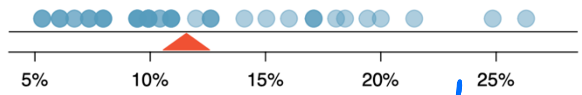
MAT1372, Classwork4, Fall2025

2.1 Examining Numerical Data

1. Dot plots and the mean



(1) Dot plots: A dot plot is a one-variable scatterplot.

(2) Mean: The mean, denoted by $\bar{x}$, is a common way to measure the center of a distribution of data.

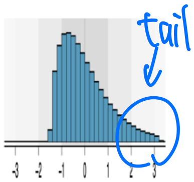
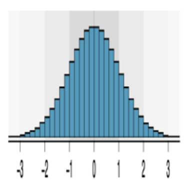
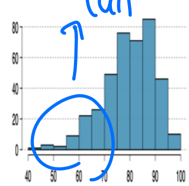
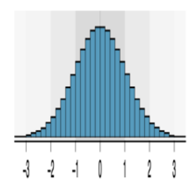
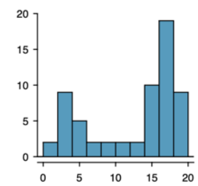
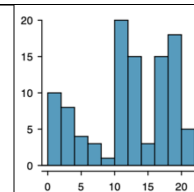
It can be computed as the sum of the observed values divided by the number of the observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$ where $x_1, x_2, \cdots, x_n$ represent the n observed values.

(3) Sample mean $\bar{x}$ and population mean μ: $\bar{x} \longrightarrow \mu$ (there is a natural bias b/c $\bar{x} \neq \mu$)

(4) Weighted mean: Some case's variable is more important than the same variable from other cases

2. Histograms and the shape: (1) Histogram: It provides a view of the data density.

(2) Shape			(3) Mode of a distribution		
tail 		tail 			
Right skew	symmetric	Left skew	unimodal	bimodal	multimodal

(2) Skewness: a distribution with a long tail

(3) Mode of a distribution: It is represented by the number of the prominent peaks.

3. Variance and Standard Deviation

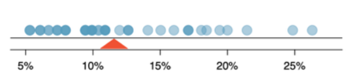
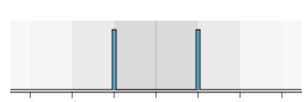
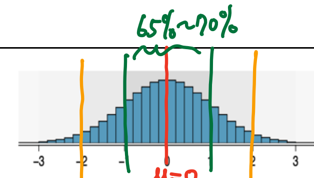
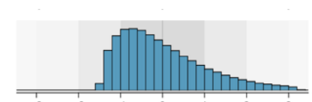
(1) Deviation	(2) Sample Variance s^2	(3) Sample standard deviation s
$x_i - \bar{x}$ for all $i = 1,2,\cdots,n$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

(4) Bessel's correction: $S^2_{n,\bar{x}} \leq S^2_{\mu}$ population variance $S^2_{\mu} = \sum_{i=1}^{n}(x_i - \mu)^2$

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

(5) Besides mean and standard variance, modality or skewness plays a role in the description of a distribution.

Why?

$\mu = 0$, $\sigma = 1$, $\sigma = 1$ $\mu = 0$, $\sigma = 1$ standard deviation

4. What can the standard deviation tell us about the data?

The standard deviation represents the typical deviation of observations from the mean. Usually about 70% of the data will be with one standard deviation of mean. About 90% of the data will be with two standard deviations of mean.

5. (Bessel's correction) Given a mid-term grade for 10 students in a certain Math course:

Amy	Bert	Barry	Doug	Emily	Howard	Leo	Penny	Raj	Wil
92	95	70	95	60	30	50	70	78	80

and a sample from these 10 grades: $x = \{70, 60, 30, 50, 70\}$. Find (a) μ, (b) $\bar{x}$, (c) $s_{n,\mu}$, (d) $s_{n,\bar{x}}$, (e) s

(a)

$$\mu = \frac{92+95+70+95+60+30+50+70+78+80}{10} = \frac{720}{10} = 72$$

(b)

$$\bar{x} = \frac{70+60+30+50+70}{5} = \frac{280}{5} = 56$$

(c)

$$S_{n,\mu} = \sqrt{\frac{\sum_{i=1}^{5}(x_i-\mu)^2}{5}} = \sqrt{\frac{(70-72)^2+(60-72)^2+(30-72)^2+(50-72)^2+(70-72)^2}{5}} = \sqrt{480}$$

(d)

$$S_{n,\bar{x}} = \sqrt{\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2}{5}} = \sqrt{\frac{(70-56)^2+(60-56)^2+(30-56)^2+(50-56)^2+(70-56)^2}{5}} = \sqrt{\frac{1120}{5}} = \sqrt{260}$$

(e)

$$S = \sqrt{\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2}{4}} = \sqrt{\frac{1120}{4}} = \sqrt{280}$$

4 ← Bessel's correction

6. Box plots, Quartiles and the median

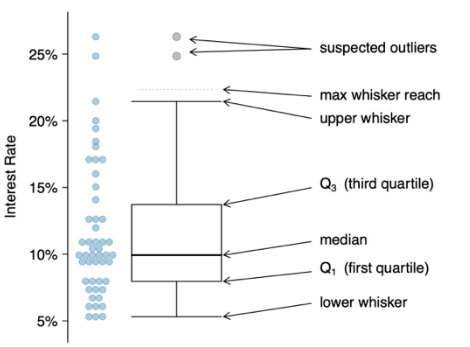


Figure 2.10: A vertical dot plot, where points have been horizontally stacked, next to a labeled box plot for the interest rates of the 50 loans.

(1) Median: ______________________________

(2) The first quartile Q_1: ______________________________

(3) The third quartile Q_3: ______________________________

(4) The interquartile range IQR = $Q_3 - Q_1$: ______________________________

(5) The whiskers: upper one ______________________________

and lower one ______________________________

(6) Outlier: ______________________________

7. Robust statistic:

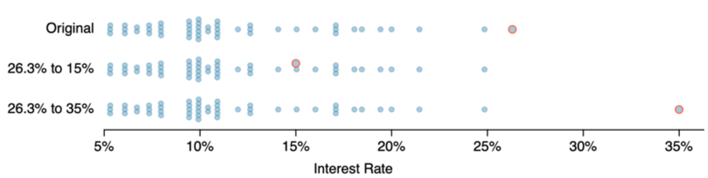


Figure 2.11: Dot plots of the original interest rate data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `interest_rate` data	9.93%	5.76%	11.57%	5.05%
move 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

Figure 2.12: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change had an extreme observations from the `interest_rate` variable been different.