

# SANNOLIKHETSLÄRA OCH STATISTIK

## FÖRELÄSNING 10

Mattias Villani

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



# ÖVERSIKT

- ▶ Bayesiansk inferens
- ▶ Binomialmodell med beta prior
- ▶ Normalmodell med normal prior
- ▶ Multinomialmodell med Dirichlet prior

# FREKVENTISTISK INFERENS

- ▶ Hittills på kursen: **frekventistisk inferens**.
  - ▶ **Parametrar  $\theta$  är fixa** (icke slumpmässiga) storheter.
  - ▶ **Data är slumpvariabler:**  $f(X_1, \dots, X_n | \theta)$ .
- ▶ Frekventistisk inferens: hur en **metod** beter sig över **upprepade stickprov** från populationen.
- ▶ **Samplingfördelningar** är i fokus. 'Vilka värden kan min estimator förväntas anta för olika stickprov?'
- ▶ **Väntevärderiktighet:** 'min skattningsmetod kommer att vara korrekt i genomsnitt' (sett över alla möjliga stickprov).
- ▶ **Konfidensintervall:** 'min intervallskattningsmetod kommer att täcka det sanna parametervärdet  $\theta$  i 95% av alla möjliga stickprov från populationen'.
- ▶ **Hypotestest:** 'min testmetod kommer bara att dra fel slutsats i 5% av alla stickprov om nollhypotesen är sann'.

# SUBJEKTIVA SANNOLIKHETER

- ▶ Du **vet inte** värdet på en populationsparameter  $\theta$ . Du är **osäker** om  $\theta$ . Påståendet  $P(\theta \leq 2)$  är meningsfullt.
- ▶ Det är **osäkerheten** som är **relevant**. Om  $\theta$  är en fix, konstant, storhet eller ej spelar ingen roll.
- ▶ Jag vet inte 10:e decimalen av  $\pi$ . Då kan jag säga

$$P(10 : \text{e decimal av } \pi = 9) = 1/10.$$

- ▶ Det är **min** osäkerhet som spelar roll. Du kanske vet 10:e decimalen av  $\pi$ . För mig är  $\pi$  osäker och jag kan prata om sannolikhetsfördelningen för 10:e decimalen av  $\pi$ .
- ▶ Sannolikheter är ett **subjektivt** mått på personlig **grad av tilltro**.
- ▶ **Bayesiansk statistik** bygger på ett subjektivt sannolikhetsbegrepp.

# THOMAS BAYES 1701-1761



# SUBJEKTIVITET I VETENSKAPEN!



# BAYESIANSK INFERENS

- ▶ Bernoullimodellen:  $X_1, \dots, X_n | \theta \sim \text{Bern}(\theta)$ . T ex slantsingling.
- ▶ Sannolikheten för krona,  $\theta$ , är okänd.
- ▶ Innan vi har börjat singla slant beskriver jag min osäkerhet om  $\theta$  med min **apriorifördelning**:  $\pi(\theta)$ .
- ▶ **a priori** = **före** (före jag har observerat data).
- ▶ Antag nu att vi har observerat ett antal slantsinglingar:  $X_1 = x_1, \dots, X_n = x_n$  (t ex 0, 0, 1, 1, 0).
- ▶ Hur bör vi **uppdatera** vår apriorifördelning med denna datainformation? Hur lär vi oss från data? **Learning**.
- ▶ **Aposteriorifördelning**:  $\pi(\theta | x_1, \dots, x_n)$ . Posterior = efter (data).
- ▶ Bayesiansk inferens **betingar på observerade data**.  $P(\text{Okänt} | \text{Känt})$ .

