Modéles continus de distributions fréquents

Antonio Falcó

Théorème Centrale Limite

Théorème

Soit $X \sim \mathcal{F}$ telle que $E[X] = \mu$ et $\mathrm{Var}(X) = \sigma^2$. Pour

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{F}$$

si n est suffisamment grand on a

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1),$$

où $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ est la moyenne empirique et $\mathcal{N}(0,1)$ est la distribution normale standard.

Soit $X \sim \mathcal{B}(10, p = 0.5)$ d'où

$$\mu = E[X] = 10 \cdot 0.5 = 5 \text{ et } \mathrm{Var}(X) = \sigma^2 = 10 \cdot 0.5 \cdot (1 - 0.5) = 2.5.$$

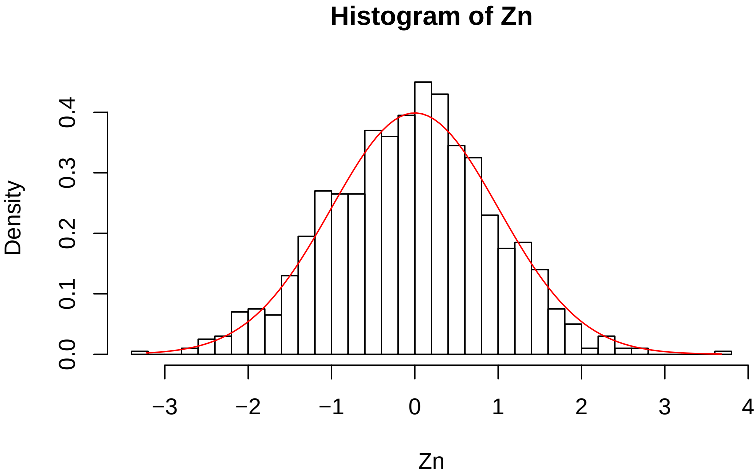
Pour calculer Z_n on a besoin de calculer $\overline{X}$. Alors on a d'obtenir

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{B}(10, p = 0.5)$$

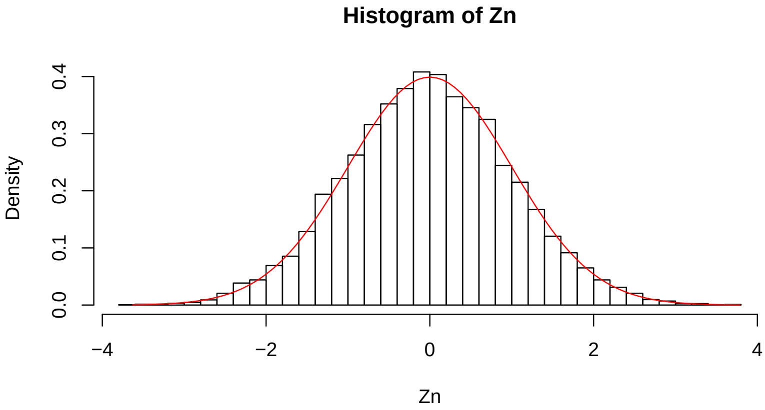
avec $n = 1000$. On peut calculer l'échatillon en utilisant un générateur du nombre aléatoire avec la fonction de R `rbinom()`.

```
n ← 1000
p ← 0.5
mu ← 10*p
sigma ← sqrt(10*p*(1-p)/n)
moyenne.empirique ← rep(0,n)
for (i in 1:n){
Xiid ← rbinom(n,10,p)
moyenne.empirique[i] ← mean(Xiid)
}
Zn ← (moyenne.empirique-mu)/sigma
```

```
hist(Zn, breaks=30, prob=TRUE)
curve(dnorm(x), from=min(Zn), to=max(Zn), add=TRUE, col='red')
```



$n = 10000$



Le TCL et la distribution binomiale

Corollaire

Soit $X \sim \mathcal{B}(p)$. Alors $E[X] = p$ et $\mathrm{Var}(X) = p(1-p)$. Pour

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{B}(p)$$

si n est suffisamment grand on a

$$Z_n = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1).$$

Regardez que dans ce cas particulier la moyenne empirique

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{\text{Cas favorables}}{\text{Cas possibles}} = \widehat{p},$$

est une probabilité empirique qui permet estimer la valeur de p.

