EBU5601
Data Design

Data and Sampling Distribution

Dr Chao Shu, Dr Xiaolan Liu
School of Electronic Engineering and Computer Science
Queen Mary University of London

Sep. 2024

Learning Outcomes

- The main outcomes are:
  - [**LO5.1**] Describe and calculate Binomial Distributions
  - [**LO5.2**] Understand the normal distribution function and properties
  - [**LO5.3**] Apply the properties of normal distribution and z-table to calculate probabilities
  - [**LO5.4**] Evaluate normality of real-life data using Q-Q plot
  - [**LO5.5**] Analyse percentages/proportions of real-life data based on its approximate normal distribution model
  - [**LO5.6**] Understand the concept and applications of descriptive and inferential statistics and relevant terminologies

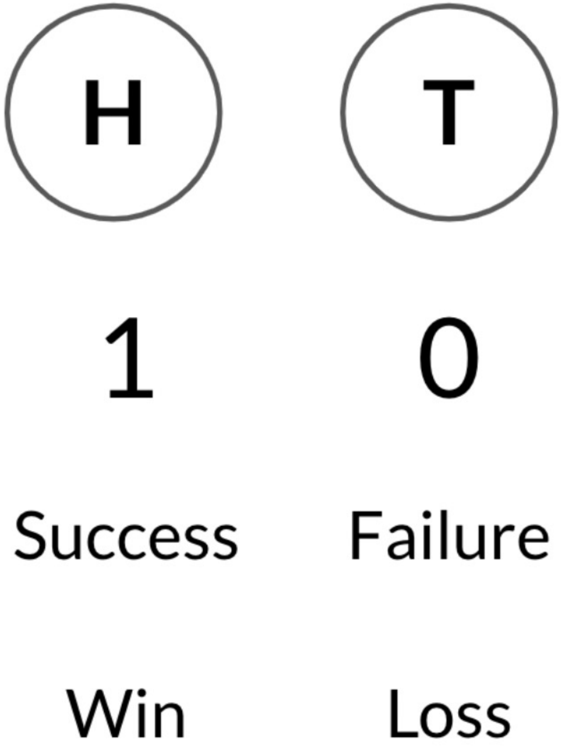
Learning Outcomes

- The main outcomes are:
  - [**LO5.7**] Apply sampling distribution and bootstrapping techniques to analyse population parameters by using Python codes
  - [**LO5.8**] Understand the Law of Large Numbers and Central Limit Theorem and limitations of the Central Limit Theorem
  - [**LO5.9**] Understand the properties of the Student's t-distribution and the relation between t-distribution and z-distribution
  - [**LO5.10**] Understand the properties of the Poisson distribution and how it can be used to model real-life scenarios
  - [**LO5.11**] Understand the properties of the Exponential distribution and the relation between the exponential distribution and the Poisson distribution

BINOMIAL DISTRIBUTION

Binomial Distribution

- Many events in real life can be modelled as random variables, and each of them has a binary outcome (an outcome with only two possible values).



H T

1 0

Success Failure

Win Loss

Binomial Distribution

Example: Coin flips

- Let’s use coin flips to demonstrate the binomial distribution, considering the questions below:
- If we flip 5 fair coins, what is the probability of having exactly 1 Head?

- If we list all possible outcomes, we can find that 5 outcomes will have exactly 1 Head.
- The probability of 1 Head (and 4 Tails):

$$P(H)^1P(T)^4 = 0.5^1\times(1-0.5)^{(5-1)}$$

- the probability of having exactly 2 Heads:

$$5\times0.5^1\times(1-0.5)^{(5-1)}=0.15625$$

Case	Result
1	HTTTT
2	THTTT
3	TTHTT
4	TTTHT
5	TTTTH

Binomial Distribution

Example: Coin flips

- If we flip 5 fair coins, what is the probability of having exactly 2 Heads?
- If we list all possible outcomes, we can find that 10 outcomes will have exactly 2 Head.
- The probability of 2 Heads (and 3 Tails):

$$P(H)^2 P(T)^3 = 0.5^2 \times (1 - 0.5)^{(5-2)}$$

- the probability of having exactly 2 Heads:

$$10 \times 0.5^2 \times (1 - 0.5)^{(5-2)} = 0.3125$$

Case	Result
1	HHTTT
2	HTHTT
3	HTTHT
4	HTTTH
5	THHTT
6	THTHT
7	THTTH
8	TTHHT
9	TTHTH
10	TTTHH

Binomial Distribution

- The **binomial distribution** with parameters n and p is the discrete probability distribution of the **number of successes** in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q=1\text{-}p$)
- In general, if the random variable X follows the binomial distribution with parameters n and p , we write $X\sim B(n,p)$. The probability of getting exactly k successes in n independent experiments trials is given by the probability mass function:

$$P(k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}$$

Binomial Distribution

Example: Coin flips

- If we flip 3 loaded coins, where P(H)=0.8. What is the probability of having exactly 1 Heads?
- Number of outcomes that will have exactly 1 Head.

$$\binom{3}{1}\ \frac{3!}{1!\,(3-1)!}=3$$

- The probability of 1 Head (and 3 Tails):

$$P(H)^1P(T)^2=0.8^1\times(1-0.8)^{(3-1)}=0.032$$

- the probability of having exactly 1 Head:

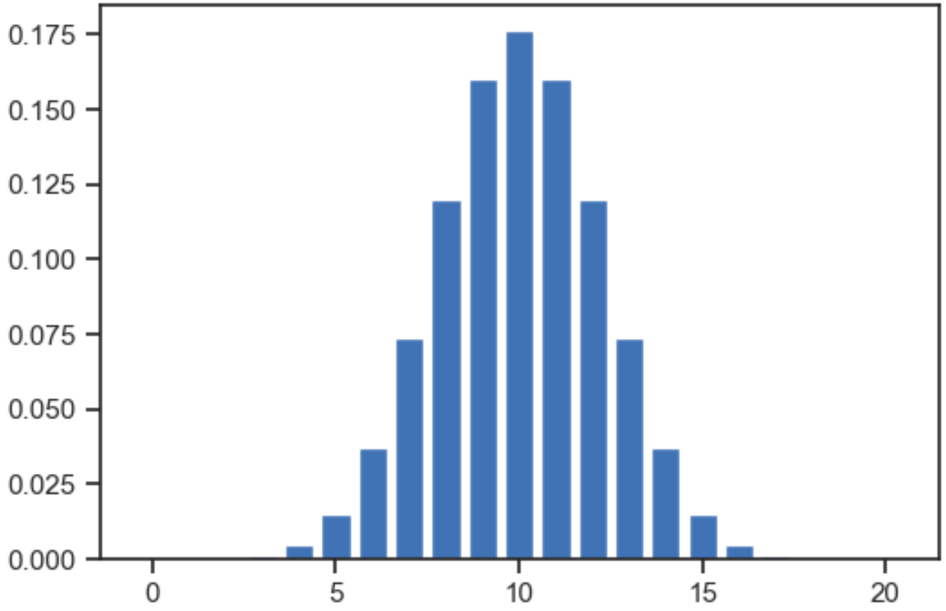
$$3\times 0.032=0.096$$

NORMAL DISTRIBUTION

Introduction

- Let's start with the binomial distribution you have learned and think about one question:
- If we flip a fair coin (P = 0.5) 20 times (n = 20). What is the number of heads (k) that maximizes the probability in the binomial distribution?

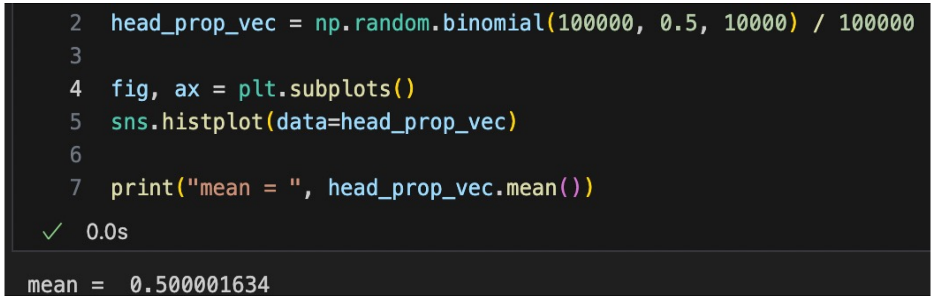
$$P(k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}$$



- The value decreases when it deviates from 10
- The curve of the plot looks like a bell, it is called a bell curve.

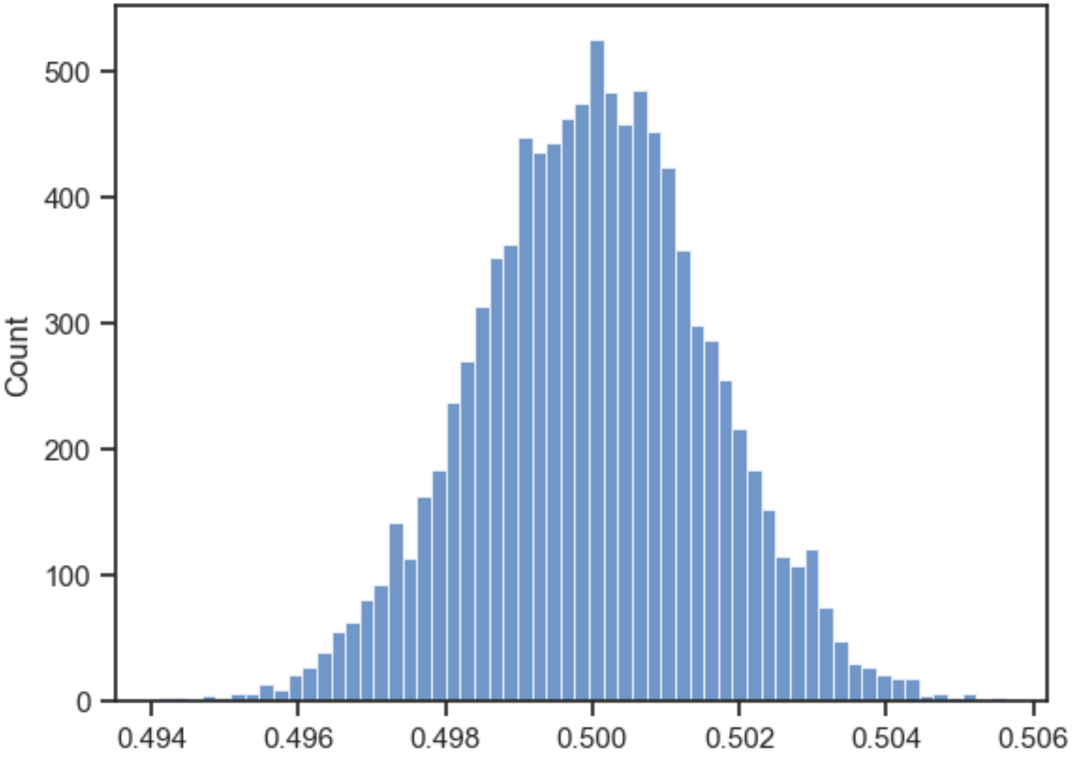
Introduction

- Let's carry out another related experiment.
- We flip a fair coin 100k times in ONE experiment. If we do this experiment 10000 times, what is the shape of the distribution of the mean value?



```python
head_prop_vec = np.random.binomial(100000, 0.5, 10000) / 100000

fig, ax = plt.subplots()
sns.histplot(data=head_prop_vec)

print("mean = ", head_prop_vec.mean())
```

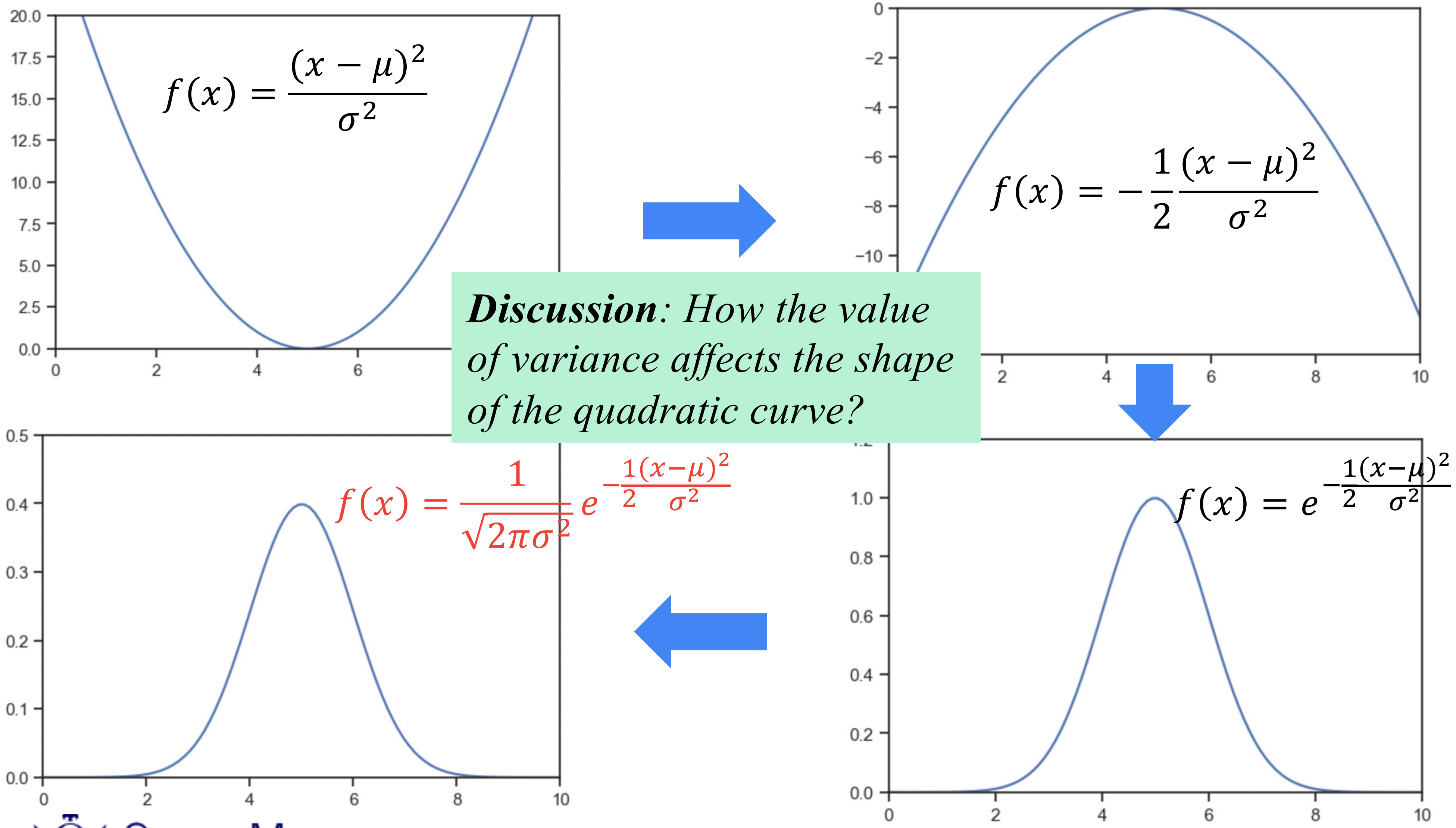
```
✓ 0.0s

mean =  0.500001634
```



- The plot also approximately forms a bell curve

Introduction

Demo: The birth of the bell curve formula



$$f(x) = \frac{(x-\mu)^2}{\sigma^2}$$

$$f(x) = -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

***Discussion**: How the value of variance affects the shape of the quadratic curve?*

$$f(x) = e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

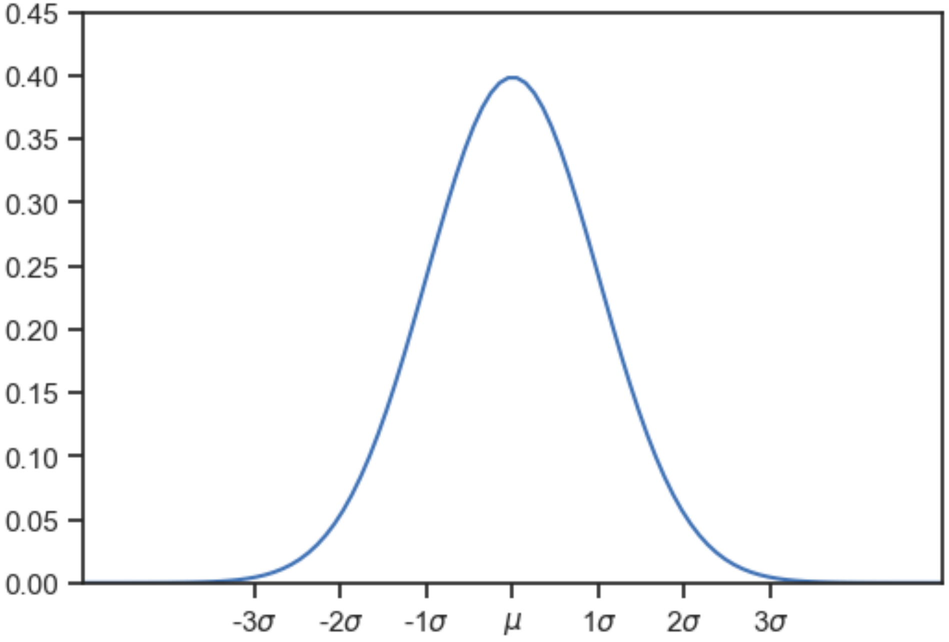
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Introduction

- A normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

where μ is the mean and σ^2 is the variance of the normal distribution.



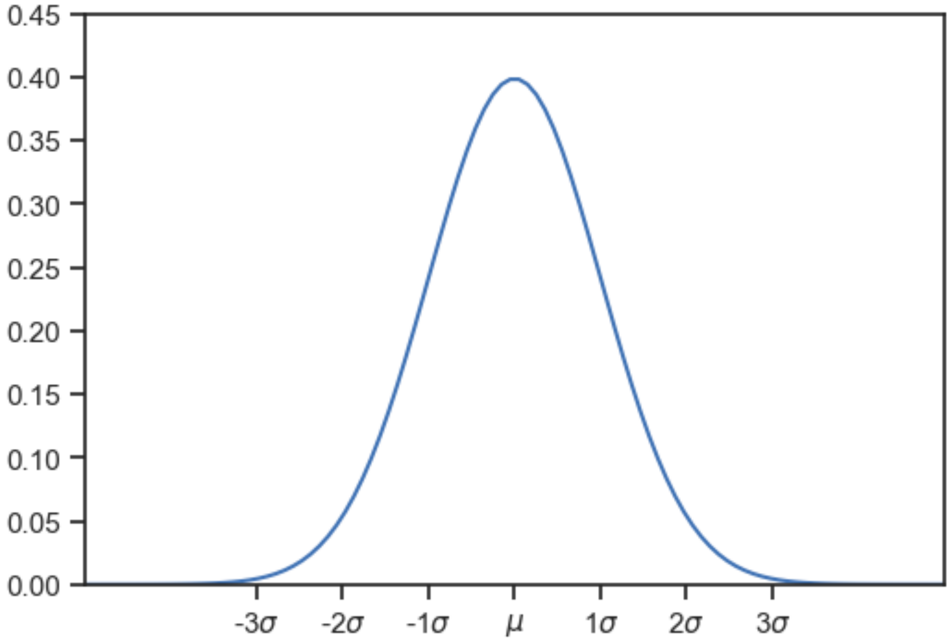
Definition and Properties

- A normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

where μ is the mean and σ^2 is the variance of the normal distribution.

A standard normal model is a normal distribution with μ=0 and σ =1.



1. The curve is symmetric at the center (i.e. around the mean, μ). The function reaches its maximum when $x = \mu$.

2. The mean, mode and median are all equal.

3. The total area under the curve is 1.

4. The shape of the curve becomes wider when σ^2 increases but the area under the curve keeps 1.

Empirical Rule

- The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:
  - 68% of the data falls within one standard deviation of the mean.
  - 95% of the data falls within two standard deviations of the mean.
  - 99.7% of the data falls within three standard deviations of the mean.

