STAT 135, Concepts of Statistics

Helmut Pitters

Parameter estimation

Department of Statistics
University of California, Berkeley

February 20, 2017

Review: Hypergeometric distribution

A population contains G good and $N - G$ bad elements.
Randomly sample $n \leq N$ elements without replacement.

$$S_n := \text{number of good elements in sample}$$

follows a hypergeometric distribution, i.e.

$$\boxed{\mathbb{P}\left\{S_n = g\right\} = \frac{\binom{G}{g}\binom{N-G}{n-g}}{\binom{N}{n}}.}$$

Recall:

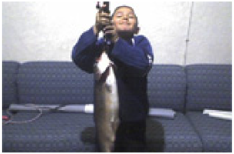
$$\boxed{\mathbb{E}S_n = np, \qquad \mathrm{Var}(S_n) = npq\frac{N-n}{N-1},}$$

where $p = G/N$, $q = (N-G)/N$.

Parameter estimation

Example (Capture-recapture)

Estimating population size N.



9-year old **Antonio Martinez** of San Lorenzo caught a 12 lb., 7 oz., 27" trout at Don Castro using power bait on 4/6/2008!!

Parameter estimation.

Example (Capture-recapture)

Catch $r = 1000$ fish, mark them red, and release them.

Later, new catch of $n = 1000$ fish is made, among which $k = 100$ are found to have red marks. What can be said about the total (unknown) number N of fish in the lake?

Heuristics:

proportion of red fish in sample $\approx$ proportion of red fish in lake,

i.e.

$$\frac{k}{n} \approx \frac{r}{N}.$$

Consequently, we expect

$$\hat{N} := \frac{n}{k} r = \frac{r}{\frac{k}{n}}$$

to be a good estimator for N.

Parameter estimation.

Example (Capture-recapture)

Clearly, $N \geq r + (n - k)$.
Before the second catch is made, the distribution

$$R(n) := \text{the number of red fish in the sample}$$

follows a hypergeometric law, i.e.

$$\mathbb{P}\{R(n) = k\} = \text{hypergeometric}(N, r, n)(k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}.$$

Parameter estimation.

Example (Capture-recapture)

We don't expect $\hat{N} = r + n - k = 1900$ to be a good guess for N. In fact, if $\hat{N}$ were the actual number of fish, the outcome of our experiment would be rather unlikely, namely it would have probability

$$\text{hypergeometric}(\hat{N}, r, n)(k) = \binom{1000}{100}\binom{900}{900} \Big/ \binom{1900}{1000}$$
$$= \frac{(1000!)^2}{100!1900!},$$

which has order of magnitude 10^{-430} according to Stirling's formula, $n! \sim \sqrt{2\pi n}(n/e)^n$.

Question: Which number $\hat{N}$ should we pick as estimate for N in order to maximize likelihood of our observation?

(Notion of *maximum likelihood estimate* goes back to R. A. Fisher.)

Parameter estimation.

Example (Capture-recapture)

Let $p_N(k) :=$ hypergeometric$(N, r, n)(k)$. Consider the ratio

$$\begin{aligned}
\frac{p_N(k)}{p_{N-1}(k)} &= \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} \frac{\binom{N-1}{n}}{\binom{r}{k}\binom{N-1-r}{n-k}} \\
&= \frac{n!(N-n)!}{N!} \frac{(N-r)!}{(n-k)!(N-r-n+k)!} \\
&\quad \times \frac{(N-1)!}{n!(N-1-n)!} \frac{(n-k)!(N-1-r-n+k)!}{(N-1-r)!} \\
&= \frac{(N-r)(N-n)}{N(N-r-n+k)}.
\end{aligned}$$

This yields $p_N(k)/p_{N-1}(k) < 1$ if $nr < Nk$, and $p_N(k)/p_{N-1}(k) > 1$ otherwise.
Thus likelihood is maximized for N the integer closest to nr/k $(= 10000)$.

Next: Confidence interval around $\hat{N}$ via normal approximation.

Parameter estimation

Example (Emission of alpha particles)

Radioactive material of certain mass is monitored with a Geiger counter.

