Review of Bayesian and Frequentist Statistics

Bertrand Clarke[1]

[1]Department of Medicine
University of Miami

NDU 2011

Outline

Basic Definitions

- There are several schools of thought in Statistics. They differ primarily in how they treat variability.
- To explain them we need some definitions.
- Population: The collection of all outcomes, real or imagined, to which one wants conclusions to apply. May be natural or artificial; must be precise.
- Examples: All people born on a Tuesday since 1950. All left handed vegetarians employed at NDU (provided vegetarian is precisely defined).
- All motorized vehicles registered in Lebanon. All Lebanese residents over age 65 as of 1 July 2011. All runs of a specific experiment that might be performed. All strings of zeros and ones.

Population vs sample

- A lot of work goes into defining a population accurately.
- A sample is a subset of size n from the population.
- Samples let us make inferences about the population they represent.
- We want the population we sampled to be the same as the population we want to study.
- A frame (if it exists) is the set from which we draw our samples.
- If we take a sample of businesses that have webpages this will not be the same as the population of all businesses.
- Even if a sample is drawn from the correct population, it doesn't mean it is representative.
- Ideal: Representative sample of the population of interest.

Representative samples

- Not always possible....
- Ideal case: The selection from the population is found by 'simple random sampling'.
- A sample of size n is taken and (i) each element of the population has the same chance of inclusion in the sample (ii) the selection of one element for the sample is unaffected by the selection of any other element.
- This means all individuals and all samples are probabilistically equivalent in the sense that they are the output of the same sampling process.
- Abbreviated: IID
- Sometimes samples are dependent or not-identical and we must model this.

Random Variables

- Random variable: The process by which a measurement is generated, X.
- Outcome: The measurement generated. $X = x$
- The process of obtaining a measurement of, say, the size of a tumor is the random variable X. The measurement is $X = .5$ cm.
- A random variable has an associated probability, P. Thus: $P(X \in A)$ is the probability that the random variable gives us a value in the set A.
- We might consider th probability that a tumor of diameter at least .5cm grows. This is $P(X \geq .5)$.
- A specific outcome does not have a probability.

Technical Point

- The way I defined RV is informal. Here's the real definition:
- $X : (\mathcal{X}, \mathcal{F}) \to (|R, \mathcal{B}(|R))$ where X is measurable i.e., $\forall B \in \mathcal{B}(|R)$, $X^{-1}(B) \in \mathcal{F}$.
- We assign a probability P_R to the observation space (the range) and pull it back to give a probability on the underlying measure space.
- Thus: $P_D(X^{-1}(A)) = P_R(A)$ for $A \in \mathcal{B}(|R)$ and:
- $P_R(A) = P_D(X \in A) = P_D(X^{-1}(A)) = P_D(\omega \in \mathcal{X} | X(\omega) \in A)$ (neglecting $\{, \}$'s) and we usually drop the subscripts on P for convenience.
- We never see $(\mathcal{X}, \mathcal{F})$ and we don't know much about it....we just more or less assume it works out OK and there are theorems guaranteeing it has the properties we want.

Parameters

- We often get n outcomes $x_1, \ldots, x_n$ of a random variable X. We may denote the n draws of X by $X_1, \ldots, X_n$.
- The collection of outcomes/measurements is the sample.
- The population is fixed, but we consider different descriptions of it.
- A description of a population is given by a probability on it.
- We don't know the correct P but we may have a collection of probabilities $\mathcal{P}$ that we are sure contains the true one.
- Often $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, $\Theta \subset I\!R^d$. θ is a parameter.
- We use a function of $x_1, \ldots, x_n$ to estimate θ. Write $\hat{\theta} = \hat{\theta}(x^n)$ where $x^n = (x_1, \ldots, x_n)$.
- We write $\hat{\theta}(X^n)$, where $X^n = (X_1, \ldots, X_n)$ when we want to emphasize that $\hat{\theta}$ can be regarded as a RV in its own right.

Inference

- The basic problem of inference is to use the data i.e., the sample, to get an estimate $\hat{\theta}$ of θ_{true}, i.e., to identify the correct description of the population.
- Sometimes the parameter means something e.g., height of people. Sometimes a parameter is just an index for a collection of probabilities.
- Here, we won't usually make a differencbetween a probability, a density, and a dsitribution function since the parameter would identify any of them.
- Not enough to annouce $\hat{\theta}$...want a description of how $\hat{\theta}$ varies.
- So, we must understand models of variability.
- There are 3 major ones and several minor ones.

Survey Sampling

- The population is finite, size N.
- **The only randomness is in which sample is chosen**. An individual, once chosen, generates a measurement with no ambiguity.
- The X is the selection of an individual from the population.
- There are $\binom{N}{n}$ possible samples of size n and when we get one we use it to get a point estimate i.e., a $\hat{\theta}$.
- If we take, say, a mean, then $\bar{X}$ has a distribution generated by considering the possible samples of size n.
- So, $E(\bar{X})$ is the sum of possible values of $\bar{X}$ weighted by the probability of choosing a sample that gives that value.

Variability

- We can also find $\mathrm{Var}(\bar{X}) = E(\bar{X} - E(\bar{X}))^2$
- Usually must have $n << N$ for decent inference.
- In this case we might get a confidence interval of the form $\bar{x} \pm z_{1-\alpha/2} FPC \frac{s}{\sqrt{n}}$.
- This means that $100(1-\alpha/2)$ of the samples of size n that we might get will give an interval of the form $\bar{x} \pm z_{1-\alpha/2} FPC \frac{s}{\sqrt{n}}$ that contains $\theta = E(X)$.
- The FPC is called the finite population correction.
- Note that the variabiity is in the sample chosen and we imagine the result of choosing all samples of size n.
- This is Frequentist in that we consider the effect of repeated samples of size n and invoke the Frequency interpretation of probability.

Frequentist

- Rests on the Frequentist interpretation of probability. That is, the probability of an event A (such as tossing a coin and getting tails) is the limit

$$P(A) = \lim_{n\to\infty} \frac{\#\text{ times observed } A}{\#\text{ times we looked}}.$$

- Not a formal limit (given an ϵ you can't find an n).
- Given P_θ and n copies of X we form confidence regions.
- A $1-\alpha$ confidence region is a random set $R(X^n)$ with the property that

$$P_\theta(\theta \in R(X^n)) = 1-\alpha.$$

- Note that we have one P_θ for each X_i and another P_θ^n for X^n formed from n copies of P_θ but we don't bother to distinguish between them

Confidence Intervals

- The question is how to find CR's. Usual approach is to form an interval.
- Suppose θ is a mean $\theta = E(X)$.
- Then, if σ is known, as $n \to \infty$, we can show

$$P_\theta(\sigma z_{\alpha/2} \leq \sqrt{n}(\bar{X} - \theta) \leq z_{1-\alpha/2}\sigma) \to 1 - \alpha$$

- That is $R(X^n) = \{\sigma z_{\alpha/2} \leq \sqrt{n}(\bar{X} - \theta) \leq z_{1-\alpha/2}\sigma)\}$ is an asymptotic $1 - \alpha$ CI.
- $R(X^n) = \{\sigma z_{\alpha/2} \leq \sqrt{n}(\bar{X} - \theta) \leq z_{1-\alpha/2}\sigma)\}$ and we have one outcome of it (from the n outcomes of X).
- Frequentist prediction comes from $P_{\hat{\theta}}(\cdot)$, i.e., A has is a $1 - \alpha$ prediction region if $P_{\hat{\theta}}(X_{n+1} \in A) = 1 - \alpha$. (Not exact because it neglects variation in $\hat{\theta}$.)

