characterize raw data — with descriptive ~~and inferential statistics~~ statistics — mean/var etc.

* linear model: $Y \sim \underline{1} + X$

command in R: lm( )

Week 3 Notes Part 1 Statistical Models

Code ▾

Author: Brendan Gongol

Last update: 24 January, 2025

1 Overview

In this Note we give more formal definition of statistical models as families of distributions or variable-to-variable relationships with unknown parameters that are determined from the observed data; we revisit the concept of independence of random variables, and then introduce linear models. This note concludes with brief semi-formal introduction of the concept of Maximum Likelihood Estimator.

2 Statistical Models

When we conduct experiments (measurements, surveys, etc), we collect raw data. What are these data used for? Clearly, the purpose of any study is to learn something. “Learning” means generalizing and/or finding trends, dependencies, and relations in the data. In contrast, representing the results of a study as 1000 data points is an exact but also the most meaningless way of conveying the findings (or lack thereof). Even in the simplest possible scenario, we want at the very least to characterize our dataset, e.g. by calculating its mean and variance. Moreover, in most cases we are studying and characterizing a sample drawn from a population and want to generalize our findings to that population. By doing so, we already make some (implicit) assumptions. For instance, if we assume that the distribution is indeed normal or reasonably close to normal (at least, some bell-shaped curve), we can use only mean and variance to characterize the distribution (higher moments are not needed), or we may want to use the t-test in order to determine whether two samples (for instance, ages of male and female patients in ALL dataset) are significantly different or not.

— inferential statistics

— descriptive stats

However, if the shape of underlying distribution is very different, e.g. has two different maxima at different locations, all our estimates would need to be done differently as well. We have seen this in Homework 2: we could still compute the mean, but it would grossly misrepresent the actual structure of the data; for instance, estimates made with t-test (which effectively looks at the means) are useless and actually misleading in this case.

When we make an assumption about underlying distribution or dependency, we say that we have a model. Since we are doing statistics and are always concerned with probability distributions of the data themselves as well as the noise, the models we are going to look at are statistical models (those that explicitly include probability distributions). Makes assumptions regarding the distribution of the population. Parametric methods presume the data have a known distribution (normal/binomial/Possion) and rely on parameters ~~to~~(mean/variance) to define the data!

In many cases, at least in this course we are going to be working with parametric models: the models where we specify general functional dependency and leave some parameters to be determined from the actual data (hence parametric). Let us consider examples:

- When we assume that underlying distribution is Gaussian, with unknown mean and variance, we are actually postulating that the functional form of the distribution function is (we use "conditional probability" notation here: the probability of x, given specific values of μ and σ: $f(x|\mu,\sigma)$)

(1)

$$\overbrace{f(x|\mu\sigma^2)} = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Since for each pair of fixed values μ and σ we have a different distribution (albeit they all share the same functional form defined above), the expression above is also often referred to as a family of distributions.

- A dependency of one random variable on another: y=f(x|Θ), where Θ is conventionally used to denote a set of parameters defining the family of dependencies under consideration. Dependencies of this type are the central topic of this week material, but before further defining it and demonstrating how this works for random variables, let us review the concept of statistical independence in general.

3 Statistical Independence

So far we have considered only distributions of random variables, i.e. we were only concerned with generalizations of sample observations to the underlying population (e.g. trying to determine mean and variance of the underlying population defined as (1) above). We have also dipped into cross-sample comparison: are the two samples "the same", in a sense that they come from the same underlying distribution (or, more precisely, from the distribution(s) with the same mean), or not? Good way to explain this.

Another interesting question to ask is: if we are given multiple random variables, X, Y, ... , is there any relationship (and if so, then how significant it is statistically) or they are completely unrelated. For example, those variables could be expression levels of individual genes measured in a sample of patients (thus, in our ALL dataset we already have 10K random variables (genes) measured across all 128 patients to look at), or time to remission in the disease, etc.

If we could find a relationship (dependency) in our data, it might provide, potentially, an important insight. We could have discovered that expression levels of gene g1 tend to be higher when expression levels of another gene g2 are also elevated, so that expression of one gene can be used to predict expression of the other. It could also happen that this correlation is mechanistic and causative in nature, and gene g2 indeed regulates expression of g1, so we could formulate a biological hypothesis for follow-up studies (note that in general correlation does not mean causation!!!). Or we could find that the time in remission depends on expression level of gene g3, so maybe expression level of this gene could serve as a biomarker for prediction of the patient response, and thus we could use it, for instance, to adjust drug dosage according to those expectations. Note that in the latter case it does not even matter if different expression levels of g3 indeed cause, biologically, different responses/times in remission or simply correlate with the latters, as long as we can use g3 to predict the response! However if we could reasonably expect/hypothesize that this dependency is indeed causal, we might be onto something even more interesting: e.g. actual biological explanation of the variation in the remission times (subject for further experimental validation, of course).

