# Science, Mathematics, and Statistics 

Professor Edsel A. Peña<br>(E-Mail: pena@stat.sc.edu)<br>University of South Carolina<br>Columbia, South Carolina 29208 USA

## University of the Philippines at Los Baños Talk June 2011

## On Science

- Understanding Nature (including ourselves) and the Universe.
- What do we mean by 'Understanding' ?
- Analogy with a complicated Chess Game.
- Does it Suffice to just KNOW the Rules of the Game?
- Or, do we Require that we be able to PLAY the game well?


## Richard Feynman: American Genius



## Richard Feynman on 'Understanding'

We can imagine that this complicated array of moving things which constitutes "the world" is something like a great chess game being played by the Gods, and we are observers of the game. We do not know what the rules of the game are; all we are allowed to do is to watch the playing. Of course, if we watch long enough, we may eventually catch on to a few of the rules. The rules of the game are what we mean by fundamental physics ...

## Feynman on Understanding ... continued

Aside from not knowing all of the rules, what we really can explain in terms of those rules is very limited, because almost all situations are so enormously complicated that we cannot follow the plays of the game using the rules, much less tell what is going to happen next. We must, therefore, limit ourselves to the more basic question of the rules of the game. If we know the rules, we consider that we "understand" the world.

## Fundamental Principles of Nature

- The Meaning of Our Existence?
- Perhaps, our Unrelenting Search for the


## Fundamental Principles (Rules of the Game)

of Nature or the Universe!?

- in Physics;
- in Chemistry;
- in Biology;
- in Medicine;
- even in the Social Sciences.


## Mathematics: Language of and Reasoning in Science

- Galileo Galilee:
"Philosophy is written in this grand book, the universe which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth."
- Eugene Wigner (1960): "The Unreasonable Effectiveness of Mathematics in the Natural Sciences."
- Mario Livio (2009): "Is GOD a Mathematician?"


## Some of Nature's Deterministic Rules

- Newton's Laws.

$$
\begin{gathered}
F=m a \\
F=\frac{G m_{1} m_{2}}{r^{2}} ; G=\text { universal gravitational constant } \\
s=\frac{1}{2} g t^{2}
\end{gathered}
$$

- Maxwell's Electro-Magnetic Equations.
- Einstein's Special and General Relativity Theories.

$$
E=m c^{2}
$$

- DNA Structure (four letters: A, G, C, T).


## Iconic Albert Einstein



## Randomness, Uncertainty, and Order

## Illustration of "Order Out of Chaos" <br> Manifestation of Weak Law of Large Numbers



Histogram of LifeTime


## Probability Spaces and Densities

- Probability Space: $(\Omega, \mathfrak{F}, P)$

$$
\begin{aligned}
\Omega & =\text { space of possible values } \\
\mathfrak{F} & =\text { (measurable) subsets of } \Omega \\
P & =\text { probability measure on }(\Omega, \mathfrak{F})
\end{aligned}
$$

- Probability Densities, $f: P(A)=\int_{A} f(x) \nu(d x), A \in \mathfrak{F}$.
- Example: Lifetime variable, $L$, described by gamma distribution:

$$
P\{L \in A\}=\int_{A} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp \{-\lambda x\} d x
$$

for some $\alpha$ and $\lambda$.

- Statistician's Goal: Discover $(\alpha, \lambda)$ based on data.


## A Most Beautiful Curve: Gaussian or Normal Density

$$
f(x)=n\left(x ; \mu, \sigma^{2}\right)=\frac{1}{\sigma \sqrt{2 \pi}} \exp \left\{-\frac{1}{2 \sigma^{2}}\left(\frac{x-\mu}{\sigma}\right)^{2}\right\}
$$

Normal Density


## Carl Friedrich Gauss: Prince of Mathematics



Professor Edsel A. Peña (E-Mail: pena@stat.sc.edu)
Science, Mathematics, and Statistics

## Ubiquity of Gaussian: Central Limit Theorem

Exponential Population


Sample Means from Exponential Population


## Some of Nature's Probabilistic Rules

- Randomness intrinsic from Quantum Mechanics.
- Einstein struggled with this notion!
- Subtle is the Lord, but Malicious He is Not.
- God Does Not Play Dice!
- Performance of Machines and Computers.
- Effectiveness of New Medical Drugs or Treatments.
- Occurrences of Events (earthquakes; accidents; insurance claims; etc.)
- DNA Inheritance [from Parents to Offsprings].
- Stock Market.
- Behavior of People.


## Measurement Errors: Signal Discovery with Noisy Data

$$
\text { Unobserved Signal : } \quad D(t)=\frac{1}{2} g t^{2}
$$

Observed Noisy Data: $Y(t)=\max (D(t)+E, 0) ; E \sim N\left(0, \sigma^{2}\right)$


## Science and Art of Statistics

- Statistics, still a relatively infant science!
- Branch of Mathematics?? Deductive versus Inductive inference.
- Discovery of deterministic signals or rules from noisy data. Indispensable in the physical and biological sciences.
- Modeling and discovery of probabilistic or stochastic rules, again based on observed data.
- Multiple decision-making in the face of uncertainty.
- Proper design of experiments and studies: Garbage In, Garbage Out!
- Statistics, a key to resolving our contemporary dilemma of

Drowning in Information but Starving for Knowledge! Rutherford D. Rogers

## Two Concrete Research Areas: Personal Interests

- Modeling and Analysis of Lifetimes.
- In particular, when event of interest is recurrent (keeps happening).
- Decision-Making (hypothesis testing).
- How to choose a decision among several competing decisions based on data?
- Multiple Decision-Making.
- How to make several decisions simultaneously and to efficiently use all the data in each of the decisions?


