

STAT 830

Non-parametric Inference Basics

Richard Lockhart

Simon Fraser University

STAT 830 — Fall 2011



The Empirical Distribution Function – EDF pp 97-99

- Suppose we have sample $X_1, \dots, X_n$ of iid real valued rvs.
- The empirical distribution function is

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

- This is a cdf and is an estimate of F , the cdf of the X s.
- People also speak of the empirical distribution:

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

- This is the probability distribution corresponding to $\hat{F}_n$.
- 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.



Bias, variance and mean squared error

- Judge estimates in many ways; focus is distribution of error $\hat{\theta} - \theta$.
- Distribution computed when θ is *true* value of parameter.
- For our non-parametric iid sampling model we are interested in

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

when F is the true distribution function of the X s.

- Simplest summary of size of a variable is root mean squared error:

$$RMSE = \sqrt{E_{\theta} [(\hat{\theta} - \theta)^2]}$$

- Subscript θ on E is important – specifies true value of θ and matches θ in the error!



MSE decomposition & variance-bias trade-off

- The MSE of any estimate is

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

- In making this calculation there was a cross product term which is 0.
- The two terms each have names: the first is the variance of $\hat{\theta}$ while the second is the square of the bias.
- **Def'n:** The **bias** of an estimator $\hat{\theta}$ is

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

- So our decomposition is

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

- In practice often find a trade-off. Trying to make the variance small increases the bias.



Applied to the EDF

- The EDF is an *unbiased* estimate of F . That is:

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

so the bias is 0.

- The mean squared error is then

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

- This is very much the most common situation: the MSE is proportional to $1/n$ in large samples.
- So the RMSE is proportional to $1/\sqrt{n}$.
- RMSE is measured in same units as $\hat{\theta}$ so is scientifically right.



Biased estimates

- Many estimates exactly or approximately averages or ftns of averages.
- So, for example,

$$\bar{X} = \frac{1}{n} \sum X_i \quad \text{and} \quad \bar{X}^2 = \frac{1}{n} \sum X_i^2$$

are unbiased estimates of $E(X)$ and $E(X^2)$.

- We might combine these to get a natural estimate of σ^2 :

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

- This estimate is biased:

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

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

$$E[\bar{X}^2] - E[(\bar{X})^2] - \sigma^2 = \mu'_2 - \mu^2 - \sigma^2/n - \sigma^2 = -\sigma^2/n.$$



Relative sizes of bias and variance

- 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!
- 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

- In any case point estimation is a silly exercise.
- Assessment of likely size of error of estimate is essential.
- A confidence interval is one way to provide that assessment.
- The most common kind is approximate:

estimate ± 2 estimated **standard error**

- This is an interval of values $L(X) < \text{parameter} < U(X)$ where U and L are random.
- Justification for the two se interval above?
- Notation $\hat{\phi}$ is the estimate of ϕ . $\hat{\sigma}_{\hat{\phi}}$ is the estimated standard error.
- Use central limit theorem, delta method, Slutsky's theorem to prove

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



Pointwise limits for $F(x)$

- Define, as usual z_α by $\Phi(z_\alpha) = 1 - \alpha$ and approximate

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

- Solve inequalities to get 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

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

- See homework to turn this into a confidence interval.



Plug-in estimates

- Now to estimate the standard error.
- It is easier to solve the inequality

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

if the term SE does not contain the unknown quantity $F(x)$.

- This is why we use an estimated standard error.
- 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: *plug-in*.
- Start with estimator, confidence interval or test whose formula depends on other parameter; plug-in estimate of that other parameter.
- Sometimes the method changes the behaviour of our procedure and sometimes, at least in large samples, it doesn't.



Pointwise versus Simultaneous Confidence Limits

- 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* (alternative jargon for 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

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

- When we use these intervals they depend on x .
- And we usually look at a plot of the results against x .
- If we pick out an x for which the confidence interval is surprising to us we may well be picking one of the x values for which the confidence interval misses its target.



Simultaneous intervals

- So we really want

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

- In that case the confidence intervals are called *simultaneous*.
- Two possible methods: one exact, but conservative, one approximate, less conservative.
- Dvoretzky-Kiefer-Wolfowitz inequality:

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

- Limit theory:

$$P_F(\exists x : |\sqrt{n}|\hat{F}_n(x) - F(x)| > y) \rightarrow P(\exists x : |B_0(x)| > y)$$

where B_0 is a *Brownian Bridge* (special Gaussian process).



