Statistical Inference Review Notes

Rui Qiu

June 3, 2018

1 Week 1 intro and generating random samples

point estimation, interval estimation, hypothesis testing (each with frequentist, bayesian, and non-parametric (frequentist) methods)

probability inverse transform approach

Note: For first 5 weeks' notes, some θ should be written as $\boldsymbol{\theta}$, notice they are vectors in fact.

2 Week 2 revision, estimators and their evaluations

Theorem 2.1. *Z follows standard normal, then $U = Z^2$ is a χ^2 distribution with 1 degree of freedom.*

Theorem 2.2. *$U_1, \ldots, U_n$ are independent and $U_i \sim \chi^2_1$ then $\sum_{i=1}^n U_i \sim \chi^2_n$*

Proof. Proof by MGF of sums of independent gamma. □

Remark 2.3. *χ^2 distribution is a gamma distribution with $\alpha = n/2, \lambda = 1/2$.*

Theorem 2.4. *$Z \sim N(0,1), U \sim \chi^2_n, Z \perp U$, then $T = Z/\sqrt{U/n} \sim t_n$*

Theorem 2.5. *$U \sim \chi^2_m, V \sim \chi^2_n, U \perp V$, then $W = \frac{U/m}{V/n} \sim F(m,n)$.*

Theorem 2.6. *$X_1, \ldots, X_n \sim iidN(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n), \bar{X} \perp S^2, (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$*

Theorem 2.7. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

Definition 2.8. *$\hat{\theta} = T(X_1, \ldots, X_n)$ is an **unbiased** estimator for θ if $E[T(\mathbf{X})] = \theta$. The **bias** of an estimator is defined as*

$$bias(\hat{\theta}) = E[T(\mathbf{X})] - \theta$$

Definition 2.9.

$$MSE = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + Bias(\hat{\theta})^2$$

Example 2.10. *$X_1, \ldots, X_n \sim iidN(\mu, \sigma^2)$, we have three estimators of σ^2:*

- *MLE est.* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, $MSE(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2}$
- *unbiased est.* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$,$MSE(S^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$
- *another est.* $\tilde{\sigma}^2 = \frac{1}{n+1}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n+1}S^2$, $MSE(\tilde{\sigma}^2) = \frac{2\sigma^4}{n+1}$
- $MSE(\tilde{\sigma}^2) < MSE(\hat{\sigma}^2) < MSE(S^2)$

Example 2.11. *$X_1, \ldots, X_n$ follows iid Bernoulli(p), The MLE $\hat{p} = \bar{X}$, bias is $bias(\hat{p}) = E[\hat{p}] - p = 0, V(\hat{p}) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$*

$$MSE(\hat{p}) = \frac{p(1-p)}{n}$$

A Bayesian estimator $\hat{p}_B = \frac{y+a}{a+b+n}$, the E and V of $\hat{p}_B$ are straightforward, then need to specify a, b to compare MSEs.

Definition 2.12. *An estimator $\hat{\theta}$ is **weakly consistent** if*

$$P(|\hat{\theta} - \theta| > \epsilon) \to 0, \ \ as\ n \to \infty\ \forall\epsilon > 0$$

Proof. By Chebyshev's inequality,

$$P(|\hat{\theta} - \theta| > \epsilon) \leq \frac{E[(\hat{\theta} - \theta)^2]}{\epsilon^2} = \frac{MSE(\hat{\theta})}{\epsilon^2} = \frac{1}{\epsilon^2}[V(\hat{\theta}) + bias(\hat{\theta})^2]$$

Thus when $V(\hat{\theta}) \to 0$ and $bias(\hat{\theta}) \to 0$ implies that $\hat{\theta}$ is consistent. □

Definition 2.13. *An estimator T^* is a **best unbiased estimator** of $\tau(\theta)$ if it satisfies $E[T^*] = \tau(\theta)$, $\forall\theta$ and for any other unbiased estimator T with $E[T] = \tau(\theta)$, T^* has the smallest variance:*

$$V(T^*) \leq V(T)\ \forall\theta$$

*T^* is called a **minimum variance unbiased estimator (MVUE)** for $\tau(\theta)$*

Definition 2.14. ***Cramer-Rao Inequality (lower bound):*** *$X_1, \ldots, X_n$ is a random sample from a distribution family with density function $f_X(x;\theta)$ where θ is a scalar parameter. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator for $\tau(\theta)$, then under certain regularity conditions:*

$$Var(T) \geq \frac{[\tau'(\theta)]^3 2}{ni(\theta)} = [\tau'(\theta)]^2 I(\theta)^{-1}$$

- $\tau'(\theta) = \frac{d}{d\theta}\tau(\theta)$
- *$I(\theta) = ni(\theta)$ is the **Expected Fisher Information***
- $I(\theta) = E[(\frac{\partial I(\theta)}{\partial\theta})^2] = -E[\frac{\partial^2 I(\theta)}{\partial\theta^2}]$

C-R Inequality Extended: *$X_1, \ldots, X_n$ be a sample (don't have to be iid) with pdf $f(\mathbf{x}|\theta)$, $T(\mathbf{X})$ be an estimator (don't have to be unbiased), then based on regularity conditions:*

$$V[T(\mathbf{X})] \geq \frac{[\frac{\partial}{\partial\theta}E[T(\mathbf{X})]]^2}{E[(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta))^2]} = \frac{[\frac{\partial}{\partial\theta}E[T(\mathbf{X})]]^2}{I(\theta)}$$

- *If unbiased* $E[T(\mathbf{X})] = \tau(\theta)$*,*

$$V[T(\mathbf{X})] \geq \frac{[\tau'(\theta)]^2}{I(\theta)}$$

- *If iid samples,*

$$V[T(\mathbf{X})] \geq \frac{[\tau'(\theta)]^2}{ni(\theta)}$$

Example 2.15. $X_1, \ldots, X_n$ *is an iid sample from Poisson(*λ*).* $\tau(\lambda) = \lambda$, $\frac{d}{d\lambda}\tau(\lambda) = 1$

$$\begin{aligned} i(\lambda) &= E[(\frac{d}{d\lambda}\ln(f(x|\lambda)))^2] \\ &= E[(\frac{X}{\lambda} - 1)^2] \\ &= \frac{1}{\lambda^2}E[(X-\lambda)^2] \\ &= \frac{1}{\lambda^2}Var(X) \\ &= \frac{1}{\lambda} \\ V_\lambda(T) &\geq \frac{1^2}{n \cdot \frac{1}{\lambda}} = \frac{\lambda}{n} \end{aligned} \tag{1}$$

This is the lower bound we need to achieve. Once achieved, we claim it is an MVUE.

Proof. The proof is based on the **Cauchy-Schwarz Inequality:** for two random variables Y, Z:

$$[Cov(Y,Z)]^2 \leq V(Y)V(Z) \implies V(Y) \geq \frac{[Cov(Y,Z)]^2}{V(Z)}$$

$$\ldots\ldots$$

□

Definition 2.16. *The* ***Fisher Information (expected Fisher Information, information number)*** *is:*

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

For iid data: $ni(\theta) = I(\theta)$

Summary:

- how to evaluate an estimator: biased or not, MSE small, if consistent
- consistent when variance and mse both close to 0
- MVUE
- C-R lower bound

3 Week 3 data reduction

Definition 3.1. ***Sufficiency Principle:*** *If $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through $T(X_1, \ldots, X_n)$.*

Definition 3.2. *A statistic $T(X_1, \ldots, X_n)$ is* ***sufficient*** *for θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $T(X_1, \ldots, X_n)$ does not depend on θ.*

In other words, what is sufficiency? When $P(X_1, \ldots, X_n | T(X_1, \ldots, X_n)$ does not depend on θ, then T is sufficient.

