Statistik och Dataanalys I

Föreläsning 16 - Sannolikhetsmodeller II

Mattias Villani

Statistiska institutionen
Stockholms universitet

mattiasvillani.com @matvil @matvil mattiasvillani

Översikt

- **Poissonfördelning**
- **Exponentialfördelning**
- **Student-*t***

Poissonfördelning

- **Poissonfördelningen** är en fördelning för **räknedata** (antal).
- Om $X \sim \text{Poisson}(\lambda)$ så

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \text{för } x = 0, 1, 2, ...$$

- $e \approx 2.71$ är Eulers tal.
- Poisson har samma **väntevärde** och **varians**:

$$E(X) = \lambda$$
$$Var(X) = \lambda$$

- Exempel:
  - antal buggar i en mjukvara
  - antal budgivare i en eBay auktion
  - antal besök till läkaren

Poissonfördelning - interaktivt

Poissonfördelningen

λ : 2

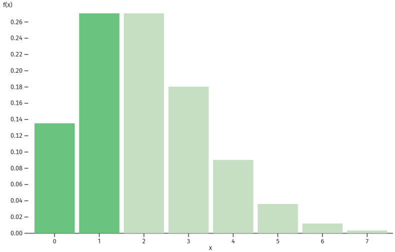
Quantile: 1

If $X \sim \text{Poisson}(2)$ then

$E(X) = \lambda = 2.00$

$Var(X) = \lambda = 2.00$

$P(X \leq 1) = 0.4060$



Mattias Villani Poisson distribution

Observable

Poissonfördelning för antal bud på eBay

- Data från 1000 **eBay-auktioner av samlarmynt**.
- `nBids` är **antalet budgivare** i en given auktion.
- Olika värdefulla och olika reservationspris (lägsta pris).
- Fokus här på de 550 observationer med lägst reservationspris.
- **Modell** för `nBids`: $X_1, \ldots, X_n \overset{\text{ober}}{\sim} \mathrm{Pois}(\lambda)$.

	nBids	PowerSeller	VerifyID	Sealed	Minblem	MajBlem	LargNeg	LogBook	MinBidShare	Sold	low_res_price
1	2	0	0	0	0	0	0	-0.224	-0.209	True	low
2	6	1	0	0	0	0	0	0.607	-0.348	True	low
3	1	1	0	0	0	0	0	0.033	0.442	True	high
4	1	0	0	0	1	0	0	0.376	0.144	True	high
5	4	0	0	0	0	0	1	1.435	-0.41	True	low
6	2	0	0	0	0	0	0	-0.914	0.632	True	high
7	2	0	0	0	1	0	0	-0.248	0.295	True	high
8	2	0	0	0	0	0	0	-0.914	0.632	True	high
9	2	1	0	0	0	0	0	0.511	0.055	True	high
10	6	0	0	1	0	0	0	-0.362	0.025	True	high
11	0	1	0	0	0	0	0	-0.224	0.477	False	high

Wegmann, B. och Villani, M. (2011). Bayesian Inference in Structural Second-Price Common Value Auctions, *Journal of Business and Economic Statistics*

Punktskattning av modellparametrar

- Modell för `nBids`: $X_1, \ldots, X_n \overset{\text{ober}}{\sim} \mathrm{Pois}(\lambda)$.
- Hur väljer vi parametern λ? **Punktskattning**. **Estimat**. $\hat{\lambda}$.
- **Momentmetoden**: Eftersom $\lambda = E(X)$ så är $\hat{\lambda} = \bar{x}$ rimligt.
- **Maximum likelihood**: välj det λ som maximerar sannolikheten för datamaterialet. 😍
- Maximum likelihood-metoden funkar för alla modeller. 😎

Maximum likelihood för Poisson - interaktivt

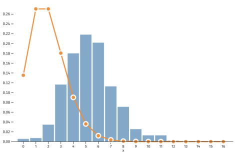
Maximum likelihood estimation - Poissonfördelning

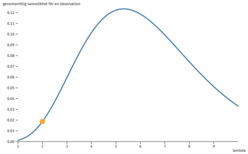
Modell: $X_1, X_2, \ldots, X_n \overset{\text{ober}}{\sim} \text{Pois}(\lambda)$

λ : 2

Visa maximum likelihood anpassning

Medelsannolikhet för observerad data med modellen Pois(λ = 2) är 0.01872





Mattias Villani Maximum likelihood - Poissonmodellen

Observable

Exponentialfördelning

- Om $X \sim \mathrm{Expon}(\lambda)$ så är **täthetsfunktionen**

$$f(x) = \lambda e^{-\lambda x}, \text{ för } x > 0$$

- **Väntevärde** och **varians**

$$E(X) = \frac{1}{\lambda} \text{ och } Var(X) = \frac{1}{\lambda^2}$$

- **Exponentialfördelning** vanlig modell för **väntetider**.
  - Tid mellan samtal till stödlinje.
  - Tid mellan mjukvarureleaser.
- Exponential och Poisson-fördelningen hänger ihop:
  - Om **antalet samtal** till stödlinje per timme är $\mathrm{Poisson}(\lambda = 6)$ så förväntar vi oss $\lambda = 6$ st samtal i timmen.
  - Då är **tiden mellan samtal** $\mathrm{Expon}(\lambda = 6)$ och vi förväntar oss $1/\lambda = 1/6$ timmar (10 minuter) mellan samtal.

