Probability (continued from last time)

Ex) Toss a coin 3 times. What is the probability of obtaining two heads?

Consider the sample space of outcomes:

$\{HHH, \underbrace{HHT}_{e_2}, \underbrace{HTH}_{e_3}, \underbrace{THH}_{e_4}, HTT, THT, TTH, TTT\}$

A: obtain two heads in 3 tosses

$P(A) = \frac{3}{8}$. Notice that $A = e_2 \cup e_3 \cup e_4$.

The probability that a composite event A will occur is the sum of probabilities of the simple events of which it is composed.

$$P(A) = P(e_2) + P(e_3) + P(e_4) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Notice that e_2, e_3, e_4 are mutually exclusive events.

Ex| What is the probability of getting at least 10 points in rolling two dice?

A_1: get 10 pts in two rolls

A_2: get 11 pts in two rolls

A_3: get 12 pts in two rolls

$A_1 \cap A_2 \cap A_3 = \phi$

C: get at least 10 points in rolling two dice

$C = A_1 \cup A_2 \cup A_3$ (union of 3 simple events)

$P(C) = P(A_1) + P(A_2) + P(A_3)$ since A_1, A_2, A_3 are mutually exclusive. They have no outcomes in common.

# of possible outcomes: $6 \times 6 = 36$ (two dice)

A_1: 6 and 4, 4 and 6, 5 and 5

$P(A_1) = \frac{3}{36}$

A_2: 6 and 5, 5 and 6

$P(A_2) = \frac{2}{36}$

A_3: 6 and 6 (get 12 pts in 2 rolls)

$P(A_3) = \frac{1}{36}$ (only one possibility)

$\Rightarrow P(C) = P(A_1) + P(A_2) + P(A_3) = \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$

$= \frac{6}{36} = \frac{1}{6}$

notice: key is that A_1, A_2, A_3 are mutually exclusive

For any two events (not necessarily mutually exclusive), one has:

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

Ex) Suppose one card is drawn at random from a deck of 52 cards.

A: get a "red ace"

B: get a "heart"

$P(A) = \frac{2}{52}$ (two kinds of red aces in 52 cards)

$P(B) = \frac{13}{52}$ (one of the four suits)

$P(A \cap B) = \frac{1}{52}$ (only a single ace of hearts card)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{52} + \frac{13}{52} - \frac{1}{52} =$$

$$= \frac{14}{52} = \frac{7}{26}$$

Notice: $A \cup B$ means get a red ace or a heart or an ace of hearts.

$A \cap B$ means get a red ace and a heart (in one draw, a single card).

when we sum A and B, we count $A \cap B$ twice, so we need to subtract $P(A \cap B)$ from $P(A \cup B)$.

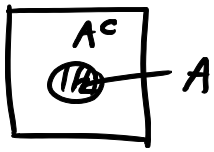
Set Relations

Independent of probability there are some fundamental relations between sets you should know.

A^c, A': complement of A

$A \cap B$ intersection 

; union $A \cup B$ 

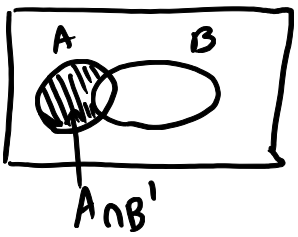


$A \cup B = (A \cap B') \cup B$

Consider x in $A \cup B$ (some element in $A \cup B$).

Then $x \in A$ or $x \in B$ or both.

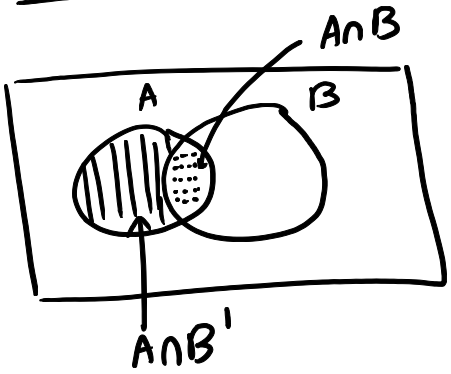
Can go through all cases and show that if an element is in $A \cup B$ it is also in $(A \cap B') \cup B$ and vice versa.



