Homework 6

Aleck Zhao

November 2, 2016

1. Prove the converse of the factorization theorem, namely prove that if T is a sufficient statistic, then the joint density can be factored as

$$f(x_1, \cdots, x_n \mid \theta) = g(T, \theta)h(x_1, \cdots, x_n)$$

Also show that if T is sufficient for θ, then the MLE must be a function of T.

Proof. If T is a sufficient statistic, then by definition the distribution of $X_1, \cdots, X_n$ given $T = t$ does not depend on θ and only depends on the data $X_1, \cdots X_n$. Thus, let

$$f(X_1, \cdots, X_n \mid T) = h(X_1, \cdots, X_n)$$

be a function of just the data. Next, let

$$g(T, \theta) = g(T(X_1, \cdots, X_n), \theta) = f(T(X_1, \cdots, X_n) \mid \theta) = f(T \mid \theta)$$

be the density of T. Thus, we have

$$\begin{aligned} g(T, \theta)h(X_1, \cdots, X_n) &= f(X_1, \cdots, X_n \mid T)f(T \mid \theta) \\ &= f(X_1, \cdots, X_n \mid \theta) \end{aligned}$$

as desired.

□

2. Let $\hat{\theta}$ be an estimator for a parameter θ, and suppose that $\text{Var}(\hat{\theta}) < \infty$. Let T be a sufficient statistic for θ. Consider the random variable

$$Y = E[\hat{\theta} \mid T]$$

Prove that

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

Explain why this suggests that a sufficient statistic can be particularly useful in parametric estimation.

Proof. We have

$$E[Y] = E\left[E[\hat{\theta} \mid T]\right] = E[\hat{\theta}]$$

so

$$\begin{aligned} E\left[(Y - \theta)^2\right] &= E[Y^2] - 2\theta E[Y] + \theta^2 \\ &= E[Y^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ E\left[\left(\hat{\theta} - \theta\right)^2\right] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

so we must show that $E[Y^2] \leq E[\hat{\theta}^2]$.

We have

$$E[\hat{\theta}^2] - E[\theta]^2 = \mathrm{Var}(\hat{\theta}) = \mathrm{Var}(E[\hat{\theta} \mid T]) + E[\mathrm{Var}(\hat{\theta} \mid T)]$$

Thus, we have

$$\begin{aligned} E[\hat{\theta}^2] &= E[\hat{\theta}]^2 + \mathrm{Var}(Y) + E[\mathrm{Var}(\hat{\theta} \mid T)] \\ &= E[\hat{\theta}]^2 + E[Y^2] - E[Y]^2 + E[\mathrm{Var}(\hat{\theta} \mid T)] \\ &= E[Y^2] + E[\mathrm{Var}(\hat{\theta} \mid T)] \end{aligned}$$

since $E[\hat{\theta}] = E[Y]$. Since a variance is always non-negative, it follows that $E[\mathrm{Var}(\hat{\theta} \mid T)] \geq 0$, so

$$E[\hat{\theta}^2] \geq E[Y^2]$$

as desired.

We have $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ is a function of the data. Thus, the density of $\hat{\theta}$ is the same as the joint density $f(X_1, \cdots, X_n)$. Since T is sufficient for θ, it follows that $f(X_1, \cdots, X_n \mid T)$ is a function of purely the data, and therefore, the density of $(\hat{\theta} \mid T)$ is a function of purely the data. Thus, $E[\hat{\theta} \mid T]$ does not depend on θ, so it can be said to be an estimator for θ. Our result shows that no matter what $\hat{\theta}$ is, the MSE of $E[\hat{\theta} \mid T]$ is at most the MSE of $\hat{\theta}$, so it is a better estimator.

□

3. Complete all the details of the example we discussed in lecture. Let $X_1, \cdots, X_n$ be iid data from a normal distribution with unknown mean θ and known variance σ^2. Suppose that θ is assumed to be random, with prior distribution also normal; assume that the mean and variance of the prior distribution of θ_0 and σ_{pr}^2, where both θ_0 and σ_{pr}^2 are known.

   (a) Compute the posterior distribution

   $$f_{\theta \mid \boldsymbol{X}}(\theta \mid x_1, \cdots, x_n)$$

   where $\boldsymbol{X} = (X_1, \cdots, X_n)$, and specify all the parameters of this distribution.

   Solution. The posterior distribution is given by

$$\begin{aligned} f_{\theta \mid X}(\theta \mid X_1, \cdots, X_n) &= \frac{f_{X\mid\theta}(X_1, \cdots, X_n \mid \theta) f(\theta)}{\displaystyle\int f_{X\mid\theta}(X_1, \cdots, X_n \mid \theta) f(\theta)\, d\theta} \\ &= \frac{\displaystyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \theta)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma_{pr}} \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_{pr}^2}\right)}{\displaystyle\int \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \theta)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma_{pr}} \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_{pr}^2}\right) d\theta} \\ &= \frac{\exp\left(-\dfrac{1}{2\sigma^2} \displaystyle\sum_{i=1}^{n} (X_i - \theta)^2 - \dfrac{1}{2\sigma_{pr}^2} (\theta - \theta_0)^2\right)}{\displaystyle\int \exp\left(-\dfrac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2 - \dfrac{1}{2\sigma_{pr}^2} (\theta - \theta_0)^2\right) d\theta} \end{aligned}$$

