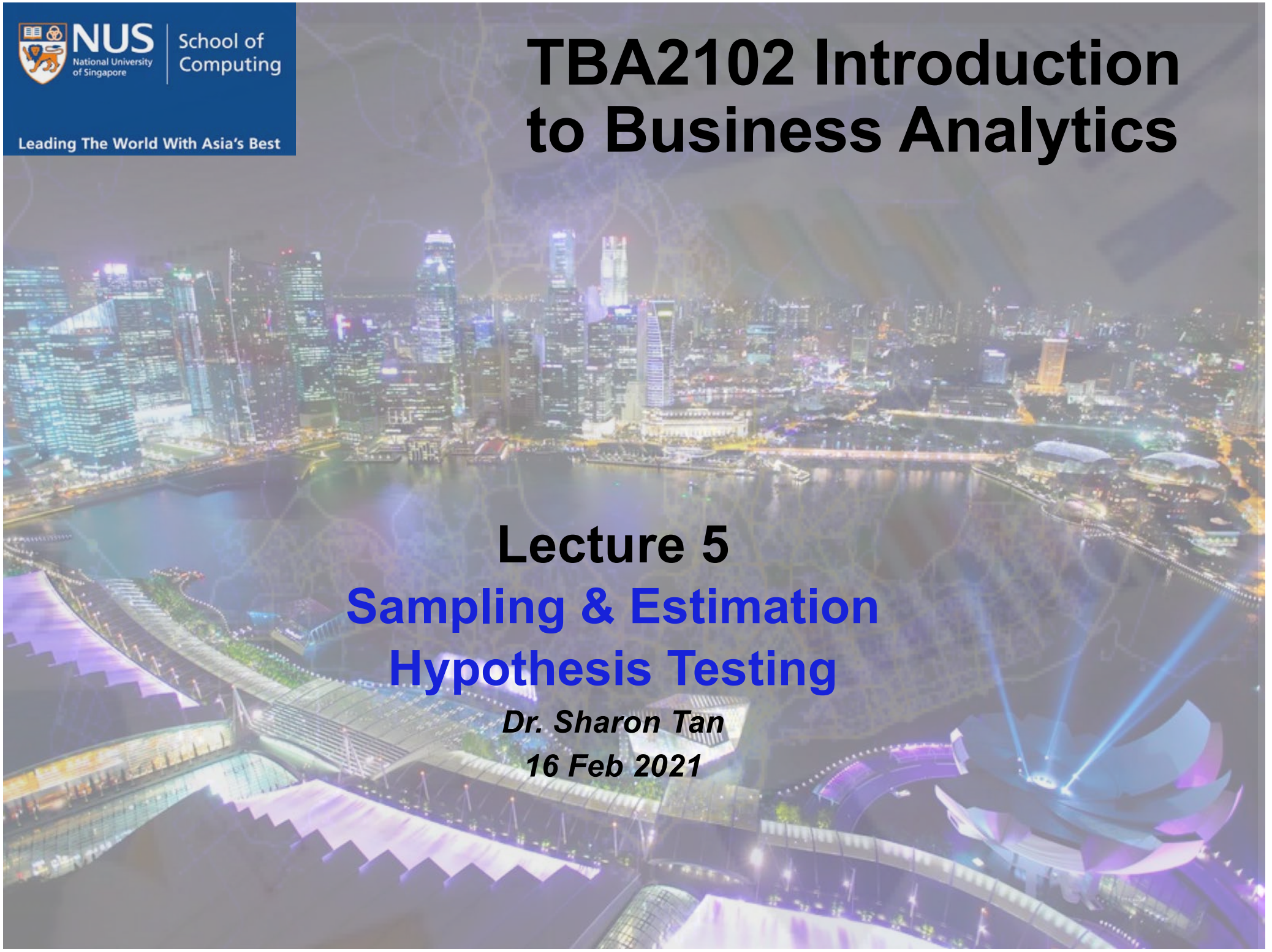
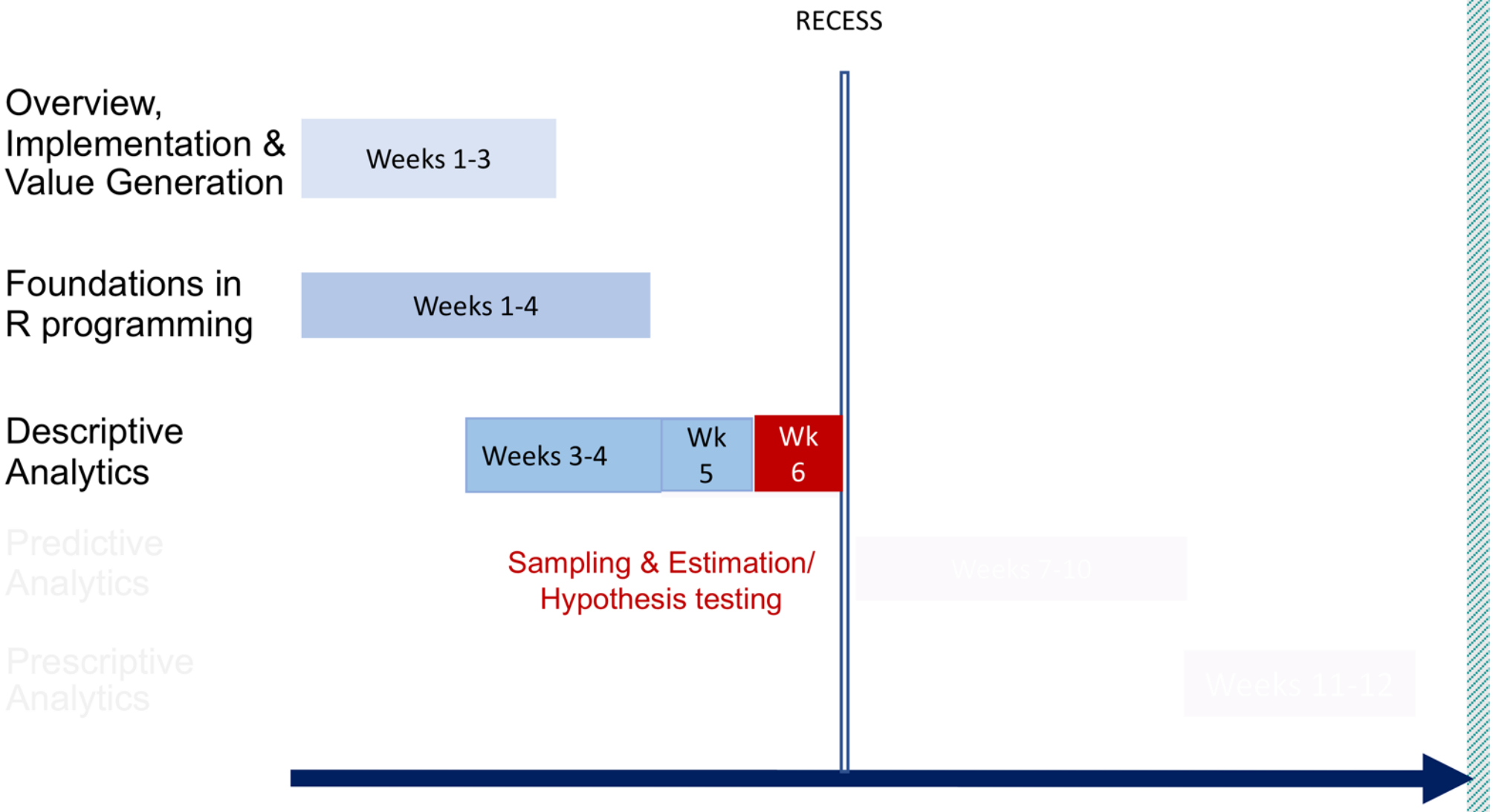
