Big 4 approaches

- Traditional/Parametric
- Bootstrapping/Randomization/Monte Carlo
- Likelihood
- Bayesian

Frequentist philosophy & Power

Frequentist approach

- More properly called Neyman-Pearson or "hypothesis-testing" approach
  - Frequentist is name given by the Bayesians
- Central idea is:
  - Test null hypothesis (H_0) vs. alternative
  - Calculate a p-value (probability of null hypothesis being true
  - Reject null, accept alternative if p "small"
- Usually two flavors:
  - Normal and nonparametric

Null hypothesis logic

- Example:
  - H_0 no treatment effect ($\mu_{control}=\mu_{treatment}$)
  - H_a treatment (e.g. increased N) increases yield ($\mu_{control}<\mu_{treatment}$)

1. $H_0 \rightarrow$ null data
2. Observe ~ null data
3. $\rightarrow \sim H_0$
4. $\rightarrow H_a$

- Step #4 is true ONLY if H_a is the true opposite (complement) of H_0
  - True in this simple example ??
  - Only if state as $\mu_{control} \neq \mu_{treatment}$

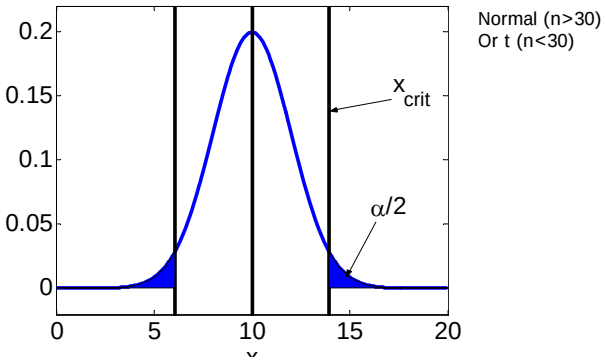
4 cases

	Reject null	Accept null
Null correct	Type I (α)	Correct
Null wrong	Correct	Type II (β)

Classical approach

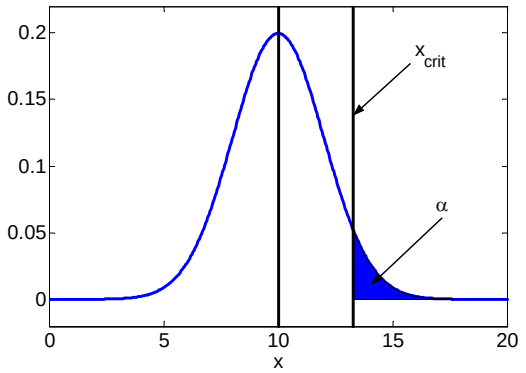
- Used for experiments
  - Null and alternative simple, opposite
- Specify p=α *a priori*
- Run experiment
- Reject null if observed effect exceeds critical value
  - =Accept alternative (what we were trying to prove)
- Fits Popperian model: two hypotheses, try to falsify and accept H_a if reject "all" others
  - Only approximate since null is not really a hypothesis

Critical values



Normal (n>30)
Or t (n<30)

Two vs one-tailed tests



Using one-tailed tests

- One main reason: has more power (get a significant result more often)
- Requirement: a priori specification of one-tailed
  - In practice in addition to stating a priori, should argue why
  - E.g. if H_a is $\mu_{treatment} > \mu_{control}$
    - H_0 becomes $\mu_{control} \leq \mu_{treatment}$
- Rice & Gaines
  - Intermediate x% of weight in left tail, 1-x% in right tail

Common misapplications

- "opposite" null often very difficult
  - Quinn & Dunn
  - What is "opposite" of succession?
- In principle
  - α specified a priori
  - Only report if above or below critical value (reject null or not)
  - Failure to reject null is NOT accepting the null
- What if want to prove equality
  - Ideal Free Distribution – organisms move so that fitness equal across heterogeneous environments

Null hypotheses

- Extremely controversial
  - Diamond found patterns in species distribution on islands and used them as evidence (observational data)
  - Various others claimed these patterns occurred by chance
  - VIGOROUS debate about the appropriate null model
- Net:
  - There is no single correct null model
  - Null models vary in how much biological realism they contain
  - But reasoning from observational data without a null model is DANGEROUS!

My take

- Three kinds of null hypothesis
  - Statistical null – traditional H_0 – logical mathematical opposite of alternative
  - Randomized null
    - Used for cases where statistical null doesn't exist
    - Has some sort of randomization (point of null is human's overdetect patterns in random data)
    - Can incorporate more or less biology – no one right null
    - Randomized null should never incorporate so much biology it becomes a "theory" that might be true – science should progress if null rejected
      - Neutral theory
      - Mid-domain effect
  - Carefully opposite models
    - Build "enough" realistic biology in but incorporate exact opposite assumption of what you are testing
- Stochastic theories
- Limit arguments

Three examples

- Randomized
  - Diamond's co-ocurrence
- Careful opposite
  - Neutral theory (genetics)
  - Neutral theory (community ecology)
  - Random branching phylogeny

Appropriate α

- Journal editors eventually settled on $p<0.05$, but report actual p values
- This is completely arbitrary!
- Is $p=0.06$ really infinitely worse than $p=0.05$?
- If $0.05<p<0.10$ can I claim anything?

Appropriate α II

- Should I use the same α if:
  - Irreplaceable habitat will be destroyed if no significant effect of habitat is found?
  - Chemical will be allowed into the environment until a significant result of harm is found?
  - Medicine may help and has little side effects but formal studies have been small?

	Reject null	Accept null
Null correct	Type I (α)	Correct
Null wrong	Correct	Type II (β)

A contrarian view

- ***"[Researchers] pay undue attention to the results of tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of magnitudes of effects which they are investigating"*** **Yates 1951**
- *""We shall marshall arguments against [signficance] testing, leading to the conclusion that it be abandoned by all substantive science …"* Guttman 1985
- *"The continued very extensive use of significance tests is alarming"* Cox 1986
- *"A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world"* Cohen 1990
- *"Are the effects of A and B different? They are always different ---for some decimal place"* Tukey 1991

Contrarian view II

- ***"...the primary product of a research inquiry is one or more measures of effect size, not p values"*** Cohen 1990
- *"Most statistician are all too familiar with the conversations [that] start:*
  Q: What is the purpose of your analysis?
  A: I want to do a significance test.
  Q: No, I mean what is the overall objective?
  A: (with a puzzled look): I want to know if my results are significant
  And so on..." Chatfield 1991
- *"...surely, god loves the 0.06 nearly as much as the 0.05." Rosnell and Rosenthal 1989*
- ***"Can you articulate even one legitimate contribution that significance testing made (or makes) to the research enterprise (i.e., any way in which it contributes to the development of cumulative scientific knowledge)? I believe you will not be able to do so"*** Schmidt 1996

