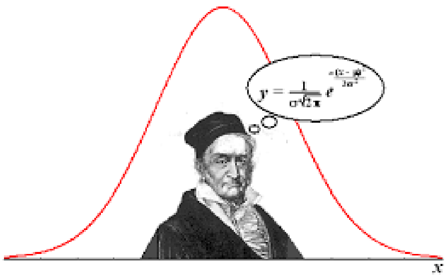


# Sufficiëntie in Exponentiële families

## Theorie en toepassingen in hypothesetoetsen


**Martial Luyts**

KU Leuven

martial.luyts@kuleuven.be


LEUVEN STATISTICS RESEARCH CENTRE

# Contents

# Part 1:

# Introductie

# 1. Leerdoelen

Na deze les kunnen jullie:

- De notie van sufficiëntie definieren en interpreteren
- De **Neyman-Fisher factorisatiestelling** toepassen om sufficiënte statistieken te herkennen
- Zelf sufficiënte statistieken afleiden voor **verdelingen uit de exponentiële familie** (bv. normaal, Poisson)
- De rol van sufficiëntie in **likelihood ratio testen** begrijpen en toepassen
- Een likelihood ratio test uitvoeren en interpreteren, zowel met als zonder gebruik van sufficiëntie

# 2. Statistisch inferentie probleem

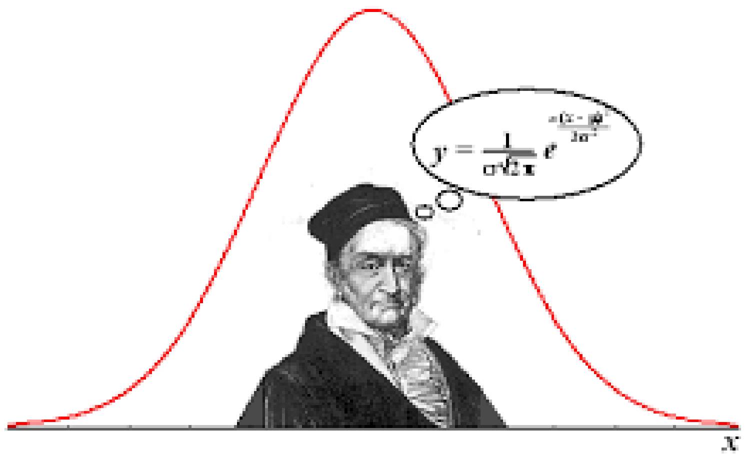
Veronderstel de volgende setting:

- Een stochastische veranderlijke (sv) uit een populatie: $X$
- Steekproef: $(X_1, \ldots, X_n)$
- Steekproefwaardes $(x_1, \ldots, x_n)$
- Steekproefruimte: verzameling van mogelijke steekproef waardes $(x_1, \ldots, x_n)$ (hier, onafhankelijke en identieke verdeelde steekproeven).

- Het **statistisch inferentie probleem**: Maak (betrouwbare) conclusies over de verdeling van $X$, gebaseerd op de data.
- Het **parametrisch schattingsprobleem**: Onderzoek de ongekende parameter $\theta$ in populatie $X$, gebaseerd op de data.

**Opmerking: Parametrische schatting**

▷ We assumeren een parametrisch model voor $X$, d.w.z.,
$X$ heeft verdeling $y = f(x; \lambda)$ dat gespecificeerd is met model parameter $\lambda$.



▷ We zijn geïnteresseerd in het onderzoek naar parameter $\theta = g(\lambda)$.

**Voorbeelden:**

$*$ $X \sim \text{Pois}(\mu)$: model parameter $\mu$ is de parameter onder interesse $\theta = \mu$,

$*$ $X \sim N(\mu, \sigma^2)$: model parameter $\lambda = (\mu, \sigma)$ (parameter vector) is de parameter onder interesse parameter, b.v. $\theta = (\mu, \sigma)$ of $\theta = \mu$ of $\theta = \sigma/\mu$ of $\theta = E(X^2) = \sigma^2 + \mu^2$.