Theorem 3.3. ***The factorization theorem/criterion:*** *Suppose $X_1, \ldots, X_n$ form a random sample from $f(x; \theta)$, then $T(\mathbf{X})$ is a sufficient statistic for θ* ***iff*** *there exists two non-negative function K_1 and K_2 such that the likelihood function $L(\theta; \mathbf{x})$ can be written as:*

$$f(\mathbf{x}; \theta) = L(\theta; \mathbf{x}) = K_1[t(\mathbf{x}); \theta] K_2[\mathbf{x}]$$

Example 3.4. *Normally dist data: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. What are the sufficient statistic(s) when μ unknown, σ^2 known? When both unknown? When μ known, σ^2 unknown?*

When μ unknown, σ^2 known:

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) \cdots f(x_n) \\ &= \exp\left(\frac{\mu}{\sigma^2} \sum x_i - \frac{n}{2\sigma^2}\mu^2\right) \cdot (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\frac{1}{2\sigma^2} \sum x_i^2\right) \end{aligned} \tag{2}$$

So we have K_1 and K_2 respectively, and $t = \sum x_i$ for K_1 is the sufficient statistic.

Similarly when both μ, σ^2 are unknown, the equation above can be written as a product of a term with knowns only and constant only as:

$$f(x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum x_i^2 + \frac{\mu}{\sigma^2} \sum x_i - \frac{n}{2\sigma^2}\mu^2\right)$$

$K_1 = f(\mathbf{x}; \mu, \sigma^2), K_2 = 1, t_1 = \sum x_i^2, t_2 = \sum x_i$ are sufficient statistics.

4 Week 4 sufficiency continued

Definition 4.1. *A sufficient statistics $T(\mathbf{X})$ is called a* ***minimal sufficient statistic*** *if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$. (not easy to use def to find one though.)*

Lemma 4.2. *$f(\mathbf{x}; \theta)$ is the pdf of a sample $\mathbf{X}$. Suppose $T(\mathbf{X})$ such that for every two sample points $\mathbf{x}, \mathbf{y}$, the ratio:*

$$\frac{L(\theta; \mathbf{x})}{L(\theta; \mathbf{y})}$$

is constant as function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$, then $T(\mathbf{X})$ is a ***minimal sufficient statistic****.*

Example 4.3. *$X_1, \ldots, X_n \sim iidN(\mu, \sigma^2)$ with both μ, σ^2 unknown, let $\mathbf{x}, \mathbf{y}$ be two sample points, let $(\bar{x}, s_x^2)$ and $(\bar{y}, s_y^2)$ be the sample means and sample variances for the samples $\mathbf{x}, \mathbf{y}$.*

$$\frac{f(\mathbf{x}|\mu, \sigma^2)}{f(\mathbf{y}|\mu, \sigma^2)} = \exp([-n(\bar{x}^2 - \bar{y}^2) + 2n\mu(\bar{x} - \bar{y}) - (n-1)(s_x^2 - s_y^2)]/(2\sigma^2))$$

The ratio will not depend on μ, σ^2 iff $\bar{x} = \bar{y}, s_x^2 = s_y^2$.
Therefore $(\bar{X}, S^2)$ are minimally sufficient for μ, σ^2.

Minimal sufficient statistics are not unique.

Theorem 4.4. ***Rao-Blackwell Theorem:*** *W be any unbiased estimator of $\tau(\theta)$, T be a sufficient statistic for θ, define $\phi(T) = E[W|T]$, then*

$$E[\phi(T)] = \tau(\theta), V[\phi(T)] \leq V[W]$$

If we have unbiased estimator and condition it on a sufficient statistic, our new statistic $\phi(T)$ is unbiased, and has the same or smaller variance!

Definition 4.5. *Let $f_T(t;\theta)$ be a family of pdfs for a statistic $T(\mathbf{x})$. The family of probability distributions is **complete** if*

$$E[h(T)] = \int h(t) f_T(t) dt = 0$$

for all θ implies that

$$P(g(T) = 0) = 1$$

for all θ.

Lemma 4.6. ***Scheffe Theorem:*** *$X_1, \ldots, X_n$ is a random sample from a distribution with pdf $f(x;\theta)$. If $T = T(\mathbf{X})$ is a complete and sufficient statistic, and $\phi(T)$ is an unbiased estimator of $\tau(\theta)$, then $\phi(T)$ is the unique MVUE of $\tau(\theta)$.*

How to find MVUEs?
1. Find or construct a sufficient and complete statistic T.
2. Find an unbiased estimator W for $\tau(\theta)$.
3. Compute $\phi(T) = E[W|T]$, then $\phi(T)$ is the MVUE.
or Find a function $h(T)$ where $E[h(T)] = \tau(\theta)$ then $h(T)$ is the MVUE.

Example 4.7. *Approach 1: $X_1, \ldots, X_n \sim$ iid Bern(θ)*
$T = \sum_{i=1}^n X_i$ is a sufficient and complete statistic for θ.
Consider $W = X_1, E[W] = \theta$, W is unbiased.
Compute $\phi(T) = E[W|T]$.

$$
\begin{aligned}
E[W|T] &= P(X_1 = 1|T = t) \\
&= \frac{P(X_1 = 1, T = t)}{P(T = t)} \\
&= \frac{P(X_1 = 1, \sum_{i=1}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)} \\
&= \frac{P(X_1 = 1, \sum_{i=2}^n X_i = t-1)}{P(\sum_{i=1}^n X_i = t)} \\
&= \frac{\theta \times [\binom{n-1}{t-1}\theta^{t-1}(1-\theta)^{(n-1)-(t-1)}]}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{t}{n} \implies \frac{T}{n} = \bar{X}
\end{aligned} \tag{3}
$$

$\bar{X}$ is the MVUE of θ.

5 Week 5 sufficiency continued and other estimators

Definition 5.1. *A random variable belongs to the* ***k-parameter exponential family of distributions*** *if its pdf can be written in the following form*

$$f(x;\theta) = \exp\left(\sum_{j=1}^k A_j(\theta)B_j(x) + C(x) + D(\theta)\right)$$

or

$$f(x;\theta) = C^*(x)D^*(\theta)\exp\left(\sum_{j=1}^k A_j(\theta)B_j(x)\right)$$

Example 5.2. *Poisson* $f(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!} = \exp\{x\ln(\lambda) - \lambda - \ln(x!)\}$

$A_1(\lambda) = \ln(\lambda)$

$B_1(x) = x$

$C(x) = -\ln(x!)$

$D(\lambda) = -\lambda$

Another form, if we define $\phi = (\phi_1, \ldots, \phi_k) = A(\theta) = \{A_1(\theta), \ldots, A_k(\theta)\}$ then ϕ is referred to as the ***canonical parameter*** *for the exponential family and*

$$f(x;\theta) = \exp\left\{\sum_{j=1}^k \phi_i B_i(x) + C(x) + D(\phi)\right\}$$

$\theta = A^{-1}(\phi), D(\phi) = D\{A^{-1}(\phi)\}$

The canonical parameter for Poisson is $\phi = \ln(\lambda)$

Lemma 5.3. *If the usual regularity condition hold, then a vector of k sufficient statistics $\mathbf{T}$ exists for a vector or parameters θ iff the distribution of $\mathbf{X}$ belong to the k-parameter exponential family.*

Proof. $f(\mathbf{x};\theta) = \exp\left\{\sum_{j=1}^{k} A_j(\theta)\left(\sum_{i=1}^{n} B_j(x_i)\right) + nD(\theta) + \sum_{i=1}^{n} C(x_i)\right\}$

Let $\mathbf{t} = (\sum_{i=1}^{n} B_1(x_i), \ldots, \sum_{i=1}^{n} B_k(x_i))$, then $K_1 = \exp\{\sum_{j=1}^{k} A_j(\theta)t_j + nD(\theta)\}, K_2 = \exp\{\sum_{i=1}^{n} C(x_i)\}$ □

Lemma 5.4. *Under the same conditions of previous lemma,* $\mathbf{T}$ *is also minimal sufficient and complete.*

Definition 5.5. ***Method of Moments:*** *By equating the moments of a distribution to the sample moments, where the distributional moments are defined as:*

$$\mu_k = E_\theta(X^k)$$

and sample moments as:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,\ k = 1, \ldots, k$$

the estimator $T(\mathbf{X}) = \tilde{\theta}$ *is the solution to the system of* k *equating moments.*

