Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

12 - Proportions et fréquences

été 2022

Fréquence

Considérons un échantillon $\{x_1, \cdots, x_n\}$, prenant des valeurs A ou B (voire davantage). Supposons que l'on s'intéresse à la fréquence d'apparition de la modalité A.

Notons $y_i = \mathbf{1}_A(x_i)$, et $\{y_1, \cdots, y_n\}$ l'échantillon prenant les valeurs 0 ou 1. La fréquence (d'apparition de A) est

$$f = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(x_i)$$

(on parle aussi parfois de proportion)

Considérons maintenant une collection de variables aléatoires indépendantes et identiquement distribuées, $Y_1, \cdots, Y_n$, de loi $\mathcal{B}(p)$. Posons

$$F = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$

Fréquence

Si les variables $Y_1, \cdots, Y_n$ sont i.i.d. de loi $\mathcal{B}(p)$

$$\mathbb{E}[F] = p \text{ et } \text{Var}[F] = \frac{p(1-p)}{n}$$

Plus précisément, comme $nF \sim \mathcal{B}(n,p)$,

$$\mathbb{P}\left(F = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \ \forall k = 0, 1, \cdots, n.$$

Si n est suffisamment grand, d'après le théorème central limite

$$Z_n = \sqrt{n}\frac{F-p}{\sqrt{p(1-p)}} \overset{\mathcal{L}}{\to} \mathcal{N}(0,1)$$

En pratique, on suppose l'approximation normale valide si $n \geq 30$, $np \geq 15$ et $n(1-p) \geq 15$

Intervalle de confiance

Modèle binomial, avec n assez grand

Intervalle de confiance, $\{x_1, \cdots, x_n\}$, $\mathcal{B}(p)$, n grand

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{B}(p)$.

$$\left[\widehat{p} \pm u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right], \text{ où } \widehat{p} = \overline{x}$$

Note parfois l'intervalle de confiance dépasse 0 ou 1

```
> sum_x = 12
> n =14
> sum_x/n + qnorm(c(.025,.975))*sqrt(sum_x*(n-sum_x)/n
   ^3)
[1] 0.6738432 1.0404425
```

Intervalle de confiance

Il existe de nombreux intervalles de confiance. Par exemple sur nos données avec 12 fois '1' et 2 fois '0' ($n = 14$),

```
> library(binom)
> binom.confint(12, 14, methods = "all")
          method  x  n      mean     lower     upper
1  agresti-coull 12 14 0.8571429 0.5881065 0.9723858
2     asymptotic 12 14 0.8571429 0.6738432 1.0404425
3          bayes 12 14 0.8333333 0.6517227 0.9853611
4        cloglog 12 14 0.8571429 0.5394482 0.9622319
5          exact 12 14 0.8571429 0.5718708 0.9822055
6          logit 12 14 0.8571429 0.5731738 0.9640393
7         probit 12 14 0.8571429 0.6007290 0.9699396
8        profile 12 14 0.8571429 0.6206505 0.9742387
9            lrt 12 14 0.8571429 0.6206560 0.9747079
10     prop.test 12 14 0.8571429 0.5615066 0.9748606
11        wilson 12 14 0.8571429 0.6005862 0.9599061
```

Intervalle de confiance

Il existe de nombreux intervalles de confiance. Par exemple sur nos données avec 12 fois ‘1’ et 2 fois ‘0’ ($n = 14$),

```
> library(DescTools)
> BinomCI(12, 14, sides = "two.sided", method = c("
    wald", "wilson", "agresti-coull", "arcsine")
                    est    lwr.ci    upr.ci
wald          0.8571429 0.6738432 1.0000000
wilson        0.8571429 0.6005862 0.9599061
agresti-coull 0.7802461 0.5881065 0.9723858
arcsine       0.8389831 0.6096856 0.9773745
```

Intervalle de confiance

Modèle binomial, avec n assez grand

Intervalle de confiance, $\{x_1,\cdots,x_n\}$, $\mathcal{B}(p)$, Agresti–Coull

Soit $\boldsymbol{x}=\{x_1,\cdots,x_n\}$ de loi $\mathcal{B}(p)$.

