UPPSALA UNIVERSITET

Föreläsningsanteckningar

Sannolikhetsteori 2

Rami Abou Zahra

Inlämningsdatum
September 27, 2023

Contents

1. Repetition

Anmärkning: Det rekommenderas starkt att läsa igenom anteckningarna från Sannolikhetsteori 1

Definition/Sats 1.1: Random trial

An event is not certain, it usually has a probability associated with it. Taking that ”risk” to see what the outcome is, is called a random trial.

Examples of random trials include throwing dice, picking cards, number of people who pass a road

Different possibilities (outcomes).
In the example of the dice, the outcomes are 1-6

Definition/Sats 1.2: Events

An event is something that happens (or does not happen) when you the random trial

You can have an event based on one outcome, or multiple.

Example (one outcome): The dice is 3 after a throw

Example (several outcomes): The card is 7 or lower (1,2,3,4,5,6, all the different colours)

Example (0 outcomes): The card shows both spades and hearts at the same time (impossible)

1.1. Probability measure.

Related to the probability that an event occurs.

Definition/Sats 1.3: Probability measure

A *probability measure* is a function which satisfies Kolmogorovs axioms and for each event gives a number $\in [0, 1]$

The number is called the *probability* of the event. Usually denoted $P = P(A)$ where $A \subset \Omega$

Definition/Sats 1.4: Kolmogorovs axioms

Let $P : 2^\Omega \to \mathbb{R}$. P is called a *probability measure* if it satisfies the following

- $P(A) \geq 0 \quad \forall A \in 2^\Omega$
- $P(\Omega) = 1$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \quad (A_i \text{ disjoint})$

1.2. **Conditional probability theory.**

One can think of conditional probability in the term of Venn Diagrams in order to create intuition. Usually, if one event has happened it will affect other events and it is of interest to take this into consideration when calculating the probability of events.

The probability of an event A occuring given that the event B has occured is denoted by

$$P(A|B)$$

Example:
Let A be the event that a person has 2 daughters, let B be the event that a person has 0 daughters, and C be the event that he has at leat 1 daughter.

The probability $P(A|B)$ is of course 0, since given that he has 0 daughters, the probability is 0 for him to have 2 at the same time as he has 0

$P(B|C)$ is also 0, using similar argument as above

$P(A|C) = \dfrac{P(A \cap G)}{P(B)}$ We cannot say much here, other than that the probability is strictly positive since we already have one child

Definition/Sats 1.5: Bayes theorem

Let $F_1, \cdots, F_n$ be disjoint events $\in \Omega$ with $P(F_i) > 0$, and $P(\bigcup F_i) = 1$.

$$P(F_j|E) = \frac{P(E|F_j) \cdot P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Example:
Suppose we have 3 different cards. The first card is red on both sides (RR), the second card is black on both sides (BB), and the third card is black and red (RB)

We draw a card at random of these three cards such that we only see one side of the card.
Now suppose the side we see is red, what is the probability that the other side is black?

We are interested in the event $P(RB|R)$:

$$\frac{P(RB \cap R)}{P(R)} = \overbrace{=}^{\text{Bayes}} = \frac{P(R|RB)P(RB)}{P(R|RR)P(RR) + P(R|RB)P(RB) + P(R|BB)P(BB)}$$

$$= \frac{(1/2)(1/3)}{1 \cdot (1/3) + (1/2) \cdot (1/3) + (0) \cdot (1/3)} = \frac{1}{3}$$

1.3. **Independent events.**

Definition/Sats 1.6: Independent events

If $P(A|B) = P(A)$, then A and B are independent

Example:
Let A be the event that 2 parents get a daughter, and B be the event that the neighbors child ate an ice cream yesterday.
Since these events do not affect each other, they are independent.

Example:

Let A be the event that the first throw of a dice yields 6, and let B be the event that the second throw is 3. Then A and B are independent since the first throw does not affect the second throw:

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

Anmärkning:
There is an equivalent definition for independance through the following:

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

Anmärkning:
Independance is a symetric relationship.

1.4. **Random variables.**

Definition/Sats 1.7: Random variable

A *random variable* is a function that for each outcome associates a number with it.

An example is a persons age, or the value of a card drawn. If the outcome is random, the number is also random.

Each random variable has a distribution function associated with it, and is defined as $F(X) = P(X \leq x)$

Anmärkning:

$$\lim_{X \to -\infty} F(X) = 0$$
$$\lim_{X \to \infty} F(X) = 1$$

If $X_1 < X_2 \Rightarrow F(X_1) \leq F(X_2)$

We also have that F is right-continuous, meaning

$$\lim_{X \to a^+} F(X) = F(a)$$

There are 2 types of random variables that will be covered in this course, discrete and continuous (there is also absolutely continuous random variables, but they will not be covered)

1.4.1. *Discrete random variables.*

Definition/Sats 1.8: Discrete random variables

Consists of a finite or countable infinite set of numbers with probabilities:

- $P(X = x_i) = P(x_i) > 0$
- $P(X = \bigcup_{i=1}^{\infty} x_i) = 1$

Anmärkning:
If we have an uncountable infinite set of possibilities, the probability would be 0. Here is where continuous variables come to play

1.4.2. *Continuous random variables.*

For a continuous random variable, $F(x)$ is differentiable so that there exists a function f such that:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

From this comes 2 important things we can derive (both from the discrete and continuous case), namely expected value and the variance

Definition/Sats 1.9: Expected value

For discrete random variables, it is defined as

$$E(X) = \sum xF(x_i)$$

For the continuous case:

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

Definition/Sats 1.10: Variance

$Var(X) = E(X - E(X))^2$ for both discrete and continuous random variables

An equivalent definition is $E(X^2) - (E(X))^2$

Anmärkning:
$E(X^2)$ is called the second moment

Anmärkning:
If the variance is small, we know that the random variable does not fluctuate a lot from the expected value.

If X is a random variable (r.v) with density function f and g is a function, then we can define a new random variable $g(X) = Y$

Y is a random variable with density function $\hat{f}$.
Then:

$$E(Y) = \int y\hat{f}(y)dy = E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

And for the discrete random variable we have:

$$E(g(X)) = \sum g(x_i)p(x_i)$$

Anmärkning:
It is often better to use the definition of the density function for X rather than Y

Another remark worth noting is that $E(X^2)$ is a special case of $\int g(x)f(x)dx$

Example:
This example considers a distribution with no expected value (∞), and therefore it has no variance.
$P(X = k) = \frac{1}{k} - \frac{1}{k+1}$, this fulfills Kolmogorovs axioms, and

$$E(X) = \sum_{k=1}^{\infty} \frac{k}{k} - \frac{k}{k+1} = \sum_{k=1}^{\infty} \left(1 - \frac{k}{k+1}\right) = \sum_{k=1}^{\infty} \frac{1}{k+1} = \infty$$

2. Transformations

2.1. Pre-knowledge.

Let X_i be independent random variables with the same mean value (expected value) μ and variance σ^2

Let $S_n = \sum_{i=1}^{n} X_i \quad \overbrace{\Longrightarrow}^{\text{Law of large numbers}} \quad \frac{S_n}{n} \to \mu$ (convergance in probability).
Of course, this is assuming some sort of equal distribution.

