Midterm Review.

probability - addition rule
multiplication rule.

Two events, A,
B.

P(A)

P(B)

P(A <u>and</u> B)

P(A <u>or</u> B or both)

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$ - multiplication rule.

$$= P(A) \times P(B).$$ - if A and B are independent.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$ - addition rule.

$$= P(A) + P(B).$$ - if A and B are mutually exclusive.

1 out of 100 is { ambidextrous. / mixed-handed

mixed handed children were twice as likely to have language difficulties.

$P(A \mid B)$

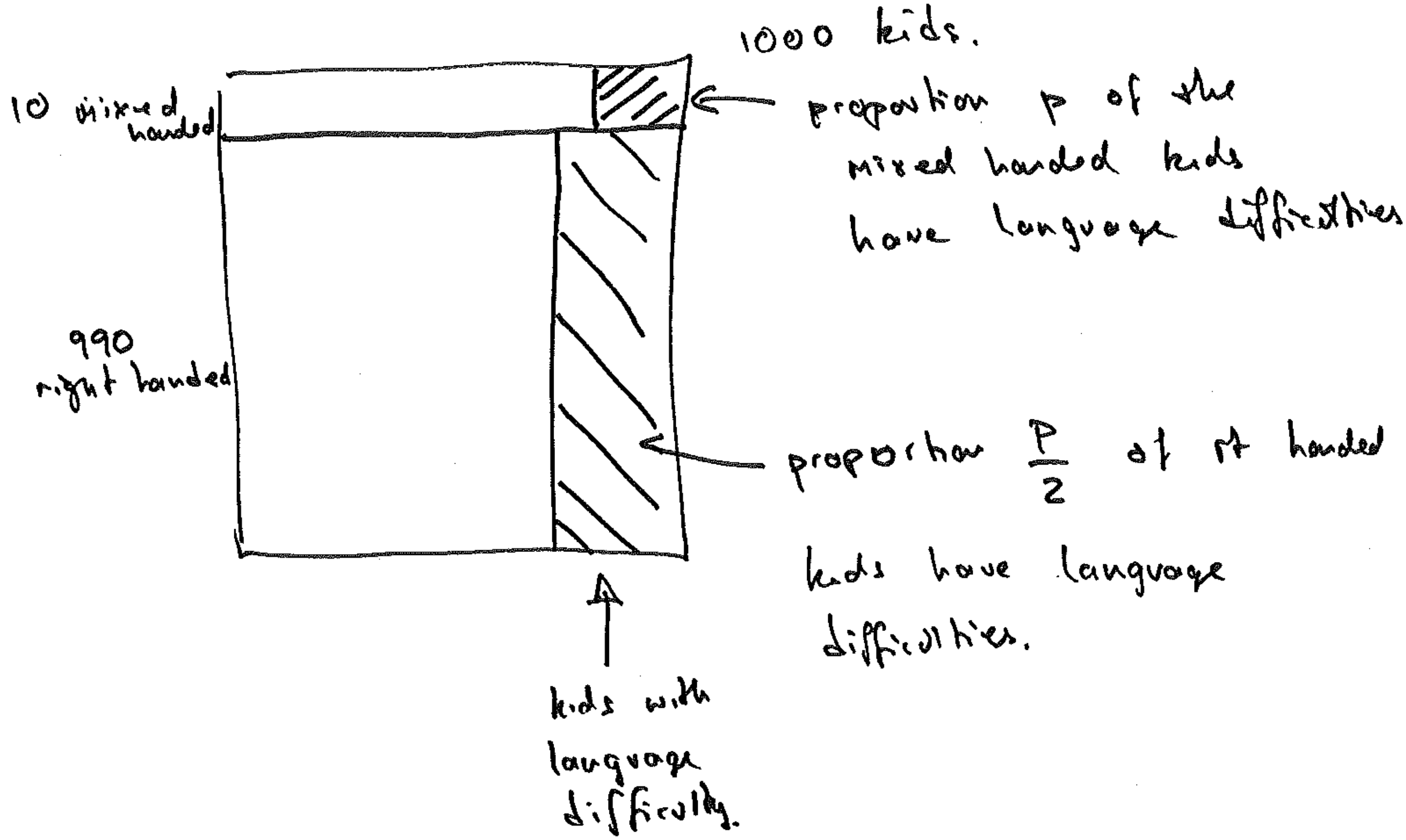
"has difficulty with language"

