BIOL309: The Jackknife & Bootstrap

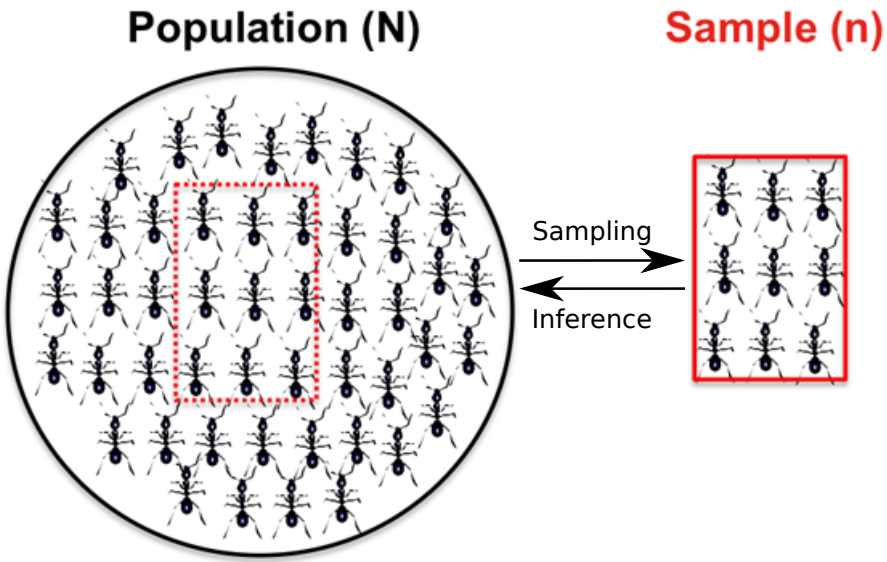
Paul Gardner

September 25, 2017

What is Resampling?

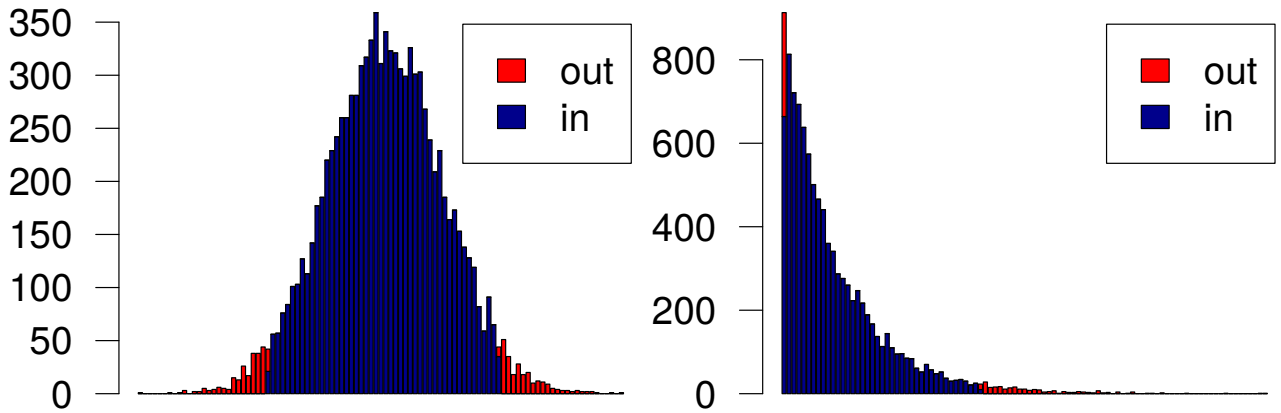
"I do not believe in any statistical test unless I can prove it with a permutation test." – R.A. Fisher

- Resampling is a statistical technique in which multiple new samples are drawn from a sample or from the population
- Statistics of interest (e.g. sample median) are calculated for each new sample. The distribution of new statistics can be analysed to investigate different properties (e.g., confidence intervals, the error, the bias) of the statistics.



First, some definitions & reminders

- Mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Standard deviation $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
- Standard error $SE_{\bar{x}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}}$
- Bias of an estimator is the difference between the estimators expected value and the true value of the parameter being estimated
- Confidence interval



Jackknifing

- a resampling technique especially useful for finding standard error, variance and bias of estimators
- the jackknife is a small, handy tool
- also called leave-one-out (LOO)
- This approach tests that some outlier datapoint is not having a disproportionate influence on the outcome.

