Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun
http://blog.ruediger-braun.net

Heinrich-Heine-Universität Düsseldorf

16. Januar 2015

Verteilungsfunktionen

Verteilungsfunktion

- X sei eine Zufallsvariable. Die Funktion

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

  ist die *Verteilungsfunktion* von X
- Die Verteilungsfunktion $F(x)$ gibt an, mit welcher Wahrscheinlichkeit ein **kleinerer** Wert als x angenommen wird (x eingeschlossen)
- Die Verteilungsfunktion wächst also monoton
- Wenn X diskret ist, weist ihre Verteilungsfunktion Sprünge auf.

Beispiel: Verteilungsfunktion von $B_{6,1/3}$

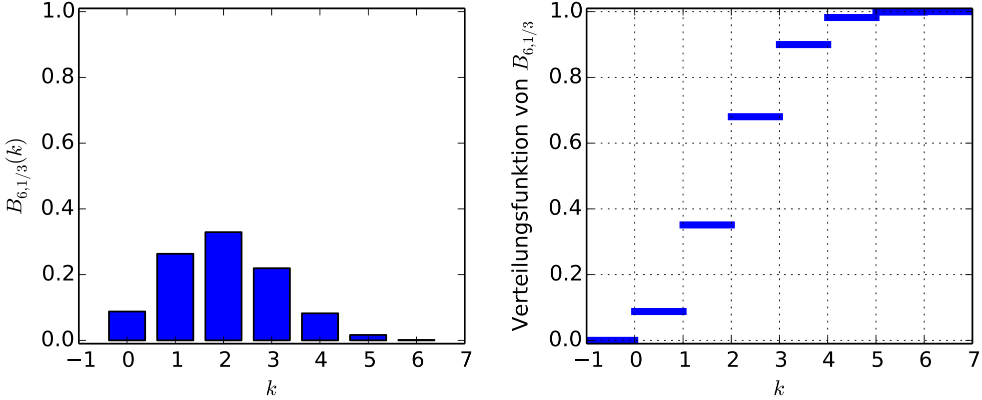
F sei die Verteilungsfunktion von $B_{6,\,1/3}$, also

$$F(r) = \sum_{k \le r} B_{6,\,1/3}(k)$$

k	$B_{6,\,1/3}(k)$	$F(k)$
0	0.0878	0.0878
1	0.2634	0.3512
2	0.3292	0.6804
3	0.2195	0.8999
4	0.0823	0.9822
5	0.0165	0.9986
6	0.0014	1.0000

Kumulierte Tabellen der Binomialverteilung zeigen die Verteilungsfunktion

Beispiel: Verteilungsfunktion von $B_{6,1/3}$



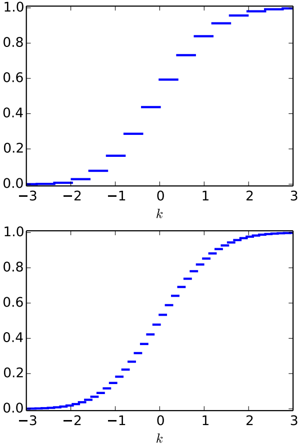
Stabdiagramm und Graph der Verteilungsfunktion von $B_{6,1/3}$

Stetige Zufallsvariable, Normalverteilung

Die standardisierte Verteilung

- Die Zufallsvariable X besitze den Erwartungswert $E(X) = \mu$ und die Varianz $Var(X) = \sigma^2$
- Setze
$$Y = \frac{X - \mu}{\sigma}$$
- Dann $E(Y) = 0$ und $Var(Y) = 1$
- Y ist die *standardisierte Zufallsvariable* zu X

Standardisierte Binomialverteilungen



Verteilungsfunktionen der Standardisierungen von $B_{n,p}$ für $p = 0.2$ und $n =$ *40, 80, 160, 320, 640* und *1280*