La densité de probabilité

Pour $-\infty < \mu < \infty$ et $\sigma > 0$ soit

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ où } -\infty < x < \infty,$$

et avec le logiciel R:

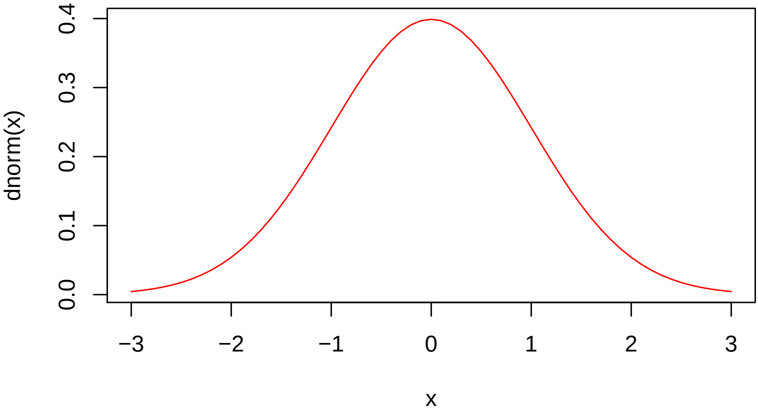
$$f(x|\mu,\sigma) = \texttt{dnorm}(x,\mu,\sigma)$$

en particulier:

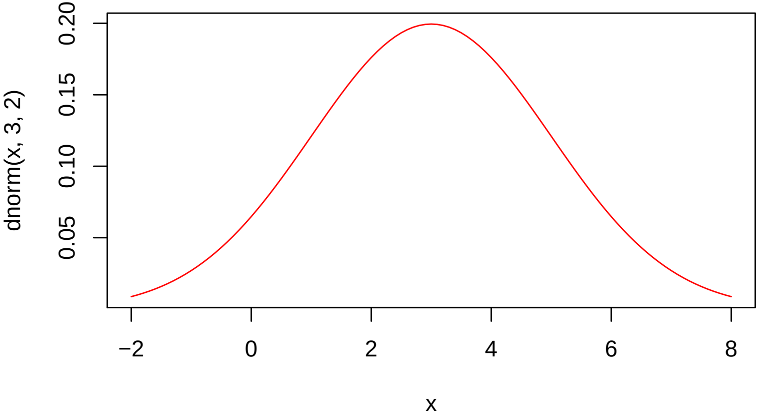
$$f(x|\mu=0,\sigma=1) = \texttt{dnorm}(x).$$

Propriété

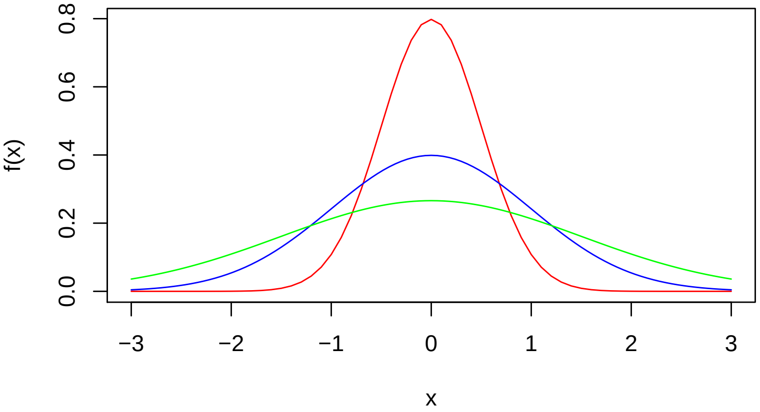
$\int_{-\infty}^{\infty} f(x|\mu,\sigma)dx = 1$ pour tout $-\infty < \mu < \infty$ et $\sigma > 0$.