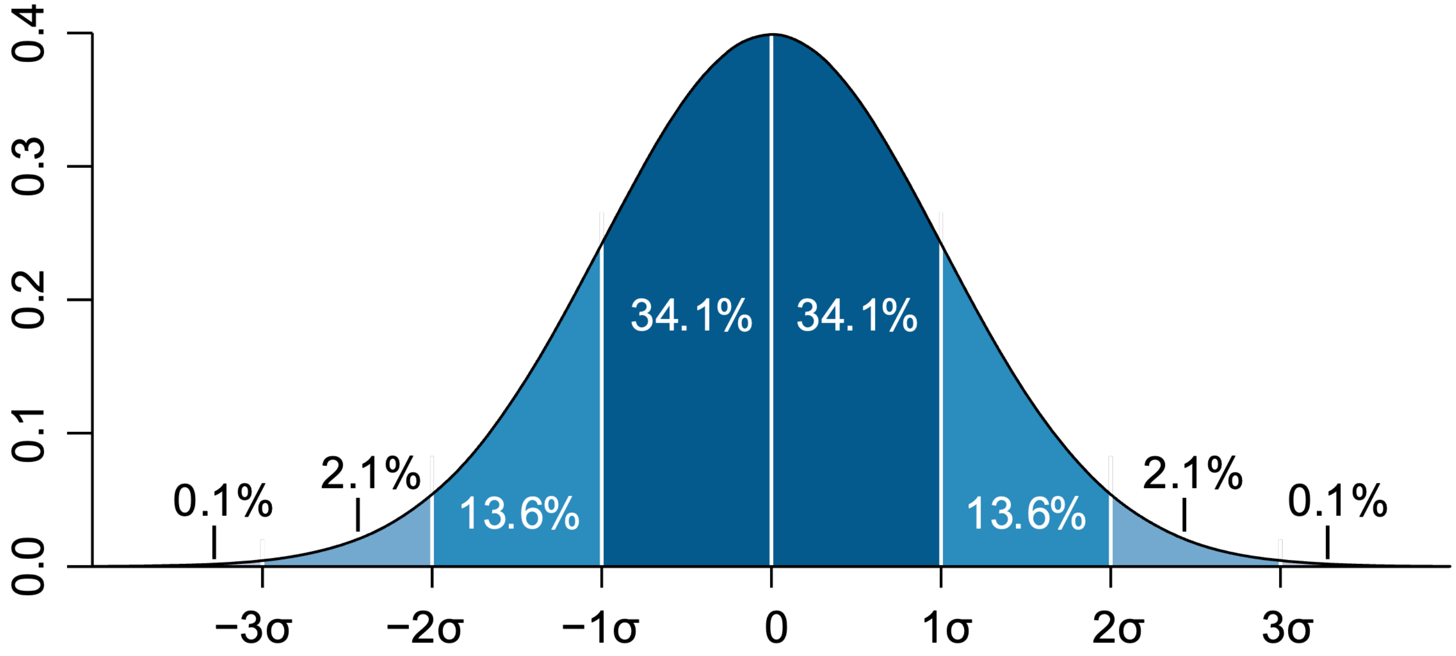


Image source: https://commons.wikimedia.org/w/index.php?curid=1903871

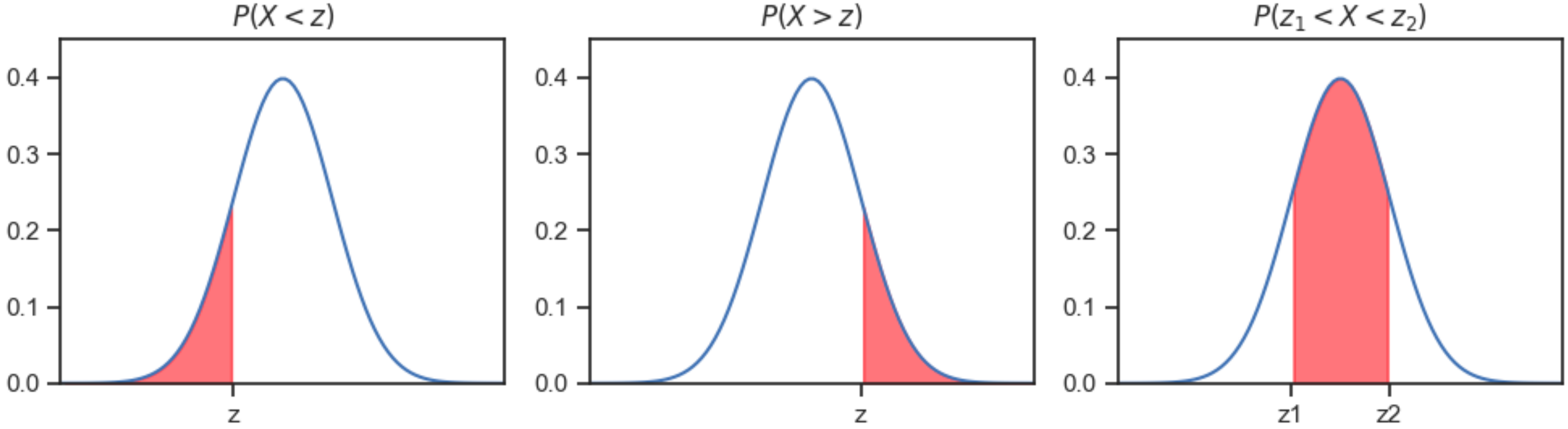
Why the normal distribution is important

- Many data follow the normal distribution. That’s why it’s widely used in business, statistics and in government bodies:
  - Heights of people.
  - Measurement errors.
  - Blood pressure.
  - Points on a test.
  - IQ scores.

Probability in Normal Distribution

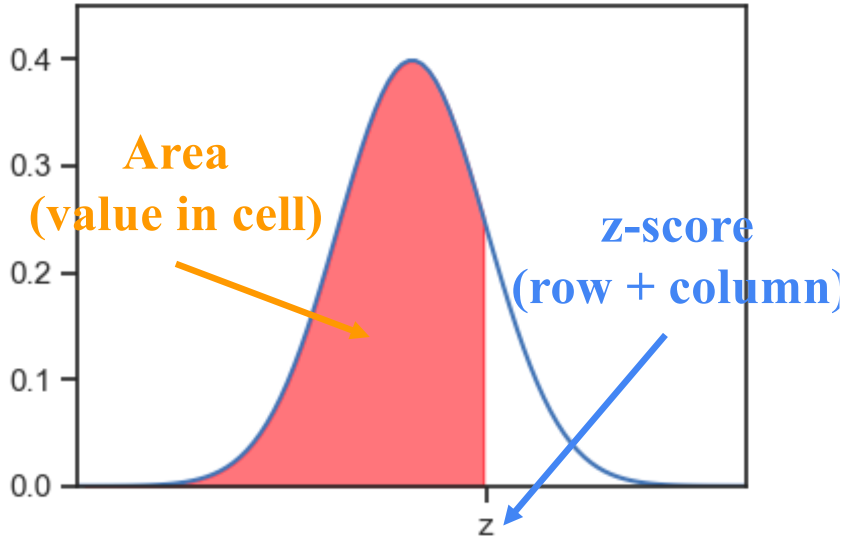
- To calculate the probability or percentage of a normal-distributed random variable that falls between a range, we can calculate the area under the bell curve between the range.
- Given $X \sim N(\mu, \sigma^2)$, the probability $P(X < z)$, $P(X > z)$ and $P(z_1 < X < z_2)$ can be calculated (*f(x)* is the PDF):

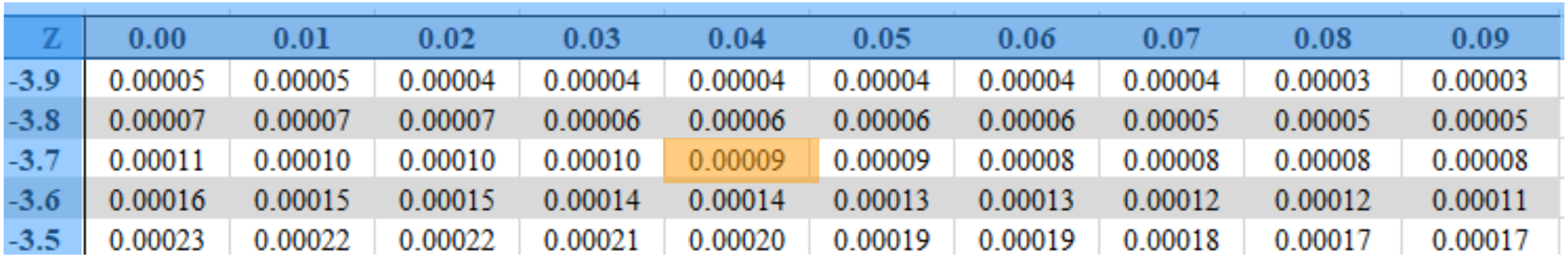
$$P(X < z) = \int_{-\infty}^{z} f(x)dx \quad P(X > z) = \int_{z}^{+\infty} f(x)dx \quad P(z_1 < X < z_2) = \int_{z_1}^{z_2} f(x)dx$$



Probability in Normal Distribution

- **Z-Table**: also known as the standard normal table, provides the area under the curve to the left of a z-score. This area represents the probability that a random variable $X \sim N(0, 1)$ falls below that z-score.
- Negative and Positive Z-table*
  - Row + Column define the z-score to the tenth's place.
  - Cells in the table represent the area under the standard normal curve to the left of the z-score.





Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017

⋮

*Only one of positive and negative table is sufficient

Probability in Normal Distribution

Example: Standard Normal Distribution

- Given $X \sim N(0,1)$, find:

(1) $P(X < -2.23)$,

(2) $P(X > -2.23)$,

(3) $P(-2.23 \ < X < 0.67)$

(4) $P(|X| > 2.23)$

Probability in Normal Distribution

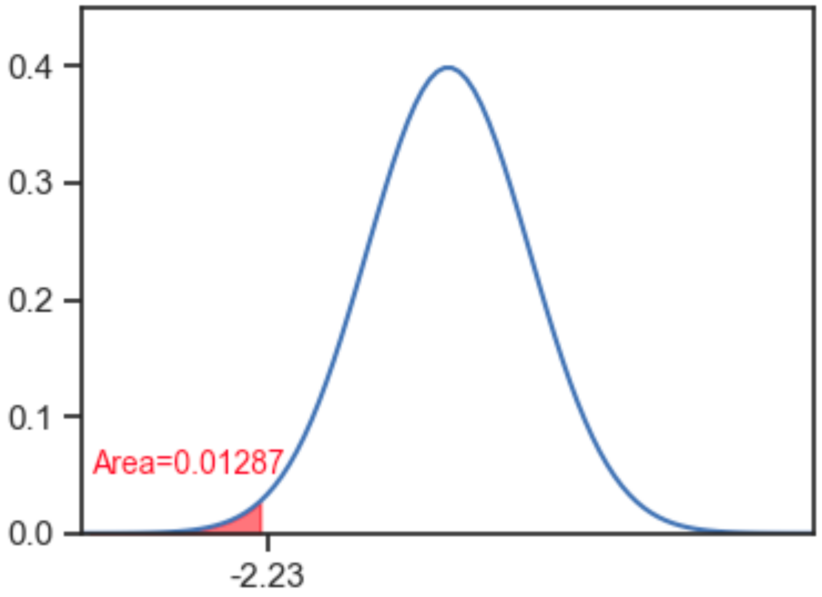
Example: Standard Normal Distribution

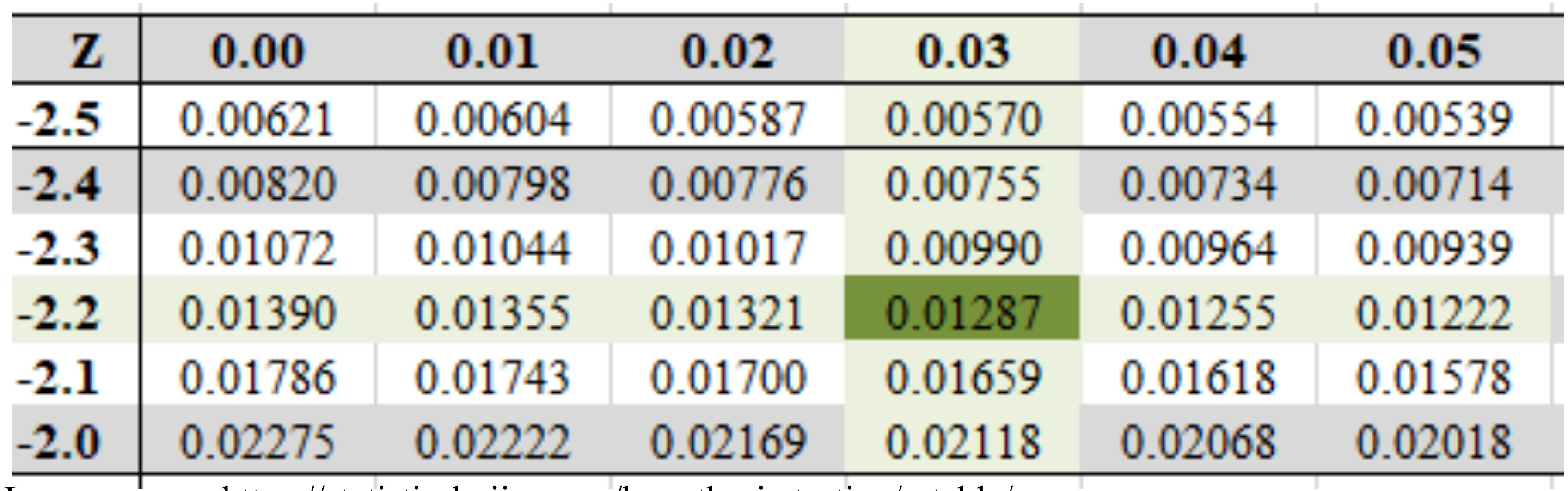
- Given $X \sim N(0,1)$, find:

  (1) $P(X < -2.23)$,

- Answer:

  (1) $P(X < -2.23) = 0.01287 = 1.287\%$





Z	0.00	0.01	0.02	0.03	0.04	0.05
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018

Image source: https://statisticsbyjim.com/hypothesis-testing/z-table/

Probability in Normal Distribution

Example: Standard Normal Distribution

- Given $X \sim N(0,1)$, find:

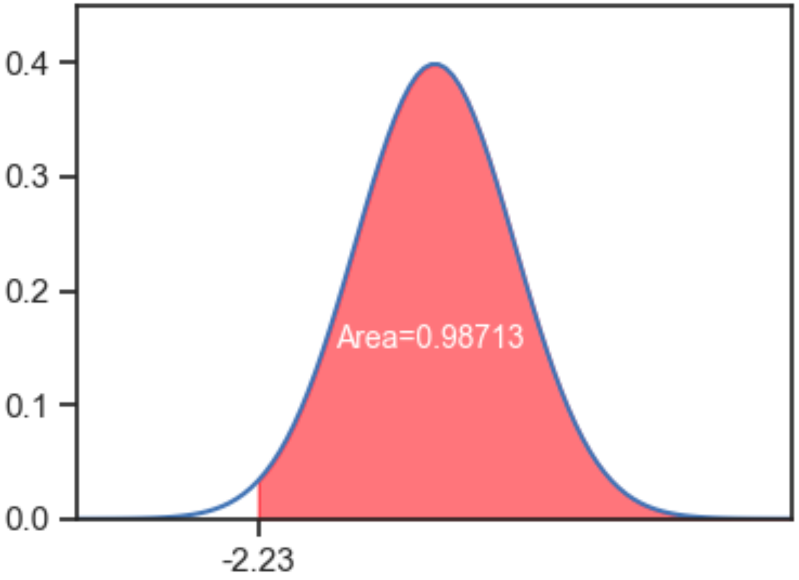
  (2) $P(X > -2.23)$,

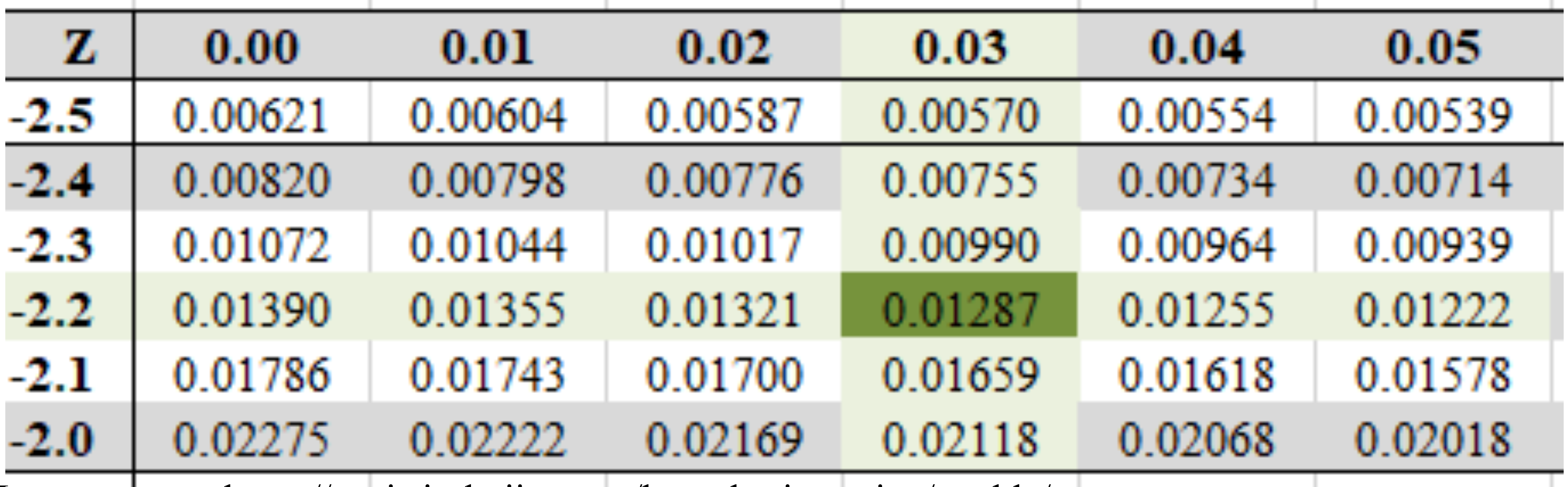
- Answer:

  (2) $P(X > -2.23) = 1 - P(X < -2.23)$

  $= 1 - 0.01287 = 0.98713$

  $= 98.713\%$





Z	0.00	0.01	0.02	0.03	0.04	0.05
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018

Image source: https://statisticsbyjim.com/hypothesis-testing/z-table/

Probability in Normal Distribution

Example: Standard Normal Distribution

- Given $X \sim N(0,1)$, find:

  (3) $P(-2.23 \ < X < 0.67)$

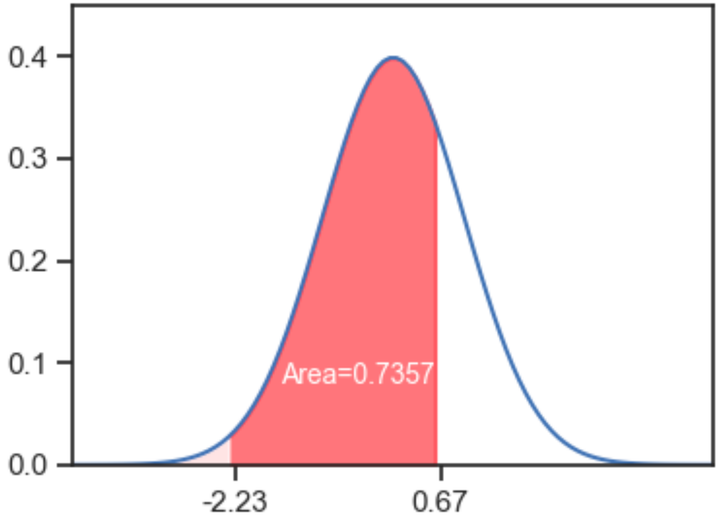
- Answer:

