

# Wahrscheinlichkeitsrechnung und Statistik für Biologen Wiederholung: Verteilungen

Noémie Becker & Dirk Metzler

31. Mai 2016

## Inhaltsverzeichnis

1	Binomialverteilung	1
2	Normalverteilung	2
3	T-Verteilung	4
4	Chi-Quadrat-Verteilung	5
5	F-Verteilung	6

## 1 Binomialverteilung

### Binomialverteilung

Sei  $X$  die Anzahl der Erfolge bei  $n$  unabhängigen Versuchen mit Erfolgswahrscheinlichkeit von jeweils  $p$ . Dann gilt für  $k \in \{0, 1, \dots, n\}$

$$\Pr(X = k) = \binom{n}{k} p^k \cdot (1 - p)^{n-k}$$

und  $X$  heißt *binomialverteilt*, kurz:

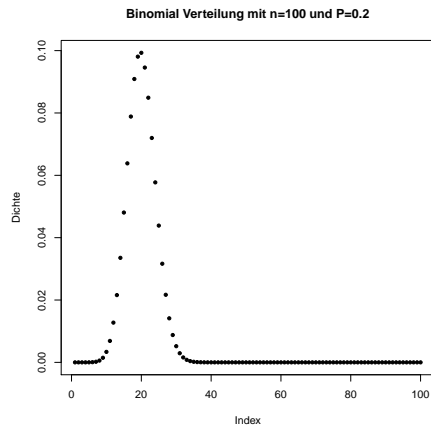
$$X \sim \text{bin}(n, p).$$

Eine Ja/Nein-Zufallsgröße nennt man auch eine Bernoulli-Zufallsgröße.

$$\mathbb{E}X = np$$

$$\text{Var } X = n \cdot p \cdot (1 - p)$$

### Dichte der Binomialverteilung



### Benutzung der Binomialverteilung

Das Problem bei der Binomialverteilung ist:  $\binom{n}{k}$  exakt zu berechnen, ist für große  $n$  sehr aufwändig. Deshalb:

Die Binomialverteilung wird oft durch andere Verteilungen approximiert.

## 2 Normalverteilung

### Normalverteilung

Eine Zufallsvariable  $Z$  mit der Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

heißt *standardnormalverteilt*.

kurz:  $Z \sim \mathcal{N}(0, 1)$

$$\mathbb{E}Z = 0$$

$$\text{Var } Z = 1$$

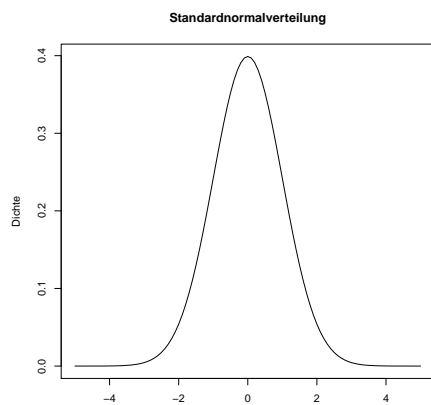
Ist  $Z \mathcal{N}(0, 1)$ -verteilt, so ist  $X = \sigma \cdot Z + \mu$  normalverteilt mit Mittelwert  $\mu$  und Varianz  $\sigma^2$ , kurz:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$X$  hat dann die Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

### Dichte der Normalverteilung



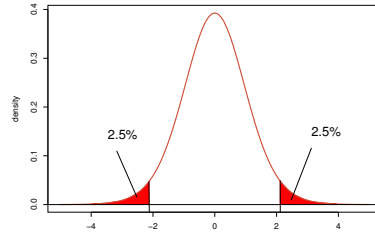
## Merkregeln der Normalverteilung

Ist  $Z \sim \mathcal{N}(\mu, \sigma^2)$ , so gilt:

- $\Pr(|Z - \mu| > \sigma) \approx 33\%$
- $\Pr(|Z - \mu| > 1.96 \cdot \sigma) \approx 5\%$
- $\Pr(|Z - \mu| > 3 \cdot \sigma) \approx 0.3\%$

## Berechnung von Quantilen

Sei  $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$  standardnormalverteilt. Für welchen Wert  $z$  gilt  $\Pr(|Z| > z) = 5\%$ ?