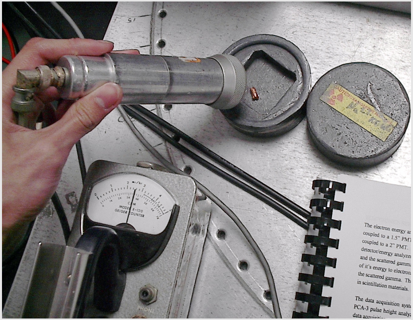


Figure: Geiger counter

Parameter estimation.

Example (Emission of alpha particles)

Assume

- rate of emission is constant over period of observation,
- particles come from large number of independent sources.

These assumptions justify model

$$N_t := \#\text{emissions during time } [0, t] \sim \mathrm{Poisson}(\alpha t),$$

for some parameter $\alpha > 0$.

Recall that $X \sim \mathrm{Poisson}(\alpha)$, i.e. X follows a Poisson distribution with parameter α, if

$$\mathbb{P}\{X = k\} = e^{-\alpha} \frac{\alpha^k}{k!}$$

for any non-negative integer k.

Review: Poisson distribution

$X \sim \mathrm{Poisson}(\alpha)$, i.e. X follows a Poisson distribution with parameter α, if

$$\mathbb{P}\{X = k\} = e^{-\alpha}\frac{\alpha^k}{k!}$$

for any non-negative integer k.

$$\mathbb{E}[X] = \alpha, \qquad \mathrm{Var}(X) = \alpha.$$

Parameter estimation.

Example (Emission of alpha particles)

Data from National Bureau of Standards.
Source of alpha particles: americium 241.
$10,220$ times between successive emissions were recorded. Total time was subdivided into 1207 intervals of 10sec. each.

$$\text{mean emission rate} = \frac{\text{\#emissions}}{\text{total time of observation}[s]} = \frac{10220}{12070s} \approx 0.839/s$$

Hence, on average 8.39 emissions are observed in an interval.

Parameter estimation.

Example (Emission of alpha particles)

Let

$$E_i := \text{\#emissions in } i\text{th interval} \sim \mathrm{Poisson}(10\alpha),$$

in particular $\mathbb{E}E_i = 10\alpha$. Data suggests that

$$\hat{\alpha} \approx 0.839/s$$

should be a good estimator for α.

Parameter estimation.

Example (Emission of alpha particles)

Next table shows summary of this data.
first column = number of counted emissions
second column = number of intervals with corresponding emission counts

Parameter estimation.

Example (Emission of alpha particles)

emission counts	number of intervals
0–2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9

Parameter estimation.

Example (Emission of alpha particles)

How could we compare the model with estimated parameter $\hat{\alpha} \approx 0.839/s$ to data? Notice that probability to have k emissions in interval i is

$$p(k) := \mathbb{P}\{E_i = k\} = e^{-8.39}\frac{(8.39)^k}{k!}$$

for any i. Thus number of intervals during which k emissions were counted is ($E_1, E_2, \dots$ are i.i.d.)

$$B_k := \sum_{i=1}^{1207} \mathbf{1}\{E_i = k\} \sim \mathrm{binomial}(1207, p(k)),$$

expected number of intervals counting k emissions is

$$\mathbb{E}B_k = 1207p(k).$$

Parameter estimation

Example (Emission of alpha particles)

emission counts	number of intervals	expected no. of intervals
0–2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9

From examining above table, our model seems to agree quite well with the data.

Ideally we'd like to quantify precisely "how well" the model fits. There might be a slightly different model that fits better?!

We will see measures for the "goodness of fit" of a model later on.

Parameter estimation. Setup

Let us think about previous examples from more abstract point of view. Have observations

$$x_1, x_2, \ldots, x_n$$

(e.g. number of emissions in certain time interval, whether or not caught fish is red, political opinion of voter, etc.)

which we regard as observed values of some random variables

$$X_1, X_2, \ldots, X_n.$$

Parameter estimation. Setup

In general, $X_1, \ldots, X_n$ are not simple random draws from finite population. They could be

- i.i.d. $\sim \mathcal{N}(\mu, \sigma)$,
- i.i.d. $\sim \mathrm{Poisson}(\lambda)$,
- i.i.d. $\sim \mathrm{Gamma}(\alpha, \lambda)$,
- generated by simple random sampling from specific population of N individuals with characteristics $x_1, \ldots, x_N$,
- generated by sampling with replacement,
- etc.

