Semester 2, 2012 (Last adjustments: August 13, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: August 13, 2012)

Tuesday, 14th August 2012

Lecture 6 - Content

- ☐ **Sets**
- ☐ **Probability and counting**
- ☐ **Conditional probability**

Sets

Before we look at probability it is necessary to understand sets because probabilities are typically described in terms of sets where an 'event' occurs.

Definition 1. The set of all possible outcomes of an experiment is called a sample space, denoted by Ω. Any subset A of the sample space Ω, denoted by $A \subset \Omega$ is called an event.

Definition 2. The counting operator $N(A)$ is a set function that counts how many elements belong to the set (event) A.

Example (Sample spaces)**.**
Coin: $\Omega = \{H, T\} \Rightarrow N(\Omega) = 2$.
Dice: $\Omega = \{1, 2, 3, 4, 5, 6\} \Rightarrow N(\{1, 2, 5\}) = 3$;
Weight: $\Omega = \mathbb{R}^+ \Rightarrow N(\mathbb{R}^+) = \infty$.

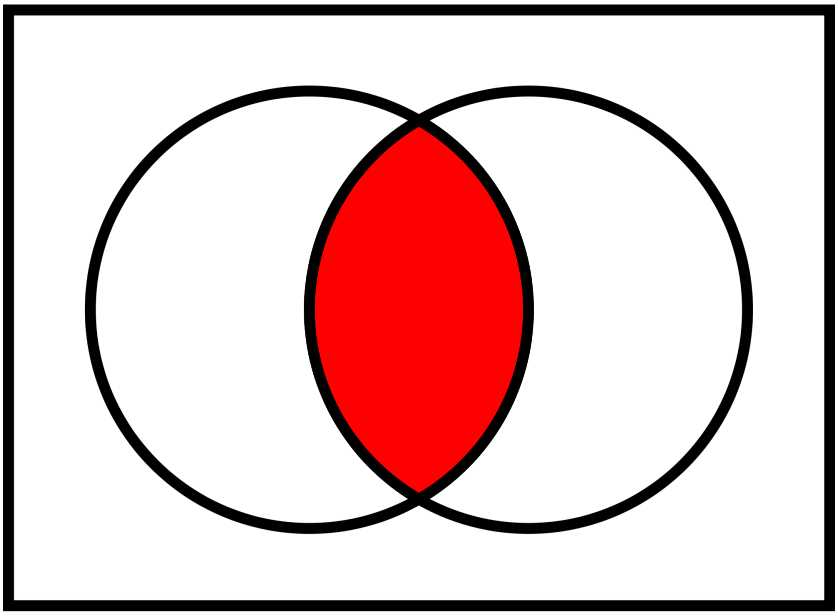
Set Notation

Before we introduce probability we need to introduce some notation. Let $A, B \subset \Omega$.

symbol	set theory	probability
Ω	largest set	certain event
$\emptyset$	empty set	impossible event
$A \cup B$	union of A and B	event A or event B
$A \cap B$	intersection of A and B	event A and event B
$A^C = \Omega \backslash A$	complement of A	not event A

Intersection Operator

The set $A \cap B$ denotes the set such that if $C \in A \cap B$ then $C \in A$ and $C \in B$ ($\cap$ is called the intersection operator).



Intersection Operator

Examples:

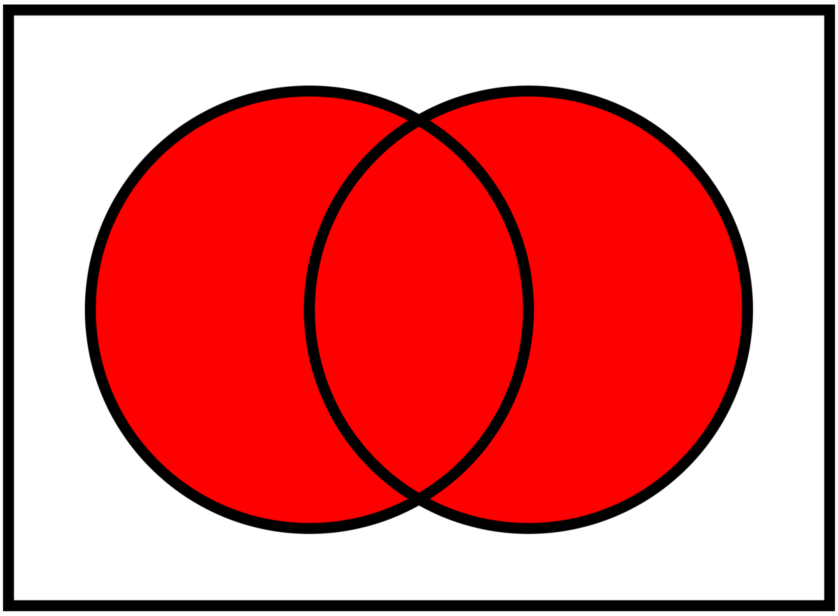
- $\{1, 2\} \cap \{\text{red}, \text{white}\} = \emptyset$.
- $\{1, 2, \text{green}\} \cap \{\text{red, white, green}\} = \{\text{green}\}$.
- $\{1, 2\} \cap \{1, 2\} = \{1, 2\}$.

Some basic properties of intersections:

- $A \cap B = B \cap A$.
- $A \cap (B \cap C) = (A \cap B) \cap C$.
- $A \cap B \subseteq A$.
- $A \cap A = A$.
- $A \cap \emptyset = \emptyset$.
- $A \subseteq B$ if and only if $A \cap B = A$.

Union Operator

The set $A \cup B$ denotes the set such that if $C \in A \cup B$ then $C \in A$ and/or $C \in B$ ($\cup$ is called the union operator).



Union Operator

Examples:

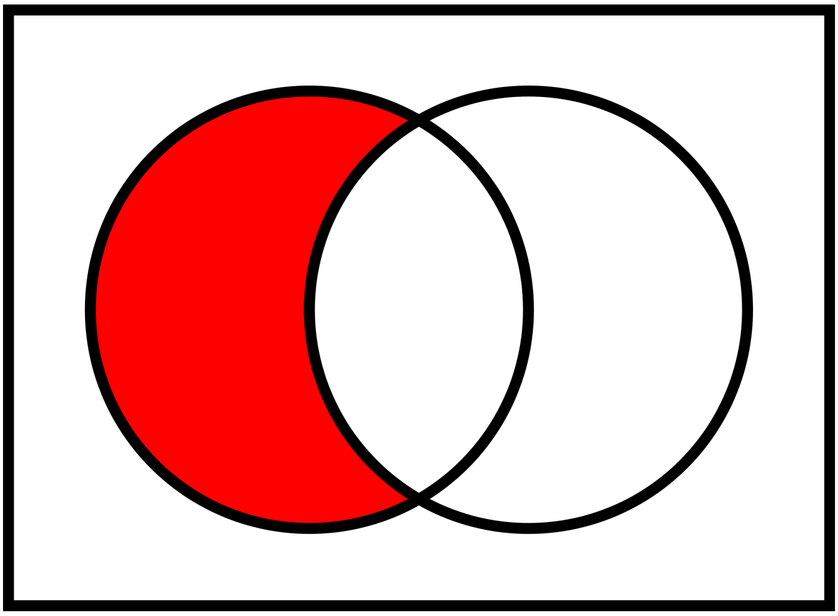
- $\{1, 2\} \cup \{$red, white$\} = \{1, 2,$ red, white$\}$.
- $\{1, 2,$ green$\} \cup \{$red, white, green$\} = \{1, 2,$ red, white, green$\}$.
- $\{1, 2\} \cup \{1, 2\} = \{1, 2\}$.

Some basic properties of unions:

- $A \cup B = B \cup A$.
- $A \cup (B \cup C) = (A \cup B) \cup C$.
- $A \subseteq (A \cup B)$.
- $A \cup A = A$.
- $A \cup \emptyset = A$.
- $A \subseteq B$ if and only if $A \cup B = B$.

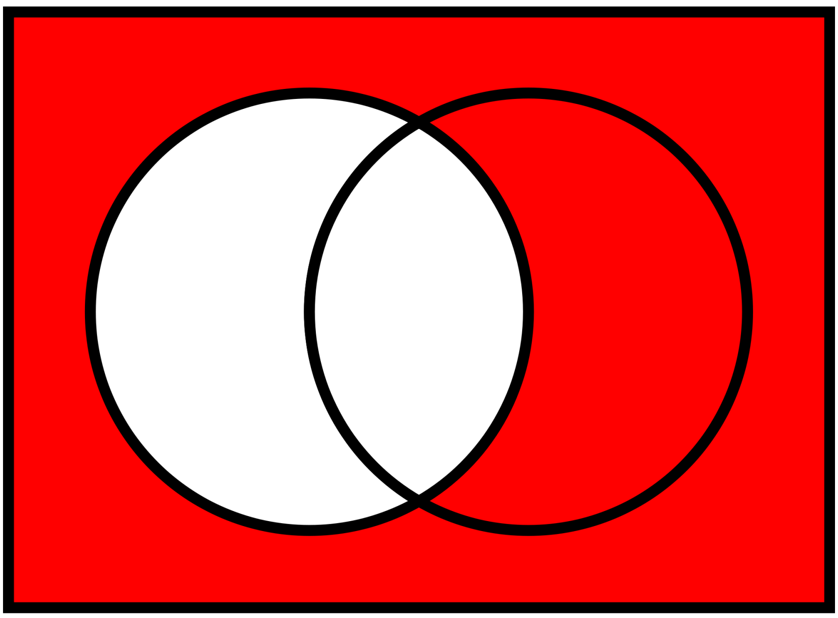
Set Minus

The set $A \setminus B$ denotes the set such that if $C \in A \setminus B$ then $C \in A$ and $C \notin B$.



Set Complement

The set $A^c = \Omega \setminus A$ denotes the set such that if $C \in A^c$ then $C \notin A$.



Set Minus

Examples:

- $\{1, 2\} \setminus \{\text{red, white}\} = \{1, 2\}$.
- $\{1, 2, \text{green}\} \setminus \{\text{red, white, green}\} = \{1, 2\}$.
- $\{1, 2\} \setminus \{1, 2\} = \emptyset$.
- $\{1, 2, 3, 4\} \setminus \{1, 3\} = \{2, 4\}$.

Some basic properties of complements:

- $A \setminus B \neq B \setminus A$.
- $A \cup A^c = \Omega$.
- $A \cap A^c = \emptyset$.
- $(A^c)^c = A$.
- $A \setminus A = \emptyset$.

Theorem 1. The complement of the union of A and B equals the intersection of the complements
$$(A \cup B)^C = (A^C) \cap (B^C).$$

Proof. Use Venn diagrams for LHS and RHS and colour areas. □

Theorem 2. de Morgan's Laws.
$$\left(\bigcap_{i=1}^{n} A_i \right)^c = \bigcup_{i=1}^{n} A_i^c$$
and
$$\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c$$

Counting – Ordered Sampling without replacement

Example (Ordered samples without replacement)**.** The number of ordered samples of size r we can draw without replacement from n objects is,

$$n \times (n-1) \times \ldots \times (n-r+1) = \frac{n!}{(n-r)!}$$

Recall: $0! = 1$.

Counting – Unordered Sampling without replacement

Example (Unordered samples without replacement)**.**

$${}^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} = n \text{ \underline{C}hoose } r.$$

Recall,

$${}^nC_r = {}^nC_{n-r}$$

since

$$\binom{n}{n-r} = \frac{n!}{(n-r)!((n-(n-r))!} = n \text{ \underline{C}hoose } r.$$

and so

$$\binom{n}{0} = \binom{n}{n} = 1$$

Sampling in R

```
# Creating ordered lists
n = 158;
x = 1:n;
set.seed(6)       # set random seed to 6 to reproduce results
sample(x)         # random permutation of nos 1,2,...,158: n! possibilities
sample(x,10)      # choose 10 numbers without replacement
sample(x,10,TRUE) # choose 10 numbers with replacement = bootstrap sampling
```


What is Probability?

1. Subjective probability expresses the strength of one's belief (the basis of Bayesian Statistics – a bit on that later).

2. Classical probability concept, mathematical answer for equally likely outcomes.

   Theorem 3. If there are n equally likely possibilities of which one must occur and s are regarded as favourable ($=$ successes), then the probability P of a success is given by s/n.

What is Probability?