Jackknifing

- The jackknife deletes each observation and calculates an estimate based on the remaining $n - 1$ values
- It uses this collection of estimates to do things like estimate the bias and the standard error

Jackknifing: definition

- Let $x_1, \ldots x_n$ be a dataset
- θ is a paramater you want to estimate from the data (e.g. mean, median, standard deviation, ...)
- Let $\hat{\theta}$ be the estimate based upon the **entire dataset**
- Let $\hat{\theta}_i$ be the estimate of θ obtained by **deleting observation** x_i
- Let $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i$
  - Sometimes $\bar{\theta}$ is written $\bar{\theta}_{(.)}$

Jackknifing: estimating bias of a method, and correcting it

- This provides an estimated correction of bias due to the estimation method. **The jackknife does not correct for a biased sample.** (Wikipedia/Jackknife_resampling)
- The jackknife estimate of bias is $B = (n-1)(\bar{\theta} - \hat{\theta})$
  - In other words, is the difference between the actual and the average of the delete-one estimates.
- We can then correct $\hat{\theta}$ (the estimator on the entire dataset), using:
  - $\hat{\theta}_{corrected} = \hat{\theta} - B$
- With the magic of algebra:
  - $\hat{\theta}_{corrected} = n\hat{\theta} - (n-1)\bar{\theta}$

Jackknifing

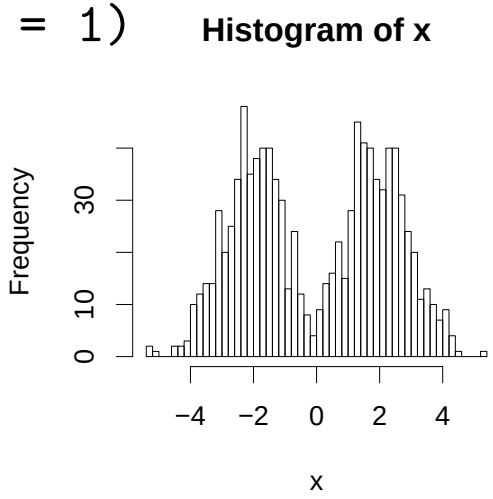
- The jackknife estimate of the standard error is:

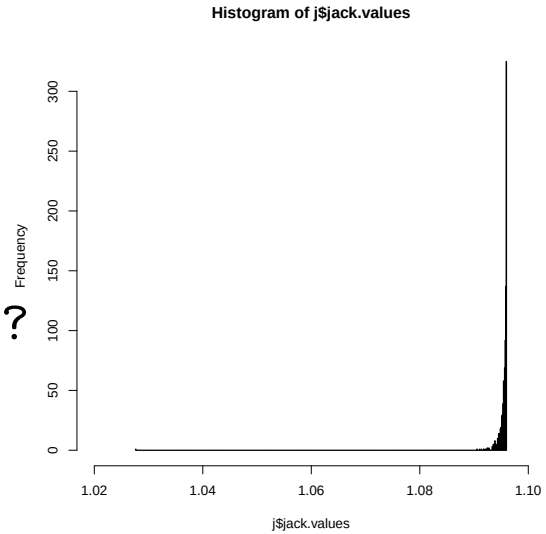
$$SE_{JK}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}$$

- This simplifies to the standard error ($SE_{\bar{x}} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n(n-1)}}$) when θ is the mean

Example

```
x1 <- rnorm(1000, mean = 2, sd = 1)
x <- c(x1,-10)
hist(x,breaks=500)

library(bootstrap)
#define theta function
theta <- function(x){sd(x)}
j <- jackknife(x,theta)

mean(j$jack.values)
#check j$jack.values is normal
hist(j$jack.values,breaks=500)

#What is the bias corrected sd?
```





CORRECTION: testing bias corrected values...

The lab example...

```
for (i in c(10,100, 1000) ){
    for (j in c(-100,-10,-1,1, 10,100) ){
        x <- c(rnorm(i, mean = 2, sd = 1),j)
        jk <- jackknife(x,sd)
        corr <- sd(x) - jk$jack.bias
        cat(paste(round(corr, digits = 2), "\t"))
      }
      cat("\n")
}
```

Jackknife bias corrected values ($i = N$, j =outlier, expected= 1)

i\j	-100	-10	-1	1	10	100
10	44.29	4.76	1.31	0.65	2.96	42.25
100	14.28	1.67	0.94	0.96	1.34	13.64
1000	4.21	1.07	0.99	1.01	1.05	4.03

Example: more bad...

