MATH3091 Statistical Modelling II
Lecture 4: Likelihood based inference

Chao Zheng

11th Feb 2022

Recap

Last time, we

- Define the expected (or Fisher) information $\mathcal{I}(\theta)$
- Proved that $\text{Var}_{\boldsymbol{\theta}}(U(\theta)) = \mathcal{I}(\theta)$
- Show the Newton–Raphson algorithm to find the MLE (local maximum)

2.3.1 Asymptotic distribution of the MLE

Asymptotic distribution of the MLE

Suppose that $y_1, \ldots, y_n$ are observations of independent random variables $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta})$ is completely specified except for the values of an unknown parameter vector $\boldsymbol{\theta}$, and that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$.

As $n \to \infty$, the distribution of $\hat{\boldsymbol{\theta}}$ tends to a multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and variance covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$.

Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the distribution of the MLE $\hat{\theta}$ tends to $N[\theta, 1/\mathcal{I}(\theta)]$.

Sketch proof (one parameter case)

(Proof not examinable)

We can write the score as

$$u(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_Y(y_i;\theta)$$

so $U(\theta)$ can be expressed as the sum of n i.i.d. random variables.

Asymptotically, as $n \to \infty$, by the central limit theorem, $U(\theta)$ is normally distributed.

But for the true θ, $E[U(\theta)] = 0$ and $\text{Var}[U(\theta)] = \mathcal{I}(\theta)$, so asymptotically

$$U(\theta) \sim N[0, \mathcal{I}(\theta)].$$

Sketch proof (continued)

A Taylor series expansion of $U(\hat{\theta})$ around the true θ gives

$$U(\hat{\theta}) = U(\theta) + (\hat{\theta} - \theta)U'(\theta) + \ldots$$

Now, $U(\hat{\theta}) = 0$, and if we approximate $U'(\theta) = H(\theta)$ by $E[H(\theta)] = -\mathcal{I}(\theta)$, and also ignore higher order terms

$$\hat{\theta} = \theta + \frac{1}{\mathcal{I}(\theta)} U(\theta).$$

As $U(\theta)$ is asymptotically $N[0, \mathcal{I}(\theta)]$, $\hat{\theta}$ is asymptotically $N[\theta, \mathcal{I}(\theta)^{-1}]$.

Approximate distribution of the MLE

For 'large enough n', we can treat the asymptotic distribution of the MLE as an approximation. The fact that $E(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$ means that the MLE is *approximately unbiased* for large samples.

Approximate distribution of the MLE

For 'large enough n', we can treat the asymptotic distribution of the MLE as an approximation. The fact that $E(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$ means that the MLE is *approximately unbiased* for large samples.

The variance of $\hat{\boldsymbol{\theta}}$ is approximately $\mathcal{I}(\boldsymbol{\theta})^{-1}$. It is possible to show that this is the smallest possible variance of any unbiased estimator of $\boldsymbol{\theta}$ (this result is called the Cramér–Rao lower bound, which we do not prove here).

Therefore the MLE is the 'best possible' estimator in large samples.

Bernoulli example

If $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p) random variables then

