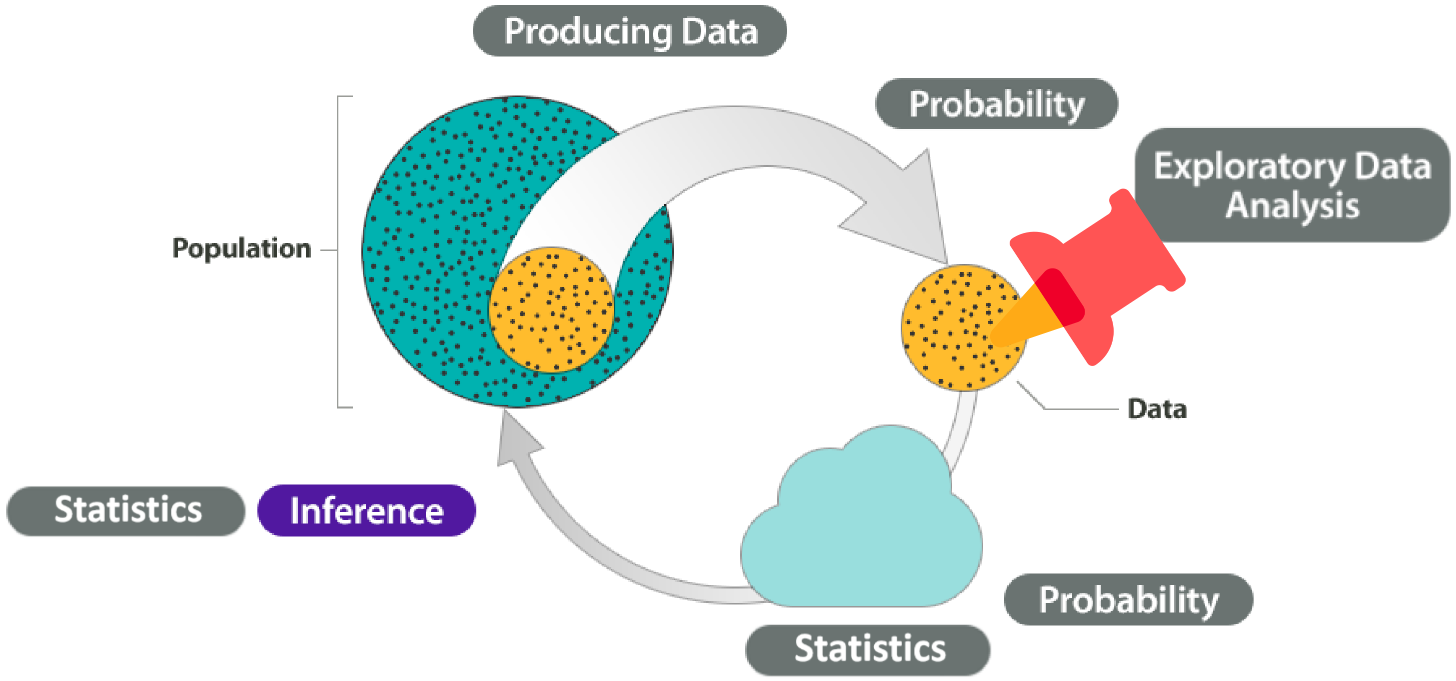
CDS 533
Statistics for Data Science

Instructor: Lisha Yu
Division of Artificial Intelligence
School of Data Science
Lingnan University

Fall 2024

Big Picture of Statistics

Exploratory Data Analysis (continued)

