Direction de la Recherche Clinique et de Innovation

Center for Methodology and Modeling
Malaria – Public Health – Infectious Diseases.
West-Africa

Biostatistiques

Partie 3 : Lois de probabilité et tests d'hypothèses

DU EPIDEMIOLOGIE DE TERRAIN - IFRISSE

Kankoé SALLAH MD, PhD

kankoe.sallah@univ-amu.fr
kankoe@skml.fr

Sept 2020

OBJECTIFS MAJEURS :

Comprendre les principales lois de probabilité et pouvoir les choisir adéquatement en fonction du phénomène à décrire.

Comprendre le raisonnement des tests d’hypothèse et les choisir convenablement en fonction des types de données, des effectifs et des conditions paramétriques

Loi binomiale

Illustration 1 : Dans un bocal contenant des boules noires et blanches, probabilité de tirer k boules blanches à l'issue de n tirages, connaissant la proportion totale (P) de boules blanches.

Illustration 2 : Dans une population de malades et non-malades, probabilité de tirer k malades à l'issue de n tirages, connaissant la proportion totale (P) de malades..

Définition : $\mathrm{P}(\mathrm{X}=\mathrm{k})=\mathrm{C}_n^k\, P^k(1-P)^{n-k}$ avec $\mathrm{C}_n^k=\dfrac{n!}{k!(n-k)!}$

Correspond à la probabilité d'obtenir k boules rouges en tirant n boules, avec remises. P est donc le pourcentage des boules rouges

Espérance: $E(X)=nP$

Variance: $\mathrm{var}(X)=nPQ$

Peut être approchée par la loi normale si nP et nQ grands (≥5)

Explications

Bernoulli Distribution (π) ("Bernoulli Trial")

A random variable Z is said to have a Bernoulli Distribution if it takes on the value 1 with probability π and takes on the value 0 with probability $(1-\pi)$.

Value of Z =	P[Z = z] =
1	π
0	$(1-\pi)$

(1) μ = Mean = E[Z] = Statistical Expectation of Z

$\mu = \pi$

(2) σ^2 = Variance = Var[Z] = $E[(Z-\mu)^2]$ = Statistical Expectation of $(Z-\mu)^2$

$\sigma^2 = \pi(1-\pi)$

Loi binomiale

Explications

Combinatorial

The number of ways to choose x items from n (without regard to order) is "n choose x":

$$ {}_{n}C_{x} = \binom{n}{x} = \frac{n!}{x!\,(n-x)!} $$

Note -

$$ \binom{n}{x} = \binom{n}{n-x} $$

Eureka! Choosing x is the same as leaving (n-x) behind.

$$ \binom{n}{0} = \binom{n}{n} = 1 $$

Loi binomiale

Explications

Binomial Distribution (n, p=π)
(n independent Bernoulli Trials)

A random variable X is said to follow a Binomial (n, p=π) distribution if it is the sum of n independent Bernoulli Distribution trials each with probability of "success" = p=π.

Value of X =	P[X = x] =
0	$(1-\pi)^n$
1	$n\pi\,(1-\pi)^{n-1}$
…	…
x	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$
…	…
n	π^n

(1) μ = Mean = E[X] = Statistical Expectation of X
$\mu = np = n\pi$

(2) σ^2 = Variance = Var[X] = E[$(X-\mu)^2$] = Statistical Expectation of $(X-\mu)^2$
$\sigma^2 = np(1-p) = n\pi\,(1-\pi)$

Explications

Binomial Probability Distribution
Formula for Calculating Probabilities

The binomial formula is the binomial distribution probability that you use to calculate a binomial probabilities of the form:

What is the probability that, among n independent Bernoulli trials, each with probability of success = π, x events of "success" occur?

The probability of obtaining exactly x events of success in n independent trials, each with the same probability of event success equal to π:

$$\Pr[X=x] = \binom{n}{x} \pi^{x} (1-\pi)^{n-x} = \left[\frac{n!}{x!\,(n-x)!}\right] \pi^{x} (1-\pi)^{n-x}$$

Explications

Another Look

Binomial Distribution (n, π)

X = sum of n *independent* Bernoulli(π) Trials Z

The n Bernoulli trials are Z_1 Z_2 … Z_n

- **Each Z_i has possible values of 1 ("success") or 0 ("failure")**
- **Pr [$Z_i = 1$] = π and**
  Pr [$Z_i = 0$] = $(1-\pi)$ for i=1, 2, …, n

The Binomial random variable is $X = Z_1 + Z_2 + \ldots + Z_n$. X is distributed Binomial(n, π)

$$X = \sum_{i=1}^{i=n} Z_i$$

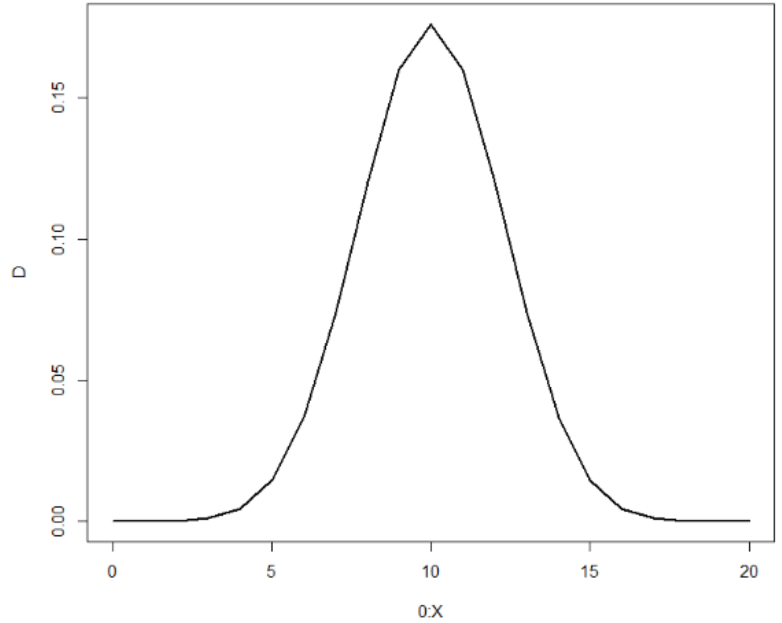
For X ~ Binomial (n, π), the probability that X =x is given by the binomial formula:

$$\text{Probability}[X=x] = \left[\frac{n!}{x!\,(n-x)!}\right] \pi^x (1-\pi)^{n-x},$$

Loi binomiale

Lançons une pièce de monnaie 20 fois.

Quelle est la probabilité d'obtenir « Face » 10 fois exactement ?



Sur R

dbinom(x, size, prob)

size=20, x=10, prob=1/2

Réponse

dbinom(x=10, size=20, prob=0.5) = 0.176

Quand on lance une pièce20 fois, la probabilité d'avoir « Face » exactement 10 fois est de 17,6%. C'est bien ce qu'on lit sur la courbe.

Loi binomiale

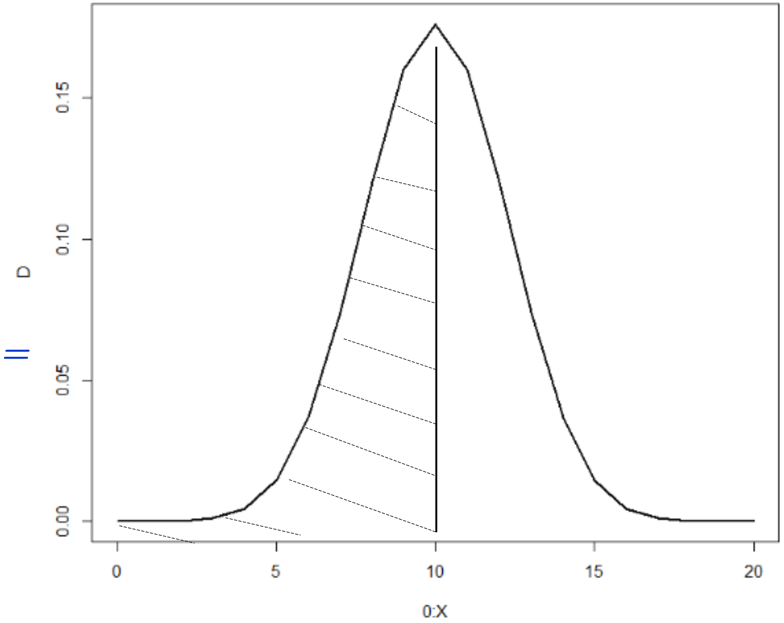
Lançons une pièce de monnaie 20 fois.

Quelle est la probabilité d'obtenir « Face » au plus 10 fois ?

Sur R

pbinom(q, size, prob, lower.tail = TRUE)

size=20, q=10, prob=1/2



Réponse

pbinom(q=10, size=20, prob=0.5, lower.tail = TRUE)= 0.58

Quand on lance une pièce 20 fois, la probabilité d'avoir « Face » exactement 0 fois, ou 1 fois ou 2 fois … Jusqu'à 10 fois est 0.58

Presque la moitié …

Loi binomiale

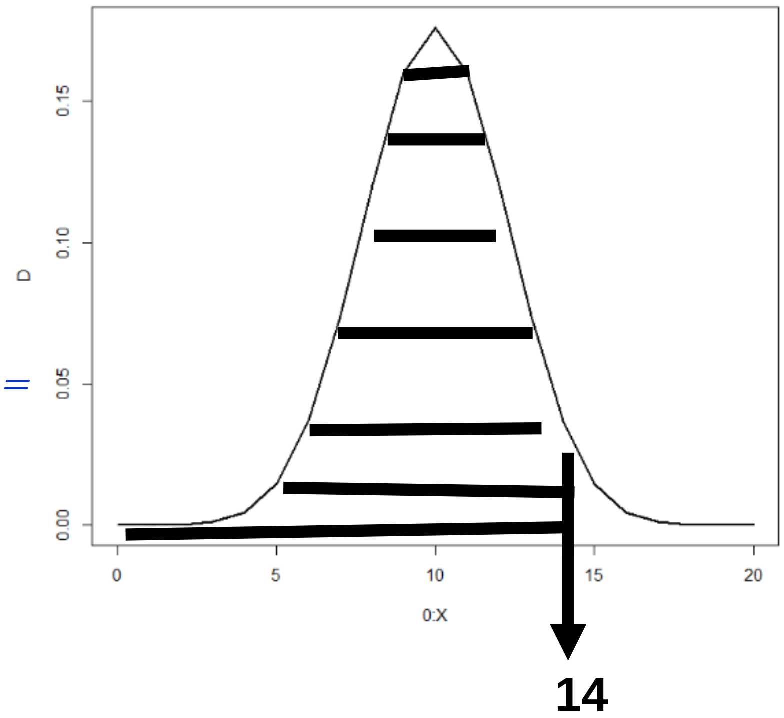
Lançons une pièce de monnaie 20 fois.

