

# Kapitel 9

## Poissonverteilung und Binomialverteilung

### 9.1 Allgemeines über Zufallsvariable mit Wertebereich in $\mathbb{Z}$

In medizinischen Untersuchungen **zählt** man unter dem Mikroskop **die Anzahl  $k$  bestimmter Zellen** oder Partikel in Blut/Urin/Lymphe/Zellgewebe **pro Flächeneinheit**, in biologischen Untersuchungen die Anzahl  $k$  von Schädlingen/seltenen Spezies pro Flächeneinheit, in chemisch-physikalischen Untersuchungen die Anzahl  $k$  gewisser Partikel (z.B. Asbest) in der Luft oder im Wasser **pro Raumeinheit** oder die Häufigkeit  $k$  eines gewissen Ereignisses **pro Zeiteinheit** (z.B. das Aussenden eines Heliumions durch ein radioaktives Präparat), in Medizin, Pharmazie, Biologie, Soziologie zählt man das Vorkommen von gewissen Individuen/Ereignissen **pro Population** oder in einer repräsentativen Auswahl dieser Population ...

Das Ergebnis der Untersuchung ist jedesmal die konkret gefundene Anzahl  $k$  der gesuchten Phänomene, aber

- entweder bezogen auf ein physikalisches Kontinuum, dessen Größe durch ein **Flächen-, Raum- oder Zeitmaß** angegeben wird,
- oder bezogen auf eine Population oder eine sonstige Gesamtheit, deren Größe durch die **Anzahl  $n$  ihrer Elemente** angegeben wird.

Wiederholt man die Auszählung an einer anderen Probe der untersuchten Fläche/des untersuchten Raums/ zu einem anderen Zeitpunkt, oder an einer anderen Auswahl aus der Population, so wird man in der Regel nicht exakt denselben Zahlwert  $k$  als Ergebnis bekommen, sondern nur einen ähnlichen, und die Zählergebnisse werden um einen häufigsten Wert herum schwanken, d.h.

$k$  ist eine Zufallsvariable, die allerdings (im Unterschied zu den normalverteilten Zufallsvariablen  $x$ , welche innerhalb  $\mathbb{R}$  variieren,) nur ganzzahlige Werte  $\geq 0$  annehmen kann.

Die Auszählung selber ist ein Experiment zur Messung von  $k$ , und wenn man dasselbe Experiment  $n$  mal durchführt, so erhält man eine Messreihe mit Messwerten  $k_1, k_2, \dots, k_n$  und einem Mittelwert  $\bar{k}$ , der nach dem 1. Grenzwertsatz<sup>1</sup> für  $n \rightarrow \infty$  gegen einen Grenzwert  $\hat{k}$ , den Erwartungswert des Experiments strebt.

**Schreibkonvention:** Während man reellwertige Zufallsvariable üblicherweise mit  $x$  und ihren Erwartungswert mit  $\mu$  bezeichnet, bezeichnet man **ganzzahlige** Zufallsvariable, wie sie bei Auszählungen auftreten und die daher nur nichtnegative Werte  $\in \mathbb{Z}$  annehmen können, üblicherweise mit  $k$ , ihren Erwartungswert mit  $\lambda$ .

Bei der Feinauswertung der Messreihe  $k_1, \dots, k_n$  beginnt man wieder mit der Strichliste, welche die absoluten Häufigkeiten  $H_j$  zählt, erhält als 1. Standardisierung ein Säulendiagramm, welches die relativen Häufigkeiten  $h_j = \frac{J_j}{n}$  darstellt, und macht sich im Rahmen der 2. Standardisierung dadurch von unterschiedlichen Säulenbreiten unabhängig, dass man die für die ganzzahlige Zufallsvariable  $k$  kleinste sinnvolle Säulenbreite, nämlich die Breite 1, wählt<sup>2</sup>, und als Mittelpunkt dieser Achsenabschnitte jedesmal eine ganze Zahl.

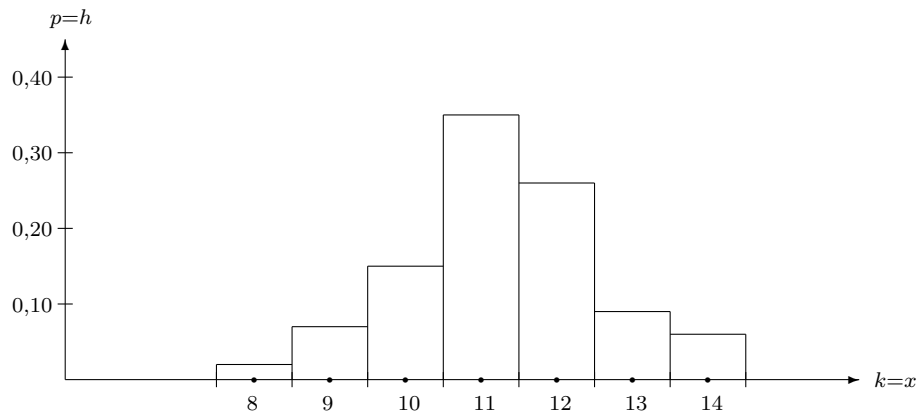


Abbildung 9.1: Graph einer Dichtefunktion bei ganzzahliger Zufallsvariabler  $k$

Nach (7.1) galt ganz allgemein<sup>3</sup>  $h_j \approx p_j \cdot \Delta x$  für kleine  $\Delta x$ .

Daraus folgt hier wegen  $\Delta x = \Delta k = 1$ , dass  $h_j \approx p_j$  ist für alle  $j$ , wobei die Näherungsformel für  $n \rightarrow \infty$  gegen Gleichheit strebt, d.h.:

**Graph der Dichtefunktion:**

Der Graph der Dichtefunktion  $z = p(x)$  zu einer Messreihe  $k_1, \dots, k_n$  einer ganzzahligen Zufallsvariablen ist gleich der Silhouette eines Säulendiagramms der Säulenbreite 1.

Da jetzt in jedem Abschnitt der waagerechten Achse genau eine ganze Zahl liegt und da  $h$  die relative Häufigkeit misst, mit der das Messergebnis in diesen Abschnitt fällt, bedeutet das inhaltlich:

<sup>1</sup>siehe 7.2, S.149

<sup>2</sup>Da ja Zwischenwerte zwischen den ganzen Zahlen bei der Auszählung gar nicht vorkommen, ist eine feinere Säuleneinteilung sinnlos.

<sup>3</sup>siehe 7.3, S.151

**Bedeutung der Dichtefunktion:**

Ist  $y = p(x)$  die empirische **Dichtefunktion** zu einer Messreihe  $k_1, \dots, k_n$  einer ganzzahligen Zufallsvariablen, so gibt  $p(k)$  für  $k \in \mathbb{Z}$  die relative Häufigkeit an, mit der die Messwerte der Messreihe den Zahlwert  $k$  haben.

Ist  $y = F(k)$  die Verteilungsfunktion zur Messreihe, so gibt diese per Definition<sup>4</sup> die relative Häufigkeit an, mit der die Messwerte einer Messreihe einen Zahlwert  $\leq k$  haben.

Daraus folgt unmittelbar: Für  $a \in \mathbb{Z}$  gilt

$$F(a) = \begin{cases} 0 & \text{falls } a < 0, \\ \sum_{k=0}^a p(k) & \text{falls } a \geq 0. \end{cases}$$

Da nur ganzzahlige Messwerte vorkommen, ist  $F(a) = F(a+0,5)$ . Betrachtet man den stufenförmigen Graphen von  $z = p(x)$  (Abb.9.1, S.178), so erkennt man: Für alle  $a \in \mathbb{Z}$  gilt

$$F(a+0,5) = F(a) = \sum_{k=0}^a p(k) = \int_{-\infty}^{a+0,5} p(k)dk$$

wie die ganz allgemein formulierte *Regel 61*<sup>5</sup> es verlangt.

Sind  $a$  und  $b$  ganzzahlig, so gilt: Die relative Häufigkeit, mit der die Messwerte der Messreihe einen Zahlwert  $a \leq k \leq b$  haben, ist gleich

$$\sum_{k=a}^b p(k) = F(b) - F(a-1).$$

Die Integralschreibweise hierzu erhält man wie folgt:  $F(b) - F(a-1) = F(b+0,5) - F(a-0,5) = \int_{a-0,5}^{b+0,5} p(k)dk$ .

