# Probability notes

## Definitions etc

Given a collection of subsets $A_i \subseteq S$ of a sample space $S$, a function $P : S \to \mathbb{R}$ is a probability if it satisfies the 3 probability axioms:

- $P(A_i) \in [0, 1] \quad \forall A_i \in S$
- If $A_i$ and $A_j$ are disjoint, then $P(A_i \cup A_j) = P(A_i) + P(A_j)$
- $P(S) = 1$

From this you can derive everything else. For example $P(\emptyset) = 1 - P(S^c) = 1 - 1 = 0$. For other things you need to be a bit sneaky, the main bit of info you have is about disjoint sets (2nd bullet), so most proofs involve writing things as disjoint sets.

### Example

For any 2 set $A, B \subseteq S$, show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$:
First write everything as disjoint sets

$$\begin{aligned} A &= (A - B) \cup (A \cap B) \\ B &= (B - A) \cup (A \cap B) \\ A \cup B &= (A - B) \cup (A \cap B) \cup (B - A) \end{aligned}$$

Then use the rules of probability:

$$\begin{aligned} P(A) &= P(A - B) + P(A \cap B) \\ P(B) &= P(B - A) + P(A \cap B) \end{aligned}$$

$$\begin{aligned} P(A \cup B) &= P[(A - B) \cup (A \cap B) \cup (B - A)] \\ &= P(A - B) + P(A \cap B) + P(B - A) \\ &= P(A - B) + P(A \cap B) + P(B - A) + [P(A \cap B) - P(A \cap B)] \\ &= [P(A - B) + P(A \cap B)] + [P(B - A) + P(A \cap B)] - P(A \cap B) \\ &= P(A) + P(B) - P(A - B) \quad \blacksquare \end{aligned}$$

The *conditional probability* is the probability of an event $A$, given event $B$ has occurred:

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

Think of it as restricting the sample space down to events where $B$ occurs, and renormalising so the probabilities within this 'restricted sample space' obey the usual rules. the thing on top, $P(AB)$ is called the *joint probability* - its the probability of events $A$ and $B$ occurring. If $A$ and $B$ are independent then the joint factors - $P(AB) = P(A)P(B)$.

Conditional probabilities are very important. You can reverse the order of the conditioning using *bayes rule*:

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

You can find $P(B)$ using the *law of total probability*. if the sample space $S$ can be partitioned into a union of disjoint sets $\{B_i\}$, then any event $A = \bigcup_{i=1}^{n} A \cap B_i$. These are all disjoint, so the rules of probability say $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$. Rearranging the definition of conditional probability gives $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$.

If $X$ is a random variable (just think of it as the outcome of an experiment, no need to get into the definition) then $P(X = x)$ is known as a *probability distribution*. For some reason distributions get different names depending on if $X$ is continuous (call it a density) or discrete (call it a distribution). The reason why we do this is because, for continuous $X$, $P(X = x) = 0$. The probability of $X$ taking on exactly the value 7.2421452 for example, all the way down to the last decimal place, is extremely unlikely. For continuous distributions it only really makes sense to talk about the probability of $X$ being in some interval - things like $P(X \in [a, b)]$. To get probabilities in the continuous case, you need to multiply the probability density by the length of the interval. Quantities of the form $P(X = x)dx$ are probabilities (here $dx$ is a small region containing $x$). The density is not a probability! It doesn't need to follow the 3 rules, and in particular it can be greater than 1.

Since distributions are just probabilities, the definitions for joint & conditional carry over. Let $f$ be some distribution, the *conditional distribution* of a random variable $X$ given another random variable $Y$ is:

$$f_{X|Y} = \frac{f_{XY}}{f_Y}$$

$f_{XY}$ is the *joint distribution* of the random variables $X$ and $Y$. If you have a joint distribution $f_{XY}$ then $f_X$ is called a *marginal distribution*. Two random variables $X$ and $Y$ are independent if the joint factors - $f_{XY} = f_X f_Y$ (just like the definition we saw for independent events).

