Karlsruher Institut für Technologie

Geodatenanalyse I: Univariate Statistik und statistisches Testen

Kathrin Menberg

Stundenplan

Vorläufiger Stundenplan		
Datum	**Thema**	**Dozent**
20.10.2021	Einführung in die Programmierung mit *Python*	Gabriel Rau
25.10.2021	Univariate Statistik und statistisches Testen	Kathrin Menberg
01.11.2021	*Feiertag*	
08.11.2021	Variablen, Datentypen und Logik eines Programms	Gabriel Rau
15.11.2021	Bivariate und schließende Statistik	Kathrin Menberg
22.11.2021	Umgang und Berechnung von Datensätzen	Gabriel Rau
29.11.2021	Multivariate Statistik	Kathrin Menberg
06.12.2021	Datenvisualisierung mit *matplotlib*	Gabriel Rau
13.12.2021	Monte-Carlo Methoden	Kathrin Menberg
20.12.2021	Datenformate, Datenspeicherung und Datenbanken	Gabriel Rau
27.12.2021	*Weihnachtsferien*	
03.01.2022	*Weihnachtsferien*	
10.01.2022	Sensitivitätsanalyse	Kathrin Menberg
17.01.2022	Analyse und Visualisierung von Geodaten	Gabriel Rau
24.01.2022	Räumliche Interpolation	Kathrin Menberg
31.01.2022	Datenethik, Lizensierung und Entwicklungstools	Gabriel Rau
07.02.2022	Regressionsanalyse	Kathrin Menberg

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:20	**Univariate Statistik**
10:20 – 11:00	Übung
11:00 – 11:10	Diskussion und Reflexion
11:10 – 11:25	Pause
11:25 – 11:45	Statistisches Testen
11:45 – 12:20	Übung
12:20 – 12:30	Diskussion und Reflexion

Lernziele Univariate Statistik

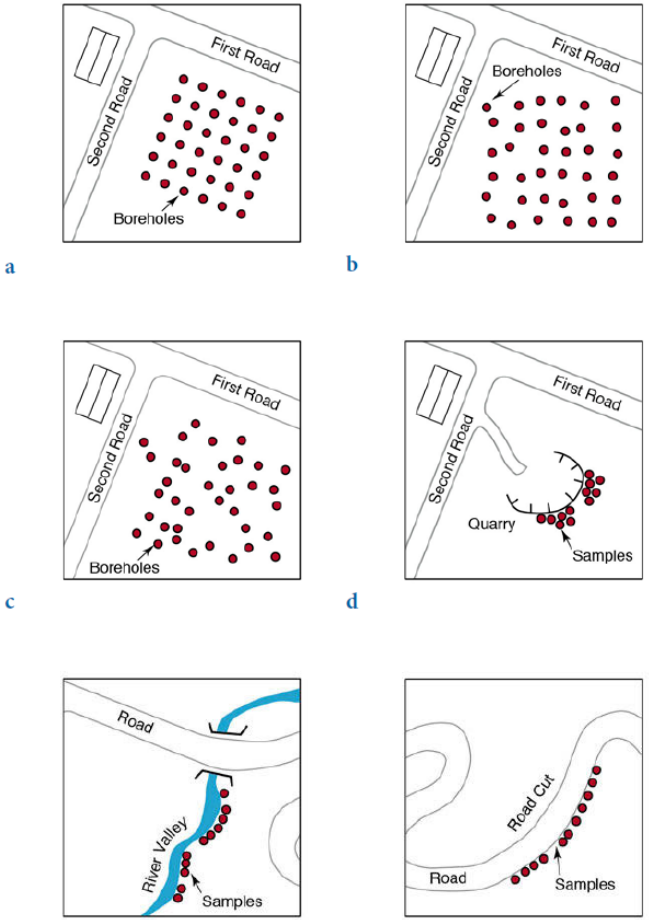
Am Ende der Stunde werden die Teilnehmer:

- … grundlegende Begriffe der univariaten Statistik und Datenanalyse kennen.
- … in Python statistische Momente bestimmen können.
- … empirische Verteilungen charakterisieren können.

