Musterlösung zum Weihnachtsblatt

erstellt von Pascal Neukirchner & Jonas Strauch

W1. ***Ein approximatives Konfidenzintervall für den Erwartungswert bei bekannter Varianz.***

(i) Als Summe von unabhängigen, identisch verteilten Zufallsvariablen ist M_n nach dem Zentralen Grenzwertsatz approximativ normalverteilt. Aus der Unabhängigkeit und der identischen Verteilung der X_i folgt

$$\sigma^2_{M_n} = \mathbf{Var}\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n^2} n\sigma^2 = \sigma^2/n.$$

Und somit folgt $\sigma_{M_n} = \frac{5}{\sqrt{n}}$.

(ii) Nach der 2σ-Regel gilt: $\mathbf{P}(M_n \in [\mu - 2\sigma_{M_n}, \mu + 2\sigma_{M_n}]) \approx 0.95$.
In Aufgabe 27 haben wir festgestellt, dass die Ereignisse $\{M_n \in [\mu - 2\sigma_{M_n}, \mu + 2\sigma_{M_n}]\}$ und $\{\mu \in [M_n - 2\sigma_{M_n}, M_n + 2\sigma_{M_n}]\}$ gleich sind.
Aus Teil i) folgt $2\sigma_{M_n} = \frac{10}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$. Also ergibt sich $c = 2$.

(iii) Statt $c = 2$ wie in Aufgabenteil (ii) wollen wir diesmal (mit einem neuen n) die halbe Intervalllänge $c = 0.2$ erhalten, d.h. dafür muss $\frac{10}{\sqrt{n}} = 0.2$ gelten (Es kann erneut die 2σ-Regel angewandt werden).
Durch Umstellen folgt $50 = \sqrt{n}$ und somit $n = 2500$.

W2.S ***Ein approximatives Konfidenzintervall für den Populationsmittelwert bei bekannter Populationsvarianz.***

a)
$$\mathbf{Var}[M_n] = \frac{1}{n^2} \cdot \mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \cdot \left(\sum_{i=1}^{n} \mathbf{Var}[X_i] + 2 \cdot \sum_{1 \le i < j \le n} \mathbf{Cov}[X_i, X_j]\right).$$

Nun kann zum einen das Ergebnis aus Aufgabe 24d übernommen werden und zur Berechung von $\mathbf{Cov}[X_1, X_2]$ verwendet man wieder den Trick, dass man den Fall $n = g$ betrachtet, dann ist nach Aufgabe 24e auch die Kovarianz gegeben. Es folgt also:

$$\mathbf{Var}[M_n] = \frac{1}{n^2} \cdot (n \cdot \sigma^2 + n \cdot (n-1) \cdot \mathbf{Cov}[X_1, X_2]) = \frac{1}{n^2} \cdot \left(n \cdot \sigma^2 + n \cdot (n-1) \cdot \frac{-\sigma^2}{g-1}\right)$$

$$= \frac{\sigma^2}{n} - \frac{(n-1) \cdot \sigma^2}{n \cdot (g-1)} = \frac{(g - 1 - n + 1) \cdot \sigma^2}{n \cdot (g-1)} = \frac{1}{n} \cdot \frac{g-n}{g-1} \cdot \sigma^2.$$

Und somit folgt $\sigma_{M_n} = \sqrt{\frac{1}{n} \cdot \frac{g-n}{g-1}} \cdot \sigma$.

b) Aus a) kennt man nun also die Varianz von M_n. Der Erwartungswert berechnet sich sehr schnell, denn die X_i haben alle aufgrund der Austauschbarkeit denselben Erwartungswert μ und damit folgt:

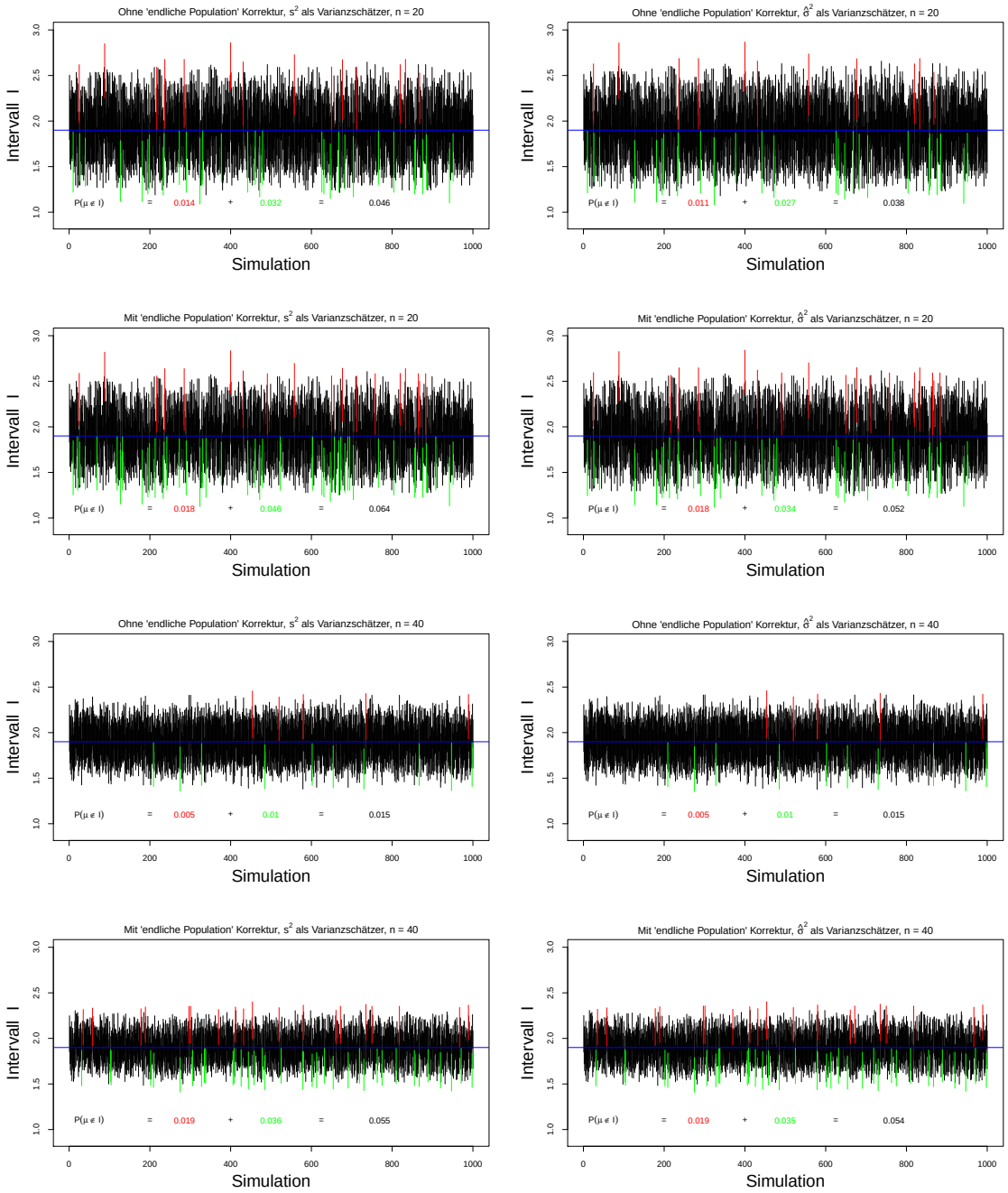
$$\mathbf{E}[M_n] = \frac{1}{n} \cdot n \cdot \mathbf{E}[X_1] = \mu.$$

Schließlich ist aus Folie V7a3 bekannt, dass M_n anähernd normalverteilt ist, da die X_i identisch verteilt sind und für großes n auch beinahe unabhängig, d.h. M_n ist $N(\mu, \frac{1}{n} \cdot \frac{g-n}{g-1} \cdot \sigma^2)$ verteilt.

c) Auch hier kann ähnlich zu Aufgabe W1(ii) die 2σ-Umgebung betrachtet werden, denn aus b) ist bekannt, dass M_n approximativ normalverteilt ist. Demnach ist das gesuchte Intervall

$I_n = [M_n - 2\sigma_{M_n}, M_n + 2\sigma_{M_n}]$.

