1.2 Epidemiologische Maßzahlen

Man kann zwei Arten unterscheiden:

Erkrankungshäufigkeiten, Krankheitsrisiken
→ binomiales Wahrscheinlichkeitsmodell

Prävalenz: Zustandsbeschreibung einer Krankheit
Kumulative Inzidenz: beschreibt Entstehung einer Krankheit

Erkrankungsraten: haben zeitlichen Bezug
→ Poisson Wahrscheinlichkeitsmodell

Inzidenz: Erkrankungsgeschwindigkeit

Population unter Risiko

Der "Nenner"

Gruppe von Individuen, die eine Krankheit entwickeln kann

Definition hängt von der Fragestellung ab; Beispiel: Brustkrebs, arbeitsmedizinische Untersuchungen

Prävalenz

Die **Prävalenz** P gibt die Wahrscheinlichkeit an, dass eine zufällig ausgewählte Person zu einem **Stichtag** an der betrachteten Krankheit erkrankt ist:

$$P = \frac{M}{N} = \frac{\text{Anzahl der Personen mit Krankheit}}{\text{Population unter Risiko}}$$

In **Querschnitts-** oder **Prävalenzstudien** wird die Prävalenz in Abhängigkeit von der derzeitigen oder auch früheren Expositionsbelastung (Expositionsprävalenz) untersucht.

Auch Prävalenz von bestimmten Charakteristiken häufig von Interesse, z.B. Risikofaktoren

Beispiel für eine Expositionsprävalenz

Rauchprävalenz nach Geschlecht im Jahr 1997

Land	Prozentsatz der Raucher	
	Männer	Frauen
Korea	68.2	6.7
Fiji	59.3	30.6
Griechenland	46.0	28.0
Ägypten	39.8	1.0
Dänemark	37.0	37.0
Schweden	22.0	24.0

Kumulative Inzidenz

Ein alternatives Maß für Krankheitsrisiko: Wahrscheinlichkeit dass eine zufällig ausgewählte Person der Population innerhalb eines zeitlich begrenzten Zeitraums (z.B. ein Jahr) an einer Krankheit neu erkrankt:

$$CI = \frac{I}{N_0} = \frac{\text{Anzahl der Neuerkrankungen im Zeitraum}}{\text{Gesunde Population zu Beginn des Zeitraums}}$$

Nicht zu verwechseln mit Inzidenzraten

Der Nenner bezieht sich nur auf den Beginn des Zeitraums

Inzidenz = Inzidenzrate

Die Inzidenzrate ist definiert als

$$I = \frac{\text{Anzahl der Neuerkrankungen im Zeitraum}}{\text{Summe der } \textbf{Personenzeiten}\text{: } \textbf{(Gesamt-)Risikozeit}}$$

Die **Personenzeit** eines Individuums ist dabei die Zeit, die ein Individuum gesund ist und sich in der Population aufhält und somit überhaupt die Möglichkeit besitzt, eine Krankheit zu entwickeln.

Typische Einheit: **Personenjahre** (engl.: "person-years")

Approximation von Inzidenzraten

Die Risikozeit, die Summe der Personenzeiten, wird manchmal approximiert durch die mittlere Populationsgröße unter Risiko, multipliziert mit der Länge des betrachteteten Zeitraums.

Somit ergibt sich auch eine Approximation der Inzidenz.

Beispiel: Krebsinzidenz bei flächendeckenden Inzidenzregistern

Beispiel für Inzidenzraten

In einer Studie in den USA wurde die Inzidenzrate des Schlaganfalls bei 118,539 Frauen im Alter von 30 bis 55 Jahren untersucht.

Rauch-kategorie	Anzahl der Fälle	Risikozeit (in Personenjahren)	Inzidenzrate (pro 100,000 Personenjahre)	90% KI
Nichtraucher	70	395,594	17.7	(14.4, 21.4)
Ex-Raucher	65	232,712	27.9	(22.6, 34.0)
Raucher	139	280,141	49.6	(43.0, 56.9)
Gesamt	274	908,447	30.2	(27.3, 33.3)

Beziehung zwischen Prävalenz und Inzidenzrate

Unter der Annahme dass die Prävalenz gering ist ("rare disease" Annahme) und sich nicht über die Zeit hinweg ändert, läßt sich diese wie folgt approximieren:

Prävalenz $\approx$ Inzidenzrate $\cdot$ durchschnittliche Dauer der Krankheit

Beispiel: Die Inzidenzrate sei 9.1 Fälle pro 100 Personenjahre und die durchschnittliche Dauer der Krankheit sei 3.3 Jahre. Dann ergibt sich die Prävalenz approximativ zu $9.1 \cdot 3.3 \approx 30$ Fälle pro 100 Personen.

