# Mann-Whitney-U-Test

Referat für PS Statistik im SS2020

PARAMETERFREIES PRÜFVERFAHREN

= STATISTISCHER TEST ZUR ÜBERPRÜFUNG, OB SIGNIFIKANTE UNTERSCHIEDE ZWISCHEN 2 GRUPPEN BESTEHEN

# Gliederung

- Anwendungsbereich
- Prinzip des Mann-Whitney-U-Test
- Durchführung des Tests an einem Beispiel

# Anwendung

- Vergleich zweier unabhängiger Stichproben
- Das besondere an dem Test ist, dass er bei abhängigen Variablen verwendet werden kann, die
  - **Ordinal skaliert** anstatt metrisch sind (d.h. wir haben eine kategorische Variable, bei der die Kategorien in eine Rangfolge gebracht werden können, die Abstände dazwischen aber nicht gleich groß sind; Mögliche Rechenoperationen sind $=, \neq$ und $<, >$ aber nicht $+, -, *, \div$)
  - **Nicht normalverteilt** sind (wir kennen entweder die Verteilung der Variable nicht oder die Stichprobe ist nicht groß genug, um mit dem Zentralen Grenzwertsatz eine Normalverteilung annehmen zu können; (Normalverteilung ist Voraussetzung bei parametrischen Testverfahren)

# Prinzip

- Problem :

Wie auf der vorherigen Folie angesprochen, kann mit ordinal skalierten Variablen nicht richtig gerechnet werden, weil die Abstände zwischen den auftretenden Werten nicht gleich groß sind (= **mangelnde Äquidistanz**).

z.B. Mittelwerte und Varianzen können nicht berechnet werden

- Lösung :

**Ersetzung** aller Werte **durch Rangplätze**, wobei die Abstände zwischen den Werten völlig vernachlässigt werden.

# Beispiel

- Untersucht wird die Fragestellung, ob sich Innsbrucker SchülerInnen und Studierende bei der Beurteilung der Funktionsfähigkeit des Corona-bedingten e-Learning Systems unterscheiden, woraus sich eine Ungleichheit der digitalen Kapazitäten unterschiedlicher Bildungsniveaus ( Schule und Universität ) ableiten ließe.

- $H_0$ = Zwischen SchülerInnen und Studierenden besteht kein Unterschied hinsichtlich der Beurteilung der Funktionsfähigkeit des e-Learning Systems.

- $H_1$ = Zwischen SchülerInnen und Studierenden besteht ein signifikanter (Irrtumswahrscheinlichkeit 0,05) Unterschied hinsichtlich der Beurteilung der Funktionsweise des e-Learning Systems, aus welchem... ließe.

# Beispiel

| Beurteilung des e-Learning | SchülerInnen | Studierende | $t_i$ | $(t_i^3 - t)$ |
|---|---|---|---|---|
| Gut | 5 | 7 | 12 | 1716 |
| Mittel | 4 | 5 | 9 | 720 |
| Schlecht | 6 | 3 | 9 | 720 |
| Gesamt | 15 = n1 | 15= n2 | 30 = N | $\sum$3156 |

# Ersetzung der Werte durch Ränge

- 1. Schritt: N = 30, d.h. wir haben 30 Ränge zu vergeben
- 2. Schritt: Den Beurteilungen Gut/Mittel/Schlecht werden Werte zugewiesen (Codierung), wobei Gut = 1, Mittel = 2, Schlecht = 3
- Insgesamt haben wir also 30 Beobachtungen, die alle den Wert 1, 2 oder 3 annehmen. Diese sind in einer Rangliste der Größe nach zu ordnen, danach vergeben wir die Ränge 1-30 an sie:

Rang

Beurteilung

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | - | 1 | 13 | - | 2 | 22 | - | 3 |
| 2 | - | 1 | 14 | - | 2 | 23 | - | 3 |
| 3 | - | 1 | 15 | - | 2 | 24 | - | 3 |
| … | - | 1 | … | - | 2 | … | - | 3 |
| 12 | - | 1 | 21 | - | 2 | 30 | - | 3 |

**1 - 12 = 1** →insgesamt haben 12 Personen das e-Learing-System als Gut beurteilt

**13 - 21 = 9** →9 P. haben System als Mittel beurteilt

**21 - 30 = 9** → 9 P. haben System als Schlecht beurteilt

# Ersetzung der Werte durch Ränge

**Gleiche Werte:** Die **Ränge 1-12** suggerieren, dass die Zahl die dem Rang 1 zugeordnet kleiner ist, als die Zahl die dem Rang 12 zugeordnet ist; da sie in Wirklichkeit alle den gleichen Wert haben (=1), müssen das auch die Ränge widerspiegeln.

