Likelihoods

History of the course:

Taught since 2017 or 2018 by

Haonan
Dan
} thanks for help!

Now: me

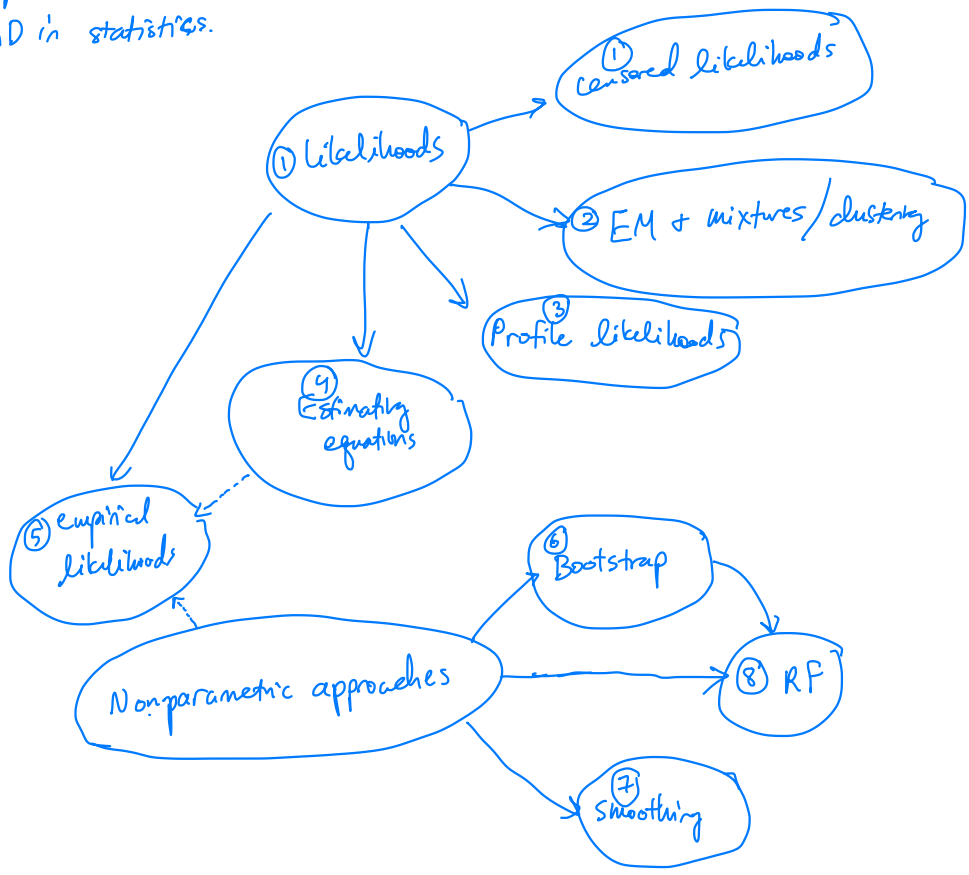
Idea: Many important topics not taught in other courses but should be "core" to a PhD in statistics.

0.1 Outline

1. Likelihoods
2. EM, k-means
3. Profile likelihood
4. Estimating equations
5. Empirical likelihood

midterm →

6. Bootstrap
7. Smoothing methods
8. Random forests.



1 Likelihood Construction and Estimation

Likelihood based methods:

MLE
LRT
likelihood based uncertainty (CIs).

Why do Statisticians love likelihood-based estimation?

1. Invariance property of the MLE: If a distribution is parametrized by θ, but interested in $\tau(\theta)$, if $\hat{\theta}$ is the MLE of θ, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

2. MLE's asymptotically unbiased; consistent

   $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) \to 0.$

3. MLE's asymptotically efficient; variance achieve Cramer-Rao lower bound

   $\sqrt{n}\left(\tau(\hat{\theta}) - \tau(\theta)\right) \to N(0, v(\theta))$ where $v(\theta)$ is CRLB.

4. Relationship w/ Fisher information matrix allows for construction of CI's (based on asymptotic properties).

Downsides?

1. Very model-based! You are assuming you know the entire distribution.

2. It often requires numerical optimization.

Still... we ♡ it!

1.1 Introduction

doesn't say iid.

Definition: Suppose random variables $\boldsymbol{Y} = (Y_1, \ldots Y_n)^\top$ has joint density or probability mass function $f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)$. Then the *likelihood function* is

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = f_{\boldsymbol{Y}}(\boldsymbol{Y}, \boldsymbol{\theta}).$$

← in general likelihood = joint

↑ likelihood is random! (because it depends on the data!)

This is somewhat obvious because we know MLE is random & we quantify its uncertainty.

Given a vector of observations $\underline{y}$, the likelihood is a function of $\underline{\theta}$.

for any valid value of $\underline{\theta}$, it returns a number (the likelihood).

$\hat{\underline{\theta}}_{MLE}$ is obtained by finding the value of $\underline{\theta}$ which yields the max likelihood value.

Key concept: In all situations, the likelihood is the joint density of the observed data to be analyzed.

Comments

(1) "density" can mean continuous pdf or pmf.

(2) "observed" data will be generalized, e.g. censored data.

1.1.1 Notation

Given $\boldsymbol{y}$, note that $L(\boldsymbol{\theta}|\boldsymbol{y}) : \mathbb{R}^b \to \mathbb{R}$.

$\underline{\theta} : (\theta_1, \ldots, \theta_b)^T$

likelihood $L(\underline{\theta}|y)$ is scalar valued.

Generally, we optimize $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|\boldsymbol{y})$.

monotone increasing

$\Rightarrow \operatorname{argmax}_{\underline{\theta}} L(\underline{\theta}|y) = \operatorname{argmax}_{\underline{\theta}} \ell(\underline{\theta})$.

How? Take derivatives, set to zero.

