Balasubramanian, Sreenithi
[redacted]

1. **Educational attainment of couples**
   a. **Probability that a man has at least a bachelor's degree**:
      P(man_atleast_bachelors) = p(man_bachelors) + p(man_graduate)
      = 0.16 + 0.09
      = 0.25

   b. **Probability that a woman has at least a bachelor's degree**:
      P(woman_atleast_bachelors) = p(woman_bachelors) + p(woman_graduate)
      = 0.17 + 0.09
      = 0.26

   c. **Probability that a man or a woman getting married has at least a bachelor's degree**:
      Assumption: a man and a woman getting a bachelor's degree are both independent events.
P(man_and_woman_atleast_bachelors) = P(man_atleast_bachelors) x P(woman_atleast_bachelors)
= 0.25 x 0.26
= 0.065

   d. **It is reasonable** because the legal age to get married will most likely be after they finish a bachelor's degree. So, they both get their degrees separately.

R Syntax:

```
	male_bachelors <- 0.16
	male_grad <- 0.09

	prob_man_atleast_bachelors <- (male_bachelors + male_grad)

	print(paste("Probability of randomly chosen man has at
least a Bachelors degree = ", prob_man_atleast_bachelors))


woman_bachelors <- 0.17
	woman_grad <- 0.09

	prob_woman_atleast_bachelors <- (woman_bachelors +
woman_grad)

	print(paste("Probability of randomly chosen woman has at
least a Bachelors degree =                                  ",
prob_woman_atleast_bachelors))
```

```
prob_man_and_woman_bachelors <- prob_man_atleast_bachelors
*          prob_woman_atleast_bachelors

print(paste("Probability of man and woman has at least a
Bachelors degree = ", prob_man_and_woman_bachelors))
```

Output:

```
[1] "Probability of randomly chosen man has at least a
Bachelors degree =  0.25"

[1] "Probability of randomly chosen woman has at least a
Bachelors degree =  0.26"

[1] "Probability of man and woman has at least a Bachelors
degree =  0.065"
```

2. **Probabilities of rolling two dice**
   a. **The sum of dice is NOT 5**
      P(sum_dice_not_5) = 1 – P(sum_dice_5)

      Possibilities for sum 5 = (1,4) (2,3) (3,2) (4,1)

      P(sum_dice_5) = 2 x [P (1,4) + P (2,3)]
      = 2 x [(1/36) + (1/36)]
      = 4/36
      =1/9

      P(sum_dice_NOT_5) = 1 – 1/9
      = **8/9**

   b. **Sum is at least 7**
      P(sum_atleast_7) = P (sum >=7)
      = P (sum = 7) + P (sum = 8) + … P (sum = 12)
      = (6+5+4+3+2+1) / 36
      = **21 / 36**

   c. **Sum is no more than 8**
      P(sum_max_8) = P(sum<=8)
      = P (sum = 2) + P(sum=3) + … P(sum=8)
      = (1+2+3+4+5+6+5)/36
      = **26/36**

R Syntax:

```
#Q2 - rolling two dice

prob_sum_dice_5 <- 2 * (2/36.0)
prob_sum_dice_not_5 <- 1 - prob_sum_dice_5

print(paste("Probability of sum dice is not 5 =",
round(prob_sum_dice_not_5,3)))

prob_sum_atleast_7 <- ((6+5+4+3+2+1)/36.0)
print(paste("Probability of sum atleast 7 =",
round(prob_sum_atleast_7,3)))

prob_sum_no_more_than_8 <- (1+2+3+4+5+6+5)/36.0
print(paste("Probability of sum no more than 8 =",
round(prob_sum_no_more_than_8,3)))
```

Output:

```
[1] "Probability of sum dice is not 5 = 0.889"
[1] "Probability of sum atleast 7 = 0.583"
[1] "Probability of sum no more than 8 = 0.722"
```