Power

Purpose of power

- A priori: how big does my experiment need to be?
- Post hoc: if we fail to reject the null hypothesis then:
  - Our sample size was too small/variance is very large making false rejection of large effect
  - The effect is nonexistent or real but trivially small
  - Want to differentiate

4 cases

	Reject null	Accept null
Null correct	Type I (α)	Correct (1- α)
Null wrong	Correct (1- β=Power)	Type II (β)

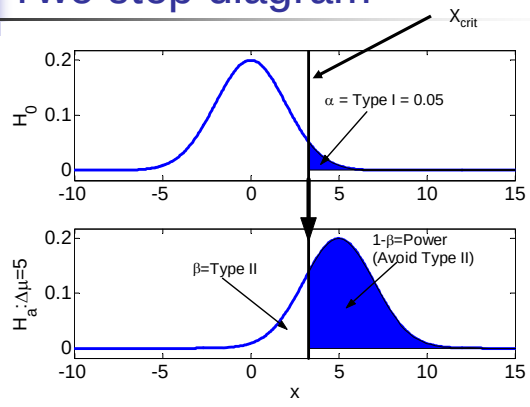
In practice

- Five parameters
  - Effect size (d)
  - Sample size (n)
  - Variance (s)
  - α (type I error)
  - Power (1-β, ranges from 0-1, close to 1 good)

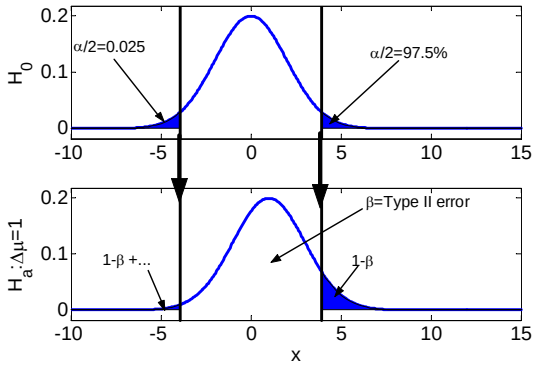
In practice II

- Given 4, can calculate the 5th
  - Usually
    - α specified (0.05)
    - Variance s/σ obtained from pilot study, related studies, generally similar studies
  - A priori ask:
    - Given d, n what is β (I can afford this experiment, what's my chance of success?)
    - Given β, d what n is needed (to be 90% successful, how big do I have to be?)
  - A posteriori ask:
    - Given β, n, what effect size (d) can be detected (could my experiment reject the null hypothesis in interesting cases?)
    - Warning: circular to ask what power of observed effect size is – just a fancy restatement of p!

Two step diagram



Two-tailed test



R example

- Suppose:
  - Pilot studies suggest $s^2=9$, $\mu_{control}=10$
  - We think our treatment will lead to an increase well above 20% ($\Delta\mu=2$)
  - How big a sample size do we need, assuming $\alpha=0.05$, $\beta=0.90$, one-tailed test?
- Background
  - Don't know if n>30, use t
  - Under H_a t is "noncentral – 3rd parameter $\lambda=\Delta\mu/\sigma=\Delta\mu/(s/\sqrt{n})=\Delta\mu/\sigma\ \sqrt{n}=d\sqrt{n}$
- Steps
  1. Find critical value of t for p=0.95, n=guess=30
  2. Find power (inverse CDF at critical value)
  3. Repeat until find n that gives $\beta=0.90$

```
n<-30;1-pt(qt(0.95,n),n,2/3*sqrt(n))
```

Or ...

```
power.t.test(delta=2,sd=3.5,power=0.8,alpha=0.05)
```

Also `power.prop.test`, `power.anova`

R example - easy

```
power.t.test(n,delta,sd=1,sig.level=0.05,power=NULL)
#also options for one-tailed, paired t, etc

power.t.test(n=50,delta=2) # gives beta w/ delta in
   units of SD
power.t.test(delta=2, power=0.9) # gives required
   sample size

power.anova.test

groupmeans <- c(120, 130, 140, 150)
power.anova.test(groups = length(groupmeans),
      between.var=var(groupmeans),within.var=500,
       power=.90) # gives n = 15.18834

model.tables(npk.aovE, "means"|"effects" [,se=T])
```

Web tools for power

- http://calculators.stat.ucla.edu/powercalc/
- http://www.stat.uiowa.edu/~rlenth/Power/

Steidel's confidence interval

- In general an idea that captures power is the confidence interval
  - High power=small confidence interval
- Gives idea of range of possible effect sizes
- Steidel argues for this

Summary

- Should ALWAYS do power analysis before starting an experiment
  - Need estimate of s
  - Need to think about d that is interesting
- Concept simple, calculations messy
- Use a preexisting package
- Always question failure to reject null hypothesis without high power

Bootstrapping & randomization

Bootstrapping

- From the image of “pick yourself up by your boot straps”
- Classical statistics:
  - Population is abstract you have a sample
  - Data is normal (or known distribution)
- Bootstrap approach:
  - You have data
  - You can resample it
  - Don’t need to know the distribution

When not to use boostrapping

- To make 4 data points look like 50
- Bootstrapping doesn’t save you from too small a sample

When to use bootstrapping

- Data not normal
  - Alternative to non-parametric, transforms
  - More intuitive
- No well developed statistical theory
  - Calculating a complicated “statistic”
    - A statistic is any calculation on the data
  - Bootstrapping can give the:
    - Mean sampling value of the statistic
    - Confidence intervals
    - The full distribution
    - Above 2 → hypothesis tests

Bootstrap application #1 - Randomization/reshuffling

- Analogous to t-test
  - Continuous variable measured on binary
  - Eg measure # ant nests in forest vs grass
- Steps
  - Randomly reassign category
  - Calculate difference in means
  - Repeat to build distribution of differences
  - If actual difference lies outside 95% interval then significant