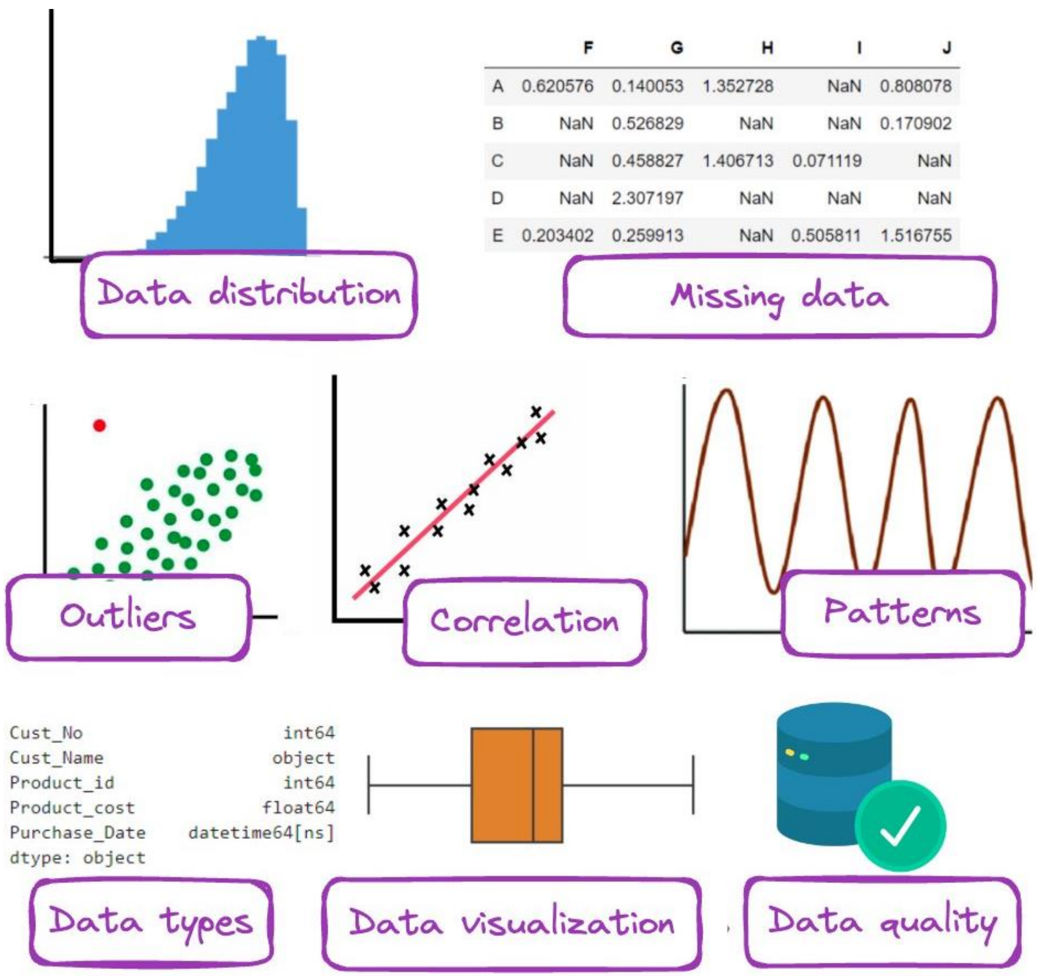


EDA Steps

- **Glance** the **whole data**.

- **Begin** by examining **each variable by itself**.
- **Then** study the **relationships among the variables**.

- **Begin** with a graph or **graphs**.
- **Then** add **numerical summaries** of specific aspects of the data.



Lab Dataset: airquality

The *airquality* dataset is built-in R object. It is a daily record of daily air quality measurements in New York, May to September 1973 with 153 observations on 6 variables.

Details: Daily readings
Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

[R: New York Air Quality Measurements (ethz.ch)]

Lab Dataset: diamonds

The *diamonds* dataset is built-in R object. Each row in the dataset is a single entry describing one diamond. There are 54,940 rows and 10 descriptive variables.

carat: The carat value of the Diamond

cut: The cut type of the Diamond, it determines the shine (Ideal' 'Premium' 'Good' 'Very Good' 'Fair')

color: The color value of the Diamond ('E' 'I' 'J' 'H' 'F' 'G' 'D')

clarity: The carat type of the Diamond ('SI2' 'SI1' 'VS1' 'VS2' 'VVS2' 'VVS1' 'I1')

depth: The depth value of the Diamond

table: Flat facet on its surface — the large, flat surface facet that you can see when you look at the diamond from above.

x: Width of the diamond

y: Length of the diamond

z: Height of the diamond

price: The price of the Diamond in USD.

[Prices of over 50,000 round cut diamonds — diamonds • ggplot2 (tidyverse.org)]

