Medical statistics – Part I *

KLMED8004 - autumn 2009

Chap. 4. Random variable
Discrete probability distribution

Harald Johnsen, sept. 2009

***) Slightly different from Norwegian version**

Random experiment

The concept of probability is relevant to experiments that have somewhat uncertain outcomes. However, the term "experiment" is not restricted to laboratory or designed experiments, but includes *any activity that results in the collection of data pertaining to phenomena that exhibits variation*.

Examples of random experiments:

a) Tossing a coin. Record outcome as "head" or "tail.

b) Rolling a die. Record outcome as number of eyes facing up.

c) Tossing a coin until "head" appears. Record outcome as number of tosses.

d) Disease. Record outcome as "recovered", "chronic ill" or "dead".

e) Measurement of haemoglobin concentration in blood.

In each of the examples, the experiment is described in terms of what aspect of the result is to be recorded.

The outcome of an individual random experiment cannot be predicted with certainty, but the set of all possible outcomes is known in advance.

> **The set of all possible distinct outcomes of an experiment is called the *sample space* of outcomes.**
> **Each distinct outcome is called a *simple event, an elementary outcome, or an element of the sample space.***

Statistical regularity

A random experiment can, in principle, be repeated numerous times under the same conditions. The outcomes of individual experiments must be independent, and must in no way be affected by any previous outcome.

Suppose a random experiment, repeatedly tossing of a coin, say. *A* is the event "head". Tossing *n* times results in "head" n_A times. Empirically, the relative frequency of *A*, n_A/n, will settle close to a certain number *p* if *n* increases, and in a repeated series n_A/n will again settle close to *p*. The relative frequency of *A* is an e*stimate* of the probability of "head in a single toss, and it is obvious that $0 \le \frac{n_A}{n} \le 1$.

Sample space, elementary outcome, and event

The collection (set) of all elementary outcomes in a random experiment is called the *sample space* (no.: utfallsrommet) , usually written *S*. Every simple event is an *element* of the sample space. In the examples above:

a) S = {Head, Tail}

b) S = {1, 2, 3, 4, 5, 6}

c) S = {1, 2, 3, 4, 5, 6, 7, 8, …….}

d) S = {Recovered, Chronically ill, Dead}

e) S = {7 – 20 g/dl} (approximately)

Note that the sample space can consist of measurable elements (numbers), or of non-measurable, qualitatively different elements.

Sample space – discrete or continuous

A sample space consisting of a finite or a *countably infinite* number of elements is called a *discrete* sample space.
When the sample space includes all the numbers in an interval of the real line, it is called a *continuous* sample space.

In examples *a, b,* and *d* the sample space has a finite number of elements.
In example c the sample space consists of all natural numbers, thereby being infinite, but the elements can be numbered. All situations above are examples of discrete sample spaces.

In example *e* we have a continuous sample space.

Event
The elements of the sample space constitute the ultimate breakdown of the results into distinct possibilities (*elementary outcomes*). Several elementary outcomes may often exhibit some common descriptive feature, and taken together, they constitute a *composite event,* or only *event,* having the stated feature. Suppose, for example, that the feature of interest when rolling a die is whether at least 4 eyes face up. Then the event is the set A={4, 5, 6}, i.e. a *subset* of the sample space S. The event A occurs if the die shows 4, 5, or 6 eyes.
When the sample space is finite, any subset of the sample space may constitute an event.

Random variable

A random variable, usually expressed as (capital) X, is a *function* of the outcome in a random experiment. For each elementary outcome $X(e)$ takes a distinct numerical value. X may have the same value for different elementary outcomes, but can have strictly one value for each outcome.

Example: Tossing a coin twice. As an example, the (elementary) outcome HT means "head" in the first toss and "tail" in the second toss, and so one.
The sample space is S = {HH, HT, TH, TT} with elementary outcomes e_1 = HH, e_2 = HT, e_3=TH and e_4 = TT. If the numbers of heads is the *event* of interest, we get:

$X(e_1)=2,\ X(e_2)=1,\ X(e_3)=1,\ X(e_4)=0$

Usually, the argument e is omitted in the expression of a random variable. In the example above X will be defined as:

$$X = \begin{cases} 0, & e = TT \\ 1, & e = TH \text{ or } HT \\ 2, & e = HH \end{cases}$$

Discrete random variable

A random variable that can assume only a finite number of values or possibly an infinite number of values that can be arranged in a sequence and counted is called a <u>discrete random variable.</u>

The probability distribution or simply, the distribution of a discrete random variable is a list of the distinct values x_i of X together with their associated probabilities $P(X = x_i)$. Often a formula can be used in place of a detailed list (see Binomial distribution later on).