## BAYES SATS UPPDATERAR PRIOR TILL POSTERIOR

- ▶ Antag att  $\theta$  bara kan anta värdena: 0.1, 0.2, ..., 0.9 (diskretisering).
- ▶ Kom ihåg: **Bayes sats** för händelser  $A$  och  $B$ :


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- ▶ Låt t ex  $A = \{\theta = 0.1\}$  och  $B = \{\mathbf{X} = \mathbf{x}\}$ .
- ▶ Bayes sats ger **posteriorfördelningen**:

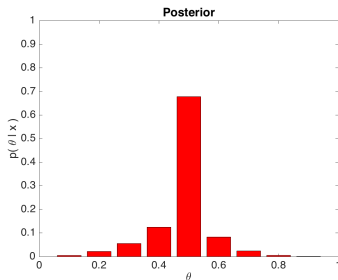
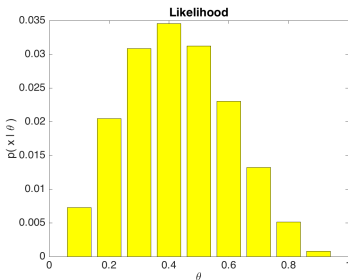
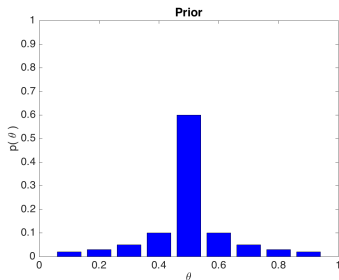
$$P(\theta = 0.1|\mathbf{x}) = \frac{P(\mathbf{x}|\theta = 0.1)P(\theta = 0.1)}{P(\mathbf{x})}$$

där satsen om total sannolikhet ger

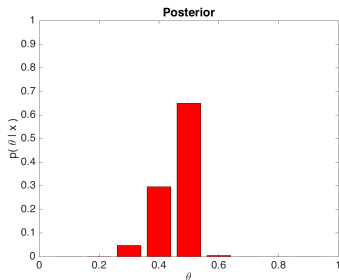
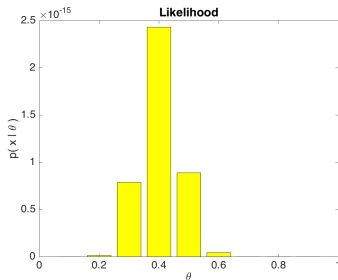
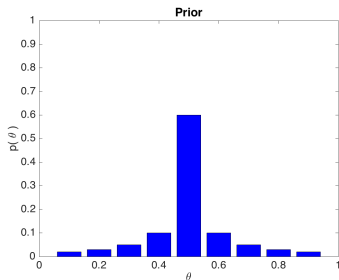
$$P(\mathbf{x}) = P(\mathbf{x}|\theta = 0.1)P(\theta = 0.1) + \dots + P(\mathbf{x}|\theta = 0.9)P(\theta = 0.9)$$



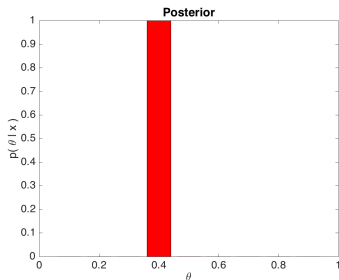
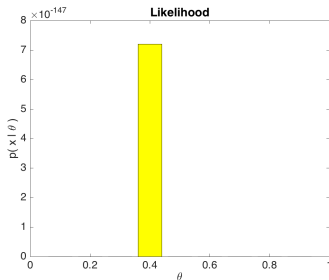
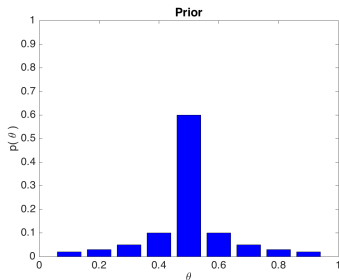
# BERNOULLIMODELL - S=2, F=3



# BERNOULLIMODELL - $s=20$ , $F=30$



# BERNOULLIMODELL - $s=200$ , $F=300$



# BAYES SATS FÖR KONTINERLIGA VARIABLER

- ▶ Diskretisering  $\theta \in \{\theta_1, \theta_2, \dots, \theta_K\}$

$$P(\theta = \theta_i | \mathbf{x}) = \frac{P(\mathbf{x} | \theta = \theta_i) P(\theta = \theta_i)}{\sum_{j=1}^K P(\mathbf{x} | \theta = \theta_j) P(\theta = \theta_j)}$$

