

Tests d'hypothèses

Goual Hafida

UBM Annaba

June 5, 2020

Table of contents

Une hypothèse est une assertion sur un paramètre.

- La population mean $\mu = 1$.
- La proportion de succès $p = .4$.
- La moyenne de la population $\mu > 2$.
- La différence entre deux populations signifie $\mu_1 - \mu_2 = 0$.
- La différence entre deux populations signifie $\mu_1 - \mu_2 > 3$.

En règle générale, il existe deux types d'hypothèses dans le cadre de test d'hypothèses:

- 1 Hypothèse nulle: ce que l'on croit avant le test. Par exemple

$$H_0 : \mu_1 = \mu_2,$$

$$H_0 : \mu_1 = 3,$$

$$H_0 : \mu_1 - \mu_2 = 4.$$

- 2 Hypothèse alternative: une hypothèse contradictoire au nul

$$H_A : \mu_1 \neq \mu_2$$

$$H_A : \mu_1 > \mu_2.$$

Deux utilisations principales du test d'hypothèse

L'hypothèse nulle est celle sur laquelle les gens se concentrent dans le test d'hypothèse classique et il existe deux constructions logiques du test d'hypothèse.

- Confirmer une théorie: en physique, on peut croire que la force est égale à la masse multipliée par l'accélération

$$F = ma$$

on peut mesurer pour des objets de masses diverses la force et l'accélération et utiliser comme hypothèse nulle et hypothèse alternative

$$H_0 : F - ma = 0$$

$$H_A : F - ma \neq 0.$$

Cette confirmation d'une hypothèse est assez courante et dans ce cas, on souhaite généralement que le null ne soit pas rejeté ou qu'il existe des preuves solides du null.

- Répudier un contrôle:
Une utilisation plus courante des tests d'hypothèse consiste à définir le null comme contrôle et à montrer qu'il existe des preuves pour le rejeter. Un exemple est que le carburant de fusée rend les voitures plus rapides. Dans ce cas, on peut prendre des vitesses de voitures dopées au carburant de fusée et calculer la moyenne de la population, μ_1 , et comparer cela à la moyenne de la population de voitures sans carburant de fusée, μ_2 , et utiliser l'hypothèse nulle et alternative suivante

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 > \mu_2.$$

Dans ce cas, nous aimerions rejeter le nul ou qu'il existe de fortes preuves contre le nul

Deux types d'hypothèses

Deux types d'hypothèses sont simples et composites:

- Simple: Une hypothèse simple est celle où la distribution sous l'hypothèse est entièrement spécifiée. Par exemple $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2 = 5)$ et l'hypothèse est

$$H : \mu = 10.$$

- Composite: Pour une hypothèse composite, la distribution sous l'hypothèse n'est pas entièrement spécifiée. Par exemple $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2 = 5)$ et l'hypothèse est

$$H : \mu > 10.$$

Il existe une variété de distributions qui satisferont à cela, toute normale avec une moyenne supérieure à 10. Généralement, le null est simple et l'alternative est composite.

Les procédures de tests

Une procédure de test comprend les éléments suivants

- Données: échantillons tirés d'une distribution

$$X_1, \dots, X_n \stackrel{iid}{\sim} p(x|\theta)$$

- L'hypothèse nulle: elle détermine le paramètre auquel on s'intéresse dans la distribution. Par exemple

$$H_0 : \mu = 20.$$

- Statistique de test: une statistique de test qui est utilisée pour estimer le paramètre auquel on s'intéresse à partir de la distribution ci-dessus. Par exemple

$$\bar{X} = \frac{1}{n} \sum_i X_i.$$

- Région de rejet: valeurs de la statistique de test pour lesquelles H_0 sera rejeté.

Il existe deux types d'erreurs dans les tests d'hypothèses.

- Type I: Rejeter l'hypothèse nulle, H_0 , lorsqu'elle est vraie. La probabilité de ce type d'erreur est désignée par

$$\alpha = \Pr(H_0 \text{ is rejected when true}).$$

- Type II: Ne pas rejeter l'hypothèse nulle, H_0 lorsqu'elle est fausse. La probabilité de ce type d'erreur est désignée par

$$\beta = \Pr(H_0 \text{ not is rejected when false}).$$

$$X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2 = 10).$$

$$H_0 : \mu = 10,$$

$$H_A : \mu = 20.$$

Disons que μ est vraiment 10 et nous sélectionnons une région de rejet de $\bar{X} > 15$. Cela signifie que pour tout $\bar{X} > 15$ supérieur à 15, nous rejetons le null.

