PROBABILITY

- Sample space and Events
- Rules of Probability
- Conditional Probability and Independence
- Random variables, distribution, mean, s.d of random variables
- Binomial, normal distribution

PROBABILITY

- Toss a coin
- P(H)= $\frac{1}{2}$ P(T)=$\frac{1}{2}$
- Tacitly assumed 'fair' coin. Two possible outcomes. Both equally likely
- S = $\{H, T\}$- Sample space

PROBABILITY

A box has :
2 tickets with # 1, 3 with # 2
4 with # 3 on it 1 with # 4 on it 5 with # 5 on it 5 with # 6 on it
Pick a ticket at random from the box

Number	Probability
1	2/20
2	3/20
3	4/20
4	1/20
5	5/20
6	5/20

A: divisible by 2 =$\{2, 4, 6\}$
B: divisible by 5 =$\{5\}$
C: divisible by 3 =$\{3, 6\}$

P(A or B) = P(A) + P(B)
P(A or C) $\neq$ P(A) + P(C)

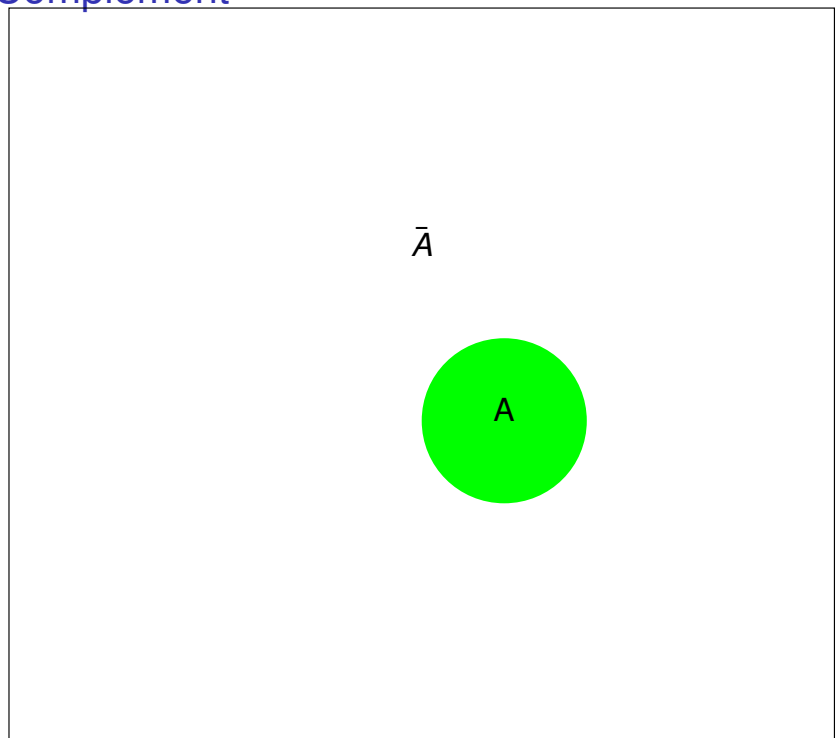
Formal definitions

- SAMPLE SPACE
- Sample space S is the set of all possible outcomes
- EVENT
- An event is a subset of the sample space. The event occurs if the outcome belongs to the subset.