- ▶ Finare och finare grid ( $\theta_{i+1} - \theta_i \rightarrow 0$ ) ger

$$f(\theta | \mathbf{x}) = \frac{P(\mathbf{x} | \theta) f(\theta)}{\int P(\mathbf{x} | \theta) f(\theta) d\theta},$$

- ▶ **Bayes sats** för kontinuerlig parameter  $\theta$

$$\pi(\theta | \mathbf{x}) = \frac{f(\mathbf{x} | \theta) \pi(\theta)}{\int f(\mathbf{x} | \theta) \pi(\theta) d\theta}$$

- ▶ **Prior:**  $\pi(\theta)$
- ▶ **Likelihood:**  $f(\mathbf{x} | \theta)$
- ▶ **Posterior:**  $\pi(\theta | \mathbf{x})$

# SUBJEKTIVITET OCH OBJEKTIVITET

- ▶  $\pi(\theta)$  är en **subjektiv** fördelning som varierar från person till person baserat på erfarenhet etc.
- ▶ **Hur vi lär oss från data**, dvs uppdaterar från prior till posterior, bestäms av Bayes sats.
- ▶ **Uppdateringsmekanismen är objektiv** (matematik).
- ▶ Resultat: när  $n \rightarrow \infty$  (**stora datamängder**) kommer alla personers posteriors att konvergera till samma fördelning. Objektivitet genom **subjektivt konsensus**.
- ▶ Vid rapportering av resultat kan man använda **icke-informativa apriorifördelningar** (dvs svag information) eller priorinformation som är lättförståelig.
- ▶ Machine learning: mycket vanligt med aprioriinformation av typen: 'Jag tror att den okända funktionen är **mjuk**, men jag vet inte mycket mer om den exakta funktionsformen'.

## BERNOULLI MED BETA PRIOR

- ▶ Bernoullimodellen:  $X_1, \dots, X_n | \theta \sim \text{Bern}(\theta)$ . **Likelihood:**  $\theta^s (1 - \theta)^f$ .
- ▶  $\theta \in [0, 1]$ . Lämplig **prior:**  $\theta \sim \text{Beta}(\alpha, \beta)$ :

$$\pi(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} \theta^{\beta-1}$$

- ▶ **Posterior**

$$\begin{aligned} \pi(\theta | \mathbf{x}) &= \frac{f(\mathbf{x} | \theta) \pi(\theta)}{\int f(\mathbf{x} | \theta) \pi(\theta) d\theta} = \frac{\theta^s (1 - \theta)^f \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} \theta^{\beta-1}}{\int \theta^s (1 - \theta)^f \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} \theta^{\beta-1} d\theta} \\ &= \frac{\theta^{\alpha+s-1} (1 - \theta)^{\beta+f-1}}{\int \theta^{\alpha+s-1} (1 - \theta)^{\beta+f-1} d\theta} = c \cdot \theta^{\alpha+s-1} (1 - \theta)^{\beta+f-1} \end{aligned}$$

där  $c = 1 / \int \theta^{\alpha+s-1} (1 - \theta)^{\beta+f-1} d\theta$  är en konstant (beror inte på  $\theta$ ).

- ▶ En täthet på formen  $c \cdot \theta^{\alpha+s-1} (1 - \theta)^{\beta+f-1}$  känns igen som en  $\text{Beta}(\alpha + s, \beta + f)$ :

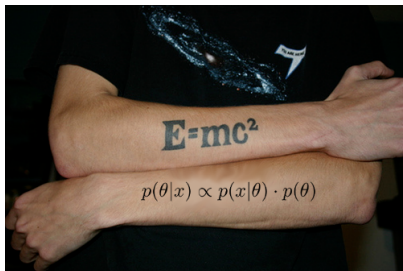
$$\pi(\theta | \mathbf{x}) = \frac{1}{B(\alpha + s, \beta + f)} \theta^{(\alpha+s)-1} \theta^{(\beta+f)-1}.$$

# BAYES SATS PÅ PROPORTIONELL FORM

- ▶ Notera att vi aldrig behövde räkna ut nämnaren i Bayes sats:  
 $\int f(\mathbf{x}|\theta)\pi(\theta)d\theta$ . Vi kände igen Beta-fördelningen ändå.
- ▶ Tätheter måste integrera till ett. Proportionalitetskonstanter kan vi "strunta i".
- ▶ Enklare form av Bayes sats:

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$



# BERNOULLI-EXEMPEL: SPAM

► George har gått igenom 4601 e-mail (elbrev). 1813 av dessa var spam.