The jackknife estimate of variance is slightly biased upward!

THE JACKKNIFE ESTIMATE OF VARIANCE

by

B. Efron and C. Stein
Stanford University

Abstract

Tukey's jackknife estimate of variance for a statistic $S(X_1, X_2, \ldots, X_n)$ which is a symmetric function of i.i.d. random variables X_i, is investigated using an ANOVA-like decomposition of S. It is shown that the jackknife variance estimate tends always to be biased upwards, a theorem to this effect being proved for the natural jackknife estimate of $\mathrm{Var}\, S(X_1, X_2, \ldots, X_{n-1})$ based on $X_1, X_2, \ldots, X_n$.

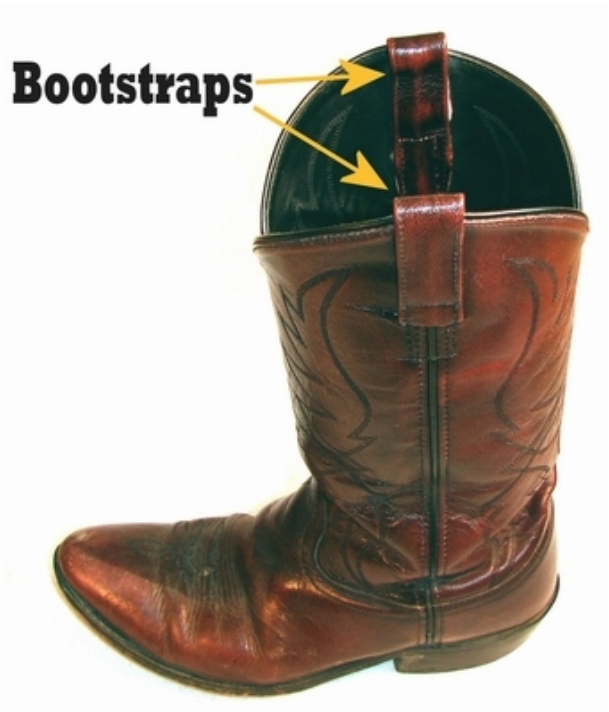
Efron & Stein (1981) The jackknife estimate of variance. *The Annals of Statistics*, pp. 586-596

Jackknifing issues

- When the estimator is not normally distributed jackknifing may fail
- May be unreliable on a small number of datasets
- This provides an estimated correction of bias due to the estimation method. **The jackknife does not correct for a biased sample.** (Wikipedia/Jackknife_resampling)
- Not great when θ is the standard deviation!

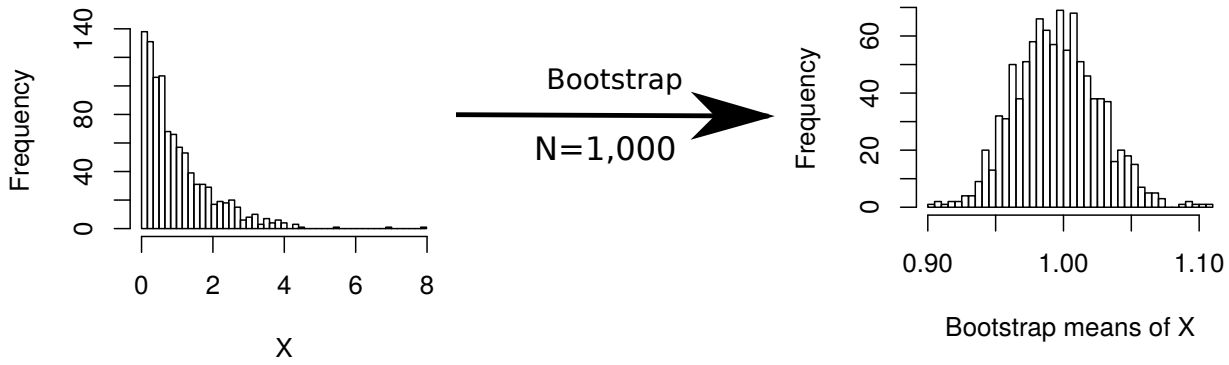
What is bootstrapping?

