Set Theory, Random Experimenets and Probability

Definition: The sample space S of a random experiment is the set of all possible outcomes.

Definition: An event E is any subset of the sample space. We say that E *occurs* if the observed outcome x is an element of E (E occurs if and only if $x \in E$).

S is the *certain* event.
$\emptyset$ is the *impossible* event.

Define

$$E_1 \cup E_2 = \{x : x \in E_1 \text{ or } x \in E_2\},$$

$$E_1 \cap E_2 = \{x : x \in E_1 \text{ and } x \in E_2\},$$

$$E' = E^c = \{x : x \notin E\},$$

and

$$E - F = \{x : x \in E \text{ and } x \notin F\} = E \cap F^{'}$$

If every element of E is an element of F then $E \subset F$.

Properties:

(i) $(E_1 \cap ... \cap E_n)' = (E_1' \cup ... \cup E_n')$

(ii) $(E_1 \cup ... \cup E_n)' = (E_1' \cap ... \cap E_n')$

(iii) $E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3)$.

(iv) $\emptyset' = S$, $S' = \emptyset$, $E \cap E = E \cup E = E$

(v) If $E \subset F$ then $E \cap F = E$ and $E \cup F = F$

Definition $E_1, ..., E_k$ are *mutually exclusive* if $E_i \cap E_j = \emptyset$ whenever $i \neq j$.

Example. A couple plans to have 2 children

(i) What is the sample space according to the gender of each children

$$S = \{GG, GB, BG, BB\}.$$

Write the event of (ii) at least one boy.

$$E_1 = \{BB, BG, GB\}.$$

(iii) One boy and one girl

$$E_2 = \{BG, GB\}.$$

(iii) At most one boy.

$$E_3 = \{BG, GB, GG\}.$$

(iv) First a boy and then a girl

$$E_4 = \{BG\}.$$

(v) Exactly two girls

$$E_5 = \{GG\}.$$

Example. Pick a point at random from the interior of the circle

$$\{(x, y) : x^2 + y^2 \leq R^2\}$$

(radius=R).

(i) What is the sample space ?

Answer:

$$S = \{(x, y) : x^2 + y^2 \leq R^2\}$$

(ii) Write the set of points that are closer to center than the boundary.

Answer:

$$E_1 = \{(x, y) : x^2 + y^2 < R^2/4\}$$

Axioms of probability: A probability measure on a sample space S is a set function P which assigns to each event $E \subseteq S$ a number $P(E)$ (called the probability of E) such that the following three properties are satisfied:

1. $P(S) = 1$

2. $P(E) \geq 0$ for any event E

3. If $E_1, E_2, ...$ are *mutually exclusive*, then

$$P(E_1 \cup E_2 \cup ...) = P(E_1) + P(E_2) + ...$$

Theorem:

1. $P(\emptyset) = 0$

2. $E_1 \subseteq E_2 \Rightarrow P(E_1) \leq P(E_2)$.

3. $0 \leq P(E) \leq 1$.

4. $P(E') = 1 - P(E)$.

Addition rule:

1.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2.

$$\begin{aligned} &P(A \cup B \cup C) = \\ &\quad P(A) + P(B) + P(C) \\ &\quad -P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad +P(A \cap B \cap C) \end{aligned}$$

Note:

$$P(A \cup B \cup C) = 1 - P(A' \cap B' \cap C')$$

1. The equiprobable model assigns the same probability to each sample point.

2. If $|E|$ is the number of distinct points in an event E

3. then $P(E) = \frac{|E|}{|S|}$.

Counting sample points.

Multiplication rule. Suppose that an experiment (procedure) E_1 has n_1 outcomes and for each of these possible outcomes an experiment (procedure) E_2 has n_2 possible outcomes. The composite experiment (procedure) E_1E_2 that consisting of performing first E_1 and then E_2 has n_1n_2 possible outcomes.

Example. How many subsets of n elements are there?

Answer. 2^n.

Example. How many different numbers of 5 digits can be made such that

(a) Digits can be repeated

(b) Digits can not be repeated

Solution. Part (a):

$$9 \times 10 \times 10 \times 10 \times 10 = 90,0000.$$

Part (b)

$$9 \times 9 \times 8 \times 7 \times 6 = 27216.$$

Example. How many different license plates are possible if a state uses

(i) Two letters followed by a four-digit integer (leading zero permissible and the letters and digits can be repeated),

(ii) Three letters followed by a three-digit integer.

Solution. (a)

$$26 \times 26 \times 10 \times 10 \times 10 \times 10 = 6,760,000$$

(b)

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17,576,000.$$

Permuation and Combination.

A collection of n different objects can be arranged in $n!$ different ways. where

$$n! = n(n-1)\cdots 1$$

fore $n \geq 1$. We define $0! = 1! = 1$.

Example. In how many different ways 10 people can stand in a row ?

Answer. 10!.

Example. In how many different orders n people can seat around a round table ?

Answer.

number of circular permutations $= \dfrac{n!}{n} = (n-1)!$.

Permuations of size r from n letters. An ordered arrangement of r objects selected from

$$a_1, \ldots, a_n$$

is a permutation of n objects taken r at a time.

Example. Write all the permutations of size 2 from 4 letters a, b, c, d.

Solution.

$$ab, ac, ad, ba, bc, bd, ca, cb, cd, da, db, dc.$$

The number of possible ordered arrangements is denoted by P_n^r.

We can generally write

$$P_n^r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Note. To write a permutation we should not repeat any object and the order that an object appears is important. For example ab and ba are different.

Example. The number of possible 4-letter codes selecting from 26 letter in which all 4 letters are different is

$$P_{26}^{4} = \frac{26!}{(26-4)!} = 358,800.$$

Combination. If r objects are selected from a set of n objects and if order of selection is not important, each of these unordered arrangements is called a combination.

Example. Write all the combinations of 2 letters from the letters $a, b, c, d,$.

Solution.

$$ab, ac, ad, bc, bd, cd.$$

Notice that we did not include ba when ab is included. Therefore ab and ba are assumed to be identical combinations.

Notation.

$\left(\begin{array}{c} n \\ r \end{array}\right)$ =number of combinations of size r from n letters.

Theorem.

$$\left(\begin{array}{c} n \\ r \end{array}\right) = \frac{n!}{r!(n-r)!}.$$

Proof. Let C denote the number of unordered arrangements of size r that can be selected from n different objects. We can obtain each of the P_n^r ordered arrangements by first selecting one of the $\left(\begin{array}{c} n \\ r \end{array}\right)$ unordered arrangement and then ordering these r objects in $r!$ ways. Therefore

$$r!\left(\begin{array}{c} n \\ r \end{array} \right) = P_n^r.$$

Binomial Theorem.

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}.$$

Some properties.

(1) Symmetry: $\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$

(2) Pascal's triangle:

$$\left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-1 \\ k \end{array} \right) = \left(\begin{array}{c} n \\ k \end{array} \right).$$

(3) $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Example. How many words (meaningful or meaningless) can we write using all the 11 letters in the word

INDEPENDENT

Answer.

$$\binom{11}{3,2,3} = \frac{11!}{3!2!3!}.$$

Example. A convex polygone of n sides has $\binom{n}{2} - n$ diagonals (why ?). How many triangles could we make using vertices of this convex polygone ?

Answer. $\binom{n}{3}$ (why ?).

Example and appllication in probability.

Example, The birthday problem: There are $c=$??? people in the class. What is the probability two or more have the same birthdate; i.e. they were born on the same day and month (like the 23'd of August) ? (Assume a year consist of 365 days).

Solution.

$$1 - P(\text{all birthdays are different})$$

$$= 1 - \frac{P_{365}^{c}}{365^c} = 1 - \frac{365(364)\cdots(365-c+1)}{365^c}$$

$$\approx \frac{1+2+\cdots+(c-1)}{365} = \frac{c(c-1)}{730}.$$

Note:

(1) Approximation is good if r is small.

(2) We need $c \leq 365$. Otherwise the probability is zero.

Example: n different letters were sent to 4 different addressees at random. Find the probability that at least one goes to the right address.

Solution. Define events

$$A_i = i^{th} \text{ letter goes to the right address}$$

for $i = 1, \ldots, 4$. We need to calculate

$$P(A_1 \cup A_2 \cup A_3 \cdots \cup A_n) = n \times \left(\frac{1}{n}\right)$$

$$-\begin{pmatrix} n \\ 2 \end{pmatrix} \times \left(\frac{1}{n(n-1)}\right) +$$

$$\begin{pmatrix} n \\ 3 \end{pmatrix} \times \left(\frac{1}{n(n-1)(n-2)}\right) - \cdots - \begin{pmatrix} n \\ n \end{pmatrix} \left(\frac{1}{n!}\right)$$

$$\approx 1 - \frac{1}{e}.$$

For large values of n. Notice that the result is not a rational number.

Definition: The *conditional probability* that an event B occurs given that event A has occurred is

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

(provided that $P(A) > 0$).

