Chapter 1: Descriptive Statistics

Detailed Solutions

1.1 Basic Concept

1.1 Data Classification

Problem: Classify variables as qualitative, quantitative discrete, or quantitative continuous.
Solution:

(a) **Failure time:** Time is a measurement that can be infinitely precise (e.g., 100.54 hours).

(b) **Number of defects:** This is a count (0, 1, 2...).

(c) **Alloy grade:** These are categories with a specific order (A $>$ B $>$ C).

(d) **Temperature:** A physical measurement.

(e) **Zip code:** These are numerical labels; mathematical operations (like average) do not make sense.

(a) Quantitative Continuous

(b) Quantitative Discrete

(c) Qualitative (Ordinal)

(d) Quantitative Continuous

(e) Qualitative (Nominal)

1.2 Measures of Center Properties

Problem: Compare Mean, Median, and Mode.
Solution:

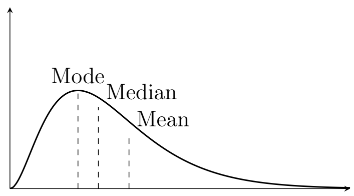
(a) **Robustness:** The Median is based on the *rank* or *position* of the data, not the magnitude of all values. Therefore, a single extreme outlier will not shift the median significantly, whereas the mean will be pulled towards the outlier.

(b) **Skewed Right Distribution:** In a positively skewed distribution, the tail extends to the right (large values). These large values pull the Mean upward. The Mode remains at the peak.

(c) **Uniqueness:** The Mean is the center of gravity and is always unique for a given dataset. The Mode is the most frequent value; there can be ties (bimodal, multimodal) or no mode at all.

> (a) **Median.** It is robust against outliers.
>
> (b) **Mean > Median > Mode.**
>
> (c) Mean: **No** (Unique). Mode: **Yes** (Can have multiple).

Right-Skewed Distribution



1.3 Parameter vs. Statistic

Problem: Identify Parameter or Statistic.
Solution:

- **Parameter:** A numerical summary of a **Population** (all members).
- **Statistic:** A numerical summary of a **Sample** (subset).

> (a) **Parameter** (All 40 mayors constitute the entire population of interest).
>
> (b) **Statistic** (Sample of 100 bags is a subset).
>
> (c) **Statistic** (50 voters are a sample from the voting population).

1.2 Intermediate

1.4 Fizzy Drinks Analysis

Problem: Calculate $n, \bar{x}, \text{Median}, S^2, S$ from frequency table.
Solution:

Data Table:

x	f	fx	x^2	fx^2
0	25	0	0	0
1	30	30	1	30
2	26	52	4	104
3	20	60	9	180
4	14	56	16	224
5	10	50	25	250
$\sum$	**125**	**248**		**788**

(a) **Sample Size:** $n = \sum f = 125$.

(b) **Sample Mean:**

$$\bar{x} = \frac{\sum fx}{n} = \frac{248}{125} = 1.984$$

(c) **Median:** Position $= \frac{n+1}{2} = \frac{126}{2} = 63^{\text{rd}}$ value. Cumulative frequency:

- $x = 0$: 25
- $x = 1$: $25 + 30 = 55$
- $x = 2$: $55 + 26 = 81$ (Values 56 to 81 are all 2)

The 63rd value falls in the $x = 2$ category. Median $= 2$.

(d) **Variance & SD:** Using computational formula:

$$S^2 = \frac{1}{n-1}\left(\sum fx^2 - \frac{(\sum fx)^2}{n}\right)$$

$$S^2 = \frac{1}{124}\left(788 - \frac{(248)^2}{125}\right) = \frac{1}{124}(788 - 492.032) = \frac{295.968}{124} \approx 2.3868$$

$$S = \sqrt{2.3868} \approx 1.545$$

(a) $n = 125$
(b) $\bar{x} = 1.984$ cans
(c) Median $= 2$ cans
(d) $S^2 \approx 2.39$, $S \approx 1.55$

1.5 Stem-and-Leaf Plot & Outliers

Problem: Data: $\{77, 78, 76, 81, 86, 51, 79, 82, 84, 99\}$.
Solution:

Ordered Data ($n = 10$): $51, 76, 77, 78, 79, 81, 82, 84, 86, 99$.

(a) **Mean & SD:**

$$\sum x = 793, \quad \bar{x} = 79.3$$

$$\sum x^2 = 64549, \quad S^2 = \frac{64549 - 10(79.3)^2}{9} = \frac{1664.1}{9} = 184.9$$

$$S = \sqrt{184.9} \approx 13.60$$

(b) **Stem-and-Leaf:**

$$\begin{array}{c|l} 5 & 1 \\ 6 & \\ 7 & 6\ 7\ 8\ 9 \\ 8 & 1\ 2\ 4\ 6 \\ 9 & 9 \end{array}$$

(c) **Outliers:** Position of $Q_1 = 0.25(11) = 2.75 \to 3^{rd}$ value $= 77$. Position of $Q_3 = 0.75(11) = 8.25 \to 8^{th}$ value $= 84$.

$$IQR = 84 - 77 = 7$$

Lower Fence: $Q_1 - 1.5(IQR) = 77 - 10.5 = 66.5$. Upper Fence: $Q_3 + 1.5(IQR) = 84 + 10.5 = 94.5$. Values outside $[66.5, 94.5]$ are outliers. Outliers: **51** and **99**.