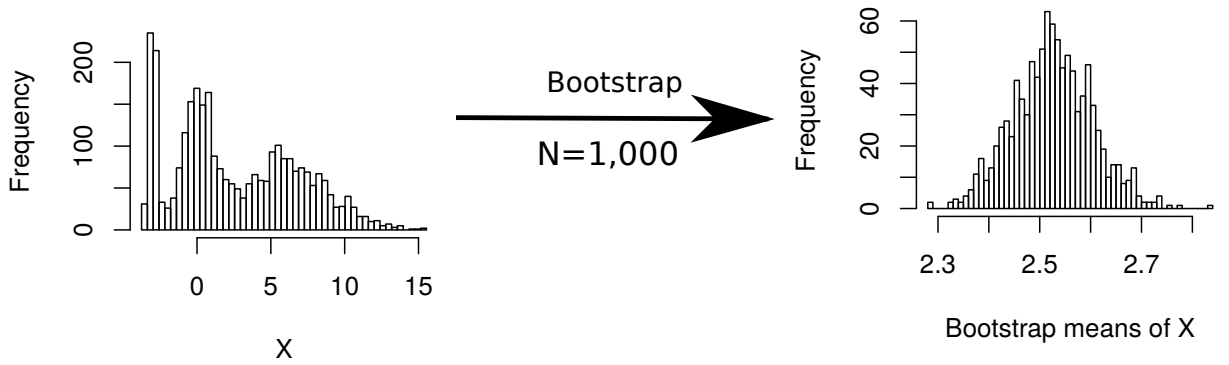
- Bootstrapping is a useful means for assessing the reliability of your data (e.g. confidence intervals, bias, variance, prediction error, ...).
- It refers to any metric that relies on **random sampling with replacement**.
- Used to estimate SE, confidence intervals, and test for significance



