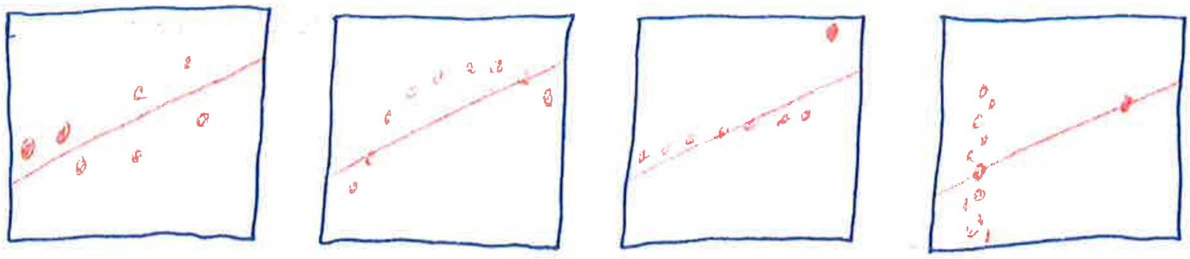
BE CAREFUL, r ALONE DOES NOT TELL THE WHOLE STORY.



ALL THESE BEST-FIT LINES ARE THE SAME, BUT THE DATASETS ARE CLEARLY VERY DIFFERENT. $r \approx 0.82$ FOR ALL OF THEM. FIGURE 3 HAS AN EXAMPLE OF AN OUTLIER, AND FIGURE 4 HAS AN EXAMPLE OF A HIGHLY INFLUENCIAL POINT (RIGHTMOST POINT IN BOTH FIGURES).

PROBABILITY

- RANDOM PHENOMENON: CANNOT BE PREDICTED WITH CERTAINTY IN ADVANCE.
- OUTCOME: SINGLE OBSERVED RESULT OF RANDOM PHENOMENON.
- SAMPLE SPACE: SET OF ALL POSSIBLE OUTCOMES.
- EMPTY SET: SET CONTAINING NO OUTCOMES.
- EVENT: SUBSET OF SAMPLE SPACE.

THE PROBABILITY OF AN EVENT IS A NUMBER $0 \leq p \leq 1$ THAT DESCRIBES HOW LIKELY IT IS THAT THE EVENT OCCURS. AN EVENT OF PROBABILITY 1 WILL HAPPEN FOR SURE; AN EVENT OF PROBABILITY 0 WILL CERTAINLY NOT HAPPEN.

$P(S)=1$, AS THE SAMPLE SPACE INCLUDES ALL POSSIBILITIES.

$P(\emptyset)=0$, AS $\emptyset$ CONTAINS NO POSSIBILITIES.

SOME PROBABILITIES CAN BE CALCULATED, OTHERS CAN BE FOUND EXPERIMENTALLY AS LONG-RUN PROPORTIONS. THEY CAN BE ADDED, PROVIDED THEY ARE DISJOINT (MUTUALLY EXCLUSIVE).

Ex: THE PROBABILITIES THAT A RANDOM STUDENT OBTAINS GRADES IN MATH 223 ARE:

F	P	C	D	HD
0.2	0.35	0.2	0.15	0.1

LET E DENOTE THE EVENT $\{C, D, HD\}$ ("CREDIT OR BETTER").

$P(E)=0.2+0.15+0.1=0.45$.

THIS IS VALID BECAUSE EVENTS $\{C\}$, $\{D\}$ AND $\{HD\}$ ARE DISJOINT (NON-OVERLAPPING).

IF ALL OUTCOMES ARE EQUALLY LIKELY, THEN

$$P(A)=\frac{|A|}{|S|},$$ WHERE $|X|$ IS THE NUMBER OF OUTCOMES IN SET X.

Ex: A COIN IS TOSSED TWICE; THE SEQUENCE OF HEADS AND TAILS IS RECORDED. $S=\{HH, HT, TH, TT\}$. LET $E=\{HH, TT\}$ DENOTE THE EVENT "SAME RESULT FOR BOTH TOSSES". SINCE ALL 4 OUTCOMES HAVE EQUAL PROBABILITY,

$$P(E)=\frac{|E|}{|S|}=\frac{2}{4}=\frac{1}{2}.$$

EX: 2 FAIR DICE ARE ROLLED. WHAT IS THE PROBABILITY THAT THE SUM OF FACES IS 4?

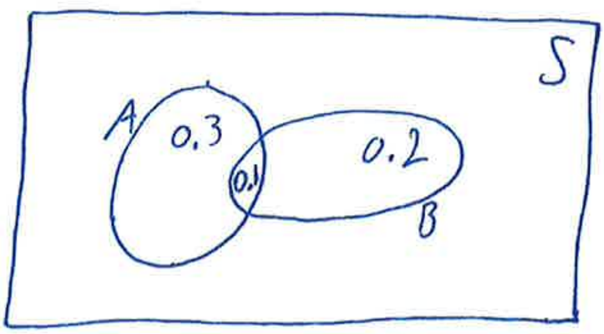
	1	2	3	4	5	6	TOTAL
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
TOTAL	1/6	1/6	1/6	1/6	1/6	1/6	1

ALL 36 EVENTS HAVE EQUAL PROBABILITY, AND 3 OF THEM MEET OUR NEEDS. LET $B = \{(1,3),(2,2),(3,1)\}$. THEN

$$P(B) = \frac{|B|}{|S|} = \frac{3}{36} = \frac{1}{12}.$$

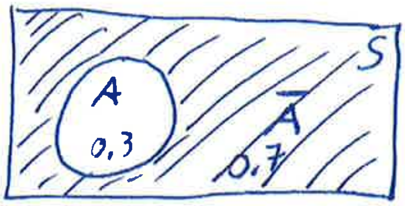
VENN DIAGRAMS

A VENN DIAGRAM REPRESENTS THE SAMPLE SPACE AND ALL EVENTS. PROBABILITIES ARE REPRESENTED AS AREAS.

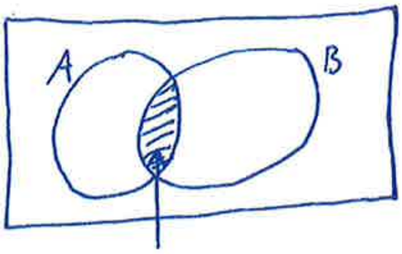


COMPLEMENT: THE COMPLEMENT OF A, DENOTED BY A^c OR $\bar{A}$, IS THE SET OF ALL OUTCOMES NOT IN A.

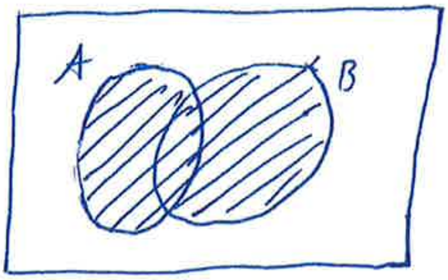
$$P(\bar{A}) = 1 - P(A).$$



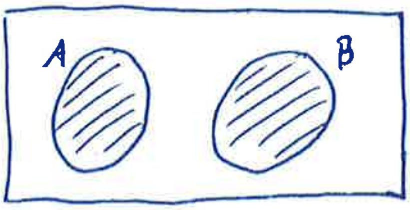
INTERSECTION: THE INTERSECTION $A \cap B$ IS THE EVENT THAT BOTH A AND B OCCUR.



UNION: THE UNION $A \cup B$ IS THE EVENT THAT A OR B (OR BOTH) OCCURS.



TWO EVENTS A AND B ARE **DISJOINT** (CANNOT OCCUR SIMULTANEOUSLY) IF $A \cap B = \emptyset$.



