Probabilistic Modeling and Statistical Computing Fall 2015

November 9, 2015

Comparing two Means

Question (= null hypothesis): **are two population means the same?** Recall the permutation test approach:

- Given two samples from two populations,
- combine the samples
- use random permutations to redistribute the combined sample to the two samples
- compute the difference of sample means for each random permutation.

This is the simulated null distribution. Use it to compute the p-value of the observed difference.

Properties of this Approach

- No additional assumptions about the distribution
- Obtain a credible distribution under the null hypothesis
- Obtain a credible p-value
- Only the null distribution is simulated, so we cannot obtain information about the actual difference

Titanic Data

Bootstrap Approach

More general question: **What can we say about the two population means? Are they the same? Estimate of the difference? Accuracy of that estimate?**

- Make many bootstrap samples from each of the two samples
- Use these to make many bootstrap versions of the difference of sample means
- Examine the bootstrap distribution of differences: median, mean, shape, spread, quantiles, confidence interval, . . .

Properties of this Approach

- No additional assumptions about the distribution
- Obtain a credible distribution of the actual test statistic
- Can obtain estimates for center and spread, confidence intervals, etc.

Titanic Data

Joint probability density function

Given a random variable X with probability mass / density function $f(x|\theta)$, where θ is some parameter. Distribution of n independent observations $X_1, \ldots, X_n$:

Joint pdf / pmf

$$f_{joint}(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdot \cdots \cdot f(x_n|\theta)$$

Probability Theory: Assume that θ is given and the x_i are variables.

Likelihood function

Given a random variable X with probability mass / density function $f(x|\theta)$, where θ is some parameter. Assume a sample of n independent observations $X = x_1, \ldots, X = x_n$ is given.

Likelihood function

$$L(\theta|x_1, \ldots, x_n) = f(x_1|\theta) \cdot \cdots \cdot f(x_n|\theta)$$

This is the same as the joint probability density/mass function. Assume now that the x_i are given and θ is unknown.

Example: Poisson distribution

Discrete distribution on $\{0, 1, 2, \dots\}$, parameter λ = intensity

The pmf is $f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ for $x = 0, 1, 2, \dots$

The joint pmf of n independent observations is

$$\begin{aligned} f_{joint}(x_1, \dots, x_n|\lambda) &= e^{-n\lambda}\frac{\lambda^{x_1}}{x_1!}\cdots\frac{\lambda^{x_n}}{x_n!} \\ &= e^{-n\lambda}\frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\dots x_n!} \\ &= L(\lambda|x_1, \dots, x_n) \end{aligned}$$

This is also the likelihood function.

Example: Exponential distribution

Continuous distribution on $[0, \infty)$, parameter λ = intensity

The pmf is $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$

The joint pmf of n independent observations is

$$\begin{aligned} f_{joint}(x_1, \ldots, x_n|\lambda) &= \lambda e^{-\lambda x_1} \ldots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda x_1 - \lambda x_2 - \ldots \lambda x_n} \\ &= L(\lambda|x_1, \ldots, x_n) \end{aligned}$$

This is also the likelihood function.

Example: Bernoulli distribution

Discrete distribution on $\{0, 1\}$, parameter p = success probability

The pmf is $f(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$

The joint pmf of n independent observations is

$$\begin{aligned} f_{joint}(x_1, \ldots, x_n|p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \\ &= L(p|x_1, \ldots, x_n) \end{aligned}$$

Likelihood function and data reduction

The likelihood function sometimes depends only on a sample statistic.

Exponential distribution:

$$L(\lambda|x_1,\ldots,x_n) = \lambda^n e^{-\lambda x_1 - \lambda x_2 \cdots - \lambda x_n}$$

depends only on $x_1 + \cdots + x_n = n\bar{x}$.

Bernoulli distribution:

$$L(\lambda|x_1,\ldots,x_n) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

depends only on $x_1 + \cdots + x_n$.

Log Likelihood

Take the logarithm of the likelihood function.

Poisson distribution

$$\log L = -n\lambda + (\sum_i x_i)\log\lambda - \sum_i \log x_i!$$

Exponential distribution

$$\log L = n\log\lambda - \lambda(\sum_i x_i)$$

Bernoulli distribution

$$\log L = \log p(\sum_i x_i) + \log(1-p)(n - \sum_i x_i)$$

Maximum Likelihood

Observe the graphs of the likelihood functions.

Where are the maxima?

Maximum Likelihood Estimation

Estimate the unknown parameter θ by using the maximum of the likelihood function,

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta} L(\theta | x_1, \ldots, x_n)$$

Use **Optimization Theory** to work out the maximum or to compute it numerically.

Examples

Poisson distribution: $\boxed{\hat{\lambda}_{MLE} = \bar{x}}$

Exponential distribution: $\boxed{\hat{\lambda}_{MLE} = \frac{1}{\bar{x}}}$

Bernoulli distribution: $\boxed{\hat{p}_{MLE} = \bar{x}}$

- Theoretical justification of intuitive choices
- Shows how to reduce data
- General method

Cauchy Distribution

Continuous distribution on $\mathbb{R}$, parameter θ = center

The pmf is $f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$ for $x \in \mathbb{R}$

The joint pmf of n independent observations is

$$f_{joint}(x_1, \ldots, x_n|\theta) = \frac{1}{\pi^n(1+(x_1-\theta)^2)\ldots(1+(x_n-\theta)^2}$$
$$= L(\theta|x_1 \ldots, x_n)$$

Difficult to minimize

Normal Distribution

Consider normal distribution $N(\mu, \sigma^2)$.