Posons $\tilde{n}=n+u_{1-\alpha/2}^2$ et $\tilde{p}=\dfrac{1}{\tilde{n}}\left(n\overline{x}+\dfrac{u_{1-\alpha/2}^2}{2}\right)$, alors un intervalle de confiance de niveau α est

$$\left[\tilde{p}\pm u_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right]$$

Une version approchée est d'utiliser $\tilde{p}=\dfrac{x_1+\cdots+x_n+2}{n+4}$

Intervalle de confiance ★★★

Comme on l'a vu dans le chapitre 11, dans un modèle binomial, avec n assez grand

Intervalle de confiance, $\{x_1, \cdots, x_n\}$, $\mathcal{B}(p)$, Wilson

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{B}(p)$.
Un intervalle de confiance de niveau α pour p est

$$\left[\frac{1}{1+\frac{u_{1-\alpha/2}^2}{n}}\left(\hat{p}+\frac{u_{1-\alpha/2}^2}{2n}\right) \pm \frac{u_{1-\alpha/2}}{1+\frac{u_{1-\alpha/2}^2}{n}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u_{1-\alpha/2}^2}{4n^2}}\right]$$

On obtient ces bornes en notant qu'elle correspondent aux p tels que $(\hat{p}-p)^2 = u_{1-\alpha/2}^2 \cdot \dfrac{p(1-p)}{n}$ qui est l'équation de degré 2

$$\left(1+\frac{u_{1-\alpha/2}^2}{n}\right)p^2 + \left(-2\hat{p}-\frac{u_{1-\alpha/2}^2}{n}\right)p + \left(\hat{p}^2\right) = 0\ .$$

Intervalle de confiance ★★★

Modèle binomial, avec n assez grand

Intervalle de confiance, $\{x_1, \cdots, x_n\}$, $\mathcal{B}(p)$, arcsinus

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{B}(p)$.
Un intervalle de confiance de niveau α pour p est

$$\left[\sin^2\left(\arcsin(\sqrt{\widehat{p}}) \pm \frac{u_{1-\alpha/2}}{2\sqrt{n}}\right)\right]$$

L'idée est de noter que comme $\text{Var}[\widehat{P}] = \dfrac{p(1-p)}{n}$,

$$\text{Var}\left(\arcsin\left(\sqrt{P}\right)\right) \approx \frac{\text{Var}(P)}{4p(1-p)} = \frac{p(1-p)}{4np(1-p)} = \frac{1}{4n}.$$

Intervalle de confiance ★★★

Modèle binomial, avec n assez grand

Intervalle de confiance pour $p_x - p_y$ $\mathcal{B}(p_x)$ et $\mathcal{B}(p_y)$, Wald

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ de loi $\mathcal{B}(p_x)$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ de loi $\mathcal{B}(p_y)$.
Un intervalle de confiance de niveau α pour $p_x - p_y$ est

$$[xxxxx]$$

Intervalle de confiance

Si on a deux échantillons, $\boldsymbol{x}$, $m = 14$ et 12 fois '1' ($\overline{x} = 0.857$) et $\boldsymbol{y}$, $n = 15$ et 11 fois '1' ($\overline{y} = 0.7333$)

```
> library(DescTools)
> BinomDiffCI(12, 14, 11, 15, sides  = "two.sided",
    method = c("wald","score"))
             est      lwr.ci     upr.ci
wald  0.1238095 -0.1654654 0.4130845
score 0.1238095 -0.1773352 0.3967329
```

Région de rejet I

Soit $n = 9$, on a dans le tableau suivant $f_\theta(x)$ pour $x \in \{0, 1, 2, \cdots, 7, 8, 9\}$ pour plusieurs valeurs possibles de θ

	0	1	2	3	4	5	6	7	8	9
0.1	0.387	0.387	0.172	0.045	0.007	0.001	0.000	0.000	0.000	0.000
0.2	0.134	0.302	0.302	0.176	0.066	0.017	0.003	0.000	0.000	0.000
0.3	0.040	0.156	0.267	0.267	0.172	0.074	0.021	0.004	0.000	0.000
0.4	0.010	0.060	0.161	0.251	0.251	0.167	0.074	0.021	0.004	0.000
0.5	0.002	0.018	0.070	0.164	0.246	0.246	0.164	0.070	0.018	0.002
0.6	0.000	0.004	0.021	0.074	0.167	0.251	0.251	0.161	0.060	0.010
0.7	0.000	0.000	0.004	0.021	0.074	0.172	0.267	0.267	0.156	0.040
0.8	0.000	0.000	0.000	0.003	0.017	0.066	0.176	0.302	0.302	0.134
0.9	0.000	0.000	0.000	0.000	0.001	0.007	0.045	0.172	0.387	0.387