The random variables X and Y are considered unrelated, or independent, when the probability to observe any value of Y does not depend on the realization of X: in other words, whatever value we have measured for X it provides us with no additional information about what values of Y we should expect (the distribution of Y stays the same). Using notations from probability theory:

(2)

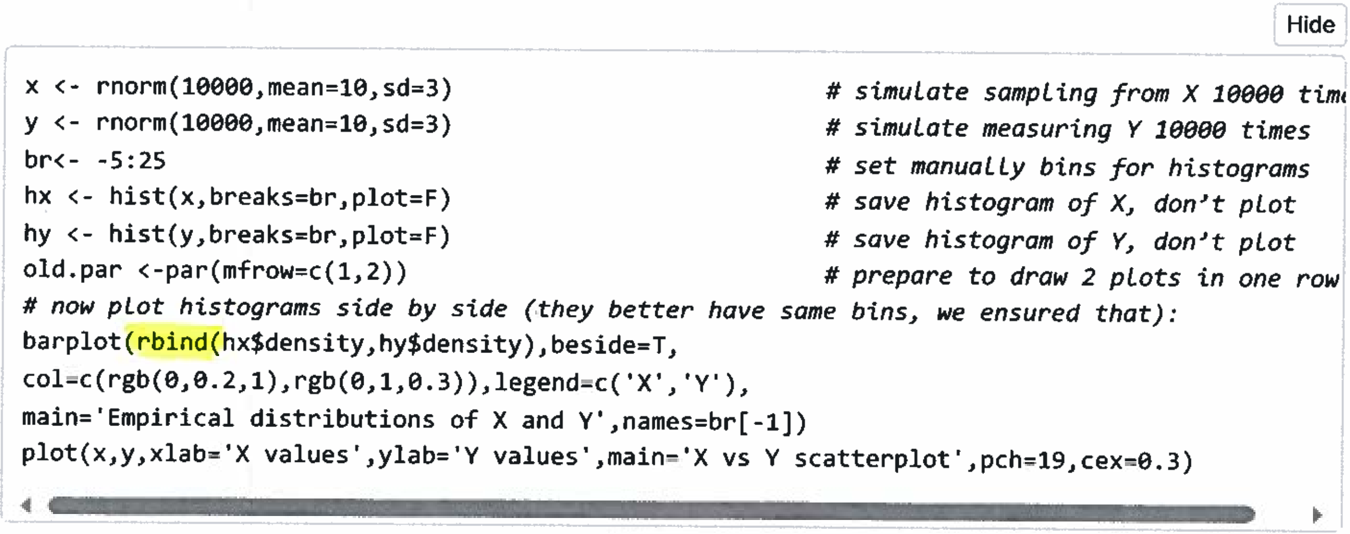
$$P(Y|X) = P(Y)$$

(probability distribution of Y conditioned on X is the same as the marginal probability distribution of Y: in other words X is irrelevant for predicting the values of Y).

Note that when talking about such conditional probabilities, we imply that realizations of the variables are matched in some (meaningful) way. For instance, if X and Y are expression levels of two genes, the particular realization (x_i, y_i) is measured in the same patient i: measuring X in 100 patients, then measuring Y in a different 100 patients and asking if the two are "dependent" in any way is quite nonsensical.

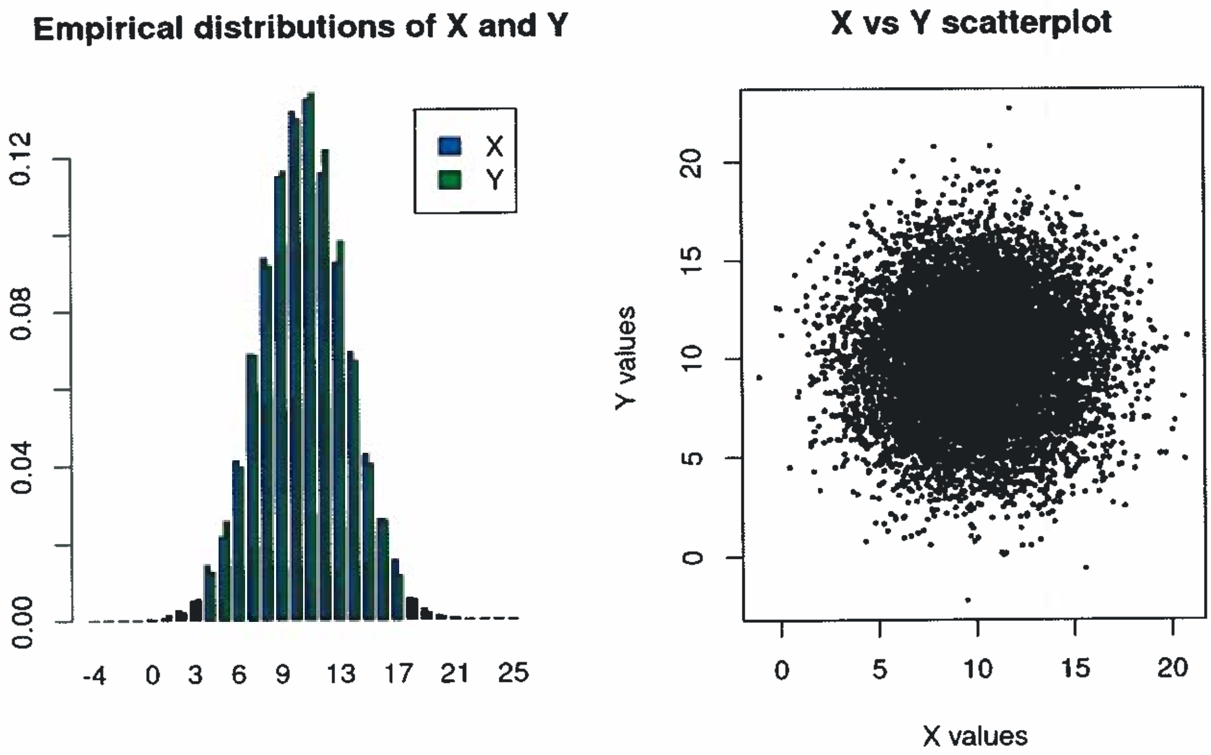
We should all be able to agree by now that R is very convenient for running data simulations (among other things). Let us stop for a moment and run some trivial demonstrations in R in order to inform our intuition and improve understanding.