Zentraler Grenzwertsatz

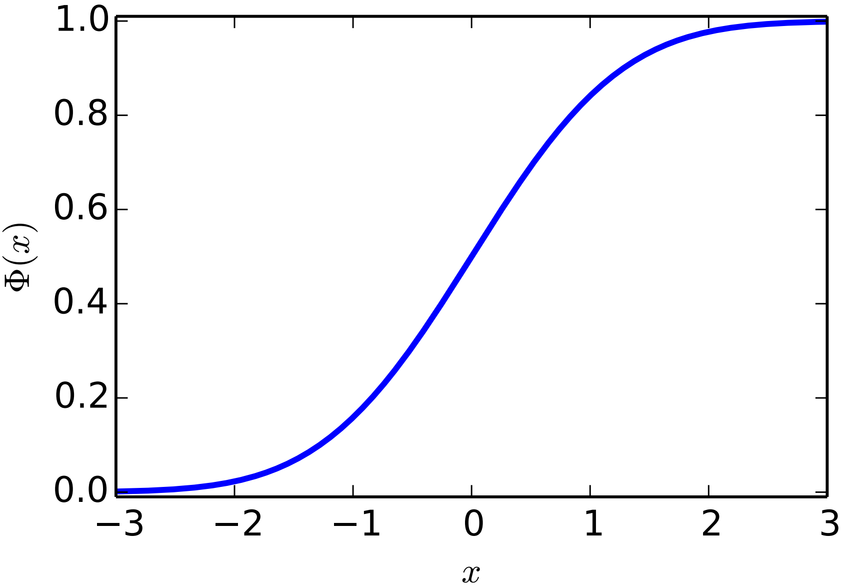
Für jeden Parameterwert p konvergiert der Grenzprozess auf der letzten Folie gegen dieselbe Funktion.

Diese Funktion heißt

Verteilungsfunktion der Standard-Normalverteilung

Sie wird mit Φ bezeichnet.

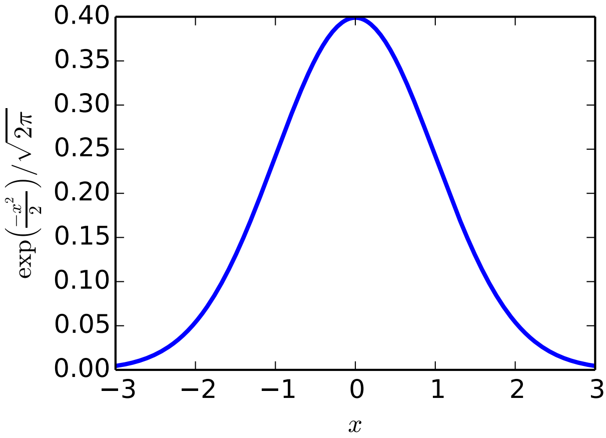
Verteilungsfunktion der Standard-Normalverteilung



Gaußsche Glockenkurve

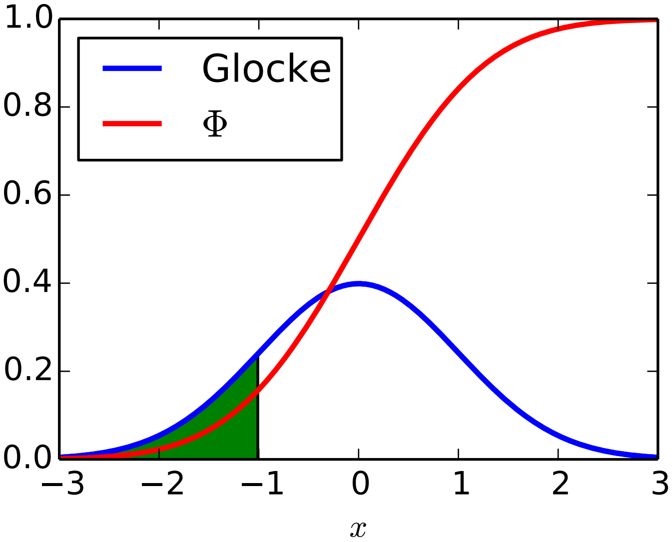
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

heißt *Gaußsche Glockenkurve*



Standard-Normalverteilung als Fläche

$\Phi(x)$ ist die Fläche links von x unter der Gaußschen Glockenkurve



Standard-Normalverteilung als Integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt$$

Für dieses Integral existiert keine geschlossene Form in Termen klassischer Funktionen

Stetige Zufallsvariable

Eine Zufallsvariable X, deren Verteilungsfunktion von der Gestalt

$$P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

ist, heißt *stetig*. Dann bezeichnet man f als die Dichte von X.

