Mathematical Statistics I

STA2212H S LEC9101

Week 2

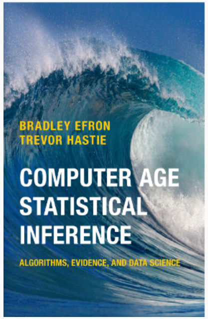
January 20 2021

Volume 88, Issue S1

Special Issue: "Data Science versus Classical Inference: Prediction, Estimation, and Attribution", honouring Prof. Brad Efron's International Prize in Statistics in 2019

Pages: S1-S224

December 2020



The Computer Age Statistical Inference book makes the distinction between the two levels of statistics, the algorithmic level and the inferential level, which is somewhat an artificial distinction but a pretty good one. It says that the first level is doing something and the second level is understanding what you did in the first level. The algorithmic level always gets more action, in particular in these days of these big prediction algorithms like deep learning. You'd think that's the only thing going on. It isn't the only thing going on. The deeper understanding of the kind of thing that Fisher and these people – Neyman, Hotelling – did for early 20th-century statistics, putting it on a solid intellectual ground so you can understand what's at stake, is terribly important.

Recap

- likelihood notation $\ell'(\theta) = U(\theta)$, $s(X, \theta)$
- score function, maximum likelihood estimate, observed and expected Fisher information $j(\hat\theta)$ $\mathcal{I}(\theta)$ $\mathcal{I}(\theta)$ $\quad \ell'(\hat\theta) = 0$
- asymptotic normality of maximum likelihood estimators
- estimating the asymptotic variance
- the delta method
- profile likelihood
- sufficient statistics $t(X)$ $\quad \ell(\theta; t(x)) + \ldots$
- Newton-Raphson method for computing $\hat{\theta}$
- irregular models
- Quasi-Newton
- EM Algorithm

notes on likelihood

$\sqrt{n}(\hat{\theta}-\theta)I_1^{1/2}(\hat{\theta}) \xrightarrow{d} N(0,1)$

$-\ell''(\hat\theta) =$ $j(\hat{\theta}), I_n(\hat{\theta})$

$\tau = g(\theta)$

see notes p.6

$U(0, \theta)$

1st order ?? proof ?

Friday

Today

1. Quasi-Newton
2. Hypothesis testing — AoS 10.1
3. Significance testing — SM 7.3.1; AoS 10.2
4. Tests based on likelihood — AoS 10.6

- January 25 3.00 – 4.00 Aleeza Gerstein — Data Science and Applied Research Series
- “Turning qualitative observation to quantitative measurement through statistical computing” Link

Notes on optimization: Tibshirani, Pena, Kolter CO 10-725 CMU

- Goal: $\max_\theta \ell(\theta; \boldsymbol{x})$
- Solve: $\ell'(\hat\theta; \underset{\sim}{x}) = 0$ mle $\frac{\partial}{\partial\theta}\ell(\theta; \underset{\sim}{x})$ $\hat\theta(\underset{\sim}{x})$
- Iterate: $\hat\theta^{t+1} = \hat\theta^t + \frac{\ell'(\hat\theta^t)}{-\ell''(\hat\theta^t)} = \hat\theta^t + [-\ell''(\hat\theta^t)]^{-1}\ell'(\hat\theta^t)$ $p \times p$ $p\times p$ $p\times 1$
- Rewrite:
- Quasi-Newton: $[\hat{j}]^{-1}(\hat\theta^{t+1} - \hat\theta^t) = \ell'(\hat\theta^t)$
- $(-\Delta\ell)^{-1}[\nabla\theta] = [\nabla\ell]$
-

replace $\hat{j}(\hat\theta^t)^{-1}$ by

by an approx$^=$

?