We will focus on models where common distribution $\mathbb{P}_\theta$ of X_i depends on some parameter, generically called $\theta > 0$.
(These are so-called *parametric models*.)

Parameter estimation. Setup

Having observed data $x = (x_1, \dots, x_n)$, what can we infer about θ? Would like to make statements about which values of θ are plausible, based on x.

Assume that

- distribution of X is known up to some parameter θ (e.g. distribution of X could be exponential with parameter θ).
- we have observed data x that we use to
  - construct a point estimate $\hat{\theta}$ of the value of θ,
  - construct a confidence interval of (plausible) values for θ,
  - test a hypothesis about θ.

Instead of "guessing" an estimator for θ as in the previous examples, we would like to have a more principled approach.

Remark

Notice that in general θ may not be a single real number, but could be an element of a more abstract set Θ (=*parameter space*).
E.g. $\mathbb{P}_{\mu,\sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma)$.

Parameter estimation. Method of moments

Random variable X has *k-th moment* (provided it exists)

$$\mu_k := \mathbb{E}[X^k].$$

Our goal will be to express θ in terms of the moments of X_1 of lowest possible order.

We will then use the *k-th sample moment*

$$\hat{\mu}_k := \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

as an estimator for μ_k, and thereby find an estimator for θ.

Once we have an estimator for θ, we will be interested in its accuracy (standard error), and, more generally, in studying its distribution, the so-called *sampling distribution*.

Parameter estimation. Method of moments

Example (Capture-recapture)

$$X_i := \begin{cases} 1 & i\text{th fish in sample has red mark} \\ 0 & \text{otherwise.} \end{cases}$$

$X_1, \dots, X_n$ is SRS from population of fishes

We are interested in parameter $\theta = N$. From

$$\mu_1 = \mathbb{E}[X_1] = \frac{r}{N},$$

we find $N = r/\mu_1$. Using $\hat{\mu}_1 = \bar{X} = k/n$ as estimator for μ, we find

$$\hat{N} = \frac{r}{\bar{X}} = r\frac{n}{k}$$

as estimator for N.

This is the so-called *method of moments estimator* for N.

Parameter estimation. Method of moments

Example

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

Interested in parameter $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$.

$$\mu_1 = \mu, \qquad \mu_2 = \mathrm{Var}(X_1) + \mathbb{E}[X_1]^2 = \sigma^2 + \mu^2,$$

thus

$$\theta_1 = \mu_1, \qquad \theta_2 = \mu_2 - \mu^2$$

and substituting sample moments, we obtain the estimators

$$\hat{\theta}_1 = \hat{\mu}_1 = \bar{X}, \qquad \hat{\theta}_2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We know: $\bar{X} \sim N(\mu, \sigma^2/n)$.
We will now see that $\hat{\theta}_1$, $\hat{\theta}_2$ are independent, and $n\hat{\theta}_2/\sigma^2 \sim \chi^2_{n-1}$.

Parameter estimation. Method of moments

Example

In order to study sampling distribution of estimator

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

need some results on distributions derived from the normal.

Review: Gamma distribution.

G has $\mathrm{Gamma}(\alpha, \lambda)$ distribution, if its probability density is given by

$$f_G(t) = \begin{cases} \frac{(\lambda t)^{\alpha-1}}{\Gamma(\alpha)} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and moments

$$\mathbb{E}[G^k] = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\lambda^k}.$$

For integer α can interpret G as waiting time until we see first event in Poisson Process of intensity λ.

