Lecture 2

Descriptive Statistics and Hypothesis Testing

Ashish Dandekar

Lecture Overview

Descriptive Statistics
Probability
Hypothesis Testing

Descriptive Statistics

Why do we need statistics?

Case 1

A real estate agent wants to estimate the average per square foot of residential property in a city.

Case 2

A manufacturing plant wants to estimate average life of their product.

Case 3

A pharmaceutical company wants to estimate the effectiveness of its new vaccine on an average human being.

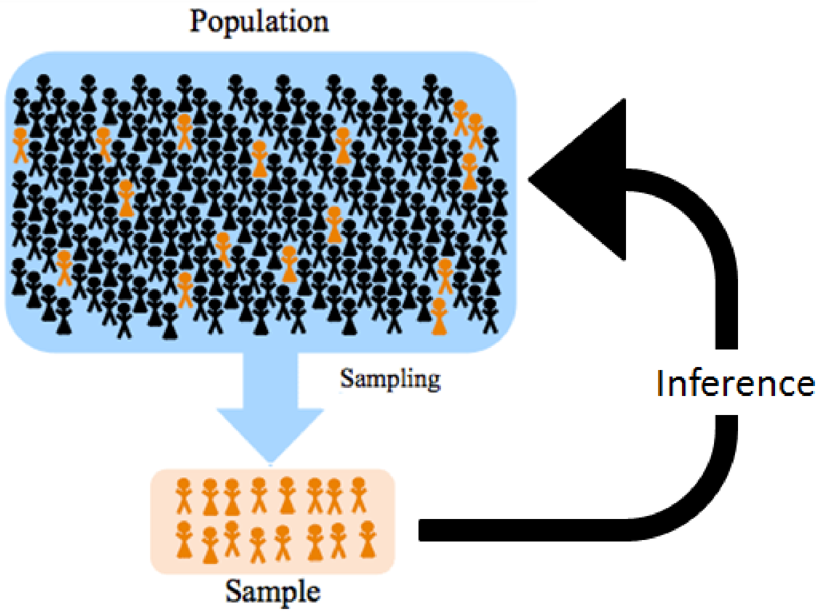
Population versus Sample

Population

It refers to a collection of entire set of measurements of any characteristic that we are interested in.

Sample

It refers to a smaller set of measurements collected from the population (*a.k.a. a subset of the population*).



Source: Towards Data Science

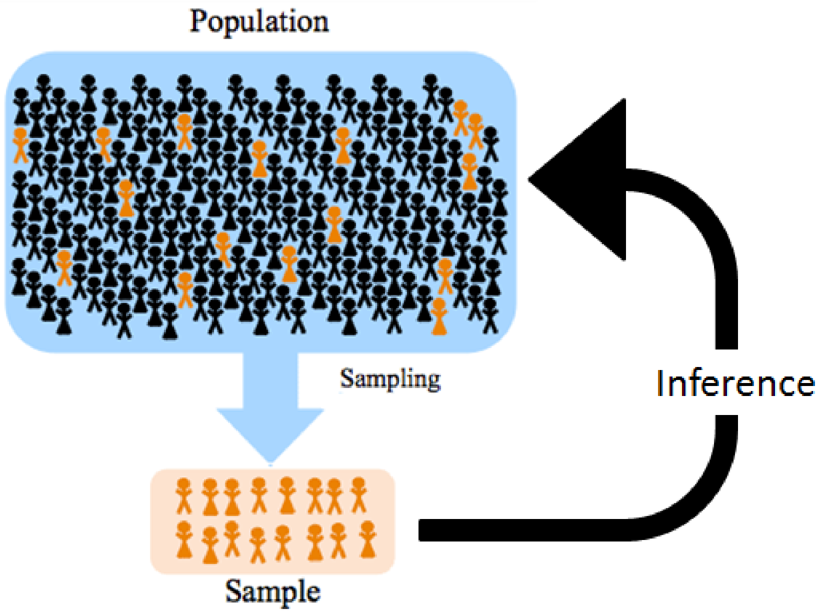
Population versus Sample

Population

It refers to a collection of entire set of measurements of any characteristic that we are interested in.

Sample

It refers to a smaller set of measurements collected from the population (*a.k.a. a subset of the population*).



Source: Towards Data Science

Beware of the Sampling Bias

A sample should truly *represent* the population.

Notation

- A **population** comprises of $x_1, x_2, ..., x_N$ datapoints.
- A **sample** is any subset of size $n \leq N$ of the population.
- A **statistic** is any function computed over the sample.

	Population	Sample
Mean	μ	$\bar{X}$
Variance	Ω^2	S^2
Proportion	π	p

Notation

- A **population** comprises of $x_1, x_2, ..., x_N$ datapoints.
- A **sample** is any subset of size $n \leq N$ of the population.
- A **statistic** is any function computed over the sample.

	Population	Sample
Mean	μ	$\bar{X}$
Variance	Ω^2	S^2
Proportion	π	p

Note.

- *Greek* symbols are used for population measures.
- *Latin* symbols are used for sample measures.
- n is used to denote total number of datapoints in a sample.
- d is used to denote the dimension of any datapoint x_i.

Measures of Location

Mean

It is the mathematical average of the datapoints computed as follows:

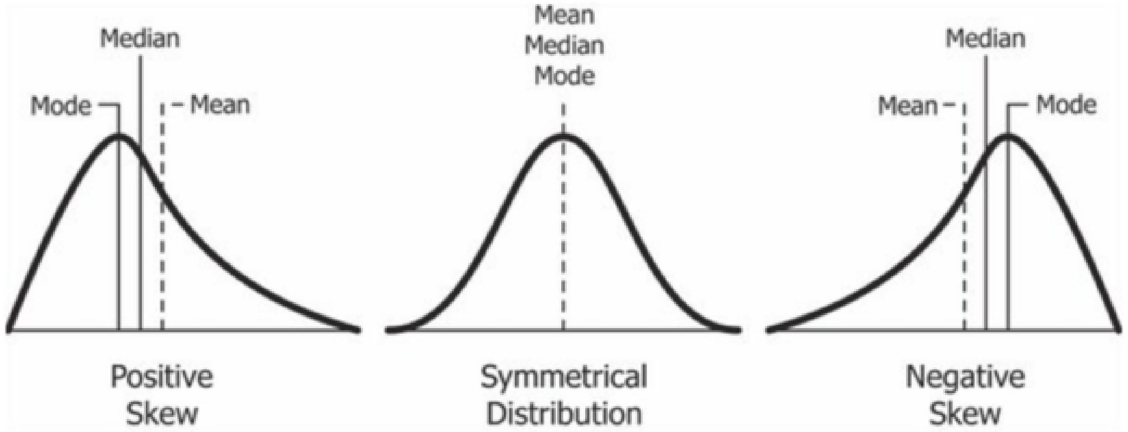
$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Median

It is the middle value in the sorted dataset.

