BIOSTAT 651
Notes #15: Bootstrap Methods

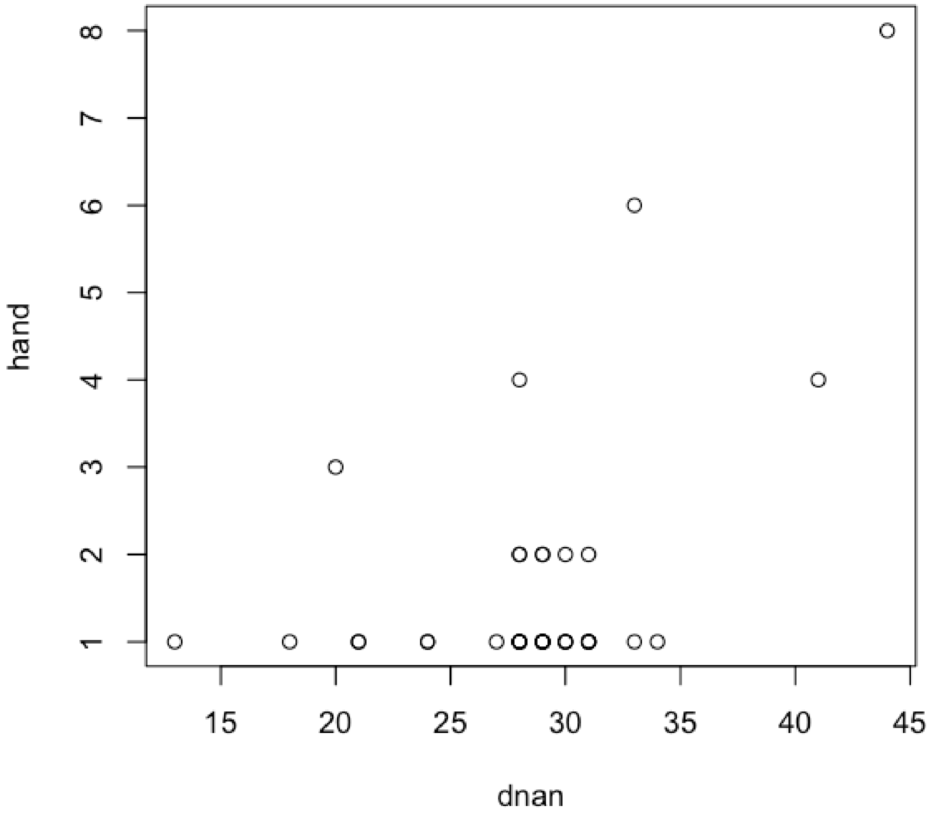
- Lecture Topics:
  - ○ Basic idea
  - ○ Hypothesis test
  - ○ Regression
- Based on following book and slides.
  - – Davison and Hinkley (1997) *Bootstrap Methods and their Application*. Cambridge University Press
  - – Davison (2006) *Bootstrap Methods and their Application*. Short course slides

Bootstrap

- Bootstrap: simulation methods to estimate sampling distribution of almost any statistics.
- Developed by Bradley Efron in ”Bootstrap methods: another look at the jackknife” (1979)
- Useful when
  - Standard assumptions are not valid (small n, invalid regression assumptions, etc)
  - Complex problem with no (reliable) theory ex. theoretical distribution of a statistic of interest is complicated or unknown
  - or (almost) anywhere else.

Example: Handedness data

- Investigate relationship between genetic measurement (dnan) and left-handedness (hand).



Example: Handedness data

- Question: Is there any dependence between dnan and hand for these $n = 37$ individuals?
- Sample coefficient $\widehat{\theta} = 0.509$
- Confidence interval (from the bivariate normal model): $CI(0.221, 0.715)$
- Issues?