Région de rejet II

Région de rejet pour un test bilatéral de niveau $\alpha = 10\%$
$H_0 : p = 1/2$ contre $H_1 : p \neq 1/2$

	0	1	2	3	4	5	6	7	8	9
0.1	0.387	0.387	0.172	0.045	0.007	0.001	0.000	0.000	0.000	0.000
0.2	0.134	0.302	0.302	0.176	0.066	0.017	0.003	0.000	0.000	0.000
0.3	0.040	0.156	0.267	0.267	0.172	0.074	0.021	0.004	0.000	0.000
0.4	0.010	0.060	0.161	0.251	0.251	0.167	0.074	0.021	0.004	0.000
0.5	0.002	0.018	0.070	0.164	0.246	0.246	0.164	0.070	0.018	0.002
0.6	0.000	0.004	0.021	0.074	0.167	0.251	0.251	0.161	0.060	0.010
0.7	0.000	0.000	0.004	0.021	0.074	0.172	0.267	0.267	0.156	0.040
0.8	0.000	0.000	0.000	0.003	0.017	0.066	0.176	0.302	0.302	0.134
0.9	0.000	0.000	0.000	0.000	0.001	0.007	0.045	0.172	0.387	0.387

Région de rejet III

Région de rejet pour un test bilatéral de niveau $\alpha = 5\%$
$H_0 : p = 1/2$ contre $H_1 : p \neq 1/2$

	0	1	2	3	4	5	6	7	8	9
0.1	0.387	0.387	0.172	0.045	0.007	0.001	0.000	0.000	0.000	0.000
0.2	0.134	0.302	0.302	0.176	0.066	0.017	0.003	0.000	0.000	0.000
0.3	0.040	0.156	0.267	0.267	0.172	0.074	0.021	0.004	0.000	0.000
0.4	0.010	0.060	0.161	0.251	0.251	0.167	0.074	0.021	0.004	0.000
0.5	0.002	0.018	0.070	0.164	0.246	0.246	0.164	0.070	0.018	0.002
0.6	0.000	0.004	0.021	0.074	0.167	0.251	0.251	0.161	0.060	0.010
0.7	0.000	0.000	0.004	0.021	0.074	0.172	0.267	0.267	0.156	0.040
0.8	0.000	0.000	0.000	0.003	0.017	0.066	0.176	0.302	0.302	0.134
0.9	0.000	0.000	0.000	0.000	0.001	0.007	0.045	0.172	0.387	0.387

Test de proportion

Modèle binomial, avec n assez grand

Test $H_0 : p = p_0$ contre $H_1 : p = p_1$, $\mathcal{B}(p)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{B}(p)$.
Pour tester $H_0 : p = p_0$ contre $H_1 : p = p_1$, on utilise

$$Z = \frac{(\overline{x} - p_0) - 1/2n}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0,1)$.

- si $p_1 > p_0$, on rejette H_0 si $z > \Phi^{-1}(1-\alpha)$
- si $p_1 < p_0$, on rejette H_0 si $z < \Phi^{-1}(\alpha)$

Test de proportion

Modèle binomial, avec n assez grand

Test $H_0 : p = p_0$ contre $H_1 : p \neq p_0$, $\mathcal{B}(p)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{B}(p)$.
Pour tester $H_0 : p = p_0$ contre $H_1 : p \neq p_0$, on utilise

$$Z = \frac{(\overline{x} - p_0) - 1/2n}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0,1)$.

- on rejette H_0 si $|z| > \Phi^{-1}(1-\alpha/2)$

Note: on peut remplacer $\widehat{p} = \overline{x}$ par $\tilde{p} = \dfrac{x_1 + \cdots + x_n + 2}{n+4}$