W3. ***Populationsvarianz geschätzt aus der Stichprobe.***
Mit dem vorliegenden R-Programm simulieren wir die Überdeckungswahrscheinlichkeiten bei 20 bzw 40 maligem Ziehen aus der Baumpopulation aus Aufgabe 20.
Wir ziehen 1000 Stichproben und berechnen jeweils das Konfidenzintervall, überdeckt dieses nicht den tatsächlichen Populationsmittelwert wird es farblich hervorgehoben, nach 1000 Versuchen berechnen wir die empirische Überdeckungswahrscheinlichkeit (Monte-Carlo Verfahren).
Um aus der geschätzten Populationsvarianz die (geschätzte) Varianz des Stichprobenmittels M_n zu erhalten können wir analog zu Aufgabe 3 vorgehen („endliche Population Korrektur“).
Hat man eine sehr große Population (beispielsweise 1000 Befragte bei einer Population von 80 Millionen) kann man stattdessen auch vereinfachend vom Ziehen mit Zurücklegen ausgehen und berechnet $\mathbf{Var}[M_n] = \frac{1}{n}\sigma^2$.
Die Idee hierbei ist, dass die Kovarianzen aufgrund der vergleichsweise großem Population sehr klein sind, bzw dass der Ausdruck $\frac{g-n}{g-1}$ aus der exakten Rechnung nahe an 1 ist.
Wie wir auch aus den Simulationen sehen, fällt bei kleinerer Populationsgröse g der Korrekturfaktor $\frac{g-n}{g-1}$ durchaus ins Gewicht - ohne ihn würde die Varianz überschätzt, die Konfidenzintervalle würden (unnötig) lang und die Überdeckungswahrscheinlichkeit wäre größer als gefordert.
Auf dem Blatt sind zwei Varianzschätzer angegeben, die erwartungstreue Stichprobenvarianz s^2 und der Maximum-Likelihood-Schätzer $\hat{\sigma}^2$.
Hinter dem Faktor $\frac{1}{n-1}$ bei der Stichprobenvarianz (stattdessen hätte man wohl einfach $\frac{1}{n}$ erwartet) steckt der Gedanke, dass im Varianzschätzer nicht mit der Abweichung vom „echten“ Mittelwert μ gerechnet wird, sondern mit dem aus der Stichprobe erreichneten M_n. Die Abweichung von diesem „eigenen“ (an die Stichprobe angepssten) Mittelwert M_n unterschätzt die tatsächliche Varianz, der größere Faktor $\frac{1}{n-1}$ gleicht das aus. Der Unterschied zwischen s^2 und $\hat{\sigma}^2$ wird geringer je größer die Stichprobe ist, da sich $\frac{1}{n-1}$ und $\frac{1}{n}$ für großes n nur noch geringfügig unterscheiden. In der Situation der Aufgabe W3 ist der Unterschied kaum merklich.



