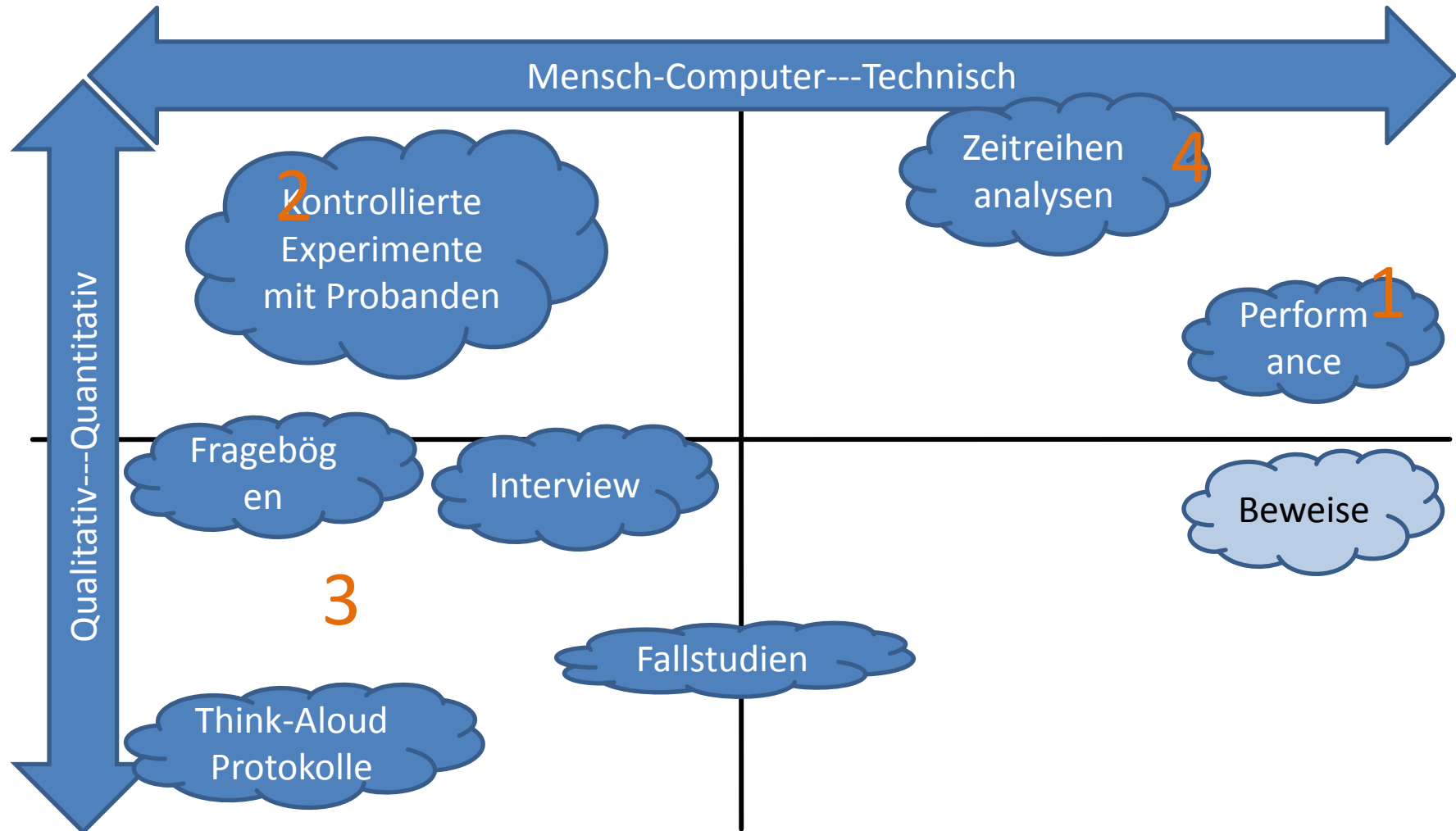


Performance Messungen

Einordnung



Aufgabe

- Bestimmen Sie die schnellste Sortierfunktion
 - Gruppe 1: Mergesort vs. Quicksort
 - Gruppe 2: Quicksort Rekursiv vs. Quicksort Iterativ
 - Gruppe 3: Quicksort Java vs. Quicksort C
 - Gruppe 4: Quicksort C++ vs. Quicksort Haskell
- Stellen Sie die Ergebnisse mit einem Poster vor
- Diskutieren Sie die Ergebnisse. Vertrauen Sie den Ergebnissen der anderen Teilnehmer?