Vaak beschrijft $\theta$ de volledige of partiële model parameter.

**Om statistische inferentie uit te oefenen, worden er 3 procedures gebruikt:**

1. Punt estimatie

   ⇒ **(punt) schatter** voor $\theta$: een statistiek (stochastische veranderlijke)

   $$T = T\left(X_1, \ldots, X_n\right)$$

   gebruiken om de populatie parameter $\theta$ te schatten of approximeren.

   ⇒ **(punt) schatting**: numerieke waarde van de schatter geevalueerd voor de gegeven dataset, m.a.w. een geobserveerde waarde

   $$T_{\text{obs}} = T\left(x_1, \ldots, x_n\right)$$

   *Voorbeeld*: parameter $\mu$, schatter $\overline{X}$, schatting $\overline{x}$.

2. Betrouwbaarheidsintervallen

3. Statistische testen

- **Doel:** Vind een goede schatter $T$ voor $\theta$.

- **Vraag:** Maar wat zijn de optimale eigenschappen van T zodat we kunnen zeggen dat $T$ een goede schatter is voor $\theta$?

- **Answer:** Basis filosofie

| | **Natuurlijke doelstellingen** voor een goede schatter $T = T(X_1, \dots, X_n)$ voor parameter $\theta$ | **Formele statistische concepten** omtrent optimale eigenschappen van een schatting |
|---|---|---|
| 1. | $T$ neemt waardes "dicht" bij de echte parameter schatting $\theta$ met hoge waarschijnlijkheid | • Unbiased schatter<br>• Min variantie unbiased schatter<br>• Mean squared error (MSE) |
| 2. | Voor grote steekproeven, $T$ zou bijna een perfect schatter moeten zijn: als $n \to \infty$,<br>• $T$ convergeert naar $\theta$<br>• $T$ convergeert naar $\theta$ zo snel als mogelijk | Asymptotische eigenschappen:<br>• Consistente schatter<br>• Asymptotische MSE<br>• Asymptotische relatieve efficiëntie |
| 3. | $T$ moet alle informatie over $\theta$ absorberen die beschikbaar is in de steekproef | • Sufficiënte schatter |

- In de volgende slides zullen dieper ingaan over het topic **sufficientie** bij schatters, specifiek in de context waar de verdeling van $X$ voortvloeit vanuit de exponentiële familie!

# Part 2:

# Sufficiëntie in de exponentiële familie

# 1. Exponentiële familie

DEFINITIE:

Een verdeling $f(x \mid \lambda)$ behoort tot de **exponentiële familie** als ze kan geschreven worden in de vorm:

$$f(x \mid \lambda) = h(x) \cdot \mathsf{exp}[\eta(\lambda) \cdot T(x) - A(\lambda)]$$

Binnen deze familie bevinden zich vele **welgekende verdelingen vanuit de statistiek** (normaal, Poisson, exponentieel, etc.).

## Voorbeeld 1: Poisson verdeling

Laat $X \sim \text{Poi}(\lambda)$, met kansverdeling

$$f(x \mid \lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

Deze verdeling behoort tot de exponentiele familie:

$$f(x \mid \lambda) = \frac{1}{x!} \cdot \exp\left[x \cdot \log(\lambda) - \lambda\right]$$

Hier:

- $h(x) = \frac{1}{x!}$
- $\eta(\lambda) = \log(\lambda)$
- $T(x) = x$
- $A(\lambda) = \lambda$

## Voorbeeld 2: Normale verdeling met gekende variantie

Laat $X \sim N(\mu, \sigma^2)$, met gekende $\sigma^2$ en kansverdeling:

$$f(x \mid \lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Herschrijven:

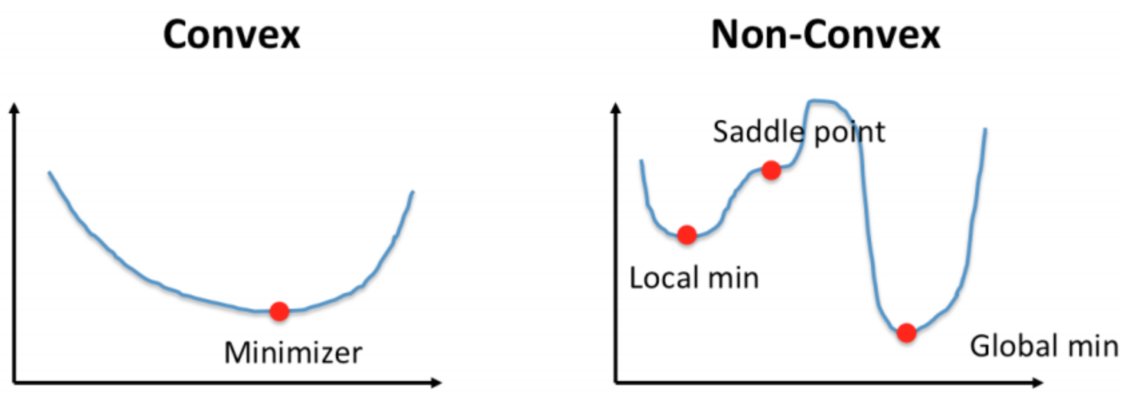
$$f(x \mid \lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\right)$$

Hier:

- $h(x) = \text{exp}\left(-\frac{x^2}{2\sigma^2}\right)$
- $\eta(\mu) = \frac{\mu}{\sigma^2}$
- $T(x) = x$
- $A(\mu) = \frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi\sigma^2}\right)$

# INTERESSANTE EIGENSCHAPPEN:

- **Sufficiëntie:** In de exponentiële familie is $T(x)$ automatisch een **sufficiënte statistiek** voor de ongekende parameter(s) (zie later)
- **Likelihood ratio testen** worden eenvoudiger (zie later)
- Exponentiële families hebben convexiteit van de log-likelihood



OEFENING:

Toon aan dat de binomiale en exponentiële verdeling behoort tot de exponentiële familie!

- $X \sim \text{Bin}(n, p) : f(x \mid p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, \qquad x = 0, ..., n$
- $X \sim \text{Exp}(\lambda) : f(x \mid \lambda) = \text{exp}\left[\text{log}(\lambda) - \lambda \cdot x\right], \qquad x \geq 0$

# 2. Sufficiëntie

- Sufficiëntie kan gezien worden als een onderdeel voor het bepalen van een goede schatter $T$ voor parameter $\theta$
- Om dit begrip uit te leggen, zullen we een **intuïtief voorbeeld** aanhalen:
  - Veronderstel dat we onafhankelijke en identieke verdeelde steekproeven $x = (x_1, \ldots, x_n)$ nemen van een gekende verdeling met ongekende parameter $\theta$.
  - Stel voor dat we 2 personen hebben:
    - Statisticus A: Kent de volledige steekproef, krijgt $n$ kwantiteiten: $x = (x_1, \ldots, x_n)$.
    - Statisticus B: Kent enkel $T(x_1, \ldots, x_n) = t$, 1 numerieke waarde welke een functie is van de steekproefwaardes. Bijvoorbeeld, de som of maximum van de steekproefwaardes.

- Heuristisch, $T(x_1, \ldots, x_n)$ is een sufficiënte statistiek als Statisticus B een even goede job kan verrichten als Statistics A, met "minder informatie".

  - Bijvoorbeeld, als de steekproefwaardes afkomstig zijn van een Bernoulli verdeling, dan is het kennen van $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ (aantal keer hoofd bij een munt opgooi) even goed als het kennen van alle individuele uitkomsten, aangezien een goede schatter het aantal keer hoofd zou zijn over de totaal aantal trials. M.a.w., we hechten geen belang aan de ordening van de uitkomsten, gewoon hoeveel hoofden tevoorschijn kwamen.

