Statistical Theory

Daniel Mao

Contents

Chapter 1

Descriptive Statistics

1.1 Population

DEFINITION (Population)**.** We define a **population**, denoted by $\mathcal{P}$, to be a finite set of elements.

DEFINITION (Unit)**.** We define a **unit**, denoted by u, to be an element of some population.

DEFINITION (Invariants and Equivariants)**.** Let $\mathcal{P}$ be a population. Let a be an attribute. We say that the attribute a is

(1) **location invariant** if

$$\forall b \in \mathbb{R}, a(\mathcal{P} + b) = a(\mathcal{P}).$$

(2) **location equivariant** if

$$\forall b \in \mathbb{R}, a(\mathcal{P} + b) = a(\mathcal{P}) + b.$$

(3) **scale invariant** if

$$\forall m \in \mathbb{R}^{\geq 0}, a(m \times \mathcal{P}) = a(\mathcal{P}).$$

(4) **scale equivariant** if

$$\forall m \in \mathbb{R}^{\geq 0}, a(m \times \mathcal{P}) = m \times a(\mathcal{P}).$$

(5) **location-scale invariant** if

$$\forall b \in \mathbb{R}, m \in \mathbb{R}^{\geq 0}, a(m \times \mathcal{P} + b) = a(\mathcal{P}).$$

(6) **location-scale equivariant** if

$$\forall b \in \mathbb{R}, m \in \mathbb{R}^{\geq 0}, a(m \times \mathcal{P} + b) = m \times a(\mathcal{P}) + b.$$

(7) **replication invariant** if

$$\forall k \in \mathbb{N}, a(\mathcal{P}^k) = a(\mathcal{P}).$$

(8) **replication equivariant** if

$$\forall k \in \mathbb{N}, a(\mathcal{P}^k) = k \times a(\mathcal{P}).$$

PROPOSITION 1.1.1. The population average is location-scale equivariant and replication invariant.

1.2 Order, Rank, and Quantiles

DEFINITION (Order Statistic)**.** Let $\mathcal{P} = \{y_1, ..., y_N\}$ be a population. We define the **order statistic**, denoted by $y_{(1)}, ..., y_{(N)}$, to be the ordered values of the variate values from $\mathcal{P}$ such that

$$y_{(1)} \leq ... \leq y_{(N)}.$$

DEFINITION (Rank Statistic)**.** Let $\mathcal{P} = \{y_1, ..., y_N\}$ be a population. We define the **rank statistic**, denoted by $r_1, ..., r_N$, to be the ranks of the variate values from $\mathcal{P}$.

So $\forall u \in \mathcal{P}, y_u = y_{(r_u)}$.

DEFINITION (Quantiles)**. Quantiles that Measure Center**:

- Median: $Q_y(1/2)$.

- Mid-hinge: $\frac{Q_y(1/4)+Q_y(3/4)}{2}$.
- Mid-range: $\frac{Q_y(1/N)+Q_y(1)}{2}$.
- Trimean: $\frac{Q_y(1/4)+Q_y(1/2)+Q_y(3/4)}{2}$.

Quantiles that Measure Spread:

- Range: $Q_y(1) - Q_y(1/N)$.
- Inter-Quartile Range: $Q_y(3/4) - Q_y(1/4)$.

1.3 Statistical Distance

DEFINITION (Kullback–Leibler (KL) Divergence)**.**

- Discrete Case:

  Let $\mathcal{X}$ be a probability space. Let P and Q be discrete probability distributions over $\mathcal{X}$. We define the **Kullback–Leibler divergence** from Q to P, denoted by $D_{KL}(P||Q)$, to be the sum given by

$$D_{KL}(P||Q) := \underset{X\sim P}{\mathbb{E}}\left[\log\frac{P(X)}{Q(X)}\right] = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}.$$

- Continuous Case:

  Let P and Q be continuous probability distributions. We define the **Kullback–Leibler divergence** from Q to P, denoted by $D_{KL}(P||Q)$, to be the integral given by

$$D_{KL}(P||Q) := \underset{X\sim P}{\mathbb{E}}\left[\log\frac{P(X)}{Q(X)}\right] = \int_{\mathbb{R}} P(x)\log\frac{P(x)}{Q(x)}dx.$$

Chapter 2

Sampling Theory

2.1 Sampling Distribution

PROPOSITION 2.1.1. Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

Proof. Part 1: show that

$$\sum_{i=1}^{n}(X_i - \mu)^2 = (n-1)s^2 + n(\bar{X} - \mu)^2.$$

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}\left[(X_i - \bar{X}) + (\bar{X} - \mu)\right]^2 \\
&= \sum_{i=1}^{n}\left[(X_i - \bar{X})^2 + (\bar{X} - \mu)^2 + 2(X_i - \bar{X})(\bar{X} - \mu)\right] \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + \sum_{i=1}^{n}2(X_i - \bar{X})(\bar{X} - \mu) \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}X_i - 2(\bar{X} - \mu)\sum_{i=1}^{n}\bar{X} \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)n\bar{X} - 2(\bar{X} - \mu)n\bar{X}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \\
&= (n-1)\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \\
&= (n-1)s^2 + n(\bar{X} - \mu)^2
\end{aligned}
$$

That is,

$$\sum_{i=1}^{n}(X_i - \mu)^2 = (n-1)s^2 + n(\bar{X} - \mu)^2,$$

as desired.

