Plan for the second part of the course:

(1) Elements of probability and statistics
(2) Time series analysis
(3) Elements of optimization theory

1. Lectures 9,10: Elements of probability and statistics

Plan:

- ⋄ Basic concepts of probability
- ⋄ Basic elements of statistics

1.1. Basic elements of probability theory ....

- ⋄ Make an experiment and measure something. The measured quantity is a *sample*. The set of all possible outcomes of a measurement is a *sample space*, S.

Examples of experiments:

- ⋄ Throwing a six-sided dice with the results: $1, 2, 3, 4, 5, 6$
- ⋄ Flipping a coin with the results: H, T
- ⋄ Flipping two coins with possible results: (H, T), (H, H), (T, H), (T, T)

For one throw of a six-sided dice, the sample space is the set $\{1, 2, 3, 4, 5, 6\}$. Whatever comes up in a given throw is a sample.

When flipping a coin, the sample space consists of heads and tails: $S = \{H, T\}$.

- ⋄ Any subset of the sample space is an *event*, E.
  - If one throws two dice, then the set $E = \{(1, 3), (2, 2), (3, 1)\}$ denotes the event "the sum equals 4".
  - If one flips two coins, then the set $E = \{(H, H), (H, T)\}$ denotes the event "the first coin is heads".

- ⋄ The *union* $E_1 \cup E_2$ consists of all the events that are either in E_1 or in E_2, where E_1 and E_2 are two events from S.
- ⋄ The *intersection* $E_1 \cap E_2$ consists of all the events that are both in E_1 and in E_2.

For example, if we toss two dice thrice and $E_1 = \{1, 1, 3\}$ and $E_2 = \{2, 2, 3\}$, then $E_1 \cup E_2 = \{1, 2, 3\}$ and $E_1 \cap E_2 = \{3\}$.

- ⋄ We define a *complement* of an event E^c such that E^c includes all the events from S that are not in E.
  Thus, $E \cup E^c = S$, while $E \cap E^c = \emptyset$.

- ⋄ Now we define the *probability* $P(E)$ of event E occurring as a function on the event space satisfying:
  - $0 \leq P(E) \leq 1$
  - $P(S) = 1$
  - $P\left(\cup_{i=1}^{N} E_i\right) = \sum_{i=1}^{N} P(E_i)$ if events E_i are mutually exclusive, i.e. $E_i \cap E_j = \emptyset$ if $i \neq j$.

Example 1. As an example, for flipping a fair coin,

$$P(\{H\}) = \frac{1}{2}, \quad P(\{T\}) = \frac{1}{2}.$$

If two coins are flipped, then

$$P(\{H, H\}) = \frac{1}{4}.$$

When throwing a dice, the probability that an even number appears is

$$P(E) = P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Since E and E^c are mutually exclusive, then $P(E \cup E^c) = P(E) + P(E^c) = 1$.

For two arbitrary events E_1 and E_2, the probability of their union is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2),$$

where we have to subtract the probability of intersection as occurring twice in calculating $P(E_1)$ and $P(E_2)$. Of course, for mutually exclusive events, the last term disappears.

Example 2. For another example, consider tossing two fair coins. The events in the sample space $S = \{(H,H), (H,T), (T,H), (T,T)\}$ are all equally likely to occur with probability 1/4.

Define $E_1 = \{(H,H), (H,T)\}$ as the event in which the first coin is the heads. Let $E_2 = \{(H,H), (T,H)\}$ be the event such that the second coin is the heads.

Then,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \\ = \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{4} + \frac{1}{4}\right) - \frac{1}{4} = 1 - \frac{1}{4} = \frac{3}{4},$$

which can also be found directly from

$$P(E_1 \cup E_2) = P(\{(H,H), (H,T), (T,H)\}) = \frac{3}{4}.$$

- ◇ *Conditional probability, $P(E_2|E_1)$,* is defined as the probability of event E_2 given we know the probability of event E_1.
  This is relevant when the two events are not independent. Some knowledge about E_1 should help in learning about E_2. That is, we want to predict the probability of some event in the case when the probability of a related event E_1 is known.