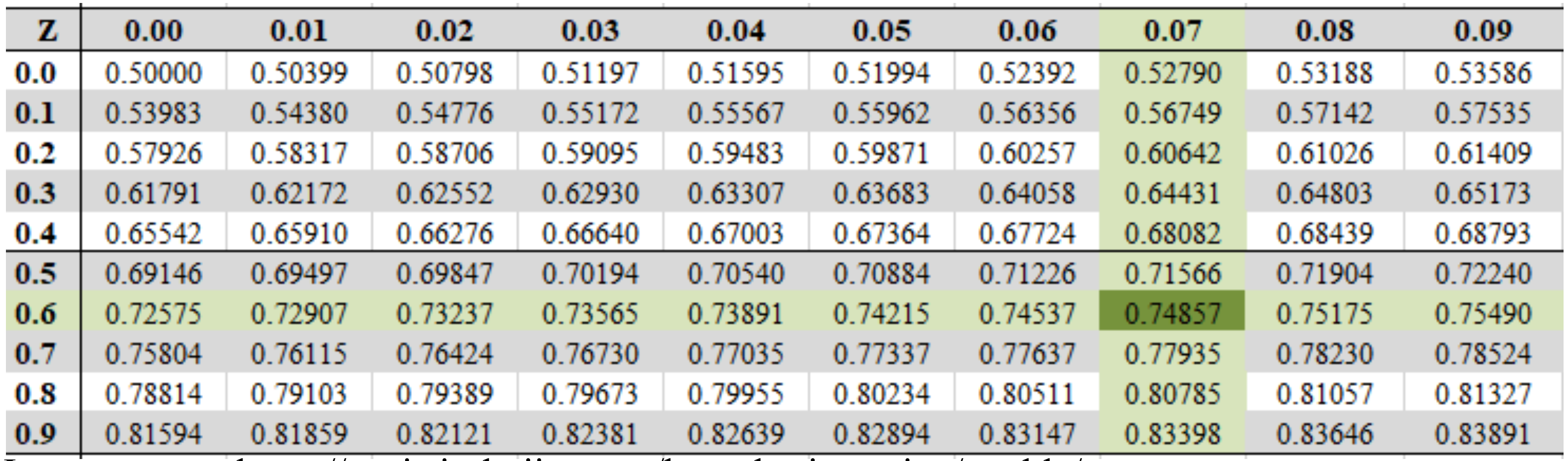
(3) $P(-2.23 < X < 0.67)$

$= P(X < 0.67) - P(X < -2.23)$

$=0.74857 - 0.01287 = 0.\ 7357$

$= 73.57\%$





Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891

Image source: https://statisticsbyjim.com/hypothesis-testing/z-table/

Probability in Normal Distribution

Example : Standard Normal Distribution

- Given $X \sim N(0,1)$, find:

  (4) $P(|X| > 2.23)$

- Answer:

(4) $P(|X| > 2.23) = P(X < -2.23) + P(X > 2.23)$

$$= P(X < -2.23) + 1 - P(X < 2.23)$$

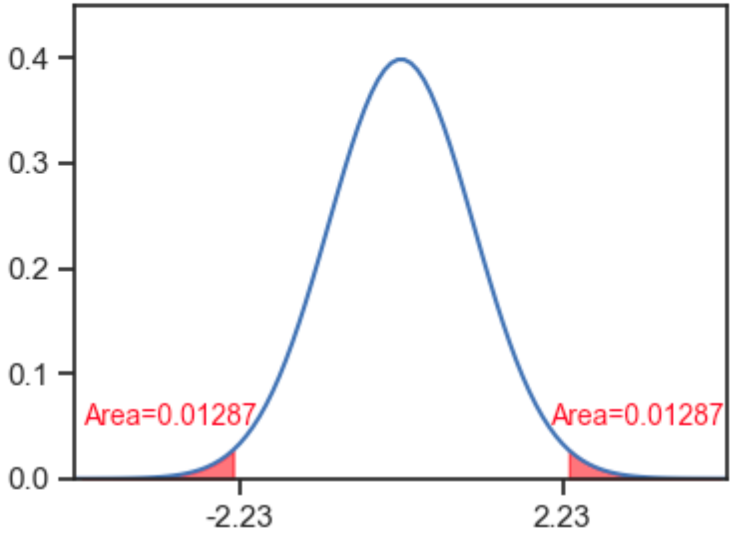
Or because we know the bell curve is symmetric about z=0

$$P(|X| > 2.23) = P(X < -2.23) + P(X > 2.23) = 2 * P(X < -2.23)$$

$$= 2 * 0.01287 = 0.02574 = 2.574\%$$

You can verify

$1 - P(X < 2.23) = P(X < -2.23)$



Probability in Normal Distribution

Example: General Normal Distribution

- Given $X \sim N(65, 9^2)$, find:

(1) $P(X < 54)$,

(2) $P(X \geq 80)$,

(3) $P(70 < X < 86)$

If $X \sim N(\mu, \sigma)$, $then\, Z = \frac{X-\mu}{\sigma} \sim N(0{,}1)$

💡 **z-score** is essentially the number of standard deviations (σ) from the mean (μ)

Probability in Normal Distribution

Example: General Normal Distribution

- Given $X \sim N(65,9)$, find:

  (1) $P(X < 54)$,

- Answer:

1. Find z-score for x = 54: $z = (x - \mu)/\sigma = (54 - 65)/9 = -1.22$
2. Look up the z-table we can get the area to the left of z=-1.22 is 0.11123.
3. $P(X < 54) = P(Z < -1.22) = 0.11123 = 11.123\%$

Probability in Normal Distribution

Exercise: General Normal Distribution

- Given $X \sim N(65,9)$, find:

(2) $P(X \geq 80)$,

Go to

www.menti.com

Enter the code

23 13 57 0

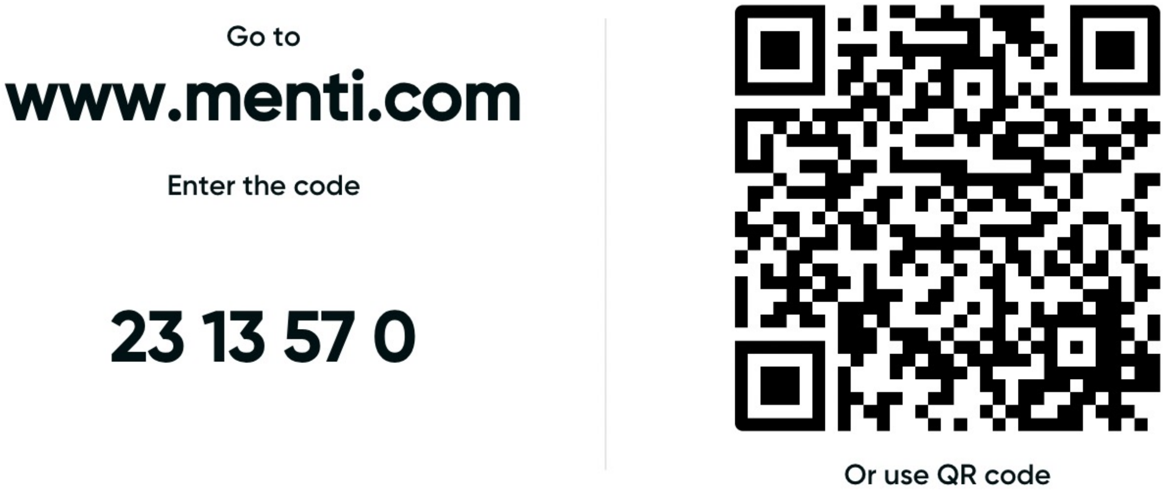
Or use QR code

Probability in Normal Distribution

Exercise: General Normal Distribution

- Given $X \sim N(65{,}9)$, find:

(3) $P(70 < X < 86)$,



Go to

www.menti.com

Enter the code

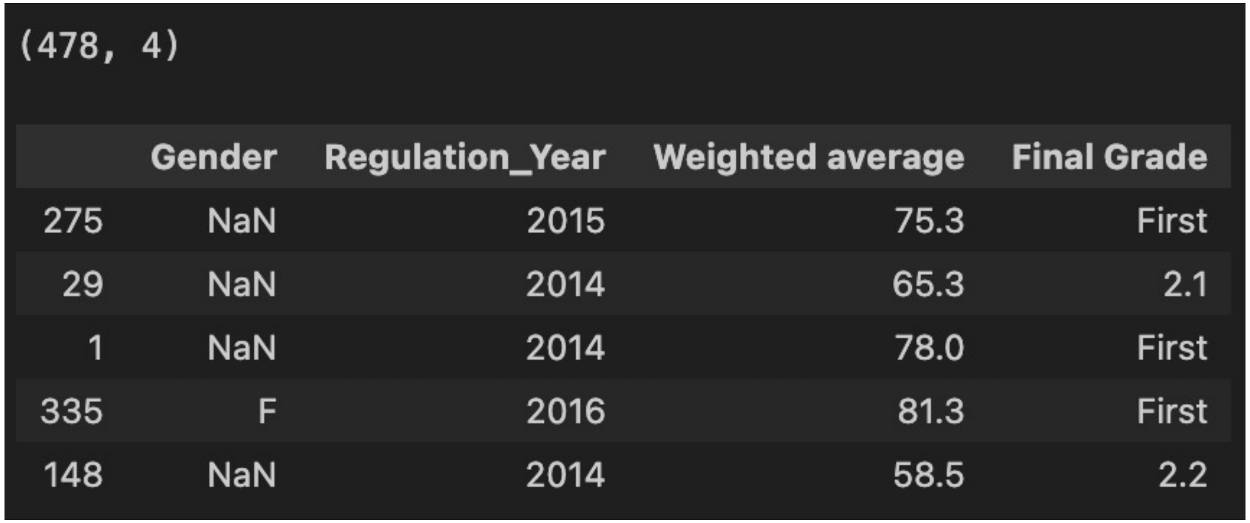
23 13 57 0

Or use QR code

Probability in Normal Distribution

Case Study: Student Degree Results

- We’ll work with real-life data that include the weighted average marks (UK scale) and final grades (degree classifications) of students who graduated between 2018 - 2020 (2014 - 2016 cohort) from an engineering programme.
- The shape of the data and 5 samples from the data are shown below.
- We'll focus on the *weighted average* marks as it is numerical data and the final grades of students are determined by their weighted average marks.



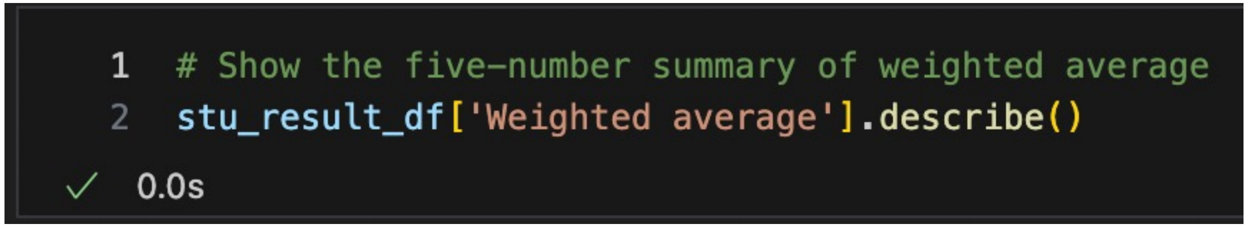
```
(478, 4)
```

	Gender	Regulation_Year	Weighted average	Final Grade
275	NaN	2015	75.3	First
29	NaN	2014	65.3	2.1
1	NaN	2014	78.0	First
335	F	2016	81.3	First
148	NaN	2014	58.5	2.2

Probability in Normal Distribution

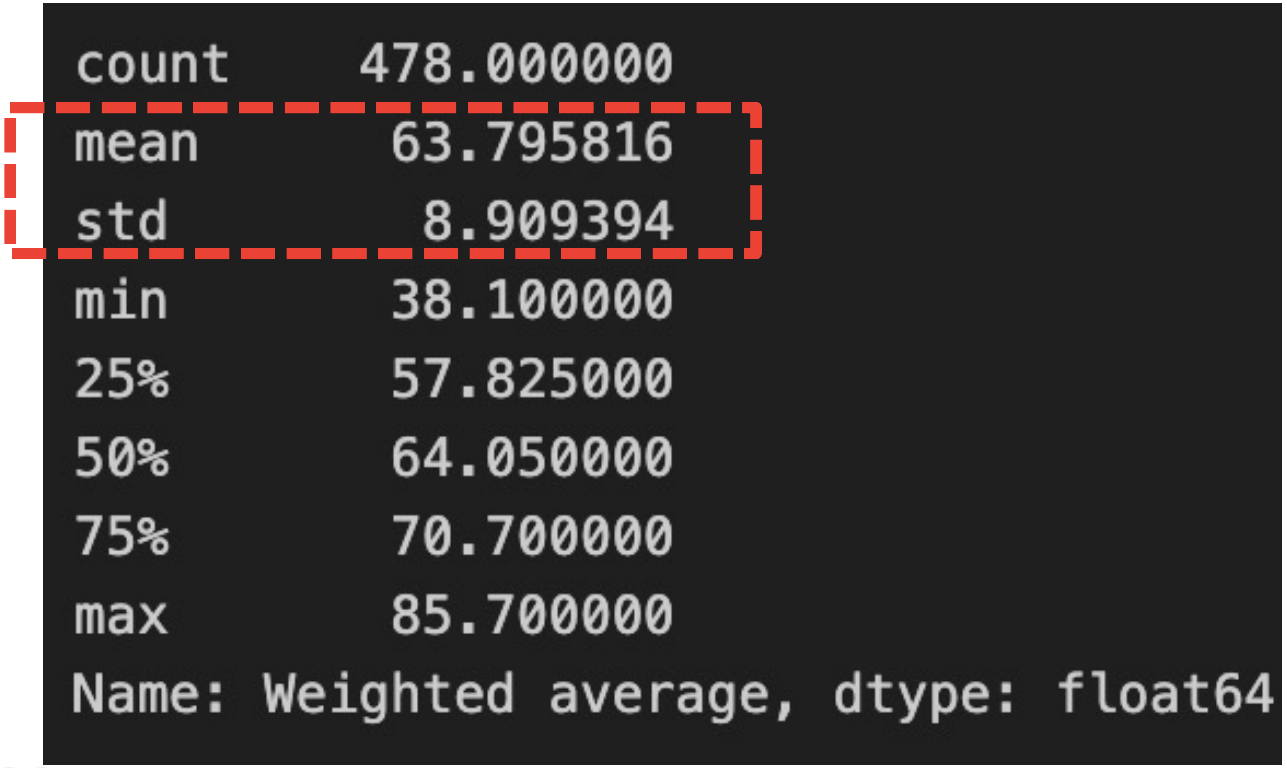
Case Study: Student Degree Results

- the five-number summary of the *Weighted average* data.



```
# Show the five-number summary of weighted average
stu_result_df['Weighted average'].describe()
```

✓ 0.0s



```
count    478.000000
mean      63.795816
std        8.909394
min       38.100000
25%       57.825000
50%       64.050000
75%       70.700000
max       85.700000
Name: Weighted average, dtype: float64
```

Probability in Normal Distribution

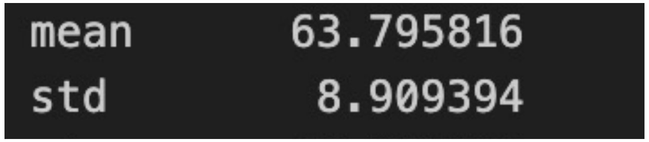
Case Study: Student Degree Results

- The *Weighted average* data closely resembles the normal distribution we produced based on the mean and standard deviation, we can use the theoretical model to approximate the actual distribution.

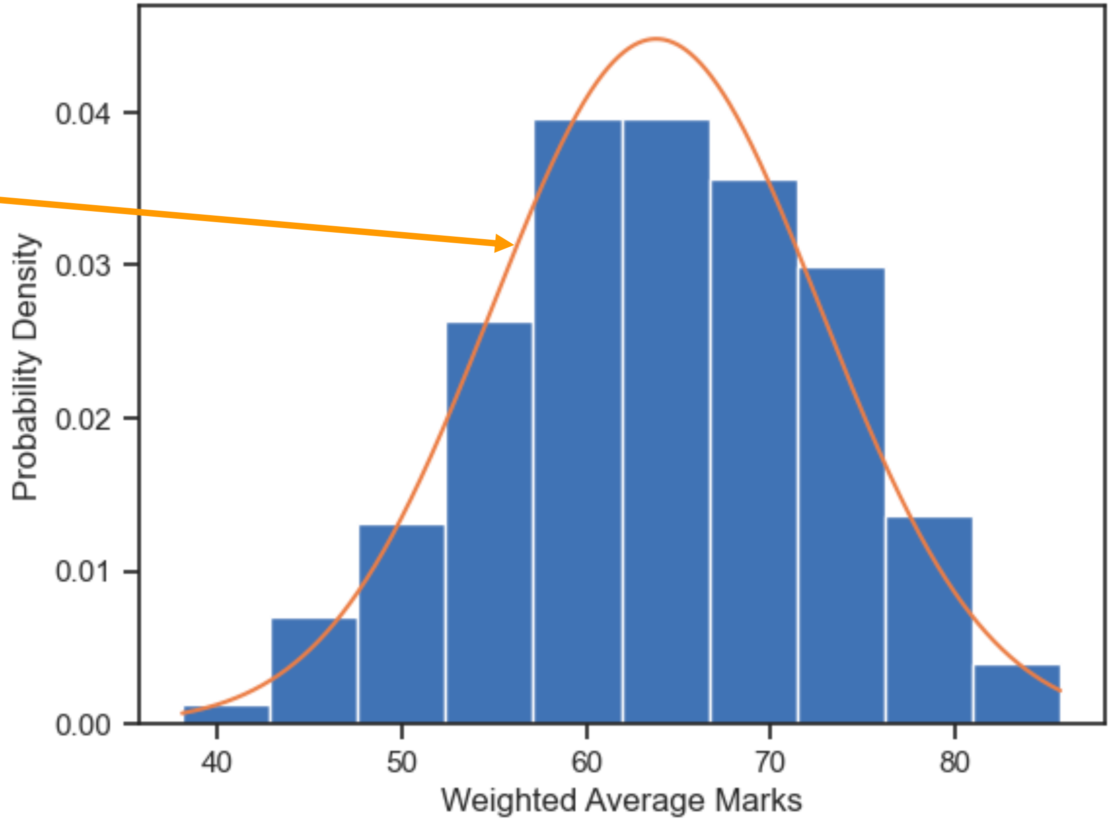
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$$\mu = 63.795816$$

$$\sigma = 8.909394$$



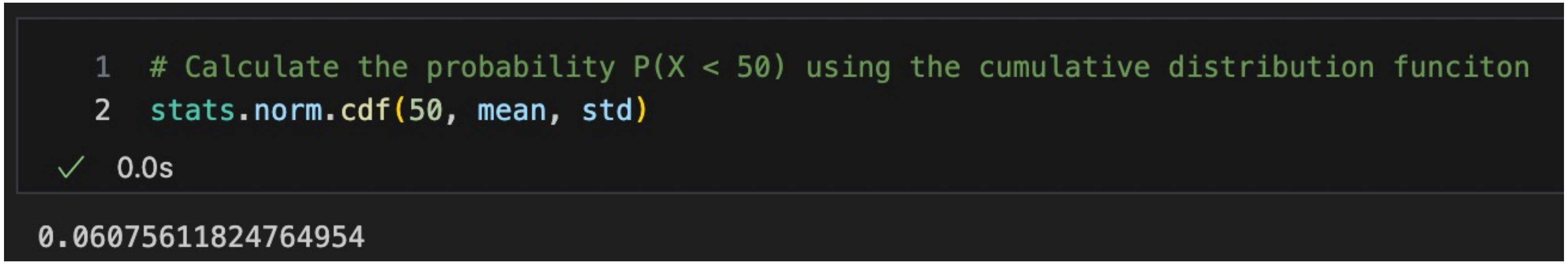
```
mean      63.795816
std        8.909394
```