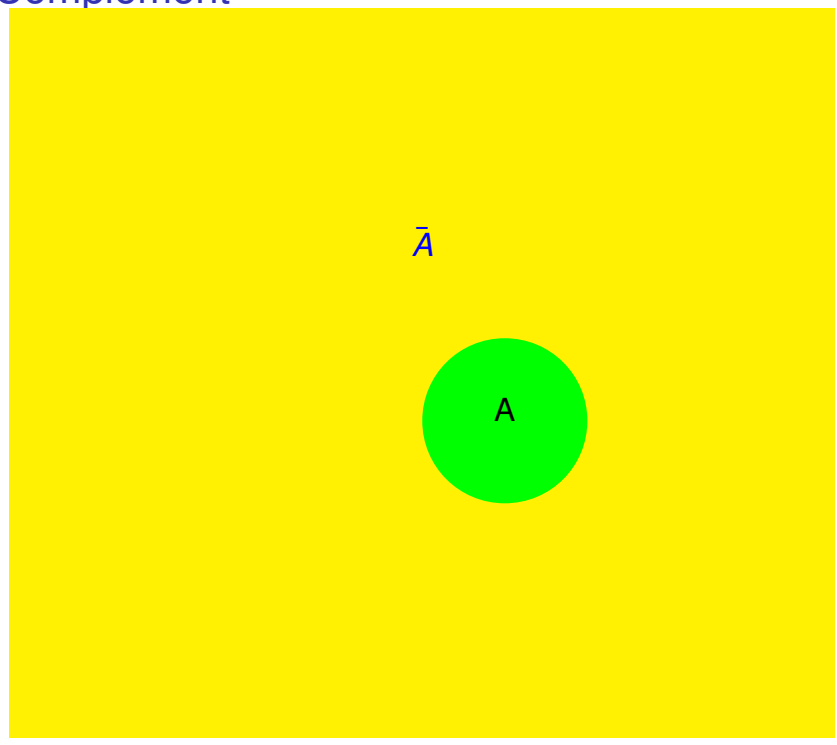
Formal definitions

- A is an event. $\bar{A}$ (A- COMPLEMENT) stands for “Not A”. A does not occur
- A, B events. $A \cup B$ (read A UNION B) At least one of A, B occurs. A or B . All outcomes in at least one of A,B
- $A \cap B$ (read A INTERSECTION B) I Both A and B occur. All outcomes which are both in A and in B. (often written as AB)

