Problem 1. Inversion Theorems

Show the one-to-one relationship between the probability distribution function and the characteristic function. Express the probability distribution function as a function of the characteristic function.

Inversion Formula[1]

There is a one-to-one correspondence between cumulative distribution functions and characteristic functions, so it is possible to find one of these functions if we know the other. The formula in the definition of the characteristic function allows us to compute ϕ when we know the distribution function F (or density f). If, on the other hand, we know the characteristic function ϕ and want to find the corresponding distribution function, then one of the following inversion theorems can be used.

Theorem: If the characteristic function ϕ_X of a random variable X is integrable, then F_X is absolutely continuous, and therefore X has a probability density function. In the univariate case (i.e., when X is scalar-valued), the density function is given by:

$$f_X(x) = F'_X(x) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{-itx} \varphi_X(t)\, dt.$$

In the multivariate case, it is:

$$f_X(x) = \frac{1}{(2\pi)^n} \int_{\mathbf{R}^n} e^{-i(t \cdot x)} \varphi_X(t) \lambda(dt),$$

where $t \cdot x$ is the dot product.

[1] *Characteristic Function (Probability Theory)*. Wikipedia. Oct. 5, 2023. URL: `https://en.wikipedia.org/w/index.php?title=Characteristic_function_(probability_theory)&oldid=1178762588` (visited on 10/31/2023).

Problem 2.

Explain why the moment generation function does not exist when a certain moment is absent. Show the absence of moments in the context of

1. Cauchy distribution
2. t distribution
3. logistic distribution.

Definition 2.3.6:[a] Let X be a random variable with cumulative distribution function F_X. The moment generating function (mgf) of X (or F_X), denoted by $M_X(t)$, is defined as

$$M_X(t) = \mathbb{E}(e^{tX}),$$

provided that the expectation exists for t in some neighborhood of 0. In other words, there exists a positive value $h > 0$ such that, for all t in $-h < t < h$, $\mathbb{E}(e^{tX})$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the moment generating function (mgf) of X as:

$$M_X(t) = \int e^{tx} f_X(x)\, dx \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum e^{tx} p(X = x) \quad \text{if } X \text{ is discrete.}$$

.....

Theorem 2.3.7: If X has the moment generating function $M_X(t)$, then

$$\mathbb{E}(X^n) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0},$$

where we define $\frac{d^n}{dt^n} M_X(t)\big|_{t=0} = M_X^{(n)}(0)$. That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.

[a] George Casella and Roger L. Berger. *Statistical Inference*. 2nd ed. Australia ; Pacific Grove, CA: Thomson Learning, 2002. 660 pp., p. 62.

Thus, the absence of a certain moment implies that the moment generation function does not exist.

Proof of Theorem 2.3.7

Proof. [a] Assuming that we can differentiate under the integral sign, we have

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int e^{tx} f_X(x)\,dx \\ &= \int \frac{d}{dt}(e^{tx}) f_X(x)\,dx \\ &= \int (xe^{tx}) f_X(x)\,dx \\ &= \mathbb{E}(Xe^{tX}).\end{aligned}$$

Thus, for $n = 1$,

$$\frac{d^n}{dt^n}M_X(t)\Big|_{t=0} = \mathbb{E}(Xe^{tX})\Big|_{t=0} = \mathbb{E}(X).$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n}M_X(t)\Big|_{t=0} = \mathbb{E}(X^n e^{tX})\Big|_{t=0} = \mathbb{E}(X^n).$$

□

[a] Casella and Berger, see n. *a*, pp. 62–63.

. . . . . .

A (standard) Cauchy random variable has the following probability density function (PDF):

$$f(x) = \frac{1}{\pi\,(1+x^2)}, \quad x \in \mathbb{R}.$$

The mean of the Cauchy distribution does not exist:[2]

For any positive number M,

$$\int_0^M \frac{x}{1+x^2}\,dx = \frac{\log(1+x^2)}{2}\Big|_0^M = \frac{\log(1+M^2)}{2}.$$

Thus,

$$\begin{aligned}\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{|x|}{\pi}\frac{1}{1+x^2}\,dx = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,dx &= \lim_{M\to\infty}\frac{2}{\pi}\int_0^M \frac{x}{1+x^2}\,dx \\ &= \frac{1}{\pi}\lim_{M\to\infty}\log(1+M^2) = \infty.\end{aligned}$$

Since $\mathbb{E}[X] = \infty$, it follows that no moments of the Cauchy distribution exist, or in other words, all absolute moments equal ∞. In particular, the moment-generating function does not exist.[3]

[2] Ibid., p. 56.