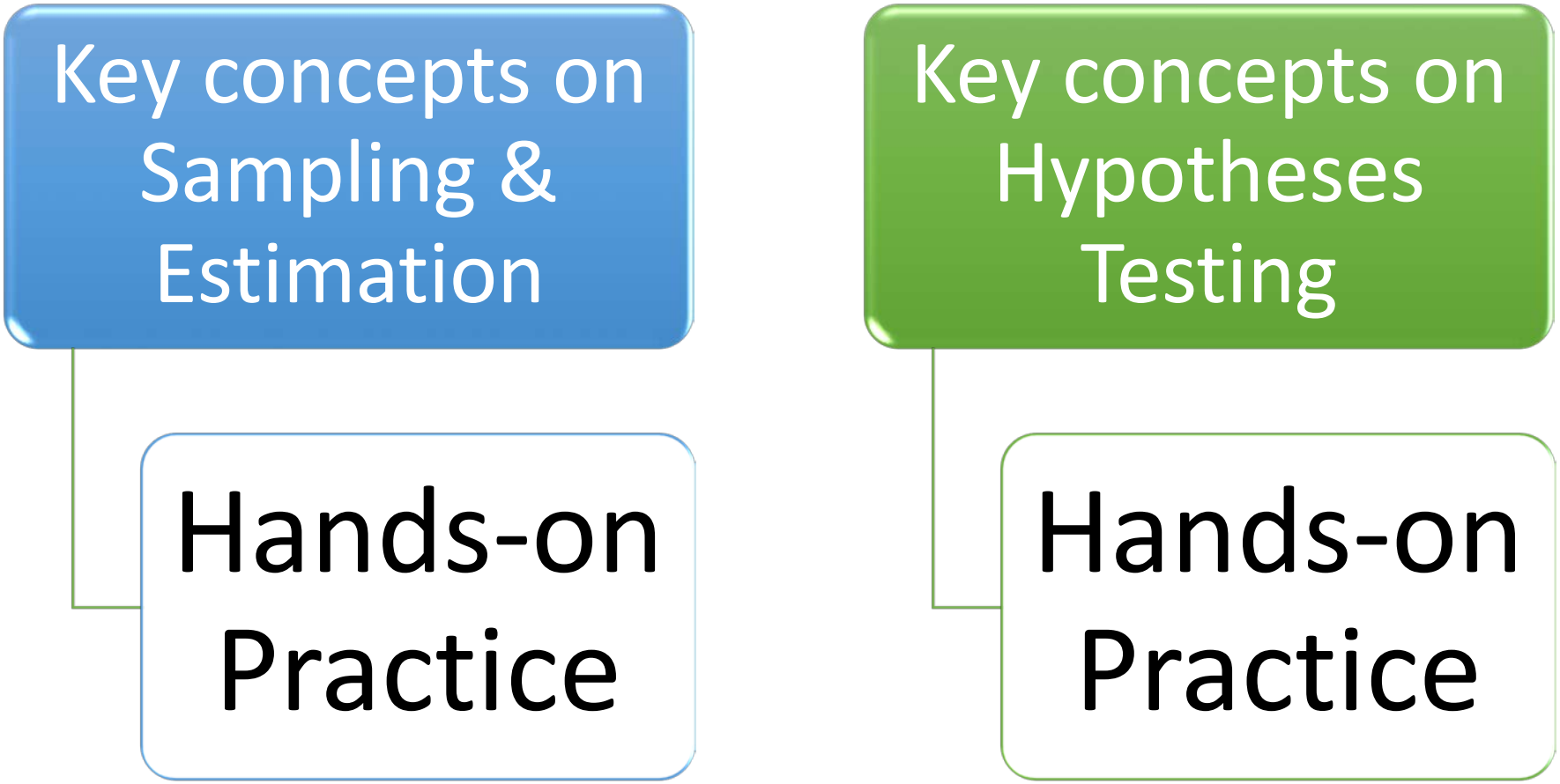