Example 3. Flip two coins. What is the probability that both results are heads H, given that at least one is H?

Solution. Sample space is $S = \{(H,H), (H,T), (T,H), (T,T)\}$ with every outcome equally likely.

Let E_2 denote the event that both coins are H.

Let E_1 denote the event that at least one of the coins is H.

Then,

$$P(E_2|E_1) = \frac{P(E_2 \cap E_1)}{P(E_1)} \\ = \frac{P(\{H,H\})}{P(\{H,H\}, \{H,T\}, \{T,H\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Example 4. We have cards numbered 1 to 10. Put them in a hat, mix, and draw one card. We are told that the drawn card has a number which is at least 5. What is the probability that the number is 10?

Solution. Sample space is $S = \{1, 2, ..., 10\}$ with every outcome equally likely (probability 1/10).

Let E_2 denote the event that the card is 10 (probability 1/10).

Let E_1 denote the event that the card is at least 5 (probability 6/10).

Then,

$$\begin{aligned} P\left(E_2|E_1\right) &= \frac{P\left(E_2 \cap E_1\right)}{P\left(E_1\right)} \\ &= \frac{1/10}{6/10} = \frac{1}{6}. \end{aligned}$$

◇ *Independent events.*

The probability of the intersection of two events $P\left(E_2 \cap E_1\right)$ is related to the conditional probability as

$$P\left(E_2 \cap E_1\right) = P\left(E_2|E_1\right) P\left(E_1\right),$$

which is a statement of the fact that the probability of occurrence of both events E_2 *and* E_1 is the product of the conditional probability of E_2 occurring given E_1 took place and of the probability that E_1 took place.

Two events are independent if

$$P\left(E_2 \cap E_1\right) = P\left(E_2\right) P\left(E_1\right).$$

As a consequence, for independent events

$$P\left(E_2|E_1\right) = P\left(E_2\right),$$

which simply states that there is really no condition on the probability of event E_2.

Example 5. Throw two six-sided dice. Each outcome has a probability of 1/36.

Let E_1 denote the event that the sum of the dice is 6 (probability 5/35: 15, 24, 33, 42, 51).

Let E_2 denote the event that the first die is 4 (probability 1/6).

Let E_3 denote the event that the sum of the dice is 7 (probability 1/6: 16, 25, 34, 43, 52, 61).

Then:

$$\begin{aligned} P\left(E_1 \cap E_2\right) &= P\left(\{(4,2)\}\right) = \frac{1}{36} \\ P\left(E_1\right) P\left(E_2\right) &= \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216} \end{aligned}$$

$$\begin{aligned} P\left(E_3 \cap E_2\right) &= P\left(\{(4,3)\}\right) = \frac{1}{36} \\ P\left(E_3\right) P\left(E_2\right) &= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}. \end{aligned}$$

We conclude that E_1 and E_2 are not actually independent, because for the sum to be 6, it is not irrelevant that the first die is 4. The first die cannot be 6.

On the other hand, E_2 and E_3 are independent because for the sum to be 7, the first die can be anything, 4 or not 4.

⋄ *Bayes formula.*

To relate the conditional probability to those of E_1 and E_2, we reason as follows. Since the two events are not independent, then $E_1 \cap E_2$ is not empty. The probability of the event in that intersection is the product of $P(E_1)$, the probability that E_1 occurs and $P(E_2|E_1)$, the probability that E_2 occurs given the E_1 occurred. Therefore, $P(E_1 \cap E_2) = P(E_1) P(E_2|E_1)$, from which we get

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}. \tag{1.1}$$

Of course, we could have reasoned equivalently that the probability of the event in $E_1 \cap E_2$ is the product of $P(E_2)$ and $P(E_1|E_2)$, in which case we would have reached a different formula

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}. \tag{1.2}$$

Division of one by the other of these formulas gives us

$$P(E_2|E_1) = \frac{P(E_1|E_2) P(E_2)}{P(E_1)}, \tag{1.3}$$

which is the Bayes formula.

