Wahrscheinlichkeitsrechnung und Statistik
im Biologie-Bachelorstudiengang der LMU

Wiederholung zum Prinzip des Statistischen Testens und dem Zusammenhang mit Konfidenzintervallen

Dirk Metzler

26. Juli 2020

Inhaltsverzeichnis

1 Prinzip des statistischen Testens **1**
1.1 Beispiel: Orientierung bei Trauerschnäppern 1
1.2 Allgemeines Prinzip 4

2 Multiples Testen **5**

3 Konfidenzintervalle **7**
3.1 Dualität von Tests und Konfidenzintervallen 7
3.2 Konfidenzintervalle für Anteile (Parameter p der Binomialverteilung) 9

4 Grundannahmen der frequentistischen Statistik **10**
4.1 Maximum-Likelihood-Schätzer 11
4.2 Vergleich zur Bayesschen Statistik 12

1 Prinzip des statistischen Testens

1.1 Beispiel: Orientierung bei Trauerschnäppern

Wir meinen: Die Farbe der Beleuchtung hat einen Einfluß auf die Orientierung

Ein Skeptiker würde erwidern: Alles nur Zufall

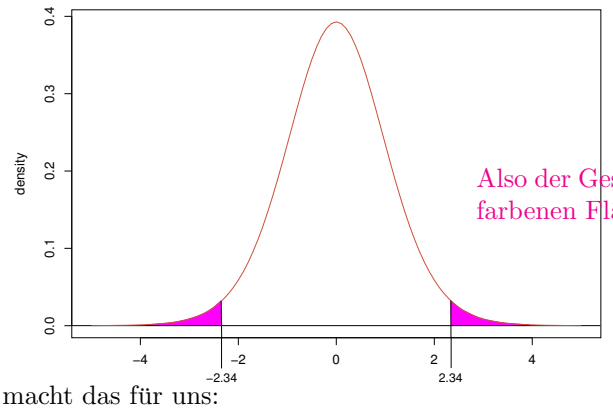
Wir wollen nun zeigen: Unter der Annahme 'Kein Einfluß' ist die Beobachtung sehr unwahrscheinlich

Nullhypothese: $\mu = 0$

Wie (un)wahrscheinlich ist nun eine mindestens so große Abweichung wie 2.34 Standardfehler?

$\Pr(|T| = 2.34) = 0$ Das bringt nichts!

Zu berechnen ist $\Pr(|T| \geq 2.34)$, der sog. p-Wert.



Also der Gesamtinhalt der magentafarbenen Flächen.

R macht das für uns:

```
> pt(-2.34,df=16)+pt(2.34,df=16,lower.tail=FALSE)
[1] 0.03257345
```

Beachte: `pt(2.34,df=16,lower.tail=FALSE)` ist dasselbe wie `1-pt(2.34,df=16)`, also der `upper tail`.

Zum Vergleich mal mit der Normalverteilung:

```
> pnorm(-2.34)+pnorm(2.34,lower.tail=FALSE)
[1] 0.01928374
```

Vollständiger t-Test mit R

```
> x <- trauerschn$gruen-trauerschn$blau
> t.test(x)

One Sample t-test

data:  x
t = 2.3405, df = 16, p-value = 0.03254
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.004879627 0.098649784
sample estimates:
 mean of x
0.05176471
```

Wir halten fest:

$$p - \text{Wert} = 0.03254$$

Wenn die **Nullhypothese** "alles nur Zufall" (hier $\mu = 0$) gilt, dann ist eine mindestens so große Abweichung sehr unwahrscheinlich.

Sprechweise:

Wir verwerfen die Nullhypothese auf dem 5%-Signifikanzniveau.

Oder:

Die Differenz zwischen grün und blau ist auf dem 5%-Niveau signifikant.

Die Nullhypothese wurde also auf dem 5%-Niveau verworfen. Welche Aussagen sind wahr/sinnvoll?