Probability in Normal Distribution

Case Study: Student Degree Results

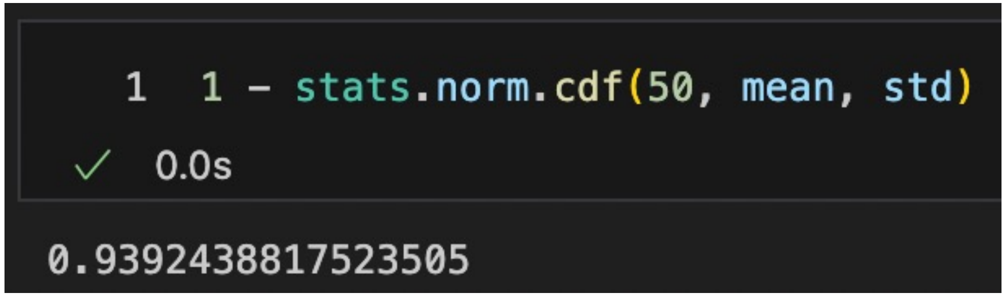
1. What percent of students had weighted average marks below 50?



```
# Calculate the probability P(X < 50) using the cumulative distribution funciton
stats.norm.cdf(50, mean, std)
```

```
✓ 0.0s

0.06075611824764954
```

2. What percent of students achieved weighted average marks above 50?



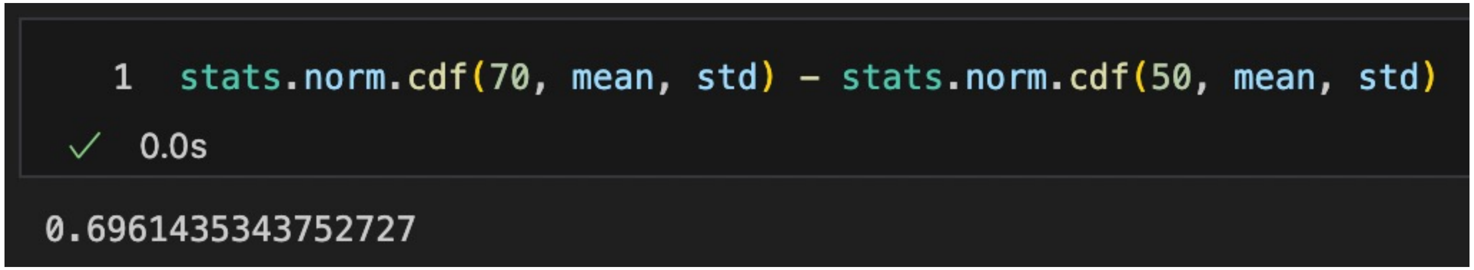
```
1 - stats.norm.cdf(50, mean, std)
```

```
✓ 0.0s

0.9392438817523505
```

Probability in Normal Distribution

Case Study: Student Degree Results

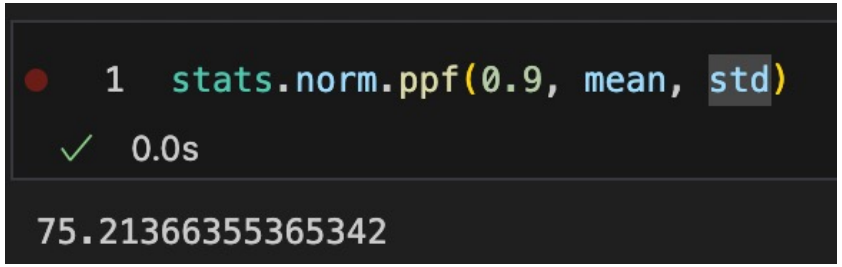
3. What percent of students achieved weighted average marks between 50 - 70?



```
stats.norm.cdf(70, mean, std) - stats.norm.cdf(50, mean, std)
✓ 0.0s

0.6961435343752727
```

4. What mark should a student achieve if he/she wants to be ranked at least the top 10% of all students in this programme?

To get the mark that 10% of students' marks are above (90% of students' marks are below) using Python (SciPy package), we can use the percent point function (ppf).

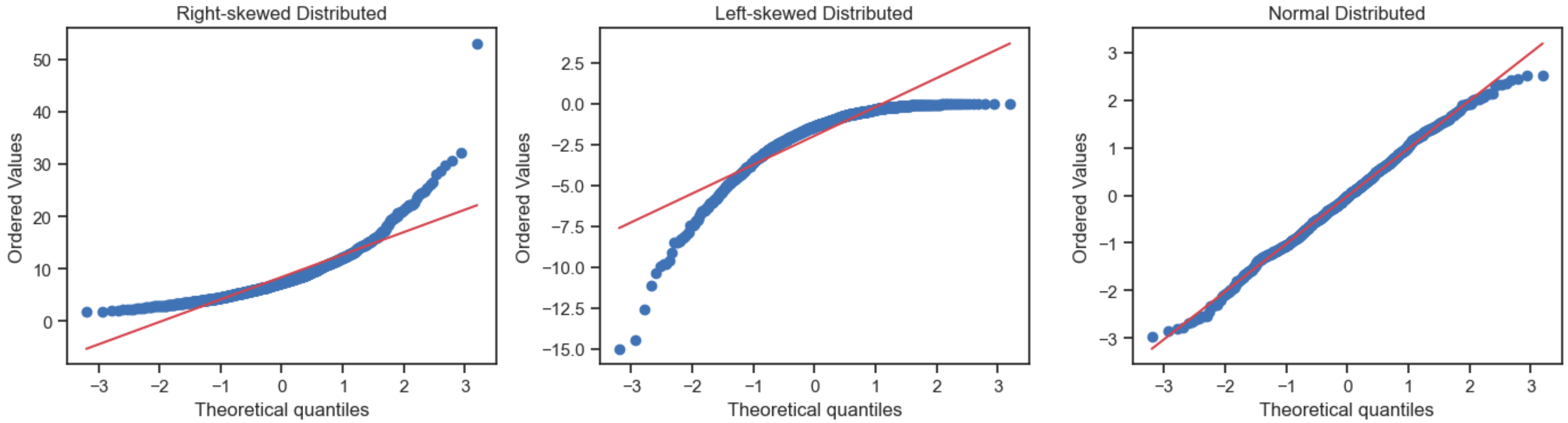


```
stats.norm.ppf(0.9, mean, std)
✓ 0.0s

75.21366355365342
```

***Discussion**: How to get the result by referring to the z-table?*

Q-Q Plot

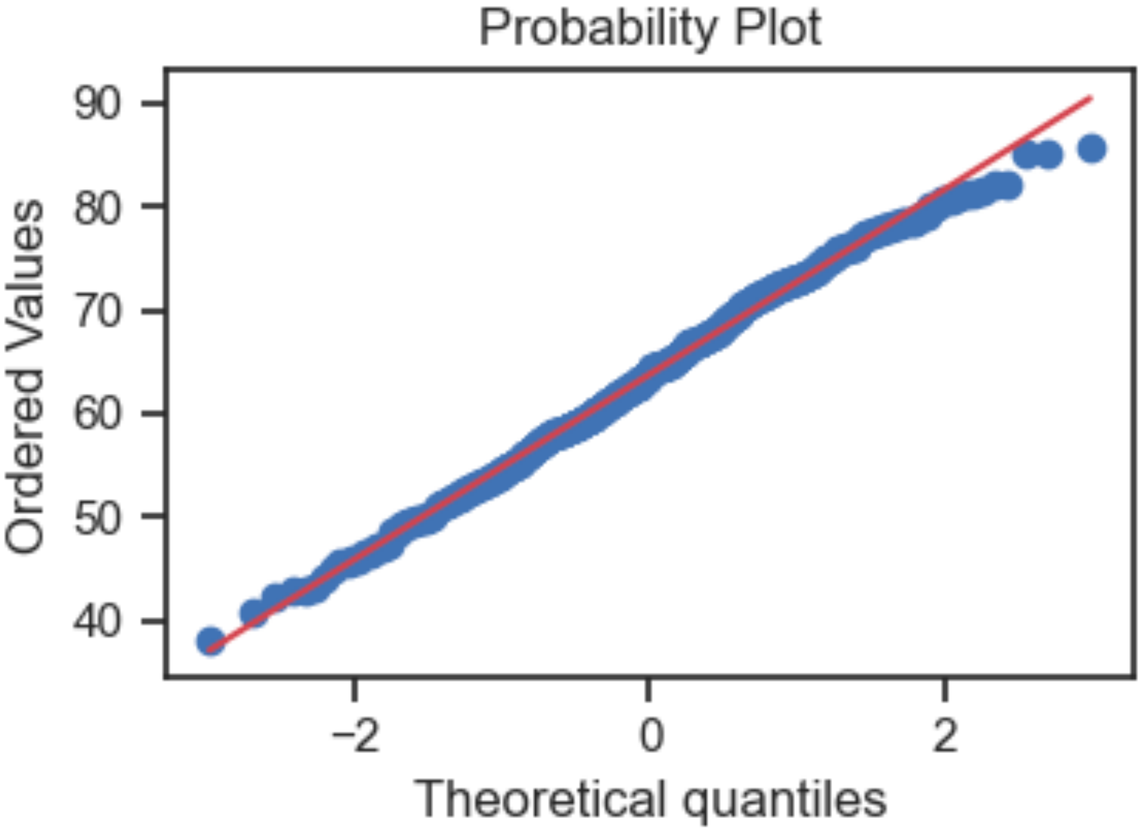
- A **Q–Q plot** (**quantile-quantile plot**) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.
  - orders the z-scores of the data from low to high and plots each value's z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value's rank. (can be normalised to the number of standard deviations away from the mean)
  - If the points roughly fall on the diagonal line (y = x), then the sample distribution can be considered close to the normal distribution



Probability in Normal Distribution

Case Study: Student Degree Results

5. Use a Q-Q plot to check whether the `Weighted average` follows the normal distribution.



DESCRIPTIVE VS. INFERENTIAL STATISTICS

Descriptive vs. Inferential Statistics

- **Descriptive Statistics** is about describing the collected data.
- **Inferential Statistics** allow you to draw conclusions about a population based on the sample data collected from that population.
- Terminologies
  - **Population**: The entire group of interest.
  - **Sample**: A subset of the population.
  - **Statistic**: Numeric summary about a sample.
  - **Parameter**: Numeric summary about a population.

Descriptive vs. Inferential Statistics

Example:

- Imagine I want to know what proportion of JP students play League of Legends. I sent a survey to all currently enrolled JP students asking the question: Do you play League of Legends?
  - 2700 JP students currently enrolled
  - 600 students responded to the survey
  - 60% of the students who responded to my survey, say they do play League of Legends.

Population: 2700 JP students.
Sample: 600 students who responded to the survey.
Statistic: 60%.
Parameter: the proportion of all 2700 students who play League of Legends.

Descriptive vs. Inferential Statistics

Exercise:

- Consider we are interested in the average sleeping time of all currently enrolled JP students (2700 students). 500 survey responses were received. The average amount of sleeping time of those that responded was 6.8 hours.

Population: all currently enrolled JP students (2700 students).

Sample: 500 students who responded to the survey.

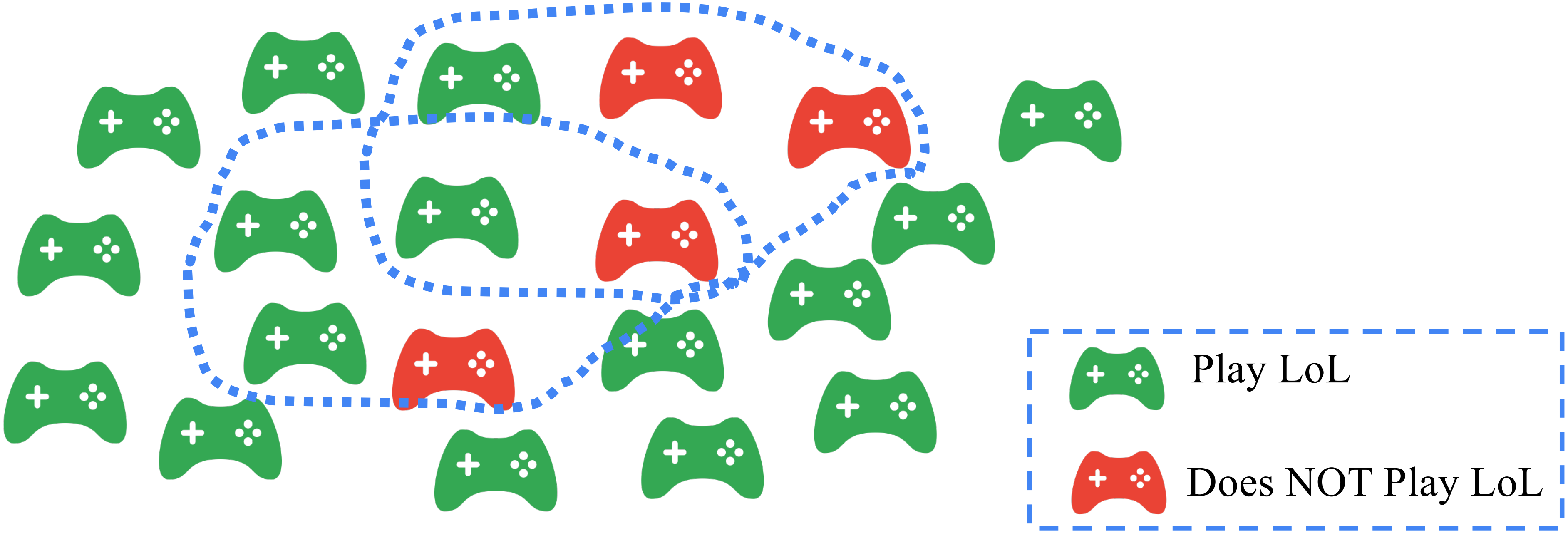
Statistic: 6.8 hours.

Parameter: We cannot know the number for sure.

SAMPLING DISTRIBUTIONS

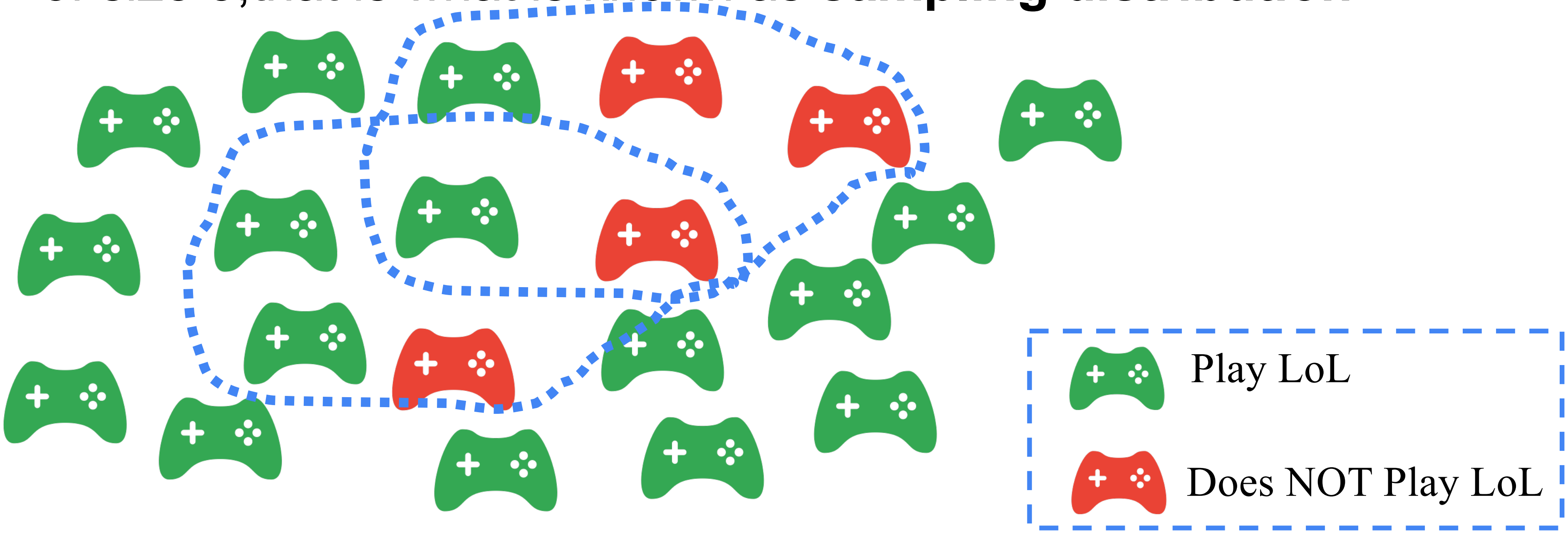
Sampling Distribution

- Consider again the game-playing habit of all JP students.
  - In the picture below, each game controller represents a JP student (20 students in total).
  - Randomly selected the students within the blue circle to ask them whether they play LoL or not
  - What would be the population, parameter, sample and statistic?



Sampling Distribution

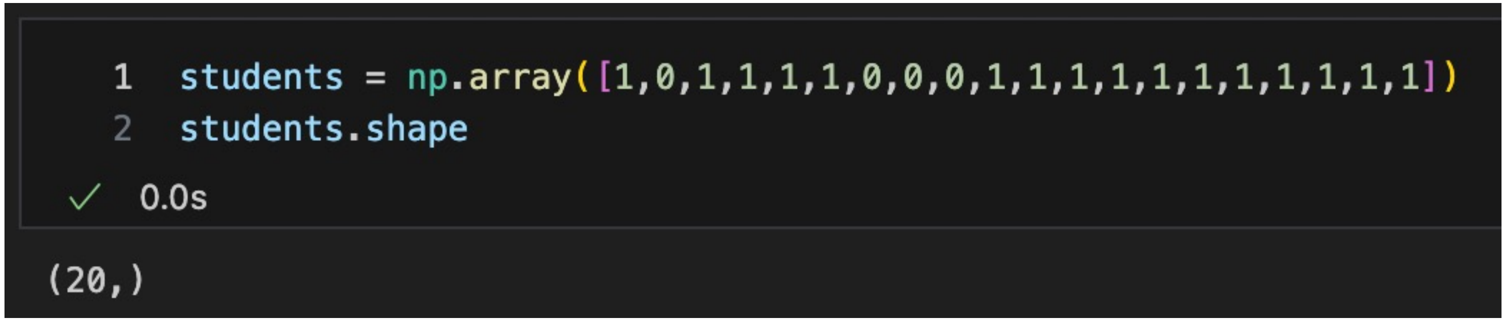
- Consider again the game-playing habit of all JP students.
  - Though our sample is still 5 students, our statistic changed.
  - We could select all possible combinations of 5 students, and re-compute the proportion of students who play LoL for each of these sample.
  - if we look at the **distribution** of the proportions across all samples of size 5,that is what is known as **sampling distribution**



Sampling Distribution

Demo

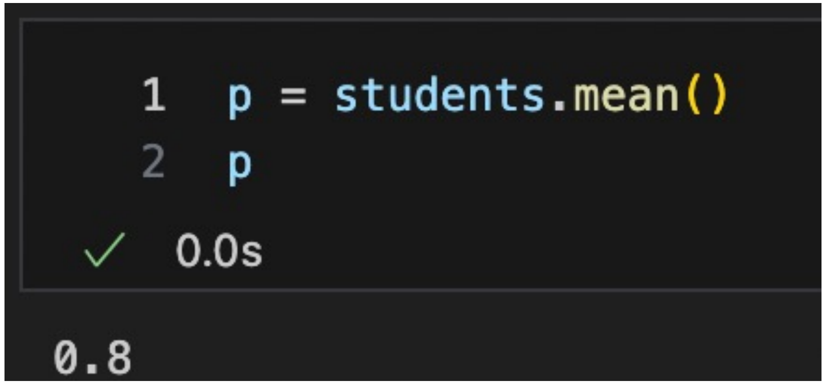
- Below is an array that represents the students' game-playing habit example in this section, where 1 represents the students who play LoL, and 0 represents the students who do not play LoL.



```
students = np.array([1,0,1,1,1,1,0,0,0,1,1,1,1,1,1,1,1,1,1,1])
students.shape
```

```
✓ 0.0s

(20,)
```

1. Find the proportion of students who play LoL in the above array. Store this value in a variable p.



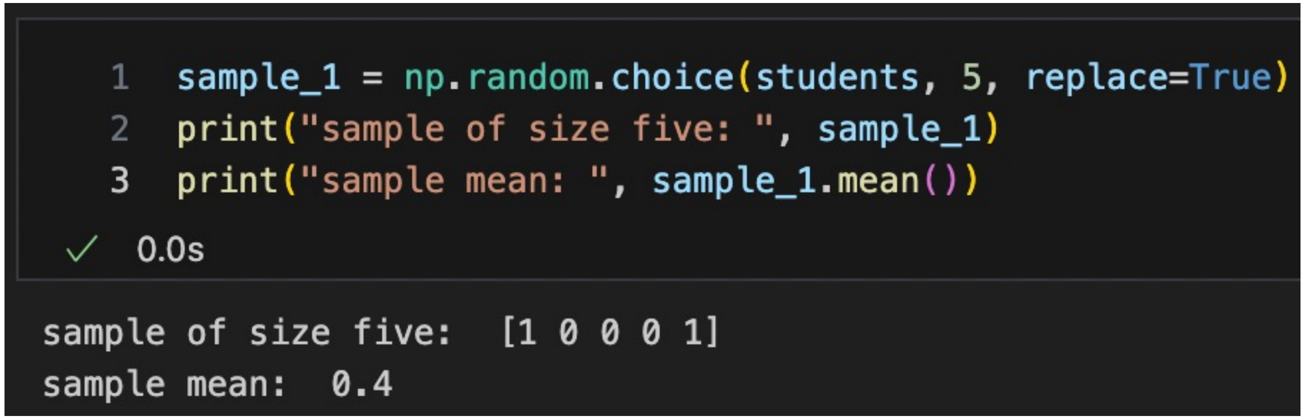
```
p = students.mean()
p
```

```
✓ 0.0s

0.8
```

Sampling Distribution

Demo

2. Use numpy's random.choice() to simulate 5 draws from the *students* array. What is the proportion of students who play LoL in your sample?