3. The frequency interpretation of probability:

   Theorem 4. The probability of an event (or outcome) is the proportion of times the event occur in a long run of repeated experiments.

   or in words:

   > If an experiment is repeated n times under **identical conditions**, and if the event A occurs m times, then as n becomes large (i.e. in the long-run) the probability of A occurring is the ratio m/n.

What is Probability?

- ☐ The constancy of the gender ratio at birth. In Australia, the proportion of male births is fairly stable at 0.51. This long run relative frequency is used to estimate the probability that a randomly chosen birth is male.
- ☐ Cancer council records show the age standardised mortality rate from breast cancer in NSW was close to 20 per 100,000 over the years 1972-2000. For a randomly chosen woman, we use 0.0002 as the probability of breast cancer.

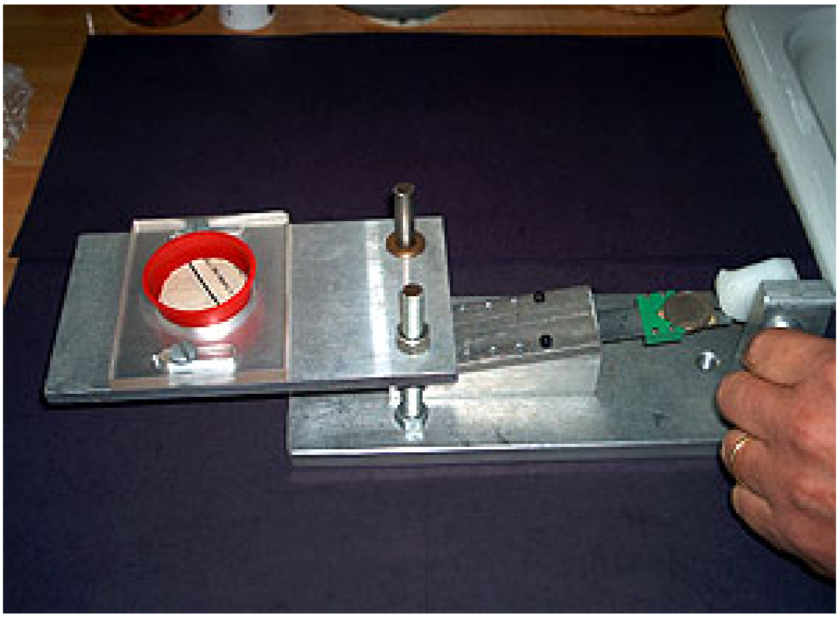
Example (Coin tossing)**.**
Buffon (1707-1788):
Pearson (1857-1936):

Coin Tossing in R

```
table(sample(c("H","T"),4040,T))/4040
table(sample(c("H","T"),24000,T))/24000
```

Coin Tossing 2010's

In the 2010's Stanford Professor Persi Diaconis developed the "Coin Tosser 3000".



However, the machine is designed to flip a coin with the same result **every time**!

What is Probability?

4. Mathematical formulation of probability

   Definition 3 (due to Andrey Kolmogorov, 1933). Given a sample space Ω $A \subset \Omega$, we define $\mathrm{P}(A)$, the probability of A, to be a value of a non-negative additive set function that satisfies the following three axioms:
   A1: For any event A, $0 \leq \mathrm{P}(A)$,
   A2: $\mathrm{P}(\Omega) = 1$,
   A3: If A and B are mutually exclusive events ($A \cap B = \emptyset$), then

   $$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

   A3': If $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of mutually exclusive events in Ω, then

   $$\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_2) + \ldots.$$

$0 \le \mathrm{P}(A) \le 1$

Theorem 5. Assume the following 3 axioms:
A1: For any event $A \subset \Omega$, $0 \le \mathrm{P}(A)$,
A2: $\mathrm{P}(\Omega) = 1$,
A3': If $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of mutually exclusive events in Ω, then

$$\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_2) + \ldots .$$

Then $0 \le \mathrm{P}(A) \le 1$.

Proof.

- $\square$
- $\square$
- $\square$

$\square$

Example (Lotto). A lotto type barrel contains 10 balls numbered $1, 2, \ldots, 10$. Three balls are drawn.

i. How many distinct samples can be drawn?

ii. Event $A = \{1, 2, \ldots, 7\}$ (all numbers less than seven).

$$\mathrm{P}(A) = \qquad\qquad .$$

iii. $B =$ all drawn numbers are even: $\mathrm{P}(B) = \frac{1}{120} \times \binom{5}{3} = \frac{10}{120} = \frac{1}{12}$.

$$\mathrm{P}(A \cap B) =$$

iv. $\mathrm{P}(A \cup B)$? To answer this we need our next theorem.

Addition Theorem

Theorem 6 (Addition Theorem)**.** If A and B are any events in Ω, then

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

Proof. Use Venn diagrams, i.e. draw pictures and colour regions.

use axioms only

$\square$

Example (Lotto)**.** A lotto type barrel contains 10 balls numbered $1, 2, \ldots, 10$. Three balls are drawn.

i. How many distinct samples can be drawn? 120.

ii. Event $A = \{1, 2, \ldots, 7\}$ (all numbers less than seven). $\mathrm{P}(A) = \frac{7}{24}$.

iii. $B =$ all drawn numbers are even: $\mathrm{P}(B) = \frac{1}{12}$.
Also $\mathrm{P}(A \cap B) = 1/120$.

iv. $\mathrm{P}(A \cup B)$?

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = \frac{7}{24} + \frac{1}{12} - \frac{1}{120} = \frac{44}{120} = \frac{11}{30}.$$

Poincarés' Theorem

Theorem 7 (Poincarés' formula, not part of M1905)**.** Let $A_1, A_2, \ldots, A_n$ be any events in Ω. Then,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k) + \ldots$$
$$+(-1)^{n-1}\,\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right).$$

(Unconditional) probability

- Recall 3 Axioms of probability.
- $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$ since $A \cap A^C = \emptyset$ hence, $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^C) = \mathrm{P}(A) + \mathrm{P}(A^C)$.
- $\mathrm{P}(\emptyset) = 0$ because $\emptyset = \Omega^C$, hence $\mathrm{P}(\emptyset) = 1 - \mathrm{P}(\Omega)$.
- etc.

Conditional Probability – Another Motivating Example

What is the probability of the important event

$$A = (\text{starting salary after uni } \geq 60\text{k})?$$

What is the sample space Ω?

Possibilities are:

$$\begin{aligned} \Omega_1 &= \{\text{all students}\}, \\ \Omega_2 &= \{\text{all male students}\}, \\ \Omega_3 &= \{\text{all students with a maths degree}\}. \end{aligned}$$

Conclusion

- ☐ Probability depends on the underlying sample space Ω!
- ☐ Hence, if it is unclear to what sample space A refers to then make it clear by writing

  $$\mathrm{P}(A|\Omega) \quad \text{instead of} \quad \mathrm{P}(A)$$

  which we read as the conditional probability of A relative to Ω or given Ω, respectively.

Definition 4. If A and B are any events in Ω and $\mathrm{P}(B) \neq 0$ then, the conditional probability of A given B is

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

Additional material for Lecture 6

A combinatorial proof of the binomial theorem

The binomial theorem says

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Consider the more complicated product

$$(x_1+y_1)(x_2+y_2)\cdots(x_n+y_n)$$

Its expansion consists of the sum of 2^n terms, each term being the product of n factors. Each term consists either x_k or y_k, for each $k = 1, \ldots, n$. For example,

$$(x_1+y_1)(x_2+y_2) = x_1x_2 + x_1y_2 + y_1x_2 + y_1y_2$$

Now, there is $1 = \binom{n}{0}$ term with y terms only, $n = \binom{n}{1}$ with one x term and $(n-1)$ y terms etc. In general, there are $\binom{n}{k}$ terms with exactly k x's and $(n-k)$ y's. The theorem follows by letting $x_k = x$ and $y_k = y$.

More on set theory

The operation of forming unions, intersections and complements of events obey rules similar to the rules of algebra. Following some examples for events A, B and C:
Commutative law: $A \cup B = B \cup A$ and $A \cap B = B \cap A$
Associative law: $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$
Distributive law: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

Monday, 20th August 2012

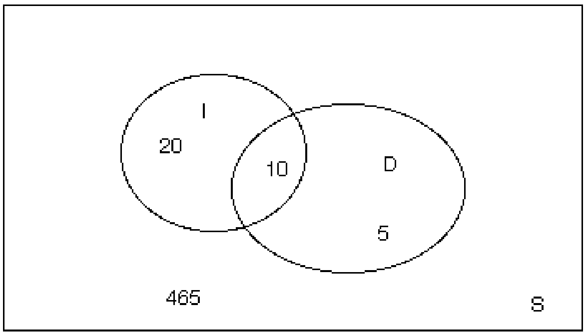
Lecture 7 - Content

- Conditional probability
- Bayes rule
- Integer valued random variables

Conditional probability (cont)

Example (Defect machine parts)**.** Suppose that 500 machine parts are inspected before they are shipped.

- $I =$ (a machine part is improperly assembled)
- $D =$ (a machine part contains one or more defective components)



Example (cont)

Assumption: equal probabilities in the selection of one of the machine parts.
$\Rightarrow$ Using the classical concept of probability we get:

General multiplication rule of probability

Theorem 8 (General multiplication rule of probability)**.** If A and B are any events in Ω, then

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(B) \times \mathrm{P}(A|B), \text{ if } \mathrm{P}(B) \neq 0, \text{ changing } A \text{ and } B \text{ yields} \\ &= \mathrm{P}(A) \times \mathrm{P}(B|A), \text{ if } \mathrm{P}(A) \neq 0. \end{aligned}$$

Proof. This holds because,

$$\mathrm{P}(A|B) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \text{ etc.}$$

□

What happens if $\mathrm{P}(A|B) = \mathrm{P}(A)$?
$\Rightarrow$ additional information of B is of no use $\Rightarrow$ special multiplication rule!

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B).$$

Definition of independence of events

Definition 5. If A and B are any two events in a sample space Ω, we say that A is independent of B if and only if

From the general multiplication rule it follows that if $\mathrm{P}(A|B) = \mathrm{P}(A)$ then $\mathrm{P}(B|A) = \mathrm{P}(B)$ and we say simply that A and B are independent.

Alternative View of Independence

Alternatively, if A and B are independent then $P(A \cap B) = P(A) \times P(B)$ and hence,

$$P(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \quad \text{(using Baye's rule)}$$

$$= \frac{\mathrm{P}(A) \times \mathrm{P}(B)}{\mathrm{P}(A)} \quad \text{(using independence)}$$

$$= \mathrm{P}(B).$$

which can also be interpreted as saying that knowing A does not effecting the probability of B.

Independence

In other words the events A and B are independent if the chance that one happens **remains the same** regardless of how the other turns out.

Example. Suppose that we toss a fair coin twice. Let
$A = \{\text{heads of the first toss}\}$
and
$B = \{\text{heads of the second toss}\}$.
Now suppose A occurred. Then

$$P(\{B \text{ knowning } A \text{ has happened}\}) = \tfrac{1}{2}.$$

Independence – Example 2

Example. Consider the following 6 boxes

1	**2**	**3**	**1**	**2**	**3**

Suppose we select a box at random, as it is drawn you see that it is **green**. Then

$$P(A = \{\text{getting a "2"}\}) = \frac{2}{6} = \frac{1}{3}$$

$$P(B = \{\text{getting a "2" if I know it is } \textbf{green}\}) = \frac{1}{3}$$

Knowing the selected box is **green** has not changed our knowledge about which numbers might be drawn.

Hence, the events A and B are independent.

Independence – Example 3

Example. Consider the following 6 boxes

1	**1**	**2**	**1**	**2**	**2**

Suppose we select a box at random, as it is drawn you see that it is **green**. Then

$$P(A = \{\text{getting a "2"}\}) = \frac{3}{6} = \frac{1}{2}$$

$$P(B = \{\text{getting a "2" if I know it is } \textbf{green}\}) = \frac{1}{3}$$

Knowing the selected box is **green** **HAS CHANGED** our knowledge about which numbers might be drawn.

Hence, the events A and B are **NOT** independent.

Independence – Example 4

Example. Two cards are drawn at random from an ordinary deck of 52 playing cards. What is the probability of getting two aces if

(a) the first card is replaced before the second is drawn?

(b) The first card is not replaced before the second card is drawn?

$\Rightarrow$ Independence is violated when the sampling is without replacement.

Independence – Example 5

Medical records indicate that the proportion of children who have had measles by the age of 8 is 0.4. The corresponding proportion for chicken pox is 0.5. The proportion who have had both diseases by the age of 8 is 0.3. An infant is randomly selected. Let A represent the event that he contracts measles, and B that he contracts chicken pox, by the age of 8 years.

