Angewandte Statistik

Julian Hinz, Timo Adam — Universität Bielefeld

Kapitel 1

Verteilungsschätzung

1.1 Empirische Verteilungsfunktion

Verteilungsschätzung — Überblick

Oft möchte man sich ein Bild von der **Verteilung einer Variablen** machen.

Zwei Fälle sind zu unterscheiden:

- diskrete Variable ⇝ plotte Häufigkeiten der Werte (hier nicht betrachtet)
- stetige Variable ⇝ etwas schwieriger (im Folgenden betrachtet)

Die Verteilung einer stetigen Zufallsvariablen X wird von der Verteilungsfunktion,

$$F(x) = P(X \leq x),$$

oder alternativ der Dichtefunktion,

$$f(x), \text{sodass } P(a \leq X \leq b) = \int_a^b f(x)dx,$$

eindeutig beschrieben sind.

Im Folgenden diskutieren wir, wie $F(x)$ bzw. $f(x)$ basierend auf einer i.i.d.[1]-Stichprobe $X_1, \ldots, X_n$ geschätzt und visualisiert werden können.

1. i.i.d. = “independent and identically distributed”

In diesem Abschnitt betrachtetes Beispiel: Daten zum 50. Hermannslauf[1]

Platz	Vorname	Nachname	Zeit (in Minuten)
1	Elias	Sansar	109.1
2	Patrick	Boehme	110.4
3	Niklas	Hänze	110.8
4	Erik	Peters	113.4
5	Markus	Scheller	114.7
6	Jan	Kaschura	116.2
….	…	…	…

1. https://www.nw.de/sport/hermannslauf/ergebnisse

Im Folgenden betrachtet:

X = Zeit in Minuten

Empirische Verteilungsfunktion

Wir führen eine Darstellung ein, welche folgende Frage beantwortet:

Wie viele Datenpunkte sind kleiner oder gleich einem bestimmten Wert x?

⇝ Beispiel beim Hermannslauf: Wie viele Läufer*innen waren schneller als 3 Stunden?

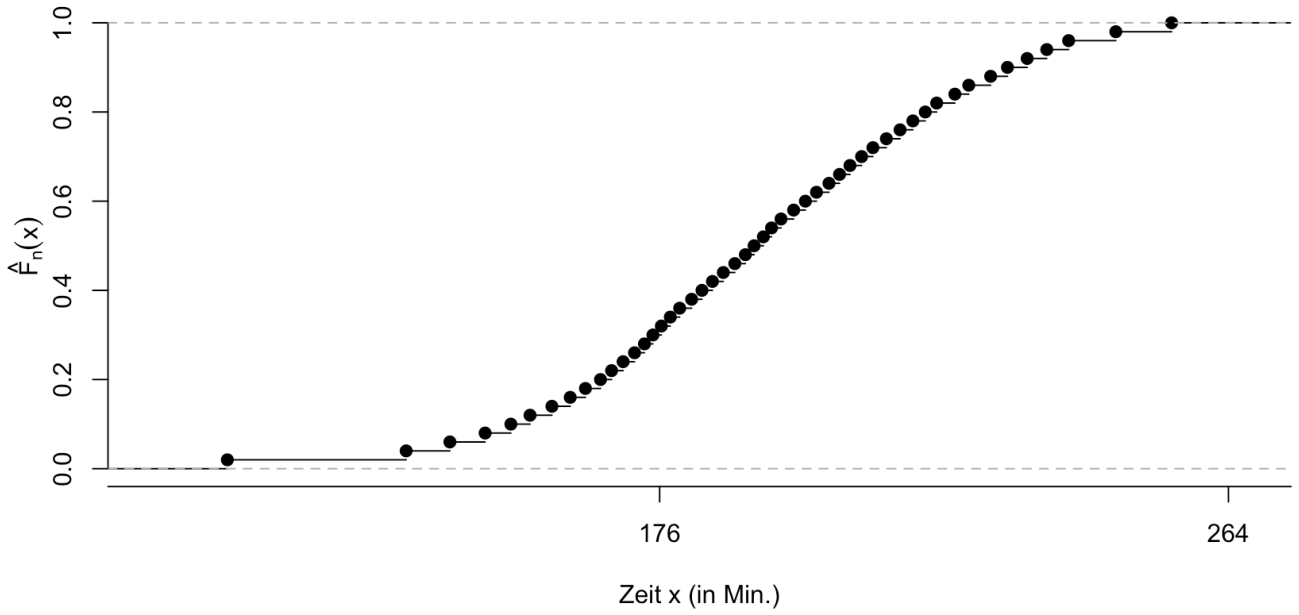
Empirische Verteilungsfunktion

Bemerkungen:

- $\hat{F}_n(x)$ ist einfach der Anteil an Beobachtungen kleiner gleich x
- in R: `ecdf(x)` bzw. `plot(ecdf(x))`
- $\hat{F}_n(x) \overset{n\to\infty}{\to} F(x)$ für alle x

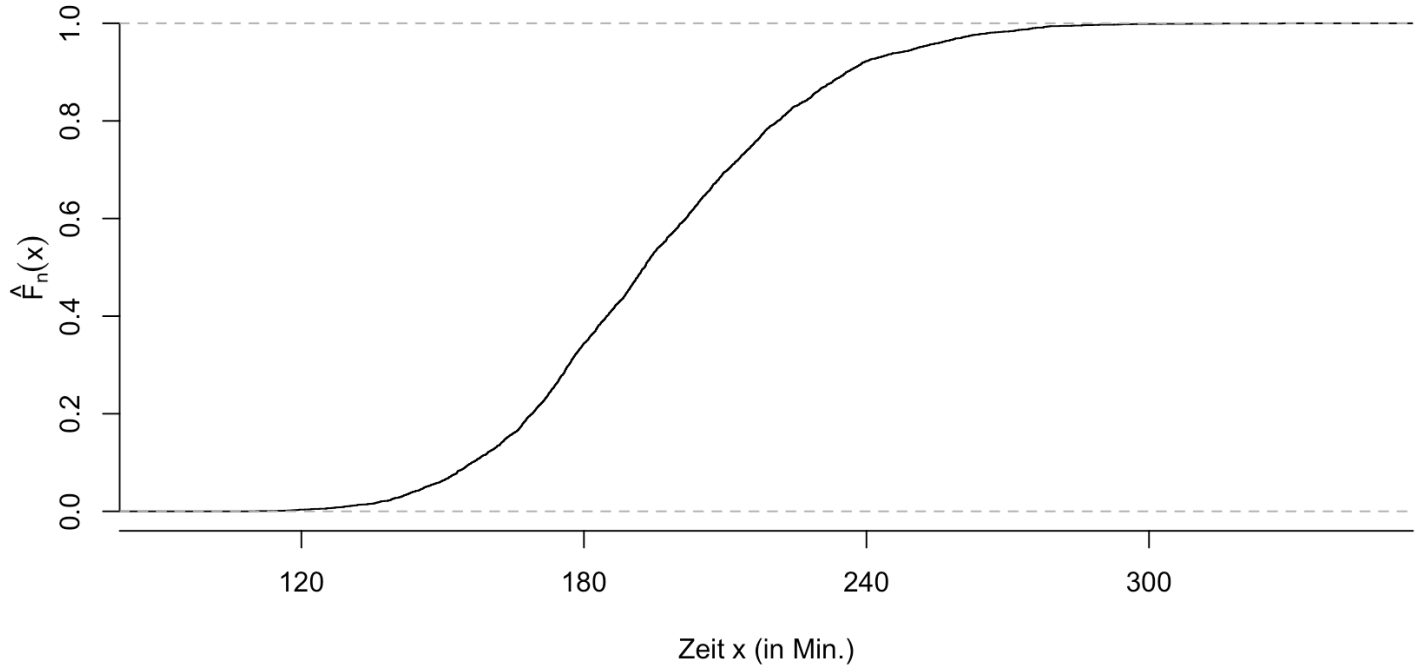
Empirische Verteilungsfunktion im Hermannslauf-Beispiel

Zur besseren Veranschaulichung betrachten wir zunächst nur die Zeiten von 50 Läufer*innen.