*Expectation values* are just weighted sums of random variables. If a random variable $X$ has distribution $f$, then the expectation is

$$EX = \int dx\, x f(x)$$

Where the integral is done over all possible values of $X$. If $X$ is discrete the integral becomes a sum. It doesn't matter what sort of density you have - marginal, conditional, joint, whatever - this definition always holds. You do need to be a bit careful about which density to use, but it's usually obvious. One slightly tricky example is the *law of iterated expectations* - the proof combines the definitions of expectation, and the relationship between conditional, marginal, and joint distributions:

$$\begin{aligned} E[E(X|Y)] &= E\left[\int dx\, x f_{X|Y}(x)\right] \\ &= \int dy \left(\int dx\, x f_{X|Y}(x)\right) f_Y(y) \\ &= \int dy \int dx\, x f_{XY}(x,y) \\ &= \int dx\, x f_X(x) = EX \quad \blacksquare \end{aligned}$$

The *variance* is a measure of spread about the mean

$$\begin{aligned} VX &= E(X-EX)^2 \\ &= EX^2 - (EX)^2 \qquad \text{[Expand the square and simplify]} \end{aligned}$$

From the definition of $EX$ we see that expectation is linear - $E(aX+b) = aEX + b$. We can use linearity of expectation to figure out what happens to the variance when we shift & scale:

$$\begin{aligned} V(aX+b) &= E(aX+b-E(aX+b))^2 \\ &= E(aX-aEX)^2 \\ &= E[a^2(X-EX)^2] \\ &= a^2E(X-EX)^2 = a^2VX \quad \blacksquare \end{aligned}$$

If you want to compute $E[g(X)]$ for some function $g$, you don't have to figure out the distribution of $g(X)$. The *law of the unconcious statistician* (LOTUS) says that $E[g(X)] = \int dx\, g(x) f_X(x)$, which is very nice.

*Covariance* extends the idea of variance to several variables:

$$cov(X,Y) = E[(X-EX)(Y-EY)]$$

Covariance is clearly symmetric, and isn't affected by shifting $X$ or $Y$ by a constant. It's also rv additive, and scales (i.e. $cov(aX+bY, Z) = a\,cov(X,Z) + b\,cov(Y,Z)$). Also from the formula you can see $VX = cov(X,X)$. You can use this to figure out the variance of a sum:

$$\begin{aligned} V(X+Y) &= cov(X+Y, X+Y) \\ &= cov(X,X) + cov(Y,Y) + 2cov(X,Y) \\ &= VX + VY + 2cov(X,Y) \end{aligned}$$

If $X$ and $Y$ are independent then $cov(X,Y) = 0$, so variances of independent rvs sum

## Common distributions

If you're after formulas, just google them. The important thing isn't to memorise the formulas for each distribution, but to understand the story behind each distribution, the assumptions needed for the distribution to be a decent model on data, and the relationships between distributions. Here's some of the most common distributions and links between them (see the next section for proofs):

### Binomial - repeated coin flips

Let $X$ be the number of successes out of $n$ trials, each with probability of success $p$. Then $X \sim Bi(n,p)$ and has distribution

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

Think of $X$ as the outcome of flipping $n$ identical coins, each with probability of success $p$. The binomial coefficient comes in because there are several different orderings for the $x$ successes. An individual coin flip follows a *bernoulli* distribution, with formula

$$P(Y=y) = p^y(1-p)^y \quad \text{[y = 1 if success, else 0]}$$

The story says that $X = \sum_{i=1}^{n} Y_i$. This is handy for working out the mean and variance - since $EY = p$ and $VY = p(1-p)$, the mean and variance for a binomial is $EX = np$, $VX = np(1-p)$.

The key assumptions for a binomial are - events are independent, each event has the same 2 outcomes with the same probabilities.