Generally convention is the derivative of function (like $\ell(\underline{\theta})$) wrt a vector (like $\underline{\theta}$) is a row vector $\ell'(\underline{\theta}) = \frac{\partial \ell(\underline{\theta})}{\partial \underline{\theta}} = \left(\frac{\partial \ell(\underline{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\underline{\theta})}{\partial \theta_b} \right)$.

Define score function $S(\underline{\theta}) = \ell'(\underline{\theta})^T$

$$= \begin{pmatrix} \frac{\partial \ell(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\underline{\theta})}{\partial \theta_b} \end{pmatrix} \quad (b \times 1)$$

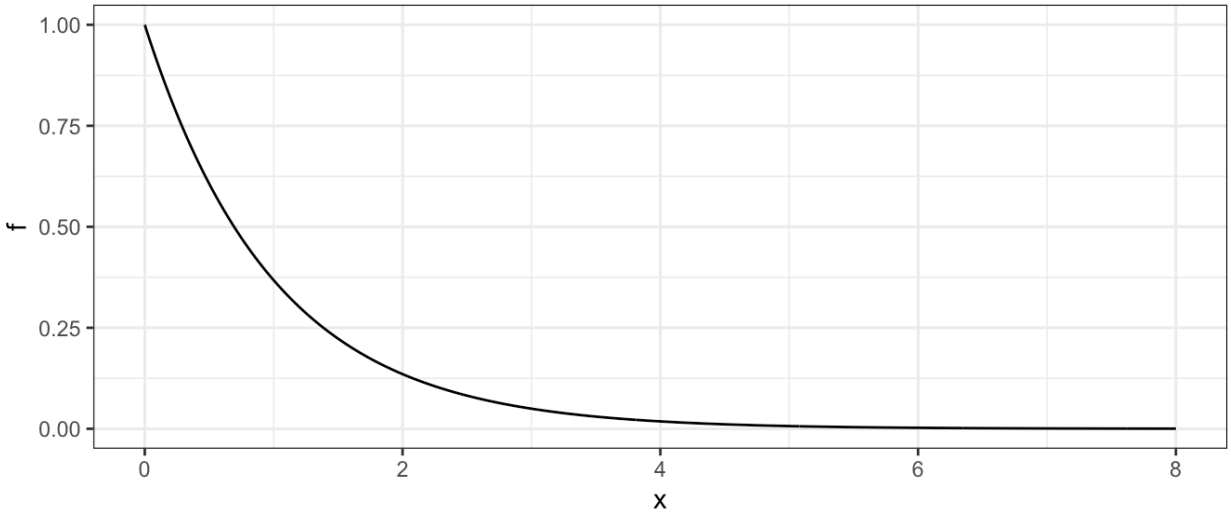
← column vector

Example: Suppose we have $Y_1, \dots Y_n \overset{iid}{\sim} \text{Exp}(\lambda)$. The likelihood function is defined as

$$L(\lambda|y) = f_y(y;\lambda)$$
$$= \prod_{i=1}^{n} f_y(y_i;\lambda)$$
$$= \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$
$$\Rightarrow \ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

```
# likelihood simulation
n <- 10
lambda <- 1

# plot of exponential(lambda) density
data.frame(x = seq(0, 8, .01)) |>
  mutate(f = dexp(x, rate = lambda)) |>
  ggplot() +
  geom_line(aes(x, f))
```



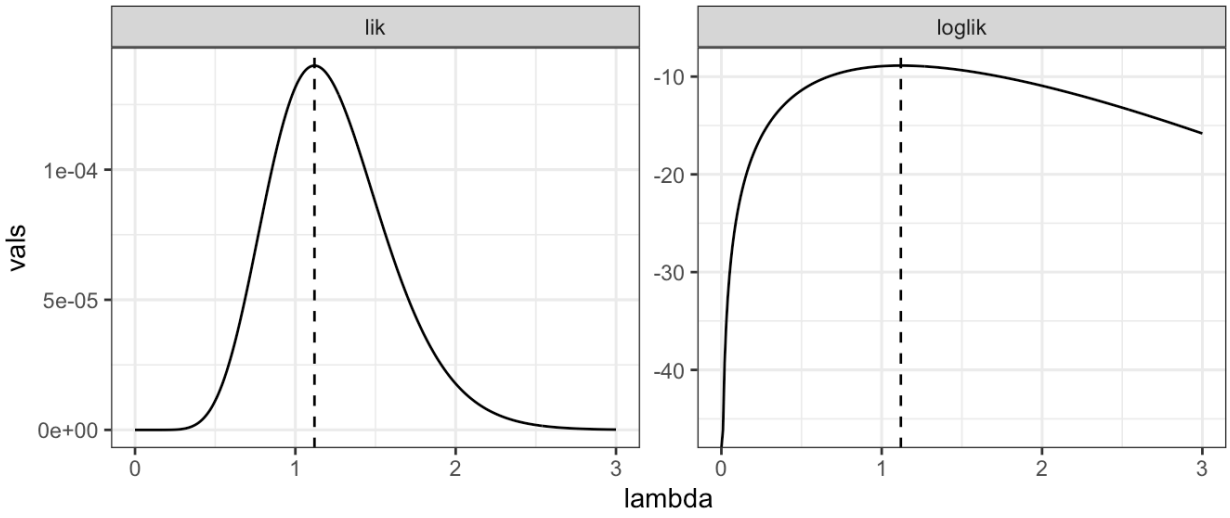
```r
# define likelihood
loglik <- function(lambda, data)
{
    lik <- prod(dexp(data, rate = lambda))
    loglik <- sum(dexp(data, rate = lambda, log = T))

    out <- data.frame(lik = lik, loglik = loglik)
    return(out)
}

# simulate data
data <- rexp(n = n, rate = lambda)

# plot likelihood and loglikelihood
data.frame(lambda = seq(0, 3, by = .01)) |>
  rowwise() |>
  mutate(loglik = loglik(lambda, data)) |>
  unnest(cols = c(loglik)) |>
  pivot_longer(-lambda, names_to = "func", values_to =
"vals") |>
  ggplot() +
  geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) + #
max likelihood estimate is 1/mean
  geom_line(aes(lambda, vals)) +
  facet_wrap(~func, scales = "free")
```

now we have realized data!



