

Statistik och Dataanalys I

Föreläsning 15 - Sannolikhetsmodeller I

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



@matvil



mattiasvillani

- Bernoulliförsök
- Geometrisk fördelning
- Binomialfördelning
- Likformig fördelning
- Normalfördelning

Bernoulliförsök

■ Bernoulliförsök

- 1 Bara **två möjliga utfall**: lyckas/misslyckas.
- 2 **Samma sannolikhet** för lyckas, p , i alla försök.
- 3 **Oberoende försök**.

■ Typexempel: **slantsingling**.

- ▶ Lyckas = Kona, Misslyckas = Klava.
- ▶ Sannolikhet $p = 0.5$ för schysst mynt.
- ▶ Utfall på en singling beror inte på andra singlar.

■ Lyckas/Misslyckas är bara en benämning.

■ Död/Levande. Hel/Trasig. Spam/Ham.

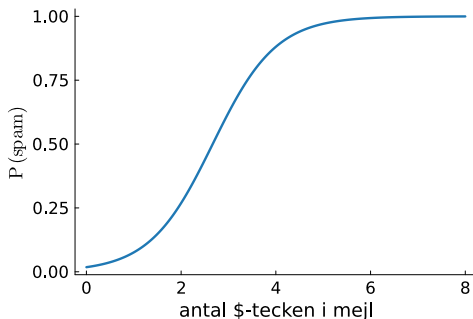


Utan återläggning \Rightarrow inte samma p i olika försök:

- ▶ $P(1:a \text{ kortet } \spadesuit) = \frac{13}{52} = \frac{1}{4}$
- ▶ $P(2:a \text{ kortet } \spadesuit) = \frac{12}{51}$ om 1:a \spadesuit eller $\frac{13}{51}$ om 1:a $\heartsuit, \diamondsuit, \clubsuit$.

Motivation - regression med binära y-variabler

- Bernoulli-fördelning med **samma sannolikhet** p .
- Spamdata: lära oss om $p = P(\text{spam})$ från data. $\hat{p} = 0.9$. 🙄
- **Spam-filter**: ska datorn skicka **just detta mejl** till Spam?
- SDAll: **Logistisk regression** där spam sannolikheten p **beror på förklarande variabler**, som i regression. 🤖

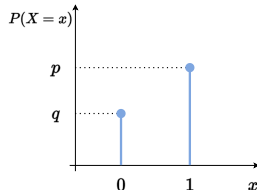


Bernoullifördelning

- Två möjliga utfall: lyckad/misslyckad. **Binär variabel**.
- Vi kan koda **lyckat = 1**, **misslyckat = 0**.

$$X = \begin{cases} 1 & \text{om Bernoulli-försök lyckat} \\ 0 & \text{om Bernoulli-försök misslyckat} \end{cases}$$

$$P(X = x) = \begin{cases} p & \text{för } x = 1 \\ q = 1 - p & \text{för } x = 0 \end{cases}$$



■ Väntevärde och Varians

$$\begin{aligned} E(X) &= \mu = \sum_{\text{alla } x} x \cdot P(x) = 0 \cdot P(X=0) + 1 \cdot P(X=1) \\ &= 0 \cdot q + 1 \cdot p = p \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \sum (x - \mu)^2 \cdot P(x) = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p \\ &= p^2 q + q^2 \cdot p = pq \underbrace{(p + q)}_1 = pq \end{aligned}$$

Geometrisk fördelning

- Email: **spam** eller **ham** (icke-spam).
 - ▶ $P(\text{spam}) = p = 0.9$
 - ▶ $P(\text{ham}) = q = 1 - p = 0.1$
- Hur många mejl måste du öppna tills du får ditt första ham?

$$P(\text{första ham på fjärde mejlet}) = \overbrace{0.9 \cdot 0.9 \cdot 0.9}^{\text{gänger pga oberoende}} \cdot \underbrace{0.1}_{\text{ham}} = 0.9^3 \cdot 0.1 = 0.0729$$

- Vad är sannolikheten för x st mejl tills första ham?

$$P(\text{första ham på } x\text{:te mejlet}) = 0.9^{x-1} \cdot 0.1$$

- **Geometrisk slumpvariabel** från Bernoulliförsök

X = antal försök **tills första lyckade** inträffar

- **Geometrisk fördelning**

$$P(X = x) = q^{x-1} p, \quad \text{för } x = 1, 2, 3, \dots$$



X inkluderar försöket där du först lyckas.

Wikipedia kallar detta för **för-första-gången-fördelning**.

