Statistics II | **Problem Set 1**
Professor: Marcelo J. Moreira | **Solutions**
TA: Luan Borelli | **July 23, 2024**

---

1. [6.11, LNs] *Prove the following (unrelated) items:*

(a) *Let X have a continuous and strictly increasing distribution $F(x)$ and define the random variable Y as $F(X)$. Find the cdf and pdf of Y.*

Solution. Denote the cdf and the pdf of Y by $F_Y(y)$ and $f_Y(y)$, respectively.

$$F_Y(y) = P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y.$$

Thus the pdf is simply $f_Y(y) = F_Y'(y) = 1$. □

(b) *Let X and Y be independent random variables. Let $U = g(X)$ and $V = h(Y)$. Is U independent of V? Prove it or give a counter example.*

Solution. By definition, X and Y are independent $\iff$ for *all* measurable sets A and B the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent. From ($\implies$) it follows that, in particular, $(X^{-1} \circ g^{-1})(C) = X^{-1}(g^{-1}(C))$ and $(Y^{-1} \circ h^{-1})(D) = Y^{-1}(h^{-1}(D))$ are independent for any arbitrary measurable sets C and D. Since C and D are arbitrary, it follows from ($\impliedby$) that $(X^{-1} \circ g^{-1})^{-1} = g(X)$ and $(Y^{-1} \circ h^{-1})^{-1} = h(Y)$ are independent. In other words, the abstract definition of independence makes this assertion trivial. Two random variables are independent if and only if the sigma-algebras they generate are independent. Because the sigma-algebra generated by a measurable function of a sigma-algebra is a sub-sigma-algebra, it follows that any measurable functions of those random variables have independent sigma-algebras, whence those functions are independent. □

(c) *Let X and Y be two random variables with $E(X) = E(Y) = 0$. Assume that $E(XY)$ exists and $E(X|Y) = 0$. Show that X and Y are uncorrelated.*

Solution. $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|Y]] = \mathbb{E}[\mathbb{E}[X|Y]Y] = 0.$ □

---

2. [7.2, LNs] *This question is on an application of the Cramér-Wold device, and uses the (multivariate) continuity theorem. Let X_n, $1 \leq n \leq \infty$, be random vectors with characteristic function $\varphi_n(t)$. (i) if $X_n \xrightarrow{d} X$ then $\varphi_n(t) \to \varphi(t)$ for all t; and, (ii) if $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at zero, then $X_n \xrightarrow{d} X$ (with X having characteristic function $\varphi(t)$.*

(a) *Use characteristic functions to prove the Crámer-Wold device: a sequence of k-dimensional random vectors S_n, $n = 1, 2, \ldots$, converges in distribution to a random vector S if and only if $\alpha' S_n \xrightarrow{d} \alpha' S$ for every fixed vector $\alpha \neq 0$.*

Solution. If $S_n \xrightarrow{d} S$, then $\varphi_n(\tau) \to \varphi(\tau)$ for all $\tau \in \mathbb{R}^k$. In particular, we can let $\tau = \alpha t$ for any arbitrary $t \in \mathbb{R}$ and every fixed vector $\alpha \in \mathbb{R}^k \backslash \{0\}$ so that

$$\mathbb{E}[e^{it\alpha' S_n}] = \mathbb{E}[e^{i(\alpha t)' S_n}] = \varphi_n(\alpha t) \to \varphi(\alpha t) = \mathbb{E}[e^{i(\alpha t)' S}] = \mathbb{E}[e^{it\alpha' S}].$$

Observe that $\varphi_n(t\alpha)$ and $\varphi(t\alpha)$ are precisely the characteristic functions of $\alpha' S_n$ and $\alpha' S$, respectively, when viewed as functions of t only (i.e., given a fixed $\alpha \in \mathbb{R}^k$). Therefore $\alpha' S_n \xrightarrow{d} \alpha' S$. Conversely, if $\alpha' S_n \to \alpha' S$ for every fixed vector $\alpha \neq 0$, then

$$\mathbb{E}[e^{i\tau\alpha' S_n}] \to \mathbb{E}[e^{i\tau\alpha' S}] \quad \text{for all } \tau \in \mathbb{R}.$$

In particular, we can let $\tau = 1$ so that

$$\varphi_n(\alpha) = \mathbb{E}[e^{i\alpha' S_n}] \to \mathbb{E}[e^{i\alpha' S}] = \varphi(\alpha) \quad \text{for all } \alpha \in \mathbb{R}^k.$$

Observe that $\varphi_n(\alpha)$ and $\varphi(\alpha)$ are precisely the characteristic functions of S_n and S, respectively, as functions of $\alpha \in \mathbb{R}^k$. Therefore $S_n \xrightarrow{d} S$. □

(b) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random vectors of dimension k with $\mathbb{E}[X_i] = \mu$ and variance $V(X_i) = \mathbb{E}[(X_i - \mu)(X_i - \mu)'] = \Sigma$. The variance Σ is positive definite ($a'\Sigma a > 0$ for any $a \neq 0$). Find the limiting distribution of $\sqrt{n}(\bar{X}_n - \mu)$.*

Solution. Let $S_n = \sqrt{n}(\bar{X}_n - \mu)$ and observe that, by the *univariate* central limit theorem, for any $\alpha \in \mathbb{R}^k \backslash \{0\}$ we have $\alpha' S_n = \sqrt{n}(\alpha' \bar{X}_n - \alpha' \mu) \xrightarrow{d} N(0, \alpha' \Sigma \alpha)$. Notice that $N(0, \alpha'\Sigma\alpha) = \alpha' N(0, \Sigma) =: \alpha' S$, where $S \sim N(0, \Sigma)$, whence it follows, by **(a)**, that $S_n \xrightarrow{d} S$. We have just proved the *multivariate* central limit theorem. □

---

3. [7.9, LNs] *Let $X \sim N(\theta, 1)$. The density of X is $g(x;\theta) = (2\pi)^{-1/2} \exp\left(-(x-\theta)^2/2\right)$.*

(a) *Show that the density ratio $g(X;\theta)/h(x)$ has mean 1 when X is being drawn from the density $h(x)$.*

Solution.

