Chapter 9

LECTURE REVIEWS FOR 1996
LECTURE #1, 1998

- Taylor’s Theorem:
  \[ g(x) = g(x_0) + g^{(1)}(x_*)(x-x_0); \]
  \[ g(x) = g(x_0) + g^{(1)}(x_0)(x-x_0) + (1/2)(x-x_0)^2 g^{(2)}(x_*). \]

- Slutsky’d Lemma
  \[ V_n \Rightarrow V \]
  \[ a_n \rightarrow a \text{ in probability} \]
  \[ b_n \rightarrow b > 0 \text{ in probability} \]
  then \( (V_n - a_n)/b_n \Rightarrow (V - a)/b \)

- Delta–Method
  \[ n^{1/2}(V_n - \theta) \Rightarrow \text{Normal}(0, \sigma^2) \text{ then} \]
  \[ n^{1/2}(g(V_n) - g(\theta)) \Rightarrow \text{Normal} \left[ 0, \sigma^2 \left\{ g^{(1)}(\theta) \right\}^2 \right]. \]

- Variance Stabilization (Poisson, \( \chi^2 \))

- Alternating Conditional Expectations
  \[ E \{ g(Y, X) \} = E \left[ E \{ g(Y, X)|X \} \right] \]
  \[ \text{var}(Y) = E \{ \text{var}(Y|X) \} + \text{var} \left\{ E(Y|X) \right\} \]

- Mean Squared Error Theorem
  \[ c = E(Y) \text{ minimizes } E(Y - c)^2; \]
  \[ c(X) = E(Y|X) \text{ minimizes } E \left\{ Y - c(X) \right\}^2; \]

- Regression when \((X_1, X_2)\) are bivariate normal
  \[ [X_2|X_1] = \text{Normal} \left\{ \mu_2 + (\rho \sigma_2/\sigma_1)(X_1 - \mu_1), \sigma_2^2(1 - \rho^2) \right\}. \]

- A statistic \( T(\mathbf{X}) \) is sufficient for a parameter \( \Theta \) based upon a random variable \( \mathbf{X} \) if the distribution of \( T(\mathbf{X}) \) given \( \mathbf{X} \) is independent of \( \Theta \).
LECTURE #2, 1998

- Neyman–Fisher Factorization Theorem. A statistic $T(\mathbf{X})$ is sufficient for $\Theta$ if there exists functions $g(t, \Theta)$ and $h(\mathbf{X})$, the latter functionally independent of $\Theta$, such that the joint density or mass function of $\mathbf{X}$ satisfies

$$f(\mathbf{x}|\Theta) = g\{T(\mathbf{x}), \Theta\} h(\mathbf{x}).$$

- The proof was a direct calculation.

- We checked a few examples. In particular, if $X_1, ..., X_n$ are iid Normal($\mu, \sigma^2$), then the joint sufficient statistic for $(\mu, \sigma^2)$ is $(\overline{X}, s^2)$, the sample mean and variance.

- For scalar $\theta$, a random variable $\mathbf{X}$ is in the one parameter exponential family (OPEF) if its density or mass function can be written as

$$f(\mathbf{x}|\theta) = S(\mathbf{x})\exp\{c(\theta)T(\mathbf{x}) + d(\theta)\}$$

Neither $S(\mathbf{x})$ nor $T(\mathbf{x})$ can depend functionally on $\theta$. The sufficient statistic for $\theta$ is $T(\mathbf{X})$.

- $c(\theta)$ is the natural parameter.

- For the Bernoulli($\theta$) case, the natural parameter is the logit,

$$c(\theta) = \log \left( \frac{\theta}{1 - \theta} \right),$$

- We checked that the Normal($\mu, \sigma^2$ known) is OPEF with natural parameter $\mu$ and $T(\mathbf{X}) = n\overline{X}/\sigma^2$. 
LECTURE #3, 1998

- We showed that if $X$ is OPEF with sufficient statistic $T(X)$, then $T(X)$ is also OPEF.

- From this we showed that if we used the natural parameterization, so that $\eta = c(\theta)$ and $d_0(\eta) = d(\theta)$, then

\[
\begin{align*}
  f(x|\eta) &= S(x)\exp \{T(x)\eta + d_0(\eta)\}; \\
  E[\exp \{sT(X)\}] &= \exp \{d_0(\eta) - d_0(\eta + s)\}; \\
  E \{T(X)\} &= -d_0^{(1)}(\eta); \\
  \text{var} \{T(X)\} &= -d_0^{(2)}(\eta);
\end{align*}
\]

- We used this result to give an easy calculation of the mean and variance of the Poisson distribution.

- We also used the result in the case that $X_1, \ldots, X_n$ are iid Beta($\alpha, \beta$) with $\beta$ known. The sufficient statistic was shown to be $T(X) = \sum_{i=1}^{n} \log(X_i)$.

- We defined consistency: a statistics $S_n$ is consistent for $\Theta$ if it converges to $\Theta$ in probability.

- In the Beta($\alpha, \beta$) with $\beta$ known problem, a consistent estimate of $\alpha$ was shown to be

\[
S_n = \frac{\beta X}{1 - X}.
\]

- We used the Delta–Method to show that

\[
n^{1/2}(S_n - \alpha) \Rightarrow \text{Normal} \left\{ 0, \frac{\alpha (\alpha + \beta)^2}{\beta (\alpha + \beta + 1)} \right\}.
\]
LECTURE #4, 1998

- We took up the problem of estimation using the method of moments.

- The first example was the $2 \times 2$ table, i.e., a multinomial with 4 calls and cell probabilities $p_{ij}$. The goal was to estimate $p_{11}$. If the cell counts from a sample of size $n$ are $N_{ij}$, the obvious estimate is $\hat{p}_{11} = N_{11}/n$. This has mean $p_{11}$ and variance $p_{11}(1-p_{11})/n$.