## Recurrent Events: Some Examples

- admission to hospital due to chronic disease
- tumor re-occurrence
- migraine attacks
- alcohol or drug (eg cocaine) addiction
- machine failure or discovery of a bug in a software
- commission of a criminal act by a delinquent minor!
- major disagreements between a couple
- non-life insurance claim
- drop of $\geq 200$ points in DJIA during trading day
- publication of a research paper by a professor


## Migratory Motor Complex (MMC) Data

Data set from Aalen and Husebye ('91) with $n=19$ subjects.


## Data Accrual: One Subject



Covariate vector: $X(s)=\left(X_{1}(s), \ldots, X_{q}(s)\right)$

## Some Aspects in Recurrent Data

- random monitoring length $(\tau)$.
- random \# of events $(K)$ and sum-quota constraint:

$$
K=\max \left\{k: \sum_{j=1}^{k} T_{j} \leq \tau\right\} \text { with } \sum_{j=1}^{K} T_{j} \leq \tau<\sum_{j=1}^{K+1} T_{j}
$$

- Basic Observable: $\left(K, \tau, T_{1}, T_{2}, \ldots, T_{K}, \tau-S_{K}\right)$
- always a right-censored observation.
- dependent and informative censoring.
- effects of covariates, frailties, interventions after each event, and accumulation of events.


## Simplest Model: One Subject

- $T_{1}, T_{2}, \ldots \stackrel{\text { IID }}{\sim} F:$ (renewal model)
- 'perfect interventions' after each event
- $\tau \sim G$
- $F$ and $G$ not related
- no covariates (X)
- no frailties (Z)
- $F$ could be parametric or nonparametric.
- Relevant Functions:

$$
\begin{aligned}
\bar{F}=1-F ; \quad \Lambda & =-\log \bar{F} ; \quad \lambda=\Lambda^{\prime} ; \quad \bar{F}=\exp (-\Lambda) \\
\lambda(t) d t & \approx P\{T \in(t, t+d t] \mid T \geq t\}
\end{aligned}
$$

- Product-Integral Representation:

$$
\bar{F}(t)=\prod_{v=0}^{t}[1-\Lambda(d v)]
$$

## Nonparametric Estimation of $F$

Some Results from Peña, Strawderman and Hollander (JASA, 01):

$$
\begin{gathered}
N(t)=\sum_{i=1}^{n} \sum_{j=1}^{K_{i}} I\left\{T_{i j} \leq t\right\} \\
Y(t)=\sum_{i=1}^{n}\left\{\sum_{j=1}^{K_{i}} I\left\{T_{i j} \geq t\right\}+I\left\{\tau_{i}-S_{i K_{i}} \geq t\right\}\right\} \\
\text { GNAE }: \tilde{\Lambda}(t)=\int_{0}^{t} \frac{d N(w)}{Y(w)} \\
\text { GPLE }: \tilde{\bar{F}}(t)=\prod_{0}^{t}\left[1-\frac{d N(w)}{Y(w)}\right]
\end{gathered}
$$

## Main Asymptotic Result

$k$ th Convolution: $F^{\star(k)}(t)=\operatorname{Pr}\left\{\sum_{j=1}^{k} T_{j} \leq t\right\}$

$$
\begin{gathered}
\text { Renewal Function: } \rho(t)=\sum_{k=1}^{\infty} F^{\star(k)}(t) \\
\nu(t)=\frac{1}{\bar{G}(t)} \int_{t}^{\infty} \rho(w-t) d G(w) \\
\sigma^{2}(t)=\bar{F}(t)^{2} \int_{0}^{t} \frac{d F(w)}{\bar{F}(w)^{2} \bar{G}(w)[1+\nu(w)]}
\end{gathered}
$$

Theorem (JASA, 01): $\sqrt{n}(\tilde{\tilde{F}}(t)-F(t)) \Rightarrow \mathrm{GP}\left(0, \sigma^{2}(t)\right)$

## Extending KG Model: Recurrent Setting

- Wanted: a tractable model with monitoring time informative about $F$.
- Potential to refine analysis of efficiency gains/losses.
- Idea: Why not simply generalize the KG model for the RCM.
- Generalized KG Model (GKG) for Recurrent Events:

$$
\exists \beta>0, \quad \bar{G}(t)=\bar{F}(t)^{\beta}
$$

with $\beta$ unknown, and $F$ the common inter-event time distribution function.

- Remark: $\tau$ may also represent system failure/death, while the recurrent event could be shocks to the system.
- Remark: Association (within unit) could be modeled through a frailty.


## Estimation Issues and Some Questions

- How to semiparametrically estimate $\beta, \Lambda$, and $\bar{F}$ ?
- Parametric estimation in Adekpedjou, Peña, and Quiton (2010, JSPI).
- How much efficiency loss is incurred when the informative monitoring model structure is ignored?
- How much penalty is incurred with Single-event analysis relative to Recurrent-event analysis?
- In particular, what is the efficiency loss for estimating $F$ when using the nonparametric estimator in PSH (2001) relative to the semiparametric estimator that exploits the informative monitoring structure?


## Basic Processes

$$
\begin{gathered}
S_{i j}=\sum_{k=1}^{j} T_{i k} \\
N_{i}^{\dagger}(s)=\sum_{j=1}^{\infty} I\left\{S_{i j} \leq s\right\} \\
Y_{i}^{\dagger}(s)=I\left\{\tau_{i} \geq s\right\} \\
R_{i}(s)=s-S_{i N_{i}^{\dagger}(s-)}=\text { backward recurrence time } \\
A_{i}^{\dagger}(s)=\int_{0}^{s} Y_{i}^{\dagger}(v) \lambda\left[R_{i}(v)\right] d v \\
N_{i}^{\tau}(s)=I\left\{\tau_{i} \leq s\right\} \\
Y_{i}^{\tau}(s)=I\left\{\tau_{i} \geq s\right\}
\end{gathered}
$$