FOR DISJOINT EVENTS A AND B,

$$P(A \cup B) = P(A) + P(B);$$

$$P(A \cap B) = 0.$$

CONDITIONAL PROBABILITY

THE CONDITIONAL PROBABILITY OF EVENT A GIVEN THAT EVENT B HAS OCCURRED IS DENOTED BY $P(A|B)$. IN GENERAL (DISJOINT OR NOT), $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$. THAT IS, FOR A AND B BOTH TO HAPPEN, ONE EVENT HAPPENS, AND THEN GIVEN THAT, THE OTHER ONE HAPPENS. THIS GIVES US A FORMULA FOR CONDITIONAL PROBABILITY.

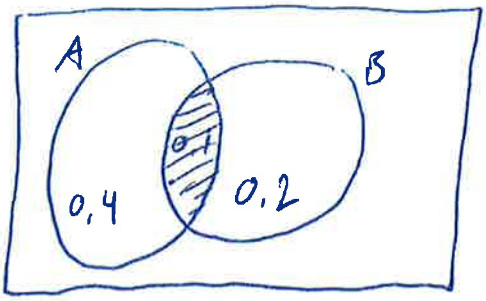
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

EX: WHAT IS THE PROBABILITY OF A ("DOUBLES") GIVEN B (SUM OF 2 DICE IS 4)?

$$P(A|B) = \frac{1/36}{3/36} = \frac{1}{3}.$$

NOTICE THAT THIS IS DIFFERENT FROM THE UNCONDITIONAL PROBABILITY $P(A) = \frac{6}{36} = \frac{1}{6}$.

IF YOU USE VENN DIAGRAMS TO CALCULATE $P(A|B)$, $\overline{B}$ IS DISCARDED AND B BECOMES THE NEW SAMPLE SPACE.



$$P(A|B) = \frac{0.1}{0.1+0.2} = \frac{1}{3}$$

PROBABILITY RULES

1. $P(S) = 1$, $P(\emptyset) = 0$.
2. $P(E) \geq 0$ FOR ANY EVENT $E \subseteq S$.
3. IF $A \cap B = \emptyset$, THEN $P(A \cup B) = P(A) + P(B)$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5. $P(\bar{A}) = 1 - P(A)$.
6. $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.

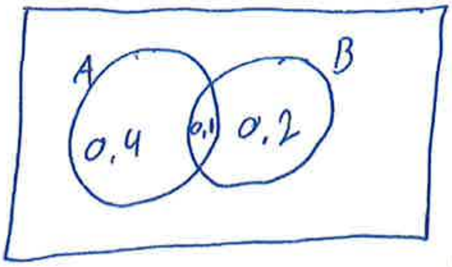
A TWO-WAY TABLE PRESENTS PROBABILITIES OF ALL POSSIBLE INTERSECTIONS.

	B	$\bar{B}$	TOTAL
A	$P(A \cap B)$	$P(A \cap \bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A} \cap B)$	$P(\bar{A} \cap \bar{B})$	$P(\bar{A})$
TOTAL	$P(B)$	$P(\bar{B})$	1

EX: USING THE PREVIOUS VENN DIAGRAM:

	B	$\bar{B}$	TOTAL
A	0,1	0,4	0,5
$\bar{A}$	0,2	0,3	0,5
TOTAL	0,3	0,7	1

TO FIND CONDITIONAL PROBABILITIES USING A TWO-WAY TABLE, DIVIDE THE INTERSECTION VALUE BY THE ROW OR COLUMN TOTAL.



	B	$\overline{B}$	TOTAL
A	0,1	0,4	0,5
$\overline{A}$	0,2	0,3	0,5
TOTAL	0,3	0,7	1

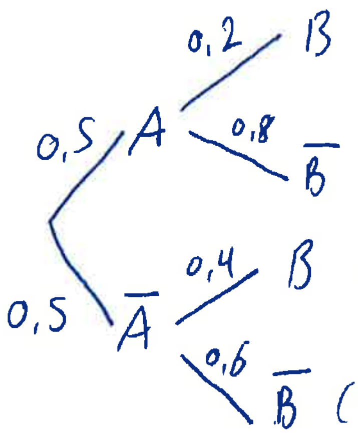
$$P(B|A) = \frac{0{,}1}{0{,}5} = \frac{1}{5} \; ; \; P(A|B) = \frac{0{,}1}{0{,}3} = \frac{1}{3}.$$

TREE DIAGRAMS

CONDITIONAL PROBABILITIES CORRESPOND TO SECOND-LEVEL (OR HIGHER) BRANCHES IN A TREE DIAGRAM.

- MULTIPLY PROBABILITIES OF ALL BRANCHES ALONG A PATH TO FIND ITS PROBABILITY.
- ADD PROBABILITIES OF ALL PATHS LEADING TO AN EVENT TO FIND ITS PROBABILITY.

EX: BASED ON THE PREVIOUS TWO-WAY TABLE:



$(0{,}5)(0{,}2) = 0{,}1 = P(A \cap B)$

$(0{,}5)(0{,}8) = 0{,}4 = P(A \cap \overline{B})$

$(0{,}5)(0{,}4) = 0{,}2 = P(\overline{A} \cap B)$

$(0{,}5)(0{,}6) = 0{,}3 = P(\overline{A} \cap \overline{B})$

EX: SAME TABLE, BRANCH FROM B FIRST.

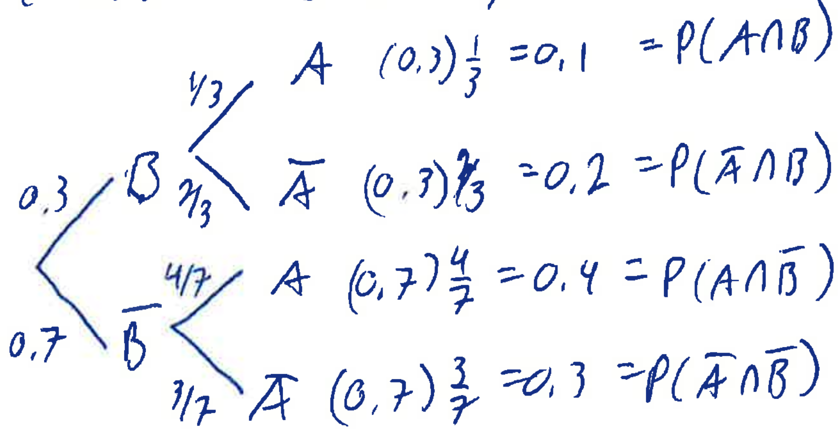
$P(B) = 0.3$, $P(\bar{B}) = 0.7$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = \frac{1}{3}.$$

$$P(\bar{A}|B) = 1 - P(A|B) = \frac{2}{3}.$$

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{0.4}{0.7} = \frac{4}{7}$$

$$P(\bar{A}|\bar{B}) = 1 - P(A|\bar{B}) = \frac{3}{7}$$



LAW OF TOTAL PROBABILITY

$P(A)$ CAN BE FOUND BY DECOMPOSING A INTO DISJOINT PIECES, THEN USING THE SUM AND PRODUCT RULES:

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$
$$= P(B)P(A|B) + P(\bar{B})P(A|\bar{B}).$$

EX: FOR THE PREVIOUS EXAMPLE,

$$P(A) = (0.3)\frac{1}{3} + (0.7)\frac{4}{7} = 0.5$$

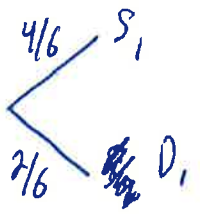
EX: 2 ITEMS ARE RANDOMLY SELECTED WITHOUT REPLACEMENT FROM A BATCH OF 6. THE BATCH CONTAINS 2 DEFECTIVE ITEMS.

LET S_i DENOTE THE EVENT THAT ITEM i INSPECTED IS SATISFACTORY.

