Answers to Exercise Problems

Q3.1 [Ans: (a) 32; (b) 16]

Q3.2 [Ans: 0.0153]

Q3.3 [Ans: 0.9728]

Q3.4 [Ans: (a) 0.94; (b) 0.95]

Q3.5 [Ans: (a) 0.612; (b) 1.0]

Q3.6 [Ans: 0.0245]

Q3.7 [Ans: (a) 0.0205; (b) 0.439; (c) 0.282 $\downarrow$]

4 Descriptive Statistics

Numerical Summary of Data

The *sample average*, also known as the *sample mean*, is a *measure of central tendency* that represents the arithmetic average of a set of data points or observations from a sample.

- Mathematically, the formula for calculating the sample average ($\bar{x}$) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  where n is the number of observations and x_i, $i = 1, 2, \cdots, n$, represents each individual data point.

- The sample average is *sensitive to extreme values or outliers* in the data, as it incorporates all data points equally.

Numerical Summary of Data (cont'd)

The *sample variance* is a *measure of how spread out* the values in a sample are from the sample mean. It quantifies the variability or dispersion of data points in the sample.

- Mathematically, the sample variance (s^2) can be calculated as:

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

  where n represents the number of observations in the sample, x_i represents each individual data point, and $\bar{x}$ is the sample mean.

- The denominator $n-1$ is the *correction* that is applied to make the sample variance an *unbiased estimator* of the population variance.
- Together, the *sample variance* and *sample average* provide important insights into the characteristics of a dataset.
- The *sample standard deviation* (s) is the square root of the sample variance, i.e., $s = \sqrt{s^2}$. It provides a measure in the *original units* and is easier to interpret.

Exercise Problems

Q4.1 A quality engineer is analyzing the lifetimes of a batch of light bulbs. The lifetimes (in hours) of a sample of 10 light bulbs are partially recorded due to some of the bulbs still functioning at the end of the study period. The observed lifetimes are as follows: $500, 600, 700, 800, 900, 1000, 1100, 1200+$, $1300+, 1400+$. Here, $1200+$, $1300+$, and $1400+$ indicate *censored data* points, meaning that those bulbs were still working at 1200, 1300, and 1400 hours, respectively, when the observation period ended. Calculate the sample mean and standard deviation of the lifetime data.

[4] Descriptive Statistics • § Numerical Summary of Data

© 2025 Indranil Hazra

Exercise Problems

Q4.1 A quality engineer is analyzing the lifetimes of a batch of light bulbs. The lifetimes (in hours) of a sample of 10 light bulbs are partially recorded due to some of the bulbs still functioning at the end of the study period. The observed lifetimes are as follows: $500, 600, 700, 800, 900, 1000, 1100, 1200+$, $1300+, 1400+$. Here, $1200+$, $1300+$, and $1400+$ indicate *censored data* points, meaning that those bulbs were still working at 1200, 1300, and 1400 hours, respectively, when the observation period ended. Calculate the sample mean and standard deviation of the lifetime data.

Q4.2 Calculate the sample mean and variance of the (*mixed*) lifetime data of a batch of electronic components as shown below.

1. *Complete data:* $5, 7, 12, 20$ hours
2. *Right-censored data:* $15+$ (component still functioning at 15 hours), $18+$ (component still functioning at 18 hours)
3. *Left-censored data:* < 3 (component failed before 3 hours), < 4 (component failed before 4 hours)
4. *Interval-censored data:* $(6, 10)$ (component failed between 6 and 10 hours), $(8, 12)$ (component failed between 8 and 12 hours)

Stem-and-Leaf Plots

A stem-and-leaf plot is a *graphical representation of a dataset* that is used to display the *distribution* and *variation* of the data values.

- It is useful for understanding the *distribution* of data, identifying *patterns*, and spotting *outliers*.
- A stem-and-leaf plot is a visual representation of a dataset where each data point is split into a *stem* (leading digits) and a *leaf* (trailing digit).
- The *stems* are typically *arranged vertically* in *ascending* order, and the *leaves* are listed *horizontally* next to their respective stems.
- *Clusters of leaves* near a stem indicate *concentrations* of data values, and outliers can be easily identified. The plot allows for a visual assessment of data patterns and irregularities.