3. **Health Coverage, Relative frequencies**
   a. **Are being in excellent health and having health coverage mutually exclusive?**

      A – excellent health; B – coverage is YES
      P (A or B) = P(A) + P(B) - P (A and B) -> if A and B are not mutually exclusive
      If A and B are mutually exclusive: P (A or B) = P(A) + P(B) which indicates P (A and B)=0

      P(A) = 0.2329
      P(B) = 0.8738
      P (A and B) = 0.2099

      Since P (A and B)! = 0, they are not mutually exclusive

   b. **Probability that a randomly chosen individual has excellent health condition**
      P(excellent) = 0.2329
   c. **Probability that a randomly chosen individual has excellent health condition given that they have health coverage**

      A – individual has excellent health condition
      B – individual has health coverage

      P(A|B) = P (A and B) / P(B)

      P (A and B) => P (excellent health AND coverage_yes) = 0.2099
      P(B) => P(coverage_YES) = 0.8738

      P(A|B) = 0.2099 / 0.8738 = 0.2402

   d. **Probability that a randomly chosen individual has excellent health condition given that they do not have health coverage**

      A – individual has excellent health condition
      B – individual does not have health coverage

      P(A|B) = P (A and B) / P(B)

      P (A and B) => P (excellent health AND coverage_NO) = 0.0230
      P(B) => P(coverage_NO) = 0.1262

P(A|B) = 0.0230 / 0.1262 = 0.1822

e. **Does having excellent health and having health coverage appear to be independent?**

   For two events A and B to be independent: P (A and B) = P(A) x P(B)
   A – excellent health; B – coverage is YES
   P(excellent_health) = 0.2329
   P(coverage_YES) = 0.8738

   Given: P (A and B) = 0.2099
   P(A) x P(B) = 0.2035

   P (A and B)! = P(A) x P(B). Hence, they are NOT independent.

R Syntax:

#Q3

```
prob_excellent_health <- 0.2329

prob_excellent_health_and_coverage <- 0.2099
prob_excellent_health_and_no_coverage  <- 0.0230
prob_coverage <- 0.8738
prob_no_coverage <- 0.1262

prob_excellent_health_given_coverage <-
prob_excellent_health_and_coverage/prob_coverage

print(paste("P(excellent health | coverage YES) = ",
round(prob_excellent_health_given_coverage,3)))

prob_excellent_health_given_no_coverage <-
prob_excellent_health_and_no_coverage/prob_no_coverage

print(paste("P(excellent health | coverage NO) = ",
round(prob_excellent_health_given_no_coverage,3)))
```

Output:

```
[1] "P(excellent health | coverage YES) =  0.24"
[1] "P(excellent health | coverage NO) =  0.182"
```

4. **Joint and Conditional Probabilities**
   P(A) = 0.3 and P(B) = 0.7

   a. **Can you compute P (A and B) if you know only P(A) and P(B)?**
      No. We do not know if A and B are independent events.

   b. **Assuming A and B are independent random processes:**
      i. **What is P (A and B)?**
         P (A and B) = P(A) x P(B)
         = 0.3 x 0.7
         = 0.21
      ii. **P (A or B) = P(A) + P(B)**
         = 0.3 + 0.7
         = 1.0
      iii. **P(A|B) = P (A and B) / P(B)**
         = 0.21 / 0.7
         = 0.3

   c. **If P (A and B) = 0.1, are the random variables giving rise to A and B independent?**
      No, because if they were independent P (A and B) = 0.21. 0.21 != 0.1, hence the variables giving rise to A and B are not independent.

   d. **Given P (A and B) = 0.1, what is P(A|B)?**
      P(A|B) = 0.1 / 0.7
      = 1/7

