Genetische Statistik

WS 2021/2022 Übung 2 - Grundlagen Statistik

Dr. Janne Pott (janne.pott@uni-leipzig.de)

November 09, 2021

Aufgabe 1: Diagnostischer Test

a) Definition **Sensitivität**, **Spezifität** und **Prävalenz**, Aufstellung einer schematische 4-Feldertafel.

b) Bestimmung von **positiv prädiktiver Wert** und **negativ prädiktiven Wert** in Abhängigkeit von Sensitivität (70%), Spezifität (95%) und Prävalenz:
   - 3% (z.B. Hausarztpraxis)
   - 20% (z.B. Altenheim)
   - 80% (z.B. Isolierstation)

c) Erstellen Sie für eine der Prävalenzen eine 4-Feldertafel (Gesamtfallzahl 1000).

Aufgabe 1: Lösung (1)

Sensitivität:

- $P(T+|K+)$
- WSK, dass Test positiv ist, unter der Bedingung, dass Erkrankung vorliegt.
- sensitivity = true positives / (true ppositives + false negatives)

Spezifität:

- $P(T-|K-)$
- WSK, dass Test negativ ist, unter der Bedingung, dass Erkrankung nicht vorliegt.
- specificity = true negatives / (true negatives + false positives)

Aufgabe 1: Lösung (2)

Prävalenz:

- $P(K+)$
- WSK einer Erkrankung in der Gesamtbevölkerung

	K+	K-
T+	true positives	false positives
T-	false negatives	true negatives

Allgemein gilt:

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

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

Aufgabe 1: Lösung (3)

Daraus folgt für uns:

$$
\begin{aligned}
P(K+|T+) &= \frac{P(K+ \cap T+)}{P(T+)} \\
&= \frac{P(T+|K+) \cdot P(K+)}{P(T+|K+) \cdot P(K+) + P(T+|K-) \cdot P(K-)} \\
&= \frac{P(T+|K+) \cdot P(K+)}{P(T+|K+) \cdot P(K+) + (1 - P(T-|K-)) \cdot (1 - P(K+))} \\
&= \frac{\text{Sens} \cdot \text{Präv}}{\text{Sens} \cdot \text{Präv} + (1 - \text{Spez}) \cdot (1 - \text{Präv})}
\end{aligned}
$$

Aufgabe 1: Lösung (4)

$$
\begin{aligned}
P(K-|T-) &= \frac{P(K-\cap T-)}{P(T-)} \\
&= \frac{P(T-|K-)\cdot(1-P(K+))}{P(T-|K-)\cdot(1-P(K+))+(1-P(T+|K+))\cdot P(K+)} \\
&= \frac{\text{Spez}\cdot(1-\text{Präv})}{\text{Spez}\cdot(1-\text{Präv})+(1-\text{Sens})\cdot\text{Präv}}
\end{aligned}
$$

Aufgabe 1: Lösung (5)