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent"),
      lower = -Inf, upper = Inf, control = list(), hessian = FALSE)
```

← 2nd

doesn't use 2nd deriv.

Notes on optimization: Tibshirani, Pena, Kolter CO 10-725 CMU

- Goal: $\max_\theta \ell(\theta; \boldsymbol{x})$
- Solve: $\ell'(\theta; \boldsymbol{x}) = 0$
- Iterate: $\hat\theta^{(t+1)} = \hat\theta^{(t)} + \{j(\hat\theta^{(t)})\}^{-1}\ell'(\hat\theta^{(t)})$
- Rewrite: $j(\hat\theta^{(t)})(\hat\theta^{(t+1)} - \hat\theta^{(t)}) = \ell'(\hat\theta^{(t)})$

$\underline{B\Delta\theta = -\nabla\ell(\theta)}$

- Quasi-Newton:
  - approximate $j(\hat\theta^{(t)})$ with something easy to invert
  - use information from $j(\hat\theta^{(t)})$ to compute $j(\hat\theta^{(t+1)})$

step size

- optimization notes add a step size to the iteration $\hat\theta^{(t+1)} = \hat\theta^{(t)} + \epsilon_t \underbrace{\{j(\hat\theta^{(t)})\}^{-1}\ell'(\hat\theta^{(t)})}$

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent"),
      lower = -Inf, upper = Inf, control = list(), hessian = FALSE)
```

Formal theory of testing

$$X_1, \dots, X_n = \underset{\sim}{X}_n \sim f(\underset{\sim}{x}; \theta)$$

$$\theta \in \Theta \subseteq \mathbb{R}^k$$

could be infinite

non par

$f(\underset{\sim}{x})$ "smooth"

$x_i = m(z_i) + \varepsilon_i$

"smooth"

- Null and alternative hypothesis
- Rejection region : $R \subseteq \mathcal{X}^n$

sample space

- Test statistic and critical value
- Type I and Type II error
- Power and Size

$H_0 : \theta \in \Theta_0$ null

$H_1 : \theta \in \Theta_1$ alternative

come back later

if $\underset{\sim}{x} \in R$: "reject H_0"

$\underset{\sim}{x} \notin R$: "don't reject H_0" (retain null)

- Null and alternative hypothesis
- Rejection region
- Test statistic and critical value
- Type I and Type II error
- Power and Size

$R \subseteq \mathcal{X} \quad x \in \mathcal{X}$

$R = \{x : t(X_n) > c\}$ 1-dim. summary

c - to be det'd

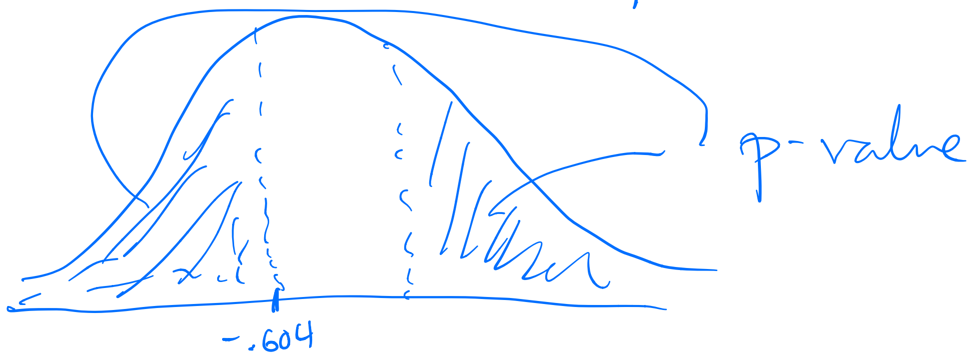
$\{X^n \in R : \theta \in \Theta_0\} = [\Pr_{H_0}\{X^n \in R\}]$ size

$\{X^n \notin R ; \theta \in \Theta_1\} = \Pr_{H_1}(X^n \notin R)$ 1-pwr

$\alpha = \Pr_{\Theta_0}\{X^n \in R\}$ pr (type 1 error)

$\beta = 1 - $ pr (type 2 error)

