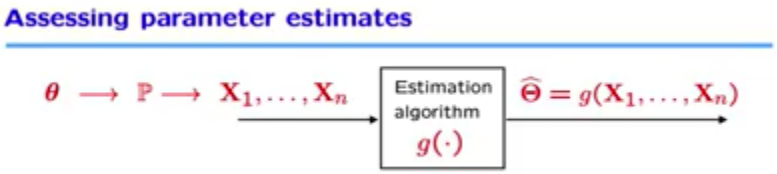
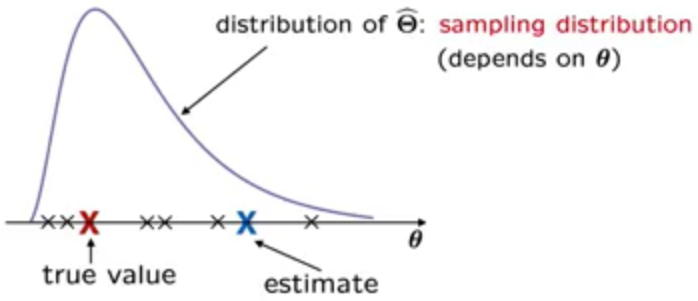
Bootstrap Sampling

Comparing different predictors is one part of performance assessment. If the prediction is all that we care about, then this is enough. On the other hand, if we are interested in estimating a model and learning some true thetas, we are also interested in assessing the accuracy of those thetas, we have seen how to do that with formulas using standard errors.

But what if we do not have some useful formulas?
For some methods, these are difficult to have. We would like to have some ways of estimating standard errors by just using the theta as opposed to formulas and this is done by a method that's called the **Bootstrap**. Let us go through this method.



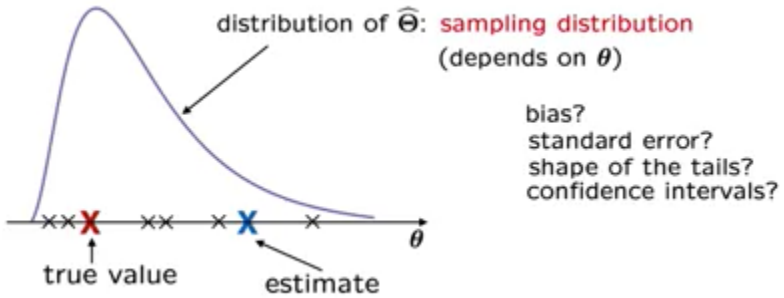
There is some phenomenon back there from which the theta is drawn and that phenomenon involves certain parameters, and our objective is to estimate those parameters. We have a data set. Some data that have been generated by this phenomenon and the distribution of those data, that distribution is affected by those parameters. We take this data and then we run an estimation algorithm. Now, notice this is an estimation algorithm that estimates parameters. This is not a predictor anymore. It takes all the data and on the basis of those data, it processes them and produces some estimates of the parameters. Now, the estimate of the parameters that are being produced is going to be random. There is a true value for the parameter.



But when we run the estimation once, we get some estimates. If we run it with a different data set, we get another estimate. If we run it with a different data set, we get another estimate, and so on. But all these data sets are drawn from the same distribution with different data sets with different parameters. So, if we take all those parameters and think about building a histogram out of them, then we get the probability distribution for a theta hat.

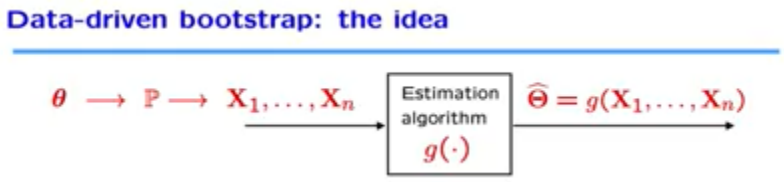
This is the same story we were discussing sometime earlier in this session. The estimates depend on noise, depend on random sampling, so they are random variables, so they are described by some probability distribution, and that probability distribution is of course affected by the method that generates the data, so it's affected by the true theta. This is called the sampling distribution.

