YANDEX SCHOOL OF DATA SCIENCE

Probability and Statistics for Data Science

Lecture 3 – Joint and conditional distributions

Today

- Joint distribution
- Conditional distribution
- Independece of random variables
- Conditional expectation
- Law of total expectation
- Law of iterated variance
- Bayes rule
- Covariance
- Correlation
- Regression line between random variables

Joint distributions - why?

- Last time we talked about random variables
- We now understand how to deal with distributions and how to use them to model real-life situations
- But what if we need to involve more than one random variable in our probabilistic model?
- Today we will extend the concepts from lecture 2 to multiple random variables

Joint distribution – discrete

- **Def:** Let X and Y be discrete random variables associated with the same experiment. The joint PMF is given by
$$P_{X,Y}(x,y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\})$$
- More generally, for any set A of pairs (x,y), then
$$P\big((X,Y) \in A\big) = \sum_{(x,y)\in \mathrm{A}} P_{X,Y}(x,y)$$
- We can obtain the distribution of X from the joint PMF,
$$P_X(x) = \sum_{y} P_{X,Y}(x,y)$$

In this context, P_X is called the marginal PMF.

Example: discrete joint PMF

Description

1st	2nd	3rd	X = longest seq.	Y = #Heads
H	H	H	3	3
H	H	T	2	2
H	T	H	1	2
H	T	T	2	1
T	H	H	2	2
T	H	T	1	1
T	T	H	2	1
T	T	T	3	0

Joint PMF

Y/X	1	2	3	$P_Y(y)$
0	0	0	1/8	1/8
1	1/8	2/8	0	3/8
2	1/8	2/8	0	3/8
3	0	0	1/8	1/8
$P_X(x)$	2/8	4/8	2/8	1

Some properties of joint PMF's

- In the previous slide, we saw that summation over the elements of the table returns 1. This is not a coincidence, and it holds that

$$\sum_x \sum_y P_{X.Y}(x,y) = 1$$

- We can generate a new RV $Z = g(X,Y)$ and then,

$$P_Z(z) = \sum_{(x,y)|g(x,y)=z} P_{X,Y}(x,y)$$

- The extension of the expected value in this case is:

$$\mathrm{E}\big(g(x,y)\big) = \sum_x \sum_y g(x,y) P_{X,Y}(x,y)$$

More than two random variables

The extension for more than two random variables is natural. For example, for the random variables X, Y and Z we have

$$P_{X,Y,Z}(x, y, z) = P(X = x, Y = y, Z = z)$$

Without loss of generality,

$$P_{X,Y}(x, y) = \sum_{z} P_{X,Y,Z}\,(x, y, z)$$

Of course,

$$\sum_{x}\sum_{y}\sum_{z} P_{X,Y,Z}(x, y, z) = 1$$

Joint distribution – continuous

- The continuous counterpart is quite intuitive.
- **Def:** We say that two continuous RV's associated with the same experiment are <u>jointly continuous</u> and can be described in terms of a <u>joint PDF</u> $f_{X,Y}$, if $f_{X,Y}$ is nonnegative and satisfies

