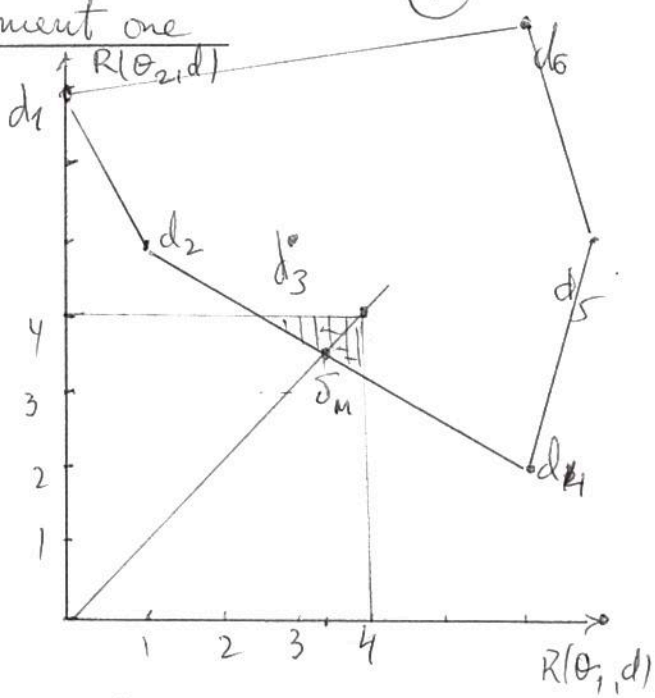
Solutions to Assignment one

Q1| a) Taking max for each of the 6 decision rules' risks we get: $(7, 5, 5, 6, 7, 8)$. Hence both d_2 and d_3 are minimax in the set D.

b) See graph → the convex hull

c) Need to find the intersection point of

$$\left.\begin{array}{l} y = x \\ y - 2 = \frac{5-2}{1-6}(x-6) \end{array}\right\} \; y = x = 3.5$$



Hence the risk point δ_M of the minimax rule in $\mathcal{D}$ is $(4.5, 4.5)$

d) Looking for α which is such that

$$\alpha * 1 + (1-\alpha)6 = 3.5 \rightarrow \alpha = \frac{1}{2}$$
$$\alpha * 5 + (1-\alpha)2 = 3.5$$

i.e. $\delta_M =$ → choose d_2 w.p. $\frac{1}{2}$
→ choose d_4 w.p. $\frac{1}{2}$

e) If the prior is $(p, 1-p)$ this leads to a line with a normal vector $(p, 1-p)$, i.e., a slope $-\frac{p}{1-p}$ and this slope should coincide with the slope of $\overline{d_2 d_4}$, i.e. $-\frac{p}{1-p} = \frac{5-2}{1-6} = -\frac{3}{5}$ should hold, i.e. $p = \frac{3}{8}$ and the least favorable prior, w.r. to which δ_M is Bayes, is $(\frac{3}{8}, \frac{5}{8})$ on (θ_1, θ_2).

f) The line with a normal vector $(\frac{5}{6}, \frac{1}{6})$ has a slope of -5. When moving such a line south-west as much as possible by retaining an intersection with the risk set, we end up with d_1, which is the corresponding Bayes rule. Its Bayes risk is $\frac{5}{6} * 0 + \frac{1}{6} * 7 = \underline{7/6}$.

g| See the shaded area.

Q2 (2)

i) $f(\bar{X}|\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^n$

$f(\bar{X}|\theta)\tau(\theta) = 3\theta^{\sum_{i=1}^{n} x_i + 2}(1-\theta)^n$. This implies that $\tau(\theta)f(\theta|\bar{X}) \propto \theta^{\sum_{i=1}^{n} x_i + 2}(1-\theta)^n$ which means that the posterior of θ given the sample is $\text{Beta}\left(\sum_{i=1}^{n} x_i + 3, n+1\right)$

Then $\hat{\theta}_{Bayes}$ = mean of this distribution = $\dfrac{\sum_{i=1}^{n} x_i + 3}{\sum_{i=1}^{n} x_i + 4 + n}$

iii) Now we have $\hat{\theta} = \frac{20}{26}$ which is in the vicinity of .75 so need the effect of the prior to see if H_0 or H_1 is more relevant. The posterior is $\text{Beta}(20, 6)$

$$P(H_0|\text{sample}) = \frac{1}{B(20,6)} \int_0^{.75} x^{19}(1-x)^5 dx \approx .3783$$

Hence H_0 should be rejected.
(The value of the integral has been calculated using your favorite program for numerical integration)

Q3|

$X_1, X_2, \dots, X_n$ are i.i.d. uniform in $(0, \theta)$

$$\Rightarrow f(x_i|\theta) = \frac{1}{\theta} I_{(x_i, \infty)}(\theta)$$

Hence $\prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n} I_{(x_{(n)}, \infty)}(\theta)$

The prior $\pi(\theta) = \begin{cases} \beta \alpha^\beta \theta^{-(\beta+1)} \\ 0 \end{cases}$, $\theta > \alpha > 0$ can also be written via indicators on "one line" as $\beta \alpha^\beta \theta^{-(\beta+1)} I_{(\alpha, \infty)}(\theta)$

The joint is a product; the product of indicators is an indicator so we end up with

$$\prod_{i=1}^{n} f(x_i|\theta)\pi(\theta) = \frac{\beta \alpha^\beta \theta^{-(\beta+1)}}{\theta^n} I_{(\max(x_{(n)}, \alpha), \infty)}(\theta) =$$

$$= \beta \alpha^\beta \theta^{-(n+\beta+1)} I_{(\max(x_{(n)}, \alpha), \infty)}(\theta)$$

Hence the Bayes estimator w.r.t. quadratic loss is

$$E(\theta|X) = \frac{\cancel{\alpha^\beta} \int_{\max(x_{(n)}, \alpha)}^{\infty} \theta^{-(n+\beta)} d\theta}{\cancel{\beta \alpha^\beta} \int_{\max(x_{(n)}, \alpha)}^{\infty} \theta^{-(n+\beta+1)} d\theta} =$$

$$= \frac{(n+\beta)}{(n+\beta-1)} \max(x_{(n)}, \alpha)$$

(Q4) The observation scheme: we have $n=1$ observation only from a geometric distribution with

$$f(x|\theta) = (1-\theta)^{x-1}\theta$$

where $\theta \in (0,1)$ is the probability of success in a single trial (since our data expresses the total number of trials until the first success).

The two priors are $\pi_1(\theta) = 6\theta(1-\theta)$ of the aviation minister and $\pi_2(\theta) = 4\theta^3$ of the prime minister.

The two corresponding posteriors are:

$h_1(\theta|x) \propto \theta^2(1-\theta)^x$ and $h_2(\theta|x) \propto \theta^4(1-\theta)^{x-1}$

These can easily be identified as

$h_1(\theta|x) \sim$ Beta $(3, x+1)$

$h_2(\theta|x) \sim$ Beta $(5, x)$

We have 2 actions available: $a_0 \equiv$ continue

$a_1 \equiv$ abandon

The losses related to these actions are:-

$$L(\theta, a_0) = \begin{cases} \frac{1}{2} - \theta & \text{if } \theta < \frac{1}{2} \\ 0 & \text{if } \theta \geq \frac{1}{2} \end{cases} \quad \text{and}$$

$$L(\theta, a_1) = \begin{cases} 0 & \text{if } \theta < \frac{1}{2} \\ \theta - \frac{1}{2} & \text{if } \theta \geq \frac{1}{2} \end{cases}$$

For an optimal Bayes decision we need to compare: $Q(x,a_0) = \int_0^{1/2}(\frac{1}{2}-\theta)h(\theta)d\theta = \frac{1}{2}\int_0^{1/2}h(\theta|x)d\theta - \int_0^{1/2}\theta h(\theta|x)d\theta$ with

$$Q(x,a_1) = \int_{1/2}^{1}(\theta-\frac{1}{2})h(\theta|x)d\theta = \int_{1/2}^{1}\theta h(\theta|x)dx - \frac{1}{2}\int_{1/2}^{1}h(\theta|x)d\theta.$$

Then a_0 would be preferred to a_1 if $Q(x,a_0)<Q(x,a_1)$ (alternatively if $Q(x,a_0)>Q(x,a_1)$ then a_1 would be preferred (and there is hesitation if $Q(x,a_0)=Q(x,a_1)$)

From the inequality

$$\frac{1}{2}\int_0^{1/2}h(\theta|x)d\theta - \int_0^{1/2}\theta h(\theta|x)dx < \int_{1/2}^{1}\theta h(\theta|x)d\theta - \frac{1}{2}\int_{1/2}^{1}h(\theta|x)d\theta$$

we see that adding $\pm\frac{1}{2}\int_0^{1/2}h(\theta|x)d\theta$, noting that $\frac{1}{2}\int_0^1 h(\theta|x)d\theta=\frac{1}{2}$ and re-arranging we get

$$\frac{1}{2}\int_0^{1/2}h(\theta|x)d\theta - \int_0^{1/2}\theta h(\theta|x)d\theta < \int_{1/2}^{1}\theta h(\theta|x)d\theta - \frac{1}{2} + \frac{1}{2}\int_0^{1/2}h(\theta|x)d\theta$$

Hence $$\frac{1}{2} < \int_0^1 \theta h(\theta|x)d\theta = E(\theta|x)$$

In other words, we choose $a_0 \equiv$ continue if $E(\theta|x)>\frac{1}{2}$ (and, of course, this decision is also intuitively appealing).

Now, for a Beta (α,β) distribution, the expected value is $\frac{\alpha}{\alpha+\beta}$ which implies in our case:

$E(\theta|x) = 3/(x+4)$ for aviation minister

$E(\theta|x) = 5/(x+5)$ for prime minister

Hence the aviation minister wants the project to continue when $x=1$, hesitates when $x=2$ and wants to stop when $x=3,4,5\ldots$. The prime minister wants to continue when $x=1,2,3,4$; hesitates when $x=5$ and wants to stop when $x=6,7,8,\ldots$. Obviously, for $x=3$ and $x=4$ we have the most serious disagreement

13/13

Question 1 We let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of n observations each with a uniform density in $[0, \theta)$

$$f(x,\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

Where $\theta > 0$ is an unknown parameter. The CDF of this function is

$$F(x,\theta) = \begin{cases} 0 & \text{if } x < 0 \\ \int_0^x \frac{1}{\theta} dx = \frac{x}{\theta} & \text{if } 0 < x < \theta \\ 1 & x > \theta \end{cases}$$

(a) Denoting the joint density as $L(X,\theta)$, we can show that the family $\{L(X,\theta)\}, \theta > 0$ has a monotone likelihood ratio in $X_{(n)}$. First, rewriting the density as

$$f(x,\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$$

Then the likelihood function is

$$L(X,\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i)$$

Which is true if and only if $\theta > max[x_1, ..., x_n]$, denoted as $X_{(n)}$, the value of the largest order statistic. The likelihood function is,

$$L(X,\theta) = \frac{1}{\theta^n} I_{[0,\theta]}(X_{(n)})$$

For $\theta_1 < \theta_2$,

$$\frac{L(X,\theta_2)}{L(X,\theta_1)} = \frac{\frac{1}{\theta_2^n} I_{[0,\theta_2]}(X_{(n)})}{\frac{1}{\theta_1^n} I_{[0,\theta_1]}(X_{(n)})} = (\frac{\theta_1}{\theta_2})^n \frac{I_{[0,\theta_2]}(X_{(n)})}{I_{[0,\theta_1]}(X_{(n)})}$$

Therefore,

$$\frac{L(X,\theta_2)}{L(X,\theta_1)} = (\frac{\theta_1}{\theta_2})^n \times \begin{cases} 1 & 0 \le X_{(n)} \le \theta_1 \\ \infty & \theta_1 \le X_{(n)} \le \theta_2 \end{cases}$$

Which is a monotone non-decreasing function of $T(X) = X_{(n)}$. ✓

(b) Given the density function of $\mathbf{X} = (X_1, X_2, ..., X_n)$ is in a family with monotone likelihood ratio in $T(X) = X_{(n)}$, then for testing $H_0 : \theta \le 2$ versus $H_1 : \theta > 2$, the UMP level α-test is given by

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 & \text{if } X_{(n)} > k \\ 0 & \text{if } X_{(n)} \le k \end{cases}$$

To find k, we need the distribution of $x_{(n)}$. Using the fact that by independence of the observations,

$$P_\theta(X_{(n)} > k) = (P_\theta(X_1 > k))^n$$

Also, $P_\theta(X_1 > k) = 1 - P(X_1 \le k) = 1 - F(k,\theta)$, then using the CDF function from part a,

$$P_\theta(X_1 > k) = \begin{cases} 1 - 0 = 1 & \text{if } k < 0 \\ 1 - \frac{k}{\theta} & \text{if } 0 < k < \theta \\ 1 - 1 = 0 & \text{if } k > \theta \end{cases}$$

Hence

$$E_\theta \varphi^* = P_\theta(X_{(n)} > k) = (P_\theta(X_1 > k))^n = \begin{cases} 1 & \text{if } k < 0 \\ 1 - (\frac{k}{\theta})^n & \text{if } 0 < k < \theta \\ 0 & \text{if } k > \theta \end{cases}$$

Therefore under the null hypothesis, $\theta = \theta_0 = 2$, so to solve for k we solve the equation $E_{\theta_0}\varphi^* = 1-(\frac{k}{\theta})^n = \alpha$,

$$\alpha = 1 - (\frac{k}{2})^n$$

$$1 - \alpha = (\frac{k}{2})^n$$

$$[2^n(1-\alpha)]^{\frac{1}{n}} = [k^n]^{\frac{1}{n}}$$

$$2(1-\alpha)^{\frac{1}{n}} = k$$ ✓

Hence, the UMP α-size test is

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 & \text{if } X_{(n)} > 2(1-\alpha)^{\frac{1}{n}} \\ 0 & \text{if } X_{(n)} \leq 2(1-\alpha)^{\frac{1}{n}} \end{cases}$$

(c) The power function is defined as part b with

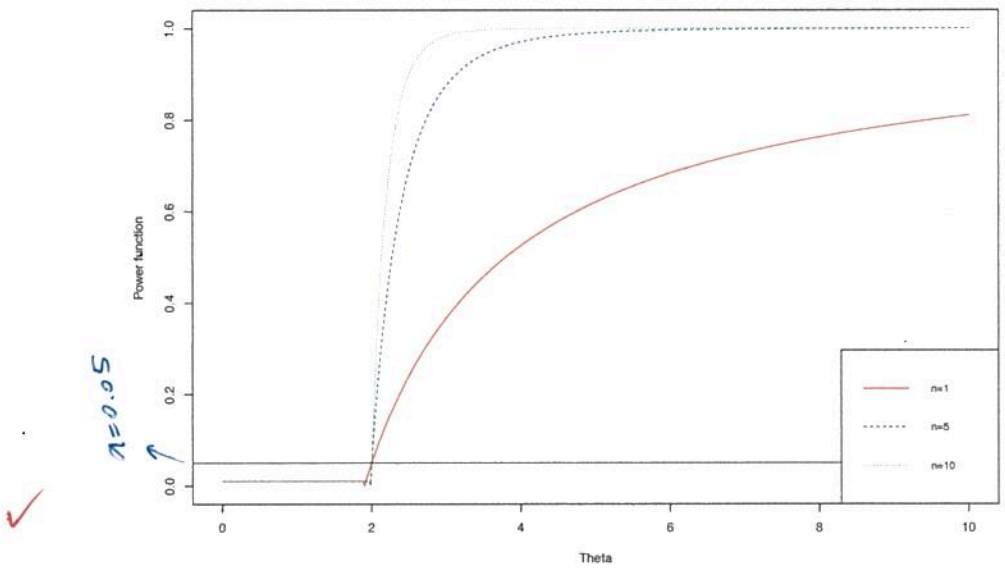
$$E_\theta\varphi^* = 1 - (\frac{k}{\theta})^n$$

Substituting in $k = 2(1-\alpha)^{\frac{1}{n}}$,

$$E_\theta\varphi^* = 1 - \frac{[2(1-\alpha)^{\frac{1}{n}}]^n}{\theta^n} = 1 - (\frac{2}{\theta})^n(1-\alpha)I_{(k,\infty)}(\theta)$$

With $E_\theta\varphi^* = \alpha$, if $\theta = 2$. Plotting this, the power function increases as θ increases, approaching 1 as θ approaches ∞. As n gets larger, the function approaches one very quickly.

Figure 1: Power Function - Uniform



(d) First, from part c, $f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n}, 0 < x < \theta$, with

$$F_{X_{(n)}}(x) = \begin{cases} 0 & \text{if } x < 0 \\ (\frac{x}{\theta})^n & \text{if } 0 \leq x < \theta \\ 1 & \text{if } x \geq \theta \end{cases}$$

Defining random variable $Y_n = n(1 - \frac{X_{(n)}}{\theta})$, then the distribution of Y_n is

$$F_{Y_n}(y) = P[Y_n \leq y] = P[n(1 - \frac{X_{(n)}}{\theta}) \leq y]$$

$$= P[1 - \frac{X_{(n)}}{\theta} \leq \frac{y}{n}]$$

$$= P[X_{(n)} \geq \theta(1 - \frac{y}{n})]$$

$$= 1 - P[X_{(n)} < \theta(1 - \frac{y}{n})]$$

$$= 1 - P[X_1 \leq \theta(1 - \frac{y}{n})] \times P[X_2 \leq \theta(1 - \frac{y}{n})] \times ... \times P[X_n \leq \theta(1 - \frac{y}{n})]$$

By mutual independence of X_i, then

$$1 - F_{X_1}(\theta(1 - \frac{y}{n})) \times F_{X_2}(\theta(1 - \frac{y}{n})) \times ... \times F_{X_n}(\theta(1 - \frac{y}{n}))$$

$$= 1 - [F_{X_n}(\theta(1 - \frac{y}{n}))]^n$$

Therefore,

$$F_{Y_n}(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \frac{(\theta(1-\frac{y}{n}))^n}{\theta} = 1 - (1 - \frac{y}{n})^n & \text{if } 0 \leq y < n \\ 1 & \text{if } y \geq n \end{cases}$$

Since,

$$lim_{n\to\infty}(1 - \frac{y}{n})^n = e^{-y}$$

Then,

$$lim_{n\to\infty} F_{Y_n}(y) = F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - e^{-y} & \text{if } y \geq 0 \end{cases}$$

Where $F_Y(y)$ is the distribution function of an exponential random variable with mean of one. Therefore Y_n converges in distribution to an exponential random variable with mean of one as $n \to \infty$. We can evaluate $P(|X_{(n)} - \theta| < \epsilon)$ directly by noting that $X_{(n)}$ cannot possible be larger than θ, so

$$P(|X_{(n)} - \theta| < \epsilon) = P(X_{(n)} > \theta - \epsilon) = 1 - P(X_{(n)} \leq \theta - \epsilon)$$

The maximum $X_{(n)}$ is less than some constant if and only if each of the random variables $X_1, ..., X_n$ is less than that constant. Therefore, since the X_i are i.i.d.,

$$P(X_{(n)} \leq \theta - \epsilon) = [P(X_1 \leq \theta - \epsilon)]^n = \begin{cases} [1 - (\epsilon/\theta)]^n & \text{if } 0 < \epsilon < \theta \\ 0 & \text{if } \epsilon \geq \theta \end{cases}$$

Since $1 - (\epsilon/\theta)$ is strictly less than 1, we conclude that no matter what positive value ϵ takes,

$$P(X_{(n)} \leq \theta - \epsilon) \to 0$$

as desired.

19/19

Question 2 We let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be i.i.d. variables with density

$$f(x,\theta) = \begin{cases} \frac{2}{\theta} x e^{\frac{-x^2}{\theta}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $\theta > 0$ is an unknown parameter.

(a) We can find the information matrix using the following formula

$$I(\theta) = -E(\frac{\partial^2 \log L(X,\theta)}{\partial \theta^2}) \quad (1)$$

First, finding the likelihood function for one observation (i.e. X_1)

$$L(X,\theta) = \frac{2X}{\theta} e^{\frac{-X^2}{\theta}}$$

With a log-likelihood function as

$$\log L(X,\theta) = \log\{\frac{2X}{\theta} e^{\frac{-X^2}{\theta}}\}$$

$$= \log 2 - \log\theta + \log e^{-\frac{1}{\theta}X^2} + \log X$$

$$= \log 2 - \log\theta - \frac{X^2}{\theta} + \log X$$

Taking the first derivative w.r.t to θ,

$$\frac{\partial \log L(X,\theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{X^2}{\theta^2}$$

Then the second derivative w.r.t to θ is

$$\frac{\partial^2 \log L(X,\theta)}{\partial \theta^2} = \frac{1}{\theta^2} - \frac{2X^2}{\theta^3}$$

Substituting this into (1), yields

$$I(\theta) = -E(\frac{1}{\theta^2} - \frac{2X^2}{\theta^3}) = -\frac{1}{\theta^2} + \frac{2}{\theta^3}E[X^2]$$

To find $E[X^2]$, we can prove the square of a Rayleigh(θ) random variable is an Exponential(θ) random variable. Considering the transformation $Y = g(X) = X^2$ is a 1-1 transformation from $X = \{x|x > 0\}$ to $Y = \{y|y > 0\}$ with inverse $X = g^{-1}(Y) = \sqrt{y}$ and Jacobian

$$\frac{dX}{dY} = \frac{1}{2}Y^{-\frac{1}{2}} = \frac{1}{2\sqrt{Y}}$$

Therefore by the transformation technique, the probability density function of Y is

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}|$$

$$= \frac{2\sqrt{y}}{\theta} e^{-\sqrt{y}^2/\theta} |\frac{1}{2\sqrt{y}}|$$

$$= \frac{1}{\theta} e^{-y/\theta}, y > 0$$

Which is the density function of the exponential distribution with expectation of θ, i.e. $X^2 \approx Exp(\theta)$ and $E[X^2] = \theta$. Therefore,

$$I(\theta) = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = -\frac{1}{\theta^2} + \frac{2}{\theta^2} = \frac{1}{\theta^2}$$

For a sample of n i.i.d. observations,

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta^2}$$ ✓

(b) To find the MLE, we find the likelihood function for the sample as

$$L(X,\theta) = \prod_{i=1}^{n} \frac{2x_i}{\theta} e^{\frac{-x_i^2}{\theta}} = \frac{2^n}{\theta^n} e^{\frac{-\sum_{i=1}^n x_i^2}{\theta}} \sum_{i=1}^{n} x_i$$

With a log-likelihood function as

$$\log L(X,\theta) = \log\{\frac{2^n}{\theta^n} e^{\frac{-\sum_{i=1}^n x_i^2}{\theta}} \sum_{i=1}^{n} x_i\}$$

$$= \log 2^n - \log \theta^n + \log e^{-\frac{1}{\theta}\sum_{i=1}^n x_i^2} + \log \sum_{i=1}^{n} x_i$$

$$= n\log 2 - n\log\theta - \frac{\sum_{i=1}^n x_i^2}{\theta} + \sum_{i=1}^{n} \log x_i$$

Taking the first derivative w.r.t to θ,

$$\frac{\partial \log L(X,\theta)}{\partial\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i^2}{\theta^2}$$

Setting to 0 to solve for the MLE, yields

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i^2}{\theta^2} = 0$$

$$\frac{\sum_{i=1}^n x_i^2}{\theta^2} = \frac{n}{\theta}$$

$$\theta \sum_{i=1}^{n} x_i^2 = n\theta^2$$

$$\sum_{i=1}^{n} x_i^2 = n\theta$$

Therefore the MLE of θ is,

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i^2}{n}$$ ✓

See part (a), we know the square of a rayleigh-distributed random variable follows an exponential distribution with parameter θ, then

$$E[\hat{\theta}] = E[\frac{\sum_{i=1}^n x_i^2}{n}] = \frac{\sum_{i=1}^n E[x_i^2]}{n} = \frac{n\theta}{n} = \theta$$

Therefore, the MLE of θ is unbiased. ✓ To obtain the variance of the MLE,

$$V[\hat{\theta}] = V[\frac{1}{n}\sum_{i=1}^{n} x_i^2] = \frac{1}{n^2}\sum_{i=1}^{n} V[x_i^2] = \frac{n\theta^2}{n^2} = \frac{\theta^2}{n}$$

Since the variance of an exponential distribution with parameter θ is θ^2. The Cramer-Rao bound is

$$\frac{1}{nI(\theta)} = \frac{1}{n\frac{1}{\theta^2}} = \frac{\theta^2}{n}$$

Therefore, the MLE of $\hat{\theta}$ attains the Cramer-Rao lower bound. ✓

(c) The asymptotic distribution of $\hat{\theta}$ is

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

Where $\frac{1}{I(\theta)} = \theta^2$. From lecture notes, page 43, $\hat{\theta} \approx N(\theta, \frac{1}{nI(\theta)})$. Using the results from part a and b,

$$\hat{\theta} \approx N(\theta, \frac{\theta^2}{n})$$

✓

(d) First, rewriting the density function as follows,

$$f(x, \theta) = \frac{2}{\theta} x e^{-\frac{x^2}{\theta}} = \frac{2}{\theta} x e^{\frac{1}{\theta} - x^2}$$

This is in the form of a one-parameter exponential family, $a(\theta)b(x)e^{[c(\theta)d(x)]}$, with

$$a(\theta) = \frac{2}{\theta}$$

$$b(x) = x$$

$$c(\theta) = -\frac{1}{\theta}$$

$$d(x) = x^2$$

Clearly this family has MLR in $T(X) = \sum_{i=1}^{n} d(X_i) = \sum_{i=1}^{n} x_i^2$ because $c(\theta)$ is increasing in θ. This also means T(X) is minimal sufficient.

(e) Given $T(X) = \sum_{i=1}^{n} x_i^2$ is a sufficient statistic for θ and the family is an MLR family, then testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ exists and the UMP level α-size test is given by rejecting H_0 if and only if $T(X) > t_0$ where $\alpha = P[T(X) > t_0 | \theta_0]$. Hence, the UMP α-size test is

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} X_i^2 \geq k \\ 0 & \text{if } \sum_{i=1}^{n} X_i^2 < k \end{cases}$$

This can be evaluated as an exact test using the Gamma distribution, that is, using the fact that $\sum_{i=1}^{n} x_i^2 \approx Gamma(n, \theta)$. This becomes difficult to integrate, so we can use the fact that $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i^2}{n} \approx N(\theta, \frac{\theta^2}{n})$, from part c,

$$P[\sum_{i=1}^{n} x_i^2 > k] = \alpha$$

$$= P[\frac{\sum_{i=1}^{n} x_i^2}{n} > \frac{k}{n}] = \alpha$$

Under the null hypothesis, $\theta = \theta_0$,

$$P[\frac{\frac{\sum_{i=1}^{n} x_i^2}{n} - \theta_0}{\theta_0/\sqrt{n}} \geq \frac{\frac{k}{n} - \theta_0}{\theta_0/\sqrt{n}}] = \alpha$$

So that $Z = \frac{\frac{\sum_{i=1}^{n} x_i^2}{n} - \theta_0}{\theta_0/\sqrt{n}}$ is approximately $N(0, 1)$. When H_0 is true,

$$\Phi = [\frac{\frac{\sum_{i=1}^{n} x_i^2}{n} - \theta_0}{\theta_0/\sqrt{n}}] = 1 - \alpha \to (\frac{\frac{\sum_{i=1}^{n} x_i^2}{n} - \theta_0}{\theta_0/\sqrt{n}}) = z_\alpha$$

Now solving for k,

$$z_\alpha = \frac{\frac{k}{n} - \theta_0}{\theta_0/\sqrt{n}}$$

$$\frac{k}{n} = \theta_0 + \frac{z_\alpha \theta_0}{\sqrt{n}}$$

$$k = n\theta_0 + \sqrt{n} z_\alpha \theta_0 = \theta_0(n + \sqrt{n} z_\alpha)$$ ✓

Therefore the UMP α-size test is

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n X_i^2 \geq n\theta_0 + \sqrt{n} z_\alpha \theta_0 \\ 0 & \text{if } \sum_{i=1}^n X_i^2 < n\theta_0 + \sqrt{n} z_\alpha \theta_0 \end{cases}$$

(f) From part e, the power function can be obtained using the normal approximation as

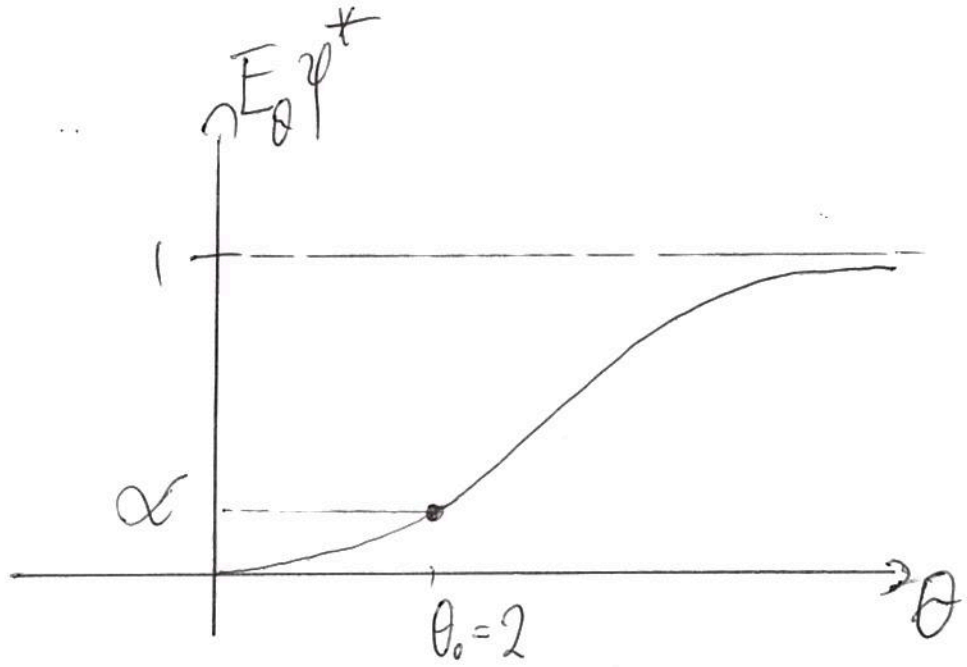
$$E_\theta \varphi^* \approx P[\frac{\sum_{i=1}^n X_i^2}{n} > \theta_0 + \frac{z_\alpha \theta_0}{\sqrt{n}})]$$

$$= P[\frac{\frac{\sum_{i=1}^n X_i^2}{n} - \theta}{\theta/\sqrt{n}} > \frac{\theta_0 + \frac{z_\alpha \theta_0}{\sqrt{n}} - \theta}{\theta/\sqrt{n}}]$$

$$= P[Z > \frac{\theta_0 + \frac{z_\alpha \theta_0}{\sqrt{n}} - \theta}{\theta/\sqrt{n}}] = 1 - P[Z \leq \frac{\theta_0 + \frac{z_\alpha \theta_0}{\sqrt{n}} - \theta}{\theta/\sqrt{n}}]$$ ✓

$$= 1 - \Phi(\frac{\frac{k}{n} - \theta}{\theta/\sqrt{n}})$$

As an example, setting $\theta_0 = 2$, and setting $\alpha = 0.05$ so that $z_{1-0.05/2} = 1.96$, then the power function can be plotted as follows. The power function increases as θ increases, approaching 1 as θ approaches ∞. As n gets larger, the function approaches one very quickly.



8/8

Question 3 **Please see MathStatica output attached for answers to Question 3, parts a, b and c as marked on the output.**

10/10

Question 4

(a) The order statistics, $X_{(1)} < X_{(2)} < X_{(3)}$, are based on a random sample size with $n = 3$, from the standard exponential family distribution with density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF of this function is

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \int_0^x e^{-x}dx = 1 - e^{-x} & x > 0 \end{cases}$$

(a) Using Theorem 7.3, given the order statistics are from a continuous population with cdf $F(X)$ and pdf $f(x)$ as defined, then the pdf of $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^{j-1}[1 - F(x)]^{n-j}$$

The pdf of $X_{(2)}$, is

$$f_{X_{(2)}}(x) = \frac{3!}{(2-1)!(3-2)!}(e^{-x})[1 - e^{-x}]^{2-1}[1 - (1 - e^{-x})]^{3-2}, x > 0$$

$$= \frac{3!}{1!^2}(e^{-x})[1 - e^{-x}](e^{-x})$$

$$= \frac{3 \times 2 \times 1}{1 \times 1}(e^{-2x})[1 - e^{-x}]$$

$$= 6[e^{-2x} - e^{-3x}] \quad \checkmark$$

With,

$$f_{X_{(2)}}(x) = \begin{cases} 6[e^{-2x} - e^{-3x}] & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

We can now find the expectation of the second order statistic,

$$E[X_{(2)}] = \int_0^\infty 6x[e^{-2x} - e^{-3x}]dx$$

Because of linearity we can evaluate the following integral terms separately

$$= \int_0^\infty 6xe^{-2x}dx - \int_0^\infty 6xe^{-3x}dx, \ (1)$$

First, solving $\int_0^\infty 6xe^{-2x}dx$ by integration by parts ($\int uv' = uv - \int u'v$),

$$\mathbf{u} = x, \mathbf{u'} = \frac{d}{dx}x = 1, \mathbf{v'} = e^{-2x}, \mathbf{v} = \int_0^\infty e^{-2x}dx = -\frac{e^{-2x}}{2}$$

Then this integral becomes,

$$= -\frac{xe^{-2x}}{2} - \int_0^\infty -\frac{e^{-2x}}{2}dx$$

$$= -\frac{xe^{-2x}}{2} + \frac{1}{2}\int_0^\infty e^{-2x}dx$$

$$= -\frac{xe^{-2x}}{2} + \frac{1}{2} \times -\frac{e^{-2x}}{2}$$
$$= -\frac{xe^{-2x}}{2} - \frac{e^{-2x}}{4}$$

Then, solving $\int_0^\infty 6xe^{-3x}dx$, in a similar manner

$$\mathbf{u} = x, \mathbf{u'} = \frac{d}{dx}x = 1, \mathbf{v'} = e^{-3x}, \mathbf{v} = \int_0^\infty e^{-3x}dx = -\frac{e^{-3x}}{3}$$

Then this integral becomes,

$$= -\frac{xe^{-3x}}{3} - \int_0^\infty -\frac{e^{-3x}}{3}dx$$
$$= -\frac{xe^{-3x}}{3} + \frac{1}{3}\int_0^\infty e^{-3x}dx$$
$$= -\frac{xe^{-3x}}{3} + \frac{1}{3} \times -\frac{e^{-3x}}{3}$$
$$= -\frac{xe^{-3x}}{3} - \frac{e^{-3x}}{9}$$

Plugging these integrals back into (1) and evaluating

$$= 6[-\frac{xe^{-2x}}{2} - \frac{e^{-2x}}{4} - (-\frac{xe^{-3x}}{3} - \frac{e^{-3x}}{9})]_0^\infty$$
$$= [-3xe^{-2x} - \frac{3}{2}e^{-2x} + 2xe^{-3x} + \frac{2}{3}e^{-3x}]_0^\infty$$
$$= -(-\frac{3}{2}e^0 + \frac{2}{3}e^0)$$
$$= \frac{3}{2} - \frac{2}{3}$$
$$= \frac{5}{6}$$ ✓

(b) First, we can find the joint density of $X_{(1)}$ and $X_{(n)}$ using Theorem 7.3,

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_x(u)f_x(v)[F_x(u)]^{i-1}[F_x(v)-F_x(u)]^{j-1-i}[1-F_x(v)]^{n-j}, 0 < u < v < \infty$$

$$= \frac{3!}{0!1!0!} f_x(x_{(1)})f_x(x_{(3)})[F_x(x_{(3)}) - F_x(x_{(1)})]$$
$$= 6(e^{-x_{(1)}})(e^{-x_{(3)}})[1 - e^{-x_{(3)}} - (1 - e^{-x_{(1)}})]$$
$$= 6e^{-x_{(1)}-x_{(3)})}[e^{-x_{(1)}} - e^{-x_{(3)}}]$$
$$f_{X_{(1)},X_{(n)}}(x_{(1)},x_{(3)}) = 6[e^{-2x_{(1)}-x_{(3)}} - e^{-x_{(1)}-2x_{(3)}}], 0 < x_{(1)} < x_{(3)} < \infty$$ ✓

We are given the transformation,

$$B = \frac{X_{(1)} + X_{(n)}}{2} = \frac{X_{(1)} + X_{(3)}}{2}$$

and defining

$$W = X_{(1)}$$

then solving for $X_{(1)}$ and $X_{(3)}$ yield

$$X_{(1)} = W$$
$$X_{(3)} = 2B - X_{(1)} = 2B - W$$

The value of the Jacobian is

$$\mathbf{J} = \begin{bmatrix} \frac{\partial X_{(1)}}{\partial B} = 0 & \frac{\partial X_{(1)}}{\partial W} = 1 \\ \frac{\partial X_{(3)}}{\partial B} = 2 & \frac{\partial X_{(3)}}{\partial W} = -1 \end{bmatrix}$$

The determinant of this matrix is

$$\frac{\partial X_{(1)}}{\partial B} \times \frac{\partial X_{(3)}}{\partial W} - \frac{\partial X_{(1)}}{\partial W} \times \frac{\partial X_{(3)}}{\partial B} = |(0 \times 1) - (2 \times 1)| = |-2|$$

Therefore, the joint density of B and W becomes

$$f_{B,W}(b,w) = 6[e^{-2W-(2B-W)} - e^{-W-2(2B-W)}]|2|$$

$$f_{B,W}(b,w) = 12[e^{-2B-W} - e^{-4B+W}]$$

The relationship $0 < x_{(1)} < x_{(3)} < \infty$ transfers into $0 < W < 2B - W < \infty$, which is equivalent to the domain of $0 < W < B$ for W. Hence the density of $B = \frac{X_{(1)}+X_{(3)}}{2}$ is

$$f_B(b) = \int_0^B 12[e^{-2B-W} - e^{-4B+W}]dW$$

Applying the linearity rule:

$$\begin{aligned} &= 12e^{-2B}\int_0^B e^{-W}dW - 12e^{-4B}\int_0^B e^{W}dW \\ &= 12e^{-2B}(-e^{-B}+1) - 12e^{-4B}(e^{B}-1) \\ &= 12e^{-2B} - 12e^{-3B} + 12e^{-4B} - 12e^{-3B} \\ &= 12e^{-4B} + 12e^{-2B} - 24e^{-3B} \end{aligned}$$

Now, evaluating $P[B > 2]$,

$$P[B > 2] = \int_2^\infty 12e^{-4B} + 12e^{-2B} - 24e^{-3B}dB$$

$$\begin{aligned} &= [-\frac{1}{4}12e^{-4B} - \frac{1}{2}12e^{-2B} + \frac{1}{3}24e^{-3B}]_2^\infty \\ &= -(-3e^{-4(2)} - 6e^{-2(2)} + 8e^{-3(2)}) \\ &= 3e^{-8} + 6e^{-4} - 8e^{-6} \end{aligned}$$ ✓

Therefore,

$$P[B > 2] = 0.09107$$ ✓

QUESTION 3 (a)

In[12]:=
```
<< mathStatica.m
f = PDF[NormalDistribution[], x]; domain[f] = {x, -∞, ∞};
g = OrderStat[r, f, 3]
domain[g] = OrderStatDomain[r, f, 3]
g /. r → {1, 2, 3} // Simplify
Prob[y, g /. r → {1, 2, 3}]
PlotDensity[g /. r → {1, 2, 3}]
Expect[x, g /. r → {1, 2, 3}]
```

Out[14]=
$$\frac{3\, e^{-\frac{x^2}{2}} \left(1 - \text{Erf}\left[\frac{x}{\sqrt{2}}\right]\right)^{3-r} \left(1 + \text{Erf}\left[\frac{x}{\sqrt{2}}\right]\right)^{-1+r}}{2\sqrt{2\pi}\ (3-r)!\ (-1+r)!}$$

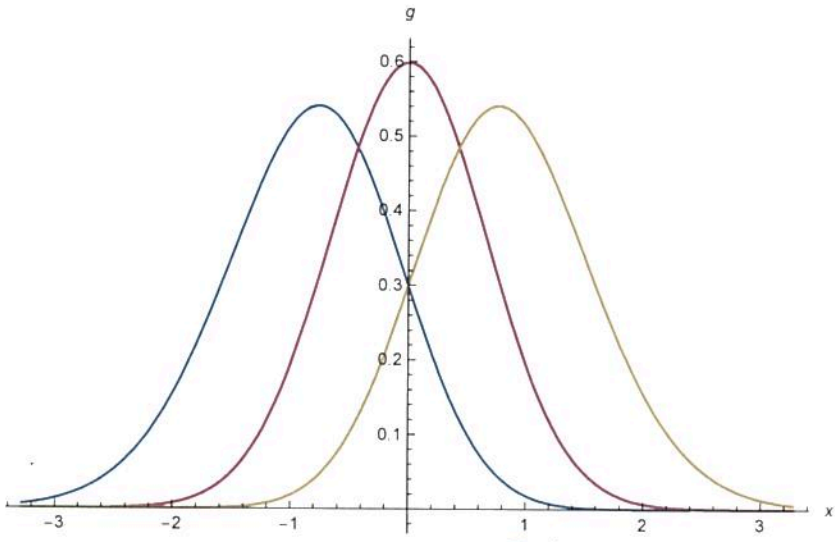
Out[15]=
$$\{x, -\infty, \infty\} \;\&\&\; \{r \in \mathbb{Z},\ 1 \le r \le 3\}$$

Out[16]=
$$\left\{\frac{3\, e^{-\frac{x^2}{2}} \left(-1 + \text{Erf}\left[\frac{x}{\sqrt{2}}\right]\right)^2}{4\sqrt{2\pi}},\ -\frac{3\, e^{-\frac{x^2}{2}} \left(-1 + \text{Erf}\left[\frac{x}{\sqrt{2}}\right]^2\right)}{2\sqrt{2\pi}},\ \frac{3\, e^{-\frac{x^2}{2}} \left(1 + \text{Erf}\left[\frac{x}{\sqrt{2}}\right]\right)^2}{4\sqrt{2\pi}}\right\}$$

→ PART i.

Out[17]=
$$\left\{1 + \frac{1}{8}\left(-1 + \text{Erf}\left[\frac{y}{\sqrt{2}}\right]\right)^3,\ -\frac{1}{4}\left(-2 + \text{Erf}\left[\frac{y}{\sqrt{2}}\right]\right)\left(1 + \text{Erf}\left[\frac{y}{\sqrt{2}}\right]\right)^2,\ \frac{1}{8}\left(1 + \text{Erf}\left[\frac{y}{\sqrt{2}}\right]\right)^3\right\}$$

Out[18]=



PART ii.

Out[19]=
$$\left\{-\frac{3}{2\sqrt{\pi}},\ 0,\ \frac{3}{2\sqrt{\pi}}\right\}$$

PART iii.

✓

QUESTION 3 (b)

In[1]:=
```
<< mathStatica.m
f = PDF[LaplaceDistribution[], x]; domain[f] = {x, -∞, ∞};
g = OrderStat[r, f, 3]
domain[g] = OrderStatDomain[r, f, 3]
g /. r → {1, 2, 3} // Simplify
Prob[y, g /. r → {1, 2, 3}]
PlotDensity[g /. r → {1, 2, 3}]
Expect[x, g /. r → {1, 2, 3}]
```

Out[3]=
$$\begin{cases} \frac{3\, e^{r x} \left(2 - e^{x}\right)^{3-r}}{4\, (3-r)!\, (-1+r)!} & x \le 0 \\ \frac{3\, e^{(3-r) x} \left(2 - e^{x}\right)^{r}}{4 \left(-1 + 2\, e^{x}\right) (3-r)!\, (-1+r)!} & \text{True} \end{cases}$$

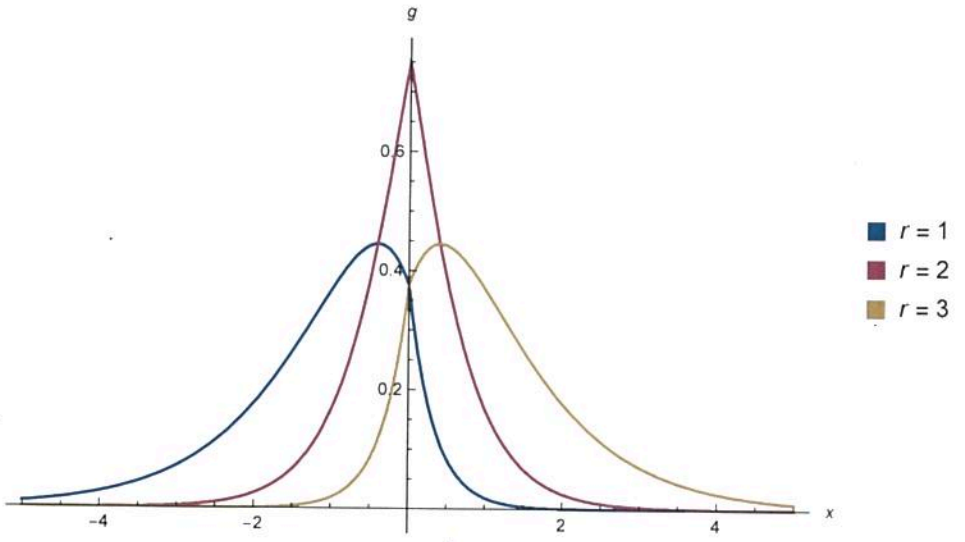
Out[4]=
$$\{x, -\infty, \infty\} \;\&\&\; \{r \in \mathbb{Z}, 1 \le r \le 3\}$$

Out[5]=
$$\begin{cases} \left\{ \frac{3}{8} e^{x} \left(-2 + e^{x}\right)^{2}, -\frac{3}{4} e^{2x} \left(-2 + e^{x}\right), \frac{3 e^{3x}}{8} \right\} & x \le 0 \\ \left\{ \frac{3\, e^{3x}}{8}, \frac{3}{4} e^{-3x} \left(-1 + 2\, e^{x}\right), \frac{3}{8} e^{-3x} \left(1 - 2\, e^{x}\right)^{2} \right\} & \text{True} \end{cases}$$

PART i

Out[6]=
$$\begin{cases} \begin{pmatrix} \frac{1}{8} e^{y} \left(12 + e^{y} \left(-6 + e^{y}\right)\right) & -\frac{1}{4} e^{2y} \left(-3 + e^{y}\right) & \frac{e^{3y}}{8} \\ \frac{1}{8} e^{y} \left(12 + e^{y} \left(-6 + e^{y}\right)\right) & -\frac{1}{4} e^{2y} \left(-3 + e^{y}\right) & \frac{e^{3y}}{8} \\ \frac{1}{8} e^{y} \left(12 + e^{y} \left(-6 + e^{y}\right)\right) & -\frac{1}{4} e^{2y} \left(-3 + e^{y}\right) & \frac{e^{3y}}{8} \end{pmatrix} & y \le 0 \\ \begin{pmatrix} 1 - \frac{e^{3y}}{8} & \frac{1}{4} \left(4 + e^{-3y} - 3\, e^{-2y}\right) & \frac{1}{8} e^{-3y} \left(-1 + 2\, e^{y}\right)^{3} \\ 1 - \frac{e^{3y}}{8} & \frac{1}{4} \left(4 + e^{-3y} - 3\, e^{-2y}\right) & \frac{1}{8} e^{-3y} \left(-1 + 2\, e^{y}\right)^{3} \\ 1 - \frac{e^{3y}}{8} & \frac{1}{4} \left(4 + e^{-3y} - 3\, e^{-2y}\right) & \frac{1}{8} e^{-3y} \left(-1 + 2\, e^{y}\right)^{3} \end{pmatrix} & \text{True} \end{cases}$$

Out[7]=



⇒ PART ii

Out[8]=
$$\left\{-\frac{9}{8}, 0, \frac{9}{8}\right\}$$

⇒ PART iii

✓

The University of New South Wales

Department of Statistics

Session 2, 2018

MATH5905 - Statistical Inference

Assignment 2

The assignment must be submitted by 5PM on Friday, 12th October 2018 at the latest (i.e., at the beginning of the lecture in week 11). Please, declare on the first page that the assignment is your own work, except where acknowledged. State also that you have read and understood the University Rules in respect to Academic Misconduct.

1) Let $X = (X_1, X_2, \ldots, X_n)$ be a sample of n observations each with a uniform in $[0, \theta)$ density

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} \text{ if } 0 < x < \theta \\ 0 \text{ else} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Denote the joint density by $L(X, \theta)$.

a) Show that the family $\{L(X, \theta)\},\ \theta > 0$ has a monotone likelihood ratio in $X_{(n)}$.

b) Show that the uniformly most powerful α-size test of $H_0 : \theta \leq 2$ versus $H_1 : \theta > 2$ is given by

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } X_{(n)} > 2(1-\alpha)^{\frac{1}{n}} \\ 0 \text{ if } X_{(n)} \leq 2(1-\alpha)^{\frac{1}{n}} \end{cases}$$

c) Find the power function of the test and sketch the graph of $E_\theta \varphi^*$ as accurately as possible.

d) Show that the random variable $Y_n = n(1 - \frac{X_{(n)}}{\theta})$ converges in distribution to the exponential distribution with mean 1 as $n \to \infty$. Hence justify that $X_{(n)}$ is a consistent estimator of θ.

2) Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ be i.i.d. random variables, each with a density

$$f(x, \theta) = \begin{cases} \frac{2}{\theta} x e^{-\frac{x^2}{\theta}}, x > 0 \\ 0 \text{ elsewhere} \end{cases}$$

where $\theta > 0$ is a parameter. (This is called the Raleigh-distribution.)

a) Find the Fisher information about θ in one observation and in the sample of n observations.

b) Find the MLE of θ. Is it unbiased? If YES, does its variance attain the Cramer Rao bound?

c) What is the asymptotic distribution of the MLE of θ?

d) Does that the family $L(\mathbf{X}, \theta)$ has a monotone likelihood ratio? If YES, in which statistic?

e) Does a UMP α test of $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ exist? If YES, outline its structure. Also, using asymptotic arguments (e.g., from c)), find the threshold constant in the definition of the test.

f) Calculate (asymptotic approximation to) the power function $E_\theta \varphi^*$ and sketch a graph.

3. (Use `mathStatica` **only** to solve parts a) and b) of this problem.) During the lab in week 7, capabilities of `mathStatica` to deal with distributions of order statistics have been demonstrated. Even more demonstrations can be found in Section 9.4 of the online `MathStatica` textbook (accessible within MATHEMATICA). Examining the files `5905demo_2018.nb` and `intrographshort.nb`) in the computing subfolder on moodle and possibly examining the material of Section 9.4, solve the following problems and attach the `mathStatica` printout to your assignment:

a) Let $X_{(1)}, X_{(2)}, X_{(3)}$ denote the order statistics of a random sample of size $n = 3$ from $X \sim N(0, 1)$.

i) Obtain the densities of each of these three order statistics. (Note: $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$.)

ii) Plot on a *single* diagram the densities of each order statistic (use the interval (-3.5,3.5)).

iii) Determine $E[X_{(r)}]$ for $r = 1, 2, 3$. In particular, you will check in this way the answer to tutorial question 5b) from the Part four exercise sheet.

b) Repeat the steps i), ii), iii) in a) for the case of order statistic of size $n = 3$ from the standard Laplace distribution with density

$$f(x) = \frac{1}{2} \exp(-|x|), x \in (-\infty, \infty).$$

To get a nicer expression about the density of the order statistic, you might need to use the option `FullSimplify`. When plotting on a single diagram, use the interval (-5,5).

4. (For this problem, you are asked to present your analytic derivations. However, if you present a `mathStatica` solution instead, you will get the marks allocated.)

Suppose $X_{(1)} < X_{(2)} < X_{(3)}$ are the order statistics based on a random sample of size 3 from the standard exponential density $f(x) = e^{-x}, x > 0$.

i) Find $E(X_{(2)})$.

ii) Find the density of the midrange $B = \frac{1}{2}(X_{(1)} + X_{(3)})$. Using this result (or otherwise), find $P(B > 2)$.

The University of New South Wales

Department of Statistics

Session 2, 2018

MATH5905 - Statistical Inference

Assignment 1

This assignment must be submitted no later than at the beginning of the lecture at 5pm on Friday, 24th August 2018. Please, declare on the first page that the assignment is your own work, except where acknowledged. State also that you have read and understood the University Rules in respect to Student Academic Misconduct.

The assignment is to be handed in as a **hard copy** (no e-mails!)

Maximal number of pages: 8

.

1. Consider a decision problem with parameter space $\Theta = \{\theta_1, \theta_2\}$ and a set of non randomized decisions $D = \{d_i, 1 \leq i \leq 6\}$ with risk points $\{R(\theta_1, d_i), R(\theta_2, d_i)\}$ as follows:

i	1	2	3	4	5	6
$R(\theta_1, d_i)$	0	1	3	6	7	6
$R(\theta_2, d_i)$	7	5	5	2	5	8

a) Find the minimax rule(s) amongst the **nonrandomized** rules in D;

b) Plot the risk set of all **randomized** rules $\mathcal{D}$ generated by the set of rules in D.

c) Find the risk point of the minimax rule in $\mathcal{D}$ and determine its minimax risk.

d) Define the minimax rule in the set $\mathcal{D}$ in terms of rules in D.

e) For which prior on $\{\theta_1, \theta_2\}$ is the minimax rule in the set $\mathcal{D}$ also a Bayes rule?

f) Determine the Bayes rule and the Bayes risk for the prior $\left(\frac{5}{6}, \frac{1}{6}\right)$ on $\{\theta_1, \theta_2\}$.

g) For a small positive $\epsilon = \frac{1}{2}$, illustrate on the risk set the risk points of all rules which are ϵ-minimax.

2. In a Bayesian estimation problem, we sample n i.i.d. observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a population with conditional distribution of each single observation being the geometric distribution

$$f_{X_1|\Theta}(x|\theta) = \theta^x(1-\theta), x = 0, 1, 2, \ldots; 0 < \theta < 1.$$

The parameter θ is considered as random in the interval $\Theta = (0, 1)$.

i) If the prior on Θ is given by $\tau(\theta) = 3\theta^2, 0 < \theta < 1$, show that the posterior distribution $h(\theta|\mathbf{X} = (x_1, x_2, \ldots, x_n))$ is in the Beta family. Hence determine the Bayes estimator of θ with respect to quadratic loss.

Hint: For $\alpha > 0$ and $\beta > 0$ the beta function $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ satisfies $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ where $\Gamma(\alpha) = \int_0^\infty exp(-x)x^{\alpha-1}dx$. A Beta (α, β) distributed random variable X has a density $f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$, with $E(X) = \alpha/(\alpha+\beta)$.

ii) Five observations form this distribution were observed: 2, 4, 7, 1, 3. Using zero-one loss, what is your decision when testing $H_0 : \theta \leq 0.75$ against $H_1 : \theta > 0.75$. (You may use the `integrate` function in `R` or another numerical integration routine from your favourite programming package to answer the question.)

3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. uniform in $(0, \theta)$ and let the prior on θ be the Pareto prior given by $\tau(\theta) = \beta\alpha^\beta\theta^{-(\beta+1)}, \theta > \alpha$. (Here $\alpha > 0$ and $\beta > 0$ are assumed to be known constants). Show that the Bayes estimator with respect to quadratic loss is given by $\hat{\theta}_{Bayes} = \max(\alpha, x_{(n)})\frac{n+\beta}{n+\beta-1}$. Justify all steps in the derivation.

4. At a critical stage in the development of a new aeroplane in the UK, a decision must be taken to continue or to abandon the project. The financial viability of the project can be measured by a parameter $\theta \in (0, 1)$, the project being profitable if $\theta > \frac{1}{2}$. Data x provide information about θ. If $\theta < 1/2$, the cost to the taxpayer of continuing the project is $(\frac{1}{2} - \theta)$(in units of \$ billion) wheres if $\theta > 1/2$ it is zero (since the project will be privatised if profitable). If $\theta > \frac{1}{2}$ the cost of abandoning the project is $(\theta - \frac{1}{2})$ (due to contractual arrangements for purchasing the aeroplane from the French), whereas if $\theta < \frac{1}{2}$ it is zero. Two actions are on the table: a_0 : continue the project and a_1 : abandon it.

Derive the Bayesian decision rule in terms of the posterior mean of θ given x. The Minister of Aviation has prior density $6\theta(1-\theta)$ for θ. The prime minister has prior density $4\theta^3$. The prototype aeroplane is subjected to trials, each independently having probability θ of success, and the data x consists of the total number of trials required for the first successful result to be obtained (i.e., one realization from a geometric distribution). For which values of x will there be most serious ministerial disagreement?

My white board writing from week 2

Discussion of the decision theoretic concepts, even in a greater detail than I presented them, is to be found in the copy of the chapter from YS that I posted on moodle. Hence I abstain from reproducing these again here (the minimalist student only needs my discussion from the notes in Lecture 2). However I will discuss in more detail the white board writing related to the

Example 2.5.8:

a) Given that X only can have 3 different values (0,1,2) it is clear that only 8 non-randomized decision rules given that the convex hull is the smallest convex set containing the risk points $\begin{pmatrix} R(\theta_1, d_i) \\ R(\theta_2, d_i) \end{pmatrix}$, $i = 1, 2, \dots, 8$ we see that it given as shown on the graph. To illustrate the calculation of the (X,Y) coordinates, we note that, for example:

$R(\theta_1, d_1) = L(\theta_1, a_1) * (.81) + L(\theta_1, a_1) * (.18) + L(\theta_1, a_1) * (.01) = 0 + 0 + 0 = 0$

$R(\theta_2, d_1) = L(\theta_2, a_1) * (.25) + L(\theta_2, a_1) * (.5) + L(\theta_2, a_1) * (.25) = 3 * 1 = 3$

hence the risk point that corresponds to d_1 is with coordinates (0,3).

Similarly, for d_2:

$R(\theta_1, d_2) = L(\theta_1, a_1) * (.81) + L(\theta_1, a_1) * (.18) + L(\theta_1, a_2) * (.01) = 0 + 1 * (.01) = .01$

$R(\theta_2, d_2) = L(\theta_2, a_1) * (.25) + L(\theta_2, a_1) * (.5) + L(\theta_2, \theta_2) * (.25) = 3 * (.25 + .5) = 3 * (.75) = 2.25$

hence the risk point that corresponds to d_2 is with coordinates (0.01, 2.25).

You can work out the rest in a similar way.

For a prior $(p_1, p_2) = (p_1, 1 - p_1)$ on (θ_1, θ_2), the risk points that have the same value b of their

Bayes risk are on the line with equation

$p_1 X + p_2 y = b$, or equivalently $p_1 X + (1-p_1) y = b$.

Note that if $X = y$ we get $X = y = b$, hence the X (and equivalently y) coordinate of the intersection of the line $p_1 X + p_2 y = b$ with the line $X = y$ also represents the value of this risk.

b) The minimax rule in the set $\mathcal{D}$ of randomized decision rules is obtained by examining the intersection of the line $y = x$ with the "most south-west" part of the convex risk set. Hence we need to solve the system:

$$\begin{cases} y = x \\ y - 0 = \dfrac{.75 - 0}{.19 - 1}(x - 1) \end{cases}$$

this gives $X = y = \boxed{\frac{25}{32}}$ for the risk point that corresponds to the minimax decision rule δ^* in the set of all randomized decision rules $\mathcal{D}$ that are generated by the set $D = \{d_1, d_2, \ldots, d_8\}$ of the non-randomized decision rules; $\boxed{\frac{25}{32}}$ is also the value of the minimax risk.

Important: if we were only looking for the minimax decision rule in the set D (not $\mathcal{D}$(!)) then the answer is different as follows:

Rule	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
$R(\theta_1, d_i)$	0	.01	.18	.19	.81	.82	.99	1
$R(\theta_2, d_i)$	3	2.25	1.5	.75	2.25	1.5	.75	0
max	3	2.25	1.5	.75	2.25	1.5	.99	1

Now the minimal of these maxima is .75 hence d_4 is the minimax decision rule with a minimax risk of .75 (and this risk is $> \frac{25}{32}$ naturally — think about why(!))

If we want to represent the minimax rule δ^* in the set $\mathcal{D}$ as a randomization of the rules d_4 and d_8, we need to find $\alpha \in (0,1)$ to say that

δ^* chooses d_4 with probability α and d_8 with probability $(1-\alpha)$.

But then $\alpha * (.19) + (1-\alpha) * 1 = \frac{25}{52}$ must hold.

Hence we get $\alpha \approx .641$ and can claim that

$$\delta^* = \begin{cases} \text{choose } d_4 \text{ with probability } .641 \\ \text{choose } d_8 \text{ with probability } .359 \end{cases}$$

c) For the least favourable prior, we need to maximize the Bayes risk when we start manipulating the priors $(p, 1-p)$. Since for any such prior the value of the Bayes risk will be geometrically represented as a X (or equivalently y) coordinate on the line that connects $(0,0)$ with $\left(\frac{25}{52}, \frac{25}{52}\right)$, we obviously maximize when we end up with $\left(\frac{25}{52}, \frac{25}{52}\right)$, i.e. the minimax solution. Here are hence looking for a prior in the form $(p, 1-p)$ for which $\left(\frac{25}{52}, \frac{25}{52}\right)$ would be the Bayes solution.

This means $(p, 1-p)$ to be $\perp$ to the line $d_4 d_8$. This requirement is the same as to ask that the slope $\frac{.75-0}{.19-1} = -\frac{25}{27}$ of this line to be the same as $-\frac{p}{1-p}$ (since the line $px + (1-p)y = \text{const}$ has a slope $-\frac{p}{1-p}$; indeed $y = -\frac{p}{1-p}x + \text{const}$)

Hence we have to satisfy $-\frac{p}{1-p} = -\frac{25}{27} \Rightarrow$ Hence $p = \frac{25}{52}$ and the least favorable prior is $\left(\frac{25}{52}, \frac{27}{52}\right)$.

d) What is the Bayes rule for the prior $(1/3, 2/3)$ over (θ_1, θ_2)?

The line $\frac{1}{3}X + \frac{2}{3}Y = \text{const}$ represents points (X, Y) in the risk set with equivalent value of their Bayes risk. The slope of this line is equal to $\frac{-1/3}{2/3} = -\frac{1}{2}$. Hence, to find the Bayes rule w.r. this prior, we need to move lines with a slope of $(-\frac{1}{2})$ "most south-west" while still having an intersection with the risk set. By doing so, you see geometrically that you end up with the rule d_8 (look carefully at the graph). Hence $d_8\,(1,0)$ represents the risk point that corresponds to the Bayesian decision rule w.r. to the prior $(\frac{1}{3}, \frac{2}{3})$ on (θ_1, θ_2). In other words, d_8 is the Bayesian decision rule w.r. to the prior $(\frac{1}{3}, \frac{2}{3})$. Its Bayes risk is equal to:

$$\frac{1}{3} * R(\theta_1, d_8) + \frac{2}{3} R(\theta_2, d_8) = \frac{1}{3} * 1 + \frac{2}{3} * 0 = \boxed{\frac{1}{3}}$$

My white board writing from week 4 – 17th August

1) Example: For $\mathbf{X} = (X_1, X_2, \dots, X_n)$ i.i.d. Bernoulli with parameter θ, i.e. $f(x_i) = P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$, $x_i = 0, 1$

we claim that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

The proof is by inspecting the original definition. We have the following partitions created by T: $\mathcal{A} = (A_0, A_1, A_2, \dots, A_n)$ where $A_r = \{\mathbf{X} \mid \sum_{i=1}^{n} X_i = r\}$ for $r = 0, 1, 2, \dots, n$.

Then $P(\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in A_r) = \dfrac{P(\mathbf{X} = \mathbf{x} \cap \mathbf{X} \in A_r)}{P(\mathbf{X} \in A_r)}$ (*)

Noting that $\mathbf{X} \in A_r$ means that $\sum_{i=1}^{n} X_i = r$ and $\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \theta)$ we have $P(\mathbf{X} \in A_r) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$

and $P(\mathbf{X} = \mathbf{x} \cap \mathbf{X} \in A_r) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} x_i \neq r \\ \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} & \text{when } \sum_{i=1}^{n} x_i = r \end{cases}$

we continue from (*):

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in A_r) = \begin{cases} 0 & \text{if } \sum x_i \neq r \\ \dfrac{\theta^r(1-\theta)^{n-r}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} = \dfrac{1}{\binom{n}{r}} & \text{if } \sum_{i=1}^{n} x_i = r \end{cases}$$

Hence this conditional probability does not depend on θ.

2) Next, I considered the proof of Neyman-Fisher's factorization criterion (for the discrete case only).

There are 2 "directions" to be shown:

($\Leftarrow$) Assuming that $L(\mathbf{x}, \theta) = g(T(\mathbf{x}), \theta)\, h(\mathbf{x})$ holds we need to check that T is sufficient for θ. Looking at

$$P(\mathbf{X} = \mathbf{x} \mid T = t) = \frac{P(\mathbf{X} = \mathbf{x} \cap T = t)}{P(T = t)} = \begin{cases} 0 & \text{if } T(\mathbf{x}) \neq t \\ \dfrac{P(\mathbf{X} = \mathbf{x})}{P(T = t)} & \text{if } T(\mathbf{x}) = t \end{cases}$$

we continue in the second case as follows:

$$\frac{P(\mathbf{X} = \mathbf{x})}{P(T = t)} = \frac{g(t, \theta)\, h(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}: T(\tilde{\mathbf{x}}) = t} g(t, \theta) h(\tilde{\mathbf{x}})} = \frac{h(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}: T(\tilde{\mathbf{x}}) = t} h(\tilde{\mathbf{x}})}$$

which does not involve θ.

($\Rightarrow$) If on the other hand T was sufficient then

$$P_\theta(X=x) = P(X=x \cap \underbrace{T=T(x)}) = P(X=x \mid T=t)\, P_\theta(T=t) =$$

(since this is the sure event; denoting $T(x)=t$)

$= h(x)\, g(t,\theta)$ where we denoted $P(X=x \mid T=t) = h(x)$ and it is known not to depend on θ by assumption, whereas $P_\theta(T=t) = g(t,\theta)$ involves the data via the value of the statistic only, i.e. the factorization is demonstrated.

—

(3) I gave several examples of using the factorization criterion to show sufficiency:

i) For Bernoulli: $T = \sum_{i=1}^n X_i$ is sufficient which follows directly from: $L(X,\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$ which involves the data only via the value of $\sum_{i=1}^n x_i = T$, so the whole RHS can be thought of as $g(t,\theta) = \theta^t (1-\theta)^{n-t}$ (and there is no need of $h(X)$ here, it can be set to the constant 1).

ii) $N(\mu, \sigma^2)$ with $\theta = \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}$.