Die Gaußsche Glockenkurve ist die Dichte der Standard-Normalverteilung.

$$P(x < X \leq y) = \int_{x}^{y} f(t)dt$$

$$P(X = x) = 0$$

$$P(x \leq X \leq y) = \int_{x}^{y} f(t)dt$$

Für **stetige** Zufallsvariable gilt also $P(x < X \leq y) = P(x \leq X \leq y)$. Bei diskreten ist das nicht so.

Verteilungsfunktion der Standard-Normalverteilung

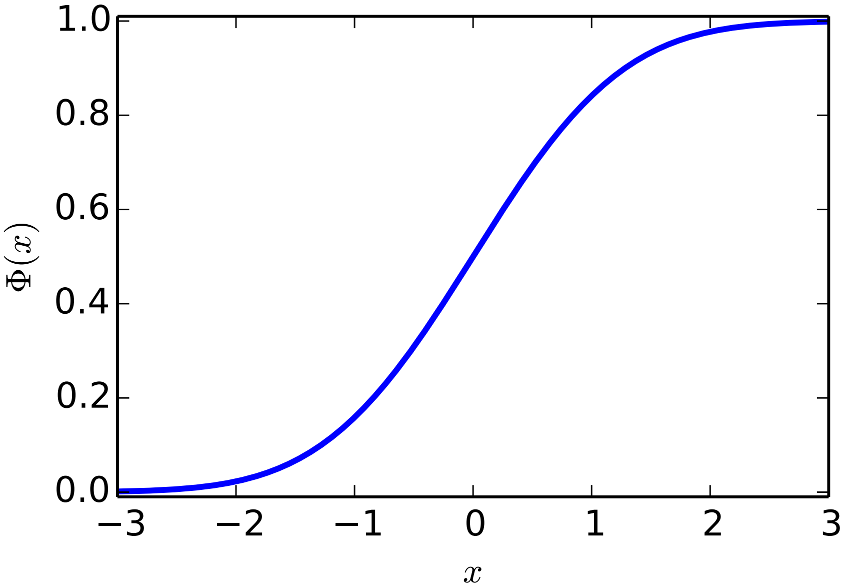


Tabelle der Standard-Normalverteilung, linke Seite

u	0.00	0.01	0.02	0.03	0.04
0.0	0,500000	0,503989	0,507978	0,511966	0,515953
0.1	,539828	,543795	,547758	,551717	,555670
0.2	,579260	,583166	,587064	,590954	,594835
0.3	,617911	,621720	,625516	,629300	,633072
0.4	,655422	,659097	,662757	,666402	,670031
0.5	,691462	,694974	,698468	,701944	,705401
0.6	,725747	,729069	,732371	,735653	,738914
0.7	,758036	,761148	,764238	,767305	,770350
0.8	,788145	,791030	,793892	,796731	,799546
0.9	,815940	,818589	,821214	,823814	,826391
1.0	,841345	,843752	,846136	,848495	,850830
1.1	,864334	,866500	,868643	,870762	,872857
1.2	,884930	,886861	,888768	,890651	,892512
1.3	,903200	,904902	,906582	,908241	,909877
1.4	0,919243	0,920730	0,922196	0,923641	0,925066

Tabelle der Standard-Normalverteilung, rechte Seite

u	0.05	0.06	0.07	0.08	0.09
0.0	0,519939	0,523922	0,527903	0,531881	0,535856
0.1	,559618	,563559	,567495	,571424	,575345
0.2	,598706	,602568	,606420	,610261	,614092
0.3	,636831	,640576	,644309	,648027	,651732
0.4	,673645	,677242	,680822	,684386	,687933
0.5	,708840	,712260	,715661	,719043	,722405
0.6	,742154	,745373	,748571	,751748	,754903
0.7	,773373	,776373	,779350	,782305	,785236
0.8	,802337	,805105	,807850	,810570	,813267
0.9	,828944	,831472	,833977	,836457	,838913
1.0	,853141	,855428	,857690	,859929	,862143
1.1	,874928	,876976	,879000	,881000	,882977
1.2	,894350	,896165	,897958	,899727	,901475
1.3	,911492	,913085	,914657	,916207	,917736
1.4	0,926471	0,927855	0,929219	0,930563	0,931888