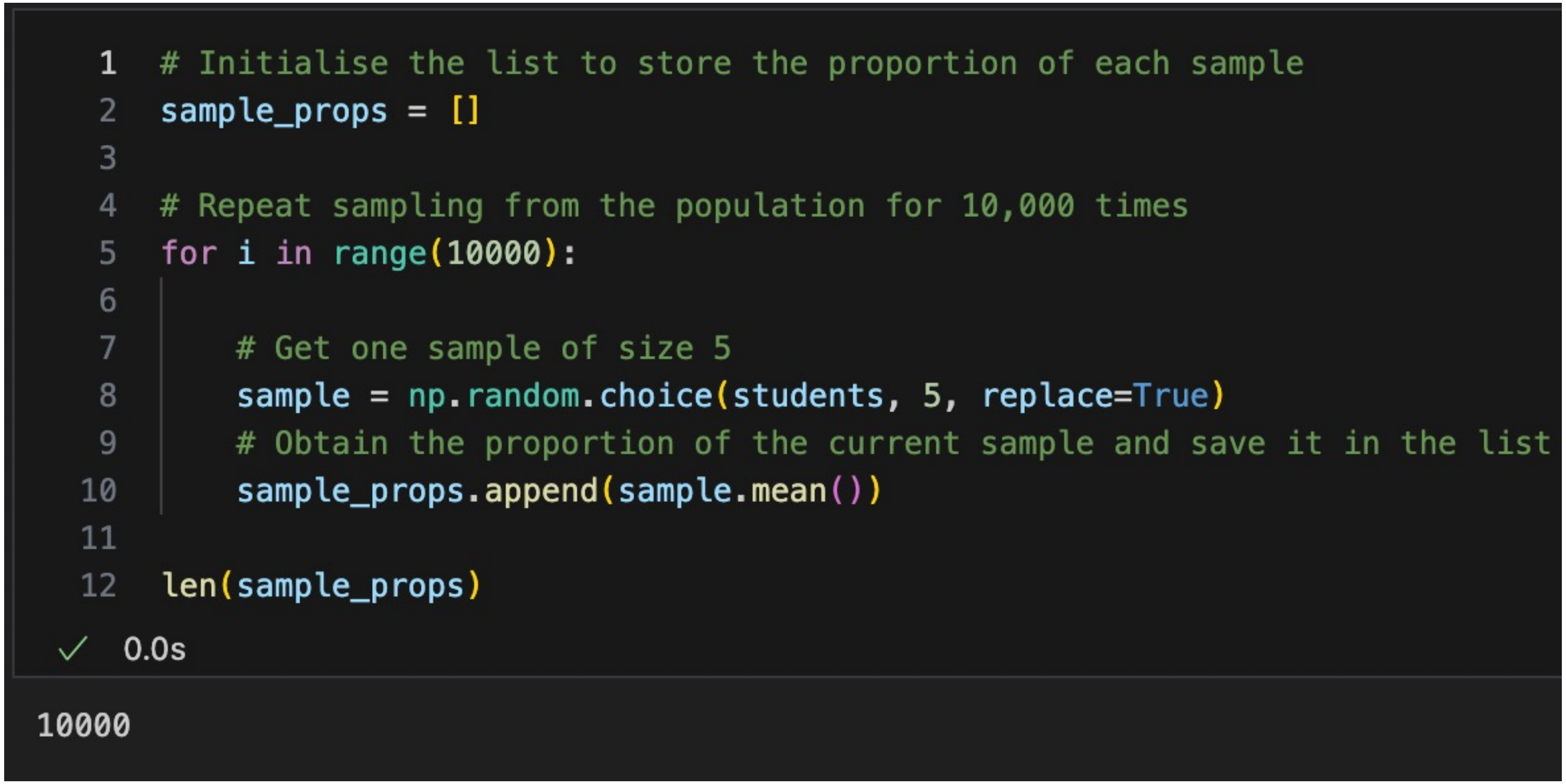
```
sample_1 = np.random.choice(students, 5, replace=True)
print("sample of size five: ", sample_1)
print("sample mean: ", sample_1.mean())
```

```
sample of size five:  [1 0 0 0 1]
sample mean:  0.4
```

Sampling Distribution

Demo

3. Repeat step 2 to obtain 10,000 proportions, where each sample is of size 5. Store these in a list called `sample_props`?



```
# Initialise the list to store the proportion of each sample
sample_props = []

# Repeat sampling from the population for 10,000 times
for i in range(10000):

    # Get one sample of size 5
    sample = np.random.choice(students, 5, replace=True)
    # Obtain the proportion of the current sample and save it in the list
    sample_props.append(sample.mean())

len(sample_props)
```

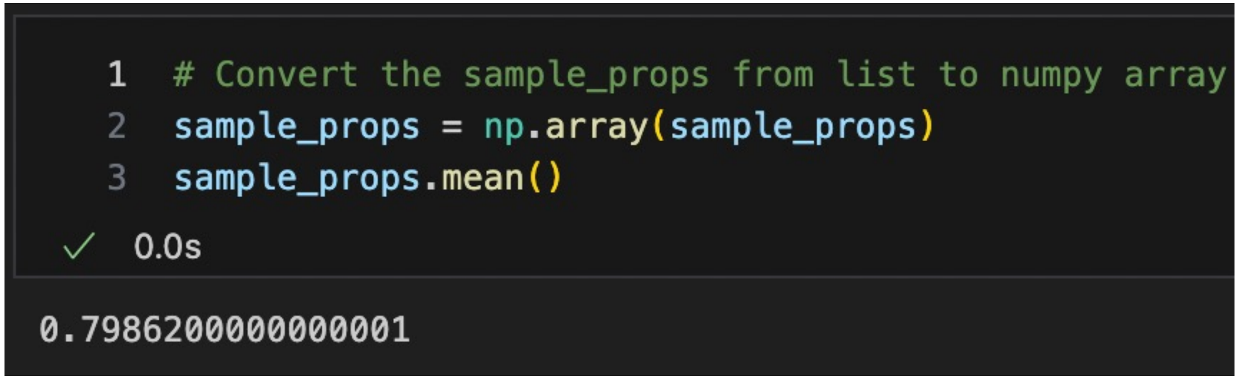
✓ 0.0s

```
10000
```

Sampling Distribution

Demo

4. What is the mean proportion of all 10,000 of these proportions? This is often called the mean of the sampling distribution.



```python
# Convert the sample_props from list to numpy array
sample_props = np.array(sample_props)
sample_props.mean()
```

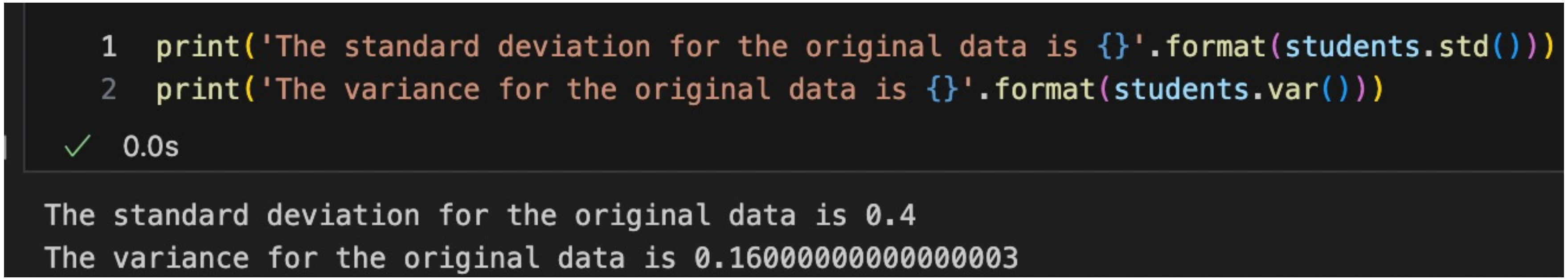
✓ 0.0s

```
0.7986200000000001
```

Sampling Distribution

Demo

5. What are the variance and standard deviation for the original 20 data values?

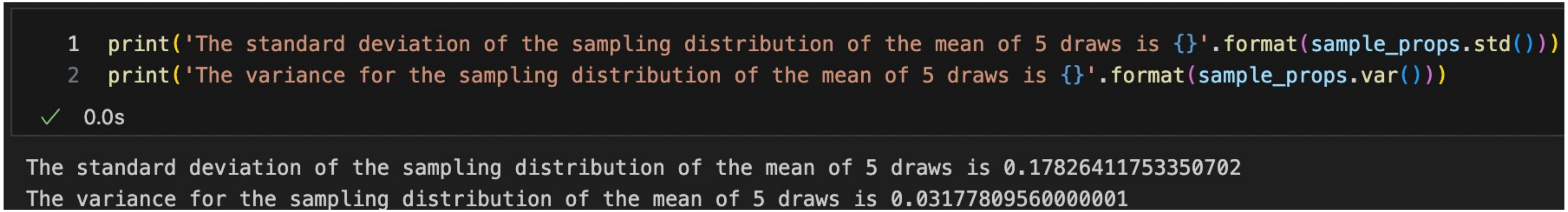


```
print('The standard deviation for the original data is {}'.format(students.std()))
print('The variance for the original data is {}'.format(students.var()))
```

✓ 0.0s

```
The standard deviation for the original data is 0.4
The variance for the original data is 0.16000000000000003
```

6. What are the variance and standard deviation for the 10,000 proportions obtained from 10,000 samples?



```
print('The standard deviation of the sampling distribution of the mean of 5 draws is {}'.format(sample_props.std()))
print('The variance for the sampling distribution of the mean of 5 draws is {}'.format(sample_props.var()))
```

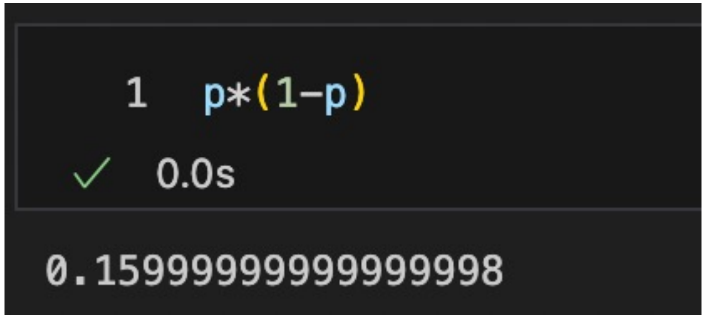
✓ 0.0s

```
The standard deviation of the sampling distribution of the mean of 5 draws is 0.17826411753350702
The variance for the sampling distribution of the mean of 5 draws is 0.031778095600000001
```

Sampling Distribution

Demo

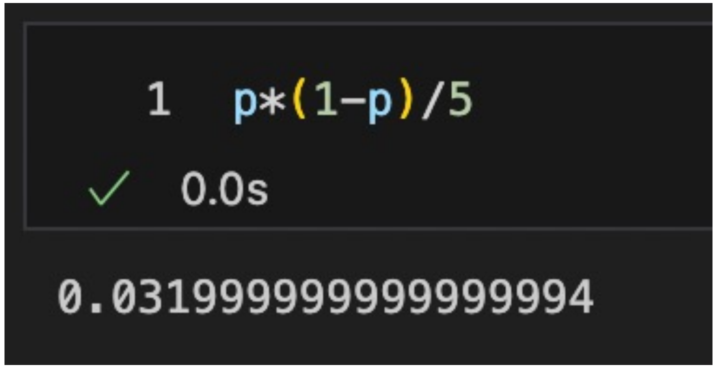
7. Compute p(1-p), which of your answers above does this most closely match?



```
p*(1-p)
```

```
✓ 0.0s

0.15999999999999998
```

8. Compute *p(1-p)/n*, which of your answers above does this most closely match?



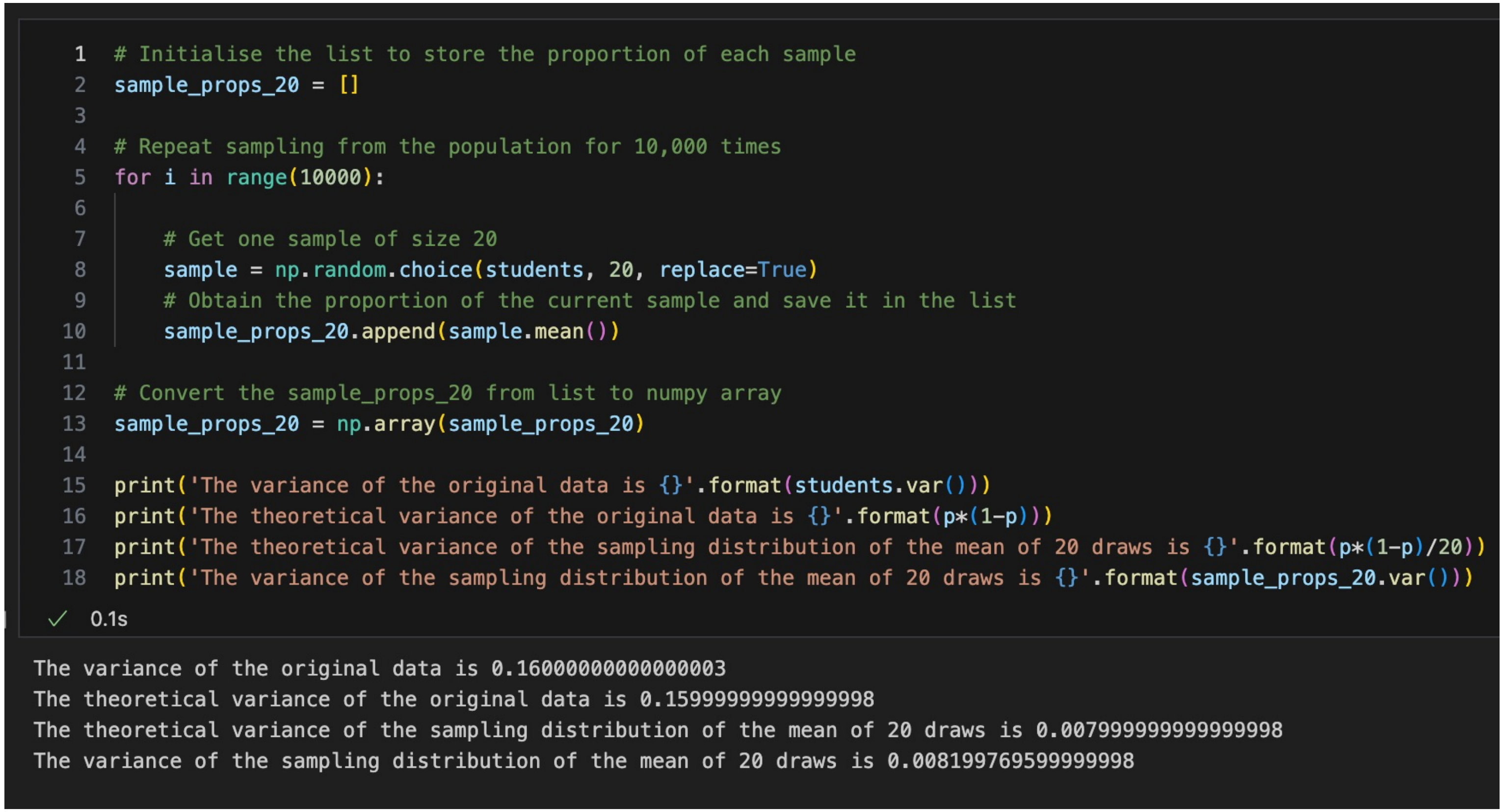
```
p*(1-p)/5
```

```
✓ 0.0s

0.031999999999999994
```

Sampling Distribution

Demo

9. If we change our sample size to be 20, what would this affect the variance of the sampling distribution? Simulate and calculate the new results in `6.` and `8.` to check that the consistency you found before still holds.



```python
# Initialise the list to store the proportion of each sample
sample_props_20 = []

# Repeat sampling from the population for 10,000 times
for i in range(10000):

    # Get one sample of size 20
    sample = np.random.choice(students, 20, replace=True)
    # Obtain the proportion of the current sample and save it in the list
    sample_props_20.append(sample.mean())

# Convert the sample_props_20 from list to numpy array
sample_props_20 = np.array(sample_props_20)

print('The variance of the original data is {}'.format(students.var()))
print('The theoretical variance of the original data is {}'.format(p*(1-p)))
print('The theoretical variance of the sampling distribution of the mean of 20 draws is {}'.format(p*(1-p)/20))
print('The variance of the sampling distribution of the mean of 20 draws is {}'.format(sample_props_20.var()))
```

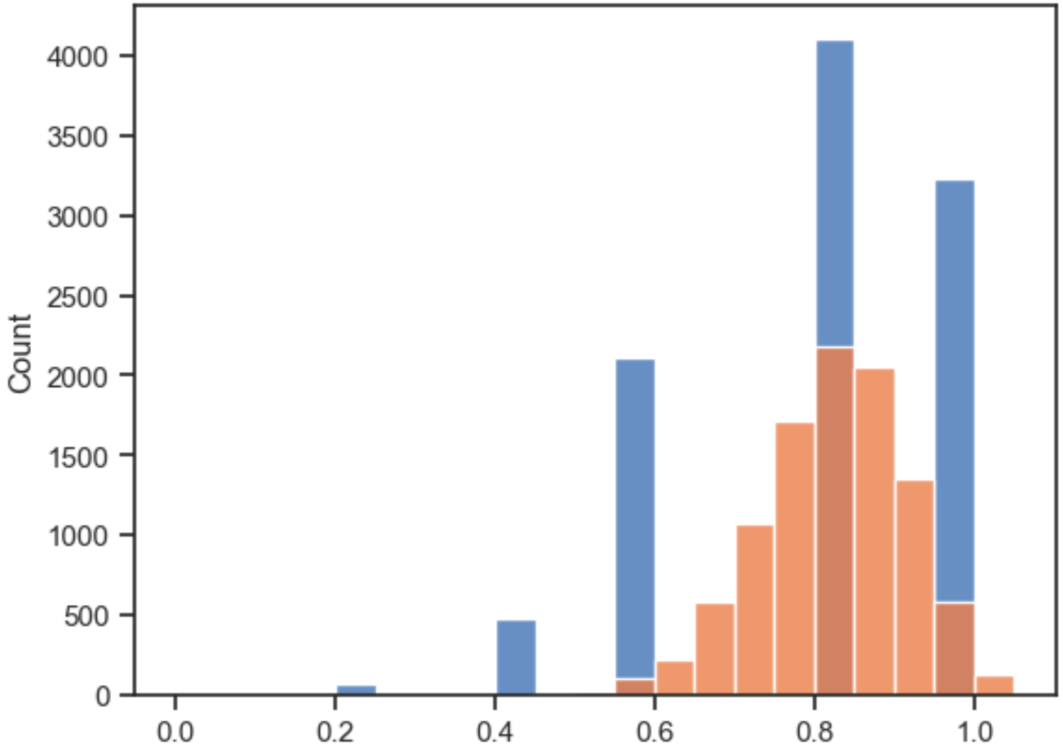
✓ 0.1s

```
The variance of the original data is 0.16000000000000003
The theoretical variance of the original data is 0.15999999999999998
The theoretical variance of the sampling distribution of the mean of 20 draws is 0.007999999999999998
The variance of the sampling distribution of the mean of 20 draws is 0.008199769599999998
```

Sampling Distribution

Demo

10. Finally, plot a histgram of the 10,000 draws from both the proportions with a sample size of 5 and the proportions with a sample size of 20. Each of these distributions is a sampling distribution. One is for the proportions of sample size 5 and the other a sampling distribution for proportions with sample size 20.