## Transformed Processes

$$
\begin{gathered}
Z_{i}(s, t)=I\left\{R_{i}(s) \leq t\right\} \\
N_{i}(s, t)=\int_{0}^{s} Z_{i}(v, t) N_{i}^{\dagger}(d v)=\sum_{j=1}^{\left.N_{i}^{\dagger}(s)\right)} I\left\{T_{i j} \leq t\right\} \\
Y_{i}(s, t)=\sum_{j=1}^{N_{i}^{\dagger}(s-)} I\left\{T_{i j} \geq t\right\}+I\left\{\left(s \wedge \tau_{i}\right)-S_{i N_{i}^{\dagger}(s-)} \geq t\right\} \\
A_{i}(s, t)=\int_{0}^{s} Z_{i}(v, t) A_{i}^{\dagger}(d v)=\int_{0}^{t} Y_{i}(s, w) \lambda(w) d w \\
\left\{M_{i}(v, t)=N_{i}(v, t)-A_{i}(v, t): v \geq 0\right\} \text { are martingales. }
\end{gathered}
$$

## Aggregated Processes

$$
\begin{aligned}
N(s, t) & =\sum_{i=1}^{n} N_{i}(s, t) \\
Y(s, t) & =\sum_{i=1}^{n} Y_{i}(s, t) \\
A(s, t) & =\sum_{i=1}^{n} A_{i}(s, t) \\
N^{\tau}(s) & =\sum_{i=1}^{n} N_{i}^{\tau}(s) \\
Y^{\tau}(s) & =\sum_{i=1}^{n} Y_{i}^{\tau}(s)
\end{aligned}
$$

## First, Assume $\beta$ Known

Via Method-of-Moments Approach, 'estimator' of $\Lambda$ :

$$
\hat{\Lambda}(s, t \mid \beta)=\int_{0}^{t}\left\{\frac{N(s, d w)+N^{\tau}(d w)}{Y(s, w)+\beta Y^{\tau}(w)}\right\}
$$

Using product-integral representation of $\bar{F}$ in terms of $\Lambda$, 'estimator' of $\bar{F}$ :

$$
\hat{\bar{F}}(s, t \mid \beta)=\prod_{w=0}^{t}\left\{1-\frac{N(s, d w)+N^{\tau}(d w)}{Y(s, w)+\beta Y^{\tau}(w)}\right\}
$$

## Estimating $\beta$ : Profile Likelihood MLE

Profile Likelihood:

$$
\begin{aligned}
& L_{P}\left(s^{*} ; \beta\right)=\beta^{N^{\tau}\left(s^{*}\right)} \times \\
& \quad \prod_{i=1}^{n}\left\{\left[\prod_{v=0}^{s^{*}}\left\{\frac{1}{Y\left(s^{*}, v\right)+\beta Y^{\tau}(v)}\right\}^{N_{i}^{\tau}(\Delta v)}\right] \times\right. \\
& \left.\quad\left[\prod_{v=0}^{s^{*}}\left\{\frac{1}{Y\left(s^{*}, v\right)+\beta Y^{\tau}(v)}\right\}^{N_{i}\left(s^{*}, \Delta v\right)}\right]\right\}
\end{aligned}
$$

Estimator of $\beta$ :

$$
\hat{\beta}=\arg \max _{\beta} L_{P}\left(s^{*} ; \beta\right)
$$

Computational Aspect: in R, we used optimize to get good seed for the Newton-Raphson iteration.

## Estimators of $\Lambda$ and $\bar{F}$

Estimator of $\Lambda$ :

$$
\hat{\Lambda}\left(s^{*}, t\right)=\hat{\Lambda}\left(s^{*}, t \mid \hat{\beta}\right)=\int_{0}^{t}\left\{\frac{N\left(s^{*}, d w\right)+N^{\tau}(d w)}{Y\left(s^{*}, w\right)+\hat{\beta} Y^{\tau}(w)}\right\}
$$

Estimator of $\bar{F}$ :

$$
\hat{\bar{F}}\left(s^{*}, t\right)=\hat{\bar{F}}\left(s^{*}, t \mid \hat{\beta}\right)=\prod_{w=0}^{t}\left\{1-\frac{N\left(s^{*}, d w\right)+N^{\tau}(d w)}{Y\left(s^{*}, w\right)+\hat{\beta} Y^{\tau}(w)}\right\}
$$

## Illustrative Data $(n=30): \operatorname{GKG}[$ Wei $(2,1), \beta=.2]$



## Estimates of $\beta$ and $\bar{F}$

$$
\hat{\beta}=.2331
$$

Estimates of SF for Illustrative Data (Blue=GKG; Red=PSH; Green=TRUE)


Professor Edsel A. Peña (E-Mail: pena@stat.sc.edu) Science, Mathematics, and Statistics

## Properties of Estimators

$$
\begin{gathered}
G_{s}(w)=G(w) I\{w<s\}+I\{w \geq s\} \\
\mathbb{E}\left\{Y_{1}(s, t)\right\} \equiv y(s, t)=\bar{F}(t) \bar{G}_{s}(t)+\bar{F}(t) \int_{t}^{\infty} \rho(w-t) d G_{s}(w) \\
\mathbb{E}\left\{Y_{1}^{\tau}(t)\right\} \equiv y^{\tau}(t)=\bar{F}(t)^{\beta}
\end{gathered}
$$

$$
\text { True Values }=\left(F_{0}, \Lambda_{0}, \beta_{0}\right)
$$

$$
y_{0}(s, t)=y\left(s, t ; \Lambda_{0}, \beta_{0}\right)
$$

$$
y_{0}^{\tau}(s)=y^{\tau}\left(s ; \Lambda_{0}, \beta_{0}\right)
$$

## Existence, Consistency, Normality

## Theorem

There is a sequence of $\hat{\beta}$ that is consistent, and $\hat{\Lambda}\left(s^{*}, \cdot\right)$ and $\hat{\bar{F}}\left(s^{*}, \cdot\right)$ are both uniformly strongly consistent.