- ☐ Estimate $P(A)$, $P(B)$ and $P(A \cap B)$.
  $P(A) = 0.4$, $P(B) = 0.5$ and $P(A \cap B) = 0.3$.
- ☐ Are A and B independent?
  $P(A) \times P(B) = 0.2 \neq P(A \cap B) = 0.3$, so NO, A and B are not independent.

Bayes rule

Example (The burgers are better...). Assume you get your burgers

- □ 60% from supplier B_1
- □ 30% from supplier B_2
- □ 10% from supplier B_3

$\Rightarrow \mathrm{P}(B_1) = 0.6$, $\mathrm{P}(B_2) = 0.3$, and $\mathrm{P}(B_3) = 0.1$.

Interested in the event $A =$(good burger).

Example (cont)

It follows that,

$$A = A \cap (B_1 \cup B_2 \cup B_3) = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3).$$

Note that $(A \cap B_1)$, $(A \cap B_2)$ and $(A \cap B_3)$ are mutually exclusive.
By Axiom 3 we get

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}((A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)) \\ &= \mathrm{P}(A \cap B_1) + \mathrm{P}(A \cap B_2) + \mathrm{P}(A \cap B_3). \end{aligned}$$

Remember the general multiplication rule:
We already know that

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(B) \times \mathrm{P}(A|B), \text{ if } \mathrm{P}(B) \neq 0, \\ &= \mathrm{P}(A) \times \mathrm{P}(B|A), \text{ if } \mathrm{P}(A) \neq 0. \end{aligned}$$

Example (cont)

So we can write

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(B_1)\cdot\mathrm{P}(A|B_1)+\mathrm{P}(B_2)\cdot\mathrm{P}(A|B_2)+\mathrm{P}(B_3)\cdot\mathrm{P}(A|B_3)\\
&= 0.6\cdot\underbrace{\mathrm{P}(A|B_1)}_{0.95,\text{ very good}}+0.3\cdot\underbrace{\mathrm{P}(A|B_2)}_{0.80,\text{ sufficient}}+0.1\cdot\underbrace{\mathrm{P}(A|B_3)}_{0.65,\text{ insufficient}}\\
&= 0.875.
\end{aligned}$$

What did the example teach us?

Strategy: decompose complicated events into mutually exclusive simple(r) events!

Total probability rule

Theorem 9 (Total probability rule). If $B_1, B_2, \ldots, B_n$ are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ then for any event $A \subset \Omega$,

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(B_i) \times \mathrm{P}(A|B_i).$$

Example (Burger, cont). We know already that supplier B_3 is bad. So what is $P(B_3|A)$ (if a burger is good is it from B_3)? By definition of the conditional probability, since $P(A) > 0$,

$$\begin{aligned}
\mathrm{P}(B_3|A) &= \frac{\mathrm{P}(A\cap B_3)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_3\cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_3)\times\mathrm{P}(A|B_3)}{\sum_{i=1}^{3}\mathrm{P}(B_i)\times\mathrm{P}(A|B_i)}\\
&= \frac{0.1\times 0.65}{0.875} = 0.074.
\end{aligned}$$

After we know that a burger is good the probability that it comes from B_3 decreases from 0.1 to 0.074.

Bayes' rule or Theorem

What we just derived is the famous formula, called Bayes' rule or theorem.

Theorem 10 (Bayes Rule)**.** If $B_1, B_2, \ldots, B_n$ are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ then for any event $A \subset \Omega$,

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(A|B_j) \times \mathrm{P}(B_j)}{\sum_{i=1}^{n} \mathrm{P}(A|B_i) \times \mathrm{P}(B_i)}.$$

The probabilities $\mathrm{P}(B_i)$ are called the priori probabilities and the probabilities $\mathrm{P}(B_i|A)$ the posteriori probabilities, $i = 1, \ldots, n$.

Reverend Thomas Bayes (1701 - 1761)

- Born in Hertfordshire (London, England),
- was a Presbyterian minister,
- studied: theology and mathematics,
- best known for Essay Towards Solving a Problem in the Doctrine of Chances ,
- where Bayes' Theorem was first proposed.
- Words: Bayes' rule, Bayes' Theorem, Bayesian Statistics.

Example of Bayes Rule – Screening test for Tuberculosis

	TB (D^+)	No TB (D^-)	
X-ray Positive (S^+)	22	51	73
X-ray Negative (S^-)	8	1739	1747
	30	1790	1820

What is the probability that a randomly selected individual has tuberculosis given that his or her X-ray is positive given that $P(D^+) = 0.000093$?

- ☐ $P(D^+) = 0.000093$ which implies that $P(D^-) = 0.999907$.
- ☐ $P(S^+|D^+) = 22/30 = 0.7333$
- ☐ $P(S^+|D^-) = 51/1790 = 0.0285$

$$\begin{aligned} P(D^+|S^+) &= \frac{P(S^+|D^+)P(D^+)}{P(S^+|D^+)P(D^+) + P(S^+|D^-)P(D^-)} \\ &= \frac{0.7333 \times 0.000093}{0.7333 \times 0.000093 + 0.0285 \times 0.999907} = 0.00239 \end{aligned}$$

Integer valued random variables

Many observed numbers are the random result of many possible numbers.

Definition 6. A random variable X is a real-valued function of the elements of a sample space Ω.

Note that such functions are denoted with capital letters and their images (outcomes) with lower case letters, e.g. x.

Examples.

- ☐ How many times (X) will you be caught speeding?
- ☐ What will your final mark (Y) for MATH1905 be?
- ☐ How old (Z, in years) do you think your stats lecturer is?

Random Variable Example – 3 Coins

Consider tossing three coins. The number of heads showing when the coins land is a random variable: it assigns the number 0 to the outcome {T, T, T}, the number 1 to the outcome {T, T, H}, the number 2 to the outcome {T, H, H}, and the number 3 to the outcome {H, H, H }.

Random Variable Example – 3 Coins

Events	Random Variable	Probability
TTT		
TTH		
THT		$P(X=0)=\frac{1}{8}$
THH	X = { Number of Heads }	$P(X=1)=\frac{3}{8}$
HTT		$P(X=2)=\frac{3}{8}$
HTH		$P(X=3)=\frac{1}{8}$
HHT		
HHH		

Random Variable Notation – 3 Coins

We use upper case letters to denote "unobserved" random variables, say X, and lower case letters to their observed values, in this case x.

For example, in the above example before the three coins land we denote the number of heads X, after the coins have landed we denote the number of coins x so that we can write $P(X = x)$.

The mother of all examples: Bernoulli trials!

Definition 7. Bernoulli trials satisfy the following assumptions:

(i) there are only two possible outcomes for each trial,

(ii) the probability of success is the same for each trial,

(iii) the outcomes from different trials are independent,

(iv) there are a fixed number n of Bernoulli trials conducted.

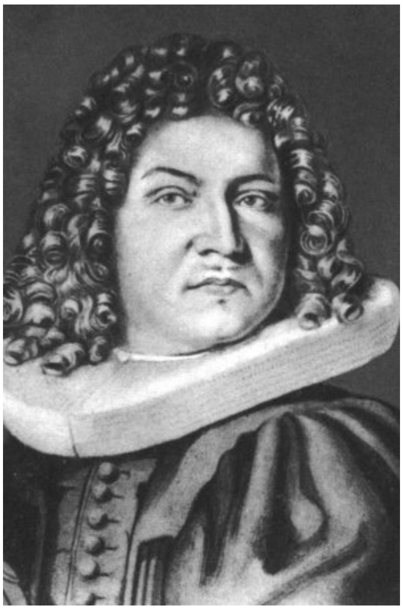
Example ($n = 1$, coin)**.** Ω: Head or Tail. We can describe the trial (before flipping the coin) in full detail. Consider a function

$$X : \{H, T\} \mapsto \{0, 1\} \quad \text{s.t.} \quad X(H) = x_H = 1 \quad \text{and} \quad X(T) = x_T = 0.$$

What is the probability that $X = x_H = 1$?

$$\mathrm{P}(X = 1) = \mathrm{P}(X = x_H) = \mathrm{P}(H) = p = 1/2 \Rightarrow \mathrm{P}(X = 0) = 1/2.$$

Jacob Bernoulli (1654–1705)



- Born in Basel (Switzerland),
- 1 of 8 mathematicians in his family,
- studied: theology $\rightarrow$ maths & astro,
- best known for Ars Conjectandi (The Art of Conjecture),
- application of probability theory to games of chance, introduction of the law of large numbers.
- Words: Bernoulli trial, Bernoulli numbers.

Tuesday, 21st August 2011

Lecture 8 - Content

- □ **Distribution of a random variable**
- □ **Binomial distribution**
- □ **Mean of a distribution**

Random Variables Reminder

Distribution of a random variable

Definition 8. The probability distribution of a integer-valued random variable X is a list of the possible values of X together with their probabilities

$$p_i = \mathrm{P}(X = i) \geq 0 \quad \text{and} \quad \sum_i p_i = 1.$$

There is nothing special with the subscript i; we could and will equally well use j, k, x etc.

Definition 9. The probability that the value of a random variable X is less than or equal to x, that is

$$F(x) = \mathrm{P}(X \leq x),$$

is called the cumulative distribution function or just the distribution function.

Also, note that for integer valued random variables that

$$P(X = x) = F(x) - F(x - 1).$$

Example ($n = 3$, IT problems). A network is fragile. By experience: $\mathrm{P}(\mathsf{F}) = 0.1 = 1-p$ that in any given week ≥ 1 major problem; $\mathrm{P}(\mathsf{S}) = 0.9 = p$ that there is none, respectively. Out of 3 weeks, how many weeks, X, had ≥ 1 problem and with what probability?

(a) All possible outcomes:

FFF SFF FSF FFS
SSF FSS SFS SSS

(b) What is the probability of each outcome? Use special multiplication rule of probability because sessions are independent!?

(c) What is the probability distribution of the number of successes, X, among the 3 sessions.

Example (cont)

$$\begin{aligned}
\mathrm{P}(X=0) &= \mathrm{P}(\mathsf{FFF}) = \mathrm{P}(\mathsf{F}) \cdot \mathrm{P}(\mathsf{F}) \cdot \mathrm{P}(\mathsf{F}) = (1-p)^3 \\
\mathrm{P}(X=1) &= \underbrace{\mathrm{P}(\mathsf{SFF} \cup \mathsf{FSF} \cup \mathsf{FFS})}_{\text{mutually exclusive events}} = \mathrm{P}(\mathsf{SFF}) + \mathrm{P}(\mathsf{FSF}) + \mathrm{P}(\mathsf{FFS}) \\
&= 3 \times (1-p)^2 p = \binom{3}{1}(1-p)^2 p, \text{ select one } S \text{ out of } 3 \text{ trials.}
\end{aligned}$$

similarly we get for $X = 2$ and $X = 3$

$$\begin{aligned}
\mathrm{P}(X=2) &= \binom{3}{2}(1-p)p^2, \text{ select two S out of } 3 \text{ trials,} \\
\mathrm{P}(X=3) &= \binom{3}{3}p^3, \text{ select three S out of } 3 \text{ trials.}
\end{aligned}$$

Binomial distribution

We can generalise this result for any $n \geq 1$ and success probability $p \in [0, 1]$.

Definition 10. The probability distribution of the number of successes $X = i$ in $n \in \mathbb{N}$ independent Bernoulli trials is called the binomial distribution,

$$p_i = \mathrm{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

The success probability of a single Bernoulli trial is p and $i = 0, 1, \ldots, n$.

To say that the random variable X has the binomial distribution with parameters n and p we write $X \sim \mathcal{B}(n, p)$.

This defines a family of probability distributions, with each member characterized by a given value of the parameter p and the number of trials n.

Binomial distribution

Since p_i, $0 \leq i \leq n$ is a probability distribution we have the identity (which we will use later on)

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1$$

for any $0 \leq p \leq 1$.

A special case of the Binomial distribution is the Bernoulli distribution where $n = 1$ and

$$\mathrm{P}(X_i = i) = p^i (1-p)^{1-i}.$$

There is another special relationship between the Bernoulli distribution and the Binomial distribution.