(d) **Correction (51 $\to$ 71):** New Data: $71, 76, 77, \ldots$. The sum increases by 20, so $\bar{x}_{new} = (793 + 20)/10 = 81.3$. The Median depends on the middle values ($5^{th}, 6^{th}$). Old Median: Avg(79, 81) = 80. New Sorted: $71, 76, 77, 78, 79, 81, 82, 84, 86, 99$. Middle is still 79, 81. Median remains **80**.

(a) $\bar{x} = 79.3, S \approx 13.60$.
(b) See Plot above.
(c) Outliers are **51** and **99**.
(d) Mean increases to 81.3; Median remains 80.

1.6 Box Plot Interpretation

Problem: Min=10, $Q_1 = 20$, Med=35, $Q_3 = 50$, Max=90.
Solution:

(a) **IQR:** $50 - 20 = 30$.

(b) **Fences:** Upper $= 50 + 1.5(30) = 95$. Lower $= 20 - 1.5(30) = -25$. (Since Min=10, whisker stops at 10).

(c) **Outlier Check:** Max=90. Since $90 < 95$, it is **not** an outlier.

(d) **Skewness:** Distance Q_1 to Median $= 35 - 20 = 15$. Distance Median to $Q_3 = 50 - 35 = 15$. (Box is symmetric). Whisker Length: Left $= 20 - 10 = 10$. Right $= 90 - 50 = 40$. The right whisker is much longer, indicating **Positive Skew (Right Skewed)**.

(a) $IQR = 30$.
(b) Fences: $[-25, 95]$.
(c) No.
(d) Positively Skewed (Skewed Right).

1.7 Combined Mean and Variance

Problem: Class A ($n = 20, \bar{x} = 80, S^2 = 25$), Class B ($n = 30, \bar{x} = 70, S^2 = 36$).
Solution:

Combined Mean:

$$\bar{x}_c = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A + n_B} = \frac{20(80) + 30(70)}{50} = \frac{1600 + 2100}{50} = \frac{3700}{50} = 74$$

Combined Variance: We need the sum of squares ($\sum x^2$) or use the ANOVA-like decomposition.

$$SS_{Total} = SS_{Within} + SS_{Between}$$

$$SS_{Within} = (n_A - 1)S_A^2 + (n_B - 1)S_B^2 = 19(25) + 29(36) = 475 + 1044 = 1519$$

$$SS_{Between} = n_A(\bar{x}_A - \bar{x}_c)^2 + n_B(\bar{x}_B - \bar{x}_c)^2$$

$$= 20(80 - 74)^2 + 30(70 - 74)^2 = 20(36) + 30(16) = 720 + 480 = 1200$$

$$SS_{Total} = 1519 + 1200 = 2719$$

$$S_c^2 = \frac{SS_{Total}}{n_{total} - 1} = \frac{2719}{49} \approx 55.49$$

Combined Mean $\bar{x}_c = 74$
Combined Variance $S_c^2 \approx 55.49$

1.8 Coefficient of Variation (CV)

Problem: Compare precision.
Solution:

Formula: $CV = \frac{S}{\bar{x}}$.

- Machine 1: $CV = \frac{0.5}{10} = 0.05$ (or 5%).
- Machine 2: $CV = \frac{2}{100} = 0.02$ (or 2%).

Lower CV indicates higher precision relative to the mean.

Machine 2 is relatively more precise (2% < 5%).

1.9 Energy Source Analysis

Problem: Pie Chart Analysis (Assuming data: Coal 25%, Gas 45%, Renew 20%, Nuclear 8%, Hydro 2%).
Solution:

(a) **Primary Source:** The largest slice is Natural Gas (45%).

(b) **Clean Energy Target:** Clean = Renewables (20%) + Nuclear (8%) + Hydro (2%) = 30%. Target is 30%. Yes, they met exactly the target.

(c) **Coal Generation:** Total = 10,000 MWh. Coal share = 25%.

$$\text{Energy}_{Coal} = 0.25 \times 10,000 = 2,500 \text{ MWh}$$

(a) Natural Gas
(b) 30%. Yes, target met.
(c) 2,500 MWh

1.10 Process Stability (Time Series)

Problem: Defect rate trend (Target $< 2.0\%$).
Solution:

(a) **Trend Jan-Jun:** The graph shows an increasing trend. The process is **deteriorating**.

(b) **Highest Month:** June appears to be the peak ($> 3.0\%$). *Hypothesis:* Summer temperatures affecting sensitive equipment, or high turnover of staff/interns in June.

(c) **Last Quarter Avg:** Reading from graph (approx): Oct (1.5%), Nov (1.4%), Dec (1.2%).

$$\text{Avg} = \frac{1.5 + 1.4 + 1.2}{3} = 1.37\%$$

(d) **Failures:** Months above 2.0 line: May, June, July, August, September. Total **5 months**.

