Preface: The Road to Statistical Inference

In a previous chapter

Type of Research Question
Make an estimate about the population
Test a claim about the population
Compare two sub-populations
Investigate a relationship between two variables in the population

Questions about the population → Requires complete population data to answer

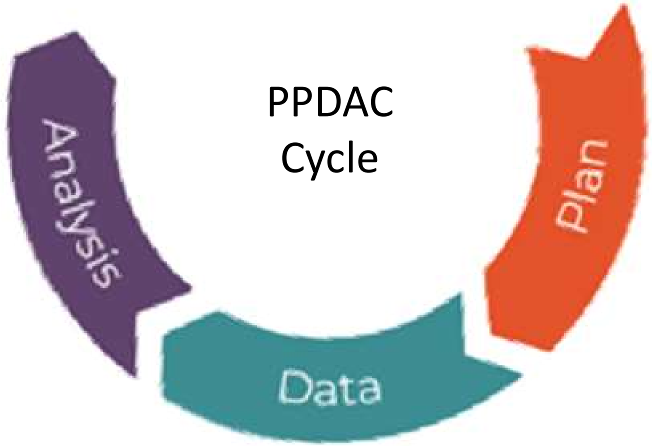
Complete population data difficult to obtain

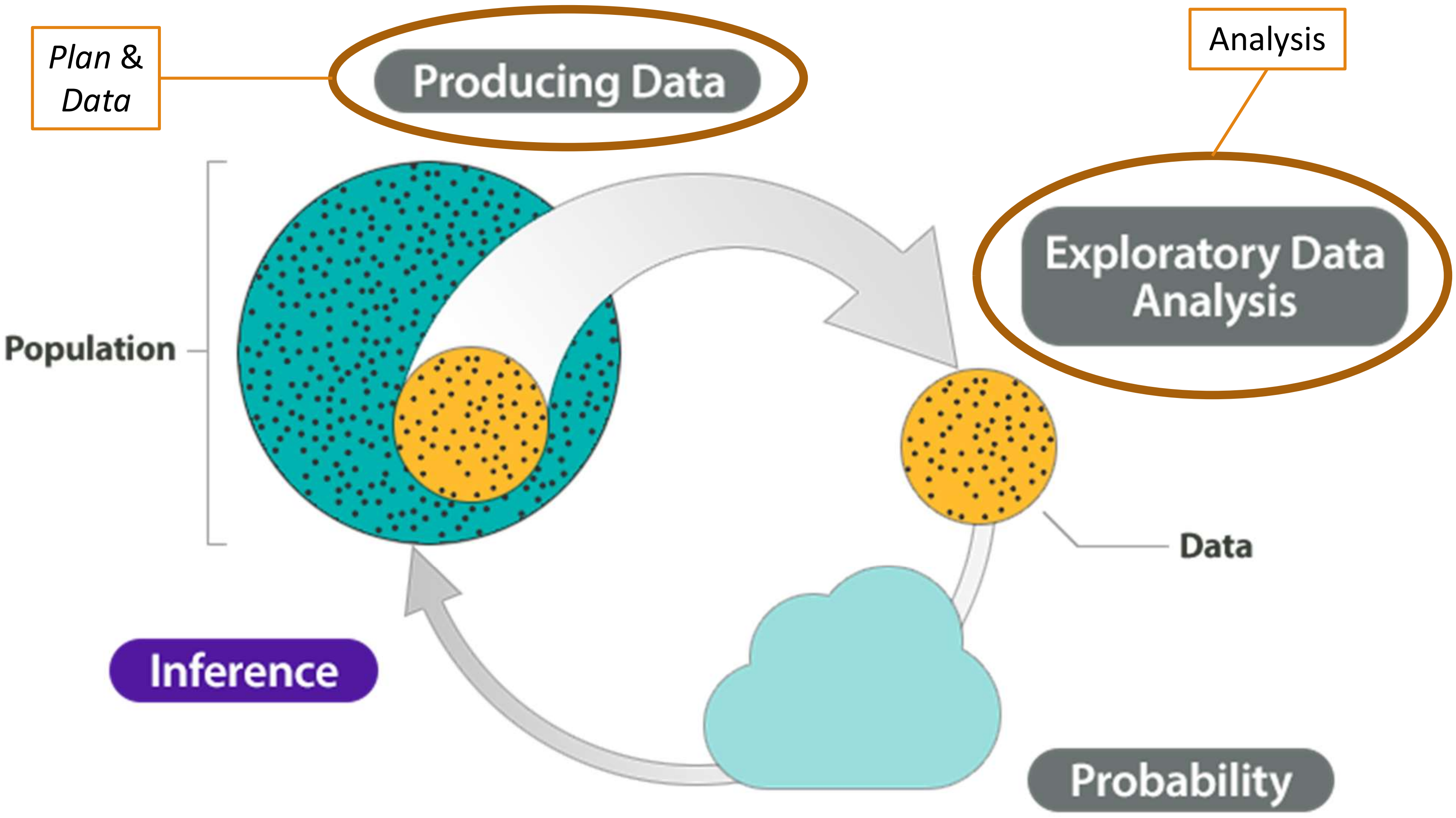
↓

Approximate from a sample of the population

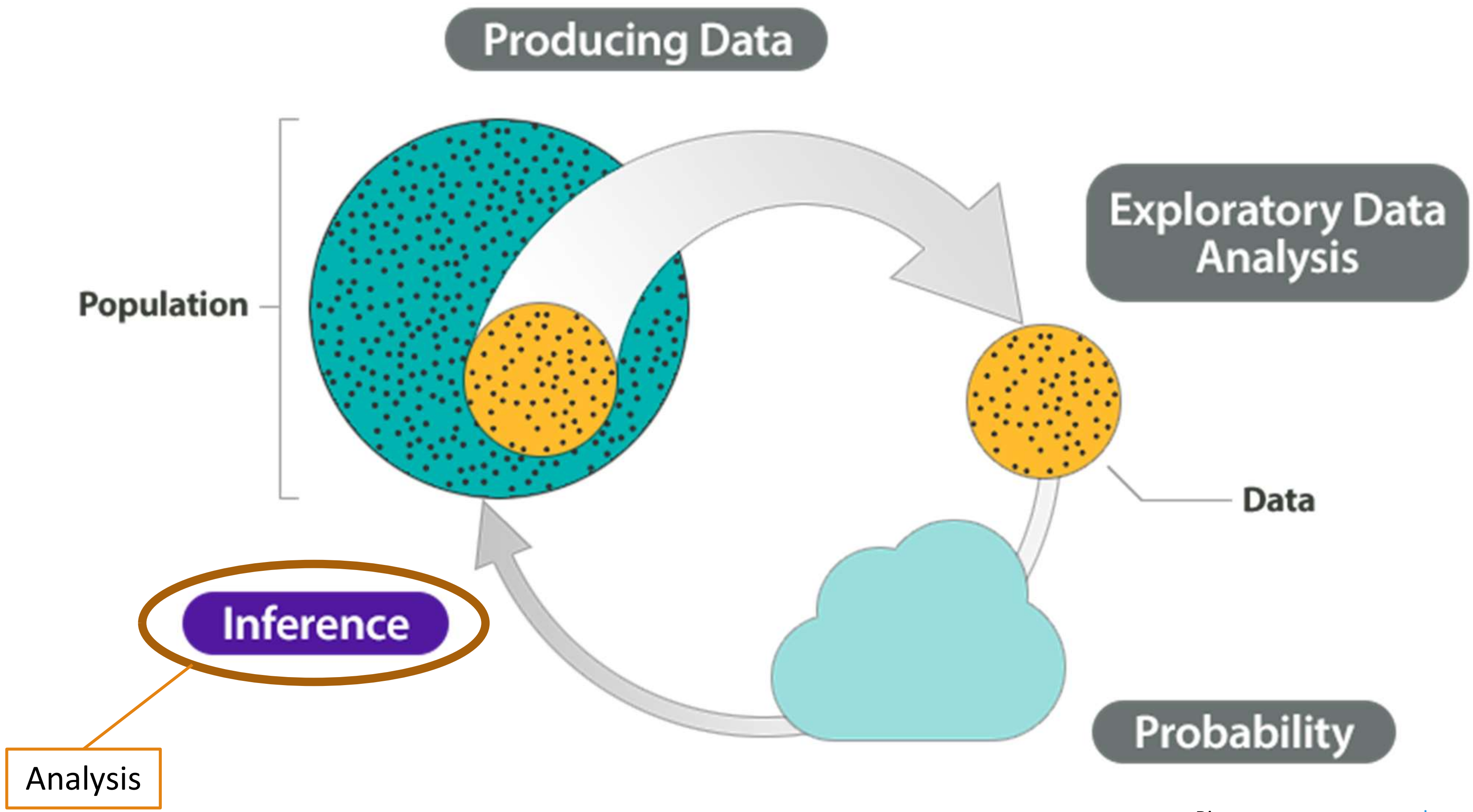
↓

Statistical inference

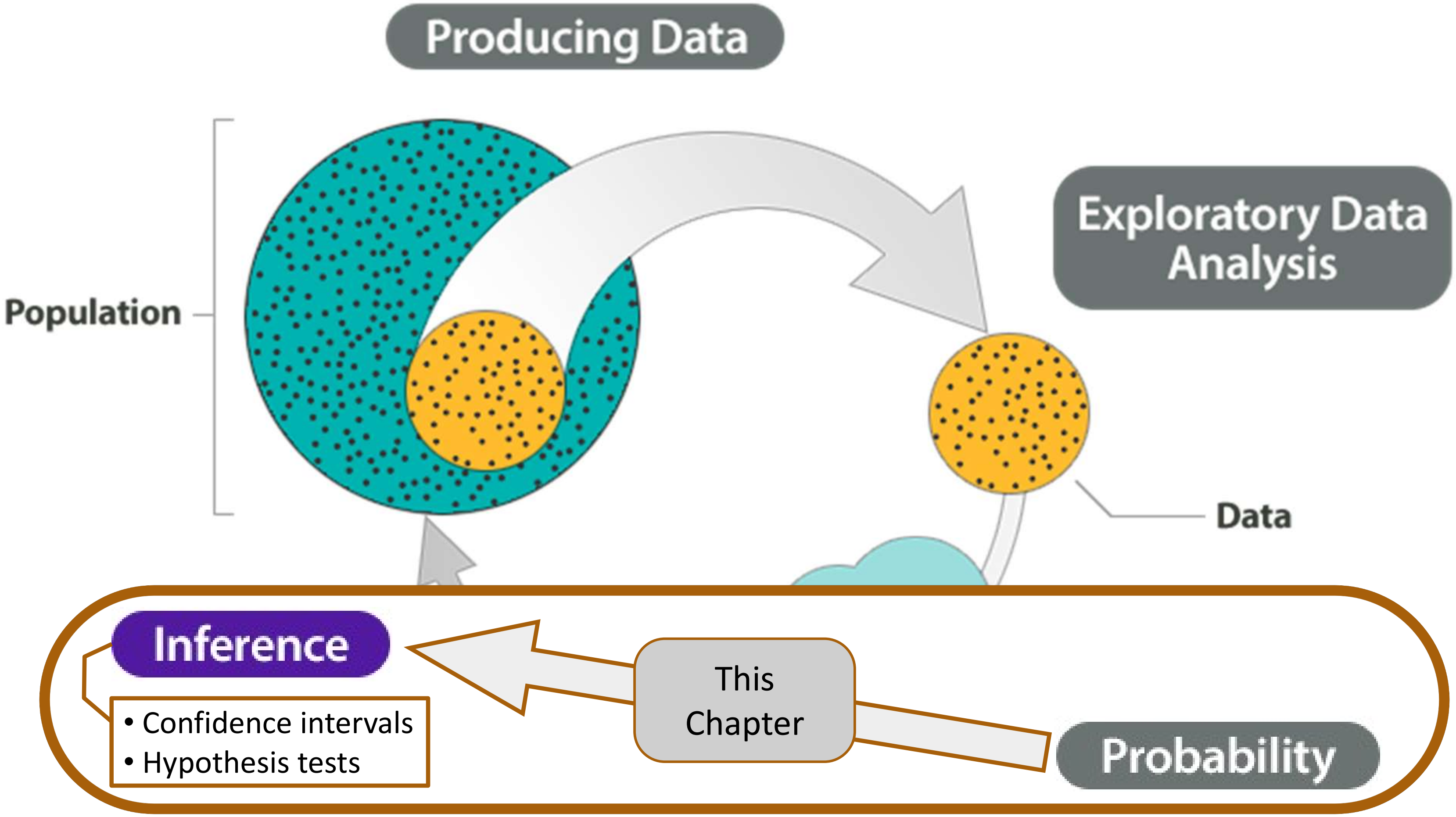

PPDAC
Cycle
Analysis
Data
Plan



Picture source: courses.lumenlearning.com

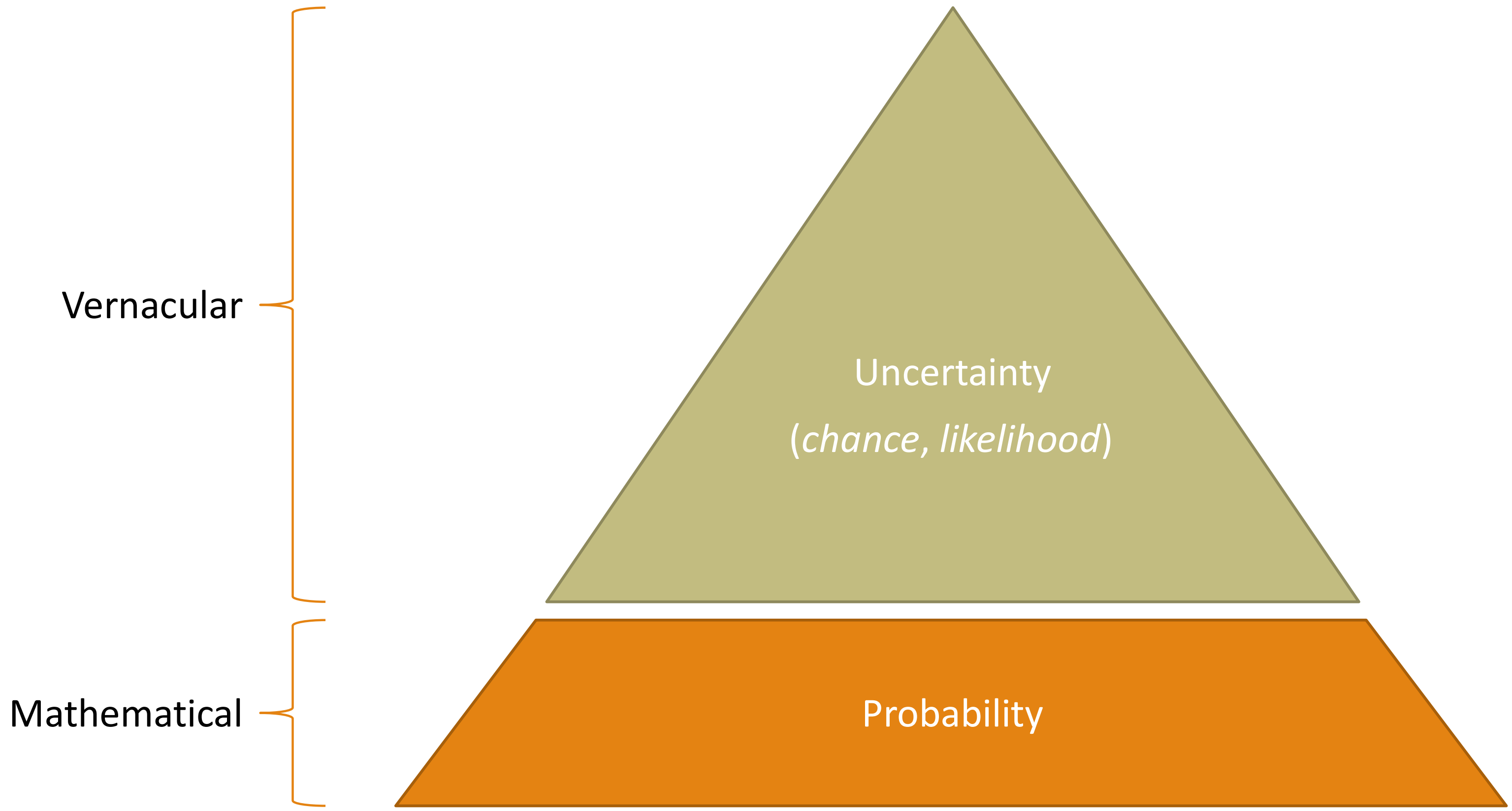


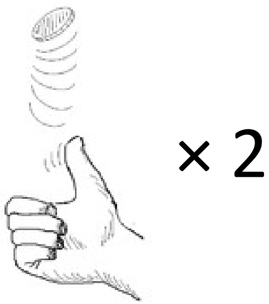
Picture source: courses.lumenlearning.com



Picture source: courses.lumenlearning.com

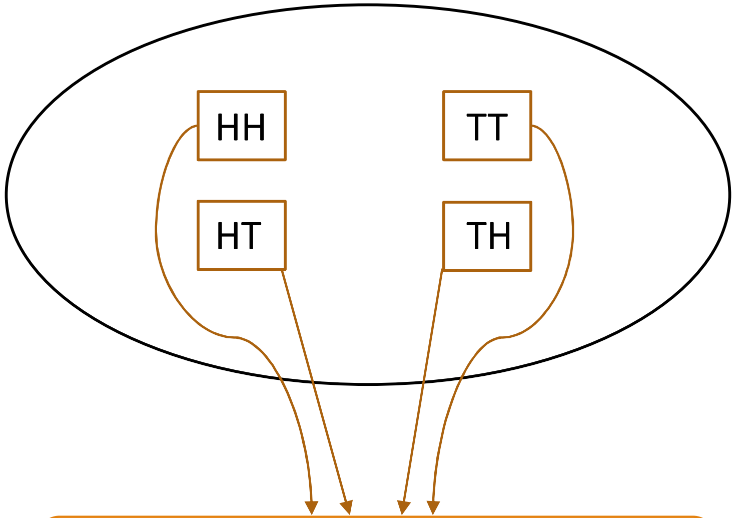
Probability:
Setting the Stage


Vernacular
Mathematical
Uncertainty
(*chance, likelihood*)
Probability



A *probability experiment*

Possibilities



Outcomes of the experiment

A probability experiment must

- Be repeatable (“as many times as you want”)
- Give rise to a precise set of outcomes

Probability experiments form the bedrock of probability

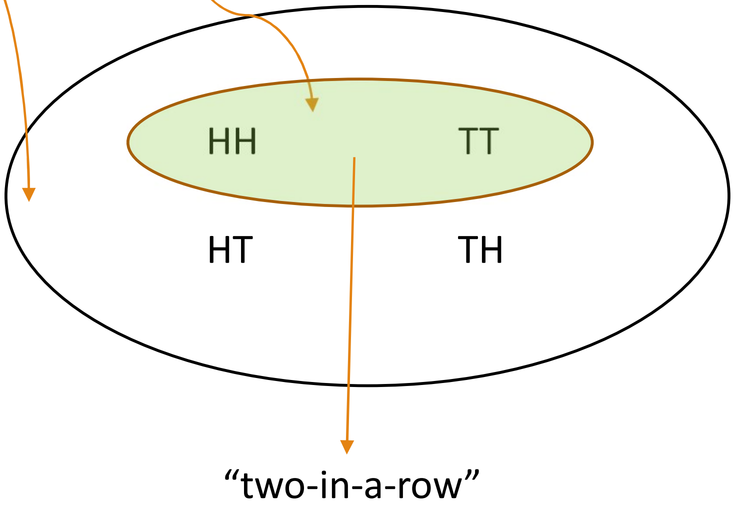
Sample space : collection of all outcomes of a probability experiment

Event : subcollection of the sample space

Context

Probability of an <u>event</u> of a <u>sample space</u>

- ➢ How likely the outcome of the probability experiment is an element of the event



Note: in practice, all outcomes are events.
But not all events are outcome!