Nous pouvons calculer α et β pour ce cas comme suit:

$$\alpha = \Pr(\bar{X} > 15 | X \sim \text{No}(\mu = 10, \sigma^2 = 10)),$$

$$\beta = \Pr(\bar{X} \leq 15 | X \sim \text{No}(\mu = 20, \sigma^2 = 20)).$$

Aucune région de rejet ne peut rendre α et β égaux à zéro. Nous recherchons donc une région de rejet qui contrôle les deux simultanément.

- Obtenez les données d'une distribution spécifiée.
- Énoncer des hypothèses nulles et alternatives.
- Indiquez la statistique de test.
- Indiquez l'erreur de type I acceptable, α . C'est ce qu'on appelle la valeur critique α .
- Calculez la région de rejet en fonction de ce qui précède.
- Vérifiez si la statistique de test se situe dans la région de rejet et rejetez la valeur Null sur cette base.

Le cas le plus typique

Le cas le plus typique n'a pas de *beta* unique. Cela est dû au fait que le test d'hypothèse typique est comme ci-dessus, mais

$$H_0 : \mu = 10,$$

$$H_A : \mu \neq 20.$$

Disons que μ est vraiment 10 et nous sélectionnons une région de rejet de $\bar{X} > 15$.
Le calcul de α est le même mais β n'est plus unique

$$\alpha = \Pr(\bar{X} > 15 | X_i \sim \text{No}(\mu = 10, \sigma^2 = 10)),$$

$$\beta = \Pr(\bar{X} \leq 15 | X_i \sim \text{No}(\mu > 10, \sigma^2 = 20)).$$

Population normale σ connue

En règle générale, le paramètre σ n'est pas connu, mais cela décrit les idées. Si $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2)$ avec σ connu. Les hypothèses nulles et alternatives sont

$$\begin{aligned}H_0 &: \mu = \mu_1, \\H_A &: \mu > \mu_1.\end{aligned}$$

Nous voulons calculer les régions de rejet pour ce test.

Nous devons d'abord spécifier le niveau d'erreur de type I ou le α critique du test, disons $\alpha = .05$. Nous spécifions la statistique de test comme \bar{X} . Nous maintenant besoin de calculer la valeur ℓ pour laquelle

$$.05 = \Pr(\bar{X} > \ell | X_i \sim \text{No}(\mu = \mu_1, \sigma^2)).$$

Pour ce faire, nous avons besoin de la "distribution de la statistique de test sous l'hypothèse nulle".

Population normale σ connue

Si l'hypothèse nulle est vraie et que les données proviennent d'une normale avec σ connue, alors nous savons que la statistique de test suivante est distribuée comme une normale standard

$$z = \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \sim \text{No}(0, 1).$$

A partir de cela, nous pouvons calculer ℓ dans ce qui suit

$$.05 = \Pr(\bar{X} > \ell | X_i \sim \text{No}(\mu = \mu_1, \sigma^2)),$$

comme $z_{.05}$. Notre région de rejet est donc $z > z_{.05}$.

Population normale σ connue

Si $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2)$ avec σ connue.

Hypothèse nulle $H_0 : \mu = \mu_1$.

Statistique de test $z = \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}}$.

- $H_A : \mu > \mu_1$: la région de rejet est $z \geq z_\alpha$.
- $H_A : \mu < \mu_1$: la région de rejet est $z \leq -z_\alpha$.
- $H_A : \mu \neq \mu_1$: la région de rejet est $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

- Identifiez le paramètre d'intérêt, la moyenne ou la variance ou le paramètre de taux.
- Déterminez la valeur nulle et énoncez les hypothèses nulles et alternatives.
- Donnez la formule de la valeur calculée de la statistique de test et branchez la valeur nulle et d'autres paramètres connus à cette étape, par exemple, μ_1 et σ dans ce qui précède.
- Sur la base de la distribution des statistiques ci-dessus, calculer la région de rejet pour un niveau de signification sélectionné α .
- Rejeter ou accepter H_0 .

Population normale σ inconnue

Si $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2)$ avec σ v. Les hypothèses nulles et alternatives sont

$$H_0 : \mu = \mu_1,$$

$$H_A : \mu > \mu_1.$$

Nous voulons calculer les régions de rejet pour ce test.