NUS
National University of Singapore
School of Computing
Leading The World With Asia's Best
TBA2102 Introduction to Business Analytics
Lecture 5
Sampling & Estimation
Hypothesis Testing
Dr. Sharon Tan
16 Feb 2021

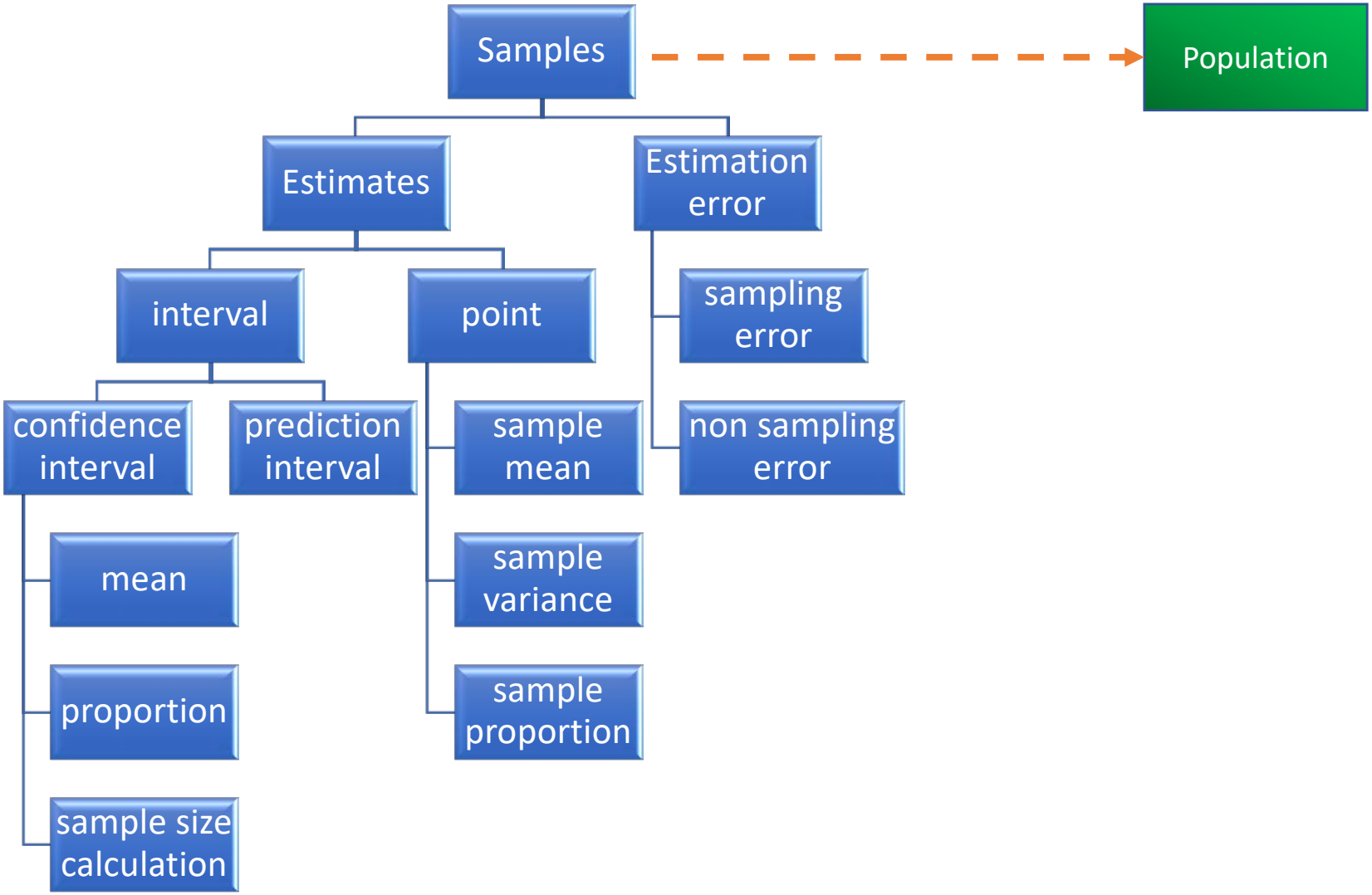
Course Map


RECESS
Overview, Implementation & Value Generation
Weeks 1-3
Foundations in R programming
Weeks 1-4
Descriptive Analytics
Weeks 3-4
Wk 5
Wk 6
Sampling & Estimation/ Hypothesis testing
Predictive Analytics
Weeks 7-10
Prescriptive Analytics
Weeks 11-12