Example: Die-rolling

Rolling a six-sided die

Sample space

1 2
3 4
5 6

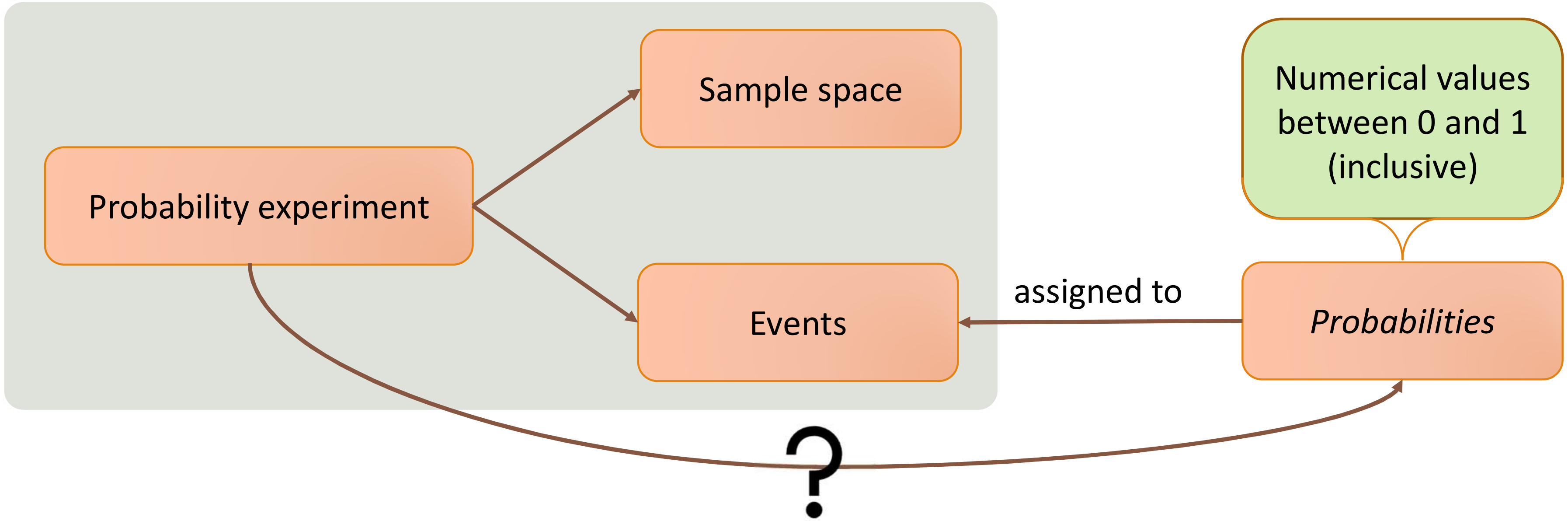
{1, 2, 3, 4, 5, 6}

A possible event

1 2
3 4
5 6

“an even-numbered face” = {2, 4, 6}

Probabilities



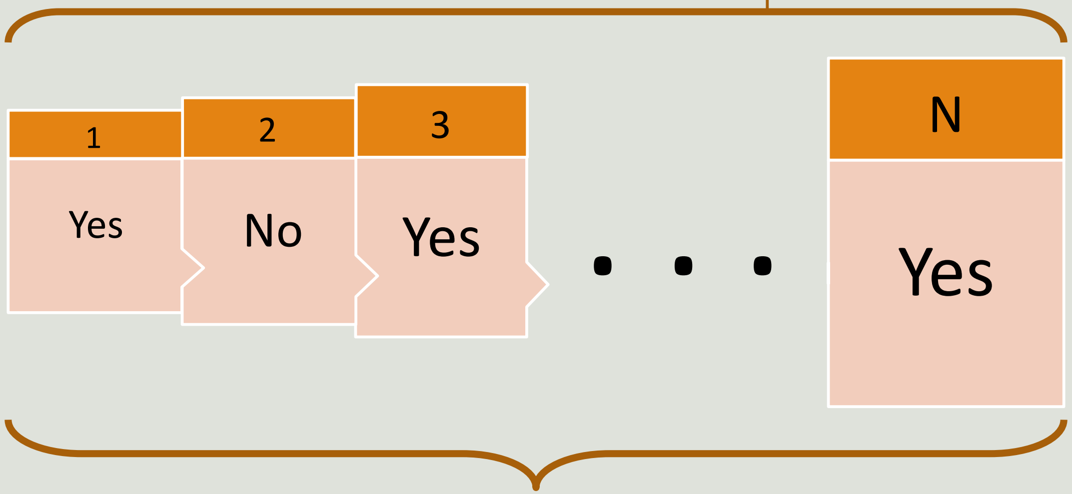
If *E* is an event that has been assigned a probability

- *P(E)*, read “the probability of *E*”, stands for the probability assigned to *E*

Simple Cases: Finite Sample Spaces

For any event E:

1. Repeat the probability experiment a large number (N) of times
2. For each repetition, check if the outcome is in E



Proportion of E = $\frac{Count\ of\ 'Yes'}{N}$ → P(E)

Every event can be assigned a probability

- Proportion of *E* estimates the true *P(E)*
- Estimate gets more accurate as *N* increases

Simple Cases: Finite Sample Spaces
An Example

Event E = “an even-numbered face”

1. Repeat the probability experiment a large number (N) of times
2. For each repetition, check if the outcome is in E

1	2	3	. . .	N
Yes	No	No		Yes

Proportion of E = $\frac{Count\ of\ 'Yes'}{N} \rightarrow$ P(E)

Every event can be assigned a probability

Rolling a six-sided die

Rules of Probabilities

For every assignment of probabilities

1. $0 \leq P(E) \leq 1$ for each event *E*
2. *P(S) = 1* if *S* is the entire sample space
3. If *E* and *F* are non-overlapping (mutually exclusive) events, then $P(E \cup F) = P(E) + P(F)$

For finite sample spaces, it is enough to assign probabilities to outcomes so that they add up to 1

For finite sample spaces, it is enough to assign probabilities to outcomes so that they add up to 1.

$$\left.\begin{array}{l} P(1) = 0.1 \\ P(2) = 0.1 \\ P(3) = 0.1 \\ P(4) = 0.1 \\ P(5) = 0.1 \\ P(6) = 0.5 \end{array}\right\} \xrightarrow{\text{add up to}} 1$$

Deriving probabilities of other events

- Repeated application of the 3rd rule of probability
- Probability of each event is the sum of probabilities of its outcomes

Rolling a biased six-sided die

For finite sample spaces, it is enough to assign probabilities to outcomes so that they add up to 1.

$$\left.\begin{aligned} P(1) &= 0.1\\ P(2) &= 0.1\\ P(3) &= 0.1\\ P(4) &= 0.1\\ P(5) &= 0.1\\ P(6) &= 0.5 \end{aligned}\right\} \xrightarrow{\text{add up to}} 1$$

Rolling a biased six-sided die

Deriving probabilities of other events

E.g. Let *E* denote the event “an odd-numbered face”.
Let *F* denote the event “an even-numbered face”.

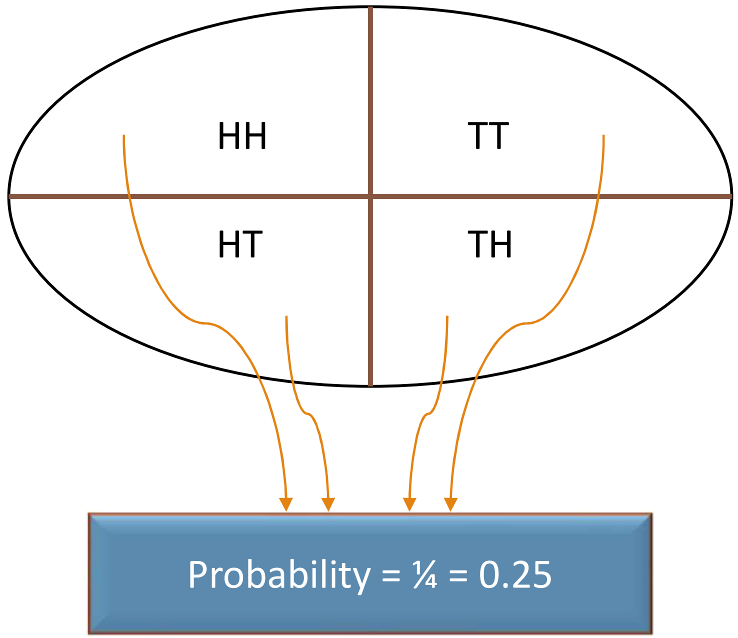
P(E)	P(F)
= P(1) + P(3) + P(5)	= P(2) + P(4) + P(6)
= 0.1 + 0.1 + 0.1	= 0.1 + 0.1 + 0.5
= 0.3	= 0.7

Uniform Probabilities and Rates

Uniform probability

Every outcome has the same probability

$$= \frac{1}{size\ of\ sample\ space}$$



Uniform Probabilities and Rates

Uniform probability

Every outcome has the same probability

$$= \frac{1}{size\ of\ sample\ space}$$

- Probability of selecting each unit

$$= \frac{1}{size\ of\ sampling\ frame}$$

- For any subgroup (event) A, P(A) = probability of selected unit being in A = rate(A)

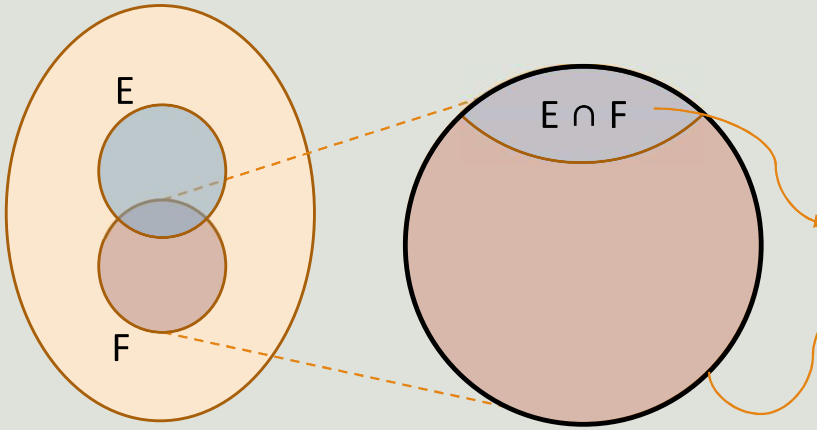
Conditional Probabilities

Conditional Probability

E and *F* are events

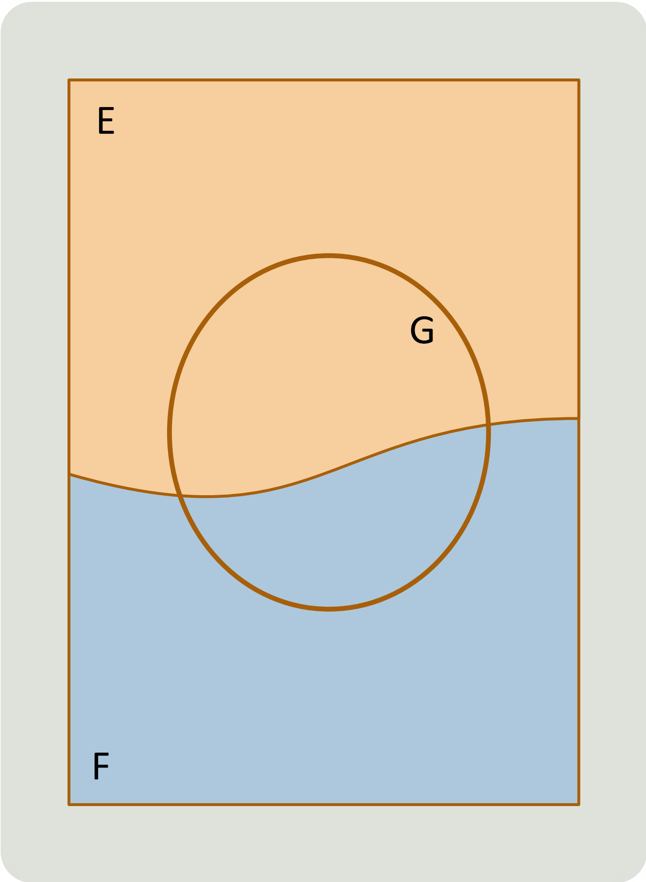
$\underbrace{\boxed{P(E \mid F)} : \text{“probability of } E \text{ given } F\text{”}}$

How likely the outcome is in E, if we know it is in F



$$\frac{P(E \cap F)}{P(F)} = P(E|F)$$

By convention, if P(F) = 0, then P(E | F) = 0 for any E.



If

1. E and F are mutually exclusive

2. $E \cup F$ is the entire sample space

Then for any event G

$$P(G) = P(G \cap E) + P(G \cap F)$$

$$\Updownarrow$$

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F)$$

Law of Total Probability

If

1. E and F are mutually exclusive

2. $E \cup F$ is the entire sample space

Then for any event G

$$P(G) = P(G \cap E) + P(G \cap F)$$

$$\Updownarrow$$

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F)$$

Conditional Probabilities as Rates

Recall:

- Sample space = sampling frame
- P(A) = rate(A), for any subgroup (event) A

A and *B* are subgroups

Does P(A | B) equal rate(A | B)?

Conditional Probabilities as Rates

Recall:

- Sample space = sampling frame
- P(A) = rate(A), for any subgroup (event) A

A and *B* are subgroups

Does P(A | B) equal rate(A | B)?

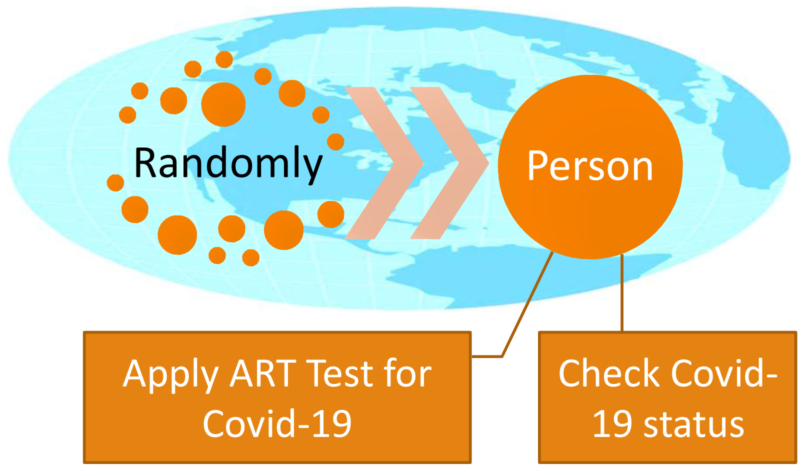
Yes!

$$\boxed{P(A \mid B)} = \frac{P(A \cap B)}{P(B)} = \frac{rate(A \cap B)}{rate(B)} = \frac{\frac{\text{size of } A \cap B}{\text{size of sampling frame}}}{\frac{\text{size of } B}{\text{size of sampling frame}}} = \frac{\text{size of } A \cap B}{\text{size of } B} = \boxed{rate(A \mid B)}$$

Example: ART for Covid-19

Antigen Rapid Test

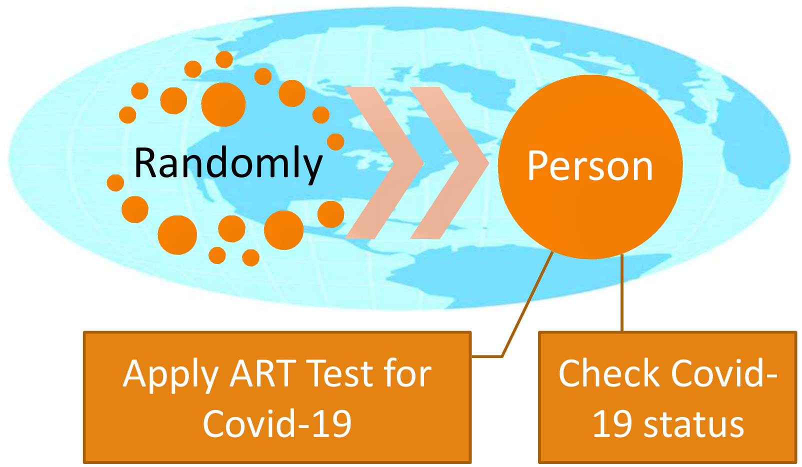
- *Sensitivity* (True positive rate) : P(+ | Covid-19) = 0.80
- *Specificity* (True negative rate) : P(- | no Covid-19) = 0.99



Example: ART for Covid-19

Antigen Rapid Test

- *Sensitivity* (True positive rate) : P(+ | Covid-19) = 0.80
- *Specificity* (True negative rate) : P(- | no Covid-19) = 0.99



Base rate of Covid-19 : P(Covid-19) = 0.01

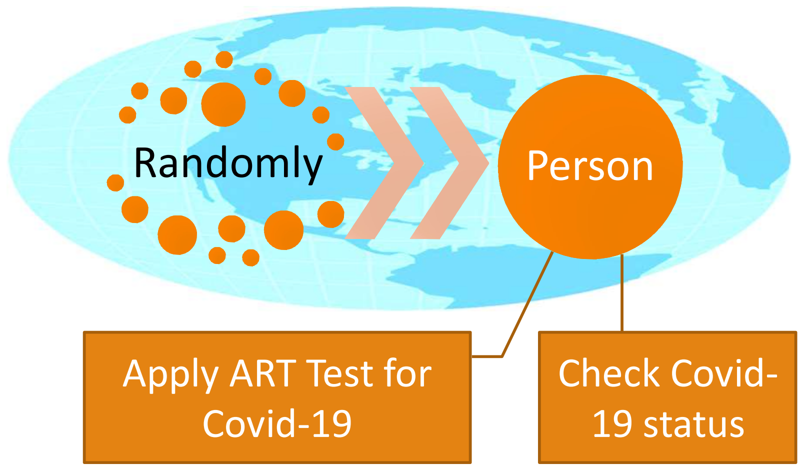
? What is P(Covid-19 | +)?

Example: ART for Covid-19

Antigen Rapid Test

Sensitivity (True positive rate) : P(+ | Covid-19) = 0.80

Specificity (True negative rate) : P(- | no Covid-19) = 0.99



	+	-	Row total
Covid-19	800	200	1000
No Covid-19	990	98010	99000
Column total	1790	98210	100000

Base rate of Covid-19 : P(Covid-19) = 0.01

Rate(Covid-19 | +) = $\frac{800}{1790}$ = $\boxed{0.447 \text{ (3 s.f.)}}$

? What is P(Covid-19 | +)?

