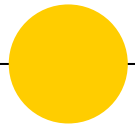
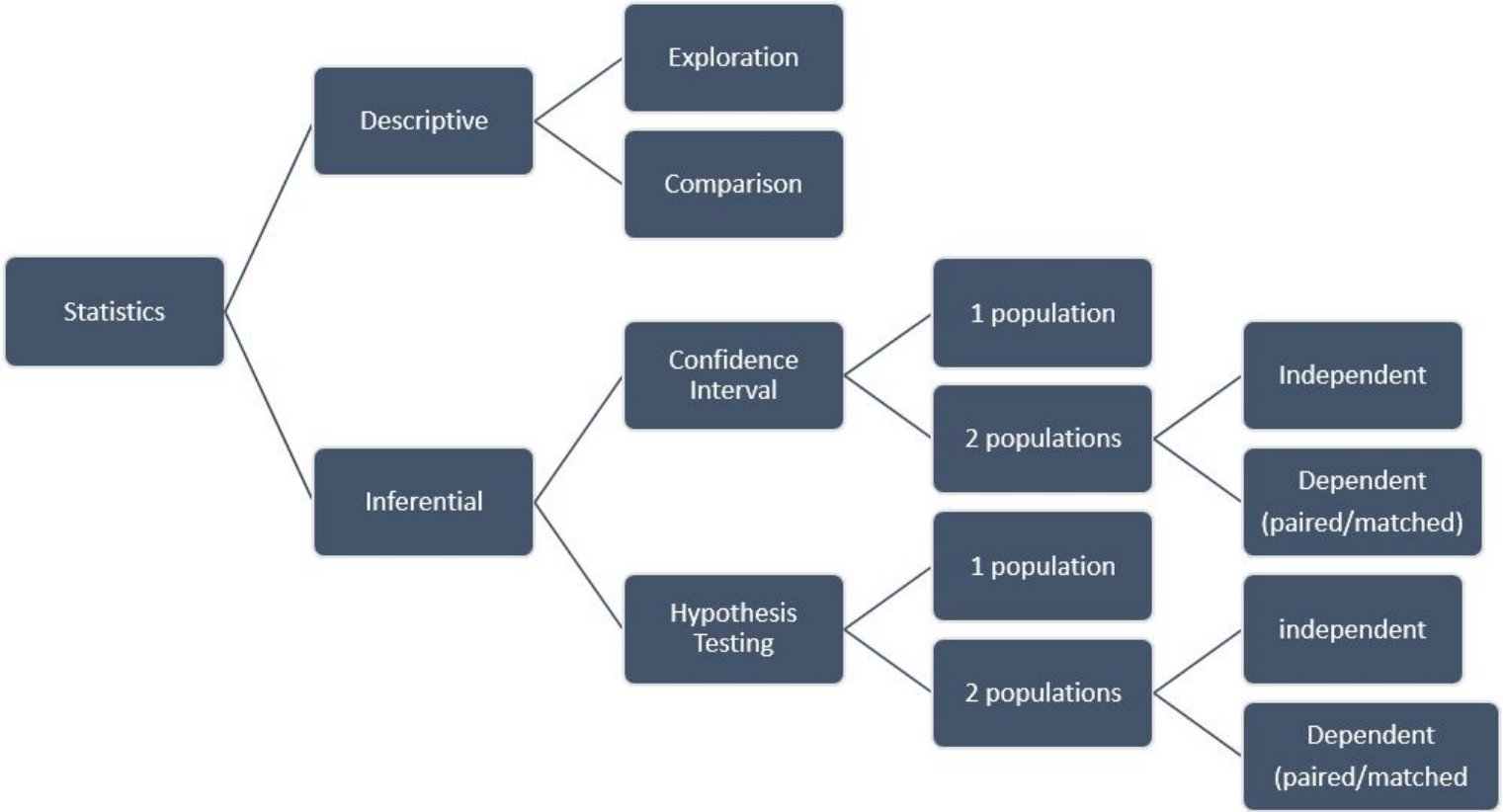
Lecture 09:
Inferences from Two Samples



Applied Statistics - STAN - 5.37 & 5.38
15 & 17 December 2020
Lecturer: Erika Siregar, SST, MS

