

Kansrekening en Statistiek met R

Erik van Zwet

12 november 2007

Inhoudsopgave

Voorwoord	7
1 Introductie in R	9
1.1 beginnen en ophouden	10
1.2 R als rekenmachine	10
1.3 variabelen	11
1.4 vectoren	12
1.5 subscripting	13
1.6 matrices	13
1.7 functies	15
1.8 plotten	15
1.9 scripting	16
2 Kansrekening	19

2.1	experiment, uitslagenruimte, gebeurtenis	19
2.2	verzamelingstheorie	20
2.3	kans	21
2.4	onafhankelijkheid	22
2.5	voorwaardelijke kansen	23
2.6	combinatoriek	24
2.7	discrete kansverdelingen	25
2.8	continue kansverdelingen	27
2.9	histogram	29
2.10	Monte Carlo	30
3	Locatie en spreiding	33
3.1	verwachting	33
3.2	variantie	34
3.3	gemiddelde	36
3.4	steekproef variantie	37
3.5	covariantie	38
3.6	correlatie	41
3.7	robuustheid: mediaan, MAD en uitbijters	41

<i>INHOUDSOPGAVE</i>	5
4 Schatten	45
4.1 parameter schatten	45
4.2 lineaire regressie	48
4.3 lineaire modellen van hogere orde	49
4.4 andere lineaire modellen	51
5 Toetsen	53
5.1 Inleiding	53
5.2 Z toets	56
5.3 Student's t toets	57
5.4 t toets, twee steekproeven	58
5.5 Wilcoxon toets	60
5.6 de F toets	60
5.7 de χ^2 toets	61
5.8 Kolmogorov-Smirnov toets	63
6 ANOVA	65
6.1 lineaire regressie	65
6.2 ANOVA; 1 factor	67
6.3 ANOVA; 2 factoren	70

6.4 ANOVA; twee factoren met interactie	71
A Kansverdelingen	75

Voorwoord

In dit dictaat behandelen we een aantal onderwerpen uit de kansrekening en statistiek. Dit doen we vooral met voorbeelden en kleine opgaven in de programmeertaal R. We moeten dus eerst wat R leren, maar dit college is géén computercursus. Wees dus niet bezorgd als je niet goed kan programmeren.

Het is aan te raden om naast dit dictaat ook een gewoon boek te gebruiken. Bijna iedere inleiding in de kansrekening en statistiek voldoet. Ik heb zelf gebruik gemaakt van J.A. Rice *Mathematical Statistics and Data Analysis*, second edition, Wadsworth 1995. Op het internet is natuurlijk ook alles te vinden.

In de eerste week doen we hoofdstukken 1 tot en met 3. In de tweede week beginnen we pas echt met statistiek, en doen we hoofdstukken 4 tot en met 6.

In het rooster staat een deeltentamen statistiek, maar dat vervalt. In plaats daarvan moeten jullie een dataset analyseren en daar een verslag over schrijven. Je *moet* daarvoor R gebruiken. Je mag met z'n tweeën werken, maar dat hoeft niet. Over twee weken, als we het hele dictaat hebben doorgewerkt, kiest elk paar een dataset uit “The Data and Story Library”

<http://lib.stat.cmu.edu/DASL/>

Het verslag moet op dinsdag 4 december uiterlijk voor middernacht per email worden ingeleverd.

Hoofdstuk 1

Introductie in R

R is een programmeer-taal met een groot aantal ingebouwde statistische functies. Het is de *open source* versie van S-plus. Wij gebruiken R dan ook omdat het gratis is. Je kunt het downloaden vanaf

<http://www.r-project.org/>

Documentatie voor R (S-plus) vind je op het web:

http://biology.leidenuniv.nl/~zandee/ab07/	door Rino Zandee
http://www.math.montana.edu/stat/tutorials/R-intro.pdf	door ?
http://www.math.montana.edu/stat/docs/sguide.ps	door Brian Ripley
http://www.math.montana.edu/stat/docs/Splus_notes.ps	door Bill Venables
http://www.stat.berkeley.edu/users/spector/intro_s.pdf	door Phil Spector

Andere, bekendere software voor statistische analyse is Excel, SPSS, SAS, Minitab, etc. Wij maken geen gebruik van Excel, omdat het de statistische functies teveel achter knoppen verborgen houdt. Wij willen juist zien hoe alles in z'n werk gaat. R kan wel eenvoudig data vanuit Excel importeren.

In dit hoofdstuk behandelen we de belangrijkste R commando's. Het is de bedoeling dat je ze zelf intypt, of met copy-paste overbrengt, en kijkt wat er gebeurt.

1.1 beginnen en ophouden

Start R op. Je krijgt een window met een prompt `>`. Hier kun je commando's typen. Om R af te sluiten, typ je

```
> q()
Save workspace image? [y/n/c]:
```

en typ n.

1.2 R als rekenmachine

```
> 2+4
[1] 6
> 3*5
[1] 15
> 2^3          # of 2**3
[1] 8
> sqrt(9)
[1] 3
> sin(pi)     # R kent pi
[1] 1.224606e-16 # nou ja, eigenlijk 0!
```

Soms is het antwoord van een berekening TRUE of FALSE

```
> 2>4          # is 2 groter dan 4?
[1] FALSE

> 3==3         # let op de dubbele ==
[1] TRUE
```

```
> 3!=3          # != staat voor ongelijkheid
[1] FALSE

> (3>2)&(3>4)    # & staat voor het logische "EN"
[1] FALSE

> (3>2)|(3>4)    # | staat voor het logische "OF"
[1] TRUE
```

opgave 1.1 Bepaal uit je hoofd

```
((3>2)|(3>4))|((3>2)&(3>4))
```

en controleer je antwoord met R Zijn de haakjes belangrijk? Wat is

```
((3>2)|(3>4)|(3>2))&(3>4)
```

1.3 variabelen

Het volgende commando maakt een variabele `a` met waarde 3.162.

```
> a = sqrt(10)
```

Kijk maar:

```
> a
[1] 3.162278
```

We kunnen ook rekenen met variabelen.

```
> b = a+5
```

maakt een variabele `b` met waarde 8.162.

Een zogeheten Boolese variabele (naar de Engelsman George Boole) neemt de waarde `TRUE` of `FALSE` aan.

```
> c = (3<4)
> c
[1] TRUE
```

Je kunt ook met Boolese variabelen rekenen. Dan geldt: `TRUE=1` en `FALSE=0`.

1.4 vectoren

Vectoren zijn het belangrijkste onderdeel van R. Probeer de volgende voorbeelden een voor een uit

```
> x = c(1,3,2,6,0)
> x = rep(3,10)
> x = seq(1, 20, by = 3)
> x = seq(10, 1, by = -2)
> x = 1:10
> x
[1] 1 2 3 4 5 6 7 8 9 10
```

We kunnen ook rekenen met vectoren

```
> y = 3*x - 12
> y
[1] -9 -6 -3 0 3 6 9 12 15 18
> z = (y<0)
> z
[1] TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
```

opgave 1.2 Het commando `x = sample(1:10,10000,replace=T)` produceert een rij van tienduizend toevallig gekozen getallen tussen 1 en 10. Voer het commando uit, en tel hoeveel vijven er in de rij voorkomen.

1.5 subscripting

Subscripting is het selecteren van elementen van een vector. We gebruiken rechte haken []

```
> y[3]
[1] -3
> y[3:5]
[1] -3 0 3
> y[c(1,3,5)]
[1] -9 -3 3
```

Negatieve selectie kan ook

```
> y[-3] # laat het derde element weg
[1] -9 -6 0 3 6 9 12 15 18
> y[-c(1,3,5)] # laat elementen 1,3 en 5 weg
[1] -6 0 6 9 12 15 18
```

We kunnen ook selecteren met behulp van een Boolese vector. We selecteren alle elementen van `y` die groter zijn dan 0 door

```
> y[y>0]
[1] 3 6 9 12 15 18
```

1.6 matrices

We maken een matrix A

```
> A = matrix(1:12, nrow = 4, ncol = 3) # de getallen 1 t/m 12 in 4 rijen
                                     # en 3 kolommen.
> A
      [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

We kunnen natuurlijk elementen van de matrix selecteren

```
> A[1,2]
[1] 5

> A[1:2,2:3]
      [,1] [,2]
[1,]    5    9
[2,]    6   10

> A[2,]          #selecteer de tweede rij
[1] 2 6 10
```

opgave 1.3 Het volgende antieke Chinese lineaire probleem bestaat uit drie vergelijkingen voor drie onbekenden (x_1, x_2 en x_3).

$$3x_1 + 2x_2 + x_3 = 39$$

$$2x_1 + 3x_2 + x_3 = 34$$

$$x_1 + 2x_2 + 3x_3 = 26$$

In matrix notatie is dit probleem van de vorm $Ax = b$ met

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \text{en } b = \begin{pmatrix} 39 \\ 34 \\ 26 \end{pmatrix}$$

De matrix A is inverteerbaar, en dus is de oplossing van het probleem $A^{-1}b$. Bepaal deze oplossing met R. De functie `solve()` berekent de inverse van een matrix, en we kunnen matrices en vectoren vermenivuldigen met `%*%`.

1.7 functies

We hebben al een aantal functies gebruikt: `q`, `sqrt`, `sin`, `c`, `rep`, `seq`, `matrix` en `solve`. R heeft bijzonder veel ingebouwde functies. De uitleg van een functie is de vinden door

```
> help(seq)
```

Hoe weet je nou of Splus een ingebouwde functie heeft voor (bijvoorbeeld) de cosinus? De beste methode is proberen te raden

```
> cosinus(pi)
Error: couldn't find function "cosinus"
> cos(pi)
[1] -1      # Aha!
```

Je kunt ook nog altijd Google of de documentatie proberen.

opgave 1.4 Stel, x is een vector. Wat is het verschil tussen `rank(x)`, `sort(x)` en `order(x)`? Probeer ze allemaal uit.

1.8 plotten

Je kunt de functies `plot()`, `points()` en `lines()` gebruiken om plaatjes te maken.

```
> x = seq(0,2*pi,by=0.1)
> plot(x,sin(x))
```

of

```
> plot(x,sin(x),type='l')
```

Let op: 'l' is een ℓ en niet een 1. Bekijk `help(plot)` voor alle mogelijkheden. De functie `plot()` maakt het bestaande plaatje eerst leeg. `lines()` en `points()` plotten over het bestaande plaatje heen.

opgave 1.5 Plot de functie $f(x) = x^2$ op het interval $[-5, 5]$. Breng vervolgens de functie $g(x) = |5x|$ aan in je grafiek.

opgave 1.6 Bedenk zelf een functie met een verticale asymptoot, en plot deze.

opgave 1.7 Voer het volgende commando uit

```
> plot(c(1,2),c(3,4),type='l',xlim=c(0,4),ylim=c(0,5))
```

Begrijp je hoe het werkt? Maak nu een plaatje van de letter F.

1.9 scripting

Een script is een lijstje met commando's. Hier is een voorbeeld van een scriptje dat de kwadraten van de getallen 1 t/m 10 uitrekent. We gebruiken een for-lus.

```
for (i in 1:10) {
  cat(i,i^2,"\n")
}
```

De functie `cat()` schrijft naar het scherm. De toevoeging `"\n"` betekent “nieuwe regel”. Laat de toevoeging weg, en kijk wat er gebeurt.

Hier is een script dat de if-constructie demonstreert. Let op alle accolades! Merk ook op hoe je door in te springen de structuur van het programma duidelijk kunt maken.