$$\mathbb{E}\left[\frac{g(X;\theta)}{h(x)}\right] = \int_{-\infty}^{\infty} \frac{g(x;\theta)}{h(x)} h(x)\, dx = \int_{-\infty}^{\infty} g(x;\theta)\, dx = 1.$$

□

(b) *Let $X_i \overset{\text{iid}}{\sim} N(0,1)$. Report on a table the sample average you found of $n^{-1} \sum_{i=1}^{n} \frac{g(X_i;\theta)}{g(X_i;0)}$ for all combinations of $n = 10^j$ for $j = 1,2,3,4,5,6$ and $\theta = 0,1,2,3,4,5,6,7,8,9$. Here, $h(x)$ is the density of a standard normal.*

n/θ	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10	1	1.5284	0.4357	0.3125	0.0001	0.0536	0.000010	0.000000008	0.000000000022	0.000000002230
100	1	0.9669	0.9024	0.5465	0.0345	0.0916	0.000478	0.000056420	0.000000021593	0.000000002721
1000	1	0.9739	1.3913	0.7259	0.4612	0.1379	0.005234	0.000055074	0.000000429973	0.000000000436
10000	1	0.9908	1.1457	0.7672	0.3040	0.4634	1.412345	0.000375722	0.000026128897	0.000000022828
100000	1	1.0009	0.9915	0.9957	1.1448	1.7508	0.273413	0.060356958	0.000537437971	0.000016813355
1000000	1	1.0009	0.9915	0.9957	1.1448	1.7508	0.273413	0.0603569584	0.0005374379711	0.0000168133549

(c) *Explain your results by plotting densities of $N(0,1)$, $N(3,1)$, $N(6,1)$, and $N(9,1)$.*

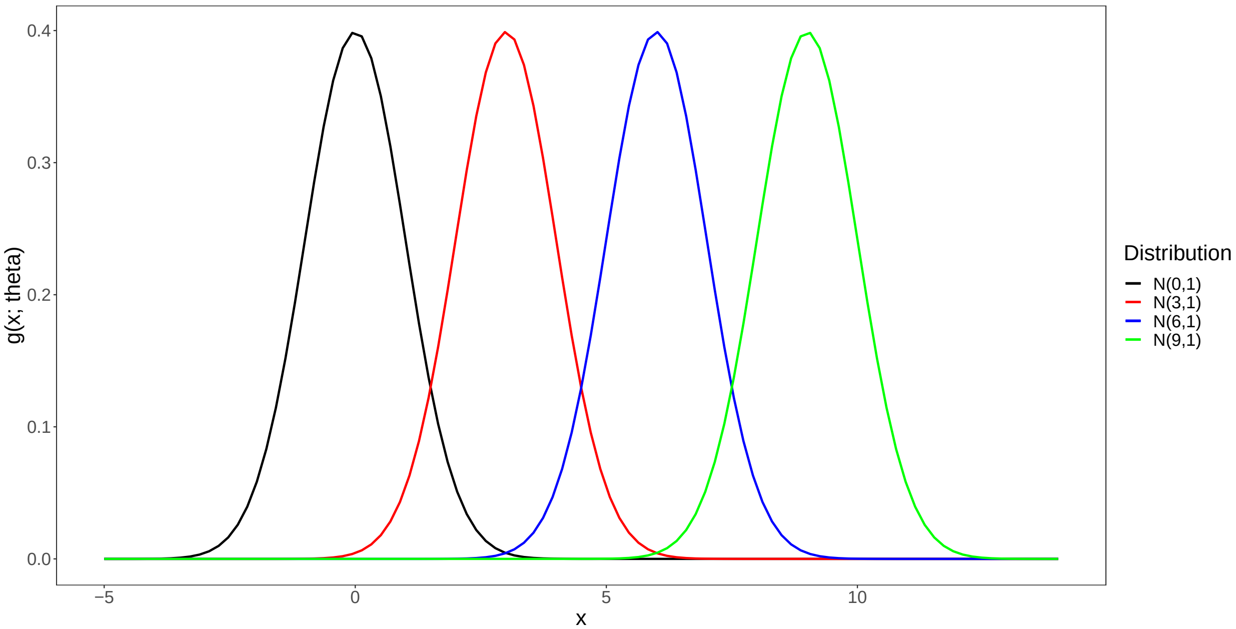
Solution. The goal of this question is to estimate the mean of a density ratio of the form $g(X;\theta)/h(X)$, where $X \sim N(0,1)$. From item **(a)**, we know that the true population mean of this density ratio is always 1, irrespective of the specific density function $h(X)$. Thus, according to the law of large numbers, we would expect the sample means from the results in table **(b)** to converge to 1 as the sample size n increases, regardless of the value of θ. Indeed, this convergence is observed in some cases, particularly for small values of θ (e.g., from 1 to 4). However, for larger values of θ, the convergence seems to break down. For instance, when $\theta = 9$, even with a very large sample size like $n = 1,000,000$, the estimated mean remains *extremely* distant from 1.[1]

What's happening? Is the law of large numbers failing for very high values of θ? Figure 1 can help us address this question. By looking at the density of $N(0,1)$ and comparing it to the density of $N(9,1)$, and considering the ratio $g(X_i;9)/g(X_i;0)$ as an example, we can observe that since $X_i \sim N(0,1)$, on average, the draws will be close to zero. Thus, for the vast majority of draws, $g(X_i;0)$ will take a high value, while $g(X_i;9)$ will take an extremely low value. As a result, in the vast majority of draws, the ratio $g(X_i;9)/g(X_i;0)$ will assume a very low value, virtually zero. However, in the occurrence of extremely rare events where positive draws of X_i deviate very far from zero, the logic will be reversed: $g(X_i;9)$ will take a high value, and $g(X_i;0)$ will take an extremely low value, causing the ratio to skyrocket to an *extremely* high value. This outlier will force the sample mean upwards, bringing it again closer to 1 and compensating for all the previous observations.

Now it becomes clear what is actually occurring: it is not that the law of large numbers fails for high values of θ; rather, the issue lies in the fact that, for high values of θ, the realizations of "crucial importance" for the convergence of the sample mean to the population mean are *extremely rare*. The convergence happens due to an exceedingly small number of extremely rare events, which carry a disproportionately significant impact compared to the rest of the events. As a consequence, in order to computationally observe the convergence for high values of θ, an enormously large number of observations would be required — practically approaching infinity! This is necessary to ensure that these exceptionally rare events happen frequently enough to "drive the convergence." However, from a computational standpoint, performing such an immense number of simulations can be impractical. □

[1] I even tested with an even larger sample size of $n = 50,000,000$, but there was hardly any difference. The estimate remains significantly far from 1.