Example

Actual	
Forest	7
Forest	5
Forest	8
Forest	8
Grass	5
Grass	6
Grass	4

forest=7, grass=5
Dif=2

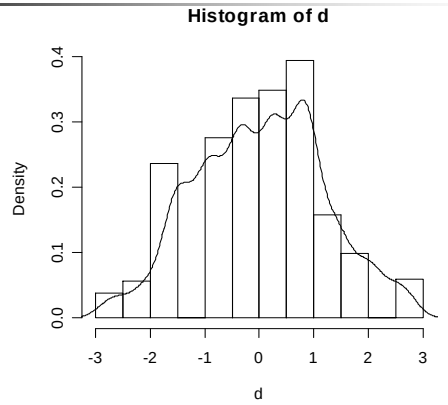
Random 1	
Forest	7
Grass	5
Forest	8
Grass	8
Forest	5
Forest	6
Grass	4

forest=6.5, grass=5.67
Dif=0.833

Random 2	
Grass	7
Forest	5
Grass	8
Forest	8
Grass	5
Forest	6
Forest	4

forest=5.75, grass=6.67
Dif=-0.917

Bootstrap replication



In R

```
t=factor(c("forest","forest","forest","forest","grass
  ","grass","grass"))
ct=c(7,5,8,8,5,6,4)
summary(lm(ct~t))
#
reps=1000
d=numeric(reps)
for (i in 1:reps) {
ts=t[sample(length(t))]
d[i]=mean(ct[ts=="forest"])-mean(ct[ts=="grass"])
}
d
hist(d,prob=T)
lines(density(d))
quantile(d,prob=0.95)
```

In R version 2 – using boot

```
library(boot)
t=factor(c("forest","forest","forest","forest","grass",
  "grass","grass"))
ct=c(7,5,8,8,5,6,4)
df=data.frame(habit=t,ct=ct)
df
 b=boot(df,function(din,f)  #cheats & uses ct
   {d=din[f,];mean(ct[d$habit=="forest"])-
   mean(ct[d$habit=="grass"])},R=1000)
b=boot(df,function(din,f)   #self contained
   {d=din[f,];mean(din$ct[d$habit=="forest"])-
   mean(din$ct[d$habit=="grass"])},R=1000,sim="permutati
   on")
b
summary(b)
boot.ci(b,conf=0.90,type="perc")
plot(b)
```

Bootstrap application #2 – Complex statistic

- You do some complex calculation on the data
  - Don't know the behavior of this
  - Contrast with mean (known ~ $N(\mu,\sigma/\sqrt{n})$
- Examples:
  - Ratio of two variables
  - Species evenness
  - Coefficient of variation

Application #2.1

- Have a large sample of individuals identified to species
- Want to estimate the H=Gini (Shannon-Weaver) measure of evenness
  - $(\Sigma p_i \log_2 p_i)/\log_2 S$
  - 0=All individuals of one species
  - 1=Every individual a new species
- Very simple to calculate H on a sample
  - How can I compare two sites?
  - How can I even assert that H is statistically different from 0 or 1?
- Don't do this on S! (or max,min)
  - badly behaved on sample w/ replacement, no variation on sample w/o replacement

Example in R

```
sid=read.table("/temp/sid.txt",h=T)
sid=sid$SID #just a vector
t(sid)  #transpose
table(sid) #crosstab
ct=as.matrix(table(sid))
hist(ct)
hist(log2(ct))
gini<-function(spid) {p<-table(spid)/length(spid);-
   sum(p*log(p))/log(length(unique(p)))}
gini(sid)
b=boot(sid,function(x,i) gini(x[i]),R=1000)
b
plot(b)
boot.ci(b,conf=0.95,type="perc")
#bbad=boot(sid,function(x,i) length(unique(x[i,])), R=1000)
```

R output

```
> b
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = sid, statistic = function(x, i) gini(x[i, ]), R = 1000)
Bootstrap Statistics :
     original      bias    std. error
t1* 0.7608427 0.008704316 0.005437653
> summary(b)
          Length Class      Mode
t0           1   -none-     numeric
t         1000   -none-     numeric
R            1   -none-     numeric
…
stype        1   -none-     character
strata    4159   -none-     numeric
weights   4159   -none-     numeric
> boot.ci(b,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
CALL :
boot.ci(boot.out = b, conf = 0.95, type = "perc")
Intervals :
Level     Percentile
95%   ( 0.7585,  0.7807 )
Calculations and Intervals on Original Scale
```

Application #2.2 (Tessier)

- Hypothesis: plant is more patchy in distribution in more disturbed wetlands
  - Measure of patchiness is CV (σ/μ) of abundance in 100 quadrats
  - Get three levels of disturbance, two plots x 100 quadrats
- Tons of work (600 quadrats!) must have enough data
- Oops – really only 2 replicates for 3 levels (6 data points) – no power!
  - CV collapses 100 plots into 1 data point
- Solution: bootstrap CV for each of 6 sites, now now confidence intervals around CV, can do hypothesis tests with power

Bootstrap application #3 – Regression confidence intervals

- Getting confidence intervals on parameters easy for GLM – hard for:
  - Robust
  - Nonlinear
- Two methods
  - Bootstrap null distribution – reshuffle data association
  - Bootstrap variability of parameter estimates
    - Use bootstrapping on residuals
    - Take the x values of input set
    - Take the estimated (least square) y values
    - Repeatedly add a randomized sample of the residuals
    - Recalculate the coefficients (new regression with new y)
    - Get a distribution of the coefficients

Example in R

```
library(quantreg)
data(Mammals)
plot(speed~log10(weight),data=Mammals)
m=rq(speed~log10(weight),tau=0.9,data=Mammals)
m
abline(m)
dat=data.frame(weight=Mammals$weight,
   speed=Mammals$speed, fitted=m$y-resid(m),
   resid=scale(resid(m),scale=F))
b=boot(dat,function(d,i)
   {dat=d;dat$speed=d$speed+d$resid[i];
   coef(update(m,data=dat))}, R=1000)
plot(b,index=1)
plot(b,index=2)
boot.ci(b,conf=0.95,type="perc",index=1)
boot.ci(b,conf=0.95,type="perc",index=2)
```

Bootstrap application #4 – Mantel test (Chapter 16 S&G)

- Correlation of distance matrices
- Example 1:
  - 4 sites
  - Distance (km) between sites
  - Change in mean temperature
  - Is distance related to temperature?

