Démarche statistique
Premiers pas avec R

David Causeur
Agrocampus Ouest
IRMAR CNRS UMR 6625
http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/david.causeur

Plan

Plan

Plan

Comparaison de deux groupes

Test de l'effet d'un facteur à deux modalités :

$$\begin{cases} \mathrm{H}_0 : \ \delta = \mu_1 - \mu_2 = 0 \\ \mathrm{H}_1 : \ \delta = \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Test de comparaison entre sexes dans `R`

Test de Student

Reprenons la statistique de Fisher :

$$F = \frac{n_1(\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet})^2 + n_2(\overline{Y}_{2\bullet} - \overline{Y}_{\bullet\bullet})^2}{\hat{\sigma}^2},$$

avec

$$\begin{aligned}
n_1(\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet})^2 &= n_1\left(\overline{Y}_{1\bullet} - \frac{n_1\overline{Y}_{1\bullet} + n_2\overline{Y}_{2\bullet}}{n_1+n_2}\right)^2, \\
&= n_1 n_2^2\left(\frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{n_1+n_2}\right)^2, \\
n_2(\overline{Y}_{2\bullet} - \overline{Y}_{\bullet\bullet})^2 &= n_2 n_1^2\left(\frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{n_1+n_2}\right)^2.
\end{aligned}$$

Test de Student

Reprenons la statistique de Fisher :

$$
\begin{aligned}
F &= \frac{n_1(\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet})^2 + n_2(\overline{Y}_{2\bullet} - \overline{Y}_{\bullet\bullet})^2}{\hat{\sigma}^2}, \\
&= \Big(\frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{\hat{\sigma}}\Big)^2 \frac{n_1 n_2 (n_1 + n_2)}{(n_1 + n_2)^2}, \\
&= \Big(\underbrace{\frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}_{T}\Big)^2.
\end{aligned}
$$

Test de Student

Reprenons la statistique de Fisher :

$$F = T^2, \text{ où } T = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}}.$$

En effet,

$$\begin{aligned} \sigma^2_{\hat{\delta}} &= \mathrm{Var}(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}), \\ &= \mathrm{Var}(\overline{Y}_{1\bullet}) + \mathrm{Var}(\overline{Y}_{2\bullet}), \\ &= \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \end{aligned}$$

Test de Student

Reprenons la statistique de Fisher :

$$F = T^2, \text{ où } T = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}}.$$

Soit $\hat{\theta}$ un estimateur de θ et $\hat{\sigma}_{\hat{\theta}}$ un estimateur de l'écart-type de $\hat{\theta}$.

Pour le test de $H_0 : \theta = \theta_0$, $T_{\theta_0} = (\hat{\theta} - \theta_0)/\hat{\sigma}_{\hat{\theta}}$ est appelée **statistique de Student**.

Test de Student

- Statistique du test de Student dans R

La distribution de T sous l'hypothèse nulle est la **loi de Student** à $n-2$ degrés de liberté, notée $\mathcal{T}_{n-2}$.

- p-value du test de Student dans R

Test unilatéral

Test de $H_0 : \delta = 0$ contre $H_1 : \delta < 0$.

Règle **unilatérale** de décision : H_0 est rejetée si T prend une valeur jugée anormalement faible sous H_0.

Ici, la p-value est la probabilité pour qu'une variable aléatoire de loi $\mathcal{T}_{n-2}$ soit plus petite que la valeur observée de T :

- ▶ Test de Student unilatéral dans `R`

Intervalle de confiance d'un paramètre

L'ensemble des valeurs θ_0 telles que $\mathrm{H}_0 : \theta = \theta_0$ n'est pas rejetée au seuil α est un intervalle de confiance de θ, de niveau de confiance $1 - \alpha$.

On en déduit l'intervalle de confiance $\mathrm{CI}_{1-\alpha}(\delta)$:

$$\begin{aligned}
\mathrm{CI}_{1-\alpha}(\delta) &= \left\{\delta_0,\ -t_{1-\alpha/2}^{(n-2)} \leq \frac{\hat{\delta} - \delta_0}{\hat{\sigma}_{\hat{\delta}}} \leq t_{1-\alpha/2}^{(n-2)}\right\}, \\
&= \left[\hat{\delta} - t_{1-\alpha/2}^{(n-2)}\hat{\sigma}_{\hat{\delta}} ; \hat{\delta} + t_{1-\alpha/2}^{(n-2)}\hat{\sigma}_{\hat{\delta}}\right].
\end{aligned}$$

- Intervalle de confiance par `t.test` dans `R`

Intervalle de confiance d'un paramètre

Pour les paramètres μ_1 et μ_2 :

$$\mathrm{CI}_{1-\alpha}(\mu_i) = \left[\hat{\mu}_i - t_{1-\alpha/2}^{(n-2)} \frac{\hat{\sigma}}{\sqrt{n_i}} ; \hat{\mu}_i + t_{1-\alpha/2}^{(n-2)} \frac{\hat{\sigma}}{\sqrt{n_i}}\right].$$

- Intervalle de confiance par `t.test` dans `R`

Puissance du test de Student

Dans quelle mesure le test de Student peut-il détecter une différence de moyennes à l'échelle de la population ?