- If I assumed however that the rows and columns were independent, then $p_{11} = p_1 p_1$, and so a sensible estimate is $\hat{p}_{11,b} = (N_1/n)(N_1/n)$. I asked you to show that this estimator is a function of the sufficient statistic under the assumption of independence.

- We then considered the Hardy–Weinberg model from genetics, with cell probabilities $\theta^2, 2\theta(1-\theta)$ and $(1-\theta)^2$. The cell counts are $(N_1, N_2, N_3 = n - N_1 - N_2)$. The data are multinomial.

- Obvious estimates of $\theta$ are $\hat{\theta}_a = \sqrt{N_1/n}$ and $\hat{\theta}_b = 1 - \sqrt{N_3/n}$. We used the delta method and showed that the limiting distributions of these estimates has variance

\[
\begin{align*}
    n^{1/2}(\hat{\theta}_a - \theta) &\Rightarrow \text{Normal} \left(0, \frac{1-\theta^2}{4}\right); \\
    n^{1/2}(\hat{\theta}_b - \theta) &\Rightarrow \text{Normal} \left(0, \frac{1-(1-\theta)^2}{4}\right).
\end{align*}
\]

- We showed that the sufficient statistic was $T(\mathbf{X}) = 2N_1 + N_2$, and that this is a member of the OPEF with $c(\theta) = \text{logit}(\theta)$ and $d(\theta) = 2n \log(1-\theta)$. This is the same form as if $T(\mathbf{X})$ were the sufficient statistic for a binomial sample with size $2n$, and hence if $\hat{\theta}_c = (2N_1 + N_2)/(2n)$,

\[
\begin{align*}
    \text{var} \{T(\mathbf{X})\} &= \frac{\theta(1-\theta)}{2n}; \\
    n^{1/2}(\hat{\theta}_c - \theta) &\Rightarrow \text{Normal} \left(0, \frac{\theta(1-\theta)}{2}\right).
\end{align*}
\]

- The last claim is not proved yet, but does make sense.
LECTURE #5, 1998

• We computed the mle for $\Theta$ in a number of special cases.

• Normal($\mu, \sigma^2$)

• Binomial($N, \theta$), first with $\theta$ known (a nonregular family) and then with $N$ known (a regular family).

• Uniform[0, $\theta$] (a nonregular family)

• Linear regression through the origen.

• We also computed the limiting distribution of $s^2$, i.e., showed that for normally distributed data, $n^{1/2}(s^2 - \sigma^2) \Rightarrow \text{Normal}(0, 2\sigma^4)$.

• We write the likelihood as $L(X|\Theta)$ and the loglikelihood as $\ell(X|\Theta)$.

• In regular families, one usual maximizes the loglikelihood. Differentiate with respect to the components of $\Theta$ and solve for a zero. This is called an estimating equation. The solution is a local maximum if the second derivative of the loglikelihood evaluated at the MLE is negative definite.
LECTURE #6, 1998

- We first showed a simple characterization of the MLE in a OPEF: Let \( c(\theta) \) be 1–1, continuously differentiable, and suppose that \( \text{var}_\theta \{ T(X) \} > 0 \) on the support of \( X \), for all \( \theta \). Then if the MLE exists, it is unique and satisfies

\[
E_\theta \{ T(X) \} = T(X).
\]

- The result was shown in a series of steps:

1. The loglikelihood is \( \eta T(X) + d_0(\eta) + \text{stuff} \)

2. Its derivative is \( T(X) + d_0^{(1)}(\eta) \)

3. Its second derivative is \( d_0^{(2)}(\eta) \)

4. But we know the mean and variance of \( T(X) \) are \( -d_0^{(1)}(\eta) \) and \( -d_0^{(2)}(\eta) \), respectively.

5. Since the second derivative of the loglikelihood is negative, any solution to the first equation maximizes the loglikelihood, and there can be only one minimizer.

- We defined the mle of a function \( g(\theta) \), and showed it to be \( g(\hat{\theta}_{ml}) \).

- We then began decision theory as a way to compare estimators and to judge their value.

- A loss function for estimating \( \theta \) by \( s \) is \( L(\theta, s) \)

- The risk function of an estimator \( S(X) \) is

\[
R(\theta, S) = E_\theta \left[ L(\theta, S(X)) \right].
\]

- We showed that uniformly best estimators in terms of risk cannot exist. For example, suppose \( L(\theta, s) = (s - \theta)^2 \), and that \( X \sim \text{Normal}(\theta, 1) \). Then you cannot beat the estimator \( S(X) = 17 \) if by chance \( \theta = 17 \).
LECTURE #7, 1998

• An estimator \( S(\mathbf{X}) \) is minimax if

\[
\max_\theta R(\theta, S) = \min_S \max_\theta R(\theta, S).
\]

• An estimator \( S_*(\mathbf{X}) \) is admissible if there exists no other estimator \( S(\mathbf{X}) \) with the property that

\[
R(\theta, S) \leq R(\theta, S_*) \quad \forall \theta
\]

with at least one strict inequality somewhere.

• Bayes estimators with respect to a prior \( \pi(\theta) \) minimize \( \int R(\theta, S)\pi(\theta)d\theta \). If \( \pi(\theta) \) is a density function, then we can write the conditional density of \( \theta \) given \( \mathbf{X} \) as \( \pi(\theta|\mathbf{X}) \). The Bayes estimator then minimizes in \( s \)

\[
\int L(\theta, s)\pi(\theta|\mathbf{X})d\theta.
\]

• We considered the case that \( X \sim \text{Normal}(\theta, 1) \), and looked at the class of estimators \( cX \) with \( 0 < c \leq 1 \). We showed that only \( c = 1 \) is minimax. I asked you to argue that all values of \( c \) are admissible. I promised that later I would show that only \( 0 < c < 1 \) are Bayes estimators with respect to proper prior densities.