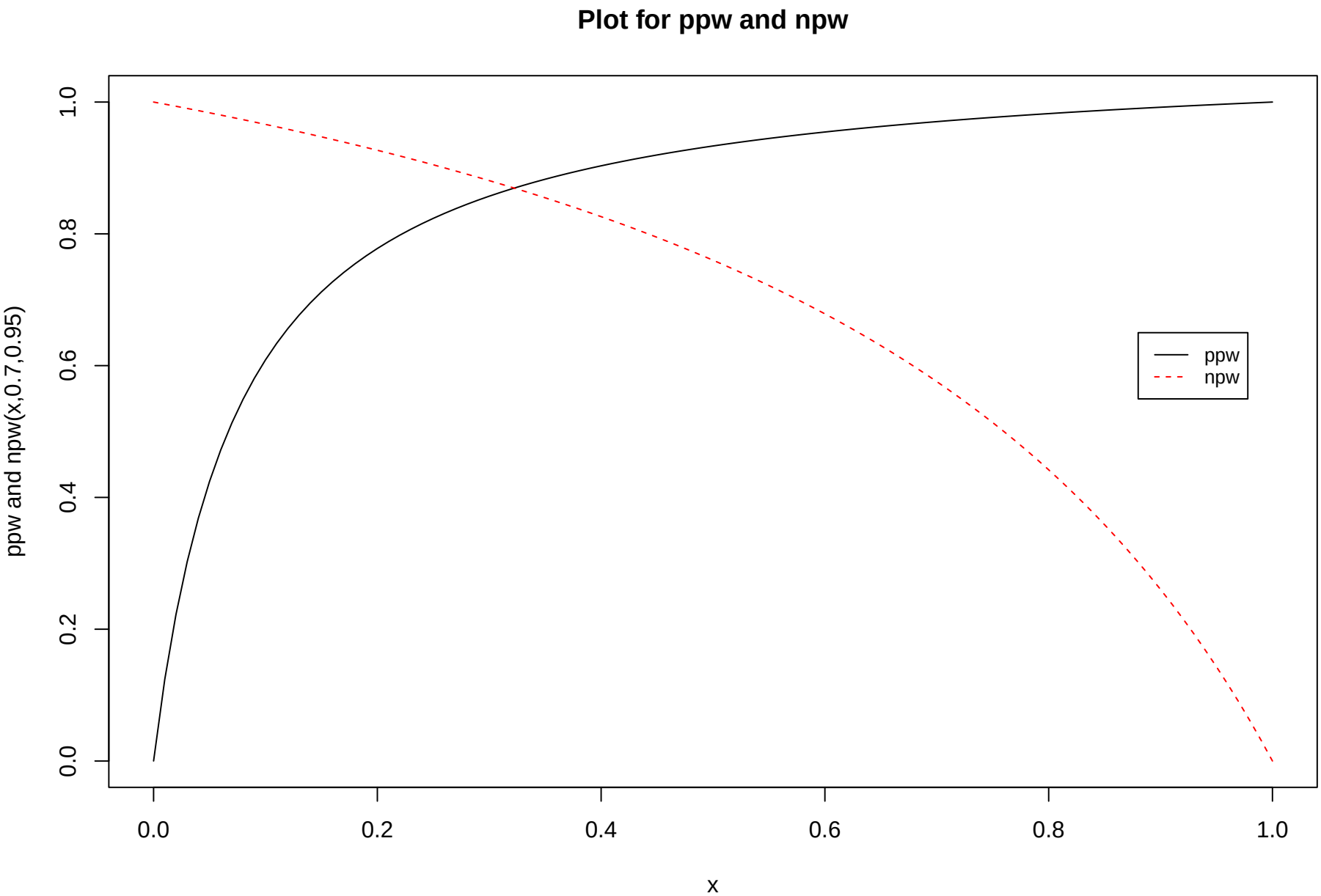
```
ppw<-function(prev,sens,spez) {
  return(sens*prev/(sens*prev+(1-spez)*(1-prev)))
}
npw<-function(prev,sens,spez) {
  return(spez*(1-prev)/(spez*(1-prev)+(1-sens)*prev))
}
ppw(c(0.03,0.2,0.8),0.7,0.95)
```

```
## [1] 0.3021583 0.7777778 0.9824561
```

```
npw(c(0.03,0.2,0.8),0.7,0.95)
```

```
## [1] 0.9903278 0.9268293 0.4418605
```

Aufgabe 1: Lösung (6)



Aufgabe 1: Lösung (7)

Prävalenz = 3% || 20% ||80%

	K+	K-		K+	K-		K+	K-	
T+	21	49	70	140	40	180	560	10	570
T-	9	921	930	60	760	820	240	190	430
	30	970	1000	200	800	1000	800	200	1000

Aufgabe 2: LogLikelihood

Bei der Genotypisierung eines biallelischen Markers in 10 diploiden Individuen haben sie viermal das Allel A beobachtet.

a) Warum n=20 Allelen? Verteilung?
b) **–Log-Likelihood** (Annahme Allelhäufigkeit 50%)?
c) **Maximum-Likelihood-Schätzer** $= k/n$?
d) **Maximum-Likelihood-Schätzer** und **–Log-Likelihood**
e) Ab wie vielen Treffern könnten Sie die Annahme, dass die wahre Allelfrequenz 50% beträgt, nicht mehr ablehnen? (Signifikanzniveau von 5%)

Aufgabe 2: Lösung (1)

- Biallelisch/diploid –> 2 Allele pro Individuum –> 2*10
- Verteilung eines Allels: ja/nein –> Bernoulli-Verteilung
- Verteilung der Treffer: Summe von Bernoulli-Ereignissen –> Binomial-Verteilung mit n=20 Allelen, k=5 Treffer und p=0.5 Allelhäufigkeit (Annahme)

Aufgabe 2: Lösung (2)

```
dbinom(4,20,0.5)
```

```
## [1] 0.004620552
```

```
-(dbinom(4,20,0.5,log=T))
```

```
## [1] 5.377241
```

Man muss also die Nullhypothese (wahre Verteilung von 0.5) verwerfen.