Using the fundamental equality $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x}-\mu)^2$ we have:

$$L(X;\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x}-\mu)^2\right]\right)$$

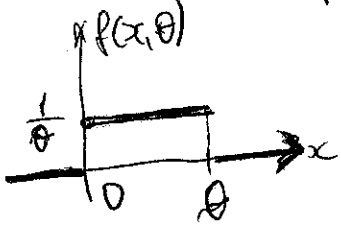
which involves the data via $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}$ only.

Hence this 2-dim vector statistic is sufficient for $\theta = \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}$

I also noted that every 1-to-1 transformation of T is also sufficient. In particular, $\tilde{T} = \begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{pmatrix}$

is also sufficient for θ (since knowing $\begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$ we can get $\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix}$ and vice versa.

iii) $\mathbf{X} = (X_1, X_2, \dots, X_n)$ uniformly distributed in $[0, \theta)$. We claim that $T = X_{(n)}$ (the n-th order statistic, equal to the maximal of the observations) is sufficient for θ.



Note that the density (show graphically on the left) can be written using indicator function as

$$f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta)$$

Then

$$L(\mathbf{X},\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(x_i,\infty)}(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(x_i,\infty)}(\theta) =$$

$$= \boxed{\frac{1}{\theta^n} I_{(x_{(n)},\infty)}(\theta)} = g(x_{(n)},\theta) * 1$$

which represents a factorization and $T = X_{(n)}$ is sufficient according to the factorization criterion.

iv) Multivariate normal:

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be n i.i.d p-dim multivariate normal data vectors $X_i \sim N_p(\mu, \Sigma)$.

We have the p-dim version of the fundamental equality:

$$\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)' = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' + n(\bar{x} - \mu)(\bar{x} - \mu)'$$

We also use properties of traces: $(x'Ax) = \mathrm{tr}(A(xx'))$

Then: $L(\mathbf{X}; \mu, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \mathrm{tr}\, \Sigma^{-1}(x_i - \mu)(x_i - \mu)'\right)$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \mathrm{tr}\left[\Sigma^{-1}\left(\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' + n(\bar{x} - \mu)(\bar{x} - \mu)'\right)\right]\right\}$$

↑ Since tr and $\sum_{i=1}^{n}$ are linear operators and order can be exchanged

which involves the data only via $T_1 = \bar{x}$ and $T_2 = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$

Hence T_1 and T_2 give a sufficient statistic for μ and Σ.

(4) I showed many examples about proving minimal sufficiency via the Lehmann-Scheffe method:

i) For Bernoulli (θ). Take 2 independent n-tuples of data $X=(X_1,X_2,\dots,X_n)$ and $Y=(Y_1,Y_2,\dots,Y_n)$ and write

$$\frac{L(Y,\theta)}{L(X,\theta)} = \frac{\theta^{\sum_{i=1}^n Y_i}(1-\theta)^{n-\sum_{i=1}^n Y_i}}{\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}} = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n Y_i - \sum_{i=1}^n X_i}$$

For this to not depend on θ $\longrightarrow$ can only happen when $\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i$ $\longrightarrow$ Hence $T=\sum_{i=1}^n X_i$ is minimal sufficient

(i.e., the sets in the minimal sufficient partition are the contours of the statistic $T(X)=\sum_{i=1}^n X_i$)

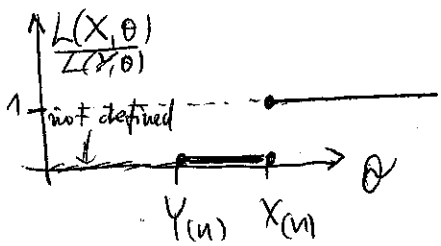
ii) $N(\mu,\sigma^2)$, $\theta=\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}$. It can be seen easily that

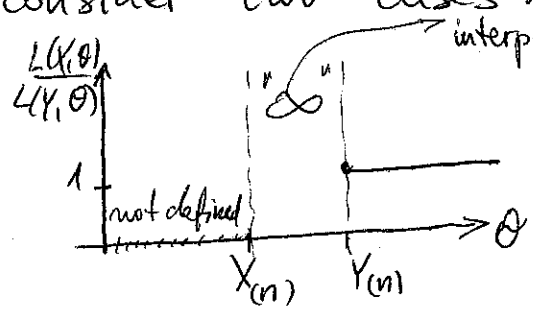
$$\frac{L(Y,\theta)}{L(X,\theta)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n X_i^2 - 2\mu\left(\sum_{i=1}^n Y_i - \sum_{i=1}^n X_i\right)\right)\right)$$

and for this to not depend on θ we need both $\left(\begin{matrix}\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i\\ \sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i^2\end{matrix}\right)$ to hold. Hence $T=\begin{pmatrix}\sum_{i=1}^n X_i\\ \sum_{i=1}^n X_i^2\end{pmatrix}$ is minimal sufficient for $\theta=\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}$ (as is also any 1-1 transformation of T)

iii) Uniform in $[0,\theta)$. For $\frac{L(X,\theta)}{L(Y,\theta)}$ we get $\frac{I_{(X_{(n)},\infty)}(\theta)}{I_{(Y_{(n)},\infty)}(\theta)}$

This is independent of θ if and only if $X_{(n)}=Y_{(n)}$ which implies that $T=X_{(n)}$ is minimal sufficient. (Indeed, if $X_{(n)} \neq Y_{(n)}$ we can consider two cases:





In both cases when $X_{(n)} \neq Y_{(n)}$, the ratio's value (where defined) depends on the position of θ, i.e., it is not independent of θ. To have it not depending on θ, we need $X_{(n)} = Y_{(n)}$ to hold.)

iv) One more example to show that not necessarily is the dimension of the minimal sufficient statistic equal to the dimension of the parameter (as it was in the previous 3 examples):

If $X_1, X_2, \dots X_n$ are i.i.d. Cauchy(θ) (i.e., with density $f(x,\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, $-\infty < x < \infty$ then

$$\frac{L(Y,\theta)}{L(X,\theta)} = \frac{\prod_{i=1}^{n}(1+(X_i-\theta)^2)}{\prod_{i=1}^{n}(1+(Y_i-\theta)^2)}$$

and we see that unless $\begin{pmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{pmatrix} = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \\ \vdots \\ Y_{(n)} \end{pmatrix}$, the ratio will depend on θ. Hence $T = \begin{pmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{pmatrix}$ is the minimal sufficient statistic in this case (although its dimension is equal to the sample size, so virtually no dimension reduction is possible in this case)

5) Next, I showed several examples of one-parameter exponential families and explained the related minimal sufficient statistics for them:

i) $f(x,\theta) = \theta \exp(-\theta x)$: $a(\theta)=\theta$, $b(x)=1$, $c(\theta)=-\theta$, $d(x)=x$
$\longrightarrow T = \sum_{i=1}^{n} X_i$ is minimal sufficient

ii) $f(x,\theta) = \frac{e^{-\theta}\theta^x}{x!} = e^{-\theta}\frac{1}{x!}e^{x\ln\theta}$ $\longrightarrow$ $\begin{cases} a(\theta)=e^{-\theta} \\ b(x)=\frac{1}{x!} \\ c(\theta)=\ln\theta \\ d(x)=x \end{cases}$

Hence $T=\sum_{i=1}^{n} x_i$ is minimal sufficient

iii) ; etc. $\rightarrow$ for your own exercise

iv) $N(0,\theta^2) \longrightarrow f(x,\theta) = \frac{1}{\sqrt{2\pi}\,\theta}\cdot e^{-\frac{1}{2\theta^2}x^2}$

Hence $a(\theta)=\frac{1}{\sqrt{2\pi}\,\theta}$, $b(x)=1$, $c(\theta)=-\frac{1}{2\theta^2}$, $d(x)=x^2$

Hence $T=\sum_{i=1}^{n} x_i^2$ is minimal sufficient

The generalization for K-parameter exponential families:

Example: $N(\mu,\sigma^2)$:

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

is 2 par. exponential and you can choose $d_1=x$ $d_2=x^2$ to end up with $\begin{pmatrix}\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2\end{pmatrix}$ as a minimal sufficient.

I also left it for you to convince yourself that

$$f(x;\theta_1,\theta_2) = \frac{1}{B(\theta_1,\theta_2)}x^{\theta_1-1}(1-x)^{\theta_2-1}, \quad x\in(0,1), \quad \theta_1,\theta_2>0$$

(the Beta density) belongs to a 2 parameter exponential family and a minimal sufficient statistic for $\theta=\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}$ is $\begin{pmatrix}\sum_{i=1}^{n}\ln x_i \\ \sum_{i=1}^{n}\ln(1-x_i)\end{pmatrix}$.

6) Then I moved over to the ancillarity principle in inference. I abstain from reproducing the discussion here since it is thorough enough in the notes. I just summarize my discussion about the Pitman estimator $\hat{\theta}_p$.

The claim is that if you consider equivariant estimators $\hat{\theta}$ (i.e. satisfying $\hat{\theta}(X_1+c, X_2+c, \dots, X_n+c) = \hat{\theta}(X_1, X_2, \dots, X_n)+c$) when dealing with location parameter estimation then you can find <u>the best</u> equivariant estimator with respect to mean-squared error (that is in the class of these estimators, there is one particular one (namely $\hat{\theta}_p$) which minimizes $E_\theta(\theta-\hat{\theta})^2$ for all θ! (i.e. has uniformly smallest risk with respect to quadratic loss). What is interesting is that in its construction you utilise the ancillary statistic $\tilde{T}_2 = (X_2-X_1, X_3-X_1, \dots, X_n-X_1)$. Starting with an <u>arbitrary</u> equivariant estimator $\tilde{\theta}$, you construct $\hat{\theta}_p$ as $\hat{\theta}_p = \tilde{\theta} - E_0(\tilde{\theta}/\tilde{T}_2)$. When the loss is quadratic, you end up with $\hat{\theta}_p$ possessing the above optimality. It turns out that

$$\hat{\theta}_p = -\frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^{n} f(X_i-\theta)\,d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(X_i-\theta)\,d\theta}$$

which you can obviously interpret as a Bayes estimator w.r. quadratic loss and w.r. <u>improper</u> prior $\pi(\theta)\equiv 1$ on R^1

Of course, this prior is improper because it is not a density over $(-\infty, \infty)$ but you can exploit the analogy.

Finally, I also convinced you that if the location family $f_\theta(x) = f(x-\theta)$ we were dealing with was the $N(\theta, 1)$ family then $\hat{\theta}_p$, as discussed above, is the familiar $\bar{X}$ (the arithmetic mean).

Indeed: $$\hat{\theta}_p = \frac{\int_{-\infty}^{\infty} \theta e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} d\theta}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} d\theta}$$

Within the integrals, We can multiply with any factor on top and bottom as long as it does not involve the θ. Hence we can write:

$$\hat{\theta}_p = \frac{\int_{-\infty}^{\infty} \theta e^{-\frac{n}{2}\theta^2 + \theta\sum_{i=1}^{n} x_i} d\theta}{\int_{-\infty}^{\infty} e^{-\frac{n}{2}\theta^2 + \theta\sum_{i=1}^{n} x_i} d\theta}$$

multiply by $e^{-\frac{n}{2}(\bar{x})^2}$ to complete the square

$$= \frac{\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \theta e^{-\frac{n}{2}(\theta-\bar{x})^2} d\theta}{\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{n}{2}(\theta-\bar{x})^2} d\theta}$$

⇐ If we interprete θ as a random variable with $\theta \sim N(\bar{x}, \frac{1}{n})$

then on top we have written the expected value of this variable (which is $\bar{x}$) and on bottom we have integrated out the density of this random variable (which gives us 1). Hence the ratio is equal to $\frac{\bar{x}}{1} = \bar{x}$.

I also mentioned that Pitman is a famous AUSTRALIAN statistician

Table of Common Distributions

Discrete Distributions

Bernoulli(p)

pmf	$P(X=x\|p)=p^x(1-p)^{1-x};\quad x=0,1;\quad 0\le p\le 1$
mean and variance	$\mathrm{E}X=p,\quad \mathrm{Var}\,X=p(1-p)$
mgf	$M_X(t)=(1-p)+pe^t$

Binomial(n,p)

pmf	$P(X=x\|n,p)=\binom{n}{x}p^x(1-p)^{n-x};\quad x=0,1,2,\ldots,n;\quad 0\le p\le 1$
mean and variance	$\mathrm{E}X=np,\quad \mathrm{Var}\,X=np(1-p)$
mgf	$M_X(t)=[pe^t+(1-p)]^n$
notes	Related to Binomial Theorem (Theorem 3.2.2). The *multinomial* distribution (Definition 4.6.2) is a multivariate version of the binomial distribution.

Discrete uniform

pmf	$P(X=x\|N)=\frac{1}{N};\quad x=1,2,\ldots,N;\quad N=1,2,\ldots$
mean and variance	$\mathrm{E}X=\frac{N+1}{2},\quad \mathrm{Var}\,X=\frac{(N+1)(N-1)}{12}$
mgf	$M_X(t)=\frac{1}{N}\sum_{i=1}^{N}e^{it}$

Geometric(p)

pmf	$P(X=x\|p)=p(1-p)^{x-1};\quad x=1,2,\ldots;\quad 0\le p\le 1$
mean and variance	$\mathrm{E}X=\frac{1}{p},\quad \mathrm{Var}\,X=\frac{1-p}{p^2}$

mgf $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad t < -\log(1-p)$

notes $Y = X - 1$ is negative binomial$(1,p)$. The distribution is *memoryless*: $P(X > s|X > t) = P(X > s - t)$.

Hypergeometric

pmf $P(X = x|N,M,K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0,1,2,\ldots,K;$
$M - (N-K) \le x \le M; \quad N,M,K \ge 0$

mean and variance $\mathrm{E}X = \frac{KM}{N}, \quad \mathrm{Var}\,X = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$

notes If $K \ll M$ and N, the range $x = 0,1,2,\ldots,K$ will be appropriate.

Negative binomial(r,p)

pmf $P(X = x|r,p) = \binom{r+x-1}{x}p^r(1-p)^x; \quad x = 0,1,\ldots; \quad 0 \le p \le 1$

mean and variance $\mathrm{E}X = \frac{r(1-p)}{p}, \quad \mathrm{Var}\,X = \frac{r(1-p)}{p^2}$

mgf $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t < -\log(1-p)$

notes An alternate form of the pmf is given by $P(Y = y|r,p) = \binom{y-1}{r-1}p^r(1-p)^{y-r}$, $y = r, r+1, \ldots$. The random variable $Y = X + r$. The negative binomial can be derived as a gamma mixture of Poissons. (See Exercise 4.34.)

Poisson(λ)

pmf $P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0,1,\ldots; \quad 0 \le \lambda < \infty$

mean and variance $\mathrm{E}X = \lambda, \quad \mathrm{Var}\,X = \lambda$

mgf $M_X(t) = e^{\lambda(e^t-1)}$

Continuous Distributions

***Beta*(α, β)**

pdf $f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \quad \alpha > 0, \quad \beta > 0$

mean and variance $\mathrm{E}X = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

mgf $M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$

notes The constant in the beta pdf can be defined in terms of gamma functions, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Equation (3.2.18) gives a general expression for the moments.

***Cauchy*(θ, σ)**

pdf $f(x|\theta,\sigma) = \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty; \quad -\infty < \theta < \infty, \quad \sigma > 0$

mean and variance do not exist

mgf does not exist

notes Special case of Student's t, when degrees of freedom $= 1$. Also, if X and Y are independent n$(0,1)$, X/Y is Cauchy.

***Chi squared*(p)**

pdf $f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}; \quad 0 \le x < \infty; \quad p = 1, 2, \ldots$

mean and variance $\mathrm{E}X = p, \quad \mathrm{Var}\, X = 2p$

mgf $M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \quad t < \frac{1}{2}$

notes Special case of the gamma distribution.

***Double exponential*(μ, σ)**

pdf $f(x|\mu,\sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = 2\sigma^2$

mgf $M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2}, \quad |t| < \frac{1}{\sigma}$

notes Also known as the *Laplace* distribution.

Exponential(β)

pdf $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \beta, \quad \mathrm{Var}\, X = \beta^2$

mgf $M_X(t) = \frac{1}{1-\beta t}, \quad t < \frac{1}{\beta}$

notes Special case of the gamma distribution. Has the *memoryless* property. Has many special cases: $Y = X^{1/\gamma}$ is *Weibull*, $Y = \sqrt{2X/\beta}$ is *Rayleigh*, $Y = \alpha - \gamma \log(X/\beta)$ is *Gumbel*.

F

pdf $f(x|\nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{x^{(\nu_1-2)/2}}{\left(1+\left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}};$
$0 \le x < \infty; \quad \nu_1, \nu_2 = 1, \ldots$

mean and variance $\mathrm{E}X = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2,$

$\mathrm{Var}\, X = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4$

moments (mgf does not exist) $\mathrm{E}X^n = \frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_2}{\nu_1}\right)^n, \quad n < \frac{\nu_2}{2}$

notes Related to chi squared ($F_{\nu_1,\nu_2} = \left(\frac{\chi^2_{\nu_1}}{\nu_1}\right) / \left(\frac{\chi^2_{\nu_2}}{\nu_2}\right)$, where the χ^2s are independent) and t ($F_{1,\nu} = t^2_\nu$).

Gamma(α, β)

pdf $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha, \beta > 0$

mean and variance $\mathrm{E}X = \alpha\beta, \quad \mathrm{Var}\, X = \alpha\beta^2$

mgf $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \quad t < \frac{1}{\beta}$

notes Some special cases are exponential ($\alpha = 1$) and chi squared ($\alpha = p/2$, $\beta = 2$). If $\alpha = \frac{3}{2}$, $Y = \sqrt{X/\beta}$ is *Maxwell*. $Y = 1/X$ has the *inverted gamma distribution*. Can also be related to the Poisson (Example 3.2.1).

Logistic(μ, β)

pdf $f(x|\mu, \beta) = \frac{1}{\beta}\frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = \frac{\pi^2\beta^2}{3}$

mgf	$M_X(t) = e^{\mu t}\Gamma(1-\beta t)\Gamma(1+\beta t), \quad \lvert t\rvert < \frac{1}{\beta}$
notes	The cdf is given by $F(x\lvert\mu,\beta) = \frac{1}{1+e^{-(x-\mu)/\beta}}$.

Lognormal(μ, σ^2)

pdf	$f(x\lvert\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}, \quad 0 \le x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$
mean and variance	$\mathrm{E}X = e^{\mu+(\sigma^2/2)}, \quad \mathrm{Var}\,X = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$
moments (mgf does not exist)	$\mathrm{E}X^n = e^{n\mu+n^2\sigma^2/2}$
notes	Example 2.3.5 gives another distribution with the same moments.

Normal(μ, σ^2)

pdf	$f(x\lvert\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$
mean and variance	$\mathrm{E}X = \mu, \quad \mathrm{Var}\,X = \sigma^2$
mgf	$M_X(t) = e^{\mu t+\sigma^2 t^2/2}$
notes	Sometimes called the *Gaussian* distribution.

Pareto(α, β)

pdf	$f(x\lvert\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad a < x < \infty, \quad \alpha > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \frac{\beta\alpha}{\beta-1}, \quad \beta > 1, \quad \mathrm{Var}\,X = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2$
mgf	does not exist

t

pdf	$f(x\lvert\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{\left(1+\left(\frac{x^2}{\nu}\right)\right)^{(\nu+1)/2}}, \quad -\infty < x < \infty, \quad \nu = 1, \ldots$
mean and variance	$\mathrm{E}X = 0, \quad \nu > 1, \quad \mathrm{Var}\,X = \frac{\nu}{\nu-2}, \quad \nu > 2$
moments (mgf does not exist)	$\mathrm{E}X^n = \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{\nu-n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\nu^{n/2}$ if $n < \nu$ and even, $\mathrm{E}X^n = 0$ if $n < \nu$ and odd.
notes	Related to F ($F_{1,\nu} = t_\nu^2$).

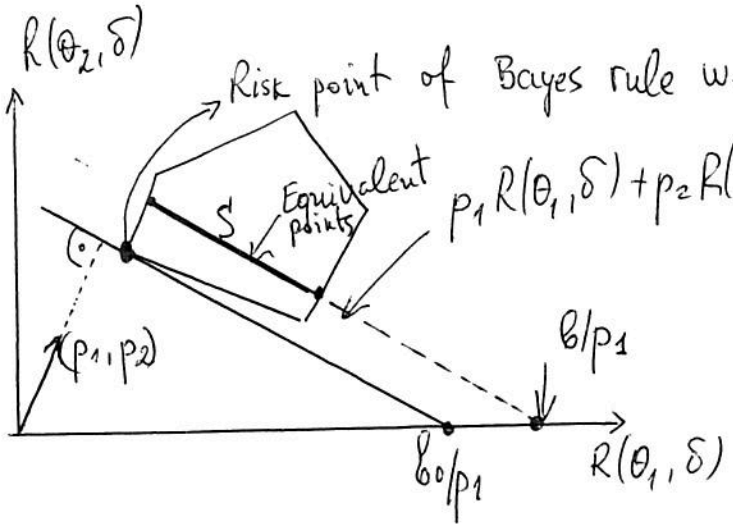
Uniform(a, b)

pdf	$f(x\|a,b) = \frac{1}{b-a}, \quad a \le x \le b$
mean and variance	$\mathrm{E}X = \frac{b+a}{2}, \quad \mathrm{Var}\, X = \frac{(b-a)^2}{12}$
mgf	$M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$
notes	If $a = 0$ and $b = 1$, this is a special case of the beta ($\alpha = \beta = 1$).

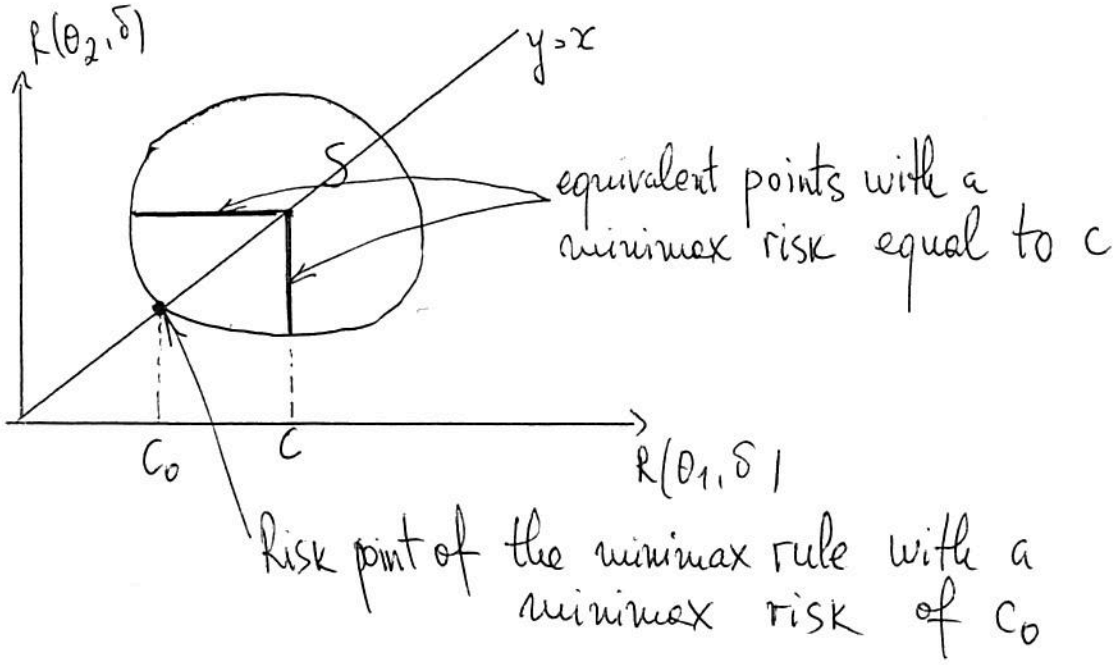
Weibull(γ, β)

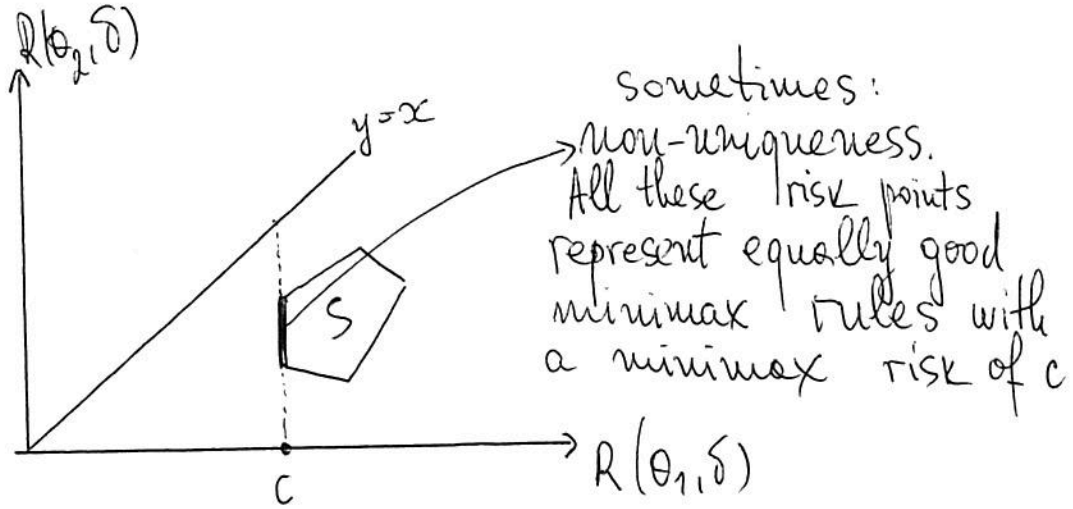
pdf	$f(x\|\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta}, \quad 0 \le x < \infty, \quad \gamma > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right), \quad \mathrm{Var}\, X = \beta^{2/\gamma}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$
moments	$\mathrm{E}X^n = \beta^{n/\gamma}\Gamma\left(1+\frac{n}{\gamma}\right)$
notes	The mgf exists only for $\gamma \ge 1$. Its form is not very useful. A special case is exponential ($\gamma = 1$).

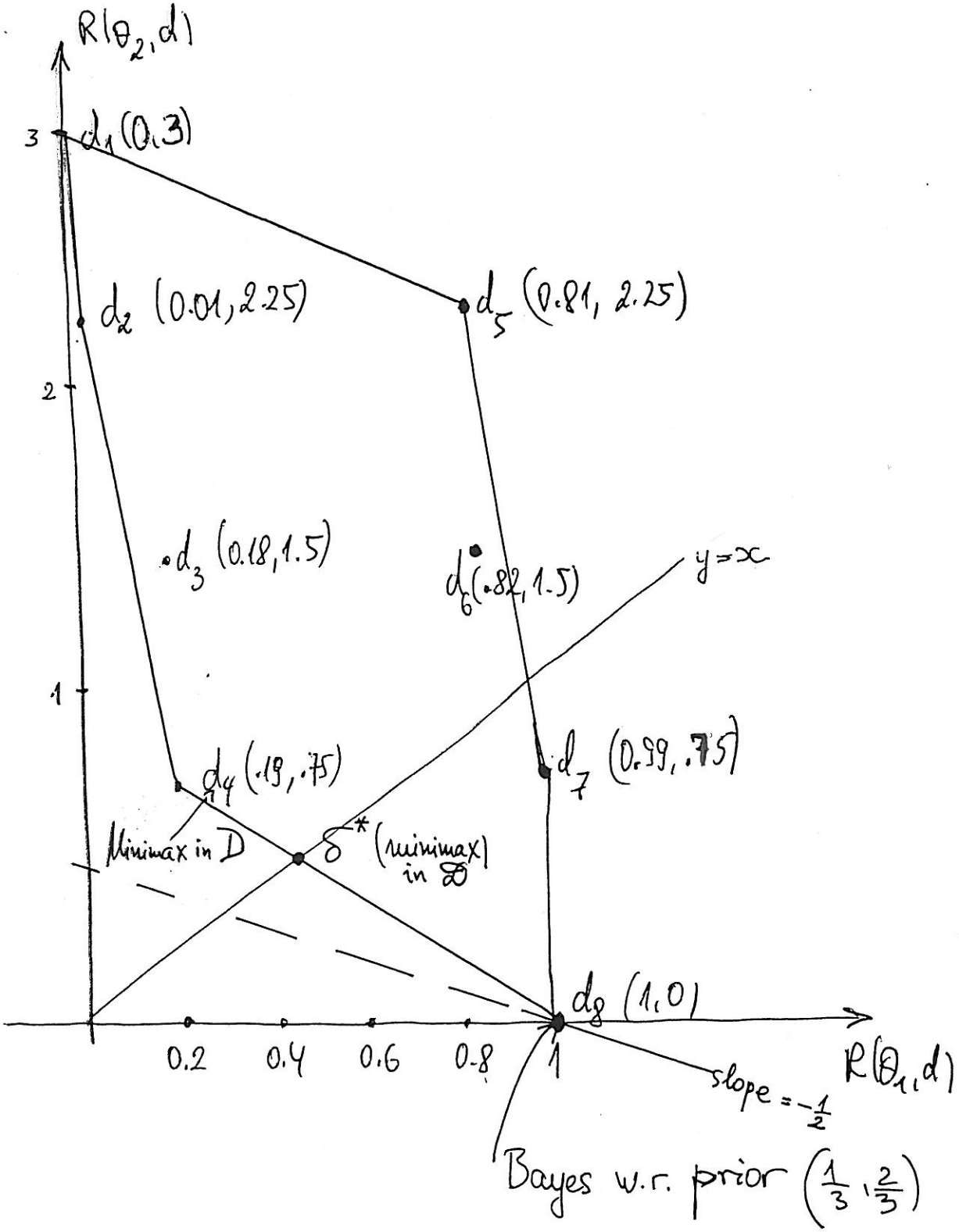
BAYES:



MINIMAX:






R(θ2, d)
3
2
1
d1 (0.3)
d2 (0.01, 2.25)
d5 (0.81, 2.25)
d3 (0.18, 1.5)
d6 (.82, 1.5)
y = x
d4 (.19, .75)
d7 (0.99, .75)
Minimax in D
δ* (minimax in D)
d8 (1, 0)
0.2
0.4
0.6
0.8
1
R(θ1, d)
slope = -1/2
Bayes w.r. prior (1/3, 2/3)

The University of New South Wales

Department of Statistics

School of Mathematics

Statistical Inference, MATH5905

Lecture notes

Associate Professor Spiridon Penev

Foreword

These notes do **not** substitute the lectures in Statistical Inference for masters students. You are strongly recommended to attend each and every lecture because the conceptual bases of the discussed methods, as well as some additional derivations and explanations will then be focused on. This volume is therefore not meant to be a substitute for a textbook, or lecture attendance.

These notes are a compilation from several sources and other notes. Some of the sources are listed in your handout.

I will appreciate if you would let me know about any ways these notes could be further improved.

1 Lecture 1: REVISION: ELEMENTS OF PROBABILITY

Some standard univariate distributions like the binomial, Poisson, normal, Cauchy, logistic, exponential, double exponential (also called Laplace distribution), are assumed to be known. These are summarised in the Table of Common Distributions on pages 621–626 of **CB** and a hard copy is given to everyone.

The revision below involves mainly the following sections of the CB reference:

- **Probability (Sections 1.2, 1.3)** Conditional probability and independence:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A)}{P(B)} P(B|A)$$

is the conditional probability of A given B.

A and B are *independent* if $P(A \bigcap B) = P(A)P(B)$

and if A and B are independent then

$$P(A|B) = P(A).$$

Bayes rule: for a partition $\{A_i\}$ of the sample space S :

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

- **Random variables and distributions (univariate) (Sections 1.4, 2.1)**

Transformations.

Theorem. Let X be a random variable with cdf function $F_X(.)$ and density $f_X(.)$. Let $Y = g(X)$ and $F_Y(.)$ be the cdf of Y. Put

$$S_X = \{x : f_X(x) > 0\}; S_Y = \{y : y = g(x) \text{ for some } x \in S_X.\}$$

i) If g is increasing on S_X then $F_Y(y) = F_X(g^{-1}(y))$ for $y \in S_y$.

ii) If g is decreasing on S_X and X is continuous random variable then $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in S_y$.

Theorem (Probability integral transform) Let X be a continuous random variable with a cdf $F_X(.)$. The the random variable $Y = F_X(X)$ is uniformly distributed on $[0, 1]$.

- **Expectations, Variances and correlations (Sections 2.2, 2.3, 4.5)**

$$E(g(X)) = \int g(x) f_X(x) dx, E(g(X)) = \sum g(x_i) P(X = x_i) = \sum g(x_i) f_X(x_i)$$

for the continuous and for the discrete case, respectively.

Variance: $Var(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2$.

Covariance: $Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$.

Correlation: $\rho = Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$.

- **Multivariate distributions (Sections 4.1, 4.2)**

Random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_p \end{pmatrix} \in R^p$, $p \geq 2$ has p different components each of which is a random variable with a cumulative distribution function *(cdf)* $F_{X_i}(x_i), i = 1, 2, ..,$p. Each of the functions $F_{X_i}(.)$ is called a *marginal distribution.* The *joint cdf* of the random vector $\mathbf{X}$ is

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p) = F_{\mathbf{X}}(x_1, x_2, .., x_p)$$

In case of a *discrete* vector of observations $\mathbf{X}$ the *probability mass function* is defined as

$$P_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

If a *density* $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, .., x_p)$ exists such that

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} .. \int_{-\infty}^{x_p} f_{\mathbf{X}}(\mathbf{t}) dt_1 \ldots dt_p$$

then $\mathbf{X}$ is a *continuous* random vector with a joint density function of p arguments $f_{\mathbf{X}}(\mathbf{x})$. In this case $f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 .. \partial x_p}$ holds. In case $\mathbf{X}$ has p independent components then

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_p}(x_p)$$

holds and, equivalently, also

$$P_{\mathbf{X}}(\mathbf{x}) = P_{X_1}(x_1) P_{X_2}(x_2) \ldots P_{X_p}(x_p), f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_p}(x_p)$$

holds.

The *marginal cdf of the first* $k < p$ components of the vector $\mathbf{X}$ is defined in a natural way as follows:

$$P(X_1 \leq x_1, X_2 \leq x_2, .., X_k \leq x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k, X_{k+1} \leq \infty, ..., X_p \leq \infty)$$

$$= F_{\mathbf{X}}(x_1, x_2, .., x_k, \infty, \infty, .., \infty)$$

The *marginal density* of the first k components can be obtained by partial differentiation:

$$\int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2, .., x_p) dx_{k+1} \ldots dx_p$$

The *conditional density* $\mathbf{X}$ **when** $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$ is defined by

$$f_{(X_1,\ldots,X_r|X_{r+1},\ldots,X_p)}(x_1,\ldots,x_r|x_{r+1},\ldots,x_p) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{X_{r+1},\ldots,X_p}(x_{r+1},\ldots,x_p)}$$

The above conditional density is interpreted as the joint density of $X_1, \ldots, X_r$ when $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$ and is only defined when $f_{X_{r+1},\ldots,X_p}(x_{r+1},\ldots,x_p) \neq 0$.

We note that in case of mutual independence the p components, all conditional distributions do **not** depend on the conditions and it holds

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} F_{X_i}(x_i), f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} f_{X_i}(x_i).$$

Moments

Given the density $f_{\mathbf{X}}(\mathbf{x})$ of the random vector $\mathbf{X}$ the joint moments of order $s_1, s_2 \ldots, s_p$ are defined, in analogy to the univariate case, as

$$E(X_1^{s_1} \ldots X_p^{s_p}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1^{s_1} \ldots x_p^{s_p} f_{\mathbf{X}}(x_1,\ldots,x_p) dx_1 \ldots dx_p$$

Density transformation formula

Assume , the p existing random variables $X_1, X_2 \ldots, X_p$ with given density $f_{\mathbf{X}}(\mathbf{x})$ have been transformed by a smooth (i.e. differentiable) one-to-one transformation into p new random variables $Y_1, Y_2 \ldots, Y_p$, i.e. a new random vector $\mathbf{Y} \in \mathbf{R^p}$ has been created by calculating

$$Y_i = y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p$$

The question is how to calculate the density $g_{\mathbf{Y}}(\mathbf{y})$ of $\mathbf{Y}$ by knowing the transformation functions $y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p$ and the density $f_{\mathbf{X}}(\mathbf{x})$ of the original random vector. Since the transformation of the X's into Y's is assumed to be one-to-one, its inverse transformation $X_i = x_i(Y_1, Y_2 \ldots, Y_p), i = 1, 2, \ldots, p$ also exists and then the following density transformation formula applies:

$$g_{\mathbf{Y}}(y_1,\ldots,y_p) = f_{\mathbf{X}}[x_1(y_1,\ldots,y_p),\ldots,x_p(y_1,\ldots,y_p)]|J(y_1,\ldots,y_p)|$$

where $J(y_1, \ldots, y_p)$ is the *Jacobian* of the transformation:

$$J(y_1,\ldots,y_p) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \ldots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \ldots & \frac{\partial x_2}{\partial y_p} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \ldots & \frac{\partial x_p}{\partial y_p} \end{vmatrix}$$

Multivariate Normal Distribution

For the purpose of this course, we will only need the density for the case of non-degenerated multivariate normal. The formula generalizes the formula from the univariate case. Looking at the term $(\frac{x-\mu}{\sigma})^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ in the exponent of the well known

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$$

for the univariate density, a natural way to generalize this is to *replace* it by $(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)$. Here $\mu = E\mathbf{X} \in \mathbf{R^p}$ is the expected value of the random vector $\mathbf{X} \in \mathbf{R^p}$ and the matrix

$$\mathbf{\Sigma} = E(\mathbf{X}-\mu)(\mathbf{X}-\mu)' = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} \in \mathbf{M_{p,p}}$$

(assumed to be positive definite) is the *covariance matrix*. On the diagonals of $\mathbf{\Sigma}$ we get the *variances* of each of the p random variables whereas $\sigma_{ij}, i \neq j$ are the *covariances* between the ith and jth random variable. Sometimes, we also denote σ_{ii} by σ_i^2.

The final result is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)/\mathbf{2}}, -\infty < x_i < \infty, i = 1, 2 \dots, p$$

2 Lecture 2: THE GENERAL INFERENCE PROBLEM

2.1 Measurement precision

The purpose in Statistical Inference is to draw relevant conclusions from given data. The conclusions could be about predicting further outcomes, evaluating risks of events, testing hypotheses, etc. In all cases, inference about the **population** is to be drawn but only a limited information contained in the sample is available. The most common situation in Statistics is the one in which an experiment has been repeatedly performed, the replicates being independent of one another. The possible results are real numbers that form a vector of observations $\mathbf{x}=(x_1, x_2, \ldots, x_n)$. The appropriate sample space is R^n. There is typically a "hidden" mechanism that generates the data and one is looking for suitable ways to identify it. **Models** will describe this mechanism in some simplistic but hopefully useful way. For the model to be more trustworthy, continuous variables, such as time, interval measurements, etc. should be treated as such, where feasible. However, in practice, only discrete events can actually be observed. Thus, the observations will be recorded with some *unit of measurement,* Δ*,* determined by the precision of the measuring instrument. This unit of measurement is always finite in any real situation.

If empirical observations were truly continuous, then, with probability one, no two observed responses would ever be identical. This fact will sometimes be used in our theoretical derivations. On the other hand, as pointed out above, the *real life empirical observations are indeed discrete*. This fact will be utilized by us to keep some of the proofs simpler. In many cases we will be dealing with the discrete case only, thus avoiding more involved measure-theoretic arguments.

2.2 Statistical Models

Having got the vector of observations we can calculate the joint density (in the continuous case)

$$L_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1).f_{X_2}(x_2)\ldots f_{X_n}(x_n) \tag{1}$$

In the discrete case this will be just the product of the probabilities for each of the measurements to be in a suitable interval of length Δ.

If the observations were **independent identically distributed (i.i.d.)** then all densities in (1) would be the same: $f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) = f(x)$. This is the most typical situation we will be discussing in our course. The need of **Statistical Inference** arises since typically, our knowledge about $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$ is **incomplete.** Given an inference problem and having collected some data, we construct one or more set of possible **models** which may help us to understand the data generating mechanism. These models are usually about the shape of the density or of the cumulative distribution function. They should represent, as much as possible, the available prior theoretical knowledge about the

data generating mechanism. The suggestion of the set of models to be validated, is the first step, perhaps the most difficult: to think of a set of suitable model functions which might reasonably describe the data generating mechanism. The step usually involves a close *collaboration* between the statistician and the people who formulated the problem. We can view the statistical model as the triplet $(\mathcal{X}, \mathcal{P}, \Theta)$ where :

$-\mathcal{X}$ is the sample space (i.e.. the set of all possible realizations $\mathbf{X}$=$(X_1, X_2, \ldots, X_n)$

$-\mathcal{P}$ is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ;

$-\Theta$ is the set of possible θ-values, i.e.. the parameter space indexing the models.

2.3 Inference problem

The statistical inference problem can be formulated as follows:

Once the random vector $\mathbf{X}$ *has been observed, what can be said about which members of* $\mathcal{P}$ *best describe how it was generated?*

The reason we are speaking about a *problem* here is that we do not know the exact shape of the distribution that generated the data. The reason that there exists a *possibility* of making inference rests in the fact that typically a given observation is much more probable under some distributions than under others (i.e. the observations give **information** about the distribution). This information should be combined with the *a priori* information about the distribution to make the inference. Note that we *always* have some a priori information. It could be more or less specific. When it is specific to such an extent that the shape of the distribution is known up to some *finite* number of parameters, we have to conduct *parametric* inference. In this case Θ is at most a (subset of) finite-dimensional Euclidean space. If Θ could only be specified as a certain *infinite dimensional function space*, we speak about *non-parametric* inference. Needless to say that nonparametric inference procedures are applicable in more general situations (which is good). However if they are applied for a situation where a parametric distributional shape gives adequate enough description of the data, nonparametric procedures may not be as efficient as s specifically tailored parametric procedure (which would be bad if the specific parametric model indeed holds).

The situation in practice is often more blurred: we may know that the populations is "close" to parametrically describable and yet "deviates a bit" from the parametric family. Going over in such cases to purely nonparametric approach would not properly address the situation since the idea about *relatively small* deviation from the baseline parametric family will be lost. The proper approach in such cases would be the *robustness* approach where we still keep the idea about the "ideal" parametric model but allow for *small* deviations from it. The aim is in such "intermediate" situations to be "close to efficient" if the parametric model holds but at the same time to be "less sensitive" to small deviations from the ideal model. These important issues will be discussed in the course.

Another way to classify the Statistical Inference procedures is determined by the way we treat the *unknown* parameter θ. If we treat it as unknown but deterministic (fixed)

then we are in a **Non-Bayesian** setting. If we consider the set of θ-values as quantities that before collecting the data, have different probabilities of occurring according to some (*a priori)* distribution, then we are speaking about *Bayesian inference.* The Bayesian approach allows us to introduce and effectively utilise any additional (prior) information about the model when such information is available. This information is entered in the model through the **prior distribution** over the set Θ of parameter values and reflects our prior belief about how likely any of the parameter values is before obtaining the information from the data. This topic will also be discussed shortly in our course.

2.4 Goals in Statistical Inference

Following are the most common goals in inference:

2.4.1 Estimation

We want to calculate a number (or a k-dimensional vector, or a single function) as an approximation to the numerical characteristic in question.

But let us point out immediately that there is little value in calculating an approximation to an unknown quantity without having an idea of how "good" the approximation is and how it compares with other approximations. Hence, immediately questions about **confidence interval** (or, more generally, **confidence set**) construction arise. To quote the famous statistician A.N. Whitehead, in Statistics we always have to "seek simplicity and distrust it".

2.4.2 Confidence set construction

After the observations have been made, further information to our a priori knowledge about the set Θ has been added, so it becomes plausible that the true distribution belongs to a smaller family than it was originally postulated, i.e. it becomes clear that the unknown θ-value belongs to a *subset* of Θ. The problem of confidence set construction arises. It means, determining a (possibly small) plausible set of θ-values and clarifying the sense in which the set is plausible.

2.4.3 Hypotheses testing

An experimenter or a statistician sometimes has a theory which when suitably translated into mathematical language becomes a statement that the true unknown distribution belongs to a smaller family than the originally postulated one. One would like to formulate this theory in the form of a hypothesis. The data can be used then to infer whether or not his theory complies with the observations or is in such a serious disarray that would indicate that the hypothesis is false.

Deeper insight in all of the above goals of inference and deeper understanding of the nature of problems involved in them is given by **Statistical Decision Theory.** Here we define in general terms what a *statistical decision rule* is and it turns out that any of the procedures discussed above can be viewed as a suitably defined decision rule. Moreover, defining optimal decision rules as solutions to suitably formulated **constrained mathematical optimization problems** will help us to find "best" decision rules in many practically relevant situations.

2.5 Statistical Decision Theoretic Approach to Inference

2.5.1 Introduction

Statistical Decision Theory studies all inference problems (estimation, confidence set construction, hypothesis testing) from a unified point of view. All parts of the decision making process are formally defined, a desired optimality criterion is formulated and a decision is considered optimal if it optimizes the criterion.

Statistical Decision Theory may be considered as the theory of a two-person game with one player being the statistician and the other one being the nature. To specify the game, we define:

Θ-set of states (of nature);

$\mathcal{A}$- set of actions (available to the statistician);

$L(\theta, a)$ - real-valued function (loss) on $\Theta \times \mathcal{A}$.

There are some important differences between mathematical theory of games (that only involves the above triplet) and Statistical Decision Theory. The most important differences are:

- In a two-person game both players are trying to maximize their winnings (or to minimize their losses) , whereas in decision theory nature chooses a state without this view in mind. Nature can not be considered as an "intelligent opponent" who would behave "rationally". Also, there is no complete information available (to the statistician) about nature$'s$ choice.

- In Statistical Decision Theory nature always has the first move in choosing the "true state" θ.

- The statistician has the chance (and this is *most important*) to gather *partial information* on nature$'s$ choice by sampling or performing an experiment. This gives him the data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ that has a distribution $L(\mathbf{X}|\theta)$ *depending* on θ. This is used by the statistician to work out his decision.

Definition 1. A (deterministic) *decision function* is a function $d : \mathcal{X} \to \mathcal{A}$ from the sample space to the set of actions.

There is a non-negative loss $L(\theta, d(\mathbf{X}))$ incurred by this action. Of course this is a random variable. Hence one defines the *RISK* $E_\theta L(\theta, d(\mathbf{X})) = R(\theta, d)$. For a fixed decision , this is a function (risk function) depending on θ. $R(\theta, d)$ is interpreted as the average loss of the statistician when the nature has a true state θ and the statistician uses decision *d.*

2.5.2 Examples.

Assume that a data vector $\mathbf{X} \sim f(\mathbf{X}, \theta)$.

a) Hypothesis testing. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ where $\theta \in R^1$ is a parameter of interest. Let $\mathcal{A} = \{a_1, a_2\}, \Theta = R^1$. Here a_1 denotes the action "accept H_0 " whereas a_2 is the action "Reject H_0." Let

$$D = \{\text{Set of all functions from } \mathcal{X} \text{ into } \mathcal{A}\}.$$

Define

$$L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta > \theta_0, \\ 0 & \text{if } \theta \leq \theta_0 \end{cases}$$

$$L(\theta, a_2) = \begin{cases} 0 & \text{if } \theta > \theta_0, \\ 1 & \text{if } \theta \leq \theta_0 \end{cases}$$

Then

$$R(\theta, d) = EL(\theta, d(\mathbf{X})) = L(\theta, a_1)P_\theta(d(\mathbf{X}) = a_1) + L(\theta, a_2)P_\theta(d(\mathbf{X}) = a_2) =$$

$$\begin{cases} P_\theta(d(\mathbf{X}) = a_1) & \text{if } \theta > \theta_0, \\ P_\theta(d(\mathbf{X}) = a_2) & \text{if } \theta \leq \theta_0. \end{cases}$$

Hence

i) if $\theta \leq \theta_0 : R(\theta, d) = P_\theta(\text{ reject } H_0) =$ Error of I type ,

ii) if $\theta > \theta_0 : R(\theta, d) = P_\theta(\text{ accept } H_0) =$ Error of II type .

We will see later when studying optimality in hypothesis testing context that the set of deterministic decision rules D is not convex and it is difficult to develop a decent mathematical optimization theory over it. It has to be extended by including the so-called randomized decision rules if we want to formulate and solve such problems.

b) Estimation.

Let now $\mathcal{A} = \Theta$ with the interpretation that each action corresponds to selecting a point $\theta \in \Theta$. Every $d(\mathbf{X})$ maps $\mathcal{X}$ into Θ and if we chose $L(\theta, d(\mathbf{X})) = (\theta - d(\mathbf{X}))^2$ (*quadratic loss*) then the decision rule d (which we can call *estimator*) has a risk function

$$R(\theta, d) = E_\theta(d(\mathbf{X}) - \theta)^2 = MSE_\theta(d(\mathbf{X})).$$

2.5.3 Randomized decision rule

The set D of deterministic decision rules is not convex and it is difficult to develop a decent mathematical optimization theory over it. This set is also very small and examples show that very often a simple randomization of given deterministic rules gives better rules in the sense of risk minimization. This explains the reason for the introduction of the randomized decision rules.

Definition 2. A rule δ which chooses d_i with probability $w_i, \sum w_i = 1$, is a randomized decision rule.

For the randomized decision rule δ we have:

$$L(\theta, \delta(X)) = \sum w_i L(\theta, d_i(X)) \text{ and } R(\theta, \delta) = \sum w_i R(\theta, d_i)$$

The set of all randomized decision rules generated by the set D in the above way will be denoted by $\mathcal{D}$.

2.5.4 Optimal decision rules

Given a game $(\Theta, \mathcal{A}, L)$ and a random vector X whose distribution depends on $\theta \in \Theta$ what (randomized) decision rule δ should the statistician choose to perform "optimally"? This is a question that is easy to pose but usually difficult to answer. The reason is that usually *uniformly* best decision rules (that minimize the risk uniformly for all θ-values) do not exist! This observations is easy to understand and explanation will be given during the lecture. It leads us to the following two ways out:

- *First way out.*

  a) Constraining the set of decision rules and try to find uniformly best in this smaller set. This corresponds to looking for optimality under restrictions- we eliminate some of the decision rules since they do not satisfy the restrictions by hoping, in the smaller set of remaining rules to be able to find a uniformly best. Sensible constraints that we introduce in the estimation context are usually unbiasedness or invariance.

  Definition 3. A decision rule d is *unbiased* if

  $$E_\theta[L(\theta', d(\mathbf{X}))] \geq E_\theta[L(\theta, d(\mathbf{X}))] \text{ for all } \theta, \theta' \in \Theta$$

  holds.

  Exercise: Show that in the context of estimation of a parameter θ with quadratic loss function, the above definition is tantamount to the requirement

  $$E_\theta d(\mathbf{X}) = \theta \text{ for all } \theta \in \Theta,$$

  that is, the new definition is equivalent to the unbiasedness from classical statistical estimation theory.

  It is obvious that the new definition of unbiasedness is more general and can be applied to broader class of loss functions. The same definition also makes sense in

hypothesis testing where we can also introduce unbiased tests in the same way (see later the separate lecture about optimality in hypothesis testing) and then look for optimality amongst all unbiased α level tests.

- *Second way out.* Reformulating the optimality criterion in a new way. Since the "uniformly best" no matter what θ-value is too strong a requirement, we can introduce

  - Bayes risk

  - minimax risk

  of a decision rule and try to find the rules that minimize these risks. This leads to Bayesian and to minimax decision rules.

2.5.5 Bayesian and minimax decision rules

a) Bayesian rule. Let us think of the θ-parameter now as of being a random variable with a given (known) prior density τ on Θ. Define the

Bayesian risk of the decision rule δ with respect to the prior τ:

$r(\tau, \delta) = E[R(T, \delta)] = \int_\Theta R(\theta, \delta)\tau(\theta)d\theta$ (here T is a random variable over Θ having a distribution with a density τ.)

Then the *Bayesian rule δ_τ with respect to the prior τ* is defined as:

$$r(\tau, \delta_\tau) = \inf_{\delta \in \mathcal{D}} r(\tau, \delta)$$

Sometimes a Bayesian rule may not exist and then we could ask for an ϵ-Bayes rule. For $\epsilon > 0$, this is any rule $\delta_{\epsilon\tau}$ that satisfies $r(\tau, \delta_{\epsilon\tau}) \leq \inf_{\delta \in \mathcal{D}} r(\tau, \delta) + \epsilon$.

b) Minimax rule. Instead of considering uniformly best rules (that usually do not exist), one can alternatively consider rules that minimize the *supremum* of the values of the risk over the set Θ. This would mean safeguarding against the worst possible performance.

The value $\sup_{\theta \in \Theta} R(\theta, \delta)$ is called *minimax risk* of the decision rule δ. Then the rule δ^* is called *minimax* in the set $\mathcal{D}$ if

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) = \textit{minimax value of the game}$$

Note that again, like in the Bayesian case, even if the minimax value is finite there may not be a minimax decision rule. Hence again we introduce the notion of ϵ-minimax rule δ_ϵ (such that $\sup_{\theta \in \Theta} R(\theta, \delta_\epsilon) \leq \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) + \epsilon$)

Note that sometimes choosing a minimax rule may turn out to be too pessimistic a strategy but experience shows that in most cases minimax rules are good rules.

2.5.6 Least favorable prior distribution

After the above definitions have been given, one can easily define the least favorable distribution (i.e. least favorable prior τ^* over the set Θ) as:

$$\inf_{\delta\in\mathcal{D}} r(\tau^*,\delta) = \sup_{\tau}\inf_{\delta\in\mathcal{D}} r(\tau,\delta)$$

It indeed deserves its name. From the above definition we see that if the statistician were told which prior distribution nature was using, he would like least to be told that τ^* was the nature$'s$ prior (since given that he always performs in an optimal way by choosing the corresponding Bayesian rule, he still has the highest possible value of the Bayesian risk as compared to the other priors).

2.5.7 Geometric interpretation of the decision rules in the case of finite Θ

Definition 4. A set $A \subset R^k$ is *convex* if for all vectors $\vec{x} = (x_1, x_2, \ldots, x_k)'$ and $\vec{y} = (y_1, y_2, \ldots, y_k)'$ in A and all $\alpha \in [0,1] : \alpha\vec{x} + (1-\alpha)\vec{y} \in$ A.

Assume that Θ has k elements all together. Now let us define the *risk set* of a set D of decision rules. This is simply the set of all *risk points* $\{R(\theta,d), \theta \in \Theta\}, d \in$D. For a fixed d, each such risk point belongs to R^k and by "moving"d within D, we get a set of such k-dimensional vectors.

Theorem 2.1. *The risk set of a set $\mathcal{D}$ of randomized decision rules generated by a given set D of non-randomized decision rules is convex.*

Proof: It is easy to see that if $\vec{y}$ and $\vec{y}'$ are the risk points of the decision rules δ and $\delta' \in D$, correspondingly, then any point in the form $\vec{z} = \alpha\vec{y} + (1-\alpha)\vec{y}'$ corresponds to (is the risk point of) the randomized decision rule $\delta_\alpha \in \mathcal{D}$ that chooses the rule δ with probability α and the rule δ' with probability $(1-\alpha)$. Hence any such $\vec{z}$ belongs to the risk set of $\mathcal{D}$.

Remark: In fact, the risk set of the set of all randomized rules $\mathcal{D}$ generated by the set D is *the smallest convex set containing the risk points of all of the non- randomized rules in D* (i.e. *the convex hull* of the set of risk points of D).

How to illustrate Bayes rules: Since $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ then the prior $\tau = (p_1, p_2, .., p_k)$ in the case we are dealing with $(p_i \geq 0, \sum_{i=1}^k p_i = 1)$. The Bayes risk of any rule δ w.r. to the prior τ is $r(\tau,\delta) = \sum_{i=1}^k p_i R(\theta_i,\delta)$. All points $\vec{y}$ in the risk set, corresponding to certain rules δ^* for which $\sum_{i=1}^k p_i y_i = r(\tau,\delta^*) =$ the same value $= b$, give rise to the same value b of the Bayesian risk and hence are equivalent from a Bayesian point of view. The value of their risk can be easily illustrated and (at least in case of $k = 2$), one can easily illustrate the point in the convex risk set that corresponds to (is the risk point of the) Bayesian rule with respect to the prior τ. (See illustration handed out).

In a similar way, the minimax rule can be illustrated in the case of finite Θ at least (See illustration).

2.5.8 Example

Let the set $\Theta = \{\theta_1, \theta_2\}$. Let X have possible values 0, 1 and 2; the set $\mathcal{A} = \{a_1, a_2\}$ and let $L(\theta_1, a_1) = L(\theta_2, a_2) = 0, L(\theta_1, a_2) = 1, L(\theta_2, a_1) = 3$. The distributions of X are tabulated as follows:

$$\left| \begin{array}{cccc} x & 0 & 1 & 2 \\ P(x|\theta_1) & .81 & .18 & .01 \end{array} \right| \left| \begin{array}{cccc} x & 0 & 1 & 2 \\ P(x|\theta_2) & .25 & .5 & .25 \end{array} \right|$$

The decision problem formulated above can be interpreted as an attempt by the statistician to guess the right state of nature. If his guess is correct, he does not lose anything but if he is wrong, he either loses \$1 or \$3 depending on the type of error he has made. In his guess he is supported by one observation X that has a different distribution under θ_1 and under θ_2.

Now, consider all possible non-randomized decision rules based on one observation:

$$\left[\begin{array}{cccccccc} x & d_1(x) & d_2(x) & d_3(x) & d_4(x) & d_5(x) & d_6(x) & d_7(x) & d_8(x) \\ 0 & a_1 & a_1 & a_1 & a_1 & a_2 & a_2 & a_2 & a_2 \\ 1 & a_1 & a_1 & a_2 & a_2 & a_1 & a_1 & a_2 & a_2 \\ 2 & a_1 & a_2 & a_1 & a_2 & a_1 & a_2 & a_1 & a_2 \end{array} \right]$$

The following questions have to be answered:

a) sketch the risk set of all randomized rules generated by $d_1, d_2, .., d_8$;

b) find the minimax rule δ^*(in $\mathcal{D}$) and compute its risk;

c) for what prior is δ^* a Bayes rule w.r. to that prior (i.e., what is the least favorable distribution) ;

d) What is the Bayes rule for the prior $\{1/3, 2/3\}$ over $\{\theta_1, \theta_2\}$? What is the value of its Bayes risk?

Solution : to be discussed at lecture.

2.5.9 Fundamental Lemma

Lemma 2.2. *If $\tau^*$is a prior on Θ and the Bayes rule δ_{τ^*} has a constant risk w.r. to θ(i.e. if $R(\theta, \delta_{\tau^*}) = c_0$ for all $\theta \in \Theta$) then:*

a) δ_{τ^*} is minimax;

b) τ^* is the least favorable distribution.

Proof: a) Let us compute the minimax risk of δ_{τ^*} and compare it to the minimax risk of any other rule δ:

$c_0 = \sup_{\theta\in\Theta} R(\theta, \delta_{\tau^*}) =$ (since constant for all θ) $= \int_\Theta R(\theta, \delta_{\tau^*})\tau^*(\theta)d\theta \leq$ (since δ_{τ^*} Bayes w.r. to τ^*) $\leq \int_\Theta R(\theta, \delta)\tau^*(\theta)d\theta \leq \sup_{\theta\in\Theta} R(\theta, \delta)$,

which means that δ_{τ^*} is minimax.

b) Now take any other prior τ. Then we have:

$\inf_\delta r(\tau,\delta) = \inf_\delta \int_\Theta R(\theta,\delta)\tau(\theta)d\theta \leq \int_\Theta R(\theta,\delta_{\tau^*})\tau(\theta)d\theta =$ (since $R(\theta,\delta_{\tau^*})$ constant and

$$\int_\Theta \tau^*(\theta)d\theta = \int_\Theta \tau(\theta)d\theta = 1) = \int_\Theta R(\theta,\delta_{\tau^*})\tau^*(\theta)d\theta = r(\tau^*,\delta_{\tau^*}),$$

hence τ^* is least favorable.

Remark. The above lemma gives a hint how to find minimax estimators. The minimax estimators turn out to be (special) Bayes estimators with respect to the least favorable prior. First we can obtain the general form of the Bayes estimator with respect to ANY given prior. Then in order to get the minimax estimator we should choose such a prior for which the corresponding Bayes rule has its (usual) risk independent of θ, i.e. constant with respect to θ.

2.5.10 Finding Bayes rules analytically

This is important in its own right but also, as seen in 2.5.9, as a device to be utilized in the search for minimax rules. It turns out that given the prior and the observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ we can find the Bayes rule point-wise (i.e. for any given $\mathbf{X}$=$\mathbf{x}$) by solving a certain minimization problem. In many practically relevant cases the solution can even be given in a closed form.

To see how, let us first introduce the following notation:

$-f(\mathbf{X}|\theta)$ is the conditional density of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ given θ;

$-\tau(\theta)$ is the prior density on θ;

$-g(\mathbf{X})$ is the marginal density of $\mathbf{X}$, i.e. $g(\mathbf{X}) = \int_\Theta f(\mathbf{X}|\theta)\tau(\theta)d\theta$;

$-h(\theta|\mathbf{X})$ is the posterior density of θ given $\mathbf{X} = (X_1, X_2, \ldots, X_n)$;

$-f(\mathbf{X},\theta)$ is the joint density of X and θ. Note: $f(\mathbf{X},\theta) = f(X|\theta)\tau(\theta) = h(\theta|\mathbf{X})g(\mathbf{X})$

$$h(\theta|\mathbf{X}) = \frac{f(\mathbf{X},\theta)}{g(\mathbf{X})} = \frac{f(\mathbf{X}|\theta)\tau(\theta)}{\int_\Theta f(\mathbf{X}|\theta)\tau(\theta)d\theta}$$

Now we formulate a **General Theorem** regarding calculation of Bayesian decision rules.

Theorem 2.3. *Define for $X \in \mathcal{X}, a \in \mathcal{A}$ and for a given prior τ, define:*

$$Q(\boldsymbol{X}, a) = \int_\Theta L(\theta,a)h(\theta|\boldsymbol{X})d\theta,$$

(remember that $L(.,.)$ was the loss function)).

Suppose that for each $\boldsymbol{X} \in \mathcal{X}$, there exists a rule $a_{\boldsymbol{X}} \in \mathcal{A}$ such that

$$Q(X, a_{\boldsymbol{X}}) = \inf_{a \in \mathcal{A}} Q(\boldsymbol{X}, a)$$

If $\delta_\tau(\boldsymbol{X}) = a_{\boldsymbol{X}}$ belongs to $\mathcal{D}$ then $\delta_\tau(\boldsymbol{X}) = a_{\boldsymbol{X}}$ is the (point wise defined) Bayes decision rule with respect to the prior τ.

Proof: For any decision rule δ we have:

$$r(\tau, \delta) = \int_\Theta R(\theta, \delta)\tau(\theta)d\theta = \int_\Theta [\int_{\mathcal{X}} L(\theta, \delta(\mathbf{X}))f(\mathbf{X}|\theta)d\mathbf{X}]\tau(\theta)d\theta =$$

$$\int_{\mathcal{X}} [\int_\Theta L(\theta, \delta(\mathbf{X}))h(\theta|\mathbf{X})d\theta]g(\mathbf{X})d\mathbf{X} =$$

$$\int_{\mathcal{X}} Q(\mathbf{X}, \delta(\mathbf{X}))g(\mathbf{X})d\mathbf{X}$$

(here we use the short-hand notation $d\mathbf{X} := dX_1 dX_2 \dots dX_n$).

But for every fixed $\mathbf{X}$-value, $Q(\mathbf{X}, \delta(\mathbf{X}))$ is smallest when $\delta(\mathbf{X}) = a_{\mathbf{X}}$. Making that way our "best choice" for each $\mathbf{X}$-value, we will, of course, minimize the value of $r(\tau, \delta)$. Hence, we should be looking for an action $a_{\mathbf{X}}$ that gives an infimum to

$$\inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a)h(\theta|\mathbf{X})d\theta$$

Now let us apply the above general theorem to the cases of estimation and hypothesis testing.

Theorem 2.4. *(case of estimation). Consider a point estimation problem for a real-valued parameter θ. The prior over θ is denoted by τ. Then:*
a) for a squared error loss $L(\theta, a) = (\theta - a)^2 : \delta_\tau(\boldsymbol{X}) = E(\theta|\boldsymbol{X}) = \int_\Theta \theta h(\theta|\boldsymbol{X})d\theta$;
b) for an absolute error loss $L(\theta, a) = |\theta - a| : \delta_\tau(\boldsymbol{X}) =$ median of $h(\theta|\boldsymbol{X})$.

Remark An example of a case in which the condition $\delta_\tau \in \mathcal{D}$ could not be satisfied is in point- estimation problems with $\Theta \equiv \mathcal{A}$- a finite set. Then $E(\theta|\mathbf{X})$ might not belong to $\mathcal{A}$, hence $E(\theta|\mathbf{X})$ would not be a function $\mathcal{X} \to \mathcal{A}$ and δ_τ would not be a legitimate estimator. But if $\Theta \equiv \mathcal{A}$ is *convex*, it can be shown that *always* $E(\theta|\mathbf{X}) \in \mathcal{A}$!