If $X_i \sim$ Bernoulli(p) for $1 \leq i \leq n$ and $Y = \sum_{i=1}^{n} X_i$ then

$$Y \sim \mathcal{B}(n, p).$$

Example (Dice)**.** Roll a fair dice 9 times. Let X be the probability of sixes obtained. Then $X \sim \mathcal{B}(9, 1/6)$; that is

With your table calculator or with R:

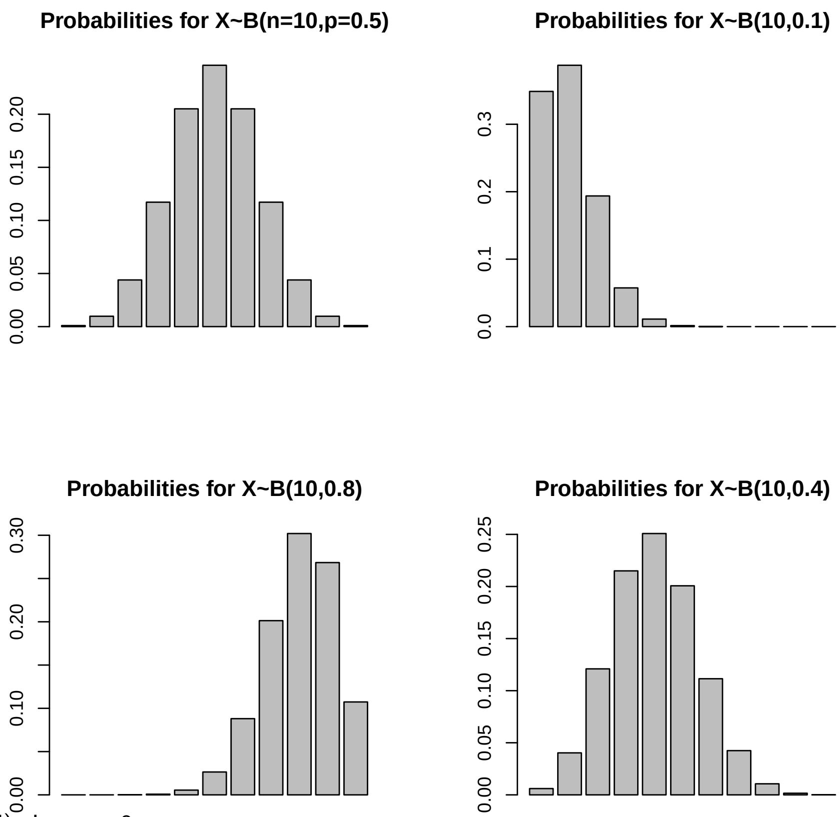
```
> n = 9;
> p = 1/6;
> round(dbinom(0:n,n,p),4) # dbinom for B(n,p) prob's
 [1] 0.1938 0.3489 0.2791 0.1302 0.0391
 [5] 0.0078 0.0010 0.0001 0.0000 0.0000
> pbinom(1,n,p) # for B(n,p) cumulative probabilities
[1] 0.5426588
```

Hence, $\mathrm{P}(X = 4) = 0.0391$ and $\mathrm{P}(X < 2) = F(1) = 0.5426588$.

Shape of the binomial distribution

- We get a binomial distribution if
  1. we are counting something over a fixed number of trials or repetitions,
  2. the trials are independent and
  3. the probability of the outcome of interest is constant across trials.
- The binomial distribution is centred at $n \times p$,
- the closer p to $1/2$ the more symmetric the distribution/histogram,
- the larger n the closer the shape to a bell (normal).

```
par(mfrow=c(2,2)); n =10 # and for n=50, 100, etc
barplot(dbinom(0:n,n,1/2))
title(main="Probabilities for X~B(n=10,p=0.5)")
barplot(dbinom(0:n,n,0.1))
title(main="Probabilities for X~B(10,0.1)")
barplot(dbinom(0:n,n,0.8))
title(main="Probabilities for X~B(10,0.8)")
barplot(dbinom(0:n,n,0.4))
title(main="Probabilities for X~B(10,0.4)")
```

Example. In a small pond there are 50 fish, 20 of which have been tagged. Seven fish are caught and X represents the number of tagged fish in the catch. Assume each fish in the pond has the same chance of being caught. Is X binomial

(a) if each fish is *returned* before the next catch?

(b) if the fish are *not returned* once they are caught?

Mean of a distribution

Definition 11. For a random variable X taking values $0, 1, 2, \ldots$ with

$$\mathrm{P}(X = i) = p_i \quad i = 0, 1, 2, \ldots$$

the mean or expected value of X is defined to be

$$\mu = \mathrm{E}(X) = \sum_i i \times p_i.$$

Interpretation of $\mathrm{E}(X)$

- ☐ Long run average of observations of X because $p_i \approx f_i/n$.
- ☐ Centre of balance of the probability density (histogram).
- ☐ Measure of location of the distribution.

Definition 12. For any function $g(X)$ we define the expected value $\mathrm{E}(g(X))$ by

$$\mathrm{E}(g(X)) = \sum_i g(i) \times p_i.$$

Expectation of a Dice Roll

Let $X = \{\text{Face showing from a dice roll}\}$ where $p_i = \mathrm{P}(X = i) = 1/6$ for $i = 1, 2, \ldots, 6$. Then

$$\begin{aligned}
\mu &= \mathrm{E}(X) \\
&= \sum_{i=1}^{6} i \times p_i \\
&= \sum_i i \times 1/6 \\
&= 3.5.
\end{aligned}$$

Note: the expected value in this case is not one of the observed values.

Mean of a distribution (cont)

Theorem 11. For constants a and b

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b.$$

Proof.

$\square$

Expectation of $X \sim \mathcal{B}(n, p)$

Theorem 12. The expectation of $X \sim \mathcal{B}(n, p)$ is $\mathrm{E}(X) = np$.

Proof.

$\square$

Example (Multiple choice section in M1905 exam is worth 35%).
20 questions and each question has 5 possible answers. A student decides to answer the questions by selecting an answer at random.

(a) What is the expected number of correct responses?

(b) Probability that the student has more than 10 correct answers?

(c) If the student scores 4 for a correct answer but -1 for a wrong response, what is his expected score?

Monday, 27th August 2012

Lecture 9 - Content

- □ **Variance of a distribution**
- □ **More integer-valued distributions**
- □ **Probability generating functions**

Expectation of a distribution – Reminders

The expectation of a distribution (or expectation of a random variable) is the mean of the probability distribution (a measure of distribution location).

Note that

- $\mathrm{E}(X) = \sum_i i \times p_i = \sum_i i \times \mathrm{P}(X = i)$ and
- $\mathrm{E}(g(X)) = \sum_i g(i) \times p_i = \sum_i g(i) \times \mathrm{P}(X = i)$.

Variance of a distribution

Example. Suppose X (e.g. number of shoes in suitcase) takes the values 2, 4 and 6 with probabilities

i	2	4	6
p_i	0.1	0.3	0.6

Hence,

What is $\mathrm{E}(X^2)$?

Suppose X (e.g. number of shoes in suitcase) takes the values 2, 4 and 6 with probabilities

i	2	4	6
p_i	0.1	0.3	0.6

What is $\mathrm{E}(X^2)$?

Solution 1: $\mathrm{E}(X^2) \overset{\text{Def}}{=} \sum g(i)p_i = \sum i^2 p_i = 26.8 \neq 5^2$.
Solution 2: $i \mapsto i^2 = j$ and $X \mapsto X^2 = Y$, use $\mathrm{E}(Y) = \sum_j j p_j$

j	4	16	36
p_j	0.1	0.3	0.6

The distribution of Y can be hard to get (e.g. for continuous rvs).

Definition 13. The variance of the random variable X is defined by

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}(X-\mu)^2 = \mathrm{E}(X^2) - \mu^2,$$

where $\mu = \mathrm{E}(X)$ and σ^2 is also a measure of spread.

This is like the large sample limit of a sample variance.
The standard deviation of X is $\sigma = \sqrt{\sigma^2}$.

Variance of a Linear Transformation

Theorem 13. For any constants a and b

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

Proof.

□

Example. If $X \sim \mathcal{B}(n, p)$ then we'll show later that $\mathrm{Var}(X) = n \times p \times (1 - p)$.

- Hence, if $p = 0$ or 1 then the variance is 0.
- the variance is largest when $p = 0.5$ and in this case it is $\sigma^2 = n/4$.

More integer-valued distributions

Geometric distribution

The binomial random variable is just one possible integer-valued random variable. Suppose we have an infinite sequence of independent trials, each of which gives a success with probability p and failure with probability $q = 1 - p$.

Definition 14. The geometric distribution with parameter p $(=$ success prob.$)$ has probabilities for the number of failures X before the first success,

$$p_i = \mathrm{P}(X = i) = q^i p, \quad i = 0, 1, 2, \ldots.$$

Note the probabilities add to 1:

$$\mathrm{P}(X = 0) + p_1 + \ldots = p + qp + q^2 p + \ldots = p(1 + q + q^2 + \ldots) = p\left(\frac{1}{1-q}\right) = 1$$

Example. A fair die is thrown repeatedly until it shows a six.

(a) What is the probability that more than 7 throws are required?

```
1-pgeom(7,1/6)        1-sum(dgeom(0:7,1/6))
```

(b) Is it more likely that an odd number of throws is required or an even number?

The Poisson approximation to the Binomial

The Poisson distribution often serves as a first theoretical model for counts which do not have a natural upper bound.

Possible examples

- modeling of number of accidents, crashes, breakdowns
- modeling radioactivity measured by the Geiger counter
- modeling of so-called rare events (meteorite impacts, heart attacks)

The Poisson distribution can be seen as the limiting distribution of $\mathcal{B}(n,p)$:

Approximation is good if $n \cdot p^2$ is small!

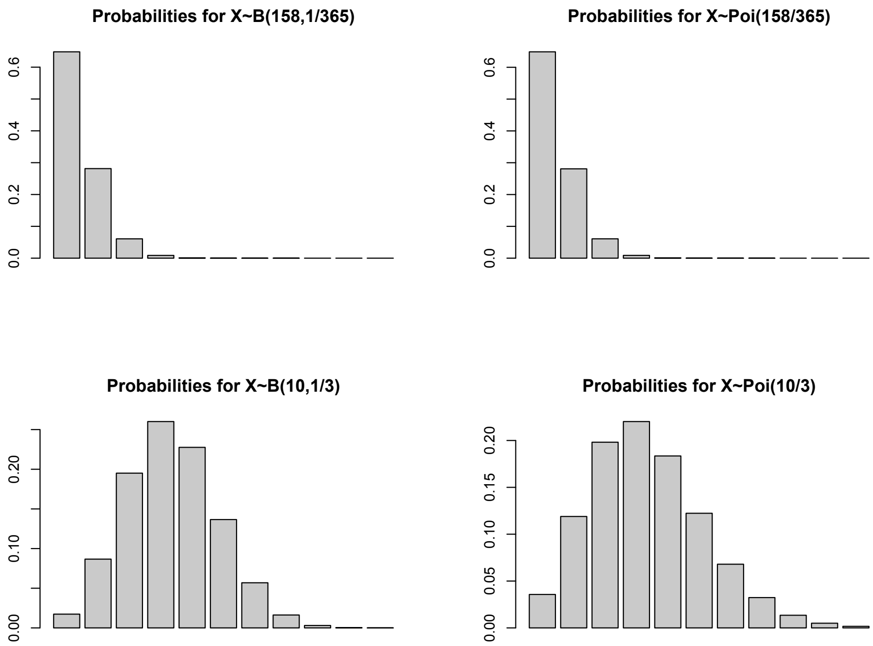
$X \sim \mathcal{B}(158, \frac{1}{365})$ and $n \cdot p^2 = 0.001186$:

```
> # What is the probability that of 158 people, exactly k have a birthday today?
> n = 158; p=1/365;
> round(dbinom(0:7,n,p),5);
[1] 0.64826 0.28139 0.06068 0.00867 0.00092 0.00008 0.00001 0.00000
> round(dpois(0:7,p*n),5);
[1] 0.64864 0.28078 0.06077 0.00877 0.00095 0.00008 0.00001 0.00000
```

But for $n = 10$

```
> n = 10; p=1/3;
> round(dbinom(0:4,n,p),5);
[1] 0.01734 0.08671 0.19509 0.26012 0.22761
> round(dpois(0:4,p*n),5);
[1] 0.03567 0.11891 0.19819 0.22021 0.18351
```


Probability distribution for $X \sim \mathcal{B}(n,p)$ and $X \sim \mathcal{P}(\lambda)$



Probability generating functions

Let $X \in \mathbb{N}$ and $p_i = \mathrm{P}(X = i)$, $i = 0, 1, 2, \ldots$

Definition 15. The probability generating function is defined as

$$\pi(s) = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \ldots.$$

Example. If X only takes a finite number of values (e.g. $X \sim \mathcal{B}(n, p)$) then $\pi(s)$ is a polynomial.