***Discussion**: What can you find from the two histograms?*

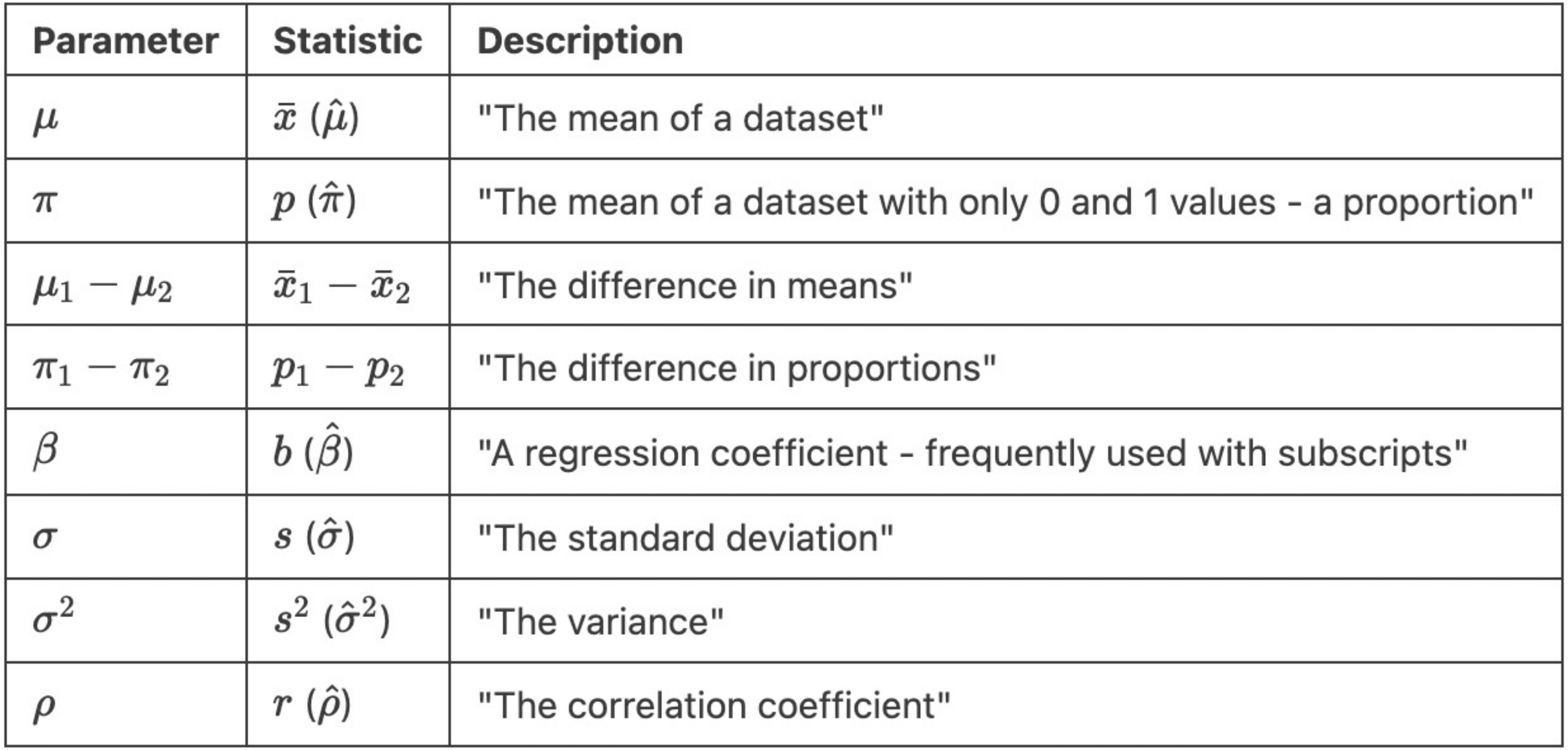
Sampling Distribution

Demo: Summary

- We have defined sampling distributions as the distribution of a statistic. We simulated the creation of sampling distributions in the demo for samples of size 5 and size 20.
- For proportions (and also means, as proportions are just the mean of 1 and 0 values), the following characteristics hold:
  - The sampling distribution is centred on the original parameter value.
  - The sampling distribution decreases its variance depending on the sample size used. Specifically, the variance of the sampling distribution is equal to the variance of the original data (population) divided by the sample size used. This is always true for the variance of a sample mean!
- **If we have a random variable, X, with a variance of σ^2, then the distribution of $\overline{x}$ (the sampling distribution of the sample mean) has a variance of σ^2/n.**

Sampling Distribution

- Most common parameters and corresponding statistics.



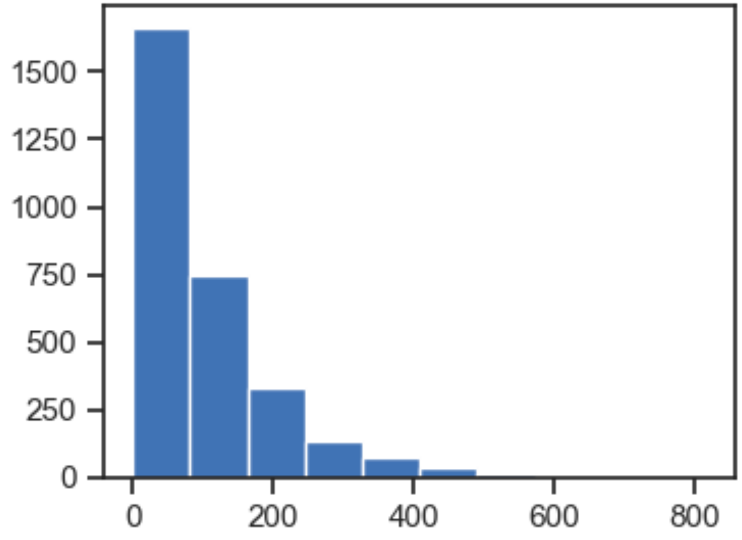
Parameter	Statistic	Description
μ	$\bar{x}$ $(\hat{\mu})$	"The mean of a dataset"
π	p $(\hat{\pi})$	"The mean of a dataset with only 0 and 1 values - a proportion"
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	"The difference in means"
$\pi_1 - \pi_2$	$p_1 - p_2$	"The difference in proportions"
β	b $(\hat{\beta})$	"A regression coefficient - frequently used with subscripts"
σ	s $(\hat{\sigma})$	"The standard deviation"
σ^2	s^2 $(\hat{\sigma}^2)$	"The variance"
ρ	r $(\hat{\rho})$	"The correlation coefficient"

TWO USEFUL THEOREMS

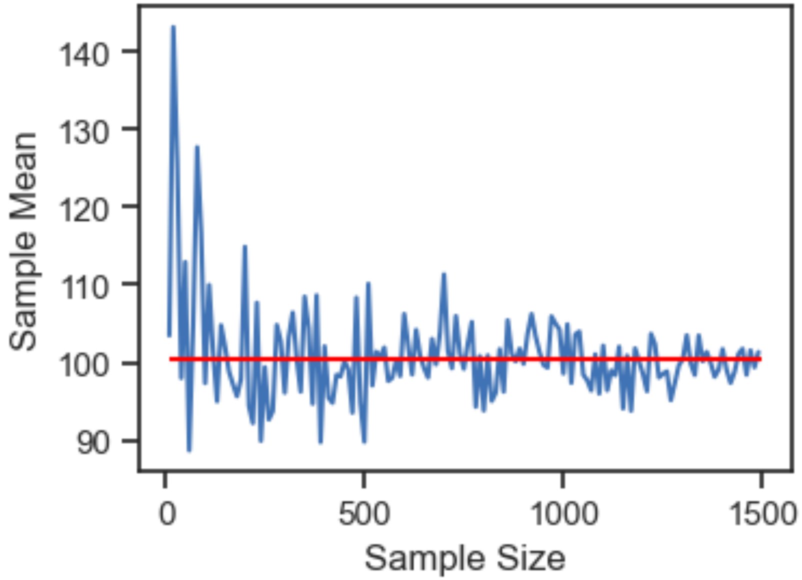
Law of Large Numbers

- The **Law of Large Numbers** says that as the sample size increases, the sample mean gets closer to the population's mean.

A dataset contains 3000 values following the log normal distribution (right-skewed) as the population dataset



draw samples from the population with different sample size, ranging from 10 to 1500 with a step of 10

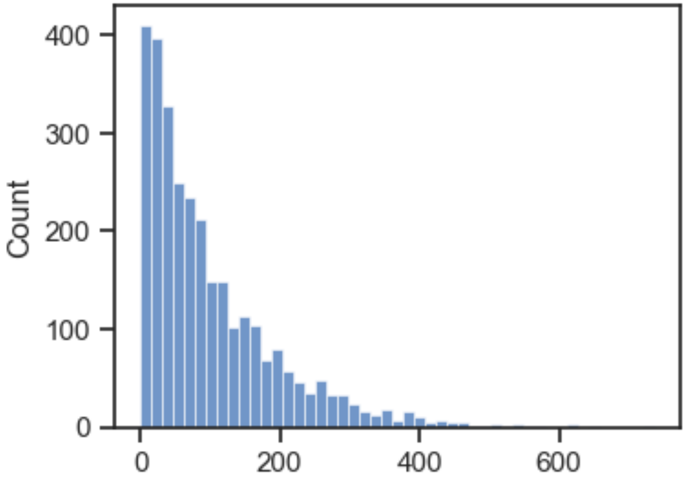


Central Limit Theorem

- The **Central Limit Theorem** states that: for independent and identically distributed (i.i.d.) random variables, the sampling distribution of the sample mean will be normally distributed given a large enough sample size.

Demo:

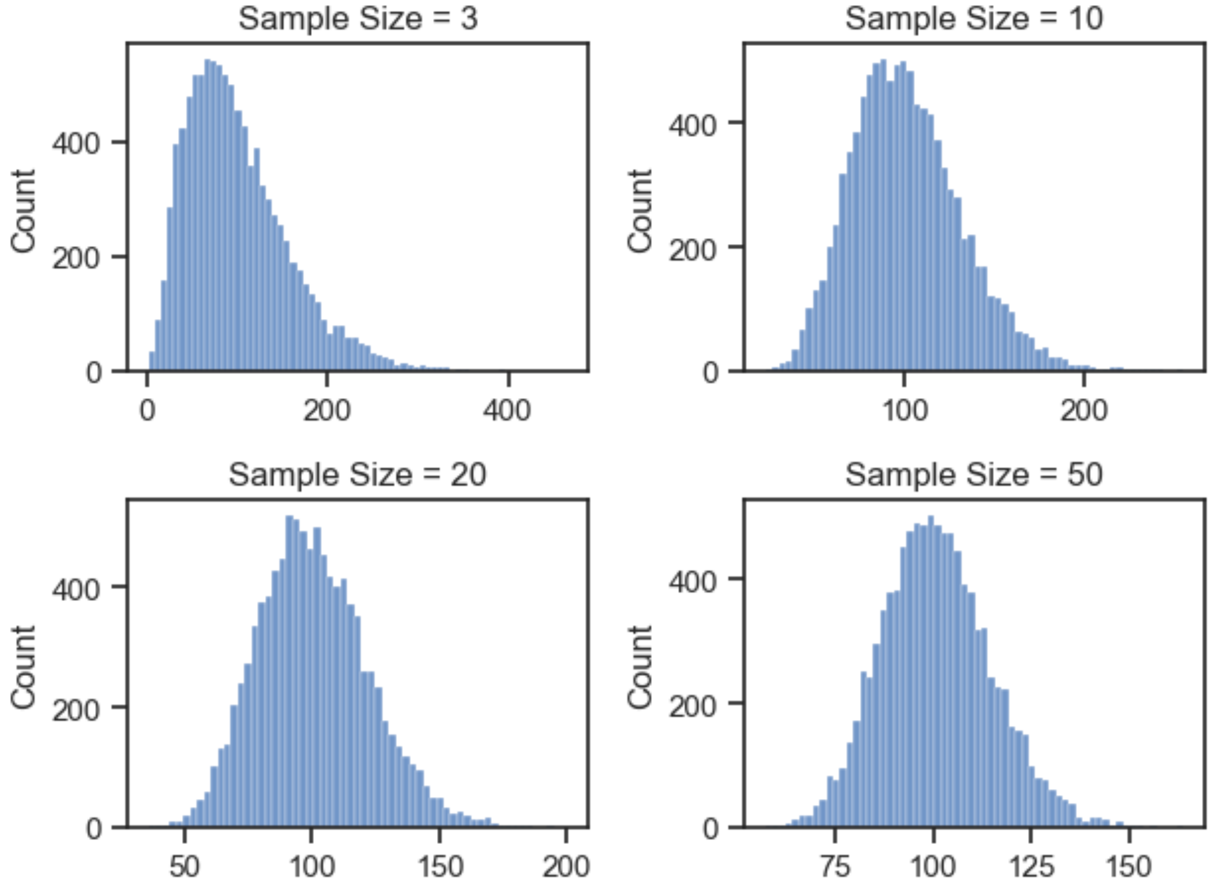
- Let's demonstrate the Central Limit Theorem (CLT) by find out how the sample size will affect the sample distribution of the sample mean.
- Let's create a population dataset following the gamma distribution (right-skewed)



Central Limit Theorem

Demo:

- Get sampling distributions of the mean with different sample size by sampling the population 10,000 times for each sample size



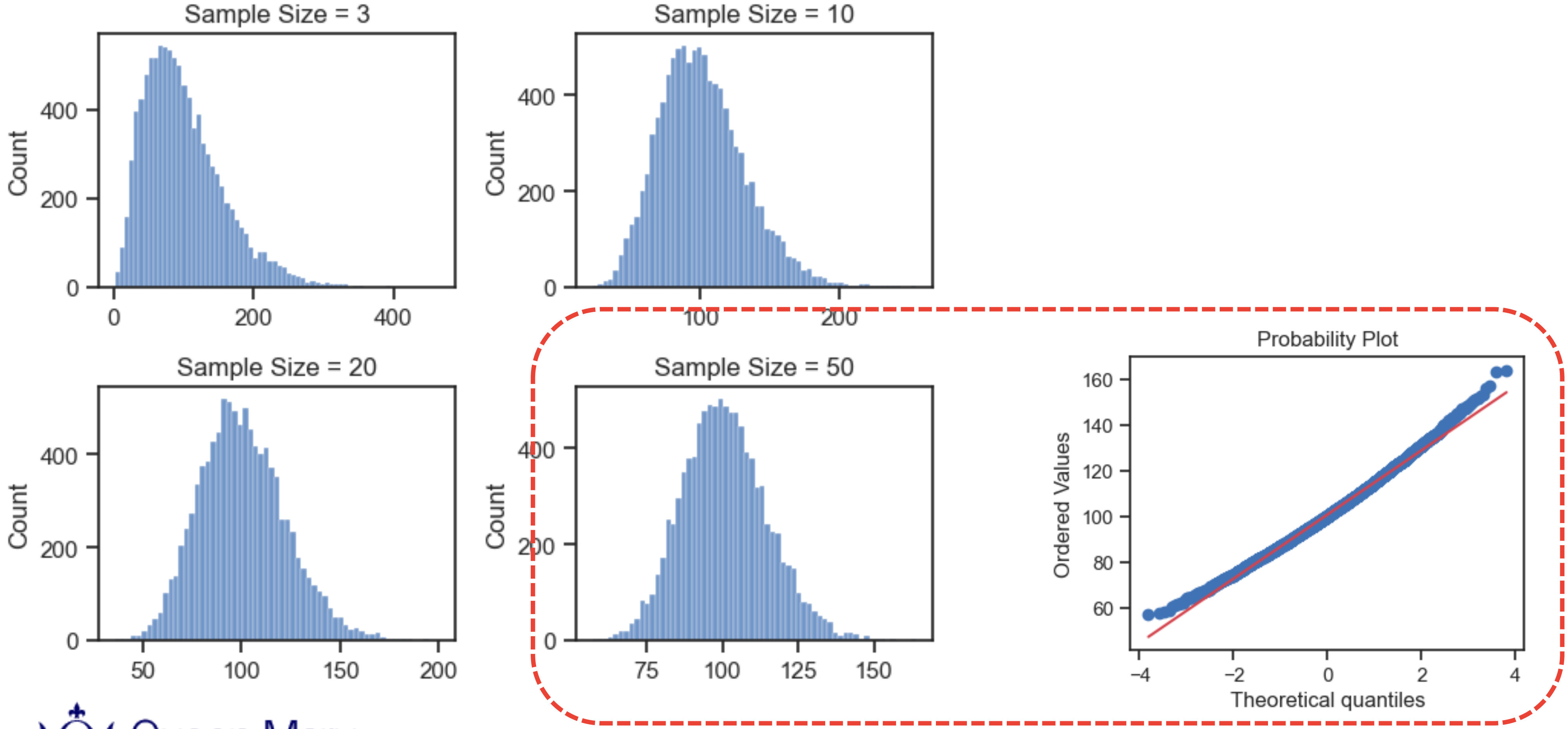
***Discussion**: By looking at the distribution of the original data and the sampling distribution of the sample mean with sample size 50, what can you find?*

***Discussion**: What tool can we use to verify if the sampling distribution of the sample mean with a sample size of 50 approximately follows the normal distribution?*

Central Limit Theorem

Demo:

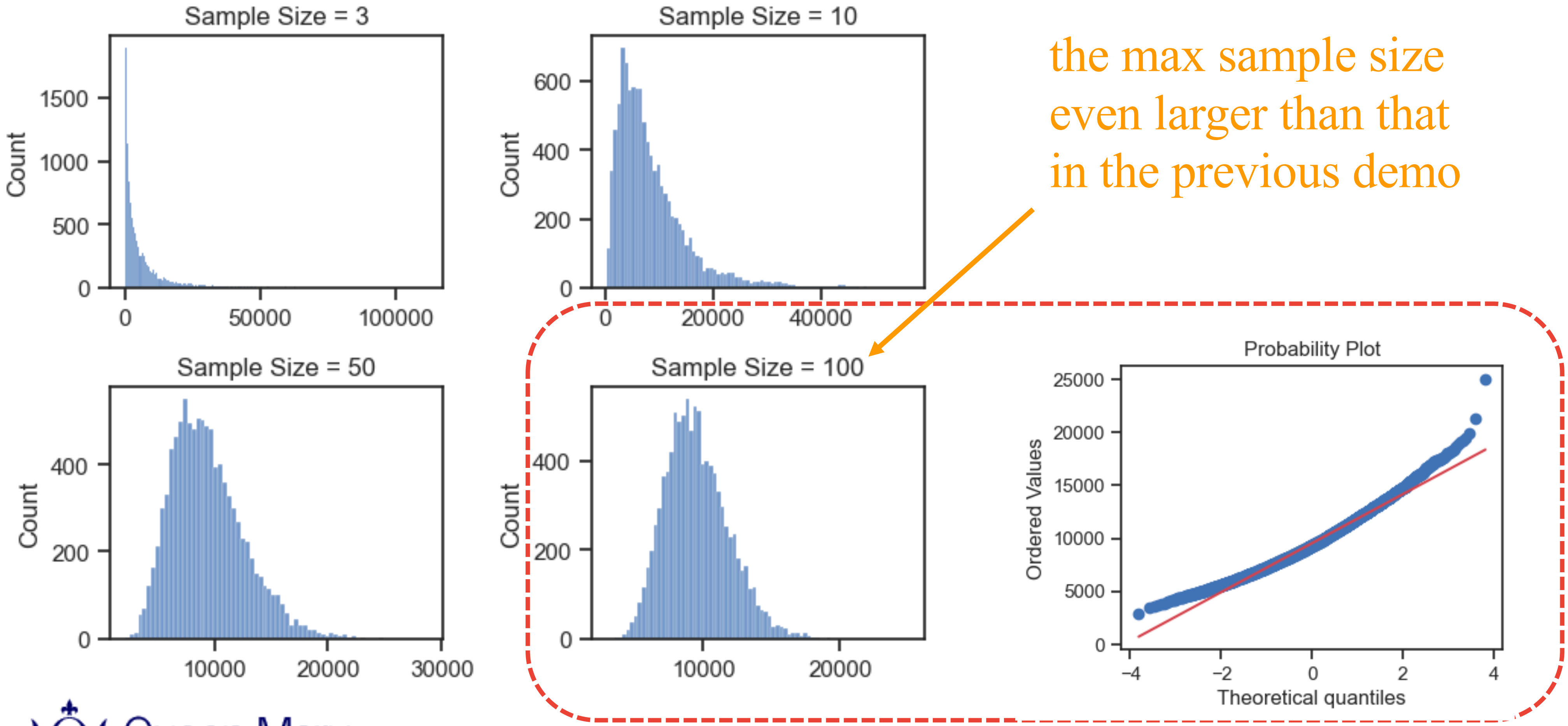
- Get sampling distributions of the mean with different sample size by sampling the population 10,000 times for each sample size



Central Limit Theorem

Demo:

- CLT doesn‘t apply to all statistics.
- Let’s continue to find out whether CLT applies to sample variance.



the max sample size even larger than that in the previous demo

Central Limit Theorem

- The Central Limit Theorem applies to these well-known statistics:
  - Sample means ($\bar{x}$)
  - Sample proportions (p)
  - Difference in sample means ($\bar{x}_1 - \bar{x}_2$)
  - Difference in sample proportions ($p_1 - p_2$).