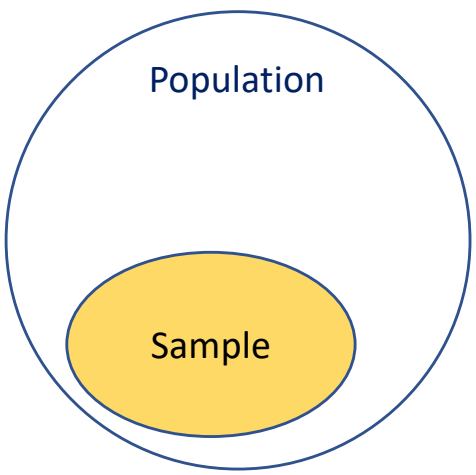
Outline for today



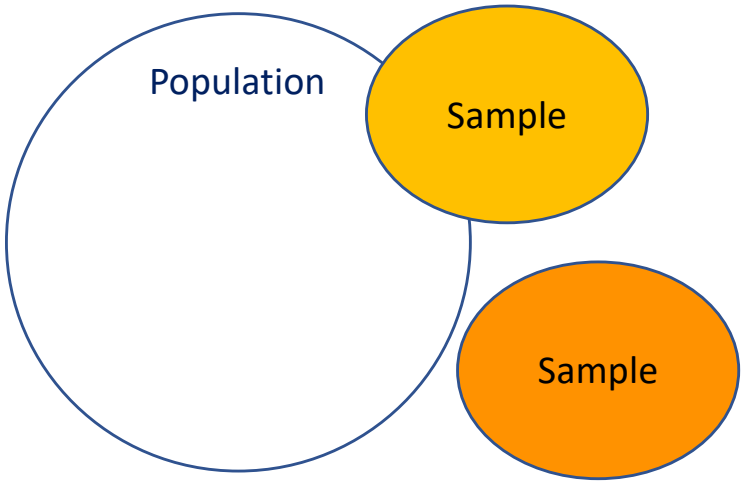
Key Concepts on Sampling and Estimation



Key Concepts on Sampling and Estimation



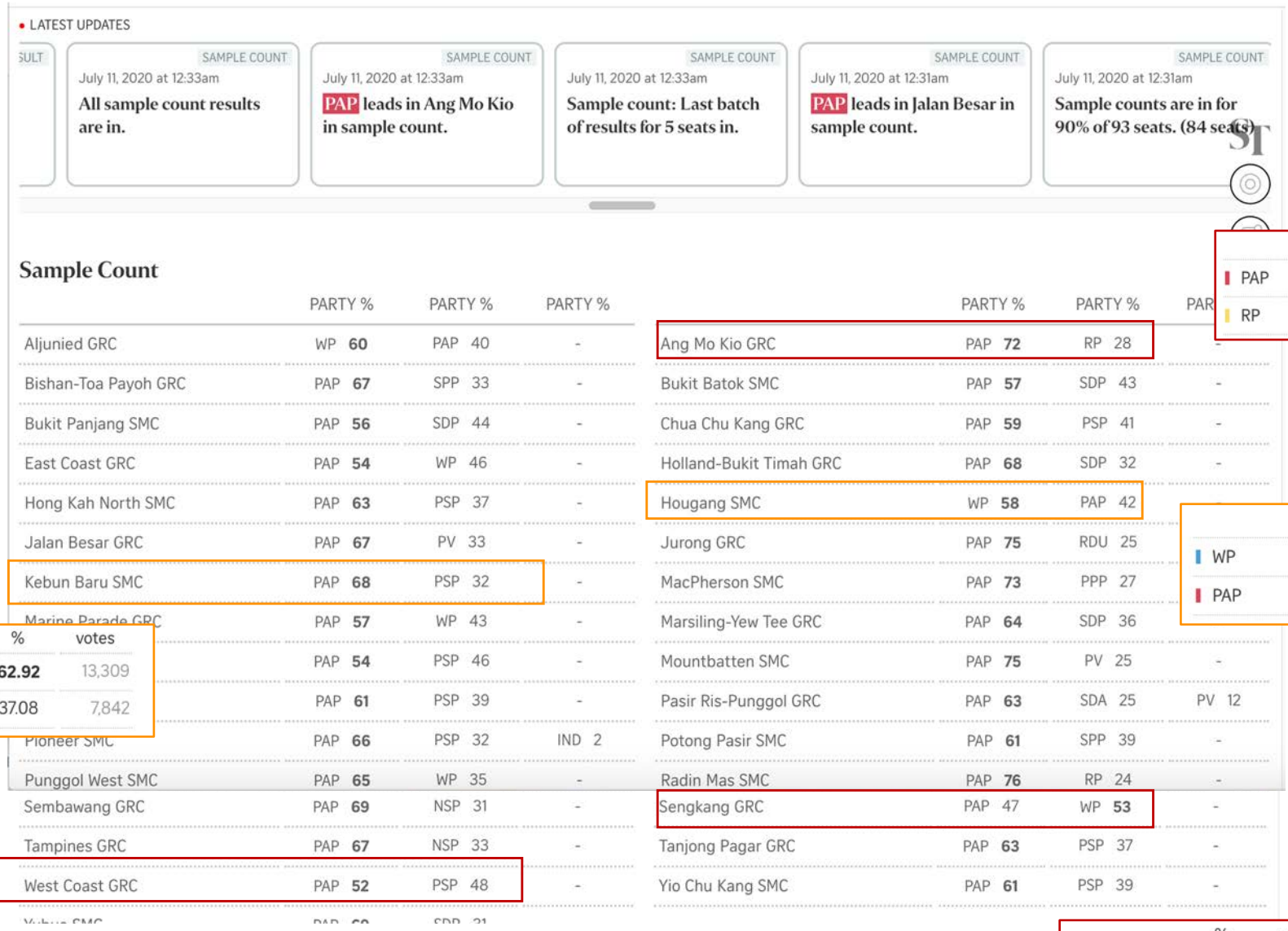
Sample is a subset of population



Sample not representative of population
(e.g. convenience sample)

SOURCE: https://www.straitstimes.com/multimedia/graphics/2020/07/singapore-general-election-ge2020-live-results/index.html

GE2020 Singapore



• LATEST UPDATES

SAMPLE COUNT · July 11, 2020 at 12:33am
All sample count results are in.

SAMPLE COUNT · July 11, 2020 at 12:33am
PAP leads in Ang Mo Kio in sample count.

SAMPLE COUNT · July 11, 2020 at 12:33am
Sample count: Last batch of results for 5 seats in.

SAMPLE COUNT · July 11, 2020 at 12:31am
PAP leads in Jalan Besar in sample count.

SAMPLE COUNT · July 11, 2020 at 12:31am
Sample counts are in for 90% of 93 seats. (84 seats)

Sample Count

	PARTY %	PARTY %	PARTY %
Aljunied GRC	WP **60**	PAP 40	-
Bishan-Toa Payoh GRC	PAP **67**	SPP 33	-
Bukit Panjang SMC	PAP **56**	SDP 44	-
East Coast GRC	PAP **54**	WP 46	-
Hong Kah North SMC	PAP **63**	PSP 37	-
Jalan Besar GRC	PAP **67**	PV 33	-
Kebun Baru SMC	PAP **68**	PSP 32	-
Marine Parade GRC	PAP **57**	WP 43	-
[illegible]	PAP **54**	PSP 46	-
[illegible]	PAP **61**	PSP 39	-
Pioneer SMC	PAP **66**	PSP 32	IND 2
Punggol West SMC	PAP **65**	WP 35	-
Sembawang GRC	PAP **69**	NSP 31	-
Tampines GRC	PAP **67**	NSP 33	-
West Coast GRC	PAP **52**	PSP 48	-
Yuhua SMC	PAP **60**	SDP 31	-