Alternatively (e.g. $X \sim \mathcal{P}(\lambda)$) $\pi(s)$ is a power series.

Properties of $\pi(s)$

Let $s \in [0, 1]$ then

- $0 \leq \pi(s) \leq 1$,
- $\pi(1) = p_0 + p_1 + \ldots = 1$,
- $\pi'(s) = p_1 + 2p_2 s + 3p_3 s^2 + \ldots \geq 0, \quad s \geq 0.$
- $\pi'(1) = p_1 + 2p_2 + 3p_3 + \ldots = \mathrm{E}(X)$ (if $\mathrm{E}(X)$ is finite),
- $\pi''(s) = 2p_2 + 6p_3 + 4 \cdot 3p_4 + \ldots$ at $s = 1$, so $\pi''(1) = \mathrm{E}(X(X-1))$ and

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2 = \pi''(1) + \pi'(1) - (\pi'(1))^2.$$

Example (Poisson distribution).

Example (Binomial distribution).

Tuesday, 28th August 2012

Lecture 10 - Content

- □ **Continuous random variables**
- □ **Chebyshev's inequality**

References from Phipps & Quine

- □ Section 2.2 pages 62-66.

Answer to Challenge Question

Show that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Let n be an integer. Then by the Binomial Theorem

$$(x+y)^n = \sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i}.$$

Let $y=1$ and $x=-\frac{\lambda}{n}$ then

$$\begin{aligned}\lim_{n\to\infty}\left(1-\tfrac{\lambda}{n}\right)^n &= \lim_{n\to\infty}\textstyle\sum_{i=0}^{n}\frac{n\times(n-1)\times\ldots\times(n-i+1)}{n^i}(-1)^i\frac{\lambda^i}{i!}\\ &= \lim_{n\to\infty}\textstyle\sum_{i=0}^{n}(-1)^i\frac{\lambda^i}{i!}\end{aligned}$$

since

$$\lim_{n\to\infty}\frac{n\times(n-1)\times\ldots\times(n-i+1)}{n^i}=1.$$

The last line is the Taylor series expansion for $e^{-\lambda}$.

Continuous random variables

Examples

Continuous random variables have images in $\mathbb{R}$, e.g.

- the speed of a car,
- the amount of alcohol in a person's blood after 4 standard drinks,
- the temperature at 1pm.

Distribution Function of a Continuous Random Variable

Definition 16. A distribution function, $F(x) = \mathrm{P}(X \leq x)$, is any function that satisfies

(i) $0 \leq F(x) \leq 1$ $\quad$ (F is a probability)

(ii) $F(x) \uparrow$, i.e. $F(x)$ is a monotonic increasing function of x.

(iii) If $a < b$ then $\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F(b) - F(a)$.

(iv) $F(-\infty) = 0$, $F(+\infty) = 1$.

(v) $F(x)$ is right-continuous; i.e. for every number x^*, $\lim_{x \downarrow x^*} F(x) = F(x^*)$.

Key Property of Continuous Random Variables

Theorem 14. A continuous random variable X attains with probability zero any value of its image. That is

$$\mathrm{P}(X = x) = 0$$

for all real numbers $x \in \mathbb{R}$.

Proof. Note that the set $A = \{X = x\}$ is a subset of $B = \{x - \epsilon < X \leq x\}$ for any $\epsilon > 0$. Since, if $A \subset B$ then $\mathrm{P}(A) \leq P(B)$ we have

$$0 \leq \mathrm{P}(X = x) \leq P(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon).$$

Due to the continuity of F we have

$$0 \leq \mathrm{P}(X = x) \leq \lim_{\epsilon \downarrow 0} F(x) - F(x - \epsilon) = 0.$$

□

Hence, if X is a continuous random variable then,

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a < X < b).$$

Probabilities for Continuous Random Variables

- Suppose that we focus on events $X \in (a, b]$, i.e. $(a, b]$ an interval of length $b - a > 0$.
- Dividing $(a, b]$ into n equal subintervals of width Δx; it follows that

$$\mathrm{P}(a < X \le b) = \lim_{n\to\infty} \sum_{i=1}^{n} \widehat{f}(i; \Delta x) \times \Delta x.$$

where

$$\begin{aligned}\widehat{f}(i; \Delta x) &= \frac{\mathrm{P}(a + (i-1)\Delta x < X \le a + i\Delta x)}{\Delta x} \\ &= \frac{F(a + i\Delta x) - F(a + (i-1)\Delta x)}{\Delta x}\end{aligned}$$

for $i = 1, \ldots, n$.

- Consider any sequence $i = i(n)$ such that

$$\lim_{n\to\infty} (a + i\Delta x) = x$$

for some $x \in (a, b]$ and let $f \ge 0$ be an integrable function in $\mathbb{R}$ such that

$$f(x) = \lim_{n\to\infty} \widehat{f}(i; \Delta x).$$

- Then

$$f(x) = \lim_{\Delta x\to 0} \frac{F(x) - F(x - \Delta x)}{\Delta x} = \frac{dF(x)}{dx}$$

and

$$\mathrm{P}(a < X \le b) = \int_a^b \underbrace{f(x)}_{=\text{ probability density function}} dx.$$

Probability density function

Definition 17. A probability density function or simply a probability density is any non-negative function $f(x) \geq 0$ such that

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Theorem 15. $F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$

As an immediate consequence of the Fundamental Theorem of Calculus,

$$f(x) = \frac{dF(x)}{dx}$$

as previously stated.

Indicator Functions

The following type of function appears quite frequently in Statistics when defining probability density functions.

Definition 18. The function $\mathbf{1}_A(x) = \mathbf{1}\{x \in A\}$ is called the indicator function of the set A. It has image 1 if $x \in A$ and image 0 if $x \notin A$.

(Although we have not yet defined the expectation of a continuous random variable it turns out that

$$E[\mathbf{1}_A(x)] = \mathrm{P}(A)$$

which is a useful property in certain contexts.)

Scaling of non-negative functions to construct density functions

Example. Find c s.t. the following non-negative function is a probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \le 0 \\ cxe^{-4x^2} & \text{for } x > 0 \end{cases} = cxe^{-4x^2} \times \mathbf{1}_{(0,\infty)}(x).$$

Moments of continuous random variables

Definition 19. Let g be any continuous function. The expected value of $g(X)$ of a continuous random variable X having probability density f is defined by

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

The mean of X is given by

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

The kth moment about the mean of X is given by

$$\mu_k = \mathrm{E}[(X-\mu)^k] = \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx.$$

The variance of X is given by

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

Useful Results

The following results, which we showed hold for integer values random variables, also hold for continuous random variables:

$$E(aX + b) = aE(X) + b$$

$$\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X).$$

Proof: Left as an exercise.

Uniform distribution

Definition 20. The uniform distribution, with parameters a and b, has

$$f(x) = \frac{1}{b-a} \cdot \mathbf{1}_{(a,b)}(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{elsewhere;} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b. \end{cases}$$

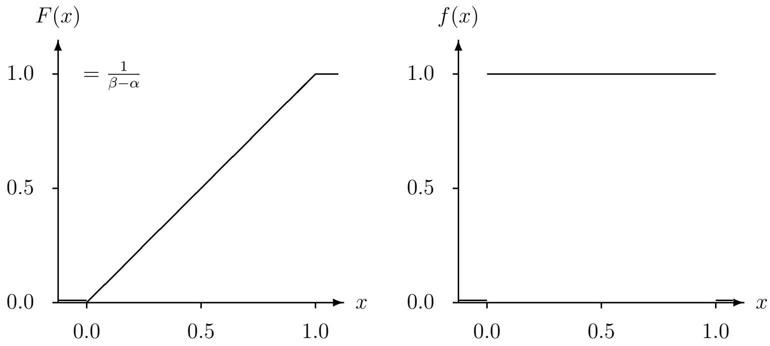
Short notation:

$$X \sim \mathcal{U}(a, b)$$

[The uniform distribution is potentially useful to model or to be applied in conjunction with rounding errors/effects, generating random variables, simulation studies.]

Uniform distribution

Example (Uniform distribution for $a = 0$ and $b = 1$).



Uniform distribution – Expectation and Variance

Theorem 16. If $X \sim \mathcal{U}(a, b)$ then,

$$\mu = \mathrm{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \mathrm{E}[(X-\mu)^2] = \frac{1}{12}(b-a)^2.$$

Proof.

Uniform distribution – R code

```
> n = 10000
> set.seed(1)
> x = runif(n)  # Generates Uniform(0,1) values
> hist(x)
> mean(x) # We should expect this value to be close to (0 + 1)/2
[1] 0.4990762
> var(x)  # We should expect this value to be close to (1 - 0)^2/12 = 1/12
[1] 0.08383338
> 1/12
[1] 0.08333333
```


Chebyshev's inequality

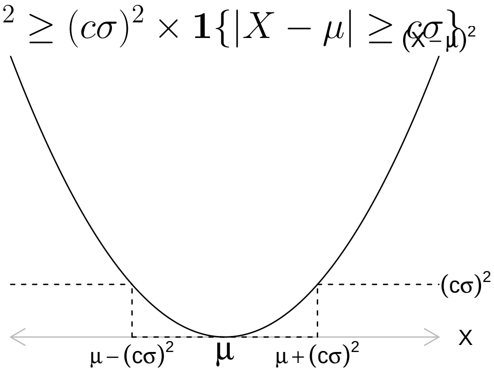
Links the three notions of probability, mean and variance.

Theorem 17. If a random variable X has mean μ and variance σ^2, then for any positive number c,

$$\mathrm{P}(|X-\mu| \geq c\sigma) \leq 1/c^2.$$

Proof. Note that

$$|X-\mu|^2 \geq (c\sigma)^2 \times \mathbf{1}\{|X-\mu| \geq c\sigma\} \quad (*)$$



From the definition of the expected value and the indicator function we have

$$\mathrm{E}[\mathbf{1}_A(X)] = \int_A f(x)dx = \mathrm{P}(A).$$

Hence, taking expectations on both sides of $(*)$ yields

$$\mathrm{E}[|X-\mu|^2] = \sigma^2 \geq (c\sigma)^2 \,\mathrm{P}(|X-\mu| \geq c\sigma).$$

$\square$

Examples

Example. Consider the IQ score where $\mu = E(X) = 100$ and $\sigma^2 = \mathrm{Var}(X) = 10^2$. What can we say about $P(X > 150)$?

Examples

Example. Suppose that $X \sim \mathcal{U}(0, 10)$. Use Chebyshev's inequality to bound the probability $\mathrm{P}(|X - 5| > 4)$.

Monday, 3rd September 2012

Lecture 11 - Content

- **Normal random variables**
- **Standardized random variables**
- **Pseudo-random numbers in R**

References from Phipps & Quine

- Section 2.3 pages 66-69.

Normal random variables

- The normal distribution or the normal probability density dates back to the 18th century.
- Abraham de Moivre (1667–1754) and Pierre-Simon Marquis de Laplace (1749–1827) find the normal distribution as an approximate distribution to the Binomial.
- Johann Carl Friedrich Gauss (1777–1852) assumed the normal distribution of errors in the context of the least squares method.

Alternative names for the normal

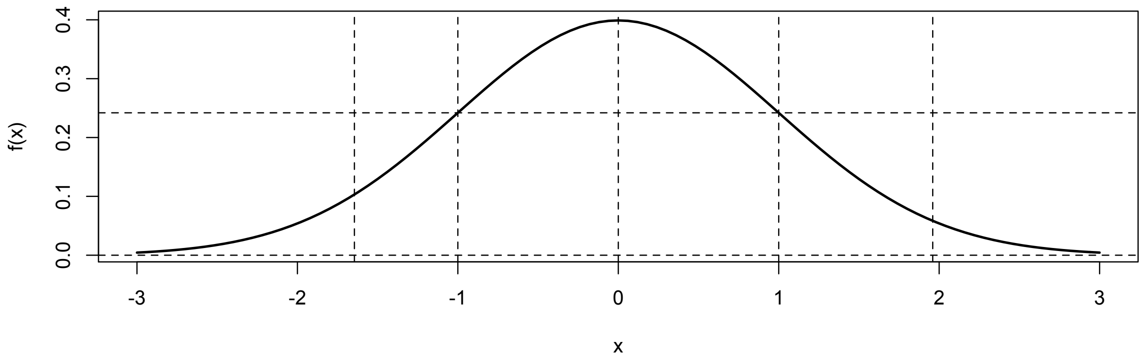
- Gaussian distribution,
- Bell distribution.