If $G_1 \sim \mathrm{Gamma}(\alpha_1, \lambda)$, $G_2 \sim \mathrm{Gamma}(\alpha_2, \lambda)$ are independent, then

$$G_1 + G_2 \sim \mathrm{Gamma}(\alpha_1 + \alpha_2, \lambda).$$

Review: chi squared distribution

Recall:
$X_1, X_2, \ldots, X_n$ i.i.d. standard normals.

$$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \mathrm{Gamma}(\frac{n}{2}, \frac{1}{2}) = \chi_n^2$$

chi squared distribution with n degrees of freedom (df).

Aside: distributions derived from normal

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

An important result states that

$\bar{X}$ and $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ are independent.[1]

This immediately implies that sample mean $\bar{X}$ and sample variance

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

are independent.

[1] We will not prove this result in STAT 135.

Moment generating function

The *moment generating function (MGF)* $M_X(t)$ of r.v. X is defined as

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

if this mean exists.

Fact

If the MGF of X exists in an open interval containing 0, it uniquely determines the probability distribution of X.

Example (MGF Gamma distribution)

$G \sim \mathrm{Gamma}(\alpha, \lambda)$.

$$\begin{aligned} M_G(t) = \mathbb{E}[e^{tG}] &= \int_0^\infty e^{ts} f_G(s) ds = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} \int_0^\infty e^{ts} s^{\alpha-1} \lambda e^{-\lambda s} ds \\ &= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^\infty s^{\alpha-1} e^{-(\lambda-t)s} ds = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda-t)^{\alpha}} = (\frac{\lambda}{\lambda-t})^{\alpha}. \end{aligned}$$

Moment generating function

Fact

Suppose r.v.s X, Y are independent and their MGFs exist on open interval containing 0. Then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

on the common interval where both M_X and M_Y exist.

The proof is straightforward.

Aside: distributions derived from normal

We now have the tools to derive an important result about the distribution of the sample variance S^2.

Theorem

For $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$,

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Remark

From this theorem we get the previous claim for the distribution of the rescaled sample

$$n\hat{\theta}_2^2/\sigma^2 \sim \chi^2_{n-1},$$

where

$$n\hat{\theta}_2^2/\sigma^2 = n\hat{\sigma}^2/\sigma^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Aside: distributions derived from normal

Proof.

Let us replace $\bar{X}$ in $(n-1)S^2/\sigma^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2$ by μ.

$$L := \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2$$

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

Notice that RHS is sum of independent r.v.s $\square$

Aside: distributions derived from normal

Proof.

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

Letting $R := n(\bar{X}-\mu)^2/\sigma^2 = (\frac{X-\mu}{\sigma/\sqrt{n}})^2$, we have

$$M_L(t) = M_{(n-1)S^2/\sigma^2}M_R(t),$$

hence

$$M_{(n-1)S^2/\sigma^2}(t) = \frac{M_L(t)}{M_R(t)} = (\frac{\frac{1}{2}}{\frac{1}{2}-t})^{\frac{n}{2}}/(\frac{\frac{1}{2}}{\frac{1}{2}-t})^{\frac{1}{2}} = (\frac{\frac{1}{2}}{\frac{1}{2}-t})^{\frac{n-1}{2}},$$

hence $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.
We used here $M_X(t) = (\frac{\frac{1}{2}}{\frac{1}{2}-t})^{\frac{n}{2}}$ for $X \sim \chi^2_n$. □

Parameter estimation

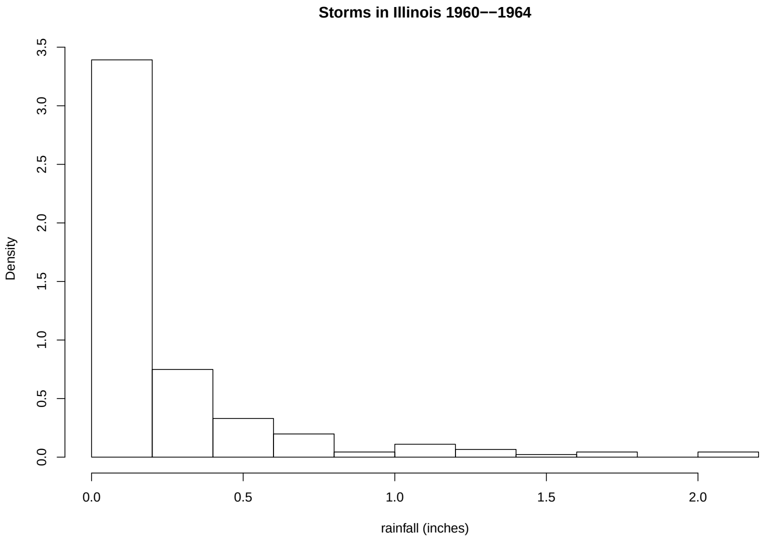


Figure: Precipitation in 227 Illinois storms.