$Z = \hat{\beta} / \hat{se}$



Example: logistic regression

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.103704   6.530014  -5.223 1.76e-07 ***
zn           -0.079918   0.033731  -2.369  0.01782 *
indus        -0.059389   0.043722  -1.358  0.17436
chas          0.785327   0.728930   1.077  0.28132
nox          48.523782   7.396497   6.560 5.37e-11 ***
rm           -0.425596   0.701104  -0.607  0.54383
age           0.022172   0.012221   1.814  0.06963 .
dis           0.691400   0.218308   3.167  0.00154 **
rad           0.656465   0.152452   4.306 1.66e-05 ***
tax          -0.006412   0.002689  -2.385  0.01709 *
ptratio       0.368716   0.122136   3.019  0.00254 **
black        -0.013524   0.006536  -2.069  0.03853 *
lstat         0.043862   0.048981   0.895  0.37052
medv          0.167130   0.066940   2.497  0.01254 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value

$N(0,1)$

	reject H_0	don't
H_0	type 1 error	✓
H_1	✓	type 2 error

$H_0: \beta_5 = 0$ vs. $H_1: \beta_5 \neq 0$

if $p < .05$ ← common

then (*)

"statistically significant"

$Pr_{\beta_5=0}(|N(0,1)| > .607)$

$0 < p < .001 < p < .01 < p < .05 < p < .1 < p \; 1$

$p < .05$ if $|z| > 1.96$

$z = \hat{\beta} / \hat{se}$

... Example: logistic regression

```
Boston.glmnull <- glm(crim2 ~ 1, family = binomial, data = Boston)
anova(Boston.glmnull, Boston.glm)
Analysis of Deviance Table

Model 1: crim2 ~ 1
Model 2: crim2 ~ (crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + medv) - crim
  Resid. Df Resid. Dev Df Deviance
1       505     701.46
2       492     211.93 13   489.54

> pchisq(489.54, 13, lower.tail = F)
[1] 2.435111e-96
```

$H_0: \beta_{(1)} = \underset{\sim}{0}$

$\Theta = (\beta_0\ \underline{\beta_1 \cdots \beta_p}) = \{\beta_0, \beta_{(1)}\}$

crim2 ~ . - crim

resid. dev. 701

212

diff $\overset{\cdot}{\sim}$ χ^2_{13}

LRT (need to prove) (coming)

```
Boston.glmpart <- glm(crim2 ~ . - crim - indus - chas - rm - lstat,
                           data = Boston, family = binomial)
anova(Boston.glmpart, Boston.glm)
Analysis of Deviance Table
Model 1: crim2 ~ (crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + medv) - crim - indus - chas -
    rm - lstat
Model 2: crim2 ~ (crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + medv) - crim
  Resid. Df Resid. Dev Df Deviance
1       496     216.22
2       492     211.93  4   4.2891
> pchisq(4.2891, 4, lower.tail = F)
[1] 0.368292
```

$H_0: \beta_{(?)} = 0$

$H_0: \beta_{(?)} \neq 0$

$? = \subseteq \{1, \dots, p\}$

- Null and alternative hypothesis: $H_0 : \theta \in \Theta_0$; $H_1 : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$
- Rejection region: $R \subset \mathcal{X}$; if $\boldsymbol{x} \in R$ "reject" H_0

  $\alpha = \Pr_{H_0}\{X \in R\}$
- Test statistic and critical value: $R = \{x \in \mathcal{X} : t(x) > c\}$ c to be chosen
- Type I and Type II error: $\Pr\{t(X) > c \mid \theta \in \Theta_0\}$, $\Pr\{t(X) \leq c \mid \theta \in \Theta_1\}$
- Power and Size: $\beta(\theta) = \Pr_\theta(X \in R)$ $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$

  Def fn 10.1

  function of θ size
- Optimal tests: among all level-α tests, find that with the highest power under H_1

  of size α $\leq$

  level-α means size $\leq \alpha$

Neyman-Pearson '33

Example: Two-sample t-test

1.2 Hypothesis Testing

Our second example concerns the march of methodology and inference for *hypothesis testing* rather than estimation: 72 leukemia patients, 47 with **ALL** (acute lymphoblastic leukemia) and 25 with **AML** (acute myeloid leukemia, a worse prognosis) have each had genetic activity measured for a panel of 7,128 genes. The histograms in Figure 1.4 compare the genetic activities in the two groups for gene 136.

