PROGRAM STUDI TEKNIK KOMPUTER DAN INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
SEMESTER 311 TAHUN AJARAN 2020/2021

IS240 – PROBABILITAS DAN STATISTIKA

Pertemuan ke 12 – Statistika Nonparametrik

Penyusun
- Tan Thing Heng, BS, MStats

Capaian Pembelajaran Mingguan Mata Kuliah (Sub-CPMK):

SUB-CPMK 8: Mahasiswa mampu menerapkan teknik statistika nonparametrik(C4)

Sub-Pokok Bahasan:
1. Uji kesamaan populasi berpasangan dengan uji Wilcoxon
2. Uji kesamaan populasi independent dengan uji Mann-Whitney
3. Korelasi Spearman

Referensi:
Black bab 17

Comparing Two Populations

Gaussian Numerical Data

- Paired population means: t-test
- Two independent population means:
  - Unequal variances: general procedure t-test
  - Equal variances: pooled t-test

Non-Gaussian Numerical Data or Categorical Data with Ordinal Scale

- Paired population: Wilcoxon Matched-Pairs Signed test
- Independent populations: Mann-Whitney U test

Paired (related) Populations

City	1979	2009
1	20.3	22.8
2	19.5	12.7
3	18.6	14.1
4	20.9	16.1
5	19.9	25.2
6	18.6	20.2
7	19.6	14.9
8	23.2	21.3
9	21.8	18.7
10	20.3	20.9
11	19.2	22.6
12	19.5	16.9
13	18.7	20.6
14	17.7	18.5
15	21.6	23.4
16	22.4	21.3
17	20.8	17.4

Independent Populations

Health Service Worker	Educational Service Worker
$20.10	$26.19
19.80	23.88
22.36	25.50
18.75	21.64
21.90	24.85
22.96	25.30
20.75	24.12
	23.45

Wilcoxon Matched-Pairs Signed Test for Related Populations for Small-sample Size

Assumptions

- The paired data are selected randomly.
- Their distributions are symmetrical.
- The sample size $n < 15$.

Example: Healthcare Spending

Family Pair	Pittsburgh	Oakland
1	$1,950	$1,760
2	1,840	1,870
3	2,015	1,810
4	1,580	1,660
5	1,790	1,340
6	1,925	1,765

A healthcare analyst uses $\alpha = .05$ to test to determine whether there is a significant difference in annual household healthcare spending between these two cities.

The following hypotheses are being tested.

$$H_0: M_d = 0$$
$$H_a: M_d \neq 0$$

Example: Healthcare Spending

Family Pair	Pittsburgh	Oakland	d	Rank
1	$1,950	$1,760	+190	+4
2	1,840	1,870	−30	−1
3	2,015	1,810	+205	+5
4	1,580	1,660	−80	−2
5	1,790	1,340	+450	+6
6	1,925	1,765	+160	+3

$$T = \text{minimum of } (T_+, T_-)$$

$$T_+ = 4 + 5 + 6 + 3 = 18$$

$$T_- = 1 + 2 = 3$$

$$T = \text{minimum of } (18, 3) = 3$$

ACTION:

STEP 7. Because $T = 3$ is greater than critical $T = 1$, the decision is not to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Not enough evidence is provided to declare that Pittsburgh and Oakland differ in annual household spending on healthcare. This information may be useful to healthcare providers and employers in the two cities and particularly to businesses that either operate in both cities or are planning to move from one to the other. Rates can be established on the notion that healthcare costs are about the same in both cities. In addition, employees considering transfers from one city to the other can expect their annual healthcare costs to remain about the same.

Wilcoxon Matched-Pairs Signed Test for Related Populations

Assumptions

- The paired data are selected randomly.
- Their distributions are symmetrical.
- For n > 15, the the test statistics is approximately Gaussian.