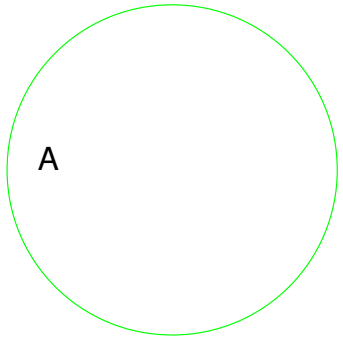
A-Complement



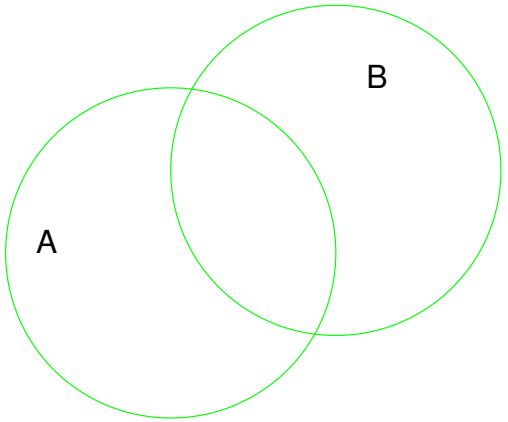
A-Complement



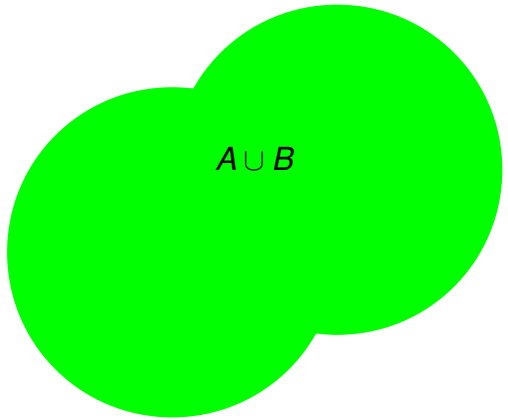
A-Union B



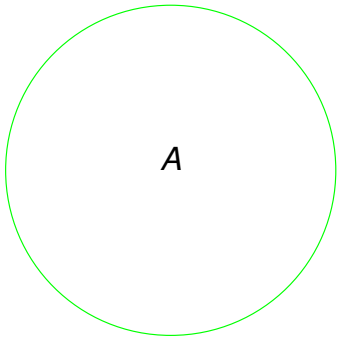
A-Union B



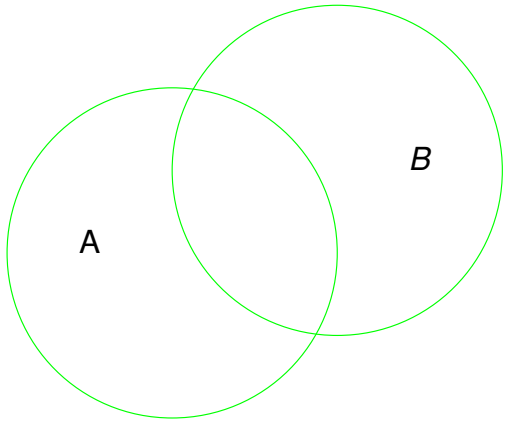
A-Union B



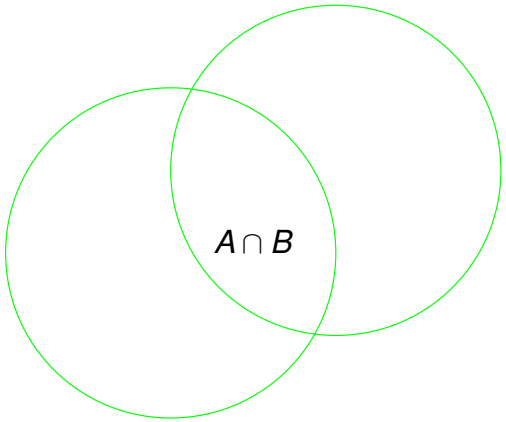
A-intersection B



A-intersection B



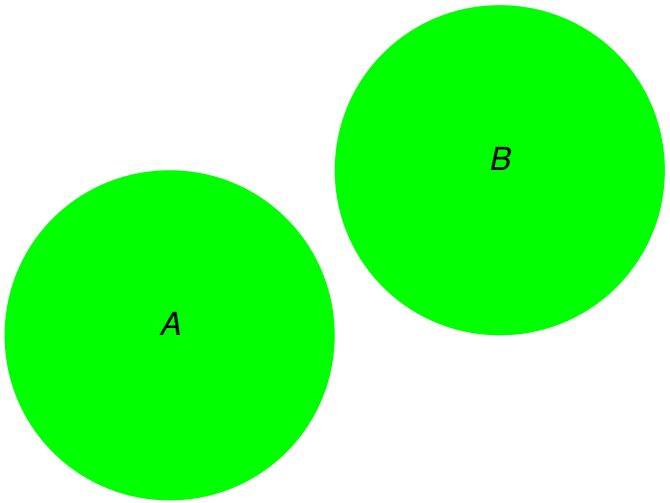
A-intersection B



Mutually Exclusive events

A and B are MUTUALLY EXCLUSIVE OR DISJOINT if
$A \cap B = \emptyset$
i.e.,A and B have no outcomes in common
Both A and B cannot occur at the same time

Mutually Exclusive



Convince yourselves that

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$

Probability: Formal Properties

Probability is an assignment of numbers P(A) to every event A, such that

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. If A and B are mutually exclusive, since $P(A \cap B) = 0$, $P(A \cup B) = P(A) + P(B)$
5. (i) and (ii) imply that $P(\overline{A}) = 1 - P(A)$

Frequency interpretation

- Repeat the experiment large number of times
- Find the relative freq of the event (getting '3')
- The relative freq. is close to probability of the event.

Bayesian interpretation

- Probability is a measure of (experimenter's) uncertainty
- Subjective, Prob depends on the experimenter
- With each observation updates the initial probability
- With more observations revisions gets close to rel. freq

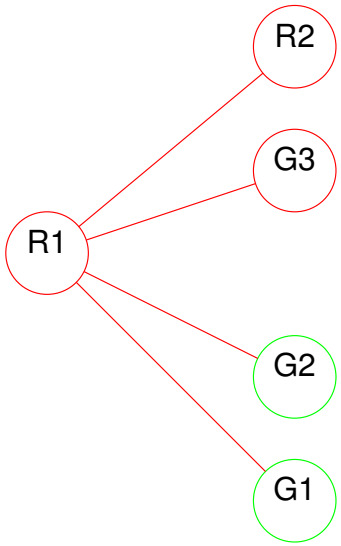
CONDITIONAL PROBABILITY

Look at the following table

	right handed	left handed	total
male	43	9	52
female	44	4	48
Total	87	13	100

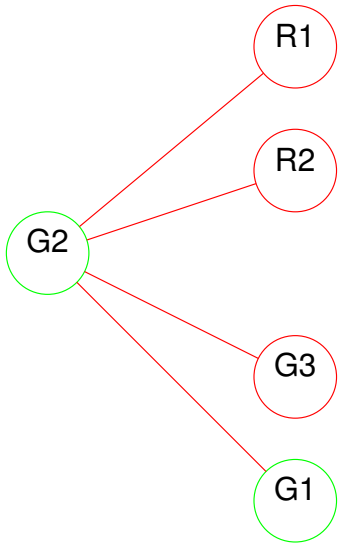
L : Left Handed
M: Male

Pick a person at random
P(ML)= 9/100
P(M) = 52/100 P(L) = 13/100
Picked person is M. What is the prob. he is L ? 9/52
Conditional probability of L given M = P(L|M)
We see P(ML)= P(M) P(L|M)



CONDITIONAL PROBABILITY

Suppose there is a box with 3 Green and 2 Red balls. Draw one at random note the color and then draw one more



Conditional Probability

- There are 5 choices for the first and 4 for the second: 20 in all
- Let A: The first draw is red : $P(A) = \frac{8}{20}$ since

$$A = \{R_1R_2, R_1, G_1, R_1G_2, R_1G_3, R_2R_1, R_2G_1, R_2G_2, R_2G_3\}$$

- Let B: The second draw is red
- Let C: The first draw is green $P(C) = \frac{12}{20}$
- Let D: The second draw is green
- $P(A \cap B) = \ = \{R_1R_2, R_2R_1\} = \frac{2}{20}$
- $P(A \cap D) = \frac{6}{20}$

$$\frac{2}{20} = \frac{2}{5} \times \frac{1}{4}$$

$P(A)$ $P(B|A)$

Prob. of B Given A

In general for any two events A, B,

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

Prob. of B Given A

Bayes rule

We know $P(A), P(B|A, P(B|A^c)$, find $P(A|B)$.