- Die Nullhypothese ist falsch. ~~Die Nullhypothese ist falsch.~~
- Die Nullhypothese ist mit 95%-iger Ws falsch. ~~Die Nullhypothese ist mit 95%-iger Ws falsch.~~
- Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in 5% der Fälle. Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in 5% der Fälle. ✓
- Die Orientierung der Vögel ist bei blau und grün verschieden. ~~Die Orientierung der Vögel ist bei blau und grün verschieden.~~
- Die Orientierung bei grün und blau ist auf dem 5%-Niveau signifikant verschieden. Die Orientierung bei grün und blau ist auf dem 5%-Niveau signifikant verschieden. ✓

Angenommen, H_0 konnte durch den Test nicht verworfen werden. Welche Aussagen sind dann richtig?

- ~~Wir müssen die Alternative H_1 verwerfen.~~
- ~~H_0 ist wahr.~~
- ~~H_0 ist wahrscheinlich wahr.~~
- ~~Es ist ungefährlich, davon auszugehen, dass H_0 zutrifft.~~
- Auch wenn H_0 zutrifft, ist es nicht sehr unwahrscheinlich, dass unsere Teststatistik einen so extrem erscheinenden Wert annimmt.✓
- Die Nullhypothese ist in dieser Hinsicht mit den Daten verträglich.✓

Man könnte auch ein anderes Signifikanzniveau α wählen. Dann müsste man zeigen, dass der p-Wert kleiner als α ist.

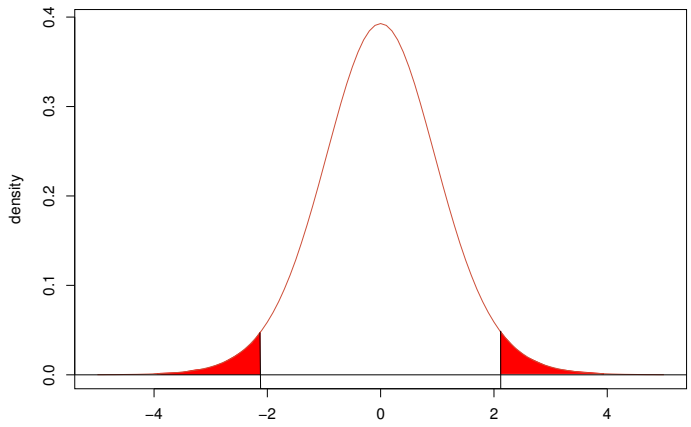
Wichtig: Wähle zuerst das Signifikanzniveau und ermittle erst dann den p-Wert! Das Signifikanzniveau je nach p-Wert zu wählen ist geschummelt.

In der Literatur wird üblicherweise 5% als Signifikanzniveau gewählt.

Beachte:

Falls die Nullhypothese zutrifft und wir auf dem 5%-Niveau testen, ist 5% die Wahrscheinlichkeit, dass wir die Nullhypothese zu Unrecht verwerfen.

Wir verwerfen also die Nullhypothese auf 5%-Niveau, wenn der Wert der t-Statistik in den roten Bereich fällt:



(hier am Beispiel der $t-$Verteilung mit df= 16 Freiheitsgraden)

Zusammenfassung gepaarter t-Test
Gegeben: gepaarte Beobachtungen

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Nullhypothese H_0: $\mu_Y = \mu_Z$
Signifikanzniveau: α (meist $\alpha = 5\%$)
Test: gepaarter t-Test (genauer: zweiseitiger gepaarter t-Test)
Berechne Differenz $X := Y - Z$
Berechne Teststatistik

$$t := \frac{\overline{X}}{s(X)/\sqrt{n}}$$

p-Wert $= \Pr(|T_{n-1}| \geq |t|)$ ($n-1$ Freiheitsgrade)
Verwirf Nullhypothese, falls p-Wert $\leq \alpha$

1.2 Allgemeines Prinzip

Statistische Tests: Die wichtigsten Begriffe

Nullhypothese H_0 : wollen wir meistens verwerfen, denn sie besagt, dass die interessanten Auffälligkeiten in den Daten nur Zufallsschwankungen sind.

Signifikanzniveau α : Wahrscheinlichkeit, dass wir, falls die Nullhypothese gilt, diese zu Unrecht verwerfen.

Teststatistik : Misst, auffällig unsere Daten von der Nullhypothese abweichen.

p**-Wert** : Für einen beobachteten Wert t der Teststatistik ist der p-Wert die Wahrscheinlichkeit, dass, falls die Nullhypothese gilt, die Teststatistik (etwa bei einer hypothetischen Wiederholung des Versuchs) einen mindestens so extremen Wert wie t annimmt. Dabei hängt es von der Art des Tests ab (z.B. einseitig/zweiseitig), was "extrem" heißt.