Empirische Verteilungsfunktion im Hermannslauf-Beispiel

ECDF für die Zeiten aller 4496 Läufer*innen:



Empirische Verteilungsfunktion im Hermannslauf-Beispiel

Beispiele dazu, was man aus dieser Funktion ablesen kann:

- _____ der Läufer*innen haben höchstens 240 Minuten benötigt
- _____ der Läufer*innen haben mehr als 180 Minuten benötigt
- (ca.) 20% der Läufer*innen haben $\leq$ _____ benötigt
- (ca.) 50% der Läufer*innen haben $>$ _____ benötigt

Verteilungsschätzung bei einer stetigen Variablen — Überblick

- im Prinzip reicht die ECDF, um sich ein umfassendes Bild zu machen
- allerdings sind Verteilungsfunktionen weniger intuitiv als Dichtefunktionen
- daher betrachtet man in der Praxis eher Schätzer der Dichtefunktion
- wir betrachten zunächst noch einmal das Histogramm, welches
  - sehr simpel zu konstruieren ist…
  - allerdings einige unerwünschte Eigenschaften aufweist…
  - welche wir zuletzt mit Hilfe eines Kerndichteschätzers überwinden werden

Kapitel 1

Verteilungsschätzung

1.2 Wdh. Histogramme

Histogramme — grundlegende Idee

Ziel: Erstellung einer sinnvollen Häufigkeitsdarstellung für stetige Merkmale.

Klar ist: Häufigkeiten ergeben nur nach Kategorisierung Sinn.[1]

Wir zerlegen also zunächst den Datenbereich in K Klassen,

$$[c_0, c_1), [c_1, c_2), \ldots, [c_{K-1}, c_K]$$

und betrachten dann Häufigkeiten innerhalb dieser Klassen.

1. sonst möglicherweise Anzahl Ausprägungen $= n$ und damit Häufigkeit $= 1$ für jede Ausprägung.

(Klassierte) Häufigkeitstabelle für die Zeit im Hermannslauf-Beispiel

Ausprägung (Klasse)	absolute Häufigkeit h_k	relative Häufigkeit f_k
$60 \leq \ldots < 120$	16	$16/4496 \approx 0.004$
$120 \leq \ldots < 150$	266	$266/4496 \approx 0.059$
$150 \leq \ldots < 180$	1265	$1265/4496 \approx 0.281$
$180 \leq \ldots < 240$	2601	$2601/4496 \approx 0.579$
$240 \leq \ldots < 300$	342	$342/4496 \approx 0.076$
$300 \leq \ldots < 360$	6	$6/4496 \approx 0.001$
$360 \leq \ldots$	0	$0/4496 = 0$

Das Histogramm wird nun so konstruiert, dass die *Fläche* über den Klassen den absoluten (oder den relativen) Häufigkeiten entspricht.

Konstruktion eines Histogramms

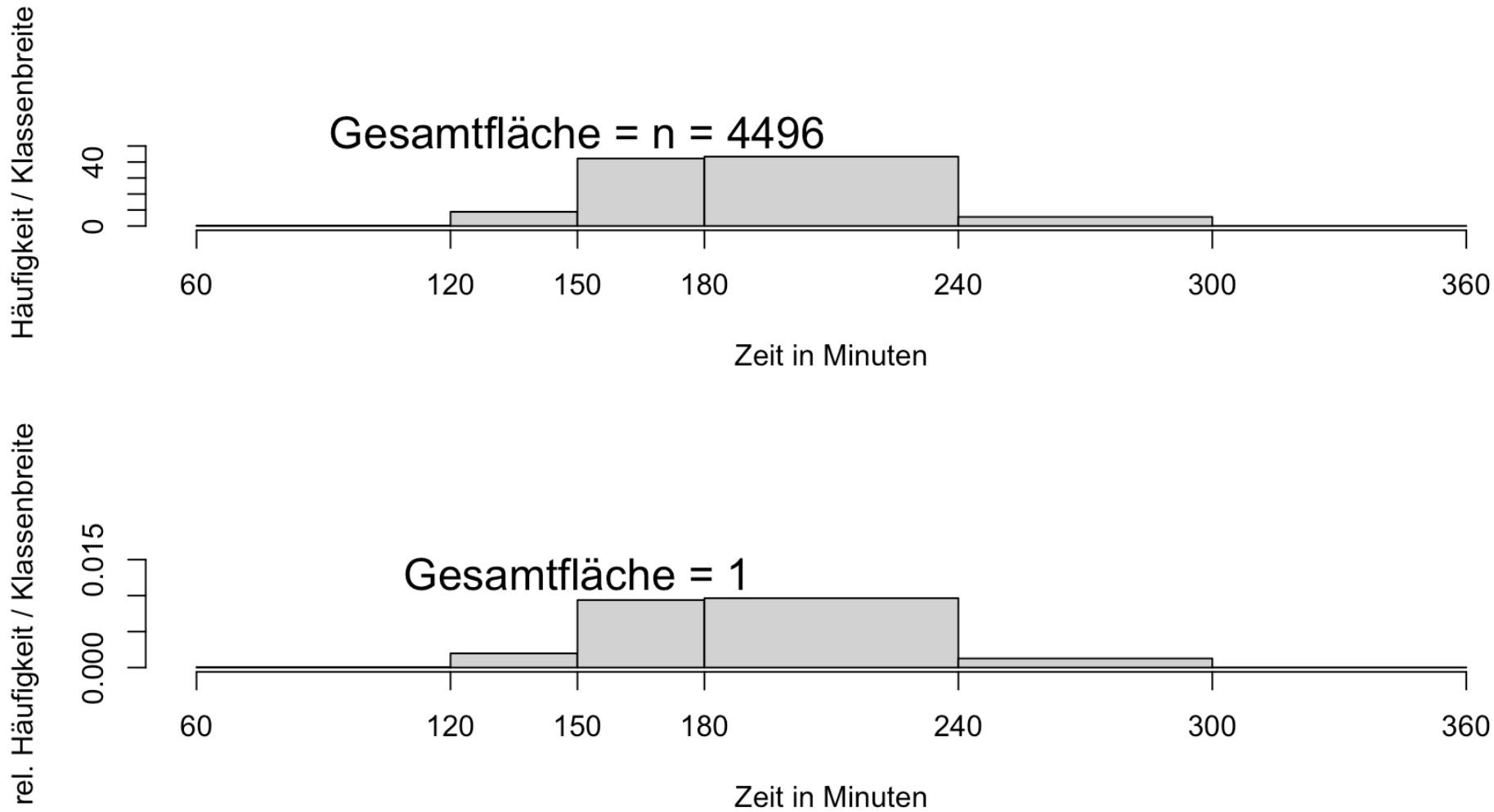
$$\text{Höhe der Säulen} = \text{Häufigkeit/Klassenbreite}$$

