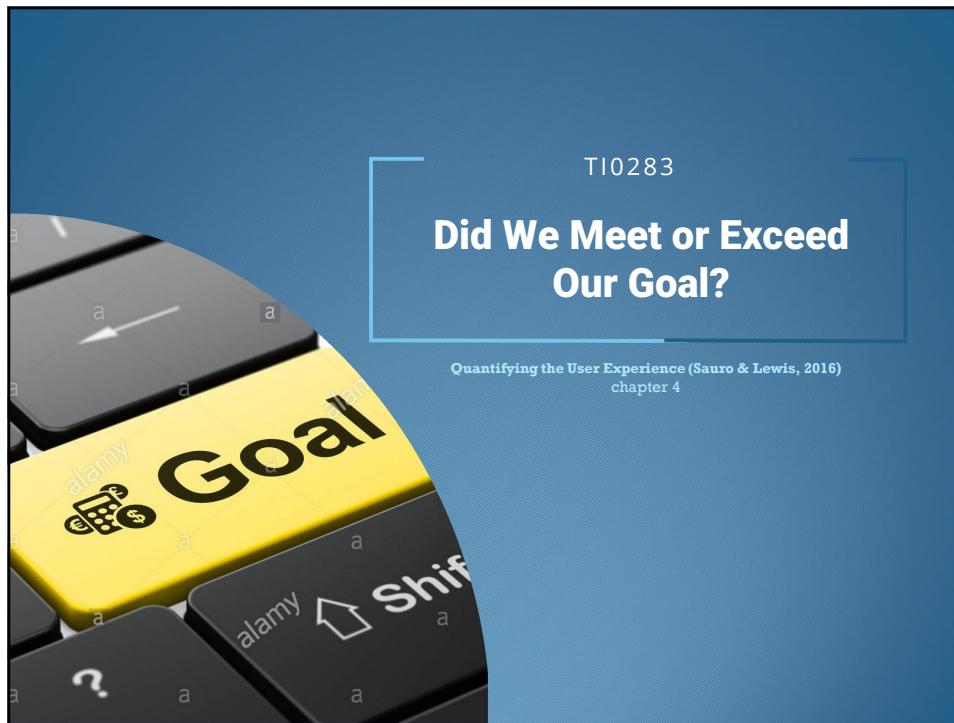




1



2

Aplikasi Statistik

The screenshot displays four statistical applications available on a mobile platform:

- Statistical Distribution** by RealXY: Contains ads. Rating 3.9. 10K+ downloads. Similar to Books & Reference.
- Statistics Calculator Pro** by Christian Göllner: Contains ads + In-app purchases. Rating 4.1. 10K+ downloads. Similar to Education.
- t Table** by feo2: Contains ads. Rating 4.4. 10K+ downloads. Similar to Education.
- Z table** by feo2: Contains ads. Rating 4.5. 10K+ downloads. Similar to Education.

Each app listing includes an UNINSTALL button, an OPEN button, and a detailed description of its features.

3

Pengantar

8 dari 9 pengguna dapat menyelesaikan task yang diberikan

Confidence Level
95%

A bar chart showing a confidence level of 95% at the 50% mark.

4/52

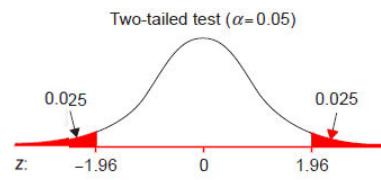
4

Pengantar

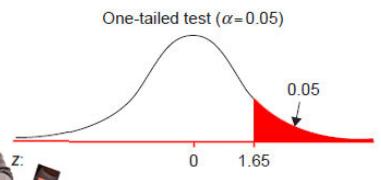
Cara lain menggunakan metode statistik yang lebih presisi
→ tipe data: discrete-binary / continuous data
→ banyaknya sampel



Two-tailed test ($\alpha=0.05$)



One-tailed test ($\alpha=0.05$)



5/52

5

Pengantar

Dua cara melakukan *one-sided test*

- Memperkirakan kemungkinan hasil observasi memenuhi benchmark
- Menentukan *confidence interval* dan menentukan batas mana (atas/bawah) yang kemungkinan besar menjadi benchmark



6/52

6

Pengantar

Contoh :

Hasil uji usabilitas suatu produk menghasilkan rata-rata task time 30 detik dengan nilai confidence interval sebesar 12 detik. Target task time yang diharapkan adalah 35 detik.



