Lectures 12: Bootstrap II.

bootstrap and its statistics reviewed

error on nonlinear function of fitted parameters?

What is the uncertainty in quantities other than the fitted coefficients:

I. Linearized error propagation

$\mathbf{b}_0$ is the MLE parameters estimate

$\mathbf{b}_1 \equiv \mathbf{b} - \mathbf{b}_0$ is the RV as the parameters fluctuate

$$f \equiv f(\mathbf{b}) = f(\mathbf{b}_0) + \nabla f\ \mathbf{b}_1 + \cdots$$

$$\langle f \rangle \approx \langle f(\mathbf{b}_0) \rangle + \nabla f\ \cancel{\langle \mathbf{b}_1 \rangle} = f(\mathbf{b}_0)$$

$$\left\langle f^2 \right\rangle - \langle f \rangle^2 \approx 2 f(\mathbf{b}_0)(\nabla f\ \cancel{\langle \mathbf{b}_1 \rangle}) + \left\langle (\nabla f\ \mathbf{b}_1)^2 \right\rangle$$

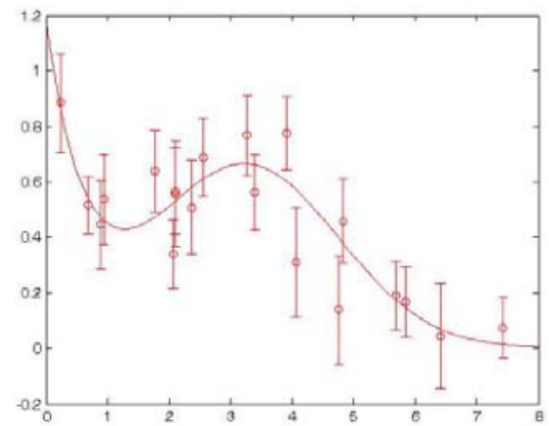
$$= \nabla f\ \left\langle \mathbf{b}_1 \mathbf{b}_1^T \right\rangle \nabla f\ ^T$$

$$= \nabla f\ \mathbf{\Sigma}\ \nabla f\ ^T$$

Linearized error propagation

In our example, if we are interested in the area of the “hump”,

```
bfit =
    1.1235    1.5210    0.6582    3.2654    1.4832
covar =
    0.1349    0.2224    0.0068   -0.0309    0.0135
    0.2224    0.6918    0.0052   -0.1598    0.1585
    0.0068    0.0052    0.0049    0.0016   -0.0094
   -0.0309   -0.1598    0.0016    0.0746   -0.0444
    0.0135    0.1585   -0.0094   -0.0444    0.0948
```



$$f = b_3 b_5$$

$$\nabla f = (0, 0, b_5, 0, b_3)$$

$$\nabla f \, \mathbf{\Sigma} \, \nabla f^T = b_5^2 \Sigma_{33} + 2 b_3 b_5 \Sigma_{35} + b_3^2 \Sigma_{55} = 0.0336$$

$$\sqrt{0.0336} = 0.18$$

So $b_3 b_5 = 0.98 \pm 0.18$ ← the one standard deviation (1-σ) error bar

Is it normally distributed?

Absolutely not! A function of normals is not normal (although, if they are all narrow, it might be close).

Sampling the posterior histogram

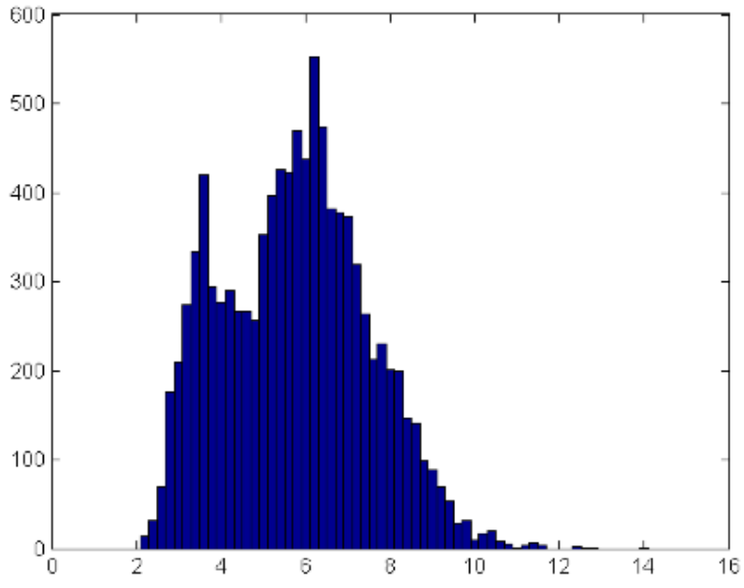
Method 2: Sample from the posterior distribution

1. Generate a large number of (vector) **b**'s

$$\mathbf{b} \sim \mathrm{MVNormal}(\mathbf{b}_0, \boldsymbol{\Sigma}_b)$$

2. Compute your $f(\mathbf{b})$ separately for each **b**

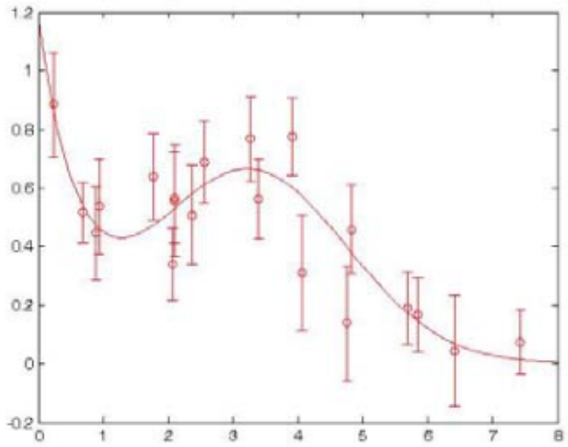
3. Histogram



Note again that **b** is typically (close to) m.v. normal because of the CLT, but your (nonlinear) f may not, in general, be anything even close to normal!