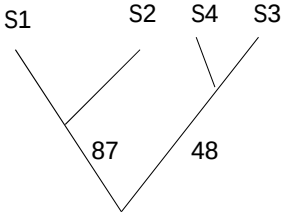
	A	B	C	D
A	0	2	1	5
B	2	0	3	4
C	1	3	0	2
D	5	4	2	0

- Example 2:
  - 4 species
  - Genetic distance
  - Morphopmetric distance
  - Are they the same
- Method:
  - Randomly reshuffle column/row for one matrix
  - Compare to other unshuffled matrix
  - Treat each cell as datapoint & calculate r
  - Get distribution of r's under null hypothesis of no special mapping of column to column (e.g. site to site)
  - Analytic version available but bootstrap often used

Bootstrap application #5 - phylogenies

- Phylogeny takes character matrix in
- Produces phylogeny
- Bootstrapping on character matrix
  - → Confidence intervals on branches (% time branch is included)

	Site 1	Site 2	Site 3	Mandible Length	Crown Height
Species1	A	G	T	50	2.4
Species2	A	G	T	48	2.6
Species3	C	T	G	65	4
Species3	C	A	G	50	4.1



Terminology

- Randomization – general term
- Bootstrapping – form of randomization where you have one variable and you resample from it
- Reshuffling or permutation – form of randomization where you have two (+) variables and you change their associations randomly
- Monte Carlo – very general term that applies to any modeling using random numbers

Bootstrap with constraints

- Sometimes there is correlation between the data
- Must incorporate these constraints into the sampling

Bootstrap summary

- Bootstrap
  - Creates mean (bias), confidence intervals, distribution of a "statistic"
- Useful when:
  - Statistic is complex
  - Statistic collapses lots of data
  - Statistic simple but data nonnormal (nonparametric is equally good alternative)
- Don't bootstrap
  - To hide small sample
  - On certain statistics (max,min, S)

Likelihood

Likelihood is one of big four

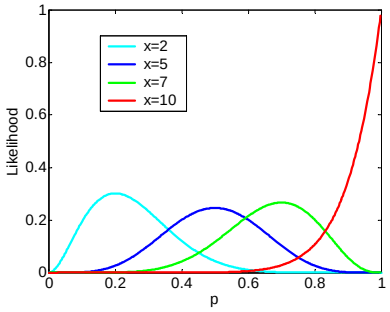
- Frequentist/Likelihood/Bayesian/Monte Carlo
- Likelihood has features in common with both Frequentist & Bayesian
- Inventor of likelihood is RA Fisher
- Great unifier – every frequentist method can be derived from scratch using likelihood

Likelihood

- Turn probability on its head
  - Normally P(data|parameters)
    - Estimate parameters (e.g. μ,σ)
    - Then look at probability of observing data given estimates or null hypothesis parameters
      - P(data|H0:μ_0=0)
    - Random variables are this way too – parameters are assumed known although they never are
  - Likelihood:
    - L(parameters)=P(parameters|data)

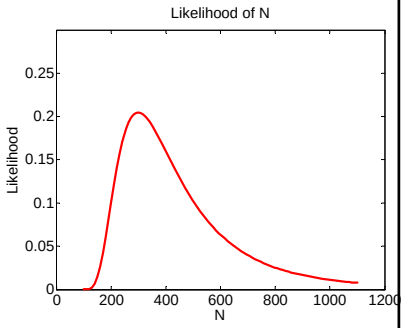
Example

- Binomial coin flips
  - Know n=10, x=2 heads, p unknown
  - What is L(p)
  - Binomial
    - L(p)=P(X=2)=(10 choose 2)$p^x(1-p)^{10-x}$



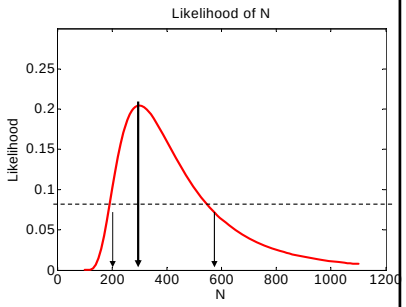
Practical application

- Mark-recapture
  - Capture 25 and tag & release
  - Capture 60, 5 tagged
  - Naively 25*60/5=300
  - What is N, population size
  - Hypergeometric distribution



How would you report N

- The whole distribution?
- Peak or maximum?
- Confidence interval
  - Equal likelihood
  - Not symmetric



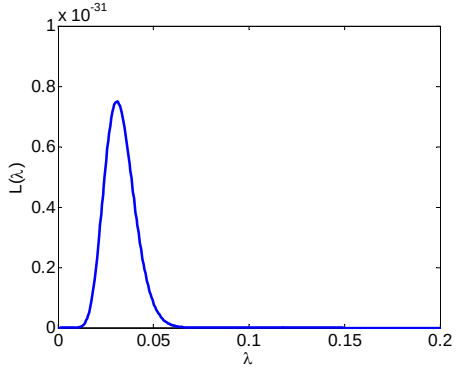
What if you have multiple points

- Binomial many trials rolled into one number
- What if we are looking at time to events (e.g. look at tree rings for fire scars, get times between fires for that tree)
- Get a sequence of inter-fire intervals:
  - 2 2 3 3 4 4 6 7 9 12 20 28 39 70 110 200
  - Treat as exponential: $p(x|\lambda)=\lambda\exp(-\lambda x)$

Use law of independent events

- $p(x_1 \& x_2 \& x_3 \& \ldots \& x_n)=$
  - $p(x_1)*p(x_2)* \ldots p(x_n)$
  - Iff independent
- In likelihood very often assume independence
- So
  - $L(\lambda|2\ 2\ \ldots\ 200)=$
  - $L(\lambda|2)*L(\lambda|2)*\ldots*L(\lambda|200)$
  - $=\lambda^n \exp(-\lambda\ 2)*\exp(-\lambda\ 2)*\ldots*\exp(-\lambda\ 200)$
  - $=\lambda^n \exp(-\lambda*(2+2+\ldots+200))$
  - $=\lambda^n \exp(-\lambda*\Sigma t_i)$

Likelihood of $\lambda|$data



Estimation

- Makes sense to use the "maximum likelihood estimate" (MLE) to estimate λ
- Can find the maximum:
  - Visually
  - Numerically
  - Calculus

Maximum likelihood estimatof fire

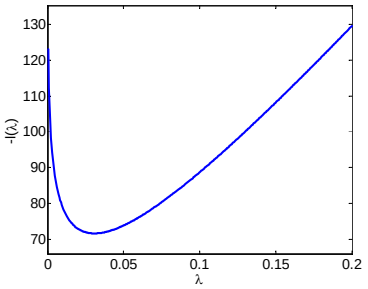
- Recall $L(\lambda|\{t_i\}) = \lambda^n \exp(-\lambda * \Sigma t_i)$
- Recall maximum wrt λ occurs where derivative=0
- So $d/d\lambda\ L(\lambda|\{t_i\})$
  - $0 = d/d\lambda\ \lambda^n \exp(-\lambda * \Sigma t_i)$
  - $= n\ \lambda^{n-1} \exp(-\lambda * \Sigma t_i) + \lambda^n \exp(-\lambda * \Sigma t_i) * (-\Sigma t_i)$
  - $= \exp(-\lambda * \Sigma t_i) * \lambda^{n-1} * (n - \lambda\ \Sigma t_i)$
  - Or $0 = (n - \lambda\ \Sigma t_i)$
  - Or $\lambda = n/\Sigma t_i$ or $1/\lambda = \Sigma t_i/n$ = average t
- MLE $\lambda = 1/32.43 = 0.0308$