Theorem 2.5. *(case of Bayesian hypothesis testing with a generalized $0-1$ loss). Let $\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \oslash$ and we are testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$. A prior τ is given on Θ. Two non-randomized actions a_0(accept H_0) and a_1(reject H_0) are possible and the losses are given by: $L(\theta, a_0) = \begin{cases} 0 \text{ if } \theta \in \Theta_0 \\ c_2 \text{ if } \theta \in \Theta_1 \end{cases}$, $L(\theta, a_1) = \begin{cases} c_1 \text{ if } \theta \in \Theta_0 \\ 0 \text{ if } \theta \in \Theta_1 \end{cases}$. Then the test $\varphi^* = \begin{cases} \text{Reject } H_0 \text{ if } P(\theta \in \Theta_0 | \boldsymbol{X}) < c_2/(c_1 + c_2) \\ \text{Accept } H_0 \text{ if } P(\theta \in \Theta_0 | \boldsymbol{X}) > c_2/(c_1 + c_2) \end{cases}$ is a Bayesian rule (Bayesian test) for the above testing problem w.r. to the prior τ.*

Proof: According to the General Theorem, we have to compare $Q(\mathbf{X}, a_0)$ and $Q(\mathbf{X}, a_1)$ and take as our action the one that gives the smaller value. Now:

$$Q(\mathbf{X}, a_0) = \int_{\Theta} L(\theta, a_0) h(\theta|\mathbf{X}) d\theta = \int_{\Theta_1} c_2 h(\theta|\mathbf{X}) d\theta = c_2 P(\theta \in \Theta_1 |\ \mathbf{X}) = c_2(1 - P(\theta \in \Theta_0 |\ \mathbf{X}))$$

$$Q(\mathbf{X}, a_1) = \int_{\Theta} L(\theta, a_1) h(\theta|\mathbf{X}) d\theta = \int_{\Theta_0} c_1 h(\theta |\ \mathbf{X}) d\theta = c_1 P(\theta \in \Theta_0 |\ \mathbf{X})$$

Hence we would reject H_0 when $Q(\mathbf{X}, a_1) < Q(\mathbf{X}, a_0)$, i.e. for

$$\{\mathbf{X} : c_1 P(\theta \in \Theta_0|\mathbf{X}) < c_2(1 - P(\theta \in \Theta_0|\mathbf{X}))\} = \{\mathbf{X} : P(\theta \in \Theta_0|\mathbf{X}) < c_2/(c_1 + c_2)\}$$

2.5.11 Example about calculating minimax estimator for the probability of success in the Bernoulli experiment.

Let given θ, the distribution of each $X_i, i = 1, 2, \ldots, n$ be Bernoulli with parameter θ, i.e. $f(\mathbf{X}|\theta) = \theta^{\Sigma X_i}(1 - \theta)^{n - \Sigma X_i}$ and assume a beta-prior τ for the (random variable) θ over $(0, 1) : \tau(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}.I_{(0,1)}(\theta)$.

Hereby $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function (with the obvious property $B(\alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-1}.B(\alpha - 1, \beta)$).

For the above prior we can find easily

$$h(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta)\tau(\theta)}{\int_0^1 f(\mathbf{X}|\theta)\tau(\theta)d\theta} = \frac{\theta^{\Sigma X_i+\alpha-1}(1 - \theta)^{n-\Sigma X_i+\beta-1}}{B(\Sigma X_i + \alpha, n - \Sigma X_i + \beta)}$$

(which is again a beta-density). Hence the Bayesian estimator is

$$\hat{\theta}_\tau = \int_0^1 \theta h(\theta|\mathbf{X}) d\theta = \frac{B(\Sigma X_i + \alpha + 1, n - \Sigma X_i + \beta)}{B(\Sigma X_i + \alpha, n - \Sigma X_i + \beta)} = \text{by the property of the beta-function} =$$

$$= \frac{\Sigma X_i + \alpha}{\alpha + \beta + n}.$$

The above derivation holds for any beta prior Beta(α, β).

Very improtant note: We do NOT in fact need to calculate the normalization $g(\mathbf{X})$ explicitly. Once we realized that $f(\mathbf{X}|\theta)\tau(\theta) \propto \theta^{\Sigma X_i+\alpha-1}(1 - \theta)^{n-\Sigma X_i+\beta-1}$ (hence up to a norming constant we are dealing with a Beta density with parameters $\Sigma X_i + \alpha$ and $n - \Sigma X_i + \beta$) then the conditional $h(\theta|\mathbf{X})$, being a density, MUST be the Beta density with these parameters. Knowing how the expected value of the beta distribution depends on the paramteres of the distribution we can immediately get the Bayesian estimator as $\frac{\Sigma X_i+\alpha}{\alpha+\beta+n}$. Such type of arguments are routinely used in Bayesian inference and save a lot of unnecessary computations of norming constants! See the tutorial problems for further illustrations of this approach.

Let us calculate the (usual) risk with respect to quadratic loss of any such Bayes estimator:

$$R(\theta,\hat{\theta}_\tau)=E(\hat{\theta}_\tau-\theta)^2=\mathrm{Var}_\theta(\hat{\theta}_\tau)+(\theta-E_\theta\hat{\theta}_\tau)^2=\frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2}+(\frac{n\theta+\alpha}{\alpha+\beta+n}-\theta)^2=\ldots=$$

$$\frac{n\theta-n\theta^2+(\alpha+\beta)^2\theta^2+\alpha^2-2\alpha(\alpha+\beta)\theta}{(n+\alpha+\beta)^2}$$

For this risk not to depend on θ, it has to hold: $\begin{cases}(\alpha+\beta)^2=n\\ 2\alpha(\alpha+\beta)=n\end{cases}$

The solution to this system is $\alpha=\beta=\sqrt{n}/2$. Hence (see 2.5.9) the minimax estimator of θ is

$$\hat{\theta}_{\text{minimax}}=\frac{\Sigma X_i+\sqrt{n}/2}{n+\sqrt{n}}.$$

2.5.12 What to do if a Bayes rule can not be found analytically.

As seen, integration techniques play a significant role in analytic determination of the Bayesian estimators and tests. Integration may be difficult to get in closed form in most of the cases and some numerical methods need to be applied in such situations. Simple Monte Carlo methods to calculate the integrals $\int_\Theta \theta f(X|\theta)\tau(\theta)d\theta$ and $\int_\Theta f(X|\theta)\tau(\theta)d\theta$ can always be applied. However, besides the simple Monte Carlo methods, there are more complicated Monte Carlo procedures which are specific and very useful in Bayesian inference. To motivate these procedures we first consider a simplified general example given in the following Lemma.

Lemma Suppose we generate random variables by the following algorithm:

i) Generate $Y\sim f_Y(y)$;

ii) Generate $X\sim f_{X|Y}(x|Y)$.

Then $X\sim f_X(x)$.

Proof: For the cumulative distribution function $F_X(x)$ we have:

$$F_X(x)=P(X\le x)=E[F_{X|Y}(x|y)]=\int_{-\infty}^{\infty}[\int_{-\infty}^{x}f_{X|Y}(t|y)dt]f_Y(y)dy=$$

$$\int_{-\infty}^{x}[\int_{-\infty}^{\infty}f_{X|Y}(t|y)f_Y(y)dy]dt=\int_{-\infty}^{x}[\int_{-\infty}^{\infty}f_{X,Y}(t,y)dy]dt=\int_{-\infty}^{x}f_X(t)dt.$$

Hence, the random variable X generated by the algorithm has a density $f_X(x)$.

The above Lemma tells us that if we wanted to calculate an expected value $E[W(X)]$ for any function $W(X)$ with $E[W^2(X)]<\infty$ then we can generate independently the

sequence $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_m, X_m)$ for a specified large value m and then by the *Law of Large Numbers* we will have

$$\bar{W} \approx E[W(X)].$$

The above simple observation can be generalized in the following algorithm of the **Gibbs sampler**:

Let m be a positive integer and X_0 an initial value. Then for $i = 1, 2, \ldots, m$:

i) Generate $Y_i|X_{i-1} \sim f_{Y|X}(y|x)$
ii) Generate $X_i|Y_i \sim f_{X|Y}(x|y)$.

In more advanced texts, it can be shown that $Y_i \to^d f_Y(y)$ and $X_i \to^d f_X(x)$ as $i \to \infty$. Therefore, intuitively, a convergence of the Gibbs sampler could be argued about in a manner similar to the Lemma.

The rigorous justification of the latter convergence is in fact a bit more involved. Indeed the Gibbs sampler algorithm is similar but not quite the same as the one in the Lemma. Let us examine the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k), (X_{k+1}, Y_{k+1})$ generated by the Gibbs sampler. On one hand, they are not generated independently, on the other hand, however, we need only the pair (X_k, Y_k) (and none of the previous $(k-1)$ pairs) to generate (X_{k+1}, Y_{k+1}). With other words, given the present, the future of the sequence is independent of the past. Such a sequence is called a **Markov chain** and for it, under quite general conditions, the distribution stabilizes (reaches an equilibrium).

The application of the Gibbs sampler in Bayesian inference can help in overcoming one of the major obstacles of this inference, namely, the fact that the prior may not be precisely known. We can in fact allow more freedom to ourselves by modeling the prior itself using another random variable. We get the so-called **hierarchical Bayes** model if we assume:

$$X|\theta \sim f(x|\theta), \Theta|\gamma \sim q(\theta|\gamma), \Gamma \sim \psi(\gamma)$$

with $q(.|.)$ and $\psi(.)$ known density functions. Here γ is called the *hyperparameter*. We keep in mind that $f(X|\theta)$ does **not** depend on γ. Keeping $g(.)$ as a generic notation for a density, we get using the Bayes formula: $g(\theta, \gamma|x) = \frac{f(x|\theta)q(\theta|\gamma)\psi(\gamma)}{g(x)}$. This conditional joint density is proportional to the product of known densities $f(x|\theta)q(\theta|\gamma)\psi(\gamma)$ hence $g(\theta|x, \gamma)$ and $g(\gamma|x, \theta)$ can (in principle) be determined. (When there is no easy analytic way of doing this then there is the *Metropolis-Hastings* algorithm to help us simulate from the conditionals. The Metropolis-Hastings algorithm is discussed in a Bayesian statistics course and we will avoid discussing it here).

We can then start a Gibbs sampler with an initial value γ_0 as follows:

i) $\Theta_i|X, \gamma_{i-1} \sim g(\theta|X, \gamma_{i-1})$
ii) $\Gamma_i|X, \theta_i \sim g(\gamma|X, \theta_i)$.

(With other words, we simulate from the *full conditionals*-the conditional distributions of each parameter given the other parameters and the data.)

Taking sufficiently large repetitions the algorithm will converge under suitable conditions as follows: $\Theta_i \to^d h(\theta|X), \Gamma_i \to^d g(\gamma|X)$ as $i \to \infty$. Hence the simple arithmetic

average of the Θ_i values (after possibly discarding some initial iterates before stabilization has occured) will converge towards the Bayes estimator with respect to quadratic loss for the given hierarchical Bayes model. In practice, we would generate the stream of values $(\theta_1, \gamma_1), (\theta_2, \gamma_2), \ldots$. Then choosing large values of m and $B > m$, our Bayes estimate of θ will be the average

$$\frac{1}{B-m} \sum_{i=m+1}^{B} \theta_i.$$

Additional note: The Gibbs sampler discussed above works fine when indeed the conditional distributions are completely known. However, quite often, these conditional distributions are only known up to a (normalizing) proportionality constant. Interestingly, the Gibbs sampler can still be used also in these cases although the drawing from the conditional distribution is a bit more involved. The best algorithm that covers this case is the **Metropolis-Hastings** algorithm. Without specifying its details (note that we have a separate course MATH5960 in Bayesian inference that deals with these details), we outline here only the essence of the algorithm. Suppose that a probability density $f(x)$ is only known up to a normalizing constant, i.e. $f(x) = c\tilde{f}(x)$ where $\tilde{f}$ is known. Choose an arbitrary, completely known, so-called proposal density $u(x'|x)$. Let the tth generated data point is $x = X^t$. For this given x define the set of points $A_x = \{x' : \tilde{f}(x')u(x|x') < \tilde{f}(x)u(x'|x)\}$.

i) Generate a value x' randomly from $u(\mathbf{X}'|x)$.
ii) If x' is not in A_x then we put $X^{t+1} = x'$ as the new simulated point. However, if x' is in A_x then we perform a further randomization and accept x' with probability $\tilde{f}(x')u(x|x')/\tilde{f}(x)u(x'|x)$. If it is accepted, we again put $X^{t+1} = x'$. Otherwise, we put $X^{t+1} = x$.

The theory of this algorithm requires some very mild conditions on $u(x'|x)$ to work but practice shows that in terms of computing time needed to run the algorithm and generating a "well mixing" sequence of simulated values, the choice of $u(x'|x)$ needs to be done carefully.

3 Lecture 3: PRINCIPLES OF DATA REDUCTION AND INFERENCE

3.1 Data Reduction in Statistical Inference

Given vector $\mathbf{X}=(X_1, X_2, .., X_n)$ of n i.i.d. random variables, each with a density $f(x;\theta)$, we are meant to conduct inference on $\theta \in \Theta$ based on the observations $x_1, x_2, .., x_n$. Let $\mathbf{X}$ takes values in $\mathcal{X}$ - the sample space. The statistician uses the information in the observations $x_1, x_2, .., x_n$ to conduct the inference. His/her wish is to summarize the information in the sample by determining a few key features of the sample values through transforming the sample values. Calculating such transformations (i.e. functions of the sample) means to calculate a **statistic.** Typically , $\dim(\mathbf{T}) << n$, i.e. using the statistic, we achieve the goal of data reduction: rather than reporting the entire sample $\mathbf{x}$, the statistic reports only that $\mathbf{T}(\mathbf{x})=\mathbf{t}$. Data reduction in terms of a particular statistic can be thought of as a *partition of the sample space*. We partition $\mathcal{X}$ into disjoint subsets $A_t = \{\mathbf{X}:\mathbf{T}(\mathbf{X})=\mathbf{t}\}$. If $\tau = \{\mathbf{t}:\mathbf{t}=\mathbf{T}(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ then the sample space $\mathcal{X}$ is represented as a union of the following disjoint sets (i.e. is partitioned) : $\mathcal{X} = \bigcup_{t\in\tau} A_t$. The ultimate goal in the data reduction is, when only using the value of the statistic $T(\mathbf{x})$ instead of the whole vector $\mathbf{x}$,"not to lose information" about the parameter of interest θ. The whole information about θ will be contained in the statistic and, in particular, we will treat as equal *any* two samples $\mathbf{x}$ and $\mathbf{y}$ that satisfy $\mathbf{T}(\mathbf{x})=\mathbf{T}(\mathbf{y})$ even though the actual sample values may be different. That way we arrive at the definition of sufficiency. The information in $\mathbf{X}$ about θ can be discussed in terms of partitions of the sample space.

Definition 1(**sufficient partition**)
Suppose for *any* set A_t in a particular partition $\mathcal{A} = \{A_t, t \in \tau\}$ we have

$$P\{\mathbf{X}=\mathbf{x} \mid \mathbf{X} \in A_t\}$$

does not depend on θ. Then $\mathcal{A}$ is a sufficient partition for θ.

Note: We have seen above that the partition is defined through a suitable statistic. If the statistic $\mathbf{T}$ is such that it generates a sufficient partition of the sample space then the statistic itself is sufficient.

3.2 Example:

$\mathbf{X}=(X_1, X_2, .., X_n)$ i.i.d. Bernoulli with parameter θ, i.e.
$P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, x_i = 0, 1$. The partition $\mathcal{A} = (A_0, A_1, \ldots, A_n)$ where $x \in A_r$ if and only if (iff) $\sum_{i=1}^n x_i = r$, is sufficient for θ. Correspondingly, the statistic $T(X) = \sum_{i=1}^n X_i$ is sufficient for θ.

Proof: At lecture.

Note that given the observed value $\mathbf{t}$ of $\mathbf{T}$, we know that the observed value $\mathbf{x}$ of $\mathbf{X}$ is in the partition set A_t. Sufficiency means that $P(X = x \mid T = t)$ is a function of x and t **only**

(i.e., is **not** a function of θ). Thus once having observed the particular realization t of T, knowing in addition the particular value $\mathbf{x}$ of $\mathbf{X}$ would not help for a better identification of θ. Hence we arrive at the **sufficiency principle:**

3.3 Sufficiency principle

The sufficiency principle implies that if T is sufficient for θ, then if x and y are such that $T(x) = T(y)$ then inference about θ should be the same whether $X = x$ or $Y = y$ is observed.

The following is a very useful criterion that helps us to check whether a statistic is sufficient or not by just looking at the joint density:

3.4 Neyman Fisher Factorization Criterion

If $X_i \sim f(x, \theta)$ then $T(X) = T(X_1, X_2, \ldots, X_n)$ is sufficient for θ iff
$L(X, \theta) = f_\theta(X_1, X_2, \ldots, X_n) = g(T(X), \theta)h(X)$(Note: X, T, θ may all be vectors, $g \geq 0, h \geq 0$).

Proof: at the lecture.

Sufficient partitions can be ordered and the **coarsest partition** (i.e. the one that contains the smallest number of sets) is called the **minimal sufficient partition.** Suppose T is sufficient and $T(X) = g_1(U(X))$ where U is a statistic and g_1 is a known function. It can be seen that U must also be sufficient for θ then. Indeed, applying the factorization criterion, we have:

$$L(X, \theta) = g(T(X), \theta)h(X) = g(g_1(U(X)), \theta)h(X) = \bar{g}(U(X), \theta)h(X)$$

which means that $U(X)$ is also sufficient. But, generally speaking, U induces a finer (or at least no coarser) partition than T since it might happen that $U_1 \neq U_2$ but yet $g_1(U_1) = g_1(U_2)$. Thus a finer partition of any sufficient partition is sufficient.

In applications one will be looking at the **coarsest** partition that is still sufficient because this means the greatest data reduction without loss of information on θ. From the above, we see that the statistic that introduces this coarsest partition will be a function of any other sufficient statistics. Such a statistic is called the **minimal sufficient** statistic.

Here we summarize some properties of sufficient statistics:

i) If T is sufficient, so is any one-to-one function of T(since it generates the same partition);

ii) If T is minimal sufficient, it is necessarily a function of all other possible sufficient statistics;

iii) If T is sufficient then $P(\mathbf{x} \mid \mathbf{t})$ does not depend on θ. The observed $\mathbf{t}$ is a summary of $\mathbf{x}$ that contains all the information about θ in the data, under the given family of models. It divides the sample space $\mathcal{X}$ into disjoint subsets A_t, each containing all possible observations $\mathbf{x}$ with the same value $\mathbf{t}$.

3.5 Examples

At lecture.

i) Bernoulli with probability of success $\theta \in (0,1)$. Sufficient statistic: $T = \sum_{i=1}^{n} X_i$.

ii) Univariate normal distribution with unknown μ and σ^2; Sufficient statistic for $\theta = (\mu, \sigma^2)'$: with two components $T_1 = \bar{X}, T_2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

iii)i.i.d. uniform in $(0,\theta)$: $f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta), x > 0, \theta > 0$. Sufficient statistic is $X_{(n)}-$ the maximal of the n observations.

iv) multivariate normal with mean vector μ and a covariance matrix Σ. Sufficient statistic: the vector $\bar{X}$ and the matrix $\sum_{i=1}^{n}(X_1 - \bar{X})(X_i - \bar{X})'$.

3.6 Lehmann and Scheffe's method for constructing a minimal sufficient partition

Often the sufficient statistic which has been found by the Factorization criterion, turns out to be minimal sufficient. Yet to find a general method for constructing a minimal sufficient statistic is a difficult task.

Consider a partition $\mathcal{A}$ of $\mathcal{X}$ by defining for any $x \in \mathcal{X}$: $A(x) = \{y : \frac{L(y,\theta)}{L(x,\theta)}$ does not depend on θ, i.e. is a function of the type $h(y,x)\}$. (Note that here we do not define explicitly the statistic that produces this partition, but in particular cases we shall always try to find a statistic whose contours produce the partition).

Theorem 3.1. *(Lehmann- Scheffe's method) The above defined sets $\{A(x), x \in \mathcal{X}\}$ indeed form a partition of $\mathcal{X}$ and this partition is minimal sufficient.*

Proof: (This proof is included for completeness only and for students who are curious to see how it really works. However the details will NOT be discussed at the lecture and are NOT required! Hence you can as well safely skip this proof.)

Proof (for simplicity- the discrete case only).

Step one. We have to show that the sets $\{A(x), x \in \mathcal{X}\}$ form a partition. To this end, we have to show that they are *either disjoint or coincide*. If we assume there exists a joint element $z \in A(x) \bigcap A(u)$ then $A(x)$ and $A(u)$ must coincide. Indeed, this is true:

i) Take any other $x_0 \in A(x)$ and any $u_0 \in A(u)$.

ii) Then $L(z,\theta)/L(x_0,\theta) = \frac{L(z,\theta)}{L(x,\theta)} / \frac{L(x_0,\theta)}{L(x,\theta)}$ is *not* a function of θ because each of the two ratios on the RHS are not.

iii)$L(z,\theta)/L(u,\theta)$ is also *not* a function of θ since $z \in A(u)$.

iv) But then $L(x_0,\theta)/L(u,\theta) = \frac{L(x_0,\theta)}{L(x,\theta)} \cdot \frac{L(x,\theta)}{L(z,\theta)} \cdot \frac{L(z,\theta)}{L(u,\theta)}$ is *not* a function of θ since the RHS is *not*.

The conclusion from this chain of statements is that an *arbitrary* $x_0 \in A(x)$ belongs to $A(u)$, too.

But in the same way as above it can be argued that $u_o \in A(x)$ and u_0 was *arbitrary* in $A(u)$. Therefore it must hold $A(x) = A(u)$ if they had one joint element z. This shows that $\{A(x), x \in \mathcal{X}\}$ **is** a partition.

Step two. We want to show the above defined partition $\mathcal{A}$ is *minimal sufficient*. Remember that we consider the discrete case only. First, we show that the partition is *sufficient*. Fix x and consider the conditional probability $P(Y = y \mid Y \in A(x))$:

$$P(Y = y \mid Y \in A(x)) = \frac{P_\theta(Y=y, Y \in A(x))}{P_\theta(Y \in A(x))} = \begin{cases} 0 \text{ if } y \text{ not in } A(x), \\ \frac{P_\theta(Y=y)}{P_\theta(Y \in A(x))} \text{ if } y \in A(x) \end{cases}$$

But since $\frac{P_\theta(Y=y)}{P_\theta(Y \in A(x))} = \frac{P_\theta(Y=y)}{\sum_{z \in A(x)} P_\theta(Y=z)} = \frac{P_\theta(y)}{\sum_{z \in A(x)} h(z,x) P_\theta(x)}$ is *not* a function of θ, we see that always $P(Y = y \mid Y \in A(x))$ does not depend on θ, i.e. $\mathcal{A}$ is a sufficient partition.

Now we want to show that $\mathcal{A}$ is a *minimal sufficient partition*. Take any $A(x)$. Fix any $y \in A(x)$. Assume that $v = v(Y)$ is also sufficient and creates a coarser partition . If y and z are such that $v(y) = v(z)$ then by the factorization theorem (v is assumed to be sufficient (!)) we have:

$$L(y, \theta) = g(v(y), \theta) h^*(y) = g(v(z), \theta) h^*(y) = \frac{L(z, \theta)}{h^*(z)} . h^*(y)$$

i.e. $L(z, \theta)/L(y, \theta)$ is not a function of θ.

But this means $\frac{L(z,\theta)}{L(x,\theta)} = \frac{L(z,\theta)/L(y,\theta)}{L(x,\theta)/L(y,\theta)}$ is *not* a function of θ, because the RHS is *not*. Hence y and z are in the same $A(x)$ class. This means that the partition $A(x)$ includes the partition generated by v and so, $\mathcal{A}$ must be the *coarsest* partition.

3.7 Examples

(at lecture).

i) Let $X_i, i = 1, 2, \ldots, n$ be i.i.d. Bernoulli with parameter $\theta \in [0, 1]$ as a probability of success. Consider the $n-$tupels $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$. Then

$$\frac{L(Y, \theta)}{L(X, \theta)} = \left\{ \frac{\theta}{1 - \theta} \right\}^{\sum_{i=1}^n Y_i - \sum_{i=1}^n X_i} .$$

Hence, given $x = (x_1, \ldots, x_n)$, the sets in the minimal sufficient partition $A(x)$ are given as $A(x) = \{y = (y_1, \ldots, y_n) : \sum_{i=1}^n x_i = \sum_{i=1}^n y_i\}$. Hence $T = \sum_{i=1}^n X_i$ is minimal sufficient for θ.

Note: Of course, in this simple example we could have argued that $T = \sum_{i=1}^n X_i$ must be minimal sufficient simply by dimension considerations (*we know that T is sufficient and is one-dimensional and you cannot further reduce the dimension that is already equal to one*). However we went directly through the original definition of minimal sufficiency, as well, to confirm out findings.

ii) i.i.d.normal with unknown μ and σ^2 : $(N(\mu, \sigma^2))$. Minimal sufficient statistic for $\theta = (\mu, \sigma^2)'$ is the vector statistic $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)'$.;

iii) i.i.d. uniform in $(0,\theta)$: $f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta), x > 0, \theta > 0$. Minimal sufficient statistic is $X_{(n)}-$ the maximal of the n observations.

iv) i.i.d. Cauchy(θ) - an example that shows that sometimes the dimension of the minimal sufficient statistics can be quite large, even equal to the sample size n itself. The minimal sufficient statistics is the vector of ordered observations

$$T = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})'$$

and its dimension n cannot be reduced any further.

3.8 A very important general example: One parameter exponential family densities

A density $f(x,\theta)$ is a **one parameter exponential family density** if $\theta \in \Theta \in R^1$ and

$$f(x,\theta) = a(\theta)b(x)\exp(c(\theta)d(x))$$

with $c(\theta)$ strictly monotone. Obviously, in this case we have:

$$\frac{L(x,\theta)}{L(y,\theta)} = \prod_{i=1}^{n} \frac{b(x_i)}{b(y_i)} \exp\{c(\theta)[\sum_{i=1}^{n} d(x_i) - \sum_{i=1}^{n} d(y_i)]\}$$

which is *not* a function of θ iff $\sum_{i=1}^n d(x_i) = \sum_{i=1}^n d(y_i)$. So, if x is any point in $\mathcal{X}$ then $A(x) = \{y : \sum_{i=1}^n d(x_i) = \sum_{i=1}^n d(y_i)\}$ and the sets in the minimal sufficient partition are contours of $\sum_{i=1}^n d(x_i)$. Hence, $T = \sum_{i=1}^n d(x_i)$ is minimal sufficient.

Note: Quite a lot of the standard distributions considered in your undergraduate courses can be seen to belong to the one parameter exponential family. Try to convince yourself that each of the following distributions is such:

- $f(x,\theta) = \theta\exp(-\theta x), x > 0, \theta > 0$
- Poisson(θ)
- Bernoulli (θ);
- $N(\theta, 1)$;
- $N(0, \theta^2)$

and others. Note, however , that there are many distributions outside the above class, too. For example, the uniform $(0,\theta)$ distribution or the Cauchy distribution do *not* belong to exponential family.

3.9 Generalization to a k- parameter exponential family)

It is natural to define the $k-$parameter exponential family ($k \geq 1$) via:

$$f(x;\theta_1,\ldots,\theta_k) = a(\theta_1,\ldots,\theta_k)b(x)\exp(\sum_{j=1}^{k} c_j(\theta_1,\ldots,\theta_k)d_j(x))$$

where $c_j(.)$ are certain smooth functions of the k-dimensional parameter vector $\theta = (\theta_1,\ldots,\theta_k)'$. In order to avoid degenerate cases, it is also requested that the $k \times k$ matrix of partial derivatives $\left\{\frac{\partial c_j}{\partial \theta_{j'}}\right\}, j = 1,\ldots,k; j' = 1,\ldots,k$ has a non-zero determinant. Minimal sufficient (vector) statistic: $T = (\sum_{i=1}^n d_1(X_i),\ldots,\sum_{i=1}^n d_k(X_i))'$.

Example of a two-parameter exponential family:

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

We have $d_1(x) = x, d_2(x) = x^2$ and $\bar{T} = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)'$ is minimal sufficient for $\theta = (\mu,\sigma^2)'$.

More details, examples and discussions-at lecture.

3.10 Ancillary Statistic. Ancillarity principle

We consider again $X = (X_1, X_2, \ldots, X_n)$: i.i.d. with $f(x,\theta), \theta \in R^k$.

Definition 2. A statistic is called ancillary if its distribution does not depend on θ.

Intuitively, *alone,* the knowledge of an ancillary statistic should not help in inference about θ. It is therefore even more interesting that an ancillary statistic, when used *in conjunction* with another statistic, sometimes *does help* in inference about θ. Inference for θ could be improved in general, if it is done *conditionally* on the ancillary statistic.

The ancillarity principle. The most important case where the above situation can occur is when a statistic $\mathbf{T}$ is minimal sufficient for θ but its dimension is *greater* than that of θ. Sometimes, we can write $\mathbf{T}$=($\mathbf{T}_1'$,$\mathbf{T}_2'$)$'$ where $\mathbf{T}_2$ has a marginal distribution not depending on θ. The distribution of $\mathbf{T}_2$ is the same for all $P_\theta \in \mathcal{P}$. Then $\mathbf{T}_2$ is ancillary and $\mathbf{T}_1$ is *conditionally sufficient given* $\mathbf{T}_2$:

$$L(\mathbf{x},\theta) \propto L_1(\mathbf{t}_1 \mid \mathbf{t}_2,\theta)L_2(\mathbf{t}_2)$$

Then ancillarity principle postulates that inference about θ then should be based on $L_1(\mathbf{t}_1 \mid \mathbf{t}_2,\theta)$, ie., the inference about θ should be based on the conditional distribution of $\mathbf{T_1}$ given $\mathbf{T_2} = \mathbf{t_2}$.

Often, the precision of the inference about θ, as provided by such conditionality, varies with values of $\mathbf{T}_2$. In a sense, this is similar to the varying precision of samples of different sizes, n.

3.11 Examples

3.11.1

Assume that n i.i.d. $X_1, \ldots, X_n$ are given from the uniform in $(\theta, 1+\theta)$ distribution. You can easily show that the minimal sufficient statistic for θ is $T = (X_{(1)}, X_{(n)})$. Since any 1-to -1 transformation is also minimal sufficient then $T^* = (X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)})$ is also minimal sufficient.

Denoting $Z_i = X_i - \theta$ we see that Z_i are i.i.d. uniformly distributed in $[0, 1]$ and their distribution does not involve θ. We see that

$$P(X_{(n)} - X_{(1)} < r) = P[(X_{(n)} - \theta) - (X_{(1)} - \theta) < r] = P(Z_{(n)} - Z_{(1)} < r).$$

Therefore only the distribution of the largest and of the smallest order statistic from *uniform in (0,1)* distribution is involved with no dependence on θ whatsoever. Hence the first component $X_{(n)} - X_{(1)}$ of the minimal sufficient statistic turns out to be ancillary statistic.

If you were to make inference about θ you would like to take note of value of $X_{(n)} - X_{(1)}$ and to condition on it. Intuitively, if this value is close to 0 then your inference about θ (based on $\frac{1}{2}(X_{(1)} + X_{(n)} - 1)$ for example) may be very unprecise but it would be very precise if $X_{(n)} - X_{(1)}$ was close to 1.

3.11.2

Inference in case of normal mixture. Assume that the density of Y is fiven by $f_Y(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-0.5(y-\mu)^2/\sigma_1^2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-0.5(y-\mu)^2/\sigma_2^2}$. If we also observe an indicator random variable C (with values 1 or 2 telling us whether the first or the second component of the mixture has been observed) then it becomes clear which is the distribution that has generated Y. Hence the joint distribution is

$$f_{C,Y}(c, y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_c}e^{-0.5(y-\mu)^2/\sigma_c^2}$$

The statistic $S = (C, Y)$ is sufficient for μ when σ_1^2, σ_2^2 are assumed known. Moreover since $P(C = 1) = P(C = 2) = 0.5$, C is ancillary. Conditioning on C can definitely help in our inference about μ.

3.11.3

Let $X = (X_1, X_2, \ldots, X_n)$ be i.i.d. from a location family with cdf $F_\theta(x) = F(x - \theta) = F_0(x - \theta), \theta \in R^k$. Consider $\mathbf{T}_2 = X_2 - X_1$. This statistic is ancillary. First, note that if $X \sim F_\theta$, then $X - \theta \sim F_0$ since:

$$F_{X-\theta}(x) = P(X - \theta < x) = P(X < x + \theta) = F_\theta(x + \theta) = F_0(x + \theta - \theta) = F_0(x)$$

Hence,the distribution of $X_i - \theta, i = 1, 2, \ldots, n$ does not depend on θ.

But $F_{\mathbf{T}_2}(y, \theta) = P_\theta(\mathbf{T}_2 < y) = P\{[(X_2 - \theta) - (X_1 - \theta)] < y\}$ and the latter expression obviously does not depend on θ. Hence $\mathbf{T}_2$ is ancillary.

Along the same lines, $\tilde{T_2} = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is also ancillary.

Definition 3. The statistic $\hat{\theta}(X)$ is called **equivariant**, if

$$\hat{\theta}(X_1 + C, X_2 + C, \ldots, X_n + C) = \hat{\theta}(X_1, X_2, \ldots, X_n) + C$$

for any vector C with appropriate dimension.

The importance of ancillarity can be illustrated in many statements about efficiency of estimators based on conditional inference. We will only formulate one of these famous results (due to an Australian Statistician).

Theorem 3.2. *If $\tilde{\theta}$ is* **any** *equivariant estimator with $E_0\tilde{\theta} < \infty$ then the estimator*

$$\hat{\theta}_P = \tilde{\theta} - E_0(\tilde{\theta}|\tilde{T_2}), \tilde{T_2} = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$$

is the **best equivariant** *estimator (i.e. with uniformly smallest risk with respect to quadratic loss among all equivariant estimators). (It is the so called* **Pitman** *estimator.)*

It can be shown that for square error loss, and $\theta \in R^1$ the Pitman estimator is given in a closed form as

$$\hat{\theta}_P = \frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^n f(X_i - \theta) d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^n f(X_i - \theta) d\theta}$$

where $f(x - \theta) = f_\theta(x)$ denotes the density of a single observation. We see that the Pitman estimator has a form of a Bayes estimator with respect to an (*improper*) prior on $(-\infty, \infty)$.

Exercise: Show that when $X_i, i = 1, 2, \ldots, n$ are in addition normally distributed, the Pitman estimator coincides with $\bar{X}$.

3.12 Maximum Likelihood Inference

3.12.1 Likelihood principle

Let $\mathbf{X} = (X_1, X_2, .., X_n)$ be i.i.d. each with density $f(x, \theta)$. Given an observation $\mathbf{x}$ of $\mathbf{X}$, we substitute in $L(\mathbf{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$ which becomes a function of θ only. This is called

the *Likelihood function*. Other functions of θ in the form $c(\mathbf{x})L(\mathbf{x},\theta)$ can also be called likelihood functions. Not that if $T(\mathbf{X})$ is sufficient for θ then $L(\mathbf{x},\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$ holds (Factorization criterion) and thus the *maximum likelihood estimator* $\hat{\theta}$(that maximizes L or, equivalently, g w.r. θ) will be a function of every sufficient statistic. In particular, the Maximum Likelihood Estimator will be a function of the *minimal sufficient statistic* when the latter exists.

Let us remember now that if two points $\mathbf{x}$ and $\mathbf{y}$ are in the same set in the minimal sufficient partition then $L(\mathbf{y},\theta) = h(\mathbf{y},\mathbf{x})L(\mathbf{x},\theta)$, which means they give rise to proportional likelihood functions and the same value of the minimal sufficient statistic. These values $\mathbf{x}$ and $\mathbf{y}$ must lead to the same inference about θ. An even stronger version of this requirement is the *weak likelihood principle:*

"Data sets with proportional likelihood functions lead to identical conclusions".

We say the version is "stronger" since it does not necessitate the sampling processes to be identical. One could have *different sampling processes* A and B that lead to likelihood functions $L_A(\mathbf{x},\theta)$ and $L_B(\mathbf{y},\theta)$. As long as $\frac{L_A(\mathbf{x},\theta)}{L_B(\mathbf{y},\theta)}$ does not depend on θ, inference about θ should be the same.

3.12.2 Example

i) In an experiment A, we observe $\mathbf{x} = (x_1, x_2, .., x_n) : n$ i.i.d. Bernoulli with parameter θ. Then $L_A(\mathbf{x},\theta) = \theta^k(1-\theta)^{n-k}$ if it happened that there were k outcomes equal to one in $\mathbf{x}$.

ii) In an experiment B, we only observe one realization $\mathbf{y}$ of a single random variable $\mathbf{Y}$ =number of successes in n i.i.d. Bernoulli trials. Then $L_B(\mathbf{y},\theta) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k(1-\theta)^{n-k}$ if it happened that $\mathbf{y}$ =k.

iii) In an experiment C we observe realization of a random variable $\mathbf{Z}$ - number of trials until k successes occurred. It is known that $P_\theta(\mathbf{Z} = z) = \begin{pmatrix} z-1 \\ k-1 \end{pmatrix} \theta^k(1-\theta)^{z-k}, z = k, k+1, \ldots$ Here the number of trials is random but if it happened that $z = n$ then $L_C(n,\theta) = \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} \theta^k(1-\theta)^{n-k}$.
Hence, in all three cases considered we would get proportional likelihood functions for specific realizations of the variables $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ and the conclusions about θ in these circumstances must be identical.

3.13 Maximum Likelihood Estimation-introduction

We have discussed some important principles: *sufficiency, ancillarity, weak likelihood principle* on which inference should be based. Each of these looks reasonable as a principle but it does not give us a *constructive procedure* for finding reasonable estimators of the

parameter of interest. Such well known procedure is the Maximum Likelihood Estimation Method. It is defined as $\hat{\theta} = \arg[\sup_{\theta\in\Theta} L(\mathbf{x},\theta)]$. In the discrete case the interpretation of the above maximization is that we look at the model that makes the observed data most likely (probable). Because here it goes about comparing different models, it makes sense to introduce the quantity $R(\mathbf{x},\theta) = \frac{L(\mathbf{x},\theta)}{L(\mathbf{x},\hat{\theta})}$ which has a range of [0,1]. This quantity is called **normed likelihood.** For a fixed $\mathbf{x}$, it is just a function of θ and we shall sometimes denote it by $R(\theta)$. For example, if a coin is tossed 100 times and yields, say, 32 heads then the maximum likelihood estimator $\hat{\theta}$ of the probability of a head to occur is $\hat{\theta} = .32$ and $R(\theta) = \frac{\theta^{32}}{.32^{32}}.\frac{(1-\theta)^{68}}{.68^{68}}$. An even more often used measure is the *deviance* $D(\theta)$ which is defined as $D(\theta) = -2\text{lnR}(\theta) = -2[\text{lnL}(\mathbf{x},\theta)\text{-lnL}(\mathbf{x},\hat{\theta})]$. The deviance is a non-negative number. The larger the deviance, the further the model under consideration from the most likely model, in the set under study, given the observed data. This observation can be used to construct confidence intervals for the parameter. We shall come back to this later.

3.14 Information and Likelihood

Now, we would like to quantify the notion of **Fisher Information** in a single observation and in the whole data vector with respect to the parameter of interest. Having done this in a proper way, we will be able to demonstarte quantitatively (with numbers) that indeed when we are using sufficient statistic, we are preserving the information about the parameter that is contained in the whole sample. On the contrary, if we are not using a sufficient statistic, we are losing some of the information that is contained in the whole sample.

3.14.1 Score function

We define $V(\mathbf{X},\theta) = \frac{\partial}{\partial\theta}\text{logL}(\mathbf{X},\theta)$ to be the *score function*. Generally speaking, it can be defined in the above way even for non-i.i.d. random variables where $L(\mathbf{X},\theta)$ is the joint density. In the case where $\mathbf{X} = (X_1, X_2, .., X_n)$ and the X_i are i.i.d. with a density $f(x,\theta)$ then $V(\mathbf{X},\theta) = \sum_{i=1}^{n} \frac{(\partial/\partial\theta)f(X_i,\theta)}{f(X_i,\theta)}$. Note also that obviously, for the maximum likelihood estimator (MLE)$\hat{\theta}$: $V(\mathbf{x},\hat{\theta}) = 0$ holds. Also, the property $E_\theta(V(\mathbf{X},\theta)) = 0$ holds under suitable regularity conditions (proof: at lecture).

3.14.2 Expected Fisher Information about θ contained in the vector X

It is denoted by $I_{\mathbf{X}}(\theta)$ and is defined as $I_{\mathbf{X}}(\theta)$ =$\text{Var}_\theta(V(\mathbf{X},\theta)) = E_\theta\{\frac{\partial}{\partial\theta}\text{lnL}(\mathbf{X},\theta)\}^2$(where we utilized the fact that $E_\theta(V(\mathbf{X},\theta)) = 0$).

3.14.3 Some properties of information

i) additivity over independent samples: if X and Y are independent random variables whose densities depend on θ then for the information in the vector $\mathbf{Z} = (X, Y)$ we have

$$I_{\mathbf{Z}}(\theta) = I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

In particular, when sampling n times, the information in the sample about the parameter equals n times the information in a single observation about the parameter: If $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ then

$$I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta).$$

ii) If $T(X)$ is sufficient for θ then $I_T(\theta) = I_X(\theta)$

iii) Under regularity conditions: $I_X(\theta) = -E(\frac{\partial^2}{\partial\theta^2} \ln L(X, \theta))$

iv) For any statistics $T(X)$ it holds: $I_T(\theta) \leq I_X(\theta)$ with equality if and only if T is sufficient for θ. This property most clearly underlines the importance of sufficiency when we try to perform data reduction without loss of information about the parameter!

Sketch of proofs (full discussion: at lecture).

i) Starting with $L_{(X,Y)}(x, y; \theta) = L_X(x; \theta)L_Y(y; \theta)$, we take logarithms of both sides first and then calculate partial derivatives with respect to θ of both sides. In the resulting equality, we square both sides and take expected values. This gives us:

$$E_\theta([\frac{\partial}{\partial\theta} \log L_{(X,Y)}(X, Y, \theta)]^2 = I_X(\theta) + I_Y(\theta) + 2E_\theta[V(X, \theta)V(Y, \theta)].$$

Since X and Y are independent:

$$E_\theta[V(X, \theta)V(Y, \theta)] = E_\theta V(X, \theta)E_\theta V(Y, \theta) = 0$$

holds (using the property of the score) and we end up with $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$.

ii) (for the discrete case) First, let us note that because of the sufficiency,

$$f_T(t, \theta) = \sum_{x:T(x)=t} f_X(x, \theta) = \sum_{x:T(x)=t} g(T(x), \theta)h(x) = g(t, \theta) \sum_{x:T(x)=t} h(x)$$

holds and hence $E_\theta[\frac{\partial}{\partial\theta} \log f_T(T; \theta)]^2 = E_\theta[\frac{\partial}{\partial\theta} \log g_T(T; \theta)]^2$ holds. Then

$$I_T(\theta) = E_\theta[\frac{\partial}{\partial\theta} \log f_T(T; \theta)]^2 = E_\theta[\frac{\partial}{\partial\theta} \log g(T; \theta)]^2 =$$

$$E_\theta[\frac{\partial}{\partial\theta}(\log g(T; \theta) + \log h(X))]^2 = E_\theta[\frac{\partial}{\partial\theta} \log L(X; \theta)]^2 = I_X(\theta).$$

iii) If $f(x, \theta)$ denotes the density of a single observation and under suitable differentiability conditions, we can write:

$$\frac{\partial^2}{\partial\theta^2}(\log f(x, \theta)) = \frac{\frac{\partial^2}{\partial\theta^2} f(x, \theta)}{f(x, \theta)} - [\frac{\frac{\partial}{\partial\theta} f(x, \theta)}{f(x, \theta)}]^2$$

For the case of a sample size $n = 1$, we see that if we take expected values in the above equality, property iii) would be shown if we are able to show that

$$E_\theta[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}] = 0$$

holds. But under suitable regularity conditions that allow for exchange of order of integration and differentiation, we have

$$E_\theta[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}] = \frac{\partial^2}{\partial\theta^2}\int f(x,\theta)dx = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

Therefore statement iii) is shown for the case $n = 1$. For the case of arbitrary sample size, we use the additivity of the information over independent samples to get $I_X(\theta) = -E(\frac{\partial^2}{\partial\theta^2}\ln L(X,\theta))$.

iv) To show this property, we need two properties of conditional expected values which we now state first. For random variables Z and Y and a function $g(z)$ we can write (under "suitable conditions")

$$E(g(Z)Y|Z = z) = g(z)E(Y|Z = z) \tag{1}$$

$$E(Y) = E_Z(E(Y|Z = z)) \tag{2}$$

Since the expected value of the square of any random variable is non-negative, we know that:

$$0 \leq E\{\frac{\partial}{\partial\theta}\log L(X,\theta) - \frac{\partial}{\partial\theta}\log f_T(T,\theta)\}^2 = I_X(\theta) + I_T(\theta) - 2E[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)] \tag{3}$$

If we were able to show in (3) that

$$E[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)] = I_T(\theta) \tag{4}$$

holds then from (3) we would have as a consequence that $I_X(\theta) - I_T(\theta) \geq 0$ holds which means that $I_T(\theta) \leq I_X(\theta)$, that is, the information in the statistic can not exceed the information in the sample. Now we concentrate on showing (4). Using properties (1) and (2) we can write

$$E[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)] = E_T[\frac{\partial}{\partial\theta}\log f_T(t,\theta)E(\frac{\partial}{\partial\theta}\log L(X,\theta)|T = t)] \tag{5}$$

Try to show now as an exercise that $E(\frac{\partial}{\partial\theta}\log L(X,\theta)|T = t) = \frac{\partial}{\partial\theta}\log f_T(t,\theta)$ holds. Then substitution in (5) shows that (4) holds.

We can also see from the derivations that the only way in which we may end up with a true equality $I_T(\theta) = I_X(\theta)$ is if we had equality in (4). But this is only possible if $\frac{\partial}{\partial\theta}\log L(X,\theta) = \frac{\partial}{\partial\theta}\log f_T(T,\theta)$ holds. In turn, this means that the difference $\log L(X,\theta) - \log f_T(T,\theta)$ does **not** depend on θ. If we denote this difference by $\log h(X)$, say, then we see that $\log L(X,\theta) = \log f_T(T,\theta) + \log h(X)$ holds. This also means that $L(X,\theta)$ can be factorized as in the Neyman Fisher criterion and T must be sufficient. That is, the only way in which the information in the statistic T can equal the information in the whole sample is if T is sufficient.

4 Lecture 4: Classical Estimation Theory

4.1 Cramer-Rao Inequality

Obtaining a point estimator of the parameter of interest is usually the first step in inference. Suppose $\mathbf{X} = (X_1, X_2, .., X_n)$ are i.i.d. from $f(x, \theta), \theta \in R$ and we use a statistic $T_n(\mathbf{X})$ to estimate θ. If $E_\theta(T_n) = \theta + b_n(\theta)$ then the quantity $b_n(\theta)$ is called *bias*. Note that it generally may depend on both θ and the sample size although this dependence may sometimes be suppressed in the notation. We would hope for a zero bias for all θ and n, called *unbiasedness*. When used repeatedly, an unbiased estimator, in the long run, will estimate the true value on average.

Caution: Note, however , that for some families an unbiased estimators may not exist or, even when they exist, may not be very useful. For example, in the case of the geometric distribution $f(x, \theta) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots$ an unbiased estimator of θ, say, $T(x)$ must satisfy $\sum_{x=1}^{\infty} T(x)\theta(1-\theta)^{x-1} = \theta$ for all $\theta \in [0, 1]$. By a polynomial expansion, the only estimator satisfying this requirement would be $T(1) = 1, T(x) = 0$ if $x \geq 2$. Cancelling θ on both sides, and setting $\tilde{\theta} = 1 - \theta$ we get

$$\sum_{x=1}^{\infty} T(x)\tilde{\theta}^{x-1} = 1 \text{ for all } \tilde{\theta} \in [0, 1]$$

Hence, the only estimator satisfying this requirement would be $T(1) = 1, T(x) = 0$ if $x \geq 2$. Having in mind the interpretation of θ (probability of success in a single trial), such an estimator is neither very reliable, nor very useful.

As an exercise, show that the MLE is $\hat{\theta} = 1/x$. It is biased but makes much more sense!

When looking for an estimator of a "good" quality, we are inclined to analyse the *mean squared error*

$$MSE_\theta(T_n) = E_\theta(T_n - \theta)^2 = Var_\theta T_n + (b_n(\theta))^2.$$

Remark: The following property holds:

$$MSE_\theta(T_n) = Var_\theta T_n + (b_n(\theta))^2$$

Indeed,

$$MSE_\theta(T_n) = E_\theta[(T_n - E_\theta T_n + E_\theta T_n - \theta]^2 =$$

$$E_\theta(T_n - E_\theta T_n)^2 - 2E_\theta[(T_n - E_\theta T_n)(\theta - E_\theta T_n)] + E_\theta(E_\theta T_n - \theta)^2 = Var(T_n) + (b_n(\theta))^2.$$

A small mean squared error as a criterion for choosing a point estimator, is in general more important than unbiasedness. To perform optimally, we would try to find an estimator that minimizes the MSE. Unfortunately, in the class of *all* estimators, an estimator that minimizes the MSE simultaneously *for all θ values*, does not exist!

(Indeed, take *any* estimator $\tilde{\theta}$. Since the parameter $\theta \in \Theta$ is unknown, there will be certain value $\theta_0 \in \Theta$ for which $MSE_{\theta_0}(\tilde{\theta}) > 0$. Then we can consider as a competitor to $\tilde{\theta}$ the estimator $\theta^* \equiv \theta_0$. Note that θ^* is not a very reasonable estimator (it does *not*

even use the data (!)) *but* for the *particular* point θ_0 we have $MSE_{\theta_0}(\theta^*) = 0$ and hence, $MSE_{\theta_0}(\tilde{\theta}) > MSE_{\theta_0}(\theta^*)$.)

With other words, when considering the class of *all* estimators, there are so many estimators available to us that to find a single one that is *uniformly* better with respect to the MSE criterion, is just impossible. Way out of this situation is either to restrict the class of estimators considered, or to change the evaluation criterion. We shall be dealing with the first way out right now (the other way was discussed *already* in the Decision theory chapter: Bayes and minimax estimation).

We choose to impose the criterion of unbiasedness. This greatly simplifies the task of minimizing the mean squared error because then, we only have to minimize the variance. In the (smaller subset of unbiased estimators) one can very often find an estimator with the smallest MSE(=Var) *for all θ values*. It is called the uniformly minimum variance unbiased estimator (UMVUE). Let us first look at a well-known result that will help us in our search of the UMVUE (the **Cramer-Rao** theorem).

Theorem 4.1. *Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ have a distribution that depends on θ and $L(\mathbf{X}, \theta)$ be the joint density. Let $\tau(\theta)$ be a smooth (i.e. differentiable) function of θ that has to be estimated. Consider any unbiased estimator $W(\mathbf{X})$ of $\tau(\theta)$, i.e. $E_\theta W(\mathbf{X}) = \tau(\theta)$. Suppose, in addition, that $L(\mathbf{X}, \theta)$ satisfies:*

$$\frac{\partial}{\partial\theta}\int ..\int h(\mathbf{X})L(\mathbf{X},\theta)dX_1..dX_n = \int ..\int h(\mathbf{X})\frac{\partial}{\partial\theta}L(\mathbf{X},\theta)dX_1..dX_n \qquad (*)$$

for any function $h(\mathbf{X})$ with $E_\theta|h(\mathbf{X})| < \infty$. Then:

$$Var_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{I_{\mathbf{X}}(\theta)}$$

for all θ holds.

Proof: The proof is elegantly simple and is a clever application of the Cauchy- Schwartz inequality. In the setting of random variables, this Inequality is equivalent to the following statement:

Cauchy-Schwartz Inequality: If Z and Y are two random variables with finite variances $Var(Z)$ and $Var(Y)$ then

$$[Cov(Z,Y)]^2 = \{E[(Z - E(Z))(Y - E(Y))]\}^2 \leq Var(Z)Var(Y)$$

holds.

To prove the Cramer-Rao Theorem, we choose W to be the Z-variable, and the score V to be the Y-variable in the Cauchy-Schwartz Inequality. Since $E_\theta V(\mathbf{X}, \theta) = 0$ holds for the score, we have that

$$[Cov_\theta(W,V)]^2 = [E_\theta(WV)]^2 \leq Var_\theta(W)Var_\theta V. \qquad (6)$$

Substituting the definition of the score, we get:

$$Cov_\theta(W,V) = E_\theta(WV) = \int ... \int W(\mathbf{X})\frac{\frac{\partial}{\partial\theta}L(\mathbf{X},\theta)}{L(\mathbf{X},\theta)}L(\mathbf{X},\theta)d\mathbf{X}$$

where $d\mathbf{X} = dX_1 dX_2 \dots dX_n$ is used as shorthand notation. Now if we utilise condition (*), we can continue to get:

$$Cov_\theta(W, V) = \frac{\partial}{\partial\theta} E_\theta W = \frac{\partial}{\partial\theta}\tau(\theta).$$

Then, Inequality (6) implies:

$$Var_\theta(W) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{I_{\mathbf{X}}(\theta)}. \tag{7}$$

Also, in lectures, we will discuss the *multivariate version of this inequality*, applicable for the case of estimating a multidimensional parameter.

4.1.1. Note: The Cramer-Rao (CR) Inequality was stated for continuous random variables. By an obvious modification of condition (*) requiring the ability to interchange differentiation and summation (instead of differentiation and integration) one can formulate this for discrete random variables, too. Note that in this case, even though $L(\mathbf{X}, \theta)$ may not be differentiable w.r. to x, it has to be assumed to be differentiable w.r. to θ.

4.1.1 Corollary for i.i.d. case.

If $\mathbf{X} = (X_1, X_2, \dots, X_n)$ are i.i.d. with $f(x, \theta)$ then $L(\mathbf{X}, \theta) = \prod_{i=1}^n f(X_i, \theta); I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$ and the CR Inequality becomes:

$$\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{nI_{X_1}(\theta)}$$

4.2 Comments on applying the CR Inequality in the search of the UMVUE

a) In case there exists an unbiased estimator of $\tau(\theta)$ whose variance is equal to the lower bound given by CR Inequality, this will be the UMVUE of $\tau(\theta)$. Such a situation occurs often in the case of observations that come from an exponential family;

b) Let us note the drawback related to the fact that condition (*) in the CR Theorem is a strong one and it often happens that it is not satisfied. A typical situation is when the range of the random variables $X_i, i = 1, 2, .., n$ depends on θ, for example, in the case of a random sample from uniform $[0, \theta)$ observations. According to the general Leibnitz$'$ rule for differentiation of parameter-dependent integrals:

$$\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{\partial}{\partial\theta}b(\theta) - f(a(\theta),\theta)\frac{\partial}{\partial\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}f(x,\theta)dx$$

holds and we see that, if a and b were genuine functions of θ, on the RHS there would be some additional non-zero terms included and condition (*) would not hold.

The following illustrative example will be discussed in detail at the lecture: Assume that the right-hand limit θ of the interval $[0, \theta)$ is to be estimated and n i.i.d. observations from the uniform in $[0, \theta)$ distribution are given. It is easy to show that the density of $Y = X_{(n)}$ is given by

$$f_Y(y, \theta) = \begin{cases} ny^{n-1}/\theta^n, \text{ if } 0 < y < \theta \\ 0 \text{ else} \end{cases}$$

Then we can calculate easily that $E[\frac{n+1}{n} X_{(n)}] = \theta$ holds, that is, $\frac{n+1}{n} X_{(n)}$ is an unbiased estimator of $\tau(\theta) = \theta$. It is also easy to calculate its variance: $Var(\frac{n+1}{n} X_{(n)}) = \frac{1}{n(n+2)}\theta^2$ holds. The latter value is **less** than the value $\frac{\theta^2}{n}$ we would get if we recklessly calculated the CR bound ignoring the fact that the regularity condition (*) is in fact **violated** in this example (since the support of the density depends on the unknown parameter).

c) Even in cases where the CR Theorem is applicable, there is no guarantee that the lower bound on the variance is attainable. In fact, looking at the proof of the theorem, we can see that the bound is achievable iff one has an equality in the Cauchy-Schwartz Inequality which means that the score $V(\mathbf{X}, \theta)$ must have a representation of the form $V(\mathbf{X}, \theta) = k_n(\theta)[W(\mathbf{X}) - \tau(\theta)]$. If, alternatively, $V(\mathbf{X}, \theta)$ can not be written in this form then no unbiased estimator of $\tau(\theta)$ would have a variance equal to the one given by the CR bound and in this case the CR Inequality would be of no use when searching for UMVUE. There still might be an UMVUE (with a variance slightly higher than the value given by the CR lower bound) but one would need to develop a method to find it in such situations.

4.3 Examples

(to be discussed at lectures)

a) Estimating the parameter θ in a Poisson(θ) distribution. In this case the CR bound is achievable and the unbiased estimator that achieves it, is $\hat{\theta} = \bar{X}$.

b) Estimating the function $\tau(\theta) = \exp(-\theta)$ from a sample of Poisson(θ) distribution. In this case no unbiased estimator of $\tau(\theta)$ has variance equal to the bound. Nevertheless, UMVUE of $\tau(\theta)$ exists.

Some details:

$$\log L(X; \theta) = \log \left\{ \frac{\theta^{\sum_{i=1}^n x_i} e^{(-n\theta)}}{\prod_{i=1}^n x_i!} \right\} = -n\theta + (\sum_{i=1}^n x_i) \log \theta - \sum_{i=1}^n \log (x_i!) \qquad (8)$$

Taking derivatives with respect to θ in (3):

$$\frac{\partial}{\partial \theta} \log L(X; \theta) = -n + \sum_{i=1}^n x_i/\theta = n \exp(\theta)[\frac{1}{\theta} \exp(-\theta)\bar{x} - \exp(-\theta)]$$

and this can not be represented as $k(\theta, n)[$ statistic $- \exp(-\theta)]$ Formal calculation of the Cramer-Rao bound gives $\frac{\theta}{n}e^{(-2\theta)}$ (check (!)) but this bound is not attainable by any unbiased estimator of $\tau(\theta) = \exp(-\theta)$.

Nevertheless UMVUE does exist and is given (as we shall see later) by $T = (1-\frac{1}{n})^{n\bar{X}}$. Another method needs to be applied to finding this UMVUE (see Theorem of Lehmann-Scheffe below).

4.4 Which are the estimators that could attain the bound?

Theorem 4.2. *If under the regularity conditions of CR Theorem there is an estimator of $\tau(\theta)$ which attains the lower bound, it should be the MLE of $\tau(\theta)$.*

Proof: At lectures.

Conclusion: When looking for UMVUE, it is a good idea to calculate the MLE first. If the MLE turns out to be unbiased and its variance equals the one given by the CR bound, the UMVUE has been constructed. Otherwise if either the MLE is biased or does not attain the bound then it is sure that the bound is not attainable at all. Very often in such situations the UMVUE (which necessarily will have variance larger than the one given by the bound) turns out to be a bias-corrected MLE.

To outline a more specific way to construct UMVUE in such more delicate situations, let us first formulate the following famous theorem:

4.5 Rao-Blackwell Theorem

Theorem 4.3. *Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\hat{\tau}(T) = E(W \mid T)$. Then $E_\theta\hat{\tau}(T) = \tau(\theta)$ and $Var_\theta\hat{\tau}(T) \leq Var_\theta W$ for all $\theta \in \Theta$, i.e. $\hat{\tau}(T)$ is uniformly better than W as an estimator of $\tau(\theta)$.*

Proof: at lecture.

Note: $\hat{\tau}(T)$ is a function of a sufficient statistic and has uniformly smaller variance than W. This theorem underlines the importance of sufficient statistics. We can therefore only consider functions of sufficient statistics when looking for UMVUE.

4.6 Uniqueness of UMVUE

Theorem 4.4. *If an estimator W is UMVUE for $\tau(\theta)$, then W is unique. Moreover, W is UMVUE iff W is uncorrelated with all unbiased estimators of zero.*

Proof: at lecture.

Note: Characterization of the estimators that are uncorrelated with any unbiased estimator of zero is therefore important. If it turned out that the family $f(T, \theta)$ of distributions

does have the property that there are *no unbiased estimators of zero* (except the constant zero itself) then our search of UMVUE will be successfully finalized. For such type of characterization , we need the notion of **completeness.**

4.7 Completeness of a family of distributions

Definition 4. Let $\tilde{f}(t,\theta), \theta \in \Theta$ be a family of distributions for a statistic $T(\mathbf{X})$. The family is called *complete* if $E_\theta g(T) = 0$ for all $\theta \in \Theta$ implies $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$. Equivalently, $T(X)$ is called a *complete statistic* for θ.

Note that completeness is a property of the *whole* family of distributions, *not* a property of a particular distribution.

Now we finally can formulate the theorem that allows us to find the UMVUE even in situations when the CR bound is not achievable.

4.8 Theorem of Lehmann-Scheffe

Theorem 4.5. *Let T be a complete sufficient statistic for a parameter θ and W be any unbiased estimator of $\tau(\theta)$. Then $\hat{\tau}(T) = E(W \mid T)$ is the unique UMVUE of $\tau(\theta)$.*

Proof: This theorem is a direct consequence and compilation of the statements we already made. Indeed, according to the Rao-Blackwell Theorem, it suffices to consider as candidates for UMVUE only unbiased functions of the *sufficient* statistic T (otherwise by conditioning T we can improve any other unbiased competitor) and (again due to Rao-Blackwell) $\hat{\tau}(T)$ is such an unbiased estimator of $\tau(\theta)$ (and, of course $\hat{\tau}(T) = E(W \mid T)$ is a function of T.) Because of the completeness of T, no further improvement of $\hat{\tau}(T)$ is possible-hence it is the *unique* UMVUE.

The theorem has very useful applications. In many situations there will be no obvious candidate for the UMVUE of $\tau(\theta)$. But the theorem suggests that if we have found *any* (even if very poor) unbiased estimator of $\tau(\theta)$ and we know a statistic T that is a *complete and sufficient* statistic for θ then $E(W \mid T)$ is the *uniformly best unbiased estimator* of $\tau(\theta)$.

4.9 Examples-

see lectures.

- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. $N(0,\theta)$, $T = \bar{X}$ is *not complete* for θ (note that θ denotes the variance in this example).
- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. Bernoulli with $\theta \in (0,1)$ denoting the probability of success, $T = \sum_{i=1}^n X_i$ is complete for θ and $\bar{X}$ is UMVUE for the expected value parameter θ. Hovewer, for the variance parameter $\theta(1-\theta)$, the UMVUE turns out to be $\bar{X}(1-\bar{X})\frac{n}{n-1}$.

- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. uniform in $[0, \theta)$, the statistic $T = X_{(n)}$ is complete and $\frac{n+1}{n} X_{(n)}$ is UMVUE for θ.
- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. Poisson (θ), the statistic $\sum_{i=1}^{n} X_i$ is complete. For $\tau(\theta) = \exp(-\theta)$, the UMVUE turns out to be $(1 - \frac{1}{n})^{n\bar{X}}$.

5 Lecture 5: Likelihood Inference. First order asymptotics

5.1 Why asymptotics

We realized that finding the UMVUE for a **fixed sample size** n could be difficult in some cases especially when the CR bound is not attainable. Finding them requires some art, and there is no easy to follow constructive algorithm for their determination. On the other hand, the MLE's are typically easy to construct by following a general recipe of optimizing either directly the Likelihood or the log-likelihood function, i.e., by following an easy general recipe. It should be pointed out that sometimes the MLE could be biased or, even if unbiased, could not attain the CR bound when outside the exponential family setting.

Note: It is easy to work with exponential families because their structure directly helps us identify a **minimal sufficient and complete** statistic: once the function $d(x)$ in the definition of the exponential family has been identified, we know that $T = \sum_{i=1}^{n} d(X_i)$ is complete and minimal sufficient. The Lehmann-Scheffe theorem can be used to construct UMVUE in such families.

Yet, it is simpler to work with the MLE's and, as shown in the many examples below, usually the UMVUE are just a bias-corrected MLE.

Indeed, the UMVUE for the variance $\theta(1-\theta)$ of Bernoulli trials was $\bar{X}(1-\bar{X})\frac{n}{n-1}$ whereas the MLE is $\bar{X}(1-\bar{X})$; the UMVUE for the endpoint θ of uniform $(0,\theta)$ distribution was $\frac{n+1}{n}X_{(n)}$ whereas the MLE is $X_{(n)}$; the UMVUE for the probability of no occurrence based on n independent Poisson random variables was $(1-\frac{1}{n})^{n\bar{X}}$ whereas the MLE is $exp(-\bar{X})$.

The bias-correction itself tends to be negligible as the sample size increases. Therefore the UMVUE$'s$ are either MLE$'s$ or "almost" MLE's. Hence, it is justified to look for a strong backing of the properties of MLE$'s$ in a general setting. This can be done using asymptotic arguments, i.e. by looking at the performance of MLE$'s$ when $n \longrightarrow \infty$, i.e. by letting the amount of information become arbitrarily large. Statistical folklore says then that "nothing can beat the MLE asymptotically".

5.2 Convergence concepts in asymptotics

We remind some stochastic convergence concepts first.

An estimator T_n of the parameter θ is said to be:

i) *consistent* (or *weakly consistent*) if

$$\lim_{n\to\infty} P_\theta(\mid T_n - \theta \mid > \epsilon) = 0$$

for all $\theta \in \Theta$ and for every fixed $\epsilon > 0$. We denote this by $T_n \xrightarrow{P} \theta$.

ii) *strongly consistent* if $P_\theta\{\lim_{n\to\infty} T_n = \theta\} = 1$ for all $\theta \in \Theta$.

iii) *mean-square consistent* if $\mathrm{MSE}_\theta(T_n) \longrightarrow_{n\to\infty} 0$ for all $\theta \in \Theta$.

It is important to note that **if the estimator is mean-square consistent then it is also consistent**. This relation has probably the most important practical consequence. The reason is that most often we are interested in weak consistency and a common method that often works in proving it, is by showing mean- square consistency first. To justify the relation between mean-square consistency and consistency we can use the **Chebyshev Inequality.** It states that for any random variable X and any $\epsilon > 0$ it holds for the k-th moment:

$$P(|X| > \epsilon) \leq \frac{E(|X|^k)}{\epsilon^k}$$

Applying this inequality for X being $T_n - \theta$ and $k = 2$ we get

$$0 \leq P(|T_n - \theta| > \epsilon) \leq \frac{MSE_\theta(T_n)}{\epsilon^2}.$$

Therefore, if an estimator T_n is mean-square consistent and the RHS tends to zero, the LHS will also tend to zero thus implying consistency.

Also, **strong consistency implies weak consistency.**

There is one more form of convergence of random variables. It is called convergence in distribution and is the weakest form of convergence. It follows from any of the three convergences discussed above. Not surprisingly it is called a *weak convergence* (or *convergence in distribution*). Assume that the sequence of random variables $X_1, X_2, .., X_n, ..$ have cumulative distribution functions $F_1, F_2, \ldots, F_n, ..$ respectively. Assume the continuous random variable X has a cdf F and that it holds for each argument $x \in R$ that $\lim_{n\to\infty} F_n(x) = F(x)$. Then we say that the sequence of random variables $\{X_n\}, n = 1, 2, \ldots$ converges weakly (or in distribution) to X and denote this fact by $X_n \xrightarrow{d} X$.

