Laboration 2

Sebastijan Babic

2024-10-23

Sammanfattning

I denna laboration testades slumptalsgeneratorn i R genom att generera och analysera data från kontinuerligt likformiga, normalfördelade och exponentialfördelade slumpvariabler.

Resultaten visade att R generator producerade data som följde förväntade fördelningar, särskilt vid stora urvalsstorlekar.

En transformation av exponentialfördelade data bekräftade också att de kunde omvandlas till en likformig fördelning på $[0, 1]$, vilket indikerar att generatorn fungerar som förväntat.

Uppgift 1

Vi delar in intervallet $[0, 1]$ i k lika stora delintervaller, såkallade klasser.

Vi generar sedan n observationer från en kontinuerlig likformig fördelning på $[0, 1]$.

Antalet observationer som hamnar i den k:te delintervall kallar vi för slumpvariabeln $\mathbb{X}_k$ för $k = 1, 2, 3, \ldots$

Teoretiska uppgifter

Upggift 1.1

För att skriva det första och sista delintervallet som $[a, b]$ där a, b uttrycks med k, dvs $\mathbb{X}_1$ respektive $\mathbb{X}_k$ kan vi skriva längden på varje delintervall som $\frac{1}{k}$ då vi har en likformig kontinuerlig fördelning.

Detta ger oss att intervallet för $\mathbb{X}_1$ är $\left[0, \frac{1}{k}\right]$. Detta får vi eftersom b ges av att gå en längd $1/k$ från 0. Dvs. $b = 0 + 1/k = 1/k$.

Intervallet för $\mathbb{X}_k$ får vi genom att uttrycka a som $\frac{1-k}{k}$ då det är den näst sista delintervall, dvs den $k-1$:te delintervall. b kan vi skriva som $\frac{k}{k} = 1$. Det ger oss intervallet $\left[\frac{1-k}{k}, 1\right]$.

Upggift 1.2

Som skrivet i uppgift 1.1 så är längden på varje sådan delintervall $1/k$ då vi behandlar en likformig kontinuerlig fördelning.

Uppgift 1.3

Vi genererar ett slumptal x från den kontinuerliga likformiga fördelningen på $[0, 1]$. Vad är då

(1) Sannolikheten att detta slumptal hamnar i $\mathbb{X}_1$?

Då vi behandlar en likformig fördelning så är sannolikheten att talet genereras i $\mathbb{X}_1$ precis $\frac{1}{k}$ per känd sats från kursen.

(2) Sannoliketen att $x \notin \mathbb{X}_1$?

Sannolikheten att detta inte sker är komplementet (1). Det vill säga $P(A^\complement) = 1 - P(A) = 1 - \frac{1}{k}$. Alltså har vi att $P(A^\complement) = \frac{k-1}{k}$.

(3) Sannolikhetsfördelningen som beskriver sannolikheten att hamna i det första intervallet eller inte? Vi ska dessutom ange parametern hös fördelningen.

Då det handlar om endast en enda observation så kan vi helt enkelt säga att sannolikheten att hamna i intervallet $\mathbb{X}_1$ är $1/k$.

Sannolikheten att observationen hamnar i det första intervallet är $1/k$. Sannolikheten att den inte hamnar där, dvs. komplementet är $\frac{k-1}{k}$. Vi har alltså en bernoullifördelning med parameter $p = 1/k$.

Upggift 1.4

Om vi nu som i beskrivningen genererar n sådana slumptal, oberoende av varandra, vad är sannolikheten att j av dessa hamnar i det första delintervallet? Dvs. för $j = 0, 1, \ldots, n$, vad är $P(X_1 = j)$? Vilken fördelning följer X_1 ? Ange parametrarna hos fördelningen.

Då varje observation kan hamna i ett delintervall med sannolikheten $1/k$, detta är vår parameter p, och det finns totalt n observationer alla oberoende av varandra som ger oss våran parameter n. Det innebär att vi har att $\mathbb{X}_1 \sim \text{Bin}(n, \frac{1}{k})$.

Via sannolikhetsfunktionen för en binomialfördelad slumpvariabel får vi att

$$P(\mathbb{X}_1 = j) = \binom{n}{j} \left(\frac{1}{k}\right)^j \left(1 - \frac{1}{k}\right)^{n-j}$$

per våran formelsamling.

Upggift 1.5

Har $X_2, \ldots, X_k$ samma sannolikhetsfördelning som X_1 ? Ar $X_1, X_2, \ldots, X_k$ oberoende av varandra? Motivera!

Alla X_k följer samma sannolikhetsfördelning eftersom de representerar antal observationer som faller inom lika stora delintervall av en likformig fördelning över $[0, 1]$. Vi ser att X_k är beroende eftersom om vi vet antalet observationer i några av delintervall så kommer antalet observationer i de återstående delintervallen också vara känd ty summan av alla observationer ska vara lika med det totala antalet observationer. Det gör dem beroende.

