K 13 Wie gut ist der ZGS?

13.1 Wahrscheinlichkeitsmetrik

Um zwei Verteilungen zu vergleichen, benutzen wir

$$d(F, G) = \sup_x |F(x) - G(x)|$$

den maximalen Unterschied zwischen den Verteilungsfunktionen. (Es gibt natürlich andere Möglichkeiten auch.)

z.B. **U(0,1) und N(0,1)** $X \sim U(0, 1)$

Um mit einer standard Normalverteilung zu vergleichen, nehmen wir

$$Y = \frac{X - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$$

weil

$$E[Y] = 0 \quad \text{und} \quad V[Y] = 1$$

$$f(y) = \frac{1}{2\sqrt{3}} \qquad -\sqrt{3} < y < \sqrt{3}$$

$$\begin{aligned} d(F(y), \phi) &= \max\left(\phi(-\sqrt{3}), \max_{(-\sqrt{3},0)}\left(\phi(y) - \frac{\sqrt{3}+y}{2\sqrt{3}}\right)\right) \\ &\approx 0.057 \end{aligned}$$

13.2 Berry-Esséen

Seien $X_1, X_2, \ldots$ u.i.v. Ist $0 < \sigma^2 = V(X_i) < \infty$ und $\gamma = E[|X - \mu|^3] < \infty$, so gilt

$$d(S_n^*, \phi) \leq \frac{0.7655\gamma}{\sigma^3\sqrt{n}}$$

$\left(S_n^*\right.$ ist die Verteilungsfunktion von $\left.\frac{\sum X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)$

13.2.1 Berry-Esséen und die Gleichverteilung

$$X_i \sim U(0,1)$$

$$\mu = \frac{1}{2} \quad \text{und} \quad \sigma^2 = \frac{1}{12}$$

$$\gamma = E[|X - \frac{1}{2}|^3] = 2\int_{\frac{1}{2}}^{1}(x - \frac{1}{2})^3 \mathrm{d}x = \frac{1}{32}$$

$$d(S_n^*, \phi) \leq \frac{0.8 * \frac{1}{32}}{(\frac{1}{12})^{\frac{3}{2}}\sqrt{n}}$$

$$n = 12 \Rightarrow d(S_{12}^*, \phi) \leq 0.3$$

$$n = 300 \Rightarrow d(S_{300}^*, \phi) \leq 0.06$$

13.2.2 Berry-Esséen und die Exponentialverteilung

$$X \sim E(\lambda)$$

$$\mu = \lambda^{-1} \quad \text{und} \quad \sigma^2 = \lambda^{-2}$$

$$\gamma = E[|X - \mu|^3] = \frac{1}{\lambda^3}(12e^{-1} - 2)$$

$$d(S_n^*, \phi) \leq \frac{0.8 * \frac{1}{\lambda^3}(12e^{-1} - 2)}{(\frac{1}{\lambda})^3\sqrt{n}}$$

$$\approx \frac{1.9316}{\sqrt{n}}$$

$$n = 12 \Rightarrow d(S_{12}^*, \phi) \leq 0.5576$$

$$n = 300 \Rightarrow d(S_{300}^*, \phi) \leq 0.1115$$

$$n = 30000 \Rightarrow d(S_{30000}^*, \phi) \leq 0.01115$$

Sei $F(s)$ die genaue Verteilungsfunktion für $n = 300$, dann sagt das Resultat, dass

$$|F(s) - \phi(s)| \leq 0.1115$$

$$\phi(s) - 0.1115 < F(s) < \phi(s) + 0.1115$$

13.3 Das Gesetz vom iterierten Logarithmus

$$\{X_i\} \quad \text{u.i.v. } E[X] = 0 \text{ und } V[X] = 1$$

$$S_n = \sum_{i=1}^{n} X_i$$

Wir wissen aus dem SGGZ, dass

$$S_n/n \to 0 \quad \text{fast sicher}$$

und aus dem ZGS, dass

$$U_n = S_n/\sqrt{n} \to N(0,1) \quad \text{in Verteilung}$$

Wie gross dürfen die (sehr seltenen) Fluctuationen von U_n sein? Es kann gezeigt werden, dass

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = 1 \quad \text{fast sicher} \quad (\liminf = -1)$$

Der Satz gilt auch für $\{X_i\}$ u.i.v. im allgemeinen. Zum Beweis muß gezeigt werden, dass das Ereignis

$$A_n = \{S_n \geq c\sqrt{2n \log\log n}\}$$

unendlich oft passiert für $c < 1$ und nur endlich oft für $c > 1$ mit Wahrscheinlichkeit 1.