Cas des données `porcs` : si la différence de moyenne entre mâles et femelles à l'échelle de la population vaut 1, le test de Student concluera-t'il que l'effet *sexe* est significatif ?

Puissance du test de Student

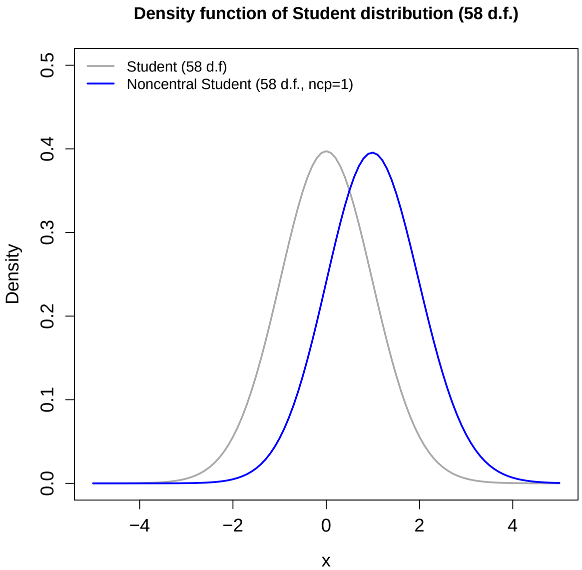
Soit le test de H_0 : $\mu_1 - \mu_2 = 0$ contre H_1 : $\mu_1 - \mu_2 = \delta \neq 0$.
Au seuil α, la **puissance** du test est la probabilité sous H_1 que le test rejette H_0 :

$$\text{Puissance}(\delta) \quad = \quad \mathbb{P}_{\mu_1 - \mu_2 = \delta}(|T| \geq t^{(n-2)}_{1-\alpha/2}).$$

Sous H_1, $T \sim \mathcal{T}_{n-2}(\lambda)$ avec :

$$\lambda \quad = \quad \frac{\delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$ [paramètre de non-centralité]

Puissance du test de Student



Puissance du test de Student

Plus λ est grand, plus le test de Student est puissant :

- Plus $|\delta|/\sigma$ est grand, plus le test est puissant.
- Plus n_1 et n_2 sont grands, plus le test est puissant.

▶ Puissance du test de Student dans R

Tests Post-hoc

L'effet d'un facteur est significatif = les moyennes de la variable réponse dans *certains* groupes sont différentes.

Quels groupes ?

$P_0 \neq P_{25} \neq P_{50}$ ou $P_0 \neq \{P_{25} = P_{50}\}$ ou ...

Tests post-hoc pour 3 groupes : 3 tests simultanés

- $H_0^{(12)}$: $\mu_1 = \mu_2$,
- $H_0^{(13)}$: $\mu_1 = \mu_3$,
- $H_0^{(23)}$: $\mu_2 = \mu_3$.

▶ Comparaisons deux à deux dans R

Tests Post-hoc

Quel risque d'erreur pour plusieurs tests simultanés ?

Probabilité d'au moins un rejet à tort (faux positif) :

$$\begin{aligned} & \mathbb{P}_{\mathrm{H}_0^{(ii')},\, i \neq i'}(\text{au moins une hypothèse } \mathrm{H}_0^{(ii')} \text{ rejetée}) \\ \leq \quad & \sum_{i \neq i'} \mathbb{P}_{\mathrm{H}_0^{(ii')}}(\mathrm{H}_0^{(ii')} \text{ rejetée}), \\ \leq \quad & 3\alpha \end{aligned}$$

Correction de Bonferroni : $\alpha^* = \alpha/3$:

▶ Correction de Bonferroni dans R

Tests sur un paramètre - Résumé

1. Test de $H_0 \ : \ \theta = \theta_0$ contre $H_1 \ : \ \theta \neq \theta_0$ ou $\theta > \theta_0$
   - Statistique du test de Student : $T = (\hat{\theta} - \theta_0)/\hat{\sigma}_{\hat{\theta}}$
   - Loi sous H_0 : $T \sim \mathcal{T}_k$, où k est le nombre de dégrés de liberté pour l'estimation de $\hat{\sigma}^2_{\hat{\theta}}$
2. Intervalle de confiance de niveau $1 - \alpha$: ensemble des θ_0 tels que H_0 n'est pas rejetée au seuil α
3. Comparaison de moyennes : $H_0 \ : \ \mu_1 = \mu_2$
4. Evaluation d'un dispositif expérimental :
   - Puissance du test (à objectif de précision fixé)
   - Taille d'échantillon nécessaire (à objectifs de précision et de puissance fixés)
5. Tests post-hoc : correction du seuil (Bonferroni) pour un contrôle global du risque d'un faux positif.

Plan