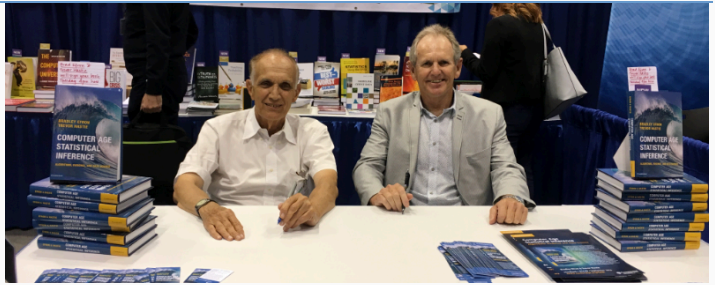
Mathematical Statistics II

STA2212H S LEC9101

Week 2

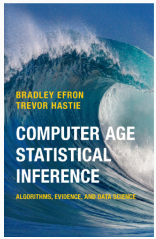
January 20 2021



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 — notes on likelihood
- score function, maximum likelihood estimate, observed and expected Fisher information
- asymptotic normality of maximum likelihood estimators — $\sqrt{n}(\hat{\theta} - \theta) I_1^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, 1)$
- estimating the asymptotic variance — $j(\hat{\theta})$, $I_n(\hat{\theta})$
- the delta method — $\tau = g(\theta)$
- profile likelihood — see notes p.6
- sufficient statistics
- Newton-Raphson method for computing $\hat{\theta}$
- irregular models — $U(0, \theta)$
- Quasi-Newton
- EM Algorithm — Friday

Today

Start Recording

1. Pareto MLE; Quasi-Newton — `Pareto.Rmd`
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:
- Iterate:
- Rewrite:
- Quasi-Newton:
-
-

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

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: $j(\hat{\theta}^{(t)})(\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}) = \ell'(\hat{\theta}^{(t)})$ $B\Delta\theta = -\nabla\ell(\theta)$
- Quasi-Newton:
  - approximate $j(\hat{\theta}^{(t)})$ with something easy to invert
  - use information from $j(\hat{\theta}^{(t)})$ to compute $j(\hat{\theta}^{(t+1)})$
- optimization notes add a step size to the iteration $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \epsilon_t\{j(\hat{\theta}^{(t)})\}^{-1}\ell'(\hat{\theta}^{(t)})$

```
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

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

Formal theory of testing

AoS 10.1

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

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
```

... 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
```

... Example: logistic regression

```
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
```

Formal theory of testing

- Null and alternative hypothesis: $H_0 : \theta \in \Theta_0$; $H_1 : \theta \in \Theta_1$, $\quad \Theta_0 \cup \Theta_1 = \Theta$
- Rejection region: $R \subset \mathcal{X}$; if $\boldsymbol{x} \in R$ "reject" H_0
- Test statistic and critical value: $R = \{x \in \mathcal{X} : t(x) > c\}$ — c to be chosen
- Type I and Type II error: $\Pr\{t(X) > c \mid \theta \in \Theta_0\}$, $\quad \Pr\{t(X) \leq c \mid \theta \in \Theta_1\}$
- Power and Size: $\beta(\theta) = \Pr_\theta(X \in R)$ $\quad \alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$
- Optimal tests: among all level-α tests, find that with the highest power under H_1

level-α means size $\leq \alpha$

Example: Two-sample t-test

EH §1.2

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.

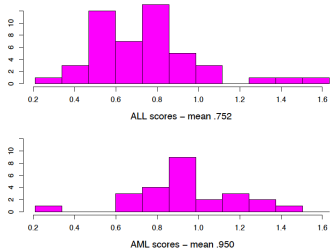


Figure 1.4 Scores for gene 136, leukemia data. Top **ALL** ($n = 47$), bottom **AML** ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

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

$$\overline{\mathbf{ALL}} = 0.752 \quad \text{and} \quad \overline{\mathbf{AML}} = 0.950. \tag{1.5}$$

```
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
```

Example: Likelihood inference

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

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

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

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

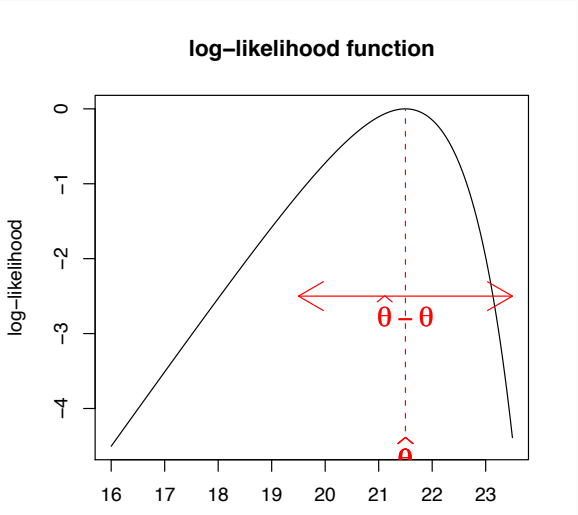
The critical region will be $\{\boldsymbol{x} : |W(\boldsymbol{x})| > z_{\alpha/2}\}$, i.e. "reject" H_0 when $|W| \geq z_{\alpha/2}$

This test has approximate size α:

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

Power? See Figure 10.1 and Theorem 10.6

... likelihood inference



Example: comparing two binomials

AoS Ex.107

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

Examples: 10.8 and 10.9

equality of means; equality of medians; Wald test

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. $Z > 1.96$)

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-values give more precise information about the null hypothesis

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

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

Example: exponential

SM Ex.7.22

$X_1, \ldots X_n$ i.i.d. $\qquad f(x;\lambda) = \lambda e^{-\lambda x}$

$H_0 : \lambda = \lambda_0$

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
```

$\longrightarrow$ Monash talk

Likelihood ratio tests

AoS 10.6