Under nice assumptions in textbook-style linear regression, the sampling distribution is going to be a normal distribution, but in general, it might be something else when we deal with more complicated ways of estimating the parameters. And we would like to know for that sampling distribution some of its properties.



- Does it have a bias?
- Is it centered around the true value?
- That is, are the estimates on the average the true values or not?
- What is the standard error of this distribution?
- What is the width of that distribution?
- That width is roughly proportional to the standard error.
- What's the shape of the tails?
- What are the probabilities that we get values that are very incorrect?
- How can we build confidence intervals?

Well, to build confidence intervals, we need to have estimates of the standard error. So, the task that we have in front of us is in such unstructured situations where we're not making any particular mathematical assumptions, how can we assess the standard error of a method?



Well, if we could run this estimation procedure a 100 different times using different data sets, we could build a histogram of the different theta hats, we would get one from each data set, and building that histogram, we would obtain this sampling distribution and once we have that sampling distribution, we can determine the standard error. Unfortunately, we only have a single data set. We do not have 100 of them.

Can we somehow do some magic, and out of a single data set, create 100 different data sets?

It looks impossible. But actually, there is a clever way of doing it. In order to build 100 different data sets, we would want 100 different data sets that are drawn from the same phenomenon. So, drawing from the same phenomenon, they are different random data sets. So, they correspond to different, they lead to different estimates. We do not have this kind of data set and we do not know the structure of the underlying

phenomenon. On the other hand, the data themselves are representative of that underlying phenomenon. So, instead of sampling 100 data sets out of this phenomenon, we can just sample data out of the data that we already have since they are representative of the underlying phenomenon and this is a very powerful idea that goes with the name of (re)sampling. We generate new data sets, and new samples by looking at the available original data sets and picking data records out of them at random.

We have an original data set that consists of n data records, and we run our estimation and it produces an estimate. What the bootstrap method does is, it creates a new data set that consists again of n data records. This n here is the same as that n.

So, it is a data set of the same size and how is that data set produced?
It is produced by looking at the original data set and picking data records at random. Now, if we pick them at random and we ask that they all must be different, then this data set would end up being identical to that data set.

However, we do the **sampling with replacement**. That is, we take at random one of those data records and put it here, then put that data record back there, choose again one at random, and put it here and repeat that n times, and thus create a data set consisting of n records. So, it is possible and it will usually be the case that some of the original data records will be selected multiple times and some of them will not be selected. So, the new data set is different from the original one, but in some ways, it is also similar. Now, what we do now is run our estimation procedure on this new data set and produce a new estimate. This is like running the same estimation procedure but on a different data set, which is also, however, generated in the background by the same underlying phenomenon, and then we repeat this process a number of times. This **m** here could be something like a 100, for example. We run it a number of times and each time that we run it, we get a different estimate. In all of these cases, all of these estimates are trying to estimate the same quantity and the same quantity as this one. However, they will all be different because they are different data sets, even though they are drawn from the same phenomenon and then we go and see how much variability we have between these different estimates. So, these estimates are all over

the place. We can form a histogram out of these and this is an approximation of the sampling distribution that we were discussing before.

$$\widehat{\Theta}_{\text{ave}} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\Theta}^i$$