Overview


Statistics
Descriptive
Exploration
Comparison
Inferential
Confidence Interval
1 population
2 populations
Independent
Dependent (paired/matched)
Hypothesis Testing
1 population
2 populations
independent
Dependent (paired/matched

Today's Agenda

- Inferences About Two Proportions
- Inferences About Two Means: σ_1 and σ_2 Unknown
- Inferences from Dependent Samples
- Comparing Variation in Two Samples

1 pop vs 2 pops

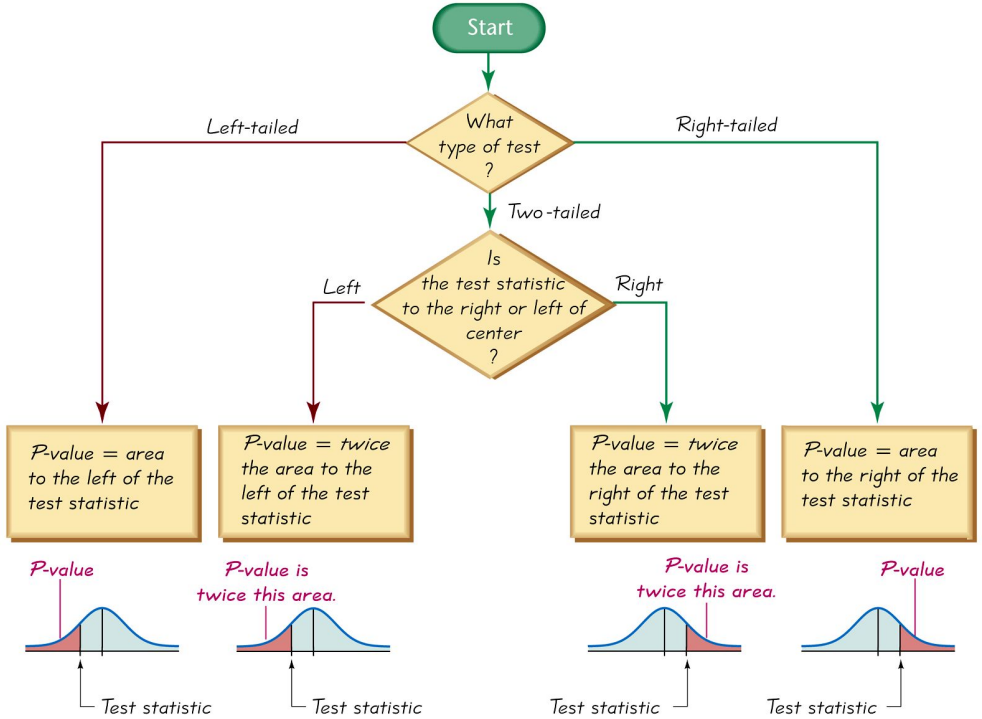
- Hypothesis of 1 population
  - $\mu < 1.000 \rightarrow$ the mean of daily covid19 cases in Indonesia is less than 1.000
  - $\sigma \neq 10 \rightarrow$ the standard deviation of midterm scores is not equal to 10.
- How about 2 populations?
  - Proportion of covid19 patients after vaccination < proportion of covid19 patients before vaccination
  - Rata-rata nilai kelas A yang rutin diberi tugas > rata-rata nilai kelas B yang jarang diberi tugas.

2 populations: Independent vs dependent

1. Independent:
   a. Samples from population 1 are **not related** to samples from population 2.
   b. **n1 ≠ n2** or **n1 = n2**
2. Dependent
   a. Samples from population 1 are **related** to samples from population 2 → ada relationshipnya (matched/paired, etc.)
   b. **n1 = n2**
   c. Example: before/after, husband/wife, mother/daughter, etc

2-Populations Hypothesis Test (Steps)

- Same as **<u>the one with 1 population</u>**, with a slight change for computing the test statistics for both populations (population 1 and population 2)
- Procedure for finding p-value



Hypothesis Test for 2 Populations

No	Estimator	Hypothesis	Test Statistics	CI
1.	Proportion	H0: p1 = p2 H1: p1 ≠ p2 or p1 < p2 or p1 > p2 **Assumption:** p1 - p2 =0	$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$ $\hat{p}_1 = \frac{x_1}{n_1}$ $\hat{p}_2 = \frac{x_2}{n_2}$ $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ $\bar{q} = 1 - \bar{p}$	$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p_1} - \hat{p_2}) + E$ $E = Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
2	Mean with σ known	H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$ **Assumption:** $\mu_1 - \mu_2 = 0$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x_1} - \bar{x_2}) + E$ where $E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Hypothesis Test for 2 Populations (Cont.)

