Mathe III

Ralf Herbrich

Verteilung von stetigen Zufallsvariablen

Überblick

1. Gleichverteilung
2. Exponentialverteilung
3. Gammaverteilung
4. Normalverteilung
5. χ^2-Verteilung
6. t-Verteilung

Überblick

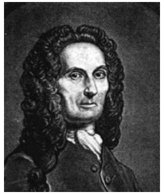
Normalverteilung

- **Definition (Normalverteilung).** Für zwei Parameter $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}_+$ bezeichnen wir die Wahrscheinlichkeitsverteilung $\mathcal{N}(\cdot; \mu, \sigma^2)$ auf dem Ereignisraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ als Normalverteilung, falls für alle $A \in \mathcal{B}(\mathbb{R})$ gilt, dass

$$\mathcal{N}(A; \mu, \sigma^2) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \int_A \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right) dx$$

- **Bemerkungen (Normalverteilung)**
  - Die Dichte der Normalverteilung ist symmetrisch und strikt positiv für alle $x \in \mathbb{R}$.
  - Die Normalverteilung $\mathcal{N}(0,1)$ wird Standardnormalverteilung genannt.
  - Die Normalverteilung wird auch Gauß-Verteilung bzw. Gaußsche Glockenkurve genannt.
  - Die kumulative Verteilungsfunktion einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ wird mit $\Phi(\cdot; \mu, \sigma^2)$ bezeichnet bzw. mit $\Phi(\cdot)$ im Falle der Standardnormalverteilung

$$\Phi(t; \mu, \sigma^2) = \int_{-\infty}^{t} \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right) dx$$



Abraham de Moivre (1667 – 1754)

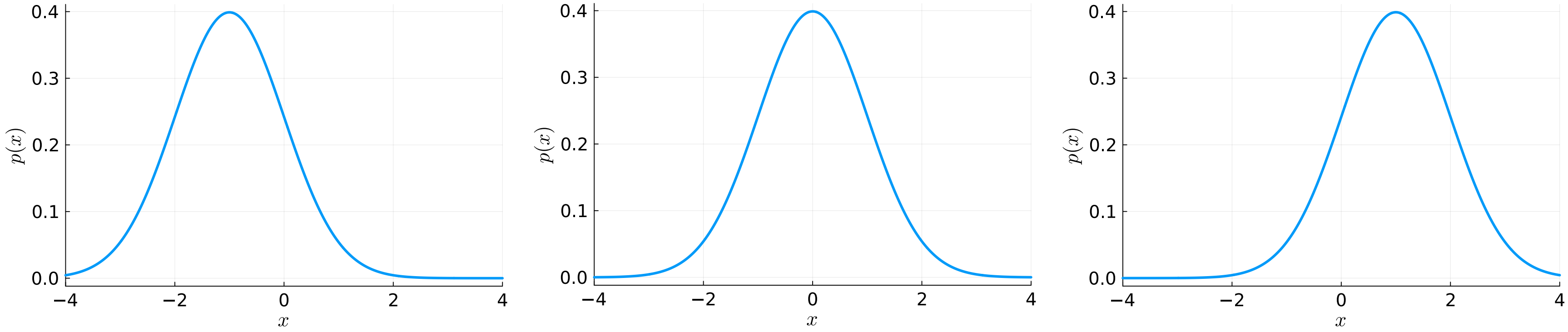
Carl Friedrich Gauss (1777 - 1855)

