Projekt 1

Josefine Sandell & Villiam Axtelius

2026-02-20

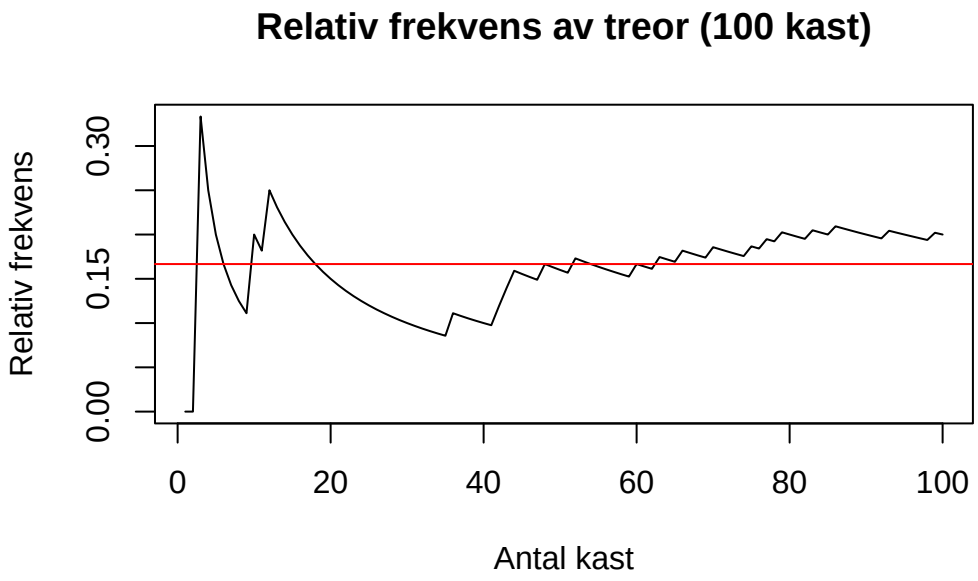
Table of contents

Sammanfattning

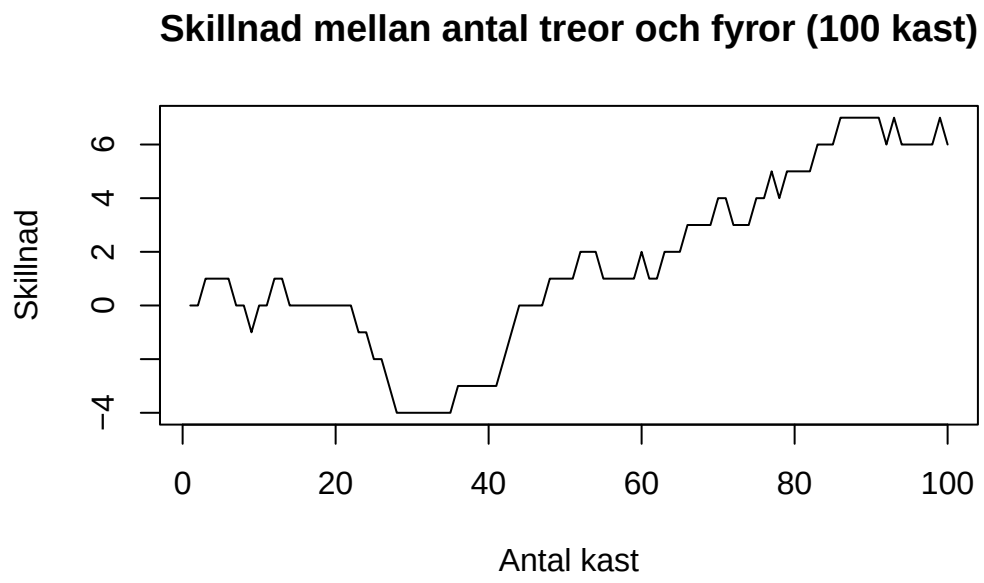
I denna rapport redovisar vi de olika basfunktionerna i R samt dess tillämpningar inom sannolikhetsteorin. Vi börjar med enkla simuleringar av tärningskast och går vidare till skattningar av sannolikhetsfunktionen, inklusive binomial- och normalfördelningen. Vi visualiserar dessa fördelningar på olika sätt och sedan tillämpar våra kunskaper i en studie om jordbävningar. Slutligen undersöker vi centrala gränsvärdessatsen i praktiken med hjälp av en dataset om Hoghkins sjukdomar.

Del 1 – Tärningssimulering

```r
set.seed(123)

# 1.1 100 kast

n <- 100
X <- floor(6 * runif(n) + 1)

# Treor
Y3 <- X == 3

# Relativ frekvens av treor
relf <- cumsum(Y3) / seq_len(n)

plot(relf, type = "l",
     main = "Relativ frekvens av treor (100 kast)",
     xlab = "Antal kast",
     ylab = "Relativ frekvens")
abline(h = 1/6, col = "red")
```

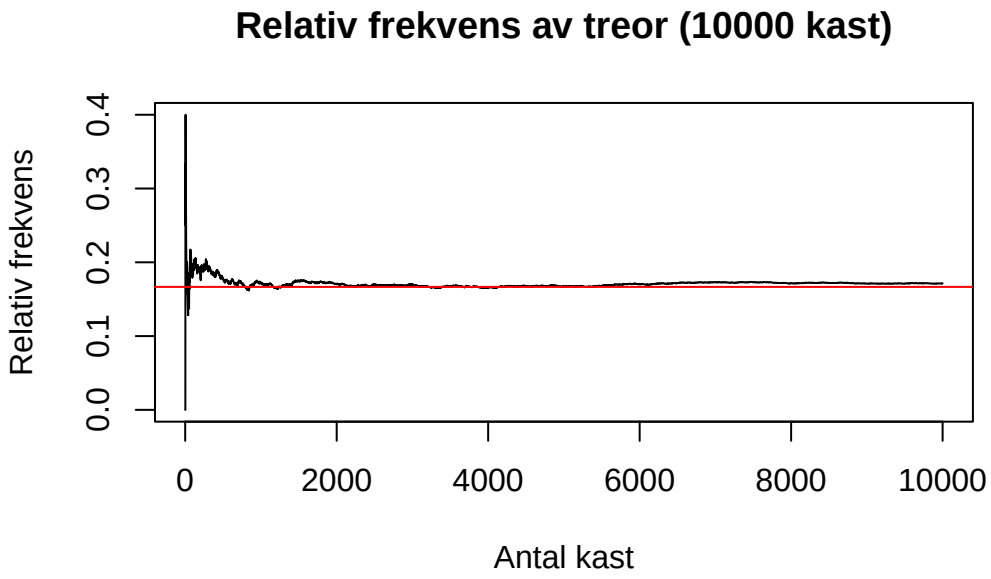


```r
# Fyror
Y4 <- X == 4

# Skillnad mellan antal treor och fyror
diff_34 <- cumsum(Y3) - cumsum(Y4)

plot(diff_34, type = "l",
     main = "Skillnad mellan antal treor och fyror (100 kast)",
     xlab = "Antal kast",
     ylab = "Skillnad")
```

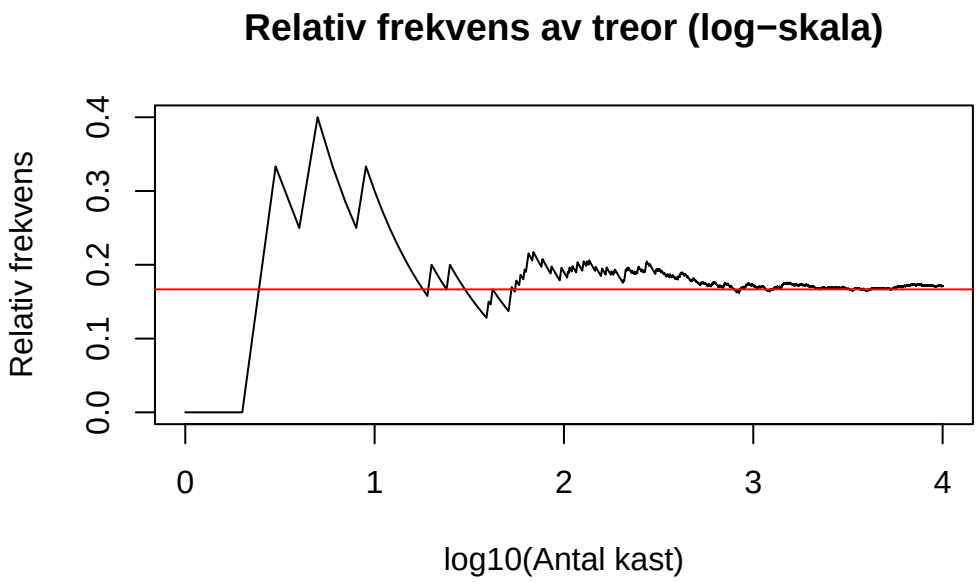



```r
# 1.2 10 000 kast

n1 <- 10000
X1 <- floor(6 * runif(n1) + 1)

Y3_1 <- X1 == 3
Y4_1 <- X1 == 4

relf1 <- cumsum(Y3_1) / seq_len(n1)
diff1 <- cumsum(Y3_1) - cumsum(Y4_1)
```

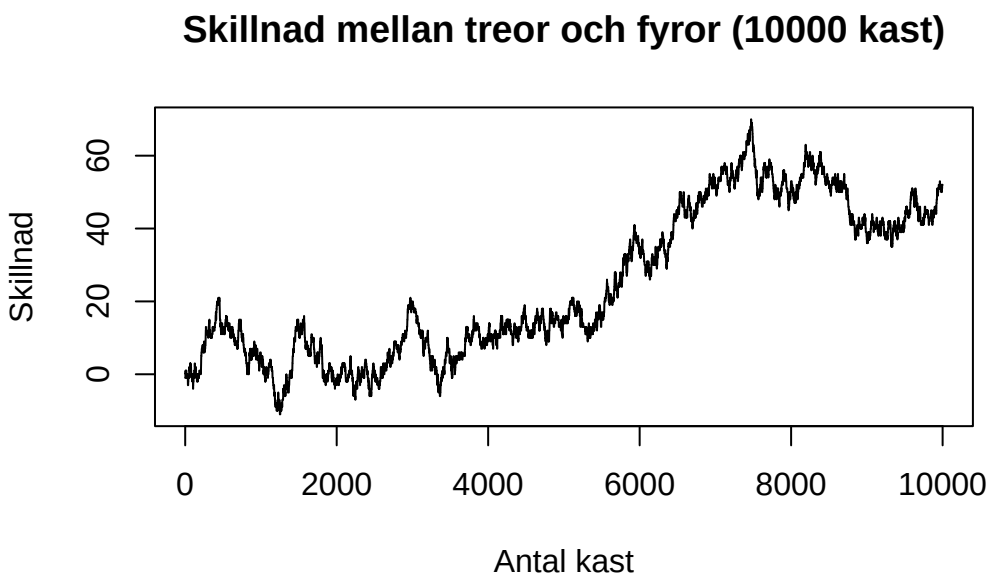
```r
# Relativ frekvens
plot(relf1, type = "l",
     main = "Relativ frekvens av treor (10000 kast)",
     xlab = "Antal kast",
     ylab = "Relativ frekvens")
abline(h = 1/6, col = "red")
```



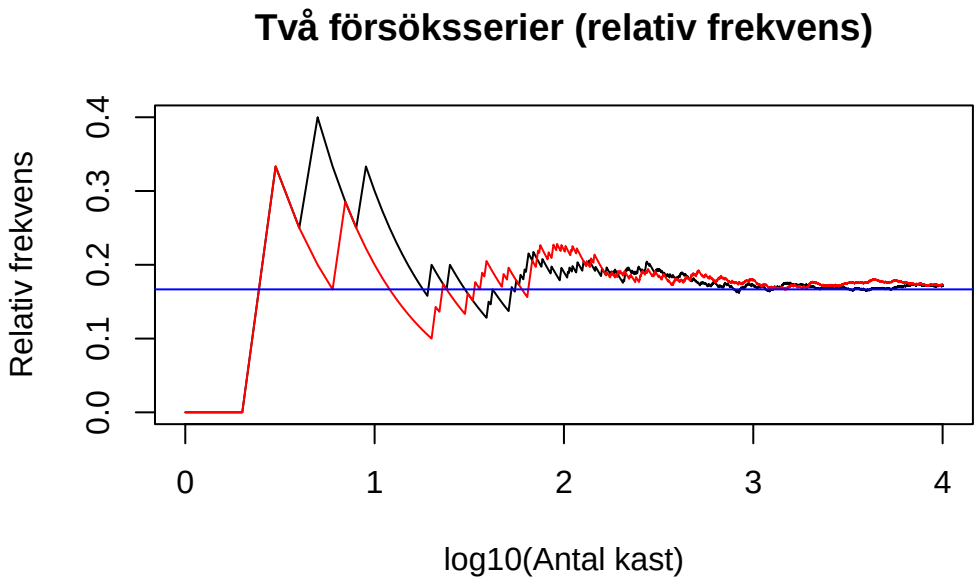

```r
# Log-skala
plot(log10(seq_len(n1)), relf1, type = "l",
     main = "Relativ frekvens av treor (log-skala)",
     xlab = "log10(Antal kast)",
     ylab = "Relativ frekvens")
abline(h = 1/6, col = "red")
```



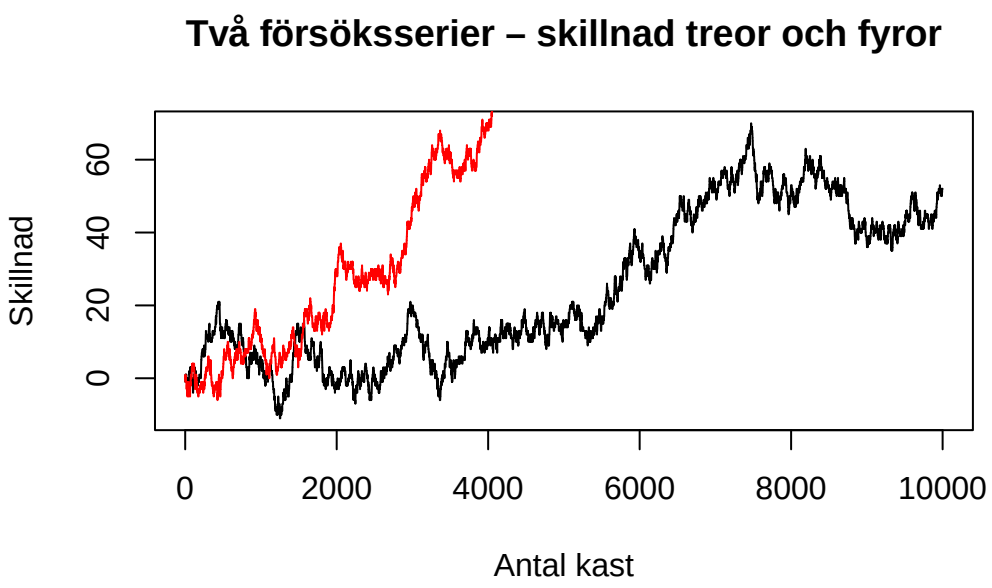

```
# Skillnad
plot(diff1, type = "l",
     main = "Skillnad mellan treor och fyror (10000 kast)",
     xlab = "Antal kast",
     ylab = "Skillnad")
```




```r
# 1.2 Andra försöksserie

X2 <- floor(6 * runif(n1) + 1)

Y3_2 <- X2 == 3
Y4_2 <- X2 == 4

relf2 <- cumsum(Y3_2) / seq_len(n1)
diff2 <- cumsum(Y3_2) - cumsum(Y4_2)

# Jämför relativ frekvens
plot(log10(seq_len(n1)), relf1, type = "l",
     main = "Två försöksserier (relativ frekvens)",
     xlab = "log10(Antal kast)",
     ylab = "Relativ frekvens")
lines(log10(seq_len(n1)), relf2, col = "red")
abline(h = 1/6, col = "blue")
```

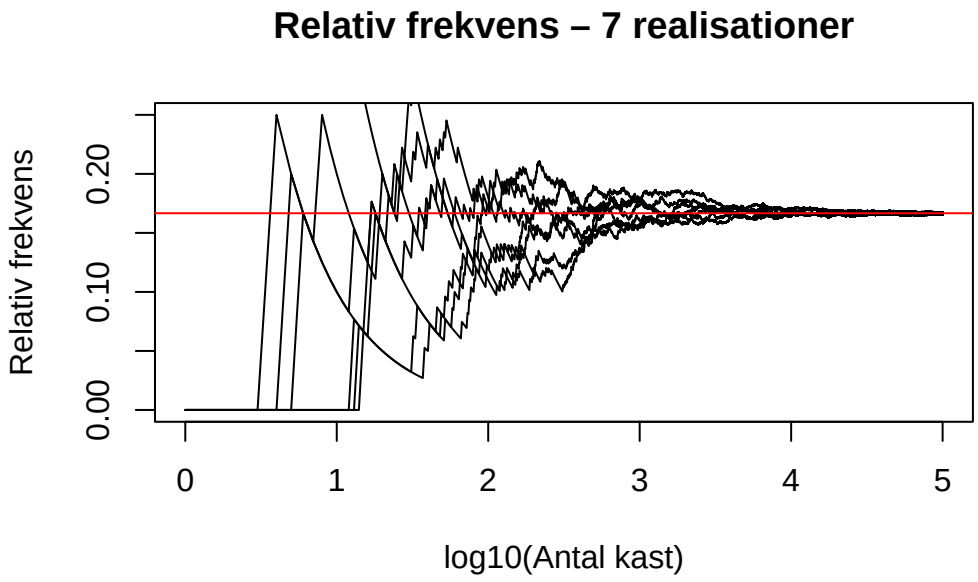



```
# Jämför skillnad
plot(diff1, type = "l",
     main = "Två försöksserier - skillnad treor och fyror",
     xlab = "Antal kast",
     ylab = "Skillnad")
lines(diff2, col = "red")
```

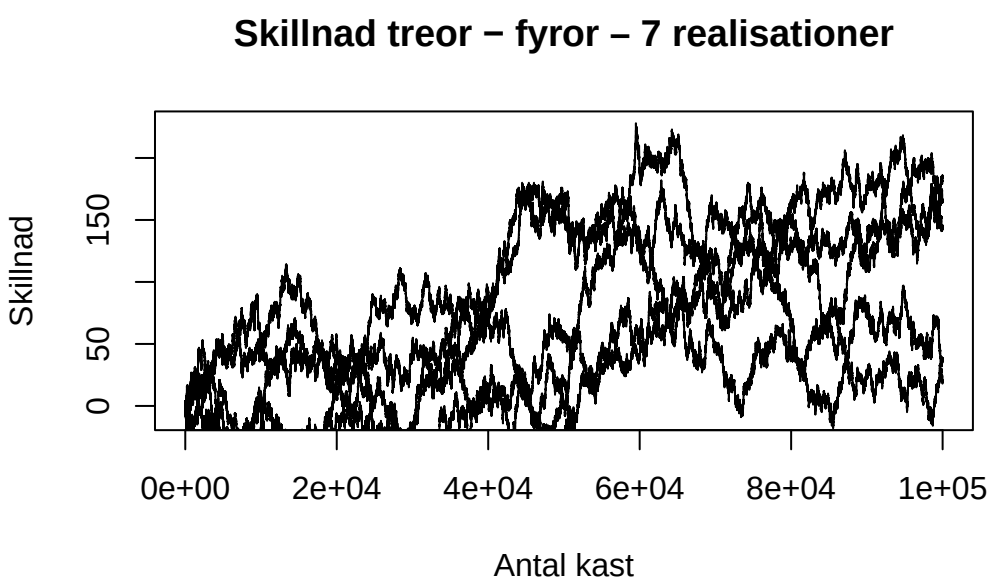



```r
# 1.3 Flera realisationer (7 st)

n_big <- 100000
runs <- 7

relf_mat <- matrix(NA, nrow = n_big, ncol = runs)
diff_mat <- matrix(NA, nrow = n_big, ncol = runs)

for (i in 1:runs) {
  Xi <- floor(6 * runif(n_big) + 1)
  Yi3 <- Xi == 3
  Yi4 <- Xi == 4

  relf_mat[, i] <- cumsum(Yi3) / seq_len(n_big)
  diff_mat[, i] <- cumsum(Yi3) - cumsum(Yi4)
}

# Plot alla relativa frekvenser
plot(log10(seq_len(n_big)), relf_mat[, 1], type = "l",
     main = "Relativ frekvens - 7 realisationer",
     xlab = "log10(Antal kast)",
     ylab = "Relativ frekvens")
```

```r
for (i in 2:runs) {
  lines(log10(seq_len(n_big)), relf_mat[, i])
}

abline(h = 1/6, col = "red")
```




```r
# Plot alla skillnader
plot(diff_mat[, 1], type = "l",
     main = "Skillnad treor - fyror - 7 realisationer",
     xlab = "Antal kast",
     ylab = "Skillnad")

for (i in 2:runs) {
  lines(diff_mat[, i])
}
```



Vid de inledande simuleringarna med 100 kast observerade vi att den relativa frekvensen av treor uppvisade relativt stora variationer och endast i begränsad utsträckning närmade sig det teoretiska värdet 1/6. Kurvan fluktuerade tydligt kring detta värde, vilket indikerar att slumpvariationerna är betydande vid små stickprov. Även skillnaden mellan antalet treor och fyror varierade kraftigt och uppvisade ingen tydlig stabilisering, vilket ytterligare bekräftar att 100 observationer är otillräckliga för att dra säkra slutsatser om sannolikheterna.