Confidence

- Interpretation: $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$ is an interval produced by a technique that ensures $100(1-\alpha)\%$ of intervals so formed will contain θ.
- Confidence is a property of the [process of producing the interval,not of the numerical interval itself.
- The distribution of $\hat{\theta} = \bar{X}$ is called the sampling distribution.
- It is the central object for Frequentist inference.
- Statements about where a parameter lies retain the randomness of the data generating mechansm. It's as if we never forget that the sample we got came from a RV.
- The outcome x is what we see. The X is like the process by which we got the outcome.

Standard Error

- A Frequentist distinguishes between the SD and the SE.
- An SD is the σ for a single out come of a RV X. This is a property of the population distribution.
- An SE is the σ for a function of n outcomes of a RV X. This is a property of the sampling distribution.
- For one X, $\mathrm{Var}(X) = \sigma^2$: The X has a distribution with a density curve and we find the SD.
- For IID $X_1, \ldots, X_n$, $\mathrm{Var}(\bar{X}) = \sigma^2/n$ and this is taken in the sampling distribution for $\bar{X}$ which is derived from the distribution of X but will be much more peaked around μ.
- For INID $X_1, \ldots, X_n$, $\mathrm{Var}(\bar{X}) = \sum_{i=1}^{n} \sigma_i^2/n$ and for dependent variables, all bets are off.

In Practice

- A Frequentist chooses a class of densities $f(x|\theta)$ integrating to 1 for each θ; $f(x^n|\theta) = f(x_1|\theta)\cdots f(x_n|\theta)$.
- The MLE is a standard estimator:

$$\hat{\theta} = \arg\max_{\theta} f(x^n|\theta).$$

- Many choices of $f(\cdot|\theta)$ have a sufficient statistic: Poisson, Binomial, normal, exponential...
- Definition of sufficiency: $T(X)$ is sufficent for θ in $f(x|\theta)$ $\iff$ inference on θ only depends on T.
- Sufficient statistics contain all the information about θ in the data so functions of them are good estimators.

Other Frequentist Techniques

- A statistic T is unbiased for θ if and only if $E_\theta T(X^n) = \theta$.
- CRLB for unbiased Statistics: $\mathrm{Var}_\theta(T(X^n)) \geq \frac{1}{I_n(\theta)}$.
- UMVUE: Any statistic that achieves the CRLB for all θ in an interval.
- It turns out that UMVUE's are unique and can be given as functions of sufficient statistics.
- Given $X_1, \ldots, X_n$ put them in order from smallest to largest: $X_{(1)}, \ldots, X_{(n)}$.
- L-Statistics: Linear combinations of order statistics.
- Decision theoretic statistics....Covariates...
- These are all ways to find statistics to generate a point estimate and a sampling distribution and hence CI.

Bayesian

- The Frequentist assumes θ is fixed and the data retain their stochastic character (via Frequency interpretation).
- Bayesians reverse this: θ is a random variable Θ and the data, once obtained are treated as fixed. So, you condition on them.
- Where does the distribution on Θ come from? We make it up. Call it the density $w(\theta)$. We still have the conditional density for X given $\Theta = \theta$ that we write as $f(\cdot|\theta)$.
- Joint density for $\Theta, X^n)$ is

$$w(\theta)f(x^n|\theta) = w(x^n|\theta)m(x^n)$$

where $m(x^n) = \int w(\theta)f(x^n|\theta)d\theta$ is called the mixture distribution or the marginal for the data.

Bayesian Inference

- Bayesians make inferences from the posterior $w(\theta|x^n)$.
- A $1-\alpha$ credible set $R = R_\alpha(x^n)$ is any set of parameter values satisfying

$$W(R|x^n) = 1 - \alpha.$$

- This does not require the Frequency interpretation.
- The interpretation of a credibility region R is that conditional on the data, we have a set that contains $1-\alpha$ posterior probability.
- The posterior density is the Bayesian's analog to the sampling distribution.
- No repeated sampling assumption (which might not be satisfied) just a direct statement about where θ is – conditional on the data

Choosing the prior

- There are two approaches to prior selection.
- Subjective: (snide) The investigator consults his/her feelings and impressions about where θ might be and draws curves to represent this trying to find a mathematical form that they fit.
- Subjective: (fair) The investigator reflects carefully on the relevant information about θ that might be available and tries to formulate a prior density that summarises this.
- Objective: The prior is chosen by some kind of auxilliary principle, e.g., noninformativity, usually an optimization or invariance criterion.
- Choose a class of priors and evaluate the stability of inferences to over the class.

Types of Bayes Estimator

- Decision theoretic: Choose a loss function $L(\cdot,\cdot)$ and find

$$\delta_B(x^n) = \arg\min_{\delta} \int L(\theta, \delta(x^n)) w(\theta|x^n) d\theta$$

  The integral is called the posterior risk of δ.
- Posterior mode: Analogous to the MLE, choose

$$\hat{\theta}_{PM} = \arg\max_{\theta} \; w(\theta|x^n)$$

  Actually, $|\hat{\theta}_{MLE} - \hat{\theta}_{PM}| = \mathcal{O}_P(1/n)$.
- Conventionally, the Bayesian wants to see the whole posterior because the shape of the curve explains the variability better than $\mathrm{Var}(\Theta|X = x)$ can.

Variability

- As noted, given $w(\theta|X^n)$, the Bayesian might use the posterior variance

$$\int (\theta - E(\Theta|X^n = x^n))^2 w(\theta|x^n) d\theta$$

  where $E(\Theta|x^n) = \int \theta w(\theta|x^n) d\theta$ is the posterior mean.
- Just as $\bar{X} \to \mu$ and $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \to N(0,1)$, posterior quantities have analogous properies.
- Bayesian LLN: $E(\bar{\Theta}|X^n) \to \mu$.
- Bayesian CLT: $w((\theta - \hat{\theta})/(\sqrt{n}\sqrt{I(\hat{\theta})}) \to N(0,1)$. (Note $\hat{\theta} = E(\bar{\Theta}|X^n)$ is one choice.)