## Theorem

As $n \rightarrow \infty$, we have

$$
\sqrt{n}\left(\hat{\beta}-\beta_{0}\right) \Rightarrow N\left(0,\left[\mathcal{I}_{P}\left(s^{*} ; \Lambda_{0}, \beta_{0}\right)\right]^{-1}\right)
$$

with

$$
\mathcal{I}_{P}\left(s^{*} ; \Lambda_{0}, \beta_{0}\right)=\frac{1}{\beta_{0}} \int_{0}^{s^{*}} \frac{y_{0}^{\tau}(v) y_{0}\left(s^{*}, v\right)}{y_{0}\left(s^{*}, v\right)+\beta_{0} y_{0}^{\tau}(v)} \lambda_{0}(v) d v .
$$

## Weak Convergence of $\hat{\Lambda}\left(s^{*}, \cdot\right)$

## Theorem

As $n \rightarrow \infty,\left\{\sqrt{n}\left[\hat{\Lambda}\left(s^{*}, t\right)-\Lambda_{0}(t)\right]: t \in\left[0, t^{*}\right]\right\}$ converges weakly to a zero-mean Gaussian process with variance function

$$
\begin{aligned}
& \sigma_{\hat{\Lambda}}^{2}\left(s^{*}, t\right)=\int_{0}^{t} \frac{\Lambda_{0}(d v)}{y_{0}\left(s^{*}, v\right)+\beta_{0} y_{0}^{\tau}(v)}+ \\
& \quad\left[\int_{0}^{s^{*}} \frac{y_{0}\left(s^{*}, v\right) y_{0}^{\tau}(v)}{\beta_{0}\left[y_{0}\left(s^{*}, v\right)+\beta_{0} y_{0}^{\tau}(v)\right]} \Lambda_{0}(d v)\right]^{-1} \times \\
& \quad\left[\int_{0}^{t} \frac{y_{0}^{\tau}(v)}{y_{0}\left(s^{*}, v\right)+\beta_{0} y_{0}^{\tau}(v)} \Lambda_{0}(d v)\right]^{2} .
\end{aligned}
$$

Remark: The last product term is the effect of estimating $\beta$. It inflates the asymptotic variance.

## Weak Convergence of $\hat{F}\left(s^{*}, \cdot\right)$ and $\tilde{\tilde{F}}\left(s^{*}, \cdot\right)$

## Corollary

As $n \rightarrow \infty,\left\{\sqrt{n}\left[\hat{\bar{F}}\left(s^{*}, t\right)-\bar{F}_{0}(t)\right]: t \in\left[0, t^{*}\right]\right\}$ converges weakly to a zero-mean Gaussian process whose variance function is

$$
\sigma_{\hat{F}}^{2}\left(s^{*}, t\right)=\bar{F}_{0}(t)^{2} \sigma_{\hat{\Lambda}}^{2}\left(s^{*}, t\right) \equiv \bar{F}_{0}(t)^{2} \sigma_{\tilde{\Lambda}}^{2}\left(s^{*}, t\right)
$$

Recall/Compare!
Theorem (PSH, 2001)
As $n \rightarrow \infty,\left\{\sqrt{n}\left[\tilde{\tilde{F}}\left(s^{*}, t\right)-\bar{F}_{0}(t)\right]: t \in\left[0, t^{*}\right]\right\}$ converges weakly to a zero-mean Gaussian process whose variance function is

$$
\sigma_{\tilde{F}}^{2}\left(s^{*}, t\right)=\bar{F}_{0}(t)^{2} \int_{0}^{t} \frac{\Lambda_{0}(d v)}{y_{0}\left(s^{*}, v\right)}
$$

## Asymptotic Relative Efficiency: $\beta_{0}$ Known

If we know $\beta_{0}$ :

$$
\begin{aligned}
& \operatorname{ARE}\left\{\tilde{\tilde{F}}\left(s^{*}, t\right): \hat{\bar{F}}\left(s^{*}, t \mid \beta_{0}\right)\right\}= \\
& \left\{\int_{0}^{t} \frac{\Lambda_{0}(d w)}{y_{0}\left(s^{*}, w\right)}\right\}^{-1} \times \\
& \quad\left\{\int_{0}^{t} \frac{\Lambda_{0}(d w)}{y_{0}\left(s^{*}, w\right)+\beta_{0} y_{0}^{\tau}(w)}\right\}
\end{aligned}
$$

Clearly, this could not exceed unity, as is to be expected.

## Case of Exponential F: $\beta_{0}$ Known

Theorem
If $\bar{F}_{0}(t)=\exp \left\{-\theta_{0} t\right\}$ for $t \geq 0$ and $s^{*} \rightarrow \infty$, then

$$
\begin{aligned}
& \operatorname{ARE}\left\{\tilde{\tilde{F}}(\infty, t): \hat{\bar{F}}\left(\infty, t \mid \beta_{0}\right)\right\}= \\
& \qquad\left\{\int_{\bar{F}_{0}(t)}^{1} \frac{d u}{\left(1+\beta_{0}\right) u^{2+\beta_{0}}}\right\}^{-1} \times \\
& \quad\left\{\int_{\bar{F}_{0}(t)}^{1} \frac{d u}{\left(1+\beta_{0}\right) u^{2+\beta_{0}}+\beta_{0}^{2} u^{1+\beta_{0}}}\right\} .
\end{aligned}
$$

Also, $\forall t \geq 0$,

$$
\operatorname{ARE}\left\{\tilde{\tilde{F}}(\infty, t): \hat{F}\left(\infty, t ; \beta_{0}\right)\right\} \leq \frac{1+\beta_{0}}{1+\beta_{0}+\beta_{0}^{2}} .
$$

## ARE-Plots; $\beta_{0} \in\{.1, .3, .5, .7, .9,1.0,1.5\}$ Known; $F=$ Exponential



## Case of $\beta_{0}$ Unknown

- As to be expected, if $\beta_{0}$ is known, then the estimator exploiting the GKG structure is more efficient.
- Question: Does this dominance hold true still if $\beta_{0}$ is now estimated?

