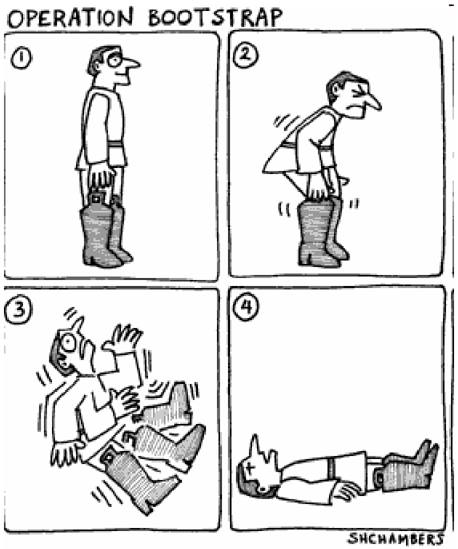
Opinionated Lessons in Statistics

by Bill Press

#23 Bootstrap Estimation of Uncertainty

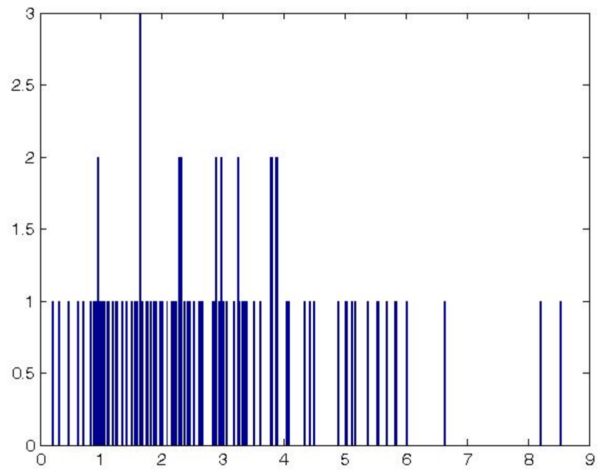
Method 3: Bootstrap resampling of the data



- We applied some end-to-end process to a data set (set of data points $\{\mathbf{x}_i, y_i\}$) and got a number f out
- The data set was drawn from a population of data points in repetitions of the identical experiment
  - which we don't get to see, unfortunately
  - we see only a sample of the population
- We'd like to draw new data sets from the population, reapply the process, and see the distribution of answers
  - this would tell us how accurate the original answer, on average, was
  - but we can't: we don't have access to the population
- However, the data set itself is an estimate of the population pdf!
  - in fact, it's the only estimate we've got!
- So we draw from the data set – with replacement – many "fake" data sets of equal size, and carry out the proposed program
  - does this sound crazy? for a long time many people thought so!
  - Bootstrap theorem [glossing over technical assumptions]: The distribution of any resampled quantity around its full-data-set value estimates (naively: "asymptotically has the same histogram as") the distribution of the data set value around the population value.

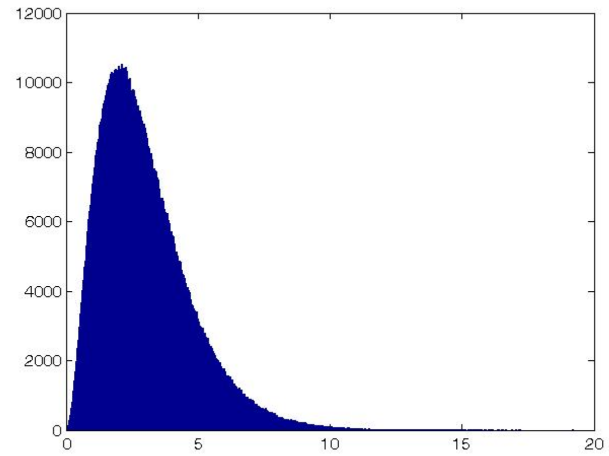
Let's try a simple example where we can see the "hidden" side of things, too.

Visible side (sample):



These happen to be drawn from a Gamma distribution.