The notation for convergance in probability is denoted by $Y_n \overset{P}{\to} a$. This follows from Markovs inequality. It is strongly suggested to look in the notes from the first course here.

From the law of large numbers, $S_n \approx n\mu$. But this does not take into account some erros that may take place in the $\frac{S_n}{n}$ side, as this does not affect the convergance to μ.
This is treated with the *Central Limit Theorem* (CLT), which says that $S_n \sim N(n\mu, n\sigma^2)$

This is equivalent to say that:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx N(0,1)$$

When n is large.

Anmärkning:
Here we talked about *convergence in distribution*

3. Multivariate random variables

It is strongly suggested to recall the random n-dimensional vector from Probability theory 1.

It is interesting to look at that the distribution of the vector, but we are often interested in a function $g(X)$

Example:
Starting with 1-dimension, and then working our way up.
Let $Y = g(X)$. Suppose g is strictly increasing of X (larger values of $X \to g(X)$ is larger).

Then

$$Y \leq y \in \mathbb{R} \Leftrightarrow g(X) \leq y \in \mathbb{R} \Leftrightarrow X \leq g^{-1}(y) = h(y)$$

If we look at the distribution function (which gives us everything, the probability the everything):

$$F_Y(y) = P(X \leq g^{-1}(y)) = P(X \leq h(y)) = F_X(h(y))$$

Anmärkning:
The inverse function g^{-1} is denoted by h

From the chain rule for derivates we get:

$$f_Y(y) = f_X(h(y)) \cdot h'(y)$$

We have gotten some information about X from Y. We can of course do the same for strictly decreasing functions:

$$Y \leq y \Leftrightarrow X \geq h(y)$$

$$\Rightarrow F_Y(y) = P(X \geq h(y)) = 1 - F_X(h(y))$$

This is good since we know that density functions are always positive. Taking the derivative gives us:

$$f_Y(y) = -f_X(h(y)) \cdot h'(y)$$

Anmärkning:
We assume g is differentiable with an inverse.

We showed that $f_Y(y) = f_X(h(y)) \cdot |h'(y)|$

3.1. Multivariate case.

Think of two n-dimensional space. One for $X = (X_1, X_2, \cdots, X_n)$, and one for $Y = (Y_1, Y_2, \cdots, Y_n)$

Suppose g is a bijective function such that g and g^{-1} are differentiable and let $Y = g(X) = (g_1(X), g_2(X), \cdots, g_n(X))$ where the component $Y_i = g_i(X = (X_1, X$

We have $X = g^{-1}(Y) = h(Y)$ (same as in 1-dimensional case)

Definition/Sats 3.11: Transformation theorem

The density of Y is given by

$$f_Y(y_1, y_2, \cdots, y_n) = f_X(h_1(y), h_2(y), \cdots, h_n(y)) \cdot |J|$$

Where J is the Jacobian matrix

$$J = \left|\frac{d(x)}{d(y)}\right| = \begin{vmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{y_n} \\ \vdots & \vdots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{vmatrix}$$

Anmärkning:
Transformation theorem corresponds to multivariate analysis change of variables

Bevis 3.1: Sketch of Transformation theorem

Let y_0 be a point in the Y-space. Choose an ε-ball C around y_0. Then we can assume that f_Y is constant in C.

The probability that our random vector Y will happen in this region is given by

$$\Delta C \cdot (f_Y(y_0) - \varepsilon) \leq P(Y \in C) \leq \Delta C \cdot (f_Y(y_0) + \varepsilon)$$

Anmärkning: ΔC is the volume/area of C

In the X-space, there then is a region which consists of all x whose $g(x)$ belongs to C.
Since g is bijective $\Rightarrow$ injective, we have that $Y \in C \Leftrightarrow X \in D = g^{-1}(C)$

This means that these probabilities are the same

$$\left|f_Y(y_0) \cdot \Delta C - \int_D f_X(x)dx\right| \leq \Delta C \cdot \varepsilon$$

As C decreases, ΔC decreases as well as ε
Since g is a nice function, D will also decreases
We let $x_0 = g^{-1}(y_0) = h(y_0)$. We can replace the integral by $f_X(x_0) \cdot \Delta D$ and obtain

$$f_Y(y_0) \cdot \Delta C \approx f_X(x_0) \cdot \Delta D \Leftrightarrow f_Y(y_0) \approx f_X(x_0) \frac{\Delta D}{\Delta C}$$

We get equality when $C \to 0$ (choosing a smaller and smaller ε)

Recall the functional determinant (Jacobian) of the matrix $\frac{\Delta x}{\Delta y} = \left|\frac{d(x)}{d(y)}\right|$ (relative volume change)

Thus, we get $f_Y(y_0) = f_X(h(y_0)) \, ||J_n(x_0, y_0)||$
Since this is true for all y, we can take away the index y_0, and we get:

$$f_Y(y) = f_X(h(y)) \cdot |J|$$

□

Example (1-dim case):
Suppose $g(X) = aX + b$. From this it is easy to see what the inverse function $h = g^{-1}$ is, namely $h(y) = \frac{y-b}{a}$.

The Jacobian is just $\frac{1}{a}$. By the transformation theorem we get

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \left|\frac{1}{a}\right|$$

The main thing is that using the density function of X we can get the density function for Y.

Example 2.4:
X, Y are independent normally distributed random variables $N(0,1)$, show that $X + Y$ and $X - Y$ are independent and determine their distribution function.

In order to solve this we do a variable substitution. Let $U = X + Y$ and $V = X - Y$.
Notice that $\frac{U+V}{2} = X$ and $\frac{U-V}{2} = Y$
We have our function $g(x, y) = (u, v) = (x + y, x - y)$
We have our inverse $g^{-1}(u, v) = (x, y) = \left(\frac{u+v}{2}, \frac{u-v}{2}\right)$
We can now use the transformation theorem:

$$f_{U,V}(u, v) = f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot |J|$$

The Jacobian can be found:

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = \frac{-1}{4} - \frac{1}{4} = \frac{-1}{2}$$

By the transformation theorem:

$$f_{U,V}(u, v) = f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot \frac{1}{2} \overset{\text{indep.}}{=} f_X\left(\frac{u+v}{2}\right) f_Y\left(\frac{u-v}{2}\right) \cdot \frac{1}{2}$$

We know their density functions since they are normally disitrbuted with $N(0,1)$:

$$= \frac{1}{\sqrt{2\pi}} e^{-(1/2)\left(\frac{u+v}{2}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(1/2)\left(\frac{u-v}{2}\right)^2} \cdot \frac{1}{2}$$

After simplification we get that $f_{U,V}$ is a product of one function of U and one function of V. This means that U, V are independent since we get them as a product of two different functions.