Theorem
Under the GKG model, for all $\left(\bar{F}_{0}, \beta_{0}\right)$ with $\beta_{0}>0, \tilde{\bar{F}}\left(s^{*}, t\right)$ is asymptotically dominated by $\hat{\bar{F}}\left(s^{*}, t\right)$ in the sense that

$$
\operatorname{ARE}\left(\tilde{\bar{F}}\left(s^{*}, t\right): \hat{\bar{F}}\left(s^{*}, t\right)\right) \leq 1
$$

## Proof.

Neat application of Cauchy-Schwartz Inequality.

## ARE-Plots; $\beta_{0} \in\{.1, .3, .5, .7, .9,1.0,1.5\}$ Unknown; $F=$ Exponential



# Simulated $\operatorname{RE}(\tilde{F}: \hat{F})$ under a Weibull $F$ with $\alpha=2$; $\beta_{0} \in\{.1, .3, .5, .7, .9,1.0,1.5\}$ but Unknown 



## Multiple Decision-Making: Area of Relevance

- Microarray data analysis: Which genes are relevant?
- Variable selection: Which of many predictors are relevant?
- Survival analysis: Which predictors affect a lifetime variable?
- Reliability: Which components in a system are relevant?
- Epidemiology: Spread of a disease in a geographical area.
- Oil (mineral) exploration: Where to dig?
- Business: Locations of business ventures.
- US Presidential Election: Where to focus resources to optimize electoral votes?
- Sporting Events: Predicting outcomes of World Cup soccer games.


## A Microarray Data: HeatMap of Gene Expression Levels

First 100 genes out of 41267 genes in a colon cancer study at USC (M Peña's Lab). Three groups (Control; 9 Days; 2 Weeks) with 6 replicates each.

HeatMap of First 100 Genes


## General Multiple Decision Problem

- Discover the value of a parameter

$$
\theta=\left(\theta_{1}, \theta_{2}, \ldots, \theta_{M}\right) \in \Theta=\{0,1\}^{M}
$$

- $\theta_{m}=1$ means $m$ th component is relevant; $\theta_{m}=0$ means $m$ th component is not relevant.
- Choose an action

$$
a=\left(a_{1}, a_{2}, \ldots, a_{M}\right) \in \mathfrak{A}=\{0,1\}^{M}
$$

- $a_{m}=1$ means declare that $\theta_{m}=1$, a discovery; $a_{m}=0$ means declare $\theta_{m}=0$, a non-discovery.


## Assessing Actions via Losses

- Family-wise error indicator (FWEI):

$$
L_{0}(a, \theta)=I\left\{\sum_{m=1}^{M} a_{m}\left(1-\theta_{m}\right)>0\right\}
$$

- False Discovery Proportion (FDP):

$$
L_{1}(a, \theta)=\frac{\sum_{m=1}^{M} a_{m}\left(1-\theta_{m}\right)}{\max \left\{\sum_{m=1}^{M} a_{m}, 1\right\}}
$$

- Missed Discovery Proportion (MDP):

$$
L_{2}(a, \theta)=\frac{\sum_{m=1}^{M}\left(1-a_{m}\right) \theta_{m}}{\max \left\{\sum_{m=1}^{M} \theta_{m}, 1\right\}}
$$

## Illustration

- $M=10$ Decisions to Make.

| Truth or State of Reality: | $\theta=(1,1,0,1,0,0,0,0,0,1)$ |
| :---: | :---: |
| Action Taken: | $a=(0,1,0,1,1,0,0,0,0,0)$ |

- Agreements and Disagreements (in Tabular Form):

| Truth $(\theta)$ | Action (a) |  |  |
| :---: | :---: | :---: | :---: |
|  | 0 | 1 | Total |
| 0 | 5 | 1 | 6 |
| 1 | 2 | 2 | 4 |
| Total | 7 | 3 | 10 |

- Family wise error indicator (FWEI) $=1$ (Type I)
- False discovery proportion (FDP) $=1 / 3=.33$ (Type I)
- Missed discovery proportion (MDP) $=2 / 4=.50$ (Type II)


## Nostalgia: Paul, the Oracle!



## Life is Tough: No Oracles, Just Statisticians!

- Obtain a BIG data (e.g., microarrays, Netflix):

$$
X \in \mathfrak{X}
$$

- Observable is Random:

$$
X \sim P=\text { Unknown Probability Function }
$$

- Marginal Components:

$$
X_{m}=z_{m}(X) \in \mathfrak{X}_{m} \quad \text { and } \quad X_{m} \sim P_{m}=P z_{m}^{-1}
$$

- Parameters of Interest:

$$
\theta_{m}=\theta_{m}\left(P_{m}\right)
$$

- Example:

$$
\theta_{m}=1 \Longleftrightarrow P_{m} \in\left\{N\left(\mu, \sigma^{2}\right): \mu>0, \sigma^{2}>0\right\}
$$

## Multiple Decision Functions

- Multiple Decision Function:

$$
\delta: \mathfrak{X} \rightarrow \mathfrak{A}=\{0,1\}^{M} \quad \text { Note }:|\mathfrak{A}|=2^{M}
$$

- Components:

$$
\begin{gathered}
\delta=\left(\delta_{1}, \delta_{2}, \ldots, \delta_{M}\right) \\
\delta_{m}: \mathfrak{X} \rightarrow\{0,1\}
\end{gathered}
$$

- $\mathfrak{D}$ : space or collection of multiple decision functions.
- $\mathcal{M}_{0}=\left\{m: \theta_{m}=0\right\}$ and $\mathcal{M}_{1}=\left\{m: \theta_{m}=1\right\}$
- Structure: $\left\{\delta_{m}(X): m \in \mathcal{M}_{0}\right\}$ is an independent collection, and is independent of $\left\{\delta_{m}(X): m \in \mathcal{M}_{1}\right\}$.
- $\left\{\delta_{m}(X): m \in \mathcal{M}_{1}\right\}$ need NOT be an independent collection.