När antalet kast ökades till 10 000 framträdde ett tydligare mönster. Den relativa frekvensen av treor närmade sig nu successivt 1/6 och fluktuationerna blev märkbart mindre. I logaritmisk skala blev denna konvergens särskilt tydlig, då kurvan stabiliserades redan efter ett relativt begränsat antal kast. Skillnaden mellan antalet treor och fyror växte visserligen i absolut tal, men gjorde det på ett långsammare och mer regelbundet sätt, vilket indikerar att de relativa skillnaderna avtar med ökande stickprovsstorlek.

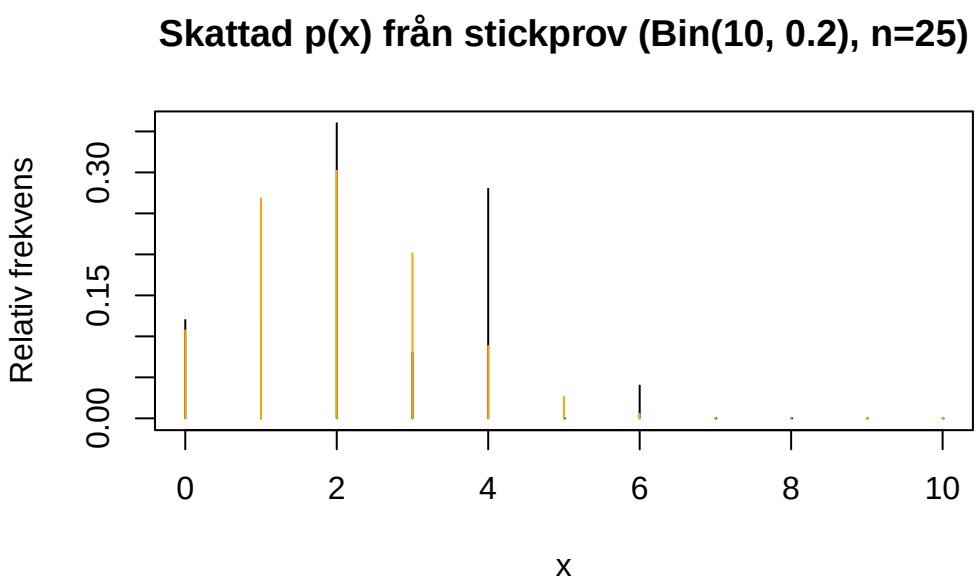
Vid jämförelsen mellan två försöksserier med 10 000 kast kunde vi konstatera att kurvorna uppvisade både likheter och skillnader. Båda serierna konvergerade mot samma teoretiska värde, men följde olika banor dit, särskilt i början av försöken. Detta illustrerar tydligt slumpens roll i varje enskild realisation, samtidigt som den långsiktiga stabiliteten bekräftar den frekventistiska tolkningen av sannolikhetsbegreppet.

Slutligen, genom att studera sju realisationer med 100 000 kast vardera, framträdde lagen om stora tal mycket tydligt. Samtliga kurvor för den relativa frekvensen samlades nära 1/6, och skillnaderna mellan realisationerna blev små. Trots att de absoluta differenserna mellan treor och fyror fortsatte att variera, skedde detta utan någon systematisk trend, vilket är förenligt med antagandet om oberoende och likformigt fördelade utfall. Sammantaget visar dessa

simuleringar hur slumpmässiga variationer dominerar vid små stickprov, men gradvis neutraliseras när antalet observationer ökar, vilket ger empiriskt stöd åt de teoretiska resultaten inom sannolikhetsteorin.

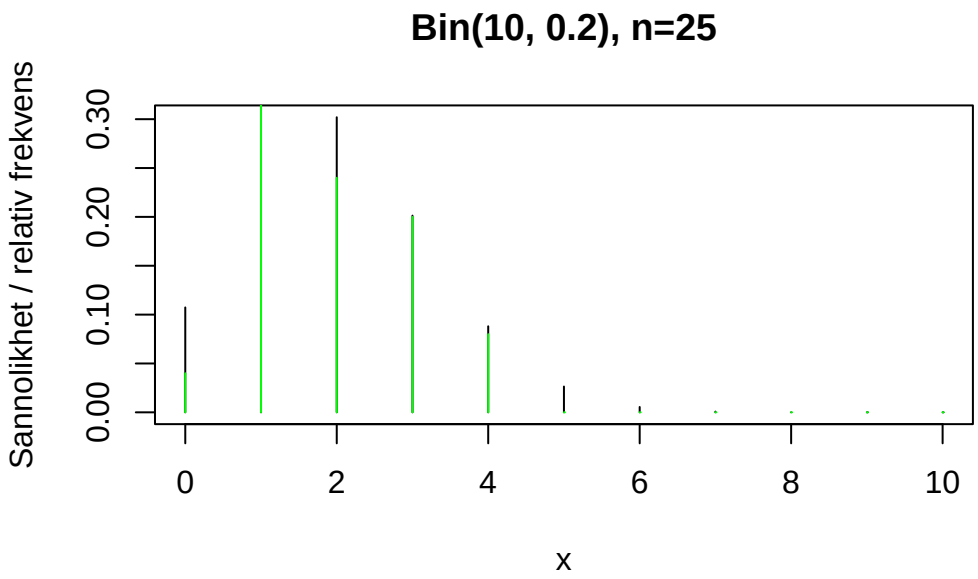
Del 2 - Binomialfördelningen

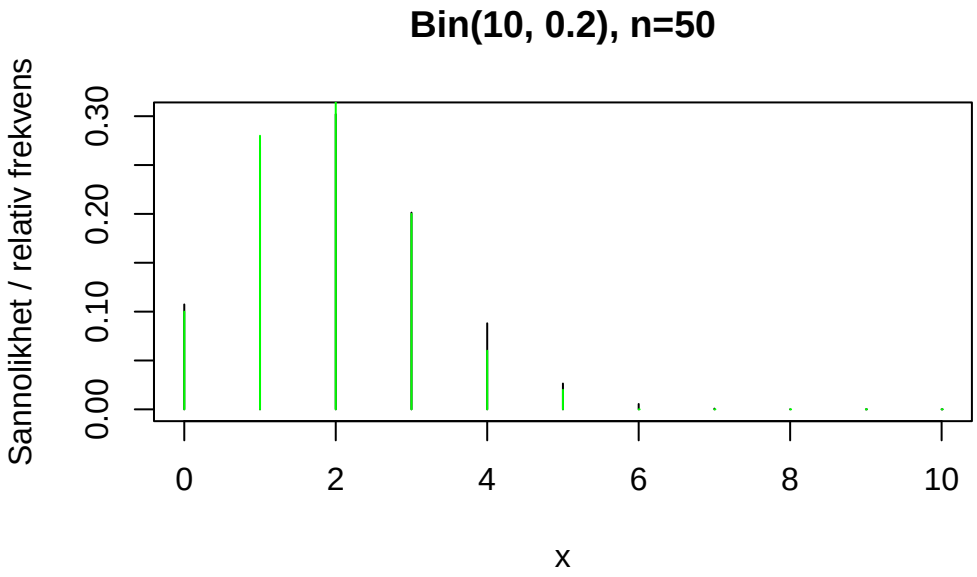
```r
set.seed(123)


# Exempel: 25 observationer från Bin(10, 0.2)


n_obs <- 25
n_trials <- 10
p <- 0.2

xbino1 <- rbinom(n_obs, n_trials, p)

# Frekvenser och relativfrekvenser
xkategori <- factor(xbino1, levels = 0:n_trials)
xfreq <- table(xkategori)
relfreq <- prop.table(xfreq)

# Utfallsrum
xomega <- 0:n_trials

# Plotta skattad sannolikhetsfunktion
plot(xomega, as.numeric(relfreq), type = "h",
     main = "Skattad p(x) från stickprov (Bin(10, 0.2), n=25)",
     xlab = "x",
     ylab = "Relativ frekvens")

# Teoretisk sannolikhetsfunktion
px <- dbinom(xomega, n_trials, p)

# Lägg till teoretisk ovanpå
points(xomega, px, type = "h", col = "orange")
```

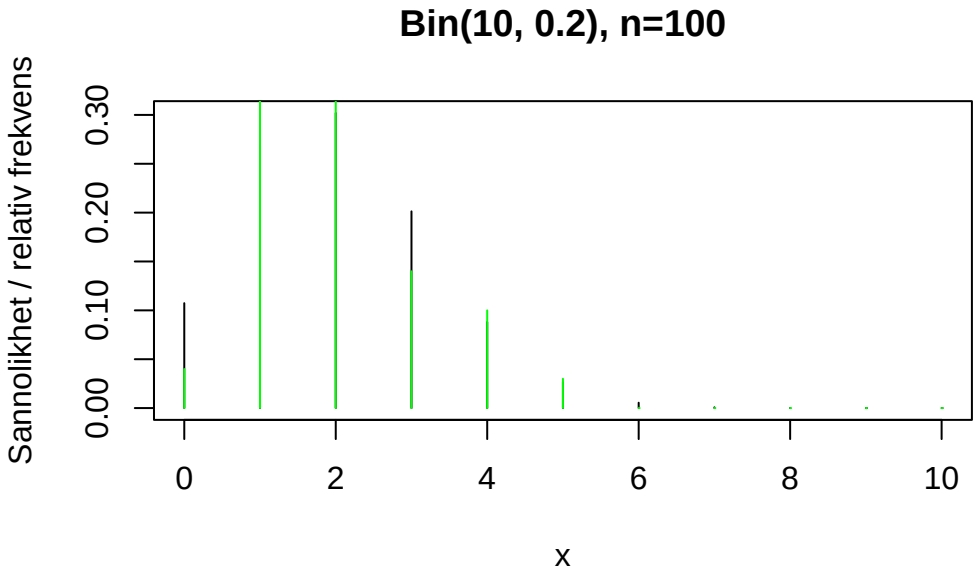


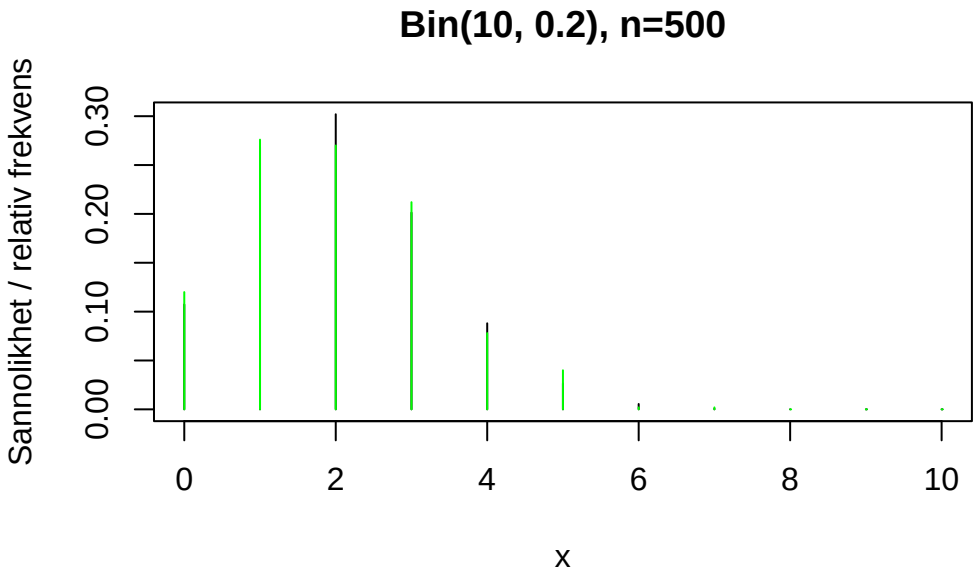

```r
# Funktion för att jämföra skattad och teoretisk p(x)


plot_bino_compare <- function(n_obs, n_trials, p) {
  x <- rbinom(n_obs, n_trials, p)
  xomega <- 0:n_trials

  xkategori <- factor(x, levels = xomega)
  relfreq <- prop.table(table(xkategori))

  px <- dbinom(xomega, n_trials, p)

  plot(xomega, px, type = "h",
       main = paste0("Bin(", n_trials, ", ", p, "), n=", n_obs),
       xlab = "x",
       ylab = "Sannolikhet / relativ frekvens")
  points(xomega, as.numeric(relfreq), type = "h", col = "green")
}


# Undersök olika stickprovsstorlekar för Bin(10, 0.2)
```

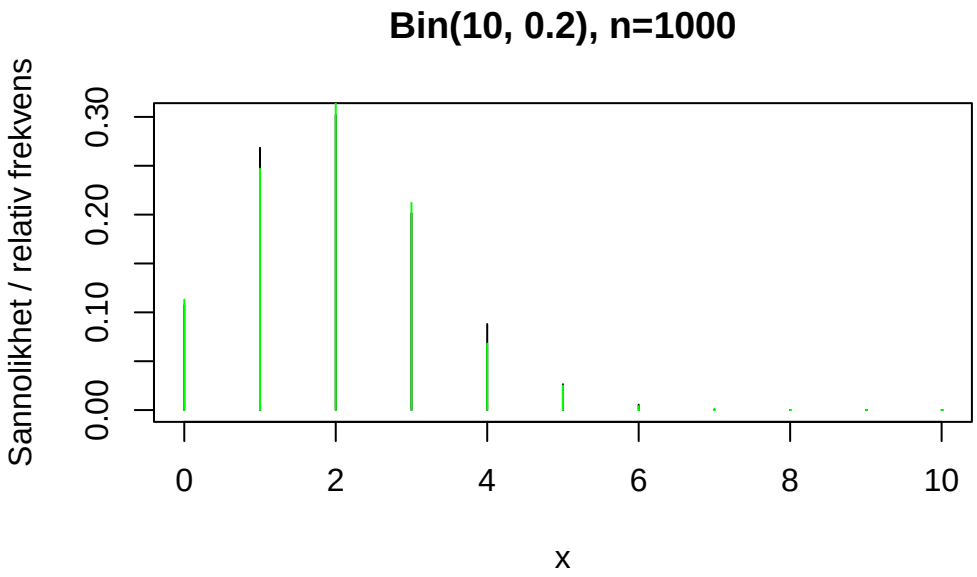
```r
n_list <- c(25, 50, 100, 500, 1000, 5000)

for (n_obs in n_list) {
  plot_bino_compare(n_obs, 10, 0.2)
}
```

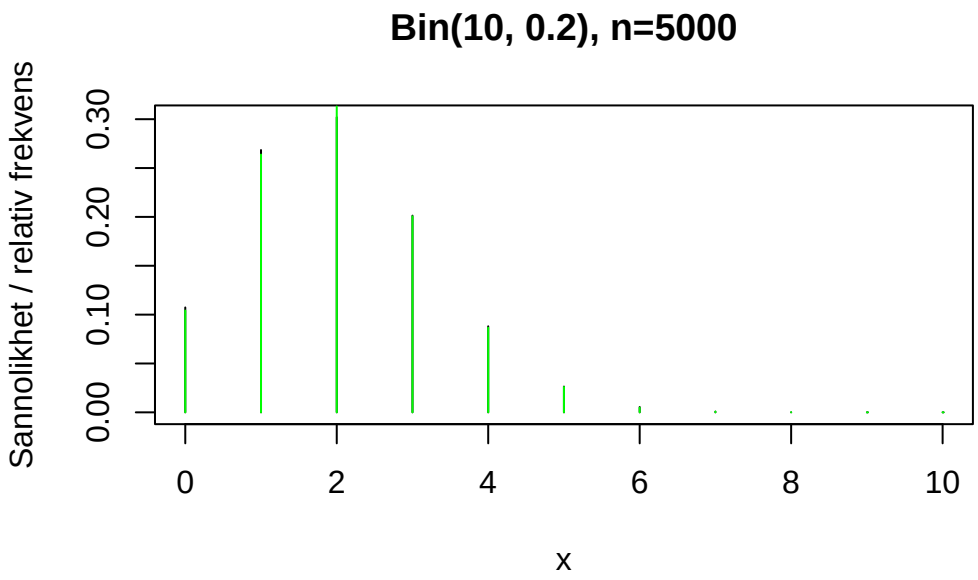



Bin(10, 0.2), n=50
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0.30
0
2
4
6
8
10
x

