

K14 Verteilungen von Ordnungsstatistiken

14.1 Dichten von Ordnungsstatistiken

X_1, X_2, \dots, X_n u.i.v. mit Dichte $f(x)$

Die gemeinsame Dichte ist

$$\prod_{i=1}^n f(x_i)$$

Wir setzen die Daten in aufsteigender Reihenfolge:

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}$$

und betrachten die Zufallsvariablen $Y_j = X_{(j)}$

Dann gilt

$$f_{Y_1, Y_2, \dots, Y_n}(y_1, y_2, \dots, y_n) = n! \prod f(y_j)$$

und die Marginaldichte für Y_k ist

$$f_{Y_k}(y_k) = n! f(y_k) \int_{-\infty}^{y_k} \dots \int_{-\infty}^{y_2} \int_{y_{n-1}}^{\infty} \dots \int_{y_k}^{\infty} \prod_{j \neq k} f(y_j) dy_j$$

weil wir über alle andere Y_j unter Aufrechterhaltung der Reihenfolge integrieren müssen.

Die Integration trennt sich in zwei, die Werte kleiner als y_k und die Werte größer als y_k . Man muss sie den Reihen nach durcharbeiten und die folgenden Resultate sind sehr hilfreich:

$$\int_{-\infty}^y F(t)^m f(t) dt = \frac{F(y)^{m+1}}{m+1}$$

$$\int_y^{\infty} (1 - F(t))^m f(t) dt = \frac{(1 - F(y))^{m+1}}{m+1}$$

Daraus folgt, dass

$$\begin{aligned} f_{Y_k}(y_k) &= \frac{n!}{(n-k)!(k-1)!} (1-F(y))^{n-k} F(y)^{k-1} f(y) \\ &= n f(y) \binom{n-1}{k-1} F(y)^{k-1} (1-F(y))^{n-k} \quad (1) \end{aligned}$$

Wir wählen 1 aus n , multiplizieren mit der Dichte $f(y)$, wählen $(k-1)$ aus den verbleibenden $(n-1)$ und verlangen, dass diese alle $< y$ sind, während die anderen $(n-k)$ alle $> y$ sind.

14.2 Die Verteilungsfunktion einer Ordnungsstatistik

$$P(Y_k \leq x \text{ und } Y_{k+1} \geq x) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

$$P(Y_k \leq x) = \sum_{j=k}^n \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$$

14.3 Warum sind Ordnungsstatistiken vom Interesse?

- Maximum — der höchste/beste Wert
- Minimum — der niedrigste/schlechteste Wert
- Median

Sport, Wetter, Rekorde

Ordnungsstatistiken hängen von n (der Stichprobengröße) ab.

14.4 Minima und Maxima

Minimum

$$f_{\min}(y) = nf(y)(1 - F(y))^{n-1}$$

$$P(\min \leq y) = 1 - P(\text{alle} > y)$$

$$F_{\min}(y) = 1 - (1 - F(y))^n \quad (1)$$

Maximum

$$f_{\max}(y) = nf(y)F(y)^{n-1}$$

$$P(\max \leq y) = P(\text{all} \leq y)$$

$$F_{\max}(y) = F(y)^n \quad (2)$$

z.B: $X \sim E(\lambda)$

$$f_{\min}(y) = n\lambda e^{-n\lambda y} \sim E(n\lambda)$$

$$f_{\max}(y) = n\lambda e^{-\lambda y} (1 - e^{-\lambda y})^{n-1}$$

$$E[\min] = \frac{1}{n\lambda} \text{ und } E[\max] = \frac{1}{\lambda} \sum_{i=1}^n \frac{1}{i}$$