Nach dem 2. Grenzwertsatz<sup>6</sup> streben Dichtefunktion und Verteilungsfunktion der empirischen Messreihe für  $n \rightarrow \infty$  gegen die Dichtefunktion bzw. Verteilungsfunktion zum Experiment, und das Experiment besteht bei ganzzahliger Zufallsvariabler in einer Auszählung nach einer bestimmten Methode. Somit gilt

**Regel 72 (Berechnung von Wahrscheinlichkeiten bei ganzzahliger Zufallsvariabler  $k \in \mathbb{Z}$ ):**

**R 72**

Ist  $k$  eine ganzzahlige Zufallsvariable und sind  $z = p(x)$  bzw.  $y = F(x)$  die Dichtefunktion bzw. Verteilungsfunktion zu einem Experiment, so gilt

$$F(a) = \sum_{k=0}^a p(k) \quad \text{für } a \geq 0, \quad F(a) = 0 \quad \text{für } a < 0.$$

Die Wahrscheinlichkeit, dass bei einer einzelnen Auszählung die Zufallsvariable  $k$  den Wert  $a$  bzw einen Wert im Bereich  $a \leq k \leq b$  annimmt, berechnet sich wie folgt :

$$P(k = a) = p(a),$$

$$P(a \leq k \leq b) = \sum_{k=a}^b p(k) \quad \text{oder auch} \quad P(a \leq k \leq b) = F(b) - F(a-1).$$

In Anlehnung an die Formulierung von *Regel 64*<sup>7</sup> läßt sich sagen:

$$P(a \leq k \leq b) = \int_{a-0,5}^{b+0,5} p(k)dk = \text{Fläche unter dem Graph der Dichtefunktion im Abschnitt } a-0,5 \leq k \leq b+0,5.$$

<sup>4</sup>siehe 7.3, S.152

<sup>5</sup>siehe 7.3, S.153

<sup>6</sup>siehe 7.3, S.153

<sup>7</sup>siehe 7.3, S.154

Die Dichtefunktionen und zugehörigen Verteilungsfunktionen der meisten ganzzahligen Zufallsvariablen  $k$  gehören einer von bloß zwei unterschiedlichen Klassen von Funktionen an, die im Folgenden vorgestellt werden.

Um welche der beiden Klassen es sich im Einzelfall handelt, hängt dabei nicht von der konkreten Zufallsvariablen  $k$  ab, sondern bloß davon, ob die mit  $k$  betriebene Auszählung sich auf ein physikalisches Kontinuum bezieht, dessen Größe durch ein Flächen-, Raum- oder Zeitmaß beschrieben wird, oder aber auf eine Population oder eine Gesamtmenge von Ereignissen, deren Größe durch die Anzahl  $n$  ihrer Elemente beschrieben wird.<sup>8</sup>

## 9.2 Poissonverteilung

**Bezeichnung:** *Bezieht sich die Auszählung auf eine Flächen-, Raum- oder Zeiteinheit und ist der Gegenstand der Auszählung in Fläche, Raum und/oder Zeit ungefähr gleichverteilt, so heißt die Zufallsvariable  $k$  poissonverteilt<sup>9</sup>, die Dichtefunktion zum Experiment heißt Dichtefunktion zur Poissonverteilung.*

**Regel 73 (Streuung und Dichtefunktion zur Poissonverteilung):**

**R 73**

*Ist  $k$  eine poissonverteilte Zufallsvariable mit Erwartungswert  $\lambda$ , so besitzt  $k$  die Streuung*

$$\sigma = \sqrt{\lambda}$$

*und die Dichtefunktion*

$$p_{\lambda}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

$p_{\lambda}(k)$  ( $k = 0, 1, 2, \dots$ ) ist für vielbenutzte Werte von  $\lambda$  tabelliert<sup>10</sup> worden, aber wenn man nur wenige Werte benötigt, geht es heute mindestens ebenso schnell mit dem Taschenrechner.

**Merke:** Poissonverteilung gehört zum Experiment **“Zählung der Vorkommnisse eines Ereignisses  $E$  pro Flächen- Raum- oder Zeiteinheit,“** wobei der Erwartungswert für die Anzahl gleich  $\lambda$  ist.

Aus den allgemeinen Resultaten des vorigen Paragraphen ergibt sich sofort

**Regel 74 (Berechnung von Wahrscheinlichkeiten bei Poissonverteilung):**

**R 74**

*Die Dichtefunktion  $p_{\lambda}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$  gibt für jeden festen ganzzahligen Wert  $k \geq 0$  die Wahrscheinlichkeit an, daß bei Poissonverteilung eine einzelne konkrete Auszählung die Anzahl  $k$  liefert, während der Erwartungswert  $= \lambda$  ist.*

*Die Wahrscheinlichkeit  $P(a \leq k \leq b)$ , daß bei Poissonverteilung mit Erwartungswert  $\lambda$  eine einzelne Auszählung eine Anzahl  $a \leq k \leq b$  liefert, ist gleich der Summe der Einzelwahrscheinlichkeiten, d.h.*

$$P(a \leq k \leq b) = \sum_{k=a}^b p_{\lambda}(k) = \left( \sum_{k=a}^b \frac{\lambda^k}{k!} \right) \cdot e^{-\lambda},$$

<sup>8</sup>siehe oben, S.177

<sup>9</sup>Simeon Denis Poisson (1781 - 1840), französischer Mathematiker und Physiker

<sup>10</sup>siehe Tabellen zur Statistik, S.226

$$\begin{aligned} \text{ausf\u00fchrlich:} \quad P(a \leq k \leq b) &= \left( \frac{\lambda^a}{a!} + \frac{\lambda^{a+1}}{(a+1)!} + \dots + \frac{\lambda^b}{b!} \right) \cdot e^{-\lambda}, \\ \text{sowie} \quad P(k \leq b) &= P(0 \leq k \leq b) \quad \text{und} \quad P(k \geq a) = 1 - P(0 \leq k \leq a-1). \end{aligned}$$

### 9.2.1 Rechenbeispiele

**Beispiel 1: Die Berechnung von Wahrscheinlichkeiten bei bekanntem Erwartungswert  $\lambda$  und mit selbstberechneten Werten  $p_\lambda(k)$ :**

Auf einer Fläche von  $5\text{cm}^2$  seien tausend Bakterien einer bestimmten Art, in etwa gleichverteilt, vorhanden.

Frage 1: Wie groß ist die Wahrscheinlichkeit, auf einer Teilfläche von  $5\text{mm}^2$  genau 6 Bakterien zu finden?

Hier ist die Zufallsvariable  $k = \text{Anzahl der Bakterien auf Teilflächen à } 5\text{mm}^2$ .

Lösung:  $5\text{mm}^2$  ist ein Hundertstel von  $5\text{cm}^2$ . Der Erwartungswert von  $k$  errechnet sich also zu  $\lambda = \frac{1000}{100} = 10$ . Die gefragte Wahrscheinlichkeit ist deshalb

$$P(k = 6) = p_\lambda(6) = p_{10}(6) = \frac{10^6}{6!} \cdot e^{-10} \approx 0,063 = 6,3\%.$$

Frage 2: Wie groß ist die Wahrscheinlichkeit, auf irgendeiner Teilfläche von  $5\text{mm}^2$  höchstens 6 Bakterien anzutreffen?

Lösung: Diese Wahrscheinlichkeit ist gleich

$$\begin{aligned} P(0 \leq k \leq 6) &= p_{10}(0) + p_{10}(1) + p_{10}(2) + p_{10}(3) + p_{10}(4) + p_{10}(5) + p_{10}(6) \\ &= \left( \frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \dots + \frac{10^6}{6!} \right) \cdot e^{-10} \\ &= \left( \frac{1}{1} + \frac{10}{1} + \frac{100}{2} + \frac{1000}{6} + \frac{10000}{24} + \frac{100000}{120} + \frac{1000000}{720} \right) \cdot e^{-10} \\ &\approx 2866,556 \cdot e^{-10} \\ &\approx 0,130 \\ &= 13,0\% \end{aligned}$$

**Beispiel 2: Eine Methode zur ungefähren Bestimmung des Erwartungswertes  $\lambda$ :**

Gegeben sei eine Bakteriensuspension. Die Bakterien seien ungefähr gleichverteilt in der Flüssigkeit, Anzahl unbekannt.

