Parameter Estimation

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics
University of Minnesota

August 30, 2020

Table of Contents

Table of Contents

Probability Distribution Reminders

A random variable X has a cumulative distribution function (CDF) denoted by $F(x) = P(X \leq x)$ that describes the probabilistic nature of the random variable X.

$F(\cdot)$ has an associated probability mass function (PMF) or probability density function (PDF) denoted by $f(x)$.

- PMF: $f(x) = P(X = x)$ for discrete variables
- PDF: $\int_a^b f(x) = P(a < X < b)$ for continuous variables

The functions $F(\cdot)$ and $f(\cdot)$ are typically assumed to depend on a finite number of parameters, where a parameter $\theta = t(F)$ is some function of the probability distribution.

Inferences and Statistics

Given a sample of n independent and identically distributed (iid) observations from some distribution F, inferential statistical analyses are concerned with inferring things about the population from which the sample was collected.

To form inferences, researchers often make assumptions about the form of F, e.g., F is a normal distribution, and then use the sample of data to form educated guesses about the population parameters.

Given a sample of data $\mathbf{x} = (x_1, \ldots, x_n)^\top$, a statistic $T = s(\mathbf{x})$ is some function of the sample of data. Not all statistics are created equal...

- Some are useful for estimating parameters or testing hypotheses

Table of Contents

Statistics are Random Variables

Assume that $x_i \overset{\text{iid}}{\sim} F$ for $i = 1, \ldots, n$, where the notation $\overset{\text{iid}}{\sim}$ denotes that the x_i are iid observations from the distribution F.

- $\mathbf{x} = (x_1, \ldots, x_n)^\top$ denotes the sample of data as an $n \times 1$ vector

Each x_i is assumed to be an independent realization of a random variable $X \sim F$, so any valid statistic $T = s(\mathbf{x})$ will be a random variable with a probability distribution.

- By "valid" I mean that T must depend on the x_i values

The sampling distribution of a statistic $T = s(\mathbf{x})$ refers to the probability distribution of T.

Sampling Distribution Properties

Suppose that we collect R independent realizations of the vector $\mathbf{x}$, and let $T_r = s(\mathbf{x}_r)$ denote the r-th realization of the statistic. The sampling distribution is the probability distribution of $\{T_r\}_{r=1}^{R}$ as the number of independent realizations $R \to \infty$.

The sampling distribution depends on the distribution of data.

- if $x_i \overset{\text{iid}}{\sim} F$ and $y_i \overset{\text{iid}}{\sim} G$, then the statistics $T = s(\mathbf{x})$ and $U = s(\mathbf{y})$ will have different sampling distributions if F and G are different.

Sometimes the sampling distribution will be known as $n \to \infty$.

- CLT or asymptotic normality of MLEs
- Question of interest is: how large does n need to be?

Table of Contents

Definition of Estimates and Estimators

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$, an estimate of a parameter $\theta = t(F)$ is some function of the sample $\hat{\theta} = g(\mathbf{x})$ that is meant to approximate θ.

An estimator refers to the function $g(\cdot)$ that is applied to the sample to obtain the estimate $\hat{\theta}$.

Standard notation in statistics, where a "hat" (i.e., $\hat{}$) is placed on top of the parameter to denote that $\hat{\theta}$ is an estimate of θ.

- $\hat{\theta}$ should be read as "theta hat"
- should interpret $\hat{\theta}$ as some estimate of the parameter θ

Examples of Estimates and Estimators

Example. Suppose that we have a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$, which denotes any generic distribution, and the population mean $\mu = E(X)$ is the parameter of interest. The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ provides an estimate of the parameter μ, so we could also write it as $\bar{x} = \hat{\mu}$.

Example. Similarly, suppose that we have a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ and the population variance $\sigma^2 = E[(X - \mu)^2]$ is the parameter of interest. The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ provides an estimate of the parameter σ^2, so we could also write it as $s^2 = \hat{\sigma}^2$. Another reasonable estimate would be $\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Table of Contents

Overview

Like statistics, not all estimators are created equal. Some estimators produce "better" estimates of the intended population parameters.