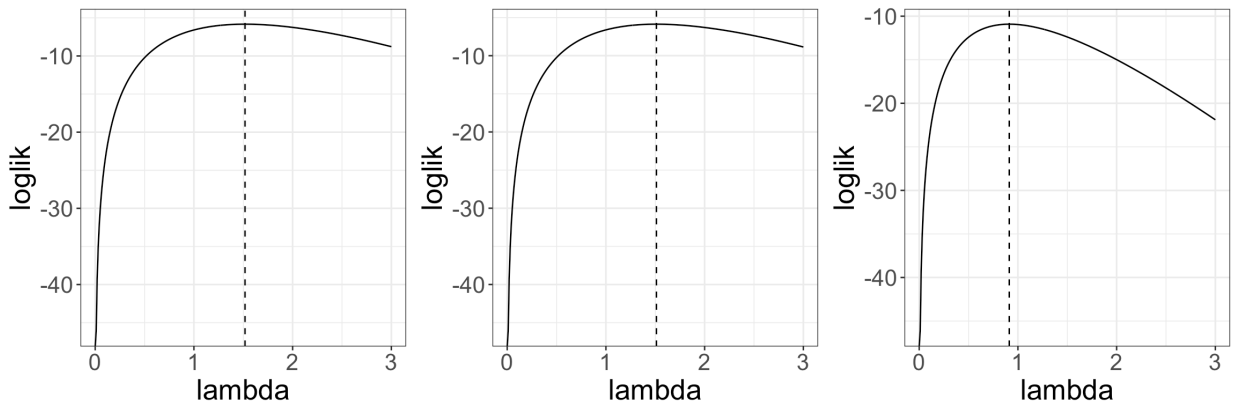
max happens at same place!

The likelihood function is random!

```
          for(i in seq_len(3)) {
            # simulate data
            data <- rexp(n = n, rate = lambda)

            # plot likelihood and loglikelihood
            data.frame(lambda = seq(0, 3, by = .01)) |>
              rowwise() |>
              mutate(loglik = loglik(lambda, data)) |>
              unnest(cols = c(loglik)) |>
              ggplot() +
              geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) +
# max likelihood estimate is 1/mean
              geom_line(aes(lambda, loglik)) +
              theme(text = element_text(size = 20)) -> p ## make
legible in notes

              print(p)
          }
```



Question : What does the likelihood integrate to?

NOT 1!

↳ function of θ, not x!

Your Turn: What is the effect of sample size on the log-likelihood function? Make a plot showing the log-likelihood function that results from $n = 10$ vs. $n = 100$ with corresponding MLE.

1.2 Construction

The use of the likelihood function in parameter estimation is easiest to understand in the case of discrete iid random variables.

1.2.1 Discrete IID Random Variables

Suppose each of the n random variables in the sample $Y_1, \ldots, Y_n$ have probability mass function $f(y; \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y_1 = y), y = y_1, y_2, \ldots$. The likelihood is then defined as:

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \text{ joint density of observed random variables}$$

In other words,

Example (Fetal Lamb Movements): Data on counts of movements in five-second intervals of one fetal lamb ($n = 240$ intervals:)

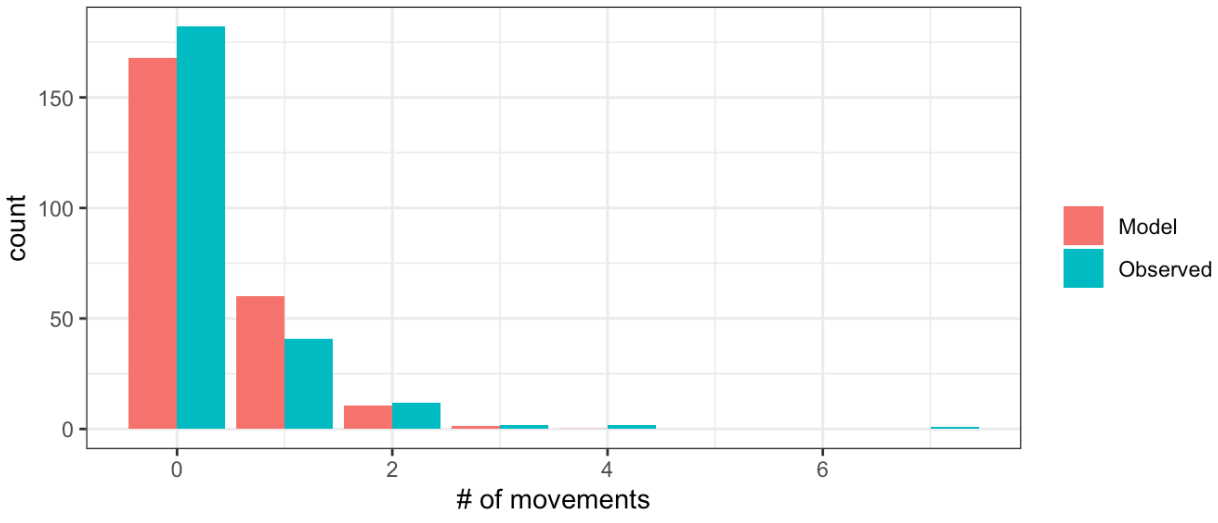
No. of Movements	0	1	2	3	4	5	6	7
Count	182	41	12	2	2	0	0	1

Assume a Poisson model: $P(Y = y) = f_Y(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$. Then the likelihood is

Equating the derivative of the loglikelihood with respect to λ to zero and solving results in the MLE

$$\hat{\lambda}_{\text{MLE}} =$$

This is the best we can do with this model. But is it good?



1.2.2 Multinomial Likelihoods

The multinomial distribution is a generalization of the binomial distribution where instead of 2 outcomes (success or failure), there are now $k \geq 2$ outcomes.

