week 1.

Data format
- numerical
- categorical

Data type
- univariate / *multivariate

	age	bp	gender	$
obs 1	—	—	—	—

- longitudinal

	bp			
obs 1	bp (w1)	bp (w2)	..	bp (wn)

  - functional
  - spatial
  - image
  - spatio-temporal
  - text

⊛ [Statistical Inference] [prediction]

- Hypothesis testing
- Conclusion

Data

	Smoking	lung cancer
	never	0
	1-3	1
100	3-10	0
	⋮	⋮

model

(1-3)/day → ? → 0.2

Claim: Smoking ↑ : lung cancer ↑

[H_0 : Smoke ⊄ lung

H_a or H_1 : ↑ Smoke : ↑ lung cancer.

1. what is your claim?

[H_0 : male = Female

H_1 : male > Female]

100

male

Femal

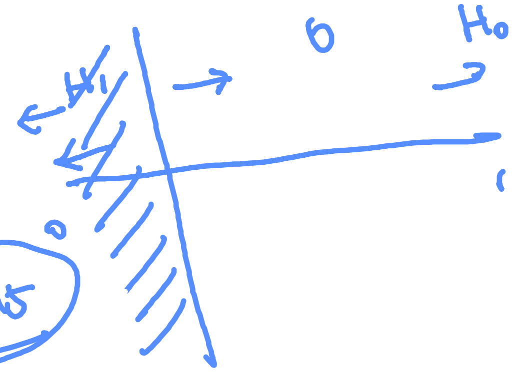
2. Choose proper test & perform

n

t-statistic =

Male	Female
$200	$10
$200	$195

$200 − $195 = $5



3. P-value

0 < <1

1%

0.01

4. Make a conclusion

n = 15, 20 small

0.05 → p

0.1

very large p-values

100.000

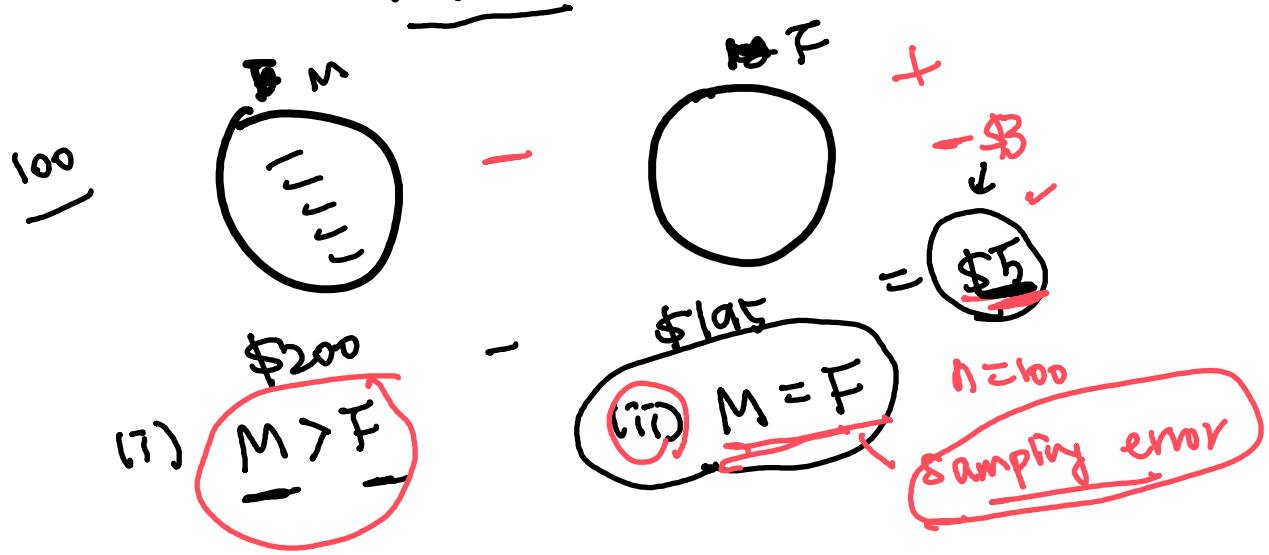
0.01

p ~ 0.056

- Statistically significant difference

↓

difference



- p-value & Effect size (n)

p-value: 0.00001

$M_M - M_F = \$1$

p-val: 0.045

$= \$20$

R

(i) install & call packages

(ii) Data import

(iii) Data sprit, filtering etc.