Nous devons d'abord spécifier le niveau d'erreur de type I ou le α critique du test, disons $\alpha = .05$. Nous spécifions la statistique de test comme \bar{X} . Nous devons maintenant calculer la valeur ℓ pour laquelle

$$.05 = \Pr(\bar{X} > \ell | X_i \sim \text{No}(\mu = \mu_1, \sigma^2)).$$

Pour ce faire, nous avons besoin de la " distribution de la statistique de test sous l'hypothèse nulle ".

Population normale σ inconnue

Si l'hypothèse nulle est vraie et que les données proviennent d'une normale avec σ inconnue, alors nous savons que la statistique de test suivante est distribuée comme une distribution t avec $n - 1$ degrés de liberté

$$t = \frac{\bar{X} - \mu_1}{S/\sqrt{n}} \sim \text{t-dist}_{n-1},$$

où S^2 est l'estimation de la variance de l'échantillon.

Nous procédons exactement comme avant mais utilisons la distribution t pour calculer la région de rejet.

Population normale σ inconnue

Si $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2)$ avec σ inconnue.

Hypothèse nulle $H_0 : \mu = \mu_1$.

Statistique de test $t = \frac{\bar{X} - \mu_1}{s/\sqrt{n}}$.

- $H_A : \mu > \mu_1$: la région de rejet est $z \geq t_{\alpha, n-1}$.
- $H_A : \mu < \mu_1$: la région de rejet est $z \leq -t_{\alpha, n-1}$.
- $H_A : \mu \neq \mu_1$: la région de rejet est $t \leq -t_{\alpha/2, n-1}$ or $t \geq t_{\alpha/2, n-1}$.

La distribution t est valide indépendamment de la taille de l'échantillon. Cela vaut pour les petits n .

Tests sur grand échantillon

Si $X_1, \dots, X_n \stackrel{iid}{\sim} p(\theta)$ où la distribution a borné la moyenne $\mu < \infty$ et la variance $\sigma^2 < \infty$.
Les hypothèses nulles et alternatives sont

$$H_0 : \mu = \mu_1,$$

$$H_A : \mu > \mu_1.$$

Nous voulons calculer les régions de rejet pour ce test.

Nous devons d'abord spécifier le niveau d'erreur de type I ou le α critique du test, disons $\alpha = .05$. Nous spécifions la statistique de test comme \bar{X} . Nous devons maintenant calculer la valeur ℓ pour laquelle

$$.05 = \Pr(\bar{X} > \ell | X_i \sim p(\theta)).$$

Pour ce faire, nous avons besoin de la " distribution de la statistique de test sous l'hypothèse nulle ".

Si l'hypothèse nulle est vraie, alors la statistique de test du théorème de la limite centrale est distribuée comme normale standard

$$z = \frac{\bar{X} - \mu_1}{S/\sqrt{n}} \sim \text{No}(0, 1),$$

ou S^2 est l'estimation de la variance de l'échantillon.

Nous procédons exactement comme avant en utilisant les valeurs z_α .

Tests sur grand échantillon

Si $X_1, \dots, X_n \stackrel{iid}{\sim} p(\theta)$ avec moyenne et variance bornées.

Hypothèse nulle $H_0 : \mu = \mu_1$.

Statistique de test $z = \frac{\bar{X} - \mu_1}{s/\sqrt{n}}$.

- $H_A : \mu > \mu_1$: la région de rejet est $z \geq z_\alpha$.
- $H_A : \mu < \mu_1$: la région de rejet est $z \leq -z_\alpha$.
- $H_A : \mu \neq \mu_1$: la région de rejet est $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

Test de grand échantillon redus

If $X_1, \dots, X_n \stackrel{iid}{\sim} p(\theta)$ avec moyenne et variance bornées.

Hypothèse nulle $H_0 : \theta = \theta_1$.

Statistique de test idéale

$$z = \frac{\hat{\theta} - \theta_1}{\sigma_{\hat{\theta}}},$$

mais $\sigma_{\hat{\theta}}$ est inconnu alors supposez $s_{\hat{\theta}} \approx \sigma_{\hat{\theta}}$ and use

$$z = \frac{\hat{\theta} - \theta_1}{s_{\hat{\theta}}}.$$

Application: proportion de la population

Étant donné une variable aléatoire $X \sim \text{bin}(p, n)$ avec n connu, nous voulons appliquer le cadre de test d'hypothèse à $H_0 : p = p_1$.

