Riya Chhikara

15 January 2023

Data Science- Basic Statistics

Q1) Identify the Data type for the Following:

Activity	**Data Type**
Number of beatings from Wife	Discrete
Results of rolling a dice	Discrete
Weight of a person	Continuous
Weight of Gold	Continuous
Distance between two places	Continuous
Length of a leaf	Continuous
Dog's weight	Continuous
Blue Color	Discrete
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Discrete

Q2) Identify the Data types, which were among the following: Nominal, Ordinal, Interval, Ratio.

Data	**Data Type**
Gender	Nominal
High School Class Ranking	Ordinal
Celsius Temperature	Interval
Weight	Ratio
Hair Color	Nominal
Socioeconomic Status	Ordinal
Fahrenheit Temperature	Interval
Height	Ratio
Type of living accommodation	Nominal
Level of Agreement	Ratio
IQ(Intelligence Scale)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Nominal
Time on a Clock with Hands	Interval
Number of Children	Ratio

Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Interval
Years of Education	Ratio

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

A3) **Sample Space-**
{HHH,HHT,HTH,THH, HTT,TTH,THT,TTT}
Total number of outcomes= 8
Total number of favorable outcomes= 3
So, there are 3 possibilities of getting two heads and one tail – {HTH,HHT,THH}
P(2H,1T) = 3/8= 0.375

Q4) Two Dice are rolled, find the probability that sum is

- **a) Equal to 1**
- **b) Less than or equal to 4**
- **c) Sum is divisible by 2 and 3**

A4) **Sample Space-**
{(1,1)(2,1)(3,1)(4,1)(5,1)(6,1)
(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)
(1,3)(2,3)(3,3)(4,3)(5,3)(6,3)
(1,4)(2,4)(3,4)(4,4)(5,4)(6,4)
(1,5)(2,5)(3,5)(4,5)(5,5)(6,5)
(1,6)(2,6)(3,6)(4,6)(5,6)(6,6)}
Total number of outcomes= 36

- a) There is no outcome where the sum is equal to 1, so the probability is 0.
- b) There are 6 outcomes where the sum of less than or equal to 6. These outcomes are- (1,1)(2,1)(3,1)(1,2)(2,2)(1,3).So, P = 6/36= 1/6= 0.167
- c) (6,6), (1,5), (5,1), (3,3), (4,2), (2,4)- There are 6 outcomes where the sum is either 6 or 36. Both of these numbers are divisible by 2 and 3.
  So, P= 6/36= 1/6= 0.167

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

A5) Total number of balls= 2+3+2=7
So, Number of ways of drawing 2 balls out of 7
=$^{7}C_{2}$
=(7×6)/ (2×1)
=21
Total favourable outcomes of drawing 2 balls, none of which is blue is as follows-
Number of outcomes where 2 balls are drawn out of (2 + 3) balls.
=$^{5}C_{2}$
=(5×4)/ (2×1)

=10

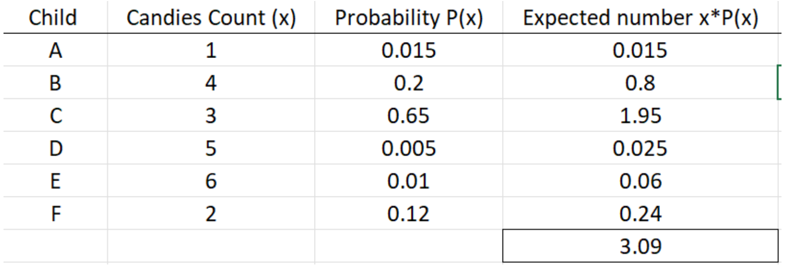
∴P(none of the balls are blue)= 10/21= 0.476

Q6) Calculate the Expected number of candies for a randomly selected child
Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

CHILD	Candies count	Probability
A	1	0.015
B	4	0.20
C	3	0.65
D	5	0.005
E	6	0.01
F	2	0.120