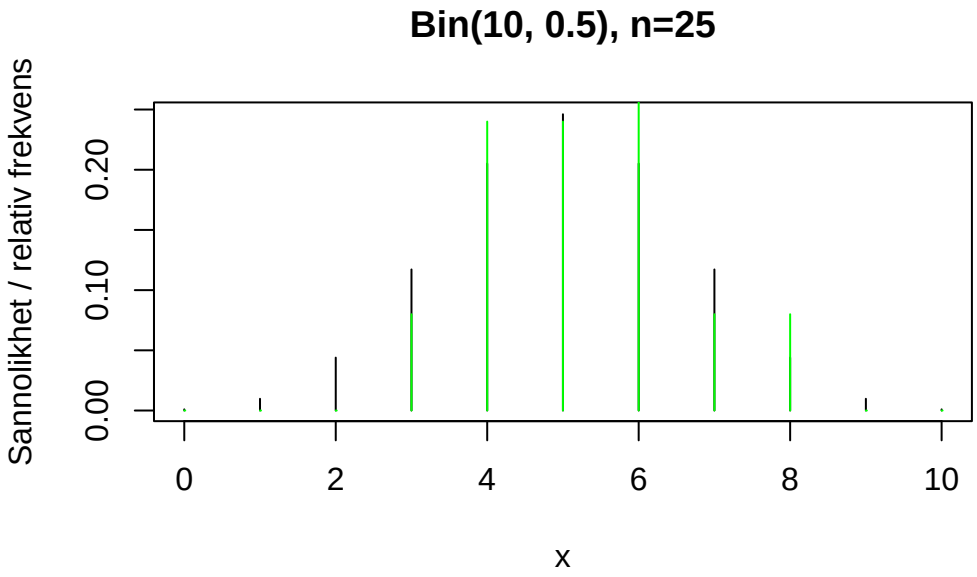

Bin(10, 0.2), n=100
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0.30
0
2
4
6
8
10
x

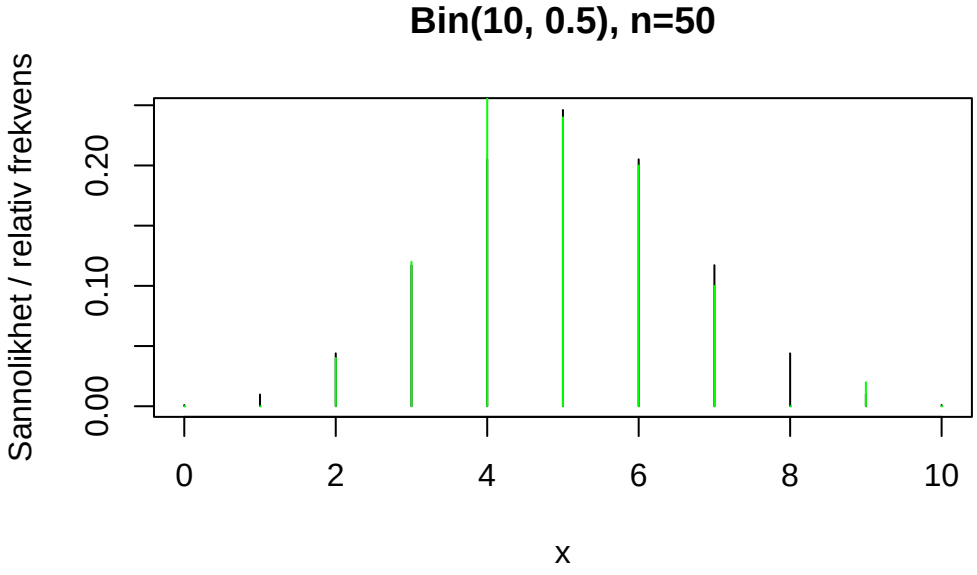

Bin(10, 0.2), n=500
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0.30
0
2
4
6
8
10
x

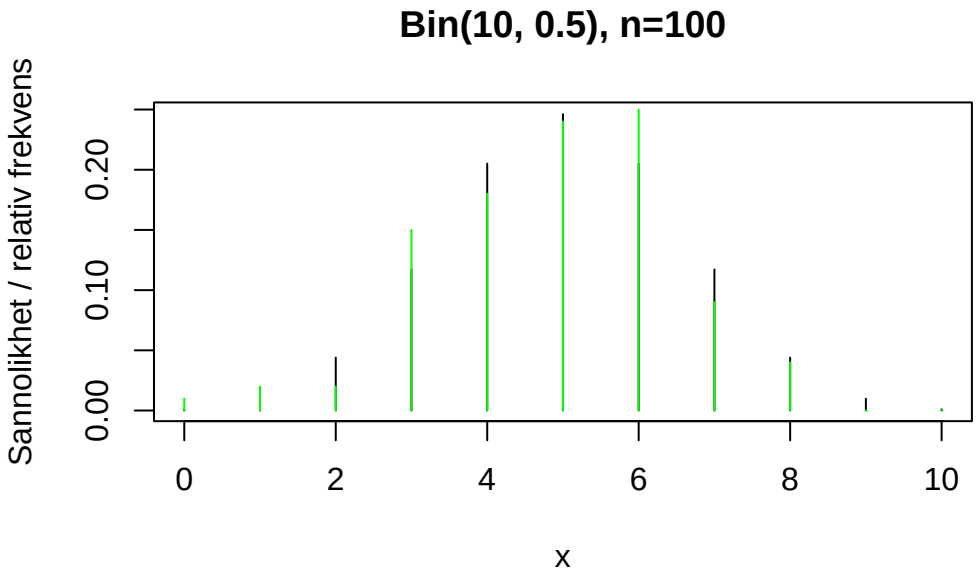

Bin(10, 0.2), n=1000
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0.30
0
2
4
6
8
10
x

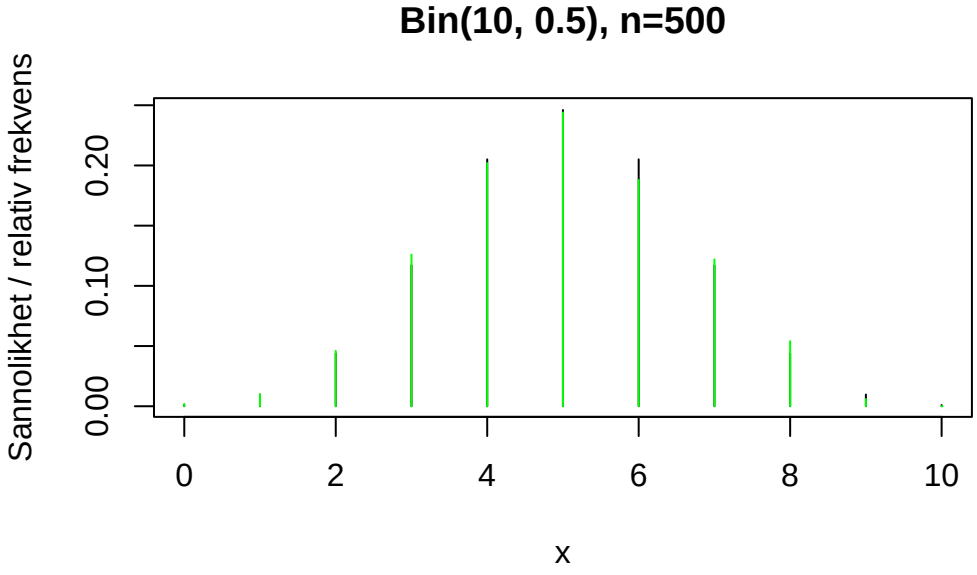


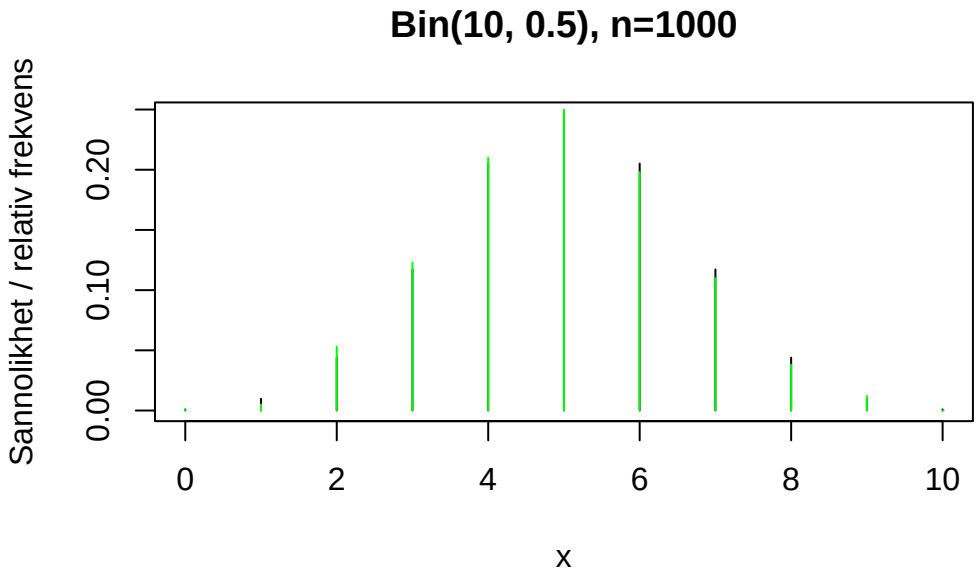

```
# Jämför även Bin(10, 0.5)


for (n_obs in n_list) {
  plot_bino_compare(n_obs, 10, 0.5)
}
```

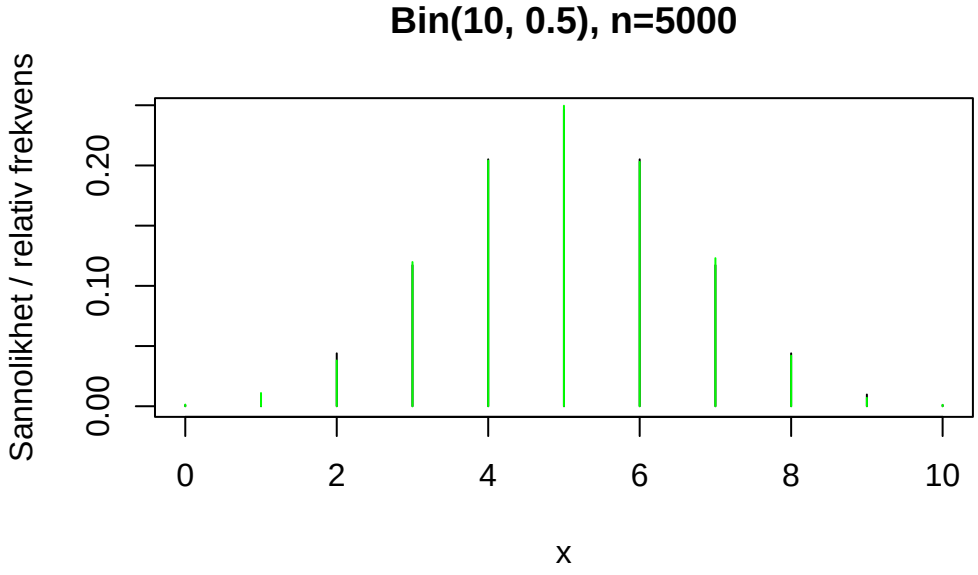

Bin(10, 0.5), n=25
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0
2
4
6
8
10
x


Bin(10, 0.5), n=50
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0
2
4
6
8
10
x


Bin(10, 0.5), n=100
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0
2
4
6
8
10
x

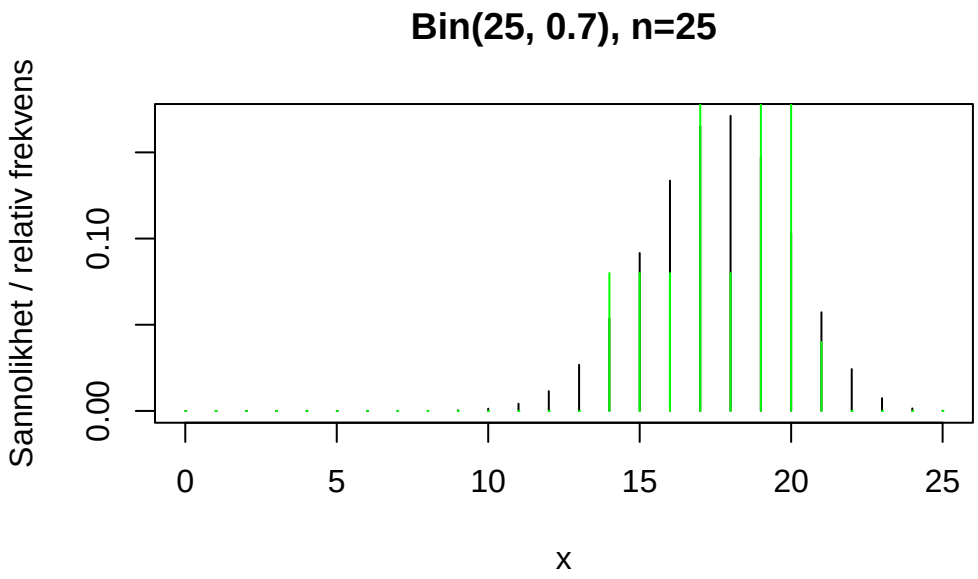

Bin(10, 0.5), n=500
Sannolikhet / relativ frekvens
0.00
0.10
0.20
0
2
4
6
8
10
x

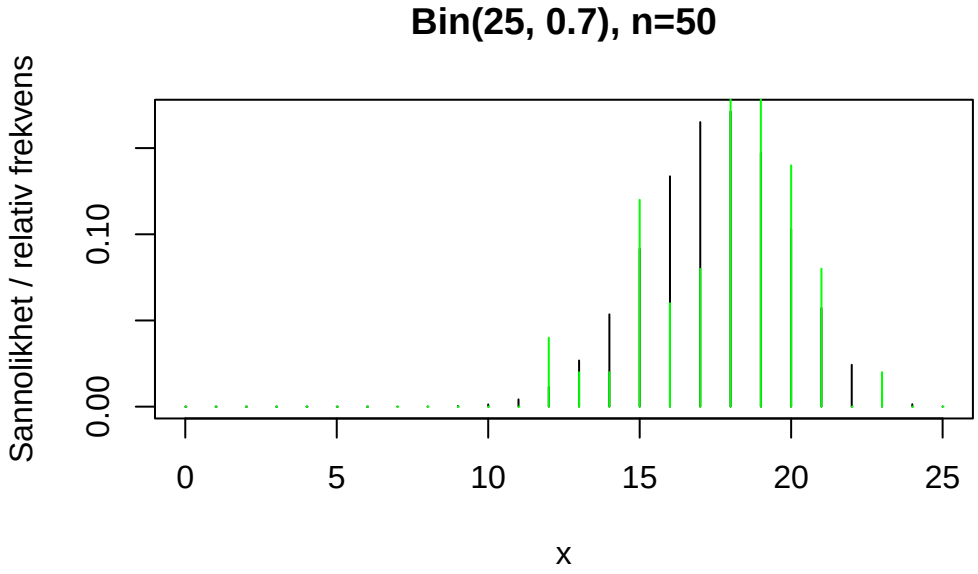


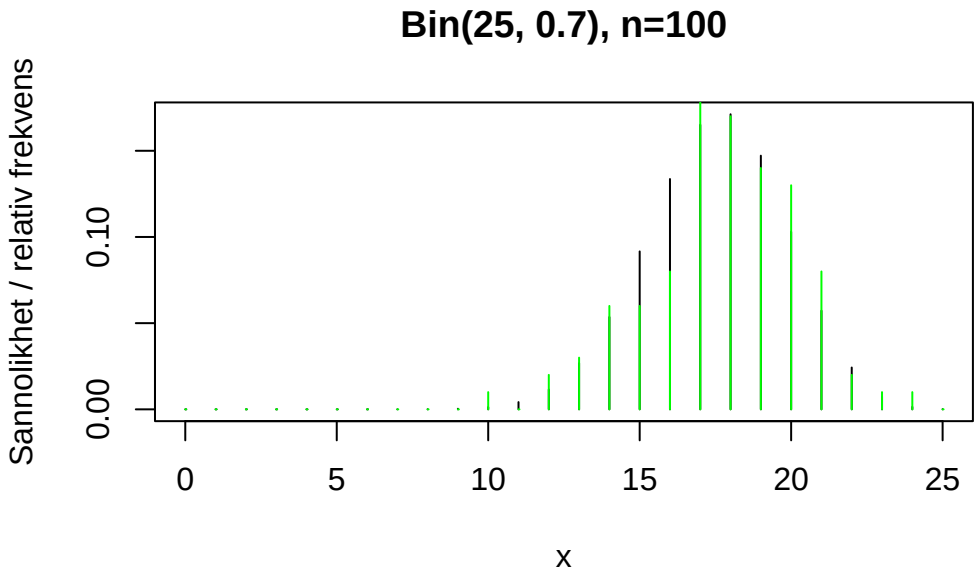


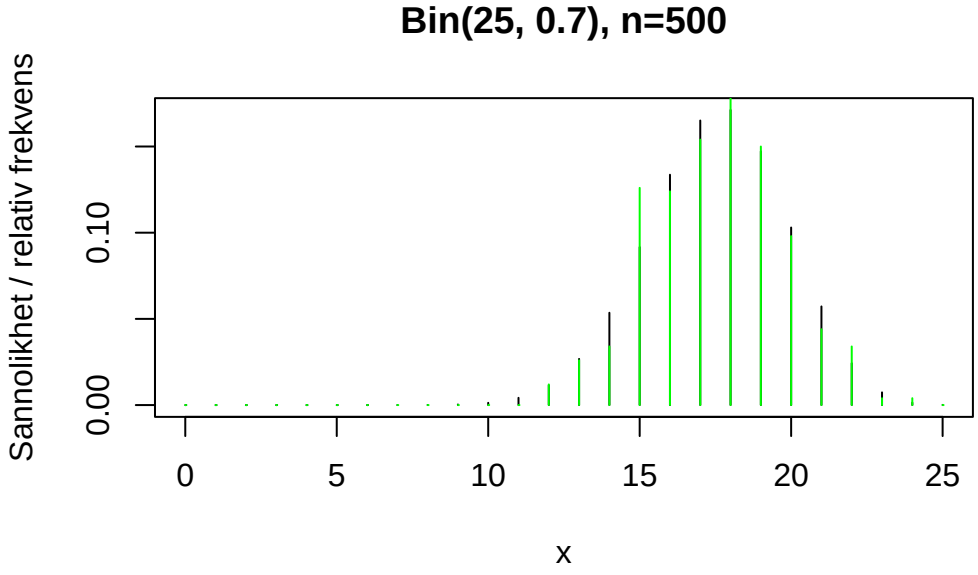

```
# Jämför även Bin(25, 0.7)


for (n_obs in n_list) {
```