Independence

Definition 1

Independence of events A, B : P(A) = $\boxed{P(A \mid B)}$

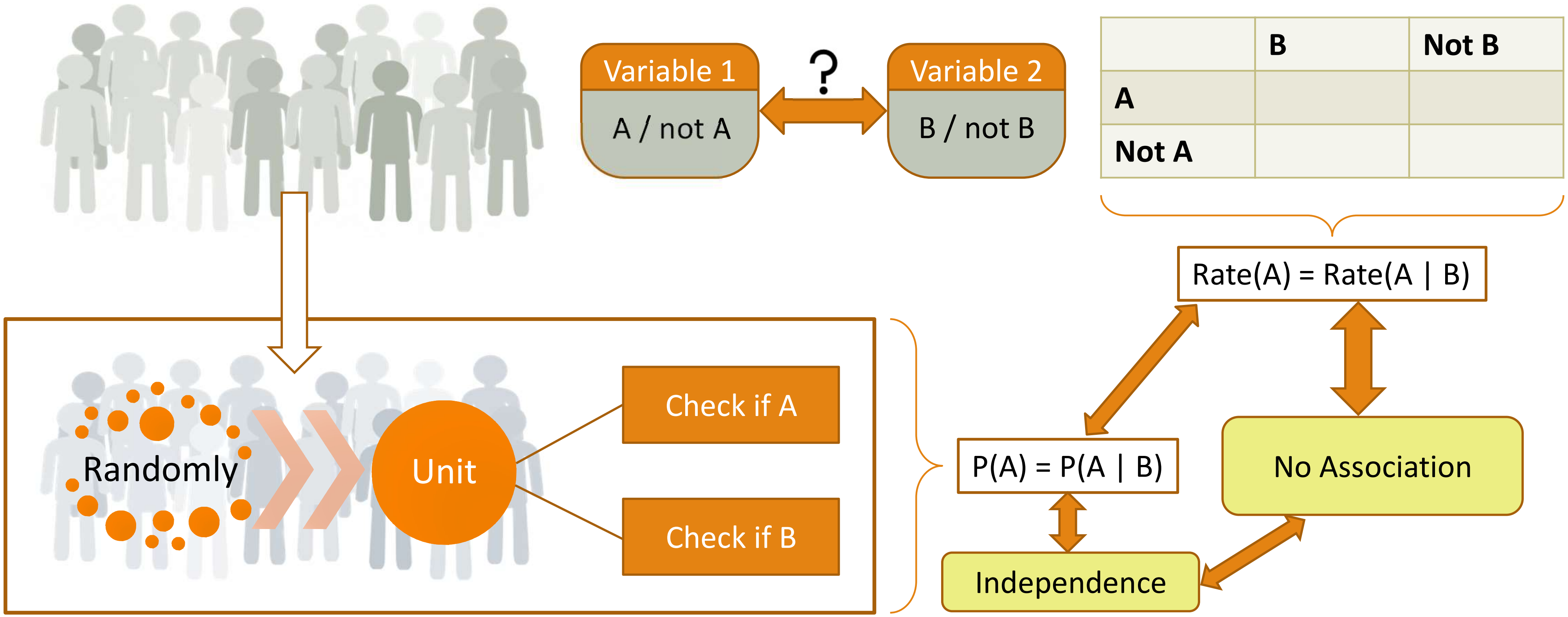
$$P(A) = \boxed{\frac{P(A \cap B)}{P(B)}}$$

Definition 2

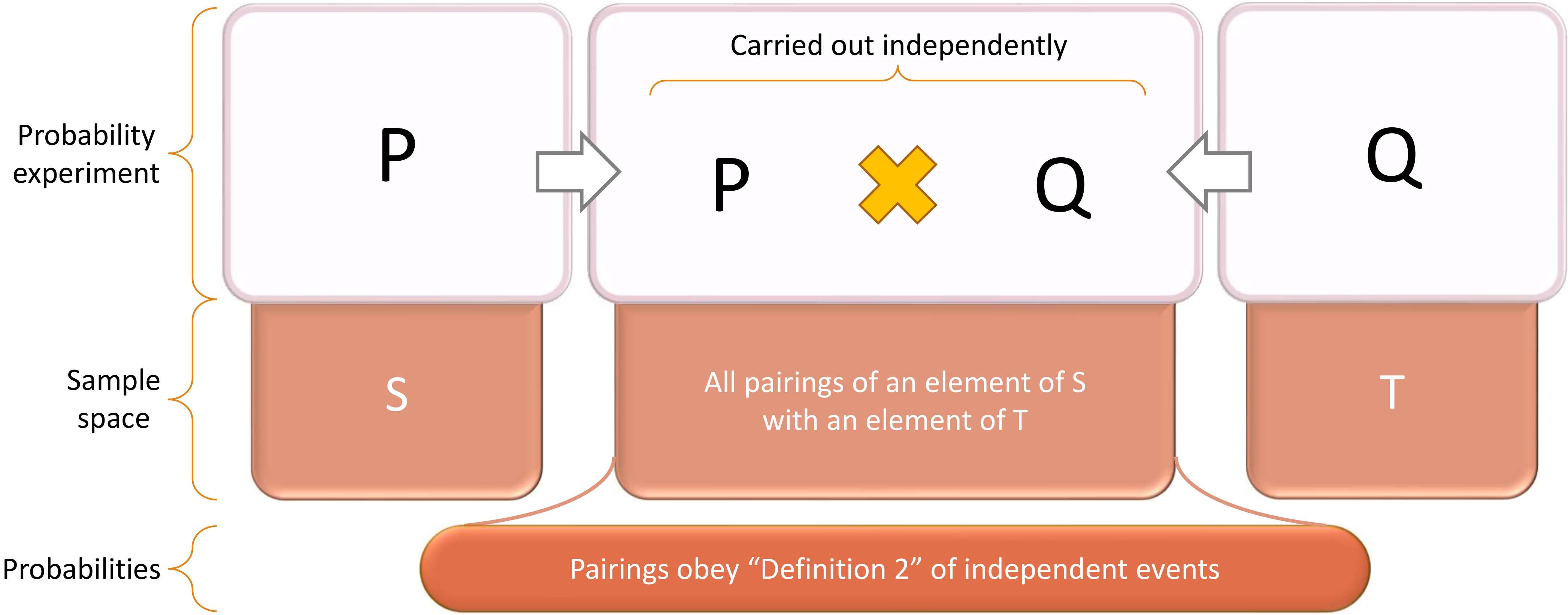
Independence of events A, B : $P(A) \times P(B) = P(A \cap B)$

Independence as Non-Association

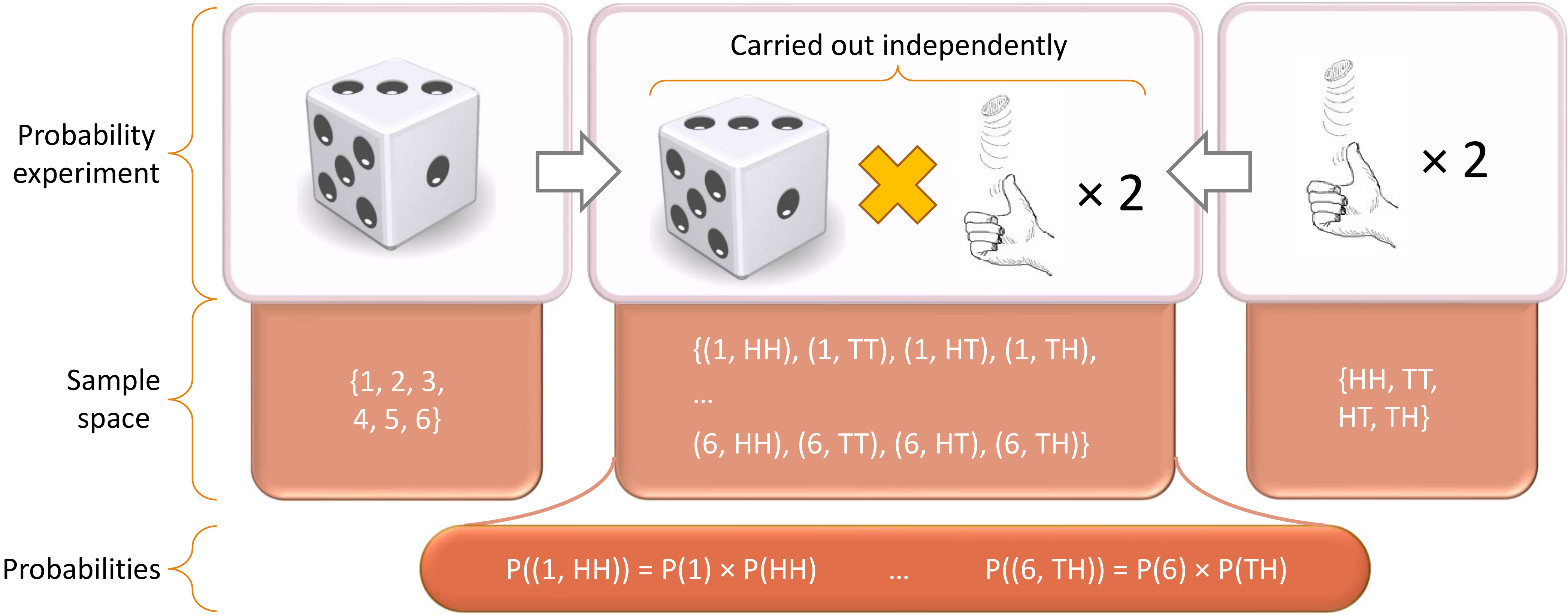
	B	Not B
A		
Not A		



Independent Probability Experiments



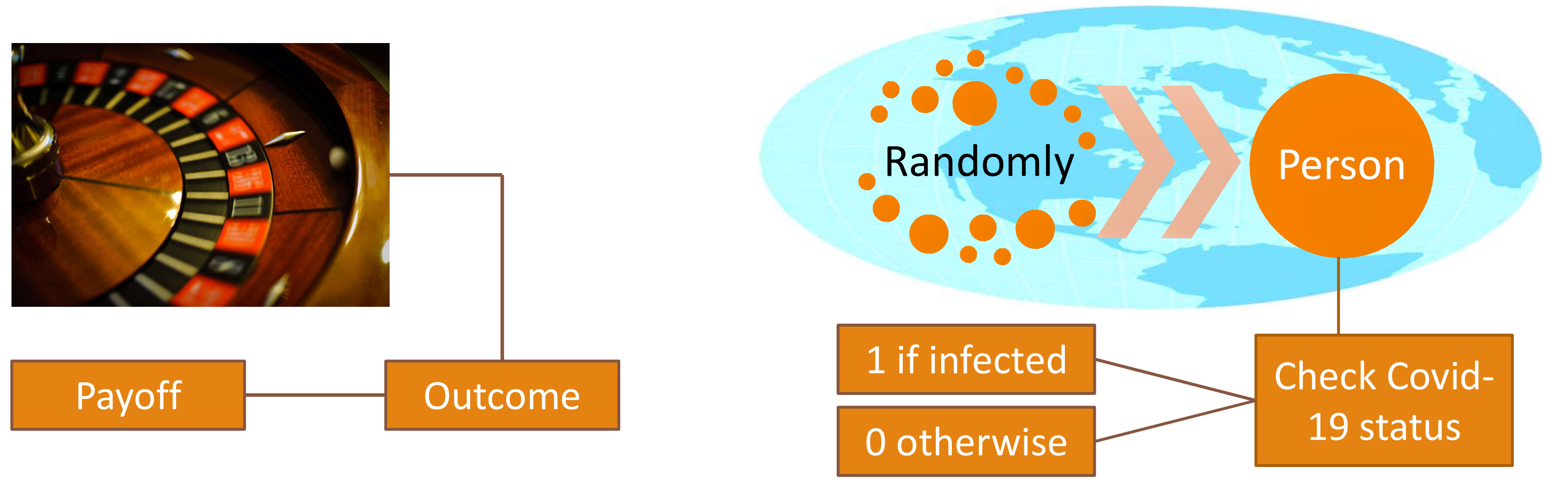
Independent Probability Experiments

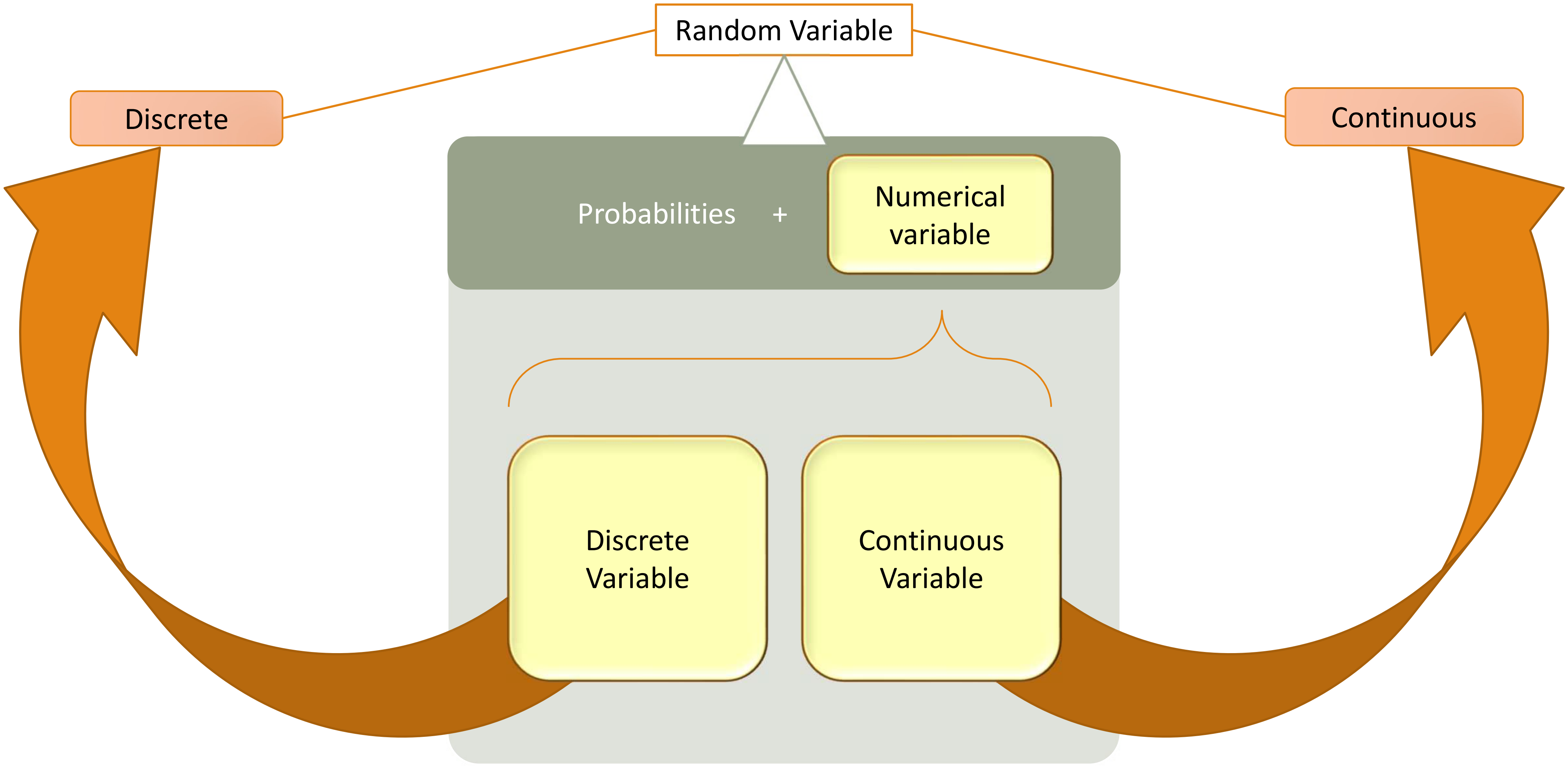


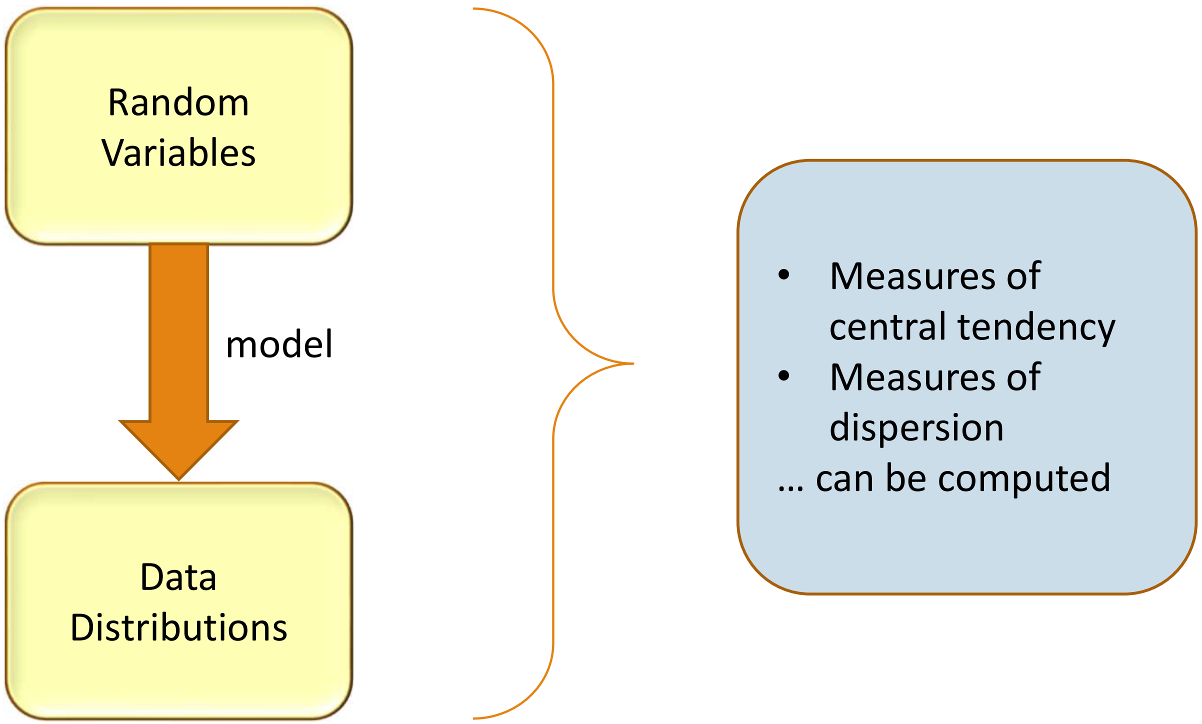
Random Variables

Probability experiment → Outcome → Numerical value

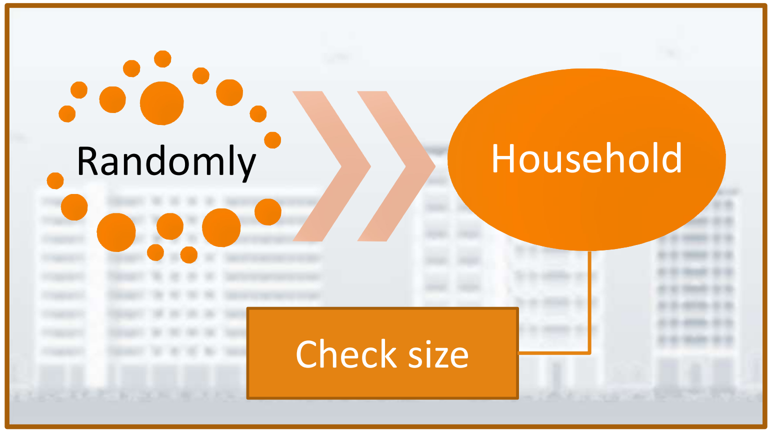
Examples




Random Variable
Discrete
Continuous
Probabilities
+
Numerical variable
Discrete Variable
Continuous Variable