$$P\big((X,Y)\in B\big)=\iint_B f_{X,Y}(x,y)dxdy$$

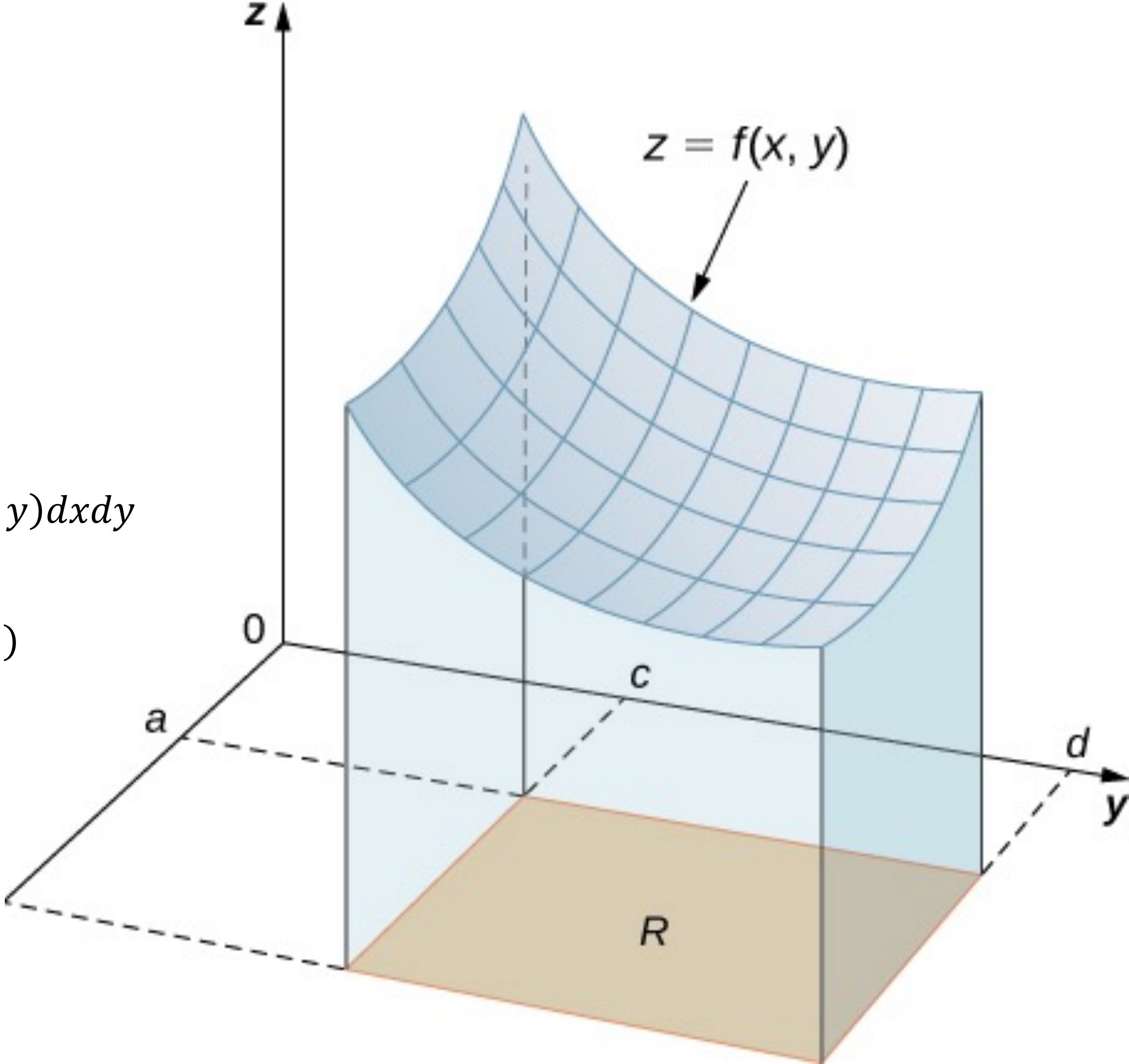
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy=1$
- The marginal density of X is

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

Interpretation of joint continuous dist.

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy$$

$$P(x \leq X \leq x + \delta_1, y \leq Y \leq y + \delta_2) \\ \approx f_{X,Y}(x, y)\delta_1\delta_2$$



More properties of joint cont.

- The expected value of a function $Z = g(X, Y)$ is

$$E\big(g(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

Joint CDF:

In general, the joint CDF is given by $F_{X,Y}(x,y) = P(X \le x, Y \le y)$.

If X and Y are described by a joint PDF, then

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

Equivalently,

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

Example

Let

$$f_{X,Y}(x,y) = c, 0 \leq x \leq 2, 1 \leq y \leq 5$$

Find c such that $f_{X,Y}$ is a joint PDF.

Sol:

$$\int_0^2 \int_1^5 c dy dx = \int_0^2 c(5-1) dx = 4c(2-0) = 8c = 1$$

Therefore,

$$c = 1/8$$

Conditioning

- Just like in the first lecture, we would like to introduce additional information to our model.
- We do that by conditioning on some event and next by another RV.
- In the following slides, we will extend the idea of conditional probability to random variables.
- Consequently, we will talk about independence of random variables as well.
- Dealing with random variables rather than events, we can talk about conditional expectation.

Conditioning on an event– discrete

- **Def:** The conditional PMF of a random variable X, conditioned on a particular event A with $P(A) > 0$ is

$$P_{X|A}(x) = P(X = x|A) = \frac{P(\{X = x\} \cap A)}{P(A)}$$

- Since the RV X forms a partition of the sample space,

$$P(A) = \sum_x P(\{X = x\} \cap A)$$

- Therefore,

$$\sum_x P_{X|A}(x) = 1$$

- Conclusion: $P_{X|A}$ is a legitimate PMF.

Example

Let X be the roll of a six-sided die and let A be the event that the roll is an even number. Compute the PMF of X given the event A.

