# Probability - quick reminder

lecture notes: https://www.stat.berkeley.edu/~aldous/134/gravner.pdf

Why do we need probability theory?

We always assume that the data we get is coming from some distribution, so there is some law in the background, that we will model using notions used in probability theory.

## Definitions

An **experiment** is any action or process whose outcome is subject to uncertainty.

The **sample space** of an experiment, $\mathcal{S}$, is the set of all possible outcomes.

An **event** is any collection of outcomes of (a subset of) $\mathcal{S}$.

An event is **simple** or **compound** if it consists of one outcome or more than one outcome, respectively.

**Example**: experiment is flipping a coin twice, then $\mathcal{S} = \{hh, ht, th, tt\}$, "the first is head" is an event, $hh$ is a simple event, while the first is head$= \{hh, ht\}$ is a compound event.

Making new events from old:

The **union** of events $A$ and $B$, denoted $A \cup B$, is the event that consists of all that outcomes that are either in $A$ or $B$ or both (the inclusive or)

The **union** of events $A_i$ $(i \in \mathbb{N})$ is the event that consists of all that outcomes that are in at least on of the $A_i'$s.

The **intersection** of events $A$ and $B$, denoted $A \cap B$, is the event that consists of all the outcomes for which both $A$ and $B$ occur. (and we can also consider the countable intersection)

The **complement** of an event $A$, denoted $\overline{A}$, is the set of all outcomes in not in $A$ (but are in $\mathcal{S}$)

The **null set** or **empty set**, denoted  is the set with nothing in it. Two sets $A$ and $B$ are **mutually exclusive** or **disjoint** if there is nothing in their intersection (i.e. $A \cap B =$)

## Some notation

We can intersect/union together more than two sets sets at a time

$$\bigcap_{i=1}^{20} A_i = A_1 \cap A_2 \cap \cdots \cap A_{19} \cap A_{20}$$

or

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \cdots$$

The sets are basically questions we can ask about something random...but how do we talk about the chances that those things will occur?

We need a function that takes events as input, and as output gives us a number between 0 and 1. This is our probability function $P(\cdot)$.

## Probability - basic notions - definition of probability

Here are the rules any probability function has to satisfy;

1. For any event $A$, $P(A) \geq 0$
2. $P(\mathcal{S}) = 1$
3. let $A_1, A_2, \ldots$ be a countably infinite collection of disjoint sets; then $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$

Here are some things you can make sure are true using only the stuff from the previous few lines:

1. $P(\emptyset) = 0$
2. $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ if $A_1, \ldots, A_n$ are all disjoint
3. $P(\overline{A}) = 1 - P(A)$ for any $A$
4. if $A \subset B$ then $P(A) \leq P(B)$
5. (in particular) for any $A$, $P(A) \leq 1$
6. for any $A$ and $B$ (not necessarily disjoint), $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

## Conditional Probability

**Motivation:**
Sometimes the probabilities of events change with the set of available information that we have.

Example: Let $A = \{\text{Google's share price increases tomorrow}\}$. Then $P(A) \approx 0.5$

But if we let $C = \{\text{the government bans access to google.com}\}$, what's $P(A \text{ given } C)$?

The notation: we'll write $P(A \text{ given } C)$ like $P(A|C)$. And here's the formula:

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

**Remarks:**

1. It helps to think of this in the context of Venn diagrams
2. $C \neq \emptyset$, otherwise we're dividing by 0 using this definition
3. sometimes we'll use $P(A|C)P(C) = P(A \cap C)$
4. it is strongly related to Bayes' theorem that we skip now

**Example**
Using $P(A|C)P(C) = P(A \cap C)$ is useful in situations where there is some sort of sequential thing going on, or there are several stages of some random process.

**Example**: you're only interested in if the stock market goes up or down. The events that it goes up on Monday, Tuesday, Wednesday, Thursday and Friday are $A_1, A_2, A_3, A_4$ and $A_5$, respectively. The changes that it goes up or goes down only depend on what happened the day before (unless it's a Monday in which case let's assume it goes up or down with probability 0.5). Say there's a 75% chance that it does the same thing as it did the day before. This means that there is a 25% chance it does the opposite. What's the probability the market goes up every day of the week?