The probability mass function is

For $N_1, \ldots, N_k, N_i =$ the number of balls in i^{th} urn,

The maximum likelihood estimator of p_i:

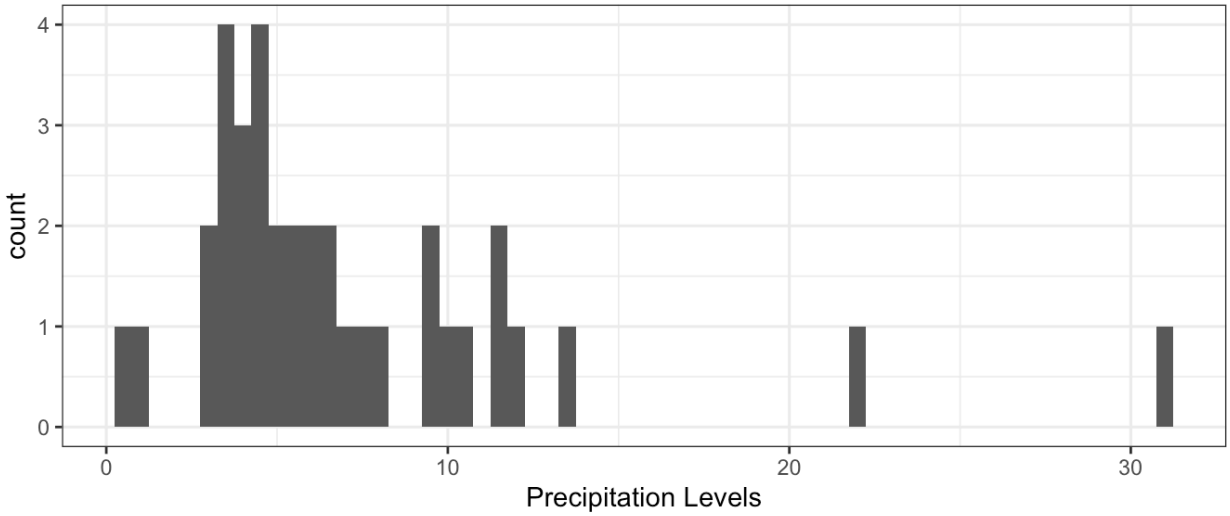
More interesting multinomial likelihoods arise when the p_i are modeled as a function of a lesser number of parameters $\theta_1, \ldots, \theta_m$, $m < k - 1$.

Example (Capture-Recapture): To estimate fish survival during a specific length of time (e.g., one month), a common approach is to use a removal design.

1.2.3 Continuous IID Random Variables

Recall: the likelihood is the joint density of data to be analyzed.

Example (Hurricane Data): For 36 hurricanes that had moved far inland on the East Coast of the US in 1900-1969, maximum 24-hour precipitation levels during the time they were over mountains.



We model the precipitation levels with a gamma distribution, which has density

$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1}\exp(-y/\beta), \quad y>0, \alpha,\beta>0.$$

This leads to the likelihood

Of course, this cannot be interpreted as a probability because

To get a probability, need to go from a density to a measure.

But it may be useful to think of the value of the likelihood as being proportional to a probability.

More formally, begin with the definition of a derivative

$$g'(x) = \lim_{h \to 0^+} \frac{g(x+h) - g(x-h)}{2h}.$$

Let F be the cumulative distribution function of a continuous random variable Y, then (if the derivative exists)

$$f(y) = \lim_{h \to 0^+} \frac{F(y+h) - F(y-h)}{2h} =$$

If we substitute this definition of a density into the definition of the likelihood

Compare this to the iid discrete case:

Example (Hurricane Data, Cont'd): Recall with a gamma model, the likelihood for this example is

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \{\Gamma(\alpha)\}^{-n}\beta^{-n\alpha}\Big\{\prod Y_i\Big\}^{\alpha-1}\exp\Big(-\sum y_i/\beta\Big),$$

and log-likelihood

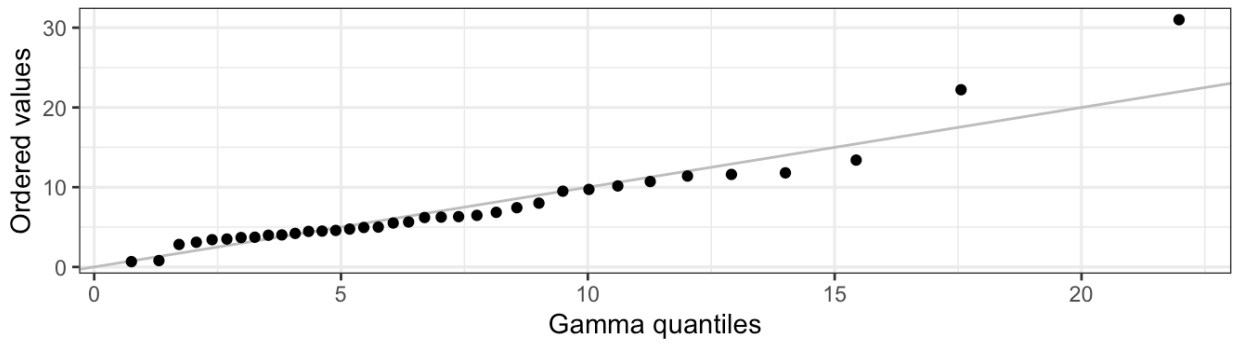
$$\ell(\boldsymbol{\theta}) =$$

```
        ## loglikelihood function
        neg_gamma_loglik <- function(theta, data) {
          -sum(log(dgamma(data, theta[1], scale = theta[2])))
        }

        ## maximize
        mle <- nlm(neg_gamma_loglik, c(1.59, 4.458), data =
hurr_rain)
        mle$estimate
```

```
## [1] 2.187214 3.331862
```

```
        ## Gamma QQ plot
        data.frame(theoretical = qgamma(ppoints(hurr_rain),
mle$estimate[1], scale = mle$estimate[2]),
                   actual = sort(hurr_rain)) |>
          ggplot() +
          geom_abline(aes(intercept = 0, slope = 1), colour =
"grey") +
          geom_point(aes(theoretical, actual)) +
          xlab("Gamma quantiles") + ylab("Ordered values")
```



1.2.4 Mixtures of Discrete and Continuous RVs

Some data Y often have a number of zeros and the amounts greater than zero are best modeled by a continuous distribution.