	PARTY %	PARTY %	PAR
Ang Mo Kio GRC	PAP **72**	RP 28	-
Bukit Batok SMC	PAP **57**	SDP 43	-
Chua Chu Kang GRC	PAP **59**	PSP 41	-
Holland-Bukit Timah GRC	PAP **68**	SDP 32	-
Hougang SMC	WP **58**	PAP 42	
Jurong GRC	PAP **75**	RDU 25	
MacPherson SMC	PAP **73**	PPP 27	
Marsiling-Yew Tee GRC	PAP **64**	SDP 36	
Mountbatten SMC	PAP **75**	PV 25	-
Pasir Ris-Punggol GRC	PAP **63**	SDA 25	PV 12
Potong Pasir SMC	PAP **61**	SPP 39	-
Radin Mas SMC	PAP **76**	RP 24	-
Sengkang GRC	PAP 47	WP **53**	-
Tanjong Pagar GRC	PAP **63**	PSP 37	-
Yio Chu Kang SMC	PAP **61**	PSP 39	-

	%	votes
PAP	**71.91**	124,597
RP	28.09	48,677

	%	votes
WP	**61.21**	15,451
PAP	38.79	9,791

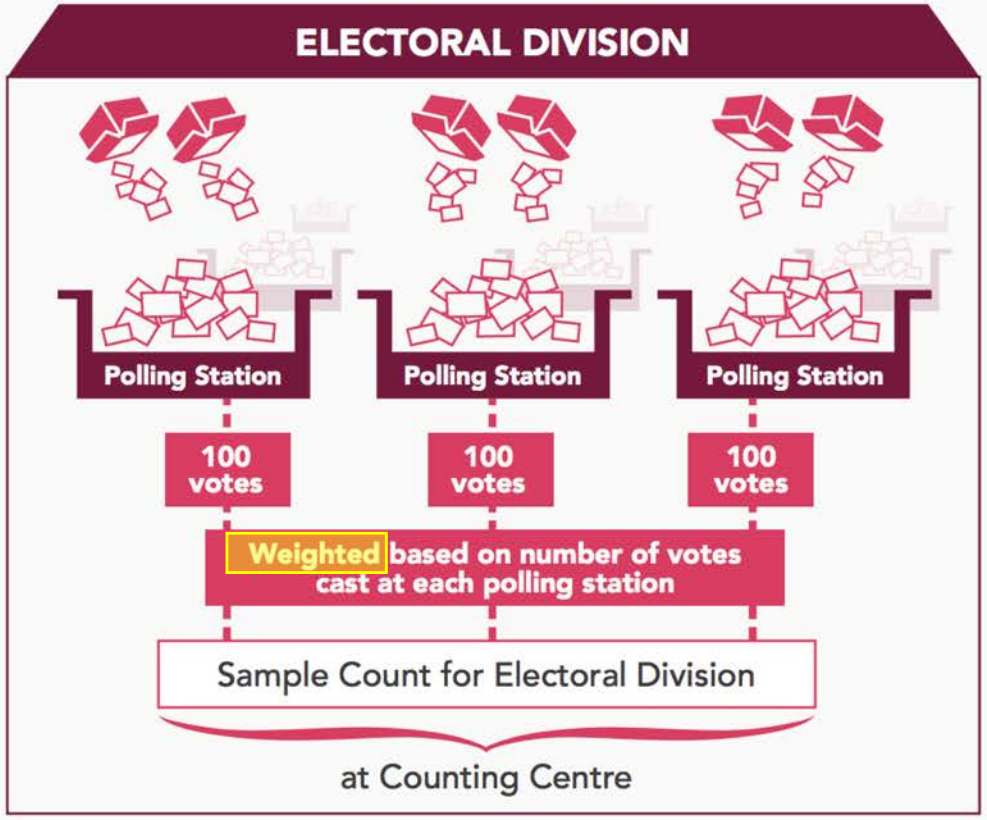
	%	votes
PAP	**62.92**	13,309
PSP	37.08	7,842

	%	votes
PAP	**51.68**	71,658
PSP	48.32	66,996

	%	votes
WP	**52.12**	60,217
PAP	47.88	55,319

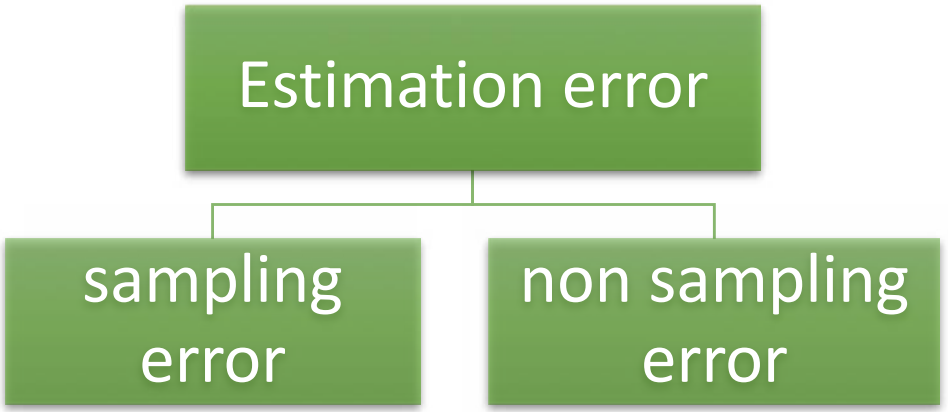
Source: https://www.eld.gov.sg/mediarelease/SampleCount_Generic.pdf

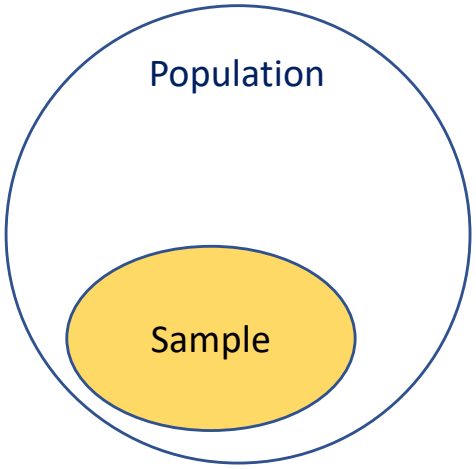
GE2020 Sample Count



- From the votes cast at each polling station, a counting assistant picks up a random bundle of 100 ballot papers (in front of the candidates and counting agents present) and counts the number of votes for each candidate (or group of candidates in the case of a GRC).
- The votes will be added up, with weightage given to account for the difference in the number of votes cast at each polling station.
- Sample count for the electoral division will be shown as a percentage of valid votes garnered by each candidate (or group of candidates).