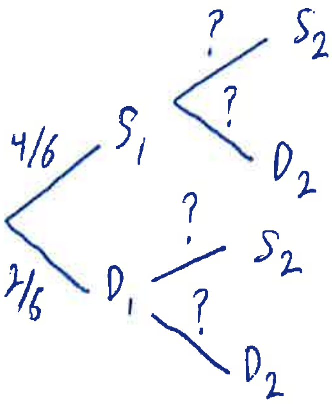
LET D_i DENOTE THE EVENT THAT ITEM i INSPECTED IS DEFECTIVE.

WHAT IS THE PROBABILITY THAT AT LEAST ONE DEFECTIVE ITEM IS FOUND?

STEP 1: INSPECT THE FIRST ITEM



STEP 2: GIVEN STEP 1, INSPECT THE SECOND ITEM



INTERSECTIONS: $P(S_1 \cap S_2) = \frac{4}{6} \cdot \frac{3}{5} = \frac{2}{5}$; $P(S_1 \cap D_2) = \frac{4}{6} \cdot \frac{2}{5} = \frac{4}{15}$;

$P(D_1 \cap S_2) = \frac{2}{6} \cdot \frac{4}{5} = \frac{4}{15}$; $P(D_1 \cap D_2) = \frac{2}{6} \cdot \frac{1}{5} = \frac{1}{15}$.

$P(\text{AT LEAST ONE DEFECTIVE}) = P(S_1 \cap D_2) + P(D_1 \cap S_2) + P(D_1 \cap D_2) = \frac{3}{5}$

OR $1 - P(\text{NO DEFECTIVES}) = 1 - P(S_1 \cap S_2) = \frac{3}{5}$.

INDEPENDENCE

IF THE PROBABILITY THAT A OCCURS IS NOT AFFECTED BY WHETHER OR NOT B OCCURS, i.e. $P(A|B) = P(A)$, WE SAY THAT A AND B ARE <u>INDEPENDENT</u>. WE HAVE THAT $P(A|B) = P(A \cap B)/P(B)$, SO A AND B ARE INDEPENDENT IFF

$$P(A \cap B) = P(A)P(B).$$

<u>EXAMPLES</u>:

- SUCCESSIVE COIN TOSSES ARE NOT AFFECTED BY PREVIOUS RESULTS, SO THE RESULTS OF DIFFERENT TOSSES ARE INDEPENDENT.
- THE EVENTS "DRUG PRESENT" AND "POSITIVE TEST RESULT" ARE NOT INDEPENDENT, AS A DRUG TEST IS MUCH MORE LIKELY TO BE POSITIVE IF THE DRUG IS PRESENT.

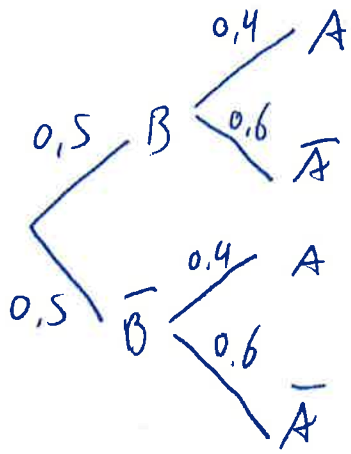
<u>EX</u>: EVENTS A AND B ARE INDEPENDENT, $P(A) = 0.4$, $P(B) = 0.5$. CONSTRUCT A TWO-WAY TABLE AND A TREE DIAGRAM.

START WITH WHAT YOU KNOW, AND USE $P(A \cap B) = P(A)P(B)$.

	B	$\bar{B}$	TOTAL
A	0.2		0.4
$\bar{A}$			
TOTAL	0.5		1

FIND THE REMAINING ENTRIES BY SUBTRACTION

	B	$\bar{B}$	TOTAL
A	0,2	0,2	0,4
$\bar{A}$	0,3	0,3	0,6
TOTAL	0,5	0,5	1



$(0{,}5)(0.4) = 0{,}2 = P(A \cap B)$

$(0{,}5)(0{,}6) = 0{,}3 = P(\bar{A} \cap B)$

$(0{,}5)(0.4) = 0{,}2 = P(A \cap \bar{B})$

$(0{,}5)(0{,}4) = 0{,}3 = P(\bar{A} \cap \bar{B})$

EX: IF $P(A \cap \bar{B}) = 0{,}36$, $P(\bar{A} \cap B) = 0{,}24$, $P(A|B) = 0{,}5$, THEN A AND B ARE

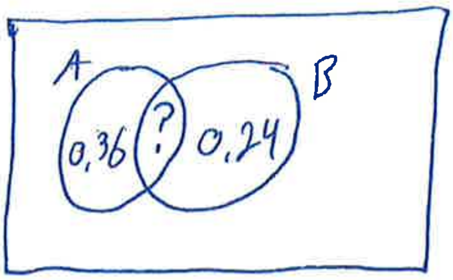
a) DISJOINT AND INDEPENDENT

b) DISJOINT AND NOT INDEPENDENT

c) INDEPENDENT AND NOT DISJOINT

d) NOT INDEPENDENT AND NOT DISJOINT

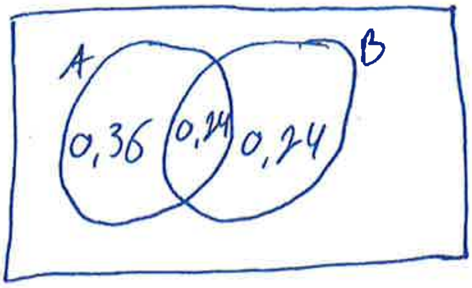
A:



RECALL THAT DISJOINT MEANS $P(A \cap B) = 0$.

$$P(A|B) = 0,5 = \frac{P(A\cap B)}{P(B)} \Rightarrow P(A\cap B) = 0.24 \neq 0.$$

$\therefore$ A AND B ARE NOT DISJOINT.



$$P(A) = 0,36 + 0.24 = 0.6 ; P(B) = 0.24 + 0.24 = 0.48$$

RECALL THAT INDEPENDENCE MEANS $P(A\cap B) = P(A)P(B)$.

$$P(A)P(B) = (0.6)(0.48) = 0.288 \neq P(A\cap B).$$

$\therefore$ A AND B ARE NOT INDEPENDENT.

BAYES' RULE: FOR EVENTS A AND B, BAYES' RULE PROVIDES A WAY TO REVERSE THE ORDER OF CONDITIONAL PROBABILITIES:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}.$$

THIS COMES DIRECTLY FROM THE DEFINITION OF CONDITIONAL PROBABILITY, THE PRODUCT RULE (NUMERATOR) AND THE LAW OF TOTAL PROBABILITY (DENOMINATOR).

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

IN TERMS OF A TREE, THE NUMERATOR DEFINES ONE PATH, THE DENOMINATOR IS THE SUM OF PATHS THAT LEAD TO A.

EX: A DRUG TEST HAS 0.96 CHANCE OF POSITIVE RESULT IF THE DRUG IS PRESENT, 0.93 CHANCE OF NEGATIVE RESULT IF THE DRUG IS NOT PRESENT. THE UNCONDITIONAL PROBABILITY OF THE DRUG BEING PRESENT IS 0.007. GIVEN A POSITIVE RESULT, WHAT IS THE PROBABILITY THAT THE DRUG IS PRESENT?

LET A = "POSITIVE TEST RESULT", B = "DRUG IS PRESENT".

$P(A|B) = 0.96$; $P(\bar{A}|\bar{B}) = 0.93$; $P(B) = 0.007$.

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

$$= \frac{(0.96)(0.007)}{(0.96)(0.007) + (1-0.93)(1-0.007)}$$

$$= 0.08815.$$

SO A POSITIVE TEST RESULT IS 91% LIKELY TO BE FALSE!!

