STAT 830

The basics of nonparametric models

The Empirical Distribution Function – EDF

The most common interpretation of probability is that the probability of an event is the long run relative frequency of that event when the basic experiment is repeated over and over independently. So, for instance, if X is a random variable then $P(X \leq x)$ should be the fraction of X values which turn out to be no more than x in a long sequence of trials. In general an empirical probability or expected value is just such a fraction or average computed from the data.

To make this precise, suppose we have a sample $X_1, \ldots, X_n$ of iid real valued random variables. Then we make the following definitions:

Definition: The empirical distribution function, or EDF, is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x).$$

This is a cumulative distribution function. It is an estimate of F, the cdf of the Xs. People also speak of the empirical distribution of the sample:

$$\hat{P}(A) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in A)$$

This is the probability distribution whose cdf is $\hat{F}_n$.

Now we consider the qualities of $\hat{F}_n$ as an estimate, the standard error of the estimate, the estimated standard error, confidence intervals, simultaneous confidence intervals and so on. To begin with we describe the best known summaries of the quality of an estimator: bias, variance, mean squared error and root mean squared error.

Bias, variance, MSE and RMSE

There are many ways to judge the quality of estimates of a parameter ϕ; all of them focus on the distribution of the estimation error $\hat{\phi} - \phi$. This distribution

is to be computed when ϕ is the *true* value of the parameter. For our non-parametric iid sampling model the estimation error we are interested in is

$$\hat{F}_n(x) - F(x)$$

where F is the true distribution function of the Xs.

The simplest summary of the size of a variable is the *root mean squared error*:

$$\text{RMSE} = \sqrt{\text{E}_\theta \left[(\hat{\phi} - \phi)^2 \right]}$$

In this definition the subscript θ on E is important; it specifies the true value of θ and the value of ϕ in the error must match the value of θ. For example if we were studying the $N(\mu, 1)$ model and estimating $\phi = \mu^2$ then the θ in the subscript would be μ and the ϕ in the error would be μ^2 and the two values of μ would be required to be the same.

The RMSE is measured in the same units as ϕ. That is if the parameter ϕ is a certain number of dollars then the RMSE is also some number of dollars. This makes the RMSE more scientifically meaningful than the more commonly discussed (by statisticians) mean squared error or MSE. The latter has, however, no square root and many formulas involving the MSE therefore look simpler. The weakness of MSE is one that it shares with the variance. For instance one might survey household incomes and get a mean of say \$40,000 with a standard deviation of \$50,000 because income distributions are very skewed to the right. The variance of household income would then be 2,500,000,000 squared dollars – ludicrously hard to interpret.

Having given that warning, however, it is time to define the MSE:

Definition: The mean squared error (MSE) of any estimate is

$$\begin{aligned} \text{MSE} &= \text{E}_\theta \left[(\hat{\phi} - \phi)^2 \right] \\ &= \text{E}_\theta \left[(\hat{\phi} - \text{E}_\theta(\hat{\phi}) + \text{E}_\theta(\hat{\phi}) - \phi)^2 \right] \\ &= \text{E}_\theta \left[(\hat{\phi} - \text{E}_\theta(\hat{\phi}))^2 \right] + \left\{ \text{E}_\theta(\hat{\phi}) - \phi \right\}^2 \end{aligned}$$

In making this calculation there was a cross product term which you should check is 0. The two terms in this formula have names: the first is the variance of $\hat{\phi}$ while the second is the square of the bias.

Definition: The **bias** of an estimator $\hat{\phi}$ is

$$\text{bias}_{\hat{\phi}}(\theta) = \text{E}_\theta(\hat{\phi}) - \phi$$

Notice that it depends on θ. The ϕ on the right hand side also depends on the parameter θ.

Thus our decomposition above says

$$\text{MSE} = \text{Variance} + (\text{bias})^2.$$

In practice we often find there is a trade-off; if we try to make the variance small we often increase the bias. Statisticians often speak of a "variance-bias trade-off".

We now apply these ideas to the EDF. The EDF is an *unbiased* estimate of F. That is,

$$\begin{aligned}\text{E}[\hat{F}_n(x)] &= \frac{1}{n}\sum_{i1=}^{n}\text{E}[1(X_i \le x)] \\ &= \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x)\end{aligned}$$

so the bias of $\hat{F}_n(x)$ is 0. The mean squared error is then

$$\text{MSE} = \text{Var}(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[1(X_i \le x)] = \frac{1}{n}F(x)[1-F(x)].$$

This is very much the most common situation: the MSE is proportional to $1/n$ in large samples. So the RMSE is proportional to $1/\sqrt{n}$.