No	Estimator	Hypothesis	Test Statistics	CI
3	Mean with σ unknown	H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$ **Assumption:** $\mu_1 - \mu_2 = 0$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$ where $E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ df = smallest[n_1 - 1, n_2-1]
4	Mean with σ unknown but it's assumed that $\sigma_1 = \sigma_2$	H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ $s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}$ df = $n_1 + n_2 - 2$ S_p^2 = pooled sample variances	$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$ where $E = t_{\alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$ df = $n_1 + n_2 - 2$

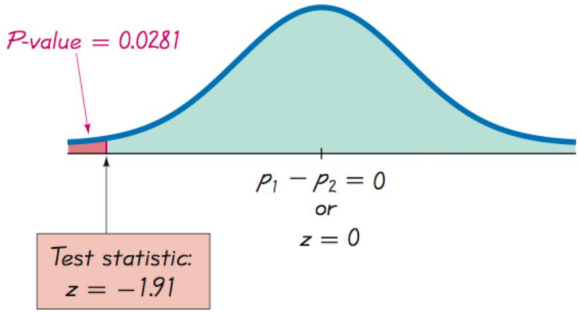
Hypothesis Test for 2 Populations (Cont.)

No	Estimator	Hypothesis	Test Statistics	CI
5	Mean for dependent population (**matched**)	H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$	$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$ Where: df = n-1 n = number of paired data $d = d_1 - d_2$ μ_d = μ of 'd' population $\bar{d}$ = μ of 'd' samples s_d = standard deviation of 'd' samples	$\bar{d} - E < \mu_d < \bar{d} + E$ where $E = t_{\alpha/2} \frac{s_d}{\sqrt{n}}$
6	Comparing Variances of 2 populations	H0: $\sigma_1^2 = \sigma_2^2$ H1: $\sigma_1^2 \neq \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 > \sigma_2^2$	F test → new probability distribution: F $F = \frac{s_1^2}{s_2^2}$ Where: $S_1^2 > S_2^2$	

Example 1 (prop)

The table below lists results from a simple random sample of **front-seat occupants involved in car crashes**. Use a **0.05** significance level to test the claim that the fatality rate of occupants is **lower** for those in **cars equipped with airbags**.

	Airbag Available	No Airbag Available
Occupant Fatalities	41	52
Total number of occupants	11,541	9,853



Answer:

p1: fatalities in cars equipped with airbags

p2: fatalities in cars not equipped with airbags.

H0: p1 = p2 → p1−p2 = 0

H1: p1 < p2 → p1−p2 < 0

α = 0.05

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{41 + 52}{11{,}541 + 9{,}853} = 0.004347$$

$\bar{q}$ = 1 − 0.004347 = 0.995653

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

$$= \frac{\left(\frac{41}{11{,}541} - \frac{52}{9{,}853}\right) - 0}{\sqrt{\frac{(0.004347)(0.995653)}{11{,}541} + \frac{(0.004347)(0.995653)}{9{,}853}}}$$

= **−1.91**

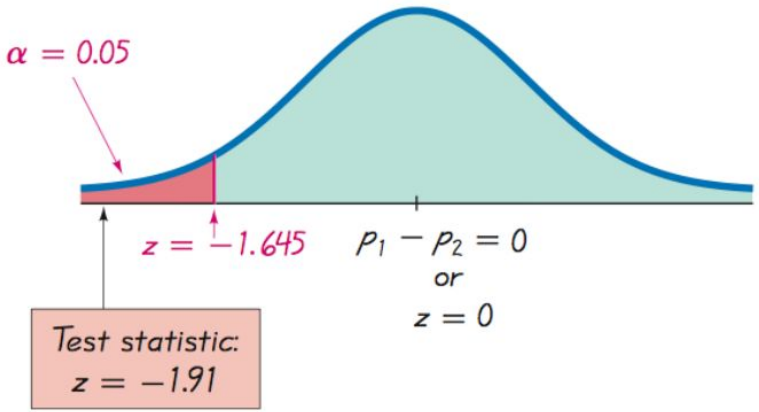
Area to left of z = −1.91 is 0.02806661 → P-value is **0.02806661**. → < 0.05 → **decision??** →