Random Variables
model
Data Distributions
• Measures of central tendency
• Measures of dispersion
… can be computed

HDB Household Size



Finite discrete variable

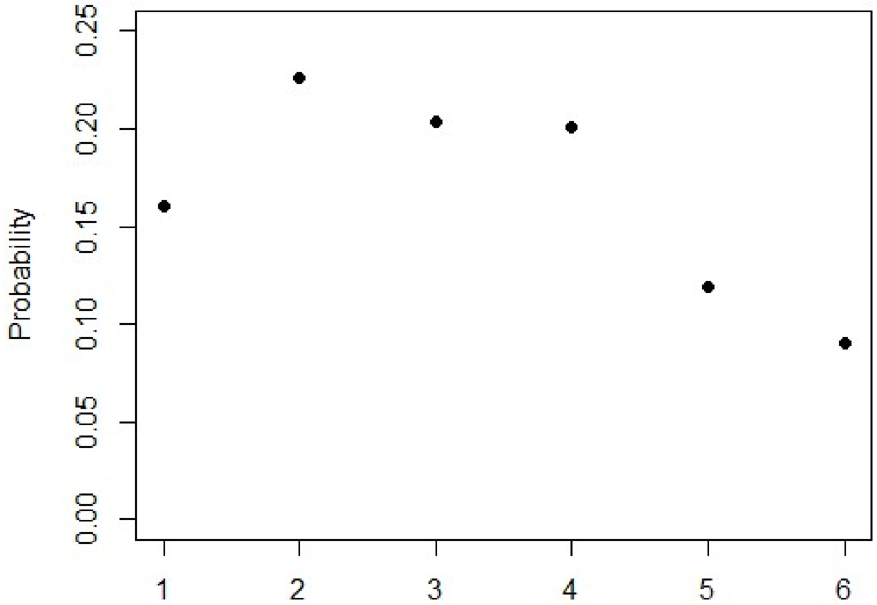
Household size	1	2	3	4	5	6
Probability	0.16	0.226	0.204	0.201	0.119	0.09

Discrete random variable

Visualisation of a Discrete Random Variable

X

Household size	1	2	3	4	5	6
Probability	0.16	0.226	0.204	0.201	0.119	0.09

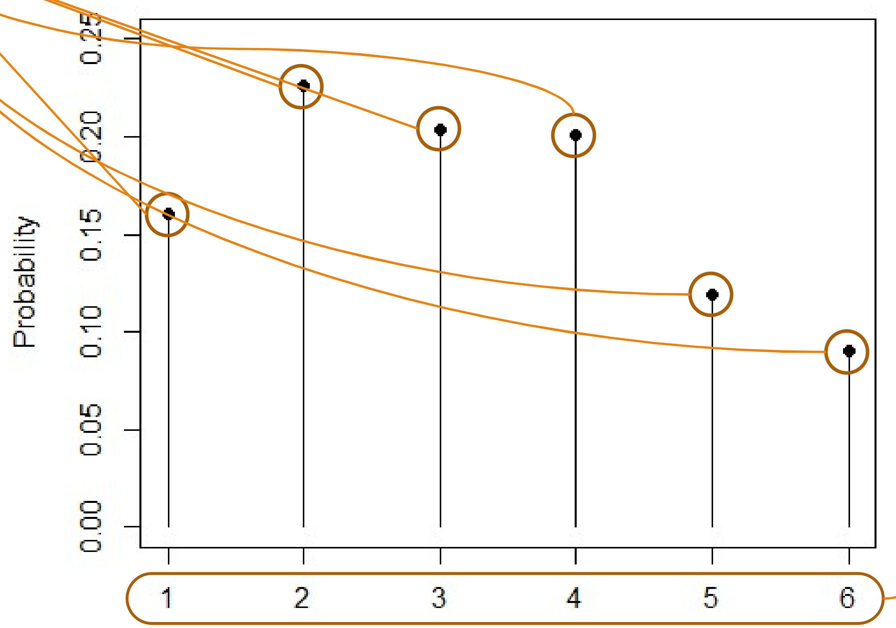


- Each point represents a possible value of X, indicated by its x-value
- y-value of a point = probability that X assumes its x-value

Visualisation of a Discrete Random Variable

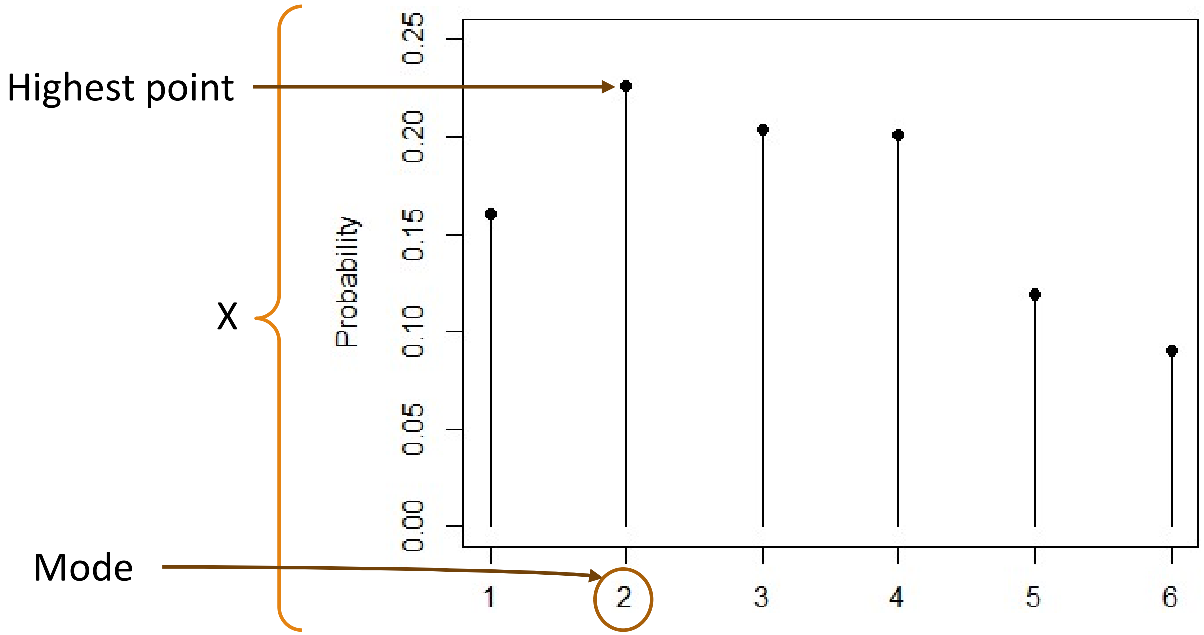
X

Household size	1	2	3	4	5	6
Probability	0.16	0.226	0.204	0.201	0.119	0.09



Discrete collection of possible values of X

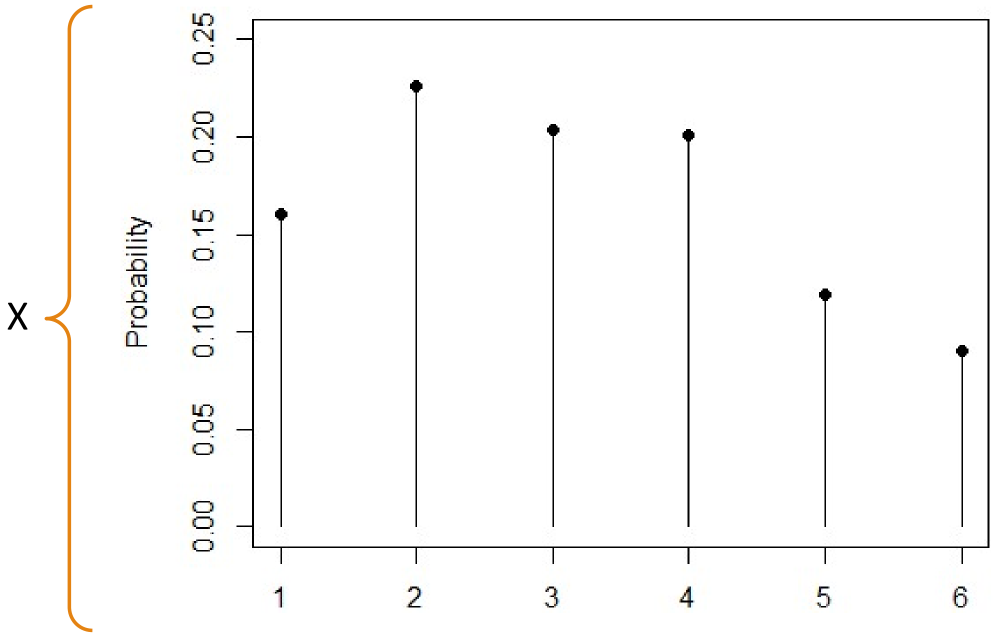
Visualisation of a Discrete Random Variable



Notes:

- Probabilities of the points add up to 1
- x-value of a highest point is a mode

Probabilities of a Discrete Random Variable

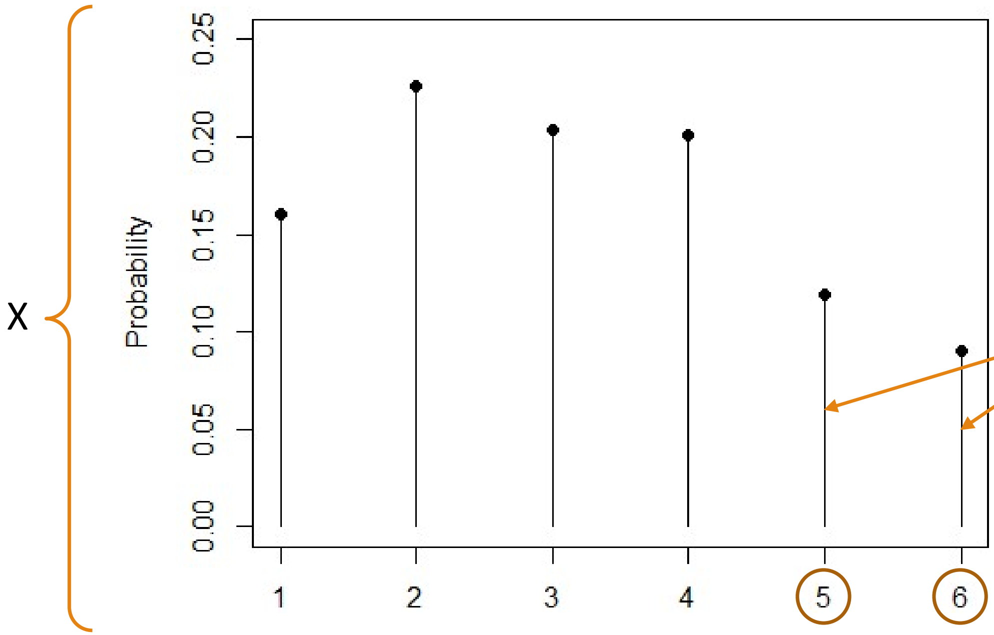


Probability that a randomly selected HDB household has size ≥ 5?

I.e.

$P(X \geq 5)$?

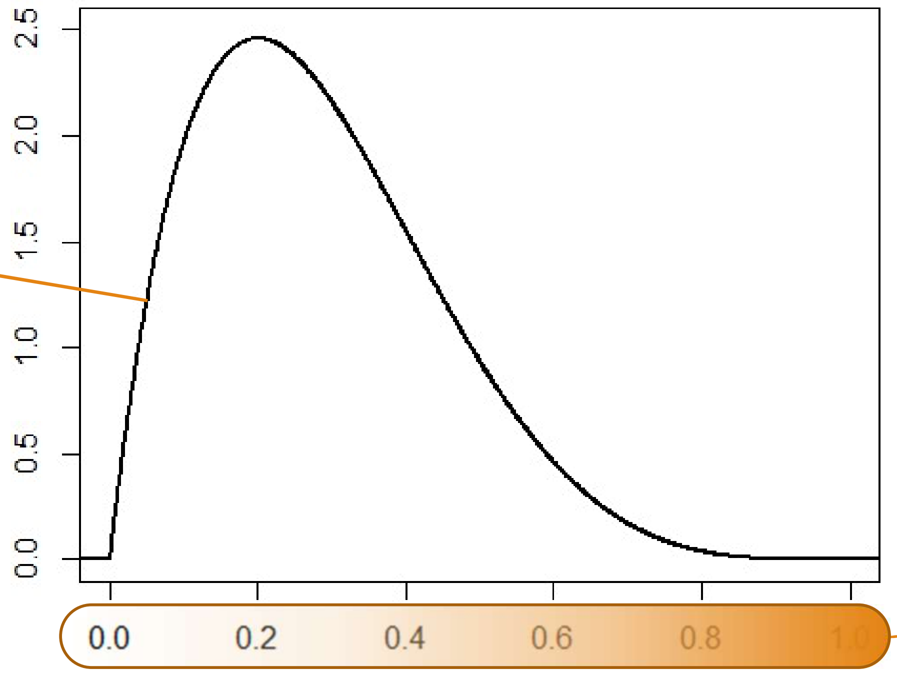
Probabilities of a Discrete Random Variable



P(X ≥ 5)
= P(5) + P(6)
= 0.119 + 0.09
= 0.209

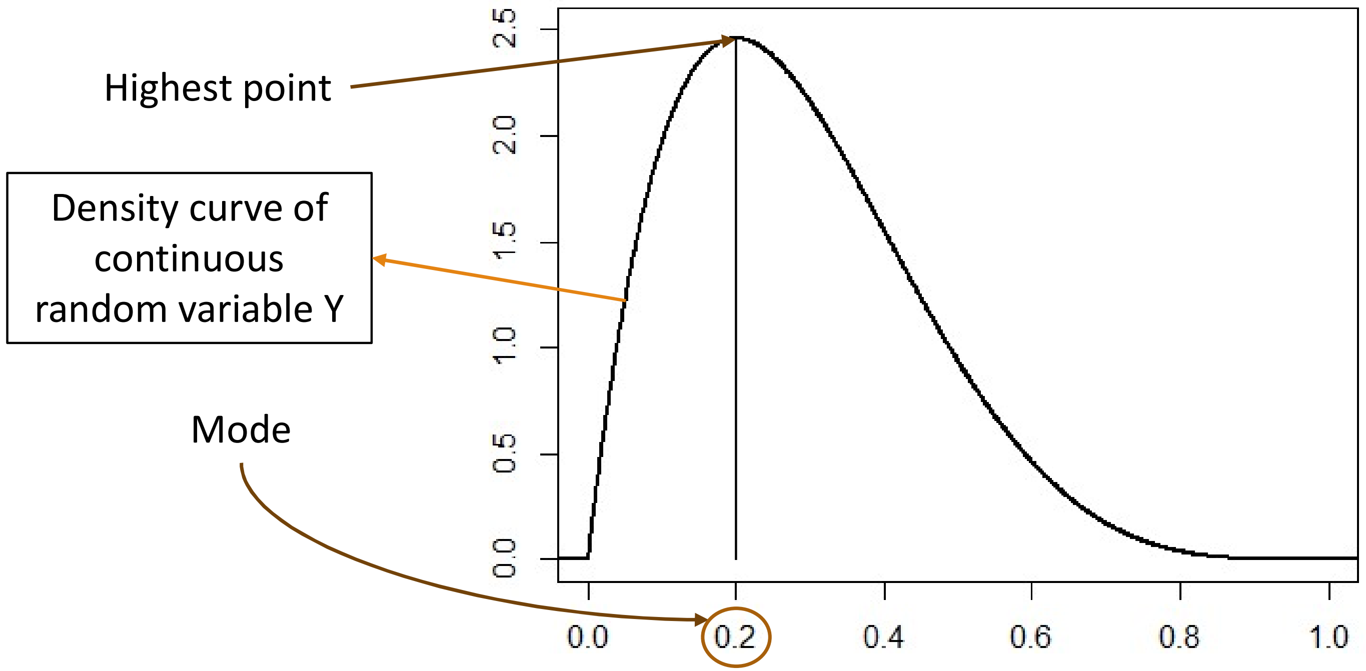
Visualisation of a Continuous Random Variable

Density curve of continuous random variable Y



Continuous range of possible values Y can assume

Visualisation of a Continuous Random Variable

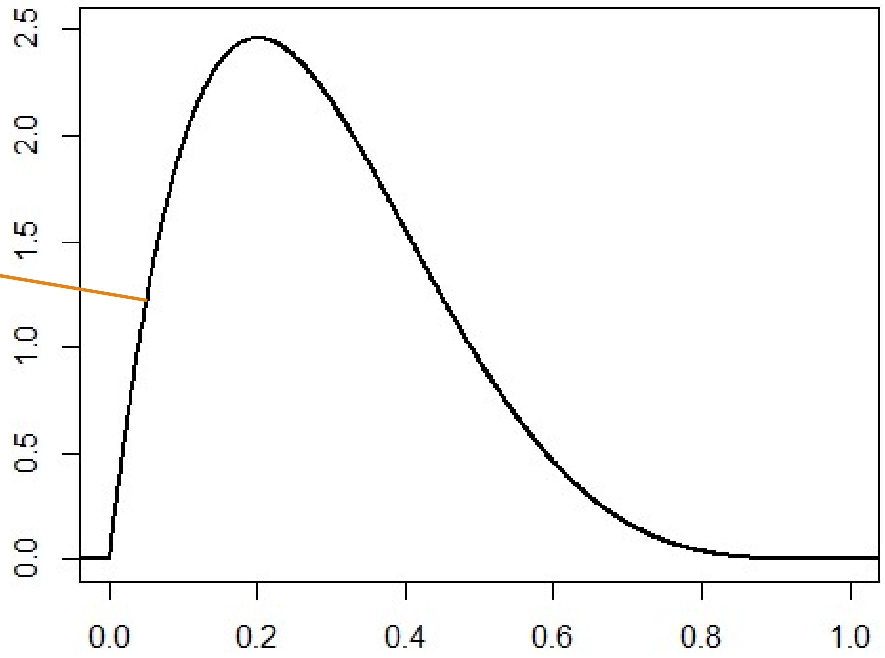


Notes:

- Area under the curve = 1
- x-value of a highest point is a mode

Probabilities of a Continuous Random Variable

Density curve of continuous random variable Y

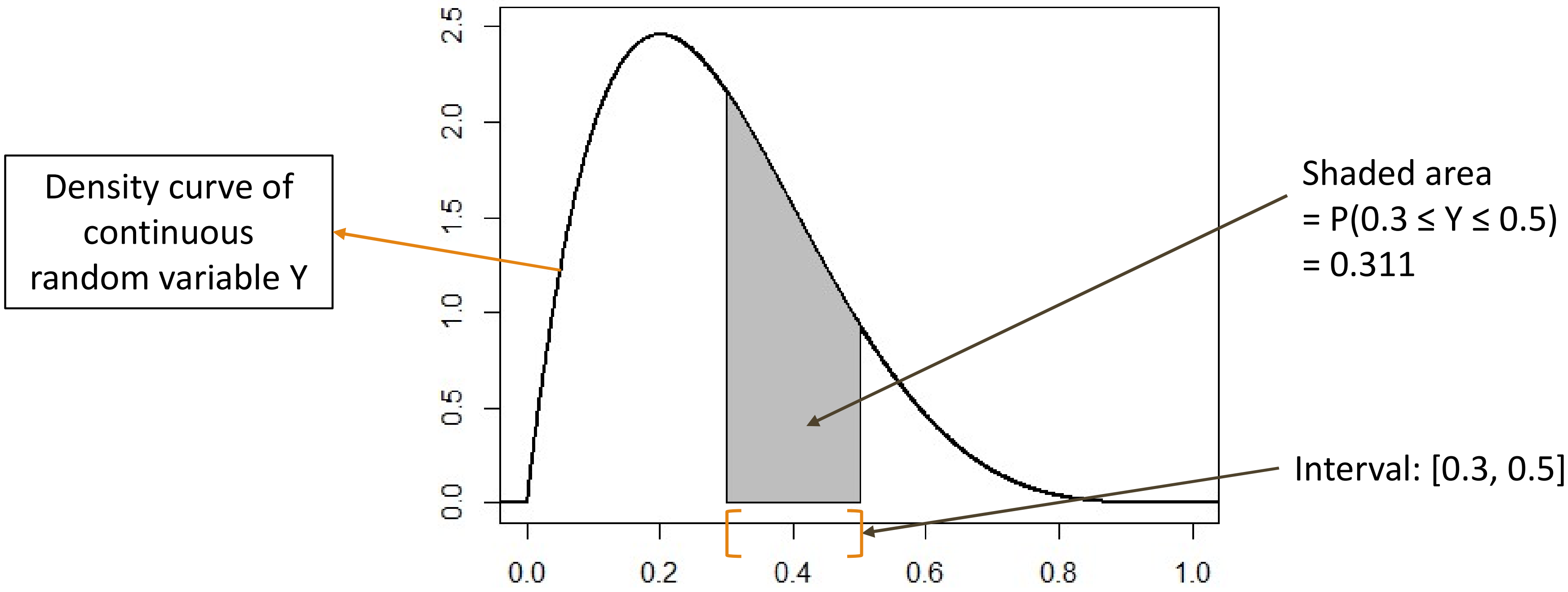


Probability that Y assumes a value between 0.3 and 0.5?