$$\text{Höhe der } k\text{-ten Säule} = \begin{cases} h_k/b_k & \text{damit Gesamtfläche} = n \\ f_k/b_k & \text{damit Gesamtfläche} = 1 \end{cases}$$

(b_k ist die Breite des k-ten Intervalls)

Im Beispiel zum Hermannslauf:

Ausprägung (Klasse)	**abs. Häuf** h_k	**rel. Häuf.** f_k	**Klassen-breite** b_k	**Höhe** h_k/b_k	**Höhe** f_k/b_k
$60 \leq \ldots < 120$	16	0.004	60	0.267	0.000067
$120 \leq \ldots < 150$	266	0.059	30	8.867	0.00197
$150 \leq \ldots < 180$	1265	0.281	30	42.12	0.0094
$180 \leq \ldots < 240$	2601	0.579	60	43.35	0.0097
$240 \leq \ldots < 300$	342	0.076	60	5.70	0.0013
$300 \leq \ldots < 360$	6	0.001	60	0.10	0.000017


Häufigkeit / Klassenbreite
0
40
Gesamtfläche = n = 4496
60
120
150
180
240
300
360
Zeit in Minuten
rel. Häufigkeit / Klassenbreite
0.000
0.015
Gesamtfläche = 1
60
120
150
180
240
300
360
Zeit in Minuten

Histogramm mit äquidistanten Klassen

Normalerweise werden die Klassenbreiten $b_k = c_k - c_{k_1}$ gleich groß gewählt. Man spricht dann von **äquidistanten Klassen** (der Breite b).

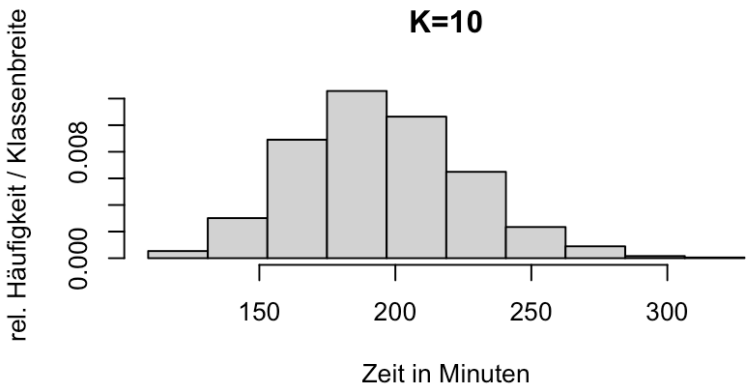
Im Falle relativer Häufigkeiten sind die Höhen der Säulen auf den Klassen dann gegeben durch