```
for (i in 1:10) {
  cat(i,i^2,"\n")
  if (i == 5) {
    cat("halverwege!\n")
  }
}
```

deze accolade sluit de if-constructie af
deze accolade sluit de for-lus af

opgave 1.8 Een for-lus is vaak niet nodig, en kost dan nodeloos veel rekentijd. Bepaal de kwadraten van de getallen 1 t/m 10 met behulp van een enkel commando.

Als je zelf een script moet schrijven, is het handig om een editor te gebruiken. Gebruik `new script` van het drop-down menu `file` om de R editor te openen.

opgave 1.9 De methode van Newton wordt gebruikt om het nulpunt van een functie f te bepalen. Begin met een start-waarde $x(1)$ die redelijk dicht in de buurt van het nulpunt is. Definieer recursief de volgende waarden

$$x(n+1) = x(n) - \frac{f(x(n))}{f'(x(n))}, \quad n = 1, 2, \dots$$

Implementeer de methode van Newton om het nulpunt te bepalen van de functie $f(x) = \cos(x)$ op het interval $[0, \pi]$. Begin met startwaarde $x(1) = 1$ en voer 5 stappen uit.

opgave 1.10 In de Fibonacci rij is ieder getal gelijk aan de som van de twee voorafgaande getallen: 1, 1, 2, 3, 5, 8, 13, 21, ... Gebruik R om de eerste 20 Fibonacci getallen te bepalen.

Hoofdstuk 2

Kansrekening

Hier volgt een korte introductie in de elementaire kansrekening. We gebruiken \mathbb{R} om steekproeven te genereren, en wat gevoel te krijgen voor het toeval.

2.1 experiment, uitslagenruimte, gebeurtenis

Bij het uitvoeren van een experiment hebben we een verzameling van mogelijke uitslagen, die we Ω noemen.

- werp tweemaal een munt: $\Omega = \{\text{kop/kop, kop/munt, munt/kop, munt/munt}\}$
- werp met een dobbelsteen: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- bepaal de massa van een planeet: $\Omega = \{\text{positieve reele getallen}\} = \mathbb{R}^+$
- bepaal een percentage: $\Omega = \{\text{reële getallen tussen 0 en 100}\} = [0, 100]$
- vraag een mening: $\Omega = \{\text{eens, oneens, geen mening}\}$

Een *gebeurtenis* A is een deelverzameling van Ω . Bijvoorbeeld, bij een worp met een dobbelsteen is $A = \{2, 4, 6\}$ de gebeurtenis de uitslag even is. We schrijven $A \subseteq \Omega$.

2.2 verzamelingstheorie

We blijven nog even bij de worp met een dobbelsteen. Zij A en B twee deelverzameling van Ω . Bijvoorbeeld $A = \{2, 4, 6\}$ en $B = \{1, 2, 5\}$.

- De *vereniging* van A en B zijn alle elementen die in A of B (of beide) zitten. Schrijf

$$A \cup B = \{1, 2, 4, 5, 6\}.$$

$A \cup B$ is weer een deelverzameling van Ω ; het is de gebeurtenis dat A of B optreedt.

- De *doorsnede* van A en B zijn alle elementen die zowel in A als in B zitten. Schrijf

$$A \cap B = \{2\}.$$

$A \cap B$ is weer een deelverzameling van Ω ; het is de gebeurtenis dat A en B beide optreden.

- Het *complement* van A zijn alle elementen van Ω die niet in A zitten. Schrijf

$$A^c = \{1, 3, 5\}.$$

A^c is weer een deelverzameling van Ω ; het is de gebeurtenis dat A niet optreedt.

opgave 2.1 Stel $\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2, 3, 4\}$, $B = \{2, 4, 6\}$ en $C = \{1, 3, 5\}$. Bepaal

1. A^c
2. $A \cap B$
3. $A \cup B$
4. $A \cap B^c$
5. $(A \cup B) \cap C$
6. $A \cup (B \cap C)$

De lege verzameling \emptyset is de verzameling zonder elementen. Twee verzamelingen heten *disjunct* als hun doorsnede leeg is. Dat wil zeggen, dat de twee verzameling geen elementen gemeenschappelijk hebben. Als een van beide optreedt, kan de andere niet ook optreden.

2.3 kans

Een kansfunctie P op een uitslagen ruimte Ω kent aan iedere deelverzameling van Ω een getal tussen 0 en 1 toe, zodat

1. $P(\emptyset) = 0$
2. $P(A) \geq 0$ voor alle $A \subseteq \Omega$
3. $P(A \cup B) = P(A) + P(B)$, als $A \cap B = \emptyset$

Het volgt dat meer in het algemeen moet gelden

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B).}$$

Er volgt ook dat $P(\Omega) = 1$. Aangezien $A^c \cup A = \Omega$ volgt ook $P(A^c) = 1 - P(A)$.

opgave 2.2 Stel, $P(A) = 0.4$, $P(B) = 0.7$ en $P(A \cap B) = 0.3$. Bepaal de kans dat A of B gebeurt.

opgave 2.3 Stel, $P(A) = 0.4$, $P(B) = 0.5$ en $P(A \cap B) = 0.1$. Bepaal de kans dat A of B gebeurt, maar niet beide.

opgave 2.4 Je hebt de finale van de grote spelshow bereikt, en Willem Ruis geeft je de keuze uit drie gesloten deuren. Achter één deur staat de auto. Je kiest willekeurig een deur, maar Ruis maakt die nog niet open. In plaats daarvan maakt hij een van de andere deuren open, en laat zien dat de prijs daar in elk geval niet staat. Er zijn nu nog twee dichte deuren over. Blijf je bij je oorspronkelijke keuze, of heeft het zin om nu die andere deur te kiezen?

Met het onderstaande script kun je het spel 1000 keer spelen.

```

deur = 1:3
wissel = F                                #strategie: wisselen of niet?
count = 0
for (i in 1:1000){
  auto = sample(deur,1)                    #auto achter willekeurige deur
  kies = sample(deur,1)                    #kies een willekeurige deur
  ruis = deur[-c(kies,auto)][1]            #ruis opent een andere deur, zonder auto
  if (wissel){
    kies = deur[-c(kies,ruis)]
  }
  count = count+(kies==auto)
}
cat(count,"keer gewonnen uit 1000 spellen.\n")

```

2.4 onafhankelijkheid

Twee gebeurtenissen heten *disjunct* als $A \cap B = \emptyset$. Disjuncte gebeurtenissen kunnen nooit tegelijk optreden en we weten

$$P(A \cup B) = P(A) + P(B), \quad \text{als } A \cap B = \emptyset.$$

Twee gebeurtenissen heten *onafhankelijk* als

$$P(A \cap B) = P(A)P(B).$$

Onafhankelijkheid is iets HEEL anders dan disjunctie, maar toch worden de twee vaak verward.

opgave 2.5 Stel $P(A) = 0.1$, $P(B) = 0.2$, $P(C) = 0.5$ en stel dat A, B en C onafhankelijk zijn. Wat is de kans dat tenminste één van de gebeurtenissen optreedt?

opgave 2.6 Werp eenmaal met een dobbelsteen. Zijn de gebeurtenissen $A = \{2, 4, 6\}$ en $B = \{1, 2\}$ afhankelijk?

opgave 2.7 Werp eenmaal met een dobbelsteen. Zijn de gebeurtenissen $A = \{2, 4, 6\}$ en $B = \{1, 2, 5\}$ afhankelijk?

2.5 voorwaardelijke kansen

opgave 2.8 Van een groep van duizend willekeurige mensen is vastgesteld of ze roken (of ooit gerookt hebben) en of ze longkanker hebben.

	longkanker	geen longkanker
roker	240	60
niet-roker	20	680

1. Wat is de kans dat een willekeurig persoon van deze groep longkanker heeft?
2. Wat is de kans dat een willekeurig persoon van deze groep longkanker heeft en rookt?
3. Gegeven dat hij of zij rookt, wat is de kans dat een willekeurig persoon van deze groep longkanker heeft?

De *voorwaardelijke kans* dat A optreedt als gegeven is dat B is opgetreden is

$$P(A | B) = \frac{P(A \cap B)}{P(B)}, \quad \text{mits } P(B) > 0.$$

opgave 2.9 Relateer deze definitie aan de voorgaande opgave.

opgave 2.10 Stel je werpt met een dobbelsteen. Bereken steeds $P(A | B)$:

1. $A = \{1, 2, 3\}$ en $B = \{4\}$.
2. $A = \{1, 2, 3\}$ en $B = \{2\}$.
3. $A = \{1, 2, 3\}$ en $B = \{1, 2, 3\}$.
4. $A = \{1, 2, 3\}$ en $B = \{1, 2, 3, 4\}$.

Als A en B onafhankelijk zijn, dan geldt $P(A \cap B) = P(A)P(B)$ en dus $P(A | B) = P(A)$. Met andere woorden, het al dan niet optreden van B verandert niets aan de kans dat A gebeurt.

De regel van totale waarschijnlijkheid is

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A | B)P(B) + P(A | B^c)P(B^c). \end{aligned}$$

In de volgende opgave pas je hem toe.

opgave 2.11 Een zeldzame ziekte komt voor bij 1 op de duizend personen. Je kunt laten testen of je de ziekte hebt. Als je de ziekte hebt, geeft de test dat in 99% van de gevallen correct aan. Helaas geeft de test ook een positief resultaat bij 0.2% van de mensen die gezond zijn. Wat is de kans dat een persoon de ziekte heeft, gegeven een positief test resultaat?

opgave 2.12 Het antwoord van de vorige opgave heeft je misschien verrast. De kans dat de persoon de ziekte heeft is erg klein, ook al heeft hij een positief test resultaat. Kun je dit verklaren?

2.6 combinatoriek

opgave 2.13 Schrijf alle manieren op om de getallen 1,2 en 3 te ordenen.

opgave 2.14 Schrijf alle manieren op om twee getallen uit de getallen 1, 2, ..., 5 te kiezen.

Bij het berekenen van kansen speelt combinatoriek soms een rol. Het aantal verschillende manieren waarop je de getallen 1, 2, ..., n kunt sorteren is $n(n-1)(n-2)\dots 1 = n!$. Spreek uit: n faculteit.

Het aantal manieren waarop je k elementen uit een verzameling van n elementen kunt kiezen is

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

spreek uit: n boven k .

opgave 2.15 Controleer je antwoorden op de vorige twee vragen met behulp van de bovenstaande formules.

opgave 2.16 Je gooit 6 keer met een munt en krijgt 4 keer kop. Op hoeveel manieren kan dat gebeuren? Stel je krijgt niet 4, maar 2 keer kop. Op hoeveel manieren kan dat gebeuren?

opgave 2.17 Je gooit 6 keer met een munt met kans p op kop. Wat is de kans dat je 4 keer kop krijgt?

opgave 2.18 Wat is de kans dat in een groep van 50 personen, tenmiste twee dezelfde verjaardag hebben. Neem aan dat ieder jaar 365 dagen heeft, en met “dezelfde verjaardag” niet ook hetzelfde geboortjaar bedoeld wordt.

2.7 discrete kansverdelingen

Zij X de numerieke waarde die we toekennen aan de uitslag van een experiment. We noemen X een *stochastische grootheid*. X neemt verschillende waarden aan met verschillende kansen.

Iedere rij niet negatieve getallen die samen tot 1 optellen is een (discrete) *kansverdeling*. Er zijn dus oneindig veel (discrete) kansverdelingen. Een aantal bekende voorbeelden zoals de Binomiale, Geometrische en Poisson verdeling, staat in de bijlage **kansverdelingen**.

Merk op dat verschillende verdelingen worden vastgelegd door verschillende *parameters*. De Binomiale verdelingen heeft twee parameters: n en p . De Exponentiele verdeling heeft er maar één: λ .

De *verdelingsfunctie* van een stochastische grootheid X is gedefinieerd als

$$F(x) = P(X \leq x).$$

De verdelingsfunctie is een stijgende functie van 0 naar 1.

Voor $0 \leq \alpha \leq 1$ is het α kwantiel gedefinieerd als de kleinste x zodat $P(X \leq x) \geq \alpha$. Het 1/2 kwantiel heet ook wel de mediaan. Kwantielen spelen een belangrijke rol in hoofdstuk 5.

In R zijn de meeste kansverdelingen voorgeprogrammeerd. Elke kansverdeling in R heeft een naam: `binom`, `geom` en `pois`. Voor de naam van de verdeling staat altijd 1 van de volgende 4 letters.

- `d` staat voor “density”. Dit is de kansverdeling zelf.
- `p` staat voor “probability”. Dit is de verdelingsfunctie.
- `q` staat voor “quantile”. Dit is de kwantiel functie.
- `r` staat voor “random”. Hiermee kunnen we een toevallige steekproef genereren.

We maken een plot van de Binomiaal(20,1/3) verdeling

