Recap

Parameter: unknown number that we are trying to get an idea about using a sample $X_1, \ldots, X_n$.

Statistic: A function of the sample. It is a random variable.

Estimator: A particular statistic that is used to get an idea about the parameter. It is a random variable.

Estimate: The value of the estimator for an observed sample $x_1, \ldots, x_n$. It is a number, known once we have observed the sample.

What are good statistics?

Cramér-Rao inequality

Best unbiased estimators

Good statistics, cont.

Sometimes mle and mom are the same, sometimes they are different. How do we choose between them?

Consider a Bernoulli experiment, performed $n$ times. Typically we do not care about the order of successes and failures, just the number $X$ of successes. In what sense does $X$ contain all the information about $p$?

Let $n=5$, $X=3$. Then

$P(01101|X = 3) =

P(11100|X = 3) =$

Sufficiency

Let $X_1, \ldots, X_n$ be iid with pdf/pmf $f(x; \theta)$. A statistic $W = h(X_1, \ldots, X_n)$ is sufficient for $\theta$ if $f_{w}(w; \theta)$ is independent of $\theta$.

Intuitively, this means that $W$ contains all the information about $\theta$ in the sample.
The factorization theorem

If we have a W we can check whether it is sufficient. But how do we find W?

Fisher-Neyman factorization criterion:

\[ W = h(X_1, \ldots, X_n) \]

is sufficient if and only if

\[ L(\theta) = g(w; \theta)s(x_1, \ldots, x_n) \]

function of parameter and \( w = h(x_1, \ldots, x_n) \)

The binomial case

\[ L(p) = p^\sum x (1 - p)^{n - \sum x} \]

The uniform case

\[ L(\theta) = \]

Is the mome a function of the sufficient statistic? Is the mle?

The beta case

\[ f_X(x; \theta) = \theta x^{\theta-1}, 0 < x < 1 \quad \text{Beta}(\theta, 1) \]

\[ L(\theta) = \prod_{i=1}^n f_X(x_i; \theta) = \theta^n \prod_{i=1}^n x_i^{\theta-1} \]

Not a statistic!

\[ \hat{\theta} = -\frac{n}{\ln(w)} \quad \hat{\theta} = \frac{\bar{x}}{1 - \bar{x}} \]

\[ E(X) = \int_0^1 x \theta x^{\theta-1} \, dx = \frac{\theta}{\theta + 1} \]
The gamma distribution

Maximum likelihood

Here is a very important fact: any mle is a function of the data only through a sufficient statistic. This is not necessarily true for the method of moments.

\[
\ell(\theta) = \ln(g(w; \theta)) + \ln(s(x_1, ..., x_n))
\]

\[
\ell'(\theta) = \frac{\partial}{\partial \theta} g(w; \theta)
\]

Sums of random variables

X~Bin(n,p) Y~Bin(m,p), X, Y independent
X+Y~

X~Po(\lambda), Y~Po(\mu), X, Y Independent
X+Y~

X~NegBin(r,p) Y~NegBin(s,p) X,Y independ.
X+Y~

X~Geom(p) Y~Geom(p) X,Y independ
X+Y~

More sums

X~\Gamma(\alpha, \beta), Y~\Gamma(\delta, \beta), X,Y independent
X+Y~

X~\chi^2(n), Y~\chi^2(m), X, Y independent
X+Y ~

X~N(\mu, \sigma^2) Y~N(\eta, \tau^2)
X+Y~

If \text{Var}(X) < \infty, X_1, ..., X_n iid as X

\[
\sqrt{n} \left( \sum_{i=1}^{n} X_i - \text{E}(X) \right) \overset{\text{d}}{\sim} N(0,1)
\]
Problems from 1/31

1. Denote the cdf for the standard normal $\Phi$. Then
$$P(X_i \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$
(a) $P(X_+ \leq \mu) = \Phi(0)^n = 2^{-n}$
(b) $P(X_- \geq \mu) = (1 - \Phi(0))^n = 2^{-n}$
(c) $P(X_- \leq \mu X_+) = P(X_+ \geq \mu) - P(X_- \geq \mu) = 1 - 2^{-n} - 2^{-n} = 1 - 2^{-(n-1)}$