Mean

It is the most frequent value in the dataset.



Source: Measures of Central Tendency

Measures of Location

Mean

It is the mathematical average of the datapoints computed as follows:

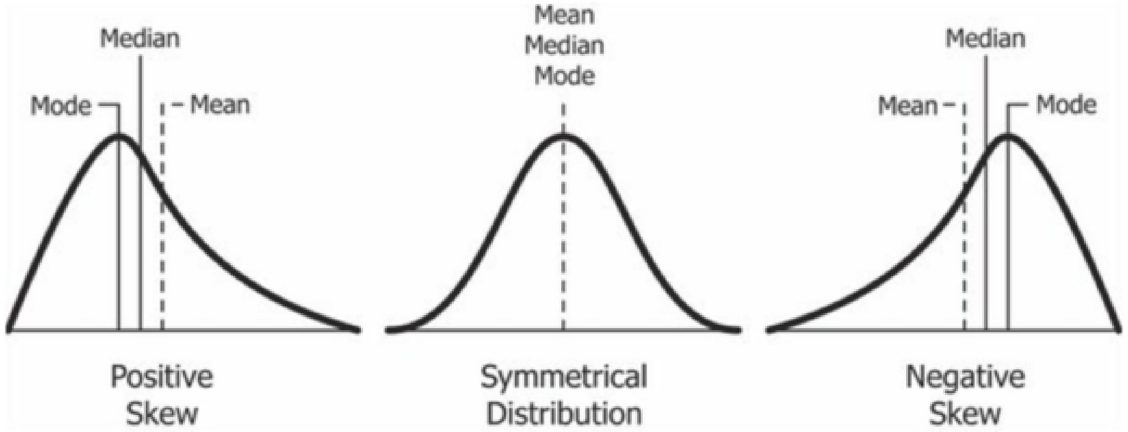
$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Median

It is the middle value in the sorted dataset.

Mean

It is the most frequent value in the dataset.



Source: Measures of Central Tendency

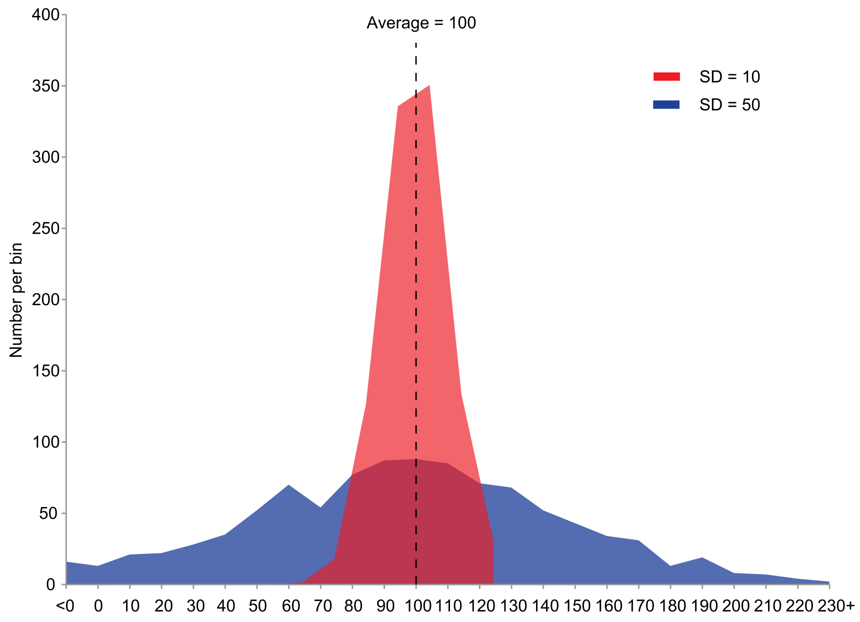
Measures of Dispersion

Variance

It is the measure of spread of the data around its mean.

$$\Omega^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$$



Standard deviation

It is the positive square root of variance.

Standardized value (Z-score)

It is the measure of distance that is **independent of the units of measurement**. It is computed as follows:

$$z_i = \frac{x_i - \mu}{\Omega}$$

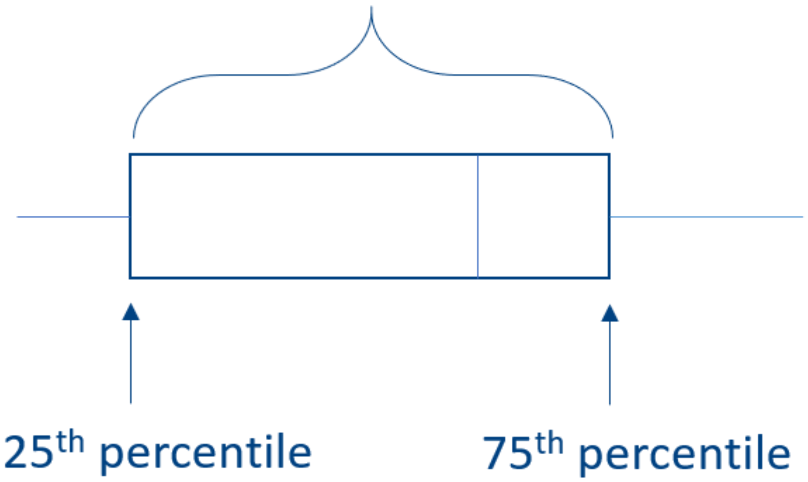
Measures of Dispersion

Range

It is the difference between the maximum and minimum value in the dataset.

Inter-quartile Range (IQR)

It is the is the difference between first and third quartile.



Think!

Which of these measures are affected by outliers in the data?

Measures of Association

Covariance

It is the measure of **linear** association between datasets.