Frage: Wie groß ist die durchschnittliche Anzahl von Bakterien in 1 ml Suspension?

Hier ist die Zufallsvariable  $k = \text{Anzahl von Bakterien pro 1 ml Suspension}$ . **Gefragt ist nach dem Erwartungswert  $\lambda$ .**

Den Erwartungswert einer Zufallsvariablen  $k$  kann man nur schätzen, wenn man eine längere Messreihe ( $n \geq 100$ ) durchgeführt hat. Man braucht also mindestens 100 Proben à 1 ml Suspension.

Die allgemeine Theorie sagt, dass dann der Mittelwert  $\bar{k}$  von  $k_1, \dots, k_n$  als Näherungswert für  $\lambda$  dienen kann. Um diesen Mittelwert berechnen zu können, müsste man also alle Proben vollständig auszählen, um die Anzahlen  $k_1, \dots, k_n$  zu gewinnen, eine äußerst mühselige Zählarbeit.

Sehr viel weniger Zählarbeit erfordert folgendes geschickte Verfahren:

100 Proben à 1 ml Suspension werden auf 100 Nährböden gegeben. Nach angemessener Zeit zählt man lediglich, wie viele der 100 Böden steril geblieben sind (d.h. dass die entsprechende Impfflüssigkeit bakterienfrei war). Man erhalte z.B. 13 sterile Böden. Daraus errechnet man  $\lambda$  wie folgt:

Bei  $n = 100$  Proben enthielten 13 Proben  $k = 0$  Bakterien. Die Messreihe  $k_1, \dots, k_{100}$  ist zahlenmäßig nicht bekannt. Aber man weiss, dass die relative Häufigkeit, mit der in der Messreihe der Wert  $k = 0$  auftrat,  $= \frac{13}{100} = 0,13$  war. Dies ist aber auch der Wert der empirischen Dichtefunktion<sup>11</sup> für  $k = 0$ . Weil  $n$  relativ groß war, ist diese Dichtefunktion schon ungefähr gleich der Poissonschen Dichtefunktion, d.h.

$$0,13 \approx p_\lambda(0) \quad | \text{ mit noch unbekanntem } \lambda$$

$$0,13 = \frac{\lambda^0}{0!} \cdot e^{-\lambda} \quad | \lambda^0 = 1, 0! = 1 \text{ einsetzen}$$

$$0,13 = e^{-\lambda} \quad | \ln$$

$$\ln 0,13 = -\lambda$$

$$-2,04 \approx -\lambda$$

d.h. die mittlere Anzahl von Bakterien in 1 ml Suspension = Erwartungswert  $\lambda \approx 2,04$ .<sup>12</sup>

**Merke:** Bei poissonverteilter Zufallsvariabler  $k$  kann man den Erwartungswert  $\lambda$  näherungsweise berechnen, indem man mittels einer längeren Messreihe den Wert von  $P(k = 0) = p_\lambda(0)$  schätzt.

**Beispiel 3: Interpolation bei der Benutzung der Wertetabelle<sup>13</sup> für  $p_\lambda(k)$ :**

Der Erwartungswert  $\lambda$  hängt noch von der gewählten räumlichen oder zeitlichen Bezugseinheit ab, ist proportional dazu (siehe Beispiel 1, wo  $\lambda$  sich von 1000 auf 10 verkleinert, wenn man nicht  $\text{cm}^2$  auszählt, sondern  $\text{mm}^2$ ). Daher kann  $\lambda$  sehr wohl eine gebrochene Zahl werden, obwohl  $k$  selbst nur ganzzahlige Werte annimmt (siehe Beispiel 2,  $\lambda = 2,04$ ). Benötigt man z.B. den Zahlwert von  $p_\lambda(1)$  für  $\lambda = 2,04$ , so liefert die Tabelle nur den Wert für  $k = 1$  und  $\lambda = 2$ :  $p_2(1) = 0,270671$ .

Tabellenwert:  $p_2(1) \approx 27,07\%$

Interpolation liefert den verbesserten Wert

$$\begin{aligned} p_{2,04}(1) &\approx p_2(1) + \frac{p_3(1) - p_2(1)}{3 - 2} \cdot (2,04 - 2) \\ &= p_2(1) + (p_3(1) - p_2(1)) \cdot 0,04 \\ &= 0,265819 \end{aligned}$$

Interpolierter Wert:  $p_{2,04}(1) \approx 26,58\%$

Den genauesten Wert erhält man allerdings bei Benutzung der Berechnungsformel für  $p_\lambda(k)$ :

$$p_{2,04}(1) = \frac{2,04^1}{1!} \cdot e^{-2,04} = 2,04 \cdot e^{-2,04} = 0,265259.$$

<sup>11</sup>siehe 9.1, S.179

<sup>12</sup>Zur Abschätzung des absoluten Fehlers  $|\Delta\lambda|$ , der bei diesem Verfahren entsteht, siehe 9.5, S.194.

<sup>13</sup>siehe Tabellen zur Statistik, S.226

Formelwert:  $p_{2,04}(1) \approx 26,53\%$

Auf ganze Prozente gerundet sind indes alle Ergebnisse gleich 27%.

## 9.3 Binomialverteilung

### 9.3.1 Zufallsvariable mit Befund positiv/negativ

Besonders in der Medizin (bei Diagnosen) treten häufig Zufallsvariable  $x$  auf, **die keine Zahlen als Wertebereich haben**, sondern bei Messungen nur **zwei mögliche Ergebnisse** liefern, nämlich ja/nein, Befund positiv/negativ. Ein Experiment besteht wieder darin, mit einer festgelegten Methode diesen Befund zu erheben.

Hier ist eine Messreihe zur Variablen  $x$  mit  $n$  Messwerten schon dadurch vollständig ausgewertet, dass man die **Anzahl  $k$  der positiven Befunde** unter  $n$  Befunden **auszählt** (= die Häufigkeit). Ist  $h_n$  die relative Häufigkeit des positiven Befundes, also

$$h_n = \frac{k}{n}, \quad (9.1)$$

so strebt  $h_n$  für  $n \rightarrow \infty$  gegen einen **Grenzwert  $p$** .

**$p$  = Wahrscheinlichkeit des positiven Befundes für  $x$ .** ( $0 \leq p \leq 1$ )

**$1 - p$  = Wahrscheinlichkeit des negativen Befundes für  $x$ .**

Die Werte  $p$  und  $1 - p$  sind die einzigen Funktionswerte der Dichtefunktion zum Experiment.

#### **Problem der Unabhängigkeit der Messergebnisse:**

Bei solchen Zufallsvariablen  $x$  mit Befund positiv/negativ kann eine große Schwierigkeit auftreten: Bei einer Messreihe zu  $x$  kommt es ja immer darauf an, die Messung wiederholt unter den gleichen Rahmenbedingungen durchzuführen. In der Medizin kann es aber leicht passieren, dass der erste Messbefund die Rahmenbedingungen für eine später durchgeführte 2. Messung verändert.

Bei Erb- und Infektionskrankheiten und allen Krankheitsbefunden mit Folgeschäden ist dies z.B. der Fall: Untersucht man 20 Schüler einer Klasse auf Sehschwäche ( $x$  = Sehschwäche ja/nein), so sind die Messbefunde sicherlich unabhängig voneinander. Untersucht man dieselben 20 Schüler auf Windpocken ( $x$  = Windpocken ja/nein) und findet man am Montag einen einzigen Erkrankten unter 20, so sind die Rahmenbedingungen für Windpockenuntersuchungen derselben Klasse am Freitag völlig verändert: Die Wahrscheinlichkeit  $p$  für den positiven Befund von  $x$  ist zwischen Montag und Freitag stark gestiegen. Da eine Zufallsvariable aber immer nur eine Wahrscheinlichkeit  $p$  besitzt, welche für sie eine spezifische Konstante ist, werden in dieser Klasse am Montag und am Freitag de facto zwei verschiedene Zufallsvariable gemessen: Am Montag handelte es sich bei  $x$  um die Zufallsvariable "Vorkommen von Windpocken in einem gesunden Milieu", am Freitag um eine ganz andere Zufallsvariable, nämlich das "Vorkommen von Windpocken in einem (mehr oder weniger) durchseuchten Milieu".