2. The mle $\hat{p} = X/n$ so we consider $n\hat{p}\hat{q}$ which has expected value
$$n \frac{E(X)}{n} - n \frac{E(X^2)}{n^2} = n(p - \frac{pq}{n} - p^2) = (n-1)pq$$

3. (a) $L(\theta) = \theta^2 (1.2 \times 1.8)^{-\theta-1}$
   (b) $L(2) = 0.40$ $L(3) = 0.41$ $L(4) = 0.34$, so the mle is 3.

4. (a) $P(Y = x | Y \geq 1) = \frac{P(Y = x)}{P(Y \geq 1)} = \frac{P(Y = x)}{1 - P(Y = 0)}$
   $$= \frac{\lambda^x e^{-\lambda}}{x! (1 - e^{-\lambda})}$$
   (b) $E(T(X)) = 2P(X \text{ even}) = 2 \sum_{j=1}^{\infty} \frac{\lambda^j e^{-\lambda}}{(2j)! (1 - e^{-\lambda})}$
   $$= 2 \frac{e^{-\lambda}}{1 - e^{-\lambda}} \left( \frac{e^\lambda + e^{-\lambda}}{2} - 1 \right) = 1 - e^{-\lambda}$$

Review

Likelihood
Maximum likelihood estimator
Method of moments
Unbiasedness
Relative efficiency
Mean squared error
Cramér-Rao lower bound
Efficient estimators
 Sufficiency
Fisher-Neyman factorization theorem
Consistency

An estimator \( \hat{\theta} \) is \textit{consistent} if for all \( \varepsilon > 0 \)
\[
\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0.
\]
We say that \( \hat{\theta} \) \textit{converges in probability} to \( \theta \).

One way to show this is to use Chebyshev’s inequality from last quarter.
So, for example, any sample average is a consistent estimate of its expected value provided the variance of the underlying distribution is finite.

Consistency of mles

Under fairly general conditions (involving smoothness of the density as a function of the parameter) all mles are consistent.

One can also show that if \( \theta_n \xrightarrow{p} \theta \) then for continuous functions \( h \)
\[
h(\theta_n) \xrightarrow{p} h(\theta)
\]
This can be used to show consistency of method of moments estimators

Binomial case

Returning to our \( n=4 \) example, here are two of the unbiased estimators we considered:
\[
\hat{p}_1 = \bar{X},
\]
\[
\hat{p}_2 = X_2
\]
Which (if any) of these are consistent?
Geometric distribution

The mle is \( \hat{p} = 1/\bar{X} \). By the law of large numbers \( \bar{X} \) converges in probability to

\[
E(X) = \frac{1}{\theta}.
\]

Uniform

The mle \( \text{max}(X) \) has pdf \( ny^{n-1}/\theta^n \). Hence

\[
P(\hat{\theta} - \theta > \epsilon) = P(\hat{\theta} > \theta + \epsilon) + P(\hat{\theta} < \theta - \epsilon)
\]

\[
= 0 + \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0
\]

Asymptotic normality

We will show that as \( n \) gets larger, the distribution of the mle approaches a certain normal distribution. One says the mle is approximately normal, with mean \( \theta \) and variance the Cramér-Rao bound. Thus mles are asymptotically efficient in most cases. The assumptions needed relate to the mle having finite mean and variance.

Geometric case
Interval estimates

The standard error of an estimator has two uses:
(1) comparison to other estimators
(2) assessment of uncertainty

An interval estimate combines an estimate and its estimated standard error into a random interval which covers the true (but unknown) value of $\theta$ with a given probability or confidence coefficient $1 - \alpha$.

For a particular sample, the interval either does or does not cover $\theta$. WE DO NOT KNOW WHICH. In the long run it covers $\theta$ in the proportion $1 - \alpha$ of all data sets.

Ideal number of children

In 1986, 1370 US adults were asked “What do you think is the ideal number of children for a family to have?”