Example 1 (Interpretation)

there is **sufficient evidence** to support the claim that the **proportion of accident fatalities** for occupants in cars **with airbags** is **less than** the proportion of fatalities for occupants in cars **without airbags**. Based on these results, it appears that **airbags are effective in saving lives**.

Now, try with:

- Traditional method



Example 1 (Using Confidence Interval

α =5% → use 90% confidence interval

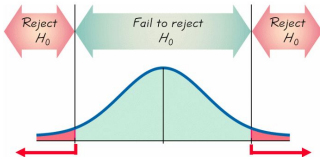
→ α = 0.1 → $z_{\alpha/2} = z_{0.05} = 1.645$

Calculate the margin of error, E

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$= 1.645\sqrt{\frac{\left(\frac{41}{11,541}\right)\left(\frac{11,500}{11,541}\right)}{11,541} + \frac{\left(\frac{52}{9,853}\right)\left(\frac{9801}{9,853}\right)}{9,853}}$$

$$= 0.001507$$



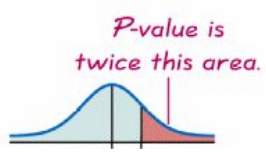


Table 8-2 Confidence Level for Confidence Interval

		Two-Tailed Test	One-Tailed Test
Significance Level for Hypothesis Test	0.01	99%	98%
	0.05	95%	90%
	0.10	90%	80%

Construct Confidence Interval

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

(0.003553 - 0.005278) - 0.001507 < (p1 - p2) < (0.003553 - 0.005278) - 0.001507

-0.003232 < (p1-p2) < -0.000218

```
> n1 <- 11541
> n2 <- 9853
> p1cap <- 41/n1
> q1cap <- (11500)/n1
> p2cap <- 52/n2
> q2cap <- 9801/n2
> print(paste("p1cap=", p1cap, "& p2cap=", p2cap))
[1] "p1cap= 0.00355255177194351 & p2cap=
0.00527758043235563"
> e <- qnorm(0.95)*(sqrt((p1cap*q1cap/n1) +
(p2cap*q2cap/n2)))
> print(paste("e = ", e))
[1] "e =  0.00150711282707152"
> leftCI <- (p1cap - p2cap) - e
> rightCI <- (p1cap - p2cap) + e
> print(paste(round(leftCI,6), " < (p1-p2) < ", round(rightCI,6)))
[1] "-0.003232  < (p1-p2) <  -0.000218"
```

Example 1 (Cont.)

- -0.003232 < (p1-p2) < -0.000218
- The CI does not contain 0 ⟶ there is a significant difference between the two proportions.
- the **fatality rate is lower** for occupants in cars **with air bags** than for occupants in cars without air bags.
- The CI also provides an estimate of the amount of the difference between the two fatality rates.

- If you want to test a claim about two population proportions, use the P-value method or traditional method; if you want to estimate the difference between two population proportions, use a confidence interval.
- the **distribution of p1 – p2** is approximately **normal**, with **mean p1 – p2** and standard deviation is the sum of their individual variances.

$$\sigma^2_{(\hat{p}_1 - \hat{p}_2)} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \quad \rightarrow \quad \sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}}$$

- When constructing CI of the difference between two proportions, we don't assume that the two proportions are equal. Thus, we estimate the standard deviation as

$$\sigma = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example 2

A headline in USA Today proclaimed that "**Men, women are equal talkers**." That headline referred to a study of the numbers of words that samples of men and women spoke in a day. Given below are the results from the study. Use a **0.05** significance level to test the claim that men and women speak the same **mean number of words** in a day. Does there appear to be a difference?