Let us assume that we are measuring two Gaussian random variables, X and Y (could be expression levels of two different genes or anything else; but we will always assume throughout this discussion that they are "matched" as explained in the discussion above). We measure them 10,000 times (e.g. across a large cohort of 10,000 patients) and end up with a sample. Let's generate such (random) sample:



```
x <- rnorm(10000,mean=10,sd=3)                     # simulate sampling from X 10000 tim
y <- rnorm(10000,mean=10,sd=3)                     # simulate measuring Y 10000 times
br<- -5:25                                         # set manually bins for histograms
hx <- hist(x,breaks=br,plot=F)                     # save histogram of X, don't plot
hy <- hist(y,breaks=br,plot=F)                     # save histogram of Y, don't plot
old.par <-par(mfrow=c(1,2))                        # prepare to draw 2 plots in one row
# now plot histograms side by side (they better have same bins, we ensured that):
barplot(rbind(hx$density,hy$density),beside=T,
col=c(rgb(0,0.2,1),rgb(0,1,0.3)),legend=c('X','Y'),
main='Empirical distributions of X and Y',names=br[-1])
plot(x,y,xlab='X values',ylab='Y values',main='X vs Y scatterplot',pch=19,cex=0.3)
```

what do breaks do in histograms?
why do we set our par to old.par?



change mean+ sd when sampling using p(norm)

Hide

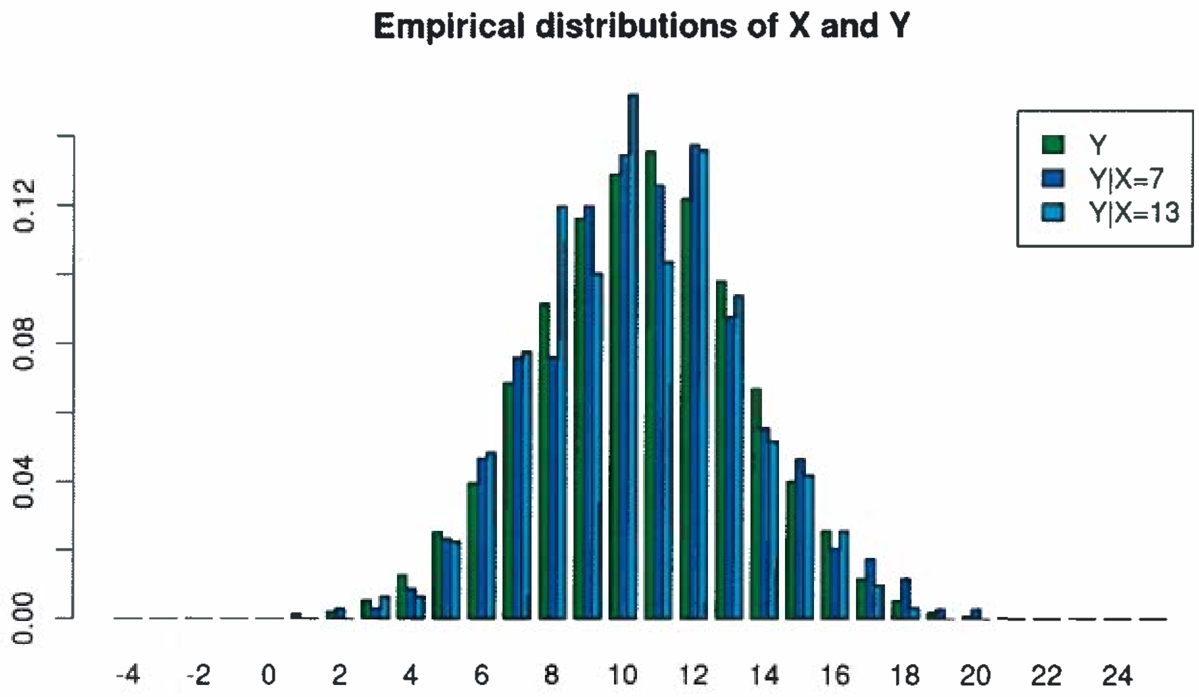
```
par(old.par)                                  # restore graphical attributes to pr
```

The resulting plot is shown below. We can see that empirical distributions of both X and Y are very close. This is not very surprising, since that's the way we sampled them in the first place (from the same distribution). We could sample X and Y from two different normal distributions as well as from two completely different distributions (you can try that), but this is not the point here, we are setting up the stage for something different, and the sampling we did will serve the purpose. Some small differences between the green and blue empirical distributions in the histogram below remain because we are dealing with finite sample size, so there is some sampling fluctuation. The scatterplot clearly shows that there is no relation between the two variables (since the observations are "matched", each point in our plot represents an observation, e.g. a patient, the x- and y-coordinates of the point being expression levels of genes X,Y in that patient). Obviously, by looking at this plot one cannot say something like "for patients with lower x, the values of y also tend to be lower".