Definition 5.6. ***Generalized method of moments:*** *based on some functions* $g_1(), \ldots, g_k()$

$$E_\theta(g_m(X)) = \frac{1}{n}\sum_{i=1}^{n} g_m(x_i),\ m = 1, \ldots, k$$

When $g_m(x) = x^m$*, then it is the standard method of moments.*

Definition 5.7. ***Maximum Likelihood Estimation:*** *We find the estimator* $\hat{\theta}$ *which maximizes the likelihood function*

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta)$$

- *If likelihood is differentiable in* θ_i*,* ***possible candidates*** *for the MLE are the values* $(\theta_1, \ldots, \theta_k)$ *that solve:*

$$\frac{\partial}{\partial\theta_i} L(\theta; \mathbf{x}) = 0,\ i = 1, \ldots, k$$

- *Possible: local vs global maximum, extrema may occur on the boundary thus the first derivative may not be* 0*,...*
- *Check the first derivative is* 0 *and the second derivative is less than* 0*.*

Don't forget to use **log-likelihood** for saving your life.

Example 5.8. *The likelihood of right censoring. Suppose* $X_1, \ldots, X_m$ *are observed,* $X_{m+1}, \ldots, X_n$ *are unobserved by time* T*, then*

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} f_X(x_i;\theta) \prod_{i=m+1}^{n} (1 - F_X(T;\theta)) \\ F_X(T) &= P(X \leq T) = \int_0^T \theta\exp(-\theta x)dx = 1 - \exp(-\theta T) \\ &\ldots \end{aligned} \tag{4}$$

General steps for multivariate (normal) distribution:

1. First-order partial derivatives (score equations) at $\hat{\theta}_1, \hat{\theta}_2$ are zero.

2. At least one second-order partial derivative is negative.

3. The determinant of the matrix of second order partial derivatives (the Hessian matrix) is positive.

$$\left.\begin{vmatrix} \frac{\partial}{\partial^2 \theta_1^2} & \frac{\partial}{\partial\theta_1 \partial\theta_2} \\ \frac{\partial}{\partial\theta_2 \partial\theta_1} & \frac{\partial}{\partial^2 \theta_2^2} \end{vmatrix}\right|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} > 0$$

Some computations about MLEs.

Definition 5.9. ***Newton-Raphson (N-R) Method in finding MLEs:***

- *N-R is a fast root finding approach, but sensitive to starting values.*
- *Consider log-likelihood $l(\theta|\mathbf{x})$, let $U(\theta) = l'(\theta|\mathbf{x})$ and $H(\theta) = l''(\theta|\mathbf{x})$. And θ_0 is the initial estimate of θ, $\hat{\theta}$ is the MLE.*
- *If we Taylor expanded $U(\theta)$ around θ_0:*

  $$U(\theta) = U(\theta_0) + (\theta - \theta_0)H(\theta_0) + \cdots$$

- *At $\theta = \hat{\theta}$, $U(\hat{\theta}) = 0$, so we have*

  $$0 = U(\theta_0) + (\hat{\theta} - \theta_0)H(\theta_0) + \cdots$$

- $$\begin{aligned} \hat{\theta} &= \theta_0 - H^{-1}(\theta_0)U(\theta_0) \\ \theta_1 &= \theta_0 - H^{-1}(\theta_0)U(\theta_0) \\ \theta_2 &= \theta_1 - H^{-1}(\theta_1)U(\theta_1) \\ &\cdots \end{aligned} \tag{5}$$

  until $|\theta_k - \theta_{k-1}| < \epsilon$ where ϵ is a threshold for convergence.

- *N-R can be extended to multivariate case, then $U(\theta)$ denotes the vector of the first partial derivatives of $l(\theta)$, $H(\theta)$ denotes the matrix of the second partial derivatives of $l(\theta)$.*

  $$\theta_{t+1} = \theta_t - H^{-1}(\theta_t)U(\theta_t)$$

Definition 5.10. ***Fisher Scoring:*** *a simple modification of N-R method, use the expectation to replace the Hessian $H(\theta)$ as*

$$E[H(\theta)] = -I(\theta)$$

*where $I(\theta)$ is **Fisher's information matrix***

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta_i}\right)\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta_j}\right)\right] = -E[\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta_i \partial\theta_j}] \\ \theta_{t+1} &= \theta_t + I^{-1}(\theta_t)U(\theta_t) \end{aligned} \tag{6}$$

Eliminates some possible convergence issues.

6 Week 6 MLE continued and EM

Definition 6.1. ***EM algorithm:*** *a general algorithm to find MLEs when some of the data are* ***missing****. Suppose we have data* $\mathbf{y} = \{y_1, \ldots, y_n\}$*, but we don't observe all* $\mathbf{y}$*'s, let* $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{miss})$

The EM seeks to maximize $l(\boldsymbol{\theta}; \mathbf{y}_{miss})$ *with respect to* $\boldsymbol{\theta}$*.*

1. ***E-Step:*** *Calculate the expectation of the complete likelihood conditional on the observed data and the current value of* $\boldsymbol{\theta}$*.*

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= E\left\{l_{comp}(\boldsymbol{\theta}; \mathbf{y}_{obs}, \mathbf{y}_{miss})|\boldsymbol{\theta}^{(r)}, \mathbf{y}_{obs}\right\} \\ &= \int [l_{comp}(\boldsymbol{\theta}; \mathbf{y}_{obs}, \mathbf{y}_{miss})] k(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \boldsymbol{\theta}) d\mathbf{y}_{miss} \end{aligned} \tag{7}$$

2. ***M-step:*** *Maximize* $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ *w.r.t.* $\boldsymbol{\theta}$*. Set* $\boldsymbol{\theta}^{(t+1)}$ *equal to the maximizer of* Q*.*

3. *Return to the E-step unless a stopping criterion has been reached.*

Lemma 6.2. ***Invariance property of MLEs:*** *Suppose* $\boldsymbol{\theta}$ *and* $\boldsymbol{\eta}$ *are two alternative parameterizations for some probability distribution and* ***eta*** *is a one-to-one function of* $\boldsymbol{\theta}$*, can write* $\boldsymbol{\eta} = \boldsymbol{g(\theta)}, \boldsymbol{\theta} = \boldsymbol{h(\eta)}$ *for some functions* $\boldsymbol{g(\cdot)}, \boldsymbol{h(\cdot)}$*.*

If $\hat{\boldsymbol{\theta}}$ *is the MLE of* $\boldsymbol{\theta}$ *then* $\hat{\boldsymbol{\eta}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$ *is the MLE for* $\boldsymbol{\eta}$*.*

If the mapping is 1-1 note $\eta = \tau(\theta) \implies \tau^{-1}(\eta) = \theta$*.*

7 Week 7 MLE asymptotics and hypothesis testing

Lemma 7.1. $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ *and let* $\hat{\theta}$ *be the MLE of* θ*. Under regularity conditions of* $f(x;\theta)$ *and thus* $L(\theta; \mathbf{x})$ *we can state:*

$$W = \frac{1}{\sqrt{n}} l'(\theta; \mathbf{x}) \xrightarrow{D} N(0, i(\theta))$$

Lemma 7.2. $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$*,* $\hat{\theta}$ *MLE of* θ*. Under regularity conditions of* $f(x;\theta)$ *and thus* $L(\theta; \mathbf{x})$*, we have*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, i(\theta)^{-1})$$

Theorem 7.3. ***Delta Method:*** *Let* Y_n *be a sequence of random variables such that*

$$\sqrt{n}(Y_n - \theta) \xrightarrow{D} N(0, \sigma^2)$$

Suppose a given function g *and a specific value* θ *suppose that* $g'(\theta)$ *exists and is not* 0*, then*

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2)$$

Lemma 7.4. $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$*,* $\hat{\theta}$ *is MLE of* θ*,* $\tau(\theta)$ *is a continuous function. Under regularity conditions (smoothness conditions) of* $f(x;\theta)$*, thus* $L(\theta; \mathbf{x})$ *we have:*

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{D} N(0, \nu(\theta))$$

where $\nu(\theta) = \frac{[\tau'(\theta)]^2}{i(\theta)}$ *is the C-R lower bound for a single data point.*
Or

$$\tau(\hat{\theta}) \sim N\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{\boldsymbol{I}(\theta)}\right)$$

from Delta method.