7

Membandingkan Completion Rate

SMALL-SAMPLE TEST:

- Jika jumlah sukses < 15 atau jumlah gagal < 15

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

x = jumlah responden yang berhasil menyelesaikan tugas

n = jumlah seluruh responden (jumlah sampel)

BINOM.DIST(number_s, trials, probability_s, cumulative)

BINOM.DIST(x,n, probability_s, FALSE)

8/52

8

Membandingkan Completion Rate



Contoh:

Dalam tahap awal tes desain, ditemukan bahwa 8 dari 9 responden berhasil menyelesaikan tugas yang diberikan. Apakah ada cukup bukti untuk menyatakan bahwa setidaknya 70% dari semua pengguna dapat menyelesaikan tugas tersebut?

$$X = 8 \quad \text{Kemungkinan mendapatkan tepat 8 responden berhasil} = \text{BINOM.DIST}(8, 9, 0.7, \text{FALSE}) = 0.15565$$

$$n = 9 \quad \text{Kemungkinan mendapatkan tepat 9 responden berhasil} = \text{BINOM.DIST}(9, 9, 0.7, \text{FALSE}) = 0.04035$$

Kemungkinan mendapatkan 8 atau 9 responden berhasil
= 0.15565 + 0.04035 = 0.1960

→ kemungkinan completion rate lebih besar
dari 70% sebesar **80.4%** (100%-19.6%)

9/52

9

Membandingkan Completion Rate



Contoh:

Dalam tahap awal tes desain, ditemukan bahwa 8 dari 9 responden berhasil menyelesaikan tugas yang diberikan. Apakah ada cukup bukti untuk menyatakan bahwa setidaknya 70% dari semua pengguna dapat menyelesaikan tugas tersebut?

$$X = 8 \quad p(8) = \frac{9!}{8!(9-8)!} 0.7^8 (1-0.7)^{(9-8)} = \frac{9!}{8!(1!)} 0.0576(0.3)^1 = 9(0.01729) = 0.1556$$

$$n = 9 \quad p(9) = \frac{9!}{9!(9-9)!} 0.7^9 (1-0.7)^{(9-9)} = \frac{9!}{9!(1!)} 0.04035(0.3)^0 = 0.04035(1) = 0.04035$$

Kemungkinan mendapatkan 8 atau 9 responden berhasil
= 0.15565 + 0.04035 = 0.1960

→ kemungkinan completion rate lebih besar
dari 70% sebesar **80.4%** (100%-19.6%)

10/52

10

Membandingkan Completion Rate



- Disebut **Exact Probability** karena hasil probabilitas dihitung secara ‘exact’ (bukan perkiraan)
- Hasil perhitungan untuk sampel yang kecil cenderung konserfatif

11/52

11

Membandingkan Completion Rate



MID-PROBABILITAS:

- *Exact probability* dipandang konservatif
- Mid-probabilitas menambahkan setengah probabilitas suatu keberhasilan

Kemungkinan mendapatkan tepat 8 responden berhasil =
 $\frac{1}{2}$ (BINOM.DIST(8,9,0.7,FALSE)) = $\frac{1}{2}$ (0.15565) = **0.07782**

Kemungkinan mendapatkan tepat 9 responden berhasil =
BINOM.DIST(9,9,0.7,FALSE) = **0.04035**

Kemungkinan mendapatkan 8 atau 9 responden berhasil
= **0.07782 + 0.04035 = 0.1182**
→ kemungkinan completion rate lebih besar
dari 70% sebesar **88.2%** (100%-**11.8%**)

12/52

12

Membandingkan Completion Rate



MID-PROBABILITAS:

- Calculator online:
<https://measuringu.com/calculators/onep/>

One Sample Proportion Calculator

Jeff Sauro • May 30, 2008

Use this calculator to generate both a one-sample confidence interval and to test against a criteria or benchmark.

