

# Statistik für Digital Humanities

## Grundannahmen Parametrischer Verfahren

Dr. Jochen Tiepmar

Institut für Informatik  
Computational Humanities  
**Universität Leipzig**

09. Novemeber 2020

[Letzte Aktualisierung: 26/11/2020, 14:09]

# Grundannahmen Parametrischer Verfahren

- Parametrische Tests weitverbreitete Grundlage statistischer Arbeit
- Parametrische Tests gehen von verschiedenen Annahmen aus
- Annahmen bzgl. Daten nicht gegeben → Test unpassend
- → Kritisch für korrekte Auswahl von Tests
- → Einschränkung der Auswahl passender Methoden

# 4 Grundannahmen

- Normalverteilung
- Homogenität der Varianzen
- Mindestens Intervalldaten
- Unabhängigkeit

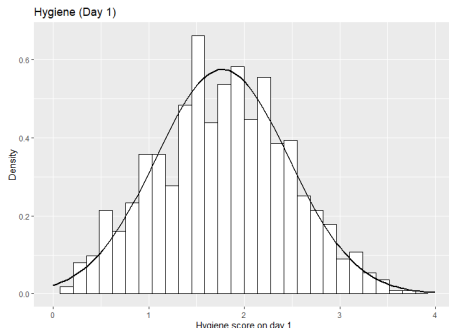
# 4 Grundannahmen

- Normalverteilung
  - Logik hinter Hypothesentests basiert meist (aber nicht immer) auf Normalverteilung (Bsp t-Test)
  - Keine Normalverteilung → Logik der Teststatistik fehlerhaft
- Homogenität der Varianzen
- Mindestens Intervalldaten
- Unabhängigkeit

# Berechnung von Normalität

- Visuell
- Vergleich von Eigenschaften der Normalverteilung (Verschiebung, Wölbung, . . . )
- Berechnung des Unterschiedes zu normaler Normalverteilung (Shapiro-Wilk Test)
- **Central Limit Theorem**
- → Wenn Stichprobe tendenziell normalverteilt dann Stichprobenverteilung ebenfalls  
if  $n > 30$ :
  - $\bar{X}_{\text{Stichprobenverteilung}} \approx \bar{X}_{\text{population}}$
  - Stichprobenverteilung tendenziell normalverteilt

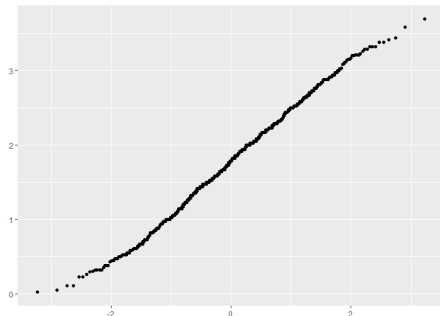
# Visuell mit Häufigkeitsverteilung



Vergleich mit Normalverteilung bei gleichem  $\bar{x}$  und  $s$

```
dlf<-read.delim("DownloadFestival.dat", header=TRUE)
dlfhistogram <- ggplot(dlf, aes(day1)) + ggtitle ("Hygiene (Day 1)")
  + xlim(0,4) + geom_histogram(aes(y=..density..), color="black", fill="white")
  + labs(x="Hygiene score on day 1", y="Density")
dlfhistogram + stat_function(fun=dnorm, args =
  list(mean = mean(dlf$day1, na.rm=TRUE), sd = sd(dlf$day1, na.rm=TRUE)))
```

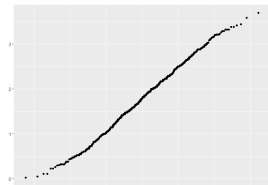
# Visuell mit Q-Q Plot



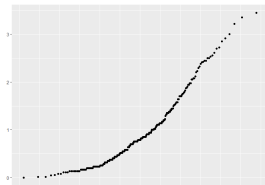
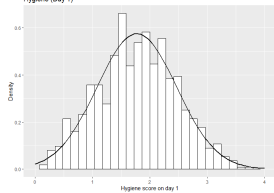
- Quantile-Quantile Plot zeichnet sortierte und kummulierte Werte der Datenverteilung gegen die einer Normalverteilung
- Je gerader die Linie desto normalverteilter die Daten

```
dlf<-read.delim("DownloadFestival.dat", header=TRUE)  
qqplot(sample=dlf$day1, stat="qq")
```