Multiplication rule:

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

Total Probability Rule:

$$\begin{aligned} P(B) &= P(A \cap B) + P(A' \cap B) \\ &= P(B \mid A)P(A) + P(B \mid A')P(A') \end{aligned}$$

If $E_1, ...E_k$ are mutually exclusive and exhaustive (i.e. $E_i \cap E_j = \emptyset$ if $i \neq j$ and $E_1 \cup .... \cup E_k = S$), then for any event B

$$\begin{aligned} P(B) &= P(B \cap E_1) + ... + P(B \cap E_k) \\ &= P(B \mid E_1)P(E_1) + ... + P(B \mid E_k)P(E_k) \end{aligned}$$

Bayes' Theorem: If $E_1, E_2, \cdots$ are mutually exclusive and exhau stive (i.e. $E_i \cap E_j = \emptyset$ if $i \neq j$ and $E_1 \cup \cdots \cup E_2 \cup \cdots = S$), then for any event B and for each i,

$$P(E_i \mid B) = \frac{P(B \mid E_i)P(E_i)}{P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots}$$

Example: Nissan sold three models of cars in North America in 1999: Sentras, Maximas and Pathfinders. Of the vehicles sold, 50% were Sentras, 30% were Maximas and 20% were Pathfinders. In the same year 12% of the Sentras, 15% of the Maximas and 25% of the Pathfinders had a defect in the ignition system.

1. I own a 1999 Nissan. What is the probability that it has the defect?

2. My 1999 Nissan has the defect. What model do you think I own? Why?

Definition: Two events A and B are *independent* if any one of the following statements is true:

1. $P(B \mid A) = P(B)$

2. $P(A \mid B) = P(A)$

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

Definition: The events $E_1, ..., E_n$ are independent if for any subcollection $E_{i_1}, E_{i_2}, ..., E_{i_k}$,

$$P(E_{i_1} \cap E_{i_2} \cap ... \cap E_{i_k}) = P(E_{i_1}) \times P(E_{i_2}) \times ... \times P(E_{i_k})$$

Example. A box contains m white chips and n black chips. Draw a chip at random and without replacement draw another chip from this box at random. What is the probability that

(i) the first chip is white

(ii) the second chip is white.

Solution. Clearly

$$P(\text{first chip is white}) = \frac{m}{m+n}.$$

Now use the total probability rule to write

$$P(\text{second chip is white})$$

$$= \frac{m-1}{m+n-1} \cdot \frac{m}{m+n} + \frac{m}{m+n-1} \cdot \frac{n}{m+n}$$

$$= \frac{m}{m+n}$$

Random Variables.

Definition: A random variable (r.v.) X is a function $X : S \to \mathbf{R}$. The range of X is the set of possible values of X.

- A random variable is discrete if its range is finite or countable infinite.

- A random variable is continuous if its range is an interval (finite or infinite).

Note: Random variables will be denoted by upper case letters X, Y, Z. Observed values will be denoted by lower case letters x, y, z.

Consider a random experiment with sample space S with probability function P. Let

$$X : S \to \mathbf{R}$$

be a random variable defined on S.

Let $A \subset \mathbf{R}$. Denote the event

$$\{s \in S : X(s) \in A\} = \{X \in A\}.$$

Then

$$P(X \in A) = P(\{s \in S : X(s) \in A\}).$$

Example 1. Roll a die. The outcome space is $S = \{1, 2, 3, 4, 5, 6\}$. For each $s \in S$, let $X(s) = s$. From definition X is a discrete random variable. Let

$$P(\{s\}) = \frac{1}{6}, s \in S$$

then we can write

$$P(2 \leq X \leq 5) = \frac{4}{6}, P(X \geq 2) = \frac{5}{6}$$

Define the random variable Y by

$$Y(s) = s^2.$$

The random variable Y is also a discrete random variable and

$$P(Y \leq 9) = P(\{1, 2, 3\}) = \frac{3}{6}.$$

Example 2. Let X equal the number of flips of a fair coin that are required to observe the first head. Values that X can take are

$$\{1, 2, 3, \ldots\}$$

which is a countable set and X is a discrete random variable. We have

$$P(X \leq 2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}, P(2 \leq X \leq 3) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}.$$

Definition: The cumulative distribution function (c.d.f.) F (F_X) of a random variable X is

$$F(x) = P(X \leq x) = P(\{s \in S : X(s) \leq x\}).$$

Properties of the c.d.f.:

1. $0 \leq F(x) \leq 1$
2. If $x \leq y$ then $F(x) \leq F(y)$.
3. $F(-\infty) = 0$, $F(+\infty) = 1$.

For any $a \leq b \in \mathbf{R}$,

$$\begin{aligned} F(b) - F(a) &= P(a < X \leq b) \\ &= P(\{s \in S : a < X(s) \leq b\}). \end{aligned}$$

Discrete Random Variables

Definition: Let X be a discrete random variable with possible values $x_1, ..., x_n$ (n may be ∞). The **probability mass function (p.m.f.)** f **(**f_X**)** of X is

$$f(x_i) = P(X = x_i) = P(\{s \in S : X(s) = x_i\}).$$

Properties of the p.m.f.:

1. $f(x_i) \geq 0$, $\forall i$

2. $\sum_{i=1}^{n} f(x_i) = 1$

3. $P(X \in A) = \sum_{i:x_i \in A} f(x_i)$

Relationship between the p.m.f. and the c.d.f of a discrete random variable X:

- $F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i)$

- $f(x) = F(x) - F(x-)$

Example: An electronic device contains three components which function independently. There is a probability of 0.1 that the first component is defective, a probability of 0.2 that the second component is defective and a probability of 0.1 that the third component is defective. Let X be the number of defective components in the device.

1. What are the possible values of X?

2. Find and graph the p.d.f. of X.

3. Find and graph the c.d.f. of X.

4. What is the probability of at least one defective component?

5. What is the probability of fewer than 2 defective components?

6. What is $P(1.2 < X \leq 2.5)$?

Example. Let $0 < p < 1$.

(i) Find c such that

$$f(x) = cp^x, x = 0, 1, 2, \ldots$$

be a probability mass function.

(ii) Find $F(x)$ for a nonnegative integer x.

Solution. (i) We need to have $c > 0$ and

$$\sum_{x=0}^{\infty} p^x = \frac{1}{c}.$$

Convergence of the series implies

$$a = 1 + p + p^2 + p^3 + \cdots.$$

Therefore

$$ap = p + p^2 + p^3 + \cdots.$$

This gives

$$a - ap = 1, a = \frac{1}{1-p}.$$

Therefore $c = 1 - p$.

(ii) We have

$$p^x + p^{x+1} + \cdots = \frac{p^x}{1-p}.$$

Therefore

$$1 - F(x) = (1-p) \sum_{k=x+1}^{\infty} p^k = p^{x+1}.$$

Therefore for $x = 0, 1, 2, \ldots$ we have

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

Definition: A random variable X is continuous if its c.d.f. F_X is a continuous function.

Definition: A probability density function (p.d.f.) f of a continuous random variable X is the derivative of the distribution function F (when it exists):

$$f(x) = \begin{cases} \frac{d}{dx}F(x) & \text{when it exists} \\ 0 & \text{otherwise} \end{cases}$$

Properties of the p.d.f.: A function f is a p.d.f. for a continuous r.v. X if

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. For $A \subseteq \mathbf{R}$, $P(X \in A) = \int_A f(x)dx$

In particular, $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$

Note: If X is a continuous r.v. with p.d.f. f, then

- For any x, $P(X = x) = 0$

- For $a, b \in \mathbf{R}$, $a < b$,
$$\begin{aligned} P(a < X < b) &= P(a \le X < b) \\ &= P(a \le X \le b) = P(a < X \le b) \\ &= \textstyle\int_a^b f(x)dx \\ &= F(b) - F(a) \end{aligned}$$

- The value given to $f(x)$ at a single point will not change the value of $\int_a^b f(x)dx$.

Example. A point is picked at random (from

$$S = \{(x, y) : x^2 + y^2 \le R^2\}$$

uniformly. For any $(a, b) \in S$, define

$$X(a, b) = \sqrt{a^2 + b^2}.$$

(i) Find $F(x)$.

(ii) Find $f(x)$.

(iii) $P(R/3 < X < R/2)$.

(iii) $P(X = R/2)$.

Solution.(i) for and $0 < x < R$,

$$F(x) = \frac{\pi x^2}{\pi R^2} = \int_0^x \frac{2t}{R^2} dt.$$

(ii) For $x \in (0, R)$ we have

$$f(x) = \frac{d}{dx} \int_0^x \frac{2t}{R^2} dt = \frac{2x}{R^2}.$$

(iii)

$$F(R/2) - F(R/3) = \frac{1}{4} - \frac{1}{9} = \frac{5}{36}.$$

(iv) $P(X = R/2) = 0$.

Example. Let X be a random variable with p.d.f.

$$f(x) = \frac{c}{1 + x^2}.$$

(i) Find c

(ii) Find $F(x)$.

(iii) Find $P(X > 1)$, $P(X = 1)$ and $P(X < 1)$.

Solution. We need to have $c > 0$ and

$$c \int_{-\infty}^{\infty} \frac{dx}{1 + x^2} = 1.$$

This gives $c = \frac{1}{\pi}$.

(ii)

$$F(x) = \int_{-\infty}^{x} \frac{dt}{\pi(1 + t^2)} = \frac{1}{\pi} \text{arctan}(x) + \frac{1}{2}.$$

(iii)

$$P(X > 1) = \int_1^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

(iv)

$$P(X = 1) = 0$$

(v)

$$P(X < 1) = 1 - P(X \geq 1) = 1 - P(X > 1) = \frac{3}{4}.$$

Example. Let X be the larger outcome when a pair of four sided dice is rolled.

(i) Find the p.m.f.

(ii) Find $P(X > 2)$ and $P(X \geq 2)$.

(iii) Find $F(x)$.