Mathe III

Unit 6b – Verteilung von stetigen Zufallsvariablen

Normalverteilung: Dichte und Implementierung in **R**

$\mathcal{N}(-1,1)$ $\mathcal{N}(0,1)$ $\mathcal{N}(1,1)$



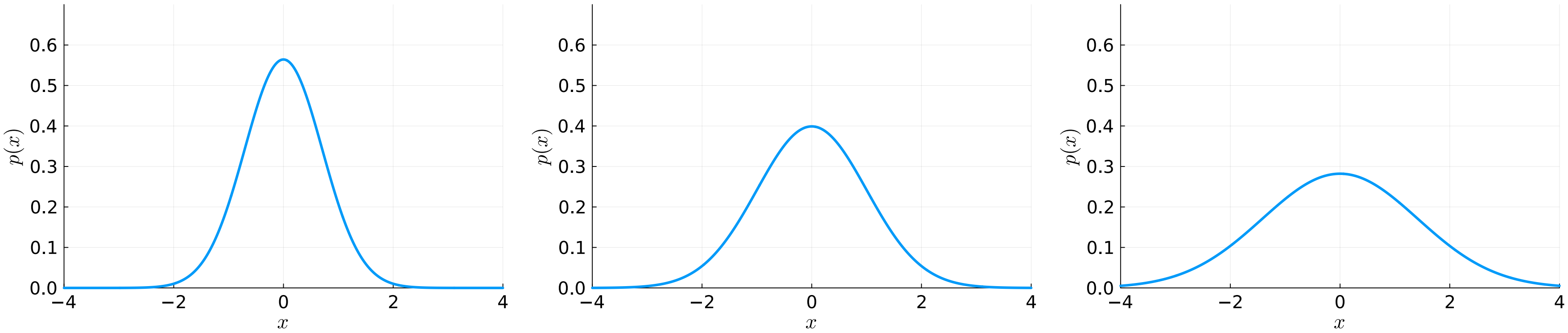
- **Bemerkungen (Normalverteilung)**
  - μ legt die Symmetrieachse der Dichte fest
  - Um μ ist der wahrscheinlichste Ausgang eines Zufallsexperiments

`dnorm(x,m,s)` ← berechnet $p(x)$ wenn $X \sim \mathcal{N}(m, s^2)$

`pnorm(x,m,s)` ← berechnet $P(X \leq x)$ wenn $X \sim \mathcal{N}(m, s^2)$

Normalverteilung: Dichte

$\mathcal{N}(0,0.5)$ $\mathcal{N}(0,1)$ $\mathcal{N}(0,2)$



- **Bemerkungen (Normalverteilung)**
  - σ legt fest, wie stark die Werte um die Symmetrieachse streuen
  - Größere Werte von σ verkleinern das Maximum der Dichte

Ursprung der Normalverteilung

- Bedeutung der Normalverteilung ist zentral in der Statistik

1. **Exakte Verteilung von physikalischen Größen**
   - Position eines Partikels während der Diffusion
   - Wahrscheinlichkeitsdichte des Basiszustands in einem harmonischen Quantenoszillator
2. **Approximative Verteilung**
   - Summe von vielen unabhängigen, identisch verteilten Zufallsvariablen
   - Zentraler Grenzwertsatz (nächste Vorlesung!)
3. **Beobachtete (empirische) Verteilung**
   - **Biologie**: Logarithmus von physiologischen Größen (z.B. Größe, Gewicht)
   - **Finanzwesen**: Logarithmus von Wechselkursen & Preisindizes
   - **Messtheorie**: Messung von physikalischen Größen (z.B. Temperatur)
   - **Meteorologie**: Monatliche Regenfallmengen

Lineartransformation normalverteilter Zufallsvariablen

- **Satz (Lineartransformation normalverteilter Zufallsvariablen).** Seien $a, b \in \mathbb{R}$ und $X \sim \mathcal{N}(\mu, \sigma^2)$ eine reelle Zufallsvariable. Dann gilt
$$a \cdot X + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$
- **Beweis**: Sei $Y = a \cdot X + b$ und F_Y die Verteilungsfunktion von Y. Für $a = 0$ ist der Satz wahr.

Für $a > 0$ gilt

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y)$$
$$= P\left(X \leq \frac{y-b}{a}\right)$$
$$= F_X\left(\frac{y-b}{a}\right)$$

Für $a < 0$ gilt

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y)$$
$$= P\left(X \geq \frac{y-b}{a}\right)$$
$$= 1 - F_X\left(\frac{y-b}{a}\right)$$

Für $\mathrm{a} < 0$ ist die Ableitung $-\frac{1}{a} p_X\left(\frac{y-b}{a}\right)$

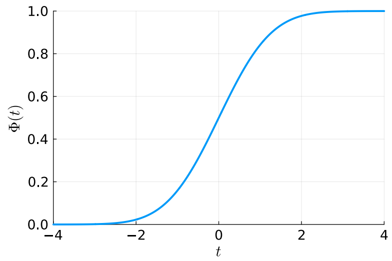
$$p_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{|a|} p_X\left(\frac{y-b}{a}\right)$$