The likelihood function depends on two parameters, μ and σ^2.

Need **calculus of several variables** to minimize.

Maximum likelihood estimates:

$$\boxed{\hat{\mu}_{MLE} = \bar{x}, \quad \hat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The estimator for μ is unbiased, the estimator for σ^2 has a non-zero bias!

Method of Moments Estimation

Given a random variable X whose distribution depends on a parameter θ. To estimate θ,

- Express a moment $\mathcal{E}(X)$ or $\mathcal{E}(X^2)$ or ... in terms of θ, e.g. $\mathcal{E}(X) = H(\theta)$
- Estimate this moment from the sample
- Solve the equation relating the moment and the parameter, e.g. solve $\bar{x} = H(\hat{\theta})$ for $\hat{\theta}$.

Similar to a plug-in estimation

Avoids calculus, only algebra is needed

Example: Beta Distribution

Continuous distribution on $(0, 1)$, parameters $\alpha, \beta > 0$

The pdf is

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

for $0 < x < 1$

Likelihood function is complicated. Calculus minimization is challenging, due to Γ function.

Estimation using Method of Moments

Known for the beta distribution:

$$\mathcal{E}(X) = \frac{\alpha}{\alpha+\beta}, \quad var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

MoM approach: Use sample mean $\bar{x}$ and sample variance $\bar{v}$. Solve the equations

$$\bar{x} = \frac{\alpha}{\alpha+\beta}, \quad \bar{v} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Resulting Estimators

After some algebra . . .

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right), \quad \hat{\beta} = (1-\bar{x})\hat{\alpha}$$

What if $\bar{v} > \bar{x}(1-\bar{x})$? The estimates then are negative!

`R` package uses a numerical method to maximize the likelihood.

Bias

Bias = systematic error

Formal Definition

Suppose $\hat{\theta}$ is an estimator (based on a random sample) for θ. The bias is defined as

$$\mathit{bias}(\hat{\theta}) = \mathcal{E}(\hat{\theta}) - \theta\,.$$

This suggests a theoretical evaluation. It also permits a simulation approach.

Example: Poisson Distribution

The maximum likelihood estimator for λ is the sample mean, $\hat{\lambda} = \bar{X}$. We know that

$$\mathcal{E}(X_i) = \lambda \implies \mathcal{E}(\bar{X}) = \lambda\,.$$

Therefore,

$$\boxed{\mathcal{E}(\hat{\lambda}) - \lambda = 0}$$

This estimator is **unbiased**.

Exponential Distribution

The maximum likelihood estimator for λ is $\hat{\lambda} = \frac{1}{\bar{X}}$. We know that

$$\mathcal{E}(X_i) = \frac{1}{\lambda} \implies \mathcal{E}(\bar{X}) = \frac{1}{\lambda}.$$

But in general

$$\mathcal{E}(\hat{\lambda}) = \mathcal{E}\left(\frac{1}{\bar{X}}\right) \neq \lambda$$

Can assess and correct the bias with a simulation (bootstrap).

Efficiency

Given two estimators $\hat{\theta}_1$, $\hat{\theta}_2$ for the same parameter. If both are unbiased, the one with smaller variance is better ("more efficient").

Relative Efficiency of $\hat{\theta}_1$ wrt. $\hat{\theta}_2$

Assuming $\mathcal{E}(\hat{\theta}_1) = \mathcal{E}(\hat{\theta}_2) = \theta$, this is defined as

$$E = var(\hat{\theta}_2)/var(\hat{\theta}_1)$$

If $\hat{\theta}_2$ is used instead of $\hat{\theta}_1$, the sample size must be increased by a factor E to get the same accuracy.

Example: Mean and Median

Consider data from a normal distribution, $N(\mu, 1)$. Can estimate μ in two ways from a sample $x = (x_1, \ldots, x_n)$:

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = median(x)$$

What is the relative efficiency?

Mean Square Error

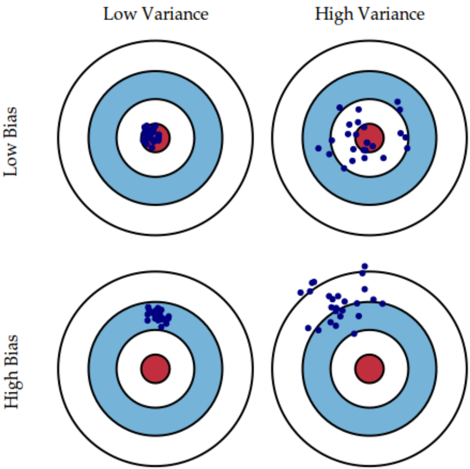
Combine variance and bias to assess quality of an estimator:

MSE

For an estimator $\hat{\theta}$,

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

Bias and Variance



Example: Uniform Distribution

Consider data from a uniform distribution $U(0, \beta)$ with unknown β. Given a sample $x = x(x_1, \ldots, x_n)$,

the ML estimator is $\boxed{\hat{\beta}_1 = \max_i x_i}$

the MoM estimator is $\boxed{\hat{\beta}_2 = 2\bar{x}}$.

- Which one is biased?
- Which one has smaller MSE?
- How does this depend on the sample size?