Passed: 8 Total Tested: 9 Is Greater Than: ▼ Test Proportion: 0.70

Submit

Results

Exact Binomial p-value = 0.196.

The probability the observed proportion 0.89 comes from a population greater than 0.70 is 88.18%.

The 95% Adjusted Wald Confidence Interval is (54.31%, 100%)

13/52

13

Membandingkan Completion Rate



LARGE-SAMPLE TEST:

- Setidaknya ada 15 responden berhasil **dan** 15 responden gagal

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

\hat{p} = proporsi completion rate

p = benchmark

n = jumlah seluruh responden (jumlah sampel)

14/52

14

Membandingkan Completion Rate



Contoh 1:

Hasil dari suatu *remote-unmoderated test* menunjukkan bahwa 85 dari 100 responden dapat menemukan produk yang diminta dan menambahkannya ke keranjang belanja. Apakah terdapat cukup bukti untuk menyimpulkan bahwa setidaknya 75% pengguna dapat menyelesaikan tugas?

$$\begin{aligned}\hat{p} &= 0.85 \\ p &= 0.75 \\ n &= 100\end{aligned} \quad z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

NORM.S.DIST(z, cumulative)

NORM.S.DIST(z, TRUE)

15/52

15

Membandingkan Completion Rate



Contoh 1:

Hasil dari suatu *remote-unmoderated test* menunjukkan bahwa 85 dari 100 responden dapat menemukan produk yang diminta dan menambahkannya ke keranjang belanja. Apakah terdapat cukup bukti untuk menyimpulkan bahwa setidaknya 75% pengguna dapat menyelesaikan tugas?

NORM.S.DIST(2.309, TRUE) = 0.9895

→ 98.95% probabilitas bahwa setidaknya 75% responden dapat menyelesaikan tugas

16/52

16

Membandingkan Completion Rate



Contoh 2:

Jika 233 dari 250 responden dapat menyelesaikan tugas yang diberikan, apakah terdapat cukup bukti untuk menyimpulkan bahwa setidaknya 90% pengguna dapat menyelesaikan tugas?

$$\begin{aligned}\hat{p} &= 0.932 \\ p &= 0.90 \\ n &= 250\end{aligned} z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.932 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{250}}} = \frac{0.032}{0.019} = 1.687$$

NORM.DIST(1.687, TRUE) = 0.9542

→ 95.42% probabilitas bahwa setidaknya 90% responden dapat menyelesaikan tugas

17/52

17

Membandingkan Satisfaction Score

SATISFACTION SCORE vs BENCHMARK:

- Satisfaction score dapat dianggap sebagai data yang kontinu → dapat menggunakan t-test untuk data yang besar maupun kecil
- Biasanya sudah ada nilai benchmark kepuasan untuk produk-produk tertentu (mis. Score kepuasan HP sekitar 67)

$$t = \frac{\hat{x} - \mu}{\frac{s}{\sqrt{n}}}$$

\hat{x} = rata-rata sampel
 μ = benchmark yang diuji
 s = standard deviasi sampel
 n = jumlah sampel



18/52

18

Membandingkan Satisfaction Score

Contoh 1:

20 responden diminta untuk melakukan beberapa tugas umum (menelepon, menambah kontak, mengirim pesan) pada sebuah desain baru HP. Di akhir percobaan, mereka diminta mengisi kuesioner SUS (10 pertanyaan).

Hasil rata-rata dari SUS score adalah 73 dengan standar deviasi 19.

Dapatkankah disimpulkan bahwa usabilitas HP baru ini lebih baik dari standar kepuasan industri sebesar 67?



19/52

19

Membandingkan Satisfaction Score

Contoh 1:

$$t = \frac{\hat{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{73 - 67}{\frac{19}{\sqrt{20}}} = \frac{6}{4.24} = 1.41$$

T.DIST(x, degree of freedom, cumulative)

T.DIST(1.41, 19, TRUE) = 0.9127

→ 91.27% probabilitas bahwa HP baru tersebut memiliki skor kepuasan yang lebih baik dibanding standar industri.



20/52

20

Membandingkan Satisfaction Score

Contoh 2:

Dalam suatu usability test baru-baru ini, 172 responden mencoba menggunakan akses situs persewaan mobil. Hasil kuesioner SUS menunjukkan rata-rata skor kepuasan adalah 80 dengan standar deviasi 23.