   The numerator can be expanded to

$$\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} X_i^2 - 2\theta \sum_{i=1}^{n} X_i + n\theta^2\right] - \frac{1}{2\sigma_{pr}^2}\left[\theta^2 - 2\theta_0\theta + \theta_0^2\right]\right)$$

and similarly for the integrand in the denominator, and in this case, we may cancel anything not involving θ since the integral treats those as constants. We may write $\sum X_i = n\bar{X}$ so the numerator (and integrand) is

$$\exp\left(\frac{2\theta n\bar{X} - n\theta^2}{2\sigma^2} + \frac{2\theta_0\theta - \theta^2}{2\sigma_{pr}^2}\right)$$

From here, we wish to complete the square within the exponent with respect to θ :

$$\begin{aligned}
\frac{2\theta n\bar{X} - n\theta^2}{2\sigma^2} + \frac{2\theta_0\theta - \theta^2}{2\sigma_{pr}^2} &= \frac{2\theta n\bar{X}\sigma_{pr}^2 - n\theta^2\sigma_{pr}^2 + 2\theta_0\theta\sigma^2 - \sigma^2\theta^2}{2\sigma^2\sigma_{pr}^2} \\
&= -\frac{1}{2\sigma^2\sigma_{pr}^2}\left[\theta^2\left(n\sigma_{pr}^2 + \sigma^2\right) - \theta\left(2n\bar{X}\sigma_{pr}^2 + 2\theta_0\sigma^2\right)\right] \\
&= -\frac{n\sigma_{pr}^2 + \sigma^2}{2\sigma^2\sigma_{pr}^2}\left[\theta^2 - 2\left(\frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}\right)\theta\right] \\
&= -\frac{n\sigma_{pr}^2 + \sigma^2}{2\sigma^2\sigma_{pr}^2}\left[\left(\theta - \frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}\right)^2 - \left(\frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}\right)^2\right]
\end{aligned}$$

Note that the final term does not include any θ, so since this same expression is in the denominator, it will also cancel. Finally, the numerator simplifies to

$$\exp\left(-\frac{n\sigma_{pr}^2 + \sigma^2}{2\sigma^2\sigma_{pr}^2}\left(\theta - \frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}\right)^2\right) = \exp\left(-\frac{\left(\theta - \frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}\right)^2}{2 \cdot \frac{\sigma^2\sigma_{pr}^2}{n\sigma_{pr}^2 + \sigma^2}}\right)$$

If we take

$$\begin{aligned}
\theta_{post} &= \frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2} \\
\sigma_{post}^2 &= \frac{\sigma^2\sigma_{pr}^2}{n\sigma_{pr}^2 + \sigma^2}
\end{aligned}$$

then this expression is just missing the factor of

$$\frac{1}{\sqrt{2\pi}\sigma_{post}}$$

in front of the exponential. Then if we do this in the denominator, the integral evaluates to 1 since it is a normal density, and finally the posterior distribution of θ is given by a normal distribution with the above parameters.

□

(b) For what value of θ is this posterior density maximized? Given this, what would you choose as an estimate for θ?

Solution. The value of θ that maximizes this posterior density is clearly the posterior mean, thus

$$\theta = \frac{n\bar{X}\sigma_{pr}^2 + \theta_0\sigma^2}{n\sigma_{pr}^2 + \sigma^2}$$

is the desired estimate (note the dependence on the data in $\bar{X}$)

□

(c) How do the prior variance σ_{pr}^2 and the posterior variance compare? Which one is larger? Does this make sense? Why?

Solution. The posterior variance is given by

$$\sigma_{post}^2 = \frac{\sigma^2 \sigma_{pr}^2}{n\sigma_{pr}^2 + \sigma^2}$$

which less than σ_{pr}^2. We can confirm this by clearing the denominators. It makes sense that the posterior variance is smaller because the data should have given us a better idea of what the actual value is.

$\square$

(d) How does the estimator you obtained in part b compare to the MLE?

Solution. The MLE for the mean of a normal distribution is simply the sample mean. For large n, the sample mean should approach its true value of θ_0, so these estimators are asymptotically the same.

$\square$

4. Suppose we are in the Bayesian framework and we wish to estimate a parameter θ with prior distribution f from some family of distributions G. If, conditional on the value of the parameter, the data have some distribution H and the posterior distribution is again in the family G, we say that G and H are conjugate.

(a) Show that if X_i are iid Bernoulli (p) and p has a Beta-distributed prior, so that

$$f_p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where, as usual,

$$\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}\,dt$$

then the Bernoulli and Beta families are conjugate.