W4. ***Rayleighverteilung.***

(i) Für eine Exp($\frac{1}{2}$)-verteilte Zufallsvariable X gilt, dass deren Verteilungsfunktion $1 - e^{-\frac{b}{2}}$ lautet. Somit folgt:

$$\mathbf{P}(R \leq b) = \mathbf{P}(\sqrt{X} \leq b) = \mathbf{P}(X \leq b^2) = 1 - e^{-\frac{b^2}{2}}$$

Da eine exponentialverteilte Zufallsvariable den Wertebereich $[0, \infty)$ hat, folgt also:

$$F(b) = \begin{cases} 1 - e^{-\frac{b^2}{2}} & b \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

(ii) Ableiten von F liefert:

$$f(b) = \begin{cases} b \cdot e^{-\frac{b^2}{2}} & b \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

(iii) In Aufgabe 18 wurde bereits der Erwartungswert einer Standard-Rayleigh-verteilten Zufallsvariable berechnet, dieser lautet $\sqrt{\frac{\pi}{2}}$.

W5. ***Die Fläche unter der Gaußschen Glockenkurve.***
Betrachte zunächst das Quadrat des Integrals, welches berechnet werden soll, dieses kann nach dem einleitenden Hinweis der Aufgabenstellung umgeformt werden:

$$\left(\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du\right)^2 = \int\int_{\mathbb{R}^{\not 2}} e^{-\frac{x^2+y^2}{2}} dxdy$$

Nun lässt sich das Integral wie in Vorlesungsfolie V7a1 lösen.Da x^2+y^2 der quadratische Abstand des Punktes (x, y) vom Ursprung ist und somit auch immer positiv ist, reicht es in den Folien $\sqrt{b}$ durch 0 zu ersetzen:

$$\int\int_{\mathbf{R^2}} e^{-\frac{x^2+y^2}{2}} dxdy = \int_0^{2\pi}\int_0^{\infty} r \cdot e^{-\frac{r^2}{2}} drd\theta$$

$$= \int_0^{\infty}\int_0^{2\pi} r \cdot e^{-\frac{r^2}{2}} d\theta dr = \int_0^{\infty} \left[\theta \cdot r \cdot e^{-\frac{r^2}{2}}\right]_0^{\infty} dr$$

$$= \int_0^{2\pi} d\theta \int_0^{\infty} r \cdot e^{-\frac{r^2}{2}} dr = 2\pi \cdot (-e^{-\frac{r^2}{2}}) \,|_0^{\infty} = 2\pi$$

Und daraus folgt schließlich $\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = \sqrt{2\pi}$.

W6.S ***Wie ändert sich der Regressionskoeffizient unter einer linearen Transformation?***
a)

$$\begin{aligned}
\kappa_{X,Y} &= \frac{\mathbf{Cov}[X,Y]}{\sigma_X \cdot \sigma_Y} \\
&= \frac{\mathbf{Cov}[2G-3, 5H+4]}{\sqrt{\mathbf{Var}[2G-3]} \cdot \mathbf{Var}[5H+4]} \\
&= \frac{\mathbf{Cov}[2G, 5H]}{\sqrt{\mathbf{Var}[2G]} \cdot \sqrt{\mathbf{Var}[5H]}} \\
&= \frac{2 \cdot 5 \cdot \mathbf{Cov}[G,H]}{\sqrt{2^2 \cdot \mathbf{Var}[G]} \cdot \sqrt{5^2 \cdot \mathbf{Var}[H]}} \\
&= \frac{2 \cdot 5 \cdot \mathbf{Cov}[G,H]}{2 \cdot 5 \cdot \sqrt{\cdot \mathbf{Var}[G]} \cdot \sqrt{\cdot \mathbf{Var}[H]}} \\
&= \frac{\mathbf{Cov}[G,H]}{\sigma_G \cdot \sigma_H} \\
&= \kappa_{G,H}
\end{aligned}$$

b) Laut Vorlesung 7a5 gilt für den Regressionskoeffizienten $\beta_1 = \frac{\sigma_Y}{\sigma_X}\kappa_{X,Y}$.
Wir kennen zwar σ_X und σ_Y nicht, dafür aber die Regressionsgerade für H auf Basis von G. Diese

hat laut Angabe den Anstieg 3, also ist $\frac{\sigma_H}{\sigma_G}\kappa_{G,H} = 3$. Für den Anstieg der Regressionsgeraden für Y auf der Basis von X berechnen wir

