UNIVERSITEIT GENT

BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

Categorische Data Analyse

Lieven Clement

2^{de} bach. in de Biologie, Chemie, Biochemie en Biotechnologie en Biomedische Wetenschappen

8.1 Inleiding

- Tot nog toe zijn modelleren van een continue uitkomst a.d.h.v. een categorische of continue predictor.
- Nu besluitvorming voor een categorische uitkomst.
- Focus associatie tussen een categorische uitkomst en een categorische predictor.
- Gebruik van *kruistabellen* om associatie voor te stellen.

8.2 Toetsen voor een proportie

Saksen-studie

- Vrij gesloten populatie (weinig immigratie en emigratie)
- Waarschijnlijk dat een ongeboren kind mannelijk is?

```
boys <- 3175
n <- 6155
```

- Op 6155 ongeboren kinderen werden 3175 jongens geobserveerd.
- Verschil in de kans dat het ongeboren kind een jongen is of een meisje.

- Gegevens voorstellen als uitkomsten van een numerieke toevalsveranderlijke X
- $X = 1$ voor jongens en
- $X = 0$ voor meisjes.
- Merk op: telprobleem omdat de uitkomst een telling (nl. het aantal jongens)
- Formeel hebben we nu een populatie van ongeboren kinderen beschouwd waarin elk individu gekenmerkt wordt door een 0 of een 1.
- De uitkomst variabele is dus binair.

Bernoulli verdeling

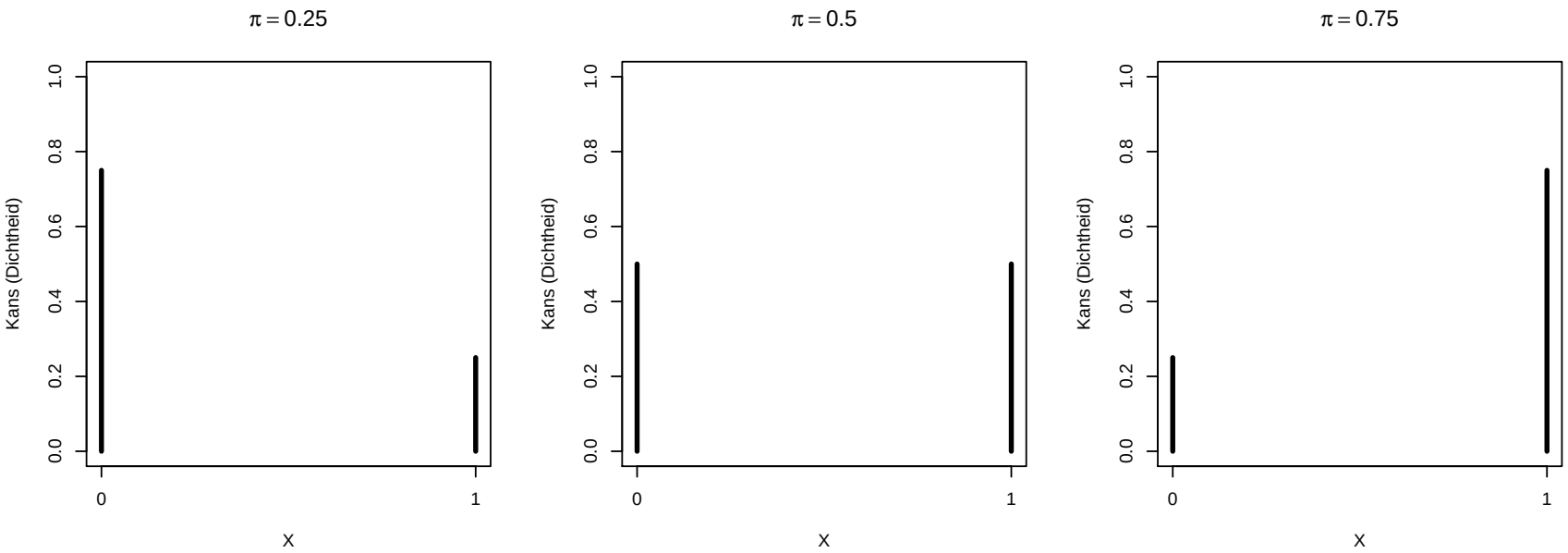
- Binaire data kan worden gemodelleerd a.d.h.v. een Bernoulli verdeling:

$$\begin{aligned} X_i &\sim B(\pi) \text{ met} \\ B(\pi) &= \pi^{X_i}(1-\pi)^{(1-X_i)}, \end{aligned}$$

- een distributie met 1 model parameter π
  - Verwachte waarde van X_i: $\mathrm{E}[X_i] = \pi$,
  - De proportie van ongeboren jongens (d.i. kinderen met een 1) in de populatie.
  - Bijgevolg is π kans dat lukraak getrokken individu een jongen is (een observatie die 1 oplevert).
- De variantie van Bernoulli data is eveneens gerelateerd aan de kans π.

$$\mathrm{Var}[X_i] = \pi(1-\pi).$$

Grafische weergave van enkele Bernoulli kansverdelingen



- In Saksenstudie worden Iukraak 6155 observaties getrokken uit de populatie.
- We schatten π als het steekproefgemiddelde :

$$\hat{\pi} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

```
pi=boys/n
pi
```

```
## [1] 0.5158408
```

In ons voorbeeld is $\bar{x} = 3175 \ / \ 6155 = 51.6\%$.

8.2.1. Binomiale test

- Geeft feit dat 51.6% van de kinderen in de studie mannelijk zijn, voldoende overtuigingskracht om te beweren dat er meer kans is dat een ongeboren kind een jongen is dan een meisje.
- Statistiche toets voor

$$H_0 : \pi = 1/2 \text{ versus } H_1 : \pi \neq 1/2,$$

- Daarvoor moeten we verdeling van de
- X en $\bar{X}$
- of van de som $S = n\bar{X}$ kennen.

- Stel $H_0 : \pi = 1/2$ is waar (voorkomen van jongens en meisjes in populatie even waarschijnlijk)
- Lukrake trekking van één individu uit de populatie, kans op een jongen

$$P(X = 1) = \pi = 1/2.$$

- Twee kinderen onafhankelijk van elkaar (en de populatie $\approx \infty$):
- Kans $\pi = 1/2$ op jongen voor zowel eerste als tweede kind (onafhankelijk van elkaar)
- Uitkomsten (x_1, x_2) voor beide kinderen hebben dan 4 mogelijke waarden: $(0,0), (0,1), (1,0)$ en $(1,1)$.
- Deze komen elk voor met kans $1/4 = 1/2 \times 1/2$.
- Toevalsveranderlijke S die som van uitkomsten weergeeft kan volgende waarden aannemen:

(x_1, x_2)	s	$P(S = s)$
(0,0)	0	1/4
(0,1), (1,0)	1	1/2
(1,1)	2	1/4

Algemeen: n onafhankelijke observaties

- Kans π op "succes" (uitkomst 1) voor elke observatie
- Totaal aantal successen S (som van alle 1-en) kan $n+1$ mogelijke waarden hebben

$$S = k, \text{ met } k = 0, \ldots, n$$

- Verdeling van S?

$$P(S = k) = \left(\begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} (\#eq : binomk) \qquad (1)$$

- $1-\pi$: kans op mislukking in 1 enkele trekking (uitkomst met 0 genoteerd) en
- binomiaalcoëfficient

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n \times (n-1) \times \ldots \times (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