There are several ways to talk about the "quality" of an estimator:

- its expected value (bias)
- its uncertainty (variance)
- both its bias and variance (MSE)
- its asymptotic properties (consistency)

MSE is typically the preferred way to measure an estimator's quality.

Bias of an Estimator

The bias of an estimator refers to the difference between the expected value of the estimate $\hat{\theta} = g(\mathbf{x})$ and the parameter $\theta = t(F)$, i.e.,

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

where the expectation is calculated with respect to F.

- An estimator is "unbiased" if $\text{Bias}(\hat{\theta}) = 0$

Despite the negative connotations of the word "bias", it is important to note that biased estimators can be a good thing (see Helwig, 2017).

- Ridge regression (Hoerl and Kennard, 1970)
- Least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996)
- Elastic Net regression (Zou and Hastie, 2005)

Bias Example 1: The Mean

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ and F has mean $\mu = E(X)$, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased estimate of the population mean μ.

To prove that $\bar{x}$ is an unbiased estimator, we can use the expectation rules from Introduction to Random Variables chapter. Specifically, note that $E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$.

Bias Example 2: The Variance (part 1)

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ and F has mean $\mu = E(X)$ and variance $\sigma^2 = E[(X-\mu)^2]$, the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of σ^2.

To prove that s^2 is unbiased, first note that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

which implies that $E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right]$.

Now note that $\sigma^2 = E(x_i^2) - \mu^2$, which implies that $E(x_i^2) = \sigma^2 + \mu^2$.

Bias Example 2: The Variance (part 2)

Also, note that we can write

$$\bar{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} x_i x_j\right)$$

and applying the expectation operator gives

$$\begin{aligned} E(\bar{x}^2) &= \frac{1}{n^2}\sum_{i=1}^{n} E(x_i^2) + \frac{2}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E(x_i)E(x_j) \\ &= \frac{1}{n}(\sigma^2 + \mu^2) + \frac{n-1}{n}\mu^2 \end{aligned}$$

given that $E(x_i x_j) = E(x_i)E(x_j)$ for all $i \neq j$ because x_i and x_j are independent, and $\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mu^2 = \frac{n(n-1)}{2}\mu^2$.

Bias Example 2: The Variance (part 3)

Putting all of the pieces together gives

$$\begin{aligned} E(s^2) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - (\sigma^2+\mu^2) - (n-1)\mu^2\right) \\ &= \sigma^2 \end{aligned}$$

which completes the proof that $E(s^2) = \sigma^2$.

This result can be used to show that $\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is biased:

- $E\left(\tilde{s}^2\right) = E\left(\frac{n-1}{n}s^2\right) = \frac{n-1}{n}E\left(s^2\right) = \frac{n-1}{n}\sigma^2$
- $\frac{n-1}{n} < 1$ for any finite n, so $\tilde{s}^2$ has a downward bias

Variance of a Estimator

The variance of an estimator refers to second central moment of the estimator's probability distribution, i.e.,

$$\text{Var}(\hat{\theta}) = E\left(\left(\hat{\theta} - E(\hat{\theta})\right)^2\right)$$

where both expectations are calculated with respect to F.

The standard error of an estimator is the square root of the variance of the estimator, i.e., $\text{SE}(\hat{\theta}) = \text{Var}(\hat{\theta})^{1/2}$.

We would like an estimator that is both reliable (low variance) and valid (low bias), but there is a trade-off between these two concepts.

Variance of the Sample Mean

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ and F has mean $\mu = E(X)$ and variance $\sigma^2 = E[(X - \mu)^2]$, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ has a variance of $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$.

To prove that this is the variance of $\bar{x}$, we can use the variance rules from the Introduction to Random Variables chapter, i.e.,

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(x_i) = \frac{\sigma^2}{n}$$

given that the x_i are independent and $\text{Var}(x_i) = \sigma^2$ for all $i = 1, \ldots, n$.