Figure 1: Normal densities $g(x;\theta)$ with $\theta = 0, 3, 6$ and 9.



(d) *Show that* $h(x) = 10^{-1}\sum_{j=0}^{9} g(x;j)$ *is a density. Furthermore, how would you draw a random sample from* $h(x)$ *in practice?*

Solution. For the sake of simplicity, let's abstract from technicalities involving the formal definition of a density function by just assuming that for $h(x)$ to be a density it suffices to show that (i) $h(x)$ integrates to 1 over $(-\infty, \infty)$ and (ii) $h(x)$ is nonnegative for all x.

We have that

$$\int_{-\infty}^{\infty} 10^{-1}\sum_{j=0}^{9} g(x;j)\ dx = \int_{-\infty}^{\infty} 10^{-1}\sum_{j=0}^{9} \exp(-(x-\theta)^2/2)\ dx \tag{1}$$

$$= 10^{-1}\sum_{j=0}^{9}\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp(-(x-\theta)^2/2)\ dx \tag{2}$$

$$= 10^{-1}\sum_{j=0}^{9} 1 = 10^{-1}10 = 1. \tag{3}$$

Moreover, since g is a density, $g(x;j) \geq 0$ for all x and $j = 0, \ldots, 9$, whence it follows that $h(x) \geq 0$ for all x. Therefore, h is a density.

In order to draw a random sample from $h(x)$ we could use inverse transform sampling. In practice, we simply generate a draw u from a *discrete* uniform distribution in the interval $[0, 9]$ and then generate a draw from a normal distribution with mean u and variance 1; that is, a draw from $N(u, 1)$. The resulting draw will be equivalent to a draw from $h(x)$. □

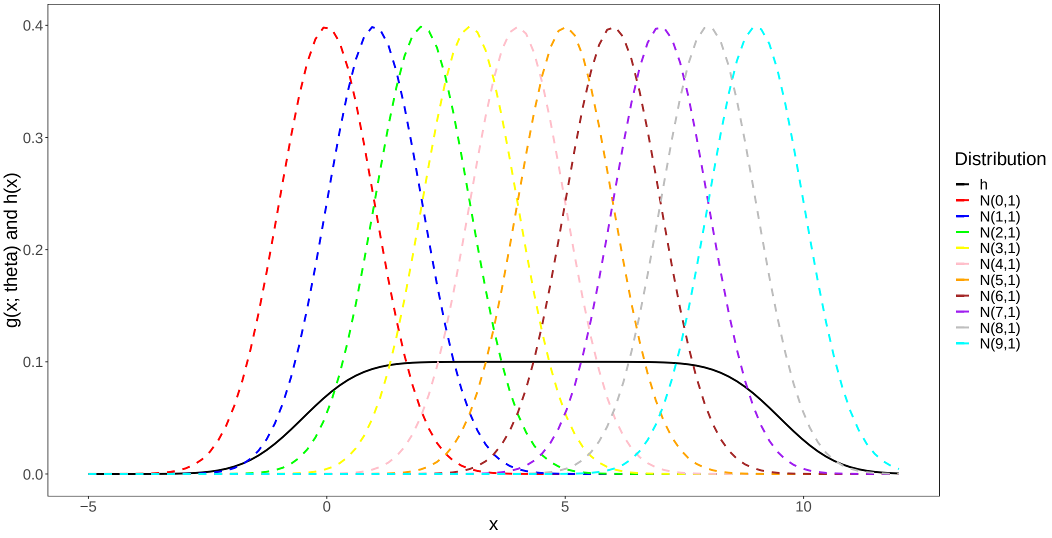
(e) *Report on a table the sample average you found of* $n^{-1}\sum_{i=1}^{n}\frac{g(X_i;\theta)}{h(X_i)}$ *for all combinations of* $n = 10^j$ *for* $j = 1, 2, 3, 4, 5, 6$ *and* $\theta = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. *Here* $h(x) = 10^{-1}\sum_{j=0}^{9} g(x; j)$.

n/θ	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10	0.713267	0.876365	0.619611	0.848035	0.998121	0.769224	0.751788	0.54929	0.641165	1.114872
100	1.068189	1.045083	1.052162	1.248939	1.022221	0.847806	0.97673	0.827878	1.182188	0.799998
1000	1.108903	0.994146	0.934179	0.980026	0.918329	1.070847	1.023598	1.035077	1.02438	0.987447
10000	0.99591	1.000614	1.000128	0.988394	0.984994	1.002482	0.983719	1.002187	1.003022	1.001579
100000	1.009509	1.001897	1.007725	0.999622	1.001571	0.996998	1.005517	0.999736	1.004651	1.003114
1000000	0.999541	0.998239	1.000453	1.000378	1.000569	0.998407	0.999694	0.999728	0.998736	0.99614

(f) *Explain your results by plotting the densities of* $h(x)$ *and of* $N(0, 1)$, $N(3, 1)$, $N(6, 1)$, *and* $N(9, 1)$.

Solution. One way to address the problem described in item **(c)** is to generate draws from a distribution that assigns greater "importance" (i.e., higher frequency) to the values that "truly matter" for determining the convergence of the sample mean of the density ratio. The distribution defined in item **(d)** accomplishes precisely that. Continuing with the example of the case $\theta = 9$ described in item **(c)**, when the draws are generated from $h(x)$ instead of $g(x; 0)$, higher values of the numerator $g(X_i; 9)$ occur more frequently, and the convergence of the sample mean to 1 happens without relying as much on the occurrence of extremely rare events. Figure 2 illustrates how the distribution $h(x)$ "prioritizes" the sampling in more important regions of $g(x; 9)$ for determining the convergence when compared to $g(x; 0)$. This method has a name: importance sampling. □

Figure 2: Normal densities $g(x; \theta)$ with $\theta = 1, \ldots, 9$ and mixture density $h(x)$.