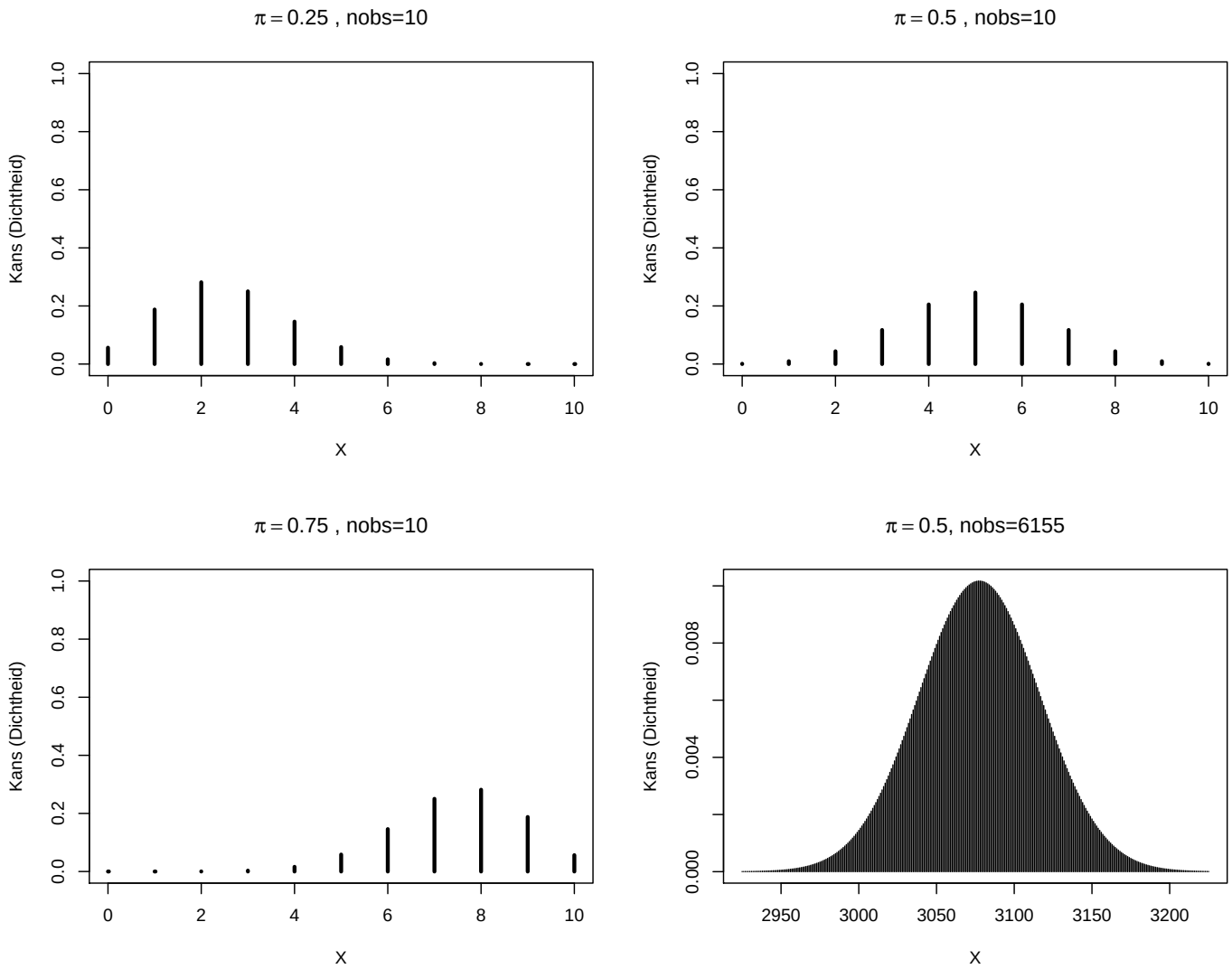
- In R kan je de kansen van binomiale verdeling voor elke $S = k$ opvragen met `dbinom(k,n,p)`

Binomiale Verdeling

Een toevalsveranderlijke S een kansverdeling in Model @ref(eq:binomk):

- *Binomiaal verdeelde toevalsveranderlijke* met bijhorende *Binomiale kansverdeling*
- parameters
  - n (d.i. het aantal trekkingen of, equivalent, de maximale uitkomstwaarde)
  - π (de kans op een 'succes' bij elke trekking).
- Kans berekenen k gebeurtenissen zich voordoen op n onafhankelijke experimenten waarbij kans op 1 zo'n gebeurtenis per experiment, π bedraagt.
- Voor analyse van gegevens die slechts 2 mogelijke waarden kunnen aannemen.
- Bijvoorbeeld: al dan niet besmet met HIV, wild type van een gen vs een mutant,...
- Gebruik: Proporties of risico's op een gebeurtenis van een bepaald type vergelijken tussen verschillende groepen.

Een grafische weergave van enkele Binomiale kansverdelingen.



Figuur 1: Binomiale verdelingen.

Toetsstatistieken voor

$$H_0 : \pi = 1/2 \text{ versus } H_1 : \pi \neq 1/2$$

- $\bar{X} - 1/2$ of, equivalent,
- $\Delta = n(\bar{X} - \pi_0) = S - s_0$.
- Verdeling van deze laatste toetsstatistiek volgt rechtstreeks uit de Binomiale verdeling:
- We observeren $s = 3175$ en dus $\delta = s - s_0 = 3175 - 6155 \times 0.5 = 97.5$.
- In veronderstelling dat jongens en meisjes even waarschijnlijk zijn (d.i. onder de nulhypothese $H_0 : \pi = 1/2$), bekomen we de bijhorende tweezijdige p-waarde:

$$p = \mathrm{P}_0 \left[S - s_0 \geq |\delta| \right] + \mathrm{P}_0 \left[S - s_0 \leq -|\delta| \right].$$

- Merk op dat we dit kunnen herschrijven in termen van S.

$$p = \mathrm{P}_0 \left[S \geq s_0 + |\delta| \right] + \mathrm{P}_0 \left[S \leq s_0 - |\delta| \right].$$

- Voor ons voorbeeld kunnen we deze kansen als volgt berekenen:

$$\begin{aligned}
\mathrm{P}_0\left[S \geq s_0 + |\delta|\right] &= P(S \geq 6155 \times 0.5 + |3175 - 6155 \times 0.5|) \\
&= P(S \geq 3175) \\
&= P(S = 3175) + P(S = 3176) + \ldots + P(S = 6155) \\
&= 0.0067
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}_0\left[S \leq s_0 - |\delta|\right] &= P(S \leq 6155 \times 0.5 - |3175 - 6155 \times 0.5|) \\
&= P(S \leq 2980) \\
&= P(S = 0) + \ldots + P(S = 2980) \\
&= 0.0067
\end{aligned}$$

- Binomiale distributie is symmetrisch als $\pi = 1/2$:

$$\mathrm{P}_0\left[S \geq s_0 + |\delta|\right] = \mathrm{P}_0\left[S \leq s_0 - |\delta|\right]$$

- Dat is niet langer het geval wanneer π afwijkt van 0.5.

```r
pi0 <- 0.5; s0 <- pi0 *n
delta <- abs(boys- s0)
delta
```

```
## [1] 97.5
```