### Hypergeometric - picking groups from 2 populations without replacement

Suppose you have $N$ deer in a forest, and you paint $n$ of them green. You then take a sample of $m$ deer without replacement, let $X$ be the number of green deer in your sample. Then $X \sim Hyp(n,N,m)$ with distribution

$$P(X=x) = \frac{\binom{n}{x}\binom{N-n}{m-x}}{\binom{N}{m}}$$

Notice that, since you're sampling without replacement, the draws aren't independent. If they were then $X \sim Bi(m, \frac{n}{N})$. The mean is $EX = \frac{mn}{N}$ (just like in the binomial case! The proof involves writing $X = \sum_i I[\text{draw i is green deer}]$ and using linearity of expectation). The variance is more involved since you need to keep track of the $\binom{m}{2}$ covariance terms, but turns out to be $VX = mp(1-p)\frac{N-m}{N-1}$. So it's like the variance for a binomial with a correction factor, and the correction factor goes to one if $N$ is large & $m$ is small relative to $N$. That makes sense, taking a tiny sample without replacement from a huge population is essentially the same as sampling with replacement - since the chances of you picking the same 2 people twice is extremely unlikely. Hypergeometric is a generalisation of binomial to non-independent trials.

### Geometric - number of failures until first success

Imagine you keep flipping the $Bern(p)$ coin until you get a success. Let $X$ be the number of failures until you get the first success, then $X \sim Geom(p)$ and is distributed:

$$P(X=x) = (1-p)^x p$$

It's called a geometric distribution because proofs involve the geometric series $\sum_i a^i = \frac{1}{1-a}$. You can find the mean in 2 ways, either use the definition (involves differentiating the geometric series), or this story proof:

There are 2 cases, either the first result is a success, or it's a failure. If it's a failure then you have 1 failure, and the problem resets so:

$$\begin{aligned} EX &= 0 \times p + (1 + EX) \times (1 - p) \\ &\implies pEX = 1 - p \\ &\implies EX = \frac{1-p}{p} \end{aligned}$$

To find the variance compute $EX(X-1)$ by differentiating the geometric series twice, and use $VX = EX^2 - (EX)^2$ to get $VX = \frac{1-p}{p^2}$.

### Negative binomial - number of failures until $r$ successes

This generalises the geometric to $r$ successes. Imagine you keep flipping the $Bern(p)$ coin until you get $r$ successes. Let $X$ be the number of failures you see before $r$th success. The sequence of outcomes must end with a success, so there are $x + r - 1$ free slots to put the $x$ failures. This means the distribution is

$$P(X = x) = \binom{x + r - 1}{x} p^r (1-p)^x$$

Write this as $X \sim NBi(r, p)$. Think of negative binomials as the sum of $r$ independent geometrics - get the first success (geometric), reset, get the second success (geometric), reset etc etc. Since $X = \sum_i Geom(p)$ we can instantly write down the mean and variance:

$$EX = \frac{r(1-p)}{p}, \qquad VX = \frac{r(1-p)}{p^2}$$

### Poisson - count data

If there is a constant rate of events per unit interval $\lambda$, then the number of events $X$ is poisson. Write $X \sim Pois(\lambda)$, with distribution:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

For a Poisson $EX = VX = \lambda$ and the most probable value is $\lambda$. Notice that for these formulas to work, *lambda* must be dimensionless. $\lambda$ is usually interpreted as a rate - the number of events in 1 hour or in 1 square km for example - but don't get too hung up on units. For example if $\lambda =$ number of events in 1 hour, then the number of events if time $t$ is $Pois(\lambda t)$. If you want to model count data but want the variance to be different than the mean, negative binomial is another option.

### Exponential - waiting times

Let $X$ be the time until the first event, where the total number of events $Y \sim Pois(\lambda t)$ ($Y$ is sometimes called a *Poisson process*). Then by definition $X \leq t \iff Y \geq 1$. This gives a cumulative

$$P(X \leq t) = P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-\lambda t}$$

Since $P(X \leq t) = \int_0^t dx\, f_X(x)$ we can differentiate to get the distribution of $X$ (replacing $t$ with $x$ to be consistent with the other formulas):

$$P(X = x) = \lambda e^{-\lambda x}$$

This is an exponential distribution, $X \sim Ex(\lambda)$. Think of $\lambda$ as a half-life, the time taken for the waiting time to decrease by a factor of $e$. For an exponential, $EX = 1/\lambda$ and $VX = 1/\lambda^2$.