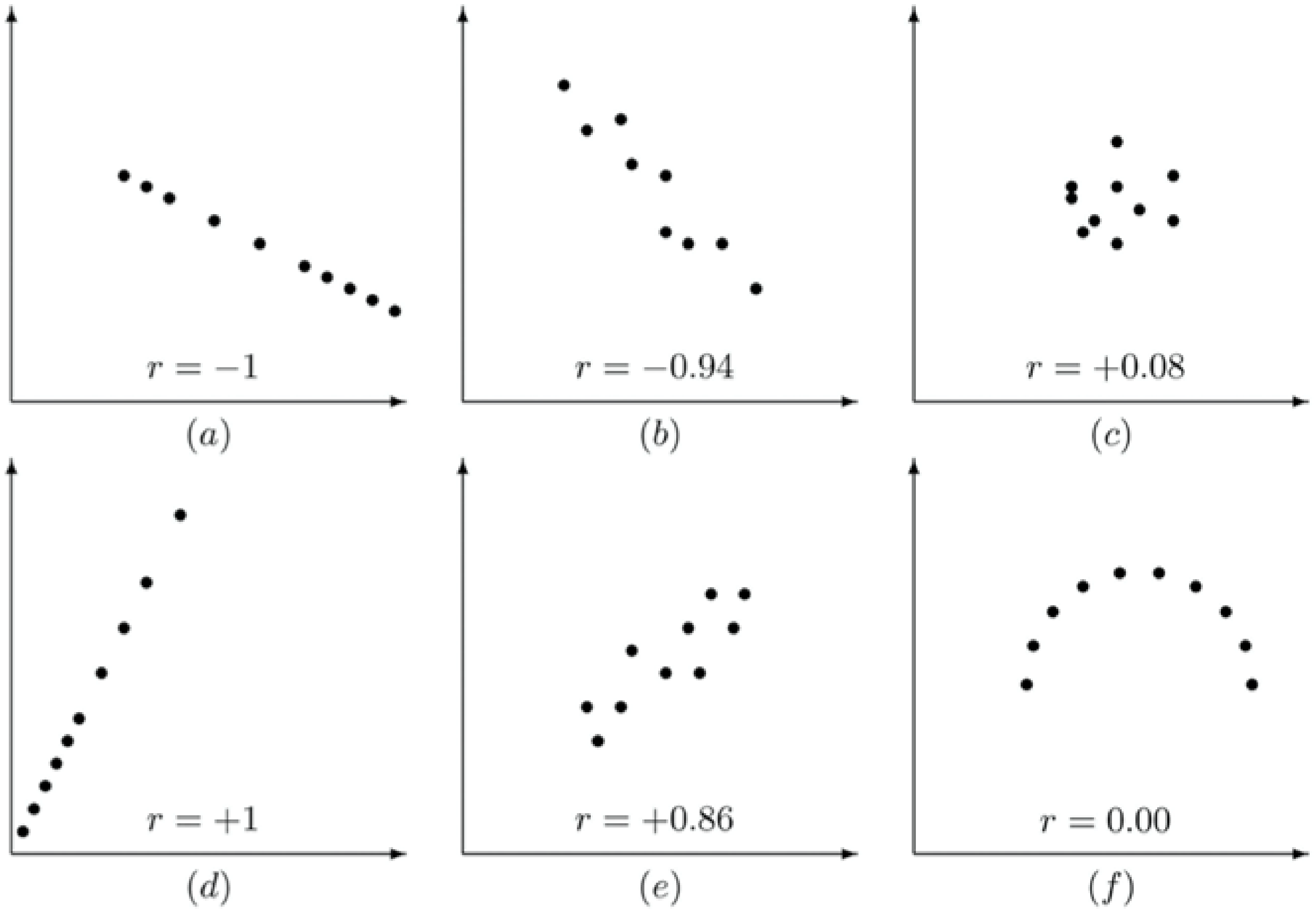
$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{Y})$$

Correlation coefficient

It is the measure of linear association between datasets that is **independent of the units of measurement**.

$$r_{XY} = \frac{cov(X,Y)}{S_X S_y}$$

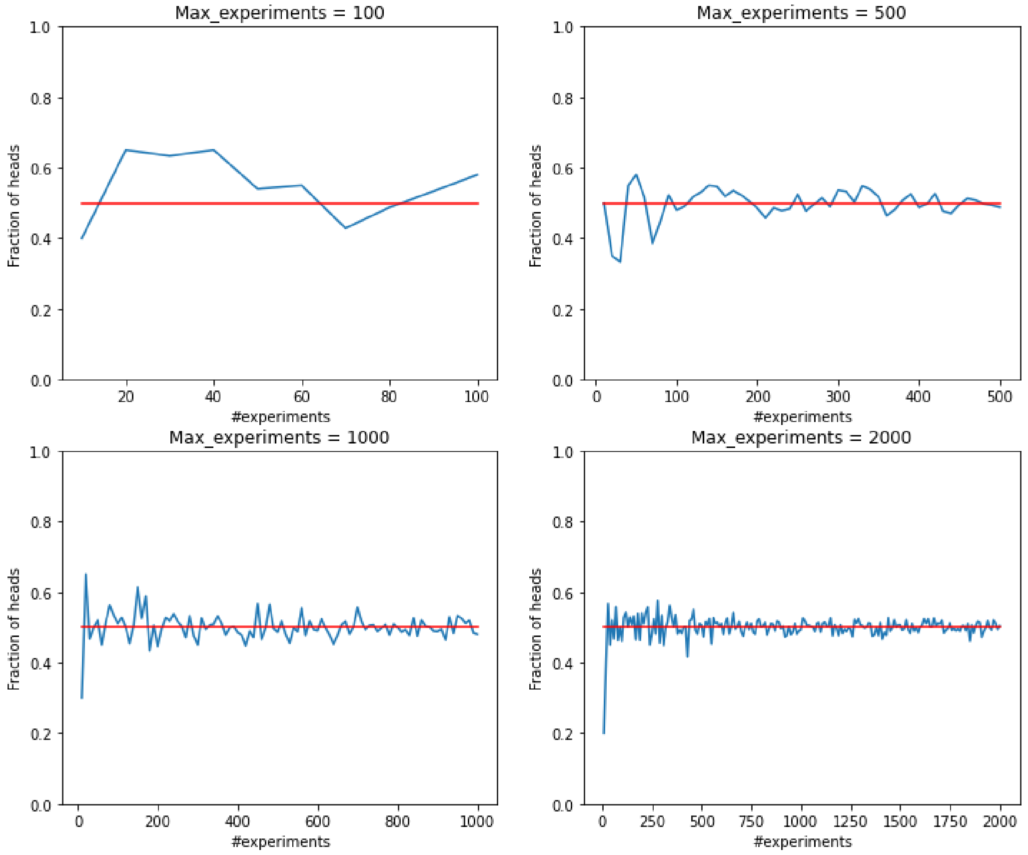
Measures of Association



Probability

From Statistics to Probablity

Experiment. An **unbiased (or fair)** coin is tossed and the number of heads are counted.