Dapatkah disimpulkan bahwa skor rata-rata populasi lebih besar dari 75?



21/52

21

Membandingkan Satisfaction Score

Contoh 2:

$$t = \frac{\hat{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{80 - 75}{\frac{23}{\sqrt{172}}} = \frac{5}{1.75} = 2.85$$

T.DIST(x, degree of freedom, cumulative)

T.DIST(2.85, 171, TRUE) = 0.9975

→ 99.75% probabilitas bahwa populasi memiliki skor kepuasan yang lebih baik dibanding standar industri.



22/52

22

Membandingkan Satisfaction Score

MENGUBAH RATING KONTINYU MENJADI DISKRET:

- Rating kepuasan dapat disederhanakan menjadi sistem binary, mis 5 skala kepuasan dapat diubah menjadi 0 (skala 1-3) dan 1 (skala 4-5)
- Kerugian :
 - presisi pengukuran berkurang
 - semakin sulit untuk mengukur perbaikan (*improvement*)
- - dibutuhkan sampel yang cukup besar untuk dapat mendeteksi perbaikan dan mencapai *benchmark*



23/52

23

Membandingkan Satisfaction Score

MENGUBAH RATING KONTINYU MENJADI DISKRET:

Contoh: 12 pengguna menguji situs Matahari Store dan memberikan tanggapan terhadap respond “Saya merasa yakin dan aman melakukan transaksi melalui situs ini”, skala 1=sangat tidak setuju, 5= sangat setuju

Hasil rating : 4, 4, 5, 5, 5, 3, 5, 1, 5, 5, 5,

Dapatkah disimpulkan bahwa 75% pengguna merasa yakin dan aman (rating 4 atau 5) bertransaksi di situs Matahari Store?

Hasil konversi : 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1

BINOM.DIST(10,12, 0.75, FALSE) = 0.23229

BINOM.DIST(11,12, 0.75, FALSE) = 0.12671

BINOM.DIST(12,12, 0.75, FALSE) = 0.03168

$\frac{1}{2}(0.23229) + 0.12671 + 0.03168 = 0.275$



→ kemungkinan 75% pengguna setuju dengan respond tersebut **72.5%**

24/52

24

Ringkasan Rumus

Table 4.1 List of Chapter 4 Formulas

Type of Evaluation	Formula	Notes
Binomial probability formula	$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$	Used in exact and mid- p binomial tests (small sample).
Normal approximation to the binomial (Wald)	$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$	Used for large-sample binomial tests (large sample if at least 15 successes and 15 failures).
One-sample t-test	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$	Used to test continuous data (e.g., satisfaction scores, completion times).
t-based confidence interval around the mean	$\bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$	Used to construct confidence interval as alternative test against a criterion for continuous data.



