim N(\mu_1, \sigma_1), Z_2 \sim N(\mu_2, \sigma_2)$ независимы, то

$$Z_1 - Z_2 \sim N(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}).$$

И следовательно

$$\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}).$$

Принимая H_0, поскольку мы полагаем $p_1 = p_2$,

То тестовая статистика равна

$(\hat{p}_1 - \hat{p}_2) / \sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2} \sim N(0,1)$

Где $\hat{p}$ - это наблюдаемая пропорция из комбинированной выборки

Аналогично 95% доверительный интервал для $p_1 - p_2$ равен

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Пример данные о сильно пьющих:

	N	**X**	$\hat{p}$
Мужчины	7180	1630	0.227
Женщины	9916	1684	0.170
	17096	3314	0.194

Проверяем гипотезы $H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2$

$$\frac{0.227 - 0.170}{\sqrt{0.194 \times (1-0.194)(1/7180 + 1/9916)}} = 9.34.$$

Отвергаем H_0 с вероятностью p-value <0.001.

Доверительный интервал для разницы

95% CI is $0.05719 \pm 1.96\sqrt{\frac{0.227 \times 0.773}{7180} + \frac{0.170 \times 0.830}{9916}}$ = (0.045, 0.069)

Пример 3: Исследуется эффект фюоридизации воды на пропорцию рождаемости мальчиков и девочек. Из 260 детей рожденных в госпитале за исследуемый период 124 – мальчики. Официальная статистика говорит что до начала флюоридизации воды процент рождаемости мальчиков был 53%.
Какое заключение можно сделать на основе этих данных?

p=вероятность рождения мальчика

H_0:*p*=0.53, H_a:*p*≠ 0.53.

$\hat{p} = 124/260 = 0.477.$

$$z = \frac{0.477 - 0.53}{\sqrt{\frac{0.53 \times 0.47}{260}}} = -1.71.$$

P(Z<-1.71)=0.0436.

Не отвергаем гипотезу H_0 с *P*-value=0.0872.