4. [10.3, LNs] *Let $W_1, \dots, W_n$ be unbiased estimators of a parameter θ with $V(W_i) = \sigma_i^2$ and $C(W_i, W_j) = 0$ if $i \neq j$.*

(a) *Show that, of all unbiased estimators of the form $\sum_{i=1}^n a_i W_i$, where the a_i's are constants, the estimator*

$$W^* = \frac{\sum_{i=1}^n W_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}$$

has minimum variance.

Solution. First, notice that unbiasedness implies

$$\mathbb{E}\left[\sum_{i=1}^n a_i W_i\right] = \sum_{i=1}^n a_i \mathbb{E}[W_i] = \sum_{i=1}^n a_i \theta = \left(\sum_{i=1}^n a_i\right)\theta = \theta.$$

Thus we must have $\sum_{i=1}^n a_i = 1$. Further, observe that the zero-covariance assumption implies

$$V\left(\sum_{i=1}^n a_i W_i\right) = \sum_{i=1}^n a_i^2 V(W_i) = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

We want to minimize $\sum_{i=1}^n a_i^2 \sigma_i^2$ with respect to a_i subject to the constraint $\sum_{i=1}^n a_i = 1$. Let λ be the Lagrange multiplier for this problem. The first order conditions are

$$2a_i\sigma_i^2 - \lambda = 0,$$

whence

$$a_i = \frac{\lambda}{2}(1/\sigma_i^2). \tag{4}$$

Summing a_i over $i = 1, \dots, n$, we obtain

$$\sum_{i=1}^n a_i = \lambda \sum_{i=1}^n \frac{1}{2\sigma_i^2},$$

whence

$$\lambda = \frac{2}{\sum_{i=1}^n (1/\sigma_i^2)}.$$

Plugging λ into (4) we obtain

$$a_i^* \equiv \frac{(1/\sigma_i^2)}{\sum_{i=1}^n (1/\sigma_i^2)}.$$

Thus

$$W^* = \sum_{i=1}^n a_i^* W_i = \sum_{i=1}^n \frac{(1/\sigma_i^2)}{\sum_{i=1}^n (1/\sigma_i^2)} W_i = \frac{\sum_{i=1}^n W_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}, \tag{5}$$

as desired. □

(b) *Show that*

$$V(W^*) = \frac{1}{\sum_{i=1}^n 1/\sigma_i^2}.$$

Solution. Taking the variance of (5) we obtain

$$V(W^*) = \frac{\sum_{i=1}^n (1/\sigma_i^2)^2 \sigma_i^2}{\left(\sum_{i=1}^n 1/\sigma_i^2\right)^2} = \frac{\sum_{i=1}^n 1/\sigma_i^2}{\left(\sum_{i=1}^n 1/\sigma_i^2\right)^2} = \frac{1}{\sum_{i=1}^n 1/\sigma_i^2}.$$

□

---

5. [11.6, LNs] *Let (X_i, Y_i) be the age and unemployment duration for person i, for $i = 1, \dots, N$. Assume that age has a normal distribution with mean 40 and standard deviation 10. Given age X_i, Y_i, the duration of unemployment in weeks is assumed to have an exponential distribution with mean $\exp(\theta \cdot X_i)$. Assume that the observations for different individuals are independent. The observations are $(20, 9)$, $(45, 27)$, $(42, 26)$, $(32, 10)$, $(52, 41)$, $(32, 25)$, $(25, 19)$, $(23, 31)$, $(40, 32)$, $(46, 44)$.*

(a) *Plot the log likelihood function for θ between -1 and 1.*

Solution. By the definition of conditional probability density functions we can write

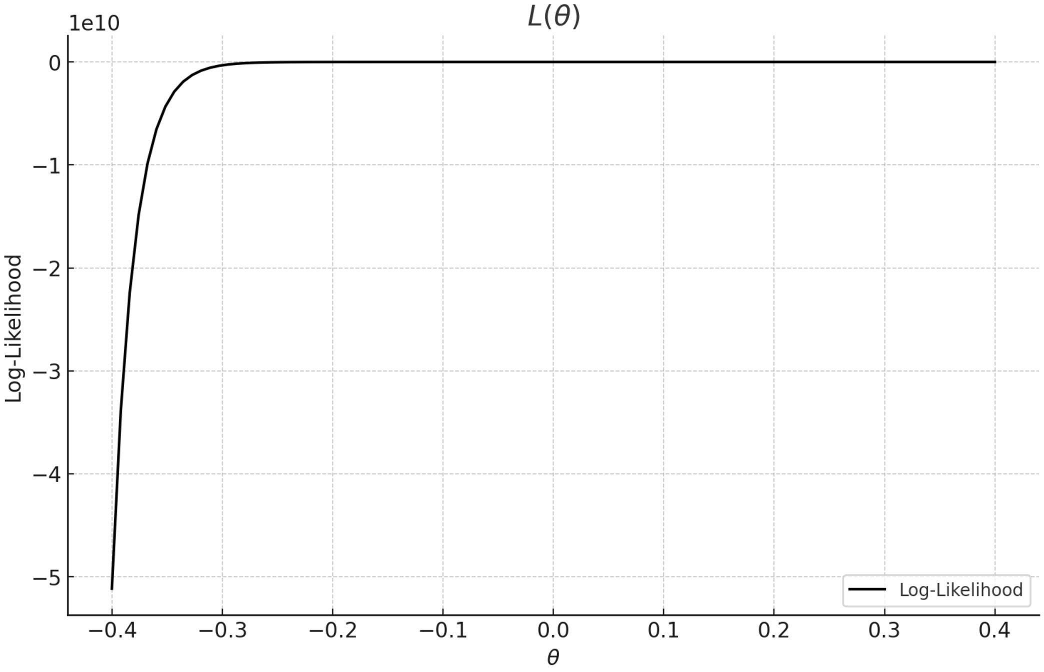
$$f_{X_i,Y_i}(x,y) = f_{Y_i|X_i}(y|x)f(x) = \exp(-\theta x)\exp(-\exp(-\theta x)y)(2\pi \cdot 100)^{-1/2}\exp\left(-\frac{1}{2}\left(\frac{x-40}{10}\right)^2\right).$$

The joint density is then $\prod_{i=1}^N f_{X_i,Y_i}(x,y)$, whence the log-likelihood function is given by

$$\begin{aligned} L(\theta) &= \sum_{i=1}^N -\theta x_i - \exp(-\theta x_i)y_i - \frac{1}{2}\ln(200\pi) - \frac{1}{2}\left(\frac{x_i - 40}{10}\right)^2 \\ &= -\sum_{i=1}^N \frac{(x_i-40)^2}{200} - \frac{n}{2}\ln(200\pi) - \theta\sum_{i=1}^N x_i - \sum_{i=1}^N \exp(-\theta x_i)y_i. \end{aligned}$$

The plot is shown below. For improved visualization, I have plotted θ in the range of -0.4 to 0.4, rather than -1 to 1.

