STAT 211 NOTES

GREG TAM

Contents

1. January 28th, 2014

1.1. Big Questions.

(1) What is truth?
We would like to estimate θ. We want to do point estimation or interval estimation. A good example of what an interval estimate is
$$\mathbb{P}(\theta \in (3.1, 9.6)) = 0.95$$
We would also want to do hypothesis testing.

(2) What is good?
There are many different criteria such as MSE, unbiasedness, coverage, admissibility, etc.

(3) What should you condition on?
Frequentists: $f(y|\theta)$ or $f_\theta(y)$. Bayesians: $\pi(\theta|y)$.

Theorem (Bayes' Theorem)**.**
$$\underbrace{\pi(\theta|y)}_{posterior} \propto \underbrace{f(y|\theta)}_{likelihood}\ \underbrace{\pi(\theta)}_{prior}$$
where $f(y|\theta)$ is called the likelihood function if viewed as a function of θ. The "constant" is $f(y)$.

(4) What is a model?
A model is a family of distributions
$$\{f(y|\theta) : \theta \in \Theta\}$$
where Θ is a parameter space. Assume a dominating measure exists. There are two types of models. Parametric: θ is finite-dimensional. Non-parametric: θ is infinite-dimensional

1.2. Bridges Between Bayesian and Frequentist Thinking.

(1) Likelihood function
$$\mathcal{L}(\theta) = f(y|\theta)$$
Bayesians need this, but frequentists also use it, e.g., MLE.

(2) Decision Theory (Analysis)
Complete class theorem.

(3) Hierarchical Models (Multi-level Models)

2. January 30th, 2014

Likelihood Function
$$\mathcal{L}(\theta) = f(y|\theta)$$
which is for one observation. If we have n iid observations, then we have
$$\mathcal{L}(\theta) = \prod_{j=1}^{n} f(y_j|\theta)$$
We can also deal with the log-likelihood
$$l(\theta) = \sum_{j=1}^{n} \log f(y_j|\theta)$$
Two likelihood functions $\mathcal{L}_1, \mathcal{L}_2$ are considered equivalent if
$$\mathcal{L}_1(\theta) = c(y)\mathcal{L}_2(\theta)$$
for all θ.

Suppose we are flipping a coin with probability p of getting heads. We observe 9 heads and 3 tails. Then the likelihood function is
$$\mathcal{L}(p) = p^9(1-p)^3$$
if this is the given specific sequence of outcomes. However, what happens if this is Binomial? That is $\#H \sim \text{Bin}(12, p)$, so
$$\mathcal{L}(p) = \binom{12}{9} p^9(1-p)^3$$
Both of these give $\hat{p} = 0.75$. Ideally we want to likelihood function to look Normal. Note that this does not mean the distribution of the parameter is Normal. Why is the Normal so good? This means the log-likelihood is just a quadratic.
$$\mathcal{L}(\theta) = e^{-(y-\theta)^2/2}$$
$$l(\theta) = -(y-\theta)^2/2$$
Because of the duality of y and θ, $y \sim \mathcal{N}(\theta, 1)$, and we sort of have "$\theta \sim \mathcal{N}(y, 1)$".

How do we get the solution to the MLE easily without using calculus? If we consider NEFs, then we have

$$\mathcal{L}(\eta) = e^{y\eta - \psi(\eta)}$$
$$l(\eta) = y\eta - \psi(\eta)$$
$$l'(\eta) = y - \psi'(\eta)$$

Set this equal to 0, so

$$\psi'(\eta) = y$$

and solve for η. We know from STAT 210, that $\hat{\mu} = y$. Use invariance to get MLE of η.

Definition 1 (Invariance of MLE)**.** *Let g be one-to-one. If the MLE of θ is $\hat{\theta}$, then the MLE of $g(\theta)$ is $g(\hat{\theta})$. This is just a reparameterization. Let $\tau = g(\theta)$. then*

$$\mathcal{L}_1(\tau) = f(y|\tau) = f(y|\theta) = \mathcal{L}_2(\theta)$$

We simply set $y = n\hat{p}$, and we know $y = 9, n = 12$.

2.1. **Sampling Schemes.**

(1) Pre-determined that there would be $n = 12$ tosses. If X is the number of heads, then $X \sim \text{Bin}(12, p)$ and

$$\mathcal{L}(p) = \binom{12}{9} p^9 (1-p)^3$$

(2) Stopped after the 9th head. This is a form of Negative Binomial.

$$\mathcal{L}(p) = \binom{11}{8} p^9 (1-p)^3$$

(3) Stopped after the 3rd tail.

$$\mathcal{L}(p) = \binom{11}{2} p^9 (1-p)^3$$

2.2. **Likelihood Principle.** This is highly controversial. All the evidence that is contained in the data about the parameter is contained in the likelihood function. There's a "weak" likelihood principle and a "strong" likelihood principle. The weak one is within one model. The strong works even if models are different.

Example 1 (p-values)**.** *Say we test $H_0 : p = \frac{1}{2}$ vs $H_1 : p > \frac{1}{2}$ using the above sampling schemes.*

(1) $\mathbb{P}(X \geq 9) = 0.073$
(2) $\mathbb{P}(\#tails \leq 3) = 0.073$
(3) $\mathbb{P}(X \geq 9) = 0.033$

Thus p-values violate the likelihood principle.

Example 2 (Censoring Data)**.** *$f_\theta(t), F_\theta(t)$, which are our PDF and CDF, without censoring. Assume we are using the Weibull distribution. If we are at the contribution of the jth individual to the likelihood function.*

$$\mathcal{L}_j(\theta) = \begin{cases} f_\theta(t_j) & \text{if } t_j \text{ observed} \\ 1 - F_\theta(c) & \text{if time survival unobserved, censoring at time } c \end{cases}$$

3. February 4th, 2014 (Sufficiency)

Let Y_1, Y_2 be two random variables which are 1 if Joe/Tirthankar smokes. We have $Y_1 = 0, Y_2 = 0$. Is it possible to do inference on this? Here, $p = 0$. Suppose we are interested in the percentage of faculty members who smoke.

In this case, we have a finite population with N units and variables $Y_1, \ldots, Y_n \in \{0, 1\}$. We are interested in

$$p = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Definition 2 (Statistic)**.** *A function of data which will be used to infer about a parameter.*

One important question to ask is "is a statistic adequate?" This leads to the topic of sufficiency.

Definition 3 (Sufficient Statistic)**.** *A statistic is **sufficient** if it can extract all information about the parameter.*

Customary notation for a vector of data is to write

$$\vec{Y} = (Y_1, \ldots, Y_n)$$

The most trivial sufficient statistic is itself ($\vec{Y} = (Y_1, \ldots, Y_n)$).

Definition 4 (Sufficient Statistic (Formal Definition))**.** *Let $T(\vec{Y})$ be a statistic. $T(\vec{Y})$ is called sufficient for parameter θ if $f(\vec{y}|T)$ is independent of θ.*

Example 3. *Suppose* $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bern}(p)$. *We claim that* $T(\vec{Y}) = \sum_{i=1}^n Y_i$ *is sufficient for* p.

$$f(\vec{y}|T=t) = \frac{1}{\binom{n}{t}}$$

Example 4. *Suppose* $Y_1, Y_2 \overset{iid}{\sim} \text{Pois}(\lambda)$ *and* $T(\vec{Y}) = Y_1 + Y_2$. *By properties of the Poisson distribution, we know*

$$T(\vec{Y}) = Y_1 + Y_2 \sim \text{Pois}(2\lambda)$$

So

$$\begin{aligned}
\mathbb{P}\Big(Y_1 = y_1, Y_2 = y_2 | T(\vec{Y}) = y\Big) &= \frac{\mathbb{P}(Y_1 = y_1, Y_2 = y - y_1)}{\mathbb{P}\Big(T(\vec{Y}) = y\Big)} \\
&= \frac{e^{-\lambda}\frac{\lambda^{y_1}}{y_1!} e^{-\lambda}\frac{\lambda^{y-y_1}}{(y-y_1)!}}{e^{-2\lambda}\frac{(2\lambda)^y}{y!}} \\
&= \frac{y!}{y_1!(y-y_1)!}\left(\frac{1}{2}\right)^{y_1}\left(\frac{1}{2}\right)^{y_2}
\end{aligned}$$

Theorem 1 (Factorization Theorem for Sufficiency)**.** $T(\vec{Y})$ *is sufficient for* θ *if and only if*

$$f_\theta(\vec{y}) = g_\theta(T(\vec{y}))h(\vec{y}) \qquad (*)$$

(Assume that $\vec{Y}$ *is a discrete random variable)*

Proof. Prove the if part first. We show that for all t

$$\mathbb{P}\Big(\vec{Y} = \vec{y} | T(\vec{Y}) = t\Big)$$

is independent of θ.

$$\begin{aligned}
\mathbb{P}\Big(\vec{Y} = \vec{y} | T(\vec{Y}) = t\Big) &= \frac{\mathbb{P}\Big(\vec{Y} = \vec{y} \cap T(\vec{Y}) = t\Big)}{\mathbb{P}\Big(T(\vec{Y}) = t\Big)} \\
&= \frac{g_\theta(y)h(\vec{y})}{\sum\limits_{\vec{y}:T(\vec{y})=t} \mathbb{P}\Big(\vec{Y} = \vec{y}\Big)} \\
&= \frac{g_\theta(y)h(\vec{y})}{\sum\limits_{\vec{y}:T(\vec{y})=t} g_\theta(y)h(\vec{y})} \qquad \text{by } (*) \\
&= \frac{h(\vec{y})}{\sum\limits_{\vec{y}:T(\vec{y})=t} h(\vec{y})}
\end{aligned}$$

which is clearly independent of θ.

Now we prove the only if part. Assume $T(\vec{Y})$ is sufficient for θ.

$$\begin{aligned}
\mathbb{P}_\theta\Big(\vec{Y} = \vec{y}\Big) &= \sum_t \mathbb{P}_\theta\Big(\vec{Y} = \vec{y} \cap T(\vec{Y}) = t\Big) \\
&= \mathbb{P}_\theta\Big(\vec{Y} = \vec{y} \cap T(\vec{Y}) = T(\vec{y})\Big) \\
&= \underbrace{\mathbb{P}_\theta\Big(\vec{Y} = \vec{y} | T(\vec{Y}) = T(\vec{y})\Big)}_{h(\vec{y})} \underbrace{\mathbb{P}_\theta\Big(T(\vec{Y}) = T(\vec{y})\Big)}_{g_\theta(T(\vec{y}))}
\end{aligned}$$

□

(1) For NEFs, $f_\theta(y) \propto \exp(y\eta - \psi(\eta))$, then $\sum_{i=1}^n Y_i$ is sufficient for η.
(2) $f_\theta(y) \propto \exp\left(\sum_{i=1}^k T_i(\vec{y})\eta_i(\theta) - \psi(\eta)\right) h(y)$, then

$$\left(\sum_{j=1}^n T_1(Y_j), \ldots, \sum_{j=1}^n T_k(Y_j)\right)$$

are jointly sufficient for $(\eta_1(\theta), \ldots, \eta_k(\theta))$.

Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}(0, \theta)$

$$\begin{aligned} f_\theta(y_1, \ldots, y_n) &= \frac{1}{\theta^n} I\{Y_1 \in [0, \theta], \ldots Y_n \in [0, \theta]\} \\ &= \underbrace{\frac{1}{\theta^n} I\{Y_{(n)} \leq \theta\}}_{g_\theta(Y_{(n)})} \end{aligned}$$

so the maximum is sufficient for θ.

3.1. **Uniqueness.** Clearly sufficient statistics are not unique as all of the data is a sufficient statistic. We want a definition that defines "minimality".

Definition 5 (Minimal Sufficient Statistic)**.** *A statistic S is **minimal sufficient** if it can be expressed as a function of any other sufficient statistic.*

Theorem 2. *A sufficient statistic $T(\vec{Y})$ is minimal sufficient if it satisfies the following condition: If for any $\vec{x}, \vec{y}$ that makes $\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})}$ independent of θ, we must have $T(\vec{x}) = T(\vec{y})$, then $T(\vec{Y})$ is minimal sufficient.*

Think of the case where $\mathcal{N}(0, \sigma^2)$. Then $T(\vec{Y}) = \sum_{i=1}^n Y_i^2$ is minimal sufficient for σ^2.

Proof. Let S be a minimal sufficient statistic and let T satisfy this condition. Choose $\vec{x} \neq \vec{y}$ such that $S(\vec{x}) = S(\vec{y})$. By the factorization theorem,

$$\frac{f_\theta(\vec{x})}{f_\theta(\vec{y})} = \frac{g_\theta(S(\vec{x}))h(\vec{x})}{g_\theta(S(\vec{y}))h(\vec{y})} \text{ is independent of } \theta \Rightarrow T(\vec{x}) = T(\vec{y})$$

This implies that if S is minimal sufficient, then T is minimal sufficient. □

4. February 6th, 2014 (More Sufficiency)

4.1. **Bayesian Version of Sufficiency.** If

$$\mathbb{P}\Big(\theta|\vec{Y}\Big) = \mathbb{P}\Big(\theta|T(\vec{Y})\Big)$$

holds for any arbitrary prior $\pi(\theta)$. Compare this to the frequentist definition that requires $f(\vec{y}|T(\vec{y}))$ be independent of θ.

Question: Is the MLE of θ minimal sufficient? By the factorization theorem, we would want this to maximize the $g_\theta(T(\vec{Y}))$ as a function of θ, that is

$$\hat{\theta}_{MLE} = \arg\max_\theta \tilde{L}(\theta|T(\vec{Y})$$

so $\hat{\theta}_{MLE}$ should be a function of the minimal sufficient statistic. The dimension of the sufficient statistic exceeds the dimension of the parameter.

Suppose we have paired observations $(X_i, Y_i)_{i=1,\ldots,n}$ iid with pdf

$$f_\theta(x, y) = \exp\left(-\theta x - \frac{y}{\theta}\right)$$

This distribution is known as Fisher's gamma-hyperbola. It's clear that

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right)$$

is jointly sufficient for θ. If we calculate the MLE, we get

$$T = \sqrt{\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}}$$

But if we move this to two dimensions, we claim that the T and U together here are jointly minimal sufficient.

$$\left(T = \sqrt{\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}}, U = \sqrt{\sum_{i=1}^n X_i}\sqrt{\sum_{i=1}^n Y_i}\right)$$

U here is independent of θ. This is called an **ancillary statistic**.

Definition 6 (Ancillary Statistic $A(\vec{Y})$)**.** *"Does not yield information about θ". More formally, it means the distribution of $A(\vec{Y})$ is independent of θ.*

Example 5. *The constant c.*

Example 6. *Consider* $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$*, then*

$$A(\vec{Y}) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_{n-1}$$

which is not dependent on μ*.*

Example 7. *Suppose* $Y_1, \dots, Y_n \overset{iid}{\sim} \exp(\lambda)$*, then*

$$A(\vec{Y}) = \frac{Y_1}{Y_1 + \cdots + Y_n}$$

We can use reasoning by representation or divide the top and bottom by λ *and the resulting distribution will not depend on* λ*.*

Example 8 (Cox (1958))**.** *Suppose there are two measurements and two instruments. One is distributed as* $\mathcal{N}(\theta, \sigma_1^2)$ *and the other is distributed as* $\mathcal{N}(\theta, \sigma_2^2)$ *with* $\sigma_1^2 \ll \sigma_2^2$ *where these values are known. Generate a random variable* $U \sim \text{Bern}(\frac{1}{2})$ *to pick which distribution to sample from. Suppose* Y *takes the Normal distribution chosen after* U*.*

When determining confidence intervals, since U *is known, we would let the confidence interval be*

$$\begin{cases} Y \pm 1.96\sigma_1 & U = 0 \\ Y \pm 1.96\sigma_2 & U = 1 \end{cases}$$

(1) Ancillary statistics provide relevant subsets of the sample space.
(2) It gives the "right" measure of variation.
(3) Provide means of reducing dimension.

Suppose there are 100 families and two brands of shampoo

	Brand 1	Brand 2
1	X	
2	X	
⋮	X	
50	X	
51		X
⋮		X
100		X
	$\bar{Y}_1$	$\bar{Y}_2$

Suppose that afterwards, we have another covariate that becomes available, say athleticism

Shampoo \Athletic	1	2	
1	40	10	50
2	10	40	50

Then this gives us further information on the two groups.

4.2. **Completeness/Complete Sufficient Statistics.**

Definition 7 (Complete Sufficient Statistic)**.** *A statistic* $T(\vec{Y})$ *is said to be* ***complete*** *if one cannot construct a non-trivial unbiased estimator of zero from it. Then*

$$\mathbb{E}_\theta[g(T)] = 0 \Rightarrow g(T) = 0 \quad a.s., \quad \mathbb{P}(g(T) = 0) = 1$$

Kalbfleisch (1975,1982) classified ancillarity in to two categories, experimental and mathematical.

Theorem 3. *A complete sufficient statistic is also minimal sufficient. However, the converse is not true.*

Proof. Let T be a complete sufficient statistic and M be a minimal sufficient statistic. Consider

$$h(T) = \mathbb{E}[T|M] - T$$

Note that $\mathbb{E}[T|M]$ is a function of T by definition of minimal sufficiency. Why does the expectation not depend on θ?

$$\mathbb{E}[h(T)] = \mathbb{E}_M[\mathbb{E}[T|M]] - \mathbb{E}[T] = 0$$

□

Theorem 4 (Basu's Theorem)**.** *If* $T(\vec{Y})$ *is a complete sufficient statistic and* $A(\vec{Y})$ *is an ancillary statistic for a parameter* θ*, then*

$$T(\vec{Y}) \perp\!\!\!\perp A(\vec{Y})$$

Proof. Consider

$$h(T) = \mathbb{P}_\theta\Big(A(\vec{Y}) \in B | T(\vec{Y})\Big) - \mathbb{P}_\theta\Big(A(\vec{Y}) \in B\Big)$$
$$\mathbb{E}[h(T)] = 0$$

So

$$\mathbb{P}_\theta\Big(A(\vec{Y}) \in B | T(\vec{Y})\Big) = \mathbb{P}_\theta\Big(A(\vec{Y}) \in B\Big)$$

□

5. February 11th, 2014

A minimal sufficient statistic is unique up to a one-to-one transformation since minimal sufficient statistics are functions of any other sufficient statistic. So, if you have two, then they are functions of each other. Suppose we have $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown.

$$\left(\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2\right)$$

is a minimal sufficient statistic. Alternatively, we can use

$$\left(\bar{y}, \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2\right)$$

Identity:

$$\sum_{i=1}^n (y_i - c)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - c)^2$$

One thing we will get to later is

$$\text{MSE} = \text{Var} + \text{Bias}^2$$

Theorem 5 ("Beautiful Property")**.** *The likelihood function itself i.e. the entire curve is a minimal sufficient statistic.*

Proof of Sufficiency.

$$\mathcal{L}(\theta) = f_\theta(y)h(y)$$

Rearranging we have

$$f_\theta(y) = \mathcal{L}(\theta)\frac{1}{h(y)}$$

□

Proof of Minimality. Let T be sufficient.

$$f_\theta(y) = g_\theta(y)h(y)$$

We can obtain $\mathcal{L}(\theta) = g_\theta(t)$. □

Example 9. *Let $y_1, \ldots, y_n \overset{iid}{\sim} f$. This is nonparametric and f is a PDF.*

$$\mathcal{L}(f) = f(y_1)f(y_2)\cdots f(y_n)$$

The order statistics $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ are minimal sufficient. (These are also complete sufficient statistics).