dnorm(x)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

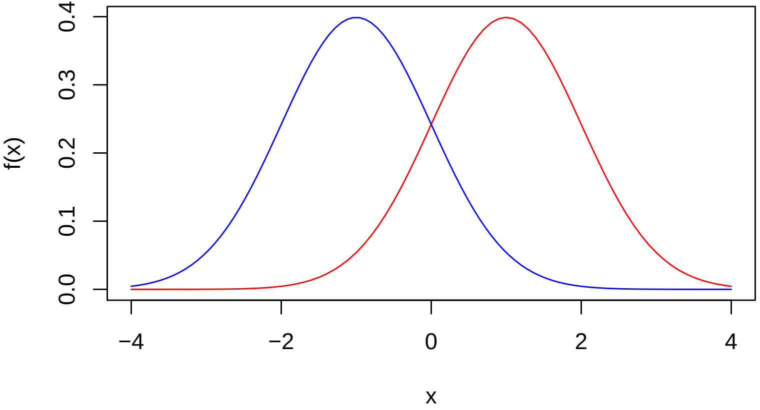
$\mu = 3$ et $\sigma = 2$.



$\mu = 3$ et $\sigma = 0.5$(rouge), 1(bleue), 1.5(vert).



$\mu = -1(\text{bleue}), 1(\text{rouge})$ et $\sigma = 1$



Définition

On dit qu'une variable X a une distribution normale, ou gaussienne, ou de Gauss, d'espérance μ et de variance σ^2 si sa densité est égale à:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ où } -\infty < x < \infty.$$

On écrit alors $X \sim \mathcal{N}(\mu, \sigma)$.

Propriétés

On peut montrer que

- $E[X] = \mu$.
- $\mathrm{Var}(X) = \sigma^2$.

Propriété

Soit $X \sim \mathcal{N}(\mu, \sigma)$ et $Y = aX + b$ où $a \neq 0$ et b sont des nombres réelles. Alors

$$Y \sim \mathcal{N}(a\mu + b, a\sigma).$$

Conséquence

Soit $X \sim \mathcal{N}(\mu, \sigma)$. Alors

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Observe que

$$Z = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma} = \underbrace{\frac{1}{\sigma}}_{=a} X + \underbrace{\left(-\frac{\mu}{\sigma}\right)}_{=b}.$$

Alors $a\mu + b = \frac{1}{\sigma}\mu + \left(-\frac{\mu}{\sigma}\right) = 0$ et $a\sigma = \frac{1}{\sigma}\sigma = 1$.

Définition

Soit $X \sim \mathcal{N}(\mu, \sigma)$. Alors on dit que la variable Z définie comme

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

suivi un distribution normal standard et l'opération ci-dessus s'appelle la standardisation.

Propriété

Soit $X \sim \mathcal{N}(\mu_X, \sigma_X)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$. Alors

$$X + Y \sim \mathcal{N}\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\mathrm{cov}(X, Y)}\right).$$

Si X et Y sont indépendantes alors

$$X + Y \sim \mathcal{N}\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

Propriété

Soit $X \sim \mathcal{N}(\mu_X, \sigma_X)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$. Alors

$$X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2 - 2\mathrm{cov}(X, Y)}\right).$$

Si X et Y sont indépendantes alors

$$X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

Les probabilités chez la distribution normal

Soit $X \sim \mathcal{N}(\mu, \sigma)$. La fonction de distribution cumulative est

$$\Phi(x|\mu,\sigma) = \Pr(X \leq x) = \int_{-\infty}^{x} f(u|\mu,\sigma)du = \texttt{pnorm}(x,\mu,\sigma).$$