Number of Words Spoken in a Day			
	Men		Women
n_1	= 186	n_2	= 210
$\bar{x}_1$	= 15,668.5	$\bar{x}_2$	= 16,215.0
s_1	= 8632.5	s_2	= 7301.2

Answer:

H0: $\mu_1 = \mu_2$

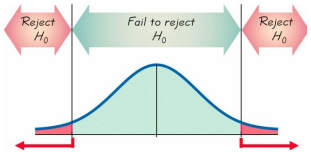
H1: $\mu_1 \neq \mu_2$

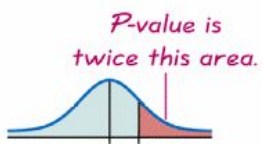
$\alpha = 0.05$

Computing test statistics

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(15{,}668.5 - 16{,}215.0) - 0}{\sqrt{\frac{8632.5^2}{186} + \frac{7301.2^2}{210}}} = -0.676$$





Computing p-value from test statistics

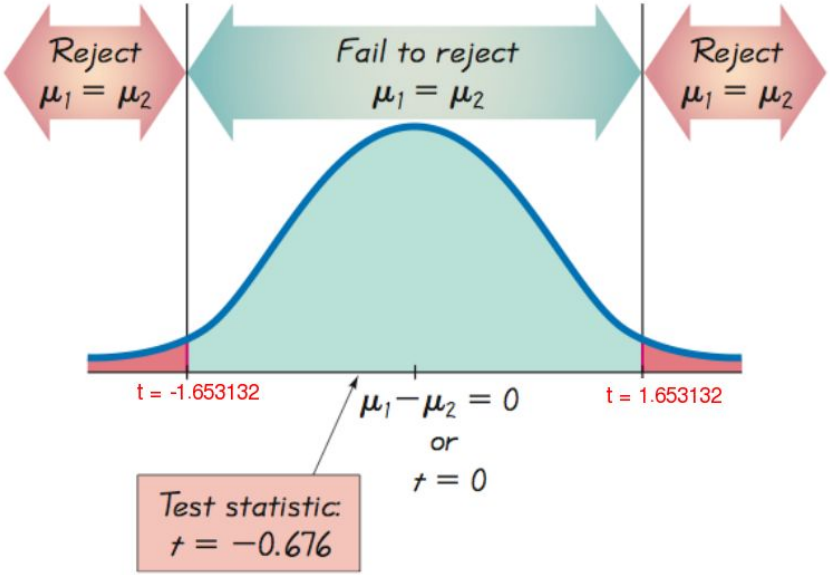
```
> pt(-0.676, 185)
[1] 0.2499424
> 2*pt(-0.676, 185)
[1] 0.4998848
```

P-value = 0.4998848 → > α → fail to reject H0

Example 2 (cont.)

Computing critical value from α = 0.05

```
> qt(p, df, lower.tail = TRUE)
> qt(0.05, 185, lower.tail = TRUE)
[1] -1.653132
> qt(0.05, 185, lower.tail = FALSE)
[1] 1.653132
```



Compare test statistics and critical value

Because the test statistic does not fall within the critical region → fail to reject H0

Interpretation

- There is not sufficient evidence to reject the claim that men and women speak the same mean number of words in a day.
- There does not appear to be a significant difference between the two means.

Example 2 (cont.)

Using the same sample data, construct a **95% confidence interval** estimate of the difference between the mean number of words spoken by men and the mean number of words spoken by women.

Number of Words Spoken in a Day	
Men	Women
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1 = 15{,}668.5$	$\bar{x}_2 = 16{,}215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

Answer:

Use $t_{\alpha/2} = 1.653132$

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.653132\sqrt{\frac{8632.5^2}{186} + \frac{7301.2^2}{210}} = 1{,}337.394$$

The CI:

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

$-1883.894 < (\mu 1 - \mu 2) < 790.894$

Example 3

Use the sample data in Table 9-1 with a **0.05** significance level to test the claim that for the population of students, the **mean change in weight** from September to April is **equal to 0** kg.