Example:
Recall the convolution formula from Probability theory 1; if X, Y are independent, we have

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

We can show this from the transformation theorem.

By the independance of X, Y, we have $f_{X,Y}(x, y) = f_X(x) f_Y(y) \forall x, y$
Let $Z = X + Y$. Then $g(X, Y) = (X + Y, X)$
The inverse is given by $Y = Z - X$. We can now use the transformation theorem:

$$f_{Z,Y}(z, x) = f_{X,Y}(h_1(z, x), h_2(z, x)) \cdot |J|$$

The Jacobian is given by

$$\begin{vmatrix} 1 & -1 \\ 1 & 0 \end{vmatrix} = 1$$

Since X, Y are independent, we get:

$$f_{Z,X}(z, x) = f_Y(z - x) f_X(x)$$

The marginal density is given by integrating away x:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z,X}(z,x)dx = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)$$

3.2. **Conditional Probabilities.**

It is suggested to do some examples from Probability theory 1.

Let X, Y be some random variables with joint discrete distribution (**TODO:** *Def*).
We look at the conditional distribution:

$$p_{Y|X=x}(y) = P(Y=y|X=x) = \frac{P_{X,Y}(x,y)}{P_X(x)}$$

If we look at the conditional probability distribution function, we have:

$$F_{Y|XX=x}(y) = \sum_{z \le y} P_{Y|X=x}(z)$$

If we now look at if X, Y have a joint continuous distribution, we have something similar, with a couple of things changed where we use the density function instead (since some probabilities are 0)

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Does this make sense for the continuous case? Well, suppose that $f_{Y|X=x_0}(y) = f_{X,Y}(x_0,y)$, would this be a natural definition?
The reason we cannot use this definition is because the probability that $X = x_0$ is very small because of continuous. But now we have this given, we have assumed this has happened, so $f_{X,Y}(x_0,y)$ could very well be too small compared to $f_{Y|X=x_0}(y)$ where we already know that $X = x_0$ happpens.
If we instead put $f_{Y|X=x_0}(y) = Kf_{X,y}(x_0,y)$ some constant to compensate, then we need to check if the properties for density functions are preserved:

$$\int_{-\infty}^{\infty} f_{Y|X=x_0}(y)dy = 1 \Leftrightarrow \text{ proper density function}$$

However, with the K in front, we get:

$$\begin{aligned}\int_{-\infty}^{\infty} f_{Y|X=x_0}(y)dy = 1 &= \int_{-\infty}^{\infty} Kf_{X,Y}(x_0,y)dy \\ &= Kf_X(x_0) = 1 \\ &\Rightarrow K = \frac{1}{f_X(x_0)}\end{aligned}$$

This is true for all $X = x_0$, so the formula we have seems to be correct!

Just as in the discrete case, we have:

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(z)dz$$

Now we have all the tools to start define conditional expectations and conditional variances.

3.3. **Conditional expected values and variances.**

There are some natural definitions

Definition/Sats 3.12: Conditional Expected Value

In the continuous case we have:

$$\mathbb{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)dy$$

In the discrete case we have:

$$\mathbb{E}(Y|X=x) = \sum_{-\infty}^{\infty} P_{Y|X=x}(y)$$

Notice that $\mathbb{E}(Y|X=x)$ is a function of x.
We now look at only the random variable $\mathbb{E}(Y|X) = g(X)$

Sometimes it is easier to look at the conditional expected value.

Definition/Sats 3.13

$\mathbb{E}(Y|X)$ has the same expected value as $\mathbb{E}(Y)$:

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(g(X)) = \mathbb{E}(Y)$$

In the discrete case, we see a variant of the law of total probabilities:

$$\mathbb{E}(Y) = \mathbb{E}(Y|X) = \sum_x \mathbb{E}(Y|X=x)P(X=x)$$

Rules for caluclations, as in the unconditional case, can be found in the book (**TODO:** *Ins*)

$$\mathbb{E}(f(X)Y|X) = f(X)\mathbb{E}(Y|X)$$

Another natural rule is if X, Y are independent:

$$\mathbb{E}(Y|X) = \mathbb{E}(Y)$$

Definition/Sats 3.14: Conditional variance

$$v(x) = Var(Y|X=x) = \mathbb{E}((Y-\mathbb{E}(Y|X=x))^2|X=x)$$
$$V(X) = Var(Y|X) = \mathbb{E}((Y-\mathbb{E}(Y|X))^2|X)$$

Definition/Sats 3.15

We define $e(X) = \mathbb{E}(Y|X)$ and $V(X) = Var(Y|X)$ and assume $g(X)$ is some function on X.
Then we have:

$$\mathbb{E}((Y-g(X))^2) = \mathbb{E}(V(X)) + \mathbb{E}((e(X)-g(X))^2)$$

Bevis 3.2

$$
\begin{aligned}
\mathbb{E}((Y-g(X))^2) &= \mathbb{E}((Y-e(X)+e(X)-g(X))^2) \\
&= \mathbb{E}((Y-e(X))^2) + 2\mathbb{E}(Y-e(X))\mathbb{E}(e(X)-g(X)) + \mathbb{E}((e(X)-g(X))^2) \\
&= \mathbb{E}(\mathbb{E}((Y-e(X))^2)|X) + 2\mathbb{E}(\mathbb{E}(Y-e(X)\mathbb{E}(e(X)-g(X)))|X) + \mathbb{E}((e(X)-g(X))^2) \\
&= \mathbb{E}(v(X)) + 2(e(X)-g(X))\mathbb{E}(Y-e(X)|X) + \mathbb{E}((e(X)-g(x))^2) \\
&= \mathbb{E}(v(X)) + 2(e(X)-g(X))\underbrace{(e(X)-e(X))}_{=0} + \mathbb{E}((e(X)-g(x))^2)
\end{aligned}
$$

In the middle term we used:

$$\mathbb{E}(\mathbb{E}(Y-e(X))|X) = \mathbb{E}(e(X)-e(X)|X) = e(X)-e(X) = 0$$

□

There is a nice corollary that follows from this:

Definition/Sats 3.16

$$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X))$$

The proof of the corollary follows from Proof 3.2 by choosing $g(X) = \mathbb{E}(Y)$.
Then the last theorem gives the result directly:

$$\mathbb{E}((Y-e(X))^2) = \mathbb{E}(v(X)) + \mathbb{E}((e(X)-e(Y))^2)$$

But $e(X)$ has the same expected value as Y, as in $\mathbb{E}(Y) = \mathbb{E}(e(X))$, so we get:

$$\mathbb{E}(v(X)) + \mathbb{E}((e(X)-e(Y))^2) = \mathbb{E}(v(X)) + Var(\mathbb{E}(Y|X))$$

Recall that ranom variables do not only have values attached to them, but also parameters. For example $X \sim N(\mu, \sigma^2)$ where μ, σ^2 are parameters.