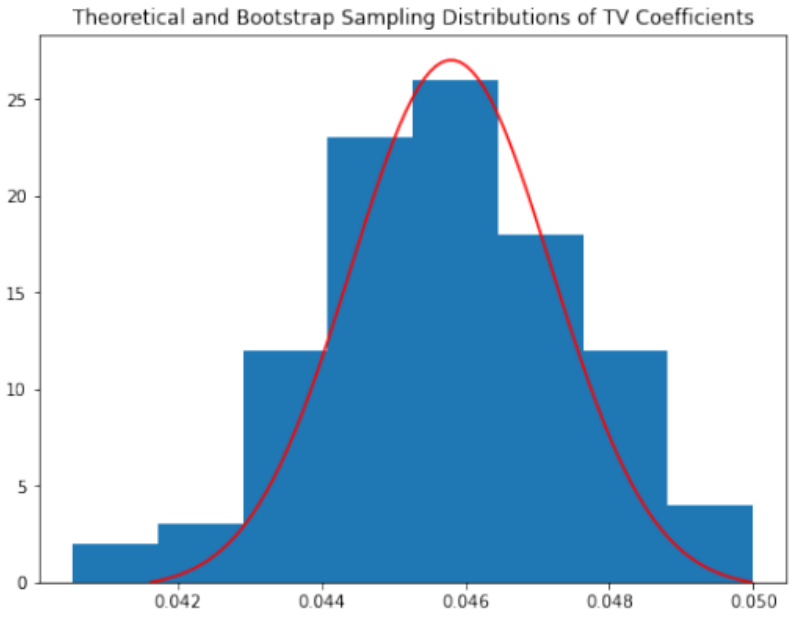
$$\widehat{\text{Var}}(\widehat{\Theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{\Theta}^i - \widehat{\Theta}_{\text{ave}}\right)^2$$

Let us look at the average of these estimates that we produced and then look at the squared distance of a typical estimate from the average. So, this is essentially the variance of this distribution that we have here and when we take the square root of that variance, we have the standard deviation of this distribution and that standard deviation is an estimate of the standard error of this particular estimation procedure. So, we have a way of estimating standard errors. We can also go and see what happens to the tails of that distribution.

Let us take the 2.5% tail left and the 2.5% tail right and we take the width of this interval and make that be the width of a confidence interval. So, this way we can estimate standard errors and we can create confidence intervals just using the data and using no formulas and no assumptions about normality or anything else. Amazingly, this method performs pretty well.

So, let us go back to our marketing example. For the television coefficients, we create many, many different data sets by resampling from the original data that we have, and this is a histogram of what the different estimates happen to be using this variety of data sets.

What does it look like?
Well, it's not too far from the normal distribution that would be predicted by the formulas if we believed that the noises were random with equal variances and all that. We do not know whether those assumptions are correct or not and in general, the histogram that we are going to get is not going to look like a normal one.



Although in this case it's not particularly far from the theoretical normal, and interestingly or curiously for this particular example, it turns out that the confidence intervals that we get from the bootstrap method are essentially the same as the ones that we got through the regression formulas and so, we also get the same confidence intervals. So, in this case, it is just a sanity check that the confidence intervals that were reported by the regression software seem to be correct and that the assumptions underlying those calculations are not violently violated.

On the other hand, in other situations, it may well be the case that the theoretical confidence intervals are quite wrong. But the bootstrap method that just uses the data, it sort of uses reality as opposed to using assumptions is going to deliver more correct and more reliable confidence intervals.

So, the bootstrap method is a very clever idea. It is a way of using just the data that we have to create estimates of standard errors and confidence intervals. We have now explained the two major ways that are being used for performance assessments.

Just to summarize, we have the validation methods and validation methods are focused on prediction performance. We use validation sets to compare different models,

methods, different parameters, and compare them in terms of the prediction performance that they yield, and when it comes to validation, we are looking at the performance that they yield on data that they had not seen before.

On the other hand, when we are interested in assessing the accuracy of parameter estimates, what we need is to calculate the standard errors of these parameter estimates. Sometimes we can do it with formulas, but when not, there is always this data-driven method, the bootstrap, that can do it for us and the general idea in both validation methods like in K-fold validation or in bootstrap is to reuse the available data in clever ways and for multiple purposes. There is certainly an art in applying those methods and there are many variations of how one applies them. But there are some key basic ideas behind them and the reuse of data and trying to do as much as we can in a data-driven way as opposed to using formulas, is one of the central ideas here.