Review

Types of data

Nominal: classification of sample (or population) units into categories.
Ex) The brand of toothpaste preferred by each individual in a sample.

Ordinal: rank orders the sample (or population) units.
Ex) Rating from 1-10 for a sample of movies.

Interval: enables comparison of sample (or pop.) units according to difference between values.
Ex) Temperature at which each sample of plastic begins to melt.

Ratio: measurements that enable the determination of how many times as much of the characteristic being measured is possessed by one unit of the sample (or pop.) than another.

Ex) Unemployment rate in US for each of past 60 months.

Qualitative vs Quantitative data

Histograms

Ex] How many hours people exercise per week

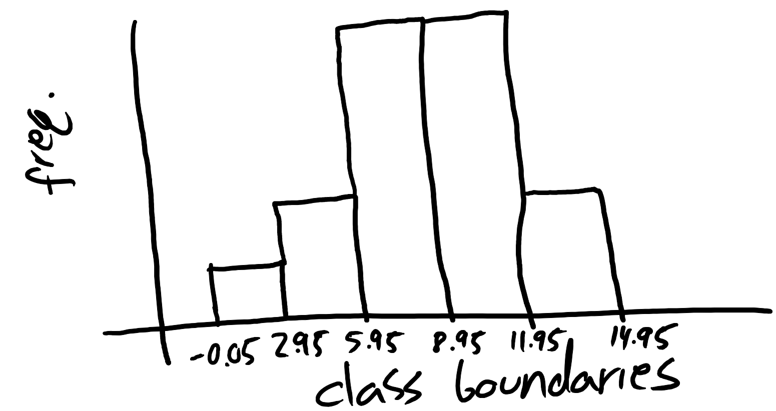
$d = \{8, 2, 4, 7.5, 10, 11, 5, 6, 8, 12, 11, 9, 6.5, 10.5, 13\}$

frequency table:

class	frequency
0 – 2.9	1
3 – 5.9	2
6 – 8.9	5
9 – 11.9	5
12 – 14.9	2

class boundaries $\{-0.05, 2.95, 5.95, 8.95, 11.95, 14.95\}$

class width $= 3$



approx. bell shaped

Five number summary and boxplots

min, Q_1, Q_2, Q_3, max

$$s = \text{sort}(d) = \{\overset{1}{2}, \overset{2}{4}, \overset{3}{5}, \overset{4}{6}, \overset{5}{6.5}, \overset{6}{7.5}, \overset{7}{8}, \overset{8}{8}, \overset{9}{9}, \overset{10}{10},$$
$$\overset{11}{10.5}, \overset{12}{11}, \overset{13}{11}, \overset{14}{12}, \overset{15}{13}\}$$

$min = 2$, $max = 13$

$$L_{25} = \left[\frac{25}{100}\right] \times 15 = 3.75 \Rightarrow \lceil L_{25} \rceil = 4$$

$$Q_1 = s[4] = 6$$

$$L_{50} = \left[\frac{50}{100}\right] \times 15 = 7.5 \Rightarrow \lceil L_{50} \rceil = 8$$

$$Q_2 = s[8] = 8$$

$$L_{75} = \left[\frac{75}{100}\right] \times 15 = 11.25 \Rightarrow \lceil L_{75} \rceil = 12$$

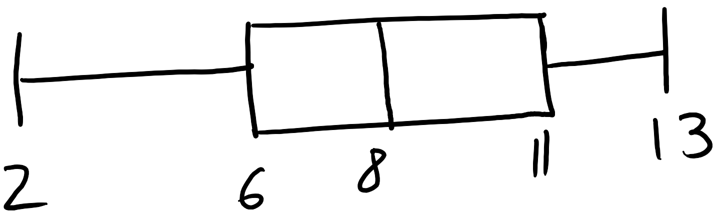
$$Q_3 = s[12] = 11$$

$$IQR = Q_3 - Q_1 = 11 - 6 = 5$$

$$\left.\begin{array}{l} Q_1 - 1.5\, IQR = -1.5 \\ Q_3 + 1.5\, IQR = 18.5 \end{array}\right\} \text{no measurements outside } (-1.5, 18.5)$$

no outliers by IQR characterization

$2, 6, 8, 11, 13 = (\min, Q_1, Q_2, Q_3, \max)$



boxplot

mean $\quad \bar{x} = \frac{\sum x_i}{n}$ (both sample and population)

std dev
variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \quad \text{sample variance}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

$\bar{x}_d = 8.23$; $s \simeq 3.13$

$\bar{x}_d \pm s = (5.10, 11.36)$

By Empirical rule, 68% of measurements should lie in this interval.

$\Rightarrow$ 10/15 lie in interval $\Rightarrow$ 66.7%

Chebyshev's thm applies to $(\bar{x}d \pm ks)$ for k whole # > 1.