[3] Ibid., p. 108.

. . . . . .

The probability density function (PDF) for Student's t distribution is given by:

$$f(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{p\pi}\Gamma\left(\frac{p}{2}\right)}\left(1+\frac{t^2}{p}\right)^{-\left(\frac{p+1}{2}\right)}$$

Where:

x is the random variable,

p is the degrees of freedom,

$\Gamma(\cdot)$ is the gamma function.

The absence of moments in the Student's t-distribution is related to the degrees of freedom (p). If there are p degrees of freedom, there are only $p-1$ moments.[4]

. . . . . .

[5] Let X be a continuous random variable which satisfies the logistic distribution:

$$X \sim \text{Logistic}(\mu, s)$$

for some $\mu \in \mathbb{R}$, $s \in \mathbb{R}^+$.

Then the moment generating function M_X of X is given by:

$$M_X(t) = \begin{cases} \exp(\mu t) B\left(1-st, 1+st\right) & |t| < \frac{1}{s} \\ \text{does not exist} & \text{o.w.} \end{cases}$$

where B denotes the beta function.

There is no absence of a moment in the logistic distribution.

[4] Casella and Berger, see n. *a*, p. 224.

[5] *Moment Generating Function of Logistic Distribution - ProofWiki.* URL: `https://proofwiki.org/wiki/Moment_Generating_Function_of_Logistic_Distribution` (visited on 10/31/2023).

Problem 3.

Show that Logistic regression has the property of 對稱性 and other models have no property of 對稱性.

As a set of independent binary regressions, to arrive at the multinomial logit model, one can imagine, for K possible outcomes, running $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome. If outcome K (the last outcome) is chosen as the pivot, the $K-1$ regression equations are:

$$\ln\left(\frac{\Pr(Y_i=k)}{\Pr(Y_i=K)}\right)=\boldsymbol{\beta}_k\cdot\mathbf{X}_i,\quad k<K.$$

This formulation is also known as the Additive Log Ratio transform, commonly used in compositional data analysis. In other applications, it's referred to as "relative risk".

If we exponentiate both sides and solve for the probabilities, we get:

$$\Pr(Y_i=k)=\Pr(Y_i=K)\cdot e^{\boldsymbol{\beta}_k\cdot\mathbf{X}_i},\quad k<K.$$

Using the fact that all K of the probabilities must sum to one, we find:

$$\Pr(Y_i=K)=1-\sum_{j=1}^{K-1}\Pr(Y_i=j)=1-\sum_{j=1}^{K-1}\Pr(Y_i=K)\cdot e^{\boldsymbol{\beta}_j\cdot\mathbf{X}_i}\Rightarrow\Pr(Y_i=K)=\frac{1}{1+\sum_{j=1}^{K-1}e^{\boldsymbol{\beta}_j\cdot\mathbf{X}_i}}.$$

We can use this to find the other probabilities:

$$\Pr(Y_i=k)=\frac{e^{\boldsymbol{\beta}_k\cdot\mathbf{X}_i}}{1+\sum_{j=1}^{K-1}e^{\boldsymbol{\beta}_j\cdot\mathbf{X}_i}},\quad k<K.$$

$$\Pr(Y_i=l)=\frac{\mathbf{I}(l=K)+\mathbf{I}(l\neq K)e^{\boldsymbol{\beta}_l\cdot\mathbf{X}_i}}{1+\sum_{j\neq K}e^{\boldsymbol{\beta}_j\cdot\mathbf{X}_i}},\quad l=1,2,\cdots,K.$$

$$\ln\left(\frac{\Pr(Y_i=l)}{\Pr(Y_i=K)}\right)=\boldsymbol{\beta}_l^{(K)}\cdot\mathbf{X}_i,\quad l\neq K.\tag{1}$$

$$\ln\left(\frac{\Pr(Y_i=l)}{\Pr(Y_i=J)}\right)=\boldsymbol{\beta}_l^{(J)}\cdot\mathbf{X}_i,\quad l\neq J.\tag{2}$$