Upggift 1.6

Vi ska ange väntevärde, varians och standardavvikelse för $Y_1 = \frac{X_1}{n}$, dvs andelen av observationerna som hamnar i första intervallet. Vi ser att X_1 följer en binomialfördelning med parameter n, antalet observationer och $p = 1/k$, sannolikheten att varje observation hamnar i det första intervallet. Då har vi alltså att

$$E(X_1) = np = n\frac{1}{k}$$

och eftersom

$$Y_1 = \frac{X_1}{n} \implies E(Y_1) = E\left(\frac{X_1}{n}\right) = \frac{E(X_1)}{n} = \frac{n1/k}{n} = \frac{1}{k}$$

För att beräkna variansen behöver vi alltså använda oss av formeln

$$\mathrm{Var}(X_1) = E(X_1^2) - E(X_1)^2$$

eller så använder vi den redan kända formeln för varians av en binomialfördelad slumpvariabel och får att

$$\mathrm{Var}(X_1) = np(1-p) = n\frac{1}{k}\left(1 - \frac{1}{k}\right)$$

som innebär att

$$\mathrm{Var}(Y_1) = \mathrm{Var}\,\frac{X_1}{n}$$

som via räkneregler för varians ger

$$\frac{\mathrm{Var}(X_1)}{n^2} = \frac{\frac{1}{k}(1-\frac{1}{k})}{n}$$

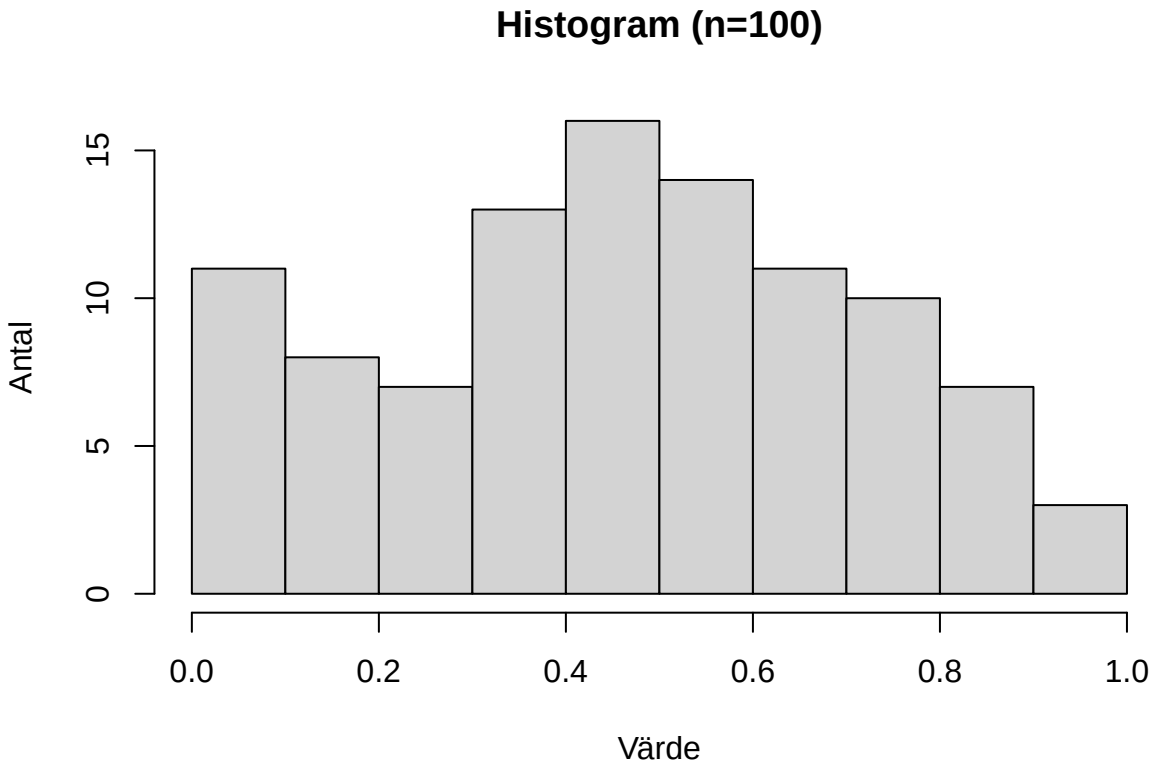
Som innebär att variansen minskar då n ökar som är inte så märkligt då fler observationer leder till mindre variationerna i andelen.