► **Modell:** Låt  $x_i = 1$  om det i:te elbrevet var spam. Antag  
 $x_i | \theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$

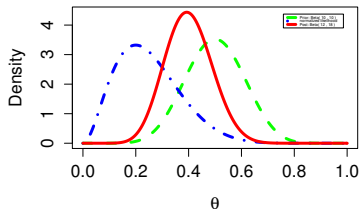
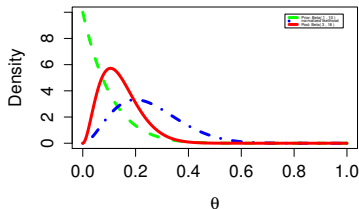
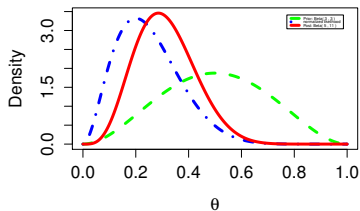
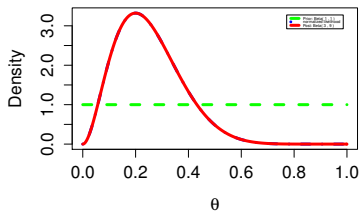
► **Prior:**  $\theta \sim \text{Beta}(\alpha, \beta)$ .

► **Posterior**

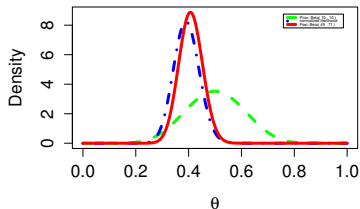
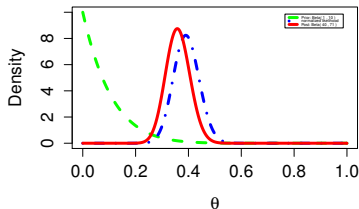
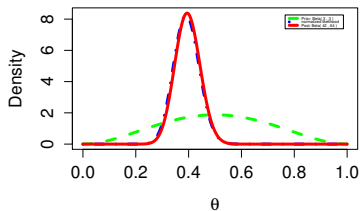
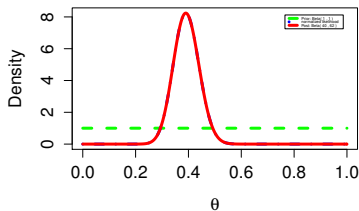
$$\theta | \mathbf{x} \sim \text{Beta}(\alpha + 1813, \beta + 2788)$$



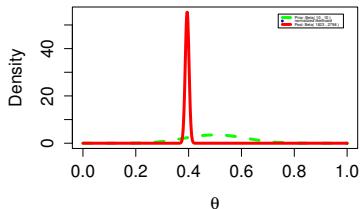
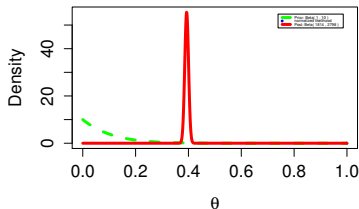
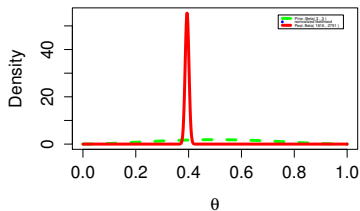
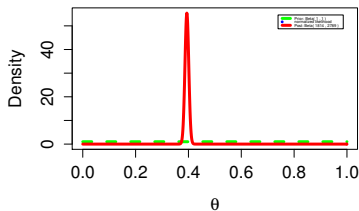
# SPAM DATA (N=10): FYRA OLIKA PRIORS



# SPAM DATA (N=100): FYRA OLIKA PRIORS



# SPAM DATA (N=4601): FYRA OLIKA PRIORS



# NORMAL DATA, KÄND VARIANS - NORMAL PRIOR

► **Modell:**  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$ ,  $\sigma^2$  känt.