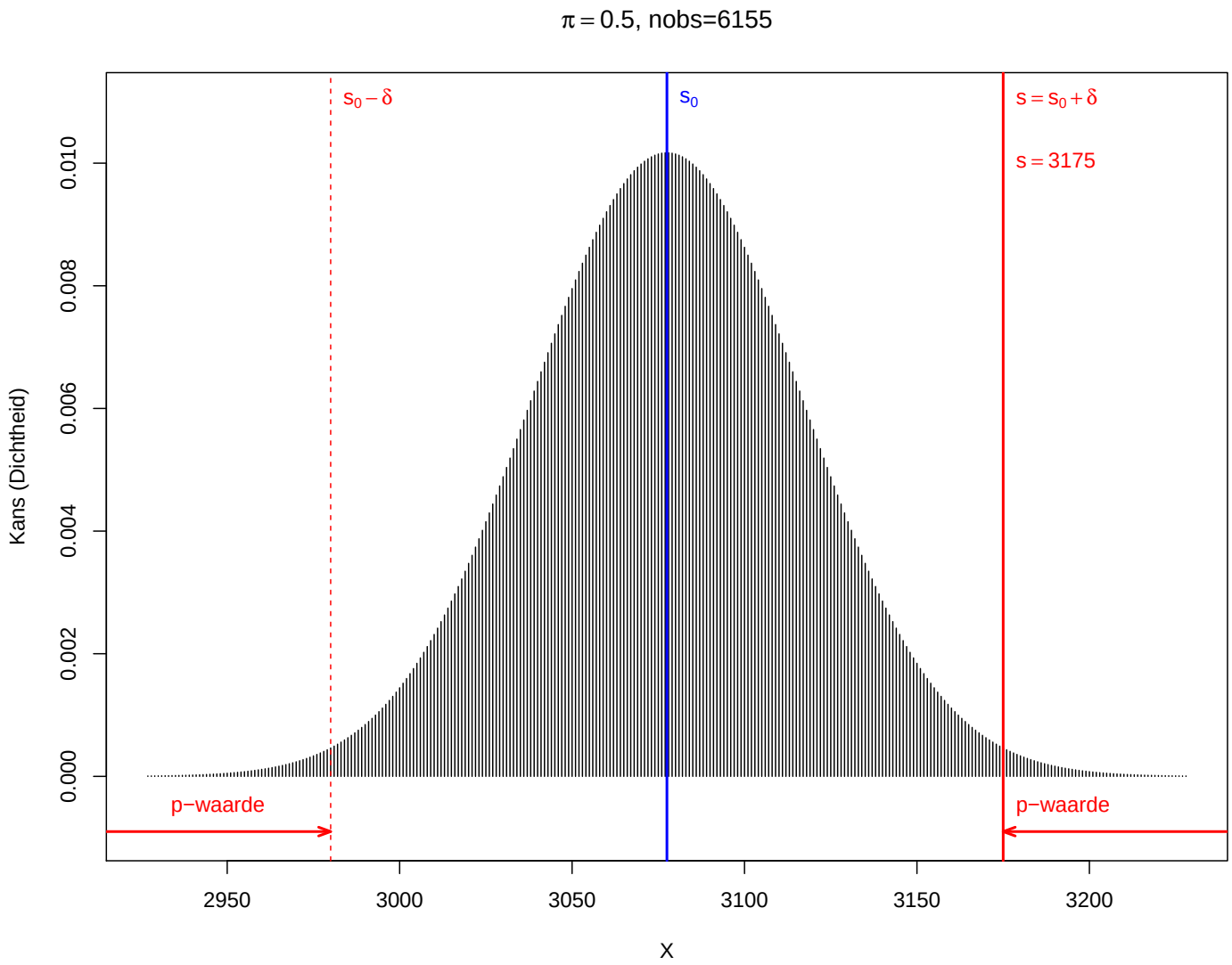
```r
sUp <- s0 + delta
sDown <- s0 -delta
c(sDown,sUp)
```

```
## [1] 2980 3175
```

```r
#Leg uit!
pUp <- 1-pbinom(sUp-1,n,pi0)
pDown <- pbinom(sDown,n,pi0)
p <- pUp+pDown
c(pUp,pDown, p)
```

```
## [1] 0.006699883 0.006699883 0.013399766
```

- Als $\pi = 1/2$, kans om door toeval minstens $\delta = 97.5$ jongens meer of minder te observeren dan het gemiddelde onder $H_0 : s_0 = 3077.5$, slechts 1.34% is: **de p-waarde van de binomiale test.**
- Heel onwaarschijnlijk om een dergelijk groot aantal jongens te observeren als in realiteit jongens en meisjes even waarschijnlijk zijn.
- Drukt uit dat de onderstelling dat jongens en meisjes even waarschijnlijk zijn, weinig gesteund wordt door de data.


π = 0.5, nobs=6155
s0 − δ
s0
s = s0 + δ
s = 3175
p−waarde
p−waarde
Kans (Dichtheid)
0.000
0.002
0.004
0.006
0.008
0.010
2950
3000
3050
3100
3150
3200
X

De test kan eveneens worden uitgevoerd a.d.h.v. de `binomial.test` functie in R.

```
binom.test(x=boys,n=n,p=pi0)
```

```
##
##  Exact binomial test
##
## data:  boys and n
## number of successes = 3175, number of trials = 6155, p-value
## 0.0134
## alternative hypothesis: true probability of success is not eq
## 95 percent confidence interval:
##  0.5032696 0.5283969
## sample estimates:
## probability of success
##              0.5158408
```

Op het 5% significantie-niveau besluiten we dat er gemiddeld meer kans is dat een ongeboren kind mannelijk dan vrouwelijk is.

8.2.2. Betrouwbaarheidsinterval op een proportie

- Schatter van de proportie van jongens in de populatie, is steekproefgemiddelde $\hat{\pi} = \bar{x} = 0.516$
- Standaard error is

$$SE_{\bar{x}} = \sqrt{\frac{\text{Var}[X]}{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We kunnen dit schatten o.b.v. de steekproef: $SE_{\bar{x}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.0064$.
- 95% BI via centrale limietstelling: $\hat{\pi} \pm 1.96 SE_{\hat{\pi}}$.

```
se=sqrt(pi*(1-pi)/n)
pi+c(-1,1)*qnorm(0.975)*se
```

```
## [1] 0.5033559 0.5283257
```

Betrouwbaarheidsinterval op een proportie in kleine steekproef?

- Inverteren van de one-sample test voor proporties.
- Stop alle waarden π_0 die niet verworpen worden door binomiale test op het 5% significantieniveau in BI
- Is geïmplementeerd in de `binom.test` functie.

```
BI <- binom.test(x=boys,n=n,p=pi0)$conf.int
BI
```

```
## [1] 0.5032696 0.5283969
## attr(,"conf.level")
## [1] 0.95
```

We verfiëren dit nu:

```
binom.test(x=boys,n=n,p=BI[1],alternative="greater")
```

```
##
##  Exact binomial test
##
## data:  boys and n
## number of successes = 3175, number of trials = 6155, p-value
## 0.025
## alternative hypothesis: true probability of success is greate
## 95 percent confidence interval:
##  0.5052779 1.0000000
## sample estimates:
## probability of success
##              0.5158408
```

```r
binom.test(x=boys,n=n,p=BI[2],alternative="less")
```

```
##
##  Exact binomial test
##
## data:  boys and n
## number of successes = 3175, number of trials = 6155, p-value
## 0.025
## alternative hypothesis: true probability of success is less t
## 95 percent confidence interval:
##  0.0000000 0.5263925
## sample estimates:
## probability of success
##              0.5158408
```

- Het exacte BI is te verkiezen boven het BI dat gebaseerd is op de CLT.
- Voor Saksen-studie ligt BI o.b.v. CLT heel dicht bij exacte BI: grote steekproef ($n = 6155$).