```
x = 0:20
plot(x,dbinom(x,20,1/3),type='s')
```

opgave 2.19 Maak ook plots van de Geom(1/5) en Poisson(3) verdelingen.

Bij de volgende opgaven kun je gebruik maken van de bijlage **kansverdelingen**.

opgave 2.20 Stel je gooit net zolang met een munt met kopkans $p = 1/5$ tot je voor de eerste keer kop krijgt. Zij X het aantal worpen dat je hebt gedaan. Wat is $P(X = 5)$? Controleer je antwoord met de R functie `dgeom()`. Let wel op dat R een iets andere definitie van de geometrische verdeling heeft. Wij hebben de verdeling gedefinieerd op $k = 1, 2, 3 \dots$ maar zij op $k = 0, 1, 2, \dots$

opgave 2.21 Stel je gooit 5 keer met een munt met kopkans $p = 1/3$. Zij X het aantal keer dat je kop hebt gegooit. Wat is $P(X = 3)$? Controleer je antwoord met `dbinom()`. Wat is $P(X \leq 3)$? Controleer je antwoord zowel met `dbinom()` als met `pbinom()`

opgave 2.22 Genereer een steekproef van grootte 10000 uit de Binomiale verdeling met parameters $n = 5$ en $p = 1/3$. Tel het aantal keer dat je $X = 3$ hebt gekregen en deel dit aantal door 10000. Vergelijk met het antwoord van de vorige vraag. Wat valt je op? Begrijp je wat je gedaan hebt?

opgave 2.23 Maak plots van de verdelingsfuncties van Binomiaal(10,1/3), Geom(1/5) en Poisson(3) verdelingen. Gebruik de functies `pbinom()`, `pgeom()` en `ppois()` en de plot optie `type='s'`.

2.8 continue kansverdelingen

Een stochastische grootte heeft een continue kansverdeling als er een functie f bestaat zodat

1. $f(x) \geq 0$, voor alle x
2. $\int_{x=-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)dx$

De functie f heet de (kans)dichtheid. Een aantal bekende continue verdelingen zijn gegeven in de bijlage **kansverdelingen**.

De verdelingsfunctie van een stochastische grootte X met een continue verdeling is gedefinieerd als

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y)dy$$

Er geldt dus ook dat $f(x) = F'(x)$. De verdelingsfunctie is een stijgende functie van 0 naar 1.

Bij de volgende opgaven kun je de bijlage **kansverdelingen** gebruiken.

opgave 2.24 Stel X heeft de uniform(1,3) verdeling. Bepaal $P(X > 2.3)$ door de dichtheid te integreren. Controleer je antwoord met `punif()`.

opgave 2.25 Stel X heeft de Exponentieel(3) verdeling. Bepaal $P(1.3 < X < 4.3)$ door de dichtheid te integreren. Controleer je antwoord met `pexp()`.

opgave 2.26 Bepaal de verdelingsfuncties van de Uniform(1,3) en de exponentieel(3) verdeling door de dichtheden te integreren. Controleer je antwoord door deze functies te plotten:

```
x = seq(-1,4,by=0.01)
plot(x,punif(x,1,3),type='l')
lines(x,pexp(x,3))
```

Merk op: In de bijlage **kansverdelingen** staat dat de twee parameters van de normale verdeling μ en σ^2 zijn. De R functie `dnorm` voor de dichtheid van de normale verdeling heeft echter als parameters μ en $\sigma = \sqrt{\sigma^2}$.

opgave 2.27 Plot de dichtheden van de volgende normale verdelingen in één plaatje. Maak gebruik van de functies `plot()`, `lines()` en `dnorm()`.

- $\mu = 0$ en $\sigma = 1$
- $\mu = 1$ en $\sigma = 0.5$
- $\mu = 0$ en $\sigma = 3$
- $\mu = -2$ en $\sigma = 1$

Wat is het effect van het veranderen van de twee parameters μ en σ van de normale verdeling?

opgave 2.28 Stel X heeft de standaard normale verdeling (dwz $\mu = 0$, $\sigma^2 = 1$). Gebruik `qnorm()` om x te bepalen zodat $P(X \leq x) = 0.05$. Controleer je antwoord met `pnorm()`.

2.9 histogram

Een steekproef is een aantal onafhankelijke trekkingen uit een kansverdeling. Een goede manier om de steekproef te visualiseren is het histogram. Het histogram verdeelt de x -as in een aantal intervallen, en telt hoeveel van de trekkingen in de steekproef in elk van de intervallen liggen.

opgave 2.29 we gebruiken de functie `rnorm` om een steekproef van 100 trekkingen uit de Normal(10,1) verdeling te trekken. Vervolgens maken we histogrammen met de functie `hist()`. Het commando `par(mfrow=c(3,2))` geeft aan dat het plaatje in 3 bij 2 delen moet worden verdeeld.

```
X = rnorm(100,10,1)
par(mfrow=c(3,2))
hist(X,breaks=5)
hist(X,breaks=10)
hist(X,breaks=25)
hist(X,breaks=50)
hist(X,breaks=75)
hist(X)
```

Wat is het effect van de parameter `breaks`?

opgave 2.30 Voer het volgende script uit, en denk na over wat je ziet.

```
x = rnorm(100,10,1)
hist(x,probability=T,xlim=c(5,15),ylim=c(0,0.4))
t = seq(5,15,by=0.1)
lines(t,dnorm(t,10,1))
```

De optie `probability=T` zorgt ervoor dat het histogram wordt geschaald zodat het oppervlak onder de grafiek gelijk is aan 1—net als bij een dichtheid.

opgave 2.31 Het volgende script genereert 6 maal 20 trekkingen uit de Binomiaal(10,0.25) verdeling. Voer het script uit

```

par(mfrow=c(3,2))
for (i in 1:6){
  x = rbinom(20,10,0.25)
  hist(x,breaks=0:10)
}

```

Voer het bovenstaande script nogmaals uit, maar dan met steekproefgrootte 50. Voer het dan nogmaals uit met steekproefgrootte 100, en dan met steekproefgrootte 1000. Wat valt je op?

2.10 Monte Carlo

Opgave 2.24 luidde: Stel X heeft de uniform(1,3) verdeling. Bepaal $P(X > 2.3)$. Met behulp van R (of natuurlijk pen en papier) is deze kans eenvoudig te bepalen:

```
1-punif(2.3,1,3)
```

Een andere methode is Monte Carlo: Neem een grote steekproef uit de uniform(1,3) verdeling en bepaal het percentage dat groter is dan 2.3.

```

x = runif(100,1,3)
p = sum(x>2.3)/100

```

opgave 2.32 Voer het bovenstaande script een aantal maal uit, en noteer welke waarden je voor p vind. Voer het script nog een aantal keer uit, maar dan met steekproefgrootte 1000, en tenslotte 10^6 .

opgave 2.33 Benader met behulp van de Monte Carlo methode de kans dat in een groep van 50 personen tenminste twee dezelfde verjaardag hebben. Om een steekproef van 50 verjaardagen te genereren gebruik je

```
x = sample(1:365,50,replace=TRUE) #let op: met teruglegging!
```

Om te bepalen of er dubbele verjaardagen zijn gebruik je de boolese uitdrukking

```
(length(unique(x))<50)
```


Hoofdstuk 3

Locatie en spreiding

3.1 verwachting

De *verwachting* is een maat voor het “midden” ofwel de locatie van een kansverdeling. De verwachting van een stochastische grootte X is

$$EX = \begin{cases} \sum_x xP(X = x), & \text{als } X \text{ een discrete verdeling heeft} \\ \int_x xf(x)dx, & \text{als } X \text{ een continue verdeling heeft} \end{cases}$$

Er geldt

1. $E(X + Y) = E(X) + E(Y)$
2. $E(aX + b) = aEX + b$, voor alle $a, b \in \mathbb{R}$

Zij $Y = g(X)$, dan geldt

$$EY = \begin{cases} \sum_x g(x)P(X = x), & \text{als } X \text{ een discrete verdeling heeft} \\ \int_x g(x)f(x)dx, & \text{als } X \text{ een continue verdeling heeft} \end{cases}$$

opgave 3.1 Stel $P(X = 1) = 1/3$, $P(X = 2) = 1/4$ en $P(X = 3) = 5/12$

1. Wat is de verwachting van X ?
2. Wat is de verwachting van $Y = 3X + 24$?
3. Wat is de verwachting van $Z = X^2$? Let op EX^2 is **niet** gelijk aan $(EX)^2$.

opgave 3.2 X is verdeeld volgens de Bernoulli(p) verdeling als $P(X = 1) = p$ en $P(X = 0) = 1 - p$. Wat is EX ?

opgave 3.3 Stel X_1, X_2, \dots, X_n zijn onafhankelijk verdeeld volgens de Bernoulli(p) verdeling. Nu geldt dat $X = \sum_i X_i$ de Binomiaal(n, p) verdeling heeft (waarom?). Wat is EX ?

opgave 3.4 We kunnen de verwachting van de binomiale verdeling met parameters $n = 20$ en $p = 1/4$ in R bepalen door middel van

```
x = 0:20
EX = sum(x*dbinom(x,20,1/4))
```

Gebruik R om de verwachting te bepalen van de Poisson verdeling met parameter $\lambda = 4$. controleer je antwoord met de bijlage **kansverdelingen**.

Voor de verwachtingen van de andere bekende verdelingen, zie de bijlage **kansverdelingen**.

3.2 variantie

De *variantie* is een maat voor de spreiding van een kansverdeling. De variantie van een stochastische grootte X is

$$\text{Var}(X) = E(X - EX)^2 = \begin{cases} \sum_x (x - EX)^2 P(X = x), & \text{als } X \text{ een discrete verdeling heeft} \\ \int_x (x - EX)^2 f(x) dx, & \text{als } X \text{ een continue verdeling heeft} \end{cases}$$

Er geldt

$$\boxed{\text{var}(X) = EX^2 - (EX)^2.}$$

Ook geldt

1. Als X en Y onafhankelijk zijn, dan $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
2. $\text{var}(aX + b) = a^2\text{var}(X)$, voor alle $a, b \in \mathbb{R}$

Als X en Y niet onafhankelijk zijn, dan geldt 1. **niet**. We komen hier later nog op terug.

De *standaard afwijking* van X is gedefinieerd als $\text{std}(X) = \sqrt{\text{var}(X)}$.

opgave 3.5 Stel $P(X = 1) = 1/3$, $P(X = 2) = 1/4$ en $P(X = 3) = 5/12$

1. Wat is de variantie van X ?
2. Wat is de variantie van $Y = 3X + 24$?

opgave 3.6 Bepaal de verwachting en variantie van de stochastische grootte $Z = (X - EX)/\text{std}(X)$.

opgave 3.7 Stel X is verdeeld volgens de Bernoulli(p) verdeling. Wat is $\text{var}(X)$? Hint: bereken eerst EX^2 .

opgave 3.8 Stel X_1, X_2, \dots, X_n zijn onafhankelijk verdeeld volgens de Bernoulli(p) verdeling. Nu geldt dat $X = \sum X_i$ de Binomiaal(n, p) verdeling heeft. Wat is $\text{var}(X)$?

Voor de varianties van de andere bekende verdelingen, zie de bijlage **kansverdelingen**.

3.3 gemiddelde

Stel X_1, X_2, \dots, X_n is een steekproef van onafhankelijke stochastische grootheden, elk met dezelfde verwachting $EX = \mu$ en variantie $\text{var}(X) = \sigma^2$. Het gemiddelde

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

is weer een stochastische grootheid, met zijn eigen verwachting en variantie. Als we de regels voor de verwachting toepassen, dan vinden we

$$E\bar{X} = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

opgave 3.9 Trek een steekproef ter grootte 100 uit de normale verdeling met verwachting $\mu = 5$ en standaard afwijking $\sigma = 2$. Bepaal het gemiddelde.

opgave 3.10 Stel \bar{X} is het gemiddelde van n onderling onafhankelijk stochastische grootheden, elk met dezelfde verwachting μ en variantie σ^2 . Wat is de variantie van \bar{X} ?

opgave 3.11 Met het onderstaande script berekenen we 100 keer het gemiddelde van een steekproef X_1, X_2, \dots, X_{10} uit de Binomiaal(20,1/4) verdeling. Voer het script een aantal maal uit. Het is de bedoeling dat het echt tot je doordringt dat \bar{X} een stochastische grootheid is, met zijn eigen kansverdeling

```
barX = 1:100
for (i in 1:100){
  X = rbinom(10,20,1/4)
  barX[i] = mean(X)
}
hist(barX,xlim=c(0,10))
```

De verwachting van \bar{X} is μ , en hangt dus niet af van de steekproefgrootte n . De variantie wel! Pas het script aan, en bereken 100 keer het gemiddelde van een steekproef $X_1, X_2, \dots, X_{1000}$ uit de binomiaal(20,1/4) verdeling. Wat valt je op?

Na de vorige opgave begrijp je (hopelijk) wat wordt bedoeld met *De Wet van de Grote Aantallen*:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow EX, \quad \text{als } n \rightarrow \infty.$$

opgave 3.12 Stel X heeft de normale verdeling met verwachting $\mu = 5$ en variantie $\sigma^2 = 4$. Benader EX^2 met behulp van de Monte Carlo methode.

opgave 3.13 Er geldt $\text{var}(X) = EX^2 - (EX)^2$. Stel X heeft de normale verdeling met verwachting $\mu = 5$ en variantie $\sigma^2 = 4$. Wat is EX^2 ?

opgave 3.14 Bepaal de verwachting en variantie van de stochastische grootheid

$$Z = \sqrt{n} \frac{\bar{X} - EX}{\text{std}(X)}.$$

3.4 steekproef variantie

Stel X_1, X_2, \dots, X_n is een steekproef van onafhankelijke stochastische grootheden, elk met dezelfde verwachting $EX = \mu$ en variantie $\text{var}(X) = \sigma^2$.

De *steekproefvariantie* is gedefinieerd als

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

S^2 is een stochastische grootheid, met zijn eigen verwachting en variantie. Er is een goede reden om door $n-1$ te delen, in plaats van door n , maar daar gaan we hier aan voorbij. Er geldt

$$ES^2 = \sigma^2,$$

en als de steekproefgrootte heel groot wordt, dan geldt

$$S_n^2 \rightarrow \sigma^2, \quad \text{als } n \rightarrow \infty.$$

opgave 3.15 Trek een steekproef ter grootte 100 uit de normale verdeling met verwachting $\mu = 5$ en standaard afwijking $\sigma = 2$. Bepaal steekproef variantie.

opgave 3.16 Voer het volgende script uit om 100 keer de steekproefvariantie van een steekproef X_1, X_2, \dots, X_{10} uit de normale(5,9) verdeling te bepalen.