Proof. The distribution of X_i given p is a Bernoulli p distribution.
We have the posterior distribution of p is

$$\begin{aligned}
f_{P|X}(p \mid x) &= \frac{f_{X|P}(x_1, \cdots, x_n \mid p) f(p)}{\int f_{X|P}(x_1, \cdots, x_n \mid p) f(p)\,dp} \\
&= \frac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\int \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}} \\
&= \frac{p^{\sum x_i}(1-p)^{n-\sum x_i} p^{\alpha-1}(1-p)^{\beta-1}}{\int p^{\sum x_i}(1-p)^{n-\sum x_i} p^{\alpha-1}(1-p)^{\beta-1}} \\
&= \frac{p^{\alpha+\sum x_i - 1}(1-p)^{\beta+n-\sum x_i - 1}}{\int p^{\alpha+\sum x_i - 1}(1-p)^{\beta+n-\sum x_i - 1}\,dp} \\
&= \frac{\frac{\Gamma(\alpha+\sum x_i)\Gamma(\beta+n-\sum x_i)}{\Gamma(\alpha+\beta+n)} p^{\alpha+\sum x_i - 1}(1-p)^{\beta+n-\sum x_i - 1}}{\int \frac{\Gamma(\alpha+\sum x_i)\Gamma(\beta+n-\sum x_i)}{\Gamma(\alpha+\beta+n)} p^{\alpha+\sum x_i - 1}(1-p)^{\beta+n-\sum x_i - 1}\,dp}
\end{aligned}$$

The integrand in the denominator evaluates to 1 since it is the density of the Beta with parameters $\alpha + \sum x_i$ and $\beta + n - \sum x_i$. Thus, the posterior distribution of p is this same Beta distribution. Thus, the Bernoulli and Beta families are conjugate, as desired.

$\square$

(b) What if the X_i are binomial with parameters n, p where n is known and p has, again, a Beta distribution? Are the binomial and Beta families conjugate?

Solution. The distribution of X_i given p is a Binomial n, p distribution.
We have the posterior distribution of p is

$$
\begin{aligned}
f_{P|X}(p \mid x) &= \frac{f_{X|P}(x_1, \cdots, x_n \mid p) f(p)}{\int f_{X|P}(x_1, \cdots, x_n \mid p) f(p)\, dp} \\
&= \frac{\displaystyle\prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\displaystyle\int \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\, dp} \\
&= \frac{p^{\sum x_i}(1-p)^{n^2 - \sum x_i} p^{\alpha-1}(1-p)^{\beta-1}}{\displaystyle\int p^{\sum x_i}(1-p)^{n^2 - \sum x_i} p^{\alpha-1}(1-p)^{\beta-1}\, dp} \\
&= \frac{p^{\alpha + \sum x_i - 1}(1-p)^{\beta + n^2 - \sum x_i - 1}}{\displaystyle\int p^{\alpha + \sum x_i - 1}(1-p)^{\beta + n^2 - \sum x_i - 1}\, dp}
\end{aligned}
$$

Note the similarity to the form in the previous problem, so we may conclude this has a Beta distribution with parameters $\alpha + \sum x_i$ and $\beta + n^2 - \sum x_i$.
Thus, the binomial and Beta families are conjugate.

$\square$

(c) Show that if X_i are iid exponential with parameter λ, and λ has a Gamma-distributed prior, then the posterior also has a Gamma distribution. What is a reasonable estimate for λ in this Bayesian setting? How doe sit compare to the MLE for the exponential?

Proof. Suppose the distribution of λ is given by

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$

The posterior distribution of λ is given by

$$\begin{aligned}
f_{L|X}(\lambda \mid x) &= \frac{f_{X|L}(x_1, \cdots, x_n \mid \lambda) f(\lambda)}{\int f_{X|L}(x_1, \cdots, x_n \mid \lambda) f(\lambda)\, d\lambda} \\
&= \frac{\displaystyle\prod_{i=1}^n \lambda e^{-\lambda x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}}{\displaystyle\int \prod_{i=1}^n \lambda e^{-\lambda x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}\, d\lambda} \\
&= \frac{e^{-\lambda \sum x_i} \lambda^{\alpha-1} e^{-\lambda\beta}}{\displaystyle\int e^{-\lambda \sum x_i} \lambda^{\alpha-1} e^{-\lambda\beta}\, d\lambda} \\
&= \frac{e^{-\lambda\left(\beta + \sum x_i\right)} \lambda^{\alpha-1}}{\displaystyle\int e^{-\lambda\left(\beta + \sum x_i\right)} \lambda^{\alpha-1}\, d\lambda} \\
&= \frac{\frac{\left(\beta + \sum x_i\right)^\alpha}{\Gamma(\alpha)} e^{-\lambda\left(\beta + \sum x_i\right)} \lambda^{\alpha-1}}{\displaystyle\int \frac{\left(\beta + \sum x_i\right)^\alpha}{\Gamma(\alpha)} e^{-\lambda\left(\beta + \sum x_i\right)} \lambda^{\alpha-1}\, d\lambda}
\end{aligned}$$

Note that the denominator is the Gamma distribution with parameters α and $\beta + \sum x_i$, so it evaluates to 1. Thus, the posterior distribution of λ is this same Gamma distribution (which is specified in the numerator).

A reasonable estimate for λ is the mean of this distribution, which is given by

$$\hat{\lambda} = \frac{\alpha}{\beta + \sum x_i}$$

The MLE for the exponential is given by

$$\hat{\lambda} = \frac{n}{\sum x_i}$$

so these are plausibly similar.

□

5. Suppose we observe an iid sample $X_1, \cdots, X_n$ from the distribution that is uniform in the interval $[-\theta, \theta]$ for some unknown $\theta > 0$.

   (a) Find the MLE for θ.