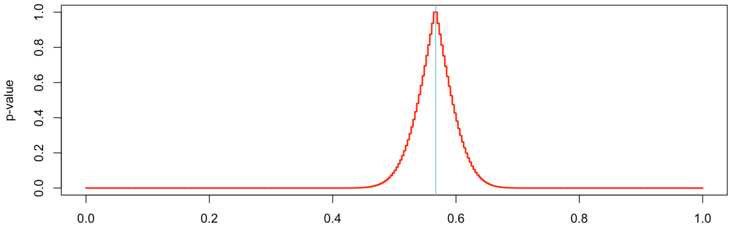
Test de proportion

Considérons un échantillon suivant une loi binomiale $\mathcal{B}(164, 3/5)$

```
> set.seed(1)
> x = sample(0:1, size=164, probability=c(.4,.6))
> binom.test(sum(x),length(x) ,0.6 , alternative ="two.sided")

  Exact binomial test

data:  sum(x) and length(x)
number of successes = 93, number of trials = 164, p-value = 0.4255
alternative hypothesis: true probability of success is not equal to 0.6
95 percent confidence interval:
 0.4875629 0.6441149
sample estimates:
probability of success
             0.5670732
```

Test de proportion

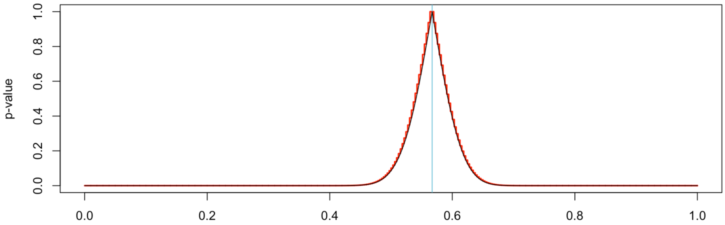
```
> binom.test(sum(x),length(x) ,0.5 , alternative ="two
    .sided")

  Exact binomial test

data:  sum(x) and length(x)
number of successes = 93, number of trials = 164, p-
    value = 0.1007
alternative hypothesis: true probability of success is
     not equal to 0.5
```



Test de proportion

On peut utiliser la p-value avec une approximation Gaussienne,

$$p - \text{value} = 2 \times \left(1 - \Phi\left(\sqrt{n} \cdot \frac{|\overline{x} - p_0|}{\sqrt{p_0(1 - p_0)}}\right)\right)$$



Test de proportion

Modèle binomial avec 2 échantillons, avec n et m assez grands

Test $H_0 : p_x - p_y = p_0$ contre $H_1 : p_x - p_y = p_1$, $\mathcal{B}(p)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ de loi $\mathcal{B}(p_x)$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ de loi $\mathcal{B}(p_y)$.
Pour tester $H_0 : p_x - p_y = p_0$ contre $H_1 : p_x - p_y = p_1$, on utilise

$$Z = \frac{(\overline{x} - \overline{y}) - p_0}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}}, \quad p = \frac{m\overline{x} + n\overline{y}}{m+n}$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0,1)$.

- si $p_1 > p_0$, on rejette H_0 si $z > \Phi^{-1}(1-\alpha)$
- si $p_1 < p_0$, on rejette H_0 si $z < \Phi^{-1}(\alpha)$

Test de proportion

Modèle binomial avec 2 échantillons, avec n et m assez grands

Test $H_0 : p_x - p_y = p_0$ contre $H_1 : p_x - p_y \neq p_0$, $\mathcal{B}(p)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ de loi $\mathcal{B}(p_x)$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ de loi $\mathcal{B}(p_y)$.

Pour tester $H_0 : p_x - p_y = p_0$ contre $H_1 : p_x - p_y \neq p_0$, on utilise

$$Z = \frac{(\overline{x} - \overline{y}) - p_0}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}}, \quad p = \frac{m\overline{x} + n\overline{y}}{m+n}$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0, 1)$.

- on rejette H_0 si $|z| > \Phi^{-1}(1 - \alpha/2)$

Quelques tests

On peut aussi utiliser les tests sur des lois binomiales dans d'autres contextes. Par exemple, on peut faire un test sur la médiane. Pour un échantillon $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, notons m la médiane.

Test $H_0 : m = m_0$ contre $H_1 : m = m_1$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de même loi (inconnue).
Pour tester $H_0 : m = m_0$ contre $H_1 : m = m_1$, on utilise

$$V = \sum_{i=1}^{n} d_i^0,\ \ d_i^0 = \mathbf{1}(x_i - m_0 > 0) = \begin{cases} 1 \text{ si } x_i > m_0 \\ 0 \text{ si } x_i < m_0 \end{cases}$$

Si H_0 est vraie, V suit une loi binomiale $\mathcal{B}(n, 1/2)$

- si $m_1 > m_0$, on rejette H_0 si $v > F_n^{-1}(1-\alpha)$
- si $m_1 < m_0$, on rejette H_0 si $v < F_n^{-1}(\alpha)$

Où F_n est la fonction de répartition de la loi $\mathcal{B}(n, 1/2)$.