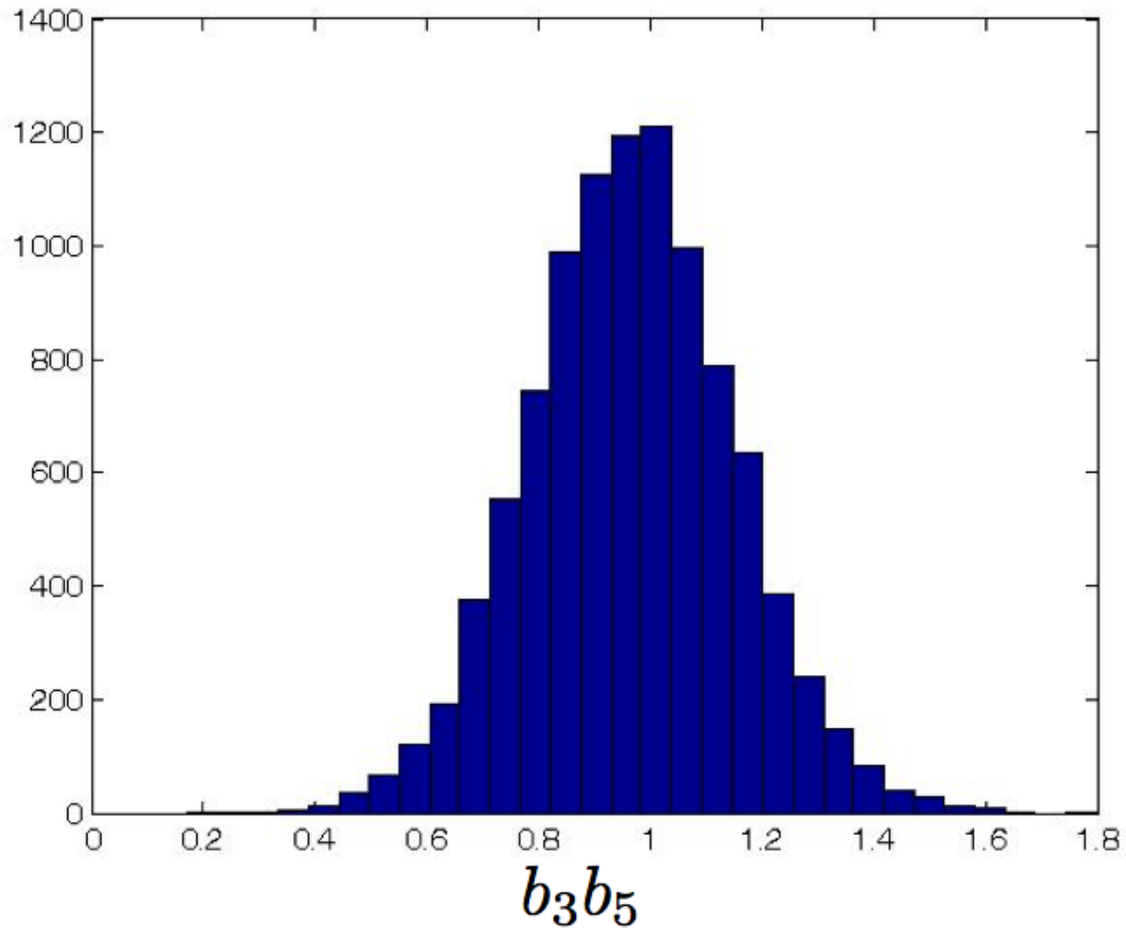
Sampling the posterior histogram

Our example:



```
bees = mvnrnd(bfit,covar,10000);
humps = bees(:,3).*bees(:,5);
hist(humps,30);
std(humps)
```

std = 0.1833



Does it matter that I use the full covar, not just the 2x2 piece for parameters 3 and 5?

comparison of linear propagation and posterior sampling:

Compare linear propagation of errors to sampling the posterior

- Note that even with lots of data, so that the distribution of the b's really → multivariate normal, a derived quantity might be very non-Normal.
  - In this case, sampling the posterior is a good idea!
- For example, the ratio of two normals of zero mean is Cauchy
  - which is very non-Normal!.
- So, sampling the posterior is a more powerful method than linear propagation of errors.
  - even when optimistically (or in ignorance) assuming multivariate Gaussian for the fitted parameters
- In fact, sampling the posterior distribution of large Bayesian models whose parameters are not at all Gaussian is, under the name MCMC, the most powerful technique in modern computational statistics.

bootstrap sampling

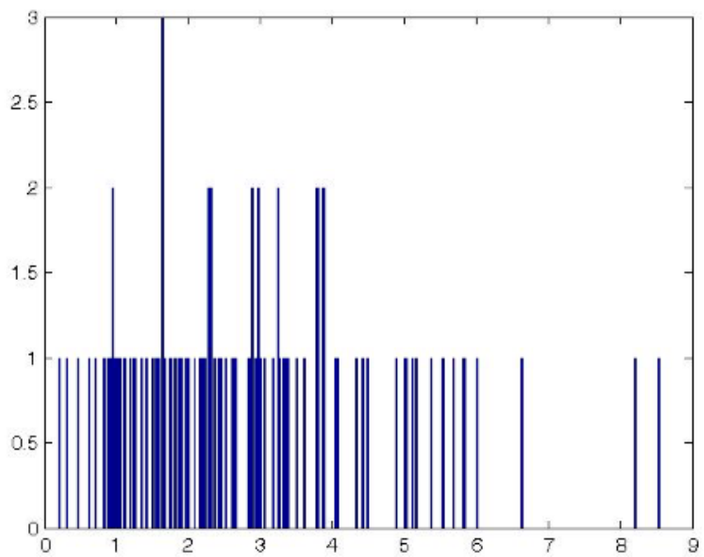
Method 3: Bootstrap resampling of the data

- We applied some end-to-end process to a data set and got a number f out
- The data set was drawn from a population of 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.