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20



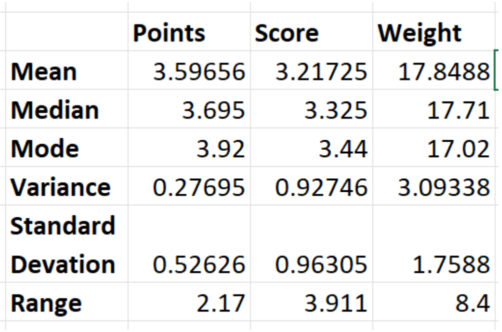
Child	Candies Count (x)	Probability P(x)	Expected number x*P(x)
A	1	0.015	0.015
B	4	0.2	0.8
C	3	0.65	1.95
D	5	0.005	0.025
E	6	0.01	0.06
F	2	0.12	0.24
			3.09

So, Expected number of candies for a randomly selected child is 3.09.

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- **For Points,Score,Weight->**
  Find Mean, Median, Mode, Variance, Standard Deviation, and Range and comment about the values/ Draw some inferences.

Use Q7.csv file



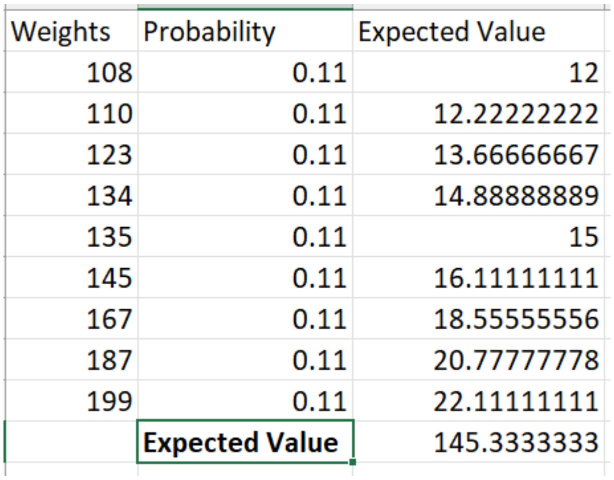
	Points	Score	Weight
Mean	3.59656	3.21725	17.8488
Median	3.695	3.325	17.71
Mode	3.92	3.44	17.02
Variance	0.27695	0.92746	3.09338
Standard Devation	0.52626	0.96305	1.7588
Range	2.17	3.911	8.4

For points, the mean, median and mode are almost the same, telling us the data is centrally dispersed. The low variance, standard deviation and range tell us that the data doesn't have much variation.

For score, while the value of the mean, median and mode are quite similar, the variance, standard deviation and range are higher compared to those of points.

For weight, the mean, median and mode are quite similar but variance, standard deviation and range are higher telling us that the data shows more variation.

Q8) Calculate Expected Value for the problem below. The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, 199. Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?



Weights	Probability	Expected Value
108	0.11	12
110	0.11	12.22222222
123	0.11	13.66666667
134	0.11	14.88888889
135	0.11	15
145	0.11	16.11111111
167	0.11	18.55555556
187	0.11	20.77777778
199	0.11	22.11111111
	Expected Value	145.3333333

Let X be the random variable denoting the weight of the patients at a clinic. Let P(X) represent the probability of the X being the weight of the randomly chosen patient.

Expected value of the weight of the randomly chosen patient is X * P(X)= 145.3.

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

	Speed	**Distance**
Skewness	-0.11395477	0.782483517
Kurtosis	-0.50899442	0.405052582

For speed, the skewness is -0.11 or negatively skewed. This tells us that there is a longer left tail and more datapoints are concentrated near the right tail. However, the very low value of the skewness tells us that it is moderately symmetrical. For speed, the kurtosis is -0.51, telling us that is platykurtic and has less datapoints spread along the tails, compared to the normal distribution.