In order to further illustrate independence of Y and X (and better understand eq. (2) above), let us follow the prescription given by Eq. (2) to the letter. Namely, let us select two different values (or actually narrow ranges since we are dealing with continuous distributions) of X and see how the distributions of Y conditioned on X will behave:

Hide

```
range.7 <- x > 6.8 & x < 7.2 # select instances where X is close to 7
range.13 <- x > 12.8 & x < 13.2 # instances where X is close to 13
# take only Y where X~7 and calculate their empirical distribution:
hy.7 <- hist(y[range.7],breaks=br,plot=F)
# select only Y where X~13 and calculate the empirical distribution:
hy.13 <- hist(y[range.13],breaks=br,plot=F)
# plot overall distribution of Y and conditional distributions,
# P(Y|X=7) and P(Y|X=13) side by side
barplot(rbind(hy$density,hy.7$density,hy.13$density),
beside=T,col=c(rgb(0,1,0.3),rgb(0,0.28,1),rgb(0,0.8,1)),
legend=c('Y','Y|X=7','Y|X=13'),
main='Empirical distributions of X and Y',names=br[-1])
```



Hide

```
t.test(y[range.7],y[range.13])
```

```
##
##  Welch Two Sample t-test
##
## data:  y[range.7] and y[range.13]
## t = 1.034, df = 648.16, p-value = 0.3015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2142288  0.6907970
## sample estimates:
## mean of x mean of y
## 10.104927  9.866642
```

Hide

```
t.test(y[range.7],y)
```

```
##
##  Welch Two Sample t-test
##
## data:  y[range.7] and y
## t = 0.55441, df = 363.43, p-value = 0.5796
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2355453  0.4205024
## sample estimates:
## mean of x mean of y
##  10.10493  10.01245
```

In the code fragment above, out of all 10,000 measurements we took for the pair X, Y of the random variables, we first identify a) cases where measured values of X are (approximately) equal to 7; b) cases where X is (approximately) equal to 13. Then we select only values of Y observed in each of those subsets of all measurements (e.g. patients) and generate empirical distributions of those subsets of data. These empirical distributions represent conditional probability distributions for observation of different values of Y in the slices of data defined by $X = 7$ or $X = 13$, respectively, i.e. $P(Y|X = 7)$ and $P(Y|X = 13)$. Next we plot the overall distribution of Y as well as those two conditional distributions side by side. The resulting plot is shown below:

The above plot demonstrates that, of course, the overall distribution of variable Y as well as distributions of the subsets of Y defined by conditions $X = 7$ or $X = 13$ are all the same (the differences observed in the plot are not impressive and likely due to sampling fluctuations: the counts of Y points falling within the slices we defined are relatively low, so random sampling fluctuations are large). There are much better and rigorous ways to compare the distributions,

the t-tests

but for now we are going to stick with just visual inspection and t-test. The last two commands in the code fragment above compare sub-samples of Y values restricted to the two slices to each other, as well as sub-sample of Y values in one slice to the whole sample of Y we have. The t-test results (large p-values, no significant difference found), while strictly speaking telling us only that the samples likely came from the distribution(s) with the same mean – t-test does not compare the shapes of the distributions!! – still strengthen our visual observation that the distribution of Y is always the same, regardless of X. In other words, we observed in our simulation that $P(Y|X = 7) = P(Y|X = 13) = P(Y)$, at least qualitatively.

2 individual rnorm() assignments.

Note that we of course expected this result because we sampled X, Y independently, but it is still an instructive example to see how actual samples behave and what simple (and naïve) methods we can use to get some insight into the data.

Let us now modify how we simulate our samples:

Hide

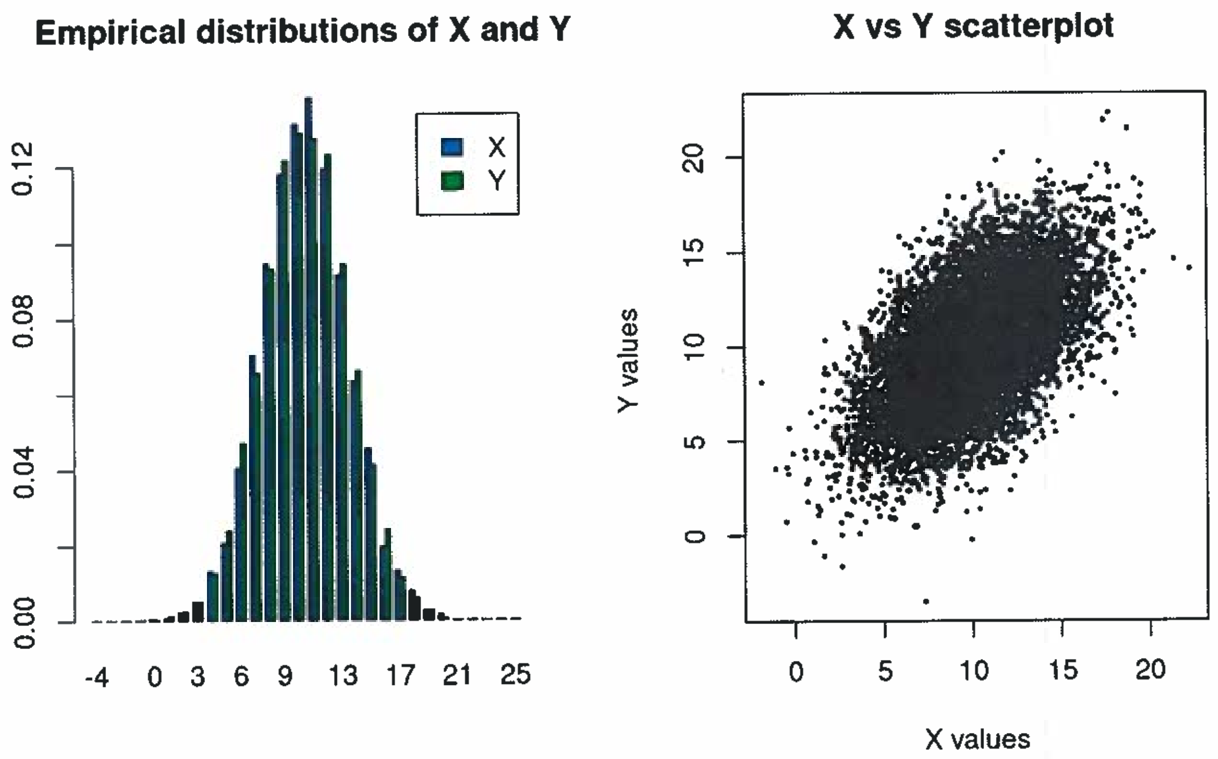
```
x <- rnorm(10000,mean=10,sd=sqrt(5))
y <- x # initialize y with x
x <- x + rnorm(10000,sd=2)
y <- y + rnorm(10000,sd=2)
```