Hidden side (population):



Statistic we are interested in happens to be (it could be anything):

$$\frac{\text{mean of distribution}}{\text{median of distribution}}$$

```
sammedian = median(sample)
sammean = mean(sample)
samstatistic = sammean/sammedian
sammedian =
    2.6505
sammean =
    2.9112
samstatistic =
    1.0984
```

How accurate is this?

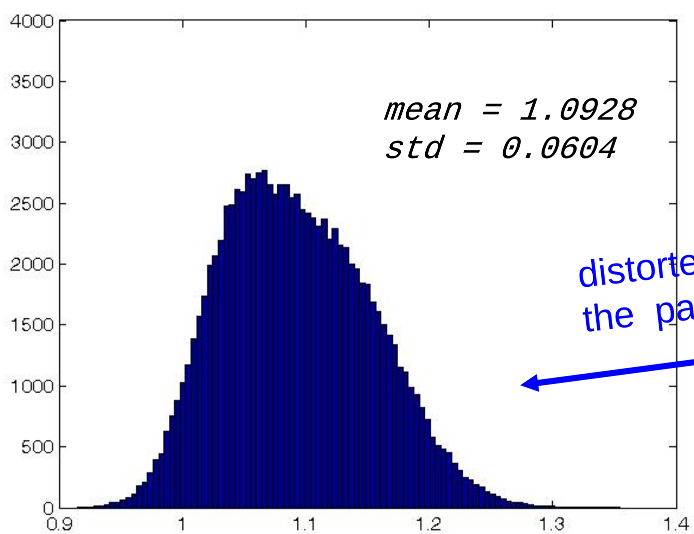
```
themedian = median(bigsample)
themean = mean(bigsample)
thestatistic = themean/themedian
themedian =
    2.6730
themean =
    2.9997
thestatistics =
    1.1222
```

To estimate the accuracy of our statistic, we bootstrap…

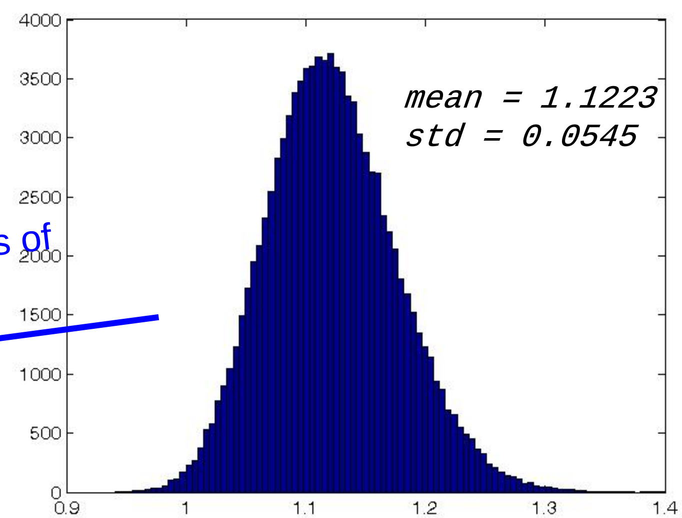
```
ndata = 100;
nboot = 100000;
vals = zeros(nboot,1);
for j=1:nboot,
   choose = randsample(ndata,ndata,true);
   vals(j) = mean(sample(choose))
          /median(sample(choose));
end
hist(vals,100)
```

new sample of integers in 1:ndata, with replacement

```
ndata = 100;
nboot = 100000;
vals = zeros(nboot,1);
for j=1:nboot,
   sam = randg(3,[ndata 1]);
   vals(j) = mean(sam)/median(sam);
end
hist(vals,100)
```



distorted by peculiarities of the particular data set

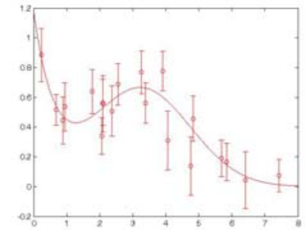


Things to notice:

The mean of resamplings does <u>not</u> improve the original estimate! (Same data!)