25/52

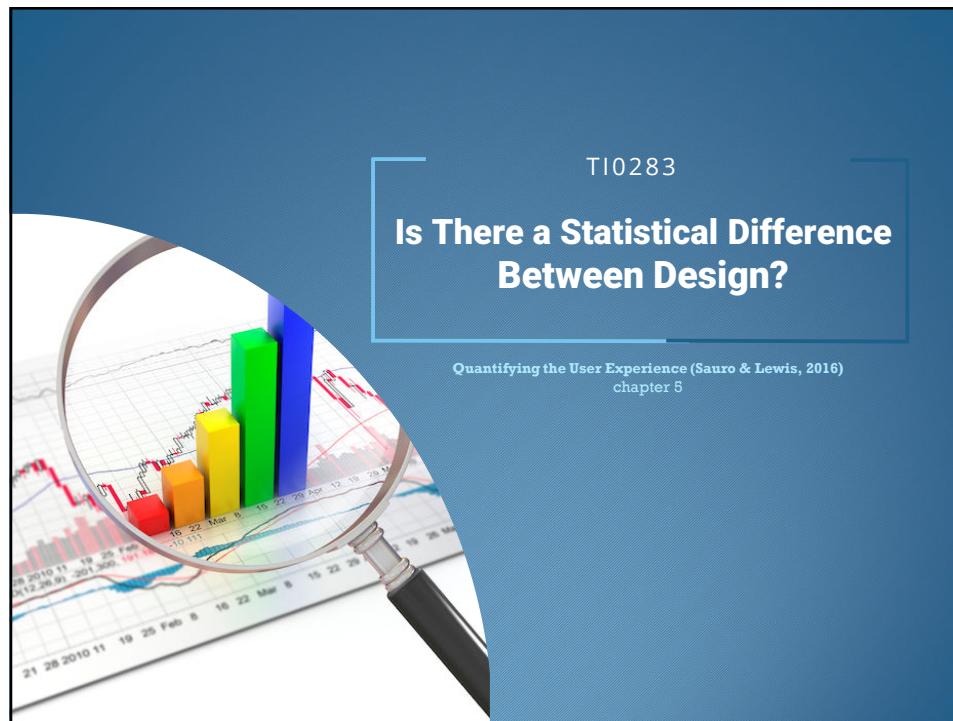
25



Hasil suatu benchmark test menunjukkan bahwa 18 dari 20 pengguna dapat menyelesaikan tugas dengan baik. Dapatkah dikatakan bahwa setidaknya 70% pengguna dapat menyelesaikan tugas?



26



27



28

Comparing Two Means

Rating Scales and Task Times

Within Subjects vs. Between-Subjects Design

	Between Subjects	Within Subjects
Condition A		
Condition B		

61 29/52

29

Comparing Two Means

Rating Scales and Task Times

WITHIN-SUBJECTS COMPARISON:

- Paired t -test

$$t = \frac{\hat{D}}{\frac{s_D}{\sqrt{n}}}$$

\hat{D} = rata-rata (mean) perbedaan skor
 s_D = standard deviasi perbedaan skor
 n = jumlah responden
 t = tes statistik

- responden melakukan pengujian terhadap beberapa produk yang berbeda

30/52

30

Comparing Two Means



Rating Scales and Task Times

WITHIN-SUBJECTS COMPARISON:

Contoh :

Pengujian terhadap dua buah aplikasi keuangan oleh 26 responden (secara acak) menghasilkan hasil kuesioner SUS seperti dalam Table 5.1

Rata-rata perbedaan skor : 29.5

Standar Deviasi : 14.125

$$t = \frac{29.5}{\sqrt{14.125}} \rightarrow T.DIST(10.649, 25, TRUE)$$

$$\sqrt{26} = 0.9999$$

$$t = 10.649$$

Disimpulkan bahwa 99.99% skor SUS aplikasi A berbeda dibanding aplikasi B Produk A *significantly higher*

Table 5.1 Pairs of SUS Scores and Their Differences for Example 1

User	A	B	Difference
1	77.5	60	17.5
2	90	62.5	27.5
3	80	45	35
4	77.5	20	57.5
5	100	80	20
6	95	42.5	52.5
7	82.5	32.5	50
8	97.5	80	17.5
9	80	52.5	27.5
10	87.5	60	27.5
11	77.5	42.5	35
12	87.5	87.5	0
13	82.5	52.5	30
14	50	10	40
15	77.5	67.5	10
16	82.5	40	42.5
17	80	57.5	22.5
18	65	32.5	32.5
19	72.5	67.5	5
20	85	47.5	37.5
21	80	45	35
22	100	62.5	37.5
23	80	40	40
24	57.5	45	12.5
25	97.5	65	32.5
26	95	72.5	22.5
Mean	82.2	52.7	29.5

31

Comparing Two Means



Rating Scales and Task Times

BETWEEN-SUBJECTS COMPARISON:

- Two-sample *t*-test

$$t = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\hat{x}_1 dan \hat{x}_2 = rata-rata dari sampel 1 dan 2

s_1 dan s_2 = standard deviasi sampel 1 dan 2

n_1 dan n_2 = jumlah sampel 1 dan 2

t = tes statistik

- Masing-masing produk di uji oleh responden yang berbeda pula

32/52

32

Comparing Two Means

Rating Scales and Task Times

!