Parameter estimation

Example (Illinois storms)

Would like to fit a probability distribution to the data. Since histogram is skewed, try to fit a gamma distribution.

Find parameters of gamma distribution via method of moments.

Parameter estimation

Example (Illinois storms)

Recall: $G \sim \Gamma(\alpha, \lambda)$ has moments

$$\mathbb{E}[G^k] = \frac{\Gamma(\alpha + k)}{\Gamma(k)\lambda^k},$$

hence (since $\Gamma(x+1) = x\Gamma(x)$)

$$\begin{aligned}
\mu_1 &= \mathbb{E}[G] = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\lambda} = \frac{\alpha}{\lambda}, \\
\mu_2 &= \mathbb{E}[G^2] = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2} = \frac{(\alpha+1)\alpha}{\lambda^2} = \frac{\alpha^2+\alpha}{\lambda^2},
\end{aligned}$$

and solving for α and λ, we have $\mu_2 = \mu_1^2 + \mu_1/\lambda$, hence

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\mu_1}{\sigma^2} \quad \text{and} \quad \alpha = \lambda\mu_1 = \frac{\mu_1^2}{\sigma^2}.$$

Parameter estimation

Example (Illinois storms)

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\mu_1}{\sigma^2} \quad \text{and} \quad \alpha = \lambda\mu_1 = \frac{\mu_1^2}{\sigma^2}$$

Substituting sample moments for population moments we obtain the method of moments estimators

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \qquad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}.$$

[show R script]

From data we find $\bar{X} = 0.224$, $\hat{\sigma} = 0.366$, hence

$$\hat{\lambda} = 1.672 \qquad \hat{\alpha} = 0.374.$$

Parameter estimation

Example (Illinois storms)

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \qquad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

Studying sampling distributions of $\hat{\lambda}$ and $\hat{\alpha}$ analytically (even working out standard errors) seems hopeless, as they are complicated functions of the observations.

Luckily, can still study sampling distribution with so-called *(parametric) bootstrap*, a versatile simulation method, thanks to computers.

Bootstrap

Example (Illinois storms)

Idea: Suppose we knew true values of α and λ, let's call them α_0 and λ_0.

Could then simulate many drawings of samples, $i = 1, \ldots, 1000$ say,

$$X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)} \sim \text{Gamma}(\alpha_0, \lambda_0)$$

of size $n = 227$ and compute estimator $\hat{\alpha}_i^*$ for each of them.

Histogram of the $\hat{\alpha}_i^*$ should then be a good approximation of the sampling distribution of $\hat{\alpha}_{\mathsf{MoM}}$.

Bootstrap

Example (Illinois storms)

Consequently, standard error

$$\mathrm{s}_{\hat{\alpha}} := \sqrt{\frac{1}{\mathrm{B}} \sum_{\mathrm{i}=1}^{\mathrm{B}} (\alpha^*_{\mathsf{MoM,i}} - \bar{\alpha})^2}, \qquad \left(\bar{\alpha} := \frac{1}{\mathrm{B}} \sum_{\mathrm{i}=1}^{\mathrm{B}} \alpha^*_{\mathsf{MoM,i}} \right)$$

of (simulated) estimators $\hat{\alpha}^*_1, \hat{\alpha}^*_2, \ldots, \hat{\alpha}^*_B$ should be good approximation of standard error of $\hat{\alpha}$.

Problem: We don't know true values α_0, λ_0.

Bootstrap

Example (Illinois storms)

However, since we don't know α_0 nor λ_0, we use their method of moments estimates instead.

