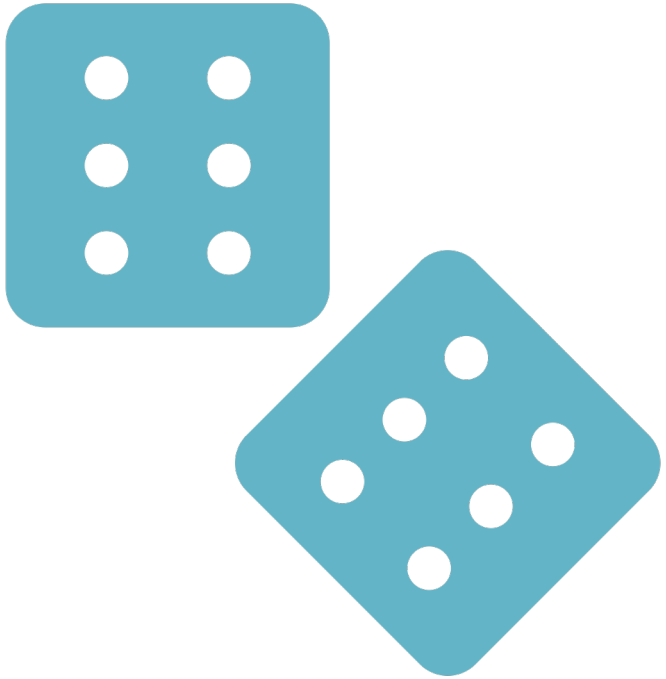
PROBABILITY

HAFIZAH AB RAHIM

AUGUST 15, 2020



CONTENT

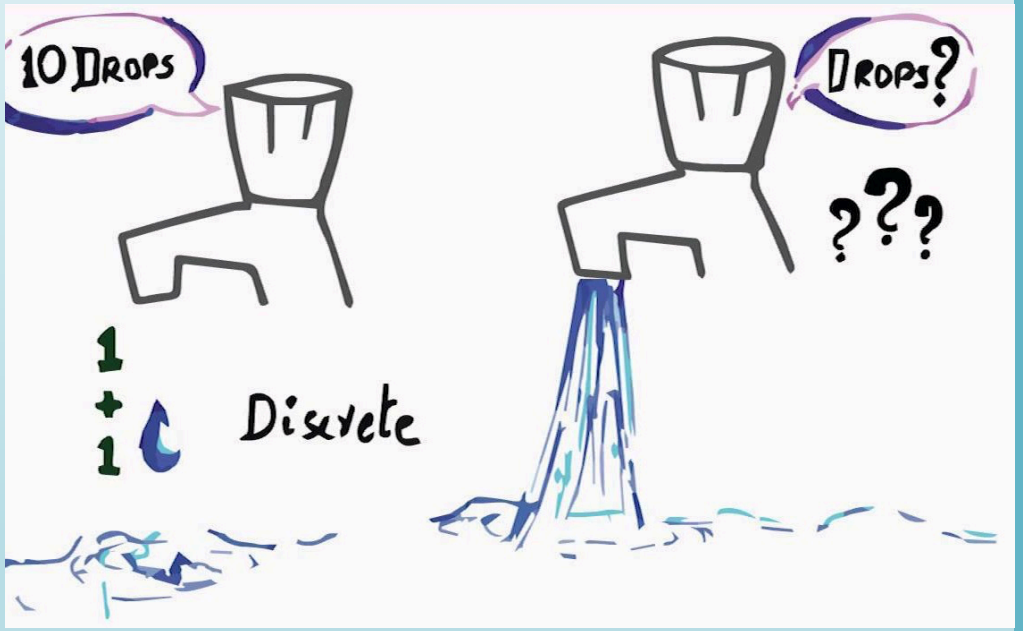
1. Random Variables
2. Discrete Probability
3. Continuous Probability
4. Central Limit Theorem
5. Concepts Application - Interest Rate

RANDOM VARIABLES

WHAT IS A RANDOM VARIABLE?

- Random variables are numeric outcomes resulting from random processes.
- A random variable can be either discrete (having specific values like 4 ft and 10 ft) or continuous (infinite number of possible values like 7.3 ft and 5.5 ft)
- Random variables are used by Risk Analysts to quantify outcomes of a random occurrence in finance.

The diagram visualizes a comparison between discrete and continuous variables



WHAT IS PROBABILITY?

- Mathematics is the logic of certainty. Probability is the logic of uncertainty.
- Probability (P) measures the chance for an event occurring.
- Probability takes values between zero and one (inclusive).
- The sum of all probabilities of an event is one.
- Probability is used in statistics, machine learning, artificial intelligence, finance and many more areas.

An event is an outcome that can occur when something happens by chance.

Example

Question

One ball will be drawn at random from a box containing: 3 cyan balls, 5 magenta balls, and 7 yellow balls.

What is the probability that the ball will be cyan?

Answer

script.R

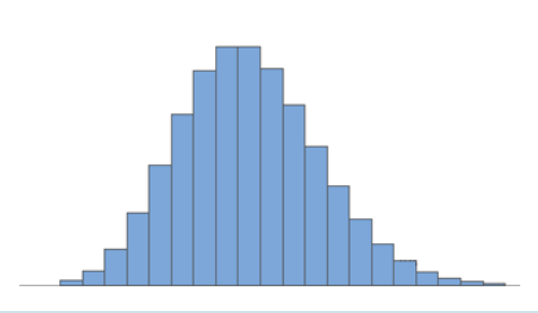
```
1  cyan <- 3
2  magenta <- 5
3  yellow <- 7
4
5  # Assign a variable `p` as the probability of choosing a cyan ball from the box
6  p= cyan/(cyan + magenta + yellow)
7  # Print the variable `p` to the console
8  p
```

```
> p
[1] 0.2
```

PROBABILITY DISTRIBUTION

- Distribution describes the probability of observing each possible outcome.

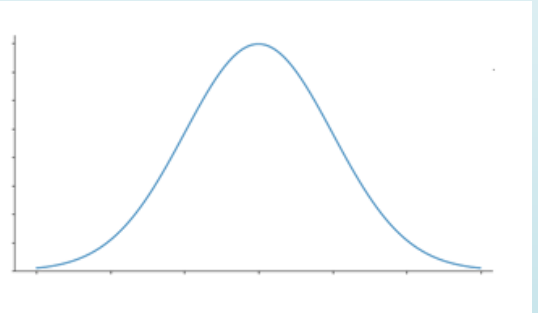
There are **TWO** types of probability distributions



Examples of Discrete Distribution

Bernoulli
Binomial
Uniform
Poisson

Discrete Distributions	Continuous Distribution
Discrete distributions have finite number of different possible outcomes	Continuous distributions have infinite many consecutive possible values
We can add up individual values to find out the probability of an interval	We cannot add up individual values to find out the probability of an interval because there are many of them
Discrete distributions can be expressed with a graph, piece-wise function or table	Continuous distributions can be expressed with a continuous function or graph
In discrete distributions, graph consists of bars lined up one after the other	In continuous distributions, graph consists of a smooth curve



Examples of Continuous Distribution

Normal
Chi-Squared
Exponential
Logistic
Students' T

DISCRETE PROBABILITY

MONTE CARLO SIMULATION

- The Monte Carlo simulation is used **to model the probability** of different outcomes by repeating a random process many times that the result is comparable to what would be observed if the process were repeated forever.