• We showed that if \( L(\theta, s) = (s - \theta)^2 \), then any estimator \( S(\mathbf{X}) \) which is not a function of the sufficient statistic \( T(\mathbf{X}) \) can be improved in terms of risk by using instead \( E \{ S(\mathbf{X})|T(\mathbf{X}) \} \).

• Jensen’s inequality says that if \( g(x) \) is a convex function, and if \( \mathbf{X} \) has its support on a convex set, then

\[
g \{ E(\mathbf{X}) \} \leq E \{ g(\mathbf{X}) \}.
\]

The inequality is strict if \( g(x) \) is strictly convex and \( \mathbf{X} \) does not equal a single value with probability 1.0.
• The Rao–Blackwell Theorem says that if the loss function $L(\theta, s)$ is convex in $s$ for each $\theta$, and if $S(X)$ is any estimator, and if the support of $T(X)$ is convex, then

$$E_\theta [L \{ \theta, G(X) \}] \leq E_\theta [L \{ \theta, S(X) \}];$$

$$G(x) = E[S(X)|T(X)].$$

The inequality is strict if $g(x)$ is strictly convex and the distribution of $T(X)$ is not concentrated at a point.

• The proof of Rao–Blackwell consisted of the following steps:

1. First remember that

$$L \{ \theta, G(X) \} = L[\theta, E[S(X)|T(X)]].$$

2. Now apply Jensen’s inequality to the conditional distribution of $X$ given $T(X)$ to note that

$$L \{ \theta, G(X) \} \leq E[L \{ \theta, S(X) \} | T(X)].$$

3. Now take expectations.

• We finished with the problem of estimating $\theta$ from a sample of size $n$ from $\text{Uniform}[0, \theta]$. We considered squared error loss, and two estimators: (a) $S_1(X) = 2X$ and $S(X) = X_{(n)}$. We showed that

$$R(\theta, S_1) = \frac{\theta^2}{3n};$$

$$R(\theta, S_2) = \frac{\theta^2}{(n + 2)(n + 1)}.$$ 

This, the function of the sufficient statistic always wins in terms of risk.

• We noted that in this example, Rao–Blackwellization was not so obvious, because it involved computing $E(2\bar{X}|X_{(n)})$. While I did not state this, we will show in the next lecture that in fact $E(2\bar{X}|X_{(n)}) = (n + 1)X_{(n)}/n$. 


LECTURE #8, 1998

- I redid the proof of Rao–Blackwell.

- $S(X)$ is unbiased for $q(\theta)$ if for all $\theta$, $E_\theta \{S(X)\} = q(\theta)$.

- An estimator of $q(\theta)$ is UMVUE if it is unbiased and if it has uniformly minimum variance among all unbiased estimators.

- This means that UMVUE uses squared error loss, and seeks to minimize risk, but it does so artificially because it restricts attention only to unbiased estimators.

- UMVUE’s have many disadvantages. They may not exist, they are not an algorithm, they are not invariant, they can be beaten.

- We defined a complete sufficient statistic $T(X)$ for $\theta$ as one for which

  $$E_\theta \{g(T)\} = 0 \ \forall \theta \Rightarrow g(T) \equiv 0.$$ 

- We then showed that if there were a complete sufficient statistic, and if $S(X)$ is unbiased for $q(\theta)$, then Rao–Blackwellization results in the unique UMVUE.

- We showed that $X_{(n)}$ is complete in the Uniform$[0, \theta]$ problem, and hence that $E(2\bar{X}|X_{(n)}) = (n + 1)X_{(n)}/n$.

- $k$–parameter exponential families are generally complete, if the range of $c_1(\Theta), \ldots, c_k(\Theta)$ contains an open, non–trivial $k$–rectangle.

- We did a few examples, mostly to convince you that this is a rather strange concept, and hard to implement. The special cases of Normal$(\theta, \theta)$ and negative exponential with mean $\theta$ and interest in estimating $q(\theta) = \exp(-t/\theta)$ were highlighted.

- We then began the work towards the information (Cramer–Rao) inequality by looking at Cauchy–Schwarz.
Cauchy–Schwartz says that

\[ \text{cov} \{S_1(X), S_2(X)\}^2 \leq \text{var} \{S_1(X)\} \text{var} \{S_2(X)\}. \]

Equality occurs if and only if the two statistics are linear functions of one another.

The Fisher information in a sample is

\[ I(\theta) = E \left( \left[ \frac{\partial}{\partial \theta} \log f(X|\theta) \right]^2 \right). \]

If the “sample” consists of a single observation, we will write the Fisher information as \( I_s(\theta) \).

If \( X \) consists of a random sample of \( n \) observations, then \( I(\theta) = nI_s(\theta) \).

(Cramer–Rao inequality). Let \( S(X) \) be any statistic with a finite variance for all \( \theta \), and with expectation \( \psi(\theta) \). Then,

\[ \text{var}_\theta \{S(X)\} \geq \frac{\left( \psi^{(1)}(\theta) \right)^2}{I(\theta)} \]

under the regularity conditions:

- the values of \( \theta \) form an open subset of the real line;
- \( f(x|\theta) \) has continuous first and second derivatives in \( \theta \).
- The support of \( X \) does not depend on \( \theta \)
- One can interchange derivatives and integrals at will.
As part of the proof, we showed the fundamental results:

\[
E \left[ \frac{\partial}{\partial \theta} \log \{f(\mathbf{X}|\theta)\} \right] = 0
\]

\[
\text{cov} \left[ \mathbf{S}(\mathbf{X}), \frac{\partial}{\partial \theta} \log \{f(\mathbf{X}|\theta)\} \right] = \psi^{(1)}(\theta).
\]

We also showed that under suitable regularity conditions,

\[
\mathcal{I}(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log \{f(\mathbf{X}|\theta)\} \right].
\]

We considered the \(\mathcal{P}(\theta)\) model, and showed that the sufficient statistic achieved the Cramer–Rao lower bound.