For example, if $X \sim N(\mu, \sigma^2)$ and $X = x$ is an outcome of this random variable, then both $x, \mu \in \mathbb{R}$ while X is a random variable.

Sometimes we want to think of the parameters as random variables.

Example:
Suppose we go to a hospital to take a blood-test and count the number of red blood cells in the sample, then we will get some value which also partly depends on some randomness.
First we look at some individual (even if I go to the hopsital several times, each time will give different results).

This seemingly random variation from the same person can be explained by the Poisson-distribution with some parameter m.
This means that if X is an observed value, $X \sim Po(m)$. The value m can be different across people.

We can think that we do a random trial in 2 steps:

- Choose a random individual to take the blood test from
- Count the amount of red blood cells from that individual

Then we can let X be the observed value for that person, and we have $X|M = m \sim Po(m)$ with M having some distribution (does not need to have Poisson distribution)

Example:
By the law of total probability, we can look at:

$$P(A) = \int_{-\infty}^{\infty} P(A|M = x) f_M(x) dx$$

Suppose now that $M \sim Exp(1)$. We then get:

$$\begin{aligned} P(X=k) &= \int_{-\infty}^{\infty} P(X=k|M=x)f_M(x)dx \\ &= \int_0^{\infty} e^{-x}\frac{x^k}{k!}e^{-x}dx = \frac{1}{k!}\int_0^{\infty} x^k e^{-2x}dx \\ \frac{1}{k!}\int_0^{\infty}\frac{y^k}{2^k}e^{-y}\frac{1}{2}dy &= \frac{1}{k!+2^{k+1}}\int_0^{\infty} y^k e^{-y}dy \\ \Rightarrow k! &\Rightarrow P(X=k) = \frac{1}{2^{k+1}} \end{aligned}$$

So X has a geometric distribution. One can also get this from the Γ distribution since it is very similar:

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

When $z = k$ then $\Gamma(z) = (k-1)!$

In the last example, we do not need to calculate the unconditional distribution of X if we just want to know the expected value/variance of X using the formulas that we proved above.

Recall that we can write $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|M))$. We said that $X|M \sim Po(m)$, and we know that a Poisson random variable has $\mathbb{E}(X|M) = M$, and we get:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|M)) = \mathbb{E}(m)$$

Also remember that $M \sim Exp(1)$, so $\mathbb{E}(M) = 1$, and we have:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|M)) = \mathbb{E}(M) = 1$$

By the corollary, we can also find the variance:

$$Var(X) = \mathbb{E}(Var(X|M)) + \underbrace{Var(\mathbb{E}(X)|M)}_{Var(X|M)\,=\,M} = \mathbb{E}(M) + \underbrace{Var(M)}_{=\,m\,=\,1} = 1+1 = 2$$

Example:
Suppose we are in a coffee shop, and every customer has a choice between coffee (with probability p) or tea (with probability $1-p$).
Suppose the number of customers during lunchtime N is $\sim Po(\lambda)$ distributed. We want to count the number of coffees ordered in total (or rather, find the distribution of the number of coffees).

We proceed by letting X be number of coffees ordered, and let N be the number of customers. We actually know because of how we assumed it, we know the amount of customers, we have a binomial distribution:

$$X|N = n \sim Bin(n,p)$$

From our example we also know that $N \sim Po(\lambda)$, we have $P(X=k|N=n) = \binom{n}{k}p^k q^{n-k}$.
Now we want to count $P(X=k)$ without $N=n$. We can use the law of total probabilities:

$$\begin{aligned} P(X=k) &= \sum_{n=0}^{\infty} P(X=k|N=n)P(N=n) \\ &= \sum_{n=0}^{\infty}\binom{n}{k}p^k q^{n-k}e^{-\lambda}\frac{\lambda^n}{n!} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!} \\ &\Rightarrow \frac{p^k}{k!}e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{(n-k)!}q^{n-k} \\ &= \frac{(\lambda p)^k}{k!}e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda q)^{n-k}}{(n-k)!} \end{aligned}$$

Let $j = n - k$, then we can rewrite the sum as:

$$\frac{(\lambda p)^k}{k!} e^{-\lambda} \underbrace{\sum_{j=0}^{\infty} \frac{(\lambda q)^j}{j!}}_{= e^{\lambda q}}$$

$$\frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda q} = \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} = \frac{(\lambda p)^k}{k!} e^{-\lambda p}$$

This looks like a Poisson distribution, but λ is changed to λp. Therefore it is a Poisson distribution with parameter $\lambda p \Rightarrow X \sim Po(\lambda p)$

4. Exercises

4.1. **Exercise 1.5.**

Show that if $X \sim C(0,1)$, then $X^2 \sim F(1,1)$

Anmärkning:
C is a Cauchy distribution, and F is a Fisher distribution. Note also that $\Gamma(1) = 1$ and that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Let $Y = X^2$. We want to show that it has distribution $\frac{1}{\pi}\frac{x^{-1/2}}{1+x}$. Taking the derivative of $F_Y(y) = P(Y \leq Y)$ yields the density function.
Want to show that this is $\frac{y^{-1/2}}{1+y}\frac{1}{\pi}$:

$$\begin{aligned} P(Y \leq y) &= P(X^2 \leq y) = P(|X| \leq \sqrt{y}) \\ &= P(-\sqrt{y} \leq X\sqrt{y}) \end{aligned}$$

From the density function of X, we can calculate this:

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\pi}\frac{1}{1+x^2} dx$$

Since we want the derivative, we can use the fundamental theorem of calculus to take the derivative of the integral:

$$f_Y(y) = \frac{d}{dy}\frac{1}{\pi}\int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{1+x^2} dx$$

$$\overset{\text{Sym.}}{=} 2\int_0^{\sqrt{y}} \frac{1}{1+x^2} dx \Rightarrow \frac{2}{\pi}\frac{1}{y} \cdot \frac{d}{dy}\sqrt{y} = \frac{2}{\pi}\frac{1}{1+y} y^{-0.5}\frac{1}{2}$$

$$= \frac{1}{\pi}\frac{y^{-0.5}}{1+y} = F(1,1)$$

Anmärkning:
One could have used the transformation theorem here to arrive at the same answer.