A trick

- log is monotonic so log(f(x)) has a maximum at the same place as maximum of f(x)
- l(params|data)= log L(params|data)
- Almost always makes calculations much easier for likelihood
  - starts with product of likelihoods which is hard to take a derivative of and turns it into sums which is easy to take derivative of
  - $L(\lambda|\{t_i\}) = \lambda^n \exp(-\lambda * \Sigma t_i)$
  - $l(\lambda|\{t_i\}) = \log L(\lambda|\{t_i\}) = n \log \lambda - \lambda\ \Sigma t_i$
  - So:
    - $d/d\lambda\ l(\lambda) = n/\lambda - \Sigma t_i = 0 \rightarrow$ MLE $\lambda = n/\Sigma t_i$

Finishing the trick

- Sometimes also take –log L
  - Then we minimize instead of maximize (just like least squares)
  - Very small numbers ($1x10^{-31}$) turn into numbers like +31 or +71.38
- This is called "support" or just l



Works for the normal too

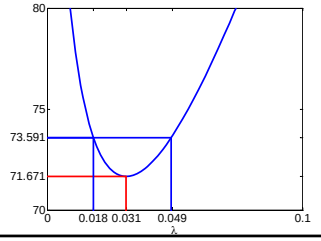
- $L(\mu,\sigma|x) = 1/(\sigma\sqrt{2\pi})\exp(-(x-\mu)^2/(2\sigma))$
- $L(\mu,\sigma|\{x_i\}) = 1/(\sigma\sqrt{2\pi})^n * \exp(-(x_1-\mu)^2/(2\sigma^2)) * \exp(-(x_2-\mu)^2/(2\sigma^2)) * \ldots$
- $L(\mu,\sigma|\{x_i\}) = 1/(\sigma\sqrt{2\pi})^n \exp(-\Sigma(x_i-\mu)^2/(2\sigma^2))$
- $l(\mu,\sigma|\{x_i\}) = \log L = -n\log \sigma\sqrt{2\pi} - \Sigma(x_i-\mu)^2/(2\sigma^2)$
- Max of l
  - $0 = d/d\mu\ l = -2\ \Sigma(x_i-\mu)/(2\sigma^2) = [2n\mu - \Sigma x_i]/2\sigma \rightarrow \mu = (\Sigma x_i)/n$

Likelihood ratio

- Say we had a null hypothesis of λ=0.05=1/20 (fire every 20 years)
- How can we contrast this to observed λ=0.0308?
- Use a likelihood ratio
  - $L(\lambda_1)/L(\lambda_2)=c$
  - Can say λ_1 is c times more likely than λ_2
  - L(0.05)/L(0.0308)=8.2e-033/7.5e-032=0.109408
  - Or L(0.0308)/L(0.05)= 7.5e-032/ 8.2e-033=9.14
- Note log[L(p1)/L(p2)]=l(p1)-l(p2)
- For large sample:
  - $-2*\log[L(p1)/L(p2)]=-2[l(p1)-l(p2)]\sim\chi_n^2$
  - n is difference in degrees of freedom (e.g. n=1)

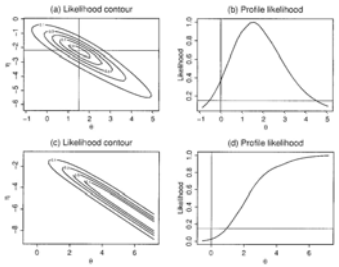
Confidence intervals

- Hudson's profile likelihood method
  - $-2*\log[L(p1)/L(p2)]=-2[l(p1)-l(p2)]\sim\chi_n^2$
  - Critical value $c^2_{1,0.95}$=3.84
  - -log likelihood difference of 3.84/2=1.92



2-parameter confidence intervals

- Option 1 – find CI holding other parameters fixed
- Option 2 – use integration to get contour intervals



Likelihood very general

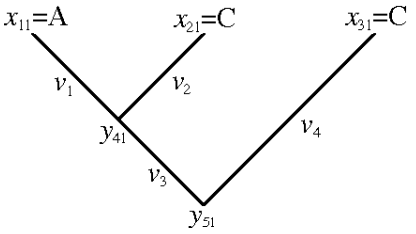
- Can use likelihood in much more complicated situations
- Dynamical models
- Phylogenies

Likelihood & dynamic models

- $N_{t+1}= \lambda N_t \exp(-cN_t)+\varepsilon_t$ (Ricker)
- Given N, model for ε, can calculate estimates for λ, c, ε
- Often assume $\varepsilon \sim$Lognormal(0,σ)
  - $N_{t+1}= \lambda N_t \exp(-cN_t+\varepsilon_t)$ (Ricker)

Likelihood & phylogenies

$$\mathbf{X}=\begin{pmatrix} A & T & T & \cdots & C \\ C & T & T & \cdots & A \\ C & G & T & \cdots & T \end{pmatrix}$$



$P(X|v,\pi,\Theta)=\pi_C \mu_{CA}(\varpi_1|\Theta)+ \pi_A * \mu_{AC}(v_4|\Theta)*\mu_{AC}(v_2|\Theta)$

Likelihood ratio test

- Use for nested models
  - Recall nested = one set of independent variables is subset of another set
  - Recall likelihood ratio $c=L_1/L_2$ is "how many times more likely" is 1 than 2
- For a nested model
  - If parameters are MLE estimates
  - Then -2 log $(L_1/L_2) \sim \chi^2_{p2-p1}$
  - Usually L_{H0}/L_{HA}, reject if $-2 \log(L_{H0}/L_{HA}) > \chi^2_{pa-p0},0.95$
- Example
  - Null hypothesis of λ=0.05=1/20 (fire every 20 years)
  - Observed λ=0.0308?
  - L(0.05)/L(0.0308)=8.2e-033/7.5e-032=0.109408
  - - 2 log (L/L)=4.4253>3.84= $\chi^2_{1-0},0.95$
  - →Reject null hypothesis

AIC (Akiake information criteria)

- Extension of likelihood ratio test for non-nested models
- Repeat NON-NESTED models
- Can compare any two models that share a common dataset and error model
- AIC=-2log(L)+2p
- The +2p comes out of deep mathematical reasons (information theory)
- AIC is a "badness of fit" (smaller AIC better)