We considered the case of independent but not identically distributed Poisson observations with means \(\exp(\theta z_i)\), and computed the Cramer–Rao lower bound for this problem.

The intention was to show you how general the inequality is: \(\mathbf{X}\) need not be an iid sample.
LECTURE #10, 1998

- We showed that if $S(X)$ achieves the Cramer–Rao lower bound, then necessarily $X$ is a member of the OPEF with sufficient statistic $S(X)$. The reason for this is that

$$\text{cov} \left[ S(X), \frac{\partial}{\partial \theta} \log \{ f(X|\theta) \} \right] = \psi^{(1)}(\theta),$$

and the Cauchy–Schwarz inequality is exact only when the two terms in the covariance are linear functions of one another.

- We also showed that the sufficient statistic in a OPEF achieves the Cramer–Rao lower bound if $c(\theta)$ is 1–1, if $c^{(1)}(\theta) \neq 0$ and if $c(\theta)$ and $d(\theta)$ are twice continuously differentiable (these conditions might be weakened, but they are sufficient).

- Not all families achieve the C–R lower bound. Let $X \sim \text{Normal}(\theta, 1)$. Consider estimating $\psi(\theta) = \theta^2$.

  - The C–R bound says that any unbiased estimator of $\psi(\theta)$ must have variance at least $4\theta^2$.
  - The UMVUE is $X^2 - 1$.
  - Direct calculations show that $\text{var}(X^2 - 1) = 4\theta^2 + 2$.
  - Note that $X$ is OPEF with sufficient statistic $X$, and thus it is only $X$ which achieves the C–R lower bound, and not any other nonlinear function of $X$.

- Class was cancelled because of a power failure.
LECTURE #11, 1998

- We proved consistency of the mle under the basic condition that the likelihood function of the data was unimodal with a unique maximum for every $n$.

- The result is essentially based on proving that
  \[ E_{\theta_t} \left[ \frac{\partial}{\partial \theta} \log \{ f(X|\theta) \} \right] \]
  is maximized at $\theta = \theta_t$. This was shown by using Jensen’s inequality, since $-\log(y)$ is a strictly convex function.

- We showed that under the conditions that $c(\theta)$ was strictly increasing (or decreasing), that $c(\theta)$ and $d(\theta)$ are twice continuously differentiable, and that $\text{var}_\theta \{ T(X) \} > 0$ for all $\theta$, that the mle was consistent since the log-likelihood function is strictly concave.

- Finally, we considered that Normal($\theta, \theta^2$) family for $\theta > 0$. There is a unique maximum of the loglikelihood for $\theta > 0$, and this loglikelihood is continuous, so that the loglikelihood is unimodal (with a unique maximum). Hence the conditions of Wald’s consistency proof apply, and the MLE is consistent. This can be shown directly though, using the WLLN.
LECTURE #12, 1998

• We showed that in iid sampling, with the information number given by \( I(\theta) \), that if \( \hat{\theta}_{ml} \) was a consistent estimator of \( \theta_t \), then

\[
n^{1/2} (\hat{\theta}_{ml} - \theta_t) \Rightarrow \text{Normal} \left\{ 0, \frac{1}{I(\theta)} \right\}.
\]

• What made this all work was a 1-term Taylor expansion, leading to the relationship that for \( \theta_* \) between the mle and \( \theta_t \),

\[
n^{1/2} (\hat{\theta}_{ml} - \theta_t) = -\frac{n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \{ f(X_i|\theta_t) \}}{n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log \{ f(X_i|\theta_t) \}}
\]

• We also noted that we could estimate \( I(\theta) \) consistently by

\[
\hat{I}_*(\hat{\theta}_{ml}) = -n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log \{ f(X_i|\hat{\theta}_{ml}) \}
\]

• This means (why?) that

\[
\frac{n^{1/2} (\hat{\theta}_{ml} - \theta_t)}{I(\hat{\theta}_{ml})} \Rightarrow \text{Normal}(0,1).
\]

• We then took up the problem that the \( X \)'s are normally distributed with mean \( \theta > 0 \) and variance \( \theta^2 \). The information number is \( I_*(\theta) = 3/\theta^2 \), and the mle is consistent. This tells us automatically the asymptotic distribution of the mle.

• On the other hand, direct computation of the limit distribution of the mle is complicated, because the mle is a function of the sample mean and variance.

• We also indicated that by Cramer–Rao type reasoning, what we have really shown is that the mle is asymptotically efficient, in the sense that no other estimator can have a limiting distribution which has a smaller variance than the limiting distribution of the mle.
Chapter 9. Lecture Reviews for 1996

Lecture #13, 1998

- We considered Bayesian estimation, with a prior \( \pi(\Theta) \).

- We showed using the Factorization Theorem that the Bayes estimator is a function of the (minimal) sufficient statistic.

- We showed that the posterior for \( \Theta \) given \( X \) is

\[
\pi(\Theta|X) = \frac{f(X|\Theta)\pi(\Theta)}{\int f(X|v)\pi(v)dv} \propto f(X|\Theta)\pi(\Theta). \tag{9.1}
\]

- For example, if \( X|\theta \sim \text{Binomial}(n, \theta) \), and if \( \pi(\theta) \) is \( \text{Beta}(a,b) \), then

\[
\pi(\theta|X) \propto \theta^{a+x-1}(1-\theta)^{n-x+b-1}
\]

and hence \( [\theta|X] \) is \( \text{Beta}(a + X, n - X + b) \). The posterior mean is

\[
E(\theta|X) = \frac{a + X}{n + a + b}.
\]

- Note that the posterior mean does not equal the mle unless \( a = b = 0 \), in which case the prior is not a proper density function.

- Note further though that the posterior mode is

\[
\text{mode}[\theta|X] = \frac{X + a - 1}{n + a + b - 2}.
\]

Thus, the mle is the posterior mode when \( a = b = 1 \), the uniform prior.