Stem-and-Leaf Plots (cont'd)

Construct a stem and leaf plot for the duration of calls that Bob makes each day.

Date	Minutes
Jan. 1	67
Jan. 2	7
Jan. 3	3
Jan. 4	47
Jan. 5	32
Jan. 6	39
Jan. 7	17
Jan. 8	62
Jan. 9	2

Stem-and-Leaf Plots (cont'd)

Construct a stem and leaf plot for the duration of calls that Bob makes each day.

Date	Minutes
Jan. 1	67
Jan. 2	7
Jan. 3	3
Jan. 4	47
Jan. 5	32
Jan. 6	39
Jan. 7	17
Jan. 8	62
Jan. 9	2

Step 1: Sort the data in minutes: 02, 03, 07, 17, 32, 39, 47, 62, 67
Step 2: Identify the stems: 0, 1, 3, 4, 6
Step 3: Identify the leaves: (0)2, (0)3, (0)7, (1)7, (3)2, (3)9, (4)7, (6)2, (6)7
Step 4: Arrange the *stems* vertically in ascending order and arrange the *leaves* horizontally next to their respective stems.
Step 5: Provide the *key.*

Stem	Leaf
0	2 3 7
1	7
2	
3	2 9
4	7
5	
6	2 7

Key: $1|7 = 17$ minutes

Stem-and-Leaf Plots (cont'd)

Construct a stem and leaf plot for the quiz scores of RE30003 students.

Date	Minutes
Student	Score
Student 1	7.6
Student 2	8.2
Student 3	7.5
Student 4	6.7
Student 5	6.2
Student 6	7.9
Student 7	8.7
Student 8	9.2

Stem-and-Leaf Plots (cont'd)

Construct a stem and leaf plot for the quiz scores of RE30003 students.

Date	Minutes
Student	Score
Student 1	7.6
Student 2	8.2
Student 3	7.5
Student 4	6.7
Student 5	6.2
Student 6	7.9
Student 7	8.7
Student 8	9.2

Step 1: Sort the data in minutes: 6.2, 6.7, 7.5, 7.6, 7.9, 8.2, 8.7, 9.2

Step 2: Identify the stems: 6, 7, 8, 9

Step 3: Identify the leaves: (6.)2, (6.)7, (7.)5, (7.)6, (7.)9, (8.)2, (8.)7, (9.)2

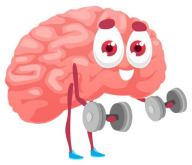
Step 4: Arrange the *stems* vertically in ascending order and arrange the *leaves* horizontally next to their respective stems.

Step 5: Provide the *key.*

Stem	Leaf
6	2 7
7	5 6 9
8	2 7
9	2
10	

Key: $6|2 = 6.2$ score

Exercise Problem



Q4.3 You are a quality engineer at a manufacturing plant that produces precision gears for automotive applications. You are asked to analyze the diameters of a sample of gears to ensure they meet quality standards. The diameters (in mm) of 10 gears are measured as follows: $54.2, 54.8, 55.0, 55.3, 55.7, 56.1, 56.5, 56.8, 57.2, 57.5$.

(a) Create a stem-and-leaf plot for the given data.
(b) Use the stem-and-leaf plot to identify any potential issues in the gear diameters.
(c) Discuss whether the diameters are centered around a specific value and if there are any outliers.

Box-and-Whisker Plots

A box-and-whisker plot, also known as a *box plot*, is a *graphical representation* of the distribution of a dataset. It provides a visual summary of the central tendency, spread, and skewness of the data.

- A box-and-whisker plot displays the distribution of a dataset using a *rectangular box and lines (whiskers)* that extend from the box to show the range of the data.
- The *length of the box* indicates the spread of the *central 50% of the data,* while the *whiskers* represent the range of the entire dataset. The position of the *median* within the box provides insight into the central tendency.
- Box-and-whisker plots effectively *highlight outliers*, making it easy to identify values that fall significantly outside the typical range of the dataset.
- Box plots are useful for *comparing the distribution* of different datasets, as differences in central tendency, spread, and skewness are readily apparent.