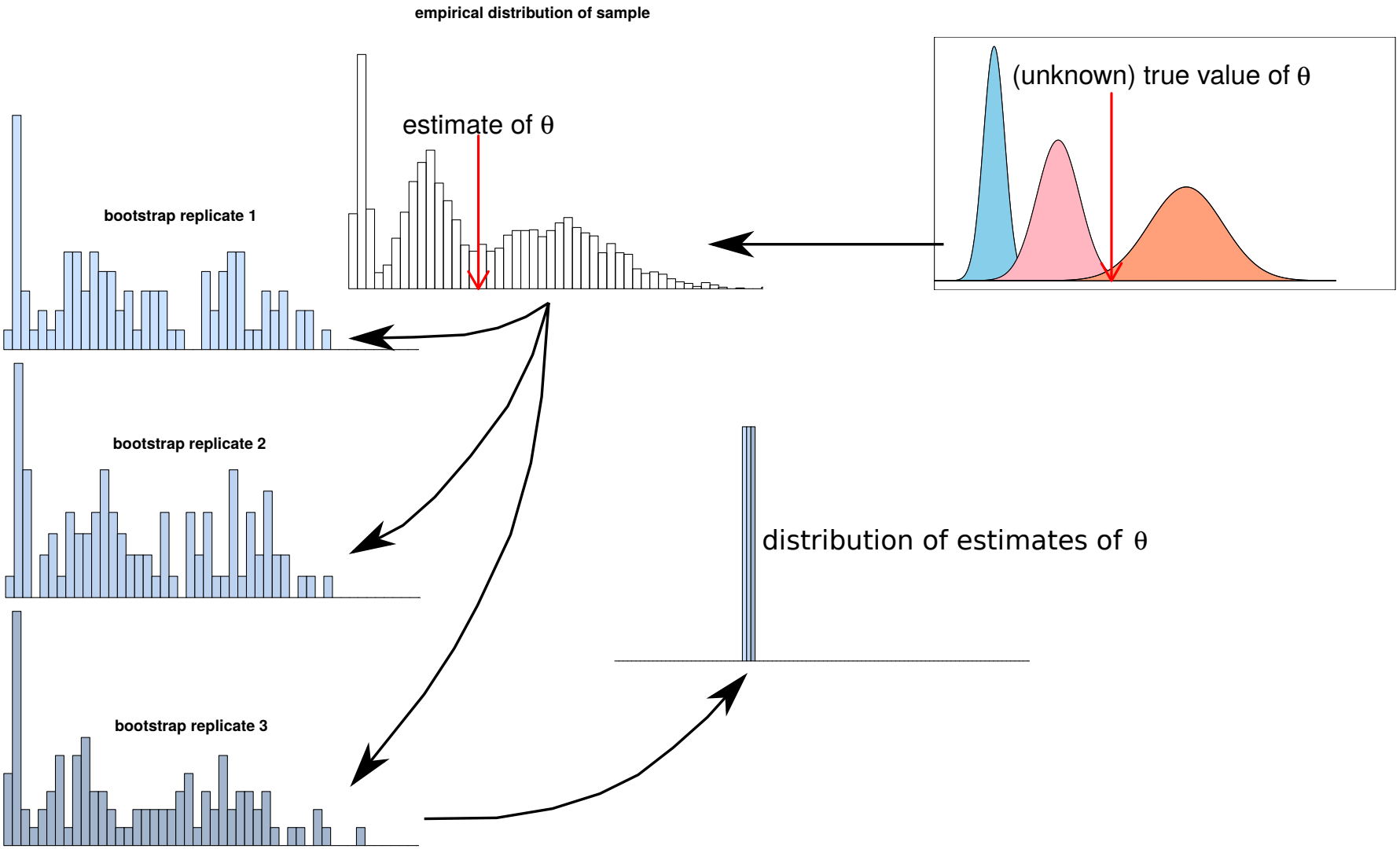
First, a definition

- Central limit theorem:
- the means from a large number of independent random samples will be approximately normally distributed, regardless of the underlying distribution