```
S2 = 1:100
for (i in 1:100){
  X = rnorm(10,5,3)
  S2[i] = var(X)
}
hist(S2,xlim=c(0,20))
```

Vergroot nu de grootte van de steekproef van 10 naar 1000 en voer het script nog eens uit.

3.5 covariantie

De volgende data komt uit de *Family Expenditure Survey*, Department of Employment, 1981 (British official statistics). In 10 gedeelten van Engeland heeft men de gemiddelde wekelijkse uitgave per gezin aan alcohol en tabak bepaald (in engelse ponden).

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76

Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51

opgave 3.17 We definiëren variabelen x (alcohol) en y (tabak). Voer de volgende commando's uit

```
x = c(6.47,6.13,6.19,4.89,5.63,4.52,5.89,4.79,5.27,6.08)
y = c(4.03,3.76,3.77,3.43,3.47,2.92,3.20,2.71,3.53,4.51)
plot(x,y,xlab="alcohol uitgave",ylab="tabak uitgave")
```

Wat valt je op?

Het is wel duidelijk dat er een verband bestaat tussen X en Y , maar hoe kunnen we dat kwantificeren? De *covariantie* tussen twee stochastische grootheden X en Y is gedefinieerd als

$$\text{cov}(X, Y) = E((X - EX)(Y - EY)).$$

Dus $\text{cov}(X, X) = \text{var}(X)$. De *steekproef covariantie* is

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

Als de steekproef covariantie groot en positief is, dan gaan x -waarden die groter zijn dan het x -gemiddelde kennelijk vaak samen met y -waarden die groter zijn dan het y -gemiddelde. En vice versa.

Als de steekproef covariantie groot en negatief is, dan gaan x -waarden die groter zijn dan het x -gemiddelde kennelijk vaak samen met y -waarden die *kleiner* zijn dan het y -gemiddelde. En vice versa.

opgave 3.18 Bepaal de steekproef covariantie voor de alcohol en tabak data. Gebruik de R functie `cov()`.

opgave 3.19 Definieer de volgende variabelen

```
a = c(6.3,4.5,6.4,4.0,2.8,5.8,5.2,4.9,5.3,4.6)
b = c(1.8,3.0,1.6,3.1,3.5,1.8,2.3,2.7,2.2,3.0)
c = c(1.9,2.4,1.9,2.7,2.7,2.0,2.1,2.4,2.1,2.4)
d = c(6.4,5.4,4.4,3.5,4.3,5.3,5.9,4.6,5.5,5.9)
```

Maak plots `plot(a,b)`, `plot(b,c)` en `plot(a,d)` en bepaal steeds de steekproef covariantie met behulp van de functie `cov()`. Zie je het verband?

opgave 3.20 Stel dat we de uitgaven aan tabak en alcohol in euros hadden gemeten in plaats van ponden.

```
xeuro = x*1.45678
yeuro = y*1.45678
```

Bepaal de steekproef covariantie tussen `xeuro` en `yeuro`. Vind je dezelfde covariantie als bij opgave 3.18?

In het algemeen geldt

$$\text{cov}(aX, bY) = ab \text{cov}(X, Y), \quad \text{voor alle } a, b \in \mathbb{R}$$

We weten al dat als X en Y onafhankelijk zijn, dan

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

In het algemeen (dus ook als X en Y niet onafhankelijk zijn) geldt

$$\boxed{\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y).}$$

Als X en Y onafhankelijk zijn dan is $\text{cov}(X, Y) = 0$.

3.6 correlatie

Aangezien $\text{cov}(aX, bY) = ab \text{cov}(X, Y)$ is de covariantie niet te gebruiken als een absolute maat voor de sterkte van het verband tussen twee stochastische grootheden. Immers, de covariantie hangt af van de eenheden waarin we meten. De *correlatie coefficient* is wèl een geschikte, absolute maat voor de sterkte van het verband tussen twee stochastische grootheden.

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\text{std}(X)\text{std}(Y)}.$$

Er geldt $-1 \leq \rho \leq 1$. De *steekproef correlatie* coefficient is gelijk aan de steekproef covariantie gedeeld door het product van de steekproef standaard afwijkingen.

opgave 3.21 Zij \mathbf{a} , \mathbf{b} , \mathbf{c} en \mathbf{d} als in opgave 16. Maak nogmaals de plots en Gebruik de functie `cor()` om de steekproef correlatie coefficienten $\rho(\mathbf{a}, \mathbf{b})$, $\rho(\mathbf{b}, \mathbf{c})$ en $\rho(\mathbf{a}, \mathbf{d})$ te bepalen.

opgave 3.22 Bepaal de correlatie tussen de uitgaven aan tabak en alcohol in ponden en in euros.

In de tabel staan de verwachting, variantie, covariantie en correlatie nog eens bij elkaar.

kansverdeling	steekproef
EX	$\bar{X} = \sum_i X_i/n$
$\text{var}(X)$	$S^2 = \sum_i (X_i - \bar{X})^2/(n-1)$
$\text{cov}(X, Y)$	$S_{XY}^2 = \sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)$
$\rho(X, Y)$	$S_{XY}^2/S_X S_Y$

3.7 robuustheid: mediaan, MAD en uitbijters

opgave 3.23 De verwachting, variantie, covariantie en correlatie zijn erg gevoelig voor “foute” waarnemingen. Stel dat je de valversnelling een aantal maal empirisch hebt vastgesteld (met een kleine meetonzekerheid). Bij het invoeren van de data vergeet je echter een keer een punt.

$g = c(9.815, 9.814, 9.798, 9.812, 9.811, 9.808, 9812)$

Bepaal het gemiddelde en steekproef standaardafwijking met en zonder de laatste waarde.

De *mediaan* is net zoiets als het gemiddelde. Het is ook een maat voor de locatie of het centrum van de steekproef. Om de mediaan te bepalen, sorteren we eerst de waarnemingen op grootte. Als er een oneven aantal waarnemingen is, dan is de mediaan gedefinieerd als de middelste waarneming. Als er een even aantal waarnemingen is, dan is de mediaan het gemiddelde van de middelste twee waarnemingen. De mediaan is veel minder gevoelig voor foute waarden. Men zegt wel dat de mediaan *robust* is tegen fouten in de data.

opgave 3.24 Bepaal de mediaan van de valversnellingsdata, met en zonder de laatste waarneming.

We zagen dat ook de variantie (en dus ook de standaardafwijking) erg gevoelig is voor fouten. De *MAD* (median absolute deviation) is een robuuste maat voor spreiding.

Als $\text{med} = \text{mediaan}(X_1, X_2, \dots, X_n)$, dan

$$MAD(X_1, X_2, \dots, X_n) = \text{mediaan}(|X_1 - \text{med}|, |X_2 - \text{med}|, \dots, |X_n - \text{med}|).$$

opgave 3.25 Bepaal de *MAD* van de valversnellingsdata, met en zonder de laatste waarneming.

Abnormale waarnemingen hoeven niet altijd fout te zijn. Het is wel altijd nodig om nader onderzoek in te stellen. Bij een onderzoek naar een bloeddruk verlagend medicijn reageren bijna alle proefpersonen met een licht verlaagde bloeddruk. Eén persoon blijkt echter allergisch te zijn voor een bestanddeel van het medicijn, raakt in shock en krijgt een gevaarlijk verlaagde bloeddruk. In dit geval is een abnormale waarde dus essentieel voor het onderzoek.

We kunnen de volgende regel gebruiken om een abnormale waarde of *uitbijter* te detecteren: X_i is een uitbijter in de steekproef X_1, X_2, \dots, X_n als $|X_i - \text{med}| > 5 MAD$.

opgave 3.26 Gebruik de bovenstaande regel om na te gaan of er uitbijters zitten tussen de valversnellingsdata.

opgave 3.27 Een andere, veel vaker gebruikte regel zegt dat X_i een uitbijter is als $|X_i - \bar{X}| > 3S$, waarbij S de steekproef standaard afwijking is. Probeer deze regel uit met de valversnellingsdata. Begrijp je wat er aan de hand is?

Hoofdstuk 4

Schatten

4.1 parameter schatten

Stel X_1, X_2, \dots, X_n is een onafhankelijke steekproef uit een normale kansverdeling. We weten (zie de bijlage **kansverdelingen**) dat de normale kansverdeling wordt bepaald door twee parameters: de verwachting $EX = \mu$ en de variantie $\text{var}(X) = \sigma^2$. Op basis van de steekproef kunnen we de parameters *schatten* met het gemiddelde $\hat{\mu} = \bar{X}$ en de steekproef standaardafwijking $\hat{\sigma}^2 = S^2$. Met een hoedje geven we aan dat het om een schatter gaat.

In de bijlage **kansverdelingen** staan schatters voor de parameters van een aantal verdelingen.

opgave 4.1 Genereer een steekproef X_1, X_2, \dots, X_{10} uit de normale verdeling met parameters $\mu = 8$ en $\sigma = 3$. Plot in één grafiek de dichtheidsfunctie van de normaal(8,3) en de normaal(\bar{X}, S) verdeling. Herhaal tenminste 10 keer.

opgave 4.2 Genereer een steekproef X_1, X_2, \dots, X_8 uit de exponentiële verdeling met parameter $\lambda = 0.3$. Plot in één grafiek de ware en de geschatte dichtheidsfunctie. Herhaal tenminste 10 keer.

Een schatter is een stochastische grootheid, want de waarde hangt af van het toeval. Een schatter heeft dus zelf ook een kansverdeling met een verwachting en een variantie. In de volgende opgave

bekijken we de kansverdeling van twee schatters.

opgave 4.3 Stel X_1, \dots, X_{13} is een onafhankelijke steekproef uit de normale verdeling met parameters $\mu = 10$ en $\sigma = 5$. We genereren 25 keer een steekproef en berekenen steeds het gemiddelde en de steekproefvariantie S^2 . voer het volgende script uit.

```
mu = 1:100
S2 = 1:100
for (i in 1:100){
  X = rnorm(13,10,5)
  mu[i] = mean(X)
  S2[i] = var(X)
}
```

Maak histogrammen van de vectoren `mu` en `S2`. Gebruik `par(mfrow=c(2,1))`.

Een schatter T van een parameter θ heet zuiver (unbiased) als $ET = \theta$. Stel X_1, X_2, \dots, X_n zijn onafhankelijke stochastische grootheden, elk met verwachting μ en variantie σ^2 . We weten al dat $E\bar{X} = \mu$. Het gemiddelde is dus een zuivere schatter van de verwachting. Men kan ook laten zien (dat doen wij hier niet) dat de steekproef variantie S^2 een zuivere schatter is voor σ^2 .

Zuiverheid is een goede eigenschap van een schatter, maar een kleine variantie is dat ook.

opgave 4.4 Stel dat de vijand $N = 5000$ tanks heeft, met serienummers $1, 2, \dots, N$. Wij weten niet wat N is, maar onze troepen hebben inmiddels 10 tanks buitgemaakt, met nummers X_1, X_2, \dots, X_{10} . De heer J. Bond van de inlichtingendienst stelt

$$T_1 = 2\bar{X} - 1$$

voor als schatter voor N . Een medewerker van afdeling Q meent dat Bond zich beter met vuurgevechten en vrouwen versieren kan bezighouden, dan met statistiek. Q stelt

$$T_2 = \frac{11}{10} \max(X_1, X_2, \dots, X_{10}) - 1$$

voor als schatter. T_1 en T_2 zijn beide zuivere schatters. Met behulp van een Monte Carlo experiment, kunnen we inzicht krijgen van het gedrag van de twee schatters.