Solution. The sample space is

$$S = \{(a,b) : a, b = 1, 2, 3, 4\}.$$

Therefore

$$P(X=1)=P(\{(1,1)\})=\frac{1}{16},$$

$$P(X=2)=P(\{(1,2),(2,1),(2,2)\})=\frac{3}{16},$$

$$P(X=3)$$

$$=P(\{(1,3),(3,1),(2,3),(3,2),(3,3)\})=\frac{5}{16},$$

and

$$P(X=4)=$$

$$P(\{(1,4),(2,4),(3,4),(4,4),(4,3),(4,2),(4,1)\})$$

$$=\frac{7}{16}.$$

(ii)

$$P(X>2)=1-P(X\leq 2)=\frac{12}{16},$$

$$P(X\geq 2)=1-P(X=1)=\frac{15}{16}.$$

(iii)

$$F(x) = \begin{cases} 0, & \text{if } \; x < 1, \\ \frac{1}{16}, & \text{if } \; 1 \leq x < 2 \\ \frac{4}{16}, & \text{if } \; 2 \leq x < 3 \\ \frac{9}{16}, & \text{if } \; 3 \leq x < 4 \\ 1, & \text{if } \; x \geq 4 \end{cases}$$

Mathematical expectation.

Definition: Let X be a continuous r.v. with the p.d.f. f. The mean or expected value of X is denoted by $\mu_X = \mu = E(X)$ and is defined by

$$\mu = \begin{cases} \sum_x x f(x), & \text{Discrete case} \\ \int_{-\infty}^{\infty} x f(x) dx, & \text{Continuous case.} \end{cases}$$

The *variance* of a random variable X is denoted by

$$\sigma_X^2 = \sigma^2 = V(X) = E[(X - u)^2]$$

and is defined by

$$\sigma^2 = \begin{cases} \sum_x (x-\mu)^2 f(x), & \text{Discrete case} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx, & \text{Continuous case} \end{cases}$$

We have

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

Proof for the continuous case: (similar for the discrete case).

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

$$= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} x f(x)dx$$

$$+\mu^2 \int_{-\infty}^{\infty} f(x)dx$$

$$= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$$

The *standard deviation* of X is

$$\sigma_X = \sigma = \sqrt{V(X)}$$

Example. Let X be a continuous random variable with the p.d.f.

$$f(x) = \begin{cases} \frac{c}{x^3}, & \text{if} \quad x > 1, \\ 0, & \text{if} \quad \text{elsewhere} \end{cases}$$

(i) Find c

(ii) Find μ and σ for the random variable X.

Solution. We need to have

$$c\int_1^\infty \frac{1}{x^3}dx = 1.$$

Therefore $c = 2$.

(ii)

$$E(X) = \int_1^\infty \frac{2}{x^2}dx = 2$$

and

$$E(X^2) = \int_1^\infty \frac{2}{x}dx = \infty.$$

Therefore $\sigma = \infty$.

Definition. Let X be a random variable with the p.d.f. of $f(x)$ and let $g(X)$ be a real valued function. Then

$$E(g(X)) = \begin{cases} \sum_x g(x)f(x), & \text{Discrete case} \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & \text{Continuous case} \end{cases}$$

Example. Let X be a random variable with the probability distribution

$$f(x) = \frac{x^2}{30}, x = 1, 2, 3, 4.$$

Find $E(X)$ and $E(X^2)$ and $Var(X)$.

Solution.

$$u = E(X) = \sum_{x=1}^{4} x\left(\frac{x^2}{30}\right) = \frac{10}{3}$$

and

$$E(X^2) = \sum_{x=1}^{4} x^2\left(\frac{x^2}{30}\right) = \frac{59}{5}.$$

We have

$$\sigma^2 = Var(X) = \frac{59}{5} - \left(\frac{10}{3}\right)^2 = \frac{31}{45}.$$

Some properties.

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

$Var(aX + b) = a^2 Var(X)$.

(ii) $E(aX + bY + c) = aE(X) + bE(Y) + c$

(iii) If X, Y are independent then

$$E(XY) = E(X)E(Y).$$

and

(iv) $Var(aX+bY+c) = a^2 Var(X)+b^2 Var(Y)$

(a, b and c are three constants).

Some discrete distributions.

(1)**Uniform distribution (discrete type).**

Let X be a random variable that takes values $x_1, \ldots, x_k$ with equal probabilities. Then the probability distribution is

$$f(x) = \frac{1}{k}, x = x_1, \ldots, x_k.$$

We have

$$E(X) = \frac{x_1 + \ldots + x_k}{k} = \bar{x}.$$

and

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

(2) **Binomial and multinomial distributions.**

Bernoulli's Experiment. A random experiment with two possible outcomes (success or failure). We have $S = \{s, f\}$. Let $Y(s) = 1$ and $Y(f) = 0$. Take $P(\{s\}) = p$ and $P(\{f\}) = q = 1 - p$. The probability distribution for Y is

y	0	1
$f(y)$	q	p

In other words

$$f(y) = p^y(1-p)^{1-y}, y = 0, 1.$$

We have

$$E(Y) = p = E(Y^2).$$

This gives $\sigma^2 = p(1-p)$. Now in a sequence of n independent of Bernoulli's experiment, define

$$X = \text{ number of successes.}$$

The random variable X takes values in $\{0, 1, \ldots, n\}$. We have

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots, n.$$

Proof. If x successes occur ($x = 0, 1, \ldots, n$), then $n-x$ failures occur. The number of ways that we can write the sequence

$$SSSS \ldots SFFF \ldots F$$

in different order is $\binom{n}{x}$. The probability of each sequence is

$$p^x (1-p)^{n-x}.$$

Therefore

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

Note. $\sum_{x=0}^{n} b(x; n, p) = 1$.

Proof. Let $q = 1 - p$. We have

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

Let $X_i = 1$ if the result of the i^{th} trial is a success and $X_i = 0$ if the result of the i^{th} trial is a failure. Then

$$X = X_1 + \cdots + X_n$$

is the number of successes in n trials. Therefore $X \sim Bin(n, p)$. We have

$$E(X_i) = E(X_i^2) = p \times 1 + (1-p) \times 0 = p.$$

and

$$Var(X_i) = p - p^2 = pq,$$

for $i = 1, 2, \ldots, p$. Therefore

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

and since $X_1, \ldots, X_n$ are independent

$$Var(X) = \sum_{i=1} Var(X_i) = npq.$$

Example. In a manufacturing system the probability that a certain item is being defective is $p = 0.05$. An inspector selects 6 items at random. Let X equal to the number of defective items in the sample. Find

(i) $\mu = E(X)$ and $\sigma^2 = Var(X)$.

(ii) $P(X = 0)$, $P(X \leq 1)$, $P(X \geq 2)$.

(iii) Find $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$.

Solution. (i) We have $X \sim Bin(6, 0.05)$. Therefore

$$\mu = E(X) = np = 6(0.05) = 0.3,$$

$$\sigma^2 = Var(X) = 6(0.05)(0.95) = 0.285.$$

(ii)

$$P(X = 0) = \binom{6}{0} 0.05^0 0.95^6 = 0.7350919,$$

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$= 0.7350919 + \binom{6}{1} 0.05^1 0.95^5$$

$$= 0.7350919 + 0.2321343 = 0.9672262.$$

(iii) We have

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9672262.$$

Multinomial Experiment. An experiment terminates in of the k disjoint classes. Suppose that the probability that an experiment terminate in the i^{th} class be p_i, for $i = 1, \ldots, k$ where

$$p_1 + p_2 + \cdots + p_k = 1.$$

We repeat the experiment n independent times and let X_i for $i = 1, 2, \ldots, k$ be number of times that the experiment terminates in class i. Then

$$P(X_1 = x_1, \ldots, X_k = x_k)$$

$$= \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

where

$$x_1 + x_2 + \cdots + x_k = n.$$

Example. In manufacturing certain item, 95% of the items are good ones, 4% are seconds and 1% are defective. In a sample of size 20 what is the probability that at least 2 seconds or at least 2 defective items are found.

Solution. Define

$$X = \text{number of seconds},$$

and

$$Y = \text{number of defectives}.$$

We need to calculate

$$P(X \geq 2 \text{ or } Y \geq 2)$$

$$= 1 - P((X = 0 \text{ or } 1) \quad \text{and} \quad (Y = 0 \text{ or } 1))$$

$$= 1 - \binom{20}{0, 0, 20}(0.04)^0(0.01)^0(0.95)^{20}$$

$$-\binom{20}{1,0,19}(0.04)^1(0.01)^0(0.95)^{19}$$

$$-\binom{20}{0,1,19}(0.04)^0(0.01)^1(0.95)^{19}$$

$$-\binom{20}{1,1,18}(0.04)^1(0.01)^1(0.95)^{18} = 0.204.$$

Remark. We have

$$(x_1 + \cdots + x_k)^n$$

$$= \sum_{\alpha_1 + \cdots \alpha_k = n} \binom{n}{\alpha_1, \cdots, \alpha_k} x_1^{\alpha_1} \cdots x_k^{\alpha_k}.$$

Hypergeometric Distribution. Consider a collection of N chips. (k chips are white and $N - k$ chips are black). A collection of n chips are selected at random and without replacement. Find the probability that exactly x chips are white.

Solution. Let

$X =$ number of white chips in the sample of n chips.

By the multiplication principle we can write:

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, x = 0, \ldots, n,\ x \leq k.$$

Example. A lot, consisting of 50 fuses, is inspected. If the lot contains 10 defective fuses what is the probability that in a sample of size 5

(i) there is no defective fuse.

(ii) There are exactly 2 defective fuses.

Solution. Let

$X =$ number of defective fuses in the sample.