BINOMIAL SCENARIO

- FIXED NUMBER OF INDEPENDENT TRIALS.
- 2 POSSIBLE OUTCOMES, "SUCCESS" AND "FAILURE".
- CONSTANT PROBABILITY OF SUCCESS FOR EACH TRIAL.
- THE QUANTITY OF INTEREST IS THE TOTAL NUMBER OF SUCCESSES.

NOTATION

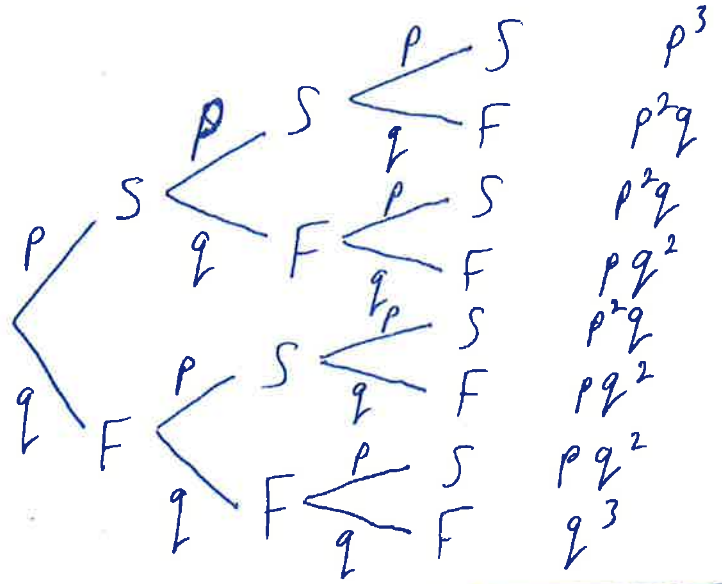
n = NUMBER OF INDEPENDENT TRIALS.

p = PROBABILITY OF SUCCESS FOR A SINGLE TRIAL, $0 < p < 1$

$q = 1 - p$ = PROBABILITY OF FAILURE.

x = NUMBER OF SUCCESSES.

FOR SMALL n, A TREE DIAGRAM CAN BE USED TO WORK OUT PROBABILITIES:



x	0	1	2	3	TOTAL
PROB.	q^3	$3pq^2$	$3p^2q$	p^3	1

FOR LARGER n, USE COMBINATORICS.

FACTORIALS: RECALL THAT THERE ARE $n!$ WAYS OF ARRANGING n OBJECTS, AND BY DEFINITION $0! = 1$. THERE IS AN $n!$ BUTTON ON MOST CALCULATORS.

R CODE:

```
factorial(n)
```

RECALL THE BINOMIAL COEFFICIENT, THE NUMBER OF WAYS TO SELECT k OBJECTS OUT OF n (ORDER NOT IMPORTANT) IS

$$\binom{n}{k} = C^n_k = \frac{n!}{k!(n-k)!}$$

EX: $\binom{3}{2} = \dfrac{3!}{2!(3-2)!} = 3$, SO THERE ARE 3 DIFFERENT WAYS OF CHOOSING 2 ITEMS FROM A SET OF 3 ITEMS.

R CODE:

```
choose(3,2)
```

AN ALTERNATIVE INTERPRETATION IS THAT THERE ARE $\binom{n}{x}$ WAYS OF ARRANGING n OBJECTS, x OF ONE TYPE (SUCCESS) AND $(n-x)$ OF ANOTHER TYPE (FAILURE): $\binom{3}{2} \to SSF, SFS, FSS$.

SINCE THE BINOMIAL SCENARIO EVENTS ARE INDEPENDENT, THE PROBABILITY OF x SUCCESSES AND $(n-x)$ FAILURES IN n TRIALS (SINGLE PATH) IS

$$\underbrace{p \cdot p \cdot \ldots \cdot p}_{x \text{ TIMES}} \; \underbrace{q \cdot q \cdot \ldots \cdot q}_{(n-x) \text{ TIMES}} = p^x q^{n-x}.$$

THE NUMBER OF SUCH PATHS IS $C^n_x = \binom{n}{x}$. SO THE PROBABILITY OF x SUCCESSES IS

$$\binom{n}{x} p^x q^{n-x}.$$

NOTE THAT THE SUM OF ALL BINOMIAL PROBABILITIES IS 1, AS IT MUST BE. WE SEE THIS BY THE BINOMIAL EXPANSION THEOREM.

$$\binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \dots + \binom{n}{n}p^n$$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (q+p)^n = (1-p+p)^n = 1.$$

EX: THE PROBABILITY THAT AN EMAIL DELIVERED TO A CERTAIN ACCOUNT IS JUNK IS 0.25, INDEPENDENTLY OF ALL OTHER MESSAGES. WHAT IS THE PROBABILITY THAT EXACTLY 5 OUT OF THE 20 MOST RECENT MESSAGES ARE JUNK?

A: $n=20$ IS FAR TOO LARGE FOR A TREE DIAGRAM, SO USE THE BINOMIAL PROBABILITY FORMULA WITH $n=20, x=5, p=0.25$.

$$P(5) = \binom{20}{5} 0.25^5 \, 0.75^{20-5} = 0.2023.$$

R CODE:

```
dbinom(5,20,0.25)
```

RANDOM VARIABLE

- A RANDOM VARIABLE IS A NUMERICAL MEASUREMENT OF THE OUTCOME OF A RANDOM PHENOMENON.
- AN UPPER-CASE LETTER, SUCH AS X, REFERS TO A RANDOM VARIABLE, WHICH CANNOT BE PREDICTED WITH CERTAINTY.
- A LOWER-CASE LETTER, SUCH AS x, REFERS TO A PARTICULAR VALUE OF THE VARIABLE.

DISCRETE PROBABILITY FUNCTION

A DISCRETE RANDOM VARIABLE HAS VALUES RESTRICTED TO SEPARATE POINTS.

THE PROBABILITY FUNCTION OF A DISCRETE RV IS DEFINED BY

$$f(x) = P[X = x].$$

SOMETIMES A SUBSCRIPT IS USED TO DISTINGUISH BETWEEN VARIABLES:

$$f_Y(5) = P[Y = 5].$$

A PROBABILITY FUNCTION MUST SATISFY

$$f(x) \geq 0 \text{ FOR ALL } x, \text{ AND } \quad \sum_x f(x) = 1.$$

THE FUNCTION MAY BE SPECIFIED BY TABLE OR BY FORMULA:

x	0	1	2	TOTAL
$f(x)$	0.3	0.55	0.15	1

$$g(x) = \binom{2}{x} p^x (1-p)^{2-x}, \quad x = 0, 1, 2.$$

BINOMIAL DISTRIBUTION

LET X BE THE NUMBER OF SUCCESSES IN n INDEPENDENT TRIALS, WITH CONSTANT PROBABILITY p OF SUCCESS. THEN X HAS A BINOMIAL PROBABILITY FUNCTION

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n, \quad q = 1 - p.$$

R CODE:

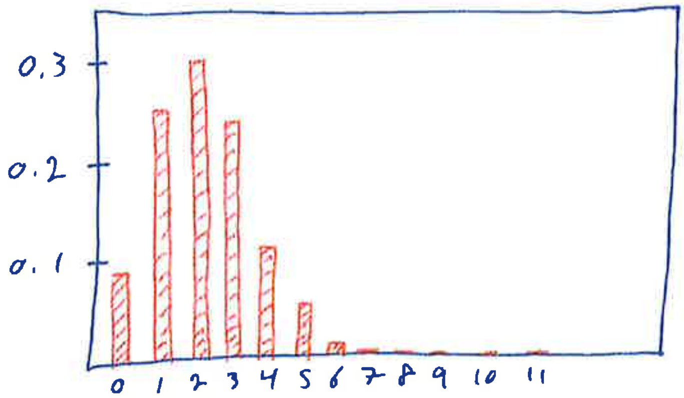
```
dbinom(x,n,p)
```

EX: A MULTIPLE CHOICE QUIZ HAS 11 QUESTIONS WITH 5 POSSIBLE ANSWERS EACH. WHAT IS THE PROBABILITY THAT A STUDENT WHO GUESSES AT EVERY QUESTION GETS A SCORE OF 4 OUT OF 11?