Quelques tests dérivés ★★★

… avec bien entendu la version bilatérale

Test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de même loi (inconnue).
Pour tester $H_0 : m = m_0$ contre $H_1 : m \neq m_0$, on utilise

$$V = \sum_{i=1}^{n} d_i^0, \;\; d_i^0 = \mathbf{1}(x_i - m_0 > 0) = \begin{cases} 1 \text{ si } x_i > m_0 \\ 0 \text{ si } x_i < m_0 \end{cases}$$

Si H_0 est vraie, V suit une loi binomiale $\mathcal{B}(n, 1/2)$

- on rejette H_0 si $v > F_n^{-1}(1 - \alpha/2)$ ou $v < F_n^{-1}(\alpha/2)$

Où F_n est la fonction de répartition de la loi $\mathcal{B}(n, 1/2)$.

Quelques tests dérivés ★★★

Classiquement, la p-value dans le cas où $H_1 : m_0 xxxx m_1$ sera $p = \mathbb{P}(V > v)$ soit $1 - F_n(v)$.

```
> loc = "http://freakonometrics.free.fr/MAT4681/
   blood_pressure.txt"
> download.file(loc,"blood_pressure.txt")
> blood_pressure = read.table("blood_pressure.txt",
   header=TRUE, sep=",")
> median(blood_pressure$mmhg)
[1] 134
```

On pourrait tester $m = 120$,

```
> mu0<-120
> d = blood_pressure$mmhg-mu0
> n = length(d[d!=0])
> v = length(d[d>0])
> mean(d[d!=0]>0)
[1] 0.6111111
> v/n
[1] 0.6111111
```

Quelques tests dérivés

On peut aussi avoir un intervalle de confiance pour m

```
> MedianCI(blood_pressure$mmhg, sides = "two.sided",
    method = "exact")
median lwr.ci upr.ci
   134    118    141
> MedianCI(blood_pressure$mmhg, sides = "two.sided",
    method = "boot")
median lwr.ci upr.ci
   134    127    150
```

Quelques tests dérivés ★★★

```
> binom.test(v,n,0.5,alternative="greater")

  Exact binomial test

data:  v and n
number of successes = 33, number of trials = 54, p-
    value = 0.06684
alternative hypothesis: true probability of success is
     greater than 0.5
95 percent confidence interval:
 0.490144 1.000000
```

Quelques tests dérivés ★★★

```
> binom.test(v,n,0.5,alternative="two.sided")

  Exact binomial test

data:  v and n
number of successes = 33, number of trials = 54, p-
    value = 0.1337
alternative hypothesis: true probability of success is
     not equal to 0.5
95 percent confidence interval:
 0.4687878 0.7408017
```

Quelques tests dérivés ★★★

Test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de même loi inconnue, de médiane m. Pour tester $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ on utilise la statistique de test

$$w_+ = \sum_{i=1}^{n} r_i \mathbf{1}_{\mathbb{R}_+}(x_i - m_0) = \sum_{i=1}^{n} r_i \mathbf{1}(x_i > m_0)$$

où r_i est le rang de x_i dans l'échantillon $\boldsymbol{x}$.
Si $n > 20$, W_+ suit (approximativement) une loi normale, i.e.

$$Z = \frac{W_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \approx \mathcal{N}(0, 1)$$

- on rejette H_0 si $z > \Phi^{-1}(1 - \alpha/2)$ ou $z < \Phi^{-1}(\alpha/2)$

Quelques tests dérivés

```
> wilcox.test(blood_pressure$mmhg,mu=120,exact=FALSE,
    correct=TRUE,alternative="two.sided")

  Wilcoxon signed rank test with continuity correction

data:  blood_pressure$mmhg
V = 1144.5, p-value = 0.0005441
alternative hypothesis: true location is not equal to
    120
```