13.4 Das Arcussinus Gesetz

$$\text{Sei} \quad P(X_i = -1) = P(X_i = 1) = 0.5$$

und

$$S_n = \sum_{i=1}^{n} X_i$$

Sei $M_n(x_1, x_2, \ldots, x_n)$ die Anzahl jener Partiellsummen S_k, die positiv sind, dann gilt

$$P\left(a \leq \frac{M_n(x_1, \ldots, x_n)}{n} \leq b\right) \to \int_a^b \frac{1}{\pi\sqrt{x(1-x)}} \mathrm{d}x$$

die Arcussinusverteilung über (0,1).

Oder in anderer Form.
Sei $L_{2N} = \max\{2n \leq 2N : S_{2n} = 0\}$ der Zeitpunkt des letzten Nullpunkts. Für alle $0 < a < b < 1$ gilt

$$\lim_{N\to\infty} P(a \leq \frac{L_{2N}}{2N} \leq b) = \int_a^b \frac{1}{\pi\sqrt{x(1-x)}} \mathrm{d}x$$

Beweis vom Arcussinus Gesetz

(1) Sei $G_n = (S_{2n} = 0, S_{2k} \neq 0$ für $1 \leq k \leq n)$, die erste Rückkehr zu 0 nach $2n$ Schritten und sei

$$u_n = 2^{-2n}\binom{2n}{n}$$

dann ist

$$P(G_n) = u_{n-1} - u_n$$

Man veranschaulicht die Pfade von

$$S_1 = 1 \text{ bis } S_{2n-1} = 1$$

und von

$$S_1 = -1 \text{ bis } S_{2n-1} = 1$$

Mit Hilfe des Reflexionsprinzip fällt das Resultat aus.

(2) Sei $G_{>n} = (S_{2n} \neq 0$ für $1 \leq k \leq n)$, keine Rückkehr während der ersten $2n$ Schritte.

$$P(G_{>n}) = u_n$$

weil $P(G_{>n}) = \sum_{i=n+1}^{\infty} P(G_n)$

(3) Sei $P(L_{2N} = 2n)$ die Wahrscheinlichkeit dass die letzte Rückkehr nach $2n$ Schritten passiert und dass es keine weitere Rückkehr bis nach $2N$ gibt.

$$= u_n * u_{N-n}$$

$$= 2^{-2N}\binom{2n}{n}\binom{2(N-n)}{(N-n)}$$

Mit Hilfe der Stirling Formel

$$n! \sim \sqrt{2\pi n} n^n e^{-n}$$

folgt das Resultat.

K14 Verteilungen von Ordnungsstatistiken

14.1 Dichten von Ordnungsstatistiken

$$X_1, X_2, \ldots, X_n \quad \text{u.i.v. mit Dichte} \quad f(x)$$

Die gemeinsame Dichte ist

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

Wir setzen die Daten in aufsteigender Reihenfolge:

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

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

Dann gilt

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

und die Marginaldichte für Y_k ist

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

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:

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

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

Daraus folgt, dass

$$f_{Y_k}(y_k) = \frac{n!}{(n-k)!(k-1)!}(1-F(y))^{n-k}F(y)^{k-1}f(y)$$

$$= nf(y)\binom{n-1}{k-1}F(y)^{k-1}(1-F(y))^{n-k} \qquad (1)$$

Wir wählen 1 aus n, multipliziern 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