A: $n = 11, x = 4, p = 0.2$

$$f(4) = P(X=4) = \binom{11}{4} 0.2^4 0.8^{11-4} = 0.1107.$$

ONE CAN CALCULATE EACH POSSIBLE OUTCOME AND OBTAIN THE BINOMIAL PROBABILITY FUNCTION GRAPH FOR THE ABOVE EXAMPLE:



(SO DON'T GUESS; YOU'LL MOST LIKELY GET 2/11. :))

CUMULATIVE PROBABILITIES

R CODE:

```
pbinom(3,11,0.2)
```

THIS GIVES $P(X \leq 3) = f(0) + f(1) + f(2) + f(3) = 0.8389$ FOR $n = 11$ AND $p = 0.2$. BY HAND, THIS IS

$$\binom{11}{0} 0.2^0 0.8^{11} + \binom{11}{1} 0.2^1 0.8^{10} + \binom{11}{2} 0.2^2 0.8^9 + \binom{11}{3} 0.2^3 0.8^8.$$

EX: IN THE PREVIOUS EXAMPLE, WHAT IS THE PROBABILITY THAT THE STUDENT GETS AT LEAST 4 OUT OF 11?

A: THE HARD WAY: $P(X \geq 4) = P(X=4) + P(X=5) + \ldots + P(X=11)$.

THE SMARTER WAY: $P(X \geq 4) = 1 - P(X < 4) = 1 - 0.8389 = 0.1611$.

THE CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF A DISCRETE RANDOM VARIABLE X, DENOTED BY $F(x)$ OR $F_X(x)$, IS DEFINED BY

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k).$$

TO AVOID SUMS WITH MANY TERMS, WE USE DIFFERENCES OF CDFs:

$$\boxed{P(a < X \leq b) = F(b) - F(a).}$$

* BE CAREFUL WITH $<$ AND $\leq$ FOR DISCRETE VARIABLES. eg:

$$P(20 \leq X \leq 25) = F(25) - F(19).$$

$F(x)$ IS FOUND BY SUMMING VALUES OF $f(k)$. TO FIND f FROM F, WE USE DIFFERENCES:

$$\begin{aligned} f(x) &= P(X = x) \\ &= P(X \leq x) - P(X < x) \\ &= F(x) - F(x-1). \end{aligned}$$

EX:

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1
$F(x)$	0.4	0.7	0.9	1

$F(2) = f(0) + f(1) + f(2) = 0.4 + 0.3 + 0.2 = 0.9$

$f(2) = F(2) - F(1) = 0.9 - 0.7 = 0.2$

$P(0 < X \leq 2) = f(1) + f(2) = F(2) - F(0)$

RELATIVE FREQUENCY AND PROBABILITY

CONSIDER n OBSERVATIONS OF A DISCRETE RV X, ON AVERAGE, WE EXPECT THE OBSERVED RELATIVE FREQUENCY $\frac{n_x}{n}$ OF A FIXED VALUE x TO BE EQUAL TO THE PROBABILITY FUNCTION $f(x) = P(X = x)$.

RECALL THE FORMULA FOR THE MEAN OF A SAMPLE:

$\bar{x} = \sum_x x \cdot \frac{n_x}{n}$. THIS LEADS TO THE FOLLOWING DEFINITION.

DEF: THE EXPECTED VALUE $E(X)$ OF A DISCRETE RV X IS DEFINED BY

$$E(X) = \sum_x x f(x).$$

$E(X)$ IS A WEIGHTED AVERAGE; GREATER WEIGHT IS ASSIGNED TO MORE LIKELY VALUES OF x.

EX: FIND THE EXPECTED VALUE OF $f(x) = 0.1(4-x)$, $x = 0, 1, 2, 3$.

A:

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1

$$E(X) = \sum_{x=0}^{3} x f(x) = 0(0.4) + 1(0.3) + 2(0.2) + 3(0.1) = \boxed{1}$$

SIMILARLY, THE EXPECTED VALUE OF $g(X)$ FOR SOME FUNCTION g IS

$E[g(X)] = \sum_x g(x) f(x)$. FOR EXAMPLE: $f(x) = 0.1(4-x)$, $x = 0, 1, 2, 3$:

$$E(X^2) = \sum_{x=0}^{3} x^2 f(x) = 2.$$

PROPERTIES OF $E(X)$

- $E(a) = a$ FOR ANY CONSTANT a.
- FOR A LINEAR TRANSFORMATION $a + bX$,

$$E(a+bX) = a + bE(X).$$

NOTE: FOR A NONLINEAR TRANSFORMATION $g(X)$, $E[g(X)]$ USUALLY DIFFERS FROM $g(E(X))$, AS IN THE LAST EXAMPLE $E(X^2) \neq [E(X)]^2$.

EX: FOR THE PREVIOUS EXAMPLE, FIND $E(7-2X)$.

A: BY DIRECT CALCULATION,

$$E(7-2X) = (7-0)0.4 + (7-2)0.3 + (7-4)0.2 + (7-6)0.1 = 5.$$

THE SMARTER WAY:

$$E(7-2X) = 7 - 2E(X) = 5.$$

MEAN AND VARIANCE

THE MEAN OF A DISCRETE RV X IS DEFINED AS

$$\mu = \mu_X = E(X).$$

FOR A LARGE SAMPLE OF OBSERVATIONS, WE EXPECT THE SAMPLE MEAN $\bar{X}$ TO BE CLOSE TO THE THEORETICAL MEAN μ.

RECALL THE SAMPLE VARIANCE IS THE AVERAGE OF SQUARED DISTANCES FROM THE SAMPLE MEAN:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

THE VARIANCE OF A RV X IS THE EXPECTED SQUARED DISTANCE FROM μ:

$$\sigma^2 = Var(X) = E[(X-\mu)^2].$$

A USEFUL ALTERNATIVE REPRESENTATION IS

$$\sigma^2 = E(X^2) - \mu^2.$$

EXERCISE: USE PROPERTIES OF $E(X)$ TO PROVE THAT

$$E[(X-\mu)^2] = E(X^2) - \mu^2.$$

THE STANDARD DEVIATION OF X IS THE POSITIVE SQUARE ROOT OF THE VARIANCE: $\sigma = \sqrt{Var(X)}$.

Ex: FIND THE VARIANCE FOR THE PREVIOUS EXAMPLE.

A: RECALL THAT $\mu = 1$ AND $E(X^2) = 2$, SO $\sigma^2 = E(X^2) - \mu^2 = 1$.

OR THE LONGER WAY:

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-1)^2 f(x)$$

$$= (0-1)^2 0.4 + (1-1)^2 0.3 + (2-1)^2 0.2 + (3-1)^2 0.1 = 1.$$

PROPERTIES OF VARIANCE

- $Var(X) \geq 0$, AND $Var(X) = 0 \Leftrightarrow X$ IS CONSTANT.
- $Var(X+a) = Var(X)$ FOR ANY CONSTANT a.
- $Var(bX) = b^2 Var(X)$ FOR ANY CONSTANT b.
- $\sigma_{a+bX} = |b| \sigma_X$.

IF A RV HAS LARGE VARIANCE, IT MEANS THAT OBSERVATIONS ARE EXPECTED TO VARY GREATLY.