5.1. **Exponential Families.** We have $y_1, \ldots, y_n \overset{iid}{\sim} f$ where f is from an exponential family.

$$f_\theta(y) = \exp\left(\sum_{i=1}^k \eta_i(\theta)T_i(y) - B(\theta)\right)h(y)$$

A minimal sufficient statistic for an exponential family is

$$\left(\sum_{i=1}^n T_1(y_i), \ldots, \sum_{i=1}^n T_k(y_i)\right)$$

We assume non-degeneracy: no linear relationship between T_i's and η_i's.

For exponential families, assume the parameter space has "full rank", i.e. contains a k-dimensional non-empty open set. Then the natural sufficient statistic is equal to the complete sufficient statistic. It is also equal to the minimal sufficient statistic.

Assume $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\ &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} y_i - \frac{n\mu^2}{2\sigma^2}\right)\end{aligned}$$

So our complete sufficient statistic is

$$\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2\right)$$

5.2. **MLE in Exponential Families.** Set the expected value of the natural sufficient statistic equal to the observed value, then solve for the parameter.

$$\mathbb{E}_\theta\left[\vec{T}(y)\right] = \vec{T}(y_{obs})$$

then we solve for θ. The solution is unique if it exists, since the log-likelihood function is strictly concave.

For the normal case, we can use the identity

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$$

so it is immediately obvious that we take $\hat{\mu} = \bar{y}$. Since

$$NSS = CSS = \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2\right)$$

then

$$n\hat{\mu} = \mathbb{E}\left[\sum_{i=1}^{n} y_i\right] = \sum_{i=1}^{n} y_i \Rightarrow \hat{\mu} = \bar{y}$$

Similarly,

$$n(\hat{\sigma}^2 + \hat{\mu}^2) = \mathbb{E}\left[\sum_{i=1}^{n} y_i^2\right] = \sum_{i=1}^{n} y_i^2 \Rightarrow \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

which gives us

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Completeness depends on both the statistic and the model.

Example 10. *Let $y_1, \ldots, y_m \overset{iid}{\sim} \text{Bin}(2, p)$. Then $\bar{y}$ is a complete sufficient statistic. Simplify this by letting $m = 1$, so $y \sim \text{Bin}(2, p)$. Here, y is a complete sufficient statistic for parameter space $(0, 1)$. Now let the parameter space be smaller, i.e. $\{\frac{1}{4}, \frac{3}{4}\}$. Is y complete for this model? Let h be such that*

$$h(0) = 3, \quad h(1) = 5, \quad h(2) = 3$$

Then

$$\mathbb{E}_p[h(y)] = h(0)(1-p)^2 + h(1)(2p(1-p)) + h(2)p^2 = 0$$

for all p in the parameter space ($\{\frac{1}{4}, \frac{3}{4}\}$). Hence, y is not complete here.

Theorem 6 (Rao-Blackwell Theorem)**.** *We can improve an estimator by conditioning on a sufficient statistic. For MSE: We are using $\hat{\theta}$ to estimate θ. Let T be a sufficient statistic. Let $\hat{\hat{\theta}} = \mathbb{E}_\theta\left[\hat{\theta}|T\right]$. It is important that T is sufficient so we can actually compute the expectation. Then Rao-Blackwell says*

$$\mathbb{E}\left[(\hat{\hat{\theta}} - \theta)^2\right] \leq \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Example 11. *Let $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. The complete sufficient statistic is $\bar{y}$. What if we just used $\hat{\theta} = y_1$. Let $\hat{\hat{\theta}} = \mathbb{E}[y_1|\bar{y}]$*

If $y_1, \ldots, y_n$ are iid then

$$\mathbb{E}\left[y_1 \middle| \sum_{i=1}^{n} y_i\right] = \mathbb{E}\left[y_2 \middle| \sum_{i=1}^{n} y_i\right] = \cdots \mathbb{E}\left[y_n \middle| \sum_{i=1}^{n} y_i\right]$$

Adding these, we get

$$n\mathbb{E}\left[y_1 \middle| \sum_{i=1}^{n} y_i\right] = \sum_{i=1}^{n} y_i$$

So $\hat{\hat{\theta}} = \mathbb{E}[y_1|\bar{y}] = \bar{y}$

Example 12. *Say* $\hat{\theta} = w_1 y_1 + \cdots + w_n y_n$, $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$. *Then*

$$\mathbb{E}\Big[\hat{\theta}|\bar{y}\Big] = w_1\bar{y} + \cdots w_n\bar{y} = \bar{y}$$

6. February 13th, 2014

6.1. **Basu's Theorem Examples.** If $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\bar{y} \perp\!\!\!\perp S^2 = \frac{1}{n-1}\sum_{j=1}^n (y_j - \bar{y})^2$$

Ways to prove this:

(1) Jacobians
(2) Orthogonal transformations, matrix theory
(3) $\bar{y} \perp\!\!\!\perp (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$. We can prove this via Multivariate Normal properties, MGFs, or conditioning arguments.
(4) Basu. If σ^2 is known, $\bar{y}$ is a complete sufficient statistic, $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ is ancillary, so they are independent. As a generalization, this is true for location families.

How do we show $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ is ancillary?

$$y_j = \mu + \sigma Z_j$$

where $Z_j \sim \mathcal{N}(0, 1)$. Then

$$(y_1 - \bar{y}, \ldots, y_n - \bar{y}) = \sigma(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$$

and so the distribution does not depend on μ.

Suppose $X \sim \text{Gamma}(a), Y \sim \text{Gamma}(b)$ and are independent. Let's show that

$$\frac{X}{X+Y} \perp\!\!\!\perp X + Y$$

This is true by Basu's theorem. Introduce a scale parameter $\frac{1}{\lambda}$. Now we have an NEF: $X + Y$ is a complete sufficient statistic for the model ($X \sim \lambda^{-1}\text{Gamma}(a) = \Gamma(a, \lambda)$ and $Y \sim \lambda^{-1}\text{Gamma}(b)$). We have

$$X = \lambda^{-1}\tilde{X}$$

where $\tilde{X} \sim \text{Gamma}(a)$ and $Y = \lambda^{-1}\tilde{Y}$. Hence

$$\frac{\lambda^{-1}\tilde{X}}{\lambda^{-1}\tilde{X} + \lambda^{-1}\tilde{Y}}$$

S^2 is unbiased for σ^2 and

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

$(\bar{y}, S^2)$ is a complete sufficient statistic.

$$\frac{1}{n-1}\sum_{j=1}^n (y_j - \bar{y})^2$$

is the UMVUE (since it is an unbiased function of a complete sufficient statistic.

$$\frac{1}{n}\sum_{j=1}^n (y_j - \bar{y})^2$$

is the MLE. Say we have

$$c\sum_{j=1}^n (y_j - \bar{y})^2$$

Then what is the "best" c?

$$\text{MSE} = \text{Var} + \text{Bias}^2$$

So we can have a trade off between Bias and Variance. The "best" c is in terms of the MSE

$$\frac{1}{n+1}\sum_{j=1}^n (y_j - \bar{y})^2$$

is the best.

6.2. **Unbiased Estimator.**

Definition 8 (UMVUE)**.** *$\hat{\theta}_1$ is UMVUE for θ if $\hat{\theta}_1$ is unbiased and $\mathrm{Var}_\theta(\hat{\theta}_1) \leq \mathrm{Var}_\theta(\hat{\theta}_2)$ for any unbiased $\hat{\theta}_2$ and is unique (if it exists).*

Proof. Let $\hat{\theta}_1, \hat{\theta}_2$ be UMVUEs. Then

$$\mathrm{Var}_\theta(\hat{\theta}_1) = \mathrm{Var}_\theta(\hat{\theta}_2) = V$$

Consider

$$\hat{\theta}_3 = \frac{\hat{\theta}_1}{2} + \frac{\hat{\theta}_2}{2}$$

which is also unbiased by linearity. Let us calculate its variance

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_3) &= \frac{1}{4}V + \frac{1}{4}V + \frac{1}{2}V\,\mathrm{Corr}(\hat{\theta}_1, \hat{\theta}_2) \\ &\leq V\end{aligned}$$

since the correlation must be less than 1 in absolute value. Equality occurs iff $\hat{\theta}_1$ and $\hat{\theta}_2$ are perfectly correlated,

$$\hat{\theta}_1 = a\hat{\theta}_2 + b \Rightarrow a = 1, b = 0$$

□

Theorem 7. *If T is a complete sufficient statistic and $\hat{\theta}$ is unbiased and is a function of T, then $\hat{\theta}$ is the UMVUE.*

Proof. Consider $\hat{\hat{\theta}} = \mathbb{E}[\theta|T]$. Is this necessarily the best everywhere? What if we had $\hat{\theta}_2$ which is better for other values of θ. If we apply Rao-Blackwell to $\hat{\theta}_2$ as well, that is

$$\hat{\hat{\theta}}_2 = \mathbb{E}\left[\hat{\theta}_2|T\right]$$

By completeness

$$\hat{\hat{\theta}} = \hat{\hat{\theta}}_2$$

is a function of T and has mean 0, so they are the same at the UMVUE. □

Example 13.

(1) $y \sim \mathrm{Bin}(n, p)$, estimate $\theta = \mathrm{logit}(p) = \log\frac{p}{1-p}$. Find an unbiased estimator $g(y)$. We then require

$$\log\frac{p}{1-p} = \mathbb{E}_p[g(y)] = \sum_{k=0}^{n} g(k)\binom{n}{k}p^k(1-p)^k$$

for all $p \in (0,1)$. Unfortunately this is impossible to solve since the RHS is a polynomial.

We know the MLE of p is $\hat{p} = \frac{y}{n}$. By invariance of the MLE we know the MLE of $\mathrm{logit}(p)$ is $\mathrm{logit}(\hat{p})$. However, this presents a lot of problems. What if $y = 0$ or $y = n$? Then the MLE does not even exist. Methods to deal with this include ignoring these extreme cases or using regularization (Bayesian technique).

(2) $y \sim \mathrm{Pois}(\lambda)$, estimate $\theta = e^{-2\lambda}$.

$$\hat{\theta} = (-1)^y$$

is unbiased and is a function of the complete sufficient statistic, so it is UMVUE. This is the only unbiased estimator, however we see that it is completely ridiculous. The MLE of this is e^{-2y}.

(3) $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \mathrm{Pois}(\lambda)$ with $n \geq 2$ and we want to estimate $e^{-\lambda} = \mathbb{P}(y_j = 0)$. We know (since it's an NEF) that the complete sufficient statistic is $\bar{y}$.

$$\hat{\theta}_1 = I(y_1 = 0)$$

is unbiased.

$$y_1|y_1 + \cdots + y_n \sim \mathrm{Bin}\left(\sum_{i=1}^{n} y_i = n\bar{y}, \frac{1}{n}\right)$$

So the UMVUE is

$$\mathbb{E}[I(y_1 = 0)|\bar{y}] = \left(1 - \frac{1}{n}\right)^{n\bar{y}}$$

The MLE is $e^{-\bar{y}}$ again by invariance.

6.3. **Fisher Weighting.** Let $\hat{\theta}_1 = \hat{\theta}_2$ be independent unbiased estimators of θ. The best unbiased

$$\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$$

where w_1 and w_2 are weights, that is $w_1 + w_2 = 1, w_j \geq 0$.

$$w_j \propto \frac{1}{\text{Var}(\hat{\theta}_j)}$$

One easy way to see why this is true is to look at sample means

$$\hat{\theta}_1 = \frac{y_1 + \cdots + y_m}{m}$$
$$\hat{\theta}_2 = \frac{y_{m+1} + \cdots + y_{m+n}}{n}$$

The weights are also proportional to Fisher Information.

7. February 18th, 2014 (Score Function and the Cramér-Rao Lower Bound)

We have already established some measures of "goodness" such as

- Unbiased
- Variance
- Consistency ($T_n(\vec{Y}) \xrightarrow{P} \theta$)

7.1. **Score Function.**

$$S(Y, \theta) = \frac{\partial}{\partial\theta} l(Y, \theta)$$

where $l(\theta, Y) = \log \mathcal{L}(\theta, Y)$. The score function has an **additive** property

$$S(\vec{Y}, \theta) = \frac{\partial}{\partial\theta} l(\vec{Y}, \theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} l(Y_i, \theta) = \sum_{i=1}^{n} S(Y_i, \theta)$$

Properties:
Under regularity conditions,

$$\mathbb{E}_\theta[S(Y, \theta)] = 0$$

The regularity conditions are that we can differentiate under the integral sign.

$$\begin{aligned}
\mathbb{E}_\theta[S(Y, \theta)] &= \int S(y, \theta) f_\theta(y)\, \mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta} l(y, \theta) f_\theta(y)\, \mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta} \{\log f_\theta(y)\} f_\theta(y)\, \mathrm{d}y \\
&= \int \frac{1}{f_\theta(y)} \frac{\partial}{\partial\theta} \{f_\theta(y)\} f_\theta(y)\, \mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta} f_\theta(y)\, \mathrm{d}y \\
&= \frac{\partial}{\partial\theta} \int f_\theta(y)\, \mathrm{d}y \\
&= 0
\end{aligned}$$

We have the variance is given by

$$\text{Var}_\theta(S(Y, \theta)) = \mathbb{E}_\theta\left[S(Y, \theta)^2\right]$$

Definition 9 (Fisher Information)**.**

$$\begin{aligned}
i(\theta, Y) &= -\frac{\partial^2}{\partial\theta^2} l(\theta, Y) \\
\mathcal{I}_1(\theta) &= \mathbb{E}_\theta[i(\theta, Y)] \qquad \textit{Expected Fisher information} \\
\mathcal{I}_n(\theta) &= n\mathcal{I}_1(\theta)
\end{aligned}$$

$i(\hat{\theta}, \vec{y})$ is called the observed Fisher information.

Example 14. *Consider* $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Pois}(\lambda)$. *Is* $\bar{Y}$ *the UMVUE?*

$$\begin{aligned}
\mathcal{L}(\lambda, \vec{Y}) &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \\
l(\lambda, \vec{Y}) &= -n\lambda + \sum_{i=1}^n y_i \log\lambda \\
S(\lambda, \vec{Y}) &= \frac{\partial}{\partial\lambda} l(\lambda, \vec{Y}) \\
&= -n + \frac{\sum_{i=1}^n y_i}{\lambda} \\
-\frac{\partial^2}{\partial\lambda^2} l(\lambda, \vec{Y}) &= \frac{\sum_{i=1}^n y_i}{\lambda^2}
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathcal{I}_n(\lambda) &= \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda} \\
i_n(\hat{\lambda}, \vec{y}) &= \frac{n\bar{y}}{\hat{\lambda}^2} = \frac{n}{\hat{\lambda}}
\end{aligned}$$

By the law of large numbers, we can actually show

$$\frac{\hat{i}}{n} = \frac{i(\hat{\theta}, \vec{y})}{n} \xrightarrow{P} \mathcal{I}_1(\theta)$$

7.2. **Variance of unbiased estimators.** Given θ,

- Parametric function $g(\theta)$
- Statistic $T(\vec{Y})$ is unbiased for $g(\theta)$
- What is the variance of $T(\vec{Y})$?

Theorem 8 (Cramér-Rao Theorem)**.** *This result gives a lower bound for the variance of unbiased estimators. Let* $T(\vec{Y})$ *be an unbiased estimator for a parametric function* $g(\theta)$*. Under regularity conditions,*

$$\text{Var}_\theta(T(\vec{Y})) \geq \frac{\left\{\frac{\partial}{\partial\theta} g(\theta)\right\}^2}{n\mathcal{I}_1(\theta)}$$

where $\mathcal{I}_1(\theta)$ *is the information for one observation.*

Proof. Consider $S(\vec{Y}, \theta)$ and $T(\vec{Y})$. By Cauchy-Schwarz

$$\text{Cov}_\theta(S(\vec{Y}, \theta), T(\vec{Y})) \leq \sqrt{\text{Var}_\theta(S(\vec{Y}, \theta))\,\text{Var}_\theta(T(\vec{Y}))}$$

Since $\mathbb{E}_\theta\Big[S(\vec{Y}, \theta)\Big] = 0$,

$$\text{Var}_\theta(S(\vec{Y}, \theta)) = \mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$$

so this becomes the denominator. Now we need to show that $\mathrm{Cov}_\theta(S(\vec{Y},\theta),T(\vec{Y})) = \frac{\partial}{\partial\theta}g(\theta)$.

$$\begin{aligned}
\mathrm{Cov}_\theta(S(\vec{Y},\theta),T(\vec{Y})) &= \int\cdots\int S(\vec{y},\theta)T(\vec{y})f_\theta(\vec{y})\,\mathrm{d}\vec{y} \\
&= \int\cdots\int T(\vec{y})\left\{\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_\theta(y_i)\right\} f_\theta(y_1)\cdots f_\theta(y_n)\,\mathrm{d}\vec{y} \\
&= \int\cdots\int T(\vec{y})\left\{\sum_{i=1}^n \frac{1}{f_\theta(y_i)}\frac{\partial}{\partial\theta}f_\theta(y_i)\right\} f_\theta(y_1)\cdots f_\theta(y_n)\,\mathrm{d}\vec{y} \\
&= \int\cdots\int T(\vec{y})\left\{\sum_{i=1}^n \frac{\partial}{\partial\theta}f_\theta(y_i)\prod_{j\neq i} f_\theta(y_j)\right\}\,\mathrm{d}\vec{y} \qquad \text{by iid} \\
&= \int\cdots\int T(\vec{y})\frac{\partial}{\partial\theta}\left\{\prod_{i=1}^n f_\theta(y_i)\right\}\,\mathrm{d}\vec{y} \qquad \text{by DUThIS} \\
&= \int\cdots\int T(\vec{y})\frac{\partial}{\partial\theta}f_\theta(\vec{y})\,\mathrm{d}\vec{y} \\
&= \frac{\partial}{\partial\theta}\int\cdots\int T(\vec{y})f_\theta(\vec{y})\,\mathrm{d}\vec{y} \\
&= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[T(\vec{Y})\right] \\
&= \frac{\partial}{\partial\theta}g(\theta)
\end{aligned}$$

□

Example 15. $Y_1,\ldots,Y_n \overset{iid}{\sim} \mathrm{Pois}(\lambda)$. $\mathcal{I}_n(\lambda) = \frac{n}{\lambda}$. *Any unbiased estimator* $T(\vec{Y})$ *of* λ *must satisfy*

$$\mathrm{Var}(T(\vec{Y})) \geq \frac{\lambda}{n}$$

$\mathrm{Var}(\bar{Y}) = \frac{\lambda}{n}$ *so this achieves the lower bound.*

Suppose $g(\lambda) = \lambda^2$. *Then* $T(\vec{Y}) = \bar{Y}^2 - \bar{Y}$ *and* $\mathbb{E}_\lambda\left[T(\vec{Y})\right] = \lambda^2 + \lambda - \lambda = \lambda^2$.

$$\mathbb{E}\left[\bar{Y}^2\right] = \mathrm{Var}(\bar{Y}) + \left(\mathbb{E}\left[\bar{Y}\right]\right)^2 = \lambda + \lambda^2$$

Example 16. *Suppose* $Y_1,\ldots,Y_n \overset{iid}{\sim} \mathcal{U}[0,\theta]$. *Then we can argue* $Y_{(n)}$ *is sufficient. We also know that* $Y_{(n)}$ *is the MLE since*

$$f_\theta(\vec{y}) = \frac{1}{\theta^n}I(Y_{(n)} \leq \theta)$$

Can we find the UMVUE of θ*?. We can show using calculus that*

$$\mathbb{E}\left[Y_{(n)}\right] = \frac{n}{n+1}\theta \Rightarrow \mathbb{E}\left[\frac{n+1}{n}Y_{(n)}\right] = \theta$$

8. February 20th, 2014

$-\frac{\partial^2}{\partial\theta^2}l(\theta,\vec{Y})$ is a matrix where the diagonals have the form $-\frac{\partial^2}{\partial\theta_i^2}l(\theta,\vec{Y})$ and the off diagonals are $-\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\theta,\vec{Y})$.