► **Prior**

$$\theta \sim N(\mu, \tau^2)$$

► **Posterior**

$$\begin{aligned} p(\theta | x_1, \dots, x_n) &\propto p(x_1, \dots, x_n | \theta, \sigma^2) p(\theta) \\ &\propto N(\theta | \mu_x, \tau_x^2), \end{aligned}$$

där

$$\frac{1}{\tau_x^2} = \frac{n}{\sigma^2} + \frac{1}{\tau^2},$$

$$\mu_x = w\bar{x} + (1 - w)\mu,$$

och

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

► Se Baron s. 344 för en härledning.

# NORMAL DATA, KÄND VARIANS - NORMAL PRIOR

$$\theta \sim N(\mu, \tau^2) \xrightarrow{x_1, \dots, x_n} \theta | \mathbf{x} \sim N(\mu_x, \tau_x^2).$$

Posterior precision = Data precision + Prior precision

Posterior väntevärde =

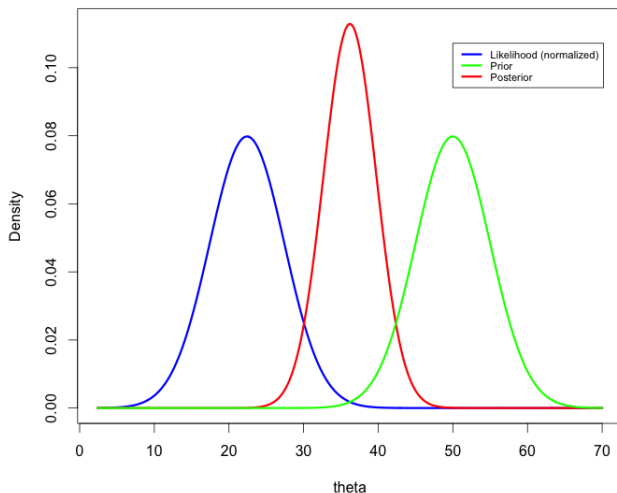
$$\frac{\text{Data precision}}{\text{Posterior precision}} (\text{Data medelvärde}) + \frac{\text{Prior precision}}{\text{Posterior precision}} (\text{Prior väntevärde})$$

# NEDLADDNINGSHASTIGHETER

- ▶ Data:  $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ .
- ▶ Modell:  $X_1, \dots, X_5 \sim N(\theta, \sigma^2)$ .
- ▶ Antag  $\sigma = 5$  (mätningar kan variera  $\pm 10$  MBit med 95% sannolikhet)
- ▶ Min prior:  $\theta \sim N(50, 5^2)$ .