- This simulation is intensively used to **forecast stock market** and **portfolio optimization**.

Example

Probability of the Celtics winning a game – Without using Monte Carlo simulation

QUESTION

Two teams, say the Celtics and the Cavs, are playing a seven game series. The Cavs are a better team and have a 60% chance of winning each game.

What is the probability that the Celtics win at least one game? Remember that the Celtics must win one of the first four games, or the series will be over!

ANSWER

script.R

```
1  # Assign the variable `p_cavs_win4` as the probability that the Cavs will win the first four
   games of the series.
2  p_cavs_win4 <-0.6^4
3
4  # Using the variable `p_cavs_win4`, calculate the probability that the Celtics win at least one
   game in the first four games of the series.
5  1- p_cavs_win4
```

```
> 1- p_cavs_win4
[1] 0.8704
```

Example

Probability of the Celtics winning a game – using Monte Carlo simulation

ANSWER

script.R

```r
# This line of example code simulates four independent random games where the Celtics either lose or win. Copy this example code to use within the `replicate` function.
simulated_games <- sample(c("lose","win"), 4, replace = TRUE, prob = c(0.6, 0.4))

# The variable 'B' specifies the number of times we want the simulation to run. Let's run the Monte Carlo simulation 10,000 times.
B <- 10000

# Use the `set.seed` function to make sure your answer matches the expected result after random sampling.
set.seed(1)

# Create an object called `celtic_wins` that replicates two steps for B iterations: (1) generating a random four-game series `simulated_games` using the example code, then (2) determining whether the simulated series contains at least one win for the Celtics.
celtic_wins <-replicate(B, {
simulated_games <- sample(c("lose","win"), 4, replace = TRUE, prob = c(0.6, 0.4))
any(simulated_games=="win")
})
# Calculate the frequency out of B iterations that the Celtics won at least one game. Print your answer to the console.
mean(celtic_wins)
```

```
> mean(celtic_wins)
[1] 0.8757
```

COMBINATIONS

- The various ways objects from a set can be arranged so that the order of selection DOES NOT matter.

$$C(n,r) = {}^{n}C_{r} = \frac{n!}{(n-r)!r!}$$

Number of items in set

Number of items selected from the set

Combinations nCr Calculator

PERMUTATIONs

$$C(n,r) = \binom{n}{r} = \frac{n!}{(r!(n-r)!)} = ?$$

n choose r

n *(objects)* = 6

r *(sample)* = 2

Clear | Calculate

Answer:

= 15

Solution:

$$C(n,r) = ?$$

$$C(n,r) = C(6,2)$$

$$= \frac{6!}{(2!(6-2)!)}$$

= 15

Example Restaurant Management

Question

A restaurant manager wants to advertise that his lunch special offers enough choices to eat different meals every day of the year. He doesn't think his current special actually allows that number of choices, but wants to change his special if needed to allow at least 365 choices.

A meal at the restaurant includes 1 entree, 2 sides, and 1 drink. He currently offers a choice of 1 entree from a list of 6 options, a choice of 2 different sides from a list of 6 options, and a choice of 1 drink from a list of 2 options.

Answer

How many meal combinations are possible with the current menu?

180 ✔

6C1 * 6C2 *2C1
= 6 * 15 * 2

PERMUTATIONS

- Various ways objects from a set can be arranged in which order of selection matters.

$$P(n,r) = \frac{n!}{(n-r)!}$$

Example Olympic Running

Question In the 200m dash finals in the Olympics, 8 runners compete for 3 medals (order matters). In the 2012 Olympics, 3 of the 8 runners were from Jamaica and the other 5 were from different countries. The three medals were all won by Jamaica (Usain Bolt, Yohan Blake, and Warren Weir).

Answer How many different ways can the 3 medals be distributed across 8 runners?

336 ✔

Permutations nPr Calculator

$$P(n,r) = \frac{n!}{(n-r)!} = ?$$

n *(objects)* = 8

r *(sample)* = 3

Clear Calculate

Answer:

$$P(n,r) = P(8,3)$$

$$= \frac{8!}{(8-3)!}$$

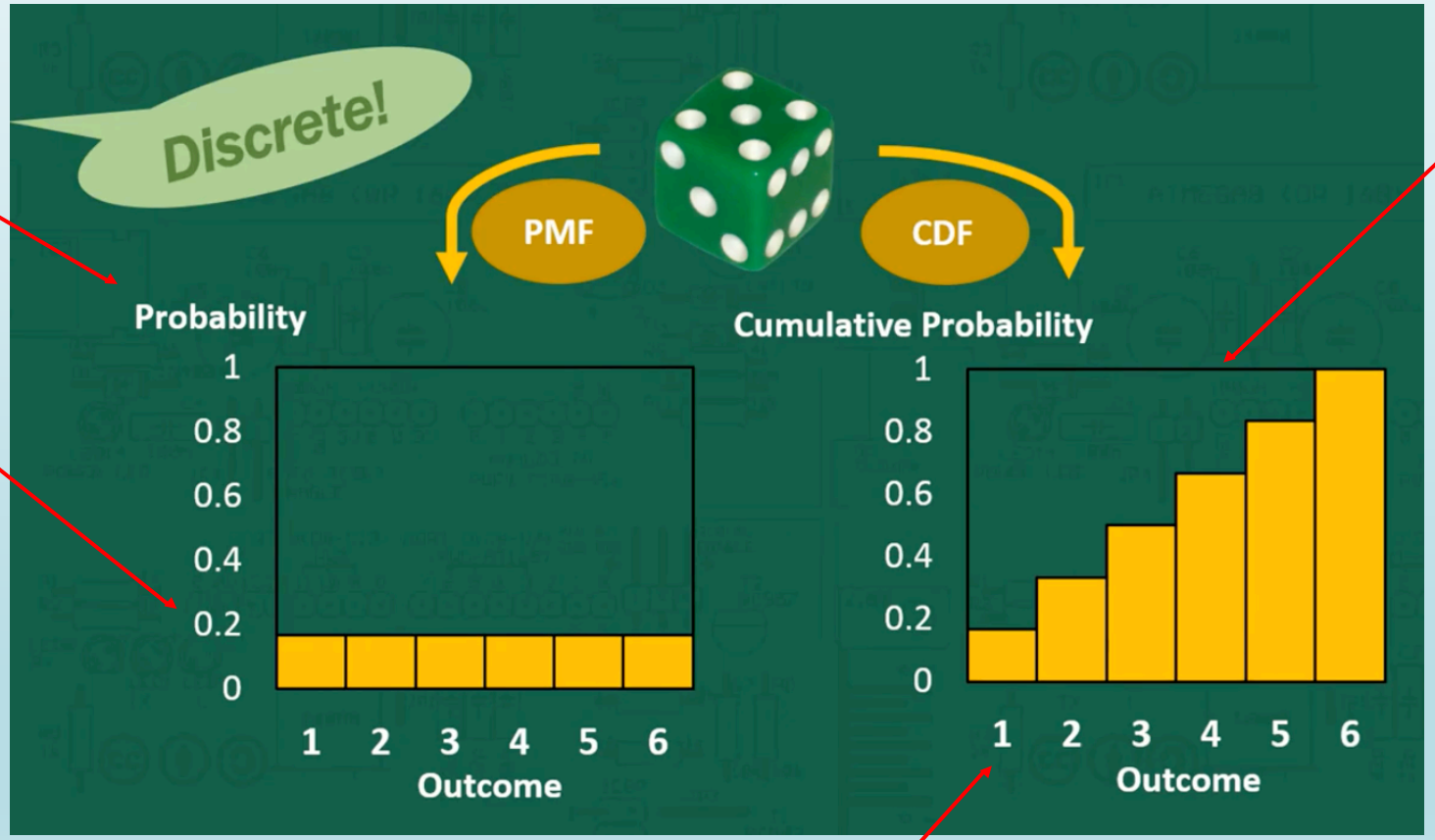
$$= 336$$

PROBABILITY MASS FUNCTION (PMF)

- PMF shows the probabilities of discrete random variables.