In this code fragment, we first sample 10,000 x values from a normal distribution with the same mean as before and with standard deviation `sqrt(5)` (we will see why in a second). In this numerical experiment we want to simulate the case when there is a dependency between Y and X, so we initialize the sample for Y with values we sampled from X. In this trivial case, whatever value we measure for X, the value for Y is exactly the same, with certainty. That's not very interesting for us, statisticians. So we make it more interesting and add random noise ε to each of the variables X and Y. Now we have some underlying exact dependency of Y on X partially washed away by the added random process (which is completely independent of X and Y). This process can represent measurement error in our experiment (if X and Y represent expression levels of two genes, then remember that our measurements of expression levels, e.g. with microarrays, are subject to noise). Or this additional random process can represent combined action of unknown yet deterministic factors. For instance, expression level Y can be determined precisely by the combination of expression level of X, expression level of Z, patient age, and the temperature in the room. But when we look at Y and X alone, those other, unaccounted factors will appear as random shifts preventing us from predicting Y precisely from X alone, so the data look like they have additional "random noise".

The choice of the standard deviations for the two random processes is dictated by the fact that the sum of two independent normally distributed random variables X+ε with variances σ_x^2 and σ_ϵ^2, respectively, is also a normally distributed random process with variance $\sigma_x^2 + \sigma_\epsilon^2$, so after our variables X and Y are augmented with the noise in our last example, they are going to have exactly the same standard deviation, 3, as in the first example we considered in this section. From this point on we can proceed using the same code as in the previous example:

Hide

```
br<- -5:25 # set manually bins for histograms
hx <- hist(x,breaks=br,plot=F) # save histogram of X, don't plot
hy <- hist(y,breaks=br,plot=F) # save histogram of Y, don't plot
old.par <-par(mfrow=c(1,2)) # prepare to draw 2 plots in one row
# now plot histograms side by side (they better have same bins, we ensured that):
barplot(rbind(hx$density,hy$density),beside=T,
col=c(rgb(0,0.2,1),rgb(0,1,0.3)),legend=c('X','Y'),
main='Empirical distributions of X and Y',names=br[-1])
plot(x,y,xlab='X values',ylab='Y values',main='X vs Y scatterplot',pch=19,cex=0.3)
```



Hide

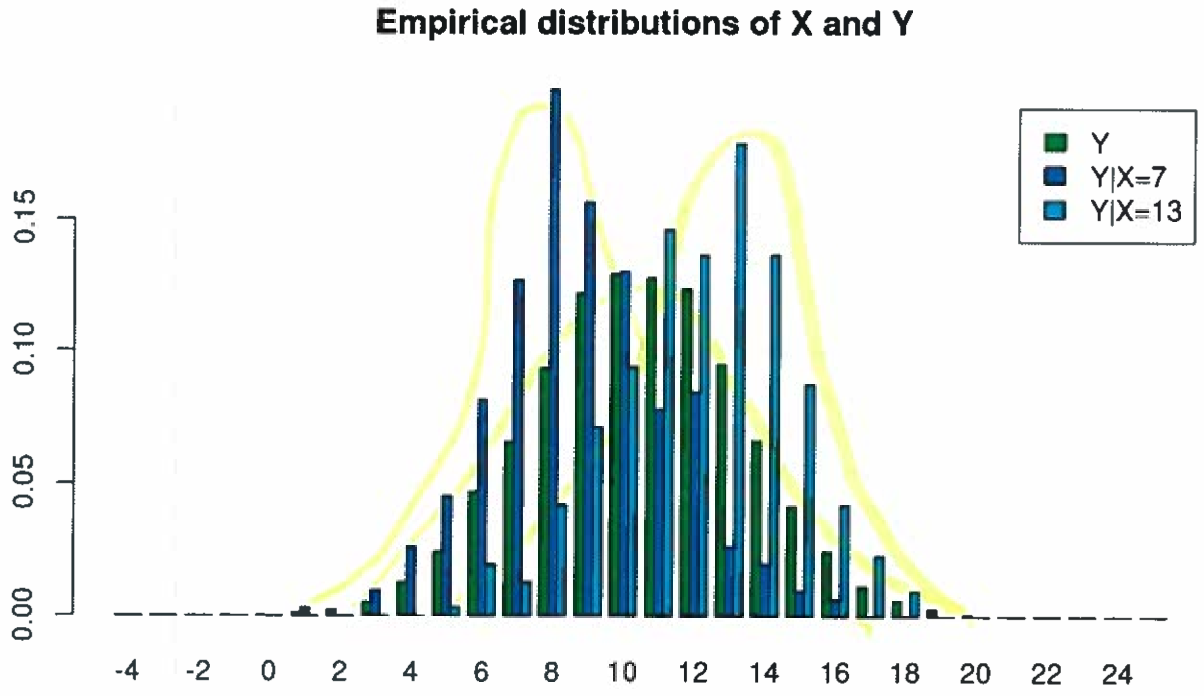
```
par(old.par) # restore graphical attributes to previous values
```