<table>
<thead>
<tr>
<th>#children</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>17</td>
<td>24</td>
<td>742</td>
<td>354</td>
<td>185</td>
<td>24</td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

The sample average is 2.60, sample sd 0.97, so $\text{ese}(\bar{X}) = \frac{0.97}{\sqrt{1370}} = 0.03$

A confidence interval for the mean ideal family size is $(2.52, 2.68) = \bar{X} \pm 2.67\text{ese}(\bar{X})$

How far off does the sample average have to be for the interval to miss the population mean? More precisely, what is $P(|\bar{X} - \mu| > 2.67 \text{ese}(\bar{X})$? The sample average is approximately normal (why?), and for large samples ese is about se (why?), we can compute this probability as $2(1 - \Phi(2.67)) = 0.0076$

Hence, $1 - 0.0076 = 0.9924$ is the probability that the interval does cover $\mu$.

We take something known and use a theoretical distribution to estimate something unknown and we compute the probability that we are correct.

The exponential case

Let $X \sim \text{exp}(\lambda)$. The mle is $\hat{\lambda} = \frac{1}{X}$. Will the interval $(0.01\hat{\lambda}, 100\hat{\lambda})$ cover the true value of $\lambda$?

How about $(0.99\hat{\lambda}, 1.01\hat{\lambda})$?

Consider the interval $(c_1\hat{\lambda}, c_2\hat{\lambda})$. How can we determine $c_1$ and $c_2$ to make this a 95% CI?
Monday’s lecture

Asymptotics = large sample size
Consistency
Asymptotic normality
Confidence intervals

Multiple intervals

A researcher constructs CIs for 15 different chemical reaction constants. Each interval has 90% confidence coefficient, and they are each constructed from independent measurements. Some may cover the true value, some may not.

What is the probability that all intervals cover their constants?

What is the most likely number covered?

About CIs

The probability involved in computing the confidence coefficient has to do with the procedure. A particular interval either covers the parameter value or not, and we do not know which.

The confidence coefficient is NOT the probability that the parameter is in the interval: the parameter is not random, the interval is.

The interval tells us something about the accuracy of our estimate. The shorter the interval, the more accurate our estimate.

Normal case

Consider a sample from $N(\mu, 1)$. We would estimate $\mu$ by $\bar{x}$, an observation of the random variable $\bar{X} \sim N(\mu, 1/n)$. Now note that $\bar{X} - \mu \sim N(0, 1/n)$ has a distribution that does not depend on $\mu$. Such a quantity is called a pivot and makes it particularly easy to create a CI.

$$1 - \alpha = P(z_{\alpha/2} \leq \sqrt{n}(\bar{X} - \mu) \leq z_{1-\alpha/2})$$

$$= P(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{z_{1-\alpha/2}}{\sqrt{n}})$$

Since $z_{\alpha/2} = -z_{1-\alpha/2}$ we get the observed CI

$$\left( \bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}}{\sqrt{n}} \right)$$
Beer preferences

100 Budweiser drinkers (polishing off at least two 6-packs per week) were subjected to a blind taste test between Schlitz and Budweiser. 46 of the subjects preferred Schlitz.

Y=# subjects preferring Schlitz. Then

\[ Y \sim \text{Binomial}(n, p) \]

From the Central Limit Theorem

\[
\frac{Y - np}{\sqrt{npq}} = \frac{\frac{Y}{n} - p}{\sqrt{\frac{pq}{n}}} = \frac{\hat{p} - p}{\text{se}(\hat{p})} \sim N(0, 1)
\]

Numerically, this is also (0.36, 0.56), so this is an easier way to do things (and there will only be a difference when \( n \) is small).

Note that in this second approach there are two approximations: approximating \( \text{se} \) by \( \text{ese} \) (LLN) and approximating the standardized distribution of \( \hat{p} \) by a normal distribution (CLT).

Were the beer drinkers able to tell the beers apart?