Now let

$$y_i = \beta_0 + \beta_1 x_i + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma^2)}$$

Here, x_i is non-stochastic and we have $(x_i,y_i)_{i=1,2,\ldots,n}$. We wish to minimize

$$\min_{\beta_0,\beta_1}\sum_{i=1}^n(y_i - \beta_0 - \beta_!x_i)^2$$

This gives us a matrix

$$\begin{pmatrix} \frac{\partial^2 l}{\partial\beta_0^2} & \frac{\partial^2 l}{\partial\beta_0\partial\beta_1} \\ \frac{\partial^2 l}{\partial\beta_0\partial\beta_1} & \frac{\partial^2 l}{\partial\beta_1^2} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

Suppose we have an experiment where we can pick 10 values of x_i in the interval $[0,1]$. What is the best choice of these 10 values? We actually choose 5 each at the extremes so that

$$\mathrm{Var}(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^n(x_i - \bar{x})^2}$$

8.1. **MLE and its Asymptotic Behaviour.**

8.1.1. *Consistency of the MLE.* Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} f_\theta(y)$ and θ_0 is the true value, then consistency means

$$\hat{\theta}_n \xrightarrow{P} \theta_0$$

A crucial assumption we must make is that the density is identifiable. Suppose $\theta_1 \neq \theta_2$. Then there must exist one y such that $f_{\theta_1}(y) \neq f_{\theta_2}(y)$. We also need that the support of the density does not depend on θ.

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} l(\theta, \vec{Y}) = \frac{1}{n} \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} l(\theta, Y_i)$$

We have

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{l(\theta, Y_i)}_{Z_i} \xrightarrow{P} \mathbb{E}_{\theta_0}[l(\theta, Y)]$$

by the weak law of large numbers. Here, $\hat{\theta}_n$ maximizes the left-hand side and θ_0 maximizes the right-hand side.

How do we show

$$\mathbb{E}_{\theta_0}[l(\theta, Y)] \leq \mathbb{E}_{\theta_0}[l(\theta_0, Y)] \qquad \forall \theta \in \Theta$$

If we consider the difference, we get

$$\begin{aligned}
\mathbb{E}_{\theta_0}[l(\theta, Y)] - \mathbb{E}_{\theta_0}[l(\theta_0, Y)] &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(y)}{f_{\theta_0}(y)}\right)\right] \\
&\leq \log \mathbb{E}_{\theta_0}\left[\frac{f_\theta(y)}{f_{\theta_0}(y)}\right] \qquad \text{by Jensen's inequality} \\
&= \log \int \frac{f_\theta(y)}{f_{\theta_0}(y)} f_{\theta_0}(y)\, \mathrm{d}y \\
&= \log \int f_\theta(y)\, \mathrm{d}y \\
&= \log 1 \\
&= 0
\end{aligned}$$

Uniform Convergence in Probability

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} l(\theta, Y_i) - \mathbb{E}_{\theta_0}[l(\theta, Y)] \right| \xrightarrow{P} 0$$

8.2. **Asymptotic distribution of $\hat{\theta}_n$.**

Conditions:

1. True value is θ_0.
2. Identifiability
3. Support not dependent on θ
4. For each $\theta \in \Theta$, $\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta} \log f_\theta(y)\right] = 0$
5. $$\mathcal{I}_1(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial \theta^2} \log f_\theta(y)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta} \log f_\theta(y)\right)^2\right] < \infty$$
6. For each $\theta \in \Theta$, the first three partial derivatives of $\log f_\theta(y)$ w.r.t. θ exist for $y \in \mathrm{supp}(f_\theta(y))$.
7. $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Think of the MLE as a solution of $S(\hat{\theta}_n, \vec{Y}) = 0$. Then

$$\begin{aligned}
S(\hat{\theta}_n, \vec{Y}) &= S(\theta_0, \vec{Y}) + (\hat{\theta}_n - \theta_0) S'(\theta_0, \vec{Y}) + \frac{(\hat{\theta}_n - \theta_0)^2}{2} S''(\theta^*, \vec{Y}) \\
&= S(\theta_0, \vec{Y}) + (\hat{\theta}_n - \theta_0) \left\{ S'(\theta_0, \vec{Y}) + \frac{1}{2} S''(\theta^*, \vec{Y})(\hat{\theta}_n - \theta_0) \right\}
\end{aligned}$$

where $\theta_0 \leq \theta^* \leq \hat{\theta}_n$. Rearranging, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} S(\theta_0, \vec{Y})}{-\frac{1}{n} S'(\theta_0, \vec{Y}) - \frac{1}{2n} S''(\theta^*)(\hat{\theta}_n - \theta)}$$

The numerator is

$$\sqrt{n}\left[\frac{1}{n} S(\theta_0, \vec{Y})\right] = \sqrt{n}\left[\frac{1}{n} \sum_{i=1}^{n} S(\theta_0, Y_i)\right] \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}_1(\theta_0)\right)$$

We also have

$$\frac{1}{n}S'(\theta_0,\vec{Y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}l(\theta,Y_i)\bigg|_{\theta=\theta_0} \xrightarrow{P} \mathbb{E}_{\theta_0}\left[-\frac{\partial}{\partial\theta^2}l(\theta,Y)\right] = \mathcal{I}_1(\theta_0)$$

Define $R_n = \frac{1}{2n}S''(\theta^*)(\hat{\theta}_n - \theta_0)$. If we can argue $|R_n| \xrightarrow{P} 0$, then by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$$

We now impose another regularity condition. For each $\theta_0 \in \Theta$, $\exists$ a function $g(y)$ such that for all θ in the neighbourhood of θ_0,

$$\left|\frac{\partial^3 \log f_\theta(y)}{\partial\theta^3}\right| \leq g(y) \qquad \forall y$$

and $\mathbb{E}[g(Y)] < \infty$.

8.3. **Computation of MLE.** One method is using Newton-Raphson, where we look for x such that $f(x) = 0$. We iterate, by doing

$$f(x^{(n+1)}) \approx f(x^{(n)}) + (x^{(n+1)} - x^{(n)})f'(x^{(n)})$$

which implies

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}$$

If $Y_1,\ldots,Y_n \overset{iid}{\sim}$ Cauchy(θ),

$$f_\theta(y) = \frac{1}{\pi\{1+(y-\theta)^2\}}$$

Then we have

$$\begin{aligned} l(\theta,\vec{Y}) &= -\sum_{i=1}^{n}\log\{1+(y_i-\theta)^2\} \\ S(\theta,\vec{Y}) &= 2\sum_{i=1}^{n}\frac{Y_i-\theta}{1+(Y_i-\theta)^2} \\ &= 2\sum_{i=1}^{n}W_i(Y_i-\theta) \qquad \text{where } W_i = \frac{1}{1+(Y_i-\theta)^2} \end{aligned}$$

and $S'(\theta,\vec{Y}) = 0$ which implies

$$\hat{\theta} = \frac{\sum_{i=1}^{n}W_iY_i}{\sum_{i=1}^{n}W_i} = \frac{\sum_{i=1}^{n}W_i(Y_i,\theta_i)Y_i}{\sum_{i=1}^{n}W_i(Y_i,\theta_i)}$$

Solving this gives

$$\hat{\theta}^{n+1} = \hat{\theta}^{(n)} - \frac{S(\vec{Y},\hat{\theta}^{(n)})}{S'(\vec{Y},\hat{\theta}^{(n)})} = \hat{\theta}^{(n)} + \frac{\sum_{i=1}^{n}W_i(Y_i,\theta)}{\sum_{i=1}^{n}(2W_i^2 - W_i)}$$

8.4. **Global Optimization Algorithms.**

- Simulated Annealing
- Genetic Annealing

9. February 25th, 2014

9.1. **Neyman-Scott Problem (1948).** Suppose Y_{ij} are independent, distributed as $\mathcal{N}(\mu_i,\sigma^2)$ where $i = 1,2,\ldots,n$, $j = 1,2,\ldots,k$. Assume the μ_i's are not of interest, but σ^2 is. We can also write this as

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

for $i = 1,\ldots,n$ and $j = 1,\ldots,k$ where $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$.

$$\mathbb{P}\big(y_{ij}\big|\vec{\mu},\sigma^2\big) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y_{ij}-\mu_i)^2}$$

$$\log\mathcal{L}(\vec{\mu},\sigma^2|\vec{y}) = -\frac{nk}{2}\log 2\pi - \frac{nk}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\mu_i)^2$$

$$\frac{\partial l(\vec{\mu},\sigma^2|\vec{y})}{\partial\sigma^2} = -\frac{nk}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\mu_i)^2$$

Equating to zero yields

$$\hat{\sigma}^2 = \frac{1}{nk}\sum_{i}^{n}\sum_{j}^{k}(y_{ij} - \hat{\mu}_i)^2$$

We also have

$$\frac{\partial l(\vec{\mu}, \sigma^2)}{\partial \mu_i} = 0 \Rightarrow \hat{\mu}_i = \frac{1}{k}\sum_{j=1}^{k} y_{ij} = \bar{y}_i \qquad i = 1, \ldots, n$$

This implies that

$$\hat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_i)^2$$

Do we have that $\hat{\sigma}^2$ is a consistent estimator of σ^2?. Since we have that for all i

$$\sum_{j=1}^{k}\underbrace{\frac{(y_{ij} - \bar{y}_i)^2}{\sigma}}_{W_i} \sim \chi^2_{k-1}$$

So

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sigma^2}{kn}\sum_{i=1}^{n} W_i \\ &\xrightarrow{P} \frac{\sigma^2}{k}(k-1)\end{aligned}$$

where the W_i's are iid with $\mathbb{E}[W_i] = k - 1$.

How do we make this consistent? One simple trick is just to adjust the bias. Let

$$\begin{aligned}\hat{\hat{\sigma}}^2 &= \frac{k}{k-1}\hat{\sigma}^2 \\ &= \frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_i)^2\end{aligned}$$

We have a property that $T_n \xrightarrow{P} \theta$ and sufficient $\mathbb{E}[T_n] \longrightarrow \theta$ and $\mathrm{Var}(T_n) \longrightarrow 0$. Then

$$\hat{\hat{\sigma}}^2 \xrightarrow{P} \sigma^2$$

If we look at $k = 2$, then $\hat{\sigma}^2 \xrightarrow{P} \frac{\sigma^2}{2}$. For $i = 1, 2, \ldots, n$, define

$$Z_i - Y_{i1} - Y_{i2} \overset{iid}{\sim} \mathcal{N}\left(0, 2\sigma^2\right)$$

The MLE of $2\sigma^2$ is $\frac{1}{n}\sum_{i=1}^{n} Z_i^2$. By invariance,

$$\hat{\sigma}^2_{MLE} = \frac{1}{2n}\sum_{i=1}^{n} Z_i^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{Z_i^2}{2\sigma^2} \xrightarrow{P} \sigma^2$$

Say $\theta = (\theta_1, \theta_2)$. If the nuisance parameter is the vector θ_1, then θ_2 is a nuisance parameter.

9.2. **Strategies for handling nuisance parameters.** Suppose we know $\mathcal{L}(\theta_1, \theta_2|y)$.

- Method 1: (Plug-in method)
  (1) Find
  $$\hat{\theta}_2 = \arg\max_{\theta_1, \theta_2 \in \Theta} \mathcal{L}(\theta_1, \theta_2|y)$$
  (2) $\mathcal{L}_{est}(\theta_1|y) = \mathcal{L}(\theta_1, \hat{\theta}_2|y)$
- Profile likelihood method
  (1) Find
  $$\hat{\theta}_2(\theta_1) = \arg\max_{\theta_2} \mathcal{L}(\theta_1, \theta_2|y)$$
  (2) Find
  $$\begin{aligned}\hat{\theta}_1 &= \arg\max_{\theta_1} \mathcal{L}_{prof}(\theta_1|y) \\ &= \arg\max_{\theta_1} \mathcal{L}(\theta_1, \hat{\theta}_2(\theta_1)|y)\end{aligned}$$

Note: Consider $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$\log \mathcal{L}(\mu, \sigma^2|y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\mathcal{L}_{est}(\mu|y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\hat{\sigma}^2_{MLE} - \frac{1}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\mathcal{L}_{prof}(\mu|y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\hat{\sigma}}^2(\mu) - \frac{1}{2\hat{\hat{\sigma}}^2(\mu)}\sum_{i=1}^{n}(y_i - \mu)^2$$

since

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\hat{\hat{\sigma}}^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

If $\vec{\theta} = (\theta_1, \theta_2)$, then

$$\mathbb{E}\left[-\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}\right] \overset{\hat{\theta}_1}{=\!=} \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}$$

Then

$$\mathrm{Var}(\hat{\theta}_1) = I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$$

Do we have that $I_{11} - I_{12}I_{22}^{-1}I_{21} = I_{11}^1 \geq 0$. Under plug-in, $(\hat{\theta}_1) = I_{11}^{-1}$.

$$\sqrt{n}(\hat{\theta}_1 - \theta_1^*) \xrightarrow{D} \mathcal{N}(0, I^{11}(\theta_1^*))$$

9.3. **Bayesian Approach.**

(1) Integrate out the nuisance parameters.
(2) Assume $\pi(\theta_1, \theta_2)$ to be a prior distribution.
(3)

$$\pi(\theta_1, \theta_2|y) \propto \mathcal{L}(\theta_1, \theta_2|y)\pi(\theta_1, \theta_2)$$

$$\pi(\theta_1|y) \propto \int \mathcal{L}(\theta_1, \theta_2|y)\pi(\theta_1, \theta_2)\,\mathrm{d}\theta_2$$

Sometimes we use the terms estimator and estimate interchangeably. However, an estimator is a statistic $T(\vec{Y})$ and its realized value, $T(\vec{y})$ is the estimate.

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big(L(\vec{Y}) \leq \theta \leq U(\vec{Y})\Big) \geq 1 - \alpha$$

$$\mathcal{L}(\theta, \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(\vec{y} - \vec{\mu})'\Sigma^{-1}(\vec{y} - \vec{\mu})\right\}$$

Then we have

$$\hat{\mu} = (\mathbf{1}_n' R^{-1}\mathbf{1}_n)^{-1}(\mathbf{1}_n' R^{-1} y)$$

10. February 27th, 2014 (Interval Estimation)

$[L(\vec{Y}), U(\vec{Y})]$ is a random interval. What we seek the find is the coverage probability, that is

$$\mathbb{P}_\theta\Big(L(\vec{Y}) \leq \theta \leq U(\vec{Y}\Big)$$

We hope that this coverage probability is at least greater than some $1 - \alpha$. If we do this over all θ, that is find

$$\inf_{\theta} \mathbb{P}_\theta\Big(L(\vec{Y}) \leq \theta \leq U(\vec{Y})\Big) \geq 1 - \alpha$$

Then this is the confidence interval with a confidence coefficient $1 - \alpha$.

Example 17. *Let $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$, then a sufficient statistic is $T = Y_{(n)}$. We can build confidence intervals here. They can be of two possible forms.*

(i) $[aY_{(n)}, bY_{(n)}]$, $1 \leq a \leq b$
(ii) $[Y_{(n)} + c, Y_{(n)} + d]$, $0 \leq c \leq d$.

Method of finding confidence intervals:

(i) In the first case, the coverage probability is

$$\mathbb{P}_\theta\big(aY_{(n)} \le \theta \le bY_{(n)}\big) = \mathbb{P}_\theta\left(\frac{1}{b} \le \frac{Y_{(n)}}{\theta} \le \frac{1}{a}\right)$$

We know the pdf of $T = \frac{Y_{(n)}}{\theta}$*, which is*

$$f_T(t) = nt^{n-1}, t \ge 0$$

so

$$\mathbb{P}_\theta\big(aY_{(n)} \le \theta \le bY_{(n)}\big) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$$

(ii) In the second case, we have the coverage probability is

$$\begin{aligned}\mathbb{P}_\theta\big(Y_{(n)} + c \le \theta \le Y_{(n)} + d\big) &= \mathbb{P}_\theta\big(c - \theta \le -Y_{(n)} \le d - \theta\big)\\ &= \mathbb{P}_\theta\left(1 - \frac{d}{\theta} \le \frac{Y_{(n)}}{\theta} \le 1 - \frac{c}{\theta}\right)\\ &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n\end{aligned}$$

but this confidence interval goes to 0 as $\theta \to \infty$*, so this is bad confidence interval. This leads us to the notion of a* ***pivot*** *or* ***pivotal quantity****.*

10.1. **Pivots/Pivotal Quantities.**

Definition 10 (Pivot)**.** *A pivot is a function* $Q(\vec{Y}, \theta)$*, which explicitly depends on* $\vec{Y}$ *and* θ*, whose distribution is independent of* θ*.*

Example 18.