Bootstrap

- Estimate distribution of $\widehat{\theta}$ using resampling
- Statistical model: data $(y_1, \ldots, y_n) \sim F$, unknown
  - Handedness data: $y =$ (dnan, hand)
  - θ: correlation coefficient
- For $r = 1, \ldots, R$
  - resample y with replacement: $y_r^* = (y_{1r}^*, \ldots, y_{nr}^*)$
  - Compute bootstrap $\widehat{\theta}_r^*$ using y_r^*
- Repeat R times !

$$\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_R^*$$

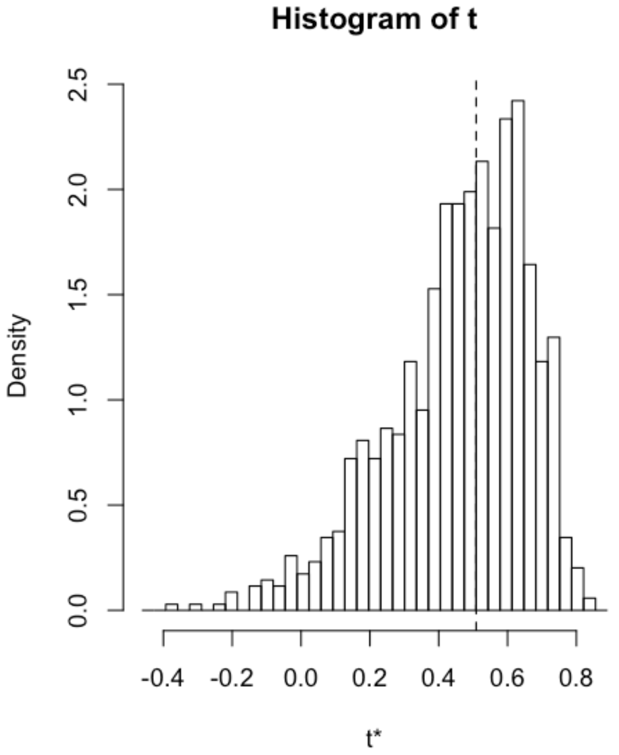
Bootstrap

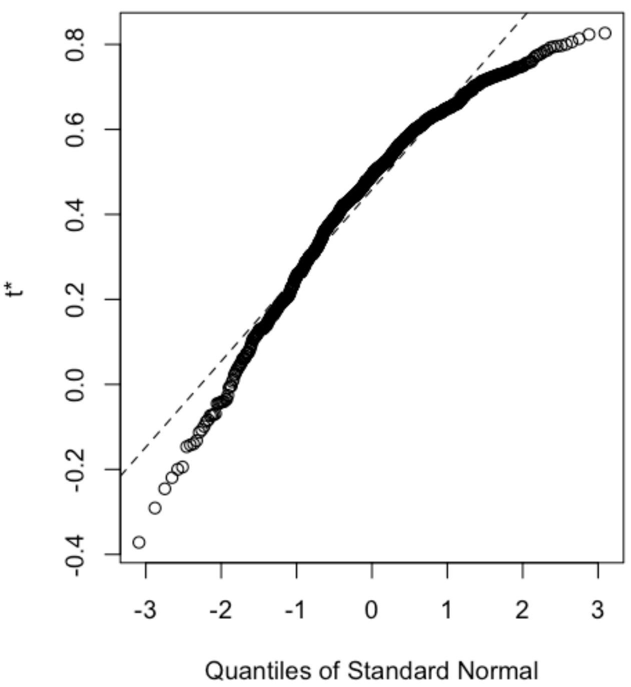
- R code (boot package)

```
library(boot)
data(claridge)

Corr <- function(d, f){
cor(d[f,1], d[f,2])
}
boot.out<-boot(claridge, Corr, R=1000)


plot(boot.out)
boot.ci(boot.out)
```

Bootstrap

- Distribution of $t = \widehat{\theta}_r^*$ (R=1000)
- CI: $(0.0912, 0.8071)$ (BCa)





Why does Bootstrap work

- Statistical model: data $(y_1, \ldots, y_n) \sim F$
- Estimate distribution of $\widehat{\theta}$
  - Key issue: what is the variability of $\widehat{\theta}$ when samples are repeatedly taken from F?
- Suppose F is known - we can answer the previous question by
  - Analytical calculation
  - Simulation

Why does Bootstrap work

- Assume F is known
- For $r = 1, \ldots, R$
  - generate random sample $y_r^* = (y_{1r}^*, \ldots, y_{nr}^*) \sim F$
  - Compute $\widehat{\theta}_r^*$ using y_r^*
- Use $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_R^*$ to estimate sampling distribution of $\widehat{\theta}$
- If $R \to \infty$, Monte Carlo error would disappear.