Solution:

Applying the formula, we have

$$P_{X|A}(x) = P(X = x|roll\ is\ even) = \frac{P(X = x, roll\ is\ even)}{P(roll\ is\ even)}$$

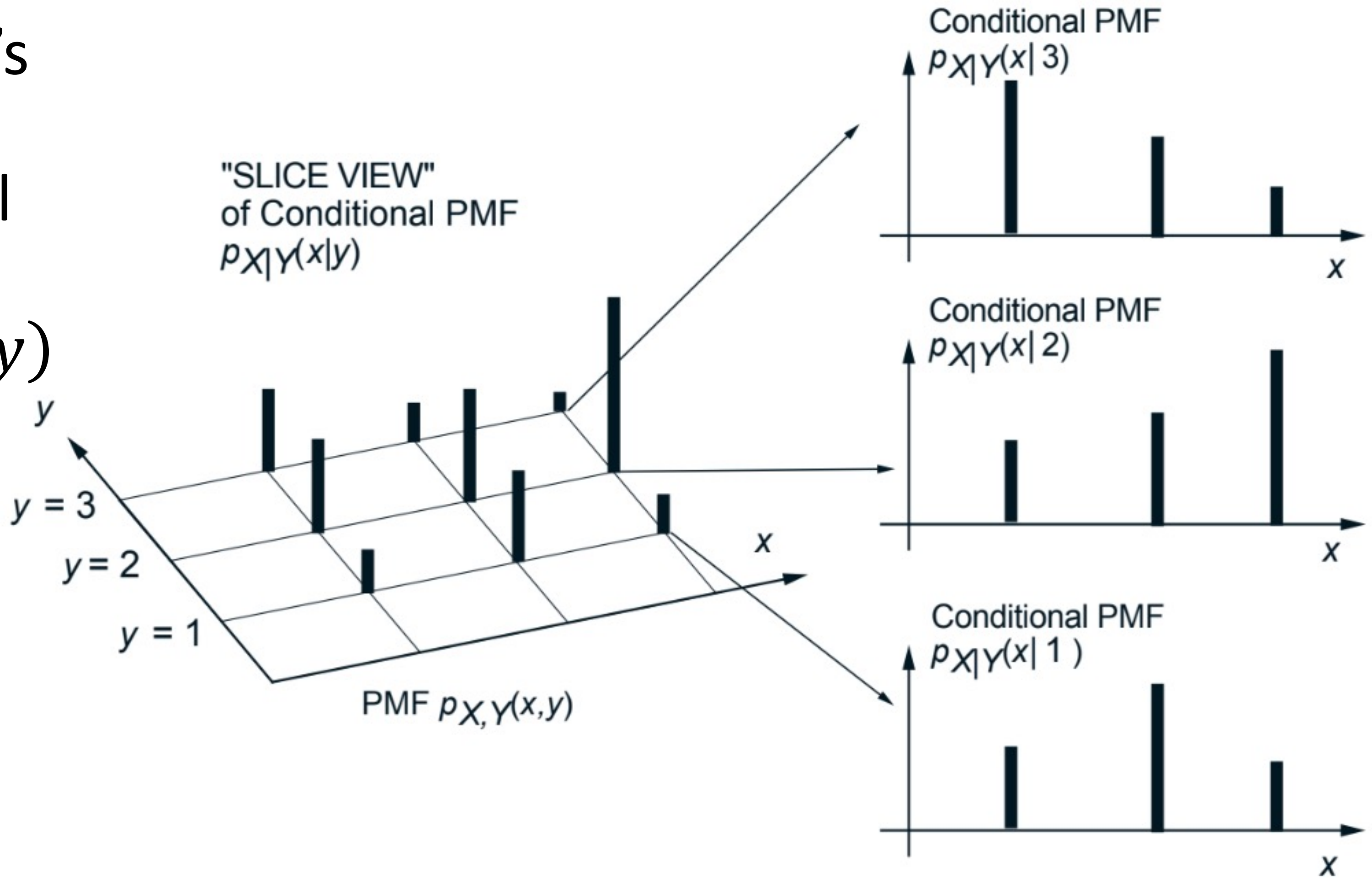
$$= \begin{cases} \dfrac{1/6}{3/6}, & if\ k = 2,4,6 \\ 0, & otherwise \end{cases}$$

Conditioning on a RV – discrete

- **Def:** Let X and Y be two RV's associated with the same experiment. The conditional PMF of X given Y is
$$P_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$
- Just like before, $P_{X|Y}$ is a legitimate PMF.
- Note that from the above definition,
$$P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y) = P_{Y|X}(y|x)P_X(x)$$
- Thus, we can obtain the marginal PMF of X by
$$P_X(x) = \sum_y P_{X|Y}(x|y)P_Y(y)$$

Conditioning on a RV – discrete

- **Def:** Let X and Y be two RV's associated with the same experiment. The conditional PMF of X given Y is
$$P_{X|Y}(x|y) = P(X = x|Y = y)$$
$$= \frac{P_{X,Y}(x,y)}{P_Y(y)}$$
- Just like before, $P_{X|Y}$ is a legitimate PMF.