Slutligen, att beräkna standardavvikelsen så har vi formeln

$$D(Y_1) = \sqrt{\mathrm{Var}(Y_1)} = \sqrt{\frac{\frac{1}{k}(1-\frac{1}{k})}{n}}$$

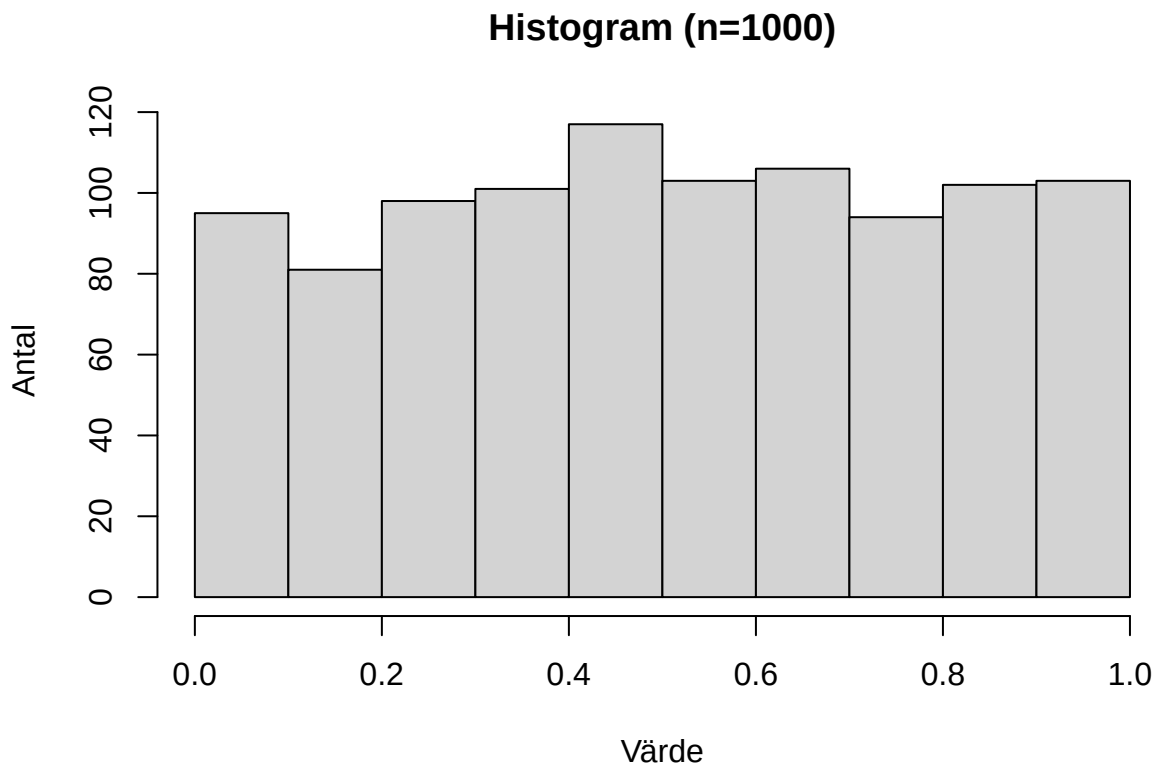
Kodrelaterade uppgifter

```r
# Sätt frö för reproducerbarhet
set.seed(20040911)

# Generera 100 observationer
slumptal_100 <- runif(100) # runif = random uniform
hist(
    slumptal_100,
    breaks = seq(0, 1, length.out = 11),
    main = "Histogram (n=100)",
    ylab = "Antal",
    xlab = "Värde"
)
```



Figur 1: Histogram för 100 slumpvis genererade observationer där slumpmässiga observationer från en kontinuerligt likformig fördelning på [0,1] har delats upp i 10 lika stora intervall.