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

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

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

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

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

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

Maximum

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

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

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

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

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

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

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

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

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

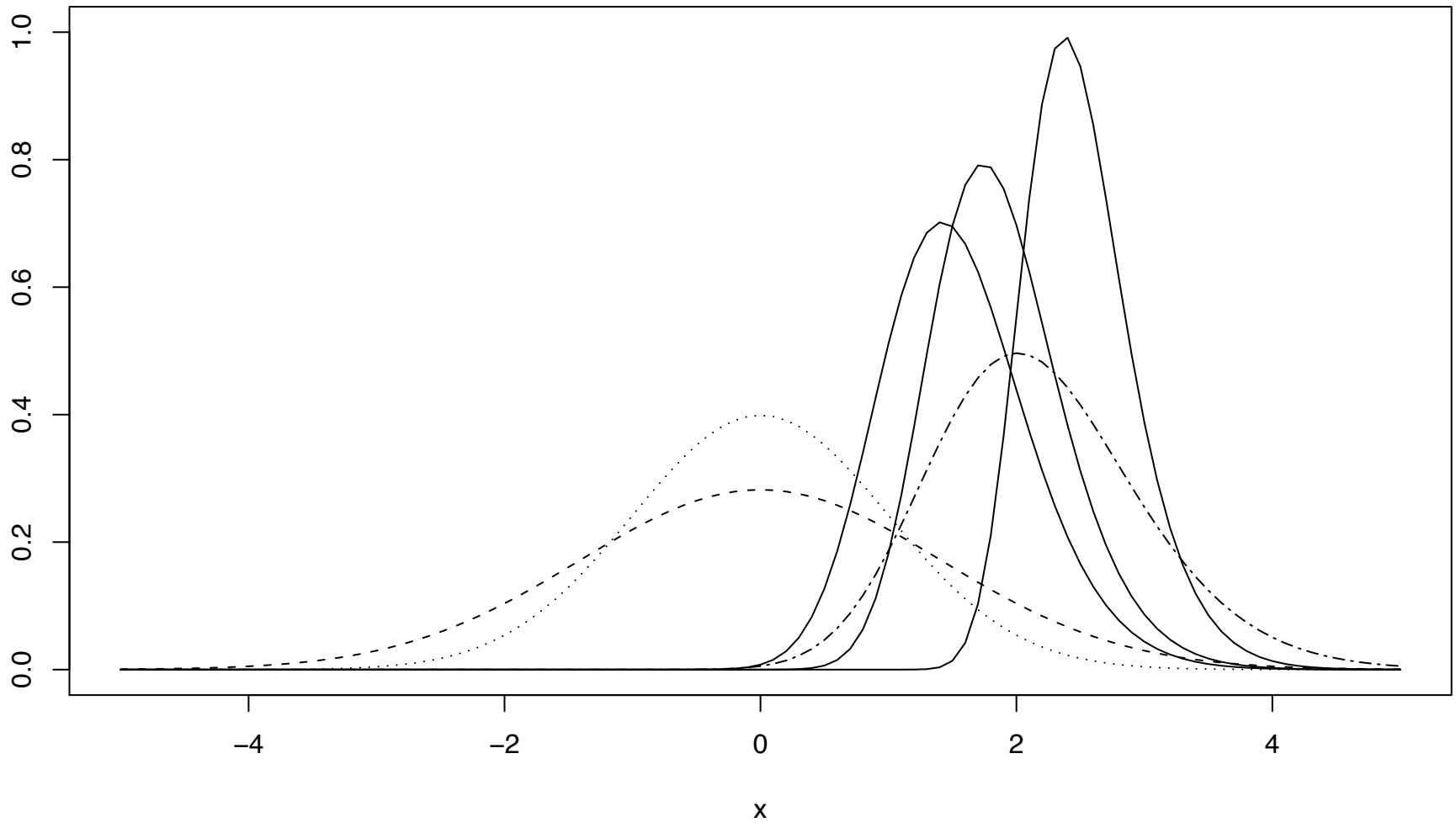
$$F_{\max_{10}}(8) = \Phi(3)^{10} = 0.98658$$

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

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

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

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)
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x

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 $\frac{1}{\lambda_A} > \frac{1}{\lambda_B}$?

$$P(\max Y > \max X) = \int_t P(\max X = t \text{ und } \max Y > t)$$

$$= \int_0^\infty n\lambda_A e^{-\lambda_A t}(1 - e^{-\lambda_A t})^{n-1}(1 - (1 - e^{-\lambda_B t})^n)\mathrm{d}t$$

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 unbekannten an einem Tag hat die Dichte

$$f(x) = \int_0^\infty \lambda e^{-\lambda\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}\mu$$

$$= \lambda e^{-\lambda x + \frac{1}{2}\lambda^2\sigma^2}(1 - \Phi(\frac{\lambda\sigma^2 - x}{\sigma}))$$