Asymptotically, MLEs are:

1. unbiased
2. achieve the C-R lower bound (efficient)
3. asymptotically normally distributed

Let's move from point estimation to hypothesis testing.

Definition 7.5. *Suppose $X_1, \ldots, X_n$ represent a simple random sample from a parametric family with $f(x;\theta)$ for some parameter $\theta \in \Theta$*

- *A statistical hypothesis is a subset of the parameter space Θ.*
- *Any statistical hypothesis of interest, often as **null hypothesis** is associated with a competing **alternative hypothesis**.*
- *A null hypo and its alt. hypo form a partition of the parameter space Θ consisting of the sets Θ_0 and*

$$\Theta_1 = \Theta_0^c \cap \Theta$$

Definition 7.6. *A **hypothesis testing procedure** or **hypothesis test** is a rule that specifies:*

1. *For which sample values the decision is made to accept H_0.*
2. *For which sample values H_0 is rejected and H_1 is accepted as true.*

*The subset of the space for which H_0 will be rejected is **rejection region** or **critical region** C. The complement of the rejection region is **acceptance region.***

Definition 7.7. ***Type I error:*** *Reject H_0 give that it is true. Thus the observations **fall in the rejection region** C when in fact that null hypo H_0 is true.*

Type II error: *Do not reject H_0 when it is false: thus the observation values **fall outside the rejection region** when in fact the null hypo is false.*

Definition 7.8. *The probability of a Type I error, α in a test of hypotheses is called the **size** or **significance level** of the test. The complement of the probability of a Type II error*

$$\eta(\theta) = 1 - \beta,$$

*is the **power** of the test.*

$$Power = 1 - P(Type - II - Error) = 1 - P(\mathbf{X} \in C^c | H_1) = P(\mathbf{X} \in C | H_1)$$

*i.e. **power** is the probability when H_1 is true we reject H_0. (which is the right thing to do here)*

Remark 7.9. ***Essential Nature of a Hypothesis Test:*** *Given H_0, H_1, data $\mathbf{x} = \{x_1, \ldots, x_n\}$.*

1. *Compute a relevant test statistic $T(\mathbf{x})$, the test statistic $T(\mathbf{x})$ should be chosen s.t. it can differentiate between H_0 and H_1 in ways they are scientifically relevant. Typically $T(\mathbf{x})$ is chosen so that $T(\mathbf{x})$ is probably small under H_0 but large under H_1.*

2. *Obtain a null distribution: a probability distribution over the possible outcomes of $T(\mathbf{X})$ under H_0. Here $\mathbf{X} = \{X_1, \ldots, X_n\}$ are potential experimental results that could have happened under H_0.*

3. *Compute the p-value: the probability under H_0 of observing a test statistic $T(\mathbf{X})$ as or more extreme than the observed statistic $t(\mathbf{x})$.*

$$p - value = P(T(\mathbf{X}) \geq t(\mathbf{x}|H_0)$$

4. *If p-value is small, evidence against H_0; if p-value is large, not evidence against H_0.*

Example 7.10. *$X_1, \ldots, X_n$ are a random sample from $N(\mu, 1)$, consider testing*

$$H_0 : \mu \leq \mu_0 \ vs \ H_1 : \mu > \mu_0$$

$C = \left\{ \frac{\bar{X} - \mu_0}{1/\sqrt{n}} \geq k \right\}$, *so* $T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{1/\sqrt{n}}$

$$p - value = P(\frac{\bar{X} - \mu_0}{1/\sqrt{n}} - \frac{\bar{x} - \mu_0}{1/\sqrt{n}}) = P(Z \geq \frac{\bar{x} - \mu_0}{1/\sqrt{n}})$$

The probability under H_0 of getting the observed test statistic or something more extreme based on the rejection region.

Definition 7.11. ***Neyman-Pearson Set-up:*** *Consider simple hypotheses – only consist of single parameter value, suppose $X_1, \ldots, X_n$ are a sample from a population with density function $f(x; \theta)$ for $\theta \in \Theta$ where $\Theta = \{\theta_0, \theta_1\}$, we focus on $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.*

Consider the likelihood ratio:

$$\lambda(\mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})},$$

the test shall define has a critical region of the form $C = \{\lambda(\mathbf{x}) \leq k\}$

The ratio of likelihood for any given sample at each of the two possible parameter values is precisely a relative measure of how plausible these two are.

If $\lambda(\mathbf{x})$ is small, strong evidence the observations arose from the alt. hypothesis rather than the null. (We favour the one with higher likelihood, it is more possible!)

*We know for a given α we could compare the power $\eta(\theta)$, would like to find a **uniformly most powerful** test s.t.*

$$\eta(\theta) \geq \eta(\theta^*),$$

N-P tests lead to UMP tests.

After comparing two single-value hypos, how to find the most power one?

Example 7.12. *$X_1, \ldots, X_n$ random sample from $N(\mu, 1)$, suppose we know $\mu \in \{0, 1\}$, wish to test $H_0 : \mu = 0$ vs $H_1 : \mu = 1$.*

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - 1)^2\right)} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n \left[X_i^2 - (X_i - 1)^2\right]\right) \\
&= \exp\left(\frac{n}{2} - \sum_{i=1}^n X_i\right) \\
C &= \left\{\exp\left(\frac{n}{2} - \sum_{i=1}^n X_i\right) \leq k\right\} \\
&= \left\{\frac{n}{2} - \sum_{i=1}^n X_i \leq \log(k)\right\} \\
&= \left\{-\sum_{i=1}^n X_i \leq \log(k) - \frac{n}{2}\right\} \\
&= \left\{\sum_{i=1}^n X_i \geq -\log(k) + \frac{n}{2}\right\} \\
&= \left\{\bar{X} \geq -\log(k)/n + \frac{1}{2}\right\} \\
&= \left\{\bar{X} \geq k^*\right\} \\
P_{H_0}(C) &= P_{H_0}(\bar{X} \geq k^*) = \alpha \\
&= P_{H_0}\left(\frac{\bar{X} - 0}{1/\sqrt{n}} \geq k^{**}\right) = \alpha \\
&= P_{H_0}(Z \geq k^{**}) = \alpha
\end{aligned} \tag{8}$$

If $\alpha = 0.05$ then we can calculate that c^{**} is 1.644854.

Lemma 7.13. ***Neyman-Pearson Lemma:*** *Suppose H_0 and H_1 are simple hypos and that the test reject H_0 whenever the likelihood ratio is less than k has significance level α.*

Then any other test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.

Now we push a little bit from N-P lemma with not-so-simple parameter value.

Example 7.14. *Suppose $X_1, \dots, X_n$ from $N(\mu, 1)$, test $H_0 : \mu = \mu_0 = 0$ vs $H_1 : \mu = \mu_1 > 0$*

$$
\begin{aligned}
\lambda(\mathbf{x}) &= \exp\left(\frac{n\mu_1^2}{2} - n\mu_1\bar{X}\right) \\
C &= \left\{\exp\left(\frac{n\mu_1^2}{2} - n\mu_1\bar{X}\right) \le k\right\} \\
&= \left\{\bar{X} > \frac{\mu_1}{2} - \frac{1}{n\mu_1}\log(k)\right\} \\
&= \left\{\bar{X} > k^*\right\} \\
&= \left\{\frac{\bar{X} - 0}{1/\sqrt{n}} \ge k^{**}\right\} = \{Z \ge k^{**}\}
\end{aligned} \tag{9}
$$

If $\alpha = 0.05$, then $k^{**} = 1.64$.