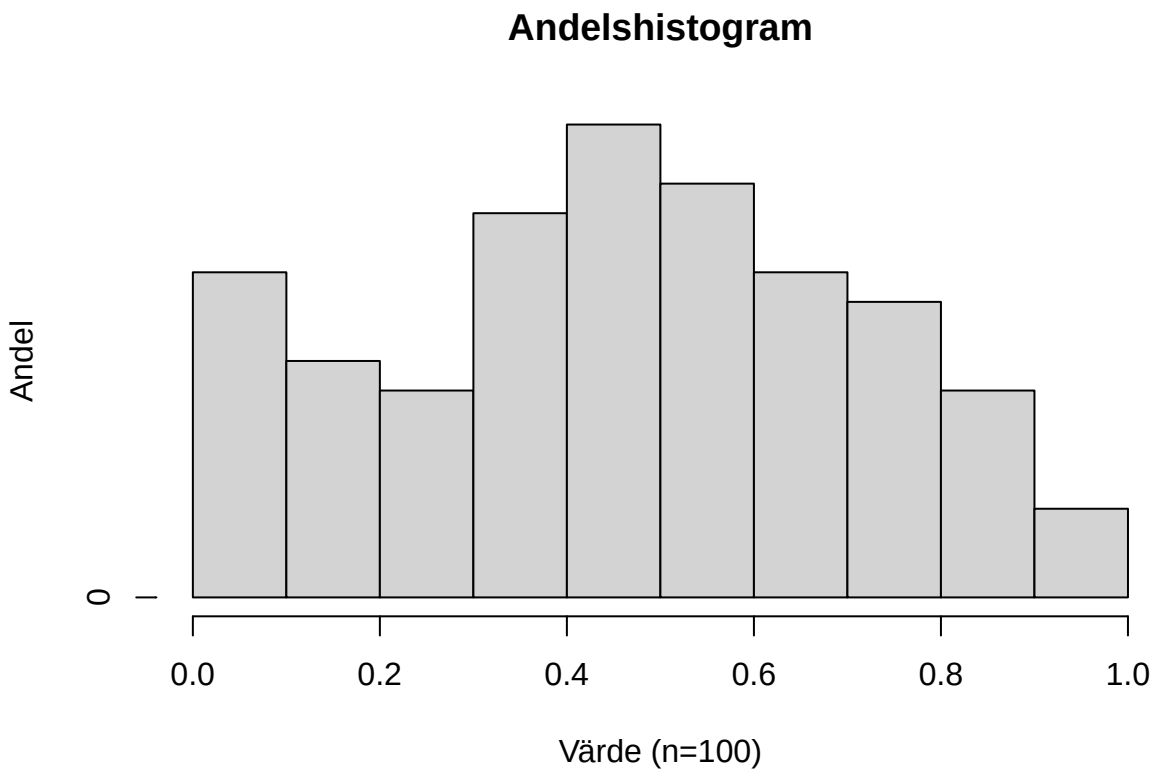
```r
# Generera 1000 observationer
slumptal_1000 <- runif(1000)
hist(
    slumptal_1000,
    breaks = seq(0, 1, length.out = 11),
    main = "Histogram (n=1000)",
    ylab = "Antal",
    xlab = "Värde"
)
```



Figur 2: Histogram för 1000 slumpvis genererade observationer där slumpmässiga observationer från en kontinuerligt likformig fördelning på [0,1] har delats upp i 10 lika stora intervall.

```r
# Funktion för andelshistogram
prop_hist <- function(x, xlab = "Värde") {
  p <- hist(x, plot = FALSE)
  p$counts <- p$counts / sum(p$counts)
  plot(p,
       main = "Andelshistogram",
       ylab = "Andel",
       xlab = xlab
      )
}
```

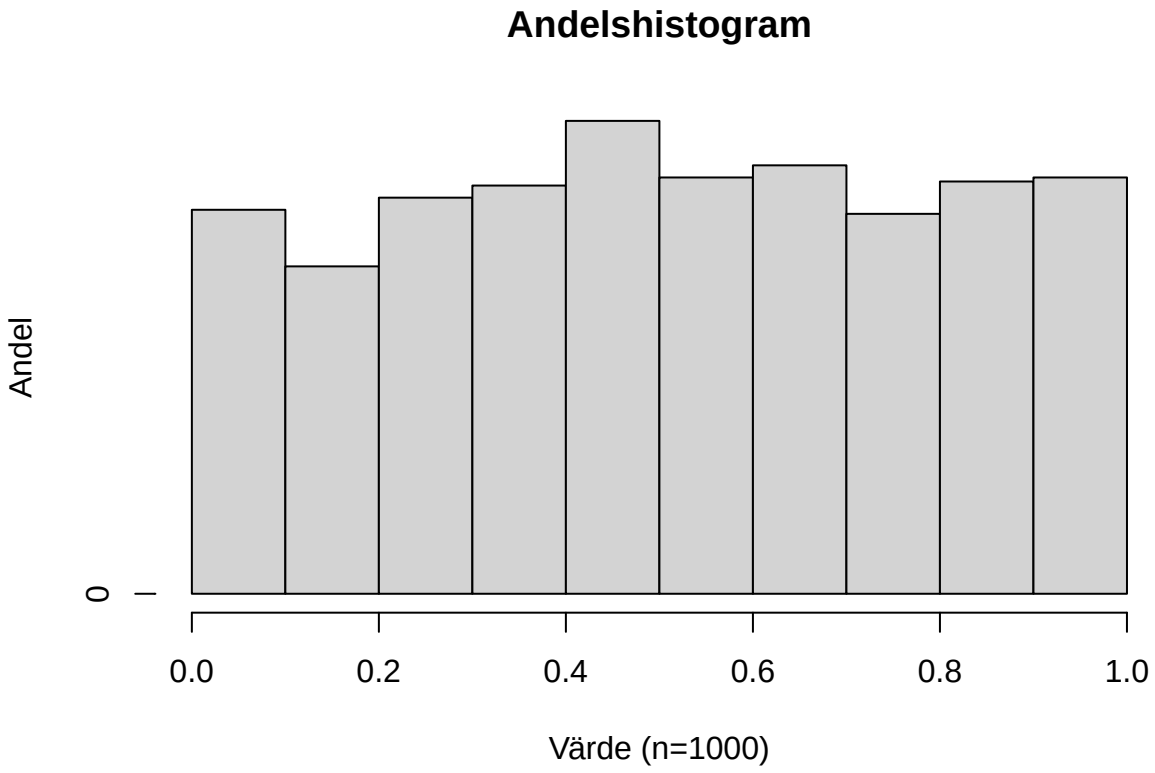
```r
# Andelshistogram för 100 observationer
prop_hist(slumptal_100,
          xlab = "Värde (n=100)"
        )
```