[show histogram and standard deviation of α_i^* in R]

Maximum likelihood estimator

Maximum likelihood estimators. Setup

Assume data to be observations of r.v.s

$$X_1, X_2, \dots, X_n.$$

Additionally, assume joint distribution $\mathbb{P}_\theta$ of $(X_1, \dots, X_n)$ has probability density f.

Given observations $X_1 = x_1, \dots, X_n = x_n$ let

$$\mathrm{lik}(\theta) := f(x_1, \dots, x_n | \theta)$$

denote the *likelihood* of θ.

Think of $x_1, \dots, x_n$ as being fixed, and of θ as varying.
We ask: what is the most likely value for θ, given the observed data?

Maximum likelihood estimators.

Definition

The value $\hat{\theta}$ that maximizes the likelihood function, i.e.

$$\hat{\theta} = \arg\max_{\theta} \mathrm{lik}(\theta)$$

is called the *maximum likelihood estimate (MLE)* for θ.

Maximum likelihood estimators.

Example (Proportion of defectives)

We know that a proportion p of products from a certain manufacturer are defective, but we don't know the value of p.

We independently sample $n = 100$ of the manufacturer's products and find $S_n = 37$ of them to be defective.

What is your "best" guess for p?

Maximum likelihood estimators.

Example (Proportion of defectives)

Now, let's find the MLE for p. Let

$$X_i := \begin{cases} 1 & \text{if } i\text{th product is defective} \\ 0 & \text{otherwise.} \end{cases}$$

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$$

We find for the likelihood function

$$\text{lik}(p) = f(X_1, \ldots, X_n | p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{S_n}(1-p)^{n-S_n},$$

where $S_n := \sum_{i=1}^{n} X_i$.

Maximum likelihood estimators.

Example (Proportion of defectives)

likelihood function

$$\text{lik}(p) = p^{S_n}(1-p)^{n-S_n}$$

The log likelihood function is

$$l(p) = \log \text{lik}(p) = S_n \log p + (n - S_n) \log(1-p),$$

and its derivative

$$\frac{d}{dp} l(p) = \frac{S_n}{p} - \frac{n - S_n}{1-p}$$

has root

$$\hat{p} = \frac{1}{n} S_n = \bar{X},$$

the MLE for p.

[Ex: Show that S_n/n is also the method of moments estimator for p.]

Maximum likelihood estimators.

Suppose that $\hat{\theta}_n$ is an estimator of θ based on a sample of size n. The sequence $(\hat{\theta}_n)$ of estimators is called *consistent*, if we have convergence

$$\theta_n \to \theta$$

(we are not specific about mode of convergence here: could be in probability, or distribution).

Example (Proportion of defectives)

Law of Large numbers implies

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbb{E}[X_1] = p$$

as $n \to \infty$, thus $\hat{p} = \hat{p}_n$ (really the sequence $(\hat{p}_n)$) is a consistent estimator for p.

Maximum likelihood estimators.

Example (MLE of normal distribution)

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma).^2$$

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2}$$

To find the maximum of the likelihood function it is often convenient to maximize the *log likelihood function*

$$\begin{aligned} l(\mu, \sigma) &:= \log \mathrm{lik}(\mu, \sigma) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \\ &= n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \end{aligned}$$

---

[2] Here we have $\theta = (\mu, \sigma)$.

Maximum likelihood estimators.

Example (MLE of normal distribution)

$$l(\mu, \sigma) = n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

Solving for roots of partial derivatives we obtain

$$0 = \frac{\partial}{\partial \mu} l(\mu, \sigma) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) \rightsquigarrow \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}}$$

$$0 = \frac{\partial}{\partial \sigma} l(\mu, \sigma) = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (X_i - \mu)^2 \rightsquigarrow \boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

as MLEs for μ and σ. Know sampling distribution of $\bar{X}$ if σ^2 is known.

Bayesian estimators

Conditional densities.

Example (Bayesian inference)

Often in medical problems it is assumed that a drug is effective with some (unknown) probability Π in each treatment, independently across treatments.

One challenge is to estimate ("learn") effectiveness Π of a drug from the results of n treatments.