4.2. **Exercise 1.13.**

Let X, Y have a joint density function given by

$$f(x,y) = \begin{cases} 1 & 0 \leq x \leq 2, \max(0, x-1) \leq y \min(1,x) \\ 0 & \end{cases}$$

determine the joint distribution functions and the joint and marginal distribution function.

We have several different cases to check in this problem depending on where (x, y) is situated. We will look at one trivial case and one non-trivial case.
In general, we know how to write out the distribution of a 2-dimensional r.v:

$$F(x,y) \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) du dv$$

In our case we have a uniform distribution, we can use this to skip a lot of the calculations. We integrate over the rectangle covered by (x, y), then the trivial case is where the area we are interested in is covered by (x, y) and the area of that area of interest is 1. The non-trivial part arises when we look at some point inside the area. Another trivial case is putting the point at $(0, 0)$, then the area covered is 0.

In the non-trivial case where we assume (x, y) is somewhere in the left part of the area. It is calculated to be $xy - \frac{y^2}{2}$

From this, we can calculate the marginal distributions pretty easily by integrating away x, y

4.3. **Exercise 1.19.**

The random vector $\mathbf{X} = (X_1, X_2, X_3)$ with density function

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{2}{2e-5} x_1^2 x_2 e^{x_1 x_2 x_3} & 0 < x_1, x_2, x_3 < 1 \\ 0 \end{cases}$$

Determine the distribution of $X_1 \cdot X_2 \cdot X_3$

We proceed using the transformation theorem. We define $y_1 = x_1 x_2 x_3$, $y_2 = x_1 x_2$, $y_3 = x_1$. The reason we write it like this is because it is easier to find the inverse of these functions:

$$x_1 = y_3$$
$$x_2 = \frac{y_2}{x_1} = \frac{y_2}{y_3}$$
$$x_3 = \frac{y_1}{x_1 x_2} = \frac{y_1}{y_3 \frac{y_2}{y_3}} = \frac{y_1}{y_2}$$

All the preparation is done, we use the transformation theorem:

$$f_Y(y) = f_X(x) \, |J|$$

$$J = \begin{vmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{y^3} & -\frac{y_2}{y_3^2} \\ \frac{1}{y_2} & -\frac{y_1}{y_2} & 0 \end{vmatrix} = \frac{-1}{y_2 y_3} \Rightarrow |J| = \frac{1}{y_2 y_3}$$

Plug and chugg:

$$f_Y(y) = \frac{2}{2e-5} y_3^2 \frac{y_2}{y_3} e^{y_1} \frac{1}{y_2 y_3} = \frac{2}{2e-5} e^{y_1}$$

Since $y_1 = x_1 x_2 x_3$ and $x_1 x_2 x_3$ are all between $0, 1$, therefore $y_1 < y_2 < y_3$.

We want to find $Y_1 = X_1 X_2 X_3$, but this is just the marginal distribution in Y_1 of $Y = (Y_1, Y_2, Y_3)$. By integrating away Y_2, Y_3 we can get this:

$$F_{Y_1}(y_1) = \int_{y_1}^{1} \int_{y_2}^{1} \frac{2}{2e-5} e^{y_1} dy_3 dy_2$$
$$= \frac{2}{2e-5} e^{y_1} \int_{y_2}^{1} (1 - y_2) dy_2$$

As an exercise, show that this integral is equal to $\frac{1}{2e-5}(1 - y_1)^2 e^{y_1}$ for $0 < y_1 < 1$

5. Transformations

NOTE: Fridays notes are not in this document. ADD

Definition/Sats 5.17

Suppose $\exists h$ such that $\psi_X(t)$ is defined $\forall t$ such that $|t| < h$, then $\mathbb{E}(|x|^r) < \infty \quad \forall r$. Note that variance is given by second moment, there are some distributions where only maybe one or two moments exists, but this theorem says that all should exist given the cases.

Bevis 5.1

If we have $t > 0$ such that $\psi_X(t) < \infty$, then we have that by the definition of the moment generating function:

$$\int_{-\infty}^{\infty} e^{tX} f_X(x)dx < \infty$$

But we also know that e^x grows faster than x^r, so if we write

$$\frac{|x|^r}{e^t x} = 0 \quad \forall r \quad x \to \infty$$

$$\Rightarrow \int_{x_1}^{\infty} |x|^r f_X(x)dx < \int_{x_1}^{\infty} e^{tX} f_X(x)dx < \infty \qquad \text{for } x \text{ is large enough}$$

$$\Rightarrow \int_{0}^{\infty} |x|^r f_X(x)dx < \infty$$

Now half of the proof is complete. The definition says we need to look from $-\infty$ to ∞. To do this, we look at other t:s

If x is instead close to $-\infty$, then we know that $e^{tx} \to 0$ for $t > 0$, and at the same time as $|x|^r \to \infty$. Thus $\lim_{x\to-\infty} \frac{|x|^r}{e^{tx}} \neq 0$. But since we assumed in the theorem that $\psi_X(t)$ is defined $\forall\, |t| < h$ (in an interval from $-h$ to h), the theorem also includes $t < 0$.

We choose such a negative t, and write it as $-t$. Then we know that we have:

$$\int_{-\infty}^{x_2} e^{-tx} f_X(dx) < \infty \qquad \forall x_2$$

We also know that when we take $\lim_{x\to-\infty} \frac{|x|^r}{e^{-tx}} = 0$. Then $\exists x_3$ such that:

$$\int_{-\infty}^{x_3} |x|^r f_X(x)dx < \int_{-\infty}^{x_2} e^{-tx} f_X(x)dx < \infty$$

We also know that the integral from x_3 to x_2 is finite, so we know the whole integral is finite. □

The moment generating function is more general than the probability generating function.