Notation

Random Experiment

It is a physical experiment whose outcome can not b predicted until it is performed.

Sample Space (Ω)

It is the set of all possible outcomes of the experiment.

Event (E)

it is any subset of the sample space.

Probability

Probability of any event E is defined as:

$$Pr(E) = \frac{|E|}{|\Omega|}$$

Example

Discrete sample space

Rolling a fair die.	
Sample Sapce	$\Omega = \{1, 2, 3, 4, 5, 6\}$
Event	An even number is rolled.
Probability	$Pr(E) = 3/6 = 0.5$

Example

Discrete sample space

Rolling a fair die.	
Sample Sapce	$\Omega = \{1, 2, 3, 4, 5, 6\}$
Event	An even number is rolled.
Probability	$Pr(E) = 3/6 = 0.5$

Continuous sample space

Rainfall in Singapore on a random day.	
Sample Sapce	$\Omega = [0, 200]$
Event	Low rainfall (≤ 20).
Probability	`How to compute?`

Note

You can't define an event with exact value for a continuous random variable!

Random Variable

Definition

Random variable is a real-valued function defined on the sample space.

Example

Suppose two fair 3-sided dice are rolled. Let X_{sum} denote the random variable that denotes the some of the digits on the dice. Thus,

$$X_{sum} : \Omega \to \{2, 3, 4, 5, 6\}$$

Question

What is the difference between an event and a random variable?

Probability Distribution

Definition

Probability distribution of a random variable is a function that assigns a probability to every possible value that the random variable takes.

Example

Consider the probability distribution of X_{sum}.

x_i	Event	$Pr[X_{sum} = x_i]$
2	$\{(1,1)\}$	1/9
3	$\{(1,2),(2,1)\}$	2/9
4	$\{(2,2),(1,3),(3,1)\}$	3/9
5	$\{(2,3),(3,2)\}$	2/9
6	$\{(3,3)\}$	1/9

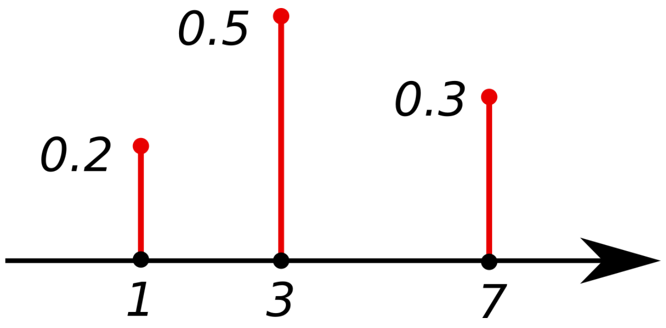
Probability Distribution

Probability mass function

It is a probability distribution function p for a **discrete** random variable X. It satisfies following rules:

$$Pr[X = x_i] = p(x_i) \geq 0.$$

$$\sum_{x_i} Pr[X = x_i] = 1.$$

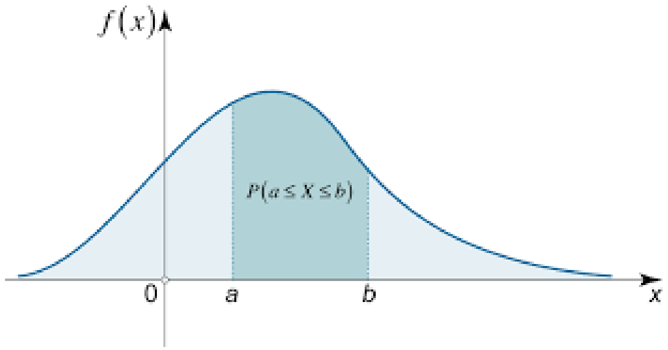


Probability density function

It is a probability distribution function of a **continuous** random variable X. It satisfies following rules:

$$Pr[X \in (a, b)] = \int_a^b p(x)dx \geq 0.$$

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$



Expected Value

Definition

Expected value $E[X]$ of a random variable X is defined as follows.

For discrete random variable. $E[X] = \sum_{x_i} x_i p(x_i)$

For continuous random variable.

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

Mean as a special case

Expected value is also known as *weighted average*. It is equal to the mean for a **uniform** distribution.
(A uniform probability distribution is the probability distribution where all values are equally probable.)

Binomial Distribution

Bernoulli Trial

- It is a random experiment with only two outcomes. It is characterised by a parameter p - the probability of observing one of the two outcomes.
- Examples.
  - Is the visitor going to purchase the product?
  - Is the new applicant a woman?

i.i.d. assumption

In data analytics, data are often assumed to be *i.i.d* samples from the data distribution. *i.i.d* stands for datapoints that independently sampled and identically distributed.

For example: Under *i.i.d.* assumption, we assume that every visitor doesn't come with any bias and has the same purchasing probability. But realistically, this assumption is violated in multiple ways such as visitors who influence each other.

Binomial Distribution

Binomial distribution

- Sum n *i.i.d.* Bernoulli trials is said to follow **Binomial distribution**.

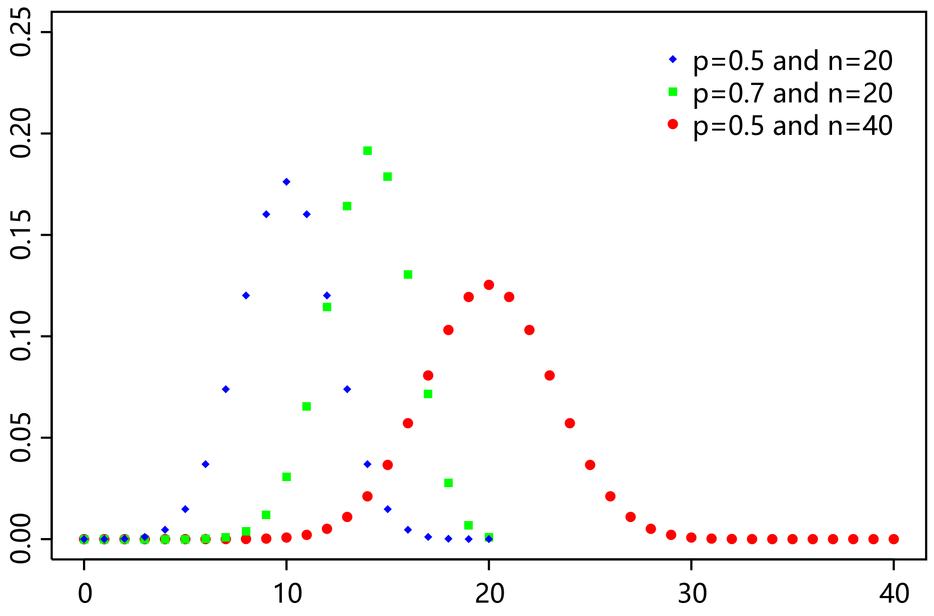
$$X \sim B(n, p)$$

- Probability of observing k *positive* outcomes is computed as follows:

$$Pr[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$$

Central Tendencies

- $\mu = np$
- $\sigma^2 = np(1 - p)$



Examples

- How many of 600 participants actually purchased the products?
- How many of 100 applicants are women?

Multinomial Distribution

Multinomial distribution

- It is an extension of the binomial distribution wherein the experiment may have > 2 outcomes.
- Let us assume that every experiment has $k > 2$ different outcomes. The experiment is performed n times.
- For an outcome i - p_i denotes the probability of observing it and x_i denote the number of times it is observed.

$$Pr[X = (x_1, x_2, ..., x_k)] = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

Examples

- In an election. what is the probability that 40% voted for Part A, 30% voted for party B and the rest for party C?

Gaussian Distribution

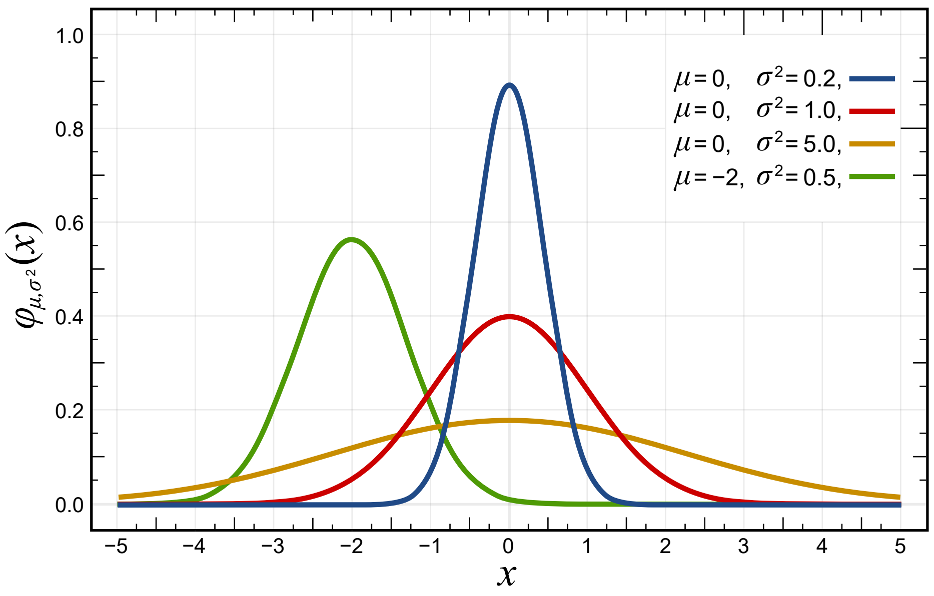
Gaussian distribution

- A random variable X following a gaussian distribution with mean μ and standard deviation σ is denoted as:

$$X \sim \mathrm{N}(\mu, sigma)$$

- The corresponding probability density function takes the following form:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2}$$



$\mathrm{N}(0, 1)$ is known as **standard normal** distribution.

Exercise

Example 1.

The probability that a sales representative makes a sale over marketing call is 0.15.

- What is the probability that no sales are made in 10 calls?
- What is the probability that more than three sales are made in 20 calls?
- If the representative makes 20 calls in a day, how sales are made with 95% probability?
- Find the least number of calls the representative should make to do 5 sales on an average per day.

Can we solve it using Python?

Exercise

Python cheatsheet

`scipy.stats` contains a large number of probability distributions, summary and frequency statistics, correlation functions. For instance:

```
z = norm(2, 3)      # creates a rv that follows Gaussian distribution
b = binom(10, 0.15) # created a rv that follows Binomial distribution
```

Every distribution supports following function:

```
z.pdf(x)    # computes the probability density for the specified point x
z.cdf(x)    # computes the cumuluative density for the specified point x
z.ppf(p)    # computes the inverse CDF for the specified probability p
z.expect(f) # computes the expected value for the specified function f
```

Central Limit Theorem

Population vs Sample (Revisited!)

Consider the following dataset of five students with their test score out of 5.

Population ($\mu = 2.6$)	
A	3
B	4
C	2
D	3
E	1

Population vs Sample (Revisited!)

Consider the following dataset of five students with their test score out of 5.

Population ($\mu = 2.6$)	
A	3
B	4
C	2
D	3
E	1

Now we will take samples of 3 such and study the *distribution* of their sample mean.

Sample mean	Samples
2.00	$\{ACE, DCE\}$
2.33	$\{ADE, BCE\}$
2.67	$\{ABE, DBE, ACD\}$
3.00	$\{ABC, DBC\}$
3.33	$\{ABD\}$

Population vs Sample (Revisited!)

Probability distribution of the sample mean

We can treat sample mean as a random variable based on the randomness introduced by the sampling procedure.

x_i	$Pr[\bar{X} = x_i]$
2.00	2/10
2.33	2/10
2.67	3/10
3.00	2/10
3.33	1/10

Population vs Sample (Revisited!)

› Motivation
Sampling Distribution
Central Limit Theorem
Confidence Interval

Probability distribution of the sample mean

We can treat sample mean as a random variable based on the randomness introduced by the sampling procedure.

x_i	$Pr[\bar{X} = x_i]$
2.00	2/10
2.33	2/10
2.67	3/10
3.00	2/10
3.33	1/10

Expected value of the sample mean

$$E[\bar{X}] = \sum_{x_i} x_i Pr[\bar{X} = x_i] = 2.6$$

Isn't it same as the population mean?