(1) Suppose $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *with* σ^2 *both unknown. Then*

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

(2) Location Family, $f(y - \theta)$*. Then* $\bar{Y} - \theta$ *is a pivot.*

$$(Y_1, \dots, Y_n) \sim (Z_1 + \theta, \dots, Z_n + \theta)$$

(3) Scale Family, $\frac{1}{\theta} f\left(\frac{y}{\theta}\right)$*. Then* $\frac{\bar{Y}}{\theta}$ *is a pivot.*
(4) Location-Scale Family. $\frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$*, then* $\frac{\bar{Y}-\mu}{s}$ *is a pivot.*

10.1.1. *Finding a pivot?* There is no strategy that always works to find a pivot. Suppose T is a statistic whose pdf can be expressed in the form

$$f_\theta(t) = g(Q(t, \theta)) \left|\frac{\partial}{\partial \theta} Q(t, \theta)\right|$$

where $Q(t, \theta)$ is a monotonic function of θ given t. Then we have $Q(T, \theta)$ is a pivotal quantity.

10.1.2. *Using a pivot to construct confidence intervals.* Fix $a < \alpha < 1$. Find a and b not depending on θ such that

$$\mathbb{P}_\theta\Big(A \le Q(\vec{Y}, \theta) \le b\Big) \ge 1 - \alpha$$

Define

$$C(\vec{Y}) = \{\theta : a \le Q(\vec{Y}, \theta) \le b\}$$

Example 19. *Suppose* $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *with* σ^2 *unknown. We wish to find a confidence interval for* μ*.*

(1) Let $T = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$*.*
(2) Then we have

$$\mathbb{P}\left(-t_{n-1,\alpha/2} \le \frac{\bar{Y} - \mu}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right) = 1 - \alpha$$

which is equivalent to

$$\mathbb{P}\left(\bar{Y} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \le \mu \le \bar{Y} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}\right) = 1 - \alpha$$

Example 20. *Suppose $Y_1, \ldots, Y_n$ which are iid with density*

$$f_\mu(y) = e^{-(y-\mu)}, \quad y \geq \mu$$

What is the $(1-\alpha)$ interval for μ?

$$\begin{aligned} f_\mu(\vec{y}) &= e^{-\sum_i (y_i - \mu)} \mathbb{1}\{y_{(1)} \geq \mu\} \\ &= \underbrace{e^{n\mu} \mathbb{1}\{Y_{(1)} \geq \mu\}}_{g(Y_{(1)}, \mu)} \underbrace{e^{-\sum_i y_i}}_{h(\vec{y})} \end{aligned}$$

Let $T = Y_{(1)}$. Then

$$f_T(t) = ne^{-n(t-\mu)}, \quad t \geq \mu$$

If we now let $U = n(T-\mu)$, then its density is given by

$$f_U(u) = ne^{-nu}, \quad u \geq 0$$

Then we have that

$$\begin{aligned} \mathbb{P}(U \leq a) &= \frac{\alpha}{2} & \mathbb{P}(U \geq b) &= \frac{\alpha}{2} \\ \int_0^a ne^{-nu}\,\mathrm{d}u &= \frac{\alpha}{2} & \int_b^\infty ne^{-nu}\,\mathrm{d}u &= \frac{\alpha}{2} \end{aligned}$$

So we have

$$Y_{(1)} + \frac{1}{n}\log(\alpha/2) \leq \mu \leq Y_{(1)} + \frac{1}{n}\log(1-\alpha/2)$$

10.1.3. *A more general strategy - Inverting CDFs of a statistic.*

Theorem 9. *Let T be a statistic with continuous cdf $F_\theta(t)$. Let $\alpha_1 + \alpha_2 = \alpha$, $0 < \alpha < 1$ be fixed. For each t in the support of F, define $\theta_L(t)$ and $\theta_U(t)$.*

(1) If $F_\theta(t)$ is monotonically decreasing in θ, then $F_{\theta_L}(t) = 1-\alpha_2$ and $F_{\theta_U}(t) = \alpha_1$.
(2) If $F_\theta(t)$ is monotonically increasing in θ, then $F_{\theta_L}(t) = \alpha_1$ and $F_{\theta_U}(t) = 1-\alpha_2$.

Then we have $[\theta_L(T), \theta_U(T)]$ is a $(1-\alpha)$ confidence interval.

Let T be a discrete statistic with CDF $F_\theta(t) = \mathbb{P}_\theta(T \leq t)$. For t in the support of F, define $\theta_L(t), \theta_U(t)$ such that

(1) If $F_\theta(t) \downarrow \theta$, $\mathbb{P}_{\theta_U(t)}(T \leq t) = \alpha_1$, $\mathbb{P}_{\theta_L(t)}(T \geq t) = \alpha_2$
(2) If $F_\theta(t) \uparrow \theta$, $\mathbb{P}_{\theta_U(t)}(T \geq t) = \alpha_1$, $\mathbb{P}_{\theta_L(t)}(T \leq t) = \alpha_2$

Example 21. *Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Pois}(\lambda)$. Then*

$$T = \sum_{i=1}^n Y_i \sim \mathrm{Pois}(n\lambda)$$

Suppose you observe $T = t_0$. Then

$$\mathbb{P}_{n\lambda_U}(T \leq t_0) = \frac{\alpha}{2} \qquad \mathbb{P}_{n\lambda_L}(T \geq t_0) = \frac{\alpha}{2}$$

If $Y \sim \mathrm{Pois}(\alpha)$

$$\begin{aligned} \mathbb{P}_\lambda(Y \geq \alpha) &= \mathbb{P}(X \leq \lambda) & X &\sim \mathrm{Gamma}(\alpha, 1) \\ \mathbb{P}_\lambda(Y \leq \alpha) &= \mathbb{P}(X \geq \lambda) & X &\sim \mathrm{Gamma}(\alpha+1, 1) \end{aligned}$$

This implies

$$\mathbb{P}(X \geq n\lambda_U) = \frac{\alpha}{2}, \quad X \sim \mathrm{Gamma}(t_0 + 1, 1)$$

11. March 4th, 2014

11.1. **Steps of Bayesian Data Analysis.**

(1) Set up a full probability model, that is we model both observable (data) and unobservable (parameters, latent variables) quantities jointly for all quantities of interest in the problem.
(2) Condition on the observed data. (Compute the posterior distribution).
(3) Evaluate fit of the model. Check whether conclusions make sense. Then do a sensitivity analysis (robustness).

One very controversial topic is choosing the prior. If nothing is known about the prior, then perhaps a Uniform is suitable, but then this would imply that the prior squared is not Uniform. However, not knowing the prior would also imply not knowing anything about the prior squared and this starts to make less sense. One other problem we have is evaluating the fit of the model. How would we know what is a "good" fit?

One way is to do posterior predictive checks, where we look at

$$y'|y$$

where y' is future data and y is the observed data. We can think of this as a Markov chain.

$$y \to \theta \to y'$$

11.2. **Nuisance Parameters.** Suppose we have (θ_1, θ_2) where θ_1 is the parameter of interest and θ_2 is a nuisance parameter. In the frequentist approach, there isn't one best approach. In the Bayesian approach, we can simply integrate it out.

$$\begin{aligned}\pi(\theta_1|y) &= \int \pi(\theta_1, \theta_2|y)\,\mathrm{d}\theta_2 \\ &\propto \int \mathcal{L}(\theta_1, \theta_2)\pi(\theta_1)\pi(\theta_2|\theta_1)\,\mathrm{d}\theta_2 \\ &= \pi(\theta_1) \underbrace{\int \mathcal{L}(\theta_1, \theta_2)\pi(\theta_2|\theta_1)\,\mathrm{d}\theta_2}_{\text{integrated likelihood}}\end{aligned}$$

The first line is already correct. We may decide to extend on it. One common mistake is that if θ_1, θ_2 are independent a priori, that does not necessarily mean that they are not independent when conditioning on y (in the posterior). In the last line, $\pi(\theta_2|\theta_1) = \pi(\theta_2)$ if $\theta_1 \perp\!\!\!\perp \theta_2$ a priori. The big question is when do we maximize, integrate, or average?

11.3. **Laplace Approximation.** Let g be a positive function. Suppose we want to approximate

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} e^{h(x)}\,\mathrm{d}x$$

where $h(x) = \log g(x)$. Let it be such that $h(x)$ is strictly concave and $h'(x_0) = 0$. Then we have

$$\begin{aligned}\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x &\approx \int_{-\infty}^{\infty} e^{h(x_0)+\frac{1}{2}h''(x_0)(x-x_0)^2}\,\mathrm{d}x \qquad \text{the linear term disappears} \\ &= e^{h(x_0)} \int_{-\infty}^{\infty} e^{\frac{1}{2}h''(x_0)(x-x_0)^2}\,\mathrm{d}x \\ &= e^{h(x_0)} \frac{\sqrt{2\pi}}{\sqrt{-h''(x_0)}} \underbrace{\int_{-\infty}^{\infty} \frac{\sqrt{-h''(x_0)}}{\sqrt{2\pi}} e^{\frac{1}{2}h''(x_0)(x-x_0)^2}\,\mathrm{d}x}_{\mathcal{N}\left(x_0, -\frac{1}{h''(x_0)}\right)} \\ &= e^{h(x_0)} \frac{\sqrt{2\pi}}{\sqrt{-h''(x_0)}}\end{aligned}$$

This is in fact one of the easiest ways to derive Stirling's formula.

Suppose we can simulate from $\pi(\theta_1, \theta_2|y)$ e.g. using MCMC, but we want $\pi(\theta_1|y)$. We can simply do this by ignoring the θ_2. This effectively marginalizes it.

Suppose we use the definition that

$$\mathbb{P}_\theta(L(y) \le \theta \le U(y)) = 0.95$$

If we have a twenty-sided die and if we get 1 through 19, our interval is the real line, but if it equals 20, then the interval is the empty set, then this confidence interval just does not make sense.

If instead, we look at this from the Bayesian perspective we have

$$\mathbb{P}(L(y) \le \theta \le U(y)|y) = 0.95$$

which is called a probability interval, this will never give you the empty set, since the empty set has probability 0.

A key thing to note about these two values is that

$$\mathbb{E}[\mathbb{P}(L(y) \le \theta \le U(y)|\theta)] = \mathbb{E}[\mathbb{P}(L(y) \le \theta \le U(y)|y)] = \mathbb{P}(L(y) \le \theta \le U(y))$$

by Adam's law.

11.4. **Conjugacy.** NEF case:

$$\mathcal{L}(\eta) = e^{r(\eta y - \psi(\eta))}$$

where r is the sample size.
The idea of conjugacy is to mimic the likelihood function:

$$g(\eta)\,\mathrm{d}\eta = k(r_0, \mu_0) e^{r_0(\eta\mu_0 - \psi(\eta))}$$

12. March 6th, 2014 (Conjugacy)

We have that

$$g(\eta)\,\mathrm{d}\eta = k(r_0, \mu_0) e^{r_0(\eta\mu_0 - \psi(\eta))}$$

is the conjugate prior on η. The posterior on η is proportional to

$$e^{(r+r_0)\left(\eta\left(\frac{r}{r+r_0}y + \frac{r_0}{r+r_0}\mu_0\right) - \psi(\eta)\right)}$$

This is same, where we replaced r_0 by $r + r_0$ and μ_0 by $\frac{r}{r+r_0}y + \frac{r_0}{r+r_0}\mu_0$. r_0 and μ_0 here are the prior quantities and $r + r_0$ and $\frac{r}{r+r_0}y + \frac{r_0}{r+r_0}\mu_0$ are the posterior quantities. Note that the second term is a linear combination (in fact a weighted average) of y and the prior mean μ_0.

What about mean μ?

$$g(\mu)\,\mathrm{d}\mu = g(\eta)\,\mathrm{d}\eta$$

Recall that

$$\mu = \psi'(\eta)$$
$$V(\mu) = \psi''(\eta)$$

Be careful since $V(\mu)$ is the variance of the data as a function of μ, not the variance of μ, as it could be in a Bayesian setting. Therefore, ψ' is differentiable and strictly increasing.

Then

$$g(\mu)\,\mathrm{d}\mu \propto e^{r_0(\eta\mu_0 - \psi(\eta))}\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\mathrm{d}\mu$$

We know

$$\frac{\mathrm{d}\eta}{\mathrm{d}\mu} = \frac{1}{V(\mu)}$$

so

$$\begin{aligned}\eta &= \int \mathrm{d}\eta \\ &= \int \frac{\mathrm{d}\eta}{\mathrm{d}\mu}\,\mathrm{d}\mu \\ &= \int \frac{1}{V(\mu)}\,\mathrm{d}\mu\end{aligned}$$

Example 22. *Suppose $y \sim \frac{1}{r}\mathrm{Bin}(r, p)$. Then the likelihood function is given by*

$$\begin{aligned}\mathcal{L}(\eta) &= p^{ry}q^{r(1-y)} \\ &= e^{ry\log p + r(1-y)\log q)} \\ &= e^{r(y\,\mathrm{logit}(p) + \log q}\end{aligned}$$

The conjugate prior for $\eta = \mathrm{logit}(p)$ is

$$e^{r_0(\mu_0\,\mathrm{logit}(p) + \log(q))}$$

To get the conjugate prior for p, multiply by $\frac{1}{V(p)} = \frac{1}{p(1-p)}$. This gives us that

$$p \sim \mathrm{Beta}\left(r_0\mu_0, r_0(1-\mu_0)\right)$$

The prior mean is

$$\frac{r_0\mu_0}{r_0\mu_0 + r_0(1-\mu_0)} = \mu_0$$

The prior sample size is

$$r_0\mu_0 + r_0(1-\mu_0) = r_0$$

12.0.1. *Stein's Identity of Normal.* If $Z \sim \mathcal{N}(0,1)$ then for all functions h,

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)]$$

provided both sides exist. The proof of this is done by integration by parts.

For NEF-QVFs,

$$\mathbb{E}[(\mu - \mu_0)h(\mu)] = \frac{1}{r_0}\mathbb{E}[h'(\mu)V(\mu)]$$

if $y|\mu \sim$ NEF-QVF $\left[\mu, \frac{V(\mu)}{r}\right]$. Then $\mu \sim$ a conjugate distribution with μ_0, r_0.

Let $h(\mu) = 1$. Then immediately

$$\mathbb{E}[\mu] = \mathbb{E}[\mu_0]$$

To get the variance, let $h(\mu) = \mu - \mu_0$. Then

$$\begin{aligned}\operatorname{Var}(\mu) &= \mathbb{E}\left[(\mu - \mu_0)^2\right] \\ &= \frac{1}{r_0}\mathbb{E}[V(\mu)]\end{aligned}$$

Using the fact that $V(\mu) = V_2\mu^2 + V_1\mu + V_0$, we have

$$\operatorname{Var}(\mu) = \frac{1}{r_0}\left(V(\mu_0) + V_2 \operatorname{Var}(\mu)\right)$$

which implies

$$\operatorname{Var}(\mu) = \frac{V(\mu_0)}{r_0 - V_2}$$

for NEF-QVFs. To get the posterior mean, we do the same thing, but replacing r_0 with $r + r_0$ and μ_0 with $\frac{r}{r+r_0}y + \frac{r_0}{r+r_0}\mu_0$, so that

$$\operatorname{Var}(\mu|y) = \frac{V\left(\frac{r}{r+r_0}y + \frac{r_0}{r+r_0}\mu_0\right)}{r + r_0 - V_2}$$

12.0.2. *Normal-Normal.* Let $y|\mu \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$.

We look for the posterior distribution $\mu|y$. It is not surprising that this would be normal since the marginal (μ) and conditional $(y|\mu)$ distributions are Normal, so (y, μ) is Multivariate Normal.

We have that

$$\mathbb{E}[\mu|y] = \frac{1/\sigma^2}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}y + \frac{\frac{1}{\tau_0^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\tau_0^2}}\mu_0$$

and

$$\operatorname{Var}(\mu|y) = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

So we have

$$\mu|y \sim \mathcal{N}\left(\frac{1/\sigma^2}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}y + \frac{\frac{1}{\tau_0^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\tau_0^2}}\mu_0, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}\right)$$

12.1. **Jeffreys Prior (Reference Prior).**

$$\pi_{Jeff}(\theta) \propto \sqrt{|\mathcal{I}(\theta)|}$$

for one-dimensions. In more than one-dimensions, we have $|\mathcal{I}(\theta)|$ is the determinant of the Fisher information matrix.

Motivation: the motivation was not to come up with something not informative, but rather something that was invariant. If $\tau = g(\theta)$ where g is differentiable and strictly increasing, so we would get

$$\pi_{Jeff}(\theta)\,\mathrm{d}\theta = \pi_{Jeff}(\tau)\,\mathrm{d}\tau$$

This is true for $\pi_{Jeff}(\theta) \propto \sqrt{\mathcal{I}(\theta)}$.

Example 23. *If $y \sim \operatorname{Bin}(r,p)$, then $\pi_{Jeff}(p)$ is* Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$. *One thing nice here is that the Jeffrey's prior is proper. It may be that it could be improper.*

Example 24. *If $y \sim \operatorname{NB}(r,p)$, then $\pi_{Jeff}(p)$ is* Beta $\left(0, \frac{1}{2}\right)$, *which is an improper prior. This is also different from the above example. This is strange since before, we conduct the experiment, the experiment we actually do will affect which prior we choose.*

Example 25 (Location Family)**.** *In a location family, $\mathcal{I}(\theta) = c$, a constant. Then $\pi_{Jeff}(\theta)$ is flat.*

Example 26 (Scale Family)**.** *Suppose $y = \theta Z$. Then $\log y = \log\theta + \log Z$. This says to use $\mathrm{d}(\log\theta) = \frac{1}{\theta}\,\mathrm{d}\theta$ for θ.*

13. March 20th, 2014

13.1. **Statistical Inference for Linear Models.** General Linear Model: We would like to predict a response y_{new} given $X_1, \ldots, X_p$ predictors and n samples $(y_1, \ldots, y_n)'$. We have

$$y = X\beta + \varepsilon$$

where

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \qquad \varepsilon \sim [0, \sigma^2 I_n]$$

We wish to infer the values of $\beta = (\beta_1, \ldots, \beta_p)$ and σ^2. Also, we normally assume that $x_{i1} = 1$ for the intercept term. To fit, we can use ordinary least square (OLS). These are obtained by minimizing the sum of the squared errors.

$$\hat{\beta} = \arg\min_{\beta} (y - X\beta')(y - X\beta) = \arg\min_{\beta} \sum (y_i - X\beta_i)$$

The result is

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

This is in fact unbiased since

$$\begin{aligned} \mathbb{E}\Big[\hat{\beta}_{OLS}\Big] &= (X'X)^{-1}X'\mathbb{E}[y] \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

The variance can similarly be computed using the property $\operatorname{Var}(a'X) = a' \operatorname{Var}(X) a$.

$$\begin{aligned} \operatorname{Var}(\hat{\beta}_{OLS}) &= \operatorname{Var}((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X' \operatorname{Var}(y) X (X'X) - 1 \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

This is the best linear unbiased estimator (BLUE). First notice that the above is a linear transformation of y. How do we use this to make a prediction of y_{new} using the OLS?

$$\begin{aligned} \hat{y}_{new} &= X\hat{\beta}_{OLS} \\ &= \underbrace{X(X'X)^{-1}X'}_{=H} y \end{aligned}$$

where H is the hat matrix. H has some nice properties:

(1) H is symmetric
(2) H is idempotent ($H^2 = H$)
(3) H is positive semi-definite ($y'Hy \geq 0$)
(4) Eigenvalues are 0 or 1 and the rank is the number of 1's and is equal to that of X.

