

SANNOLIKHETSLÄRA OCH STATISTIK

FÖRELÄSNING 3

Mattias Villani

**Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet**



ÖVERSIKT

- ▶ Fördelningsfamiljer för diskreta variabler
- ▶ Bernoulli, binomial, multinomial
- ▶ Geometrisk
- ▶ Poisson

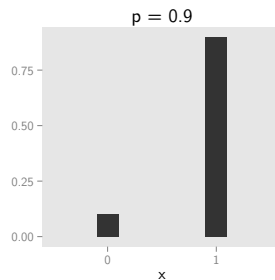
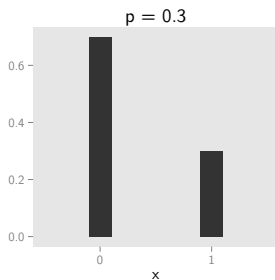
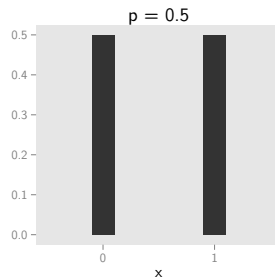
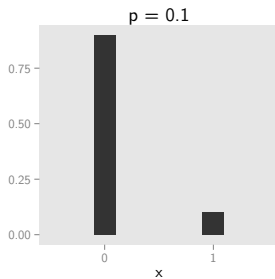
BERNOULLIFÖRDELNINGEN

- En fördelningsfamilj är en mängd olika sannolikhetsfördelningar som indexeras med en eller flera parametrar.

Definition. En **Bernoullivariabel** X kan anta två olika värden, 0 och 1. Om X är **Bernoullifördelad** ($X \sim \text{Bernoulli}(p)$) så gäller att $P(X = 1) = p$ och $P(X = 0) = q = 1 - p$.

- Genom att ändra parametern p får vi en mängd olika sannolikhetsfördelningar på $\{0, 1\}$. Se **ManipDistributions.R**

BERNOULLIFÖRDELNINGEN



BERNOULLIFÖRDELNINGEN

- ▶ Pmf för $X \sim \text{Bernoulli}(p)$

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

- ▶ Om $X \sim \text{Bernoulli}(p)$

$$\mathbb{E}X = p$$

$$\text{Var}(X) = p \cdot q.$$

- ▶ Bernoulliförsök: en sekvens av oberoende Bernoulli variabler, alla med sannolikhet p . Slantsingling.

BINOMIALFÖRDELNINGEN

Definition. Antalet lyckade ($X = 1$) i en sekvens av n Bernoulliförsök med sannolikheten p följer en **binomialfördelning** med parametrar n och p .
 $X \sim \text{Binomial}(n, p)$.

► Pmf

$$P(x) = \binom{n}{x} p^x q^{n-x},$$

för $x = 0, 1, 2, \dots, n$.

- $\binom{n}{x}$ är antalet sekvenser av längd n med exakt x lyckade försök.
Binomialkoefficienten.
- Om t ex $n = 3$ och $x = 2$, så leder alla tre sekvenserna $(0, 1, 1)$, $(1, 0, 1)$ och $(1, 1, 0)$ till utfallet $x = 2$.
- Sekvensen $(0, 1, 1)$ har sannolikheten $q \cdot p \cdot p = p^2 q$.
- Sekvensen $(1, 0, 1)$ har sannolikheten $p \cdot q \cdot p = p^2 q$.
- Sekvensen $(1, 1, 0)$ har sannolikheten $p \cdot p \cdot q = p^2 q$.
- Se `dbinom(x, size, prob)` och `ManipDistributions.R`

BINOMIALFÖRDELNINGEN

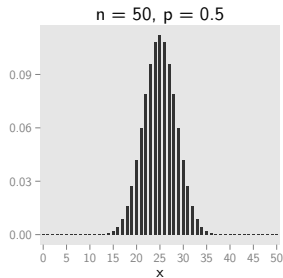
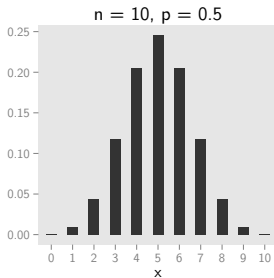
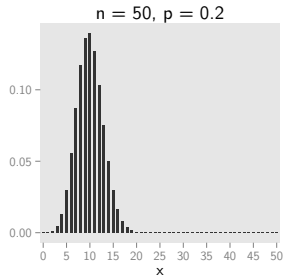
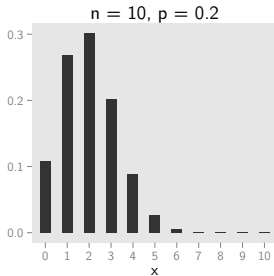
- ▶ Binomialfördelningen passar data som:
 - ▶ **diskreta icke-negativa heltal**
 - ▶ kan anta alla **heltal mellan 0 och n** .
- ▶ Passande: hur många elever i klass 5A kan simma?
- ▶ Inte passande: hur många mål gör IFK Norrköping på lördag? (ingen naturlig övre gräns) eller längdmätningar (kontinuerliga).
- ▶ **Väntevärde och varians** för $X \sim \text{Binomial}(n, p)$:

$$\mathbb{E}X = n \cdot p$$

$$\text{Var}(X) = n \cdot p \cdot q$$

- ▶ Bevis: $X \sim \text{Binomial}(n, p)$ innebär att X är en summa av n oberoende Bernoullivariabler. Väntevärde och varians av summan av oberoende variabler.