EX: FOR THE EXAMPLE $f(x) = 0.1(4-x)$, WE FOUND $\sigma^2 = 1$.

- $Var\left(\frac{X}{2}\right) = \left(\frac{1}{2}\right)^2 Var(X) = \frac{1}{4}$
- $Var(X+6) = Var(X) = 1$
- $Var(6-2X) = (-2)^2 Var(X) = 4$

FOR A BINOMIAL DISTRIBUTION WITH n TRIALS AND p PROBABILITY OF SUCCESS, WE FIND THAT

$$\mu = E(X) = np$$

$$\sigma^2 = Var(X) = np(1-p)$$

$$\sigma = \sqrt{Var(X)} = \sqrt{np(1-p)}$$

EX: FIND THE MEAN AND STANDARD DEVIATION OF THE NUMBER X OF HEADS OBTAINED IN 100 TOSSES OF A COIN.

A: BINOMIAL DISTRIBUTION, $n = 100$, $p = 0.5$.

$$\mu = np = 50$$

$$\sigma = \sqrt{np(1-p)} = 5.$$

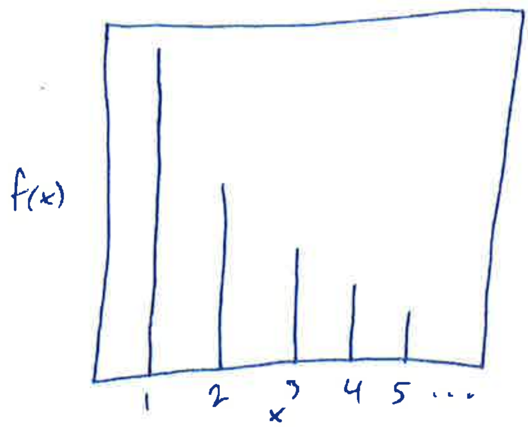
SO ALTHOUGH X WILL BE ABOUT 50 ON AVERAGE, IT WOULD NOT BE UNUSUAL TO OBSERVE VALUES BETWEEN 45 AND 55 ($\mu - \sigma$ AND $\mu + \sigma$).

GEOMETRIC DISTRIBUTION

THE GEOMETRIC DISTRIBUTION ARISES WHEN COUNTING THE NUMBER X OF TRIALS UNTIL THE FIRST SUCCESS. THE FINAL (SUCCESSFUL) TRIAL IS ALSO COUNTED, SO $X \in \{1, 2, \ldots, n\}$. THE DISCRETE PROBABILITY FUNCTION IS

$$f(x) = q^{x-1} p \,, \quad x = 1, 2, 3, \ldots, n.$$

THE SHAPE OF THE GRAPH IS STRONGLY SKEWED TO THE RIGHT.



TO FIND CDF OF $f(x) = q^{x-1} p$, WE GET

$$F(x) = p + qp + q^2 p + \ldots + q^{x-1} p$$

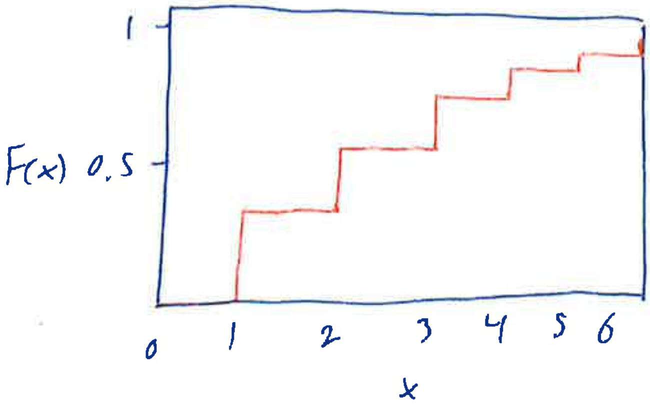
$$= p \, \frac{1-q^x}{1-q} = \boxed{1 - q^x} \,, \quad x = 1, 2, 3, \ldots, n.$$

IF CONVENIENT, YOU CAN ALSO USE $F(x) = P(X \leq x) = 1 - P(X > x)$, WHERE $P(X > x)$ IS THE PROBABILITY OF NO SUCCESSES IN THE FIRST x TRIALS.

WE CAN CHECK THE CDF RESULT BY USING THE FORMULA $f(x) = F(x) - F(x-1)$.

$$f(x) = F(x) - F(x-1)$$
$$= 1 - q^{x} - (1 - q^{x-1})$$
$$= q^{x-1} - q^{x}$$
$$= q^{x-1}(1-q) = q^{x-1}p.$$

THE CDF (GEOMETRIC OR OTHERWISE) APPROACHES 1 AS x INCREASES.



THE MEAN μ OF A GEOMETRIC RV X INVOLVES A SUMMATION TRICK THAT WE WILL NOT PROVE HERE.

$$E(X) = \sum_{x=1}^{\infty} x q^{x-1} p = p(1 + 2q + 3q^2 + \dots)$$
$$= p \frac{d}{dq}(1 + q + q^2 + q^3 + \dots)$$
$$= p \frac{d}{dq}\left(\frac{1}{1-q}\right) = \frac{p}{(1-q)^2}$$
$$= \boxed{\frac{1}{p}}$$

EX: A STUDENT GUESSES AT QUESTIONS WITH 5 MULTIPLE CHOICE ANSWERS EACH. LET X BE THE NUMBER OF THE FIRST QUESTION ANSWERED CORRECTLY. WHAT IS $E(X)$, AND WHAT IS THE PROBABILITY THAT $X > E(X)$?

A: X IS GEOMETRIC WITH $p = 0.2$, SO

$$E(X) = \frac{1}{p} = 5.$$

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F(5)$$

$$= 1 - \left[1 - (1-0.2)^5\right] = 0.3277.$$

SO ON AVERAGE, IT WILL TAKE THE STUDENT 5 QUESTIONS TO GET 1 RIGHT, AND THERE'S A 33% CHANCE THAT IT WILL TAKE MORE THAN 5 QUESTIONS.

POISSON DISTRIBUTION

POISSON IS A DISCRETE DISTRIBUTION THAT APPLIES WHEN WE COUNT THE NUMBER OF POINTS IN A GIVEN TIME/AREA/DISTANCE/VOLUME. FOR INSTANCE, THE NUMBER OF CARS THAT GO THROUGH AN INTERSECTION IN 10 MINUTES, OR THE NUMBER OF RUST SPOTS IN A $10\,m^2$ AREA.

LET λ BE THE AVERAGE RATE OF OCCURRENCES PER UNIT TIME (OR DISTANCE, AREA, VOLUME). THEN THE EXPECTED NUMBER μ OF OCCURRENCES IN AN INTERVAL OF LENGTH t IS

$$\mu = \lambda t.$$

FOR THE POISSON PROCESS, WE ASSUME

1) IF $N_{\Delta t}$ IS THE NUMBER OF OCCURRENCES IN A VERY SHORT TIME Δt, THEN

$$P(N_{\Delta t}=1) \simeq \lambda \Delta t,$$

$$P(N_{\Delta t}>1) \simeq 0.$$

2) COUNTS IN TIME INTERVALS THAT DON'T OVERLAP ARE INDEPENDENT.

TWO RVs X AND Y ARE <u>INDEPENDENT</u> IF

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b) \text{ FOR ALL } a, b.$$

i.e. PROBABILITIES INVOLVING ONE VARIABLE ARE UNAFFECTED BY INFORMATION ABOUT THE OTHER VARIABLE.

IF X AND Y ARE INDEPENDENT, THEN $E(XY)=E(X)E(Y)$.

THE PROCESS IS THE FOLLOWING.