The EDF is a sample average and all sample averages are unbiased estimates of their expected values. There are many estimates in use – most estimates really – which are biased. Here is an example. Again we consider a sample $X_1, \ldots, X_n$. The sample mean is

$$\bar{X} = \frac{1}{n}X_i.$$

The sample second moment is

$$\bar{X^2} = \frac{1}{n}X_i^2.$$

These two estimates are unbiased estimates of $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$. We might combine them to get a natural estimate of σ^2 if we remember that

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2.$$

It would then be natural to use $\bar{X}^2$ to estimate μ^2 which would lead to the following estimate of σ^2:

$$\hat{\sigma}^2 = \bar{X^2} - \bar{X}^2.$$

This estimate is biased, however, because it is a non-linear function of $\bar{X}$. In fact we find

$$\mathrm{E}\left[(\bar{X})^2\right] = \mathrm{Var}(\bar{X}) + \left[\mathrm{E}(\bar{X})\right]^2 = \sigma^2/n + \mu^2.$$

So the bias of $\hat{\sigma}^2$ is

$$\mathrm{E}\left[\bar{X^2}\right] - \mathrm{E}\left[(\bar{X})^2\right] - \sigma^2 = \mu_2' - \mu^2 - \sigma^2/n - \sigma^2 = -\sigma^2/n.$$

In this case and many others the bias is proportional to $1/n$. The variance is proportional to $1/n$. The squared bias is proportional to $1/n^2$. So in large samples the variance is more important than the bias!

Remark: The biased estimate $\hat{\sigma}^2$ is traditionally changed to the usual sample variance $s^2 = n\hat{\sigma}^2/(n-1)$ to remove the bias.
WARNING: the MSE of s^2 is larger than that of $\hat{\sigma}^2$.

Standard Errors and Interval Estimation

Traditionally theoretical statistics courses spend a considerable amount of time on finding good estimators of parameters. The theory is elegant and sophisticated but point estimation itself is a silly exercise which we will not pursue here. The problem is that a bare estimate is of very little value indeed. Instead assessment of the likely size of the error of our estimate is essential. A confidence interval is one way to provide that assessment.

The most common kind of confidence interval is approximate:

$$\text{estimate} \pm 2 \text{ estimated } \textbf{standard error}$$

This is an interval of values $L(X) <$ parameter $< U(X)$ where U and L are random because they depend on the data.

What is the justification for the two SE interval above? In order to explain we introduce some notation.

Notation: Suppose $\hat{\phi}$ is the estimate of ϕ. Then $\hat{\sigma}_{\hat{\phi}}$ denotes the estimated standard error.

We often use the central limit theorem, the delta method, and Slutsky's theorem to prove

$$\lim_{n\to\infty} P_F\left(\frac{\hat{\phi}-\phi}{\hat{\sigma}_{\hat{\phi}}} \le x\right) = \Phi(x)$$

where Φ is the standard normal cdf:

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du.$$

Example: We illustrate the ideas by giving first what we will call *pointwise* confidence limits for $F(x)$. Define, as usual, the notation for the upper α critical point z_α by the requirement $\Phi(z_\alpha) = 1-\alpha$. Then we approximate

$$P_F\left(-z_{\alpha/2} \le \frac{\hat{\phi}-\phi}{\hat{\sigma}_{\hat{\phi}}}\ lez_{\alpha/2}\right) \approx 1-\alpha.$$

Then we solve the inequalities inside the probability to get the usual interval.

Now we apply this to $\phi = F(x)$ for one fixed x. Our estimate is $\hat{\phi} \equiv \hat{F}_n(x)$. The random variable $n\hat{\phi}$ has a Binomial distribution. So $\text{Var}(\hat{F}_n(x)) = F(x)(1-F(x))/n$. The standard error is

$$\sigma_{\hat{\phi}} \equiv \sigma_{\hat{F}_n(x)} \equiv \text{SE} \equiv \frac{\sqrt{F(x)[1-F(x)]}}{\sqrt{n}}.$$

According to the central limit theorem

$$\frac{\hat{F}_n(x) - F(x)}{\sigma_{\hat{F}_n(x)}} \xrightarrow{d} N(0,1)$$

(In the homework I ask you to turn this into a confidence interval.)