(i)

$$P(X = 0) = \frac{\binom{10}{0}\binom{40}{5}}{\binom{50}{5}}.$$

(ii)

$$P(X=2) = \frac{\binom{10}{2}\binom{40}{3}}{\binom{50}{5}}.$$

Properties.

$$E(X) = n\left(\frac{k}{N}\right), Var(X) = n\left(\frac{k}{N}\right)\left(1-\frac{k}{N}\right)\frac{N-n}{N-1}.$$

If sampling is with replacement then

$$P(X=x) = \binom{n}{x}\left(\frac{k}{N}\right)^x\left(1-\frac{k}{N}\right)^{n-x}.$$

In this case

$$E(X) = n\left(\frac{k}{N}\right), Var(X) = n\left(\frac{k}{N}\right)\left(1-\frac{k}{N}\right).$$

For large values of N sampling with replacement and without replacement are identical and we can easily see that

$$\frac{N-n}{N-1} \to 1 \text{ as } N \to \infty.$$

Like the multinomial case we can generalize the hypergeometric distribution to more than one variable.

Geometric Distribution. In a sequence of independent Bernoulli trials let

$$P(\{s\}) = p, P(\{f\}) = q, p + q = 1.$$

Define

$X =$ number of trials needed to observe the first success.

To have $X = x$ for a given value $x = 1, 2, 3, \ldots$ we need to have a sequence of $x - 1$ failures follows by a success. Since experiments are independent we can write

$$P(X = x) = pq^{x-1}, x = 1, 2, 3, \ldots.$$

Since

$$1 + q + q^2 + \cdots = \frac{1}{1-q} = \frac{1}{p}. \qquad (*)$$

we have

$$\sum_{x=1}^{\infty} pq^{x-1} = 1.$$

To calculate $E(X)$ we need to find

$$\sum_{x=1}^{\infty} xpq^{x-1} = p \sum_{x=1}^{\infty} xq^{x-1}.$$

Since

$$\sum_{x=1}^{\infty} xq^{x-1} = \frac{d}{dq}\sum_{x=1}^{\infty} q^x = \frac{d}{dq}\left(\frac{q}{1-q}\right) = p^{-2},$$

($p = 1 - q$) we have

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

Similarly

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

$$= p\sum_{x=1}^{\infty} x(x-1)q^{x-1} = pq\sum_{x=1}^{\infty} x(x-1)q^{x-2} = 2qp^{-2}.$$

Therefore

$$E(X^2) = \frac{1}{p} + \frac{2q}{p^2}.$$

This gives

$$Var(X) = \frac{1}{p} + \frac{2q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.$$

Example. The probability that an applicant for driver's license passes the road test is 75%.

(i) What is the probability that an applicant passes the test on his fifth try ?

(ii) What is the average and variance for the number of trials until he passes the road test ?

Solution. (i) We have a sequence of independent Bernoulli trials with $p = 0.75, q = 0.25$. We have

$$P(X = 5) = (0.25)^4(0.75).$$

(ii)

$$E(X) = \frac{1}{0.75} = \frac{4}{3}, Var(X) = \frac{0.25}{0.75^2} = \frac{1}{225}.$$

Example. An inspector examines trucks to check if they emit excessive pollutants. The probability that a truck emits excessive pollutant is 0.05. In average how many truck should he examine to find the first truck which emits excessive pollutants.

Solution.

$$E(X) = \frac{1}{p} = \frac{1}{0.05} = 20.$$

Poisson distribution. In a binomial distribution let $p = \frac{\lambda}{n}$. We have

$$P(X = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

As $n \to \infty$ we have

$$\lim_{n\to\infty} \frac{n!}{(n-x)!n^x} = 1$$

and

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

Therefore

$$\lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = f(x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Definition. A discrete random variable X has Poisson distribution if its p.m.f. is of the form

$$P(X = x) = f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots.$$

Notice that since

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!},$$

we have

$$\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = 1.$$

We can prove that

$$E(X) = \lambda, Var(X) = \lambda.$$

Example. Telephone calls enter a college switchboard on the average of 2 every 3 minutes. Let X denote the number of calls in a 9 minute period. Calculate

$$P(X \geq 5).$$

Solution. In average we have 6 calls for every 9 minutes. Therefore $\lambda = 6$ and

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{e^{-6}6^x}{x!} = 0.715.$$

Example. A certain type of aluminum screen that is 2 feet wide has on the average one flaw in a 100-foot roll. Find the probaility that a 50-foot roll has no flaws.

Solution. In average we have $\lambda = 0.5$ flaws in every 50-foot roll. Therefore

$$P(X = 0) = \frac{e^{-0.5}0.5^0}{0!} = e^{-0.5} \approx 0.61.$$

We saw that when n is large and $p = \lambda/n$ is small we can use the Poisson distribution to approximate the binomial distribution.

Example. Records show that the probaility is 0.00005 that a car will have a flat tire while crossing a certain bridge. Among 10,000 cars crossing this bridge find the probability that

(a) Exactly two will have a flat tire

(b) at most one car has flat tire.

Solution. Number of cars with flat tire among 10,000 cars has $Bin(10,000, 0.00005)$. Since $n = 10,000$ is large and $p = 0.00005$ is small so we can approximate binomial distribution with Poisson distribution with the mean $\lambda = 0.5$. Therefore

(a)

$$P(X = 2) = \frac{e^{-0.5}0.5^2}{2!} = 0.0758$$

and (b)

$$P(X \leq 1) = \frac{e^{-0.5}0.5^0}{0!} + \frac{e^{-0.5}0.5^1}{1!} = 1.5e^{-0.5} \approx 0.91$$

Some Continuous distribution.

Uniform distribution. A point is drwan at random from the interval $[A, B]$ with the uniform density function. We have

$$f(x) = \begin{cases} c, & \text{if} \quad A \leq x \leq B, \\ 0, & \text{if} \quad \text{elsewhere} \end{cases}$$

where c is a constant. We need to have

$$\int_A^B c dx = c(B - A) = 1.$$

Therefore $c = \frac{1}{B-A}$.

Example. Customers arrive randomly at a bank teller's window. Given that one customer arrived during a particular 10-minutes period and let X equal the time within 10 minutes that the customer arrived. If X has a uniform distribution in $[0, 10]$ find the probaility that

(a) $P(X \geq 8)$

(b) $P(2 \leq X < 8)$.

Solution. We have

$$f(x) = \begin{cases} \frac{1}{10}, & \text{if} \quad 0 \leq x \leq 10, \\ 0, & \text{if} \quad \text{elsewhere} \end{cases}$$

This gives

$$P(X \geq 8) = \int_0^8 \frac{1}{10} dx = 0.8$$

and

$$P(2 \leq X < 8) = \int_2^8 \frac{1}{10} dx = 0.6.$$

Mean and Variance. We have

$$\mu = E(X) = \int_A^B x \left(\frac{1}{B-A}\right) dx = \frac{A+B}{2}$$

and

$$E(X^2) = \int_A^B x^2 \left(\frac{1}{B-A}\right) dx = \frac{A^2+B^2+AB}{3}.$$

Therefore

$$\sigma^2 = Var(X) = \frac{A^2+B^2+AB}{3} - \left(\frac{A+B}{2}\right)^2$$

$$= \frac{(B-A)^2}{12}.$$

Definition: A *standard* normal random variable Z is a normal random variable with

$$E(Z) = 0, Var(Z) = 1.$$

Its p.d.f. is

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

Its c.d.f is

$$\begin{aligned}\Phi(z) &= P(Z \leq z) \\ &= \int_{-\infty}^{z} n(z)dz \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz\end{aligned}$$

Theorem: If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal random variable. Therefore,

$$\begin{aligned}F(x) &= P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right) \\ &= \Phi(z)\end{aligned}$$

where $z = \frac{x-\mu}{\sigma}$ is known as the z-value obtained by *standardizing* Z.

Note: If $X \sim N(\mu, \sigma^2)$, then

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Values of $\Phi(z)$ may be found in the Appendix, Table A-3, pages 670-671.

- ϕ is symmetric about the origin.
- $\Phi(-z) = 1 - \Phi(z)$
  $(P(Z \leq -z) = P(Z \geq z))$

Example:
Suppose that $Z \sim N(0, 1)$. Find the following:

1. $P(.53 < Z < 2.06)$
2. $P(-2.63 \leq Z \leq -.51)$
3. $P(|Z| > 1.96)$

4. Find c such that $P(|Z| \le c) = .95$

5. Find c such that $P(|Z| > c) = .10$

Example. Let $X \sim N(\mu, \sigma^2)$. Find the following

1. $P(\mu - \sigma < X < \mu + \sigma)$.

2. $P(\mu - 2\sigma \le X \le \mu + 2\sigma)$.

3. $P(\mu - 3\sigma \le X \le \mu + 3\sigma)$.

Definition: Let X be the time to the first arrival in a Poisson process with rate $\frac{1}{\beta}$. X has the exponential distribution with parameter β ($X \sim$ exponential, β).

- $$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{for } 0 < x \\ 0 & \text{for } x \leq 0 \end{cases}$$

- $$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-x/\beta} & \text{for } 0 \leq x \end{cases}$$

- $E[X] = \beta$

- $V(X) = \beta^2$

- The time between any two successive arrivals in a Poisson process with rate β has the exponential distribution with parameter β.

- "Lack of memory":

$$P(X > s + t \mid X > t) = P(X > s)$$

Normal approximation to binomial distribution. We saw that when $X \sim bin(n, p)$ for a large n and a small p, the binomial distribution can be approximated with the Poisson distribution with mean $\lambda = np$.

What if n is large but p is not small ?