Lesehinweise

- Auf dem Weblog gibt es die komplette Tabelle im Format A4 unter http://www.math.uni-duesseldorf.de/~braun/bio1415/tabNorm.pdf
- Beispiel $\Phi(0.31) = 0.621720$
- Wegen der Punktsymmetrie, also wegen

$$\Phi(-x) = 1 - \Phi(x)$$

  kann man die Tabelle auch für negative Argumente verwenden. Also beispielsweise $\Phi(-1.34) = 1 - \Phi(1.34) = 1 - 0.909877 = 0.090123$
- Durch Aufsuchen des Funktionswertes erhält man die Umkehrfunktion: Gesucht u mit $\Phi(u) = 0.79$. Man liest ab: $u = 0.81$
- In der letzten Zeile der ausgedruckten Tabelle sind die Werte der Umkehrfunktion für ausgewählte Argumente angegeben. Man bezeichnet sie als *Quantile*

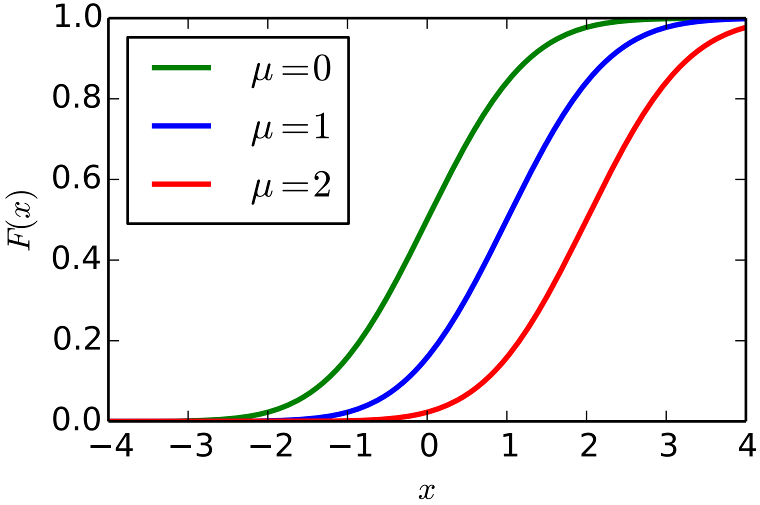
Tabelle der Standard-Normalverteilung, linke Seite

u	0.00	0.01	0.02	0.03	0.04
0.0	0,500000	0,503989	0,507978	0,511966	0,515953
0.1	,539828	,543795	,547758	,551717	,555670
0.2	,579260	,583166	,587064	,590954	,594835
0.3	,617911	,621720	,625516	,629300	,633072
0.4	,655422	,659097	,662757	,666402	,670031
0.5	,691462	,694974	,698468	,701944	,705401
0.6	,725747	,729069	,732371	,735653	,738914
0.7	,758036	,761148	,764238	,767305	,770350
0.8	,788145	,791030	,793892	,796731	,799546
0.9	,815940	,818589	,821214	,823814	,826391
1.0	,841345	,843752	,846136	,848495	,850830
1.1	,864334	,866500	,868643	,870762	,872857
1.2	,884930	,886861	,888768	,890651	,892512
1.3	,903200	,904902	,906582	,908241	,909877
1.4	0,919243	0,920730	0,922196	0,923641	0,925066

Normalverteilungen

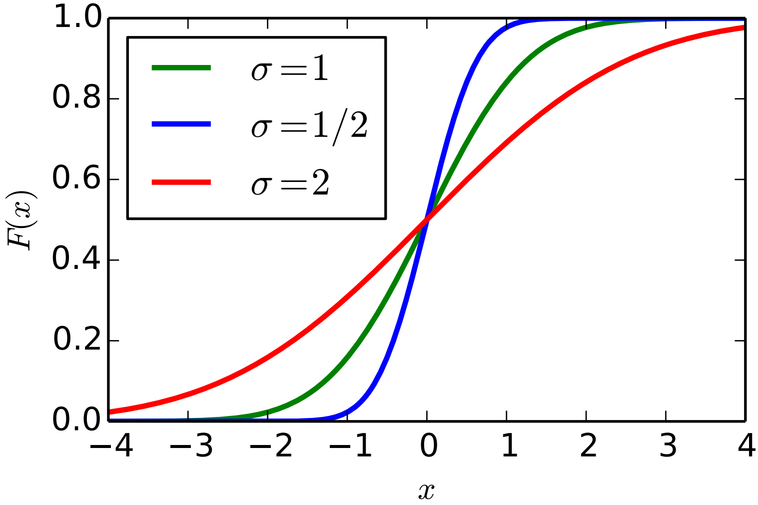
- Die Zufallsvariable X heißt *normalverteilt* zum Erwartungswert μ und der Varianz σ^2, wenn
$$Y = \frac{X - \mu}{\sigma}$$
standard-normalverteilt ist. Man sagt dann, X sei $N(\mu, \sigma^2)$-verteilt
- $X = \mu + \sigma \cdot Y$
- Y ist die Standardisierung von X
- Normalverteilungen werden beispielsweise zur Modellierung von Messfehlern benutzt