$$H_0 : M_d = 0$$
$$H_a : M_d \neq 0$$

$$\mu_T = \frac{(n)(n+1)}{4}$$

$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}}$$

$$z = \frac{T - \mu_T}{\sigma_T}$$

Example: Airline ticket cost in various cities

City	1979	2009	d	Rank
1	20.3	22.8	−2.5	−8
2	19.5	12.7	+6.8	+17
3	18.6	14.1	+4.5	+13
4	20.9	16.1	+4.8	+15
5	19.9	25.2	−5.3	−16
6	18.6	20.2	−1.6	−4
7	19.6	14.9	+4.7	+14
8	23.2	21.3	+1.9	+6.5
9	21.8	18.7	+3.1	+10
10	20.3	20.9	−0.6	−1
11	19.2	22.6	−3.4	−11.5
12	19.5	16.9	+2.6	+9
13	18.7	20.6	−1.9	−6.5
14	17.7	18.5	−0.8	−2
15	21.6	23.4	−1.8	−5
16	22.4	21.3	+1.1	+3
17	20.8	17.4	+3.4	+11.5

$$T = \text{minimum of } (T_+, T_-)$$
$$T_+ = 17 + 13 + 15 + 14 + 6.5 + 10 + 9 + 3 + 11.5 = 99$$
$$T_- = 8 + 16 + 4 + 1 + 11.5 + 6.5 + 2 + 5 = 54$$
$$T = \text{minimum of } (99, 54) = 54$$

The T value is normally distributed for large sample sizes, with a mean and standard deviation of

$$\mu_T = \frac{(n)(n+1)}{4} = \frac{(17)(18)}{4} = 76.5$$
$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(17)(18)(35)}{24}} = 21.1$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{54 - 76.5}{21.1} = -1.07$$

Example: Airline ticket cost in various cities

ACTION:

STEP 7. The critical z value for this two-tailed test is $z_{.025} = \pm 1.96$. The observed $z = -1.07$, so the analyst fails to reject the null hypothesis. There is no significant difference in the cost of airline tickets between 1979 and 2009.

BUSINESS IMPLICATIONS:

STEP 8. Promoters in the airline industry can use this type of information (the fact that ticket prices have not increased significantly in 30 years) to sell their product as a good buy. In addition, industry managers could use it as an argument for raising prices.

R Code: Large-Sample Wilcoxon Matched-Pairs Rank Sign Test

```
library(coin)
```

```
## Loading required package: survival
```

```
(wt <- coin::wilcoxsign_test(airline$year1979 ~ airline$year2009))
```

```
##
##  Asymptotic Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = 1.0416, p-value = 0.2976
## alternative hypothesis: true mu is not equal to 0
```

```
(wmp <- wilcox.test(airline$year1979, airline$year2009, paired = TRUE, mu = 0))
```

```
## Warning in wilcox.test.default(airline$year1979, airline$year2009, paired =
## TRUE, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  airline$year1979 and airline$year2009
## V = 98.5, p-value = 0.3087
## alternative hypothesis: true location shift is not equal to 0
```

EXERCISES

Paired Populations

- Black p 17.15 – 17.18

Independent Populations

- Black p685 #17.9 – 17.12

Mann-Whitney U Test for Independent Populations for Small-sample Size

H_0: The two populations are identical.
H_a: The two populations are not identical.

Assumptions:

- The scale of measurement is at least ordinal.
- The samples are independent.
- The sample size n < 10.

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - W_2$$

$$U' = n_1 \cdot n_2 - U$$

Table A.13 contains *p*-values for *U*. To determine the *p*-value for a *U* from the table, let n_1 denote the size of the smaller sample and n_2 the size of the larger sample. Using the particular table in Table A.13 for n_1, n_2, locate the value of *U* in the left column. At the intersection of the *U* and n_1 is the *p*-value for a one-tailed test. For a two-tailed test, double the *p*-value shown in the table.

Example: Hourly Rate

Health Service Worker	Educational Service Worker
$20.10	$26.19
19.80	23.88
22.36	25.50
18.75	21.64
21.90	24.85
22.96	25.30
20.75	24.12
	23.45

H_0: The health service population is identical to the educational service population on employee compensation.

H_a: The health service population is not identical to the educational service population on employee compensation.

