Lecture Outline

(1) Maximum Likelihood and Normal Inference

- Confidence Intervals based on the MLE
- Delta Method – useful reparameterizations
- The curse of dimensionality in terms of nuisance parameters

(2) Failure of the Likelihood Principle with fixed α-level tests

Defn.: Let $\theta^{\boldsymbol{*}}$ denote the argmax of the likelihood function $\mathbf{p}(X \mid \theta)$, the *maximum likelihood estimate* (*MLE*) of the parameter.

Main Theorem (under general regularity conditions on the statistical model):

For large (*iid*) samples of size *n*.

$$\mathbf{P}(\theta^{\boldsymbol{*}} \mid \theta_0) \approx \boldsymbol{N}\left(\theta_0, [\boldsymbol{I}_X(\theta^{\boldsymbol{*}})]^{-1}\right) = \boldsymbol{N}\left(\theta_0, [n\boldsymbol{I}_{X_i}(\theta^{\boldsymbol{*}})]^{-1}\right)$$

$$\text{where } \boldsymbol{I}_{\boldsymbol{X}}(\theta) = \mathrm{E}_\theta\left[-\frac{\partial^2(\ln \mathbf{p}(X \mid \theta))}{\partial\theta^2}\right].$$

This leads to the *Classical Inference* procedure that,

when $\theta \in \Re$ there is the convenient 95% *Confidence Interval* based on the MLE:

$$\boldsymbol{CI} = [\theta^{*} - 2se,\ \theta^{*} + 2se]$$

where $se = [n\boldsymbol{I}_{\boldsymbol{X}_i}(\theta^{*})]^{-1/2}$

Example – yet more coin flipping

Data: $x = < x_1, \ldots , x_n > $ *iid* Bernoulli trials ***Model***: $P(X_i = 1) = \theta$, with $\theta \in \Theta = [0, 1]$.

***Likelihood function*:**

$$\mathbf{P}(x_n \mid \theta) = \Pi_i \, \mathbf{P}(x_i \mid \theta) = \Pi_i \, \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\Sigma_i x_i} (1-\theta)^{n-\Sigma_i x_i}$$

Thus, $\ln(\mathbf{P}(x_n \mid \theta)) = \Sigma_i x_i \ln(\theta) + (n-\Sigma_i x_i)\ln(1-\theta)$ and evidently $\theta^* = \Sigma_i x_i/n$

$$-\frac{\partial^2(\ln \mathbf{p}(x \mid \theta))}{\partial\theta^2} = x/\theta^2 + (1-x)/(1-\theta)^2$$

$$\boldsymbol{I}_{\boldsymbol{X}}(\theta) = \mathrm{E}_\theta\left[-\frac{\partial^2(\ln \mathbf{p}(X \mid \theta))}{\partial\theta^2}\right] = [\theta\,(1-\theta)]^{-1}$$

So: $$\theta^* \approx \boldsymbol{N}(\theta,\ \theta^*(1-\theta^*)/n]$$

And the 95% *Confidence Interval* based on the MLE is:

$$\theta^* \pm\ 2[\theta^*(1-\theta^*)/n]^{1/2}$$

But the length of the interval depends upon the parameter. Can this be controlled?

The Delta Method

When $\sqrt{n}(Y - \mu) \approx \boldsymbol{N}(0, \sigma^2)$

we have the following good approximation for (differentiable) transformations $\mathbf{g}(\bullet)$.

$$\sqrt{n}(\mathbf{g}(Y) - \mathbf{g}(\mu)) \approx \boldsymbol{N}(0, \sigma^2[\mathbf{g}'(\mu)]^2).$$

We can use this to create ***variance stabilizing transformations***.

In the coin-tossing case, $\sqrt{n}(\bar{X} - \mu) \approx \boldsymbol{N}(0, \theta(1-\theta))$

So, we want a transformation such that $\mathbf{g}'(\theta) = 1/\sqrt{[\theta(1-\theta)]}$,

with a solution $\mathbf{g}(\theta) = 2\,arcsin(\sqrt{\theta})$.

Then $\sqrt{n}(\, 2\,arcsin(\sqrt{\bar{X}}) - 2\,arcsin(\sqrt{\theta})\,) \approx \boldsymbol{N}(0, 1)$

and we may control the length of the interval estimate, independent of θ.

Nuisance Parameters and the MLE – the *curse of dimensionality* (Neyman-Scott, 1948)

Let (X_i, Y_i) be *iid* $\mathbf{N}(\mu_i, \sigma^2)$ $(i = 1, 2, \ldots.)$

σ^2 is the parameter of interest -- common variance,
the μ_i are nuisances -- the unknown means.