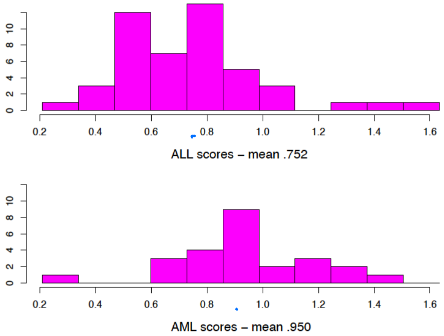


Figure 1.4 Scores for gene 136, leukemia data. Top **ALL** ($n = 47$), bottom **AML** ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

The **AML** group appears to show greater activity, the mean values being

$$\overline{\text{ALL}} = 0.752 \quad \text{and} \quad \overline{\text{AML}} = 0.950. \qquad (1.5)$$

$X_{1i} \sim N(\mu_1, \sigma^2)$

$X_{2i} \sim N(\mu_2, \sigma^2)$

gp 1 values

ALL

x_i "genetic activity" for patient i on gene ⟺ row (136)

gp 2

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

$\boxed{t(\underset{\sim}{X}) > c}$

when

$\mu_1 \neq \mu_2$

$\Pr_{H_0}\{t(\underset{\sim}{X}) > c\}$

$= \alpha$

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\hat{se}(\bar{x}_1 - \bar{x}_2)} = t$$

$s^2 = \frac{\sum(x_i-\bar{x})^2 + \sum(y_i-\bar{y})^2}{n_1+n_2-2}$ $s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

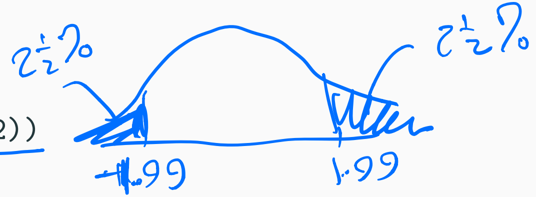
$\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}$

```
leukemia_big <- read.csv
  ("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
oneline <- leukemia_big[136,]
one <- c(1:20, 35:61) #  I had to extract these manually,
two <- c(21:34, 62:72) #  couldn't figure out the data frame
n1 <- length(one); n2 <- length(two)
mean_one <- sum(oneline[1,one])/n1.  ##[1] 0.7524794
mean_two <- sum(oneline[1,two])/n2.  ##[1] 0.9499731
var_one <-  sum((oneline[1,one]-mean_one)^2)/(n1-1)
var_two <- sum((oneline[1,two]-mean_two)^2)/(n2-1)
pooled <- ((n1-1)*var_one + (n2-1)*var_two)/(n1+n2-1)
taos <- (mean_one-mean_two)/sqrt((var_one/n1)+(var_two/n2))
##[1] -3.132304
tbe <- (mean_one-mean_two)/sqrt(pooled*((1/n1)+(1/n2)))
##[1] -3.035455
```

Pr_{H_0} (reject H_0)

$= Pr_{H_0}(|T_{70}| \geq c_\alpha)$

$c_\alpha = 1.99$



"reject H_0" at level .05

$n_1+n_2=72$ d.f. $[n_1+n_2-2 = 70]$

"p-value < 0.05" statistically sign. (at level .05)

"p-value" (2-sided) = .002

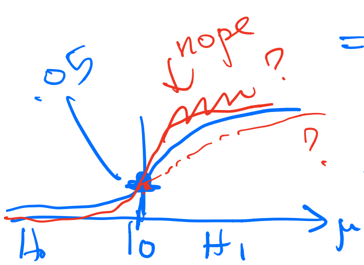
$X_1, \dots X_n$ iid $N(\mu, \sigma^2)$ $\underline{\sigma^2 \text{ known}}$

Let's use $R = \{\bar{X} > c\}$ sample mean

$Pr_{H_0}(\bar{X} > c) = \alpha$ under $H_0: \mu \le 0$

unif. most powerful $\qquad A_1: \mu \ge 0$

$\beta(\mu) = Pr_\mu\{\bar{X} > c\}$



$$= Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

$$= Pr\left(Z > \frac{c - \mu}{\sigma/\sqrt{n}}\right) = \alpha \qquad \equiv 0$$