### 9.3.2 Theorie der Binomialverteilung

**Merke:** Im folgenden sei immer unterstellt, dass die Ergebnisse wiederholter Messungen der Zufallsvariablen  $x$  mit Befund positiv/negativ sich nicht untereinander beeinflussen können, sondern garantiert **unabhängig voneinander** sind. Das bedeutet mathematisch:  
Es wird im Folgenden vorausgesetzt, dass die Wahrscheinlichkeit  $p$  für den positiven Befund bei allen Messungen gleich groß ist (d.h.  $p = \text{const}$ ).

Ist nun  $x$  eine Zufallsvariable mit konstanter Wahrscheinlichkeit  $p$  für den positiven Befund und eine empirischen Messreihe zu  $x$  mit  $n$  Messungen gegeben, so ist nach (9.1) die relative Häufigkeit des positiven Befunds

$$h_n = \frac{k}{n}.$$

Daraus folgt durch Multiplikation beider Seiten mit  $n$

$$k = h_n \cdot n.$$

Je größer  $n$ , umso näher strebt  $k$  gegen seinen Erwartungswert  $\lambda$ , zugleich  $h_n$  gegen seinen Grenzwert  $p$ . Je größer  $n$ , umso genauer gilt also die Abschätzung

$$k = h_n \cdot n \approx p \cdot n \tag{9.2}$$

Sei nun für  $n \in \mathbb{N}$  ein **fester** Zahlwert gewählt. Betrachtet man zum selben Experiment nun alle erdenklichen Messreihen **derselben festen Länge**  $n$ , so gilt zwar stets (9.2), aber  $h_n$  schwankt doch zufällig um den Wert  $p$  herum. Folglich schwankt  $k$  zufällig um den Wert  $p \cdot n$  herum.

**Regel 75 (Erwartungswert für die Anzahl  $k$  positiver Befunde bei  $n$  Messungen):** **R 75**  
*Sei  $x$  eine Zufallsvariable, die nur die Befunde positiv/negativ liefert. Sei  $p$  die Wahrscheinlichkeit für den positiven Befund,  $n$  eine feste natürliche Zahl und*

**$k$  = Anzahl der positiven Befunde für  $x$  bei einer Messreihe der Länge  $n$ .**

*Dann ist  $k$  eine ganzzahlige Zufallsvariable mit dem*

$$\text{Erwartungswert } \lambda = p \cdot n \text{ und der Streuung } \sigma = \sqrt{p \cdot (1 - p) \cdot n}$$

*Eine solche Zufallsvariable heißt **binomialverteilt**.*

Hierfür gelten die schon in Kapitel 1 als Beispiel für die Anwendung von Binomialkoeffizienten genannten *Regeln für unabhängig voneinander wiederholte Experimente*<sup>14</sup>, die aber jetzt erst beweisbar sind.

**Regel 76 (Unabhängigkeitsregel):**

*Sind Ereignisse  $E_1, E_2, \dots, E_k$  in ihrem Auftreten unabhängig voneinander und ist die Wahrscheinlichkeit für ihr einzelnes Auftreten gleich  $p_1, p_2, \dots, p_k$ , so ist die Wahrscheinlichkeit für ihr gemeinsames (kumulatives) Auftreten gleich dem Produkt der Einzelwahrscheinlichkeiten, also gleich  $p_1 \cdot p_2 \cdot \dots \cdot p_k$ .* **R 76**



**Regel 77 (Dichtefunktion zur Binomialverteilung):****R 77**

Hat der einzelne positive Befund zu  $x$  die Wahrscheinlichkeit  $p$ , ( $0 \leq p \leq 1$ ), sind die Befunde voneinander unabhängig und ist die Zufallsvariable  $k$  definiert als die Anzahl der positiven Befunde bei  $n$  Messungen, so hängt die zu  $k$  gehörige Dichtefunktion nicht von der speziellen gemessenen Variablen  $x$  ab, sondern **einzig von der Wahrscheinlichkeit  $p$  und der Anzahl der Messungen  $n$  und lautet**

$$p_{n,p}(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \quad (0 \leq k \leq n).$$

**Bezeichnung:**  $z = p_{n,p}(k)$  heißt die **Dichtefunktion der Binomialverteilung zu den Parametern  $n$  und  $p$ .**

*Beweis von Regel 77:* Wir betrachten eine Messreihe zu  $x$  mit  $n$  durchnummerierten Ergebnissen  $x_1, \dots, x_n$  (jedes Ergebnis lautet entweder "positiv" oder "negativ").

Bezeichnet man mit  $E_j$  das Ereignis, dass bei der  $j$ -ten Messung ein positiver Befund auftritt, dass also  $x_j = \text{positiv}$  gilt. Dann haben alle Ereignisse  $E_j$  dieselbe Wahrscheinlichkeit  $p_j = p$  ( $j = 1, 2, \dots, n$ ).

Analog bezeichne  $\bar{E}_j$  das Ereignis, dass bei der  $j$ -ten Messung ein negativer Befund auftritt, dass also  $x_j = \text{negativ}$  gilt. Dann haben alle Ereignisse  $\bar{E}_j$  dieselbe Wahrscheinlichkeit  $p_j = 1 - p$  ( $j = 1, 2, \dots, n$ ).

Für die Zufallsvariable  $k$  sei nun ein **fester Wert**  $a \in \mathbb{Z}$ ,  $0 \leq a \leq n$ , betrachtet, es sei also  $k = a$ . Nach *Regel 72*<sup>15</sup> gibt die Dichtefunktion zum Experiment jeweils gerade die Wahrscheinlichkeit  $P(k = a)$  an, im vorliegenden Fall also:

$$p_{n,p}(a) = P(k = a).$$

Um den Wert von  $p_{n,p}(a)$  zu bekommen, müssen wir daher die Wahrscheinlichkeit  $P(k = a)$  ausrechnen, dass bei  $n$  Messungen von  $x$  genau  $a$  positive Befunde auftreten. Weil nach Voraussetzung die Befunde zu  $x$  voneinander unabhängig sind, kann hierzu die Unabhängigkeitsregel angewandt werden:

Die Wahrscheinlichkeit dafür, dass die ersten  $a$  Befunde  $x_1, \dots, x_a$  positiv sind, die restlichen  $(n - a)$  Befunde negativ, ist die Wahrscheinlichkeit dafür, daß bei der Auswertung der Messreihe zu  $x$  folgende Ereignisse kumulativ auftreten:  $E_1, E_2, \dots, E_a, \bar{E}_{a+1}, \bar{E}_{a+2}, \dots, \bar{E}_n$ . Dieses Resultat hat nach der *Unabhängigkeitsregel* die Wahrscheinlichkeit

$$(p_1 \cdot p_2 \cdot \dots \cdot p_a) \cdot (p_{a+1} \cdot p_{a+2} \cdot \dots \cdot p_n)$$

In diesem Produkt haben die ersten  $a$  Faktoren alle den Wert  $p$ , die restlichen  $n - a$  Faktoren alle den Wert  $1 - p$ . Also ist der Wert dieses Produkts gleich

$$p^a \cdot (1 - p)^{n-a}.$$

Der Fall  $E_1, E_2, \dots, E_a, \bar{E}_{a+1}, \bar{E}_{a+2}, \dots, \bar{E}_n$  ist aber nur eine von vielen Varianten, wie es sich ereignen kann, dass bei  $n$  Messungen genau  $a$  positive Befunde auftreten. Wie viele solche Varianten gibt es? So viele, wie es Möglichkeiten gibt, aus der  $n$ -elementigen Menge von

<sup>14</sup>siehe 1.2.4, S.10

<sup>15</sup>siehe 9.1, S.179

Messungen  $x_1, \dots, x_n$  eine genau  $a$ -elementige Menge von Messungen herauszugreifen, deren Messergebnisse positiv sein sollen.