## DEFINITIE:

Een statistiek $T(X)$ is sufficiënt voor parameter $\theta$ als de voorwaardelijke verdeling van $X$ gegeven $T(X)$ onafhankelijk is van $\theta$:

$$f(x_1, \ldots, x_n \mid \theta, T = t) = f(x_1, \ldots, x_n \mid T = t)$$

**Intuïtief:** Alle informatie over $\theta$ die in de data $X$ zit, is vervat in T(X)

## Ons voorbeeld:

- **Herinner:** Statisticus A heeft alle steekproefwaardes $x_1, \ldots, x_n$ maar Statisticus B heeft enkel 1 numerieke waarde $t = T(x_1, \ldots, x_n)$.

- Het idee is, Statisticus B kent enkel T = t, maar aangezien T sufficiënt is, heeft hij/zij $\theta$ niet nodig om nieuwe steekproeven $X'_1, \ldots, X'_n$ te genereren van de verdeling.

- Dit is omdat $f(x_1, \ldots, x_n \mid \theta, T = t) = f(x_1, \ldots, x_n \mid T = t)$ en aangezien hij/zij T=t kent, kent hij de voorwaardelijke verdeling (kan nieuwe steekproeven generen!).

- Nu heeft Statisticus B $n$ steekproeven van de verdeling, idem aan Statisticus A. Door deze steekproeven $X'_1, \ldots, X'_n$ te gebruiken, kan Statisticus B een even goede job verrichten als Statisticus A met steekproeven $X_1, \ldots, X_n$ (gemiddeld). M.a.w. geen enkele heeft een nadeel!

- **Probleem:** De formele definitie is vaak moeilijk te checken in de praktijk.
- Er bestaat een stelling dat helpt bij het checken van sufficiëntie bij statistieken.
- Deze stelling noemt de **Neyman-Fisher factorisatiestelling**!

# 3. Neyman-Fisher Factorisatiestelling

STELLING:

Veronderstel $X_1, \ldots, X_n$ onafhankelijke en identiek verdeelde stochastische veranderlijke met verdelingsfunctie $f_X(x \mid \theta)$.

Een statistiek $T = T(X_1, \ldots, X_n)$ is sufficiënt voor $\theta$ a.s.a. er een niet-negatieve functie $g$ and $h$ bestaat zodat de gezamelijke verdelingsfunctie $f_X(X_1, \ldots, X_n \mid \theta)$ gefactoriseerd kan worden als volgt:

$$f_X(X_1, \ldots, X_n \mid \theta) = h(X_1, \ldots, X_n) \cdot g(T(X_1, \ldots, X_n); \theta)$$

De gezamelijke verdelingsfunctie kan dus gesplitst worden in een product met 2 functies:

- Functie $h$: Hangt af van de gehele data, maar niet $\theta$
- Functie $g$: Hangt af van $\theta$, maar enkel op data doorheen the sufficiënte statistiek $T$.

M.a.w., $T$ is het enigste aspect dat een interactie toelaat tussen $X_1, \ldots, X_n$ en $\theta$!

**Opmerking:** De gezamelijke verdeling $f_X(X_1, \ldots, X_n \mid \theta)$ noemen we de **likelihood van de data**, vaak genoteerd als $L(\theta \mid X_1, \ldots, X_n)$.

- Aangezien $X_1, \ldots, X_n$ onafhankelijke en identiek verdeeld stochastische veranderlijke zijn met verdeling $f_X(x; \theta)$, krijgen we:

$$L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

## Voorbeeld 1: Uniforme verdeling

Veronderstel $x_1, \ldots, x_n$ onafhankelijke en identiek verdeelde steekproeven van Unif$(0, \theta)$. Toon aan dat $T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ een sufficiënte statistiek is voor $\theta$.

## Oplossing:

$$
\begin{aligned}
L(\theta \mid x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{\{x_i \leq \theta\}} = \frac{1}{\theta^n} I_{\{x_1, \ldots, x_n \leq \theta\}} = \frac{1}{\theta^n} I_{\{max\{x_1, \ldots, x_n\} \leq \theta\}} \\
&= \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}
\end{aligned}
$$

Kies $h(x_1, \ldots, x_n) = 1$ en $g(T(x_1, \ldots, x_n); \theta) = \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}$.