Therefore H is a projection matrix. The OLS solution projects the data vector onto the linear space spanned by the column space of X. The residual error is

$$\begin{aligned} e &= y - \hat{y} \\ &= y - Hy \\ &= (I - H)y \\ &= Py \end{aligned}$$

So we have $H \perp I - H$ therefore $P'H = 0$.

Theorem 10 (Gauss-Markov)**.** *Consider a scalar parameter $\theta = c'\beta$. Then $\hat{\theta}_{OLS} = c'\hat{\beta}_{OLS}$ is the best linear unbiased estimator (BLUE) of θ.*

Proof. Consider any unbiased linear estimator $\theta^* = a'y$, which means that $\mathbb{E}[a'y] = \theta = c'\beta$. This tells us that $a'X\beta = c'\beta$ for all β and so we have that $a'X = c'$

$$\begin{aligned} \operatorname{Var}(\theta^*) &= \operatorname{Var}(a'y) \\ &= \sigma^2 a'a \end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{OLS}) &= \mathrm{Var}(c'\hat{\beta}_{OLS}) \\ &= \mathrm{Var}(c'(X'X)^{-1}y) \\ &= \sigma^2 c'(X'X)^{-1}c \\ &= \sigma^2 a'X(X'X)^{-1}X'a \\ &= \sigma^2 a'Ha\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Var}(\theta^*) - \mathrm{Var}(\hat{\beta}_{OLS}) &= \sigma^2 a(I-H)a \\ &= \sigma^2 a'Pa \\ &\geq 0\end{aligned}$$

□

What is an unbiased estimator of σ^2? Since

$$\begin{aligned}\mathbb{E}[e'e] &= \mathbb{E}[y'P'Pe] \\ &= \mathbb{E}[y'Py] \\ &= \mathbb{E}[\mathrm{tr}(Pyy')] \\ &= \mathrm{tr}(P\mathbb{E}[yy']) \\ &= (n-p)\sigma^2\end{aligned}$$

so that $\frac{e'e}{n-p}$ is an unbiased estimator of σ^2. Let us place some distributional assumptions. Assume $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then

$$\mathcal{L}(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}|\sigma^2 I|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$$

Then

$$\begin{aligned}\hat{\beta}_{MLE} &= \hat{\beta}_{OLS} \\ &= (X'X)^{-1}X'y\end{aligned}$$

Notice that

$$\hat{\sigma}^2_{MLE} = \frac{e'e}{n}$$

is biased.

13.2. **Inference for LM.** We have that
(1) $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X'X)^{-1})$
(2) $\frac{1}{\sigma^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \sim \chi^2_p$
(3) $\frac{e'e}{\sigma^2} \sim \chi^2_{n-p}$
(4) $\hat{\beta}$ and e are independent.
We can get the following pivot

$$\frac{\frac{1}{\sigma^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/p}{\frac{e'e}{\sigma^2}/(n-p)} \sim \frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/p}{\mathrm{MSE}} \sim F_{p,n-p}$$

The numerator is the sum of squares due to regression and $e'e$ is the sum of squares due to residuals SSE/SSE and $MSE = \frac{e'e}{\sigma^2}$. We can then obtain a confidence region

$$C(\beta) = \left\{\beta : \frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/p}{\mathrm{MSE}} \leq F_{p,n-p,1-\alpha}\right\}$$

Proof.
(1) is trivial
(2) Use $Y \sim \mathcal{N}_m(\mu, \Sigma)$ then

$$(Y - \mu)'\Sigma^{-1}(Y - \mu) \sim \chi^2_m$$

We can use the spectral decomposition theorem to write

$$\Sigma = T\Lambda Y' \iff T'\Sigma T = \Lambda \Rightarrow \Sigma^{-1} = T\Lambda^{-1}T'$$

where T is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$. Then let $Z = T'(Y - \mu)$ and we have that $\mathbb{E}[Z] = 0$ and $\text{Var}(Z) = T'\Sigma T = \Lambda$. Then

$$(Y-\mu)'\Sigma^{-1}(Y-\mu) = Z'\Lambda^{-1}Z = \sum_{i=1}^{m} \frac{Z_i^2}{\lambda_i} = \chi^2_m$$

(3) This is the same.

$$\frac{e'e}{\sigma^2} = \frac{y'P'Py}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right)$$

so we have that if $Y \sim \mathcal{N}(0, I_m)$ and B is a symmetric idempotent matrix of rank $q < m$ then

$$Y'BY \sim \chi^2_q$$

(4) We need to show that $\text{Cov}(\hat{\beta}, e) = 0$. We have

$$\begin{aligned}
\text{Cov}(\hat{\beta}, e) &= \text{Cov}((X'X)^{-1}X'y, Py) \\
&= (X'X)^{-1}\,\text{Var}(y)P' \\
&= \sigma^2(X'X)^{-1}X'P' \\
&= 0
\end{aligned}$$

as $PX = 0$.

□

13.3. **LM without independence.** We have $y = X\beta + \varepsilon$ where $\varepsilon \sim (0, \sigma^2\Omega)$. Is $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ still the BLUE of β? Not reallym we need to get rid of $\Omega = T\Lambda T'$. Now let us define

$$\bar{y} = \Lambda^{-1/2}T'y = \underbrace{\Lambda^{-1/2}T'X}_{=\bar{X}}\beta + \underbrace{\Lambda^{-1/2}T'\varepsilon}_{=\varepsilon}$$

Then

$$\begin{aligned}
\text{Var}(\varepsilon) &= \sigma^2\Lambda^{-1/2}T'\Omega T\Lambda^{-1/2} \\
&= \sigma^2 I
\end{aligned}$$

So now we are back in the Gauss-Markov scenario. Therefore the BLUE is

$$\begin{aligned}
\hat{\beta} &= (\bar{X}'\bar{X})^{-1}\bar{X}'\bar{y} \\
&= (X'\Omega^{-1}X)^{-1}(X'\Omega y) \\
&= \hat{\beta}_{GLS}
\end{aligned}$$

which is the generalized least square (GLS) estimate. How do we do this in practice?

(1) Fit the OLS of $\hat{\beta}_{OLS}$ and obtain the error term e.
(2) Plot the residuals.

Then for example we would consider an AR(1)

$$\varepsilon_t = \psi\varepsilon_{t-1} + z_t$$

where $z_t \overset{iid}{\sim} [0, \sigma^2]$. Then

$$\Omega = \hat{\sigma}^2 \begin{pmatrix} 1 & \psi & \psi^2 & \cdots & \psi^{n-1} \\ \psi & & & & \\ \psi^2 & & & & \\ & & & & 1 \end{pmatrix}$$

We can then estimate everything recursively

$$\phi : \phi \to \phi^{(1)}$$

and

$$\begin{aligned}
\Omega^{(1)} &= \hat{\sigma}^2\Omega(\psi^{(1)}) \\
\hat{\beta}_{GLS} &= (X{\Omega^{(1)}}^{-1}X)^{-1}(X'{\Omega^{(1)}}^{-1}y)
\end{aligned}$$

13.4. **Side note on MLE for truncation.** Say we've been told that our data set is such that all values of y_i such that $y_i < 3$ have been omitted, then

$$f(t|T > 3) = \frac{f(t)}{\mathbb{P}(T > 3)}$$

14. March 25th, 2014 (Hypothesis Testing)

14.1. **Hypothesis Testing.** Suppose we have a hypothesis test of the form

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta = \theta_1 \quad \theta_1 > \theta_0$$

This is called a "point-null" hypothesis as it will fully specify the distribution. Compare this to the " point alternative". What is a "test"? Typically, we have a rejection rule, based on the data, which tells us whether we should reject or not. We observe data $\vec{Y} = (Y_1, \ldots, Y_n)$. We partition Ω into two disjoint regions, which are our rejection and acceptance regions. The rejection region is where $\vec{Y} \in R$ and and acceptance region is where $\vec{Y} \notin R$. Therefore, we essentially reject H_0 if $\vec{Y} \in R$ and do not reject otherwise.

Type I error	Rejecting H_0 when H_0 is true
Type II error	Not rejecting H_0 when H_0 is not true

In tabular form, we have

Action \ Reality	H_0 True	H_0 False
Reject H_0	Type I Error	✓
Do not Reject H_0	✓	Type II Error

We wish to choose a rejection region R such that the type I error is such that

$$\mathbb{P}\Big(\vec{Y} \in R \Big| H_0\Big) \leq \alpha$$

where α refers to the level of significance. We also wish to find a test which is the test that is most powerful among all size α.

(1)

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

Most Powerful Test: Rejection rule R is most powerful among all size α-tests if

$$\mathbb{P}\Big(\vec{Y} \in R \Big| H_1\Big) \geq \mathbb{P}\Big(\vec{Y} \in \tilde{R} \Big| H_1\Big)$$

where $\tilde{R}$ satisfies size-α condition, i.e. $\mathbb{P}\Big(\vec{Y} \in \tilde{R} \Big| H_0\Big) = \alpha$.

(2)

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

Uniformly Most Powerful Tests: A test R that is most powerful for every $\theta_1 \in \Theta_1$ is uniformly most powerful (UMP).

Lemma (Neyman-Pearson Lemma)**.** *Suppose we have data $\vec{Y}$. Assume $f_\theta(\vec{y})$ is the density of $\vec{Y}$, where we have*

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

Then define rejection rule R.

$$R = \left\{ \vec{Y} : \frac{f_{\theta_1}(\vec{y})}{f_{\theta_0}(\vec{y})} \geq k \right\}$$

where k is such that the size-α condition is satisfied, then R is most powerful (MP) for testing H_0 vs H_1.

Example 27. *Suppose we have $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, 1)$, where we have*

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

By the Neyman-Pearson lemma, we reject H_0 if

$$\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i - \mu_i)^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i - \mu_0)^2\right)} \geq k \Longrightarrow \exp\left\{ n\bar{y}(\mu_1 - \mu_0) - \frac{n}{2}(\mu_1^2 - \mu_0^2) \right\} \geq k \Longrightarrow \bar{y} \geq k^*$$

that is we can set our rejection region to be of the form $[k^*, \infty)$. To find k^*, we have

$$\mathbb{P}(\bar{y} \geq k^* | \mu = \mu_0) = \alpha \Rightarrow \mathbb{P}\big(\sqrt{n}(\bar{y} - \mu_0) \geq \sqrt{n}(k^* - \mu_0)\big) = \alpha$$

which tells us

$$\sqrt{n}(k^* - \mu_0) = z_\alpha$$

Rearranging this gives

$$k^* = \mu_0 + \frac{1}{\sqrt{n}} z_\alpha$$

However, this does not work for the alternative $H_1 : \mu < \mu_0$. Hence if we have a hypothesis test of the form $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$. and by the Neyman-Pearson lemma, we cannot find a UMP test.

Suppose we have two rejection regions R and $\tilde{R}$ which have size α, then by definition, we have

$$\int_R f_{\theta_0}(y)\,\mathrm{d}y = \int_{\tilde{R}} f_{\theta_0}(y)\,\mathrm{d}y = \alpha$$

We wish to show taht

$$\int_R f_{\theta_1}(y)\,\mathrm{d}y = \int_{\tilde{R}} f_{\theta_1}(y)\,\mathrm{d}y$$

Note that we have

$$\begin{aligned}\text{LHS} &= \int_{R\cap\tilde{R}^c} f_{\theta_1}(y)\,\mathrm{d}y + \int_{R\cap\tilde{R}} f_{\theta_1}(y)\,\mathrm{d}y \\ &= \int_{R\cap\tilde{R}^c} \frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} f_{\theta_0}\,\mathrm{d}y\end{aligned}$$

Suppose we have

$$H_0 : \theta = \theta_0 \text{ vs.} H_1 : \theta \in R$$

Then we can use the likelihood ratio statistic

$$\Lambda(\vec{Y}) = \frac{\sup_\theta \mathcal{L}(\theta)}{\mathcal{L}(\theta_0)} = \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)}$$

and we can reject if $\Lambda(\vec{Y}) \geq k$ for some k.

What if we have

$$Y_1, \ldots, Y_n \overset{ind}{\sim} \lambda_i e^{-\lambda_i y_o}$$

where they are independent, but not identically distributed and we wish to test

$$H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_n \text{ vs. } H_1 : \text{not equal}$$

Instead, we have

$$\Lambda(\vec{Y}) = \frac{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta)} = \frac{\mathcal{L}(\hat{\theta})}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta)}$$

The numerator is $\hat{\lambda}_i = \frac{1}{y_i}$ and in the denominator we have $\hat{\lambda} = \frac{1}{\bar{Y}}$.

15. March 27th, 2014 (Hypothesis Testing Continued)

Recall that we have the setup $y_1, y_2, \ldots, y_n$ are iid data, and we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

Questions to ask:

(0) How should you choose α? Why are type I and type II treated differently?
(1) How should you choose H_0 and H_1? Can you swap them? Should you do one-sided or two-sided alternative?
(2) Statistical significance vs practical or scientific significance, entanglement with sample size issues.
(3) What should a Bayesian do?
(4) Relationship with interval estimation.

15.1. Famous tests for the above hypothesis test.

Let $\hat{\theta}$ be the MLE

(1) Likelihood ratio test:

$$T = \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)}$$

is the test statistic. We reject H_0 if T is large. "Large " seems arbitrary, but it depends on the value of α. We find c such that

$$\mathbb{P}(T \geq c | H_0) = \alpha$$

What if we have discrete data? We cannot make it equal to α. Then we have

$$\begin{cases} T > c & \text{reject } H_0 \\ T < c & \text{retain } H_0 \\ T = c & \text{flip coin with some probability of heads} \end{cases}$$

Usually we take the log:

$$l(\hat{\theta}) - l(\theta_0)$$

This is invariant to transformations.

(2) Score test: Use $S(y,\theta_0)$ as the test statistic. Reject H_0 if $|S(y,\theta_0)|$ is large. However, it is hard to find an exact threshold value as the distribution of the score function is not easy to get. For large n and under H_0,

$$S(y,\theta_0) \dot{\sim} \mathcal{N}(0, \mathcal{I}(\theta_0))$$

The advantage of this test over the others is that it does not require knowing $\hat{\theta}_0$. It is also the "locally most-powerful one-sided".

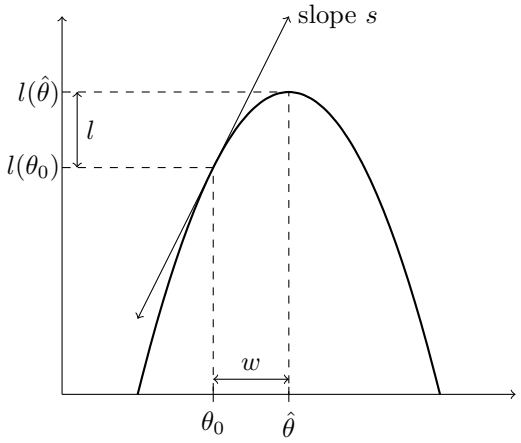
(3) Wald test: Directly use the MLE $\hat{\theta}$ as the test statistic. Again the exact distribution of the MLE is usually difficult to get, however for large n and under H_0,

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta_0, \frac{1}{\mathcal{I}(\theta_0)}\right)$$

This is not invariant to transformations.

Asymptotically, these three tests are equivalent. Usually the first two are better unless you're sure the asymptotics for the Wald test have kicked in.

Ideal Case: log-likelihood is quadratic



$l(\hat{\theta}) - l(\theta_0)$: likelihood-ratio test, length

$\hat{\theta} - \theta_0$: Wald test, width

s : score test, slope

Theorem 11 (Wilks' Theorem)**.** *Under regularity,*

$$2\log(LR) \xrightarrow{D} \chi^2_1$$

More generally, for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ where $\Theta_0 \subseteq \Theta_1$, then

$$2\log(LR) \xrightarrow{D} \chi^2_{\dim(\Theta_1)-\dim(\Theta_0)}$$

where

$$LR = \frac{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta_1)}{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta_0)}$$

15.2. **Bayesian Hypothesis Testing.** Goal: Find $\mathbb{P}(H_0|\text{data})$. Suppose we want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Assume there are only 2 possibilities.

$$\underbrace{\frac{\mathbb{P}(H_0|y)}{\mathbb{P}(H_1|y)}}_{\text{posterior odds for } H_0} = \underbrace{\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}}_{\text{prior odds for } H_0} \times \underbrace{\frac{f(y|\theta_0)}{f(y|\theta_1)}}_{\text{likelihood ratio}}$$

If we have $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$, where $\Theta_1 = \Theta_0^c$.

$$\frac{\mathbb{P}(H_0|y)}{\mathbb{P}(H_1|y)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \underbrace{\frac{f(y|H_0)}{f(y|H_1)}}_{\text{Bayes' factor}}$$

where

$$f(y|H_0) = \int_{\Theta_0} f(y|\theta)\pi(\theta|\theta\in\Theta_0)\,\mathrm{d}\theta = \int_{\Theta_0} f(y|\theta)\pi(\theta)\,\mathrm{d}\theta$$

so

$$\mathbb{P}(H_0) \to \int_{\Theta_0} \pi(\theta)\,\mathrm{d}\theta$$

16.1. **Jeffreys Lindleys Paradox.** Suppose $y_1, \dots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ^2 known. So $\bar{y}$ is the MLE, a UMVUE, and a CSS. We test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. In this situation, the Wald test, score test, and likelihood ratio tests are all the same, where we reject H_0 if $|\bar{y}|$ is large. Suppose n is very large and $\alpha = 0.05$. Let

$$Z_n \equiv \sqrt{n}\frac{\bar{y}}{\sigma} = 3$$

Then in the frequentist setting, we reject H_0.