Nach der *Kombinatorischen Regel 1*<sup>16</sup> ist die Anzahl der verschiedenen  $a$ -elementigen Teilmengen einer  $n$ -elementigen Menge gleich  $\binom{n}{a}$ . Also ist die gefragte Anzahl der Varianten gleich  $\binom{n}{a}$ , und jede dieser Varianten hat dieselbe Wahrscheinlichkeit  $p^a(1-p)^{n-a}$ .

Die **Summe** der Wahrscheinlichkeiten aller dieser Varianten,  $\binom{n}{a} \cdot p^a(1-p)^{n-a}$ , ergibt nun die Wahrscheinlichkeit dafür, dass (egal durch welche Variante) bei  $n$  Messungen genau  $a$  positive Befunde auftreten. Damit erhalten wir insgesamt:

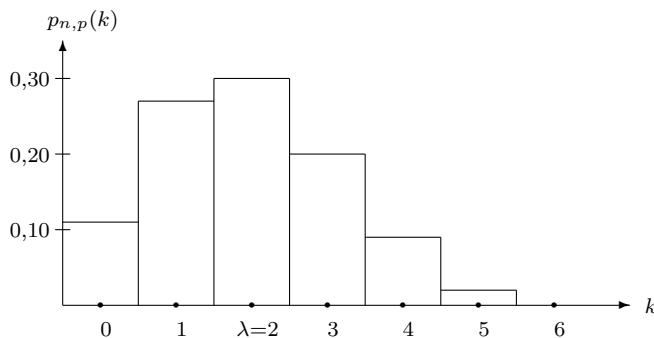
$$p_{n,p}(a) = P(k = a) = \binom{n}{a} \cdot p^a(1-p)^{n-a} \quad \text{für } 0 \leq a \leq n.$$

□

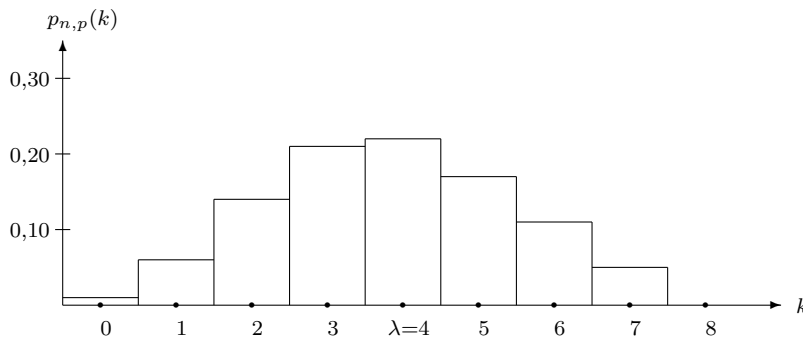
Graphisch veranschaulicht man die Dichtefunktionen der Binomialverteilung durch Säulendiagramme mit Säulen der Breite 1 über den Zahlen  $k = 0, 1, \dots, n$ .<sup>17</sup>

**Beispiel:** Sei  $x$  eine Zufallsvariable mit ja/nein-Befund und der Wahrscheinlichkeit  $p = 0,20$  (= 20%) für positiven Befund.

Säulendiagramme der Dichtefunktionen  $p_{n,p}(k)$  für  $n = 10, 20$  und  $40$ :



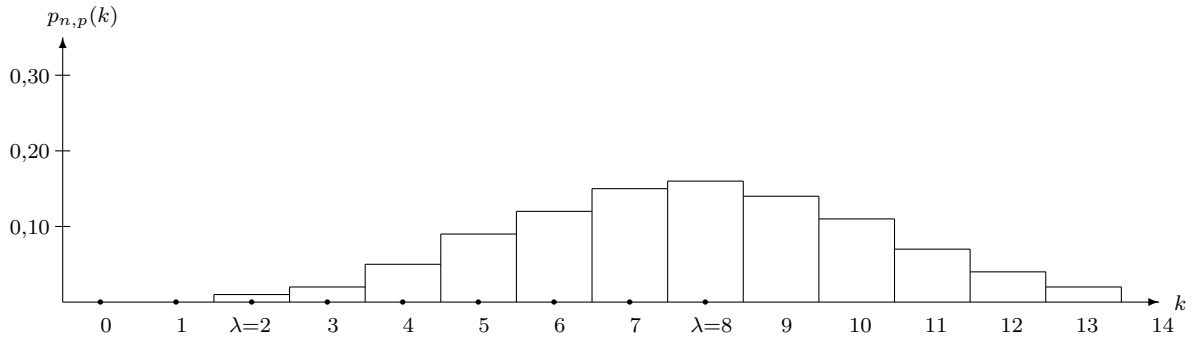
Graph von  $p_{n,p}(k)$  für  $p = 0,20$  und  $n = 10$ , Erwartungswert  $\lambda = p \cdot n = 2$



Graph von  $p_{n,p}(k)$  für  $p = 0,20$  und  $n = 20$ , Erwartungswert  $\lambda = p \cdot n = 4$

<sup>16</sup>siehe 1.2.3, S.8

<sup>17</sup>siehe 9.1, S.178



Graph von  $p_{n,p}(k)$  für  $p = 0,20$  und  $n = 40$ , Erwartungswert  $\lambda = p \cdot n = 8$

Man sieht,

- dass der Erwartungswert  $\lambda = n \cdot p$  stets in dem Abschnitt der  $k$ -Achse mit der höchsten Säule liegt, d.h. dass  $\lambda$  der häufigste Wert für  $k$  ist,
- wie der Erwartungswert  $\lambda = n \cdot p$  bei gleichbleibendem  $p$  mit wachsendem  $n$  mitwächst, so dass sich das Säulendiagramm immer weiter nach rechts verschiebt,
- dass das Diagramm mit wachsendem  $n$  immer flacher und breiter, aber zugleich auch immer symmetrischer wird, mit senkrechter Symmetrieachse bei  $k = \lambda$ .

Aus der allgemeinen *Regel 72*<sup>18</sup> folgt die konkrete Regel für den Fall der Binomialverteilung:

**Regel 78 (Berechnung von Wahrscheinlichkeiten bei Binomialverteilung):**

R 78

Ist  $x$  eine Zufallsvariable mit ja/nein-Befund, hat der einzelne positive Befund zu  $x$  die Wahrscheinlichkeit  $p$ , ( $0 \leq p \leq 1$ ), sind die Befunde voneinander unabhängig und ist die Zufallsvariable  $k$  definiert als die Anzahl der positiven Befunde bei  $n$  Messungen, so gilt:

Die Dichtefunktion  $p_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$  gibt für jeden festen ganzzahligen Wert  $k \geq 0$  die Wahrscheinlichkeit an, dass eine einzelne empirische Messreihe der Länge  $n$  zu  $x$  genau  $k$  positive Befunde ergibt.

Die Wahrscheinlichkeit  $P(a \leq k \leq b)$ , dass eine solche Messreihe eine Anzahl  $a \leq k \leq b$  liefert, ist gleich der Summe der Einzelwahrscheinlichkeiten, d.h.

$$P(a \leq k \leq b) = \sum_{k=a}^b p_{n,p}(k) = \sum_{k=a}^b \binom{n}{k} p^k (1-p)^{n-k},$$

ausführlich:

$$P(a \leq k \leq b) = \binom{n}{a} p^a (1-p)^{n-a} + \binom{n}{a+1} p^{a+1} (1-p)^{n-a-1} + \dots + \binom{n}{b} p^b (1-p)^{n-b} \quad (9.3)$$

sowie  $P(k \leq b) = P(0 \leq k \leq b)$  und  $P(k \geq a) = 1 - P(0 \leq k \leq a-1)$ .

Die Formel (9.3) für die Binomialverteilung ist ziemlich umständlich zu berechnen, insbesondere wenn  $p$  besonders klein und/oder  $n$  sehr groß ist, oder auch, wenn die Summe über viele Summanden gebildet werden muss.