Dans 95% des cas, nous sommes sûrs d’obtenir « Face » au plus n fois ?
Estimer n

Sur R

qbinom(p, size, prob, lower.tail = TRUE)

size=20, p=95/100, prob=1/2



Réponse

qbinom(p=95/100, size=20, prob=0.5) = 14

Quand on lance une pièce 20 fois, dans 95% des cas on a moins de 14 « Faces »

Loi binomiale

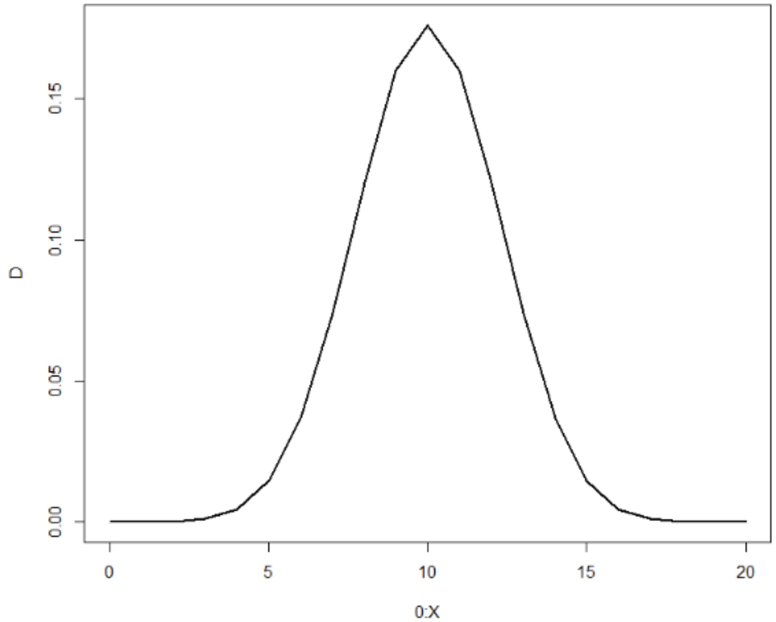
Condition d'approximation de la loi binomiale par la loi normale

$$\boxed{n \geq 30,\ np \geq 5 \text{ et } n(1-p) \geq 5}$$

Loi binomiale et loi normale

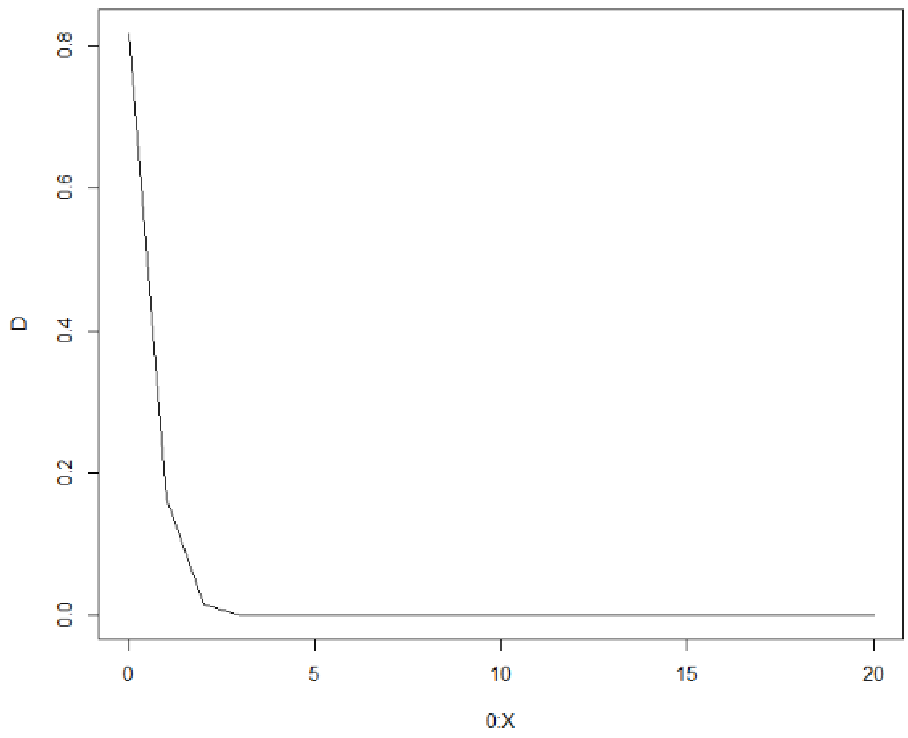
```
dbinom(x=3, size=20, prob=0.5, log = FALSE)
```

```
X=20
D=c()
for (x in 0:X){D=c(D, dbinom(x=x, size=X, prob=0.5, log = FALSE))}
plot(0:X,D, type="l", lty=1)
```