The UMP test has the same rejection as our previous example. This test is actually UMP for $H_0 : \mu = 0$ vs $H_1 : \mu > 0$. Can also be shown that the test is UMP for $H_0 : \mu \le 0$ vs $H_1 : \mu > 0$.

Maximum Likelihood Ratio Test is not UMP. (later)

Actually, we can compute the powers of those two previous tests:

1. $H_0 : \mu = 0$ vs $H_1 : \mu > 0$: $\eta(\mu) = P\left(\frac{\bar{X}-\mu_0}{1/\sqrt{n}} \ge 1.64\right) = \dots$
2. $H_0 : \mu = 0$ vs $H_1 : \mu < 0$: $\eta(\mu) = P\left(\frac{\bar{X}-\mu_0}{q/\sqrt{n}} \le -1.64\right) = \dots$

8 Week 8 MLRT

Definition 8.1. *The **likelihood ratio test** for testing $H_0 : \theta \in \omega$ vs $H_1 : \theta \in \Omega - \omega$.*

$$
\lambda(\mathbf{x}) = \frac{\max_{\Theta\in\omega} L(\theta; \mathbf{x})}{\max_{\Theta\in\Omega} L(\theta; \mathbf{x})}
$$

Above is a restricted maximization, below is an unrestricted maximization. Construct a test of the form:

$$
C = \{\mathbf{x} : \lambda(\mathbf{x}) \le k\}
$$

$0 \le \lambda \le 1$, $\lambda \to 1$ if H_0 is true, $0 \le k \le 1$.

Example 8.2. *$X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$, test $H_0 : \theta = 0\theta_0$ vs $H_1 : \theta \ne \theta_0$.*

$\hat{\theta} = \bar{X}$

$$
\begin{aligned}
\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp[-\sum(x_i - \theta_0)^2/2]}{(2\pi)^{-n/2}\exp[-\sum(x_i - \bar{x})^2/2]} \\
&= \exp[-n(\bar{x} - \theta_0)^2/2]
\end{aligned} \tag{10}
$$

$$\begin{aligned}
C &= \{\lambda(\mathbf{x}) \leq k\} \\
&= \left\{(\bar{x}-\theta_0)^2 > [-2\log(k)]/n\right\} \\
&\implies \left\{|\bar{x}-\theta_0| > \sqrt{[-2\log(k)]/n}\right\} \\
&= \left\{\frac{|\bar{x}-\theta_0|}{1/\sqrt{n}} > \frac{\sqrt{[-2\log(k)]/n}}{1/\sqrt{n}}\right\} \\
&= \left\{|Z| > \frac{\sqrt{[-2\log(k)]/n}}{1/\sqrt{n}}\right\} \\
&= \{|Z| > k^*\}
\end{aligned} \tag{11}$$

$$\begin{aligned}
P(|Z| > k^*) &= P(Z > k^*) + P(Z < -k^*) = \alpha \\
&= 2P(Z < -k^*) = \alpha \\
&= P(Z < -k^*) = \alpha/2 \\
&= P(Z < k^{**}) = \alpha/2
\end{aligned} \tag{12}$$

Theorem 8.3. ***LRT asymptotics theorem:*** *For testing $H_0 : \theta \in \omega$ vs $H_1 : \theta \in \Omega - \omega$, $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$ and $\hat{\theta}$ is the MLE of θ and pdf under regularity conditions. Then under H_0, as $n \to \infty$,*

$$-2\log[\lambda(\mathbf{x})] \overset{D}{\to} \chi^2_1$$

If we reject when $\{\lambda \leq k\}$ then we reject when $\{-2\log(\lambda) > -2\log(k)\} = \{-2\log(\lambda) > k^*\}$, then use this to find k^*:

$$P(-2\log(\lambda) > k^*) = 0.05$$

Theorem 8.4. ***Theorem A (extended):*** *can be extended for more parameters as*

$$-2\log(\lambda) \overset{D}{\to} \chi^2_\nu$$

where ν is the number of constraints set in H_0.

Another way: p be the number of parameters estimated under H_1, p_0 be the number of parameters estimated under H_0, then $\nu = p - p_0$.

9 Week 9 MLRT, other tests, and interval estimation

Computational method: find λ first, then randomly generate samples with parameter from H_0 and set $P(\lambda_{H_0} \leq k) = 0.05$, use quantile function to find k.

The MLRT

- is asymptotically **most powerful unbiased**
- is asymptotically **similar**

- is asymptotically **efficient**

Definition 9.1. *Test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. A test of size α is **unbiased** if $\eta(\boldsymbol{\theta}) \geq \alpha$ for all $\boldsymbol{\theta} \in \Theta_1$*

Definition 9.2. *Test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. A test of size α is **similar** if $\eta(\boldsymbol{\theta}) = \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$*

Definition 9.3. *Test two simple hypotheses H_0 vs H_1, if n_1 and n_2 are the minimum possible sample sizes for test 1 and 2 for which we can achieve a size α and power$\geq \eta$, then the **relative efficiency** of test 1 compared to test 2 is n_2/n_1*

Definition 9.4. ***The Score Test:***

$$\boldsymbol{u(\theta)} = \left(\frac{\partial l}{\partial \theta_1}, \frac{\partial l}{\partial \theta_2}, \ldots, \frac{\partial l}{\partial \theta_k}\right)^T$$

Test $H_0 : \boldsymbol{\theta = \theta_0}$ vs $H_1 : \Omega - \{\boldsymbol{\theta_0}\}$.
Test statistic

$$\boldsymbol{u(\theta)^T I_{\theta_0}^{-1} u(\theta)} \sim \chi^2_{df=k}$$

Definition 9.5. ***The Wald Test:*** *Also testing..., test statistic:*

$$\boldsymbol{(\hat{\theta} - \theta_0)^T I_{\hat{\theta}} (\hat{\theta} - \theta_0)} \sim \chi^2_{df=k}$$

The MLRT, the Score Test and the Wald Test are asymptotically equivalent.

Definition 9.6. *Suppose that $\{f(x; \boldsymbol{\theta}); \theta \in \Omega\}$ defines a family of distributions, if $S_{\mathbf{x}}$ is a subset of Ω, depending on $\mathbf{X}$, s.t.*

$$P(\mathbf{X} : \boldsymbol{\theta} \subset S_{\mathbf{X}}) = 1 - \alpha$$

*then $S_{\mathbf{X}}$ is a **confidence set** for $\boldsymbol{\theta}$ with **confidence coefficient** $1 - \alpha$.*

Strong relationship between hypothesis testing and interval estimation. Every confidence set corresponds to a test and vice versa.
What is the relationship?

Example 9.7. *$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 unknown, test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$,*

$$C = \left\{ |\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| \geq z_{\alpha/2} \right\}$$

We know that under H_0, $P(C) = \alpha$, then the probability that H_0 is accepted is $1 - \alpha$:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Then we fix α and solve for μ:

$$P(\bar{X} - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n})) = 1 - \alpha$$

Then we have a $100(1 - \alpha)\%$ confidence estimator for μ.

