Lösungsvorschläge für die Übungsaufgaben der Vorlesung “Statistik“

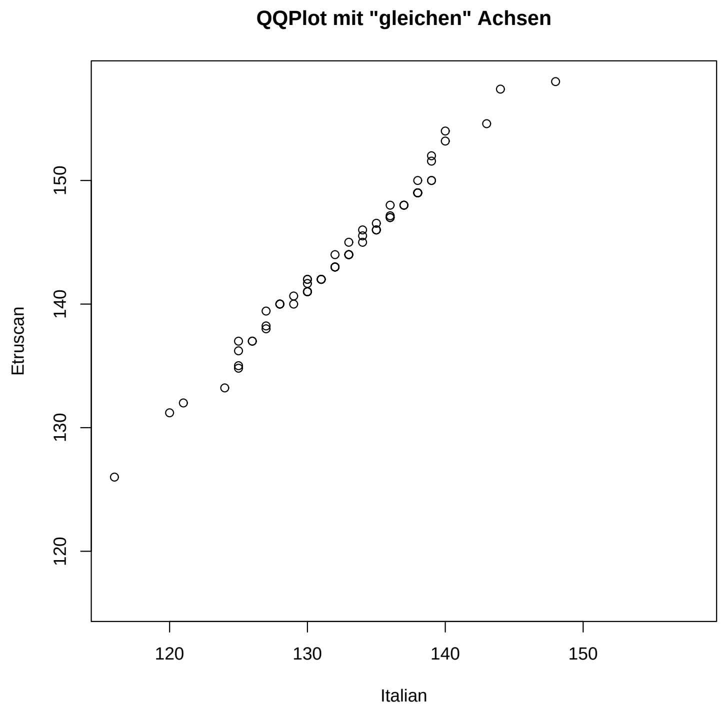
Thomas Rupp
thomas@7t7.de

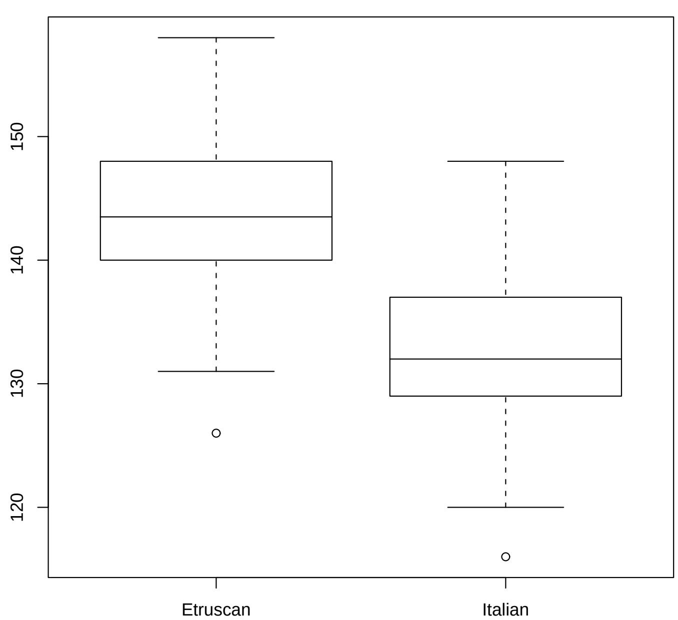
6. Dezember 2001

Aufgabe 5

a)

Als erstes sehen wir uns einen QQPlot und einen Boxplot an:





Bei beiden Bildern lässt sich das gleiche Ergebnis ablesen:
Die Etruskerschädel stammen wohl aus der gleichen Verteilung, sind aber verscho-

ben (sind größer).
Dies untermauert auch der Wilcoxen Test: Er testet ob zwei Stichproben aus der gleichen Verteilung stammen. Er testet aber nur auf die Alternative einer verschobenen Verteilung.
Stammen die Stichproben nicht aus der gleichen oder einen verschobenen Verteilung, so liefert der Test nur bei sehr kleinen p-Werten ein brauchbares Ergebnis. In diesem Fall erhalten wir ein p-value von $4 \cdot 10^{-19}$. Da wir gesehen haben, dass die Verteilung wohl verschoben ist, können wir die Nullhypothese, dass beide Schädelgruppen aus dem gleichen Topf stammen einwandfrei ablehnen.