(a) Deteriorating.
(b) June. (Heat/Staffing).
(c) $\approx 1.37\%$.
(d) 5 months.

1.11 Precision (Dot Plot)

Problem: Deviations from 0.00 mm. Spec ± 0.05.
Solution:

(a) **Center:** The dots cluster symmetrically around 0.00. The process is centered.

(b) **Range:** The smallest dot is at -0.04, largest at $+0.04$. Range is $[-0.04, 0.04]$.

(c) **Outliers:** No points are isolated far from the main cluster. No outliers.

(d) **Defects:** All points fall within $[-0.04, 0.04]$, which is inside the spec limits $[-0.05, 0.05]$. Defect rate $= 0\%$.

(a) Centered at 0.00.
(b) Range ≈ 0.08 mm.
(c) None.
(d) 0%.

1.3 Challenge

1.12 Minimizing Squared Deviations

Problem: Prove $g(a) = \sum(x_i - a)^2$ is minimized at $a = \bar{x}$.
Solution:

To minimize, take the derivative with respect to a and set to 0.

$$\frac{d}{da}\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n} 2(x_i - a)(-1) = -2\sum_{i=1}^{n}(x_i - a)$$

Set to 0:

$$-2\left(\sum x_i - \sum a\right) = 0$$

$$\sum x_i - na = 0 \implies na = \sum x_i \implies a = \frac{\sum x_i}{n} = \bar{x}$$

Check 2nd derivative: $\frac{d^2 g}{da^2} = \sum 2 = 2n > 0$, confirming a minimum.

$$\boxed{\text{Proven: } a = \bar{x}}$$

1.13 Sum of Deviations

Problem: Prove $\sum(x_i - \bar{x}) = 0$.
Solution:

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

Since $\bar{x}$ is a constant:

$$= \sum x_i - n\bar{x}$$

Substitute $\bar{x} = \frac{\sum x_i}{n}$:

$$= \sum x_i - n\left(\frac{\sum x_i}{n}\right) = \sum x_i - \sum x_i = 0$$

$$\boxed{\text{Proven: Sum is 0}}$$

1.14 Linear Transformation of Variance

Problem: If $y_i = ax_i + b$, prove $S_y^2 = a^2 S_x^2$.
Solution:

First, find $\bar{y}$:

$$\bar{y} = \frac{\sum(ax_i + b)}{n} = a\bar{x} + b$$

Now substitute into variance formula:

$$S_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\sum((ax_i + b) - (a\bar{x} + b))^2$$

The b terms cancel out:

$$= \frac{1}{n-1}\sum(a(x_i - \bar{x}))^2 = \frac{1}{n-1}\sum a^2(x_i - \bar{x})^2$$

Factor out a^2:

$$= a^2\left[\frac{1}{n-1}\sum(x_i - \bar{x})^2\right] = a^2 S_x^2$$

$$\boxed{\text{Proven: } S_y^2 = a^2 S_x^2}$$

1.15 Computational Formula for Variance

Problem: Derive $S^2 = \frac{1}{n-1}(\sum x^2 - n\bar{x}^2)$.
Solution:

Start with numerator $\sum(x_i - \bar{x})^2$:

$$\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2$$

Substitute $\sum x_i = n\bar{x}$ and $\sum \bar{x}^2 = n\bar{x}^2$:

$$= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2$$

$$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

Dividing by $n-1$ gives the computational formula.

$$\boxed{\text{Proven.}}$$

1.4 Application

1.16 Excel Formulas

Problem: Data in A1:A50.
Solution:

(a) Mean: `=AVERAGE(A1:A50)`

(b) Sample SD: `=STDEV.S(A1:A50)`

(c) Median: `=MEDIAN(A1:A50)`

(d) Count > 50: `=COUNTIF(A1:A50, ">50")`

See formulas above.

1.17 Data Analysis Toolpak Interpretation

Problem: Skewness = 0.05, Kurtosis = 1.5.
Solution:

(a) **Symmetry:** Skewness is very close to 0 (0.05). The distribution is approximately **symmetric**.

(b) **Heavy Tails:** Excess Kurtosis is 1.5 (positive). This indicates a **Leptokurtic** distribution, meaning it has heavier tails (more outliers) and a sharper peak than a normal distribution (which has excess kurtosis 0).

(a) Symmetric.
(b) Yes, heavy tails (Leptokurtic).

1.18 Exploratory Data Analysis (EDA)

Problem: Python output analysis. Data has outlier 25.0.
Solution:

(a) **Stats:** Sum of normal values $\approx 10.1 \times 14 = 141.4$. Total Sum $= 141.4 + 25 = 166.4$. Mean $\approx 166.4/15 = 11.09$. Max value is 25.0.

(b) **Mean vs Median:** The outlier (25.0) pulls the **Mean** upwards significantly (11.09 vs typical ~ 10.1). The **Median** remains robust, likely around 10.2.

(c) **Boxplot:** The value 25.0 will be displayed as an individual **point (dot)** well above the top whisker, indicating it is an outlier.

(a) Mean ≈ 11.1, Max $= 25.0$.
(b) Mean is inflated; Median is unaffected.
(c) Represented as a dot (outlier).