14.4.1 Maxima von normalverteilten Zufallsvariablen

$$F_{\max}(x) = F(x)^n$$

$$X_i \sim N(\mu, \sigma^2)$$

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

z.B. $X \sim N(5, 1)$

$$F(8) = \Phi(3) = 0.99865$$

$$F_{\max_{10}}(8) = \Phi(3)^{10} = 0.98658$$

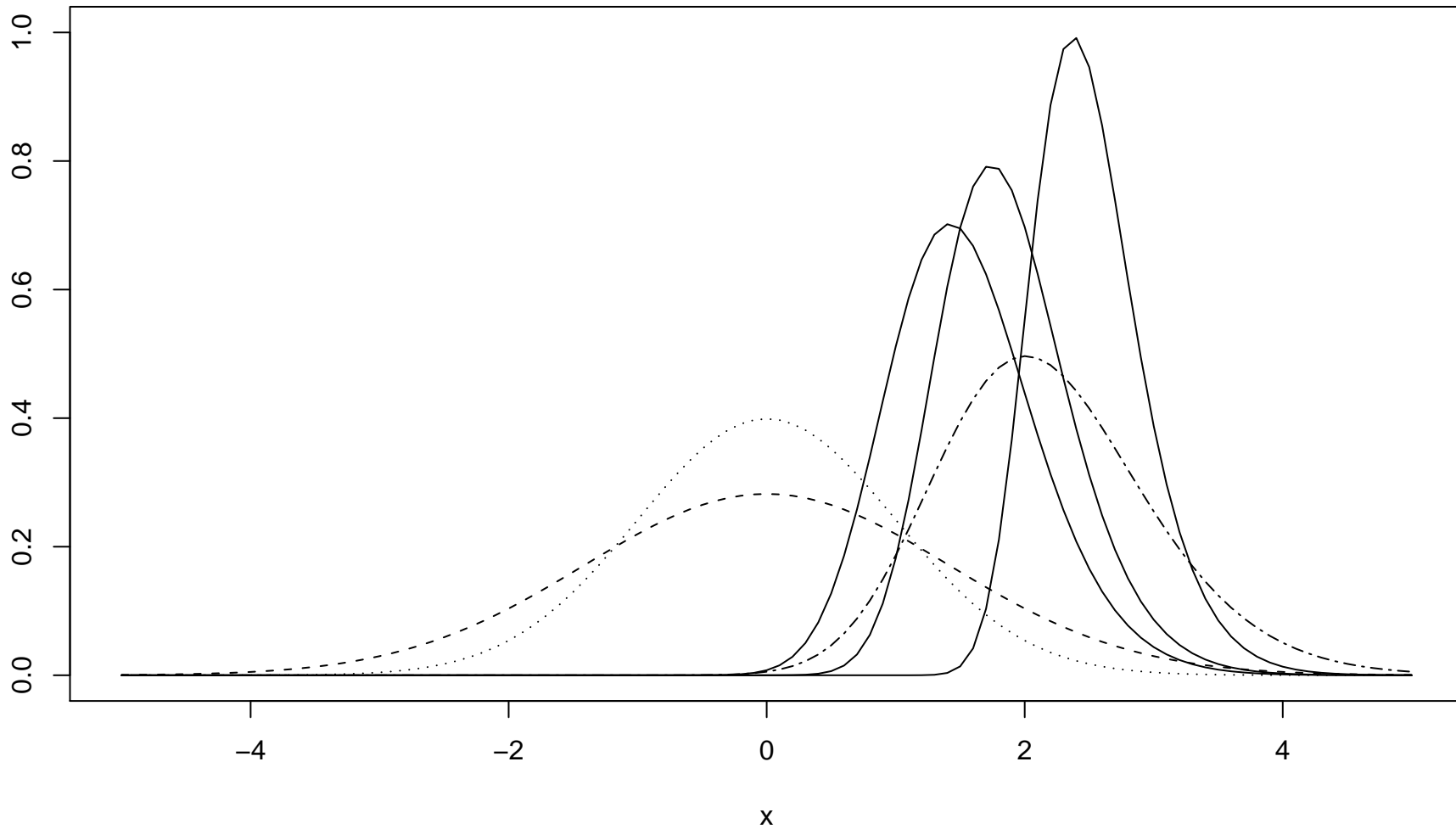
$$F_{\max_{20}}(8) = \Phi(3)^{20} = 0.97334$$

$$F_{\max_{100}}(8) = \Phi(3)^{100} = 0.87364$$

$$F_{\max_{1000}}(8) = \Phi(3)^{1000} = 0.25901$$

Die Wahrscheinlichkeit eines Werts größer 8 ist 0.00135, aber die Wahrscheinlichkeit, dass das Maximum aus 100 unabhängigen Stichproben größer 8 ist, liegt bei 0.1264 und für 1000 ist die Wahrscheinlichkeit knapp unter 0.75.