8.2.3. Conclusie

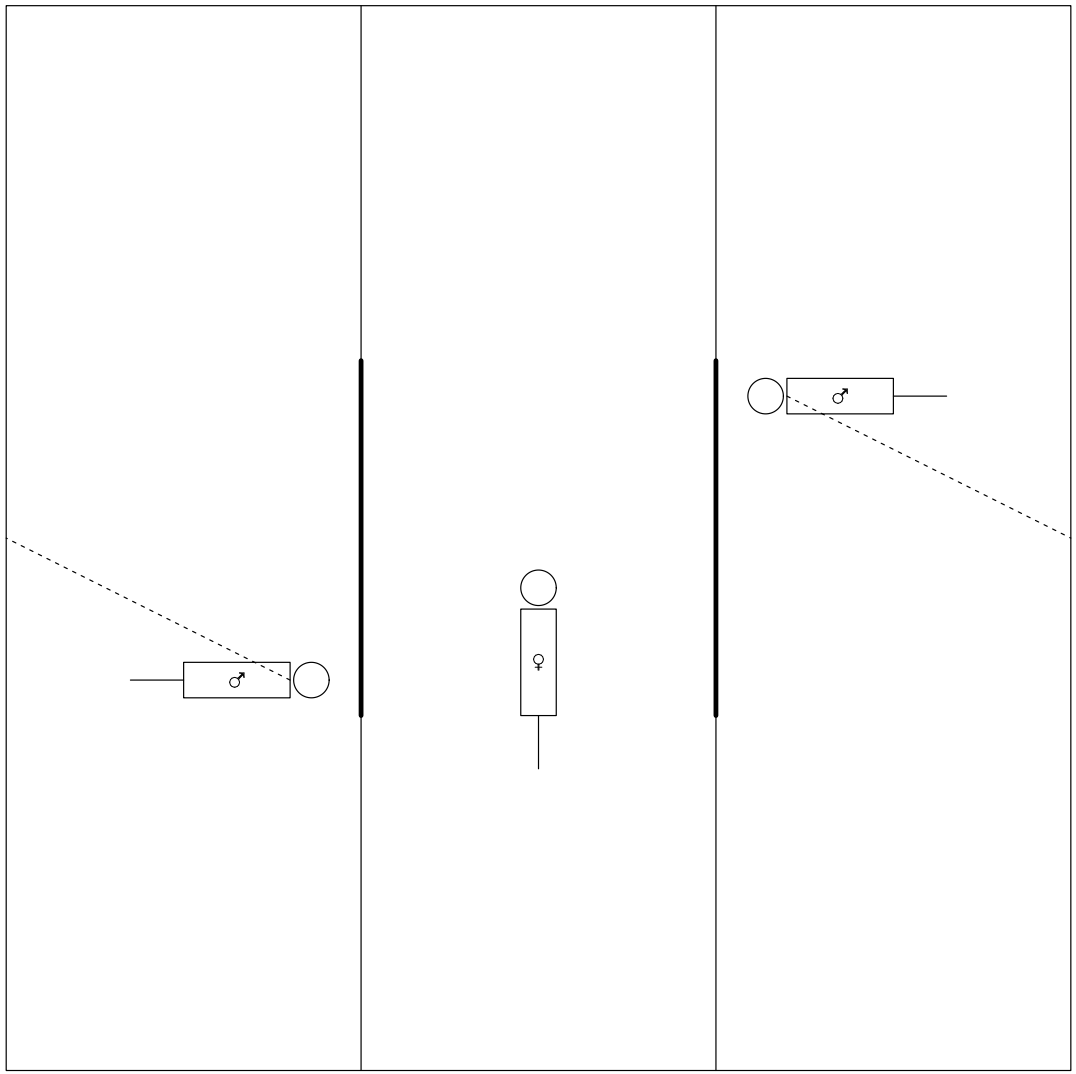
- Merk op dat het testen voor een proportie kan gezien worden als het equivalent van een one-sample t-test voor binaire data.
- Voor de Saksen populatie besluiten we op het 5% significantieniveau dat er meer kans is dat een ongeboren kind mannelijk dan vrouwelijk is ($p = 0.013$). De kans dat een ongeboren kind mannelijk is, bedraagt 51.6% (95% BI [50.3,52.8]%).

8.3. Toets voor associatie tussen 2 kwalitatieve variabelen

8.3.1. Gepaarde gegevens

- 2 keer zelfde individu meten
- bijvoorbeeld, vóór en na blootstelling aan de experimentele stof
- telkens de categorische uitkomst te observeren.
- Hier enkel: *gepaarde binaire uitkomsten*
- Statistische analyse moet rekening houden met de paring.

8.3.1.1. Voorbeeld: partnerkeuze van seksueel mature vrouwelijke *Campbelli* dwerghamster (Rogovin et al. 2017)



Voorbeeld: partnerkeuze van seksueel mature vrouwelijke *Campbelli* dwerghamster (Rogovin et al. 2017)

- Na 3 minuten, scheidingswand weg
- aggressief vs niet-agressief mannentje
- Elk vrouwtje onderging tweemaal de test: na verblijf in
  - vijandige omgeving (hoge populatie, weinig voedsel, veel concurrentie)
  - vriendelijkere omgeving

Tabel 2: Kruistabel van partnerkeuze bij dwerghamster.

	vriendelijk-agressief	vriendelijk-niet-agressief	totaal
vijandig-agressief	3 (e)	17 (f)	20
vijandig-niet-agressief	1 (g)	13 (h)	14
totaal	4	30	34

	vriendelijk-agressief	vriendelijk-niet-agressief	totaal
vijandig-agressief	3 (e)	17 (f)	20
vijandig-niet-agressief	1 (g)	13 (h)	14
totaal	4	30	34

- $\pi_1 = P[\text{agressief mannetje} \mid \text{verblijf vijandige omgeving}$
- $\hat{p}i_1 = (e+f)/n$, waarbij $n = e+f+g+h$.
- $\pi_0 = P[\text{agressief mannetje} \mid \text{verblijf vriendelijke omgeving}$
- $\hat{p}i_0 = (e+g)/n$ -Absoluut riscoverschil (ARV)

$$\widehat{\text{ARV}} = \hat{\pi}_1 - \hat{\pi}_0 = \frac{e+f}{n} - \frac{e+g}{n} = \frac{f-g}{n}$$

- Enkel beïnvloed door aantallen discordante paren f en g

- Standaard error op ARV

$$\text{SE}_{\widehat{\text{ARV}}} = \frac{1}{n}\sqrt{f + g - \frac{(f-g)^2}{n}}$$

- Als er voldoende gegevens zijn, kan men een $(1-\alpha)100\%$ BI op ARV

$$\left[\widehat{\text{ARV}} - z_{\alpha/2}\text{SE}_{\widehat{\text{ARV}}}, \widehat{\text{ARV}} - z_{\alpha/2}\text{SE}_{\widehat{\text{ARV}}}\right]$$

of

$$\left[\frac{f-g}{n} - \frac{z_{\alpha/2}}{n}\sqrt{f + g - \frac{(f-g)^2}{n}}, \frac{f-g}{n} + \frac{z_{\alpha/2}}{n}\sqrt{f + g - \frac{(f-g)^2}{n}}\right]$$

```r
hamster <- matrix(c(3,17,1,13),ncol=2,byrow=TRUE)
rownames(hamster) <- c("vijandig-agressief", "vijandig-niet-agre
colnames(hamster) <- c("vriendelijk-agressief","vriendelijk-niet
```

```r
f=hamster[1,2]; g=hamster[2,1] ;n=sum(hamster)
riskdiff=(f-g)/n
riskdiff
```

```
## [1] 0.4705882
```

```r
se=sqrt(f+g-(f-g)^2/n)/n
se
```

```
## [1] 0.09517144
```

```r
bi=riskdiff+c(-1,1)*qnorm(0.975)*se
bi
```

```
## [1] 0.2840556 0.6571208
```

$$\widehat{\text{ARV}} = \frac{17-1}{34} = 0.471$$

of 47.1%. - De standaard error

$$\text{SE}_{\widehat{\text{ARV}}} = \frac{1}{34}\sqrt{17+1-\frac{(17-1)^2}{34}} = 0.0952$$

- Een 95% betrouwbaarheidsinterval voor het absolute risicoverschil op de keuze van een agressief mannetje tussen een verblijf in een vijandige en vriendelijke omgeving is bijgevolg

$$[0.471 - 1.96 \times 0.0952, 0.471 + 1.96 \times 0.0952] = [0.284, 0.658]$$

- We hebben dus geschat dat het absolute risico met 95% kans in het interval [28.4,65.8]% ligt.