## Voorbeeld 2: Poisson verdeling

Veronderstel $x_1, \ldots, x_n$ onafhankelijke en identiek verdeelde steekproeven van Poi$(\theta)$. Toon aan dat $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ een sufficiënte statistiek is voor $\theta$.

## Oplossing:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!} = \frac{e^{-n\theta} \cdot \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = \frac{1}{\prod_{i=1}^n x_i!} \cdot e^{-n\theta} \cdot \theta^{T(x_1,\ldots,x_n)}$$

Kies $h(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$ en $g(T(x_1, \ldots, x_n); \theta) = e^{-n\theta} \cdot \theta^{T(x_1,\ldots,x_n)}$.

OEFENING:

Veronderstel $x_1, \ldots, x_n$ onafhankelijke en identiek verdeelde steekproeven van Bern$(\theta)$. Toon aan, met gebruik van de Neyman-Fisher Factorisatiestelling, dat $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ een sufficiënte is voor $\theta$.

## STELLING:

Stel dat $X_1, \ldots, X_n$ onafhankelijke en identieke verdeelde steekproeven zijn afkomstig uit de exponentiële familie:

$$f(x \mid \lambda) = h(x) \cdot \exp[\eta(\lambda) \cdot T(x) - A(\lambda)]$$

Dan is $T(X) = \sum_{i=1}^{n} T(X_i)$ een sufficiënte statistiek voor $\lambda$.

Bewijs:

De likelihood kunnen we noteren als:

$$
\begin{aligned}
L(\lambda \mid x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i \mid \lambda) = \prod_{i=1}^{n} h(x_i) \cdot \mathsf{exp}[\eta(\lambda) \cdot T(x_i) - A(\lambda)] \\
&= \left[\prod_{i=1}^{n} h(x_i)\right] \cdot \mathsf{exp}\left[\eta(\lambda) \sum_{i=1}^{n} T(x_i) - nA(\lambda)\right]
\end{aligned}
$$

Noem:

- $T(X) = \sum_{i=1}^{n} T(X_i)$
- $h(x) = \prod_{i=1}^{n} h(x_i)$
- $g(T(x_1, \ldots, x_n); \lambda) = \exp\left[\eta(\lambda) \sum_{i=1}^{n} T(x_i) - nA(\lambda)\right]$

Dan:

$$f_X(X_1, \ldots, X_n; \lambda) = h(x) \cdot g(T(x), \lambda)$$

$\longrightarrow$ Dit is exact de vorm vereist door de **Neyman-Fisher Factorisatiestelling**!

**Opmerkingen:**

- In vele gevallen is $T(x_i) = x_i$, zoals bij de Poisson en normale verdeling (slides 13 & 15, respectievelijk)

- Dit kan perfect uitgebreid worden naar de **meerdimensionale exponentiële familie**:

$$f(x \mid \boldsymbol{\theta}) = h(x) \cdot \exp[\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{T}(x) - A(\boldsymbol{\theta})]$$

## Voorbeeld: Normaal verdeling met onbekende $\mu$ en $\sigma^2$

$$f(x \mid \mu, \sigma^2) = \text{exp}\left[\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2} \cdot \log(2\pi\sigma^2)\right]$$

Hier:

- $\boldsymbol{T}(x) = \binom{x}{x^2}$
- $\boldsymbol{\eta}(\mu, \sigma^2) = \binom{\mu/\sigma^2}{-1/(2\sigma^2)}$
- $A(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \cdot \log(2\pi\sigma^2)$
- $h(x) = 1$

Voor een steekproef $X_1, \ldots, X_n$, $\boldsymbol{T}(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is een sufficiënte statistiek voor $(\mu,\ \sigma^2)$!