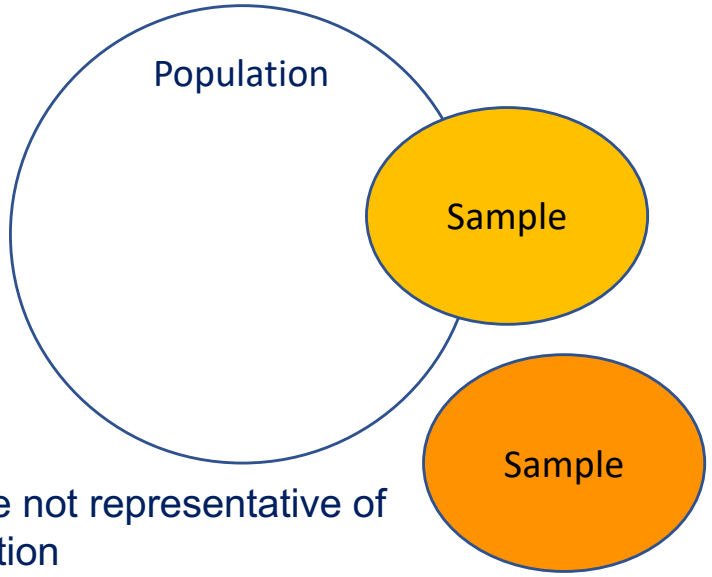
Key Concepts on Sampling and Estimation





Sample is a subset of population

Increase sample size



Sample not representative of population

unbiased sampling

Key Concepts on Sampling and Estimation

Sample Estimates

- interval
  - confidence interval
  - prediction interval
  - sample size calculation
    - mean
    - proportion
- point
  - sample mean — $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ — eg: salary, scores
  - sample std deviation — $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
  - sample proportion — $p = \frac{x}{n}$, where x: # successes, n: sample size — eg: votes

Sample size for mean:

$$n \geq (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$$

Sample size for proportion:

$$n \geq (z_{\alpha/2})^2 \frac{\pi(1-\pi)}{E^2}$$

Key Concepts on Sampling and Estimation

interval estimates

provides a range for a population characteristic based on a sample

confidence interval

range of values between which the value of the population parameter is believed to be, along with a probability that the interval correctly estimates the true (unknown) population parameter

mean

A 100(1 – α)% C.I. for mean with known population sd: $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$

A 100(1 – α)% C.I. for mean with unknown population sd: $\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n})$

proportion

A 100(1 – α)% C.I. for proportion:

$$\hat{p} \pm z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

prediction interval

range for predicting value of a new observation from same population.

While confidence interval is associated with sampling distribution of a statistic, a prediction interval is associated with the distribution of random variable itself.

A 100(1 – α)% prediction interval for a new observation:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1+\frac{1}{n}}\right)$$

Sampling Distribution of Means & Central Limit Theorem

- Sampling distribution of mean:
  - Distribution of means of all possible samples of a fixed size *n* from some population (sample size is n, not n number of samples)
- Standard error of the mean:
  - Standard deviation of sampling distribution of the mean

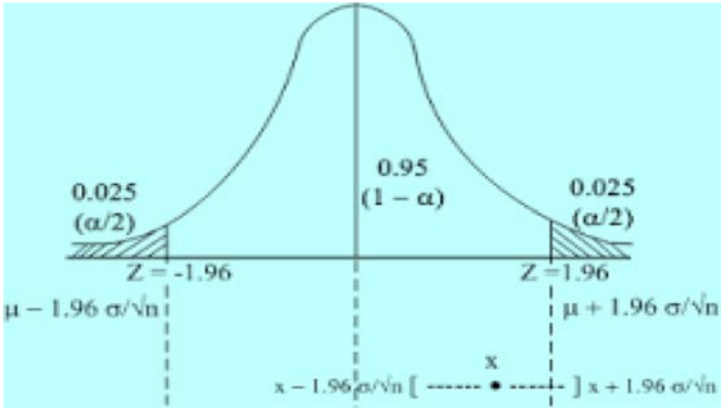
$\sigma/\sqrt{n}$ as n increases, se decreases diminishing returns

- Frequency/Probability distribution of random variable
- Standard deviation of variable σ

How does Central Limit Theorem (CLT) apply?

Theorem states:
- If sample size is large enough, sampling distribution of the mean is approximately normally distributed *regardless* of population distribution, and sample means will be equal to population mean
- If population is normally distributed, sampling distribution is also normally distributed for any sample size

CLT allows us to
- make assumption about the distribution of the sampling means
- use theory on computing probabilities for normal distributions to draw conclusions about sample means