Geometrisk fördelning

Geometrisk fördelning

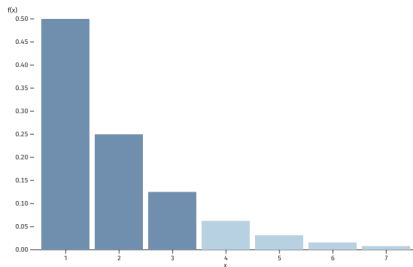
p : 
Kvantil: 

Om $X \sim \text{Geo}(0.5)$ så gäller att

$$E(X) = \frac{1}{p} = 2.00$$

$$\text{Var}(X) = \frac{1-p}{p^2} = 0.250$$

$$P(X \leq 3) = 0.8750$$



Geometrisk fördelning i R

- $X \sim \text{Geom}(p = 0.4)$. Sannolikheten p kallas `prob` i R.

Beräkning	R kommando
$P(X = 2)$	<code>dgeom(x = 2, prob = 0.4)</code>
$P(X \leq 2)$	<code>pgeom(q = 2, prob = 0.4)</code>
Kvantil	<code>qgeom(p = 0.5, prob = 0.4)</code>
10 slumpstal	<code>rgeom(n = 10, prob = 0.4)</code>

⚠ R använder Wikipedias definition av geometrisk fördelning. X räknar **antalet misslyckade försök innan** första lyckade. Fix:

```
y = rgeom(n = 100, prob = 0.5) # y is number of trials BEFORE first success
x = y + 1                      # x is number of trials INCLUDING first success
```

- Se programkoden [geometric.R](#) på kurssidan.

Binomialfördelning

■ Geometrisk fördelning:

- ▶ Hur många Bernoulli-försök tills första lyckade?
- ▶ Antal försök är slumpmässigt.

■ Binomialfördelning:

- ▶ Hur många lyckade i n Bernoulli-försök med sannolikhet p .
- ▶ Antal försök n är förbestämt och fixerat.
- ▶ Antal lyckade är slumpmässigt.

■ Vi skriver $X \sim \text{Bin}(n, p)$ och säger:

■ “ X är binomialfördelad med parametrar n och p .”

■ Binomial: summan av n oberoende Bernoullivariabler

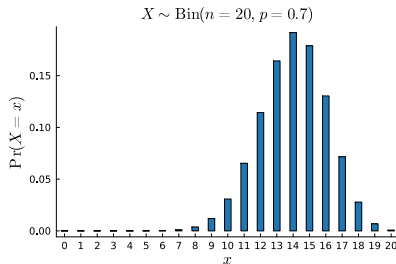
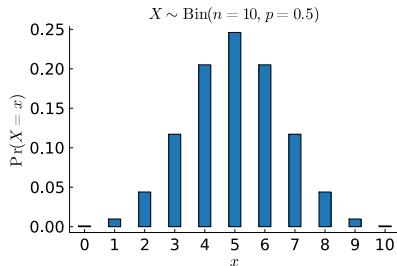
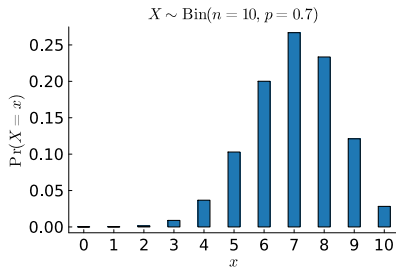
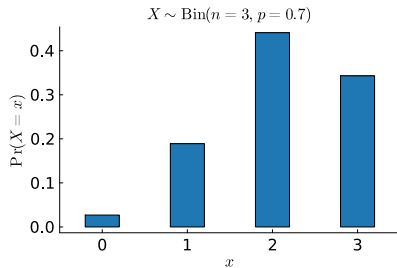
$$X = X_1 + X_2 + \dots + X_n$$

■ Exempel: $n = 3$ försök med resultat:

$X_1 = 1$ (Krona första), $X_2 = 1$ (Krona andra) och $X_3 = 0$ (Klave tredje).

$$X = 1 + 1 + 0 = 2 \text{ st lyckade (Krona).}$$

Binomialfördelning



Binomialfördelning - väntevärde

- Väntevärde i en binomialfördelning? 🤪

$$E(X) = \sum_{x=0}^n x \cdot P(x)$$

Väntevärde - summa av slumpvariabler.

$$E(X_1 + X_2 + \dots, X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

- Väntevärde för varje Bernoulli-variabel: $E(X_i) = p$.