If we have no prior knowledge about Π, seems reasonable to assume it is a random number distributed uniformly on $[0, 1]$. [In Bayesian jargon, this is called the *uninformative prior.*]

$$X := \text{\# effective treatments}^{[3]}$$

Goal: "Update" distribution of Π given the observed number of effective treatments X.

[3] What is your guess for the distribution of X?

Conditional densities.

Example (Bayesian inference)

More formally, we want to find conditional distribution

$$f_\Pi(p|X = x)$$

of effectiveness given $X = x$ effective treatments.

Maybe we can find joint density of (Π, X) first?
From the setup of the experiment

$$f_X(x|\Pi = p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (0 \le x \le n)$$

hence, joint density of (Π, X) is

$$f(p, x) = f_X(x|\Pi = p) f_\Pi(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

for $0 \le p \le 1, 0 \le x \le n$, since $\Pi \sim \mathrm{U}(0,1)$.

Conditional densities.

Example (Bayesian inference)

For $a, b > 0$ the distribution on $(0,1)$ with density

$$\boxed{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1}$$

is called the *beta(a, b) distribution*. Notice that this implies

$$\int_0^1 x^{a-1}(1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

$$\begin{aligned} f_X(x) &= \int_0^1 f(p,x)dp = \binom{n}{x} \int_0^1 p^x (1-p)^{n-x} dp \\ &= \frac{n!}{x!(n-x)!} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{1}{n+1}, \end{aligned}$$

i.e. the number of effective treatments has uniform distribution.

Conditional densities.

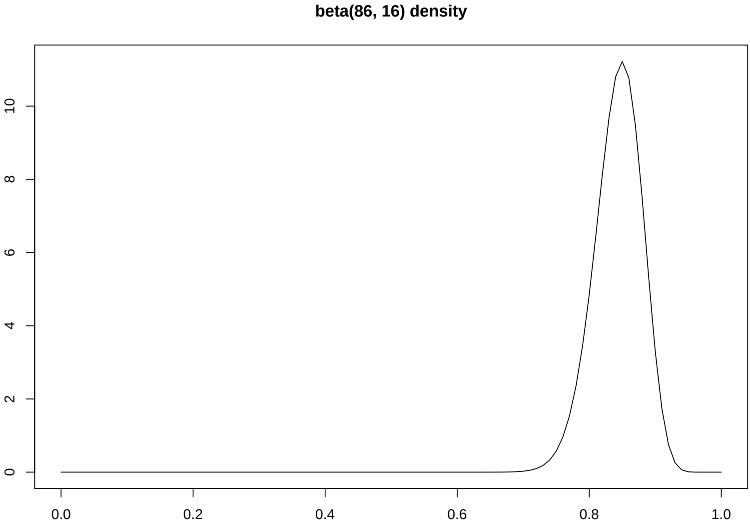
Example (Bayesian inference)

For the density of the "updated effectiveness" we obtain

$$f_\Pi(p|X = x) = \frac{f(p, x)}{f_X(x)} = (n+1)\binom{n}{x} p^x (1-p)^{n-x}$$
$$= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^x (1-p)^{n-x},$$

the density of a beta$(x+1, n-x+1)$ distribution.

$X = 85$ out of $n = 100$ treatments are found to be effective.
Given this observation, we update our prior distribution to $(\Pi|X = 85) \sim$ beta$(86, 16)$, the so-called *posterior distribution*.


beta(86, 16) density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0

Conditional densities.

Example (Bayesian inference)

Posterior distribution:

$$(\Pi|X = x) \sim \mathsf{beta}(x + 1, n - x + 1)$$

A natural estimator for the effectiveness Π given that out of n treatments $X = x$ where effective, is the mean of the posterior distribution[4]

$$\begin{aligned}\mathbb{E}[\Pi|X = x] &= \int_0^1 t \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x} dt \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)} = \frac{x+1}{n+2}\end{aligned}$$

This is called the *posterior mean* for Π.

[4] Or the mode.

More examples

Parameter estimation

Following example appears in different disguises in science, medicine, engineering, etc. We consider settings where observations can be assigned to different categories (categorial data).