- THE INTERVAL OF INTEREST $[0,t]$ IS SUBDIVIDED INTO n SUBINTERVALS OF LENGTH $\Delta t = t/n$.
- N_t IS APPROXIMATED BY COUNTING HOW MANY SUBINTERVALS CONTAIN AT LEAST ONE POINT.
- BY INDEPENDENCE OF SUBINTERVALS, N_t HAS AN APPROXIMATELY BINOMIAL DISTRIBUTION WITH $p = \lambda t/n$.
- THIS GIVES $\mu = np = \lambda t$, AND $\sigma^2 = np(1-p) = \lambda t(1-\lambda t/n)$.
- AS $n \to \infty$, WE HAVE $\sigma^2 \to \lambda t$. SO $\sigma^2 = \mu$ AT THE LIMIT.
- LETTING $n \to \infty$, WE HAVE $P(N_t = x) = \binom{n}{x}\left(\frac{\lambda t}{n}\right)^x \left(1-\frac{\lambda t}{n}\right)^{n-x} \to \frac{(\lambda t)^x}{x!} e^{-\lambda t}$.
- WITH $\mu = \lambda t$, WE OBTAIN THE <u>POISSON PROBABILITY FUNCTION</u>:

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \ldots$$

THERE ARE INFINITELY MANY NONZERO PROBABILITIES, BUT THEY STILL ADD UP TO 1:

$$\sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} = e^{-\mu}\left(1+\mu+\frac{\mu^2}{2}+\frac{\mu^3}{6}+\dots\right) = e^{-\mu}e^{\mu} = 1.$$

FOR A POISSON RV, $E(X) = Var(X) = \mu$.

EX: THE UOW SWITCHBOARD RECEIVES ON AVERAGE 0.6 CALLS PER MINUTE. FIND THE PROBABILITY THAT IN A 4-MINUTE INTERVAL THERE WILL BE (i) EXACTLY 3 CALLS, (ii) AT LEAST 3 CALLS.

A: THE RATE OF CALLS IS $\lambda = 0.6$ PER MINUTE, SO

$\mu = \lambda t = 0.6 \cdot 4 = 2.4$.

(i) $P(X=3) = \frac{2.4^3}{3!} e^{-2.4} \simeq 0.209$

R CODE:

```
dpois(3,2.4)
```

(ii) $P(X \geq 3) = 1 - P(X<3)$.

$P(X<3) = f(0)+f(1)+f(2)$

$$= \frac{2.4^0}{0!}e^{-2.4} + \frac{2.4^1}{1!}e^{-2.4} + \frac{2.4^2}{2!}e^{-2.4} = 0.5697$$

$\therefore P(X \geq 3) = 1 - 0.5697 = 0.4303$.

R CODE:

```
1-ppois(2,0.24)
```

EX: LET A = "AT LEAST 4 CALLS ARRIVE BETWEEN 10:00am AND 10:04am",

B = "EXACTLY 5 CALLS ARRIVE BETWEEN 10:00am AND 10:05am".

WHAT IS THE CONDITIONAL PROBABILITY OF A GIVEN B?

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

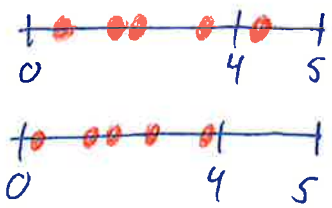
WE CAN FIND $P(B)$ DIRECTLY WITH $\mu = 0.6 \cdot 5 = 3$, $x = 5$.

$$P(B) = \frac{3^5}{5!}e^{-3}.$$

THERE ARE TWO POSSIBILITIES FOR $A\cap B$:

1) 4 CALLS IN $[0,4]$ AND 1 CALL IN $(4,5]$

2) 5 CALLS IN $[0,4]$ AND 0 CALLS IN $(4,5]$



SO WE NEED POISSON PROBABILITIES FOR EACH INTERVAL, WITH MEANS $0.6 \cdot 4 = 2.4$ FOR THE 4-MINUTE INTERVAL AND 0.6 FOR THE 1-MINUTE INTERVAL. BY INDEPENDENCE OF NON-OVERLAPPING COUNTS,

$$P(4 \text{ CALLS IN } [0,4] \text{ AND } 1 \text{ CALL IN } (4,5]) = \frac{2.4^4}{4!}e^{-2.4}\frac{0.6}{1!}e^{-0.6} = \frac{2.4^4 \cdot 0.6}{4!}e^{-3}$$

$$P(5 \text{ CALLS IN } [0,4] \text{ AND } 0 \text{ CALLS IN } (4,5]) = \frac{2.4^5}{5!}e^{-2.4}\frac{0.6^0}{0!}e^{-0.6} = \frac{2.4^5}{5!}e^{-3}$$

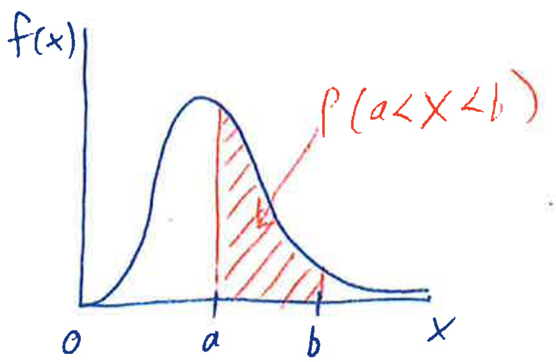
ADDING THESE TOGETHER, $P(A\cap B) = \dfrac{2.4^4 \cdot 0.6}{4!}e^{-3} + \dfrac{2.4^5}{5!}e^{-3}$

$$\therefore P(A|B) = \frac{2.4^4 \cdot 0.6/4! + 2.4^5/5!}{3^5/5!} = \boxed{0.7373}$$

CONTINUOUS RANDOM VARIABLES

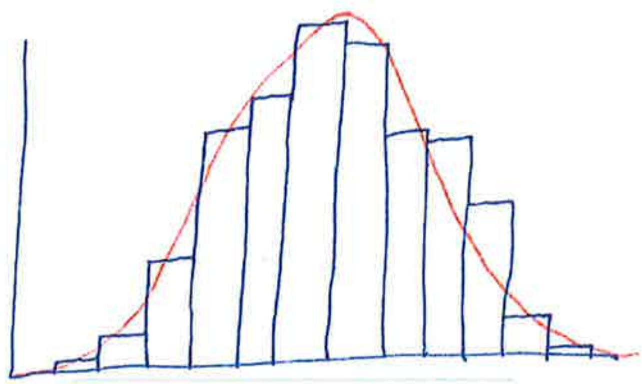
A RV IS CONTINUOUS IF THE SET OF ITS POSSIBLE VALUES IS ONE OR MORE INTERVALS. eg. THE SET OF POSSIBLE VALUES T = TIME IT TAKES FOR A PIZZA TO ARRIVE TO YOUR HOUSE IS (HOPEFULLY) $\{t : 15 \le t \le 45 \text{ min}\}$.

THE PROBABILITY DENSITY FUNCTION (PDF) $f(x)$ OF A CONTINUOUS RV X IS THE FUNCTION SUCH THAT $P(a < X < b)$ IS THE AREA $\int_a^b f(x)\,dx$ UNDER THE CURVE $y = f(x)$ BETWEEN $x = a$ AND $x = b$.



NOTE: THE AREA IS THE SAME FOR $P(a \le X < b)$, $P(a < X \le b)$, $P(a \le X \le b)$.

A PDF CAN BE VIEWED AS A HISTOGRAM AS SAMPLE SIZE TENDS TO ∞, AND CLASS WIDTH TENDS TO ZERO. THE TOTAL AREA UNDER THE WHOLE CURVE MUST BE 1.