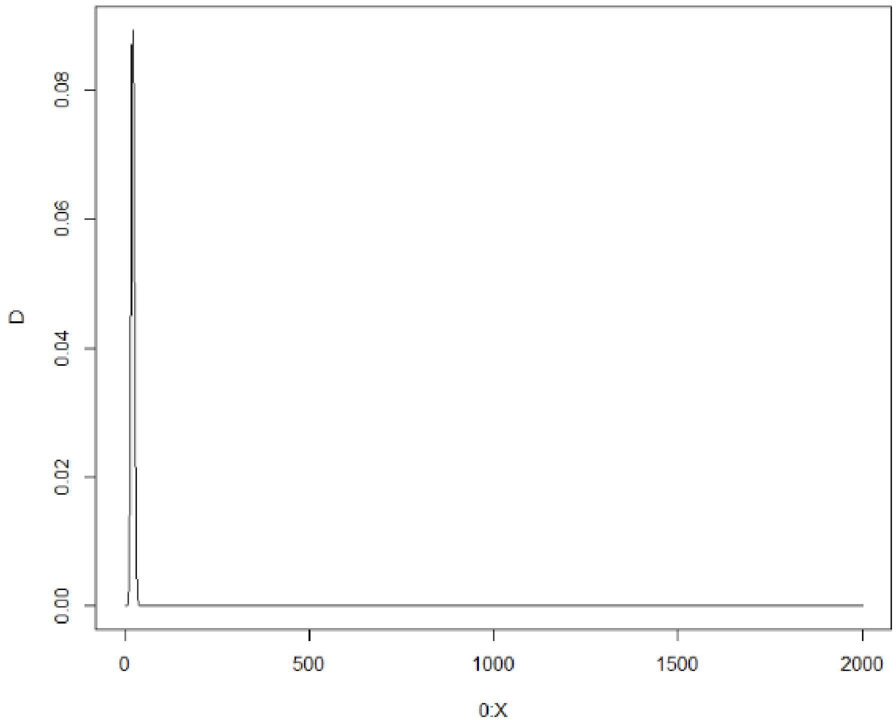
Loi binomiale et loi normale

```
N=20
D=c()
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=0.01, log = FALSE))}
plot(0:N,D, type="l", lty=1)
```



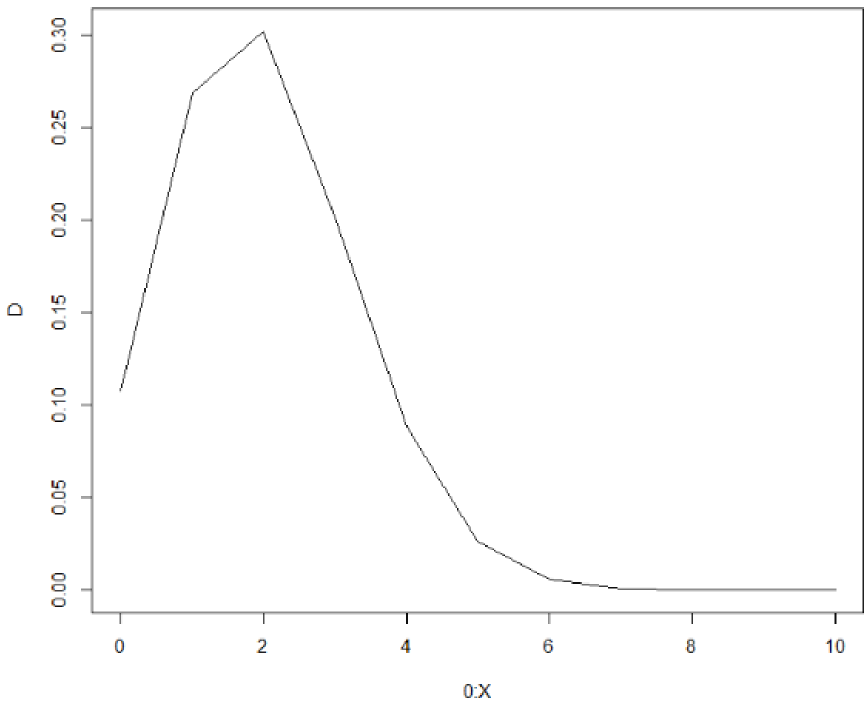
Loi binomiale et loi normale

```
N=2000
D=c()
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=0.01, log = FALSE))}
plot(0:N,D, type="l", lty=1)
```



Loi binomiale et loi normale

```
N=10
D=c()
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=0.2, log = FALSE))}
plot(0:X,D, type="l", lty=1)
```

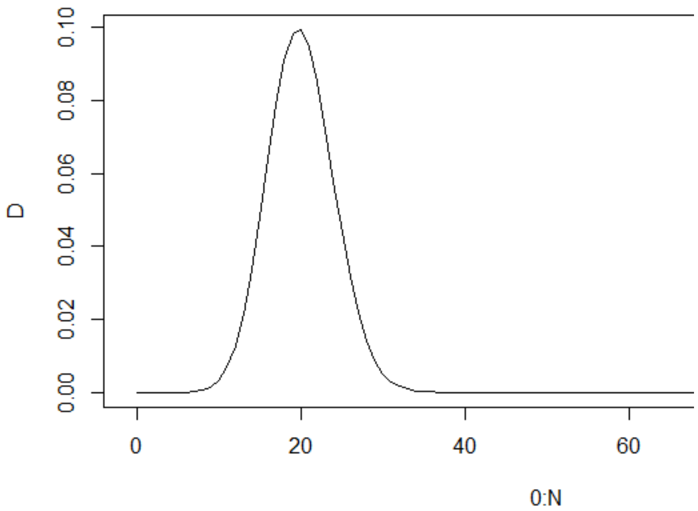


Condition d 'approximation

$$n \geq 30,\ np \geq 5 \text{ et } n(1-p) \geq 5$$

Loi binomiale et loi normale

```
N=100
D=c()
for (x in 0:N){D=c(D, dbinom(x=x, size=N, prob=0.2, log = FALSE))}
plot(0:N,D, type="l", lty=1)
```



Condition d 'approximation

$$n \geq 30,\ np \geq 5 \text{ et } n(1-p) \geq 5$$

Loi normale

Normal Distribution (μ, σ^2)