$$\begin{aligned}
P(\text{goes up every day}) &= P(A_5 \cap A_4 \cap \cdots \cap A_1) \\
&= P(A_5|A_4 \cap A_3 \cap A_2 \cap A_1) \cdot P(A_4|A_3 \cap A_2 \cap A_1) \cdot P(A_3|A_2 \cap A_1) \cdot P(A_2|A_1) \cdot P(A_1) \\
&= P(A_5|A_4) \cdot P(A_4|A_3) \cdot P(A_3|A_2) \cdot P(A_2|A_1) \cdot P(A_1) \\
&= (0.75) \cdot (0.75) \cdot (0.75) \cdot (0.75) \cdot (0.5)
\end{aligned}$$

## Independence

### Motivation + definition

When we were talking about conditional probability before we usually talked about examples where either $P(A) > P(A|B)$ or $P(A) < P(A|B)$. Quite often we'll assume that this isn't true when we're putting together statistical models, though.

Two events $A$ and $B$ are **independent** if $P(A|B) = P(A)$.

Two events $A$ and $B$ are **dependent** if $P(A|B) \neq P(A)$

**Equivalently:**

Events $A$ and $B$ are **independent** if

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

this one is less intuitive but more general; it holds when we're talking about events that might have probability 0. It also extends more easily to when we talk about independence between more than two things at a time.

$A_1, \ldots A_n$ are **mutually independent** if for every $k = 1, 2, \ldots, n$ and every subset of indices you can make with such a $k$ $i_1, \ldots, i_k$

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k})$$

Note that this is stronger than pairwise independence.

## Random variables

### Motivation

Even though sometimes one might think of a random variable $X$ as just a random number, technically they're actually functions. We'll define random variables in a more technical way so that a) you won't be totally confused when you get to measure-theoretic probability and b) we can classify rvs further (which helps us with modelling and doing stuff in statistics).

**Definitions:**

A **random variable** (rv) is a function whose domain is the sample space $\mathcal{S}$ and whose range is a subset of the real numbers $\mathbb{R}$

A **Bernoulli** (or indicator) rv is one that maps into $\{0, 1\}$

**Example**: Let $\mathcal{S}$ be the set of outcomes of how many gas pumps are active at each of two gas stations. Let's say the first gas station has five pumps, and the second has 6 pumps. Then $\mathcal{S} = \{1, 2, 3, 4, 5\} \times \{1, 2, 3, 4, 5, 6\}$.

Let $X$ be the sum of the number of active gas pumps, then we can write $X$ as $X(s_1, s_2) = s_1 + s_2$.

But $Y$ can be the absolute difference between the number of gas pumps active, or $U$ be the maximum number of gas pumps active.

We can classify rvs by what their range is.

A **discrete** rv is one whose range is finite or countably infinite.

A **continuous** rv is one whose a) range is a union of real intervals, and b) has no mass on any single number (i.e. if we call our rv $X$ then for any number $c$, $P(X = c) = 0$)

(Note: If you've taken a real analysis course before, you might know about the different types of infinity. For this class, we won't be doing analysis; all this helps to recognize certain probability models/distributions. Also, when we do stuff with discrete random variables we're typically using sums and differences. If we're doing stuff with continous rvs, then we're probably going to be using integrals and derivatives.)

**Examples**

Our last gas station example was a collection of discrete random variables.

Examples of continuous rvs: tomorrows return for a stock, or a randomly selected persons height.

### Distribution of discrete variables

First we work with a general discrete rvs. We need to do this before we start using specific discrete rvs (i.e. Binomial, Poisson, etc.)

A **probability distribution** is a function that shows how the total probability of 1 is distributed among the possible values of a rv (say $X$).

A **probability mass function** is a probability distribution for a discrete random variable. For each possible number $x$ that $X$ can take on, it gives you $P(X = x)$.

1. we talked about probability function on $\mathcal{S}$ before. How does that relate to this probability function on the range of $X$? Well, for a discrete rv: $P(X = x) = P(\{s \in \mathcal{S} : X(s) = x\})$

2. All those rules we talked about still apply. Check as many as you want to.

3. For the probability mass function we have that it it non-negative (zero except for a countably infinite set) and the sum of the values is 1. And it exactly characterizes the discrete probability mass functions.

4. Knowing the distribution allows you to do essentially anything you want to (i.e. find the probability of any event, compute moments, etc)

When we're dealing with random variables and we want to specify a distribution, typically it isn't feasible to just pick numbers for the probability of each outcome...what if there are an infinite amount? Fortunately there are many probability distributions where we only have to specify one or two **parameters**, and then it fills out the rest.