Where is the variability?

- Importance of variability: 1 m $\pm$ 1cm vs, 1m $\pm$ 1 km.
- The Bayesian says the data, once obtained, are no longer stochastic. They are the fixed outcomes of a RV and so you condition on them.
- The Bayesian says the variability is transmuted from the data to the parameter by way of the posterior distribution for the parameter that is conditional on the data.
- Thus, the Survey Sampler thinks in terms of subsets of a specific population; a Frequentist thinks in terms of repeated sampling; a Bayesian thinks of what the resuling posterior says about the parameter given the data.
- Outcomes remain stochastic for the Frequentist, not the Bayesian.

Special Cases

- Conjugate priors: Choose a prior from a class so that the posterior is in the same class. Depends on the likelihood.
- Non-parametric Bayes is exactly the same structure as parametric Bayes.
- Bayesian prediction comes from the predictive distribution:

$$m(x_{n+1}|x^n) = \int f(x_{n+1}|\theta) w(\theta|x^n) d\theta.$$

That is, A is a $1-\alpha$ prediction region if $M_{n+1}(X_{n+1} \in A|x^n) = 1-\alpha$, still conditional on x^n, like Frequentist case.

- In IID cases, Bayes and Frequentist methods for estimation are often asymptotically equivalent.

Testing

- Bigger differences in hypothesis testing: The p-value is obtained from the sampling distribution and has a very different meaning than a Bayes factor.
- Bayes testing is]decision-theoretic using 0-1 loss.
- Bayes testing of $\mathcal{H}_0 : \theta \in S$ vs $\mathcal{H}_1 : \theta \in S^c$ based on $W(S|x^n)$ or, equivalently, on the Bayes factor

$$BF(1;2) = \frac{W(S|x^n)/W(S)}{W(S^c|x^n)/W(S^c)}.$$

- This is the ratio of the prior odds to the posterior odds.
- Contrast: Frequentist testing uses the Neyman-Pearson Lemma which is an optimization of $P(\text{reject } \mathcal{H}_0|\mathcal{H}_1 \text{true })$ subject to $P(\text{reject } \mathcal{H}_0|\mathcal{H}_0 \text{true }) \leq 1 - \alpha$, both proabilities are in the sampling distribution of the test statistic.

Likelihood; Information-Theoretic

- Likelihood $=$ conditional density for X given θ but regarded as a function of θ for fixed $X = x$, $L(\theta|x) = f(x|\theta)$.
- LP: All inferences should come only from the Likelihood.
- Get intervals like $\{\theta|\ L(\hat{\theta}|x^n) - L(\theta|x^n) \leq t\}$ for thresholds t. No notion of confidence or credibility.
- Information-theoretic: The idea is that the central features of models and data are expressible in terms of measures of complexity (Kolmogorov, VC-dimension, codelength).
- e.g., choose a model $\hat{p}$

$$\hat{p} = \arg\min_{p \in \mathcal{P}}\ L(p) + L(x^n|p)$$

where $L(\cdot)$ is the Shannon codelength for x^n given p or p.
- Includes maxent, rel. entropy criteria, MML, MDL, etc etc.

Predictive

- Prequential Principle, Dawid (1984).
- 'The method of evaluation of a predictor should be disjoint from its method of construction, e.g., depend only on the predictions it makes and the future data.'
- Typically look at something like

$$CPE = \sum_{i=n_1}^{n_2} (\hat{Y}_{i+1}(x_{i+1}; x^i) - Y_{i+1}(x_{i+1}))^2.$$

  as a way to evaluate a predictor $\hat{Y}_{i+1}$.
- Inference comes from prediction errors.
- Other principles: variance/bias, robustness, complexity etc.
- Predictive criteria are most important with complex and high dimensional data where modeling is impossible.

Another perspective...

- Fiducialist: Wang, Hannig, Iyer (2011). This was a weird idea due to Fisher that never worked but from time to time people try to make it work.
- Regard X as $X = G(\theta, U)$, U = error distribution. Define $Q(x, u) = \{\theta |\ G(\theta, u) = x\}$; Q is like a G^{-1} for fixed θ.
- Fiducial distribution for θ is: $Q(x, U^*)|Q(x, U^*) \neq \phi)$, where U is an IID copy of U.
- Still fairly complicated and under development, but an interesting alternative.

Frequentist

- $X \sim N(\mu, \sigma^2)$ has density

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

- n IID outcomes $X^n = x^n$ satisfy:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad (n-1)S^2/\sigma^2 = \frac{1}{\sigma^2}\sum_{i-1}^{n}(x_i - \bar{x})^2 \sim \chi^2_{n-1}$$

- $\bar{X}$ and S^2 are independent. So, $\sqrt{n}(\bar{X}-\mu)/\sigma \sim N(0,1)$ and $\sqrt{n}(\bar{X}-\mu)/s \sim t_{n-1}$ gives CI's for μ and σ.
- $t_{n-1} = N(0,1)/\sqrt{\chi^2/n}$; tends to $N(0,1)$ as $n \to \infty$ but otherwise has heavier tails; t_{k+1} has moments of order at most k ($t_{k+1} \sim \mathcal{O}(x^{k+1})$).

Bayes version

- X IID $N(\mu, \sigma^2)$...assume σ fixed, estimate μ. (If we use $\bar{X}$, set $\sigma^2 = \sigma'^2/n$.)
- Use a prior $w(\mu)$ on μ: $\mu \sim N(\theta, \tau^2)$, θ, τ known.
- Joint distribution:

$$h(\mu, x) = w(\mu) f(x|\mu) = \frac{1}{2\pi\sigma\tau} e^{-(1/2)[(\mu-\theta)/\tau^2 + (x-\mu)/\sigma^2]}$$

- Bayes rule: $w(\theta) f(x|\mu) = w(\mu|x) m(x)$....so to find $w(\theta|x)$ let $\rho = (\tau^2 + \sigma^2)/\tau^2\sigma^2$ and complete the square:

$$(\mu-\theta)/\tau^2 + (x-\mu)/\sigma^2 = \frac{\rho}{2}[\mu - \frac{1}{\rho}(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2})] + \frac{(\theta - x)^2}{2(\sigma^2 + \tau^2)}.$$

This gives a useful form for $h(\mu, x)$.