Likelihood function $L(\mu_i, \sigma^2)$:

$$\frac{1}{(2\pi)^n}\sigma^{-2n}\exp\{-\frac{1}{2\sigma^2}\sum_i[(x_i-\mu_i)^2+(y_i-\mu_i)^2]\}$$

And $\ln(L(\mu_i, \sigma^2))$:

$$-2n\ln\sigma-\frac{1}{2\sigma^2}[2\sum_i(\frac{x_i+y_i}{2}-\mu_i)^2+\frac{1}{2}\sum_i(x_i-y_i)^2]$$

The MLEs are calculated from this equation by setting first partial derivatives to 0, resulting in the MLE estimates:

$$\mu^*_{i,n} = (X_i + Y_i)/2 \qquad \sigma^{2*}_n = \Sigma_i (X_i - Y_i)^2/4n$$

Since $\quad (X_i - Y_i) = Z_i \sim \boldsymbol{N}(0, 2\sigma^2)$

we find that $\quad \sigma^{2*}_n \Rightarrow \sigma^2/2$

The MLE for σ^2 is inconsistent, converging to the wrong value.

Thus the *nice* convergence properties of the MLE do not extend (automatically) to the case with unlimited numbers of nuisance parameters!

We need consider ways to keep the statistical model finite dimensional.

Two approaches to resolving this anomaly

- *Classical (easy!)*: Reparameterize so that the infinity of nuisance factors are confined to one portion of the data, and there are enough data remaining for informative inference

Transform from (X_i, Y_i) to the equivalent pair (Z_i, W_i)

where $Z_i = (X_i - Y_i)$ and $W_i = (X_i + Y_i)$

$Z_i \sim \boldsymbol{N}(0, 2\sigma^2)$ and $W_i \sim \boldsymbol{N}(2\mu_i, 2\sigma^2)$ then use only the Z_i *!!*

This amounts to a transformation that permits factoring the likelihood function

$$\mathbf{P}(\langle Z, W\rangle | \sigma^2, \mu_1, \mu_2, \ldots) = \mathbf{P}(Z \mid \sigma^2)\, \mathbf{P}(W \mid \sigma^2, \mu_1, \mu_2, \ldots)$$

so that one term, $\mathbf{P}(Z \mid \sigma^2)$, involves only a finite- (one-) dimensional statistical model, including the parameter of interest

while the other term, $\mathbf{P}(W \mid \sigma^2, \mu_1, \mu_2, \ldots)$ is infinite dimensional.

• *Bayes approach – this could be hard*:

Complete the Bayes' model by adding the (possibly infinite dimensional) prior for the nuisance factors $(\mu_1, \mu_2, \ldots)$ and integrate them out using Bayes' theorem.

$$\mathbf{p}(\sigma^2 \mid <X, Y>) \propto \iint \ldots \mathbf{p}(<X, Y> \mid \sigma^2, \mu_1, \mu_2, \ldots)\mathbf{p}(\mu_1, \mu_2, \ldots \mid \sigma^2) d\mathbf{P}(\mu_1, \mu_2, \ldots \mid \sigma^2)$$

This can become tractable if, for example, the μ_i can be give a simple (conjugate) distribution, e.g., if μ_i are *iid* $\boldsymbol{N}(\theta, \tau^2)$, which gives the nuisance factors a finite dimensional statistical model.

Another matter of experimental design

Let $Y \sim \boldsymbol{N}(\mu, \sigma^2)$ with σ^2 known.

A ***statistical test*** $\delta(\boldsymbol{y})$ of a simple statistical (*null*) hypothesis H_0: $\theta = 0$ versus the *alternative* hypothesis H_1: $\theta = 1$, based on $\boldsymbol{Y}$ is defined by a ***critical region*** C, where the null hypothesis is rejected if and only if $\boldsymbol{Y} \in C$.

- The prob. of a type-1 error, $\alpha = \mathbf{P}(C \mid H_0)$.
- The prob. of a type-2 error, $\beta = \mathbf{P}(C^{\mathbf{c}} \mid H_1)$.

By the Neyman-Pearson lemma, for each value of σ^2 and a, there exists a Most Powerful test of H_0 versus the alternative H_1.

Question: What becomes of a (*Classical Statistical*) convention always to choose the *Most Powerful* test with a fixed α-level, say, $\alpha = .05$?