bootstrap sampling

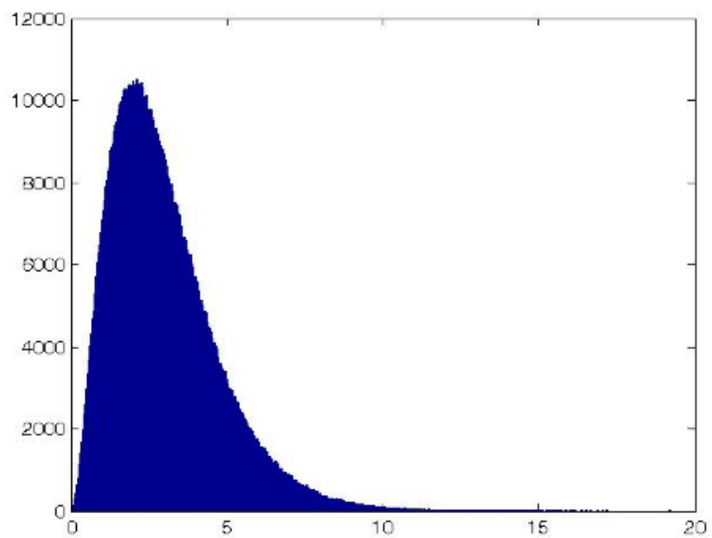
Let's try a simple example where we can see the "hidden" side of things, too.

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
```

bootstrap sampling

Gamma distribution:

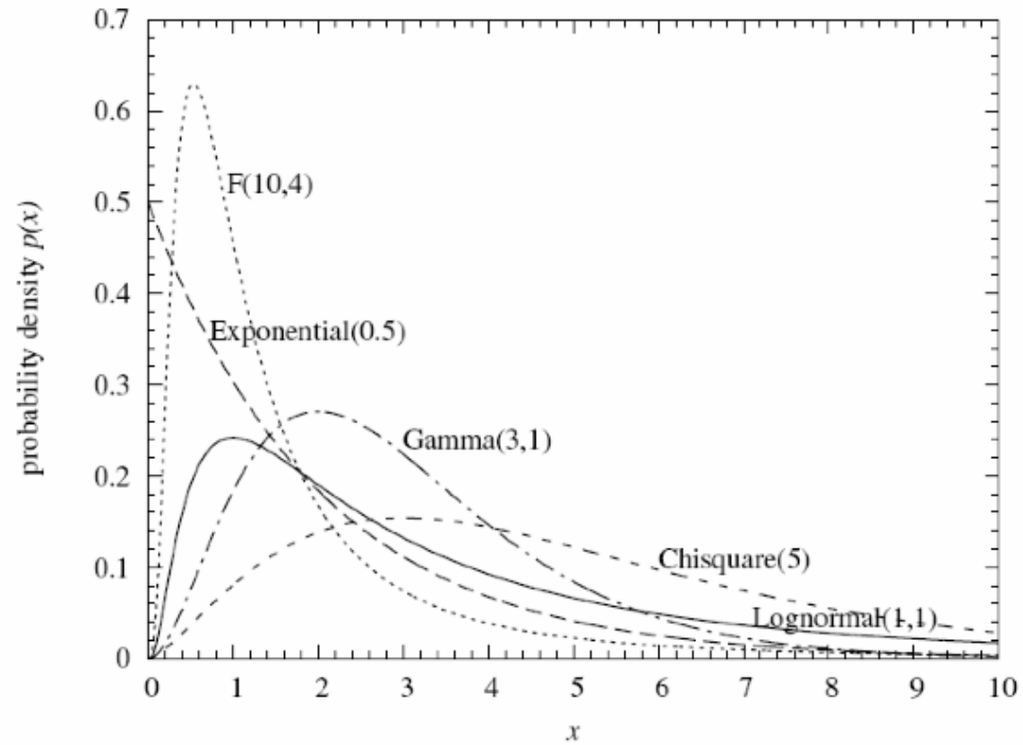
$$x \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0$$

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

$$\text{Mean}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta$$

$$\text{Var}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta^2$$

When $\alpha \geq 1$ there is a single mode at $x = (\alpha - 1)/\beta$

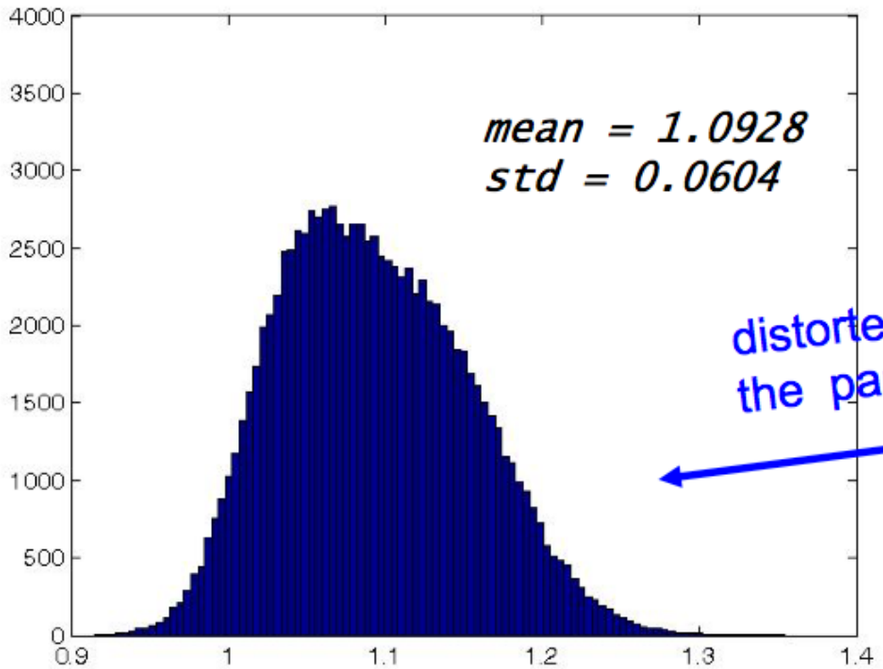


bootstrap sampling

To estimate the accuracy of our statistic, we bootstrap

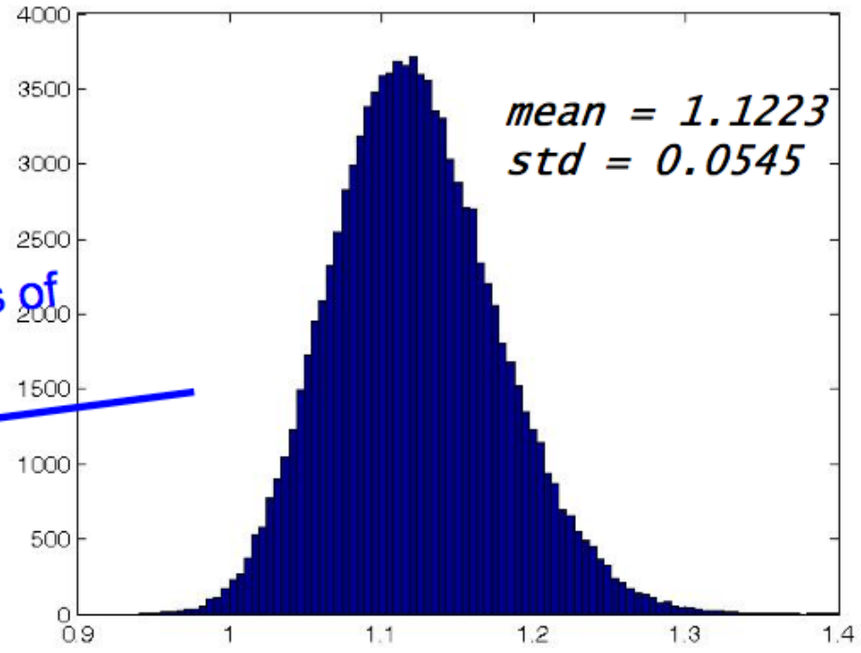
```
ndata = 100;
nboot = 100000;
vals = zeros(nboot,1);
for j=1:nboot,
   choose = randsample(ndata,ndata,true);
   vals(j) = mean(sample(choose))
          /median(sample(choose));
end
hist(vals,100)
```

new sample of integers in 1:ndata, with replacement



distorted by peculiarities of the particular data set