Statistical Functionals

- Not all parameters are created equal.
- 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.
- But every distribution has a median and other quantiles:

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

- If r is bounded ftn then every distribution has value for parameter

$$\phi \equiv E_F(r(X)) \equiv \int r(x) dF(x).$$

- Most distributions have a mean, variance and so on.
- A ftn from set of all cdfs to real line is called a *statistical functional*.
- Example: $E_F(X^2) - [E_F(X)]^2$.



Statistical functionals

- 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

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

- This estimate is unbiased and has variance

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

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

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



Plug-in estimates of functionals; bootstrap standard errors

- When $r(x) = x$ we have $T(T) = \mu_F$ (the mean)
- The standard error is $\sigma/\sqrt{n}$.
- Plug-in estimate of SE replaces σ with sample SD (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

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

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 can simulate.
- **Example:** To compute $P(|X| > 2)$ use software to generate some number, say M , of replicates: $X_1^*, \dots, X_M^*$ all having same distribution as X .
- Estimate desired probability using sample fraction.
- R code: `x=rnorm(1000000) ; y =rep(0,1000000); y[abs(x)>2] =1 ; sum(y)`
- Produced 45348 when I tried it. Gives $\hat{p} = 0.045348$.
- Correct answer is 0.04550026.
- Using a million samples gave 2 correct digits, error of 2 in third digit.
- Using $M = 10000$ is more common. I got $\hat{p} = 0.0484$.
- SE of $\hat{p}$ is $\sqrt{p(1-p)}/100 = 0.0021$. So error of up to 4 in second significant digit is likely.



The bootstrap

- In bootstrapping X is replaced by the whole data set.
- Generate new data sets (X^*) from distribution F of X .
- Don't know F so use $\hat{F}_n$.
- **Example:** Interested in distribution of t pivot:

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

- Have data $X_1, \dots, X_n$. Don't know μ or cdf of X s.
- Replace these by quantities computed from $\hat{F}_n$.
- Call $\mu^* = \int x d\hat{F}_n(x) = \bar{X}$.
- Draw $X_{1,1}^*, \dots, X_{1,n}^*$ an iid sample from the cdf $\hat{F}$.
- Repeat M times computing t from $*$ values each time.



Bootstrapping the t pivot

- 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))
}
```



Bootstrapping a pivot continued

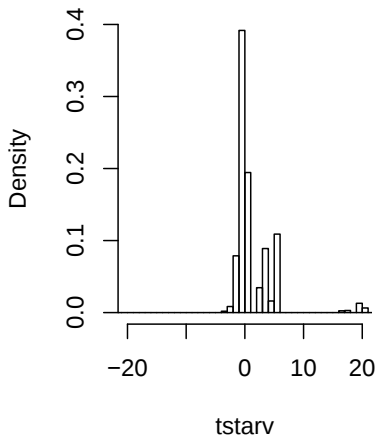
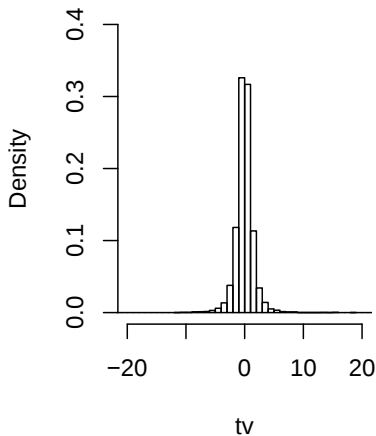
- Loop does two simulations.
- in x_n and t_v we do *parametric bootstrapping*: simulate t -pivot from parametric model.
- x_{star} is bootstrap sample from population x .
- t_{starv} is t -pivot computed from x_{star} .
- Original data set is

(0.7432447, 0.8355277, 0.8502119, 0.3499080, 0.8229354)

- So $mustar = 0.7203655$
- Side by side histograms of t_v and t_{starv} on next slide.



Bootstrap distribution histograms



Using the bootstrap distribution

- 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?
- Theoretical analysis: let C_n be resulting interval.
- Assume number of bootstrap reps is so large that we can ignore simulation error.
- Compute

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

- Method is *asymptotically valid* (or calibrated or accurate) if this limit is $1 - \alpha$.
- Simulation analysis: generate many data sets of size 5 from Uniform.
- Then bootstrap each data set, compute C_n .
- Count up number of simulated uniform data sets with $0.5 \in C_n$ to get coverage probability.
- Repeat with (many) other distributions.



R code

```
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)  
}
```



R code

```
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])
```



Results

- 804 out of 1000 intervals based on $\bar{X} - \mu$ cover the true value of 0.5.
- 972 out of 1000 intervals based on t statistics cover true value.
- This is the uniform distribution.
- Try another distribution. For exponential I get 909, 948.
- Try another sample size. For uniform $n = 25$ I got 921, 941.