- **Väntevärde för $X \sim \text{Bin}(n, p)$**

$$E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = \underbrace{p + p + \dots + p}_{n \text{ st}} = np$$

Binomialfördelning - varians

- Varians i en binomialfördelning? 🤔🤔🤔

$$E(X) = \sum_{x=0}^n (x - \mu)^2 \cdot P(x)$$

Varians - summa av oberoende slumpvariabler.


$$V(X_1 + X_2 + \dots, X_n) = Var(X_1) + Var(X_2) + \dots + Var(X_n)$$


- Bernoulliförsök är oberoende. ✓
- Varians för varje Bernoulli-variabel: $Var(X_i) = pq$.
- **Varians för $X \sim \text{Bin}(n, p)$**


$$Var(X) = Var(X_1) + \dots + Var(X_n) = \underbrace{pq + pq + \dots + pq}_{n \text{ st}} = npq$$

Binomialfördelning - interaktivt

Binomialfördelningen

n : 

p : 

Kvantil: 

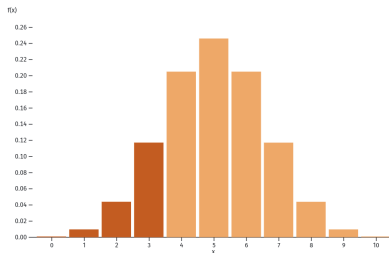
Visa
normalapproximation ☐

Om $X \sim \text{Binom}(10, 0.5)$ så gäller att

$$E(X) = np = 5.00$$

$$\text{Var}(X) = np(1-p) = 2.50$$

$$\text{Exakt: } P(X \leq 3) = 0.1719$$



Binomialfördelningens sannolikheter

- Om $X \sim \text{Bin}(n, p)$ - vad är egentligen $P(X = x)$?
- Sannolikheten att få $\{1, 1, 0\}$ i $n = 3$ försök?

$$p \cdot p \cdot q = p^2 q^1$$

- Det finns dock **flera sätt att få $X = 2$** i $n = 3$ försök:

1:a försök	2:a försök	3:e försök	X	$P(X = x)$
1	1	0	2	$p^2 q$
1	0	1	2	$p^2 q$
0	1	1	2	$p^2 q$

- Eftersom dessa tre olika sätt att få $X = 2$ är **disjunkta**:

$$P(X = 2) = 3 \cdot p^2 q$$

- På samma sätt

$$P(X = 0) = P(\{0, 0, 0\}) = 1 \cdot q^3$$

$$P(X = 1) = P(\{1, 0, 0\}, \{0, 1, 0\}, \{0, 0, 1\}) = 3 \cdot p q^2$$

$$P(X = 2) = P(\{1, 1, 0\}, \{1, 0, 1\}, \{0, 1, 1\}) = 3 \cdot p^2 q$$

$$P(X = 3) = P(\{1, 1, 1\}) = 1 \cdot p^3$$

Binomialfördelningens sannolikheter

- **Sannolikhetsfördelning** $X \sim \text{Bin}(3, p)$

x	0	1	2	3
$P(x)$	q^3	$3 \cdot pq^2$	$3 \cdot p^2q$	p^3

- Kolla att summan av alla sannolikheter är ett:

$$q^3 + 3 \cdot pq^2 + 3 \cdot p^2q + p^3 = (p + q)^3 = 1^3 = 1$$

- Allmänna fallet $X \sim \text{Bin}(n, p)$

$$P(X = x) = {}_nC_x \cdot p^x q^{n-x}$$

- ${}_nC_x$ är antalet sätt ordna x st 1:or bland n observationer.

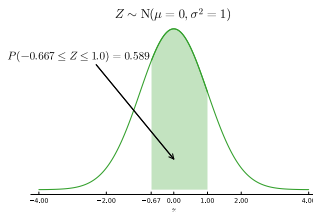
Kombinationer och permutationer

Hur många sätt att välja k element bland n element?		
	med återläggning	utan återläggning
med ordning	n^k	${}_nP_k = \frac{n!}{(n-k)!}$
utan ordning	ej på kurs	${}_nC_k = \frac{n!}{(n-k)!k!}$

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

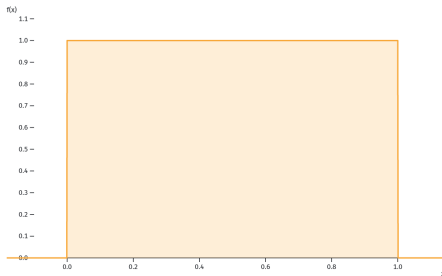
a : 
 b : 
Kvantil: 

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

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

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

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



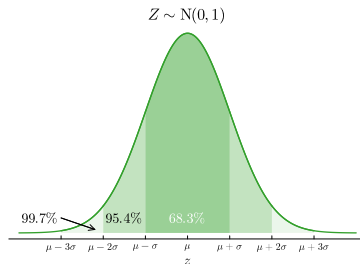
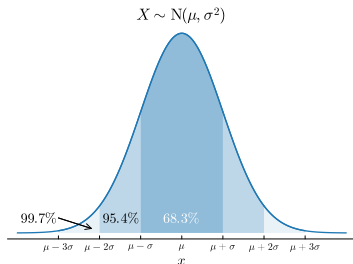
Normalfördelning