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = \underline{\underline{-1.96}}$$

$\tilde{\mu}$ = median $(X_1, \dots, X_n)$ $\qquad Pr_{H_0}(\underbrace{\tilde{\mu} > c_\alpha}_{R}) = \alpha$

Example: Likelihood inference

AoS

$(\tilde{\theta}-\theta)/\widetilde{se} \equiv$ "Wald"

$X_1, \ldots, X_n$ i.i.d. $f(x;\theta)$; $\hat{\theta}(X_n)$ is maximum likelihood estimate. From last week:

$$(\hat{\theta}-\theta)/\widehat{se} \dot{\sim} N(0,1)$$

logistic regression

$\hat{se} = \sqrt{j^{-1}(\hat{\theta})}$

or $= \sqrt{I_n^{-1}(\hat{\theta})}$

To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ we could use

Wald

$$T = W = W(X_n) = (\hat{\theta}-\theta_0)/\widehat{se},$$

Wald means mle $\hat{\theta}$

The approx. critical region will be $\{\mathbf{x} : |W(\mathbf{x})| > z_{\alpha/2}\}$, i.e. "reject" H_0 when $|W| \geq z_{\alpha/2}$

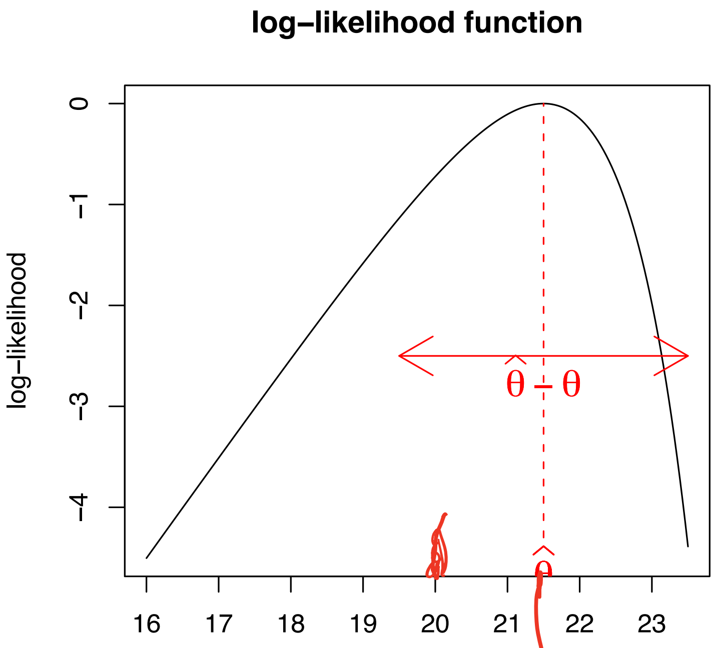
This test has approximate size α:

$\alpha = .05 \quad z_{\alpha/2} = 1.96$

$$\Pr_{H_0}(|W| > z_{\alpha/2}) \doteq \alpha.$$

$|W|$ "large" when H_0 is false.

Power? See Figure 10.1 and Theorem 10.6



$\ell(\theta; x)$ $\quad \hat\theta = 21.5$

$f(x;\theta) \overset{?}{=} e^{x-\theta-e^{(x-\theta)}}$

> 1.96 or < -1.96

$H_0: \theta = 20$

$\dfrac{(\hat\theta^{21.5} - 20)}{j^{1/2}(\hat\theta)} \sim N(0,1)$

$\boxed{\dfrac{21.5-20}{??}}$

$\hat{se}$

p-value =

Example: comparing two binomials

X ind't of Y

$X \sim Bin(n_1, p_1), \quad Y \sim Bin(n_2, p_2), \quad \underline{\delta = p_1 - p_2}, \quad H_0 : \delta = 0$

mle $\quad \hat{p}_1 = \frac{X}{n_1} \quad \hat{p}_2 = \frac{Y}{n_2} \quad \hat{\delta}_{mle} = \frac{X}{n_1} - \frac{Y}{n_2} = \hat{p}_1 - \hat{p}_2$

$\hat{se}(\hat{\delta}) = \sqrt{var(\hat{p}_1 + \hat{p}_2)} \leftarrow$ known

$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Normal approx?