Dichten von Normalverteilung Maxima: $N(0,1)$ ($n=10,20,100$) & $N(0,2)$ ($n=10$)



14.5 Die Verteilung des Medians

Wir müssen zwei Fälle nach der Größe der Stichprobe unterscheiden:

$$n = 2m + 1 \quad \text{und} \quad n = 2m$$

Für $n = 2m + 1$ ist $Y_{m+1} = X_{(m+1)}$ der Median und

$$f_{Y_{m+1}}(y) = (2m + 1)f(y) \binom{2m}{m} F(y)^m (1 - F(y))^m$$

Für $n = 2m$ wird der Median als $\frac{1}{2}(X_{(m)} + X_{(m+1)})$ definiert. Wir brauchen zuerst die gemeinsame Dichte von $X_{(m)}$ und $X_{(m+1)}$

$$f_{Y_m, Y_{m+1}}(u, v) = \binom{2m}{2} f(u) f(v) \binom{2m-2}{m-1} F(u)^{m-1} (1 - F(v))^{m-1}$$

Dann müssen wir die Dichte von $W = \frac{1}{2}(Y_m + Y_{m+1})$ unter der Bedingung $Y_m \leq Y_{m+1}$ berechnen.

14.6 Ordnungsstatistiken und Rekorde

$$P(\text{Max aus } n > r) = 1 - F(r)^n$$

Exponentialverteilungen

$$X \sim E(\lambda) \Rightarrow P(\max > r) = 1 - (1 - e^{-\lambda r})^n$$

Gegeben zwei Teilnehmer mit Leistungsverteilungen

$$X \sim E(\lambda_A) \text{ und } Y \sim E(\lambda_B)$$

und n Chancen für beide, wie groß ist die Wahrscheinlichkeit, dass Y gewinnt, wenn $\frac{1}{\lambda_A} > \frac{1}{\lambda_B}$?

$$\begin{aligned} P(\max Y > \max X) &= \int_t P(\max X = t \text{ und } \max Y > t) \\ &= \int_0^\infty n\lambda_A e^{-\lambda_A t} (1 - e^{-\lambda_A t})^{n-1} (1 - (1 - e^{-\lambda_B t})^n) dt \end{aligned}$$

Um die Frage zu beantworten: "Ist Y besser als X , wenn $\max X > \max Y$?", brauchen wir apriori Verteilungen für λ_A und λ_B .

14.6.1 Normalverteilungen

Sei Leistung modelliert als

$$X_i \sim N(\mu_i, \sigma^2)$$

Eine Möglichkeit für die Verteilung von μ unter der Bevölkerung wäre

$$\mu \sim E(\lambda)$$

d.h. die meisten bringen nichts oder sehr wenig, aber es gibt einige ausgezeichnete.

Die Leistung von einem unbekanntem an einem Tag hat die Dichte

$$\begin{aligned} f(x) &= \int_0^{\infty} \lambda e^{-\lambda\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} d\mu \\ &= \lambda e^{-\lambda x + \frac{1}{2}\lambda^2\sigma^2} \left(1 - \Phi\left(\frac{\lambda\sigma^2 - x}{\sigma}\right)\right) \end{aligned}$$

Die a posteriori Dichte für $\mu|X_i = x$ ist nach Bayes

$$f(\mu|x) = \frac{f(x|\mu)g(\mu)}{\int_{\mu} f(x|\mu)g(\mu)d\mu}$$

$$\begin{aligned}
&= \frac{\lambda e^{-\lambda\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\lambda e^{-\lambda x + \frac{1}{2}\lambda^2\sigma^2} \left(1 - \Phi\left(\frac{\lambda\sigma^2 - x}{\sigma}\right)\right)} \\
&= \left(1 - \Phi\left(\frac{\lambda\sigma^2 - x}{\sigma}\right)\right)^{-1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu - (x - \lambda\sigma^2)}{\sigma}\right)^2}
\end{aligned}$$

eine Normalverteilung $N(x - \lambda\sigma^2, \sigma^2)$, die auf $[0, \infty)$ beschränkt ist.

Da wir annehmen können, dass die Schwächeren nicht daran teilnehmen, könnten wir $g(\mu)$ ersetzen mit

$$\begin{aligned}
h(\mu) &= \lambda e^{-\lambda(\mu-a)} & \mu > a \\
&= 0 & \text{sonst}
\end{aligned}$$

Aber nicht alle Guten nehmen daran teil, so dass eine weitere Verfeinerung wäre

$$\begin{aligned}
P(\text{Teilnahme}|\mu) &= 1 - e^{-\gamma(\mu-a)} & \mu > a \\
&= 0 & \text{sonst}
\end{aligned}$$

Vergleich zwischen $N(\mu=5)$ und $N(\mu \text{ exponential-verteilt})$