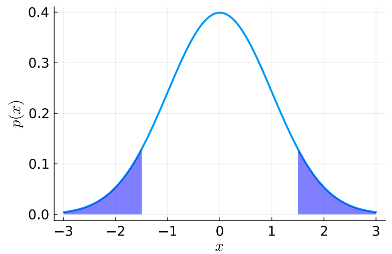
$$p_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2 a^2}} \exp\left(-\frac{\left(y - (a\mu + b)\right)^2}{2\sigma^2 a^2}\right)$$

Dichte von $\mathcal{N}(a\mu + b, a^2\sigma^2)$

Standardnormalverteilung

- **Korollar (Standardnormalverteilung)**. Sei eine reelle Zufallsvariable $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist die Variable $Z \sim \mathcal{N}(0,1)$ standardnormalverteilt, wobei
$$Z = \frac{X - \mu}{\sigma}$$
- **Beweis**: Folgt direkt aus der Lineartransformation von normalverteilten Zufallsvariablen.
- **Bemerkungen (Standardnormalverteilung)**
  - Zur Berechnung der Verteilungsfunktion $P(X \leq t)$ wenn $X \sim \mathcal{N}(\mu, \sigma^2)$ reicht die Standardverteilungsfunktion
$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{t} \exp\left(-\frac{x^2}{2}\right) dx$$
  - Für beliebige $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$ gilt $\Phi(t; \mu, \sigma^2) = \Phi\left(\frac{t-\mu}{\sigma}\right)$
  - Die Funktion $\Phi(t)$ hat die Eigenschaft
$$\underbrace{\Phi(-t)}_{P(X \leq -t)} = \underbrace{1 - \Phi(t)}_{P(X > t)}$$





Mathe III

Unit 6b – Verteilung von stetigen Zufallsvariablen

Erwartungswert der Normalverteilung

- **Satz (Erwartungswert der Normalverteilung).** Eine reelle Zufallsvariable $X \sim \mathcal{N}(0, \sigma^2)$ hat den Erwartungswert
$$E[X] = 0$$
- **Beweis**: Aus der Definition des Erwartungswertes folgt
$$E[X] = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left[-\sigma^2 \exp\left(-\frac{x^2}{2\sigma^2}\right) \Big|_{-\infty}^{+\infty} \right] = \frac{1}{\sqrt{2\pi}} \cdot [0 - 0] = 0$$
- **Korollar (Erwartungswert der Normalverteilung).** Eine reelle Zufallsvariable $X \sim \mathcal{N}(\mu, \sigma^2)$ hat den Erwartungswert
$$E[X] = \mu$$

Varianz der Normalverteilung

- **Satz (Varianz der Normalverteilung)**. Eine reelle Zufallsvariable $X \sim \mathcal{N}(0,1)$ hat die Varianz

$$V[X] = 1$$

- **Beweis**: Mit Hilfe der partiellen Integration sieht man, dass

$$E[X^2] = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} \underbrace{x}_{u(x)} \cdot \underbrace{\left(x \cdot \exp\left(-\frac{x^2}{2}\right)\right)}_{v'(x)} dx$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \left[\underbrace{x \cdot \left(-\exp\left(-\frac{x^2}{2}\right)\right)\Bigg|_{-\infty}^{+\infty}}_{= 0 - 0 = 0} + \underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx}_{= \sqrt{2\pi}} \right]$$

Partielle Integration

$$\int_a^b u(x) \cdot v'(x)\, dx = u(x) \cdot v(x)\Big|_a^b - \int_a^b u'(x) \cdot v(x)\, dx$$

- **Korollar (Varianz der Normalverteilung)**. Eine reelle Zufallsvariable $X \sim \mathcal{N}(\mu, \sigma^2)$ hat die Varianz

$$V[X] = \sigma^2$$

Summe von Normalverteilungen

- **Satz (Summe von normalverteilten Zufallsvariablen).** Seien $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ unabhängige reelle Zufallsvariablen. Dann gilt
$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- **Beispiel (Lineartransformation und Faltung normalverteilter Zufallsvariablen).** Seien $X_1 \sim \mathcal{N}(3,4)$ und $X_2 \sim \mathcal{N}(0,1)$ unabhängige Zufallsvariablen. Wie ist die Verteilung von $3 \cdot X_1 + 2 \cdot X_2 + 5$?
- **Lösung**. Sei $Y = 3 \cdot X_1 + 2 \cdot X_2 + 5$. Dann ist Y normalverteilt nach den beiden Parametern
$$\mu = 3 \cdot 3 + 2 \cdot 0 + 5 = 14$$
und
$$\sigma^2 = 3^2 \cdot 4 + 2^2 \cdot 1 = 40$$