Why does Bootstrap work

- But we don't know F!
  - Estimate F using the empirical distribution function $\widehat{F}_n$
  - Generate random samples from $\widehat{F}_n$
- For $r = 1, \ldots, R$
  - generate random sample $y_r^* = (y_{1r}^*, \ldots, y_{nr}^*) \sim \widehat{F}_n$
  - Compute $\widehat{\theta}_r^*$ using y_r^*
- Bootstrap (re)samples are iid samples from $\widehat{F}_n$

Bootstrap estimators

- Variance

$$Var_B(\widehat{\theta}) = \frac{1}{R-1}\sum_{i=1}^{R}(\widehat{\theta}_r^* - \widehat{\theta}_{(.)}^*)^2$$

$$\widehat{\theta}_{(.)}^* = \frac{1}{R}\sum_{i=1}^{R}\widehat{\theta}_r^*$$

- Bias
  - Bias: $E(\widehat{\theta}) - \theta$
  - Bootstrap estimator of Bias:

$$Bias_B = \frac{1}{R}\sum_{i=1}^{R}\widehat{\theta}_r^* - \widehat{\theta}$$

- Bias corrected estimator

$$\widehat{\theta}_{BC} = \widehat{\theta} - Bias_B$$

Bootstrap Confidence Intervals

- There are several versions of CI
- Normal confidence internval
  - If $\widehat{\theta}$ approximately normal, then $\widehat{\theta} \sim N(\theta + \beta, v)$, where β is a bias
  - With known β and v, $(1 - 2\alpha)$ CI is

$$\theta - \beta \pm Z_\alpha v^{1/2}$$

  - Replace β and v to their Bootstrap estimates.
- Percentile interval
  - Estimate CI nonparametically
  - Use α and $1 - \alpha$ quantiles of bootstrap samples to estimate CI

$$\widehat{\theta}^*_{(R+1)\alpha}, \quad \widehat{\theta}^*_{(R+1)(1-\alpha)}$$

- Studentized-t (Percentile-t) Bootstrap Confidence Interval
  - Generalize Student-t statistic to bootstrap setting
  - Require variance formula V for $\widehat{\theta}$ computed from $(y_1, \cdots, y_n)$
  - R bootstrap copies of $(\widehat{\theta}^*_r, \widehat{V}_r^{*1/2})$

$$T_1^* = \frac{(\widehat{\theta}^*_1 - \widehat{\theta})}{\widehat{V}_1^{*1/2}}, \ldots, T_r^* = \frac{(\widehat{\theta}^*_r - \widehat{\theta})}{\widehat{V}_r^{*1/2}}$$

  - CI

$$\widehat{\theta} - \widehat{V}^{1/2} T^*_{(R+1)\alpha}, \quad \widehat{\theta} - \widehat{V}^{1/2} T^*_{(R+1)(1-\alpha)}$$

- Bias corrected, accelerated (BCa) percentile interval
  - Shift and scale the percentile bootstrap confidence interval to compensate for bias
  - Replace percentile interval with

$$\widehat{\theta}^*_{(R+1)\alpha_1}, \quad \widehat{\theta}^*_{(R+1)(1-\alpha_2)}$$

    where α_1 and α_2 were chosen to improve CI.

Handedness data

- Bias: -0.0401
- SE: 0.208
- CI:
  - Normal $(0.1631, 0.9551)$
  - Percentile $(-0.0402, 0.7465)$
  - BCa $(0.0912, 0.8071)$

Hypothesis test

- Testing problem
  - data $(y_1, \ldots, y_n)$
  - Model M_0 to be tested
  - test statistic $T_{obs} = T(y_1, \ldots, y_n)$, with large values giving evidence against H_0
- P-value: $Pr(T \geq T_{obs} | M_0)$
  - Small p-value indicates evidence against M_0
- Issue: P-values are often hard to calculate