Da ist es gut zu wissen, dass sie unter gewissen Voraussetzungen durch die Formel für die Poissonverteilung<sup>19</sup> oder durch die Formel für die  $N(\lambda; \sigma)$ -Verteilung ersetzt werden kann:

<sup>18</sup>siehe 9.1, S.179

<sup>19</sup>siehe 9.2, S.180, *Regel 74*

### 9.3.3 Seltene Ereignisse: Poisson- statt Binomialverteilung

Je seltener ein positiver Befund ist, umso mehr Messungen zu  $x$  müssen stattfinden, damit dieses Ereignis überhaupt einmal eintritt: Wenn  $p \leq 1$  Prozent ist, muss schon  $n \geq 300$  sein, damit der Erwartungswert  $\lambda = p \cdot n$  wenigstens  $= 3$  ist, wenn  $p \leq 1$  Promille ist, muss  $n \geq 3\,000$  sein usw.

Die Berechnung von Wahrscheinlichkeiten mit *Regel 78* kann dann für Mensch und Taschenrechner äußerst arbeitsaufwendig werden, evtl. sogar die Rechnerkapazität sprengen.

Gerade in der Medizin und Pharmakologie muss aber die Wahrscheinlichkeit von seltenen Ereignissen (Komplikationen, Nebenwirkungen, Spätfolgen von Medikationen u.a.) besonders oft berechnet werden. Da macht man sich folgende Gesetzmäßigkeit zunutze: Wenn die Population, auf die sich die Wahrscheinlichkeit bezieht, sehr groß ist ( $n$  vier, fünf, sechsstellig oder größer), und das seltene Ereignis sich darin ungefähr gleichverteilt, dann verhält sich dieses Ereignis statistisch wie ein Teilchen im physikalischen Raum, d.h. die Zufallsvariable  $k$  zur Auszählung des Ereignisses ist poissonverteilt!

Der Übergang von  $k$  aus einer binomialverteilten in eine poissonverteilte Zufallsvariable ist gleitend. Mathematisch gesprochen drückt sich dieses Phänomen so aus:

**Regel 79 (1. Grenzwertregel für die Binomialverteilung):**

**R 79**

Sei  $\lambda = p \cdot n$  der Erwartungswert für die binomialverteilte Zufallsvariable  $k$ . **Je kleiner  $p$**  (etwa  $\leq 0,01$ ) **und je größer  $n$** , umso mehr strebt die Dichtefunktion

$$z = p_{n,p}(k) \text{ der Binomialverteilung}$$

gegen die Dichtefunktion

$$z = p_{\lambda}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \text{ der Poissonverteilung.}$$

Als praktische Nutzenanwendung erhält man

**Regel 80 (Schätzregel für seltene Ereignisse):**

**R 80**

Ist  $x$  eine Zufallsvariable mit ja/nein-Befund, ist der positive Befund zu  $x$  selten ( $p \leq 0,01$ ), sind die Befunde voneinander unabhängig, ist die Zufallsvariable  $k$  definiert als die Anzahl der positiven Befunde bei  $n$  Messungen und ist

$$\lambda = p \cdot n \text{ der Erwartungswert für } k,$$

so ist  $P(k = a)$  bzw.  $P(a \leq k \leq b)$  **bequem und in guter Näherung berechenbar durch**

$$P(k = a) \approx p_{\lambda}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{bzw.}$$

$$P(a \leq k \leq b) \approx \left( \frac{\lambda^a}{a} + \frac{\lambda^{a+1}}{(a+1)!} + \dots + \frac{\lambda^b}{b} \right) \cdot e^{-\lambda}.$$

Am Beispiel  $\lambda = 2$  soll nachfolgend gezeigt werden, wie der in *Regel 79* genannte Grenzprozess konkret abläuft:

Es sind kolonnenweise die Wertetabellen einiger binomialer Dichtefunktionen  $z = p_{n,p}(k)$  mit demselben Erwartungswert  $\lambda = p \cdot n = 2$  aufgeföhrt. Simultan mit  $p \rightarrow 0$  findet  $n \rightarrow \infty$  statt (weil  $n = \frac{2}{p}$ ).

Dabei kann man beobachten, dass

$$p_{n,p} \xrightarrow{n \rightarrow \infty} p_\lambda = p_{n \cdot p}$$

$\lambda = 2$	$p_{n,p}(k)$					$\dots \rightarrow$	$p_\lambda(k)$
$p$	0,50	0,20	0,10	0,05	0,01	$\dots \rightarrow$	0
$n$	4	10	20	40	200	$\dots \rightarrow$	$\infty$
$k$							
0	0,062 500	0,107 374	0,121 577	0,128 512	0,133 980	$\dots \rightarrow$	0,135 335
1	0,250 000	0,268 435	0,270 170	0,270 552	0,270 666	$\dots \rightarrow$	0,270 671
2	0,375 000	0,301 990	0,285 180	0,277 672	0,272 033	$\dots \rightarrow$	0,270 671
3	0,250 000	0,201 327	0,190 120	0,185 114	0,181 355	$\dots \rightarrow$	0,180 447
4	0,062 500	0,088 080	0,084 498	0,090 122	0,090 220	$\dots \rightarrow$	0,090 224
5		0,026 424	0,031 921	0,034 151	0,035 723	$\dots \rightarrow$	0,036 089
6		0,005 505	0,008 276	0,010 485	0,011 727	$\dots \rightarrow$	0,012 030
7		0,000 786	0,001 970	0,002 680	0,003 283	$\dots \rightarrow$	0,003 437
8		0,000 090	0,000 356	0,000 582	0,000 800	$\dots \rightarrow$	0,000 859
9		0,000 004	0,000 053	0,000 109	0,000 172	$\dots \rightarrow$	0,000 191
10			0,000 006	0,000 018	0,000 033	$\dots \rightarrow$	0,000 038
11				0,000 003	0,000 006	$\dots \rightarrow$	0,000 007
12					0,000 001	$\dots \rightarrow$	0,000 001

### 9.3.4 Viele Messungen: Normal- statt Binomialverteilung

Sei  $k$  wieder eine binomialverteilte Zufallsvariable mit Erwartungswert  $\lambda = p \cdot n$  und Streuung  $\sigma = \sqrt{p \cdot (1 - p) \cdot n}$ . Ganz unabhängig davon, ob  $p$  klein, der positive Befund also selten ist, oder nicht, gilt:

Mit wachsender Anzahl von Messungen  $n$  wird sowohl der Erwartungswert  $\lambda$  als auch die Streuung  $\sigma$  immer größer. Da kann schnell z.B. eine Frage wie die Folgende aufkommen:

**Beispiel:**

Die Wahrscheinlichkeit für den positiven Befund bei einer Messung sei  $p = 0,22$ . Wie wahrscheinlich ist es, dass bei  $n = 378$  Messungen die Anzahl  $k$  der positiven Befunde zwischen 91 und 120 liegt? Nach *Regel 78*<sup>20</sup> muss dazu

$$P(91 \leq k \leq 120) = \sum_{k=91}^{120} \binom{378}{k} 0,22^k \cdot 0,78^{378-k}$$

berechnet werden, eine Summe mit 30 (!) unangenehm zu berechnenden Summanden.