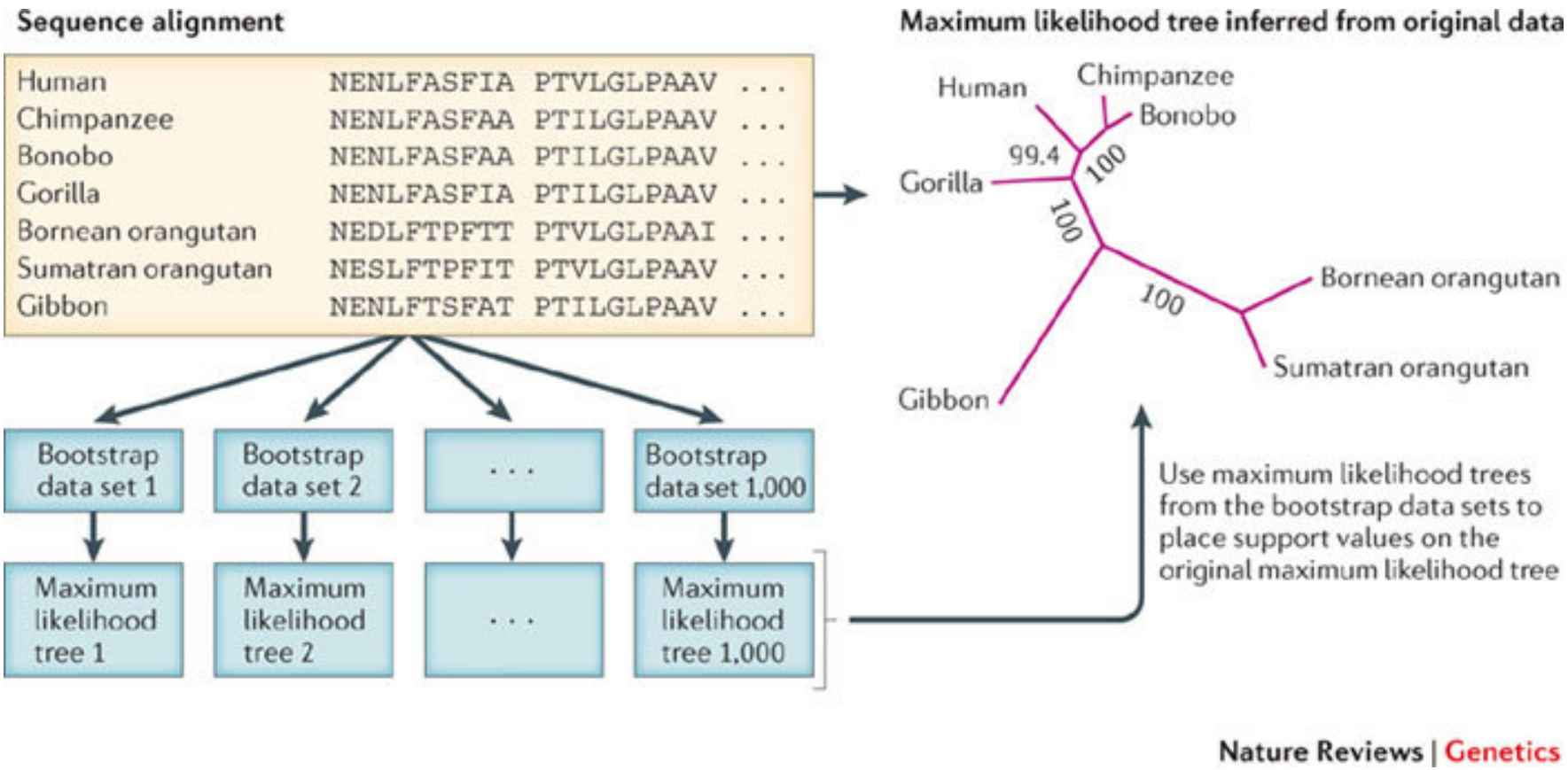


Bootstrapping illustrated



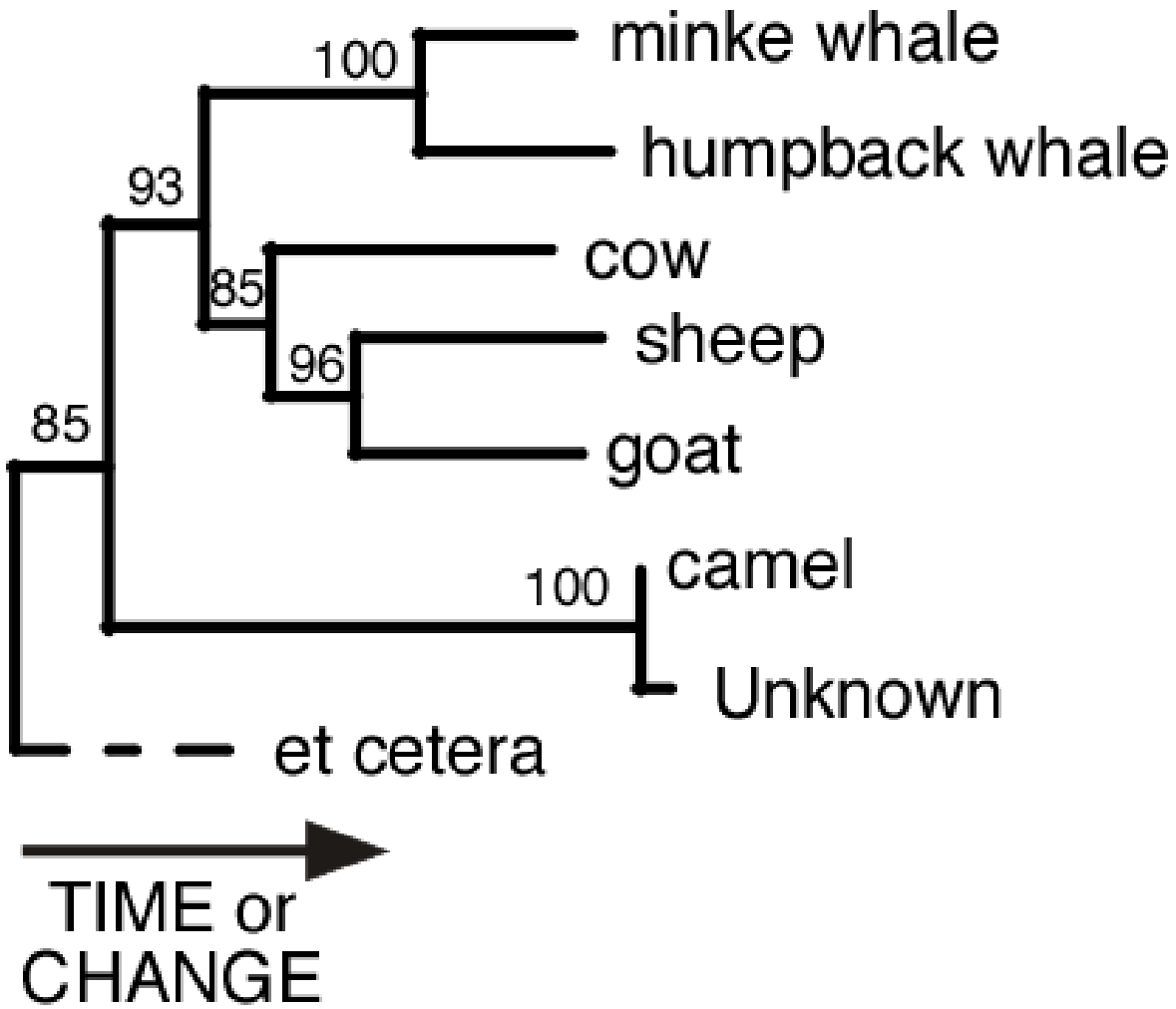
- Bootstrap sampling from a distribution (a mixture of 3 normal distributions) to estimate the variance of the mean