Hypothesis test

- Estimate P-values by simulating from the fitted null hypothesis model $\widehat{M}_0$
- For $r = 1, \ldots, R$
  - generate random sample $y_r^* = (y_{1r}^*, \ldots, y_{nr}^*) \sim \widehat{M}_0$
  - Compute T_r^* from y_r^*
- P-value estimate

$$\widehat{p} = \frac{1 + \#[T_r^* \geq T_{obs}]}{1 + R}$$

Hypothesis test

- Handedness data: are dnan and hand positively associated?
- Observed Correlation: $\widehat{\theta} = 0.509$

$$T_{obs} = 0.509^2 = 0.259$$

- Null hypothesis: independence

$$F(dnan, hand) = F_1(dnan)F_2(hand)$$

- For $r = 1, \ldots, R$
  - Simulate bootstrap samples independently from $\widehat{F}_1(dnan_1, \ldots, dnan_n)$ and $\widehat{F}_2(hand_1, \ldots, hand_n)$, then put them together $(dnan_1^*, hand_1^*), \ldots, (dnan_n^*, hand_n^*)$
  - Calculate $T_r^* = \widehat{\theta}_r^{*2}$
- P-value estimate (R=10000)

$$\widehat{p} = \frac{1 + \#[T_r^* \geq T_{obs}]}{1 + R}$$

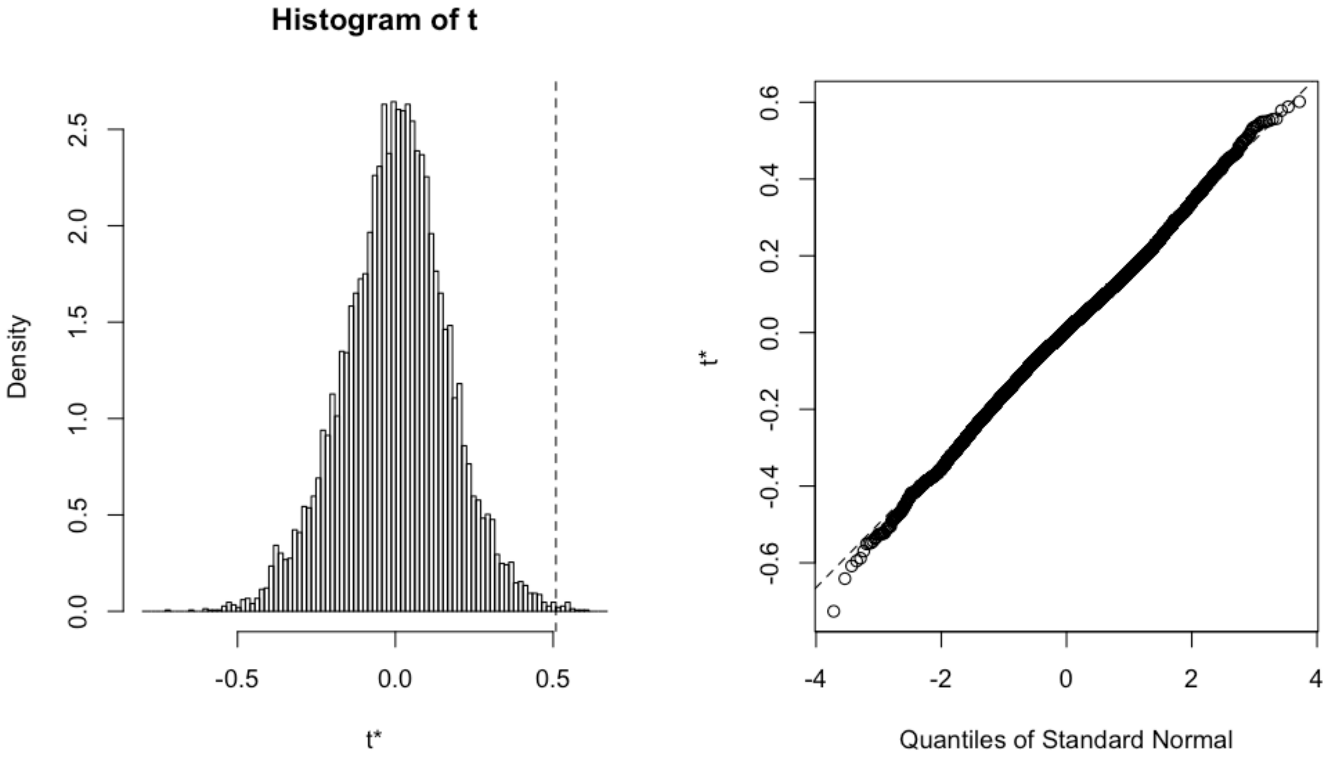
Hypothesis test

- R code (boot package)

```
 set.seed(100)
# Bootstrap p-values
R<-10000
New.D<-c(claridge$dnan, claridge$hand)
Corr1 <- function(d, f, n){
x<-d[f]
cor(x[1:n], d[(n+1):(2*n)])
}
boot.out1<-boot(New.D, Corr1, R=R
, strata=rep(c(1,2), c(n,n)), n=n)


n1<-sum(boot.out1$t^2 >= boot.out1$t0^2)
Pval.boot = (n1+1)/(R+1)
```

Hypothesis test

- Bootstrap p-value: 0.0041
- Histogram under the null (bootstrap)



Hypothesis test

- Alternatively confidence interval can be used for hypothesis test
- Handedness data:
  - 95% CI does not include 0
  - Can reject H_0 at $\alpha = 0.05$

Hypothesis test

- Instead of using Bootstrap, Permutation can be used for the hypothesis test
- For $r = 1, \ldots, R$
  - Take samples from