In this case we can use the central limit theorem as follows.

Theorem. Let $X \sim bin(n, p)$ where $0 < p < 1$. As $n \to \infty$ then

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

has the standard normal distribution.

Notes:

(i) Notice that

$$E(X) = np \text{ and } Var(X) = np(1-p).$$

(ii) A rough guide to use normal approximation is that

$$np \geq 5, n(1-p) \geq 5.$$

(iii) For an inteher k we use

$$P(X=k) \approx P\left(\frac{k-0.5}{\sqrt{np(1-p)}} < Z < \frac{k+0.5}{\sqrt{np(1-p)}}\right).$$

Example. Let Y be number of heads in flips of an unbiased coin $n = 10$ times. Find

$$P(3 \leq Y < 6)$$

(i) Accurately

(ii) Using normal approximation.

Solution. (i) We use the table A.I. from the textbook to find

$$P(3 \leq Y < 6) = P(3 \leq Y \leq 5)$$

$$= \sum_{k=5}^{6} \binom{10}{k} 0.5^k 0.5^{10-k} = 0.5683.$$

(ii) Since $E(X) = 5$ and $Var(X) = 2.5$ we can write

$$P(3 \leq Y < 6) = P(3 \leq Y \leq 5) = P(2.5 < Y < 5.5)$$

$$= P\left(\frac{2.5 - 5}{\sqrt{2.5}} < Z < \frac{5.5 - 5}{\sqrt{2.5}}\right)$$

$$= P(Z < 0.316) - P(Z < -1.581)$$

$$= 0.6240 - 0.057 = 0.567.$$

Example. Find the probability that more than 30 but less than 35 of the next 50 births at a particular hospital will be boys.

Solution. Let $X =$ number of boys. We have $X \sim bin(50, 0.5)$ and

$$E(X) = 25, \sigma = \sqrt{12.5} \approx 3.54.$$

Therefore we can write

$$P(31 \leq X \leq 34) = P(30.5 < X < 34.5)$$

$$\approx P\left(\frac{30.5 - 25}{3.54} < Z < \frac{34.5 - 25}{3.54}\right)$$

$$= P(1.55 < Z < 2.68) = P(Z < 2.68) - P(Z < 1.55)$$

$$= 0.0569.$$

Sampling distribution and Statistical inference

Definition: If $X_1, ..., X_n$ are independent and identically distributed with common distribution F, we call $(X_1, ..., X_n)$ a **random sample** from the distribution F. The **sample size** is n. After the data is collected, the observed values of the random variables will be denoted by $x_1, ..., x_n$.

Definition: The common distribution of the random variables in a random sample is sometimes referred to as the **population.**

Definition: A **statistic** is any function of the random variables in a random sample.

Example: Let $X_1, ..., X_n$ be a random sample from a distribution F. Three statistics are:

- the **sample mean:**

$$\overline{X} = \frac{X_1 + ... + X_n}{n}$$

- the **sample variance:**

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Note:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 + \bar{x}^2 - 2x_i\bar{x}\right)$$

$$= \frac{n\sum x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}.$$

- the **sample standard deviation:**

$$S = \sqrt{S^2}$$

Definition: A statistic is a random variable. Its distribution is referred to as a **sampling distribution.**

The Central Limit Theorem (CLT): Let $X_1, ..., X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then as the sample size $n \to \infty$,

$$P(\bar{X} \leq x) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{x - \mu}{\sigma/\sqrt{n}}) \to \Phi(\frac{x - \mu}{\sigma/\sqrt{n}}).$$

In other words, for large values of n (say, $n \geq$ 25 or 30),

$$\begin{aligned} P(a \leq \bar{X} \leq b) &\approx P(\frac{a - \mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b - \mu}{\sigma/\sqrt{n}}) \\ &= \Phi(\frac{b - \mu}{\sigma/\sqrt{n}}) - \Phi(\frac{a - \mu}{\sigma/\sqrt{n}}) \end{aligned}$$

Example. A soft drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 200 milliliters and a standard deviation of 15 milliliters. Waht is the probability that the average amount dispensed in a random sample of size 36 is at least 204 milliliters.

Solution.

$$P(\bar{X} \geq 204) = P\left(Z \geq \frac{\sqrt{36}(204 - 200)}{15}\right)$$

$$= P(Z \geq 1.6) = 0.0548.$$

Example. An electronic company manufactures resistors that have a mean resistance of 100 Ω and a standard deviation of 10 Ω. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than 95 Ω.

Solution.

$$P(\bar{X} < 95) = P\left(Z < \frac{\sqrt{25}(95 - 100)}{10}\right)$$

$$= P(Z < -2.5) = 0.0062.$$

Importnat Notes:

Let X and Y be two independent random variables and

$$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$$

then (i)

$$aX + bY + c \sim N(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b\sigma_2^2)$$

(ii) $aX_1 + b \sim N(a\mu_1 + b, a^2\sigma_1^2)$.

(a, b and c are given constants).

Example. Let $X \sim N(0, 1)$ and $Y \sim N(1, 1)$. Find the distribution for

$$3X - 2, X - Y, 2X + Y - 1, X + Y + 1.$$

Solution.

$$3X - 2 \sim N(-2, 9), X - Y \sim N(-1, 2),$$

$$2X + Y - 1 \sim N(0, 5), X + Y + 1 \sim N(2, 2).$$

Example: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$.

(i) Find the distribution for $\frac{X_1+X_2}{2}$.

(ii) Find the distribution for

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

(iii) If $\mu = 1, \sigma = 4$ and $n = 16$ find

$$P(0.9 < \bar{X} < 1.1).$$

Theorem. Let $X_1, \ldots, X_m$ be an independent sample from a population with the mean μ_1 and the standard deviation σ_1^2. Draw at random another independent sample $Y_1, \ldots, Y_n$ independently from another population with the mean μ_2 and the standard deviation σ_2^2. Then for large values of m and n we have

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

approximately.

Gamma and χ^2 distribution. For $\alpha > 0$, define

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

This gives $\Gamma(1) = 1$. Use integration by parts to conclude

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1).$$

Therefore if α is integer we have

$$\Gamma(\alpha) = (\alpha - 1)!.$$

Take $\beta > 0$ and use the change of variable $x = \frac{y}{\beta}$ to write

$$\int_0^\infty y^{\alpha-1} e^{-\frac{y}{\beta}} dy = \Gamma(\alpha)\beta^\alpha.$$

This shows that

$$g(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}}$$

is a p.d.f. on $(0, \infty)$ (gamma distribution). It is not difficult to show that

$$E(X) = \alpha\beta, Var(X) = \alpha\beta^2.$$

When

$$\alpha = \frac{r}{2}, \beta = 2$$

then

$$X \sim \chi^2(r).$$

r=Degrees of freedom.

$$E(\chi^2(r)) = r, Var(\chi^2(r)) = 2r$$

Probabilities for the chi-square distribution can be calculated from the table A.5 of the textbook.

Example. The effective life of a certain manufactured product is a random variable with mean 5000 hr and standard deviation of 40 hr. A new company manufactures a similar component but claims that the mean life is increased to 5050 hr and decreases the standard deviation to 30 hr. A random sample of size $m = 16$ and $n = 25$ are selected from these companies respectively. What is the probability that the difference in the sample mean is at least 25 hr.

Solution. Approximately

$$\bar{Y} - \bar{X} \sim N(5050 - 5000, \frac{900}{25} + \frac{1600}{16})$$

$$\bar{Y} - \bar{X} \sim N(50, 136).$$

$$P(\bar{Y} - \bar{X} > 25) \approx P\left(Z > \frac{25 - 50}{\sqrt{136}}\right)$$

$$= P(Z > -2.14) = 0.9838$$

Notation.

$$P(t(n) > t_\alpha(n)) = \alpha$$

Example. Calculate

(i) $P(-1.96 < t_{50} < 1.96)$

0.95

(ii) $t_{0.05}(12)$, $t_{0.01}(12)$.

1.782, 2.681

Definition: If $X_1, ..., X_n$ is a random sample from a distribution F which depends on an unknown parameter θ, any statistic

$$\widehat{\Theta} = h(X_1, ..., X_n)$$

used to estimate θ is called a **point estimator** of θ. After the sample has been selected, the observed values $x_1, ..., x_n$ are used to obtain a numerical value

$$\widehat{\theta} = h(x_1, ..., x_n)$$

which is called the **point estimate** of θ.

Example: If the distribution F has mean μ and variance σ^2, then $\overline{X}$ is a point estimator of μ and S^2 is a point estimator of σ^2.

Properties of Estimators

Definition: $\widehat{\Theta}$ is an *unbiased* estimator for θ if

$$E[\widehat{\Theta}] = \theta.$$

The *bias* of the estimator is

$$E[\hat{\Theta}] - \theta.$$

Examples: Let $X_1, ..., X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2.

- The sample mean $\bar{X}$ is an unbiased estimator for the population mean μ.

- The sample variance S^2 is an unbiased estimator for the population mean σ^2.

- Occasionally the following estimator is used for σ^2:

  $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

  Find its bias.

Solution. We have

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Therefore

$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sum_{i=1}^{n} E(X_i - \mu)^2 - nVar(\bar{X})$$

$$= n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = (n-1)\sigma^2.$$

Therefore

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sigma^2$$

and

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{(n-1)\sigma^2}{n}.$$

Note: If two different estimators are unbiased for θ, the one with the smaller variance is best.

Definition: Considering all unbiased estimators for θ, the one with the smallest variance is the *mininum variance unbiased estimator* and is called the most efficient estimator of θ (MVUE).