The distribution around the mean is not <u>identical</u> to that of the population. But it is <u>close</u> and would become identical asymptotically for large *ndata* (not *nboot*!).

You often have to customize your own bootstrap code, but it is not a big deal



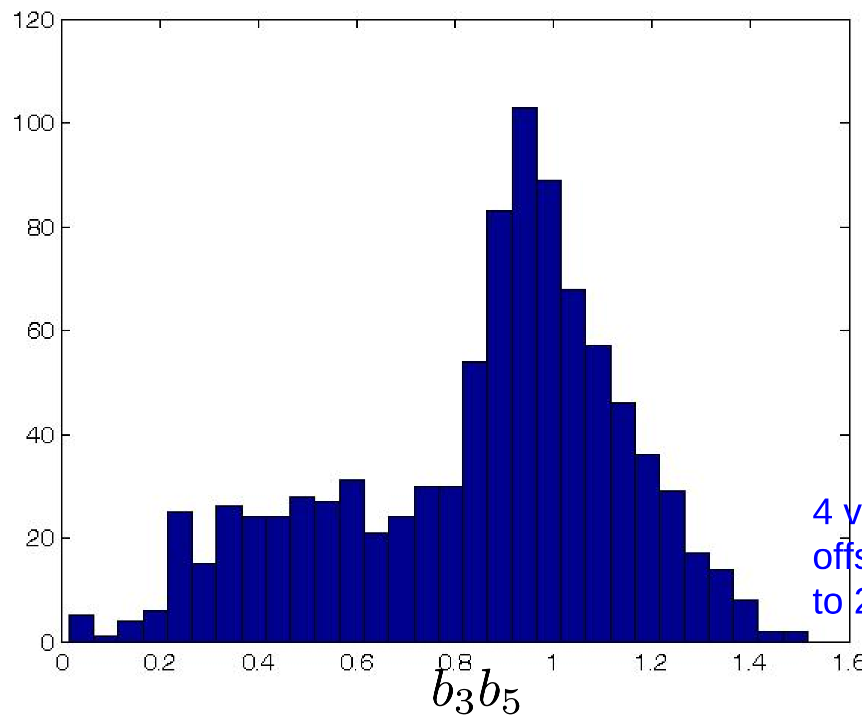
```
ndata = 20;
nboot = 1000;
vals = zeros(nboot,1);
ymodel = @(x,b) b(1)*exp(-b(2)*x)+b(3)*exp(-(1/2)*((x-b(4))/b(5)).^2);
for j=1:nboot,
   samp = randsample(ndata,ndata,true);
   xx = x(samp);
   yy = y(samp);
   ssig = sig(samp);
   chisqfun = @(b) sum(((ymodel(xx,b)-yy)./ssig).^2);
   bguess = [1 2 .7 3.14 1.5];
   options = optimset('MaxFunEvals',10000,'MaxIter',
          10000,'TolFun',0.001);
   [b fval flag] = fminsearch(chisqfun,bguess,options);
   if (flag == 1), vals(j) = b(3)*b(5);
   else vals(j) = 100; end
end
hist(vals(vals < 2),30);
std(vals(vals < 2))

      0.2924
```

new sample of integers in 1:ndata, with replacement

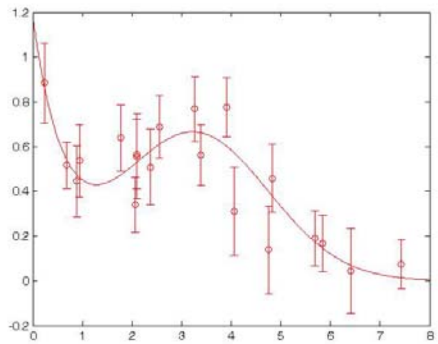
here is the embedded "whole statistical analysis of a data set" inside the bootstrap loop