A random variable X that is distributed normal with mean=μ and variance=σ^2 has probability density function

$$f_X(X=x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \text{ where}$$

x = Value of X

Range of possible values of X: -∞ to +∞

Exp = e = Euler's constant = 2.71828 … ***note:*** $e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots$

π = mathematical constant = 3.14 ***note: π = (circumference/diameter) for any circle***

μ = Expected value of X ("the long run average")

σ^2 = Variance of X. ***Recall – this is the expected value of*** $[X - \mu]^2$

Standard Normal Distribution (μ=0, σ^2=1)

Z-scores are distributed Normal(0,1)

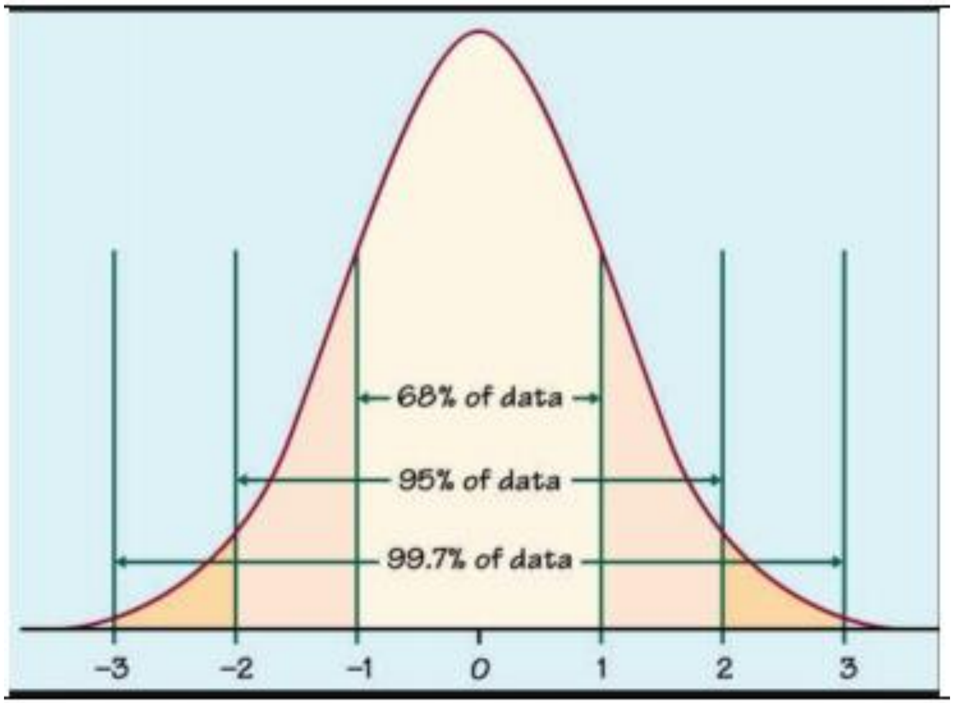
A random variable Z that is distributed standard normal has probability density function

$$f_Z(Z=z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right]$$

$$\text{Z-score} = \frac{X - \mu}{\sigma}$$

From: X distributed Normal(μ,σ^2) To: Z-score distributed Normal(0,1)	**From: Z-score distributed Normal(0,1) To: X distributed Normal(μ,σ^2)**
Use this when you want to calculate probabilities of event occurrence *(areas under the curve)*	**Use this when you want to obtain the values of percentiles of interest, eg; median, P25, etc.**
$\text{Z-score} = \frac{X - \mu}{\sigma}$	$X_{ptile} = \sigma\left[\text{Z-score}_{ptile}\right] + \mu$

Loi normale



Important to Remember (okay, I happen to think so)

A quick reminder about the sampling distribution of the average.

standard error $[\overline{X}] = \sqrt{\text{variance}[\overline{X}]} = \text{SE}[\overline{X}] = (s/\sqrt{n})$

```
# Calculate Pr[Normal(mean=0, sd=1) <= 1.82], two ways:
pnorm(1.82)
round(pnorm(1.82,mean=0,sd=1,
lower.tail=TRUE),digits=4))
```

```
# Calculate Pr[Normal(mean=0, sd=1) >= 2.38], two ways: simple, fancy
round(pnorm(2.38,mean=0,sd=1,lower.tail=FALSE),digits=4))

# Calculate Pr[-2.58 <= Normal(mean=0, sd=1) <= 0.58]
round(pnorm(0.58,mean=0,sd=1) - pnorm(-2.58,mean=0,sd=1),digits=4))
```

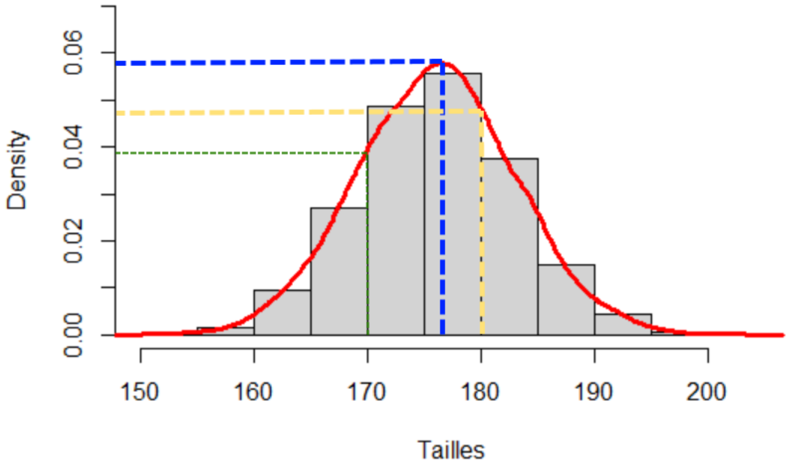
Loi normale

1.	**If a "generic" random variable Y is distributed Normal with** $\mu_{Y} = E(Y)$ $\sigma_{Y}^{2} = Var(Y)$	$\text{z-score} = \dfrac{Y\text{-}E(Y)}{\sqrt{var(Y)}}$ **is distributed Normal(0,1)**
2.	**If X is distributed Normal (μ, σ^2)**	$\text{z-score} = \dfrac{X-\mu}{\sigma}$ **is distributed Normal(0,1)**
3.	**If $\bar{X}$ is distributed Normal (μ, σ^2/n)**	$\text{z-score} = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ **is distributed Normal(0,1)**