Bayes continued

- To find $w(\mu|x)$, divide $h(x,\mu)$ by $m(x)$:

$$m(x) = \int h(x,\mu)d\mu = \frac{1}{\sqrt{2\pi\rho\sigma\tau}} e^{-(\mu-x)^2/2(\sigma^2+\tau^2)} = N(\mu, \sigma^2+\tau^2).$$

- Now, $w(\mu|x) = h(x,\mu)/m(x)$ equals

$$\sqrt{\frac{\rho}{2\pi}} e^{-\rho/2(\mu-(1/\rho)(\theta/\tau^2+x/\sigma^2))^2}.$$

- Thus, $(\mu|x)$ has a $N(\mu(x), 1/\rho)$ distribution where

$$\mu(x) = \frac{\sigma^2}{\sigma^2+\tau^2}\theta + \frac{\tau^2}{\sigma^2+\tau^2}x$$

- We can now get credible sets from $N(\mu(x), 1/\rho)$ for any x.
- If σ not known, we still get the t_{n-1} distribution...

Bayesian, unknown σ

- Let $X_i \sim N(\mu, \sigma^2)$ IID for $i = 1, \ldots, n$ and assume $w(\mu, \sigma^2) \propto 1/\sigma^2$.
- It can be shown that $w(\mu|\sigma^2, x^n) \sim N(\bar{x}, \sigma^2)$.
- Write $w(\mu|x^n)$, then

$$w(\mu|x^n) = \int w(\mu|\sigma^2, x^n) d\sigma^2.$$

- It can be shown that $w(\mu|x^n) \sim t_{n-1}(\bar{x}, s^2/n)$. That is, $w((\mu - \bar{x})/(s/\sqrt{n})|x^n) \sim t_{n-1}$.
- From $w(\mu, \sigma|x^n) \propto w(\mu, \sigma^2) f(x^n|\mu, \sigma^2)$ we can get

$$w(\sigma^2|x^n) \propto \int \left(\frac{1}{\sigma^2}\right) e^{-(1/2\sigma^2)(n-1)s^2 + n(\bar{x}-\mu)^2} d\mu,$$

i.e., an Inv-χ^2_{n-1,s^2} = Inv-gamma$((n-1)/2, (n-1)s^2/2)$.

Information-theory

- The entropy of a RV is $H(X) = -\int p(x) \log p(x) dx$; the entropy is the minimal noiseless codelength.
- The normal family can be derived from the following.
- Consider k functions f_1,...,f_k and numbers a_1, ..., a_k and suppose $E(f_j(X)) = a_j$ for $j = 1, \ldots, k$. The maximum entropy distribution for X, if it exists, is

$$p(x) = ce^{\sum_{j=1}^{k} \lambda_j f_j(x)}$$

  where c and the λ_j's are found to make $\int p(x)dx = 1$.
- Given the first two moments, solving this gives the normal. That is, if $k = 2$, let f_1 be X $f_2 = X^2$.
- Another coding argument gives estimates for μ and σ by two-stage coding Barron and Cover (1990).

Predictive

- Use x^n to predict X_{n+1} – analog to estimating μ.
- Consider sample mean $\bar{x}$. If no model is assumed, set $\mu = EX_i$ and $\sigma^2 = \mathrm{Var}(Y_i)$.
- Use standard inequalities (Markov, triangle, Cauchy-Schwarz) to obtain for given $\tau > 0$.

$$\begin{aligned} P & \left(|\bar{Y} - Y_{n+1}| \geq \frac{\sigma(1 + (1/\sqrt{n}))}{\tau} \right) \\ & \leq \frac{\tau}{\sigma(1 + (1/\sqrt{n}))} \left((E|\bar{Y} - \mu|^2)^{1/2} + (E|\mu - Y_{n+1}|^2)^{1/2} \right) \\ & \leq \tau. \end{aligned}$$

- For $\tau < 1$, the Frequentist PI for known σ is

$$\bar{Y} \pm \sigma(1 + (1/\sqrt{n}))/\tau.$$

Normal Case

- If $X_i \sim N(\mu, \sigma^2)$, $\bar{X} - X_{n+1} \sim N(0, \sigma^2(1 + \frac{1}{n}))$ and the prediction interval becomes $\bar{X} \pm z_{1-\alpha}\sigma(1 + (1/n))^{1/2}$ where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of $N(0,1)$.
- So, if $\tau = 1/z_{1-\alpha}$, the only difference between the general case and the normal case is $(1 + (1/n))^{1/2}$ versus $(1 + 1/\sqrt{n})$ which is asymptotically negligible.
- If σ is unknown, the PI's become

$$\bar{Y} \pm \hat{\sigma}\frac{\sqrt{n-1}\sqrt{1+(1/n)}}{\tau\sqrt{n-3}},$$

- An exact form for the normal can be found in this case too. See Geisser (1995), Chap. 2.

Bayes Prediction

- As above, the posterior mean is $E(\Theta|x^n) = \tau^2/(\sigma^2/n+\tau^2)\bar{x} + (\sigma^2/n)/(\sigma^2/n+\tau^2)\mu$ and the posterior variance is $1/\rho = \tau^2\sigma^2/(n\tau^2+\sigma^2) = \mathcal{O}(1/n)$. So,

$$\begin{aligned} m(x_{n+1}|x^n) &= N(E(X_{n+1}|X^n = x^n), \mathrm{Var}(X_{n+1}|X^n = x^n)), \\ E(X_{n+1}|X^n = x^n) &= E(\Theta|X^n = x^n), \end{aligned}$$

  and

$$\mathrm{Var}(X_{n+1}|X^n = x^n) = \sigma^2 + \mathrm{Var}(\Theta|X^n = x^n).$$

- PI is now $E(\Theta|X^n = x^n) \pm z_{\alpha/2}\mathrm{Var}(X_{n+1}|X^n = x^n)$.
- As before, this can be extended to the case that σ is not known by putting a prior on it.

Exponential Famlies

- A parametric family $f(x|\theta)$ is of exponential form $\iff$ $\exists K$ $f(x|\theta)$ can be written as

$$f(x|\theta) = h(x)c(\theta)e^{\sum_{k=1}^{K} w_k(\theta)t_k(x)}.$$

- Support of X is independent of θ.
- The K functions t_k are sufficent for θ in X, even as n increases.
- Convenient with independence: exponents add.
- Natural form: Replace $w_k(\theta)$ by η.
- $c(\theta)$ is the normalizing constant, $h(x)$ independent of x
- Examples: Normal, Poisson, Binomial, Exponential, Gamma, Dirichlet, χ^2_k....etc
- Non-exponential families: $Unif[0,\theta]$, Cauchy, Laplace, t_n,...

Sufficiency

- $T(X)$ is sufficent for θ in $f(x|\theta) \iff$ the conditional distribution for X given $T(X)$ is independent of θ.
- Formally, T is sufficent $\iff$ $P(X = x|T(X) = t, \theta) = P(X = x|T(X) = t)$.
- Fisher's Factorization criterion for sufficiency:

$$\text{T is sufficient for } \theta \iff f(x^n|\theta) = g(T(x^n)|\theta)h(x^n)$$

- T partitions the sample space into sets $\tau_t = \{x^n|T(x^n) = t\}$ so two x^n's in the same τ lead to the same infrences.
- If T is sufficient, Frequentists like to use it for inference (esp.if minimal).
- If T is sufficent than $w(\theta|x^n) = w(\theta|t(x^n))$.