I.e.

$P(0.3 \leq Y \leq 0.5)$?

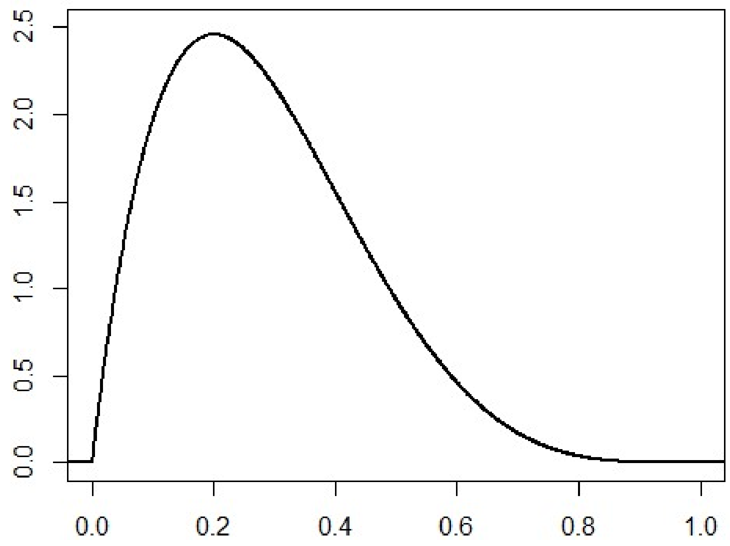
Probabilities of a Continuous Random Variable



Probabilities of a Continuous Random Variable

In general

Probability that a continuous random variable takes on a value in interval $[a, b]$
= area under its density curve from *a* to *b*



Normal Distributions

Two normal distributions can only differ by their means or their variances.

$N(x, y)$: the normal distribution with mean *x* and variance *y*

Common properties:

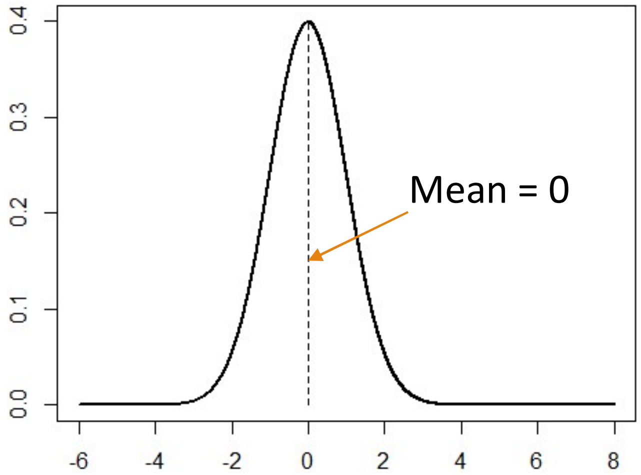
- Bell-shaped curve
- Peak of the curve occurs at the mean
- Curve is symmetrical about the mean

Mean = mode = median

Two normal distributions can only differ by their means or their variances.

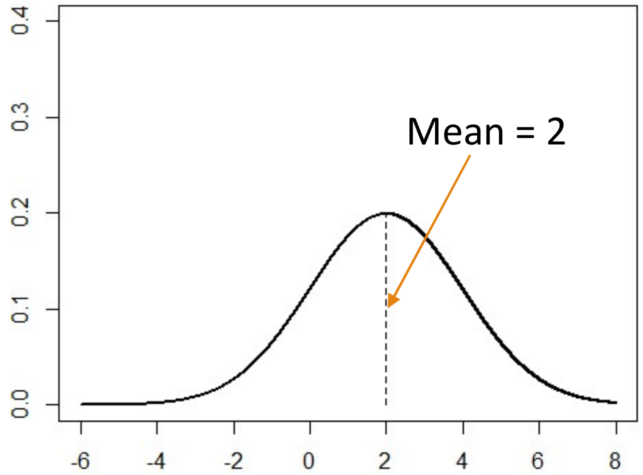
$N(x,y)$: the normal distribution with mean *x* and variance *y*

$N(0,1)$: *standard normal*



- Smaller variance → thinner bell shape

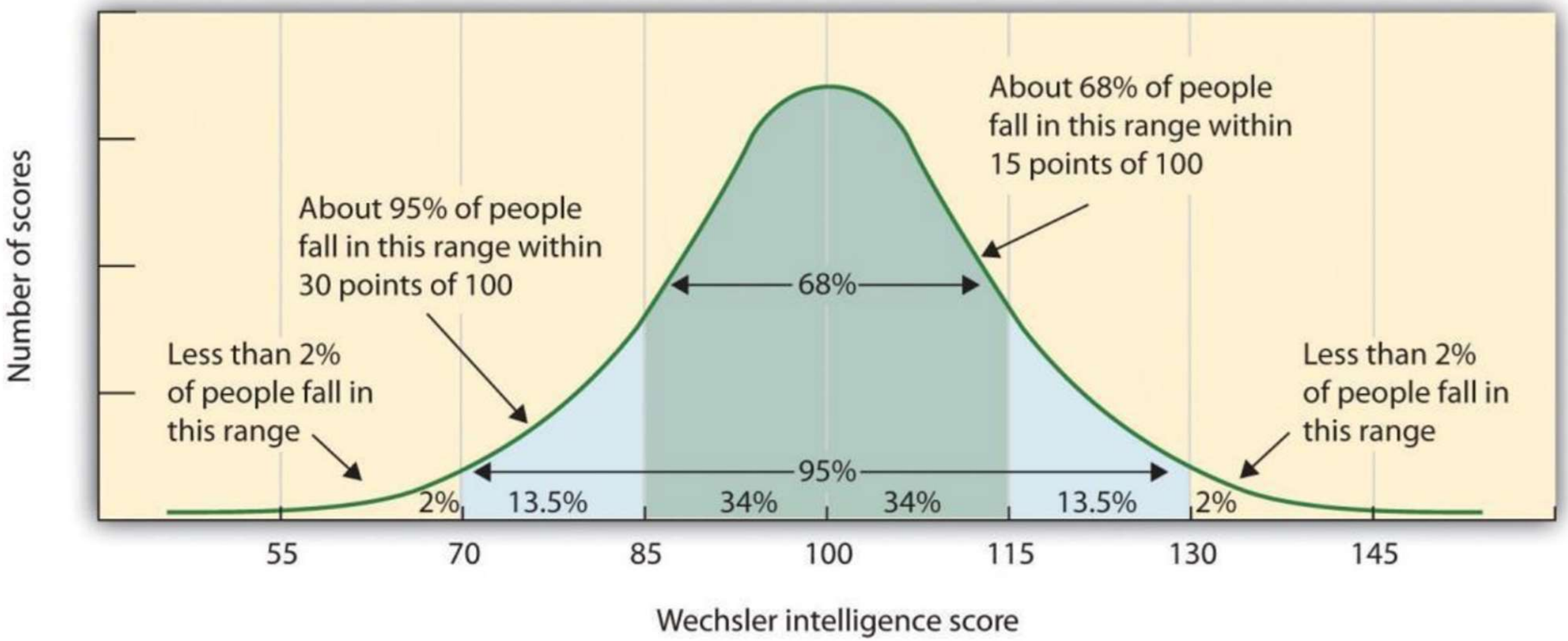
$N(2,2)$



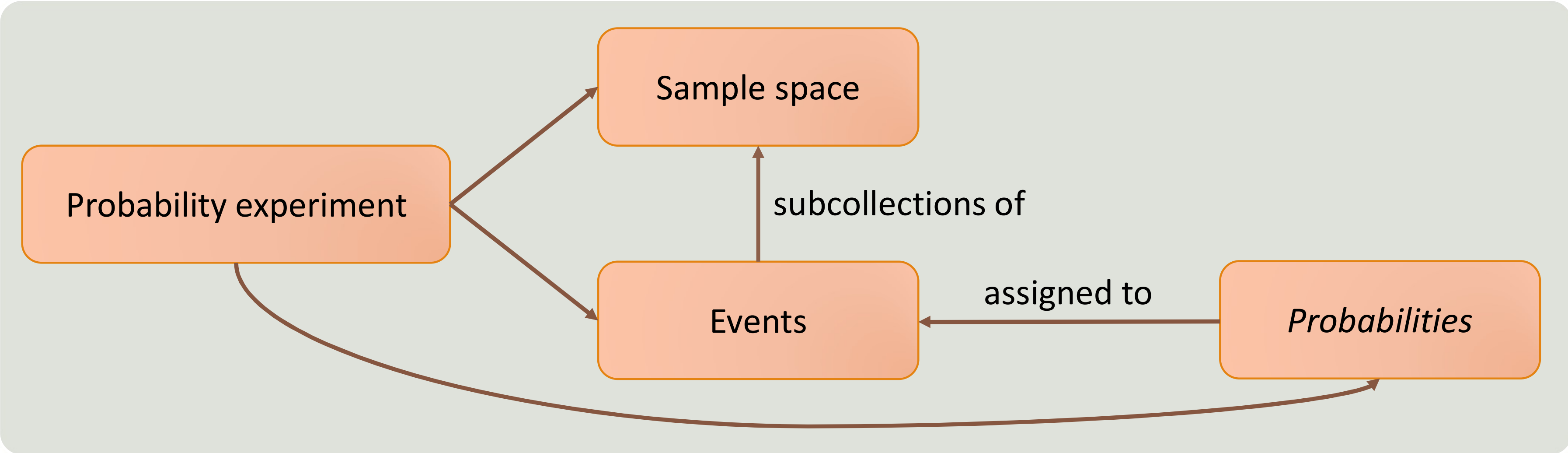
- Greater variance → fatter bell shape

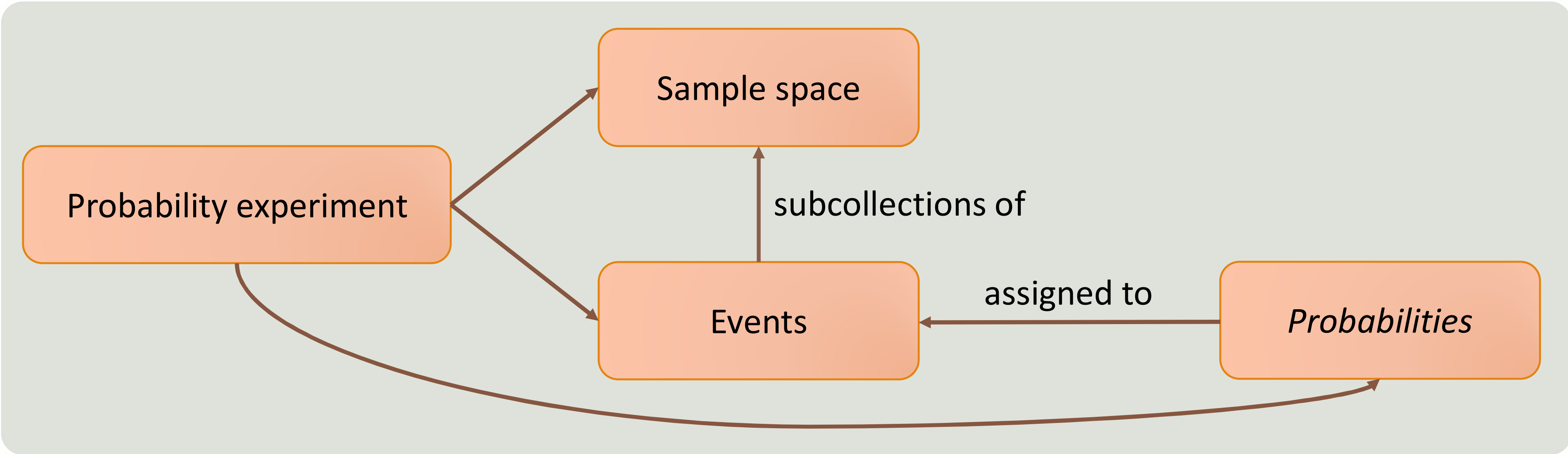
Area under curve kept constant → a fatter curve compensates by being shorter

Normally Distributed: IQ



Summary

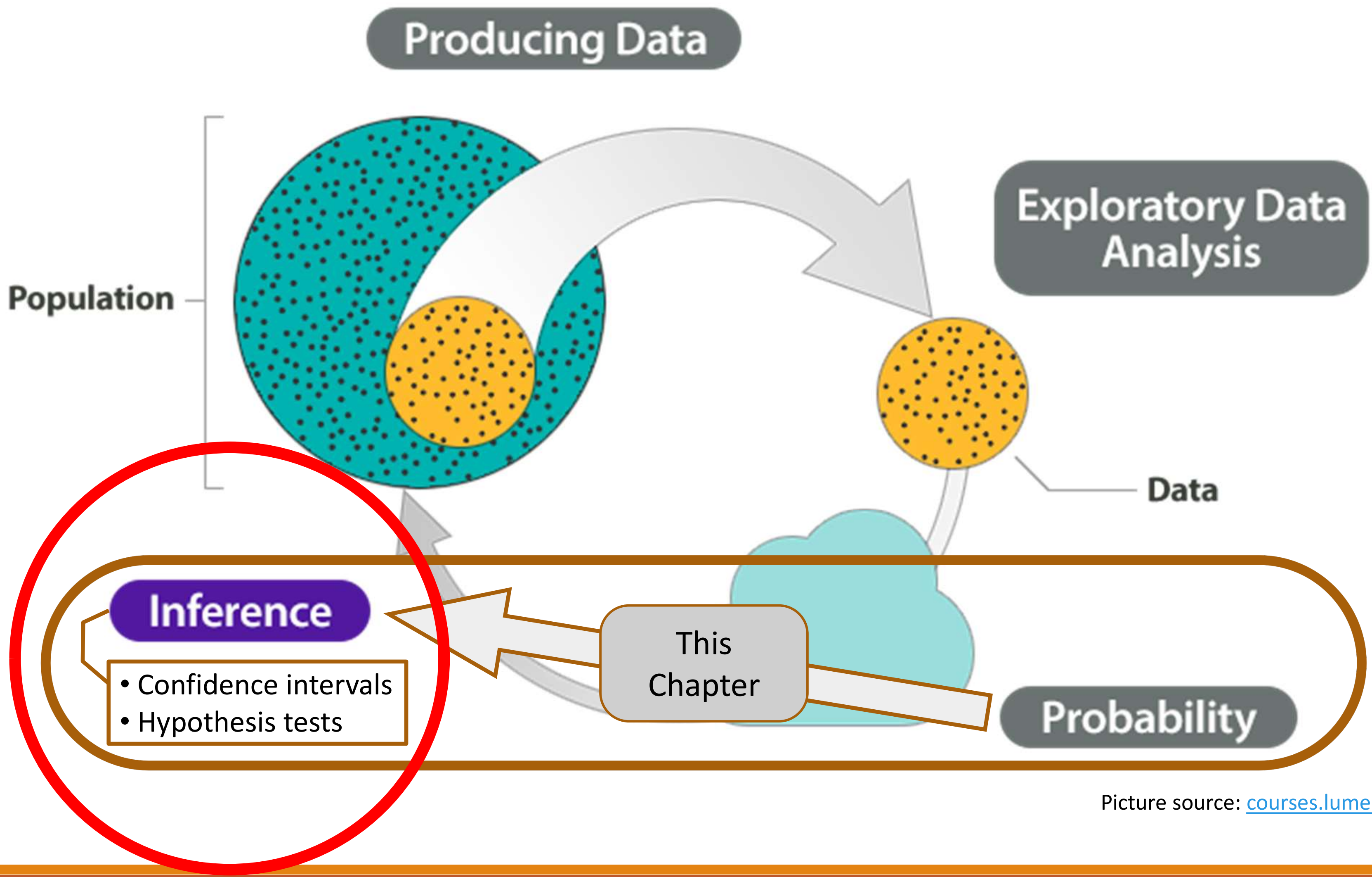

Sample space
Probability experiment
subcollections of
Events
assigned to
Probabilities



- Conditional probability
  - Independence
    - Independent events
    - Independent probability experiments

- Random variables
  - Discrete random variables
  - Continuous random variables
    - Normal distributions