## Risk Functions: Averaged Losses over Possible Data

- Given a $\delta \in \mathfrak{D}$ :
- Family-Wise Error Rate (FWER):

$$
R_{0}(\delta, P)=E\left[L_{0}(\delta(X), \theta(P))\right]
$$

- False Discovery Rate (FDR):

$$
R_{1}(\delta, P)=E\left[L_{1}(\delta(X), \theta(P))\right]
$$

- Missed Discovery Rate (MDR):

$$
R_{2}(\delta, P)=E\left[L_{2}(\delta(X), \theta(P))\right]
$$

- Expectations are with respect to $X \sim P$.
- Our Goal: Choose $\delta \in \mathfrak{D}$ with small risks, whatever $P$ is.


## Special Case: A Pair of Choices $(M=1)$

- $\theta \in \Theta=\{0,1\}$
- $a \in \mathfrak{A}=\{0,1\}$
- $L_{0}(a, \theta)=L_{1}(a, \theta)=a l(\theta=0)$
- $L_{2}(a, \theta)=(1-a)!(\theta=1)$
- $X \sim P$ with $P \in\left\{P_{0}, P_{1}\right\}$.
- Assume $P_{0}$ and $P_{1}$ have respective densities:

$$
f_{0}(x) \text { and } f_{1}(x)
$$

- $R_{0}(\delta, \theta)=R_{1}(\delta, \theta)=P_{0}(\delta(X)=1)\ulcorner(\theta=0)$
- $R_{2}(\delta, \theta)=\left[1-P_{1}(\delta(X)=1)\right] /(\theta=1)$


## Types I and II Errors, Power, and Optimality

- $R_{0}(\delta, \theta)$ : Type I error probability.
- $R_{2}(\delta, \theta)$ : Type II error probability.
- Note

$$
R_{2}(\delta, \theta=1)=1-\pi(\delta)
$$

where

$$
\pi(\delta)=P_{1}(\delta(X)=1)=\text { POWER of } \delta
$$

- Desired Goal: Given Type I level $\alpha \in[0,1]$, find $\delta^{*}(\cdot ; \alpha)$ with

$$
R_{0}\left(\delta^{*}, \theta\right) \leq \alpha, \quad \text { for all } \theta,
$$

and

$$
R_{1}\left(\delta^{*}, \theta\right) \leq R_{1}(\delta, \theta), \quad \text { for all } \theta
$$

for any other $\delta$ with $R_{1}(\delta, \theta) \leq \alpha, \forall \theta$.

## Neyman-Pearson MP Test $\delta_{\alpha}^{*}$

- Neyman and Pearson (1933): optimal [most powerful] decision function of form

$$
\delta_{\alpha}^{*}(x)=\left\{\begin{array}{ccc}
1 & \text { if } & f_{1}(x)>c(\alpha) f_{0}(x) \\
\gamma(\alpha) & \text { if } & f_{1}(x)=c(\alpha) f_{0}(x) \\
0 & \text { if } & f_{1}(x)<c(\alpha) f_{0}(x)
\end{array}\right.
$$

where $c(\alpha)$ and $\gamma(\alpha)$ satisfy

$$
R_{0}\left(\delta_{\alpha}^{*}, \theta=0\right)=\alpha
$$

- Remark: Depends on $\alpha$, hence power depends on $\alpha$.
- Leads to the notion of a decision process.


## Concrete Example of a Decision Process

- Model: $X=\left(X_{1}, X_{2}, \ldots, X_{n}\right) \stackrel{\text { IID }}{\sim} N\left(\mu, \sigma^{2}\right)$.
- Problem: Test $H_{0}: \mu \leq \mu_{0}[\theta=0]$ vs $H_{1}: \mu>\mu_{0}[\theta=1]$
- Decision Function: $t$-test of size $\alpha$ given by

$$
\delta(X ; \alpha)=I\left\{\frac{\sqrt{n}\left(\bar{X}-\mu_{0}\right)}{S} \geq t_{n-1 ; \alpha}\right\}
$$

- Decision function depends on the size index $\alpha$.
- Decision Process:

$$
\Delta=(\delta(\alpha) \equiv \delta(\cdot ; \alpha): \alpha \in[0,1])
$$

- Size Condition:

$$
\sup \left\{E_{P}[\delta(X ; \alpha)]: \theta(P)=0\right\} \leq \alpha
$$

## Multiple Decision Process

- Multiple decision problem with $M$ components.
- Multiple Decision Process:

$$
\boldsymbol{\Delta}=\left(\Delta_{m}: m \in \mathcal{M}=\{1,2, \ldots, M\}\right)
$$

- Decision Process for $m$ th Component:

$$
\Delta_{m}=\left(\delta_{m}(\alpha): \alpha \in[0,1]\right)
$$

- Example: $t$-test decision process for each component.
- Usual Approach: Pick a $\delta_{m}$ from $\Delta_{m}$ using the same $\alpha$.
- Common Choices for $\alpha$ : (weak) FWER Threshold of $q$ use:

Bonferroni: $\alpha=q / M$
Sidak: $\quad \alpha=1-(1-q)^{1 / M}$

## Notion of Size Functions

- Size Function:

$$
A:[0,1] \rightarrow[0,1]
$$

continuous, strictly increasing, $A(0)=0$ and $A(1) \leq 1$, and possibly differentiable.