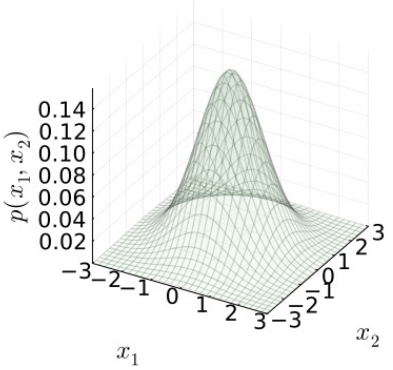
Exkurs: Bivariate Normalverteilung

- Oft unterliegt nicht nur eine, sondern mehrere Größen einer zufälligen Verteilung. Diese können dennoch gemeinsam betrachtet werden.

- **Beispiel (Bivariate Normalverteilung).** Ein Bogenschütze schießt auf die Mitte einer Zielscheibe. Trotz viel Übung trifft er nicht immer ins Gelbe. Die Zufallsvariablen X und Y geben die horizontale bzw. vertikale Abweichung in cm an und werden als unabhängig normalverteilt mit $\mu_X = \mu_Y = 0$ und $\sigma_X^2 = \sigma_Y^2 = 1$ angenommen. Wie können wir eine gemeinsame Dichte für X und Y angeben?
  - **Lösung**. Für zwei unkorrelierte normalverteilte Zufallsvariablen X und Y können wir die gemeinsame Dichte $f_{X,Y}$ für alle $(x, y) \in \mathbb{R}^2$ angeben als

$$p_{X,Y}(x,y) = \frac{1}{2 \cdot \pi \cdot \sigma_X \cdot \sigma_Y} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right)$$

  - Wir sprechen von einer **bivariaten Normalverteilung** für unabhängige Zufallsvariablen.



Exkurs: Kovarianzmatrix

- **Definition (Kovarianzmatrix).** Für ein $n \in \mathbb{N}_+$ sowie n Zufallsvariablen $X_1, X_2, \dots, X_n$ bezeichnen wir $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ als Kovarianzmatrix, falls für alle $i, j \in \{1, \dots, n\}$ gilt, dass

$$\Sigma_{i,j} = \mathrm{Cov}\left(X_i, X_j\right)$$

- **Bemerkungen (Kovarianzmatrix)**
  - Die Hauptdiagonale der Kovarianzmatrix enthält die Varianzen $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ der einzelnen Zufallsvariablen.
  - Eine Kovarianzmatrix ist symmetrisch entlang der Hauptdiagonalen.
- **Beispiel (Kovarianzmatrix)**

	X_1	X_2	X_3
X_1	2	0	1
X_2	0	0.5	2
X_3	1	2	3

	X_1	X_2	X_3
X_1	4	0	0
X_2	0	1	0
X_3	0	0	2

Exkurs: Multivariate Normalverteilung

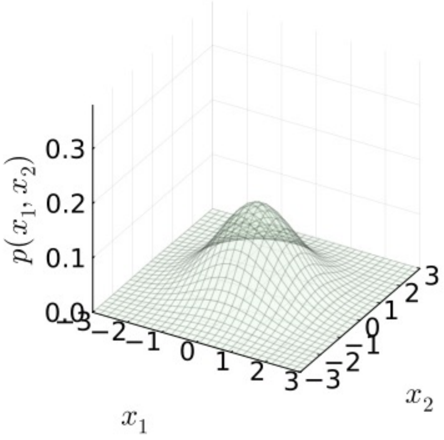
- **Definition (Multivariate Normalverteilung).** Sei $n \in \mathbb{N}_+$ und $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ eine Kovarianzmatrix sowie $\boldsymbol{\mu} \in \mathbb{R}^n$ ein Vektor. Dann bezeichnen wir die Wahrscheinlichkeitsverteilung$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ auf dem Ereignisraum $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ als multivariate Normalverteilung, falls für alle $A \in \mathcal{B}(\mathbb{R}^n)$ gilt, dass

$$\mathcal{N}(A; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_A \frac{1}{\sqrt{(2 \cdot \pi)^n \cdot |\boldsymbol{\Sigma}|}} \cdot \exp\left(-\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right) d\boldsymbol{x}$$