Lernziele

- Aussagekräftige Performance-Analysen selbst durchführen können
- Performance-Analysen bewerten können



Warum Performanceanalyse?

- Alternativen vergleichen
- Einfluss eines Features
- System Tuning
- Relative Performance erkennen (über Zeit)
- Absolute Performance für ausgewählte Fälle
- Erwartungen setzen
- Analyse von Systemverhalten

Analysetechniken

- Messen
 - keine vereinfachenden Annahmen
 - i.d.R. am glaubwürdigsten
 - inflexibel, spezielles System
- Simulation
 - Abstraktion
 - Flexibel
- Analytisches Modellieren
 - Mathematische Beschreibung des Systems
 - Starke Abstraktion, i.d.R. kaum glaubwürdig
 - Insbesondere zur frühen Validierung

Benchmark

- Ausführen realer Programme/Hardwarekomponenten in realen Umgebungen (keine analytische Simulation)
- Messen von Performance, Speicherverbrauch, usw.
- Automatisierbar
- Kein menschlicher Einfluss

Benchmark - Beispiele

- 3DMark (Grafikkarte/System)
- TCP-H (Datawarehouse)
- TCP-C (OLTP)
- Sintel (Video-Encoder)

Was messen?

- Ausführungszeit
- CPU-Zyklen
- MIPS (Million instructions per second)
- MFLOPS (Million floating-point operations per second)
- SPEC (System Performance Evaluation Cooperative)
- QUIPS (Quality improvements per second)
- Transaktionen pro Sekunde

Aufgabe

- Welche Kriterien sollte eine gute Metrik erfüllen?
- Sind die vorgestellten Metriken gute Metriken nach Ihren Kriterien?

Kriterien

Kriterium	Execution Time	CPU Zyklen	MIPS	MFLOPS	SPEC	QUIPS	Transactions/second
Linearität	+	-	-	+	-	+	+
Reliabilität	+	-	-	-	-	-	+
Wiederholbarkeit	+	+	+	+	+	+	+
Einfache Messbarkeit	+	+	+	+	+	+	+
Konsistenz	+	+	-	-	+	+	+
Unabhängigkeit	+	+	+	-	-	+	+

Beispiel für Prüfungsfrage: Welche Metrik(en) würden Sie benutzen, um den schnellsten Sortieralgorithmus zu bestimmen?

Störvariablen

- Beeinflussen das Messergebnis systematisch oder unsystematisch
- Beispiele:
 - Hintergrundprozesse
 - Hardwareunterschiede
 - Temperaturunterschiede
 - Eingabedaten, zufällig?
 - Heap-Size
 - Hardware-Plattform
 - System-Interrupts
 - Parallelität in Single- und Multicore-Systemen
 - Garbage Collection

Aufgabe

- Wie kann man diese Störvariablen kontrollieren?

Typisches Vorgehen: Bester Wert

- Wiederholen
- Bester, zweitbesten oder schlechtesten Wert
- Bsp: Antwortzeiten für Programmieraufgabe
- R: Daten einlesen
 - `data <- read.csv("rt.csv", header=TRUE, sep = ";", dec = ".")`
 - header: gibt an, ob Variablen/Spaltennamen in der ersten Zeile stehen
 - sep: Separator für Datensätze in der selben Zeile
 - dec: Dezimaltrennzeichen
 - `min(data)/max(data)`

Typisches Vorgehen: Mittelwert

- Messung wiederholen
- Mittelwert bilden

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- R:
 - mean(data)

Median

- Wert, der in der Mitte liegt
- Robust gegen Ausreißer
- R:
 - `median(data)`
- Bei gerader Anzahl an Messwerten:
 - Arithmetisches Mittel der beiden mittleren Werte
 - Einen der beiden mittleren Werte angeben

Median oder Mittelwert?