Bootstrapping is used a lot in phylogenetics



Yang & Rannala (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*.

Application: DNA surveillance



http://dna-surveillance.fos.auckland.ac.nz/

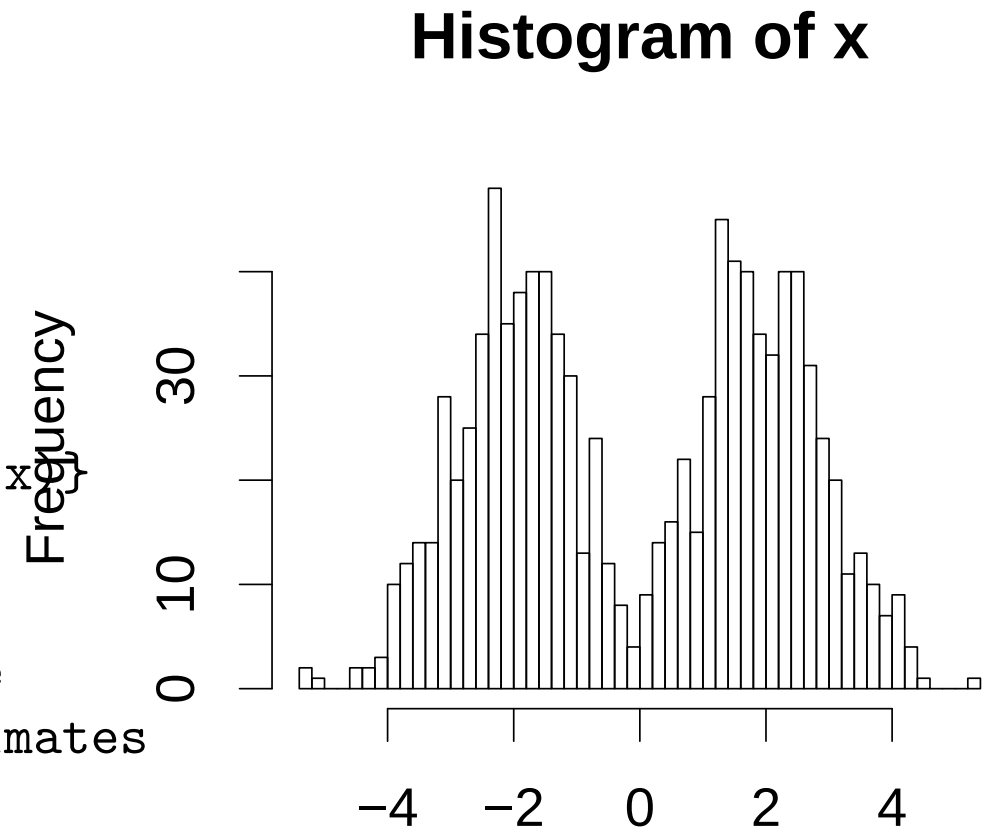
Bootstrap sampling

To infer the error in a quantity, θ, estimated from a dataset $x_1, x_2, \ldots x_N$ we do the following R times (e.g. $R = 1,000$):

1. Draw a "bootstrap sample" by sampling n times with replacement from the sample. Call these $X_1^*, X_2^*, \ldots X_n^*$. Note that some points are represented more than once in the bootstrap samples, some once, some not at all.
2. Estimate θ from the bootstrap sample, call this $\hat{\theta}_k^*$ $(k = 1, 2, \ldots R)$.
3. When all R bootstrap samples have been done, the distribution of $\hat{\theta}_k^*$ estimates the distribution one would get if one were able to draw repeated samples of n points from the unknown true distribution.