8.3.1.2. McNemar test

	vriendelijk-agressief	vriendelijk-niet-agressief	totaal
vijandig-agressief	3 (e)	17 (f)	20
vijandig-niet-agressief	1 (g)	13 (h)	14
totaal	4	30	34

- Toetsen of de risico's verschillen tussen de vijandige en vriendelijke omgeving.
- Enkel de discordante paren leveren hier informatie over.
- $f > g$ indicatie tegen H_0 : *partnerkeuzenietgeassocieerdmetomgeving*
- Kans evalueren dat in een lukraak discordant paar, vrouwtje na verblijf in een vijandige omgeving kiest voor het agressieve mannetje.
- Deze kans wordt geschat als

$$\frac{f}{f+g}$$

$$
\begin{aligned}
\mathrm{E}\left[f/(f+g)\right] &\overset{H_0}{=} 0.5 \\
f &\overset{H_0}{\sim} \mathrm{Binom}(n = f+g, \pi = 0.5) \\
\mathrm{SE}_{\frac{f}{f+g}} &\overset{H_0}{=} \sqrt{(f+g) \times 0.5 \times 0.5} = \frac{\sqrt{f+g}}{2}
\end{aligned}
$$

- Asymptotisch one-sample z-test (o.b.v. normale verdeling)

$$z = \frac{f - (f+g)/2}{\sqrt{f+g}/2} = \frac{f-g}{\sqrt{f+g}}$$

- De Normale benadering is goed als

$$f \times g/(f+g) \geq 5$$

- In kleine steekproeven is het meer aangewezen om een continuïteitscorrectie te gebruiken d.m.v. de toetsingsgrootheid

$$\frac{|f-g|-1}{\sqrt{f+g}}$$

De **Mc Nemar test** analogon van de gepaarde t-test voor binaire, kwalitatieve i.p.v. continue variabelen.

We voeren nu de analyse uit voor het hamstervoorbeeld in R:

```
correct=f*g/(f+g)
correct
```

```
## [1] 0.9444444
```

```
#continuiteitscorrectie
t= (abs(f-g)-1)/sqrt(f+g); t
```

```
## [1] 3.535534
```

```
p=(1-pnorm(t))*2; p
```

```
## [1] 0.000406952
```

- Voor het dwerghamster voorbeeld observeren we dat $f \times g/(f+g) = 0.944 < 5 \rightarrow$ continuïteitscorrectie
- De kans dat een Normaal verdeelde toevalsveranderlijke groter is dan 3.54 of kleiner is dan -3.54 bedraagt 0.0407%: *p-waarde*

In R kan de analyse ook worden uitgevoerd a.d.h.v. de `mcnemar.test` functie

```
mcnemar.test(hamster)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  hamster
## McNemar's chi-squared = 12.5, df = 1, p-value = 0.000407
```

- We verwerpen bijgevolg de nulhypothese op het 5% significantieniveau en
- Besluiten dat de parternkeuze extreem significant geassocieerd is met de omgeving.
- We zien dat hier eveneens de continuïteitscorrectie werd uitgevoerd en dat we exact dezelfde p-waarde bekomen.

- Normale benadering van deze toetstatistiek niet ideaal is omdat $f \times g/(f+g) = 0.944 < 5$.
- Aangewezen om een exacte toets te gebruiken op basis van binomiale test

```
binom.test(x=f,n=f+g,p=0.5)
```

```
##
##  Exact binomial test
##
## data:  f and f + g
## number of successes = 17, number of trials = 18, p-value =
## 0.000145
## alternative hypothesis: true probability of success is not eq
## 95 percent confidence interval:
##  0.7270564 0.9985944
## sample estimates:
## probability of success
##              0.9444444
```

8.3.1.3. Conclusie

- Op basis van de exacte test besluiten we eveneens dat de parternkeuze extreem significant geassocieerd is met de omgeving ($p < 0.001$).
- De kans op de keuze van een agressief mannetje ligt 47.1% hoger als een dwerghamster vrouwtje zich in een vijandige omgeving bevindt dan wanneer ze zich in een vriendelijke omgeving bevindt (95% BI [28.4,65.7]%).

8.3.2. Ongepaarde gegevens

Genetische associatie studie (zie Sectie 3.6.2)

- Genetische associatiestudie polymorfismen in het BRCA1 gen geassocieerd is met borstkanker?
- Retrospectieve case-controle studie met 800 borstkankercases en 572 controles
- R object is opgeslagen in de file `brca.rda`

```
load("dataset/brca.rda")
head(brca)
```

```
##    cancer variant variant2
## 1 control pro/pro   andere
## 2 control pro/pro   andere
## 3 control pro/pro   andere
## 4 control pro/pro   andere
## 5 control pro/pro   andere
```

Genotype	Controles	Cases	Totaal
Pro/Pro	266 (a)	342 (d)	608 (a+d)
Pro/Leu	250 (b)	369 (e)	619 (b+e)
Leu/Leu	56 (c)	89 (f)	145 (c+f)
Totaal	572 (a+b+c)	800 (d+e+f)	1372 (n)

- In case-controle studies kiest men een vast aantal cases en controles en spoort men voor hen op welke blootstellingen ze in het verleden ondervonden hebben.
- Dergelijke studies noemt men ook retrospectief
- Onmogelijk om het risico's and riscoverschillen op borstkanker te schatten: proportie van cases en controles weerspiegelt populatie niet!

Genotype	Controles	Cases	Totaal
Pro/Pro	266 (a)	342 (d)	608 (a+d)
Pro/Leu	250 (b)	369 (e)	619 (b+e)
Leu/Leu	56 (c)	89 (f)	145 (c+f)
Totaal	572 (a+b+c)	800 (d+e+f)	1372 (n)

- Wel mogelijk om kans te schatten om allel Leu/Leu
  - cases: $\pi_1 = f/(d+e+f) = 89/800 = 11.1\%$
  - controles:$\pi_0 = c/(a+b+c) = 56/572 = 9.8\%$
- Relatief risico op blootstelling voor cases versus controles is bijgevolg $11.1/9.8 = 1.14$.
- Vrouwen met borstkanker hebben dus 14% meer kans om de allelcombinatie Leu/Leu te hebben op het BRCA1 gen dan vrouwen zonder borstkanker.
- Dit suggereert een associatie, maar drukt iet uit hoeveel hoger het risico op borstkanker is voor vrouwen met de allelcombinatie Leu/Leu dan voor andere vrouwen
- Andere risicomaat?

$$Odds = \frac{p}{1-p}$$

waarbij p de kans is op die gebeurtenis.

Transformatie van het risico, met volgende eigenschappen:

- de odds neemt waarden aan tussen nul en oneindig.
- de odds is gelijk aan 1 als en slechts als de kans zelf gelijk is aan $1/2$.
- de odds neemt toe als de kans toeneemt.
- populair bij gokkers: hoeveel waarschijnlijker het is om te winnen dan om te verliezen

Genotype	Controles	Cases	Totaal
Pro/Pro	266 (a)	342 (d)	608 (a+d)
Pro/Leu	250 (b)	369 (e)	619 (b+e)
Leu/Leu	56 (c)	89 (f)	145 (c+f)
Totaal	572 (a+b+c)	800 (d+e+f)	1372 (n)

Odds op allel Leu/Leu

- Cases: $\text{odds}_1 = \frac{f/(d+e+f)}{(d+e)/(d+e+f)} = f/(d+e) = 89/711 = 0.125$.
  Vrouwen met borstkanker hebben ongeveer 8 keer meer kans om de allelcombinatie Leu/Leu niet te hebben op het BRCA1 gen dan om het wel te hebben.
- Controles: $\text{odds}_2 = c/(a+b) = 56/516 = 0.109$.
- Associatie tussen blootstelling en uitkomst:

$$OR_{Leu/Leu} = \frac{\text{odds}_T}{\text{odds}_C} = \frac{f/(d+e)}{c/(a+b)} = \frac{f/(d+e)}{c/(a+b)} = 1.15$$

Genotype	Controles	Cases	Totaal
Pro/Pro	266 (a)	342 (d)	608 (a+d)
Pro/Leu	250 (b)	369 (e)	619 (b+e)
Leu/Leu	56 (c)	89 (f)	145 (c+f)
Totaal	572 (a+b+c)	800 (d+e+f)	1372 (n)

- Was de bovenstaande studie echter een volledig lukrake steekproef geweest (waarbij het aantal cases en controles niet per design werden vastgelegd),
- dan konden we daar ook de odds ratio op borstkanker berekenen voor mensen met versus zonder het allel Leu/leu.

$$OR_{case} = \frac{\frac{f}{c}}{\frac{(d+e)}{(a+b)}} = \frac{f(a+b)}{c(d+e)} = OR_{Leu/Leu} = 1.15,$$

- OR is een symmetrische maat! OR op borstkanker kan wel worden geschat!
- De odds op borstkanker is bijgevolg 15% hoger bij vrouwen met die specifieke allelcombinatie.