   Solution. Since these are uniform variables, we must have

   $$\theta \geq \max_{1 \leq i \leq n} |X_i|$$

   otherwise there would be an impossible data element. The likelihood function is given by

   $$f(X_1, \cdots, X_n \mid \theta) = \prod_{i=1}^{n} \frac{1}{2\theta} = \frac{1}{2^n \theta^n}$$

   which is a decreasing function in θ, so the MLE is in fact given by

   $$\hat{\theta} = \max_{1 \leq i \leq n} |X_i|$$

   □

   (b) Show that the pair $T = \max\{X_1, \cdots, X_n\}$ and $S = \min\{X_1, \cdots, X_n\}$ are sufficient for θ.

   Proof. The distribution of X_i given T and S is simply a uniform distribution from S to T. We know they were initially drawn from a uniform distribution, but we don't know anything about its endpoints, so if we are given S and T as the endpoints, the distribution is uniform $[S, T]$, which in particular does not depend on θ. Thus, T and S are sufficient for θ.

   □

6. Suppose (U, V) is a uniformly distributed point in the unit circle $\{ (x, y) \mid x^2 + y^2 \leq 1 \}$ in the plane.

   (a) Determine the marginal PDFs of U and V and expectations $E[U]$ and $E[V]$. Also determine the covariance $\operatorname{Cov}(U, V)$ and decide if U, V are independent.

   Solution. The area of the unit circle is π, so the joint density is given by

   $$f_{U,V}(u, v) = \frac{1}{\pi}$$

   The marginal density of u is given by

   $$f_U(u) = \int f_{U,V}(u, v)\, dv = \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} \frac{1}{\pi}\, dv = \frac{2\sqrt{1-u^2}}{\pi}$$

   Similarly, the marginal density of v is given by

   $$f_V(v) = \frac{2\sqrt{1-v^2}}{\pi}.$$

   It's easy to see that these densities are symmetric about the origin, so $E[U] = E[V] = 0$. The covariance is given by

   $$\begin{aligned}
   \operatorname{Cov}(U, V) &= E[UV] - E[U]E[V] = E[UV] \\
   &= \int \int uv \cdot f_{U,V}(u, v)\, dv\, du \\
   &= \frac{1}{\pi} \int_{-1}^{1} \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} uv\, dv\, du \\
   &= 0
   \end{aligned}$$

but the product of the marginal densities is

$$f_U(u)f_V(v) = \frac{2\sqrt{1-u^2}}{\pi} \cdot \frac{2\sqrt{1-v^2}}{\pi} = \frac{4(1-u^2)(1-v^2)}{\pi^2} \neq f_{U,V}(u,v)$$

so U and V are not independent.

□

(b) Let $W = U^2 + V^2$. Compute the density $f_W(w)$ for W.

Solution. Consider the probability $F_W(w) = P(W \leq w) = P(U^2 + V^2 \leq w)$. This is a circle of radius w centered at the origin, but $U^2 + V^2$ can be anywhere in the unit circle, so this probability is given by

$$P(W \leq w) = \frac{w^2\pi}{\pi} = w^2$$

so the density is given by

$$f_W(w) = \frac{d}{dw} F_W(w) = \frac{d}{dw}\left[w^2\right] = 2w, \quad 0 \leq 1 \leq w$$

□

(c) Let $R = \theta U$, and $T = \theta V$, where $\theta > 0$ is some non-random parameter. Compute the joint distribution of (R, T).

Solution. We have

$$f_{R,T}(r,t) = f_{U,V}(u,v)\left|\frac{d(u,v)}{d(r,t)}\right|$$

where $U = R/\theta$ and $V = T/\theta$, so the joint density of R, T is given by

$$f_{R,T}(r,t) = \frac{1}{\pi}\left|\begin{bmatrix} 1/\theta & 0 \\ 0 & 1/\theta \end{bmatrix}\right| = \frac{1}{\theta^2\pi}$$

□

7. Suppose we observe independent pairs (X_i, Y_i) where each (X_i, Y_i) has a uniform distribution in the circle of unknown radius θ and centered at $(0, 0)$ in the plane.

(a) Show that $(X_i/\theta, Y_i/\theta)$ has a uniform distribution in the unit circle, and find the PDF of $X_i^2 + Y_i^2$.

Proof. The joint density of X_i, Y_i is given by

$$f_{X_i,Y_i}(x_i, y_i) = \frac{1}{\theta^2\pi}$$

so letting $X_i = \theta A, Y_i = \theta B$, we have the joint density of A, B is

$$f_{A,B}(a,b) = f_{X_i,Y_i}(x_i,y_i)\left|\frac{d(x_i,y_i)}{d(a,b)}\right| = \frac{1}{\theta^2\pi}\left|\begin{bmatrix} \theta & 0 \\ 0 & \theta \end{bmatrix}\right| = \frac{1}{\pi}$$

which is exactly the joint density of a uniform distribution on the unit circle, as desired.
Let $W = X_i^2 + Y_i^2$. Then the CDF of W is given by

$$F_W(w) = P(W \leq w) = P(X_i^2 + Y_i^2 \leq w)$$

which is a circle of radius w centered on the origin, and since X_i, Y_i is uniformly distributed on a circle of radius θ, this probability is

$$F_W(w) = \frac{w^2\pi}{\theta^2\pi} = \frac{w^2}{\theta^2}.$$

Thus, the density of W is given by

$$f_W(w \mid \theta) = \frac{d}{dw} F_W(w) = \frac{d}{dw}\left[\frac{w^2}{\theta^2}\right] = \frac{2w}{\theta^2}, \quad 0 \le w \le \theta$$

$\square$

(b) Show that $(X_1^2 + Y_1^2, \cdots, X_n^2 + Y_n^2)$ is a sufficient statistic for θ.