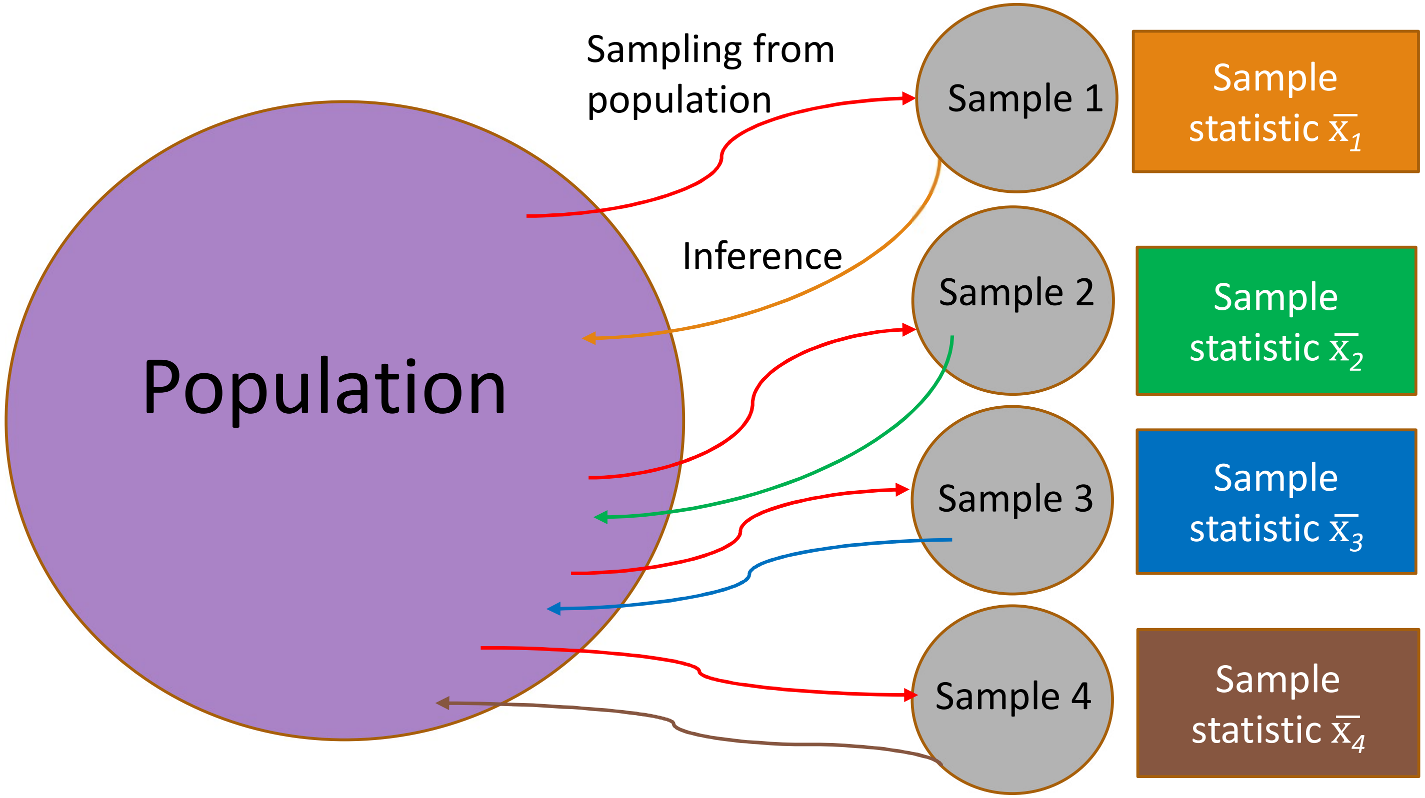
Introduction to Statistical Inference



Picture source: courses.lumenlearning.com

Sample statistic = population parameter + bias + random error

Sample statistic = population parameter + random error



Sample statistic = population parameter + random error

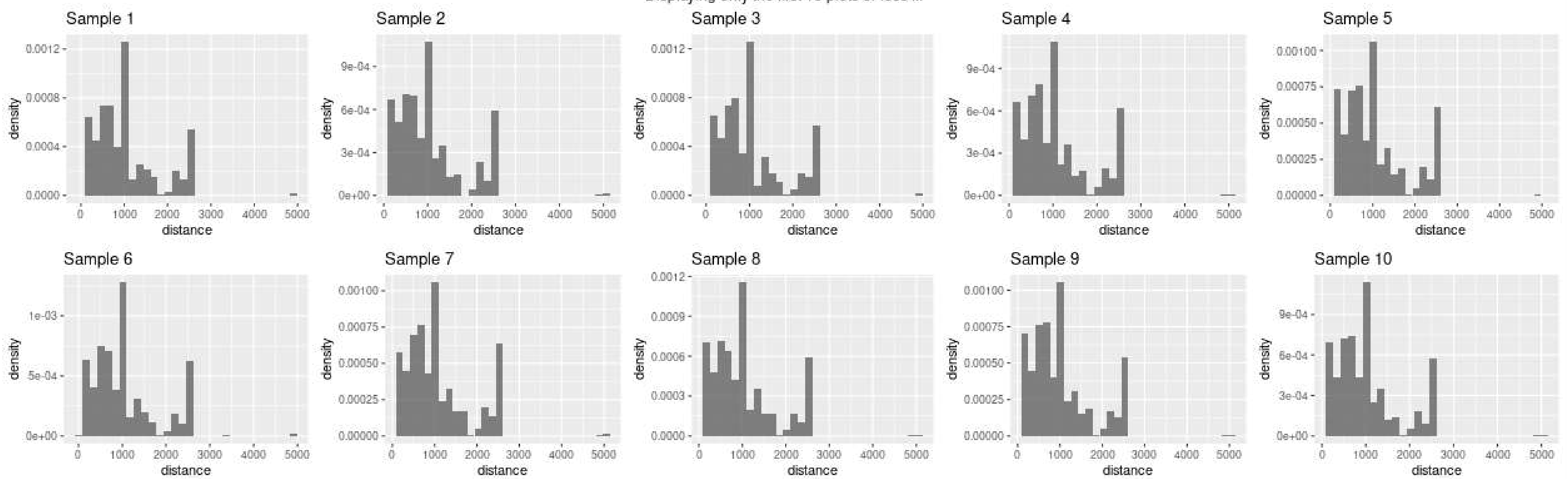
```
Mean & SD for samples generated

          S1      S2      S3      S4      S5      S6      S7      S8      S9
mean 1042.98 1075.58 1048.25 1061.51 1040.78 1017.94 1019.73 1045.18 1039.53
sd    731.88  756.74  712.79  729.99  723.69  720.84  711.04  736.02  716.08
         S10
mean 1061.15
sd    732.36
```

2 types of inferences

- Confidence Intervals
- Hypothesis Testing

Displaying only the first 15 plots or less ...



Confidence Intervals for proportions

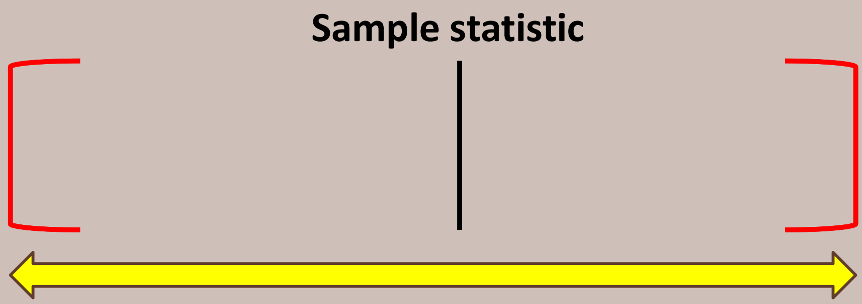
Sample statistic = **population parameter** + **random error**

What we have

What we want to estimate

What we don't know

Confidence Interval



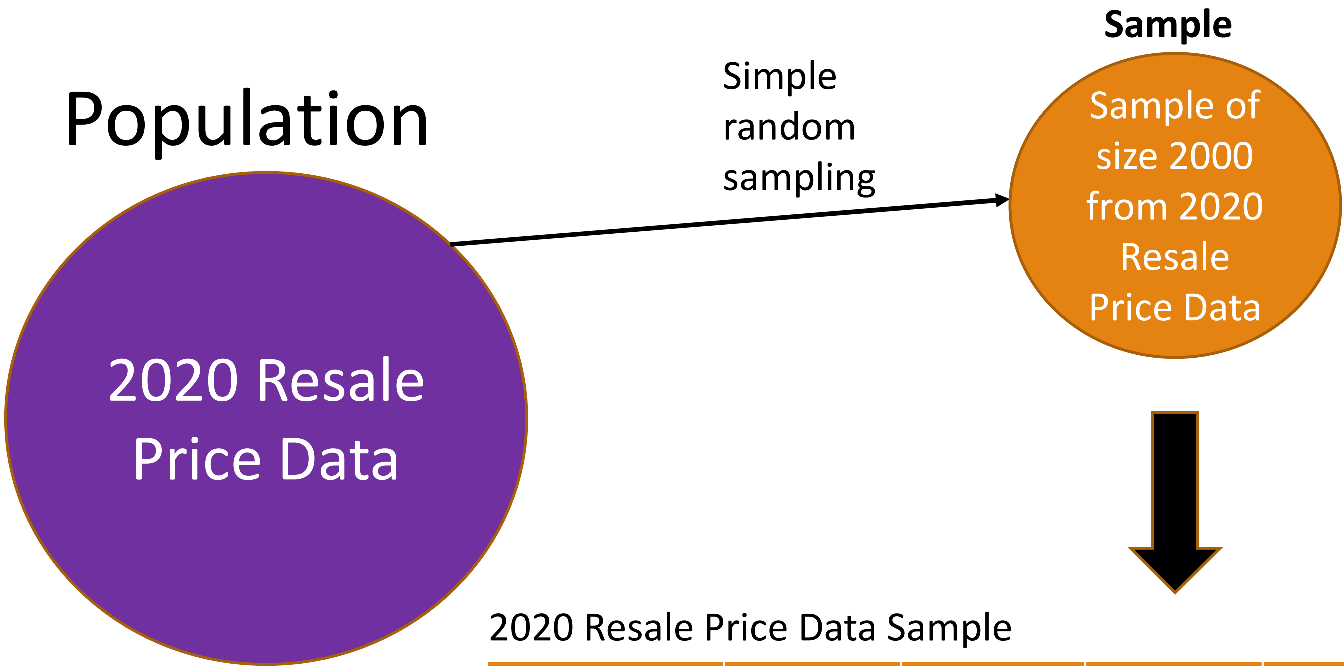
Reasonably certain that population parameter lies within this range of values

Confidence Intervals for:

- Population proportions
- Population means

2020 Resale Price Data

1	year	month	town	flat_type	block	street_na	lstorey	ustorey	floor_area	flat_mode	lease_con	remaining	resale_pri	price_psm
2	2020	1	ANG MO K	3 ROOM	208	ANG MO K	4	6	73	New Gene	1976	55.58333	265000	3630.137
3	2020	1	ANG MO K	3 ROOM	307C	ANG MO K	19	21	70	Model A	2012	91.66667	470000	6714.286
4	2020	1	ANG MO K	3 ROOM	319	ANG MO K	1	3	73	New Gene	1977	56.33333	230000	3150.685
5	2020	1	ANG MO K	3 ROOM	216	ANG MO K	4	6	73	New Gene	1976	55.25	280000	3835.616
6	2020	1	ANG MO K	3 ROOM	556	ANG MO K	7	9	68	New Gene	1980	59.08333	220000	3235.294
7	2020	1	ANG MO K	3 ROOM	536	ANG MO K	10	12	68	New Gene	1980	59.08333	280000	4117.647
8	2020	1	ANG MO K	3 ROOM	560	ANG MO K	4	6	67	New Gene	1980	59.08333	240000	3582.09
9	2020	1	ANG MO K	3 ROOM	463	ANG MO K	4	6	82	New Gene	1980	59.16667	301000	3670.732
10	2020	1	ANG MO K	3 ROOM	476	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
11	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	233000	3477.612
12	2020	1	ANG MO K	3 ROOM	578	ANG MO K	7	9	67	New Gene	1980	59	250000	3731.343
13	2020	1	ANG MO K	3 ROOM	442	ANG MO K	4	6	67	New Gene	1979	58.58333	245000	3656.716
14	2020	1	ANG MO K	3 ROOM	445	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
15	2020	1	ANG MO K	3 ROOM	435	ANG MO K	4	6	67	New Gene	1979	58	288000	4298.507
16	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	235000	3507.463
17	2020	1	ANG MO K	3 ROOM	466	ANG MO K	7	9	67	New Gene	1984	63.66667	268000	4000
18	2020	1	ANG MO K	3 ROOM	570	ANG MO K	4	6	67	New Gene	1979	58.41667	240000	3582.09
19	2020	1	ANG MO K	3 ROOM	345	ANG MO K	4	6	88	New Gene	1978	57.33333	320000	3636.364
20	2020	1	ANG MO K	3 ROOM	213	ANG MO K	4	6	67	New Gene	1976	55.25	250000	3731.343
21	2020	1	ANG MO K	3 ROOM	126	ANG MO K	1	3	67	New Gene	1978	57.75	250000	3731.343
22	2020	1	ANG MO K	3 ROOM	587	ANG MO K	1	3	67	New Gene	1979	58.41667	265000	3955.224
23	2020	1	ANG MO K	3 ROOM	121	ANG MO K	10	12	67	New Gene	1978	57.75	3.00E+05	4477.612
24	2020	1	ANG MO K	3 ROOM	345	ANG MO K	4	6	73	New Gene	1978	57.25	320000	4383.562
25	2020	1	ANG MO K	3 ROOM	212	ANG MO K	10	12	67	New Gene	1977	56.25	255000	3805.97
26	2020	1	ANG MO K	3 ROOM	301	ANG MO K	4	6	88	New Gene	1978	57.16667	355000	4034.091
27	2020	1	ANG MO K	3 ROOM	120	ANG MO K	4	6	67	New Gene	1978	57.75	270000	4029.851



2020 Resale Price Data Sample

flat_type	1-room	2-room	3-room	4-room	5-room	Executive	Multi-Generational
Frequency of each flat_type	2	41	464	819	508	165	1
Proportion	0.001	0.0205	0.232	0.4095	0.254	0.0825	0.0005

2020 Resale Price Data Sample

flat_type	1-room	2-room	3-room	4-room	5-room	Executive	Multi-Generational
Frequency of each flat_type	2	41	464	819	508	165	1
Proportion	0.001	0.0205	0.232	0.4095	0.254	0.0825	0.0005

Sample proportion

Question: What is p, the population proportion of 5-room flat_type in the population of transacted HDB flats in 2020?

Population parameter

Census of 23334 transacted HDB flats in 2020 in Singapore

flat_type	1-room	…	Multi-Generational
Frequency of each flat_type	?	…	…
Proportion	?	…	…

0.254 = population proportion of 5-room flat_type + random error

population proportion of
1-member households

0.254

0.254 = population proportion of 5-room flat_type + random error

2020 Resale Price Data Sample

flat_type	1-room	2-room	3-room	4-room	5-room	Executive	Multi-Generational
Frequency of each flat_type	2	41	464	819	508	165	1
Proportion	0.001	0.0205	0.232	0.4095	0.254	0.0825	0.0005

Sample proportion p^* = 0.254

Confidence interval for population proportion

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

Sample proportion, p^*
Sample size, n
Value from standard normal distribution, z^*

Confidence interval for population proportion

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

Sample proportion p^*: 0.254
Sample size n: 2000
Value from standard normal distribution, z^*

Value from standard normal distribution z^*

- For 90% confidence interval, z^* is 1.645
- For 95% confidence interval, z^* is 1.96

Standard Normal Probabilities

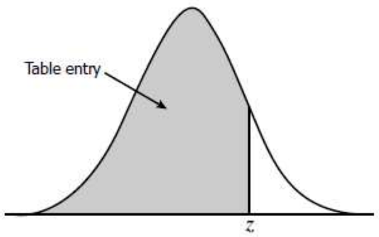


Table entry for z is the area under the standard normal curve to the left of z.

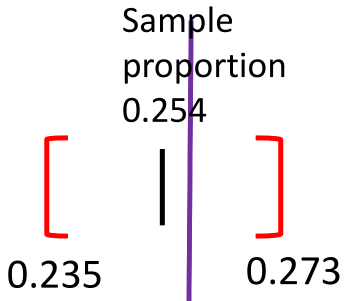
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Confidence interval for population proportion of 5-room flat_type:
95% CI: 0.254 ± 0.0191