Shows the PMF of a rolling dice.

P(x=1) =1/6 = **0.167**



Graph shows the Cumulative Density Function (CDF) of a rolling dice.

CDF is calculated by continuously adding the value of the probability before the targeted outcome.

To find the probability of dice rolling less than or equal to 3:

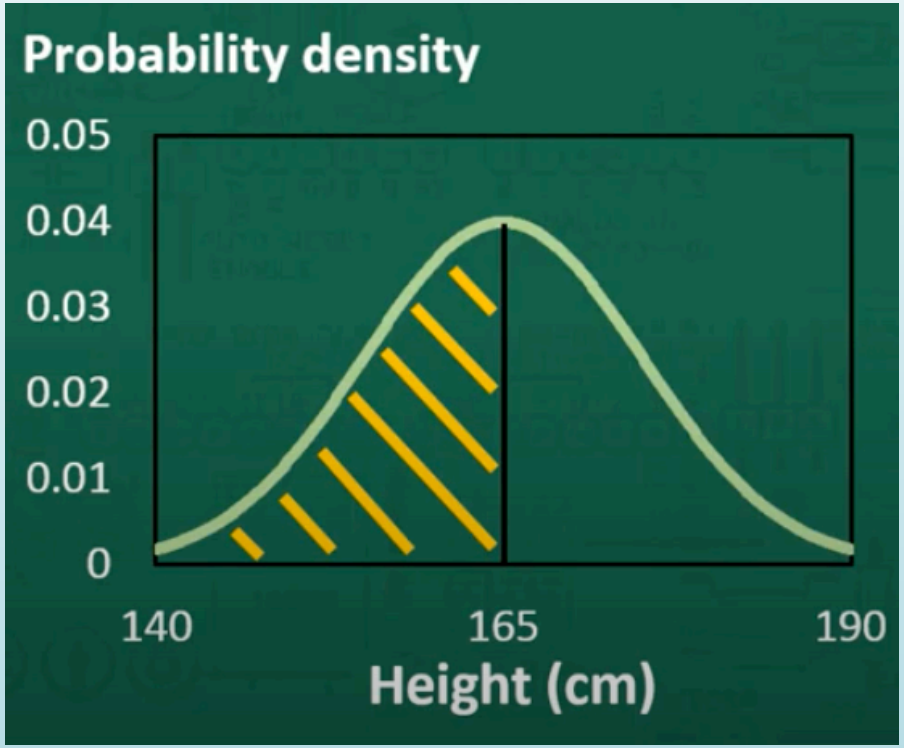
$P(x \leq 3) = P(x=1) + P(x=2) + P(x=3)$

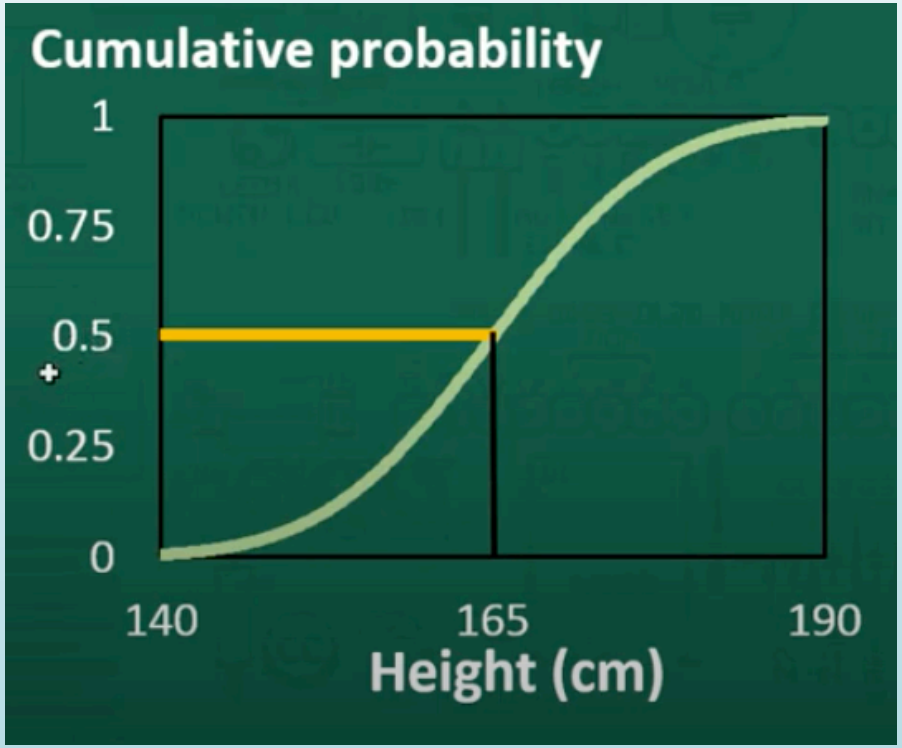
$P(x=2) = 1/6 + 1/6 =$ **0.333**

CONTINUOUS PROBABILITY

PROBABILITY DENSITY FUNCTION (PDF)

- PDF shows the probability of continuous distribution.