- Is verschil groot genoeg zodat we het effect die we in de steekproef zien kunnen veralgemenen naar de populatie toe.
- Hiertoe zullen we de kruistabel eerst herschrijven tot een 2x2 tabel

Genotype	Controles	Cases	Totaal
andere	516 (a)	711 (c)	1227 (a+c)
Leu/Leu	56 (b)	89 (d)	145 (b+d)
Totaal	572 (a+b)	800 (c+d)	1372 (n)

8.3.3. De Pearson Chi-kwadraat test voor ongepaarde gegevens

- Testen voor associatie tussen de categorische blootstelling (bvb. variant, X) en de categorische uitkomst (bvb. ziekte, Y).

  H_0 : Er is geen associatie tussen X en Y vs H_1 : X en Y zijn geassocieerd

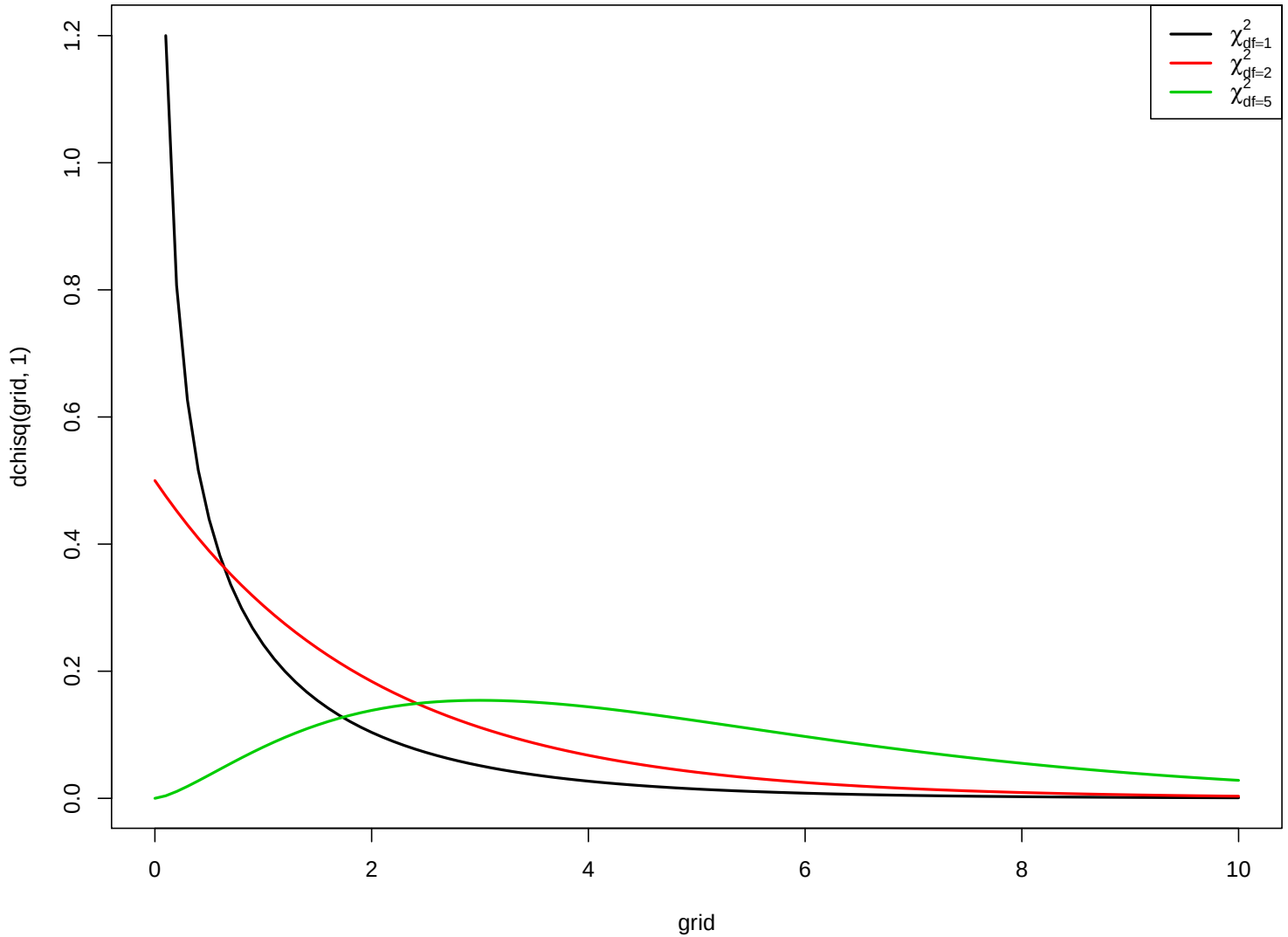
- Beschouw de rijtotalen $n_{\text{andere}} = a + c$, $n_{\text{leu,leu}} = b + d$ enerzijds en
- de kolomtotalen $n_{\text{contr}} = a + b$ en $n_{\text{case}} = c + d$ anderzijds.
- Zij verstrekken informatie over de *marginale verdeling* van de blootstelling (bvb. variant, X) en de uitkomst (bvb. ziekte, Y), maar niet over de associatie tussen die veranderlijken.
- Onder H_0 zijn X en Y onafhankelijk zijn en verwacht men een proportie $(b + d)/n$ van $a + b$ controles met een Leu/Leu variant, of dat $(a + b)(b + d)/n$ een Leu/Leu variant hebben
- Analoog kan men verwachte aantal E_{ij} berekenen dat onder de nulhypothese in *elke cel* van de 2×2 tabel zou liggen.

- E_{11} = het verwachte aantal onder H_0 in de (1,1)-cel = $1227 \times 572/1372 = 511.5$;
- E_{12} = het verwachte aantal onder H_0 in de (1,2)-cel = $1227 \times 800/1372 = 715.5$;
- E_{21} = het verwachte aantal onder H_0 in de (2,1)-cel = $145 \times 572/1372 = 60.45$;
- E_{22} = het verwachte aantal onder H_0 in de (2,2)-cel = $145 \times 800/1372 = 84.55$;

Toetsstatistiek:

$$X^2 = \frac{(|O_{11} - E_{11}| - .5)^2}{E_{11}} + \frac{(|O_{12} - E_{12}| - .5)^2}{E_{12}} + \frac{(|O_{21} - E_{21}| - .5)^2}{E_{21}} + \frac{(|O_{22} - E_{22}| - .5)^2}{E_{22}}$$

$$X^2 \xrightarrow{H_0} \chi^2(df = 1)$$


dchisq(grid, 1)
grid
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
$\chi^2_{df=1}$
$\chi^2_{df=2}$
$\chi^2_{df=5}$