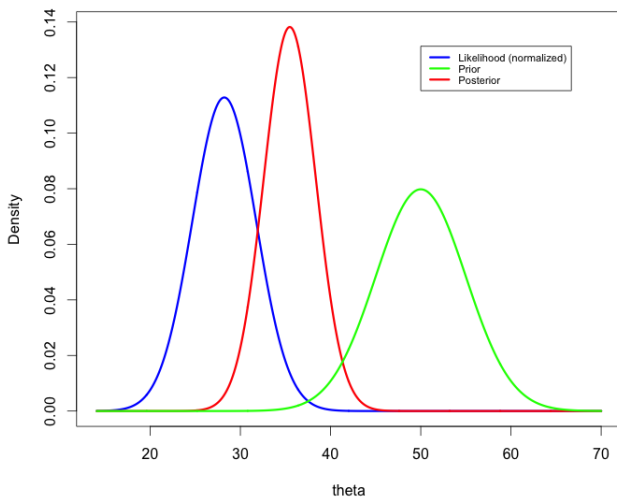
# NEDLADDNINGSHASTIGHETER N=1

Download speed data:  $x=(22.42)$



# NEDLADDNINGSHASTIGHETER N=2

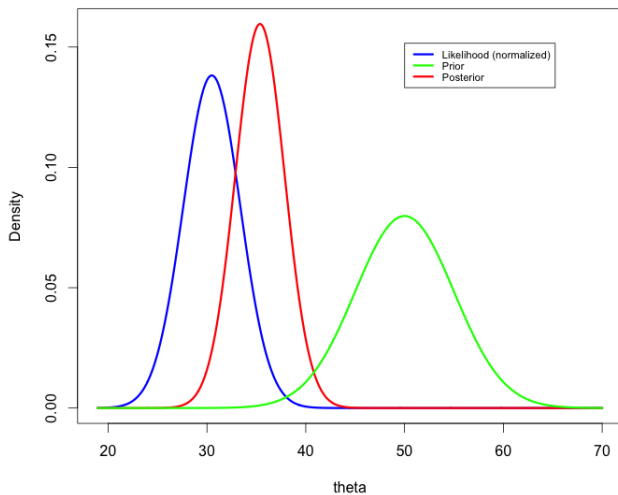
Download speed data:  $x=(22.42, 34.01)$





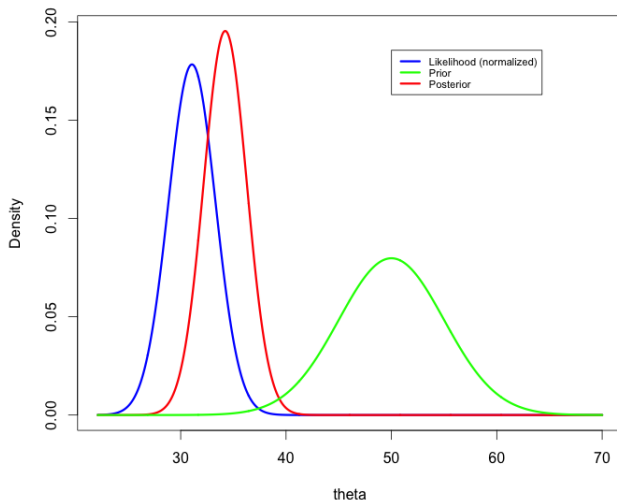
# NEDLADDNINGSHASTIGHETER $N=3$

Download speed data:  $x=(22.42, 34.01, 35.04)$



# NEDLADDNINGSHASTIGHETER N=5

Download speed data:  $x=(22.42, 34.01, 35.04, 38.74, 25.15)$



# MULTINOMIAL MODELL MED DIRICHLET PRIOR

- ▶ *Data*:  $y = (y_1, \dots, y_K)$ .  $y_k$  = antalet obs i den  $k$ :te klassen.
- ▶ Exempel:  $K = 8$ ,  $y_k$  antal som röstar på parti  $k$  i en valundersökning med  $n = \sum_{k=1}^K y_k$  tillfrågade personer.
- ▶ **Multinomial modell**:

$$p(y|\theta) \propto \prod_{k=1}^K \theta_k^{y_k}, \text{ där } \sum_{k=1}^K \theta_k = 1.$$

- ▶ **Konjugerad prior**:  $\text{Dirichlet}(\alpha_1, \dots, \alpha_K)$

$$p(\theta) \propto \prod_{k=1}^K \theta_k^{\alpha_k - 1}.$$

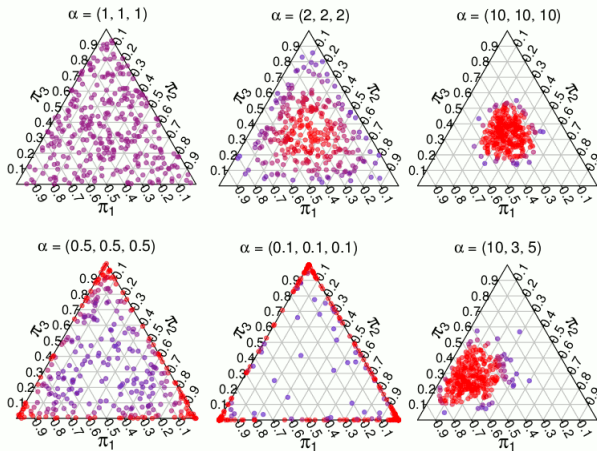
- ▶ **Väntevärde** för  $\theta = (\theta_1, \dots, \theta_K)' \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$

$$\mathbb{E}(\theta_k) = \frac{\alpha_k}{\sum_{j=1}^K \alpha_j}$$

- ▶ Variansen minskar för större  $\alpha$ -värden. **Icke-informativ** prior har små värden, t ex  $\alpha_k = 1$  för alla  $k$ .