Aufgabe 2: Lösung (3)

Ableiten nach p –> Binomialkoeffizient kann vernachlässigt werden:

$$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \Rightarrow p^k \cdot (1-p)^{n-k}$$

Mittels Log-Transformation vereinfachen (ändert das Maximum nicht):

$$\Rightarrow ln(p^k \cdot (1-p)^{n-k}) = ln(p^k) + ln((1-p)^{n-k}) = k \cdot ln(p) + (n-k) \cdot ln(1-p)$$

Aufgabe 2: Lösung (4)

Ableiten nach p

$$
\begin{aligned}
\Rightarrow k\frac{1}{p} + (n-k)\frac{-1}{1-p} &\stackrel{!}{=} 0 \\
\Rightarrow \frac{k}{p} &= \frac{n-k}{1-p} \\
\Leftrightarrow k \cdot (1-p) &= (n-k) \cdot p \\
\Leftrightarrow k &= (n-k) \cdot p + kp = np \\
\Leftrightarrow p &= k/n
\end{aligned}
$$

Aufgabe 2: Lösung (5)

Wenn man nun $p = k/n = 0.2$ verwendet, kann man die Nullhypothese (wahre Verteilung von 0.2) nicht mehr ablehnen:

```
dbinom(4,20,0.2)
```

```
## [1] 0.2181994
```

```
-(dbinom(4,20,0.2,log=T))
```

```
## [1] 1.522346
```

Aufgabe 2: Lösung (6)

Kumulative Wahrscheinlichkeit

```
# VARIANTE 1: per Hand
n=20;p=0.5;alla=c()
a = for(k in 0:20) {
  a= choose(n,k)*p^k*(1-p)^(n-k)
  alla = c( alla, a)
}
table(cumsum(alla)>=0.05)
```

```
##
## FALSE  TRUE
##     6    15
```

```
cumsum(alla)[7]
```

```
## [1] 0.05765915
```

Aufgabe 2: Lösung (7)

Kumulative Wahrscheinlichkeit

```
# VARIANTE 2: Mittels pbinom
pbinom(6,20,0.5)
```

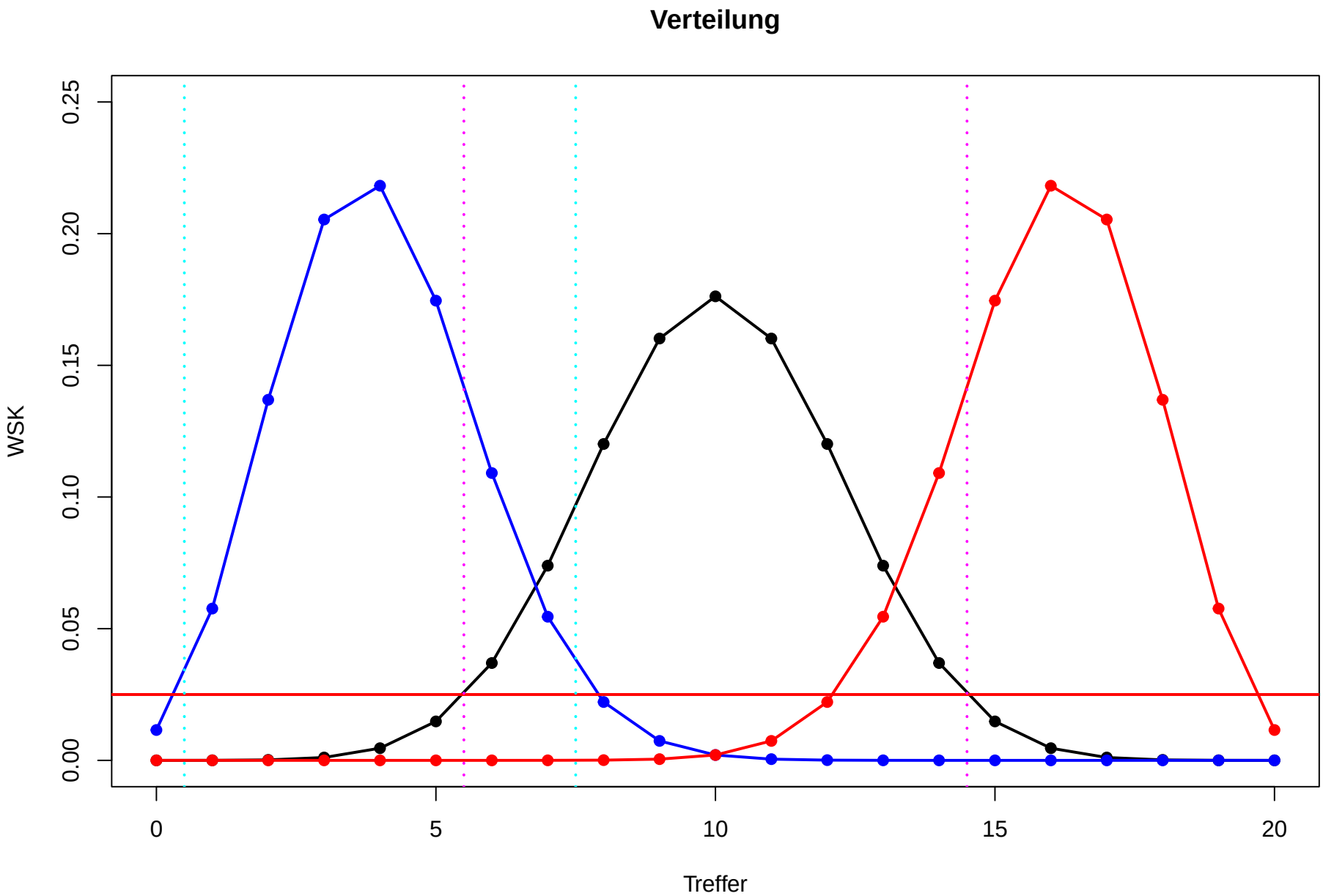
```
## [1] 0.05765915
```

```
# VARIANTE 3: Mittels binom.test
test<-binom.test(6,n=20,p=0.5,alternative="less")
test$p.value
```

```
## [1] 0.05765915
```