Quelques tests dérivés ★★★

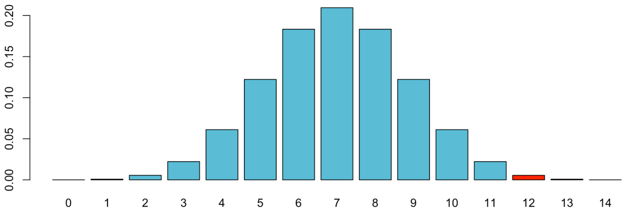
Comment faire quand on a peu d'obervations (n) ? Paul de Poulpe
"*sur 14 prédictions au total, 12 se sont révélées exactes*"

On peut vouloir tester $H_0 : p = 1/2$ contre $H_1 : p > 1/2$.

```
> paul = c(rep(1,12),rep(0,2))
> binom.test(12 ,14 ,0.5 , alternative ="greater")

  Exact binomial test

data:  12 and 14
number of successes = 12, number of trials = 14, p-
    value = 0.00647
alternative hypothesis: true probability of success is
     greater than 0.5
95 percent confidence interval:
 0.6146103 1.0000000
sample estimates:
probability of success
             0.8571429
```

Quelques tests dérivés ★★★

```
> 1 - pbinom(11,size = 14, prob = .5)
[1] 0.006469727
```



Intervalle de confiance pour des comptages

Intervalle de confiance, loi de Poisson $\mathcal{P}(\mu)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{P}(\mu)$. Si n est grand, un intervalle de confiance de niveau α pour μ est

$$\left[\overline{x} \pm u_{1-\alpha/2}\sqrt{\frac{\overline{x}}{n}}\right]$$

```
> set.seed(1)
> x = rpois(n = 60, lambda = 5)
> mean(x)
[1] 5.066667
> PoissonCI(sum(x), length(x), sides = "two.sided",
    method = c("exact", "score", "wald"))
           est   lwr.ci   upr.ci
exact 5.066667 4.513061 5.669440
score 5.066667 4.528228 5.669130
wald  5.066667 4.497114 5.636219
```

Intervalle de confiance pour des comptages ★★★

Intervalle de confiance, loi de Poisson $\mathcal{P}(\mu)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{P}(\mu)$. Si n est grand, un intervalle de confiance de niveau α pour μ est

$$\left[\frac{1}{2}Q_{2n\overline{x}}^{-1}(\alpha/2)\ ;\ \frac{1}{2}Q_{2(n\overline{x}+1)}^{-1}(1-\alpha/2)\right]$$

où $Q_{\nu}^{-1}(u)$ est le quantile de niveau u de la loi du chi-deux à ν degrés de liberté.

Test pour des comptages

Modèle de Poisson avec 1 échantillon avec n assez grand

Test $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1$, $\mathcal{P}(\mu)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ de loi $\mathcal{P}(\mu)$.
Pour tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1$, on utilise

$$xxxxx$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0, 1)$.

- on rejette H_0 si $|z| > \Phi^{-1}(1 - \alpha/2)$

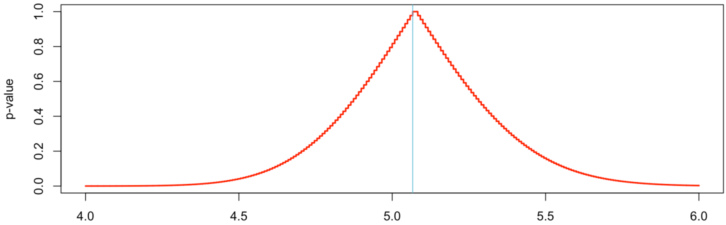
Test pour des comptages

```
> poisson.test(sum(x), length(x), r = 5, alternative = "two.sided")

  Exact Poisson test

data:  sum(x) time base: length(x)
number of events = 304, time base = 60, p-value = 0.8173
alternative hypothesis: true event rate is not equal to 5
95 percent confidence interval:
 4.513061 5.669440
sample estimates:
event rate
  5.066667
```

Test pour des comptages

```
> poisson.test(sum(x), length(x), r = 6, alternative =
    "two.sided")

number of events = 304, time base = 60, p-value =
   0.002657
alternative hypothesis: true event rate is not equal
   to 6
```

On peut visualiser l'évolution de la p-value du test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, en fonction de μ_0



Test pour des comptages

Modèle de Poisson avec 2 échantillons, avec n et m assez grands

Test $H_0 : \mu_x - \mu_y = p_0$ contre $H_1 : \mu_x - \mu_y \neq p_0$, $\mathcal{P}(\mu)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ de loi $\mathcal{P}(\mu_x)$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ de loi $\mathcal{P}(\mu_y)$.
Pour tester $H_0 : \mu_x = \mu_y$ contre $H_1 : \mu_x \neq \mu_y$, on utilise

$$Z = \frac{\overline{x} - \overline{y}}{\sqrt{m\overline{x} + n\overline{y}}}$$

Si H_0 est vraie, $Z \sim \mathcal{N}(0, 1)$.