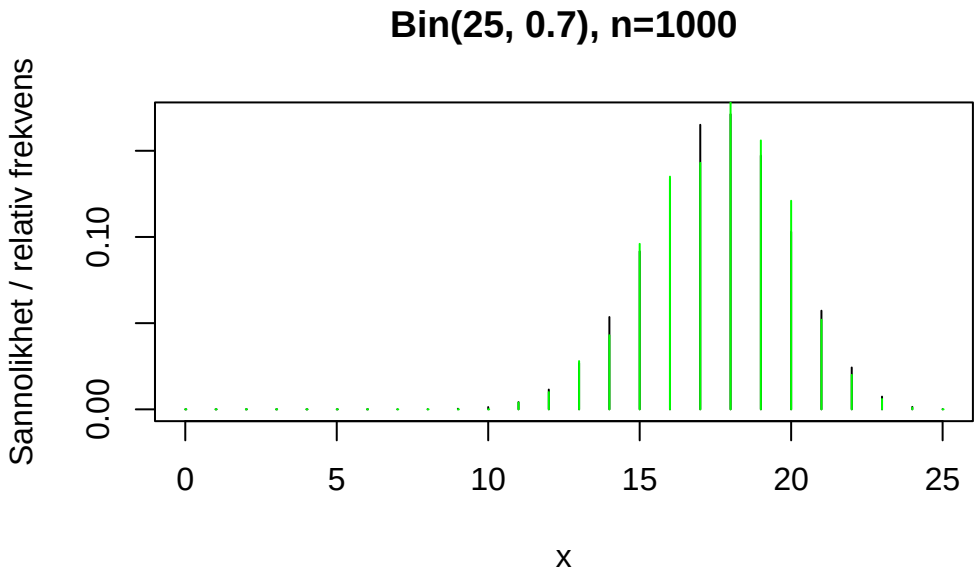
```
  plot_bino_compare(n_obs, 25, 0.7)
}
```

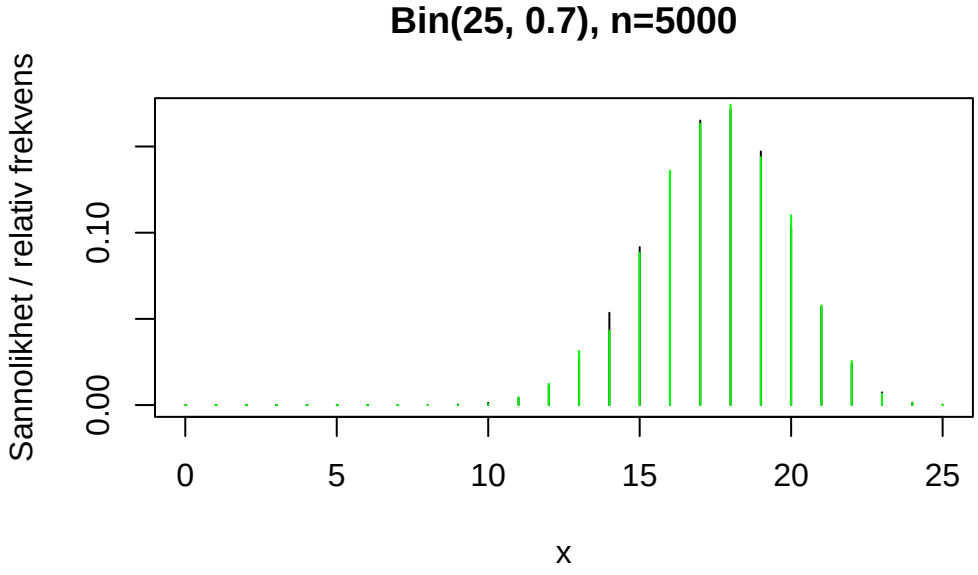





Bin(25, 0.7), n=100
Sannolikhet / relativ frekvens
0.00
0.10
0
5
10
15
20
25
x


Bin(25, 0.7), n=500
Sannolikhet / relativ frekvens
0.00
0.10
0
5
10
15
20
25
x






```
# Kontroll: vilka utfall kom med i stickprovet?


x <- rbinom(25, 10, 0.2)
```

```r
sort(unique(x))
```

```
[1] 0 1 2 4 6
```

```r
x <- rbinom(25, 10, 0.5)
sort(unique(x))
```

```
[1] 3 4 5 6 7 9
```

```r
x <- rbinom(25, 25, 0.7)
sort(unique(x))
```

```
[1] 13 14 15 16 17 18 19 20 21
```

Vid den inledande simuleringen med 25 observationer från en binomialfördelning med parametrarna n = 10 och p = 0.2 framgick tydligt att den skattade sannolikhetsfunktionen avvek relativt kraftigt från den teoretiska fördelningen. I stolpdiagrammet saknades vissa möjliga utfall helt, medan andra var över- eller underrepresenterade. Detta illustrerar hur ett litet stickprov ger en mycket osäker skattning av den underliggande sannolikhetsfunktionen och hur slumpvariationer kan dominera resultatet.

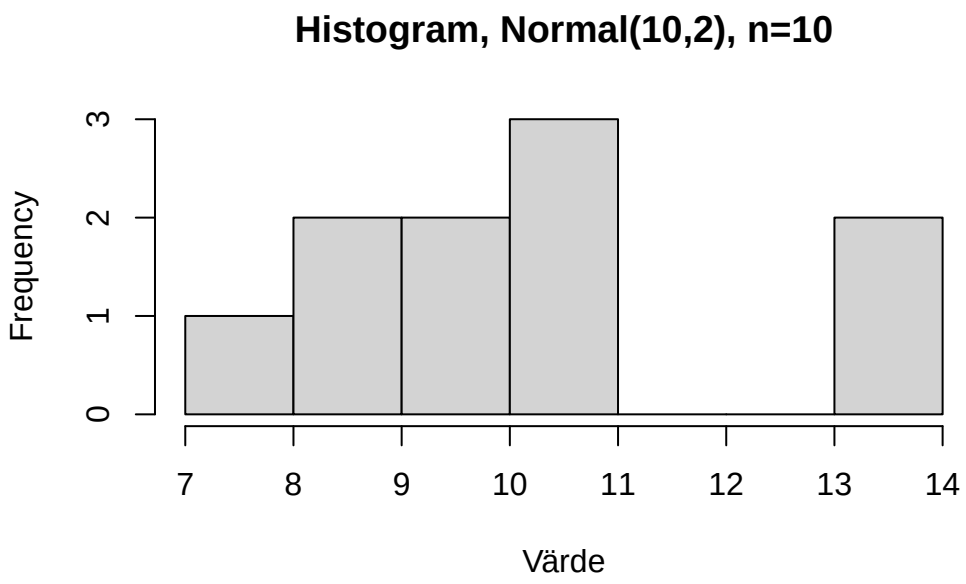
När stickprovsstorleken successivt ökades till 50, 100, 500, 1 000 och 5 000 observationer framträdde en allt tydligare överensstämmelse mellan den skattade och den teoretiska fördelningen för Bin(10, 0,2). De empiriska staplarna började nu följa den oranga teoretiska kurvan på ett mer systematiskt sätt, och variationerna kring de förväntade sannolikheterna minskade successivt. Särskilt vid de största stickproven kunde fördelningens karakteristiska form urskiljas tydligt, vilket ger ett konkret exempel på hur frekvensskattningar konvergerar mot de sanna sannolikheterna.

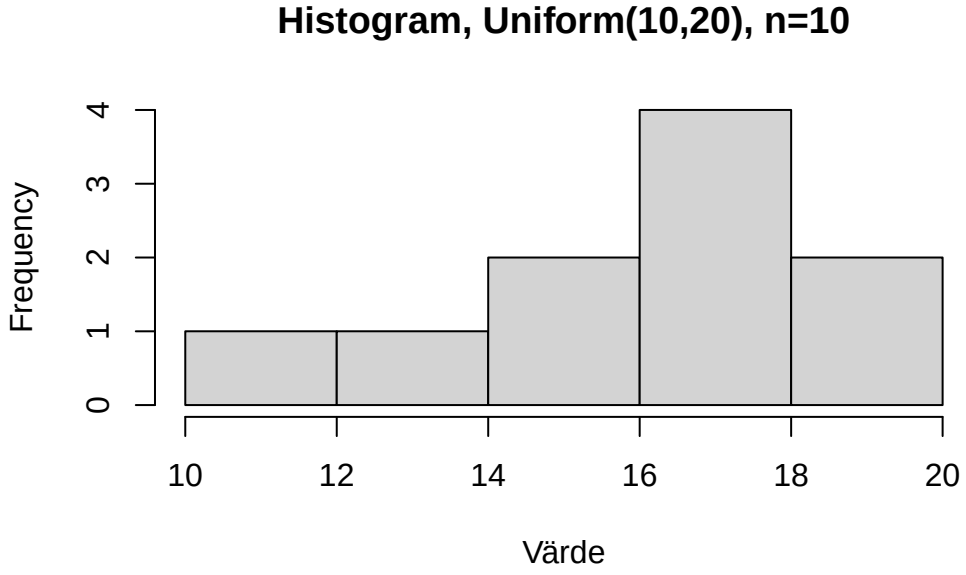
Motsvarande analys för Bin(10, 0.5) visade ett liknande mönster, men med en mer symmetrisk fördelning kring medelvärdet. Redan vid måttliga stickprovsstorlekar uppvisade den skattade fördelningen en relativt god överensstämmelse med teorin, vilket kan förklaras av att sannolikhetsmassan i detta fall är mer jämnt fördelad över utfallen. Fördelningen Bin(25, 0.7) uppvisade däremot en tydligare skevhet, och här krävdes generellt större stickprov för att den empiriska fördelningen skulle stabiliseras kring den teoretiska.

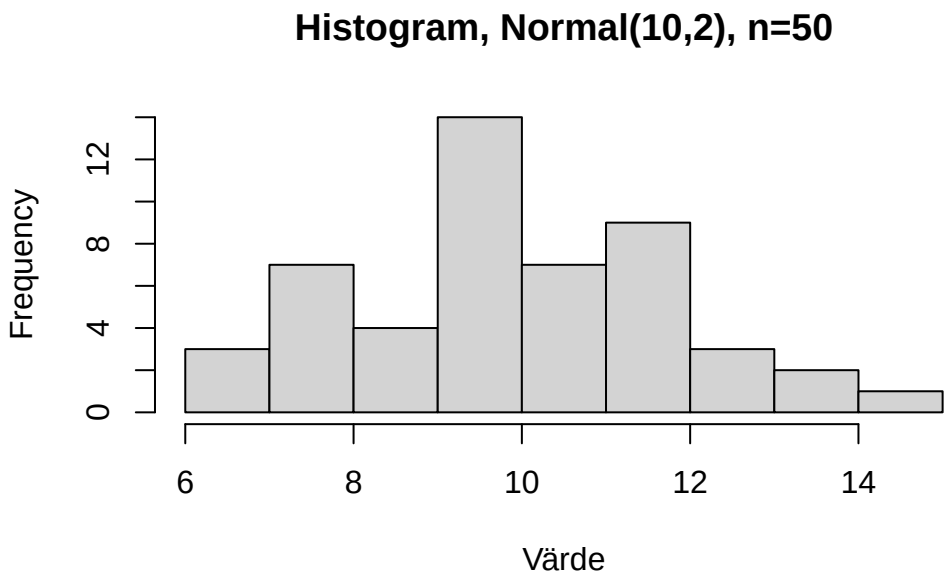
Genom att undersöka vilka utfall som faktiskt förekom i små stickprov kunde vi även konstatera att flera möjliga värden ofta saknas helt vid låga observationstal. Detta understryker ytterligare den osäkerhet som är förknippad med små stickprov och visar att frånvaron av ett visst utfall inte nödvändigtvis innebär att dess sannolikhet är låg, utan snarare kan bero på slumpvariation.

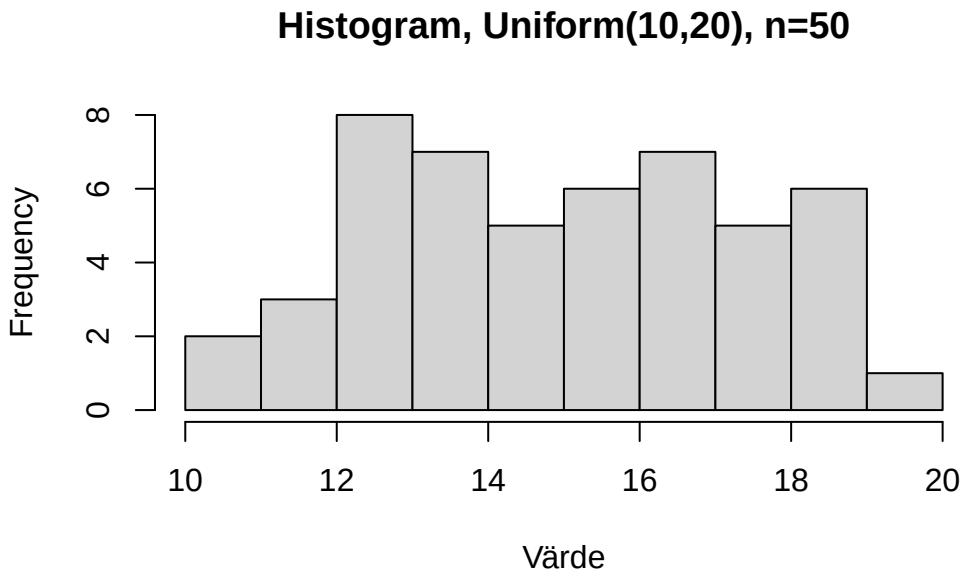
Del 3 - Normalfördelning

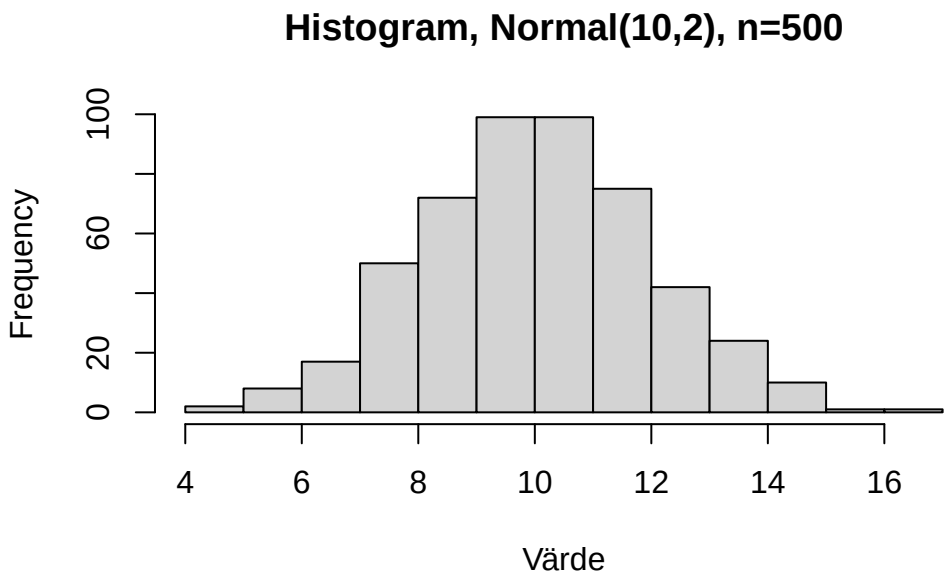
```r
set.seed(123)


# Uppgift 1: n = 10, 50, 500
# Normal(10, 2) och Uniform(10, 20)


n_list <- c(10, 50, 500)

for (n in n_list) {
  normal <- rnorm(n, 10, 2)
  uni <- runif(n, 10, 20)

  hist(normal,
       main = paste0("Histogram, Normal(10,2), n=", n),
       xlab = "Värde")

  hist(uni,
       main = paste0("Histogram, Uniform(10,20), n=", n),
       xlab = "Värde")
}
```

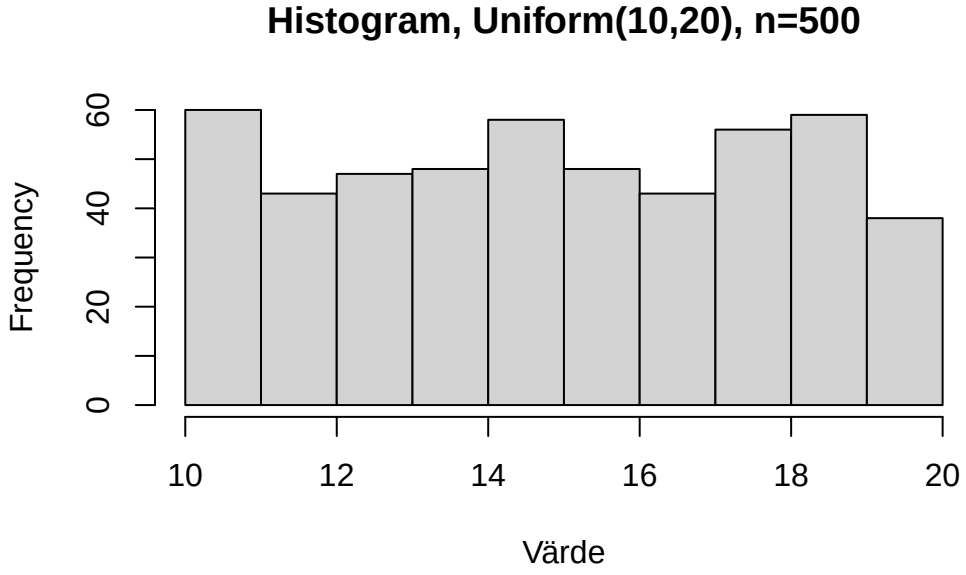

Histogram, Normal(10,2), n=10
Frequency
0
1
2
3
7
8
9
10
11
12
13
14
Värde


Histogram, Uniform(10,20), n=10
Frequency
0
1
2
3
4
10
12
14
16
18
20
Värde


Histogram, Normal(10,2), n=50
Frequency
0
4
8
12
6
8
10
12
14
Värde

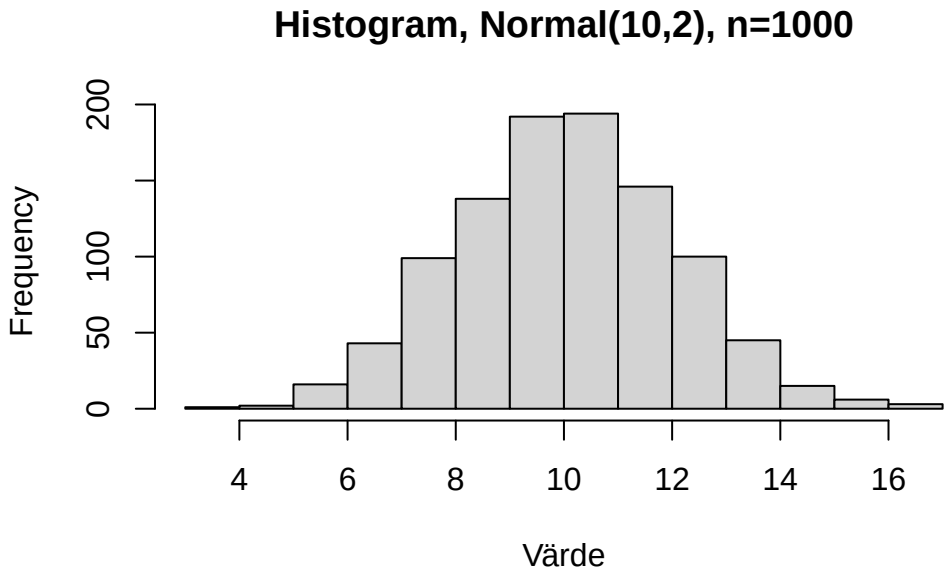

Histogram, Uniform(10,20), n=50
Frequency
0
2
4
6
8
10
12
14
16
18
20
Värde



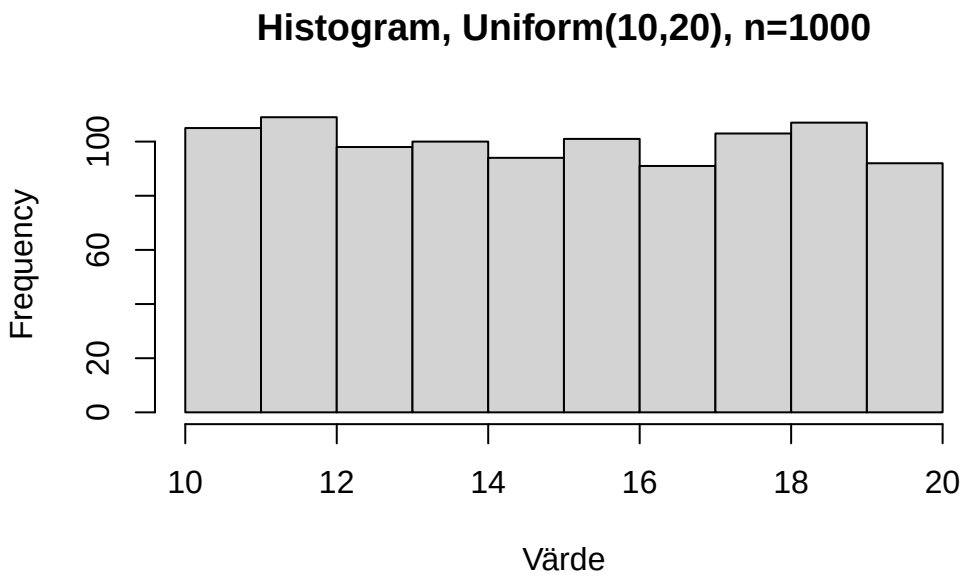



```
# Uppgift 2: n = 1000
# Normal(10, 2), Uniform(10, 20), Exp med väntevärde 1
```