Box-and-Whisker Plots (cont'd)

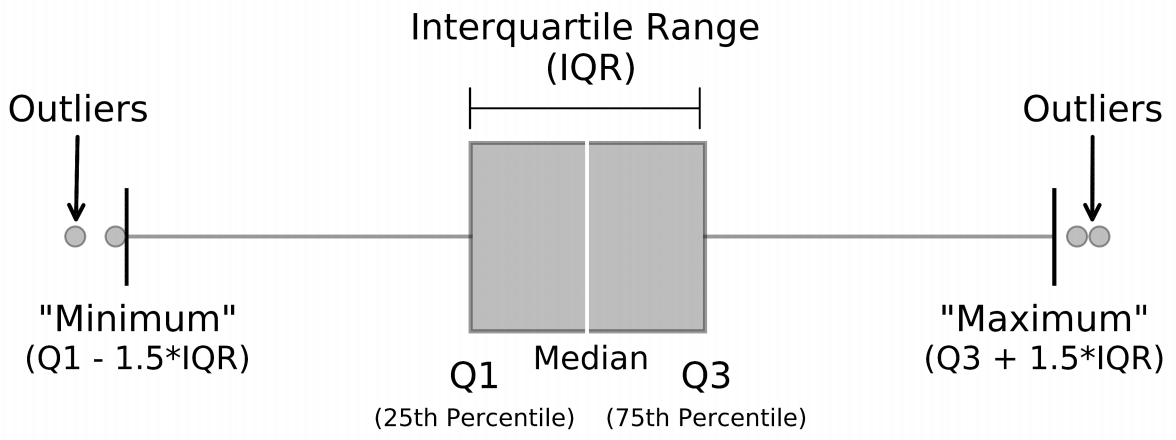
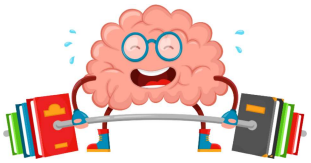


Figure: The box represents the *interquartile range (IQR)*, with the median marked by a horizontal line inside the box. Whiskers extend from the box to the minimum and maximum values within a specified range (often 1.5 times the IQR). Outliers, individual data points beyond the whiskers, may be plotted individually. The symmetry or skewness of the data distribution can be visually assessed by observing the placement and length of the whiskers.

Exercise Problem



Q4.4 A precisely engineered opening or hole on the leading edge of the aircraft wing is necessary for various purposes, including structural, aerodynamic, or other functional reasons. Create a box-and-whisker plot for the aircraft wing leading edge hole diameter (mm) data from 12 aircraft as shown below.

120.5	120.4	120.7	120.9
120.2	121.1	120.3	120.1
120.9	121.3	120.5	120.8

Box-and-Whisker Plots (cont'd)

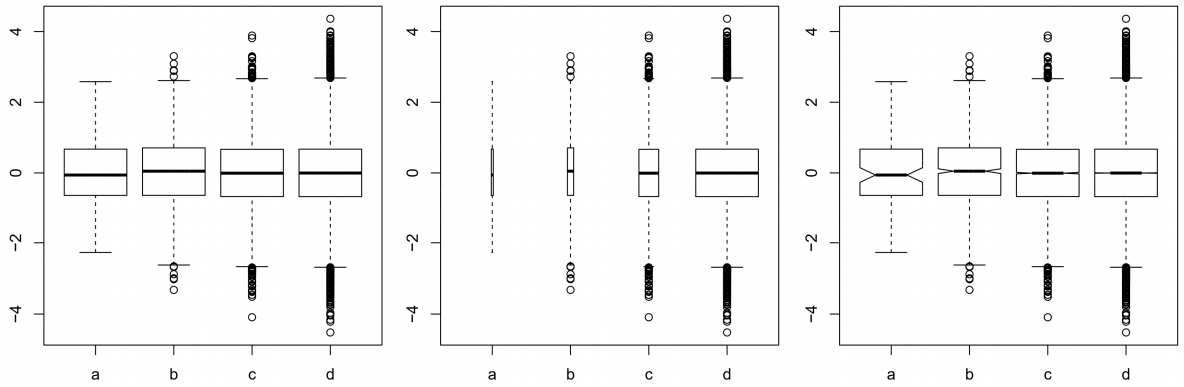


Figure: Box plots are relatively robust to the shape of the distribution and are particularly useful for comparing datasets of varying sizes. Box plot variations showing 100, 1000, 10000, and 100000 numbers drawn from a standard normal distribution. In a *regular box plot* the only hint that the groups are different sizes is the number of outliers (*Left*). A *variable-width box plot* shows the differences in group size (*Middle*). The *notched box plots* displays the error associated with the estimate of the median (*Right*).