CAUTION

Asumsi pada two-sample t-test:

- Kedua kelompok merupakan sampel yang representatif terhadap populasi induk
- Kedua kelompok tidak berhubungan satu dengan yang lain
- Kedua kelompok terdistribusi secara normal
- Variance dari setiap kelompok sama

33/52

33

Comparing Two Means

Rating Scales and Task Times

NORMALITY:

- Data terdistribusi secara normal
 $\mu=0, \sigma=1 \rightarrow$ standard normal distribution
- Masing-masing produk di uji oleh responden yang berbeda pula

34/52

34

Comparing Two Means

Rating Scales and Task Times

EQUALITY OF VARIANCE:

- Asumsi bahwa variance dalam tiap populasi sama

$$S^2 = \frac{\sum(X-\bar{x})^2}{N-1}$$
$$SD = S = \sqrt{S^2}$$

$S^2 = 5$
 $S^2 = 5$
Homogeneous

$S^2 = 4$
 $S^2 = 9$
Heterogeneous

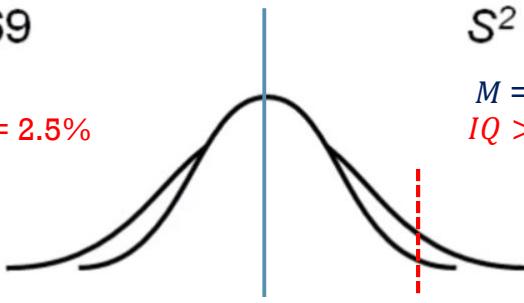
35/52

35

Comparing Two Means

Rating Scales and Task Times

EQUALITY OF VARIANCE:

$$S^2 = 169$$
$$M = 100$$
$$IQ > 130 = 2.5\%$$
$$S^2 = 289$$
$$M = 100$$
$$IQ > 130 = 7.5\%$$


- Aturan umum : Variance kedua populasi BOLEH tidak sama, SELAMA ratio perbedaan kedua standar deviasi < 2 mis: SD group A = 4, SD group B = 12 → ratio = 3 !!!

36/52

36

Comparing Two Means

Rating Scales and Task Times

BETWEEN-SUBJECTS COMPARISON:

Contoh :
Pengujian terhadap dua buah aplikasi CRM.
Aplikasi A diuji oleh 11 responden
Aplikasi B diuji oleh 12 responden
Hasil evaluasi SUS dicatat dalam Table 5.3

Rerata skor SUS aplikasi A : 51.6 (SD = 4.07)
Rerata skor SUS aplikasi B : 49.6 (SD = 4.63)

$$t = \frac{51.6 - 49.6}{\sqrt{\frac{4.07^2}{11} + \frac{4.63^2}{12}}} \rightarrow T.DIST(1.102, 20, TRUE) = 0.2835$$

$t = 1.102$

Disimpulkan bahwa kemungkinan completion time aplikasi A berbeda dari aplikasi B hanya sebesar 71.65%

Table 5.3 Data for Comparison of SUS Scores from Independent Groups

A	B
50	50
45	52.5
57.5	52.5
47.5	50
52.5	52.5
57.5	47.5
52.5	50
50	50
52.5	50
55	40
47.5	42.5
	57.5
51.6	49.6

37/52

37

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS:

- Dalam statistik terapan, perbandingan variabel biner dua populasi yang independen merupakan hal yang sering dilakukan
- Sampel yang besar gunakan **chi-square test**
Sampel yang kecil gunakan **Fisher exact test**

38/52

38

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Chi-square Test of Independence

- Tidak ada asumsi mengenai populasi induk (*parent population*) → distribution free & non parametric

	Pass	Fail	Total
Design A	a	b	m
Design B	c	d	n
Total	r	s	N

CHISQ.DIST.RT(x, degree of freedom)

39/52

39

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Chi-square Test of Independence

Contoh 1:

	Pass	Fail	Total
Design A	40	20	60
Design B	15	20	35
Total	55	40	95

$$\chi^2 = \frac{(40 \times 20 - 20 \times 15)^2 \times 95}{60 \times 35 \times 55 \times 40} \rightarrow \text{CHISQ.DIST.RT}(5.1406, 1) = 0.0234$$

Disimpulkan bahwa terdapat perbedaan yang significant secara statistic terhadap completion rate responden Desain A dan Desain B