- Wir verwerfen H_0, falls der p-Wert kleiner als α wird. (Üblich ist $\alpha = 0.05$).
- Damit ergibt sich, dass wir nur in einem Anteil α der Fälle, in denen H_0 gilt, diese (fälschlicherweise) verwerfen.
- Auch wer immer nur Daten ananlysiert, in denen außer Zufallsschwankungen nichts steckt, wird in einem Anteil α der Tests die Nullhypothese verwerfen.
- Ein schwerer Verstoß gegen die Wissenschaftlichkeit ist daher, so lange statistische Tests durchzuführen, bis mal H_0 auf einem Signifikanzniveau von 5% verworfen werden kann, und dann nur letzteres zu veröffentlichen.

Reine Lehre des statistischen Testens

- Formuliere eine Nullhypothese H_0, z.B. $\mu = 0$.
- Lege ein Signifikanzniveau α fest; üblich ist $\alpha = 0.05$.
- Lege ein Ereignis $\mathcal{A}$ ("extreme Abweichung") fest, so dass

  $$\Pr_{H_0}(\mathcal{A}) = \alpha$$

  (oder zumindest $\Pr_{H_0}(\mathcal{A}) \leq \alpha$). z.B. $\mathcal{A} = \{\overline{X} > q\}$ oder $\mathcal{A} = \{|\overline{X} - \mu| > r\}$
- ERST DANN: Betrachte die Daten und überprüfe, ob $\mathcal{A}$ eintritt.
- Dann ist die Wahrscheinlichkeit, dass H_0 verworfen wird, wenn H_0 eigentlich richtig ist ("Fehler erster Art") , lediglich α.

Fake Science: Verstöße gegen das Prinzip des Testens

"Beim zweiseitigen Testen kam ein p-Wert von 0.06 raus. Also hab ich einseitig getestet, da hat's dann funktioniert."

genauso problematisch:

"Beim ersten Blick auf die Daten habe ich sofort gesehen, dass $\overline{x}$ größer ist als μ_{H_0}. Also habe ich gleich einseitig getestet"

Wichtig
Die Entscheidung, ob einseitig oder zweiseitig getestet wird, darf nicht von den konkreten Daten abhängen, die zum Test verwendet werden. Allgemeiner: Ist $\mathcal{A}$ das Ereignis, dass zum Verwerfen von H_0 führt (falls es eintritt), so muss die Festlegung von $\mathcal{A}$ (und H_0) stattfinden bevor man die Daten betrachtet hat.

Mehr oder weniger bewusstes Fälschen von Forschungsergebnissen:

p-Hacking: Anpassen der Testmethoden bis p-Wert $< 5\%$ (z.B. welcher Test, welche Daten werden als Ausreißer entfernt, welche Co-Variablen in linearem Modell)

HARKing: Hypothesis After Result Known; suche nach Auffälligkeiten in den Daten und zeige Signifikanz ohne multiple-Testing-Korrektur

Mögliche Maßnahmen zur Vermeidung:

Trennung von explorativer Datenanalyse und Testen: Explorative Datenanalyse führt zu Hypothesen, die mit anderen Daten getestet werden.

Veröffentlichung der Hypothesen und Testmethoden vor Erhebung der Daten, mit denen die Tests durchgeführt werden

Registrierter Bericht: Fachzeitschriften entscheiden anhand Versuchsplanung über Veröffentlichung bevor Daten bekannt sind

Wenn Sie zu diesem Thema mal ein Video in einem ganz anderen Stil sehen wollen, empfehle ich Ihnen das maiLab-Video "Wissenschaftler irren":

`https://www.youtube.com/watch?v=DHyRaUeHcGY`

(p-Hacking und HARKing kommen in dem Video nach etwa 11 Minuten)

2 Multiples Testen

Die Varianzanalyse zeigte, dass es signifikante Unterschiede zwischen den Laboren gibt.
Aber welche Labore unterscheiden sich signifikant?
p-Werte aus paarweisen Vergleichen mittels t-Tests:

```
        Lab2    Lab3    Lab4    Lab5    Lab6    Lab7
Lab1 0.05357 0.00025 0.00000 0.00017 0.00055 0.04657
Lab2         0.84173 0.02654 0.25251 0.25224 0.97985
Lab3                 0.00001 0.03633 0.05532 0.86076
Lab4                         0.09808 0.16280 0.01944
Lab5                                 0.94358 0.22336
Lab6                                         0.22543
```

Wir haben 21 paarweise Vergleiche; auf dem 5%-Niveau zeigen einige davon Signifikanz an.