Geostatistik – auf Mathematik basierende Methoden zur Analyse quantitativer Daten mit Raumbezug (Geodaten)

Geodaten

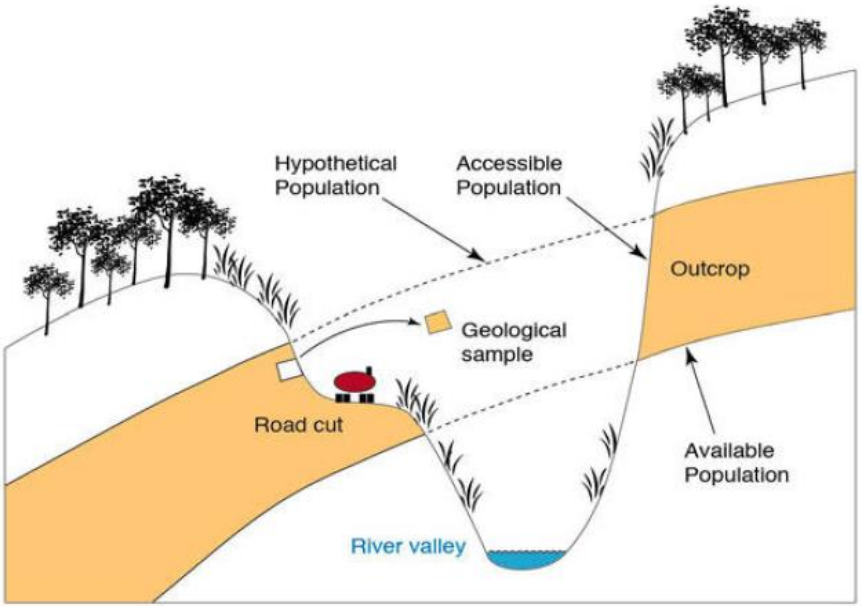
- Geschätzt 80% aller Daten haben einen Raumbezug!
- Sammeln von Daten (Feld, Labor, Satellitendaten, usw.)
  - Begrenzte Probenzahl (n)
  - Messunsicherheit



Trauth (2015) Fig. 1.2

Geodaten

- Grundgesamtheit und Stichprobe
  (engl. population, sample)



Trauth (2015) (Fig. 1.1)

Typische Arten von Geodaten

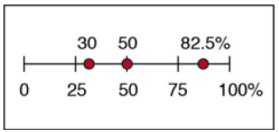
Nominale Daten
(nominal data)

Cyclotella ocellata
C. meneghiniana
C. ambigua
C. agassizensis
Aulacoseira granulata
A. granulata var. curvata
A. italica
Epithemia zebra
E. sorex
Thalassioseira faurii

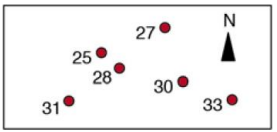
Ordinale Daten
(ordinal data)

1. Talc
2. Gypsum
3. Calcite
4. Flurite
5. Apatite
6. Orthoclase
7. Quartz
8. Topaz
9. Corundum
10. Diamond

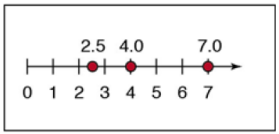
Geschlossene Daten
(closed data)



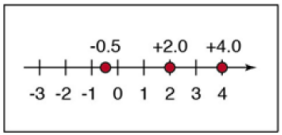
Räumliche Daten
(spatial data)



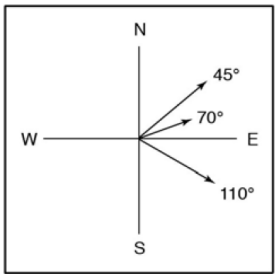
Verhältnisdaten
(ratio data)

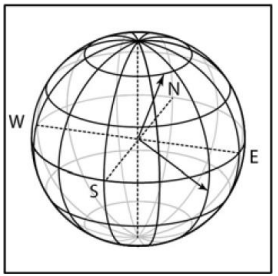


Intervalldaten
(interval data)



Richtungsabhängige Daten
(directional data)





Trauth (2015) Fig. 1.3

Weitere Grundbegriffe

- Messgröße (measured variable)
- Zufallsgröße (random variable)
- Diskrete Daten, bzw. Funktionen (discrete data)
- Stetige Daten, bzw. Funktionen (continuous data)
- Parameter
- Variable
- Freiheitsgrade (degrees of freedom)
- Wahrscheinlichkeit (probability)
- Unsicherheit (uncertainty)

