

Statistik och Dataanalys I

Föreläsning 15 - Sannolikhetsmodeller II

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



[@matvil](https://twitter.com/matvil)



[@matvil](https://mastodon.social/@matvil)



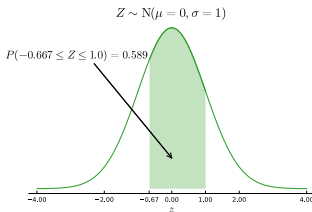
[mattiasvillani](https://github.com/mattiasvillani)

- Likformig fördelning
- Normalfördelning
- Poissonfördelning
- Exponentialfördelning
- Student- t

Kontinuerliga slumpvariabler och täthetsfunktionen

- **Kontinuerlig slumpvariabel** antar alla värden, men $P(X = x) = 0$ för alla x ! 🤖
- **Täthetsfunktion**: $f(x)$.
- Positiv $f(x) > 0$ för alla x .
- Täthetsfunktion ger **inte** sannolikheter. OK om $f(x) > 1$.
- **Täthetsfunktionen** används för att **beräkna sannolikheter**:




$P(a \leq X \leq b) = \text{arean under } f(x) \text{ mellan } a \text{ och } b$



- **SDAIII**: räkna arean under funktion med **integration**.

Likformig fördelning

Likformig fördelning

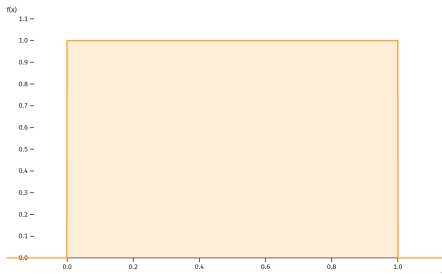
a : 
 b : 
Kvantil: 

Om $X \sim \text{Uniform}(0, 1)$ så gäller att

$$E(X) = \frac{a+b}{2} = 0.500$$

$$\text{Var}(X) = \frac{(b-a)^2}{12} = 0.0833$$

$$P(X \leq 1) = 1.000$$



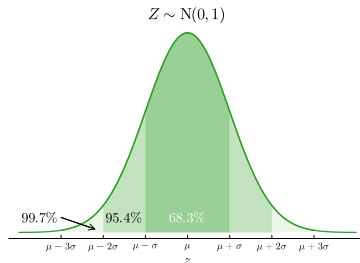
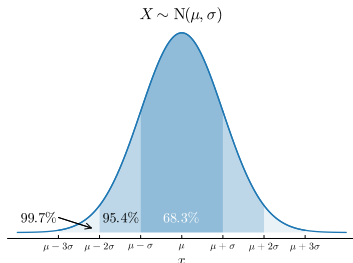
Normalfördelning

■ $X \sim N(\mu, \sigma)$

$$E(X) = \mu$$

$$SD(X) = \sigma$$

■ 68-95-99.7% regeln



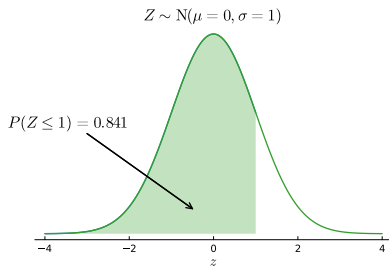
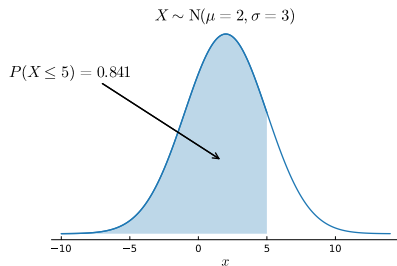
Normalfördelning - standardisering

■ Standardisering

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

■ Sannolikhet via standardisering för $X \sim N(2, 3)$

$$P(X \leq 5) = P(X - 2 \leq 5 - 2) = P\left(\frac{X - 2}{3} \leq \frac{5 - 2}{3}\right) = P(Z \leq 1)$$



Normalfördelning - Z-tabell

Normalfördelning

Tabellen ger sannolikheten $\Phi(z) = P(Z \leq z)$ för olika z där Z är standardnormal, $Z \sim N(0, 1)$.

Sannolikheter i den vänstra svansen fås genom symmetri: $P(Z \leq -z) = 1 - P(Z \leq z)$.