16.1.1. *Bayesian Approach.* In this situation, we would need a prior. The prior we choose puts mass at 0. If it is non-zero, then we have a continuous prior. Assume prior, $\mathbb{P}(H_0) = \frac{1}{2}$. We know that $\mu|H_0 = 0$ trivially. To get the distribution conditioned on the alternative, we have

$$\mu|H_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{r}\right)$$

We wish to find the posterior odds of H_1, which is the prior odds multiplied by the Bayes factor. This equals

$$\frac{\displaystyle\int_{-\infty}^{\infty} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2} \frac{\sqrt{r}}{\sqrt{2\pi}\sigma} e^{-\frac{r}{2\sigma^2}} \,\mathrm{d}\mu}{\dfrac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{n\bar{y}}{2\sigma^2}\right)}$$

If we look at the numerator, we see that this is exactly a convolution. Recall that convolutions take the form

$$h(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\,\mathrm{d}x$$

The numerator is exactly a convolution of $\mathcal{N}\left(0, \frac{\sigma^2}{r}\right)$ and $\mathcal{N}\left(0, \frac{\sigma^2}{r}\right)$, which implies the numerator is

$$\frac{1}{\sqrt{2\pi}\sigma\sqrt{\frac{1}{n}+\frac{1}{r}}} \exp\left(-\frac{\bar{y}^2}{2\sigma^2\left(\frac{1}{n}+\frac{1}{r}\right)}\right)$$

Substituting this in and cancelling denominator terms, we get that the posterior odds is

$$\text{odds}(H_1|\text{data}) = \sqrt{\frac{r}{r+n}} \exp\left(\frac{n}{n+r}Z_n^2\right) \approx 0$$

16.2. **Decision Theory.** Imagine a game of nature vs statistician. Nature chooses $\theta \in \Theta$. Statistician, who does not get to observe θ, observes data, then takes an action a, e.g.

(1) Provides point estimate $\hat{\theta}$.
(2) Provides interval estimate.
(3) Decides whether or not to reject some null hypothesis.

Each player only goes once, then the game ends. There is a loss/cost $C(\theta, a)$.

Suppose we are testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. We also have a_0 is retain H_0 and a_1 is reject H_0.

	a_0	a_1
θ_0	0	c_0
θ_1	c_1	0

where $c_0, c_1 > 0$ as we need to have some penalty for choosing wrong. First assume no data. The Bayesian approach with prior $\pi_c = \mathbb{P}(\theta = \theta_i)$. The goal is choose the action that minimizes the expected loss. If we take action 0, then the cost is $\pi_1 c_1$. If we take action 1, then the cost is $\pi_0 c_0$. So the decision rule is to take action 1 if $\pi_1 c_1 > \pi_0 c_0$ or equivalent if

$$\frac{\pi_1 c_1}{\pi_0 c_0} > 1$$

This is called the no data problem. Suppose we now have data $y_1, \dots, y_n$. Update the prior to posterior and then do as above, that is, take action 1 if

$$\frac{\pi_1 f(y|\theta_1)}{\pi_0 f(y|\theta_0)} > 1$$

This is the Bayesian analog of Neyman-Pearson. In a more general setting (say if we have a composite hypothesis), then we replace the likelihood ratio by the Bayes factor.

The loss function is

$$C(\theta, \delta(y))$$

where δ is the decision rule, which determines which action we take. The risk function is

$$R_\delta(\theta) = \mathbb{E}_\theta[C(\theta, \delta(y))]$$

which is dependent on a decision rule.

16.2.1. *Decision Rules.*

Definition 11 (Minimaxity)**.** *δ_0 is minimax if it minimizes*

$$\sup_\theta R_\delta(\theta)$$

Definition 12 (Admissibility)**.** *δ_1 dominates δ_2 if for all θ*

$$R_{\delta_1}(\theta) \leq R_{\delta_2}(\theta)$$

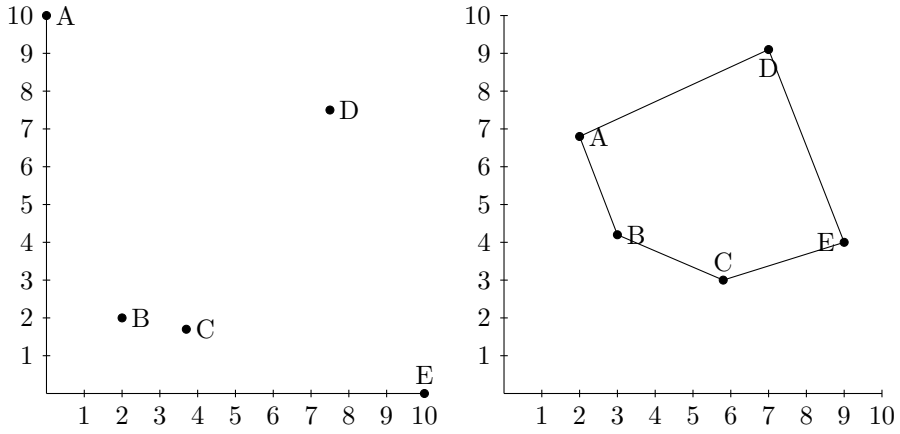
This is strict for at least one θ value. δ is admissible if no $\tilde{\delta}$ dominates it.

Theorem 12. *Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ and π be a non-dogmatic prior, that is a distribution that does not assign zero probability to any θ_i. Let δ be a Bayes rule, where we would want to minimize the expected posterior loss*

$$\mathbb{E}[C(\theta, \delta(y))|y]$$

Then δ is admissible.

17. April 3rd, 2014



These graphs are called two risk diagrams where the x-axes are $R(\theta_1, \delta)$ and the y-axes are $R(\theta_2, \delta)$.

$$R_\delta(\theta) = \mathbb{E}[C(\theta, \delta(y)]$$

Consider the case where Θ is finite, $\{\theta_1, \ldots, \theta_n\}$. The risk function is then a point in $\mathbb{R}^n$.

In the first graph, the admissible values are all but D. The minimax value is B. In the second graph, the admissible values are A, B, and C and the lines connecting them. Intersect it with 45° line.

Definition 13 (Bayes Risk)**.** *The Bayes risk of a procedure δ with respect to prior π is*

$$\begin{aligned}\mathbb{E}[C(\theta, \delta(y)] &= \iint C(\theta, \delta(y)) f(y|\theta)\pi(\theta)\,\mathrm{d}y\mathrm{d}\theta \\ &= \int R_\delta(\theta)\pi(\theta)\,\mathrm{d}\theta\end{aligned}$$

so the expectation is averaged over both θ and y. How would we choose δ to minimize this?

Minimizing the Bayes risk is equivalent to minimizing the posterior expected loss. By Fubini's theorem, we have

$$\begin{aligned}\mathbb{E}[C(\theta, \delta(y)] &= \iint C(\theta, \delta(y)) f(y)\pi(\theta|y)\,\mathrm{d}\theta\mathrm{d}y \\ &= \int \underbrace{\left(\int C(\theta, \delta(y))\pi(\theta|y)\,\mathrm{d}\theta\right)}_{\text{posterior expected loss}} f(y)\,\mathrm{d}y\end{aligned}$$

Now we can see that we minimize the posterior expected loss since outside of that integral, there are no θ terms and so these two are equivalent.

Now we prove that non-dogmatic Bayes implies admissible.

Proof. Let δ be a non-dogmatic Bayes with respect to π. Suppose $\tilde{\delta}$ dominates δ. Compute Bayes risks. Since δ is the Bayes rule,

$$\sum_{j=1}^{N} R_\delta(\theta_j)\pi_j \leq \sum_{j=1}^{N} R_{\tilde{\delta}}(\theta_j)\pi_j \qquad (\star)$$

However, this contradicts since

$$R_{\tilde{\delta}}(\theta) \leq R_\delta(\theta)$$

sometimes strict. The only case in which this would not imply $(\star)$ is if some of the $\pi_j = 0$. However, this cannot happen since we have a non-dogmatic prior. □

Theorem 13 (Complete Class Theorem)**.** *Let $\Theta = \{\theta_1, \ldots, \theta_n\}$. Then any admissible rule is a Bayes rule.*

Suppose we have the following decision table.

	a_0	a_1
θ_0	101	100
θ_1	3	100

Would you want to choose action 0 or 1? At a glance, without assuming any prior, we would probably choose action 0. However, a_1 is the minimax solution, which seems like a bad choice. Suppose now instead, we have a \$20 choice under θ_1.

	a_0	a_1
θ_0	101	100
θ_1	23	120

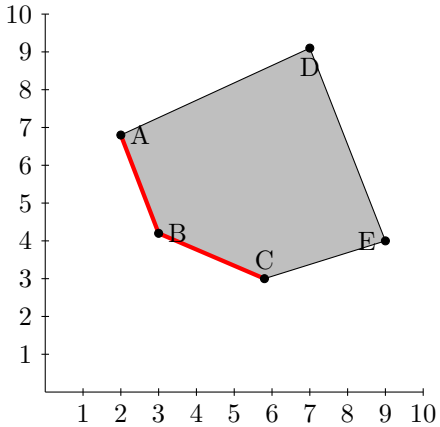
In this case, a_0 is the minimax. What this is implying is that the minimax decision changes if you add a constant to a row.

17.0.2. *Minimax Regret.* The regret table is set up such that an entry in a row takes on the value of the difference between its original value and the lowest value in the row. That is, if we choose a_0 under θ_0, then we have regret $101 - 100 = 1$.

	a_0	a_1
θ_0	1	0
θ_1	0	97

Back to the first two risk diagram, the minimax regret rule is B. Suppose now instead, we eliminate A. Now, the minimax regret rule is C. This violates what is called the "independence of of irrelevant alternatives".

Example 28. *You are at a restaurant and you have to choose between a soup or salad and you choose soup. Later, the waiter offers a pasta dish and you choose to have salad instead. This doesn't make sense.*



The Bayes rule with respect to π is the δ such that we minimize

$$\pi_1 r_1 + \pi_2 r_2$$

However, this is like the equation of a line $ax + by = c$. The prior controls the slope. So for a given prior/slope, we change the intercept until it hits the frontier (red lines). In higher dimensions we would use something called the supporting hyperplane theorem.

18. April 8th, 2014

18.1. **Permutation Tests vs Randomization Tests.**

Permutation tests $\longrightarrow$ Super populations, weak distributional assumptions

Randomization tests $\longrightarrow$ Finite population of experimental units $\longrightarrow$ "Randomization"

Suppose we have 6 units and random pick 2 of them to give the "old method" and the other 4 we give the "new method". Suppose our values are $x_1 = 6, x_2 = 8$ for the "old method" and $y_3 = 7, y_4 = 18, y_5 = 11, y_6 = 9$ for the "new method". How would we test that this test was significant? We could use a two sample t-test, where

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{s^2\left(\frac{1}{2} + \frac{1}{4}\right)}}$$

and we can compare

$$t^{obs} = 1.17 > t_{4,0.5}$$

Let

$$I_i = \begin{cases} 1 & i \in X \\ 0 & i \in Y \end{cases}$$

Then we have a table of potential outcomes/counterfactuals

Unit i	$Z_i(X)$	$Z_i(Y)$	I_i^{obs}	Z_i^{obs}
1	$Z_1(X)$	$Z_1(Y)$	1	6
2	$Z_2(X)$	$Z_2(Y)$	1	8
3	$Z_3(X)$	$Z_3(Y)$	0	7
4	$Z_4(X)$	$Z_4(Y)$	0	18
5	$Z_5(X)$	$Z_5(Y)$	0	11
6	$Z_6(X)$	$Z_6(Y)$	0	9

From what we observe, we have $Z_1(X) = 6, Z_2(X) = 8, Z_3(Y) = 7, Z_4(18), Z_5(11), Z_6(9)$. Suppose we have H_0 be the treatment having no effect. How do we find a test statistic? We can use t^{obs} as we did before. But now, we could also use

$$\begin{aligned} W^{obs} &= \text{sum of the ranks of } r_{obs} \\ &= 2 + 4 + 5 + 6 \\ &= 17 \end{aligned}$$

To impute the missing data under H_0, we simply have that $Z_i(X) = Z_i(Y)$.

If we have $t^{obs} = 1.17$, then the p-value is $\frac{2}{15}$.

W	10	$\cdots$	17	18	total
Prob	1	$\cdots$	1	1	15

So the p-value is $\frac{2}{15}$.

Suppose $X_1, \ldots, X_m \overset{iid}{\sim} F_X(x)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} F_Y(x)$. We wish to test $H_0 : F_X(x) = F_Y(x)$.

18.1.1. *Test Procedure.*

(1) $N = m + n$ units. Randomly assign m units to X and n units to Y.
(2) Run the experiment, obtain (I^{obs}, Z^{obs}).
(3) Assume Z^{obs} is fixed (condition on Z^{obs}), and test statistic $T = T(Z)$, so $T^{obs} = T(Z^{obs})$.
(4) Consider all permutations M_N pf Z^{obs}. Let $\tilde{Z}_j$ be the jth permutation. Compute $\tilde{Y}_j = T(\tilde{Z}_j)$.
(5) Compute

$$p = \frac{1}{M_N} \sum_{j=1}^{M_N} I\left\{\tilde{T}_j \geq T^{obs}\right\}$$

(6) Reject H_0 if $p < \alpha$.

Claim. *The test is of level α for H_0.*

(1) Under H_0, Z^{obs} has the same distribution as $\tilde{Z}_1, \ldots, \tilde{Z}_{M_N}$. (Exchangeability of units)
(2) T^{obs} has the same distribution as $T(\tilde{Z}_j)$.
(3) Ranks of $T^{obs}, T(\tilde{Z}_1), \ldots, T(\tilde{Z}_{N_M})$ are uniform.

How do we actually generate the distribution of T and obtain cut-offs?

(1) Use Monte-Carlo methods.
(2) Use asymptotic approximations.
- Asymptotic Normality (Ranked-based statistics, e.g. W) If

$$W = \sum_{m+1}^{n} Y_j$$

this is called the Wilcoxon Rank-Sum statistic.

- Box-Anderson test

18.2. **Linear Statistics.**

$$Z^{obs} = (Z_1^{obs}, \dots, Z_N^{obs})'$$
$$\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_N)$$

where $\tilde{Z}$ is any permutation of Z^{obs}. A linear statistic is of the form

$$T = \sum_{j=1}^{N} c_j \tilde{Z}_j$$

We can write the rank-sum statistic as

$$W = \sum_{j=1}^{N} \tilde{R}_j$$

where

$$R^{obs} = (R_1^{obs}, \dots, R_N^{obs})'$$
$$\tilde{R} = (\tilde{R}_1, \dots, \tilde{R}_N)'$$

again where $\tilde{R}$ is any permutation of R^{obs}. We also have

$$c_1 = \cdots = c_m = 0 \quad c_{m+1} = \cdots = c_N = 1$$

Under H_0,

$$\mathbb{P}\Big(\tilde{Z}_i = Z_s^{obs}\Big) = \frac{1}{N} \qquad \forall s$$

For $i \neq j$,

$$\mathbb{P}\Big(\tilde{Z}_i = Z_s^{obs}, \tilde{Z}_j = Z_t^{obs}\Big) = \frac{1}{N(N-1)} \qquad s \neq t$$

We can then get

$$\mathbb{E}\Big[\tilde{Z}_1 | Z^{obs}\Big] = \bar{Z}^{obs}$$
$$\mathrm{Var}(\tilde{Z}_i | Z^{obs}) = \frac{1}{N} \sum_{s=1}^{N} (Z_s^{obs} - \bar{Z}^{obs})^2$$
$$\mathrm{Cov}(\tilde{Z}_i, \tilde{Z}_j | Z^{obs}) = -\frac{1}{N(N-1)} \sum_{s=1}^{N} (Z_s^{obs} - \bar{Z}^{obs})^2$$

This will tell us that

$$\mathbb{E}\big[T | Z^{obs}\big] = N \bar{Z}^{obs} \bar{c}$$
$$\mathrm{Var}(T | Z^{obs}) = \frac{1}{N-1} \sum_{s=1}^{N} (Z_s^{obs} - \bar{Z})^2 \sum_{i=1}^{N} (c_I - \bar{c})^2$$

Using the fact that

$$1 + 2 + \cdots + N = \frac{N(N+1)}{2}$$
$$1^2 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}$$

we obtain

$$\mathbb{E}\big[W | R^{obs}\big] = \frac{n(N+1)}{2}$$
$$\mathrm{Var}(W | R^{obs}) = \frac{mn(N+1)}{12}$$

Theorem 14 (Hajek)**.** *Let $T = \sum_{i=1}^{N} c_i R_i$ where the rank R comes from data vector Z that is continuous (no ties with probability 1) and exchangeable and if $c_1, \dots, c_n$ satisfy*

$$\frac{\sum_{i=1}^{N} (c_i - \bar{c})^2}{\max_{1 \leq i \leq N} (c_i - \bar{c})^2} \longrightarrow \infty$$

as $N \to \infty$, then T is approximately $\mathcal{N}(\mathbb{E}[T], \mathrm{Var}(T))$ for large N.

19. April 10th, 2014

This lecture will have **slides** to accompany it.

20. April 15th, 2014

Suppose we have $y_1 \sim \mathcal{N}(\theta_1, 1), y_2 \sim \mathcal{N}(\theta_2, 1), \ldots, y_k \sim \mathcal{N}(\theta_k, 1)$, where $k \geq 3$ and $y_1, \ldots, y_k$ are independent. Our goal is to estimate $\theta = (\theta_1, \ldots, \theta_k)$, where the loss function is the total squared error. One obvious estimator is simply $y = (y_1, \ldots, y_k)$ as this is the MLE. This is also the UMVUE and even is the minimax. A great advantage here is that it is very simply, however it is **inadmissible**. This is dominated by

$$\hat{\theta}_{JS} = \underbrace{\left(1 - \frac{k-2}{\|y\|^2}\right)}_{\text{Shrinkage towards the origin}} y$$

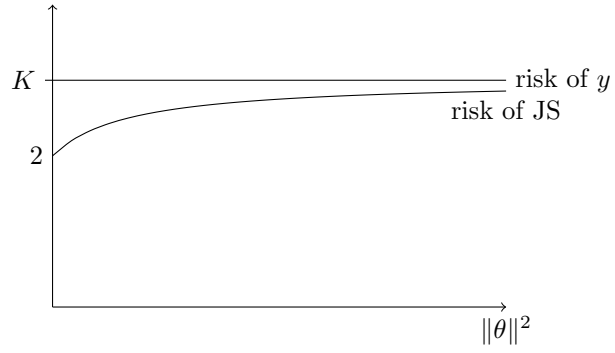
which is known as the **James-Stein estimator**. However, this itself is inadmissible, since the shrinkage term can be potentially negative. It we set if to zero if this term is negative, then this turns out to dominated the James-Stein estimator.

In the naive case, the risk function is

$$\sum_{j=1}^{k} \mathbb{E}\big[(y_j - \theta_j)^2\big] = k$$

so we have constant risk. In the James-Stein case, our risk function is a function of $\|\theta\|^2$ (this is proven elsewhere) and we have

$$R_{JS}(\|\theta\|^2) < k$$



Example 29. *Suppose we wish to estimate $Y = \|\theta\|^2$. Then the MLE is*

$$\|y\|^2 = \sum_j y_j^2$$

which has a non-central χ^2 distribution. This gives

$$\mathbb{E}\big[\|y\|^2\big] = \|\theta\|^2 + k$$

If k is large, then we have a large bias. However, we simply have that $\|y\|^2 - k$ is unbiased.