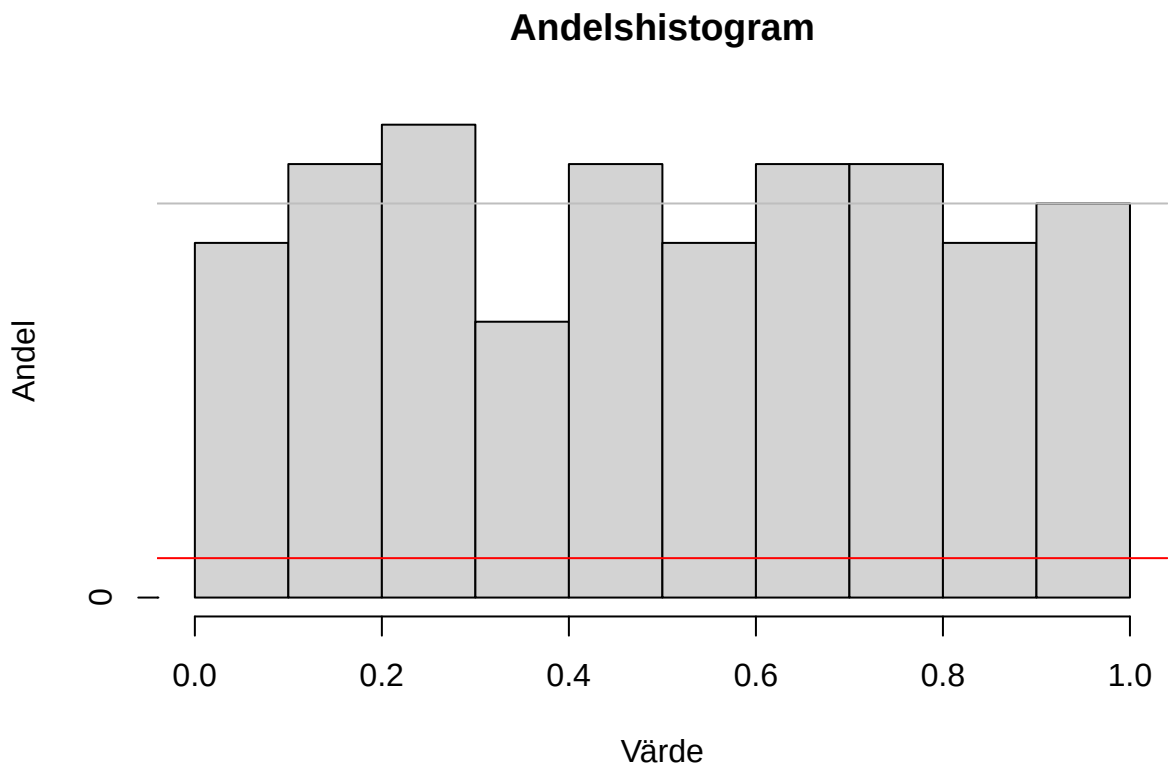
Figur 3: Andelshistogram för 100 slumpvis genererade observationer där slumpmässiga observationer från en kontinuerligt likformig fördelning på [0,1] har delats upp i 10 lika stora intervall.

```r
# Andelshistogram för 1000 observationer
prop_hist(slumptal_1000,
          xlab = "Värde (n=1000)"
        )
```

```r
k <- 10 # Antal intervall
n <- 100 #
slumptal_100 <- runif(100)

E <-  1/k # Väntevärdet, beräkna själv genom de teoretiska uppgifterna
D <-  sqrt((1 / k) * (1 - 1 / k) / n) # Standardavvikelsen, beräkna själv genom de teoretiska uppgifter

prop_hist(slumptal_100)
abline(a = E, b = 0, col = "grey") # Väntevärdet
abline(a = E + 3 * D, b = 0, col = "red") # Väntevärdet + 3 standardavvikelser
abline(a = E - 3 * D, b = 0, col = "red") # Väntevärdet - 3 standardavvikelser
```



Figur 4: Andelshistogram för 1000 slumpvis genererade observationer där slumpmässiga observationer från en kontinuerligt likformig fördelning på [0,1] har delats upp i 10 lika stora intervall.



Figur 5: Andelshistogram för n=100. Väntevärdet (grå linje) visar den förväntade andelen observationer per intervall, och de röda linjerna visar gränserna för tre standardavvikelser från väntevärdet. Detta ger en indikation på om observerade andelar ligger inom det förväntade intervallet.