Auch ein Blick auf die Differenz der Mediane von 11.5 und der Mittelwerte von 11.33 bei einer Standardabweichung von 5.9 bei den Etruskern und 5.75 bei den Italienern bestätigen obiges Ergebnis (diese Std.Abweichungen machen eine derart große, zufällige Verschiebung sehr unwahrscheinlich).

b)

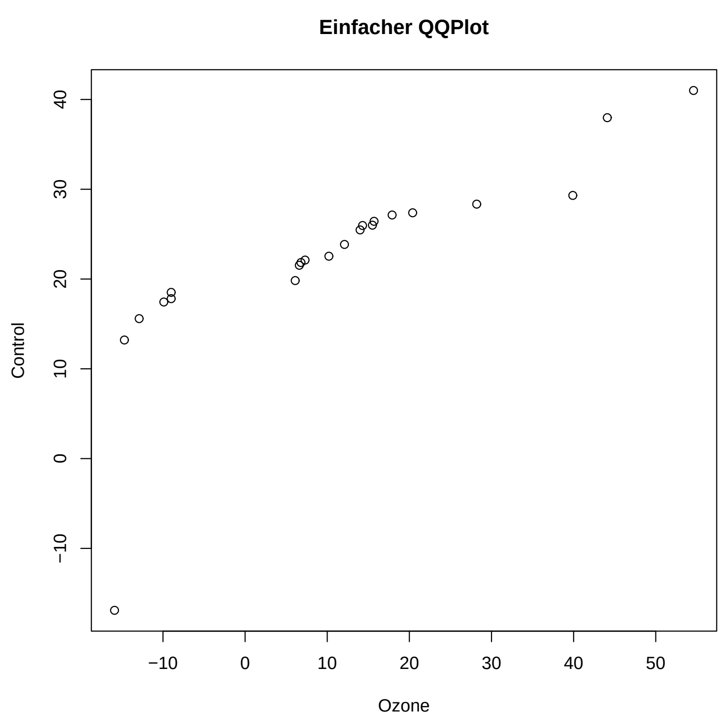
Es gibt aber einige Argumente, die eine derartige Nullhypothese sehr unplausibel erscheinen lassen:

- Ist die Kontrollgruppe wirklich vergleichbar (sicherlich wurde nicht jeder Etrusker so bestattet, dass deren Schädel als etruskisch identifiziert werden konnte)

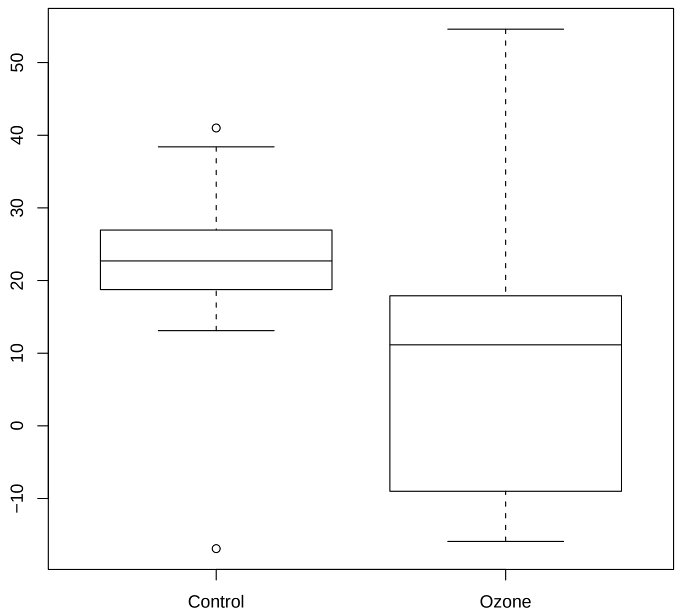
- Gibt es überhaupt einen gesicherten Zusammenhang zwischen Schädelumfang und ethnischen Stammbaum (Integration anderer Stämme, grundlegende Frage über Aussagefähigkeit, etc.)

Diese beiden Hauptpunkte könnten noch weiter ausgeführt werden.

Aufgabe 6 Sehen wir uns ersteinmal den qqplot an:



Eine Verschiebung ist hier eigentlich nicht zu erkennen. Wenn überhaupt, dann lässt sich eine grössere Streuung vermuten. Ein Boxplot gibt mehr Aufschluss:



Hier können wir sowohl eine grössere Streuung (höhere Varianz) und eine leichte Verschiebung erkennen. Da letztere aber eher die kleinere Rolle spielt und der Wilcoxtest nur bei möglichen Verschiebungen sinnvoll ist, sollte man diesem Test mit Vorsicht geniessen (vor allem bei größeren p-Werten). Er liefert eine p-Wert von 0.0028, welcher noch klein genug ist und somit die Nullhypothese (Ozon hat keine Auswirkungen) unglaubwürdig macht.

Als Fazit lässt sich festhalten, dass Ozon wohl etwas zur Gewichtsverringerung beiträgt und die Änderungen der Gewichte der Ratten verstärkt.

Aufgabe 7

a)

Der Median ist definiert durch $\mu := F^{-1}(0.5)$ bzw. $F(\mu) \leq \frac{1}{2}$ und $F(\mu+) \geq \frac{1}{2}$.

Die Verteilungsfunktion ist rechtsstetig und die linken Limiten existieren (also $F(x) = \nu((-\infty, x))$.