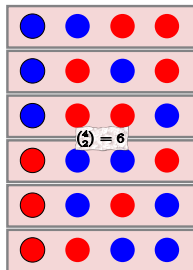
BINOMIALFÖRDELNINGEN



MULTINOMIALFÖRDELNINGEN

- ▶ Bernoullidata: n personer utfrågas om vilket partiblock de föredrar (röd eller blå). n_1 personer svarar röd, n_2 personer svarar blå.
- ▶ Antal sätt vi kan få dessa data: $\binom{n}{n_1} = \frac{n!}{n_1!n_2!}$
- ▶ Sannolikheten för att få n_1 röda i n försök:

$$P(n_1) = \binom{n}{n_1} p^{n_1} q^{n_2},$$

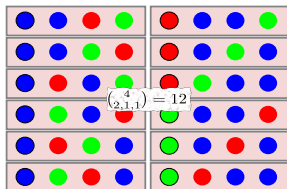


MULTINOMIALFÖRDELNINGEN

- ▶ Multinomials data: n personer utfrågas om vilket partiblock de föredrar (röd, blå eller grön). n_1 personer svarar blå, n_2 personer svarar röd och n_3 personer svarar grön.
- ▶ Antal sätt vi kan få dessa data ges av **multinomialkoefficienten**:
 $\binom{n}{n_1 n_2 n_3} = \frac{n!}{n_1! n_2! n_3!}$ och

$$P(n_1, n_2, n_3) = \binom{n}{n_1 n_2 n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3},$$

- ▶ Notera att multinomialfördelningen är en simultanfördelning för **tre** slumpvariabler: N_1 , N_2 och N_3 .



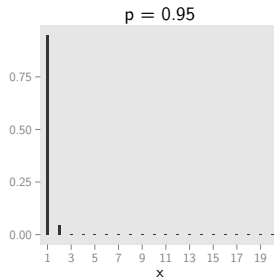
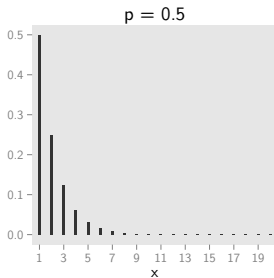
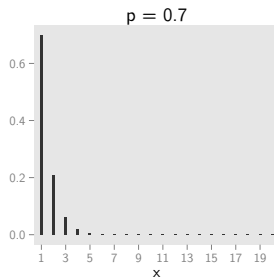
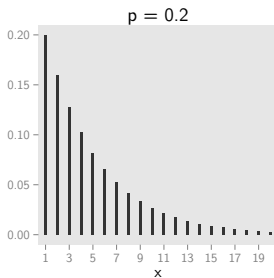
GEOMETRISK FÖRDELNING

- ▶ Låt X_1, X_2, \dots vara en sekvens Bernoulli försök med sannolikhet p .
- ▶ $X =$ antalet försök för att få ett lyckat försök.
- ▶ $X \sim \text{Geo}(p)$ med pmf

$$P(x) = (1 - p)^{x-1} p, x = 1, 2, \dots$$

- ▶ Geometrisk fördelning passar data:
 - ▶ som antar **diskreta icke-negativa heltal**: $0, 1, 2, \dots$
 - ▶ som **inte har en övre gräns** (jfr binomial)
 - ▶ med **monotont avtagande pmf**.
- ▶ Egenskaper för $X \sim \text{Geo}(p)$
 - ▶ $\mathbb{E}X = 1/p$
 - ▶ $\text{Var}(X) = \frac{1-p}{p^2}$
- ▶ Väntevärde och varians beräknas med hjälp av den geometriska serien, se sid 61 i Baron.

GEOMETRISK FÖRDELNING



EXEMPEL: LEVEL UP!

- ▶ Sannolikheten att du klarar en nivå på ett spel är p . De olika försöken är oberoende. Förväntat antal spel innan du klarar nivån? Svar: $1/p$.
- ▶ Antag nu att du klarar en nivå vid r :te försöket med sannolikheten $1 - (1 - p)^r$. Förväntat antal spel? Svar: inte geometrisk.

```
# Function that simulates the number of game plays when you get better over  
# time.
```

```
SimGameVaryingProbs <- function(p) {  
  success <- FALSE  
  r <- 0  
  while (success == FALSE) {  
    r = r + 1  
    if (runif(1) < 1 - (1 - p)^r)  
      success = TRUE  
  }  
  return(r)  
}
```

```
nSim <- 500 # Number of simulations  
numberOfTries <- matrix(NA, nSim, 1) # Setting up storage  
for (i in 1:nSim) {  
  numberOfTries[i] <- SimGameVaryingProbs(p = 0.01)  
}  
mean(numberOfTries)
```

```
## [1] 11.53
```

POISSONFÖRDELNING

Definition. En Poissonfördelad slumpvariabel med frekvens λ , $X \sim Po(\lambda)$, har pmf

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

- ▶ Egenskaper: Om $X \sim Po(\lambda)$
 - ▶ $\mathbb{E}X = \lambda$
 - ▶ $Var(X) = \lambda$
- ▶ Väntevärde och varians beräknas med Taylorutvecklingen:

$$e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

- ▶ Poissonfördelningen passar data:
 - ▶ som antar **diskreta icke-negativa heltal**: 0,1,2,...
 - ▶ som **inte har en övre gräns** (jfr binomial)
 - ▶ vars väntevärde och varians är ungefär lika

POISSONFÖRDELNING

- ▶ Exempel 1: antalet upptäckta buggar i en kod.
- ▶ Exempel 2: antalet döda i trafiken under år 2014.
- ▶ Poissonfördelning med $\lambda = n \cdot p$ kan användas för att approximera binomialfördelningen när $n \geq 30$ and $p \leq 0.05$.

POISSONFÖRDELNING