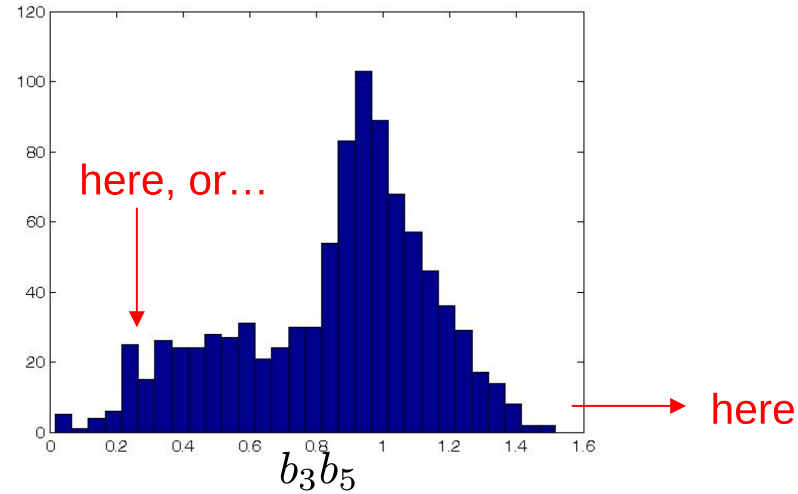
So we get the peak around 1, as before, but a much broader distribution.

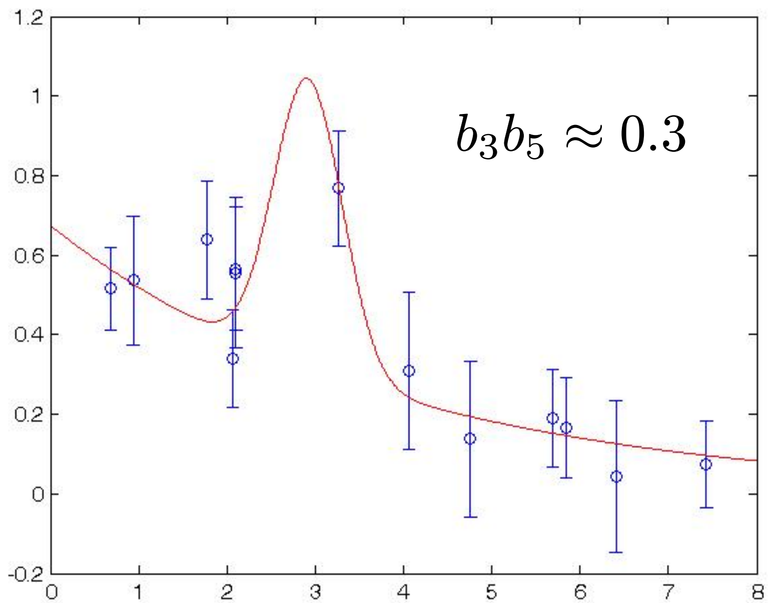


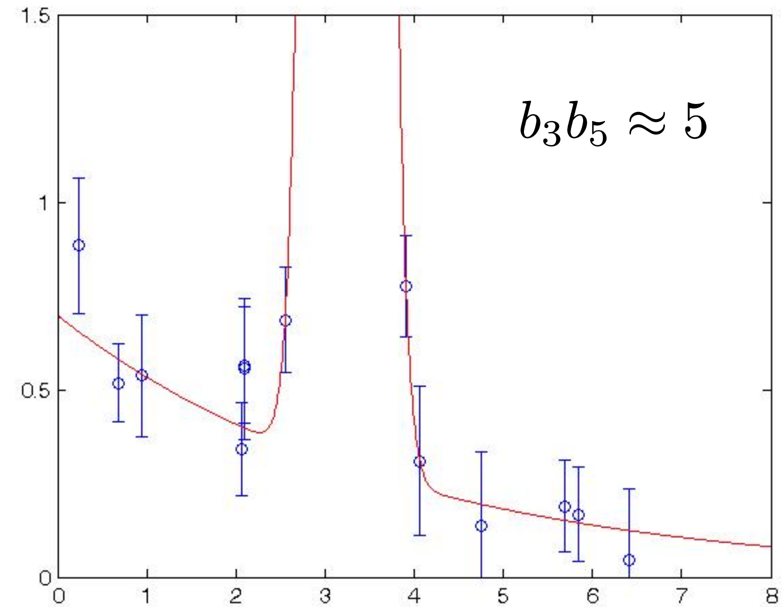
4 values offscale (up to 27!)