Large sample CIs from mle

We have seen several instances of CIs based on the mle of the form

\[ \hat{\theta} \pm z_{1-\alpha/2} \text{ese}(\hat{\theta}) \]

This is based on the asymptotic normality of the mle, and we often can use the general formula

\[ \text{ese}(\hat{\theta}) = \left( -\frac{\partial^2}{\partial \theta^2} \ell_n(\theta) \right)^{-1/2} \]
The exponential case

\[ \ell(\lambda) = n \ln \lambda - \frac{n}{\lambda} \sum_{i=1}^{n} x_i \]

\[ \ell'(\lambda) = \frac{n}{\lambda} - \frac{1}{\lambda} \sum_{i=1}^{n} x_i \]

\[ \ell''(\lambda) = -\frac{n}{\lambda^2} \Rightarrow (\ell''(\hat{\lambda}))^{-1} = \frac{\lambda}{\sqrt{n}} \]

so the 95% approximate confidence interval is

\[ \hat{\lambda}(1 \pm \frac{1.96}{\sqrt{n}}) \]

What sample size do we need?

The US Commission on Crime wants to estimate the proportion of crimes related to firearms in an area with one of the highest crime rates in the country. They intend to draw a random sample of files of recently committed crimes in the area, and want to know the proportion of cases with firearms to within 5% of the true proportion with probability at least 90%. How many files do they need to look at?

Problem solutions

1. (a) \( 0 \leq X \leq \theta \Rightarrow 0 \leq U \leq 1 \)

   \[ P(U \leq x) = P(X \leq x\theta) = \frac{x\theta}{\theta} = x \]

(b) \( P(V_n \leq x) = P(U_1 \leq x, \ldots, U_n \leq x) = x^n \)

   \[ P(n(1 - V_i) \leq x) = P(V_n \geq 1 - \frac{x}{n}) = 1 - \left(1 - \frac{x}{n}\right)^n \rightarrow 1 - e^{-x} \]

(c) \( P(n(\theta - \hat{\theta}_n) \leq x) = P(n\theta(1 - V_i) \leq x) \rightarrow 1 - e^{-x/\theta} \)

2. \( \text{se}(\hat{\lambda}_n) = \sqrt{\frac{\lambda}{100}} \) and the two estimators are independent, so

\[ \text{se}(\hat{\lambda}_1 - \hat{\lambda}_2) = \sqrt{\text{se}(\hat{\lambda}_1)^2 + \text{se}(\hat{\lambda}_2)^2} \]

which can be estimated by

\[ \sqrt{\frac{9.4 + 8.6}{100}} = 0.42 \]

\( E(\hat{\lambda}_1 - \hat{\lambda}_2) = \lambda - \lambda = 0 \)

and

\[ \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\text{ese}(\hat{\lambda}_1 - \hat{\lambda}_2)} = \frac{0.8}{0.42} = 1.89 \]

It is not too surprising to see this (3 ese difference would be surprising).

3. (a) Since the \( Y_i \) count when there is a new mark, the sum must be the number marked.

(b) Given \( m \) marked individuals, the probability of drawing a marked one is \( m/\theta \).
(c) Given what happened in draws 1,...,i-1 there are m_{i-1} marked and we either draw a marked (y_i = 0) or an unmarked (y_i = 1) with the probability of the first given in (b) and of the second being 1- that. The joint probability of all n draws is the product of the conditional probabilities.

(d) Note that y_1 = 1. Suppose n = 5 and we draw 1 0 1 1 0. Using the formula in (b) we get

\[ \frac{1}{\theta} \times \frac{\theta - 1}{\theta} \times \frac{\theta - 2}{\theta} \times \frac{2}{\theta} \]

\[ \times \left( \frac{\theta - 1}{\theta} \times \frac{\theta - 2}{\theta} \right)^{n-1} \]

The general formula is obtained in the same way, and we see that r is sufficient by the factorization criterion.