More generally, suppose we have mutually exclusive events $E_1, E_2, \ldots, E_N$ such that $\cup E_i = S$ and that exactly one of the events occurs. Any event H will occur together with one of the events E_i,

$$P(H) = \sum_{i=1}^{N} P(H \cap E_i) = \sum_{i=1}^{N} P(H|E_i) P(E_i).$$

This means that the probability of H is the weighted average of conditional probabilities $P(H|E_i)$ of H given E_i occurs with the weights equal to the probabilities of E_i occurring.

Using this, the Bayes formula can be written as

$$P(E_i|H) = \frac{P(H \cap E_i)}{P(H)} = \frac{P(H|E_i) P(E_i)}{\sum_{i=1}^{N} P(H|E_i) P(E_i)}. \tag{1.4}$$

Let's look at some examples of its use.

Example 6. We have a coin and two boxes with white and black balls. The box 1 contains 2 white and 7 black balls, box 2 contains 5 white and 6 black balls. We flip the coin and if heads – take a ball from box 1, if tails – from box 2.

The question: What is the probability that the outcome of the toss was heads if a white ball was selected?

Solution. To answer this question, define W as the event that a white ball was taken, and H that heads was tossed. Then we are asked to find the conditional probability $P(H|W)$. It is calculated as follows:

$$\begin{aligned} P(H|W) &= \frac{P(W|H) P(H)}{P(W)} \\ &= \frac{P(W|H) P(H)}{P(W|H) P(H) + P(W|T) P(T)} = \\ &= \frac{\frac{2}{9} \cdot \frac{1}{2}}{\frac{2}{9} \cdot \frac{1}{2} + \frac{5}{11} \cdot \frac{1}{2}} = \frac{22}{67}. \end{aligned}$$

Example 7. Now we consider a medical example. Suppose a lab test is 95% effective in determining a certain disease when applied to sick people. It can also give a false positive with 1% of healthy people. Suppose we know that 0.5% of the population is actually sick with this disease. The question is: if a test is positive, what is the probability that the person is actually sick?

Solution. To answer, let D be the event that the person tested has the disease, and that E is the event that the test was positive. What we need to find is $P\left(D|E\right)$.

Using Bayes,

$$P\left(D|E\right)=\frac{P\left(E|D\right)P\left(D\right)}{P\left(E|D\right)P\left(D\right)+P\left(E|D^c\right)P\left(D^c\right)},$$

where D^c is the event that the person is healthy.

Here

$$\begin{aligned}
P\left(E|D\right)=0.95 - &\ \text{probability of a positive test given the person is sick}\\
P\left(D\right)=0.005 - &\ \text{probability of a person being sick}\\
P\left(E|D^c\right)=0.01 - &\ \text{probability of a positive test given the person is healthy}\\
P\left(D^c\right)=0.995 - &\ \text{probability that a person is healthy}
\end{aligned}$$

Then

$$P\left(D|E\right)=\frac{0.95\cdot 0.005}{0.95\cdot 0.005+0.01\cdot 0.995}\approx 0.323.$$

That is, the test is pretty bad despite its nice look. A good test would give close to 100%, and the fact that the test is good in 95% of cases with sick people means nothing here. The 1% false results on healthy people make this test basically useless.

With the form (1.4), we can analyze the following situation. Suppose we have an experimental sample, and we know that we have some probability distribution depending on parameter ϑ, which we consider as a random variable. That is, we do not know what value ϑ takes, but have an idea *a priori* that the set of possible values of ϑ can be viewed as a random variable with a density g_{old} (the *"prior" distribution*). Now that we have made an experiment and have some data, we should be able to learn from these data how to improve the prior ideas and obtain a new density g_{new}, the *"posterior" distribution*, which improves the "prior" in light of the data.