"is mixed handed" / "is right handed"

p(language difficulty | mixed handed).

p(language difficulty | right handed)

Told that the first is twice as large as the second.



fraction of kids with language difficulty that are mixed handed $= \dfrac{10p}{990 \times \frac{p}{2} + 10p}$

$$= \frac{10}{495 + 10} = \frac{10}{505} = \text{[illegible]}$$

$$P(\text{kid with language difficulties is mixed handed}) = \frac{1}{51} \simeq \frac{1}{50}$$

$$P(\qquad \text{is right handed}) = \frac{50}{51} \simeq \frac{49}{50}$$

If you are mixed handed, you are twice as likely to have a language difficulty than if you are right handed

<u>BUT</u>

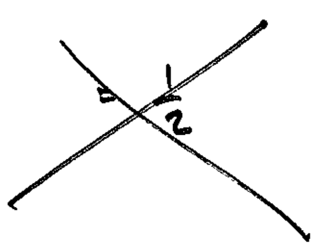
if you have a language difficulty, you are <u>50</u> times more likely to be right handed than mixed handed.

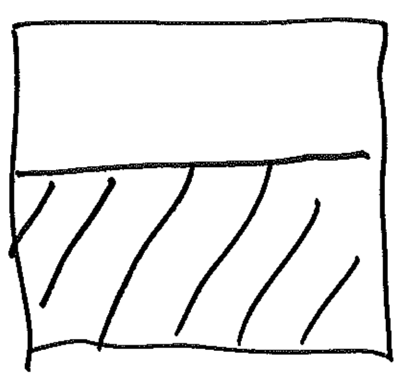
$$P(A \mid B) \neq P(B \mid A)$$

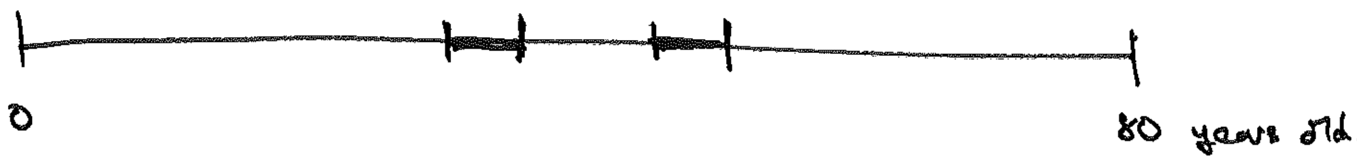
A = person is pregnant

B = person is woman.

$P(A \mid B) = P(\text{person is pregnant} \mid \text{person is a woman})$.







if average woman has 2 kids

pregnant for $1\frac{1}{2}$ years

if meet this woman at a random point in time, the chance that we meet her when she's pregnant is $\frac{1\frac{1}{2}}{80} = 1.875\%$

$\simeq \frac{1}{50}$

$p(\text{pregnant} \mid \text{woman}) \simeq \frac{1}{50}$

$p(\text{woman} \mid \text{pregnant}) = 1$

SD.

1/ find mean.

2/ subtract the mean from all the observations

3/ square the results.

4/ find the mean of the squared differences

5/ take the square root

1, 2, 3, 4, 5, 6, 7

mean $\frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$

$(1-4)^2$	9
$(2-4)^2$	4
$(3-4)^2$	1
$(4-4)^2$	0
$(5-4)^2$	1
$(6-4)^2$	4
$(7-4)^2$	9

$\frac{9+4+1+0+1+4+9}{7} = \frac{28}{7} = 4$

$SD = \sqrt{4} = 2.$

---

[1] [2] [3] [4] [5] [6]

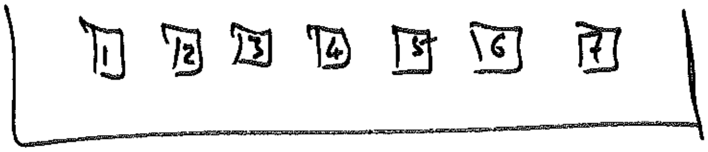
ave of box 3.5

SD of box 1.7...

[1] . . . . [7]

ave of box 4

SD 2.