Diskrete Daten

Stetige Daten

Methoden zur Datenanalyse

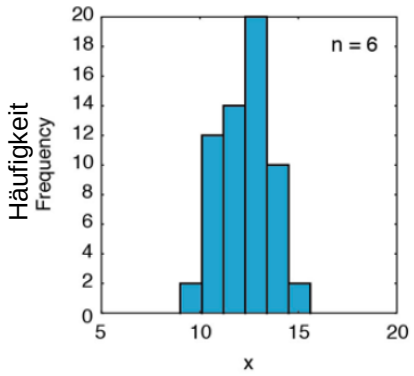
- Univariate Methoden: eine unabhängige Messgröße
- Bivariate Methoden: zwei abhängige Messgrößen
- Multivariate Methoden: mehrdimensionale Datensätze
- Zeitreihenanalyse: Datenwerte als Funktion der Zeit
- Räumliche Analyse: Daten mit Koordinaten in 2D oder 3D

- … und viele mehr.

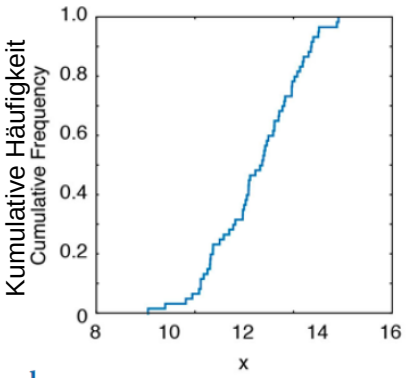
Beschreibende Statistik

- Statistische Charakterisierung von Stichproben
- Empirische Verteilungen von Messwerten (empirical distribution)
- Graphische Darstellung

Empirische Häufigkeitsverteilung



Kumulative Häufigkeitsverteilung



Trauth (2015) (Fig. 3.1)

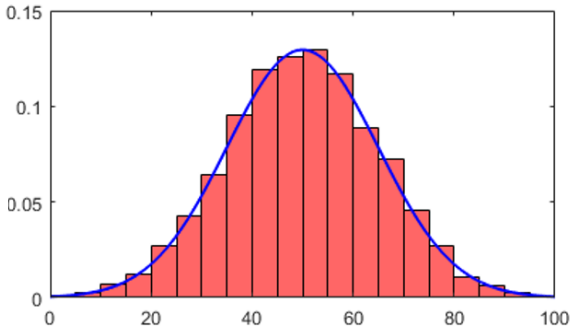
- … auch deskriptive Statistik genannt (descriptive statistics)

Schließende Statistik

- Analyse der Grundgesamtheit
- Theoretische Verteilungen von Zufallsvariablen (theoretical distribution)

Empirische Häufigkeitsverteilung

Theoretische (angepasste) Wahrscheinlichkeitsverteilung



- … dazu später mehr!

Charakterisierung von Stichproben

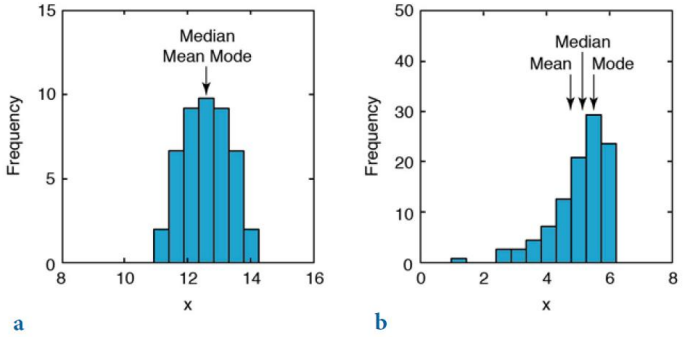
Statistische Parameter (statistical measures)

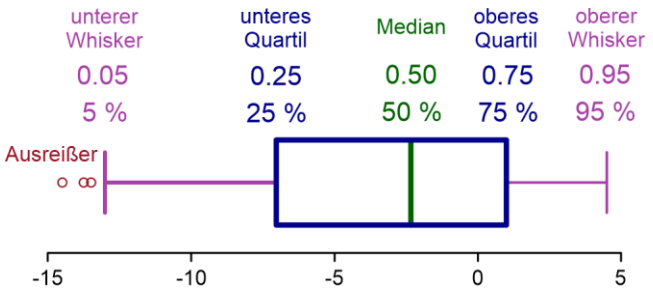
Trauth (2015) (Fig. 3.2)

1. Lageparameter (central tendency)

- Arithmetisches Mittel (mean)
- Geometrisches Mittel
- Harmonisches Mittel

- Median (median)
- Modus (mode)
  - Nur für diskrete, bzw. nominale Daten!
- Quartile, Quantile, Perzentile
- usw.