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

$$P(B) = P(AB) + P(A^cB) = P(A)P(B|A) + P(A^c)P(B|A^c)$$

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

Bayes rule: Application to HIV testing

- Tests are not foolproof.
- Two random quantities
  - status of the subject : A – infected, A^c– not infected
  - Test shows positive– B , shows negative B^c
- Question of interest: if the test is Positive, what is the probability that the subject is infected? P(A|B)
- The answer to this question is the theme of Bayes Rule.
- What data are relevant and are generally available? Let us see.

Bayes rule: Application to HIV testing

- A - is the incidence rate of the disease (say .0001)
- Administer the test on subjects known to be not infected. The proportion of false positives gives an estimate of $P(B|A^c)$, say .1
- Administer the test on subjects known to be infected. The proportion of positives gives an estimate of $P(B|A)$ say .95
- Bayes theorem gives

$$P(A|B) = \frac{.0001 \times .95}{.0001 \times .95 + .9999 \times .1} = .09$$

Independence

- A, B are independent
- Knowing A has occurred gives no information about B
- Formally, $P(B|A) = P(B)$
- $P(A \cap B) = P(A)P(B)$

- If A, B are mutually exclusive (disjoint) can they be independent?
- Being mutually exclusive $P(A \cap B) = 0$
- If A, B are independent $P(A \cap B) = P(A)P(B) = 0$!!
- If mutually exclusive they cannot be independent

Summary Formulas

- $P(\overline{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A)P(B|A)$
- $P(A \cap B) = P(B)P(A|B)$
- If A, B are independent
  $P(A \cap B) = P(A)P(B)$

To include an issue in the ballot, a certain number of signatures are required. The supporters of the issue stand at various public places and collect signatures in a sheet. Each sheet has 20 lines but all the lines may not be filled. At the end, when all the signatures are submitted,the election officer, in order to validate the submission,chooses a certain number of signatures at random and follows through.

- draw sheets at random
- from each chosen sheet, pick a signature at random

10 signatures A | 20 signatures B

Draw a box at random and then draw a signature.

$$P(A) = .5 \times \frac{1}{10} \qquad P(A) = .5 \times \frac{1}{20}$$

- There are 185 contestants. Number them as 001,002,....011,....185.
- Pick a digit 0 or 1 at random ,
- pick a digit between 0 and 8 at random
- pick a digit between 0 and 5 at random
- make up three digits.
- is this reasonable?

Topics

- Discrete random variables
  - Distribution, Mean, Standard deviation
  - Binomial
- Continuous random variables
  - probability density, mean, s.d. examples
  - Normal distribution

Example

Go back to the example

Number	Probability
1	2/20
2	3/20
3	4/20
4	1/20
5	5/20
4	6/20

If X stands for the number of a randomly picked ticket

- X is called a random variable
- The above table is called the Probability distribution of X

A random variable is

- an uncertain quantity whose values depend on chance
- Formally, is a function defined on a sample space
- A probability model specifies a way of calculating probabilities involving the random variable

Discrete Random Variables

A random variable is

- DISCRETE: if it can assume finitely many or countably many values
- CONTINUOUS: if it can assume values in a continuum, Ex. Time, Height . Typically if it takes a large number of distinct values, it is taken as continuous, Ex. Income

X	x_1	x_2	x_3	$\cdots$	
Prob.	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\cdots$	

1. $P(x_i) \geq 0$
2. $\sum_i P(x_i) = 1$

All relevant information about a discrete random variable is contained in its probability distribution

- List of values
- corresponding probabilities

Expected value, Variance and S.D

Back to the example:

Number	Probability
1	2/20
2	3/20
3	4/20
4	1/20
5	5/20
6	5/20

The average of the numbers is

$$\frac{1 \times 2 + 2 \times 3 + 3 \times 4 + 4 \times 1 + 5 \times 5 + 6 \times 5}{20}$$

$$1\times(2/20)+2\times(3/20)+3\times(4/20)+4\times(1/20)+5\times(5/20)+6\times(5/20)$$

$$1\times P(X=1)+2\times P(X=2)+3\times P(X=3)+4\times P(X=4)+5\times P(X=5)$$

X	x_1	x_2	x_3	$\cdots$	x_n
Pr.	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\cdots$	$P(x_n)$