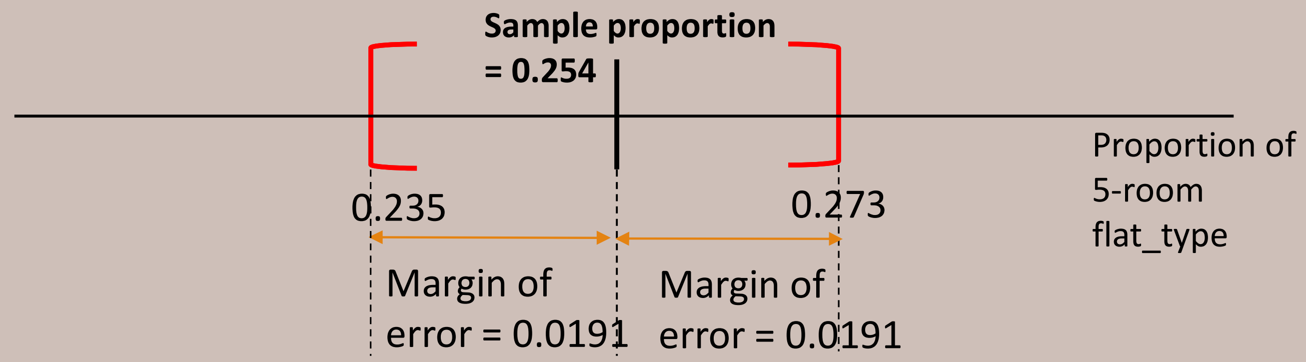
Interpretation of Confidence Intervals

0.254 = population proportion of 5-room flat_type + random error

population proportion of 5-room flat_type



Confidence Interval



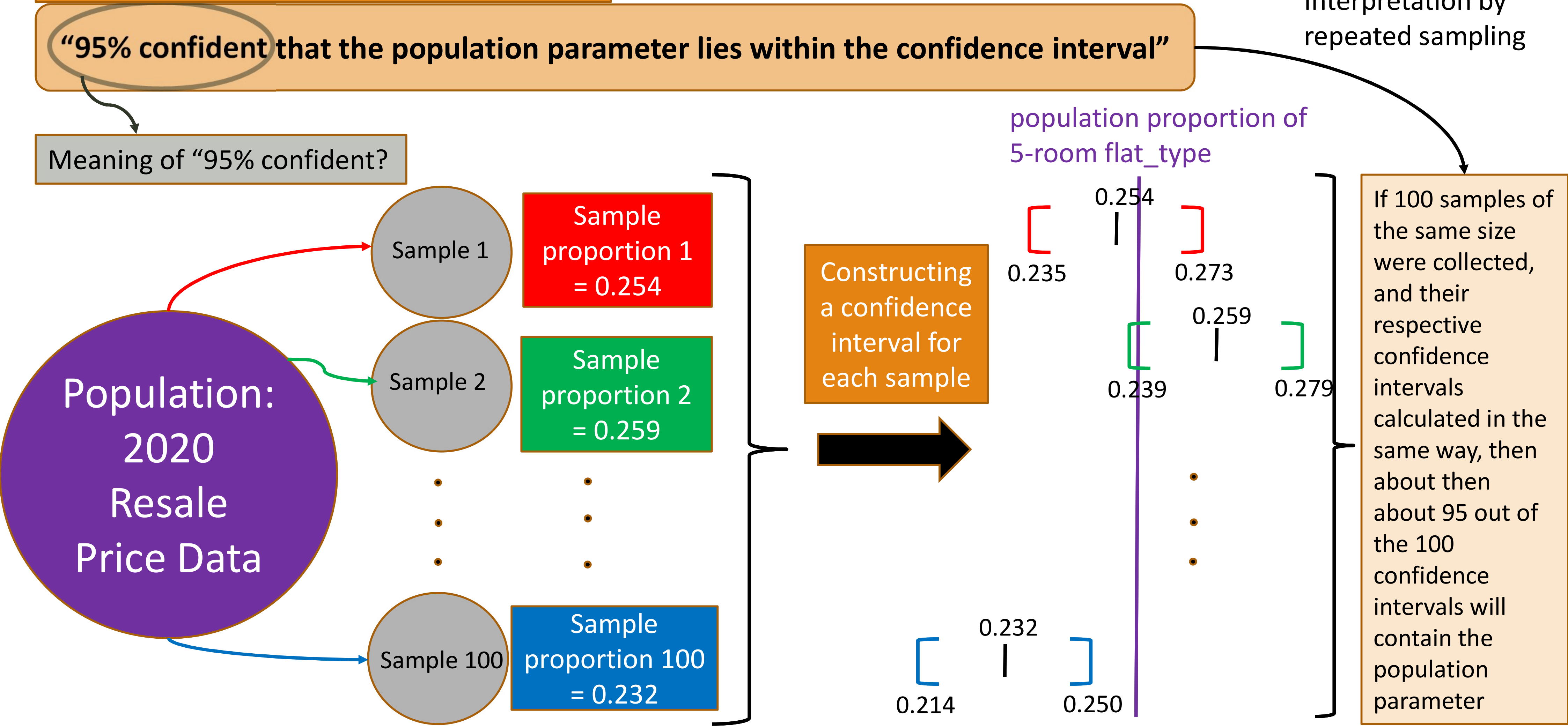
95% CI: 0.254 ± 0.0191

Confidence level

Interval

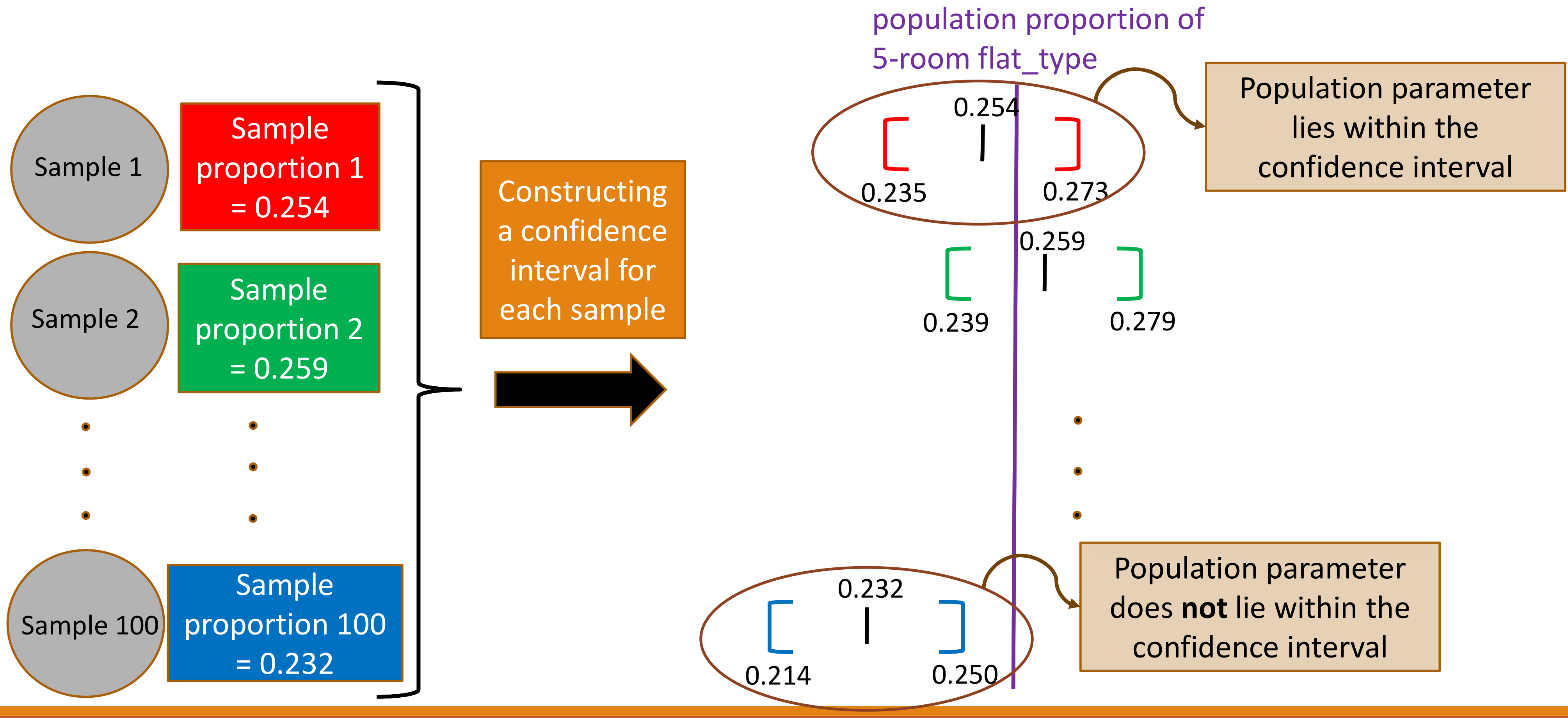
“95% confident that the population parameter lies within the confidence interval”

95% CI: 0.254 ± 0.0191

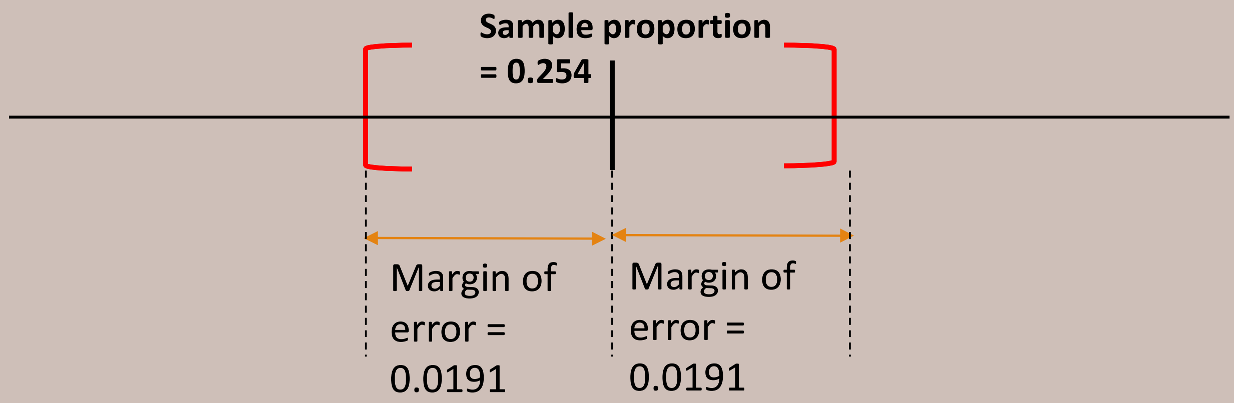


Sample 100 - 95% CI: 0.232 ± 0.0185

“95% confident that the population parameter lies within the confidence interval”



Confidence Interval



95% CI: 0.254 ± 0.0191

95% chance that the population proportion of 5-room flat_type lies between 0.235 and 0.273?

95% confident that the population proportion of 5-room flat_type lies between 0.235 and 0.273

The chances are in the sampling procedure, not the parameter.

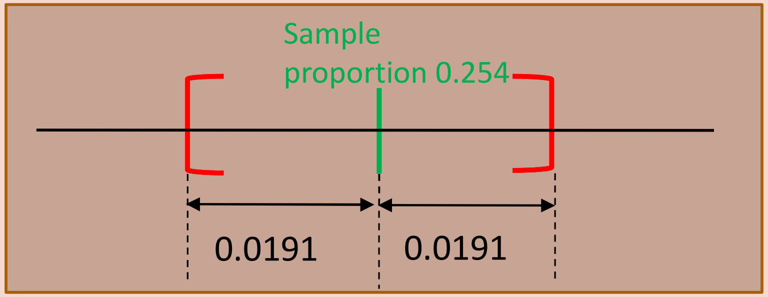
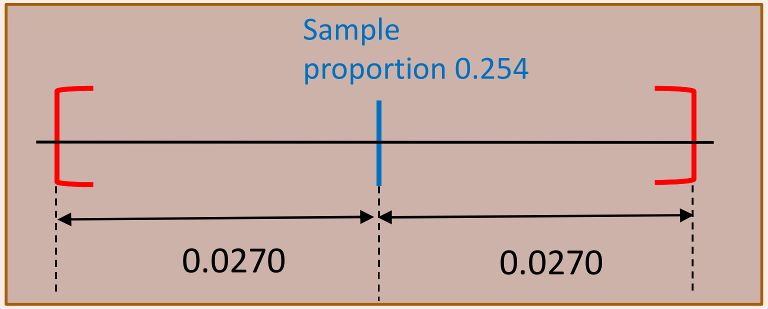
Properties of confidence intervals

Sample of size 2000

Sample proportion of 5-room flat_type = 0.254

Sample of size 1000

Sample proportion of 5-room flat_type = 0.254

Sample size	**Sample proportion**	**Confidence level**	**Confidence interval**
2000	0.254	95%	95% CI: 0.254 ± 0.0191 
1000	0.254	95%	95% CI: 0.254 ± 0.270 

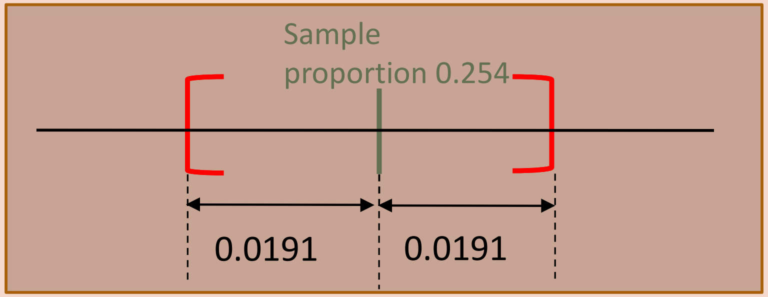
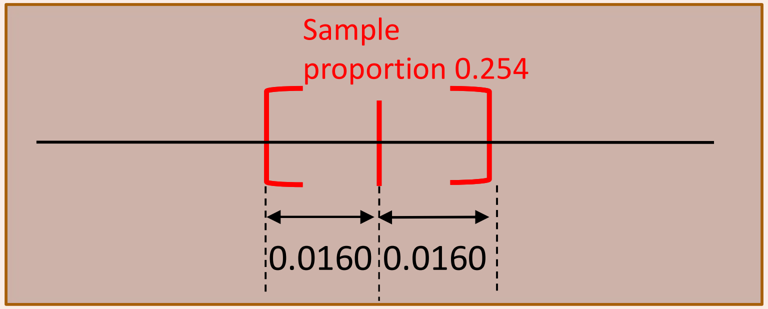
Properties of confidence intervals

Sample of size 2000

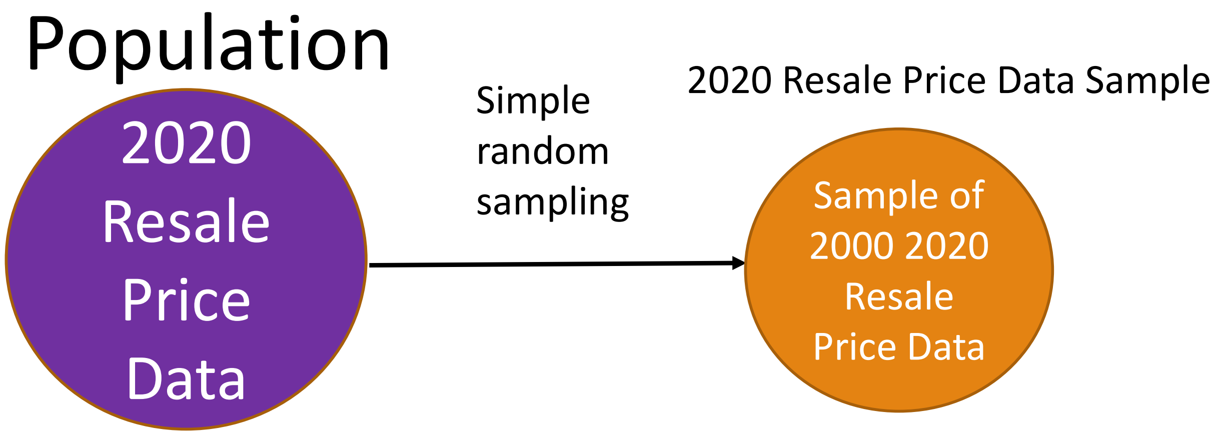
Sample proportion of 5-room flat_type = 0.254

Sample of size 2000

Sample proportion of 5-room flat_type = 0.254

Sample size	Sample proportion	Confidence level	Confidence interval
2000	0.254	95%	95% CI: 0.254 ± 0.0191 
2000	0.254	90%	90% CI: 0.254 ± 0.0160 

Confidence Intervals for means



1	year	month	town	flat_type	block	street_nai	lstorey	ustorey	floor_area	flat_mode	lease_con	remaining	resale_pr	price_psm
2	2020	1	ANG MO K	3 ROOM	208	ANG MO K	4	6	73	New Gene	1976	55.58333	265000	3630.137
3	2020	1	ANG MO K	3 ROOM	307C	ANG MO K	19	21	70	Model A	2012	91.66667	470000	6714.286
4	2020	1	ANG MO K	3 ROOM	319	ANG MO K	1	3	73	New Gene	1977	56.33333	230000	3150.685
5	2020	1	ANG MO K	3 ROOM	216	ANG MO K	4	6	73	New Gene	1976	55.25	280000	3835.616
6	2020	1	ANG MO K	3 ROOM	556	ANG MO K	7	9	68	New Gene	1980	59.08333	220000	3235.294
7	2020	1	ANG MO K	3 ROOM	536	ANG MO K	10	12	68	New Gene	1980	59.08333	280000	4117.647
8	2020	1	ANG MO K	3 ROOM	560	ANG MO K	4	6	67	New Gene	1980	59.08333	240000	3582.09
9	2020	1	ANG MO K	3 ROOM	463	ANG MO K	4	6	82	New Gene	1980	59.16667	301000	3670.732
10	2020	1	ANG MO K	3 ROOM	476	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
11	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	233000	3477.612
12	2020	1	ANG MO K	3 ROOM	578	ANG MO K	7	9	67	New Gene	1980	59	250000	3731.343
13	2020	1	ANG MO K	3 ROOM	442	ANG MO K	4	6	67	New Gene	1979	58.58333	245000	3656.716
14	2020	1	ANG MO K	3 ROOM	445	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97

2020 Resale Price Data Sample

Sample mean resale_price, $\bar{x}$ = $448,727

2020 Resale Price Data

Census

Population mean resale_price μ

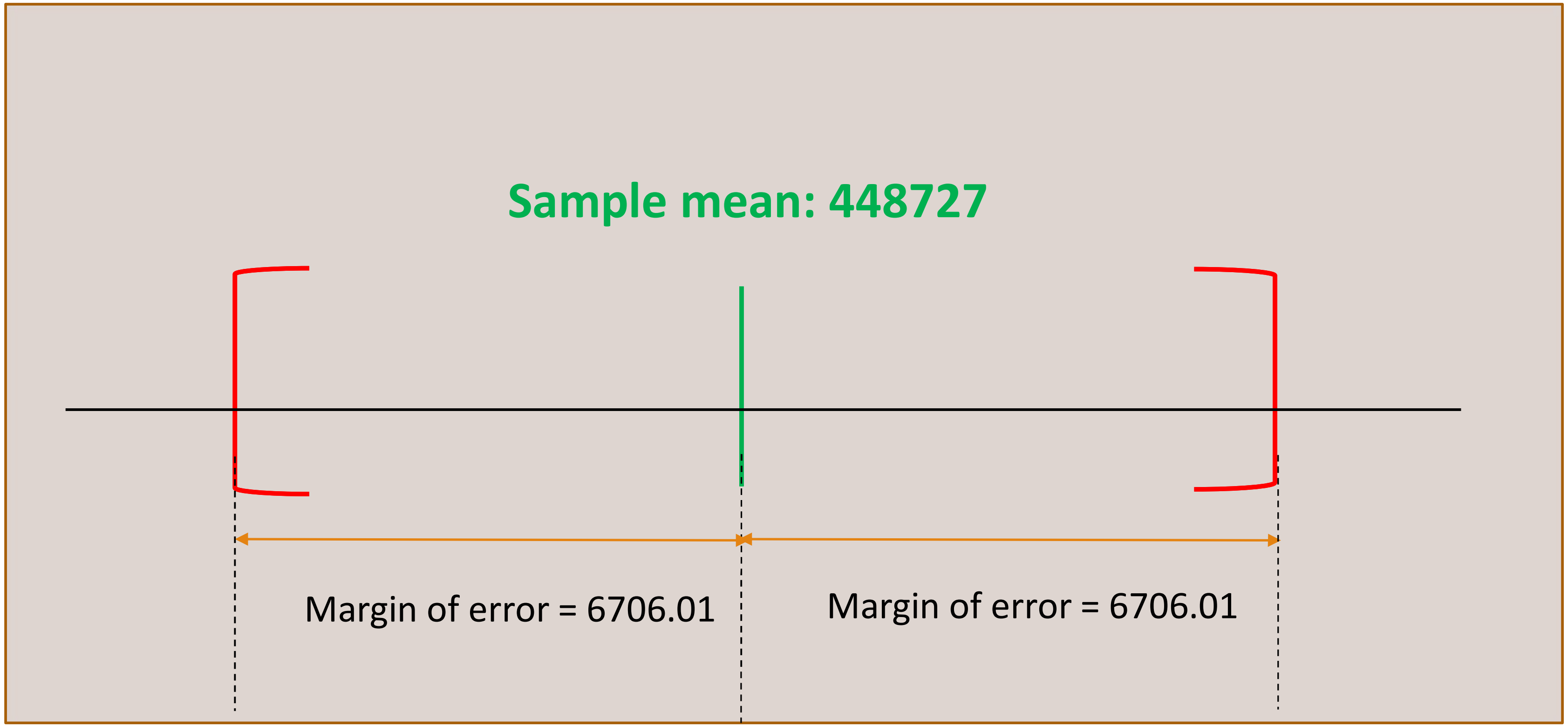
Question: What is μ, the population mean of resale_price of all transacted HDB flats in 2020?

Population parameter

Confidence interval for population mean

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Sample mean, $\bar{x}$: 448727
Sample size, n: 2000
Sample standard deviation, s
Value from t-distribution, t^*



Sample mean: 448727
Margin of error = 6706.01
Margin of error = 6706.01

Summary of topics covered for confidence intervals

1. Confidence intervals as a way to quantify random error
2. Understanding of confidence intervals via repeated sampling
3. Properties of confidence intervals and its relation to size of random error
4. Construction of confidence intervals from data
5. Construction of confidence intervals for population means and population proportion.

Hypothesis testing

- Sometimes we are faced with a ("yes/no") decision problem involving a large population.

- For example, from the government's point of view, is a vaccine effective and safe enough to be rolled out to the entire populace?

- One way we can answer such questions from sample data, is through hypothesis testing.