Normal probability density

Definition 21. The normal probability density is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \Rightarrow X \sim \mathcal{N}(\mu,\sigma^2).$$

It has location parameter $\mu = \mathrm{E}(X)$ and scale parameter $\sigma^2 = \mathrm{Var}(X)$.



Some useful facts

- The density function of the normal distribution has the shape of a symmetric bell curve.
- Its maximum is at $x = \mu$ and it has inflection points at

Why is the normal distribution so famous?

- If $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow$ simple results and theorems!
- Central limit theorem: the mean of many independent random variables $X_1, X_2, \ldots$ (having finite variances) is approximately normally distributed

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right) \approx \mathcal{N}(0, 1).$$

- The distribution of measurement errors is often very similar to the normal distribution

Standard normal random variable

Definition 22. The normal with mean 0 and variance 1 is called the standard normal random variable and is generally denoted by Z. Thus

$$Z \sim \mathcal{N}(0, 1)$$

with

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty.$$

Remark

The integral $\int_{-\infty}^{\infty} e^{-z^2} dz$ is called the Euler-Poisson integral and equals $\sqrt{\pi}$. See additional slides at the end of this lecture.

Normal distribution function

The normal distribution function $F(x;\mu,\sigma^2)$ has no closed form, thus

$$\begin{aligned} F(x;\mu,\sigma^2) &= \mathrm{P}(X \le x) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt. \end{aligned}$$

In practice the normal distribution function needs to be approximated numerically.

There are several nice ways of doing this, but they rely on transforming the integral into "standard form".

Standardised random variables

Theorem 18. Let $X \sim \mathcal{N}(\mu,\sigma^2)$, then the centred and standardised random variable

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$
$$\Rightarrow \mathrm{P}(Z \le z) := \Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}\,dt.$$

Proof. The proof is left as an exercise. Begin with $F(x;\mu,\sigma^2)$, substitute $Z = g(X) = \frac{X-\mu}{\sigma}$, continue with calculus knowledge till you get $F(z;0,1)$. $\square$

Thanks to the theorem it is sufficient to know the (tabulated) probabilities of the standard normal distribution e.g. from the formula sheet, software, or any other source.

Standardizing random variables

Definition 23. If X is any random variable with mean μ and variance σ^2 then

$$Z = \left(\frac{X-\mu}{\sigma}\right)$$

is called the standardized version of X.

Theorem 19. If $Z = \left(\frac{X-\mu}{\sigma}\right)$ with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ then,

$$\mathrm{E}(Z) = 0 \quad \text{and} \quad \mathrm{Var}(Y) = 1.$$

Proof. Follows from the definitions of E and Var and from the identity

$$\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X).$$

□

Useful identities for the normal

- $\phi(-z) = \phi(z)$ because of symmetry of ϕ.
- $\Phi(-z) = 1 - \Phi(z)$ because of symmetry of $\phi \geq 0$ and $\int \phi(t)dt = 1$.
- $\mathrm{P}(|Z| \leq z) = 2\Phi(z) - 1$ because

$$\mathrm{P}(|Z| \leq z) = \mathrm{P}(-z \leq Z \leq z) = \Phi(z) - \Phi(-z).$$

Example. $X \sim \mathcal{N}(3, 2^2)$. Find $\mathrm{P}(X \leq 4)$ and $\mathrm{P}(X < 1.24)$.

Example. $X \sim \mathcal{N}(5, 3^2)$. Find c such that $\mathrm{P}(X > c) = 0.1$.

Exponential distribution and friends

Definition 24. The exponential distribution, with parameter λ, has probability density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0,\ \lambda > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and distribution function given by

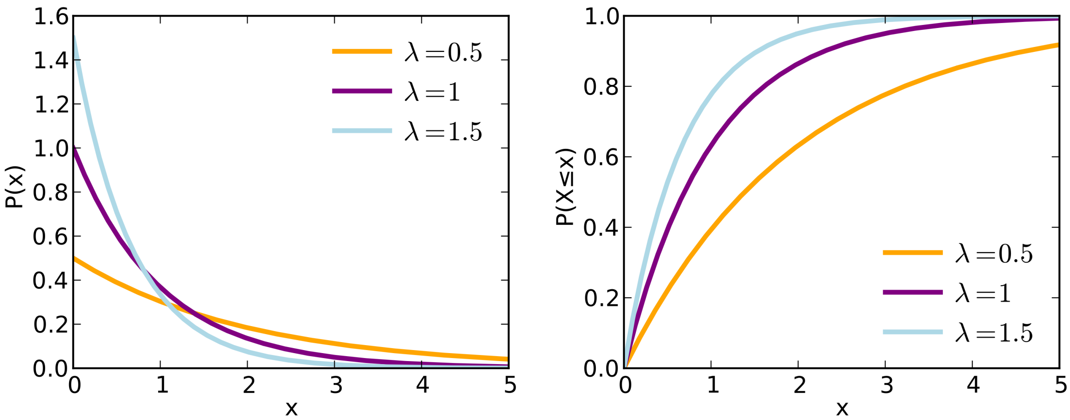
$$F(x) = 1 - e^{-\lambda x} = \int_0^x \lambda e^{-\lambda t} dt \quad t > 0.$$

To say that the random variable X has the exponential distribution with parameter $\lambda > 0$ we write

$$X \sim \mathcal{E}(\lambda).$$

Sometimes an alternative parameterisation is used where $\beta = 1/\lambda$ becomes the parameter of the distribution.

Plots of the Exponential Distribution



Properties and applications

- The mean and variance of $X \sim \mathcal{E}(\lambda)$ equals $\mathrm{E}(X) = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.
- Waiting times X between two events, failure distribution with underlying constant failure rate, distance between roadkill on a street etc are often modelled by $X \sim \mathcal{E}(\lambda)$.
- The exponential distribution is memoryless

  $$\mathrm{P}(X > t + h) = \mathrm{P}(X > t)\,\mathrm{P}(X > h), \quad t, h > 0,$$

  and therefore

  $$\mathrm{P}(X > t + h | X > t) = \mathrm{P}(X > h),\ t, h > 0$$

Example – Exponential Distribution

Example. Suppose that the amount of time one spends in a bank is exponentially distributed with mean 10 minutes, $\lambda = 1/10$. What is the probability that a customer will spend more than 15 minutes in the bank? What is the probability that a customer will spend more than 15 minutes in the bank given that he is still in the bank after 10 minutes?

Pseudo-random numbers in R

```
> # generating samples of 'independent' continous 'random' variables
> set.seed(010909)  # set random seed to 01 Sep 09
> n = 10            # choose sample size of 10
> rnorm(n)          # 10 pseudo-standard-normal random numbers
 [1] -1.6657 -0.1583 -0.2662 -0.9809 -1.0117 -1.2175  0.0986  0.7802  2.3596 -0.3192
> runif(10)         #       ...-uniform [0,1]
 [1] 0.1784 0.8924 0.7842 0.4014 0.7271 0.2366 0.1984 0.0003 0.7880 0.8027
> rexp(10)          #       ...-exponential with mean 1
 [1] 0.5629 0.4597 0.1792 0.5607 0.5740 0.7506 2.4387 0.7580 0.2380 0.0726
> # hence the r... in front of norm, unif, exp signifies drawing random numbers
> # the d signifies density, the p = P(X <= x), the q returns the quantile.
> curve(dnorm,from=-3,to=3)
> pnorm(95,mean=100,sd=10) # qnorm(0.3085375,mean=100,sd=10) = 95
[1] 0.3085375
> 1-pnorm(95,mean=100,sd=10)
[1] 0.6914625
> pnorm(95,mean=100,sd=10,lower.tail = FALSE)
[1] 0.6914625
```


The Gamma Distribution

A generalisation of the exponential distribution leads to the family of gamma distributions.

Definition 25. The gamma distribution, with parameters α and β, has probability density

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x, \alpha, \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

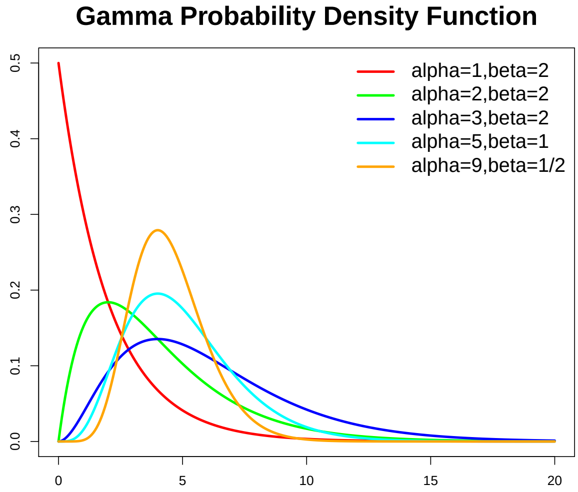
If $f(x)$ has the above density then we write $X \sim \text{Gamma}(\alpha, \beta)$.

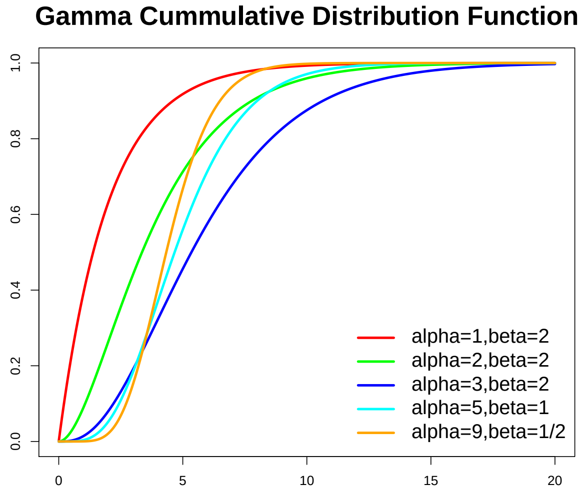
Note that $\Gamma(\alpha)$ is a value of the gamma function, defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

It is a generalisation of $n!$, $n \in \mathbb{N}$.

Plots of the Gamma Distribution





Friends of the Gamma Distribution

Depending on special choices of the parameters α and β the gamma distribution becomes

- for $\alpha = 1$ the exponential distribution (with $\beta = 1/\lambda$),
- for $\alpha = 1/2$ and $\beta^{-1} = \sigma^2/2$ the distribution of $Y = X^2$, if $X \sim \mathcal{N}(0, \sigma^2)$,
- for $\alpha = m/2$, $m \in \mathbb{N}$, and $\beta = 2$ the chi-square distribution.

Properties of the Gamma Distribution

Theorem. If $X \sim \text{Gamma}(\alpha, \beta)$ then, the mean and variance of X equals

$$\mu = E[X] = \alpha\beta \quad \text{and} \quad \sigma^2 = E[(X-\mu)^2] = \alpha\beta^2.$$

Proof. For the mean we have (proof for variance is similar):

$$\mu = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x \cdot x^{\alpha-1} e^{-x/\beta} dx$$

$$\overset{y=x/\beta}{\Rightarrow} \mu = \frac{\beta}{\Gamma(\alpha)} \underbrace{\int_0^\infty y^\alpha e^{-y} dy}_{=\Gamma(\alpha+1)=\alpha\Gamma(\alpha)} = \alpha\beta$$

□

The Gamma Function

Let

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Using integration by parts shows that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for any $\alpha > 1$.
Remember: Integration by parts: $\int fG = FG - \int Fg$
For $f(x) = e^{-x}$ and $G(x) = x^\alpha$ it follows:

$$\begin{aligned}
\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x} dx = [-x^\alpha e^{-x}]_0^\infty - \int_0^\infty \alpha x^{\alpha-1}(-1)e^{-x} dx \\
&= -\lim_{x\to\infty} x^\alpha e^{-x} + \underbrace{0^\alpha e^{-0}}_{=0,\text{ since } \alpha>0} + \alpha \underbrace{\int_0^\infty x^{\alpha-1} e^{-x} dx}_{=\Gamma(\alpha)} \\
&= -\lim_{x\to\infty} \left(x^{-\alpha} \sum_{k=0}^\infty \frac{x^k}{k!} \right)^{-1} + \alpha \cdot \Gamma(\alpha) = \alpha \cdot \Gamma(\alpha)
\end{aligned}$$

Now we have the proof for $\Gamma(\alpha+1) = \alpha!$, $(\alpha \in \{1, 2, 3, \ldots\})$ if and only if $\Gamma(1+1) = 1\Gamma(1) = 1! = 1$.