Definition/Sats 5.18

Suppose there exists h such that $\psi_X(t)$ is defined for $|t| < h$, then we have that $\mathbb{E}(|X|^n) = \psi_X^{(n)}(0)$ $\forall n \in \mathbb{N}$

Bevis 5.2

This comes naturally from the Taylor expansion of e^{tX}:

$$e^{tX} = \sum_{n=0}^{\infty} \frac{t^n X^n}{n!}$$

Recall that $\psi_X(t) = \mathbb{E}(e^{tX})$:

$$\Rightarrow \mathbb{E}(e^{tX}) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right) = 1 + \sum_{k=1}^{\infty} \frac{t^n}{n!}\mathbb{E}(X^n)$$

Derivating yields:

$$\psi_X'(t) = \sum_{n=1}^{\infty} \frac{t^{n-1}\mathbb{E}(X^n)}{(n-1)!} = \mathbb{E}(X) + \sum_{n=2}^{\infty} \frac{t^{n-1}\mathbb{E}(X^n)}{(n-1)!}$$

Plugging in $t = 0$, we see that:

$$\psi_X'(0) = \mathbb{E}(X) + 0 = \mathbb{E}(X)$$

Looking at the second moment instead:

$$\frac{d}{dt}\left(\mathbb{E}(X) + \sum_{n=2}^{\infty} \frac{t^{n-1}\mathbb{E}(X^n)}{(n-1)!}\right) = 0 + \sum_{n=2}^{\infty} \frac{t^{n-2}\mathbb{E}(X^n)}{(n-2)!}$$
$$= \mathbb{E}(X^2) + \sum_{n=3}^{\infty} \frac{t^{n-2}\mathbb{E}(X^n)}{(n-2)!}$$

Plugging in $t = 0$ yields $\psi_X''(0) = \mathbb{E}(X^2)$, this will happen for all derivatives. □

We shall now look at some moment generating functions (mgf) for some discrete functions:

- $X \sim Be(p) \quad \mathbb{E}(e^{tX}) = pe^t + (1-p)e^{t0} = pe^t + 1 - p = \psi_X(t)$
  Looking at the derivatives we have $\psi_X'(t) = pe^t$, and $\psi_X'(0) = p = \mathbb{E}(X)$
  We can find the variance, since it is the second moment minus the first moment squred, and we get $p - p^2 = p(1-p)$

- $X \sim Bin(n,p) \quad \mathbb{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{\infty}(pe^t)^k\binom{n}{k}(1-p)^{n-k}$
  From the binomial theorem, we can write this as $(pe^t + (1-p))^n$
  Observe that $(pe^t + (1-p))^n$ is the mgf of n independent $Be(p)$ because of corollary 3.2.1

- Geometric and Poisson distribution: We know that if the probability generating function exsits (positive integral values), then we can write $\psi_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}((e^t)^X)$. Letting $t = e^t$, we have $= g_X(e^t)$
  If we trust the existance, we can use that the probability generating function for the geometric distribution $g_X(t) = \dfrac{p}{1-(1-p)t}$ for $|t| < 1$. This implies:

  $$\Rightarrow \psi_X(t) = \frac{p}{1-(1-p)e^t}$$

  For the Poisson distribution, we saw that $g_X(t) = e^{m(t-1)}$, we can use this again:

  $$\psi_X(t) = e^{m(e^t-1)}$$

We will now look at some common continuous distributions as well as their moment generating functions:

- If $X \sim U(a,b) \Rightarrow \psi_X(t) = \int_a^b e^{tx}\dfrac{1}{b-a}dx = \dfrac{e^{tb}-e^{ta}}{t(b-a)}$

  Anmärkning: If $\psi_X'(0)$ yields something indeterminate, then take the limit as $t \to 0$

- For the exponential distribution we have:
$$\psi_X(t) = \int e^{tx}\frac{1}{a}e^{-x/a}dx = \frac{1}{a}\frac{1}{\frac{1}{a}-t} = \frac{1}{1-at}$$

- Gamma distribution: We know that the mgf for a sum of independent random variables is the same as the product of the mgf:s, and we know a theorem which says that the gamma distribution is a sum of exponential random variables.
  For integers p, $\Gamma(p,a)$ is a sum of p independent exponentially disitrbuted random variables with parameter a. So we can use what we know about mgf:s for exponentially disitrbuted functions:
$$\psi_X(t) = \frac{1}{(1-at)^p}$$

- Normal distrbution ($N(0,1)$):
$$\psi_X(t) = \int e^{tx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int e^{tx-(1/2)x^2}dx$$
$$= \frac{1}{\sqrt{2\pi}}\int e^{-(1/2)(x-t)^2-t^2}dx = \underbrace{\frac{1}{\sqrt{2\pi}}\int e^{-(1/2)(x-t)^2}}_{=1}e^{t^2/2}dx \Rightarrow e^{t^2/2}$$

  Looking at the general case $Y \sim N(\mu,\sigma^2) \Rightarrow Y = \sigma X + \mu$:
$$\psi_Y(t) = \mathbb{E}(e^{\sigma X t+\mu t}) = e^{\mu t}\mathbb{E}(e^{\sigma X t}) = e^{\mu t}\psi_X(\sigma t) = e^{\mu t}e^{(\sigma t)^2/2}$$

Theorem 3.3 a) tells that if the moment generating function exists $\Rightarrow$ all absolute moments exist and $\mathbb{E}(|X|^r) < \infty$, but does the opposite apply? The answer to that is infact no, a counter example is something called a log-normal (LN) distribution:
$$X \sim LN(\mu,\sigma^2) \Leftrightarrow X = e^Y \quad (Y \sim N(\mu,\sigma^2))$$

Note that $\mathbb{E}(X^r) = \mathbb{E}(e^{Yr}) = \psi_Y(r) < \infty$.
One can however show that $\psi_X(t) = \mathbb{E}(e^{tX}) = \infty \ \forall t \neq 0$

Definition/Sats 5.19: Moment Generating function for a multivariate variable

$$\psi_{X_1,X_2,\cdots,X_n}(t_1,t_2,\cdots,t_n) = \mathbb{E}(e^{t_1x_1+t_2x_2+\cdots+t_nx_n})$$

Definition/Sats 5.20: Characteristic function

$$\varphi_X(t) == \mathbb{E}(e^{itX}) = \mathbb{E}(\cos(tX) + i\sin(tX))$$

From the definition, there are some rules that we can derive for the characteristic function:

- $|\varphi_X(t)| \leq 1 \quad \forall t$
- $\varphi_X(t) = \varphi_X(-t)$
- Is uniformly continuous
- $\varphi_S(t) = \Pi\varphi_{X_k}(t)$ where $S = X_1 + X_2 + \cdots + X_n$
- $\varphi_X^{(k)} = i^k\mathbb{E}(X)$

Let us go back to talking about distirbutions with random parameters.