Probability distribution of a discrete random variable – probability mass function

Rolling a fair die. Random variable X is number of eyes facing up. Possible values of X are $x=1, 2, 3,..., 6$. $P(X = x) = 0.5$ for all values of x (uniform distribution.)

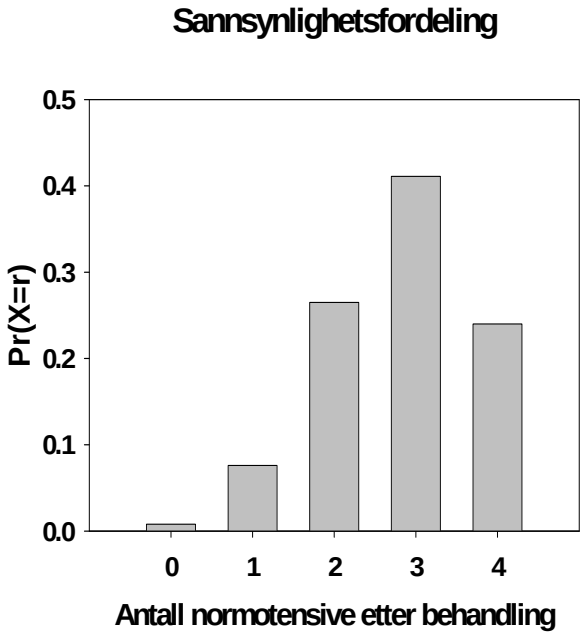
x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

Probability distribution – hypertension-control example

Ex. 4.6: Probability distribution for number of patients out of 4 patients achieving normotension after treatment

x	0	1	2	3	4
P(X=x)	0.008	0.076	0.265	0.411	0.240

Note that the probability for a distinct value *x* always is a number between 0 and 1, and that the all probabilities always add up to exactly 1.



The Expected value of a Discrete random variable

The expected value (expectation, "mean") is a measure of the "centre of gravity" of the probability mass distribution.

Imagine a metal block cut in the shape of the probability histogram. Then the expected value represents the point on the base at which the block will balance.

The expected value is calculated as a weighed mean of all possible vales of the random variable. The weights are the probabilities of each the different value of the variable.

$$\mu = E[X] = x_1 P(X = x_1) + x_2 P(X = x_2) + \ldots + x_n P(X = x_n) = \sum_{i=1}^{n} x_i \, P(X = x_i)$$

A special case: If the distribution is *uniform*, the expectation is equivalent to the arithmetic mean of all possible values of the random variable.

Expected value - example

Rolling a die, uniform distribution. *X* is possible number of eyes, *x* is possible value of *X*

x	1	2	3	4	5	6
P(X =x)	1/6	1/6	1/6	1/6	1/6	1/6

$$E[X] = \sum_{x=1}^{6} x \, P(X = x)$$

$$= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3)$$

$$+ 4 \cdot P(X = 4) + 5 \cdot P(X = 5) + 6 \cdot P(X = 6)$$

$$= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = \tfrac{21}{6} = 3.5$$

The expected value is the arithmetic mean of the possible values of the number of eyes, and corresponds to the mean number of eyes observed in (infinitely) repeated number of times.

Note that the expected value doesn't belong to the sample space.

Expected value - example

Ex. 4.6: Expected value of the number of patients out of 4 patients achieving normotension after treatment.

x	0	1	2	3	4
P(X=x)	0.008	0.076	0.265	0.411	0.240

$$\mu = E[X] = \sum_{x=0}^{4} x\,P(X=x)$$

$$= 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3) + 4 \cdot P(X=4)$$

$$= 0 \cdot 0.008 + 1 \cdot 0.076 + 2 \cdot 0.265 + 3 \cdot 0.411 + 4 \cdot 0.240 = 2.80$$

Note that this is the *population mean*. It is calculated without using real data, nor doing any experiment.

Variance of a Discrete random variable

The variance, denoted *Var(X)* or simply σ^2, is a measure of the deviation of *X* from μ, or equivalently a measure of spread in the distribution of *X*.