Examples of Sufficiency

- If X^n are IID from a full rank K natural parameter exponential family, then then $T(x^n) = (\sum_{i=1}^n t_1(x_1), \ldots, \sum_{i=1}^n t_K(x_1))$ is (minimal) sufficient for θ.
- Sufficent statistics are not unique – any one-one function of a sufficient statistic is also sufficient.
- A sufficent statistic T is minimal $\iff$ for any other sufficient statistic S, $\exists g$ so that $T = g(S)$.
- Test for minimality: Suppose:

$$\frac{f(x^n|\theta)}{f(y^n|\theta)} \quad \text{constant as a function of } \theta \iff T(x^n) = T(y^n).$$

Then T is minimal sufficient.

An Unusual Example

- Let $X_i \sim Unif[\theta, \theta + 1]$ be IID. Joint pdf is

$$f(x^n|\theta) = \begin{cases} 1, & x_{(n)} - 1 < \theta < x_{(1)} \\ 0, & otherwise \end{cases}$$

- Check the condition of the theorem:

$$\frac{f(x^n|\theta)}{f(y^n|\theta)} \text{ constant in } \theta \iff \begin{cases} x_{(n)} = y_{(n)} \\ x_{(1)} = x_{(1)} \end{cases}$$

- So, $T(X^n) = (X_{(1)}, X_{(n)})$ is minimal sufficient.
- T is two-dimensional but θ is unidimensional, f is not an exponential family.
- In a full rank exponential family of order K, there are K sufficient statistics.

Sufficiency and Ancillarity

- One extreme: The ordered data $X_{(1)}$, ..., $X_{(n)}$ are always a sufficent statistic.
- The other extreme is that exponential famlies are almost characterized by having a finite dimensional sufficient statistic. Suppose X^n are IID $f(\cdot|\theta)$.
- Then, f_θ has support independent of θ and there is a K dimensional sufficient statistics $\iff$ $f)\cdot|\theta)$ is of exponential form.
- What about functions of the data that are not sufficient? The question is whether they depend on the parameter...or have other information about the parameter.
- A statistic $S(x^n)$ is ancillary $\iff$ distribution of $S(x^n)$ does not depend on θ.

Properties of Ancillarity

- Example: Location scale family: Suppose we have a RV Z with density f. Write $X = \sigma Z + \mu$. Now, X has density $f(x|\mu,\sigma) = (1/\sigma)f((x-\mu)/\sigma)$
- Assume IID data from $f(x|\mu,\sigma)$ and let $R = X_{(n)} - X_{(1)}$ be the range. DF of R is

$$\begin{aligned} F_{R,\mu}(r) &= P_\mu(R \le r) = P_\mu(X_{(n)} - X_{(1)} \le r) \\ &= P_\mu((X_{(n)} - \mu) - (X_{(1)} - \mu) \le r) \\ &= P_f(Z_{(n)} - Z_{(1)} \le r) \end{aligned}$$

  which is constant in μ. So, R is ancillary.
- In a scale family, any statistic that is a function of X_i/X_n will be ancillary. (Think of the distribution function of $(1/X_n)(X_1,\ldots,X_{n-1})$.

Completeness

- A family of densities $f(T|\theta)$ for a statistic T is complete $\iff \forall\theta\ E_\theta g(T) = 0$ implies $g = 0$ a.e.
- $\bar{X}$ is complte for $N(\mu, 1)$.
- In a full rank exponential family $T(X^n) = (\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_K(X_i)$ is complete.
- Basu's theorem: If T is complete and minimal sufficient then it is independent of every ancillary statistic.
- Curious fact: The distribution of an ancillary statistic does not depend on θ, but there can be information in the ancillary about θ (Fraser-Monette example). Basu's theorem eliminates this possibility.
- A minimal sufficent statistic therefore has all the information in the data about the parameter.

Natural Sufficient Statistic

- Let X^n be an IID sample from an exponential family and write $T_k(X^n) = \sum_{i=1}^n T_k(X_i)$.
- Suppose $(w_1(\theta), \ldots, w_K(\theta))$ contains an open set in $I\!R^K$ and $(T_1(X^n), \ldots, T_K(X^n))$ contains an open set in $I\!R^K$.
- Then: The distribution of $(T_1, \ldots, T_K)$ as RV's is

$$f_T(u_1, \ldots, u_K|\theta) = H(u_1, \ldots, u_K)c^n(\theta)e^{\sum_{k=1}^K w_k(\theta)U_k}.$$

- The sampling distribution of the complete, minimal sufficient summary statistics of an exponential family is of exponential form.

MOM

- Suppose X^n is IID $f(x|\theta_1, \ldots, \theta_K)$.
- Equate the first K sample moments to the first K population moments and solve for estimates of the θ_k's.
- Write $m_1 = (1/n)\sum_{i=1}^n X_i$, ..., $m_K = (1/n)\sum_{i=1}^n X^K$.
- Write $\mu_1(\theta_1^K) = E(X)$, ..., $\mu_K(\theta_1^K) = E(X^K)$.
- Solve K equations: $M_k = \mu_k(\theta_1^K)$ for the K unknowns θ_1^K.
- $N(\mu, \sigma^2)$: We get $\hat{\mu} = \bar{X}$ and $E(X^2) = (1/n)\sum_{i=1}^n X_i^2$. So, $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n X_i^2 - \hat{\mu}^2$.
- $X_i \sim Bin(n, p)$: $\bar{X} = E(X) = np$ and $\bar{X^2} = np(1-p) + (np)^2$. Solving gives $\hat{p} = \bar{X}/\hat{n}$ and

$$\hat{n} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum_{i=1}^n (X_i - \bar{X})^2}. \leftarrow \text{can be} < 0 \text{ !}$$

- Can use CLT's to get asymptoic normality and CI's

MLE's

- Recall the MLE is $\hat{\theta} = \arg\max_\theta f(x^n|\theta)$.
- Properties: Not always unique: $X \sim Unif[\theta, \theta+1]$

$$L(\theta|x) = \begin{cases} 1, & \theta < 1 < \theta + 1 \\ 0, & otherwise \end{cases}$$

so any $\theta < 1$ is an MLE.

- Discrete case: $X_i \sim Bin(n, p)$. MLE for n? Then

$$L(n, p|x^n) = \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i}(1-p)^{1-x_i}.$$

- Observe $L(n|x^n) = 0$ for $n < \max_i x_i$ so $n \geq max_i x_i$
- Want least n satisfying $L(n|x^n) \geq L(n-1|x^n)$ and $L(n+1|x^n) \leq L(n|x^n)$, then argue uniqueness.

