Tests statistique

M2 Radiophysique médicale, INSTN, 2023

Clément GAUCHY (clement.gauchy@cea.fr)
CEA Saclay

Sommaire

1. Définitions & généralités

2. Démarche d'un test et quantification de l'erreur

3. Choix du test

4. Test du rapport de vraisemblance et boson de Higgs

Objectif

On cherche à prendre des décisions à partir de données

Il va donc s'agir de décider si les différences observées entre un modèle posé a priori et les observations sont *significatives* ou bien sont dus au hasard

Réaliser un *test statistique* consiste à

1. Confronter une hypothèse avec les observations réelles
2. Prendre une décision par la suite

Généralités sur les tests statistiques

Un test statistique est une **procédure de décision** entre 2 hypothèses au vu d'un échantillon d'observations.

Généralités sur les tests statistiques

Un test statistique est une **procédure de décision** entre 2 hypothèses au vu d'un échantillon d'observations.

On appelle **l'hypothèse nulle** notée $\mathcal{H}_0$ une question Oui/Non que l'on cherche à valider ou refuter à l'aide des données.

Exemples:

- "Le médicament utilisé est il efficace ?"
- "Le processus de fabrication est il conforme ?"
- "La variable aléatoire X suit une loi normale ?"

L'hypothèse alternative est notée $\mathcal{H}_1$.

⚠ Les deux hypothèses n'ont pas des rôles symmétrique. Par analogie avec la justice, $\mathcal{H}_0$ est la *présomption d'innocence*.

Modélisation mathématique

On définit généralement à partir des observations une **statistique de test** notée S tel que

- S résume l'information de l'échantillon,
- On connaît la loi de S **en supposant $\mathcal{H}_0$ vraie.**

Avec un jeu de données $(x_i)_{1 \leq i \leq n}$, on évalue la statistique $S(x_1, \ldots, x_n)$ et on regarde si elle est "cohérente" avec $\mathcal{H}_0$.

On appelle W la région critique du test l'ensemble des valeurs de $S(x_1, \ldots, x_n)$ pour lesquelles $\mathcal{H}_0$ est rejeté.

Exemple: Moyenne d'une loi normale avec écart-type connu

On suppose observer un échantillon $(X_1, \dots, X_n)$ suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec σ connu.

Exemple: Moyenne d'une loi normale avec écart-type connu

On suppose observer un échantillon $(X_1, \dots, X_n)$ suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec σ connu.

Hypothèse: $\mathcal{H}_0 : \ \mu = \mu_0, \ \mathcal{H}_1 : \ \mu > \mu_0$ (test unilatéral)

Exemple: Moyenne d'une loi normale avec écart-type connu

On suppose observer un échantillon $(X_1, \ldots, X_n)$ suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec σ connu.

Hypothèse: $\mathcal{H}_0 : \mu = \mu_0$, $\mathcal{H}_1 : \mu > \mu_0$ (test unilatéral)

Statistique de test: La moyenne empirique $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Exemple: Moyenne d'une loi normale avec écart-type connu

On suppose observer un échantillon $(X_1, \ldots, X_n)$ suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec σ connu.

Hypothèse: $\mathcal{H}_0 : \mu = \mu_0$, $\mathcal{H}_1 : \mu > \mu_0$ (test unilatéral)

Statistique de test: La moyenne empirique $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Critère de choix: On rejette $\mathcal{H}_0$ si $\sqrt{n}(\bar{X}_n - \mu_0)/\sigma > q_\alpha$ où q_α est le quantile de niveau $1 - \alpha$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

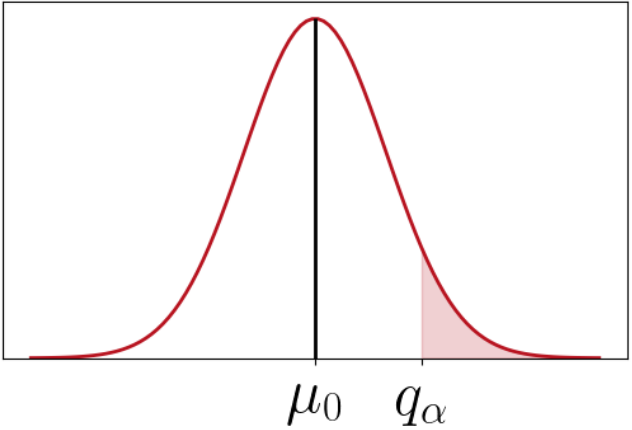
Exemple: Moyenne d'une loi normale avec écart-type connu

On suppose observer un échantillon $(X_1, \dots, X_n)$ suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec σ connu.