Table 1. *The "best" β-values for twelve α-values and six experiments*

σ =	.250	.333	.400	.500	1.000	1.333
α	β-values					
.010	.047	.250	.431	.628	.908	.942
.020	.026	.172	.327	.521	.854	.904
.030	.017	.131	.268	.452	.811	.871
.040	.012	.106	.227	.401	.773	.841
.045	.011	.096	.210	.380	.756	.828
.050	.009	.088	.196	.361	.740	.814
.055	.008	.080	.184	.344	.725	.802
.060	.007	.074	.172	.328	.710	.789
.070	.006	.064	.153	.300	.683	.766
.080	.005	.055	.137	.276	.657	.744
.090	.004	.049	.123	.255	.633	.722
.100	.003	.043	.111	.236	.611	.702

Consider tests based on two different sample sizes, e.g., $\sigma = 4/3$ and $\sigma = 1/3$.
With the larger sample size, $\sigma = 1/3$, consider tests with (α,β) values

Test T1 with operating characteristics (.050, .814).
Test T2 with operation characteristics (.070, .766).

With the smaller sample size, $\sigma = 4/3$, consider tests with (α,β) values

Test T3 with operating characteristics (.050, .088).
Test T4 with operating characteristics (.030, .131).

The convention – choose the MP test with a = .05 regardless – has an *incoherence* associated with it exposed by looking at the two mixed tests

Test T5 = .5T1 $\oplus$ .5T3 with operating characteristics (.050, .451).
Test T6 = .5T2 $\oplus$ .5T4 with operating characteristics (.050, .449).

Thus T5 is inadmissible, as T6 has better power at the same .05 level.

However, T5 is the mixture of MP .05-level tests. Thus, the MP .05 level mixed test will not be a mixture of .05-level MP tests, and *Ancillarity* fails with mixed tests!

The Bayes analysis of this phenomenon

The figure below displays the curve of available MP tests in this problem at three values of σ: σ = 4/3, = .5, and = 1/3, and the tangents to these curves for tests with α = .05.

The Bayes' prior for H_0 associated with a specific MP test is identified by the tangent to the curve at that point on the curve.

In order to be *coherent* tests chosen at different σ-values must have parallel tangents, meaning that they associate with the same (implicit) Bayes' prior for H_0.

In order to keep the tangents parallel (to maintain *coherence*),

as sample size increases (as σ decreases), α-levels must decrease as well!

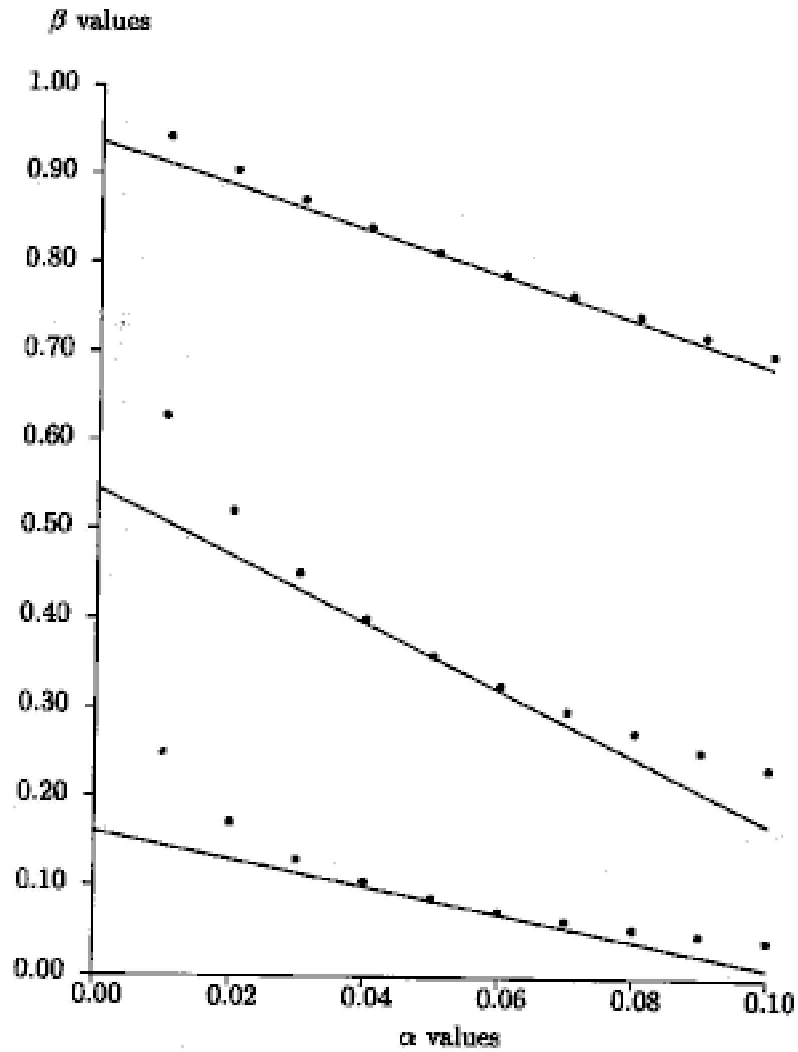

Figure 3. Families of (α, β) pairs for undominated tests