Man beachte dass sich die Zeiteinheit "Jahre" in dieser Rechnung wegkürzt.

1.3 Likelihood-Inferenz für Inzidenzraten

$\rightsquigarrow$ Poisson Likelihood

Daraus:

- ML-Schätzer
- Standardfehler
- Konfidenzintervalle (KI)

  Likelihood-Intervalle
  Wald-KI basierend auf quadratischer Approximation

Likelihood für Inzidenzraten

Sei Y die Anzahl der im Zeitraum Erkrankten und T die Gesamt-Risikozeit. Man nimmt nun an dass Y Poisson-verteilt ist mit Erwartungswert λT (λ ist also die unbekannte Inzidenzrate), d.h. Y hat Wahrscheinlichkeitsfunktion

$$f(y;\lambda) = \frac{(\lambda T)^y}{y!}\exp(-\lambda T), \quad y = 0, 1, \ldots$$

Die **Likelihoodfunktion** von λ ist also

$$L(\lambda) = \lambda^y \exp(-\lambda T),$$

da alle übrigen multiplikativen Faktoren nicht von λ abhängen.

Die Log-Likelihood

Die **Log-Likelihood** ergibt sich somit zu

$$l(\lambda) = y\log(\lambda) - \lambda T,$$

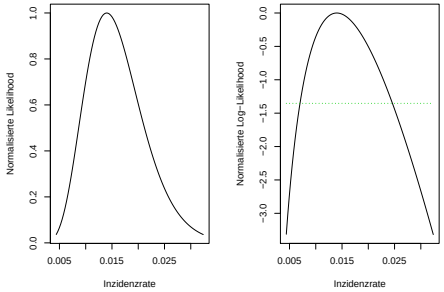
mit ML-Schätzer $\widehat{\lambda} = y/T$ bzw. $\widehat{\lambda} = Y/T$.

Die **normalisierte Log-Likelihood** $\tilde{l}(\lambda) = l(\lambda) - l(\widehat{\lambda})$ ist somit (unter der Annahme $y > 0$)

$$\tilde{l}(\lambda) = y\log\left(\frac{\lambda T}{y}\right) - \lambda T + y.$$

Beispiel für eine normalisierte (Log)-Likelihood

Angenommen wir beobachten $y = 7$ Fälle und die Gesamt-Risikozeit ist $T = 500$ Personenjahre. $\rightarrow \widehat{\lambda} = y/T = 1.4$ pro 100 Personenjahre.



Likelihood-Intervall

Ein **Likelihood-Intervall** für λ ist die Menge aller Werte von λ, für die die normierte Log-Likelihood größer als ein bestimmter Wert c ist: $\{\theta : \widetilde{l}(\theta) > c\}$

Üblicherweise wird $c = -\frac{1}{2}\chi^2_{1,\,(1-\alpha)}$ gewählt, was ein approximatives KI zum Niveau $1 - \alpha$ liefert

α	c
$0,1$	$-1,35$
$0,05$	$-1,92$
$0,01$	$-3,32$

Die Berechnung von Likelihood-Intervallen benötigt i.A. numerische Verfahren

Der Standardfehler von $\widehat{\lambda}$ und $\log \widehat{\lambda}$

Der Standardfehler des ML-Schätzers $\widehat{\lambda} = Y/T$ ist

$$se(\widehat{\lambda}) = \sqrt{Y}/T$$

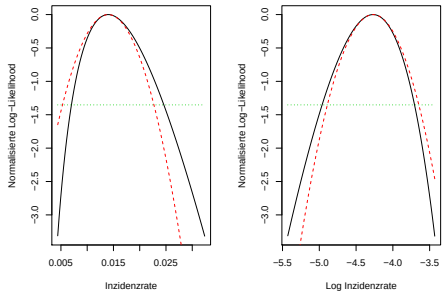
Betrachte nun die logarithmierte Inzidenzrate $\gamma = \log \lambda$. Der Standardfehler von $\widehat{\gamma} = \log(Y/T)$ ist

$$se(\widehat{\gamma}) = 1/\sqrt{Y}$$

(Herleitung über Δ-Regel)

