

# Statistik och Dataanalys I

## Föreläsning 14 - Sannolikhetsmodeller I

**Oskar Gustafsson**

Statistiska institutionen  
Stockholms universitet

- Bernoulliförsök
- Geometrisk fördelning
- Binomialfördelning

# 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.

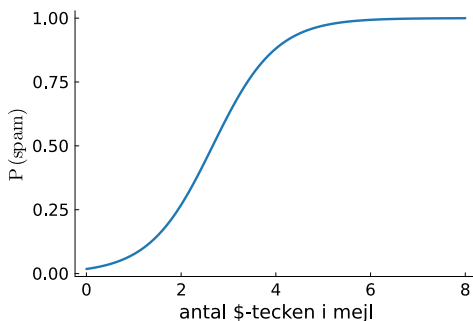
# Bernoulliförsök forts.

- Lyckas/Misslyckas är bara en benämning utan någon mänsklig värdering.
  - Samma med positivt/negativt resultat.
  - 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  $\heartsuit, \diamondsuit, \clubsuit$ .
- ▶ Vi bryter emot både regel 2 och 3 på föregående sida.

# 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. 🤔

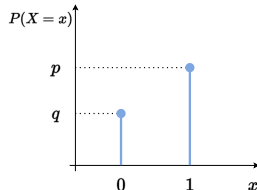


# Bernoullifördelning

- 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

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

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

## Exempel - Basketspelarens straffkast

- Situation: En basketspelare kastar ett straffkast.

- ▶ Lyckas = Träff. Misslyckas = Miss.
- ▶ Sannolikheten att spelaren sätter sit kast är  $p = 0.8$ .
- ▶ Sannolikheten för att misslyckas är  $q = 1 - p = 0.2$ .

- Slumpvariabeln  $X$ :

$$X = \begin{cases} 1 & \text{om spelaren träffar, med } P(X = 1) = p = 0.8 \\ 0 & \text{om spelaren missar, med } P(X = 0) = q = 0.2 \end{cases}$$

- Väntevärde  $E(X)$  :

- ▶  $E(X) = \sum_{\text{alla } x} x \cdot P(x) = (1 \cdot p) + (0 \cdot q)$
- ▶  $E(X) = (1 \cdot 0.8) + (0 \cdot 0.2) = 0.8$
- ▶ Tolkning: I genomsnitt 0.8 poäng per kast.

- Varians  $\text{Var}(X)$ :

- ▶  $\text{Var}(X) = p \cdot q$
- ▶  $\text{Var}(X) = 0.8 \cdot 0.2 = 0.16$
- ▶ Tolkning: Ett mått på osäkerheten i varje kasts utfall.

# Geometrisk fördelning

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

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

- 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 fördelning

- **Geometrisk slumpvariabel** från Bernoulliförsök

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

- **Geometrisk fördelning**

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



$X$  **inkluderar** försöket där du först lyckas.

Wikipedia kallar detta för [för-första-gången-fördelning](#).

# Geometrisk fördelning

## Geometrisk fördelning

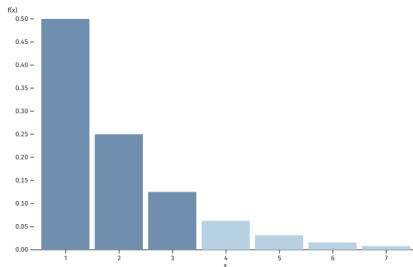
$p$  :    
Kvantil:  

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.

## Basketspelarens straffkast, forts.

- Vi har samma basketspelare, men nu definierar vi ett "lyckat" försök som en miss.
  - ▶ Sannolikheten för att "lyckas" är nu  $p = P(\text{Miss}) = 0.2$ .
  - ▶ Sannolikheten för att "misslyckas" (dvs. att träffa) är  $q = P(\text{Träff}) = 0.8$
- $Y$  är antalet kast som krävs för att få den första **missen**.
- $E(Y) = 1/p = 1/0.2 = 5$
- **Tolkning:** Vi förväntar oss att det i genomsnitt krävs  $Y = 5$  kast innan spelaren missar för första gången.
- $\text{Var}(Y) = q/p^2 = 0.8/(0.2)^2 = 0.8/0.04 = 20$
- **Tolkning:** Variansen är nu mycket högre nu! Det betyder att det är mycket större osäkerhet kring när den första missen kommer.

# 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 att: “ $X$ är binomialfördelad med parametrar $n$ och $p$ .”

## ■ Binomial: summan av $n$ oberoende Bernoullivariabler

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

# Binomialfördelning

- Exempel 1:  $n = 3$  försök med resultat:  
 $X_1 = 1$  (Krona första),  $X_2 = 1$  (Krona andra) och  $X_3 = 0$  (Klave tredje).