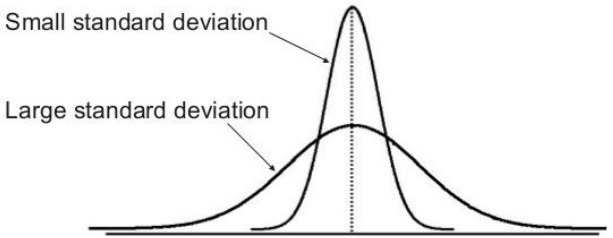


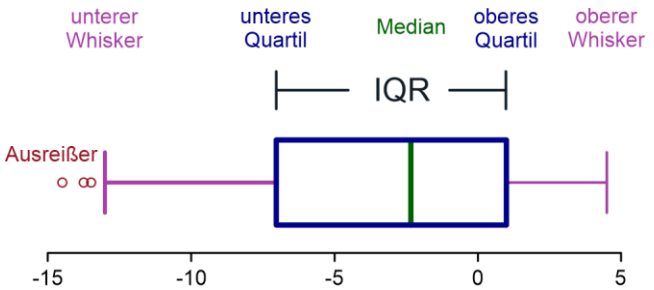


Charakterisierung von Stichproben

2. Streuungsmaß (dispersion)

- Range (Maximum – Minimum)
- empirische Varianz (σ^2)
- empirische Standardabweichung (σ)
- (Inter)Quartilabstand (IQR)
- usw.

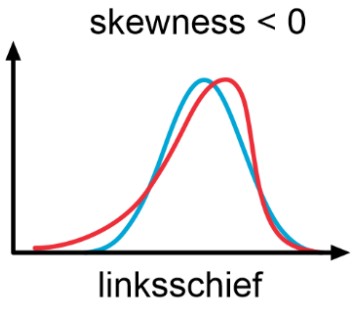


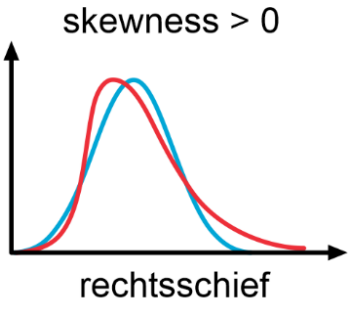


Charakterisierung von Stichproben

3. Schiefe (skewness)

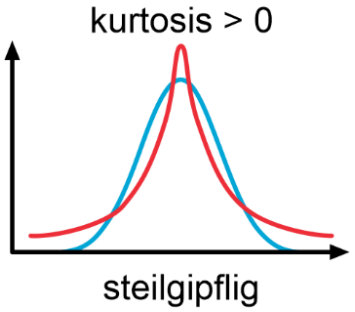
- nach Pearson
- nach Fisher
- Quantil-basiert
- usw.

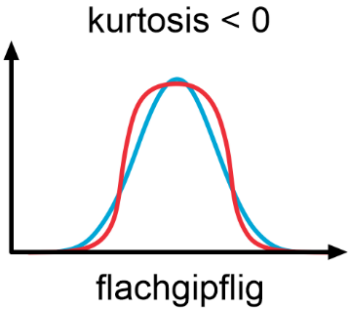




4. Wölbung (kurtosis)

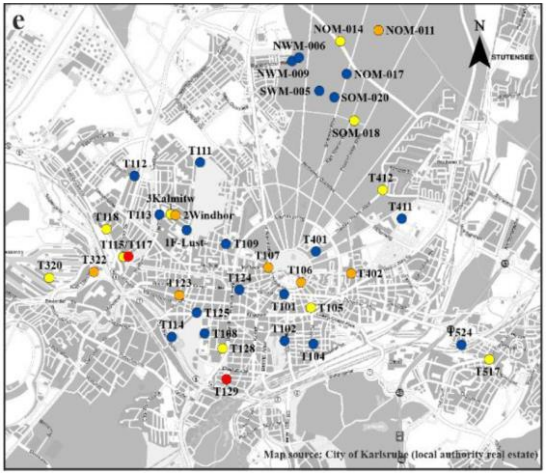
- Im Vergleich zu einer Normalverteilung
- nach Fisher