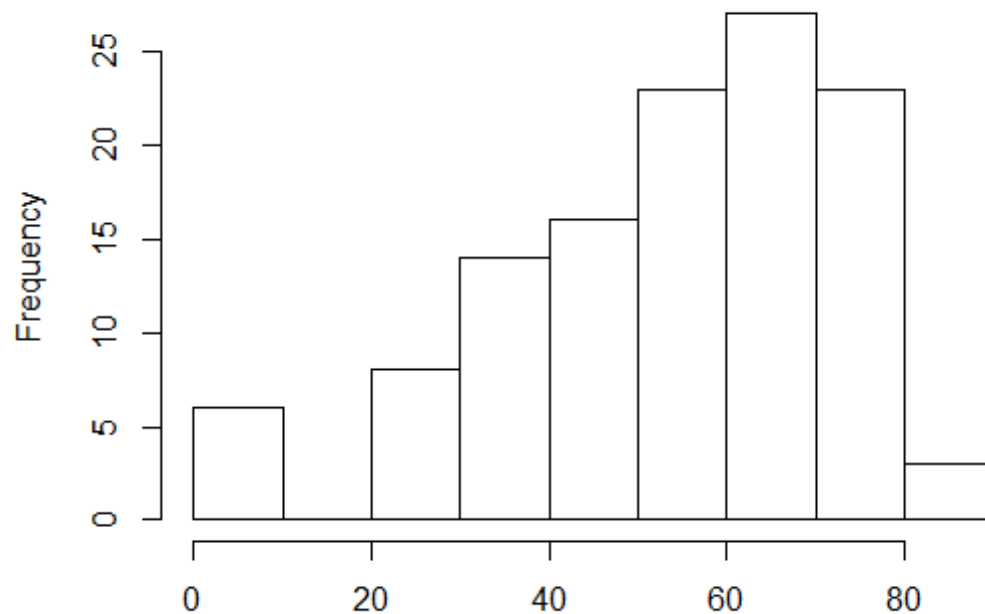
- Median statt arithmetisches Mittel, wenn
 - Ordinale Daten*
 - Wenig Messwerte
 - Asymmetrische Verteilung
 - Ausreißer
- *Skalenniveaus
 - Nominal (z.B. Geschlecht)
 - Ordinal (z.B. Platzierungen)
 - Metrisch (z.B. Temperatur, Antwortzeit)

Daten anschauen

- Überblick verschaffen
- Verteilung und Ausreißer einschätzen

Histogramme

- Häufigkeit von Messwerten in festgelegten Bereichen



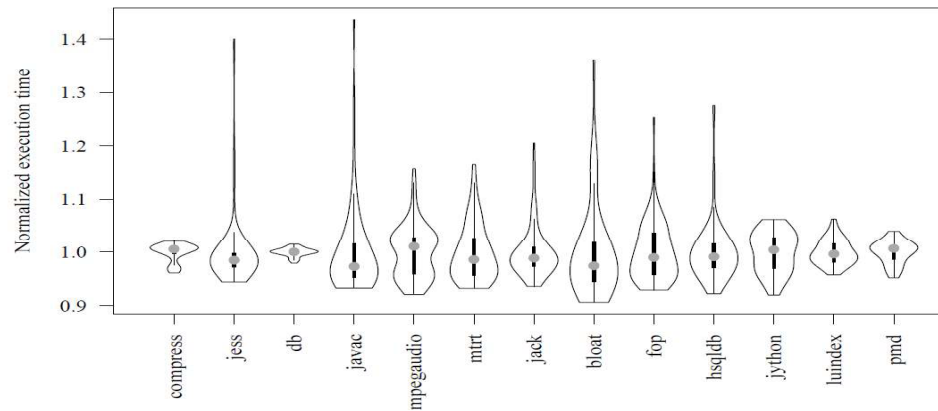
R: `hist(data)`

Boxplots

- Boxplot zeigt
 - Median als breite Linie
 - Quartile als Box (50% aller Werte in der Box)
 - Whiskers
 - Ausreisser als Punkte
- Graphische Darstellung von Verteilungen
- R: `boxplot(data)`

Violin-Plot

- Zeigt zusätzlich zu Boxplot die Verteilung der Daten



R: library(vioplot)
vioplot(data)

Recap

- Benchmarks
- Metriken zur Performance-Messung
- Daten visualisieren

Messmodel

- $y = \tau + \varepsilon$
- y : beobachteter Wert
- τ : wahrer Wert
- ε : Fehler
- Population: griechische Buchstaben
- Stichprobe: deutsche Buchstaben