### Normal - most physical measurements

Pretty much all the everyday biological measurements - height, weight, blood pressure etc - are normally distributed. $X \sim N(\mu, \sigma^2)$ has mean $EX = \mu$, variance $VX = \sigma^2$, and distribution

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

Any normal distribution can be transofmred into a *standard normal* - $N(0, 1)$ - by subtracting the mean and dividing by the standard deviation.

### Chi squared - squares of standard normals

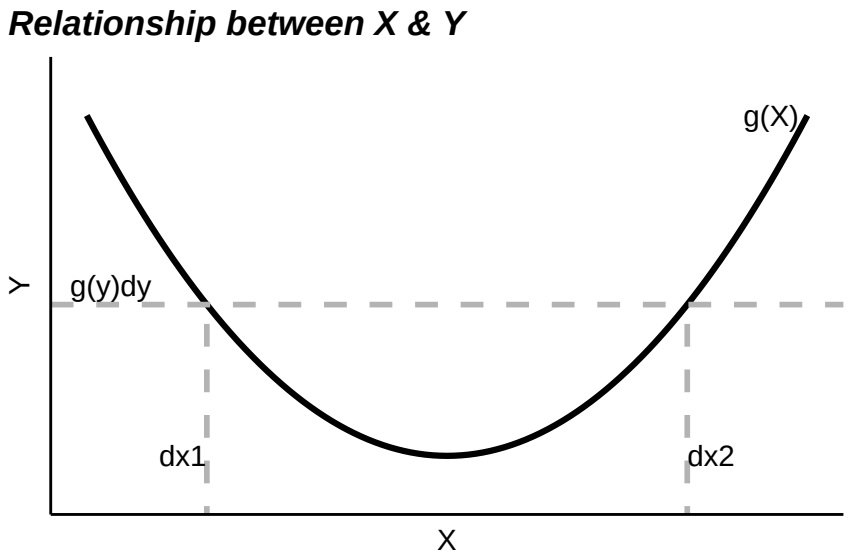
The square of a standard normal is distributed $\chi_1^2$ - ie $X \sim N(0, 1) \implies X^2 \sim \chi_1^2$. Proof is in the next section. The parameter on the bottom is called the *number of degrees of freedom* (see inference notes). The general form for a $\chi_n^2$ distribution is

$$P(X = x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-x/2}$$

It has mean $EX = n$ and variance $VX = 2n$.

## Sums & functions of random variables

Let $Y = g(X)$. What is the distribution of $Y$? As a warm up lets prove the link between chi squared and squares of standard normals:




**Example**

$X \sim N(0,1)$, $Y = g(X) = X^2$. The picture shows a rough schematic of how the distributions are related to each other:

From the picture we can see that Y will get probability contributions at the points $x_1, x_2$, so

$$\begin{aligned} g(y)dy &= f(x_1)dx_1 + f(x_2)dx_2 \\ g(y) &= f(x_1)\left|\frac{dx}{dy}\right|_{x_1} + f(x_2)\left|\frac{dx}{dy}\right|_{x_2} \end{aligned}$$

Where in the last line we put in absolute values because densities can't be negative. Now we just need to find the points $x_1, x_2$ and the derivative by solving $y = x^2$:

$$\begin{aligned} & y = x^2 \\ \implies & x_{1,2} = \pm\sqrt{y} \\ \implies & \frac{dy}{dx} = 2x \\ \implies & \frac{dx}{dy} = \frac{1}{2x} \end{aligned}$$

Putting this into the equation for $g(y)$ gives

$$\begin{aligned} g(y) &= \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}} \\ &= \frac{1}{2^{1/2}\Gamma(1/2)}y^{-1/2}e^{-y/2} = \chi^2_1(y) \end{aligned}$$

Where in the last line we used the fact that $\Gamma(1/2) = \sqrt{\pi}$.