Normalverteilungen für verschiedene Erwartungswerte



Verteilungsfunktionen für $N(\mu, 1)$-verteilte Zufallsvariable

Normalverteilungen für verschiedene Streuungen



Verteilungsfunktionen für $N(0, \sigma^2)$-verteilte Zufallsvariable

Umrechnung auf Standardnormalverteilung

Die Zufallsvariable X sei $N(\mu, \sigma^2)$-verteilt. Dann ist $\frac{X-\mu}{\sigma}$ standard-normalverteilt und für $a < b$ gelten

$$P(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$P(a < X) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$P(X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Beispiel: natürliche Variabilitäten

- Roggenpflanzen erreichen eine mittlere Höhe von $0.98m$. Dabei streut die Höhe um $19cm$. Welcher Prozentsatz aller Pflanzen erreicht mindestens $1.10m$ Höhe?
- $X =$ Höhe der Pflanze
- Wir rechnen in Metern. Dann $E(X) = 0.98$ und $\sigma = 0.19$
- Wir suchen

$$P(1.1 < X) = 1 - \Phi\left(\frac{1.1 - 0.98}{0.19}\right) = 1 - \Phi(0.6316) = 1 - 0.735653 = 0.2644$$

- ca 26% der Pflanzen sind mindestens $1.10m$ hoch

Kritische Betrachtung des Modells

- Das Modell erlaubt auch den unsinnigen Fall, dass Roggenpflanzen eine negative Höhe aufweisen
- Mit welcher Wahrscheinlichkeit geschieht das?

$$
\begin{aligned}
P(X < 0) = \Phi\left(\frac{-0.98}{0.19}\right) &= \Phi(-5.158) = 1 - \Phi(5.158) \\
&\leq 1 - \Phi(5) = 1 - (1 - 2.867 \cdot 10^{-7}) = 2.867 \cdot 10^{-7}
\end{aligned}
$$

- Das Modell sagt für weniger als eine unter 3 Millionen Pflanzen eine negative Höhe voraus
- Damit können wir leben

Beispiel: IQ-Tests

- IQ-Tests sind so skaliert, dass die Werte in der Population normalverteilt mit Erwartungswert $\mu = 100$ und Streuung $\sigma = 15$ sind
- Welcher Anteil der Bevölkerung hat einen IQ über 130?
- X messe den IQ
- X ist $N(100, 225)$-verteilt.
- Also

$$P(130 < X) = 1 - \Phi\left(\frac{130 - 100}{15}\right) = 1 - \Phi(2) = 1 - 0.977250 = 0.02275$$

- 2.275% der Population weist einen IQ von mindestens 130 auf

Normalapproximation

Normalapproximation

- Die Zufallsvariable X sei $B(n, p)$-verteilt
- $E(X) = n \cdot p$ und $Var(X) = n \cdot p \cdot (1 - p)$
- Die standardisierte Zufallsvariable zu X ist

$$Y = \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}}$$

- Der zentrale Grenzwertsatz sagt aus, dass Y für große n annähernd standardnormalverteilt ist
- "groß" bedeutet

$$n \cdot p \cdot (1 - p) > 9$$

Normalapproximation: Formel

- Die Zufallsvariable X sei $B(n,p)$-verteilt mit $n \cdot p \cdot (1-p) > 9$
- Dann gilt näherungsweise für natürliche Zahlen $a < b$

$$P(a \leq X \leq b) \simeq \Phi\left(\frac{b+\frac{1}{2}-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right) - \Phi\left(\frac{a-\frac{1}{2}-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right)$$