Hands-on Practice

Bank Credit Risk Analyses

Bank Credit Risk Data

	Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
4	Small Appliance	$0	$739	13	12	M	Single	23	Own	3	Unskilled	Low
5	Furniture	$0	$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
6	New Car	$0	$389	19	119	M	Single	38	Own	4	Management	High
7	Furniture	$638	$347	13	14	M	Single	36	Own	2	Unskilled	High
8	Education	$963	$4,754	40	45	M	Single	31	Rent	3	Skilled	Low
9	Furniture	$2,827	$0	11	13	M	Married	25	Own	1	Skilled	Low
10	New Car	$0	$229	13	16	M	Married	26	Own	3	Unskilled	Low
11	Business	$0	$533	14	2	M	Single	27	Own	1	Unskilled	Low
12	Small Appliance	$6,509	$493	37	9	M	Single	25	Own	2	Skilled	High
13	Small Appliance	$966	$0	25	4	F	Divorced	43	Own	1	Skilled	High
14	Business	$0	$989	49	0	M	Single	32	Rent	2	Management	High
15	New Car	$0	$3,305	11	15	M	Single	34	Rent	2	Unskilled	Low
16	Business	$322	$578	10	14	M	Married	26	Own	1	Skilled	Low
17	New Car	$0	$821	25	63	M	Single	44	Own	1	Skilled	High
18	New Car	$396	$228	13	26	M	Single	46	Own	3	Unskilled	Low
19	Used Car	$0	$129	31	8	M	Divorced	39	Own	4	Management	Low
20	Furniture	$652	$732	49	4	F	Divorced	25	Own	2	Skilled	High
21	New Car	$708	$683	13	33	M	Single	31	Own	2	Skilled	Low
22	Repairs	$207	$0	28	116	M	Single	47	Own	4	Skilled	Low
23	Education	$287	$12,348	7	2	F	Divorced	23	Rent	2	Skilled	High
24	Furniture	$0	$17,545	34	16	F	Divorced	22	Own	4	Skilled	High
25	Furniture	$101	$3,871	13	5	F	Divorced	26	Rent	4	Skilled	High
26	Furniture	$0	$0	25	23	M	Married	19	Own	4	Skilled	High
27	Furniture	$0	$485	37	23	F	Divorced	27	Own	2	Management	High
28	New Car	$0	$10,723	11	15	M	Single	39	Rent	2	Unskilled	Low
29	Business	$141	$245	22	33	M	Single	26	Own	3	Skilled	Low
30	Used Car	$0	$0	19	58	M	Single	50	Other	4	Skilled	High
31	Used Car	$2,484	$0	49	46	M	Single	34	Other	1	Skilled	Low
32	Small Appliance	$237	$236	37	24	M	Single	23	Rent	4	Skilled	Low
33	Small Appliance	$0	$485	19	12	M	Single	23	Own	2	Skilled	Low
34	Education	$335	$1,708	37	7	M	Single	46	Other	4	Skilled	High
35	Small Appliance	$3,565	$0	31	32	M	Single	35	Own	3	Skilled	Low
36	Small Appliance	$0	$407	13	2	F	Divorced	28	Own	2	Skilled	Low
37	Business	$16,647	$895	16	34	M	Single	25	Rent	4	Skilled	Low
38	Business	$0	$150	49	46	F	Divorced	36	Rent	4	Skilled	High
39	Small Appliance	$0	$490	5	41	M	Single	41	Own	1	Unskilled	Low
40	Furniture	$0	$162	25	1	M	Divorced	54	Own	1	Skilled	High
41	Small Appliance	$940	$715	9	40	F	Divorced	43	Own	2	Unskilled	Low
42	Small Appliance	$0	$323	49	42	M	Married	33	Own	1	Skilled	High
43	New Car	$0	$128	13	74	M	Single	34	Own	3	Skilled	High
44	Other	$218	$0	49	0	M	Single	39	Other	4	Unemployed	Low
45	Used Car	$0	$109	25	26	M	Single	34	Own	3	Unskilled	Low
46	Small Appliance	$16,935	$189	37	60	M	Single	30	Own	2	Skilled	Low
47	Furniture	$664	$537	31	33	M	Single	48	Own	2	Skilled	High
48	Furniture	$150	$6,520	12	1	F	Divorced	19	Own	1	Skilled	Low
49	Small Appliance	$0	$138	7	119	M	Married	29	Rent	2	Skilled	Low
50	Furniture	$216	$0	19	3	F	Divorced	26	Rent	3	Skilled	High
51	New Car	$0	$660	17	75	M	Single	42	Rent	4	Skilled	High
52	Business	$0	$724	25	8	M	Single	30	Rent	2	Skilled	High

Hands-on Practice

Bank Credit Risk Analyses

The credit risk manager wants to have a more thorough investigation of the customer age profile. Particularly how it compares across different Loan Purpose. Today, we will assume that the 425 records are a random sample of records that have been pulled out for analyses.

i) Develop 95% confidence interval for mean `Age`. Interpret your results.

ii) Develop 95% confidence interval for proportion of `Age`>50. Interpret your results.

iii) Develop 99% prediction intervals for `Age` of a new customer. Interpret your results.

Key Concepts on Hypothesis Testing

statistical inference

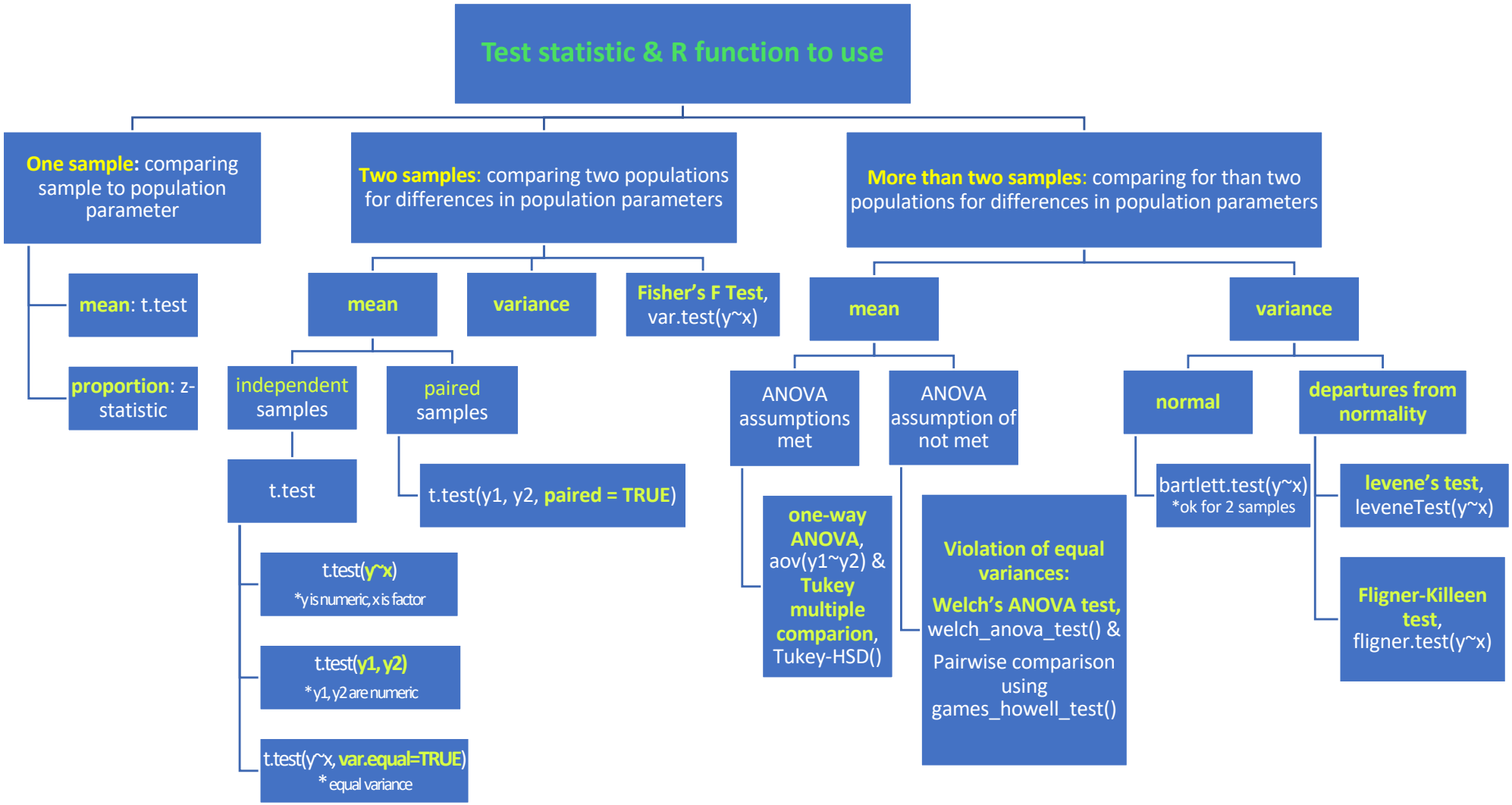
- estimation of population parameters
- hypothesis testing
  - drawing conclusions about value of the parameter of one or more populations based on sample data