Hypothèse:
$\mathcal{H}_0 : \ \mu = \mu_0, \ \mathcal{H}_1 : \ \mu > \mu_0$
(test unilatéral)

Statistique de test: La moyenne empirique $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Critère de choix: On rejette $\mathcal{H}_0$ si $\sqrt{n}(\bar{X}_n - \mu_0)/\sigma > q_\alpha$ où q_α est le quantile de niveau $1 - \alpha$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$.



Sommaire

1. Définitions & généralités
2. Démarche d'un test et quantification de l'erreur
3. Choix du test
4. Test du rapport de vraisemblance et boson de Higgs

Un test statistique, ça trompe énormément !

Décision \ Vérité	$\mathcal{H}_0$ est vrai	$\mathcal{H}_0$ est fausse
$\mathcal{H}_0$ acceptée	$1 - \alpha$	$1 - \beta$
$\mathcal{H}_0$ rejetée	$\alpha = \mathbb{P}(\mathcal{H}_0$ rejetée$\vert\mathcal{H}_0$ vraie$)$	$\beta = \mathbb{P}(\mathcal{H}_0$ rejetée$\vert\mathcal{H}_0$ fausse$)$

Le *seuil* α d'un test statistique est la probabilité d'avoir un *faux-positif* (on rejette $\mathcal{H}_0$ alors qu'elle est vraie). On l'appelle aussi *erreur de première espèce*.

La *puissance* β d'un test statistique est la probabilité de rejeter $\mathcal{H}_0$ à raison.

La probabilité $1 - \beta$ d'accepter $\mathcal{H}_0$ alors qu'elle est fausse s'appelle *l'erreur de deuxième espèce*.

La probabilité $1 - \alpha$ d'accepter $\mathcal{H}_0$ alors qu'elle est vraie est appellée *niveau de confiance*

Un test statistique, ça trompe énormément !

Vérité / Décision	$\mathcal{H}_0$ est vrai	$\mathcal{H}_0$ est fausse
$\mathcal{H}_0$ acceptée	$1-\alpha$	$1-\beta$
$\mathcal{H}_0$ rejetée	$\alpha = \mathbb{P}(\mathcal{H}_0 \text{ rejetée} \mid \mathcal{H}_0 \text{ vraie})$	$\beta = \mathbb{P}(\mathcal{H}_0 \text{ rejetée} \mid \mathcal{H}_0 \text{ fausse})$

Le *seuil* α d'un test statistique est la probabilité d'avoir un *faux-positif* (on rejette $\mathcal{H}_0$ alors qu'elle est vraie). On l'appelle aussi *erreur de première espèce*.

La *puissance* β d'un test statistique est la probabilité de rejeter $\mathcal{H}_0$ à raison.

La probabilité $1-\beta$ d'accepter $\mathcal{H}_0$ alors qu'elle est fausse s'appelle *l'erreur de deuxième espèce*.

La probabilité $1-\alpha$ d'accepter $\mathcal{H}_0$ alors qu'elle est vraie est appellée *niveau de confiance*

Question: Que souhaite t'on maximiser/minimiser entre α ou β ?

Un test statistique, ça trompe énormément !