- This shows a general fact: the mle is the posterior mode when \( \pi(\theta) \) is constant in \( \theta \).

- We also consider the case of iid normal sampling where \( X|\theta \sim \text{Normal}(\theta,1) \), and with a normal prior with mean \( \theta_0 \) and variance \( \sigma_0^2 \).

- One can do brute force calculation to show that

\[
[\theta|\bar{X}] \sim \text{Normal}\left\{ w\bar{X} + (1-w)\theta_0, w/n \right\};
\]

\[
w = \frac{\sigma_0^2}{\sigma_0^2 + 1/n}.
\]
• I asked you to check this, by formally manipulating (9.1).

• Note that as \( n \to \infty \), the posterior mean and mode both look more and more like \( \bar{X} \) since \( w \to 1 \). The posterior mean and mode are biased estimate of \( \theta \) in the frequentist sense, but they have smaller variance than the mle.

• Again, if we set \( \sigma_0^2 = \infty \), then the prior is constant (improper) and the posterior mean and mode are the mle.

• The normal calculation can be done without brute force, using a simple idea which has applications in many other contexts.

• We can write

\[
\theta = \theta_0 + \epsilon_1,
\]

where \( \epsilon_1 \sim \text{Normal}(0, \sigma_0^2) \).

• We can also write

\[
[\bar{X}|\theta] = \theta + \epsilon_2 = \theta_0 + \epsilon_1 + \epsilon_2,
\]

where \( \epsilon_2 \sim \text{Normal}(0, 1/n) \) is independent of \( \epsilon_1 \).

• From this it is easy to see that \( \bar{X}, \theta \) have a joint bivariate normal distribution with common mean \( \theta_1 \), covariance \( \sigma_0^2 \) and variances \( \sigma_0^2 + 1/n \) and \( \sigma_0^2 \).

• The posterior distribution for \( [\theta|\bar{X}] \) then follows from standard bivariate normal calculations.

• I asked you to check this too.
LECTURE #14, 1998

- We had a lot of toil and trouble concerning the following result. Suppose that the data follow a OPEF with sufficient statistics \( T(\mathbf{X}) \) and parameterization \( c(\theta) \) and \( d(\theta) \). Let

\[
\pi(\theta) \propto \exp\{a_1c(\theta) + a_2d(\theta)\}
\]

with the restriction that the prior cannot depend on \( n \). Then

\[
\pi(\theta|\mathbf{X}) \propto \exp\left\{\{T(\mathbf{X}) + a_1\}c(\theta) + (a_2 + 1)d(\theta)\right\}.
\]

- In this case, since the prior and the posterior fall in the same family, the prior is said to be a conjugate prior.

- We fumbled around with the case of iid random sampling from the Negative Exponential with mean \( 1/\theta \). Here we have \( c(\theta) = -\theta, \quad d(\theta) = n\log(\theta), \) and \( T(\mathbf{X}) = n\bar{X}. \)

- The conjugate prior then is

\[
\pi(\theta) \propto \exp\left\{-a_1\theta + a_2n\log(\theta)\right\}.
\]

To make this prior independent of \( n \), we set \( a_2 = b_2/n \). In this case,

\[
\pi(\theta) \propto \exp\left\{-a_1\theta + b_2\log(\theta)\right\} = \theta^{(b_2+1)-1}\exp(-a_1\theta),
\]

which is a Gamma random variable with mean \( (b_2+1)/a_1 \).

- Because the prior is conjugate, the posterior is

\[
\pi(\theta|\mathbf{X}) \propto \exp\left\{-(a_1 + n\bar{X})\theta + (b_2 + n)\log(\theta)\right\},
\]

which is also Gamma with mean \( (b_2 + n + 1)/(a_1 + n\bar{X}) \).

- We also showed that the Gamma is the conjugate prior for the Poisson distribution.

- We reviewed some facts from earlier in the semester.

- A loss function for estimating \( \theta \) by \( s \) is \( L(\theta, s) \)
• The risk function of an estimator $S(X)$ is

$$R(\theta, S) = E_{\theta} [L \{\theta, S(X)\}] .$$

• Bayes estimators with respect to a prior $\pi(\theta)$ minimize $\int R(\theta, S) \pi(\theta) d\theta$. If $\pi(\theta)$ is a density function, then we can write the conditional density of $\theta$ given $X$ as $\pi(\theta|X)$. The Bayes estimator then minimizes in $s$

$$\int L(\theta, s) \pi(\theta|X) d\theta .$$

• We showed that if $L(\theta, s) = (s - \theta)^2$, then the Bayes estimator is $E\theta|X)$. 
LECTURE #15, 1998

- We proved that if an estimator is constant risk and if it is Bayes against any prior, then it is also minimax.

- We computed the Bayes estimator in the Binomial problem with a conjugate prior and the loss function \( L(\theta, s) = (s - \theta)^2/\{\theta(1 - \theta)\} \).

- We showed in this problem that the mle is Bayes, that it has constant risk, and that it is hence also minimax.
LECTURE #16, 1998

- Exam #1 was given
LECTURE #17, 1998

- We began the problem of testing whether a parameter $\theta$ is in one of two sets, call then $\mathcal{Q}_0$ and $\mathcal{Q}_1$.

- We showed that for $0 - -1$ loss functions, the Bayes procedure for making this decision is to find the set with the greater posterior probability.

- We applied this to the binomial problem.

- We also showed that the risk function can be characterized as

\[
R(\theta, S) = \text{pr(“reject” $\mathcal{Q}_0$ when $\theta \in \mathcal{Q}_0$)};
\]

\[
R(\theta, S) = \text{pr(“accept” $\mathcal{Q}_0$ when $\theta \in \mathcal{Q}_1$)}.
\]

- We defined Type I and Type II errors probabilities.