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

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

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

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

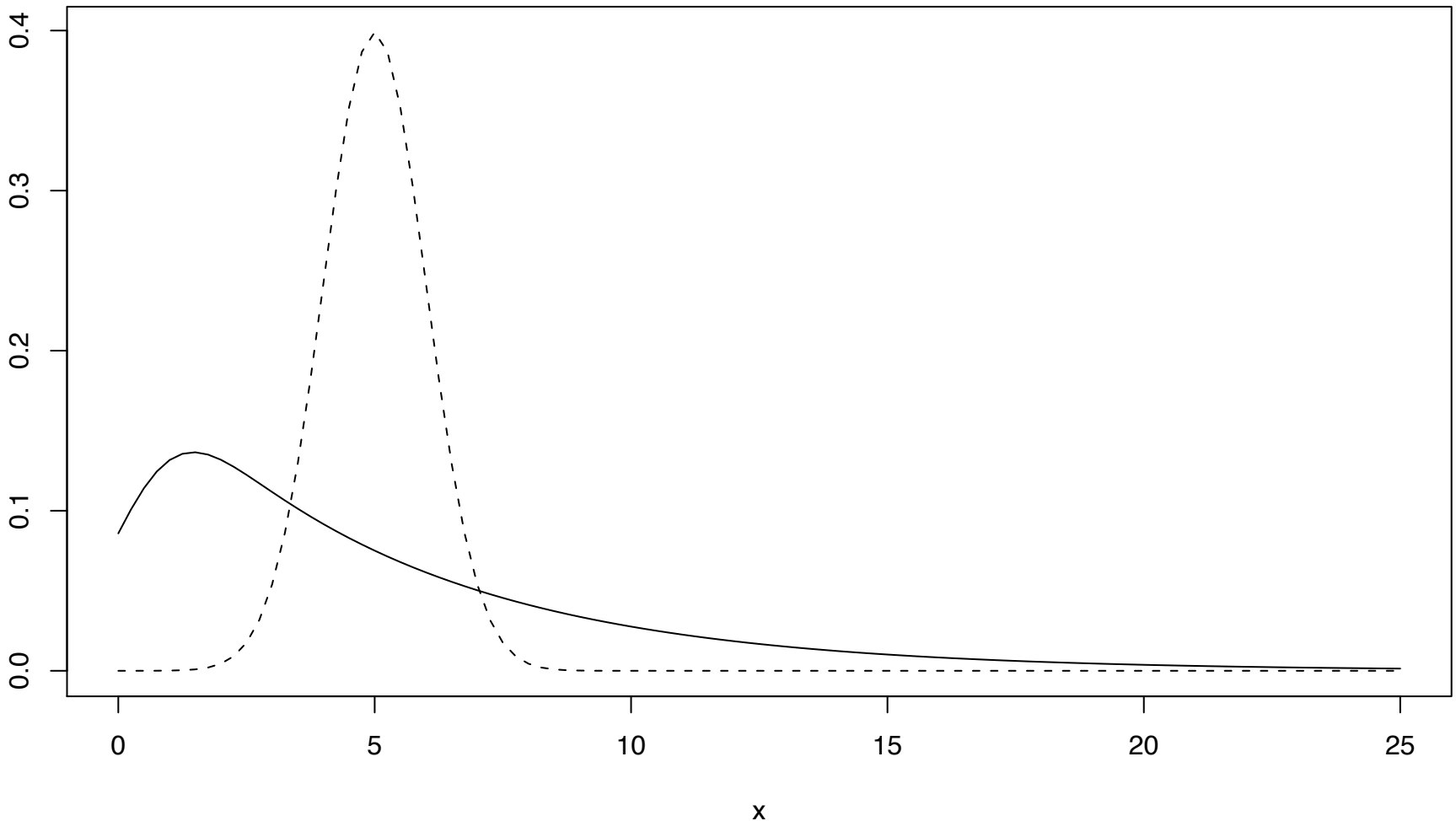
eine Normalverteilung $N(x-\lambda\sigma^2,\sigma^2)$, die auf $[0,\infty)$ beschränkt ist.

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 exponential−verteilt)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
x