The example shows us what the general case looks like - the probability of the transformed variable $g(y)dy$ is the sum of all contributions from $f$:

$$g(y)dy = \int_S dx\, f(x)$$

Where $S$ is the set of $X$ space which contribute to $Y$. Using this formula (and being careful about limits of integration) you can show that the density for $Z = X + Y$ is given by a *convolution* - $g(z) = \int dx\, f(x, z - x)$ and the density of $Z = XY$ is given by a *Mellin transform* - $g(z) = \int \frac{dx}{x} f(x, \frac{z}{x})$.

In the most general case where the joints have several variables, they transform with a Jacobian scale factor:

$$g(\underline{y}) = \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)} f(\underline{x})$$

If you're working with sums of random variables, it's usually easiest to work with the *moment generating function* defined as

$$M(t) = Ee^{tX} = \int dx\, e^{tx} p(x)$$

To see why this is called a MGF, taylor expand the exponential and compare to the general power series formula:

$$M(t) = \sum_{n=0}^{\infty} \frac{EX^n}{n!} t^n \quad \text{compare with} \quad M(t) = \sum_{n=0}^{\infty} M^{(n)}(0) \frac{t^n}{n!}$$

So the moments are given by

$$EX^n = M^{(n)}(0)$$

Which can be used to find means & variances. MGFs can also be used to find distributions for sums of random variables - it relies on the incredibly important fact that *if $X$ and $Y$ have the same MGF, then they have the same distribution.* This is particularly nice if the random variable is a sum of other independent random variables, as the MGF will factor in some way. The example shows this in more detail:

**Example**

The MGF for $X \sim Pois(\lambda)$ is

$$\begin{aligned} M(t) = Ee^{tX} &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^x)^t}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

The first & second derivatives are $M' = \lambda e^t e^{\lambda(e^t-1)}$, $M'' = (\lambda + \lambda^2) e^t e^{\lambda(e^t-1)}$. Evaluating these at zero gives $EX = \lambda$, $EX^2 = \lambda + \lambda^2$, so the variance is

$$VX = EX^2 - (EX)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

which is exactly the results we shown without proof in the distribution section.

If $Y = X_1 + X_2$ where $X_1 \sim Pois(\lambda)$ and $X_2 \sim Pois(\mu)$ then the MGF of Y is

$$M(t) = Ee^{tY} = Ee^{tX_1+tX_2} = Ee^{tX_1}e^{tX_2} = Ee^{tX_1}Ee^{tX_2} = M_{X_1}(t)M_{X_2}(t) = e^{(\lambda+\mu)(e^t-1)}$$

We're allowed to split up the expectation of the product into the product of expectations because $X_1$ and $X_2$ are independent. That's the MGF of a $Pois(\lambda + \mu)$ random variable, so the sum of 2 Poissons is also Poisson with rate = sum of the rates.

## Multivariate normal - marginal & conditional distributions

The multivariate version of the normal is given by

$$P(X = x) = \frac{1}{\det(2\pi\Sigma)} \exp\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Where $x$ is a k-dimensional vector, $\mu$ is the vector of means, and $\Sigma$ is the covariance matrix with elements $\Sigma_{ij} = cov(x_i, x_j)$. We can partition everything up:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The idea is to partition the term in the exponential into a term involving $X_1$ and $X_2$, and a term just involving $X_1$. Then the exponential factors into a product, and we can interpret that as $f_{X_1X_2} = f_{X_2|X_1}f_{X_1}$. For simplicity we will set $\mu = 0$, we can just shift the results at the end to include it.