■ $X \sim N(\mu, \sigma^2)$

$$E(X) = \mu$$

$$\text{Var}(X) = \sigma^2$$

■ 68-95-99.7% regeln



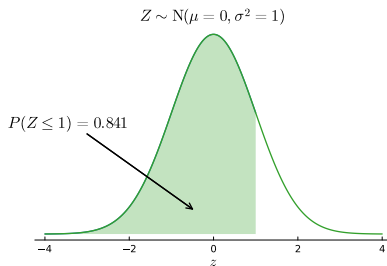
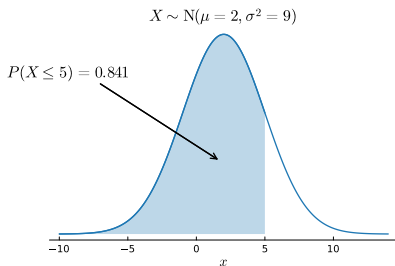
Normalfördelning - standardisering

■ Standardisering

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

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

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

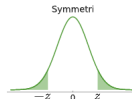


Normalfördelning - Z-tabell

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: $P(Z \leq -z) = 1 - P(Z \leq z)$.



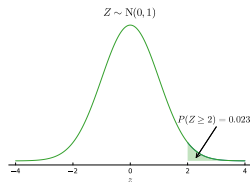
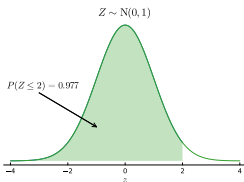
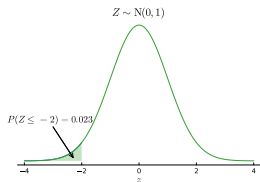
Andra decimalen i z

	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

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



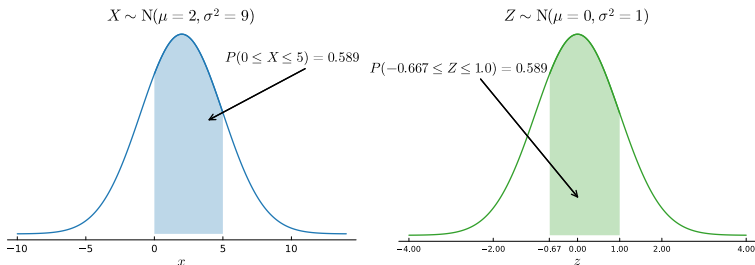
Normalfördelning - intervall via standardisering

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

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


och pga **symmetri**


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




Normalfördelningen - interaktivt

Normalfördelningen

μ : 

σ : 

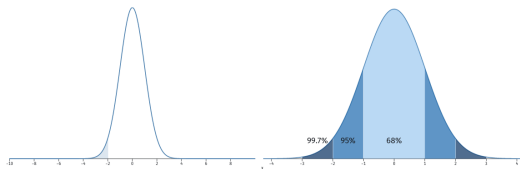
Kvantil: 

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 $X \sim N(\mu, \sigma^2)$ och $Y = c + aX$ så gäller

$$Y \sim N(c + a\mu, a^2\sigma^2)$$

Summa av oberoende normalfördelade slumpvariabler.

Om $X \sim N(\mu_X, \sigma_X^2)$ och $Y \sim N(\mu_Y, \sigma_Y^2)$ är oberoende slumpvariabler så är även summan normalfördelad:

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

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

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

Approximera binomialfördelning med normal

- Om $X \sim \text{Bin}(n, p)$ så

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

och

$$\text{Var}(X) = \sigma^2 = npq$$

- **Normalapproximation** av binomialfördelning

$$X \overset{\text{approx}}{\sim} N(np, npq)$$


- Approximationen är tillräckligt bra om

$$np \geq 10 \text{ och } nq \geq 10$$


- Man kan också göra en **kontinuitetskorrektur** som korrigerar för att vi approximerar en diskret fördelning (binomial) med en kontinuerlig (normal), se SDM-boken kapitel 15.5.

Normalapproximation av binomial - interaktivt

Binomialfördelningen

n : 

p : 

Kvantil: 

Visa
normalapproximation ☒

Om $X \sim \text{Binom}(10, 0.5)$ så gäller att

$$E(X) = np = 5.00$$

$$\text{Var}(X) = np(1-p) = 2.50$$

Exakt:

$$P(X \leq 3) = 0.1719$$

Normal approx:

$$P(X \leq 3) = 0.1030$$

Normal approx med kontinuitetskorrektion:

$$P(X \leq 3) = 0.1714$$