Vérité / Décision	$\mathcal{H}_0$ est vrai	$\mathcal{H}_0$ est fausse
$\mathcal{H}_0$ acceptée	$1 - \alpha$	$1 - \beta$
$\mathcal{H}_0$ rejetée	$\alpha = \mathbb{P}(\mathcal{H}_0 \text{ rejetée} \mid \mathcal{H}_0 \text{ vraie})$	$\beta = \mathbb{P}(\mathcal{H}_0 \text{ rejetée} \mid \mathcal{H}_0 \text{ fausse})$

Le *seuil* α d'un test statistique est la probabilité d'avoir un *faux-positif* (on rejette $\mathcal{H}_0$ alors qu'elle est vraie). On l'appelle aussi *erreur de première espèce*.

La *puissance* β d'un test statistique est la probabilité de rejeter $\mathcal{H}_0$ à raison.

La probabilité $1 - \beta$ d'accepter $\mathcal{H}_0$ alors qu'elle est fausse s'appelle *l'erreur de deuxième espèce*.

La probabilité $1 - \alpha$ d'accepter $\mathcal{H}_0$ alors qu'elle est vraie est appellée *niveau de confiance*

Question: Que souhaite t'on maximiser/minimiser entre α ou β ?

Réponse: Parmis tout les tests de niveau α, on cherche celui maximisant la puissance.

Exemple

Soit μ la moyenne du niveau de radioactivité de l'eau en picocuries par litres. La valeur $\mu_0 = 5$ est considérée comme une valeur seuil entre eau potable et non potable. On peut tester $\mathcal{H}_0$: "$\mu \geq 5$" contre $\mathcal{H}_1$: "$\mu < 5$".

L'erreur de première espèce (faux-positif) conduirait de laisser boire de l'eau non potable.

L'erreur de deuxième espèce (faux-négatif) conduirait à déclarer non potable de l'eau saine.

$\hookrightarrow$ Asymmétrie entre les deux types d'erreurs ! Rejeter $\mathcal{H}_0$ à raison (1ere espèce) a beaucoup plus de conséquence que de la conserver à tort ...

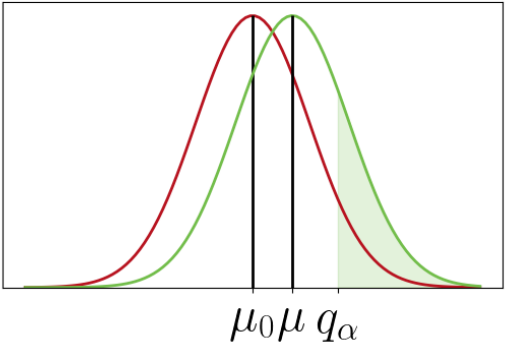
Puissance du test

Exemple pour le test sur la moyenne d'une Gaussienne avec écart-type connu.

La puissance $\beta(\mu)$ dépend donc de la moyenne ! Elle se calcul de la façon suivante:

$$\beta(\mu) = \mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}(\bar{X}_n > q_\alpha)$$

La puissance est une fonction de μ car tout les $\mu > \mu_0$ sont dans l'hypothèse alternative $\mathcal{H}_1$



Puissance d'un test

La puissance d'un test portant sur la valeur d'un paramètre réel θ est la fonction de θ définie par:

$$\begin{array}{rcccc} \beta & : & \mathbb{R} & \to & [0,1] \\ & & \theta & \mapsto & \mathbb{P}_\theta(S(X_1,\ldots,X_n) \in W) \end{array}$$

Le seuil du test est $\alpha = \sup_{\mathcal{H}_0} \beta(\theta)$. Cela correspond à la probabilité maximale de rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie.

Démarche d'un test statistique

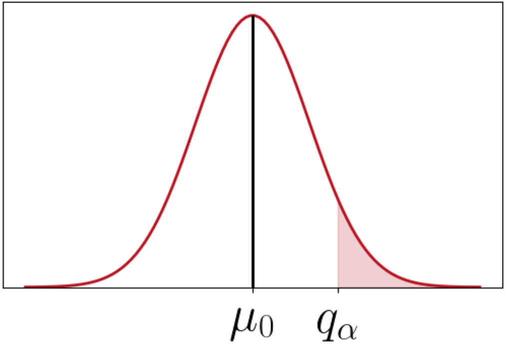
- Choix de $\mathcal{H}_0$ et $\mathcal{H}_1$. Fixer le niveau α
- Détermination de la statistique de test $S(X_1, \ldots, X_n)$
- Allure de la région critique en fonction de $\mathcal{H}_1$
- Calcul de la région critique en fonction de α et $\mathcal{H}_0$
- Calcul de la valeur observée de la statistique de test
- Rejet ou acceptation de $\mathcal{H}_0$ au seuil α
- Si possible, calcul de la puissance du test