Example (Emergency room)

Emergency room of large hospital assigns patients to one of three categories:

1. Stable. No immediate treatment required.
2. Serious. Immediate treatment not required, but patient needs to be monitored until physician available.
3. Critical. Patient's life endangered without immediate treatment.

Parameter estimation

Example (Emergency room)

Hospital records over past week show that

- 300 patients were classified as stable,
- 180 patients classified serious,
- 120 patients classified critical,

To ensure optimal organization, administration needs to estimate long-run frequency p_i of patients that are classified in category i.

Find estimators for p_1, p_2, p_3.
(Make a guess before we proceed formally!)

Parameter estimation

Review: multinomial distribution. Think of n different marbles that we paint in c different colors. We paint marbles independently one after the other, with

$$\mathbb{P}\left\{\text{a particular marble is painted in color } i\right\} = p_i$$

($\sum_i p_i = 1$, $p_i \geq 0$). Let

$$X_i := \text{\# marbles of color } i.$$

Then[5]

$$\boxed{\mathbb{P}\left\{X_1 = x_1, \ldots, X_c = x_c\right\} = \binom{n}{x_1, \ldots, x_n} \prod_{i=1}^{c} p_i^{x_i}}$$

if $\sum_i x_i = n$ and $= 0$ otherwise. The vector $(X_1, \ldots, X_n)$ has a *multinomial distribution* with parameters $(n, p_1, \ldots, p_c)$ denoted

$$(X_1, \ldots, X_c) \sim \text{multinomial}(n, p_1, \ldots, p_c).$$

[5] $\binom{n}{x_1,\ldots,x_c} = n!/(x_1!x_2!\cdots x_c!)$ is the multinomial coefficient.

Parameter estimation

Let us work out the MLE for $p_1, p_2, \ldots, p_c$.
Notice that $(X_1, \ldots, X_c)$ are not i.i.d.

Want to maximize log likelihood

$$
\begin{aligned}
l(p_1, \ldots, p_c) &= \log f(x_1, \ldots, x_c | p_1, \ldots, p_c) \\
&= \log \binom{n}{x_1, \ldots, x_c} \prod_{i=1}^{c} p_i^{x_i} \\
&= \log n! - \sum_{i=1}^{c} \log x_i! + \sum_{i=1}^{c} x_i \log p_i
\end{aligned}
$$

subject to $p_1 + p_2 + \cdots + p_c = 1$.

Maximize instead

$$
L(p_1, \ldots, p_c; \lambda) = \log n! - \sum_{i=1}^{c} \log x_i! + \sum_{i=1}^{c} x_i \log p_i + \lambda \left(\sum_{i=1}^{c} p_i - 1 \right).
$$

Parameter estimation

Maximize instead

$$L(p_1, \ldots, p_c; \lambda) = \log n! - \sum_{i=1}^{c} \log x_i! + \sum_{i=1}^{c} x_i \log p_i + \lambda \left(\sum_{i=1}^{c} p_i - 1 \right).$$

For any $j = 1, \ldots, c$ the partial derivative

$$\frac{d}{dp_j} L = \frac{x_j}{p_j} + \lambda$$

has root

$$\hat{p}_j = -\frac{x_j}{\lambda},$$

and summing both sides w.r.t. j we obtain

$$1 = -\frac{1}{\lambda} \sum_{j=1}^{c} x_j = -\frac{n}{\lambda} \text{ hence } \lambda = -n.$$

where we used the constraint $\sum_j \hat{p}_j = 1$. MLE $\boxed{\hat{p}_j = \frac{x_j}{n}.}$

What can we say about the sampling distribution of $\hat{p}_j$?

Parameter estimation

Example (Emergency room)

Recall hospital records:

- 300 patients were classified as stable,
- 180 patients classified serious,
- 120 patients classified critical.

Find MLEs

$$\hat{p}_1 = \frac{300}{600} = 50\% \qquad \hat{p}_2 = \frac{180}{600} = 30\% \qquad \hat{p}_3 = \frac{120}{600} = 20\%.$$

Ex: Work out the method of moments estimator for p_i.