Richer Displays of Densities

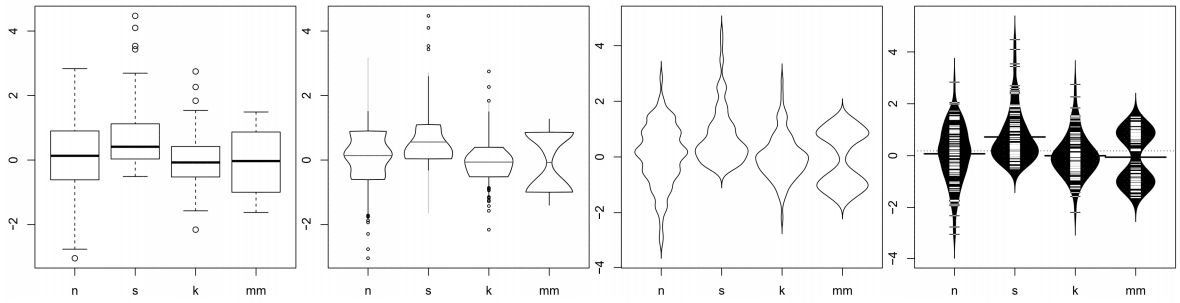


Figure: From left to right: *box plot, vase plot, violin plot and bean plot*. Within each plot, the distributions from left to right are: *standard normal (n), right-skewed (s), leptikurtic (k), and bimodal (mm).* These variations attempt to display more information about the distribution, maintaining the compact size of the box plot, but bringing in the richer distributional summary of the data. In a vase plot, the box is replaced with a symmetrical display of estimated density. Violin plots are very similar, but display the density for all data points, not just the middle half. The bean plot is an enhancement that adds a rug showing every value and a line for the mean.

Histograms

A histogram is a graphical representation of the distribution of a dataset. It displays the *frequency of data points within specified intervals*, providing a visual summary of the variation and shape of the data.

- The *shape* of the histogram (e.g., *symmetrical, skewed, uniform*) provides insights into the underlying distribution of the data. *Peaks and valleys* in the histogram indicate concentration or dispersion of values.
- The center of the histogram corresponds to the mode of the dataset. The width of the bars and the overall spread of the histogram illustrate the variability or spread of the data.
- *Skewness is visually apparent* in the asymmetry of the histogram. *Positive* skewness has a long tail to the right, while *negative* skewness has a long tail to the left.
- Histograms facilitate the *comparison of different datasets* and can reveal similarities or differences in their distributions.
- The *choice of bin width* can influence the appearance of the histogram. Adjusting bin width allows for a balance between capturing details and smoothing out noise in the data.

Histograms (cont'd)

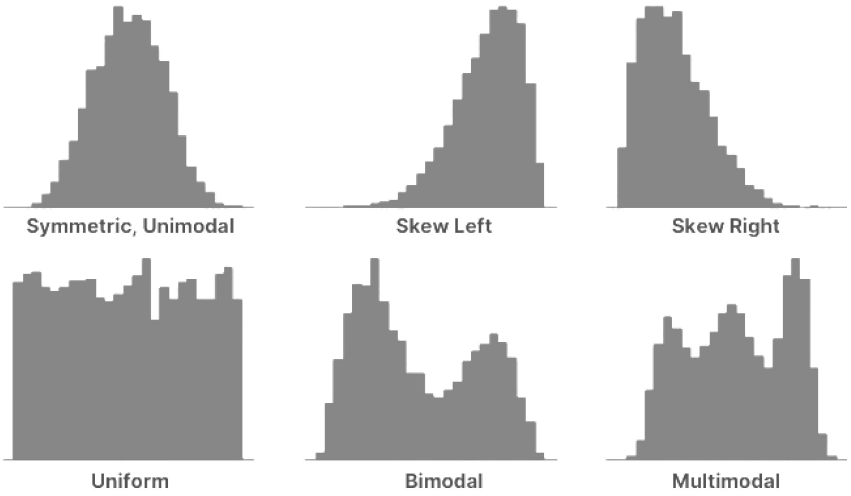


Figure: A *histogram* is constructed by dividing the data range into intervals, or bins, and counting the number of data points that fall into each bin. The *x-axis represents the range of values* in the dataset, divided into intervals. The *y-axis* represents the *frequency or relative frequency* of data points within each interval. Bars are drawn above each interval, with the height of the bar corresponding to the frequency of data points in that interval.