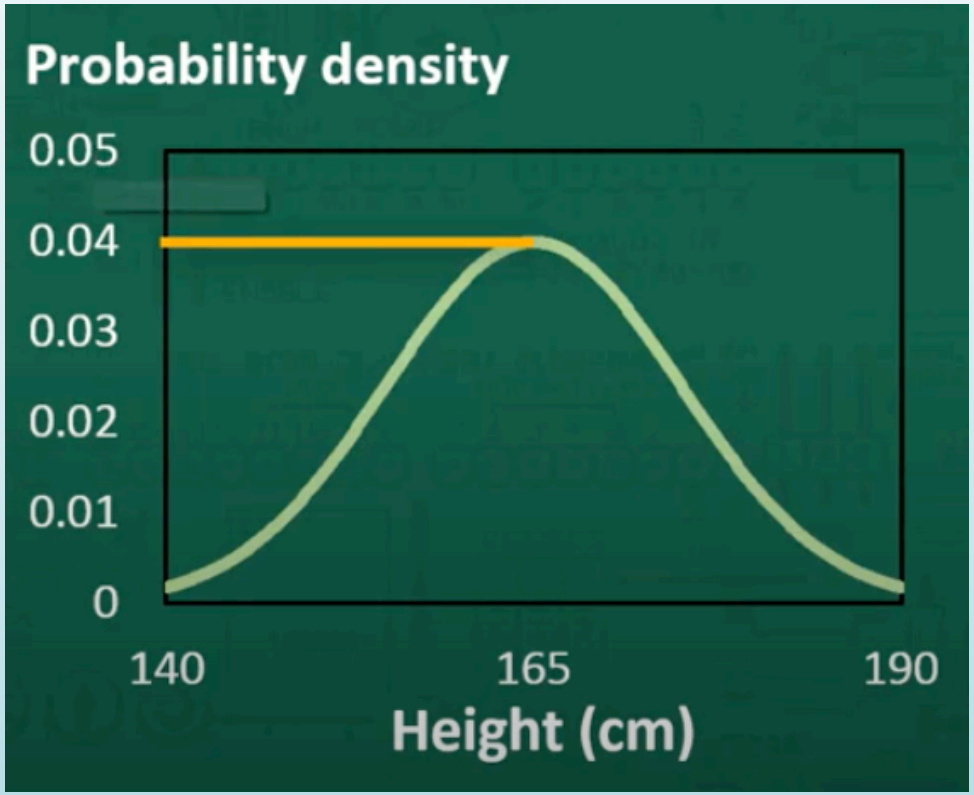


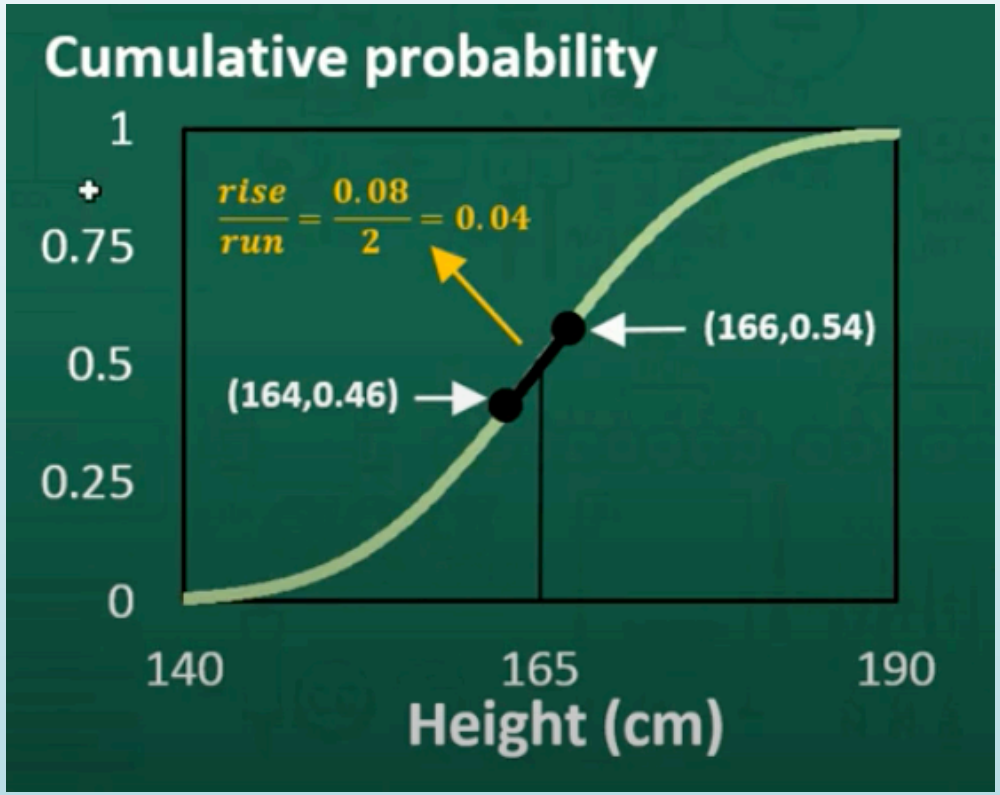


The graph above shows the distribution of female heights with a mean of 165 cm. As you approach towards -∞ and +∞, the probability of certain heights decreases.

One half of the area under PDF is 0.5 which can be shown on the CDF as well.

- The gradient of CDF at 165 cm is 0.04 which is also shown on the y-axis of the PDF.





- Looking at CDF, the gradient of the s-curve increases and then decreases after 165 cm. This pattern is also seen on the probability of CDF.

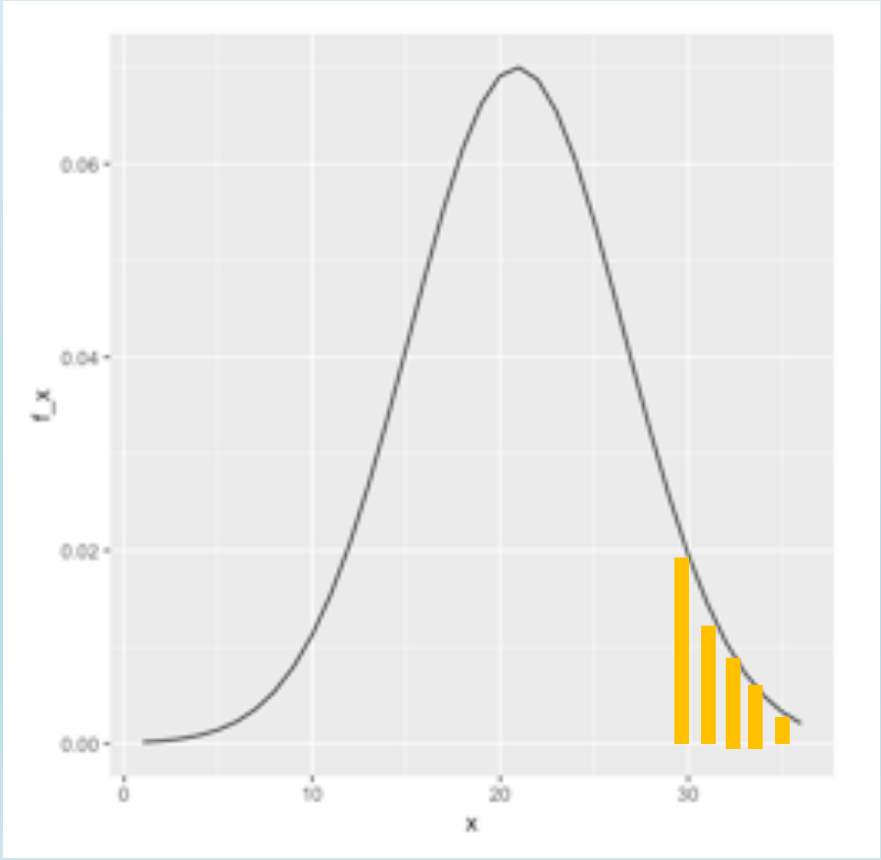
Example

ACT Scores

The ACT is a standardized college admissions test used in the United States. The four multi-part questions in this assessment all involve simulating some ACT test scores and answering probability questions about them.

For the three year period 2016-2018, ACT standardized test scores were approximately normally distributed with a mean of 20.9 and standard deviation of 5.7. (Real ACT scores are integers between 1 and 36, but we will ignore this detail and use continuous values instead.)

Set the seed to 16, then use `rnorm()` to generate a normal distribution of 10000 tests with a mean of 20.9 and standard deviation of 5.7.



In `act_scores`, what is the probability of an ACT score greater than 30?