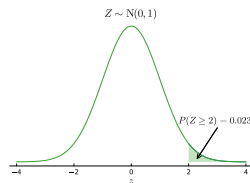
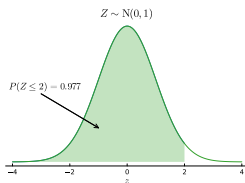
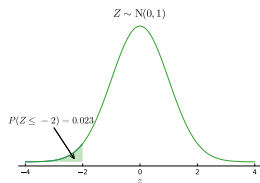
Andra decimalen i z

| | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |

Normalfördelning - symmetri

- **Negativa z-värden** finns inte i Z-tabellen.
- Vi utnyttjar normalfördelningens **symmetri** för negativa z

$$P(Z \leq -2) = 1 - P(Z \leq 2)$$



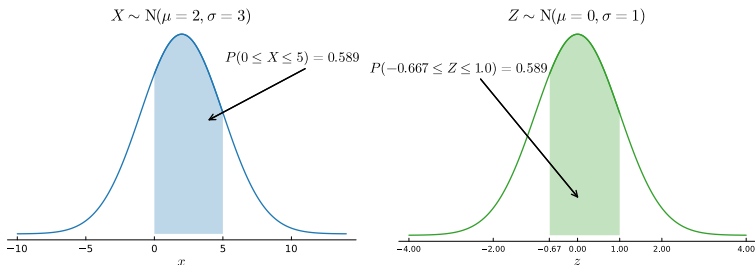
Normalfördelning - intervall via standardisering

■ Sannolikhet via standardisering för $X \sim N(2, 3)$

$$\begin{aligned}P(0 \leq X \leq 5) &= P\left(\frac{0-2}{3} \leq \frac{X-2}{3} \leq \frac{5-2}{3}\right) \\&= P(-0.667 \leq Z \leq 1) \\&= P(Z \leq 1) - P(Z \leq -0.667)\end{aligned}$$

och pga **symmetri**

$$P(Z \leq -0.667) = 1 - P(Z \leq 0.667)$$



Normalfördelningen - interaktivt

Normalfördelningen

μ :

σ :

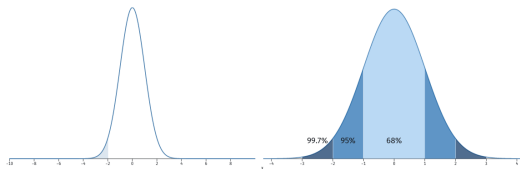
Kvantil:

Om $X \sim N(0, 1)$ så gäller att

$$E(X) = \mu = 0.00$$

$$Var(X) = \sigma^2 = 1.00$$

$$P(X \leq -1.96) = 0.02500$$



Normalfördelning - egenskaper

Linjärkombination av normalfördelad slumpvariabel.

Om $X \sim N(\mu, \sigma)$ och $Y = c + aX$ så gäller

$$Y \sim N(c + a\mu, |a|\sigma)$$

Summa av oberoende normalfördelade slumpvariabler.

Om $X \sim N(\mu_X, \sigma_X)$ och $Y \sim N(\mu_Y, \sigma_Y)$ är oberoende slumpvariabler så är även summan normalfördelad:

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

■ **Fördelningarna** för linjärkombination och summa är **normal**!

■ Summan är fortfarande normal om **X och Y är beroende**.

Poissonfördelning

- **Poissonfördelningen** är en fördelning för **räknedata** (antal):

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

- Om $X \sim \text{Pois}(\lambda)$ så

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

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

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda$$

Poissonfördelning - interaktivt

Poissonfördelningen

λ : 

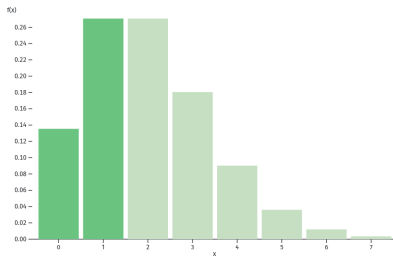
Quantile: 

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

$$E(X) = \lambda = 2.00$$

$$\text{Var}(X) = \lambda = 2.00$$

$$P(X \leq 1) = 0.4060$$



 Mattiias Villani Poisson distribution

 Observable

Poissonfördelning för antal bud på eBay

sda123 ebaycoins

- 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: $X_1, \dots, X_n \overset{\text{ober}}{\sim} \text{Pois}(\lambda)$.

| | 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: $X_1, \dots, X_n \overset{\text{ober}}{\sim} \text{Pois}(\lambda)$.
- 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: $X_1, X_2, \dots, X_n \overset{\text{ober}}{\sim} \text{Pois}(\lambda)$

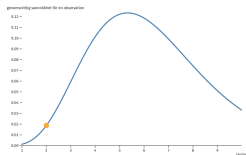
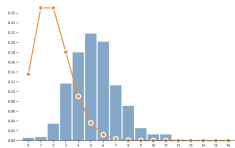
λ :