For distance, the skewness is -0.78, or it is negatively skewed. This tells us that there is a longer left tail and more datapoints are concentrated near the right tail. However, the very low value of the skewness tells us that it is moderately symmetrical. For distance, the kurtosis is 0.41, telling us that are more datapoints spread along the tails compared to the normal distribution.

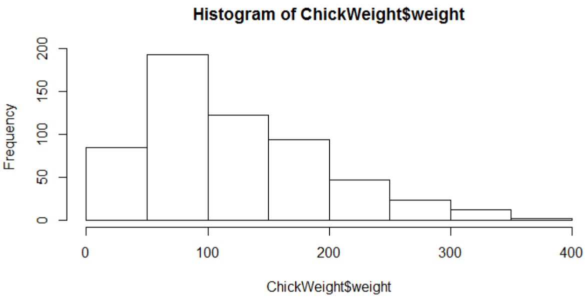
SP and Weight(WT)

	SP	**WT**
Skewness	1.581454	-0.60331
Kurtosis	2.977329	0.950291

For SP, the skewness is 1.58, which is greater than 1, indicating that the distribution shows positive skewness. This means that it has a longer tail on the right, indicating that there are more datapoints distributed along the left. For SP, the kurtosis is 2.9 which is very close to 3. Since the value is lesser than 3, we can say that the distribution has lighter tails, but since it's very close to 3, the distribution is almost mesokurtic and close to the regular-belled shaped curve of a normal distribution.

For weight, the skewness is -0.61 telling us it's negatively skewed with a longer tail and since the value is a little over 0.5, we can say that the skewness is moderate. There are more datapoints along the right tail resulting in a longer left tail. For weight, the kurtosis is 0.95 telling us that it has a positive kurtosis, that is, distribution has heavier tails and sharper peak than a regular bell-shaped normal distribution.

Q10) Draw inferences about the following boxplot & histogram.



Interpretation of the histogram:

With right-skewed distribution (also known as "positively skewed" distribution), most data falls to the right, or positive side, of the graph's peak. Thus, the histogram skews in such a way that its right side (or "tail") is longer than its left side. On a right-skewed histogram, the mean, median, and mode are all different.

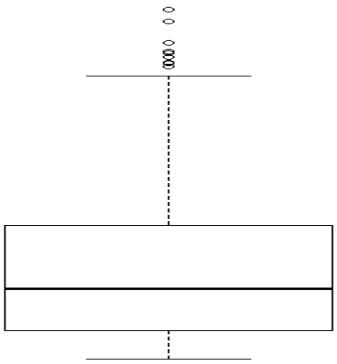


Figure 1 Boxplot

Interpretation of the boxplot:

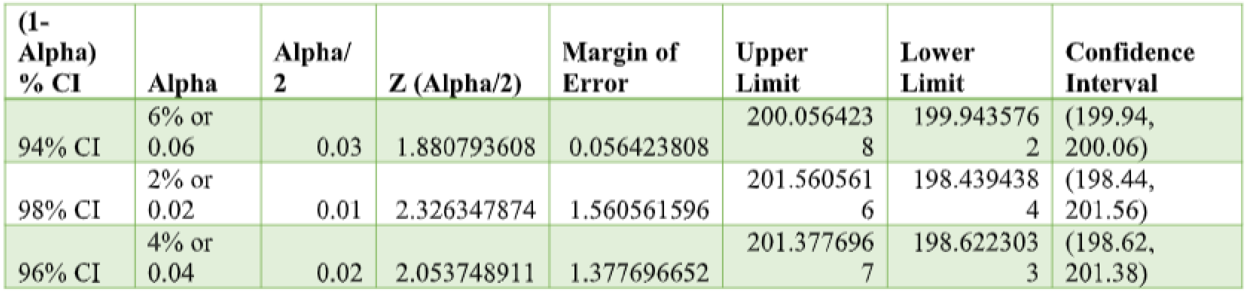
Top-most line or the whisker marks the maximum. The upper whisker is given by UW= Q3+ 1.5*IR. IR is the inter-quartile range (Q3- Q1).There are outliers lying outside/above the upper whisker, thus their value is greater than Q3+ 1.5* IR. This indicates that there are more datapoints with extremely high values, that cannot be accounted for by the calculations used in constructing the box-plot. The bottom most line is the lower whisker or the minimum, given by LW= Q1-1.5*IR.