The resulting figure shown above confirms (at least visually for now) that the marginal distributions of X and Y (i.e. taken alone, as in $P(X)$ and $P(Y)$, without any regard for additional, jointly measured, variables) are the same, and do not differ from the distributions we have seen in the first example. Looking at X or Y alone (e.g. measuring only one of the genes), we thus see a “usual” normally distributed random variable with no interesting properties. However, the scatterplot reveals that there is a correlation between the values: we can clearly see that in patients with higher expression values of X, the expression level of Y also tends to be higher (again, we just built this toy dataset this way, that was the whole purpose of the simulation).

Let us now look at conditional probability distributions, again using the same code as before:

Hide

```
range.7 <- x > 6.8 & x < 7.2
range.13 <- x > 12.8 & x < 13.2
hy.7 <- hist(y[range.7],breaks=br,plot=F)
hy.13 <- hist(y[range.13],breaks=br,plot=F)
barplot(rbind(hy$density,hy.7$density,hy.13$density),
beside=T,col=c(rgb(0,1,0.3),rgb(0,0.28,1),rgb(0,0.8,1)),
legend=c('Y','Y|X=7','Y|X=13'),
main='Empirical distributions of X and Y',names=br[-1])
```



The resulting empirical distributions are quite different as the figure below shows. Indeed, the two blue distributions are obtained by taking two different narrow vertical cuts in the scatterplot above, at $x = 7$ and $x = 13$, and considering distributions of values of Y located within these cuts. We see that now there is a clear conditioning of the distribution of variable Y on the value observed for variable X: while at the different fixed values of X the variable Y is still random and has a whole probability distribution to itself, those distributions are very different from each other and from the marginal distribution $P(Y)$. This is precisely the meaning of equation (2): when $P(Y|X)$ differs from $P(Y)$, the variables X, Y are not independent, and we just illustrated that with an example.

4 Linear Models

There are many ways to look for dependencies and relationships in the data (and we used couple of most trivial ones above: visual inspection of distributions and conditional distributions, scatterplots, t-test on subsamples of one variable conditioned on the other), and the choice of the "correct" method is not necessarily a trivial task, as it is often the case with statistics. Each method has its own underlying assumptions, statistical power, strengths and weaknesses; some methods can be also intrinsically related but historically developed as "different" ones. In practice, it is often useful to try different methods and visualizations and summarize the results. But for now we will concentrate on a single method of linear models. Despite its deceptively simple formulation, it is a very useful and versatile approach (and not even necessarily "linear" at all!) that allows for detecting and quantifying dependencies.

In line with what we have started our discussion from, linear models are parametric statistical models, which implies that we are looking for functional relation of predefined form with (unknown) parameter(s). Consider two random variables X and Y (you can think of variables we have simulated in the previous section) and let us assume that we are given matched realizations of size N (samples) of both variables (i.e. we have expression levels of two genes measured in N patients): x_i, y_i where $i = 1...N$. We set up the simplest linear model by assuming the dependency:

(3)

$$\gamma_1 = \beta_o + \beta_1 x_1 + \varepsilon_1$$

at every point i (so strictly speaking this is a system of equations at i=1…N, with some fixed but yet unknown β which we want to determine from the data).

Eq. (3) is a parametric model with unknown parameters β_0 (intercept) and β_1 (slope). The term εi is referred to as noise, error, or residual. The variable X is called predictor variable (also known as input, independent variable, feature), and Y is the response variable (also known as outcome, dependent variable).

A linear model (3) is understood as follows: we are looking for a single line that goes "as close as possible" to all the data points at the same time. In other words, we are looking for the slope and intercept such that measured values x_i predict values y_i most accurately. Imagine an idealized case $y_i = \beta_0 + \beta_1 x_i$ (without the error term). In this case each value yi can be predicted precisely if we already measured xi . Clearly this is the most accurate prediction ever possible (and there is no need for statistical approaches since there is no variability/randomness left!). In real life, there is always noise and/or effect of some other factors which are not accounted for (either in the whole experiment or in the particular model we are adopting), and we are trying to find the parameters that provide best possible prediction of the response Y from the variable X included in the model. In other words, we want to simultaneously minimize all the residuals εi in some sense: it would not do much good if we had precise predictions for some values xi, but got huge errors at the others. In principle, there are many different ways to define how the residuals have to be minimized, ε depending on the specifics of the problem. For instance, one might look for the slope and intercept such that the sum of absolute values of all the residuals reaches its minimum possible value. But by far the most common approach is to minimize the sum of squares of the errors: $\sum_i \epsilon_i^2$ (we will see in a second what the meaning of this requirement is). You might recognize the least squares requirement in this last condition.