Histograms (cont'd)

Layer Thickness (Å) on Semiconductor Wafers

438	450	487	451	452	441	444	461	432	471
413	450	430	437	465	444	471	453	431	458
444	450	446	444	466	458	471	452	455	445
468	459	450	453	473	454	458	438	447	463
445	466	456	434	471	437	459	445	454	423
472	470	433	454	464	443	449	435	435	451
474	457	455	448	478	465	462	454	425	440
454	441	459	435	446	435	460	428	449	442
455	450	423	432	459	444	445	454	449	441
449	445	455	441	464	457	437	434	452	439

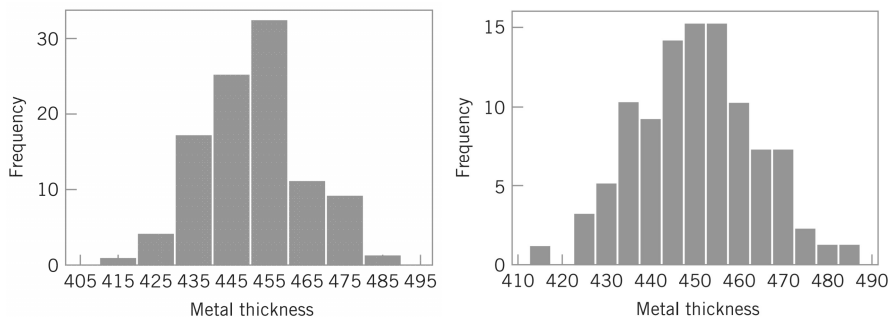


Figure: The table shows metal layer thickness data on 100 silicon wafers resulting from a chemical process in a semiconductor plant. Note the reasonably symmetric or bell-shaped distribution of the metal thickness data. *Choosing the number of bins* approximately equal to the *square root of the number ($\sqrt{n}$) of observations* often works well in practice ($\sqrt{100} = 10$). There is no universal agreement about how to select the number of bins for a histogram. Some suggests *Sturges's rule*, which sets the number of bins equal to $1 + \log_2 n$, where n is the sample size. There are also many variations of Sturges's rule.

Probability Distributions

A probability distribution is a mathematical function that describes the *likelihood of various outcomes* in a random experiment. It assigns probabilities to different events or values of a *random variable*.

- Probability distributions are associated with *random variables*, which are variables whose values are outcomes of a random process. Random variables can be *discrete* or *continuous*.
- When the variable being measured is expressed on a continuous scale, its probability distribution is called a *continuous distribution*.
- When the parameter being measured can only take on certain values, such as the integers, $0, 1, 2, \cdots$, the probability distribution is called a *discrete distribution*.

Probability Distributions (cont'd)