p-valeur

La p-valeur est la probabilité d'observer en supposant $\mathcal{H}_0$ vraie d'être dans la zone de rejet.

La p-valeur permet d'avoir une information quantitative sur le rejet de $\mathcal{H}_0$.

On considère le rejet de $\mathcal{H}_0$ significatif à partir de $p < 0.05$. ⚠ Cela dépend des applications ! En physique des particules, on cherche $p < 10^{-7}$



La zone rouge correspond à la p-valeur.

$$p = \mathbb{P}_{\mathcal{H}_0}(\bar{X}_n > q_\alpha)$$

Sommaire

1. Définitions & généralités

2. Démarche d'un test et quantification de l'erreur

3. Choix du test

4. Test du rapport de vraisemblance et boson de Higgs

Zoologie des tests statistique

- **Test paramétriques**
  - Un échantillon
    - Test sur la moyenne, variance connue (Gaussienne)
    - Test sur la moyenne, variance inconnue et estimée (Student)
    - Test sur une proportion (Binomiale)
  - Deux échantillons indépendants
    - Comparaison des deux moyennes (Student)
    - Comparaison des deux variances (Fisher)
- **Test d'adéquation (non paramétrique)**
  - Comparaison de deux distributions (χ^2)
  - Normalité d'une distribution (Kolmogorov, Shapiro Wilks)

Tests unilatéral ou bilatéral

Exemple avec le test de la moyenne d'une Gaussienne:

Test unilatéral $\mathcal{H}_0 : \mu \leq \mu_0$ contre $\mathcal{H}_1 : \mu > \mu_0$

Test bilatéral $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu \neq \mu_0$

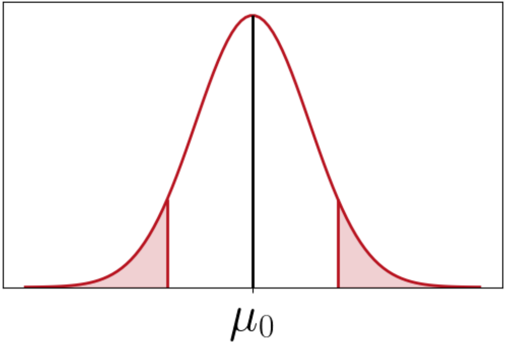


Figure 1: Zone de rejet du test bilatéral en rouge

Test de Student

On observe $(X_1, \ldots, X_n)$ i.i.d. tel que $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ^2 inconnues.

On souhaite tester $\mathcal{H}_0 : \ \mu = \mu_0$ contre $\mathcal{H}_1 : \ \mu \neq \mu_0$

Sous $\mathcal{H}_0$, on a $T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n-1}}$ (avec $\bar{X}_n$ la moyenne empirique et S_n^2 la variance empirique) qui suit la loi de Student $\mathcal{T}_{n-1}$ à $n-1$ degrés de liberté.

On rejette ainsi $\mathcal{H}_0$ au seuil α si $T_n > t_{n-1,(1+\alpha)/2}$ ou $T_n < t_{n-1,(1-\alpha)/2}$.

On peut calculer la puissance $\beta(\mu)$ par simulation Monte-Carlo. (Exercice)

Test d'ajustement (ou d'adéquation)

Durant tout le cours, on a construit des tests portant sur le paramètre θ d'un modèle statistique $\mathcal{M} = \{p_{\theta / \theta \in \Theta}\}$.

Désormais, on cherche à tester si la fonction de répartition $F(x) = \mathbb{P}(X \leq x)$ d'un échantillon $(X_1, \ldots, X_i)$ est égale à une fonction de répartition connue F_0.