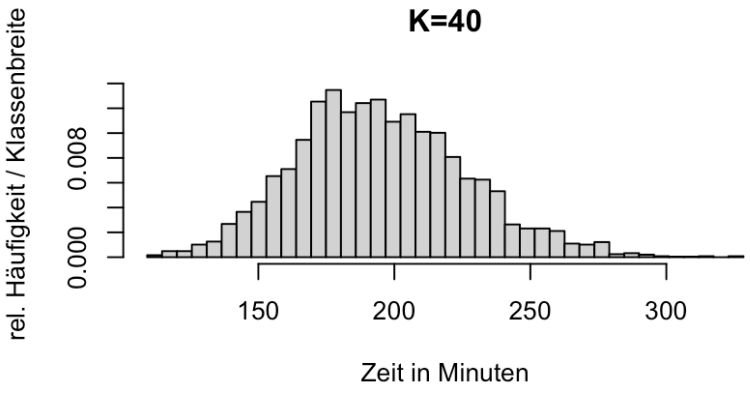
$$\frac{f_k}{b} = \frac{h_k/n}{b} = \frac{h_k}{nb}$$

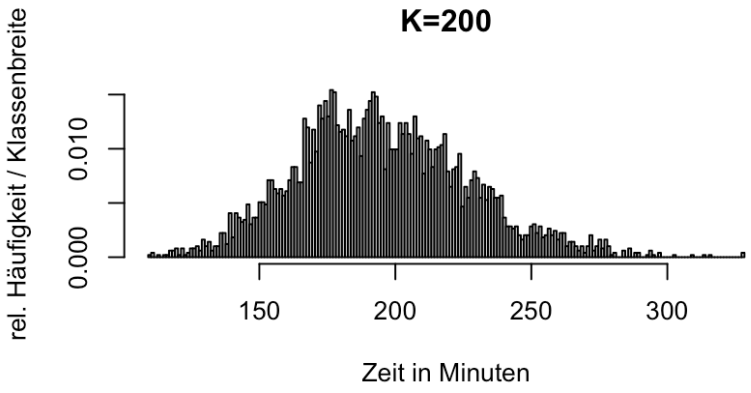
In der Praxis wählt man die Anzahl der Klassen K, Startpunkt c_0 und Endpunkt c_k. Die **Wahl von K hat einen starken Einfluss** auf das Erscheinungsbild.

In R: `hist(x, breaks = ..., prob = TRUE)`

Verschiedene Klassenbreiten im Hermannslauf-Beispiel







Die Wahl von K und der Bias-Varianz-Trade-off

Die Werte $K = 10, 40$ liefern bessere Ergebnisse als $K = 2, 200$ — warum?

- bei sehr großen Klassenbreiten ($K = 2$) geht viel Information verloren — durch Mittelung über großen Bereich liegt man teilweise komplett falsch
- bei sehr kleinen Klassenbreiten ($K = 200$) geht wenig Information verloren, dafür bekommt man sehr unruhige Histogramme

Man spricht von einem Bias-Varianz-Trade-off:

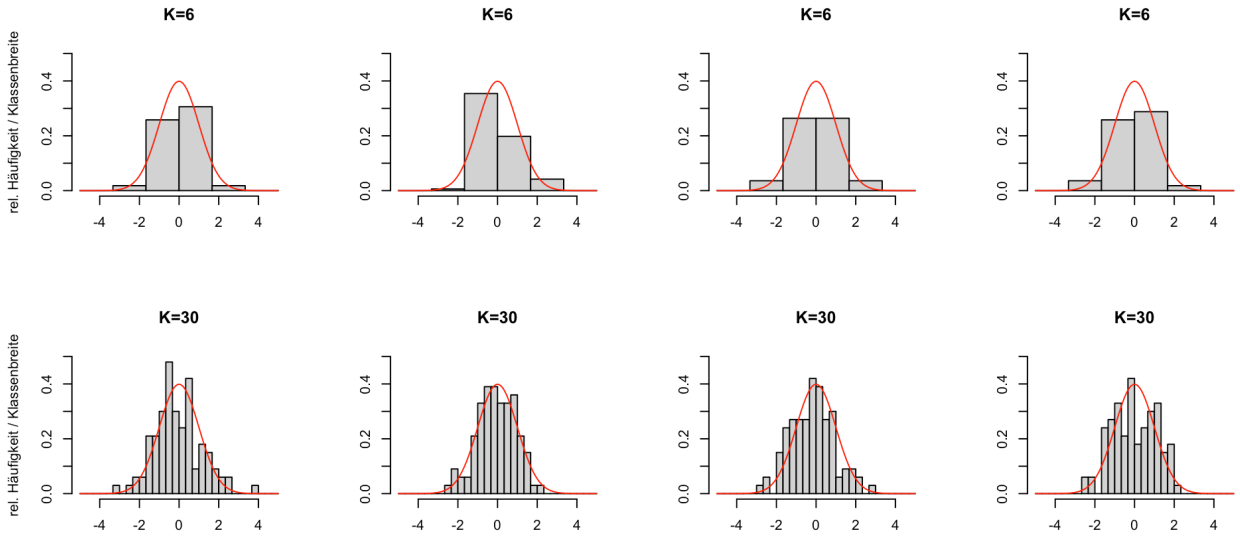
in den Extremen hat man entweder…

- hohe Varianz ($K = 200$)
- oder hoher Bias (d.h. Verzerrung; $K = 2$)

Optimale Lösung (moderate Verzerrung & Variabilität) liegt irgendwo dazwischen.

Simulationsexperiment zur Illustration des Bias-Varianz-Trade-off

- Histogramm von 100 zufällig gezogenen Beobachtungen aus der $N(0, 1)$-Verteilung
- in rot: Dichtefunktion der $N(0, 1)$-Verteilung



Histogramm — Zusammenfassung der Eigenschaften & Ausblick

- bei Nutzung der relativen Häufigkeiten ist die Gesamtfläche des Histogramms gleich 1, d.h. das Histogramm definiert eine Dichte!
- insb. kann das Histogramm als Schätzer $\hat{f}(x)$ der uns interessierenden Dichte $f(x)$ betrachtet werden
- Einfluss von b (bzw. K) auf das Verhalten des Schätzers:
  - b klein (d.h. K groß) $\rightsquigarrow$ Bias klein, Varianz groß
  - b groß (d.h. K klein) $\rightsquigarrow$ Bias groß, Varianz klein
- konstant auf vom Anwender festgelegten Intervallen (restriktiv!)
- nicht stetig (unschön!)

Per Modifikation des Histogramms führen wir den **Kerndichteschätzer** ein, welcher beide Probleme beseitigt.

Kapitel 1

Verteilungsschätzung

1.3 Dynamische Klassengrenzen

Dynamische Klassengrenzen

Das Histogramm (mit Gesamtfläche = 1 bzw. Häufigkeiten h_k/nb) können wir schreiben als:

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathcal{I}(x_i \text{ und } x \text{ im gleichen Intervall})$$

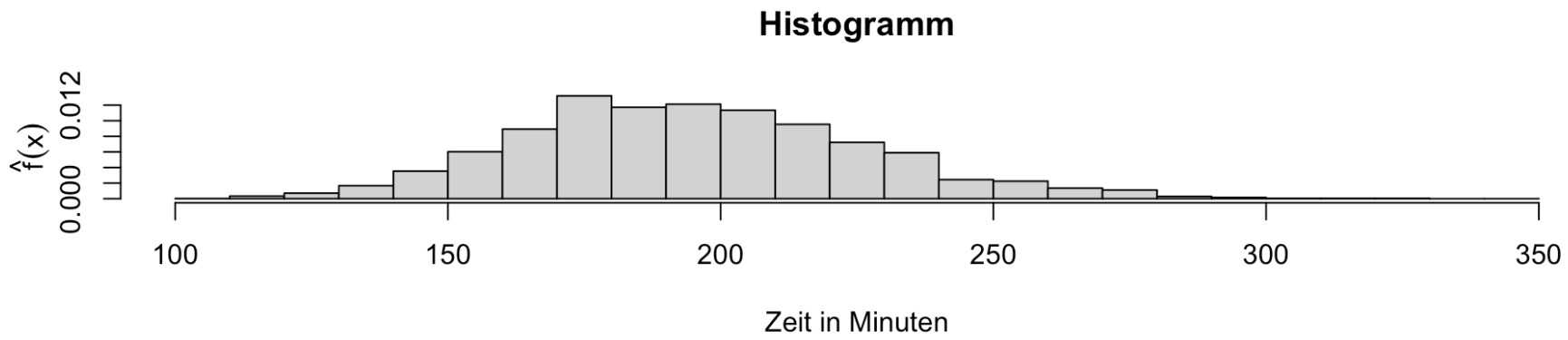
Hier sind die Intervalle *fest* und der Dichteschätzer konstant auf diesen.