Proof. Let $W_i = X_i^2 + Y_i^2$. Then the likelihood function is given by

$$f(W_1, \cdots, W_n \mid \theta) = \prod_{i=1}^n f(W_i \mid \theta) = \prod_{i=1}^n \frac{2W_i}{\theta^2} = \frac{2^n}{\theta^{2n}} \prod_{i=1}^n W_i$$

Since

$$T\left[(X_1, Y_1), \cdots, (X_n, Y_n)\right] = (X_1^2 + Y_1^2, \cdots, X_n^2 + Y_n^2) = (W_1, \cdots, W_n)$$

is a function of the data, we may write $h\left[(X_1, Y_1), \cdots, (X_n, Y_n)\right] = 1$ and

$$g(T, \theta) = g[(W_1, \cdots, W_n), \theta)] = \frac{2^n}{\theta^{2n}} \prod_{i=1}^n W_i$$

so T is a sufficient statistic by the Factorization theorem.

$\square$

(c) Find the MLE and determine its density function and its bias. Are the regularity assumptions were require on the MLE satisfied here?

Solution. As above, the joint density

$$f\left[(X_1, Y_1), \cdots, (X_n, Y_n) \mid \theta\right] = \prod_{i=1}^n f\left[(X_i, Y_i) \mid \theta\right] = \prod_{i=1}^n \frac{1}{\theta^2 \pi} = \frac{1}{\theta^{2n} \pi^n}$$

Since $X_i^2 + Y_i^2 \le \theta^2$, the MLE $\hat{\theta}$ is

$$\hat{\theta} = \max_{1 \le i \le n} \sqrt{X_i^2 + Y_i^2}$$

Consider the CDF of $\hat{\theta}$

$$\begin{aligned} F(t) = P(\hat{\theta} \le t) &= P\left(\max_{1 \le i \le n} \sqrt{X_i^2 + Y_i^2} \le t\right) \\ &= P\left(\sqrt{X_1^2 + Y_1^2}, \cdots, \sqrt{X_n^2 + Y_n^2} \le t\right) \\ &= \prod_{i=1}^n P\left(\sqrt{X_i^2 + Y_i^2} \le t\right) \\ &= \prod_{i=1}^n P(W_i \le t^2) = \prod_{i=1}^n \frac{t^2}{\theta^2} = \frac{t^{2n}}{\theta^{2n}} \end{aligned}$$

and the density of $\hat{\theta}$ is the derivative of this wrt to t :

$$f_{\hat{\theta}}(t) = \frac{\partial}{\partial t}\left[\frac{t^{2n}}{\theta^{2n}}\right] = \frac{2nt^{2n-1}}{\theta^{2n}}$$

Then $E[\hat{\theta}]$ is given by

$$\begin{aligned} E[\hat{\theta}] &= \int_0^\theta t \frac{2nt^{2n-1}}{\theta^{2n}}\, dt = \int_0^\theta \frac{2nt^{2n}}{\theta^{2n}}\, dt \\ &= \frac{2nt^{2n+1}}{\theta^{2n}(2n+1)}\bigg|_0^\theta = \frac{2n\theta}{(2n+1)} \end{aligned}$$

so the bias of $\hat{\theta}$ is

$$E[\hat{\theta}] - \theta = \frac{2n\theta}{2n+1} - \theta = -\frac{\theta}{2n+1}.$$

The support of the distribution of (X_i, Y_i) is

$$\{\, (x_i, y_i) \mid f[(x_i, y_i) \mid \theta] > 0 \,\} = \{\, (x_i, y_i) \mid 1/\theta^2\pi > 0 \,\}$$

which is the entire domain, and doesn't depend on θ, so MLE satisfies the regularity conditions.

□

(d) Compute the variance of the MLE and simplify it so that it is clear how this variance decays with the sample size n.

Solution. The variance of the MLE is given by

$$\mathrm{Var}(\hat{\theta}) = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

where

$$\begin{aligned} E[\hat{\theta}^2] &= \int_0^\theta t^2 \frac{2nt^{2n-1}}{\theta^{2n}}\, dt = \int_0^\theta \frac{2nt^{2n+1}}{\theta^{2n}}\, dt \\ &= \frac{2nt^{2n+2}}{\theta^{2n}(2n+2)}\bigg|_0^\theta = \frac{n\theta^2}{n+1} \end{aligned}$$

so the variance is

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \frac{n\theta^2}{n+1} - \left(\frac{2n\theta}{2n+1}\right)^2 \\ &= \theta^2\left(\frac{n}{n+1} - \frac{4n^2}{(2n+1)^2}\right) = \frac{n\theta^2}{(n+1)(2n+1)^2} \end{aligned}$$

Clearly, this diminishes very quickly as n increases.