Sampling Distribution

Sampling Distribution

Every random-sampling based statistic follows the probability distribution called as the **sampling distribution**.

Sample mean

Let us consider a population of datapoints with mean μ and standard deviation σ. We take multiple samples of size n from the population. The sample of mean $\bar{X}$ follows the sampling distribution with expected value and `standard error`

$$E[\bar{X}] = \mu$$

$$std(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution

Sampling Distribution

Every random-sampling based statistic follows the probability distribution called as the **sampling distribution**.

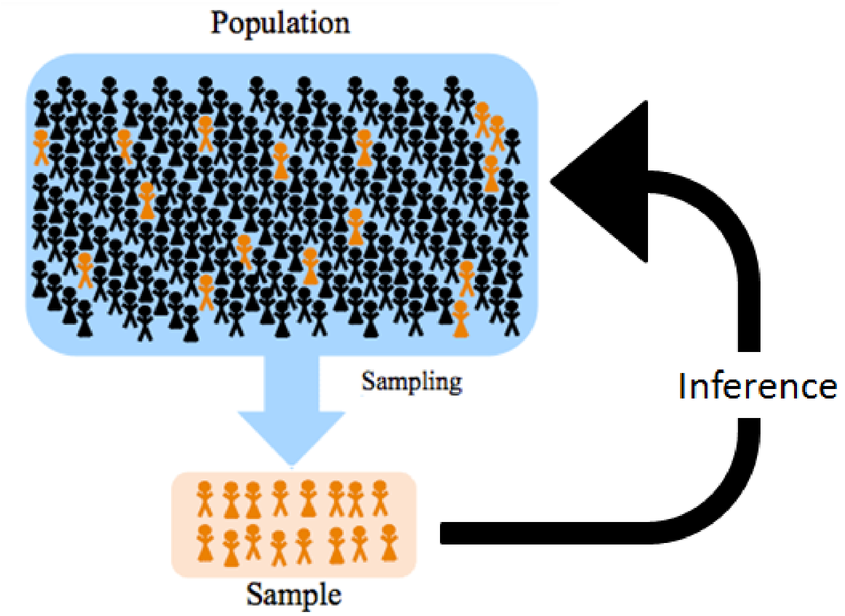
Sample mean

Let us consider a population of datapoints with mean μ and standard deviation σ. We take multiple samples of size n from the population. The sample of mean $\bar{X}$ follows the sampling distribution with expected value and `standard error`

$$E[\bar{X}] = \mu$$

$$std(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Why do we study statistics?



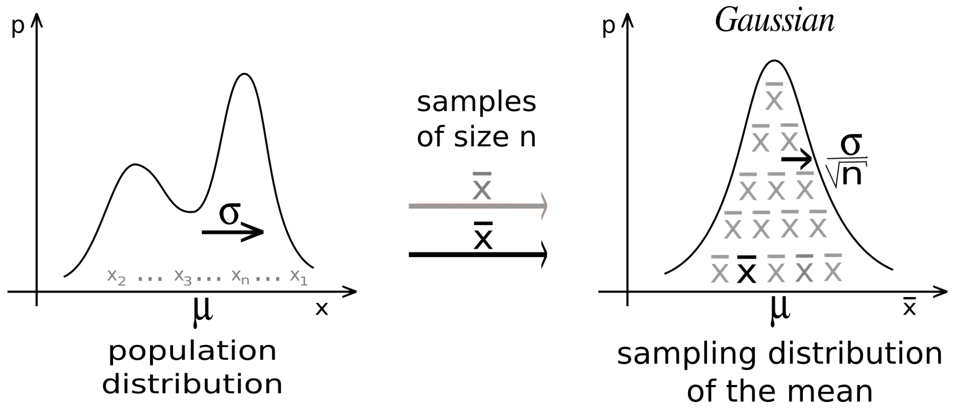
Source: Towards Data Science

Central Limit Theorem

Theorem

The sampling distribution of the **sample mean** any **sufficiently large** samples drawn from a population with mean μ and standard deviation σ follows standard normal distribution with mean μ and standard deviation $\mu/\sqrt{n}$.

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



What is a sufficiently large sample size?

Theoretically, the sampling distribution tends to the Gaussian distribution as the sample size increases ($n \to \infty$).
Practically, it has been observed that sample size of 30 or more is sufficient for central limit theorem to hold.

Confidence Interval

Example 2: Average salary

As a student of IT5006, your first assignment is to conduct a survey and find the average salary of graduates at NUS. What will be your methodology? What will be the underlying assumptions?

1. Assume that all graduates salaries come from an unknown distribution with mean μ and standard deviation σ.
2. Since, you can not interview every single graduate, ***randomly*** choose a *sufficiently large* sample.
3. Compute the average of the sample $\bar{X}$.

What to do now? Is this the true average?

Confidence Interval

Interval estimation

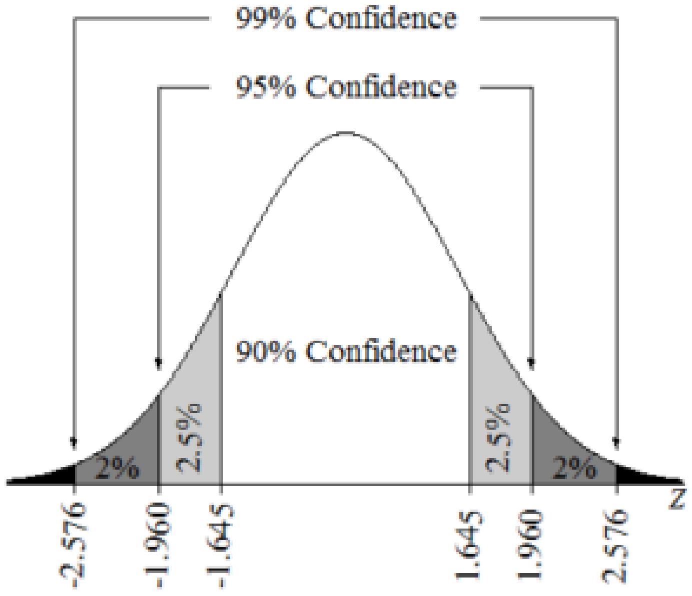
$[a, b]$ is said to be $(1 - \alpha) * 100\%$ confidence interval of a random variable X if and only if

$$Pr[a \leq X \leq b] = (1 - \alpha)$$

CI Simulator

Question

How do we compute it for standard normal distribution?