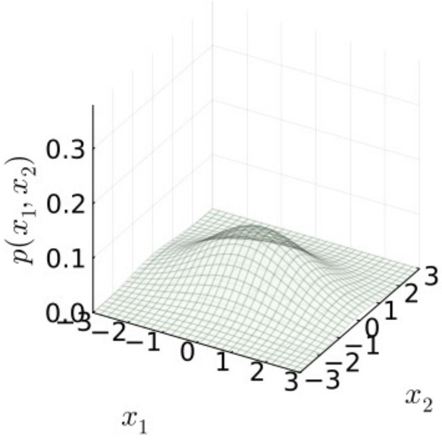
- **Bemerkungen (Multivariate Normalverteilung)**
  - Die Randverteilungen einer multivariaten Normalverteilung sind normalverteilt.
  - Die einzelnen Zufallsvariablen einer multivariaten Normalverteilung sind genau dann unkorreliert, wenn sie unabhängig sind.

Exkurs: Multivariate Normalverteilung in Bildern

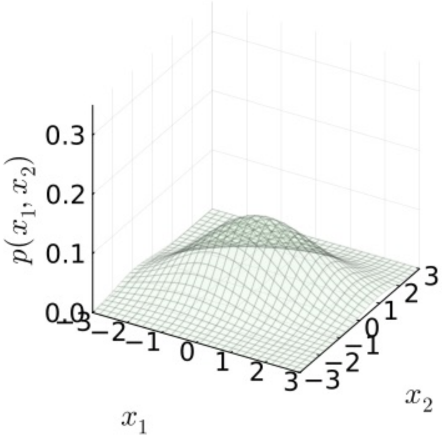
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



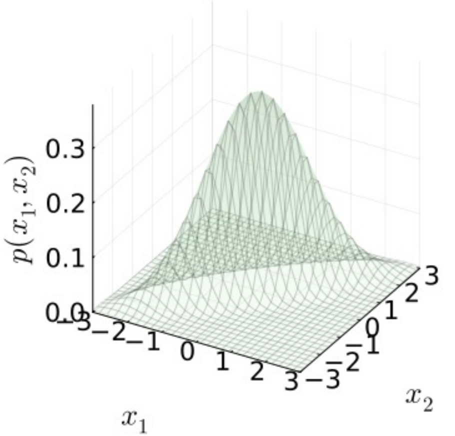
$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

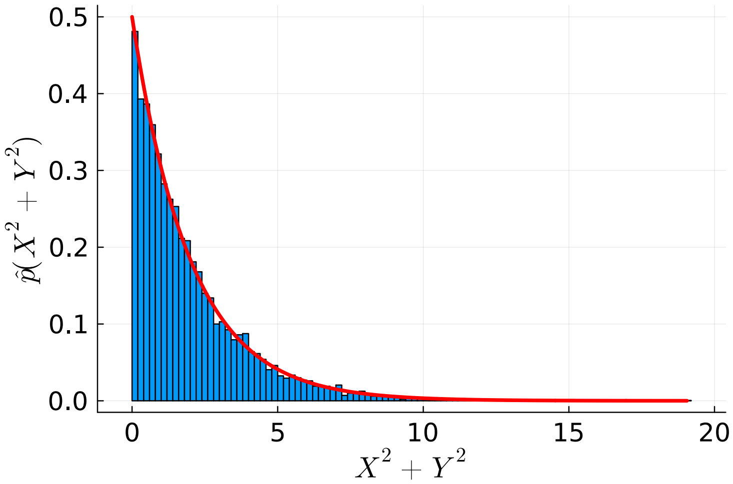


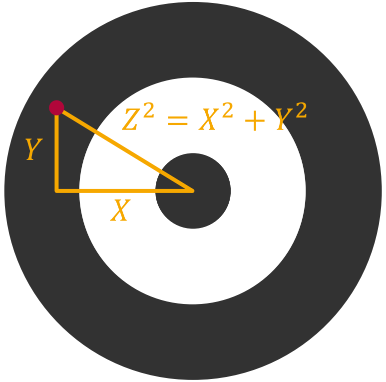
Überblick

Beispiel: χ^2-Verteilung

- **Beispiel (χ^2-Verteilung).** Ein Bogenschütze schießt auf die Mitte einer Zielscheibe. Trotz viel Übung trifft er nicht immer ins Schwarze. Die Zufallsvariablen X und Y geben die horizontale bzw. vertikale Abweichung in cm an und werden als unabhängig standardnormalverteilt angenommen.