The box plot shows a longer upper whisker and a shorter lower whisker, with outliers lying above the upper whisker. This means that there is greater range in the dataset, since the value of the maximum is quite high compared to the minimum.

The presence of outliers towards the upper whisker tells us that there are more datapoints with very high values.

Q11) Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

A11) The sample size is 2000 and the population size is 3000000. The sample mean of weight 200lbs. The sample mean of the weight is 200 pounds. The sample deviation is 30 pounds.



(1-Alpha) % CI	Alpha	Alpha/2	Z (Alpha/2)	Margin of Error	Upper Limit	Lower Limit	Confidence Interval
94% CI	6% or 0.06	0.03	1.880793608	0.056423808	200.0564238	199.9435762	(199.94, 200.06)
98% CI	2% or 0.02	0.01	2.326347874	1.560561596	201.5605616	198.4394384	(198.44, 201.56)
96% CI	4% or 0.04	0.02	2.053748911	1.377696652	201.3776967	198.6223033	(198.62, 201.38)

Q12) Below are the scores obtained by a student in tests

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

1) Find mean, median, variance, standard deviation.
2) What can we say about the student marks?

Mean	41
Median	40.5
Variance	24.11111
Standard Deviation	4.910307

The mean and median are almost the same, thus the distribution is quite symmetrical and balanced. The variance is moderately high, which means there is more dispersion along the central tendencies. (mean, median, mode)

Q13) What is the nature of skewness when mean, median of data are equal?
When the mean and median of a dataset are equal, there is no skewness in the data. The data is perfectly symmetrical with zero skewness.

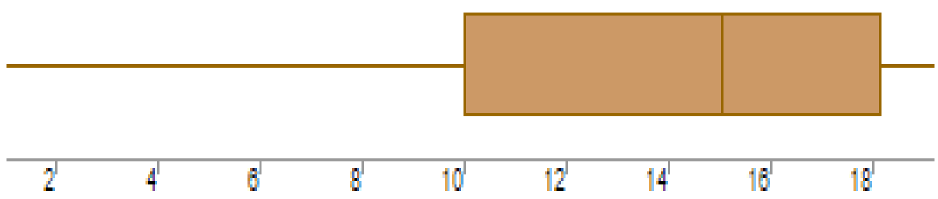
Q14) What is the nature of skewness when mean > median ?
The distribution is right-skewed or positively- distributed. It has a longer tail on the right end.

Q15) What is the nature of skewness when median > mean?
When the median is greater than the mean, it is left skewed or negatively distributed.

Q16) What does positive kurtosis value indicates for a data ?
It indicates that the distribution is leptokurtic, that is it shows more peak in the curve and has lighter tail.

Q17) What does negative kurtosis value indicates for a data?
It indicates that the distribution is leptokurtic, that the distribution shows a flatter peak, so the tail is heavier.

Q18) Answer the below questions using the below boxplot visualization.

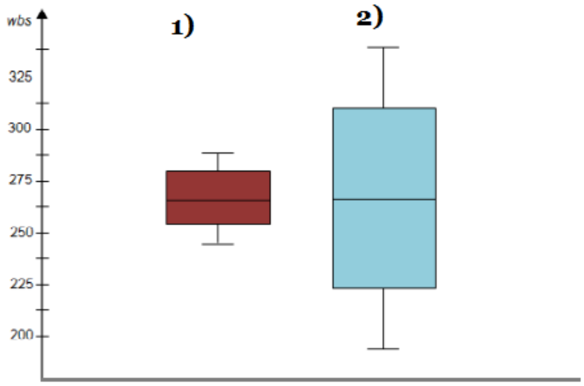


a) What can we say about the distribution of the data?
The boxplot has a longer lower whisker and a shorter upper whisker. The median is more towards the right. Also, there is greater range along the left with fewer datapoints as shown by the longer whisker.