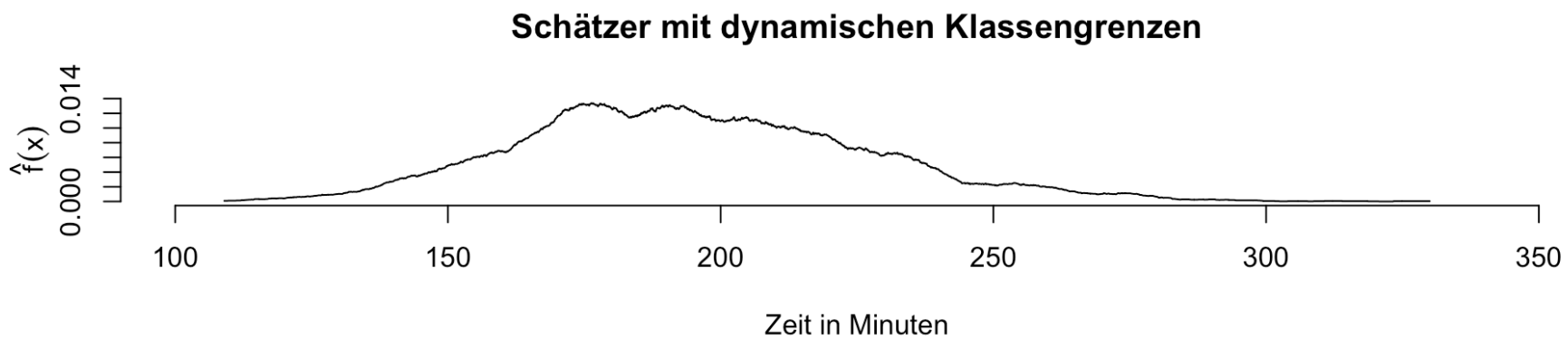
Betrachte stattdessen sich *mit x bewegende Intervalle*:

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathcal{I}(x_i \in [x - \frac{b}{2}, x + \frac{b}{2}])$$

Der resultierende Dichteschätzer ist nicht konstant auf vorher festgelegten Intervallen, da wir **dynamische Klassengrenzen** betrachten.

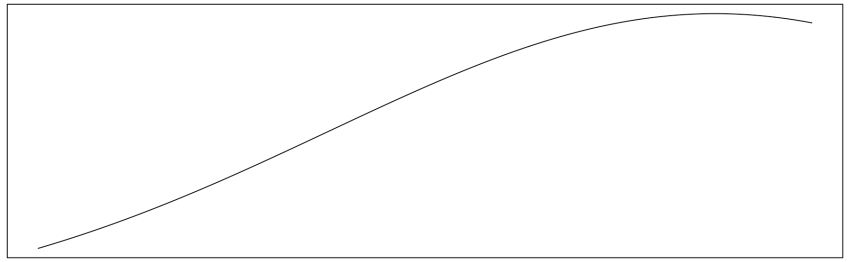
Dynamische Klassengrenzen im Hermannslauf-Beispiel

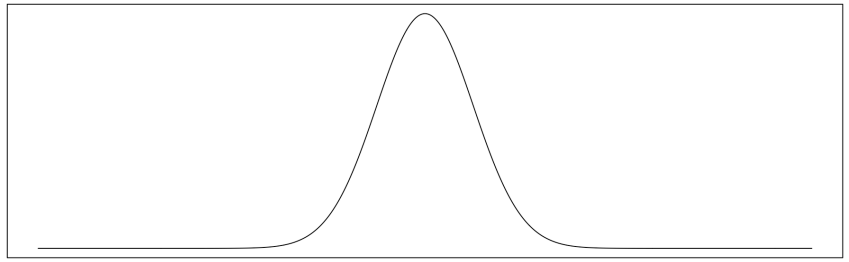


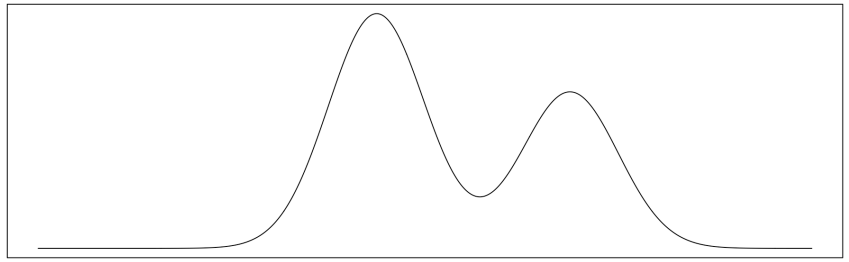


Erwartungswert des Schätzers mit dynamischen Klassengrenzen: (Vorlesung)