---

<sup>20</sup>siehe 9.3, S.187

Eine große Erleichterung bietet in dieser Situation die

**Regel 81 (2. Grenzwertregel für die Binomialverteilung):**

**R 81**

Sei  $k$  eine binomialverteilte Zufallsvariable mit Erwartungswert  $\lambda = p \cdot n$  und Streuung

$$\sigma = \sqrt{p \cdot (1 - p) \cdot n}$$

**Je größer**  $\sigma$  (etwa  $\sigma > 3$ , was gleichbedeutend ist mit  $n > \frac{9}{p(1-p)}$ ), umso mehr strebt die Verteilungsfunktion der Binomialverteilung

$$P(k \leq a) = \sum_{k=0}^a \binom{n}{k} p^k (1-p)^{n-k}$$

gegen die Verteilungsfunktion der  $N(\lambda; \sigma)$ -Verteilung

$$P(k \leq a) = \Phi\left(\frac{a - \lambda}{\sigma}\right).$$

Für die Praxis folgt daraus

**Regel 82 (Schätzregel für zahlreiche Messungen):**

**R 82**

Ist  $x$  eine Zufallsvariable mit ja/nein-Befund, sind die Befunde voneinander unabhängig, ist die Zufallsvariable  $k$  definiert als die Anzahl der positiven Befunde bei  $n$  Messungen und ist die **Bedingung**

$$n > \frac{9}{p(1-p)}$$

**erfüllt**, so berechne

$$\lambda = p \cdot n$$

sowie

$$\sigma = \sqrt{\lambda \cdot (1 - p)}.$$

Dann gilt in guter Näherung

$$P(a \leq k \leq b) \approx \Phi\left(\frac{b - \lambda}{\sigma}\right) - \Phi\left(\frac{a - 1 - \lambda}{\sigma}\right).$$

Gelegentlich sind die Bedingungen "seltenes Ereignis" ( $p \leq 0,01$ ) und "zahlreiche Messungen" ( $n > \frac{9}{p(1-p)}$ ) beide erfüllt. Dann stellt sich die Frage, welche Schätzregel vorteilhafter, d.h. bequemer in der Anwendung ist. Als ungefähre Richtlinie kann gelten:

- Ist  $P(k = a)$  zu berechnen oder  $P(a \leq k \leq b)$  mit nahe beieinander liegenden  $a$  und  $b$ , so dass die Summe  $\sum_{k=a}^b$  nicht viel mehr als 3 bis 5 Summanden umfasst, so macht die Schätzregel mittels Poissonverteilung wahrscheinlich weniger Arbeit.
- Die Anwendung der Schätzregel mittels Normalverteilung ist hingegen immer vorteilhafter, wenn andernfalls sehr viele Summanden  $p_{\lambda}(k)$  oder gar  $p_{n,p}(k)$  berechnet und addiert werden müssten.

**Anwendung auf das eingangs genannte Beispiel:**

Aus  $p = 0,22$  und  $n = 378$  folgt

$$\lambda = 0,22 \cdot 378 = 83,16 \text{ und } \sigma = \sqrt{0,22 \cdot 0,78 \cdot 378} = \sqrt{64,8648} \approx 8,054.$$

Also ist *Regel 82* anwendbar.

$$\text{Aus } a = 91 \text{ folgt } \Phi\left(\frac{a - \lambda}{\sigma}\right) = \Phi\left(\frac{91 - 83,16}{8,054}\right) = \Phi(0,97) = 0,8340$$

Aus  $b = 120$  folgt  $\Phi\left(\frac{b - \lambda}{\sigma}\right) = \Phi\left(\frac{120 - 83,16}{8,054}\right) = \Phi(4,57)$ . Dieser Wert ist größer als der größte Tabellenwert  $\Phi(3,62) = 0,9999$ , und kann damit  $\approx 1$  gesetzt werden.

Damit ergibt sich insgesamt  $P(91 \leq k \leq 120) \approx 1 - 0,8340 = 0,166 = 16,6\%$ .

**9.4 Schätzung von unbekanntem  $p$  bei Binomialverteilung**

Sei  $x$  eine Zufallsvariable, die nur die Befunde positiv/negativ liefert und  $p$  die konstante, aber unbekannte Wahrscheinlichkeit des positiven Befunds.

Macht man eine Messreihe mit  $n$  Messungen und ist dabei  $H_n$  die Anzahl der positiven Befunde, so weiß man zwar, dass die relative Häufigkeit

$$h_n = \frac{H_n}{n}$$

für  $n \rightarrow \infty$  gegen den Grenzwert  $p$  strebt, aber wie weit ist dieser Näherungswert für  $p$  im konkreten Fall noch von  $p$  entfernt?

**Bezeichnung:** Ist  $\gamma$  eine beliebige Wahrscheinlichkeit mit  $0 < \gamma < 1$ , so gibt es Zahlen  $a$  und  $\varepsilon > 0$  derart, dass  $p$  mit der Sicherheit  $\gamma$  im Bereich

$$a - \varepsilon \leq p \leq a + \varepsilon$$

liegt. Dieser Bereich heißt **ein Vertrauensbereich für  $p$  zur Sicherheit  $\gamma$** .

In diesem Fall ist  $a$  ein Näherungswert für  $p$ , und mit der Sicherheit  $\gamma$  gilt für den absoluten Fehler

$$|\Delta p| = |a - p| = \varepsilon.$$

**Regel 83 (Schätzregel für unbekanntes  $p$ ):**

Ist  $H_n$  die Anzahl der positiven Befunde bei  $n$  Messungen, so berechnet sich der Vertrauensbereich für  $p$  zu einer beliebigen Sicherheit  $\gamma$

$$a - \varepsilon \leq p \leq a + \varepsilon$$

wie folgt:

$$1. \text{ Schritt: } \text{Berechne } \frac{1 + \gamma}{2}$$

2. Schritt: *Schlage in der  $\Phi$ -Tabelle denjenigen  $x$ -Wert nach, für den*

$$\Phi(x) = \frac{1 + \gamma}{2}$$

*ist und nenne diesen  $x$ -Wert gleich  $c$ .*

*Hinweis: Für  $\gamma = 0,95$  ist  $c = 1,96$ , für  $\gamma = 0,99$  ist  $c = 2,576$ , für  $\gamma = 0,999$  ist  $c = 3,30$ .*

3. Schritt: *Berechne*

$$a = \frac{H_n + 0,5 \cdot c^2}{n + c^2} \quad (= \text{Näherungswert für } p)$$

*sowie*

$$\varepsilon = \frac{c}{n + c^2} \cdot \sqrt{\frac{H_n \cdot (n - H_n)}{n} + \frac{c^2}{4}} \quad (= |\Delta p|).$$

**Zusatz:**

- *Stets ist mit der Sicherheit  $\gamma$*

$$|\Delta p| = |a - p| \leq \frac{c}{2\sqrt{n + c^2}}.$$

- *Ist  $p \leq 0,10$  und  $n \geq 100$ , so gilt mit der Sicherheit  $\gamma$  die schärfere Abschätzung*

$$|\Delta p| = |a - p| \leq \frac{c}{\sqrt{n}} \cdot \sqrt{0,09 + \frac{c^2}{400}}.$$

Der Zusatz gestattet folgende

**Anwendung:**

Um zu wissen, wie groß die Anzahl  $n$  der Messungen gewählt werden muss, damit der wahre Wert von  $p$  vom Schätzwert  $a$  mit der Sicherheit  $\gamma$  um höchstens  $\pm \frac{x}{100}$  abweicht, berechnet man  $n$  je nach Sachlage aus einer der beiden Bestimmungsgleichungen

$$\frac{c}{2\sqrt{n + c^2}} = \frac{x}{100}$$

oder, wenn  $p \leq 0,10$  und  $n \geq 100$ ,

$$\frac{c}{\sqrt{n}} \cdot \sqrt{0,09 + \frac{c^2}{400}} = \frac{x}{100}.$$

Man errechnet, dass es in jedem Fall genügt,

$$n = \left( \frac{2500}{x^2} - 1 \right) \cdot c^2$$

Messungen zu machen.

Ist voraussichtlich  $p \leq 0,10$  und wählt man bestimmt  $n \geq 100$ , so ergibt sich, dass schon die kleinere Anzahl

$$n = \frac{c^2}{x^2} \cdot (90 + 10 \cdot c^2)$$

von Messungen ausreicht.



**Beispiel:** Es soll das Risiko  $p$  einer gewissen Nebenwirkung eines Medikaments so genau eingeschätzt werden, dass mit 99% Sicherheit das wahre Risiko vom Schätzwert um höchstens  $\pm 0,5\%$  abweicht.

Hier ist  $\gamma = 0,99$ , also  $c = 2,576$  und  $x = 0,5$ . Es genügt daher jedenfalls, an

$$n = \left( \frac{2500}{0,5^2} - 1 \right) \cdot 2,576^2 \approx 66351$$

Patienten das Medikament zu verabreichen und dabei die absolute Häufigkeit  $H_n$  der fraglichen Nebenwirkung zu ermitteln.