- Bonferroni size function: $A(\alpha)=\alpha / M$
- Sidak size function: $A(\alpha)=1-(1-\alpha)^{1 / M}$
- S: collection of possible size functions.
- For decision process $\Delta$ and size function $A$, choose decision function from $\Delta$ via

$$
\delta[A(\alpha)] .
$$

## Multiple Decision Size Function

- For a multiple decision problem with $M$ components, a multiple decision size function is

$$
\mathbf{A}=\left(A_{m}: m \in \mathcal{M}\right) \quad \text { with } \quad A_{m} \in \mathfrak{S} .
$$

- Condition:

$$
1-\prod_{m \in \mathcal{M}}\left[1-A_{m}(\alpha)\right] \leq \alpha
$$

- Given a $\Delta=\left(\Delta_{m}: m \in \mathcal{M}\right)$ and an $\mathbf{A}=\left(A_{m}: m \in \mathcal{M}\right)$, multiple decision function is

$$
\delta(\alpha)=\left(\delta_{m}\left[A_{m}(\alpha)\right]: m \in \mathcal{M}\right)
$$

- Weak FWER of $\delta(\alpha)$ :

$$
R_{0}(\delta(\alpha), P)=1-\prod\left[1-A_{m}(\alpha)\right] \leq \alpha
$$

## Neyman-Pearson Paradigm

- Control Type I error rate; minimize Type II error rate.
- Desired Type I error threshold: $q \in(0,1)$
- Weak Control: For $P$ with $\theta_{m}(P)=0$ for all $m$, want a $\delta$ with

$$
R_{0}(\delta, P) \leq q \quad \text { or } \quad R_{1}(\delta, P) \leq q
$$

- Strong Control: Whatever $P$ is, want a $\delta$ such that

$$
R_{0}(\delta, P) \leq q \quad \text { or } \quad R_{1}(\delta, P) \leq q
$$

- And, if above Type I error control is achieved, we want to have $R_{2}(\delta, P)$ small, if not optimal.


## Towards Strong FWER Control

Given a MDP $\Delta=\left(\Delta_{m}\right)$ and MDS $\mathbf{A}=\left(A_{m}\right)$, for the chosen $\delta$ at $\alpha$, its FWER is

$$
\begin{aligned}
R_{0}(\delta, P) & =E_{P}\left\{I\left(\sum \delta_{m}\left[A_{m}(\alpha)\right]\left[1-\theta_{m}(P)\right]>0\right)\right\} \\
& =P\left\{\sum_{\mathcal{M}_{0}} \delta_{m}\left[A_{m}(\alpha)\right]>0\right\} \\
& =1-\prod_{\mathcal{M}_{0}}\left[1-A_{m}(\alpha)\right] \\
& =1-\prod\left[1-A_{m}(\alpha)\right]^{1-\theta_{m}(P)}
\end{aligned}
$$

Question: Given a threshold of $\boldsymbol{q}$, what is the best $\alpha$ ?

## 'Best' Choice of $\alpha$

- Oracle Paul's Choice:

$$
\alpha^{\dagger}(q ; P)=\inf \left\{\alpha \in[0,1]: \prod\left[1-A_{m}(\alpha)\right]^{1-\theta_{m}(P)}<1-q\right\}
$$

- But, $P$ is unknown, hence $\theta_{m}(P)$ is also unknown. But we could estimate $\theta_{m}(P)$ by

$$
\delta_{m}\left[A_{m}(\alpha)-\right] .
$$

- The Oracle's choice is then estimated by

$$
\alpha^{\dagger}(q)=\inf \left\{\alpha \in[0,1]: \prod\left[1-A_{m}(\alpha)\right]^{1-\delta_{m}\left[A_{m}(\alpha)-\right]}<1-q\right\}
$$

## Strong FWER-Controlling MDF

- Chosen Multiple Decision Function:

$$
\delta^{\dagger}(q)=\left(\delta_{m}\left[A_{m}\left(\alpha^{\dagger}(q)\right)\right]: m \in \mathcal{M}\right)
$$

- Theorem

Given a $\Delta=\left(\Delta_{m}\right)$ and an $\mathbf{A}=\left(A_{m}\right)$, the $\delta^{\dagger}(q)$ defined above has

$$
R_{0}\left(\delta^{\dagger}(q), P\right) \leq q
$$

whatever $P$ is. That is, $\delta^{\dagger}(q)$ is an MDF achieving strong FWER control at level $q$.

## FDR Control

- Let

$$
\alpha^{*}(q)=\sup \left\{\alpha \in[0,1]: \sum A_{m}(\alpha) \leq q \sum \delta_{m}\left[A_{m}(\alpha)\right]\right\}
$$

- Chosen Multiple Decision Function:

$$
\delta^{*}(q)=\left(\delta_{m}\left[A_{m}\left(\alpha^{*}(q)\right)\right]: m \in \mathcal{M}\right)
$$

- Theorem

Given a pair $(\Delta, \mathbf{A})$, the $\operatorname{MDF} \delta^{*}(q)$ achieves $F D R$ control at level $q$ in that

$$
R_{1}\left(\delta^{*}(q), P\right) \leq q
$$

## Famous FDR-Controlling MDF

- Benjamini-Hochberg (JRSS B, '95) paper. Most well-known FDR-controlling procedure.
- Let $P_{1}, P_{2}, \ldots, P_{M}$ be the ordinary $P$-values from the $M$ tests.
- Let $P_{(1)}<P_{(2)}<\ldots<P_{(M)}$ be the ordered $P$-values.
- For FDR-threshold equal to $q$, define

$$
K=\max \left\{k \in\{0,1,2, \ldots, M\}: P_{(k)} \leq \frac{q k}{M}\right\} .
$$

- BH MDF $\delta^{B H}(q)=\left(\delta_{m}^{B H}: m \in \mathcal{M}\right)$ has

$$
\delta_{m}^{B H}(X)=I\left\{P_{m} \leq P_{(K)}\right\}, m \in \mathcal{M} .
$$

- Simple and easy-to-implement, but is it the BEST?


## BH Procedure on Two-Group Microarray Data

- Agilent Technology microarray data set from M. Peña's lab. Jim Ryan of NOAA did the microarray analysis.
- $M=41267$ genes.
- 2 groups, each group with 5 replicates.
- Applied $t$-test for each gene, using logged expression values. $P$-values obtained.
- Applied Benjamini-Hochberg Procedure with $q=.15$ to pick out the significant genes from the $M=41267$ genes.
- Procedure picked out 2599 significant genes.
- Further analyzed the top (wrt to their p-values) 200 genes from these selected genes.
- Performed a cluster analysis on these 200 genes.