  Wie ist die quadratische Distanz vom Auftreffpunkt zur Mitte der Zielscheibe verteilt?





χ^2-Verteilung

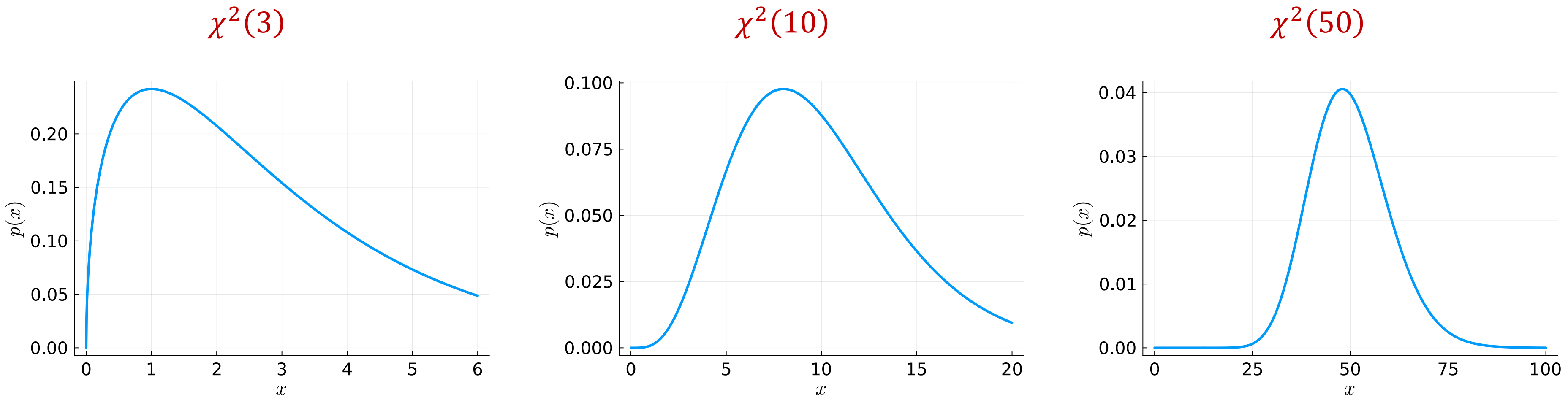
- **Definition (χ^2-Verteilung).** Sei $n \in \mathbb{N}_+$ und seien $X_1, \ldots, X_n$ unabhängige standardnormalverteilte Zufallsvariablen. Dann folgt die Zufallsvariable $Z = \sum_{i=1}^{n} X_i^2$ einer χ^2-Verteilung mit n Freiheitsgraden, auch als $\chi^2(n)$ bezeichnet. Für alle $A \in \mathcal{B}(\mathbb{R})$ gilt, dass

$$\chi^2(A; n) = \int_A \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot x^{\frac{n}{2}-1} \cdot \exp\left(-\frac{x}{2}\right) \cdot \mathbb{I}_{[0,\infty)}(x)\, dx$$

$$\mathrm{Gam}(A; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \cdot \int_A \lambda \cdot \exp(-\lambda x)\, (\lambda x)^{\alpha-1} \cdot \mathbb{I}_{[0,\infty)}(x)\, dx$$

- **Bemerkungen (χ^2-Verteilung)**
  - Die Dichte entspricht dem Spezialfall der Gammaverteilung mit $\alpha = \frac{n}{2}$ und $\lambda = \frac{1}{2}$.
  - Intuitiv ziehen wir unabhängig n standardnormalverteilte Zufallsvariablen, quadrieren und addieren diese.
  - Die χ^2-Verteilung wird primär für statistische Tests verwendet. Wir werden die Verteilung in den nächsten Wochen für den Mann-Whitney-Test benötigen.