Fehlermodell

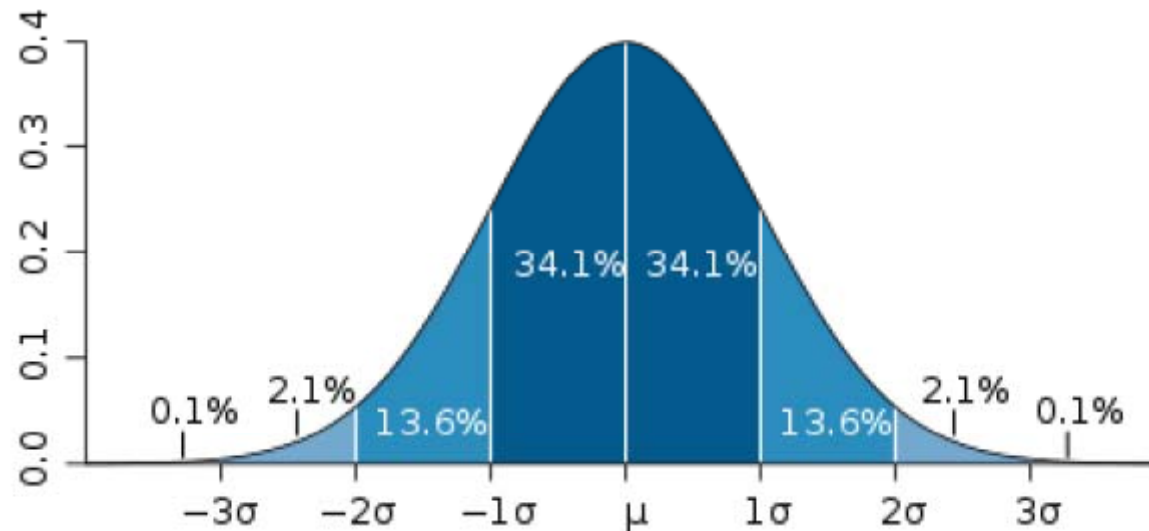
- Echter Mittelwert: 10
- 1 zufälliger Fehler, Einfluss ± 1
- Messwerte: 9 (50%) und 11 (50%)
- 2 zufällige Fehler, je ± 1
- Messwerte: 8 (25%), 10 (50%) und 12 (25%)
- 3 zufällige Fehler, je ± 1
- Messwerte: 7 (12.5%), 9 (37.5%), 11 (37.5%), 13 (12.5%)
- N zufällige Fehler, je ± 1
- Normalverteilung

Normalverteilung



Standardabweichung

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$



Standardabweichung: Anwendung

- Ausreißer definieren
- Hochbegabung definieren
- Entdeckung des Higgs-Boson verkünden

Konfidenzintervall

$$\bar{x} - z_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}}; \bar{x} + z_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$$

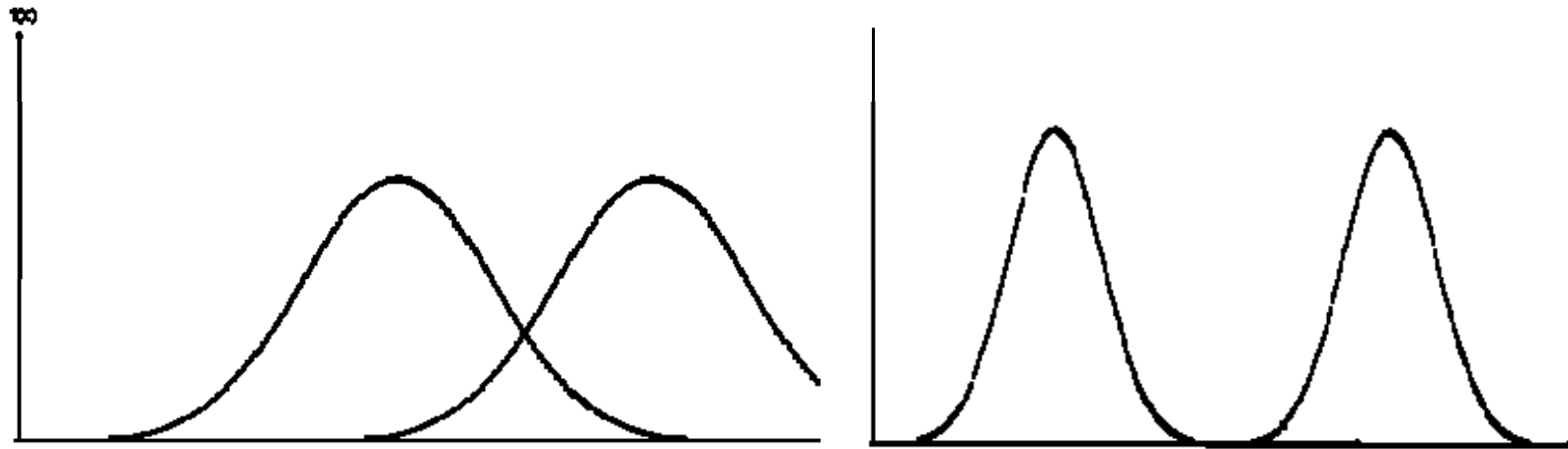
- \bar{x} : Mittelwert
- z : x-Wert der Standardnormalverteilung
- α : 1-Wert des Konfidenzintervall (z.B. 95%)
- s : Standardabweichung
- n : Anzahl der Messungen