Example 8. (Chorin & Hald) Let x_1 and x_2 be two *i.i.d.* (*independent and identically distributed*) random variables with

$$x_1, x_2 = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

The sum $x_1 + x_2$ will take values

$$x_1 + x_2 = \begin{cases} 2 & \text{with probability } p^2 \\ 1 & \text{with probability } 2p\left(1-p\right) \\ 0 & \text{with probability } \left(1-p\right)^2. \end{cases}$$

Suppose now that the "prior" distribution is such that $p = 1/4$ with probability $P\left(p = 1/4\right) = 1/4$ and $p = 1/2$ with probability $P\left(p = 1/2\right) = 3/4$. We make an experiment and find that $x_1 + x_2 = 1$. Now we use Bayes to improve these distributions of values of p.

Let:
A be the event that $x_1 + x_2 = 1$,
E_1 be the event that $p = 1/4$, and
E_2 be the event that $p = 1/2$.
Note $E_1 \cup E_2 = S$ as their probabilities add to 1.

Then, the probability that $p = 1/4$ in light of this new information in the event A, is obtained using

$$P\left(E_1|A\right) = \frac{P\left(A|E_1\right)P\left(E_1\right)}{P\left(A|E_1\right)P\left(E_1\right) + P\left(A|E_2\right)P\left(E_2\right)}.$$

Here

$$\begin{aligned} P\left(A|E_1\right) &= 2\cdot\frac{1}{4}\cdot\frac{3}{4} - \text{ probability of } x_1 + x_2 = 1 \text{ given } p = \frac{1}{4} \\ P\left(E_1\right) &= \frac{1}{4} - \text{ probability of } p \text{ being equal to } \frac{1}{4} \\ P\left(A|E_2\right) &= 2\cdot\frac{1}{2}\cdot\frac{1}{2} - \text{probability of } x_1 + x_2 = 1 \text{ given } p = \frac{1}{2} \\ P\left(E_2\right) &= \frac{3}{4} - \text{probability of } p \text{ being equal to } p = \frac{1}{2}. \end{aligned}$$

Then we find

$$P\left(E_1|A\right) = \frac{\frac{3}{32}}{\frac{3}{32} + \frac{3}{8}} = \frac{1}{5},$$

which differs from 1/4 of the "prior". Why is it smaller?

1.2. **Mean and variance.** If we observe N samples of variable x, then the sample mean is the average of those samples:

$$m = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

In terms of probabilities p_i of observing each sample x_i, we have the expected value:

$$m = E\left[x\right] = \sum_{i=1}^{N} p_i x_i = p^T x.$$

$E\left[x\right]$ tells what to expect, and μ tells what we got.

The *variance* σ^2 measures the expected distance squared from the expected mean $E\left[x\right]$.

On the other hand, the *sample variance* S^2 measures the actual distance squared from the actual sample mean.

Standard deviation is σ or S, and both measure how spread the variables are relative to their mean.

To know σ and $E\left[x\right]$ requires knowing probabilities, while S and m require measured samples.

Note that

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - m\right)^2 = \frac{1}{N-1}\left[\sum x_i^2 - 2m\sum x_i - Nm^2\right] =$$

$$= \frac{1}{N-1}\left[\sum x_i^2 - 2mNm - Nm^2\right] = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - Nm^2\right].$$

Similarly, with probabilities

$$\sigma^2 = E\left[x^2\right] - \left(E\left[x\right]\right)^2 = \sum p_i x_i^2 - \left(\sum p_i x_i\right)^2.$$

Two important theorems of probability theory are the *law of large numbers* and the *central limit theorem*. Loosely, the first says that the probability of an event is the proportion of times it occurs in many trials, when the number of trials tends to infinity. The second theorem states that for a large enough sampling of data, any distribution of random variables will tend to the normal distribution. This is stated more precisely later.