- We defined the power function as the probability of rejecting that $\theta \in \mathcal{Q}_0$, and showed that it is impossible to minimize power in $\mathcal{Q}_0$ and maximize it in $\mathcal{Q}_1$.

- This led us to the idea of restricting the number of possible procedures $S(X)$ to those whose power function does not exceed a predetermined number $\alpha$ for all $\theta \in \mathcal{Q}_0$. 


A decision rule $S(X)$ is a random variable taking on the values 0 and 1.

A randomized decision rule is one for which, given $X$, $S(X)$ is not specified exactly. This is useful when $X$ is discrete.

The overall level of a test $S(X)$ is

$$\sup_{\theta \in \Theta} \Pr\{S(X) = 1 = \text{mistake}\}.$$ 

A test $S_*(X)$ is UMP of level $\alpha$ if it has overall level $\leq \alpha$ and if, for any other test $S(X)$ with level $\leq \alpha$,

$$\Pr_\theta\{S(X) = 1\} \leq \Pr_\theta\{S_*(X) = 1\} \forall \theta \in \Theta_1.$$ 

The likelihood ratio statistic for comparing $\theta_0$ to $\theta_1$ is defined as

$$\mathcal{L}(x; \theta_0, \theta_1) = \frac{f(x|\theta_1)}{f(x|\theta_0)}.$$ 

The Neyman Pearson test statistic is defined by

$$S_{NP}(X) = 1 \text{ w.p. 1 if } \mathcal{L}(x; \theta_0, \theta_1) > \kappa;$$
$$S_{NP}(X) = 1 \text{ w.p. } \gamma \text{ if } \mathcal{L}(x; \theta_0, \theta_1) = \kappa;$$
$$S_{NP}(X) = 1 \text{ w.p. 0 if } \mathcal{L}(x; \theta_0, \theta_1) < \kappa.$$ 

We proved the Neyman–Pearson Lemma. Consider any Neyman–Pearson test statistic, and let its overall level equal $\alpha$ for comparing $\Theta_0 = \{\theta_0\}$ to $\Theta_1 = \{\theta_1\}$. Then any other test of level $\leq \alpha$ has power $\leq$ that of the Neyman–Pearson test.

We constructed the Neyman–Pearson tests for some special cases.
Lectures #19–#20, 1998

- A (possibly randomized) decision rule or test $S(X)$ takes on the values 0 and 1. $S(X) = j$ means that you think $\theta \in \Theta_j$.

- The level of a test is

$$\max_{\theta \in \Theta_0} \Pr_\theta \{ S(X) = 1 \}.$$ 

- A test $S_{\text{ump}}(X)$ is said to be uniformly most powerful (UMP) of level $\alpha$ if it is a level $\alpha$ test and if for any test $S_*(X)$ of level $\leq \alpha$,

$$\Pr_\theta \{ S_{\text{ump}}(X) = 0 \} \leq \Pr_\theta \{ S_*(X) = 0 \} \text{ for all } \theta \in \Theta_1.$$ 

- A simple hypothesis is one in which $\Theta$ has a single value.

- We did a series of examples, which had in common that monotone transformations of the likelihood ratio statistics were made so that the Neyman-Pearson test was in the form

some random variable $> c$,

and then we computed $c$ to make the test of level $\alpha$.

- From the proof of the Neyman-Pearson Lemma, we can show that Neyman-Pearson tests are unique: if a test for simple null versus simple alternative is UMP of its level, it is a Neyman-Pearson test.

- $f(x|\theta)$ has a monotone likelihood ratio (MLR) in $T(x)$ if for any $\theta_0 < \theta_1$, the likelihood ratio $L(x; \theta_0, \theta_1)$ is monotone nondecreasing in $T(x)$ on the set of $x$’s such that $f(x; \theta_0) > 0$ or $f(x; \theta_1) > 0$.

- OPEF’s have MLR if $c(\theta)$ is monotone nondecreasing in $\theta$.

- We showed via a graphical argument that if MLR in $T(x)$, then the following is a Neyman-Pearson test:

$$S_{\text{MLR}}(X) = 1 \text{ if } T(X) > \kappa_{\text{MLR}};$$
\[ S_{MLR}(X) = 1 \text{ with probability } \gamma_{MLR}(X) \text{ if } T(X) = \kappa_{MLR}; \]
\[ S_{MLR}(X) = 0 \text{ if } T(X) < \kappa_{MLR}. \]

**Theorem:** The monotone likelihood ratio test has three properties: (a) \( \Pr_{\theta}(S_{MLR}(X) = 1) \) is monotone nondecreasing as a function of \( \theta \); (b) if \( \Theta_0 = \{\theta \leq \theta_0\} \) and \( \Theta_1 = \{\theta > \theta_0\} \), then the level of the MLR test is \( \Pr_{\theta_0}(S_{MLR}(X) = 1) \); and (c) the MLR test is UMP of its level.

**Proof, Step #1:** Suppose that (a) is not true. Then there exists \( \theta_L < \theta_U \) such that
\[ \alpha_* = \Pr_{\theta_L}(S_{MLR}(X) = 1) > \Pr_{\theta_U}(S_{MLR}(X) = 1). \tag{9.2} \]
Consider the test \( S_{ray}(X) \) which equals 1 with probability \( \alpha_* \) independent of \( X \). Then we have
\[ \alpha_* = \Pr_{\theta_L}(S_{ray}(X) = 1) = \Pr_{\theta_U}(S_{ray}(X) = 1). \tag{9.3} \]
However, because \( S_{MLR}(X) \) is a Neyman-Pearson test of level \( \alpha_* \), from the Neyman-Pearson lemma we have that
\[ \alpha_* = \Pr_{\theta_U}(S_{ray}(X) = 1) \leq \Pr_{\theta_U}(S_{MLR}(X) = 1). \tag{9.4} \]
The combination of (9.2) and (9.4) forms a contradiction, thus proving (a).