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$

Note that the variance is the expectation of $(X - \mu)^2$ and so can be expressed as

$$Var(X) = E\left[(X-\mu)^2\right] = E(X^2 - 2\mu X + \mu^2)$$

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

$$= E(X^2) - \mu^2 = E(X^2) - \left(E(X)\right)^2$$

Variance - example

Ex. 4.9: Variance of the number of patients out of 4 patients achieving normotension after treatment

x	0	1	2	3	4
P(X=x)	0.008	0.076	0.265	0.411	0.240

$$E(X) = \mu = 2.8$$

$$Var(X) = \sum_{x=0}^{4} (x-\mu)^2 \cdot P(X=x) = E(X^2) - \mu^2$$

$$= 0^2 \cdot P(X=0) + 1^2 \cdot P(X=1) + 2^2 \cdot P(X=2)$$

$$+ 3^2 \cdot P(X=3) + 4^2 \cdot P(X=4) - 2.8^2$$

$$= 1^2 \cdot 0.0076 + 2^2 \cdot 0.265 + 3^2 \cdot 0.411 + 4^2 \cdot 0.240 - 2.8^2 = 0.835$$

Standard deviation

The positive root of the variance is called the *standard deviation*:

$$SD(X) \equiv +\sqrt{Var(X)} = +\sqrt{\sigma^2} = \sigma$$

Example 4.9 (hypertension):

$$Var(X) = 0.835$$

$$SD(X) = \sqrt{0.835} = 0.914$$

Note that alike the variance, the standard deviation too is a non-negative quantity. Accordingly, it is nonsense to give the standard deviation as ± 0.914.

Probability distribution and frequency distribution

Example 4.8: Sample-frequency-distribution and the theoretical-probability distribution for the hypertension example. Each of 100 physicians treats 4 patients each.

Number of hypertonics under control = x	Probability distribution P(X=x)	Frequency distribution (observed)
0	0.008	0.00 =0/100
1	0.076	0.09 =9/100
2	0.265	0.240 =24/100
3	0.411	0.480 =48/100
4	0.240	0.190 =19/100

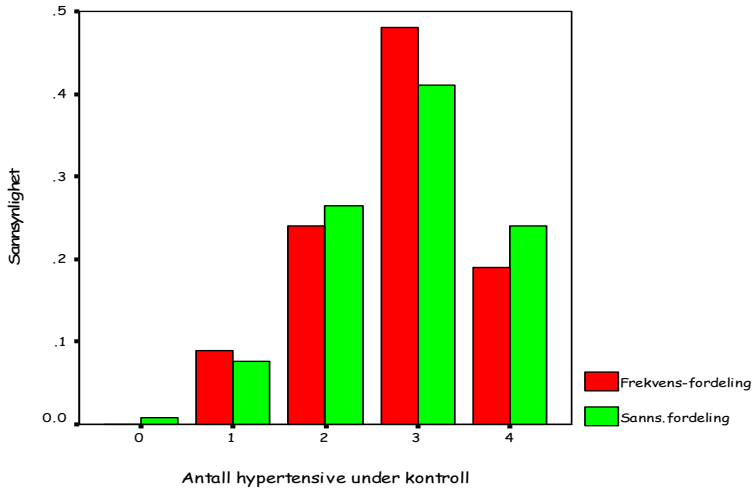
Expected value (estimated expectation) based on sample data:

$$\sum_{x=0}^{4} x\,P(X=x)$$

$$=0\cdot P(X=0)+1\cdot P(X=1)+2\cdot P(X=2)+3\cdot P(X=3)+4\cdot P(X=4)$$

$$=0+1\cdot 0.09+2\cdot 0.24+3\cdot 0.48+4\cdot 0.190=2.77$$

Comparison of the frequency and true probability distribution in the hypertension-control example



Counting and its use in uniform probability models

Uniform probability models are in particular useful in experiments where all elementary outcomes in the sample space *S* are equally likely. The probability of an event *A* is given by

$$P(A)=\frac{Number\ of\ elements\ in\ A}{Number\ of\ element\ in\ S}$$

When convenient methods of counting are available, we can omit listing all elements in *S*.

Example: There are *r* red balls and *b* black balls in a box (urn). One ball is drawn at random. What is the probability to draw a red ball? All balls have equal probability to be drawn (uniform probability).

$$P(Red\ ball)=\frac{Number\ of\ red\ balls}{Number\ of\ balls}=\frac{r}{r+b}$$

Combinations and permutations

Combinations of experiments

Two examples:

1) **A cafeteria offers a dinner consisting of 3 dishes with possibility to choose one starter among 3, one main dish among 6, and one dessert among 4. How many different dinners are possible? Answer:** $3\cdot 6\cdot 4=72$.

2) **Football betting. Twelve matches, each with 3 possible outcomes. Number of possible rows in a betting slip (tippekupong):** $3\cdot 3\cdot 3\cdot\ldots\cdot 3=3^{12}=531441$.