Theorem 9. *(Strong law of large numbers). If $x_1, x_2, \ldots x_n$ is a sequence of independent random variables having a common distribution and $E\left[x_i\right] = \mu$. Then, with probability 1,*

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \to \mu \text{ as } n \to \infty.$$

To illustrate, suppose a sequence of independent trials is performed, and E is a fixed event with $P\left(E\right)$ its probability in any given trial. Let

$$x_i = \begin{cases} 1, & \text{if } E \text{ occurs in trial } i \\ 0, & \text{if } E \text{ does not occur in trial } i. \end{cases}$$

The theorem states that the average tends to the probability $P\left(E\right)$:

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \to E\left[x\right] = P\left(E\right).$$

Since $x_1 + x_2 + \ldots + x_n$ is the number of times the event occurred, then $P\left(E\right)$ is just equal to the limiting proportion of times the event occurs.

1.3. **Discrete probability distributions.** The main discrete distributions are:

- ◇ *Binomial* – comes from tossing a coin n times.
- ◇ *Poisson* – has to do with rare events.

Binomial. Each trial has a binary outcome: $x = 1$ (success) or $x = 0$ (failure), with probability p and $q = 1 - p$, respectively.

The probability of k successes in n trials is

$$p_{k,n} = C_k^n p^k (1-p)^{n-k}, \quad C_k^n = \frac{n!}{k!\,(n-k)!},$$

because out of n trials one can choose C_k^n different sequences of k successes.

For example, if we have $n = 3$ trials and want exactly $k = 2$ successes, then the possibilities are: 110, 101, 011, each with probability $p^2(1-p)$. Then $C_2^3 = \frac{3!}{2!1!} = 3$ and $p_{2,3} = 3p^2(1-p)$.

The mean value of x in n trials is

$$\mu_n = \sum_{k=1}^{n} k p_{k,n} = \sum_{k=1}^{n} \frac{kn!}{k!\,(n-k)!} p^k (1-p)^{n-k} =$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-1-(k-1))!} p^{k-1} (1-p)^{n-1-(k-1)} =$$

$$= np \underbrace{\sum_{k=0}^{n-1} \frac{(n-1)!}{k!\,(n-1-k)!} p^k (1-p)^{n-1-k}}_{=(p+(1-p))^{n-1}=1 \text{ by binomial formula for } (a+b)^n} = np$$

while the variance in n trials is $\sigma_n^2 = np(1-p)$.

Poisson. We take binomial distribution and take a limit: $n \to \infty$, $p \to 0$, but demand that $np \to \lambda < \infty$. That is we look at a large number of unlikely events: $n \gg 1$ and $p \ll 1$. Applications include rare events over long times.

If we have k successes in n trials of binomial distribution, then

$$p_{k,n} = C_k^n p^k (1-p)^{n-k} \overset{p=\lambda/n}{=} C_k^n \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= C_k^n \frac{\lambda^k}{n^k} \frac{1}{(1-\lambda/n)^k} \left(1 - \frac{\lambda}{n}\right)^n \underset{n\to\infty}{\to} \frac{\lambda^k}{k!} e^{-\lambda},$$

as

$$C_k^n \frac{1}{n^k} = \frac{n!}{k!\,(n-k)!} \frac{1}{n^k} = \frac{n!}{n^k (n-k)!} \frac{1}{k!} = \frac{n(n-1)\ldots(n-(k-1))}{n^k} \frac{1}{k!} \to \frac{1}{k!} \text{ as } n \to \infty \text{ and } k \text{ fixed},$$

since $n(n-1)\ldots(n-(k-1)) \sim n^k$ as $n \gg 1$, and

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda} \text{ as } n \to 8.$$

Thus the probability of k successes in n trials is given by

$$P_k = \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda = np$ and p is the probability of a single event. The Poisson distribution has the mean $\mu = \lim_{n\to\infty} np = \lambda$ and variance $\sigma^2 = \lim_{n\to\infty} np(1-p) = \lambda$.