**Proof, Step #2:** The claim (b) is a consequence of (a) and the definition of the level of a test.

**Proof, Step #3:** To prove claim (c), we again use the contrapositive. Suppose that the MLR test is of level \( \alpha \), but it is not UMP for the composite null versus the composite alternative. Then there must be a test \( S_s(X) \) of level \( \leq \alpha \) with the property that for some \( \theta_1 > \theta_0 \),
\[ \Pr_{\theta_1}(S_s(X) = 1) > \Pr_{\theta_1}(S_{MLR}(X) = 1). \tag{9.5} \]
Because of the levels we have defined, we have
\[ \Pr_{\theta_0}(S_{MLR}(X) = 1) = \alpha \geq \Pr_{\theta_0}(S_s(X) = 1). \tag{9.6} \]
The equations (9.5)–(9.6) form a contradiction because the MLR test is a Neyman-Pearson test of level \( \alpha \) for testing the simple null that \( \Theta_0 = \{\theta_0\} \) versus \( \Theta_1 = \{\theta_1\} \).
LECTURE #21, 1998

- We considered the case of iid sampling from the Binomial \((n, \theta)\) distribution, with both \(n\) and \(\theta\) unknown.

- This is a non-regular family, and so of course all the optimality theory goes out the window.

- We computed a method of moments estimator and showed it to be very unstable.

- We computed the MLE for \((n, \theta)\) and found it to be unstable as well.

- We computed the Bayes posterior mode with a \(\text{Beta}(a, b)\) prior for \(\theta\) and a noninformative (uniform) prior for \(n\) on a range.

- With \(a = b = 2\), we found the posterior mode to be quite stable.

- With \(a = 0, b = 1\), we found the Bayes posterior mode to be very unstable. This is to be expected.

- We also computed a conditional estimator.

- The idea is simple: \(\theta\) is a nuisance parameter, while \(n\) is the parameter of interest. We found a sufficient statistic for \(\theta\) pretending that \(n\) was known, and then computed the conditional distribution of the data given this statistic.

- Obviously, the result is a distribution which is independent of \(\theta\).

- I indicated and asked you to check that the conditional mle for \(n\) was the same as the Bayes posterior mode with \(a = 0, b = 1\).
We considered pairs of independent Bernoulli random variables $(Y_{i1}, Y_{i2})$, with means $H(\alpha_i + x_{ij}\beta)$, where $x_{ij}$ is a fixed constant, $H(v)$ is the logistic distribution function, and $(\beta, \alpha_1, \alpha_2, \ldots)$ are the unknown parameters.

Here $\beta$ is the interesting parameters, and the $\alpha$'s are nuisance parameters. The sufficient statistic for $\alpha_i$ is $Y_{i1} + Y_{i2}$.

There is no information about $\beta$ in concordant pairs, i.e., when $Y_{i1} = Y_{i2}$.

We then computed the probability that $Y_{i1} = 1$ and $Y_{i2} = 0$ given that $Y_{i1} + Y_{i2} = 1$. These are the discordant pairs. If we write $Z_i = 1$ when $Y_{i1} = 1$ and $Y_{i2} = 0$, then among the discordant pairs, $Z_i$ is Bernoulli with success probability depending on $(\beta, x_{i1}, x_{i2})$.

I asked you to investigate the case that $x_{i1} = 0$ and $x_{i2} = 1$, and to show in detail how to estimate $\beta$. 
LECTURE #23, 1998

- We considered the problem of testing for non-one-sided hypotheses.

- In general, the parameter space is $\Theta$, and we are considering deciding between $\Theta_0$ and $\Theta_1$. We have that $\Theta$ is $k$-dimensional, that $\Theta_0$ is $q < k$ dimensional, and $\Theta_1$ is the complement of $\Theta_0$ in $\Theta$.

- We considered forming the generalized likelihood ratio statistic

$$L(X) = \frac{\sup \{ f(X|\theta), \theta \in \Theta \}}{\sup \{ f(X|\theta), \theta \in \Theta_0 \}}$$

- Asymptotically, we claimed that for regular families, the test which chooses $\Theta_1$ when

$$2 \log \{ L(X) \} > \chi^2_{k-q}(1 - \alpha)$$

has level $\alpha$ in the sense that for any $\theta \in \Theta_0$,

$$\alpha = \lim_{n \to \infty} \Pr_{\theta} \left[ 2 \log \{ L(X) \} > \chi^2_{k-q}(1 - \alpha) \right].$$

- We sketched a proof of this result for the case that $q = 0$ and $k = 1$, i.e., when $\Theta_0 = \{ \theta_0 \}$ and $\Theta_1 = \{ \theta \neq \theta_0 \}$.

- The proof used two terms of a Taylor series.

- We then constructed the test statistic in the special case that $X_1, ..., X_n$ are iid Normal($\theta$, 1).
LECTURE #24, 1998

- We constructed the GLR test for $X_1, \ldots, X_n$ being iid Normal($\theta, \sigma^2$), with $\theta$ and $\sigma^2$ both unknown.

- GLR tests are generally preferred because they keep their level reasonably well even in hard nonlinear problems.

- Their drawback is that they require the fitting of multiple models.

- You have to fit the model under $\Theta_0$, and then turn around and fit the model for $\Theta$, the union of $\Theta_0$ and $\Theta_1$.

- Because this “double-work” can be inconvenient with multiple parameters and hence multiple possible models, most computer packages try to do only a single fit, namely the fit under $\Theta$. They then construct what is known as Wald’s Tests.