40

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Chi-square Test of Independence

- Chi-Square tidak direkomendasikan untuk ukuran sampel yang kecil
- Gunakan Chi-Square jika minimum expected cell counts lebih dari 5

	Pass	Fail	Total
Design A	a	b	m
Design B	c	d	n
Total	r	s	N

	Pass	Fail	Total
Design A	40	20	60
Design B	15	20	35
Total	55	40	95

$$\frac{(r \times m)}{N} = \frac{(55 \times 60)}{95} = 34.74$$

$$\frac{(s \times m)}{N} = \frac{(40 \times 60)}{95} = 25.26$$

$$\frac{(r \times n)}{N} = \frac{(55 \times 35)}{95} = 20.26$$

$$\frac{(s \times n)}{N} = \frac{(40 \times 35)}{95} = 14.74$$

41/52

41

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Chi-square Test of Independence

Contoh 2:

	Pass	Fail	Total
Design A	11 (a)	1 (b)	12 (m)
Design B	5 (c)	5 (d)	10 (n)
Total	16 (r)	6 (s)	22 (N)

$$\chi^2 = \frac{(ad - bc)^2 N}{mnrs}$$

$$\chi^2 = \frac{(11 \times 5 - 1 \times 5)^2 \times 22}{12 \times 10 \times 16 \times 6}$$

$$\chi^2 = 4.7743$$

$$\rightarrow \text{CHISQ.DIST.RT}(4.7743, 1) = 0.0288$$

42/52

42

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Chi-square Test of Independence

Contoh 2:

Table 5.7 Conversion Rates for Two Designs

	Pass	Fail	Total
Design A	11 (a)	1 (b)	12 (m)
Design B	5 (c)	5 (d)	10 (n)
Total	16 (r)	6 (s)	22 (N)

Expected cell counts:

$$\frac{(r \times m)}{N} = \frac{(16 \times 12)}{22} = 8.73$$
$$\frac{(s \times m)}{N} = \frac{(6 \times 12)}{22} = 3.27$$
$$\frac{(r \times n)}{N} = \frac{(16 \times 10)}{22} = 7.27$$
$$\frac{(s \times n)}{N} = \frac{(6 \times 10)}{22} = 2.73$$

43/52

43

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



BETWEEN-SUBJECTS: Fisher Exact Test

- Tidak seperti chi-square distribution dan t-distribution, Fisher Exact Test menggunakan probabilitas yang tepat (*exact*) bukan perkiraan (*approximation*)
- Calculator online:
<https://measuringu.com/calculators/fisher/>

$$p = \frac{m!n!r!s!}{a!b!c!d!N!}$$

44/52

44

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



WITHIN-SUBJECTS: McNemar Exact Test

- Menggunakan table 2x2 seperti pada between-subject test TETAPI yang diukur adalah jumlah responden yang berubah dari sukses menjadi gagal, atau sebaliknya (pada saat mencoba desain yang berbeda)

	Design B Pass	Design B Fail	Total
Design A Pass	<i>a</i>	<i>b</i>	<i>m</i>
Design A Fail	<i>c</i>	<i>d</i>	<i>n</i>
Total	<i>r</i>	<i>s</i>	<i>N</i>

$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$

$\chi =$ jumlah pasangan discordant positif/negatif (yg terkecil)
 $n =$ tot. jumlah pasangan discordant ($b + c$)
 $p = 0.5$

45/52

45

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



WITHIN-SUBJECTS: McNemar Exact Test

Contoh:

User	Design A	Design B
1	1	0
2	1	1
3	1	1
4	1	0
5	1	0
6	1	1
7	1	1
8	0	1
9	1	0
10	1	1
11	0	0
12	1	1
13	1	0
14	1	1
15	1	0
Comp Rate	87%	53%

46/52

46

Comparing Rates

Completion Rates, Conversion Rates, A/B Testing



WITHIN-SUBJECTS: McNemar Exact Test

Contoh:

	Design B Pass	Design B Fail	Total
Design A Pass	7 (a)	6 (b)	13 (m)
Design A Fail	1 (c)	1 (d)	2 (n)
Total	8 (r)	7 (s)	15 (N)

Concordant Pairs

- Seven users completed the task on both designs (cell *a*).
- One user failed on Design A and failed on Design B (cell *d*).