```
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)
```



Things to notice:

The mean of resamplings does not improve the original estimate! (Same data!)

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

statistics reviewed

The Empirical density function

Statistical inference concerns learning from experience: we observe a random sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and wish to infer properties of the complete population $\mathcal{X}$ that yielded the sample. A complete knowledge is obtained from the population density function $F(.)$ from which $\mathbf{x}$ has been generated $F \rightsquigarrow \mathbf{x} = (x_1, x_2, \cdots, x_n)$

Definition

The empirical density function $\hat{F}(.)$ is defined as:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

where $\delta(\cdot)$ is the Dirac delta function. So the probability of $x = x_j$ is :

$$\int \hat{F}(x_j)\, dx = \int \frac{1}{n} \sum_{i=1}^{n} \delta(x_j - x_i)\, dx = \begin{cases} \frac{1}{n}, & x_j \in \{x_1, \cdots, x_n\} \\ 0, & \text{otherwise} \end{cases}$$

statistics reviewed

Parameters

Definition

A parameter, θ, is a function of the probability density function (p.d.f.) F, e.g.:

$$\theta = t(F)$$

if θ is the mean

$$\theta = \mathbb{E}_F(x) = \int_{-\infty}^{+\infty} x\ F(x)dx = \mu_F$$

if θ is the variance

$$\theta = \mathbb{E}_F[(x - \mu_F)^2] = \int_{-\infty}^{+\infty} (x - \mu_F)^2\ F(x)dx = \sigma_F^2$$

α=3 β=1 a parameter θ on F: θ=mean/median

statistics reviewed

Gamma distribution:

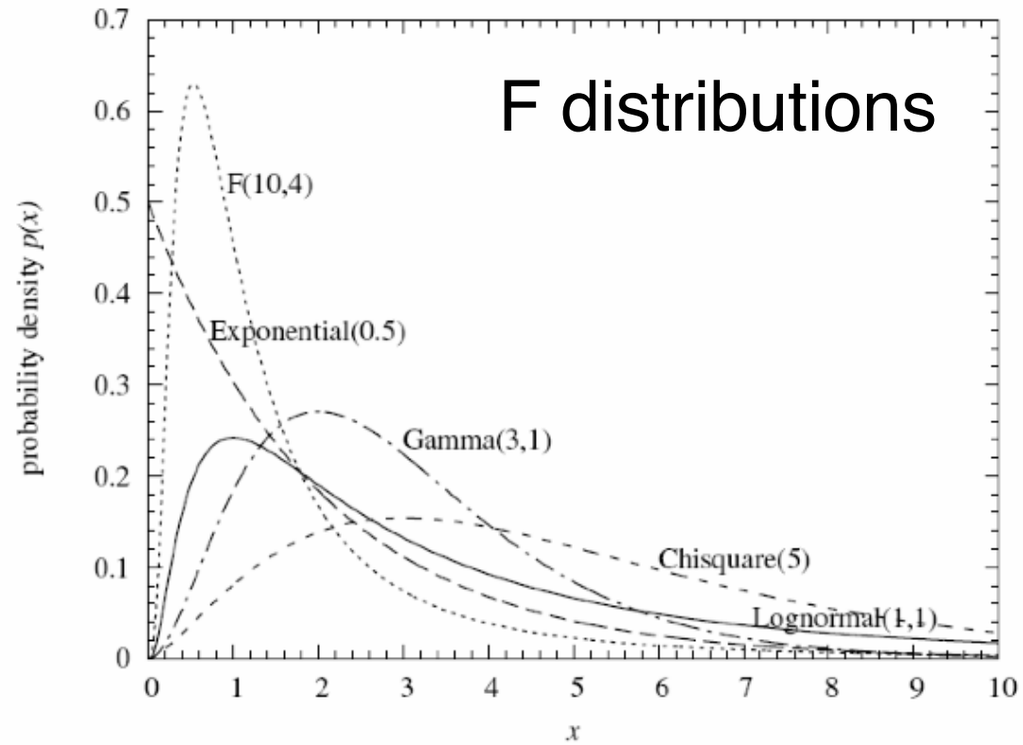
$$x \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0$$

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

$$\text{Mean}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta$$

$$\text{Var}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta^2$$

When $\alpha \geq 1$ there is a single mode at $x = (\alpha - 1)/\beta$

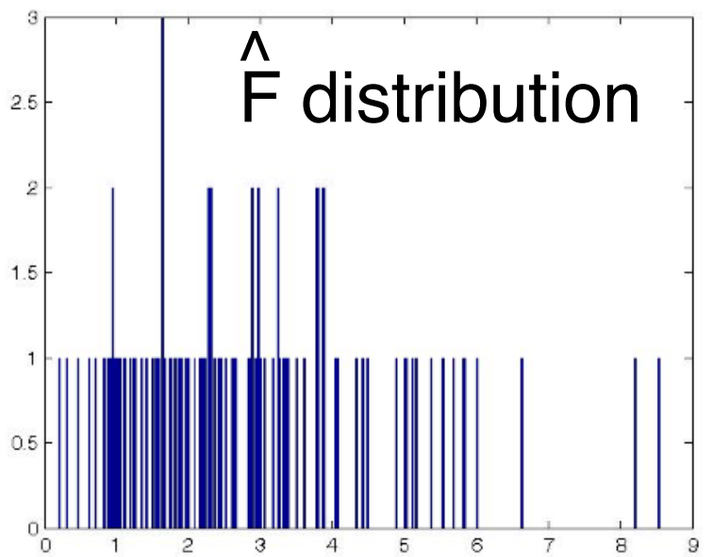


α=3 β=1 a parameter θ on F: θ=mean/median

bootstrap sampling

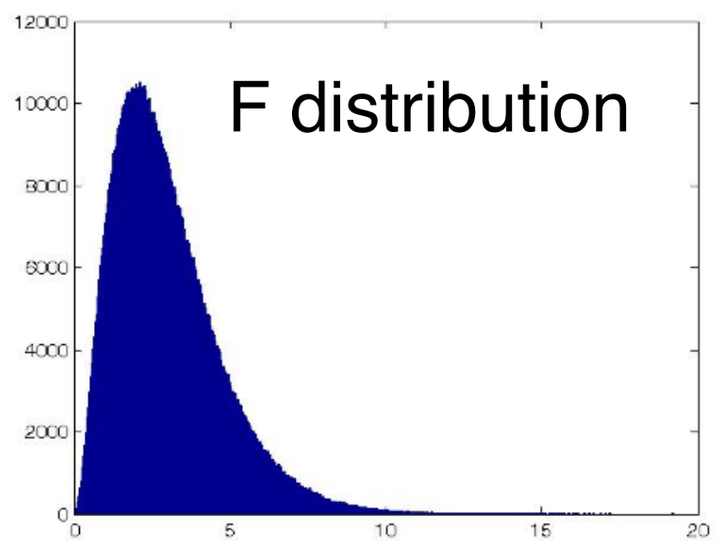
Let's try a simple example where we can see the "hidden" side of things, too.

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):

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

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

How accurate is this?

$\hat{\theta}$ statistic

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

→ θ parameter

statistics reviewed

Statistics or estimates

Definition

A statistic (also called estimates, estimators) $\hat{\theta}$ is a function of $\hat{F}$ or the sample $\mathbf{x}$, e.g.:

$$\hat{\theta} = t(\hat{F})$$

or also written $\hat{\theta} = s(\mathbf{x})$.

if $\hat{\theta}$ is the mean:

$$\begin{aligned} \hat{\theta} = t(\hat{F}) &= \int_{-\infty}^{+\infty} x\ \hat{F}(x)dx \\ &= \int_{-\infty}^{+\infty} x\ \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\ dx \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= s(\mathbf{x}) = \overline{x} \end{aligned}$$

statistics reviewed

Statistics or estimates

if θ is the variance:

$$\begin{aligned}\hat{\theta} &= \int_{-\infty}^{+\infty}(x-\bar{x})^2\ \hat{F}(x)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \hat{\sigma}^2\end{aligned}$$

statistics reviewed