Example: Hourly Rate

Total Employee Compensation	Rank	Group
$18.75	1	H
19.80	2	H
20.10	3	H
20.75	4	H
21.64	5	E
21.90	6	H
22.36	7	H
22.96	8	H
23.45	9	E
23.88	10	E
24.12	11	E
24.85	12	E
25.30	13	E
25.50	14	E
26.19	15	E

$$W_1 = 1 + 2 + 3 + 4 + 6 + 7 + 8 = 31$$

$$W_2 = 5 + 9 + 10 + 11 + 12 + 13 + 14 + 15 = 89$$

$$U_1 = (7)(8) + \frac{(7)(8)}{2} - 31 = 53$$

$$U_2 = (7)(8) + \frac{(8)(9)}{2} - 89 = 3$$

Because U_2 is the smaller value of U, we use $U = 3$ as the test statistic for Table A.13. Because it is the smallest size, let $n_1 = 7$; $n_2 = 8$.

ACTION:

STEP 7. Table A.13 yields a *p*-value of .0011. Because this test is two tailed, we double the table *p*-value, producing a final *p*-value of .0022. Because the *p*-value is less than $\alpha = .05$, the null hypothesis is rejected. The statistical conclusion is that the populations are not identical.

BUSINESS IMPLICATIONS:

STEP 8. An examination of the total compensation figures from the samples indicates that employers pay educational service workers more per hour than they pay health service workers.

Mann-Whitney U Test for Independent Populations

Assumptions:

- The scale of measurement is at least ordinal.
- The samples are independent.
- For sample size n is at least 10, the U test statistic is approximately Gaussian.

$$\mu_U = \frac{n_1 \cdot n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 \cdot n_2(n_1 + n_2 + 1)}{12}}, \quad z = \frac{U - \mu_U}{\sigma_U}$$

Example: Income of TV Viewers

PBS	Non-PBS
$24,500	$41,000
39,400	32,500
36,800	33,000
43,000	21,000
57,960	40,500
32,000	32,400
61,000	16,000
34,000	21,500
43,500	39,500
55,000	27,600
39,000	43,500
62,500	51,900
61,400	27,800
53,000	
$n_1 = 14$	$n_2 = 13$

H_0: The incomes of PBS and non-PBS viewers are identical.
H_a: The incomes of PBS and non-PBS viewers are not identical.

Example: Income of TV Viewers

Income	Rank	Group	Income	Rank	Group
$16,000	1	Non-PBS	39,500	15	Non-PBS
21,000	2	Non-PBS	40,500	16	Non-PBS
21,500	3	Non-PBS	41,000	17	Non-PBS
24,500	4	PBS	43,000	18	PBS
27,600	5	Non-PBS	43,500	19.5	PBS
27,800	6	Non-PBS	43,500	19.5	Non-PBS
32,000	7	PBS	51,900	21	Non-PBS
32,400	8	Non-PBS	53,000	22	PBS
32,500	9	Non-PBS	55,000	23	PBS
33,000	10	Non-PBS	57,960	24	PBS
34,000	11	PBS	61,000	25	PBS
36,800	12	PBS	61,400	26	PBS
39,000	13	PBS	62,500	27	PBS
39,400	14	PBS			

H_0: The incomes of PBS and non-PBS viewers are identical.
H_a: The incomes of PBS and non-PBS viewers are not identical.

$$W_1 = 4 + 7 + 11 + 12 + 13 + 14 + 18 + 19.5 + 22 + 23 + 24 + 25 + 26 + 27 = 245.5$$

Then W_1 is used to compute the U value. Because $n_1 = 14$ and $n_2 = 13$, then

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1 = (14)(13) + \frac{(14)(15)}{2} - 245.5 = 41.5$$

Because $n_1, n_2 > 10$, U is approximately normally distributed, with a mean of

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{(14)(13)}{2} = 91$$

and a standard deviation of

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(14)(13)(28)}{12}} = 20.6$$

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{41.5 - 91}{20.6} = \frac{-49.5}{20.6} = -2.40$$

ACTION:

STEP 7. The observed value of z is -2.40, which is less than $Z_{\alpha/2} = -1.96$ so the results are in the rejection region. That is, there is a difference between the income of a PBS viewer and that of a non-PBS viewer. Examination of the sample data confirms that in general, the income of a PBS viewer is higher than that of a non-PBS viewer.