## Part 3:

## Belang van sufficiëntie binnen likelihood ratio testen

# 1. Introductie

- Naast punt estimatie wensen statistici graag uitspraken te maken over $\theta$
- Statistici doen dit a.d.h.v. statistische testen
- Een vaak gebruikte test hiervoor is de **likelihood ratio test**
- Terwijl sufficiëntie vaak gebruikt wordt om een goede schatter te definiëren, heeft het ook voordelen binnen deze test

- In de volgende slides leg ik uit:
  - Wat een likelihood ratio test is
  - Hoe die eruit ziet met sufficiënte statistiek
  - Hoe dat moeilijker wordt zonder die statistiek
  - Een uitgewerkt voorbeeld met de Poisson verdeling

# 2. Likelihood ratio test

- Veronderstel dat we de volgende hypothese wilt testen:

$$\boxed{H_0 : \theta = \theta_0 \qquad H_a : \theta \neq \theta_0}$$

- De **likelihood ratio test** is gedefinieerd als volgt:

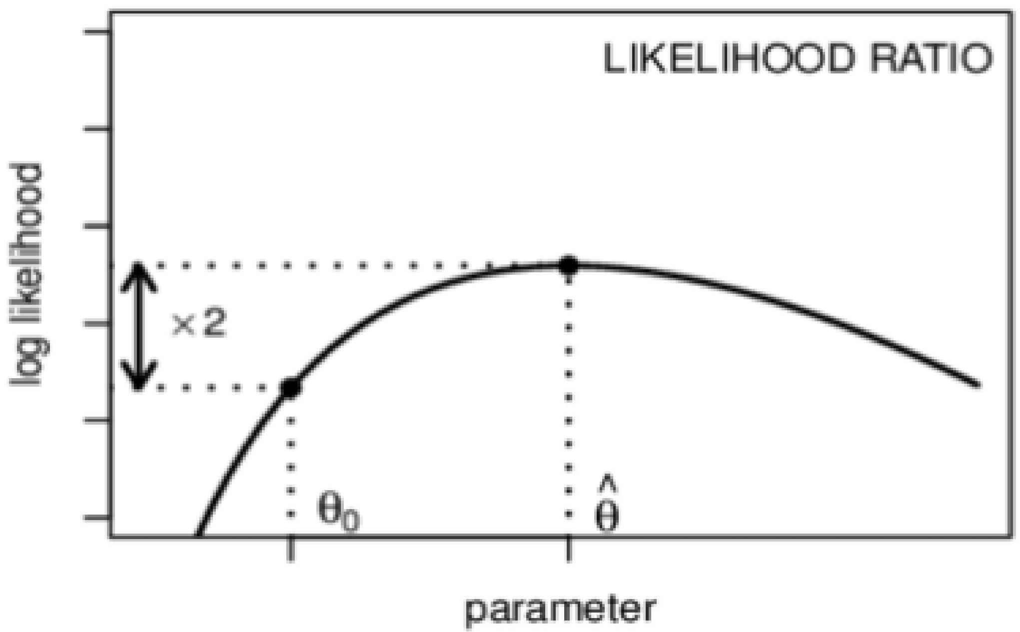
$$\Lambda = \frac{\max_{\theta \in \Theta_0} L(\theta; x)}{\max_{\theta \in \Theta} L(\theta; x)}$$

  - $\Theta$: De totale parameter ruimte van $\theta$
  - $\Theta_0$: Deelverzameling van $\Theta$ onder de nulhypothese

- In de praktijk gebruikt men vaak de log-likelihood ratio:

$$-2\log\Lambda$$

die (onder $H_0$) asymptotisch een $\chi^2_{\dim(\Theta)-\dim(\Theta_0)}$-verdeling.

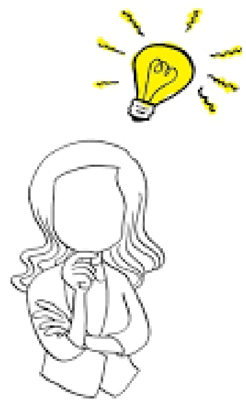


- **Vraag:** Waarom is sufficiëntie belangrijk hierbij?