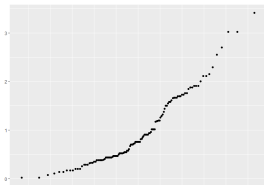
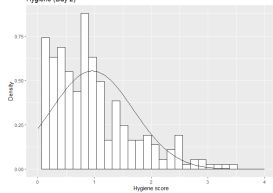
# Visuell



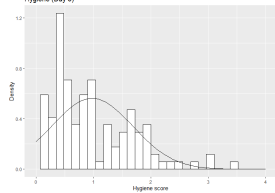
Hygiene (Day 1)



Hygiene (Day 2)



Hygiene (Day 3)





# Vergleich von Eigenschaften

- R Paket pastecs
- z-Scores  $\frac{skew}{2SE}$  und  $\frac{kurt}{2SE}$  zeigen signifikante Wölbung oder Verschiebung bei Werten
  - $< -1$  und  $> 1$  bei  $p = 0.05$
  - $< -1.29$  und  $> 1.29$  bei  $p = 0.01$
- Signifikanz nur bei kleinen Samples sinnvoll ( $< 200$ )

```
library(pastecs)
round(stat.desc(cbind(dlf$day1,dlf$day2,dlf$day3),basic=FALSE,norm=TRUE),digits=3)
      V1      V2      V3
median 1.790 0.790 0.760
mean   1.771 0.961 0.977
SE.mean 0.024 0.044 0.064
CI.mean.0.95 0.048 0.087 0.127
var     0.481 0.520 0.504
std.dev 0.694 0.721 0.710
coef.var 0.392 0.750 0.727
skewness -0.004 1.083 1.008
skew.2SE -0.026 3.612 2.309
kurtosis -0.422 0.755 0.595
kurt.2SE -1.228 1.265 0.686
normtest.W 0.996 0.908 0.908 <-- Ergebnisse des Shapiro-Wilk Test
normtest.p 0.032 0.000 0.000 <--
```

# Shapiro-Wilk Test

Shapiro, S.S. & Wilk, M.B. (1965): *An Analysis of Variance Test for Normality*

- Teststatistik zur Signifikanz der Abweichung der Daten von einer Normalverteilung
- Maximale Stichprobengröße: 50
- Generell je größer Stichprobe, desto mehr Typ 1 Fehler, deshalb zusätzlich visuelle Analyse sowie Skew und Kurtosis in Betracht ziehen
- $H_1$  Es liegt keine Normalverteilung vor
- $H_0$  Es liegt eine Normalverteilung vor
- **Achtung:** Der R Befehl `shapiro.test(data)` liefert nicht den eigentlichen Test sondern den von Patrick Royston (1982) für  $n > 50$
- p-Wert bei `shapiro.test(data)`  $< 0.05$  → Daten signifikant anders als Normalverteilung

# Shapiro-Wilk Test

## Berechnung

- X sortieren
- $W = \frac{b^2}{S^2}$ 
  - $b = \sum_{i=1}^k \alpha_i * (y_{n-i+1} - y_i)$
  - $S^2 = \sum x_i^2 - \frac{1}{n} * (\sum x_i)^2$
  - $k = \frac{n}{2}$  wenn n gerade,  $\frac{n-1}{2}$  sonst
  - $\alpha_i$  aus Shapiro-Wilk Tabelle ablesen (auf passendes n achten)
- Vergleiche  $W$  mit Grenzwert  $W_{kr}$  für 0.5-Level aus Tabelle

## Interpretation

- Wenn  $W > W_{kr}$  :  $H_0$  wahrscheinlich (Test findet keinen Hinweis gegen Normalverteilung)

# Shapiro-Wilk Test

Beispiel: Like/Dislike Verhältnis auf Youtube  $X = \{6, 1, -4, 8, -2, 5, 0\}$

- Sortiert:  $X = \{-4, -2, 0, 1, 5, 6, 8\}$
- $S^2 = \sum x_i^2 - \frac{1}{7} * (\sum x_i)^2 = 146 - 28 = 118$
- $b = 0.6233 * (8 + 4) + 0.3031 * (6 + 2) + 0.1401 * (5 - 0) = 10.6049$
- $W = 10.6049^2 / 118 = 0.9530$
- $\rightarrow W$  wesentlich größer als  $W_{kr}(0.928)$
- $\rightarrow$  Kein Beweis gegen Normalverteilung gefunden