Loi normale

La taille moyenne des hommes est de 176 cm avec un écart type de 7 cm.

```
dnorm(x, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1,
lower.tail = TRUE)
```



Q? Quelle est la probabilité qu'un homme pris au hasard ait exactement une taille de 170 cm ? 176 cm ?

```
dnorm(x=170, mean=176, sd=7)
dnorm(x=176, mean=176, sd=7)
```

0.039
0.057

Q? Quelle est la limite de taille au dessous de laquelle on retrouve 75% des hommes ?

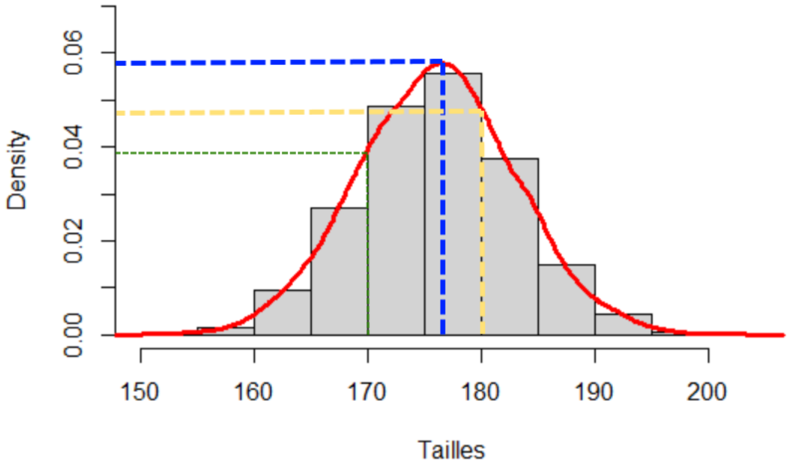
```
qnorm(p=.75, mean=176, sd=7)
```

180.72 cm

Loi normale

La taille moyenne des hommes est de 176 cm avec un écart type de 7 cm.

```
pnorm(q, mean = 0, sd =
1, lower.tail = TRUE)
```



Q? Quelle est la probabilité qu'un homme ait une taille de moins de 1,80 m ?

`pnorm(180,mean=176,sd=7)` 0.716

Q? Quelle est la probabilité qu'un homme ait une taille de moins de 1,70 m ?

`pnorm(170,mean=176,sd=7)` 0.195

Q? Quelle est la probabilité qu'un homme ait une taille entre 1,70m et 1,80 m ?

`pnorm(180,mean=176,sd=7) - pnorm(170,mean=176,sd=7)` 0.520

Distribution des moyennes d'échantillonnage

The Central Limit Theorem

IF

1) **We have an independent random sample of n observations $X_1 \ldots X_n$; and**
2) **the $X_1 \ldots X_n$ are all from the same distribution, *whatever that is.*; and**
3) **this distribution has mean = μ and variance = σ^2**

THEN as $n \to \infty$

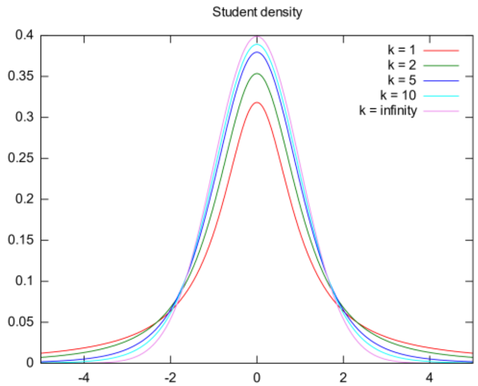
the sampling distribution of $\bar{X}_n = \left[\frac{\sum_{i=1}^{n} X_i}{n} \right]$ **is eventually**

Normal with mean = μ and variance = σ^2/n

<u>In words</u>:
"In the long run, averages have distributions that are well approximated by the Normal"

"The sampling distribution of $\bar{X}_n$, upon repeated sampling, is eventually Normal $\left(\mu, \frac{\sigma^2}{n} \right)$

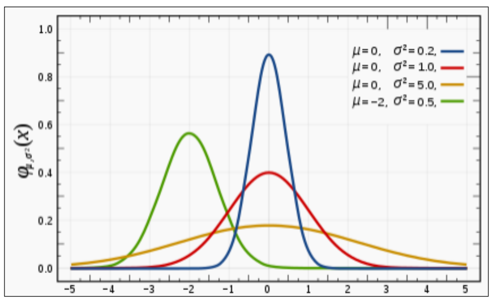
Loi de Student, loi du Chi 2



Loi de Student

La densité de T, notée f_T, est donnée par :