AIC variations

- AIC=2k+n*log(RSS/n)
- AICc=AIC+2k(k+1)/(n-k-1)
  - Corrects for small sample size

Model comparison

- Three levels of approach
  - Null hypothesis HA vs H0
  - Nested with null at bottom H3 vs H2 vs H1 vs H0
  - Comparison – non-nested models, all biologically realistic
- AIC makes model comparison possible
  - Revolutionizing statistics
- AIC frees from nestedness

AIC continued

- Smallest AIC is best model
  - (least bad fit)
- Absolute AIC irrelevant (smaller AIC with bigger sample size)
- Delta (difference) matters
- $\Delta_i = AIC_i - AIC_{min}$
  - χ^2 distributed – can get p-values
  - $\Delta_i > 3.84$ is significant
    - Also likelihood ratio>1.92
  - Burnham & Anderson
    - Δ_i 0-2 Substantial support (relative to base)
    - Δ_i 4-7 Considerably less
    - >10 Essentially no support

AIC continued

- Weights
  - $w_i = \exp(-\Delta_i/2)/\Sigma_j \exp(-\Delta_j/2)$
  - w_i=probability for that model
  - Can use for blending of models
    - $\mu = \mu_1 w_1 + \mu_2 w_2 + \ldots$
    - Predicted y also
- Other related measures
  - BIC (Bayesian Information Criteria)
  - Small sample AIC
  - QIC
  - TIC

Species area example

	K	Likelihood	AIC	Δ_i	w_i	R^2
aA^b	3	221.64	227.64	813.12	0	0.962
a+blog(A)	3	85.56	91.56	677.04	0	0.986
$a(1-e^{-bA})$	3	526.17	529.17	1114.65	0	0.624
$a(1-e^{-bA})^c$	4	-50.85	-42.85	542.63	0	0.995
$A[1-e^{-b(x-c)^d}]$	5	-575.48	-585.48	0.00	1	0.999

Likelihood Summary

- p(parameter|data)
- Very general
  - Includes normal frequentist as special case
  - Extends to much more complicated scenarios
- Usually –log Likelihood
- Plot likelihood surfaces
- MLE (maximum likelihood estimate)
- Likelihood ratio
  - Tests nested models
- AIC
  - Tests non-nested models

In R

```
x=c(1,2,3,4)
y=c(0,2,3,5)
plot(x,y)
summary(lm(y~x))
library(bbmle)
reg_nll<-function(b0,b1,sig) -sum(dnorm(y,mean=b0+b1*x,sd=sig,log=T))
m<-mle2(reg_nll,list(b0=1,b1=1,sig=1))
summary(m)
p<-profile(m)
plot(p)
confint(p)
AIC(m)
```

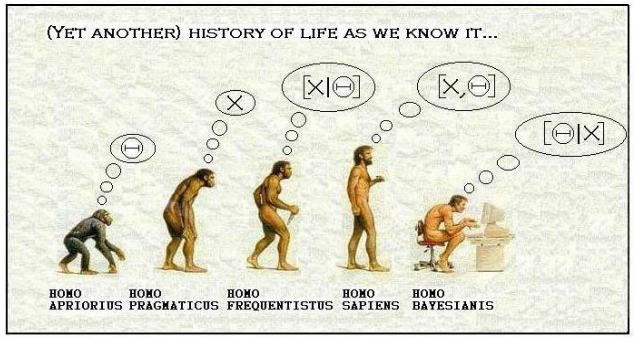
Bayesian analysis

Bayesian approach

- Calculate p(θ|data) (like likelihood)
  - Get probability distribution for θ, not probability of data given null θ
- Use a "prior"
  - Prior expectations about θ
  - Makes Bayesian very controversial – loss of objectivity!?

Bayes Theorem

- $p(\theta|data)=p(\theta)*p(data|\theta)/p(data)$
  $=p(\theta)*p(data|\theta)/\int p(\theta)p(data/\theta)d\theta$
- Posterior=prior*"likelihood"/c
- Likelihood is not the same sense as used by likelihood school
- Posterior & prior are:
  - Probability distributions for θ
  - Sum up to one across θ



Mike West - Duke

What prior

- Informative
  - Have previous data→previous expectation
- Uninformative
  - Well known distribution (e.g. normal) but with ***very*** high variance
- Improper uninformative:
  - uniform – all values equally likely
  - Not integrable->improper
- Conjugate –prior chosen so that calculations are easy
  - Well-known pairs (beta/binomial, Gaussian/inverse-Chi2, etc)
  - Can be informative or unininformative

Result

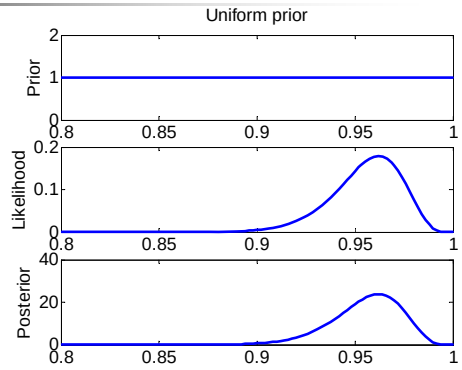
- Posterior = distribution of θ
- Can get expected value, confidence intervals on θ
- Solution of Bayes formula, 2 methods:
  - By calculus (usually need conjugate prior)
  - By computer
    - Gibbs distribution & Markov Chain Monte Carlo (MCMC)
    - Can simulate random sample from posterior without knowing formula
    - Many replicates gives approximation to distribution

Clark's example (S&G ch 17)

- Estimating mortality rate of *Acer rubrum* in southern Appalachians
- 127/132 trees survive in study site→μ=0.962
- But:
  - mortality rare, how much confidence do we have?
- Frequentist approach:
  - Binomial distribution with μ=p=0.962
  - Derive confidence intervals on μ
  - Alternatively compare μ=p=0.962 vs. μ_0
- Likelihood approach:
  - Derive likelihood surface
- $p(data|q)=(n,k)\theta^k\theta^{n-k}$ (i.e. binomial)

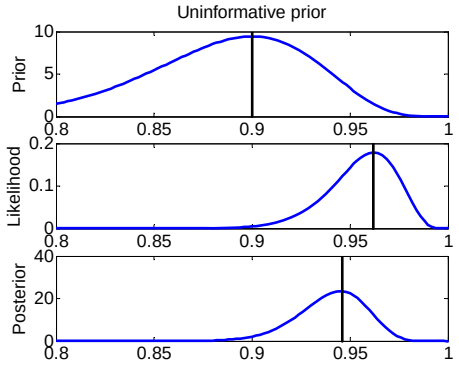
Uniform (noninformative) prior

- Prior uniform (p(x)=1)
- Note: posterior= likelihood rescaled
  - Generally true with large n
  - →same as frequentist!