□

(e) Find the MSE of the MLE. As $n \to \infty$, which term contributes more to the MSE, the squared bias or the variance?

Solution. The MSE is given by

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= \mathrm{Var}(\hat{\theta}) + \Big(E[\hat{\theta} - \theta]\Big)^2 \\ &= \frac{n\theta^2}{(n+1)(2n+1)^2} + \left(-\frac{\theta}{2n+1}\right)^2 \\ &= \frac{\theta^2}{(2n+1)^2}\left(\frac{n}{n+1} + 1\right) \qquad (1) \\ &= \frac{\theta^2}{(n+1)(2n+1)} \end{aligned}$$

In (1), since

$$\frac{n}{n+1} \to 1$$

as $n \to \infty$, the squared bias and the variance contribute equally to the MSE.

□

(f) Find a method of moments estimator for θ based on the X_i and call this $\hat{\theta}_X$.

Solution. The marginal density of X_i is given by

$$f_X(x) = \frac{2\sqrt{\theta^2 - x^2}}{\theta^2 \pi}$$

which is symmetric about the origin, so $\mu_1 = E[X_i] = 0$. Then

$$\mu_2 = E[X_i^2] = \int_{-\theta}^{\theta} x^2 \cdot \frac{2\sqrt{\theta^2 - x^2}}{\theta^2 \pi}\, dx = \frac{\theta^2}{4}$$

according to Wolfram, so the method of moments estimate is

$$\hat{\theta}_x = 2\sqrt{\hat{\mu}_2}.$$

□

(g) Compare the performance of the MLE and the method of moments estimator as follows: In R, do the following 10000 times. Sample the uniform distribution in the unit circle using a sample of size 10, and compute the three estimators (MLE, MoM X_i, MoM Y_i). Compute estimates of the bias, the variance, and the MSE of each. Estimate the correlation coefficient between $\hat{\theta}_x$ and $\hat{\theta}_y$. Assuming your estimate in the previous parts are correct, how much should we impove the variance of one of $\hat{\theta}_x$ or $\hat{\theta}_y$ by averaging them?

(h) Show that for the method of moments estimator and the MLE, is it the case that the distribution of $\hat{\theta}/\theta$ does not depend on θ. Explain why this means we can write

$$MSE_\theta(\hat{\theta}) = \theta^2 \left(MSE_{\theta=1}(\hat{\theta}) \right)$$

From this, explain why it suffices that we compare the two estimators when $\theta = 1$.

Proof. The MLE was

$$\hat{\theta} = \max_{1 \le i \le n} \sqrt{X_i^2 + Y_i^2}$$

so consider the probability

$$P(\hat{\theta}/\theta \le t) = P(\hat{\theta} \le t\theta) = \frac{(t\theta)^{2n}}{\theta^{2n}} = t^{2n}$$

from the density derived in part (c), so the density of $\hat{\theta}/\theta$ does not depend on θ. □

Chapter 9: Testing Hypotheses and Assessing Goodness of Fit

2. Which of the following hypotheses are simple, and which are composite?

   a. X follows a uniform distribution on $[0, 1]$.

      Answer. This is simple, since it specifies the entire distribution of X.

   b. A die is unbiased.

      Answer. This is simple, since it specifies the distribution of the roll (each has probability $1/6$).

   c. X follows a normal distribution with mean 0 and variance $\sigma^2 > 10$.

      Answer. This is composite, since the variance is not specified entirely.

   d. X follows a normal distribution with mean $\mu = 0$.

      Answer. This is composite, because the variance is not specified at all.

5. True or false, and state why:

   a. The significance level of a statistical test is equal to the probability that the null hypothesis is true.

      Answer. This is $\boxed{\text{false.}}$ The significance level is the probability of a type I error, not the probability the null hypothesis is true.

   b. If the significance level of a test is decreased, the power would be expected to increase.

      Answer. This is $\boxed{\text{false.}}$ Decreasing the significance level increases the chance of a type II error, which decreases the power.

   c. If a test is rejected at the significance level α, the probability that the null hypothesis is true equals α.

      Answer. This is $\boxed{\text{false.}}$ We know nothing about the probability that the null hypothesis is true, only that the test statistic falls into the rejection region.

   d. The probability that the null hypothesis is falsely rejected is equal to the power of the test.

      Answer. This is $\boxed{\text{false.}}$ Falsely rejecting the null hypothesis means rejecting the null when it is actually true. This is a type I error and is the significance level, not the power.

   e. A type I error occurs when the test statistic falls in the rejection region of the test.

      Answer. This is $\boxed{\text{false.}}$ When the statistic falls in the rejection region, we reject the null hypothesis. This says nothing about whether the null was actually true to begin with or not.

   f. A type II error is more serious than a type I error.

      Answer. This can't be answered definitively, but it is not in general true.

   g. The power of a test is determined by the null distribution of the test statistic.

      Answer. This is $\boxed{\text{false.}}$ The power of a test is determined by the alternative distribution.

   h. The likelihood ratio is a random variable.

      Answer. This is $\boxed{\text{true.}}$ It is a function of the data, which are random variables.

4. Let X have one of the following distributions:

X	H_0	H_A
x_1	0.2	0.1
x_2	0.3	0.4
x_3	0.3	0.1
x_4	0.2	0.4

a. Compare the likelihood ratio, Λ, for each possible value X and order the x_i according to Λ.