5.3 Consistency and asymptotic normality of MLE.

The basic statement about asymptotic properties of MLE follows.

Theorem 5.1. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *be i.i.d. from* $f(x, \theta), \theta \in \Theta \in R^1, \Theta-$ *open interval. Assume, following regularity conditions are satisfied:*

1) $\frac{\partial f}{\partial \theta}(x, \theta), \frac{\partial^2 f}{\partial \theta^2}(x, \theta), \frac{\partial^3 f}{\partial \theta^3}(x, \theta)$ *exist for all* x *and all* $\theta \in \Theta$.

2) $\frac{\partial}{\partial \theta} \int f(x, \theta)dx = \int \frac{\partial}{\partial \theta} f(x, \theta)dx$; $\frac{\partial^2}{\partial \theta^2} \int f(x, \theta)dx = \int \frac{\partial^2}{\partial \theta^2} f(x, \theta)dx$

3) $0 < I(\theta) = E_\theta(\frac{\partial ln f}{\partial \theta}(x, \theta)^2) < \infty$ *for all* $\theta \in \Theta$

4) $|\frac{\partial^3 ln f}{\partial \theta^3}(x, \theta)| \leq H(x)$ *for all* $\theta \in \Theta$ *with* $E_\theta H(X) = \int H(x) f(x, \theta)dx \leq C$, C *not depending on* $\theta \in \Theta$.

Let θ_0 *be the "true" value of* θ. *Then the MLE* $\hat{\theta}_n$ *of* θ_0 *is strongly consistent and asymptotically normal, i.e.*

a) $P_{\theta_0}(X : \hat{\theta_n} \to \theta_0) = 1$

b) $\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$

Proof a) i) First step: notice that

$$\frac{1}{n}\log L(X,\hat{\theta_n}) \geq \frac{1}{n}\log L(X,\theta_0) \tag{9}$$

where $L(.,.)$ denotes the joint density of n independent identically distributed (i.i.d.) observations, each with a density $f(.,.)$.

ii) Notice that by Jensen's Inequality:

$$E_{\theta_0}[\log\frac{L(X,\theta)}{L(X,\theta_0)}] < \log E_{\theta_0}[\frac{L(X,\theta)}{L(X,\theta_0)}] = 0$$

which implies

$$E_{\theta_0}[\frac{1}{n}\log L(X,\theta)] < E_{\theta_0}[\frac{1}{n}\log L(X,\theta_0)]$$

The Law of Large numbers implies then that for a fixed $\theta \neq \theta_0$ we should have

$$P_{\theta_0}\{\lim_{n\to\infty}\frac{1}{n}\log L(X,\theta) < \lim_{n\to\infty}\frac{1}{n}\log L(X,\theta_0)\} = 1 \tag{10}$$

Comparing (9) and (10) we see that we need to have $P_{\theta_0}(\hat{\theta}_n \to \theta_0) = 1$.

b) Since

$$0 = \frac{\partial}{\partial\theta}\log L(\mathbf{X},\theta)_{|\theta=\hat{\theta}_n} = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i,\theta)_{|\theta=\hat{\theta}_n}$$

holds, after Taylor expansion around θ_0 and recollection of terms we get

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = \frac{(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta_0))/[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)]}{1+\frac{1}{2}(\hat{\theta}_n-\theta_0)\frac{\frac{1}{n}\sum_{i=1}^{n}\eta_i H(x_i)}{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)}}.$$

Here η_i are intermediate values: $|\eta_i| < 1$. Now we just have to use:

-the Law of large numbers regarding the term $[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)]$,

-the central limit theorem regarding the term $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta_0)$

-and the uniform bound assumption 4 of the theorem

$$\frac{1}{n}\sum_{i=1}^{n}|\eta_i H(X_i)| \leq \frac{1}{n}\sum_{i=1}^{n}H(X_i) \to E_{\theta_0}H(X_i) \leq C$$

to finish the argument.

Note: The statement of the above theorem can be extended to the *multivariate case.* This is, of course, a crucial step regarding practical applications of the maximum likelihood methodology since in predominant majority of cases, the parameter vector of interest is multi-dimensional. Let now $f(x,\theta), \theta \in \Theta \in R^p$.

To formulate it, we need to extend the notion of Fisher information in a parameter-**vector** $\overrightarrow{\theta}$. For such a vector, we define a Fisher information **matrix in the whole sample** $I_{\mathbf{X}}(\overrightarrow{\theta})$ whose (i,j)th element is defined as

$$E(\frac{\partial}{\partial\theta_i}log\mathbf{L}\frac{\partial}{\partial\theta_j}log\mathbf{L}) = -E(\frac{\partial^2}{\partial\theta_i\partial\theta_j}log\mathbf{L}), i = 1,2,\ldots,p; j = 1,2,\ldots,p.$$

(*For simplicity of notation, we skip the arrow over the parameter even though we are in the multidimensional case*)

Then, for an inner point θ_0 (the "true value" of the parameter space Θ) we have under some regularity conditions on the density (similar to the ones listed in Theorem 5.1):

a) $P_{\theta_0}(X : \hat{\theta}_n \to_{n\to\infty} \theta_0) = 1$

b) $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{X_1}^{-1}(\theta_0))$ (asymptotic normality).

Here $I_{X_1}^{-1}(\theta_0)$ is the information in **one** observation and that $I_{X_1}^{-1}(\theta_0) = nI_{\mathbf{X}}^{-1}(\theta_0)$ holds. Hence, result b) can be also written roughly as

$$\hat{\theta}_n \approx N(\theta_0, I_{\mathbf{X}}^{-1}(\theta_0)) \approx N(\theta_0, \frac{1}{n} I_{X_1}^{-1}(\theta_0)).$$

Even more can be said. It turns out that the limiting variance-covariance matrix $I_{X_1}^{-1}(\theta_0)$ in b) is the *smallest possible*. This is to be interpreted in the sense that (under regularity conditions) any other limiting matrix A_{θ_0} (related to another possible estimator) is such that the difference

$$A_{\theta_0} - I_{X_1}^{-1}(\theta_0) \tag{11}$$

is non-negative definite, that is, has non-negative eigenvalues. We also denote this as $A_{\theta_0} \geq I_{X_1}^{-1}(\theta_0)$.

5.4 Additional comments on asymptotic properties of MLE.

We indicate how the above result can be interpreted as "asymptotic efficiency" of the MLE. For simplicity, start with the case of a *one-dimensional parameter*. Formally, the asymptotic normality of the MLE and the form of the asymptotic variance show that $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_{mle}).[nI_{X_1}(\theta)] = 1$ for all $\theta \in \Theta$ which means that the MLE "asymptotically achieve the CR bound on variance". In fact, there are some additional obstacles in formulating such a claim. First of all, the asymptotic normality claim is about *convergence in distribution* (which was the weakest type of convergence) and it does not immediately follow from this result that also the variances converge. The second obstacle is the existence of the so-called *superefficiency phenomena* (first example of a superefficient estimator has been suggested by Hodges). The examples of superefficient estimators show that it is possible to construct estimators for which the above limit can be even less than one *for a certain small number of θ values*. Nevertheless, by imposing some further reasonable regularity conditions and a suitable interpretation of the convergence to the asymptotic distribution, the above obstacles can be overcome. The details are subtle and will not be discussed in our course. We finish our discussion with the words (since the above difficulties can be overcome): "folklore says that MLE are asymptotically the best (*asymptotically efficient*) estimators meaning that they are asymptotically unbiased and with the smallest possible asymptotic variance".

The asymptotic efficiency in the case of *multi-dimensional* parameter vector is interpreted in a similar way. Using the fact that the MLE is asymptotically centered at the "true value" θ_0 and asymptotically is less spread around this true value than any other of its competitors because of (11) we can claim that the MLE is asymptotically efficient.

5.5 Delta method

5.5.1 Invariance property of MLE.

The main theorem about asymptotic properties of MLE was related to the estimation of the parameter θ itself. Sometimes, certain smooth function (a transformation) of the parameter θ is of interest to us, as we already had a chance to see earlier in this course. If we denote by $h(\theta)$ such a transformation, it is useful to know two things:

- what is the MLE of the new parameter $h(\theta)$.
- what is the asymptotic distribution of the MLE of the new parameter $h(\theta)$.

The answer to the first question shows one of the very useful properties of the MLE. The claim is that the MLE of $h(\theta)$ can be obtained by substitution (plug-in) of the MLE $\hat{\theta}$ of θ in the transformation formula: that is, $h(\hat{\theta})$ is the MLE of $h(\theta)$. This is the *invariance* or better to say *transformation invariance* property of the MLE.

Let us now assume in addition that the transformation $h(\theta)$ is smooth enough. Then we are also able to find the **asymptotic distribution** of $h(\hat{\theta})$. This is a very important result called "the delta method". Let us discuss it below.

5.5.2 Delta method

Since the transformation is assumed to be smooth, we can expand $h(\hat{\theta})$ around the true parameter θ_0 :

$$h(\hat{\theta}_{mle}) = h(\theta_0) + (\hat{\theta}_{mle} - \theta_0)\frac{\partial h}{\partial \theta}(\theta_0) + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2.\frac{\partial^2 h(\theta_0)}{\partial \theta^2} + \dots$$

From here we get the convergence in distribution:

$$\sqrt{n}(h(\hat{\theta}_{mle}) - h(\theta_0)) \xrightarrow{d} N(0, [\frac{\partial h}{\partial \theta}(\theta_0)]^2 I^{-1}(\theta_0))$$

This result is called **"the delta method"**. Roughly, we shall also say that the distribution of $h(\hat{\theta}_{mle})$ can be approximated by

$$N(h(\theta_0), \frac{1}{n}[\frac{\partial h}{\partial \theta}(\theta_0)]^2 I^{-1}(\theta_0)).$$

The delta method has a version applicable for the case where $h(\theta)$ is a smooth transformation of a $p-$dimensional parameter-vector $\overrightarrow{\theta}$.

If we introduce the vector of partial derivatives

$$\nabla h(\overrightarrow{\theta}) = (\frac{\partial}{\partial \theta_1} h(\overrightarrow{\theta}), \dots, \frac{\partial}{\partial \theta_p} h(\overrightarrow{\theta}))'$$

then the distribution of $h(\hat{\overrightarrow{\theta}}_{mle})$ can be approximated by

$$N(h(\overrightarrow{\theta}_0), \nabla h(\overrightarrow{\theta}_0)' I_{\mathbf{X}}(\overrightarrow{\theta}_0)^{-1} \nabla h(\overrightarrow{\theta}_0).$$

(Note that $I_{\mathbf{X}}(\overrightarrow{\theta}) = nI_{X_1}(\overrightarrow{\theta})$ and hence $I_{\mathbf{X}}(\overrightarrow{\theta})^{-1} = \frac{1}{n} I_{X_1}(\overrightarrow{\theta})^{-1}$ in agreement with the one-dimesional case $p = 1$).

5.5.3 Examples

Example 1. (details at lecture). For estimating the parameter $\sqrt{\lambda}$ of the Poisson (λ) distribution using MLE, we get $\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{d} N(0, \frac{1}{4})$. There is an interesting additional observation that should be made in relation to this example. Although the asymptotic normal distribution of the MLE for the parameter λ has a variance that **depends** on the (unknown) parameter λ itself, the asymptotic normal distribution of the *transformed* parameter $h(\lambda) = \sqrt{\lambda}$ has a *constant* variance of $\frac{1}{4}$ independent of the value of λ. Such a transformation $h(.)$ that makes the asymptotic variance independent of the parameter, is called **variance stabilising transformation.** Variance stabilising transformations are actively sought after sometimes, especially for the purpose of constructing confidence intervals with a more precise coverage accuracy.

Example 2 (exponential transformation) Consider $h(\theta) = e^{\theta}$. Assume $X_1, X_2, \ldots, X_n$ are i.i.d. (not necessarily normal) with $E(X_i) = \theta, Var(X_i) = \sigma^2, i = 1, 2, \ldots, n$. Here $h'(\theta) = e^{\theta}$. The delta method tells us that the distribution of $e^{\bar{X}}$ can be approximated by **normal** with mean e^{θ} and variance $\frac{1}{n}\sigma^2 e^{2\theta}$. Using the data, this variance could be estimated by $\frac{1}{n} S^2 e^{2\bar{X}}$ where S^2 is just the sample variance.

Example 3 (Asymptotic distribution of a ratio estimator, this is a variant of the Example 5.5.27 in CB) Suppose X and Y are bivariate normally distributed with a known 2×2 covariance matrix $\Sigma = (\sigma_{ij}, i = 1, 2, j = 1, 2)$ and a mean vector $(\mu_X, \mu_Y)'$. A sample of n observation pairs $(X_i, Y_i)', i = 1, 2, \ldots, n$ is given. We are interested in the asymptotic distribution of the MLE $\bar{X}/\bar{Y}$ of the ratio $\frac{\mu_X}{\mu_Y}$.

Solution: First we note that $h(\mu_X, \mu_Y) = \frac{\mu_X}{\mu_Y}$ and $\frac{\partial}{\partial \mu_X} = \frac{1}{\mu_Y}, \frac{\partial}{\partial \mu_Y} = \frac{-\mu_X}{\mu_Y^2}$. From the first order Taylor expansion we have $E(\frac{\bar{X}}{\bar{Y}}) \approx \frac{\mu_X}{\mu_Y}$. The inverse of the information matrix is

$$I_n(\mu_X, \mu_Y)^{-1} = \frac{1}{n} \left\{ \begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{array} \right\}$$

(WHY(!)) hence aplying the delta method we get

$$\sqrt{n}(\bar{X}/\bar{Y} - \mu_X/\mu_Y) \xrightarrow{d} N(0, (\frac{1}{\mu_Y}, \frac{-\mu_X}{\mu_Y^2}) \left\{ \begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{array} \right\} \left\{ \begin{array}{c} 1/\mu_Y \\ \frac{-\mu_X}{\mu_Y^2} \end{array} \right\})$$

Completing the matrix multiplication we get for the asymptotic variance the expression

$$\frac{\mu_X^2}{\mu_Y^2}(\frac{\sigma_{11}}{\mu_X^2} + \frac{\sigma_{22}}{\mu_Y^2} - 2\frac{\sigma_{12}}{\mu_X \mu_Y}).$$

Another way in which we can state the same result is to say that

$$\bar{X}/\bar{Y} \approx N(\mu_X/\mu_Y, \frac{1}{n}\frac{\mu_X^2}{\mu_Y^2}(\frac{\sigma_{11}}{\mu_X^2} + \frac{\sigma_{22}}{\mu_Y^2} - 2\frac{\sigma_{12}}{\mu_X\mu_Y}).$$

Note that it would have been quite difficult to get exact closed-form expression for the variance whereas the delta method is routinely applicable here.

5.5.4 Exact and asymptotic distributions for the deviance

The deviance $D(\theta) = -2\log\left(\frac{L(\mathbf{X},\theta)}{L(\mathbf{X},\hat{\theta}_{MLE})}\right)$ has an important role in constructing confidence intervals and testing hypotheses about unknown parameters. It is a function of the observations, as well of the (unknown) parameters of interest. For a fixed value of the parameters, the deviance is only a function of the observations (i.e. statistic). Its distribution is of great interest because of the applications mentioned above.

Examples (details at lecture):

- For n i.i.d. observations from a $N(\mu, \sigma^2)$ with σ^2 known, the deviance is $D(\mu) = \frac{n(\bar{x}-\mu)^2}{\sigma^2}$. Suppose the true mean is μ_0 so that $\bar{X} \sim N(\mu_0, \sigma^2/n)$. Then $D(\mu_0) \sim \chi^2(1)$(chi-squared with one d.f.) (this is an *exact* (not asymptotic) result).

- Chi square *approximation* of the deviance. Assume that $\theta = \theta_0$ is the "true" value of the population parameter. Expand the Log-likelihood in Taylor series around $\theta = \hat{\theta}_{mle}$:

  $\text{logL}(\mathbf{X}, \theta_0) = \text{logL}(\mathbf{X}, \hat{\theta}_{mle}) + (\theta_0 - \hat{\theta}_{mle})\frac{\partial \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta} + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2 . \frac{\partial^2 \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta^2} + \ldots$

  Because the second summand in the RHS vanishes at $\hat{\theta}_{mle}$, ignoring higher order terms, we get: $D(\theta_0) \approx (\hat{\theta}_{mle} - \theta_0)^2\{-\frac{\partial^2 \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta^2}\}$.

  Using the results about the normal approximation to the distribution of MLE, we get the deviance has an asymptotic χ^2 distribution with one degree of freedom. More generally, if the parameter $\theta_0 \in R^p$ then, asymptotically,

  $$(\theta_0 - \hat{\theta}_{mle})'\{-\frac{\partial^2 \text{logL}(\mathbf{X}, \hat{\theta}_{mle})}{\partial\theta\partial\theta'}\}(\theta_0 - \hat{\theta}_{mle}) \sim \chi^2_p$$

  Hence, we can easily suggest a **test** of the null hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ with an asymptotic level equal to α. (How (?)) Discussion of this example will be continued later in the course under the heading "Generalized likelihood ratio tests".

- As a further example, we apply the above results in the case of the exponential distribution. Assume that $x_1, x_2, .., x_n$ are i.i.d. realizations from the density $f(x, \theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta}), x > 0$. Then $L(\mathbf{x}, \theta) = \frac{1}{\theta^n}\exp(-\sum_{i=1}^n x_i/\theta)$; the MLE is $\hat{\theta} = \bar{x}$ and the exact deviance is easily seen to be $D(\theta) = 2n[\frac{\bar{x}}{\theta} - \ln(\frac{\bar{x}}{\theta}) - 1]$. The expected information (Fisher information) is $I_{\mathbf{X}}(\theta) = n/\theta^2$ and hence the chi square approximation to the deviance is $D_{approx}(\theta) \approx \frac{n(\theta-\bar{x})^2}{\theta^2}$. Comparing this with the exact value

$D(\theta)$ we see that if we expand $\log(1+y) \approx y - \frac{y^2}{2}$ for $y = \frac{\bar{x}-\theta}{\theta}$ in the exact formula, we would get the above chi square approximation for the deviance. Both statistics ($D(\theta_0)$ and $D_{approx}(\theta_0)$) can be used to test the hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.

6 Lecture 6: Hypothesis Testing

6.1 Motivation

Assume $\mathbf{X} = (X_1, X_2, .., X_n)$ are i.i.d. from $f(x, \theta), \theta \in \Theta$. Point estimation of θ will give an estimated value of θ which will be in general different from the true θ. In fact, if Θ was not a finite set but an interval (as it often happens) then the estimator and the true value will coincide with probability zero! This observation alone is convincing enough to claim that it is not enough just to give a single estimated value of the parameter. The problems of constructing confidence intervals (if the parameter was one-dimensional), or confidence sets (if the parameter was multi-dimensional), and the problems of testing hypotheses about θ naturally arise.

We will mainly constrain ourselves to hypothesis testing and will avoid the thorough discussion of confidence sets. Besides lack of time, the following argument can be put forward to defend this decision. In your introductory Statistics courses you have studied the interrelationship between Hypothesis testing and the construction of Confidence intervals. In particular, given certain α size test of a hypothesis $H_0 : \theta = \theta_0$, and having obtained the sample, we can take the set of the parameter values for which the test "answers" with an acceptance when the sample is substituted in the test statistic. This set of parameter values is a confidence set at level $1 - \alpha$. Symbolically, we can say that the subset in Θ defined via

$$\{\theta' | H_0 : \theta = \theta' \text{ is accepted given realization } \mathbf{X} = \mathbf{x} \text{ of the sample } \}$$

represents a confidence set at level $(1 - \alpha)$ for the unknown parameter θ.

In other words, knowing how to construct tests, we basically also know how to construct confidence sets. Moreover, the usefulness of the relationship between testing hypotheses and confidence sets is further exemplified by the fact that some optimality results carry over. It can be shown quite generally that the above procedure of constructing confidence sets leads to confidence sets with optimality properties if the hypothesis test used in the construction was optimally designed.

6.2 General terminology in relation to hypothesis testing.

Let us start with the case of testing a *simple* hypothesis against a *simple* alternative. This is the easiest case to discuss. Besides, the technique that is being used in this simple case (the **Neyman-Pearson lemma** below) is indeed fundamental and serves as a basis to deal with the more difficult cases, too.

Assume that the unknown parameter θ can be one of the two values $\{\theta_0, \theta_1\}$ only. In other words, we are testing a *simple hypothesis* $H_0 : \theta = \theta_0$ versus a *single alternative* $H_1 : \theta = \theta_1$. A *test* $\varphi(\mathbf{x})$ is defined as

$$\varphi(\mathbf{x}) = P(\text{ reject } H_0 \mid \mathbf{X} = \mathbf{x}).$$

Generally, we would prefer deterministic decisions, i.e. we would like $\varphi(\mathbf{x})$ to be equal to either zero or one. Based on the observations we calculate $\varphi(\mathbf{x})$ and according to its value,

we reject H_0 (if it happens that $\varphi(\mathbf{x}) = 1$) or do not reject it (if $\varphi(\mathbf{x}) = 0$). This means that we need to decompose the sample space $\mathcal{X}$ into two regions A and S :

$$\mathcal{X} = A \cup S; A \cap S = \oslash$$

so that depending on $\mathbf{x} \in A$ or $\mathbf{x} \in S$ we decide for either H_0 or H_1. We shall be looking for a definition of $\varphi(\mathbf{x})$(or equivalently, to a decomposition of $\mathcal{X}$ into A and S) in some sort of optimal way. To define reasonably what optimality could mean in this setting, we need to examine the sorts of errors that we can encounter when deciding to reject or accept H_0. As is well-known from your introductory Statistics course, we can commit 2 types of errors:

- to reject H_0 given that H_0 is correct (first type of error)
- to accept H_0 given that H_1 is correct (second type of error).

The corresponding probabilities are denoted as follows:

$$P(\text{ reject } H_0 \mid H_0 \text{ correct }) = \text{level of the test (significance)}$$

$$P(\text{ accept } H_0 \mid H_1 \text{ correct }) = 1 - (\text{power of the test})$$

It is also well known that minimizing level and maximizing power simultaneously is **not possible**. Because of this well known fact, one possible way out is to formulate a **constrained optimization** problem as follows: we decide to fix certain (small) value α(e.g. $\alpha = 0.005, 0.01, 0.05, 0.10$) (level of significance) that is not allowed to be exceeded for the first type error and in the set of all tests having first type error equal to α, we are looking for the one with a smallest possible second type error (or equivalently the highest possible power).

In the signal processing literature, the first type error is called, not unreasonably, the "false alarm". This is because if the null hypothesis is about no (enemy) signal in the generally recorded noise level then the rejection of the null implies that a false alarm has been raised. You can imagine now that choosing the probability level for a false alarm (α) too high means that too many false alarms could be raised; on the other hand, choosing it too low implies that we are increasing the chance for a signal to be missed. This is why choosing the suitable level α represents a compromise. There is no unique recommendation for the choice of α and this choice is often related to the particular field of study. Indeed, very often we have an idea what highest first type error we could tolerate. Despite the above comments, the value $\alpha = 0.05$ is often considered as a "default".

Once having decided on the level of significance, we would like to "perform optimally". Of course, after having constructed the optimal test, we would still like to examine its power to see if it is not too low for the purpose of our analysis. If this turns out to be the case, we might need to increase the **sample size** in order to improve the power.

Unfortunately, even this constrained optimization problem turns out sometimes not to have a solution in cases of **discrete** observations. Sometime, in the discrete case, it is not possible to decompose $\mathcal{X}$ in such a way that the first type error is exactly equal to α(we can not "exhaust the level"). To be able to do this, one needs to introduce *randomized*

tests by allowing $\varphi(\mathbf{x})$ to take *any value* in [0,1]. With this extension of φ, the two types of errors discussed above have the following interpretation:

$$P(\text{reject } H_0 \mid H_0 \text{ correct }) = \int .. \int P(\text{ reject } H_0 \mid \mathbf{X} = \mathbf{x}) L(\mathbf{x}, \theta_0) d\mathbf{x} =$$

$$\int .. \int \varphi(\mathbf{x}) L(\mathbf{x}, \theta_0) d\mathbf{x} = E_{\theta_0} \varphi$$

(where $d\mathbf{x}$ is a shorthand notation for $d\mathbf{x} = dx_1 dx_2 .. dx_n$)

$$P(\text{ accept } H_0 \mid H_1 \text{ true }) = \text{ similarly to the argument above} = 1 - E_{\theta_1} \varphi$$

These definitions of the two types of errors can be easily interpreted also in cases of composite hypotheses and will be used from now on.

6.3 Fundamental Lemma of Neyman- Pearson

You must have heard about it from your earlier statistics courses.

Lemma 6.1. *i) For every $\alpha \in (0,1)$ there exists a constant C and a test*

$$\varphi^* = \begin{cases} 1 \text{ if } x \in S = \{x : L(x,\theta_1)/L(x,\theta_0) > C\}, \\ \gamma \text{ if } x \in R = \{x : L(x,\theta_1)/L(x,\theta_0) = C\}, \\ 0 \text{ if } x \in A = \{x : L(x,\theta_1)/L(x,\theta_0) < C\} \end{cases}$$

with $E_{\theta_0}\varphi^* = \alpha$. The constant $\gamma \in (0,1)$ in the definition of the test is equal to $\gamma = \frac{\alpha - P_{\theta_0}(S)}{P_{\theta_0}(R)}$;

ii)φ^* is the best α-test, i.e. $E_{\theta_1}\varphi^*$ is maximal among all tests $\varphi \in \Phi_\alpha = \{\varphi \mid E_{\theta_0}\varphi \leq \alpha\}$.

iii) φ^* is essentially unique, i.e. all other "best"α-tests in the sense of ii) must coincide with φ^* on S and A.

Proof (sketch, details at lecture):

i) Given α, we define C to be the smallest value on the real line for which $P_{\theta_0}\{\frac{L(X,\theta_1)}{L(X,\theta_0)} > C\}$ is still $\leq \alpha$ (in the continuous case we would actually have precisely $P_{\theta_0}\{\frac{L(X,\theta_1)}{L(X,\theta_0)} > C\} = \alpha$ but equality might not be possible in the discrete case). The constant C which we choose in this manner has a specific name-it is called the **upper $\alpha * 100\%$−point** of the distribution of $\frac{L(X,\theta_1)}{L(X,\theta_0)}$ when θ_0 is the true parameter. Then, looking at the definition of φ^* and using the definition of γ we see that

$$E_{\theta_0}\varphi^* = 1 * P_{\theta_0}(\mathbf{X} \in \mathbf{S}) + \gamma * \mathbf{P_{\theta_0}}(\mathbf{X} \in \mathbf{R}) = \cdots = \alpha$$

ii) Take *any other* $\alpha-$test φ and divide the sample space $\mathcal{X}$ into $\mathcal{X} = \mathcal{X}^+ \bigcup \mathcal{X}^- \bigcup \mathcal{X}^=$ with:

$$\mathcal{X}^+ = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) > 0\}$$

$$\mathcal{X}^- = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) < 0\}$$
$$\mathcal{X}^= = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) = 0\}$$

Analyzing the expression $Z(X) = (\varphi^*(X) - \varphi(X))(L(X, \theta_1) - CL(X, \theta_0))$ separately for values of $X \in \mathcal{X}^+, X \in \mathcal{X}^-$ and $X \in \mathcal{X}^=$, we see that always $Z(X) \geq 0$ holds! (Why (!)). But then, of course,

$$\int_{\mathcal{X}} Z(X)dX \geq 0 \tag{12}$$

holds. Substituting back the value of $Z(X)$ in (12) we get: $E_{\theta_1}\varphi^* \geq E_{\theta_1}\varphi$. Since φ was arbitrarily chosen in the set of $\alpha-$tests, this implies that φ^* can not be improved with respect to power, that is, it is the best $\alpha-$ size test.

iii) If $\bar{\varphi}$ is another "best" α test (that is $E_{\theta_1}\varphi^* = E_{\theta_1}\bar{\varphi}$ holds) then according to our discussion in ii), we necessarily need to have $Z(X) \equiv 0$. Since $Z(X)$ is a product of two factors, we either have one of the factors being zero: $\varphi^*(X) = \bar{\varphi}(X)$ or, if not then the other one must be zero (which means $X \in R$). Hence, always when X is not in R but in S or in A, we must have $\varphi^*(X) = \bar{\varphi}(X)$.

6.4 Comments related to the Neyman-Pearson Lemma

Any test φ with $E_{\theta_0}\varphi = \alpha$ is called an *α-test* (equivalently, an $\alpha-$size test. The region S is called a *rejection region (critical region).* The optimal test φ^* has the highest power among all tests of size $\leq \alpha$. Looking at the structure of φ^* given in Part i) of Lemma 6.1 we see that it has a very simple and intuitively appealing interpretation: we look at the ratio of likelihoods for the sample under the alternative and under the null hypothesis. When this ratio is large enough, the alternative is more likely to have generated the sample and we "vote" for the alternative. If the ratio is small enough, the hypothesis is more likely to have generated the sample and we "vote" for the hypothesis. In the "intermediate case" of the ratio being equal to C , we are in doubt and this is why our decision is random (we decide for the alternative with a probability $\gamma \in (0, 1)$). The choice of C and γ is tailored to make the test have exactly a size equal to α, since by exhausting the level given in advance to us we are hoping to maximize the power. All this simple reasoning finds its rigorous support in the Fundamental Lemma given above.

6.5 Simple $\mathbf{H}_0$ versus composite $\mathbf{H}_1$-the "simple case"

We consider now the more realistic situation where we have a *collection* of alternatives instead of a simple one. This is a new situation, not covered by the Neyman-Pearson Lemma. One simple case can be handled immediately. If we obtain the same size α best critical region for all θ-values in the alternative set then the optimal Neyman-Pearson test φ^* that one can construct for *one concrete alternative value* θ_1 will be *uniformly most powerful* (UMP) α-size test for the simple hypothesis versus the *collection* of alternatives.

Example: $\mathbf{X} = (X_1, X_2, .., X_n)$: i.i.d. $N(\theta, 1)$. Consider $H_0 : \theta = \theta_0 \in R^1$ versus $H_1 : \theta > \theta_0$. We want to find the UMP α-test of H_0 versus H_1, i.e. we want to find a test φ^* which is such

that for any other test $\varphi \in \Phi_\alpha : E_\theta\varphi^* \geq E_\theta\varphi$ *for all* $\theta > \theta_0$ holds. Take and fix *any* $\theta_1 > \theta_0$ and consider testing H_0 versus $\bar{H}_1 : \theta = \theta_1$. Then the Neyman-Pearson (NP) lemma can be applied and after simple transformations, we get the rejection region S of the best (NP test) for H_0 versus $\bar{H}_1$ in the form $S = \{\mathbf{x} : \bar{x} \geq \theta_0 + (z_\alpha/\sqrt{n})\}$(which obviously does not depend on the specific θ_1 in the alternative. Then $\varphi^* = \begin{cases} 1 \text{ if } \bar{x} \geq \theta_0 + (z_\alpha/\sqrt{n}), \\ 0 \text{ if } \bar{x} < \theta_0 + (z_\alpha/\sqrt{n}) \end{cases}$ will be the uniformly most powerful α-test of H_0 versus H_1.

6.6 Composite $\mathbf{H}_0$ versus composite $\mathbf{H}_1$

In general, for such type of hypothesis testing problems, there is no UMP α-size test. But for *some types of distributions* and specific hypotheses/alternatives, like intervals on the real line, one can find UMP α-tests:

6.6.1 MLR family of distributions

The family $L(\mathbf{x}, \theta), \theta \in R$ has a *monotone likelihood ratio* (MLR) in the statistic $T(\mathbf{X})$ if for any fixed θ' and θ'' such that $\theta' < \theta''$, it holds that $\frac{L(\mathbf{x},\theta'')}{L(\mathbf{x},\theta')}$ is a *non-decreasing function* of $T(\mathbf{x}) = T(x_1, x_2, .., x_n)$.

Note: typical examples are from the one-parameter exponential family: if $f(x, \theta) = a(\theta)b(x)\exp(c(\theta)d(x))$ and $c(\theta)$ is strictly monotone increasing then

$$\frac{L(\mathbf{x}, \theta'')}{L(\mathbf{x}, \theta')} = \frac{a^n(\theta'')}{a^n(\theta')}.\exp\{[(c(\theta'') - c(\theta')]\sum_{i=1}^{n} d(x_i)\}$$

and clearly, this family has a MLR in $T(\mathbf{X}) = \sum_{i=1}^{n} d(X_i)$.

6.6.2 Theorem of Blackwell & Girshick

Suppose $\mathbf{X} \sim L(\mathbf{x}, \theta)$ and the family is with MLR in $T(\mathbf{X})$. Then for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, the α-test φ^* with the structure:

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } T(\mathbf{x}) > k \\ \gamma \text{ if } T(\mathbf{x}) = k \\ 0 \text{ if } T(\mathbf{x}) < k \end{cases}$$

(k being the upper $\alpha.100\%$ point of the P_{θ_0}-distribution of $T(\mathbf{X})$) has an increasing power function $E_\theta\varphi^*$(i.e. its power as a function of θ is increasing) and the test is UMP α-test.

Note: There is an obvious variant of the theorem: under the same conditions on the family, the test that rejects $H_0 : \theta \geq \theta_0$ in favour of $H_1 : \theta < \theta_0$ when $T(\mathbf{x}) < k$ ($\alpha = P_{\theta_0}(T < k) + \gamma P_{\theta_0}(T = k)$) is the UMP α-test.

6.6.3 Examples

(see lectures):

6.7 Unbiasedness. UMPU α-tests.

6.7.1 General discussion and definition.

We have already seen that for a one parameter exponential family, for example, a UMP α-test for composite alternatives exists. If no UMP test exists, we should think about another criterion to make the optimal choice among possible tests (i.e. a further *restriction* of the set of α-tests is necessary in order to find an optimal solution in a *smaller* set of competing tests). For example, one can:

-choose a 'typical' alternative in the set of alternatives and use the most powerful test for that alternative;

-maximize the power locally, by considering only the θ-values from the alternative set that are close to the hypothetical θ-values. This leads to the notion of a *locally most powerful test* (not to be discussed in this course);

-maximize some weighted average of power for the different alternatives.

One more solution is mathematically very attractive and leads to very reasonable tests. This is to restrict ourselves to the set of *unbiased tests.*

Definition 1. A test φ of $H_0 : \theta \in \Theta_0 (\Theta_0 \subset \Theta)$ versus $H_1 : \theta \in \Theta \backslash \Theta_0$ is an *unbiased* size α-test if $E_\theta \varphi \leq \alpha$ for all $\theta \in \Theta_0$ and $E_\theta \varphi \geq \alpha$ for all $\theta \in \Theta \backslash \Theta_0$.

Basically, the above definition of unbiasedness ensures that there exist no alternatives for which acceptance of the hypothesis is more probable than in cases when the null hypothesis is true. This is a very reasonable requirement (don$'t$ you think) so that asking in addition for it to be satisfied would not restrict too seriously the set of tests of interest (the majority of reasonable tests of size α would still be allowed to compete).

6.7.2 Basic results

Theorem 6.2. *Suppose* $\mathbf{X} \sim L(\mathbf{x}, \theta)$ *with* $L(\mathbf{x}, \theta) = (a(\theta))^n \prod_{i=1}^n b(x_i) \exp[c(\theta). \sum_{i=1}^n d(x_i)]$ *and* $T(\mathbf{X}) = \sum_{i=1}^n d(x_i)$. *Then, for testing* $H_0 : \theta_1 \leq \theta \leq \theta_2$ *versus* $H_1 : \theta < \theta_1$ *or* $\theta > \theta_2$, *the test* φ^* *is* $:\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } T(\mathbf{x}) \notin [c_1, c_2] \\ \gamma_i \text{ if } T(\mathbf{x}) = c_i, i = 1, 2 \\ 0 \text{ if } c_1 < T < c_2 \end{cases}$ *where* $c_1, \gamma_1, c_2, \gamma_2$ *are determined by the conditions* $E_{\theta_1} \varphi^* = E_{\theta_2} \varphi^* = \alpha$.

Moreover,the power function has a minimum somewhere within (θ_1, θ_2) *and is monotone outside* (θ_1, θ_2).

Example: negative exponential (to be considered at lecture)

Theorem 6.3. *Consider the same family like in the previous Theorem 6.2. Then , for testing* $H_0 : \theta = \theta_0$ *versus* $H_2 : \theta \neq \theta_0$ *an UMPU* α*-test exists with the structure:*

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } T(\mathbf{x}) < c_1 \text{ or } T(\mathbf{x}) > c_2 \\ \gamma_i \text{ if } T(\mathbf{x}) = c_i, i = 1, 2 \\ 0 \text{ if } c_1 < T < c_2 \end{cases}$$

. The constants c_i, γ_i *satisfy:* $Power(\theta_0) = \alpha = E_{\theta_0}\varphi^*$ *and* $\frac{\partial}{\partial\theta} Power(\theta_0) = 0 = \frac{\partial}{\partial\theta} E_\theta \varphi^* \mid_{\theta=\theta_0}$ *.*

Note: The latter Theorem 6.3 can be justified as a limiting case of Theorem 6.2 when the interval $[\theta_1, \theta_2]$ collapses to a single point θ_0.

6.8 Examples

a) Assume, a "sample" of one observation ($n = 1$) from an exponential family with density $f(x, \theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta}), x > 0$ is available. The parameter $\theta > 0$ is to be tested. One would like to test $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$.

According to Theorem 6.3, the UMPU α-test φ^* has the structure:

$\varphi^*(x) = \begin{cases} 1 \text{ if } T < C_1 \text{ or } T > C_2 \\ 0 \text{ if } C_1 \leq T \leq C_2 \end{cases}$ where $T(x) = x$ in this case (one-parameter exponential family, $d(x) = x, n = 1$). We only need to find C_1 and C_2 in order to uniquely specify the above test. Note that $E_\theta\varphi^* = P_\theta(x \notin (C_1, C_2)) = 1 - \exp(-C_1/\theta) + \exp(-C_2/\theta)$(since the cdf is $F(x, \theta) = 1 - \exp(-x/\theta), x > 0$). The two conditions on $E_\theta\varphi^*$ are:

- $E_\theta\varphi^* \mid_{\theta=1} = \alpha = 1 - \exp(-C_1) + \exp(-C_2)$
- $\frac{\partial}{\partial\theta} E_\theta\varphi^* \mid_{\theta=1} = -\frac{C_1}{\theta^2}\exp(-\frac{C_1}{\theta}) + \frac{C_2}{\theta^2}\exp(-\frac{C_2}{\theta}) \mid_{\theta=1} = -C_1\exp(-C_1) + C_2\exp(-C_2) = 0$

We get a system of two equations with respect to C_1 and C_2. It can be solved numerically (iteratively) given the level α and hence the UMPU α-test will be completely specified.

b) If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\theta, 1)$ then for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ there exists an UMPU α-test (according to Theorem 6.3). We want to show that it coincides with the well-known test $\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } \sqrt{n}|\bar{\mathbf{x}} - \theta_0| \geq z_{\alpha/2} \\ 0 \text{ if } \sqrt{n}|\bar{\mathbf{x}} - \theta_0| < z_{\alpha/2} \end{cases}$

$E_\theta\varphi^* = P_\theta(\bar{\mathbf{x}} \leq C_1 \text{ or } \bar{\mathbf{x}} \geq C_2) = P_\theta\{\sqrt{n}(\bar{\mathbf{x}}-\theta) \leq \sqrt{n}(C_1-\theta) \text{ or } \sqrt{n}(\bar{\mathbf{x}}-\theta) \geq \sqrt{n}(C_2-\theta)\} =$

$P\{N(0,1) \leq \sqrt{n}(C_1-\theta)\} + P\{N(0,1) \geq \sqrt{n}(C_2-\theta)\} = \Phi(\sqrt{n}(C_1-\theta)) + 1 - \Phi(\sqrt{n}(C_2-\theta))$.

Here Φ is the cdf of the standard normal distribution. The two equations that have to be satisfied, are:

$$E_{\theta_0}\varphi^* = \alpha$$

and

$$\frac{\partial}{\partial\theta}E_\theta\varphi^*|_{\theta=\theta_0} = 0.$$

This leads us to the following two equations:

$$\Phi(\sqrt{n}(C_1-\theta_0)) + 1 - \Phi(\sqrt{n}(C_2-\theta_0) = \alpha$$
$$\Phi'(\sqrt{n}(C_1-\theta_0)) = \Phi'(\sqrt{n}(C_2-\theta_0))$$

From the second equation we get (since $C_1 \neq C_2$ and the standard normal density is symmetric around zero) : $C_1 + C_2 = 2\theta_0$. Substituting into the first equation, we get: $2[1-\Phi(\sqrt{n}(C_2-\theta_0))] = \alpha$. The latter relation means that $C_2 = \theta_0 + \frac{z_{\alpha/2}}{\sqrt{n}}$ and $C_1 = 2\theta_0 - C_2 = \theta_0 - \frac{z_{\alpha/2}}{\sqrt{n}}$

Hence the form of φ^* is indeed $\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } \sqrt{n}|\bar{\mathbf{x}}-\theta_0| \geq z_{\alpha/2} \\ 0 \text{ if } \sqrt{n}|\bar{\mathbf{x}}-\theta_0| < z_{\alpha/2} \end{cases}$

6.9 Locally most powerful tests

Another way to handle the situation in which no UMP test exists is to restrict attention to values of the alternative parameter, i.e. we look at tests which have high power at some *particular alternatives*. In most cases, we consider the behaviour of the power for alternative parameter values that are close to the null hypothesis. One reckons that deriving a test that works well in *such "difficult" situations* when the hypothesis and alternative are close to each other, is most essential for the applications (when hypothesis and alternative are relatively far from each other, hopefully many tests would do a good job).

Definition 2 (locally most powerful test).

Test φ^* with power function $E_\theta\varphi^*$ is (at its size) locally most powerful (LMP) for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ if for any other test φ with $E_{\theta_0}\varphi = E_{\theta_0}\varphi^*$, there exists $\Delta > 0$ such that $E_\theta\varphi^* \geq E_\theta\varphi$ for every $\theta \in (\theta_0, \theta_0+\Delta)$.

Note: In most practical situations, the tests we consider have differentiable power functions. In such cases, a LMP test will obviously maximize:

$$\frac{\partial}{\partial\theta}E_\theta\varphi_{|\theta=\theta_0} = \int..\int \varphi(\mathbf{x})\frac{\partial}{\partial\theta}L(\mathbf{x},\theta)_{|\theta=\theta_0}d\mathbf{x} \text{ under the constraint } \int \varphi(\mathbf{x})L(\mathbf{x},\theta_0)d\mathbf{x} = \alpha = E_{\theta_0}\varphi.$$

But note that the structure of this optimization problem is the same as in the NP Lemma. So, proceeding along the same lines (please, reread the proof of the NP Lemma) we arrive at following optimal solution:

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } (\partial/\partial\theta)L(\mathbf{x},\theta)_{|\theta=\theta_0} > kL(\mathbf{x},\theta_0) \\ 0 \text{ if } (\partial/\partial\theta)L(\mathbf{x},\theta)_{|\theta=\theta_0} \leq kL(\mathbf{x},\theta_0) \end{cases}$$

and the above optimal test is unique. If we denote, as usual, $V(\mathbf{x},\theta)$ to be the score function then obviously φ^* has the following simple and appealing structure:

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } V(\mathbf{x}, \theta_0) > k \\ 0 \text{ if } V(\mathbf{x}, \theta_0) \leq k \end{cases}$$

Example. $\mathbf{X}$=$(X_1, X_2, .., X_n)$ be i.i.d. Cauchy$(\theta, 1)$ variables with density $f(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}, x \in R, \theta$ being an unknown parameter. Test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ (see discussion at lecture).

6.10 Likelihood ratio tests

We have now looked at a variety of *ad hoc* criteria for optimal tests, including unbiasedness, locally most powerful etc. Other criteria, not discussed here, include similarity, invariance etc. The reason to consider so many different criteria comes from the lack of universal criterion for comparing sets of models. But if we agree with the strong likelihood principle and believe the set of models to be best represented by its most likely member given the observed data, we can arrive at a relatively simple and universal procedure also in a testing context.

When no points in the parameter space specified by H_0 are preferred to others, the likelihood function can be maximized under the null and alternative hypotheses:

6.10.1 General formulation

Assume, for example, that $\Theta \subseteq R^{r+s} = R^k; \Theta_0 = \{\theta \in \Theta \mid \theta_1 = \theta_{10}, .., \theta_r = \theta_{r0}\}, H_0 : \theta \in \Theta_0$and $H_1 : \theta \in \Theta \backslash \Theta_0$. Let us define the statistic

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L(\mathbf{X}, \theta)}{\sup_{\theta \in \Theta} L(\mathbf{X}, \theta)}$$

which is obviously in the interval [0,1]. Intuitively, it makes sense to define the rejection region as $S = \{\mathbf{x} \mid \lambda(\mathbf{x}) \leq C\}$ for a certain constant C. However, the optimum properties of likelihood ratios for simple hypotheses, as discussed in the NP lemma, no longer apply, except asymptotically. In addition, the exact distribution of $\lambda(\mathbf{X})$(that is needed to determine the constant C) is also difficult to obtain without using asymptotic approximations. Thus, the fact that the *deviance statistic* has an asymptotic distribution which is well known is generally used to obtain significance levels or to get the constant C for α given in advance.

Bearing in mind the derivations related to the deviance in Lecture 5, we can formulate the following, slightly more general, statement:

Theorem 6.4. *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *be a random sample from* $f(x, \theta), \theta \in R^{r+s}$. *Suppose the regularity conditions for consistency and asymptotic normality of MLE under* H_0 *and* H_1 *hold. Then under* $H_0 : -2 \ln \lambda(\mathbf{X}) \to^d \chi^2_r$ *(*r *is interpreted as the difference between the number of free parameters specified by* $\theta \in \Theta$ *and the number of free parameters specified by* $\theta \in \Theta_0$*).*

Examples:

i) For testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for a sample of n i.i.d. $N(\mu, \sigma^2), (\sigma^2$ known) one has $-2\ln\lambda(\mathbf{x}) = \frac{n(\bar{\mathbf{x}}-\mu_0)^2}{\sigma^2} = D(\mu_0) \sim \chi^2_1$ (and the result is *exact).*

ii) Normal sample with both μ and σ^2 unknown. Testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The MLR test is equivalent to the classical t-test with a rejection region

$$S = \{\mathbf{x}||\frac{(\bar{\mathbf{x}} - \mu_0)\sqrt{n}}{s} |\geq k\}$$

Here $s = (s^2)^{1/2} = [\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2]^{1/2}$ and to make the size equal to α, we must choose $k = t_{\alpha/2,(n-1)}$(the upper $(\alpha/2).100\%$ point of the t distribution with $(n-1)$ degrees of freedom.)

Thus, a popular test (known to be also the UMPU α-test for the above problem) could be constructed using the Likelihood Ratio Test construction.

We shall only note here that the asymptotic distribution result in Theorem 6.4 is formulated under the hypothesis only but more sophisticated reasoning can be used to derive the asymptotic distribution under alternative parameter values, too. This distribution is a non-central χ^2 and can be used for (approximate asymptotic) power computations. We omit the details.

6.11 Alternatives to the GLRT.

The GLRT is widely used but there are circumstances where other test procedures may be preferred. We define two of these test procedures for the case $s = 0$, that is, $k = r$.

6.11.1 Score test.

It uses $S = V(\mathbf{X}, \theta_0)' I^{-1}_{\mathbf{X}}(\theta_0) V(\mathbf{X}, \theta_0)$ instead of $-2\log\lambda(\mathbf{X})$.

6.11.2 Wald test.

It uses $(\hat{\theta} - \theta_0)' I_{\mathbf{X}}(\hat{\theta})(\hat{\theta} - \theta_0)$ instead of $-2\log\lambda(\mathbf{X})$ where θ_0 is the hypothetical vector and $\hat{\theta}$ is the MLE.

For both the Score test, and the Wald test, the *asymptotic* distribution of the test statistic under H_0 is the *same* as the distribution of the GLRT statistic (that is, chi-square with $r = k$ degrees of freedom). Score tests have a numerical advantage in comparison to GLRT and the Wald test, that they do *not* require the MLE to be calculated! Specifically in the econometrics literature, the Score test is known as **Lagrange Multiplier Test**. The name comes from its alternative derivation in which the Likelihood function is maximized subject to the restrictions of H_0 and the maximization method uses Lagrange multipliers.

Much research has been devoted to selecting one of the three tests as a preferred test in a particular situation for relatively small sample sizes. We will only say that this is a difficult task and will not be discussed in our course.

Exercise. Let X be the number of successes in a binomial experiment with a probability of success p. We wish to test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Denote $\hat{p} = \frac{X}{n}$. Then the Score statistic is

$$S = \frac{(X - np_0)^2}{np_0(1 - p_0)},$$

the Wald statistic is

$$W = \frac{(X - np_0)^2}{n\hat{p}(1 - \hat{p})}$$

and

$$-2 \ln \lambda = 2\{X \ln \frac{\hat{p}(1 - p_0)}{p_0(1 - \hat{p})} + n \ln \frac{1 - \hat{p}}{1 - p_0}\}.$$

7 Lecture 7: Order Statistics

7.1 Motivation

Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ denote a random sample from a population with a continuous distribution function F_X. Since F_X is assumed to be *continuous*, the probability of any two of these random variables assuming the same value is zero. After reordering the n values we get $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)})$ in which, as mentioned, the $\leq$ sign could also be replaced by $<$. These values are collectively termed the *order statistic* of the random sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$. The subject of order statistics generally deals with properties of $X_{(r)}$ $(r = 1, 2, \ldots, n)$ which is called the r-th order statistic.

Order statistics are particularly useful in nonparametric statistics because of the following:

Theorem 7.1. (Probability-integral transformation). *If the random variable X has a continuous cdf F_X then the random variable $Y = F_X(X)$ has the uniform probability distribution over the interval (0,1). Further, given a sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ of n i.i.d. random variables with cdf F_X, the transformation $U_{(r)} = F_X(X_{(r)})$ produces a random variable $U_{(r)}$ which is the r-th order statistic from the uniform population in (0,1), regardless of what F_X is, i.e. $U_{(r)}$ is distribution-free.*

Proof: It is your textbook and was also discussed in the introductory Lecture 1.

Note: The above theorem has also an extremely important practical application in the generation **(computer simulation)** of observations from any specific continuous distribution function. There are several well-developed **uniform random number generators** that implement methods to generate sequences of uniform in (0,1) pseudo-random numbers. These numbers are **pseudo** since in fact they are generated by a deterministic algorithm (therefore are not random) **but** look as **random** (hence the word pseudo-random) in the sense that they pass usual statistical tests about randomness of the generated sequence. Every program system (Fortran, SPLUS, C, SAS, etc.) has such uniform random number generators and we will not discuss their specific implementation here. What we would like to discuss is how we could use these **uniform** random number generators to generate random numbers with **arbitrary continuous** cumulative distribution function F_X. The answer is:

1) Generate Y as uniformly distributed in (0,1) using the uniform random number generator

2) Calculate $\xi = F_X^{-1}(Y)$.

Then ξ is distributed according to $F_X(.)$ since its cumulative distribution function is:

$$P(F_X^{-1}(Y) < x) = P(F_X^{-1}(F_X(X)) < x) = P(X < x) = F_X(x).$$

Some further important obvious applications of order statistics are listed below:

- $X_{(n)}$ is of interest in studying floods, earthquakes and other extreme phenomena, sports records, financial markets etc.

- $X_{(1)}$ is useful, for example, in estimating strength of a chain that would depend on the weakest link.
- the sample median defined as $X_{[(n+1)/2]}$ for n odd and any number between $X_{(n/2)}$ and $X_{(n/2+1)}$ for n even, is a measure of location and an estimate of the population central tendency.
- the sample midrange $(X_{(n)} + X_{(1)})/2$ is also a measure of central tendency, whereas the sample range $X_{(n)} - X_{(1)}$ is a measure of dispersion.

7.2 Multinomial distribution.

At this point, we shall give the definition of the multinomial distribution. It is used to prove in an easy way many of the results related to distributions of order statistics and for that reason, it will be given first:
Suppose a single trial can result in k ($k \geq 2$) possible outcomes numbered $1, 2, \ldots, k$ and let $w_i = P(\text{a single trial results in outcome } i)$ ($\sum_{i=1}^{k} w_i = 1$). For n independent trials, let X_i denote the number of trials resulting in outcome i (then $\sum_{i=1}^{k} X_i = n$). Then we say that the distribution of $(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n; w_1, w_2, \ldots, w_k)$ and it holds

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} w_1^{x_1} w_2^{x_2} \ldots w_k^{x_k}, 0 < w_i < 1, \sum_{i=1}^{k} w_i = 1$$

Note that when $k = 2$ this is just the familiar Binomial distribution, i.e. the Multinomial can be considered as its generalization. It is also easy to see that

$$E(X_i) = nw_i; Var(X_i) = nw_i(1 - w_i), i = 1, 2, \ldots, k$$

holds by noting that the **marginal** distribution of the i-th component is in fact binomial ($\text{Bin}(n; w_i)$). In a bit more complicated way (for example, using moment generating functions), one can also show that $Cov(X_i, X_j) = -nw_i w_j$ holds for $j \neq i$.

7.3 Distributions related to order statistics

A lot of material discussed in this lecture can be found in Section 5.4 of the textbook.

Let X be a random variable with a density $f_X(x)$ and a cumulative distribution function $F_X(x)$ and let there be n independent copies $X_1, X_2, \ldots, X_n$ of X.

Theorem 7.2. *The joint density* $f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ *is given by:*

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = n! \prod_{i=1}^{n} f_X(x_{(i)}) \text{ for } x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

Proof: We consider a trial consisting of $k = 2n+1$ possible outcomes that is repeated independently n times. Each of the outcomes is a realization in one of the $2n+1$ intervals:

$$(-\infty, x_{(1)}), [x_{(1)}, x_{(1)} + \Delta x_{(1)}), [x_{(1)} + \Delta x_{(1)}, x_{(2)}), \ldots, [x_{(n-1)} + \Delta x_{(n-1)}, x_{(n)}),$$

$$[x_{(n)}, x_{(n)} + \Delta x_{(n)}), [x_{(n)} + \Delta x_{(n)}, \infty)$$

where $\Delta x_{(i)}, i = 1, 2, \ldots, n$ are chosen sufficiently small so that no overlap of the intervals occurs. The trial's outcome can be interpreted as a realization of multinomial

$$\text{Multin}(n; F_X(x_{(1)}), F_X(x_{(1)}+\Delta x_{(1)})-F_X(x_{(1)}), F_X(x_{(2)})-F_X(x_{(1)}+\Delta x_{(1)}), \ldots, (1-F_X(x_{(n)}+\Delta x_{(n)}))$$

distribution. We are looking at the probability for one very particular outcome of the trial, namely realization in the second, fourth, $2n$th interval for this multinomial distribution. On one hand, this probability is

$$\frac{n!}{0!1!0!1!\ldots 0!1!0!}[F_X(x_{(1)} + \Delta x_{(1)}) - F_X(x_{(1)}] \ldots [F_X(x_{(n)} + \Delta x_{(n)}) - F_X(x_{(n)}].$$

On the other hand, it is just $P(x_{(i)} \leq X_{(i)} < x_{(i)} + \Delta x_{(i)}, i = 1, 2, \ldots, n)$. Having in mind the definition

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = \lim_{\Delta x_{(i)} \to 0} \frac{P(x_{(i)} \leq X_{(i)} < x_{(i)} + \Delta x_{(i)}, i = 1, 2, \ldots, n)}{\prod_{i=1}^{n} \Delta x_{(i)}}$$

we get

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = n! \prod_{i=1}^{n} f_X(x_{(i)}) \text{ for } x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

Theorem 7.3. *It holds:*

$$f_{X_{(n)}}(y_n) = n[F_X(y_n)]^{n-1} f_X(y_n)$$

$$f_{X_{(1)}}(y_1) = n[1 - F_X(y_1)]^{n-1} f_X(y_1)$$

$$f_{X_{(r)}}(y_r) = \frac{n!}{(r-1)!(n-r)!}[F_X(y_r)]^{r-1}[1 - F_X(y_r)]^{n-r} f_X(y_r)$$

Proof: Since the previous theorem gives us the joint distribution, the marginal distributions formulated in Theorem (7.2) can be obtained through integration. This method is a straightforward as an idea but the integration is tiresome. A much simpler method can be used which appeals to probability theory instead to pure mathematical integration. It will be illustrated at the lecture.

Note: From Theorem (7.3) we realize that for the particular case of the F_X being uniform distribution on (0,1), we get

$$f_{X_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} y^{r-1}(1-y)^{n-r}, y \in (0,1)(\text{ and zero elsewhere})$$

which is the density of the *Beta distribution* with parameters r and $n-r+1$. In particular, using the properties of the beta distribution, we can now show that for the rth order statistic $X_{(r)}$ of the beta distribution we have $EX_{(r)} = \frac{r}{n+1}, Var(X_{(r)}) = \frac{r(n+1-r)}{(n+1)^2(n+2)}$.

Joint densities of couples $(X_{(i)}, X_{(j)})$ can also be derived through integration of the joint density of all the n order statistics. But it is again easier to use some probabilistic arguments instead. The idea is illustrated in the following result:

Theorem 7.4. *It holds for* $1 \leq i < j \leq n$*:*

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

for $-\infty < u < v < \infty$*.*

Proof (idea): We obtain the cdf $F_{X_{(i)},X_{(j)}}(u,v)$ first and then find its partial derivative $f_{X_{(i)},X_{(j)}}(u,v) = \frac{\partial^2}{\partial u \partial v} F_{X_{(i)},X_{(j)}}(u,v)$. To obtain $F_{X_{(i)},X_{(j)}}(u,v)$, let U be a random variable that counts the number of $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ that are less or equal to u and V be the random variable that counts the number of $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ greater than u and less or equal to v. Then

$$(U, V, n - U - V) \sim \text{Multinomial}(n; F_X(u), F_X(v) - F_X(u), 1 - F_X(v)).$$

Then, obviously

$$F_{X_{(i)},X_{(j)}}(u,v) = P(U \geq i, U + V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U = k, V = m) + P(U \geq j) =$$

$$\sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \frac{n!}{k!m!(n-k-m)!} [F_X(u)]^k [F_X(v) - F_X(u)]^m [1 - F_X((v)]^{n-k-m} + P(U \geq j)$$

Looking at this expression, we realize that when calculating $\frac{\partial^2}{\partial u \partial v} F_{X_{(i)},X_{(j)}}(u,v)$, the probability $P(U \geq j)$ is irrelevant since this term only depends on u and *not* on v. Carefully calculating the mixed partial derivative of the other term, yields the formula given in the Theorem.

Of course, joint densities of three or more order statistics could also be derived using similar arguments like above but the calculations will be more exhausting. We could use mathStatica's help to alleviate the calculations.

Note: The above derivations have very important applications. For instance, the density of the sample median, or of the sample range can now easily be derived by using standard formulae for density of transformed random variables. In the case of the range $R = X_{(n)} - X_{(1)}$, for example, we know from Theorem (7.4) that

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)[F_X(y) - F_X(x)]^{n-2} f_X(x) f_X(y)$$

and if we make the transformation $u = y - x, v = y$ (with absolute value of Jacobian equal to one), we obtain for the range: for all $u > 0$:

$$f_R(u) = \int_{-\infty}^{\infty} n(n-1)[F_X(v) - F_X(v-u)]^{n-2} f_X(v-u) f_X(v) dv$$

For the uniform distribution, the integrand is non-zero for the intersection of the regions $0 < v - u < 1$ and $0 < v < 1$ which is just the region $0 < u < v < 1$. Therefore after integration, we get in this case $f_R(u) = n(n-1)u^{n-2}(1-u)$ for $0 < u < 1$.
The "trick" in the above derivation was to introduce, on top of the transformation $u = y - x$ of interest, one more transformation ($v = y$) so that we could tranform the original

vector $(X_{(1)}, X_{(n)})$ in a new random vector (U, V) whose first component is the statistic of interest. This allows us, by using the density transformation formula for random vectors, to get the **joint** density of (U, V) first. Then we integrated out the unwanted variable V to obtain the marginal density of the main component of interest (the component U). Such a trick is used very often when working with order statistics. Clearly, the variable V plays an intermediate role here and in its choice, we are mainly guided by convenience of the calculations. For example, we could have chosen $v = x$ instead as a component of our tranformation. You are advised to go along similar lines by using this new transform instead, derive the marginal of U again and convince yourself that you get the same result for the density of the range.

8 Lecture 8: Higher order asymptotics

8.1 Motivation

General results in estimation and in other inference procedures can usually be defended on asymptotic grounds only (i.e., when the sample size n tends to infinity). An inference procedure that is optimal for a finite sample size is very rarely possible to construct and it depends heavily on the specific distributional assumptions about the sample that has given rise to the data. Hence such a procedure is too much individually tailored and can not be offered as a general methodological tool to be used in situations where the distributional assumptions have been changed. General inference procedures could be offered on asymptotic grounds only and we have seen how useful they can be when discussing the MLE, for example.

Most useful in statistical practice are the so-called **first-order** asymptotic results. They state that under some regularity conditions asymptotic normality of certain estimator holds (or that under regularity conditions, the null distribution of a certain test-statistic is asymptotically normal or chi-squared etc.) We have already discussed many such results in Lectures 5 and 6.

The techniques used to show such results are usually a combination of central limit theorem (CLT) and Taylor expansions. Again, in our Lecture 5 about properties of MLE we have seen these techniques demonstrated at work. The final product of such results is the statement about asymptotic normality of a suitably normalized statistic. However, sometimes, the sample sizes used in practice are not large enough to warrant that the accuracy achieved by the asymptotic normal approximation is precise enough. Then it is worth trying to include **higher order expansions** for the distribution of the statistic of interest hoping that these more complex expressions will bring about a better approximation for the distribution when the sample size is not as large.

To take an example to illustrate our point, assume that we are dealing with the distribution of the sample mean $\bar{X}$ taken from a (not necessarily normal) population with finite mean μ and variance σ^2. Let $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ and $F_n(z) = P(Z_n \leq z)$. Then the CLT says that $F_n(z) \to_{n\to\infty} \Phi(z)$ for every $z \in R$. If we also assume that a third finite moment γ exists then the famous Berry-Esseen theorem states that

$$|F_n(z) - \Phi(z)| = O(\frac{1}{\sqrt{n}})$$

uniformly in z where the bound on the right hand side depends on the absolute third moment. More precise statement can be obtained in the form

$$|F_n(z) - \Phi(z) - \frac{C_1(F)p_1(z)\phi(z)}{\sqrt{n}}| = O(\frac{1}{n}) = o(\frac{1}{\sqrt{n}})$$

uniformly in z where $\phi(z)$ is the density of the standard normal, C_1 is a suitably chosen constant and $p_1(z)$ is certain first degree polynomial. Expansions in the form

$$F_n(z) = \Phi(z) + \sum_{s=1}^{k} \frac{q_s(z)}{n^{s/2}} + o(n^{-k/2})$$

can often be obtained for the distribution $F_n(.)$ of suitably normalized statistics of interest and are called **Edgeworth expansions.** Here $q_s(z)$ are some polynomials multiplied by the standard normal density. The coefficients of the polynomials depend on the cumulants of $F(z)$ and the cumulants are related to the moments of the distribution $F(.)$. In such a way, going beyond the normal expansion given by the CLT, the effects of skewness, kurtosis and of higher order moments on the approximation of the statistic Z_n can be captured.

8.2 Moments and cumulants

As is known, the moment generating function (**MGF**) of a random variable X is defined as $M_X(t) = E[\exp(tX)]$ whenever the expectation is finite. It is obvious that always $M_X(0) = 1$ holds. The MGF may only be defined in a small neigbourhood of 0 (or even only for $t = 0$). If it exists in an open interval around 0 then all moments of the random variable exist and can be obtained via $E(X^r) = M_X^{(r)}(t)_{|t=0} = \mu'_r$. We then have the Taylor expansion

$$M_X(t) = 1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \cdots + \mu'_r \frac{t^r}{r!} + O(t^{r+1})$$

when $t \to 0$. In addition, we define the cumulant generating function (**CGF**) of the random variable X:

$$K_X(t) = \log[M_X(t)],$$

(defined on the same interval as the MGF). The Taylor series expansion

$$K_X(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots + \kappa_r \frac{t^r}{r!} + O(t^{r+1})$$

when $t \to 0$ holds in this interval and defines the **cumulants** $k_s, s = 1, 2, \ldots$ of the distribution of X. Clearly $\kappa_r = K_X^{(r)}(t)_{|t=0}$ holds.

Moreover, by equating the coefficients in the expansions of $\exp[K_X(t)]$ and of $M_X(t)$, the relationships between moments and cumulants can be established. In particular, the relationships $\mu'_1 = \kappa_1 = E(X), \kappa_2 = Var(X) = \mu'_2 - (\mu'_1)^2$ and $\kappa_3 = 2\mu'^3_1 - 3\mu'_1\mu'_2 + \mu'_3$ hold. The `mathStatica` functions `CumulantToRaw` and `RawToCumulant` can be used to express analytically the conversions between moments and cumulants and vice versa (consult Section 2.4G of the book).

Per **definition** the third cumulant is the skewness and the fourth cumulant is the curtosis of the distribution of X. To summarise:

- the first cumulant is the first moment
- the second cumulant is the variance
- the third cumulant is the skewness
- the fourth cumulant is the kurtosis

of the distribution.

The reason we introduce the cumulants in our discussion is that although the moments are a more familiar concept from your introductory statistics lectures, the higher order asymptotic expansions we are about to discuss in this lecture, can in fact be presented in more compact form by using the cumulants.

Exercise 1. Show that if $X_1, X_2, \ldots, X_n$ are i.i.d. then $K_{\sum_{i=1}^n X_i}(t) = nK_{X_1}(t)$ holds. Also, for any constants a, b we have $K_{aX_1+b}(t) = bt + K_{X_1}(at)$.