To further understand the structure and nature of the error term, remember that observed values are just a particular random sample from the distribution of the random variable. If we want to be statistically minded, we have to interpret our linear model in terms of random variables and use probabilistic language of samples drawn according to the distribution functions of those variables. The second example from the previous section (in which we simulated dependent variables) becomes useful here: in statistical sense, we expect that a specific value x of random variable X cannot predict an exact value of y, but instead defines a distribution over possible values of y that we can measure given x, i.e. $P(Y|X = x)$. In this sense, the narrower the latter distribution, the more accurately we can predict y given value of x. Hence, in our case the "remaining" distribution of the output Y after $X = x$ is measured, $P(Y|X = x)$, is actually the distribution of noise. The noise is the random process in itself, and the observed residual values εi in the expression (3) are a sample from that process.

The distribution $P(Y|X = x)$ (the random process from with the noise terms ε are drawn) can be also referred to as unexplained variation. If we revisit the example from the previous section, we can recall that even though we explicitly simulated the (imperfect) dependence of Y on X, the marginal distribution $P(Y)$ looked just like a usual normal distribution, indistinguishable from a distribution from a truly independent variable that we built in the first example. This means that if someone tried to measure only variable Y, they would observe some variation in the measured values and discover the distribution $P(Y)$ but would be unable to do much more than that. The observed variation in Y would be the end result of their study (if we measure only one variable, all we can do is to characterize its distribution, more or less). However, when a better informed experimenter measures both variables X and Y and discovers that there is some dependency between the two, she might set up a linear model as in (3). The part of the model without the error term defines explained variation of the response variable Y (i.e. "the expected value of Y changes from patient to patient because the measured value of X changed" – completely explained part). The error terms is what governs the remaining uncertainty: measured value x predicts where the corresponding measured value $\bar{y} = \beta_o + \beta_1 x$ should be found (i.e. it's the expectation), but the actually observed value y is only somewhere around this value (determined by the width of the error distribution $\varepsilon = P(Y|X = x))$ rather than exactly equal to it, so there is this additional variation left unexplained by our model, which contributes to the total variation of Y.

In the textbooks and literature, models are often denoted using alternative, statistics-specific notation, rather than general mathematical notation of Equation (3). A model, in which response variable Y linearly depends on X could be denoted simply as Y~X (do you see where the 'tilde' in R formulas comes from?). Note that

- In this notation we (more appropriately) define a conceptual relation between random variables; in the functional form (3) we have to define a model in terms of particular realization
- This notation does imply that we want (3) to hold for any specific realization
- We usually do not explicitly write down coefficients in this notation (although sometimes people do), the formula above already implies all the coefficients written down in (3); the more common variation is to use $Y \sim 1 + X$ where "1" explicitly specifies that we are looking for a model where intercept is allowed to be non-zero. Note that we write '1' , not actual parameter β0 (just like we simply write X omitting the slope coefficient).

To conclude this section, we have to mention a very important (and confusing) difference between interpretations of the word "linear" in statistical linear models and in the "rest of the world". Normally, in math we call a function $f(x) = ax + b$ linear, and, e.g., functions $g(x) = ax2 + bx + c$ and $h(x) = logx$ - non-linear. In other words, the concept of "linearity" refers to the function argument, while coefficients are some fixed constants. In statistics, we can fit the model (i.e. determine intercept and slope) only when we have measured data, i.e. a sample of values x_I, yi drawn from the underlying distributions. But after the data are measured, all those values are fixed constants. It is the coefficients (intercept, slope) that we need to solve the equations (3) for, so a "linear model" in statistics is actually linear with respect to its parameters, not random variables. The latters can be subjected to non-linear transformations and the model will remain "linear" anyway (so linear models in statistics represent a really broad and powerful class of models). Consider, for instance, the following model:

$$Y \sim 1 + X + X2$$

(in statistical notation introduced above), or in functional notation:

$$y_i = \beta_o + \beta_1 x_1 + \beta_2 x_i^2 + \varepsilon$$

In this equation, x_i^2 is just another set of values (derived from xi but this does not matter) and we need to fit coefficients β against all those observed values of y_i, x_i and x_i^2 - it is still a linear model and one needs to solve a system of linear equations in order to find β!! We could similarly argue in terms of random variables: X^2 is simply another random variable, we could call it Z for all we know, and the fact that Z and X are dependent is not going to make the linear model invalid, it can handle that.

5 Maximum Likelihood Estimator (MLE)

How do we fit linear models to the data (or for that matter define what is a good vs bad fit in the first place)? There are different methods and procedures suitable for different specific situations, but here we consider only one, which is very common and also important as it covers most of not-too-exotic cases.