- Een grote waarde van de toetsingsgrootheid geeft een indicatie van een afwijking van de nulhypothese.
- Concreet zal een toets op het $\alpha 100\%$ significantieniveau de nulhypothese verwerpen zodra de geobserveerde waarde van de toetsingsgrootheid het $100\%(1-\alpha)$-percentiel, $\chi^2_{1,\alpha}$, van de χ^2_1-verdeling overschrijdt.
- Ze kan niet verwerpen in het andere geval.
- De p-waarde voor een 2-zijdige toets is in dit geval de kans om een grotere waarde voor de toetsingsgrootheid te observeren dan de geobserveerde waarde x^2 als de nulhypothese waar is.
- Dit is de kans dat een χ^2_1-verdeelde toevalsveranderlijke waarden groter dan x^2 aanneemt.

```r
expected <- matrix(0,nrow=2,ncol=2)
for (i in 1:2)
    for (j in 1:2)
        expected[i,j] <-
            sum(brcaTab2[i,])*sum(brcaTab2[,j])/sum(brcaTab2)
expected
```

```
##          [,1]     [,2]
## [1,] 511.5481 715.4519
## [2,]  60.4519  84.5481
```

```r
x2 <- sum((abs(brcaTab2-expected) - .5)^2/expected)
1-pchisq(x2,1)
```

```
## [1] 0.481519
```

- Omdat de observaties O_{ij} in feite discrete getallen zijn, kan de toetsingsgrootheid X^2 slechts discrete waarden aannemen en kan een continue verdeling zoals de χ^2_1-verdeling slechts een benadering zijn voor haar werkelijke verdeling.
- Om de discrete verdeling beter bij de continue χ^2_1-verdeling te doen aansluiten, heeft men in de uitdrukking van de toetsingsgrootheid voor elke cel telkens 0.5 afgetrokken.
- Dit wordt een *continuïteitscorrectie* genoemd.
- In dit geval gaat het om de correctie van Yates en noemt men deze toets dan ook de *Pearson Chi-kwadraat toets met Yates correctie*.
- Wanneer de correctie niet gebruikt wordt (d.w.z. wanneer de getallen '0.5' in de uitdrukking voor X^2 door 0 vervangen worden), dan spreekt men van de *Pearson Chi-kwadraat toets*.

In R kan je deze toetsen uitvoeren door de optie `correct` op TRUE of FALSE te zetten:

```
chisq.test(brcaTab2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  brcaTab2
## X-squared = 0.49542, df = 1, p-value = 0.4815
```

```
chisq.test(brcaTab2,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  brcaTab2
## X-squared = 0.62871, df = 1, p-value = 0.4278
```

- Zelfs met continuïteitscorrectie is χ^2_1 benadering slechts verantwoord als in geen enkele van de cellen het verwachte aantal onder H_0 kleiner is dan 5.
- Wanneer de χ^2-benadering niet verantwoord is, kan men een *Fisher's exact test* uitvoeren.
- De nulhypothese van deze test is eveneens dat X en Y onafhankelijk zijn, en de alternatieve hypothese dat X en Y afhankelijk zijn.
- Een nadeel van de exacte test, is dat ze conservatiever is

```
fisher.test(brcaTab2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  brcaTab2
## p-value = 0.4764
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7998798 1.6738449
## sample estimates:
```

8.3.3.1. Uitbreiding naar categorische variabelen met meerdere niveaus

- χ^2-toets kan ook als minstens 1 van de discrete variabelen X en Y meer dan 2 mogelijke waarden aanneemt
- Opnieuw: nulhypothese $H_0 : X$ en Y zijn onafhankelijk (niet-geassocieerd), ten opzichte van het tweezijdig alternatief $H_A : X$ en Y zijn niet onafhankelijk (geassocieerd).
- Als de variabele voorgesteld op de rijen r mogelijke uitkomsten heeft en die op de kolommen c mogelijke uitkomsten, dan noemt men de kruistabel die X tegenover Y uitzet, een $r \times c$ tabel.
- Zoals voorheen vergelijkt men het aantal geobserveerde waarden in cel (i, j), O_{ij} genoteerd, met het aantal verwachte waarden onder de nulhypothese, E_{ij} -Opnieuw is E_{ij} product van het i-de rijtotaal met het j-de kolomtotaal gedeeld door het algemene totaal.

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Men kan aantonen dat ze een Chi-kwadraat verdeling volgt met $(r-1) \times (c-1)$ vrijheidsgraden als de nulhypothese waar is.
- De continuïteitscorrectie wordt meestal niet gebruikt bij meer dan 2 rijen of kolommen.
- **Pearson χ^2 test** is analogon van de one-way variantie-analyse voor kwalitatieve i.p.v. continue variabelen.

```r
brcaTab <- table(brca$variant,brca$cancer)
chisq.test(brcaTab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  brcaTab
## X-squared = 2.0551, df = 2, p-value = 0.3579
```

- Om te onderzoeken of het BRCA1 gen geassocieerd is met borstkanker, berekenen we de Pearson chi-kwadraat toets voor de case-controle studie uit Tabel @ref(tab:leu3).
- De toetsingsgrootheid bedraagt nu 2.055 en volgt een Chi-kwadraat verdeling met 2 vrijheidsgraden. De kans dat zo'n χ^2- verdeelde toevalsveranderlijke extremer is dan 2.055, bedraagt 36%.
- Op het 5% significantieniveau kunnen we dus niet besluiten dat het BRCA1 gen geassocieerd is met borstkanker.

8.4. Logistische regressie

- Raamwerk voor het modelleren van binaire data (vb. kanker vs geen kanker): *logistische regressie-modellen*.
- Binaire gegevens modelleren a.d.h.v. continue en/of dummy variabelen.
- De modellen veronderstellen dat de observaties voor subject $i = 1, \ldots, n$ onafhankelijk zijn en een Bernoulli verdeling volgen.
- Het logaritme van de odds wordt dan gemodelleerd d.m.v. een lineair model, ook wel lineaire predictor genoemd:

$$\begin{cases} Y_i & \sim & B(\pi_i) \\ \log \frac{\pi_i}{1-\pi_i} & = & \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} \end{cases} \tag{2}$$

8.4.1. Categorische predictor

- Borstkanker voorbeeld: is BRCA 1 variant geassocieerd is met het krijgen van borstkanker.
- Net zoals in de anova context, factor in het regressieraamwerk d.m.v. dummy variabelen.
- 1 dummy variable minder nodig hebben dan er groepen zijn.
- Voor het BRCA 1 voorbeeld zijn dus twee dummy variabelen nodig en kunnen we de data dus modelleren met onderstaande lineaire predictor:

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- Waarbij de predictoren dummy-variabelen zijn:

$$x_{i1} = \begin{cases} 1 & \text{als subject } i \text{ heterozygoot is, Pro/Leu variant} \\ 0 & \text{als subject } i \text{ homozygoot is, (Pro/Pro of Leu/Leu variant)} \end{cases}.$$

$$x_{i2} = \begin{cases} 1 & \text{als subject } i \text{ homozygoot is in de Leucine mutatie: Leu/Leu} \\ 0 & \text{als subject } i \text{ niet homozygoot is in de Leu/Leu variant} \end{cases}$$

- Homozygositeit in het wild type allel Pro/Pro wordt voor dit model de **referentiegroep**.

Het model wordt als volgt in R gefit:

```
brcaLogit <- glm(cancer~variant,data=brca,family=binomial)
summary(brcaLogit)
```

```
##
## Call:
## glm(formula = cancer ~ variant, family = binomial, data = brc
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.379  -1.286   1.017   1.017   1.073
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.25131    0.08175   3.074  0.00211 **
## variantpro/leu  0.13802    0.11573   1.193  0.23302
## variantleu/leu  0.21197    0.18915   1.121  0.26243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
```

```r
anova(brcaLogit,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cancer
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     1371     1863.9
## variant  2   2.0562      1369     1861.9   0.3577
```

De χ^2-test op het logistische regressiemodel geeft eveneens aan dat er geen significante associatie is tussen de uitkomst (voorkomen van kanker) en de factor (de genetische variant van het BRCA gen) ($p = 0.358$). De p-waarde is bijna equivalent aan de p-waarde van de χ^2-test uit de vorige sectie.