Wegen der Symmetrie bzgl der y-Achse gilt

$$\Pr(|Z| > z) = \Pr(Z < -z) + \Pr(Z > z) = 2 \cdot \Pr(Z < -z)$$

Finde also  $z > 0$ , so dass  $\Pr(Z < -z) = 2.5\%$ .

```
> qnorm(0.025, mean=0, sd=1)
```

```
[1] -1.959964 Antwort:  $z \approx 1.96$ , also knapp 2 Standardabweichungen
```

## Normalapproximation

Für große  $n$  und  $p$ , die nicht zu nahe bei 0 oder 1 liegen, kann man die Binomialverteilung durch die Normalverteilung mit dem entsprechenden Erwartungswert und der entsprechenden Varianz approximieren:

Ist  $X \sim \text{bin}(n, p)$  und  $Z \sim \mathcal{N}(\mu = n \cdot p, \sigma^2 = n \cdot p \cdot (1 - p))$ , so gilt

$$\Pr(X \in [a, b]) \approx \Pr(Z \in [a, b])$$

(eine Faustregel: für den Hausgebrauch meist okay, wenn  $n \cdot p \cdot (1 - p) \geq 9$ )

## Zentraler Grenzwertsatz

Eine etwas allgemeinere *Normalapproximation* beschreibt der **Zentrale Grenzwertsatz**.

Der zentrale Grenzwertsatz besagt, dass die Verteilung von Summen

**unabhängiger und identisch verteilter**

Zufallsvariablen in etwa die Normalverteilung ist.

## Zentraler Grenzwertsatz

Die  $\mathbb{R}$ -wertigen Zufallsgrößen  $X_1, X_2, \dots$  seien unabhängig und identisch verteilt mit endlicher Varianz  $0 < \text{Var } X_i < \infty$ . Sei außerdem

$$Z_n := X_1 + X_2 + \dots + X_n$$

die Summe der ersten  $n$  Variablen.

Dann ist die zentrierte und reskalierte Summe im Limes  $n \rightarrow \infty$  standardnormalverteilt, d.h.

$$\frac{Z_n - \mathbb{E}Z_n}{\sqrt{\text{Var } Z_n}} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

bei  $n \rightarrow \infty$ .

Formal: Es gilt für alle  $-\infty \leq a < b \leq \infty$

$$\lim_{n \rightarrow \infty} \Pr\left(a \leq \frac{Z_n - \mathbb{E}Z_n}{\sqrt{\text{Var } Z_n}} \leq b\right) = \Pr(a \leq Z \leq b),$$

wobei  $Z$  eine standardnormalverteilte Zufallsvariable ist.

### 3 T-Verteilung

#### T-Verteilung

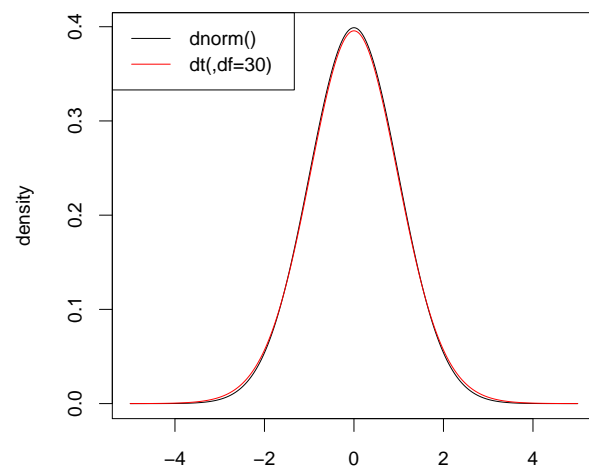
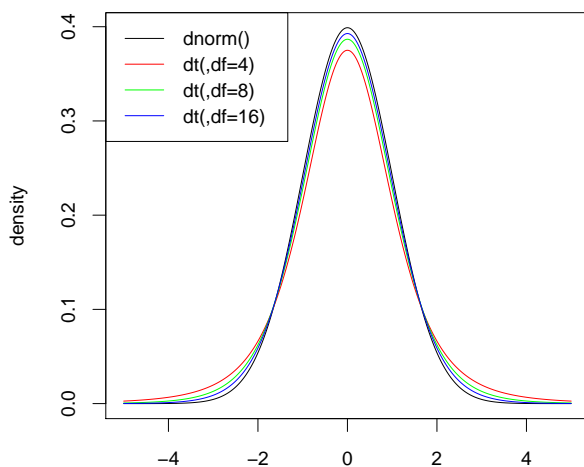
Sind  $X_1, \dots, X_n$  unabhängig aus einer Normalverteilung mit Mittelwert  $\mu$  gezogen, so ist