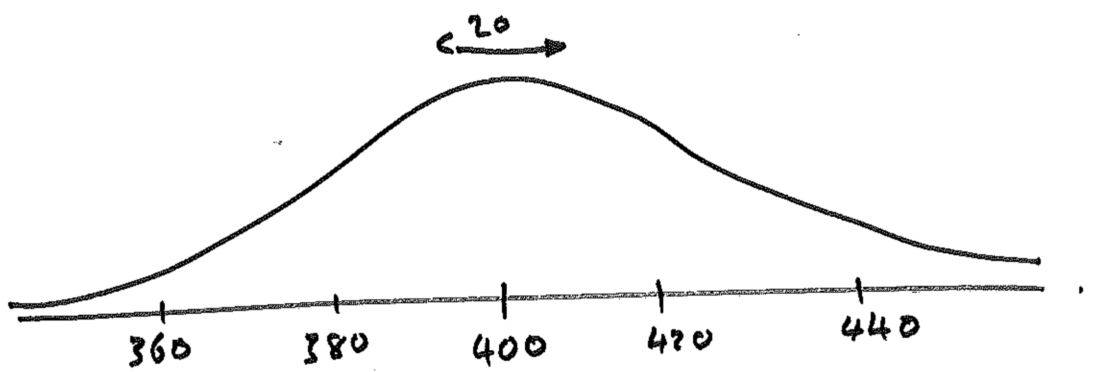
mean = 4

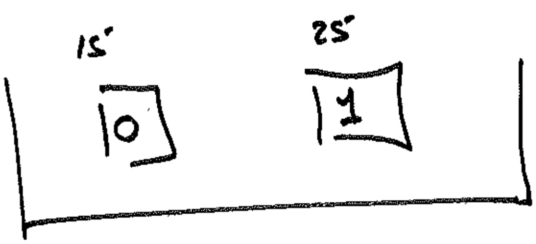
SD = 2

Roll 7 sided die 100 times.

expected value is $100 \times 4 = 400$

standard error is $\sqrt{100} \times 2 = 20$





$$SD_{box} = \left(\text{big number} - \text{small number}\right) \times \sqrt{\text{fraction of tickets with big number} \times \text{fraction with small number}}$$

$$= (1 - 0)\sqrt{\frac{25}{40} \times \frac{15}{40}}$$

4. At the start of the first class we rolled a regular 6-sided die 10 times and recorded the total number of pips. The frequency distribution of the recorded totals is given in the table below.

Class Interval (total # pips)	Frequency	%	%-per-pip
<6	~~1~~		
6-15	0		
16-25 (width 10)	4	2.4	0.24
26-30 (width 5)	23	14	2.8.
31-35 (width 5)	60	36.6	7.3
36-40	51 (139)	31.1	6.2
41-50 (width 10)	26	15.9	1.6.
51-60	0		
>60	~~4~~		

164 → 170 including outliers.

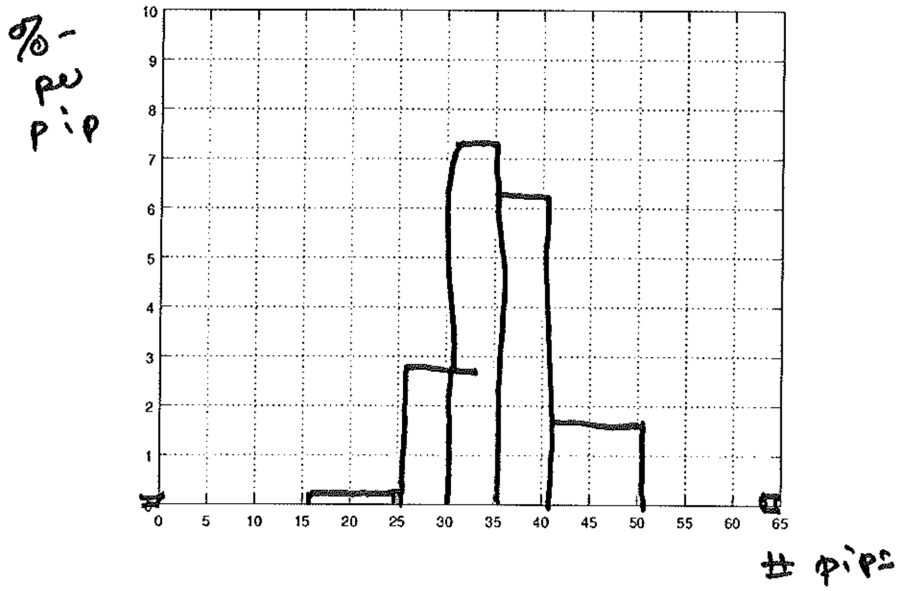
divide by. width of class interval.