0.0527 ✔

```
> B<-1000
> set.seed(16, sample.kind = "Rounding")
Warning message:
In set.seed(16, sample.kind = "Rounding") :
  non-uniform 'Rounding' sampler used
> act_scores<-rnorm(10000,20.9,5.7)
> mean(act_scores)
[1] 20.84012
> sd(act_scores)
[1] 5.675237
> mean(act_scores > 30)
[1] 0.0527
```

Example

Question

In `act_scores`, what is the probability of an ACT score less than or equal to 10?

0.0279 ✔

Question

What is the probability of a Z-score greater than 2 (2 standard deviations above the mean)?

0.02275 ✔

Question

What ACT score value corresponds to 2 standard deviations above the mean (Z = 2)?

32.26 ✔

Question

Use `qnorm()` to determine the expected 95th percentile, the value for which the probability of receiving that score or lower is 0.95, given a mean score of 20.9 and standard deviation of 5.7.

What is the expected 95th percentile of ACT scores?

30.276 ✔

```
> qnorm(0.975, 20.9, 5.68)
[1] 32.0326
```

Z-score = 2

Probability of x<32.26: 0.97725

Probability of x>32.26: 0.02275

Probability of 20.9<x<32.26: 0.47725

Steps:

$$\text{Z score} = \frac{x - \mu}{\sigma} = \frac{32.26 - 20.9}{5.68} = 2$$

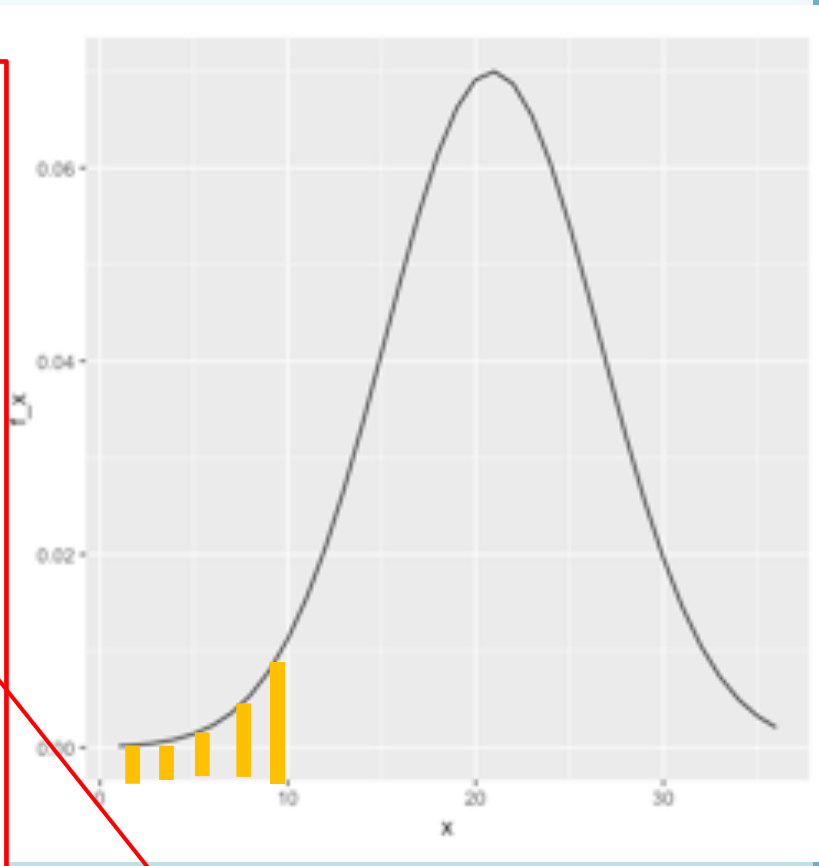
P-value from Z-Table:

P(x<32.26) = 0.97725

P(x>32.26) = 1 - P(x<32.26) = 0.02275

P(20.9<x<32.26) = P(x<32.26) - 0.5 = 0.47725

Raw Score, **x**	32.26
Population Mean, **μ**	20.9
Standard Deviation, **σ**	5.68



Z-score estimates how far the data point from the mean. It measures how many standard deviations below or above the population mean.

Example

Question

Make a vector containing the quantiles for `p <- seq(0.01, 0.99, 0.01)`, the 1st through 99th percentiles of the `act_scores` data. Save these as `sample_quantiles`.

In what percentile is a score of 26?

Your answer should be an integer (i.e. 60), not a percent or fraction. Note that a score between the 98th and 99th percentile should be considered the 98th percentile, for example, and that quantile numbers are used as names for the vector `sample_quantiles`.