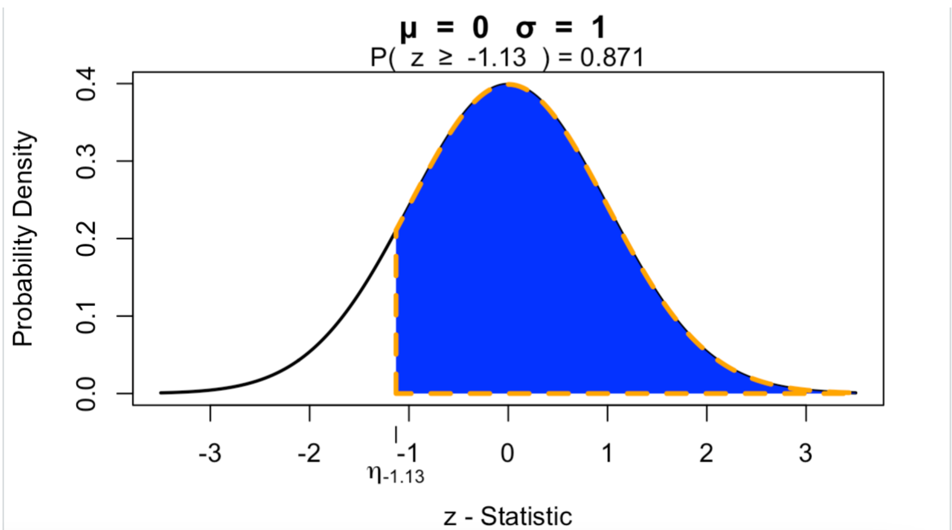
R Syntax

```
#Q4
p_A <- 0.3
p_B <- 0.7

p_A_and_B <- p_A * p_B
print(paste("P(A and B) = ", p_A_and_B))

p_A_or_B <- p_A + p_B
print(paste("P(A or B) = ", p_A_or_B))

P_A_given_B <- p_A_and_B / p_B
print(paste("P(A|B) = ", P_A_given_B))
```

```
Output:
[1] "P(A and B) =  0.21"
[1] "P(A or B) =  1"
[1] "P(A|B) =  0.3"
```

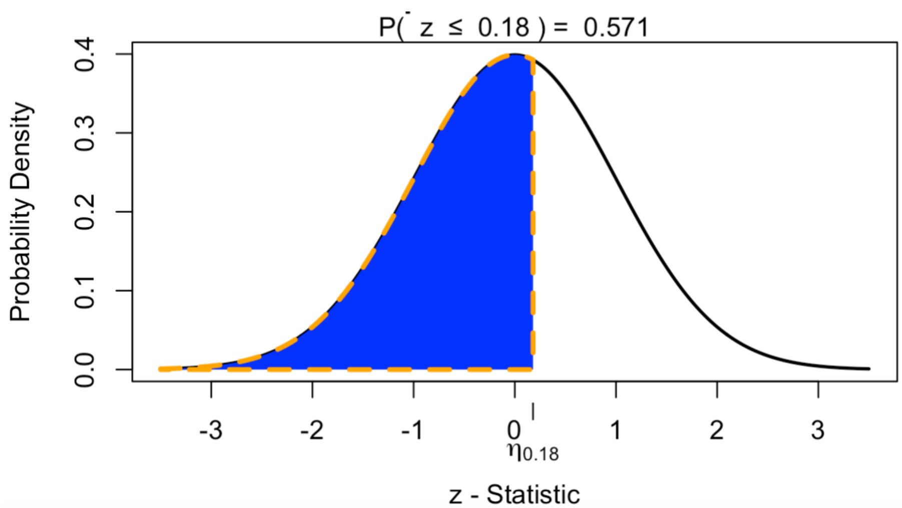
5. **Area under the curve – what percent is found under each region**
    a. Z > -1.13