Uniqueness Result for MLE's

- May need tricks to find MLE's...can't always solve $(\log L(\theta|x))' = 0$. Discrete parameters, and boundary problems of parameter space may be problems.
- Often helpful to take logs...
- In exponential families we have MLE's: Let

$$f(x|\theta) = e^{\sum_{k=1}^{K} w_k(\theta) T_k(X) + \log c(\theta) + \log h(X)}$$

- If C is the interior of the range of $(w_1(\theta), \ldots, w_K(\theta)) \subset I\!R^K$ for $\theta \in \Theta$,
- and the equations $E_\theta T_k(X) = T_k(x^n)$ have a solution say $\hat{\theta} = (\hat{\theta}_1(x^n), \ldots, \hat{\theta}_1(x^n))$ for which $(c_1(\hat{\theta}_1), \ldots, c_K(\hat{\theta}_k)) \in C$
- Then the MLE is unique.

Invariance

- If $\hat{\theta} = \text{MLE}(\theta)$, then for any $\tau(\theta)$, $\text{MLE}(\tau) = \tau(\hat{\theta})$.
- Proof: If τ is one-one and $\eta = \tau(\theta)$ then take sup's on both sides o $L^*(\eta|x^n) = L(\tau^{-1}(\eta)|x^n) = L(\theta|x^n)$.
- If τ is not one-one, let $L^*(\eta|x^n) = \sup_{\{\theta|\ \tau(\theta)=\eta\}} L(\theta|x^n)$ and proceed as before.
- Example: Suppose $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$
- For fixed σ, can find MLE's for β_0 and β_1 from minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This is the exponent in the normal so $\text{MLE}(\beta_0, \beta_1)$ is the LSE. Putting $\hat{\beta}_0$ and $\hat{\beta}_1$ into $L'(\hat{\beta}_0, \hat{\beta}_1, \sigma^2|Y_1^n) = 0$ gives $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

Asymptotics of the MLE I

- A sequence of estimators $W_n = W_n(X^n)$ is consistent for $\tau(\theta) \iff \forall\theta\ W_n \xrightarrow{p} \tau(\theta)$.
- A sequence of estimators W_n is asymptotically efficient for $\tau(\theta) \iff W_n$ achieves the CRLB in the limit of large n:

$$\lim_{n\to\infty} \frac{\mathrm{Var}_\theta(W_n)}{(\tau'(\theta)^2/nI(\theta))}$$

  where $I(\theta) = E_\theta((\partial/\partial\theta)\log f(X|\theta))^2$.
- Asymptotic normality means $\exists V >$ such that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, V)$.
- MLE's are consistent, asymptotically normal and efficient.

Asymptotics of the MLE II

- Under 'regularity' conditions we have that 1) $\hat{\theta} \xrightarrow{p_\theta} \theta$ 2) $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1/I(\theta))$
- Regularity conditions: (i) $(\partial/\partial\theta)^i f(\cdot|\theta)$ exists for $i = 1, 2, 3$, (ii) $(\partial/\partial\theta)^i f(\cdot|\theta)$ bounded by integrable functions of x locally in θ i=1, 2, 3 , (iii) $E[(\partial/\partial\theta) \ln f(X|\theta)]^2 < \infty$.
- $h(\hat{\theta})$ is also $AN(h(\theta), (h'(\theta)/nI(\theta))^2)$.
- Same result holds with $\hat{I}(\hat{\theta}) = (1/n)\sum_{i=1}^{n}(\partial^2/\partial\theta^2) \ln f(\cdot|\theta) \underset{\rightarrow}{p_\theta} I(\theta)$.
- Multivariate version too: Under regularity conditions, $(\hat{\theta}_1, \ldots, \hat{\theta}_K) \xrightarrow{p_\theta} (\theta_1, \ldots, \theta_K)$
- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, I(\theta)^{-1})$.

Best Unbiased

- The MSE of an estimator $\hat{\theta}$ of θ is

$$E_\theta(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}))^2 + (E\hat{\theta} - \theta)^2$$

  i.e., MSE = Variance plus bias-squared.
- Let's search the collection of unbiased estimators for the one with the smallest variance.
- Definition: $\hat{\theta}$ is the best unbiased estimator of $\theta \iff$ $E_\theta\hat{\theta} = \theta$ and $\forall\theta$ and $\forall$ unbiased θ^* $\text{Var}_\theta(\hat{\theta}) \leq \text{Var}_\theta(\theta^*)$.
- In Poisson(λ), $\hat{\lambda} = \bar{X}$ and S^2 both estimate $\text{Var}_\lambda(X) = \lambda$. Can show $\bar{X}$ has a lower variance than S^2.

- For IID data, the smallest variance given by CRLB:

$$\mathrm{Var}_\theta(W_n) \geq (E_\theta W_n)'/nI(\theta).$$

- If $X \sim f_\theta$ independent of $Y \sim g_\theta$ then $I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta)$.
- Fisher information for Poisson(λ) is $I(\lambda) = 1/\lambda$. Since $\lambda' = 1$ and $\mathrm{Var}_\lambda(\bar{X}) = \lambda/n$, we see that $\bar{X}$ attains the CRLB and so is best unbiased or UMVU.
- Roughly, if X_i's are IID $f(\cdot|\theta)$ and $W(X^n)$ is unbiased for $\tau(\theta)$ then W_n attains the CRLB $\iff$ $\exists a(\theta)$ so that

$$a(\theta)(W(X^n) - \tau(\theta)) = \frac{\partial}{\partial\theta} \ln f(x^n|\theta).$$

- Strange fact: The best unbiased estimator for σ^2 in $N(\mu, \sigma^2)$ is $(1/n)\sum(X_i - \mu)^2$, so, if μ is unknown, you can't attain the CRLB.

UMVU and Sufficiency

- Rao-Blackwell Theorem: Let W be unbiased for $\tau(\theta)$ and let T be sufficient for θ. Let $\phi(T) = E(W|T)$. Then: (1) $E\phi(T) = \tau(\theta)$ and (2) $\forall\theta$ we have $\text{Var}_\theta(\phi(T)) \leq \text{Var}_\theta(W)$.
- UMVUE's are unique
- Let W be unbaised for $\tau(\theta)$. Then τ is UMVU $\iff$ W is uncorrelated with any unbiased estimator of 0.
- Theorem: Let T be complete nd sufficient for θ and let $\phi(T)$ be an estimator. Then, $\phi(T)$ is the unique UMVUE for $E_\theta\phi(T)$.
- Lehmann-Scheffe Theorem: Let T be complete and sufficient for θ and let $h(X)$ be unbiased for $\tau(\theta)$. Then $W = \phi(T) = E(h(X)|T)$ is UMVUE for $\tau(\theta)$.