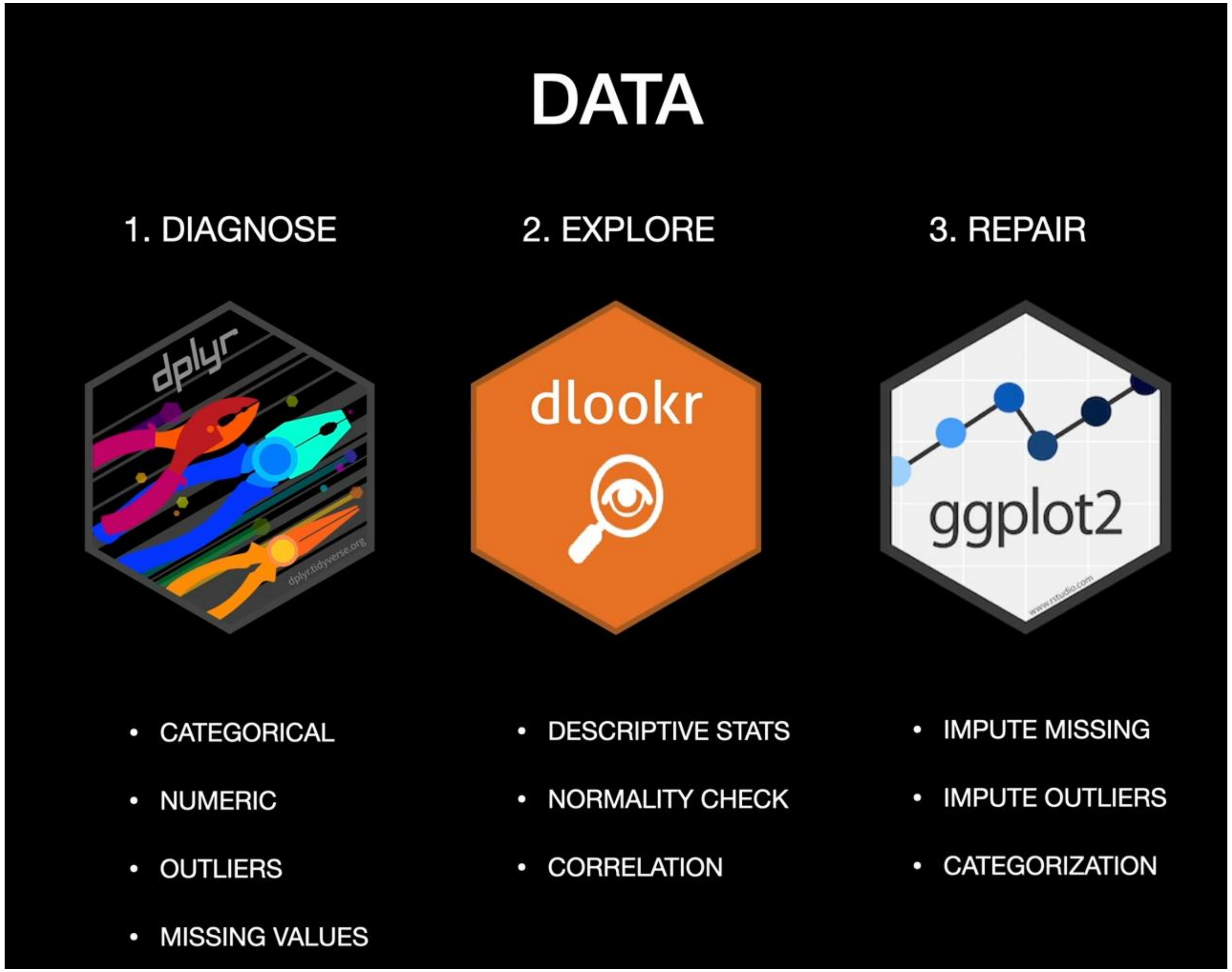
R Tools: DataExplorer

As data scientists spent exploring data and preparing it for analysis. Why not speed this up?

- Data
- Exploratory Data Analysis
  - Missing values
  - Distributions
    - Bar Charts
    - Histograms
    - QQ Plot
  - Correlation Analysis
  - Principal Component Analysis
  - Slicing & dicing
    - Boxplots
    - Scatterplots
- Feature Engineering
  - Replace missing values
  - Group sparse categories
  - Dummify data (one hot encoding)
  - Drop features
  - Update features
- Data Reporting

Introduction to DataExplorer (r-project.org)

R Tools: dlookr



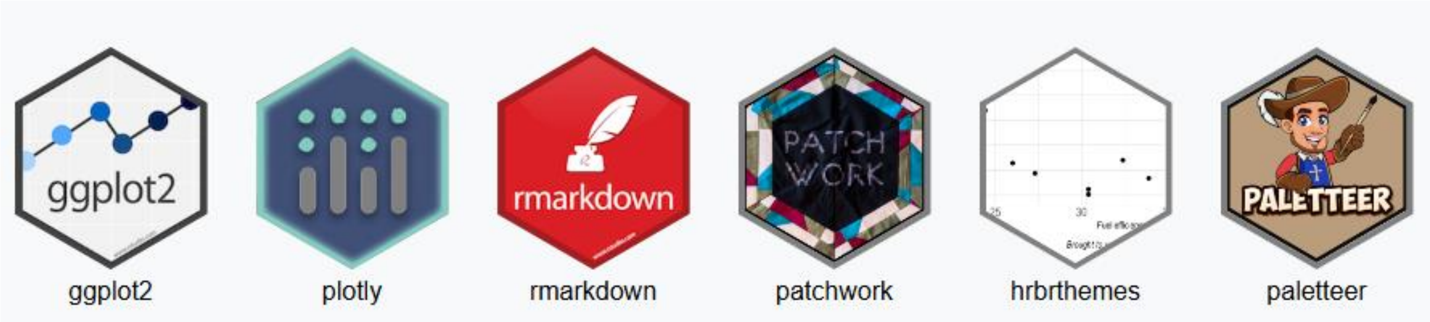
Download EDA-R from Moodle

[Introduce dlookr • dlookr (choonghyunryu.github.io)](https://choonghyunryu.github.io)

R Tools: EDA

Functions may find useful

- DataExplorer
- dlookr
- SmartEDA
- tidyverse
- bcdstats
- psych
- …



[From data to Viz | Find the graphic you need (data-to-viz.com)](#)

[The best R packages for data visualization (r-graph-gallery.com)](#)

The R community is always evolving, with new packages and functions emerging all the time. Keep an eye out for these new developments—they could **inspire and empower** your own projects!