Exemple

Soit $X = \texttt{poids}$ d'un population avec $\mu = 67$ kg et $\sigma = 2.5$ kg. Quelle la probabilité de trouver une personne avec un poids inférieure à 62 kg dans la population?

$$\Phi(62|\mu = 67, \sigma = 2.5) = \Pr(X \leq 62) = \texttt{pnorm}(62, 67, 2.5) = 0.0227501.$$

```
mu ← 67
sigma ← 2.5
x ← 62
pnorm(x,mu,sigma)
```

```
[1] 0.02275013
```

```
z ← (x-mu)/sigma
pnorm(z)
```

```
[1] 0.02275013
```

Soit $X \sim \mathcal{N}(\mu, \sigma)$. Alors

$$\Pr(X \leq x) = \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Pr\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

avec $Z \sim \mathcal{N}(0,1)$.

Exemple

Soit $X = \texttt{poids}$ d'un population avec $\mu = 67$ kg et $\sigma = 2.5$ kg. Quelle la probabilité de trouver une personne avec un poids supérieure à 62 kg dans la population?

L'événement contraire

$$\overline{\{X > 62\}} = \{X \leq 62\}.$$

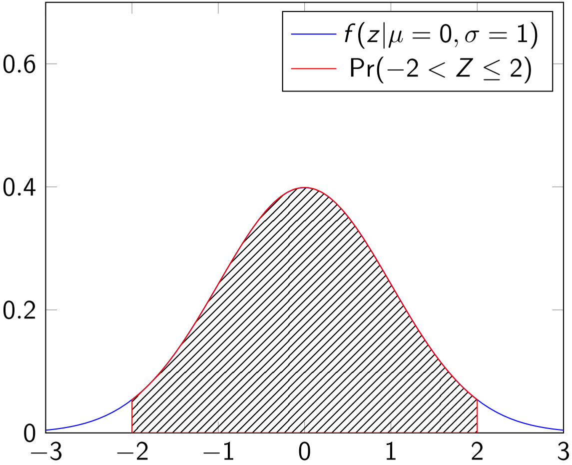
Alors

$$\begin{aligned} \Pr(X > 62) &= 1 - \Pr(\overline{\{X > 62\}}) = 1 - \Pr(X \leq 62) \\ &= 1 - \texttt{pnorm}(62, 67, 2.5) = 0.9772499. \end{aligned}$$

Une propriété valable pour toute distribution normale

Soit $Z \sim \mathcal{N}(0,1)$ et $0 < \alpha < 1$ fixée. On cherche $z_\alpha > 0$ telle que

$$\Pr(-z_\alpha < Z \leq z_\alpha) = 1 - \alpha$$



Une propriété valable pour toute distribution normale

Problème

Soit $Z \sim \mathcal{N}(0,1)$ et $0 < \alpha < 1$ fixée. On cherche $z_\alpha > 0$ telle que

$$\Pr(-z_\alpha < Z \leq z_\alpha) = 1 - \alpha$$

Exemple

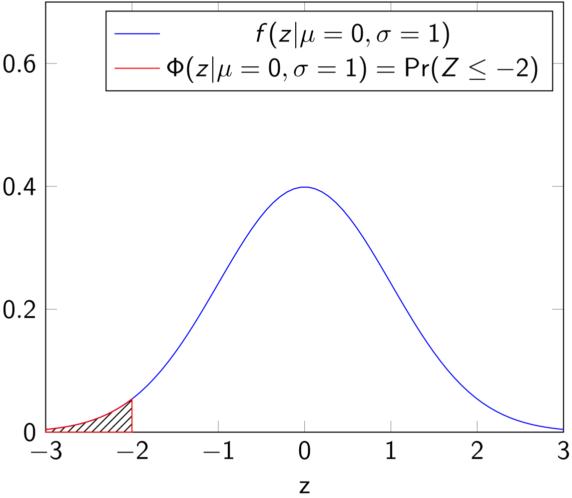
Soit $\alpha = 0.05 > 0$, alors on cherche $z_{\alpha=0.05} > 0$ telle que