Vergleich dazu das abweichende Ergebnis des R Skripts

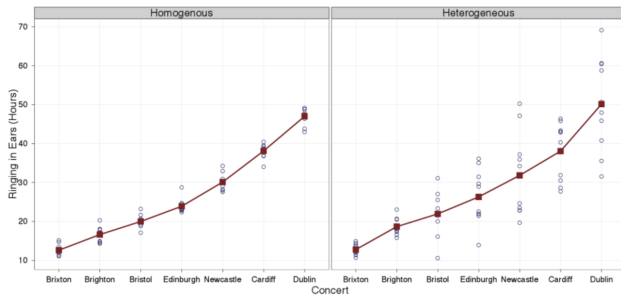
```
data <- c(6,1,-4,8,-2,5,0)
shapiro.test(data)
```

$W = 0.90428$ ,  $p\text{-value} = 0.3998$

# 4 Grundannahmen

- Normalverteilung
- Homogenität der Varianzen
  - Bei Gruppendesigns: Varianz einer Variable zwischen verschiedenen Gruppen sollte gleich sein
  - Messwiederholungsdesign: Varianz einer Variable sollte gleich bleiben bei Variation einer anderen → *Siehe VO Vergleich zweier Mittelwerte*
- Mindestens Intervalldaten
- Unabhängigkeit

# Homogenität der Varianz



# Berechnung von Homogenität der Varianz

- Bei Kontinuierlicher Messung: Visuelle Analyse
- Bei Gruppendesigns: Levene's Test, Hartleys Varianz-Ratio

# Levene's Test

Levene, H. (1960): *Robust tests for equality of variances*

- $H_0$ : Varianzen in verschiedenen Gruppen sind homogen / Der Unterschied ist nicht signifikant
- $p < 0.05$  :  $H_0$  ist nicht korrekt, signifikante Unterschiede zwischen den Varianzen verschiedener Gruppen
- Berechnung: One-Way Anova (Einweg-Varianzanalyse) → später
- Generell je größer Stichprobe ( $n \geq 50$ ), desto mehr Typ 1 Fehler –  $>$  Hartleys  $F_{max}$  ebenfalls anwendbar

```
library(car)
rexam <- read.delim("rexam_factor.dat",header=TRUE)
> leveneTest(rexam$exam, rexam$uni)
  Levene's Test for Homogeneity of Variance (center = median)
    Df F value Pr(>F)
group 1  2.0886 0.1516
    98
> leveneTest(rexam$numeracy, rexam$uni, center= mean)
...
group 1  7.3681 0.007846 **
    98
```



# Hartley's $F_{max}$ (Varianz-Ratio)

Pearson, E.S. & Hartley, H.O. (1954): *Biometrika tables for statisticians*

- $H_0$ : kleinster und größter Wert sind keine Outlier
- Test auch allgemein zur Outlier-Analyse geeignet
- $F_{max} = \frac{\max(s^2)}{\min(s^2)}$
- $F_{max} < \text{Kritischer Wert} \rightarrow \text{Test nicht signifikant} \rightarrow H_0 \text{ gilt} \rightarrow \text{Varianzen homogen}$

# Hartley's $F_{max}$ (Varianz-Ratio)

Varianz der Freundesanzahlen bei Facebook, StudiVZ, Steam, Friendster

–  $V = \{22, 40, 53, 57\}$

–  $n_{pergroup} = 10$

$H_0$ : kleinster und größter Wert sind keine Outlier

$min = 22$

$max = 57$

$F_{max} = \frac{\max(s^2)}{\min(s^2)} = \frac{57}{22} = 2.59 < \text{Kritischer Wert (6,31)} \rightarrow \text{Test nicht signifikant} \rightarrow H_0 \text{ gilt}$

$\rightarrow$  Varianzen homogen

$V = \{9, 40, 53, 57\}$  wäre mit  $F_{max} = 6.33$  nicht homogen

# 4 Grundannahmen

- Normalverteilung
- Homogenität der Varianzen
- Mindestens Intervalldaten
  - Daten sollten zumindest intervallskaliert sein
  - Ordnung der Werte & aussagekräftiger Abstand
  - Absoluter Nullpunkt optional
- Unabhängigkeit

# 4 Grundannahmen

- Normalverteilung
- Homogenität der Varianzen
- Mindestens Intervalldaten
- Unabhängigkeit
  - Variablenwerte unabhängig voneinander, beeinflussen sich nicht
  - Bei Messwiederholungsdesigns Variablenwerte verschiedener Probanden unabhängig voneinander

# Annahmen nicht gegeben

## Grundsätzlich 2 Möglichkeiten der Abweichung

- Daten passen nicht → Extremwerte, Outlier verzerren die Verteilung
- Testverfahren passt nicht → Alternativen möglich?