- **Antwoord:**

  Als u weet dat een statistiek $T$ sufficiënt voor $\theta$, dan is het voldoende om de likelihood ratio te baseren op een samenvattende statistiek $T$ i.p.v. de volledige steekproef $X_1, \ldots, X_n$.

  **Waarom?** Statistiek $T$ bevat alle informatie over $\theta$. Dit maakt de berekening van $\Lambda$ eenvoudiger! Als je de likelihood $L(\theta; x)$ zou proberen uit te rekenen op basis van de volledige steekproef, vereist dit het optimaliseren over alle $x_i$'s mee!



## Voorbeeld: Poisson verdeling

- Stel dat $X_1, ..., X_n$ onafhankelijke en identieke verdeelde steekproeven zijn afkomstig uit Poi$(\lambda)$

- Hypothese test:

$$\boxed{H_0 : \lambda = \lambda_0 \qquad H_a : \lambda \neq \lambda_0}$$

- **Likelihood:**

$$L(\lambda \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda \cdot e^{-\lambda}}{x_i!} = \lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda} \cdot \prod_{i=1}^{n} \frac{1}{x_i!}$$

- **Sufficiënte statistiek:** $T = \sum_{i=1}^{n} X_i$
- **Likelihood ratio test:** We herschrijven nu de likelihood in termen van $T$
  - Onder $H_0$: $L(\lambda_0; x) \propto \lambda_0^T \cdot e^{-n\lambda_0}$
  - Onder $H_a$: maximum likelihood bij $\hat{\lambda} = \bar{X} = T/n$, dus:

$$L(\hat{\lambda} \mid x) \propto \left(\frac{T}{n}\right)^T \cdot e^T$$

$$\to \Lambda = \frac{\lambda_0^T \cdot e^{-n\lambda_0}}{(T/n)^T \cdot e^T} = \left(\frac{\lambda_0}{T/n}\right)^T \cdot e^{T-n\lambda_0}$$

**Let op:** De hele test hangt **enkel af van T**!

# 3. Conclusie

Sufficiëntie laat toe om een test op te bouwen op basis van een samenvattende statistiek $T$ die alle informatie bevat rond een ongekende parameter $\theta$.

Dit vereenvoudigt:

- De berekening van de likelihood ratio $\Lambda$
- De implementatie in software pakketten
- De simulatie of afleiding van de verdeling onder $H_0$
  - De verdeling van de test statistiek onder $H_0$ kan je afleiden of simuleren op basis van $T$ alleen, i.p.v. de volledige steekproef X.

Dit heeft 2 voordelen:

- **Voordeel 1:** Lagere dimensionale ruimte, dus eenvoudige simulatie!

  **Voorbeeld:**

  Voor $X_1, ..., X_n \sim$ Poi$(\lambda_0)$ volstaat het om de verdeling van $T = \sum_{i=1}^{n} X_i \sim$ Poi$(n\lambda_0)$ te gebruiken.

  I.p.v. honderden vectoren $(X_1, ..., X_{100})$ te genereren, simuleer je enkel 1 getal per steekproef: $T$