Interval	Confidence
$[-2.57, 2.57]$	99%
$[-1.96, 1.96]$	95%
$[-1.64, 1.64]$	90%

Confidence Interval

Motivation
Sampling Distribution
Central Limit Theorem
❯ Confidence Interval

Going back to Example 2

1. Let's assume that we look at the historical employment data and find the standard deviation of the graduate salaries in general and **assume** it to be σ.

2. We can provide the 95% confidence interval for the graduate salaries as follows:

$$Pr[-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96] = 95$$

.

$$Pr[(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}) \leq \mu \leq (\bar{X} + 1.96\frac{\sigma}{\sqrt{n}})] = 95$$

.

Hypothesis Testing

Motivation

Example 3: Average Purchase Value

Suppose an e-commerce company wants to investigate whether the average purchase value on their website is significantly different from a target value of 50 dollars. They collect a random sample of 100 purchase transactions and calculate the mean purchase value to be 52.50 dollars, with a known population standard deviation of 8.00 dollars.

What can they say about target value? Is it scientifically valid to say that the platform offers higher cost?

Motivation

Example 3: Average Purchase Value

Suppose an e-commerce company wants to investigate whether the average purchase value on their website is significantly different from a target value of 50 dollars. They collect a random sample of 100 purchase transactions and calculate the mean purchase value to be 52.50 dollars, with a known population standard deviation of 8.00 dollars.

What can they say about target value? Is it scientifically valid to say that the platform offers higher cost?

Motivation

Example 3 (solution): Average Purchase Value

1. Let's assume that $\mu = 50$.
2. As per the given information, $\bar{X} = 52.5$, $n = 50$ and $\sigma = 8$.
3. Let's construct 95% confidence interval around the mean.

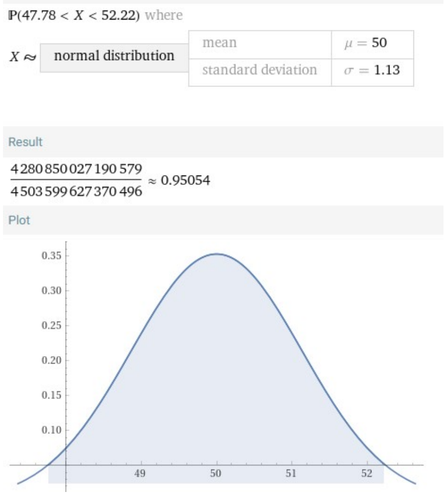
$$Pr[(50 - 1.96\frac{8}{\sqrt{50}}) \leq \bar{X} \leq (50 + 1.96\frac{8}{\sqrt{50}})] = 95$$

.

$$Pr[47.78 \leq \bar{X} \leq 52.22] = 95$$

.

We can reject the assumption that average cost is 50 with 95% confidence.



Hypothesis Testing

Null Hypothesis H_0

- Something that is already established.
- Something that you want to challenge.

Alternate Hypothesis H_1

- Something that you want to assess.
- Something that challenges the current establishment.

	H_0 is True	H_0 is False
Accept H_0	No Error	Type II error (β)
Reject H_0	Type I Error (α)	No Error

Hypothesis Testing

How is the Type I error related to the confidence interval?

Let's say: H_0: $\mu = \mu_0$ and we set Type I error to be 5%.

- We construct the 95% confidence interval around μ_0.
- If the observed mean $\bar{X}$ lies outside the interval
  - We reject H_0.
  - In doing so, we would have committed an error of 5%.
- If the observed mean $\bar{X}$ lies within the interval
  - We do not sufficient evidence to reject H_0.
  - We are 95% confident H_0 is correct.

p-value

Definition

It is the probability of observing results as extreme as the observed, under the assumption that the null hypothesis is correct.

- In simple words p-value computes the error of commiting error.
- If p-value is less than or equals to 0.05
  - We reject H_0.
  - It means that the error is within the tolerance.
- If p-value is more than 0.05
  - We do not have sufficient evidence to reject H_0

Motivation

Example 3 (p-value solution): Average Purchase Value

1. Let's assume that $\mu = 50$.
2. As per the given information, $\bar{X} = 52.5$, $n = 50$ and $\sigma = 8$.
3. Let's compute the p-value. To do so, let's compute the probability that the output to be worse that the observed.

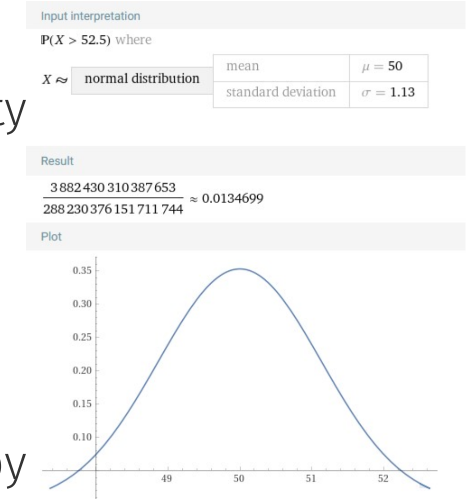
$$Pr[\bar{X} > 52.5] = 0.013$$

.

4. Since we are checking the alternate hypothesis that $\mu \neq 50$, by symmetry

$$p - value = 2 * 0.013 = 0.026$$

5. Since p-value is less than 0.05, we can reject the null hypothesis.



Two ways to solve HT

Using the confidence interval

1. Setup the null hypothesis.
2. Assume Type I error tolerance α.
3. *Choose the sampling distribution.*
4. Construct a $(1 - \alpha) * 100\%$ confidence interval under the null hypothesis.
5. If
   - **Observation lies outside CI.** Reject H_0.
   - **Otherwise.** No sufficient evidence to reject H_0.

Two ways to solve HT

Using the confidence interval

1. Setup the null hypothesis.
2. Assume Type I error tolerance α.
3. *Choose the sampling distribution.*
4. Construct a $(1 - \alpha) * 100\%$ confidence interval under the null hypothesis.
5. If
   - **Observation lies outside CI.** Reject H_0.
   - **Otherwise.** No sufficient evidence to reject H_0.

Using p-value

1. Setup the null hypothesis.
2. Assume Type I error tolerance α.
3. *Choose the sampling distribution.*
4. Compute p-value.
5. If
   - $p \leq \alpha$. Reject H_0.
   - **Otherwise.** No sufficient evidence to reject H_0.