8.3 Asymptotic expansions

We should explicitly stress here that the Edgeworth type expansions are representative of what we call *asymptotic expansions* in statistics. They arise in the following general way. We want to represent a set of functions $f_n(z)$ (indexed by $n = 1, 2, \ldots$) in the form

$$f_n(z) = \gamma_0(z)b_{0,n} + \gamma_1(z)b_{1,n} + \cdots + \gamma_k(z)b_{k,n} + o(b_{k,n}) \tag{13}$$

The expansion is considered as $n \to \infty$. Typical choices of the doubly indexed coefficients $b_{k,n}$ are $\{1, n^{-1/2}, n^{-1}, \ldots, n^{-k/2}\}$ or $\{1, n^{-1}, n^{-2}, \ldots, n^{-k}\}$. The idea is to have contributions progressively falling in order with sample size and essentially requesting for this effect the condition $b_{r+1,n} = o(b_{r,n})$ as $n \to \infty$ for each $r = 0, 1, 2, \ldots, k-1$. In particular examples of asymptotic expansions, the function $f_n(z)$ may be a density or a CDF of a statistic calculated for a sample size of n at a point z on the real axis. The $\gamma_0(z)$ would be the density of the CDF of the standard normal depending on the context.

It is important to notice that the above expansion **is not necessarily a convergent series** for $f_n(z)$ for any fixed z! This means that if n is fixed, just increasing the number of terms k does not necessarily lead to convergence and to improved approximation of $f_n(z)$. One has to increase the sample size, too.

8.3.1 Edgeworth expansion for cdf

A major condition to deriving Edgeworth expansions of the form (13) is the **Cramér condition**. A cdf $F(.)$ on the real line is said to satisfy Cramér's condition if for the characteristic function $\chi_F(t) = E_F(e^{itX})$

$$\limsup_{t\to\infty} |E_F(e^{itX})| < 1 \tag{14}$$

holds. This condition is needed to deal with some more singular cases. It can be shown that *all* continuous distributions do satisfy Cramér's condition. On the other hand, the lattice distributions (whose whole mass is concentrated on a lattice in the form

$$a + bn, n \text{ integer}; a, b \text{ constants})$$

do *not*. An easy intuitive explanation for this phenomenon is that for lattice distributions, the cdf of $F_n(.)$ would have jump discontinuities for any fixed sample size n whereas the Edgeworth expansion is a smooth function and accuracy of the type shown in (13) can not be expected to hold. To give more explicit expression for the terms involved in the Edgeworth expansion, we will define it here for the particular case of the CDF (not the density) and $k = 2$.

Theorem 8.1. *Suppose that $F(.)$ satisfies Cramér's condition and $E_F(X^4) < \infty$ holds. Then*

$$F_n(z) = \Phi(z) - \frac{C_1(F)p_1(z)\phi(z)}{\sqrt{n}} - \frac{C_2(F)p_2(z) + C_3(F)p_3(z)}{n}\phi(z) + o(n^{-1}). \tag{15}$$

Here $C_1(F) = \frac{E(X-\mu)^3}{6\sigma^3}$ (skewness correction), $C_2(F) = \frac{\frac{E(X-\mu)^4}{\sigma^4} - 3}{24}$ (kurtosis correction), $C_3(F) = \frac{C_1^2(F)}{2}$, $p_1(z) = z^2 - 1$, $p_2(z) = z^3 - 3z$, $p_3(z) = z^5 + 15z - 10z^3$. The polynomials p_1, p_2 and p_3 are the second, third and fifth Hermite polynomials $H_j, j = 2, 3, 5$. Their general definition is given by $H_j(z) = \frac{(-1)^j \phi^{(j)}(z)}{\phi(z)}, j = 0, 1, 2, \ldots$ where $\phi^{(j)}(z)$ is the j-th order derivative of the standard normal density.

About the proof: The proof is subtle and here we only sketch the idea behind it. The main ingredient in proving such result is *Esseen's smoothing lemma*. This lemma bounds the closeness of a cdf and of its approximation by closeness of their Fourier transforms. For a cdf F and a differentiable function G with G' uniformly bounded and such that $\int |F(t) - G(t)|dt < \infty$, Esseen's lemma tells us that for each $T > 0$

$$\sup_z |F(z) - G(z)| \leq \frac{1}{\pi}\int_{-T}^{T} |\frac{\chi_F(t) - \chi_G(t)}{t}|dt + \frac{A}{T}\sup_t |G'(t)| \tag{16}$$

where A is an absolute constant and $\chi_F(t) = \int_{-\infty}^{\infty} e^{itx}dF(x), \chi_G(t) = \int_{-\infty}^{\infty} e^{itx}dG(x)$.

If we set

$$G(z) = \Phi(z) - \frac{C_1(F)p_1(z)\phi(z)}{\sqrt{n}} - \frac{C_2(F)p_2(z) + C_3(F)p_3(z)}{n}\phi(z)$$

and $F = F_n$ to be the CDF of Z_n then it is possible to control from above the difference $\chi_F(t) - \chi_G(t)$ on the right hand side of (16) thus getting control from above of the left hand side.

The control on the difference $\chi_F(t) - \chi_G(t)$ is possible to achieve for the following reason. Since $EZ_n = 0$ and $Var(Z_n) = 1$, one can show that starting with the expansion

$$E(e^{itZ_n}) = 1 + remainder \tag{17}$$

with remainder involving the moments of Z_n, using the fact that $\log(1+z)$ can be approximated by z for small z, we can first take logs of both sides in (17), apply the approximation and then exponentiate back to get

$$E(e^{itZ_n}) = \exp\{n\sum_{j=2}^{4} \frac{(it/\sqrt{n})^j \rho_j}{j!} + \frac{t^4}{n}\theta(t, n)\} \tag{18}$$

where $\sup_{n\geq 1}\sup_{|t|\leq \epsilon\sqrt{n}} |\theta(t, n)| \to 0$ as $\epsilon \to 0$ and $\rho_j = \frac{\kappa_j}{\kappa_2^{j/2}}$ are the *standardised cumulants*. We see that the leading term in the summation in the exponent (for $j = 2$) is $-t^2/2$. Hence the RHS of (18) is in the form

$$\exp(-t^2/2)\{1 + (...)/\sqrt{n} + (..)/n + \text{ remainder }\}$$

and the remainder is "under control". When we calculate $\chi_G(t) = \int e^{itx} dG(x)$ we get again

$$\exp(-t^2/2)\{1 + (...)/\sqrt{n} + (..)/n\}$$

with the same expressions in the brackets (..) in front of the $n^{-1/2}$ and n^{-1} powers.

At this point we can apply the smoothing lemma. We apply it with $T = n^4$, however when evaluating the RHS we subdivide the integral part into two regions. More precisely, we take $T_1 = \epsilon\sqrt{n}$ for a suitably small but fixed $\epsilon > 0$ and evaluate the integral separately for regions $t \in [-T_1, T_1]$ and for $T_1 \leq |t| \leq T$. For the evaluation in $T_1 \leq |t| \leq T$ we need Cramér's condition. The details of these evaluations are subtle and are omitted. The final evaluation from above of the right hand side helps us to bound the left hand side:

$$\sup_z |F(z) - G(z)| = o(n^{-1}).$$

8.3.2 Edgeworth expansion for the density of Z_n

Formally, by differentiating both sides in the the expansion from Theorem 8.1 we can get the two term expansion for the density of $f_{Z_n}(z)$ of Z_n. The final result is:

$$\phi(z)\{1 + \frac{\rho_3}{6\sqrt{n}} H_3(z) + \frac{1}{n}[\frac{\rho_4 H_4(z)}{24} + \frac{\rho_3^2 H_6(z)}{72}]\} + o(n^{-1}) \tag{19}$$

and the expansion (19) holds uniformly in $z \in R^1$. Here we have denoted by $\rho_r = \kappa_r/\kappa_2^{r/2}$ the *standardised cumulants*. As you can easily convince yourself, $H_4(z) = z^4 - 6z^2 + 3, H_6(z) = z^6 - 15z^4 + 45z^2 - 15$ holds.

The interpretation of (19) is that the leading term is the standard normal density in lieu with the Central Limit Theorem. Then there are higher order correction terms whose relevance becomes important for small to moderate sample sizes. The correction terms represent simultaneous adjustment of the normal approximation by using the information for the standardised skewness and kurtosis. It is also interesting to note that if the distribution of X is symmetric (hence $\rho_3 = 0$) the normal approximation is more accurate in the sense that the correction term is then of order n^{-1} rather than the $n^{-1/2}$ when $\rho_3 \neq 0$. Other important observation to be made is that the accuracy of the approximation depends on the position of the argument z. When we are in the tails (i.e., for $|z|$ large enough, the Edgeworth approximation (19) may worsen quite a lot and may become even negative. The same observation holds also for (15) which may become either negative or bigger than one in the tails. The saddlepoint approximation to be discussed next is meant to improve upon Edgeworth especially in the tails.

8.3.3 Cornish-Fisher expansions for quantiles

One major application of the Edgeworth expansion in practice is in fact in "reversing" it to construct better confidence intervals for small to moderate sample sizes. When studying coverage probabilities of confidence intervals, we need to solve, given confidence level α, equations of the type $F_{Z_n}(z_\alpha) = 1 - \alpha$ as accurately as possible. We know that Z_n is "close to standard normally distributed" and hence believe, not unreasonably, that z_α is in a vicinity of $u_\alpha : \Phi(u_\alpha) = 1 - \alpha$. If an Edgeworth expansion of the cdf $F_{Z_n}(.)$ has been

obtained already, one can of course try to replace $F_{Z_n}(z_\alpha)$ by its Edgeworth expansion and then equate this Edgeworth expansion to $\Phi(u_\alpha)$. Solving the resulting equation w.r. to the argument (that is, "inverting" it), gives an approximation for z_α. This inversion will involve the powers of u_α and is called *Cornish-Fisher* expansion η_α. The precise result (again if we contain ourselves to order $o(n^{-1})$ only) is given as follows:

Theorem 8.2. *Let $F(.)$ be the cdf of a single observation having a finite moment generating function in a neighbourhood of 0 and satisfying Cramér's condition. Then*

$$\eta_\alpha = u_\alpha + \frac{(u_\alpha^2-1)\rho_3}{6\sqrt{n}} + \frac{(u_\alpha^3-3u_\alpha)\rho_4}{24n} - \frac{(2u_\alpha^3-5u_\alpha)\rho_3^2}{36n} + o(n^{-1})$$

holds.

The accuracy of the above two-term Cornish-Fisher expansion is usually quite impressive! One numerical illustration follows now.

Example . Assume that $W_n \sim \chi_n^2$ is chi-squared with n d.f. Then we know that per definition. W_n/n can be represented as an average $\bar{X^2}$ of n i.i.d. squared standard normals $X_i, i = 1, 2, \ldots, n$. You can calculate the leading cumulants of the square of a standard normal directly (or you can help yourself with `mathStatica` as illustrated in the handout). You get for the standardised cumulants: $\rho_3 = 2\sqrt{2}, \rho_4 = 12$. Since $\sqrt{n}(W_n/n-1)/\sqrt{2} = (W_n-n)/\sqrt{2n}$ is about standard normal by the central limit theorem, the first order approximation of the α -quantile of W_n would be just $n+u_\alpha\sqrt{2n}$. The second order approximation using Theorem above will be

$$n + \sqrt{2n}[u_\alpha + \frac{(u_\alpha^2-1)2\sqrt{2}}{6\sqrt{n}} + \frac{(u_\alpha^3-3u_\alpha)12}{24n} - \frac{(2u_\alpha^3-5u_\alpha)8}{36n}] =$$

$$n + \sqrt{2n}u_\alpha + \frac{(u_\alpha^2-1)2}{3} + \frac{(u_\alpha^3-7u_\alpha)}{9\sqrt{2n}}$$

For $\alpha = 0.01$ we have $u_\alpha = 2.326$. Applying the Central Limit Theorem-based approximation $n+u_\alpha\sqrt{2n}$ for $n = 5$ gives 12.36. The first order approximation $n+\sqrt{2n}[u_\alpha+\frac{(u_\alpha^2-1)2\sqrt{2}}{6\sqrt{n}}]$ gives 15.296 whereas the second order approximation delivers 15.16. The true 99th percentile of W_n for $n = 5$ is known to be 15.09. Thus clearly in this case, the CLT-based approximation is much poorer and the two Cornsih-Fisher approximations are quite good, with the second order approximation "almost hitting" the true value.

8.3.4 Edgeworth expansions for other statistics

Our discussion was related to Edgeworth expansion for the normalised distribution of the sample mean because this is the easiest case to discuss. However, from practical point of view, Edgeworth expansions are of interest also for many other statistics. In particular, Edgeworth expansions for the MLE are of particular interest and can be obtained.

8.3.5 Saddlepoint density approximation for the mean

As mentioned already, the Edgeworth expansion of the density of Z_n may not be very accurate particularly in the tails where it could even become negative. One explanation

for this phenomenon is that the error approximation in the Edgeworth expansion is only *absolute* instead of *relative* and in the tails, where the true density is small to start with, even a small absolute error may turn out to be quite large relative to the true density's value and can cause serious error of the approximation. This is the reason to look for alternative expansions where the error of approximation is actually relative. The *Saddlepoint approximation method* offers such alternatives. It provides very accurate numerical approximations for densities and tail areas of statistics of interest down to surprisingly small sample sizes such as 5 or 10 far out in the tails.

There are several approaches to introduce the saddlepoint approximation. The historically first one is linked to the original paper by Henry Daniels in *The Annals of Mathematical Statistics* in 1954. This approach also explains the name of the method. It is based on Fourier inversion of the moment generating function and trying to approximate this inversion (represented as integral over the imaginary axis) by exploiting the biggest contribution to the integral coming form a small region around the *saddlepoint* of the integrand.

However, it is our feeling that the above approach is more difficult to comprehend. The second approach seems to be easier and leads to the same result so we only outline this second approach here. We start with the observation that since $Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$, if we use the density transformation formula, we can rewrite (19) in terms of the variable $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as follows:

$$f_{\bar{X}}(\bar{x}) = \frac{\sqrt{n}}{\sigma}\phi(z)\{1 + \frac{\rho_3}{6\sqrt{n}}H_3(z) + \frac{1}{n}[\frac{\rho_4 H_4(z)}{24} + \frac{\rho_3^2 H_6(z)}{72}] + o(n^{-1})\} \tag{20}$$

where $z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$. Because at the centre $z = 0$ the Hermite polynomials are $H_3(0) = 0, H_4(0) = 3, H_6(0) = -15$ we can see that the approximation (20) is in fact **more accurate** when $\bar{x} = \mu$, i.e., at the mean, since then the $O(\frac{1}{\sqrt{n}})$ correction term disappears. The explicit expression for the approximation there is:

$$f_{\bar{X}}(\bar{x}) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\{1 + \frac{1}{n}(\frac{1}{8}\rho_4 - \frac{5}{24}\rho_3^2) + O(n^{-2})\}. \tag{21}$$

The idea is then, if we want similar better accuracy for other values of $\bar{x} \neq \mu$, to *tilt* the distribution (apply the so-called *exponential tilting*≡ Esscher transform) to shift the mean so that $\bar{x}$ coincides with the new mean.

More specifically, we notice first of all that for any random variable X with a density $f(x)$ and a cumulant generating function $K_X(t)$ that converges on an open neighbourhood A of the point 0, we can define the tilted density family $f(x;t) = e^{tx-K_X(t)}f(x), t \in A$. This means that in fact we are embedding the original density $f(x)$ in a one-parameter exponential family parameterized (indexed) by t. Of course, we can also write

$$f(x) = e^{K_X(t)-tx}f(x;t) \text{ for any } t \in A \tag{22}$$

thus allowing us a choice of a suitable $t \in A$ when representing $f(x)$.

We can also easily see (DO IT (!)) that if X_t denotes a random variable with the density $f(x;t)$ then for the its mean, variance and standardised cumulants we have:

$$E[X_t] = K'_X(t), Var[X_t] = K''_X(t), \rho_i(t) = \frac{K_X^{(i)}(t)}{K''_X(t)^{i/2}}, i \geq 3. \tag{23}$$

Our strategy to derive the saddlepoint approximation then is as follows:

Step one: Given x, choose a suitable $\hat{t}$ such that the Edgeworth expansion for $f(x;\hat{t})$ is most accurate (and this means, as seen, that it is applied at the mean of $X_{\hat{t}}$).

Step two: Transform back to calculate $f(x)$ using the formula (22) by replacing t by $\hat{t}$ there.

To implement Step one, means that the fixed value of the argument x has to be the mean of X_t, that is,

$$E[X_t] = K'_X(\hat{t}) = x \tag{24}$$

The equation (24) is called the saddlepoint equation. Then we apply the Edgeworth expansion at the mean as given in (21). This gives

$$f(x;\hat{t}) \approx \frac{1}{\sqrt{2\pi K''_X(\hat{t})}}\{1 + [\frac{1}{8}\rho_4(\hat{t}) - \frac{5}{24}\rho_3^2(\hat{t})]\}.$$

Then in the second step we substitute the above approximation in the right hand side of (22) to get the approximation:

$$f(x) = e^{\{-\hat{t}x + K_X(\hat{t})\}} f(x;\hat{t}) \approx \frac{1}{\sqrt{2\pi K''_X(\hat{t})}} e^{[K_X(\hat{t}) - \hat{t}x]}\{1 + [\frac{1}{8}\rho_4(\hat{t}) - \frac{5}{24}\rho_3^2(\hat{t})]\}. \tag{25}$$

The discussion until now involved essentially *any* random variable X. The asymptotic point of view is brought forward when (25) is applied to the random variable $\bar{X}$. The density $f(\bar{x})$ of $\bar{X}$ is tilted in the exponential family

$$f(\bar{x};t) = e^{nt\bar{x} - nK_X(t)} f(\bar{x})$$

with $\bar{X}_t \sim f(\bar{x};t)$. The mean of $\bar{X}_t$ is, of course, just $K'_X(t)$ and the optimal tilt is obtained by solving $K'_X(\hat{t}) = \bar{x}$. It is easily seen that the standardised cumulants for $\bar{X}_{\hat{t}}$ are just $\hat{\rho}_i = n^{1-i/2}\rho_i(\hat{t})$, where $\rho_i(t)$ are given in (23). Hence we get

$$\hat{f}(\bar{x}) \approx \sqrt{\frac{n}{2\pi K''_X(\hat{t})}} e^{\{nK_X(\hat{t}) - n\hat{t}\bar{x}\}}\{1 + [\frac{1}{8n}\hat{\rho}_4 - \frac{5}{24n}\hat{\rho}_3^2]\}. \tag{26}$$

The above approximation (26) is called the *second order saddlepoint density approximation* and is *extremely accurate*! Sometimes, for the sake of numerical simplicity, even just the *first order saddlepoint approximation*

$$\hat{f}(\bar{x}) \approx \sqrt{\frac{n}{2\pi K''_X(\hat{t})}} e^{\{nK_X(\hat{t}) - n\hat{t}\bar{x}\}} \tag{27}$$

is used and it is still very accurate!

Final note: The saddlepoint approximation (26) is quite different from the Edgeworth expansion. It is an asymptotic expansion on powers of n^{-1} instead of $n^{-1/2}$, as in the Edgeworth. This implies that already the simple approximation (27) has already absorbed the skewness correction. The leading term (27) is clearly *not* the normal (and in fact any other) density; it may not integrate to one (although it can be renormalised to integrate to one). The saddlepoint approximation is more accurate than Edgeworth, especially in

the tails. These advantages are achieved due to the fact that for the derivation of the saddlepoint, we used the *whole cumulant generating function* of the distribution of X_1 whereas for Edgeworth expansion we only need the four leading cumulants. Also, the saddlepoint method is computationally more intensive: the saddlepoint equation $K'_X(\hat{t}) = \bar{x}$ has to be solved for each value of the argument $\bar{x}$. To summarise, via the saddlepoint approximation "by requiring more we achieve more".

8.3.6 Saddlepoint for CDF and examples

First, we note that a very accurate asymptotic approximation exists also for the cumulative distribution function (CDF) of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $K_X(t)$ is the cumulant generating function of X_1. This approximation is derived via two layers of approximation: first the density is approximated via (27). The CDF is approximated as integral of this approximated density then. Finally, the resulting integral is approximated using the *Temme approximation*. The latter represents an approximate integration by parts formula. We omit the details and only give the final formula. It is called the *Lugannani-Rice* formula:

$$P(\bar{X} \leq \bar{x}) = \Phi(\hat{w}_n) + \phi(\hat{w}_n)[\frac{1}{\hat{w}_n} - \frac{1}{\hat{u}_n}] + O(\frac{1}{n})(\text{ for } \bar{x} \neq E(\bar{X}).$$

Here $\hat{w}_n = sgn(\hat{t})\sqrt{2n[\hat{t}\bar{x} - K_X(\hat{t})]}, \hat{u}_n = \hat{t}\sqrt{nK''_X(\hat{t})}$ and $\hat{t}$ is the saddlepoint, that is, the solution to the equation: $K_X(\hat{t})' = \bar{x}$. Like in the density case, the term of order $O(\frac{1}{n})$ can also be calculated but is more complicated and we do not give its explicit form here.

Examples Typically, it is difficult to give closed form formulae for the saddlepoint approximations in particular cases since the saddlepoint equation is typically non-linear and it is solved using iterative numerical methods. Some very simple cases can be dealt with.

Example 1. The saddlepoint approximation for the sample mean is exact for the standard normal distribution. $f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Indeed, we have $K_X(t) = \frac{t^2}{2}, \hat{t} = K'_X(\hat{t}) = \bar{x}, K''_X(t) \equiv 1$ hence (27) gives

$$\hat{f}(\bar{x}) = \sqrt{\frac{n}{2\pi}}e^{-\frac{n\bar{x}^2}{2}}$$

which is the density of $N(0, \frac{1}{n})$. Also, the Lugannani-Rice formula gives $P(\bar{X} \leq \bar{x}) = \Phi(\sqrt{n}\bar{x})$ which is the CDF of $N(0, \frac{1}{n})$.

Example 2. Saddlepoint approximation for the density of the sample mean of Gamma(α, 1) density. Here $f(x) = \frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}, x > 0$.

We have $K_X(t) = -\alpha\log(1-t), t < 1, K''_X(t) = \frac{\alpha}{(1-t)^2}$ and the saddlepoint equation $\frac{\alpha}{1-t} = \bar{x}$ has a root $\hat{t} = 1 - \frac{\alpha}{\bar{x}}$. This implies $K''_X(\hat{t}) = \bar{x}^2/\alpha$. Hence we get from (27)

$$\hat{f}(\bar{x}) = \sqrt{\frac{n\alpha}{2\pi\bar{x}^2}}exp[-n\alpha\log(1-\hat{t}) - n\hat{t}\bar{x}] = (\sqrt{\frac{2\pi}{n\alpha}}(n\alpha)^{n\alpha}e^{-n\alpha})^{-1}(n\bar{x})^{n\alpha-1}e^{-n\bar{x}}n$$

The expression $\frac{1}{\Gamma(n\alpha)}(n\bar{x})^{n\alpha-1}e^{-n\bar{x}}n$ is the exact density of $\bar{X}$ and we see that the difference between the exact and the saddlepoint approximation is just that the exact normalising constant $\frac{1}{\Gamma(n\alpha)}$ has been replaced by the constant $(\sqrt{\frac{2\pi}{n\alpha}}(n\alpha)^{n\alpha}e^{-n\alpha})^{-1}$. However,

the famous *Stirling approximation* of the Gamma function states precisely the fact that $\sqrt{\frac{2\pi}{n\alpha}}(n\alpha)^{n\alpha}e^{-n\alpha} \approx \Gamma(n\alpha)$ hence the ratio of the two normalising constants tends to one when sample size is increased. In fact if we renormalised the saddlepoint approximation, we would again recover the exact density!

8.4 Extensions of the saddlepoint method

If it was only about calculating precise approximations for the density of the sample mean **only**, the saddlepoint method would not have got very widespread. However, it turns out that the saddlepoint approximation idea can be extend and can be applied for approximating densities of maximum likelihood estimators in exponential families, of likelihood ratio or score statistics, of Bayes estimators etc. Many statistics can be approximated by sums of i.i.d. random variables (as seen from the representation

$$T(F_n) \approx T(F) + \frac{1}{n}\sum_{i=1}^{n} a(X_i) + remainder$$

in our discussion about estimating statistical functionals. For these, again the saddlepoint idea can be applied. There are also multivariate extensions of the method to approximate the joint distribution of vector-statistics.

We do not discuss these because of lack of time. We will finish with a formula (called *Barndorff-Nielsen's formula*) about the saddlepoint approximation of the density of the MLE in a k-parameter exponential family:

If $F_X(x;\theta) = e^{\theta' t(x) - \psi(\theta) - d(x)}$ is the density of a single observation, $L(\theta)$ is the joint likelihood of the sample, if $J(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}$ then for the density of the MLE of $\theta \in R^k$, the following approximation can be derived:

$$\hat{f}(\theta) = (2\pi)^{-k}|J(\hat{\theta})|^{1/2}\{L(\theta)/L(\hat{\theta})\}[1 + O(n^{-1})] \tag{28}$$

(We note that the MLE $\hat{\theta}$ is a one-to-one transformation of the statistic $T = \sum_{i=1}^{n} t(X_i)$ in this case).

If in fact the constant $(2\pi)^{-k}$ is replaced by the true renormalisation constant then the order of the approximation in (28) improves to $O(n^{-3/2})$. The accuracy of formula (28) for approximating the density of the MLE, especially after renormalisation, is so impressive that it has been termed "the magic formula" by prominent statisticians (in an overview paper with discussions: "R. A. Fisher in the 21st Century" by B. Efron (*Statistical Science*, 1998, Vol. 13, No. 2, 95-122). Applications of this formula outside the exponential family, have been also been investigated by Barndorff-Nielsen.

9 Lecture 9: Robustness. Estimating statistical functionals.

9.1 Motivation. Basic idea of robustness

Along this course, we have studied theories about how to construct optimal procedures (be they Likelihood-based or Bayesian) when certain parametric model $F(X, \theta)$ is given. These theories say nothing about the behaviour of the optimal procedures when the models are only approximately valid. Going over in such cases directly to purely Nonparametric approach would also not address properly the situation since the idea about (relatively small) deviation form a baseline parametric model would be lost. The proper approach would be the robustness approach where we still keep the idea about the ideal parametric model but allow for deviations from it. Speaking loosely, nonparametric statistics allows "all" possible probability distributions and reduces the ignorance about them only by one or a few dimensions. Classical parametric statistics allows only a very "thin" finite-dimensional subset of probability distributions, i.e., the ideal parametric model of interest for which usually optimal inferences are available. Robust statistics allows a full-dimensional neighbourhood of a parametric model, thus being more realistic and yet, at a price of a relatively small loss of efficiency at the ideal model, provides almost the same advantages as a strict parametric model in a "broader" neighbourhood of the ideal parametric model.

The problem in robustness is to construct estimators that are **close to efficient** if the parametric model holds but are at the same time **less sensitive** to small deviations from the ideal model.

9.1.1 Simple example

One of the simple examples to start with, is estimating the location parameter of a continuous symmetric distribution. Assume a sample $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ is available from a location parameter family $F(x, \theta) = F(x - \theta), \theta \in R^1$. Denote the density by $f(x, \theta) = f(x - \theta)$. If F is a normal distribution then θ coincides with its mean, median and mode. As we know, in this case the estimator $\bar{\mathbf{x}}$ is efficient for θ for any fixed sample size. But assume now that F is Cauchy with a density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. Then $f(x, \theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$. The parameter θ in this model does **not** coincide with the mean of the distribution (in fact, the Cauchy distribution does not have a finite mean) but coincides with its median and mode. It can be shown (see lecture) that $\bar{\mathbf{x}}$ has **the same distribution** as the distribution of a single observation from the Cauchy model! Therefore $\bar{\mathbf{x}}$ is even **not** consistent for θ in the Cauchy model! The reason for the good behaviour of $\bar{\mathbf{x}}$ as an estimator of location parameter θ in the normal family and for its "bad" behaviour in the Cauchy family are the **heavy tails** of the Cauchy distribution, i.e. it allows with a large probability for very large (in absolute value) realizations to occur. Because of this observation, we would decide to ignore the observations with a large absolute value and use the **empirical median** instead when estimating the location parameter of the Cauchy distribution. The empirical median $\tilde{\theta}_n$ is **not sensitive** to large realizations in the tail of the distribution, hence it is more robust as a location parameter estimator.

Assume for simplicity that sample size n is odd. From theoretical derivations to be demonstrated later (in the Section about influence functions in this lecture) we know that for a symmetric F (i.e. $F(0) = 1/2$) with a density $f(0) > 0$,

$$\sqrt{n}(\tilde{\theta}_n - \theta) \to^d N(0, \frac{1}{4f^2(0)}). \tag{29}$$

Hence, if F is a standard normal, the asymptotic variance of the median will be $\frac{\pi}{2}$ whereas the variance of $\bar{x}$ would be one in this case. This means $\tilde{\theta}_n$ is not asymptotically efficient when the family $f(x, \theta)$ is the normal family but it is consistent and even asymptotically normal as a location parameter estimator **simultaneously** for both the normal **and** the Cauchy family.

Often, in practice it would be not sure if the family is normal, of Cauchy, or another location family with tails that are heavier than the ones of the normal but less heavy than the ones of the Cauchy family. So how to estimate the location parameter then? A compromise between the mean and the median is the α**- trimmed mean** $\bar{x}_\alpha = \frac{1}{n-2k}[x_{(k+1)} + x_{(k+2)} + \ldots x_{(n-k)}]$ where $k/n = \alpha < 1/2$ (i.e. we trim symmetrically $2\alpha 100\%$ of the observations and average the rest). It can be shown that this estimator also has an asymptotically normal distribution and when α is small, it has a high efficiency at the normal family.

9.1.2 Extension of the discussion. The notion of an M-estimator.

Another compromise is suggested by the following observation. It is well known that $\bar{x}$ minimizes the sum $\sum_{i=1}^n (x_i - a)^2$ for all possible values of a whereas $\tilde{\theta}_n$ minimizes the sum $\sum_{i=1}^n |x_i - a|$. When the information is incomplete, it would be a good idea to choose to minimize a function in the form $\sum_{i=1}^n \rho(x_i - a)$ where ρ is symmetric nonnegative and $\rho(0) = 0$. An example of such a function is given by: $\rho(x) = \begin{cases} \frac{1}{2}x^2 \text{ if } |x| \le k \\ k|x| - \frac{1}{2}k^2 \text{ if } |x| \ge k, \end{cases}$ ' k being a positive constant. This is Huber's famous ρ-function. If k is growing, $\rho(x)$ will coincide with $\frac{1}{2}x^2$ on almost the whole interval and the estimator will approach $\bar{x}$. In the opposite case, when k is getting smaller, one gets values closer to $\tilde{\theta}_n$. As a compromise values, $k = 1.5$ and $k = 2$ are suggested. The estimators one gets through the minimization:

$$\min_a \sum_{i=1}^n \rho(x_i - a)$$

have the common name **M-estimators**. When ρ is differentiable (such is the case with the Huber function) they can also be considered as a solution of the equation

$$\sum_{i=1}^n \psi(x_i - a) = 0.$$

To find such estimators, one needs to apply iterative procedures. Under certain regularity conditions, they are asymptotically normal with asymptotic variance equal to $\sigma^2(F, \psi) = \frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2}$. The derivation of this result is in fact not difficult and follows the same steps as the proof of asymptotic normality of the maximum likelihood estimators. You

can find the details in Casella and Berger, p. 486. Obviously, when $\rho(x) = -lnf(x)$ (in the ideal case when f was known), one gets the MLE estimator for the location parameter of the family $f(x, \theta)$.

The most common formulation of the robustness approach is the following. Instead of assuming that F is known precisely, we assume that F is in a ϵ -neighbourhood of certain distribution G, i.e. $F(x) = (1-\epsilon)G(x) + \epsilon H(x) = F_H(x)$ where $\epsilon < 1/2$ and G are given. They reflect the "amount of contamination" (ϵ) of the "ideal" G. We are interested in finding the M-estimator that makes $\inf \sup_{F_H} \sigma^2(F_H, \psi)$ (i.e., minimax approach). Here $\sup_{F_H}$ is taken over all symmetric continuous distributions H. Then it can be shown that if G is the standard normal , the minimax M-estimator is a special ρ- estimator of Huber where the constant k can be determined as a function of the contamination ϵ and is the root of the equation

$$\frac{1}{1-\epsilon} = \int_{-k}^{k} \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})dx + \frac{2}{k}\frac{1}{\sqrt{2\pi}} exp(-k^2/2)$$

9.2 Robustness approach based on influence functions

Huber's approach to optimality in robustness as described above is very attractive from theoretical point of view. However it does have the drawback that it requires in the case of the contaminated normal the contamination itself to be symmetric. This is unrealistic. An alternative approach to robustness based on Influence functions was introduced by Hampel and nowadays it dominates the robustness theory. It does not need the assumption about symmetric contamination and has broader applicability, notably in regression.

Roughly speaking, the influence function (IF) describes the effect of an additional observation in any point x on a statistic $T(X)$, given a large sample with distribution F. To give a more precise idea, we consider paremeters of interest as functionals.

Note: Indeed, virtually every parameter of interest of a given distribution can be presented as a functional of the distribution: for example, the mean μ is $\mu = \int xdG(x)$, the variance σ^2 is $\sigma^2 = \int (x - \int xdG(x))^2 dG(x)$, etc. Then it becomes natural to define estimators of parameters as resulting from applying the coresponding defining functional on the empirical distribution function **(edf)** (i.e. $T_n(G_n) = T(G_n)$ where G_n is the empirical distribution function of the sample). Since it is known that the edf is a very good estimator of the true distribution it is then not unreasonable to expect that $T_n(G_n) = T(G_n)$ would be a good estimator of the parameter.

We say that T(G) is the **asymptotic value** of $T_n, n \geq 1$ at G. We shall also assume that the functional under study are **Fisher consistent**, i.e. $T(F_\theta) = \theta$ (which means that on the ideal family they estimate the right parameter asymptotically). Under regularity conditions, one can assume that the limit:

$$lim_{t\to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int a(x)dG(x) \tag{30}$$

holds which equivalently can be written as

$$\frac{\partial}{\partial t}[T((1-t)F + tG)]_{t=0} = \int a(x)dG(x) = L_F(G)$$

and we call it the *Gâteaux derivative* of the functional T at F in direction G. By putting G=F in the latter equality, we also see from (30) that $\int a(x)dF(x) = 0$ holds. If we put for G the empirical measure that puts the whole mass 1 at the point x we get the *influence function* of the functional T :

$$IF(x;T,F) = \lim_{t\to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t} = a(x). \tag{31}$$

Heuristically, the IF describes the effect of infinitesimal contamination at the point x on the estimate, standardized by the mass of contamination. It measures the asymptotic bias caused by contamination.

(Note: The IF of a functional can naturally be estimated via the empirical influence function

$$IF(x;T,\hat{F}_n)$$

where $\hat{F}_n$ is the edf of the data.)

From the definition (30) we see that, since for large n, F_n and F are close, we can write then approximately

$$T(F_n) \approx T(F) + \int IF(x;T,F)dF_n(x) + remainder \approx T(F) + \frac{1}{n}\sum_{i=1}^{n} a(X_i) + remainder$$

If the remainder behaves well under further regularity conditions (the so-called Hadamard differentiability of the functional), from here we also get the asymptotic variance of the estimator. The role of the remainder would be negligible in the calculation of the asymptotic variance and we would end up the statement that $T_n = T(F_n)$ has the property that

$$\sqrt{n}[T_n - T(F)]$$

tends in distribution to $N(0, V(T,F))$ with $V(T,F) = \int IF(x;T,F)^2 dF(x)$. Of course, for large n, we can estimate $\hat{V}(T,F) = V(T,F_n) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)^2$ and we have the approximate result

$$\sqrt{n}(T(F_n) - T(F)) \approx N(0, \hat{V}(T,F)). \tag{32}$$

Then a confidence interval for $T(F)$ can easily be constructed.

Note. The statement (32) is often referred to as the *nonparametric delta method.* The reason behing this name is that indeed in (32) we are "transferring" a result related to the asymptotic distribution of $\sqrt{n}(F_n(t) - F(t)), t \in (-\infty, \infty)$ as a stochastic process to a distribution of a functional T applied to its paths. It is a known fact from probability theory that the stochastic process $G_n(t) = \sqrt{n}(F_n(t) - F(t))$ converges to a zero mean Gaussian process with a covariance function $E[G(s)G(t)] = F(\min(s,t)) - F(s)F(t)$.

9.2.1 Simple Examples

1) Let $T(F) = \int xdF(x) = \mu(F)$ (the mean). An estimator of this functional is $T(F_n) = \int xdF_n(x) = \bar{X}$. We also see that

$$T((1-t)F + t\Delta_x) - T(F) = (1-t)T(F) + tx - T(F) = t(x - T(F))$$

and from here we get $a(x) = x - T(F) = x - \mu$. If we want to estimate it we get $\hat{a}(x) = x - \bar{X}$ and then $\hat{V}(T, F) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. An asymptotic confidence interval for $T(F)$ at level $(1-\alpha)$ would be, not surprisingly:

$$\bar{X} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

2) Second example. Let $F(x)$ be strictly increasing with a positive density $f(x)$. The $p-th$ quantile ($p \in (0,1)$) of F is defined as the point $q_p : F(q_p) = p$ and clearly it is very important object in Statistical Inference. Symbollically, you can also denote $q_p(F) = T(F) = F^{-1}(p)$. The $p-th$ quantile is used to evaluate threshold constants in designing tests and it is also important in its own right. It became hugely prominet in financial applications (the so-called **VaR** (value at risk) is just a quantile). Estimating this quantile is an important problem in Statistical inference. One obvious candidate is just the empirical quantile $\hat{q}_p$ defined as a solution to the equation $\tilde{F}_n(\hat{q}_p) = p$ where $\tilde{F}_n$ is (linearised version of) the empirical distribution function. We want to derive the influence function of $q_p(F) = T(F) = F^{-1}(p)$. Defining the function

$$H(y, x) = H(y - x) = \left\{ \begin{array}{l} 1 \text{ if } y \geq x \\ 0 \text{ otherwise} \end{array} \right\}$$

we can represent a perturbation of $F(y)$ via point-mass in x as follows:

$$F_t(y) = (1-t)F(y) + tH(y - x), t \in [0, 1].$$

Now, per definition $p = F_t(q_p(F_t))$ holds and taking derivatives from both sides, and applying the chain rule, we get:

$$\frac{d}{dt}p_{|t=0} = 0 = \frac{d}{dt}F_t(q_p(F_t))_{|t=0} = f(q_p)\frac{d}{dt}q_p(F_t)_{|t=0} - p + H(q_p(F) - x)$$

This implies

$$\frac{d}{dt}q_p(F_t)_{|t=0} = IF(x; T, F) = \left\{ \begin{array}{l} \frac{p-1}{f(q_p)} \text{ if } x \leq q_p(F) \\ \frac{p}{f(q_p)} \text{ if } x > q_p(F) \end{array} \right\}$$

From here, we also get the asymptotic variance of $\sqrt{n}(T_n - T(F))$ for the empirical $p-th$ quantile as $V(T, F) = p\frac{(p-1)^2}{f(q_p(F))^2} + (1-p)\frac{p^2}{f(q_p(F))^2} = \frac{p(1-p)}{f(q_p(F))^2}$

Note Applying this result for the case of the median $p = 1/2$ and for F being the standard normal distribution, we discover (29). Note also that if we wanted to use the above result to approximate the asymptotic variance of the sample quantile estimator, we sould need to involve a non-parametric density estimator $\hat{f}$. More stable and simpler estimate of this asymptotic variance can be obtained using the bootstrap method.

9.3 Using the influence function in practice of robust inference.

A sample analogue of the influence function of an estimator $\hat{\theta}_n$ is its so-called **sensitivity curve (SC).** It is defined as

$$SC_{n-1}(x) = n[\hat{\theta}_n(x_1, x_2, \ldots, x_{n-1}, x) - \hat{\theta}_{n-1}(x_1, x_2, \ldots, x_{n-1})].$$

The SC is a function of x and it measures how much an estimator can change when an observation with value x is added to a data set consisting of $(n-1)$ observations $x_1, x_2, \ldots, x_{n-1}$. Robust estimators are the ones for which the sensitivity curve is bounded.

Note: An advantage of the SC in comparison to the IF is that the former can be calculated from the sample whereas the latter is just a theoretical concept. However the SC is sample dependent therefore it is not uniquely defined. One can in fact consider the SC as a (non-parametric) estimator of the IF.

9.3.1 Example

Find the SC for the sample mean and for the sample median. In particular show that the former is unbounded whereas the latter is bounded in x.

i) For the mean $\bar{X} = \hat{\theta}_n$:

$$SC_{(n-1)}(x) = n[\frac{\sum_{i=1}^{n-1} x_i + x}{n} - \frac{\sum_{i=1}^{n-1} x_i}{n-1}] = \frac{(n-1)x - \sum_{i=1}^{n-1} x_i}{n-1} = x - \bar{x}_{n-1}.$$

If we let $n \to \infty$ we get the influence function $x - \mu$ from Section 9.2.1 thus confirming again that the sensitivity curve is just a finite sample analogue of the influence function.

ii) For the median, we first consider the case of an odd sample size $n = 2k+1$. Then $\hat{\theta}_{n-1} = \frac{1}{2}[x_{(k)} + x_{(k+1)}]$ whereas $\hat{\theta}_n(x_1, x_2, \ldots, x_{n-1}, x) = \left\{ \begin{array}{l} x_{(k)} \text{ if } x < x_{(k)} \\ x \text{ if } x_{(k)} \leq x \leq x_{(k+1)} \\ x_{(k+1)} \text{ if } x > x_{(k+1)} \end{array} \right\}$ Hence for $n[\hat{\theta}_n - \hat{\theta}_{(n-1)}]$ we get:

$$SC_{(n-1)}(x) = \left\{ \begin{array}{l} -\frac{n}{2}[x_{(k+1)} - x_{(k)}] \text{ if } x < x_{(k)} \\ \frac{n}{2}[2x - x_{(k)} - x_{(k+1)}] \text{ if } x_{(k)} \leq x \leq x_{(k+1)} \\ \frac{n}{2}[x_{(k+1)} - x_{(k)}] \text{ if } x > x_{(k+1)} \end{array} \right\}$$

This implies that no matter what the value of x, we end up with

$$|SC_{(n-1)}(x)| \leq \frac{n}{2}[x_{(k+1)} - x_{(k)}].$$

The case of even sample $n = 2k$ is similar and is left for you to do it.

As we have seen, the IF can be used in assessing the asymptotic variance of an estimator. Having derived the influence function of an estimator, one can evaluate its robustness properties. The rough guide is that for an estimator to be robust, its IF must be bounded. However, finer robustness properties can also be investigated by studying the IF. For example:

- The **gross error sensitivity** $\gamma = \sup_x |IF(x : T, F)|$ measures the worst influence which a small amount of contamination can have on the value of the estimator. For robustness purposes, it has to be finite.
- Rejection of outliers in terms of IF means that IF should vanish outside certain area.

- **Breakdown point** is the maximal amount of model misspecification before an estimator breaks down (meaning that its bias becomes infinite). Formally the breakdown point is the value $\epsilon^* = \inf_{\epsilon \in (0,1/2]} |bias(\hat{\theta}, F_\theta, \epsilon) = \infty$. The breakdown point of the sample mean is 0 but for the median is the maximal possible ($= 1/2$.) WHY!

10 Lecture 10: Introduction to the Bootstrap

10.1 Motivation

In Statistical Inference we are "learning from experience": we observe a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and wish to infer properties of the complete population that yielded the sample. A complete knowledge about the population is obtained from the *population distribution function* $F(.)$. This (unknown (!)) function is usually estimated in a non-parametric framework (i.e. when no additional information about the population is available except the sample itself) by the *empirical distribution function* (EDF)

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i).$$

Basically, when calculating the EDF, we look at the empirical proportion of the realizations in the sample that happen to be $\leq x$ and use this expirical proportion to *estimate* the true unknown probability $F(x)$ for a realisation in the interval $(-\infty, x]$. The remarkable fact about the EDF is that despite its simplicity, it has been defended over the years by prominent statisticians as *the* asymptotically optimal estimator of the population distribution function. A statement first proved in a paper by Dvoretsky, Kiefer and Wolfowitz in 1956, it has been shown that the EDF is *asymptotically minimax* among the collection of all continuous distributions. As Millar (1979) notes, "This paper has stood for over 20 years as one of the pivotal achievements of nonparametric decision theory".

We will not discuss the precise meaning of the above mentioned asymptotic optimality in this course. Note that it goes about estimating a function (not a finite-dimensional parameter) and justifying the optimality requires defining it in a suitably chosen function space, with a suitable loss over the functions in this space etc. This requires some preparation in itself. The convergence of

$$\sqrt{n}(\hat{F}(x) - F(x))$$

towards a limiting Gaussian process (See Lecture 8) helps and is utilised but you probably can appreciate that the details are subtle and are skipped here.

However, encouraged by the existence of such results in the literature, we can be tempted to estimate some interesting aspect of $F(.)$ (such as its mean or median, higher order moment etc.) by using the corresponding aspect of $\hat{F}(.)$. For example, we would be tempted to estimate the population mean $\mu = \int x dF(x)$ by the sample mean $\bar{X} = \int x d\hat{F}(x)$. This approach to estimation is called *plug-in principle* (we plug-in the good estimator $\hat{F}$ instead of F in the formula for the theoretical mean and hope to get a good estimator of the theoretical mean itself). In this example, we *were* successful- we got the sample mean which is known to be a good estimator of the theoretical mean. More generally, any parameter of interest θ can usually be written as a statistical functional $\theta = t(F)$ and the obvious plug-in point estimator for θ would be $\hat{\theta} = t(\hat{F})$ then.

The (nonparametric) *bootstrap method* is an *application of the above discussed plug-in principle*. Originally, the bootstrap was suggested by B. Efron as a method to derive an estimate for the the *standard error* of arbitrary estimator. Finding the standard error of

an estimator is a significant activity for every statistician since, not being satisfied with a point estimator only, he/she is always looking for the *variablity* of the estimator. He/she would be interested in the bias, standard error or *even in the complete distribution* of the estimator itself. If available, these can be used to construct confidence intervals, to test hypotheses etc. for the parameter of interest.

Since theoretical investigation of the standard error is possible only in a limited number of "textbook cases" and even in these "textbook cases" the treatment is only asymptotic, Efron's idea was to use the bootstrap as a means to get standard error estimates of $\hat{\theta} = t(\hat{F})$ in an "*automatic way*", no matter how complicated the functional mapping $\theta = t(F)$ may be. We describe the idea below.

10.2 Nonparametric bootstrap.

In the original settings of the (nonparametric) bootstrap we assume that we need to estimate the *standard error* of the estimator $\hat{\theta} = t(\hat{F})$. *If we had several estimators* of size n of $t(F)$ then of course we could use their empirical standard error as an estimator for the unknown standard error! Unfortunately, we only have got *one sample* of size n that allows us to construct *one* plug-in estimator $\hat{\theta} = t(\hat{F})$. How could we the evaluate the sampling accuracy of this estimator when we only have one only realisation? The bootsrap helps us here in a very simple manner by helping us to generate **many** samples that look like the original sample hence to get many versions of the estimator by evaluating it on each of these samples.

To this end we consider the so called *bootstrap sample* of size n drawn *from* $\hat{F}$ instead of F! It is defined as $\mathbf{x}^* = (x_1^*, x_2^*, .., x_n^*)$, the star indicating that this is not the original data set but a randomized, or *resampled* version of **x.** Hence the bootstrap data points are a random sample *with replacement* from the population of n objects $(x_1, x_2, .., x_n)$ rather than from the original population (but we use them as if they were a new sample from the entire population). Correspondingly to the bootstrap sample, we can get a *bootstrap replication* $\hat{\theta}^*$ via plug-in. The process of getting new bootstrap samples and then plug-in, can be repeated as many times as we like! As a consequence, we get B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, .., \mathbf{x}^{*B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$. To each such sample, we calculate the corresponding $\hat{\theta}^*(b), b = 1, 2, .., B$ by the plug- in method and finally, we can calculate an estimator of the standard error $s_F(\hat{\theta})$ by applying the classical formula $\hat{s}_B = \{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \bar{\hat{\theta}^*}]^2/(B-1)\}^{1/2}, \bar{\hat{\theta}^*} = \sum_{b=1}^{B}\hat{\theta}^*(b)/B$.

Note that for a given sample size n, letting $B \to \infty$ means that we get in a limit the ideal bootstrap estimate $s_{\hat{F}}(\hat{\theta}^*)$ of $s_F(\hat{\theta})$, i.e. $\lim_{B\to\infty}\hat{s}_B = s_{\hat{F}}(\hat{\theta}^*)$. This ideal bootstrap estimate $s_{\hat{F}}(\hat{\theta}^*)$ is called a *nonparametric bootstrap estimate of the standard error*. In computer simulations, using the cheap computer power, the value B can be taken very large so that indeed $\hat{s}_B \approx s_{\hat{F}}(\hat{\theta}^*)$ holds and because of this, sometimes the value $\hat{s}_B$ itself is called a *nonparametric bootstrap estimate of the standard error*. Our hope is that when n is large and B is large, $\hat{s}_B$ will be close to $s_{\hat{F}}(\hat{\theta}^*)$ which will be close to $s_F(\hat{\theta})$. Much of the theory in bootstrap has been developed to justify such type of statements. From the discussion so far we see that the bootstrap can be considered as a *data- based simulation method for statistical inference*.

The use of the term bootstrap derives from the phrase *to pull oneself up by one's*

bootstrap from the eighteenth century adventures of Baron Münhausen. The method is extremely powerful and has made Efron an instant celebrity. Efron mentions that he thought about calling the method the *shotgun* since it " .. can blow the head off any problem if the statistician can stand the resulting mess". This quotation relates to bootstrap's wide applicability in combination to the, generally speaking, large volume of numerical work associated with its application.

It should be noted that when enough bootstrap resamples have been generated, *not only the standard error* but *any aspect of the distribution of the estimator* $\hat{\theta} = t(\hat{F})$ *could be estimated!* One can, for example, have a *histogram* of the distribution of $\hat{\theta} = t(\hat{F})$ by constructing a histogram based on the observed $\hat{\theta}^*(b), b = 1, 2, .., B$ values! This histogram can be used as an estimator of the density of the es timator $t(\hat{F})$. Specifically, the histogram would be very useful since it will indicate, for example, if normal approximation to the distribution of $t(\hat{F})$ is justified for the given fixed sample size n.

10.3 Parametric bootstrap

Of course, it is to be expected that in situations where there **is** more information about the population's distribution function F other than that provided by the sample, the plug-in principle would not be very good and the standard bootstrap method will need some modifications in order to be applied. For example, if there existed a *parametric model* for F, exact (or approximate asymptotic) analytic formulae may exist for the standard errors of some estimators and we could use them instead of the nonparametric bootstrap. But even in parametric situations the bootstrap's idea *still* can be used *parametrically*. We can apply the so called *parametric bootstrap* and the results are similar (and sometimes even better) than the textbook formulae. The parametric bootstrap estimate of standard error is defined as $s_{\hat{F}_{\text{par}}}(\hat{\theta}^*)$ where $\hat{F}_{\text{par}}$ is an estimate of F derived from a *parametric model of the data*. That is, in this case, instead of drawing independent samples with replacement from the data, we draw B samples of size n from the parametric estimate of the function F. After generating the B samples, we proceed in the usual way outlined in *1.* to calculate the estimator of the standard error. When used in the above parametric mode, bootstrap can provide more accurate results than approximate textbook formulae results based on asymptotic normality approximations and it also provides answers in problems for which no textbook formulae exist.

10.4 Numerical illustration

(B. Efron and R. Tibshirani give the example below to illustrate a non-trivial application of the bootstrap and to compare the accuracy of the method with the accuracy of alternative methods that exist to deal with this example). The goal is to **evaluate standard error of the estimator of the correlation coefficient**. A sample of size $n = 15$ is available from a set of American law schools participating in a large study of admission practices.

School	LSAT (X)	GPA (Y)
1	576	3.39
2	635	3.3
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

Two measurements are available on the entering classes of each school : LSAT and GPA. It is believed that these scores are highly correlated. One is interested in estimating the correlation coefficient **and** in obtaining an estimate of the standard error of the estimator. **Precise** theoretical formula for the standard error of the estimator is **unavailable** but an **asymptotic approximation formula exists**. As an alternative to it, and especially when the sample size is **not very large** ($n = 15$ only), the bootstrap estimate is considered. The numerical values are compared and show very close values.

A bivariate scatterplot indicates relatively strong linear relationship. Methods for testing bivariate normality suggest the data is likely to have been generated from a bivariate normal distribution. The parameter of interest is the correlation coefficient $\rho = \frac{E[(X-EX)(Y-EY)]}{\{E[X-EX]^2.E[Y-EY]^2\}^{1/2}}$. Its typical estimator is $\hat{\rho} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{[\Sigma x_i^2 - n\bar{x}^2]^{1/2}.[\Sigma y_i^2 - n\bar{y}^2]^{1/2}} = .776$ (the sample correlation coefficient).To estimate its standard error, B bootstrap samples of $n = 15$ points selected at random with replacement from the actual sample are performed and the coefficients $\hat{\rho}^*(1), \hat{\rho}^*(2), .., \hat{\rho}^*(B)$ are obtained. Efron and Tibshirani experiment with different values of B. They notice a stabilization of the empirical standard errors of these B bootstrap replications towards 0.132. Closeness is observed already at values of B around 1000. This illustrates the **nonparametric** bootstrap approach.

To perform a **parametric bootstrap** in this example, we assume that the the LSAT and the GPA results do follow a bivariate normal distribution. We estimate this distribution by $\hat{F}_{\text{norm}}$ via substitution of the empirical estimators of the mean vector and of the covariance matrix in the formula for the bivariate normal density. Then B samples of size 15 each from $\hat{F}_{\text{norm}}$ are simulated and the correlation coefficient is computed for each sample. The parametric bootstrap estimate for $B = 3200$ repetitions has been .124, close to the value obtained by the nonparametric bootstrap.

Finally, for this example, there exists a celebrated theoretical result which states that asymptotically, for a sample from a bivariate normal, the standard error of the empirical correlation coefficient is approximated by $(1 - \hat{\rho}^2)/\sqrt{n-3}$. Substituting $\hat{\rho} = .776$ gives .115 in our case. Furthermore, another celebrated result in this direction is **Fisher's** z

transformation (a variance stabilizing transformation (!)) which implies that

$$z = 0.5.\log(\frac{1+\hat{\rho}}{1-\hat{\rho}}) \sim N(0.5\log(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

holds asymptotically. This result is used to make inference (test hypotheses, construct confidence intervals etc.) for ρ by first making inference about $0.5\log(\frac{1+\rho}{1-\rho})$ and then transforming back to inference about ρ. The empirical standard deviation of the 3200 z values obtained from the transformation of the parametric bootstrap's $\hat{\rho}^*$ values has been equal to .290 which is very close to the value $\frac{1}{\sqrt{15-3}} = .289$.

These numerical values are quite convincing about the accuracy of the bootstrap values obtained (where **no textbook results are necessary to be used !**). Also, one more merit of bootstrap should be appreciated. Note that one of the main reasons for making parametric assumptions in traditional statistical inference is to facilitate the derivation of analytically tractable formulae for standard errors. This restricts the applicability (we could only apply the parametric method to cases where indeed its assumptions are believed to hold for the data available (!)) while still possibly leading to quite painful theoretical derivations. **But in bootstrap approach** we **do not need** these formulae, hence we can also **avoid making** restrictive parametric assumptions.

10.5 Bootstrap estimate of bias

The bootstrap was first introduced as a method for evaluating standard errors of estimators. It should be noted, however, that it can easily be adapted to estimate the **bias** of an estimator and therefore can be applied as a bias correction procedure.

To illustrate this, let us assume again that a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is available from an unknown distribution F and $\theta = t(F)$ is a parameter to be estimated by a statistic $\hat{\theta} = s(\mathbf{x})$. We would like to estimate the (unknown) bias $b_F(\hat{\theta}, \theta) = E_F[s(\mathbf{x})] - t(F)$. The bootstrap estimate of the theoretical bias is obtained naturally by plugging-in $\hat{F}$ instead of F in the above bias formula, i.e.: $b_{\hat{F}}(\hat{\theta}^*, \theta) = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})$. Hence, we can get the *bias-adjusted* estimate as:

$$t(\hat{F}) - \{E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})\} = 2t(\hat{F}) - E_{\hat{F}}[s(\mathbf{x}^*)].$$

It is an easy exercise for you to show that if the parameter θ is *the mean :* $\theta = t(F) = \int xdF(x)$ and $\hat{\theta} = \bar{\mathbf{x}}$ then $b_{\hat{F}}(\hat{\theta}, \theta) = 0$ and there is no bias correction necessary. This is but a rather exceptional case. In most other situations a bias will exist and bias correction via $b_{\hat{F}}$ makes sense. This is demonstrated with the next example.

Example: When estimating the variance σ^2 by $s(x) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ the bias is well known to be $-\frac{1}{n}\sigma^2$ and in this case $b_{\hat{F}}(s^*, \sigma^2) = -\frac{1}{n^2}\sum_{i=1}^n (x_i - \bar{x})^2$ (show it (!)). After the bias correction: $s(x) - b_{\hat{F}}(s, \sigma^2) = \frac{n+1}{n^2}\sum_{i=1}^n (x_i - \bar{x})^2$ and you can see that now the bias of the corrected estimator is $-\frac{1}{n^2}\sigma^2$! This is of much smaller order than the bias $-\frac{1}{n}\sigma^2$ before the corrective action thus illustrating the effect of the bias correction.

In the above example, because of its simplicity, we were able to give a **theoretical** bootstrap estimate of the bias. In more involved situations this will not be possible but again simulation can help us to avoid the difficulties! We can do the bias correction without any theoretical derivations by simply:

- generate B independent bootstrap samples of size n;
- evaluate the bootstrap replications $\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), b = 1, 2, .., B$;
- approximate the bootstrap expectation $E_{\hat{F}}[s(\mathbf{x}^*)]$ by $\frac{1}{B}\sum_{b=1}^{B} s(\mathbf{x}^{*b})$;
- get the estimate (approximation) of the bootstrap estimate of bias based on B replications by $\hat{b}_B(\hat{\theta}, \theta) = \frac{1}{B}\sum_{b=1}^{B} s(\mathbf{x}^{*b}) - t(\hat{F})$

Of course, in the last step we could calculate **both** the estimated standard deviation $\hat{s}_B$ **and** $\hat{b}_B(\hat{\theta}, \theta)$ from the same set of bootstrap replications thus being in a position to give a bootstrap estimate of the mean squared error of the estimator,too.

10.6 The jackknife estimate of bias.

In fact , the original method proposed for bias reduction was not the bootstrap but the **jackknife** (M. Quenouille as early as in the mid $1950's$). We shall discuss it briefly here and also its relation to the bootstrap will be pointed out. Given the original sample $\mathbf{x}$, one defines n *jackknife samples* $\mathbf{x}_{(i)} = (x_1, x_2, .., x_{i-1}, x_{i+1}, .., x_n)$ and the ith *jackknife replication* $\hat{\theta}_{(i)}$ of the statistic $\hat{\theta} = s(\mathbf{x})$ is defined as $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$. Then the *jackknife estimate of bias* is defined as $\hat{b}_{\text{jack}} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta})$ where $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$. Note also that the jackknife method has been also extended to estimating the standard error by using the formula

$$\hat{s}_{\text{jack}} = [\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2]^{1/2}$$

Finally , one **warning** should be given: **sometimes** bias correction could be dangerous in practice. It may happen that it reduces the bias at too high a price,namely by increasing the variance quite significantly thus increasing the mean squared error. Fortunately, we have means to check if this unpleasant effect has occurred! Indeed, (again using the bootstrap), we can estimate the mean squared error, as well! Then, bias correction should be only applied when it does not blow up significantly the estimated mean squared error.

10.7 Relation of bootstrap and jackknife methods.

The bootstrap and the jackknife **standard error estimators** look also quite similar except that the factor $(n-1)/n$ in the jackknife estimator formula is much larger than the $1/(n-1)$ factor in the bootstrap estimator formula. This is so since the jackknife deviations $(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2$ tend to be smaller than the bootstrap deviations $[\hat{\theta}^*(b) - \bar{\hat{\theta}}^*]^2$ (intuitively, the jackknife sample is more similar to the original sample than is the bootstrap sample). More refined arguments show that the jackknife can be considered as a **linear approximation to the bootstrap:** i.e. it agrees with bootstrap for a certain *linear statistic in the form* const$+\frac{1}{n}\sum_{i=1}^{n} \alpha(x_i)$ that approximates the possibly nonlinear statistic $\hat{\theta}$. In fact, the accuracy of the jackknife estimate depends essentially on how close is $\hat{\theta}$ to a linear statistic. In terms of our discussion in the robustness lecture, if the remainder in

the approximation of the functional $T(F)$ via $T(F_n)$ "behaves well" then the results when using bootsrap and jackknife for bias correction are very similar.

As an upshot, the jackknife provides a simple approximation to the bootstrap for estimating bias and standard error. However, the jackknife can fail if the statistic is significantly non-linear.

10.8 Confidence intervals based on the bootstrap.

The ultimate goal in evaluating the standard deviation of an estimator is to utilize this standard deviation for constructing *confidence intervals*. Given $\alpha \in (0,1)$ the naive **asymptotic** $(1-\alpha).100\%$ confidence interval would be $[\hat{\theta} - z_{\alpha/2}.s_F(\hat{\theta}), \hat{\theta} + z_{\alpha/2}.s_F(\hat{\theta})]$. Despite its simplicity, this confidence interval does not have much merit. It has only approximate $(1-\alpha).100\%$ coverage since the $(1-\alpha).100\%$ coverage is **only** obtained in a limit. It is possible to derive **better** (still approximate but with better coverage accuracy) bootstrap confidence intervals based on the following observation. The distribution of $Z = \frac{\hat{\theta}-\theta}{s_F(\hat{\theta})}$ would not follow exactly the $N(0,1)$ law **but this very same distribution can be estimated directly from the data at hand.** A corresponding bootstrap table can be constructed by generating B bootstrap samples that give rise to B different Z values (the empirical percentiles can be obtained from the empirical distribution of these Z values). More specifically, we:

- generate B bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, .., \mathbf{x}^{*B}$
- for each sample, we calculate $Z^*(b) = (\hat{\theta}^*(b) - \hat{\theta})/\hat{s}^*(b), b = 1, 2, .., B$
- evaluate the αth percentile by the value $\hat{t}^{(\alpha)}$such that $\#(Z^*(b) \leq \hat{t}^{(\alpha)})/B = \alpha$.
- the **bootstrap t confidence interval** at level $(1-2\alpha)$ is then $(\hat{\theta} - \hat{t}^{(1-\alpha)}s_F(\hat{\theta}), \hat{\theta} - \hat{t}^{(\alpha)}s_F(\hat{\theta}))$(or with the empirical value $s_B(\hat{\theta})$ substituted if $s_F(\hat{\theta})$ or $s_{\hat{F}}(\hat{\theta}^*)$ unknown).

A subtle theoretical investigation of the procedure shows that in large samples the coverage of the bootstrap t based confidence interval is closer to the desired level than the coverage of the standard normality based interval or the interval based on the t-distribution table. The price of this accuracy is that generality is lost- the bootstrap t table can be applied **only for the data at hand.** The table will need to be generated from scratch for a new set of observations / new problem. Note also that the bootstrap t percentiles **are not necessarily symmetric around zero** and the resulting confidence intervals are not necessarily symmetric around $\hat{\theta}$!

The following problems are encountered when applying the bootstrap t and should be mentioned here:

i) for more complicated statistics than the arithmetic mean, we need to estimate $\hat{s}^*(b)$ and this needs in fact a **second nested level of bootstrapping.** This might increase the computational costs **dramatically.**

ii) for small samples, the bootstrap t interval may perform erratically. It may be preferable to work with the **bootstrap distribution of** $\hat{\theta}$ itself instead of working with

the bootstrap distribution of Z. This leads to the so called "BC_a" procedure. Its description is outside the scope of this course, however.

10.9 Extensions of bootstrap outside the i.i.d. setting

Many practically important models are **not** based on the assumption of availability of i.i.d. observations. These models are exactly the more complicated ones for which no textbook solutions for standard errors of estimators etc. exist. Hence extensions of the bootstrap method beyond the i.i.d. setting become of crucial importance. We do not have time to discuss these extensions here but will mention that many such extensions exist and cover situations such as:

- regression (model-based bootstrap)
- autoregressive type time series (block bootstrap)
- other weakly dependent series (sieve bootstrap)
- bootstrap in the frequency domain
- bootstrap methods tailored specifically for Markov Processes
- bootstrap for long range dependent data
- bootstrap for spatial data

These more involved extensions of the bootstrap may not yet exist in traditional packages such as SAS of SPLUS but have realisations in packages such as *R*.

Example related to LMP tests:

$X = (X_1, X_2, \dots, X_n)$ i.i.d. Cauchy$(\theta, 1)$ $H_0: \theta \le 0$ vs $H_1: \theta > 0$

$$L(X, \theta) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{1+(x_i-\theta)^2}$$

$V(X, \theta)\big|_{\theta=0}$ is proportional to $\sum_{i=1}^{n} \frac{2(X_i-\theta)}{1+(x_i-\theta)^2}\Big|_{\theta=0} = \sum_{i=1}^{n} \frac{2X_i}{1+X_i^2}$

Hence LMP test:

$$\varphi^* = \begin{cases} 1 & T = \sum_{i=1}^{n} \frac{2X_i}{1+X_i^2} > K \\ 0 & \text{—"—} \le K \end{cases}$$

How to choose K:

If $U = \frac{2x}{1+x^2}$, $x \sim C(0,1) \Rightarrow EU = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{2x\,dx}{(1+x^2)^2} = 0$

$$\operatorname{Var} U = EU^2 = \frac{4}{\pi} \int_{-\infty}^{+\infty} \frac{x^2 dx}{(1+x^2)^3} = \frac{2}{\pi} \int_{-\infty}^{+\infty} \frac{x\, d(1+x^2)}{(1+x^2)^3} =$$

$$= -\frac{1}{\pi} \int_{-\infty}^{+\infty} x\, d \frac{1}{(1+x^2)^2} = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{1+x^2}{(1+x^2)^2} dx = \dots = \frac{1}{2}$$

Hence $U_i = \frac{2X_i}{1+X_i^2}$ have mean 0 and $\operatorname{Var} U_i = \frac{1}{2} < \infty$

Hence $\frac{\sqrt{2}\,T}{\sqrt{n}} \xrightarrow{L} N(0,1)$ (CLT)

$$T > K \Leftrightarrow \frac{\sqrt{2}\,T}{\sqrt{n}} > \frac{\sqrt{2}\,K}{\sqrt{n}} \approx Z_\alpha \longrightarrow K = \frac{Z_\alpha \cdot \sqrt{n}}{\sqrt{2}}$$

Solutions to MST-2018- Inference

1)

a) First, we take an unbiased estimator of $\tau(\theta)$ as,

e.g. $W = I_{\{X_1=1\}}(\mathbf{X}) = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{if } X_1 \neq 1 \end{cases}$

$E_\theta W = 1 * P(X_1 = 1) = \theta e^{-\theta} = \tau(\theta)$ obviously holds.