Lab Time

Download Data Descriptions and Datasets
from Moodle

MBA Admission Class 2025

Laptop Price

Student Performance

Employee dataset

AI-Powered Job Market Insights

Digital Wallet Transaction

Lab Time

Two types of questions will always be useful for making discoveries. You can loosely word these questions as:

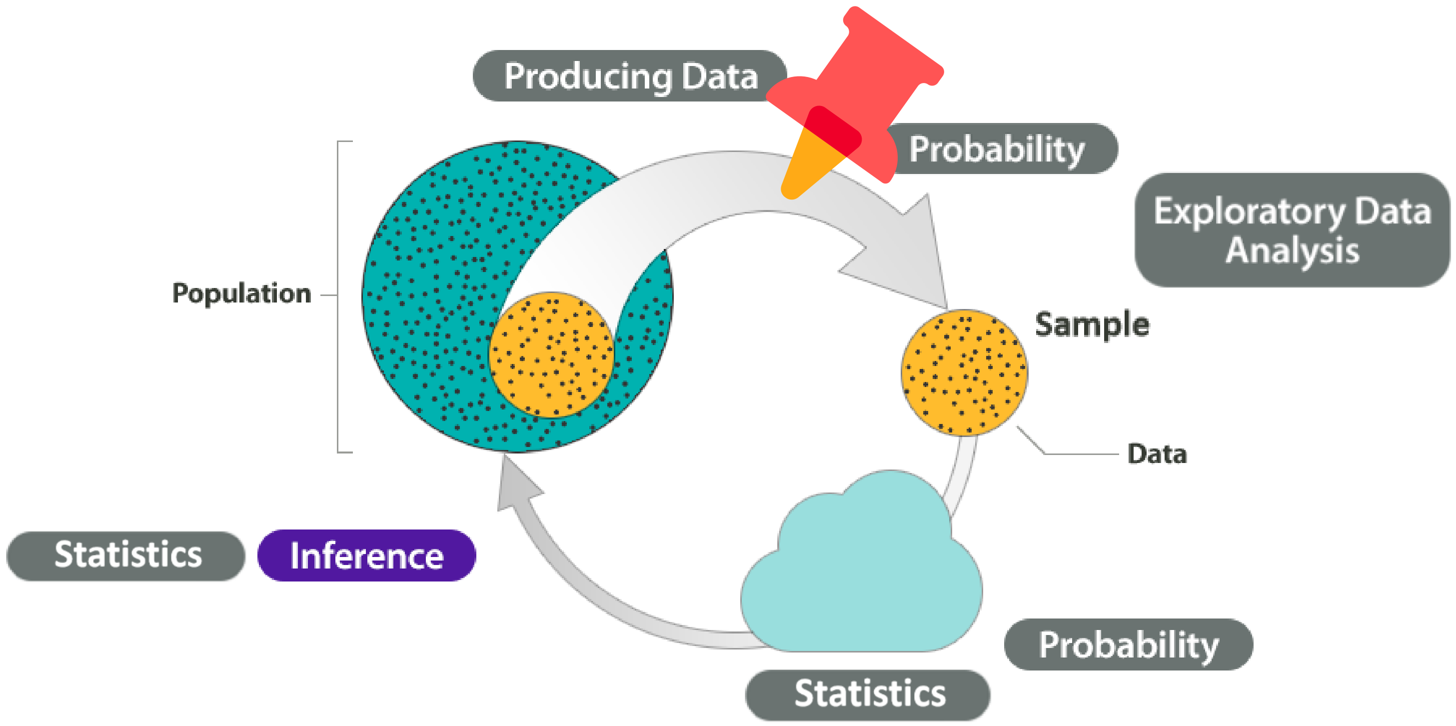
1. **What type of variation occurs within my variables?**
2. **What type of covariation occurs between my variables?**

EDA Cycle

EDA is an **iterative cycle**:

1. Generate questions about your data
2. Search for answers by visualizing, transforming, and modelling your data
3. Use what you learn to refine your questions and/or generate new questions

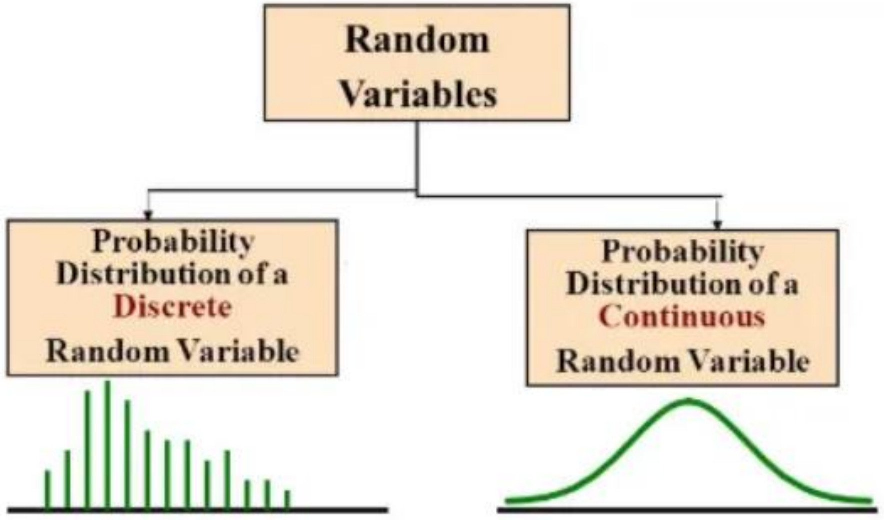
Big Picture of Statistics



Random Variables

Random Variables - Random outcomes corresponding to subjects randomly selected from a population.