    $$(dnan_1, hand_{1^*}), \ldots, (dnan_n, hand_{n^*})$$

    where $(1^*, \ldots, n^*)$ is random permutation of $(1, \ldots, n)$
  - Calculate $T_r^* = \widehat{\theta}_r^{*2}$
- Handedness data, permutation p-value=0.0049
  - Nearly identical to bootstrap p-value=0.0041

Linear Regression

- Independent data $(x_1, y_1), \ldots, (x_n, y_n)$ with

$$y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2)$$

- Studentized residuals

$$e_i = \frac{y_i - x_i^T \widehat{\beta}}{(1 - h_i)^{1/2}} \sim (0, \sigma^2)$$

- Two main resampling schemes
- Model based resampling (or residual resampling)

$$y_i^* = x_i^T \beta + \epsilon_i^*, \quad \epsilon_i^* \sim EDF(e_1 - \bar{e}, \ldots, e_n - \bar{e})$$

  - Fixed design X, but not robust to model failure

- Case resampling

$$(x_i, y_i)^* \sim EDF[(x_1, y_1), \ldots, (x_n, y_n)]$$

  - Varying design X, but robust
  - Assume (x_i, y_i) sampled from population

GLM

- Case resampling can be used for GLM.
- There exist approximation methods for residual resampling.

GLM: Seizure count

- Seizure count data (Overdispersion!)

```
 > dat<-read.table("./seizure1.txt", header=FALSE)
> colnames(dat)<-c("Y1","Y2","Y3","Y4", "Z","base", "age")
> dat$Y<-dat$Y1+dat$Y2+dat$Y3+dat$Y4
>
> out<-glm(Y ~ age+base+Z, data=dat, family=poisson)
> summary(out)

Call:
glm(formula = Y ~ age + base + Z, family = poisson, data = da

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8949  -2.0883  -0.9471   0.7746  11.0049

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.072832   0.115817  17.897  < 2e-16 ***
age          0.018678   0.003336   5.599 2.15e-08 ***
base         0.022615   0.000510  44.346  < 2e-16 ***
Z           -0.184221   0.046487  -3.963 7.41e-05 ***
---
```