Lemma 9.8. *Suppose that $\bar{C}(\theta_0)$ is the acceptance region for a test of size α, test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in \Omega - \theta_0$.*

*Then a **confidence set** for θ with **confidence coefficient** $(1-\alpha)$, is given by*

$$S_{\mathbf{x}} = \{\theta_0 : \mathbf{x} \in \bar{C}(\theta_0)\}$$

Example 9.9. *Suppose that 2.6, 1.2 and 4.9 are a random sample from a normal distribution whose mean is zero and whose variance σ^2 is unknown. Derive and compute a central 99% confidence interval for σ^2.*

Approach 1:

$$\begin{aligned}
\left(\frac{X_i - \mu}{\sigma}\right)^2 &= \left(\frac{X_i}{\sigma}\right)^2 \sim Z^2 = \chi^2_1 \\
\sum_{i=1}^{3} \left(\frac{X_i}{\sigma}\right)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{3} X_i^2 \sim \chi^2_3 \\
Y &= \sum_{i=1}^{3} X_i^2 \\
P\left(\chi^2_{\alpha/2,3} \le \frac{Y}{\sigma^2} \le \chi^2_{1-\alpha/2,3}\right) &= 1-\alpha \\
P\left(\frac{1}{\chi^2_{\alpha/2,3}} \ge \frac{\sigma^2}{Y} \ge \frac{1}{\chi^2_{1-\alpha/2,3}}\right) &= 1-\alpha \\
P\left(\frac{Y}{\chi^2_{1-\alpha/2,3}} \le \sigma^2 \le \frac{Y}{\chi^2_{\alpha/2,3}}\right) &= 1-\alpha
\end{aligned} \tag{13}$$

Approach 2:
Know that

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

so

$$\begin{aligned}
P\left(\chi^2_{\alpha/2,2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2,2}\right) &= 1-\alpha \\
P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2,\alpha}}\right) &= 1-\alpha
\end{aligned} \tag{14}$$

Approach 3:

$$\begin{aligned}
\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = \left(\frac{\bar{X}}{\sigma/\sqrt{n}}\right)^2 &= \frac{n\bar{X}^2}{\sigma^2} = Z^2 \sim \chi^2_1 \\
P\left(\chi^2_{\alpha/2,1} \le \frac{n\bar{X}^2}{\sigma^2} \le \chi^2_{1-\alpha/2,1}\right) &= 1-\alpha \\
P\left(\frac{n\bar{X}^2}{\chi^2_{1-\alpha/2,1}} \le \sigma^2 \le \frac{n\bar{X}^2}{\chi^2_{\alpha/2,1}}\right) &= 1-\alpha
\end{aligned} \tag{15}$$

Definition 9.10. *A random variable $g(\mathbf{X}, \boldsymbol{\theta})$ is a **pivotal quantity** (or a pivot) if the distribution of $g(\mathbf{X}, \boldsymbol{\theta})$ is independent of all parameters.*

If θ is a scalar, some definitions require that $g()$ be a monotonic function of θ.

Basic idea, the known distribution of a pivot quantity $g(\mathbf{X}, \theta)$ can be used to write a probability statement:

$$P(g_1 \leq g(\mathbf{X}, \theta) \leq g_2) = 1 - \alpha,$$

solve for θ (if g is monotonic then it's even easier)

$$P(\theta_1(\mathbf{X}) \leq \theta \leq \theta_2(\mathbf{X})) = 1 - \alpha$$

$$\hat{\theta} \sim N(\theta, I(\theta)^{-1})$$

$$\frac{\hat{\theta} - \theta}{1/\sqrt{I(\theta)}} \sim N(0, 1)$$

We have a pivotal quantity, asymptotic $100(1-\alpha)\%$ ci as:

$$\left[\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}}\right]$$

If we are interested in a function of θ say $\tau(\theta)$ we have:

$$\tau(\hat{\theta}) \sim N\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{I(\theta)}\right)$$

$$\frac{\tau(\hat{\theta}) - \tau(\theta)}{\sqrt{\frac{[\tau'(\theta)]^2}{I(\theta)}}} \sim N(0, 1)$$

Construct an asymptotic $100(1-\alpha)\%$ ci as:

$$\left[\tau(\hat{\theta}) - z_{\alpha/2}\frac{\tau'(\hat{\theta})}{\sqrt{I(\hat{\theta})}}, \tau(\hat{\theta}) + z_{\alpha/2}\frac{\tau'(\hat{\theta})}{\sqrt{I(\hat{\theta})}}\right]$$

Definition 9.11. ***Interval estimation - CDF method:*** *Pivoting the CDF, a pivot g leads to a confidence set:*

$$S_{\mathbf{X}} = \{\theta_0 : a \leq g(\mathbf{X}; \theta_0) \leq b\}$$

If for every $\mathbf{x}$ the pivot is a monotone function fo θ then the confidence set $C(\mathbf{x})$ is guaranteed to be an interval.

Theorem 9.12. *T be a statistic with a continuous cdf $F_T(t; \theta)$, let $\alpha_1 + \alpha_2 = \alpha, 0 < \alpha < 1$. Suppose that for each $t \in T$ the function $\theta_L(t)$ and $\theta_U(t)$ can be defined as:*

1. If $F_T(t; \theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by:

$$F_T(t;\theta_U(t)) = \alpha_1, F_T(t;\theta_L(t)) = 1-\alpha_2$$

2. If $F_T(t;\theta)$ is increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by:

$$F_T(t;\theta_L(t)) = \alpha_1, F_T(t;\theta_U(t)) = 1-\alpha_2$$

Then the interval $[\theta_L(t), \theta_U(t)]$ is a $1-\alpha$ confidence interval for θ.

10 Week 10 Bayesian

The key distinction between Bayesian and sampling theory stats is the issue of what is to be regarded as random and what is to be regarded as fixed.

- bayesian: parameters are random, and data once observed are fixed.
- sampling theorist: data are random even after observed, parameters are fixed.

Definition 10.1. *Bayesian inference is simply the application of Bayes' Rule to infer about parameters:*

T(...) is sufficient for θ iff posterior of θ given (...) is ~~a~~ the same as posterior dist of θ given T.

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \\ &\propto f(\mathbf{x}|\theta)\pi(\theta)\end{aligned} \tag{16}$$

- $\pi(\theta|\mathbf{x})$ *is the* ***posterior distribution*** *for θ*
- $f(\mathbf{x}|\theta)$ *is the* ***joint sampling distribution*** *for* $\mathbf{X}$ *or the* ***likelihood*** *for θ*
- $\pi(\theta)$ *is the* ***prior distribution*** *for θ*
- $m(\mathbf{x})$ *is the* ***marginal distribution*** *for* $\mathbf{X}$.

$L(\theta)$
$= K_1[t|\theta]K_2[\bar{X}]$

$$p(\theta|x_1,\ldots,x_n) \propto p(\mathbf{x}|\theta)p(\theta)$$

As $\pi(\theta|\mathbf{x})$ is a pdf, consider a number of summaries for θ after we observe the data: posterior mode, posterior median and posterior mean.

The posterior mean of θ:

$$\hat{\theta}_B = E[\theta|\mathbf{x}] = \int_\Theta \theta\pi(\theta|\mathbf{x})d\theta$$

Posterior mean of function of θ, $\tau(\theta)$:

$$\tau(\hat{\theta})_B = E[\tau(\theta)|\mathbf{x}] = \int_\Theta \tau(\theta)\pi(\theta|\mathbf{x})d\theta$$

Definition 10.2. *Let F denote the class of pdfs $f(x|\theta)$, a class P of prior distributions is a **conjugate family** for F if the posterior distribution is in the class P for all $f \in F$, all priors in P and all $x \in X$.*

In other words, if the prior distribution is the same family of distributions as the posterior, then it is a conjugate prior distribution.

For $f(\mathbf{x}|\theta)$ examine the kernel with regard to θ and see if your recognize the distribution. This will be the conjugate distribution.