It is easier to solve the inequality

$$\left|\frac{\hat{F}_n(x) - F(x)}{\text{SE}}\right| \le z_{\alpha/2}$$

if the term SE does not contain the unknown quantity $F(x)$. In the example above it did but we will modify the SE term by estimating the standard error. The method we follow uses a so-called *plug-in* procedure.

In our example we will estimate $\sqrt{F(x)[1-F(x)]/n}$ by replacing $F(x)$ by $\hat{F}_n(x)$:

$$\hat{\sigma}_{F_n(x)} = \sqrt{\frac{\hat{F}_n(x)[1-\hat{F}_n(x)]}{n}}.$$

This is an example of a general strategy in which we start with an estimator, a confidence interval or a test statistic whose formula depends on some other parameter; we plug-in an estimate of that other parameter to the formula and then use the resulting object in our inference procedure. Sometimes the method changes the behaviour of our procedure and sometimes, at least in large samples, it doesn't.

In our example Slutsky's theorem shows

$$\frac{\hat{F}_n(x) - F(x)}{\hat{\sigma}_{F_n(x)}} \xrightarrow{d} N(0,1).$$

So there was no change in the limit *law* (which is common alternative jargon for the word *distribution*).

We now have two pointwise 95% confidence intervals:

$$\hat{F}_n(x) \pm z_{0.025}\sqrt{\hat{F}_n(x)[1-\hat{F}_n(x)]/n}$$

or

$$\{F(x) : \left|\frac{\sqrt{n}(\hat{F}_n(x) - F(x))}{\sqrt{F(x)[1-F(x)]}}\right| \leq z_{0.025}\}$$

When we use these intervals they depend on x. Moreover, we usually look at a plot of the results against x. This leads to a problem. If we pick out an x for which the confidence interval is surprising or interesting to us we may well be picking one of the x values for which the confidence interval misses its target. After all, 1 out of every 20 confidence intervals with 95% coverage probabilities misses its target.

This suggests that what we really want is

$$P_F(L(X,x) \leq F(x) \leq U(X,x) \text{ for all } x) \geq 1-\alpha.$$

In that case the confidence intervals are called *simultaneous*. Thee are at least two possible methods: one is exact (meaning that the coverage probability of a 95% confidence interval is at least 95% for *every* choice of F, but conservative (meaning that the coverage is often quite a bit larger that 95% so that the interval is unnecessarily wide); the other method is approximate and less conservative. Here are some incomplete details.

The exact, conservative, procedure is based on the Dvoretsky-Kiefer-Wolfowitz inequality:

$$P_F(\exists x : |\hat{F}_n(x) - F(x)| > \sqrt{\frac{-\log(\alpha/2)}{2n}}) \leq \alpha$$

The use of this inequality to generate confidence intervals is quite uncommon – the homework problems ask you to compare it to the next interval and to criticize its properties.

The approximate procedure is based on large sample limit theory. The following assertion is a famous piece of probability theory:

$$\lim_{n\to\infty} P_F(\exists x : |\sqrt{n}|\hat{F}_n(x) - F(x)| > y) = P(\exists t \in [0,1] : |B_0(t)| > y)$$

where B_0 is a *Brownian Bridge*. A Brownian Bridge is a stochastic process; in particular it is a Gaussian process, with mean $\mathrm{E}(B_0(x)) \equiv 0$ and covariance function

$$\mathrm{Cov}(B_0(x), B_0(y)) = \min\{x, y\} - xy.$$

I won't be describing precisely what that all means. You might consult some book or other which I will eventually cite I hope. It is possible, however, to choose y so that the probability on the right hand side above is α.

Statistical Functionals

Not all parameters are created equal. Some of them have a meaning for all or at least most distribution functions or densities while others really only have a meaning inside some quite specific model. For instance, in the Weibull model density

$$f(x; \alpha, \beta) = \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\{-(x/\beta)^\alpha\} 1(x > 0).$$

there are two parameters: shape α and scale β. These parameters have no meaning in other densities; that is if the real density is normal we cannot say what α and β are. But every distribution has a median and other quantiles:

$$p^{\text{th}}\text{-quantile} = \inf\{x : F(x) \geq p\}.$$

Too, if r is a bounded function then every distribution has a value for the parameter

$$\phi \equiv \mathrm{E}_F(r(X)) \equiv \int r(x)dF(x).$$

Similarly, most distributions have a mean, variance and so on.

Definition: A function from set of all cdfs to real line is called a *statistical functional*.

Example: The quantity $T(F) \equiv \mathrm{E}_F(X^2) - [\mathrm{E}_F(X)]^2$ is a statistical functional, namely, the variance of F. It is not quite defined for all F but it is defined for most.