- The mean or expected value of X is

$$\mu_X = \sum_i^n x_i P(x_i)$$

  μ_X denoted by μ
- Variance

$$V(X) = \sum_i^n (x_i - \mu)^2 P(x_i) = \sigma^2$$

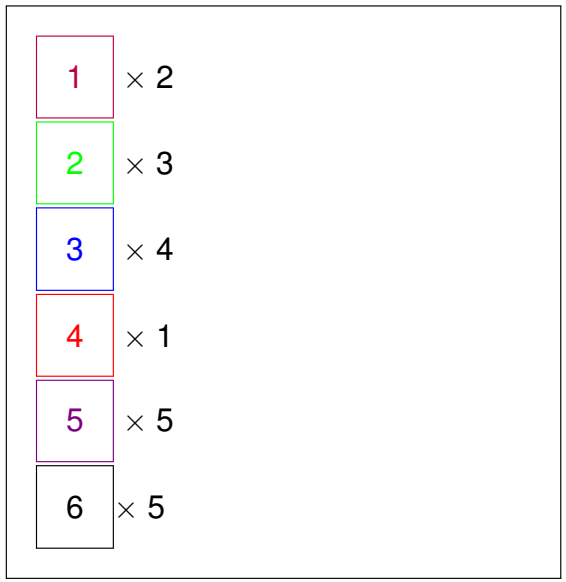
- Standard deviation of X
  = $\sqrt{V(X)}$ usually denoted by σ.

X	x_1	x_2	x_3	$\cdots$	x_n
Prob.	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\cdots$	$\frac{1}{n}$

- $\mu_X = \frac{\sum_i^n x_i}{n} = \bar{x}$
- $V(X) = \frac{\sum_i^n (x_i - \bar{x})^2}{n}$
  Same as (population) variance of
  $x_1, x_2, x_3, \ldots, x_n$

Properties

- $E(aX + b) = a\mu_X + b$
- $V(aX + b) = a^2 V(X)$
- s.d$(aX + b) = |a|s.d.(X)$



Number	Prob
1	2/20
2	3/20
3	4/20
4	1/20
5	5/20
6	5/20

Interpretation

Only proportions matter

Population Distribution

1	× 2000
2	× 3000
3	× 4000
4	× 1000
5	× 5000
6	× 5000

Prob. Distribution

Number	Prob	
1	2/20	.1
2	3/20	.15
3	4/20	.2
4	1/20	.05
5	5/20	.25
6	5/20	.25

- Drawing n samples (data) gives n observations of X
- If repeated random samples were drawn from the population
- (if repeated observations were made of X)
- proportion of 1 will be approx. 10%
- proportion of 2 will be approx 15%
- .....
- average of the data will be approx μ_X
- s.d. of data will be approx $s.d(X)$
- can use computer to simulate

Interpretation

- How do random variables arise?
- Draw a sample from a population and observe the value of X for the sample
- Population distn $\longrightarrow$ Distribution of X
- population average $\longrightarrow$ μ_X
- population s.d $\longrightarrow$ s.d (X)

Group testing

During the second world war it was desired to test the soldiers for the presence of syphillis. Each section consists of 10 soldiers. Think of the following two methods.

1. take blood sample from each soldier and test
2. take blood sample from each soldier in a section. Take a part of the sample from each soldier, mix them and test. If it is negative then there is no need for further test. If positive then test each soldier in the section. Assume that incidence of syphillis is 1 in 100. Which of the two methods would you recommend?

Incidence and Expected number of tests

In method one the expected number of tests is 10
In method two it is 1 with probability $(.99)^{10} = .9$ and 11 with probability .1. So the expected value is $.9 + 1.1 = 2$

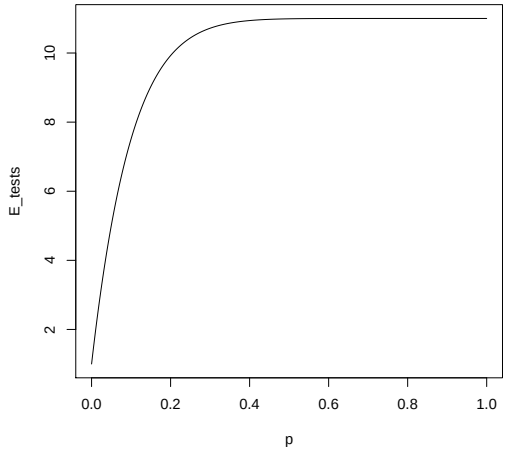


Figure: Incidence Vs expected tests

Binomial

- A large population consists of objects of two types Call them S(success), F (failure)
- Want to estimate the proportion of S in the population from a sample
- estimate the proportion of one type gives estimate for the other

- Draw n samples at random.
  With replacement after each draw Binomial Model
- Draw n samples at random.
  without Replacement after each draw Hyper Geometric Model
- Fix a number k. Keep drawing till you obtain k Green balls. How many draws did you need. Negative Binomial Model

Relation between with and without replacement

If N, the population size is large compared to n then can the probabilities arising from with and without replacement are nearly equal

Binomial

An X that counts the number of successes in many independent bernoulli trials is called a binomial random variable. The two parameters are