The Plug-in principle

Definition

The Plug-in estimate of a parameter $\theta = t(F)$ is defined to be:

$$\hat{\theta} = t(\hat{F}).$$

The function $\theta = t(F)$ of the probability density function F is estimated by the same function $t(.)$ of the empirical density $\hat{F}$.

- $\overline{x}$ is the plug-in estimate of μ_F.
- $\hat{\sigma}$ is the plug-in estimate of σ_F

bootstrap review and bias

Bootstrap samples and replications

Definition

A bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $\mathbf{x} = (x_1, x_2, \cdots, x_n)$.

Considering a sample $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, some bootstrap samples can be:

$$\mathbf{x}^{*(1)} = (x_2, x_3, x_5, x_4, x_5)$$
$$\mathbf{x}^{*(2)} = (x_1, x_3, x_1, x_4, x_5)$$
$$\text{etc.}$$

Definition

With each bootstrap sample $\mathbf{x}^{*(1)}$ to $\mathbf{x}^{*(B)}$, we can compute a bootstrap replication $\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)})$ using the plug-in principle.

bootstrap review and bias

How to compute Bootstrap samples

Repeat B times:

1. A random number device selects integers $i_1, \cdots, i_n$ each of which equals any value between 1 and n with probability $\frac{1}{n}$.
2. Then compute $\mathbf{x}^* = (x_{i_1}, \cdots, x_{i_n})$.

Some matlab code available on the web

See BOOTSTRAP MATLAB TOOLBOX, by Abdelhak M. Zoubir and D. Robert Iskander,
http://www.csp.curtin.edu.au/downloads/bootstrap_toolbox.html

bootstrap review and bias

How many values are left out of a bootstrap resample ?

Given a sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and assuming that all x_i are different, the probability that a particular value x_i is left out of a resample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is:

$$\mathcal{P}(x_j^* \neq x_i, 1 \leq j \leq n) = \left(1 - \frac{1}{n}\right)^n$$

since $\mathcal{P}(x_j^* = x_i) = \frac{1}{n}$. When n is large, the probability $\left(1 - \frac{1}{n}\right)^n$ converges to $e^{-1} \approx 0.37$.

bootstrap review and bias

The Bootstrap algorithm for Estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Estimate the standard error $\mathrm{se}_F(\hat{\theta})$ by the standard deviation of the B replications:

$$\hat{\mathrm{se}}_B = \left[\frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right]^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$

bootstrap review and bias

Bootstrap estimate of the standard Error

Example A

From the distribution F: $F(x) = 0.2\,\mathcal{N}(\mu{=}1,\sigma{=}2) + 0.8\,\mathcal{N}(\mu{=}6,\sigma{=}1)$. We draw the sample $\mathbf{x} = (x_1, \cdots, x_{100})$:

$$\mathbf{x} = \begin{pmatrix}
7.0411 & 4.8397 & 5.3156 & 6.7719 & 7.0616 \\
5.2546 & 7.3937 & 4.3376 & 4.4010 & 5.1724 \\
7.4199 & 5.3677 & 6.7028 & 6.2003 & 7.5707 \\
4.1230 & 3.8914 & 5.2323 & 5.5942 & 7.1479 \\
3.6790 & 0.3509 & 1.4197 & 1.7585 & 2.4476 \\
-3.8635 & 2.5731 & -0.7367 & 0.5627 & 1.6379 \\
-0.1864 & 2.7004 & 2.1487 & 2.3513 & 1.4833 \\
-1.0138 & 4.9794 & 0.1518 & 2.8683 & 1.6269 \\
6.9523 & 5.3073 & 4.7191 & 5.4374 & 4.6108 \\
6.5975 & 6.3495 & 7.2762 & 5.9453 & 4.6993 \\
6.1559 & 5.8950 & 5.7591 & 5.2173 & 4.9980 \\
4.5010 & 4.7860 & 5.4382 & 4.8893 & 7.2940 \\
5.5741 & 5.5139 & 5.8869 & 7.2756 & 5.8449 \\