Das gelingt, indem wir den Mittelwert der Ränge 1-12 bilden:

(1+2+…+12)/12= **6.5**

**→ statt den Rängen 1-12 nehmen wir 12-mal den Rang 6.5 (=Rangziffer)**

# Ersetzung der Werte durch Ränge

- Das gleiche gilt für die **Ränge 13-21**, denen allen der Wert 2 zugeordnet wurde:

mittlerer Rang = **17 → statt den Rängen 13-21 nehmen wir 9-mal den Rang 17**

- Dasselbe für **die Ränge 22-30**, denen allen der Wert 3 zugeordnet wurde:

mittlerer Rang = **26 → statt den Rängen 22-30 nehmen wir 9-mal den Rang 26**

# Ermittlung der Rangsumme

- Die Rangsummen werden ermittelt, indem wir die Anzahl der Beobachtungen für einen Wert mit dem zugeordneten Rang multiplizieren und die Werte pro Gruppe dann summieren.

| **Beurteilung** | **SchülerInnen = $n_1$** | **Studierende = $n_2$** |
|---|---|---|
| 1 = Gut | 5*6.5= 32.5 | 7*6.5 = 45.5 |
| 2 = Mittel | 4*17= 68 | 5*17 = 85 |
| 3 = Schlecht | 6*26 = 156 | 3*26 = 78 |
| Rangsumme R | 256.5 = $R_1$ | 208.5 = $R_2$ |
| Mittlere Rangsumme | 256.5/15= 17.1 | 208.5/15 = 13.9 |

- Es gilt: $R_1 + R_2$ = (N ( N +1 )) / 2 = 256.5+208.5= [(30*(30+1)]/2 = 465

→ entspricht der Summe der einzelnen Ränge also 1+2+…+30

- Exkurs: Gaußsche Summenformel: 1+2+3+4+n = (n+1)*n/2 = $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

# Bestimmung der Testmaßzahl U

$$U_1 = R_1 - \frac{( n_1 * ( n_1 + 1 ) )}{2}$$

$$U_2 = R_2 - \frac{( n_1 * ( n_1 + 1 ) )}{2}$$

$$U_1 = 256{,}5 - \frac{( 15 * 16 )}{2}$$

$$U_2 = 208{,}5 - \frac{( 15 * 16 )}{2}$$

$\boxed{U_1 = 136{,}5 \qquad U_2 = 88{,}5}$

- Die Prüfgröße (=Teststatistik) des U Tests ist immer der kleinere der beiden U Werte

- $U = \text{Minimum} ( U_1 , U_2 ) = 88{,}5$

# Kritischer U - Wert

- $U_{0,05;15;15} = 64$
- Um signifikant zu ein, muss U kleiner oder gleich dem kritischen U-Wert sein. Da der errechnete U-Wert größer als der tabellierte U-Wert ist (88.5 >64), wird $H_0$ beibehalten. Es folgt keine Interpretation von $H_1$
- Es besteht kein signifikanter Unterschied zwischen SchülerInnen und Studierenden,  hinsichtlich der Beurteilung der Funktionsfähigkeit des Corona bedingten e-Learningsystems

**Table 3** Critical values of *U* (5% significance).

| $n_1$ \ $n_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 3 | | | | | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 |
| 4 | | | | 0 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 11 | 12 | 13 | 13 |
| 5 | | | 0 | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 20 |
| 6 | | | 1 | 2 | 3 | 5 | 6 | 8 | 10 | 11 | 13 | 14 | 16 | 17 | 19 | 21 | 22 | 24 | 25 | 27 |
| 7 | | | 1 | 3 | 5 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 | 32 | 34 |
| 8 | | 0 | 2 | 4 | 6 | 8 | 10 | 13 | 15 | 17 | 19 | 22 | 24 | 26 | 29 | 31 | 34 | 36 | 38 | 41 |
| 9 | | 0 | 2 | 4 | 7 | 10 | 12 | 15 | 17 | 20 | 23 | 26 | 28 | 31 | 34 | 37 | 39 | 42 | 45 | 48 |
| 10 | | 0 | 3 | 5 | 8 | 11 | 14 | 17 | 20 | 23 | 26 | 29 | 33 | 36 | 39 | 42 | 45 | 48 | 52 | 55 |
| 11 | | 0 | 3 | 6 | 9 | 13 | 16 | 19 | 23 | 26 | 30 | 33 | 37 | 40 | 44 | 47 | 51 | 55 | 58 | 62 |
| 12 | | 1 | 4 | 7 | 11 | 14 | 18 | 22 | 26 | 29 | 33 | 37 | 41 | 45 | 49 | 53 | 57 | 61 | 65 | 69 |
| 13 | | 1 | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 33 | 37 | 41 | 45 | 50 | 54 | 59 | 63 | 67 | 72 | 76 |
| 14 | | 1 | 5 | 9 | 13 | 17 | 22 | 26 | 31 | 36 | 40 | 45 | 50 | 55 | 59 | 64 | 67 | 74 | 78 | 83 |
| 15 | | 1 | 5 | 10 | 14 | 19 | 24 | 29 | 34 | 39 | 44 | 49 | 54 | 59 | 64 | 70 | 75 | 80 | 85 | 90 |
| 16 | | 1 | 6 | 11 | 15 | 21 | 26 | 31 | 37 | 42 | 47 | 53 | 59 | 64 | 70 | 75 | 81 | 86 | 92 | 98 |
| 17 | | 2 | 6 | 11 | 17 | 22 | 28 | 34 | 39 | 45 | 51 | 57 | 63 | 67 | 75 | 81 | 87 | 93 | 99 | 105 |
| 18 | | 2 | 7 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 55 | 61 | 67 | 74 | 80 | 86 | 93 | 99 | 106 | 112 |
| 19 | | 2 | 7 | 13 | 19 | 25 | 32 | 38 | 45 | 52 | 58 | 65 | 72 | 78 | 85 | 92 | 99 | 106 | 113 | 119 |
| 20 | | 2 | 8 | 13 | 20 | 27 | 34 | 41 | 48 | 55 | 62 | 69 | 76 | 83 | 90 | 98 | 105 | 112 | 119 | 127 |