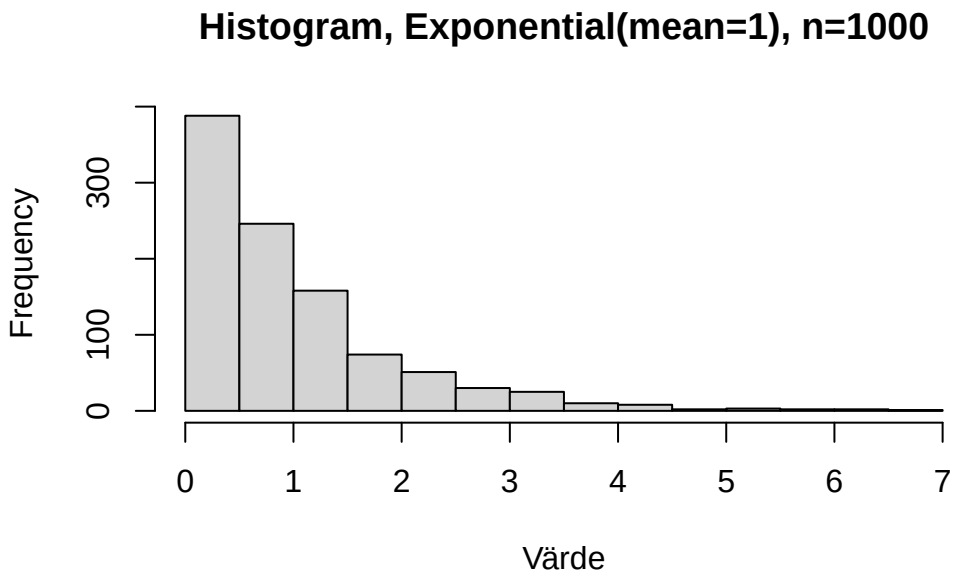
```r
n <- 1000

norm1 <- rnorm(n, 10, 2)
uni1 <- runif(n, 10, 20)
exp1 <- rexp(n, 1)

hist(norm1,
     main = "Histogram, Normal(10,2), n=1000",
     xlab = "Värde")
```




```r
hist(uni1,
     main = "Histogram, Uniform(10,20), n=1000",
     xlab = "Värde")
```




```
hist(exp1,
     main = "Histogram, Exponential(mean=1), n=1000",
     xlab = "Värde")
```



Vid de inledande histogrammen för stickprovsstorlekarna $n = 10$, $n = 50$ och $n = 500$ framgick tydligt hur stickprovsstorleken påverkar möjligheten att urskilja den underliggande fördelningen. För normalfördelningen med parametrarna $\mu = 10$ och $\sigma = 2$ uppvisade histogrammet för $n = 10$ en mycket oregelbunden form, där enstaka observationer fick stort genomslag. Fördelningens karakteristiska klockform var i detta fall svår att identifiera. Vid $n = 50$ började däremot en tydligare symmetri kring medelvärdet framträda, och vid $n = 500$ kunde normalfördelningens typiska form observeras relativt tydligt.

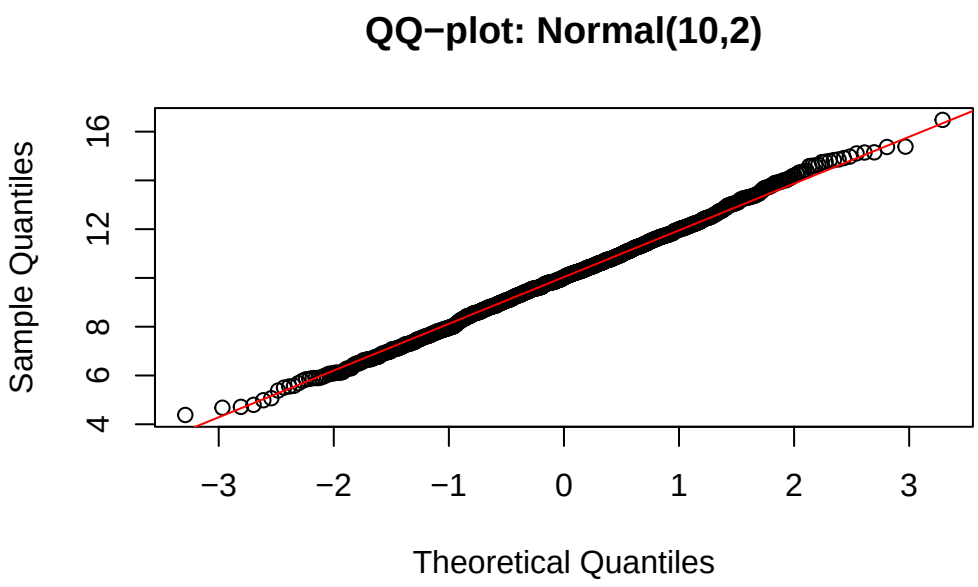
Motsvarande utveckling kunde även observeras för den likformiga fördelningen på intervallet $[10, 20]$. Vid små stickprov var histogrammen ojämna och fragmenterade, medan de vid större stickprov successivt närmade sig en mer jämn och rektangulär form. Detta illustrerar hur även enkla fördelningar kan vara svåra att identifiera empiriskt vid begränsade datamängder.

När stickprovsstorleken ökades till $n = 1000$ blev överensstämmelsen mellan empiriska histogram och teoretiska fördelningar tydlig för samtliga tre fördelningstyper. Normalfördelningen uppvisade då en välutvecklad symmetrisk form, den likformiga fördelningen framstod som jämnt fördelad över intervallet, och exponentialfördelningen visade sin karakteristiska högerskevhet med en tydlig koncentration av observationer nära noll. I synnerhet för exponentialfördelningen blev det vid detta stickprov möjligt att urskilja den snabbt avtagande tätheten.

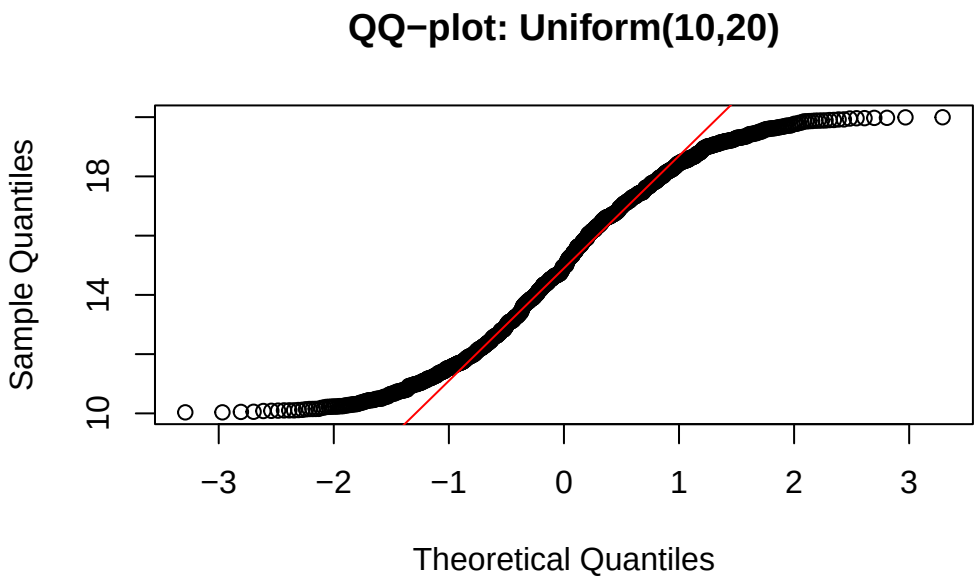
Del 4 - Empirisk fördelningsfunktion och QQ-plot

```
set.seed(123)


# Skapa stickprov (n = 1000)


n <- 1000

norm1 <- rnorm(n, 10, 2)
uni1 <- runif(n, 10, 20)
exp1 <- rexp(n, 1)


# QQ-plot för normalfördelning


qqnorm(norm1,
```

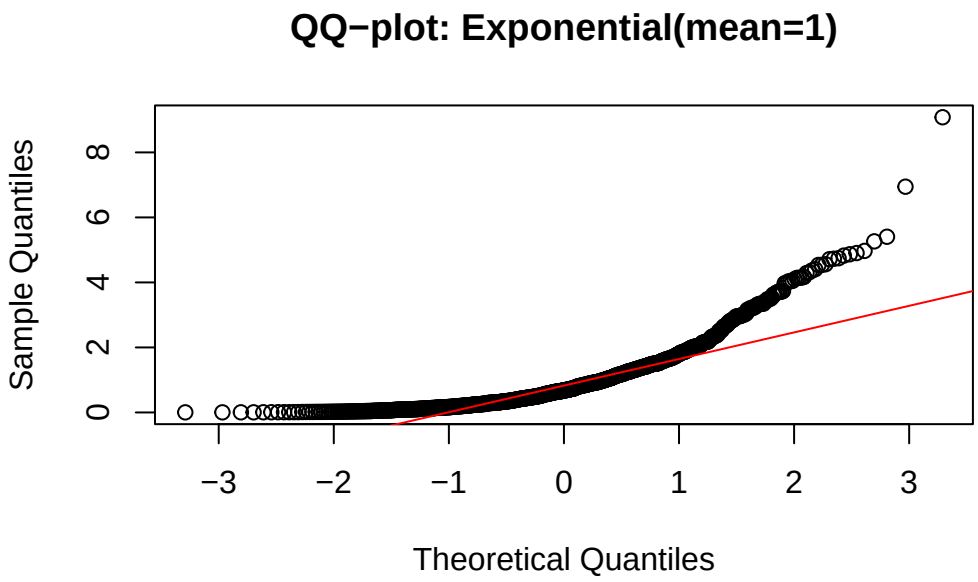
```
       main = "QQ-plot: Normal(10,2)")
qqline(norm1, col = "red")
```



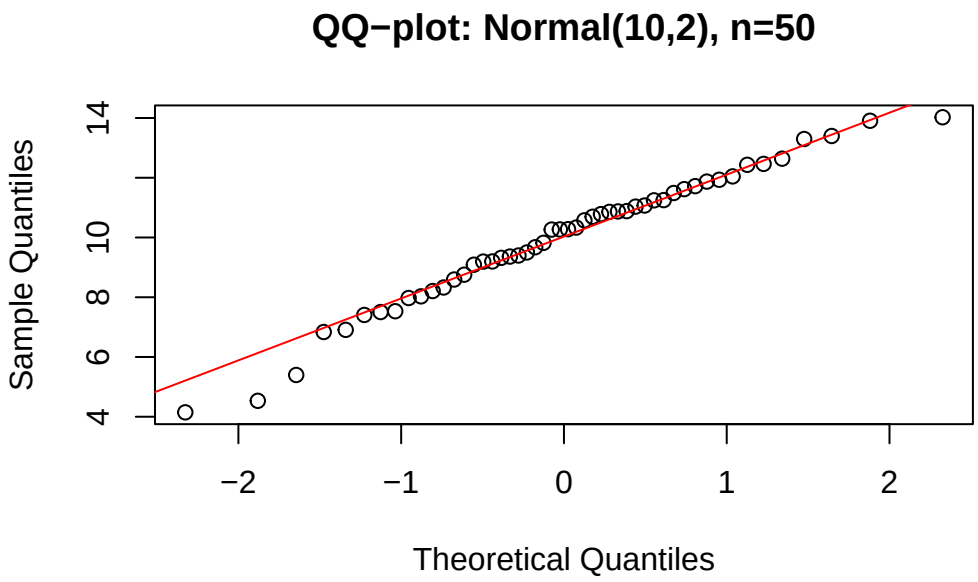

```
# QQ-plot för likformig fördelning


qqnorm(uni1,
       main = "QQ-plot: Uniform(10,20)")
qqline(uni1, col = "red")
```



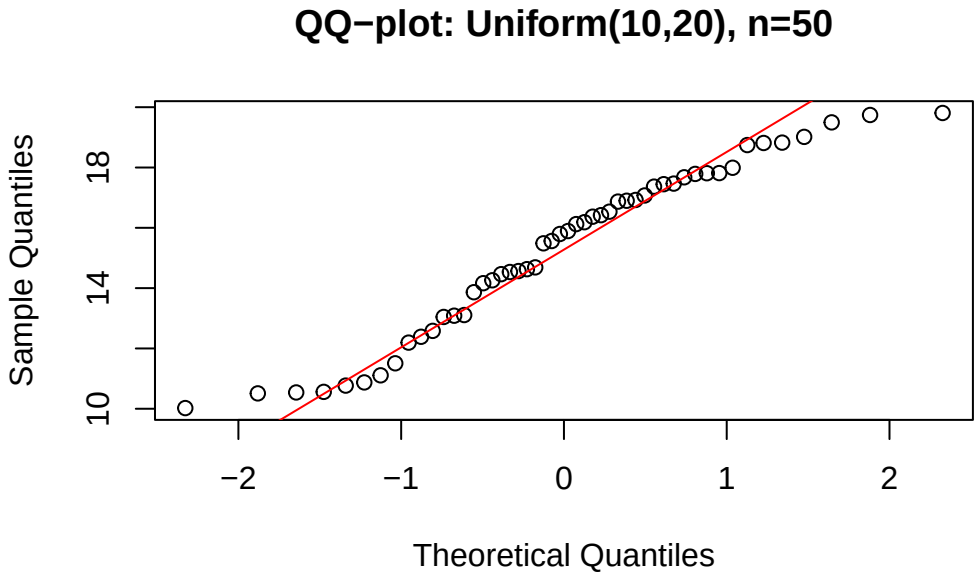

```r
# QQ-plot för exponentialfördelning


qqnorm(exp1,
       main = "QQ-plot: Exponential(mean=1)")
qqline(exp1, col = "red")
```



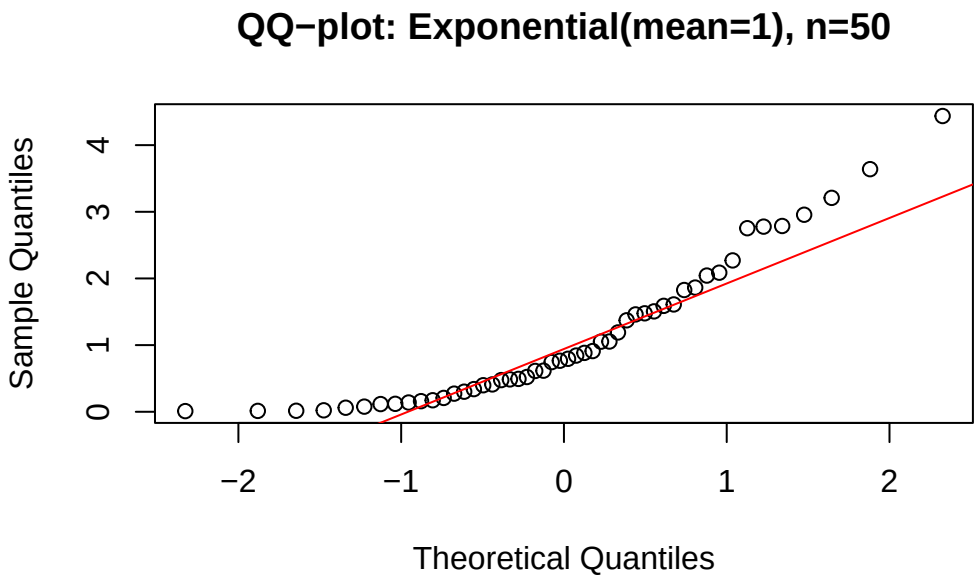

```r
# Jämförelse med mindre stickprov


n_small <- 50

norm_small <- rnorm(n_small, 10, 2)
uni_small <- runif(n_small, 10, 20)
exp_small <- rexp(n_small, 1)

qqnorm(norm_small,
       main = "QQ-plot: Normal(10,2), n=50")
qqline(norm_small, col = "red")
```




```
qqnorm(uni_small,
       main = "QQ-plot: Uniform(10,20), n=50")
qqline(uni_small, col = "red")
```



```r
qqnorm(exp_small,
       main = "QQ-plot: Exponential(mean=1), n=50")
qqline(exp_small, col = "red")
```



Vid analysen av QQ-plots för stickprov med storlek $n = 1000$ framgick tydligt hur väl olika fördelningar överensstämmer med normalfördelningen. För det normalfördelade stickprovet med parametrarna $\mu = 10$ och $\sigma = 2$ låg observationerna till största delen nära referenslinjen. Endast mindre avvikelser kunde observeras i ytterkanterna, vilket indikerar att data i hög grad följer en normalfördelning. Detta bekräftar att QQ-plots är ett effektivt verktyg för att visuellt bedöma normalitet.