$$\Pr(-z_\alpha < Z \leq z_\alpha) = 1 - 0.05 = 0.95.$$

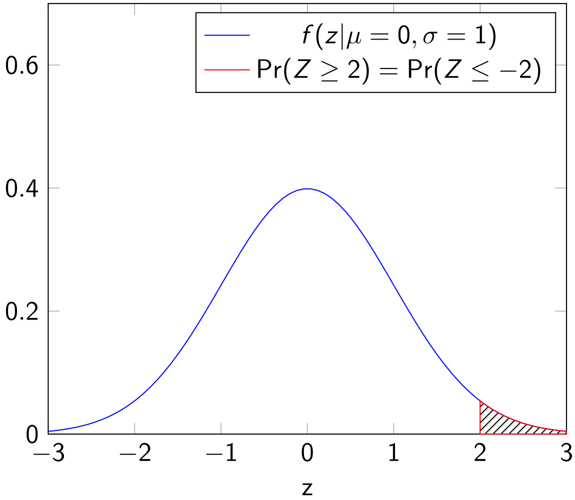
cet-à-dire l'intervalle $[-z_{\alpha=0.05}, z_{\alpha=0.05}]$ compris le 95% des probabilités. En conséquence, $z_{\alpha=0.05} > 0$ est la solution de l'équation

$$\Pr(Z \leq -z_{\alpha=0.05}) = \Pr(Z > z_{\alpha=0.05}) = 0.05/2 = 0.025.$$

Une propriété valable pour toute distribution normale



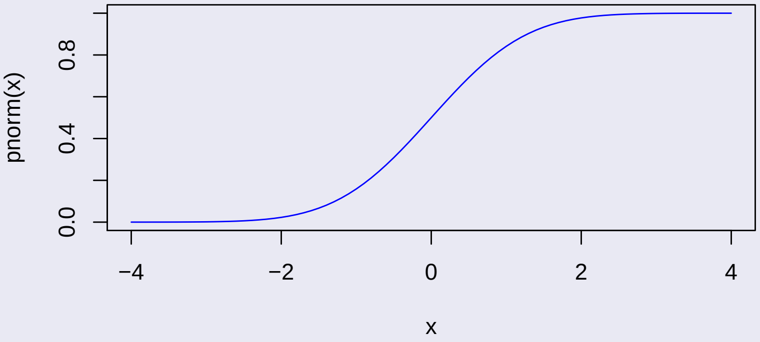
Une propriété valable pour toute distribution normale



Une propriété valable pour toute distribution normale

La solution de l'équation $P(Z \leq -z_\alpha) = \alpha/2$

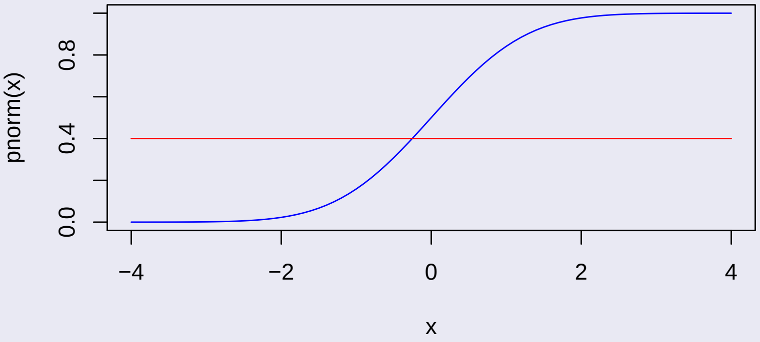
La fonction $\Phi(z|\mu = 0, \sigma = 1) = \Pr(Z \leq z)$ est strictement croissante:



Une propriété valable pour toute distribution normale

La solution de l'équation $P(Z \leq -z_\alpha) = \alpha/2$

Pour chercher z telle que $\Pr(Z \leq z) = 0.4$:



Une propriété valable pour toute distribution normale

La solution de l'équation $P(Z \leq -z_\alpha) = \alpha/2$

On utilise la fonction inverse de $\Pr(Z \leq z) = \texttt{pnorm}(z) = \alpha/2$:

$$\texttt{qnorm}(\alpha/2) = z \Leftrightarrow \Pr(Z \leq z) = \alpha/2.$$

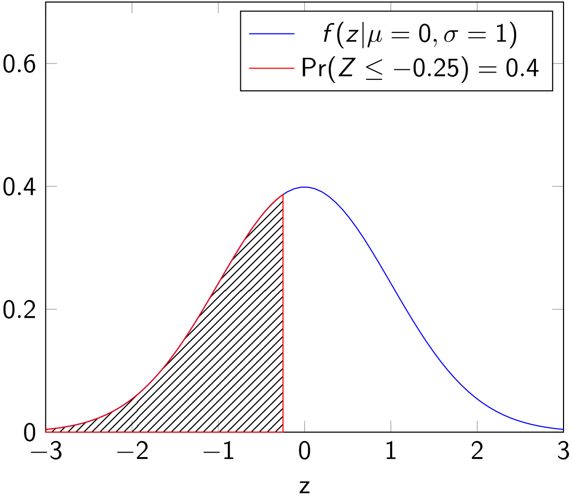
Exemple

Soit $\alpha/2 = 0.4$ on cherche z telle que $\Pr(Z \leq z) = 0.4$:

$$z = \texttt{qnorm}(0.4) = -0.2533471.$$

En conséquence, $\Pr(Z \leq -0.2533471) = 0.4$.

Une propriété valable pour toute distribution normale



Une propriété valable pour toute distribution normale

Problème

Soit $Z \sim \mathcal{N}(0,1)$ et $0 < \alpha < 1$ fixée. On cherche $z_\alpha > 0$ telle que

$$\Pr(-z_\alpha < Z \leq z_\alpha) = 1 - \alpha$$

Solution

$$-z_\alpha = \texttt{qnorm}(\alpha/2) \Leftrightarrow \Pr(Z \leq z_\alpha) = \alpha/2.$$

Exemple

Trouvez $z_{\alpha=0.05}$:

$$-z_{\alpha=0.05} = \texttt{qnorm}(0.05/2) = -1.959964.$$

Alors

$$\Pr(-1.959964 < Z \leq 1.959964) = 1 - 0.05 = 0.95.$$

Une propriété valable pour toute distribution normale

Cas général

Soit $X \sim \mathcal{N}(\mu, \sigma)$. Alors on cherche $x_\alpha, y_\alpha > 0$ telle que

$$\Pr(x_\alpha < X \leq y_\alpha) = \Pr\left(\frac{x_\alpha - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{y_\alpha - \mu}{\sigma}\right) = 1 - \alpha.$$

En conséquence la solution est:

$$-z_\alpha = \frac{x_\alpha - \mu}{\sigma} \Leftrightarrow x_\alpha = \mu - z_\alpha\,\sigma$$
$$z_\alpha = \frac{y_\alpha - \mu}{\sigma} \Leftrightarrow y_\alpha = \mu + z_\alpha\,\sigma$$

où $\Pr(-z_\alpha < Z \leq z_\alpha) = 1 - \alpha$ avec $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Exemple

Question

Soit $X = \texttt{taille}$ d'un population où $\mu = 176$ cm et $\sigma = 4.2$ cm. Quelle est l'intervalle que contiens le 95% des tailles dans la population?

On cherche $x_\alpha, y_\alpha > 0$ telle que

$$\Pr(x_\alpha < X = \texttt{taille} \leq y_\alpha) = 0.95 = 1 - \alpha \Rightarrow \alpha = 0.05.$$

Comme $-z_{\alpha=0.05} = \texttt{qnorm}(0.05/2) = -1.959964$, alors

$$x_\alpha = \mu - z_\alpha \cdot \sigma = 176 - 1.959964 \cdot 4.2 = 167.7681513$$
$$y_\alpha = \mu + z_\alpha \cdot \sigma = 176 + 1.959964 \cdot 4.2 = 184.2318487$$

On peut affirmer que dans cette population le 95% des personnes ont une taille dans l'intervalle $[167.7681513, 184.2318487]$ cm.