Variance of the Sample Variance

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ and F has mean $\mu = E(X)$ and variance $\sigma^2 = E[(X-\mu)^2]$, the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ has a variance of

$$\text{Var}(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

where $\mu_4 = E[(X-\mu)^4]$ is the fourth central moment of X.

- The proof of this is too tedious to display on the slides
- Bonus points for anyone who can prove this formula

The above result can be used to show that

- $\text{Var}(\tilde{s}^2) = \text{Var}\left(\frac{n-1}{n}s^2\right) = \frac{(n-1)^2}{n^3}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$

Mean Squared Error of an Estimator

The mean squared error (MSE) of an estimator refers to the expected squared difference between the parameter $\theta = t(F)$ and the estimate $\hat{\theta} = g(\mathbf{x})$, i.e.,

$$\text{MSE}(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

where the expectation is calculated with respect to F.

Although not obvious from its definition, MSE can be decomposed as

$$\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

where the first term is squared bias and the second term is variance.

MSE = Bias^2 + Variance

To prove this relationship holds for any estimator, first note that $(\hat{\theta} - \theta)^2 = \hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2$, and applying the expectation operator gives

$$E\left((\hat{\theta} - \theta)^2\right) = E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2$$

given that the parameter θ is assumed to be an unknown constant.

Next, note that we can write the squared bias and variance as

$$\text{Bias}(\hat{\theta})^2 = \left(E(\hat{\theta}) - \theta\right)^2 = E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2$$

$$\text{Var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$$

and adding these two terms together gives

$$\begin{aligned}\text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}) &= E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2 + E(\hat{\theta}^2) - E(\hat{\theta})^2 \\ &= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2\end{aligned}$$

which is the form of the MSE given on the previous slide.

Consistency of an Estimator

Given a sample of data $x_1, \ldots, x_n$ with $x_i \overset{\text{iid}}{\sim} F$, an estimator $\hat{\theta} = g(\mathbf{x})$ of a parameter $\theta = t(F)$ is said to be consistent if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

The notation $\xrightarrow{p}$ should be read as "converges in probability to", which means that the probability that $\hat{\theta} \neq \theta$ goes to zero as n gets large.

Note that any reasonable estimator should be consistent. Otherwise, collecting more data will not result in better estimates.

All of the estimators that we've discussed (i.e., $\bar{x}$, s^2 and $\tilde{s}^2$) are consistent estimators.

Efficiency of an Estimator

Given a sample of data $x_1, \ldots, x_n$ with $x_i \overset{\text{iid}}{\sim} F$, an estimator $\hat{\theta} = g(\mathbf{x})$ of a parameter $\theta = t(F)$ is said to be efficient if it is the best possible estimator for θ using some loss function.

The chosen loss function is often MSE, so the most efficient estimator is the one with the smallest MSE compared to all other estimators of θ.

If you have two estimators $\hat{\theta}_1 = g_1(\mathbf{x})$ and $\hat{\theta}_2 = g_2(\mathbf{x})$, we would say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $\text{MSE}(\hat{\theta}_1) < \text{MSE}(\hat{\theta}_2)$.

- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased, the most efficient estimator is the one with the smallest variance

Table of Contents

Least Squares Estimation

A simple least squares estimate of a parameter $\theta = t(F)$ is the estimate $\hat{\theta} = g(\mathbf{x})$ that minimizes a least squares loss function of the form

$$\sum_{i=1}^{n} (h(x_i) - \theta)^2$$

where $h(\cdot)$ is some user-specified function (typically $h(x) = x$).

Least squares estimation methods can work well for mean parameters and regression coefficients, but will not work well for all parameters.

- Variance parameters are best estimated using other approahces

Least Squares Estimation Example

Given a sample of data $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$, suppose that we want to find the least squares estimate of $\mu = E(X)$.

The least squares loss function is

$$\text{LS}(\mu|\mathbf{x}) = \sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is the observed data vector.

Taking the derivative of the function with respect to μ gives

$$\frac{d\text{LS}(\mu|\mathbf{x})}{d\mu} = -2 \sum_{i=1}^{n} x_i + 2n\mu$$

and setting the derivative to 0 and solving for μ gives $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

- The sample mean $\bar{x}$ is the least squares estimate of μ

Method of Moments Estimation

Assume that $X \sim F$ where the probability distribution F depends on parameters $\theta_1, \ldots, \theta_p$.