```

T1 = 1:1000
T2 = 1:1000
for (i in 1:1000){
  X = sample(1:5000,10,replace=F) # trek 10 uit N=5000 tanks, zonder teruglegging
  T1[i] = 2*mean(X)-1 # bereken T1
  T2[i] = (11/10)*max(X)-1 # bereken T2
}

```

1. Maak histogrammen van de vectoren T1 en T2. Gebruik `par(mfrow=c(2,1))`.
2. Bepaal de gemiddelden en de standaard afwijkingen van de vectoren T1 en T2.
3. Welke schatter is beter?

Een goede maat voor de kwaliteit van een schatter T voor de parameter θ is de verwachte kwadratische fout (mean squared error)

$$\text{MSE}(T) = E(T - \theta)^2 = \text{var}(T) + (ET - \theta)^2.$$

De MSE kijkt dus zowel naar de zuiverheid, als de variantie.

opgave 4.5 Stel X_1, X_2, \dots, X_{20} zijn onafhankelijk verdeeld volgens de Poisson(λ) verdeling met $\lambda = 2$. Het volgt (zie kansverdelingen) dat $P(X = 0) = e^{-\lambda}$. Stel dat λ ons onbekend is, en we deze kans willen schatten. Twee schatters liggen voor de hand:

$$T_1 = \frac{\#\{X_i = 0\}}{n} \quad \text{en} \quad T_2 = e^{-\bar{X}}.$$

1. Waarom “liggen deze twee schatters voor de hand”?
2. T_1 is een zuivere schatter, maar T_2 is dat niet. Voer een Monte Carlo experiment uit in de stijl van de vorige opgave. Maak histogrammen, en bereken de gemiddelden, de steekproef varianties en de MSE van beide schatters. Welke van de twee schatters vind je beter?

4.2 lineaire regressie

De volgende data komt uit de *Family Expenditure Survey*, Department of Employment, 1981 (British official statistics). In 10 gedeelten van Engeland heeft men de gemiddelde wekelijkse uitgaven per gezin aan alcohol en tabak bepaald (in engelse ponden). We definiëren variabelen x (uitgave aan alcohol) en y (uitgave aan tabak).

```
x = c(6.47,6.13,6.19,4.89,5.63,4.52,5.89,4.79,5.27,6.08)
y = c(4.03,3.76,3.77,3.43,3.47,2.92,3.20,2.71,3.53,4.51)
plot(x,y,xlab="alcohol uitgave",ylab="tabak uitgave")
```

In de plot zien we dat er wel eens een lineair (rechtevenredig) verband zou kunnen bestaan tussen X en Y . We stellen het volgende *model* op voor de uitgaven aan tabak

$$Y = a + bX + \varepsilon.$$

Het lineaire deel $a + bX$ stelt een recht lijn voor, terwijl ε de toevallige verstoring of meetfout is. We nemen aan dat ε normaal verdeeld is met verwachting 0 en onbekende variantie σ^2 .

opgave 4.6 Welke rechte lijn past het beste bij de data? Gebruik je timmermansoog om geschikte waarden voor a en b te vinden. Gebruik de functie `abline(a,b)` om rechte lijnen door je plot te trekken.

Je timmermansoog is natuurlijk niet echt exact. De absoluut *beste* rechte lijn wordt bepaald door die a en b te vinden zodat de kwadraat som (*sum of squares*)

$$SS = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

zo klein mogelijk is. Dit heet de methode van de kleinste kwadraten (KK).

opgave 4.7 Probeer te begrijpen waarom SS minimaliseren een geschikte rechte lijn zou opleveren.

R heeft een functie `lm()` (staat voor: linear model) om de methode van de kleinste kwadraten uit te voeren. Het R commando

```
fit = lm(y ~ x)
```

levert een R “object” op. Dit object heeft een aantal “attributen”. Het belangrijkste attribuut is “coefficients”; de geschatte coëfficiënten a en b van de rechte lijn. Met behulp van het `$` teken krijgen we deze coëfficiënten te zien:

```
coef = fit$coefficients
```

Met andere woorden, we kennen de coëfficiënten van de kleinste kwadraten fit toe aan de variabele `coef`. Merk op dat `coef` een vector is met twee elementen. We kunnen nu als volgt de beste rechte lijn door de data trekken

```
plot(x,y)
abline(coef,col='red')
```

opgave 4.8 Voer de bovenstaande commando's uit, en vergelijk de optimale waarden voor a en b met de waarden die je met je timmermansoog vond.

De kleinste kwadraten schatters voor de coëfficiënten a en b zijn zuiver. Hoe nauwkeurig zijn ze? De functie `summary.lm()` geeft de standaard afwijking van onze schatters

```
summary.lm(fit)
```

opgave 4.9 Wat zijn de standaard afwijkingen van de geschatte a en b ?

4.3 lineaire modellen van hogere orde

Een rechte lijn is niet altijd een goede beschrijving van de data. Soms is bijvoorbeeld een kwadratisch verband beter.

$$Y = a + bX + cX^2 + \varepsilon.$$

Het kwadratische deel $a + bX + cX^2$ stelt een (berg of dal-) parabool voor, terwijl ε de toevallige verstoring of meetfout is. Y hangt lineair af van de parameters a , b en c en dus is methode van kleinste kwadraten even eenvoudig toepasbaar als hiervoor.

```
x2 = x^2
fit = lm(y ~ x + x2)
coef = fit$coefficients
```

De vector `coef` heeft nu 3 elementen. We plotten de beste parabool door de data

```
plot(x,y)
t = seq(4.5,6.5,by=0.01)
lines(t,coef[1] + coef[2]*t + coef[3]*t^2,col='red')
```

opgave 4.10 Voer de bovenstaande commando's uit, en bekijk de plot. Heeft het toevoegen van de extra term veel zin gehad?

Hoe meer termen je aan het model toevoegt (3-e macht, 4e macht,...) hoe beter de kromme bij de data past. Als je tenslotte evenveel termen als datapunten toevoegt, krijg je een lijn die precies door alle datapunten heen kronkelt. Dat is natuurlijk niet de bedoeling!

opgave 4.11 We genereren een steekproef uit een lineair model.

```
x = rnorm(50,10,10)
y = 10 + 3*x + 0.2*x^2 + 0.05*x^3 + rnorm(50,0,20)
plot(x,y)
```

Fit een rechte lijn, en een tweede- en derde-machts kromme. Plot ze in de grafiek.

4.4 andere lineaire modellen

Het komt vaak voor dat het verband tussen twee grootheden niet in een lineair model te vangen is, maar dat dat wel mogelijk is na een geschikte transformatie. Een voorbeeld is de Arrhenius vergelijking

$$\alpha = Ce^{-e_A/KT}.$$

Hierin is α de snelheid is van een chemische reactie, C een onbekende constante, e_A de activatie energie van de reactie, $K = 1.38 \times 10^{-23} \text{m}^2 \text{kg s}^{-2} \text{K}^{-1}$ de Boltzmann constante en T de absolute temperatuur. Als we de reactie een aantal keer laten verlopen bij verschillende temperaturen en we meten de snelheid α , dan kunnen we C en e_A schatten. Helaas is de vergelijking zoals hij hierboven staat wel lineair in C , maar niet in e_A . We kunnen de methode van kleinste kwadraten dus niet zomaar toepassen. Maar, als we de natuurlijke logaritme nemen, dan geldt

$$\ln(\alpha) = \ln(C) - e_A \frac{1}{KT}$$

We zien dat $\ln(\alpha)$ lineair is in $\ln(C)$ en e_A . We kunnen dus $1/KT$ plotten versus $\ln \alpha$. De helling en het snijpunt met de y -as van de beste rechte lijn door deze plot, levert ons schatters voor $-e_A$ en $\ln(C)$.

opgave 4.12 We hebben de snelheid is van een chemische reactie gemeten bij verschillende temperaturen. Pas de logaritmische transformatie toe, plot de getransformeerde data en schat de constanten C en e_A .

```
T = c(338,342,343,414,417,418,475,480,550,555,560,648,652,653)
a = c(1,2,1,4,5,3,12,7,19,20,16,52,53,51)
plot(T,a)
```


Hoofdstuk 5

Toetsen

5.1 Inleiding

Gooi met een munt met kop-kans p . Als kop valt krijg je een euro, als munt valt ben je een euro kwijt. De meeste mensen zullen dit spel alleen willen spelen als $p > 1/2$. Stel nu dat je eerst 10 keer met de munt mag gooien om vast te stellen of p groter dan $1/2$ is.

Zij X het aantal koppen in de eerste 10 worpen. Het ligt voor de hand om de volgende regel te hanteren:

Kies een zekere K . Als $X > K$, dan zijn we bereid het spel te spelen.

Het is nu nog niet duidelijk hoe we K moeten kiezen, maar daar komen we later op terug.

opgave 5.1 Waarom ligt deze regel voor de hand?

In het jargon van de statistiek hebben we de volgende situatie. Stel X is verdeeld volgens de Binomiale verdeling met parameters $n = 10$ en onbekende p . We toetsen de *nulhypothese*

$$H_0 : p \leq \frac{1}{2}$$

versus het *alternatief*

$$A : p > \frac{1}{2}.$$

We besluiten de nulhypothese te *verwerpen* als $X > K$. We noemen X de *toetsingsgrootte* en K de *kritieke waarde*

De nulhypothese is de “grondtoestand”. We verwerpen de nulhypothese pas als we echt genoeg reden om eraan te twijfelen. Als het experiment geen duidelijke uitslag oplevert, dan krijgt de nulhypothese het voordeel van de twijfel.

Nog een keer voor alle duidelijkheid: Als we de nulhypothese niet kunnen verwerpen, betekent dat *niet* dat de nulhypothese waar is. Het betekent alleen maar dat we (nog) niet genoeg reden hebben om aan de nulhypothese te twijfelen.

Het volgende script genereert 1000 trekkingen uit de binomiaal($n = 10, p = 0.6$) verdeling. In dit voorbeeld is de nulhypothese dus **niet** waar. Als het aantal koppen groter dan $K = 5$ is, verwerpen we de nulhypothese.