Clear that $(A \cap B') \cup B$ corresponds to

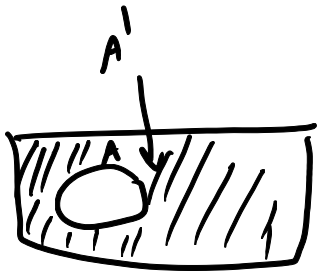
 : $A \cup B$

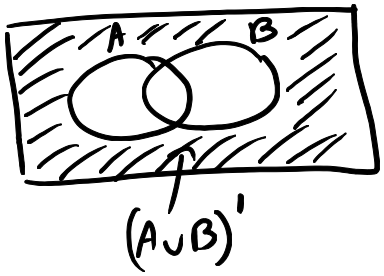
$A = (A \cap B') \cup (A \cap B)$



the combined region is set A

$(A' \cap B') = (A \cup B)'$ (de Morgan's law)



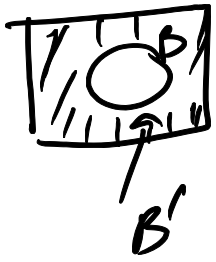


Application:

$P[(A' \cap B')] =$

$P[(A \cup B)'] = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B) - P(A \cap B)]$



Conditional probability

If $P(B) \neq 0$, then the conditional probability of A relative to B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(probability of A occurring given B)

Ex) A: randomly selected student comes from two parent home
B: student does poorly in school (avg $< D^+$)

$P(A) = 0.75$, $P(A \cap B) = 0.18$

What is the probability that a randomly selected student will be a low achiever given that he or she comes from a two-parent home.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.18}{0.75} \approx 0.24$$

$P(B|A)$: student does poorly given that student comes from two parent home

Connection to event independence

Ex) Consider two flips of a fair coin

sample space: {HH, HT, TH, TT}

A: get head on first flip

B: get head on second flip

$P(A) = \frac{2}{4} = \frac{1}{2}$; $P(B) = \frac{1}{2}$

$P(B|A) = \frac{P(B \cap A)}{P(A)}$ ← notice the order of B and A is important

$P(B \cap A) = P(A \cap B) = P(HH) = \frac{1}{4}$

$\Rightarrow P(B|A) = \frac{1/4}{1/2} = \frac{1}{2}$ (makes sense since A and B are independent!)

$P(B|A) = P(B) = \frac{1}{2}$

In this case, the probability of event B is the same regardless of whether event A has occured.

Ex) The probabilities that it will rain or snow in a given city on Christmas or New Years, or on both days are $P(C) = 0.60$, $P(N) = 0.60$, $P(C \cap N) = 0.42$

Check whether events N and C are independent.

$$P(N|C) = \frac{P(C \cap N)}{P(C)} = \frac{0.42}{0.60} = 0.70 \neq P(N)$$

Since $P(N|C) \neq P(N)$ the two events are dependent.
This makes sense; if there is a storm system it can influence weather for a week.

Ex) In rolling two fair dice, if the sum of two values is 7, what is the probability that one of the values is a 1.

A: one of the values is a 1

B: the sum of two rolls is 7

Review

Chapter 1 Intro, sampling techniques

Population vs sample. Biased vs unbiased.
(parameter) (statistic)

Ex) voluntary response surveys likely to be biased.
Only people who feel strongly about it respond.