Probability Distributions - A listing of the possible outcomes and their probabilities (discrete r.v.) or their densities (continuous r.v.).



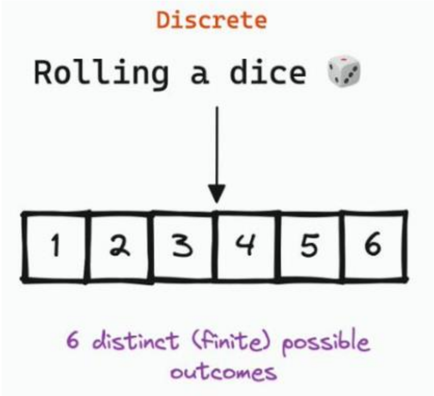
Discrete Probability Distributions

The **probability distribution** of a **discrete random variable** X lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

$$0 \le P(x) \le 1 \qquad \sum_{all\ x} P(x) = 1$$

- Discrete random variable
  - finite or countable



Example

Example: A bag contains 10 chips. 3 of the chips are red, 5 of the chips are white, and 2 of the chips are blue. Three chips are selected, with replacement. Find the probability that you select exactly one red chip.

Answer:

p = the probability of selecting a red chip = 0.3

$q = 1 – p = 0.7$ [redacted]

$n = 3$ [redacted]

$x = 1$ [redacted]

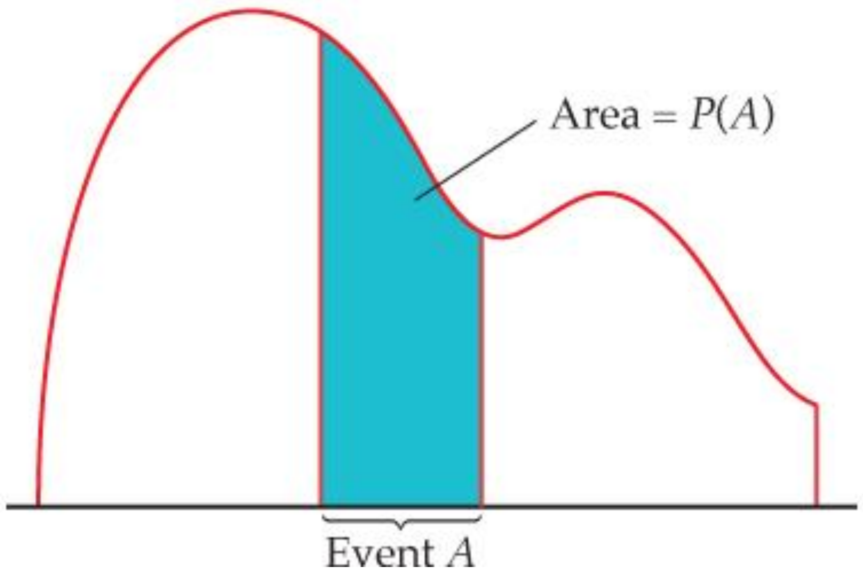
x	*P* (*x*)
0	0.240
1	0.412
2	0.265
3	0.076
4	0.008

In a **binomial experiment**, the probability of exactly x successes in n trials is

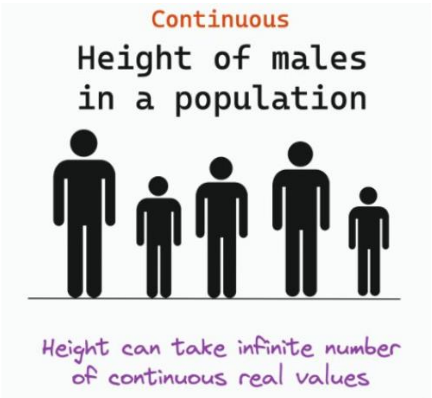
$$P(x) = {}_nC_x p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}.$$

Continuous Probability Distributions

A **continuous random variable** takes on all values in an **interval of numbers**. The probability distribution of *Y* is described by a **density curve**.



- Continuous random variable
  - infinitely many values
  - and the collection of values if not countable

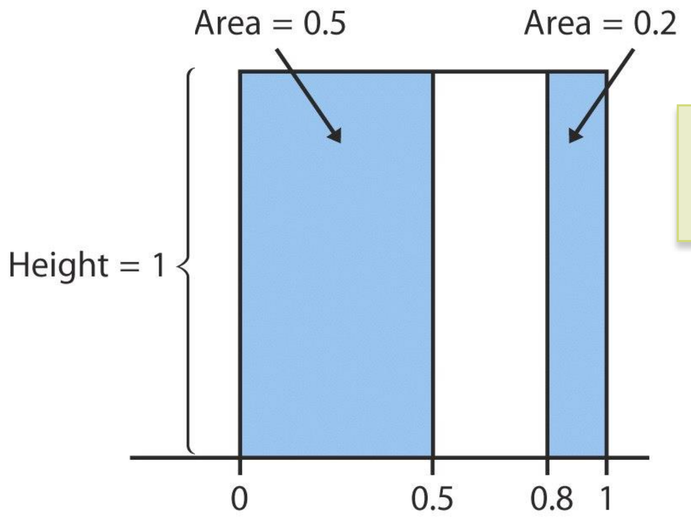


Example

A **continuous probability model** assigns probabilities as areas under a density curve. The area under the curve and above any range of values is the probability of an outcome in that range.