On peut par exemple tester si les données suivent la loi normale $\mathcal{N}(0, 1)$.

Test du χ^2

Le test du χ^2 est un test d'adéquation pour les lois dites discrètes.

$(X_1, \ldots, X_k)$ est un échantillon de variables aléatoires i.i.d. tel que X_1 prend ses valeurs dans $\{1, \ldots, k\}$. On se donne alors le vecteur $(p_i)_{1 \leq i \leq k}$ tel que $p_i \geq 0$ et $\sum_i p_i = 1$. On souhaite tester

$\mathcal{H}_0$: Pour tout i de 1 à k, $\mathbb{P}(X_1 = i) = p_i$

contre

$\mathcal{H}_1$: Il existe i de 1 à k tel que $\mathbb{P}(X_1 = i) \neq p_i$.

Exemple: On cherche à déterminer si un dé est biaisé au risque de 1%. Soit X la variable aléatoire qui donne le chiffre obtenu à chaque lancer de dé. On va donc tester $\mathbb{P}(X = i) = 1/6$ pour i de 1 à 6.

Test du χ^2, statistique de test

On note N_i l'effectif observé de la valeur i tandis que np_i correspond à l'effectif espéré de cette valeur sous $\mathcal{H}_0$. On définit la statistique de test par

$$D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \ .$$

On admet que D^2 suit asymptotiquement selon n une loi du χ^2 à $k-1$ degrés de liberté (d'où le nom du test).

On peut alors définir la région critique pour le test du χ^2 de seuil α par $D^2 > q_{\chi^2_{k-1},1-\alpha}$ tel que $\mathbb{P}(D^2 < q_{\chi^2_{k-1},1-\alpha} | \mathcal{H}_0 \text{ vraie}) = 1 - \alpha$

Test de Kolmogorov

Le test de Kolmogorov est employé pour tester si la loi de probabilité de X à valeurs réelles à pour fonction de répartition F_0.

On va tester $\mathcal{H}_0 : \; F = F_0$ contre $\mathcal{H}_1 : \; F \neq F_0$. On va pour cela utiliser l'estimateur empirique de $F(x) = \mathbb{P}(X \leq x)$:

$$
\begin{aligned}
F(x) &= \mathbb{P}(X \leq x) \\
F(x) &= \mathbb{E}[1_{X \leq x}] \\
F(x) &\approx \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}
\end{aligned}
$$

On a donc un estimateur empirique de la fonction de répartition $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}$.

Test de Kolmogorov, statistique de test

La statistique de test est la variable aléatoire

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right| ,$$

en effet, D_n converge en loi vers la distribution de probabilité de Kolmogorov-Smirnov

On définit la région critique pour le test de Kolmogorov-Smirnov au seuil α par $D_n > q_{\mathrm{KS},1-\alpha}$ tel que $\mathbb{P}(D_n < q_{KS,1-\alpha} | \mathcal{H}_0 \text{ vraie}) = 1 - \alpha$.

Question: Comment calculer la puissance de ce test ?

Toujours plus de tests

- Test de Shapiro-Wilk, de Lilliefors, d'Agostino: L'hypothèse nulle est le caractère gaussien des données
- Test d'Anderson-Darling: Même objectif que le test de Kolmogorov
- Test de Mann-Whitney, de Wilcoxon, de Kruskal-Wallis: L'hypothèse nulle est l'égalité des lois de deux variables aléatoire X et Y.

Sommaire

1. Définitions & généralités

2. Démarche d'un test et quantification de l'erreur

3. Choix du test

4. Test du rapport de vraisemblance et boson de Higgs

Test du rapport de vraisemblance

On va considérer un modèle statistique $\mathcal{M} = \{f_\theta,\ \theta \in \Theta\}$ où f_θ désigne la densité de probabilité de X.

On a un échantillon $(X_1, \ldots, X_n)$ i.i.d. distribué selon f_{θ_*}. On veut tester si $\theta_* \in \Theta_0$ où $\Theta_0 \subset \Theta$. On a donc $\mathcal{H}_0 :\ \theta_* \in \Theta_0$ contre $\mathcal{H}_1 :\ \theta_* \notin \Theta_0$.