    ```
    pnorm(-1.13, lower.tail = FALSE)
    Output: 0.8707619 (87.07%)
    ```



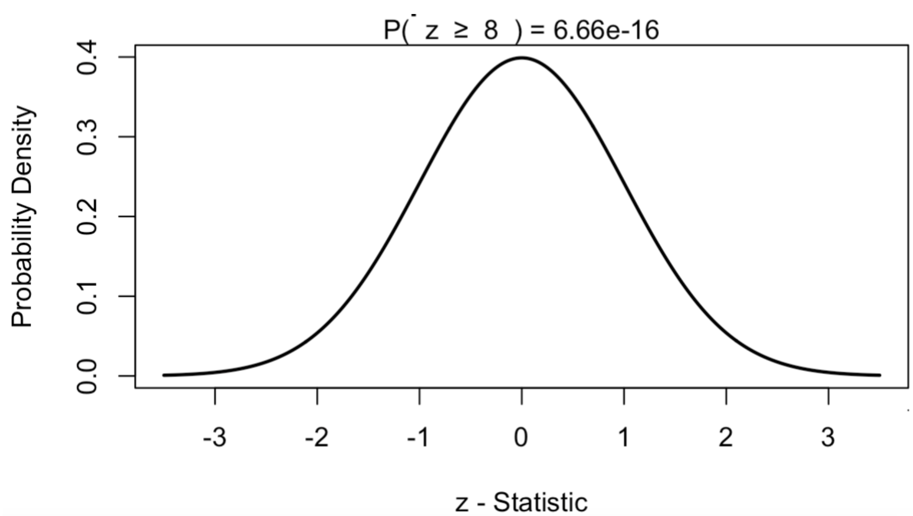
b. Z < 0.18

```
pnorm(0.18)
Output: 0.5714237 (57.1%)
```



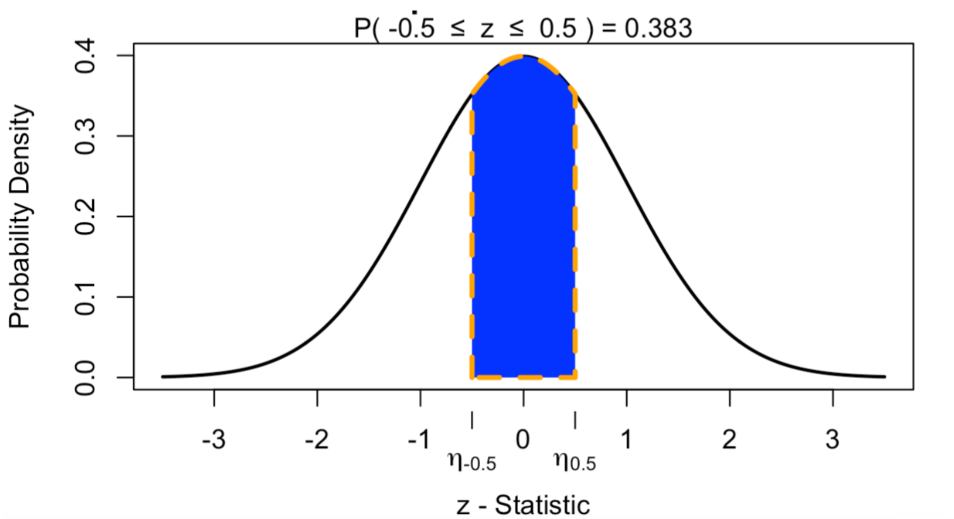
c. Z > 8

```
pnorm(8, lower.tail = FALSE)
Output: 6.220961e-16
```