```
p = 0.6
K = 5
X = rbinom(1000,10,p)
verwerp = (X>K)
cat("We verwerpen de nulhypothese in ",sum(verwerp)," uit 1000 keer.\n")
```

opgave 5.2 Voer het script uit. De nulhypothese is hier niet waar. Wordt deze dan ook altijd verworpen? Probeer K zo in te stellen dat de nulhypothese altijd wordt verworpen.

opgave 5.3 Verander het script zodat $p = 0.5$ en $K = 5$. Nu is de nulhypothese wèl waar. Wordt deze dan ook nooit verworpen? Probeer K zo in te stellen dat de nulhypothese nooit wordt verworpen.

Zoals je ziet, kun je twee soorten fouten maken bij het toetsen van een hypothese:

1. fout van de eerste soort: nulhypothese is waar, maar wordt toch verworpen.

2. fout van de tweede soort: nulhypothese is niet waar, maar wordt niet verworpen.

Door K aan te passen kunnen we de kans op een fout van de ene soort klein krijgen. Helaas gaat daardoor de kans op een fout van de andere soort omhoog.

analogie Vergelijk de situatie met een rookalarm. De nulhypothese is de normale toestand dat er geen brand is. Het alternatief is dat er wel brand is. Een gevoelig rookalarm gaat al af als je een eitje bakt: een fout van de eerste soort. Geergerd haal je de batterij eruit. De kans op een fout van de eerste soort is omlaag gebracht (naar nul), maar de kans op een fout van de tweede soort gaat daardoor omhoog (naar 1).

De algemeen geaccepteerde oplossing voor dit dilemma is van te voren vast te stellen hoe groot de kans op een fout van de *eerste* soort mag zijn. Dit heet het *significantieniveau* α . Met andere woorden, het significantie niveau is de kans dat de nulhypothese wordt verworpen, terwijl deze juist is. Gebruikelijke waarden voor α zijn 1% of 5%.

opgave 5.4 Stel $\alpha=5\%$. Gebruik het simulatie script om K zo te kiezen dat de nulhypothese in ongeveer 5% van de gevallen wordt verworpen als $p = 1/2$.

Bij de vorige opgave vond je dat $K = 7$ een toets oplevert die in ongeveer 5% van de gevallen de nulhypothese verwerpt, wanneer deze in feite juist is. Met andere woorden $P(X > 7) \approx 0.05$ als de kop-kans $p = 1/2$.

opgave 5.5 Gebruik `R` om $P(X > 7)$ te bepalen, onder de aanname dat $p = 1/2$. Dit is het exacte significantie niveau van de toets.

De zogeheten P -waarde lijkt erg op het significantie niveau, maar mag daarmee niet verward worden. De P -waarde is de kans, onder de nulhypothese, dat de toetsingsgrootte een waarde aanneemt die tenminste zo onwaarschijnlijk is onder de nulhypothese als de geobserveerde waarde.

Bijvoorbeeld: Stel dat we 9 koppen in 10 worpen waarnemen. De P -waarde is $P(X \geq 9) = 0.011$. De P -waarde geeft aan hoe onwaarschijnlijk de observatie is, als de nulhypothese waar zou zijn. Een kleine P -waarde betekent dus dat de observatie erg onwaarschijnlijk is onder de nulhypothese—reden om deze te verwerpen. Er geldt:

De nulhypothese wordt verworpen dan en slechts dan als de P -waarde van de observatie kleiner is

dan het significantie niveau.

Je kan echter NIET zeggen dat de P -waarde de kans is dat de nulhypothese waar is.

5.2 Z toets

Stel X_1, X_2, \dots, X_{25} is een onafhankelijke steekproef uit een normale kansverdeling met onbekende verwachting μ en bekende variantie $\sigma^2 = 4$. We willen de nulhypothese $H_0 : \mu = 3$ toetsen versus het alternatief $A : \mu > 3$. We kiezen als toetsingsgroottheid

$$Z = \sqrt{25} \frac{\bar{X} - 3}{2}.$$

Onder de nulhypothese heeft Z de normale verdeling met verwachting 0 en variantie 1.

opgave 5.6 We verwerpen H_0 als Z groot is. Waarom verwerpen we H_0 niet als Z groot en negatief is?

We kiezen significantie niveau $\alpha = 1\%$ en we willen de kritieke waarde K bepalen zodat $P(Z > K) = 0.01$, onder de nulhypothese.

opgave 5.7 Bepaal K met behulp van de kwantielfunctie `qnorm()`.

opgave 5.8 We proberen de K uit die je zojuist hebt bepaald. Als het goed is zul je de nulhypothese in ongeveer 1% van de gevallen onterecht verwerpen. Voer het onderstaande script een paar keer uit.

```
K = ... # gebruik K die je bij opgave 9 bepaald hebt
z = 1:10000
for (i in 1:10000){
  x = rnorm(25,3,2) # steekproef onder de nulhypothese
  z[i] = 5*(mean(x)-3)/2
}
verwerp = (z>K)
cat("We verwerpen de nulhypothese in ",sum(verwerp)," uit 10000 keer.\n")
```

opgave 5.9 Stel, op basis van een steekproef X_1, X_2, \dots, X_{25} vinden we $Z = 1.65$. Wat is de P waarde van deze waarneming? Wordt de nulhypothese verworpen op niveau $\alpha = 0.05$? En op niveau $\alpha = 0.01$?

Stel dat we de nulhypothese $H_0 : \mu = 3$ willen toetsen versus het alternatief $A : \mu < 3$. Nu verwerpen we H_0 als Z juist groot en negatief is. We kiezen K zodat $P(Z < K) = \alpha$, onder de nulhypothese.

opgave 5.10 Stel, $x = \text{rnorm}(38, 4, 2)$. Kies significantie niveau 1% en toets de nulhypothese $H_0 : \mu = 3$ versus het alternatief $A : \mu < 3$. Gebruik de `qnorm()` om de kritieke waarde te bepalen.

Stel dat we de nulhypothese $H_0 : \mu = 3$ willen toetsen versus het *tweezijdige* alternatief $A : \mu \neq 3$. Nu verwerpen we H_0 als Z groot *of* klein is. We kiezen $K > 0$ zodat $P(Z = K) + P(Z > K) = \alpha$, onder de nulhypothese. De standaard normale verdeling symmetrisch om 0, en het volgt dat

$$P(Z = K) + P(Z > K) = 2P(Z = K), \quad K > 0.$$

opgave 5.11 Stel we willen toetsen $H_0 : \mu = 3$ versus het tweezijdig alternatief $A : \mu \neq 3$. We kiezen significantie niveau $\alpha = 0.05$. Bepaal de kritieke waarde K .

5.3 Student's *t* toets

De Z toets wordt niet zo vaak gebruikt, want het komt eigenlijk nooit voor dat de variantie bekend is, maar de verwachting niet. De t toets die we nu bespreken, wordt daarentegen héél vaak gebruikt.

Stel X_1, X_2, \dots, X_{25} is een onafhankelijke steekproef uit een normale kansverdeling met onbekende verwachting μ en onbekende variantie σ^2 . We willen toetsen $H_0 : \mu = 3$ versus $A : \mu > 3$.

Omdat we σ^2 niet kennen, moeten we hem schatten. Dat doen we met S^2 (zie **kansverdelingen**). Definieer

$$T = \sqrt{25} \frac{\bar{X} - 3}{S}.$$

We verwerpen H_0 als T groot is. Onder de nulhypothese heeft T de zogeheten t -verdeling met $(n - 1)$ vrijheidsgraden. Deze verdeling staat niet in de bijlage **kansverdelingen**. Hij heeft 1 parameter,

te weten het aantal vrijheidsgraden. De t verdeling lijkt erg op de standaard normale verdeling. R kent de functies `dt`, `pt` `qt` en `rt`.

opgave 5.12 Maak plots van de standaard normale verdeling, en van de t verdeling met 1,5,10 en 100 vrijheidsgraden.

opgave 5.13 Stel we kiezen significantie niveau $\alpha = 0.05$. Bepaal de kritieke waarde K zodat $P(T > K) = \alpha$.

opgave 5.14 Stel, op basis van een steekproef X_1, X_2, \dots, X_{25} berekenen we $T = 1.65$. Wat is de P waarde? Wordt de nulhypothese verworpen op niveau $\alpha = 0.05$? En op niveau $\alpha = 0.01$?

5.4 t toets, twee steekproeven

De t -toets wordt vaak gebruikt om te toetsen of de verwachting van de ene populatie groter is dan de andere. Stel dat 20 patienten een medicijn krijgen dat wordt geacht de bloeddruk te verlagen, en dat een controle groep van 15 patienten een placebo (niet werkzame stof) krijgt toegediend. Zij X_1, X_2, \dots, X_{20} de bloeddruk van de patienten na toediening van het medicijn, en zij Y_1, Y_2, \dots, Y_{15} de bloeddruk van de mensen in de controlegroep na toediening van het placebo.

opgave 5.15 Waarom neemt men eigenlijk de moeite om de controle groep een placebo te geven?

We simuleren een dataset, op zo'n manier dat er inderdaad een verschil in verwachtingswaarden is tussen de patienten en de controle groep.

```
x = rnorm(20,180,3) #20 patienten met verwachting 180 en standaardafwijking 3
y = rnorm(15,182,3) #15 controles met verwachting 182 en standaardafwijking 3
```

We kunnen de twee steekproeven vergelijken met behulp van zogeheten boxplots.

```
boxplot(list(x,y))
```

Een boxplot geeft de mediaan en de 5% en 95% kwantielen weer als een “doos”. De “snorharen” van de doos geven de grootste en kleinste waarden in de steekproef.

Door naar de twee boxplots te kijken begrijp je dat niet alleen het verschil in verwachtingswaardes, maar ook de grootte van de standaard afwijking belangrijk is voor het succes van de toets.

opgave 5.16 Verander de standaard afwijking van 3 in 0.5 en 5 en maak boxplots.

In statistisch jargon hebben we het volgende probleem: Stel X_1, X_2, \dots, X_n zijn normaal verdeel met verwachting μ_1 en (onbekende) variantie σ^2 en Y_1, Y_2, \dots, Y_m zijn normaal verdeeld met verwachting μ_2 en dezelfde variantie σ^2 . We willen toetsen

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad A : \mu_1 < \mu_2.$$

We gebruiken als toestingsgroottheid

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

waarbij

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2},$$

met

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2.$$

We kiezen K een verwerpen H_0 als $T < K$. Onder de nulhypothese heeft T de t verdeling met $n + m - 2$ vrijheidsgraden.

opgave 5.17 Genereer data $\mathbf{x} = \text{rnorm}(20, 180, 3)$ en $\mathbf{y} = \text{rnorm}(15, 182, 3)$ en toets $H_0 : \mu_1 = \mu_2$ versus $A : \mu_1 < \mu_2$ op niveau $\alpha = 5\%$. Gebruik de functie `qt()` om de kritieke waarde te bepalen. Gebruik de functie `pt()` om ook de P -waarde te bepalen. Herhaal een aantal keer met verschillende parameters (n, m, μ_1, μ_2 en σ^2)

opgave 5.18 Herhaal de vorige opgave een aantal keer. Bekijkt steeds de boxplots, en probeert te raden of de nulhypothese verworpen kan worden of niet.

opgave 5.19 Herhaal de vorige opgave maar neem nu $\mathbf{x} = \text{rnorm}(2000, 180, 3)$ en $\mathbf{y} = \text{rnorm}(1500, 182, 3)$

opgave 5.20 De R functie `t.test()` neemt je al het werk uit handen. Genereer data zoals hierboven, en gebruik deze ingebouwde functie om de toets uit te voeren.

5.5 Wilcoxon toets

De Wilcoxon toets wordt net als de t toets gebruikt om te toetsen of twee populaties dezelfde verwachting hebben. De Wilcoxon toets maakt echter geen gebruik van de aanname dat de data normaal verdeeld zou zijn. En zelfs al is de data normaal verdeeld, dan nog is de Wilcoxon toets bijna even goed als de t toets. In de praktijk wordt bijna altijd de t toets gebruikt, en vaak zonder na te gaan of de data wel de normale verdeling heeft. Dat is dus heel erg dom.

De Wilcoxon toets is in R geïmplementeerd als `wilcox.test()`.

5.6 de F toets

Stel X_1, X_2, \dots, X_n zijn normaal verdeeld met verwachting μ_1 en (onbekende) variantie σ_1^2 en stel de Y_1, Y_2, \dots, Y_m zijn normaal verdeeld met verwachting μ_2 en dezelfde variantie σ_2^2 . We willen toetsen

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad A : \sigma_1^2 > \sigma_2^2.$$

We gebruiken als toestingsgroottheid

$$F = \frac{S_X^2}{S_Y^2}.$$

We verwerpen de nul hypothese als F groot is. Onder de nulhypothese heeft F de zogeheten F -verdeling met $(n - 1, m - 1)$ vrijheidsgraden. R kent de functies `df`, `pf` `qf` en `rf`.

opgave 5.21 genereer zelf data naar keuze, kies zelf het significantie niveau, en voer de F toets uit.

De F toets is in R geïmplementeerd als `var.test()`.

5.7 de χ^2 toets

Stel dat we een steekproef X_1, X_2, \dots, X_n hebben uit een discrete kansverdeling p op de uitslagen verzameling $\{1, 2, \dots, K\}$. We willen de nulhypothese toetsen de kansverdeling gegeven wordt door zekere $p_0(1), p_0(2), \dots, p_0(K)$.

Zij E_k het verwachte aantal keer dat de waarde k wordt waargenomen, en zij O_k het werkelijk aantal keer dat k wordt waargenomen. Onder de nulhypothese geldt $E_k = np_0(k)$ (waarom?). Als toetsingsgrootte gebruiken we

$$\chi^2 = \sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k},$$

en we verwerpen de nulhypothese als χ^2 groot is.

opgave 5.22 Waarom is het logisch dat we de nulhypothese verwerpen als χ^2 groot is?

Onder de nulhypothese heeft χ^2 (bij benadering) de zogeheten χ^2 verdeling (zeg chi kwadraat verdeling) met $(K - 1)$ vrijheidsgraden. R kent de functies `dchisq`, `pchisq`, `qchisq` en `rchisq`.

voorbeeld Stel we hebben een steekproef X_1, X_2, \dots, X_{100} uit een discrete kansverdeling p op de uitslagenruimte $\{1, 2, 3, 4\}$. We willen toetsen

$$H_0 : p = p_0 \quad \text{versus} \quad A : p \neq p_0,$$

waar $p_0 = (1/3, 1/4, 1/6, 1/4)$. We kiezen niveau $\alpha = 5\%$

We bekijken eerste de situatie dat de nul hypothese juist is. We simuleren een steekproef ter grootte $n = 100$ uit de kansverdeling $p = p_0$.