First of all lets look at factoring a quadratic polynomial - this is essentially just completing the square. We would like to write

$$x^2 + axy + by^2 = (x + c)^2 + d$$

The first term should have all the $x$ terms, then the second one will have all the leftovers - that is, it'll just include $y$ terms. Expanding and comparing terms gives

$$\begin{aligned} &ay = 2c, \qquad by^2 = c^2 + d \\ &\implies c = \frac{ay}{2}, \quad d = by^2 - c^2 = y^2\left(b - \frac{a^2}{4}\right) \end{aligned}$$

ie

$$x^2 + axy + by^2 = \left(x + \frac{ay}{2}\right)^2 + y^2\left(1 - \frac{a^2}{4}\right)$$

A term involving $x$ & $y$, and a term involving $y$ - just what we want. The $x^T\Sigma^{-1}x$ is a quadratic form, the matrix version of the polynomial we just factored. Working with inverses directly is quite hard so define $V = \Sigma^{-1}$. Then we try to factor like we did above

$$x^T V x = (x_2 + a)^T B(x_2 + a) + c$$

And expand

$$x_1^T V_{11} x_1 + x_1^T V_{12} x_2 + x_2^T V_{21} x_1 + x_2^T V_{22} x_2 = x_2^T B x_2 + x_2^T B a + a^T B x_2 + a^T B a + c$$

Comparing terms gives $B = V_{22}$, $a = V_{22}^{-1} V_{21} x_1$, and $c = x_1^T (V_{11} - V_{12} V_{22}^{-1} V_{21}) x_1$. That means we can write the exponential term as

$$x^T\Sigma^{-1}x = (x_2 + V_{22}^{-1} V_{21} x_1)^T V_{22} (x_2 + V_{22}^{-1} V_{21} x_1) + x_1^T (V_{11} - V_{12} V_{22}^{-1} V_{21}) x_1$$

Putting this into $f_{X_1 X_2}$ gives (ignoring the normalisation factor out front)

$$f_{X_1 X_2} = \exp\left[(x_2 + V_{22}^{-1} V_{21} x_1)^T V_{22} (x_2 + V_{22}^{-1} V_{21} x_1)\right] \exp\left[x_1^T (V_{11} - V_{12} V_{22}^{-1} V_{21}) x_1\right]$$

Comparing to $f_{X_1 X_2} = f_{X_2|X_1} f_{X_1}$ shows that the marginal and conditionals are both multivariate normal. Specifically

$$\begin{aligned} X_2|X_1 &\sim N(\mu_2 - V_{22}^{-1} V_{21} x_1,\, V_{22}^{-1}) \\ X_1 &\sim N(\mu_1,\, (V_{11} - V_{12} V_{22}^{-1} V_{21})^{-1}) \end{aligned}$$

The very last thing to do is rewrite the $V$ terms in terms of $\Sigma$. $\Sigma V = I$ gives 4 equations

$$\begin{aligned} \Sigma_{11} V_{11} + \Sigma_{12} V_{21} &= I & \Sigma_{11} V_{12} + \Sigma_{12} V_{22} &= 0 \\ \Sigma_{21} V_{11} + \Sigma_{22} V_{21} &= 0 & \Sigma_{21} V_{12} + \Sigma_{22} V_{22} &= I \end{aligned}$$

Which you can use to express the $V$s in terms of $\Sigma$s

$$\begin{aligned} X_2|X_1 &\sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1} x_1,\, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \\ X_1 &\sim N(\mu_1,\, \Sigma_{11}) \end{aligned}$$

In the case of 2 variables $x_1, x_2$, the elements of the covariance matrix become $\Sigma_{11} = \sigma_1^2$, $\Sigma_{22} = \sigma_2^2$, $\Sigma_{12} = \Sigma_{21} = cor(x_1, x_2) = \rho\sigma_1\sigma_2$, so the marginal & conditionals become

$$\begin{aligned} X_2|X_1 &\sim N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1} x_1,\, \sigma_2^2(1 - \rho^2)\right) \\ X_1 &\sim N(\mu_1,\, \sigma_1^2) \end{aligned}$$

Which is the result quoted without proof in the slides.