Figure 3: Log likelihood function for $\theta \in [-0.4, 0.4]$.



□

(b) *Show that the log likelihood function is concave.*

Solution. The first order derivative of $L(\theta)$ with respect to θ is

$$\frac{\partial L(\theta)}{\partial \theta} = -\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \exp(-\theta x_i) x_i y_i.$$

Thus, the second order derivative of $L(\theta)$ with respect to θ is

$$\frac{\partial^2 L(\theta)}{\partial \theta} = -\sum_{i=1}^{N} \exp(-\theta x_i) x_i^2 y_i.$$

Since $\exp(-\theta x_i)$, x_i^2, and y_i are nonnegative, we have that $\frac{\partial^2 L(\theta)}{\partial \theta} \leq 0$ for all possible values of θ. Therefore $L(\theta)$ is concave. □

(c) *Find the maximum likelihood estimate of θ by starting at $\theta_0 = 0$ and using the Newton-Raphson algorithm for finding a maximum of a concave function*

$$\theta_{k+1} = \theta_k - \frac{\partial^2 L}{\partial \theta^2}(\theta_k)^{-1} \cdot \frac{\partial L}{\partial \theta}(\theta_k)$$

where $L(\theta)$ is the log likelihood function. Report the sequence of values θ_k and $L(\theta_k)$ for $k = 1, 2, 3, \ldots, 10$.

Solution. The results are reported below.

k	θ_k	$L(\theta_k)$
1	0.02293535	-150.7605715
2	0.045716507	-100.5635007
3	0.066394956	-84.78310755
4	0.081178714	-81.38430538
5	0.087204246	-81.08629081
6	0.087975548	-81.08259457
7	0.087986528	-81.08259385
8	0.08798653	-81.08259385
9	0.08798653	-81.08259385
10	0.08798653	-81.08259385

Table 1: θ and log likelihood values for each Newton-Raphson step.

□

---

6. [12.4, LNs] *This question is on (minimally) sufficient statistics.*

(a) *Find the sufficient statistic for $X_i \sim N(\theta, 1)$.*

Solution. Write the probability density function of $X = (X_1, X_2, \dots, X_n)$ as

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2)}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\theta\sum_{i=1}^n x_i - \sum_{i=1}^n \theta^2}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2} + \theta\sum_{i=1}^n x_i - \frac{n\theta^2}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right) \exp\left(\theta\sum_{i=1}^n x_i - \frac{n\theta^2}{2}\right). \end{aligned}$$

Let

$$h(x) \equiv \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right), \quad g(T(x), \theta) \equiv \exp\left(\theta T(x) - \frac{n\theta^2}{2}\right), \quad T(x) \equiv \sum_{i=1}^n x_i.$$

By the Factorization Theorem, $T(x) = \sum_{i=1}^n x_i$ is a sufficient statistic for $X_i \sim N(\theta, 1)$. □

(b) *Find the sufficient statistic for $X_i \sim N(\theta, \theta)$.*

Solution. Write the probability density function of $X = (X_1, X_2, \dots, X_n)$ as

$$\begin{aligned} f(x) &= \frac{1}{(2\pi\theta)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\theta}\right) \\ &= \frac{1}{(2\pi\theta)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2)}{2\theta}\right) \\ &= \frac{1}{(2\pi\theta)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\theta\sum_{i=1}^n x_i - \sum_{i=1}^n \theta^2}{2\theta}\right) \\ &= \frac{1}{(2\pi\theta)^{n/2}} \exp\left(\sum_{i=1}^n x_i - \frac{1}{2\theta}\sum_{i=1}^n x_i^2 - \frac{n\theta}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(\sum_{i=1}^n x_i\right) \exp\left(-\frac{1}{2\theta}\sum_{i=1}^n x_i^2 - \frac{n\theta}{2}\right). \end{aligned}$$

Let

$$h(x) \equiv \frac{1}{(2\pi)^{n/2}} \exp\left(\sum_{i=1}^n x_i\right), \quad g(T(x), \theta) \equiv \exp\left(-\frac{1}{2\theta}T(x) - \frac{n\theta}{2}\right), \quad T(x) \equiv \sum_{i=1}^n x_i^2.$$

By the Factorization Theorem, $T(x) = \sum_{i=1}^n x_i^2$ is a sufficient statistic for $X_i \sim N(\theta, \theta)$. □

(c) *Find the sufficient statistic for $X_i \sim N(\theta, \theta^2)$.*

Solution. Write the probability density function of $X = (X_1, X_2, \dots, X_n)$ as

$$\begin{aligned} f(x) &= \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\theta^2}\right) \\ &= \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2)}{2\theta^2}\right) \\ &= \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\theta\sum_{i=1}^n x_i - \sum_{i=1}^n \theta^2}{2\theta^2}\right) \\ &= \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(\frac{1}{\theta}\sum_{i=1}^n x_i - \frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 - \frac{n}{2}\right) \\ &= \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(-\frac{n}{2}\right) \exp\left(\begin{bmatrix}\theta^{-1} & (-2\theta^2)^{-1}\end{bmatrix} \begin{bmatrix}\sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2\end{bmatrix}\right). \end{aligned}$$

Let

$$h(x) \equiv \exp\left(-\frac{n}{2}\right), \quad g(T(x), \theta) \equiv \frac{1}{(2\pi\theta^2)^{n/2}} \exp\left(\begin{bmatrix}\theta^{-1} & (-2\theta^2)^{-1}\end{bmatrix} T(x)\right),$$

$$\text{and} \quad T(x) \equiv \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)'.$$

By the Factorization Theorem, $T(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)'$ is a sufficient statistic for $X_i \sim N(\theta, \theta^2)$. □