Example: Find the probability of getting a random number that is less than or equal to 0.5 OR greater than 0.8.

Answer:
$P\,(X \le 0.5 \text{ or } X > 0.8)$
$= P\,(X \le 0.5) + P(X > 0.8)$
$= 0.5 + 0.2$
$= 0.7$



Uniform distribution

Normal Distribution

Normal Distribution

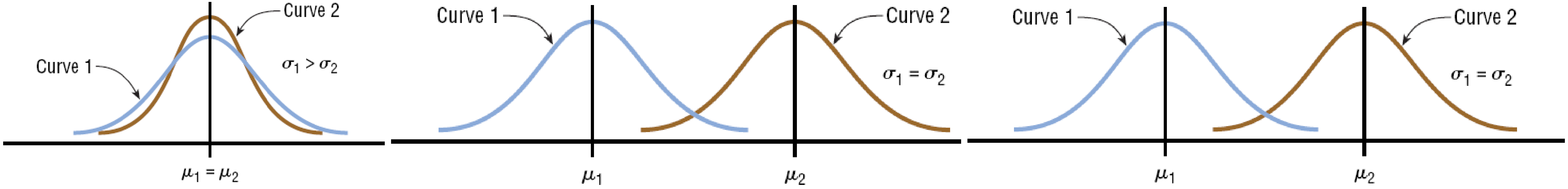
The mathematical equation for the normal distribution is:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{(2\sigma^2)}}}{\sigma\sqrt{2\pi}} = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} \qquad X \sim N(\mu, \sigma)$$

where $e \approx 2.718, \pi \approx 3.14$

$$\mu = E(X) = \sum xP(x)$$

$$\sigma = \sqrt{E(X-\mu)^2} = \sqrt{\sum (x-\mu)^2 P(x)} = \sqrt{\sum x^2 P(x) - \mu^2}$$



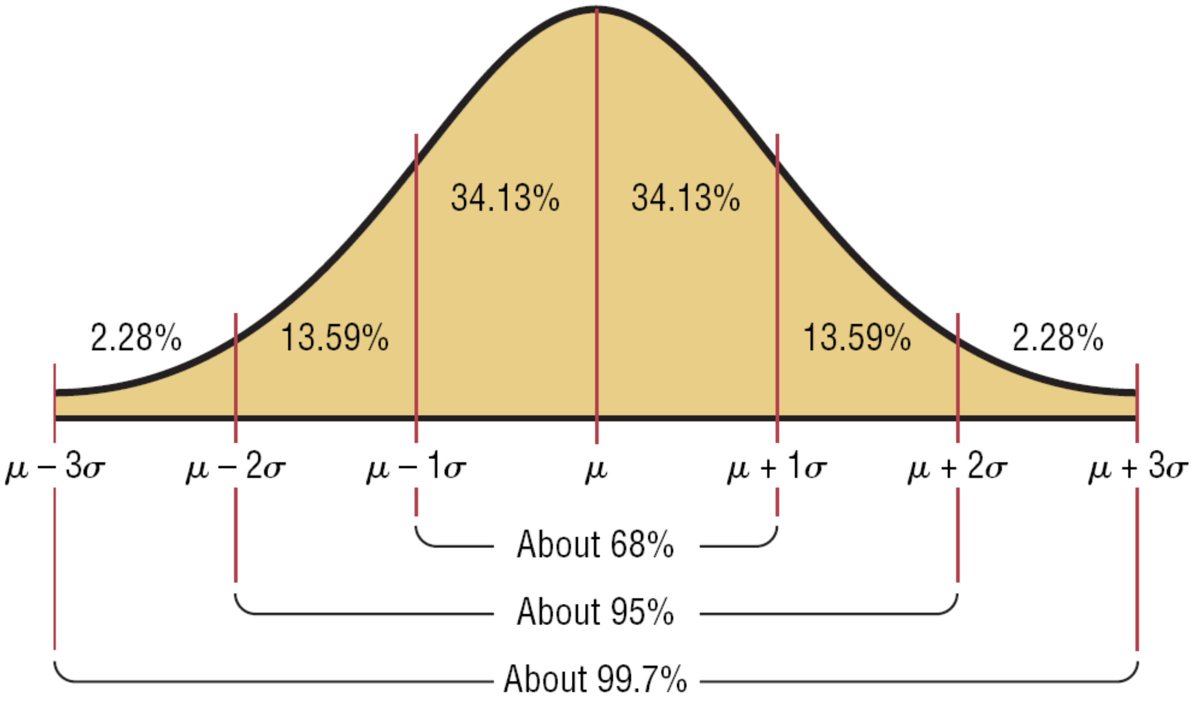
(a) Same means but different standard deviation