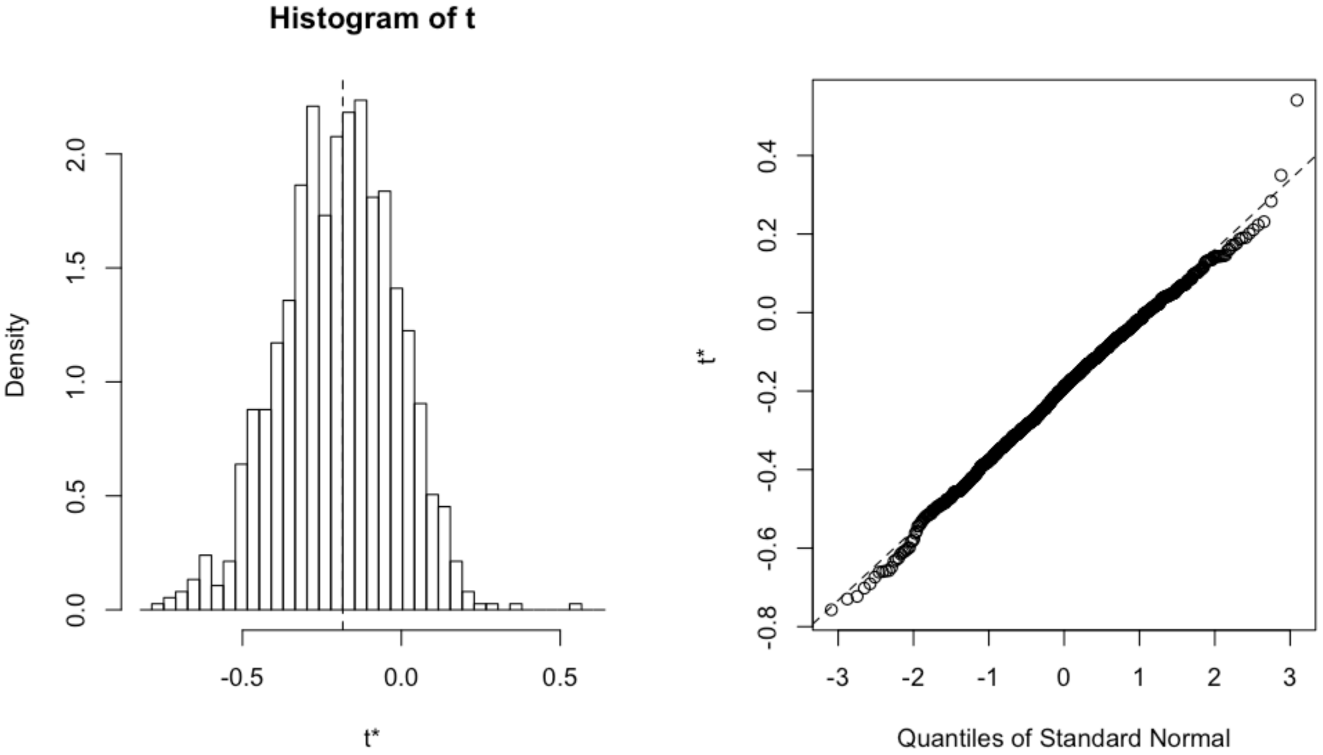
GLM: Seizure count

- Bootstrap $R = 1000$

```
> # case resampling
> Seizure <- function(d, f){
+ d1 = d[f,]
+ out<-glm(Y ~ age+base+Z, data=d1, family=poisson)
+ out1<-summary(out)$coefficients[4,1]
+ return(out1)
+
+ }
> boot.out<-boot(dat, Seizure, R=1000)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = dat, statistic = Seizure, R = 1000)

Bootstrap Statistics :
      original       bias    std. error
t1* -0.1842214 -0.01292668   0.1786147
```

GLM: Seizure count

- CI: $(-0.5437, 0.1412)$ (BCA)



SAS example (jackboot macro)

```
data seizure1;
  infile "~/BIOSTAT651/seizure1.txt";
input Y1 Y2 Y3 Y4 Z base age;
Y_tot=y1+y2+y3+y4;
idnum=_N_;
run;

%inc "~/BIOSTAT651/jackboot.sas";

%macro analyze(data=,out=);
    options nonotes;
proc genmod data=&data;
  model Y_tot = age base Z / dist=Poisson  link=log;
  ods output ParameterEstimates=&out(drop=DF StdErr
LowerWaldCl UpperWaldCL ChiSq ProbChiSq);
  %bystmt;
run;

    options notes;
%mend;
ODS SELECT NONE;
%boot(data=seizure1,samples=1000, id=Parameter, random=123);
%bootci(bca,alpha=.05, id=Parameter)
```

```
ODS SELECT ALL;


proc print data=BOOTSTAT;
run;

proc print data=BOOTCI;
run;

/* Get Bootstrap dist for Z */
data BOOTDIST1;
set BOOTDIST;
if Parameter ne "Z" then delete;
run;

proc UNIVARIATE data=BOOTDIST1;
var Estimate;
histogram;
run;
```