- Wenn $a = 0$ oder $b = n$ ist, braucht man nur einen Term

$$P(a \leq X) \simeq 1 - \Phi\left(\frac{a-\frac{1}{2}-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right)$$

$$P(X \leq b) \simeq \Phi\left(\frac{b+\frac{1}{2}-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right)$$

- Die Terme $\frac{1}{2}$ in den Formeln bezeichnet man als Stetigkeitskorrektur

Beispiel zur Normalapproximation

- Heilversuch mit 94 Fischen; Heilerfolgswahrscheinlichkeit im Einzelfall $p = 0.85$
- Mit welcher Wahrscheinlichkeit werden mindestens 80 Fische geheilt?
- Die Anwendung der Normalapproximation ist gerechtfertigt, denn

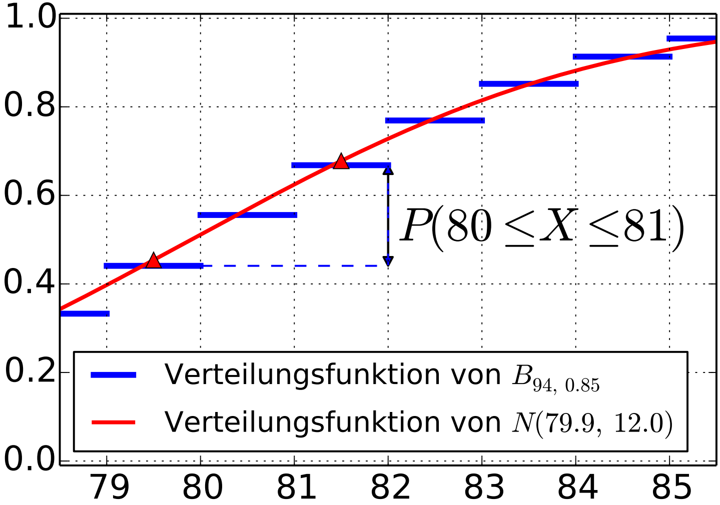
$$n \cdot p \cdot (1 - p) = 94 \cdot 0.85 \cdot 0.15 = 11.985$$

- Also wegen $n \cdot p = 79.9$

$$P(X \geq 80) \cong 1 - \Phi\left(\frac{80 - 0.5 - 79.9}{\sqrt{11.985}}\right)$$
$$= 1 - \Phi(-0.1155) = \Phi(0.1155) \cong \Phi(0.12) = 0.548$$

- Die Wahrscheinlichkeit, dass mindestens 80 Fische geheilt werden, beträgt ungefähr 55%

Stetigkeitskorrektur: grafische Erklärung



Beispiel: Zuchtlachse

- Zuchtlachsen wird Fischabfall und vegetarisches Futter zur Auswahl angboten
- 11% aller Lachse bevorzugen das Gemüse
- Die Lachse werden stufenweise an vegetarische Kost gewöhnt
- Die Frage

  Bevorzugen sie das Gemüse nach der Umgewöhnung mit höherer Wahrscheinlichkeit als vorher?

  soll zum Signifikanzniveau $\alpha = 0.05$ beantwortet werden
- Einseitiger, oberer Binomialtest mit Nullhypothese $H_0 = \{p \leq p_0\}$, wobei $p_0 = 0.11$
- Stichprobenumfang $n = 380$

Zuchtlachse: Bestimmung des kritischen Werts

- Kritischer Wert c bestimmt durch

$$\sum_{k=0}^{c} B_{n,\,p_0}(k) \geq 1-\alpha$$

- Wende Normalapproximation an

$$\Phi\left(\frac{c+\frac{1}{2}-n\cdot p_0}{\sqrt{n\cdot p_0\cdot(1-p_0)}}\right) \cong 0.95$$

wobei $n\cdot p_0 = 41.8$ und $n\cdot p_0\cdot(1-p_0) = 37.2$

- Tabelle:

$$\frac{c+\frac{1}{2}-n\cdot p_0}{\sqrt{n\cdot p_0\cdot(1-p_0)}} \cong 1.645$$

- Also

$$c+\frac{1}{2}-41.8 = 1.645\cdot\sqrt{37.2} = 10.0$$

und daher $c = 52$ (aufgerundet). Bei mindestens 53 vegetarischen Lachsen wird H_0 abgelehnt.