Example: Let $X_1, ..., X_n$ be a random sample from a distribution with mean μ and variance σ^2.The standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}.$$

If σ is unknown, the estimated standard error of $\bar{X}$ is

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}.$$

Confidence Intervals

We have a population with distribution F. Suppose that the population variance σ^2 is known but the mean μ is not. We take a sample $X_1, ..., X_n$ from F and use the sample mean

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

to estimate μ.

How close is $\bar{X}$ to the true value of μ?

By the CLT, if $n \geq 25$

$$\begin{aligned} 2\Phi(z) - 1 &\approx P\left(-z \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \\ &= P\left(\bar{X} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Let $P(|Z| > z_{\alpha/2}) = \alpha$. Then $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ and

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

The interval

$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

has random endpoints and there is a probability of $(1 - \alpha)$ that it will contain the true value of μ.

When the sample has been selected, the observed interval

$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

is called a $100(1-\alpha)\%$ confidence interval for μ.

Definition: Let θ be an unknown parameter. Suppose there exist random variables L and U based on an estimator $\widehat{\Theta}$ of θ such that

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

- If l, u are the observed values of L and U, we call $[l, u]$ a $100(1 - \alpha)\%$ **confidence interval** for θ.

- l and u are called the **lower and upper confidence limits**.

- $(1-\alpha)$ is the **confidence coefficient** of the interval.

- The half-interval length is the **precision** of the interval.

Sample Size. To estimate μ by $\bar{x}$ with a specified error e with $100(1-\alpha)\%$ confidence we need to have

$$e = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Solve for n (the necessary sample size) to get

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Example. If a random sample of size $n = 20$ from a normal population with the variance $\sigma^2 = 225$ has the mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

Solution. Since

$$n = 20, \sigma^2 = 225, z_{0.025} = 1.96$$

we have

$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

$$= \left[64.3 - 1.96\left(\frac{15}{\sqrt{20}}\right), 64.3 + 1.96\left(\frac{15}{\sqrt{20}}\right)\right]$$

$$= [57.7, 70.9]$$

Example. We would like to estimate the mean thermal conductivity of a certain iron with error less than 0.1, with 95% confidence. From the previous investigations we know $\sigma = 0.3$. Find the sample size required.

Solution. We have

$$n = \left(\sigma\frac{z_{\alpha/2}}{e}\right) = \left[\frac{(1.96)0.3}{0.1}\right]^2 = 34.57.$$

Therefore we need to have at least 35 samples.

Confidence Interval for mean when variance is unknown

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. observations from a normal population. We saw that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1).$$

Therefore

$$P\left(-t_{\alpha/2}(n-1) < \frac{\sqrt{n}(\bar{X} - \mu)}{S} < t_{\alpha/2}(n-1)\right) = 1 - \alpha.$$

Thus

$$P\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Then a $100(1-\alpha)\%$ confidence interval for μ when σ is unknown is

$$\left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right].$$

Example. The content of $n = 7$ similar container of sulfuric acid are

$$9.8, 10.2, 10.4, 9.8, 10, 10.2, 9.6.$$

Find a 95% confidence interval for μ.

Solution. We have

$$\bar{x} = \frac{9.8 + 10.2 + 10.4 + 9.8 + 10 + 10.2 + 9.6}{7} = 10$$

and

$$S^2 = \frac{1}{6}((9.8-10)^2 + (10.2-10)^2 + (10.4-10)^2 + (9.8-10)^2$$

$$+(10 - 10)^2 + (10.2 - 10)^2 + (9.6 - 10)^2) = 0.08$$

and $s = 0.283$. Therefore the 95% confidence interval for μ is

$$\left[10 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right)\right]$$

$$= [9.74 < \mu < 10.26].$$

Testing statistical hypothesis.

Example. The Acme Lightbulb Company has a problem. It has found a crate of 10,000 unlabelled light bulbs. It produces two types of light bulb: regular and longlife. The lifetimes of both types of bulbs are normally distributed with standard deviation 500 hours. Regular bulbs have a mean lifetime of 1,000 hours, while longlife bulbs have a mean lifetime of 1,500 hours. Acme would like to sell these light bulbs, but it must decide what label to put on the packages. If the bulbs are erroneously labelled longlife, the company's reputation will be damaged, and they will lose a substantial market share. It would clearly be safer to sell them as regular bulbs. However, regular bulbs have a much lower selling price, and Acme would lose a significant amount of money if in fact they are longlife. Thus, Acme will reject the "null hypothesis" that the mean lifetime is $\mu = 1000$ and accept the "alternative hypothesis" that $\mu = 1500$ if it feels that it has strong evidence to do so. Otherwise, Acme will not reject the null hypothesis that $\mu = 1000$.

Problem:

To test $H_0 : \mu = 1000$ vs. $H_1 : \mu = 1500$

A sample of 10 light bulbs is taken, and the lifetime of each is observed. The sample mean $\bar{X}$ is calculated. H_0 will be rejected if the observed value $\bar{x}$ is large (i.e. if $\bar{x} > c$, where c is some constant).

Question: How large should c be? c is the **critical value** of $\bar{x}$.

Answer: This depends on what sort of risk Acme is willing to take that it will make a **type I error** by rejecting H_0 when in fact it is true. This probability is the **significance level** of the test and is denoted by α.

What is the probability of a **type II error**: i.e. that Acme does not reject H_0 when H_1 is true? This probability is denoted by β.

1. Suppose Acme decides that it is willing to take a 5% chance of a type I error. For what values of $\bar{x}$ will Acme reject H_0?

   Solution. Under the null hypothesis (H_0)

$$\bar{X} \sim N\left(1000, \frac{500^2}{10}\right).$$

   Therefore

$$0.05 = P(\bar{X} > c | \mu = 1000, \sigma = 500)$$
$$= P\left(Z > \frac{\sqrt{10}(c - 1000)}{500}\right).$$

This gives

$$c = 1000 + 1.645 \left(\frac{500}{\sqrt{10}} \right) = 1260.0973.$$

Therefore we reject H_0 if and only if

$$\bar{X} > 1260.0973.$$

2. What is the probability of a type II error if $\alpha = 0.05$? If $\alpha = 0.01$?

For $\alpha = 0.05$ we get $c = 1260.0973$. Therefore

$$\beta = P(\bar{X} \leq 1260.0973 | \mu = 1500, \sigma = 500)$$

$$= P\left(Z < \frac{\sqrt{10}(1260.0973 - 1500)}{500} \right) = 0.0646.$$

For $\alpha = 0.01$,

$$c = 1000 + 2.33 \left(\frac{500}{\sqrt{10}} \right) = 1368.4053$$

and

$$\beta = P(\bar{X} \leq 1368.4053 | \mu = 1500, \sigma = 500)$$

$$= P\left(Z \leq \frac{\sqrt{10}(1368.4053 - 1500)}{500} \right) = 0.203.$$

Notice that we have a larger β (for a smaller α).

3. If we increase the sample size to 25, what is the appropriate critical value for $\alpha = .01$? What is the probability of a type II error?

   For $\alpha = 0.01$,

$$c = 1000 + 2.33\left(\frac{500}{\sqrt{25}}\right) = 1233$$

   and

$$\beta = P\left(Z \leq \frac{\sqrt{25}(1233 - 1500)}{500}\right) = 0.038.$$

4. A sample of size 10 is taken, and we observe a mean life of 1300 hours. What conclusion can be drawn? What is the probability that we would get a value of $\bar{X}$ at least this extreme if in fact H_0 is true?

$$p - value = P(\bar{X} > 1300|\mu = 1000, \sigma = 500)$$

$$= P\left(Z > \frac{\sqrt{10}(1300 - 1000)}{500}\right) = 0.02888976.$$

   For $\alpha = 0.05$ we should reject H_0 and for $\alpha = 0.01$ we should accept H_0.

Hypothesis Testing

Definitions:

- A *statistical hypothesis* is a statement about one or more population parameters. It is *simple* if it assigns exactly one value to the population parameter(s) (eg. $\theta = \theta_0$). If more than one value is assigned, it is *composite* (eg. $\theta \leq \theta_0$).

- The *null hypothesis* H_0 is the statement to be rejected or not rejected. It will always be stated as a simple hypothesis of the form $H_0 : \theta = \theta_0$.

- The *alternative hypothesis* H_1 is the statement which is accepted when H_0 is rejected. A composite alternative may be two-sided (eg. $\theta \neq \theta_0$) or one-sided (eg. $\theta > \theta_0$).

- A *test of a statistical hypothesis* is a procedure leading to a decision on whether or not to reject the null hypothesis. It will be based on the *test statistic.*

- Those values of the test statistic for which H_0 is rejected is known as the **critical region**.

- Those values of the test statistic for which H_0 is not rejected is known as the **acceptance region**.

- The boundary point(s) between the acceptance and critial regions are the **critical values** of the test statistic.

- A **type I error** is committed if H_0 is rejected when it is true. Let

$$\begin{aligned}\alpha &= P(\text{reject } H_0 \mid H_0 \text{ true}) \\ &= P(\text{reject } H_0 \mid \theta = \theta_0) \\ &= P(\text{type I error})\end{aligned}$$

  Then α is the *significance level* or the *size* of the test.

- A **type II error** is committed if H_0 is not rejected when it is false. Let $\theta_1 \in H_1$.

$$\begin{aligned}\beta(\theta_1) &= P(\text{do not reject } H_0 \mid \theta = \theta_1) \\ &= P(\text{type II error} \mid \theta = \theta_1)\end{aligned}$$

  The value of β will vary for a composite alternative.