What if we tried the Bayesian approach with flat priors on θ_j, which are independent? Then the posterior mean is

$$\begin{aligned}\mathbb{E}\big[\|\theta\|^2 | y\big] &= \sum_j \mathbb{E}\big[\theta_j^2 | y\big] \\ &= \|y\|^2 + k\end{aligned}$$

since if $y_j|\theta_j \sim \mathcal{N}(\theta_j, 1)$ and θ_j is a flat prior, then $\theta_j|y_j \sim \mathcal{N}(y_j, 1)$.

Proof.

$$R_{JS} = K - \mathbb{E}\left[\frac{(k-2)^2}{k - 2 + 2J}\right]$$

where $J \sim \text{Pois}\left(\frac{\|\theta\|^2}{2}\right)$.

Use Stein's Lemma: If $X \sim \mathcal{N}(\mu, 1)$, then

$$\mathbb{E}[(X - \mu)h(X)] = \mathbb{E}[h'(X)]$$

Suppose we estimate θ with $y - g(y)$ where $g : \mathbb{R}^k \mapsto \mathbb{R}^k$. We find the risk function using Stein's Lemma.

$$\begin{aligned}\|y - g(y) - \theta\|^2 &= \sum_j (y_j - \theta_j - g_j(y))^2 \\ &= \sum_j (y_j - \theta_j)^2 + \sum_j g_j(y)^2 - 2\sum_j (y_j - \theta_j) g_j(y)\end{aligned}$$

Taking the expectation on both sides, we get the risk of $y - g(y)$ is

$$\mathbb{E}\big[\|y - g(y) - \theta\|^2\big] = k + \mathbb{E}\big[\|g(y)\|^2\big] - 2\sum_j \mathbb{E}[(y_j - \theta_j) g_j(y)]$$

If we took away the expectation, i.e.

$$k + \|g(y)\|^2 - 2\sum_j \frac{\partial g_j(y)}{\partial y_j}$$

then this is called the **Stein's Unbiased Risk Estimate (SURE)**.

Next, we have that

$$g(y) = \frac{k-2}{\|y\|^2} y \qquad g_j(y) = \frac{k-2}{\|y\|^2} y_j$$

and so

$$\|g(y)\|^2 = \frac{(k-2)^2}{\|y\|^4}\|y\|^2 = \frac{(k-2)^2}{\|y\|^2}$$

Then we get

$$\frac{\partial g_j(y)}{\partial y_j} = (k-2)\left(\frac{\|y\|^2 - y_j}{\|y\|^4}\right)$$

If we take the sum, we get

$$\sum_j \frac{\partial g_j(y)}{\partial y_j} = \frac{(k-2)^2}{\|y\|^4}\|y\|^2 = \frac{(k-2)^2}{\|y\|^2}$$

Therefore the risk is

$$k - \mathbb{E}\left[\frac{(k-2)^2}{\|y\|^2}\right]$$

□

21. April 17th, 2014

Our setup was that $y_j|\mu_j \sim \mathcal{N}(\mu_j, V_j)$, $1 \le j \le k$ where $k \ge 3$ and assuming V_j are known. A more general loss function is

$$\sum_j (\hat{\mu}_j - \mu_j)^2 w_j$$

where $w_j > 0$ are "weights", e.g. $w_j = 1$ or $w_j = \frac{1}{V_j}$ (also known as "statistician's weights").

Now imagine we are not in the context of the weights. Another approach is to normalize all our variables such that they each have the same variance. Let

$$Z_j \equiv \frac{y_j}{\sqrt{V_j}} \sim \mathcal{N}\left(\frac{\mu_j}{\sqrt{V_j}}, 1\right)$$

Now, in this case we minimize

$$\min \sum_j \left(\frac{\hat{\mu}_j}{\sqrt{V_j}} - \frac{\mu_j}{\sqrt{V_j}}\right)^2 = \min \sum_j (\hat{\mu}_j - \mu_j)^2 \frac{1}{V_j}$$

21.1. **Two-Level Model.**

21.1.1. *Descriptive.* How do we extend this to the two-level model?

$$\begin{aligned} y_j|\mu_j &\overset{iid}{\sim} \mathcal{N}(\mu_j, V_j) \\ \mu_j &\overset{iid}{\sim} \mathcal{N}(\mu_0, A)\end{aligned}$$

Suppose that μ_0 and A are known.

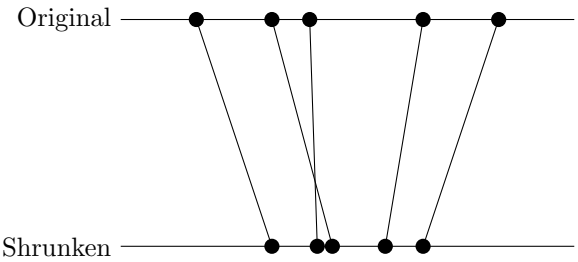
21.1.2. *Inferential.* In this case, we are interested in

$$y_j \sim \mathcal{N}(\mu_0, V_j + A)$$
$$\mu_j | y_j \sim \mathcal{N}((1 - B_j)y_j + B_j\mu_0, (1 - B)V_j)$$

where

$$B_j = \frac{V_j}{V_j + A}$$

is called the shrinkage factor.



Note that we have a crossover here when we do the shrinkage. This has to do with the fact that the V_j are not constant, so the ordering is not necessarily preserved. Suppose we have two baseball players. Player 1 has a batting average of 1/1. Player 2 has a batting average of 45/100. How can we determine which player is better? It is hard to tell because the amount of data is very small. If we have prior knowledge about baseball, we could say Player 2's statistics are impressive.

Let us go back to the case where

$$y_j | \mu_j \overset{iid}{\sim} \mathcal{N}(\mu_j, V)$$
$$\mu_j \sim \mathcal{N}(0, A)$$

where now $V_j = V$ and $\mu_0 = 0$. If A is known, then

$$\hat{\mu}_j = (1 - B)y_j \qquad B = \frac{V}{V + A}$$

If A is unknown, then $A \sim h$ where h is a pdf on $(0, \infty)$.

$$\begin{aligned}\mathbb{E}[\mu_j | y] &= \mathbb{E}[\mathbb{E}[\mu_j | y, A] \, | y] \\ &= \mathbb{E}[(1 - B)y_j | y] \\ &= (1 - \mathbb{E}[B|y])y_j\end{aligned}$$

We get the James-Stein if

$$A \sim \mathcal{U}(-V, \infty)$$

which is a ridiculous, improper prior. However, this gives the correct result. We can improve this if we take

$$A \sim \mathcal{U}(0, \infty)$$

which is known as Stein's harmonic prior.

21.2. **Empirical Bayes.** A problem with the hierarchical models is when to stop adding levels. The empirical Bayes uses the data to estimate the hyperparameters.

If $y \sim \mathcal{N}(0, V + A)$,

$$S = \sum_{i=1}^{n} y_i^2 \sim (V + A)\chi_k^2$$

In expectation,

$$\mathbb{E}\left[\frac{S}{K}\right] = V + A$$

so we can use

$$\frac{S}{K} - V$$

to estimate A.

We have $S \sim (V + A)\chi_k^2$, but we want to estimate $B = \frac{V}{V+A}$. Instead, we can look at $\frac{1}{S}$. Since

$$\mathbb{E}\left[\frac{1}{\chi_k^2}\right] = \frac{1}{k - 2}$$

we get that

$$\mathbb{E}\left[\frac{1}{S}\right] = \frac{1}{(V+A)(k-2)}$$
$$\mathbb{E}\left[\frac{(k-2)V}{S}\right] = \frac{V}{V+A}$$
$$= B$$

Let $\hat{B} = \frac{(k-2)V}{S}$. Use

$$\hat{\mu}_j = (1-\hat{B})y_j$$

This is the James-Stein estimator.

22. April 22nd, 2014

23. April 24th, 2014

Recall that

$$\hat{\theta}_{MLE} = \arg\max_{\theta} l(\theta|x)$$

23.1. **Mixture of Normal Densities.** $Y_1, \ldots, Y_n$ are iid from a 50-50 mixture of $\mathcal{N}(\mu_0, 1)$, $\mathcal{N}(\mu_1, 1)$. Let $\vec{\mu} = (\mu_0, \mu_1)$.

$$\mathcal{L}(\mu_0, \mu_1|y) = \prod_{j=1}^{n} \left\{ \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_j-\mu_0)^2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_1)^2} \right\}$$

- Logistic Regression:

$$Y_i \sim \text{Bern}\left(\frac{e^{\vec{X_i}'\vec{\beta}}}{1+e^{\vec{X_i}'\vec{\beta}}}\right)$$

- Probit:

$$Y_i \sim \text{Bern}(\Phi(\vec{X_i}'\vec{\beta}))$$

and the likelihood is

$$\mathcal{L}(\vec{\beta}, \vec{Y}, \vec{X}) = \prod_{i=1}^{n} \left[\Phi(\vec{X_i}'\vec{\beta})^{Y_i} \left\{ 1 - \Phi(\vec{X_i}'\vec{\beta}) \right\}^{1-Y_i} \right]$$

Suppose you are in a factory and you have a line of items and you need to determine whether there are defects. You stop the process if you encounter a defect. Because inspection is costly/destructive, you cannot inspect each one individually. Instead, we inspect at intervals $m = 10$. Suppose the process becomes faulty with probability p.

Past a certain point, the proportion of faulty ones is $\pi \in (0, 1]$. Suppose we look at the first 17 and they are all defect free. If the first defect is found at the 30th one, then we call this a cycle. T_I is the length of the cycle and C_i is the cost associated with each cycle.

$$T_i \sim f_{p,\pi}$$

$T_1, T_2, T_3, \ldots$ is known as a renewal process. If we look at $(T_1, C_1), (T_2, C_2), (T_3, C_3), \ldots$ is known as a renewal-reward process.

- $f(\vec{y}|\theta)$: observed/incomplete
- $f(\vec{y}, \vec{z}|\theta)$: unobserved/complete

We would like to maximize these. Note that we can link this two by

$$f(\vec{y}, \vec{z}|\theta) = f(\vec{y}|\theta) f(\vec{z}|\vec{y}, \theta)$$

which implies

$$f(\vec{y}|\theta) = \frac{f(\vec{y}, \vec{z}|\theta)}{f(\vec{z}|\vec{y}, \theta)}$$

so that

$$l(\theta|\vec{y}) = l(\theta|\vec{y}, \vec{z}) - \log(f(\vec{z}|\vec{y}, \theta))$$

The idea here is to take the conditional expectation with respect to the density of $\vec{z}$ given $\vec{y}$, $\theta^{(t)}$.

$$l(\theta|\vec{y}) = \underbrace{\mathbb{E}_{\theta^{(t)}}[l(\theta|\vec{y}, \vec{z})|\vec{y}]}_{Q(\theta|\theta^{(t)})} - \mathbb{E}_{\theta^{(t)}}[\log(f(\vec{z}|\vec{y}, \theta))|\vec{y}]$$

23.2. **Expectation-Maximization Algorithm (EM Algorithm).**

(1) Compute $Q(\theta|\theta^{(t)})$
(2) Find $\arg\max_\theta Q(\theta|\theta^{(t)})$.

23.2.1. *Properties of EM.* The sequence $\{\theta^{(t)}\}$ results in an increasing sequence $\{l(\theta^{(t)}|\vec{y})\}$.

Proof. To prove this, it suffices to show

(a) $\mathbb{E}_{\theta^{(t)}}\left[l(\theta^{(t+1)}|\vec{y},\vec{z})\middle|\vec{y}\right] \geq \mathbb{E}_{\theta^{(t)}}\left[l(\theta^{(t)}|\vec{y},\vec{z})\middle|\vec{y}\right]$

(b) $\mathbb{E}_{\theta^{(t)}}\left[\log(f(\vec{z}|\vec{y},\theta^{(t+1)}))\middle|\vec{y}\right] \leq \mathbb{E}_{\theta^{(t)}}\left[\log(f(\vec{z}|\vec{y},\theta^{(t)}))\middle|\vec{y}\right]$

(a) is obvious by construction. (b) follows from the fact that the difference between the two related to the Kullback-Leibler divergence.

$$\begin{aligned}
\mathbb{E}_{\theta^{(t)}}\left[\log(f(\vec{z}|\vec{y},\theta^{(t+1)}))\middle|\vec{y}\right] - \mathbb{E}_{\theta^{(t)}}\left[\log(f(\vec{z}|\vec{y},\theta^{(t)}))\middle|\vec{y}\right] &= \mathbb{E}_{\theta^{(t)}}\left[\log\left(\frac{f(\vec{z}|\vec{y},\theta^{(t+1)})}{f(\vec{z}|\vec{y},\theta^{(t)})}\right)\right] \\
&\leq \log \mathbb{E}_{\theta^{(t)}}\left[\frac{f(\vec{z}|\vec{y},\theta^{(t+1)})}{f(\vec{z}|\vec{y},\theta^{(t)})}\right] \qquad \text{by Jensen's inequality} \\
&= \log \int \frac{f(\vec{z}|\vec{y},\theta^{(t+1)})}{f(\vec{z}|\vec{y},\theta^{(t)})} f(\vec{z}|\vec{y},\theta^{(t)})\,\mathrm{d}\vec{z} \\
&= 0
\end{aligned}$$

□

Suppose $Q(\theta|\theta^{(t)})$ is continuous in both θ and $\theta^{(t)}$. Then all the limit points of the sequence $\{\theta^{(t)}\}$ converge monotonically to $l^* = l(\theta^*|y)$, where θ^* is some stationary point.

Theorem 15. *Suppose $l(\theta|\vec{y})$ is unimodal with θ^* being the only stationary point and that $\frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)})$ is continuous. Then for any EM sequence $\{\theta^{(t)}\}$, $\theta^{(t)}$ converges to θ^*.*

Let us go back to the mixture model, where $Y_i, \ldots, Y_n$ are iid from a mixture of $\mathcal{N}(\mu_0, 1)$ and $\mathcal{N}(\mu_1, 1)$. Let $Z_1, \ldots, Z_n$ denote indicators, where if $Z_i = 1$, then $Y_i \sim \mathcal{N}(\mu_1, 1)$ and if $Z_i = 0$, $Y_i \sim \mathcal{N}(\mu_0, 1)$.

$$f(y_j, z_j|\vec{\mu}) = \{\phi(y_j - \mu_0)\}^{1-z_j}\{\phi(y_j - \mu_0)\}^{z_j}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

$$\begin{aligned}
l(\vec{\mu}|\vec{y},\vec{z}) &= \sum_{j=1}^{n} \log f(y_j, z_j|\vec{\mu}) \\
&= -\frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{n}(1-z_j)(y_j-\mu_0)^2 - \frac{1}{2}\sum_{j=1}^{n} z_j(y_j-\mu_1)^2
\end{aligned}$$

Let $\mu(t) = (\mu_1(t), \mu_2(t))$ be the EM-value at the tth step. Denote

$$E^{(t)}(Z_j) = \mathbb{E}_{\vec{\mu}(t)}[Z_j|Y_j]$$

Then the E-step is

$$Q(\vec{\mu}|\vec{\mu}^{(t)}) \approx \frac{1}{2}\sum_{j=1}^{n}(1 - E^{(t)}(Z_j))(y_j-\mu_0)^2 - \frac{1}{2}\sum_{j=1}^{n} E^{(t)}(Z_j)(y_j-\mu_1)^2$$

The M-step is

$$\frac{\partial}{\partial\mu_1}Q(\vec{\mu}|\vec{\mu}^{(t)}) = 0 \Rightarrow \mu_1^{(t+1)} = \frac{\sum_{j=1}^{n} y_j E^{(t)}(Z_j)}{\sum_{j=1}^{n} E^{(t)}(Z_j)}$$

and

$$\frac{\partial}{\partial\mu_0}Q(\vec{\mu}|\vec{\mu}^{(t)}) = 0 \Rightarrow \mu_0^{(t+1)} = \frac{\sum_{j=1}^{n} y_j\{1 - E^{(t)}(Z_j)\}}{\sum_{j=1}^{n}\{1 - E^{(t)}(Z_j)\}}$$

So then

$$\begin{aligned}
E^{(t)}(Z_j) &= \mathbb{P}\Big(Z_j = 1|Y_j, \mu^{(t)}\Big) \\
&= \frac{f(y_j|Z_j = 1, \mu^{(t)})\mathbb{P}(Z_j = 1)}{f(y_j|Z_j = 1, \mu^{(t)})\mathbb{P}(Z_j = 1) + f(y_j|Z_j = 0, \mu^{(t)})\mathbb{P}(Z_j = 0)} \\
&= \frac{\frac{1}{2}\phi(y_j - \mu_1^{(t)})}{\frac{1}{2}\phi(y_j - \mu_1^{(t)}) + \frac{1}{2}\phi(y_j - \mu_0^{(t)})}
\end{aligned}$$

Let $y_1, \dots, y_n \overset{iid}{\sim} F$. Suppose we have some parameter of interest θ, with an estimator $\hat{\theta}$. We want to estimate the variance and bias of $\hat{\theta}$. There are 3 challenges here:

(1) Variance, bias depend on θ in general.
(2) Math often gets hard, $\hat{\theta}$ may not even be in closed form.
(3) We only have one data set, so we only have one observed $\hat{\theta}$.

How does this make sense? $\hat{\theta}$ is a single value. As a simple example, let $\hat{\theta} = \bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$. The σ^2 here is the same as the variance as the individual y_i. So if we estimate σ^2, this gives us an estimate for $\hat{\theta}$.

Example 30. *Suppose we have data* $5.71, 3.24, 0.97, 1.10, 3.05$. *When we do bootstrap replications, we sample from these, e.g.*

$$\begin{aligned}\hat{\theta}_1^* &: 3.24, 1.10, 1.10, 3.05, 5.71\\ \hat{\theta}_2^* &: 0.97, 3.05, 3.24, 5.71, 3.05\\ &\vdots\\ \hat{\theta}_B^*\end{aligned}$$

where $B = 10^3$ *for example. Then we have*

$$\mathrm{Var}(\hat{\theta}) \approx \frac{1}{B}\sum_i (\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2$$

Real world

$$F \longrightarrow y_1, \dots, y_n \longrightarrow \hat{\theta}$$

Bootstrap world

$$\hat{F} \longrightarrow y_1^*, \dots, y_n^* \longrightarrow \hat{\theta}^*$$

In the real world, there are problems with the bias.

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}_F[\hat{\theta}] - \theta$$

but both F and θ are unknown. In the bootstrap world, we have

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}_{\hat{F}}[\hat{\theta}] - \hat{\theta}_{obs}$$

As mentioned before, you cannot always make the estimator unbiased by subtracting off the bias, because we do not always necessarily known it.