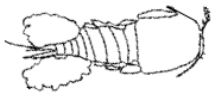
Übung 1: Univariate Statistik

- Grundwasserdatensatz Karlsruhe
  - Messwerte zu Grundwassertemperatur, hydrochemische Parameter, faunistische Daten
  - Datentypen
  - Bestimmung statistischer Parameter
- Aufgaben in Jupyter Notebook: 01_Univariate_Statistik_loesung



Koch et al. (2020) HESS-D

Crustaceen



Oligochaeten

Aufgabenbesprechung

Parameter	Datentyp
Pegel	Nominale Daten
Tiefe	Verhältnisdaten (stetig)
Sauerstoff	Verhältnisdaten (stetig)
Temperatur	Intervalldaten (stetig)
Elektrische Leitfähigkeit	Verhältnisdaten (stetig)
pH Wert	Verhältnisdaten (stetig)
Eisen, Mangan, Phosphat, Nitrat	Verhältnisdaten (stetig)
Detritus, Sediment	Ordinale Daten (diskret)
Geologische Einheit, Flächennutzung	Ordinale Daten (diskret)
Anzahl Arten, Anzahl Individuen	Verhältnisdaten (stetig)
Anteil Crustaceen, Anteil Oligochaeten	Geschlossene Daten (stetig)

Aufgabenbesprechung

Variable	Python-Datentyp
GWT	list
n	int

Variable	Wert
arithm. Mittel	13.5
mean	13.5
Median_1	14.1
Median_2	14.0
Mode (Geologie) (nur für diskrete Daten!)	4
Quartile	[11.4, 14. 0, 15.0]
Range	7.0
IQR	3.6
Standardabweichung	2.11
Varianz	4.45

Aufgabenbesprechung

Variable	Python-Datentyp
Skewness Pearson	-0.27
Skewness Fisher	0.29
Kurtosis	3.53

… noch Fragen?

Pause

… bis 11:25 Uhr

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:20	Univariate Statistik
10:20 – 11:00	Übung
11:00 – 11:10	Diskussion und Reflexion
11:10 – 11:25	Pause
11:25 – 11:45	**Statistisches Testen**
11:45 – 12:20	Übung
12:20 – 12:30	Diskussion und Reflexion

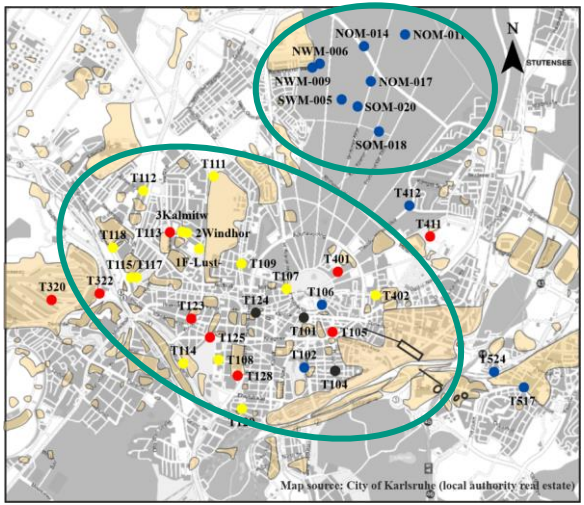
Lernziele statistisches Testen

Am Ende der Stunde werden die Teilnehmer :

- … die theoretischen Grundlagen des klassischen statistischen Testens kennen.
- … verschiedene statistische Tests für unterschiedliche Zwecke in Python kennen und anwenden können.
- … die Testergebnisse in Bezug auf Signifikanz und p-Wert bewerten und kritisch diskutieren können.