That is easy:

$$\Gamma(1+1) = \int_0^\infty x e^{-x} dx = \int_0^\infty e^{-x} dx = [-e^{-x}]_0^\infty = 1.$$

On the Euler-Poisson or Gaussian integral

The Euler-Poisson integral is the improper integral

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

and exists because $\exp(-x^2)$ is continuous and bounded, i.e. $0 \leq e^{-x^2} < e^{-|x|+1}$ noting that $\int_{-\infty}^{\infty} e^{-|x|+1} dx = 2e$.

Instead of calculating I we show that

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right) \times \left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2} dxdy = \pi.$$

For any point $(x, y) \in \mathbb{R}^2$ we have the alternative coordinate notation $x = r\cos\theta$ and $y = r\sin\theta$. Hence,

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} \times |J| dr d\theta,$$

where $|J|$ denotes the determinant of the Jacobi matrix, i.e. matrix of partial derivatives:

$$J = \begin{pmatrix} \partial x/\partial r & \partial y \partial r \\ \partial x/\partial \theta & \partial y/\partial \theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \Rightarrow |J| = r(\cos^2\theta + \sin^2\theta) = r.$$

By substituting $r^2 = u$ we obtain,

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\, dr d\theta = \int_0^{2\pi}\int_0^{\infty} \frac{1}{2} e^{-u}\, du d\theta = \int_0^{2\pi} \frac{1}{2}\, d\theta = \pi.$$

Tuesday, 4th September 2012

Lecture 12 - Content

- ☐ **Joint distributions**
- ☐ **Independent random variables**
- ☐ **Central limit theorem**

References from Phipps & Quine

- ☐ Section 2.4 pages 69-72.
- ☐ Section 3.2 pages 73-75.

Standard Normal Distribution

Let $Z \sim N(0, 1)$ then

- ☐ The probability density function at z is given by

```
> dnorm(z)
```

- ☐ The (cumulative) distribution function at z, $\Phi(z) = \mathrm{P}(Z < z)$, is given by

```
> pnorm(z)
```

- ☐ The inverse (cumulative) distribution function at t, $\Phi^{-1}(t)$ or the value of z such that $\Phi(z) = t$, is given by

```
> qnorm(t)
```

- ☐ To generate n random values from $Z \sim N(0, 1)$ we use

```
> rnorm(n)
```


Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ then

- ☐ The probability density function at x is given by

```
> dnorm(x,mu,sigma)
```

- ☐ The (cumulative) distribution function at x, $\Phi((x - \mu)/\sigma) = \mathrm{P}(X < x)$, is given by

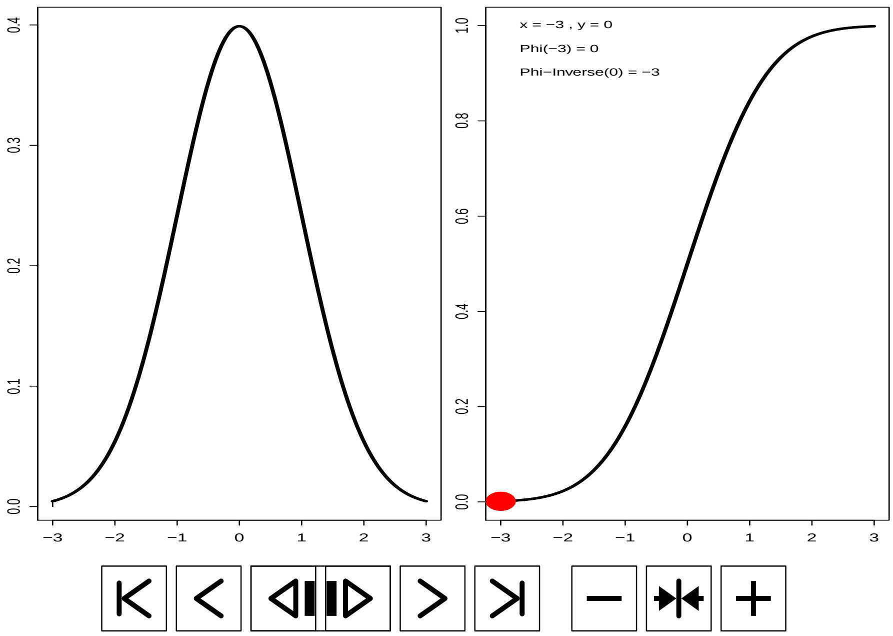
```
> pnorm(x,mu,sigma) # OR
> pnorm( (x-mu)/sigma )
```

- ☐ The inverse (cumulative) distribution function at t, the value of x such that $\mathrm{P}(X < x) = t$, is given by

```
> qnorm(t,mu,sigma)
```

- ☐ To generate n random values from $X \sim N(\mu, \sigma^2)$ we use

```
> rnorm(n,mu,sigma)
```


x = −3 , y = 0
Phi(−3) = 0
Phi−Inverse(0) = −3

Joint distributions

Independence of random variables

Let X be a real-valued random variable (e.g. normal, exponential, binomial) and $x \in \mathbb{R}$ any number, then

$$A = \{X \leq x\}$$

represents an event. Let Y be another real-valued random variable and

$$B = \{Y \leq y\}, \quad y \in \mathbb{R}.$$

Recall the definition of independence of events: A and B are independent iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

which is a special case of the general multiplication rule,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) = \mathrm{P}(B)\,\mathrm{P}(A|B) \quad \text{if} \quad \mathrm{P}(A), \mathrm{P}(B) \neq 0.$$

Definition 26. Two random variables X and Y are independent if and only if for any numbers x and y the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events.

Example.

- ☐ (X = 'height', Y = 'weight') from a random person are not independent.
- ☐ X_1 = 'lottery numbers next draw' and X_2 = 'lottery numbers in three weeks time' are
- ☐ X_1 = 'todays rainfall' and X_2 = 'tomorrows rainfall' are

From the above Definition 26 we easily get the joint cumulative distribution function and joint probability density function of independent random variables.

Joint distribution functions and densities

Definition 27. The joint cumulative distribution function of two random variables X and Y is

$$F_{X,Y}(x,y) := \mathrm{P}(X \leq x, Y \leq y)$$

and the joint density function is denoted $f_{X,Y}(x,y)$.

Note that, if X and Y are continuous random variables, then $F_{X,Y}(x,y)$ and $f_{X,Y}(x,y)$ are related via

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

and

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t) dt ds.$$

Marginal distribution functions and densities

Definition 28. If $F_{X,Y}(x,y)$ is the joint cumulative distribution function of two random variables X and Y then, $F_X(x)$ and $F_Y(y)$ are called the marginal cumulative distribution functions of X and Y, respectively.

For integer valued random variables the marginal probability mass functions can be calculated via

$$P(X=x)=\sum_y P(X=x,Y=y) \qquad \text{and} \qquad P(Y=y)=\sum_x P(X=x,Y=y)$$

while for continuous random variables the marginal density functions can be calculated via

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \qquad \text{and} \qquad f_Y(y)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

From these the marginal cumulative distribution functions can be calculated in the usual way.

Expectations of Joint Distributions

Let $g(x,y)$ be a bivariate function and let X and Y be random variables with joint density function $f_{X,Y}(x,y)$.

If X and Y are discrete random variables then

$$E[g(X,Y)]=\sum_x\sum_y g(x,y)f_{X,Y}(x,y).$$

If X and Y are continuous random variables then

$$E[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy.$$

Independence

Definition 29. Let $F_X(x) = \mathrm{P}(X \leq x)$ and $F_Y(y) = \mathrm{P}(Y \leq y)$ be the cumulative distribution functions of the independent random variables X and Y then, the joint cumulative distribution function is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\,\mathrm{P}(Y \leq y) = F_X(x)F_Y(y).$$

Definition 30. Let $f_X(x)$ and $f_Y(y)$ be the probability density functions of the independent random variables X and Y then, the joint probability density function is given by

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Independent random variables: rules for expectations and variances

Theorem 20 (Properties of E and Var)**.** Let X and Y be random variables then

1. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
2. if X and Y are independent then, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
3. if X and Y are independent then, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note that for any two, not necessarily independent, random variables

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

where

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))].$$

Proof of 1. discrete case only:

Central limit theorem

Many observed phenomena can be modelled as the sum of several random variables:

- □ total weight of passengers in a lift,
- □ total of available funds

or means of random variables

- □ average class mark,
- □ average height and weight,
- □ average temperature in Sydney.

The central limit theorem is useful in these types of situations.

Some useful facts about the normal distribution

Theorem 21. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, let X and Y be independent and let a and b be two real numbers. Then

$$Z = aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2).$$

Proof: Not in MATH 1905.

In general, let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent and a_i be real numbers for $1 \leq i \leq n$ then

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, a_i^2\sigma_i^2\right).$$

Example

Example (Mean and variance of the sample mean $\overline{X}$)**.** Let the n random variables $X_1, X_2, \ldots, X_n$ be pairwise independent and each have the same distribution with mean μ and variance σ^2. Then for the sample mean, that is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have,

i) mean: $\mu_{\overline{X}} = E(\overline{X}) = \mu$

ii) variance: $\sigma^2_{\overline{X}} = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

Sums of normal random variables

Theorem 22. If all $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then,

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

[This is only true for normal rvs; in STAT2911 moment generating functions are introduced that make a simple proof available].

Example. X_1, X_2, X_3 are independent random variables with

i	0	1	3
p_i	$1/3$	$1/3$	$1/3$

$\overset{T_2=X_1+X_2}{\Longrightarrow}$

i	0	1	2	3	4	6
p_i	$1/9$	$2/9$	$1/9$	$2/9$	$2/9$	$1/9$

i	0	1	2	3	4	5	6	7	9
p_i	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

(Note, the distribution of T_3 clusters around the mean $\mathrm{E}\, T_3 = 4$.)

Theorem 23 (CLT, central limit theorem)**.** If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance $0 < \sigma^2 < \infty$ then,

$$\mathrm{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) = \mathrm{P}(Z \leq x) \quad \text{as} \quad n \to \infty.$$

Proof. Postponed to second year... □

Thus for n large (here $n \geq 25$) the following are approximately true:

$$\begin{aligned} T &= \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \\ \overline{X} &= \frac{1}{n} T \sim \mathcal{N}(\mu, \sigma^2/n). \end{aligned}$$

The closer the distribution of X_i is to the normal the better the approximation for small n values.

Example (PQ, p71). Steel rods, made with diameter $R \sim \mathcal{N}(4.90, 0.03^2)$ (in cm), are to fit into sockets, made with diameter $S \sim \mathcal{N}(5.00, 0.04^2)$ (in cm). For a satisfactory fit the socket diameter should exceed the rod diameter, but by no more than 0.20 cm. If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory?

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) What is the probability that the average length of 25 independent tibia lengths will be less than 7.6 mm?

Solution (i):

Because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(ii) What is the prob. that the average will differ from 7.8 by more than 0.1?

Solution (ii):

Note we can show that

$$\mathrm{P}(|L - 7.8| > 0.1) = 0.7414$$

so the average varies much less than the individual measurements.

Again, because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

(i) What is the probability that the average BP is greater than 125 mm hg?

Solution (i):

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

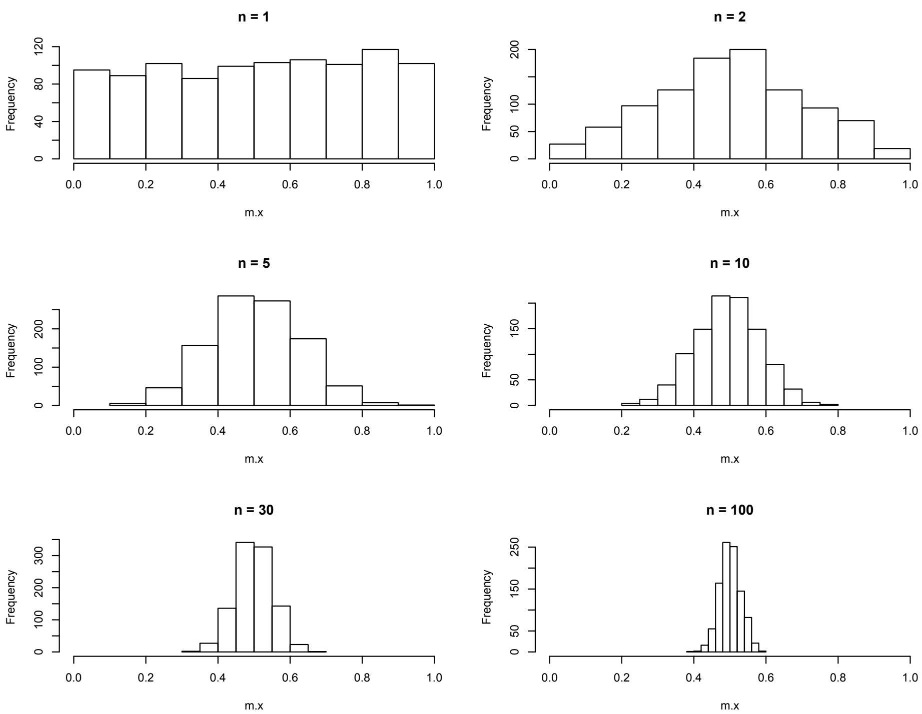
(ii) If the sample size increases to 40 how changes the answer to (i)?