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

**t-verteilt** mit  $n - 1$  Freiheitsgraden (df=*degrees of freedom*). Eine t-verteilte Zufallsvariable bezeichnen wir meist mit  $T$ .

Die t-Verteilung heißt auch **Student-Verteilung**. Die t-Verteilung wurde 1908 von William Gosset veröffentlicht, während Gosset in einer Guinness-Brauerei arbeitete. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie unter dem Pseudonym *Student*.

#### Dichte der t-Verteilung



#### T-Test

Gepaarter t-test

Ein-Stichproben t-test

Zwei-Stichproben t-Test, ungepaart mit gleichen Varianzen

Welch-t-Test, die Varianzen dürfen ungleich sein

**T test : Zweiseitig oder einseitig testen?**

In den meisten Fällen will man testen, ob zwei Stichproben sich signifikant unterscheiden.  $\rightsquigarrow$  **zweiseitiger Test**

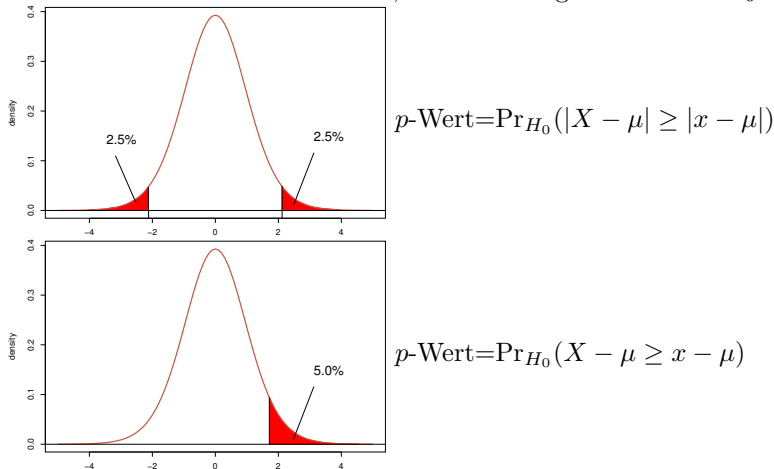
In manchen Fällen

- kann man von vornherein ausschließen, dass die erste Stichprobe kleinere Werte als die zweite Stichprobe hat. Dann will man testen, ob die erste Stichprobe signifikant größer ist.
- will man nur testen, ob der Mittelwert der ersten Stichprobe signifikant größer ist (bzw. kleiner).

↪ einseitiger Test

### T test : Zweiseitig oder einseitig testen?

Wir beobachten einen Wert  $x$ , der deutlich größer als der  $H_0$ -Erwartungswert  $\mu$  ist.



## 4 Chi-Quadrat-Verteilung

### Chi-Quadrat-Verteilung

Seien  $X_1, X_2, \dots, X_n$   $n$  unabhängige standardnormalverteilte Zufallsvariablen, so ist

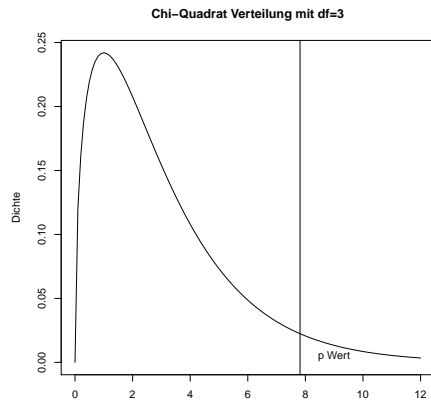
$$Y = \sum_i X_i^2$$

Chi-Quadrat-verteilt mit  $n$  Freiheitsgraden.

$$\mathbb{E}Y = n$$

$$\text{Var } Y = 2n$$

### Dichte der Normalverteilung



### Chi-Quadrat-Test

Gegeben Abweichungen zwischen Daten und eine Verteilung oder zwischen zwei Verteilungen. Wir messen die Abweichungen durch die  $X^2$ -Statistik:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

wobei  $E_i$  = erwartet Anzahl in Klasse  $i$  und  $O_i$  = beobachtete (engl. *observed*) Anzahl in Klasse  $i$ .