- The *power function* $K(\theta)$ gives the probability of rejecting the null hypothesis. It is a function of the value of the unknown parameter. On H_O, $K(\theta_0) = \alpha$. For $\theta_1 \in H_1$, $K(\theta_1) = 1 - \beta(\theta_1)$.

- The p-value associated with an observation of the test statistic is the probability of the test statistic taking on a value at least as extreme, if H_0 is true. H_0 is rejected if the p-value is $\leq \alpha$.

Note: Rejecting the null hypothesis is a *strong* conclusion. Not rejecting the null hypothesis is a weak conclusion. **Inference about the mean of a population, know variance**

$X_1, ..., X_n$ is a random sample from a population with mean μ (unknown) and variance σ^2 (known). Either the underlying distribution is known to be normal or $n \geq 25$, so that the CLT is valid. Therefore

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(approximately, in the case of the CLT).

We wish to test $H_0 : \mu = \mu_0$ against one of the following three alternatives:
a) $H_1 : \mu \neq \mu_0$ (two-sided alternative)
b) $H_1 : \mu > \mu_0$ (one-sided alternative)
b) $H_1 : \mu < \mu_0$ (one-sided alternative)

Test statistic:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

If H_0 is true, $Z_0 \sim N(0, 1)$.

For $\alpha > 0$, define z_α to be that value such that $P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha$.

Critical region for a test of significance level α: a) $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$ (i.e. $\mid z_0 \mid > z_{\alpha/2}$). This is a two-sided or *2-tailed test.* The critical values of the test statistic are $z_{\alpha/2}$ and $-z_{\alpha/2}$.
b) $z_0 > z_\alpha$. This is one-sided *upper-tailed* test. The critical value is z_α.
c) $z_0 < -z_\alpha$. This is one-sided *lower-tailed* test. The critical value is $-z_\alpha$.

p**-value** when z_0 is the observed value of Z_0:
a) $P(z_0) = P(\mid Z_0 \mid > \mid z_0 \mid) = 2(1 - \Phi(\mid z_0 \mid))$
b) $P(z_0) = P(Z_0 > z_0) = 1 - \Phi(z_0)$
b) $P(z_0) = P(Z_0 < z_0) = \Phi(z_0)$

Note: Equivalently we could have used the test statistic $\bar{X}$ and the critical regions
a) $\mid \bar{x} - \mu_0 \mid > z_{\alpha/2}\sigma/\sqrt{n}$
b) $\bar{x} > \mu_0 + z_\alpha\sigma/\sqrt{n}$
c) $\bar{x} < \mu_0 - z_\alpha\sigma/\sqrt{n}$

Type II errors: Let $\delta = \mu_1 - \mu_0$. If the true mean is μ_1, then

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$$

and

$$Z_0 \sim N\left(\frac{\delta}{\sigma/\sqrt{n}}, 1\right).$$

a) $\beta(\mu_1) = \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right)$

b) $\beta(\mu_1) = \Phi\left(z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right)$

c) $\beta(\mu_1) = 1 - \Phi\left(-z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right)$

Power function: $K(\mu_1) = 1 - \beta(u_1)$

Note: If α is not specified, it is assumed to be 5%.

Inference about the mean of a population, unknown variance

$X_1, ..., X_n$ is a random sample from a population with mean μ (unknown) and variance

σ^2 (unknown). In this case we use

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1).$$

(Replace σ by S). Therefore we make decision based on T instead of Z. The rest of arguments remains the same.

Example. Let X equal the growth in 20 days of a tumor induced in a mouse in millimeters. Let $X \sim N(\mu, \sigma^2)$. Test

$$H_0 : \mu = 4 \text{ against } H_1 : \mu \neq 4.$$

In a sample of $n = 9$ observations with $\bar{x} = 4.3$ and $s = 1.2$ with a significance level of $\alpha = 0.1$ should we accept or reject H_0. Calculate the p-value.

Solution. We should reject H_0 if

$$|T| = \left|\frac{\bar{x} - 4}{S/\sqrt{n}}\right| > t_{\alpha/2}(n-1).$$

Since $t_{0.05}(8) = 1.86$ we have

$$\left|\frac{4.3 - 4}{1.2/3}\right| = 0.75 < 1.86$$

and we accept H_0. The pvalue is

$$p-\text{value} = 2P(t(8) > 0.75) = 0.475$$

Note: We can not calculate this $p-$value exactly from the table. We calculated this with a computer program.

Inference on a population proportion

$X \sim B(n, p)$ where p is unknown.

Point estimator of p:

$$\hat{P} = \frac{X}{n}$$

Test at level α $H_0 : p = p_0$ vs.

1. $H_1 : p \neq p_0$
2. $H_1 : p > p_0$
3. $H_1 : p < p_0$

Test statistic for n large enough that $np \geq 5$ and $n(1-p) \geq 5$:

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\widehat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

If H_0 is true, then Z_0 is approximately $N(0,1)$.

Reject $H_0 : p = p_0$ and accept
1. $H_1 : p \neq p_0$ if $\mid z_0 \mid > z_{\alpha/2}$
2. $H_1 : p > p_0$ if $z_0 > z_\alpha$
3. $H_1 : p < p_0$ if $z_0 < z_{1-\alpha} = -z_\alpha$

P-value of z_0**=**
1. $P(\mid Z_0 \mid \geq z_0 \mid H_0) = 2(1 - \Phi(\mid z_0 \mid)$
2. $P(Z_0 \geq z_0 \mid H_0) = 1 - \Phi(z_0)$
3. $P(Z_0 \leq z_0 \mid H_0) = \Phi(z_0)$

Example: The Acme Lightbulb Company does not want the proportion of defective lightbulbs which it produces to exceed .05. A sample of 100 bulbs is taken from a large lot and a decision on whether to accept or reject the lot will be based on X, the number of defectives in the sample. The company is willing to take a 10% risk of rejecting the lot when in fact $p = .05$.

1. Formulate an appropriate test of hypothesis, giving the test statistic and the critical region.

$$H_0 : p = p_0 = 0.05,$$

$$H_1 : p > p_0 = 0.05.$$

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{X - 5}{\sqrt{4.75}}.$$

Since $z_\alpha = z_{0.10} = 1.282$, reject H_0 if

$$Z_0 > 1.282.$$

2. The sample contains 8 defectives. What action will Acme take? Find the p-value of

the test statistic.

$$z_0 = \frac{8 - 100(0.05)}{\sqrt{100(0.05)(0.95)}} = 1.3765$$

therefore we should reject H_0.

$$p - \text{value} = P(Z > 1.3765) = 0.0843.$$

Confidence intervals for p

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \\
&\approx P\left(-z_{\alpha/2} \le \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right) \\
&= P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \right. \\
&\qquad \left. \le \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \\
&\approx P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \right. \\
&\qquad \left. \le \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)
\end{aligned}$$

$\Rightarrow$ approximate $(1-\alpha)100\%$ two-sided confidence interval for p:

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Sample size To ensure that

$$P(|\ \widehat{P} - p\ | \leq E) \geq 1-\alpha,$$

we must have

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq E \Leftrightarrow n \geq \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p)$$

Since p is unknown, note that $p(1-p) \leq 1/4$, so

$$n \geq \frac{1}{4}\left(\frac{z_{\alpha/2}}{E}\right)^2 \Rightarrow P(|\ \widehat{P} - p\ | \leq E) \geq 1-\alpha$$

Example: The Acme Lightbulb Company is conducting a market survey to estimate the proportion p of consumers who prefer Acme products.

1. If in a sample of 100 households it is found

that 32 prefer Acme products, construct a 98% confidence interval for p.

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.32 \pm 2.0536\sqrt{\frac{0.32(0.68)}{100}}$$

$$= 0.32 \pm 0.0958.$$

2. How large a sample should be taken if Acme wants to be 98% certain that the estimated proportion is within .05 of p?

$$n \geq \frac{1}{4}\left(\frac{z_{\alpha/2}}{E}\right)^2 = \frac{1}{4}\left(\frac{2.0536}{0.05}\right)^2 = 421.7273.$$

Take $n = 422$.

Chapter 11, Simple linear regression and correlation. Let observations be paired in the sense that an (x, y) pair arise from the same sampling unit. For n sampling units, we can write the measurement pairs as

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

A major purpose for collecting bivariate data is to answer the following questions:

(i) Are the variables x and y related ?

(ii) What type of relationship is indicated by the data ?

(iii) Can we find a quantity for the strength of their relationship ?

(iv) Can we predict one variable from the other and how accurate is our prediction?

Model. Let

$$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where

$$\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

The least square criterion. For the points in a scatter diagram usually there is no single line that passes through all those points. We would like to find the line of **best** fit. The best line is the line with smallest sum of squared errors.

The principlpe of Least Squares.