Ab 6 Treffern kann man die Nullhypothese p=0.5 nicht mehr ablehnen.

Aufgabe 2: Lösung (8)



Aufgabe 3: Konfidenzintervall

Berechnung der Konfidenzintervalle für $p_1 = 0.05$ und $p_2 = 0.5$!

Untersuchungsanlage	
Grundgesamtheit	Wahlberechtigte Bevölkerung ab 18 Jahren
Stichprobe	Repräsentative Zufallsauswahl / Dual Frame (Relation Festnetz-/Mobilfunknummern 60:40), Disproportionaler Ansatz (West/Ost 70:30)
Erhebungsverfahren	Telefoninterviews (CATI)
Fallzahl	1027 Befragte
Erhebungszeitraum	31.08. bis 02.09.2020
Schwankungsbreite	1.4 bis 3.1 Prozentpunkte (bei einem Anteilswert von 5 bzw. 50 Prozent)

Aufgabe 3: Lösung (1)

Frage: Würden Sie Partei x wählen –> ja/nein –> Bernoulli-Verteilung

$Var(X) = p * (1 - p), E(X) = p$

$P(E(X) \in [EW - 1.96 \cdot SE, EW + 1.96 \cdot SE]) = 0.95$

$SE = \sqrt{((Var(X))/N)} = \sqrt{((p \cdot (1 - p))/N)}$

$p_1 = 0.05$:

$SE = \sqrt{((0.05 \cdot 0.95)/1027)} = 6.80 \cdot 10^{-3} \Rightarrow SE * 1.96 = 0.0133 \approx 1.4\%$

$p_2 = 0.5$:

$SE = \sqrt{((0.5 \cdot 0.5)/1027)} = 0.0156 \Rightarrow SE * 1.96 = 0.0306 \approx 3.1\%$

Aufgabe 4: FDR & FWER

a) Definition **false discovery rate** (FDR) und **family-wise error rate** (FWER).

b) Schranken & Signifikanz

SNP-ID	p-Wert	Bonferroni	Bonferroni-Holm	Benjamini-Hochberg
rs1001	0.023			
rs1002	0.006			
rs1003	0.025			
rs1004	0.350			
rs1005	0.300			
rs1006	0.040			
rs1007	0.200			
rs1008	0.002			
rs1009	0.015			

Aufgabe 4: Lösung (1)

	H_A wahr	H_0 wahr	Total
Test sig.	S = true positives	V = false positives	R
Test nicht sig.	T = false negatives	U = true negatives	$m - R$
Total	$m - m_0$	m_0	m

m_0 ist die Anzahl wahrer H_0, die in der Regel unbekannt ist.