(1)-(2)

$$\ln\left(\frac{\Pr(Y_i=l)}{\Pr(Y_i=K)}\right)-\ln\left(\frac{\Pr(Y_i=l)}{\Pr(Y_i=J)}\right)=\ln\left(\frac{\Pr(Y_i=J)}{\Pr(Y_i=K)}\right)=(\boldsymbol{\beta}_l^{(K)}-\boldsymbol{\beta}_l^{(J)})\cdot\mathbf{X}_i=\boldsymbol{\beta}_J^{(K)}\cdot\mathbf{X}_i,\quad K\neq J.$$

$$\boldsymbol{\beta}_l^{(J)} = \boldsymbol{\beta}_l^{(K)} - \boldsymbol{\beta}_J^{(K)}, \quad K \neq J.$$

$$\boldsymbol{\beta}_l^{(J)} = \boldsymbol{\beta}_l^{(K')} - \boldsymbol{\beta}_J^{(K')}, \quad K' \neq J.$$

On the other hand, for an arbitrary function

$$g\left(\frac{\Pr(Y_i = l)}{\Pr(Y_i = K)}\right) = \boldsymbol{\beta}_l^{(K)} \cdot \mathbf{X}_i,$$

the property holds only if $g(\cdot) = \ln(\cdot)$.

Therefore, we can conclude that the property holds if and only if in the logistic regression model.

Problem 4.

(b.1) Log-linear model:

$$P(Y=y)=C\left(\theta_1,\theta_2\right)\exp\left(Y^T\theta_1+W^T\theta_2\right),$$

where

- $W=\left(Y_1Y_2,Y_1Y_3,\cdots,Y_{m-1}Y_m,\cdots,Y_1Y_2\cdots Y_m\right)^T$,
- $\theta_1=\left(\theta_1^{(1)},\cdots,\theta_m^{(1)}\right)$
- $\theta_2=\left(\theta_{12}^{(2)},\cdots,\theta_{m-1m}^{(2)},\cdots,\theta_{12\cdots m}^{(m)}\right)$
- $C(\theta_1,\theta_2)$ is a function of θ_1 and θ_2 that normalizes the p.d.f. to integrate to one.

Transformation:

$$\left(\theta_1,\theta_2\right)\rightarrow\left(\mu,\theta_2\right),\mu=\left(\mu_1,\ldots,\mu_m\right)\triangleq\mu\left(\theta_1,\theta_2\right).$$

Model assumption:

$$\text{logit}\left(\mu_j\right)=X_j^T\beta.$$

The score equation for β under this parameterization takes the GEE form:

$$\left(\frac{\partial\mu}{\partial\beta}\right)^T[V(Y)]^{-1}(Y-\mu)=0,$$

where

$$\frac{\partial\mu}{\partial\beta}=\left(\frac{\partial\mu_1}{\partial\beta},\ldots,\frac{\partial\mu_m}{\partial\beta}\right)^T.$$

Remark:

The conditional odds ratios are not easily interpreted when the association among responses is itself a focus of the study.

Properties:

1. From

$$M_Y(t)=E\left[e^{t^TY}\right]=\sum_y C\left(\theta_1,\theta_2\right)\exp\left(y^T\left(t+\theta_1\right)+w^T\theta_2\right)=\frac{C\left(\theta_1,\theta_2\right)}{C\left(\theta_1+t,\theta_2\right)},$$

one has

- $$\mu=\left.\frac{\partial M_Y(t)}{\partial t}\right|_{t=0}=-\frac{\frac{\partial}{\partial\theta_1}C\left(\theta_1,\theta_2\right)}{C\left(\theta_1,\theta_2\right)},$$
- $$E\left[YY^T\right]=\left.\frac{\partial^2M_Y(t)}{\partial t\partial t^T}\right|_{t=0}=-\frac{\frac{\partial^2}{\partial\theta_1\partial\theta_1^T}C\left(\theta_1,\theta_2\right)}{C\left(\theta_1,\theta_2\right)}+2\mu\mu^T,$$
- $$V(Y)=-\frac{\frac{\partial^2}{\partial\theta_1\partial\theta_1^T}C\left(\theta_1,\theta_2\right)}{C\left(\theta_1,\theta_2\right)}+\mu\mu^T.$$

2. Let

$$l(\theta_1, \theta_2) = \ln P(Y = y) = \ln\{C(\theta_1, \theta_2) + (Y^T\theta_1 + W^T\theta_2)\}.$$