A **parameter** (for a discrete rv) is a quantity that can change (certain) distribution (probability mass) functions $P(X = x) = p(x)$. Let's say there is only one parameter in our situation, and call it $\alpha$. To make this dependence on $\alpha$ explicit, sometimes we write our pmf as $p_\alpha(x)$ or $p(x; \alpha)$.

**Example:** We observe the gender of newborns in a hospital until a boy ($B$) is born. So $\mathcal{S} = \{B, GB, GGB, GGGB, \ldots\}$. There are a countably inifinite number of possible situations. What if we just assume that there is one probability that an individual birth results in a boy, and all births are independent from one another? Let's call $P(B) = p$.

$X$ = number of births observed. Then $p(x) = (1 - p)^{x-1}p$, $x > 0$

If our assumptions are true, we just have to pick a good value for $p$. This distribution is called the **geometric distribution** of paramter $p$.

The **cumulative distribution function (cdf)**: $F_X(x) = P(X \leq x) = \sum_{k \leq x} p(k)$

Instead of going from pmf to cdf, we can also go from cdf to pmf also: For any $a$ and $b$, with $a \leq b$

$$P(a \leq X \leq b) = F_X(b) - F_X(a-)$$

By a=b, we get the pmf.

Note: $F_X(x-) = \lim_{\epsilon \to 0} F_X(x-\epsilon)$

Note: A lot of times $F_X(x-) = F_X(x-1)$

### Expectation and variance

Let $X$ be a discrete rv with range $D$ and pmf $p(x)$. The **expected value** or **mean** of X, $E(X)$ is

$$E(X) = \sum_{x \in D} x \cdot p(x)$$

We say it exists if it exists and finite.

**Motivation:**

Say we have $X$. We can make a new rv $Y = h(X)$ with some function $h(\cdot)$. It would be true that $E(Y)$ could be found using the formula above, but we would need $p_Y(y)$ to do that. We would have to find that from $p_X(x)$. That's a pain. Good news though: we don't have to find the new distribution, though.

Say we start out with $X$ and $p_X(x)$. Then for any function $h(\cdot)$,

$$E(Y) = E[h(X)] = \sum_x h(x)p_X(x)$$

(we're assuming here that these expected values exist i.e. that they're finite)

This is called the law of the unconscious statistician (LOTUS).

A lot of times $h(\cdot)$ is a linear transformation. In this case

$$E[aX+b] = aE(X)+b$$

where $a$ and $b$ are constants

Here's the **definition** of population variance. Let $D$ be the range of a rv $X$. Let $\mu = E(X)$ (it's easier to write it this way). Then the **variance** of $X$, call it $\sigma^2(X)$ is:

$$\sigma^2(X) = \sum_{x \in D} (x-\mu)^2 \cdot p(x) = E[(X-\mu)^2]$$

**Standard deviation** is just the square root of this.

This is an average again, but we're not taking the average of $X$. We're taking the average of a nonlinear transformation of $X$: $(X-\mu)^2$.

Sometimes we use this formula:

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

We also have this:

$$\sigma^2(aX+b) = a^2\sigma^2(X)$$

### Introduction of binomial distribution - motivation

Many experiments conform either exactly or approximately to the following list of requirements:

1. The experiment consists of a sequence of $n$ *trials* ($n$ is nonrandom)

2. Each trial results in one of two outcomes (success or failure)

3. Trials are independent of one another

4. The probability of a success is the same for all trials

An experiment for which these conditions are satisfied is called a **binomial experiment**

A **binomial rv** is the random variable denoting the number of successes for such an experiment For a binomial rv, the pmf is

$$p(x; n, p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \dots, n$$

and the cdf is

$$F_X(x) = P(X \leq x) = \sum_{k=0}^{x} p(k; n, p)$$

**Example**

A basketball player is a 70% free-throw shooter. He has to make both of two shots to win the game for his team. What is the probability he makes one shot? What is the probability he doesn't win the game for his team?

$$P(\text{he makes one}) = p(1; 2, 0.7)$$

$$P(\text{loses}) = P(X < 2) = P(X \leq 1) = p(0; 2, 0.7) + p(1; 2, 0.7)$$

**Mean, Variance:**

A few things about the binomial distribution

1. if $n = 1$, it's called a Bernoulli distribution/rv
2. in general, $EX = np$ and $V(X) = np(1-p)$

proofs are left as exercise

**Motivation:**

A lot of discrete rvs arise from simple experiments consisting of trials with a finite number of possible outcomes (e.g. binomial, multinomial, hypergeometric, negative binomial, etc). This one doesn't, but it arises from taking a limit of a binomial rv.