The Lehmann-Scheffe theorem then tells us that $\hat{\tau} = E(W|T=t)$ is the UMVUE of $\tau(\theta)$.

Now $\hat{\tau} = 1 * P(W=1|T=t) = \frac{P(W=1 \cap T=t)}{P(T=t)}$ and we know that $\sum_{i=1}^{n} X_i \sim Po(n\theta)$. Hence we have

$$\hat{\tau} = \frac{P(X_1=1 \cap \sum_{i=2}^{n} X_i = t-1)}{P(\sum_{i=1}^{n} X_i = t)} = \frac{\theta e^{-\theta} \, e^{-(n-1)\theta} ((n-1)\theta)^{t-1}}{(t-1)! \, e^{-n\theta} n^t \theta^t / t!}$$

$$= \frac{t}{n}\left(\frac{n-1}{n}\right)^{t-1} = \bar{X}\left(1-\frac{1}{n}\right)^{n\bar{X}-1}$$

Numerically: $\sum_{i=1}^{8} x_i = 14 \quad n = 8 \rightarrow 1.75\left(\frac{7}{8}\right)^{13} = 0.3084$

b) CR Bound is $\frac{(\tau'(\theta))^2}{I_{\mathbf{X}}(\theta)}$. Now $(\tau'(\theta))^2 = (\theta-1)^2 e^{-2\theta}$

For $I_{\mathbf{X}}(\theta)$: $L(\mathbf{X}, \theta) = \frac{e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$

$\ln L(\mathbf{X}, \theta) = -n\theta + \sum_{i=1}^{n} x_i \ln\theta + \text{const}$

$\frac{\partial}{\partial\theta} \ln L(\mathbf{X}, \theta) = V(\mathbf{X}, \theta) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$

$-\frac{\partial^2}{\partial\theta^2} \ln L(\mathbf{X}, \theta) = \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \Rightarrow E\left[-\frac{\partial^2}{\partial\theta^2} \ln L(\mathbf{X}, \theta)\right] = \frac{n}{\theta}$

Hence CR Bound $= \frac{\theta(\theta-1)^2 e^{-2\theta}}{n}$

Bound is not attainable since the score $V(\mathbf{X}, \theta) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$ cannot be factorized into $K(n, \theta)(\text{statistic} - \tau(\theta))$:

We have $V(\mathbf{X}, \theta) = \frac{ne^{\theta}}{\theta}(\bar{x}e^{-\theta} - \tau(\theta))$

c) From $V(X,\theta)=0$ we get $\hat{\theta}_{MLE}=\bar{X}$

Hence $\widehat{\tau(\theta)}_{MLE}=\bar{X}e^{-\bar{X}}$ Numerically this gives

$\widehat{\tau(\theta)}_{MLE}=0.3041$ → close to the UMVUE

d) $f(X|\theta)\,\text{prior}(\theta) \propto \theta^{\sum_{i=1}^{n}x_i+1}e^{-(2+n)\theta}$

where $\text{prior}(\theta)\propto\theta^{2-1}e^{-2\theta}$

This reveals the posterior as $\text{Gamma}\left(\sum_{i=1}^{n}x_i+2,\ \frac{1}{2+n}\right)$

Hence the Bayes estimator is the mean of this posterior: $\hat{\theta}_{Bayes}=\frac{\sum_{i=1}^{n}x_i+2}{2+n}$. Numerically $\hat{\theta}_{Bayes}=1.6$

e) $P(H_0|X)=\frac{10^{16}}{\Gamma(16)}\int_0^{1.5}\exp(-10x)x^{15}dx=7647.164\times 0.000056$

$=0.4282<\frac{1}{2}$

Hence we reject H_0.

2) a) cdf of a single observation is $F_{X_1}(x)=\begin{cases}0 & x<0\\ \left(\frac{x}{\theta}\right)^2 & 0<x<\theta\\ 1 & x>\theta\end{cases}$

Then $F_{Z_n}(z)=P(X_{(n)}<z)=P(X_1<z\cap X_2<z\cap\dots\cap X_n<z)=$

$=\begin{cases}0 & z<0\\ \left(\frac{z}{\theta}\right)^{2n} & 0<z<\theta\\ 1 & z>\theta\end{cases}$

Hence $f_{Z_n}(z)=\begin{cases}\frac{2nz^{2n-1}}{\theta^{2n}} & 0<z<\theta\\ 0 & \text{else}\end{cases}$

b) - Sufficiency: write $f(x;\theta)=\frac{2x}{\theta^2}I_{(0,\theta)}(x)$. Then $L(X;\theta)=2^n\frac{\prod_{i=1}^{n}x_i}{\theta^{2n}}I_{(0,\theta)}(X_i)$

$=\frac{2^n}{\theta^{2n}}\prod_{i=1}^{n}x_i\,I_{(0,\theta)}(X_{(n)})$ which can be factorized in

$g(\theta,X_{(n)})\cdot h(X)$ with $g(\theta,X_{(n)})=\frac{1}{\theta^{2n}}I_{(0,\theta)}(X_{(n)})$, $h(X)=2^n\prod_{i=1}^{n}x_i$

Hence $X_{(n)}$ is sufficient.

- Completeness: take $g(\cdot)$ with $E_\theta g(X_{(n)}) = 0 \ \forall \theta$

This implies $\frac{2n}{\theta^{2n}} \int_0^\theta g(t) t^{2n-1} dt = 0 \ \forall \theta$ and since $\frac{2n}{\theta^{2n}} \neq 0 \ \forall \theta > 0$

we have $\int_0^\theta g(t) t^{2n-1} dt = 0 \ \forall \theta > 0 \Rightarrow$ Take derivative w.r.$\theta \Rightarrow$ $g(\theta)\theta^{2n-1} = 0 \ \forall \theta > 0 \Rightarrow g(\theta) = 0 \ \forall \theta > 0$

i.e. $P_\theta(g(t) = 0) = 1$ which implies completeness.

c) We first calculate $E Z_n = \int_0^\theta x \frac{2n x^{2n-1}}{\theta^{2n}} dx = \frac{2n}{2n+1}\theta \neq \theta$

i.e. Z_n is biased for θ but $\boxed{\frac{2n+1}{2n} Z_n}$ is unbiased and is a function of complete and sufficient statistic.

Hence $E\left(\frac{2n+1}{2n} Z_n \,/\, Z_n\right) = \boxed{\frac{2n+1}{2n} Z_n}$ is UMVUE

(this is Lehmann-Scheffé's theorem)

Total Marks (36) = 24 + 12
Q1 Q2

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

MID SESSION TEST - 2017 -Tuesday, 5th September (Week 7)

MATH5905

Time allowed: 75 minutes

1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. Bernoulli (θ) random variables (that is,

$$f(x, \theta) = \theta^x (1-\theta)^{1-x}, x = \{0, 1\}; \theta \in (0, 1)).$$

   a) Justify that that $T = \sum_{i=1}^n X_i$ is sufficient and complete for θ.

   b) Derive the UMVUE of $h(\theta) = \theta^2$. Justify each step in your answer.

   c) Calculate the Cramer-Rao bound for the minimal variance of an unbiased estimator of $h(\theta) = \theta^2$. Does the variance of the UMVUE of $h(\theta)$ attain this bound? Give reasons.

   d) Find the MLE $\hat{h}$ of $h(\theta)$. Justify your answer.

   e) When testing $H_0 : \theta \leq 0.6$ versus $H_1 : \theta > 0.6$ with a 0-1 loss in Bayesian setting with the prior $\tau(\theta) = 12\theta^2(1-\theta)$, what is your decision when $T = 5$ and $n = 8$. (You may use: $\int_0^{0.6} x^7(1-x)^4 dx = 0.00011$.)
   Note: The continuous random variable X has a beta density f with parameters $\alpha > 0$ and $\beta > 0$ if $f(x; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, x \in (0, 1)$. It holds: $E(X) = \frac{\alpha}{\alpha+\beta}$. Here

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx, B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \Gamma(\alpha+1) = \alpha\Gamma(\alpha).$$

2. Let $X_1, X_2, \ldots, X_n$ be independent random variables, with a density

$$f(x; \theta) = \begin{cases} \frac{\theta}{x^2}, x > \theta, \\ 0 \text{ else} \end{cases}$$

   where $\theta > 0$ is an unknown parameter. If $Z_n = X_{(1)}$, (i.e., the minimal of the n observations) then

   a) Argue that Z_n is a sufficient statistic for θ.

   b) Show that the density of Z_n is

$$f_{Z_n}(z) = \begin{cases} \frac{n\theta^n}{z^{n+1}}, z > \theta, \\ 0 \text{ else} \end{cases}$$

   (**Hint:** You may find the cdf first by using $P(X_{(1)} > x) = P(X_1 > x \cap X_2 > x \ldots \cap X_n > x)$.)

   c) Find the MLE of θ. Justify your answer.

   d) (*) Given that $X_{(1)}$ is complete for θ, find the UMVUE of θ.

Solution to Question 1

Part a). 4 marks

Approach 1: Using property of one parameter exponential family (see p25 of lecture notes).

Indeed, we observe that,

$$\begin{aligned} f(x,\theta) &= \theta^x(1-\theta)^{1-x} \\ &= (1-\theta)\exp\left(x\log\left(\frac{\theta}{1-\theta}\right)\right). \end{aligned}$$

Thus, Bernoulli belongs to one parameter exponential family. This then implies $T = \sum_{i=1}^n X_i$ is (minimal) sufficient, and complete.

Marking criteria:

- 2 marks for notifying that binomial belongs to a one parameter exponential family.
- 1 mark for making the conclusion that the test statistic is sufficient.
- 1 mark for notifying that the test statistic is complete.

Approach 2: Lehmann and Scheffe's method (see p23 of lecture notes) and definition of completeness (see p38 of lecture notes).

We calculate the proportion of the joint density

$$\begin{aligned} \frac{L(X,\theta)}{L(Y,\theta)} &= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}} \\ &= \theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}(1-\theta)^{\sum_{i=1}^n y_i - \sum_{i=1}^n x_i}. \end{aligned}$$

This is independent of θ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Thus, $T = \sum_{i=1}^n X_i$ is (minimal) sufficient.

To show it is also complete, we see that

$$\begin{aligned} \mathbb{E}_\theta(g(T)) &= \sum_{k=0}^n \binom{n}{k}\theta^k(1-\theta)^{n-k}g(k) \\ &= (1-\theta)^n \sum_{k=0}^n \binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k g(k) \\ &= (1-\theta)^n \sum_{k=0}^n \binom{n}{k}\eta^k g(k) = 0, \end{aligned}$$

where $\eta = \frac{\theta}{1-\theta}$, holds for all $\theta \in (0,1)$, and all $\eta \in (0,\infty)$ only if implies $g(k) = 0$ for all $k = 0, ..., n$. Notice that $\binom{n}{k} = \frac{n!}{k!(n-k)!} > 0$, $1-\theta > 0$, and η^k is a polynomial. As a consequence, we have $\mathbb{P}(g(T) = 0) = 1$.

Marking criteria:

- 1 mark for calculating the proportion.
- 1 mark for proving the sufficiency.
- 1 mark for using the definition of completeness.

- 1 mark for proving the test statistic is complete.

Approach 3: First Investigation (see p21 of lecture notes)

By definition of sufficient statistics, it is enough to show that

$$\mathbb{P}\Big(X_1 = x_1, ..., X_n = x_n \Big| \sum_{i=1}^{n} X_i = k\Big) \quad = \quad \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \frac{1}{\binom{n}{k}}.$$

is independent of θ, which is obvious. For completeness, see Approach 2.

Marking criteria:

- 1 mark for using the definition of sufficiency.
- 1 mark for proving the test statistic is sufficiency.
- 1 mark for using the definition of completeness.
- 1 mark for proving the test statistic is complete.

Part b). 4 marks

Since T is complete and sufficient, we need to find an unbiased estimator of τ and apply Theorem of Lehmann-Scheffe (see p38 lecture notes). Let

$$\tau \quad = \quad X_1 X_2, \tag{1}$$

we see that

$$\mathbb{E}(\tau) \quad = \quad \Big(\mathbb{E}(X_1)\Big)^2 = \theta^2, \tag{2}$$

which is unbiased. Now, we apply Theorem of Lehmann-Scheffe, this then yields

$$\begin{aligned}
a(k) &= \mathbb{E}\Big(X_1X_2 \Big| \sum_{i=1}^{n} X_i = k\Big) \\
&= \mathbb{P}\Big(X_1 = 1, X_2 = 1 \Big| \sum_{i=1}^{n} X_i = k\Big) \\
&= \frac{\mathbb{P}\Big(X_1 = 1, X_2 = 1, \sum_{i=1}^{n} X_i = k\Big)}{\mathbb{P}\Big(\sum_{i=1}^{n} X_i = k\Big)} \\
&= \frac{\mathbb{P}\Big(\sum_{i=3}^{n} X_i = k-2\Big)\theta^2}{\mathbb{P}\Big(\sum_{i=1}^{n} X_i = k\Big)} \\
&= \frac{\binom{n-2}{k-2}\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} \\
&= \frac{k(k-1)}{n(n-1)} \\
&= \bar{X}(\bar{X} - \frac{1}{n})\frac{n}{n-1}.
\end{aligned}$$

Marking criteria:

- 1 mark for selecting an estimator.

- 1 mark for proving the estimator is unbiased.
- 1 mark for applying Theorem of Lehmann-Scheffe.
- 1 mark for getting the final UMVUE.

Part c). 4 marks

we can calculate the Cramer-Rao bound as:

$$\begin{aligned} Var_\theta(a(k)) &\geq \frac{\left(\frac{\partial}{\partial\theta}\theta^2\right)^2}{-\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(X,\theta)\right)} \\ &\geq \frac{4\theta^2}{-\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}\right)} \\ &\geq \frac{4\theta^2}{-\mathbb{E}_\theta\left(\frac{\sum_{i=1}^n x_i}{\theta^2}+\frac{n-\sum_{i=1}^n x_i}{(1-\theta)^2}\right)} \\ &\geq 4\theta^2\frac{\theta(1-\theta)}{n}. \end{aligned}$$

Next, we see that

$$V(X,\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n-\sum_{i=1}^n x_i}{(1-\theta)} = \frac{n}{\theta^2(1-\theta)}\left(\bar{X}\theta-\theta^2\right).$$

and, $\bar{X}\theta$ is not a statistics so the bound is not attainable.

Marking criteria:

- 2 marks for calculating the Cramer-Rao bound.
- 2 marks for make the right conclusion.

Part d). 2 marks

We first calculate the log-likelihood:

$$\begin{aligned} \log\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i} &= \log\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \sum_{i=1}^n x_i\log\theta + (n-\sum_{i=1}^n x_i)\log(1-\theta). \end{aligned}$$

Differentiating with respect to θ and set this to zero yields:

$$\sum_{i=1}^n x_i\frac{1}{\theta} - (n-\sum_{i=1}^n x_i)\frac{1}{(1-\theta)} = 0.$$

Thus, we have the MLE $\hat{\theta}_{MLE} = \bar{X}$. By the invariance property,

$$\hat{h} = \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 = (\bar{X})^2.$$

Marking criteria:

- 1 mark for getting the MLE for θ.
- 1 mark for using the invariance property.

Part e). 3 marks

We note that, from p17 of lecture notes,

$$\begin{aligned} P(\theta \leq 0.6) &= \int_0^{0.6} h(\theta|X)d\theta \\ &= \int_0^{0.6} \frac{f(X|\theta)\tau(\theta)}{\int_0^1 f(X|\theta)\tau(\theta)d\theta}d\theta \\ &= \int_0^{0.6} \frac{\theta^7(1-\theta)^4}{B(8,5)}d\theta \\ &= 0.4356 < 0.5 \end{aligned} \tag{3}$$

(note the threshold for 0-1 loss is 0.5). Thus, we reject the H_0.

Marking criteria:

- 1 mark for getting the expression for $P(\theta \leq 0.6)$.
- 1 mark for getting the correct probability.
- 1 mark for making the correct conclusion.

Solution to Question 2

Part a). 3 marks

We first calculate the joint density

$$L(X,\theta) = \prod_{i=1}^n f(x_i,\theta) = \frac{\theta^n}{\prod_{i=1}^n x_i^2} I_{(\theta,\infty)}(X_{(1)}).$$

Thus, Z_n is sufficient by the Neyman Fisher Factorization Criterion (see p22 of lecture notes).

Marking criteria:

- 1 mark for writing the joint.
- 1 mark for making a correct calculation.
- 1 mark for noting the Neyman Fisher Factorization Criterion.

Part b). 3 marks

We first calculating the survival function.

$$\begin{aligned} 1 - F_{Z_n}(z) &= \mathbb{P}(X_{(1)} > z) \\ &= \mathbb{P}(X_1 > z, ..., X_n > z) \\ &= \prod_{i=1}^n \mathbb{P}(X_i > z) \\ &= \frac{\theta^n}{z^n}. \end{aligned}$$

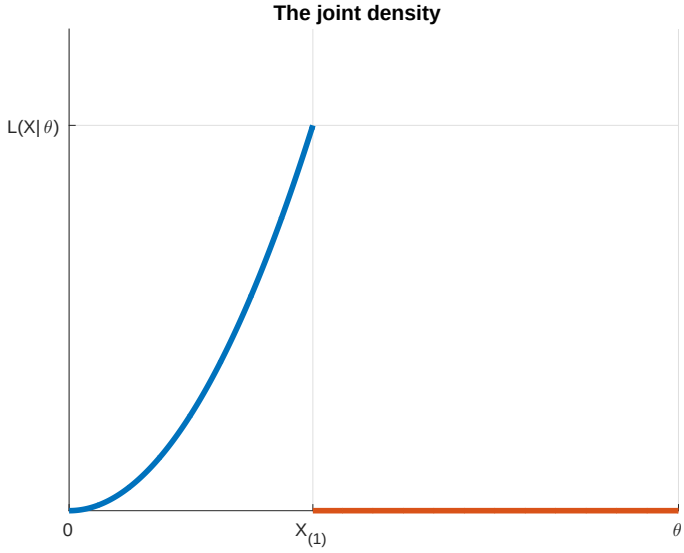


Figure 1: Graph of Joint Density

Then, we obtain the density

$$f_{Z_n}(z) \quad = \quad \frac{n\theta^n}{z^{n+1}} I_{(\theta,\infty)}(z).$$

Marking criteria:

- 2 marks for calculating the survival function (or the cdf).
- 1 mark for obtaining the density via differentiation.

Part c). 3 marks

We first calculate the joint:

$$L(X,\theta) \quad = \quad \frac{\theta^n}{\prod_{i=1}^n x_i^2} I_{(\theta,\infty)}(X_{(1)}),$$

Thus, the likelihood will be maximized at $\theta = \min(x_1, ..., x_n) = X_{(1)}$. See also Figure 1.

Marking criteria:

- 1 mark for calculating the joint.
- 1 mark for giving the correct mle.
- 1 mark for giving the correct reasoning.

Part d). 2 marks

We first check if this is a unbiased estimator.

$$\mathbb{E}(X_{(1)}) = \int_\theta^\infty z \frac{n\theta^n}{z^{n+1}} dz = \frac{n}{n-1}\theta.$$

The UMVUE is then given by

$$\mathbb{E}(\frac{n-1}{n} X_{(1)} | X_{(1)}) = \frac{n-1}{n} X_{(1)}.$$

Marking criteria:

- 1 mark for checking that $X_{(1)}$ is not unbiased and needs a bias-correction.
- 1 mark for finding the UMVUE.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

MID SESSION TEST - 2018 -Friday, 7th September (Week 7)

MATH5905

Time allowed: 75 minutes

1. In a sequence of consecutive years $1, 2, \ldots, n$, an annual number of high-risk events is recorded by a bank. The random counts $X_i, i = 1, 2, \ldots, n$ of high-risk events in a given year is modelled via Poisson(θ) distribution and can be assumed independent from year to year. Within the last eight years counts were $0, 3, 1, 1, 2, 2, 4, 1$.

   a) Given that $T = \sum_{i=1}^{n} X_i$ is sufficient and complete for θ, derive the UMVUE of $\tau(\theta) = \theta e^{-\theta}$, i.e., the probability that exactly one extremal event in a given year will emerge. Justify your answer and evaluate the probability using the given data.

   b) Calculate the Cramer-Rao bound for the minimal variance of an unbiased estimator of $\tau(\theta) = \theta e^{-\theta}$. Does the variance of the UMVUE of $\tau(\theta)$ attain this bound? Give reasons.

   c) Find the MLE $\hat{\tau}$ of $\tau(\theta)$. Compare the numerical values in a) and c) and comment.

   d) The prior on θ is Gamma$(2, 0.5)$. Determine the Bayesian estimator of θ w.r.t. quadratic loss.
   Note: You may use that for known $\alpha > 0, \beta > 0$, the $Gamma(\alpha, \beta)$ density is given by:

   $$f(x; \alpha, \beta) = \frac{e^{-\frac{x}{\beta}} x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}, x > 0.$$

   Here $\Gamma(\alpha) = \int_0^{\infty} e^{-x} x^{\alpha-1} dx$ is the gamma function. If X is distributed $Gamma(\alpha, \beta)$ then $EX = \alpha\beta, V(X) = \alpha\beta^2$ holds. $\diamond$

   e) The bank claims that the intensity θ is less than 1.5. Test the bank's claim via Bayesian testing with a zero-one loss. You can use: $10^{16}/\Gamma(16) = 7647.164, \int_0^{1.5} exp(-10x) * x^{15} dx = 0.000056$.

2. Let $X_1, X_2, \ldots, X_n$ be independent random variables, with a density

   $$f(x; \theta) = \begin{cases} \frac{2x}{\theta^2}, 0 < x < \theta, \\ 0 \text{ else} \end{cases}$$

   where $\theta > 0$ is an unknown parameter. If $Z_n = X_{(n)}$, then

   a) Show that the density of Z_n is

   $$f_{Z_n}(z; \theta) = \begin{cases} \frac{2nz^{2n-1}}{\theta^{2n}}, 0 < z < \theta, \\ 0 \text{ else} \end{cases}$$

   (**Hint:** find the cdf $F_{Z_n}(z; \theta)$ of Z_n first).

   b) Argue that Z_n is a sufficient and complete statistic for θ.

   c) Find the UMVUE of the parameter θ as a function of Z_n.

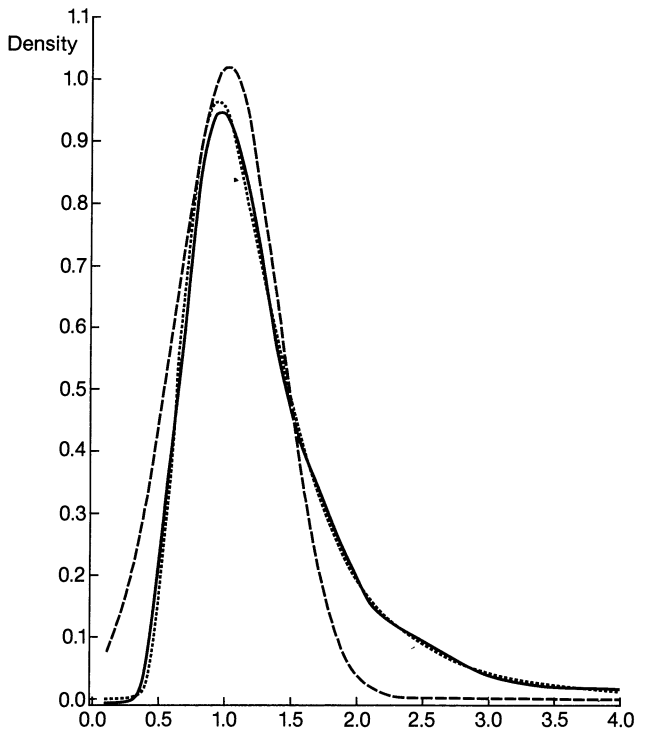


FIG. 1. *Saddlepoint approximation (dotted line) to the density of the maximum likelihood estimator of the shape parameter of a gamma distribution, based on a sample of size* 10. *The "exact" density (solid line) was estimated from* 10,000 *simulations. The normal approximation (dashed line) is shown for comparison.*

To construct a confidence interval for ν it is necessary either to carry out repeated numerical integration or to expand (13) and invert it algebraically. For the general one-parameter case, the details of the expansion are carried out in Barndorff-Nielsen (1985b) and McCullagh (1984a). Unfortunately, the relatively simple approximations to tail probabilities discussed in Daniels (1987) (cf. also Section 6.3) cannot be directly applied to this example because $\hat{\theta}$ is not a one-to-one function of a sample average. It may be possible to adapt the conditional probability tail approximation of Skovgaard (1988b) to this example.

Outside the exponential family setting, the maximum likelihood estimator will not be a one-to-one function of the minimal sufficient statistic, so even if we contemplated using the righthand side of (13) we would not be able to write $L(\theta; x)$ for example, as a function only of θ and $\hat{\theta}$. However, (13) does continue to provide an approximation to a conditional density of $\hat{\theta}$, as is illustrated in the next example.

Example 2. Location-scale family. Suppose $f_X(x; \theta)$ is an arbitrary continuous density on R^1, with θ as a two-dimensional location-scale parameter (μ, σ), so that for an independent, identically distributed sample,

$$f(x_1, \cdots, x_n; \mu, \sigma) = \prod_{i=1}^{n} \sigma^{-1} f_X\{(x_i - \mu)/\sigma\}.$$

Without further assumptions about f, the minimal sufficient statistic is the order statistic $(x_{(1)}, \cdots, x_{(n)})$. It can be separated into two components, $\hat{\theta}$, the maximum likelihood estimate of θ, and $a = (a_1, \cdots, a_n)$, where $a_i = (x_{(i)} - \hat{\mu})/\hat{\sigma}$. The vector a has $n - 2$ independent components, and is ancillary; i.e., its distribution does not depend on θ. The conditional distribution of $\hat{\theta}$, given a, is

$$f_{\hat{\Theta}|A}(\hat{\theta} \mid a; \theta) = c_0(a)\hat{\sigma}^{n-2}\Pi f_X(\hat{\sigma}a_i + \hat{\mu}; \mu, \sigma)$$

which can be re-expressed as

$$(14) \qquad f(\hat{\theta} \mid a; \theta) = c(a)\,|j(\hat{\theta})|^{1/2}\{L(\theta)/L(\hat{\theta})\},$$

using the fact that $|j(\hat{\mu}, \hat{\sigma})| = \hat{\sigma}^{-4}d(a)$, where $d(\cdot)$ depends on the derivatives of $\log f$.

Note the similarity of (14) to approximation (13), and also that (14) is the exact conditional density of $\hat{\theta}$, given the maximal ancillary a. Fisher (1934) derived (14) and argued that inference for θ should be based on this conditional distribution; see also Cox and Hinkley (1974, page 115). Different versions of formula (14) have been derived by several authors, including Pitman (1938), Fraser (1968), Efron and Hinkley (1978), and Barndorff-Nielsen (1980, 1983). Barndorff-Nielsen (1983) emphasized the similarity of (13) and (14), and showed further that (14) provides an expression for the conditional density of the maximum likelihood estimate in any transformation model, i.e., any model generated by a group.

That the same formula provides either a highly accurate approximation or an exact expression for the distribution of the maximum likelihood estimator in full exponential families or transformation families is rather surprising. Exponential families and transformation families are usually considered to be quite different statistical objects, but this suggests that there may be a close connection between them. McCullagh (1987, Chapter 8) has investigated to what extent an arbitrary family of densities can be made to "look like" an exponential family, by conditioning on some approximately distribution-free statistic. Also relevant is Mitchell (1988) in which the geometry of a subclass of transformation models, the elliptic families, is studied. This geometry has some striking similarities to the geometry of exponential families outlined in Amari (1982; 1985, Chapter 2) and Efron (1978).

What about densities that are not members of exponential or transformation families? Remarkably, the same formula continues to provide an approximation to the conditional distribution of the maximum likelihood estimate, conditioned on an approximately ancillary statistic a. Approximately ancillary is taken to mean that the distribution of a depends on θ only in terms of $O(n^{-1})$ or higher, for θ

UNIVERSITY OF NEW SOUTH WALES • SCHOOL OF MATHEMATICS AND STATISTICS

Statistical Tables

t distribution critical values

Key: Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
	.50	.60	.70	0.80	.90	.95	.96	.98	.99	.995	.998	.999
	Probability C											

Standard normal probabilities

Key: Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

χ^2 distribution critical values

Key: Table entry for p is the critical value with probability p lying to its right.

	Upper tail probability p									
df	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000039	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17

F distribution critical values

Key: p=Upper tail probability p, df_n=degrees of freedom in numerator, df_d=degrees of freedom in denominator, $*$ Multiply by 10, † Multiply by 100.

	df_n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_d	p																			
1	.05	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
	.025	648	800	864	900	922	937	948	957	963	969	977	986	993	997	1001	1006	1010	1014	1018
	.01	405*	500*	540*	563*	576*	586*	593*	598*	602*	606*	611*	616*	621*	624*	626*	629*	631*	634*	637*
	.005	162†	200†	216†	225†	231†	234†	237†	239†	241†	242†	244†	246†	248†	249†	250†	251†	253†	254†	255†
2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	.005	199	199	199	199	199	199	199	199	199	199	199	199	199	200	200	200	200	200	200
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.08
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
24	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
30	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Additional Exercises for MATH5905, Statistical Inference

Part one: Decision theory. Bayes and minimax rules

1. Suppose d_1, d_2, d_3 and d_4 are nonrandomized decision rules with risks as given in the following table:

i	1	2	3	4
$R(\theta_1, d_i)$	0	1	2	3
$R(\theta_2, d_i)$	6	5	3	5

   a) Find the minimax rule(s) amongst the **nonrandomized** rules $D = \{d_1, d_2, d_3, d_4\}$;

   b) Obtain the minimax rule in the set of randomized rules $\mathcal{D}$ generated by the set of rules in D. State the minimax risk of this rule.

   c) Find the Bayes rule and the Bayes risk for the prior $(\frac{1}{3}, \frac{2}{3})$ on (θ_1, θ_2).

   d) Express the randomized decision rule with risk point (2,5) using the given non-randomized decision rules.

   e) Calculate all priors for which d_1 is a Bayes rule.

2. A decision rule d is called admissible in a class of rules if there is no other decision rule d^* in the class such that $R(\theta, d^*) \leq R(\theta, d)$ for all $\theta \in \Theta$ and $R(\theta, d^*) < R(\theta, d)$ for at least one value of $\theta \in \Theta$. Let X be uniformly distributed on $[0, \theta]$ where $\theta \in (0, \infty)$ is an unknown parameter (i.e., $\Theta = [0, \infty)$). Let the action space be $[0, \infty)$ and the loss function $L(\theta, a) = (\theta - a)^2$ where a is the chosen action (the action now is estimation so $a = d(X)$ for given observation X and decision d). Consider the set of decision rules $d_\mu(x) = \mu x, \mu \geq 0$. For what value of μ is d_μ unbiased? Show that $\mu = 3/2$ is necessary condition for d_μ to be admissible.

3. Suppose $X_1, X_2, \ldots, X_n$ have conditional joint density

$$f_{X_1, X_2, \ldots, X_n|\Theta}(x_1, x_2, \ldots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}, x_i > 0 \text{ for } i = 1, \ldots, n; \theta > 0$$

   and a prior density is given by $\tau(\theta) = ke^{-k\theta}, \theta > 0$, where k is a known constant.

   i) Calculate the posterior density of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

   ii) Find the Bayesian estimator of θ with respect to squared error loss.

4. Suppose a **single** observation x is available from the uniform distribution with a density $f(x|\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta), \theta > 0$. The prior on θ is with a density $\tau(\theta) = \theta \exp(-\theta), \theta > 0$. Find the Bayes estimator of θ:

   i) with respect to quadratic loss;

   ii) with respect to absolute value loss $L(\theta, a) = |\theta - a|$.

   iii)(*) with respect to the loss $L_\eta(\theta, a) = (\theta - a)(\eta - I(\theta - a < 0))$ where $\eta \in (0, 1)$ is a fixed weight.

5. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal density with mean μ and variance 1. Consider estimating μ with a squared-error loss. Assume that the prior $\tau(\mu)$ is a normal density with mean μ_0 and variance 1. Show that the Bayes estimator of μ is $\frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}$.

6. As part of a quality inspection program, five components are selected at random from a batch of components to be tested. From past experience, the parameter θ (the probability of failure), has a beta distribution with density

$$\tau(\theta) = 30\theta(1 - \theta)^4, 0 \leq \theta \leq 1.$$

We wish to test the hypothesis $H_0 : \theta \leq 0.2$ against $H_1 : \theta > 0.2$ using Bayesian hypothesis testing with a 0-1 loss. What is your decision if:

i) In a batch of five, no failures were found

ii) In a batch of five, one failure was found.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Additional Exercises for MATH5905, Statistical Inference

Part two: Data reduction. Sufficient statistics. Classical estimation

Sufficiency

1. Use the factorization criterion to find a sufficient statistic for the parameter when $X_1, X_2, \ldots, X_n$ are independent random variables each with distribution

   a) $N(\mu, 1)$,

   b) $N(0, \sigma^2)$,

   c) Uniform $(\theta, \theta + 1)$,

   d) Poisson (λ).

   Check your answer in (d) using the definition of sufficiency.

2. Let X_1, X_2, X_3 be a sample of size 3 from the Bernoulli (p) distribution. Consider the 2 statistics $S = X_1 + X_2 + X_3$ and $T = X_1X_2 + X_3$. Show that S is sufficient for p and T is not.

3. A random variable $X = (X_1, X_2)$ has the following distribution (with $1 < \theta < 3$):

(x_1, x_2)	(0,0)	(0,1)	(1,0)	(1,1)
$P(X_1 = x_1,\ X_2 = x_2)$	$\frac{1}{12}(12 - 7\theta + \theta^2)$	$\frac{\theta}{12}(4 - \theta)$	$\frac{\theta}{12}(4 - \theta)$	$\frac{\theta}{12}(\theta - 1)$

   Check whether $X_1 + X_2$ or X_1X_2 is sufficient for θ.

4. If $X_1, X_2, \ldots, X_n$ are independent Bernoulli (p) random variables, prove that X_1 is not sufficient for p.

5. Given that θ is an integer and that X_1 and X_2 are independent random variables which are Uniformly distributed on the integers $1, 2, \ldots, \theta$, prove that $X_1 + X_2$ is not sufficient for θ.

6. Suppose $X_1, X_2, \ldots, X_n$ are independent discrete random variables each with probability function $f(x; \theta), \theta$ unknown. Prove that $(X_1, X_2, \ldots, X_{n-1})$ is not sufficient for θ.

7. Find a minimal sufficient statistic for the parameter when $X_1, X_2, \ldots, X_n$ are independent random variables each with distribution

   a) Poisson (λ),

   b) $N(0, \sigma^2)$,

   c) Gamma (α), (With a density $f(x, \alpha) = \frac{1}{\Gamma(\alpha)} \exp(-x)x^{\alpha-1}, x > 0$. (Here the Gamma function is defined as $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$ and has the property $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. In particular, for a natural number $n : \Gamma(n + 1) = n!$ holds).

   d) Uniform $(0, \theta)$.

   e) Uniform $(\theta, \theta + 1)$.

   f) Uniform (θ_1, θ_2).

8. If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with densities $f(x;\theta)$ given below, find a sufficient statistics for θ.

   a) $f(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x), \theta \in (0,\infty)$.

   b) $f(x;\theta) = \frac{1}{6\theta^4} x^3 e^{-x/\theta} I_{(0,\infty)}(x), \theta \in (0,\infty)$.

9. Show that the minimal sufficient statistic T_n for the parameter σ of the Scale-Cauchy family $f(x,\sigma) = \frac{\sigma}{\pi(x^2+\sigma^2)}$ has dimension n and is equal to $T_n = (X^2_{(1)}, X^2_{(2)}, \ldots, , X^2_{(n)})$ where $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ is the variation sequence.

10. Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from a scale parameter family $\{F_\sigma(X)\}, \sigma > 0$ with $F_\sigma(x) = F(x/\sigma), \sigma > 0 (F(.)$- a given continuous cumulative distribution function.) Show that any statistic that depends on the sample through the $n-1$ values $X_1/X_n, X_2/X_n, \ldots, X_{n-1}/X_n$ is an ancillary statistic.

Cramer-Rao Bound. UMVUE

11. Calculate the Cramer-Rao lower bound for the variance of an unbiased estimator of θ and find a statistic with variance equal to the bound when $X_1, X_2, \ldots, X_n$ are independent random variables each with distribution

    a) $f(x,\theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0$,

    b) Bernoulli (θ),

    c) $N(\theta, 1)$,

    d) $N(0, \theta)$.

    e) Prove that no unbiased estimator of θ has variance equal to the bound when the distribution is $N(0, \theta^2)$.

12. If $X_1, X_2, \ldots, X_n$ are independent Poisson (λ) random variables, find the *umvue* of $e^{-2\lambda}$. Check that the estimator has mean $e^{-2\lambda}$ and compare the variance of the estimator with the Cramer-Rao lower bound for the variance of an unbiased estimator of $e^{-2\lambda}$.

13. Suppose random variables X and Y have joint density

$$f_{X,Y}(x,y) = 8xy, 0 < y < x, 0 < x < 1.$$

    For this pair of random variables, verify directly the lemma which states that if $a(x) = \mathbf{E}(Y|X = x)$, then $\mathbf{E}a(X) = \mathbf{E}(Y)$ and $Var\{a(X)\} \leq Var(Y)$.

14. Find the *umvue* of θ^2 when $X_1, X_2, \ldots, X_n$ are independent Bernoulli (θ) random variables. Check that your estimator does have mean θ^2.

15. Find the *umvue* of θ^2 when $X_1, X_2, \ldots, X_n$ are independent random variables each with density

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x > 0; \theta > 0.$$

    Hint: consider $\bar{X}^2$.

16. Suppose $X_1, X_2, \ldots, X_n$ are independent Uniform $(0, \theta)$ random variables.

    a) Find the *umvue* of θ^2 and calculate its variance.

    b) Find the *umvue* of $\frac{1}{\theta}$.

17. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables, each with density

$f(x;\theta) = \theta e^{-\theta x}, x > 0, \theta > 0$. Let $T = \sum_{i=1}^n X_i$.

a) Prove (*) that the density of T is given by $f(t;\theta) = \frac{1}{\Gamma(n)}\theta^n t^{n-1} e^{-\theta t}, t > 0$.

b) Prove that the indicator function of the event $\{X_1 > k\}$ is an unbiased estimator of $e^{-k\theta}$, where k is a known constant.

c) If $T = \sum_{i=1}^n X_i$, take for granted (or try to prove using a) (*)) that the conditional density of X_1 given $T = t$ is

$$f(x_1|t) = \frac{(n-1)}{t}(1 - \frac{x_1}{t})^{n-2}, 0 < x_1 < t < \infty.$$

Then find the *umvue* of $e^{-k\theta}$.

18. The random variable X takes values 0,1,2,3 with probabilities

P(X=0)	P(X=1)	P(X=2)	P(X=3)
$2\theta^2$	$\theta - 2\theta^3$	θ^2	$1 + 2\theta^3 - 3\theta^2 - \theta$

The range of the parameter θ is $\theta \in \Theta = (0, \frac{1}{5})$. Is the family of distributions $\{P_\theta(X)\}, \theta \in \Theta$, complete? Give reasons for your answer.

19. Is the following statistic complete:

a) $T = \bar{X}$ when the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are i.i.d. $N(0, \theta)$.

b) $T = \sum_{i=1}^n X_i$ when the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are i.i.d. Bernoulli (θ).

c) $T = \sum_{i=1}^n X_i$ when the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are i.i.d. Poisson (θ).

d) $T = X_{(n)}$ when the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are i.i.d. Uniform $(0, \theta)$.

Answers:

11. a) bound: $\frac{\theta^2}{n}$, UMVUE: $\bar{X}$;
b) bound: $\frac{\theta(1-\theta)}{n}$, UMVUE: $\bar{X}$;
c) bound: $\frac{1}{n}$, UMVUE: $\bar{X}$;
d) bound: $\frac{2\theta^2}{n,}$ UMVUE: $\frac{\sum_{i=1}^n X_i^2}{n}$;
e) bound: $\frac{\theta^2}{2n,}$ the score is $-\frac{n}{\theta} + \frac{\sum_{i=1}^2 X_i^2}{\theta^3}$ and can not be written as $K(\theta, n)(T - \theta)$.
12. $(1 - \frac{2}{n})^{\sum_{i=1}^n X_i}$, CR bound: $\frac{4\lambda e^{-4\lambda}}{n}$ and is smaller than the variance of the UMVUE.
13. $Var(a(X)) = \frac{8}{675}, Var(Y) = \frac{11}{225}$.
14. $\frac{T(T-1)}{n(n-1)}, T = \sum_{i=1}^n X_i$.
15. $\frac{T^2}{n(n+1)}, T = \sum_{i=1}^n X_i$.
16. a) $\frac{(n+2)T^2}{n}, T = X_{(n)}$.
17. UMVUE: $\{\frac{T-k}{T}\}^{n-1} I_{(k,\infty)}(T), T = \sum_{i=1}^n X_i$.
18. Not complete.
19. a) Not complete; b) Complete; c) Complete; d) Complete.

MLE and their properties. Asymptotic properties of estimators

20. A sample of size n_1 is to be drawn from a normal population with mean μ_1 and variance σ_1^2. A second sample of size n_2 is to be drawn from a normal population with mean μ_2 and variance σ_2^2. What is the MLE of $\theta = \mu_1 - \mu_2$? If we assume that the total sample size $n = n_1 + n_2$ is fixed, how should the n observations be divided between the two populations in order to minimize the variance of the MLE?

21. Let $X_1, X_2, \ldots, X_n$ be a sample from the density $f(x;\theta) = \theta x^{-2} I_{[\theta,\infty)}(x)$ where $\theta > 0$.

a) Find the MLE of θ.

b) Is $Y = X_{(1)}$ a sufficient statistic?

22. Let $X_1, X_2, \ldots, X_n$ be a sample from the density function $f(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x)$ where $\theta > 0$.

a) Find the MLE of $\tau(\theta) = \frac{\theta}{1+\theta}$.

b) State the asymptotic distribution of the MLE of $\tau(\theta)$ in a).

c) Find a sufficient statistic, and check completeness. Is $\sum_{i=1}^n X_i$ a sufficient statistic?

d) Is there a function of θ for which there exists an unbiased estimator whose variance coincides with the Cramer-Rao lower Bound? What is the Cramer-Rao lower bound?

23. Let $X_1, X_2, \ldots, X_n$ be a sample from normal distribution $N(\mu, \sigma^2)$ where μ is known and σ^2 is the parameter to be estimated.

a) Find the MLE and state its asymptotic distribution.

b) Assume now that σ is to be estimated. Find the MLE and state its asymptotic distribution.

24. Consider n i.i.d. observations from a Poisson (λ) distribution.

a) Suppose the parameter of interest is $\tau(\lambda) = \frac{1}{\lambda}$.
i) What is the MLE of $\tau(\lambda)$?
ii) What is its variance?
iii) What is its asymptotic variance?

b) Assume that the parameter of interest is $\tau(\lambda) = \sqrt{\lambda}$.
i) State the asymptotic distribution of the MLE of $\sqrt{\lambda}$. In particular, show that the asymptotic variance does not depend on λ (we say in that case that $\sqrt{\lambda}$ is a *variance stabilising transformation*).
ii) For a given small value of $\alpha \in (0,1)$ and using the result in i), how would you construct a confidence interval for λ that asymptotically has a level $1-\alpha$.

Answers:

20. MLE: $\bar{X}_{n_1} - \bar{Y}_{n_2}; n_1 = \frac{\sigma_1}{\sigma_1+\sigma_2} n$

21. $\hat{\theta}_{mle} = X_{(1)}$, sufficient.

22. a) $\frac{n}{n-\sum_{i=1}^n \log X_i}$; b) $N(0, \frac{\theta^2}{(1+\theta)^4})$
c) $\sum_{i=1}^n \log X_i$ is sufficient and complete; $\sum_{i=1}^n X_i$ is **not** sufficient.
d) $\tau(\theta) = \frac{1}{\theta}$ is such a function. The attainable bound in this case is $\frac{1}{n\theta^2}$.

23. a) MLE of σ^2 is $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$. The asymptotic distribution : $N(0, 2\sigma^4)$.
b) MLE is $\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2}$. The asymptotic distribution is $N(0, \frac{1}{2}\sigma^2)$.

24. a) Variance of MLE is infinite but **asymptotic** variance of MLE is finite and equals $\frac{1}{\lambda^3}$.
b) Asymptotic variance is 0.25.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Additional exercises for MATH5905, Statistical Inference

Part three: Hypothesis testing

1. For each of the families $L(\mathbf{x},\theta)=\prod_{i=1}^{n} f(x_i,\theta)$ below suggest a statistic $T(\mathbf{X})$ with respect to which the family has the MLR property:
   a) $f(x;\theta)$ is $N(\theta,1)$
   b) $f(x;\theta)=\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0;\theta>0.$
   c) $f(x;\theta)=\theta e^{-x\theta}, x>0;\theta>0.$
   d) $f(x;\theta)$ is $N(0,\theta^2)$
   e) $f(x;\lambda)$ is Poisson $(\lambda), \lambda>0.$
   f) $f(x;p)$ is Bernoulli $(p), p\in(0,1).$
   g) $f(x;\theta)$ is Uniform $(0,\theta).$

2. Find the *ump* size α test of $H_0:\sigma\le\sigma_0$ versus $H_1:\sigma>\sigma_0$ based on n i.i.d. observations from $N(0,\sigma^2)$ population. Sketch a graph of its power function. Answer the same question in case that $H_0:\sigma\ge\sigma_0$ versus $H_1:\sigma<\sigma_0$ was to be tested.

3. Let X be a single observation from the density
$$f(x;\theta)=\theta x^{\theta-1}, 0\le x\le 1,\ \theta>0\ .$$
   a) For testing $H_0:\theta\le 1$ versus $H_1:\theta>1$, find the power function and size of the test with rejection region $x\ge\frac{1}{2}$.
   b) Find the most powerful test of size $\alpha=0.05$ of $H_0:\theta=2$ versus $H_1:\theta=1$.
   c) Find the *ump* size α test of $H_0:\theta\ge 2$ versus $H_1:\theta<2$ and calculate its power function.
   d) Find the generalized likelihood-ratio test of $H_0:\theta=1$ versus $H_1:\theta\ne 1$ with size $\alpha=0.1$.

4. Suppose X_1 and X_2 are independent random variables, each with density

   $f(x;\theta)=\theta x^{\theta-1}$, $0\le x\le 1$. For testing $H_0:\theta\le 1$ versus $H_1:\theta>1$, find the size and the power function of the test with rejection region $3x_1\le 4x_2$. Would you use this test? What alternative test could you suggest?

5. For a random sample of size n from the density $f(x;\theta)=e^{-(x-\theta)}, x\ge\theta$, construct the *ump* size α test of $H_0:\theta\le\theta_0$ versus $H_1:\theta>\theta_0$. Calculate the power function of the test and sketch its graph.

6. For a sample of size $n=10$ from a Poisson (λ) family construct the *ump* $\alpha=.10$ size test of $H_0:\lambda\le 1$ versus $H_1:\lambda>1$. You may utilize the following extract of a table of Poisson (10) probabilities:

x	12	13	14	15	16
$P(X\le x)$	0.7915	0.8644	0.9165	0.9512	0.9729

7. Find the form of the rejection region of the *ump* test of $H_0:\theta\le\theta_0$ versus $H_1:\theta>\theta_0$ based on independent random variables $X_1,X_2,\ldots,X_n$ each with density
$$f(x;\theta)=\frac{1}{\theta}e^{-x/\theta}, x>0,\theta>0\ .$$
   Use the Central Limit Theorem to determine approximately the constant specifying the rejection region for a size α test. Hence find an appropriate expression for the power function.

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where σ^2 is known. Let Λ denote the generalized likelihood ratio for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Find the exact distribution of $-2\log\Lambda$, and compare it with the corresponding asymptotic distribution when H_0 is true.

9. Suppose $X_1, X_2, \ldots, X_m$ are independent $N(\mu_1, 1)$ random variables and $Y_1, Y_2, \ldots, Y_n$ is an independent set of independent $N(\mu_2, 1)$ random variables.

   a) Show that when $\mu_1 = \mu_2 = \mu$, say, the MLE of μ is $\tilde{\mu} = \frac{m\bar{X}+n\bar{Y}}{m+n}$.

   b) Prove that when the MLE's of μ_1 and μ_2 are $\hat{\mu}_1 = \bar{X}$ and $\hat{\mu}_2 = \bar{Y}$.

   c) Derive the generalized likelihood ratio statistic $\Lambda_{m,n}$ for testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ and show that, when H_0 is true, $-2\log\Lambda_{m,n}$ has precisely χ^2_1 distribution for every m, n.

   Answers:

   1) MLR in: a) $T(\mathbf{X}) = \sum_{i=1}^n X_i$, b) $T(\mathbf{X}) = \sum_{i=1}^n X_i$, c) $T(\mathbf{X}) = -\sum_{i=1}^n X_i$, d) $T(\mathbf{X}) = \sum_{i=1}^n {X_i}^2$, e) $T(\mathbf{X}) = \sum_{i=1}^n X_i$, f) $T(\mathbf{X}) = \sum_{i=1}^n X_i$, g) $T(\mathbf{X}) = X_{(n)}$.

   2) $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } \sum_{i=1}^n X_i^2 \geq \sigma_0^2\chi^2_{n,\alpha} \\ 0 \text{ if } \sum_{i=1}^n X_i^2 < \sigma_0^2\chi^2_{n,\alpha}. \end{cases}$ with a power function $p(t) = P(\chi^2_n \geq \frac{\sigma_0^2}{t^2}\chi^2_{n,\alpha})$. But in the case of $H_0 : \sigma \geq \sigma_0$ versus $H_1 : \sigma < \sigma_0$, the ump $\alpha-$ test changes to

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } \sum_{i=1}^n X_i^2 \leq \sigma_0^2\chi^2_{n,1-\alpha} \\ 0 \text{ if } \sum_{i=1}^n X_i^2 > \sigma_0^2\chi^2_{n,1-\alpha}. \end{cases}$$

   with a power function $p(t) = P(\chi^2_n \leq \frac{\sigma_0^2}{t^2}\chi^2_{n,1-\alpha})$.

   3) a) $E_\theta\varphi = 1 - \frac{1}{2^\theta}, \theta > 0$, size at $\theta_0 = 1 : E_{\theta_0}\varphi = 0.5$.

   b) $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } X \leq 0.2236 \\ 0 \text{ if } X > 0.2236 \end{cases}$

   c) same test as in b)

   d) $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } X \leq 0.05 \text{ or } X > 0.95 \\ 0 \text{ else.} \end{cases}$

   4) Too high size, test is not to be recommended. Test based on $T = \log(X_1) + \log(X_2)$ should be used instead.

   5) $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } X_{(1)} > k \\ 0 \text{ if } X_{(1)} \leq k \end{cases}$ where $k = \theta_0 - \frac{ln(\alpha)}{n}$.

   6) $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } T = \sum_{i=1}^{10} X_i > 14 \\ 0.317 \text{ if } T = 14 \\ 0 \text{ if } T < 14 \end{cases}$

   7) Rejection region: $\{\mathbf{X} : \sum_{i=1}^n X_i \geq k = \theta_0\gamma_{n,\alpha}\}$ where $\gamma_{n,\alpha}$ is the upper $\alpha * 100\%$ point of the gamma (n) distribution (exact result). *Asymptotically,* it holds $k \approx n\theta_0 + \sqrt{n}\theta_0 z_\alpha$.

   8) See your lecture.

   9) c) If $T = \frac{mn}{m+n}(\bar{X} - \bar{Y})^2$ then the generalized likelihood ratio test is

   $\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } T \geq \chi^2_{1,\alpha} \\ 0 \text{ if } T < \chi^2_{1,\alpha} \end{cases}$

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Additional exercises for MATH5905, Statistical Inference

Part four: Multinomial distribution. Order statistics. Robustness

1. a) X_1, X_2, X_3 has a multinomial $(8; 0.2, 0.3, 0.5)$ distribution. Find $P(X_1 = 2, X_2 = 2, X_3 = 4)$, $E(X_2), Var(X_2)$ and $Cov(X_1, X_3)$.

   b) X_1, X_2, X_3 has a multinomial $(6; 0.5, 0.2, 0.3)$ distribution. Find $P(X_1 = 3, X_2 = 1, X_3 = 2)$ and $P(X_1 + X_2 = 2)$.

2. Find the probability density function of the second order statistic $X_{(2)}$ in a random sample of size four from a population with the density function

$$f(x) = \begin{cases} e^{1-x}, 1 < x < \infty, \\ 0 \text{ elsewhere} \end{cases}$$

3. Find the probability density function of $X_{(4)}$ in a random sample of size five from a population with the density function

$$f(x) = \begin{cases} \frac{1}{x^2}, 1 \leq x < \infty, \\ 0 \text{ elsewhere} \end{cases}$$

4. The opening prices per share of two similar stocks Y_1 and Y_2 are independent random variables, each with density function

$$f(y) = \begin{cases} \frac{1}{2} e^{-\frac{y-4}{2}}, y \geq 4, \\ 0 \text{ elsewhere} \end{cases}$$

   On a given morning Mr. A is going to buy shares of whichever stock is less expensive. Find the probability density function and the expected value for the price per share that Mr. A will have to pay.

5. Find the expected value of the largest order statistic in a random sample of size 3 from:

   a) the exponential distribution with density $f(x) = e^{-x}, x \geq 0$.

   b) the standard normal distribution

6. Electric components of a certain type have lifetime Y with probability density given by

$$f(y) = \begin{cases} \frac{1}{100} e^{-\frac{y}{100}}, y > 0, \\ 0 \text{ elsewhere} \end{cases}$$

   a) Suppose that two such components operate independently and in series in a certain system (that is, the system fails when either component fails). Find the density function for X, the lifetime of the system.

   b) Now suppose that the components operate in parallel (that is, the system does not fail until both components fail). Find the density function for X, the lifetime of the system.

7. A continuous random variable X has a standard exponential distribution

$$f(x) = \begin{cases} e^{-x}, x > 0, \\ 0 \text{ elsewhere} \end{cases}$$

   For a random sample of size 3, let $X_{(1)}, X_{(2)}, X_{(3)}$ denote the ordered sample.

a) Write down the joint distribution of $X_{(1)}$ and $X_{(3)}$.

b) Obtain the distributions of $X_{(1)}$ and $X_{(3)}$.

c) Evaluate $EX_{(1)}$ and $EX_{(3)}$.

d) Find the sampling distribution of the range $R = X_{(3)} - X_{(1)}$.

8. A random sample of size 3 is taken from a population with density

$$f(x) = \begin{cases} 2x, 0 \leq x < 1, \\ 0 \text{ elsewhere} \end{cases}$$

Find the sampling distribution of the range R.

9. For a random sample of size 2 from a standard normal distribution, find the distribution of the range.

10. The Cauchy-Schwartz Inequality tells us that for random variables X, Y with $E(X^2) < \infty$ and $E(Y^2) < \infty$, $\{E(|XY|)\}^2 \leq E(X^2)E(Y^2)$ holds. Using this Inequality, show that when estimating the location parameter θ in $f(x-\theta)$ by using an M-estimator (defined by its ψ function):

a) we get a loss in asymptotic efficiency in comparison to the MLE estimator (i.e. the Inequality

$$\sigma^2(F, \psi) = \frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2} \geq \frac{1}{\int [\frac{f'(x)}{f(x)}]^2 f(x) dx}$$

holds).

b) Equality holds only when the M-estimator coincides with the MLE. (**But: despite the above observation we still may decide to use the M-estimator instead of the MLE because of the better robustness properties of the former in comparison to the latter**).

11. Assume that we are estimating the location parameter θ_0 for the family $f_\theta(x) = f(x-\theta)$ in a robustness context. The "true" value θ_0 to be estimated in robustness context is usually taken as the solution of the Equation $E_{\theta_0}\psi(X - \theta_0) = 0$. Derive the asymptotic normality statement from the lecture notes for the M-estimator a defined as the solution to the equation $\sum_{i=1}^n \psi(X_i - \hat{\theta}_M) = 0$:

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \rightarrow^d N(0, \frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2}).$$

Answers

1 a) 0.0945, 2.4, 1.68, -0.8 b) 0.135 0.059535

2) $12exp(3(1 - x_{(2)}))(1 - exp(1 - x_{(2)})), 1 \leq x_{(2)} < \infty$

3) $20(x_{(4)} - 1)^3 / x_{(4)}^6, 1 \leq x_{(4)} < \infty$

4) $exp(-y_{(1)} - 4), y_{(1)} \geq 4$. The expected value is 5.

5) a) 11/6, b) $\frac{3}{2\sqrt{\pi}}$.

6) a) $\frac{1}{50}exp(-\frac{x}{50})$; b) $\frac{1}{50}(1 - exp(-\frac{x}{100}))exp(-\frac{x}{100})$

7) a) $6(e^{-x_{(1)}} - e^{-x_{(3)}})e^{-(x_{(1)}+x_{(3)})}, 0 < x_{(1)} < x_{(3)} < \infty$.

b) $3exp(-3x_{(1)}), x_{(1)} > 0; 3(1 - exp(-x_{(3)}))^2 exp(-x_{(3)}), x_{(3)} > 0$.

c) $1/3, 11/6$; d) $2exp(-u)(1 - exp(-u)), u > 0$.

8) $12u(1-u)^2, 0 < u < 1$.

9) $\frac{1}{\sqrt{\pi}}exp(-\frac{u^2}{4}), u > 0$.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Additional exercises for MATH5905, Statistical Inference

Part five: Higher order asymptotics

1. Consider n i.i.d. observations from Poisson (λ) distribution.

   i) Show that the cumulant generating function for a single observation is $K_X(t) = \lambda[e^t - 1]$.

   ii) Show that the one-term saddlepoint approximation for the density of the sum of n i.i.d. observations $Y = \sum_{i=1}^n X_i$ from this distribution is given by

   $$\hat{g}(y) = \frac{1}{\sqrt{2\pi}} e^{-n\lambda} \frac{e^y (n\lambda)^y}{y^{y+1/2}}, \; y = 0, 1, 2, \ldots.$$

   (Having in mind Stirling's approximation of $n! \approx \sqrt{2\pi n} e^{-n} n^n$ for large n, the above approximation coincides up to a normalising constant with the exact density (which is Poisson $(n\lambda)$) and the constant tends to 1 when $n \to \infty$.)

   Solution:

   1) Hint: Using the formula (27) from lecture 8 (Higher order asymptotics), get the saddlepoint density approximation of the mean $\bar{Y}$ and then use density transformation formula.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Solutions to selected exercises for MATH5905, Statistical Inference

Part three: Hypothesis Testing

1) Check the answers and try to explain to yourself why the MLR property holds. Study the examples from lectures first.

2) Using 1d), we know that the rejection region of the ump-α test of $H_0 : \sigma \leq \sigma_0$ versus $H_1 : \sigma > \sigma_0$ is in the form $\{\sum_{i=1}^n X_i^2 \geq k\}$. To determine k, we have to "exhaust the level", that is:

$$\alpha = P(\sum_{i=1}^n X_i^2 \geq k | \sigma = \sigma_0) = P(\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} \geq \frac{k}{\sigma_0^2}) = P(\chi_n^2 \geq \frac{k}{\sigma_0^2}).$$

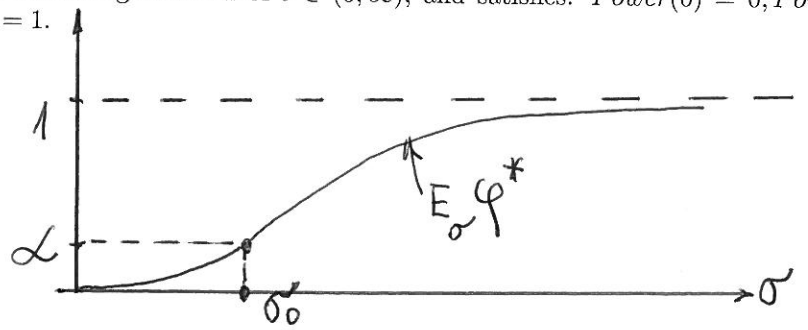
Hence $k/\sigma_0^2 = \chi_{n,\alpha}^2 \Rightarrow k = \sigma_0^2 \chi_{n,\alpha}^2$ and the ump-α test is:

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 \text{ if } \sum_{i=1}^n X_i^2 \geq \sigma_0^2 \chi_{n,\alpha}^2 \\ 0 \text{ if } \sum_{i=1}^n X_i^2 < \sigma_0^2 \chi_{n,\alpha}^2 \end{cases}$$

The power function:

$$Power(t) = P(\sum_{i=1}^n X_i^2 > \sigma_0^2 \chi_{n,\alpha}^2 | \sigma = t) = P(\chi_n^2 \geq (\frac{\sigma_0}{t})^2 \chi_{n,\alpha}^2).$$

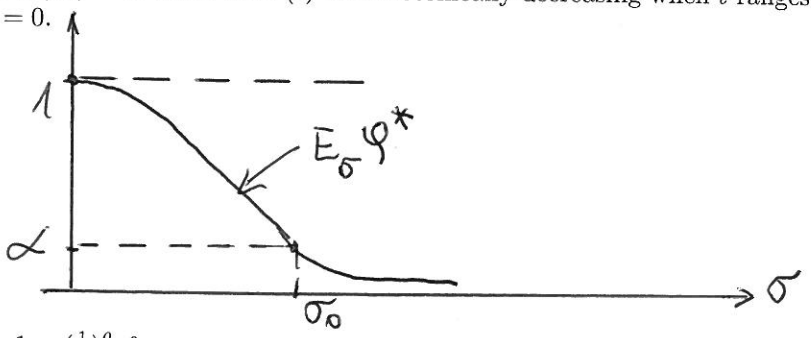
Thus $Power(t)$ is increasing function of $t \in (0, \infty)$, and satisfies: $Power(0) = 0, Power(\sigma_0) = \alpha$, and $lim_{t \to \infty} Power(t) = 1$.



If $H_0 : \sigma \geq \sigma_0$ versus $H_1 : \sigma < \sigma_0$ was to be tested, then using the **Note** after the Blackwell-Girshick theorem, we know that the ump-α test exists and that now the rejection region is $\{\sum_{i=1}^n X_i^2 < \sigma_0^2 \chi_{n,1-\alpha}^2\}$. The power function is

$$Power(t) = P(\chi_n^2 \leq (\frac{\sigma_0}{t})^2 \chi_{n,1-\alpha}^2).$$

The graph will be "reversed" now since the hypothetical and the alternative region have been changed. We will have: $Power(0) = 1, Power(\sigma_0) = \alpha$, and $Power(t)$ is monotonically decreasing when t ranges from 0 to ∞ with $lim_{t \to \infty} Power(t) = 0$.



3) a) $E_\theta \varphi = \int_{1/2}^1 \theta x^{\theta-1} dx = 1 - (\frac{1}{2})^\theta, \theta > 0$ is the power function. The size is obtained at $\theta_0 = 1$, so $E_{\theta_0} \varphi = \frac{1}{2}$.

b) By the Neyman-Pearson lemma, for $H_0 : \theta = 2$ versus $H_1 : \theta = 1$, the best α-test is the one with a rejection region in the form $\{\frac{L(\mathbf{X};1)}{L(\mathbf{X};2)} \geq k\}$. Here, the sample size is $n = 1$ and we have $\frac{L(\mathbf{X};1)}{L(\mathbf{X};2)} = \frac{1x^0}{2x} \geq k$. Equivalently: $x \leq \frac{1}{2k} = k'$. To make it an α-test, we need

$$\alpha = 0.05 = P(X \leq k'|\theta = 2) = \int_0^{k'} 2xdx = (k')^2.$$

This implies that $k' = \sqrt{0.05} \approx .2236$. Hence the best 0.05-size test of $H_0 : \theta = 2$ versus $H_1 : \theta = 1$ is

$$\varphi^* = \begin{cases} 1 \text{ if } x \leq .2236 \\ 0 \text{ if } x > .2236 \end{cases}$$

c) $f(x;\theta) = \theta e^{(\theta-1)lnx}$ is obviously a member of one-parameter exponential family with $d(x) = lnx$ and a monotonically increasing $c(\theta) = \theta - 1$. Hence, according to the **Note** to the Blackwell-Girshick theorem, the ump-α test exists and has a rejection region $S = \{x : lnx \leq k\}$. But $lnx \leq k \Longleftrightarrow x \leq k'$. For $\alpha = 0.05$ we get: $\alpha = 0.05 = P(x \leq k'|\theta = 2) = (k')^2$ and we again get $k' = 0.2236$. This means that the same test as in b) is ump-α size test of $H_0 : \theta \geq 2$ versus $H_1 : \theta < 2$. (We could have also argued about this by noticing that in b) the rejection region **did not** depend on the θ value under the alternative, hence the same test as in b) will be an **ump**-0.05 test.

d)

$$\lambda = \frac{max_{\theta\in\Theta_0} L(x;\theta)}{max_{\theta\in\Theta} L(x;\theta)} = \frac{1}{(-\frac{1}{lnx})x^{-\frac{1}{lnx}-1}}$$

(on the bottom we have replaced the argument that gives rise to the maximum by the value of the MLE $\hat{\theta}_{mle} = -\frac{1}{lnx}$. We now observe that $g(x) = \frac{1}{(-\frac{1}{lnx})x^{-\frac{1}{lnx}-1}} = -exlnx$ tends to zero when $x \to 0$ or $x \to 1$. Therefore, $\lambda \leq$ constant is equivalent to $\{x \leq k_1 \text{ or } x \geq k_2\}$. The values of k_1 and k_2 must be such that

$$0.1 = \alpha = P(X \leq k_1 \text{ or } x \geq k_2|\theta = 1) = k_1 + 1 - k_2.$$

Several choices are possible for k_1 and k_2. If we want an **equal-tailed** test then $k_1 = 0.05$ and $k_2 = 1 - 0.05 = 0.95$ should be chosen, that is:

$$\varphi = \begin{cases} 1 \text{ if } x \leq 0.05 \text{ or } x > 0.95 \\ 0 \text{ else} \end{cases}$$

4) Draw a diagram of the first quadrant with axis OX_1 and OX_2 and try to represent the rejection region S as a subset of the unit square. This helps to understand the calculations below.
The joint density, because of the assumed independence, is given by

$$f_{X_1,X_2}(x_1, x_2) = \theta^2 x_1^{\theta-1} x_2^{\theta-1}.$$

Hence

$$E_\theta\varphi = \int\int_S f_{X_1,X_2}(x_1, x_2)dx_1dx_2 = \int_0^1 (\int_{3x_1/4}^1 \theta^2(x_1x_2)^{\theta-1}dx_2)dx_1 = \ldots = 1 - \frac{1}{2}(\frac{3}{4})^\theta.$$

The size of this test is $E_\theta\varphi_{|\theta=1} = \frac{5}{8}$. Note that the test is **not good** since the size is too high and $E_\theta\varphi_{|\theta=0} = \frac{1}{2}$ which is also very high. Test based on the statistic $T = lnX_1 + lnX_2$ should be used instead.