1.4. **Continuous probability distributions.** Main continuous distributions are:

- ◇ *Exponential* – forgets the past
- ◇ *Gaussian* (=*Normal*) – averages of many tries
- ◇ *Log-normal* – logarithm of the random variable has normal distribution
- ◇ *Chi-squared* – distance squared in n dimensions
- ◇ *Multivariate Gaussian* – probabilities for a vector

If a random variable takes on continuous values, than we need a *probability distribution* $p(x)$ to tell us how likely it is to find the value of the variable near x. More precisely,

$$p(x)\,dx$$

is the probability to find the random variable in the range from x to $x+dx$.

The mean and the variance are given by

$$m = E[x] = \int xp(x)\,dx,$$
$$\sigma^2 = E\left[(x-m)^2\right] = \int (x-m)^2 p(x)\,dx.$$

The probability that the variable x takes values in the interval (a,b) is given by

$$P(a<x<b) = \int_a^b p(x)\,dx.$$

Clearly, the probability for x to take a particular value $x=a$ is 0. And

$$P(-\infty<x<\infty) = \int_{-\infty}^{\infty} p(x)\,dx = 1.$$

Cumulative distribution is given by

$$\Phi(x) = \int_{-\infty}^{x} p(x)\,dx.$$

Uniform distribution. For example, if x is equally likely to take any value in $[0,a]$, then we have a uniform distribution with $p=1/a$. In that case, $m=a/2$ and $\sigma^2 = \int_0^a \frac{1}{a}\left(x-\frac{a}{2}\right)^2 dx = \frac{a^2}{12}$.

Exponential distribution. The pdf (*probability density function*) is

$$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Then the mean is

$$\mu = \int_0^\infty xp(x)\,dx = \frac{1}{\lambda}$$

and the variance is

$$\sigma^2 = \int_0^\infty \left(x-\frac{1}{\lambda}\right)^2 p(x)\,dx = \frac{1}{\lambda^2}.$$

The *cumulative distribution* is

$$F(t) = \int_0^t p(x)\,dx = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t},$$

and it gives *the probability that an event occurs before time t.* In this context, the mean is the *average waiting time.*

The exponential distribution has *no memory* in the sense that the probability of waiting for at least y hours more is independent of having already waited x hours. That is

$$P\{(t>x+y)\,|\,(t>x)\} = P\{(t>y)\}.$$

This follows from

$$\frac{\int_{x+y}^{\infty} \lambda e^{-\lambda t} dt}{\int_{x}^{\infty} \lambda e^{-\lambda t} dt} = e^{-\lambda y}.$$

Gaussian (Normal) distribution. Gaussian (normal) distribution is given by

$$p\left(x\right) = N\left(m, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(x-m\right)^2}{2\sigma^2}\right]$$

with the mean m and variance σ. The so-called "standard normal distribution" has $m = 0$ and $\sigma = 1$, and

$$p = N\left(0,1\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right].$$

Some common numbers associated with the normal distribution:

◇ Between $-\sigma$ and σ, the area under $p\left(x\right)$ is 0.67.

◇ Between -2σ and 2σ, the area under $p\left(x\right)$ is 0.95.

◇ Between -3σ and 3σ, the area under $p\left(x\right)$ is 0.997.

Cumulative distribution of the Gaussian is

$$\Phi\left(x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right],$$

where the error function is defined as

$$erf\left(s\right) = \frac{2}{\sqrt{\pi}} \int_{0}^{s} e^{-\tau^2} d\tau.$$

Log-normal distribution

◇ The distribution of x is *log-normal* if $y = \ln x$ is normal, which requires $x > 0$. Since y is normally distributed, then

$$p\left(y\right) dy = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right] dy =$$

$$\frac{1}{\sqrt{2\pi\sigma^2} x} \exp\left[-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}\right] dx = p\left(x\right) dx,$$

which shows the distribution of the log-normal variable x.