To make it precise, if $p = \frac{\lambda}{n}$, then binomial distribution tends to the Poission distribution, where a rv $X$ is said to have a **Poisson distribution** with parameter $\lambda > 0$ if its pmf is

$$p(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad x = 0, 1, \dots$$

### Continuous random variables

Our discrete rvs had a finite or countably infinite number of possible values it could take on. Now we'll talk about continuous rvs. They can take on a whole interval/range of possible values.

**Definitions:**

Let $X$ be a continuous rv. The **probability density function** pdf of $X$ is the function $f(x)$ such that for any $a \leq b$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

...it also has these properties

1. $f(x) \geq 0$ for all $x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

The **cumulative distribution function (cdf)** for a cts rv $X$ is this function

$$F(x) = \int_{-\infty}^{x} f(s)ds$$

Properties:

1. $\lim_{x\to-\infty}$ F(x)$= 0$
2. $\lim_{x\to\infty}$ F(x)$= 1$
3. monotone increasing
4. (in general: continuous from the right)

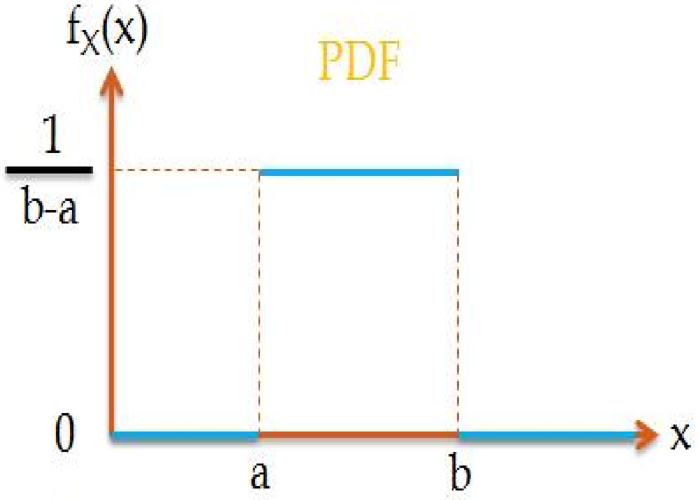
If $X$ is continuous with pdf and cdf $f(x)$ and $F(x)$, respectively, then

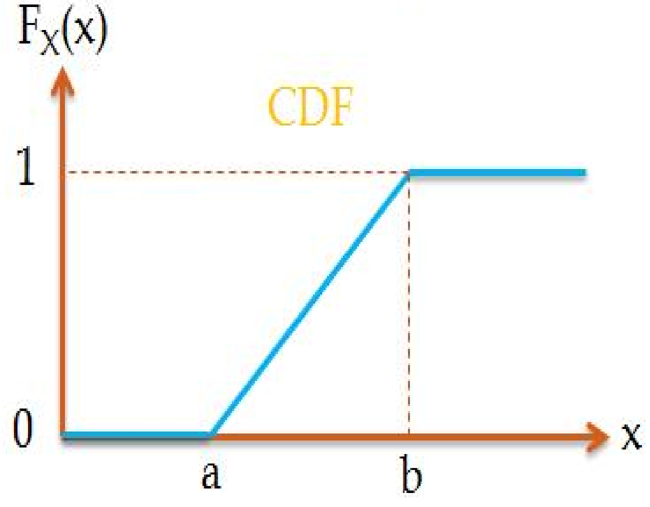
$$f(x) = \frac{d}{dx}F(x)$$

(because of the fundamental theorem of calculus)

Example of continous rv No. 1: the **uniform distribution** on the interval $[a, b]$:

$$f(x) = f(x; a, b) = (b-a)^{-1}, \quad a \le x \le b$$





assuming $a \le x \le b$...

$$F(x) = \int_{-\infty}^{x} f(s)ds = \int_{a}^{x} \frac{1}{b-a}ds = \frac{1}{b-a}(s)|_{s=a}^{s=x} = \frac{x-a}{b-a}$$

Here is an important thing to remember about continuous rvs:

$$P(X = c) = \int_c^c f(x)dx = \lim_{\epsilon \to 0} \int_{c-\epsilon}^{c+\epsilon} f(x)dx = 0$$

so for any $a, b$

$$P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b)$$

(we don't have to be careful about our inequalities like we do with discrete rvs...)