Table 9-1 Weight (kg) Measurements of Students in Their Freshman Year

Sample from population 1	April weight	66	52	68	69	71
Sample from population 2	September weight	67	53	64	71	70
	Difference d = (April weight) − (September weight)	−1	−1	4	−2	1

Answer:

H0: $\mu_d = 0$

H1: $\mu_d \neq 0 \rightarrow$ 2 tails

$\alpha = 0.05$

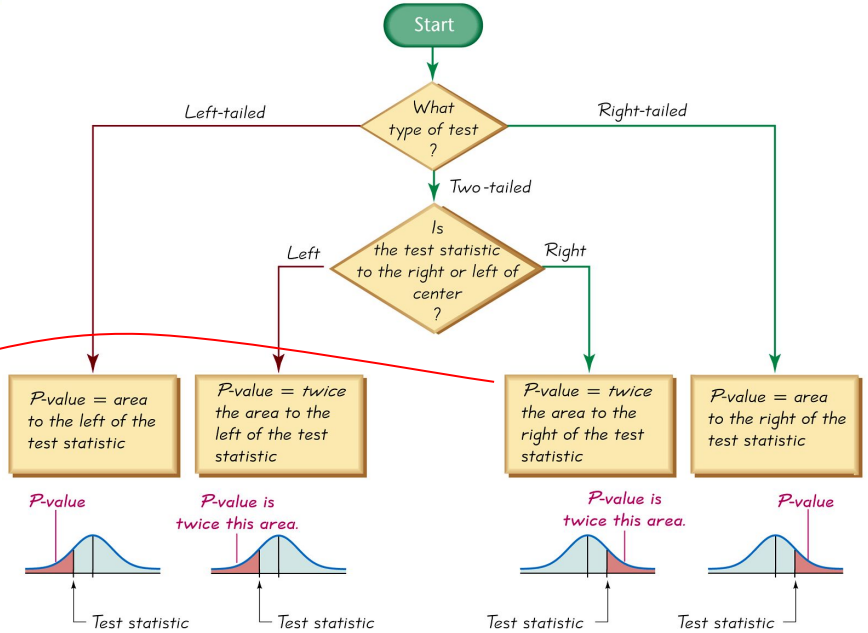
Test Statistics

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{0.2 - 0}{\frac{2.4}{\sqrt{5}}} = 0.186$$

P-value vs α

> p-value of the test statistics
> pt(0.186, 4)
[1] 0.5692518
> #test statistics twice area to the right of center
> 2*(1 - pt(0.186, 4))
[1] 0.8614964

P-value > α → fail to reject H0

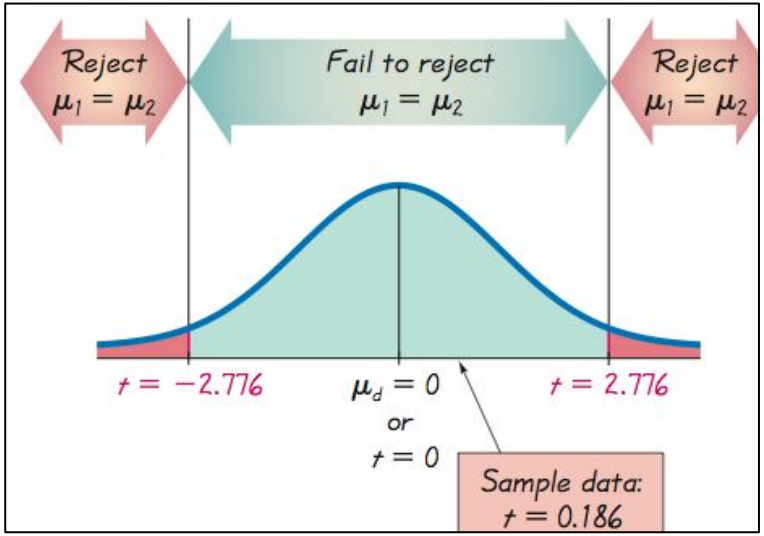


Example 3 (critical value method)

Test Statistics vs Critical Value

Critical Value (using R):
> α =0.05
> qt(0.025, df=4)
[1] -2.776445

REMEMBER: 2-tailed test has 2 critical values: **-2.776445 & 2.776445**



Because the test statistic does not fall in the critical region, we **fail to reject H0**.

Interpretation:
We conclude that there is **not sufficient evidence** to warrant rejection of the claim that for the population of students, the **mean change** in weight from September to April is **equal to 0 kg**. There **does not appear** to be a **significant weight gain** from September to April.