2

Visa maximum
likelihood
anpassning



Medelsannolikhet för observerad data med modellen $\text{Pois}(\lambda = 2)$ är 0.01872



Mattias Villani Maximum likelihood - Poissonmodellen

Observable

Exponentialfördelning

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

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

- **Väntevärde** och **varians**

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

- **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 $\text{Poisson}(\lambda = 6)$ så förväntar vi oss $\lambda = 6$ st samtal i timmen.
- ▶ Då är **tiden mellan samtal** $\text{Expon}(\lambda = 6)$ och vi förväntar oss $1/\lambda = 1/6$ timmar (10 minuter) mellan samtal.

Exponentialfördelning

Exponentialfördelningen

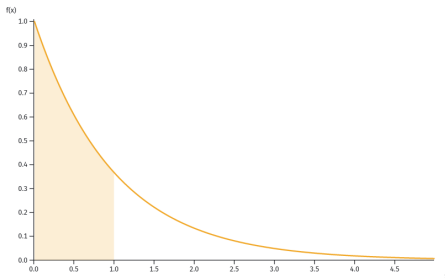
λ : 
Kvantil: 

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

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

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

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



Exponentialfördelning i R

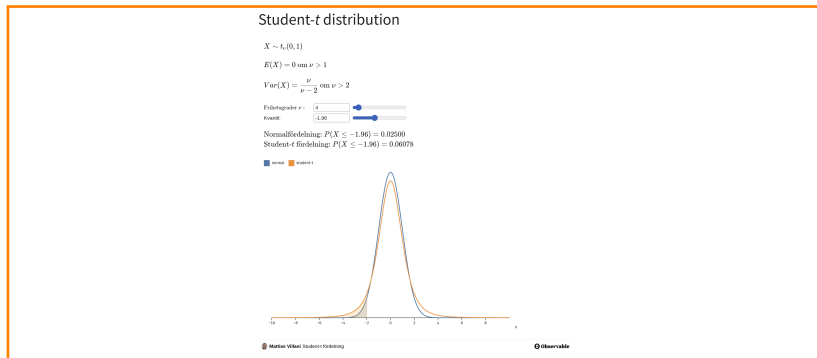
- $X \sim \text{Expon}(\lambda = 3)$. Parametern λ kallas `rate` i R.

| Beräkning | R kommando | Kommentar |
|-----------------|--------------------------------------|----------------------|
| $f(0.5)$ | <code>dexp(x = 0.5, rate = 3)</code> | $f(x)$ vid $x = 0.5$ |
| $P(X \leq 0.5)$ | <code>pexp(q = 0.5, rate = 3)</code> | |
| Kvantil | <code>qexp(p = 0.5, rate = 3)</code> | Medianen |
| 10 slumpstal | <code>rexp(n = 10, rate = 3)</code> | |

- **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 alltmer lik normalfördelning när ν ökar.



Varför student- t är viktig för inferens

- X_1, X_2, \dots, X_n är oberoende data från $N(\mu, \sigma)$.
- Stickprovmedelvärdet

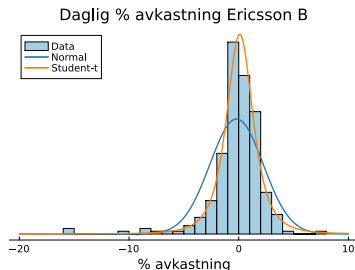
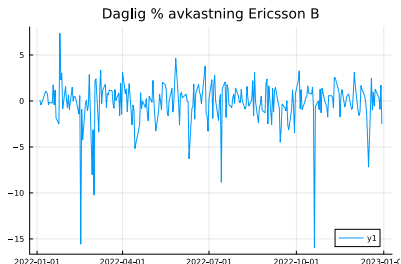
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- 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 fördelat** med $\nu = n - 1$ frihetsgrader.

- Student- t fördelningen kommer visa sig viktig för inferens för väntevärdet μ i en normalpopulation. F17.
- Student- t är en bra modell för data med extremvärden.
- Daglig avkastning Ericsson B aktie under hela år 2022.
- 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

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

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

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

| | | |
|-----------------------|----------------------------------|---|
| Läge μ : | <input type="text" value="2"/> |  |
| Skala ϕ : | <input type="text" value="1.5"/> |  |
| Frihetsgrader ν : | <input type="text" value="4"/> |  |
| Kvantil: | <input type="text" value="0"/> |  |

visa
normalfördelning ☐

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