CLT applies to additional statistics, but NOT ALL statistics.

BOOTSTRAPPING

Bootstrapping

- **Bootstrapping** is a statistical procedure that estimate the sampling distribution of a statistic by drawing additional samples with *replacement* from the sample itself and recalculating the statistic for each resample.
  - Do not rely on the theorems
  - Do not necessarily involve any assumptions about the data or the sample statistic
- The algorithm for a bootstrap resampling of the mean, for a sample of size n, is as follows:
  - Draw a sample value, record it, and then replace it.
  - Repeat n times.
  - Record the mean of the n resampled values.
  - Repeat steps 1–3 R times.
  - Use the R results (means) as the sampling distribution of the sample mean.

Bootstrapping

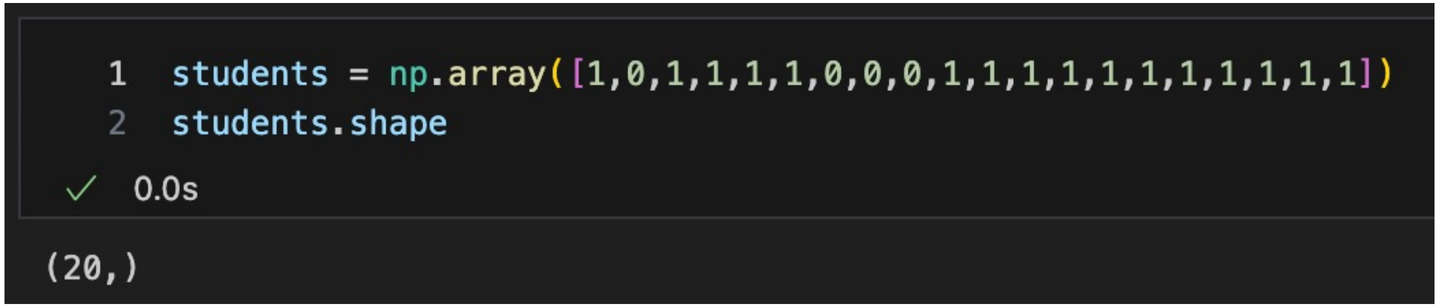
- We actually have been bootstrapping to create sampling distributions in the demonstrations in this lesson! We bootstrap the sampling distribution from the population.
- With **bootstrapping**, we can find where a parameter is by using a sampling distribution created from only **ONE sample** (if the sample is representative of our population).

Make Inference!

Bootstrapping

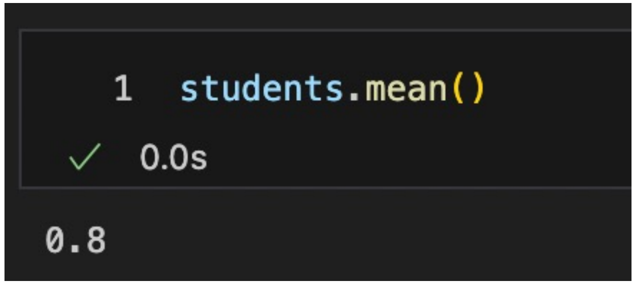
Demo:

- Let's use the student game-playing habit data again to demonstrate the idea of bootstrapping.



```
students = np.array([1,0,1,1,1,1,0,0,0,1,1,1,1,1,1,1,1,1,1,1])
students.shape
✓ 0.0s

(20,)
```

- The mean of the original **sample**.

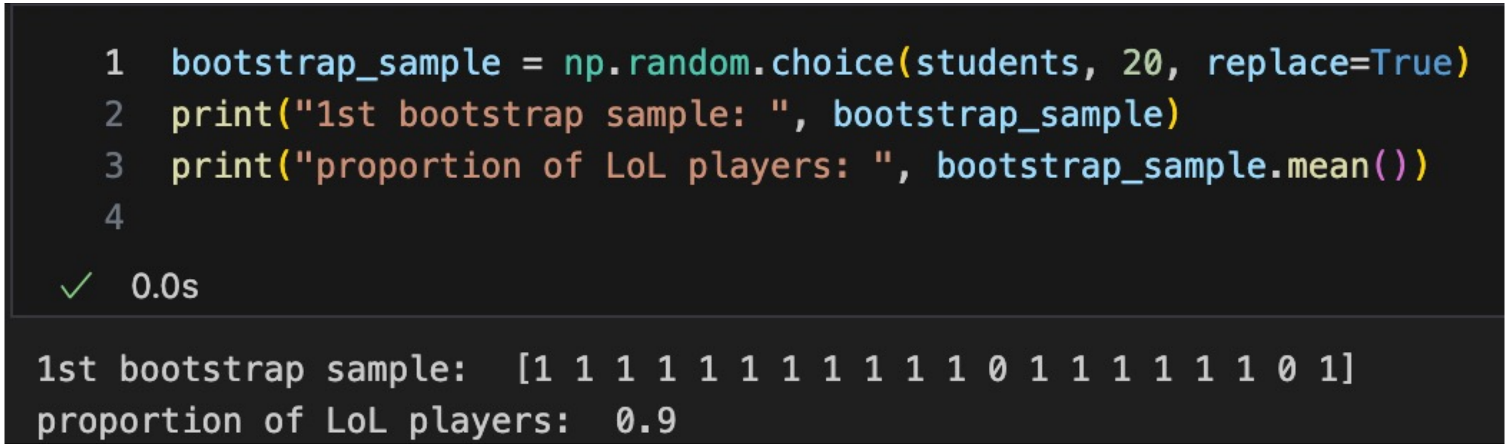


```
students.mean()
✓ 0.0s

0.8
```

Bootstrapping

Demo:

- Take the first bootstrap sample of the same size as the original sample and calculate the proportion of LoL players.



```python
bootstrap_sample = np.random.choice(students, 20, replace=True)
print("1st bootstrap sample: ", bootstrap_sample)
print("proportion of LoL players: ", bootstrap_sample.mean())

```

✓ 0.0s

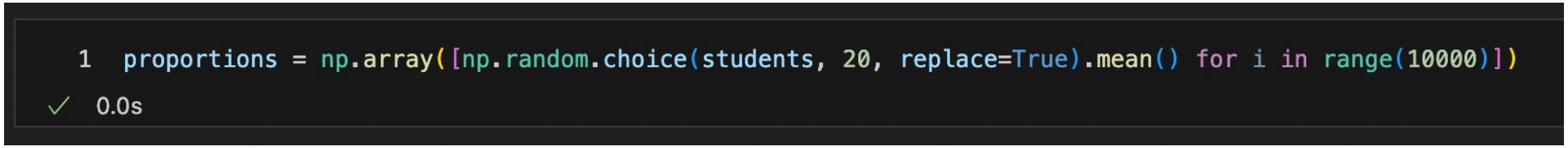
```
1st bootstrap sample:  [1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1]
proportion of LoL players:  0.9
```

***Discussion**: What can you find about the sample and the sample mean from the new sample created by bootstrapping and why?*

Bootstrapping

Demo:

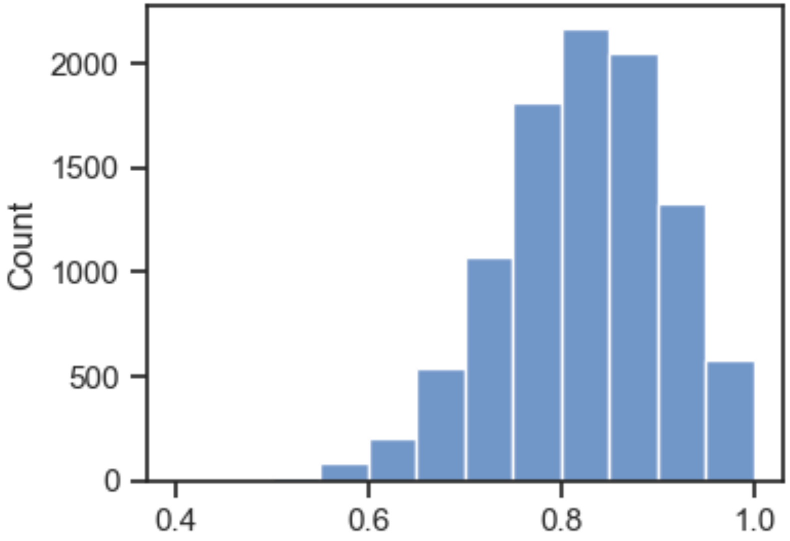
- Repeat the same bootstrap sampling 10,000 times to see how the proportion will change, which is the **sampling distribution of the proportion**



```
proportions = np.array([np.random.choice(students, 20, replace=True).mean() for i in range(10000)])
```

```
✓ 0.0s
```

List comprehension



Bootstrapping

- The application of the bootstrapping technique goes beyond even the use cases in this lesson. It has been used for leading **machine learning** algorithms, such as Random Forest and Stochastic Gradient Boost, etc.
- The bootstrapping technique enable us to make inference about a population parameter by only performing repeated sampling within our existing sample

STUDENT’S T-DISTRIBUTION

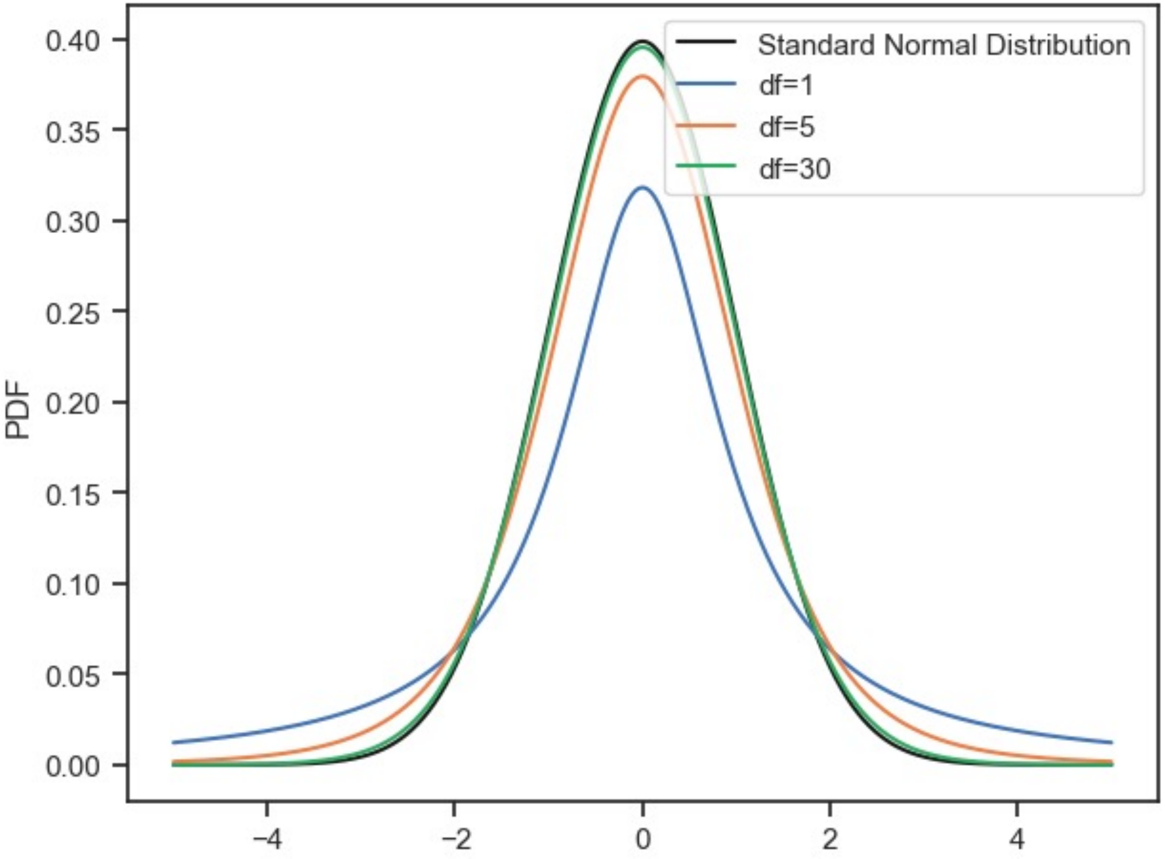
Student’s t-Distribution

- The **t-distribution** is a continuous probability distribution that is symmetric and bell-shaped like the normal distribution but with a shorter peak and thicker tails. It was designed to factor in the greater uncertainty associated with small sample sizes.
- The t-distribution describes the variability of the distances between sample means and the population mean when the population standard deviation is unknown and the sample data approximately follow the normal distribution. This distribution has only one parameter, the **degrees of freedom**, based on (but not equal to) the sample size.
- **Degree of Freedom (df)**: A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and numbers of groups.

Student’s t-Distribution

- t-distributions with smaller degrees of freedom have thicker tails and lower peaks. At around 30 degrees of freedom, the t-distribution closely approximates the standard normal distribution (z-distribution).

- Extensively used in depicting distributions of sample statistics. You should use the t-distribution when you need to assess the mean and do not know the population standard deviation.
- It’s particularly important to use it when you have a small (n < 30) sample size.



Student’s t-Distribution

- From the CLT, we know that the sampling distribution of the sample mean follows a normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- However, most of the time, the population standard deviation σ is unknown. If the sample size is larger enough, we consider the sample standard deviation s is a good estimate of σ

$$Z \approx \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1), if\ n > 30$$

- If the sample size is small n < 30), $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ will not follow the standard normal distribution, but will follow a t-distribution with a particular degree of freedom.

POISSON DISTRIBUTION

Poisson Processes

- A **Poisson process** is a process where events appears to happen at a certain rate but at completely random:
  - Number of people arriving at a bubble tea shop per hour
  - Number of animals adopted from an animal shelter per week
  - Number of earthquakes in California per year

- Time unit is irrelevant, as long as you use the same unit when talking about the same situation.

Poisson Distribution

- The **Poisson distribution** is a **discrete probability distribution** that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
  - Probability of 12 people arriving at a bubble tea shop per hour
  - Probability of ≥ 5 animals adopted from an animal shelter per week
  - Probability of < 20 earthquakes in California per year

Poisson Distribution

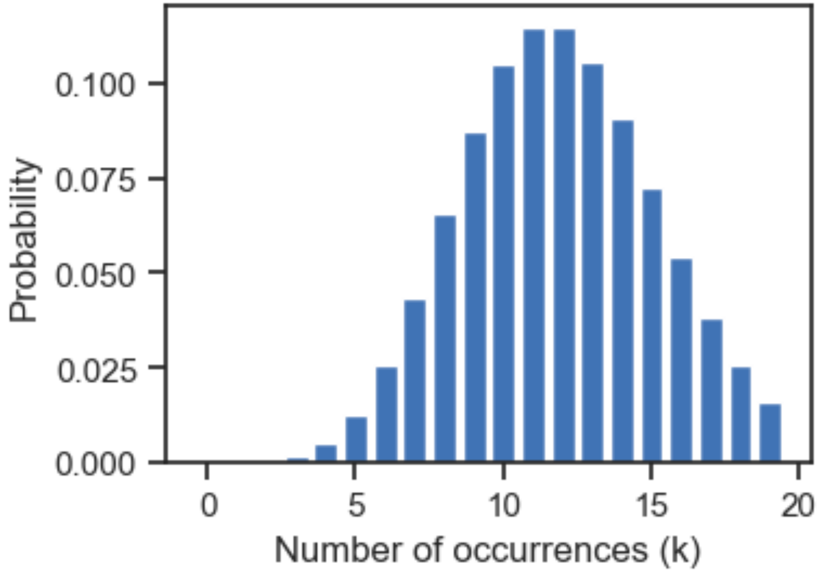
- **Probability Mass Function**.
  - A discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$, if it has a probability mass function given by:

$$f(k;\lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k is the number of occurrences (k = 0, 1, 2, ...) and λ is the average number of events per time interval.

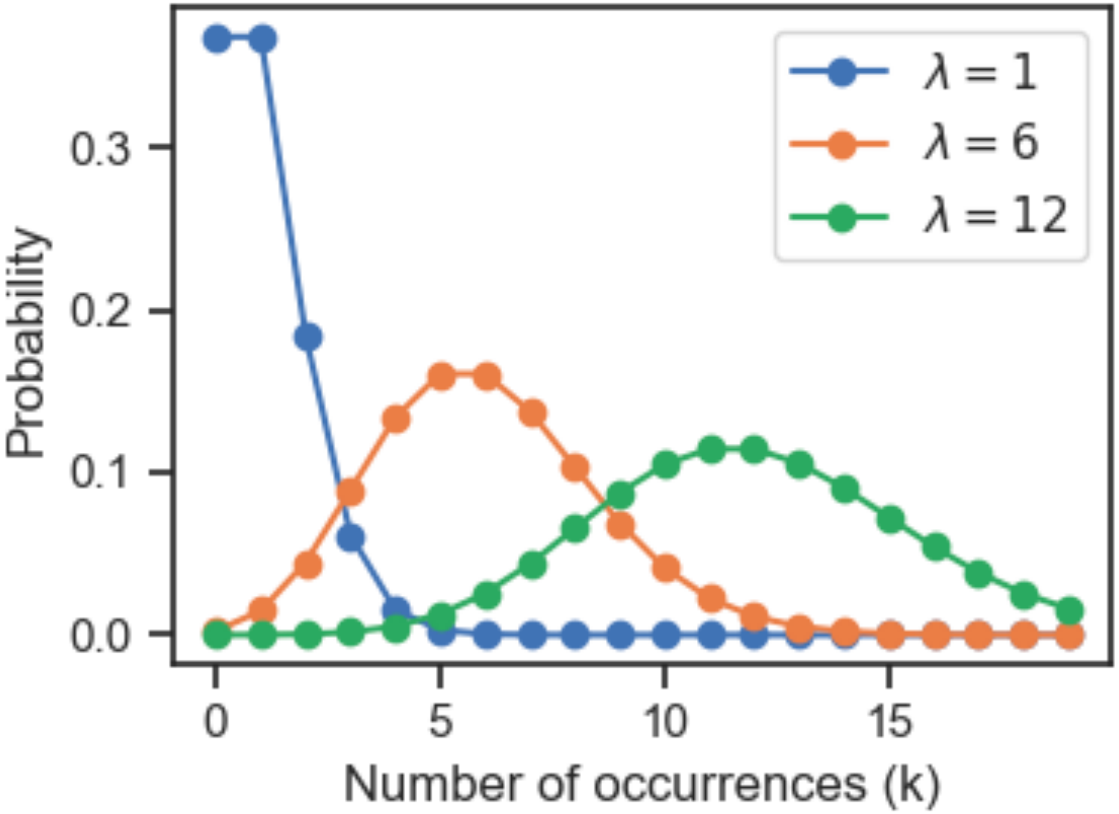
Example

- If the average number of people arriving at a bubble tea shop per hour is 12, the corresponding Poisson distribution looks like the bar chart
- 11 and 12 are the most likely number of customers arriving at the bubble tea shop in an hour.