Ex:

In other words, they have positive probability of taking a value of exactly zero, but continuous distribution otherwise.

A sensible model would assume Y_i are iid with cdf

$$F_Y(y; p, \boldsymbol{\theta}) = \begin{cases} 0 & y = 0 \\ p & y = 0 \\ p + (1-p)F_T(y; \boldsymbol{\theta}) & y > 0 \end{cases}$$

where $0 < p \leq 1$ is $P(Y = 0)$ and $F_T(y; \boldsymbol{\theta})$ is a distribution function for a continuous positive random variable.

Another way to write this:

How to go from here to get a likelihood?

One approach: let n_0 be the number of zeroes in the data and $m = n - n_0$ be the number of non-zero Y_i. This leads to an intuitive way to construct the likelihood for iid $Y_1, \ldots, Y_n$ distributed according to the above distribution:

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \lim_{h \to 0^+} \left(\frac{1}{2h} \right)^m \prod_{i=1}^{n} \{F_Y(Y_i + h; p, \boldsymbol{\theta}) - F_Y(Y_i - h; p, \boldsymbol{\theta})\}$$

Feels a little arbitrary in how we are defining different weights on our likelihood for discrete and continuous parts.

Turns out, it doesn't matter! (Need some STAT 630/720 to see why.)

Definition (Absolute Continuity) On $(\mathbb{X}, \mathcal{M})$, a finitely additive set function ϕ is *absolutely continuous* with respect to a measure μ if $\phi(A) = 0$ for each $A \in \mathcal{M}$ with $\mu(A) = 0$. We also say ϕ is *dominated* by μ and write $\phi \ll \mu$. If ν and μ are measures such that $\nu \ll \mu$ and $\mu \ll nu$ then μ and ν are *equivalent*.

Theorem (Lebesgue-Randon-Nikodym) Assume that ϕ is a σ-finite countably additive set function and μ is a σ-finite measure. There exist unique σ-finite countably additive set functions ϕ_s and ϕ_{ac} such that $\phi = \phi_{ac} + \phi_s, \phi_{ac} \ll \mu$, ϕ_s and μ are mutually singular and there exists a measurable extended real valued function f such that

$$\phi_{ac}(A) = \int_A f d\mu, \qquad \text{for all } A \in \mathcal{M}.$$

If g is another such function, then $f = g$ a.e. wrt μ. If $\phi \ll \mu$ then $\phi(A) = \int_A f d\mu$ for all $A \in \mathcal{M}$.

Definition (Radon-Nikodym Derivative) $\phi = \phi_{ac} + \phi_s$ is called the *Lebesgue decomposition*. If $\phi \ll \mu$, then the density function f is called the *Radon-Nikodym derivative* of ϕ wrt μ.

So what?

1.2.5 Proportional Likelihoods

Likelihoods are equivalent for point estimation as long as they are proportional and the constant of proportionality does not depend on unknown parameters.

Why?

Consider if $Y_i, i = 1, \ldots, n$ are iid continuous with density $f_Y(y; \boldsymbol{\theta})$ and $X_i = g(Y_i)$ where g is increasing and continuously differentiable. Because g is one-to-one, we can construct Y_i from X_i and vice versa.

More formally, the density of X_i is $f_X(x; \boldsymbol{\theta}) = f_Y(h(x); \boldsymbol{\theta}) h'(x)$, where $h = g^{-1}$, and

$$L(\boldsymbol{\theta}|\boldsymbol{X}) =$$

Example (Likelihood Principle): Consider data from two different sampling plans:

1. A binomial experiment with $n = 12$. Let $Y_i = 1$ if i^{th} trial is a success and 0 otherwise.

$$L_1(p|\boldsymbol{Y}) = \binom{12}{S} p^S (1-p)^{12-S}, \text{ where } S = \sum_{i=1}^{n} Y_i$$

2. A negative binomial experiment, i.e. run the experiment until three zeroes are obtained.

$$L_2(p|\boldsymbol{Y}) = \binom{S+2}{S} p^S (1-p)^3.$$

The ratio of these likelihoods is

$$\frac{L_1(p|\boldsymbol{Y})}{L_2(p|\boldsymbol{Y})} =$$

Suppose $S = 9$. Is all inference equivalent for these likelihoods? Debatable.

The likelihood principle states all the information about $\boldsymbol{\theta}$ from an experiment is contained in the actual observation $\boldsymbol{y}$. Two likelihood functions for $\boldsymbol{\theta}$ (from the same or different experiments) contain the same information about $\boldsymbol{\theta}$ if they are proportional.

1.2.6 Empirical Distribution Function as MLE

Recall the empirical cdf:

Suppose $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ are the order statistics of an iid sample from an unknown distribution function F_Y. Our goal is to estimate F_Y.

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y \geq y_{(i)})$$

Is this a “good” estimator of F_Y?

Yes, because it's MLE.

Suppose $Y_1, \ldots, Y_n$ are iid with distribution function $F(y)$. Here $F(y)$ is the unknown parameter.

An approximate likelihood for F is

$$L_h(F|\boldsymbol{Y}) = \prod_{i=1}^{n} \{F(Y_i + h) - F(Y_i - h)\}$$

1.2.7 Censored Data

Censored data occur when the value is only partially known. This is different from *truncation*, in which the data does not include any values below (or above) a certain limit.