FWER: Family-wise error rate, $FWER = P(V \geq 0)$

Ziel: WSK falsch positive zu haben möglichst klein halten.

FDR: False discovery rate, $FDR = E(V/R)$

Ziel: Anteil an falsch positiven unter einem bestimmten Schwellenwert q zu halten.

Aufgabe 4: Lösung (2)

Bonferroni-Verfahren:

- Vergleich der p-Werte mit 5 %/Anzahl der Tests
- Schranke = Niveau/#Hypothesen (0.05/9 = 0.00556).
- Adjustierter p-Wert: $p_{adj} = p * m$.
- Nur p_8 ist kleiner, d.h. es gibt nur eine signifikante Assoziation.
- Kontrolle der FWER.

Aufgabe 4: Lösung (3)

Bonferroni-Holm-Verfahren:

- P-Werte zunächst der Größe nach sortieren, dann mit wachsender Schranke vergleichen.
- Schranke = Niveau/(#Hypothesen-Rank+1).
- Adjustierter p-Wert: $p_{adj} = p * (m - rank + 1)$
- Sowohl p_8 als auch p_2 sind kleiner als ihre jeweiligen Schranken, d.h. die SNPs rs1008 und rs1002 sind signifikant assoziiert.
- Kontrolle der FWER.

Aufgabe 4: Lösung (4)

Benjamini-Hochberg-Verfahren:

- P-Werte zunächst der Größe nach sortieren, dann mit wachsender Schranke vergleichen.
- Schranke = (Niveau/#Hypothesen)*Rank)).
- Adjustierter p-Wert: $p_{adj} = (p * m)/rank$
- Hier sind p_8, p_2, p_9, und p_3 kleiner als ihre jeweiligen Schranken. Alle SNPs mit einem kleineren Rank als der letzte gültige SNP sind laut dieser Methode signifikant assoziiert, d.h. auch rs1001 ist ist assoziiert, obwohl dessen p-Wert größer als seine Schranke ist.
- Kontrolle der FDR.

Aufgabe 4: Lösung (5)

```
pvec<-c(0.023,0.006,0.025,0.35,0.3,0.04,0.2,0.002,0.015)
pid<-c("rs1001","rs1002","rs1003","rs1004","rs1005",
       "rs1006","rs1007","rs1008","rs1009")
myDat<-data.table(pid,pvec)
setorder(myDat,pvec)

myDat$p.adj.1<-round(p.adjust(myDat$pvec,method = "bonferroni"
myDat$p.adj.2<-round(p.adjust(myDat$pvec,method = "holm"),4)
myDat$p.adj.3<-round(p.adjust(myDat$pvec,method = "BH"),4)
```

Aufgabe 4: Lösung (6)

Table 3: Ausgefüllte Tabelle zu Aufgabe 4. B=Bonferroni, BonH=Bonferroni-Holm, BenH=Benjamini-Hochberg

SNP	p-Wert	adj. p B	adj. p BonH	adj. p BenH
rs1008	0.002	0.018	0.018	0.0180
rs1002	0.006	0.054	0.048	0.0270
rs1009	0.015	0.135	0.105	0.0450
rs1001	0.023	0.207	0.138	0.0450
rs1003	0.025	0.225	0.138	0.0450
rs1006	0.040	0.360	0.160	0.0600
rs1007	0.200	1.000	0.600	0.2571
rs1005	0.300	1.000	0.600	0.3375
rs1004	0.350	1.000	0.600	0.3500

Aufgabe 4: Lösung (7)

Die adjustierten p-Werte werden noch angepasst:

- Werte >1 werden auf 1 gesetzt (z.B. *adj. p B* von rs1007)
- Das Ranking bleibt gleich, d.h.
  - wenn beim Bonferroni-Holm-Verfahren ein adjustierter p-Wert nach Korrektur kleiner ist als ein Rang-niedriger werden beide bzw. alle dazwischen liegenden SNPs auf den höheren Wert des Rang-niedrigeren SNPs gesetzt (z.B. *adj p BonH* von rs1007, rs1005, und rs1004)
  - Wenn beim Benjamini-Hochberg-Verfahren ein adjustierter p-Wert nach Korrektur kleiner ist als ein Rang-niedriger werden beide bzw. alle dazwischen liegenden SNPs auf den niedrigeren Wert des Rang-höheren SNPs gesetzt (z.B. *adj p BenH* von rs1001 und rs1003)