Let $X$ be a cts rv with pdf $f(x)$ and cdf $F(x)$. Then for any $a$

$$P(X > a) = 1 - F_X(a)$$

$$P(a \leq X \leq b) = P(a < X \leq b) = F_X(b) - F_X(a)$$

A **percentile** $\eta(p)$ associated with some percent $p$ of a cts rv $X$ is some number such that

$$p = F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(s)ds$$

i.e. it's some number such that $p$ % of the data is behind it

Note: **quantiles** are just special cases of percentiles (e.g. quartiles = 25th,50th,75th percentiles) (e.g. deciles = 10th,20th,30th,40th,50th,60th,70th,80th,90th percentiles)

Note: the **median** is the 50th percentile

A cts rv $X$ is **symmetric** if there is some point $c$ such that

$$f(c - s) = f(c + s)$$

for all $s$

## Normal distribution

A good example for continuous rv is the normal distribution that is the most important one in all of probability and statistics."

a cts rv $X$ has a **normal distribution** with parameters $\mu$ and $\sigma^2$ if it has the pdf

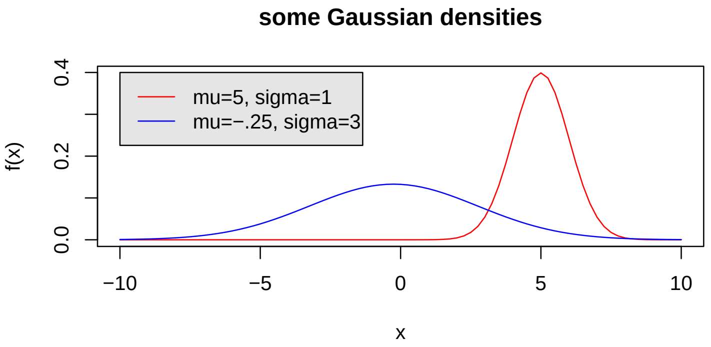
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

where $-\infty < \mu < \infty$, $\sigma^2 > 0$ and $-\infty < x < \infty$

shorthand: $X \sim \mathcal{N}(\mu, \sigma^2)$

Here's a picture of two examples of this function

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



**Expectation, variance:**

The **expected value** of a cts rv $X$ with pdf $f(x)$ is

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

and the expectation of any function $h(X)$ (LOTUS) is

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx$$

The **variance** of a cts rv with mean $\mu$ is

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

and this works again too:

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

**Propositions**
And these are true still as well:

$$E[aX+b] = aE[X]+b$$

$$\sigma^2[aX+b] = a^2\sigma^2[X]$$

A special instance of the normal distribution is the **standard normal distribution**. You just set $\mu = 0$ and $\sigma^2 = 1$

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$$

where $-\infty < z < \infty$

shorthand: $Z \sim \mathcal{N}(0,1)$

It's convention to use capital $Z$ when we're talking about standard normal rvs
When we're talking about probabilities, we can "do algebra" inside the parentheses. E.g:

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

or

$$P\left(\frac{1}{X} > -c\right) = P\left(X > \frac{1}{-c}\right)$$

with $X, c > 0$ for the last one.
We'll generalize this a bit later, but if

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

then

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

That's why we go back and forth between $X = \mu + \sigma Z$ and $Z = \frac{X-\mu}{\sigma}$ a lot. They are both normal, but sometimes one will have more convenient parameters than the other.

**Joint distributions - Definition**

When multiple rvs vary together, we need a **joint distribution** to fully describe them.

For discrete rvs, we have the **joint probability mass function**

$$p(x,y) = P(X = x \cap Y = y)$$

also, if $A$ is a two-dimensional set:

$$P[(X,Y)\in A]=\sum_{(x,y)\in A}\sum p(x,y)$$

The **marginal probability mass functions** can be obtained from the joint via summation...

$$p_X(x)=\sum_y p(x,y)$$

and

$$p_Y(y)=\sum_x p(x,y)$$

Let $X$ and $Y$ be two cts rvs. Then $f(x,y)$ is the **joint probability density function** if for any $A$

$$P[(X,Y)\in A]=\iint_A f(x,y)dxdy$$

also:

1. $\iint_{\mathcal{S}} f(x,y)dxdy=1$
2. $f(x,y)\geq 0$

We say

$$P[(X,Y)\in A]=\iint_A f(x,y)dxdy$$

...but doing this in practice is more difficult. We need to be very careful when we find the bounds of integration since we're dealing with more than one random variable!!!