- "n" the number of trials
- " p" the probability of success in a trial

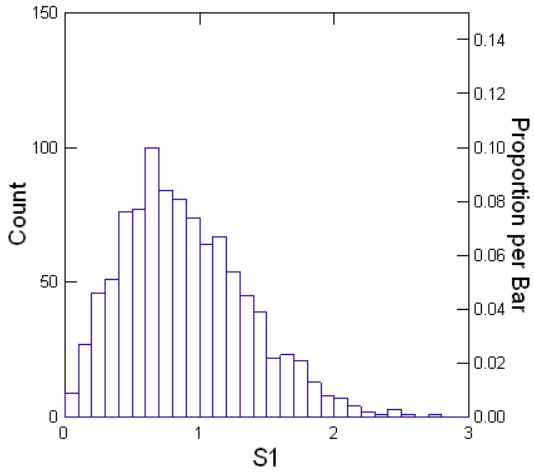
- "n" trials
- Each trial has only two outcomes, Success or Failure
- trials are independent
- In all trials, prob of success = p
- $X = \#$ successes is Bin (n,p)

Issues

- Recognizing a binomial
- $\mu_X = np$, $V(X) = np(1-p)$
  s.d (X) $= \sqrt{(np(1-p)}$
- Calculating probabilities: from table, software

Continuous Random variables

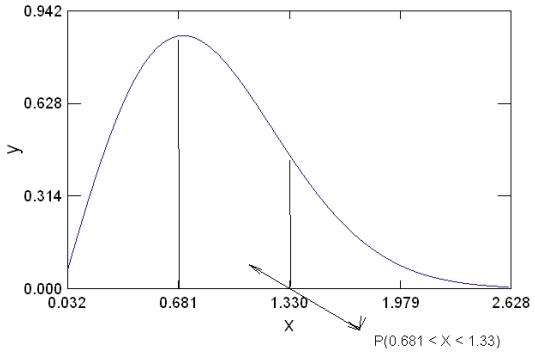
- How to we model a continuous random variable?
- specify prob of each possible value? NO
- The probabilistic structure is described by a probability density function



- histogram.jpg

Continuous Random variables

- For continuous random variables, probability is modelled by a PROBABILITY DENSITY FUNCTION.
- Area under the curve between two points gives the probability of X lying between those two points



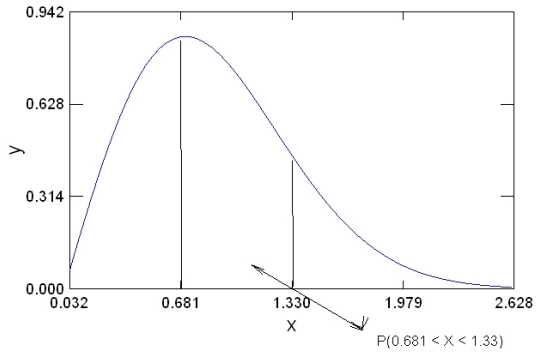


Figure: probability density

Interpretation

- If we draw a histogram of the population it will be close to the density

- If repeated random samples were drawn from the population
- (if repeated observations were made of X)
- proportion of values between a and b will be approximately area under the density between a and b.
- histogram of the observations will be close to the density
- average of the data will be approx μ_X
- s.d. of data will be approx $s.d(X)$
- can use computer to simulate

uniform random variable

- Suppose X is a number picked at 'random ' from the interval $[0, 1]$
- $P(X = x) = 0$ for all x
- But Probability X falls in an interval is equal to the length of the interval,and is nonzero.