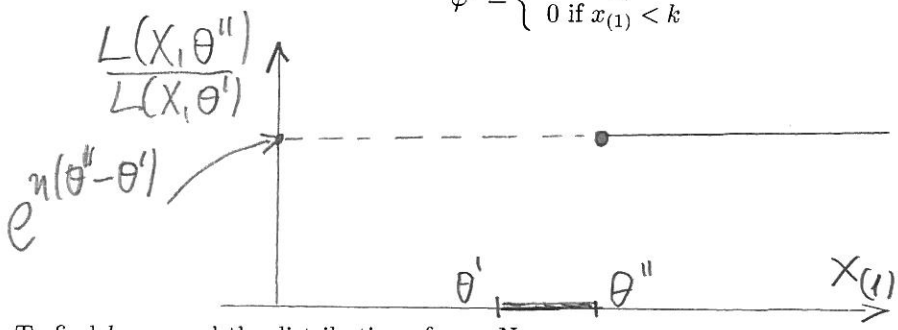
5) The likelihood function is

$$L(\mathbf{x},\theta) = I_{(\theta,\infty)}(x_{(1)})e^{[-\sum_{i=1}^n x_i + n\theta]}$$

The family has a **monotone likelihood ratio in** $X_{(1)}$ (show this by examining the behaviour of $\frac{L(\mathbf{X},\theta'')}{L(\mathbf{X},\theta')}, \theta'' > \theta'$ as a function of $x_{(1)}$ and convince yourself that the ratio is zero when $x_{(1)} \in (\theta', \theta'')$ but is equal to a positive constant $e^{n(\theta''-\theta')}$ when $x_{(1)} > \theta''$).

This means that a ump-α test exists and has the form

$$\varphi^* = \begin{cases} 1 \text{ if } x_{(1)} > k \\ 0 \text{ if } x_{(1)} < k \end{cases}$$



To find k we need the distribution of $x_{(1)}$. Now:

$$P_\theta(X_{(1)} > k) = (P_\theta(X_1 > k))^n$$

But

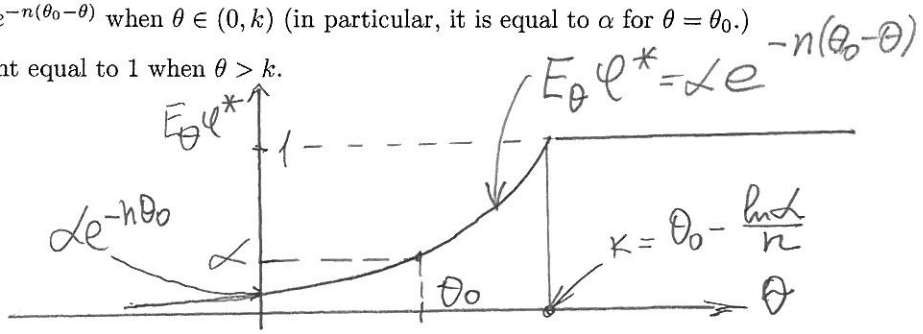
$$P_\theta(X_1 > k) = \begin{cases} 1 \text{ if } k \leq \theta \\ 1 - \int_\theta^k e^{-(t-\theta)}dt = e^{-(k-\theta)} \text{ if } \theta < k \end{cases}$$

Hence

$$E_\theta\varphi^* = P_\theta(X_{(1)} \geq k) = \begin{cases} 1 \text{ if } k \leq \theta \\ e^{-n(k-\theta)}, \text{ if } \theta < k \end{cases}$$

To find k we solve the equation $E_{\theta_0}\varphi^* = e^{-n(k-\theta)} = \alpha$. This gives $k = \theta_0 - \frac{ln\alpha}{n}$.
The powerfunction is defined on the positive half axis and is:

- equal to $\alpha e^{-n\theta_0}$ for $\theta = 0$.
- equal to $\alpha e^{-n(\theta_0-\theta)}$ when $\theta \in (0,k)$ (in particular, it is equal to α for $\theta = \theta_0$.)
- is a constant equal to 1 when $\theta > k$.



6) According to 1e) we have a MLR property in $T = \sum_{i=1}^{10} X_i$. Moreover, $T \sim Po(10\lambda)$. For $\lambda_0 = 1$ this is the Poisson distribution with parameter 10. Blackwell-Girshick theorem tells us that a ump-α test ($\alpha = 0.1$) exists and is in the form

$$\varphi^* = \begin{cases} 1 \text{ if } T > 14 \\ \gamma \text{ if } T = 14 \\ 0 \text{ if } T < 14 \end{cases}$$

The value of γ is

$$\gamma = \frac{.1 - .0835}{.9165 - .8644} = \frac{.0165}{.0521} = .317$$

7) We use the density transformation formula:

$$f_Y(y) = f_X(w^{-1}(y))|\frac{dw^{-1}(y)}{dy}|$$

Since $X \sim f(x;\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, x > 0, \theta > 0$ we get: $Y = \frac{X}{\theta}$ has a standard exponential density $f_Y(y) = e^{-y}, y > 0$ and then, using the properties of Gamma distribution,

$$\sum_{i=1}^{n} X_i/\theta \sim gamma(n)$$

with the density $f_{gamma(n)}(x) = \frac{e^{-x}x^{n-1}}{\Gamma(n)}, x > 0$. Hence, the ump-$\alpha$ test is given by

$$\varphi^* = \begin{cases} 1 \text{ if } \sum_{i=1}^n X_i \geq k \\ 0 \text{ if } \sum_{i=1}^n X_i < k \end{cases}$$

and

$$P(\frac{1}{\theta_0}\sum_{i=1}^n X_i \geq \frac{k}{\theta_0}|\theta = \theta_0) = \int_{k/\theta_0}^{\infty} \frac{e^{-x}x^{n-1}}{\Gamma(n)}dx.$$

Hence, the threshold is $k = \theta_0\gamma_{n,\alpha}$ where $\gamma_{n,\alpha}$ is the upper $\alpha * 100\%$ point of the $gamma(n)$ density. This is an **exact** result.

On the other hand, asymptotically (for large n), by using the Central Limit Theorem (CLT) and the fct that $EX_i = \theta_0, VarX_i = \theta_0^2$ we can get the following approximate value for the threshold:

$$\alpha = P(\sum_{i=1}^n X_i \geq k|\theta = \theta_0) = P(\frac{\sqrt{n}(\bar{X} - \theta_0)}{\theta_0} \geq \frac{\sqrt{n}(\frac{k}{n} - \theta_0)}{\theta_0}) \approx 1 - \Phi(\frac{\sqrt{n}(\frac{k}{n} - \theta_0)}{\theta_0})$$

which implies that $(\frac{k}{n} - \theta_0)\sqrt{n} = \theta_0 z_\alpha$. Hence $k \approx n\theta_0 + \sqrt{n}\theta_0 z_\alpha$ should be chosen in order to have the size asymptotically equal to α.

8) Discussed at lectures.

9)

$$L(\mathbf{X},\mathbf{Y}; \mu_1, \mu_2) = (2\pi)^{-\frac{m+n}{2}} e^{-\frac{1}{2}[\sum_{i=1}^m (x_i-\mu_1)^2 + \sum_{i=1}^n (y_i-\mu_2)^2]}.$$

Unrestricted maximisation with respect to both μ_1 and μ_2 leads to $\bar{X}, \bar{Y}$ as solutions. Now, **restricted** maximisation under the restriction $\mu_1 = \mu_2 = \mu$ leads to $\hat{\mu}_{mle(restricted)} = \frac{\sum_{i=1}^m x_i + \sum_{i=1}^n y_i}{m+n}$. Therefore,

$$2ln\Lambda_{m,n} = \sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^m (x_i - \frac{m\bar{x} + n\bar{y}}{m+n})^2 - \sum_{i=1}^n (y_i - \frac{m\bar{x} + n\bar{y}}{m+n})^2 =$$

$$-m\bar{x}^2 - n\bar{y}^2 + 2\frac{m\bar{x} + n\bar{y}}{m+n}(\sum_{i=1}^m x_i + \sum_{i=1}^n y_i) - (m+n)(\frac{m\bar{x} + n\bar{y}}{m+n})^2 = \ldots = -\frac{mn(\bar{x} - \bar{y})^2}{m+n}$$

Hence

$$-2ln\Lambda_{m,n} = \frac{mn(\bar{x} - \bar{y})^2}{m+n} = T$$

(which can be seen directly to be distributed as chi-square with one degree of freedom under the hypothesis of equal means. Our Generalised LRT test is then:

$$\varphi = \begin{cases} 1 \text{ if } T \geq \chi^2_{1,\alpha} \\ 0 \text{ if } T < \chi^2_{1,\alpha} \end{cases}$$

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Solutions to selected exercises for MATH5905, Statistical Inference

Part one: Decision theory. Bayes and minimax rules

Question 1 Answers: Please draw carefully the graph of the risk set before doing anything else.

a) d_3 since the minimal between the four values $\{6, 5, 3, 5\}$ is 3. ;
b) The rule d_3 again. Its minimax risk is 3.
c) The rule d_3 again. Its Bayes risk is equal to $\frac{1}{3} \times 2 + \frac{2}{3} \times 3 = 2\frac{2}{3}$.
d) Chooses d_2 and d_4 with probability 1/2 each.
e)All priors in the form $(p, 1-p)$ with $1 > p > 3/5$. Explanation: the slope $-\frac{p}{1-p}$ should be smaller than the slope $-\frac{3}{2}$ of $\overline{d_1 d_3}$.

Question 2 Note that for X uniformly distributed in $[0, \theta)$ we have the density $f(x, \theta) = \frac{1}{\theta} I_{[0,\theta)}(x)$ and from here we have easily $E(X) = \frac{\theta}{2}, E(X^2) = \frac{\theta^2}{3}$. The rule is unbiased when $\mu = 2 : E(2X) = \theta$ holds.

For any fixed value of μ we have $E(\theta - \mu X)^2 = \theta^2(1 - \mu + \mu^2/3)$. When $\mu = \frac{3}{2}$ the latter mean squared error is equal to $\frac{\theta^2}{4}$. Now, we get

$$E(\theta - \mu X)^2 - E(\theta - \frac{3}{2}X)^2 = \frac{\mu^2\theta^2}{3} - \mu\theta^2 + \frac{3\theta^2}{4} = \frac{\theta^2}{12}(2\mu - 3)^2 \geq 0$$

the rule $\frac{3}{2}X$ will be uniformly better than any other rule in the form μX (that is, any rule in the form μX would be inadmissible unless $\mu = 3/2$).

Question 3 i)

$$f(\mathbf{x}|\theta)\tau(\theta) = k\theta^n e^{-\theta(\sum_{i=1}^n x_i + k)}$$

$$g(\mathbf{x}) = \int_0^\infty f(\mathbf{x}|\theta)\tau(\theta)d\theta = k\int_0^\infty \theta^n e^{-\theta(\sum_{i=1}^n x_i + k)}d\theta$$

Now we change the variables: set $\theta(\sum_{i=1}^n x_i + k) = y, d\theta = \frac{dy}{(\sum_{i=1}^n x_i + k)}$ and get:

$$g(\mathbf{x}) = \frac{k}{(\sum_{i=1}^n x_i + k)^{n+1}} \int_o^\infty y^n e^{-y} dy = \frac{k\Gamma(n+1)}{(\sum_{i=1}^n x_i + k)^{n+1}}$$

Hence

$$h(\theta|\mathbf{x}) = \frac{\theta^n e^{-\theta(\sum_{i=1}^n x_i + k)}}{\Gamma(n+1)(\frac{1}{\sum_{i=1}^n x_i + k})^{n+1}}, \theta > 0.$$

Recalling the general definition of a Gamma(α, β) density:

$$f(x; \alpha, \beta) = \frac{e^{-\frac{x}{\beta}} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, x > 0,$$

we see that $h(\theta|\mathbf{x}) \sim$ Gamma$(n+1, \frac{1}{\sum_{i=1}^n x_i + k})$.

NOTE: We **did NOT REALLY HAVE** to determine that normalising constant the way I showed above. Here is an **EASIER APPROACH**. Indeed just by looking at the joint density

$$f(\mathbf{x}|\theta)\tau(\theta) = k\theta^n e^{-\theta(\sum_{i=1}^n x_i + k)}$$

we can identify that up to a normalising constant this is a $Gamma(n+1, \frac{1}{\sum_{i=1}^n x_i+k})$ density hence the posterior $h(\theta|\mathbf{x})$ HAS to be $Gamma(n+1, \frac{1}{\sum_{i=1}^n x_i+k})$.

ii) For a Bayes estimator with respect to quadratic loss, we have $\hat{\theta} = E(\theta|\mathbf{X})$, and for a Gamma (α, β) density it is known that the expected value is equal to $\alpha\beta$ hence we get immediately $\hat{\theta} = \frac{n+1}{\sum_{i=1}^n x_i+k}$. Of course, we could also calculate directly:

$$\hat{\theta} = \int_0^\infty \theta h(\theta|\mathbf{x})d\theta = \frac{(\sum_{i=1}^n x_i + k)^{n+1}}{\Gamma(n+1)} \int_0^\infty \theta^{n+1} e^{-\theta(\sum_{i=1}^n x_i+k)} d\theta$$

and after changing variables: $\theta(\sum_{i=1}^n x_i + k) = y, d\theta = \frac{dy}{(\sum_{i=1}^n x_i+k)}$ we can continue the evaluation:

$$\hat{\theta} = \frac{\int_0^\infty e^{-y} y^{n+1} dy}{\Gamma(n+1)(\sum_{i=1}^n x_i + k)} = \frac{\Gamma(n+2)}{\Gamma(n+1)(\sum_{i=1}^n x_i + k)} = \frac{n+1}{\sum_{i=1}^n x_i + k}$$

Question 4 Note that we have a SINGLE observation X only. Now: $f(x|\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta)$ implies that

$$g(x) = \int_0^\infty f(x|\theta)\tau(\theta)d\theta = \int_x^\infty \frac{1}{\theta}\theta e^{-\theta} d\theta = e^{-x}, x > 0.$$

Hence

$$h(\theta|x) = \frac{f(x|\theta)\tau(\theta)}{g(x)} = \begin{cases} e^{x-\theta}, \text{ if } \theta > x \\ 0 \text{ if } 0 < \theta < x \end{cases}$$

i) With respect to quadratic loss: The Bayesian esimator $\delta_\tau(x)$ is given by:

$$\delta_\tau(x) = \int_x^\infty \theta h(\theta|x)d\theta = \int_x^\infty \theta e^{x-\theta} d\theta = e^x \int_x^\infty \theta e^{-\theta} d\theta = e^x(xe^{-x} + e^{-x}) = x + 1.$$

ii) With respect to absolute value loss: The Bayesian estimator m solves the equation:

$$\int_m^\infty e^{x-\theta} d\theta = \frac{1}{2}$$

and we get: $e^{x-m} = \frac{1}{2} \Longrightarrow m - x = ln2 \Longrightarrow m = x + ln2$.

Question 5 Let $\mathbf{X} = (X_1, \ldots, X_n)$ are the random variables. Setting $\mu_0 = x_0$ for convenience of the notation, we can write:

$$h(\mu|\mathbf{X}{=}\mathbf{x}) \propto e^{-\frac{1}{2}\sum_{i=0}^n (x_i-\mu)^2} \propto e^{-\frac{n+1}{2}[\mu^2 - 2\mu\frac{\sum_{i=0}^n x_i}{n+1}]}$$

Of course this also means (by completing the square with expression that does not depend on μ)

$$h(\mu|\mathbf{X}{=}\mathbf{x}) \propto e^{-\frac{n+1}{2}[\mu - \frac{\sum_{i=0}^n x_i}{n+1}]^2}$$

which implies that $h(\mu|\mathbf{X}{=}\mathbf{x})$, (being a density), MUST be the density of $N(\frac{\sum_{i=0}^n x_i}{n+1}, \frac{1}{n+1})$. Hence, the Bayes estimator (being the posterior mean) would be

$$(\sum_{i=0}^n x_i)/(n+1) = (\mu_0 + \sum_{i=1}^n x_i)/(n+1) = \frac{1}{n+1}\mu_0 + \frac{n}{n+1}\bar{X},$$

that is, the Bayes estimator is a convex combination of the mean of the prior and of $\bar{X}$. In this combination, the weight of the prior information diminishes quickly when the sample size increases. The **same** estimator is obtained with respect to absolute value loss.

Question 6i) $X \sim Bin(5, \theta)$. We have:

$$P(X = 0|\theta) = (1-\theta)^5,$$

which means that the posterior of θ given the sample is $\propto (1-\theta)^5\theta(1-\theta)^4 = \theta(1-\theta)^9$. Hence

$$h(\theta|X=0) = 110\theta(1-\theta)^9.$$

(Note: $\frac{\Gamma(12)}{\Gamma(10)\Gamma(2)} = \frac{11!}{9!1!} = 110$.) Then we get for the posterior probability given the sample:

$$P(\theta \in \Theta_0|X=0) = \int_0^{0.2} 110\theta(1-\theta)^9 d\theta = .6779$$

and we **accept** H_0 since the above posterior probability is $> \frac{1}{2}$.

ii) Now

$$P(X=1|\theta) = 5(1-\theta)^4\theta,$$

which implies that the posterior of θ given the sample is $\propto (1-\theta)^4\theta(1-\theta)^4\theta = (1-\theta)^8\theta^2$. Hence

$$h(\theta|X=1) = \frac{\Gamma(12)}{\Gamma(9)\Gamma(3)}(1-\theta)^8\theta^2 = 495\theta^2(1-\theta)^8.$$

Then we get for the posterior probability given the sample:

$$P(\theta \in \Theta_0|X=1) = \int_0^{0.2} 495\theta^2(1-\theta)^8 d\theta = .3826 < \frac{1}{2}.$$

and we **reject** H_0 since the above posterior probability is $< \frac{1}{2}$.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Solutions to selected exercises for MATH5905, Statistical Inference

Part two: Data reduction. Sufficient statistics. Classical estimation

Question 1 a) Denoting $T = \sum_{i=1}^{n} X_i$, you can factorise $L(\mathbf{X}, \mu)$ with

$$h(\mathbf{X}) = exp(-\frac{1}{2}\sum_{i=1}^{n} X_i^2),$$

$$g(T, \mu) = exp(-\frac{n}{2}\mu^2)exp(T\mu)\frac{1}{(\sqrt{2\pi})^n}.$$

b) Denoting $T = \sum_{i=1}^{n} X_i^2$, you can factorise $L(\mathbf{X}, \sigma^2)$ with

$$h(\mathbf{X}) = 1, g(T, \sigma^2) = exp(-\frac{1}{2\sigma^2}T)\frac{1}{(\sqrt{2\pi}\sigma)^n}.$$

c) For a point x and a set A, we use the notation $I_A(x) = I(x \in A) = \left\{ \begin{array}{l} 1 \text{ if } x \text{ is in A,} \\ 0 \text{ if } x \text{ is not in } A \end{array} \right.$ Then

$$L(\mathbf{X}, \theta) = \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) = I_{(\theta,\theta+1)}(x_{(n)})I_{(\theta,\theta+1)}(x_{(1)}) = I_{(x_{(n)}-1,\infty)}I_{(-\infty,x_{(1)})}(\theta).$$

Hence $T = \left\{ \begin{array}{c} X_{(1)} \\ X_{(n)} \end{array} \right\}$ can be taken as sufficient vector-statistic.

d) Denoting $T = \sum_{i=1}^{n} X_i$, you can factorise $L(\mathbf{X}, \lambda)$ with $g(T, \lambda) = exp(-n\lambda)\lambda^T$ and $h(\mathbf{X}) = \frac{1}{\prod_{i=1}^{n} X_i!}$. According to the factorisation criterion, T is sufficient.

Now, using the definition **and** noting that $T = \sum_{i=1}^{n} X_i \sim Po(n\lambda)$ we have:

$$P(\mathbf{X} = \mathbf{x}|T = t) = \frac{P(\mathbf{X}{=}\mathbf{x} \bigcap T = t)}{P(T = t)} = \left\{ \begin{array}{l} 0 \text{ if } \sum_{i=1}^{n} x_i \neq t \\ \frac{P(\mathbf{X}=\mathbf{x})}{P(\sum_{i=1}^{n} X_i = t)} \text{ if} \sum_{i=1}^{n} x_i = t \end{array} \right.$$

Since $\sum_{i=1}^{n} X_i \sim Po(n\lambda)$, the latter expression on the right is easily seen to be equal to $\frac{t!}{n^t \prod_{i=1}^{n} x_i!}$ and obviously does not depend on λ. Hence $T = \sum_{i=1}^{n} X_i$ is sufficient according to the original definition of sufficiency.

Question 2 For $S = X_1 + X_2 + X_3$ we already know ($n = 3$ is a special case of the general case considered at the lecture.) To show that $T = X_1X_2 + X_3$ is **not** sufficient, it suffices to show that, say, $f_{(X_1,X_2,X_3|T=1)}(0, 0, 1|1)$ **does** depend on p. You can easily see that

$$f_{(X_1,X_2,X_3|T=1}(0, 0, 1|1)) = \frac{P(X_1 = 0 \bigcap X_2 = 0 \bigcap X_3 = 1 \bigcap T = 1)}{P(T = 1)} = \frac{(1-p)^2 p}{3p^2(1-p) + p(1-p)^2} = \frac{1-p}{1+2p}$$

Hence $T = X_1X_2 + X_3$ is not sufficient for p.

Question 3 We will show that $T_1 = X_1 + X_2$ is sufficient but $T_2 = X_1X_2$ is **not** sufficient. By a direct check we have

$$P(X_1 = 0 \bigcap X_2 = 0|X_1 + X_2 = 0) = 1,$$

$$P(X_1 = 1 \bigcap X_2 = 0|X_1+X_2 = 0) = P(X_1 = 1 \bigcap X_2 = 1|X_1+X_2 = 0) = P(X_1 = 0 \bigcap X_2 = 1|X_1+X_2 = 0) = 0$$

$$P(X_1 = 1 \bigcap X_2 = 0 | X_1 + X_2 = 1) = \frac{\theta(4-\theta)/12}{\theta(4-\theta)/6} = \frac{1}{2} = P(X_1 = 0 \bigcap X_2 = 1 | X_1 + X_2 = 1)$$

$$P(X_1 = 0 \bigcap X_2 = 0 | X_1 + X_2 = 1) = 0 = P(X_1 = 1 \bigcap X_2 = 1 | X_1 + X_2 = 0)$$

$$P(X_1 = 1 \bigcap X_2 = 1 | X_1 + X_2 = 2) = \frac{\theta(\theta-1)/12}{\theta(\theta-1)/12} = 1$$

$$P(X_1 = 0 \bigcap X_2 = 1 | X_1 + X_2 = 2) = P(X_1 = 1 \bigcap X_2 = 0 | X_1 + X_2 = 2) = 0$$

$$P(X_1 = 0 \bigcap X_2 = 0 | X_1 + X_2 = 2) = 0$$

and we see that in all possible cases the conditional distribution does not involve θ.
However, for $T_2 = X_1 X_2$ we can easily see, following the same pattern, that

$$P(X_1 = 1 \bigcap X_2 = 0 | X_1 X_2 = 0) = \frac{4\theta - \theta^2}{\theta - \theta^2 + 12}.$$

This clearly depends on θ hence T_2 is not sufficient.

Question 4 The conditional probability $P(\mathbf{X}{=}\mathbf{x}|X_1 = x_1)$ is the probability $P(X_2 = x_2 \bigcap \ldots \bigcap X_n = x_n)$ and it clearly depends on p since for each i we have $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$.

Question 5 We need to show that at least in some cases there is explicit dependence of the conditional distribution of the vector $\left\{ \begin{array}{c} X_1 \\ X_2 \end{array} \right\}$ given the statistic $T = X_1 + X_2$. We note that possible realisations of T are $t = 2, 3, \ldots, 2\theta$. We examine $P(\left\{ \begin{array}{c} X_1 \\ X_2 \end{array} \right\} = \left\{ \begin{array}{c} x_1 \\ x_2 \end{array} \right\} | X_1 + X_2 = x)$. Of course, if $x_1 + x_2 \neq x$, this conditional probability is zero and does not involve θ.
Let us study the case $x_1 + x_2 = x$ now. We have two scenarios:
First scenario: $2 \leq x \leq \theta$. Then

$$P(\left\{ \begin{array}{c} X_1 \\ X_2 \end{array} \right\} = \left\{ \begin{array}{c} x_1 \\ x_2 \end{array} \right\} | X_1 + X_2 = x) = \frac{P(X_1 = x_1 \bigcap X_2 = x - x_1)}{\sum_{i=1}^{x-1} P(X_1 = i \bigcap X_2 = x - i)} = \frac{(1/\theta)^2}{(x-1)(1/\theta)^2} = \frac{1}{x-1}$$

which does not involve θ.
Second scenario: $\theta < x \leq 2\theta$. Then:

$$P(\left\{ \begin{array}{c} X_1 \\ X_2 \end{array} \right\} = \left\{ \begin{array}{c} x_1 \\ x_2 \end{array} \right\} | X_1 + X_2 = x) = \frac{P(X_1 = x_1 \bigcap X_2 = x - x_1)}{\sum_{i=x-\theta}^{\theta} P(X_1 = i \bigcap X_2 = x - i)} = \frac{(1/\theta)^2}{(-x + 2\theta + 1)(1/\theta)^2} = \frac{1}{2\theta - x + 1}$$

In the second case, the conditional distribution explicitly involves θ hence $T = X_1 + X_2$ can not be sufficient for θ.

Question 6 Similar to Q4 above. Left as exercise for you.
Question 7 a)

$$\frac{L(\mathbf{x}, \lambda}{L(\mathbf{y}, \lambda} = \lambda^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i} \frac{\prod_{i=1}^n (y_i)!}{\prod_{i=1}^n (x_i)!}$$

and this would not depend on λ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Hence $T = \sum_{i=1}^n X_i$ is minimal sufficient.
b)

$$\frac{L(\mathbf{x}, \sigma^2)}{L(\mathbf{y}, \sigma^2)} = exp(-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2)).$$

This would not depend on σ^2 iff $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$. Hence $T(\mathbf{X}) = \sum_{i=1}^n X_i^2$ is minimal sufficient.
c) Similarly, $\rightarrow T = \prod_{i=1}^n X_i$ is minimal sufficient. (We can also take $\tilde{T} = \sum_{i=1}^n logX_i$ as minimal sufficient).

d) We have $\frac{L(\mathbf{X},\theta)}{L(\mathbf{Y},\theta)} = \frac{I_{(x_{(n)},\infty)}(\theta)}{I_{(y_{(n)},\infty)}(\theta)}$. This has to be considered as a function of θ for fixed $x_{(n)}$ and $y_{(n)}$. Assume that $x_{(n)} \neq y_{(n)}$ and, to be specific, let $x_{(n)} > y_{(n)}$ first. Then the ratio $\frac{L(\mathbf{X},\theta)}{L(\mathbf{Y},\theta)}$ is:

- not defined if $\theta \leq y_{(n)}$,
- equal to zero when $\theta \in [y_{(n)}, x_{(n)})$.
- equal to one when $\theta > x_{(n)}$.

In other words, the ratio's value depends on the position of θ on the real axis, that is, it is a function of θ. Similar conclusion will be reached if we had $x_{(n)} < y_{(n)}$ (do it yourself). Hence, iff $x_{(n)} = y_{(n)}$ will the ratio not depend on θ. This implies that $T = X_{(n)}$ is minimal sufficient.

e) $T = (X_{(1)}, X_{(n)})$ is minimal sufficient. We know from 1c) that $L(\mathbf{x}, \theta)$ depends on the sample via $x_{(n)}$ and $y_{(n)}$ only. If $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are such that either $x_{(1)} \neq y_{(1)}$ or $x_{(n)} \neq y_{(n)}$ or both then $\frac{L(\mathbf{x},\theta)}{L(\mathbf{y},\theta)}$ will have different values in different intervals, that is, will depend on θ. For this **not** to happen, $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$ must hold.

f) Similar to e). $T = (x_{(1)}, x_{(n)})$ is minimal sufficient.

8) a)$L(\mathbf{x}, \theta) = \theta^n (\prod_{i=1}^n x_i)^{\theta - 1}$ and we see by the factorisation criterion that $T = \prod_{i=1}^n x_i$ is sufficient. Note that $\tilde{T} = \sum_{i=1}^n log x_i$ is also sufficient since it is an 1-to-1 transformation of T.
b) $L(\mathbf{x}, \theta) = \frac{1}{(6\theta^4)^n} (\prod_{i=1}^n x_i^3) e^{-(\sum_{i=1}^n x_i)/\theta}$. We can factorise with $h(\mathbf{x}) = \prod_{i=1}^n x_i^3$, $g(t, \theta) = \frac{1}{(6\theta^4)^n} e^{-t/\theta}$, where $t = \sum_{i=1}^n x_i$.

Questions 9, 10 left for you as exercises. I have treated the location case for the Cauchy family in the lectures, the scale case is along the same lines.

Question 11 All of these are easy, the answers are given to you and you should be able to get them.

Question 12 Take $\hat{\tau} = I_{\{X_1 = 0 \bigcap X_2 = 0\}}(\mathbf{X})$. Then obviously $E\hat{\tau} = e^{-2\lambda}$ (that is, $\hat{\tau}$ is unbiased for $\tau(\lambda) = e^{-2\lambda}$). Then the UMVUE would be $E(\hat{\tau} | \sum_{i=1}^n X_i = t) = 1 * P(\hat{\tau} = 1 | \sum_{i=1}^n X_i = t)$. We know that $\sum_{i=1}^n X_i \sim Po(n\lambda)$. The unbiased estimate is

$$a(t) = \frac{P(X_1 = 0 \bigcap X_2 = 0 \bigcap \sum_{i=1}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)} = \frac{P(X_1 = 0 \bigcap X_2 = 0 \bigcap \sum_{i=3}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)} =$$

$$\frac{(n-2)^t}{n^t} = (1 - \frac{2}{n})^t.$$

We can check directly that this estimator is unbiased for $\tau(\lambda)$ (*although this is not necessary: we have stated a general theorem that Rao-Blackwellization preserves the unbiasedness property. I have included the calculations below just as an additional exercise:*

$$Ea(T) = \sum_{t=0}^{\infty} (1 - \frac{2}{n})^t \frac{e^{-n\lambda}(n\lambda)^t}{t!} = e^{-n\lambda} \sum_{t=0}^{\infty} \frac{[\lambda(n-2)]^t}{t!} = e^{-2\lambda}.$$

The variance given by the Cramer-Rao lower bound is:

$$\frac{(\tau'(\lambda))^2}{n I_{X_1}(\lambda)} = \frac{\lambda(-2e^{-2\lambda})^2}{n} = \frac{4\lambda e^{-4\lambda}}{n}$$

For the variance of the unbiased estimator, we have:

$$V(a(T)) = \sum_{t=0}^{\infty} (1 - \frac{2}{n})^{2t} \frac{e^{-n\lambda}(n\lambda)^t}{t!} - (e^{-2\lambda})^2 = e^{-n\lambda} \sum_{t=0}^{\infty} \frac{(n-2)^{2t}\lambda^t}{n^t t!} - e^{-4\lambda} =$$

$$e^{-n\lambda} e^{(n-4+\frac{4}{n})\lambda} - e^{-4\lambda} = e^{-4\lambda}[e^{4\lambda/n} - 1] > 0.$$

The latter value is strictly larger than the bound:

$$e^{-4\lambda}[e^{4\lambda/n} - 1 - \frac{4\lambda}{n}] = e^{-4\lambda}(\frac{1}{2!}(\frac{4\lambda}{n})^2 + \frac{1}{3!}(\frac{4\lambda}{n})^3 + \ldots) > 0.$$

Question 13 This is again just to refresh some required, useful technical skills.

$$f_X(x) = \int_0^x 8xydy = 4x^3 \text{ if x in (0,1) (and zero else)}$$

$$f_Y(y) = \int_y^1 8xydx = 4y - 4y^3 \text{ if y in (0,1) (and zero else)}$$

$$f_{Y|X}(y|x) = \frac{8xy}{4x^3} = \frac{2y}{x^2} \text{ if } 0 < y < x, 0 < x < 1 \text{ (and zero else)}$$

$$a(x) = E(Y|X = x) = \int_0^x y f_{Y|X}(y|x)dy = \frac{2x}{3}, 0 < x < 1$$

$$E(a(X)) = \int_0^1 a(x) f_X(x)dx = \int_0^1 \frac{2x}{3} 4x^3 dx = \frac{8}{15}$$

$$EY = 4\int_0^1 y(y - y^3)dy = \frac{8}{15}$$

Similarly $Ea^2(x) = \frac{8}{27}, V(a(X)) = \frac{8}{27} - (\frac{8}{15})^2 = \frac{8}{675}$

$$E(Y^2) = \frac{1}{3}, V(Y) = \frac{11}{225}$$

and we see directly that indeed $V(a(X)) < V(Y)$ holds.
Again note that the fact that by conditioning we reduce the variance was proved quite generally in the lectures. In this problem we are just checking that indeed $V(a(X)) < V(Y)$ on a particular example.

Question 14) Steps:

a) $T = \sum_{i=1}^n X_i$ is complete and sufficient for θ.

b) If $\hat{\tau} = X_1X_2$ then $E\hat{\tau} = \theta^2$ (that is, $\hat{\tau}$ is unbiased for θ^2).

c) $a(t) = E(\hat{\tau}|T = t) = \ldots = \frac{t(t-1)}{n(n-1)}$ which is the UMVUE.

We can also check directly the unbiasedness of this estimator:

$$E(a(T)) = E[\bar{X}(\frac{n}{n-1}\bar{X} - \frac{1}{n-1})] = \frac{n}{n-1}E(\bar{X})^2 - \frac{E(\bar{X})}{n-1} =$$

$$\frac{n}{n-1}[Var(\bar{X}) + (E(\bar{X}))^2] - \frac{\theta}{n-1} = \frac{n}{n-1}(\frac{\theta(1-\theta)}{n} + \theta^2) - \frac{\theta}{n-1} = \theta^2.$$

Question 15 $f(x;\theta)$ is an **one-parameter exponential** family, with $d(x) = x$. Using our general statement from the lecture, we can claim that $T = \sum_{i=1}^n X_i$ is **complete and minimal sufficient** for θ. We also know that for this distribution $E(X_1) = \theta, Var(X_1) = \theta^2$ holds. Let us calculate:

$$E(\bar{X}^2) = Var(\bar{X}) + (E(\bar{X}))^2 = \frac{Var(X_1)}{n} + (EX_1)^2 = \frac{n+1}{n}\theta^2 \neq \theta^2.$$

After bias-correction, by Lehmann-Scheffe's theorem:

$$\frac{n(\bar{X})^2}{n+1} = \frac{T^2}{n(n+1)}$$

is unbiased for θ **and** since T is complete and sufficient, we conclude that $\frac{T^2}{n(n+1)}$ is UMVUE for θ^2.

Question 16 a) $T = X_{(n)}$ is complete and sufficient for θ, with $f_T(t) = \frac{nt^{n-1}}{\theta^n}, 0 < t < \theta$. Hence $ET^2 = \frac{n}{n+2}\theta^2$. Hence $T_1 = \frac{n+2}{n}T^2$ is unbiased estimator of θ^2. By Lehmann-Scheffe, $\frac{n+2}{n}T^2$ is the UMVUE.

Its variance:

$$E(\frac{n+2}{n}T^2)^2 - \theta^4 = (\frac{n+2}{n})^2 ET^4 - \theta^4 = (\frac{n+2}{n})^2 n\int_0^\theta \frac{t^{n+3}}{\theta^n}dt - \theta^4 =$$

$$\theta^4[\frac{(n+2)^2}{n}\frac{1}{n+4}-1]=\frac{4\theta^4}{n(n+4)}.$$

b) Similar to a). $\frac{n-1}{n}\frac{1}{T}$ is the UMVUE; its variance is $\frac{1}{n(n-2)\theta^2}$.

Question 17 This is a *more difficult (*) question.* It is meant to challenge the better students. Do not be too upset if you have a difficulty with it.

a) The density $f(t;\theta)$ in 7a) is also called $Gamma(n,\theta)$ **density**. To show the result,we could use convolution. Reminder: the **convolution formula** for the density of the sum of two independent random variables X, Y :

$$f_{X+Y}(t)=\int_{-\infty}^{\infty}f_X(x)f_Y(t-x)dx$$

In particular, if the random variables are non-negative, the above formula simplifies to:

$$f_{X+Y}(t)=\int_0^t f_X(x)f_Y(t-x)dx,\ \text{if } t>0 \text{ (and 0 elsewhere).}$$

Applying it for the two non-negative random variables in our case, we get:

$$f_{X_1+X_2}(t)=\theta^2\int_0^t e^{-\theta x}e^{-t\theta+\theta x}dx=\theta^2e^{-t\theta}\int_0^t dx=\theta^2te^{-t\theta}.$$

which means that for $n=2$ the claim is proved (note that $\Gamma(2)=1$.) We apply **induction** to show the general case. Assume that for $T=\sum_{i=1}^k X_i$, the formula is also true and we want to show that then it is true for $k+1$. We apply for $\sum_{i=1}^{k+1}X_i=\sum_{i=1}^k X_i+X_{k+1}$ the convolution formula and we get:

$$f_{\sum_{i=1}^{k+1}X_i}(t)=\frac{t^k\theta^{k+1}e^{-\theta t}}{\Gamma(k+1)},$$

that is, the claim is true for $k+1$.

(**Note:** It is possible to give an alternative proof by using the moment generating functions approach. Try it if you feel familiar enough with moment generating functions.)

b) Consider the estimator $\hat{\tau}=I_{\{X_1>k\}}(\mathbf{X})$. Obviously, $E\hat{\tau}=1*P(X_1>k)=\int_k^\infty\theta e^{-\theta x}dx=e^{-k\theta}$.

c) $T=\sum_{i=1}^n X_i$. Consider for small enough Δx_1 :

$$f_{X_1|T}(x_1|t)\Delta x_1=\frac{f_{X_1,T}(x_1,t)\Delta x_1\Delta t}{f_T(t)\Delta t}\approx$$

$$\frac{P[x_1<X_1<x_1+\Delta x_1;t<\sum_{i=1}^n X_i<t+\Delta t]}{\frac{1}{\Gamma(n)}\theta^n t^{n-1}e^{-\theta t}\Delta t}\approx$$

$$\frac{P[x_1<X_1<x_1+\Delta x_1;t-x_1<\sum_{i=2}^n X_i<t-x_1+\Delta t]}{\frac{1}{\Gamma(n)}\theta^n t^{n-1}e^{-\theta t}\Delta t}\approx$$

$$\frac{P(x_1<X_1<x_1+\Delta x_1)P(t-x_1<\sum_{i=2}^n X_i<t-x_1+\Delta t)}{\frac{1}{\Gamma(n)}\theta^n t^{n-1}e^{-\theta t}\Delta t}\approx$$

$$\frac{\theta e^{-\theta x_1}\frac{1}{\Gamma(n-1)}\theta^{n-1}(t-x_1)^{n-2}e^{-\theta(t-x_1)}\Delta x_1\Delta t}{\frac{1}{\Gamma(n)}\theta^n t^{n-1}e^{-\theta t}\Delta t}=(n-1)\frac{(t-x_1)^{n-2}}{t^{n-1}}\Delta x_1.$$

Going to the limit as Δx_1 tends to zero, we get

$$f_{X_1|T}(x_1|t)=\frac{n-1}{t}(1-\frac{x_1}{t})^{n-2},0<x_1<t<\infty.$$

Now we can find the UMVUE. It will be:

$$E(I_{(k,\infty)}(X_1)|T=t)=\int_k^\infty f_{X_1|T}(x_1|t)dx_1=\int_k^t\frac{n-1}{t^{n-1}}(t-x_1)^{n-2}dx_1=(\frac{t-k}{t})^{n-1}.$$

That is,

$$(\frac{T-k}{T})^{n-1} I_{(k,\infty)}(T)$$

with $T = \sum_{i=1}^{n} X_i$ is the UMVUE of $e^{-k\theta}$.

Question 18 The restriction $\theta \in (0, 1/5)$ makes sure that the probabilities calculated as a function of θ indeed belong to $[0, 1]$. Let $E_\theta h(X) = 0$ for all $\theta \in (0, 1/5)$. This means:

$$h(0)2\theta^2 + h(1)(\theta - 2\theta^3) + h(2)\theta^2 + h(3)(1 + 2\theta^3 - 3\theta^2 - \theta) = 0.$$

We rewrite the above relationship as follows:

$$[2h(3) - 2h(1)]\theta^3 + [2h(0) + h(2) - 3h(3)]\theta^2 + [h(1) - h(3)]\theta + h(3) = 0$$

for all $\theta \in (0, 1/5)$. The main theorem of algebra implies then that the coefficients in front of each power of the 3rd order polynomial in θ must be equal to zero. Hence $h(3) = 0 \Longrightarrow h(1) - h(3) = 0 \Longrightarrow h(1) = 0 \Longrightarrow 2h(0) + h(2) = 0$. The latter relationship **does not** necessarily imply that both $h(0) = 0, h(2) = 0$ must hold. Hence the family of distributions is **not** complete.

Question 19 Parts 19a), 19b), 19c) were treated in lecture and are easy. We consider 19d) here. We have to show that $T = X_{(n)}$ is complete. We know that the density of T is

$$f_T(t) = nt^{n-1}/\theta^n, 0 < t < \theta \text{ (and 0 else)}.$$

Let $E_\theta g(T) = 0$ for all $\theta > 0$. This implies:

$$\int_0^\theta g(t)\frac{nt^{n-1}}{\theta^n}dt = 0 = \frac{1}{\theta^n}\int_0^\theta g(t)nt^{n-1}dt$$

for all $\theta > 0$ must hold. Since $\frac{1}{\theta^n} \neq 0$ we get $\int_0^\theta g(t)nt^{n-1}dt = 0$ for all $\theta > 0$. Differentiating both sides with respect to θ we get

$$ng(\theta)\theta^{n-1} = 0$$

for all $\theta > 0$. This implies $g(\theta) = 0$ for all $\theta > 0$. This also means $P_\theta(g(T) = 0) = 1$. In particular, this result implies that $S = \frac{n+1}{n}X_{(n)}$ is the UMVUE of $\tau(\theta) = \theta$ in this model since $E_\theta S = \theta$ holds (see previous lectures) **and** S is a function of sufficient **and** complete statistic.

Question 20 We have

$$L(\mathbf{X},\mathbf{Y};\mu_1,\sigma_1^2,\mu_2,\sigma_2^2) = \frac{1}{(\sqrt{2\pi})^n\sigma_1^{n_1}\sigma_2^{n_2}}e^{\{-\frac{1}{2}\sum_{i=1}^{n_1}\frac{(x_i-\mu_1)^2}{\sigma_1^2}-\frac{1}{2}\sum_{i=1}^{n_2}\frac{(y_i-\mu_2)^2}{\sigma_2^2}\}}$$

$$lnL = -nln(\sqrt{2\pi}) - n_1\sigma_1 - n_2\sigma_2 - \frac{1}{2}\sum_{i=1}^{n_1}\frac{(x_i-\mu_1)^2}{\sigma_1^2} - \frac{1}{2}\sum_{i=1}^{n_2}\frac{(y_i-\mu_2)^2}{\sigma_2^2}$$

Solving the equation system

$$\frac{\partial}{\partial\mu_1}lnL = 0 \& \frac{\partial}{\partial\mu_2}lnL = 0$$

delivers $\hat{\mu_1} = \bar{X}_{n_1}, \hat{\mu_2} = \bar{Y}_{n_2}$ for the MLE. Using the transformation invariance property, we get $\hat{\theta} = \bar{X}_{n_1} - \bar{Y}_{n_2}$ for the maximum likelihood estimator of θ. Further:

$$Var(\hat{\theta}) = Var(\bar{X}_{n_1}) + Var(\bar{Y}_{n_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n-n_1} = f(n_1).$$

To find the minimum, we set the derivative with respect to n_1 to be equal to zero and solve the resulting equation. This gives: $\frac{\sigma_1}{\sigma_2} = \frac{n1}{n2}$. With other words, the sample sizes must be proportional to the standard deviations. In particular, if n is fixed, we get $n_1 = \frac{\sigma_1}{\sigma_1+\sigma_2}n$.

Question 21 i) $L(\mathbf{X};\theta) = \theta^n \prod_{i=1}^n x_i^{-2} I_{[\theta,\infty)}(x_{(1)})$. We consider L as a function of theta after the sample has been substituted. When θ moves on the positive half-axis, this function first grows monotonically (when θ moves between 0 and $x_{(1)}$) and then drops to zero onward since the indicator becomes equal to zero. Hence L is a discontinuous function of θ and its maximum is attained at $x_{(1)}$. This means that $\hat{\theta}_{mle} = X_{(1)}$.

ii) Using the factorisation criterion, we see that $X_{(1)}$ is sufficient. It is also minimal sufficient due to dimension considerations. The minimal sufficiency can also be shown by directly examining the ratio $\frac{L(\mathbf{X};\theta)}{L(\mathbf{Y};\theta)}$.

Question 22 a)

$$L(\mathbf{X};\theta) = \theta^n (\prod_{i=1}^n x_i)^{\theta-1}.$$

$$lnL(\mathbf{X};\theta) = nln\theta + (\theta-1)\sum_{i=1}^n lnx_i.$$

$$\frac{\partial}{\partial\theta} lnL = \frac{n}{\theta} + \sum_{i=1}^n lnx_i = 0$$

gives the root $\hat{\theta} = \hat{\theta}_{mle} = \frac{-n}{\sum_{i=1}^n lnx_i}$. Then, using the translation invariance property, we get

$$\hat{\tau(\theta)} = \frac{\hat{\theta}}{\hat{\theta}+1}.$$

b)$\sqrt{n}(\hat{\theta} - \theta \to^d N(0, \frac{1}{I_{X_1}(\theta)})$. We need to find $I_{X_1}(\theta)$. To this end, we take:

$$lnf(x;\theta) = ln\theta + (\theta-1)lnx; \frac{\partial}{\partial\theta} lnf(x;\theta) = \frac{1}{\theta} + lnx; \frac{\partial^2}{\partial\theta^2} lnf(x;\theta) = -\frac{1}{\theta^2}.$$

This means that $I_{X_1}(\theta) = \frac{1}{\theta^2}$ and $\sqrt{n}(\hat{\theta} - \theta) \to^d N(0, \theta^2)$.

Since $\tau(\theta) = \frac{\theta}{\theta+1}$, by applying the delta method we get

$$\sqrt{n}(\hat{\tau} - \tau) \to^d N(0, \frac{\theta^2}{(1+\theta)^4}).$$

c) According to the factorisation criterion, $\prod_{i=1}^n X_i$ is sufficient (also, $\sum_{i=1}^n lnX_i$ is sufficient). Since the density belongs to an one-parameter exponential (WHY(!)) we do have completeness, as well.

$T = \sum_{i=1}^n X_i$ is **not** sufficient. Consider for example $0 < t < 1, n = 2, T = X_1 + X_2$. Using the convolution formula (see previous tutorial sheet) we have:

$$f_{X_1+X_2}(t) = \theta^2 \int_0^t x^{\theta-1}(t-x)^{\theta-1} dx.$$

Changing the variables: $x = ty, dx = tdy$, we can continue to obtain:

$$f_{X_1+X_2}(t) = t^{2\theta-1}\theta^2 \int_0^1 y^{\theta-1}(1-y)^{\theta-1} dy = t^{2\theta-1}\theta^2 B(\theta,\theta).$$

Then the conditional density becomes:

$$f_{(X_1,X_2)|T}(x_1,x_2|t) = \frac{\theta^2(x_1x_2)^{\theta-1}}{t^{2\theta-1}\theta^2 B(\theta,\theta)}$$

(if $x_1 + x + 2 = t$, and, of course, zero elsewhere). Hence the conditional density of the sample given the value of the statistic does depend on the parameter.

d)Looking at $\frac{\partial}{\partial\theta} lnL = -n(\frac{-\sum_{i=1}^n lnx_i}{n} - \frac{1}{\theta})$ we see that for $\frac{1}{\theta}$ the CRLB will be attained. This means that $\frac{1}{\theta}$ can be estimated by the UMVUE $T = -\frac{\sum_{i=1}^n lnX_i}{n}$. The attainable bound is easily seen to be

$\frac{1}{n\theta^2}$.

Question 23 a) The density of a single observation is $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ where only σ^2 is assumed unknown. Then

$$lnL(\mathbf{X};\sigma^2) = -nln((\sqrt{2\pi}) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}$$

Then the equation

$$\frac{\partial}{\partial\sigma^2}lnL = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^4} = 0$$

has a root $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ which is also the MLE (Why(!)).
Further,

$$lnf(x;\mu,\sigma^2) = -ln((\sqrt{2\pi}) - \frac{1}{2}ln(\sigma^2) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2},$$

$$\frac{\partial}{\partial\sigma^2}lnf = -\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(x-\mu)^2}{\sigma^4},$$

$$\frac{\partial^2}{\partial\sigma^2\partial\sigma^2}lnf = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

Taking $-E(\ldots)$ in the last equation gives $I_{X_1}(\sigma^2) = \frac{1}{2\sigma^4}$. Hence:

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \longrightarrow^d N(0, 2\sigma^4).$$

b) We apply the delta method. First, we notice that $\hat{\sigma}_{mle} = \hat{\sigma} = \sqrt{\hat{\sigma^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}$ is the MLE (due to the transformation invariance property). Now:

$$\sqrt{n}(\hat{\sigma} - \sigma) \longrightarrow^d N(0, ((\frac{\partial}{\partial\sigma^2}h)^2 2\sigma^4)$$

where $h(\sigma^2) = \sqrt{\sigma^2}$. Hence $\frac{\partial}{\partial\sigma^2}h(\sigma^2) = \frac{1}{2\sigma}$ and we get, after substitution:

$$\sqrt{n}(\hat{\sigma} - \sigma) \longrightarrow^d N(0, \sigma^2/2).$$

Question 24 a) i) The MLE of λ is $\bar{X}$ hence of $\tau(\lambda) = \frac{1}{\lambda}$ would be $\hat{\tau} = \frac{1}{\bar{X}}$.
ii) Since $P(\bar{X} = 0) > 0$, we get that even the first moment is infinite (not to mention the second) and there is no finite variance.
iii) The delta method gives us:

$$\sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\lambda}) \longrightarrow^d N(0, \frac{1}{\lambda^4}I^{-1}_{X_1}(\lambda)))$$

(since in our case $h(\lambda) = \frac{1}{\lambda}$, $\frac{\partial}{\partial\lambda}h(\lambda) = -\frac{1}{\lambda^2}$.) But, as you can easily see (and we discussed at lectures), for $Po(\lambda)$, we have $I_{X_1}(\lambda) = \frac{1}{\lambda}$, therefore

$$\sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\lambda}) \longrightarrow^d N(0, \frac{1}{\lambda^3}).$$

(Comparing the outcomes in (ii) and (iii) we see that although the finite variance does not exist, the asymptotic variance is well defined ($= \frac{1}{\lambda^3}$.))

b) i) $\sqrt{\bar{X}}$ is the MLE and, using the delta method, we get

$$\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \longrightarrow^d N(0, (\frac{1}{2\sqrt{\lambda}})^2\lambda) = N(0, \frac{1}{4}).$$

(Since the asymptotic variance becomes constant ($= \frac{1}{4}$) and does not depend on the parameter, we call the transformation $h(\lambda) = \sqrt{\lambda}$ a **variance stabilising transformation**).

ii) $\sqrt{\bar{X}} \pm \frac{z_{\alpha/2}}{2\sqrt{n}}$ would be the confidence interval for $\sqrt{\lambda}$ and

$$((\sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}})^2, (\sqrt{\bar{X}} + \frac{z_{\alpha/2}}{2\sqrt{n}})^2)$$

would be the confidence interval for λ.

THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

Solutions to selected exercises for MATH5905, Statistical Inference

Part four: Multinomial distribution. Order statistics. Robustness

1). This is just substitution in the formula.

a) $P(X_1=2,X_2=2,X_3=4)=\frac{8!}{2!2!4!}(.2)^2(.3)^2(.5)^4=0.0945$ The marginal distributions are Binomial which means that $X_2\sim Bin(8,0.3)$ and therefore $E(X_2)=8*.3=2.4, Var(X_2)=8*.3*.7=1.68$. $Cov(X_1,X_3)=-8*(0.2)*(0.5)=-0.8$.

b) $P(X_1=3,X_2=1,X_3=2)=\frac{6!}{3!1!2!}(0.5)^3(0.2)^1(0.3)^2=0.135$.

A little "trick" helps to do calculations quicker: we notice that $P(X_1+X_2)=2=P(X_3=4)$. Since $X_3\sim Bin(6,0.3)$ we get $P(X_1+X_2=2)=\frac{6!}{4!2!}(0.3)^4(0.7)^2=0.059535$.

2).The general formula is

$$f_{X_{(i)}}(x)=\frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x)$$

Here we have $n=4,i=2,f(x)=e^{1-x},x>0$. We get from here $F(x)=\int_1^x e^{1-y}dy=1-e^{1-x},x>1$. Then

$$f_{X_{(2)}}(x)=\frac{4!}{1!2!}(1-e^{1-x})e^{2(1-x)}e^{1-x}=12e^{3(1-x)}(1-e^{1-x}),x>1.$$

3). We use the general formula from Problem 4. Here we have $n=5,i=4,f(x)=\frac{1}{x^2},x>1\rightarrow F(x)=\int_1^x y^{-2}dy=1-\frac{1}{x},x>1$. Hence

$$f_{X_{(4)}}(x)=\frac{5!}{1!3!}(1-\frac{1}{x})^3\frac{1}{x}\frac{1}{x^2}=\frac{20}{x^3}(1-\frac{1}{x})^3,x>1.$$

4). We use the general formula: $n=2,f(y)=\frac{1}{2}e^{-\frac{y-4}{2}},y\geq 4\rightarrow F(y)=1-e^{-\frac{1}{2}(y-4)},y>4$. Hence

$$f_{y_{(1)}}(y)=n[1-F(y)]^{n-1}f(y)=2e^{-\frac{1}{2}(y-4)}\frac{1}{2}e^{-\frac{y-4}{2}}=e^{-(y-4)},y>4.$$

Then

$$E(Y_{(1)})=\int_4^\infty ye^{-(y-4)}dy=\int_4^\infty (y-4)e^{-(y-4)}d(y-4)+4\int_4^\infty e^{-(y-4)}d(y-4)=\Gamma(2)+4=5.$$

5). The general formula gives for the density of the largest order statistic: $g_{X_{(n)}}(x)=nF^{n-1}(x)f(x)$.

a) Here $f(x)=e^{-x},x>0\rightarrow F(x)=1=e^{-x}$. We get: $f_{X_{(3)}}(x)=3(1-e^{-x})^2e^{-x}$. Then we can get the expected value:

$$EX_{(3)}=3\int_0^\infty xe^{-x}(1-e^{-x})^2dx=3\int_0^\infty xe^{-x}(1-2e^{-x}+e^{-2x})dx=$$

$$3[\Gamma(2)-\frac{2}{4}\int_0^\infty 2xe^{-2x}d(2x)+\frac{1}{9}\int_0^\infty 3xe^{-3x}d(3x)]=3\Gamma(2)(1-\frac{1}{2}+\frac{1}{9})=\frac{11}{6}.$$

b) This part is more "tricky" and uses some specific properties of the CDF and the density of the standard normal distribution. We start with the general statement:

$EX_{(3)}=3\int_{-\infty}^\infty xF^2(x)f(x)dx$ where $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},F(x)=\int_{-\infty}^x f(u)du$. We note that $f(-x)=f(x),F(-x)=1-F(x)$ holds. Hence

$$EX_{(3)}=3[\int_{-\infty}^0 ..+\int_0^\infty ...]=3\int_0^\infty (-u)F^2(-u)f(u)du+3\int_0^\infty uF^2(u)f(u)du=$$

$$3\int_0^\infty u[F^2(u)-(1-F(u))^2]f(u)du = 3\int_0^\infty u(2F(u)-1)f(u)du = 6\int_0^\infty uF(u)f(u)du - 3\int_0^\infty uf(u)du.$$

Now we note that $f'(u) = -uf(u)$ holds and therefore $\int uf(u)du = -f(u)$. We get then:

$$EX_{(3)} = -6\int_0^\infty F(u)df(u) + 3[f(\infty)-f(0)] = 6\frac{1}{2}\frac{1}{\sqrt{2\pi}} + 6\int_0^\infty \frac{e^{-u^2}}{2\pi}du - \frac{3}{\sqrt{2\pi}} =$$

$$\frac{6\sqrt{\frac{1}{2}}}{\sqrt{2\pi}}\int_0^\infty \frac{e-\frac{u^2}{2\frac{1}{2}}}{\sqrt{2\pi}\sqrt{\frac{1}{2}}}du$$

But the integral is equal to 1/2 (WHY!), hence $EX_{(3)} = \frac{3}{2\sqrt{\pi}}$.
6). a) $X = min(Y_1, Y_2)$. Hence $f_X(x) = 2e^{-\frac{x}{100}}\frac{1}{100}e^{-\frac{x}{100}} = \frac{1}{50}e^{-\frac{x}{50}}$.
b) $X = max(Y_1, Y_2)$. Hence $f_X(x) = 2(1-e^{-\frac{x}{100}})\frac{1}{100}e^{-\frac{x}{100}}$.

7). a) It holds $g(x_{(1)}, x_{(n)}) = n(n-1)(F(x_{(n)} - F(x_{(1)})^{n-2}f(x_{(1)})f(x_{(n)}),\ x_{(1)} < x_{(n)}$. In the particular case considered here, we have $f(x) = e^{-x}, F(x) = 1-e^{-x}, x > 0$. Since $n = 3$, we get:

$$g(x_{(1)}, x_{(3)}) = 6(e^{-x_{(1)}} - e^{-x_{(3)}})e^{-x_{(1)}}e^{-x_{(3)}}$$

for $0 < x_{(1)} < x_{(3)} < \infty$.
b) We can integrate out the unwanted variable $x_{(3)}$ in the joint density from a) to get the marginal of $x_{(1)}$:

$$g(x_{(1)}) = \int_{x_{(1)}}^\infty 6(e^{-x_{(1)}} - e^{-x_{(3)}})e^{-x_{(1)}}e^{-x_{(3)}}dx_{(3)} = \dots$$

but, of course, we could also use directly the formula for the marginal density:

$$g(x_{(1)}) = 3[1-(1-e^{-x_{(1)}})]^2e^{-x_{(1)}} = 3e^{-3x_{(1)}}, x_{(1)} > 0.$$

Similarly, we could integrate out the $x_{(1)}$ variable and get the marginal of $x_{(3)}$. But, of course, we could directly use the formula

$$g(x_{(3)}) = 3[1-e^{-x_{(3)}})]^2e^{-x_{(3)}}, 0 < x_{(3)} < \infty.$$

c) $EX_{(1)} = \int_0^\infty x * 3e^{-3x}dx = 1/3$,

$$EX_{(3)} = \int_0^\infty x * 3e^{-3x}(1-2e^{-x}+e^{-2x})dx = 3(\Gamma(2) - \frac{1}{2}\Gamma(2) + \frac{1}{9}\Gamma(2)) = \frac{11}{6}.$$

d) We define the transform:

$$U = X_{(3)} - X_{(1)}, V = X_{(1)}.$$

The joint density of $X_{(1)}$ and $X_{(3)}$ is (see a)):

$$g(x_{(1)}, x_{(3)}) = 6(e^{-x_{(1)}} - e^{-x_{(3)}})e^{-x_{(1)}}e^{-x_{(3)}}, 0 < x_{(1)} < x_{(3)} < \infty.$$

We get: $x_{(3)} = u+v, x_{(1)} = v$ with a Jacobian of the transformation equal to (-1). We get the joint density

$$f_{(U,V)}(u,v) = 6(e^{-u} - e^{-(u+v)})e^{-v}e^{-(u+v)} * |-1|$$

The relationship $0 < x_{(1)} < x_{(3)} < \infty$ transfers into $0 < v < u+v < \infty$ which is equivalent to $0 < u < \infty, 0 < v < \infty$. Hence the density of the range $R = U = X_{(3)} - X_{(1)}$ is

$$f_R(u) = \int_0^\infty (6e^{-3v-u} - 6e^{-2u-3v})dv = 2(1-e^{-u})e^{-u}, u > 0.$$

8). The transformation $V = X_{(3)}, U = X_{(3)} - X_{(1)}$ has an inverse defined as $x_{(1)} = v - u, x_{(3)} = v$. The absolute value of the Jacobian is 1. Hence

$$f_{(U,V)}(u,v) = n(n-1)[F(v) - F(v-u)]^{n-2} f(v-u) f(v)$$

where $F(.)$ denotes the cdf of a single observation. The region $0 < x_{(1)} < x_{(3)} < 1$ is transformed into $0 < u < v < 1$ for the new variables. Hence we get for the density $f_R(.)$ of the range:

$$f_R(u) = \int_u^1 6[v^2 - (v-u)^2]2(v-u)2vdv = 24\int_u^1 (-u^2 + 2uv)(v^2 - uv)dv = \ldots = 12u(1-u)^2, 0 < u < 1.$$

Try to get the same result by using the transform $V = X_{(1)}, U = X_{(3)} - X_{(1)}$.

9). This problem is a bit more technical. Let $\phi(.)$ denote the standard normal density. We transform: $-\infty < X_{(1)} < X_{(2)} < \infty$ into $X_{(2)} - X_{(1)} = U, X_{(1)} = V$. The region for U and V becomes: $-\infty < v < \infty, 0 < u < \infty$. The joint density $h(u,v)$ of (U,V) becomes:

$$h(u,v) = 2\phi(v)\phi(u+v) = \frac{1}{\pi} e^{-\frac{v^2}{2} - \frac{(u+v)^2}{2}}$$

Hence, for the density $f_R(u)$ of the range we get:

$$f_R(u) = \frac{1}{\pi} e^{-\frac{u^2}{2}} \int_{-\infty}^{\infty} e^{-v^2 - uv} dv$$

Completing the square, we get finally:

$$f_R(u) = \frac{1}{\sqrt{2}}\sqrt{2\pi}\frac{1}{\pi} e^{-\frac{u^2}{2} + \frac{u^2}{4}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}/\sqrt{2}} e^{-\frac{1}{2*\frac{1}{2}}(v + \frac{u}{2})^2} dv = \frac{1}{\sqrt{\pi}} e^{-u^2/4}, u > 0.$$

(The integral above is equal to one since it is if fact the integral of the $N(-\frac{u}{2}, \frac{1}{2})$ density .)
10) a) If X is a random variable with density $f(x)$ we have to show that

$$\frac{E\Psi^2(X)}{[E\Psi'(X)]^2} \geq \frac{1}{E[\frac{f'}{f}(X)]^2}$$

holds. Since

$$E\Psi'(X) = \int_{-\infty}^{\infty} \Psi'(x) f(x) dx = \int_{-\infty}^{\infty} f(x) d\Psi(x) = -\int_{-\infty}^{\infty} \Psi(x) f'(x) dx$$

we see that we need to show that

$$[\int_{-\infty}^{\infty} \Psi(x) f'(x) dx]^2 \leq E\Psi^2(X) E[[\frac{f'}{f}(X)]^2$$

holds. But indeed, by applying Cauchy-Schwartz Inequality, we have

$$[\int_{-\infty}^{\infty} \Psi(x) f'(x) dx]^2 = [\int_{-\infty}^{\infty} \Psi(x) \frac{f'(x)}{f(x)} f(x) dx]^2 = [E\Psi(X)\frac{f'}{f}(X)]^2 \leq E\Psi^2(X) E[\frac{f'}{f}(X)]^2.$$

b) The equality means that equality in the Cauchy-Schwartz must hold and this means that we must have $\Psi(x) = c\frac{f'(x)}{f(x)}$ with certain constant c (which constant, without loss of generality, can also be set to one). Then $\Psi(x;\theta) = \frac{f'(x-\theta)}{f(x-\theta)} = \frac{\partial}{\partial\theta} lnf(x;\theta)$. Hence the equation $\sum_{i=1}^n \Psi(x_i;\theta) = 0$ that defines the M-estimator is the same as the equation for the score $V(\mathbf{X},\theta) = 0$ that defines the MLE.

11) Start with a Taylor expansion along θ_0 :

$$0 = \sum_{i=1}^n \psi(x_i - \hat{\theta}_M) = \sum_{i=1}^n \psi(x_i - \theta_0) - (\hat{\theta}_M - \theta_0)\sum_{i=1}^n \psi'(x_i - \theta_0) + \ldots$$

where θ_0 is defined as in the formulation of the problem, $\hat{\theta}_M$ is the solution of the M-estimator equation and we ignore the higher order terms. We can now rearrange terms, divide through $\sqrt{n}$ both sides and ignore remainder terms (HOT) of higher order to get

$$\sqrt{n}(\hat{\theta}_M - \theta_0) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i - \theta_0)}{\frac{1}{n} \sum_{i=1}^{n} \psi'(x_i - \theta_0)} + \text{HOT}$$

For the expression on top of the RHS the central limit theorem can be applied to show it converging in law to $N(0, E_{\theta_0}(\psi(X_1 - \theta_0)^2)$.