Uppgift 1 KOD

För ett litet värde av n, ser det ut som en likformig fördelning? Motivera varför genom att hänvisa till de teoretiska resultaten.

Vi ser att om $n = 100$ har vi ganska ojämna staplar, detta kan synas i figur 1. Att ha ojämna staplar innebär här att vi inte riktigt har en likformig fördelning på intervallet $[0, 1]$. Vi får exempelvis värdet 0 cirka 11 gånger medan vi får 0.5 får vi 15. Vi kan även se att vi får värdet 1 cirka 3 gånger. Alltså är avvikelsen mycket stora från väntevärdet.

Uppgift 2 KOD

Vad händer då antalet slumptal n blir stort? Experimentera gärna med olika stora värden.

Vi ser som per figur 2 att vi får en histogram mycket mer likt en likformig fördelning mellan antal och värden. Vi bildar nästan en linje precis på antal 100 som ses på y-axeln.

Uppgift 3 KOD

Jämför andelen observationer i klasserna (intervallerna) med de förväntade andelarna och se om avvikelserna verkar stora (jämfört med väntevärde ± 3 standardavvikelser).

Om staplarna i histogrammet, dvs. andelarna av observationer i varje intervall främst ligger inom det grå (väntevärde) och röda området (± 3 standardavvikelser), kan vi säga att andelarna stämmer överens med de förväntade värdena. Då vi ökar n så ser vi att vi får mer och mer staplar emellan det grå och röda.

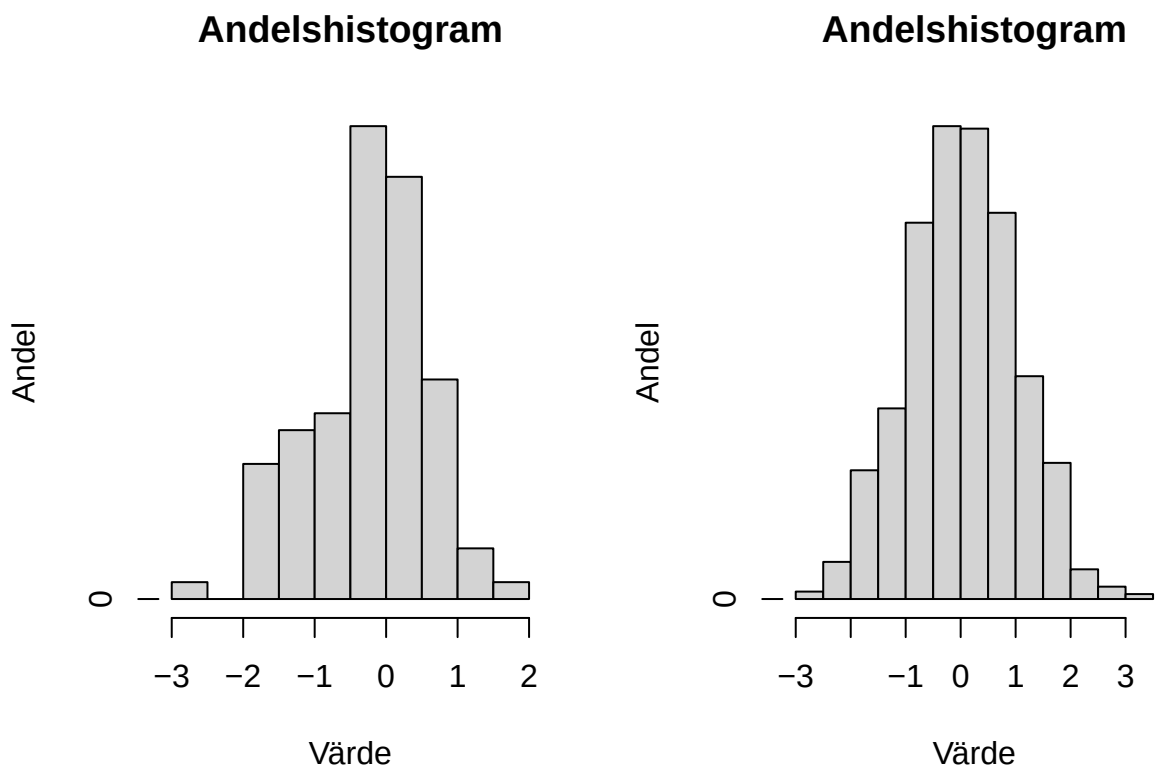
Uppgift 4 KOD

Då det är ändå en majoritet av staplarna inom intervallet mellan väntevärdet och standardavvikelsen kan vi nog säga att slumptalsgeneratorn får godkänt givet standardavvikelse ± 3 räcker för det som undersöks.