Problem des Multiplen Testens: Wenn die Nullhypothese ("alles nur Zufallsschwankungen") stimmt, verwirft man im Schnitt bei 5% der Tests die Nullhypothese zu Unrecht. Testet man mehr als 20 mal und gelten jeweils die Nullhypothesen, wird man im Schnitt mehr als eine Nullhypothese zu Unrecht verwerfen. Daher sollte man bei multiplen Tests mit korrigierten p-Werten arbeiten.

Eine Möglichkeit bei Varianzanalysen: Tukey's Honest Significant Differences (HSD).

```
> TukeyHSD(chlor.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Gehalt ~ Labor, data = chlor)

$Labor
       diff          lwr          upr     p adj
2-1 -0.065 -0.147546752  0.017546752 0.2165897
3-1 -0.059 -0.141546752  0.023546752 0.3226101
4-1 -0.142 -0.224546752 -0.059453248 0.0000396
5-1 -0.105 -0.187546752 -0.022453248 0.0045796
6-1 -0.107 -0.189546752 -0.024453248 0.0036211
7-1 -0.064 -0.146546752  0.018546752 0.2323813
3-2  0.006 -0.076546752  0.088546752 0.9999894
4-2 -0.077 -0.159546752  0.005546752 0.0830664
5-2 -0.040 -0.122546752  0.042546752 0.7578129
6-2 -0.042 -0.124546752  0.040546752 0.7140108
7-2  0.001 -0.081546752  0.083546752 1.0000000
4-3 -0.083 -0.165546752 -0.000453248 0.0478900
5-3 -0.046 -0.128546752  0.036546752 0.6204148
6-3 -0.048 -0.130546752  0.034546752 0.5720976
7-3 -0.005 -0.087546752  0.077546752 0.9999964
5-4  0.037 -0.045546752  0.119546752 0.8178759
6-4  0.035 -0.047546752  0.117546752 0.8533629
7-4  0.078 -0.004546752  0.160546752 0.0760155
6-5 -0.002 -0.084546752  0.080546752 1.0000000
7-5  0.041 -0.041546752  0.123546752 0.7362355
7-6  0.043 -0.039546752  0.125546752 0.6912252
```

Wir erhalten Konfidenzintervalle [lwr,upr] für die Unterschiede zwischen den Labormittelwerte und p-Werte für die Nullhypothese, dass diese Unterschiede 0 sind; alles bereits korrigiert für multiples Testen.

Einschränkung: Tukeys HSD-Methode ist streng genommen nur für *balancierten Versuchsplänen* (engl. *balanced design*) anwendbar, d.h. wenn in jeder Gruppe die selbe Anzahl von Messungen vorliegt. (Außerdem geht HSD wie die ANOVA selbst von gleichen Varianzen in allen Gruppen aus.)

Das ist bei dem Laborvergleich der Fall, da jedes Labor 10 Messungen durchgeführt hat.Die Blutgerinnungsdaten sind jedoch nicht balanciert, da Behandlung 1 an vier Ratten und Behandlung 2 an 8 Ratten erprobt wurde.

Was können wir verwenden, wenn die Bedingungen für Tukeys HSD nicht erfüllt sind?

Eine ganz allgemeine Korrektur für multiples Testen ist die **Bonferroni-Methode**: Multipliziere jeden p-Wert mit der Anzahl n der durchgeführten Tests.