Rule:
When an experiment consists of K parts, each having k_i distinct results, and if we wish to combine results, one form each part, then the total number of possible result of the experiment is $k_1\cdot k_2\cdot,,,\cdot k_K$.

Sampling with replacement when order matters

In a box (urn) there are *n* balls numbered from *1* to *n*. A ball is selected *with replacement* *k* times from the box. In how many ways can that be done?

At the first drawing there are *n* possibilities. The ball is put back in the box, and in the next selection there are again *n* possibilities, and so one. This is done *k* times. In particular, the same ball may be selected several times. The total number of combinations of *k* numbered balls will be $n \cdot n \cdot n \cdot \ldots \cdot n = n^k$ **.**

Rule:

When k objects is drawn with replacement from a collection of n distinct objects the total number of combinations is n^k ***.***

Example: In the international alphabet (A, B, C,…,Z) there are 26 different letters. For a car registration number with 2 letters $26^2 = 676$ **pairs of letters can be created if the same letter is allowed to appear both times.**

Sampling without replacement when order matters

In a box (urn) there are *n* balls numbered from 1 to n. A ball is selected *without replacement* *k* times from the box. In how many ways can that be done?

At the first selection there are *n* possibilities, at the next one *(n-1)* possibilities, thereafter *(n-2)* and so on.

(n-(k-1)) balls remain in the box before the k^{th} **selection giving *(n-(k-1))=(n-k+1)* possibilities for the last selection. So, when the order of selection matters, the total number of combinations of *k* numbered balls is**

$$ {}_nP_k = n(n-1)\cdot\ldots\cdot(n-k+1) $$

Rule:

The number of possible orderings or arrangements of k objects selected from n distinct objects is ${}_nP_k = n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)$.

Example: From the international alphabet it can be created ${}_{26}P_4 = 26(26-1)\cdot\ldots\cdot(26-4+1) = 26\cdot 25\cdot 24\cdot 23 = 358800$ **different codes consisting of 4 different letters if the order matters.**

The rule above has as a consequence that if all balls , i.e. *k=n*, are drawn without replacement, it can be done in

$$ {}_nP_n = n(n-1)(n-2)\cdot\ldots\cdot 3\cdot 2\cdot 1 $$

different ways.

The expression $n(n-1)(n-2)\cdot\ldots\cdot 3\cdot 2\cdot 1$ **is written *n!* and read "n factorial" (no.: n fakultet). By definition, *0!=1.***

Corollary: The number of permutations of n objects is n!

Sampling without replacement when order doesn't matter

If *k* balls are drawn out of *n* balls without replacement, there are ${}_nP_k = n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)$ **ordered samples of *k* balls. These *k* balls can afterwards be permuted *k!* times. If the order doesn't matter, but only which balls are drown (as in Lotto), the question to be addressed is how many non-ordered samples of k balls can appear. We call this number** ${}_nC_k$**. Every non-ordered sample can be permuted *k!* times giving the equality**

$$ {}_nP_k = {}_nC_k \cdot k! $$

This gives

$$ {}_nC_k = \frac{{}_nP_k}{k!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} $$

$$ = \frac{n(n-1)(n-2)\cdots(n-k+1)\cdot(n-k)!}{k!\cdot(n-k)!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} $$

The shorthand notation $\binom{n}{k}$ is called the *binomial coefficient,* read *"n over k"* or *"k out of n"*, and must no be confused with $\frac{n}{k}$.

Rule 4: The number of possible collection of k object chosen from a group of n distinct objects is

$$ {}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} $$

Example: Lotto. Seven numbered balls out of 34 are drown without replacement. Number of possible combinations:

$$ {}_{34}C_7 = \binom{34}{7} = \frac{34!}{7!(34-7)!} = \frac{34!}{7!\cdot 27!} = \frac{34\cdot 33\cdot \ldots \cdot 28}{7!} = 5379616 $$

Binomial series of trials (no.: binomisk forsøksrekke)

An experiment characterized by

i) **The experiment consists of *n* independent trials**
ii) **In each trial it recorded whether one specific event *A* occurs or not (A*)**
iii) ***P(A)= p* (success probability) in all trials (implies that $P(A^*)=1-p=q$)**

is called a *binomial series of trials.*

Examples of binomial series of trials:

1) **Tossing a coin *n* times recording number of heads ("head" is success)**
2) **Sowing *n* seeds and recording how many seeds sprouting after a certain time ("sprouting" is success)**
3) **Recording how many childbirths resulting in girl ("girl" is success)**