The statistical functional

$$T(F) = \int r(x)dF(x)$$

is linear. The sample variance is not a linear functional.

Statistical functionals are often estimated using plug-in estimates so that

$$\hat{T(F)} = \int r(x)d\hat{F}_n(x) = \frac{1}{n}\sum_1^n r(X_i).$$

This estimate is unbiased and has variance

$$\sigma^2_{\hat{T(F)}} = n^{-1}\left[\int r^2(x)dF(x) - \left\{\int r(x)dF(x)\right\}^2\right].$$

This variance can in turn be estimated using a plug-in estimate:

$$\hat{\sigma}^2_{\hat{T(F)}} = n^{-1}\left[\int r^2(x)d\hat{F}_n(x) - \left\{\int r(x)d\hat{F}_n(x)\right\}^2\right].$$

And of course from that estimated variance we get an estimated standard error.

Bootstrap standard errors

When $r(x) = x$ we have $T(T) = \mu_F$ (the mean). The standard error of this estimate is $\sigma/\sqrt{n}$. The plug-in estimate of the standard error replaces σ with the sample standard deviation (but with n not $n-1$ as the divisor).

Now consider a general functional $T(F)$. The plug-in estimate of this is $T(\hat{F}_n)$. The plug-in estimate of the standard error of this estimate is

$$\sqrt{\mathrm{Var}_{\hat{F}_n}(T(\hat{F}_n))}.$$

which is hard to read and seems hard to calculate in general. The solution is to simulate, particularly to estimate the standard error.

Basic Monte Carlo

To compute a probability or expected value we can simulate.

Example: To compute $P(|X| > 2)$ for some random variable X we use software to generate some number, say M, of replicates: $X_1^*, \ldots, X_M^*$ all having same distribution as X. Then we estimate the desired probability using the sample fraction. Here is some `R` code:

```
x=rnorm(1000000)
y =rep(0,1000000)
y[abs(x) >2] =1
sum(y)
```

This produced 45348 when I tried it which gives me the estimate $\hat{p} = 0.045348$. Using `pnorm` I find the correct answer is 0.04550026. So using a million samples gave 2 correct significant digits and an error of 2 in the third digit. Using $M = 10000$ has traditionally been more common, though I think that is changing. Using 10000, I got $\hat{p} = 0.0484$. In fact, the SE of $\hat{p}$ is $\sqrt{p(1-p)}/100 = 0.0021$. So error of up to 4 in second significant digit is reasonably likely.

The bootstrap

In the previous section we were drawing samples from some specific distribution – the normal distribution in the example. In bootstrapping the random

variable X is replaced by the whole data set and we simulate by drawing samples from the distribution $\hat{F}_n$.

The idea is to generate new data sets (I will use a superscript $*$ as in X^* to indicate these are newly generated data sets) from the distribution F of X. Bu we don't know F so we use $\hat{F}_n$.

Example: Suppose we are interested in confidence intervals for the mean of a distribution. We will get them by simulating the distribution of the t pivot:

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{s}.$$

We have data $X_1, \ldots, X_n$ and as usual for statisticians we don't know μ or the cumulative distribution function F of the Xs. So we replace these by quantities computed from $\hat{F}_n$. Call $\mu^* = \int x d\hat{F}_n(x) = \bar{X}$. Then draw $X^*_{1,1}, \ldots, X^*_{1,n}$ an iid sample from the cdf $\hat{F}$. Repeat this sampling process M times computing t from the * values each time. Here is R code:

```
x=runif(5)
mustar = mean(x)
tv=rep(0,M)
tstarv=rep(0,M)
for( i in 1:M){
  xn=runif(5)
  tv[i]=sqrt(5)*mean(xn-0.5)/sqrt(var(xn))
  xstar=sample(x,5,replace=TRUE)
  tstarv[i]=sqrt(5)*mean(xstar-mustar)/sqrt(var(xstar))
}
```

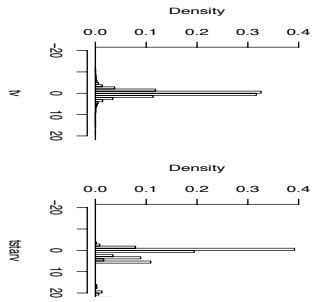
This loop does two simulations. First, the variables `xn` and `tv` implement *parametric bootstrapping*: they simulate the t-pivot from a parametric model, namely, the Uniform[0,1] model. On the other hand `xstar` is bootstrap sample from the population `x` and `tstarv` is the t-pivot computed from `xstar`.