R Code

```
library(coin)
```

```
## Loading required package: survival
```

```
stacked_income <- stack(income)
(wt <- coin::wilcox_test(values ~ ind, data = stacked_income))
```

```
##
##  Asymptotic Wilcoxon-Mann-Whitney Test
##
## data:  values by ind (PBS, Non-PBS)
## Z = 2.4024, p-value = 0.01629
## alternative hypothesis: true mu is not equal to 0
```

```
(mw <- wilcox.test(income$PBS, income$`Non-PBS`, mu = 0, paired = FALSE))
```

```
## Warning in wilcox.test.default(income$PBS, income$`Non-PBS`, mu = 0, paired =
## FALSE): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  income$PBS and income$`Non-PBS`
## W = 140.5, p-value = 0.0174
## alternative hypothesis: true location shift is not equal to 0
```

EXERCISES

Paired Populations

- Black p 17.15 – 17.18

Independent Populations

- Black p685 #17.9 – 17.12

Asosiasi, Korelasi dan Kausasi

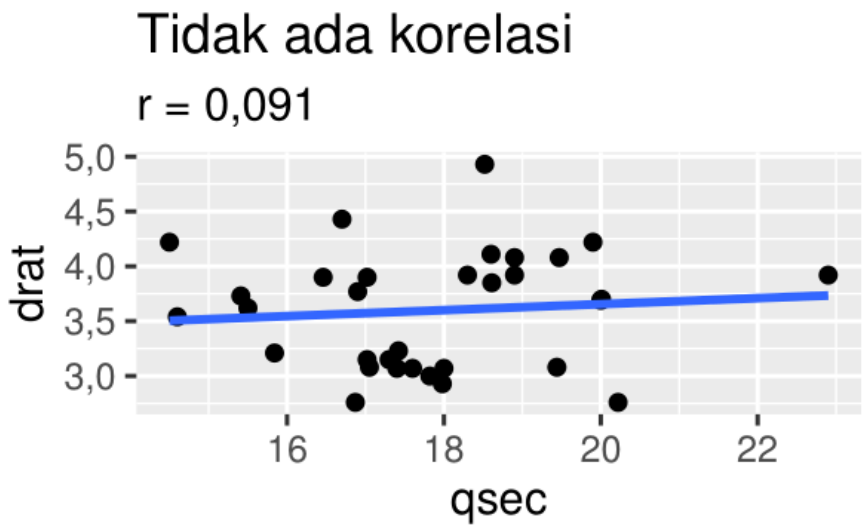
- Asosiasi adalah hubungan antara dua buah variabel (independen atau dependen).
- Korelasi adalah kekuatan hubungan linier antara 2 buah variabel numerik.
- $cor(X, Y) = cor(Y, X)$
- Kausasi **(causation)** adalah hubungan sebab akibat.
- Korelasi **tidak sama** dengan kausalitas. Koefisien korelasi yang tinggi hanya menunjukkkan kuatnya suatu hubungan, **bukan** arah kausalitas.
- **Spurious correlation** terjadi bila terdapat korelasi yang kuat antara 2 variabel numerik yang tidak ada hubungannya. Contoh: korelasi antara penjualan es krim dengan skor kuis Statistika.