Conditional expectation

- As the conditional PMF is a legitimate PMF, we can compute the associated expected value.
- The conditional expectation of X given A is defined by

$$E(X|A) = \sum_x x P_{X|A}(x)$$

- The conditional expectation of X given a value y of Y is

$$E(X|Y = y) = \sum_x x P_{X|Y}(x|y)$$

- If the events $A_1, \dots, A_n$ are a partition of the sample space,

$$E(X) = \sum_{i=1,\dots,n} E(X|A_i)P(A_i)$$

For X given Y we have,

$$E(X) = \sum_y E(X|Y = y)P_Y(y) = E\big(E(X|Y)\big)$$

Example

Let $X \sim Geo(p)$. That is, X is a random variable that counts the number of trials until (and including) the first success.

We will use conditional expectation to compute the expected value of X.

Let $A_1 = \{X = 1\}, A_2 = \{X > 1\}$.

$$E(X|A_1) = 1, E(X|A_2) = 1 + E(X)$$

Thus,

$$E(X) = E(X|A_1)P(A_1) + E(X|A_2)P(A_2)$$
$$= 1 \cdot p + \big(1 + E(X)\big)(1 - p) \Rightarrow E(X) = 1/p$$

Independence

- **Def:** The random variables X and Y are <u>independent</u> if for all x, y,
$$P_{X,Y}(x, y) = P_X(x)P_Y(y)$$
or equivalently if $P_{X|Y}(x|y) = P_X(x)$.
- **Theorem:** If X and Y are independent, then
$$E(XY) = E(X)E(Y) \text{ and } Var(X + Y) = Var(X) + Var(Y)$$
- **Lemma:** If X and Y are independent, then f(X) and $g(Y)$ are independent.
- The definitions of conditional independence and independence of more than two random variables extend similarly.

Conditional distributions - continuous version

- **Def:** The conditional PDF $f_{X|A}$ of a continuous random variable X given an event A satisfies

$$P(X \in B|A) = \int_B f_{X|A}(x)dx$$

- If $A \subset \mathbb{R}$ with $P(X \in A) > 0$, then

$$f_{X|\{X \in A\}}(x) = \frac{f_X(x)}{P(X \in A)}, if\ x \in A\ and\ 0\ otherwise$$

- If $A_1, \ldots, A_n$ is a partition of the sample space,

$$f_X(x) = \sum_{i=1,\ldots,n} P(A_i) f_{X|A_i}(x)$$

Conditional distributions - continuous version

- **Def:** For two continuous random variable X and Y with joint PDF $f_{X,Y}$, the conditional PDF of X given $Y = y$ is defined by
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$
- The normalization property holds for the conditional PDF:
$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) = 1$$
- The facts about conditional expectation extend naturally.
- What’s flawed in our theory?

Example: Exponential RV is memoryless

The time until a lightbulb burns out is denoted by X and has exponential distribution with parameter λ. John leaves the room and returns t time units later. The light in the room is still on. Let T be the additional time until the light bulb burns out. What is the CDF of T given the above event?

Sol: We are interested in the probability $P(T > x|X > t)$.

$$P(T > x|X > t) = P(X > x + t|X > t) = \frac{P(X > x + t, X > t)}{P(X > t)}$$

$$= \frac{P(X > x + t)}{P(X > t)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = P(X > x)$$

Independence of continuous RV's

- **Def:** The continuous random variables X and Y are independent if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all x, y.

  Equivalently, if $f_{X|Y}(x|y) = f_X(x)$.
- It follows that if X and Y are independent,
$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

Specifically, $F_{X,Y}(x, y) = F_X(x) F_Y(y)$.

- **Convolution:** Let X and Y be two continuous RV's and define $Z = X + Y$. Then,
$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z - x) dx$$

Law of iterated variance

A convenient theorem that follows from all the above theory of conditional expectation is the following:

$$Var(X) = E\big(Var(X|Y)\big) + Var\big(E(X|Y)\big)$$

Where $Var(X|Y) = E(X^2|Y) - E^2(X|Y)$.

Example: We toss a coin n independent times with probability of H being a random variable $Y \sim U(0,1)$. What is the variance of the number of Heads X?