Quadratische Approximation der Log-Likelihood

Für Inzidenzrate λ (links) und logarithmierte Inzidenzrate γ(rechts)



Wald-Konfidenzintervalle für $\widehat{\lambda}$

Sei $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung

Somit ist

$$\begin{aligned} \widehat{\lambda} &\pm z_{1-\alpha/2} \cdot se(\widehat{\lambda}) \quad \text{d.h.} \\ Y/T &\pm z_{1-\alpha/2} \cdot \sqrt{Y}/T \end{aligned}$$

ein approximatives $(1-\alpha)$-KI für λ.

Wald-Konfidenzintervalle für $\log \widehat{\lambda}$ und $\widehat{\lambda}$

Alternativ kann man

$$\widehat{\gamma} \pm z_{1-\alpha/2} \cdot se(\widehat{\gamma})$$

verwenden, was durch Exponieren das approximative $(1-\alpha)$-KI

$$[\widehat{\lambda}/\exp(z_{1-\alpha/2}/\sqrt{Y}), \widehat{\lambda} \cdot \exp(z_{1-\alpha/2}/\sqrt{Y})],$$

für λ liefert.

Man nennt den Term $\exp(z_{1-\alpha/2}/\sqrt{Y})$ den **Fehlerfaktor** (engl.: "error factor")

Beispiel: Konfidenzintervalle für Raten

Angenommen wir beobachten $y = 7$ Fälle und die Gesamt-Risikozeit ist $T = 500$ Personenjahre. ML-Schätzer: $\widehat{\lambda} = 1.4$ pro 100 Personenjahre

Typ	90%-KI
	(pro 100 Personenjahre)
Likelihood	0.700, 2.459
Wald-Typ für λ	0.530, 2.270
Wald-Typ für $\log \lambda$	0.752, 2.607

Kohortenstudien

Eine Gruppe von Individuen wird über einen gewissen Zeitraum verfolgt: **prospektive Kohortenstudie**

Typischerweise gibt es verschiedene Expositionsgruppen.

Von Interesse ist die Inzidenzrate in den verschiedenen Gruppen.

Bei einer **retrospektiven Kohortenstudie** wird die Kohorte mit Hilfe historischer Daten rekonstruiert.

Beispiel

- Eine Kohortenstudie wurde durchgeführt, um die Beziehung zwischen der Energieaufnahme und dem Auftreten von ischämischen Herzkrankheiten (IHD) zu untersuchen.
- Die Kohorte bestand aus 337 Männern mit mittlerem Follow-up von 13.7 Jahren.
- 45 Fälle von IHD wurden beobachtet.
- Niedrige Energieaufnahme ist ein Surrogat für Bewegungsarmut → Expositionsgruppe

Daten

	Energieaufnahme	
	< 2750 kcal (exponiert)	≥ 2750 kcal (nicht exponiert)
Risikozeit (in Personenjahren)	1857.5 (T_1)	2768.9 (T_0)
Neue Fälle	28 (y_1)	17 (y_0)
Geschätzte Rate (pro 1000 Personenjahre)	15.1	6.1
90% KI (Wald-Typ für $\log \hat{\lambda}$)	(11.1, 20.6)	(4.1, 9.1)

Wie kann man quantitativ untersuchen, ob der Unterschied zwischen den beiden Gruppen "signifikant" ist?

Die Inzidenzdifferenz

Zielgröße: Differenz der Inzidenzraten $\lambda_1 - \lambda_0$ in den beiden Gruppen (engl.: "rate difference")

Auch **Exzess** oder **Absolutes Risiko** genannt (besser: **Exzess Rate**)

Idee:

– Fälle, die man der "natürlichen" Inzidenz zuschreiben kann, sind in beiden Gruppen in der selben Intensität vorhanden.
– Fälle, die der Exposition zugeschrieben werden können, sind in der **Exzess Rate** repräsentiert.

Beispiel

In obigem Beispiel ergibt sich die Inzidenzdifferenz zu $28/1857.5 - 17/2768.9 = 8.93$ Fälle pro 1000 Personenjahren.