Example

- E.g., $X_i \sim Bin(n, p)$ IID. Let $\tau(p) = P_p(\text{one success})$. Then, $T(X^n) = \sum X_i$ is complete and sufficient, but not unbiased.
- Let $h(X) = \chi X = 1$ then $E(h(X)) = \tau(\theta)$
- Theorem $\Rightarrow \phi(T) = E(h(X)|T)$ is UMVU for $\tau(\theta)$.
- Work out what ϕ is:
  $\phi(t) = E(h(X)|T = t) = P(X_1 = 1|T = t)$
- Writing out the conditiona probability and simplifying gives

$$\phi(T) = n\binom{n-1}{T-1}/\binom{n}{T}.$$

Neyman-Pearson Framework

- Basic hypothesis testing problem: $\mathcal{H} : \theta \in \Omega_H$ vs $\mathcal{K} : \theta \in \Omega_K$, $\Omega_H \cap \Omega_K = \varnothing$.
- Suppose we base our decision on an outcome of X.
- A test is defined by a region S and $x \in S$ means do not reject $\mathcal{H}$ while $x \in S^c$ means reject $\mathcal{H}$.
- 4 cases: $\theta \in \mathcal{H}, \mathcal{K}$ and we choose $\mathcal{H}, \mathcal{K}$.
- Want $P(\text{Type I error}) = P_{\mathcal{H}}(\text{reject } \mathcal{H})$ low.
- Want $P(\text{Type II error}) = 1 - P_{\mathcal{K}}(\text{reject } \mathcal{H})$ high.
- Power function is $P_\theta(\text{reject } \mathcal{H})$, a function of θ. Want power low on $\Omega_{\mathcal{H}}$ and high on $\Omega_{\mathcal{K}}$.

NPFL intuition

- A test defined by a region S is level $\alpha \Leftrightarrow P_\theta(S^c) \leq \alpha$ for all $\theta \in \Omega_H$.
- Suppose $\Omega_H = P_0$ and $\Omega_K = P_1$. Then, we want to find points for a set S so that we can reject on S^c and

$$\sum_{x \in S^c} P_0(x) \leq \alpha \quad \text{and} \quad \sum_{x \in S^c} P_1(x) \text{ maximum.}$$

- The most valuable points have a high value of $P_1(x)/P_2(x)$.
- So, rank the points in decreasing order by $P_1(x)/P_2(x)$ and then start putting them in until you hit α in terms of P_0.
- This leads to the Most Powerful test in the simple-vs-simple case.

NPFL

- Let P_0 and P_1 have pdf's p_0 and p_1.
- For testing $\mathcal{H} : P_0$ vs $\mathcal{K} : P_1$ there is a critical function ϕ and a constant k so that (1) $E(\phi) = \alpha$ and (2)

$$\phi(x) = \begin{cases} 1 & p_1(x) > kp_0(x) \\ 0 & p_1(x) < kp_0(x) \end{cases}$$

- If a ϕ satisfies (1) and (2) for some k, it is MP for P_0 vs P_1.
- If ϕ is a MP level α test for P_0 vs P_1 then $\exists k$ so that (2) is satisfied.

MP and Sufficiency

- Consider $\mathcal{H} : \theta = \theta_0$ vs $\mathcal{K} : \theta = \theta_1$. Suppose T is sufficient for θ with density $g(T|\theta)$. Then:
- Any test based on T with rejection region S is an MP level α test if it satifies

$$\begin{cases} g(t|\theta_1) > kg(t|\theta_0) \Rightarrow t \in S \\ g(t|\theta_1) < kg(t|\theta_0) \Rightarrow t \in S^c \end{cases}$$

  for some $k > 0$ where $\alpha = P_{\theta_0}(T \in S)$
- We rarely test point nulls aginst point nulls...so we want a concept of uniformly most powerful i.e., tests that are good for composite hypotheses.

Uniform NPFL

- Consider $\mathcal{H} : \theta \in \Omega_H$ vs $\mathcal{K} : \theta \in \Omega_K$, with $\Omega_H = \Omega_K^c$ and suppose we have a test based on a sufficient statistic T with pdf $g(t|\theta)$ and rejection region S.
- If (1) the test is level α, (2) $\exists \theta_0 \in \Omega_H$ with $P_{\theta_0}(S) = \alpha$, and (3) For the same θ_0 as in (2) we have $\forall \theta^* \in \Omega_K$ there is a $K > 0$ so that

$$\begin{cases} g(t|\theta^*) > k g(t|\theta_0) \Rightarrow t \in S \\ g(t|\theta^*) < k g(t|\theta_0) \Rightarrow t \in S^c. \end{cases}$$

  Then: This test is UMP level α for $\mathcal{H}$ vs $\mathcal{K}$.
- This result works for many one-sided tests, in particular for one-dimensional exponential families.
- Also....Monotone Likelihood Ratio, unbiasedness, etc.

Examples

- We already saw that a normal prior on a normal likelihood gave a normal posterior for estimating the mean.
- Let $\mathcal{P}$ be a class of prior densities and let $\mathcal{F}$ be a class of densities from a parametric family.
- $\mathcal{P}$ is conjugate to $\mathcal{F} \iff \forall f \in \mathcal{F}, \forall w \in \mathcal{P}, \forall x^n w(\theta|x^n) \in \mathcal{P}$. That is, the posterior is in the same class as the prior.
- Suppose X_i are IID Poisson(λ) so $f(x^n|\lambda) = \lambda^{n\bar{x}} e^{-n\lambda} / \prod_{i=1}^n x_i!$. Let $\lambda \sim Gamma(\alpha, \beta)$. Then, the joint density is

$$h(x^n, \lambda) = \frac{\lambda^{n\bar{x}+\alpha-1} e^{-\lambda(n+1/\beta)}}{\Gamma(\alpha)\beta^\alpha \pi x_i!}.$$

- So, $w(\lambda|x^n)$ is a Gamma$(n\bar{x} + \alpha, (n + 1/\beta)^{-1})$.

Conjugacy and Exponential Families

- Recall the exponential family $f(x^n|\theta) = e^{\sum_{k=1}^{K} c_k(\theta) \sum_{i=1}^{n} T_k(x_i) + \sum_{i=1}^{n} S(x_i) + nd(\theta)}$.
- We obtain a conjugate prior by using the sufficient statistics and n.
- Let $t_k = \sum_{i=1}^{n} T_k(x_i)$ for $k = 1, \ldots, K$ and $t_{K+1} = n$.
- Let $w(t_1, \ldots, t_{k+1}) = \int e^{\sum_{k=1}^{K} c_k(\theta) t_k + t_{K+1} d(\theta)} d\theta$.
- Let $\Omega = \{t_1^{K+1} |\ w(t_1^{K+1}) < \infty\}$
- The $K+1$ parameter exponential family given by

$$w(\theta|t_1^{K+1}) = e^{-\sum_{k=1}^{K} c_k(\theta) t_k + t_{K+1} d(\theta) - \ln w(t_1^{K+1})}$$

is conjugate to $f(x^n|\theta)$ and $w(t_1^{K+1})$ is the normalizing constant so the data is fixed.