$$\begin{aligned}\beta_1 &= \frac{\sigma_Y}{\sigma_X}\kappa_{X,Y}\\ &= \frac{\sigma_{5H+4}}{\sigma_{2G-3}}\kappa_{G,H}\\ &= \frac{5\sigma_H}{2\sigma_G}\kappa_{G,H}\\ &= \frac{5}{2}\cdot\frac{\sigma_H}{\sigma_G}\kappa_{G,H}\\ &= \frac{15}{2}.\end{aligned}$$

c) Es gilt $\beta_0 = \mu_Y - \beta_1\mu_X$. Wie in b) mit der Standardabweichung geschehen wollen wir nun mittels der Erwartungswerte μ_G und μ_H sowie der Linearität des Erwartungswertes μ_X bzw μ_Y berechnen.
Um μ_H zu berechnen nutzen wir die gleiche Formel für $\tilde{\beta}_0$ der Regressionsgerade für H auf Basis von G, es ist also:

$$\begin{aligned}\tilde{\beta}_0 &= \mu_H - \tilde{\beta}_1\mu_g\\ 2 &= \mu_H - 3\cdot 3\\ \mu_H &= 11\end{aligned}$$

Damit berechnen wir

$$\begin{aligned}\beta_0 &= \mu_Y - \beta_1\mu_X\\ &= \mu_{5H+4} - \frac{15}{2}\mu_{2G-3}\\ &= 5\mu_H + 4 - \frac{15}{2}(2\mu_G - 3)\\ &= 5\cdot 11 + 4 - \frac{15}{2}(2\cdot 3 - 3)\\ &= \frac{73}{2}.\end{aligned}$$

W7. ***Der Münzwurf lässt grüßen.***
a) $X_1 + X_2$ ist Binom$(n_1 + n_2, p)$-verteilt. Denn:
X_1 zählt die Anzahl der Erfolge bei n_1 Würfen, X_2 bei n_2 Würfen, da die Erfolgswahrscheinlichkeit bei beiden gleich ist werfen wir also insgesamt $n_1 + n_2$-mal und zählen mit $X_1 + X_2$ die Gesamtanzahl der Erfolge.

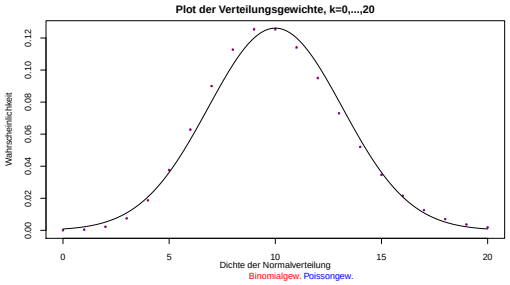
b)(i)

$$\begin{aligned}
\mathbf{P}(Y_1+Y_2=k) &= \sum_{\ell=0}^{k} \mathbf{P}(Y_1=\ell, Y_2=k-\ell) \\
&= \sum_{\ell=0}^{k} \mathbf{P}(Y_1=\ell)\cdot\mathbf{P}(Y_2=k-\ell) && (Y_1, Y_2 \text{ unabhängig}) \\
&= \sum_{\ell=0}^{k} e^{-\alpha}\frac{\alpha^\ell}{\ell!}\cdot e^{-\beta}\frac{\beta^{k-\ell}}{(k-\ell)!} \\
&= e^{-(\alpha+\beta)}\sum_{\ell=0}^{k}\alpha^\ell\beta^{k-\ell}\cdot\frac{1}{\ell!\cdot(k-\ell)!}\cdot\frac{k!}{k!} \\
&= e^{-(\alpha+\beta)}\frac{1}{k!}\sum_{\ell=0}^{k}\alpha^\ell\beta^{k-\ell}\cdot\binom{k}{l} \\
&= e^{-(\alpha+\beta)}\frac{1}{k!}\cdot(\alpha+\beta)^k = e^{-(\alpha+\beta)}\frac{(\alpha+\beta)^k}{k!},
\end{aligned}$$