(d) *Find the sufficient statistic for $X_i \sim N(\mu, \sigma^2)$.*

Solution. Write the probability density function of $X = (X_1, X_2, \dots, X_n)$ as

$$\begin{aligned}
f(x) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\mu + \mu^2)}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\mu\sum_{i=1}^n x_i - \sum_{i=1}^n \theta^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\mu^2)^{n/2}} \exp\left(\frac{\theta}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n\mu^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\mu^2)^{n/2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(\begin{bmatrix}\mu/\sigma^2 & (-2\sigma^2)^{-1}\end{bmatrix}\begin{bmatrix}\sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2\end{bmatrix}\right).
\end{aligned}$$

Let $\theta \equiv (\mu, \sigma^2)'$ and

$$h(x) \equiv 1, \quad g(T(x), \theta) \equiv \frac{1}{(2\pi\mu^2)^{n/2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(\begin{bmatrix}\mu/\sigma^2 & (-2\sigma^2)^{-1}\end{bmatrix} T(x)\right),$$

$$\text{and} \quad T(x) \equiv \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)'.$$

By the Factorization Theorem, $T(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)'$ is a sufficient statistic for $X_i \sim N(\mu, \sigma^2)$. □

---

7. *Show that the likelihood ratio (LR), score (LM), and Wald tests are asymptotically equivalent for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.*

Solution. I shall prove the asymptotic equivalence of these three tests based on a likelihood estimator of θ. However, it is important to note that this equivalence extends to more general classes of estimators, such as generalized method of moments (GMM) estimators. For a comprehensive discussion on the asymptotic equivalence of these tests in broader contexts, I refer the reader to Newey and McFadden (1994).[2] For generality, I will assume θ is a vector.

Given the maximum likelihood estimator $\hat{\theta} \equiv \arg\max_{\theta\in\Theta} L(\theta)$ of θ_0 and a consistent estimator $\hat{\mathcal{I}}_\theta$ for the Fisher information at θ, the likelihood ratio (LR), score (LM), and Wald test statistics are defined as follows:

$$\begin{aligned}
LR &\equiv 2(L(\hat{\theta}) - L(\theta_0)), \\
LM &\equiv N^{-1}\left[\nabla_\theta L(\theta_0)\right]' \hat{\mathcal{I}}_{\theta_0}^{-1} \left[\nabla_\theta L(\theta_0)\right], \\
W &\equiv N(\hat{\theta} - \theta_0)'\hat{\mathcal{I}}_{\hat{\theta}}(\hat{\theta} - \theta_0).
\end{aligned}$$

[2] Newey, Whitney K., and Daniel McFadden. "Large sample estimation and hypothesis testing." Handbook of Econometrics, vol. 4 (1994): 2111-2245.

To demonstrate the asymptotic equivalence of the three tests, we need to show that they share the same asymptotic distribution. Specifically, I will show that W, LR, and LM all asymptotically follow a chi-squared distribution with k degrees of freedom, where $k = \dim(\theta)$.

For the LR test statistic, consider a second-order Taylor expansion of $L(\hat{\theta})$ around θ_0,

$$L(\hat{\theta}) = L(\theta_0) + (\hat{\theta} - \theta_0)' \nabla_\theta L(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' \nabla_{\theta\theta'} L(\theta_0)(\hat{\theta} - \theta_0) + o_p(1),$$

whence

$$2(L(\hat{\theta}) - L(\theta_0)) = 2\sqrt{n}(\hat{\theta} - \theta_0)' \frac{1}{\sqrt{n}} \nabla_\theta L(\theta_0) + \sqrt{n}(\hat{\theta} - \theta_0)' \frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1). \tag{6}$$

Now, observe that a first-order Taylor expansion of $\nabla_\theta L(\hat{\theta})$ around θ_0 gives

$$\nabla_\theta L(\hat{\theta}) = \nabla_\theta L(\theta_0) + \nabla_{\theta\theta'} L(\theta_0)(\hat{\theta} - \theta_0) + o_p(1).$$

Notice that by definition of $\hat{\theta}$ we must have $\nabla_\theta L(\hat{\theta}) = 0$, whence

$$\frac{1}{\sqrt{n}} \nabla_\theta L(\theta_0) = -\frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

Plugging this back into (6) gives

$$\begin{aligned} LR &= 2\sqrt{n}(\hat{\theta} - \theta_0)' \left(-\frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \right) \sqrt{n}(\hat{\theta} - \theta_0) + \sqrt{n}(\hat{\theta} - \theta_0)' \frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \\ &= \sqrt{n}(\hat{\theta} - \theta_0)' \left(-\frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \right) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1). \end{aligned}$$

Observe that by the Central Limit Theorem, $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} X \sim N(0, \mathcal{I}_{\theta_0}^{-1})$, and by the Law of Large Numbers $-\frac{1}{n} \nabla_{\theta\theta'} L(\theta_0) \xrightarrow{p} \mathcal{I}_{\theta_0}$. Therefore, by Slutsky's Theorem,

$$LR \xrightarrow{d} X' \mathcal{I}(\theta_0) X \sim \chi^2(k).$$

Alternatively, one could perform an expansion of $L(\theta_0)$ around $\hat{\theta}$,

$$\begin{aligned} L(\theta_0) &= L(\hat{\theta}) + (\theta_0 - \hat{\theta})' \nabla_\theta L(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' \nabla_{\theta\theta'} L(\hat{\theta})(\theta_0 - \hat{\theta}) + o_p(1) \\ &= L(\hat{\theta}) + \frac{1}{2}(\hat{\theta} - \theta_0)' \nabla_{\theta\theta'} L(\hat{\theta})(\hat{\theta} - \theta_0) + o_p(1). \end{aligned}$$

Then, plugging $L(\theta_0)$ into the expression for the LR statistic,

$$\begin{aligned} LR &= -(\hat{\theta} - \theta_0)' \nabla_{\theta\theta'} L(\hat{\theta})(\hat{\theta} - \theta_0) + o_p(1) \\ &= \sqrt{n}(\hat{\theta} - \theta_0)' \left(\frac{1}{n} \nabla_{\theta\theta'} L(\hat{\theta}) \right) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \xrightarrow{d} X' \mathcal{I}(\theta_0) X \sim \chi^2(k). \end{aligned}$$