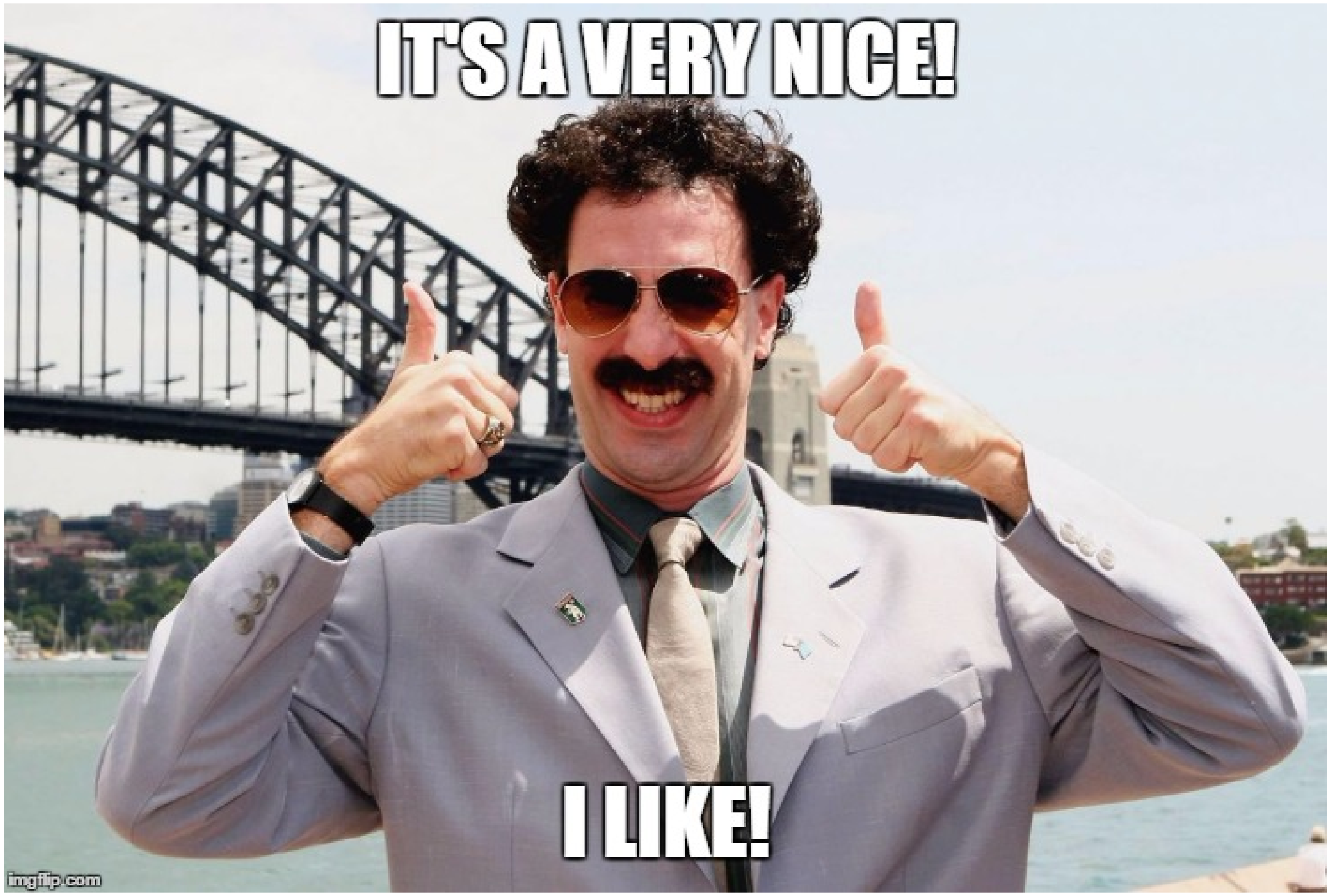
- **Voordeel 2:** Soms exacte verdeling beschikbaar

  Omdat sufficiënte statistieken vaak een bekende verdeling hebben onder $H_0$, kunt u de verdeling van de test statistiek exact afleiden.

  **Voorbeeld:**

  $T = \sum_{i=1}^{n} X_i \sim \text{Poi}(n\lambda_0) \rightarrow$ Exacte verdeling onder $H_0$

  U moet dus de verdeling onder $H_0$ niet af te leiden van een complexere functie van de data, of deze benaderen via resampling of permutaties bij kleinere steekproeven!


IT'S A VERY NICE!
I LIKE!
imgflip.com

# REFERENTIES


- Casella, G., and Berger, R. L. (2002). *Statistical Inference (2nd ed.)*. Chapman and Hall/CRC.
- Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media
- Lehmann, E. L., and Casella, G. (2006). *Theory of Point Estimation & Testing Statistical Hypotheses*. Springer Science & Business Media