Can you guess what the extreme bootstrap cases look like, compared to the full data?



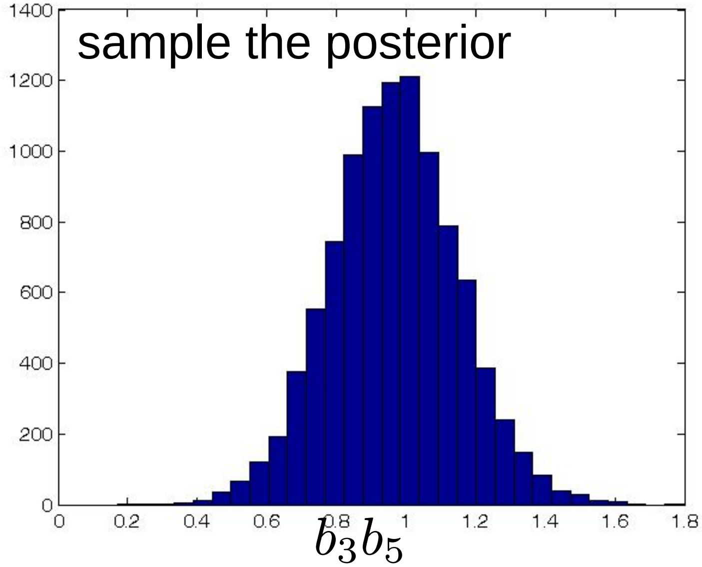


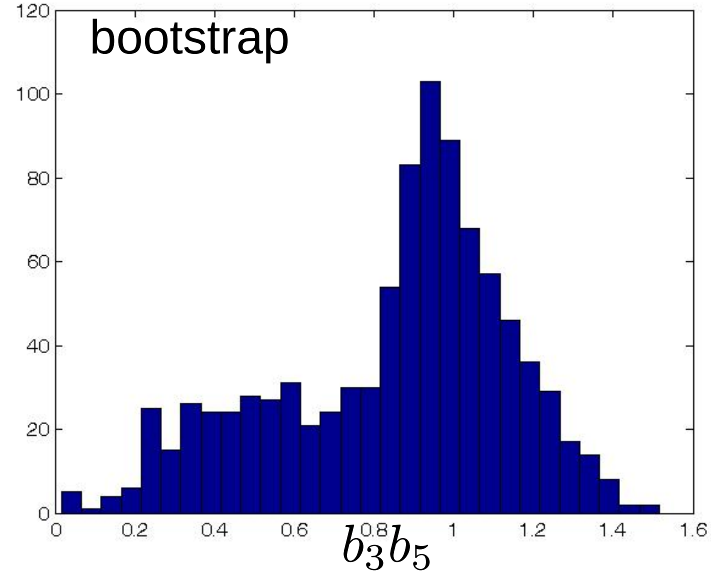




Is it “fair” for our estimate of b_3b_5 to have its accuracy “impuned” by data sets that “don’t look like” the full data? Deep frequentist philosophical question!

Let's compare boostrap-from-a-sample to sample-the-posterior:





- We could increase number of samples of posterior, and of bootstrap, to make both curves very smooth.
  - the histograms would not converge to each other!
- We could increase the size of the underlying data sample
  - from 20 (x,y) values to infinity (x,y) values
  - the histograms <u>would</u> converge to each other (modulo technical assumptions)
- For finite size samples, each technique is a valid answer to a different question
  - Frequentist: Imagining repetitions of the experiment, what would be the range of values obtained?
    - And. conservatively, I shouldn't expect my experiment to be better than that, should I?
  - Bayesian: For exactly the data that I see, what is the probability distribution of the parameters?
    - Because maybe I got lucky and my data set really nails the parameters!