?"Wald?" test at level α reject $H_0: \delta = 0$

$\boxed{\text{if } |\hat{\delta}| / \hat{se}(\hat{\delta}) > 1.96}$

equality of means; equality of medians; Wald test

power $\Pr_\delta\left\{\frac{\hat\delta}{\hat{se}(\hat\delta)} > 1.96\right\}$ $\longrightarrow$ N approx. to binom.

$$= \Pr_\delta\left\{\hat\delta > 1.96\cdot\hat{se}(\hat\delta)\right\}$$

$$= \Pr_\delta\left\{\hat\delta - \delta > 1.96\,\hat{se}(\hat\delta) - \delta\right\}$$

$\beta(\delta)$

$$= \Pr_\delta\left\{\frac{\hat\delta - \delta}{\hat{se}(\hat\delta)} > 1.96 - \frac{\delta}{\hat{se}(\hat\delta)}\right\}$$

10.9 $X_1,\dots,X_n$ iid $f_1(\cdot)$ $Y_1,\dots,Y_m$ iid $f_2(\cdot)$

$H_0: \text{med}_1 = \text{med}_2$ $H_1: \text{med}_1 \neq \text{med}_2$

$m_1 - m_2 = \Delta$

find a f= $T = t(X^n)$

know at least $T \underset{H_0}{\sim} :($ some ~~th~~ density $)$

"Reject" H_0 iff $\mathbb{P}_{H_0}(T \in R) \leq \alpha$

size ≈ 0.05

need T to be "sensitive" to H_1 $\mathbb{P}_{H_1}(T \in \{T \notin R\})$ sm.

The formal theory of testing imagines a decision to "reject H_0" or not, according as $X \in R$ or $X \notin R$, for some defined region R (e.g. $|Z| > 1.96$)

"at level .05" $\Pr_{H_0}(T\ 1)$

This is useful for deriving the form of optimal tests, but not useful in practice.

Doesn't distinguish between $Z = 1.97$ and $Z = 19.7$, for example.

$p = 0$

P-values give more precise information about the null hypothesis

AoS definition: p-value $= \inf\{\alpha : T(X_n) \in R_\alpha\}$ Def 10.11

$p \simeq .048$

SM definition $p_{obs} = \Pr_{H_0}\{T(X_n) \geq t_{obs}\}$

observed value if Ho true

"pr {getting a result as or more extreme than data, if Ho is true}"

$H_0 : \mu = 0$

$\cancel{\bar{x}}^{obs} = 7.$

$pr(\bar{X} \geq 7; \mu = 0) = Pr(N(0,1) > 7) \leq 10^{-?}$

(known $\sigma = 1$)

$t^{obs} = t(X_n) = \frac{1}{n}\Sigma(X_i)$

Example: exponential

$X_1, \dots X_n$ i.i.d. $\quad f(x;\lambda) = \lambda e^{-\lambda x}$ $\quad L(\lambda; \underset{\sim}{x}) = \lambda^n e^{-\lambda \sum x_i}$

$H_0 : \lambda = \lambda_0$

$\checkmark \; t(\underset{\sim}{X}_n) = \sum X_i \sim \Gamma(n, \lambda)$

$Pr_{H_0}(t(\underset{\sim}{X}_n) \geq t_{obs}) = \dots$

… Example: logistic regression

```
Boston.glmnull <- glm(crim2 ~ 1, family = binomial, data = Boston)
anova(Boston.glmnull, Boston.glm)
Analysis of Deviance Table

Model 1: crim2 ~ 1
Model 2: crim2 ~ (crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + black + lstat + medv) - crim
  Resid. Df Resid. Dev Df Deviance
1       505     701.46
2       492     211.93 13   489.54

> pchisq(489.54, 13, lower.tail = F)
[1] 2.435111e-96
```