PROPERTIES OF PDF

- FOR VALUES OF x THAT ARE NEVER OBSERVED, $f(x)=0$.
- $f(x) \geq 0$ FOR ALL X.
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

EX: SHOW THAT $f(x) = 3x^2$, $x \in (0,1)$ IS A VALID PDF. FIND $P(0.2<X<0.9)$.

i) SINCE THE DOMAIN IS ONLY $(0,1)$, WE MAY SET $f(x)=0$ FOR ALL $x \notin (0,1)$.

ii) $3x^2 \geq 0$ FOR ALL x.

iii) $\int_{-\infty}^{\infty} 3x^2\,dx$? NO

$$\int_{-\infty}^{\infty} f(x)\,dx = \underbrace{\int_{-\infty}^{0} f(x)\,dx}_{0} + \int_{0}^{1} f(x)\,dx + \underbrace{\int_{1}^{\infty} f(x)\,dx}_{0} = \int_{0}^{1} 3x^2\,dx = x^3\Big|_0^1 = 1.$$

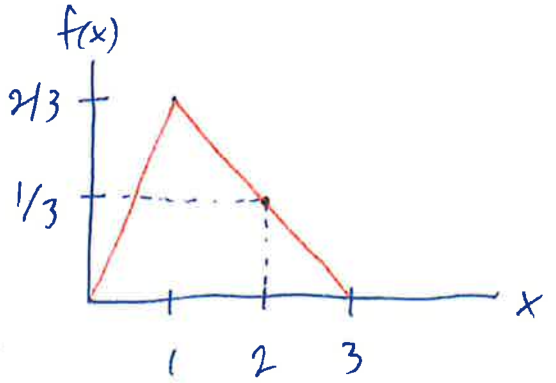
$$P(0.2<X<0.9) = \int_{0.2}^{0.9} 3x^2\,dx = 0.9^3 - 0.2^3 = 0.721.$$

THE CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF A CONTINUOUS RV X IS DEFINED BY

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

IF YOU HAVE CDFs AVAILABLE, YOU CAN SOMETIMES AVOID INTEGRATION BY USING $P(a<X<b) = F(b) - F(a)$.

EX: LET $F(x)$ BE THE CDF OF THE CONTINUOUS RV WHOSE PROBABILITY FUNCTION $f(x)$ IS GRAPHED BELOW. FIND $F(2)$ AND $P(1<X<2)$.



A: TOTAL AREA UNDER f IS 1, SO THE AREA TO THE LEFT OF 2 IS

$$1-(\text{RIGHT AREA}) = 1-\frac{1/3}{2} = \frac{5}{6} = F(2).$$

FOR $P(1<X<2)$, EITHER FIND THE AREA BETWEEN 1 AND 2:

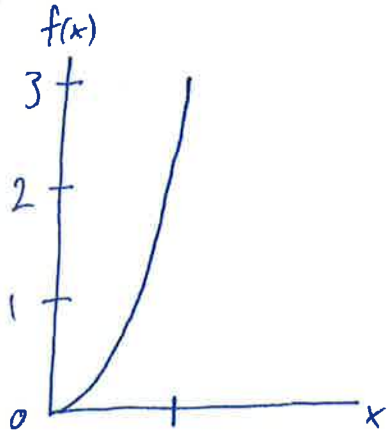
$$1\cdot\frac{2/3+1/3}{2} = \frac{1}{2}$$

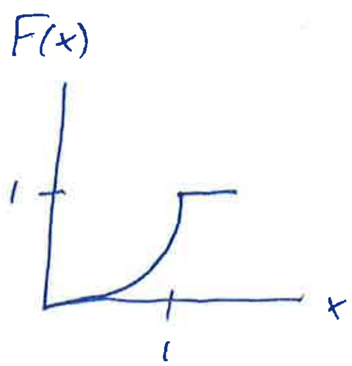
OR USE $F(2)-F(1) = \frac{5}{6}-\frac{1}{3} = \frac{1}{2}$.

PROPERTIES OF CDF

- AS $F(x)$ IS A PROBABILITY, IT MUST LIE BETWEEN 0 AND 1.
- AS x INCREASES, THE EVENT $\{X \le x\}$ INCLUDES MORE OUTCOMES, SO $F(x)$ IS AN INCREASING FUNCTION OF X.
- FOR CONTINUOUS X, F IS CONTINUOUS. (FOR DISCRETE X, F IS A STEP FUNCTION).

EX: FOR THE PDF $f(x)=3x^2$, $0<x<1$, F IS AS FOLLOWS.





WE OBTAIN F FROM f BY INTEGRATION, SO TO OBTAIN f FROM F, WE USE DIFFERENTIATION. BY THE FUNDAMENTAL THEOREM OF CALCULUS,

$$\frac{d}{dx}F(x) = \frac{d}{dx}\int_{-\infty}^{x} f(t)\,dt = f(x).$$

$$\text{i.e. } f(x) = \frac{d}{dx}F(x).$$

<u>Ex:</u> IF $F(x) = x^3, 0<x<1$, THEN

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}x^3 = 3x^2, \; 0<x<1.$$

<u>MEAN, VARIANCE AND STANDARD DEVIATION</u>

THE EXPECTED VALUE OF A FUNCTION OF A CONTINUOUS RV IS

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

THIS IS THE LONG-RUN AVERAGE VALUE OF $g(X)$ FOR A LARGE NUMBER OF EVALUATIONS. AS IN THE DISCRETE CASE, WE HAVE

- MEAN $\mu = E(X)$,
- VARIANCE $\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$,
- STANDARD DEVIATION $\sigma = \sqrt{Var(X)}$.

<u>Ex</u>: FOR $f(x) = 3x^2, 0<x<1$, FIND μ, $E(4X-2)$, $E\left(\frac{1}{X}\right)$, $\frac{1}{E(X)}$, $E(X^2)$, $Var(X)$ AND σ.

A: $\mu = E(X) = \int_0^1 x\, f(x)\,dx = \int_0^1 3x^3\,dx = \frac{3x^4}{4}\Big|_0^1 = \frac{3}{4}$

$E(4X-2) = 4E(X) - 2 = 1$

$E\left(\frac{1}{X}\right) = \int_0^1 \frac{1}{x}\cdot 3x^2\,dx = \frac{3x^2}{2}\Big|_0^1 = \frac{3}{2}$ (NOTE THAT $E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$.)

$E(X^2) = \int_0^1 x^2\cdot 3x^2\,dx = \frac{3x^5}{5}\Big|_0^1 = \frac{3}{5}$

$Var(X) = E(X^2) - \mu^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}$

$\sigma = \sqrt{\frac{3}{80}} = \sqrt{\frac{3}{5}}\cdot\frac{1}{4}$

MEDIAN

THE MEDIAN OF A CONTINUOUS RV X IS FOUND BY SOLVING $P(X \leq Q_2) = F(Q_2) = 0.5$. WE WANT HALF THE AREA UNDER THE PDF TO BE ON EITHER SIDE OF $x = Q_2$. SIMILARLY, THE UPPER AND LOWER QUARTILES SATISFY $F(Q_3) = 0.75$, $F(Q_1) = 0.25$.

IF f IS SYMMETRIC ABOUT μ, THEN $Q_2 = \mu$.

EX: LET $F(x) = x^3$, $0 < x < 1$. FIND μ AND Q_2. WHAT DOES THIS SAY ABOUT f?

A: $f(x) = \frac{d}{dx}F(x) = 3x^2$. FROM THE PREVIOUS EXAMPLE,

$\mu = \frac{3}{4}$.

$F(Q_2) = 0.5 \Rightarrow Q_2^3 = 0.5 \Rightarrow Q_2 = 0.5^{1/3} \approx 0.7937$.

THE MEDIAN IS BIGGER THAN THE MEAN, SO f IS SKEWED TO THE LEFT.