```

n = 100          # steekproefgrootte n=100
K = 4           # K=4 elementen in uitslagenruimte
p = c(1/3,1/4,1/6,1/4) # ware kansverdeling p
p0 = c(1/3,1/4,1/6,1/4) # nul hypothese kansverdeling p0
x = sample(1:K,n,replace=T,prob=p) # steekproef uit verdeling p=p0

```

We maken een histogram van de data

```
hist(x,prob=T,breaks=0:4)
```

Vervolgens berekenen we de χ^2 statistiek

```

E = n*p0        # verwachte aantallen onder p0
O = 1:K
for (k in 1:K){
  O[k] = sum(x==k) # het aantal keer dat waarde k voorkomt
}
X2 = sum((O-E)^2/E) # de chi kwadraat statistiek

```

We bepalen de P waarde, ofwel de kans op een nog grotere waarde voor χ^2

```

pvalue = 1-pchisq(X2,K-1)
cat("P value =",pvalue,"\n")

```

Als de P waarde kleiner is dan het niveau $\alpha = 0.05$, verwerpen we de nulhypothese (hoewel deze, in dit voorbeeld, juist was.)

opgave 5.23 Voer de bovenstaande commando's uit. Herhaal het hele experiment een aantal keer. Bepaal steeds of je de nulhypothese verwerpt.

opgave 5.24 Gebruik weer $p = (1/3,1/4,1/6,1/4)$ om de data te genereren, maar toets nu de (valse) nul hypothese dat de verdeling van de data uniform is $p_0(1) = p_0(2) = p_0(3) = p_0(4) = 1/4$.

De χ^2 toets is in R geïmplementeerd als `chisq.test()`.

5.8 Kolmogorov-Smirnov toets

Stel we hebben een steekproef X_1, X_2, \dots, X_n uit een onbekende kansverdeling. We vragen ons af de verdeling wellicht de normale is. We toetsen

H_0 : kansverdeling is normaal versus A : kansverdeling is *niet* normaal.

We bepalen eerst welke normale verdeling het beste bij de data past; we schatten $\hat{\mu} = \bar{X}$ en $\hat{\sigma}^2 = S^2$. Vervolgens vergelijken we de verdelingsfunctie van de geschatte normale verdeling, met de zogeheten empirische verdelingsfunctie. De empirische verdelingsfunctie is gedefinieerd als

$$F_n(x) = \frac{\#\{i : X_i \leq x\}}{n}.$$

De Kolmogorov-Smirnov toets is geïmplementeerd in R als `ks.test()`.

opgave 5.25

1. Genereer een steekproef ter grootte 40 uit de normale verdeling met verwachting 11 en standaard afwijking 3.
2. Schat μ en σ^2 en maak een plot van de geschatte normale verdelingsfunctie. Gebruik `pnorm()`
3. Het volgende scriptje voegt een plot van de empirische verdelingsfunctie toe.

```
n = 40
lines(sort(x), (1:n)/n, type='s')
```

voer het script uit (en probeer het te begrijpen).

4. Voer de Kolmogorov-Smirnov toets uit:

```
ks.test(x, 'pnorm', mean(x), sd(x))
```

Verwerp je de nulhypothese?

opgave 5.26 Genereer een steekproef ter grootte $n = 100$ uit de exponentiële verdeling met parameter $\lambda = 0.2$ en toets de nulhypothese dat de steekproef uit de exponentiële verdeling komt

op niveau 5%. Maak ook een plot van de geschatte exponentiële verdeling versus de empirische verdeling.

opgave 5.27 Genereer een steekproef ter grootte $n = 100$ uit de exponentiële verdeling met parameter $\lambda = 0.2$ en toets de (valse) nulhypthese dat de steekproef uit de normale verdeling komt op niveau 5%. Maak ook een plot van de geschatte normale verdeling versus de empirische verdeling.

Hoofdstuk 6

ANOVA

ANOVA staat voor *analysis of variance*, ofwel variantie analyse. Het doel is te bepalen of variatie in de data systematisch is, of alleen maar toevallig.

6.1 lineaire regressie

We bekijken nog een keertje de alcohol en tabak uitgaven.

```
x = c(6.47,6.13,6.19,4.89,5.63,4.52,5.89,4.79,5.27,6.08)
y = c(4.03,3.76,3.77,3.43,3.47,2.92,3.20,2.71,3.53,4.51)
plot(x,y,xlab="alcohol uitgave",ylab="tabak uitgave")
```

Het lineaire model is

$$Y_i = a + bX_i + \varepsilon_i \quad (6.1)$$

waarin de ε_i onafhankelijk verdeeld zijn volgens de normale verdeling met verwachting nul en (onbekende) variantie σ^2 . We kunnen a en b schatten door middel van

```
fit = lm(y ~ x)
```

Is er een relatie tussen X en Y ? Kunnen we dat hard maken? Het *lijkt* in de plot wel dat er een systematisch verband is, maar misschien is dat wel gewoon toeval!

opgave 6.1 Genereer en plot onafhankelijke steekproeven en voer een lineaire regressie uit.

```
xx = rnorm(10,mean(x),sd(x))
yy = rnorm(10,mean(y),sd(y))
plot(xx,yy)
fitt = lm(yy ~ xx)
abline(fitt$coefficients,col='red')
```

Herhaal tenminste 10 keer. Het lijkt af en toe net alsof er een verband is tussen xx en yy , terwijl dat toch puur toeval is.

Nu zijn X en Y onafhankelijk dan en slechts dan als $b = 0$. We willen dus toetsen

$$H_0 : b = 0 \quad \text{versus} \quad A : b \neq 0.$$

Onder de nulhypothese is ons model dus eigenlijk

$$Y_i = a + \varepsilon_i$$

en de kleinste kwadraten schatter van a in dit model is \bar{Y} . Dat wil zeggen, de beste horizontale lijn door de punten wolk is $y = \bar{Y}$. De kwadraat som die hierbij hoort is

$$SS_{H_0} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Onder het alternatief hebben we te maken met het volledige model (6.1). De kleinste kwadraten schatters zijn \hat{a} en \hat{b} met kwadraat som

$$SS_A = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i)^2.$$

Als SS_{H_0} veel groter is dan SS_A , dan levert de toevoeging van de parameter b kennelijk een veel betere fit op; reden om H_0 te verwerpen. We gebruiken als toestingsgrootheid

$$F = (n - 2) \frac{SS_{H_0} - SS_A}{SS_A}$$

en verwerpen H_0 als F groot is. Onder de nulhypothese heeft F de F verdeling met $(1, n - 2)$ vrijheidsgraden.

opgave 6.2 Bepaal de F statistiek voor de alcohol/tabak data en bereken de P waarde. Controleer je antwoord met `summary(fit)`. Verwerp je de nulhypothese op niveau $\alpha = 0.01$?

Een andere statistiek die vaak wordt genoemd is

$$R^2 = \frac{SS_{H_0} - SS_A}{SS_{H_0}}.$$

Dit is de relatieve verbetering als gevolg van de toevoeging van de parameter b .

6.2 ANOVA; 1 factor

Als we twee steekproeven hebben, weten we inmiddels hoe we kunnen toetsen of de ene steekproef “groter” is dan de andere: we kunnen de t toets of (bij voorkeur) de Wilcoxon toets gebruiken. Wat moeten we doen als we meer dan twee steekproeven hebben?

voorbeeld Een nieuwe semi-automatische methode voor het meten van de hoeveelheid chloorpheniramine maleaat in tabletten is ontwikkeld. In 7 laboratoria heeft men de nieuwe methode toegepast op 10 tabletten, die elk 4 mg van de stof bevatten. Het doel van de studie is de variabiliteit tussen de laboratoria, en de variabiliteit in het meetproces te bepalen.

`y1 = c(4.13,4.07,4.04,4.07,4.05,4.04,4.02,4.06,4.10,4.04)`

`y2 = c(3.86,3.85,4.08,4.11,4.08,4.01,4.02,4.04,3.97,3.95)`

`y3 = c(4.00,4.02,4.01,4.01,4.04,3.99,4.03,3.97,3.98,3.98)`

`y4 = c(3.88,3.88,3.91,3.95,3.92,3.97,3.92,3.90,3.97,3.90)`

`y5 = c(4.02,3.95,4.02,3.89,3.91,4.01,3.89,3.89,3.99,4.00)`

`y6 = c(4.02,3.86,3.96,3.97,4.00,3.82,3.98,3.99,4.02,3.93)`

`y7 = c(4.00,4.02,4.03,4.04,4.10,3.81,3.91,3.96,4.05,4.06)`

opgave 6.3 Plot de data door middel van `boxplot(list(y1,y2,y3,y4,y5,y6,y7))`

Stel

$$Y_{ij} = \text{de } j\text{-de meting in het } i\text{-de laboratorium } (i = 1, 2, \dots, 7, j = 1, 2, \dots, 10)$$

Het statistische model voor de metingen is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

waarbij μ het totale gemiddelde is, α_i het “effect” van het i -de laboratorium en ε_{ij} de toevalige fout in de (i, j) -de meting. We nemen aan de ε_{ij} allen onafhankelijk, normaal verdeeld zijn met verwachting 0 en (onbekende) variantie σ^2 .

Omdat het alleen gaat om verschillen tussen de laboratoria stellen we

$$\alpha_1 = 0.$$

Het is niet moeilijk μ en de α_i te schatten. We schatten eerst

$$\hat{\mu} = \frac{1}{10} \sum_{j=1}^{10} Y_{1j} = \bar{Y}_{1\bullet}$$

en vervolgens

$$\hat{\alpha}_i = \frac{1}{10} \sum_{j=1}^{10} (Y_{ij} - \hat{\mu}) = \bar{Y}_{i\bullet} - \bar{Y}_{1\bullet}.$$

Zoals je ziet geven we met een \bullet aan dat we over een bepaalde index middelen.

opgave 6.4 Waarom liggen deze schatters voor de hand? Bepaal $\hat{\mu}$ en de $\hat{\alpha}_i$ ($i = 1, 2, \dots, 7$).

We kunnen de schatters ook door R laten bepalen met behulp van de functie `lm()`. Dat zullen we in een aantal stappen doen.

opgave 6.5 We maken één lange data vector, en een vector `labs` van dezelfde lengte die aangeeft welke waarneming bij welk laboratorium hoort.

```
y = c(y1,y2,y3,y4,y5,y6,y7)
```

```
labs = c(rep(1,10),rep(2,10),rep(3,10),rep(4,10),rep(5,10),rep(6,10),rep(7,10))
```

Voer deze commando's uit. Een korter commando is `labs = floor(seq(1,7.9,by=0.1))`. Begrijp je hoe dat werkt?

opgave 6.6 We kunnen nu de functie `lm()` gebruiken.

```
plot(labs,y)
fit = lm(y ~ labs)
abline(fit$coefficients,col='red')
```

Is dit een zinvolle regressie om de relatie tussen de laboratoria en de metingen te bepalen? Wat gebeurt er als we de laboratoria anders hadden genummerd?

De nummering van de laboratoria is niet relevant. Het gaat er alleen om dat de data in 7 groepen moet worden verdeeld. Dat doen we door van de vector `labs` een *factor* te maken.

opgave 6.7 Voer de onderstaande regressie uit, en vergelijk het resultaat met het antwoord van opgave 4.