För den likformiga fördelningen på intervallet $[10, 20]$ uppvisade QQ-ploten däremot en tydlig S-formad struktur, där observationerna avvek systematiskt från referenslinjen. Särskilt i ändpunkterna framträdde markanta avvikelser. Detta mönster visar att den likformiga fördelningen skiljer sig fundamentalt från normalfördelningen och att dessa skillnader enkelt kan identifieras med hjälp av QQ-plots.

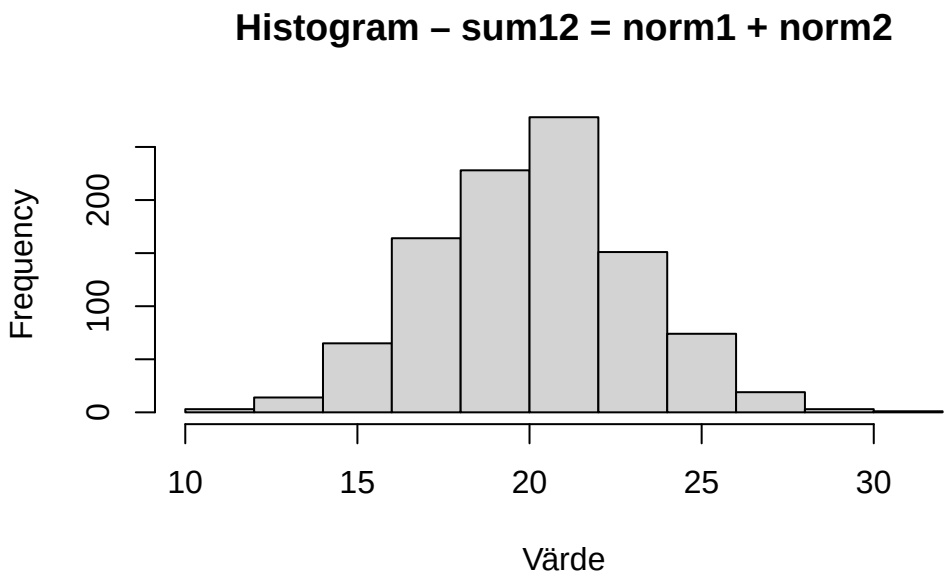
QQ-ploten för den exponentialfördelade datan uppvisade ännu tydligare avvikelser från normalfördelningen. Observationerna låg kraftigt böjda i förhållande till referenslinjen, vilket speglar fördelningens tydliga högerskevhet. Detta indikerar att data inte kan approximeras väl med en normalfördelning och att normalantagandet i detta fall vore olämpligt.

Vid jämförelsen med mindre stickprov, $n = 50$, blev dessutom slumpvariationens inverkan tydlig. För det normalfördelade stickprovet var spridningen kring referenslinjen större än för

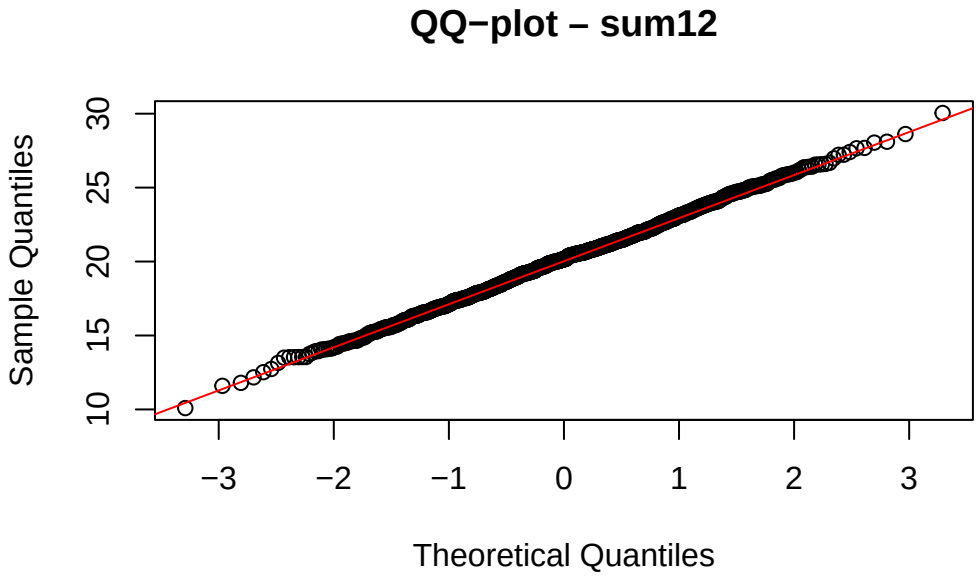
$n = 1000$, och enskilda observationer fick större genomslag. Motsvarande mönster observerades även för de likformiga och exponentialfördelade stickproven, där avvikelserna framstod som mer oregelbundna. Detta visar att tolkningen av QQ-plots blir mer osäker vid små stickprov.

Del 5 - Summa av två normala

```r
set.seed(123)

# Skapa två stickprov (1000 obs vardera) från N(10, 2)


n <- 1000
mu <- 10
sigma <- 2

norm1 <- rnorm(n, mu, sigma)
norm2 <- rnorm(n, mu, sigma)


# Bilda summan


sum12 <- norm1 + norm2


# Undersök fördelningen: histogram och QQ-plot


hist(sum12,
     main = "Histogram - sum12 = norm1 + norm2",
     xlab = "Värde")
```




```
qqnorm(sum12,
       main = "QQ-plot - sum12")
qqline(sum12, col = "red")
```



```r
# Teoretiska värden


mu_theory <- mu + mu
sd_theory <- sqrt(sigma^2 + sigma^2)

mu_theory
```

```
[1] 20
```

```r
sd_theory
```

```
[1] 2.828427
```

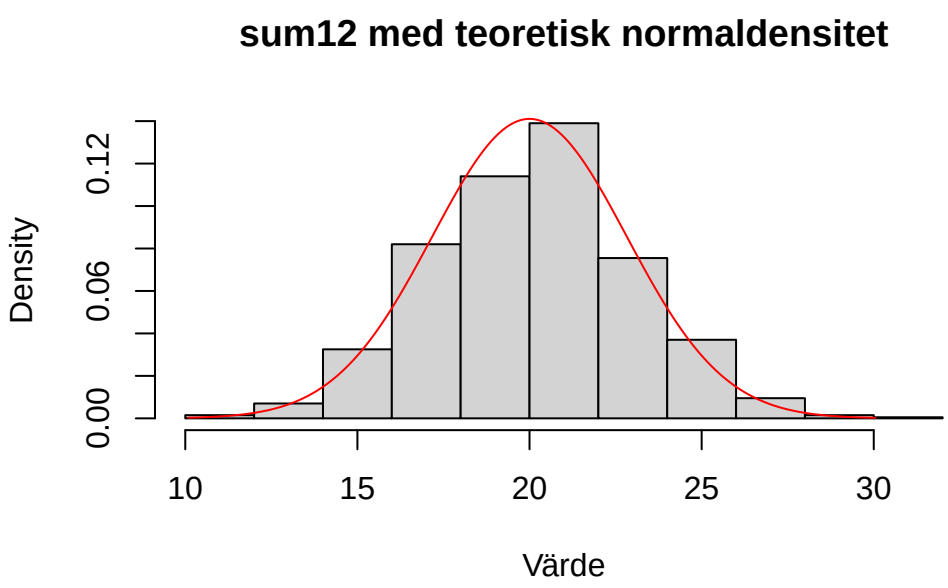
```r
# Stickprovsmedelvärde och stickprovs-standardavvikelse


mean(sum12)
```

```
[1] 20.11719
```

```r
sd(sum12)
```

```
[1] 2.95031
```

```r
# Jämför med en teoretisk normalfördelning i samma figur (valfritt)


hist(sum12, freq = FALSE,
     main = "sum12 med teoretisk normaldensitet",
     xlab = "Värde")

xgrid <- seq(min(sum12), max(sum12), length.out = 500)
lines(xgrid, dnorm(xgrid, mean = mu_theory, sd = sd_theory), col = "red")
```



I denna del bildade vi summan \$ `sum12` = `norm1` + `norm2`\$, där båda stickproven består av $n = 1000$ observationer från en normalfördelning med parametrarna $\mu = 10$ och $\sigma = 2$. Redan när vi studerade histogrammet för `sum12` framträdde en tydligt symmetrisk och ”klockformad” fördelning, centrerad kring ungefär 20. Den visuella bilden är i sig en stark indikation på att summan är normalfördelad, vilket också är precis vad teorin förutsäger: om $X \sim N(\mu_1, \sigma_1^2)$ och $Y \sim N(\mu_2, \sigma_2^2)$ är oberoende, så gäller att $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. I vårt fall innebär detta att `sum12` bör följa $N(20, 2^2 + 2^2)$, det vill säga en normalfördelning med väntevärde $E[\texttt{sum12}] = 20$ och standardavvikelse $\sqrt{8} \approx 2.828$.

När vi sedan gick vidare till QQ-ploten för `sum12` blev detta antagande ytterligare förstärkt. Punkterna låg i stort sett längs en rät linje, utan någon systematisk böjning. Eventuella avvikelser fanns främst i ytterkanterna (svansarna), vilket är väntat i ett stickprov: extremvärden uppträder slumpmässigt och kan ge små lokala avvikelser även när grundfördelningen är helt normal. Det centrala mönstret i QQ-ploten var emellertid mycket tydligt linjärt, vilket stödjer slutsatsen att summan i praktiken beter sig som en normalfördelad variabel.

vi beräknade därefter de teoretiska värdena explicit. Förväntat väntevärde är $\mu_{\text{theory}} = 10 + 10 = 20$, och förväntad standardavvikelse är $\sigma_{\text{theory}} = \sqrt{2^2 + 2^2} = \sqrt{8} \approx 2.828$. Dessa teoretiska värden kan direkt jämföras med de empiriska stickprovsstorheterna: stickprovsmedelvärdet blev $\overline{x} = 20.11719$ och stickprovsstandardavvikelsen blev $s = 2.95031$. Medelvärdet ligger mycket nära 20, vilket är precis vad man bör förvänta sig vid $n = 1000$: avvikelsen 0.117 är liten relativt skalan på observationerna och helt förenlig med slumpvariation. Standardavvikelsen ligger däremot något högre än det teoretiska värdet. Även detta är inte oväntat; stickprovsstandardavvikelsen är en slumpvariabel och kan i en given realisation

bli både något större eller något mindre än σ_{theory}. En rimlig tolkning är alltså att både $\overline{x}$ och s ligger inom vad man kan betrakta som normal, stickprovsbetingad variation, och att de i stort överensstämmer med teorin.

För att ytterligare konkretisera jämförelsen lade vi även in en teoretisk normaldensitet ovanpå histogrammet. Detta var särskilt informativt: kurvans centrum och huvudsakliga form överensstämde väl med histogrammets ”massa”, vilket förstärker slutsatsen att summan beter sig normalfördelat och att parametrarna är rimliga. Om man vill vara petig kan man notera att histogrammets form påverkas av val av klassindelning (antal och bredd på bins), men trots detta framstår matchningen som robust, vilket tyder på att resultatet inte är en artefakt av plottningen utan en faktisk egenskap hos datan.

Del 6 - Fördelningsanpassning (5.1 + 5.2)

```r
set.seed(123)


# 5.1 Empiriska fördelningsfunktionen F*(x)
# Exempel: 100 obs från N(2,1)


n <- 100
X <- rnorm(n, 2, 1)

sortX <- sort(X)

# Plotposition enligt instruktionen: i/(n+1)
Fn <- seq_len(length(sortX)) / (length(sortX) + 1)

plot(sortX, Fn, type = "s",
     main = "Empirisk fördelningsfunktion F*(x)",
     xlab = "x",
     ylab = "F*(x)")
grid()

# Skatta lägesmått från stickprovet
mean(X)
```

```
[1] 2.090406
```

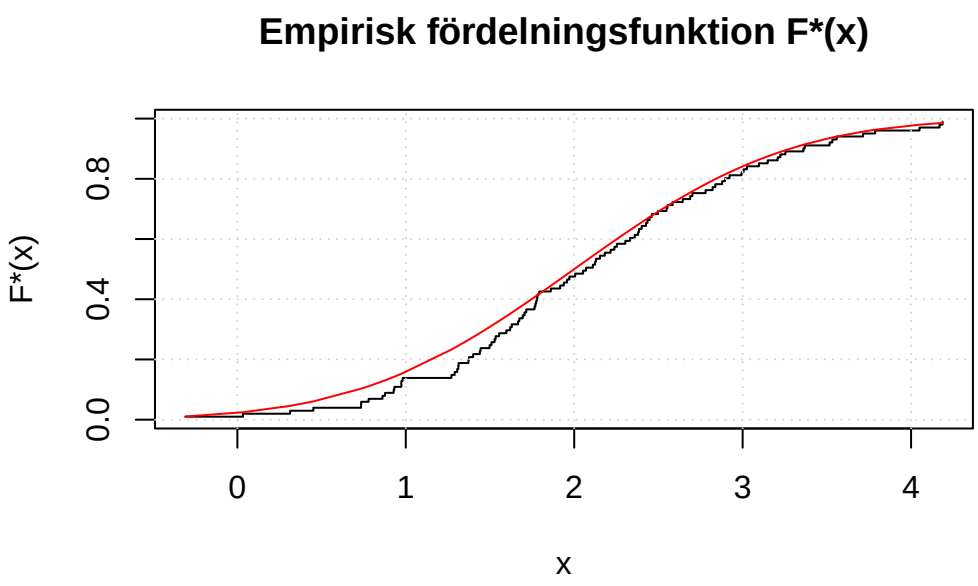
```r
median(X)
```

```
[1] 2.061756
```

```r
quantile(X, probs = c(0.25, 0.75))
```

```
     25%      75%
1.506146 2.691819
```

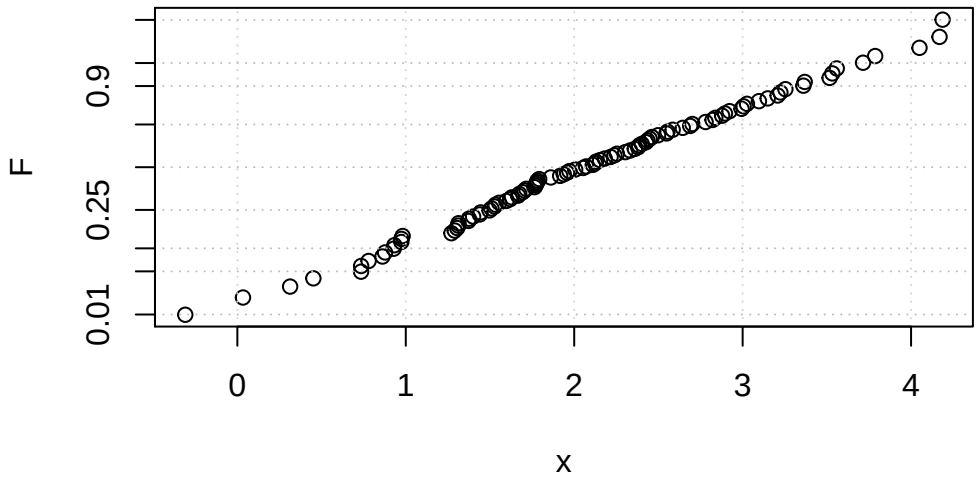
```r
# Lägg till teoretiska fördelningsfunktionen F_X(x) = pnorm(x,2,1)
points(sortX, pnorm(sortX, 2, 1), type = "l", col = "red")
```




```r
# 5.2 Normplot i nsRFA (om paketet finns)


if (!requireNamespace("nsRFA", quietly = TRUE)) {
  install.packages("nsRFA")
}

library(nsRFA)
```

```
Warning: package 'nsRFA' was built under R version 4.5.2
```

```r
# Normalfördelningspapper
normplot(X)
```




```r
# Uppgift 4: skatta mu och sigma från normalplot-idén
# (linjär anpassning: X   mu + sigma * z)


# Teoretiska normal-kvantiler för samma plotpositioner
z <- qnorm(Fn)

fit <- lm(sortX ~ z)

mu_hat <- coef(fit)[1]
sigma_hat <- coef(fit)[2]

mu_hat
```

```
(Intercept)
   2.090406
```

```
sigma_hat
```

```
      z
0.94299
```

```
# Jämför med stickprovsmedel och stickprovs-sd
mean(X)
```

```
[1] 2.090406
```

```
sd(X)
```

```
[1] 0.9128159
```

I denna del studerade vi ett stickprov med storlek $n = 100$ från en normalfördelning med parametrarna $\mu = 2$ och $\sigma = 1$, med fokus på att undersöka hur den empiriska fördelningsfunktionen kan användas för att beskriva och analysera data. Genom att sortera observationerna och beräkna plotpositionerna $F_n(x_i) = \frac{i}{n+1}$ erhölls en skattning av den kumulativa fördelningsfunktionen $F^*(x)$. Den resulterande trappstegsgrafen gav en konkret bild av hur sannolikhetsmassan fördelas över observationsintervallet.

Vid jämförelsen mellan den empiriska fördelningsfunktionen och den teoretiska normalfördelningens fördelningsfunktion $F(x) = \Phi!\left(\frac{x-\mu}{\sigma}\right)$ kunde vi konstatera att kurvorna i huvudsak följde varandra väl. Den röda teoretiska kurvan låg nära den empiriska trappfunktionen över större delen av intervallet, vilket indikerar att stickprovet är representativt för den underliggande normalfördelningen. Samtidigt förekom mindre avvikelser, särskilt i ytterkanterna, vilket är förväntat vid ett relativt begränsat stickprov. Dessa skillnader kan tolkas som ett uttryck för slumpvariation snarare än som systematiska avvikelser från modellen.

De beräknade läges- och spridningsmåtten, såsom stickprovsmedelvärdet $\bar{x} = 2.090406$ och medianen 2.061756, låg nära det teoretiska värdet $\mu = 2$. Även kvartilerna låg på rimliga nivåer i förhållande till fördelningens form. Detta stärker intrycket av att stickprovet i stort sett följer den antagna normalfördelningen, även om vissa variationer är oundvikliga vid $n = 100$.

Vid användningen av normplot (normalfördelningspapper) blev detta mönster ytterligare tydligt. Observationerna placerade sig huvudsakligen längs en approximativt rät linje, vilket indikerar att datan är förenlig med en normalfördelning. Mindre avvikelser kunde noteras, särskilt i ändpunkterna, men dessa var inte tillräckligt systematiska för att ifrågasätta normalantagandet. Normploten framstår därmed som ett effektivt komplement till histogram och QQ-plots vid bedömning av fördelningsform.