When I ran the code the first time I got the original data set

$$\texttt{x} = (0.7432447, 0.8355277, 0.8502119, 0.3499080, 0.8229354)$$

So `mustar` =0.7203655. Now let us look at side-by-side histograms of `tv` and `tstarv`:

Figure 1: Histograms of simulations. The histogram on the left is of t pivots for 1,000,000 samples of size 5 drawn from the uniform distribution. The one on the right is for t pivots computed for 1,000,000 samples of size 5 drawn from the population with just 5 elements as specified in the text



Confidence intervals: based on t-statistic: $T = \sqrt{n}(\bar{X} - \mu)/s$.
Use the bootstrap distribution to estimate $P(|T| > t)$.
Adjust t to make this 0.05. Call result c. Solve $|T| < c$ to get interval

$$\bar{X} \pm cs/\sqrt{n}.$$

Get $c = 22.04$, $\bar{x} = 0.720$, $s = 0.211$; interval is -1.36 to 2.802. Pretty lousy interval. Is this because it is a bad idea? Repeat but simulate $\bar{X}^* - \mu^*$. Learn

$$P(\bar{X}^* - \mu^* < -0.192) = 0.025 = P(\bar{X}^* - \mu^* > 0.119)$$

Solve inequalities to get (much better) interval

$$0.720 - 0.119 < \mu < 0.720 + 0.192$$

Of course the interval missed the true value!

Monte Carlo Study

So how well do these methods work? We can do either a theoretical analysis or a simulation study. To describe the possible theoretical analysis let C_n be resulting interval. Usually we assume the number of bootstrap repetitions is so large that we can ignore that simulation error. Now we use theory (more sophisticated than in this course) to compute

$$\lim_{n\to\infty} P_F(\mu(F) \in C_n)$$

We say the method is *asymptotically valid* (or calibrated or accurate) if this limit is $1 - \alpha$.

The other way to assess this point is via simulation analysis: we generate many data sets of size 5 from say the Uniform[0,1] distribution. Then we carry out the bootstrap method for each data set and compute the interval C_n. Finally we count up the number of simulated uniform data sets with $0.5 \in C_n$ to get an *empirical* coverage probability. Since we will be using the method *without* knowing the true distribution we repeat the process with samples from (many) other distributions. We try to select enough distributions to give us a pretty good idea of the overall behaviour.

Remark: : Some statisticians never do anything except the simulation part. I think this is somewhat perilous – you are hard pressed to guarantee that your simulation covered all the realistic possibilities. But then, I do theory for a living.

Here is some `R` code which carries out a bit of that Monte Carlo study:

```
tstarint = function(x,M=10000){
  n = length(x)
  must=mean(x)
  se=sqrt(var(x)/n)
  xn=matrix(sample(x,n*M,replace=T),nrow=M)
  one = rep(1,n)/n
  dev= xn%*%one - must
  tst=dev/sqrt(diag(var(t(xn)))/n)
  c1=quantile(dev,c(0.025,0.975))
  c2=quantile(abs(tst),0.95)
  c(must-c1[2],must-c1[1], must -c2*se,must+c2*se)
}

lims=matrix(0,1000,4)
count=lims
for(i in 1:1000){
  x=runif(5)
  lims[i,]=tstarint(x)
}
count[,1][lims[,1]<0.5]=1
count[,2][lims[,2]>0.5]=1
count[,3][lims[,3]<0.5]=1
count[,4][lims[,4]>0.5]=1
sum(count[,1]*count[,2])
sum(count[,3]*count[,4])
```

The results for the study I did are these. For samples of size 5 from the Uniform[0,1] distribution the empirical coverage probability for the true mean of 1/2, using the bootstrap distribution of the error $\bar{X} - \mu$ is 80.4% in 1000 Monte Carlo trials using 10,000 bootstrap samples in each trial. This compares to coverage of 97.2% under the same conditions using the t pivot $\sqrt{n})(\bar{X} - \mu)/s$. (Strictly speaking t is only an approximate pivot.) For samples of size 25 I got 92.1% and 94.8%. I also tried exponential data. For $n = 5$ I got ? and ? while for $n = 25$ I got 92.1% and 94.1%.

Remark: : It is possible to put standard errors on these Monte Carlo estimates and to assess the inaccuracy induced by using only 10,000 bootstrap samples instead of infinitely many. Ignoring the latter you should be able to

add standard errors to each of the percentages given. They are all roughly 0.007 or 0.7 percentage points.