For the score (LM) test statistic, write

$$LM \equiv \left[\frac{1}{\sqrt{N}}\nabla_\theta L(\theta_0)\right]' \hat{\mathcal{I}}_{\theta_0}^{-1}\left[\frac{1}{\sqrt{N}}\nabla_\theta L(\theta_0)\right].$$

Observe that by the Central Limit Theorem, $\frac{1}{\sqrt{N}}\nabla_\theta L(\theta_0) \xrightarrow{d} Z \sim N(0, \mathcal{I}(\theta_0))$, and by the Law of Large Numbers and the Continuous Mapping Theorem $\hat{\mathcal{I}}_{\theta_0}^{-1} \xrightarrow{p} \mathcal{I}_{\theta_0}^{-1}$. Therefore, by Slutsky's Theorem,

$$W \xrightarrow{d} Z'\mathcal{I}_{\theta_0}^{-1}Z \sim \chi^2(k).$$

For the Wald statistic, write

$$W = \sqrt{N}(\hat{\theta} - \theta_0)'\hat{\mathcal{I}}_{\hat{\theta}}\sqrt{N}(\hat{\theta} - \theta_0).$$

Observe that by the Central Limit Theorem, $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} X \sim N(0, \mathcal{I}_{\theta_0}^{-1})$, and by the Law of Large Numbers $\hat{\mathcal{I}}_{\hat{\theta}} \xrightarrow{p} \mathcal{I}_{\theta_0}$. Therefore, by Slutsky's Theorem,

$$W \xrightarrow{d} X'\mathcal{I}_{\theta_0}X \sim \chi^2(k).$$

□

---

8. [7.5, LNs] *Let $X_1, X_2, \ldots,$ be i.i.d. random variables with cdf F. Let*

$$\hat{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} I[X_i \leq x]$$

be the empirical distribution. Use the SLLN to show that $\hat{F}_N(x) \xrightarrow{a.s.} F(x)$, and use the CLT for i.i.d. random variables to find the limiting distribution of

$$\sqrt{N}\left(\begin{bmatrix}\hat{F}_N(x_1)\\ \vdots \\ \hat{F}_N(x_k)\end{bmatrix} - \begin{bmatrix}F_N(x_1)\\ \vdots \\ F_N(x_k)\end{bmatrix}\right).$$

Solution. Observe that since $X_1, X_2, \ldots, X_n$ are i.i.d., so are $I(X_1 \leq x), I(X_2 \leq x), \ldots, I(X_N \leq x)$. Therefore, by the Strong Law of Large Numbers, it follows that

$$\hat{F}_N = \frac{1}{N}\sum_{i=1}^{N} I(X_i \leq x) \xrightarrow{\text{a.s.}} \mathbb{E}[I(X_i \leq x)] = P(X_i \leq x) = F(x).$$

For the limiting distribution, observe that

$$
\begin{aligned}
\sqrt{N}\left(\begin{bmatrix}\hat{F}_N(x_1)\\ \vdots \\ \hat{F}_N(x_k)\end{bmatrix}-\begin{bmatrix}F_N(x_1)\\ \vdots \\ F_N(x_k)\end{bmatrix}\right) &= \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}I(X_i\leq x_1)\\ \vdots \\ I(X_i\leq x_k)\end{bmatrix}-\begin{bmatrix}F_N(x_1)\\ \vdots \\ F_N(x_k)\end{bmatrix}\right)\\
&=: \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_i-\mathbb{E}[\mathcal{I}_i]\right)\\
&=: \sqrt{N}\left(\bar{\mathcal{I}}-\mathbb{E}[\bar{\mathcal{I}}]\right)\xrightarrow[\text{CLT}]{d} N(0,\mathbb{E}[(\mathcal{I}_i-\mathbb{E}[\mathcal{I}_i])(\mathcal{I}_i-\mathbb{E}[\mathcal{I}_i])']).
\end{aligned}
$$

The (r,s)-th entry of the variance matrix $\Sigma\equiv\mathbb{E}[(\mathcal{I}_i-\mathbb{E}[\mathcal{I}_i])(\mathcal{I}_i-\mathbb{E}[\mathcal{I}_i])']$ can be neatly expressed as

$$
\begin{aligned}
\Sigma_{r,s} &= \mathbb{E}[(I(X_i\leq x_r)-F_n(x_r))(I(X_i\leq x_s)-F_n(x_s))]\\
&= \mathbb{E}[I(X_i\leq x_r)I(X_i\leq x_s)]-\mathbb{E}[I(X_i\leq x_r)]\mathbb{E}[I(X_i\leq x_s)]\\
&= \mathbb{E}[I(X_i\leq x_r)I(X_i\leq x_s)]-F_N(x_r)F_N(x_s)\\
&= \mathbb{E}[I(X_i\leq \min\{x_r,x_s\})]-F_N(x_r)F_N(x_s)\\
&= F_N(\min\{x_r,x_s\})-F_N(\min\{x_r,x_s\})F_N(\max\{x_r,x_s\})\\
&= F_N(\min\{x_r,x_s\})[1-F_N(\max\{x_r,x_s\})].
\end{aligned}
$$

Therefore,

$$
\Sigma=\begin{bmatrix}
F_N(x_1)[1-F_N(x_1)] & F_N(x_1)[1-F_N(x_2)] & \cdots & F_N(x_1)[1-F_N(x_k)]\\
F_N(x_1)[1-F_N(x_2)] & F_N(x_2)[1-F_N(x_2)] & \cdots & F_N(x_2)[1-F_N(x_k)]\\
\vdots & \vdots & \ddots & \vdots\\
F_N(x_1)[1-F_N(x_k)] & F_N(x_2)[1-F_N(x_k)] & \cdots & F_N(x_k)[1-F_N(x_k)]
\end{bmatrix}.
$$

□

---

9. [16.1, LNs] *Show that $tr(C'D)=vec(C)'vec(D)$ and that $tr(C'D)=tr(DC')$ for $p\times q$ matrices C and D.*