Offensichtlich ist $g(x)$ eine konvexe Funktion und O.B.d.A. beschränkte. Daher besitzt sie ein Minima x, deren linksseitige Ableitung negativ und rechtsseitige Ableitung positiv ist. Die Ableitung von links ist

$$\begin{aligned}
\lim_{h\to 0} \frac{g(x-h)-g(x)}{-h} &= \lim_{h\to 0} \left[\int_{-\infty}^{x-h} \frac{(x-h-z)-(x-z)\nu(dz)}{-h} \right. \\
&\quad \left. + \int_{x-h}^{\infty} \frac{(z-x+h)-(z-x)\nu(dz)}{-h} \right] \\
&= \int 1_{(-\infty,x]} - 1_{(x,\infty)} \nu(dz) \\
&= F(x) - (1-F(x)) \\
&\leq 0 \Leftrightarrow F(x) \leq \frac{1}{2}
\end{aligned}$$

Analog zeigt man über die rechte Ableitung, dass $F(x+) \geq \frac{1}{2}$.
Also ist dieses Minima $x =: \mu$ der Median.

b)

$$\begin{aligned}
|\mu - \tilde{\mu}| &= |\mathbb{E}(X - \tilde{\mu})| \\
&= |\mathbb{E}(X) - \tilde{\mu}| \\
&\leq \mathbb{E}|X - \tilde{\mu}| \\
&\leq \mathbb{E}|X - \mu| \qquad \text{a)} \\
&\leq \sqrt{\mathbb{E}\left[(X - \mathbb{E}(X))^2\right]} \qquad \text{Höldersche Ungl.: } \mathbb{E}(A)\mathbb{E}(B) \leq \sqrt{E(A^2)\mathbb{E}(B^2)} \\
&= \sigma
\end{aligned}$$

Aufgabe 8
Wir wollen zeigen, dass $\sqrt{n}(M_n - \tilde{\mu}) \to \mathfrak{N}(0, (2f(\tilde{\mu}))^{-2}$ konvergiert. Dazu sehen wir uns

$$\lim P(\sqrt{n}(M_n - \tilde{\mu}) \leq a) \Leftrightarrow \lim P(M_n \leq \frac{a}{\sqrt{n}} + \tilde{\mu})$$

an. Dieses Ereignis $\{M_n \leq \frac{a}{\sqrt{n}} + \tilde{\mu}\}$ können wir auch, da wir nur ungerade n betrachten, schreiben als $\{\sum Y_i \geq \frac{n+1}{2}\}$. Dabei ist Y_i eine Bernoulli-Zufallsvariable, die genau dann 1 ist, wenn $x_i \leq \frac{a}{\sqrt{n}} + \tilde{\mu}$ und Null sonst.
Die Summe der Y_i zählt also die Anzahl der Ausgänge, an denen X kleiner als $\frac{a}{\sqrt{n}} + \tilde{\mu}$ ausfällt. Sind das mehr als die Hälfte, dann ist auch der Stichproben-Median M_n kleiner als dieser Wert. Weiterhin ist $P(Y_i = 1) = F(\frac{a}{\sqrt{n}} + \tilde{\mu}) =: p_n$.

Somit erhalten wir

$$P(\sqrt{n}(M_n - \tilde{\mu}) \leq a) = P(\sum Y_i \geq \frac{n+1}{2}) = P\left(\frac{\sum Y_i - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{n/2 + 1/2 - np_n}{\sqrt{np_n(1-p_n)}}\right).$$

Für $n \to \infty$ konvergiert $p_n \to \frac{1}{2} = F(\tilde{\mu})$ und somit gilt für das Argument der RHS der oberen Zeile:

$$\begin{aligned}
\frac{n/2 + 1/2 - np_n}{\sqrt{np_n(1-p_n)}} &= \frac{1}{2\sqrt{n}\sqrt{p_n - p_n^2}} + \frac{n}{\sqrt{p_n - p_n^2}} \cdot \frac{F(\tilde{\mu}) - F(\mu - a/\sqrt{n})}{\sqrt{n}(\frac{-a}{\sqrt{n}} \frac{\sqrt{n}}{-a})} \\
&\to -2aF'(\mu)
\end{aligned}$$

Dabei haben wir benutzt, dass $F(\tilde{\mu})$ differenzierbar ist, demnach dort keine Sprungstelle und somit $= 1/2$ ist. Für den Differentialquotienten haben wir $a/\sqrt{n}$ benutzt.
Nach dem zentralen Grenzwertsatz ist $Z := \lim \frac{\sum Y_i - np_n}{\sqrt{np_n(1-p_n)}}$ standardnormalverteilt.
Da diese symmetrisch ist, gilt

$$\begin{aligned}
P(\sqrt{n}(M_n - \tilde{\mu}) \leq a) &\to P(Z \geq -2af(\mu)) = P(Z \leq 2af(\mu)) \\
&= P(\frac{Z}{2f(\mu)} \leq a)
\end{aligned}$$

Da $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ und $\mathbb{E}(Z) = 0$, ist $Z(2f(\mu))^{-2}$ also $\mathfrak{N}(0, \frac{1}{(2f(\mu))^2})$-verteilt.