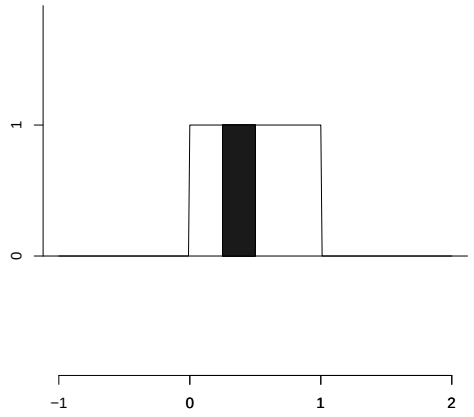
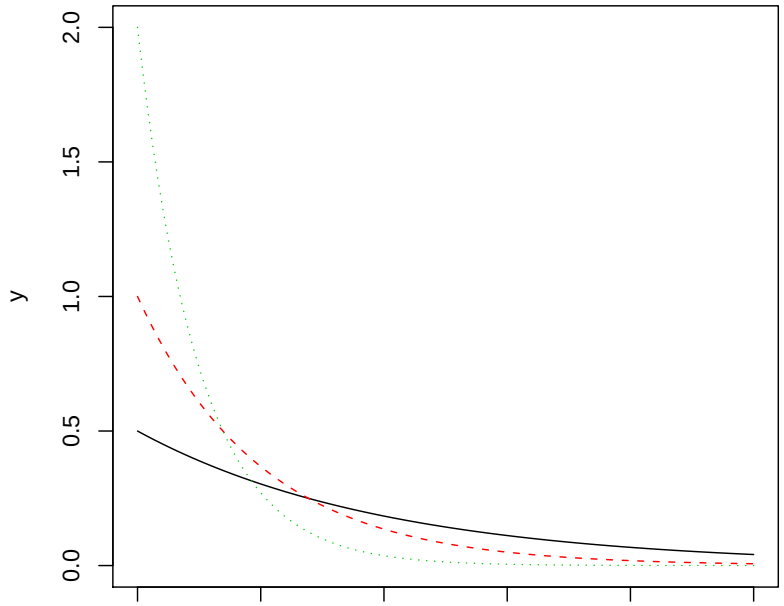
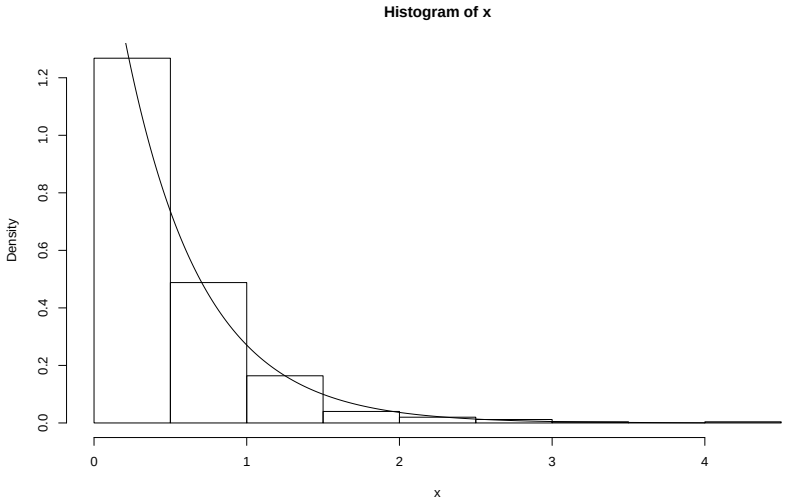


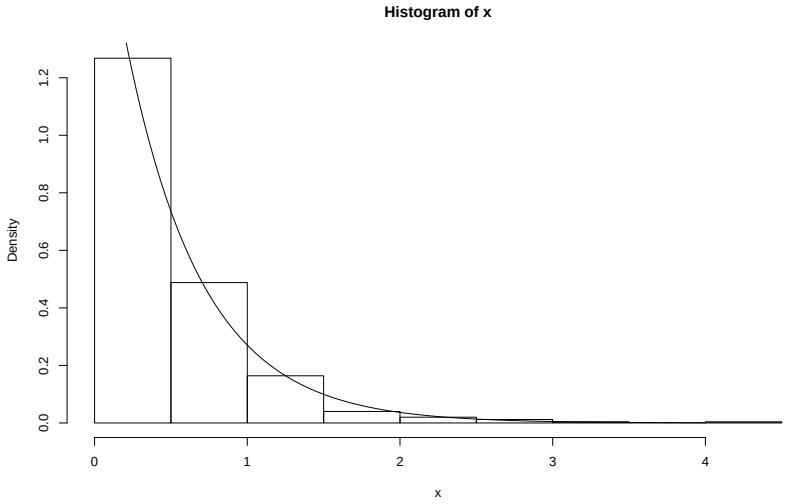
Figure: Uniform density

Exponential density

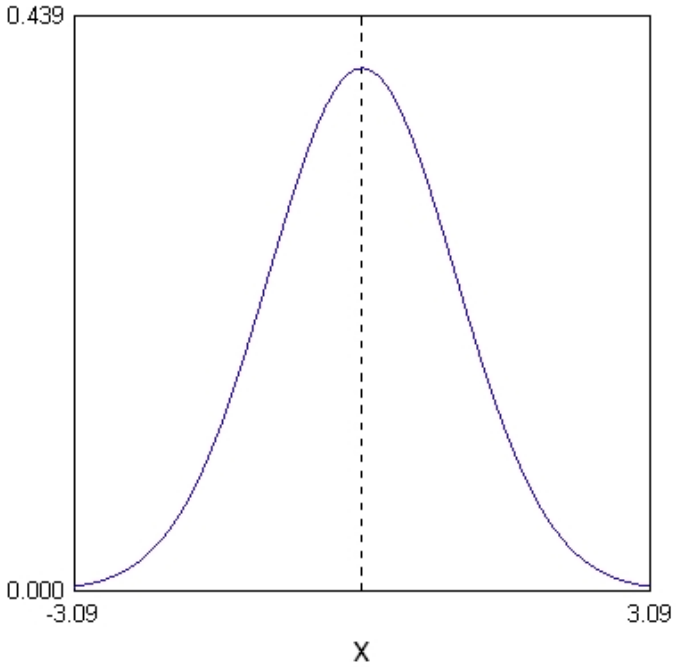


- When should we use exponential as a model?
- usually used as a model for life time, occurrence of events like earth quakes etc.
- Memory less property
- if we have past data, and if the histogram looks like exponential