- on rejette H_0 si $|z| > \Phi^{-1}(1 - \alpha/2)$

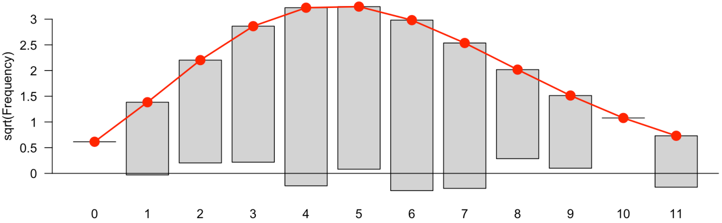
Test pour des comptages ★★★

```
> mean(x)
[1] 5.066667
> library(vcd)
> goodfit(x, type = "poisson", method = "ML")

Observed and fitted values for poisson distribution
with parameters estimated by ‘ML’

 count  observed     fitted pearson residual
     0         0  0.3782038      -0.61498275
     1         2  1.9162325       0.06051336
     2         4  4.8544557      -0.38781027
     3         7  8.1986364      -0.41861679
     4        12 10.3849394       0.50117205
     5        10 10.5234053      -0.16134664
     6        11  8.8864311       0.70901059
     7         8  6.4320835       0.61822577
     8         3  4.0736529      -0.53195130
     9         2  2.2933157      -0.19368829
    10         0  1.1619466      -1.07793628
    11         1  0.5351997       0.10909110
```

Test pour des comptages ★★★

```
> plot(goodfit(x, type = "poisson", method = "ML"))
```

On peut comparer l'histogramme empirique des x_i, et la fréquence théorique de la loi de Poisson $\mathcal{P}(\mu)$,



Test

$$X^2 = \sum_{j=1}^{k} \frac{\big((\text{observed number of } i) - (\text{expected number of } i)\big)^2}{(\text{expected number of } i)}$$

	observed		total	expected ($\perp$)	
	men	women		men	women
right-handed	934	1070	2004	956	1048
left-handed	113	92	205	98	107
ambidextrous	20	8	28	13	15
total	1067	1170	2237	1067	1170

$$n \cdot \mathbb{P}(N_{rm}^{\perp}) = n \cdot \mathbb{P}(N_r)\mathbb{P}(N_m) = n\frac{n_r}{n}\frac{n_m}{n} = 2237\frac{2004}{2237}\frac{1067}{2237} \simeq 956$$

Hypothesis: left-handedness equally common for men and women

$$X^2 = \frac{22^2}{956} + \frac{22^2}{1048} + \frac{15^2}{98} + \frac{15^2}{107} + \frac{7^2}{13} + \frac{7^2}{15} \simeq 12$$

The probability of getting a probability of 12 with a χ_2^2 is 0.2%

Surgery versus Radiation Therapy

Let $\widehat{p}_A$ and $\widehat{p}_B$ be the empirical frequency favoring surgery.

- $\widehat{p}_A$ is (roughly) normally distributed, with mean p_A and standard deviation $\sqrt{p_A(1-p_A)/n}$, that can be approximated by $\sqrt{\widehat{p}_A(1-\widehat{p}_A)/n} \simeq \sqrt{0.5^2/80} = 0.056$,
- $\widehat{p}_B$ is (roughly) normally distributed, with mean p_B and standard deviation $\sqrt{p_B(1-p_B)/n}$, that can be approximated by $\sqrt{\widehat{p}_B(1-\widehat{p}_B)/n} \simeq \sqrt{0.84 \cdot 0.16/87} = 0.039$,

Assuming that $p_A = p_B$ (assumption H_0), $\widehat{p}_A - \widehat{p}_B$ is (roughly) normally distributed, with mean 0 and standard deviation approximated by $\sqrt{0.056^2 + 0.039^2} = 0.068$.

$$Z = \frac{\widehat{p}_A - \widehat{p}_B}{0.068} = \frac{0.50 - 0.84}{0.068} = -5$$