Der Standardfehler der Inzidenzdifferenz ist

$$se(\hat{\lambda}_1 - \hat{\lambda}_0) = \sqrt{se(\lambda_1)^2 + se(\lambda_0)^2} = \sqrt{Y_1/(T_1)^2 + Y_0/(T_0)^2}$$

Im Beispiel ergibt sich 3.21 Fälle pro 1000 Personenjahre.

Beispiel

Brustkrebsinzidenz in Island in zwei Geburtskohorten

Geburtsjahr	Alter (in Jahren)				
	40-49	50-59	60-69	70-79	80-89
1880 - 1909	65.9	95.1	129.5	140.1	227.9
1840 - 1879	38.7	53.8	71.7	81.1	136.9
Inzidenzdifferenz	27.2	41.3	57.8	59.0	91.0
Inzidenzquotient	1.70	1.78	1.81	1.73	1.66

Man beachte, dass die Inzidenzquotienten relativ stabil sind, während die Inzidenzdifferenz sich mehr als verdreifacht.

Der Inzidenzquotient

Der Inzidenzquotient (engl.: "rate ratio") ist häufig ein besserer Indikator der Stärke des Zusammenhangs, da er **relativ** zu einer Baseline Inzidenz ausgedrückt wird, also dimensionslos ist.

Der Inzidenzquotient ist häufig konstant in **Strata** (Schichten)

Likelihood für den Inzidenzquotienten

Reparametrisiere die Raten λ_0 und λ_1 in den beiden Gruppen zu λ_0 und $\lambda_1 = \theta\lambda_0$.

Der Inzidenzquotient $\theta = \lambda_1/\lambda_0$ ist der interessierende Parameter.

Unter Unabhängigkeit ergibt sich die Log-Likelihood

$$\begin{aligned} l(\lambda_0, \theta) &= y_0 \log \lambda_0 - \lambda_0 T_0 + y_1 \log \lambda_1 - \lambda_1 T_1 \\ &= y_0 \log \lambda_0 - \lambda_0 T_0 + y_1 \log(\lambda_0 \theta) - \lambda_0 \theta T_1 \\ &= (y_0 + y_1) \log \lambda_0 + y_1 \log \theta - \lambda_0 (T_0 + \theta y_1) \end{aligned}$$

Profile Likelihood von θ

Maximierung von $l(\lambda_0, \theta)$ bzgl. λ_0 (bei festem θ) ergibt

$$\widehat{\lambda}_0(\theta) = \frac{y_0 + y_1}{T_0 + \theta T_1}.$$

Einsetzen ("plug-in") von $\widehat{\lambda}_0$ in $l(\lambda_0, \theta)$, führt (unter Ignorierung von additiven Konstanten) zu der **Profile Log-Likelihood**

$$l_P(\theta) = l(\widehat{\lambda}_0(\theta), \theta) = y_1 \log \theta - (y_0 + y_1) \log(T_0 + \theta T_1)$$

Profile Likelihood von θ

Maximierung von $l_P(\theta)$ führt zu dem ML-Schätzer

$$\hat{\theta} = \frac{y_1/y_0}{T_1/T_0} = \frac{y_1/T_1}{y_0/T_0} \text{ bzw. } \hat{\theta} = \frac{Y_1/T_1}{Y_0/T_0}$$

Der Standardfehler von $\log(\hat{\theta})$ ergibt sich zu

$$se(\log(\hat{\theta})) = \sqrt{1/Y_1 + 1/Y_0}.$$

IHD Beispiel

Hier ergibt sich ein Inzidenzquotient von 2.48 mit Standardfehler von $\log \hat{\theta}$ gleich $\sqrt{1/28 + 1/17} = 0.3075$.

Der 90% Fehlerfaktor is somit $\exp(1.645 \cdot 0.3075) = 1.66$

Somit ergibt sich das 90% Konfidenzintervall $[1.49, 4.12]$ für den Inzidenzquotienten.

Ein alternativer bedingter Likelihoodansatz

Idee: Die Gesamtanzahl von Ereignissen $Y = Y_0 + Y_1$ ist irrelevant für θ, nur die Anteile Y_0/Y and Y_1/Y sind wichtig. Wir können daher so tun, als ob Y fest wäre.