Part 2

Since

$$\sum_{i=1}^{n}(X_i - \mu)^2 = (n-1)s^2 + n(\bar{X} - \mu)^2,$$

we get

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

Since $X_i \sim N(\mu, \sigma^2)$, we get

$$\frac{X_i - \mu}{\sigma} \sim N(0,1).$$

So

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$$

and hence

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).$$

Since $\bar{X} \sim N(\mu, \sigma^2/n)$, we get

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

So

$$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1).$$

Notice $\frac{(n-1)s^2}{\sigma^2}$ and $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ are independent, it follows that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

■

2.2 Inclusion Probability

Simple Random Sampling Without Replacement

- inclusion probability: $\pi_u = \frac{n}{N}$.
  Derivation:

$$\begin{aligned}\pi_u &= \sum_{S \in \mathcal{P}_S} P(S) I(u \in S)\\ &= \frac{1}{\binom{N}{n}} \binom{N-1}{n-1}\\ &= \frac{1}{\frac{N!}{n!(N-n)!}} \frac{(N-1)!}{(n-1)!(N-n)!}\\ &= \frac{n}{N}.\end{aligned}$$

- joint inclusion probability: $\pi_{uv} = \frac{n(n-1)}{N(N-1)}$
  Derivation:

$$\begin{aligned}\pi_{uv} &= \sum_{S \in \mathcal{P}_S} P(S) I(u, v \in S)\\ &= \frac{1}{\binom{N}{n}} \binom{N-2}{n-2}\\ &= \frac{1}{\frac{N!}{n!(N-n)!}} \frac{(N-2)!}{(n-2)!(N-n)!}\\ &= \frac{n(n-1)}{N(N-1)}.\end{aligned}$$

Chapter 3

Estimation Theory

3.1 Maximum Likelihood Estimation

DEFINITION (Likelihood Function)**.** Let Ω be a parameter space. Let θ be an unknown parameter in Ω. Let $\boldsymbol{Y}$ be a random variable and $\boldsymbol{y}$ be an observed sample. Let f denote the probability function of Y. We define the **likelihood function** for θ, denoted by $\mathcal{L}$, to be a function from Ω to $\mathbb{R}$ given by

$$\mathcal{L}(\theta; \boldsymbol{y}) := f(\boldsymbol{y}; \theta).$$

DEFINITION (Maximum Likelihood Estimate)**.** Let Ω be a parameter space. Let θ be an unknown parameter in Ω. Let $\boldsymbol{Y}$ be a random variable and $\boldsymbol{y}$ be an observed sample. Let $\mathcal{L}$ be the likelihood function for θ. We define the **maximum likelihood estimate** for θ, denoted by $\hat{\theta}$, to be a parameter given by

$$\hat{\theta} := \underset{\theta \in \Omega}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{y}).$$

DEFINITION (Relative Likelihood Function)**.** Let Ω be a parameter space. Let θ be an unknown parameter in Ω. Let $\boldsymbol{Y}$ be a random variable and $\boldsymbol{y}$ be an observed sample. Let $\mathcal{L}$ be the likelihood function for θ. We define the **relative likelihood**

function for θ, denoted by $\mathcal{R}$, to be a function from Ω to $\mathbb{R}$ given by

$$\mathcal{R}(\theta; \boldsymbol{y}) := \frac{\mathcal{L}(\theta; \boldsymbol{y})}{\mathcal{L}(\hat{\theta}; \boldsymbol{y})}.$$

DEFINITION (Log Likelihood Function)**.** Let Ω be a parameter space. Let θ be an unknown parameter in Ω. Let $\boldsymbol{Y}$ be a random variable and $\boldsymbol{y}$ be an observed sample. Let $\mathcal{L}$ be the likelihood function for θ. We define the **log likelihood function** for θ, denoted by ℓ, to be a function from Ω to $\mathbb{R}$ given by

$$\ell(\theta; \boldsymbol{y}) := \ln \mathcal{L}(\theta; \boldsymbol{y}).$$

3.2 Examples

PROPOSITION 3.2.1 (MLE of Binomial Distribution)**.** Let $n \in \mathbb{N}$. Let θ be an unknown parameter. Let y be an observed value from a Binomial(n, θ) distribution. Then the likelihood function for θ is

$$\mathcal{L}(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

PROPOSITION 3.2.2 (MLE of Poisson Distribution)**.** Let θ be an unknown parameter. Let $y_1, ..., y_n$ be an observed random sample from a Poisson(θ) distribution. Then the likelihood function for θ is