Suppose that $X|N = n \sim Bin(n,p)$ where $N \sim Po(\lambda)$. Previously, we tried to find the unconditional distribution.
We will try to solve this using the probability generating function $g_X(t)$.
We can do this by looking at $g_X(t) = \mathbb{E}(t^X)$ and rewriting it to include the conditional:
$$g_X(t) \Rightarrow \mathbb{E}(\mathbb{E}(t^X|N))$$

We know that $\mathbb{E}(t^X|N=n) = (q+pt)^n$ since $X|N=n \sim Bin(n,p)$

$$\Rightarrow \mathbb{E}(t^X|N) = (q+pt)^N \Rightarrow g_X(t) = \mathbb{E}((q+pt)^N) = g_N(q+pt)$$

Since we know the distribution of $N \sim Po(\lambda)$, we know the probability generating function for it too:

$$g_N(t) = e^{\lambda(t-1)} \Rightarrow g_N(q+pt) = e^{\lambda(q+pt-1)} = e^{\lambda(1-p+pt-1)} = e^{\lambda p(t-1)}$$

We have shown that $g_X(t)$ is equal to $e^{\lambda p(t-1)} \Rightarrow X \sim Po(\lambda p)$

If one of the distributions is not of integer value, we have to use the moment generating function instead.

In example 2.3.1 we have that $X|M=m \sim Po(m)$ where $M \sim Exp(1)$
We apply the mgf:

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}(\mathbb{E}(e^{tX}|M))$$

When $M = m \Rightarrow X|M=m \sim Po(m)$. We know that what the mgf for Poisson variables is:

$$= e^{m(e^t-1)}$$

$$\Rightarrow \psi_X(t) = \mathbb{E}((e^{tX}|M)) = \mathbb{E}(e^{M(e^t-1)})$$

This looks like something that we know, in of itself it looks like a mgf because we have e^M where M is a random variable with some real values. Well, our intuition is somewhat correct:

$$= \psi_M(e^t-1)$$

We know $M \sim Exp(1) \Rightarrow$ the mgf for $M = \dfrac{1}{1-t}$. From this, we can now find the mgf for X since we expressed in the terms of M:

$$\psi_X(t) = \frac{1}{1-(e^t-1)} = \frac{1}{2-e^t}$$

Now we want to find the distribution for X, so we can look in the table and see that a random variable that has the closest looking mgf is if $X \sim Geo$. We do have to do some rewriting for it to look exact:

$$\frac{1/2}{1-(1/2)e^t} = Geo(1/2)$$

Since if $X \sim Geo(p)$, then the mgf is $\dfrac{p}{1-qe^t}$.
This means that the function we found is the moment generating function for a geometric distribution with parameter 1/2.

We shall look at one more example. Suppose that $X|\Sigma = y \sim N(0,y)$ with $\Sigma \sim Exp(1)$. We want to find the moment generating function for X:

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}(\mathbb{E}(e^{tX}|\Sigma))$$

We look for an mgf for $N(0,y)$, which is $e^{yt^2/2}$. This means that:

$$\mathbb{E}(e^{tX}|\Sigma) \quad (e^{\Sigma t^2/2})$$

$$\Rightarrow \psi_X(t) = \mathbb{E}(e^{\Sigma t^2/2}) = \psi_\Sigma(t^2/2)$$

Since $\Sigma \sim Exp(1)$, we know that the mgf for Σ is $\dfrac{1}{1-t}$, but now instead of t we have $t^2/2$, so we have:

$$\psi_\Sigma(t^2/2) = \psi_X(t) = \frac{1}{1-t^2/2}$$

We now look in the appendix to see if this looks like some known distribution, and it turns out it looks like a Laplace distribution:

$$\frac{1}{1-a^2t^2} \sim \mathcal{L}(a) \Rightarrow X \sim \mathcal{L}\left(\frac{1}{\sqrt{2}}\right)$$

5.1. **Sums of random number of terms.**

Suppose that $X_1, X_2, \cdots$, is a sequence of independent and equally distributed random variables.
We look at the partial sum $S_N = X_1 + X_2 + \cdots + X_n$ and study $S_N = X_1 + \cdots + X_N$.
Assume the number of terms is independent of all X_i. We first at the special case where X_i are positive integer valued random variables.
Then we have that the probability generating function for S_N exists, so we can write:

$$g_{S_N}(t) = \mathbb{E}(t^{S_n})$$

In that case we can use the law of total probability and show that it is the following sum:

$$\mathbb{E}(t^{S_n}) = \sum_{n=0}^{\infty} \mathbb{E}(t^{S_N} | N = n) \cdot P(N = n)$$

$$\sum_{n=0}^{\infty} \mathbb{E}(t^{S_n}) \cdot P(N = n) \qquad \text{(since all } X_i \text{ in } S_n \text{ are independent of } n\text{)}$$

We know that $\mathbb{E}(t^{S_n}) = g_{S_n} = (g_X(t))^n$:

$$g_{S_N}(t) = \sum_{n=0}^{\infty} (g_X(t))^n \cdot P(N = n) = \mathbb{E}\left(g_X(t)\right)^n$$

This means that this is almost a probability generating function because it is something to the power of N:

$$g_N((g_X(t)))$$

We can use this to find $\mathbb{E}(S_N)$:

$$\mathbb{E}(S_N) = g'_{S_N}(1) = g'_N(g_X(1))g'_X(1)$$

Notice that this is a product of 2 expected values, since $g'_X(1) = \mathbb{E}(X)$ and $g'_N(1) = \mathbb{E}(N)$. Always when you plug in 1 in a probability generating function you will get 1, so $g_X(1) = 1$, and we get:

$$g'_N(1)g'_X(1) = \mathbb{E}(N)\mathbb{E}(X)$$

If $\operatorname{Var}(N) < \infty$ and $\operatorname{Var}(X) < \infty$,

$$\operatorname{Var}(S_N) = g''_{S_N}(1) + g'_{S_N}(1) - \left(g'_{S_N}(1)\right)^2$$

We use the following formula for the variance:

$$\operatorname{Var}(S_N) = \mathbb{E}(\operatorname{Var}(S_n)) + \operatorname{Var}(\mathbb{E}(S_N|N))$$
$$\operatorname{Var}(S_N|N = n) = \operatorname{Var}(S_n) = n\operatorname{Var}(X)$$
$$\Rightarrow \operatorname{Var}(S_N|N) = N\operatorname{Var}(X)$$

We also have that $\mathbb{E}(S_N|N = n) = n\mathbb{E}(X)$:

$$\Rightarrow \mathbb{E}(S_N|N) = N\mathbb{E}(X)$$

Now we can actually use this to look at our general formula. From the first part, we know that $\mathbb{E}(\operatorname{Var}(S_N|N)) = \mathbb{E}(N\operatorname{Var}(X))$. But $\operatorname{Var}(X)$ is just a constant, so this we only get $\mathbb{E}(N)\operatorname{Var}(X)$ from the first part.
From the second part, we get:

$$\operatorname{Var}(\mathbb{E}(S_N|N)) = \operatorname{Var}(N\mathbb{E}(X))$$

Now we also remember from Probability theory 1, that $\operatorname{Var}(CX) = c^2\operatorname{Var}(X)$, so we get:

$$\operatorname{Var}(N\mathbb{E}(X)) = \mathbb{E}^2(X)\operatorname{Var}(N)$$

Then we have our general formula:

$$\operatorname{Var}(S_N) = \mathbb{E}(N)\operatorname{Var}(X) + \operatorname{Var}(N)\,\mathbb{E}^2(X)$$

Definition/Sats 5.21

Let $X_1, \cdots$ be i.i.d.r.v whose mgf exists for $|t| < h$ for some $h > 0$.