Problemstellung

- Grundwassertemperaturen in Karlsruhe und im Hardtwald:



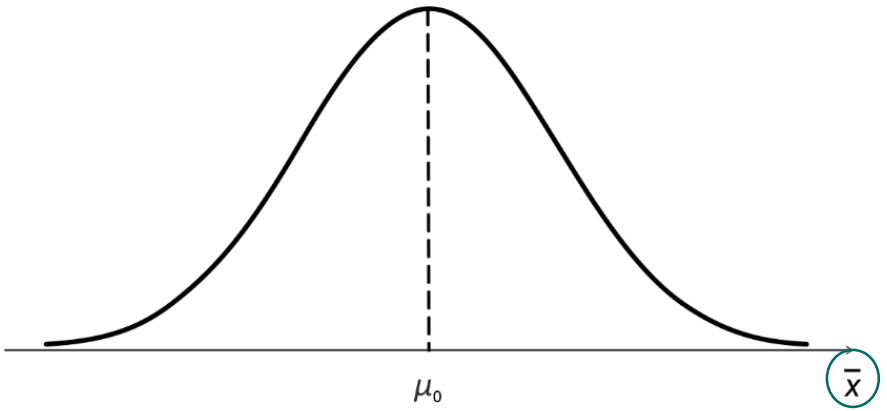
- … passt unsere ursprüngliche Vermutung, dass die Temperatur im Wald 11°C beträgt?
- … sind die Temperaturen in der Stadt höher, oder doch eher gleich?

Klassisches statistisches Testen

- Aufstellen einer Hypothese
  - z.B. „Die mittlere Grundwassertemperatur im Hardtwald beträgt 11°C“.
- Prüfen der Hypothese
  - Vergleich von dem was man sieht, mit dem was man beobachten würde, wenn die Hypothese stimmt.
  - Je besser die Beobachtung zur Hypothese passt, desto eher wird man ihr vertrauen
- Eine Hypothese kann nicht endgültig bestätigt oder widerlegt werden!
- … wir können uns aber dafür entscheiden sie anzunehmen oder abzulehnen

Ablauf eines statistisches Tests

- ***Nullhypothese H_0***: „Die mittlere Temperatur beträgt 11°C.“
- Hypothetischer Wert $\mu_0 = 11$
- Alternative Hypothese: „Die mittlere Temperatur beträgt **nicht** 11°C.“

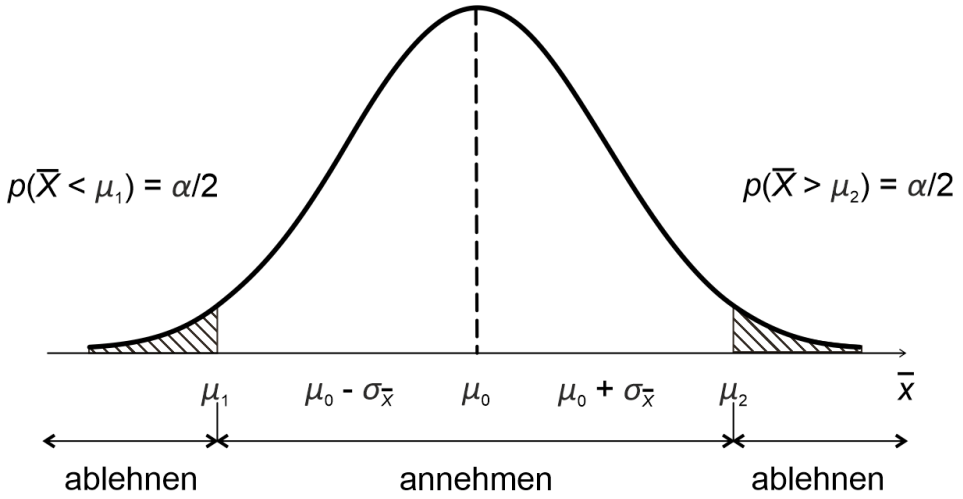


Mittelwert der Stichprobe

Tschirk (2014)

Ablauf eines statistisches Tests

- Definition eines Annahme, bzw. Ablehnungsbereichs
- Bedingte Wahrscheinlichkeit, dass wir H_0 ablehnen, obwohl H_0 stimmt
- Signifikanzniveau α (oftmals 0.05 oder 0.01)



Tschirk (2014)