d. |Z| < 0.5

```
pnorm(0.5) - pnorm(-0.5)
Output: 0.3829249 (38.2%)
```



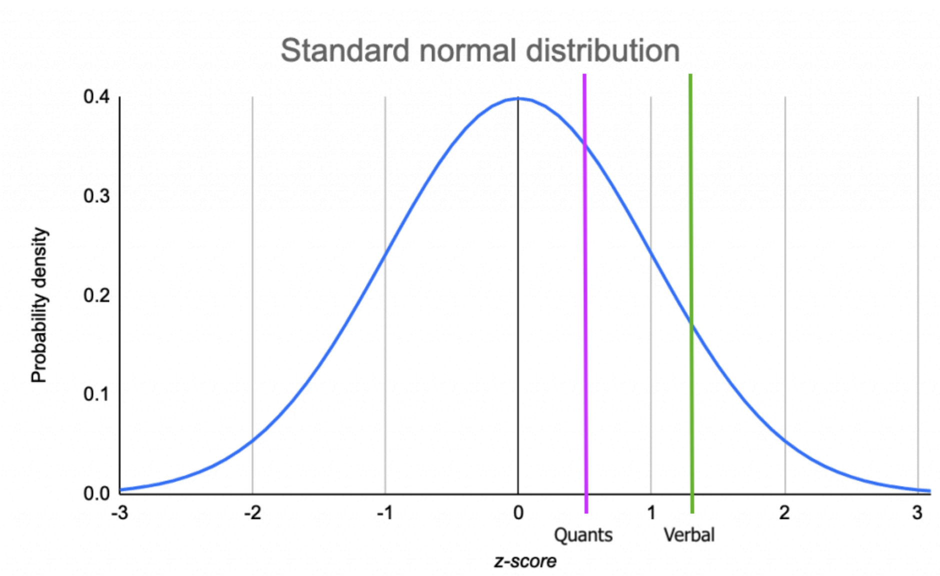
6. **GRE Scores**

Data - data distribution is normal

Section	Score	Mean	Standard Deviation
Verbal	160	151	7
Quants	157	153	7.67

a. **Shorthand for the normal distributions**:
   i. Verbal: V ~ N (151, 7)
   ii. Quants: Q ~ N (153, 7.67)

b. **Z-scores**
   i. Verbal: Z(verbal) = [V – mean(V)] / std(V)
      = [160 – 151] / 7
      = 1.28
   ii. Quants: Z(quants) = [Q – mean(Q)] / std(Q)
      = [157 – 153] / 7.67
      = 0.52



c. The Z-score of a variable X tells us how many standard deviations above or below the mean does X fall under, given a distribution. In this case it tells us how much better or how much worse Sophia has scored in GRE when compared to the average scores.

d. Relative to others, she did better in the Verbal section – since her Z-score for her verbal is higher than the quants Z-score. It indicates she did better than average in verbal compared to quants.
e. Percentile Scores:

Quants: ~70th percentile
Verbal: ~90th percentile

f. 30% of people did better than Sophia in the Quants section and 10% of people did better than Sophia in the verbal section.
g. Simply comparing raw scores to judge which section she did better on could be misleading because these scores are comparative to everyone who took the test not just one individual. To see where she stands with other people who took the test, just comparing raw data would not be correct. Her scores need to be compared with the average scores calculated across all test takers to properly judge which section she did better in. This is where mean and standard deviation become helpful.
h. Since we don't know what distribution it is – we only know it is not nearly normal. If we know the mean and standard deviation, we can calculate the questions related to Z-scores [questions (b) - (d)]. To calculate percentile, we need to know the kind of distribution it is to find out what percentile the scores fall in. So there might not be enough information to calculate questions [e] and [g] if we don't know what distribution it is.

<u>R Syntax</u>:

```
#Q6

verbal <- 160
quants <- 157

mean_verbal <- 151
std_verbal <- 7

mean_quants <- 153
std_quants <- 7.67

z_score_verbal <- (verbal - mean_verbal)/std_verbal
print(paste("VERBAL Z-score = ",round(z_score_verbal,3)))

z_score_quants <- (quants - mean_quants)/std_quants
print(paste("QUANTS Z-score = ",round(z_score_quants,3)))
```

```
Output
[1] "VERBAL Z-score =  1.286"
[1] "QUANTS Z-score =  0.522"
```

```
#Percentile
    pnorm(z_score_verbal)
    pnorm(z_score_quants)

    Output:
    [1] 0.9007286
    [1] 0.6989951
```