Falls die Nullhypothese gilt und die Erwartungswerte  $E_i$  nicht zu klein sind (Faustregel: sie sollten alle  $\geq 5$  sein), ist  $X^2$  *ungefähr*  $\chi^2$ -verteilt. Die  $\chi^2$ -Verteilung hängt ab von der Anzahl der Freiheitsgrade **df**.

## 5 F-Verteilung

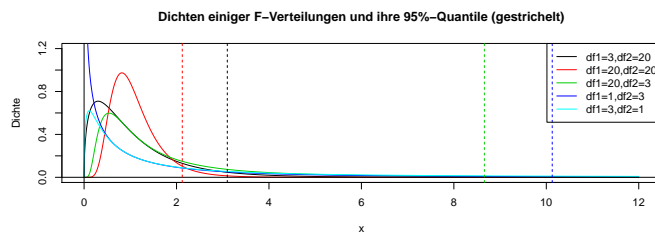
### F-Verteilung

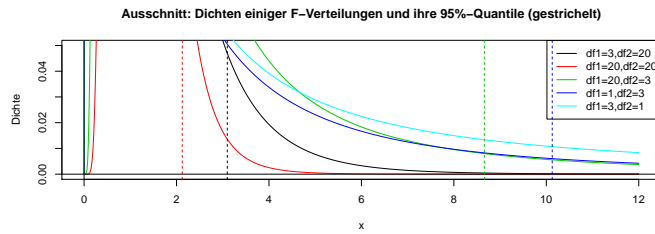
Sind X und Y unabhängige  $\chi^2$ -verteilte Zufallsvariablen mit Freiheitsgraden m für X und n für Y, so ist

$$F = \frac{X/m}{Y/n}$$

**F-verteilt** mit  $m$  und  $n$  Freiheitsgraden.

### Dichte der F-Verteilung





Das 95%-Quantil der F-Verteilung mit  $df_1 = 3$  und  $df_2 = 1$  passte leider nicht in diese Abbildung. Es beträgt 215.7

### F-Test

$X_{ij}$  =  $j$ -te Beobachtung in der  $i$ -ten Gruppe,  $j = 1, \dots, n_i$ , Modellannahme:  $X_{ij} = \mu_i + \varepsilon_{ij}$ .

$\mathbb{E}[\varepsilon_{ij}] = 0$ ,  $\text{Var}[\varepsilon_{ij}] = \sigma^2$

$SS_{\text{innerh}} = \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$     Quadratsumme innerhalb d. Gruppen,  
 $n - I$  Freiheitsgrade

$SS_{\text{zw}} = \sum_{i=1}^I n_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$     Quadratsumme zwischen d. Gruppen,  
 $I - 1$  Freiheitsgrade

$$F = \frac{SS_{\text{zw}} / (I - 1)}{SS_{\text{innerh}} / (n - I)}$$

Unter der Hypothese  $H_0 : \mu_1 = \dots = \mu_I$  („alle  $\mu_i$  sind gleich“) ist  $F$  Fisher-verteilt mit  $I - 1$  und  $n - I$  Freiheitsgraden

(unabhängig vom tatsächlichen gemeinsamen Wert der  $\mu_i$ ).

### F-Test

$F$ -Test: Wir lehnen  $H_0$  zum Signifikanzniveau  $\alpha$  ab, wenn  $F \geq q_\alpha$ , wobei  $q_\alpha$  das  $(1 - \alpha)$ -Quantil der Fisher-Verteilung mit  $I - 1$  und  $n - I$  Freiheitsgraden ist.