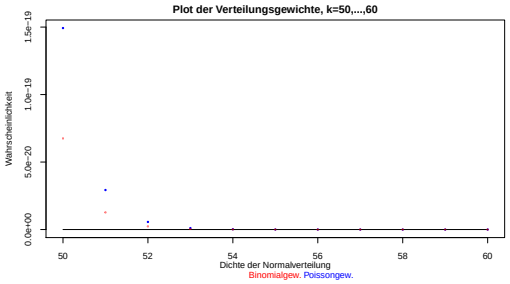
wobei wir in der vorletzten Gleichheit den Binomischen Lehrsatz verwendet haben.
(ii) Y_1+Y_2 ist also Poisson zum Parameter $\alpha+\beta$ verteilt.

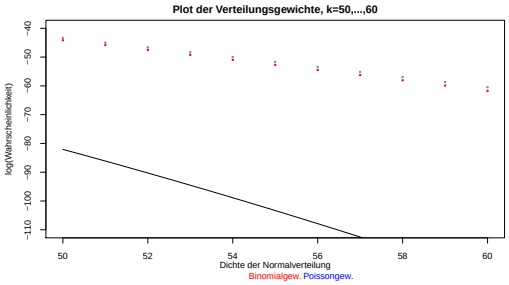
W8. ***Bernoulli, Poisson und Gauß geben sich die Hand.***
a) Wir haben zwei Möglichkeiten kennengelernt die Verteilungsgewichte einer binomial zu den Parametern $n=1000$ und $p=\frac{1}{100}$ verteilten Zufallsvariable zu approximieren. Erstens die Poissonapproximation, Approximation mittels einer zum Parameter $\lambda=n\cdot p=10$ poisson verteilten Zufallsvariable. Zweitens die Normalapproximation, eine Folgerung aus dem zentralen Grenzwertsatz, Approximation mittels einer zu den Parametern $\mu=n\cdot p=10$ und $\sigma^2=n\cdot p\cdot(1-p)=9.9$ normalverteilten Zufallsvariable.
So sehen wir direkt, dass binom$(1000,\frac{1}{100})$ und pois(10) approximativ gleich sind, mit der Normalapproximation erhalten wir so eine $\mathrm{N}(10,9.9)(\approx \mathrm{N}(10,10))$-verteilte Zufallsvariable.

b)(i)

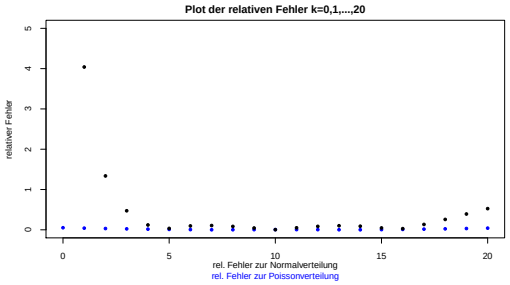


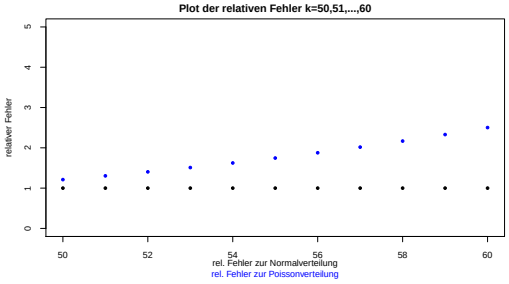
(ii)





c)





d) Wir sehen dass man bei kleinem p eher die Poissonapproximation verwenden sollte: die Gewichte der Binomialverteilung sind vergleichsweise „unsymmetrisch“ was auch im interessanten Bereich rund um den Erwartungswert zu „sichtbaren“ Fehlern bei der Normalapproximation führt.