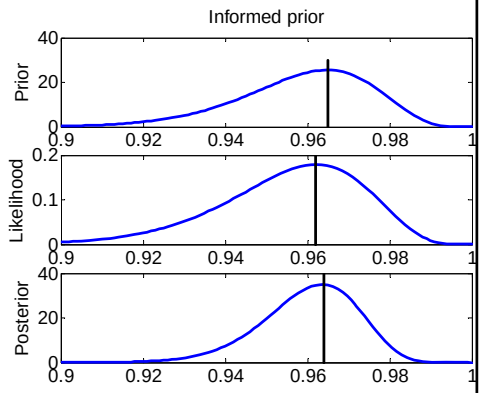
Weakly informative prior

- Mean 0.9, large variance

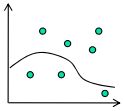


Informed prior

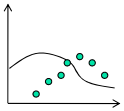
- Choose conjugate (beta)
- USFIA gives 137/142 for larger area
- Prior(q)= Beta(137,142)
- Posterior much narrower



MCMC



- Recall: $p(\theta|data)=p(\theta)*p(data|\theta)/p(data)$
  $=p(\theta)*p(data|\theta)/\int p(\theta)p(data/\theta)d\theta$
- Can be a tough integral
- Monte Carlo integration
  - $\int$=area=%points under f(x)*area of box
  - Breaks in many dimensions
- MCMC
  - Points are random walk w/ rejection rule
  - Gibb's sampler, Metropolis Hastings
- In practice
  - Ignore beginning (burn in)
  - Run for a long time!
  - Run a couple of times with different seeds



Bayesian hypothesis testing

- Use posterior
  - But p(b=0) is 0
- Credible intervals
  - Similar to confidence intervals
- BIC/Schwarz=-2log(L)+k*log(n)
  - cf log(n) vs 2 (in AIC)
- DIC
  - Deviance $D(\theta)=-2\log(p(y|\theta))+C$
  - Dbar=Eq[$D(\underline{\theta})$) badness of fit
  - pD=Dbar-$D(\underline{\theta}$bar) model complexity
  - DIC=pD+Dbar
  - Easy to calculate from MCMC/Winbugs

Net net

- Uninformative prior
  - Likelihood (data) dominates
  - Weakly informative may modify somewhat
  - *In limit with uniform prior, large data, same as likelihood*
  - MCMC allow solution of complex models
- Informative prior
  - Balance depends on amount of data in prior vs likelihood – a form of weighted average

Bayesian tools

- BUGS
  - http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.html
- BATS (time series)
  - http://www.stat.duke.edu/~mw/bats.html

Bayesian in WinBugs (Poisson, Poisson Lognormal)

```
#Quercus rubra trees per quadrat counted, Poisson distribution, want to estimate lambda (mean density)
#box 3.4 in McCarthy's book Bayesian Methods for Ecology
model
{
                for (i in 1:10) {
                                y[i]~dpois(lambda)  #each data point is sampled from poisson w/ mean lambda
                }
                lambda~dlnorm(0.0,1.0E-6) #uninformative prior - lognormal since lambda>0, 1.0E-6=precision, var=1/precision
}

list(y=c(6,0,1,2,1,7,1,5,2,0))

list(lambda=5) #ball park initial value for lambda
```

1. Open Winbugs
2. New file, enter,save
3. Model/specification...
   1. Dblclick on model, Check Model
   2. Dblclick on data, Load data
   3. Compile model
   4. Load or generate initial values
4. Model/Update...
5. Inference/Samples...
   1. Set variables to track
6. Run updates (from #4)
7. View history, distribution, stats (from #5)

```
#Poisson but lambda should vary across quadrats - wet, dry, etc
#let lambda be lognormal across quadrats
model
{
                for (i in 1:10)
                {
                                lambda[i]~dlnorm(mu,tau)
                                y[i]~dpois(lambda[i])
                }
                mu~dnorm(0,1.0E-3) #prior on lognormal mu 1.0E-6 makes too diffuse - cannot bracket error
                sd~dunif(0,10)
                tau<-1/(sd*sd)  #tau is precision for lognormal, precision=1/var, sd is uniform on 0-10
}

list(y=c(6,0,1,2,1,7,1,5,2,0))
```

From a leading advocate of Bayesian methods in ecology

- This importance of philosophy seems to be reinforced by examples aimed at demonstrating, comparing, and/or contrasting classical vs. SB. Such examples show that, with similar underlying assumptions (e.g. vague priors), classical methods and SB yield near identical confidence envelopes –with a lot more work, one can expect to obtain a Bayesian credible interval (CI) that is not importantly different from a classical confidence interval. … If one gets essentially the same answer, philosophy must be the motivation. Despite the counsel to start from one's view of probability, ecologists having no philosophical axe to grind might question the point; if Bayes requires more work to arrive at the same interpretation, why bother with Bayes?
- Those already possessing a healthy scepticism of classical hypothesis tests can continue to estimate classical confidence intervals without offending many Bayesians. The focus of this paper stems from an alternative view that the emergence of modern Bayes has little to do with philosophy, but comes rather from pragmatism. ….
- A growing number of practitioners are willing to let the choice between frequentist and Bayes rest on complexity…. The power of HB comes from a capacity to accommodate complexity.
- Clark 2005 Ecology Letters

Bayesian summary

- Prior+likelihood(of data)→Posterior
- Posterior is a probability distribution of parameter
- Various types of priors
- Benefits:
  - Allows increase of power with prior data
  - Allows balance of one experiment vs. world of known data
  - No silly null model
    - Even uninformative prior gets enhanced by data

Comparison summary

- Parametric → distribution of parameter (parametric e.g. t), emphasize p
  1. Normality assumed for model (usually)
  2. Estimator has a formula
  3. SE of estimator has a distribution
  4. →P values, Confidence Intervals
- Monte Carlo →distribution of statistic, emphasize p
  1. Define statistic
  2. Define randomization
  3. Calc statistic over many randomizations
  4. Compare observed statistic to distribution
1. Likelihood→likelihood distribution of parameter (often complex), emphasize whole distribution
   1. Define probability model
   2. Write likelihood
   3. Maximize, confidence interval, likelihood ratio
2. Bayesian→Posterior distribution of parameter (may be MCMC), emphasize whole distribution
   1. Define prior
   2. Define model
   3. Likelihood
   4. Calculate posterior =likelihood*prior
   5. Credible intervals, Bayes factors, BIC/DIC

Is there a difference?