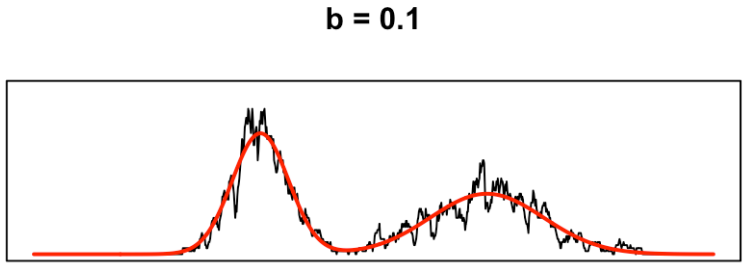
$$E\left(\hat{f}(x)\right) =$$

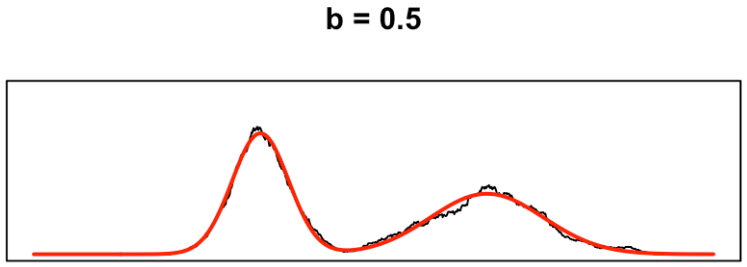


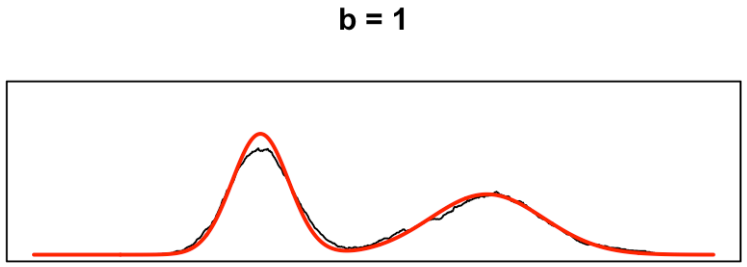


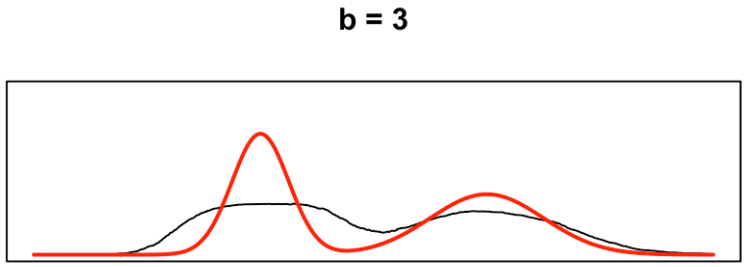


Simulationsexperiment: wahre Dichtefunktion und Dichteschätzer basierend auf 1000 zufällig generierten Beobachtungen, für verschiedene Werte von b.









Schätzer mit dynamischen Klassengrenzen — Eigenschaften

- Bias verschwindet bei $b \to 0$:

$$E\Big(\hat{f}(x)\Big) = \frac{F(x + \frac{b}{2}) - F(x - \frac{b}{2})}{b} \overset{b \to 0}{\longrightarrow} F'(x) = f(x)$$

- aber: je kleiner b, desto weniger Daten pro Klasse, d.h. hohe Varianz
- umgekehrt erhalten wir bei hohem b einen Bias — äußert sich durch "Überschätzung von Tälern" & "Unterschätzung von Gipfeln"
- Einfluss von b auf Verhalten des Schätzers daher wie zuvor:
  - b klein ⇝ Bias klein, Varianz groß
  - b groß ⇝ Bias groß, Varianz klein
- keine festen Intervalle mehr, aber **immer noch nicht stetig** (unschön!)