(b) Different means but same standard deviations

(b) Different means but same standard deviations

Normal Distribution Properties

- Bell-shaped, symmetric family of distributions
- Classified by 2 parameters: Mean (μ) and standard deviation (σ).
- **(68-95-99 Rules)**



$$P(X \geq \mu) = 0.50$$

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

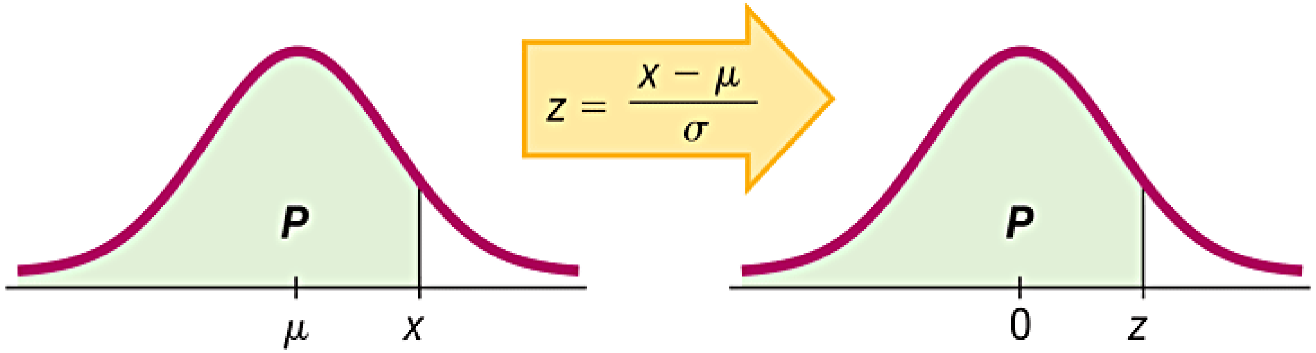
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

Standard Normal (Z) Distribution

Problem: Unlimited number of possible normal distributions ($-\infty < \mu < \infty$, $\sigma > 0$)

Solution: Standardize the random variable to have mean 0 and standard deviation 1

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



(a) Nonstandard Normal Distribution

(b) Standard Normal Distribution

Sampling Distributions

Sampling Distribution of a Statistic

Sample statistics based on random samples are also **random variables** and have **sampling distributions** that are probability distributions for the statistic (outcomes that would vary across samples)

Parameter		Statistics
μ	Mean	$\bar{x}$
σ	Standard deviation	s
π	Proportion	p
N	Size	n

Central Limit Theorem :

https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

The Central Limit Theorem

When samples are large and measurements independent then many estimators have normal sampling distributions (CLT):

Sample Mean: $$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sample Proportion: $$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- For ***n* > 30**, the distribution of the sample means can be approximated reasonably well by a normal distribution as *n* becomes larger.
- If the original population is **normally distributed**, then for **any** sample size *n*, the sample means will be normally distributed (not just the values of *n* larger than 30).

Example

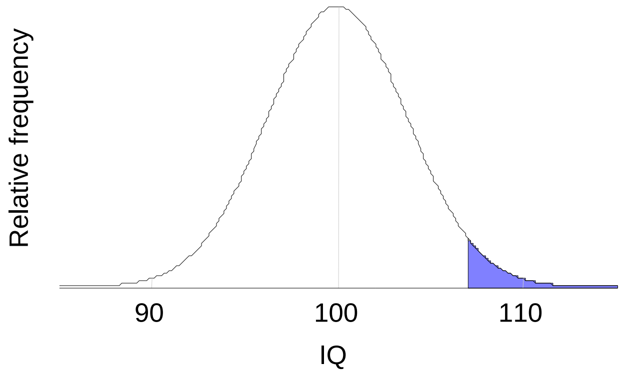
Example: Suppose the IQ of the population is distributed normally with a mean of 100 and a standard deviation of 15. If we draw 16 people at random from the population, what is the probability that the mean IQ of this sample will be greater than 107?

Answer: We know that the sampling distribution of the mean with n=16 will have a mean and standard deviation of:

$$u_{\overline{X}} = u_X = 100 \qquad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{15}{\sqrt{16}} = 3.75$$

The z-score for 107 is therefore

$$z = \frac{\overline{X} - u_{\overline{X}}}{\sigma_{\bar{x}}} = \frac{107-100}{3.75} = 1.867$$



The area under the normal distribution above z=1.86 is 0.0314
So there is a less than 5% chance of observing a sample mean greater than 107.

Finite Population Correction

Correction for a Finite Population

The formula for standard error of the mean is accurate when the samples are drawn with replacement or are drawn without replacement from a very large or infinite population.