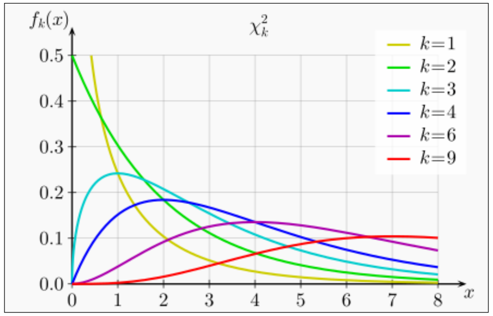
$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \text{ pour } k > 0.$$



Loi normale

Fonction densité

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



Loi du Chi 2

En fonction de variables de loi normale

$$\sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

Les tests d'hypothèse

La démarche hypothético-déductive

Etape 1 : Formuler une hypothèse nulle (H0) : « Pas de différence »

Etape 2 : Formuler une hypothèse alternative (H1) : « Différence »

Etape 3 : Prévoir le comportement d'une loi de probabilité si H0 est vraie

Etape 4 : Observer le comportement de la loi de probabilité sur l'échantillon observé : données de l'étude

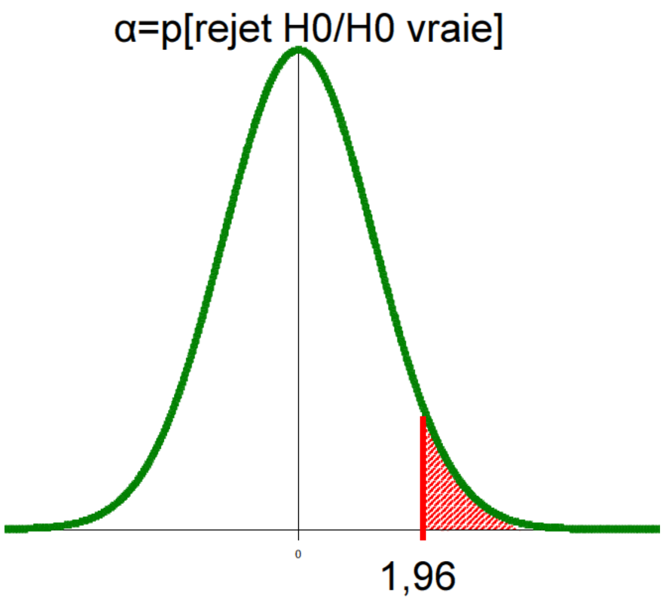
Etape 5 : Confronter les observation aux prédictions sous H0

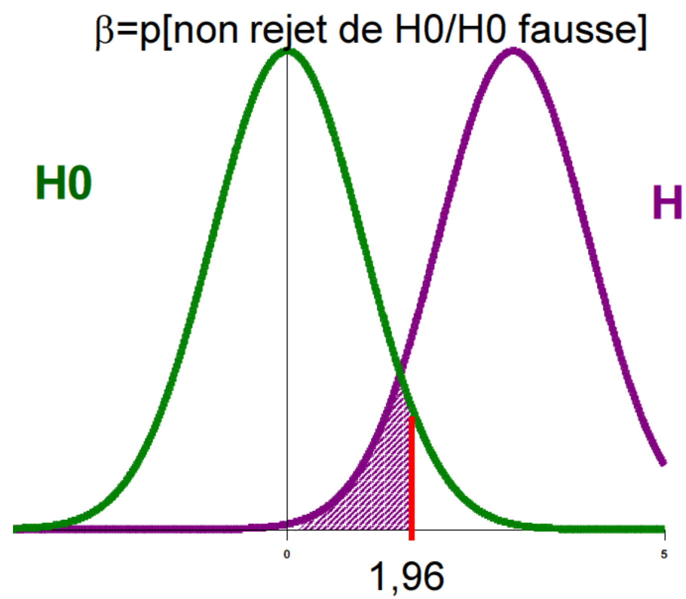
Etape 6 : Conclure :

- **Si les observations sont conformes aux prédictions : Non rejet de H0**
- **Si les observations divergent des prédictions : Rejet de H0**

p-value = probabilité que les observations soient conformes aux prédictions sous H0

Les tests d'hypothèse





α = risque de rejeter H0 alors que H0 est vraie

β = risque de ne pas rejeter H0 alors que H0 est fausse

Les tests d'hypothèse

On conclut en assumant un risque de se tromper

Usuellement on fixe **α** = 5%

Si p-value < 0,05 alors le test est significatif

On met en évidence une différence

Si p-value > 0,05 alors le test est non significatif

On ne met pas en évidence une différence

Une comparaison numérique entre observations et prédictions permet de préciser le sens (ordre) de la différence

Les tests d'hypothèse

Lien entre 2 variables qualitatives

Ex. 2 EU Masters. Si on choisit le 1er, a-t-on plus de chances de choisir le 2ème ?

- Test de Chi2 (paramétrique)

$$\chi^2 = \sum \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

Suit chi2 à (l-1)(c-1) ddl

Conditions

- $C_{ij} > 5$
- Indépendance des individus

inégalité de Bonferroni

$$\alpha_{total} \leq \sum_{i=1}^{k\ tests} \alpha_i \longrightarrow$$

Si tests multiples

- Test exact de Fisher

(non paramétrique, combinatoire)

Probabilité d'observer 1 tableau donné, si H0 vraie

$$p_i = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

Probabilité d'observer tableau au moins aussi grand, si H0 vraie: "p"

$$p = \sum_{i=1}^{K} p_i$$

Les tests d'hypothèse

Exemple : la déficience en G6PD est-elle dépendante du sexe ?

```
chisq.test(data$sex, data$G6PD)

K=chisq.test(data$sex, data$G6PD)
K$expected
```

Exemple : la déficience en G6PD est-elle dépendante du sexe ?

```
Si Min[K$expected]<5

fisher.test(data$sex, data$G6PD)
```

Les tests d'hypothèse

Lien entre 2 variables quantitatives

Ex. le Poids et le Taille sont-ils liés par une relation linéaire ?