- In the case that $q = 0$, $k = 1$, with iid sampling, the Wald test takes the following form:

$$\hat{\theta}_{MLEW} = \text{the mle}$$

$$I_*(\theta) = \text{information in a single observation}$$

$$\hat{t} = \frac{n^{1/2}|\hat{\theta}_{mle} - \theta_0|}{\{1/I_*(\hat{\theta}_{mle})\}^{1/2}}$$

$$S_{Wald}(\mathbf{X}) = 1 \quad \text{if } \hat{t} > \Phi^{-1}(1 - \alpha/2)$$

$$S_{Wald}(\mathbf{X}) = 0 \quad \text{otherwise}.$$  

- In the case that $q = 0$, $k = 1$, with more general sampling, the Wald test takes the following form:

$$\hat{\theta}_{MLEW} = \text{the mle}$$

$$\mathcal{I}(\theta) = \text{information in the sample}$$

$$\hat{t} = \frac{n^{1/2}|\hat{\theta}_{mle} - \theta_0|}{\{n/\mathcal{I}(\hat{\theta}_{mle})\}^{1/2}}$$

$$S_{Wald}(\mathbf{X}) = 1 \quad \text{if } \hat{t} > \Phi^{-1}(1 - \alpha/2)$$

$$S_{Wald}(\mathbf{X}) = 0 \quad \text{otherwise}.$$
- We then constructed the test statistic in the special case that $X_1, ..., X_n$ are iid Normal($\theta, 1$).
- We then compared GLR and Wald tests in the special case that $X_1, ..., X_n$ are iid Bernoulli($\theta$).
LECTURE #25, 1998

- In this lecture, we constructed the so-called scores tests. We discussed in detail the motivation for these procedures, namely to make a decision about a parameter without having to compute the MLE of that parameter.

- I mentioned some work in genetics that needed this problem.

- Remember that the score is

\[ \frac{\partial}{\partial \theta} \log \{ f(\mathbf{X}|\theta) \} . \]

- The variance of the score is simply \( \mathcal{I}(\theta) \).

- Hence, if we are trying to decide between \( \Theta_0 = \{ \theta_0 \} \) and \( \Theta_1 = \Theta_0^c \), it makes sense to choose the latter when

\[ \frac{|\frac{\partial}{\partial \theta} \log \{ f(\mathbf{X}|\theta) \}|}{\{ \mathcal{I}(\theta) \}^{1/2}} > z(1 - \alpha/2). \]

- We showed that the score, Wald and GLR tests coincide for the Normal(\( \theta, 1 \)) problem.

- We showed that the score, Wald and GLR tests are all different for the Bernoulli(\( \theta \)) problem.

- We also considered the case that \( X_1, \ldots, X_n \) are independent normals with mean zero and variance \( \exp(\gamma z_i) \), where \( z_1, \ldots, z_n \) are fixed constants. The score test was derived for this problem.
We considered the canonical exponential family in natural form:

\[ f(x, \eta) = \exp\{\eta T(x) + d_0(\eta) + S(x)\}. \]

Using the fact that \( T(X) = -d^{(1)}(\hat{\eta}_{ml}) \), we showed that the GLR statistic for considering \( \eta_0 \) equals

\[ Q(\hat{\eta}_{ml}, \eta_0) = (\hat{\eta}_{ml} - \eta_0)\{ -d^{(1)}(\hat{\eta}_{ml}) \} + d_0(\hat{\eta}_{ml}) - d_0(\eta_0). \]

A simple analysis shows that \( Q(\eta, \eta_0) \) as a function of \( \eta \) achieves its minimum at \( \eta_0 \) and is decreasing for \( \eta < \eta_0 \) and increasing for \( \eta > \eta_0 \). This means that the GLR test rejects \( \eta_0 \) if \( \hat{\eta}_{ml} > b \) or \( \hat{\eta}_{ml} < a \) for some \( a < b \).

It is interesting to note that as a function of \( \eta \), \( Q(\hat{\eta}_{ml}, \eta) \) is convex with minimum at \( \hat{\eta}_{ml} \). In what follows, this can be used to show that the GLR confidence set is an interval containing \( \hat{\eta}_{ml} \).

We defined a confidence set of level \( 1 - \alpha \) as a set \( C(X) \) such that \( \Pr_{\theta}\{\theta \in C(X)\} = 1 - \alpha \).

We showed how to construct such intervals from decision procedures, as follows. For any \( \theta_0 \), consider the decision between \( \Theta_0 = \{\theta_0\} \) and \( \Theta_1 = \Theta_0^c \), at the level \( \alpha \). Then a \( 1 - \alpha \) confidence set is the set of all \( \theta_0 \) values in which we do not choose \( \Theta_1 \).

We showed in the Normal(\( \theta, 1 \)) case based on a sample of size \( n \) and using the GLR test that the resulting confidence set is necessarily of the form \( \bar{X} \pm z(1 - \alpha/2)/\sqrt{n} \).
LECTURE #27, 1998

• We proved the following result. Let $\pi(\theta|\mathbf{X})$ be a strictly unimodel posterior density. Define $Q(c, \mathbf{X}) = \{\theta : \pi(\theta|\mathbf{X}) \geq c\}$, and suppose that $\text{pr}(\theta \in Q(c, \mathbf{X})|\mathbf{X}) = 1 - \alpha$. Then $Q(c, \mathbf{X})$ is the shortest interval with posterior probability $1 - \alpha$, and is called the Highest Posterior Density (HPD) region.

• The proof is in Casella & Berger. We reproduced that proof, and also gave the proof graphically.

• In the case that the posterior $\pi(\theta) \equiv$ constant, we showed that the HPD region was also the GLR confidence set.

• Finally, we considered estimating equations. Let $E_\theta\{\psi(X, \theta) = 0$. Consider solving

$$0 = n^{-1/2} \sum_{i=1}^{n} \psi(X_i, \hat{\theta}).$$

Then we showed that as $n \to \infty$,

$$n^{1/2}(\hat{\theta} - \theta) \Rightarrow \text{Normal} \left[0, \frac{E_\theta\{\psi^2(X, \theta)\}}{\left\{E_\theta\left\{\frac{\partial}{\partial \theta}\psi(X, \theta)\right\}\right)^2} \right]$$
LECTURE #28, 1998

- No class