◇ If we have several variables, x_1, x_2, $\ldots x_n$, each normally distributed with mean 0 and variance 1, then their sum $x_1 + x_2 + \ldots + x_n$ is also normally distributed with mean 0 and variance n. Indeed, the mean is $E\left[\sum x_i\right] = \sum E\left[x_i\right] = 0$, and the variance is found from $E\left[\sum x_i^2\right] = \sum E\left[x_i^2\right] = n\sigma^2 = n$.

Chi-squared distribution

Start with a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. Then the variable $s = x^2$ is distributed with chi-squared distribution.

Question: what is the probability that s is between y and $y + dy$? For this to happen, we must have $x = \sqrt{s}$ to be between $\sqrt{y}$ and $\sqrt{y+dy}$ or $-\sqrt{y+dy}$ and $-\sqrt{y}$. These two are equally likely, hence a factor of two will show up. We will also use $\sqrt{y+dy} = \sqrt{y} + \frac{1}{2}\frac{dy}{y} + ...$ Then

$$P\left\{s \in \left(y, y+dy\right)\right\} = 2P\left\{\sqrt{s} \in \left(\sqrt{y}, \sqrt{y} + \frac{1}{2}\frac{dy}{y}\right)\right\} = \frac{2}{\sqrt{2\pi}} e^{-\left(\sqrt{y}\right)^2/2} \frac{dy}{2\sqrt{y}}.$$

Therefore, the distribution of $s = x^2$ is

$$p_1(s) = \frac{1}{\sqrt{2\pi s}} e^{-s/2}, s > 0 \quad \chi_1^2 \text{ distribution} .$$

Consider the sum of two squares, $s_2 = x_1^2 + x_2^2$, with x_1 and x_2 being standard normal variables with mean 0 and variance 1.

Then, the probability (distribution) p_2 of s_2 is the sum of probabilities of x_1^2 being s and x_2^2 being $s_2 - s$, that is it is the integral

$$p_2(s_2) = \int_0^{s_2} p_1(s)\, p_1(s_2 - s)\, ds,$$

which is the *convolution* of p_1 with p_1. Using the previous equation for p_1, we get

$$p_2(s_2) = \frac{1}{2} e^{-s_2/2},$$

which is just the exponential distribution.

Continuing with the same idea, we can find the distribution of sum of n squares as

$$p_n(s_n) = \int_0^{s_n} p_1(s)\, p_{n-1}(s_n - s)\, ds.$$

This integral has the form $C s_n^{(n-2)/2} \exp(-s_n/2)$ that has to satisfy $\int_0^\infty p_n ds_n = 1$, which gives the constant C. At the end of the day

$$p_n = C s_n^{(n-2)/2} \exp(-s_n/2) \quad \text{with}$$
$$C = \frac{1}{2^{n/2}\Gamma(n/2)}$$

is the probability distribution of $s_n = x_1^2 + x_2^2 + ... + x_n^2$ with all x_i standard normal.

The use of chi-squared can be illustrated as follows. Suppose, we manufacture some parts and then test them for accuracy. We have samples of thickness x_1, x_2, ..., x_n, for which we calculate the mean and variance:

$$\bar{x} = \frac{1}{n}\sum x_i, \quad S^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 .$$

Then, for $n = 1$ we have $\bar{x} = x_1$, $S = 0$, for $n = 2$ we have $\bar{x} = \frac{1}{2}(x_1 + x_2)$ and $S^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 = \frac{1}{2}(x_1 - x_2)^2$, and so on. The probability distribution of S^2 is given by p_{n-1}, χ_{n-1}^2 chi-square distribution.

Theorem 10. *(Central limit theorem).* Let $x_1, x_2, \ldots x_n$ be a sequence of independent and identically distributed (i.i.d.) random variables each with mean μ and variance σ^2. Then the distribution of the quantity

$$\xi = \frac{x_1 + x_2 + \ldots + x_n - n\mu}{\sigma\sqrt{n}} = \frac{\frac{x_1+x_2+\ldots+x_n}{n} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$ so that

$$P\{\xi \le a\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

as $n \to \infty$.