Uppgift 2 - Normal- och exponentialfördelade slumptal

```r
# Ange parametrar för normal- och exponentialfördelningen
m <- 0  # Medelvärde för normalfördelningen
s <- 1  # Standardavvikelse för normalfördelningen

# Set seed for reproducibility
set.seed(20040911)

# Skapa layout för två diagram sida vid sida
par(mfrow = c(1, 2))

# Normalfördelade slumptal med n = 100
n <- 100
normal_data_100 <- rnorm(n, mean = m, sd = s)
prop_hist(normal_data_100)


# Normalfördelade slumptal med n = 1000
n <- 1000
```

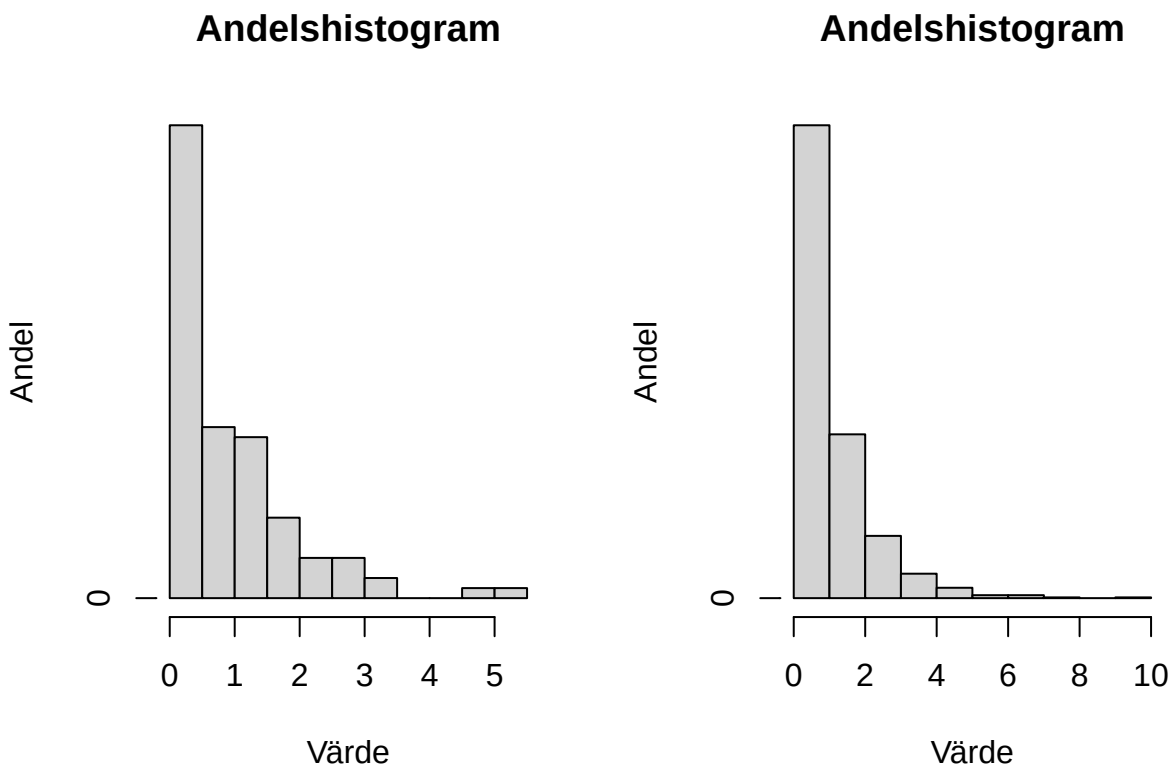
```r
normal_data_1000 <- rnorm(n, mean = m, sd = s)
prop_hist(normal_data_1000)
```



Figur 6: Andelshistogram på normalfördelade vektorer med n = 100 till vänster respektive n = 1000 höger. Högerhistogramen börjar nu likna den klassiska bell curve som skapas vid tillräckligt stora n i en normalfördelning.

```r
a <- 1  # Parametern för exponentialfördelningen

# Skapa layout för två diagram sida vid sida
par(mfrow = c(1, 2))

# Exponentialfördelade slumptal med n = 100
n <- 100
exponential_data_100 <- rexp(n, rate = a)
prop_hist(exponential_data_100)

# Exponentialfördelade slumptal med n = 1000
n <- 1000
exponential_data_1000 <- rexp(n, rate = a)
prop_hist(exponential_data_1000)
```

Ser de slumpade vektorerna normal- respektive exponentialfördelade ut? Motivera era svar:



Figur 7: Andelshistogram på exponentiellt fördelade vektorer med n = 100 till vänster respektive n = 1000 höger. Både visar en högersvans, en karaktär hos en exponentiell fördelning men vi ser att vid större n får vi mindre oväntade andel såsom vi fick vänster för värde 5.

Givet att n tillräckligt stort så får vi precis normal- respektive exponentialfördelningar. Detta syns mycket tydligt i figur 6 respektive 7. Vi vet att dessa fördelningar har täthetsfunktionerna:

$$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

respektive

$$\beta e^{-\beta x}$$

per formelsamlingen, dessa täthetsfunktioner har precis de beteende som syns i figurerna 6 och 7. Alltså jämför vi resultatena med täthetsfunktionen i både fall.