For example, we might sample only hourseholds that have an income above a limit, L_0. If all incomes have distribution $F(x;\boldsymbol{\theta})$, then for $y > L_0$,

$P(Y_1 \leq y | Y_1 > L_0) =$

The likelihood is then

1.2.7.1 Type I Censoring

Suppose a random variable X is normally distributed with mean μ and variance σ^2, but whenever $X \leq 0$, all we observe is that it is less than or equal to 0. If the sample is set to 0 in the censored cases, then define

$$Y = \begin{cases} 0 & \text{if } X \leq 0 \\ X & \text{if } X > 0. \end{cases}$$

The distribution function of Y is

Suppose we have a sample $Y_1, \ldots, Y_n$ and let n_0 be the number of sample values that are 0. Then $m = n - n_0$ and

We might have censoring on the left at L_0 and censoring on the right at R_0, but observe all values of X between L_0 and R_0. Suppose X has density $f(x; \boldsymbol{\theta})$ and distribution function $F(x; \boldsymbol{\theta})$ and

$$Y_i = \begin{cases} L_0 & \text{if } X_i \leq L_0 \\ X_i & \text{if } L_0 < X_i < R_0 \\ R_0 & \text{if } X_i \geq R_0 \end{cases}$$

If we let n_L and n_R be the number of X_i values $\leq L_0$ and $\geq R_0$ then the likelihood of the observed data $Y_1, \ldots, Y_n$ is

We could also let each X_i be subject to its own censoring values L_i and R_i. For the special case of right censoring, define $Y_i = \min(X_i, R_i)$. In addition, define $\delta_i = \mathbb{I}(X_i \leq R_i)$. Then the likelihood can be written as

Example (Equipment failure times): Pieces of equipment are regularly checked for failure (but started at different times). By a fixed date (when the study ended), three of the items had not failed and therefore were censored.

y	2	72	51	50	33	27	14	24	4	21
delta	1	0	1	0	1	1	1	1	1	0

Suppose failure times follow an exponential distribution $F(x;\sigma) = 1 - \exp(-x/\sigma), x \geq 0$. Then

$$L(\sigma|\boldsymbol{Y}) =$$

1.2.7.2 Random Censoring

So far we have considered censoring times to be fixed. This is not required.

This leads to random censoring times, e.g. R_i, where we assume that the censoring times are independent of $X_1, \ldots, X_n$ and iid with distribution function $G(t)$ nd density $g(t)$.

Let's consider the contributions to the likelihood:

which results in

$$L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\delta}) =$$

1.3 Likelihoods for Regression Models

We will start with linear regression and then talk about more general models.

1.3.1 Linear Model

Consider the familiar linear model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n,$$

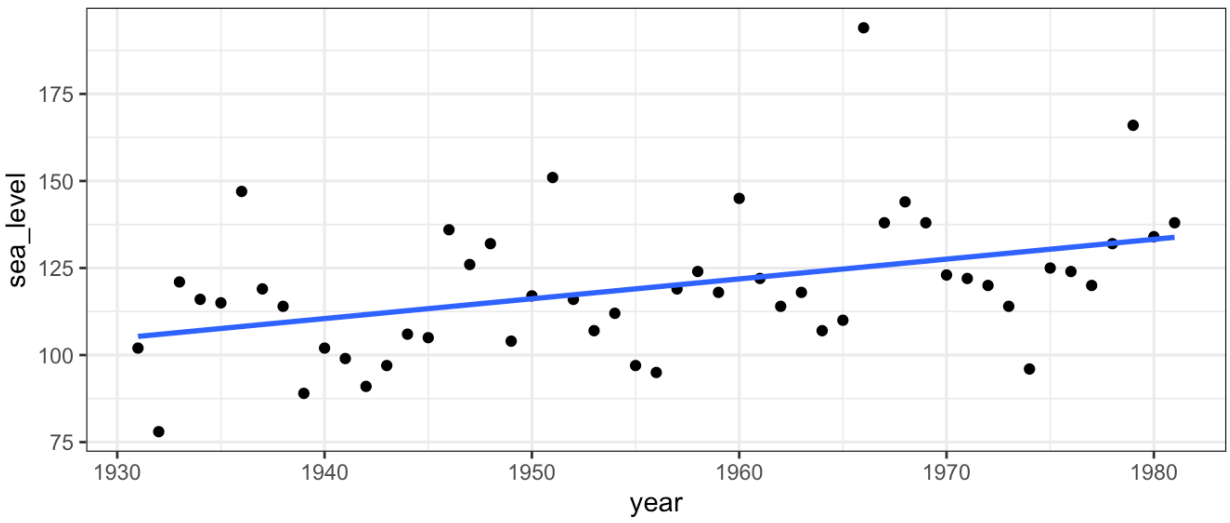
where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are known nonrandom vectors.

For likelihood-based estimation,

$$L(\boldsymbol{\beta}, \sigma | \{Y_i, \boldsymbol{x}_i\}_{i=1}^n) =$$

What do you do when ϵ_i are not Gaussian?

Example (Venice sea levels): The annual maximum sea levels in Venice for 1931–1981 are :



1.3.2 Additive Errors Nonlinear Model

1.3.3 Generalized Linear Models

Imagine an experiment where individual mosquitos are given some dosage of pesticide. The response is whether the mosquito lives or dies. The data might look something like:

Goal: Model the relationship between the predictor and response.

Question: What would a curve of best fit look like?

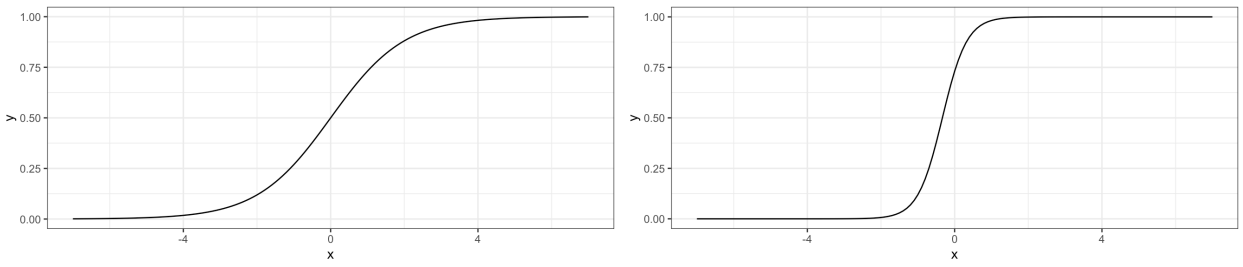
Refined Goal:

Let's build a sensible model.