This results holds for any distribution of x_i. Thus for a large enough sampling of data, one should observe the Gaussian distribution.

1.5. **Joint distributions.** Now we look at events (say two) that take place jointly, that is both E_1 and E_2 occur and we want to know the probability of that joint event, $P\left(E_1 \cap E_2\right)$. Or, when dealing with continuous variables, we want to know the joint probability density $p\left(x, y\right)$ of x taking values between x and $x + dx$ and y taking values between y and $y + dy$.

In terms of cumulative distribution, the probability that $x \leq a$ and $y \leq b$ is given by

$$P\left\{\left(x \leq a, y \leq b\right)\right\} = \int_{-\infty}^{a} dx \int_{-\infty}^{b} dy\, p\left(x, y\right).$$

If we *marginalize*, i.e. take the limits $b \to \infty$ or $a \to \infty$, we obtain the single variable distributions

$$P_x\left\{x \leq a\right\} = P\left\{\left(x \leq a, y \leq \infty\right)\right\} = \int_{-\infty}^{a} dx \int_{-\infty}^{\infty} dy\, p\left(x, y\right) = \int_{-\infty}^{a} dx p_x\left(x\right),$$

where

$$p_x\left(x\right) = \int_{-\infty}^{\infty} p\left(x, y\right) dy.$$

Similarly with P_y and p_y.

The variable x and y will be *independent* if

$$p\left(x, y\right) = p_x\left(x\right) p_y\left(y\right)$$

and $P\left(a, b\right) = P_x\left(a\right) P_y\left(b\right)$.

1.6. **Parametric estimation.** The problem is to identify the pdf of a random variable η for which we have some samples $x_1, x_2, ...x_n$. Suppose we know (assume) the functional form of this distribution, but do not know the parameters in it. How do we identify them?

That is, we want to estimate a parameter ϑ of the pdf by a "statistic" $\hat{\vartheta}\left(x_1, x_2, ..., x_n\right)$ ("statistic" is a term that denotes any function of the sample).

An estimate is called "unbiased" if $E\left[\hat{\vartheta}\right] = \vartheta$, i.e. on average the estimate is exact. For example, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is an unbiased estimate of the mean, whereas the variance

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

is not unbiased, while

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

is an unbiased estimate of the variance.

Next, we show a method to find good estimators.

Suppose, the pdf of the random variable η that gave us the sample $x_1, x_2, ..., x_n$ is $f\left(x|\vartheta\right)$ with a parameter ϑ.

Question: what the best parameter ϑ given we have the sample?

To approach the question, suppose we know ϑ. Then the probability of getting the sample is proportional to

$$L = \prod_{i=1}^{n} f\left(x_i|\vartheta\right) \qquad \text{likelihood function.}$$

It seems reasonable to assume that a good estimate of ϑ is the one that maximizes L. This estimate is called the "maximum likelihood estimate".

Example. Suppose we assume that $x_i's$ are independent samples of a Gaussian with mean m and variance σ^2. Then

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(x_i - m\right)^2}{2\sigma^2}\right].$$

To find the maximum of L, we find the maximum of $\ln L$:

$$\ln L = \sum \left[-\frac{\left(x_i - m\right)^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\right) - \ln\sigma\right].$$

Then

$$\frac{\partial \ln L}{\partial m} = \sum \frac{x_i - m}{\sigma^2} = 0 \quad \Rightarrow nm = \sum x_i,$$

so the maximum likelihood estimate of m is the sample mean:

$$\hat{m} = \frac{1}{n}\sum x_i.$$

Similarly,

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum \frac{\left(x_i - m\right)^2}{\sigma^3} = 0 \quad \Rightarrow n\sigma^2 = \sum \left(x_i - m\right)^2,$$

so that the maximum likelihood estimate of the variance of a Gaussian variable is the sample (biased) variance

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