Sol: We know that $E(X|Y) = nY, Var(X|Y) = nY(1-Y)$. Thus,

$$Var(X) = Var(nY) + E\big(nY(1-Y)\big) = \cdots$$

Bayes rule

- First, we have that for continuous X and Y,
$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t)f_{Y|X}(y|t)dt}$$
- Now, if the purpose is to learn from/about a discrete random variables N using continuous measurements Y, we can use the formulas
$$f_Y(y)P(N = n|Y = y) = P_N(n)f_{Y|N}(y|n)$$

Resulting in
$$P(N = n|Y = y) = \frac{P_N(n)f_{Y|N}(y|n)}{f_Y(y)} = \frac{P_N(n)f_{Y|N}(y|n)}{\sum_i P_N(i)f_{Y|N}(y|i)}$$

And the formula for $f_{Y|N}$ is analogous.

Example: signal detection

A binary signal S is transmitted, and we know that

$$P(S=1)=p, P(S=-1)=1-p$$

The received signal is $Y=N+S$, where $N \sim N(0,1)$, independent of Y. What is the probability that $S=1$ as a function of the observed value y of Y?

Sol: Conditioned on $S=s$, Y has $N(s,1)$ distribution. Therefore,

$$P(S=1|Y=y)=\frac{P_S(1)f_{Y|S}(y|1)}{P_S(1)f_{Y|S}(y|1)+P_S(-1)f_{Y|S}(y|-1)}$$

$$=\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-1)^2}{2}}\cdot p}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-1)^2}{2}}\cdot p+\frac{1}{\sqrt{2\pi}}e^{-\frac{(y+1)^2}{2}}\cdot(1-p)}=\frac{pe^y}{pe^y+(1-p)e^{-y}}$$

Covariance

- f_X contains all the information regarding X.
- $f_{X,Y}$ contains all the information regarding X and Y together.
- $E(X)$ summarizes the mean value of X
- $Var(X)$ summarizes the way in which X changes.
- $E\big(g(X,Y)\big)$ summarizes the mean value of a function of X and Y.
- How can we summarize the change in X and Y together?
- **Def:** The covariance of X and Y is

$$Cov(X,Y) = E\Big(\big(X - E(X)\big)\big(Y - E(Y)\big)\Big)$$

Properties of covariance

- $Cov(X,Y) = Cov(Y,X) = E(XY) - E(X)E(Y)$
- If X and Y are independent, then $Cov(X,Y) = 0$. The other way around is not necessarilly true!
- $Cov(X,X) = Var(X)$
- $Cov(aX + b, cY + d) = ac \cdot Cov(X,Y)$
- $Cov(X, Y + Z) = Cov(X,Y) + Cov(X,Z)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Correlation

- Def: For the random variables X and Y, their correlation coefficient is given by

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

Properties:

1. The value of the correlation is not affected by measures of unit.
2. $Corr(aX+b, cY+d) = sign(ac)Corr(X,Y)$
3. $-1 \leq Corr(X,Y) \leq 1$
4. $|Corr(X,Y)| = 1$ if and only if Y is a linear function of X.

The regression line between random variables

Say that we want to predict the value of Y using the value of X with the linear formula

$$\hat{Y} = aX + b$$

We would like to find a, b such that the mean squared error between the prediction and the real value is minimized. That is,

$$a_{min}, b_{min} = argmin_{a,b} E\left(Y - \hat{Y}\right)^2$$

It turns out that

$$a_{min} = Corr(X,Y)\frac{SD(X)}{SD(Y)}, b_{min} = E(Y) - a_{min}E(X)$$

The line $y = a_{min}x + b_{min}$ is called the regression line of Y on X.

The regression line between random variables

- Note that

$$E\left(Y-\hat{Y}\right)^2 = Var(Y-aX) + E^2(Y-b-aX)$$

The first expression does not depend on b, and the second expression vanishes for our choice of b_{min}.

- If we plug-in a_{min} in the first expression, we get (after some algebra) that

$$E\left(Y-\hat{Y}\right)^2 = Var(Y)\left(1-Corr^2(X,Y)\right)$$

Proving that $|Corr| \leq 1$ but also that the correlation is an indicator for the strength of the linear relationship (either positive or negative) between X and Y.

http://guessthecorrelation.com/

References

- Bertsekas, Dimitri P. and Tsitsiklis, John N.. "Introduction to Probability." 2008 .
- Ross, Sheldon M. A First Course in Probability / Sheldon Ross. Eighth edition, global edition. Harlow: Pearson Education Limited, 2010.
- Haviv, Moshe. Introduction to Descriptive Statistics and Probability, 2021.
- Wasserman, Larry. *All of statistics : a concise course in statistical inference*. New York: Springer, 2010.