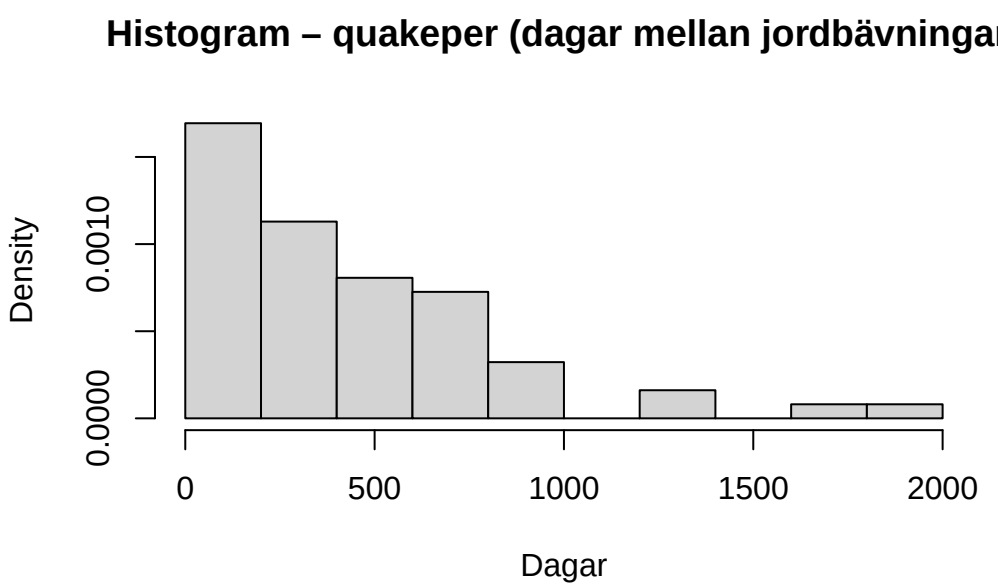
Genom att använda normalplot för att skatta parametrarna μ och σ via en linjär regression mellan sorterade observationer och teoretiska normal-kvantiler erhölls skattningarna $\hat{\mu} = 2.090406$ och $\hat{\sigma} = 0.94299$. Dessa värden överensstämmer väl med stickprovsmedelvärdet respektive stickprovsstandardavvikelsen $s = 0.9128$. Att flera olika metoder ger liknande resultat stärker tilltron till skattningarna och visar hur olika statistiska verktyg kan användas parallellt för att analysera samma data.

Del 7 - Jordbävningar (5.3 + 5.4, Uppgift 5–7)

```
# Läs in data


load("C:/Users/josie/Documents/STAG35/STAG31 Projekt 1/Quakes.RData")

# Kontrollera att quakeper finns
length(quakeper)
```

```
[1] 62
```

```
# Grundläggande beskrivning


hist(quakeper, freq = FALSE,
     main = "Histogram - quakeper (dagar mellan jordbävningar)",
     xlab = "Dagar")
```




```
m <- mean(quakeper)
med <- median(quakeper)
s <- sd(quakeper)
s2 <- var(quakeper)
myrange <- range(quakeper)

m
```

```
[1] 437.2097
```

```
med
```

```
[1] 331.5
```

```
s
```

```
[1] 399.9273
```

```
s2
```

```
         [,1]
[1,] 159941.8
```

```
myrange
```

```
[1]    9 1901
```

```
# Uppgift 5: > 5 år?


five_years <- 5 * 365
max(quakeper)
```

```
[1] 1901
```

```
any(quakeper > five_years)
```

```
[1] TRUE
```

```
sum(quakeper > five_years)
```

```
[1] 1
```

```
mean(quakeper > five_years)
```

```
[1] 0.01612903
```

```
# Hur ofta inträffar de? (återkomsttid och intensitet)
Tr <- m
lambda_hat <- 1 / Tr
Tr
```

```
[1] 437.2097
```

```
lambda_hat
```

```
[1] 0.002287232
```

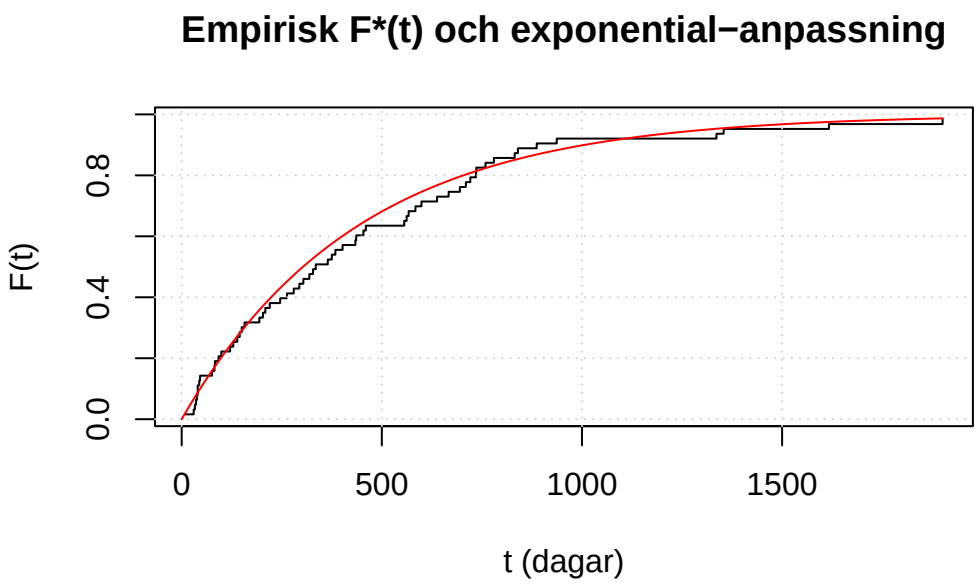
```r
# Uppgift 6: P(T < 1000) och P(T > 200)


less1000 <- quakeper[quakeper < 1000]
p_less1000 <- length(less1000) / length(quakeper)
p_less1000
```

```
[1] 0.9354839
```

```r
p_more200 <- mean(quakeper > 200)
p_more200
```

```
[1] 0.6612903
```

```r
# Uppgift 7: Empirisk F*(t) vs teoretisk exponential
# Teoretisk: F_T(t) = 1 - exp(-t/m)


sortX <- sort(quakeper)
Fn <- seq_len(length(sortX)) / (length(sortX) + 1)

plot(sortX, Fn, type = "s",
     main = "Empirisk F*(t) och exponential-anpassning",
     xlab = "t (dagar)",
     ylab = "F(t)")
grid()

taxis <- seq(0.01, max(quakeper), length.out = 1000)
FT <- 1 - exp(-taxis / m)

points(taxis, FT, type = "l", col = "red")
```



I denna del analyserade vi tidsintervallen mellan större jordbävningar med hjälp av variabeln `quakeper`, som består av $n = 62$ observationer. Histogrammet över datamaterialet visade en tydligt högerskev fördelning, där majoriteten av observationerna är relativt små medan ett fåtal mycket långa intervall förekommer. Detta mönster är typiskt för väntetidsdata och indikerar att långa perioder utan jordbävningar är ovanliga men möjliga.

De deskriptiva måtten gav ytterligare information om fördelningens karaktär. Stickprovsmedelvärdet blev $\bar{x} = 437.21$ dagar, medan medianen var 331.5 dagar. Att medelvärdet är större än medianen bekräftar fördelningens högerskevhet. Standardavvikelsen $s = 399.93$ är dessutom nästan lika stor som medelvärdet, vilket tyder på stor spridning i observationerna. Spannet $[9, 1901]$ visar att väntetiderna varierar kraftigt, från mycket korta intervall till perioder på över fem år.

Vid undersökningen av extremvärden framgick att endast ett intervall överstiger fem år (1825 dagar), vilket motsvarar cirka 1.6% av observationerna. Detta visar att mycket långa perioder utan större jordbävningar är sällsynta men inte omöjliga. Medelvärdet kan därmed tolkas som en ungefärlig återkomsttid, $T_r \approx 437$ dagar, vilket motsvarar drygt ett år. Den skattade intensiteten kan skrivas som

$$\hat{\lambda} = \frac{1}{T_r} \approx 0.0023$$

vilket ger ett mått på hur ofta större jordbävningar inträffar i genomsnitt.

Vid beräkningen av sannolikheter erhölls $P(T < 1000) \approx 0.94$ och $P(T > 200) \approx 0.66$. Detta innebär att det i de flesta fall går mindre än tre år mellan större jordbävningar, men att det samtidigt är relativt vanligt att väntetiden överstiger cirka ett halvår. Dessa resultat ger en kvantitativ bild av hur oregelbundet jordbävningar inträffar och hur stora variationerna är mellan olika perioder.

För att undersöka om en exponentialfördelning är en rimlig modell jämförde vi den empiriska fördelningsfunktionen $F^*(t)$ med den teoretiska exponentialfördelningen

$$F_T(t) = 1 - \exp\left(-\frac{t}{m}\right)$$

där $m = \bar{x}$. Den empiriska trappfunktionen följde den teoretiska kurvan relativt väl över stora delar av intervallet, särskilt för små och medelstora värden på t. Vissa avvikelser kunde dock observeras för större väntetider, där enskilda extremvärden får stort genomslag. Dessa skillnader är inte oväntade med tanke på det begränsade stickprovsantalet.

Den övergripande överensstämmelsen mellan empirisk och teoretisk fördelning tyder ändå på att exponentialmodellen är en rimlig första approximation för dessa data. Detta är förenligt med antagandet om ett minneslöst förlopp, där sannolikheten för en jordbävning inte beror på hur lång tid som gått sedan den förra. Samtidigt bör man vara medveten om att verkliga geologiska processer är mer komplexa än vad en enkel exponentialmodell kan fånga, och att modellen därför endast bör ses som en grov förenkling.

Sammantaget visar denna del hur statistiska modeller kan användas för att analysera och tolka verkliga fenomen. Genom att kombinera deskriptiv statistik, sannolikhetsberäkningar och fördelningsanpassning har vi kunnat ge en sammanhängande bild av hur ofta större jordbävningar inträffar och hur stora variationerna är. Resultaten illustrerar både styrkan och begränsningarna hos enkla sannolikhetsmodeller vid analys av naturdata.

Del 8 - Hodkin-data

```
# Läs in data

load("C:/Users/josie/Documents/STAG35/STAG31 Projekt 1/Hodgkindata.RData")

# Kontrollera att variablerna finns
length(Hodgkin)
```

```
[1] 35
```

```r
length(NonHodgkin)
```

```
[1] 35
```

```r
# Grundläggande beskrivning

summary(Hodgkin)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  171.0   292.5   554.0   758.9   994.5  2415.0 
```

```r
summary(NonHodgkin)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  116.0   187.5   375.0   408.3   496.0  1252.0 
```

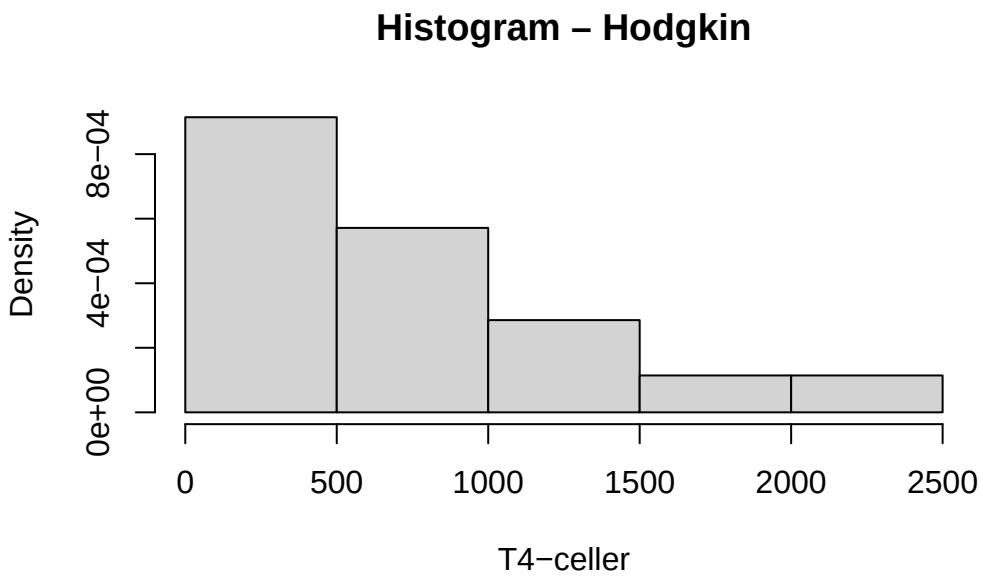
```r
sd(Hodgkin)
```

```
[1] 596.1023
```

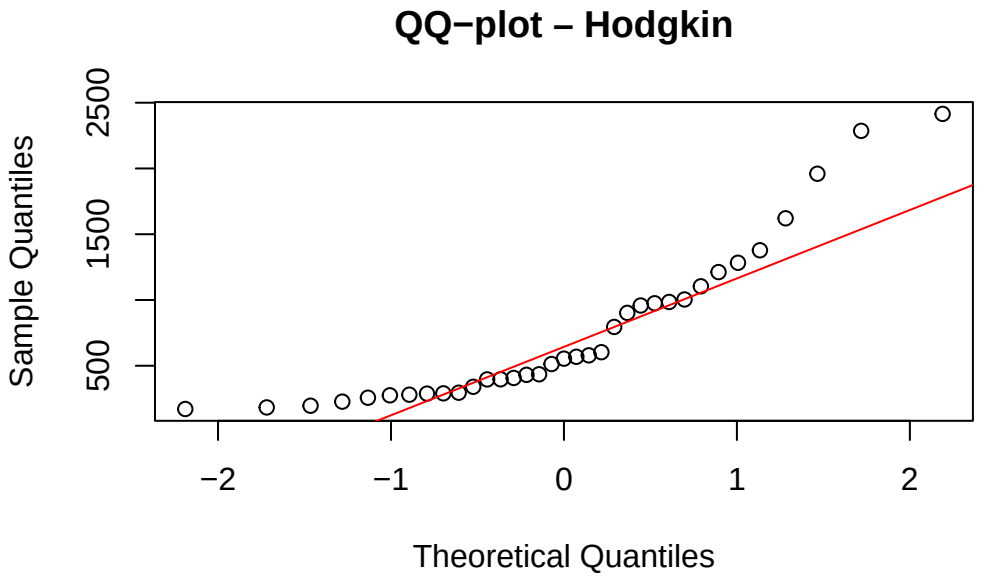
```r
sd(NonHodgkin)
```

```
[1] 268.9818
```

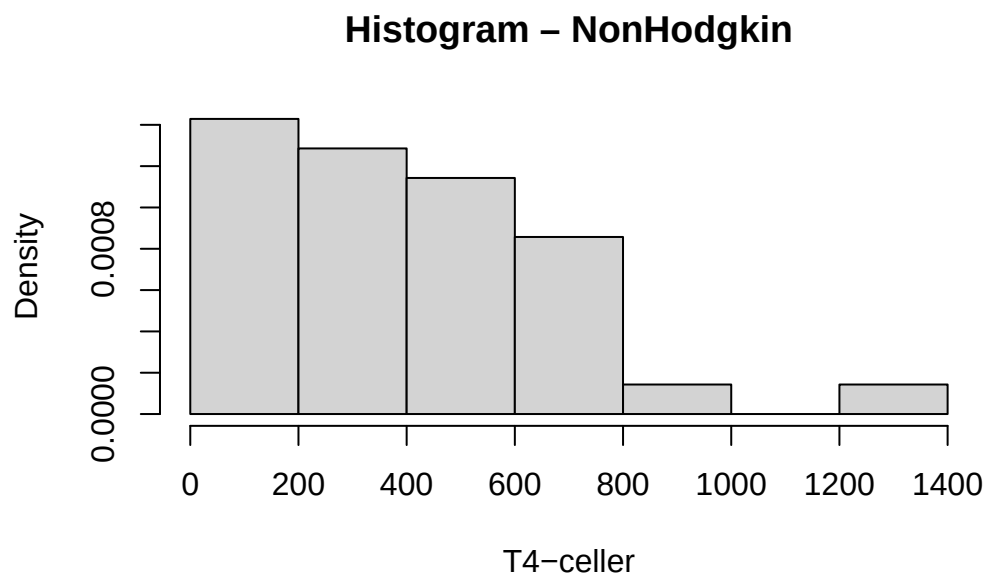
```r
# Histogram + QQ-plot (normalitet) för båda grupperna


hist(Hodgkin, freq = FALSE,
     main = "Histogram - Hodgkin",
     xlab = "T4-celler")
```



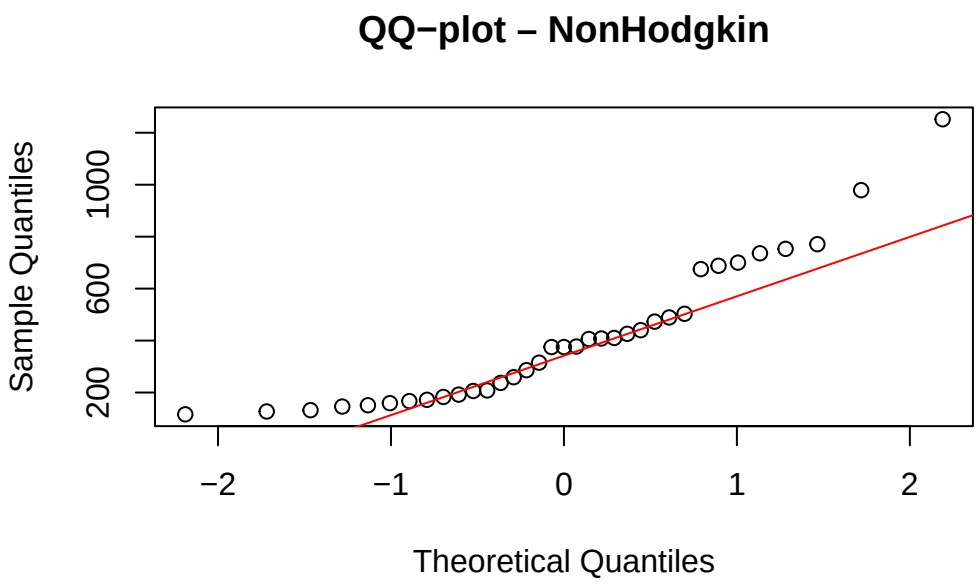

```
qqnorm(Hodgkin, main = "QQ-plot - Hodgkin")
qqline(Hodgkin, col = "red")
```