Example 3 (CI method)

Construct a 95% confidence interval estimate of d , which is the mean of the "April–September" weight differences of college students in their freshman year.

$\bar{d}$ = 0.2, s_d = 2.4, n = 5, $t_{a/2}$ = 2.776

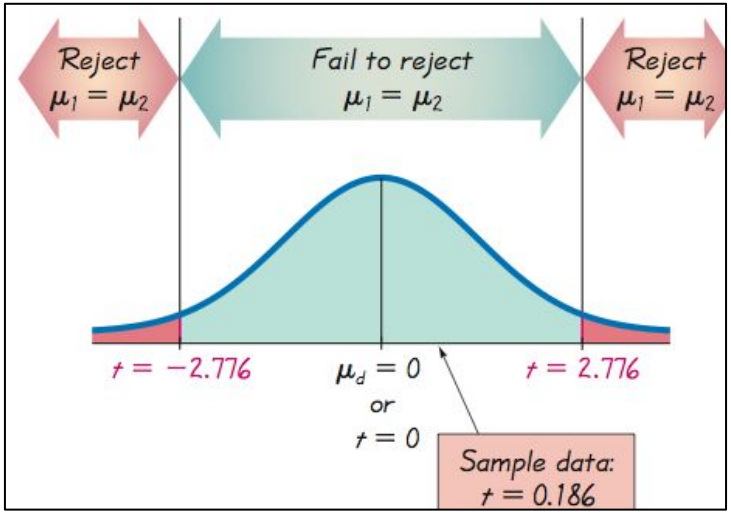
$$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$ = 2.776*2.4/√5 = 2.979516

CI:

$$\bar{d} - E < \mu_d < \bar{d} + E$$

0.2 - 2.979516 < μ_d < 0.2 + 2.979516

-2.779516 < μ_d < 3.179516



Interpretation:
We have 95% confidence that the **limits** of -2.779516 kg and 3.179516 kg c**ontain the true value** of the mean weight change from September to April. In the long run, 95% of such samples will lead to confidence interval limits that actually do contain the true population mean of the differences. → **fail to reject H0**.

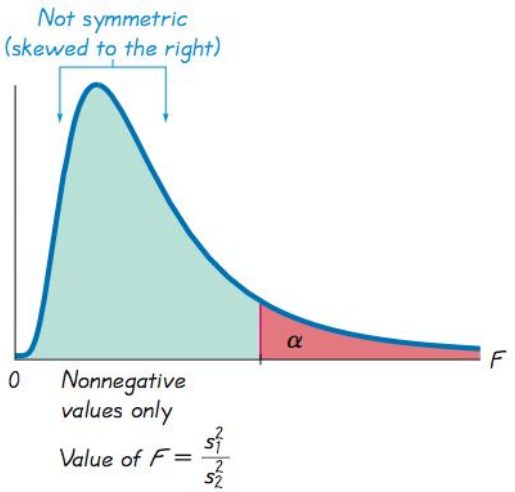
Hypothesis vs CI

- **hypothesis test** and **confidence interval** are equivalent in the sense that they result in the **same conclusions**.
- Consequently, **H0: $\mu_1 - \mu_2 = 0$** can be tested by determining whether the **confidence interval includes 0**.

Several Continuous Prob. Distribution

No	Criteria	z	t	$\mathcal{X}^2$	F
1	curve	Symmetric, Bell shaped	Symmetric, Bell shaped	asymmetric	Asymmetric. The exact shape of the F distribution depends on the two different degrees of freedom
2	Negative value	yes	yes	no	no
2	μ	0	0	varies with sample size n	varies with sample size n
3.	σ	1	> 1, varies with sample size *n*	varies with sample size n	varies with sample size n
4	**df**	**no**	**yes**	**yes**	**Yes (2)**
5	Table reading	Area to the left of z	Area to the left of t	Area to the right of $\mathcal{X}^2$	
5	R syntax	pnorm(z), qnorm(p)	pt(), qt()	pchisq(), qchisq()	pf(df1,df2) qf()