$\% = \frac{\text{frequency.}}{\text{total \# entries}} \times 100$

(a) Which of these can be considered to be outliers? How did you decide they were outliers?

(b) On the graph paper on page 5 plot a histogram of the distribution of the data, excluding the data you consider as outliers. Label the axes.

[TURN OVER]

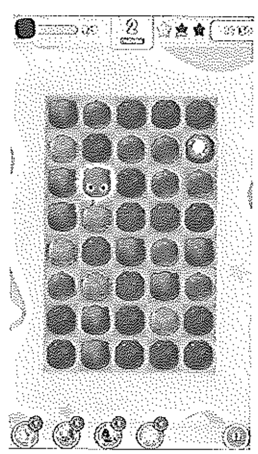


5. At the end of a level of "Jelly Splash" if the player has any remaining moves, they are converted into *"splashes"*.

   If there is one remaining move, one position is chosen at random and all the jellies in the row and column occupied by that position are exploded. 2500 bonus points are scored for the jelly under the chosen position, and 250 bonus points are scored for each of the other exploded jellies.

   If there are two remaining moves, two distinct positions are chosen at random, and all the jellies in the affected row(s) and column(s) are exploded, scoring bonus points as above (2500 for the jelly under the chosen position, and 250 for each of the other exploded jellies).

   For the board shown here



The Standard Error is the size of the Chance Error after many draws.

$$\text{SE for the sum} = \sqrt{\text{number of draws}} \times \text{SD of the box}$$

$$\text{SE for average} = \frac{\text{SE for sum}}{\text{number of draws}} = \frac{\text{SD of the box}}{\sqrt{\text{number of draws}}}$$

$$\text{SE for count} = \text{SE for sum from a 0-1 box}$$

$$\text{SE for percent} = \frac{\text{SE for count}}{\text{number of draws}} \times 100\% = \frac{\text{SD of the box}}{\sqrt{\text{number of draws}}} \times 100\%$$

Expected values and Standard errors

If we draw many times from a Box model we might add the values of draws or calculate the average of draws.

> The expected value for the sum of draws = number of draws × average of the box

> The expected value of the average of draws = average of the box

The normal approximation can be used to calculate the chance of getting specific sum or averages.

Doing calculations with the normal curve requires the use of a table. Tables are available for the standard normal curve and they require that observations be transformed to standard units.

Given a list of numbers, we convert to standard units by subtracting the average and dividing by the SD

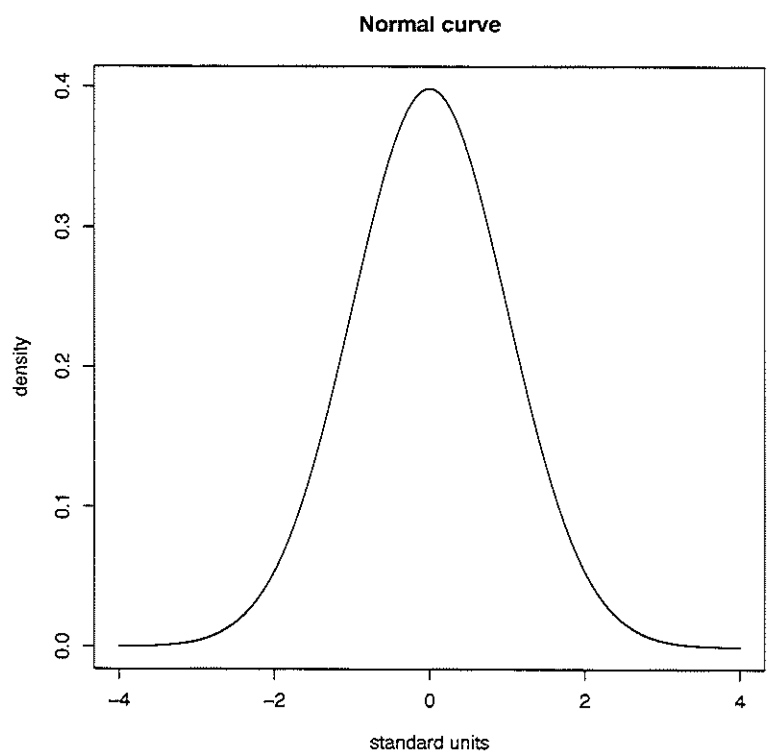
- $P((0, z)) = 1/2 \times P((-z, z))$
- $P((-z, x)) = P((-z, 0)) + P((0, x))$
- $P(> z) = 1/2 \times (P(< -z) + P(> z))$
- $P(< -z) + P(> z) = 1 - P((-z, z))$
- $P(< z) = P(< 0) + P((0, z))$
- $P((z, x)) = 1/2 \times (P((-x, x)) - P((-z, z))$