Key Concepts on Hypothesis Testing

Hypothesis Testing Procedure

Steps in conducting a hypothesis test:

1. Identify the population parameter and formulate the hypotheses to test.
2. Select a level of significance (the risk of drawing an incorrect conclusion).
3. Determine the decision rule on which to base a conclusion.
4. Collect data and calculate a test statistic.
5. Apply the decision rule and draw a conclusion.


Test statistic & R function to use
One sample: comparing sample to population parameter
mean: t.test
proportion: z-statistic
Two samples: comparing two populations for differences in population parameters
mean
variance
Fisher's F Test, var.test(y~x)
independent samples
paired samples
t.test
t.test(y1, y2, paired = TRUE)
t.test(y~x)
*y is numeric, x is factor
t.test(y1, y2)
*y1, y2 are numeric
t.test(y~x, var.equal=TRUE)
* equal variance
More than two samples: comparing for than two populations for differences in population parameters
mean
variance
ANOVA assumptions met
ANOVA assumption of not met
one-way ANOVA, aov(y1~y2) & Tukey multiple comparion, Tukey-HSD()
Violation of equal variances:
Welch's ANOVA test, welch_anova_test() &
Pairwise comparison using games_howell_test()
normal
departures from normality
bartlett.test(y~x)
*ok for 2 samples
levene's test, leveneTest(y~x)
Fligner-Killeen test, fligner.test(y~x)

Let's continue with our Hands-on Practice

Bank Credit Risk Analyses

The credit risk manager was to have a more thorough investigation of the customer age profile. Particularly how it compares across different Loan Purpose.

The loan service manager makes the following claims:

i) mean age of all their customers is 35

ii) mean age of all their customers is less than 40

iii) proportion of customers with age > 50 is at least 0.18

Test his claims with the data you have.

Let's continue with our Hands-on Practice

Bank Credit Risk Analyses

The credit risk manager wants to drill down to 5 types of Loan Purpose:

Used Car, New Car, Small Appliance, Furniture, Business

The CR manager wants to see if the mean age differs between customers applying to these Loans:

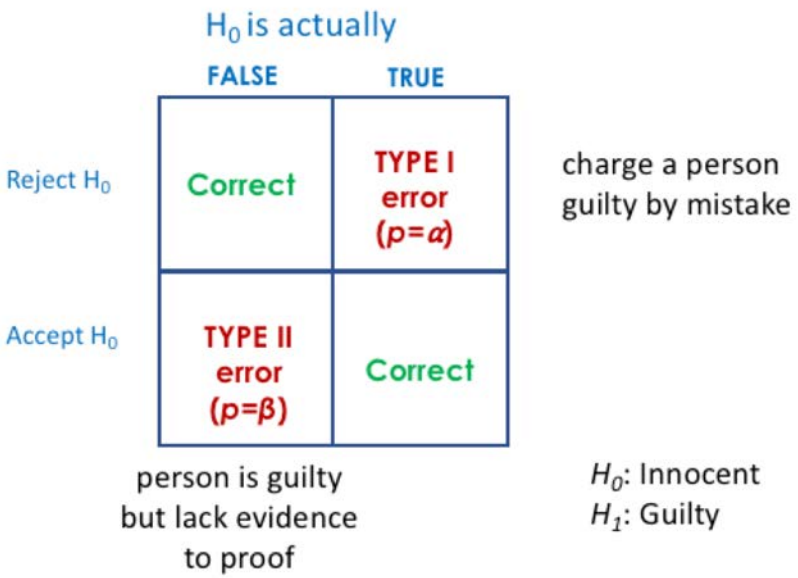
i) Used Car vs New Car

ii) Business vs Small Appliance

iii) Used Car, New Car, Small Appliance, Furniture, Business

Set up the hypotheses and test each of them with your data.

Key Concepts on Hypothesis Testing



bigger α : probability of rejecting H0 is higher, probability of type 2 error decreases (increase power of analyses)

small α : probability of type 2 error increases; therefore need larger sample to increase power of test

power of test: Probability of not committing type II error (1- β)

References:

- test of variances: http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r
- One-way ANOVA in r - http://www.sthda.com/english/wiki/one-way-anova-test-in-r
- Welch's ANOVA: https://rdrr.io/cran/rstatix/man/welch_anova_test.html
- Homogeneity of variance: https://www.datanovia.com/en/lessons/homogeneity-of-variance-test-in-r/

THE END!

Thank You for Your Attention!