```
labs = factor(labs)
fit = lm(y ~ labs)
```

Bekijk nog eens de boxplot. We vragen ons af of de verschillen tussen de laboratoria systematisch zijn, of dat het alleen maar toeval is. We toetsen

$$H_0 : a_1 = a_2 = \dots = a_7 = 0 \quad \text{versus} \quad A : \text{tenminste één } a_i \text{ is niet nul.}$$

Het idee achter de toets is precies hetzelfde als bij de simpele lineaire regressie. We vergelijken de fit van het model onder de nulhypothese met de fit onder het alternatief. De toetsingsgrootte is weer het quotient van kwadraat sommen, met de F verdeling onder de nulhypothese.

opgave 6.8 Doe `summary(fit)` en bekijk de output. In de laatste regel staat de F statistiek en de P waarde. Kies niveau $\alpha = 0.05$. Verwerp je de nulhypothese dat er geen systematisch verschil tussen de laboratoria is?

Het commando `anova(fit)` is iets overzichtelijker. Het geeft alleen de relevante kwadraatsommen en de F statistiek met zijn P waarde.

6.3 ANOVA; 2 factoren

We bestuderen data ontleend aan J. Rice, *Mathematical Statistics and Data Analysis*, tweede editie, Wadsworth 1995. Men wil onderzoeken welke van twee soorten ijzer (Fe^{2+} of Fe^{3+}) beter door het lichaam wordt opgenomen, en dus geschikter is als voedingssupplement. Men heeft 108 muizen toevallig onderverdeeld in 6 groepen van 18. De eerste drie groepen krijgen Fe^{3+} , maar elk in een andere concentratie: 10.2, 1.2 en 0.3 mmol. De tweede drie groepen krijgen Fe^{2+} in dezelfde concentraties.

De data staat op blackboard onder “course documents” als `ijzermuizen.txt`. Sla het bestand op in je computer, en lees het in R

```
A = scan(file="???/ijzermuizen.txt")      # lees data als 1 lange vector
A = t(matrix(A,6,18))                    # maak weer matrix
```

Hierin staat ??? voor de volledige naam van de map (directory) waar je `ijzermuizen.txt` hebt opgeslagen. Als alles goed is, heb je nu een matrix A met 18 rijen en 6 kolommen. Laten we eens kijken

```
namen = list("Fe3 10.2", "Fe3 1.2", "Fe3 0.3", "Fe2 10.2", "Fe2 1.2", "Fe2 0.3")
boxplot(list(A[,1], A[,2], A[,3], A[,4], A[,5], A[,6]), names=namen)
```

Het eerste dat opvalt is dat hoe groter het gemiddelde is, hoe groter ook de spreiding. Dat kunnen we verhelpen door de natuurlijke logaritme te nemen.

opgave 6.9 Neem de logaritme en maak een nieuwe boxplot.

We hebben nu te maken twee factoren: soort ijzer en dosis. Een lineair model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, 2 \quad j = 1, 2, 3 \quad \text{en} \quad k = 1, 2, \dots, 18,$$

waarin de ε_{ijk} onafhankelijk normaal verdeeld zijn met verwachting nul en (onbekende) variantie σ^2 . In dit model stellen we $\alpha_1 = 0$ en $\beta_1 = 0$. De α_i zijn het effect van de soort ijzer, de β_j zijn het effect van de dosis.

We voeren een twee factor variantie analyse uit.

```

logA = log(A) # neem de logaritme
logA = as.vector(logA) # rek uit tot een lange vector
soort = c(rep(3,54),rep(2,54)) # de bijbehorende soort ijzer
dosis = c(rep(10.2,18),rep(1.2,18),rep(0.3,18))
dosis = c(dosis,dosis) # de bijbehorende dosis
soort = factor(soort)
dosis = factor(dosis)
fit = lm(logA ~ soort + dosis) # fit een twee factor ANOVA
anova(fit)

```

We vinden de volgende ANOVA tabel

```

Response: logA
          Df Sum Sq Mean Sq F value    Pr(>F)
soort      1  2.074    2.074   5.9735  0.01620
dosis      2 15.588    7.794  22.4504 7.854e-09
Residuals 104 36.106    0.347

```

Als we willen toetsen of de soort ijzer effect heeft, toetsen we feitelijk

$$H_0 : \alpha_1 = \alpha_2 = 0 \quad \text{versus} \quad A : \alpha_2 \text{ is niet nul}$$

Volgens de tabel vinden we $F = 5.97$ met een P waarde van 0.016. Op niveau 0.01 is dit dus niet significant.

Als we willen toetsen of de dosis effect heeft, toetsen we feitelijk

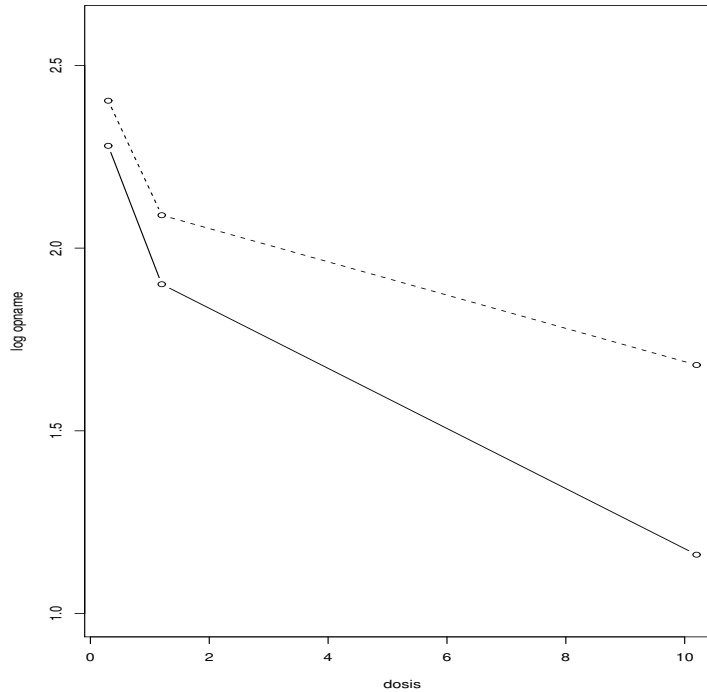
$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{versus} \quad A : \beta_2 \text{ of } \beta_3 \text{ is niet nul}$$

Volgens de tabel vinden we $F = 22.45$ met een P waarde van $7.8 \cdot 10^{-9}$. Op niveau 0.01 wordt de nul hypothese overtuigend verworpen. Het effect van de dosis is duidelijk significant. Dat hadden we in de boxplot natuurlijk ook al gezien.

6.4 ANOVA; twee factoren met interactie

Het is soms ook interessant om te onderzoeken of er een *interactie* bestaat tussen de factoren. Het zou zo kunnen zijn dat juist bij een bepaalde dosis de ene soort ijzer beter wordt opgenomen dan de

andere, terwijl bij een andere dosis dit niet het geval is. We plotten de gemiddelde (log)ijzeropname voor de 6 groepen. De doorgetrokken lijn is Fe^{3+} , de gestippelde lijn is Fe^{2+} .



Figuur 6.1: Dosis versus log ijzeropname voor Fe^{3+} (doorgetrokken) en Fe^{2+} (gestippeld)

Als er geen interactie was tussen de soort ijzer en de dosis, zou je verwachten dat de lijnen in de grafiek steeds dezelfde afstand houden (waarom?). De lijnen in onze grafiek doen dat niet, maar het is de vraag of de afwijking systematisch is of gewoon toeval. Ons model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, 2 \quad j = 1, 2, 3 \quad \text{en} \quad k = 1, 2, \dots, 18,$$

waarin de ε_{ijk} onafhankelijk normaal verdeeld zijn met verwachting nul en (onbekende) variantie σ^2 . De γ_{ij} modelleren het interactie effect.

We hebben nu dus 3 factoren: de soort ijzer (**soort** met 2 niveaus), de dosis (**dosis** met 3 niveaus) en de interactie ($2 \times 3 = 6$ niveaus). De interactie kunnen we definiëren in R als **soort:dosis**.

opgave 6.10 Bekijk de factor `soort:dosis` en begrijp dat er geen wezenlijk verschil is met de andere factoren `soort` en `dosis`.

We voeren een variantie analyse uit met interactie.

```
interactie = soort:dosis
fit = lm(logA ~ soort + dosis + interactie) # twee factor, met interactie
anova(fit)
```

opgave 6.11 Voer de bovenstaande variantie analyse uit. Toets de nulhypothese dat er geen interactie is op niveau $\alpha = 0.05$.

Bijlage A

Kansverdelingen

1. De Binomiale verdeling

model: aantal koppen in n worpen met een munt met kop-kans p .

parameters: n en p .

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

$$EX = np \quad \text{en} \quad \text{var}(X) = np(1 - p)$$

Schatter van p op basis van een trekking X uit de Binomiaal(n, p) verdeling

$$\hat{p} = \frac{X}{n}.$$

De Binomiaal($n = 1, p$) verdeling wordt ook wel de Bernoulli(p) verdeling genoemd.

2. De Geometrische verdeling

model: aantal keer gooien met munt met kop-kans p tot de eerste keer kop.

parameters: p

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

$$EX = \frac{1}{p} \quad \text{en} \quad \text{var}(X) = \frac{1-p}{p^2}$$

Schatter op basis van steekproef X_1, X_2, \dots, X_n

$$\hat{p} = \frac{1}{\bar{X}}.$$

3. De Poisson verdeling

model: aantal oproepen in een telefooncentrale per tijdseenheid.

parameters: λ

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

$$EX = \lambda \quad \text{en} \quad \text{var}(X) = \lambda$$

Schatter op basis van steekproef X_1, X_2, \dots, X_n

$$\hat{\lambda} = \bar{X}.$$

4. De Uniforme verdeling

model: toevallig gekozen reël getal in een interval

parameters: de grenzen van het interval: $[a, b]$

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

$$EX = \frac{a+b}{2} \quad \text{en} \quad \text{var}(X) = \frac{(b-a)^2}{12}$$

Schatter op basis van steekproef X_1, X_2, \dots, X_n

$$\hat{a} = \min(X_1, X_2, \dots, X_n) \quad \text{en} \quad \hat{b} = \max(X_1, X_2, \dots, X_n).$$

5. De Exponentiële verdeling

model: Levensduur (van een transistor)

parameters: $\lambda > 0$

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

$$EX = \frac{1}{\lambda} \quad \text{en} \quad \text{var}(X) = \frac{1}{\lambda^2}$$

Schatter op basis van steekproef X_1, X_2, \dots, X_n

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

6. De Normale verdeling

model: verdeling van toevallige meetfout

parameters: μ en σ^2

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

$$EX = \mu \quad \text{en} \quad \text{var}(X) = \sigma^2$$

Schatter op basis van steekproef X_1, X_2, \dots, X_n

$$\hat{\mu} = \bar{X} \quad \text{en} \quad \hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2.$$

7. De rest

Bekende discrete verdelingen die we hier niet uitwerken zijn: discreet uniform, hypergeometrisch, multinomiaal en negatief binomiaal. Andere bekende continue verdelingen zijn: Beta, Gamma, χ^2 (chi-kwadraat), Student's t , F , lognormaal, en Cauchy.

Index

- alternatief, 52
- analysis of variance (ANOVA), 63
- ANOVA, 63
- Boolese variabele, 10
- boxplot, 57
- combinatoriek, 22
- complement, 18
- correlatie coefficient, 39
- covariantie, 36
- deelverzameling, 17
- dichtheidsfunctie, 25
- disjunct, 20
- doorsnede, 18
- experiment, 17
- factor, 65
- faculteit, 22
- for lus, 14
- fout van de eerste soort, 52
- fout van de tweede soort, 53
- gebeurtenis, 17
- gemiddelde, 34
- histogram, 27
- hypothese toets, 51
- if constructie, 15
- interactie, 69
- kansdichtheid, 25
- kansfunctie, 19
- kansrekening, 17
- kansverdeling
 - F* verdeling, 58, 73
 - χ^2 verdeling, 59, 73
 - t* verdeling, 55, 73
 - Bernoulli, 32, 73
 - Binomiaal, 23, 73
 - Exponentieel, 23, 73
 - Geometrisch, 23, 73
 - Normaal, 26, 73
 - Poisson, 23, 73
 - Uniform, 73
- Kansverdling
 - Uniform, 25
- kleinste kwadraten, 46
- kritieke waarde, 52
- kwadraat som, 46, 64
- kwantiel, 24
- lege verzameling, 18
- lineaire regressie, 46, 63
- MAD, 39
- matrix, 11
- mean squared error, 45
- mediaan, 24, 39
- Monte Carlo, 28
- MSE, 45
- nulhypothese, 51

- onafhankelijk, 20
- P-waarde, 53
- parameter, 43
- plot, 13
- R, 7
- R functies, 13
- regel van totale waarschijnlijkheid, 22
- robuust, 39
- schatten, 43
- script, 14
- significantieniveau, 53
- standaard afwijking, 33
- steekproef correlatie coefficient, 39
- steekproef covariantie, 37
- steekproef variantie, 35
- stochastische grootheid, 23
- subscripting, 11
- sum of squares, 46
- toets, 51
 - F toets, 58, 65
 - Z toets, 54
 - χ^2 toets, 59
 - t toets, 55
 - Kolmogorov-Smirnov toets, 61
 - Wilcoxon toets, 58
- uitbijter, 39
- uitslagenruimte, 17
- variantie, 32
- variantie analyse (ANOVA), 63
- verdelingsfunctie, 23, 25
- vereniging, 18
- verwachte kwadratische fout, 45
- verwachting, 31
- verwerpen, 52
- voorwaardelijke kans, 21
- vrijheidsgraden, 55, 58, 59
- zuivere schatter, 44