A **correction factor** is necessary for computing the standard error of the mean for samples drawn **without replacement from a finite population.**

- Sampling without replacement and the sample size n is greater than 5% of the finite population of size N (that is, $n > 0.05N$)

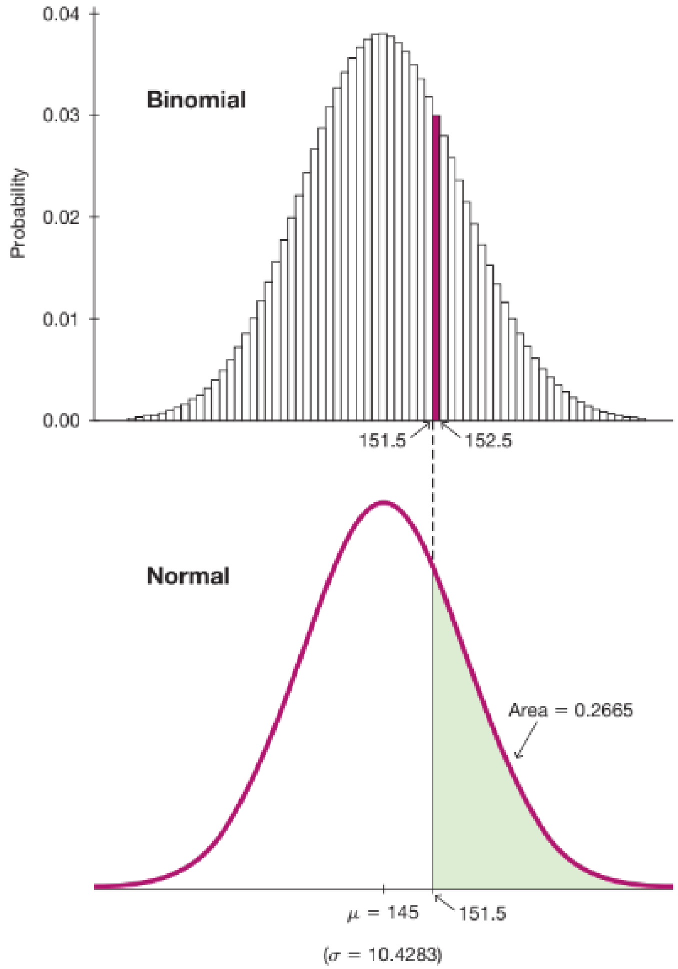
Finite population correction factor $\sqrt{\dfrac{N-n}{N-1}}$

The standard error of the mean $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

$$Z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}}}$$

Normal as Approximation to Binomial

Recall: The sampling distribution of a sample proportion tends to approximate a normal distribution.



Requirements

1. The sample is a simple random sample of size n from a population in which the proportion of successes is p
2. $np \geq 5$ and $nq \geq 5$

(The requirements of $np \geq 5$ and $nq \geq 5$ are common, but some recommend using 10 instead of 5.)

Continuity Correction

The Normal Approximation to the Binomial Distribution

If normal approximation requirements are satisfied, then the probability distribution of the random variable x can be approximated by a normal distribution with these parameters:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

Binomial	**Normal**
When finding:	Use:
$P(X = a)$	$P(a - 0.5 < X < a + 0.5)$
$P(X \geq a)$	$P(X > a - 0.5)$
$P(X > a)$	$P(X > a + 0.5)$
$P(X \leq a)$	$P(X < a + 0.5)$
$P(X < a)$	$P(X < a - 0.5)$

For all cases, $\mu = np, \sigma = \sqrt{npq}, np \geq 5, nq \geq 5$

Example

$\mu = np, \sigma = \sqrt{npq}, np \geq 5, nq \geq 5$

Example: Assume that 6% of American drivers text while driving. If 300 drivers are selected at random, find the probability that exactly 25 say they text while driving. (Use Normal approximation)

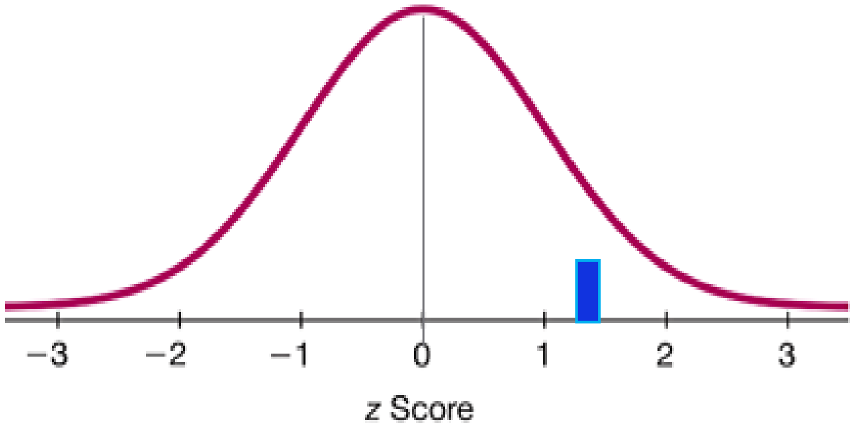
Answer: $n = 300,\ p = 0.06$

$$x \sim Binom(300,0.06)$$

Check $np = 300 \times 0.06 = 18 \geq 5$ and $nq = 300 \times 0.94 = 282 \geq 5$

$$\mu = np = 300 \times 0.06 = 18$$
$$\sigma = \sqrt{npq} = \sqrt{300 \times 0.06 \times 0.94} = 4.1134$$

$$P(x = 25) = P(24.5 < x < 25.5)$$
$$= P(\frac{24.5-18}{4.11} < z < \frac{25.5-18}{4.11})$$
$$= 0.9656\text{-}0.9428$$
$$=0.0227$$



END OF LECTURE!