Example 10.3. *$X_1, \ldots, X_n$ follow Poisson(λ), determine a conjugate prior distribution and then determine the posterior distribution for λ.*

$$f(\mathbf{x}|\lambda) = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!} \implies \lambda^{\sum x_i} e^{-n\lambda}$$

This is a kernel for a gamma distribution, thus we have the following conjugate prior: $\lambda \sim gamma(a, b)$

$$\begin{aligned} p(\lambda|\mathbf{x}) &\propto p(\mathbf{x}|\lambda)p(\lambda) \\ &\propto [\lambda^{\sum x_i} e^{-n\lambda}][\lambda^{a-1} e^{-\lambda/b}] \\ &= \lambda^{\sum x_i + a - 1} e^{-\lambda(n+1/b)} \\ &= \lambda^{a^*-1} e^{-\lambda/b^*} \\ [\lambda|\mathbf{x}] &\sim gamma(a^*, b^*) \\ a^* &= \sum x_i + a \\ b^* &= \frac{b}{bn+1} \end{aligned} \tag{17}$$

Definition 10.4. ***Bayesian testing:*** *$H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. Not like classical statistician, Bayesians consider θ to be random and it's quite natural to consider:*

$P(\theta \in \Theta_0|\mathbf{x})$ vs $P(\theta \in \Theta_1|\mathbf{x})$

One approach to Bayesian testing is to reject H_0 if

$$P(\theta \in \Theta_1|\mathbf{x}) > P(\theta \in \Theta_0|\mathbf{x})$$

Example 10.5. *$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Let $\theta \sim N(\mu, \tau^2)$, where σ^2, μ, τ^2 are known. Consider testing:*

$$H_0 : \theta \leq \theta_0 \text{ vs } H_1 : \theta > \theta_0$$

$$[\theta|\mathbf{x}] \sim N\left(\frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

To compare we simple check:

$$\int_{-\infty}^{\theta_0} p(\theta|\mathbf{x})d\theta \text{ vs } \int_{\theta_0}^{\infty} p(\theta|\mathbf{x})d\theta$$

We cannot test a single point such as θ_0 since it is a continuous distribution.

Approach 1: If the scientific question concerns whether a parameter can be exactly θ_0 or not, we should model as:

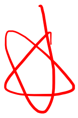
$$\theta \sim p\mathbf{1}_{\theta=\theta_0} + (1-p)N(\mu,\tau^2), 0 \le p \le 1.$$

*Approach 2: take $\beta = 0$ as **Bayes factors**. Rephrase hypothesis testing as choosing between competing models:*

Model 1 (M_1): $y_i = \alpha + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0,\sigma^2), \theta_1 = \{\alpha, \sigma^2\}$

Model 2 (M_2): $y_i = \alpha + \beta\mathbf{x}_i + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0,\sigma^2), \theta_2 = \{\alpha, \beta, \sigma^2\}$

The posterior probability for a given model i:



$$\begin{aligned}
\pi(M_i|\mathbf{x}) &= \frac{f(\mathbf{x}|M_i)\pi(M_i)}{m(\mathbf{x})} \\
f(\mathbf{x}|M_i) &= \int_\theta f(\mathbf{x}|M_i)\pi(\theta|M_i)d\theta_i \\
m(\mathbf{x}) &= \sum_{i=1}^{2} f(\mathbf{x}|M_i)\pi(M_i) \qquad (18)\\
\frac{\pi(M_2|\mathbf{x})}{\pi(M_1|\mathbf{x})} &= \frac{f(\mathbf{x}|M_2)}{f(\mathbf{x}|M_1)} \times \frac{\pi(M_2)}{\pi(M_1)} \\
&= BF(M_2;M_1) \times \frac{\pi(M_2)}{\pi(M_1)}
\end{aligned}$$

*$BF(M_2;M_1)$ is called **the Bayes factor***

Empirically, if BF is from 1 to 3.2, not worth more than a bare mention that evidence against model 1 (H_0);

if $3.2 < BF < 10$, substantial evidence

if $10 < BF < 100$, strong evidence

if $BF > 100$, decisive evidence against model 1.

Bayesian Interval Estimation

To obtain an interval, simply consider

$$P\pi(\theta|x)(C) = \int_C \pi(\theta|\mathbf{x})d\theta = 1-\alpha$$

3 main choices of C:

1. Equal tailed:

$$\int_{-\infty}^{\theta_L} \pi(\theta|\mathbf{x})d\theta = \alpha/2, \int_{\theta_U}^{\infty} \pi(\theta|\mathbf{x})d\theta = \alpha/2$$

2. Smallest length: Choose C to minimize $\theta_U - \theta_L$.

3. Highest posterior density region (HPD), define C to be that set with posterior probability $1-\alpha$ which satisfies the criterion.

$$\theta_1 \in C \text{ and } \pi(\theta_2|\mathbf{x}) > \pi(\theta_1|\mathbf{x}) \implies \theta_2 \in C$$

So C as the set

$$C = \{\theta \in \Theta : \pi(\theta|\mathbf{x}) > c_\alpha\}$$

If the posterior is unimodal then HPD is also the smallest length interval.

Definition 10.6. *A statistic $T(\mathbf{X})$ is* ***sufficient*** *for θ iff the posterior distribution of θ given $\mathbf{X}$ is the same as the posterior distribution of θ given $T(\mathbf{X})$.*

11 Week 11 decision theory and non-parametric

Definition 11.1. *θ denotes the true state of nature.*

Suppose we are able to observe some data, a draw from the random variable $\mathbf{X}$ whose distribution depends on θ. Sometimes no data are available.

A ***decision procedure*** *δ specifies what action to take for each value of $\mathbf{X}$. If we observe $\mathbf{X} = \mathbf{x}$, then we should adopt the procedure $\delta(\mathbf{x})$.*

Whether $\delta(\mathbf{x})$ is good depends on the ***loss function*** *which measures the loss from $\delta(\mathbf{x})$ when θ holds*

$$L_S(\theta, \delta(\mathbf{x}))$$

Note that the negative of a loss function is a utility function.

Frequentists want to minimize expected loss. Bayesians want to minimize posterior expected loss.

Definition 11.2. *The* ***risk function*** *$R(\theta, \delta(\mathbf{x}))$ is defined as* example

$$R(\theta, \delta(\mathbf{x})) = \int L_S(\theta, \delta(\mathbf{x}))L(\theta, \mathbf{x})d\mathbf{x}$$

This is the expected loss with respect to the joint distribution (likelihood).

Definition 11.3. *A procedure δ_1 is* ***inadmissible*** *if there exists another procedure δ_2 such that*

$$R(\theta, \delta_1) \geq R(\theta, \delta_2) \forall \theta$$

Definition 11.4. *The* ***minimax procedure*** *is such that $\max_\theta R(\theta, \delta)$ is minimized.*

Definition 11.5. *A* ***Bayes*** *procedure is such that the* ***Bayes risk***

$$\int R(\boldsymbol{\theta}, \delta)p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

is minimized. Expected prior loss is minimized.

A Bayes procedure with constant risk for θ is minimax.

A minimax procedure is generally a Bayes procedure for some prior distribution. In particular, the so called **least favourable prior distribution.**

Some simple loss functions:

1. Zero-one loss:

$$L_S(\theta, \delta = \hat{\theta}) = 0 \text{ when } |\hat{\theta} - \theta| < b; = a \text{ when } |\hat{\theta} - \theta| \geq b \text{ where } a, b > 0$$

2. Absolute error loss:

$$L_S(\theta, \delta = \hat{\theta}) = a|\hat{\theta} - \theta|$$

3. Squared Error loss (quadratic loss):

$$L_S(\theta, \delta = \hat{\theta}) = a(\hat{\theta} - \theta)^2$$

Non-parametric methods:

The empirical distribution function $\hat{F}$ is the CDF of a new discrete random variable X^*. Can be shown that $\hat{F}$ is a **sufficient statistic** for F (based on a random sample)

$\hat{F}$ and X^* mimics the relationship between F and X. This leads to studying $(\hat{F}, X^*)$ to learn about (F, X).

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(x_i \leq x)}$$

while $F(x) = P(X \leq x)$.

$$\begin{aligned} E(\hat{F}(x)) &= E\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(x_i \leq x)}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left\{\mathbb{I}_{(x_i \leq x)}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} P(X_i \leq x) \\ &= \frac{1}{n}\sum_{i=1}^{n} F(x) \\ &= F(x) \\ &= \frac{n_x}{n} \end{aligned} \tag{19}$$

where n_x is the number of observed data which are less than or equal to the value x.