- All get distribution of a parameter/statistic, but calculate/describe it different ways
- Tendency to use it different ways
  - =philosphy
  - But not required! Get a p-value in Bayesian, a distribution in parametric

	Frequentist (p) (p(data|param)	"Distributional" (CI) (p(param|data)	Model Selection	Subjective Prior	Dist (statistic)
Parametric	√	√ (distribution, e.g. t)	=likelihood		
Monte Carlo	√	√ (histogram)	√* (r^2)		√
Likelihood	√	√ (I surface)	√ (IC)		
Bayesian	√+	√ (MCMC or conjugate posterior)	√ (IC)	√	

*=cross validation, jacknife correct for overfitting
+=one-tailed test on posterior

Pros/cons

	Pros	Cons	Use when
Parametric	Familiarity Analytical solutions	Normal assumption Simple models only	Whenever you can (i.e. simple enough)
Monte Carlo	Extremely flexible (computer solve any model) Intuitive	Requires programming No theoretical framework	Too complicated Summary statistic that boils down lots of data Interesting null
Likelihood	General (others often special cases) Powerful	Requires advanced math	Medium complexity
Bayesian	Allows prior Extremely flexible (computer solve any model)	Complex Propaganda Priors subjective	Too complicated Have a prior

General setup

- Do all 4 styles of regression in 4

```
x=c(1,2,3,4)
y=c(0,2,3,5)
plot(x,y)
summary(lm(y~x))
```

Regression I - Parametric

- Parametric: Assume e~Normal
  - Use least-squares(=MLE) estimator
  - →estimates for β=cov(x,y)/var(x) and σ=RSS/df
  - β and σ ~ t-distribution
  - MSS/RSS~F

Parametric in R

```
xin<-cbind(1,x)
temp<-chol2inv(chol(t(xin)%*%xin)) #%*%=matrix mult
bhat<-temp%*%t(xin)%*%y
res<-as.vector(y-xin%*%bhat)
df<-(nrow(xin)-ncol(xin))
sse<-as.numeric(res%*%res)
tss<-sum((y-mean(y))^2)
seb=sqrt(diag(sse*temp)/df)
b1_ci=bhat[2]+c(qt(0.025,df=df),qt(0.975,df=df))*seb[2]
r2<-1-sse/tss
F<-((tss-sse)/1)/(sse/df)
pF<-pf(F,df1=1,df2=df)

curve(dt((x-bhat[2])/seb[2],df=df),0,4)#plot t-
distribution of b1
```

Regression II - Likelihood

- Likelihood
  - Use maximum-likelihood estimate
    - (→least-squares)
  - Likelihood surface of β and σ
  - Likelihood ratio test
    - (=F-test)

Likelihood In R

```
reg_nll<-function(p) -sum(dnorm(y,mean=p[1]+p[2]*x,sd=p[3],log=T))
o<-optim(c(1,1,1),reg_nll)
o$par
o$val
# 1-d plot of support surface (-log) for sigma
vals=seq(0.05,1,0.01)
z<-matrix(NaN,1,length(vals))
for (i in 1:length(vals)) {
    z[i]=reg_nll(c(o$par[1],o$par[2],vals[i]))
}
plot(vals,z,type="l")
plot(vals,exp(-z),type="l") #actual likelihood

#2-d plot of b0 vs b1
vals<-seq(-5,5,0.1);
z<-matrix(NaN,length(vals),length(vals))
for (i in 1:length(vals)) {
    for (j in 1:length(vals)) {
        z[i,j]=reg_nll(c(vals[i],vals[j],o$par[3]))
    }
}
image(vals,vals,z,col=rainbow(12),xlab="b0",ylab="b1")
contour(vals,vals,z,add=T)
```

Likelihood in R II

```
#Confidence interval
uniroot(function(x)reg_nll(c(o$par[1],o$par[2],x))-
   1.92,c(0,o$par[3]))
uniroot(function(x)reg_nll(c(o$par[1],o$par[2],x))-
   1.92,c(o$par[3],2))

#null test (likelihood ratio)
null_ll<-function(p) -sum(dnorm(y,mean=p[1],sd=p[3],log=T))
on<-optim(c(1,1,1),null_ll)
on
lik=prod(dnorm(y,mean=o$par[1]+o$par[2]*x,sd=o$par[3]))
nullik=prod(dnorm(y,mean=on$par[1],sd=on$par[3]))
lr<- -2*log(nullik/lik)
pchisq(lr,2)
```

Regression III - Bootstrap

```
#variation around null – significance testing
library(boot)
boot.regr <- function(data, indices){
    y <- data[indices,1]
    x<-data[,2]
    mod<-lm(y~x)
    coefficients(mod) # return coefficient vector
}
df<-data.frame(cbind(y=y,x=x)) #boot likes a dataframe input
b=boot(df,boot.regr,R=1000)
plot(b,index=1)
plot(b,index=2)
boot.ci(b,conf=0.95,type="perc",index=1)
boot.ci(b,conf=0.95,type="perc",index=2)

#variation around parameter estimate w/ residuals – distribution generation
m<-lm(y~x)
resid<-scale(resid(m),scale=F) #center but do not rescale residuals
df<-data.frame(x=x,y=y,resid=resid)
b=boot(df,function(d,i) {dat=d;dat$y=d$fitted+d$resid[i]; coef(lm(y~x,data=dat))}, R=1000)
plot(b,index=1) #note variance fairly low since bootstrapping doesn't create data!!
plot(b,index=2)
boot.ci(b,conf=0.95,type="perc",index=1)
boot.ci(b,conf=0.95,type="perc",index=2)
```

Bayesian in R

```
post<-function(beta,x,y) {
    if (beta[3]<0) beta[3]=0
    llik<- sum(dnorm(y,x%*%beta[1:2],sd=beta[3],log=T))
    prior<- sum(dnorm(beta,0,10,log=T))
    llik+prior
}

library(MCMCpack)
 out <- MCMCmetrop1R(post, c(1,1,1), mcmc=10000,x=xin,y=y)
plot(out)
summary(out)
```

Bayesian in winbugs

```
model
{
    #note: in dnorm 1E-6 is precision=1/var
     b0~dnorm(0,1.0E-6) #prior for intercept
     b1~dnorm(0,1.0E-6) # prior for slope
     sig~dgamma(0.1,0.001) #prior for sigma - gamma since must be >0

     for (i in 1:4) {
            mean[i]<-b0+b1*x[i]   #linear component
            y[i]~dnorm(mean[i],sig)  #error component
     }
}


list(x=c(1,2,3,4),
     y=c(0,2,3,5))
```