Poisson Distribution

- The shape of the Poisson distribution changes as λ changes, and λ is the peak of the distribution.

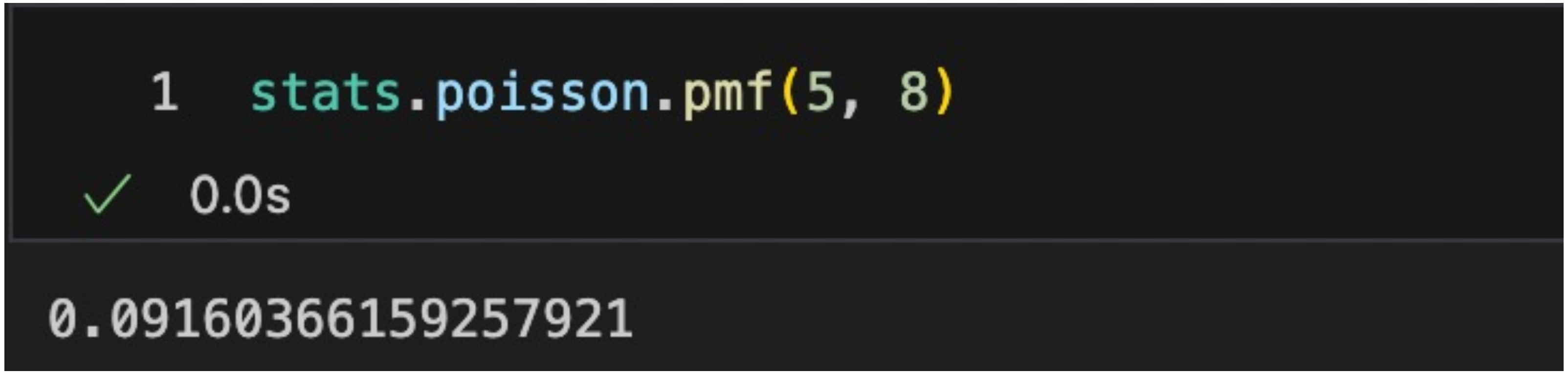


Poisson Distribution

Example/Demo

- If the average number of customers arriving at a bubble tea shop per hour is 8, answer the following questions.

1. What is the probability of 5 customers arriving at the shop in an hour ($P(k = 5)$)?



```python
stats.poisson.pmf(5, 8)
```

✓ 0.0s

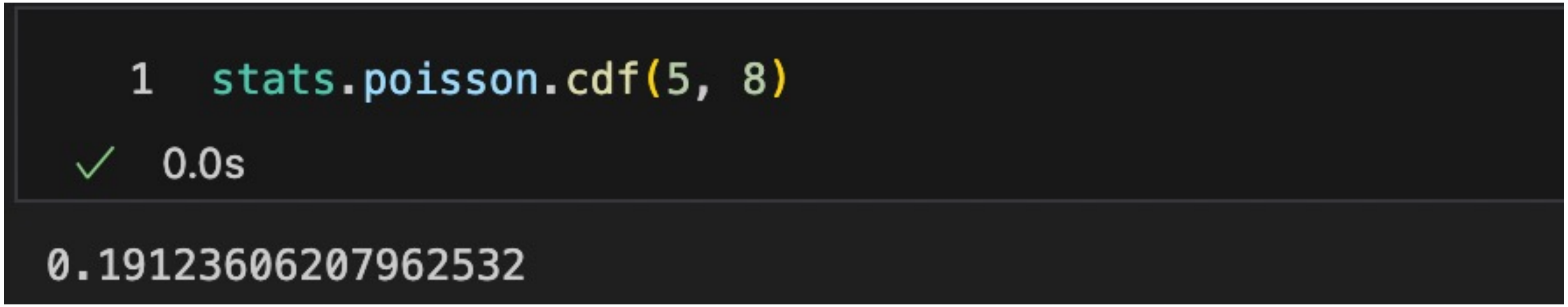
```
0.09160366159257921
```

Poisson Distribution

Example/Demo

- If the average number of customers arriving at a bubble tea shop per hour is 8, answer the following questions.

2. What is the probability of no more than 5 customers arriving at the shop in an hour ($P(k \leq 5)$)?



```
stats.poisson.cdf(5, 8)
```

```
✓ 0.0s
```

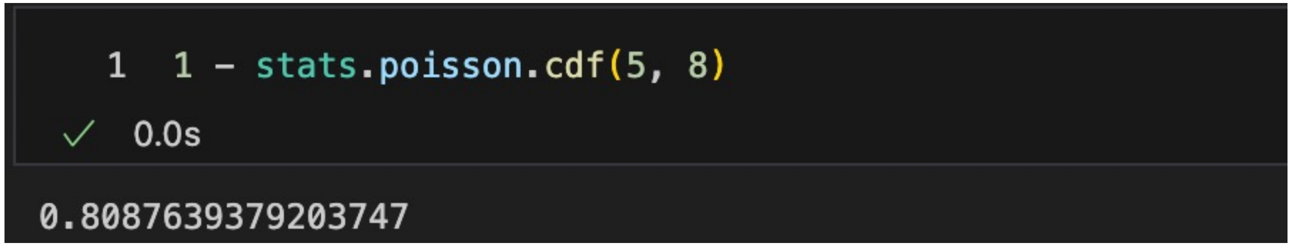
```
0.19123606207962532
```

Poisson Distribution

Example/Demo

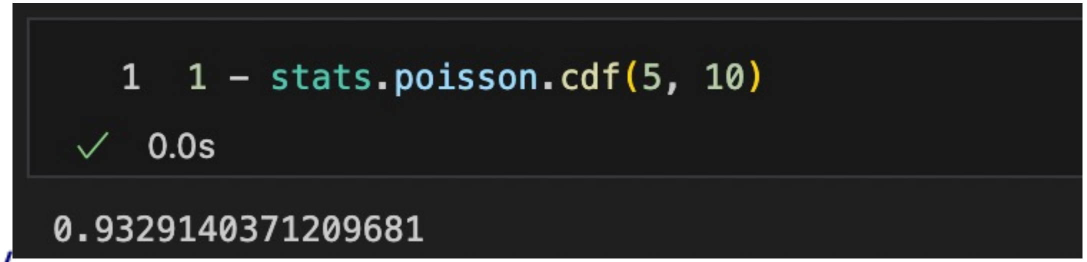
- If the average number of customers arriving at a bubble tea shop per hour is 8, answer the following questions.

3. What is the probability of more than 5 customers arriving at the shop in an hour ($P(k > 5)$)?



```
1 - stats.poisson.cdf(5, 8)
✓ 0.0s

0.8087639379203747
```

4. If the average number of customers arriving at a bubble tea shop per hour is 10, what is the probability of more than 5 customers arriving at the shop in an hour?



```
1 - stats.poisson.cdf(5, 10)
✓ 0.0s

0.9329140371209681
```

EXPONENTIAL DISTRIBUTION

Exponential Distribution

- The **exponential distribution** is the probability distribution of the time between events in a Poisson point process.
- Examples
  - Probability of < 10 minutes between customer arrivals in a bubble tea shop
  - Probability of > 1 day between adoptions
  - Probability of 6-8 months between earthquakes
  - Time unit is irrelevant, as long as it is consistent.

Exponential Distribution

- The probability density function (pdf) of an exponential distribution is:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}$$

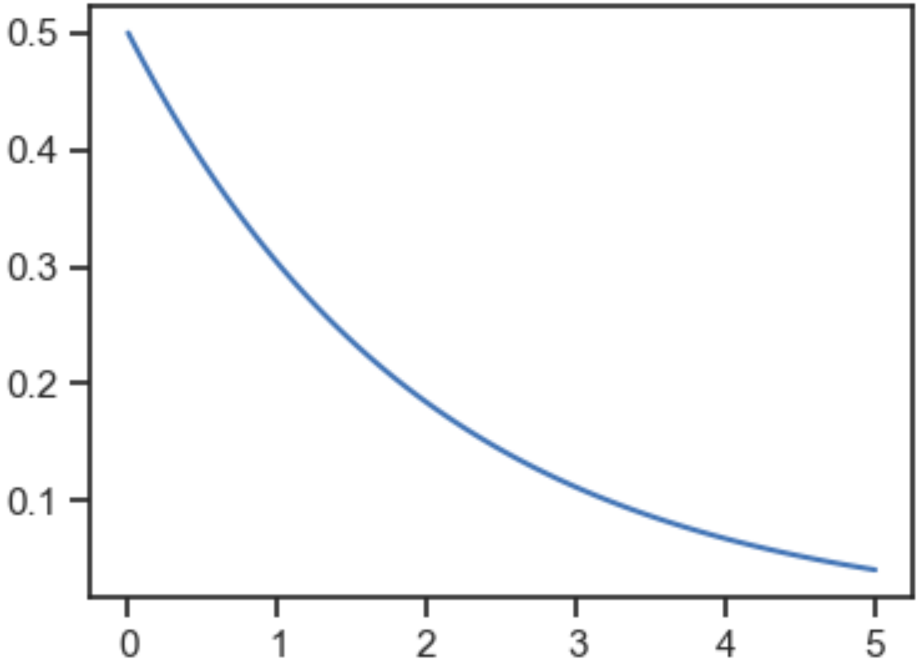
where λ is the parameter of the distribution, often called the rate parameter. It is the same λ value as that the Poisson does, which represents the rate.

- But the exponential distribution is continuous, unlike the Poisson distribution, since it represents time.

Exponential Distribution

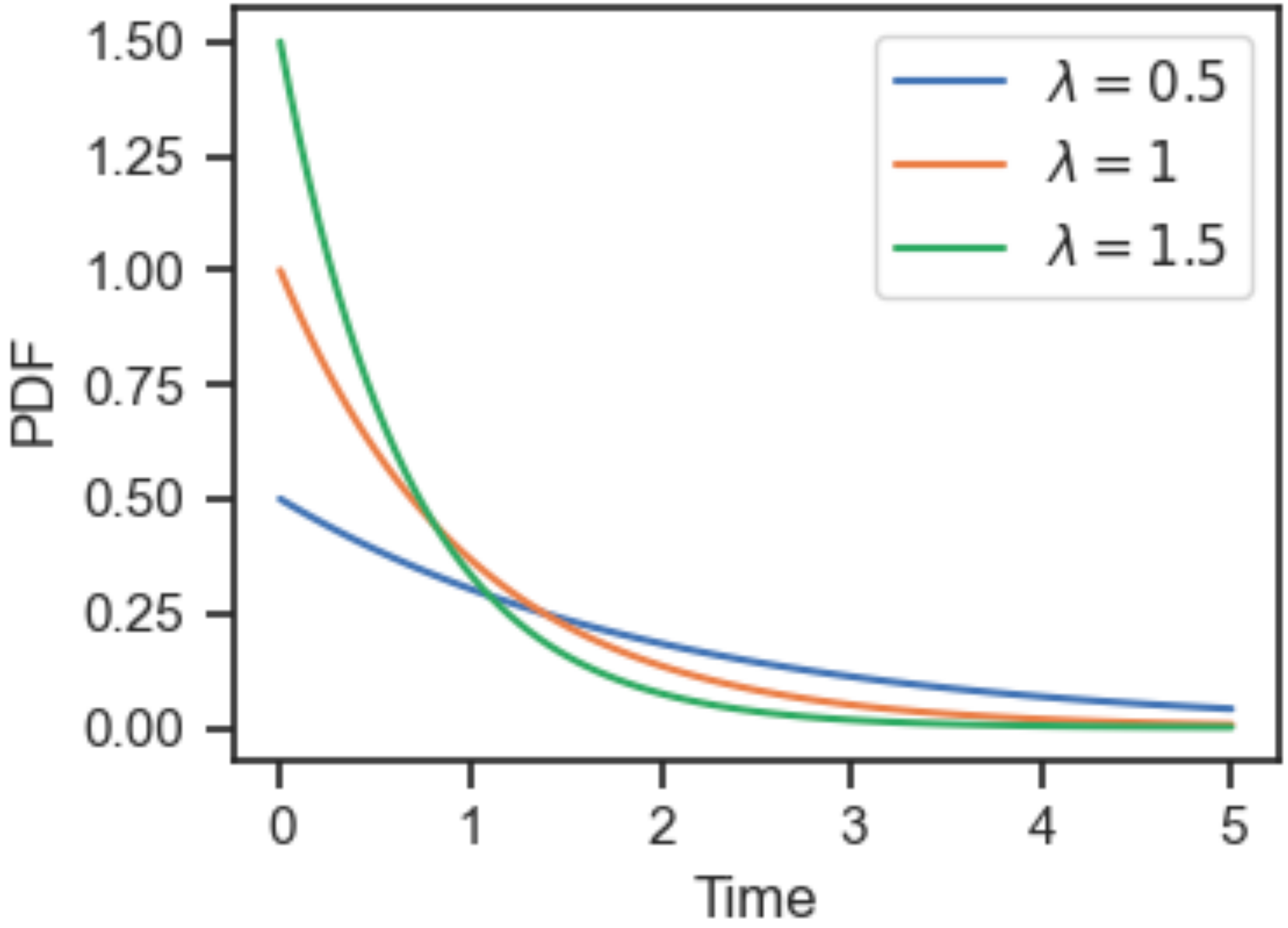
Example:

- On average, one customer service ticket is created every 2 minutes
  - $\lambda = 0.5$ customer service tickets created per minute on average
- The exponential distribution with a rate of 0.5 looks like:



Exponential Distribution

- The rate affects the shape of the distribution and how steeply it declines:



Exponential Distribution

- Poisson distribution measures frequency in terms of rate or number of events per time interval. The average number of events per time interval is λ
- The exponential distribution measures frequency in terms of time between events. The average time between events is $1/\lambda$.

Example:

In our customer service ticket example:

- In terms of rate (Poisson):
  - the average number of requests per minute is 0.5
- In terms of time between events (exponential):
  - the average time between requests is $\frac{1}{\lambda} = \frac{1}{0.5} = 2$ min

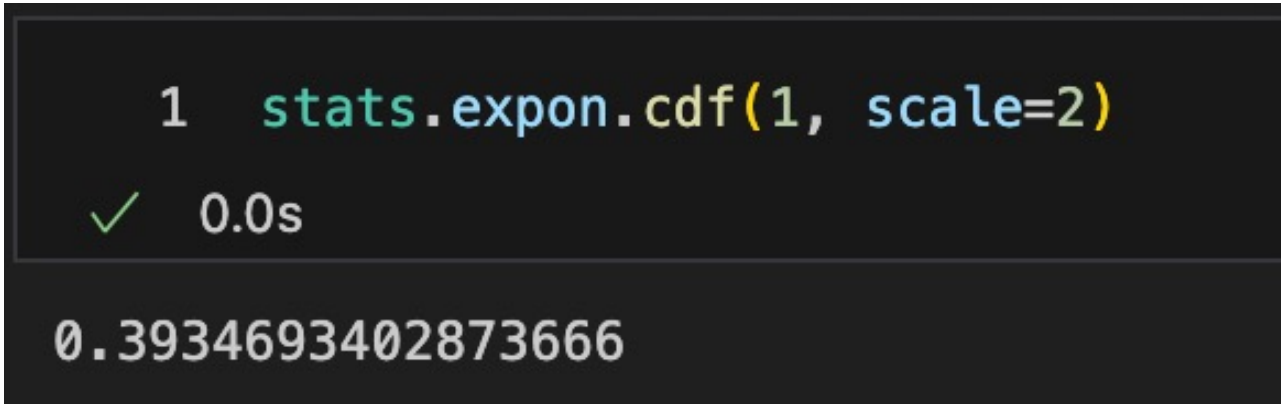
Exponential Distribution

Example:

Given one customer service ticket is created every 2 minutes on average:

1. What is the probability of waiting less than 1 minute for a new request ($P(wait < 1min)$)

1 ticket every 2 minutes means the rate $\lambda = 1 / 2 = 0.5$ (0.5 requests per minute), so scale = $1 / \lambda = 2$



```
stats.expon.cdf(1, scale=2)
```

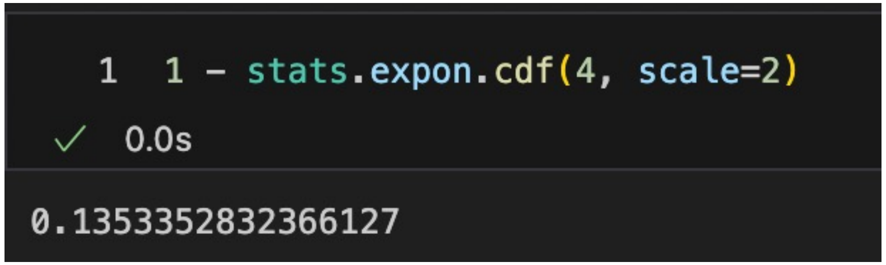
✓ 0.0s

```
0.3934693402873666
```

Exponential Distribution

Example:

2. What is the probability of waiting more than 4 minute for a new request ($P(wait > 4\,min)$)

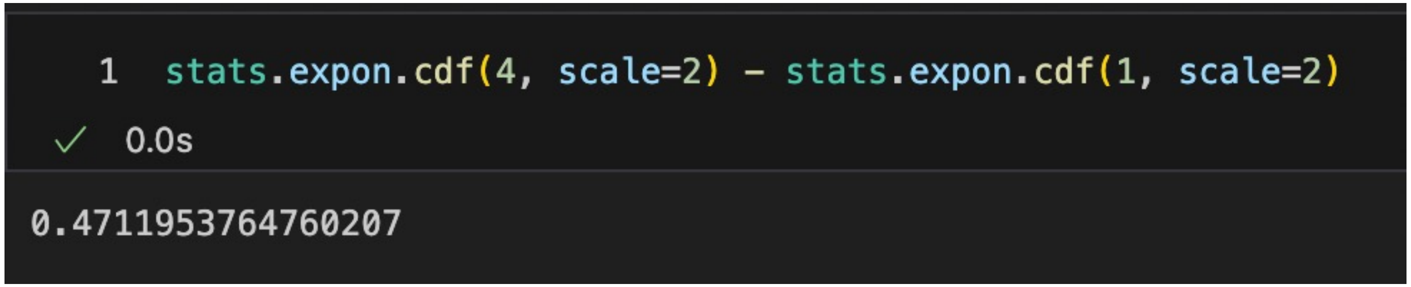


```
1 - stats.expon.cdf(4, scale=2)
```

✓ 0.0s

```
0.1353352832366127
```

3. What is the probability of waiting between 1 and 4 minute for a new request ($P(1 < \; wait < 4\,min)$)



```
stats.expon.cdf(4, scale=2) - stats.expon.cdf(1, scale=2)
```

✓ 0.0s

```
0.4711953764760207
```

RECAP

Data and Sampling Distributions

Recap

- **Binomial Distribution**
  - PMF formula
- **Normal Distribution**
  - PDF formula
  - Empirical Rule
  - Probability of Normal Distribution
    - z-table
  - Q-Q Plot
- **Descriptive Statistics and Inferential Statistics**
  - Population, Parameters
  - Sample, Statistics

Recap

- **Sampling Distribution**
- **Two Useful Theorems**
  - Law of Large Number
  - Central Limit Theorem
- **Bootstrapping**
- **Other Distributions**
  - Student’s t-distribution
  - Poisson Distribution
  - Exponential Distribution

Questions

Use student forum on QM+

chao.shu@qmul.ac.uk
xiaolanliu@qmul.ac.uk