Determine the values of slope (β) and intercept (α) such that

$$SSE = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

is minimized. Let $a = \widehat{\alpha}$ and $b = \widehat{\beta}$ be the least square estimates for α and β and denote

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\beta} x_i, i = 1, 2, \ldots, n$$

as predicted response (fitted values) and

$$e_i = \text{Observed response-Predicted response}$$

$$= y_i - \widehat{y}_i$$

as errors (residuals). Here is how we can find a and b. Differentiate SSE with respect to a and b to get

$$\frac{\partial SSE}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

and

$$\frac{\partial SSE}{\partial \beta} = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0$$

and solve for α and β to get

$$\widehat{\alpha} = a = \bar{y} - b\bar{x}$$

and

$$\widehat{\beta} = b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

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

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

Example 2. Latitudes and magnitudes of earthquakes occurred in 13 spots are recorded in the following table.

	x	y	xy	x^2	y^2
	60	4.1	246	3600	16.81
	77.5	4	310	6006.25	16
	50.7	2.6	131.82	2570.49	6.76
	65.6	2.8	183.68	4303.36	7.84
	48.2	0.9	43.38	2323.24	0.81
	63.5	2.2	139.70	4032.25	4.84
	49.2	3	147.6	2420.64	9
	60.3	4.1	247.23	3636.09	16.81
	52.6	1.2	63.12	2766.76	1.44
	52.8	1.1	58.08	2787.84	1.21
	64.3	5.5	353.65	4134.49	30.25
	49.3	2.7	133.11	2430.49	7.29
	48.3	0.9	43.47	2332.89	0.81
Total	742.3	35.1	2100.84	43344.79	119.87

We get

$$\bar{x} = \frac{742.3}{13} = 57.1, \bar{y} = \frac{35.1}{13} = 2.7$$

and

$$b = \frac{(13)(2100.84) - (742.3)(35.1)}{(13)(43344.79) - (742.3)^2} = 0.1007$$

and

$$a = 2.7 - 0.1007(57.1) = -3.04997.$$

Therefore the regression line is

$$y = -3.04997 + 0.1007x.$$

Notice that we can find residulas (e_i) easily. For example

$$e_1 = 4.1 + 3.04997 - 0.1007(60) = 1.10797$$

Distribution for a and b. We have

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i Y_i$$

where

$$Y_i \overset{i.i.d.}{\sim} N(\alpha + \beta x_i, \sigma^2), i = 1, \ldots, n$$

and

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

Since b is a linear combination of independent normal random variables we have

$$b = \sum_{i=1}^{n} c_i Y_i \sim N(\sum_{i=1}^{n} c_i(\alpha + \beta x_i), \sigma^2 \sum_{i=1}^{n} c_i^2).$$

Since

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

and

$$\sum_{i=1}^{n} c_i x_i = 1,$$

we have

$$\sum_{i=1}^{n} c_i(\alpha + \beta x_i) = \beta.$$

This shows that $E(\widehat{\beta}) = \beta$. Also

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

This gives

$$b = \widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Similarly we have

$$\widehat{\alpha} = a = \sum_{i=1} (\frac{1}{n} - c_i \bar{x}) Y_i = \sum_{i=1} d_i Y_i$$

where

$$d_i = \frac{1}{n} - c_i \bar{x}, i = 1, \ldots, n.$$

Therefore

$$\widehat{\alpha} \sim N(\sum_{i=1}^{n} d_i(\alpha + \beta x_i), \sigma^2 \sum_{i=1}^{n} d_i^2).$$

Now since

$$\sum_{i=1}^{n} d_i = 1, \sum_{i=1}^{n} d_i x_i = 0$$

and

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x}\right)^2 = \sum_{i=1}^{n} \left(\frac{1}{n^2} + c_i^2 \bar{x}^2 - \frac{2 c_i \bar{x}}{n}\right)$$

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

This gives

$$a = \widehat{\alpha} \sim N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)\right)$$

Notice that

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

Notations. Define

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Therefore

$$b = \frac{S_{xy}}{S_{xx}}.$$

Since $a = \bar{y} - b\bar{x}$ we can write

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y} - b(x_i - \bar{x}))^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 + b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2b \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= S_{yy} - 2bS_{xy} + b^2 S_{xx} = S_{yy} - bS_{xy}.$$

Theorem. We have

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2).$$

Therefore

$$E\left(\frac{SSE}{n-2}\right) = \sigma^2.$$

Therefore

$$S^2 = \frac{SSE}{n-2}$$

is an unbiased estimate for σ^2.

Inference on regression coefficients. Since

$$b = \widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

and

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

we can write

$$\frac{\widehat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t(n-2).$$

Note: $\widehat{\beta}$ is independent from S.

A $100(1-\alpha)\%$ confidence interval for β is

$$b - \frac{t_{\alpha/2}(n-2)S}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2}(n-2)S}{\sqrt{S_{xx}}}.$$

Similarly since

$$a = \hat{\alpha} \sim N\left(\alpha, \sigma^2\left(\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

a $100(1-\alpha)\%$ confidence interval for α is

$$a - \frac{t_{\alpha/2}(n-2)S\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + \frac{t_{\alpha/2}(n-2)S\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}.$$

Example. (i) For the following data values find the formula for the regression line.

	x	y	x^2	y^2	xy	e
	3	9	9	81	27	1.85
	3	5	9	25	15	-2.15
	4	12	16	144	48	2.11
	5	9	25	81	45	-3.63
	6	14	36	196	84	-1.37
	6	16	36	256	96	0.63
	7	22	49	484	154	3.89
	8	18	64	324	144	-2.85
	8	24	64	576	192	3.15
	9	22	81	484	198	-1.59
Total	59	151	389	2651	1003	0.04

This gives

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$= \frac{10(1003) - (59)(151)}{10(389) - (59)^2} = 2.74$$

and since

$$\bar{x} = 5.9, \bar{y} = 15.1$$

we get

$$a = \bar{y} - b\bar{x} = 15.1 - (2.74)(5.9) = -1.07.$$

Therefore the regression line is

$$y = -1.07 + 2.74x.$$

We have

$$S^2 = \frac{1}{n-2}\sum_{i=2}^{n} e_i^2 = 7.956601,$$

and

$$\sum_{i=1}^{n} x_i^2 = 389$$

Since $t_{0.025}(8) = 2.306$ and $S^2 = 7.956601$, $S = 2.820745$ and $S_{xx} = 40.9$ a $100(1-\alpha)\%$ confidence interval for α and β are

$$-1.07 \pm 2.306\sqrt{7.956601/40.9}$$

$$= -1.07 \pm 1.017095.$$

and

$$2.74 \pm 2.306\sqrt{\frac{7.956601(389)}{10(40.9)}}.$$

Hypothesis testing. To test

$$H_0 : \beta = \beta_0 \text{ against } \beta > \beta_0$$

Use

$$\frac{\widehat{\beta} - \beta_0}{S/\sqrt{S_{xx}}} > t_\alpha(n-2) \Rightarrow RH_0.$$

Similarly to test

$$H_0 : \beta = \beta_0 \text{ against } \beta \neq \beta_0$$

use

$$\left| \frac{\widehat{\beta} - \beta_0}{S/\sqrt{S_{xx}}} \right| > t_{\alpha/2}(n-2) \Rightarrow RH_0.$$

Similarly to test

$$H_0 : \alpha = \alpha_0 \text{ against } H_1 : \alpha > \alpha_0$$

we use

$$\frac{\widehat{\alpha} - \alpha_0}{S\sqrt{\sum_{i=1}^n x_i^2/(nS_{xx})}} > t_\alpha(n-2) \Rightarrow RH_0.$$

and to test

$$H_0 : \alpha = \alpha_0 \text{ against } H_1 : \alpha \neq \alpha_0$$

we use

$$\left| \frac{\widehat{\alpha} - \alpha_0}{S\sqrt{\sum_{i=1}^n x_i^2/(nS_{xx})}} \right| > t_{\alpha/2}(n-2) \Rightarrow RH_0.$$

Confidence interval for $E(Y|x)$ and prediction interval. We first find $Cov(\widehat{\alpha}, \widehat{\beta})$. From

$$\widehat{\alpha} = \sum_{i=1}^{n} (\frac{1}{n} - w_i \bar{x}) Y_i$$

and

$$\widehat{\beta} = \sum_{i=1}^{n} w_i Y_i$$

where

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

we can write

$$Cov(\widehat{\alpha}, \widehat{\beta}) = \sum_{i=1}^{n} (\frac{1}{n} - w_i \bar{x}) w_i \sigma^2 = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

For a future observation x_0 we would like to construct a $100(1-\alpha)\%$ confidence interval for $E(Y_0) = \alpha + \beta x_0$. An unbiased estimate for $\alpha + \beta x_0$ is $\widehat{\alpha} + \widehat{\beta} x_0$. Since

$$\widehat{\alpha} + \widehat{\beta} x_0 \sim N\left(\alpha + \beta x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]\right).$$

Therefore a $100(1-\alpha)\%$ confidence interval for $\alpha+\beta x_0$ is

$$\hat{\alpha}+\hat{\beta}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}.$$

Prediction Interval for Y_0.

we can show that

$$\hat{Y_0}-Y_0 \sim N\left(0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\right).$$

Therefore a $100(1-\alpha)\%$ prediction interval for Y_0 is

$$\hat{\alpha}+\hat{\beta}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}.$$

Correlation. The correlation coefficient for pairs

$$(x_1,y_1),\ldots,(x_n,y_n)$$

is defined by

$$r=\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

From

$$\widehat{\beta} = r\frac{S_y}{S_x}$$

where

$$S_y = \sqrt{S_{yy}}, S_x = \sqrt{S_{xx}}$$

we can conclude that

$$sign(\widehat{\beta}) = sign(r).$$

We can show that

$$-1 \leq r \leq 1.$$

To see this note that

$$0 \leq Q(t) = \sum_{i=1}^{n} [(x_i - \bar{x}) + t(y_i - \bar{y})]^2$$

$$= t^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2t \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

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

This is a nonnegative qudratic function of t. Therefore it can not have any root. This implies that

$$(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2 \leq \sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Equality holds if

$$y_i - \bar{y} = k(x_i - \bar{x})$$

(i.e. there is an exact linear relationship between x and y). In this case $r^2 = 1$ ($r = \pm 1$).