Konfidenzintervall: Bedeutung

- Bedeutung
- Vertrauensintervall
- Wahrer Mittelwert liegt in 95% im Intervall
- Technischer: Bei grosser Anzahl von Wiederholungen des Experiments liegt in 95% der Fälle der wahre Mittelwert in dem jeweils berechneten Konfidenzintervall

Konfidenzintervall: Anwendung

- Überlappung: vmtl kein Unterschied
- Keine Überlappung: vmtl Unterschied



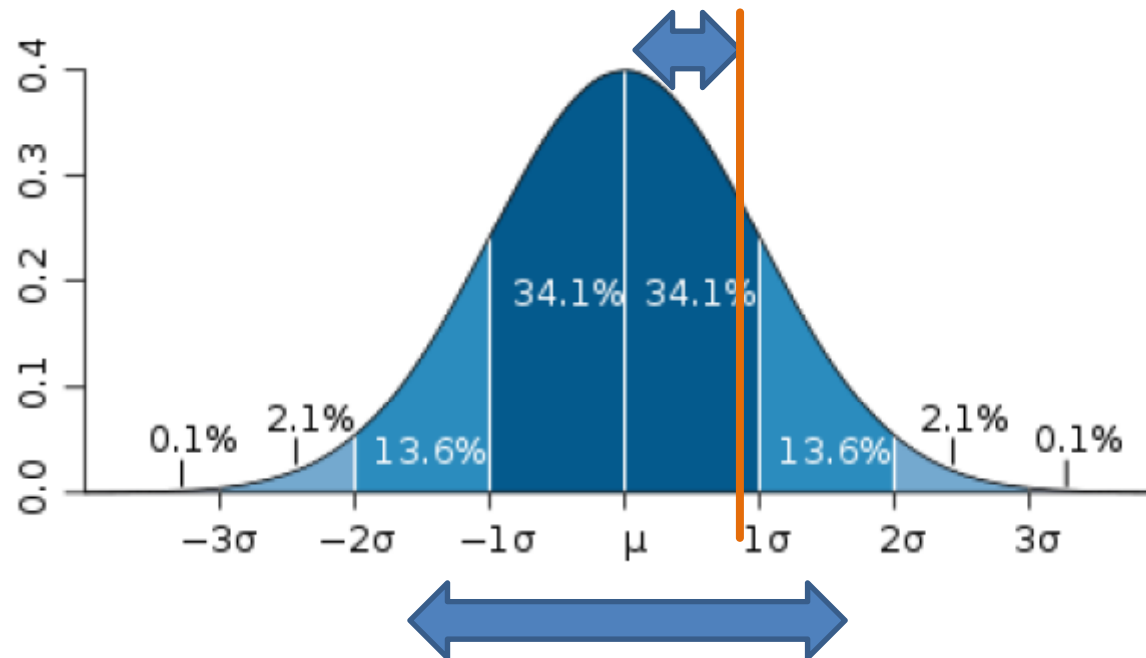
- Mehr Messungen verkleinert Intervalle

Genauigkeit vs. Präzision

Genauigkeit:

Abweichung beobachteter Mittelwerte vom wahren Mittelwert

Wichtig bei Zeitmessungen



Präzision:

Streuung um Stichprobenmittelwert

Ursache von Messfehlern unklar

Zufällige vs. Systematische Fehler

- Systematische Fehler: Fehler des Experiments/der Messmethode
 - CPU Speed: Messung bei unterschiedliche Temperaturen
 - Zustand nicht zurückgesetzt für zweite Messung
 - Geringe Varianz, bis konstant über alle Messungen
 - Im Design ausschließen, braucht Erfahrung
 - Genauigkeit
- Zufällige Fehler
 - Nicht kontrollierbar
 - Stochastische Methoden
 - Präzision

Signifikanztests

- Rigoroser als Vergleich von Konfidenzintervallen
- Zur Evaluierung, ob Messreihen unterschiedlich sind
- Z.B. t-Test