The normal density

The Gaussian or normal curve corresponds to the following formula

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad e = 2.71828\ldots$$

and corresponds to the graph



The area below the curve is equal to one. We observe that the curve is symmetric around zero and that most of the area is concentrated between -4 and 4. The probability of an interval is the corresponding area under the curve.

Assumptions in the application of the binomial formula

1. The value of n must be fixed in advance
2. p must be equal from trial to trial
3. The trials are independent

The binomial formula

Suppose hat n is the number of trials, as for example, rolling a die ten times. Let k be equal to the number of times a given event is to occur, for example, getting two ones, and p is the probability that the event will occur on any particular trial. The *binomial formula* can be written as

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

The addition rule

Two events are mutually exclusive or disjoint when the occurrence of one prevents the occurrence of the other.

If two events are disjoint then, the probability that at least one will happen is obtained by adding the probabilities of each event.

The mathematical notation for this is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Notation

Consider an event A, then the probability of A is denoted as

$$P(A)$$

Consider two events, A and B, then the conditional probability of A given B is denoted as

$$P(A|B)$$

The multiplication rule can be written as

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

A and B are independent if

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B)$$

When two events are independent the multiplication rule is

$$P(A \text{ and } B) = P(A)P(B)$$

The chances of an event are equal to the ratio of the number of outcomes corresponding to the event over the number of all possible outcomes

The probability that two events will happen equals the probability that the first will happen times the probability that the second will happen given that the first one has happened

Two events are independent if the probabilities of the second given the first are the same regardless of the outcome of the first event

Drawing at random with replacement produces independent events. Drawing without replacement produces dependent events

Probability

How do we quantify chance?

The chance of a given event is the percentage of times the event is expected to happen when the process is repeated over and over independently and under the same conditions

The chance of a given event is the amount you would be willing to bet in favour of that event to obtain a reward of one unit if the event happens and nothing if it doesn't happen

chance has to be a number between 0% and 100% (or between 0 and 1).

If an event has a given chance p of happening, the opposite has chance $1 - p$.

Average and standard deviation

The average of a list of numbers equals their sum, divided by how many they are

The median of a histogram is the value with half the area to the left and half to the right. In a symmetric histogram the median and the average coincide.

The SD of a list of numbers measures how far away they are from their average

SD = r.m.s. deviation from average.

In a histogram, the areas of the blocks represent percentages

Variables can be classified as:

- Quantitative data. Correspond to observations measured on a numerical scale. This can be:
  - Discrete when the values can differ by fixed amounts like in size.
  - Continuous differences in values can be arbitrarily small like in age.
- Qualitative data. Correspond to observations classified in groups or categories like in sex and marital status.

Collecting data: Sample Surveys

A population is a class of individuals that an investigator is interested in.

A full examination of a population requires a census. If only one part of the population is examined, then we are looking at a sample. There are usually some numerical characteristics of the population that we are interested in. These are called parameters. Parameters are unknown quantities which are estimated using statistics, which are numbers that can be computed from the sample.

> Taking a large number of samples with a biased procedure does not improve the results

When considering the quality of a survey keep in mind two possible sources of bias:

- Selection bias
- Non-response bias

Review

Collecting data: design of experiments

To eliminate bias, subjects are assigned to each group at random and the experiment is run double blind.

This is called a Controlled Experiment and allows to establish a causal effect of the treatment on the response.

In an observational study the subjects assign themselves to the different groups

Association is not causation

Relationships between percentages in subgroups can be reversed when the subgroups are combined