Step 1: Find a function that behaves the way we want.

```
# understanding the logistic function
# first, theta just equals x
x <- seq(-7, 7, .1)
theta <- x
y <- exp(theta)/(1 + exp(theta))
ggplot() + geom_line(aes(x, y))

# now, let theta be a linear function of x
theta <- 1 + 3*x
y <- exp(theta)/(1 + exp(theta))
ggplot() + geom_line(aes(x, y))
```



Step 2: Build a stochastic mechanism to relate to a binary response.

Step 3: Put Step 1 and Step 2 together.

Fitting our model: Does OLS make sense?

Consider the likelihood contribution.

$$L_i(p_i|Y_i) =$$

So the log-likelihood contribution is

$$\ell_i(p_i) =$$

Recall, we said $p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ was sensible.

Which gives us,

$$\ell_i(\theta_i) =$$

So the log-likelihood is

$$\ell(\theta_1, \ldots, \theta_n) =$$

To optimize?

```
        ## data on credit default
        data("Default", package = "ISLR")
        head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

```
        ## fit model with ML
        m0 <- glm(default ~ balance, data = Default, family =
binomial)
        tidy(m0) |> kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-10.6513306	0.3611574	-29.49221	0
balance	0.0054989	0.0002204	24.95309	0

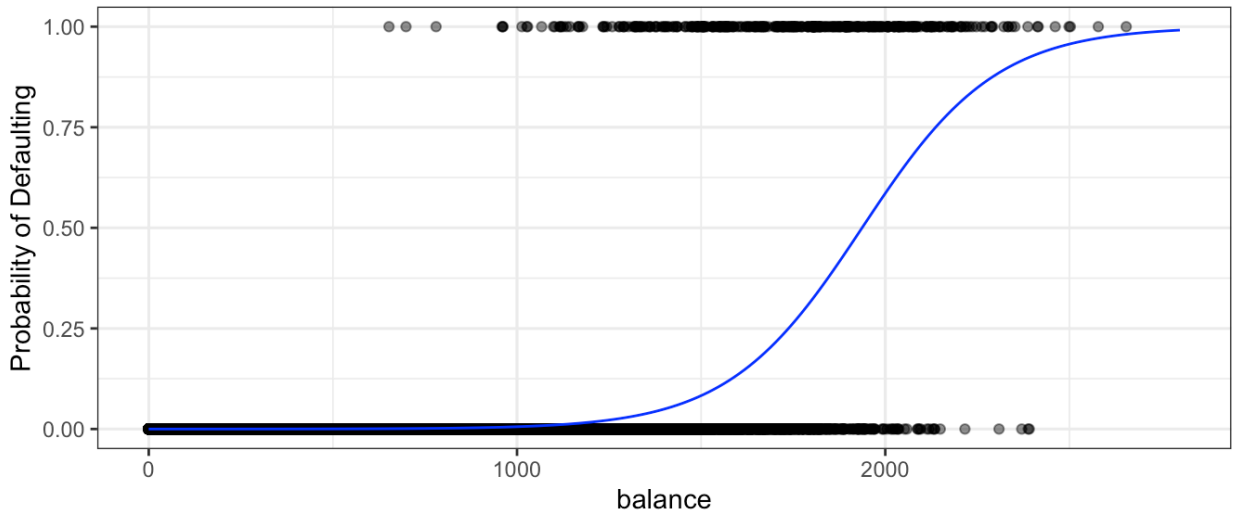
```
        glance(m0) |> kable()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
2920.65	9999	-798.2258	1600.452	1614.872	1596.452	9998	10000

```
        ## plot the curve
        x_new <- seq(0, 2800, length.out = 200)
        theta <- m0$coefficients[1] + m0$coefficients[2]*x_new
        p_hat <- exp(theta)/(1 + exp(theta))

        ggplot() +
          geom_point(aes(balance, as.numeric(default) - 1), alpha =
0.5, data = Default) +
```