- Significante associatie? Post-hoc tests om te evalueren welke odds ratio's verschillend zijn.
- Voor het BRCA1 voorbeeld zouden we uiteraard geen post-hoc testen
- Toch illustratie zodat jullie over de code beschikken

```r
library(multcomp)
posthoc=glht(brcaLogit,linfct=mcp(variant = "Tukey"))
posthocTests=summary(posthoc)
posthocTests
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: glm(formula = cancer ~ variant, family = binomial, data
##
## Linear Hypotheses:
##                          Estimate Std. Error z value Pr(>|z|)
## pro/leu - pro/pro == 0  0.13802    0.11573   1.193    0.449
## leu/leu - pro/pro == 0  0.21197    0.18915   1.121    0.493
## leu/leu - pro/leu == 0  0.07395    0.18922   0.391    0.917
## (Adjusted p values reported -- single-step method)
```

- BI's kunnen als volgt worden teruggetransformeerd naar odds ratios:

```
OR=exp(posthocBI$confint)
OR
```

```
##                      Estimate       lwr      upr
## pro/leu - pro/pro 1.148000 0.8771158 1.502543
## leu/leu - pro/pro 1.236111 0.7962014 1.919075
## leu/leu - pro/leu 1.076752 0.6934417 1.671942
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 2.325567
```

- De odds ratios die worden bekomen met het logistisch regressiemodel zijn exact gelijk aan de odds ratios die we zouden bekomen op basis van Tabel:
- vb. $OR_{\text{Leu/Leu-Pro/Pro}} = 89 \times 266/(56 \times 342) = 1.236$.
- Merk op dat de statistische besluitvorming bij logistische modellen beroep doet op asymptotische theorie.

8.4.2. Continue predictor

- Toxicologisch effect van koolstofdisulfide (CS_2) op kevers.
- De centrale onderzoeksvraag is of de concentratie van CS_2 een effect heeft op de mortaliteit (i.e. kans op sterven) van de kevers?

Design - 32 onafhankelijk experimenten - Telkens 1 kever blootgesteld aan één van 8 concentraties (mg/l) van CS_2 voor een gegeven periode. - De uitkomst van het experiment is: de kever sterft ($y = 1$) of de kever overleeft ($y = 0$).

```
load("dataset/kevers.rda")
head(kevers)
```

```
##    dosis status
## 1 169.07      1
## 2 169.07      0
## 3 169.07      0
## 4 169.07      0
## 5 172.42      1
```

We bouwen nu een logistisch regressiemodel waarbij we de log odds modelleren in functie van de dosis x_i:

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 \times x_i.$$

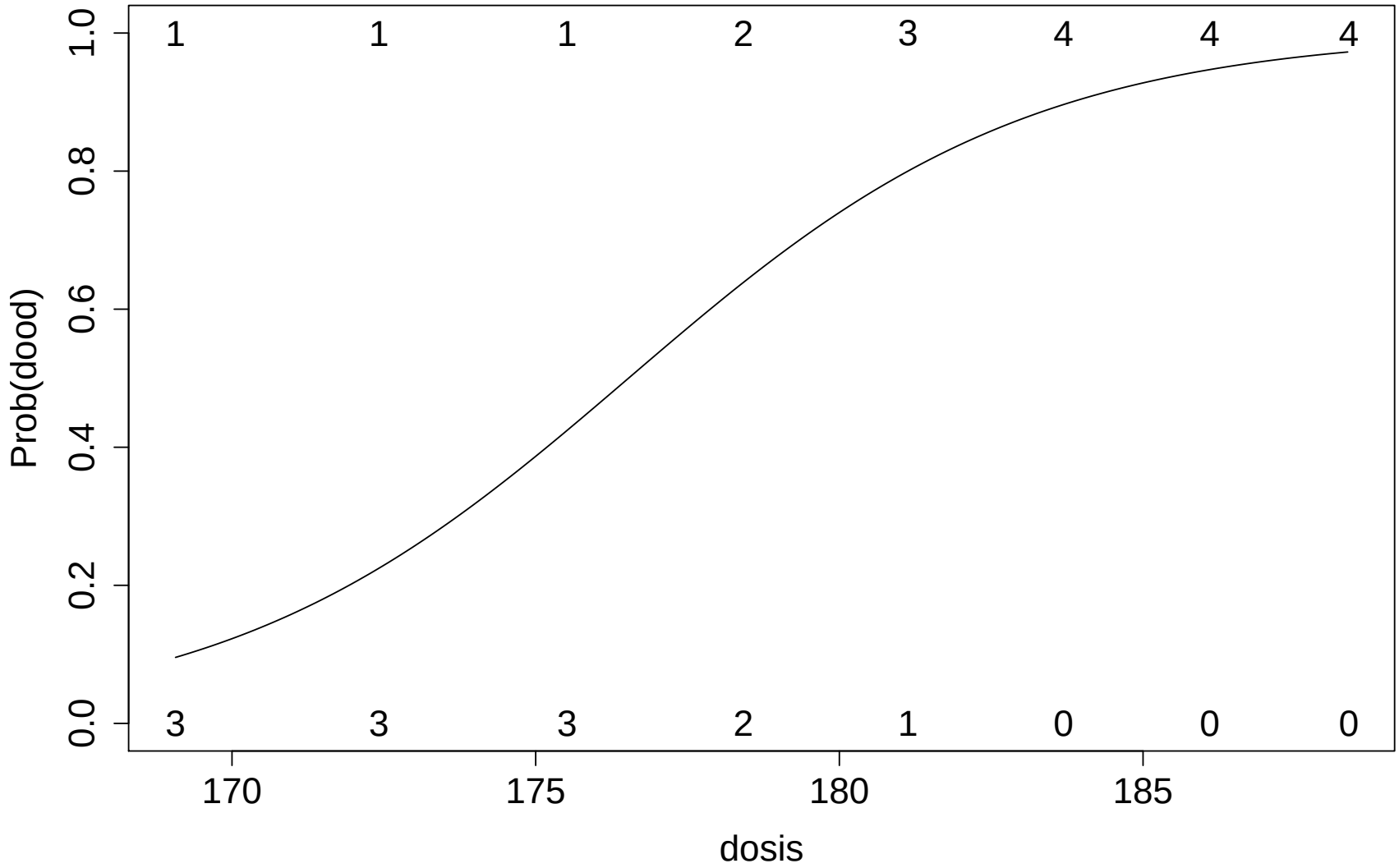
```
keverModel<-glm(status~dosis,data=kevers,family=binomial)
summary(keverModel)
```

```
##
## Call:
## glm(formula = status ~ dosis, family = binomial, data = kever
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7943  -0.7136   0.2825   0.5177   2.1670
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -53.1928    18.0046  -2.954  0.00313 **
## dosis         0.3013     0.1014   2.972  0.00296 **
##
```

- Intercept heeft als betekenis de log odds op mortaliteit wanneer er geen CS_2 gas wordt toegediend.
- Erg lage odds op sterfte ($\pi/(1-\pi) = \exp(-53.2)$) en dus op een kans die nagenoeg nul is.
- Merk op: heel sterke extrapolatie: minimum dosis in de dataset 169.07 mg/l.
- Geschatte odds ratio voor het effect van dosis op de mortaliteitskans is $\exp(0.3013) = 1.35$.
- Dus bij een toename van de dosis CS_2 met 1 mg/l, is de odds ratio voor de mortaliteit 1.35.

- We besluiten dat dit effect heel significant is ($p = 0.003$).
- Een toename in de CS_2 dosis doet de kans op sterven toenemen.

```
dosisGrid=seq(min(kevers$dosis),max(kevers$dosis),.1)
piHat=predict(keverModel,
          newdata=data.frame(dosis=dosisGrid),
          type="response")
```


Prob(dood)
0.0
0.2
0.4
0.6
0.8
1.0
1 1 1 2 3 4 4 4
3 3 3 2 1 0 0 0
170
175
180
185
dosis