Solution. Let c_i and d_i denote the i-th columns of C and D, respectively. By partitioning $C=\begin{bmatrix}c_1 & c_2 & \cdots & c_q\end{bmatrix}$ and $D=\begin{bmatrix}d_1 & d_2 & \cdots & d_q\end{bmatrix}$ we have

$$
C'D=\begin{bmatrix}c_1'\\ c_2'\\ \vdots\\ c_q'\end{bmatrix}\begin{bmatrix}d_1 & d_2 & \cdots & d_q\end{bmatrix}=\begin{bmatrix}
c_1'd_1 & c_1'd_2 & \cdots & c_1'd_q\\
c_2'd_1 & c_2'd_2 & \cdots & c_2'd_q\\
\vdots & \vdots & \ddots & \vdots\\
c_q'd_1' & c_q'd_2 & \cdots & c_q'd_q
\end{bmatrix}
$$

Thus

$$
tr(C'D)=\sum_{j=1}^{q}c_j'd_j=\begin{bmatrix}c_1' & c_2' & \cdots & c_q'\end{bmatrix}\begin{bmatrix}d_1\\ d_2\\ \vdots\\ d_q\end{bmatrix}=vec(C)'vec(D).
$$

It is also easy to see that

$$tr(DC') = tr\left(\sum_{j=1}^{q} d_j c_j'\right) = \sum_{j=1}^{q} tr(d_j c_j') = \sum_{j=1}^{q}\sum_{i=1}^{p} d_{ij}c_{ij} = \sum_{i=1}^{q}\sum_{j=1}^{p} c_{ij}d_{ij} = \sum_{i=1}^{q} c_j' d_j = tr(C'D),$$

where c_{ij} and d_{ij} denote the (i,j) entries of C and D, respectively. □

---

10. [16.4, LNs] *Let n_{ij} and m_{ij} be the (i,j) elements of $N = X(X'X)^{-1}X'$ and $M = I - N$.*

(a) *Show that $0 \le n_{ii} \le 1$ and $0 \le m_{ii} \le 1$.*

Solution. Let e_i denote the i-th canonical vector. By spectral decomposition of N, we can write

$$n_{ii} = e_i'Ne_i = e_i'S\Lambda S'e_i = (S'e_i)'\Lambda S'e_i = v'\Lambda v,$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, S is orthogonal and $v \equiv S'e_i$. Here, we arrange eigenvalues in increasing order: $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. Observe that $v'v = e_i'SS'e_i = e_i'e_i = 1$, so by denoting v_i as the i-th element of v, we can write

$$\lambda_1 = \lambda_1 v'v = \sum_{i=1}^{n} \lambda_1 v_i^2 \le \underbrace{\sum_{i=1}^{n} \lambda_i v_i^2}_{v'\Lambda v} \le \sum_{i=1}^{n} \lambda_n v_i^2 = \lambda_n v'v = \lambda_n.$$

Since N is idempotent, all of its eigenvalues are either 0 or 1. Therefore $0 \le n_{ii} \le 1$. Since $m_{ii} = 1 - n_{ii}$, it also follows that $0 \le m_{ii} \le 1$.

An alternative one-line proof is

$$0 = \lambda_1 = \min_x \frac{x'Nx}{x'x} \le \frac{\overbrace{e_i'Ne_i}^{n_{ii}}}{\underbrace{e_i'e_i}_{1}} \le \max_x \frac{x'Nx}{x'x} = \lambda_n = 1.$$

Both inequalities follow from the Rayleigh quotient. □

(b) *Find all of the eigenvalues of N and M.*

Solution. As argued in **(a)**, since N is idempotent, all of its eigenvalues are either 0 or 1. The same holds for M, as it is also idempotent. Here I shall prove this result. Let A be any idempotent matrix. By eigendecomposition, $A = H\Lambda H^{-1}$, whence

$$AA = H\Lambda H^{-1}H\Lambda H^{-1} = H\Lambda\Lambda H^{-1} = H\Lambda^2 H^{-1}.$$

Therefore $\Lambda = \Lambda^2$, and hence $\lambda_i = \lambda_i^2$ for all $i = 1, \ldots, n$. Thus each λ_i must be equal to either zero or one. □

(c) *Interpret geometrically the vectors Ny and My.*

Solution. Let $Y = X\beta + e$. Observe that

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

$$\text{and} \quad MY = (I - X(X'X)^{-1}X')Y = Y - X(X'x)^{-1}X'Y = Y - X\hat{\beta}_{OLS} = \hat{e}.$$

Therefore

$$NY + MY = \hat{Y} + \hat{e}. \quad (= Y)$$

Observe that $NY + MY = (N + M)Y = IY = Y$, so $\hat{Y} = NY$ is the "part" of Y that is in the column space of X, while $\hat{e} = MY$ is the "part" of Y that is orthogonal to the column space of X. To visualize, examine Figure 4. This displays the case $n = 3$ and $k = 2$. Displayed are three vectors Y, X_1, and X_2, which are each elements of $\mathbb{R}^3$. The plane created by X_1 and X_2 is the column space of X. Regression-fitted values are linear combinations of X_1 and X_2 and so lie on this plane. The fitted value $\hat{Y}$ is the vector on this plane closest to Y. The residual $\hat{e} = Y - \hat{Y}$ is the difference between the two. The angle between the vectors $\hat{Y}$ and $\hat{e}$ is 90°, and therefore they are orthogonal as shown.

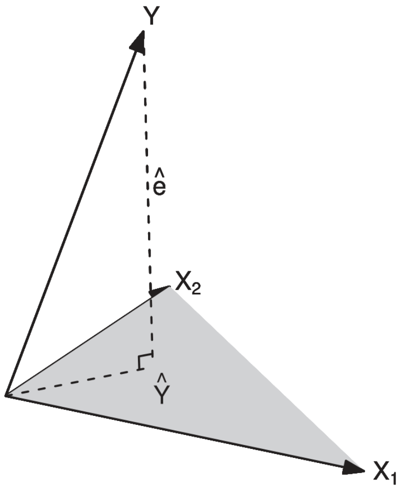


Figure 4: Projection of Y onto X_1 and X_2.

□

(d) *Show that the null space of N is the column space of M.*

Solution. Let $v \in \mathbb{R}^n$. If v is in the column space of M, then $Mx = v$ for some x. Hence $Nv = NMx = 0$. Thus v is in the null space of N. Conversely, if v is in the null space of N, then $Nv = 0$ and hence $-Nv = 0$, whence $v - Nv = v$. Thus $(I - N)v = v$. Therefore $Mv = v$, which means that v is in the column space of M. □