We can obtain marginal pdfs via integration

$$f_X(x)=\int_y f(x,y)dy$$

$$f_Y(y)=\int_x f(x,y)dx$$

**Propositions**

Let $X$ and $Y$ be jointly distributed random variables with pmf $p(x,y)$ or pdf $f(x,y)$ (according to whether the rvs are discrete or continuous). Then

$$E[h(X,Y)]=\sum_x\sum_y h(x,y)p(x,y)$$

if they're jointly discrete, or

$$E[h(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)f(x,y)dxdy$$

if they are continuous.

Note: even though $h(X,Y)$ is a random variable itself, and it has a new density, we don't have to use that to find it's expected value. (LOTUS)

Two random variables are **independent**, if the cdf of their joint distribution is the product of the cdf's.

Let $X_1,X_2,\ldots,X_n$ be independent random variables and assume that all the expected values we write down exist. Then

$$E[h_1(X_1)h_2(X_2)\cdots h_n(X_n)]=E[h_1(X_1)]\times\cdots\times E[h_n(X_n)]$$

### Covariance and correlation

The **covariance** between two rvs $X$ and $Y$ is

$$\text{Cov}(X,Y) = E[(X - EX)(Y - EY)]$$

if these are jointly cts, then

$$\text{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_Y)f(x,y)dxdy$$

if discrete then

$$\text{Cov}(X,Y) = \sum_x\sum_y(x - \mu_X)(y - \mu_Y)p(x,y)$$

If $\text{Cov}(X,Y) > 0$ that means $(X - \mu_X)$ and $(Y - \mu_Y)$ tend to be the same sign (both negative, or both positive).

If $\text{Cov}(X,Y) < 0$ that means $(X - \mu_X)$ and $(Y - \mu_Y)$ tend to be opposite signs (ones positive and the other is negative, and vice versa)

Remember that this is a probability weighted average. If a covariance is positive, for example, that tells you they have the same sign on *average*.

Covariance formula

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

It is sometimes easier to use this formula! For example, if you have the first two derivatives of a moment generating function.

**Proposition**

The following is called *bilinearity*. If $a$ and $b$ are constants, $X$, $Y$ and $Z$ are three rvs...

$$\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$$

$$\text{Cov}(X,Y+Z) = \text{Cov}(X,Y) + \text{Cov}(X,Z)$$

$$\text{Cov}(aX,Y) = \text{Cov}(X,aY) = a\,\text{Cov}(X,Y)$$

What's $\text{Cov}(X,X)$?

$$\begin{aligned}\text{Cov}(X,X) &= E[(X - EX)(X - EX)]\\ &= E[(X - EX)^2]\\ &= \sigma^2[X]\end{aligned}$$

**Correlation - Motivation**

"It would appear that the relationship example A is quite strong since $\text{Cov}(X,Y) = 1875$, whereas in the example B $\text{Cov}(X,Y) = -\frac{2}{75}$ would seem to imply quite a weak relationship."

This isn't true, though. You can't compare covariances. The number depends on the scale. For instance, if you compared the covariance between prices of one item with another item, it totally depends on whether you're talking about dollars or cents.

You can see this formally just using bilinearity. For any $X$, $Y$, and $a, b \neq 1$, $\text{Cov}(aX,bY) = ab\,\text{Cov}(X,Y) \neq \text{Cov}(X,Y)$.

That's why we talk about **correlation**

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

It might help to think about it like this:

$$\mathrm{Corr}(X,Y) = \mathrm{Cov}\left[\frac{X}{\sigma(X)}, \frac{Y}{\sigma(Y)}\right].$$

A few more things

1. if $a$ and $c$ have the same sign, $\mathrm{Corr}(aX+b, cY+d) = \mathrm{Corr}(X,Y)$
2. $-1 \leq \mathrm{Corr}(X,Y) \leq 1$
3. if $X$ and $Y$ are independent, then $\mathrm{Corr}(X,Y) = 0$
4. if $\mathrm{Corr}(X,Y) = 0$, $X$ and $Y$ can still be dependent
5. $|\,\mathrm{Corr}\,| = 1$ if and only if $Y = aX + b$ for some scalars $a, b$ with $a \neq 0$

**Proof of 3**

Prove that “independence" is stronger than “uncorrelated." That is prove that independence implies $\mathrm{Corr}(X,Y) = 0$.

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$$

So $\mathrm{Corr}(X,Y) = 0$.