Binomial distribution

Let *X* (random) be the number of *n* independent trials resulting in the event *A*. *P(A)=p*, P(A*)=*1-p*. As an example, if n=3, we investigate the distribution of X. The sample space S has 2^3 =8 elementary outcomes

Sample space (S)	Elementary outcomes	$P(e_i)$	X
A*A*A*	e_1	$(1-p)^3$	0
A*A*A	e_2	$p(1-p)^2$	1
A*A A*	e_3	$p(1-p)^2$	1
A A*A*	e_4	$p(1-p)^2$	1
A A A*	e_5	$p^2(1-p)$	2
A A*A	e_6	$p^2(1-p)$	2
A *A A	e_7	$p^2(1-p)$	2
A A A	e_8	p^3	3

Probability mass distribution :

$$ P(X=0)=(1-p)^3,\ P(X=1)=3(1-p)^2p,\ P(X=2)=3(1-p)p^2,\ P(X=3)=p^3 $$

Probability mass function:

$$ P(X=k) = \binom{3}{k} p^k (1-p)^{3-k}, \quad k=0, 1, 2, 3 $$

In general, in *n* trials the probability to get *k* outcomes *A*

$$ P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} $$

Example: Suppose that the probability of "girl" in a single childbirth is *p = 0.5* . What is the probability of 9 girls in 18 independent childbirths?

Binomial model:

***n=18* independent trials**
***A* = "girl" in a trial**
***P(A)=p* =0.5 in each trial**
***X* = number of *A* in *n=18* trials**

$$ P(X=9) = \binom{18}{9} 0.5^9 (1-0.5)^{18-9} = \binom{18}{9} 0.5^9 0.5^9 = \binom{18}{9} 0.5^{18} = 0.186 $$

In practice, it is seldom of interest to calculate the probability of exactly *k* successes, but rather probabilities of the type $P(X \le k)$ or $P(X > k)$.
We the have the formula

$$P(X \le k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \qquad i=0, 1, 2,\ldots,k$$

Example, cont.: A fortune-teller claims she is clairvoyant and claims she is able to predict the sex in a childbirth. She predicts correctly in 14 of 18 cases. Does she simply guess?

> **Question to be addressed: What is the probability to predict correctly in *at least* 14 out of 18 times *given that she simply guesses, i.e. p=0.5*?**

$$P(X \ge 14) = P(X=14) + P(X=15) + P(X=16) + P(X=17) + P(X=18)$$
$$= 0.0117 + 0.0031 + 0.0006 + 0.0001 + 0.0000 = 0.0155 \quad \text{(fra tabell)}$$

The probability to predict correctly just by chance at least 14 out of 18 times is "low", "unlikely", thus indicating supernatural power.

Expectation and variance of indicator variable

Trial with two possible outcomes

Indicator variable: $I = \begin{cases} 1 \text{ at outcome } A, & P(A) = p \\ 0 \text{ at outcome } A^*, & P(A^*) = 1-p = q \end{cases}$

$$E(I) = 1 \cdot P(I=1) + 0 \cdot P(I=0) = p$$

$$E(I^2) = 1^2 \cdot P(I=1) + 0 \cdot P(I=0) = p$$

$$Var(I) = E(I^2) - \big(E(I)\big)^2 = p - p^2 = p(1-p) = pq$$

Expectation and variance in binomial distribution

Number of A in n independent trials: $X = I_1 + I_2 + .. + I_n = \sum_{i=1}^{n} I_i$

Expectation:
$E(X) = E(I_1) + E(I_2) + \ldots + E(I_n) = p + p + \ldots + p = np$

Independent trials (independence, no covariances) result in:

$$Var(X) = Var(I_1) + Var(I_2) + \ldots + Var(I_n) = pq + pq + \ldots + pq = npq = np(1-p)$$

Poissonfordeling – ekstrastoff (bonus, ikke pensum)

Noen ganger har man å gjøre med en binomisk forsøksrekke der *n* er stor og *p* er svært liten. Forventet antall enkeltutfall er som alltid $\mu = np$. Under betingelsene ovenfor kan det da vises at

$$P(X=k) = \frac{\mu^k}{k!} e^{-\mu}$$

Dette er *poissonfordelingen*, og det kan videre vises at $E(X) = Var(x) = \mu$

Eksempler på poissonfordelte hendelser:

i) **Utsending partikler fra en radioaktiv kilde over et visst tidsrom**
ii) **Antall kollisjoner i en sterkt trafikkert veikryss over et visst tidsrom**
iii) **Antall sjeldne celler i et synsfelt under mikroskopet**
iv) **Antall tilfeller av en sjelden sykdom i en stor populasjon over en viss tidsperiode**