NORMAL DISTRIBUTION



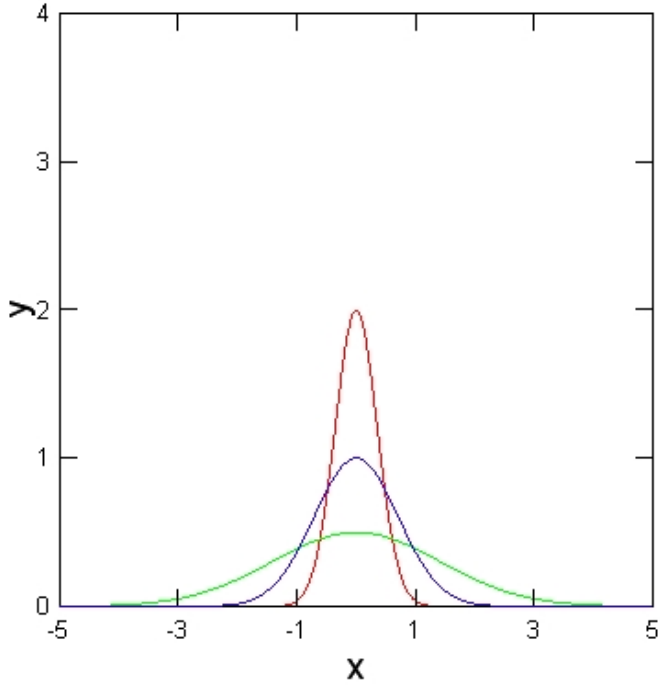


Figure: Normal – different sd

Properties

- symmetric around μ
- $E(X) = \mu, s.d(X) = \sigma$
- models phenomenon where the random variable varies from the centre symmetrically. Small deviations have high probability and large deviations have small probability
- The normal density with $\mu = 0, \sigma = 1$ is called standard normal distribution

Normal Distribution-Properties

- When measured in the scale of standard deviations, with origin at the mean, all normal densities reduce to standard normal.
- If X is $N(\mu, \sigma)$ then $Z = \frac{x-\mu}{\sigma}$ is standard normal.
- $$P(a < N(\mu, \sigma) < b) = P(\frac{a-\mu}{\sigma} < N(0,1) < \frac{b-\mu}{\sigma}$$
- The most appropriate scale for normal random variables is " standard deviations" with the mean as the origin
- the density of standard normal is usually denoted by ϕ and the distibution function by Φ

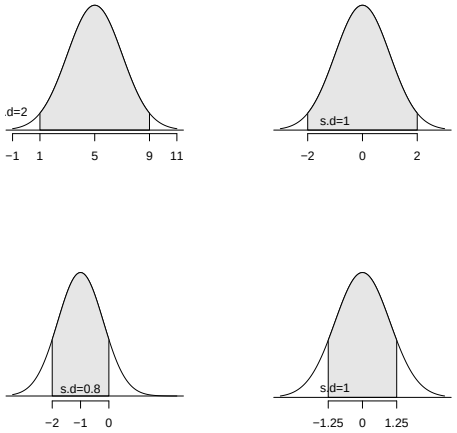


Figure: Normal areas

Normal distribution

Formally, If X is $N(\mu, \sigma)$ then

$$Z = \frac{X-\mu}{\sigma} \text{ is } N(0,1)$$

s.d's away from mean

Chebyshev's Inequality

- $P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

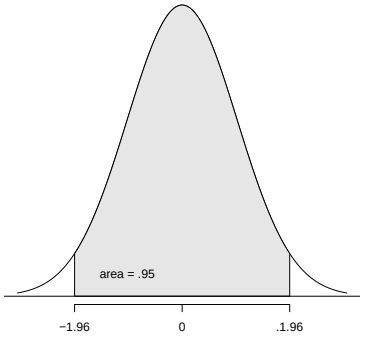
In particular, take $k = 3$,

$$P(|X - \mu| \leq 3\sigma) \geq 1 - \frac{1}{9} \approx 90\%$$

i.e, with probability larger than 90% the observation will be within 3 standard deviation of the mean.

Normal distribution

When, in addition to mean and s.d. we also assume normal population then:



X lies within 1.96 standard deviations of the mean with 95% probability