b) What is nature of skewness of the data?
Left skewed or negatively skewed. This is because, the median is along 15.7 and the upper quartile is at 18. This tell us that more data is on the riht

c) What will be the IQR of the data (approximately)?
IQR= Q3-Q1
Here, Q3=18 and Q1= 10.
Thus, IQR= 18-10=8

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Boxplot 2 is normally distributed since whiskers are equally long on both sides and the median is right in the middle.

Q 20) Calculate probability from the given dataset for the below cases. Data _set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

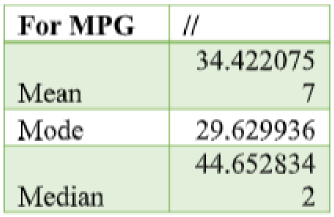
a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

	True	**Total**	**Probability**
If MPG>38	33	81	0.407407407
If MPG<40	61	81	0.75308642

If 20<MPG<50	69	81	0.851851852

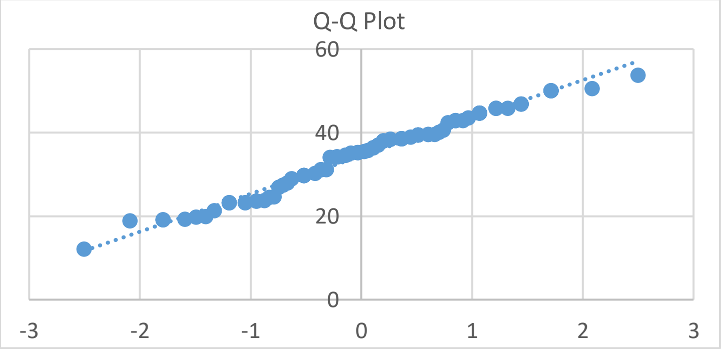
Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
   Dataset: Cars.csv



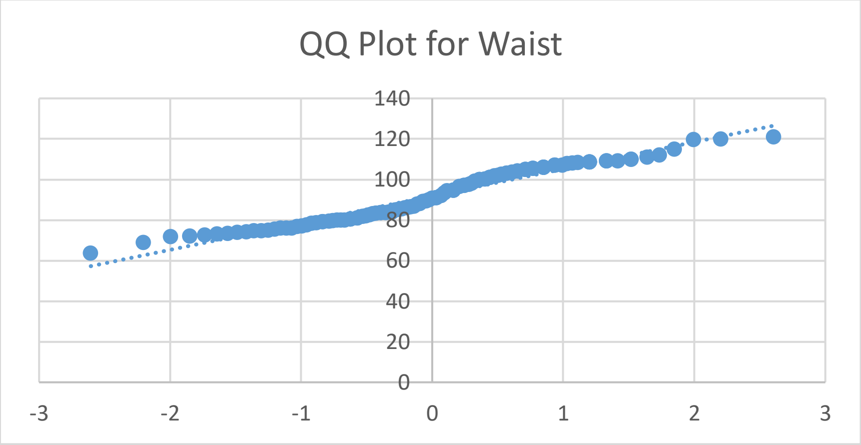
For MPG	//
Mean	34.4220757
Mode	29.629936
Median	44.6528342

In a normal distribution, the mean, median and mode are the same. Here, all three values are different, indicating that MPG is not normally distributed. Further, since the median is greater than the mean, we can say that the distribution has a negative or left-skew with a longer tail on the left.