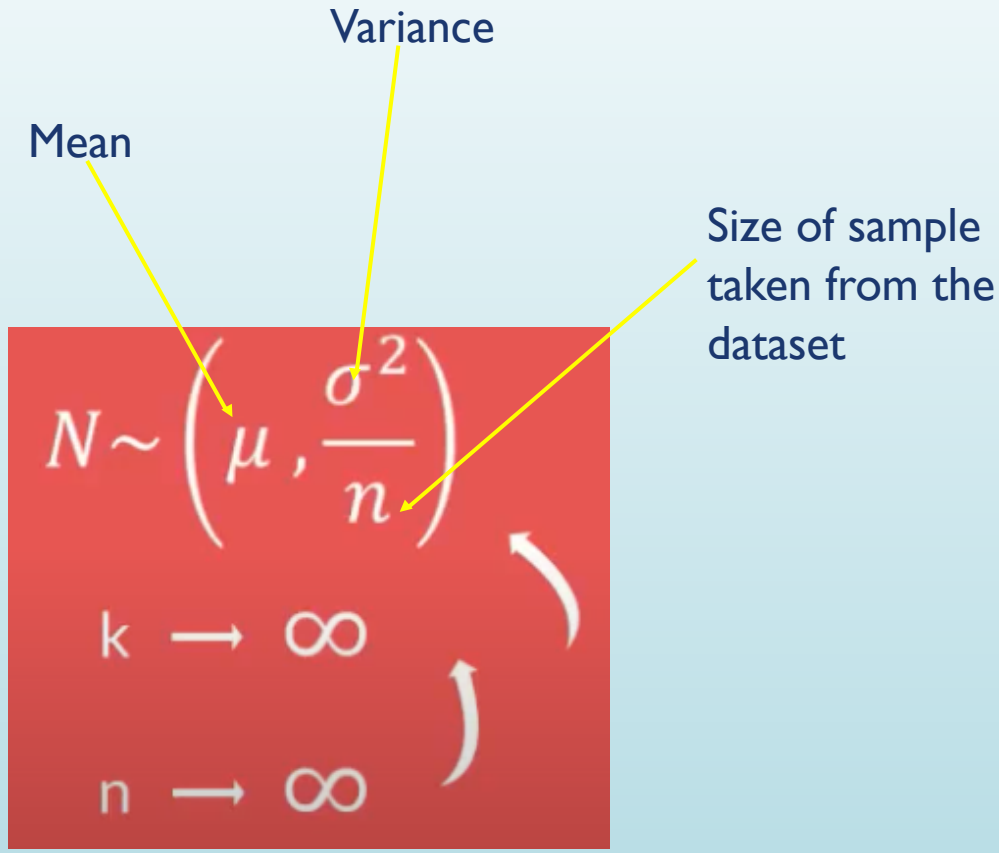
82 ✔

Answer

```
> p<-seq (0.1, 0.99, 0.01)
> sample_quantile <-quantile(act_scores, p)
> sample_quantile
     10%      11%      12%      13%      14%      15%      16%      17%      18%      19%
13.60494 13.89507 14.19862 14.46664 14.71360 14.98274 15.24947 15.47853 15.72999 15.94121
     20%      21%      22%      23%      24%      25%      26%      27%      28%      29%
16.12200 16.32100 16.52447 16.67580 16.89245 17.07520 17.24362 17.40823 17.60173 17.76615
     30%      31%      32%      33%      34%      35%      36%      37%      38%      39%
17.92766 18.07248 18.22973 18.34931 18.52808 18.67542 18.80498 18.94604 19.09691 19.25335
     40%      41%      42%      43%      44%      45%      46%      47%      48%      49%
19.38348 19.53081 19.66486 19.79833 19.94863 20.09278 20.24325 20.37247 20.51231 20.66070
     50%      51%      52%      53%      54%      55%      56%      57%      58%      59%
20.79946 20.92734 21.08109 21.23459 21.38429 21.53814 21.66232 21.82880 21.99768 22.13444
     60%      61%      62%      63%      64%      65%      66%      67%      68%      69%
22.27607 22.43218 22.56342 22.71359 22.85486 23.00470 23.16246 23.32844 23.49964 23.66440
     70%      71%      72%      73%      74%      75%      76%      77%      78%      79%
23.81223 23.98439 24.10999 24.29705 24.48122 24.68499 24.89074 25.06032 25.23528 25.41534
     80%      81%      82%      83%      84%      85%      86%      87%      88%      89%
25.60035 25.80893 25.99266 26.21095 26.42293 26.69641 26.93145 27.23751 27.50394 27.81494
     90%      91%      92%      93%      94%      95%      96%      97%      98%      99%
28.10234 28.43191 28.86662 29.22168 29.60805 30.17991 30.68391 31.42929 32.56188 34.00656
```

CENTRAL LIMIT THEOREM

- Central Limit Theorem - CLT states that the mean taken from sampling the dataset will approximate a normal distribution regardless of the type of the distribution of the entire dataset.
- The more and larger samples being extracted, the closer the sample mean will be to a normal distribution!
- The distribution will have the same mean as the original dataset with a smaller variance.
- CLT is powerful when you have a huge dataset to analyze. It is sufficient to take a small sample for analysis, with the assumption that the dataset follows a normal distribution!



SUBSET OR SAMPLING

Mean of sample k

$\bar{x}_1 = 555.2$

$\bar{x}_2 = 439.5$

$\bar{x}_3 = 625.3$

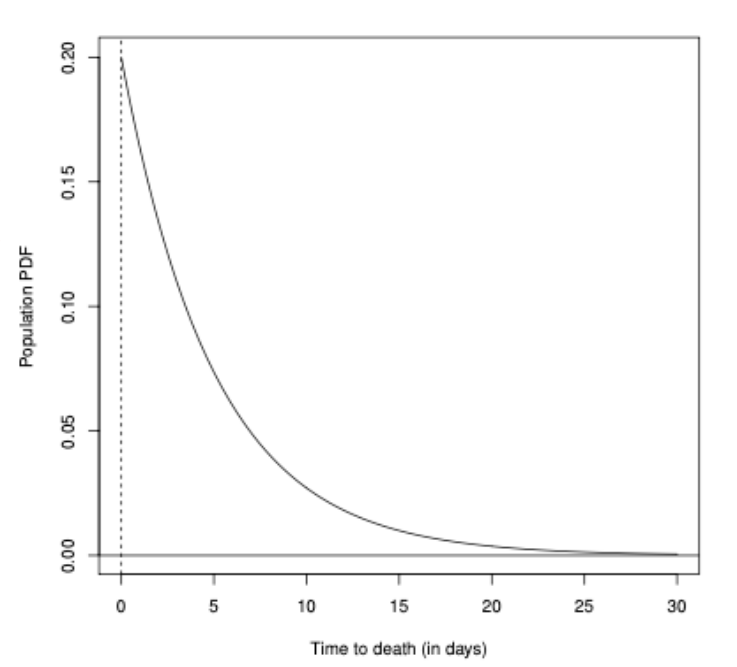
$\bar{x}_k = 567.5$

The blue boxes show the sampling process from the dataset. This helps to understand the details of the dataset.

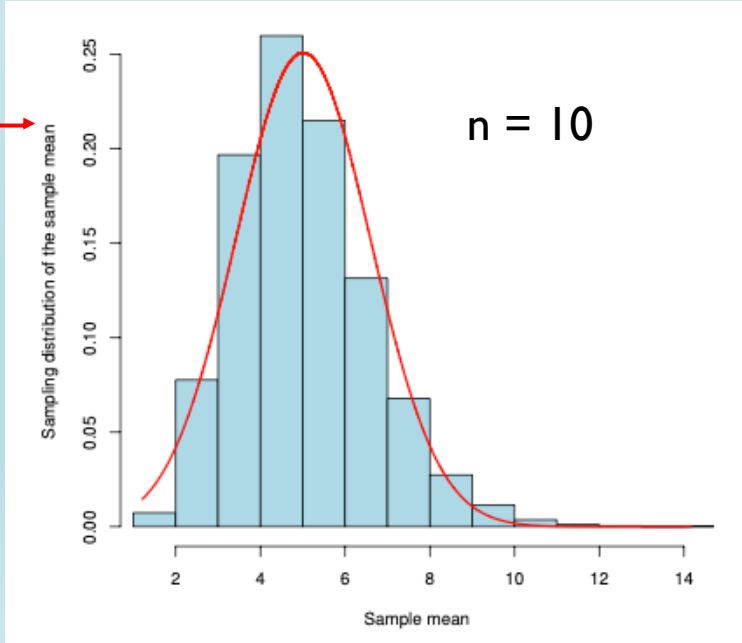
After the process of sampling, the mean of each sample is calculated before applying the CLT.