Solution. We have

$$\Lambda_1 = \frac{P(X = x_1 \mid H_0)}{P(X = x_1 \mid H_A)} = \frac{0.2}{0.1} = 2$$
$$\Lambda_2 = \frac{P(X = x_2 \mid H_0)}{P(X = x_2 \mid H_A)} = \frac{0.3}{0.4} = \frac{3}{4}$$
$$\Lambda_3 = \frac{P(X = x_3 \mid H_0)}{P(X = x_3 \mid H_A)} = \frac{0.3}{0.1} = 3$$
$$\Lambda_4 = \frac{P(X = x_4 \mid H_0)}{P(X = x_4 \mid H_A)} = \frac{0.2}{0.4} = \frac{1}{2}$$

So the ordering from least to greatest is

$$\Lambda_4 < \Lambda_2 < \Lambda_1 < \Lambda_3$$

□

b. What is the likelihood ratio test of H_0 versus H_A at the level $\alpha = 0.2$? What is the test at the level $\alpha = 0.5$?

Solution. If H_0 is true, then the PMF of Λ is given by

Λ	0.5	0.75	2	3
$p(\lambda)$	0.2	0.3	0.2	0.3

Thus, at the level $\alpha = 0.2$, we have

$$P(\Lambda \leq c \mid H_0) = 0.2$$

so $c \in [0.5, 0.75)$ defines the upper bound of the rejection region. We would reject if we picked x_4. At the level $\alpha = 0.5$, we have

$$P(\Lambda \leq c \mid H_0) = 0.5$$

so $c \in [0.75, 2)$. We would reject if we picked x_2 or x_4.

□

c. If the prior probabilities are $P(H_0) = P(H_A)$, which outcomes favor H_0?

Solution. If $P(H_0) = P(H_A)$, then the likelihood ratio is given by

$$\frac{P(X \mid H_0)}{P(X \mid H_A)} = \frac{P(H_A)}{P(H_0)} \cdot \frac{P(H_0 \mid X)}{P(H_A \mid X)} = \frac{P(H_0 \mid X)}{P(H_A \mid X)}$$

The outcomes that favor H_0 are the outcomes where this ratio is greater than 1. These occur exactly when the probability of x_i under H_0 is greater than the probability under H_a, which are when $X = x_1$ and $X = x_3$.

□

7. Let $X_1, \cdots, X_n$ be a sample from a Poisson distribution. Find the likelihood ratio for testing $H_0 : \lambda = \lambda_0$ versus $H_a : \lambda = \lambda_1$, where $\lambda_1 > \lambda_0$. Use the fact that the sum of independent Poisson random variables follows a Poisson distribution to explain how to determine a rejection region for a test at level α.

Solution. Let $S_n = \displaystyle\sum_{i=1}^{n} X_i$. The likelihood ratio is given by

$$\begin{aligned}
\Lambda &= \frac{P(X_1, \cdots, X_n \mid \lambda = \lambda_0)}{P(X_1, \cdots, X_n \mid \lambda = \lambda_1} \\
&= \frac{\displaystyle\prod_{i=1}^{n} \frac{\lambda_0^{X_i} e^{-\lambda_0}}{X_i!}}{\displaystyle\prod_{i=1}^{n} \frac{\lambda_1^{X_i} e^{-\lambda_1}}{X_i!}} \\
&= e^{-n(\lambda_0 - \lambda_1)} \left(\frac{\lambda_0}{\lambda_1}\right)^{S_n}
\end{aligned}$$

To determine a rejection region, we consider the probability

$$P(\Lambda \leq c \mid H_0) = \alpha$$

which is

$$\begin{aligned}
P(\Lambda \leq c \mid H_0) &= P\left(e^{-n(\lambda_0 - \lambda_1)} \left(\frac{\lambda_0}{\lambda_1}\right)^{S_n} \leq c \,\middle|\, \lambda = \lambda_0\right) \\
&= P\left(\left(\frac{\lambda_0}{\lambda_1}\right)^{S_n} \leq c e^{n(\lambda_0 - \lambda_1)}\right) \\
&= P\left(S_n \log\left(\frac{\lambda_0}{\lambda_1}\right) \leq n(\lambda_0 - \lambda_1) \log c\right) \\
&= P\left(S_n \geq n(\lambda_0 - \lambda_1) \frac{\log c}{\log\left(\frac{\lambda_0}{\lambda_1}\right)}\right) \\
&= P\left(S_n \leq n(\lambda_1 - \lambda_0) \frac{\log c}{\log \lambda_0 - \log \lambda_1}\right)
\end{aligned}$$

Since S_n is the sum of Poisson random variables, its density is given by

$$f(s) = \frac{(n\lambda_0)^s e^{-n\lambda_0}}{s!}$$

if we assume that $\lambda = \lambda_0$. We must have $c < 1$ otherwise the RHS will be negative, and the probability is 0. Suppose M is the largest integer less than or equal to the RHS, so the probability is

$$P\left(S_n \leq n(\lambda_1 - \lambda_0) \frac{\log c}{\log \lambda_0 - \log \lambda_1}\right) = \sum_{s=0}^{M} \frac{(n\lambda_0)^s e^{-n\lambda}}{s!} = \alpha$$

so we may solve explicitly for c in terms of α since M is a function of c.