Discordant Pairs

- Six users passed on Design A but failed on Design B (cell *b*).
- One user failed on Design A and passed on Design B (cell *c*).

47/52

47

Ringkasan Rumus

Name of Formula	Formula	Notes
Paired <i>t</i> -test (dependent means)	$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$	Used for all sample sizes when the same users are used in both groups.
Confidence interval around the difference between paired means	$\bar{D} \pm t_{\alpha/2} \frac{s_D}{\sqrt{n}}$	Used for all sample sizes.
Two-sample <i>t</i> -test (independent means)	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Used for all sample sizes when different users are in each sample. It is robust to violations of normality and unequal variances especially when using the Welch-Satterthwaite procedure to adjust the degrees of freedom.
Welch-Satterthwaite procedure adjustment to degrees of freedom	$df' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$	Adjusts the degrees of freedom used in a two-sample <i>t</i> -test, which makes the test more robust to violations of normality and unequal variances.
Confidence interval around two independent means	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	Used for all sample sizes.
<i>N</i> – 1 chi-square test for comparing two independent proportions (equal to the <i>N</i> – 1 two-proportion test)	$\chi^2 = \frac{(ad - bc)^2 (N-1)}{mnrs}$	The test is the same as the standard chi-square test except it is adjusted by multiplying the numerator by <i>N</i> – 1. The test is algebraically equivalent to the <i>N</i> – 1 two-proportion test. It works well as long as the expected cell counts are greater than 1 (otherwise use the Fisher exact test).



48

Ringkasan Rumus

Table 5.21 Formulas Used in This Chapter

Name of Formula	Formula	Notes
$N - 1$ two-proportion test for comparing two independent proportions	$z = \frac{(\hat{p}_1 - \hat{p}_2) \sqrt{\frac{N-1}{N}}}{\sqrt{PQ \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	The test is the same as the standard two-proportion test except it is adjusted by multiplying the numerator by $\sqrt{\frac{N-1}{N}}$. The test is algebraically equivalent to the $N - 1$ chi-square test. It works well as long as the expected cell counts are greater than 1 (otherwise use the Fisher exact test).
Fisher exact test on two independent proportions	$p = \frac{m!n!r!s!}{a!b!c!d!N!}$	Only recommended when expected cell counts are less than 1 (which doesn't happen a lot). Software computes the p -values by finding all possible combinations of tables equal to or more extreme than the marginal totals observed.
Adjusted-Wald confidence interval for the difference between independent proportions	$(\hat{p}_{adj1} - \hat{p}_{adj2}) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{adj1}(1-\hat{p}_{adj1})}{n_{adj1}} + \frac{\hat{p}_{adj2}(1-\hat{p}_{adj2})}{n_{adj2}}}$	The adjustment is to add a quarter of a squared z-critical value to the numerator and half a squared z-critical value to the denominator when computing each proportion.
McNemar exact test for matched proportions	$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$	This is the binomial probability formula, which is used on the proportion of discordant pairs. See the chapter for the process of using this and the mid-p-value.
Adjusted-Wald confidence interval for difference between matched proportions	$(\hat{p}_{2adj} - \hat{p}_{1adj}) \pm z_{\alpha/2} \sqrt{\frac{(\hat{p}_{12adj} + \hat{p}_{21adj}) - (\hat{p}_{21adj} - \hat{p}_{12adj})}{N_{adj}}}$	The interval is adjusted by adding $\frac{1}{4}$ to each cell. For a 95% confidence level this is about 0.5.



49



20 pengguna diminta untuk menambahkan contact pada suatu aplikasi CRM. 11 pengguna dari group pertama dapat menyelesaikan tugas tersebut pada versi lama, sementara hanya 9 pengguna dari group kedua yang dapat menyelesaikan tugas tersebut pada versi yang baru.



50

Activity

Task Time kedua group dapat dilihat pada tabel berikut:

Old	New
18	12
44	35
35	21
78	9
38	2
18	10
16	5
22	38
40	30
77	
20	

Dapatkah disimpulkan bahwa terjadi penurunan rerata ?



51



52