6.6439 & 4.5224 & 5.5028 & 4.5672 & 5.8718 \\
6.0919 & 7.1912 & 6.4181 & 7.2248 & 8.4153 \\
7.3199 & 5.1305 & 6.8719 & 5.2686 & 5.8055 \\
5.3602 & 6.4120 & 6.0721 & 5.2740 & 7.2329 \\
7.0912 & 7.0766 & 5.9750 & 6.6091 & 7.2135 \\
4.9585 & 5.9042 & 5.9273 & 6.5762 & 5.3702 \\
4.7654 & 6.4668 & 6.1983 & 4.3450 & 5.3261
\end{pmatrix}$$

We have $\mu_F = 5$ and $\overline{x} = 4.9970$.

bootstrap review and bias

Bootstrap estimate of the standard Error

Example A

1. $B = 1000$ bootstrap samples $\{\mathbf{x}^{*(b)}\}$
2. $B = 1000$ replications $\{\overline{x}^*(b)\}$
3. Bootstrap estimate of the standard error:

$$\widehat{\text{se}}_{B=1000} = \left[\frac{\sum_{b=1}^{1000}[\overline{x}^*(b) - \overline{x}^*(\cdot)]^2}{1000-1}\right]^{\frac{1}{2}} = 0.2212$$

where $\overline{x}^*(\cdot) = 5.0007$. This is to compare with $\hat{se}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = 0.22$.

bootstrap review and bias

Distribution of $\hat{\theta}$

When enough bootstrap resamples have been generated, not only the standard error but any aspect of the distribution of the estimator $\hat{\theta} = t(\hat{F})$ could be estimated. One can draw a histogram of the distribution of $\hat{\theta}$ by using the observed $\hat{\theta}^*(b),\ b = 1, \cdots, B$.

Example A

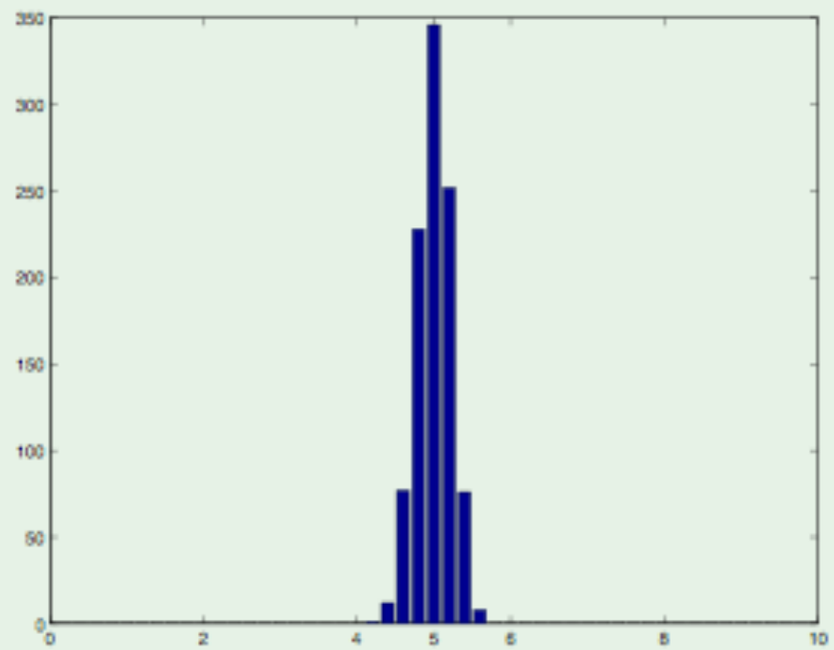


Figure: Histogram of the replications $\{\overline{x}^*(b)\}_{b=1\cdots B}$.

bootstrap review and bias

Bootstrap estimate of the standard error

Definition

The ideal bootstrap estimate $\mathrm{se}_{\hat{F}}(\theta^*)$ is defined as:

$$\lim_{B \to \infty} \hat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\theta^*)$$

$\mathrm{se}_{\hat{F}}(\theta^*)$ is called a non-parametric bootstrap estimate of the standard error.

bootstrap review and bias

Bootstrap estimate of the standard Error

How many B in practice ?

you may want to limit the computation time. In practice, you get a good estimation of the standard error for B in between 50 and 200.

Example A

B	10	20	50	100	500	1000	10000
$\widehat{\text{se}}_B$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187

Table: Bootstrap standard error w.r.t. the number B of bootstrap samples.

bootstrap review and bias

Bootstrap estimate of bias

Definition

The bootstrap estimate of bias is defined to be the estimate:

$$\begin{aligned}\text{Bias}_{\hat{F}}(\hat{\theta}) &= \mathbb{E}_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F}) \\ &= \theta^*(\cdot) - \hat{\theta}\end{aligned}$$

Example A

B	10	20	50	100	500	1000	10000
$\mathbb{E}_{\hat{F}}(\overline{x}^*)$	5.0587	4.9551	5.0244	4.9883	4.9945	5.0035	4.9996
$\widehat{\text{Bias}}$	0.0617	-0.0419	0.0274	-0.0087	-0.0025	0.0064	0.0025

Table: $\widehat{\text{Bias}}$ of $\overline{x}^*$ ($\overline{x} = 4.997$ and $\mu_F = 5$).

bootstrap review and bias

Bootstrap estimate of bias

1. B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Approximate the bootstrap expectation :

$$\hat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b) = \frac{1}{B}\sum_{b=1}^{B}s(\mathbf{x}^{*(b)})$$

4. the bootstrap estimate of bias based on B replications is:

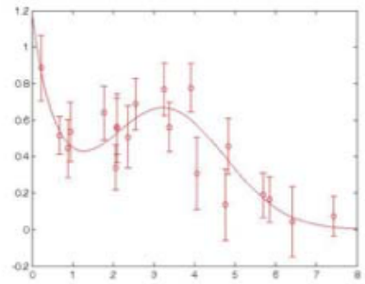
$$\widehat{\mathrm{Bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$$

bootstrap sampling