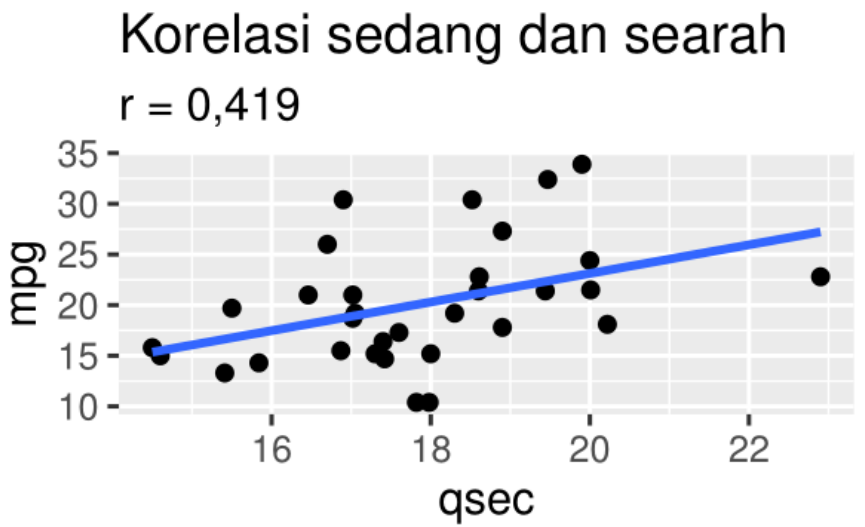
Tipe Asosiasi

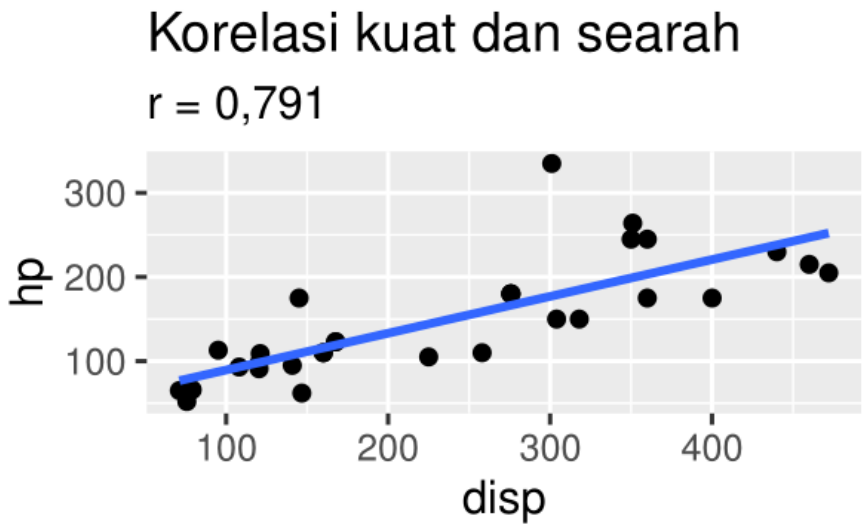
Uji	Tipe Variabel	Syarat Penggunaan
Pearson	Numerik & Numerik	distribusi data Gaussian
Spearman	Numerik & Numerik	distribusi data **tidak** Gaussian
Spearman	Kategorikal & Kategorikal	skala pengukuran ordinal
Kendall	Kategorikal & Kategorikal	s.p. ordinal, bila tdp *(tied rank)*
Chi-Kuadrat	Kategorikal & Kategorikal	s.p. nominal

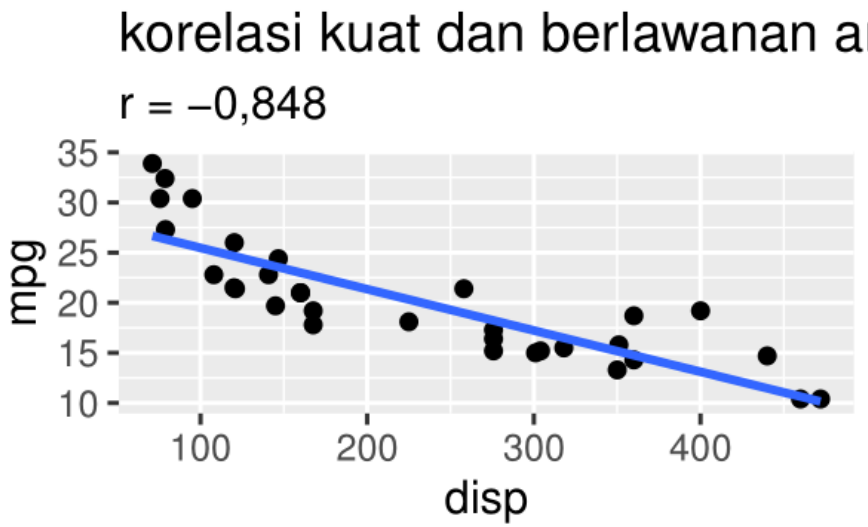
Tipe asosiasi yang dibahas pada pertemuan ini adalah korelasi Pearson dan korelasi Spearman.