La moyenne empirique

Soit $X \sim \mathcal{N}(\mu, \sigma)$ et on considère un échantillonnage aléatoire simple dans ce population:

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma).$$

L'objectif est de trouvé des estimateurs pour μ et σ^2. Pour estimer μ on considère la moyenne empirique:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Leftrightarrow \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

La variance empirique corrigée

Pour estimer σ^2 on considère la variance empirique corrigée:

$$S^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{1}{n-1}\sum_{k=1}^{n}(X_i - \overline{X})^2.$$

d'où

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{k=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Propriété

Soit $X \sim \mathcal{N}(\mu, \sigma)$ et on considère un échantillonnage aléatoire simple dans ce population:

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma).$$

Alors

$$\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma} \text{ i.i.d. } \sim \mathcal{N}(0, 1)$$

Remarque

Observe que

$$\left(\frac{X_1-\mu}{\sigma}\right)^2+\cdots+\left(\frac{X_n-\mu}{\sigma}\right)^2$$

et une somme de carres d'une échantillonnage aléatoire simple d'une population $\mathcal{N}(0,1)$.

Propriété 1

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\overline{X})-(\mu-\overline{X})]^2\\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}(\mu-\overline{X})^2-2(\mu-\overline{X})\sum_{i=1}^{n}(X_i-\overline{X})\\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2+n(\mu-\overline{X})^2-2(\mu-\overline{X})\left(\sum_{i=1}^{n}X_i-n\overline{X}\right)\end{aligned}$$

d'où

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2$$

$(\overline{X}-\mu)^2=(\mu-\overline{X})^2$ et $\sum_{i=1}^{n}X_i=n\overline{X}$.

Propriété 2

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 + \frac{n}{\sigma^2}(\overline{X}-\mu)^2$$

d'où

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + \left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

et

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

Définition

Soit $Z \sim \mathcal{N}(0,1)$ et on considère un échantillonnage aléatoire simple dans ce population:

$$Z_1, \dots, Z_n \text{ i.i.d. } \sim \mathcal{N}(0,1).$$

Soir alors

$$Y = Z_1^2 + \cdots + Z_n^2.$$

La variable Y a une distribution χ^2 à n-degrées de liberté, ce qu'on note

$$Y \sim \chi_n^2.$$

Propriétés

- $E[Y] = 2n.$
- $\mathrm{Var}(Y) = 2n.$

Distribution des probabilités

Soit $X \sim \chi^2_n$. Alors

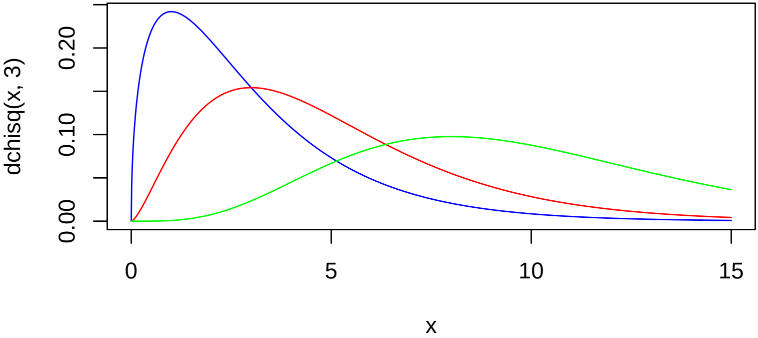
- $f(x|n) = \texttt{dchisq}(x, n)$
- $\Phi(x|n) = \Pr(X \leq x) = \texttt{pchisq}(x, n)$.