χ^2-Verteilung: Dichte und Implementierung in **R**



- **Bemerkung (χ^2-Verteilung)**
  - Die Dichte nähert sich einer symmetrischen Verteilung mit steigenden n

`dchisq(x,n)` ← berechnet $p(x)$ wenn $X \sim \chi^2(n)$

`pchisq(x,n)` ← berechnet $P(X \leq x)$ wenn $X \sim \chi^2(n)$

Momente der χ^2-Verteilung

- **Satz (Erwartungswert und Varianz der χ^2-Verteilung).** Sei $X \sim \chi^2(n)$ eine reelle Zufallsvariable. Dann gilt für den Erwartungswert $E[X] = n$ und für die Varianz $V[X] = 2n$.
  - **Beweis**: Folgt direkt aus den Eigenschaften der Gamma-Verteilung: wenn $Y \sim \mathrm{Gam}(\alpha, \lambda)$ dann ist $E[Y] = \frac{\alpha}{\lambda}$ und $V[Y] = \frac{\alpha}{\lambda^2}$. Mit $\alpha = \frac{n}{2}$ und $\lambda = \frac{1}{2}$ ist der Satz bewiesen.
- **Satz (Summe von χ^2-verteilten Zufallsvariablen).** Seien $X_1 \sim \chi^2(n_1)$ und $X_2 \sim \chi^2(n_2)$ zwei unabhängige reelle Zufallsvariablen. Dann gilt

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

  - **Beweis**: Seien $Y_1, \dots, Y_{n_1}, Y_{n_1+1}, \dots, Y_{n_1+n_2}$ unabhängige standard-normalverteilte Zufallsvariablen, $Y_i \sim \mathcal{N}(0,1)$, $i = 1, \dots, n_1 + n_2$. Sei außerdem

$$X_1 = Y_1^2 + \cdots + Y_{n_1}^2 \sim \chi^2(n_1)$$

$$X_2 = Y_{n_1+1}^2 + \cdots + Y_{n_1+n_2}^2 \sim \chi^2(n_2)$$

Definition der χ^2 Verteilung

Dann ist aber per Definition

$$X_1 + X_2 = Y_1^2 + \cdots + Y_{n_1}^2 + Y_{n_1+1}^2 + \cdots + Y_{n_1+n_2}^2 \sim \chi^2(n_1 + n_2)$$

Überblick

1. Gleichverteilung
2. Exponentialverteilung
3. Gammaverteilung
4. Normalverteilung
5. χ^2-Verteilung
6. **t-Verteilung**

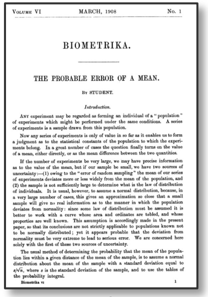
t-Verteilung

- **Definition (t-Verteilung).** Sei $Z \sim \mathcal{N}(0,1)$ und $Y \sim \chi^2(n)$. Dann folgt die reelle Zufallsvariable $X = Z/\sqrt{\frac{Y}{n}}$ einer t-Verteilung mit n Freiheitsgraden. Für eine t-Verteilung mit n Freiheitsgraden gilt für alle $A \in \mathcal{B}(\mathbb{R})$, dass

$$t(A;n) = \int_A \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n \cdot \pi} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$$

- **Bemerkungen (t-Verteilung)**
  - Die t-Verteilung wurde von William Sealy Gosset, der für die Guinness Brauerei arbeitete, entdeckt und unter dem Pseudonym „Student“ veröffentlicht (daher nennt man die Verteilung auch „Student t-Verteilung“).
  - Die t-Verteilung wird primär für statistische Tests verwendet.
  - Wir werden die Verteilung in den nächsten Wochen für den t-Test benötigen.

William Sealy Gosset (1876 – 1937)