# Asymptotische Signifikanz

- Neben dem Vergleich der Prüfgröße U mit dem kritischen U-Wert, gibt es noch den **asymptotischen Test**, um festzustellen, ob ein signifikanter Unterschied zwischen den beiden Gruppen vorliegt; damit kann die **exakte Irrtumswahrscheinlichkeit** berechnet werden.
- Kann nur bei ausreichend großem Stichprobenumfang durchgeführt werden (Unterschiedliche Meinungen: für beide Gruppen mindestes n=6, insgesamt mindestens n=50, n1+n1>30)
- Bei diesem Test wird die Teststatistik (Prüfgröße U) in einen z-Wert umgerechnet (z-Standardisierung), der dann asymptotisch normalverteilt ist. Nachdem z bestimmt wurde, wird mittels Integral die asymptotische Signifikanz (exakte Irrtumswahrscheinlichkeit) berechnet.

# Asymptotische Signifikanz

Formel für z-Wert: $$z = \frac{U - \mu_U}{\sigma_U}$$

Da Messwerte auftreten, die wiederholt vorkommen, wird in die Formel ein Korrektur-Term eingebaut, sodass sie dann wie folgt lautet:

$$z = \frac{U - \frac{n1 * n2}{2}}{\sqrt{\frac{n1 * n2}{12 * N * (N-1)} * \left(N^3 - N - \sum_{i=1}^{m} (t_i^3 - t_i)\right)}}$$

z = -1,059 bzw. als Betrag 1,059

Der z-Wert wird auf Signifikanz geprüft, indem er mit dem kritischen Wert der Standard-normalverteilung verglichen wird. Ist der Betrag der Teststatistik höher als der kritische Wert, dann ist der Unterschied signifikant.

Für das zweiseitige Signifikanzniveau 0,05 beträgt dieser Grenzwert +- 1,96.

Da 1,059 < 1,96 besteht kein signifikanter Unterschied.

# Berechnung der Irrtumswahrscheinlichkeit

$$\text{sig} = 1 - \left[ \left( \frac{1}{\sqrt{2\pi}} \right) \cdot \int_{-z}^{+z} e^{\frac{-x^2}{2}} \, dx \right] =$$

Unser Beispiel : $Z = -1{,}059$ (siehe vorige Folie)

sig = 0,289

sig = 0,289 = 28,9 % Irrtumswahrscheinlichkeit

$\Rightarrow$ Da sig > 0,05 ist, wird $H_0$ beibehalten.

# Schlussfolgerung

- Zwischen SchülerInnen und Studierenden besteht kein signifikanter Unterschied hinsichtlich der Beurteilung des Corona-bedingten e-Learning Systems.

- WÜRDE ein Unterschied bestehen, dann müsste man die mittleren Ränge interpretieren:
- Man betrachtet die mittleren Ränge:

| | **SchülerInnen = n1** | **Studierende = n2** |
|---|---|---|
| **Mittlerer Rang** | 17.1 | 13.9 |

# Schlussfolgerung

- Der einzelne mittlere Rang sagt nichts aus und wird so auch nicht interpretiert. Betrachtet man jedoch die Höhe des mittleren Ranges, so kann festgestellt werden, dass der mittlere Rang bei den SchülerInnen höher ist als jener, der Studierenden.
- Die ursprüngliche Codierung der Beurteilung war:
  - 1 = gut
  - 2 = mittel
  - 3 = schlecht
- Ein höherer Wert sagt also aus, dass "schlechter beurteilt" wurde. Bei H1 (ist in unserem Beispiel nicht der Fall !) hieße das, dass SchülerInnen die Funktionsfähigkeit des e-Learnings etwas schlechter beurteilen als Studierende.

# Zusätzliche Quellen


- „Mann-Whitney-U-Test“. https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/mann.html#2.2._Berechnung_der_Teststatistik (12. Mai 2020).
- Mann-Whitney U-Test (Tutorial). https://www.youtube.com/watch?v=3aTOd8yKRno (12. Mai 2020).
- „Verteilungsunabhängige Tests / nichtparametrische Tests - Statistik Wiki Ratgeber Lexikon“. Statistik Nachhilfe Ratgeber. https://www.statistik-nachhilfe.de/ratgeber/statistik/induktive-statistik/signifikanztests-hypothesentests/verteilungsunabhaengige-tests-nichtparametrische-tests (12. Mai 2020).