T-Test

- Entwickelt von Student (William Sealy Gosset)
- Vergleich von 2 Messreihen

Nullhypothese (H_0)	Alternativhypothese (H_1)
Statistische Hypothese	
Messreihen sind gleich, i.e., Daten von beiden Messreihen stammen aus der selben Population	Daten beider Messreihen stammen aus unterschiedlichen Populationen
Formal: $H_0 : \bar{x}_1 = \bar{x}_2$	Formal: $H_1 : \bar{x}_1 \neq \bar{x}_2$

T-Test: Ergebnis

- Bestimmt Wahrscheinlichkeit, das beobachtete Ergebnis unter Annahme der H_0 zu erhalten -> bedingte Wahrscheinlichkeit
- Wenn Wahrscheinlichkeit kleiner ist als:
 - 0.001 sehr sehr signifikant
 - 0.01 sehr signifikant
 - 0.05 typisches Signifikanniveau
 - 0.10 oft bei explorativen/initialen Untersuchungenmuss Nullhypothese falsch sein
- Signifikanzniveau
 - Vorher definieren!

T-Test: Aussage

- Was bedeutet signifikantes Ergebnis?
- Ist Nullhypothese falsch? -> Nein
- Ist Alternativhypothese richtig? -> Nein

- Kein Gegenbeweis für Gültigkeit der Nullhypothese gefunden
- Aufschreiben:
 - Ablehnen/nicht ablehnen der Nullhypothese
 - Nie: Bestätigen der Null-/Alternativhypothese

T-Test: Berechnung von Hand (1)

- Berechnung der Kenngröße

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)}}$$

Datensatz (ProgramComprehensionRT):
t = 1.477

$$\hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

T-Test: Berechnung von Hand (2)

- Freiheitsgrade (Degrees of freedom, df)
 - für t-Test: $n_1 + n_2 - 2$ (hier: 11)

- Tabelle mit t-Verteilung (z.B. wikipedia)

$$t_{\alpha/2, df=11} = 2,201$$

- Vergleich mit beobachtetem Wert ($t_{\text{emp}} = 1.477$)
 - ist $t_{\text{emp}} > t_{\alpha/2, df=11}$?
 - nein, darum nicht signifikant

T-Test: Einseitig vs. Zweiseitig

- Zweiseitig:
 - Keine Kenntnisse über Richtung des Effekts (z.B., welches System schneller ist)
 - Signifikanzniveau halbieren
- Einseitig:
 - Vermutung, das ein System schneller ist
 - Signifikanzniveau muss nicht halbiert werden

$$t_{\alpha, df=11} = 1,796$$

T-Test: R

- `t.test(dataPC1, dataPC2)`

- Ausgabe:

```
Welch Two Sample t-test
```

```
data: dataPC1 and dataPC2
```

```
t = 1.5222, df = 10.566, p-value = 0.1573
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-5.095727 27.583584
```

```
sample estimates:
```

```
mean of x mean of y
```

```
50.74243 39.49850
```

- p-Wert: Bedingte Wahrscheinlichkeit, Ergebnis unter Annahme der H_0 beobachtet zu haben
- Wenn p-Wert kleiner als definiertes Signifikanzniveau ist, ist Ergebnis signifikant

T-Test: Varianten

- T-Test für unabhängige Stichproben:
 - Zusammensetzung der Stichproben ohne gegenseitige Beeinflussung
 - Z.B. zufällige Zuteilung von Probanden in einer oder andere Stichprobe
- T-Test für abhängige Stichproben:
 - Zusammensetzung einer Stichprobe hängt von Zusammensetzung anderer Stichprobe ab
 - Z.B.: Wiederholungsmessungen, zuteilen von Ehepartnern in unterschiedliche Stichproben

T-Test: Voraussetzungen

- Metrisches Skalenniveau
- Normalverteilte Daten (z.B. Shapiro-Wilk)
- Oder: $n \geq 30$

Mann-Whitney-U

- Nicht parametrischer Test
- Bei ordinalen Daten (oder nicht-normalverteilten metrischen Daten)
- Berechnung der Kenngröße:

$$U = n_1 \bullet n_2 + \frac{n_1(n_1 + 1)}{2} + T_1$$

$$T = \sum_{i=1}^n r_i \quad - r_i: \text{Rangplätze in der Stichprobe}$$

Lernziele

- Aussagekräftige Performance-Analysen selbst durchführen können
- Performance-Analysen bewerten können



Literatur

- David Lilja. *Measuring Computer Performance: A practitioner's guide*. Cambridge University Press. 2000
- Paper 1, 2 & 3
- Beliebiges Statistikbuch