Also, suppose that the first p moments of X can be written as

$$\mu_j = E(X^j) = m_j(\theta_1, \ldots, \theta_p)$$

where $m_j(\cdot)$ is some known function for $j = 1, \ldots, p$.

Given data $x_i \overset{\text{iid}}{\sim} F$ for $i = 1, \ldots, n$, the method of moments estimates of the parameters are the values $\hat{\theta}_1, \ldots, \hat{\theta}_p$ that solve the equations

$$\hat{\mu}_j = m_j(\hat{\theta}_1, \ldots, \hat{\theta}_p)$$

where $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j$ is the j-th sample moment for $j = 1, \ldots, p$.

Method of Moments: Normal Distribution

Suppose that $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. The first two moments of the normal distribution are $\mu_1 = \mu$ and $\mu_2 = \mu^2 + \sigma^2$.

The first two sample moments are $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$ and $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n x_i^2 = \bar{x}^2 + \tilde{s}^2$, where $\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

Thus, the method of moments estimates of μ and σ^2 are given by $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \tilde{s}^2$.

Method of Moments: Uniform Distribution

Suppose that $x_i \overset{\text{iid}}{\sim} U[a, b]$ for $i = 1, \ldots, n$. The first two moments of the uniform distribution are $\mu_1 = \frac{1}{2}(a + b)$ and $\mu_2 = \frac{1}{3}(a^2 + ab + b^2)$.

Solving the first equation gives $b = 2\mu_1 - a$ and plugging this into the second equation gives $\mu_2 = \frac{1}{3}\left(a^2 - 2a\mu_1 + 4\mu_1^2\right)$, which is a simple quadratic function of a.

Applying the quadratic formula (see here) gives $a = \mu_1 - \sqrt{3}\sqrt{\mu_2 - \mu_1^2}$, and plugging this into $b = 2\mu_1 - a$ produces $b = \mu_1 + \sqrt{3}\sqrt{\mu_2 - \mu_1^2}$.

Using $\hat{\mu}_1$ and $\hat{\mu}_2$ in these equations gives the methods of moments estimates of a and b.

Likelihood Function and Log-Likelihood Function

Suppose that $x_i \overset{\text{iid}}{\sim} F$ for $i = 1, \ldots, n$ where the distribution F depends on the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$.

The likelihood function has the form

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

where $f(x_i|\boldsymbol{\theta})$ is the probability mass function (PMF) or probability density function (PDF) corresponding to the distribution function F.

The log-likelihood function is the logarithm of the likelihood function:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \log\left(L(\boldsymbol{\theta}|\mathbf{x})\right) = \sum_{i=1}^{n} \log\left(f(x_i|\boldsymbol{\theta})\right)$$

where $\log(\cdot) = \ln(\cdot)$ is the natural logarithm function.

Maximum Likelihood Estimation

Suppose that $x_i \overset{\text{iid}}{\sim} F$ for $i = 1, \ldots, n$ where the distribution F depends on the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$.

The maximum likelihood estimates (MLEs) are the parameter values that maximize the likelihood (or log-likelihood) function, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|\mathbf{x})$$

where $\boldsymbol{\Theta} = \Theta_1 \times \cdots \times \Theta_p$ is the joint parameter space with Θ_j denoting the parameter space for the j-th parameter, i.e., $\theta_j \in \Theta_j$ for all j.

Maximum likelihood estimates have desirable large sample properties:

- *consistent*: $\hat{\theta}_{\text{MLE}} \to \theta$ as $n \to \infty$
- *asymptotically efficient*: $\text{Var}(\hat{\theta}_{\text{MLE}}) \leq \text{Var}(\hat{\theta})$ as $n \to \infty$
- *functionally invariant*: if $\hat{\theta}_{\text{MLE}}$ is the MLE of θ, then $h(\hat{\theta}_{\text{MLE}})$ is the MLE of $h(\theta)$ for any continuous function $h(\cdot)$

MLE for Normal Distribution

Suppose that $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. Assuming that $X \sim N(\mu, \sigma^2)$, the probability density function can be written as

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

This implies that the log-likelihood function has the form

$$\ell(\mu, \sigma^2|\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log(\sigma^2) - c$$

where $c = (n/2)\log(2\pi)$ is a constant with respect to μ and σ^2.