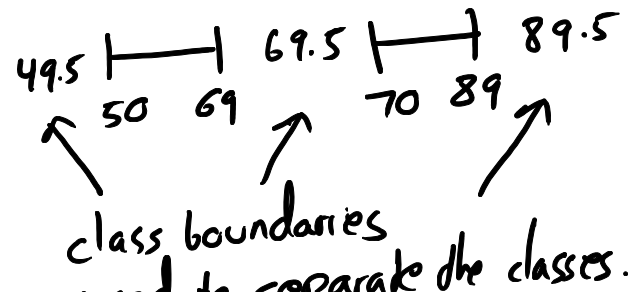
prepare and sample → analyze → conclude (statistical vs practical significance)

Chapter 2

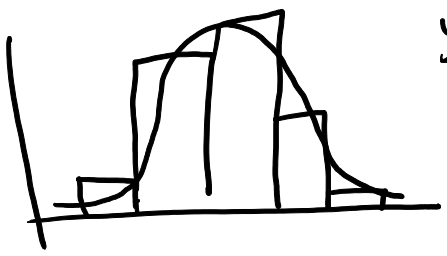
Frequency distributions and histograms

Ex1) (external) Notice that number of classes is either supplied or chosen by you. know lower/upper class limits and boundaries.

IQ score	frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1



Is it approximately normal?

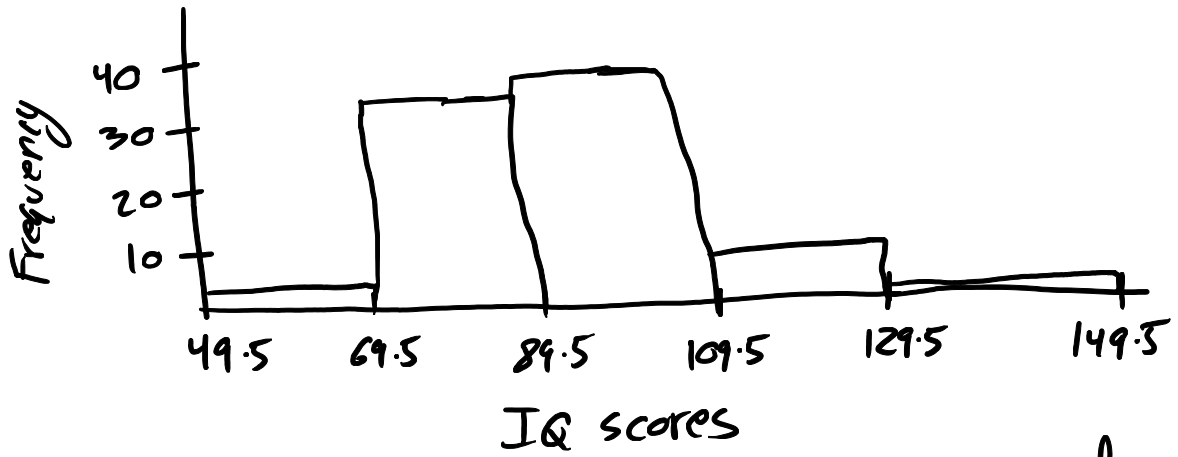


yes, approx. bell shaped

Convert frequency distribution into a cumulative distribution.

Histograms

Can be drawn from a frequency distribution after class boundaries are defined.



The most important use of histograms is to judge the distribution of the values in the data set.

Chapter 3 Measures of central tendency and variation.

Identification of outliers.

Compare mean and median (external).

mean = $\frac{\Sigma x}{n}$ median is middle of sorted data set.

mode is the measurement which occurs most frequently in the data set.

weighted mean : $\overline{x}_w = \frac{\Sigma(w \cdot x)}{\Sigma w}$ divide by sum of the weights

range = max(x) − min(x)

sample std dev : $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

population sample deviation: $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

s^2 is sample variance; σ^2 is population variance

Ex) calculate the std deviation of the following sample

$\{2, 3, 3, 3, 4\}$

$\bar{x} = \frac{2+3+3+3+4}{5} = 3$

$s = \sqrt{\frac{(2-3)^2+(3-3)^2+(3-3)^2+(3-3)^2+(4-3)^2}{5-1}} = \sqrt{\frac{2}{4}} = \sqrt{0.5}$

"shortcut" formula for sample variance

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

$\bar{x}$ and s are estimators for μ, σ

x	x^2
2	4
3	9
3	9
3	9
4	16
$\Sigma x = 15$	$\Sigma x^2 = 47$

$\Rightarrow s^2 = \frac{47 - \frac{(15)^2}{5}}{5-1} = \frac{47-45}{4} = \frac{2}{4} = 0.5$

percentiles, median, quartiles

The median is the value of middle of sort(x) when n is odd, and mean of the two middle items when n is even.