Beispiel: Paarweise Vergleiche (mittels t-Test) für die Blutgerinnungszeiten bei vier verschiedenen Behandlungen, zunächst ohne Korrektur für multiples Testen:

```
        B       C       D
A 0.00941 0.00078 1.00000
B         0.17383 0.00663
C                 0.00006
```

Nun mit Bonferroni-Korrektur (alle Werte mit 6 multiplizieren):

```
        B       C       D
A 0.05646 0.00468 6.00000
B         1.04298 0.03978
C                 0.00036
```

Nach Bonferroni-Korrektur führen folgende Paare von Behandlungen zu jeweils signifikant unterschiedlichen Ergebnissen: A/C, B/D sowie C/D. (Der Bonferroni-korrigierte p-Wert von 6.0 für den Vergleich der Behandlungen A und D ist natürlich nicht als echter p-Wert zu interpretieren.)

Die Bonferroni-Methode ist sehr *konservativ*, d.h. um auf der sicheren Seite zu sein, lässt man sich lieber die eine oder andere Signifikanz entgehen.

Eine Verbesserung der Bonferroni-Methode ist die **Bonferroni-Holm-Methode**: Ist k die Anzahl der Tests, so multipliziere den kleinsten p-Wert mit k, den zweitkleinsten mit $k-1$, den drittkleinsten mit $k-2$ usw.

In R gibt des den Befehl p.adjust, der p-Werte für multiples Testen korrigiert und dabei defaultmäßig Bonferroni-Holm verwendet:

```
> pv <- c(0.00941, 0.00078, 1.00000, 0.17383,
+         0.00663, 0.00006)
> p.adjust(pv)
[1] 0.02823 0.00390 1.00000 0.34766 0.02652 0.00036
> p.adjust(pv, method="bonferroni")
[1] 0.05646 0.00468 1.00000 1.00000 0.03978 0.00036
```

Für paarweise t-Tests gibt es ebenfalls eine R-Funktion, die per default die Bonferroni-Holm-Korrektur verwendet:

```
> pairwise.t.test(rat$bgz,rat$beh,pool.sd=FALSE)

Pairwise comparisons using t tests with non-pooled SD

data:  rat$bgz and rat$beh

  A       B       C
B 0.02823 -       -
C 0.00391 0.34766 -
D 1.00000 0.02654 0.00035

P value adjustment method: holm
```

3 Konfidenzintervalle

Konfidenzintervall für den wahren Mittelwert

Ziel: Bestimme das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α, also das $(1-\alpha)$-Konfidenzintervall.

Das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α ist ein aus den Daten $X = (X_1, \ldots, X_n)$ geschätztes (zufälliges) Intervall

$$\big[a(X),\ b(X)\big]$$

mit folgender Eigenschaft: Ist der wahre Mittelwert gleich μ und ist $(X_1, \ldots, X_n)$ eine Stichprobe aus der Grundgesamtheit (mit Mittelwert μ), so gilt

$$\Pr_{\mu}\left(\mu \in \big[a(X),\ b(X)\big]\right) \geq 1-\alpha.$$

Selbstverständlich wollen wir das Konfidenzintervall möglichst klein wählen.

3.1 Dualität von Tests und Konfidenzintervallen

Die wechselseitige Beziehung zwischen Test und Konfidenzintervall untersuchen wir am Beispiel des folgenden Datensatzes:

```
> X
 [1] 4.111007 5.023229 5.489230 4.456054 4.343212
 [5] 5.431928 3.944405 3.471677 4.337888 5.412292
> n <- length(X)
> m <- mean(X)
```

```
> sem <- sd(X)/sqrt(n)
> t <- -qt(0.025,n-1)
> konf <- c(m-t*sem,m+t*sem)
> konf
[1] 4.100824 5.103360
```

