

Statistik och Dataanalys I

Föreläsning 15 - Diskreta sannolikhetsmodeller

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



@matvil



mattiasvillani

- Bernoulliförsök
- Geometrisk fördelning
- Binomialfördelning
- Poissonfördelning

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 = Klave.
- ▶ 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:

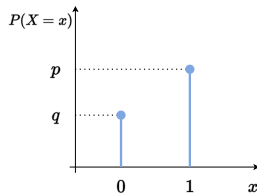
- ▶ $P(1:a \text{ kortet } \spadesuit) = \frac{13}{52}$
- ▶ $P(2:a \text{ kortet } \spadesuit) = \frac{12}{51}$ om 1:a \spadesuit eller $\frac{13}{51}$ om 1:a $\heartsuit, \diamondsuit, \clubsuit$.

Bernoullifördelning

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

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

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



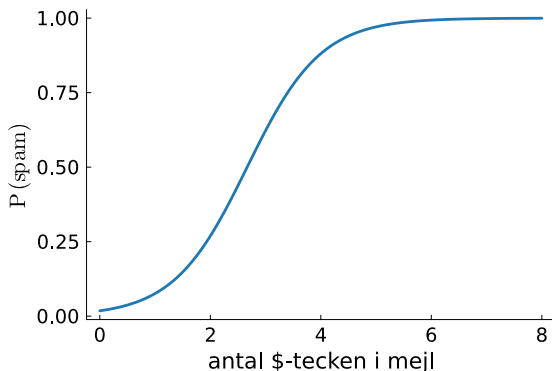
■ Väntevärde och Varians

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

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

Motivation - regression med binära y-variabler

- Bernoulli-fördelning med **samma sannolikhet** p .
- Spamdata: kan lära oss om p 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. 🤖



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

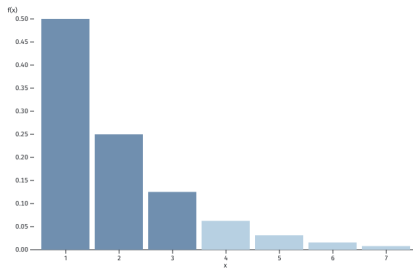
p : 
Kvantil: 

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 = 0.3, 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 .”

■ Binomialvariabeln X är summan av n Bernoullivariabler

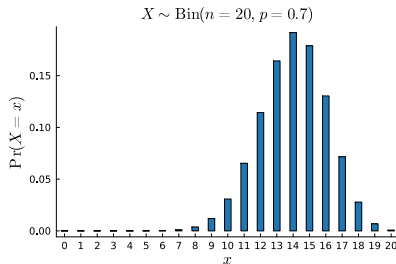
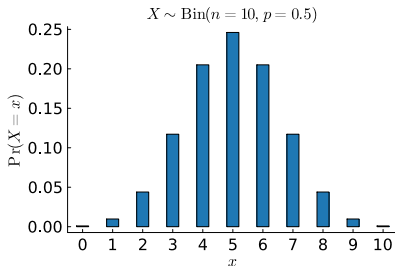
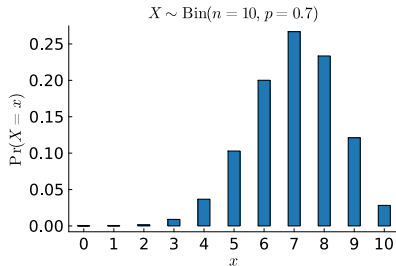
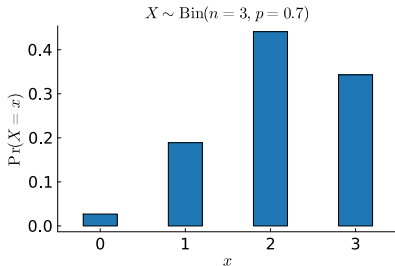
$$X = X_1 + X_2 + \dots + X_n$$

■ 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: 

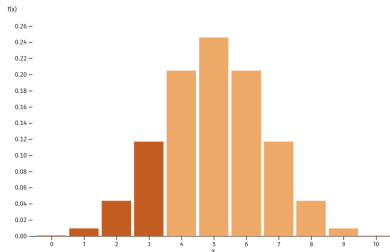
Visa
normalapproximation ☐

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	${}_nC_k = \frac{n!}{(n-k)!k!}$

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 **kontinuitetskorrektion** 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:

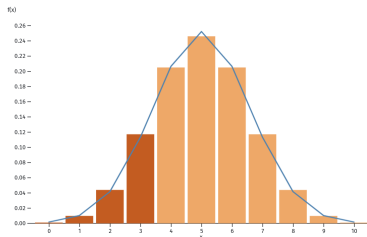
$$P(X \leq 3) = 0.1719$$

Normal approx:

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

Normal approx med kontinuitetskorrektion:

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



Poissonfördelning

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

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

- Poisson har samma **väntevärde** och **varians**:

$$E(X) = \lambda$$

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

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

Poissonfördelning - interaktivt

Poissonfördelningen

λ : 

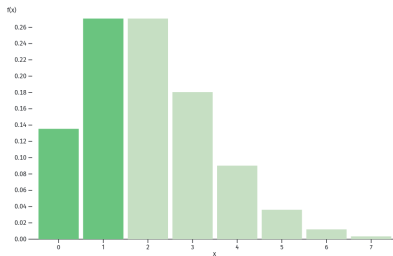
Quantile: 

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

- 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 with Bertil Wegmann, [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 $E(X) = \lambda$ så är $\hat{\lambda} = \bar{x}$ rimligt.
- **Maximum likelihood**: välj det λ som maximerar sannolikheten för datamaterialet. 🥰
- Maximum likelihood-metoden funkar för alla modeller. 😎
- **Minsta-kvadrat-metoden** för regression:
Regressionslinjen $\hat{y} = b_0 + b_1 \cdot x$ är en skattning av populationens regressionslinje: $\beta_0 + \beta_1 \cdot x$. Mer om det i F22.
- För normalfördelade regressionsdata (F22) är b_0 och b_1 faktiskt också maximum likelihood-skattningar!