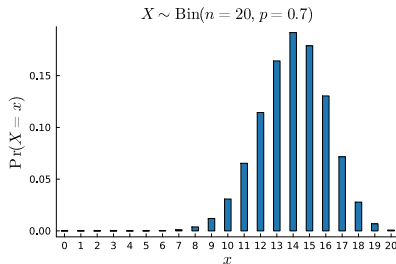
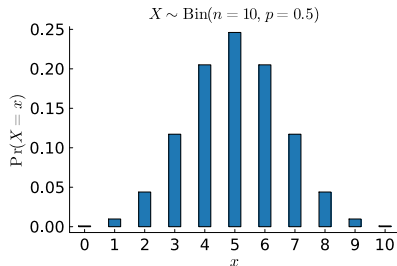
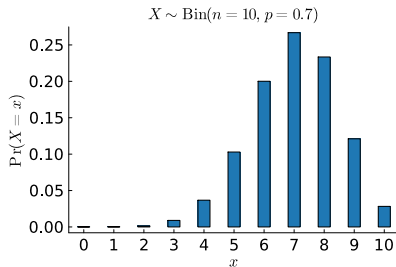
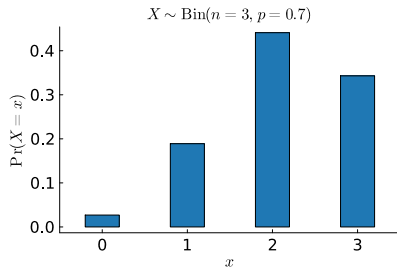
$$X = 1 + 1 + 0 = 2 \text{ st lyckade (Krona).}$$

- Exempel 2: observera  $n = 5$  förbipasserande bilar, våra Bernoulli fördelade variabler,  $Y_1, \dots, Y_5$ , antar värdet 1 om bilen är vit, 0 annars.
- Experimentet resulterade i:  
 $Y_1 = 1$  (vit bil),  $Y_4 = 1$  (vit bil) och övriga bilar inte vita.

$$\text{Slumpvariabel: } X = \sum_{i=1}^n Y_i \text{ antalet vita bilar.}$$

$$\text{Utfall: } x = \sum_{i=1}^n y_i = 2 \text{ vita bilar.}$$

# 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? 🤪🤪🤪

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

**Varians - summa av oberoende slumpvariabler.**


$$\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)$$


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


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

# Binomialfördelning - interaktivt

## Binomialfördelningen

$n$  :  

$p$  :  

Kvantil:  

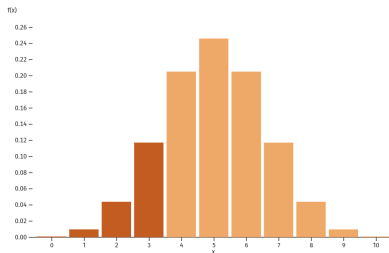
Visa  
normalapproximation ☐

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.

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

# Binomialfördelningen bil exemplet

- Exempel 2: observera  $n = 5$  förbipasserande bilar, våra Bernoulli fördelade variabler,  $Y_1, \dots, Y_5$ , antar värdet 1 om bilen är vit, 0 annars. Antag att sannolikheten att observera en vit bil (lyckat försök) är  $p = 0.3$
- Vad är sannolikheten att vi observerar exakt 2 vita bilar,  $P(X = 2)$ , som i exemplet?

$$\begin{aligned}P(X = 2) &= {}_n C_x p^x q^{(n-x)} = \frac{n!}{(n-x)!x!} (1-p)^{(n-x)} p^x \\&= \frac{5 \times 4 \times 3 \times 2}{(3 \times 2) \times 2} (1-0.3)^{(5-2)} 0.3^2 = \frac{120}{12} 0.7^3 0.3^2 = 0.3087\end{aligned}$$

- Vad är väntevärde och varians för antal bilar?
- Bara att använda formelbladet!  $E(X) = np = 5 \times 0.3 = 1.5$  och  $V(X) = npq = 5 * 0.3 * 0.7 = 1.05$

# Approximera binomialfördelning med normal

- Om  $X \sim \text{Bin}(n, p)$  så

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

och

$$SD(X) = \sigma = \sqrt{npq}$$

- Normalapproximation** av binomialfördelning

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




- 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:

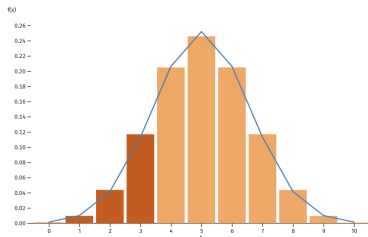
$$P(X \leq 3) = 0.1719$$

Normal approx:

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

Normal approx med kontinuitetskorrektion:

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



Dessa slides skapades för kursen statistik och dataanalys 1 av Mattias Villani HT 2023, och har modifierats av Oscar Oelrich VT 2024, och Oskar Gustafsson för VT 2025.