$[4.100824, 5.103360]$

```
> t.test(X,mu=4)

One Sample t-test

data:  X
t = 2.7172, df = 9, p-value = 0.02372
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!

Dualität Tests $\leftrightarrow$ Konfidenzintervalle

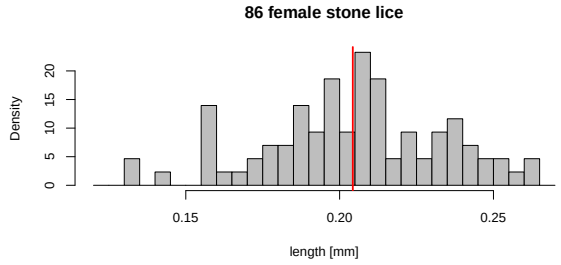
Ist $[a, b]$ ein $(1 - \alpha)$-Konfidenzintervall für einen Parameter θ, so erhält man einen Test mit Signifikanzniveau α, wenn man die Nullhypothese $\theta = x$ genau dann verwirft, wenn $x \notin [a, b]$.[0.5cm]

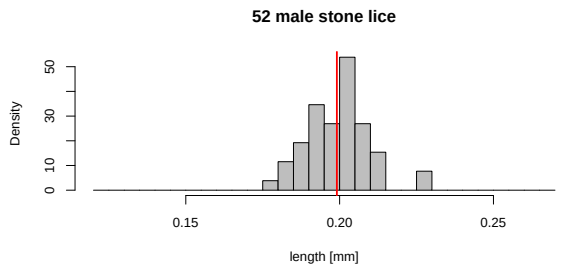
Ist umgekehrt T_x ein Test mit Nullhypothese $\theta = x$ und Signifikanzniveau α, so bilden alle Werte x, für die die Nullhypothese $\theta = x$ *nicht* verworfen wird, ein $(1 - \alpha)$-Konfidenzintervall für θ.

Konfidenzintervalle sind auch und gerade dann hilfreich, wenn ein Test *keine* Signifikanz anzeigt.

Beispiel: Gibt es bei Steinläusen geschlechtsspezifische Unterschiede in der Körperlänge?

Datenlage: die Längen von 86 weiblichen (F) und 52 männlichen (M) Steinläusen.