Zuchtlachse: Power des Tests

- Was ist die Power des Tests, wenn tatsächlich nach der Umgewöhnungsphase 15% der Lachse vegetarische Kost bevorzugen?
- Die Power ist die Wahrscheinlichkeit, bei Vorliegen der Alternative die Nullhypothese auch tatsächlich abzulehnen
- H_0 wird abgelehnt bei mindestens 53 Erfolgen
- Im Beispiel $p = 0.15$
- Also ist die Power gleich

$$\sum_{k=53}^{380} B_{380,0.15}(k) \cong 1 - \Phi\left(\frac{53 - \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}\right) = 1 - \Phi\left(\frac{52.5 - 57}{\sqrt{48.5}}\right)$$
$$= 1 - \Phi(-0.646) = 1 - (1 - \Phi(0.646)) = \Phi(0.646) = 0.742$$

- Die Power beträgt fast 75%.

Zuchtlachse: Vergleich

Alle Tests zum Signifikanzniveau $\alpha = 0.05$

n	c	Power bei $p = 0.15$
38	8	0.106
380	53	0.742
760	98	0.942
1140	143	0.989
1520	187	0.998

Die 3-Sigma Regel

Die 3σ-Regel

- $X_1, \ldots, X_n$ unabhängig, alle mit derselben Verteilung
- $\mu = E(X_1) = \cdots = E(X_n)$ und $\sigma^2 = Var(X_1) = \cdots = Var(X_n)$
- Arithmetisches Mittel der X_j

$$Y = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- n ausreichend groß, dann näherungsweise

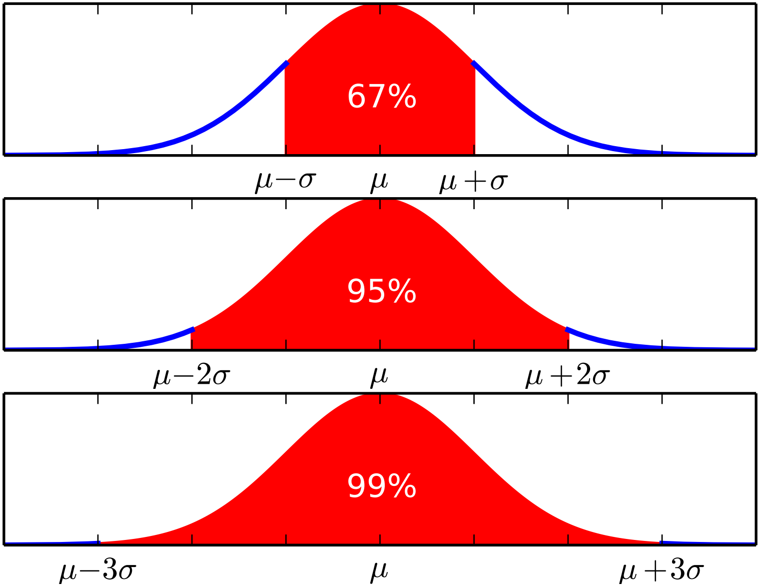
$$P\left(\left|Y - \mu\right| \leq \frac{\sigma}{\sqrt{n}}\right) \geq \frac{2}{3}$$
$$P\left(\left|Y - \mu\right| \leq \frac{2\sigma}{\sqrt{n}}\right) \geq 0.95$$
$$P\left(\left|Y - \mu\right| \leq \frac{3\sigma}{\sqrt{n}}\right) \geq 0.99$$

- Die 3σ-Regel ist eine Faustregel.

Erläuterung der 3σ-Regel

- $E(Y) = \mu$
- $Var(Y) = n\sigma^2$
- Also ist $n(Y - \mu)$ die Standardisierung von Y
- Für große n sieht die Standardisierung aus wie Φ

3σ-Regel, Zeichnung



3σ-Regel, Beispiel

- Der Wirkstoffgehalt von Düngetabletten wird gemessen (in mg). Die Streuung betrage $8mg$
- Wie viele Messungen schreibt die 3σ-Regel vor, um mit 95% Wahrscheinlichkeit einen Messfehler von weniger als $4mg$ zu haben?
- Wir brauchen
$$4 = \frac{2\sigma}{\sqrt{n}}$$
- Also
$$4 = \frac{2 \cdot 8}{\sqrt{n}}$$
- Das bedeutet $n = 16$