Visualisasi Korelasi dengan Diagram Acak









Koefisien Korelasi Pearson

```
##                mpg       disp         hp       drat
## mpg   1,0000000 -0,8475514 -0,7761684  0,6811719
## disp -0,8475514  1,0000000  0,7909486 -0,7102139
## hp   -0,7761684  0,7909486  1,0000000 -0,4487591
## drat  0,6811719 -0,7102139 -0,4487591  1,0000000
## wt   -0,8676594  0,8879799  0,6587479 -0,7124406
```

Interpretasi Koefisien Korelasi Pearson

Nilai r	Kekuatan hubungan antara 2 variabel numerik
$r = 0$	Tidak terdapat hubungan
$0 \leq r < 0,2$	Korelasi searah yang dapat diabaikan
$0,2 \leq r < 0,4$	Hubungan searah yang lemah
$0,4 \leq r < 0,6$	Hubungan searah dengan kekuatan sedang
$0,6 \leq r < 0,8$	Hubungan searah yang kuat
$0,8 \leq r \leq 1$	Hubungan searah yang sangat kuat

Interpretasi Koefisien Korelasi Pearson

Nilai r	Kekuatan hubungan antara 2 variabel numerik
$r = 0$	Tidak terdapat hubungan
$-0,2 \leq r < 0$	Korelasi berlawanan arah yang dapat diabaikan
$-0,4 \leq r < -0,2$	Hubungan berlawanan arah yang lemah
$-0,6 \leq r < -0,4$	Hubungan berlawanan arah dengan kekuatan sedang
$-0,8 \leq r < -0,6$	Hubungan berlawanan arah yang kuat
$-1 \leq r < -0,8$	Hubungan berlawanan arah yang sangat kuat

Rumus Korelasi Sample Pearson

$$r = \frac{s_{XY}}{s_X s_Y}$$
$$= \frac{\sum(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum(x_j - \bar{x})^2}\sqrt{\sum(y_j - \bar{y})^2}}$$

di mana

- x_j = observasi ke j
- $\bar{x}$ = rata-rata X
- s_x = simpangan baku X

Koefisien r seringkali disebut sebagai *Pearson product-moment correlation coefficient*.

Rumus Korelasi Spearman

$$r_s = 1 - \frac{6 \sum d^2}{n(n-1)}$$

di mana

- d = selisih rank antara data untuk sebuah pasangan (x, y)
- n = ukuran sampel

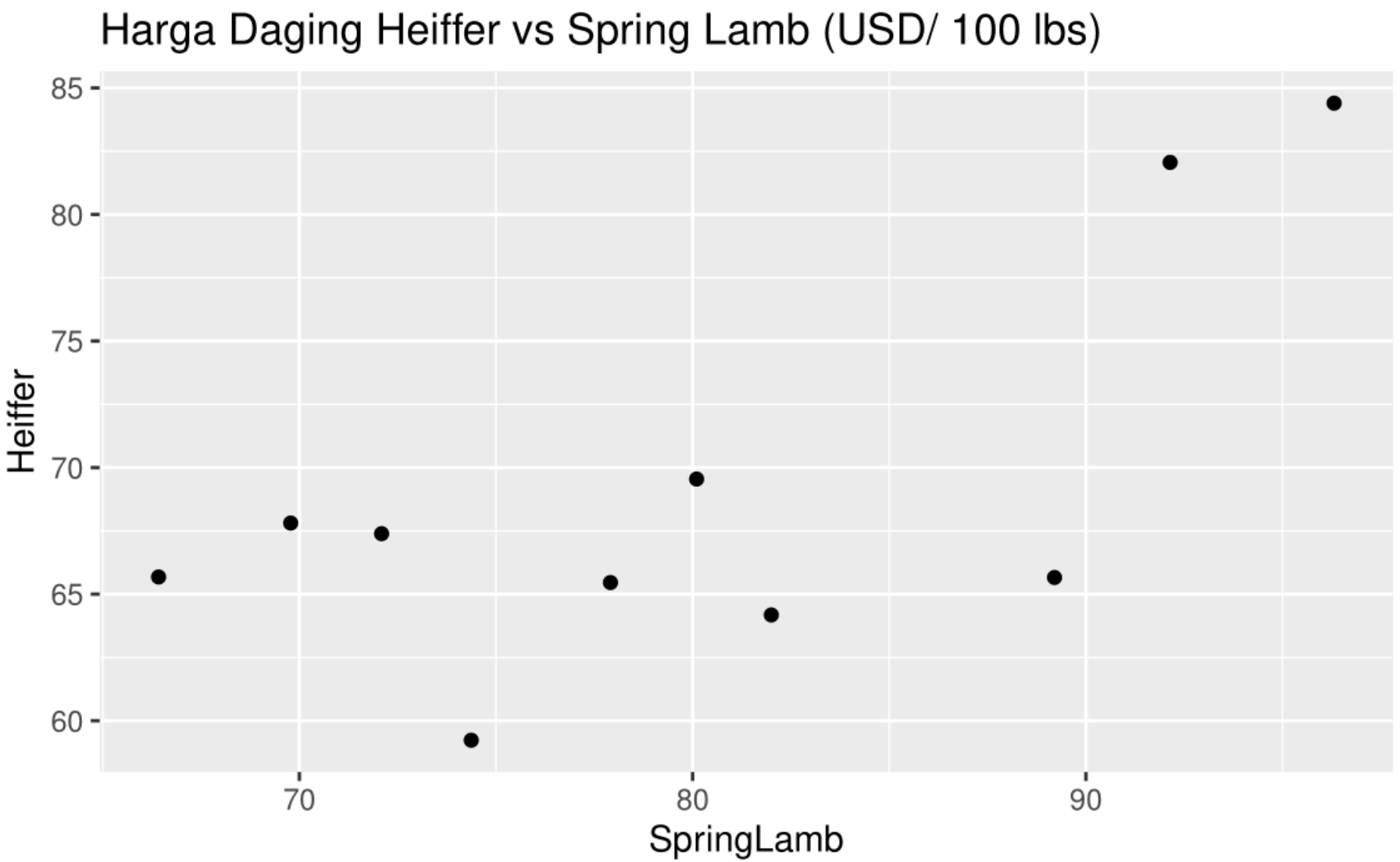
Korelasi Spearman digunakan untuk mengukur korelasi