Notre estimation est $\hat{p} = \frac{X}{n}$ et nous savons p sous l'hypothèse nulle donc $\sigma_{\hat{p}} = \sqrt{p_1(1 - p_1)/n}$ ce qui se traduit par la statistique de test

$$z = \frac{\hat{p} - p_1}{\sqrt{p_1(1 - p_1)/n}},$$

qui si $\min(np_1, n(1 - p_1)) > 10$ est normale standard.

Test de proportion de la population

Si $X_1, \dots, X_n \stackrel{iid}{\sim} p(\theta)$ avec moyenne et variance bornées.

Hypothèse nulle $H_0 : p = p_1$.

Statistique de test

$$z = \frac{\hat{p} - p_1}{\sqrt{p_1(1 - p_1)/n}}.$$

- $H_A : p > p_1$: la région de rejet est $z \geq z_\alpha$.
- $H_A : p < p_1$: la région de rejet est $z \leq -z_\alpha$.
- $H_A : p \neq p_1$: la région de rejet est $z \leq -z_\alpha/2$ or $z \geq z_\alpha/2$.

Dans le cadre de test d'hypothèse, nous avons considéré le cas du calcul de la région de rejet étant donné une valeur α .

La valeur de p est l'échantillon ou le niveau de signification observé. Il indique que si nous utilisons l'exemple de valeur statistique pour déterminer la région de rejet, quelle est la valeur *alpha*.

Exemple

Si $X_1, \dots, X_n \stackrel{iid}{\sim} \text{No}(\mu, \sigma^2)$ avec σ connue. Les hypothèses nulles et alternatives sont

$$\begin{aligned}H_0 &: \mu = \mu_1, \\H_A &: \mu > \mu_1.\end{aligned}$$

Pour un ensemble de données particulier, nous pouvons calculer les éléments suivants

$$z_{\text{data}} = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}},$$

on peut alors calculer

$$\text{p-value} = \Pr(z > z_{\text{data}})$$

qui est l'erreur de type I si nous définissons la région critique sur z_{data} .

Pour les tests d'hypothèse utilisant la statistique z .

- Test unilatéral droite: $P = 1 - \Phi(z_{\text{data}})$.
- Test unilatéral gauche: $P = \Phi(z_{\text{data}})$.
- Test bilatéral: $P = 2[1 - \Phi(|z_{\text{data}}|)]$.

Definition

La **P-value** est le plus petit niveau de signification auquel H_0 serait rejeté lorsque la procédure de test spécifiée est appliquée à un ensemble de données donné. Une fois la valeur P déterminée, on peut rejeter ou accepter l'hypothèse nulle à un niveau α en comparant la valeur P à alpha

- Rejeter H_0 : $P\text{-value} \leq \alpha$
- Ne rejetez pas H_0 : $P\text{-value} > \alpha$.

Definition

La **P-value** est la probabilité, en supposant que H_0 est vraie, d'obtenir une valeur statistique de test au moins aussi contradictoire à H_0 que la valeur obtenue sur les données. Plus la valeur P est petite, plus les données à H_0 sont contradictoires.

Tests de deux échantillons

Une situation très courante est la configuration suivante

1

$$X_1, \dots, X_m \stackrel{iid}{\sim} p_1(\theta_1).$$

2

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} p_2(\theta_2).$$

3

X et Y sont indépendants.

Questions:

$$\mu_1 - \mu_2 = \Delta_0$$

$$\mu_1 - \mu_2 > \Delta_0$$

$$\mu_1 - \mu_2 < \Delta_0$$

$$\mu_1 - \mu_2 \neq \Delta_0.$$

Population normale σ connue

$$X_1, \dots, X_m \stackrel{iid}{\sim} \text{No}(\mu_1, \sigma_1^2)$$

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{No}(\mu_2, \sigma_2^2)$$

Les hypothèses nulles et alternatives sont

$$H_0 : \mu_1 - \mu_2 = \Delta_0,$$

$$H_A : \mu_1 - \mu_2 > \Delta_0.$$

Nous voulons calculer les régions de rejet pour ce test.

Nous devons d'abord spécifier le niveau d'erreur de type I ou le α critique du test, disons $\alpha = .05$. Nous spécifions la statistique de test comme $\bar{X} - \bar{Y}$. Nous devons maintenant calculer la valeur ℓ pour laquelle

$$.05 = \Pr(\bar{X} - \bar{Y} > \ell | X_i \sim \text{No}(\mu = \mu_1, \sigma_1^2) \text{ and } Y_j \sim \text{No}(\mu = \mu_2, \sigma_2^2)).$$

Pour ce faire, nous avons besoin de la "distribution de la statistique de test sous l'hypothèse nulle".