Solution (ii):

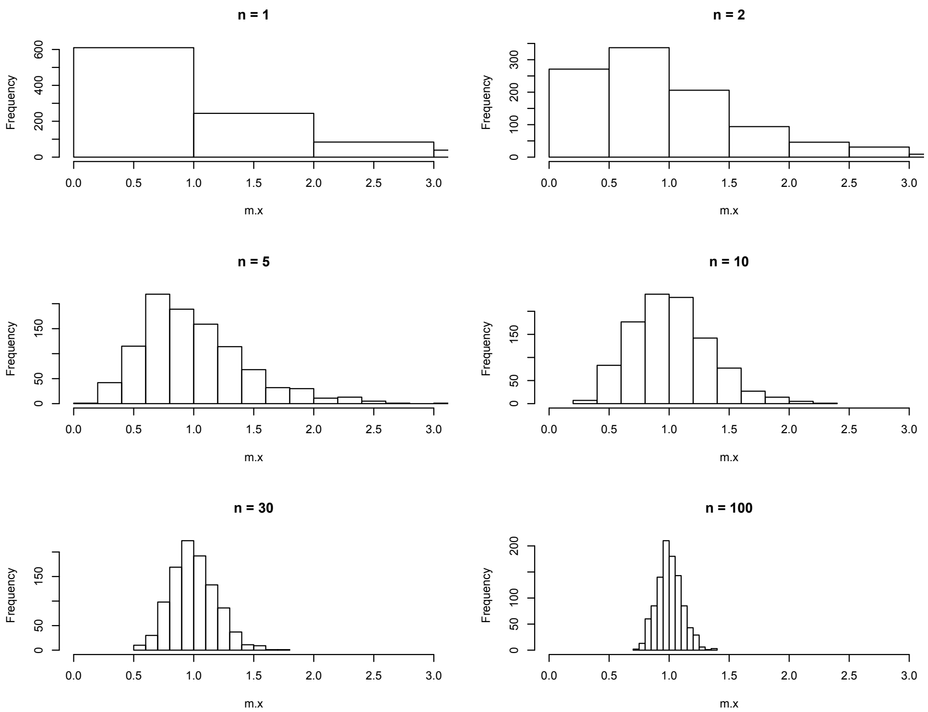
CLT in R

```
# Uniform distribution and CLT
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = runif(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,1),main="n = 1")


# Exponential distribution and CLT
par(mfrow=c(3,2))
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = rexp(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,3),main="n = 1")

# choose n in {1,2,5,10,30,100}

qqnorm(m.x)          # produces a Q-Q-plot
plot(density(m.x))   # produces a smooth estimated density
```


CLT for $X \sim \mathcal{U}$



CLT for $X \sim \mathcal{E}$



More on functions of random variables

In many applications interest focuses on some function $g(X)$ of the random variable X. E.g. change scale from meters to millimeters, logarithm of daily exchange rate changes or squared body height (BMI). In STAT2911 you will learn more. In the following I show some results that you are likely to understand with what you already know.

Theorem. (A simple version of the transformation theorem for densities) Let the random variable X have probability density f_X and let $g : \mathbb{R} \to \mathbb{R}$ be some monotone function and $h = g^{-1}$ be the inverse of g with

$$\frac{\partial h(y)}{\partial y} = h'(y).$$

Then, the density function of $Y = g(X)$ is given by $f_Y(y) = f_X(h(y)) \cdot |h'(y)| \cdot 1_{g(\mathbb{R})}(y)$.

Proof. From the definition of the probability density of Y and by applying the chain rule we get for a non-decreasing function g, that

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} \mathrm{P}(Y \le y) \\ &= \frac{\partial}{\partial y} \mathrm{P}(g(X) \le y) = \frac{\partial}{\partial y} \mathrm{P}(X \le h(y)) \\ &= \frac{\partial}{\partial y} F_X(h(y)) = f_X(h(y)) h'(y) \\ &= f_X(h(y)) \cdot |h'|. \end{aligned}$$

For a non-increasing function $g(\cdot)$ the proof is essentially the same. $\square$

Example. Let $X \sim \mathcal{U}(0,1)$ and $Y = g(X) = X^c, \quad c > 0$. The inverse of g equals $h(y) = y^{\frac{1}{c}}$, its derivative is $\frac{\partial h(y)}{\partial y} = \frac{1}{c} \cdot y^{\frac{1}{c}-1}$. From the transformation theorem it follows that

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| = 1 \cdot \frac{1}{c} \cdot y^{\frac{1}{c}-1} \cdot 1_{(0,1)}(y).$$

Monday, 10th September 2012

Lecture 13 - Content

- □ **Normal approximation to the Binomial**
- □ **Sampling distributions**

References from Phipps & Quine

- □ Section 3.1 pages 72-73.
- □ Section 3.3 pages 75-78.

Normal approximation to the Binomial

Let X_i be independent random variables (outcomes of Bernoulli trials), defined as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a S,} \\ 0 & \text{if the } i\text{th trial is a F,} \end{cases}$$

and let $p = \mathrm{P}(\mathsf{S})$ on the ith trial.

Theorem 24. Let $X = X_1 + \ldots + X_n \sim \mathcal{B}(n, p)$ with $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = n\,\mathrm{Var}(X_1) = np(1-p)$. Then, X is approximately $\mathcal{N}(np, np(1-p))$.

Proof. Postponed to second year... □

- □ The approximation is quite good if $np \geq 5$ and $n(1-p) \geq 5$!
- □ The closer p is to 0.5 the better the approximation for small n.

Example. $(X \sim \mathcal{B}(12, 0.5))$

$$\mathrm{P}(X=3) = \binom{12}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^9 = \frac{12\times 11\times 10}{1\times 2\times 3}\cdot\frac{1}{2^{12}} = 0.0537.$$

Comparing to the area under the approximating normal curve, e.g.

$$X \simeq Y \sim \mathcal{N}(6,3)$$

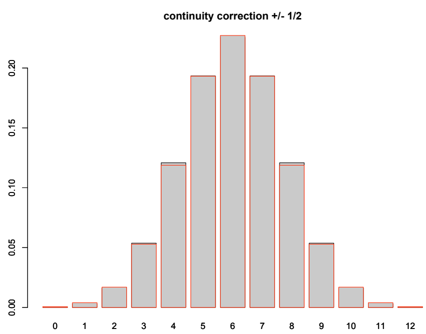
$$\mathrm{P}(X=3) \simeq \mathrm{P}\left(3-\lambda < Y < 3+(1-\lambda)\right); \quad \lambda \in [0,1].$$

Most authors choose $\lambda = 1/2$ which in the above example is clearly closer to the true value of $\mathrm{P}(X=3)$:

```
> mu=6; sd=sqrt(3);
> pnorm(4.0,mu,sd) - pnorm(3.0,mu,sd)
[1] 0.08247428
> pnorm(3.5,mu,sd) - pnorm(2.5,mu,sd)
[1] 0.05280327
> pnorm(3.0,mu,sd) - pnorm(2.0,mu,sd)
[1] 0.03117159
```


Example (cont)

- Note that `pnorm(3.5176,mu,sd) - pnorm(2.5176,mu,sd)` comes very close to `dbinom(3,12,0.5)` but is only optimal for this particular example.
- Overall performance of $\lambda = 1/2$ is best.



Continuity correction

- To approximate binomial probabilities using the normal consider areas of corresponding rectangles.
- Adjust the normal probability statement by adding or subtracting 0.5 to the constant to increase the area under the normal curve.

$$\begin{aligned} \mathrm{P}(X = x) &\simeq \mathrm{P}(x - 0.5 < Y < x + 0.5) \\ &= \mathrm{P}\left(\underbrace{\frac{x - 0.5 - \mu}{\sigma}}_{z_l} < Z < \underbrace{\frac{x + 0.5 - \mu}{\sigma}}_{z_u}\right) \\ &= \Phi(z_u) - \Phi(z_l). \end{aligned}$$

- For $\mathrm{P}(X \geq x)$ repeat the above step by noting:

$$\mathrm{P}(X \geq x) = \sum_{i \geq x} \mathrm{P}(X = i).$$

Example. If $X \sim \mathcal{B}(12, 0.5)$ find $\mathrm{P}(2 \leq X < 5)$.

```
> pnorm((4.5-6)/sqrt(3)) - pnorm((1.5-6)/sqrt(3))
[1] 0.1885507
> sum(dbinom(2:4,12,0.5))
[1] 0.1906738
```

Example (PQ, p80 Q21). It is known that 80% of patients with a certain disease can be cured with a certain drug. What is the probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug.

Example. The proportion of children having a particular type of birth defect born to Pima Indian women is 0.05. Calculate the probability that in 785 independent births no more than 21 children have the birth defect.

Sampling distributions

- ☐ How do statistics vary across samples?
- ☐ Height for randomly selected $n = 4$ adult males.
- ☐ What is the distribution of $\overline{X}$ and S^2?

Model: Assume 4 independent readings of
Observations: X_1, X_2, X_3, X_4

$$\Rightarrow \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$$

The mean: because $\mathrm{E}\, X_i = 178$ and $\mathrm{Var}\, X_i = 8^2$ it follows $\overline{X} \sim \mathcal{N}(178, 4^2)$.
The sample variance: but $s^2 \not\sim \mathcal{N}$!

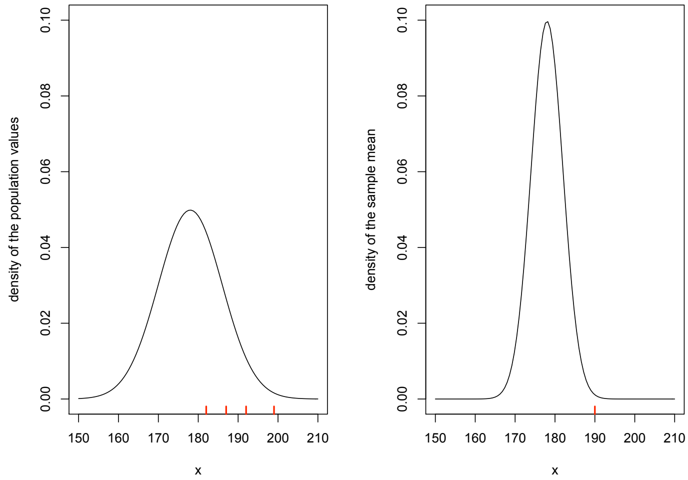
$$S^2 = \frac{1}{4-1}\left((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + (X_4 - \overline{X})^2\right).$$

Sampling distribution for S^2 and non-normal models

- ☐ Use CLT for large n and non-normal models.
- ☐ Knowing the sampling distribution helps identify unusual statistic values.
- ☐ E.g. if $\overline{X}$ was 190 (four basketball players):

```
# sampling distribution and extreme observations
x = c(182,187,192,199);
x.m = mean(x)
dnorm2 = function(x){
return(dnorm(x,mean=178,sd=8))}
dnorm3 = function(x){
return(dnorm(x,mean=178,sd=4))}
par(mfrow=c(1,2))
curve(dnorm2,from=150,to=210,ylim=c(0,0.1),ylab="density of the population values")
rug(x,col=2,lwd=2)
curve(dnorm3,from=150,to=210,ylim=c(0,0.1),ylab="density of the sample mean")
rug(x.m,col=2,lwd=2)
```

Distribution of population and mean



There must be something special with those 4 observations!

Sampling distributions – Movie 1

(Loading bootstrap.mp4)

Sampling distributions – Movie 2

(Loading bootstrap.mp4)