Ex) $X = \{16, 10, 14, 13, 20, 11, 17\}$

$sort(x) = \{10, 11, 13, 14, 16, 17, 20\}$

$median(x) = 14$

what if we add 30 to the end?

$sort(x_{new}) = \{10, 11, 13, \underline{14, 16}, 17, 20, 30\}$

$median(x_{new}) = \frac{14+16}{2} = 15$

percentiles and quartiles

k^{th} percentile : at least $k\%$ of data equal or less than value. at least $100-k\%$ equal or greater than value.

five number summary — first, obtain sort(x)

$min(x)$, Q_1 (25^{th} percentile), Q_2 (median), Q_3 (75^{th} percentile), $max(x)$

converting k-th percentile to value: $[1, 2, 3, 4, \underline{5, 6}, 7, 8, 9, 10]$

sort(x) → compute $L = \left(\frac{k}{100}\right) n$ where n = # of values

(a) if L is whole # → $val = \frac{[sort(x)]_L + [sort(x)]_{L+1}}{2}$

(b) if L is fraction → round up to $\lceil L \rceil$ → $val = [sort(x)]_{\lceil L \rceil}$

Ex) consider the following temperature readings for June.

$X = \{90, 75, 86, 77, 85, 72, 78, 79, 94, 82, 74, 93\}$

$Sord(x) = \{72, 74, 75, 77, 78, 79, 82, 85, 86, 90, 93, 94\}$ (indices 1 2 3 4 5 6 7 8 9 10 11 12)

Q_1 is 25^{th} percentile ; $n = 12$; $\min(x) = 72$, $\max(x) = 94$

$$L_{25} = \left(\frac{25}{100}\right)12 = \frac{1}{4} \cdot 12 = \frac{12}{4} = 3$$

$$Q_1 = \frac{sort(x)_3 + sort(x)_4}{2} = \frac{75+77}{2} = 76$$

$$Q_2 = \text{median} = \frac{79+82}{2} = 80.5 = 50^{th} \text{ percentile}$$

$$L_{75} = \left(\frac{75}{100}\right)12 = \left(\frac{15}{20}\right)12 = \left(\frac{3}{4}\right)12 = 9$$

$$Q_3 = \frac{sord(x)_9 + sort(x)_{10}}{2} = \frac{86+90}{2} = 88$$

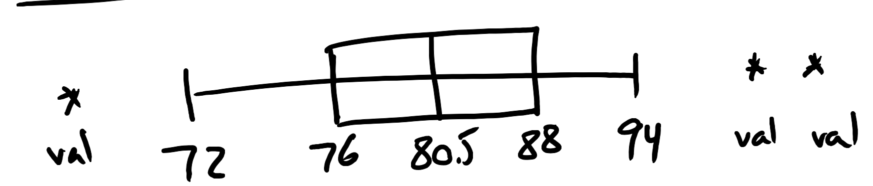
$$IQR = Q_3 - Q_1 = 88 - 76 = 12$$

outliers are values greater than $Q_3 + 1.5 IQR$ or less than $Q_1 - 1.5 IQR$.

Recall also z-sores: $z = \frac{x - \bar{x}}{s}$

For approx. normally distributed data, can use z scores to characterize outliers ($z < -2$, $z > 2$). In that case, it is roughly equivalent to the above definitions.

Modified boxplots



any outliers would be marked with asterisks

Empirical Rule and chebyshev's theorem (external)