# DIRICHLETFÖRDELNINGEN

Draws from a 3-dimensional Dirichlet with different  $\alpha$



# MULTINOMIAL MODEL WITH DIRICHLET PRIOR

## ► *Uppdatering från prior till posterior:*

*Modell:*  $y = (y_1, \dots, y_K) \sim \text{Multin}(n; \theta_1, \dots, \theta_K)$

*Prior :*  $\theta = (\theta_1, \dots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$

*Posterior :*  $\theta|y \sim \text{Dirichlet}(\alpha_1 + y_1, \dots, \alpha_K + y_K)$ .

## ► **Simulering** från en Dirichlet-fördelning:

- Slumpa  $x_1 \sim \text{Gamma}(\alpha_1, 1), \dots, x_K \sim \text{Gamma}(\alpha_K, 1)$ .
- Beräkna  $z_k = x_k / (\sum_{j=1}^K x_j)$ .
- $\mathbf{z} = (z_1, \dots, z_K)$  är nu en slumpvektor från  $\text{Dirichlet}(\alpha_1, \dots, \alpha_K)$ -fördelningen.

# EXEMPEL: MARKNADSANDELAR

- ▶ En undersökning bland 513 smartphone-ägare gav:
  - ▶ 180 föredrar en iPhone
  - ▶ 230 föredrar en Androidtelefon
  - ▶ 62 föredrar en Blackberrytelefon
  - ▶ 41 föredrar något annat märke
- ▶ Tidigare undersökning: iPhone 30%, Android 30%, Blackberry 20% och Annat 20%.
- ▶  $P(\text{Android har störst marknadsandel} \mid \text{Data})$
- ▶ Prior:  $\alpha_1 = 15, \alpha_2 = 15, \alpha_3 = 10$  och  $\alpha_4 = 10$  (prior info motsvarar en undersökning med 50 svarande)
- ▶ Posterior:  $(\theta_1, \theta_2, \theta_3, \theta_4) \mid \mathbf{y} \sim \text{Dirichlet}(195, 245, 72, 51)$

# R KOD FÖR MARKNADSANDELAR

```
# Setting up data and prior
y <- c(180,230,62,41) # The cell phone survey data (K=4)
alpha <- c(15,15,10,10) # Dirichlet prior hyperparameters
nIter <- 1000 # Number of posterior draws

# Defining a function that simulates from a Dirichlet distribution
SimDirichlet <- function(nIter, param){
  nCat <- length(param)
  thetaDraws <- as.data.frame(matrix(NA, nIter, nCat)) # Storage.
  for (j in 1:nCat){
    thetaDraws[,j] <- rgamma(nIter,param[j],1)
  }
  for (i in 1:nIter){
    thetaDraws[i,] = thetaDraws[i,]/sum(thetaDraws[i,])
  }
  return(thetaDraws)
}

# Posterior sampling from Dirichlet posterior
thetaDraws <- SimDirichlet(nIter,y + alpha)
```

# R KOD FÖR MARKNADSANDELAR

```
# Posterior mean and standard deviation of Androids share (in %)
message(mean(100*thetaDraws[,2]))

## 43.5281011565501

message(sd(100*thetaDraws[,2]))

## 2.12159111747558

# Computing the posterior probability that Android is the largest
PrAndroidLargest <- sum(thetaDraws[,2] > max(thetaDraws[,c(1,3,4)]))/nIter
message(paste('Pr(Android has the largest market share) = ', PrAndroidLargest))

## Pr(Android has the largest market share) = 0.907
```



# R CODE FOR MARKET SHARE EXAMPLE, CONT