# Umgang mit Problemwerten

Zuallererst

- Daten auf offensichtliche (Tipp-)Fehler prüfen

Weitere Optionen nach umgekehrter Präferenz sortiert

- Problemfälle löschen
- Datentransformation
- Score des Problemfalls ändern

# Umgang mit Problemwerten

Weitere Optionen nach umgekehrter Präferenz sortiert

- Problemfälle löschen
  - Werte, die sehr wahrscheinlich nicht zur Population gehören kann man löschen
  - Katze hat gebellt → War wahrscheinlich ein verkleideter Hund
- Datentransformation
- Score des Problemfalls ändern

# Umgang mit Problemwerten

Weitere Optionen nach umgekehrter Präferenz sortiert

- Problemfälle löschen
- Datentransformation
  - Manche Analysen erlauben Datentransformationen (aller Werte!)
  - Bspw Relativer Abstand analysiert, aber absoluter Abstand egal → Umskalierung der Skala unproblematisch
  - Beispiele:
    - Log-Transformation  $\log X$  verkürzt rechten Tail der Verteilung, reduziert pos. skew & Varianz
    - Wurzel-Transformation  $\sqrt{X}$  bringt jeden Wert näher ans Zentrum, reduziert (pos.) skew & Varianz
    - Reziproke Transformation  $\frac{1}{X}$  normalisiert auf  $-1 \dots 1$ , reduziert Einfluss großer Werte (aber dreht Höhe der Werte um), reduziert pos. skew & Varianz (obviously)
    - Umgekehrter Score  $X_r = x_{max} - X$  oder  $x_{max} - X + 1$  erlaubt Korrektur von negativem Skew mit erwähnten Mitteln.
  - **Für Interpretation unbedingt wieder rückrechnen**
- Score des Problemfalls ändern



# Umgang mit Problemwerten

Weitere Optionen nach umgekehrter Präferenz sortiert

- Problemfälle löschen
- Datentransformation
- Score des Problemfalls ändern
  - Wert sehr unrepräsentativ → Ändern kleineres Übel
  - Nahester Score  $\pm 1$  Einheit
    - Reihenfolge bleibt, problematischer Abstand wird annulliert
  - Mittelwert  $\pm 3 * s$  (folgt aus z-Score)
  - Mittelwert  $\pm 2 * s$

# SuperGAU Handling

Wenn selbst Datenkorrektur nicht hilft oder zu "messy" wird:

- Gerade Normalität oft schwer objektiv bestimmbar
- Bootstrapping (Hochrechnen der Daten anhand gegebener Verteilung)
- Manche parametrische Tests gelten als **robust**, funktionieren also auch wenn nicht alle Annahmen erfüllt sind
  - Trimmed Mean → k kleinste und größte Werte löschen (k mit angeben)
  - M-Schätzer → k empirisch bestimmt
  - Bootstrap → Stichprobe in kleinere Proben mit Normalverteilung zerlegen, Stichprobenwerte abschätzen
- Konsequenzen von Transformationen eventuell schwerwiegender als ein Bruch mit den Annahmen
- Nichtparametrische Testverfahren haben keine Grundannahmen über die Daten, sind aber sehr eingeschränkt anwendbar
- Siehe R Package WRS und Wilcox, R.R. (2005): *Introduction to robust estimation and hypothesis testing*

# Zusammenfassung

- Parametrische Tests basieren auf 4 Grundannahmen über die Daten
  - Normalverteilung → Shapiro-Wilk Test
  - Homogenität der Varianzen → Levene und Hartley Test
  - Mindestens Intervallskalierung
  - Unabhängigkeit
- Wenn Annahmen nicht gegeben sind können folgende zunehmend unangenehme Reperaturmaßnahmen helfen
  - Daten auf offensichtliche (Tipp-)Fehler prüfen
  - Problemfälle löschen
  - Datentransformation
  - Score des Problemfalls ändern
- Wenns alles nix hilft
  - Nichtparametrische Tests
  - Robuste Tests
  - Schadensabschätzung
  - Kreative Argumentation

# Aktuelle Beispiele

Nick Redfern (2012): *The log-normal distribution is not an appropriate parametric model for shot length distributions of Hollywood films*

- Sind Analysen auf Basis einer Annahme einer Lognormal-Verteilungen bei der Betrachtung von Schnittlängen von Filmszenen wirklich angemessen?
- Gilt die Annahme der Lognormal-Verteilung hier?

Mike Baxter (2012): *On the distributional regularity of shot lengths in film*

- Welche methodischen Fehler hat Redfern (2012) begangen?