Let also N be a positive integer independent random varaiable of all all X_i, then

$$S_n = \sum_{k=0}^{n} X_i \Rightarrow \psi_{S_N}(t) = g_N(\psi_X(t))$$

6. Multivariate Normal Distributions

Usually, an observed value is a funcftiojn of many things at once, such as:

$$W = f(\text{thing}_1, \text{thing}_2, \cdots)$$

As long as there are small variations around a mean value, then f is a linear function of all their variables, and the variation is a sum of the contributions.

If one wants instead to measure several properties at once for an individual then one can look at the d-dimensional vector and show that it also varies in a similar way. This vector will be *multivariate normal distributed.*

The mean value is given by:

$$\mathbb{E}(X) = [\mathbb{E}(X_1), \mathbb{E}(X_2), \cdots, \mathbb{E}(X_n)]^T$$
$$\mu = [\mu_1, \mu_2, \cdots]^T$$

Well, what about the variance? If they are independent than sure, look at at the vector with all variances. But that does not really tell us a lot about how the variances play together for the different things we are measuring.
For dependance, we have something called the *covariance matrix*, which corresponds to the variance in the 1-dimensional case:

$$(\mathrm{Cov}\,(X))_{ik} = \mathrm{Cov}\,(X_i, X_j)$$

Anmärkning: Since the covariance is a symmetric operator, the covariance matrix is a symmetric matrix. It is also positive semi-definite.

If B is an $m \times n$ matrix and b is an m-vector, then we can look at $Y = BX + b = m$-dimensional vector. Recall from the 1-dimensional case that we had the following:

$$\mathbb{E}(Y) = B\mathbb{E}(X) + b \qquad \mathrm{Var}\,(Y) = B^2\mathrm{Var}\,(X)$$

In the multivariate case we have:

$$\mathbb{E}(Y) = [\mathbb{E}(Y_1), \mathbb{E}(Y_2), \cdots, \mathbb{E}(Y_n)] \qquad \mathrm{Cov}\,(Y) = \mathrm{Cov}\left(\sum_{i,j} B_{ij}X_j + b_i, \sum_{l,e} B_{le}X_e + b_l\right)$$

$$\text{Scalar product of vector with matrix yeilds:} \qquad \sum_{k=1}^{n}\sum_{p=1}^{n} B_{ik}B_{je}\mathrm{Cov}\,(X_k, X_e)$$

$$Y_i = \sum_{k=1}^{n} B_{ik}X_k + b_i \Rightarrow \mathbb{E}(Y_i) = \sum_{k=1}^{n} B_{ik}\mathbb{E}(X_k) + b_i \qquad \sum_{k=1}^{n} B_{ik} \sum_{l=1}^{n} \mathrm{Cov}\,(X_k, X_l)\, B_{ej}^T$$

$$\Rightarrow \mathbb{E}(Y) = B\mathbb{E}(X) + b \qquad BXB^T$$

Before we continue, it is useful to stop and recall the spectral theorem as well as some properties:

Definition/Sats 6.22: Spectral Theorem

A symmetric matrix A has n orthogonal eigenvectors $c_1, \cdots, c_n$, with associated eigenvalues d_i.

The matrix collection of the eigenvectors is an orthogonal matrix:

$$C^T C = I = C^{-1}C$$

If one wants to express x in the basis $c_1, \cdots, c_n$, we find a vector z such that $x = z_1c_1 + \cdots + z_nc_n = Cz$, where c_i is column vectors in C:

$$\Rightarrow z = C^T x$$

Let D be an orthogonal matrix where the diagonal elements are the eigenvalues. We get:

$$Ax = A\sum_{i=1}^{n} z_i c_i = CDz = CDC^T x$$
$$\Rightarrow A = CDC^T$$

Definition/Sats 6.23

When a matrix A is positive semi-definite and symmetric (like our covariance matrix), then $\exists\sqrt{A} = B$ such that $B \times B = A$

This follows from the spectral theorem.

Bevis 6.1

Since A is symmetric, then A can be diagonalized:

$$A = CDC^T$$

Since A is positive semi-definite, all $d_i \geq 0$. We can now construct a diagonal matrix consisting of $\sqrt{d_i}$ on the diagonal. Call this:

$$\widetilde{D} = C\begin{pmatrix} \sqrt{d_1} & \cdots & 0 \\ \vdots & \sqrt{d_2} & 0 \\ 0 & \cdots & \sqrt{d_n} \end{pmatrix} C^T$$
$$\Rightarrow B \times B = C\widetilde{D}C^T C\widetilde{D}C^t = CDC^T = A = B \times B = B^T B$$

□

A symmetric positive definite matrix (not semi-definite) has an inverse ⇔ the inverse is given by:

$$C\begin{pmatrix} \frac{1}{d_1} & \cdots & 0 \\ \vdots & \frac{1}{d_2} & 0 \\ 0 & \cdots & \frac{1}{d_n} \end{pmatrix} C^T = A^{-1}$$

Note that the inverse also has a square root:

$$C\begin{pmatrix} \frac{1}{\sqrt{d_1}} & \cdots & 0 \\ \vdots & \frac{1}{\sqrt{d_2}} & 0 \\ 0 & \cdots & \frac{1}{\sqrt{d_n}} \end{pmatrix} C^T = A^{-1/2}$$

Definition/Sats 6.24

The n-dimensional vector X is multivariate normal ⇔ $\forall n$ vectors a, the 1-dimensional vector $a^T x$ is normal

We write:

$$X \sim N_n(\mu, \Sigma)$$

Several properties from this definition follows directly:

- Every compontent in X is N_1 distributed
- $X_1 + \cdots + X_n \sim N_1$ distributed (let $a0(1, \cdots, 1)$)

- Each subset of $(X_1, \cdots, X_n)$ is also multivariate normal (mvn) since one can just pick a sufficient a

Definition/Sats 6.25

If $X \sim N(\mu, \Sigma)$ and $Y = BX + b$, then $Y \sim N(B\mu + b, B\Sigma B^t)$

Bevis 6.2

Show that it follows the definition:

$$a^T Y = a^T(BX + b) = a^T BX + a^T b = C^T X + d$$

Where $C^T X \sim N_1$ distributed and d is just a constant. □