$$\mathcal{L}(\theta) = \theta^{n\bar{y}} e^{-n\theta} \text{ for } \theta \in \mathbb{R}_+,$$

and the maximum likelihood estimation of θ is

$$\hat{\theta} = \bar{y}.$$

PROPOSITION 3.2.3 (MLE of Exponential Distribution)**.** Let θ be an unknown parameter. Let $y_1, ..., y_n$ be an observed random sample from a Exponential$(\frac{1}{\theta})$ distribution. Then the likelihood function for θ is

$$\mathcal{L}(\theta) = \theta^{-n} e^{-n\bar{y}/\theta},$$

and the maximum likelihood estimate $\hat{\theta}$ of θ is

$$\hat{\theta} = \bar{y}.$$

3.3 Confidence Intervals

DEFINITION. Let X be a random variable from a probability distribution with statistical parameter $\theta \in \Omega$. We define a **confidence interval** for the parameter θ with confidence level γ to be an interval with random endpoints $u(X)$ and $v(X)$ such that

$$\forall \theta \in \Omega, \quad \mathbb{P}_\theta(u(X) < \theta < v(X)) = \gamma.$$

Chapter 4

Hypothesis Testing

4.1 p-values

DEFINITION. Let H_0 be a null hypothesis. Let D be a discrepancy measure. Let $\boldsymbol{y}$ be an observed value of the data. We define the **p-value** of the test of the hypothesis H_0, using the discrepancy measure D, to be the probability given by

$$p\text{-value} := \mathbb{P}(D(\boldsymbol{Y}) \geq D(\boldsymbol{y}); H_0).$$

Chapter 5

Asymptotic Theory

In statistics: asymptotic theory, or large sample theory, is a framework for assessing properties of estimators and statistical tests. Within this framework, it is often assumed that the sample size n *may grow indefinitely; the properties of estimators and tests are then evaluated under the limit of* $n \to \infty$. *In practice, a limit evaluation is considered to be approximately valid for large finite sample sizes too. (Wikipedia, 2021-07-11)*

In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed. The LLN is important because it guarantees stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game. It is important to remember that the law only applies (as the name indicates) when a large number of observations is considered. There is no principle that a small number of observations will coincide with the expected value or that a streak of one value will immediately be "balanced" by the others (see the gambler's fallacy). (Wikipedia, 2021-07-11)

5.1 Law of Large Numbers

THEOREM 5.1 (The Weak Law of Large Numbers (Khinchin's Law))**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables. Suppose that they all have finite expected value and standard variation. Let μ denote their common expected value. Then the sample average converges weakly (in probability) to the common expected value:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{weakly}}{\longrightarrow} \mu.$$

Proof. Note that

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \text{ and}$$

$$\operatorname{var}[\overline{X}_n] = \operatorname{var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

So by the Chebyshev's inequality, we get

$$\Pr\left[\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq \frac{1}{n}\varepsilon\right] \leq \frac{n^2\sigma^2}{\varepsilon^2}.$$

■

THEOREM 5.2 (The Strong Law of Large Nubmers (Kolmogorov's Law))**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables. Suppose that they all have finite expected value and standard variation. Let μ denote their common expected value. Then the sample average converges strongly (almost surely) to the common expected value:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{strongly}}{\longrightarrow} \mu.$$

5.2 Central Limit Theorem

THEOREM 5.3 (Central Limit Theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed Lebesgue integrable random variables. Let μ denote their common expectation and σ denote their common standard deviation. Then we have

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. When MGF's exist:

$$\begin{aligned}
M_n(t) &= \mathbb{E}\left[\exp\left\{t\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right\}\right], \quad Y_i := \frac{X_i - \mu}{\sigma} \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left\{\frac{tY_i}{\sqrt{n}}\right\}\right] \\
&= \left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[M_Y(0) + M_Y'(0)\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}M_Y''(0)\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\
&= \left[1 + \mathbb{E}[Y_i]\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}\mathbb{E}[Y_i^2]\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\
&= \left[1 + \mathbb{E}[Y_i]\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}\big(\operatorname{var}[Y_i] + (\mathbb{E}[Y_i])^2\big)\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\
&= \left[1 + 0\cdot\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}(1+0)\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\
&= \left[1 + \frac{t^2/2}{n} + o\left(\frac{t^2}{n}\right)\right]^n \\
&\to \exp\left\{\frac{t^2}{2}\right\}.
\end{aligned}$$

■

Chapter 6

Power Transformation

DEFINITION (Box-Cox Transformation)**.** We define the one-parameter **Box-Cox transformation** of y, denoted by $T_\lambda(y)$, to be a function given by

$$T_\lambda(y) := \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \textbf{if } \lambda \neq 0 \\ \ln y, & \textbf{if } \lambda = 0. \end{cases}$$

PROPOSITION 6.0.1. The Box-Cox transformation is monotonic increasing.