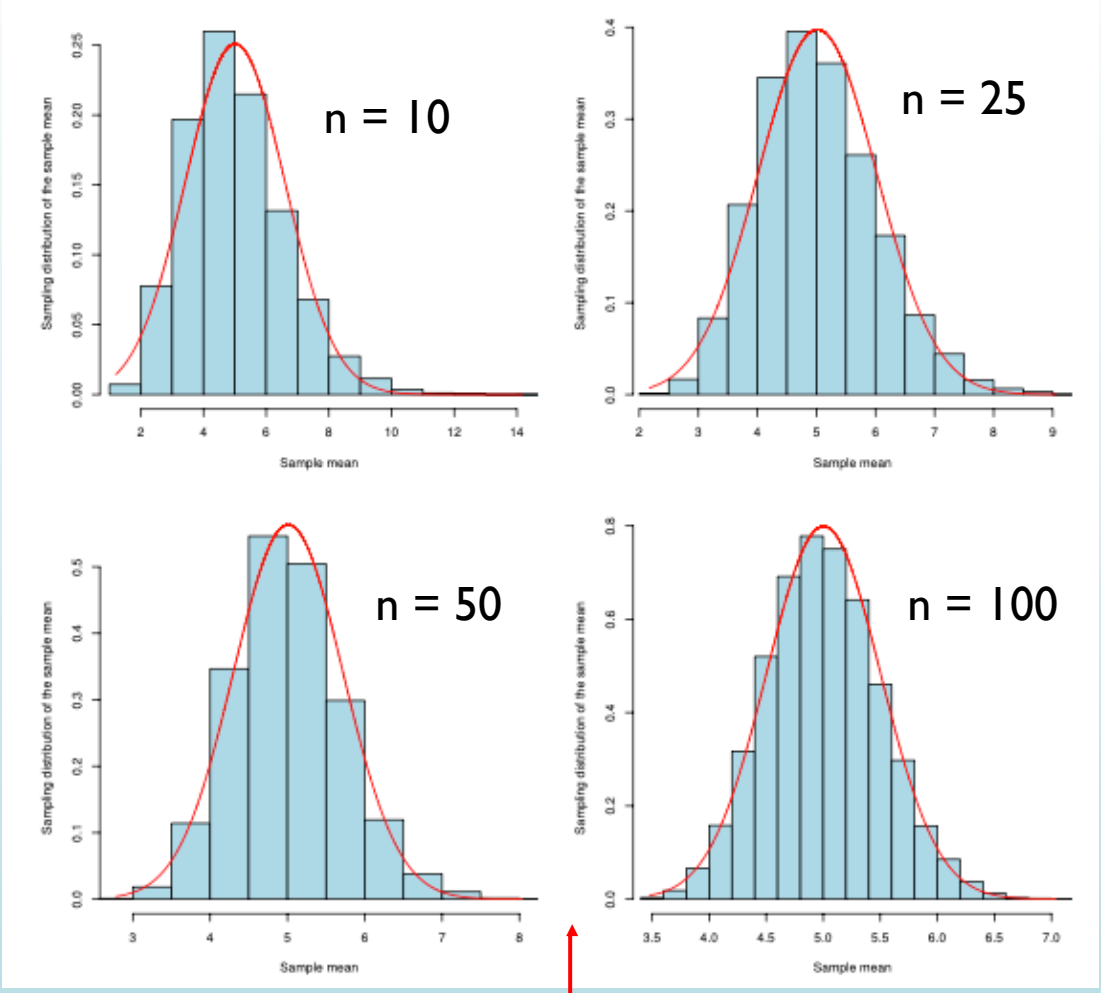
Example

The diagram shows an exponential graph comparing the population of rats to their time of death.



The diagram shows a histogram obtained after simulating 10,000 random samples, each of size n =10 rats.





By increasing the sample size, the normal approximation of the sample distribution improved.

Example

SAT Testing

An old version of the SAT college entrance exam had a -0.25 point penalty for every incorrect answer and awarded 1 point for a correct answer. The quantitative test consisted of 44 multiple-choice questions each with 5 answer choices. Suppose a student chooses answers by guessing for all questions on the test.

Question Use the Central Limit Theorem to determine the probability that a guessing student scores 8 points or higher on the test.

Answer

```
> pnorm(8,0,3.317)
[1] 0.9920634
> 1-pnorm(8,0,3.317)
[1] 0.007936603
```

The probability is **0.00794**

Question Set the seed to 21, then run a Monte Carlo simulation of 10,000 students guessing on the test. What is the probability that a guessing student scores 8 points or higher?

Answer

```
> set.seed(21, sample.kind = "Rounding")
Warning message:
In set.seed(21, sample.kind = "Rounding") :
  non-uniform 'Rounding' sampler used
> B <- 10000
> n <- 44
> p <- 0.2
> tests <- replicate(B, {
+     X <- sample(c(1, -0.25), n, replace = TRUE, prob = c(p, 1-p))
+     sum(X)
+ })
> mean(tests >= 8)
[1] 0.008
```

The probability is **0.008**

Both methods produced very close estimations!

Example

Betting on Roulette

A casino offers a House Special bet on roulette, which is a bet on five pockets (00, 0, 1, 2, 3) out of 38 total pockets. The bet pays out 6 to 1. In other words, a losing bet yields -$1 and a successful bet yields $6. A gambler wants to know the chance of losing money if he places 500 bets on the roulette House Special.

Question

What is the expected value of the payout for one bet?

-0.0785 ✔

```
p <- 5/38
a <- 6
b <- -1
mu <- a*p + b*(1-p)
mu
```

Shows the formula to find expected value of the payout, where "mu" represents the expected value

Question

What is the standard error of the payout for one bet?

2.366 ✔

```
sigma <- abs(b-a) * sqrt(p*(1-p))
```

Shows the formula to find standard error of the payout, where "sigma" represents the expected value

Example

Question What is the expected value of the sum of 500 bets?

-39.45 ✔

```
n*mu
```

Shows the formula to find expected value of the payout, where "mu" represents the expected value

Question What is the standard error of the sum of 500 bets?

52.90 ✔

```
sqrt(n) * sigma
```

Shows the formula to find standard error of the payout, where "sigma" represents the expected value

Question

Use `pnorm()` with the expected value of the sum and standard error of the sum to calculate the probability of losing money over 500 bets, $\Pr(X \leq 0)$.

0.7721 ✔

"pnorm" function calculates cumulative distribution function of a normal distribution.

Answer

```
> pnorm(0,-39.45,52.90)
[1] 0.7720898
```

CONCEPTS APPLICATION

INTEREST RATE

- Interest rates for loans are set using the **probability** of loan defaults to calculate a rate that minimizes the probability of losing money.
- We can define the outcome of loans as a **random variable**. We can also define the sum of outcomes of many loans as a random variable.
- The **Central Limit Theorem** can be applied to fit a normal distribution to the sum of profits over many loans.
- We can use properties of the **normal distribution** to calculate the interest rate needed to ensure a certain probability of losing money for a given probability of default.

- The Central Limit Theorem states that the **sum of independent draws of a random variable follows a normal distribution**. However, when the draws are not independent, this assumption does not hold.

- If an event changes the probability of default for all borrowers, then the probability of the bank losing money changes.

- **Monte Carlo simulations** can be used to model the effects of unknown changes in the probability of default.

Example

Bank Earnings

Question

Say you manage a bank that gives out 10,000 loans. The default rate is 0.03 and you lose $200,000 in each foreclosure.

Create a random variable S that contains the earnings of your bank. Calculate the total amount of money lost in this scenario.

Answer

script.R | solution.R

```r
# Assign the number of loans to the variable `n`
n <- 10000

# Assign the loss per foreclosure to the variable `loss_per_foreclosure`
loss_per_foreclosure <- -200000

# Assign the probability of default to the variable `p_default`
p_default <- 0.03

# Use the `set.seed` function to make sure your answer matches the expected result after random sampling
set.seed(1)

# Generate a vector called `defaults` that contains the default outcomes of `n` loans
defaults <- sample( c(0,1), n, replace = TRUE, prob=c(1-p_default, p_default))

# Generate `S`, the total amount of money lost across all foreclosures. Print the value to the console.
S <- sum(defaults * loss_per_foreclosure)
S
```

```
> S
[1] -6.3e+07
```

Shows "sample" function generating a vector called "defaults" that contains "n" samples from a vector of "c(0,1)"