Basic Setup

- If you have a good idea about the various costs of the ways to be wrong you can make decisions that minimize the average cost of errors.
- We have a model $f(\cdot|\theta)$, a parameter space Θ, data x from a sample space. When we get data we make a decision about where θ is by using a rule δ. The collection of rules we allow ourselves is $\mathcal{A}$, the action space. We measure cost by a loss function $L : \Theta \times \mathcal{A} \to I\!R$
- $\mathcal{A}$ can be Θ (estimation) or $\{\text{accept}\mathcal{H}, \text{reject}\mathcal{H}\}$ (testing).
- $L(\theta, \delta(x))$ is the cost of $\delta(x)$ when θ is the state of nature.
- How to choose a good δ?
- Try averaging: $R(\theta, \delta) = E_\theta L(\theta, \delta(X))$, the expected loss.

Bayes Optimality

- $R(\theta, \delta)$ is called the risk of δ.
- Smaller risk is better, so we can compare curves $g_\delta(\theta) = R(\theta, \delta)$ for various δ. If the curve for a δ does not admit a uniform improvement then δ is called admissible. This is hard to work with.
- Assume a prior $w(\theta)$ and form the Bayes risk:

$$R(w, \delta) = E_w R(\Theta, \delta) = \int w(\theta) R(\theta, \delta) d\theta$$

$$= \int m(x) w(\theta|x) L(\theta, \delta(x)) dx d\theta = E_m \left[E_{w(\cdot|x)} L(\Theta, \delta(X)) \right]$$

The posterior risk (in [ ]'s) gives the Bayes estimator:

$$\delta_B = \arg\min R(w, \delta) = \arg\min R(w, \delta|x)$$

- The Bayes or posterior risk (they are equivalent) depends on w, not θ.
- An alternative is the minimax (mM) approach: Minimize the maximum risk. That is find

$$\delta = \arg\min_{\delta} \max_{\theta} R(\theta, \delta) = \arg R_{mM}.$$

Or, maximize the minimum risk (maximin, Mm):

$$\delta = \arg\max_{w} \min_{\delta} R(w, \delta) = \arg R_{Mm}.$$

- Both are global criteria; the Mm risk is based on the Bayes estimator, δ_B.
- $R_{mM} \leq R_{Mm}$. The Game Theorem asserts they are equal.

Two Loss Functions

- Given a loss function we can, in principle, work out the Bayes estimator (and the mM, Mm, admissible etc).
- Suppose $L(\theta, \delta) = (\theta - \delta)^2$.

$$\arg\min_{\delta} \int w(\theta|x)(\theta - \delta(x))^2 d\theta = E(\Theta|x).$$

- Suppose $L(\theta, \delta) = |\theta - \delta|$.

$$\arg\min_{\delta} \int w(\theta|x)|\theta - \delta(x)| d\theta = \text{median}(w(\theta|x)).$$

- Other loss functions give percentiles (asymmetric absolute value) or $m(x)$ (relative entropy), LINEX loss etc etc

Generalized 0-1 Loss

- Bayes testing is the Bayes action under generlaized 0-1 loss, i.e., it has the smallest Bayes risk.
- Consider $\mathcal{H} : \theta \in \Omega_H$ vs. $\mathcal{K} : \theta \in \Omega_K$. Let a_0 (a_1) be the action that we choose $\mathcal{H}$ ($\mathcal{K}$). So, given a rule δ, the acceptance region for $\mathcal{H}$ is $\{x|\delta(x) = a_0\}$.
- So, define the loss

$$L(\theta, a) = \begin{cases} 0 & \theta \in \Omega_H, a = a_0 \\ 0 & \theta \in \Omega_K, a = a_1 \\ c_1 & \theta \in \Omega_H, a = a_1 \\ c_2 & \theta \in \Omega_K, a = a_0 \end{cases}$$

c_1 (c_2) is the cost of a Type I (II) error.

Risk of a Test

- Let $R = \{x|\ \delta(x) = a_1\}$ and let $\beta(\theta) = P_\theta(X \in R)$. Then

$$E(\theta, \delta) = \begin{cases} c_1\beta(\theta) & \theta \in \Omega_H \\ c_2(1 - \beta(\theta)) & \theta \in \Omega_K \end{cases}$$

  So, if $c_1 = c_2 = 1$ we have the usual power function of a Frequentist test.
- Theorem: Under generalized 0-1 loss, any test of the form 'reject $\mathcal{H} : \theta \in \Omega_H$ when $W(\Omega_K|x) < c_2/(c_1 + c_2)$' is Bayes optimal.
- That is, $\delta_B(x) = \chi(W(\Omega_K|x) < c_1/(c_1 + c_2))$ is the Bayes optimal test and achieves $\min_\delta R(w, \delta|x)$.

A Final Example

- Suppose $X \sim N(\theta, \sigma^2)$ and w is $N(\mu, \tau^2)$ with μ, σ and τ known. Let $\eta = \sigma^2/(\sigma^2 + \tau^2)$.
- $(\Theta|\bar{X})$ has distribution $N(E(\Theta|\bar{X}), \mathrm{Var}(\Theta|\bar{X})) = N((1-\eta)\bar{x} + \eta\mu, \tau^2\eta)$.
- Test $\mathcal{H} : \theta \geq \theta_0$ vs $\mathcal{K} : \theta < \theta_0$.
- $W(\mathcal{H}|x) = W(\Theta \geq \theta_0|x) = P(N(0,1) > \frac{\theta_0 - (1-\eta)\bar{x} - \eta\mu}{\tau\sqrt{\eta}}|\bar{x})$. Now

$$\begin{cases} W(\mathcal{H}|x) \leq c_2/(c_1 + c_2) = \alpha \in (0,1) \\ \iff (\theta_0 - (1-\eta)\bar{x} - \eta\mu)/\tau\sqrt{\eta} > z_\alpha \\ \iff \bar{x} < \theta_0 - \frac{\eta(\mu - \theta_0) - z_\alpha\tau\sqrt{\eta}}{1-\eta} \end{cases}$$

- Thus, rejecting $\mathcal{H}$ for small values of $\bar{x}$ is Bayes optimal where the threshold depends on c_1, c_2 and the prior.