We can derive that

$$\frac{\partial l\left(\theta_1, \theta_2\right)}{\partial \theta_1} = \frac{\frac{\partial}{\partial \theta_1} C\left(\theta_1, \theta_2\right)}{C\left(\theta_1, \theta_2\right)} + Y = (Y - \mu),$$

and hence,

$$\frac{\partial l\left(\theta_1, \theta_2\right)}{\partial \beta} = \left(\frac{\partial \mu}{\partial \beta}\right)^T \frac{\partial \theta_1}{\partial \mu} \left(\frac{\partial l\left(\theta_1, \theta_2\right)}{\partial \theta_1}\right) = \left(\frac{\partial \mu}{\partial \beta}\right)^T \left(\frac{\partial \mu}{\partial \theta_1}\right)^{-1} (Y - \mu) = \left(\frac{\partial \mu}{\partial \beta}\right)^T [V(Y)]^{-1}(Y - \mu).$$

Validate the final expression:

$$\mu = -\frac{\frac{\partial}{\partial \theta_1} C\left(\theta_1, \theta_2\right)}{C\left(\theta_1, \theta_2\right)}.$$

$$\begin{aligned}
\frac{\partial \mu}{\partial \theta_1} &= \frac{\partial}{\partial \theta_1}\left\{-\frac{\frac{\partial}{\partial \theta_1} C\left(\theta_1, \theta_2\right)}{C\left(\theta_1, \theta_2\right)}\right\} \\
&= -\frac{C\left(\theta_1, \theta_2\right) \frac{\partial^2}{\partial \theta_1 \partial \theta_1^T} C\left(\theta_1, \theta_2\right) - \frac{\partial}{\partial \theta_1} C\left(\theta_1, \theta_2\right)\left[\frac{\partial}{\partial \theta_1} C\left(\theta_1, \theta_2\right)\right]^T}{C\left(\theta_1, \theta_2\right)^2} \\
&= -\frac{\frac{\partial^2}{\partial \theta_1 \partial \theta_1^T} C\left(\theta_1, \theta_2\right)}{C\left(\theta_1, \theta_2\right)} + \mu\mu^T \\
&= V(Y).
\end{aligned}$$

Problem 5. Show orthonormal

(b.2) The Bahadur representation:
Let

$$r_j = \frac{Y_j - \mu_j}{\sqrt{\mu_j\left(1-\mu_j\right)}}, j = 1, \cdots, m,$$

$$\rho_{jk} = E\left[r_j r_k\right], \rho_{jkl} = E\left[r_j r_k r_l\right], \cdots, \rho_{12\cdots m} = E\left[r_1 r_2 \cdots r_m\right]$$

$$P(Y=y) = \prod_{j=1}^{m} \mu_j^{y_j}\left(1-\mu_j\right)^{1-y_j}\left(1 + \sum_{j<k}\rho_{jk} r_j r_k + \sum_{j<k<l}\rho_{jkl} r_j r_k r_l + \cdots + \rho_{12\cdots m} r_1 r_2 \cdots r_m\right).$$

Remark.
1. the joint probability density function is expressed in terms of the marginal means, pairwise correlations, and higher moments of the standardized variables r_j.
2 . The correlations among binary responses are constrained in complicated ways by the marginal means.

.....

Let $P_{[1]}(Y=y) = \prod_{j=1}^{m} \mu_j^{y_j}\left(1-\mu_j\right)^{1-y_j}, g(y) = P(Y=y)/P_{[1]}(Y=y)$, and V be a vector space of real-valued functions f on Y_1 (2^m possible values of y).
Here, V is regarded as an inner-product space with

$$< f_1, f_2 > \triangleq E_{P_{[1]}}[f_1 f_2] = \sum_{y\in Y_1} f_1(y) f_2(y) P_{[1]}(y).$$

It follows easily that the set of functions $S = \{1, r_1, \ldots, r_m; r_1 r_2, \ldots, r_{m-1} r_m, \ldots, r_1 r_2 \cdots r_m\}$ on Y_1 is **orthonormal** and, thus, is a basis in V. Since $g(y)$ is a function on Y_1, there exists a unique representation as a linear combination of functions in S, namely,

$$g(y) = \sum_{f\in S} < g, f > f.$$

$$\because < g, f > = \sum_{y\in Y_1} g(y) f(y) P_{[1]}(y) = \sum_{y\in Y_1} f(y) P(Y=y) = E_P[f]\,\forall f, \text{ and}$$

$$E_P[1] = 1, E_P\left[r_j\right] = 0, E_P\left[r_j r_k\right] = \rho_{jk}, \cdots, \text{ and } E_P\left[r_1 \cdots r_m\right] = \rho_{12\cdots m}.$$

$$\therefore g(y) = \left(1 + \sum_{j<k}\rho_{jk} r_j r_k + \sum_{j<k<l}\rho_{jkl} r_j r_k r_l + \cdots + \rho_{12\cdots m} r_1 r_2 \cdots r_m\right).$$