Exemple

Soit $X \sim \chi^2_n$ avec $n = 5$.

- $f(3|n = 5) = \texttt{dchisq}(3, 5) = 0.1541803$.
- $\Phi(3|n = 5) = \Pr(X \leq 3) = \texttt{pchisq}(3, 5) = 0.3000142$.

$X \sim \chi^2_n$ bleue($n = 3$),rouge($n = 5$) et vert($n = 10$).



L'estimateur de la variance

Propiété

Soit $X \sim \mathcal{N}(\mu, \sigma)$ et on considère un échantillonnage aléatoire simple dans ce population:

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma).$$

Alors $\overline{X}$ et S^2 sont indépendantes et

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n.$$

En conséquence

$$\frac{(n-1)S^2}{\sigma^2} \sim\sim \chi^2_{n-1}.$$

Dépendance par rapport les paramètres

La distribution de la variance empirique corrigée dépend de la variance de la population:

$$\frac{S^2}{\frac{\sigma^2}{n-1}} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

et la distribution de la moyenne empirique corrigée dépend de la moyenne et la variance de la population:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

Question:

Est-ce qu'on peut connaître la distribution de:

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \text{ où } S = \sqrt{\frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}}.$$

Remarque

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}} \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{S^2}{n}}{\frac{\sigma^2}{n}}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1}\left(\frac{(n-1)S^2}{\sigma^2}\right)}}$$

où

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = Z_0 \sim \mathcal{N}(0,1)$$

et

$$\frac{(n-1)S^2}{\sigma^2} = Z_1^2 + \cdots + Z_{n-1}^2 \sim \chi^2_{n-1},$$

où

$$Z_1, \ldots, Z_{n-1} \text{ i.i.d. } \sim \mathcal{N}(0,1).$$

Soient $Z_0, Z_1, \ldots, Z_{n-1}$ des variables normales standard indépendantes. Soit alors

$$T = \frac{Z_0}{\sqrt{\frac{1}{n-1}(Z_1^2 + \cdots + Z_{n-1}^2)}}$$

La variable T a une distribution t de Student à n degrés de liberté, ce qu'on note

$$T \sim t_{n-1}.$$

Propriété

Soit $X \sim \mathcal{N}(\mu, \sigma)$. Alors $\overline{X}$ et S^2 sont indépendantes et

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Propriété

Soit $T \sim t_n$. Alors

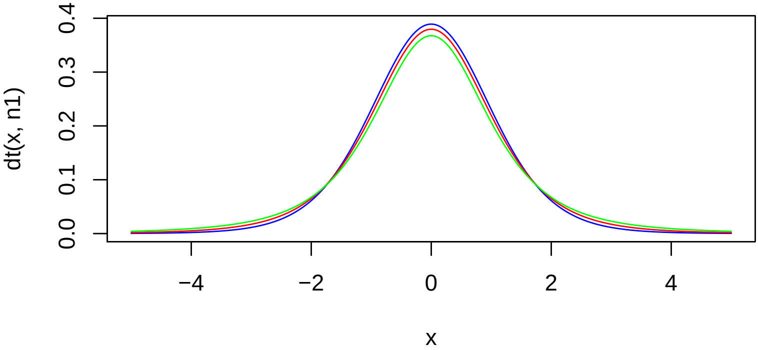
- $E[T] = 0$.
- $\mathrm{Var}(T) = \frac{n}{n-2}$ pour $n > 2$.

Distribution de probabilité

Soit $X \sim t_n$. Alors

- $f(x|n) = \mathtt{dt}(x, n)$
- $\Phi(x|n) = \Pr(X \leq x) = \mathtt{pt}(x, n)$

$X \sim t_n$ bleue($n = 10$),rouge($n = 5$) et vert($n = 3$).



Propriétés

- La distribution t est symétrique autour de 0,
- lorsque $n \to \infty$ la densité d'une variable $T \sim t_n$ tend vers une densité normale standard.