Exponentialfördelning

Exponentialfördelningen

λ : 1.01

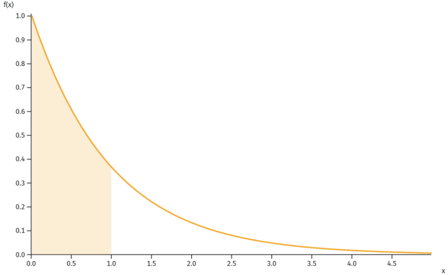
Kvantil: 1

Om $X \sim \text{Expon}(1.01)$ så gäller att

$$E(X) = \frac{1}{\lambda} = 0.990$$

$$Var(X) = \frac{1}{\lambda^2} = 0.980$$

$$P(X \leq 1) = 0.6358$$



Mattias Villani Exponentialfördelningen **Observable**

Exponentialfördelning i R

- $X \sim \mathrm{Expon}(\lambda = 3)$. Parametern λ kallas `rate` i R.

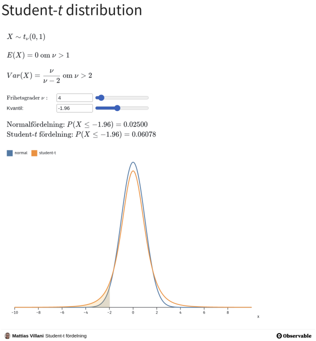
Beräkning	R kommando	Kommentar
$f(0.5)$	`dexp(x = 0.5, rate = 3)`	$f(x)$ vid $x = 0.5$
$P(X \leq 0.5)$	`pexp(q = 0.5, rate = 3)`	
Kvantil	`qexp(p = 0.5, rate = 3)`	Medianen
10 slumptal	`rexp(n = 10, rate = 3)`	

- **Täthetsfunktion** heter **density function** på engelska. Därav namnet `dexp`.

- Se programkoden exponential.R på kurssidan.

Student-t fördelning (standard)

- $X \sim t_\nu(0,1)$ är en **student-*t*** fördelning med ν **frihetsgrader**.
- **Kontinuerliga symmetriska** variabler över $(-\infty, \infty)$.
- Student-t har mer sannolikhet på **extrema utfall**.
- **Student-*t*** fördelning blir alltmer lik normal när ν ökar.



Varför student-t är viktig för inferens

- $X_1, X_2, \ldots, X_n$ är oberoende data from $N(\mu, \sigma^2)$.
- Stickprovmedelvärdet

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

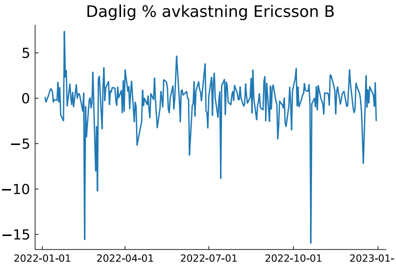
- Inferens: fördelningen för det **standardiserade medelvärdet**

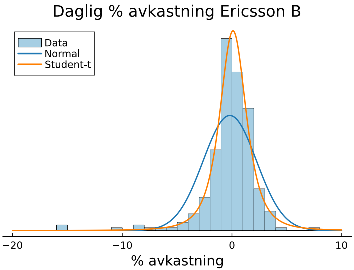
$$\frac{\bar{X} - \mu}{SD(\bar{X})}$$

- Om variansen i populationen σ^2 **är känd** så är det **standardiserade medelvärdet normalfördelat**.
- Om variansen i populationen σ^2 **är okänd**, och måste skattas med s^2, så är det **standardiserade medelvärdet student-t fördelad** med $\nu = n - 1$ frihetsgrader.

Student-t som modell för aktieavkastning

- Daglig avkastning Ericsson B aktie under hela år 2022. sda1 ericsson
- Finansiella data har ofta extremvärden. **Tunga svansar**.
- Maximum likelihood: $\mu = 0.094$, $\phi = 1.279$ och $\nu = 2.706$.





Allmän Student-t fördelning för datamodellering

Allmän Student-t distribution

$X \sim t_\nu(\mu, \phi^2)$

$E(X) = \mu$ om $\nu > 1$

$Var(X) = \dfrac{\nu}{\nu - 2}\phi^2$ om $\nu > 2$

Läge μ :	2
Skala ϕ :	1.5
Frihetsgrader ν :	4
Kvantil:	0

visa normalfördelning ☐

Student-t fördelning: $P(X \leq 0) = 0.1266$