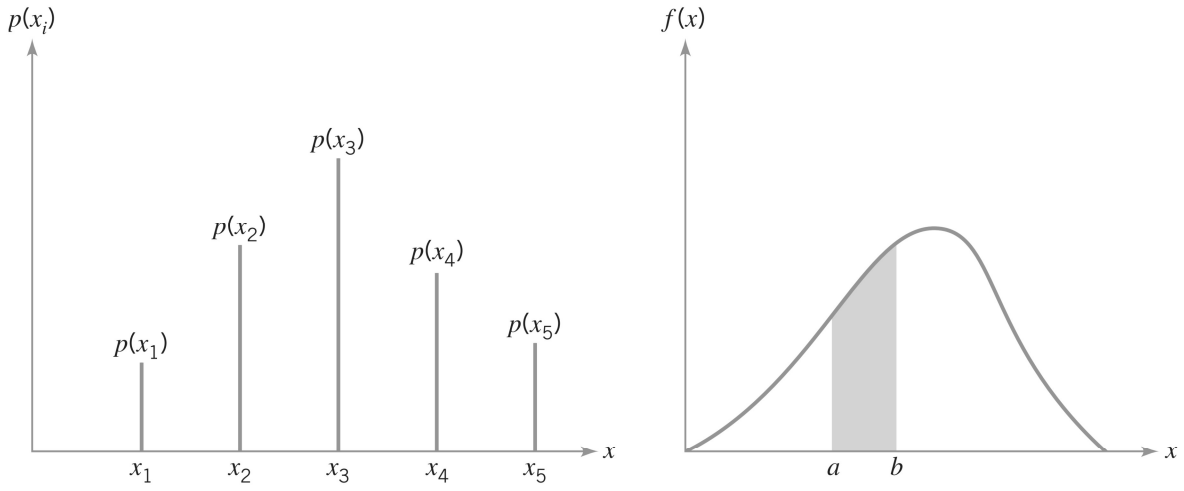


Figure: *Discrete (left) & continuous (right) probability distributions*. For discrete random variables, the probability distribution is expressed as a *probability mass function (PMF)*. PMF provides the probability of each possible value of a discrete random variable. For continuous random variables, the probability distribution is described by a *probability density function (PDF)*. PDF represents the likelihood of a random variable falling within a particular range.

Probability Distributions (cont'd)

The *probability mass function (PMF)* of a discrete random variable gives the *probability of each specific outcome*. The probability that a discrete random variable X takes on a specific value x_i can be written as

$$P\{X = x_i\} = p(x_i)$$

The *probability density function (PDF)* of a continuous random variable gives the probability density at a particular value. The appearance of a continuous distribution is that of a smooth curve, with the *area under the curve equal to probability*, so that the probability that X lies in the interval from a to b can be obtained as

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx$$

Probability Distributions (cont'd)

The *mean* μ of a probability distribution is a measure of the central tendency in the distribution, or its location. The mean is defined as

$$\mu = \begin{cases} \sum_{i=1}^{n} x_i p(x_i) & \text{(Discrete)} \\ \int_X x f(x)dx & \text{(Continuous)} \end{cases}$$

The scatter, spread, or variability in a distribution is expressed by the *variance* σ^2, which is defined as

$$\sigma^2 = \begin{cases} \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i) & \text{(Discrete)} \\ \int_X (x - \mu)^2 f(x)dx & \text{(Continuous)} \end{cases}$$

The *standard deviation* σ of a probability distribution is a measure of spread or scatter in the population expressed in the original units.

Probability Distributions (cont'd)

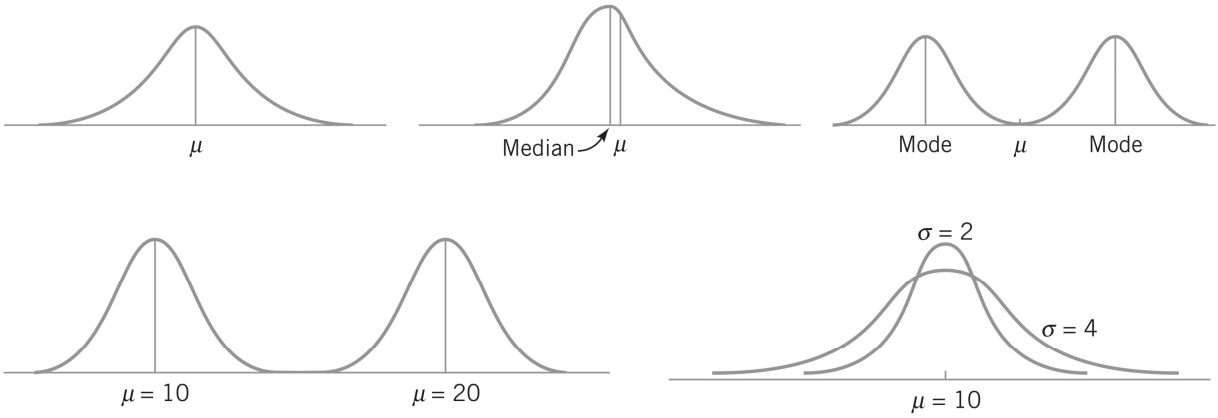
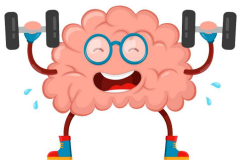


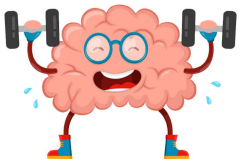
Figure: The mean, median, and mode of a distribution *(top)*; two probability distributions with different means *(bottom left)*; two probability distributions with the same mean but different standard deviations *(bottom right)*. The *mean* is not necessarily the fiftieth percentile of the distribution (which is the *median*), and it is not necessarily the most likely value of the variable (which is called the *mode*). The mean determines the location of the distribution, whereas, the standard deviation is a measure of spread or scatter in the population.