```
> t.test(F,M)
```

```
Welch Two Sample t-test

data:  F and M
t = 0.7173, df = 122.625, p-value = 0.4746
alternative hypothesis: true difference in means is
                                      not equal to 0
95 percent confidence interval:
 -0.004477856  0.009567353
sample estimates:
mean of x mean of y
0.2018155 0.1992707
```

Wie berichten wir über das Ergebis des Tests?

- Es gibt keinen Unterschied zwischen männlichen und weiblichen Steinläusen. ~~Es gibt keinen Unterschied zwischen männlichen und weiblichen Steinläusen.~~
- Männliche und weibliche Steinläuse sind im Mittel gleich lang. ~~Männliche und weibliche Steinläuse sind im Mittel gleich lang.~~
- Die Daten zeigen keine signifikanten Unterschiede zwischen den mittleren Längen männlicher und weiblicher Steinläuse. Die Daten zeigen keine signifikanten Unterschiede zwischen den mittleren Längen männlicher und weiblicher Steinläuse.✓
- Ein 95%-Konfidenzbereich für die Differenz zwischen der mittleren Länge der Weibchen und der Männchen ist [-0.0045,0.0096]. Ein 95%-Konfidenzbereich für die Differenz zwischen der mittleren Länge der Weibchen und der Männchen ist [-0.0045,0.0096].✓

3.2 Konfidenzintervalle für Anteile (Parameter p der Binomialverteilung)

Das Konfidenzintervall

$$\left[\widehat{p} - 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\ ,\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\right]$$

nennt man auch Wald-Konfidenzintervall.

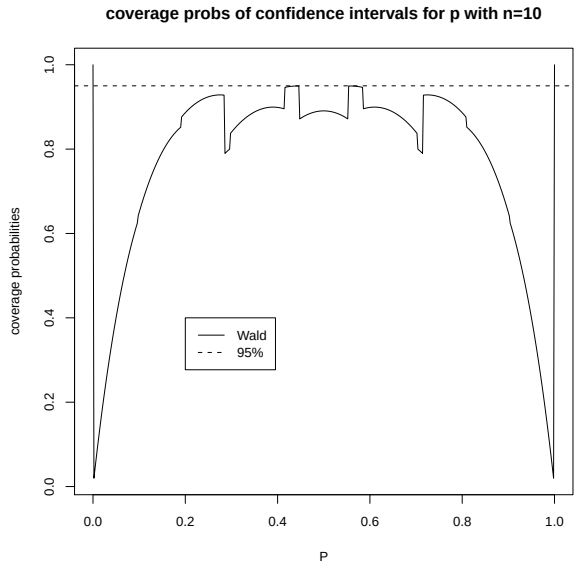
Es sollte gelten: Das Konfidenzintervall überdeckt (d.h. enthält) den wahren Parameterwert mit einer Wahrscheinlichkeit von mindestens 95%.

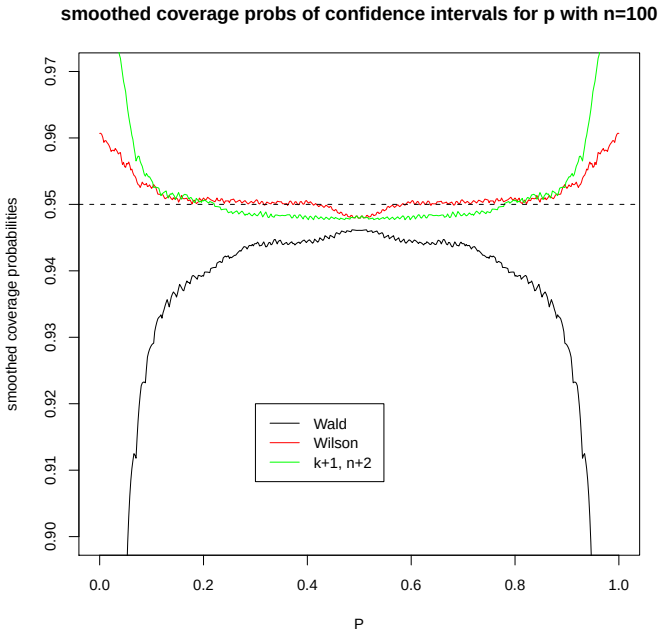
Diese *Überdeckungswahrscheinlichkeit* kann man berechnen, und das tun wir nun für $n = 10$ mit Werten für p zwischen 0 und 1.

Genauer: Wir zeichnen die Funktion

$$p \mapsto \Pr\left(p \in \left[\widehat{p} - 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\ ,\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\right]\right)$$

wobei $\widehat{p} = X/n$ und X binomialverteilt ist mit Versuchslänge n und Erfolgswahrscheinlichkeit p .





4 Grundannahmen der frequentistischen Statistik

- Parameter sind unbekannt aber nicht zufällig.
- Daten hängen von den Parametern und vom Zufall ab (gemäß Modellannahmen).
- frequentistischer Wahrscheinlichkeitsbegriff: Wenn ein Ereignis eine Wahrscheinlichkeit p hat, dann bedeutet das, dass es auf lange Sicht im Anteil p aller Fälle eintritt.
- Wenn ich meine Tests mit Signifikanzniveau α durchführe, verwerfe ich die Nullhypothese zu Unrecht nur in einem Anteil α der Fälle. (auf lange Sicht)

- Wenn ich 95%-Konfidenzintervalle angebe, enthalten 95% meiner Konfidenzintervalle den tatsächlichen Parameterwert. (auf lange Sicht)

4.1 Maximum-Likelihood-Schätzer

- Auch wenn es allgemein sinnvoll ist, Konfidenzintervalle für Parameterschätzungen anzugeben, möchte man manchmal einen einzelnen Schätzwert für einen Parameter angeben, und die frequentistische Statistik hat auch hierfür eine bevorzugte Methode, die *Maximum-Likelihood*-Schätzung (kurz ML).
- Es ist sinnlos, nach dem "wahrscheinlichsten" Parameterwert zu fragen, denn Parameter sind (aus Sicht der frequentistischen Statistik) nicht zufällig und haben daher auch keine Wahrscheinlichkeit.
- Statt dessen sucht man den Parameterwert, der die Daten am wahrscheinlichsten macht. Die *Likelihood* eines Werts x für einen Parameter θ ist die Wahrscheinlichkeit der beobachteten Daten D, falls $\theta = x$ gilt:
$$L_D(x) := \Pr_{\theta=x}(D)$$
- Die *Likelihood* eines Werts x für einen Parameter θ ist die Wahrscheinlichkeit der beobachteten Daten D, falls $\theta = x$ gilt:
$$L_D(x) := \Pr_{\theta=x}(D)$$
- Der *Maximum-Likelihood-Schätzer* (ML-Schätzer) ist der Parameterwert $\widehat{\theta}$, für den die Funktion L_D maximal wird:
$$\widehat{\theta} = \arg\max_x L_D(x)$$
also dasjenige x, für das $L_D(x)$ maximal wird

Beispiel: Auf einem mtDNA-Abschnitt der Länge 100 werden zwischen Mensch und Schimpanse 7 Unterschiede festgestellt. Wie hoch ist die Wahrscheinlichkeit p, auch an der 101. Position einen Unterschied zu sehen?

Naheliegender Schätzer 7/100

ML-Schätzer: Modelliere die Anzahl K der beobachteten Mutationen als binomialverteilt mit $n = 100$ und unbekanntem p. Dann gilt

$$L(p) = \Pr_p(K = 7) = \binom{100}{7} p^7 \cdot (1-p)^{93}$$

und

$$\begin{aligned} \widehat{p} &= \arg\max_p \binom{100}{7} p^7 \cdot (1-p)^{93} = \arg\max_p p^7 \cdot (1-p)^{93} \\ &= \arg\max_p \log\left(p^7 \cdot (1-p)^{93}\right) \end{aligned}$$

Gesucht ist also die Maximalstelle von

$$f(p) := \log\left(p^7 \cdot (1-p)^{93}\right) = 7 \cdot \log(p) + 93 \cdot \log(1-p).$$

Wir finden Sie wie üblich durch Nullsetzen der Ableitung:

$$0 = f'(p) = 7 \cdot \frac{1}{p} + 93 \frac{1}{1-p} \cdot (-1)$$

(dabei hilft es, zu wissen dass $\log'(x) = 1/x$.) Löst man die Gleichung nach p so erhält man:

$$\widehat{p} = 7/100$$

Wir haben also eine theoretische Begründung für den naheliegenden Schätzer 7/100 gefunden.

Der ML-Schätzer ist in vielen Fällen *konsistent*, d.h. wenn genügend viele Daten vorliegen und die Modellannahmen erfüllt sind, wird er den tatsächlichen Parameterwert finden.

Wenn eher wenig Daten vorhanden sind, ist manchmal ein anderer Schätzer zu bevorzugen.

Beispiel: ist $X_1, \ldots, X_n$ eine Stichprobe aus einer Normalverteilung, so ist $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ der ML-Schätzer für die Varianz σ^2. Meistens wird aber der Bias-korrigierte Schätzer $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ bevorzugt.

4.2 Vergleich zur Bayesschen Statistik

Grundannahmen der Bayesschen Statistik

- Parameter werden auch als zufällig betrachtet
- Die *a-priori-Wahrscheinlichkeitsverteilung* eines Parameters gibt an, für wie wahrscheinlich man die möglichen Parameterwerte hält, **bevor** man die Daten gesehen hat.
- Mit der Bayes-Formel erhält man die *a-posteriori-Verteilung*, also die bedingte Wahrscheinlichkeitsverteilung der Parameterwerte θ gegeben die Daten D.

$$\Pr(\theta_0|D) \;=\; \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\Pr(D)} \;=\; \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

  Das Ganze geht nur, wenn die a-priori-Wahrscheinlichkeiten $\Pr(\theta)$ definiert sind. $\Pr(D|\theta_0)$ ist gerade die Likelihood $L_D(\theta)$ aus der frequentistischen Statistik. In der Regel hat man es mit kontinuierlichen Parameterräumen zu tun. Dann sind die a-priori- und a-posteriori-Wahrscheinlichkeiten durch Dichten und die Summe durch ein Integral zu ersetzen.

- Wenn man a-posteriori-Verteilungen für Parameter berechnen oder simulieren kann, kann man sich ein Bild davon machen, welche Parameterwerte angesichts der Daten in Frage kommen.
- Statt des ML-Schätzers verwendet man zur Parameterschätzung den Erwartungswert der a-posteriori-Verteilung oder den Wert mit der höchsten a-posteriori-Wahrscheinlichkeit(sdichte) [MAP=maximum a-posteriori].
- Analog zu den Konfidenzintervallen der frequentistischen Statistik gibt es in der Bayesschen Statistik die **Kredibilitätsbereiche**. Ein 95%-Kredibilitätsbereich ist ein Parameterbereich, in dem gemäß der a-posteriori-Verteilung der wahre Parameter mit 95% liegt.