Entre 2 variables quantitatives :

-Test du coeff de corrélation de Pearson

$$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \rightarrow T_{n-2}$$

Loi de Student

Conditions d'applications

Relation linéaire entre X et Y ++

Loi conditionnelle Normale $L(Y/X) \rightarrow N$

Variance conditionnelle constante $\mathrm{var}(Y/X)$

Independence des individus

-Test du coeff de corrélation de Spearman

(Evalue la corrélation des rangs. CA : relation monotone)

Les tests d'hypothèse

Exemple : le taux d'Hb0 est-il lié à l'âge ?

```
plot(data$sex, data$G6PD)
abline(lm(data$hb0~data$age))
cor.test(data$age, data$hb0, method = "pearson")
```

Si conditions non paramétriques :

```
cor.test(data$age, data$hb0, method = “sperman")
```

Les tests d'hypothèse

Vous étudiez le lien :

Entre une variable qualitative à 2 modalités et une variable quantitative

Ex. La taille moyenne des Hommes est-elle identique à celle des Femmes ?

-Test de Student (robuste ++)

$$T = \frac{|\mu_{NF} - \mu_F|}{\sqrt{\sigma^2\left(\frac{1}{n_{NF}} + \frac{1}{n_F}\right)}} \rightarrow T_{n_{NF}+n_F-2}$$

Loi de Student

Conditions d'applications

$X \rightarrow N(\mu;\sigma^2)$ OU $n_{NF} > 30\ ET\ n_F > 30$

ET

$\sigma^2_{NF} = \sigma^2_F$

ET

Indépendance des individus

-Test Wilcoxon / Mann-Whitney (texte l'égalité de la somme des rangs pour chaque groupe après mixage des données des 2 groupes)

Les tests d’hypothèse

Exemple : la variation du taux d’hémoglobine est plus élevée chez les femmes ?

```
boxplot(data$age~data$sex)
t.test(data$age~data$sex)
```

Si conditions non paramétriques (ex absence de normalité) :

```
wilcox.test(data$age~data$sex)
```

Les tests d'hypothèse

Vous étudiez le lien :

Entre une variable qualitative à plus de 2 modalités et une variable quantitative : Notes 3 disciplines

-Test de l'ANOVA

$$F = \frac{\sigma_A^2}{\sigma_R^2} \rightarrow F_{n-k}^{k-1}$$

Loi de Fisher

Conditions d'applications

- Loi de X Normale dans chaque groupe (ou effectifs >30 dans chaque groupe)
- Variance de X constante
- Independence des individus

inégalité de Bonferroni

$$\alpha_{total} \leq \sum_{i=1}^{k\ tests} \alpha_i \longrightarrow$$

Si des comparaisons en sous-groupes sont effectuées ensuite

-Test Kruskal–Wallis $\longrightarrow$ Si les conditions paramétriques ne sont pas remplies

Les tests d’hypothèse

Exemple : la variation du taux d’hémoglobine dépend du traitement reçu ?

```
boxplot(data$deltadj~data$TTT)
A=aov(data$deltadj~data$TTT)
summary(A)
```

Conditions

```
qqnorm(data$deltadj[data$TTT=="A"])
qqline(data$deltadj[data$TTT=="A"], col="red")
qqnorm(data$deltadj[data$TTT=="AB"])
qqline(data$deltadj[data$TTT=="AB"], col="blue")
```

Si conditions non paramétriques (ex absence de normalité) :

```
kruskal.test(data$deltadj~data$TTT)
```

Les tests d'hypothèse

Test paramétrique

Requiert un modèle à fortes contraintes (normalité des distributions, égalité des variances)

Test non paramétrique

Test dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon.

-Ne nécessite aucune hypothèse de distribution (surtout pas de contrainte de normalité)

-Peut néanmoins exiger des conditions d'application (certes peu contraignantes) : « peu » d'ex-aequo pour les tests basés sur les rangs, relation monotone, et toujours indépendance des individus

-Puissance faible

-Calculs nombreux et difficiles (moins perceptible avec les logiciels + processeurs puissants)

-Exemples: test de Kolmogorov-Smirnov (rangs), test exact Fisher (combinatoire)

Nombre de sujets nécessaire (NSN)

Nombre minimal de sujets nécessaire pour déceler une différence entre 2 moyennes

nombre de sujets nécessaires par groupe :

connaissance

$$n_A = 2 \times \frac{s^2}{\Delta^2} \times \left(z_{\alpha/2} - z_{puis}\right)^2$$

À définir

α=5% ⇔1,96

1-β=80% ⇔-0,842

Différence pertinente pour un « praticien »

Nombre minimal de sujets nécessaire pour déceler une différence entre 2 moyennes

α=5% ⇔1,96

1-β=80% ⇔-0,842

$$n_A = \frac{\left(z_{\alpha/2} - z_{puis}\right)^2}{2\left(\arcsin\sqrt{P_1} - \arcsin\sqrt{P_2}\right)^2}$$

connaissance

À définir

Puissance d'un test statistique : capacité à montrer un effet lorsqu'il existe. Habituellement fixé à **80%**

Merci
Fin de la Partie 3 : Lois de probabilités et tests d’hypothèses

A suivre : Modèles multivariés et fonctions « avancées »