Population normale σ connue

Si l'hypothèse nulle est vraie et que les données proviennent d'une normale avec σ connue, alors nous savons que la statistique de test suivante est distribuée comme une normale standard

$$z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \text{No}(0, 1).$$

Comme nous l'avons fait dans le cas du test à un échantillon, nous utilisons des valeurs z_α .

$$X_1, \dots, X_m \stackrel{iid}{\sim} \text{No}(\mu_1, \sigma_1^2)$$

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{No}(\mu_2, \sigma_2^2)$$

Hypothèse nulle $H_0 : \mu_1 - \mu_2 = \Delta_0$.

Statistique de test

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \text{No}(0, 1).$$

- $H_A : \mu_1 - \mu_2 > \Delta_0$: la région de rejet est $z \geq z_\alpha$.
- $H_A : \mu_1 - \mu_2 < \Delta_0$: la région de rejet est $z \leq -z_\alpha$.
- $H_A : \mu_1 - \mu_2 \neq \Delta_0$: la région de rejet est $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

$$X_1, \dots, X_m \stackrel{iid}{\sim} p_1(\theta_1).$$

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} p_2(\theta_2).$$

avec moyenne et variance bornées pour les deux distributions.

Hypothèse nulle $H_0 : \mu_1 - \mu_2 = \Delta_0$.

Statistique de test

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \sim \text{No}(0, 1).$$

- $H_A : \mu_1 - \mu_2 > \Delta_0$: la région de rejet est $z \geq z_\alpha$.
- $H_A : \mu_1 - \mu_2 < \Delta_0$: la région de rejet est $z \leq -z_\alpha$.
- $H_A : \mu_1 - \mu_2 \neq \Delta_0$: la région de rejet est $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

Population normale σ inconnue

$$X_1, \dots, X_m \stackrel{iid}{\sim} \text{No}(\mu_1, \sigma_1^2)$$

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{No}(\mu_2, \sigma_2^2)$$

mais σ_1 et σ_2 sont inconnues. Les hypothèses nulles et alternatives sont

$$H_0 : \mu_1 - \mu_2 = \Delta_0,$$

$$H_A : \mu_1 - \mu_2 > \Delta_0.$$

Nous voulons calculer les régions de rejet pour ce test.

Nous devons d'abord spécifier le niveau d'erreur de type I ou le α critique du test, disons $\alpha = .05$. Nous spécifions la statistique de test comme $\bar{X} - \bar{Y}$. Nous devons maintenant calculer la valeur ℓ pour laquelle

$$.05 = \Pr(\bar{X} - \bar{Y} > \ell | X_i \sim \text{No}(\mu = \mu_1, \sigma_1^2) \text{ and } Y_j \sim \text{No}(\mu = \mu_2, \sigma_2^2)).$$

Pour ce faire, nous avons besoin de la "distribution de la statistique de test sous l'hypothèse nulle".

Population normale σ inconnue

Si l'hypothèse nulle est vraie et que les données proviennent d'une normale avec σ inconnue, alors nous savons que la statistique de test suivante est distribuée comme une distribution t avec une formule très compliquée pour les degrés de liberté

$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}},$$

ou S_1^2 and S_2^2 sont les estimations des variances d'échantillon.
Les degrés de liberté ν est

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}.$$

Remarque pour $s_1 \approx s_2$ et $m \approx n$ alors $\nu \approx 2n$.

$$X_1, \dots, X_m \stackrel{iid}{\sim} \text{No}(\mu_1, \sigma_1^2)$$

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{No}(\mu_2, \sigma_2^2)$$

mais σ_1 et σ_2 sont inconnues.

Hypothèse nulle $H_0 : \mu_1 - \mu_2 = \Delta_0$.

Statistique de test

$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}},$$

- $H_A : \mu_1 - \mu_2 > \Delta_0$: la région de rejet est $t \geq t_{\alpha, \nu}$.
- $H_A : \mu_1 - \mu_2 < \Delta_0$: la région de rejet est $t \leq -t_{\alpha, \nu}$.
- $H_A : \mu_1 - \mu_2 \neq \Delta_0$: la région de rejet est $t \leq -t_{\alpha/2, \nu}$ or $t \geq t_{\alpha/2, \nu}$.