F curve



back

Some Properties of F

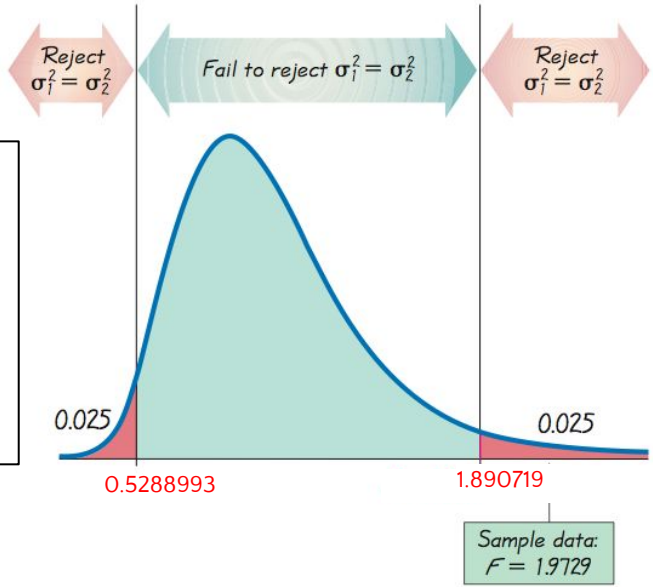
- large value of F will be evidence against the conclusion of equality of the population variances .
- Consequently, a value of F near 1 will be evidence in favor of the conclusion that $\sigma_1^2 = \sigma_2^2$.

Example 4

Take a look at sample statistics listed below. When designing coin vending machines, we must consider the standard deviations of pre-1964 quarters and post-1964 quarters. Use a 0.05 significance level to test the claim that the weights of pre-1964 quarters and the weights of post-1964 quarters are from populations with the **same standard deviation**.

Pre-1964 Quarters	Post-1964 Quarters
$n = 40$ $s = 0.08700$ g	$n = 40$ $s = 0.06194$ g

$$F = \frac{s_1^2}{s_2^2} = \frac{0.08700^2}{0.016194^2} = 1.9729$$



P-value vs α

> p-value of the test statistics
> pf(1.9729, 39, 39)
[1] 0.9816261
> #test statistics twice area to the right of center
> 2* (1 - pf(1.9729, 39, 39))
[1] 0.03674778
P-value < α → reject H0

Critical Values:

> qf(0.025, 39, 39, lower.tail = TRUE)
[1] **0.5288993**
> qf(0.025, 39, 39, lower.tail = FALSE)
[1] **1.890719**

Try the CI method !!!

Exercise 1

Berikut adalah data tentang volume otak (cm^3) dari twin babies pada saat dilahirkan. Ujilah hipotesis bahwa tidak ada perbedaan pada rata-rata volume otak 2 bayi kembar yang dilahirkan.

1st born	1005	1035	1281	1051	1034	1079	1104	1439	1029	1160
2nd born	963	1027	1272	1079	1070	1173	1067	1347	1100	1204

Exercise 2

Sebuah survei dilakukan untuk menentukan efektivitas penggunaan kelambu terhadap pencegahan malaria. Diketahui dari 343 bayi yang menggunakan kelambu, 15 terkena malaria. Sementara 294 bayi yang tidur tanpa kelambu, 27 mengalami malaria. Dengan significance level 0.01, ujilah claim yang mengatakan bahwa *incidence of malaria* akan turun dengan penggunaan kelambu.

Exercise 3

Sebuah survei dilakukan untuk membandingkan kadar nikotin dalam darah orang yang **tidak merokok tapi sehari-hari terekspos dengan asap rokok**, dengan orang yang **tidak merokok dan tidak terekspos dengan asap rokok**.

Untuk kelompok I, diperoleh statistik sebagai berikut: n= 40, xbar = 60.58, s = 138.08.

Untuk kelompok II, diperoleh statistik sebagai berikut: n = 40, xbar= 16.35, s = 62.53.

Ujilah klaim yang menyatakan bahwa nonsmokers namun terekspos asap rokok memiliki rata-rata kadar nikotin dalam darah lebih tinggi dibandingkan nonsmokers yang tidak terekspos asap rokok. Diasumsikan kedua kelompok memiliki varians populasi yang sama dan significance level 0.05.

Thanks!

Any questions ?