MLE for Normal Distribution (part 2)

Maximizing $\ell(\mu, \sigma^2|\mathbf{x})$ with respect to μ is equivalent to minimizing

$$\ell_1(\mu|\mathbf{x}) = \sum_{i=1}^{n}(x_i - \mu)^2$$

which is the least squares loss function that we encountered before.

We can use the same approach as before to derive the MLE:

- Take the derivative of $\ell_1(\mu|\mathbf{x})$ with respect to μ
- Equate the derivative to zero and solve for μ

The MLE of μ is the sample mean, i.e., $\hat{\mu}_{\text{MLE}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

MLE for Normal Distribution (part 3)

Maximizing $\ell(\mu, \sigma^2|\mathbf{x})$ with respect to σ^2 is equivalent to minimizing

$$\ell_2(\sigma^2|\hat{\mu}, \mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\log(\sigma^2)$$

Taking the derivative of $\ell_2(\sigma^2|\hat{\mu}, \mathbf{x})$ with respect to σ^2 gives

$$\frac{d\ell_2(\sigma^2|\hat{\mu}, \mathbf{x})}{\sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n}{\sigma^2}$$

Equating the derivative to zero and solving for σ^2 reveals that $\hat{\sigma}^2_{\text{MLE}} = \tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

MLE for Binomial Distribution

Suppose that $x_i \overset{\text{iid}}{\sim} B[N,p]$ for $i = 1, \ldots, n$. Assuming that $X \sim B[N,p]$, the probability density function can be written as

$$f(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x} = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}$$

This implies that the log-likelihood function has the form

$$\ell(p|\mathbf{x}, N) = \log(p) \sum_{i=1}^{n} x_i + \log(1-p) \left(nN - \sum_{i=1}^{n} x_i \right) + c$$

where $c = n \log(N!) - \sum_{i=1}^{n} [\log(x_i!) + \log((N-x_i)!)]$ is a constant.

MLE for Binomial Distribution (part 2)

Taking the derivative of the log-likelihood with respect to p gives

$$\frac{d\ell(p|\mathbf{x}, N)}{dp} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(nN - \sum_{i=1}^{n} x_i\right)$$

Setting the derivative to zero and multiplying by $p(1-p)$ reveals that the MLE satisfies

$$(1-p)n\bar{x} - pn\left(N - \bar{x}\right) = 0 \quad \rightarrow \quad \bar{x} - pN = 0$$

Solving the above equation for p reveals that the MLE of p is

$$\hat{p}_{\text{MLE}} = \frac{1}{nN}\sum_{i=1}^{n} x_i = \frac{\bar{x}}{N}$$

MLE for Uniform Distribution

Suppose that $x_i \overset{\text{iid}}{\sim} U[a,b]$ for $i = 1, \ldots, n$. Assuming that $X \sim U[a,b]$, the probability density function can be written as

$$f(x|a,b) = \frac{1}{b-a}$$

This implies that the log-likelihood function has the form

$$\ell(a,b|\mathbf{x}) = -\sum_{i=1}^{n} \log(b-a) = -n\log(b-a)$$

Maximizing $\ell(a,b|\mathbf{x})$ is equivalent to minimizing $\log(b-a)$ with the requirements that $a \leq x_i$ for all $i = 1, \ldots, n$ and $b \geq x_i$ for all i.

- MLEs are $\hat{a}_{\text{MLE}} = \min(x_i) = x_{(1)}$ and $\hat{b}_{\text{MLE}} = \max(x_i) = x_{(n)}$

References

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology 13*, 1–19.

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Tibshirani, R. (1996). Regression and shrinkage via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*, 301–320.