Vier mögliche Ergebnisse

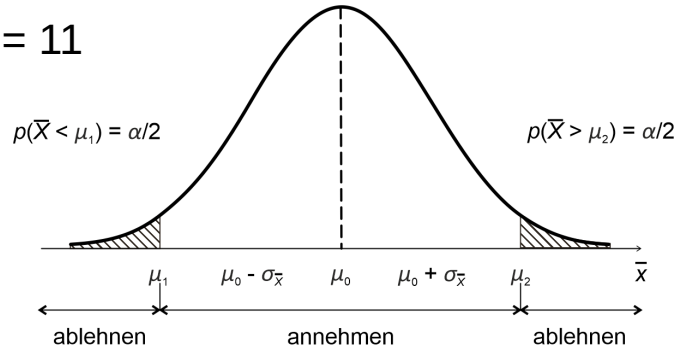
	H_0 ist richtig	H_0 ist falsch
H_0 wird angenommen	richtig entschieden	β-Fehler
H_0 wird abgelehnt	α-Fehler	richtig entschieden

Tschirk (2014)

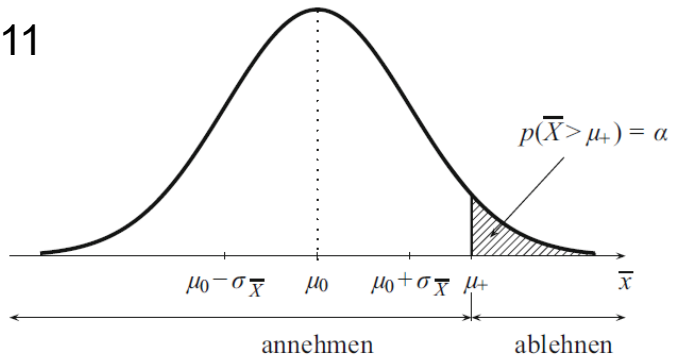
- 2 mögliche Fehler:
  - Ablehnen einer richtigen Nullhypothese: α-Fehler oder Fehler 1. Art
  - Annehmen einer falschen Nullhypothese: β-Fehler oder Fehler 2. Art

Ein- und zweiseitige Tests

- Zweiseitig: Nullhypothese $\mu_0 = 11$



- Einseitig: Nullhypothese $\mu_0 \leq 11$



Tschirk (2014)

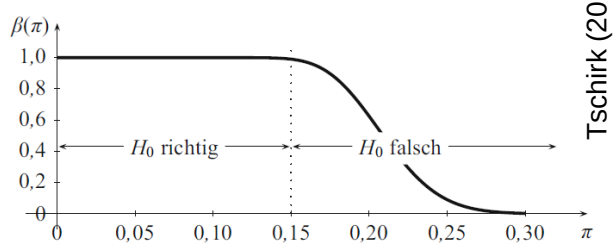
Interpretation des Testergebnisses

- Signifikanzwert, *p*-Wert
- Wahrscheinlichkeit den beobachteten Wert zu erhalten, unter der Bedingung dass H_0 stimmt
- Ablehnen der Nullhypothese, wenn *p*-Wert ≤ α
  - Annahme der alternativen Hypothese
  - „statistisch signifikant“ = „überzufällig“
- Ermöglicht Vergleich verschiedener Testergebnisse

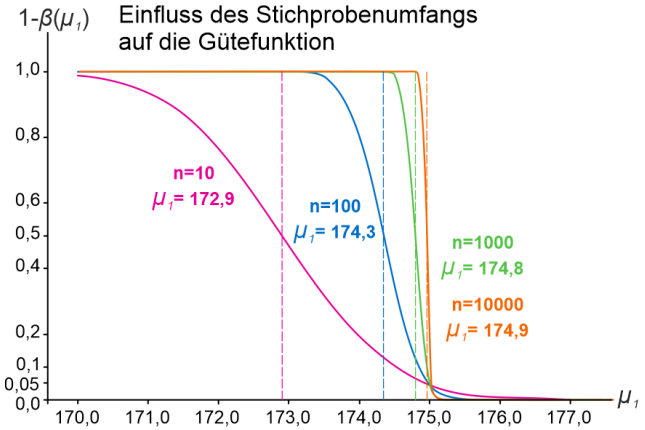
- **Gibt keine Aussage über die Größe des wahre Effekts**
- **Sagt nicht aus wie wahrscheinlich die Nullhypothese ist**

Trennschärfe eines Tests

- β: Wahrscheinlichkeit, dass H_0 korrekterweise abgelehnt wird



Tschirk (2014)



- Funktion 1 – β: auch Güte, Stärke, engl. power
- Abhängig von Anzahl der Proben

Übersicht einiger typischer Tests

Student‘s t-test (one sample)	Mittelwert einer Verteilung entspricht einem bestimmten Wert
Student‘s t-test (two sample)	Mittelwert zweier Verteilungen sind identisch
F-Test	Vergleicht die Varianz zweier Proben
Mann-Whitney U-Test	Differenz des Median zweier Verteilungen
Shapiro-Wilk Test	Test auf Normalverteilung

- … viele mehr!