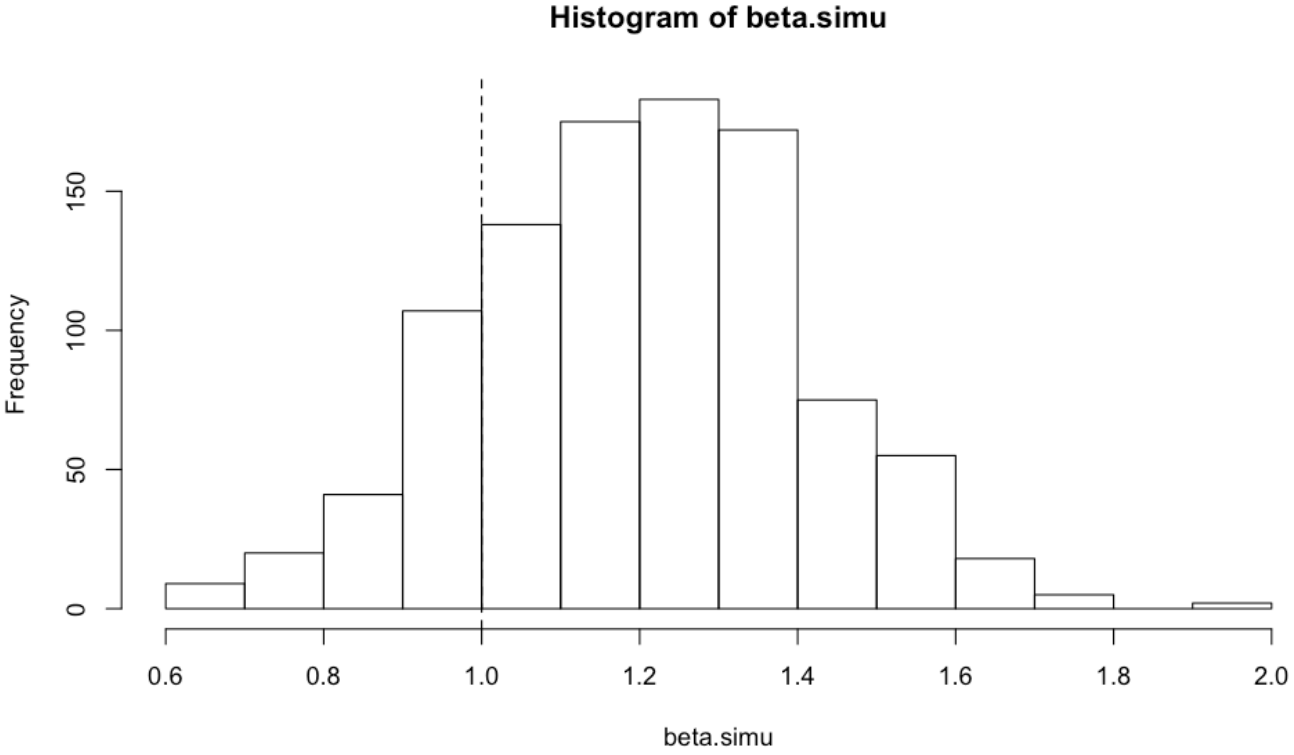
GLM: too many strata

- 100 strata, each has 6 samples
- In each stratum, 3 individuals received treatment ($X_{ki} = 1$) and 3 received placebo ($X_{ki} = 0$).
- Logistic regression model:

  $$logit(\pi_{ki}) = \alpha_k + \beta X_{ki}, \quad (k = 1, \ldots, 100; i = 1, \ldots, 6)$$

GLM: too many strata

- The true $\beta = 1$
- Generate data 1000 times and get the distribution of $\widehat{\beta}$
- Mean $\widehat{\beta} = 1.2 \rightarrow$ Bias $= 0.2$