Let S be defined as the set of functions such that:

$$S = \left\{r_1^{a_1} r_2^{a_2} \cdots r_m^{a_m} \mid a_i \in \{0,1\}, \text{ for } i = 1, \ldots, m\right\},$$

where s_a and s_b are elements of S:

$$s_a = r_1^{a_1} r_2^{a_2} \cdots r_m^{a_m} \in S,$$
$$s_b = r_1^{b_1} r_2^{b_2} \cdots r_m^{b_m} \in S.$$

To show S is orthonormal, is to show that:

$$< s_a, s_b > = 1, \text{ if } s_a = s_b,$$
$$< s_a, s_b > = 0, \text{ if } s_a \neq s_b.$$

$$
\begin{aligned}
r_j = \frac{Y_j - \mu_j}{\sqrt{\mu_j\left(1-\mu_j\right)}} &\overset{iid}{\sim} N(0,1) \\
\mathbb{E}[r_j] &= 0 \\
\mathbf{Var}[r_j] &= 1 \\
\mathbb{E}[r_j^2] &= 1
\end{aligned}
$$

(1)

Show that each is normalized:

$$
\begin{aligned}
< s_a, s_a > = E_{P_{[1]}}\left[s_a s_a\right] &= \sum_{y \in Y_1} s_a(y) s_a(y) P_{[1]}(y) \\
&= \mathbb{E}[r_1^{2a_1} r_2^{2a_2} \cdots r_m^{2a_m}] \\
&= \mathbb{E}[\prod_{i=1}^{m} [r_i^2]^{a_i}], \\
&= \prod_{i=1}^{m} \mathbb{E}[[r_i^2]^{a_i}], \quad a_i \in \{0,1\} \\
&= 1.
\end{aligned}
$$

(2)

Show orthogonal:

For $s_a \neq s_b$, $\exists i$ s.t, $a_i \neq b_i$, $\implies a_i + b_i = 1$, thus

$$
\begin{aligned}
< s_a, s_b > = E_{P_{[1]}}\left[s_a s_b\right] &= \sum_{y \in Y_1} s_a(y) s_b(y) P_{[1]}(y) \\
&= \mathbb{E}[r_1^{a_1+b_1} r_2^{a_2+b_2} \cdots r_i^{a_i+b_i} \cdots r_m^{a_m+b_m}] \\
&= \prod_{i \in \{i | a_i \neq b_i\}} \mathbb{E}[r_i] \prod_{i \in \{i | a_i = b_i\}} \mathbb{E}[[r_i^2]^{a_i}], \quad a_i \in \{0,1\} \\
&= 0.
\end{aligned}
$$

Paper[6] Summary

- **Research Problem**: Traditional cross-sectional analysis has its limitation. This paper addresses the challenge of analyzing longitudinal data on disease markers in AIDS, which involves problems of (left and right) censoring and (left) truncation.
- **Methodology**: The paper proposes a likelihood-based approach that models the joint distribution of the disease markers ($Z(t)$, such as T4 count and T4/T8 ratio), the time of infection, and the time to AIDS using parametric models. The paper also handles the censoring and truncation problems using standard survival analysis techniques and provides a method for predicting the time to AIDS given a series of disease marker measurements.
- **Data and Application**: The paper applies the proposed method to the Toronto AIDS cohort study, which consists of 159 cases who were infected with HIV or seroconverted during the 5.5 years' follow-up. The paper analyzes the longitudinal data on T4 counts and T4/T8 ratio as disease markers and compares three models with different assumptions about the infection time.
- **Results and Conclusion**: The paper finds that there is a significant association between the rate of decline of T4 counts or T4/T8 ratio and the time to AIDS, and that the scaled incubation time may be a more natural scale for viewing the progression of HIV infection. The paper also shows that the disease marker information improves the prediction of the time to AIDS. The paper concludes that the proposed method is useful for understanding the natural history of AIDS and developing treatment strategies.

[6] Yudi Pawitan and Steve Self. "Modeling Disease Marker Processes in AIDS". in: *Journal of the American Statistical Association* 88.423 (Sept. 1993), pp. 719–726.