```
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))
```

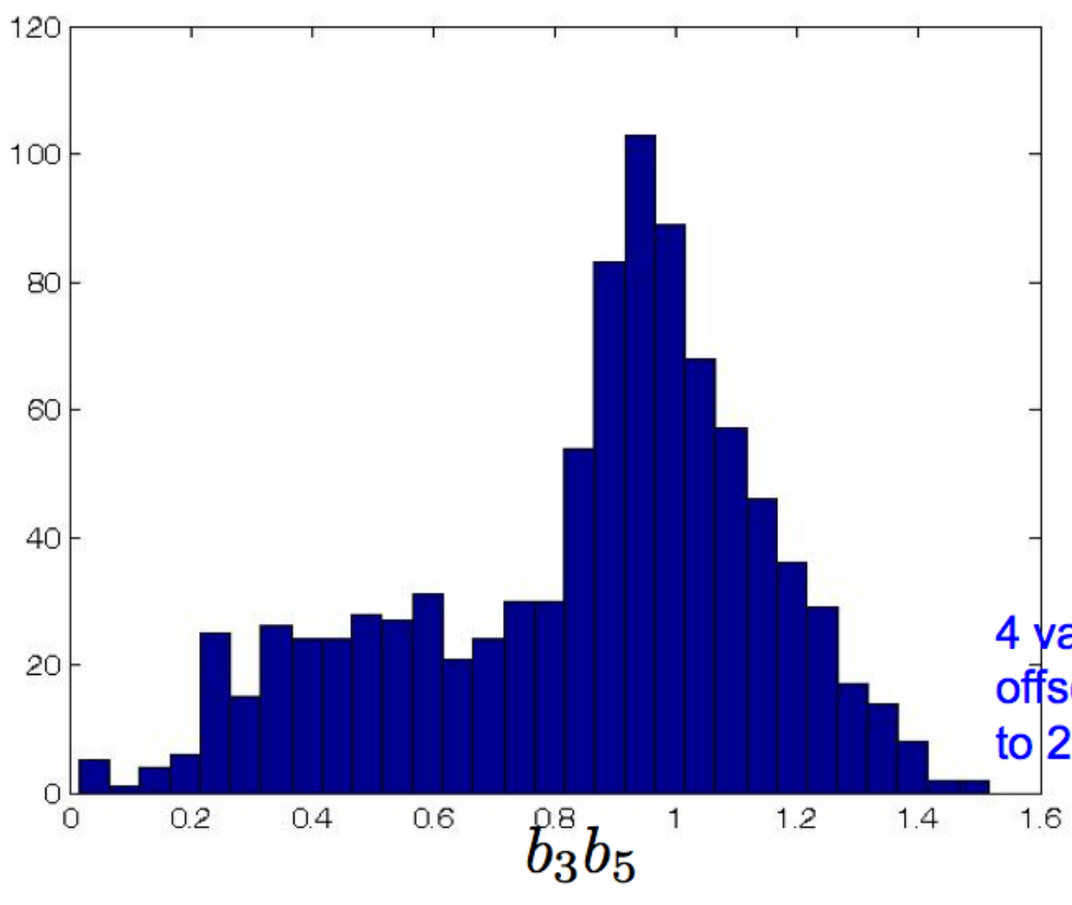
0.2924

new sample of integers in 1:ndata, with replaceme



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

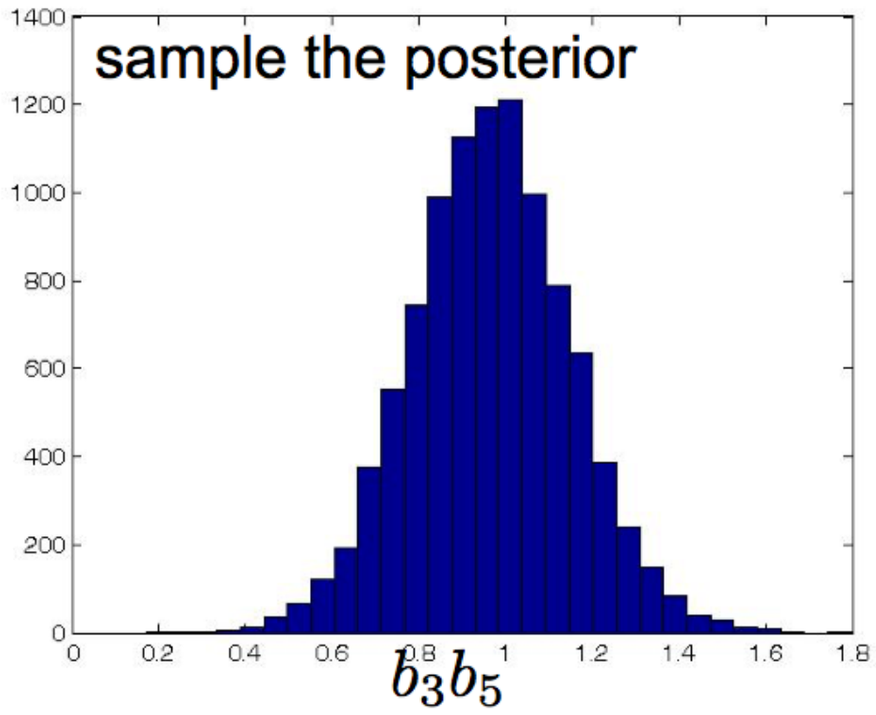
So we get the peak around 1, as before, but a much broader distribution.

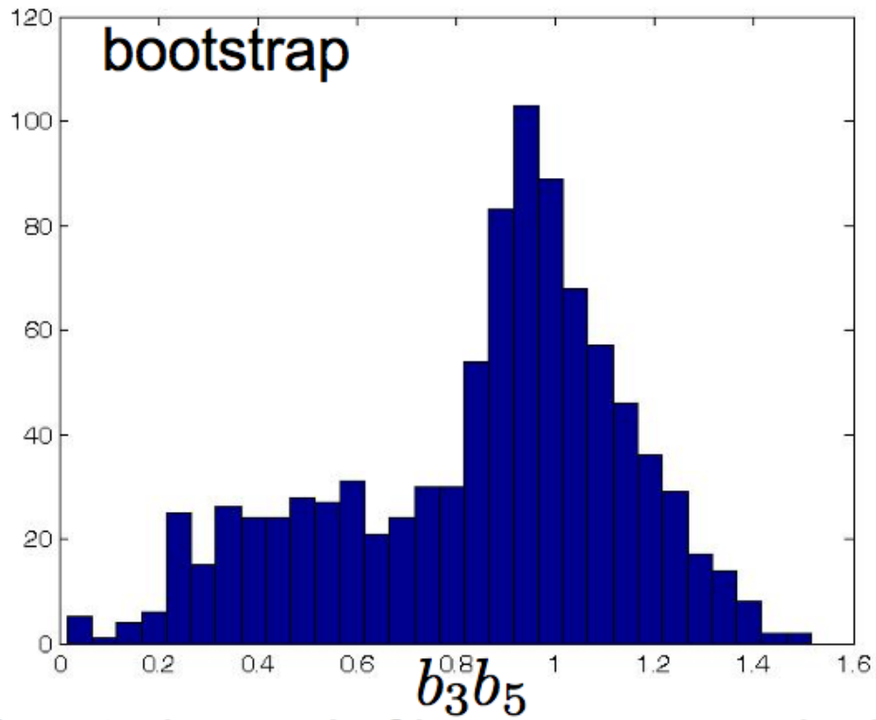


4 values offscale (up to 27!)

bootstrap sampling

We previously compared bootstrap-from-sample to bootstrap-from-population.
More relevant, let's compare boostrap-from-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 would 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!

bootstrap sampling

Note that sampling the posterior "honors" the stated measurement errors. Bootstrap doesn't. That can be good!

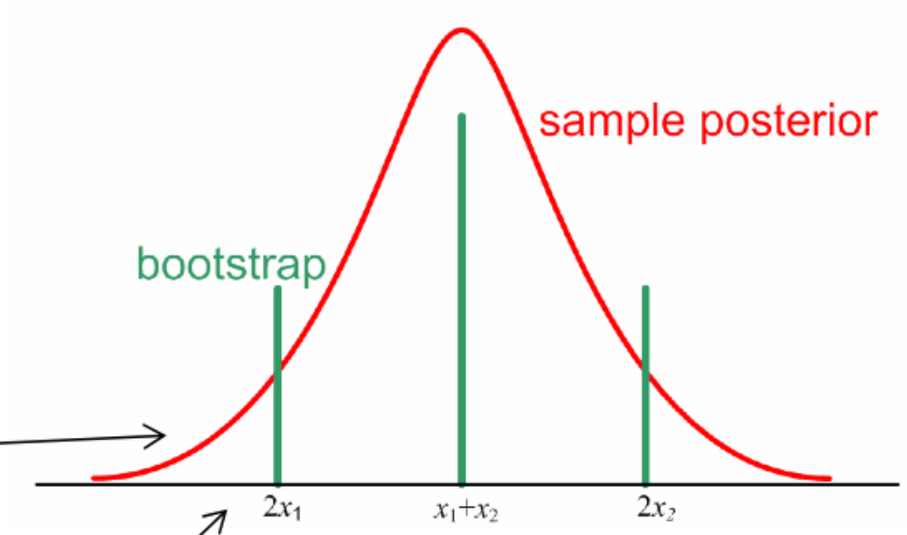
Suppose (very toy example) the "statistic" is

$$s = x_1 + x_2$$

then the posterior probability is

$$P(s) \propto \exp\left[-\frac{1}{2}\frac{(s - x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2}\right]$$

Note that this depends on the σ's!



The bootstrap (here noticeably discrete) doesn't depend on the σ's. In some sense it estimates them, too.

So, if the errors were badly underestimated, sampling the posterior would give too small an uncertainty, while bootstrap would still give a valid estimate.

If the errors are right, both estimates are valid. Notice that the model need not be correct. Both procedures give estimates of the statistical uncertainty of parameters of even a wrong (badly fitting) model. *But for a wrong model, your interpretation of the parameters may not mean anything!*