Parametrische und nicht-parametrische Tests

- Parametrische Tests setzen eine Normalverteilung der Stichproben voraus → Überprüfen!
- ggfs. müssen Datensätze normalisiert, bzw. standardisiert werden
- Parametrische Test:
  - Student‘s t-test
  - F-test
  - Analysis of Variance (ANOVA)
  - …
- Nicht-Parametrische Test:
  - Mann-Whitney U-test
  - …

Limitierungen statistischer Tests

- Eine Hypothese kann nicht endgültig bestätigt oder widerlegt werden
- Prüfung der Übereinstimmung von Stichprobe und Hypothese
- Der Test bevorzugt die Nullhypothese (kleinere *p*-Werte)
- Die Nullhypothese muss von der Stichprobe unabhängig sein

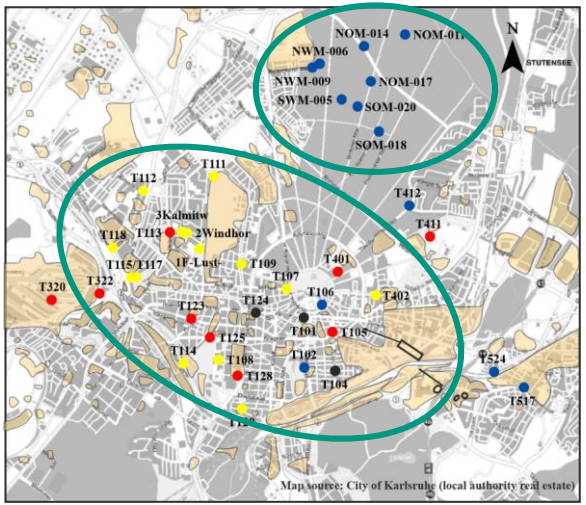
… was ein Test nicht kann

12 Missverständnisse zu *p*-Werten (Goodman, 2008):

- 1. mit $p = 0.05$ hat die Nullhypothese eine Chance von 5% wahr zu sein.
- 2. ein nicht-signifikanten Unterschied ($p > 0.05$) bedeutet, dass kein Unterschied zwischen den Gruppen besteht.
- 3. ein statistisch signifikantes Ergebnis ist wissenschaftlich bedeutsam.
- 7. $p = 0.05$ und $p \leq 0.05$ bedeuten das Gleiche.
- …

Übung 2: Statistisches Testen

- Grundwasserdatensatz Karlsruhe
  - Hypothesen testen
  - Verschiedene Tests
  - p-Werte bestimmen



- Aufgaben in Jupyter Notebook:
  02_Statistisches_Testen_loesung

Aufgabenbesprechung

- Hypothese 1:
  - Temperatur im Wald normalverteilt → H_0 annehmen
  - Temperatur = 11°C? → H_0 nicht annehmen (p = 0.005)
  - Mittelwert Temperatur = 10.7°C, n = 8 → Trennschärfe!
- Hypothese 2:
  - Sauerstoffsättigung normalverteilt → H_0 annehmen
  - F = 1.79, p = 0.12 → H_0 annehmen
  - T = 3.46, p = 0.0007, und T > T_kritisch → H_0 ablehnen
- Mann-Whitney U-test:
  - z.B. Phosphat, nicht normalverteilt, gleiche Verteilung in Wald und Stadt

Literatur

- Trauth (2015): MATLAB Recipes for Earth Sciences (4th Ed.), Springer
- Koch et al. (2020) Groundwater fauna in an urban area: natural or affected?, Hydrology and Earth System Sciences Discussions
- Tschirk (2014) Statistik: Klassisch oder Bayes, Springer
- Steve Goodman (2008) A Dirty Dozen: Twelve P-Value Misconceptions

Nützliche Links:

- https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/


thank you
danke
gracias
merci
obrigado
teşekkür ederim
go raibh maith agat
mochchakkeram
dziękuję
bedankt
sukriya
kop khun krap
arigatō
dakujem
grazie
terima kasih
спасибо
мерси