"0" indicates payment and "1" indicates default

Total amount of money lost is **- $63M**

Example

Bank Earnings with Monte Carlo simulation

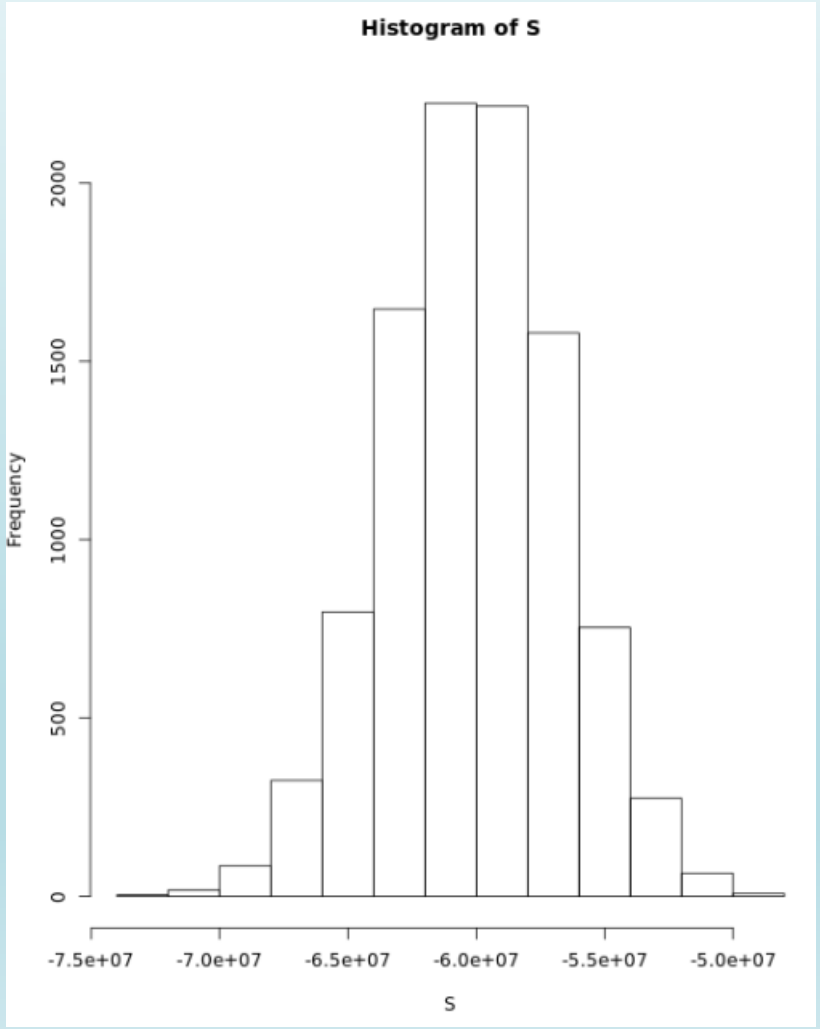
Question

Run a Monte Carlo simulation with 10,000 outcomes for S, the sum of losses over 10,000 loans. Make a histogram of the results.

Answer

script.R | solution.R

```r
# Assign the number of loans to the variable `n`
n <- 10000

# Assign the loss per foreclosure to the variable `loss_per_foreclosure`
loss_per_foreclosure <- -200000

# Assign the probability of default to the variable `p_default`
p_default <- 0.03

# Use the `set.seed` function to make sure your answer matches the expected result after random sampling
set.seed(1)

# The variable `B` specifies the number of times we want the simulation to run
B <- 10000

# Generate a list of summed losses 'S'. Replicate the code from the previous exercise over 'B' iterations to generate a list of summed losses for 'n' loans. Ignore any warnings for now.

S <- replicate(B, {
    defaults <- sample( c(0,1), n, prob=c(1-p_default, p_default), replace = TRUE)
  sum(defaults * loss_per_foreclosure)
})
```



The histogram on the left shows the distribution of losses across 10000 loans

Example

Bank Earnings – Expected value & Standard Error

Question

What is the expected value of S, the sum of losses over 10,000 loans? For now, assume a bank makes no money if the loan is paid.

What is the standard error of S?

Answer

script.R | solution.R

```
# Assign the number of loans to the variable `n`
n <- 10000

# Assign the loss per foreclosure to the variable `loss_per_foreclosure`
loss_per_foreclosure <- -200000

# Assign the probability of default to the variable `p_default`
p_default <- 0.03

# Calculate the expected loss due to default out of 10,000 loans
n*(p_default*loss_per_foreclosure + (1-p_default)*0)
```

```
[1] -6e+07
```

```
# Compute the standard error of the sum of 10,000 loans
sqrt(n) * abs(loss_per_foreclosure) * sqrt(p_default*(1 - p_default))
```

```
[1] 3411744
```

Shows the formula to find expected value of the sum of n draws of a random variable.

$$n \times (ap + b(1-p))$$

Shows the formula to find standard error of the sum of n draws of a random variable.

$$\sqrt{n} \times |b-a| \sqrt{p(1-p)}$$

The expected value is **- $60M** and the standard error is **$3.4M**

Example

Bank Earnings – Interest Rate Part 1

Question

So far, we've been assuming that we make no money when people pay their loans and we lose a lot of money when people default on their loans. Assume we give out loans for $180,000. How much money do we need to make when people pay their loans so that our net loss is $0?

In other words, what interest rate do we need to charge in order to not lose money?

Answer

script.R | solution.R

```
1  # Assign the loss per foreclosure to the variable `loss_per_foreclosure`
2  loss_per_foreclosure <- -200000
3
4  # Assign the probability of default to the variable `p_default`
5  p_default <- 0.03
6
7  # Assign a variable `x` as the total amount necessary to have an expected outcome of $0
8  x <- -(loss_per_foreclosure*p_default) / (1 - p_default)
9
10 # Convert `x` to an interest rate, given that the loan amount is $180,000. Print this value to the console.
11 x / 180000
```

```
[1] 0.03436426
```

Below shows the explanation how to find total amount needed to get a net loss of $0

We can calculate the amount x to add to each loan so that the expected value is 0 using the equation $lp + x\,(1 - p) = 0$. Note that this equation is the definition of expected value given a loss per foreclosure l with foreclosure probability p and profit x if there is no foreclosure (probability $1 - p$).

The interest needed is **3.4%**

Example

Bank Earnings – Interest Rate Part II

Question

With the interest rate calculated in the last example, we still lose money 50% of the time. What should the interest rate be so that the chance of losing money is 1 in 20?

In math notation, what should the interest rate be so that $\Pr(S < 0) = 0.05$?

Answer

script.R | solution.R

```
1  # Assign the number of loans to the variable `n`
2  n <- 10000
3
4  # Assign the loss per foreclosure to the variable `loss_per_foreclosure`
5  loss_per_foreclosure <- -200000
6
7  # Assign the probability of default to the variable `p_default`
8  p_default <- 0.03
9
10 # Generate a variable `z` using the `qnorm` function
11 z <- qnorm(0.05)
12
13 # Generate a variable `x` using `z`, `p_default`, `loss_per_foreclosure`, and `n`
14 x <- -loss_per_foreclosure*( n*p_default - z*sqrt(n*p_default*(1 - p_default)))/ ( n*(1 - p_default) + z*sqrt
   (n*p_default*(1 - p_default)))
15
16 # Convert `x` to an interest rate, given that the loan amount is $180,000. Print this value to the console.
17 x / 180000
```

"qnorm" function is used to compute a continuous variable at given quantile of the distribution to solve for "z"

```
[1] 0.03768738
```

The interest needed is **3.8%**