Z-scores $\quad z_i = \frac{x_i - \bar{x}}{s}$

weighted mean

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

Elements with a high weight contribute more to the weighted mean than do elements with a low weight.

Probability

Ex) Suppose a ball is drawn at random from a box containing 6 red balls, 4 white balls, and 5 blue balls.

Determine the probability that it is

(a) red, (b) white, (c) blue, (d) not red

(e) red or white.

$$P(R) = \frac{6}{15} = \frac{2}{5}$$

$$P(W) = \frac{4}{15}$$

$$P(B) = \frac{5}{15} = \frac{1}{3}$$

$$P(R') = 1 - P(R) = 1 - \frac{2}{5} = \frac{3}{5}$$

$$P(R \text{ or } W) = P(R \cup W)$$

note that $R \cap W = \emptyset$ (mutually exclusive events)

$$P(R \cup W) = P(R) + P(W) = \frac{2}{5} + \frac{4}{15} = \frac{2}{3}$$

Also:

$$P(R \cup W) = P(B') = 1 - P(B) = 1 - \frac{1}{3} = \frac{2}{3}$$

Also

$$P(R \cup W) = \frac{6+4}{6+4+5} = \frac{10}{15} = \frac{2}{3}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$
$$- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)$$
$$+ P(A_1 \cap A_2 \cap A_3)$$

Ex) A doctor has 520 patients of which

(1) 230 are hypertensive

(2) 185 are diabetic

(3) 35 are hypochondriac and diabetic

(4) 25 are all three

(5) 150 are none

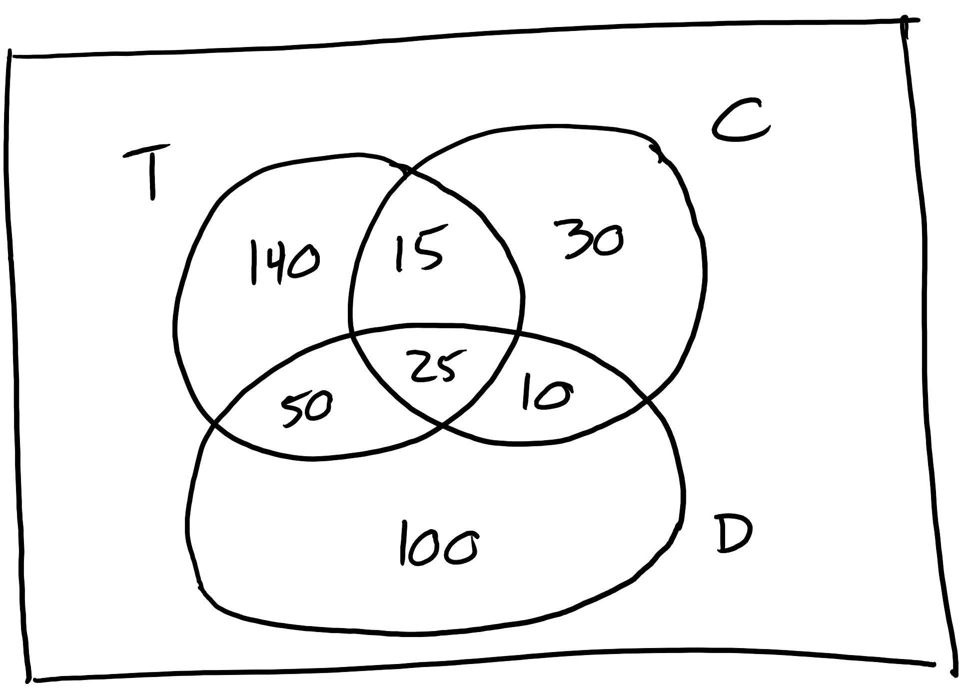
(6) 140 are only hypertensive

(7) 15 are hypertensive and hypochondriac but not diabetic.

Find the probability that the doctor's next appointment is hypochondriac but neither diabetic nor hypertensive.

Let T, C, D denote the events that the next appointment is hypertensive, hypochondriac, and diabetic, respectively.

Venn diagram:



desired probability is $\frac{30}{520} \approx 0.06$

Ex) In tossing four dice, what is the probability of at least one 3?

A: event of at least one 3

A^c: the event of no 3 in tossing 4 dice.

$$P(A^c) = \frac{N(A^c)}{N} = \frac{5^4}{6^4}$$ (tosses of dice independent)

$$P(A) = 1 - \frac{5^4}{6^4} = 1 - \frac{625}{1296} \approx 0.52$$

Birthday problem

In a class of size n, the probability that no two students have the same birthday is:

$$P(n) = \frac{365 \times 364 \times \ldots \times [365-(n-1)]}{365^n}$$

$$1 - P(50) \approx 0.970$$