## Histogram of the $P$-Values from the $t$-Tests

Histogram of data\$P.CTFL


Professor Edsel A. Peña (E-Mail: pena@stat.sc.edu)
Science, Mathematics, and Statistics

## Scatterplot of the Pairwise Gene Means

Significant and Chosen Genes


## Obtaining a Better MDF than BH

- IDEA: Given MDP $\Delta=\left(\Delta_{m}: m \in \mathcal{M}\right)$, find optimal MDS $\mathbf{A}^{*} \equiv \mathbf{A}^{*}(\Delta) \in \mathfrak{S}$ achieving smallest MDR

$$
R_{2}\left[(\Delta \circ \mathbf{A})(\alpha), P_{1}\right]=\frac{1}{M} \sum\left\{1-\pi_{m}\left[A_{m}(\alpha)\right]\right\}
$$

- $\pi_{m}(\alpha)=$ POWER of $\delta_{m}(\alpha)$
- FWER-controlling MDF:

$$
\delta^{\dagger}(q)=\delta^{\dagger}\left(q ; \Delta, \mathbf{A}^{*}(\Delta)\right)
$$

- FDR-controlling MDF:

$$
\delta^{*}(q)=\delta^{*}\left(q ; \Delta, \mathbf{A}^{*}(\Delta)\right)
$$

- Use the best MDP $\Delta$, e.g., MPs; UMPs; UMPUs; UMPIs.


## Simple Nulls and Simple Alternatives

- Neyman-Pearson Most Powerful Decision Process for each $m$.
- ROC Functions:

$$
\alpha \mapsto \pi_{m}(\alpha)
$$

- ROC functions are concave, continuous, and increasing.
- Assume that they are also twice-differentiable.


## Theorem

Multiple decision size function $\left(\alpha \mapsto A_{m}(\alpha): m \in \mathcal{M}\right)$ is optimal if it satisfies the $M+1$ equilibrium conditions

$$
\begin{gathered}
\forall m \in \mathcal{M}: \quad \pi_{m}^{\prime}\left(A_{m}\right)\left(1-A_{m}\right)=\lambda \quad \text { for some } \lambda \in \Re ; \\
\sum_{\mathcal{M}} \log \left(1-A_{m}\right)=\log (1-\alpha) .
\end{gathered}
$$

## Example: Optimal Multiple Decision Size Function

- $M=2000$
- For each $m: X_{m} \sim N\left(\mu_{m}, \sigma=1\right)$
- Multiple Decision Problem: To test

$$
H_{m 0}: \mu_{m}=0 \quad \text { versus } \quad H_{m 1}: \mu_{m}=\gamma_{m} .
$$

- Effect Sizes: $\gamma_{m} \stackrel{\text { IID }}{\sim}|N(0,3)|$
- For each $m$, Neyman-Pearson MP decision process.

$$
\begin{gathered}
\Delta_{m}=\left(\delta_{m}(\alpha): \alpha \in[0,1]\right) \\
\delta_{m}\left(x_{m} ; \alpha\right)=I\left\{x_{m} \geq \Phi^{-1}(1-\alpha)\right\}
\end{gathered}
$$

- Power or ROC Function for the $m$ th NP MP Decision Process:

$$
\alpha \mapsto \pi_{m}(\alpha)=1-\Phi\left[\Phi^{-1}(1-\alpha)-\gamma_{m}\right]
$$

## Optimal Test Sizes vs Effect Sizes



## Economic Aspect: A Size-Investing Strategy

- Do not invest your size on those where you will not make discoveries (small power) or those that you will certainly make discoveries (high power)!
- Rather, concentrate on those where it is a bit uncertain, since your differential gain in overall discovery rate would be greater!
- Some Wicked Consequences
- Departmental Merit Systems.
- Graduate Student Advising.


## BH MDF versus $\delta^{*}(q): q^{*}=.1 ; \quad ; 1000$ Reps

| $\nu$ | $p$ | $\delta_{F}^{*}$-FDR | $\delta_{F}^{*}$-MDR | $\delta^{B H}-$ FDR | $\delta^{B H_{-}-\text {MDR }^{*}}$ |
| :---: | :---: | ---: | ---: | ---: | ---: |
| 1 | 0.1 | 8.03 | 70.80 | 8.43 | 72.64 |
| 1 | 0.2 | 7.55 | 79.64 | 8.77 | 81.99 |
| 1 | 0.4 | 6.05 | 77.47 | 6.65 | 80.30 |
| 2 | 0.1 | 7.70 | 54.42 | 8.43 | 55.80 |
| 2 | 0.2 | 7.39 | 56.32 | 7.59 | 57.31 |
| 2 | 0.4 | 6.47 | 47.82 | 6.21 | 49.38 |
| 4 | 0.1 | 9.14 | 8.62 | 9.48 | 10.30 |
| 4 | 0.2 | 7.80 | 7.34 | 6.97 | 9.20 |
| 4 | 0.4 | 6.15 | 3.58 | 5.65 | 5.53 |

## Concluding Remarks

- Interplay among science, mathematics, and statistics.
- Recurrent event modeling and analysis.
- Multiple decision-making.
- Statistics, a highly promising intellectual enterprise, especially with current technology.


## Acknowledgements

- To my collaborators (current or former students of mine):
- Wensong Wu (joining as Asst Professor at Florida International University)
- Joshua Habiger (Asst Professor at Oklahoma State University)
- Akim Adekpedjou (Asst Professor at Missouri University of Science and Technology)
- Jonathan Quiton (Asst Professor at Western Kentucky University)
- Support from National Science Foundation (NSF), National Institutes of Health (NIH), and Environmental Protection Agency (EPA).
- Thanks to Dr Albacea (Zita) and the faculty at UPLB, and to my former professors and students for their hospitality during our visit here. It is wonderful to be back!