- data numerik yang distribusinya tidak Gaussian atau
- data kategorikal yang skala pengukurannya ordinal

Contoh Soal: Lamb

```
## # A tibble: 10 x 3
##     Year SpringLamb Heiffer
##    <dbl>      <dbl>   <dbl>
##  1     1       77.9    65.5
##  2     2       82      64.2
##  3     3       89.2    65.7
##  4     4       74.4    59.2
##  5     5       66.4    65.7
##  6     6       80.1    69.6
##  7     7       69.8    67.8
##  8     8       72.1    67.4
##  9     9       92.1    82.1
## 10    10       96.3    84.4
```

Contoh Soal: Lamb



Contoh Soal: Lamb

Year	Rank: Lamb	Rank: Heifer	d	d^2
1	5	3	2	4
2	7	2	5	25
3	8	4	4	16
4	4	1	3	9
5	1	5	−4	16
6	6	8	−2	4
7	2	7	−5	25
8	3	6	−3	9
9	9	9	0	0
10	10	10	0	0
				$\Sigma d^2 = 108$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(108)}{10(10^2 - 1)} = 0.345$$

Uji Korelasi Spearman

```
##
##  Spearman's rank correlation rho
##
## data:  lamb$SpringLamb and lamb$Heiffer
## S = 108, p-value = 0,3305
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0,3454545
```

Koefisien korelasi Spearman tidak berbeda secara signifikan dari 0.

Latihan

- Pearson
  - Black hal 468 nomor 12.3 - 12.5
  - Black hal 510 nomor 12.59, 12.61
  - Hayter hal 599 nomor 12.9.5, 12.9.10 - 12.9.14, 12.12.24
- Spearman
  - Black hal 709 nomor 17.31, 17.35, 17.37

References and Bibliography

- Ken Black, Business Statistics for Contemporary Decision Making 6 ed, John Wiley & Sons, Inc
- User's guide to correlation oleh Haldun Akuglu https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969
- A guide to appropriate use of Correlation coefficient in medical research oleh MM Mukaka https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830
- Correlation Coefficients: Appropriate Use and Interpretation oleh Schober et al https://journals.lww.com/anesthesia-analgesia/Fulltext/2018/05000/Correlation_Coefficients__Appropriate_Use_and.50.aspx
- Association, correlation and causation oleh Altman dan Krzywinski https://www.nature.com/articles/nmeth.3587
- Peter Bruce and Andrew Bruce (2017), Practical Statistics for Data Scientists: 50 Essential Concepts, O'Reilly.
- Hadley Wickham dan Garret Grolemund (2017), R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, O'Reilly.