The Maximum Likelihood approach works as follows. Suppose we have some measured data D (which includes both independent and dependent variables) and a parameterized model $M(\Theta)$ that we want to fit to the data (i.e. we want to find optimal parameters of the model Θ, such that independent variables predict the outcomes in the best possible way). Statistical models are formulated in probabilistic terms, so all we can do is compute the probability, under the model M, to observe data given any fixed set of parameter values, $P_M(D|\Theta)$. If specific set of parameter values Θ is such that it would be extremely unlikely to observe the data the way they are, it is natural to believe that the data lend little evidence in favor of such parameters. In contrast, if at some values of the parameters observation of the data we actually obtained becomes very likely, then these are the parameter values that are supported by the observed data. Hence the optimal parameters of the model are those that maximize the likelihood of observed data $P_M(D|\Theta)$. The formal basis for this sort of reasoning is the Bayes theorem. Consider joint distribution of the data and the model parameters, $P(D, \Theta)$. It can be written in two equivalent ways, via the conditional probability distribution of data given parameters or conditional probability distribution of parameters given the data:

$$P(\Theta, D) = P(D|\Theta)P(\Theta) = P(\Theta|D)P(D)$$

Hence the probability distribution of parameters given the observed data is

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)} \propto P(D|\Theta)$$

i.e. the most likely (highest probability) set of parameters given the data are those, for which the probability of observed data reaches its maximum. The probability distribution of the parameter values on the left hand side of the equation is also known as likelihood function, or simply likelihood.

Let us illustrate this using linear model as an example. The model is y=ax+b+ε. Let's assume that we have a measured sample (data) xi, yi, where i=1...N. Under the assumption that the model is correct, we need to come up with probability to observe the data given the parameters a,b:

$$L(\Theta|x_1, y_1; x_2, y_2; \ldots; x_N, y_N) = P(x_1, y_1; x_2, y_2; \ldots; x_N, y_N|\Theta)$$

But our repetitive measurements (samplings) of random variables X,Y are independent: values xi, yi observed in the particular instance, e.g. particular patient, might depend on each other – at least that's what we believe when we try fitting a linear model to the data, but those values are absolutely independent of the values xk, yk observed in a different patient! So the joint probability is simply a product of probabilities at each data point:

$$L(x_1, y_1; x_2, y_2; \ldots; x_N, y_N|\Theta) = P(x_1, y_1\Theta) \cdot P(x_2, y_2\Theta) \cdot \ldots \cdot P(x_N, y_N\Theta)$$

In other words the probability to observe all the data (given the parameters) is just the product of the probabilities to observe each data point (given the same parameters). Thus all we need is to compute the probability of a pair of values x,y if our linear model is correct. Finishing this part of the derivation is left for the homework, but we will give here a starting point. If we are given the value x, then according to our linear model there is an exact, predicted value y'=ax+b defined by x. The difference between actually observed value y and the predicted one, y-y', is distributed according to the error probability f(ε)=P(Y-y'|X=x). This error probability distribution tells us how (un)likely is any particular deviation of observed value y from the value y' predicted from x (at given model parameters). In other words, it is nothing but the probability to observe the datapoint (x,y): we measured x, and thus expect y' (according to the model), but observed y instead, thus the probability to observe (x,y) under our linear model is just the probability to have the error (discrepancy) equal to y-y'. This sounds like a circular argument (since we do not know what the probability distribution of the error is), but it is not. Assuming that the error distribution is centered at 0 (which is natural:

we expect errors to occur in either direction equally) and adopting a specific shape of that distribution can be enough! Here we make the simplest and widely used assumption that the remaining unexplained noise is Gaussian, i.e.

$$P(x, y|\Theta) = f(\epsilon) \sim \frac{1}{\sigma_\epsilon} exp(-\frac{\epsilon^2}{2\sigma^2})$$

where σε is the magnitude (spread) of the noise. Note that we do not need to specify the standard deviation of the noise as a number, nor should we assume anything about it, just the fact that the noise is Gaussian will let us solve our model fitting problem completely! Using the error distribution to quantify how unlikely are deviations of observed y from the value y' predicted by observed x, and taking the logarithm of the likelihood function (4), we arrive to log-likelihood

$$logL = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

We need to find parameters of the model a, b that maximize the likelihood L, and thus also maximize its logarithm (i.e. we are asking at what parameters a, b the observed data become most likely under our model). Differentiating log L with respect to parameters a, b and solving for a, b at which $logL$ reaches maximum (i.e. solving $\varphi logL/\varphi a = 0$ and $\varphi logL/\varphi b = 0$), we obtain optimal (in MLE sense) parameter estimates for the observed dataset.

Note that in the Eq. (5) the first term is constant, and the second term is exactly the sum of squared residuals (residual=observed – predicted, so it's another name for the unexplained variance). Since this term has a negative sign, $logL$ is at its largest when the sum of squared residuals is at its smallest – but the latter is exactly the requirement of the least squares method as we learned it at school. So we now see that linear model parameters optimized using least squares criterion are actually correct MLE estimates, under the assumption of the normality of the noise. We can also see that the assumption of normality was sufficient to find the optimal parameters without specifying what the variance of the noise actually is; hence the circularity we worried about earlier is broken. In practice we get the data, perform least squares estimate (i.e. just assume normality of the noise), then using the obtained optimal model parameters a,b we calculate predicted values y' for each measured value of x, compute the actual residuals $r_i = y_i - y'_I$ and then from these residuals we determine their variance, i.e. the variance of the remaining, unexplained noise. Statistics of the residuals is in fact an important criterion of the goodness of fit of our linear model. We will get to this in much greater detail soon enough!