For the expression on bottom of the RHS, the Law of Large Numbers can be applied to show convergence to $E_{\theta_0}\psi'(X_1 - \theta_0)$ in probability. Hence

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \rightarrow^d N(0, \frac{E_{\theta_0}\psi(X_1 - \theta_0)^2}{[E_{\theta_0}\psi'(X_1 - \theta_0)]^2}).$$

Since we are dealing with a location family, the expression about the variance on the RHS is the same for all values of the location parameter hence it is equal to $\frac{\int \psi^2(x) f(x) dx}{(\int \psi'(x) f(x) dx)^2}$.

Some of my white board writing from week 11

1.) I focused on the relationships between moments and cumulants.

The moment generating function (MGF) $M_X(t) = E(\exp(tX))$ has the obvious property $M_X^{(r)}(t)\big|_{t=0} = E(X^r) = \mu_r'$ with μ_r' being a short-hand notation for the raw moment $E(X^r)$ (as opposed to $\mu_r = E(X-\mu_1')^r$ used to denote the central moments).

We have by simple Taylor expansion then

$$M_X(t) = 1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots + \mu_r' \frac{t^r}{r!} + O(t^{r+1})$$

as $t \to 0$. The cumulant generating function is $K_X(t) = \log M_X(t)$ (i.e. $e^{K_X(t)} = M_X(t)$ holds).

Since $K_X(0) = 0$ we get the Taylor expansion

$$K_X(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots + \kappa_r \frac{t^r}{r!} + O(t^{r+1}).$$

Substituting in the relation we get:

$$e^{\kappa_1 t}\, e^{\kappa_2 \frac{t^2}{2!}}\, e^{\kappa_3 \frac{t^3}{3!}} \cdots = 1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots \quad (1)$$

and expanding the exponents on the LHS we get.

$$\left(1 + \kappa_1 t + \kappa_1^2 \frac{t^2}{2!} + \cdots\right)\left(1 + \kappa_2 \frac{t^2}{2!} + \frac{1}{2!}\left(\kappa_2 \frac{t^2}{2!}\right)^2 + \cdots\right)\left(1 + \kappa_3 \frac{t^3}{3!} + \frac{1}{2!}\left(\kappa_3 \frac{t^3}{3!}\right)^2 + \cdots\right)$$

Equating the coefficients in front of the powers of t in the LHS and RHS of (1) we get:

$\kappa_1 = \mu_1'$

$\kappa_2 + \kappa_1^2 = \mu_2' \rightarrow \kappa_2 = \mu_2' - (\mu_1')^2 = E(X^2) - (EX)^2 = \mathrm{Var}X = \sigma^2$

$\kappa_3 = 2\mu_1'^3 - 3\mu_1'\mu_2' + \mu_3'$

$\kappa_4 = -6(\mu_1')^4 + 12(\mu_1')^2\mu_2' - 3(\mu_2')^2 - 4\mu_1'\mu_3' + \mu_4'$

In particular for $N(0,1)$ we get $\kappa_4 = 0$ since $\mu_4' = 3(\mu_2')^2$ holds for $N(0,1)$

To summarize: - the first cumulant is the first moment, the second cumulant is the variance, the third cumulant is the skewness, the fourth cumulant is the kurtosis.

We also introduce:

- STANDARDIZED SKEWNESS: $\rho_3 = \kappa_3/(\kappa_2^{3/2})$
- Standardized kurtosis $\rho_4 = \kappa_4/(\kappa_2^2)$

These are useful in the forthcoming Edgeworth expansions.

2) Then I spoke about Cramer's condition and its importance and discussed the details in the formulation of Theorem 8.1. I specifically stressed that the constants $C_1(F), C_2(F), C_3(F)$ can be expressed by using ρ_3 and ρ_4. Namely: $C_1(F) = \frac{1}{6}\frac{\kappa_3}{\sigma^3} = \frac{1}{6}\rho_3$; $C_2(F) = \frac{1}{24}\rho_4$

$C_3(F) = \frac{C_1^2(F)}{2} = \frac{\rho_3^2}{72}$.

In section 8.3.2, when discussing formula (19), I also wrote all the first 6 Hermite polynomials explicitly:

$H_1(z) = z$; $H_2(z) = z^2 - 1$; $H_3(z) = z^3 - 3z$; $H_4(z) = z^4 - 6z^2 + 3$,

$H_5(z) = z^5 - 10z^3 + 15z$; $H_6(z) = z^6 - 15z^4 + 45z^2 - 15$

3) Next, I discussed the Cornish-Fisher expansion. I gave heuristic justification to the formula given in Theorem 8.2. It goes as follows:

Since $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \approx$ standard normal, the quantile z_α, which is theoretically defined as the solution of $F_{Z_n}(z_\alpha) = 1-\alpha$, should be in a vicinity of the u_α quantile defined as the solution of $\Phi(u_\alpha) = 1-\alpha$.

I also discussed some "famous" u_α quantiles such as $u_{0.01} = 2.326$, $u_{0.025} = 1.96$, $u_{0.05} = 1.645$, $u_{0.1} = 1.28$

Then the argument, by using Taylor expansion, goes as follows:

$$1-\alpha = F_{Z_n}(z_\alpha) \overset{\text{using Theorem 8.1}}{\approx} \Phi(z_\alpha) - \frac{G_1(F)p_1(z_\alpha)\varphi(z_\alpha)}{\sqrt{n}} - \frac{G_2(F)p_2(z_\alpha)+G_3(F)p_3(z_\alpha)}{n}$$

$$= \Phi(u_\alpha) + \varphi(u_\alpha)(z_\alpha - u_\alpha) + \varphi(u_\alpha)[\text{polynomials containing } z_\alpha, u_\alpha]$$

↑ apply Taylor by expanding "everywhere" around u_α

Since $\Phi(u_\alpha) = 1-\alpha$, we cancel with $(1-\alpha)$ on the LHS and get

$$\varphi(u_\alpha)\underbrace{[\text{some polynomials of } z_\alpha, u_\alpha]}_{H_0} = 0$$

Set this $= 0$ and express z_α by using u_α from the resulting relation. In this way we finally obtain the expression given in Theorem 8.2

$$z_\alpha = u_\alpha + \frac{(u_\alpha^2-1)\rho_3}{6\sqrt{n}} + \frac{(u_\alpha^3-3u_\alpha)\rho_4}{24n} - \frac{(2u_\alpha^3-5u_\alpha)\rho_3^2}{36n} + o(n^{-1})$$

4) I then applied this Theorem to illustrate the power and accuracy of the Cornish-Fisher expansion on an example given on p. 69:

Example from p.69 in detail:

$\frac{W_n}{n}$ is an average of n i.i.d. squared standard normals.
Hence the CLT will give us: (Note: for χ_1^2 r.v. $E\chi_1^2 = 1$, $Var\chi_1^2 = 2$)

$$\frac{\sqrt{n}\left(\frac{W_n}{n} - 1\right)}{\sqrt{2}} = \frac{W_n - n}{\sqrt{2n}}$$ is about standard normal (this is the first order asymptotics). Then since

$$P(W_n < z_\alpha) = P\left(\frac{W_n - n}{\sqrt{2n}} < \frac{z_\alpha - n}{\sqrt{2n}}\right)$$ we know that

$\frac{z_\alpha - n}{\sqrt{2n}}$ should be "close" to u_α so then $z_\alpha = n + \sqrt{2n}\,u_\alpha$ is the first order approximation of the $(1-\alpha)100\%$ quantile of W_n.
To improve it by using higher order Cornish-Fisher expansion we proceed as follows:
The mgf of the χ_n^2 random variable (denoted generically as X here) is known to be $M_X(t) = (1-2t)^{-\frac{n}{2}}$
Hence $K_X(t) = -\frac{n}{2}\log(1-2t)$ and we get:

$$K_X'(t) = \frac{n}{1-2t},\quad K_X''(t) = \frac{2n}{(1-2t)^2},\quad K_X'''(t) = \frac{8n}{(1-2t)^3},\quad K_X^{IV}(t) = \frac{48n}{(1-2t)^4}$$

Hence $K_X'(0) = n$, $K_X''(0) = 2n$, $K_X'''(0) = 8n$, $K_X^{IV}(0) = 48n$
In our case we only need to specialise this for χ_1^2 random variable so we get $K_X'(0) = 1$, $K_X''(0) = 2$, $K_X'''(0) = 8$, $K_X^{IV}(0) = 48$

Then $\rho_3 = \frac{8}{2^{3/2}} = 2\sqrt{2}$, $\rho_4 = \frac{48}{2^2} = 12$

This leads to $$\eta \approx n + \sqrt{2n}\left[u_\alpha + \frac{(u_\alpha^2 - 1)2\sqrt{2}}{6\sqrt{n}} + \frac{(u_\alpha^3 - 3u_\alpha)12}{24n} - \frac{(2u_\alpha^3 - 5u_\alpha)8}{36n}\right]$$

We apply these approximations e.g., for $n = 5$ to get:

$$5 + \sqrt{2*5} * 2.326 = 12.36$$

$$5 + \sqrt{2*5}\left(2.326 + \frac{2.326^2 - 1}{6\sqrt{5}} \cdot 2\sqrt{2}\right) = 15.296$$

$$5 + \sqrt{2*5}\left(2.326 + \frac{2.326^2 - 1}{6\sqrt{5}} 2\sqrt{2} + \frac{(2.326^3 - 3*2.326)12}{24*5} - \frac{(2*2.326^3 - 5*2.326)8}{36*5}\right)$$

$$= 15.16.$$

The true value of the quantile is 15.09 and we see the increasing precision popping up when we increase the order of the approximation.

5.) I then started discussing the idea of the saddlepoint method. I believe that the derivations, as presented in the lecture notes, are detailed enough and I did not write anything specific on the white board. I still need to go through pages 72-73 of this lecture at the beginning of lecture in week 12.

Some of my white board writing from week 12

1) I continued the discussion after formula (25) on p. 71 of the notes. I pointed out that when applying this formula for a specific random variable which is the arithmetic mean of n i.i.d. variables $X_1, X_2, \ldots, X_n$, and by utilizing the relationship between cumulant generating functions as given in Exercise 1 on p. 66 (i.e. $K_{\sum_{i=1}^{n} X_i}(t) = nK_{X_1}(t)$) we get the saddlepoint approximation formula for the density $f(\bar{x})$ of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as:

$$\hat{f}(\bar{x}) \approx \sqrt{\frac{n}{2\pi K_X''(\hat{t})}}\, e^{\{nK_X(\hat{t}) - n\hat{t}\bar{x}\}} \left\{1 + \left[\frac{1}{8n}\hat{\rho}_4 - \frac{5}{24}\hat{\rho}_3^2\right]\right\}$$

Here K_X is the cumulant generating function of a single observation X_i, $\hat{t}$ is the saddlepoint (i.e. it has to be re-calculated for every value $\bar{x}$ of the argument as the solution to $K_X'(\hat{t}) = \bar{x}$, and $\hat{\rho}_i = n^{1-\frac{i}{2}}\rho_i(\hat{t})$, with $\rho_i(t) = \frac{K_X^{(i)}(t)}{[K_X''(t)]^{i/2}}$, $i \geq 3$

Even the simpler version of the above second order saddlepoint approximation, namely the first order one:

$$\hat{f}(\bar{x}) \approx \sqrt{\frac{n}{2\pi K_X''(\hat{t})}} \cdot e^{nK_X(\hat{t}) - n\hat{t}\bar{x}}$$

is extremely accurate for sample sizes such as 5, 6, 10.

2) Then I discussed the structure (not the derivation) of a similar formula for the CDF of $\bar{X}$ (called the Lugannani–Rice formula):

$$F_{\bar{X}}(\bar{x}) = P(\bar{X} \le \bar{x}) = \Phi(\hat{w}_n) + \varphi(\hat{w}_n)\left(\frac{1}{\hat{w}_n} - \frac{1}{\hat{u}_n}\right) + O\left(\frac{1}{n}\right)$$

Here $\hat{w}_n = \mathrm{sgn}(\hat{t})\sqrt{2n[\hat{t}\bar{x} - K_X(\hat{t})]}$, $\hat{u}_n = \hat{t}\sqrt{n K_X''(\hat{t})}$ where again $\hat{t}$ is the saddlepoint, i.e. $K_X'(\hat{t}) = \bar{x}$ holds.

3.) I then plugged in the formulae above on 2 occasions:

i) Example 1: (saddlepoint approximation for the sample mean of the standard normal: in this case the approximation is precise, i.e. no error):

$f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. For this, the cgf is known to be $K_X(t) = \log M_X(t) = \log\left(e^{\frac{t^2}{2}}\right) = \frac{t^2}{2}$

Hence $K_X'(t) = t$; $K_X''(t) = 1$. The saddlepoint equation gives $\hat{t} = \bar{x}$, hence the first order approximation for the density becomes $\hat{f}(\bar{x}) = \sqrt{\frac{n}{2\pi}}\, e^{n\frac{(\bar{x})^2}{2} - n\frac{(\bar{x})^2}{2}} = \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\bar{x})^2}{2}}$

which is precisely the density of $N(0, \frac{1}{n})$ (and we know that $\bar{X} \sim N(0, \frac{1}{n})$ in this case).

For the cdf we get: $\hat{w}_n = \mathrm{sgn}(\bar{x})\sqrt{2n\left((\bar{x})^2 - \frac{(\bar{x})^2}{2}\right)} = \bar{x}\sqrt{n}$ and $\hat{u}_n = \hat{w}_n = \bar{x}\sqrt{n}$, hence $\varphi(\hat{w}_n)\left(\frac{1}{\hat{w}_n} - \frac{1}{\hat{u}_n}\right) = 0$ in this case and we get

$F_{\bar{X}}(\bar{x}) = \Phi(\bar{x}\sqrt{n})$ which is the cdf of $N(0, \frac{1}{n})$ (again precise)

ii) Example 2 Saddlepoint approximation for the density of the sample mean of n i.i.d. observations from the Gamma$(\alpha,1)$ density.

The density of a single observation is $f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$, $x > 0$.

Hence if $t<1$: $M_X(t) = \frac{1}{\Gamma(\alpha)} \int_0^\infty e^{-(1-t)x} x^{\alpha-1} dx = \frac{1}{\Gamma(\alpha)(1-t)^\alpha} \int_0^\infty e^{-y} y^{\alpha-1} dy$

set $(1-t)x = y$, $dx = \frac{dy}{1-t}$

$= \frac{\Gamma(\alpha)}{\Gamma(\alpha)} \cdot \frac{1}{(1-t)^\alpha} = (1-t)^{-\alpha}$

Hence $K_X(t) = -\alpha \log(1-t)$, $K_X'(t) = \frac{\alpha}{1-t}$, $K_X''(t) = \frac{\alpha}{(1-t)^2}$

Saddlepoint equation is: $\frac{\alpha}{1-\hat{t}} = \bar{x}$, i.e. $\hat{t} = 1 - \frac{\alpha}{\bar{x}}$.

Hence $K_X''(\hat{t}) = \frac{(\bar{x})^2}{\alpha}$ and substituting in the first-order saddlepoint approximation formula we get

$$\hat{f}(\bar{x}) = \sqrt{\frac{n\alpha}{2\pi(\bar{x})^2}} \exp\left[-n\alpha \log\left(\frac{\alpha}{\bar{x}}\right) - n(\bar{x}-\alpha)\right] =$$

$$= \left(\sqrt{\frac{2\pi}{n\alpha}} (n\alpha)^{n\alpha} e^{-n\alpha}\right)^{-1} (n\bar{x})^{n\alpha-1} e^{-n\bar{x}} \cdot n$$

Now: $\frac{1}{\Gamma(n\alpha)} (n\bar{x})^{n\alpha-1} e^{-n\bar{x}} \cdot n$ is the exact density of $\bar{X}$

(because $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, and $\sum_{i=1}^n X_i \sim$ Gamma$(n\alpha, 1)$).

Hence the difference between the exact and the approximation is that the norming $\frac{1}{\Gamma(n\alpha)}$ has been replaced by $\frac{1}{\sqrt{\frac{2\pi}{n\alpha}} (n\alpha)^{n\alpha} e^{-n\alpha}}$

However the famous Stirling approximation of the Gamma function tells us precisely that

$$\Gamma(n\alpha) \approx \sqrt{\frac{2\pi}{n\alpha}} (n\alpha)^{n\alpha} e^{-n\alpha}$$

Hence the difference is only that the norming constant in the true density $\left(\frac{1}{\Gamma(n\alpha)}\right)$ has been replaced by its very accurate approximation given by Stirling.

3) In the lecture about robustness, I first discussed the fact that $\bar{X}$ is a disastrously bad "estimator" for the location parameter of the Cauchy distribution.
Indeed, it is known that the characteristic function of a single observation $X_1 \sim$ Cauchy (θ) with density $f(x;\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$ is $\varphi_{X_1}(t) = e^{i\theta t - |t|}$
Hence for the characteristic function of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we have

$$\varphi_{\bar{X}}(t) = E\, e^{i\frac{t}{n}\left(\sum_{i=1}^{n} X_i\right)} \overset{\text{using independence}}{=} \left[E\, e^{i\left(\frac{t}{n}\right)X_1}\right]^n = \left(e^{i\frac{\theta}{n}t - \frac{|t|}{n}}\right)^n =$$

$$= e^{i\theta t - |t|} = \varphi_{X_1}(t)$$

that is, $\bar{X}$ has the same distribution as a single X_1 no matter what the sample size! Hence $\bar{X}$ can not be consistent for θ.
With respect to the remaining discussions, I believe that my derivations in the lecture notes are detailed enough to reproduce them again here. Finally, I note that the asymptotic variance for the sample quantile as given by $V(T,F) = \frac{p(1-p)}{f(q_p)^2}$ on p. 78 of the notes,

specialises to

$$\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{f(0)^2}$$

when applied for the median ($p=\frac{1}{2}$) of the location family $f(x-\theta) = f(x,\theta)$. In particular this justifies the result

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f^2(0)}\right)$$

which I was using on p.75 of the notes (formula (29)). For the normal location family this gave us:

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{\pi}{2}\right),$$ i.e. the empirical median for the standard normal is asymptotically normally distributed with asymptotic variance $\frac{\pi}{2} > 1$ and (as we know), is less efficient than the empirical (sample) mean for which we have precisely

$$\sqrt{n}(\bar{X} - \theta) \sim N(0, 1).$$

This illustrates the claim that we "pay for robustness by a slight loss of efficiency".

My white board writing from week 6

1) I started with the Poisson(θ) example, $\underline{X} = (X_1, X_2, \dots, X_n)$ i.i.d Poisson(θ) : $f(x,\theta) = \frac{e^{-\theta}\theta^x}{x!}$, $x = 0,1,2,\dots$

$L(\underline{X},\theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n x_i!}$, hence $V(X,\theta) = \frac{\partial}{\partial\theta}\log L = -n + \frac{\sum_{i=1}^n X_i}{\theta}$

I considered two cases for a parameter $\tau(\theta)$ of interest:

a) $\tau(\theta) = \theta$. In this case $V(X,\theta) = \frac{n}{\theta}(\bar{X} - \theta)$ here we represents a factorization in which we have a statistic (i.e. transformation of the data only, no parameter involved). Hence the CR bound is attainable and the statistic $\bar{X}$ attains it. We can check also directly the attainability in this case:

$-\frac{\partial^2}{\partial\theta^2}\log L = \frac{\sum_{i=1}^n X_i}{\theta^2}$ implies $I_X(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\log L\right) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}$

and the CR bound for variance of an unbiased estimator of $\tau(\theta) = \theta$ is $\frac{(\tau'(\theta))^2}{I_X(\theta)} = \frac{1}{n/\theta} = \boxed{\frac{\theta}{n}}$

And a direct check shows: $Var(\bar{X}) = \frac{1}{n^2}Var\left(\sum_{i=1}^n X_i\right) = \frac{n\theta}{n^2}$

b) $\tau(\theta) = e^{-\theta} = P(X_1 = 0)$. Then $V(X,\theta) = ne^{\theta}\left(\left(\frac{1}{\theta}e^{-\theta}\bar{X} - e^{-\theta}\right)$ Since the quantity in the factorization now does depend on θ (i.e., is not a statistic), the CR Bound is not attainable by any unbiased estimator of $\tau(\theta) = e^{-\theta}$

HOWEVER, as we will see a bit later, $\left(1 - \frac{1}{n}\right)^{n\bar{X}}$ is an unbiased estimator of $\tau(\theta) = e^{-\theta}$ that is the UMVUE (just that its variance, even if the smallest possible, is > then the Bound.

2) I then went through the proof of the Rao Blackwell theorem:

i) $E\hat{\tau}(T) = E_T(E(W|T)) = EW = \tau(\theta)$

"iterative" property of expected value

(hence $\hat{\tau}(T)$ is unbiased for $\tau(\theta)$).

ii) We first show that "always" $\text{Var}(Y|X) \le \text{Var}\, Y$ (i.e. the variance is never increased after conditioning)

Let $a(X) = E(Y|X)$. Then:

$$\text{Var}\, Y = E(\underbrace{Y - a(X)} + \underbrace{a(X) - EY})^2 = E(Y-a(X))^2 + E(a(X)-EY)^2 + 2E[(Y-a(X))(a(X)-EY)]$$

Next: $E[(Y-a(X))(a(X)-EY)] = E_X[E((Y-a(X))(a(X)-EY)|X]$

$= E_X\{(a(X)-EY)E(Y-a(X)|X)\} =$

$= E_X\{(a(X)-EY)(\underbrace{E(Y|X)}_{a(X)} - a(X))\} = 0$

Hence $\text{Var}\, Y = E(Y-a(X))^2 + E(a(X)-EY)^2 \ge$

$\ge E(a(X)-EY)^2 = E(a(X)-E(a(X)))^2 = \text{Var}(a(X))$

i.e. $\text{Var}(Y|X) \le \text{Var}\, Y$

3). I discussed sufficiency and completeness and its role in justifying the Lehmann-Scheffe theorem which gives us the recipe to obtain UMVUE by Rao-Blackwellising an unbiased estimator by conditioning on complete and sufficient statistic. This discussion is in the lecture notes.

NEXT, I solved a variety of illustrative examples.

Some of my white board writing in week 6

Completeness, Lehmann-Scheffe:

• $X=(X_1,X_2,\dots,X_n)$ i.i.d. $N(0,\theta)$. Show, $T=\bar{X}$ is not complete for θ. It suffices to find a counterexample. Here it is: take $g(t)=t\neq 0$. We have

$E_\theta g(T)=E_\theta(\bar{X})=0 \ \forall \theta>0$ but $g(t)\neq 0$

• $X=(X_1,X_2,\dots,X_n)$ i.i.d. Bernoulli with parameter θ. The statistic $T=\sum_{i=1}^{n}X_i$ is complete for $\theta\in(0,1)$:

We know $T\sim Bin(n,\theta) \longrightarrow P_\theta(T=t)=\binom{n}{t}\theta^t(1-\theta)^{n-t}$

Take $E_\theta g(T)=0 \ \forall \theta\in(0,1) \Rightarrow \sum_{t=0}^{n} g(t)\binom{n}{t}\theta^t(1-\theta)^{n-t}=0$

$\Rightarrow \underbrace{(1-\theta)^n}_{\neq 0}\cdot\sum_{t=0}^{n} g(t)\binom{n}{t}\eta^t=0 \quad \forall \ \eta=\frac{\theta}{1-\theta}\in(0,\infty).$

Then all coefficients $g(t)\binom{n}{t}=0$ must hold and

Since $\binom{n}{t}\neq 0 \Rightarrow g(t)=0, t=0,1,2,\dots,n \Rightarrow P_\theta(g(T)=0)=1$

and T is complete.

We also know: T is sufficient. Hence if we start with an unbiased estimator W of $\tau(\theta)=\theta(1-\theta)$ and calculate $E(W|T)$, in a second step, we will get the UMVUE of $\tau(\theta)=\theta(1-\theta)$.

Suggestions for W: $W=X_1(1-X_2)$, (or $\tilde{W}=I_{(X_1=1)}(X)I_{(X_2=0)}(X)$).

We see: $E_\theta W=E_\theta X_1-E_\theta(X_1X_2)=\theta-E_\theta X_1\,E_\theta X_2=\theta-\theta^2=\theta(1-\theta)$

(Similarly $E_\theta\tilde{W}=EI_{(X_1=1)}(X)\cdot EI_{(X_2=0)}(X)=$

$=P(X_1=1)P(X_2=0)=\theta(1-\theta)$.

Now we get:

$$E_\theta(W|T=t) = 1 \cdot P_\theta(W=1|T=t) + 0 =$$

$$= \frac{P(W=1 \cap T=t)}{P(T=t)} = \frac{P(X_1=1 \cap X_2=0 \cap \sum_{i=3}^{n} X_i = t-1)}{P(\sum_{i=1}^{n} X_i = t)} =$$

↑ made sure to have intersection of independent events

$$= \frac{P(X_1=1)\,P(X_2=0)\,P(Bin(n-2,\theta)=t-1)}{P(Bin(n,\theta)=t)} = \frac{\theta(1-\theta)\binom{n-2}{t-1}\theta^{t-1}(1-\theta)^{n-t-1}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} =$$

$$= \frac{\binom{n-2}{t-1}}{\binom{n}{t}} = \boxed{\bar{X}(1-\bar{X})\frac{n}{n-1}} \text{ being the UMVUE of } \tau(\theta)=\theta(1-\theta)$$

- For uniform in $[0,\theta)$ distribution, $Y=\frac{n+1}{n}X_{(n)}$ is the UMVUE of the parameter $\tau(\theta)=\theta$. Justification:
From the previous lectures we know that Y is unbiased for θ and that $X_{(n)}$ is a sufficient statistic for θ. Now we show that $X_{(n)}$ is also <u>complete</u>. Recall that $f_{X_{(n)}}(t,\theta) = \begin{cases} \frac{n t^{n-1}}{\theta^n}, & 0<t<\theta \\ 0 & \text{else} \end{cases}$ (derived last lecture)

Take $E_\theta g(X_{(n)}) = 0 \;\forall\, \theta \in (0,\infty) \Rightarrow \frac{n}{\theta^n}\int_0^\theta g(t)t^{n-1}dt = 0$

$0 = \frac{d}{d\theta}\int_0^\theta g(t)t^{n-1}dt = g(\theta)\theta^{n-1}$. But since $\theta>0$, this implies $g(\theta)=0$ for all $\theta>0$. Equivalently: $P_\theta(g(X_{(n)})=0)=1$ and $X_{(n)}$ is complete.

Now: $Y=\frac{n+1}{n}X_{(n)}$ is unbiased for θ and is a function of complete and sufficient statistic. Conditioning Y on $X_{(n)}$ does not change it $E(Y/X_{(n)}) = \frac{n+1}{n}X_{(n)}$
Hence, by Lehmann-Scheffe: $Y=\frac{n+1}{n}X_{(n)}$ is UMVUE of $\tau(\theta)=\theta$.

• the Poisson example continued: For $\tau(\theta)=e^{-\theta}$,

UMVUE was advertised as being $\left(1-\frac{1}{n}\right)^{n\bar{X}}$ and below I justify this claim:

First we note that $T=\sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$ (this is a known property of the Poisson distribution. $T=\sum_{i=1}^{n} X_i$ is known to be sufficient for θ from our previous lectures. Now we will show that it is also complete:

Take $E_\theta g(T)=0 \ \forall \theta>0 \Rightarrow \sum_{t=0}^{\infty} g(t)\, e^{-n\theta}\frac{(n\theta)^t}{t!}=0$

for all $\theta>0$. This means $e^{-n\theta} \neq 0$ and $\sum_{t=0}^{\infty} g(t)\frac{(n\theta)^t}{t!}=0 \ \forall \theta>0$ (this polynomial of θ must be then $\equiv 0 \ \forall \theta>0$ → the coefficients $\frac{g(t)n^t}{t!}$ must be all $=0$ which implies $g(t)=0, t=0,1,\ldots$

i.e. $P_\theta(g(T)=0)=1$ and $T=\sum_{i=1}^{n} X_i$ is complete.

To find an unbiased starting estimator for $\tau(\theta)=e^{-\theta}$ we use the interpretation $e^{-\theta}=P_\theta(X_1=0)$ of $\tau(\theta)$. Hence $W=I_{(X_1=0)}(X)$ would be unbiased for $\tau(\theta)$: $E_\theta W = 1\cdot P_\theta(X_1=0)=e^{-\theta}$. If we now condition on the complete & sufficient $T=\sum_{i=1}^{n} X_i$, we will get the UMVUE:

$$E_\theta(W\mid T=t)=1\cdot P(W=1\mid T=t)=\frac{P(W=1\cap T=t)}{P(T=t)}=$$

$$=\frac{P(X_1=0\cap \sum_{i=1}^{n} X_i=t)}{P(\sum_{i=1}^{n} X_i=t)}=\frac{P(X_1=0\cap \sum_{i=2}^{n} X_i=t)}{P(\sum_{i=1}^{n} X_i=t)}=$$

$$=\frac{e^{-\theta}\cdot e^{-(n-1)\theta}\frac{((n-1)\theta)^t}{t!}}{e^{-n\theta}\frac{(n\theta)^t}{t!}}=\left(\frac{n-1}{n}\right)^t=\left(1-\frac{1}{n}\right)^{n\bar{X}} \quad \text{q.e.d.}$$

- I also justified why for the uniform distribution in $[0,\theta]$, the maximal observation $X_{(n)}$ is the MLE, i.e. $\hat{\theta}_{MLE} = X_{(n)}$.

I noticed that $L(\bar{X},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$ is not differentiable for all θ, hence instead of trying to solve the equation $\nabla(\bar{X},\hat{\theta}_{MLE}) = 0$ to find the MLE (which is what we would do in "regular" cases), we look directly into the shape of $L(\bar{X},\theta)$ to see which is the argument that maximizes it.

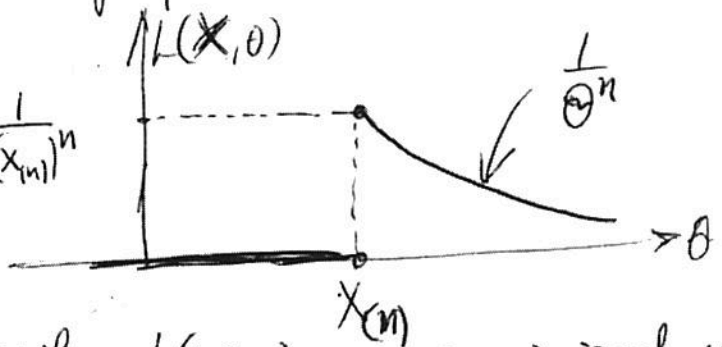
Since $f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta)$ then

$$L(\bar{X},\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(x_i,\infty)}(\theta) =$$

$$\boxed{\frac{1}{\theta^n} I_{(x_{(n)},\infty)}(\theta)}$$

↑ using properties of indicators

If we now graph $L(\bar{X},\theta)$ we get after plugging the sample:



And clearly $L(\bar{X},\theta)$ is maximized when $\theta = X_{(n)}$, i.e. $\hat{\theta}_{MLE} = X_{(n)}$ by direct inspection

Note that the $\hat{\theta}_{MLE} = X_{(n)}$ is different from the UMVUE $Y = \frac{n+1}{n} X_{(n)}$ that we derived earlier.

Finally, I stressed that usually parameters of interest such as $\tau(\theta) = \theta(1-\theta)$ for the Bernoulli, $\tau(\theta) = e^{-\theta} = P(X_1 = 0)$ for the Poisson; $\tau(\theta) = \theta e^{-\theta} = P(X_1 = 1)$ for the Poisson, etc., typically have some probabilistic interpretation that can be exploited to suggest a (simple) unbiased estimator W which then can be Rao-Blackwellized to obtain the UMVUE. The tutorial questions (set 2) contain a lot of such exercises. By doing them, you can get a feeling how to proceed in a particular situation.

ONE MORE REMARK I made at the end: If we have realized that we are dealing with a one-parameter exponential family density $f(x,\theta) = a(\theta)b(x)\exp(c(\theta)d(x))$ then the statistic $T(X) = \sum_{i=1}^{n} d(X_i)$ is complete and minimal sufficient. We do not need to separately check completeness for such families.

Some of my white board writing from week 10

I discussed the generalized likelihood ratio test for two examples related to the normal distribution:

(a) Testing $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$ for a sample of n i.i.d. $N(\mu, \sigma^2)$ (σ^2 assumed known)

In this case

$$-2\left[\ln L(\mathbf{x}, H_0) - \ln L(\mathbf{x}, \bar{x})\right] = -2\left[-\sum_{i=1}^{n}\frac{(x_i-\mu_0)^2}{2\sigma^2} + \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2}\right] =$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2 - \sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{n(\bar{x}-\mu_0)^2}{\sigma^2}$$

This should be $\sim \chi^2_1$ asymptotically but in this case, because of dealing with normal, the result is precise, (not only asymptotic). Indeed, we know that under H_0:

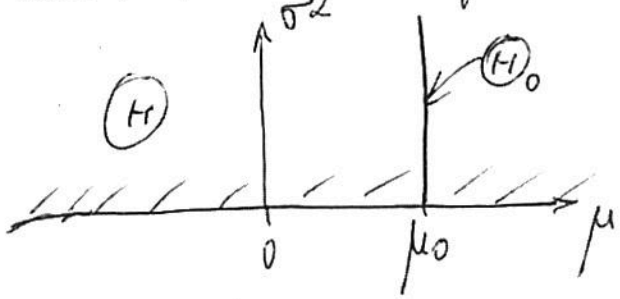
$$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0,1) \Rightarrow \frac{n(\bar{x}-\mu_0)^2}{\sigma^2} \sim \chi^2_1$$

The GLRT is: $\varphi^* = \begin{cases} 1 & \text{if } \frac{n(\bar{x}-\mu_0)^2}{\sigma^2} > \chi^2_{\alpha,1} \\ 0 & \text{if } \frac{n(\bar{x}-\mu_0)^2}{\sigma^2} \leq \chi^2_{\alpha,1} \end{cases}$ and

is equivalent to the standard Z-test in this case.

(b) Testing $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$ again but when σ^2 is unknown. Hence we are testing in effect:

$H_0: \begin{cases} \mu = \mu_0 \\ \sigma^2 > 0 \end{cases}$ vs $H_1: \begin{cases} \mu \neq \mu_0 \\ \sigma^2 > 0. \end{cases}$ In terms of the notation of GLRT: we have a parameter vector $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$ and



H_0 is one-dimensional subspace as sketched, whereas H is "anything" above the $0x$ axis.

The dimensions K, r, S, as discussed in Section 6.10.1 of Lecture 6, p. 56, are $K=2$, $r=1$, $S=1$.
To perform the GLRT we need to maximize $L(X,\theta)$ under the Hypothesis and under the alternative.
i) under the hypothesis: $\mu=\mu_0$, so we need to optimize w.r. to σ only. $\ln L = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2 + \text{const}$

$$\frac{\partial}{\partial\sigma^2}\ln L = 0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu_0)^2 \quad \text{implies}$$

$$\hat{\sigma}^2_{H_0} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2$$

and $$\sup L|H_0 = \frac{1}{(\sqrt{2\pi})^n(\hat{\sigma}^2_{H_0})^{n/2}}\exp\left(-\frac{n}{2}\right)$$

ii) without the restriction of H_0, we have to maximise $\ln L$ w.r. to both μ and σ^2, i.e. solve the system

$$\left.\begin{aligned}\frac{\partial}{\partial\mu}\ln L &= 0\\ \frac{\partial}{\partial\sigma^2}\ln L &= 0\end{aligned}\right| \quad \text{This leads to} \quad \begin{aligned}\hat{\mu} &= \bar{X}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\end{aligned}$$

When plugging-in, we get the $\sup L$ without any restriction and it is $\sup L = \frac{1}{(\sqrt{2\pi})^n(\hat{\sigma}^2)^{n/2}}\exp\left(-\frac{n}{2}\right)$

Hence $-2\log\Lambda = -2\log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_{H_0}}\right)^{n/2} = n\log\left(\frac{\hat{\sigma}^2_{H_0}}{\hat{\sigma}^2}\right)$ and the GLRT is

$$\varphi^* = \begin{cases}1 & \text{if } n\log\left(\frac{\hat{\sigma}^2_{H_0}}{\hat{\sigma}^2}\right) > \chi^2_{\alpha,1}\\ 0 & \text{if } n\log\left(\frac{\hat{\sigma}^2_{H_0}}{\hat{\sigma}^2}\right) \le \chi^2_{\alpha,1}\end{cases}$$

(the degrees of freedom are $=1$ since in this case $r=1=K-S$ $(K=2, S=1)$.
Note that now the convergence of $-2\log\Lambda$ to the limiting χ^2_1 is only asymptotic (not precise as in case a)).
But $-2\log\Lambda = n\log\left(1+\frac{(\bar{X}-\mu_0)^2}{\hat{\sigma}^2}\right) \approx \frac{n(\bar{X}-\mu_0)^2}{S^2}$, so it is "almost" equivalent to the standard t-test for $\mu=\mu_0$ when σ^2 is unknown.

Continuation:

Some of my white board writing in week 10

1.) Regarding the multinomial distribution:

I explained the formula

$$P(X_1=x_1, X_2=x_2, \dots, X_K=x_K) = \frac{n!}{x_1!\,x_2!\dots x_K!}\, w_1^{x_1} w_2^{x_2} \cdots w_K^{x_K},$$

$0 < w_i < 1$, $\sum_{i=1}^{K} w_i = 1$ for calculating the probability of a particular outcome $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix}$ with $x_1+x_2+\cdots+x_K=n$, n being the number of independent trials.

I also discussed two examples:

i) If a die is tossed 6 times, what is the probability that each number $(1,2,3,4,5,6)$ turns up once.

Applying the above formula with $K=6$, $x_i=1$, $i=1,2,\dots,6$, and $w_i = \frac{1}{6}$, $i=1,2,\dots,6$ we get $6!\left(\frac{1}{6}\right)^6 = \frac{5}{324}$

ii) Out of 7 tosses, what is the probability that each number $(1,2,3,4,5,6)$ turns up at least once.

Answer: $6 \cdot \frac{7!}{2!\,(1!)^6} \cdot \left(\frac{1}{6}\right)^7 = \frac{35}{648}$

2:) I discussed in detail the proof of Theorem 7.2. (p. 60) but I see that all details are presented in the lecture note so I will abstain from reproducing them again here.

3.) I discussed a simple method to derive the density of the r-th order statistic as stated in Theorem 7.3, p. 61

Details: we introduce a discrete random variable
$Y = \{$ number of realisations $X_1, X_2, \dots, X_n$ that happen to be $\le x\}$
Then $Y \sim Bin(n, F_X(x))$.
Now we first derive $F_{X_{(r)}}(x)$ (the cdf of $X_{(r)}$) and then differentiate it to find the density. The main observation we make is that

$$F_{X_{(r)}}(x) = P(X_{(r)} \le x) = P(Y \ge r)$$

Hence we can state that

$$F_{X_{(r)}}(x) = \sum_{k=r}^{n} \binom{n}{k} (F_X(x))^k (1 - F_X(x))^{n-k}$$

Now to get the density we need to differentiate each of the summands in $\sum_{k=r}^{n}$ by applying the $(uv)' = u'v + v'u$ formula each time.
We get:

$$f_{X_{(r)}}(x) = \binom{n}{r} r f_X(x) F_X(x)^{r-1} (1 - F_X(x))^{n-r} - \binom{n}{r}(n-r) F_X(x)^r (1 - F_X(x))^{n-r-1} f_X(x)$$
$$+ \binom{n}{r+1}(r+1) f_X(x) F_X(x)^r (1 - F_X(x))^{n-r-1} - (\quad)$$
$$+ (\quad) - (\quad)$$
$$+ (\quad) - \quad - \quad = (\quad)$$
$$+ (\quad) \qquad + (n-n) * (\quad) = 0$$

Huge cancellation happens!! and, because of the equality $\binom{n}{r}(n-r) = \binom{n}{r+1}(r+1)$ each of the summands after the first one disappears. Hence

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} f_X(x) F_X(x)^{r-1} (1 - F_X(x))^{n-r} \text{ holds}$$

4) I also discussed the idea of the proof of Theorem 7.4 on p. 62. Again, we first get the cdf and then find the mixed partial derivative

$$\frac{\partial^2}{\partial u \partial v} F_{X_{(i)},X_{(j)}}(u,v)$$ to calculate the density $f_{X_{(i)},X_{(j)}}(u,v)$.

With the discrete variables U and V as introduced on p. 62 we see that

$$(U, V, n-U-V) \sim \text{Multinomial}\left(n; F_X(u), F_X(v)-F_X(u), 1-F_X(v)\right)$$

Then we observe that

$$F_{X_{(i)},X_{(j)}}(u,v) = P(U \geq i \cap U+V \geq j) =$$

$$= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U=k, V=m) + P(U \geq j)$$

Since the second summand does not involve v, its mixed partial derivative w.r. u and v will be zero and hence

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{\partial^2}{\partial u \partial v} \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \frac{n!}{k!\,m!\,(n-k-m)!} \left(F_X(u)\right)^k \left(F_X(v)-F_X(u)\right)^m \left(1-F_X(v)\right)^{n-k-m}$$

Again, a huge cancellation happens when we calculate the partial derivatives by using the product rule for differentiation and we end up with

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!} f_X(u) f_X(v) \left(F_X(u)\right)^{i-1} \left(F_X(v)-F_X(u)\right)^{j-1-i} \left(1-F_X(v)\right)^{n-j}$$

for $n \geq j > i \geq 1$, $-\infty < u < v < \infty$ (and $= 0$ else)

(5) I also discussed in detail the example stating that for the range $R = X_{(n)} - X_{(1)}$ for order statistic $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ from the uniform $(0,1)$ distribution, it holds $f_R(u) = \begin{cases} n(n-1)u^{n-2}(1-u), & 0<u<1 \\ 0 & \text{else} \end{cases}$

Proof:

To this end, first we note that by using the formula from Theorem 7.4 we have (with $i=1$ and $j=n$):

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)\left(F_X(y)-F_X(x)\right)^{n-2} f_X(x) f_X(y)$$

We introduce the variable of interest

$U = X_{(n)} - X_{(1)}$ and one auxiliary variable

$V = X_{(n)}$ so that we could apply

the density transformation formula

$$f_{\binom{U}{V}}(u,v) = f_{\binom{X_{(1)}}{X_{(n)}}}(x(u,v), y(u,v)) \left| \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \right|$$

Note: $\begin{cases} X_{(1)} = V - U =: x \\ X_{(n)} = V =: y \end{cases}$ and $\left| \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \right| = \left| \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} \right| = 1$

Hence $f_{\binom{U}{V}}(u,v) = n(n-1)\left(F_X(v)-F_X(v-u)\right)^{n-2} f_X(v-u) f_X(v) \cdot 1$

To get $f_U(u)$ (which we are interested in) we need to integrate out the unwanted variable V from the joint density $f_{\binom{U}{V}}(u,v)$. We need to be careful with the integration range when doing this:

Since $0 < X_{(1)} < X_{(n)} < 1$ we get

$$\Downarrow$$

$$0 < V-U < V < 1$$

$$\Updownarrow$$

$$\underline{0 < U < V < 1}$$

. This means that for a fixed U, V ranges in the interval $(U, 1)$.

Therefore:

$$f_R(u) = \int_u^1 f_{(U,V)}(u,v)\,dv = \int_u^1 n(n-1)\,(v-(v-u))^{n-2}\,dv$$

$$= \begin{cases} n(n-1)u^{n-2}(1-u) & \text{if } 0<u<1 \\ 0 & \text{else} \end{cases}$$

I also advised you to repeat this exercise by using $X_{(1)} = V$ as an auxiliary variable. The intermediate calculations will be slightly different but at the end after integrating out V again (BUT THIS TIME in the range $(0, 1-u)$ (!)) you will get the same final result for the density $f_R(u)$.

6.) I also discussed one more problem in class (<u>(7/d) from tutorial sheet 4</u>) but because it is completely solved in the solutions to tutorial set 4, I abstain from reproducing the derivation here.

Some of my white board writing in weeks 8&9

1) In week 8, I first discussed the examples on p. 45-47 of the notes. Looking at the amount of material provided in the notes, I believe that you should be able to reconstruct the details of these examples by yourself.

In week 9, I started with recalling the Neyman-Pearson Lemma.

2) Proof of the Neyman-Pearson Lemma: Again, I did complete derivation on the white board but looking at the content that is put on p. 50-51 of the notes, I believe that you should be able to reconstruct the details of the proof yourself. Then I did:

3.) Example about uniformly most powerful (UMP) α-test for the normal distribution:

$\mathbf{X} = (X_1, X_2, \dots, X_n)$ i.i.d. $N(\theta, 1)$. Consider $H_0: \theta = \theta_0 \in R^1$ versus a composite $H_1: \theta > \theta_0$ (all these are alternative)

We are looking for UMP α-test φ^* which means: if we take any competitor $\varphi \in \Phi_\alpha = \{$set of all tests φ such that $E_{\theta_0}\varphi \le \alpha\}$ then we claim that $E_\theta \varphi^* \ge E_\theta \varphi$ for all $\theta > \theta_0$

We first simplify the problem by considering testing a simple $H_0: \theta = \theta_0$ versus simple $H_1: \theta = \theta_1$ for a fixed $\theta_1 > \theta_0$. Because this is a Neyman-Pearson lemma-type problem, for it we have the most powerful α test and it is given by

$$\varphi^* = \begin{cases} 1 & \text{if } L(\mathbf{X}, \theta_1)/L(\mathbf{X}, \theta_0) > C \\ 0 & \text{if } L(\mathbf{X}, \theta_1)/L(\mathbf{X}, \theta_0) \le C \end{cases}$$

Notice that

$$\frac{L(X,\theta_1)}{L(X,\theta_0)} = \exp\left((\theta_1-\theta_0)\sum_{i=1}^{n} X_i + \frac{n}{2}(\theta_0^2-\theta_1^2)\right)$$

Since $\theta_1-\theta_0>0$, $\frac{L(X,\theta_1)}{L(X,\theta_0)}$ is monotonically increasing in $T=\sum_{i=1}^{n} X_i$ and $\frac{L(X,\theta_1)}{L(X,\theta_0)} > C$ is equivalent to $\sum_{i=1}^{n} X_i > C_1$ or, by renaming constants, to $\bar{X} > \bar{\bar{C}}$.

To find $\bar{\bar{C}}$, we must exhaust the given level α which means $E_{\theta_0}\varphi^* = 1 * P_{\theta_0}(\bar{X}>\bar{\bar{C}}) = \alpha$ must hold (see the statement of the NP Lemma).

But $E_{\theta_0}\varphi^* = P_{\theta_0}(\bar{X}>\bar{\bar{C}}) = P_{\theta_0}\left(\frac{\sqrt{n}(\bar{X}-\theta_0)}{1} > \frac{\sqrt{n}(\bar{\bar{C}}-\theta_0)}{1}\right)$

$= P\left(Z > \frac{\sqrt{n}(\bar{\bar{C}}-\theta_0)}{1}\right) = \alpha$ where $Z\sim N(0,1)$.

This implies that $\frac{\sqrt{n}(\bar{\bar{C}}-\theta_0)}{1} = z_\alpha$ must hold where z_α is the upper $\alpha*100\%$ point of the $N(0,1)$.

Then $\bar{\bar{C}} = \theta_0 + \frac{z_\alpha}{\sqrt{n}}$ and φ^* becomes:

$$\varphi^*(X) = \begin{cases} 1 & \text{if } \bar{X} > \theta_0 + \frac{z_\alpha}{\sqrt{n}} \\ 0 & \text{if } \bar{X} \le \theta_0 + \frac{z_\alpha}{\sqrt{n}} \end{cases}$$

NOW WE NOTICE that the resulting $\varphi^*(X)$ above, although having been constructed for a particular $H_1: \theta=\theta_1$, DOES NOT involve this $\theta_1>\theta_0$ in its shape. Hence the SAME test φ^* will be the most powerful α-test for any chosen $\theta_1>\theta_0$! Therefore, $\varphi^*(X)$ will be the UNIFORMLY most powerful for testing also $H_0: \theta=\theta_0$ versus $H_1: \theta>\theta_0$.

Notice that we used the monotonicity of the likelihood ratio in our argument.

This example was generalized in the Blackwell-Girshick (BG) Theorem (p.52 of the notes).

I also gave an example of applying the BG theorem to derive UMP α tests.

Example: Assume that $\mathbf{X}=(X_1, X_2, \dots, X_n)$ are i.i.d. from $f(x,\theta) = \begin{cases} 2x/\theta^2, & 0<x<\theta \\ 0 & \text{else.} \end{cases}$ Construct a UMP α-test of $H_0: \theta \le \theta_0$ versus $H_1: \theta > \theta_0$.

Solution: First we want to show that the family $L(\mathbf{X},\theta)$ in this case is a MLR family in the statistic $T = X_{(n)}$.

Indeed $L(\mathbf{X},\theta) = \frac{2^n \prod_{i=1}^n x_i}{\theta^{2n}} I_{(X_{(n)},\infty)}(\theta)$. Now take two values $0<\theta'<\theta''$ and consider

$$\frac{L(\mathbf{X},\theta'')}{L(\mathbf{X},\theta')} = \left(\frac{\theta'}{\theta''}\right)^{2n} \frac{I_{(X_{(n)},\infty)}(\theta'')}{I_{(X_{(n)},\infty)}(\theta')}$$

Putting $X_{(n)}$ on the OX axis we have the graph:

Hence we have MLR property in $T = X_{(n)}$. Then BG theorem tells us that UMP α test of H_0 vs H_1 exists and is given by

$$\varphi^* = \begin{cases} 1 & \text{if } X_{(n)} > K \\ 0 & \text{if } X_{(n)} \le K \end{cases}.$$

To find K we need to exhaust the level, i.e. must satisfy $E_{\theta_0}\varphi^* = \alpha$. However $E_{\theta_0}\varphi^* = P_{\theta_0}(X_{(n)} > K) = 1 - P_{\theta_0}(X_{(n)} \le K)$

$= 1 - [P_{\theta_0}(X_1 \le K)]^n = 1 - \left(\frac{K}{\theta_0}\right)^{2n} = \alpha \Rightarrow \boxed{K = \theta_0(1-\alpha)^{\frac{1}{2n}}}$ and φ^* is completely determined.

Some of my white board writing in weeks 8 & 9 continued

I finished the discussion of the example:

$\mathbf{X} = (X_1, X_2, \dots, X_n)$ are i.i.d from $f(x,\theta) = \begin{cases} 2x/\theta^2, & 0<x<\theta \\ 0 & \text{else} \end{cases}$

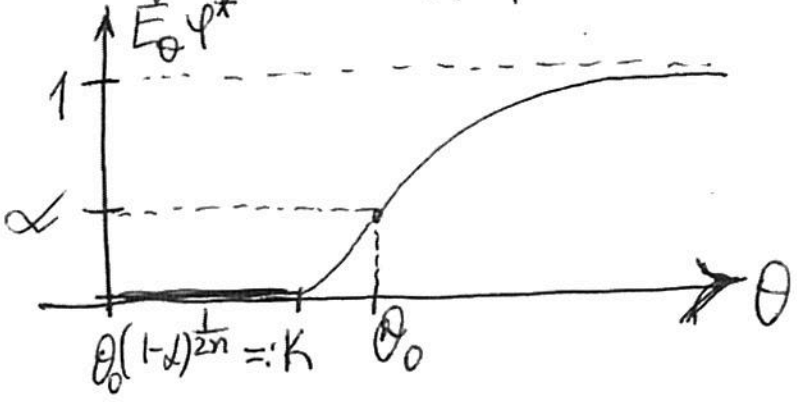
We constructed the UMP α-test of $H_0: \theta \le \theta_0$ vs $H_1: \theta > \theta_0$. We ended up with

$$\varphi^* = \begin{cases} 1 & \text{if } X_{(n)} > K = \theta_0 (1-\alpha)^{\frac{1}{2n}} \\ 0 & \text{if } X_{(n)} \le K = \theta_0 (1-\alpha)^{\frac{1}{2n}} \end{cases}$$

I discussed the graph of the resulting power function: Since

$$E_\theta \varphi^* = \begin{cases} 1 - \left(\frac{\theta_0}{\theta}\right)^{2n} (1-\alpha), & 0<K<\theta \\ 0 & K \ge \theta \end{cases}$$

we can graph this function $E_\theta \varphi^*$, $\theta > 0$ and the graph looks like:



Next, I discussed the construction of UMP α test in the case of DISCRETE data where randomization was necessary.

Problem: $n = 25$ i.i.d. observations from Bernoulli (θ). $\alpha = 0.01$ is chosen for level of significance. Construct the UMP α test of $H_0: \theta \le .15 = \theta_0$ vs $H_1: \theta > .15$

Solution: First of all, UMP α test exists because of the Blackwell-Girshick theorem since,

$$f(x,\theta) = \theta^x (1-\theta)^{1-x} = (1-\theta) \cdot e^{x \ln\left(\frac{\theta}{1-\theta}\right)}$$

is an one-parameter exponential family with

$c(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$ monotone increasing in θ, and $d(x) = x$.
Hence we have the MLR property in the statistic $T = \sum_{i=1}^{n} X_i$ and Blackwell-Girshick's theorem tells us that UMP α test exists and has the structure

$$\varphi^* = \begin{cases} 1 & \text{if } T > K \\ \gamma & \text{if } T = K \\ 0 & \text{if } T < K \end{cases}$$

$\left(T = \sum_{i=1}^{25} X_i \text{ here}\right)$

To determine K and γ we have to find the smallest natural number for which still $P_{\theta_0}(T > K) < \alpha$.

Since $T \sim Bin(25, .15)$ under the borderline $\theta_0 = .15$ value, we can find any of the probabilities $P_{\theta_0}(T = t) = \binom{25}{t} (.15)^t (.85)^{25-t}$, $t = 0, 1, 2, \dots, 25$ and tabulate the whole cdf (or "ask" the computer to do it for us) the extract from the distribution of T follows:

x	...	7	8	9	...
$P(T \le x)$		.974532	.99207	.99786	

We need $P(T > x) < 0.01$ and x should be the smallest with this property $\Rightarrow K = x = 8$. Then

$$\gamma = \frac{\alpha - P_{\theta_0}(\le)}{P_{\theta_0}(=)} = \frac{0.01 - (1 - .99207)}{.99207 - .9745} = .114$$

Hence the UMP 0.01-test is completely determined:

$$\varphi^* = \begin{cases} 1 & \text{if } \sum_{i=1}^{25} X_i > 8 \\ .114 & \text{if } \sum_{i=1}^{25} X_i = 8 \\ 0 & \text{if } \sum_{i=1}^{25} X_i < 8 \end{cases}$$

I then discussed Q2 from tutorial sheet 3 but because a complete solution to it is given on moodle I will not reproduce it here (see it there)

I then moved over to discuss the notion of an uniformly most powerful <u>unbiased</u> (UMPU) α-test. This discussion is well and thoroughly represented in the notes (p. 53–54 there) and I am not reproducing it here again.

I also went through the two examples in Section 6.8 (p. 54-55). Both are related to constructing UMPU α tests of $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$.

- the first example treats the case of sample of size $n=1$ from the exponential distribution

- the second example treats the testing of the same hypothesis for a sample of size n from $N(\theta, 1)$ distribution. For this second problem, it is known that statistical folklore advises to use the so-called "Z-test", i.e.

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 & \text{if } \sqrt{n}\,|\bar{x} - \theta_0| \geq z_{\frac{\alpha}{2}} \\ 0 & \text{if } \sqrt{n}\,|\bar{x} - \theta_0| < z_{\frac{\alpha}{2}} \end{cases}$$

I explained <u>why</u> this test $\varphi^*(\mathbf{x})$ was in fact the UMPU α test for the testing problem $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$

Pages 54-55 contain the detailed argument and I abstain from reproducing it here again.

Some of my white board writing from week 3

I discussed in details the main statements (Theorems 2.3, 2.4 and 2.5) from lecture 2.
Theorem 2.3 is presented in sufficient details in the notes and I am not reproducing it again here.
Theorem 2.4 was based on the following two simple observations:

a) If Y is a random variable with $E(Y^2) < \infty$ then the constant a^* that minimizes $E(Y-a)^2 \xrightarrow[a \in R^1]{}$ is $a^* = E(Y)$.

This is because $\frac{\partial}{\partial a}\left[E(Y-a)^2\right] = \frac{\partial}{\partial a}\left[E(Y^2) - 2(EY)a + a^2\right] =$

$= -2E(Y) + 2a = 0$ implies $a^* = E(Y)$ for the stationary point and obviously a^* gives rise to a minimum.

b) If $E|Y| < \infty$ then the constant b^* that minimizes $E|Y-b| \xrightarrow[b \in R^1]{}$ is $b^* = \text{median}(Y)$.

This is because

$$\frac{\partial}{\partial b}\left(E|Y-b|\right) = \frac{\partial}{\partial b}\left[\int_{-\infty}^{b}(b-y)f(y)dy + \int_{b}^{\infty}(y-b)f(y)dy\right] =$$

$$= \frac{\partial}{\partial b}\left[b\,F(b) - \int_{-\infty}^{b} y f(y)dy + \int_{b}^{\infty} y f(y)dy - b\left(1-F(b)\right)\right] =$$

$$= F(b) + b\cancel{f(b)} - b\cancel{f(b)} - b\cancel{f(b)} - \left(1-F(b)\right) + b\cancel{f(b)} =$$

$= 2F(b) - 1 = 0$ implies $F(b^*) = \frac{1}{2}$ for the stationary point, i.e. b^* is the median. And obviously b^* gives rise to a minimum.

The proof of Theorem 2.5 is also presented in details in the notes.

Then I was illustrating applications of Theorems 2.4 (estimation) and 2.5 (hypothesis testing) in Bayesian context.

Example 2.5.11 about Bayesian estimator of the parameter θ of the Bernoulli distribution when using a Beta(α,β) prior, leading to $\hat{\theta}_{Bayes} = \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha+\beta+n}$ by following step-by-step the theory from Theorem 2.4 is thoroughly presented in the notes. However, I stressed on the fact that an easier derivation that does not need the calculation of the marginal distribution $g(x)$ of the data, is recommended and is often applied in Bayesian inference.

Namely: knowing that $g(x)$ serves just as a norming constant for the conditional density of $\theta|X$:

$$h(\theta|x) = \frac{f(x|\theta)\tilde{\tau}(\theta)}{g(x)} \propto f(x|\theta)\tilde{\tau}(\theta),$$

we only needed to examine the expression on the top: $f(x|\theta)\tilde{\tau}(\theta) = \theta^{\sum_{i=1}^{n} x_i + \alpha - 1}(1-\theta)^{n-\sum_{i=1}^{n} x_i + \beta - 1}$

This already identifies $h(\theta|x)$ as Beta$\left(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta\right)$

But for any beta distributed random variable Y with parameters κ, β, is known that $EY = \frac{\kappa}{\kappa+\beta}$ holds.

Hence we get immediately that $\hat{\theta}_{Bayes} = E(\theta|x) = \frac{\sum_{i=1}^{n} x_i + \alpha}{\sum_{i=1}^{n} x_i + \alpha + n - \sum_{i=1}^{n} x_i + \beta}$

$$= \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha+\beta+n}$$

I also noticed that as $n \to \infty$ $\hat{\theta}_{Bayes} = \frac{\bar{x} + \frac{\alpha}{n}}{1 + \frac{\alpha+\beta}{n}}$ becomes very close to $\bar{x}$ as expected.

I also discussed the "approach based on α" once again in Problem 3 from the Set 1 of tutorial exercises. There we have $h(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{g(x)} \propto \theta^n e^{-\left(\sum_{i=1}^n x_i + K\right)\theta}$. Comparing this conditional density with the Gamma density (the latter being defined as $Y \sim \text{Gamma}(\alpha, \beta)$ having density

$$f(y) = \frac{y^{\alpha-1} e^{-y/\beta}}{\Gamma(\alpha)\,\beta^{\alpha}}, \quad y > 0$$

and $EY = \alpha\beta$) we see immediately that $h(\theta|x)$ must be $\text{Gamma}\left(n+1, \frac{1}{\sum_{i=1}^n x_i + K}\right)$ and hence $\hat{\theta}_{\text{Bayes}} = \frac{n+1}{\sum_{i=1}^n x_i + K}$ must hold.

I then discussed Question 6 from the Set 1 to illustrate that the "approach based on α" also helps a lot in Bayesian hypothesis testing. I discussed in detail on the white board the solution to Question 6. However, the solution is presented very thoroughly also in the file containing the solutions to the Set 1 of tutorial exercises. This set is available on moodle hence I abstain from reproducing the solution to Question 6 once again here.

Some white board writing from week 5

I continued lectering about inference principles. I discussed the weak likelihood principle. I also presented an example with 3 different experiments that give rise to proportional likelihoods. These discussions are presented in details in the notes and I abstain from reproducing them here.

Then I introduced the notion of (Fisher) Information as the variance of the score: Information in the sample $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ of n i.i.d. observations from distribution with a density $f(x,\theta)$, $\theta \in R^1$ as

$$I_X(\theta) = Var_\theta(V(X,\theta)) = E_\theta \left(\frac{\partial}{\partial\theta} \ln L(X,\theta)\right)^2 =$$

$$= E_\theta \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i,\theta)\right)^2.$$

I also discussed the properties of the Information quantity introduced above such as:

i) additivity over independent samples

ii) preservation of information by a sufficient statistic, i.e. $I_T(\theta) = I_X(\theta)$ when T is sufficient

iii) alternative way to calculate the information in the sample: under smoothness regularity conditions:

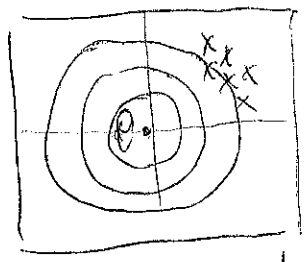
$$I_X(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln L(X,\theta)\right)$$

iv) For <u>any</u> statistic $T(X)$ it holds $I_T(\theta) \le I_X(\theta)$, with equality if and only if $T(X)$ is sufficient for θ.

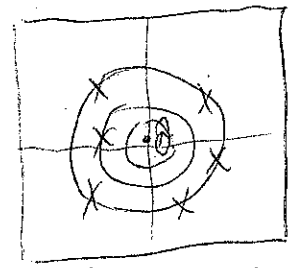
I went through the proofs of all these statements but the proofs are thoroughly presented in the notes and I am not reproducing them here.

Then I started discussing unbiasedness and CR inequality.

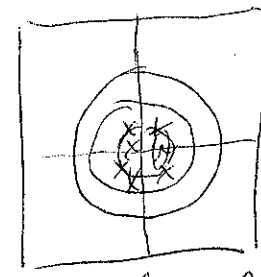
1) First I discussed the relevance of the notion of unbiasedness:



Biased estimator



Unbiased estimator with high variance



Unbiased estimator with small variance

Thanks to the decomposition

$$MSE_\theta(T_n) = E_\theta(T_n - \theta \pm ET_n) = Var_\theta T_n + (b_n(\theta))^2$$

and the graphs above, it does make sense to look for the the estimator with the smallest possible variance in the class of unbiased estimators.

I did make a cautions remark that sometimes an unbiased estimator may not be that useful: Take $f(x,\theta) = \theta(1-\theta)^{x-1}$, $x = 1,2,\dots$ & $T(x)$ based on $n=1$ observation would mean that $\sum_{x=1}^{\infty} T(x)\theta(1-\theta)^{x-1} = \theta$ holds $\forall\, \theta \in (0,1)$. Cancel θ set $\eta = 1-\theta$ and we see:

$$T(1) + \eta T(2) + \eta^2 T(3) + \dots = 1 \quad \forall\ \eta \in (0,1)$$ must hold. Hence $T(1) = 1$ and $T(2) = T(3) = T(4) = \dots = 0$!

Note: the estimator $\tilde{T}(x) = \frac{1}{x}$ (which happens to be the MLE in this example) is much more useful.

2) Example illustrating that when condition (*) is violated, we could have estimators which are unbiased and have a variance < CR bound:

Let $X_1, X_2, \dots, X_n$ be i.i.d. uniform in $(0,\theta)$ (so that the support of the density depends on θ & (*) is violated)

$$f_{X_1}(x) = \frac{1}{\theta} I_{(0,\theta)}(x) \longrightarrow F_{X_1}(x) = \begin{cases} 0 & x \le 0 \\ \frac{x}{\theta} & 0 < x \le \theta \\ 1 & x > \theta \end{cases}$$ Then:

$$F_{X_{(n)}}(y) = P(X_{(n)} \le y) = P(X_1 < y \cap X_2 < y \cap \dots X_n < y) = \left(\frac{y}{\theta}\right)^n, \quad 0 < y < \theta$$

Hence $f_{X_{(n)}}(y) = \frac{n y^{n-1}}{\theta^n}$, $0 < y < \theta$ (and zero else)

Take $EX_{(n)} = \int_0^\theta y \frac{n y^{n-1}}{\theta^n} dy = \frac{n}{n+1}\theta + \theta$,

i.e. $X_{(n)}$ is biased for estimating θ. BUT:

$T = \frac{n+1}{n} X_{(n)}$ is unbiased for estimating θ.

$\mathrm{Var}\, T = E(T^2) - \theta^2 = \left(\frac{n+1}{n}\right)^2 \int_0^\theta y^2 \frac{n y^{n-1}}{\theta^n} dy - \theta^2 = \dots = \frac{\theta^2}{n(n+2)}$

But for $f_{X_1}(\theta) = \frac{1}{\theta}, 0 < \theta$ we have $\ln f_{X_1}(\theta) = -\ln \theta$

$\frac{\partial}{\partial \theta} \ln f_{X_1}(\theta) = -\frac{1}{\theta}$ $\quad E\left(\frac{\partial}{\partial \theta} \ln f_{X_1}(\theta)\right)^2 = \frac{1}{\theta^2}$ and

reckless application of CR Bound would imply

$\mathrm{CR\,Bound} = \frac{\theta^2}{n}$. As we see now:

$\mathrm{Var}(T) = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}$

Reason for this seeming contradition: the condition (*) was violated in this example!

3) Score function for the Poisson(θ) example:

$X_1, X_2, \dots, X_n$ i.i.d. Poisson(θ)

$P(X_i = x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0, 1, 2, \dots$

$L(\mathbf{X}, \theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$. Hence

$V(\mathbf{X}, \theta) = \frac{\partial}{\partial \theta} \log L(\mathbf{X}, \theta) = -n + \frac{\sum_{i=1}^n x_i}{\theta}$

If $\tau(\theta) = \theta \to V(\mathbf{X}, \theta) = \frac{n}{\theta}(\bar{X} - \theta) \to$ factorization possible and $\bar{X}$ is the UMVUE of θ that attains the CR Bound