Normalfördelningen är symmetrisk kring sitt medelvärde, i detta fall $m = 0$. Majoriteten av alla värden ligger inom ± 1 standardavvikelse från mittpunkten.

Exponentialfördelningen:s täthetsfunktion är en avtagande funktion, detta är uppenbarligen fallet som vi har här i figur 7. Det högsta värdet bör antas i $x = 0$ och även det uppfylls av figur 7.

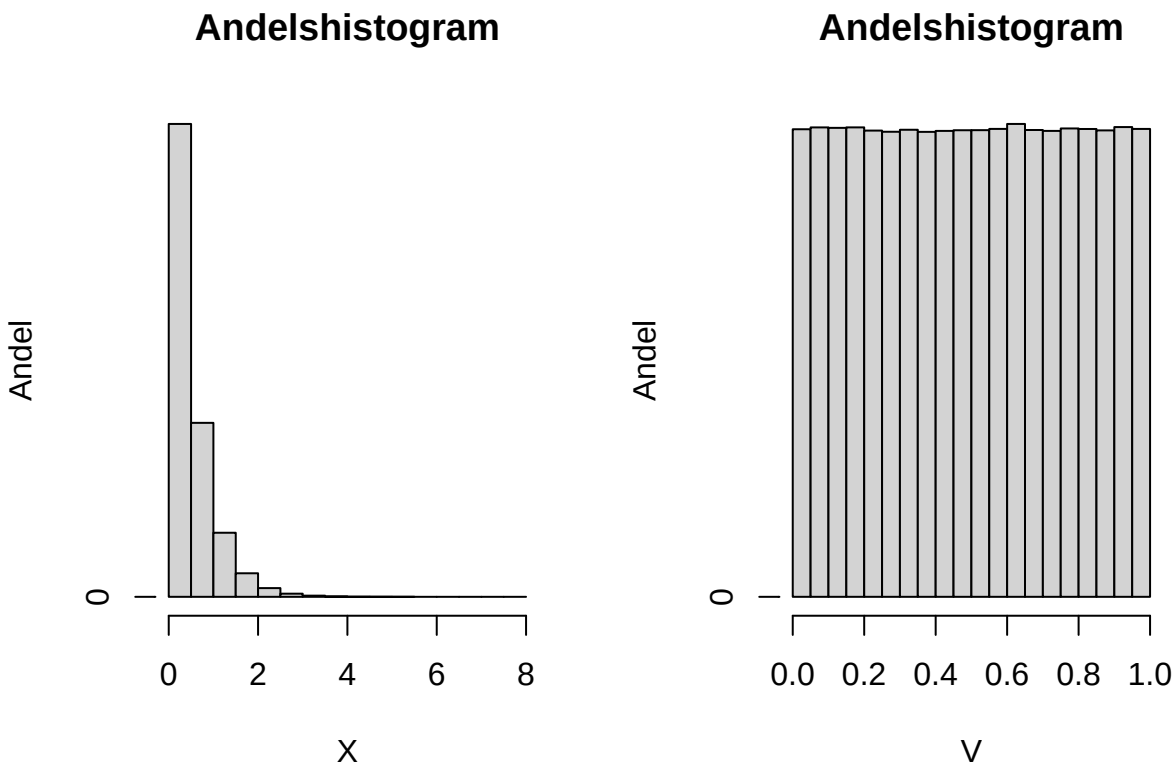
Uppgift 3

```r
set.seed(20040911) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.
n = 1000000

x <- rexp(n, rate = 2) # välj värde för n
v <- 1 - exp(-2 * x)

# Plotta x och v bredvid varandra
par(mfrow = c(1, 2))
prop_hist(x, xlab = "X")
prop_hist(v, xlab = "V")
```

Jämför de två andelshistogrammen. Ser det ut som att påstående stämmer? Dvs., leder transformationen av de exponentialfördelade slumptalen till en likformig fördelning? Motivera! Tänk på vad ni redan gjort i denna labb!

Histogrammet över X, den vänsta diagramen bör visa en exponentialfördelning, dvs vi ska ha en tydlig högersvans som avtar exponentiellt och det är precis det som sker.

Den högra diagramen bör visa en likformig fördelning på $[0, 1]$ och vi har det på ett ungefär då andelen för varje värde är ungefär lika. Givet tillräckligt stort n får vi precis en likformig fördelning, här tar vi $n = 100000$ och får en diagram som är mycket representativ på en likformig fördelning.



Figur 8: Figur som visar två andelshistogram därav den vänstra visar fördelningen för X, dvs. en exponentiell fördelning. Den högra andelshistogramen visar den transformerade exponentiella fördelningen med väntevärde 1/2 som visar en ungefärlig likformig fördelning.