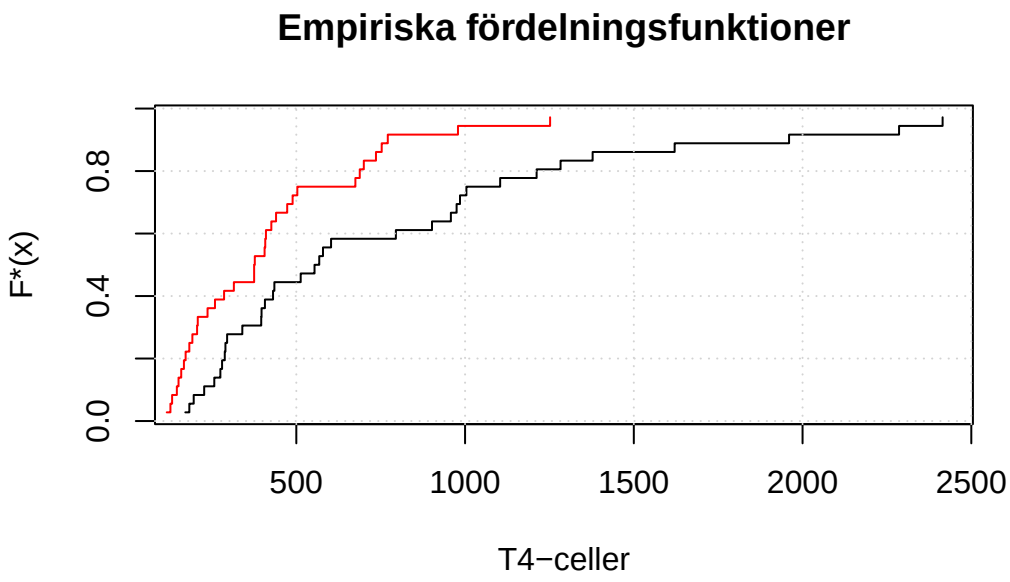
```r
hist(NonHodgkin, freq = FALSE,
     main = "Histogram - NonHodgkin",
     xlab = "T4-celler")
```



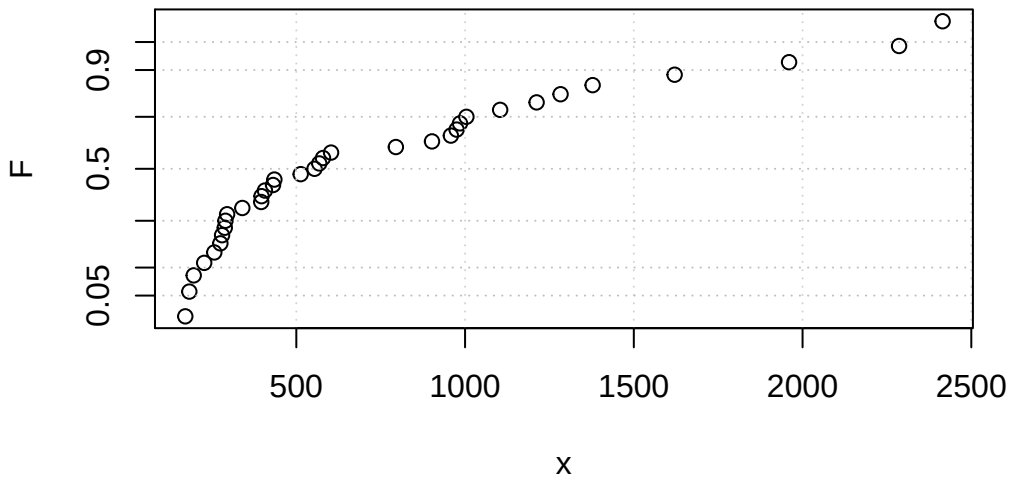

```r
qqnorm(NonHodgkin, main = "QQ-plot - NonHodgkin")
qqline(NonHodgkin, col = "red")
```



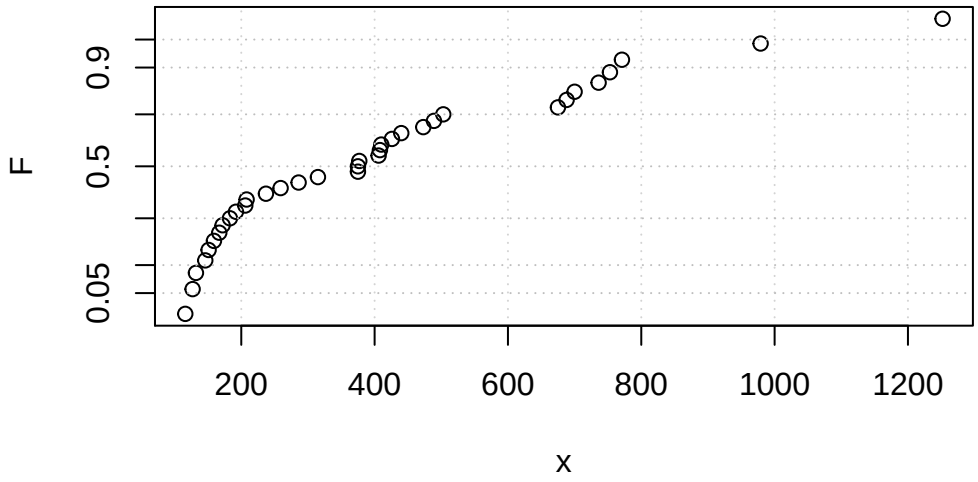

```r
# Empiriska fördelningsfunktioner (ECDF) - jämför grupperna


sortH <- sort(Hodgkin)
FnH <- seq_len(length(sortH)) / (length(sortH) + 1)

sortN <- sort(NonHodgkin)
FnN <- seq_len(length(sortN)) / (length(sortN) + 1)

plot(sortH, FnH, type = "s",
     main = "Empiriska fördelningsfunktioner",
     xlab = "T4-celler",
     ylab = "F*(x)")
grid()
points(sortN, FnN, type = "s", col = "red")
```



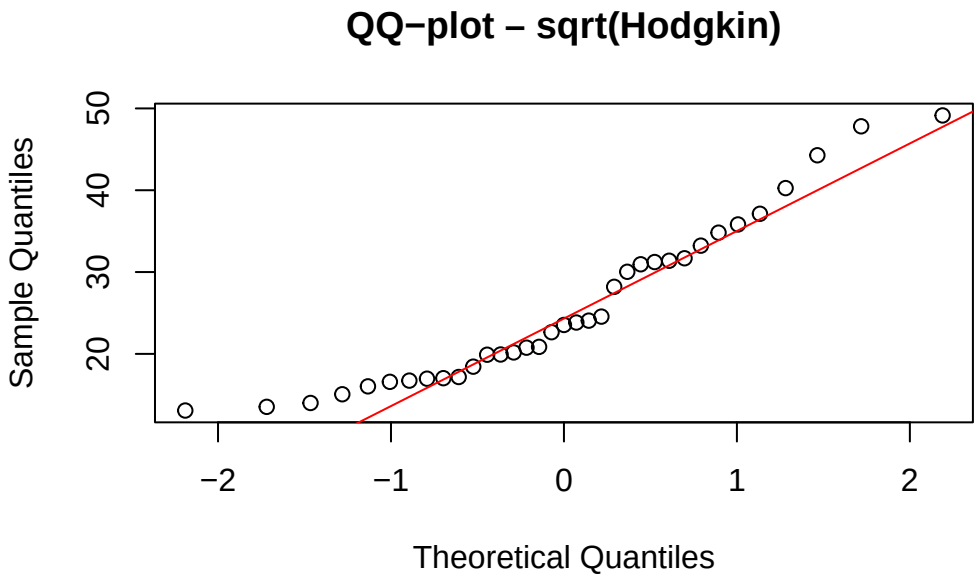

```
# Normplot (normalfördelningspapper) via nsRFA


if (!requireNamespace("nsRFA", quietly = TRUE)) {
  install.packages("nsRFA")
}
library(nsRFA)

normplot(Hodgkin)
```



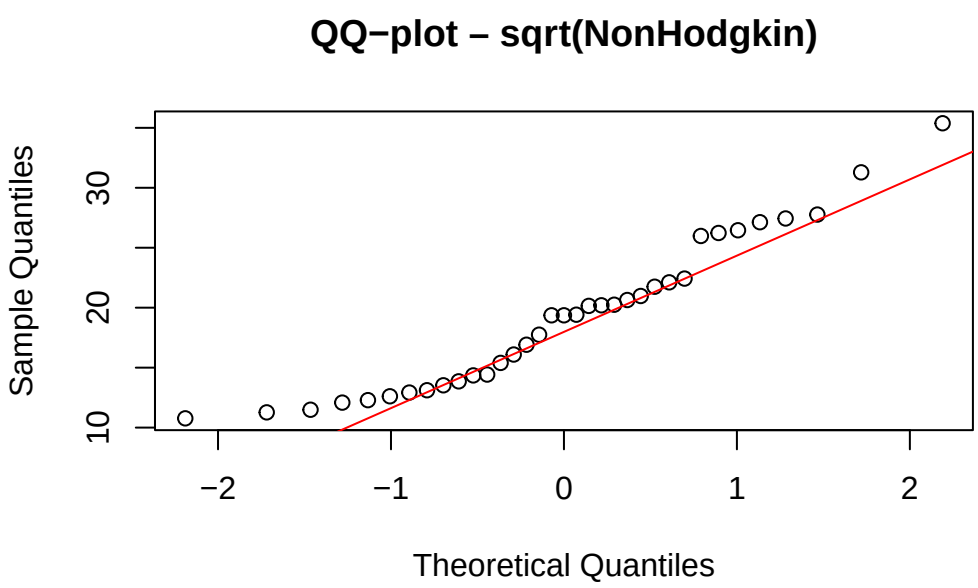

```
normplot(NonHodgkin)
```



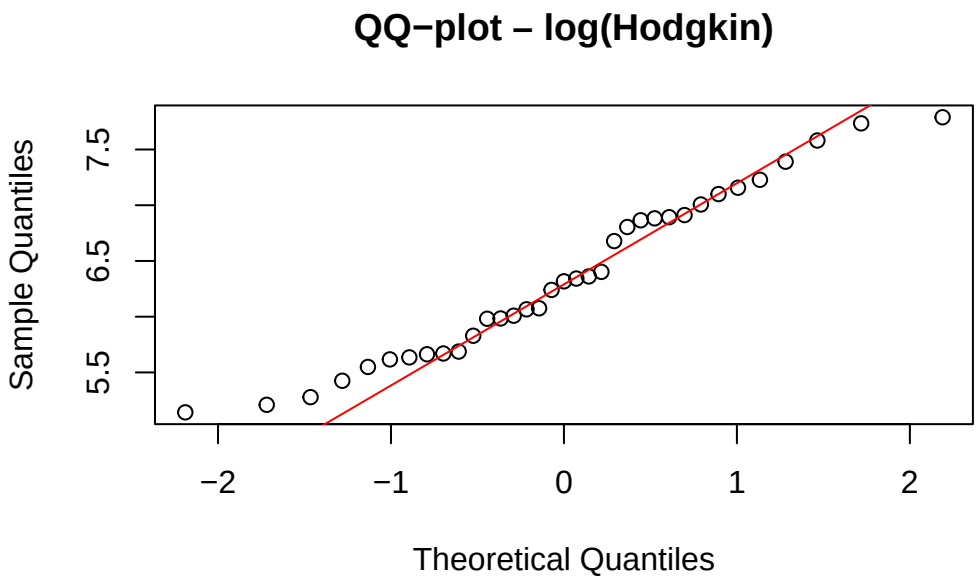
```r
# Transformationer: sqrt och log


sqrtH <- sqrt(Hodgkin)
sqrtN <- sqrt(NonHodgkin)

qqnorm(sqrtH, main = "QQ-plot - sqrt(Hodgkin)")
qqline(sqrtH, col = "red")
```

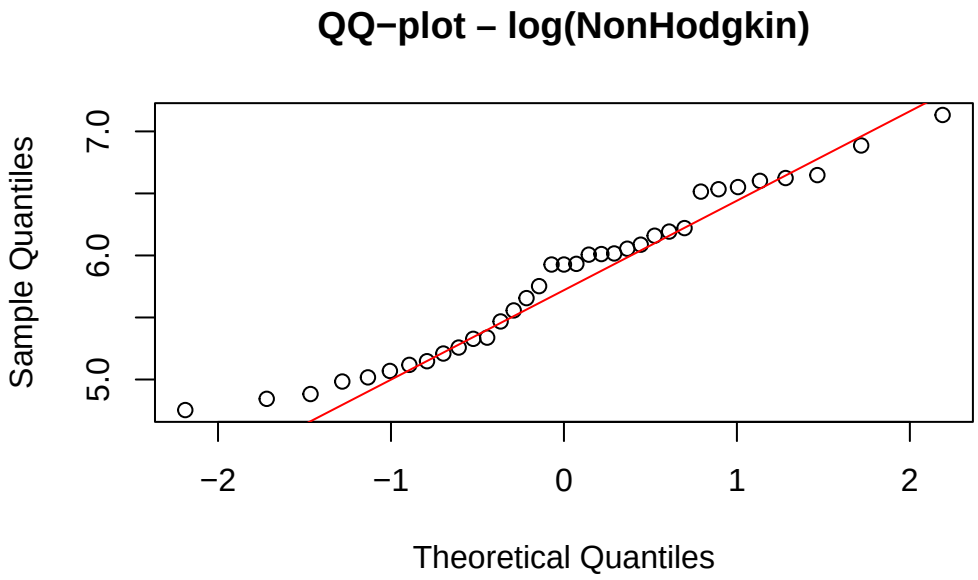



```r
qqnorm(sqrtN, main = "QQ-plot - sqrt(NonHodgkin)")
qqline(sqrtN, col = "red")
```




```r
if (all(Hodgkin > 0) && all(NonHodgkin > 0)) {
  logH <- log(Hodgkin)
  logN <- log(NonHodgkin)

  qqnorm(logH, main = "QQ-plot - log(Hodgkin)")
  qqline(logH, col = "red")

  qqnorm(logN, main = "QQ-plot - log(NonHodgkin)")
  qqline(logN, col = "red")
}
```






```
# Jämförelse av medelvärden och CLT


mH <- mean(Hodgkin)
```

```r
mN <- mean(NonHodgkin)
diff_means <- mH - mN

mH
```

```
[1] 758.9429
```

```r
mN
```

```
[1] 408.3429
```

```r
diff_means
```

```
[1] 350.6
```

```r
# Standard error för differens i medelvärden (oberoende stickprov)
se_diff <- sqrt(var(Hodgkin) / length(Hodgkin) + var(NonHodgkin) / length(NonHodgkin))
se_diff
```

```
[1] 110.5427
```

```r
# En enkel normal approx-intervall för differensen (CLT)
ci_low <- diff_means - 1.96 * se_diff
ci_high <- diff_means + 1.96 * se_diff

ci_low
```

```
[1] 133.9363
```

```r
ci_high
```

```
[1] 567.2637
```

```r
# Samma idé efter sqrt-transformation


mH_s <- mean(sqrtH)
mN_s <- mean(sqrtN)
diff_means_s <- mH_s - mN_s

se_diff_s <- sqrt(var(sqrtH) / length(sqrtH) + var(sqrtN) / length(sqrtN))

diff_means_s
```

```
[1] 6.504328
```

```r
se_diff_s
```

```
[1] 1.991901
```

```r
ci_low_s <- diff_means_s - 1.96 * se_diff_s
ci_high_s <- diff_means_s + 1.96 * se_diff_s

ci_low_s
```

```
[1] 2.600202
```

```r
ci_high_s
```

```
[1] 10.40845
```

```r
# Samma idé efter log-transformation (om möjligt)


if (all(Hodgkin > 0) && all(NonHodgkin > 0)) {
  mH_l <- mean(logH)
  mN_l <- mean(logN)
  diff_means_l <- mH_l - mN_l

  se_diff_l <- sqrt(var(logH) / length(logH) + var(logN) / length(logN))

  diff_means_l
```

```r
  se_diff_l

  ci_low_l <- diff_means_l - 1.96 * se_diff_l
  ci_high_l <- diff_means_l + 1.96 * se_diff_l

  ci_low_l
  ci_high_l
}
```

```
[1] 0.873799
```

I denna del analyserade vi T4-cellsnivåer för två grupper, `Hodgkin` och `NonHodgkin`, med syfte att undersöka fördelningsform, normalitet och skillnad i medelvärde mellan grupperna. De inledande histogrammen visade att båda grupperna uppvisar en viss grad av skevhet, särskilt i den högra svansen. Detta indikerar att normalantagandet inte är självklart uppfyllt på individnivå.

QQ-plots för de ursprungliga observationerna visade att punkterna inte låg helt längs referenslinjen, särskilt i svansarna. Avvikelserna var dock inte extrema, vilket tyder på att datan inte är fundamentalt oförenlig med normalfördelning, men att vissa systematiska skevheter förekommer. Detta är viktigt eftersom flera statistiska metoder bygger på normalantaganden.

För att undersöka om en transformation kunde förbättra normalanpassningen prövade vi både kvadratrotstransformation och logaritmisk transformation. Efter transformationerna blev QQ-plotarna mer linjära, särskilt för log-transformationen. Detta tyder på att fördelningen närmar sig symmetri och att variationen stabiliseras, vilket är ett typiskt syfte med sådana transformationer.

Vid jämförelse av medelvärden erhölls stickprovsmedelvärdena $\bar{x}_H$ och $\bar{x}_N$ för respektive grupp. Skillnaden i medelvärde kan skrivas som

$$\Delta = \bar{x}_H - \bar{x}_N.$$

För att bedöma osäkerheten i denna skattning beräknade vi standardfelet

$$SE(\Delta) = \sqrt{\frac{s_H^2}{n_H} + \frac{s_N^2}{n_N}},$$

där s_H^2 och s_N^2 är stickprovsvarianserna och n_H, n_N respektive stickprovsstorlekar.

Med hjälp av centrala gränsvärdessatsen kan fördelningen av Δ approximeras med en normalfördelning när stickprovsstorlekarna är tillräckligt stora, även om de underliggande fördelningarna inte är exakt normala. Ett approximativt 95%-konfidensintervall kan därför skrivas som

$$\Delta \pm 1.96 \cdot SE(\Delta).$$

Detta intervall ger ett mått på den statistiska osäkerheten kring skillnaden i medelvärde mellan grupperna.

Resultaten visar att även om individdata uppvisar viss skevhet, möjliggör centrala gränsvärdessatsen en normalapproximation för medelvärdesskillnaden. Detta illustrerar en central princip inom inferensstatistiken: normalitet på observationsnivå är inte nödvändig för att medelvärden ska kunna analyseras med normalbaserade metoder, förutsatt att stickprovsstorleken är tillräckligt stor.

Sammantaget visar denna del hur grafisk analys, transformationer och teoretiska resultat såsom centrala gränsvärdessatsen samverkar i praktisk statistisk analys. Genom att kombinera dessa verktyg kan man både diagnostisera fördelningsform och genomföra inferens på ett teoretiskt välgrundat sätt.

AI Statement

Ingen AI användes i detta projekt.