Hypothesis Testing

- There are a few key steps to hypothesis testing.
  - Step 1: Identify the question and state the null hypothesis and alternative hypothesis.
  - Step 2: Collecting relevant data. Decide on the relevant test statistic.
  - Step 3: Determining the level of significance and computing the p-value.
  - Step 4: Making conclusion about the null hypothesis.

Hypothesis Testing : Formulating hypotheses

- Step 1 of the hypothesis testing is formulating hypotheses.

- We produce 2 hypotheses namely the null hypothesis and alternative hypothesis.

- The null hypothesis takes a stance of no difference or no effect. This hypothesis assumes that any differences seen are due to the variability inherent in the population and could have occurred by random chance.

- Alternative hypothesis is typically what we wish to confirm and pit against the null hypothesis.

- The two hypotheses must be *mutually exclusive*: they cannot both be true.

Hypothesis Testing : Formulating hypotheses

- Let's say we suspect a particular coin is biased towards heads. Then we can formulate our null hypothesis and alternative hypothesis as such:

- Null hypothesis H_0 : "The coin is fair."

- Alternative hypothesis H_1 : "The coin is biased towards heads."

Hypothesis Testing : Formulating hypotheses

- Do note that in this module, we could formulate our null hypothesis in numbers, instead of words. Let us use the same example as the previous slide.

- Let H be the event that the coin lands on heads, in a single coin toss. We can write our hypothesis as:

- Null hypothesis H_0 : P(H) = 0.5.

- Alternative hypothesis H_1 : P(H) > 0.5.

Hypothesis Testing : Collect Relevant Data

- Step 2 of the hypothesis testing is collecting relevant sample data for the experiment.

- For example, when conducting a hypothesis test on the fairness of a coin, we can toss the coin 8 times and observe all the outcomes.

- We are sampling 8 observations – one for each coin toss.

Hypothesis Testing : Relevant random variable, test statistic.

- We have previously defined what is a random variable.

- In our example, our random variable is the number of heads out of 8 independent coin tosses, assuming the coin is fair.

- The test statistic is a value computed using data which you use to determine whether to reject the null hypothesis or not.

- In our example, the test statistic is the observed number of heads. We will talk more about the test statistic in the next unit.

Hypothesis Testing : Set the significance level: a number between 0 and 1

- The lower the significance level, the greater the evidence needs to be in order to conclude the alternative hypothesis over the null.

- A commonly used level of significance is 0.05, or 5% level of significance.

- Other commonly used level of significance is 0.10 (10% level of significance) or 0.01 (1% level of significance)

Hypothesis Testing : Computing the p-value.

- Now we compute the p-value.

- The p-value is the probability of obtaining a test result at least as extreme as the result we observed, assuming the null hypothesis is true.

- We will illustrate it using an example in the next slide.

- The p-value can also be thought of as the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis was true.

Hypothesis Testing : Computing the p-value.

- Let's continue with our example on the coin. Say we observed 7 heads out of 8 toss.

- p-value = P(obtaining a result at least as extreme as observed | null hypothesis is true)

  = P(7 heads out of 8 toss| null hypothesis is true) +

  P(8 heads out of 8 toss | null hypothesis is true)

  $= 8\,(1/2)^8 + (1/2)^8 = 9(1/2)^8 = 0.035156.$

Hypothesis Testing : Making conclusion.

- ❑ Now we determine whether to reject or not to reject the null hypothesis.
- ❑ Decide between one of the ONLY two options:
  - ❑ Reject the null hypothesis in favour of the alternative, if p-value < significance level.
  - ❑ Do not reject the null hypothesis, if p-value ≥ significance level.
- ❑ Recall in our example, our p-value = 0.035156. With this p-value, we can reject the null hypothesis at the 5% level of significance but not at the 3% level of significance.

Hypothesis Testing : Computing the p-value.

- You might be wondering what we mean by "at least as extreme as observed".

- Suppose we want to test if a coin is biased towards heads. Let X be the number of heads we see out of 8 independent tosses of the coin. Let us say we observed 3 heads out of 8 coin tosses.

- Null hypothesis H_0 : "The coin is fair," P(H) = 0.5,

  Alternative hypothesis H_1 : "The coin is biased towards heads," P(H) > 0.5.

- Recall our definition of p-value. "The probability of obtaining a test result at least as extreme as the one observed, assuming the null hypothesis is true."

Hypothesis Testing : Computing the p-value.

- What is the range of test results at least as extreme as observed?

- In the context of computing p-value, “at least as extreme” is interpreted as “at least as favorable to the alternative hypothesis”.

- In our scenario, since the alternative hypothesis is P(H) > 0.5, the greater the value X assumes, the more favorable the case is to the alternative hypothesis.

- Thus, the range of test results at least as extreme as that observed will be all possible values of X greater than what is observed, i.e $3 \leq X \leq 8$.

Hypothesis Testing : Computing the p-value.

❑ Now suppose we want to test if the coin is biased towards tails. We define X following the previous example. We formulate our hypothesis:

❑Null hypothesis H_0 : "The coin is fair," P(H) = 0.5,

Alternative hypothesis H_1 : "The coin is biased towards tails," P(H) < 0.5.

❑ Let's say we observed 3 heads out of 8 coin tosses. What is the range of test results at least as extreme as observed?

Hypothesis Testing : Computing the p-value.

- In this case, the range of test results at least as extreme as observed would be $0 \leq X \leq 3$.

- From the previous two examples, we can see that "at least as extreme as observed" is dependent on what the alternative hypothesis is.

- From this we also observe that the same experiment can give us two different p-values.

- Thus, when making computation on p-values, it is important that we know what the null hypothesis and alternative hypothesis is.

Hypothesis Testing : Common Misconceptions

- If p-value is not lower than the level of significance, we cannot reject the null hypothesis which means we don’t know if the observation is due to chance or not.

- Not rejecting the null hypothesis doesn’t mean the null hypothesis is true.

- There does not exist a scenario where we attempt to reject alternative hypothesis. The p-value is calculated based on the null hypothesis; we can’t reject the alternative hypothesis.

- We only carry out hypothesis tests when working with sample data. When given population data, we do not conduct hypothesis tests.

Common hypothesis tests

Common Hypothesis Tests

- The previous unit outlined a general framework for carrying out a hypothesis test. However, there are some important details missing.
    - How do we go about deciding on which test statistic to use?

Common Hypothesis Tests

- ❑ These decision are made based on:
  - ❑ The kinds of hypotheses we want to test.
  - ❑ The distribution of the test statistics.

- ❑ Let's go through two common hypothesis tests in this unit.

Common Hypothesis Tests : One-sample t-test (Example)

❑ Suppose we are interested in the average price of HDB flats in Singapore in the year 2020. Researcher A claims the average price of HDB flats in Singapore is 600 thousand dollars. However, researcher B believes it is higher than 600 thousand dollars.

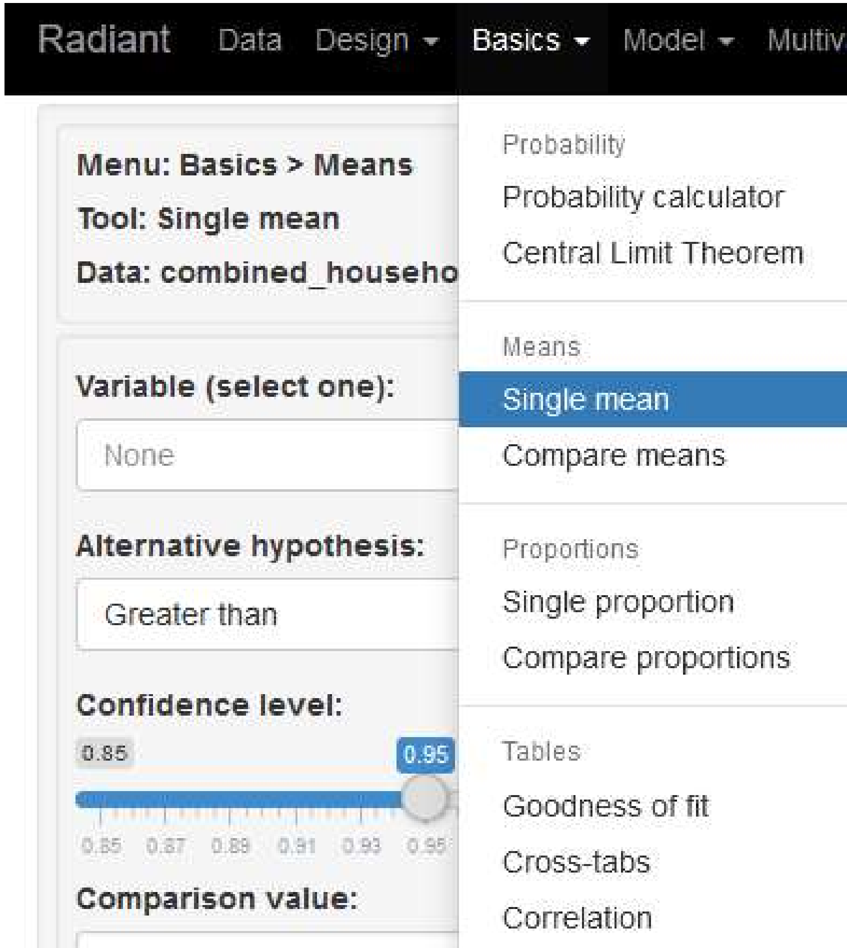
❑ Null hypothesis, H_0: μ = 600 thousand.

❑ Alternative hypothesis, H_1: μ > 600 thousand.

❑ Now we draw a simple random sample of 1000 households from 1.3 million households in Singapore.

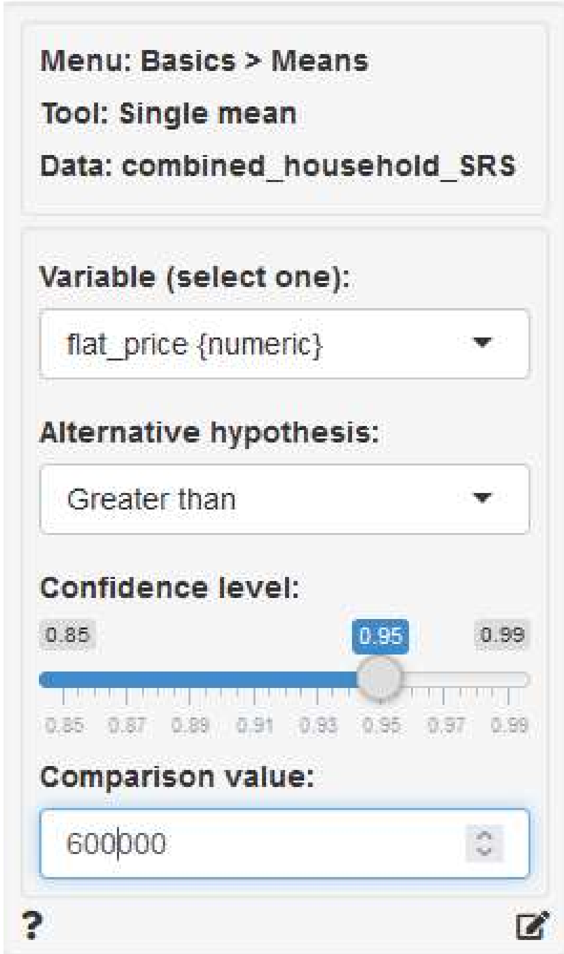
Common Hypothesis Tests : One-sample t-test (Example)

- ❑ We use Radiant to do a one-sample t-test. First, load the relevant dataset into Radiant.
- ❑ On the toolbar, click on the tab “Basics” -> “Single Means”.



Common Hypothesis Tests : One-sample t-test (Example)

- ❑ Under “Variable”, click “flat_price”.
- ❑ Under “Alternative hypothesis”, click “Greater than”.
- ❑ Under “Comparison value”, type “600,000”.



Common Hypothesis Tests : One-sample t-test (Example)

- ❑ Radiant would compute the p-value for us when we key in the values in the previous slide.

- ❑ We see that the p-value is less than 0.05.

- ❑ At the 5% level of significance, we have sufficient evidence to reject the null hypothesis and accept the alternative hypothesis.

```
Single mean test
Data       : combined_household_SRS
Variable   : flat_price
Confidence: 0.95
Null hyp. : the mean of flat_price = 600000
Alt. hyp. : the mean of flat_price is > 600000

       mean     n n_missing          sd         se        me
 616,326.799 1,000        0 155,981.331 4,932.563 9,679.372

    diff       se t.value p.value  df       5% 100%
 16326.8 4932.563    3.31  < .001 999 608205.9  Inf ***
```

Common Hypothesis Tests : Criteria for conducting a t-test

- We will list down the criteria needed to run a t-test.
  - The population distribution should be approximately normal if n, the sample size, is smaller than 30.
  - The data used is produced randomly.

Common Hypothesis Tests : Chi-squared test

- The second test we are going to cover is the chi-squared test.

- It is commonly used to check whether two categorical variables, A and B are associated at the population level.

- Let us illustrate the chi-squared test with an example.

Common Hypothesis Tests : Chi-squared test (Example)

- Let us use the same sample of 1000 households from 1.3 million households in Singapore that we use in the example on one-sample t-test.

- In our sample, let us classify all flats that cost more than equal to 600,000 dollars as "Expensive", and less than 600,000 dollars as "Not Expensive".

- Let us classify all households with more than 3 members as "Large", and all households with 3 or less members as "Small".

Common Hypothesis Tests : Chi-squared test (Example)

- We then formulate our null and alternative hypothesis based on the data given.

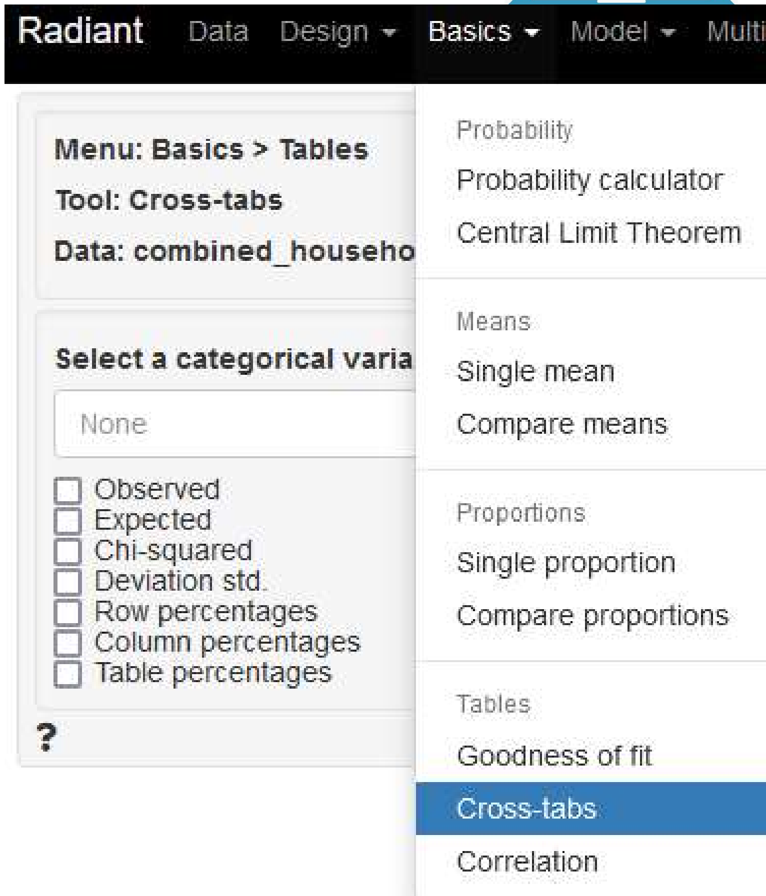
- Null hypothesis, H_0: Household type is not associated with flat value in Singapore.

- Alternative hypothesis, H_1: Household type is associated with flat value in Singapore.

Common Hypothesis Tests : Chi-squared test (Example)

- ❑ We use Radiant to compute the p-value for this hypothesis test.

- ❑ On the toolbar, go to “Basics” -> “Cross-tabs”.

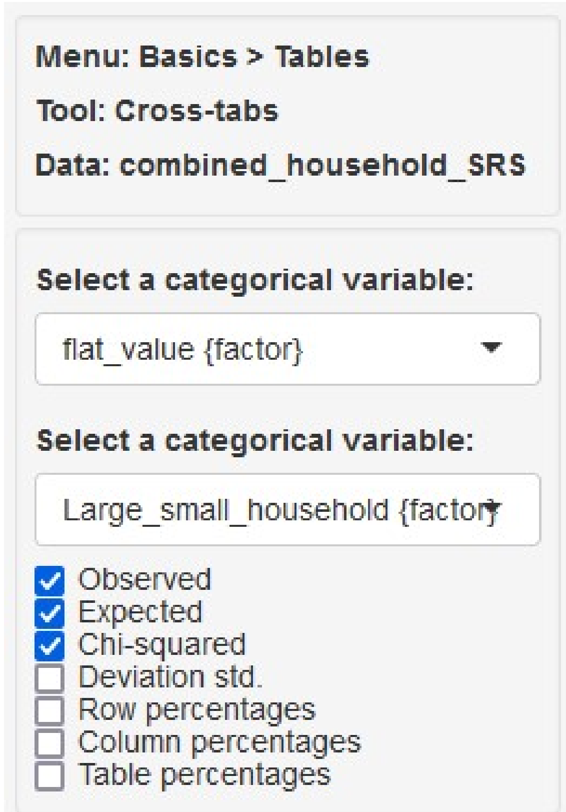


Common Hypothesis Tests : Chi-squared test (Example)

❑ Since we select “Cross-tabs”, we are able to make 2 selections of categorical variables.

❑ Choose “flat_value” and “Large_small_household” as our categorical variables.

❑ Tick the options “Observed”, “Expected”, “Chi-squared”.



Common Hypothesis Tests : Chi-squared test (Example).

❑ Radiant will run the chi-squared test for us.

❑ The p-value is 0.278. Since 0.278 > 0.05, we do not have enough evidence to reject the null hypothesis at the 5% level of significance.

❑ Thus, we cannot conclude that household type is associated with flat value in Singapore.

```
Observed:
                    Large_small_household
flat_value           Large Small Total
  Expensive            193   313   506
  Not Expensive        205   289   494
  Total                398   602 1,000

Expected: (row total x column total) / total
                    Large_small_household
flat_value           Large     Small     Total
  Expensive            201.39    304.61    506.00
  Not Expensive        196.61    297.39    494.00
  Total                398.00    602.00 1,000.00

Contribution to chi-squared: (o - e)^2 / e
                    Large_small_household
flat_value           Large Small Total
  Expensive          0.35  0.23  0.58
  Not Expensive      0.36  0.24  0.59
  Total              0.71  0.47  1.17

Chi-squared: 1.175 df(1), p.value 0.278
```

Common Hypothesis Tests : Criteria for Chi-squared test.

- We will list down the criteria needed to run a chi-squared test.

  - The data must be counts for the categories of a categorical variable.

  - Similar to the one sample t-test, the sample taken from the population should be a random sample before the test can be deployed.

Common Hypothesis Tests : Summary

One Sample t-test	Chi-squared Test
-Mainly used when testing for significant difference between sample mean and a known/hypothesized mean	-Mainly used when testing for association between two categorical variables
-population distribution should be approximately normal if n, the sample size, is smaller than 30.	-The data given is the count for the categories of a categorical variable.
-Data used is acquired randomly.	-Data used is acquired randomly.