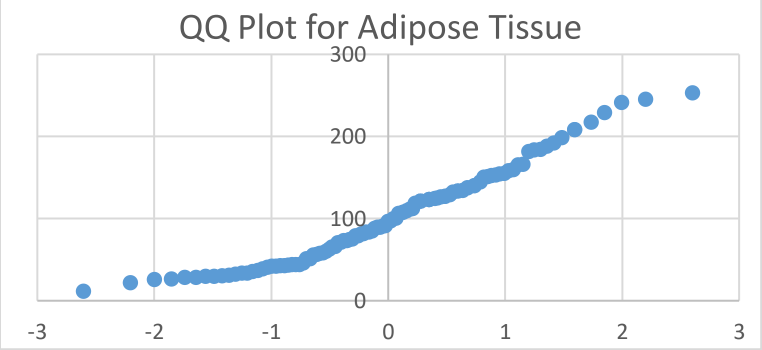


The datapoints are closer to the 45-degree line, indicating that the distribution is close to normal. However, when we look towards the left, the data seems to deviate from the straight 45-degree line more, telling us the distribution shows a slight left skew.

b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
   Dataset: wc-at.csv

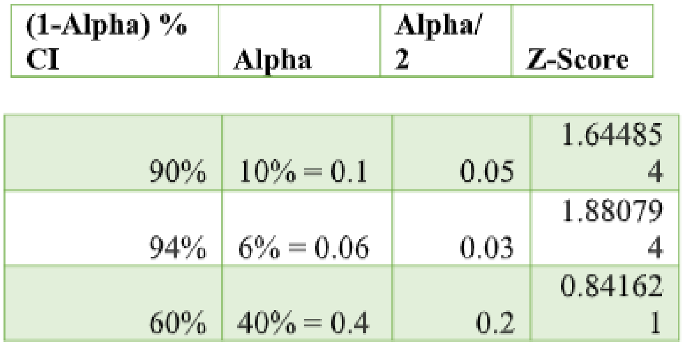


According to the Q-Q plot for Waist, most of the datapoints lie along the 45-degree line, making it close to normal distribution, but there are more deviations towards the ends. So, the distribution of Waist is close to a normal distribution.



According to the Q-Q plot for Adipose Tissue, most of the datapoints, especially along the ends are not following the 45-degree line. So, we can conclude that the datapoints are not normally distributed.

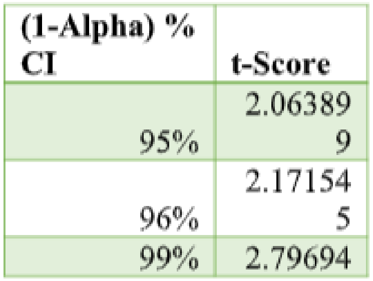
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval



(1-Alpha) % CI	Alpha	Alpha/ 2	Z-Score
90%	10% = 0.1	0.05	1.644854
94%	6% = 0.06	0.03	1.880794
60%	40% = 0.4	0.2	0.841621

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

Sample size/n= 25
Degrees of freedom= n-1=25-1=24



(1-Alpha) % CI	t-Score
95%	2.063899
96%	2.171545
99%	2.79694

Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint: rcode → pt(tscore,df)

df → degrees of freedom

Population mean = 270 days

Sample Size = 18

Sample mean= 260 days

To solve this problem, we will use a normal distribution. First, we need to calculate the mean and the standard deviation of the distribution. Since we are sampling 18 bulbs, the mean of the distribution will be the same as the average life of the bulbs that the government company claims, which is 270 days. The standard deviation of the distribution will be the standard deviation of the sample divided by the square root of the sample size, which is 90/sqrt(18)=15 days.

Now that we have the mean and standard deviation of the distribution, we can use the normal distribution formula to calculate the probability that the average life of 18 randomly selected bulbs is no more than 260 days. This probability can be calculated as follows:

P(x <= 260) = 1 - P(x > 260) = 1 - 0.5 * (1 + erf((260 - 270) / (15 * sqrt(2)))) = 0.0228

Therefore, the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.0228, or 2.28%.