4. Chebyshev says \( P(|Y|\geq c) \leq \frac{E(Y^2)}{c^2} \). Let \( Y = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \). Then \( E(Y^2) = 1 \), and \( P(|Y|\leq c) \geq 1 - \frac{1}{c^2} \). In other words

\[ P(\hat{\theta} - c \text{se}(\hat{\theta}) \leq \theta \leq \hat{\theta} + c \text{se}(\hat{\theta})) \geq 1 - \frac{1}{c^2} \]

so \( \hat{\theta} \pm \frac{\text{se}(\hat{\theta})}{\sqrt{\alpha}} \) is at least a 1-1/c^2 confidence interval.

**Another pivot**

\( X_1,...,X_n \) iid density \( f_X(x; \theta) = \frac{x}{\theta^2} e^{-x/\theta} \)

From midterm, \( \hat{\theta} = \frac{1}{n} \sum x_i / 2 \)

\( Y = \sum_{i=1}^n X_i / \theta \sim \)

**One-sided CIs**

Airplanes are inspected for corrosion every 10 years. A company has inspected 5 of their fleet of 200 planes, finding no corrosion, and would like a 95% lower confidence bound on the probability \( p \) of no corrosion in 10 years, i.e. a value \( p_L(x,1-\alpha) \) such that

\[ P(p_L(X,1-\alpha) \leq p) \geq 1-\alpha. \]
One-sided CIs, cont.

Poisson approximation?

Another approach (we will justify it later) is to use

\[ C(x) = \{ p : P_p(X \geq x) > 1 - \alpha \} \]

Bayesian methods

Recall Bayes' formula

\[ P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^{n} P(B | A_i)P(A_i)} \]

In the context of continuous random variables X and Y this becomes

\[ f_{y|x}(y|x) = \frac{f_{x|\theta}(x|\theta)f_{\theta}(\theta)}{\int f_{x|\theta}(x|\theta)f_{\theta}(\theta)d\theta} \]

Expressing uncertainty

In the Bayesian approach to statistic we describe anything that is uncertain by a probability distribution. The likelihood is the conditional distribution of data X given the parameter (now a random variable) \( \Theta \):

\[ L(\theta) = f_{x|\theta}(x|\theta) \]

Before we collect data we assign a prior distribution to \( \Theta \): \( f_{\theta}(\theta) \)

After observing data \( x \), we compute the posterior distribution of \( \Theta \): \( f_{\theta|x}(\theta|x) \)

Where does the prior come from?

Previous experience
Expert knowledge
Mathematical convenience

The posterior depends on the prior. But if you have a lot of data, the posterior will look similar to the likelihood.
Production

The error rate in production of a computer chip is about 9%. The proportion P of faulty chips has a prior distribution proportional to \( p^{10}(1-p)^9 \). From a large batch, we sample 100 chips and find 16 defective.

\[ L(p) = \]

\[ f(p|x) = \]

What if we make another experiment?

Use the posterior from the past experiment as the prior for the new one. This is the same as using the original prior and then performing the combined experiment.

Exponential case

Let the prior be \( \text{Exp}(\alpha) \) and the data \( \text{Exp}(\lambda) \). Then the posterior is

\[ f(\lambda|x) \propto \alpha^\lambda \exp \left\{ -\lambda \left( \sum_{i=1}^{n} x_i + \alpha \right) \right\} \]

This is a gamma density with shape parameter \((n+1)\) and scale parameter \( \sum x_i + \alpha \). 

\( \alpha \) is called a hyperparameter, and is set by the statistician based on prior expectations.
Conjugate priors

A mathematically convenient prior is one where the prior and posterior are in the same parametric family. Poisson likelihood:

\[ L(\lambda) = \lambda^\sum x \exp(-n\lambda) \]

If we look at things involving \( x \)'s (or \( n \)) as parameters, and things involving \( \lambda \) as the dummy variable, we choose a prior of the form \( f(\lambda) = \lambda^\alpha \exp(-\beta \lambda) \) which is a gamma density. Then the posterior density will also be gamma, with shape parameter \( \alpha + \sum x_i \) and shape parameter \( \beta + n \).

Credible interval

Find an interval on the posterior distribution such that the probability that the parameter falls in that interval is 95%. Common way: high posterior density interval

Computer chips

![Computer chips graph]