One-sided test

Example 4: Average Purchase Value

Suppose an e-commerce company wants to investigate whether the average purchase value on their website is ~~significantly different~~ more than the target value of 50 dollars. They collect a random sample of 50 purchase transactions and calculate the mean purchase value to be 52.50 dollars, with a known population standard deviation of 8.00 dollars.

In this case we can set up the hypotheses as follows:

$$H_0: \mu \leq 50$$

$$H_1: \mu > 50$$

One-sided test

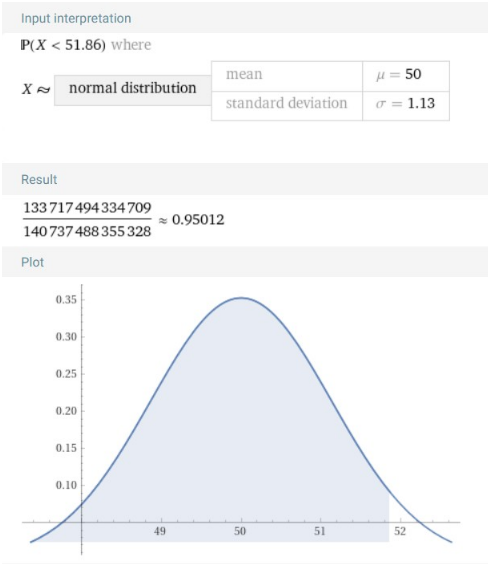
Example 4 (Solution CI): Average Purchase Value

We need to construct a one-sided 95% confidence interval around 50. To do so,

$$Pr[\frac{\bar{X} - 50}{8/\sqrt{50}} \leq 1.64] = 0.95$$

$$Pr[\bar{X} \leq 51.86] = 0.95$$

.



Since 52.5 lies outside the CI, we reject the null hypothesis.

One-sided test

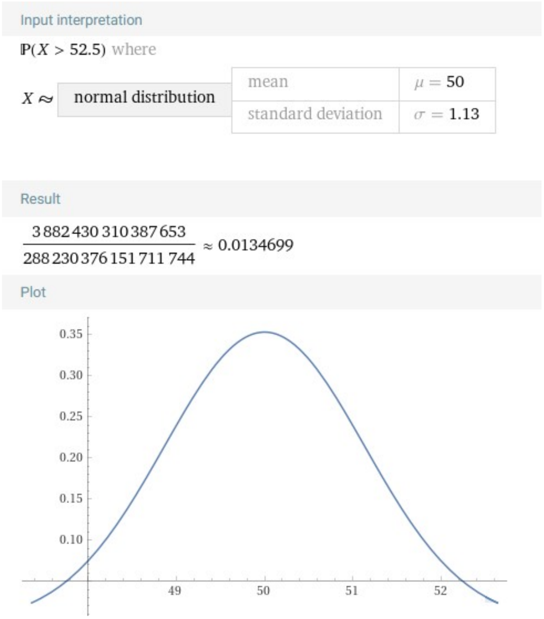
Example 4 (Solution p-value): Average Purchase Value

We had previously computed the error as follows:

$$Pr[\bar{X} > 52.5] = 0.013$$

.

Since the p-value is less than 0.05, we reject the null hypothesis.



z-test

- The tests that we have performed so far are known as z-tests.
- The name stems from the use of standard normal distribution as the sampling distribution.

It's easy with Python!

- z-test using Python

t-test

Example 5:

Suppose an e-commerce company wants to investigate whether the average purchase value on their website is significantly different more than the target value of 50 dollars. They collect a random sample of ~~50~~ 22 purchase transactions and calculate the mean purchase value to be 52.50 dollars, with a known population standard deviation of 8.00 dollars.

Question

What is the sampling distribution of the mean for the sample size smaller than 30?

One sample t-test

Example 5 (solution):

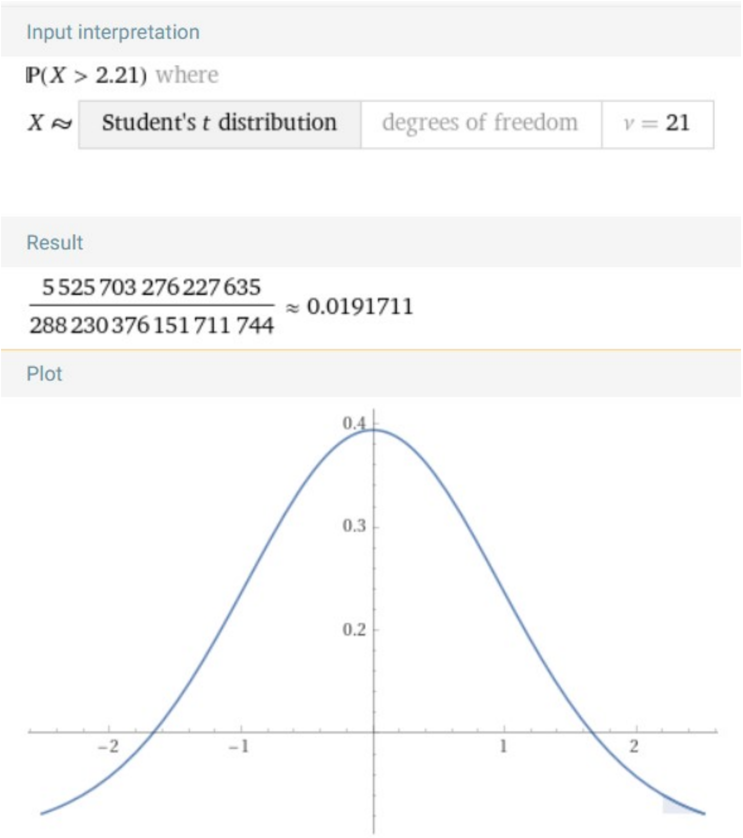
Let's compute the t i.e.

$$\frac{52.5 - 50}{8/\sqrt{22}} = 2.21$$

Thus, the p-value is $2 * 0.019 = 0.038$.

We reject the null hypothesis.

> **Think!**
>
> Earlier with a sample of 50, the p-value was 0.026. With the sample of 22, it has become 0.038. Does this make sense?



Summary

Summary

Descriptive Statistics

- Central tendencies

Probability

- Formalism
- Probability distribution

Central Limit Theorem

- Interval Estimation

Hypothesis Testing

- Type I and Type II Error
- Some popular tests

Thank you!

Feel free to reach out to me at

dcsashi (at) nus (dot) edu (dot) sg