Example: confidence intervals for the median

```
x1=rnorm(500, mean =  2, sd = 1)
x2=rnorm(500, mean = -2, sd = 1)
x=c(x1,x2)
hist(x,breaks=50)

summary(x)

library(bootstrap)
#define theta function
theta = function(x){median(x)}
bs = bootstrap(x,50,theta)
summary(bs$thetastar)
#What is the 50% confidence
#interval for boostrap estimates
#of median?

boott(x,theta,nboott=1000,perc=c(0.025,0.975))
```



Example: regression (I)

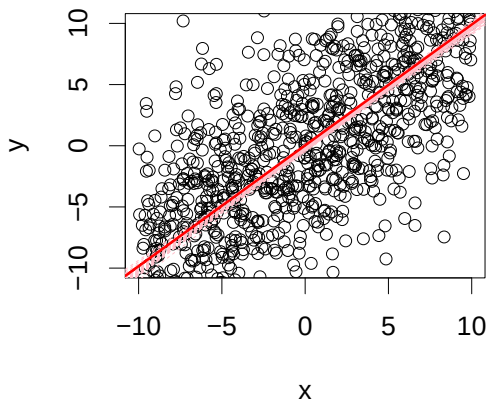
```
#create a simulated dataset, sampling from a normal distribution
x<-runif(1000,-10,10)

#generate a y dataset with a little noise:
#y = m * x + c
y<-rnorm(length(x),1,0.1)*x + rnorm(length(x),mean=0,sd=1)

#plot a regression
reg1<-lm(y ~ x)
plot(x,y,type="p")
abline(reg1,col="red",lwd=3)
```

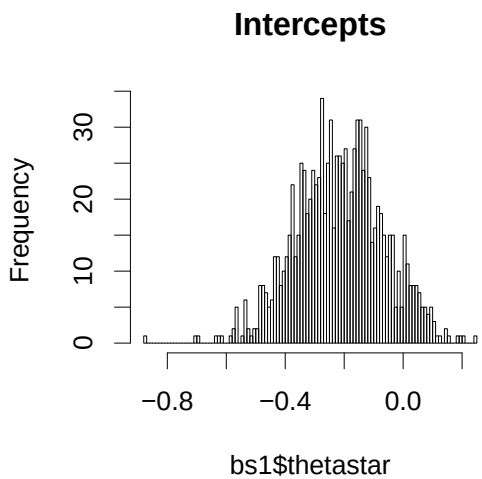
Example: regression (II)

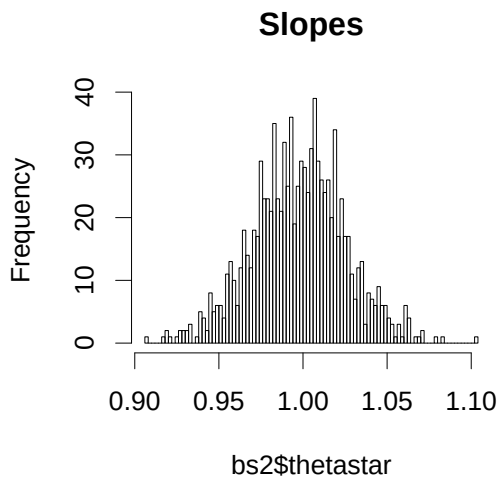
```
library(bootstrap)

#column bind x & y
xdata <- cbind(x,y)

#create functions, theta1 & theta2,
#1 returns the intercept, 2 returns the slope
theta1 <- function(i,xdata){
      coef(lm(xdata[i,2] ~ xdata[i,1]))[1]
      }
theta2 <- function(i,xdata){
      coef(lm(xdata[i,2] ~ xdata[i,1]))[2]
      }

#bootstrap!
bs1=bootstrap(1:length(x),1000,theta1,xdata)
bs2=bootstrap(1:length(x),1000,theta2,xdata)

quantile(bs2$thetastar,probs = c(0.025,0.975))
```

Example: regression (III)

```
#plot the resulting lines:
for (i in 1:length(bs1$thetastar)){
   abline(bs1$thetastar[i],bs2$thetastar[i], lty=2,col="pink")
}
abline(reg1,col="red",lwd=3)
hist(bs1$thetastar,breaks=100,main="Intercepts")
hist(bs2$thetastar,breaks=100,main="Slopes")
```







Bootstrap issues

- Need a large number of bootstrap samples (e.g. $R \geq 1000$). The larger the number, the better the estimates.
- If θ is hard to calculate (e.g. tree building) then bootstrapping can be very computationally intensive.