On note $L(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$ la vraisemblance, on définit la statistique du rapport du vraisemblance de la façon suivante

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \notin \Theta_0} L(\theta)}$$

Test du rapport de vraisemblance

Le statisticien Abraham Wald a démontré le théorème suivant dans les années 40:

Théorème (Loi asymptotique du rapport de vraisemblance)

Soit un échantillon $(X_1, \ldots, X_n)$ *i.i.d. distribué selon* f_{θ_*} *et* $\mathcal{M} = \{f_\theta, \ \theta \in \Theta\}$. *On a* $\Theta_0 = \{\theta_*\}$ *(i.e. on cherche à tester* $\theta = \theta_0$*) alors on a la convergence en loi suivante:*

$$-2\log(\Lambda) \xrightarrow[n \to +\infty]{\mathcal{L}} \chi_1^2$$

On peut donc construire un test asymptotique à partir du rapport de vraisemblance.
Question (rhétorique): Quel est l'intérêt de ce test ?

On rejette $\mathcal{H}_0$ au seuil α si $-2\log(\Lambda) > q_{\alpha,\chi_1^2}$ avec q_{α,χ_1^2} quantile de niveau α de la loi du χ_1^2.

Lemme de Neyman-Pearson

Un test statistique est dit uniformément le plus puissant s'il admet la plus grand puissance parmis tout les tests de seuil α.

Définition formelle: Pour tout test de seuil $\alpha' \leq \alpha$, on a $\forall \theta \notin \Theta_0, \ \beta'(\theta) \leq \beta(\theta)$.

Lemme de Neyman-Pearson: Le test du rapport de vraisemblance est uniformément le plus puissant.

Test de détection du boson de Higgs

Les physiciens des particules se basent sur une théorie que l'on appelle le Modèle Standard.

Le challenge est de déterminer à partir de quantités massives de données (issues du LHC par exemple) de l'existence de nouvelles particules ou non

Problème statistique: Déterminer, avec la plus grande puissance, si les données suggèrent l'existence de nouvelles particules.

Problème statistique

Les données observés sont généralement des comptages. On fait une hypothèse Poissonienne.

$$\mathcal{D} = (X_i)_{1 \leq i \leq n} \text{ i.i.d.}, \; X_1 \sim \mathcal{P}(\lambda)$$

On va effectuer un test d'hypothèses avec $\mathcal{H}_0 : \; \lambda = b$ et $\mathcal{H}_1 : \; \lambda = \mu_H + b$, où b est l'intensité du "bruit de fond" et μ_H l'intensité du signal attribué au boson de Higgs.

Statistique de test

Motivé par le lemme de Neyman-Pearson, on effectue un test du rapport de vraisemblance:

$$S = -2 \log \left(\frac{L(\mathcal{H}_0)}{L(\mathcal{H}_1)} \right) ,$$

où $L(\mathcal{H}_0)$ et $L(\mathcal{H}_1)$ correspondent respectivement à la vraisemblance sous $\mathcal{H}_0$ et $\mathcal{H}_1$.

<u>Exercice:</u> Écrire S en utilisant le modèle de Poisson de la planche précédente.

p-valeur du test

La communauté des physiciens des particules s'accordent pour rejeter l'hypothèse nulle avec une p-valeur très faible (de l'ordre de 10^{-7}).

Dans un cadre Gaussien, cela correspondrait à un écart de 5 fois l'écart type à la moyenne ! D'où le terme de 5σ que l'on entend parfois

Références

- Site web wikistat.fr, `http://wikistat.fr/pdf/st-l-inf-tests.pdf`
- Site web wikistat.fr, `http://wikistat.fr/pdf/st-m-inf-test.pdf`
- E. Gross, *Praticle statistics for high energy physics*, `https://indico.cern.ch/event/614672/contributions/2605123/attachments/1519560/2375162/StatESHEP.pdf`