Ein Ereignis in der Expositionsgruppe ist somit binomialverteilt mit bedingte Wahrscheinlichkeit $\theta T_1/(T_0 + \theta T_1)$

Eine binomiale Likelihood mit obiger Wahrscheinlichkeit ist identisch zur Profile Likelihood für θ

Confounding

Epidemiologische Studien vergleichen die Krankheitsinzidenz in Gruppen mit unterschiedlicher Exposition.

Wegen der fehlenden Randomisierung gibt es immer die Möglichkeit dass eine wichtige weitere Einflußvariable systematisch zwischen den Expositionsgruppen variiert.

Daher ist es möglich, dass Teile eines scheinbaren Effektes durch solche Unterschiede erklärt werden können, man spricht dann von **Confounding**.

IHD Daten

Daten nach Alter geschichtet:

	Exponiert		Nicht exponiert		geschätzter
Alter	Y_1	T_1 (Anteil)	Y_0	T_0 (Anteil)	Inzidenzquotient
40-49	2	311.9 (0.17)	4	607.9 (0.21)	0.97
50-59	12	878.1 (0.47)	5	1272.1 (0.46)	3.48
60-69	14	667.5 (0.36)	8	888.9 (0.32)	2.33
Total	28	1857.5 (1.0)	17	2768.9 (1.0)	2.45

$\rightarrow$ Exponierte Gruppe ist eher älter. Kann dies (zumindest zu einem Teil) die Unterschiede in den Inzidenzraten erklären?

Stratifizierter Vergleich von Inzidenzraten

Die Daten liegen nun in Strata, indiziert durch i, vor, wobei die Anzahl der Erkrankungen und die zugehörige Risikozeit gleich (Y_{1i}, T_{1i}) (exponiert) bzw. (Y_{0i}, T_{0i}) (nicht exponiert) sind.

Wir nehmen im folgenden an, dass die Inzidenzraten λ_{1i} und λ_{0i} in den einzelnen Strata i unterschiedlich sind, der Inzidenzquotient $\theta = \lambda_{1i}/\lambda_{0i}$ aber nicht von i abhängt.

ML-Schätzung von θ kann über log-lineare Poisson-Regression berechnet werden.

Beispiel

Folgendes GLM-Programm in R berechnet den ML-Schätzer $\hat{\theta}$ in der IHD Studie:

```
Y <- c(2,12,14,4,5,8)
T <- c(311.9, 878.1, 667.5,607.9, 1272.1, 888.9)
group <- c(1,1,1,0,0,0)
age <- as.factor(c(1,2,3,1,2,3))
model1 <- glm(Y ~ group + age + offset(log(T)), family = poisson)
summary.glm(model1)
```

Die Ergebnisse sind:

```
Coefficients:   Estimate   Std. Error   z value      Pr(>|z|)
(Intercept)     -5.4177       0.4421   -12.256   < 2e-16 ***
group            0.8697       0.3080     2.823    0.00476 **
age2             0.1290       0.4754     0.271       0.78609
age3             0.6920       0.4614     1.500       0.13366
```

$\rightsquigarrow \hat{\theta} = \exp(0.8697) \approx 2.39$

Der Mantel-Haenszel Schätzer

Gute Approximation zum ML-Schätzer (siehe Vorlesung):

$$\hat{\theta}_{MH} = \frac{\sum_i Y_{1i}T_{0i}/T_i}{\sum_i Y_{0i}T_{1i}/T_i} = \frac{\sum_i Q_i}{\sum_i R_i} = \frac{Q}{R} \text{ mit } T_i = T_{0i} + T_{1i}$$

Für den Standardfehler von $\log \hat{\theta}_{MH}$ gilt:

$$se(\log \hat{\theta}_{MH}) = \sqrt{V/(Q \cdot R)}$$

mit

$$V = \sum_i V_i = \sum_i (Y_{0i} + Y_{1i}) \frac{T_{0i}T_{1i}}{T_i^2}$$

Beispiel

Für die IHD Daten ergibt sich der MH-Schätzer $\hat{\theta} = 2.40$ (der ML-Schätzer war $\exp(0.8697) = 2.39$) mit Standardfehler für $\log \hat{\theta}_{MH}$ von 0.311 (ML: 0.308). Die Ergebnisse liegen also sehr nahe beim ML-Schätzer.