Exercise Problems



Q4.5 A manufacturing process produces thousands of semiconductor chips per day. On the average, 1% of these chips do not conform to specifications. Every hour, an inspector selects a random sample of 25 chips and classifies each chip in the sample as conforming or nonconforming. What is the probability of finding one or fewer nonconforming parts in the sample?

Exercise Problems



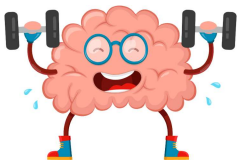
Q4.5 A manufacturing process produces thousands of semiconductor chips per day. On the average, 1% of these chips do not conform to specifications. Every hour, an inspector selects a random sample of 25 chips and classifies each chip in the sample as conforming or nonconforming. What is the probability of finding one or fewer nonconforming parts in the sample?

Q4.6 Suppose that X is a random variable representing the actual content (kg) of an LPG cylinder. The probability distribution of X is assumed to be:

$$f_X(x) = \frac{1}{1.5}, \quad 15.5 \leq x \leq 17.0$$

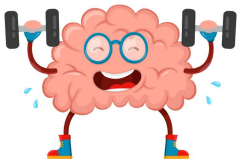
This is a continuous distribution, since the range of X is in the interval $[15.5, 17.0]$. Determine the probability of a gas cylinder containing less than 16.0 kg LPG.

Exercise Problems (cont'd)



Q4.7 A quality assurance team at a beverage company is designing a system to ensure that the fill levels of their soda cans fall within a certain specification. Suppose the fill level Y (in ml) of these soda cans is uniformly distributed between 340 ml and 360 ml. To maintain high quality, the team wants to set a threshold such that only 5% of the cans are filled more than this threshold. Determine the threshold fill level t that meets this criterion.

Exercise Problems (cont'd)



Q4.7 A quality assurance team at a beverage company is designing a system to ensure that the fill levels of their soda cans fall within a certain specification. Suppose the fill level Y (in ml) of these soda cans is uniformly distributed between 340 ml and 360 ml. To maintain high quality, the team wants to set a threshold such that only 5% of the cans are filled more than this threshold. Determine the threshold fill level t that meets this criterion.

Q4.8 A research laboratory is testing the purity levels of a new chemical compound. The purity level P (in %) of the samples is uniformly distributed between 95% and 105%. To ensure that the compound meets quality standards, the laboratory wants to establish a range of acceptable purity levels such that 90% of the samples fall within this range. Determine the lower threshold and the upper threshold such that 90% of the samples have a purity level between these two thresholds.

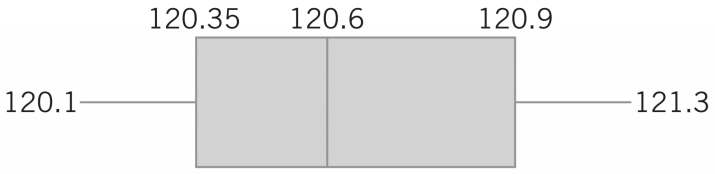
Answers to Exercise Problems

Q4.1 [Ans: (a) 950; (b) 302.77]

Q4.2 [Ans: (a) 10.2; (b) 35.06]

Q4.3 [Ans: (1) The diameters appear to be centered around 55 to 56 mm, which indicates good consistency in the manufacturing process. There is a fairly symmetric distribution around these values. (2) There are no extreme values (outliers) that deviate significantly from the rest of the data, suggesting that the process is stable and in control.]

Q4.4 [Ans: Median ≈ 120.6 mm; $Q_1 \approx 120.35$ mm; $Q_3 \approx 120.9$ mm; $IQR \approx 0.55$ mm]



Q4.5 [Ans: ≈ 0.9742]

Q4.6 [Ans: ≈ 0.3333]

Q4.7 [Ans: ≈ 359 ml]

Q4.8 [Ans: $\approx 95.5\%, 104.5\%$]