□

9. Let $X_1, \cdots, X_{25}$ be a sample from a normal distribution having a variance of 100. Find the rejection region for a test at level $\alpha = 0.10$ of $H_0 : \mu = 0$ versus $H_A : \mu = 1.5$. What is the power of the test? Repeat for $\alpha = 0.01$.

Solution. The variance is 100, so the density is given by

$$f(x) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{200}\right)$$

Thus, the likelihood ratio is given by

$$\begin{aligned}
\Lambda &= \frac{P(X_1, \cdots, X_{25} \mid H_0)}{P(X_1, \cdots, X_{25} \mid H_A)} \\
&= \frac{\displaystyle\prod_{i=1}^{25} \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{X_i^2}{200}\right)}{\displaystyle\prod_{i=1}^{25} \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{(X_i-1.5)^2}{200}\right)} \\
&= \exp\left(-\frac{1}{200}\left[\sum_{i=1}^{25} X_i^2 - \sum_{i=1}^{25}(X_i - 1.5)^2\right]\right) \\
&= \exp\left(-\frac{1}{200}\left[3\sum_{i=1}^{25} X_i - 25(1.5)^2\right]\right) \\
&= \exp\left(-\frac{1}{200}\left[3(25\bar{X}) - 25(1.5)^2\right]\right) \\
&= \exp\left(\frac{2.25 - 3\bar{X}}{8}\right)
\end{aligned}$$

At the level $\alpha = 0.10$, we have the rejection region given by

$$\begin{aligned}
P(\Lambda \le c \mid H_0) &= P\left(\exp\left(\frac{2.25 - 3\bar{X}}{8}\right) \le c\right) \\
&= P\left(\frac{2.25 - 3\bar{X}}{8} \le \log c\right) \\
&= P\left(\bar{X} \ge \frac{2.25 - 8\log c}{3}\right) = 0.10
\end{aligned}$$

If we assume the null hypothesis to be true, that is $\mu = 0$, then the distribution of $\bar{X}$ is

$$N\left(0, \frac{10^2}{25}\right) = N(0, 2^2).$$

Thus, the probability is

$$\begin{aligned}
P\left(\bar{X} \ge \frac{2.25 - 8\log c}{3}\right) &= P\left(\frac{\bar{X}}{2} \ge \frac{2.25 - 8\log c}{6}\right) = 1 - P\left(\frac{\bar{X}}{2} < \frac{2.25 - 8\log c}{6}\right) \\
&= 1 - \Phi\left(\frac{2.25 - 8\log c}{6}\right) = 0.10
\end{aligned}$$

so at this point we may solve for c :

$$\begin{aligned}
\Phi\left(\frac{2.25-8\log c}{6}\right) &= 0.90 \\
\implies \frac{2.25-8\log c}{6} &= 1.282 \\
\implies c &= 0.506
\end{aligned}$$

Thus, we reject H_0 if $\Lambda \in (0, 0.506]$. Let β be the probability of a type II error, that is,

$$\beta = P(\Lambda > 0.506 \mid H_A)$$

If the alternative is true, then $\mu = 1.5$, and the distribution of $\bar{X}$ is $N(1.5, 2^2)$. Thus the probability is given by

$$\begin{aligned}
P(\Lambda > 0.506 \mid H_A) &= P\left(\exp\left(\frac{2.25-3\bar{X}}{8}\right) > 0.506\right) \\
&= P\left(\frac{2.25-3\bar{X}}{8} > \log 0.506\right) \\
&= P\left(\bar{X} < 2.567\right) \\
&= P\left(\frac{\bar{X}-1.5}{2} < \frac{2.567-1.5}{2}\right) \\
&= \Phi(0.533) = 0.71
\end{aligned}$$

so the power is $1-\beta = 1-0.71 = 0.29$.

At the level $\alpha = 0.01$, basically everything is the same, except the distribution of $\bar{X}$ is

$$N\left(0, \frac{10^2}{100}\right) = N(0,1)$$

so

$$\begin{aligned}
P\left(\bar{X} \geq \frac{2.25-8\log c}{3}\right) = 1-\Phi\left(\frac{2.25-8\log c}{3}\right) &= 0.01 \\
\Phi\left(\frac{2.25-8\log c}{3}\right) &= 0.99 \\
\implies \frac{2.25-8\log c}{3} &= 2.327 \\
\implies c &= 0.554
\end{aligned}$$

Thus, we reject H_0 if $\Lambda \in (0, 0.554]$. If the alternative is true, then the distribution of $\bar{X}$ is $N(1.5, 1)$, so the probability of a type II error is

$$\begin{aligned}
P(\Lambda > 0.554 \mid H_A) &= P\left(\exp\left(\frac{2.25-3\bar{X}}{8} > 0.554\right)\right) \\
&= P\left(\frac{2.25-3\bar{X}}{8} > \log 0.554\right) \\
&= P(\bar{X} < 2.325) \\
&= P(\bar{X}-1.5 < 2.325-1.5) \\
&= \Phi(0.825) = 0.795
\end{aligned}$$

so the power is $1-\beta = 1-0.795 = 0.205$.

□