Also can treat n_x as binomially distributed with n trials and a success probability $p = P(X_i \leq x) = F(x)$.

$$E(\hat{F}) = E(\frac{n_x}{n}) = \frac{E(n_x)}{n} = \frac{np}{n} = \frac{nF(x)}{n} = F(x)$$

$$V(\hat{F}) = V(\frac{n_x}{n}) = \frac{1}{n^2}V(n_x) = \frac{1}{n^2}np(1-p) = \frac{1}{n}F(x)(1 - F(x))$$

Definition 11.6. ***Bootstrap:*** *Construct an estimate of the true population parameter $\theta(\hat{F})$ using the re-sampled data, arriving at $\theta(\hat{F}^*)$ as the estimator for $\theta(\hat{F})$.*

Define the bootstrap estimators of bias and variance as

$$\hat{B}_B = E_{\hat{F}}\left\{\theta(\hat{F}^*)\right\} - \theta(\hat{F})$$

$$\hat{V}_B\left\{\theta(\hat{F}^*)\right\} = E_{\hat{F}}\left\{\theta(\hat{F}^*)^2\right\} - [E_{\hat{F}}\left\{\theta(\hat{F}^*)\right\}]^2$$

We create B bootstrap (re-sampled) data sets:

$$\{X^*_{1,1}, \dots, X^*_{n,1}\}, \dots, \{X^*_{1,b}, \dots, X^*_{n,b}\}, \dots, \{X^*_{1,B}, \dots, X^*_{n,B}\}$$

with each sample we estimate $\hat{\theta}^b = \theta(\hat{F}^*_b)$ then we can compute:

$$\hat{B}_B \approx \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_b - \hat{\theta}$$

$$\hat{V}_B\left\{\theta(\hat{F}^*)\right\} \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b\right)^2$$

$$\hat{\sigma}_B(\hat{\theta}) = \sqrt{\frac{1}{B-1}(\hat{\theta}^*_b - \bar{\hat{\theta}}^*)^2}$$

Definition 11.7. ***bootstrap interval estimation:***

$$[\hat{\theta} - z_{\alpha/2}\hat{\sigma}_B(\hat{\theta}), \hat{\theta} + z_{\alpha/2}\hat{\sigma}_B(\hat{\theta})]$$

For small samples, we still might be in trouble, so use Student's t-distribution quantiles instead of the standard normal quantiles

Non-parametric Testing:

1. Permutation/randomization tests: suppose we have a two sample problem, the two data sets are drawn from two potentially different probability distribution F and G:

$$\begin{aligned} F &\to \mathbf{y} = (y_1, \dots, y_n) \\ G &\to \mathbf{x} = (x_1, \dots, x_n) \end{aligned} \tag{20}$$

Consider testing whether H_0: F(treatment)=G(control), H_1: the distributions are not equal.

Need a test statistic: $\theta = \bar{T} - \bar{C}$

Let n and m be sizes of treatment and control.

Permutation tests: how many possible ways are there to permute? $N = n + m$:

$$\mathbf{z} = \{y_1, \dots, y_n, x_1, \dots, x_m\}$$

$$\frac{N!}{n!m!} = \binom{N}{n}$$

can have an exact p-value if we construct our test based on every possible permutation. But if we randomly sample from the N choose n possible permutation, then we have a **randomization test** and p-value is approximate.

p-value:

$$P(T(\mathbf{X}, \mathbf{Y}) \geq T(\mathbf{x}, \mathbf{y}))$$

2. Bootstrap tests: same setup.

1. Draw B samples of size $N = n+m$ with replacement from $\mathbf{z} = \{y_1, \dots, y_n, x_1, \dots, x_m\}$. Call the first n observations $\mathbf{y}^*$ and the remaining m obs. $\mathbf{x}^*$.

2. Evaluate the test statistic $t(\cdot)$ on each bootstrap sample:

$$t(\mathbf{y}^*, \mathbf{x}^*) = \bar{y}^* - \bar{x}^*$$

3. Construct an approximate p-value:

$$p \approx \#\{t(\mathbf{X}^*, \mathbf{Y}^*) \geq t_{obs}\}/B$$

12 Week 12 random sampling and MH

Inverse CDF method - generating random variables: Based on the uniform-exponential relationship we can generate the following:

- Sums of iid exponential random variables have a gamma distribution

$$Y = -\beta \sum_{j=1}^{a} \log(U_j) \sim gamma(a, \beta)$$

- If $\beta = 2$, then the distribution is a χ^2 random variable:

$$Y = -2 \sum_{j=1}^{v} \log(U_j) \sim \chi^2_{2v}$$

- The ratio of sums of exponentials is a beta distribution:

$$Y = \frac{\sum_{j=1}^{a} \log(U_j)}{\sum_{j=1}^{a+b} \log(U_j)} \sim Beta(a, b)$$

Then how can we randomly generate $N(0,1)$? If we could generate χ^2_1, then we have normal.

Example 12.1. ***Box-Muller Algorithm:*** *Generate $U_1, U_2 \sim Unif(0,1)$, set*

$$R = \sqrt{-2\log(U_1)}, \theta = 2\pi U_2$$

$$X = R\cos(\theta), Y = R\sin(\theta)$$

Then $X, Y \overset{iid}{\sim} N(0,1)$. If we want χ^2_1 then we sample X^2, Y^2.

What about discrete distributions? If Y is a discrete random variable taking on values:

$$y_1 < y_2 < \cdots < y_k$$

then we write:

$$P[F_Y(y_i) < U \leq F_Y(y_{i+1})] = F_Y(y_{i+1}) - F_Y(y_i) = P(Y = y_{i+1})$$

Using this idea, we can easily generate discrete RVs. To generate $Y_i \sim f_Y(y)$.

1. Generate $U \sim U(0,1)$,

2. If $F_Y(y) < U \leq F_Y(y_{i+1})$, set $Y = y_{i+1}$. Define $y_0 = -\infty, F_Y(y_0) = 0$.

Theorem 12.2. ***The Accept/Reject Algorithm:*** *let $Y \sim f_Y(y)$ and $V \sim f_V(v)$ where densities have common support and*

$$M = \sup_y \frac{f_Y(y)}{f_V(y)} < \infty$$

Suppose we want to sample from Y and are able to sample from V.

1. Generate $U \sim Unif(0,1)$ and $V \sim f_V$ independently.

2. If $U < \frac{1}{M}\frac{f_Y(V)}{f_V(V)}$, set $Y = V$; otherwise, return to step (1).

Note envelope= $Mf_V(v) \geq f_Y(v)$

For the standard A/R algorithm need a good envelope. when a good envelope is not available, MCMC can aid in sampling for a desired target distribution.

Interested in modelling data where $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \xi)$ where $\xi = \frac{1}{\sigma^2}$ is the precision, then

$$f_X(x|\theta,\xi) = \left(\frac{\xi}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2}\xi(x-\theta)^2\right)$$

Model the priors as being independent.

$$p(\theta,\xi) = p(\theta)p(\xi)$$

The prior $\theta \sim N(\theta_0, \tau_0)$, $\xi \sim gamma(\alpha_0, \lambda_0)$
Here we have:

$$p(\theta,\xi|\mathbf{x}) \propto p(\mathbf{y}|\theta,\xi)p(\theta)p(\xi)$$

In a **Metropolis-Hastings sampling scheme:**
We propose a new value of θ, say θ^* and decide to accept or reject.
We propose a new value of ξ, say ξ^* and decide to accept or reject.
$\theta^* \sim N(\theta, \delta_1)$, symmetric proposal

$$\rho = \frac{p(\mathbf{y}|\theta^*,\xi)p(\theta^*)p(\xi)}{p(\mathbf{y}|\theta,\xi)p(\theta)p(\xi)} = \frac{p(\mathbf{y}|\theta^*,\xi)p(\theta^*)}{p(\mathbf{y}|\theta,\xi)p(\theta)}$$

$\xi^* \sim unif(\xi - \delta_2, \xi + \delta_2)$. If $\xi^* < 0$ then reflect the value to the positive line.

$$\rho = \frac{p(\mathbf{y}|\theta,\xi^*)p(\theta)p(\xi^*)}{p(\mathbf{y}|\theta,\xi)p(\theta)p(\xi)} = \frac{p(\mathbf{y}|\theta,\xi^*)p(\xi^*)}{p(\mathbf{y}|\theta,\xi)p(\xi)}$$