```
          geom_line(aes(x_new, p_hat), colour = "blue") +
ylab("Probability of Defaulting")
```



In general, a GLM is three pieces:

1. The random component

2. The systemic component

3. A linear predictor

Remarks:

Example (Poisson regression):

Consider a general family of distributions:

$$\log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

Example (Normal model):

We can learn something about this distribution by considering it's mean and variance. Because we don't have an explicit form of the density, we rely on two facts:

1. $\mathrm{E}\left[\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial \theta_i}\right] = 0.$

2. $\mathrm{E}\left[\frac{\partial^2 \log f(Y_i;\theta_i,\phi)}{\partial \theta_i^2}\right] + \mathrm{E}\left[\left(\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial \theta_i}\right)^2\right] = 0.$

For $\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)$,

Example (Bernoulli model):

$$f(y_i; p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

Finally, back to modelling. Our **goal** is to build a relationship between the mean of Y_i and covariates $\boldsymbol{x}_i$.

Example (Bernoulli model, cont'd):

1.4 Marginal and Conditional Likelihoods

Consider a model which has $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ are the parameters of interest and $\boldsymbol{\theta}_2$ are nuisance parameters.

One way to improve estimation for $\boldsymbol{\theta}_1$ is to find a one-to-one transformation of the data $\boldsymbol{Y}$ to $(\boldsymbol{V}, \boldsymbol{W})$ such that either

The key feature is that one component of each contains only the parameter of interest.

Example (Neyman-Scott problem): Let $Y_{ij}, i = 1, \ldots, n, j = 1, 2$ be independent normal random variables with possible different means μ_i but the same variance σ^2.

Our goal is to estimate σ^2. Should we be able to?

Following the usual arguments,

$$\hat{\mu}_{i,\mathrm{MLE}} = \frac{Y_{i1} + Y_{i2}}{2}$$

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \hat{\mu}_{i,\mathrm{MLE}})^2$$

$$\mathrm{E}[\hat{\sigma}^2_{\mathrm{MLE}}] =$$

A reworking of the data seems more promising. Let,

$$V_i = \frac{Y_{i1} - Y_{i2}}{\sqrt{2}} \qquad \text{and} \qquad W_i = \frac{Y_{i1} + Y_{i2}}{\sqrt{2}}$$

For conditional likelihoods, we can often exploit the existence of sufficient statistics for the nuisance parameters under the assumption that the parameter of interest is known.

Example (Exponential Families): The structure of exponential families is such that it is often possible to exploit their properties to eliminated nuisance parameters. Let Y have a density of the form

$$f(y;\boldsymbol{\eta}) = h(y)\exp\left\{\sum_{i=1}^{s}\eta_i T_i(y) - A(\boldsymbol{\eta})\right\},$$

then

Thus, exponential families often provide an automatic procedure for finding $\boldsymbol{W}$ and $\boldsymbol{V}$.

Example (Logistic Regression): For binary Y_i, the standard logistics regression model is

$$P(Y_i = 1) = p_i(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

and the likelihood is

$$L(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) =$$

1.5 The Maximum Likelihood Estimator and the Information Matrix

We have now talked about how to construct likelihoods in a variety of settings, now we can use those constructions to formalize how we make inferences about model parameters.

Recall the score function

$$S(\boldsymbol{Y}, \boldsymbol{\theta}) =$$

Generally, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is the value of $\boldsymbol{\theta}$ where the maximum (over the parameter space Θ) of $L(\boldsymbol{\theta}|\boldsymbol{Y})$ is attained.

Under the assumption that the log-likelihood is continuously differentiable, then

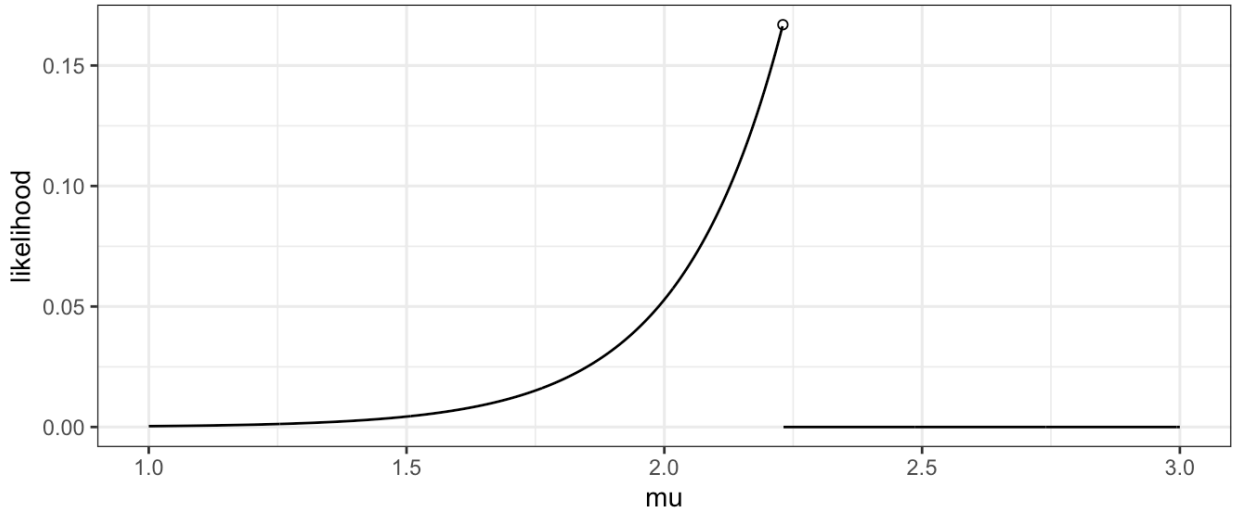
But not always (?!).

Example (Exponential threshold model): Suppose that $Y_1, \ldots, Y_n$ are iid from the exponential distribution with a threshold parameter μ,

$$f(y; \mu) = \begin{cases} \exp\{-(y-\mu)\} & \mu < y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

for $\infty < \mu < \infty$.

Consider the artificial data set $\boldsymbol{y} = [2.47, 2.35, 2.23, 3.53, 2.36]$.



1.5.1 The Fisher Information Matrix

The Fisher information matrix $I(\boldsymbol{\theta})$ is defined as the $b \times b$ matrix where

$$I_{ij}(\boldsymbol{\theta}) =$$

In matrix form,

$$I(\boldsymbol{\theta}) =$$

Fisher information facts:

1. The Fisher information matrix is the variance of the score contribution.

2. If regularity conditions are met,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \theta) \xrightarrow{d} \mathrm{N}_b(0, I(\boldsymbol{\theta})^{-1}).$$

3. If $b = 1$, then any unbiased estimator must have variance greater than or equal to $\{nI(\boldsymbol{\theta})\}^{-1}$

4. The information matrix is related to the curvature of the log-likelihood contribution.

1.5.2 Observed Information

The information matrix is not random, but it is also not observable from the data.

Let $Y_1, \ldots, Y_n$ be iid with density $f_Y(y_i; \boldsymbol{\theta})$. The log likelihood is defined as

taking two derivatives and dividing by n results in

Definition: The matrix $n\bar{I}(Y;\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ is called the sample information matrix, or the *observed information matrix*.

Why use $I(\boldsymbol{\theta}) = \mathrm{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f(Y_1;\boldsymbol{\theta})\right]$ as the basis for an estimator, rather than $I(\boldsymbol{\theta}) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log f(Y_1;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(Y_1;\boldsymbol{\theta})\right\}\right]$?

Now let's prove the asymptotic normality of the MLE (in the scalar case).