24.0.2. *Nonparametric Bootstrap.* $\hat{F}_n$ is the empirical distribution.

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^n I(Y_i \le y)$$

This essentially treats the sample as if it were the population, i.e. the bootstrap is generating iid samples from $\hat{F}_n$. By the strong law of large numbers,

$$\hat{F}_n(y) \to F(y) \text{ a.s.}$$

that is it converges pointwise. A stronger notion is that

$$\sup_y |\hat{F}_n - F(y)| \to 0 \text{ a.s.}$$

by Glivenko-Cantelli. However, this still relies on asymptotics. An even better notion of convergence is for all $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_y |\hat{F}_n(y) - F(y)| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}$$

(DKW inequality).

The bootstrap is a plugin estimator. We are interested in a functional $\theta = T(F)$. (The mean, variance, etc. are all examples of functionals). We estimate it as $\hat{\theta} = T(\hat{F})$.

$$\mu = \int_{-\infty}^{\infty} x \, \mathrm{d}F(x)$$

The plugin estimator is

$$\begin{aligned}\hat{\mu} &= \int_{-\infty}^{\infty} x \, \mathrm{d}\hat{F}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

24.0.3. *Parametric Bootstrap.* Suppose we have F_τ and we want to estimate τ as $\hat{\tau}$ and work with respect to $F_{\hat{\tau}}$ in the bootstrap world.

Example 31. $y_1, \ldots, y_n \sim \text{Pois}(\lambda)$,

$$\theta = \mathbb{P}(Y_j = 0) = e^{-\lambda}$$

Recall that the unbiased estimator is completely ridiculous here, so we use the MLE, which is $\hat{\theta} = e^{-\bar{y}}$*. What are the bias and variance of* $\hat{\theta}$*? The bootstrap approach would involve estimating* λ *by* $\bar{y}$ *and then generating samples from a* $\text{Pois}(\bar{y})$ *to estimate* λ*.*

- *Parametric Bootstrap*
- *Delta Method*

  $$g(\bar{y}) \approx g(\lambda) + (\bar{y} - \lambda)g'(\lambda)$$

  Then, if $g(y) = e^{-y}$*, then we get*

  $$e^{-\bar{y}} \approx e^{-\lambda} - (\bar{y} - \lambda)e^{-\lambda}$$

  So

  $$\text{Var}(e^{-\bar{y}}) \approx e^{-2\lambda}\frac{\lambda}{n}$$

  To get the bias, we have

  $$e^{-\bar{y}} \approx e^{-\lambda} - (\bar{y} - \lambda)e^{-\lambda} + \frac{1}{2}(\bar{y} - \lambda)^2 e^{-\lambda}$$

  The bias is simply $e^{-\bar{y}} - e^{-\lambda}$*. So we have*

  $$\text{Bias} \approx \frac{1}{2}e^{-\lambda}\frac{\lambda}{n}$$

  To get the exact bias, use the MGF.

Bootstrap Confidence Intervals are a mess! One of Efron's favourites is called the BCa. The simplest is called the percentile interval, which uses the quantiles of the bootstrap distribution $\hat{\theta}^*$.

24.0.4. *Bootstrap with Regression.*

$$y_i = X_i'\beta_i + \varepsilon_i$$

where we assume $\varepsilon_1, \ldots, \varepsilon_n$ are iid with mean 0. y_i are independent but not iid. Bootstrap resampling the y_i's is bad. We hope that we would be able to do bootstrap on the ε_i, but they are unobservable. An alternative is to do bootstrap on the residuals e_i.

24.1. **Iav's stuff.** The above examples show that context really matters!

Lets consider the regression example, consider $E[y|\mu] = \mu$, we are trying to predict μ given y. What we want is $E[\mu|y] \overset{???}{=} y$, however this isn't generally true, we can't even define this if we are using a one level model. Consider we have some data x and output y, we then use a linear model (after standardisation), then $y = mx$. Now what if we want to predict y from x? In maths we would say $x = 1/my$, however this is wrong, in stats we would use $x = my$ because m is the covariance $m = \rho\frac{\sigma_y}{\sigma_x}$. The ρ is the shrinkage towards the mean.

Consider the following graphical model:

$$A \to \mu \to y$$

Back to the simpler example: $y_j|\mu_j \sim (\mu_j, V)$ and $\mu_j \sim N(0, A)$. Let as assume A is known, then we can take the full Bayes approach to $\hat{\mu}_j = (1 - B)y_j$ and $B = V/(V + A)$. However, if A is unknown, let us have $A \sim h$ where h is a pdf on $(0, \infty)$, then $E[\mu_j|y] = E[E[\mu_j|y, A]|y] = E[(1 - B)y_j|y] = (1 - E[B|y]y_j)$. This is almost the same as when we know what A is known, except we just take the expectation, i.e. we estimate B.
Question: For what choice of B do we get the JS?
Answer: $A \sim unif(-V, \infty)$.
This gives us a way of improving it, for example let $A \sim unif(0, \infty)$, this is often called *Stein's harmonic prior*.

24.2. **Empirical Bayes.** In empirical bayes you use the data to estimate the hyper-parameters. Then use that for the prior. Clearly a lot of Bayesians don't like it, for point estimation it is a good approximation to the fully Bayesian approach.As before, let $y_j|\mu_j \sim (\mu_j, V)$ and $\mu_j \sim N(0, A)$ and $y \sim N(0, V + A)$. Then $S = \sum_i y_i^2 \sim (V + A)\chi_k^2$, then $E[S]/k = V + A$ so we could use $S/K - V$ to estimate A.

However, we could do this a smarter way, $S \sim (V + A)\chi_k^2$ we want to estimate $B = V/(V + A)$, so it makes sense to look at $1/S$ which has an inverted χ_k^2 then, using LOTUS,

$$E[1/S] = E[1/\chi_k^2] = \frac{1}{k-2} \quad \Rightarrow E\left[\frac{(k-2)V}{S}\right] = \frac{V}{V+A} = B$$

So let $\hat{B} = \frac{(k-2)V}{S}$ and use $\hat{\mu}_j = (1 - \hat{B})y_j$, well this is exactly the JS estimator, except last time we had $V = 1$.

25. Computational strategies

"Statistics is physics unleashed" C Morris.

25.1. **Rejection Sampling.** The goal is to simulate from some complicated distribution with PDF f, often we only have this up to a constant. The idea is we find some simpler distribution g that we have access to, for example a norma distribution. We then have the following method

Data: pdf f bounded by pdf g
Result: N exact draws from f
for *k in 1 to N* **do**
  Propose $x \sim g$;
  Accept with probability $\frac{f(x)}{g(x)}$;
end

Algorithm 1: Rejection Sampling

Lemma. *Probability is proportional to area.*

25.2. **Importance Sampling.** Very closely related method to rejection sampling, again we want draws from some distribution f, however we want these to approximate $E_f[h(X)]$ for some h. However, we only have draws from a distribution g. Assume for now that both f and g are properly normalised, for the discrete case

$$E_f[h(X)] = \sum_x h(x)f(x) = \sum_x \frac{h(x)f(x)g(x)}{g(x)} = E_g\left[\frac{h(x)f(x)}{g(x)}\right]$$

To make sure we have no 0 problems we assume $supp(f) \subset supp(g)$. We then generate $X_1, \ldots, X_n \sim g$ for n large and then use the law of large numbers to get

$$E_f[h(X)] \approx \frac{1}{n}\sum_j \frac{h(x_j)f(x_j)}{g(x_j)}$$

Problems: We wrote it as if we new the normalising constant, but in reality we don't. However, we can replace this with a biased version, let

$$W_j = \frac{f(X_j)}{g(X_j)} \quad E_g[W_j] = 1$$

and use

$$E_f[h(X)] \approx \frac{\sum_j h(X_j)W_j}{\sum_j W_j}$$

Asymptotically they will be the same, however it is also bias. The main advantage of this is that we don't need the normalise constant! This is a huge advantage!
Advice: Make the tails of g heavier then the tails of f.

Example 32 (Normal Example)**.** *Assume we want to generate from $f : N(0, 2)$ and we can generate $N(0, 1)$. (Clearly this is silly as we can just multiply by the square root of 2) Lets see what happens with importance sampling*

$$W = \frac{f(X)}{g(X)}$$

What is the variance?

$$var(w) = E[W^2] - 1$$

Then

$$E[W^2] = \int \frac{f(x)^2}{g(x)^2} g(x)dx = c \int \frac{e^{-x^2/2}}{e^{-x^2/2}} dx = \infty$$

Clearly this is terrible.

25.3. **Metropolis-Hastings.** Same set up as before, we want samples from $\pi(x)$ which we can't directly simulate from. As always we may not be able to compute the constant. Assume that we know how to run a Markov Chain on state space of interest $K(x, y)$. Often this is a simple random walk.

Data: pdf π chain $K(x, y)$, starting point x^0
Result: Draws from f
for *k in 1 to N* **do**
  Propose $x^k \sim K(x^{k-1}, x^k)$;
  Accept with probability $\min\left(\frac{\pi(x^k)K(x^k, x^{k-1})}{\pi(x^{k-1})K(x^{k-1}, x^k)}, 1\right)$;
end

Algorithm 2: Metropolis Hastings

Lemma. *If $s(x)K(x, y) = s(y)K(y, x)$ for all x, y where $s(x) \geq 0$ and $\sum_x s(x)$ then s is the stationary distribution for the chain K. Then K satisfies detailed balance.*

Theorem 16. *The above defined MH algorithm satisfies DB with respect to π stationary.*

Example 33 (Simple MH)**.** *Assume $K(x, y) = K(y, x)$ and try to find acceptance probability that will work. $x \neq y$*

$$P(x \to y) = \pi(x)K(x, y)a(x, y) = \pi(x)K(y, x)a(y, x)$$

Which tells us that

$$a(x, y) = \frac{\pi(y)}{\pi(x)} a(y, x) \leq \frac{\pi(y)}{\pi(x)}$$

Where $a(x, y)$ is the probability of accepting.

25.4. **Gibbs Sampler.** Consider the two dimensional version, where we want to sample pairs (x, y) from some bivariate distribution of interest. The basic idea is that we change one coordinate at a time from its conditional distribution given the rest.

Data: pdf $\pi(x|y)$, $\pi(y|x)$
Result: Draws from f
for *k in 1 to N* **do**
  Sample $x^t \sim \pi(x|y^{t-1})$;
  Sample $y^t \sim \pi(y|x^t)$;
end

Algorithm 3: Gibbs

You could also consider a the above where you randomly select a co-ordinate. Then we can relate it to the MH where the proposal is "pick a random coordinate and update it using its conditional distribution".

Example 34 (Data Augmentation - Chicken-egg problem)**.** *Chicken lays $N \sim Pois(\lambda)$ and $X|N \sim Bin(N,p)$ hatch and $Y = N - X$ doesn't hatch. $X \perp\!\!\!\perp Y$ and $X \sim Pois(\lambda p)$ and $Y \sim Pois(\lambda q)$. Now suppose we can only observe X and not N, but we want to estimate p and further assume that λ is known. Assume $p \sim beta(a,b)$ and then we can use Gibbs to compute this,*

$$\pi(p|X) \propto e^{-\lambda p}(\lambda p)^x p^{a-1}(1-p)^{b-1} \propto e^{-\lambda p} p^{a+x-1}(1-p)^{b-1}$$

Gibbs idea is to alternate

$$p|X,N \quad \textit{beta binomial conjugacy}$$

and

$$N|p,X$$

is also easy, then you just sample from each of these and it is a lot easier.

25.5. **Expectation Maximisation (EM).** Assume we want to minimise the log likelihood:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} l(\theta|Y)$$

EM tries to increase the dimensions to simplify it. For example you can always introduce latent variables. The idea, if we have $y \sim f(y|\theta)$ which is hard to use and we might introduce z such that $f(y,z|\theta)$ is easier to work with.

Example 35 (Mixture of normal densities)**.** *Assume $Y_1, \ldots, Y_n$ are iid from a 50-50 mixture of a $N(\mu_0,1), N(\mu_1,1)$. We want to estimate $\mu = (\mu_0,\mu_1)$. The likelihood is given by*

$$L(\mu_0,\mu_1|y) = \prod_{j=1}^{n}\left[\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(y_j-\mu_0)^2/2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(y_j-\mu_1)^2/2}\right]$$

This is a very nasty likelihood Let us introduce $Z_1, \ldots Z_n$ such that $Z_i = 1$ implies $Y_i \sim N(\mu_1,1)$ and $Z_i = 0$ implies that $Y_i \sim N(\mu_0,1)$. Then

$$f(y_j,z_j|\mu) = \phi(y_j-\mu_1)^{1-z_j}\phi(Y_j-\mu_0)^{z_j}$$

Where $\phi(x)$ is the standard normal then

$$\begin{aligned} l(\mu|y,z) &= \sum_j \log f(y_j,z_j|\mu) \\ &= \frac{-n}{2}\log 2\pi + \frac{1}{2}\sum_j (1-z_j)(y_j-\mu_0)^2 - \frac{1}{2}\sum_j z_j(y_j-\mu_1)^2 \end{aligned}$$

Let $\mu(t) = (\mu_0^{(t)}, \mu_1^{(t)})$ by the EM value at the t^{th} step and denote $E^t(z_j) = E_{\mu(t)}(z_j) = E_{\mu(t)}(Z_j|Y_j)$. Then the E-step is

$$Q(\mu|\mu(t)) = \frac{1}{2}\sum_j (1-E^t(z_j))(y_j-\mu_0)^2 - \frac{1}{2}\sum_j E^t(z_j)(y_j-\mu_1)^2$$

Then the M-step is

$$\frac{\partial Q}{\partial \mu_1} = 0 \Rightarrow \mu_1^{(t+1)} = \frac{\sum_j y_j E^t(z_j)}{\sum_j E^t(z_j)}$$

And similarly for μ_0 you get

$$\mu_-^{(t+1)} = \frac{\sum_j y_j(1-E^t(z_j))}{\sum_j (1-E^t(z_j))}$$

Then

$$\begin{aligned} E^t(z_j) = P(Z_j = 1|Y_j,\mu^{(t)}) &= \frac{f(y_j|Z_j=1,\mu^{(t)})P(Z_j=1)}{f(y_j|Z_j=1,\mu^{(t)})P(Z_j=1) + f(y_j|Z_j=0,\mu^{(t)})P(Z_j=0)} \\ &= \frac{\frac{1}{2}\phi(y_j-\mu_1^{(t)})}{\frac{1}{2}\phi(y_j-\mu_1^{(t)}) + \frac{1}{2}\phi(y_j-\mu_0^{(t)})} \end{aligned}$$

Example 36 (Logistic regression)**.** *$Y_1, \dots, Y_n$ observations and $X_1, \dots, X_n$ such that*

$$Y_i \sim Bern\left(\frac{e^{X_i'\beta}}{1+e^{X_i'\beta}}\right)$$

or could also consider probit Φ regression, the the likelihood function is

$$L(\beta|Y,X) = \prod_i \left[\Phi(X_i'\beta)^{Y_i}\Phi(1-X_i'\beta)^{1-Y_i}\right]$$

Example 37 (Faulty items)**.** *You are only inspecting items from an assembly line at an interval of $m = 10$. The process becomes faulty with probability p. After a fault the prob of a mistake is $0 < \pi \leq 1$. The idea is, supposed we are inspecting at intervals of 10. Suppose that the first 17 are defect free, however you don't pick it up at the 20th inspecting but you pick it up at 30. So from 0-30 you have a cycle, after this you start a new cycle. Notice that there is a cost to inspect, however there is also a cost of missing mistakes. Let T_i be the length of each cycle (i.e. the number produced in each cycle). Let C_i be the cost associated with every cycle. Then $T_i \sim f_{p,\pi}$. Notice that this basically becomes a Renewal process, when you tag along the costs you get a Renewal-Reward process. How do we find the optimum m? We do this by minimising the expected cost (or loss) per cycle. There is a nice theorem that says the long term expected loss will tend to*

$$\frac{E(C)}{E(T)}$$

Notice that we don't know when the process broke down, so the likelihood function is very nasty.

Some notation, we have observed $f(y|\theta)$ but this is incomplete, we actually have the complete data $f(y,z|\theta)$ which is unobserved.

$$f(y,z|\theta) = f(y|\theta)f(z|y,\theta)$$

Which gives us

$$f(y|\theta) = \frac{f(y,z|\theta)}{f(z|y,\theta)} \Rightarrow l(\theta|y) = l(\theta|y,z) - \log f(z|y,\theta)$$

Idea: Take the conditional expectation with respect to the density of Z given Y and $\theta^{(t)}$.

$$l(\theta|y) = \underbrace{E_{\theta^{(t)}}[l(\theta|y,z)|y]}_{=Q(\theta|\theta^{(t)})} - E_{\theta^{(t)}}[\log f(z|y,\theta)|y]$$

Step 1: Compute $Q(\theta|\theta^{(t)})$
Step 2: Fine $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$

Theorem 17 (Properties of EM)**.** *$l(\theta|Y)$. The sequence $\{\theta^{(t)}\}$ obtained using the EM gives us a increasing sequence of $\{l(\theta^{(t)}|y)\}$.*

Proof. It suffices to show (a)

$$E_{\theta^{(t+1)}}[l(\theta|y,z)|y] \geq E_{\theta^{(t)}}[l(\theta|y,z)|y]$$

well this is true by construction of steps 2! and (b)

$$E_{\theta^{(t)}}[\log f(z|y,\theta^{(t+1)})|y] \leq E_{\theta^{(t)}}[\log f(z|y,\theta^{(t)})|y]$$

Look at the KL divergence by taking the difference

$$\begin{aligned}
E_{\theta^{(t)}}[\log f(z|y,\theta^{(t+1)})|y] - E_{\theta^{(t)}}[\log f(z|y,\theta^{(t)})|y] &= E_{\theta^{(t)}} \log\left(\frac{f(z|y,\theta^{(t+1)})}{f(z|y,\theta^{(t)})}\right) \\
&\leq \log E_{\theta^{(t)}}\left(\frac{f(z|y,\theta^{(t)+1})}{f(z|y,\theta^{(t)})}\right) \\
&= \log \int \left(\frac{f(z|y,\theta^{(t+1)})}{f(z|y,\theta^{(t)})}\right) f(z|y,\theta^{(t)})dz \\
&= 0
\end{aligned}$$

□

Theorem 18. *(1)Suppose $Q(\theta|\theta^{(t)})$ is continues in both θ and $\theta^{(t)}$. Then all of the limit points of the sequence $\{\theta^{(t)}\}$ converges monotonically to $l^* = l(\theta^*|y)$ where θ^* is some stationary point.*
(2) Suppose $l(\theta|y)$ is unimodal with θ^* being the only stationary point and that $\frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)})$ is continues. Then for any EM sequence $\{\theta^{(t)}\}$ converges to θ^* which is the unique maximiser.