Sodann errechnet man den gesuchten Schätzwert für das Risiko nach *Regel 83*:

$$p \approx a = \frac{H_n + 0,5 \cdot c^2}{n + c^2} \approx \frac{H_n + 3,3}{66356,6}.$$

Sei z.B. bei 995 der 66351 Patienten die fragliche Nebenwirkung aufgetreten. Dann ist also

$$p \approx a \approx \frac{995 + 3,3}{66356,6} \approx 0,01504$$

und mit *Regel 83* folgt für den absoluten Fehler mit 99% Sicherheit:

$$\begin{aligned} |\Delta p| &= \varepsilon = \frac{c}{n+c^2} \cdot \sqrt{\frac{H_n \cdot (n-H_n)}{n} + \frac{c^2}{4}} \\ &\approx \frac{2,576}{66356,6} \cdot \sqrt{\frac{995 \cdot 65356}{66351} + 1,66} \\ &\approx \frac{2,576}{66356,6} \cdot \sqrt{980,08 + 1,66} \\ &\approx \frac{80,7}{66356,6} \\ &\approx 0,00122 \end{aligned}$$

Der tatsächlich gemachte absolute Fehler erweist sich im vorliegenden Fall als noch kleiner als die geforderten 0,5%: Er beträgt de facto nur 0,12%. (Dieses erfreuliche Phänomen tritt umso deutlicher auf, je näher  $p$  bei 0 oder bei 1 liegt.) Die Auswertung der Messreihe bringt somit folgendes Ergebnis: Das Risiko für die untersuchte Nebenwirkung liegt mit 99% Sicherheit zwischen 1,382% und 1,626%.

Ist aufgrund von Erfahrungen mit dem Medikament schon im Voraus die Vermutung erlaubt, dass das Risiko  $p$  höchstens 10% beträgt, so kann man die erforderliche Anzahl wesentlich günstiger wie folgt berechnen:

$$n = \frac{c^2}{x^2} \cdot (90 + 10 \cdot c^2) = \frac{2,576^2}{0,25} \cdot (90 + 10 \cdot 2,576^2) \approx 4150,2$$

Man macht also nur eine Messreihe mit  $n = 4151$  Patienten und ermittelt hierfür die Häufigkeit  $H_n$  der Nebenwirkung. Ergäbe sich tatsächlich auch eine relative Häufigkeit von rund 1,5%, also beispielsweise eine absolute Häufigkeit von  $H_n = 64$  Fällen, so erhielte man

$$p \approx a = \frac{H_n + 0,5 \cdot c^2}{n + c^2} \approx 0,01619 = 1,619\%$$

sowie  $|\Delta p| = \frac{c}{n+c^2} \cdot \sqrt{\frac{H_n \cdot (n-H_n)}{n} + \frac{c^2}{4}} \approx 0,00498 = 0,489\%$

d.h., der absolute Fehler von  $p$  bleibt jetzt nur noch ganz knapp unter der geforderten Schranke von 0,5%.

Aus dem Beispiel kann man einen pauschalen Zusammenhang zwischen der Länge der Messrei-

he und der Größe des absoluten Fehlers ersehen: Bei der langen Messreihe mit  $n = 66351$  war  $n$  rund 16 mal so groß wie bei der kürzeren mit  $n = 4151$ . Der absolute Fehler war hingegen mit 0,12% nur 1/4 so groß wie der der kürzeren Reihe. Dahinter steht ein allgemeineres Gesetz:

**Merke:** Will man den absoluten Fehler bei der Schätzung von  $p$  um den Faktor  $1/m$  verkleinern, so muss man die Anzahl der Messungen um den Faktor  $m^2$  vergrößern.

## 9.5 Schätzung von unbekanntem $\lambda$ bei Poissonverteilung

Sei eine poissonverteilte Zufallsvariable  $k$  gegeben, bezogen auf eine fest vorgegebene Raum-Zeiteinheit ("RZE"). Ihr Erwartungswert  $\lambda$  sei unbekannt.

Wir wissen, dass

$$p_\lambda(0) = e^{-\lambda} \quad (9.4)$$

ist, also kann man  $\lambda$  näherungsweise ausrechnen, wenn man  $p_\lambda(0)$  näherungsweise kennt.<sup>21</sup> Hier soll die Frage des dabei auftretenden absoluten Fehlers  $|\Delta\lambda|$  untersucht werden.

Sei  $x$  die ja/nein-Zufallsvariable, ob die Auszählung einer Portion der RZE im konkreten Einzelfall ja oder nein den Wert  $k = 0$  ergibt, und  $p$  die (unbekannte) Wahrscheinlichkeit des positiven Befundes. Damit ist  $1 - p$  die Wahrscheinlichkeit, dass die Auszählung im Einzelfall eine Anzahl  $k \geq 1$  ergibt. Es gilt die Gleichung

$$p = p_\lambda(0). \quad (9.5)$$

Nun macht man eine Messreihe, in der für  $n$  Portionen der RZE nicht die Anzahlen  $k_1, \dots, k_n$  korrekt ausgezählt werden (das wäre eine Auswertung der Messreihe für die poissonverteilte Zufallsvariable  $k$ , und es wäre sehr viel Zählarbeit), sondern wo pro Portion lediglich überprüft wird, ob, ja oder nein, die Anzahl  $k = 0$  vorliegt (= Auswertung der Messreihe für die ja/nein-Zufallsvariable  $x$ : Sobald erkennbar wird dass  $k \geq 1$  ist, ist die Portion fertig überprüft). Sei  $H_n$  die Anzahl der positiven Befunde für  $x$ .

Jetzt kann man mit der *Schätzregel für unbekanntes  $p$* <sup>22</sup> einen Schätzwert für  $p$  und den zugehörigen absoluten Fehler  $|\Delta p|$  mit jeder gewünschten Sicherheit  $\gamma$  bestimmen. Außerdem kann man mittels des Zusatzes zu dieser Regel vorausberechnen, wie groß die Anzahl  $n$  der Messungen sein muss, um die gewünschte Genauigkeit mit der gewünschten Sicherheit  $\gamma$  garantiert zu erzielen.

Sind nun  $p$  und  $|\Delta p|$  mit der gewünschten Sicherheit  $\gamma$  ausgerechnet, so folgt aus (9.4) und (9.5)  $p = e^{-\lambda}$ , also

$$\lambda = -\ln p \quad (9.6)$$

Nach *Regel 16*<sup>23</sup> für Fehlerfortpflanzung bei einer ungenauen Variablen gilt

$$|\Delta\lambda| = \left| \frac{d\lambda}{dp} \right| \cdot |\Delta p|.$$

<sup>21</sup>siehe das Rechenbeispiel in 9.2.1, S.181

<sup>22</sup>siehe *Regel 83* in 9.4, S.191

<sup>23</sup>siehe 4.5.2, S.53

Angewandt auf (9.6) ergibt sich hieraus

$$|\Delta\lambda| = \frac{1}{p} \cdot |\Delta p| \quad (9.7)$$

In (9.6) und (9.7) besitzt man nun Formeln zur Berechnung von  $\lambda$  samt dem absoluten Fehler  $|\Delta\lambda|$  zur gewünschten Sicherheit  $\gamma$ .

Die Größe der Raum-Zeiteinheit  $RZE$  kann man ja frei wählen. Aus (9.7) ist ersichtlich, was zweckmäßig ist:

$|\Delta\lambda|$  ist antiproportional zu  $p$ . Damit  $|\Delta\lambda|$  möglichst klein bleibt, sollte also  $p = p_\lambda(0)$  möglichst groß sein. Das erreicht man dadurch, dass man die Raum-Zeiteinheit so klein wie technisch möglich wählt. Hiermit wird zugleich der Arbeitsaufwand zur Feststellung, ob, ja oder nein, der positive Befund  $k = 0$  eingetreten ist, möglichst klein gehalten.

Gleichzeitig bewirkt eine solche Wahl der Raum-Zeiteinheit natürlich auch, dass der Erwartungswert  $\lambda$  für  $k$  klein wird, im Allgemeinen einstellig. Hieraus erklärt sich wiederum, wieso gedruckte Wertetabellen für die Dichtefunktion  $p_\lambda(k)$  sich auf den Bereich  $0,1 \leq \lambda \leq 9$  beschränken.