Volume VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

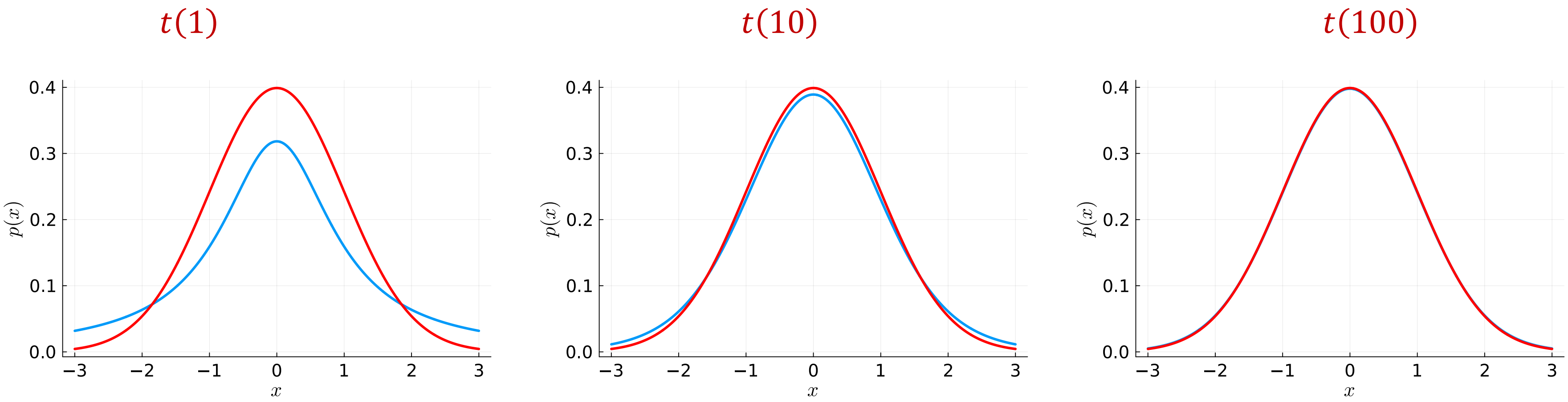
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

Biometrika VI 1

t-Verteilung: Dichte und Implementierung in R



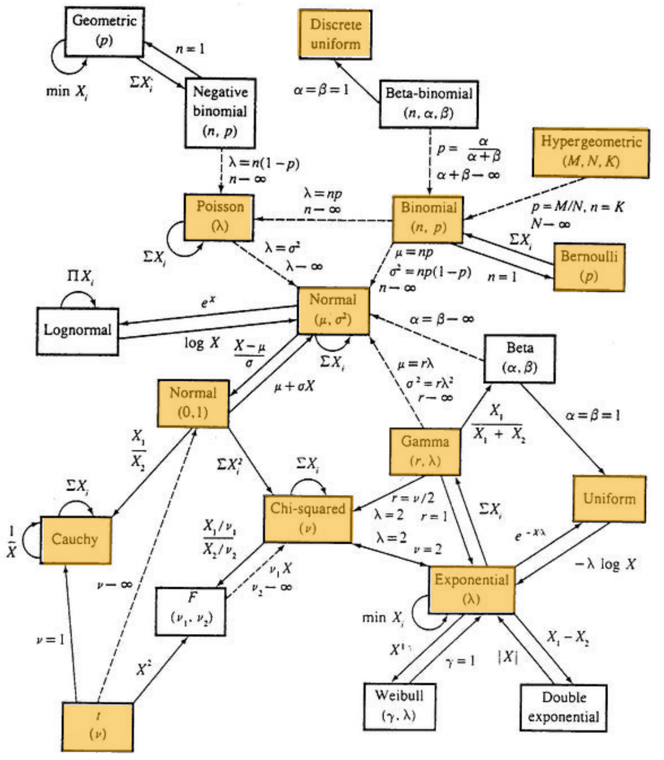
- **Bemerkungen (t-Verteilung)**
  - Für $t = 1$ ist die t-Verteilung die Cauchy-Verteilung und hat keinen Erwartungswert!
  - Mit steigendem n nähert sich die Dichte der Standardnormalverteilung an.

`dt(x,n)` ← berechnet $p(x)$ wenn $X \sim t(n)$

`pt(x,n)` ← berechnet $P(X \leq x)$ wenn $X \sim t(n)$

Beziehung zwischen Verteilungen

- Neben den bereits bekannten Verteilungen existiert noch eine Vielzahl weiterer bekannter Standardmodelle, die in Relation zueinander stehen.
- Für Interessierte: https://www.math.wm.edu/~leemis/chart/UDR/UDR.html



Relationships among common distributions. Solid lines represent transformations and speci cases, dashed lines represent limits. Adapted from Leemis (1986).

Grafik: George Casella & Roger L. Berger: Statistical Inference. 2. Auflage. Duxbury, Pacific Grove 2001.

Viel Spaß bis zur nächsten Vorlesung!