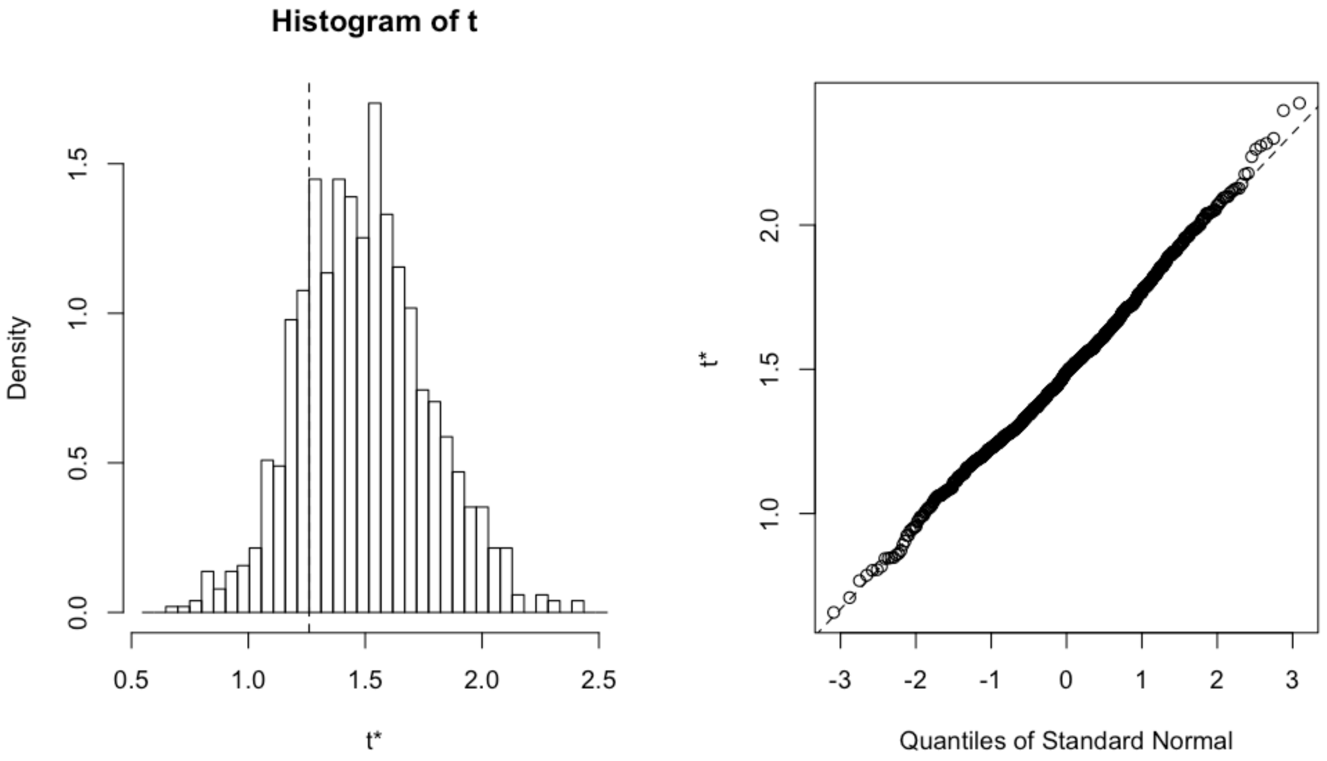
GLM: too many strata

- Carry out 1000 bootstrap to estimate the bias

```
> # case resampling
> StrataDat <- function(d, f){
+ d1 = d[f,]
+ out<-glm(Y ~ X + factor(Strata) -1, data=d1
, family=binomial)
+ out1<-summary(out)$coefficients[1,1]
+ return(out1)
+
+ }
> boot.out<-boot(dat, StrataDat, R=1000)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = dat, statistic = StrataDat, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 1.260071 0.2332586   0.2748678
```

GLM: too many strata

- Bootstrap bias estimate = 0.233



Bootstrap

- Bootstrap is a very useful tool to estimate sampling distribution of statistics
- In regression model, you can use either case-resampling or residual-resampling
  - In GLM, residual-sampling can be done (using approximation), but case-sampling is more widely used.
- There exist several R packages (ex. boot package)
- In SAS, you can use jackboot macro.