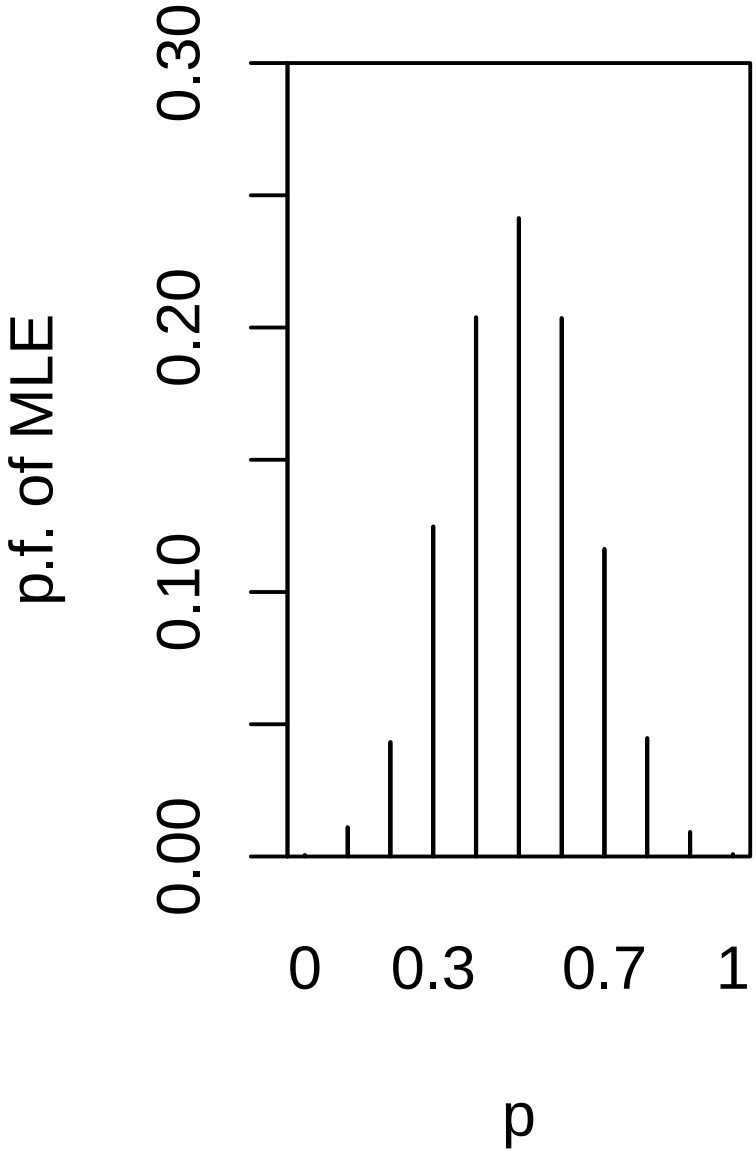
$$\mathcal{I}(p) = \frac{n}{p(1-p)},$$

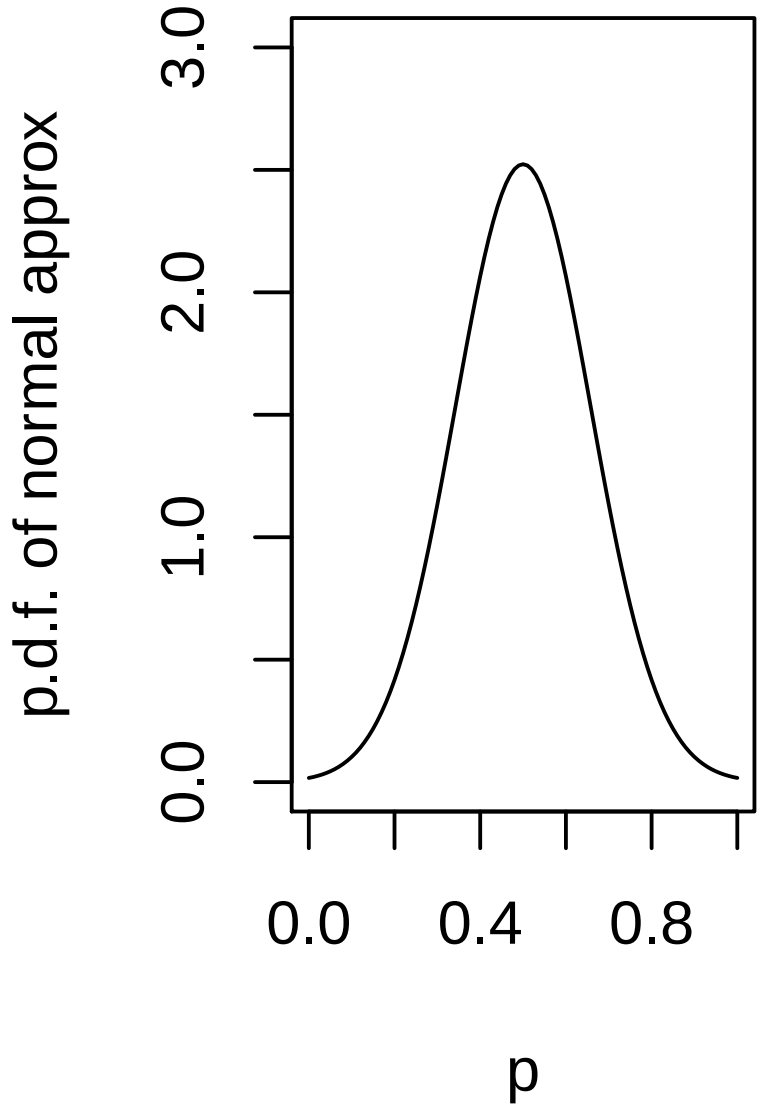
so asymptotically $\hat{p} = \bar{Y}$ has a $N(p, p(1-p)/n)$ distribution.

What does this result mean in practice?

e.g. if $n = 10$ and $p = 0.5$, how does the distribution of the MLE compare with a $N(0.5, 0.025)$ distribution?

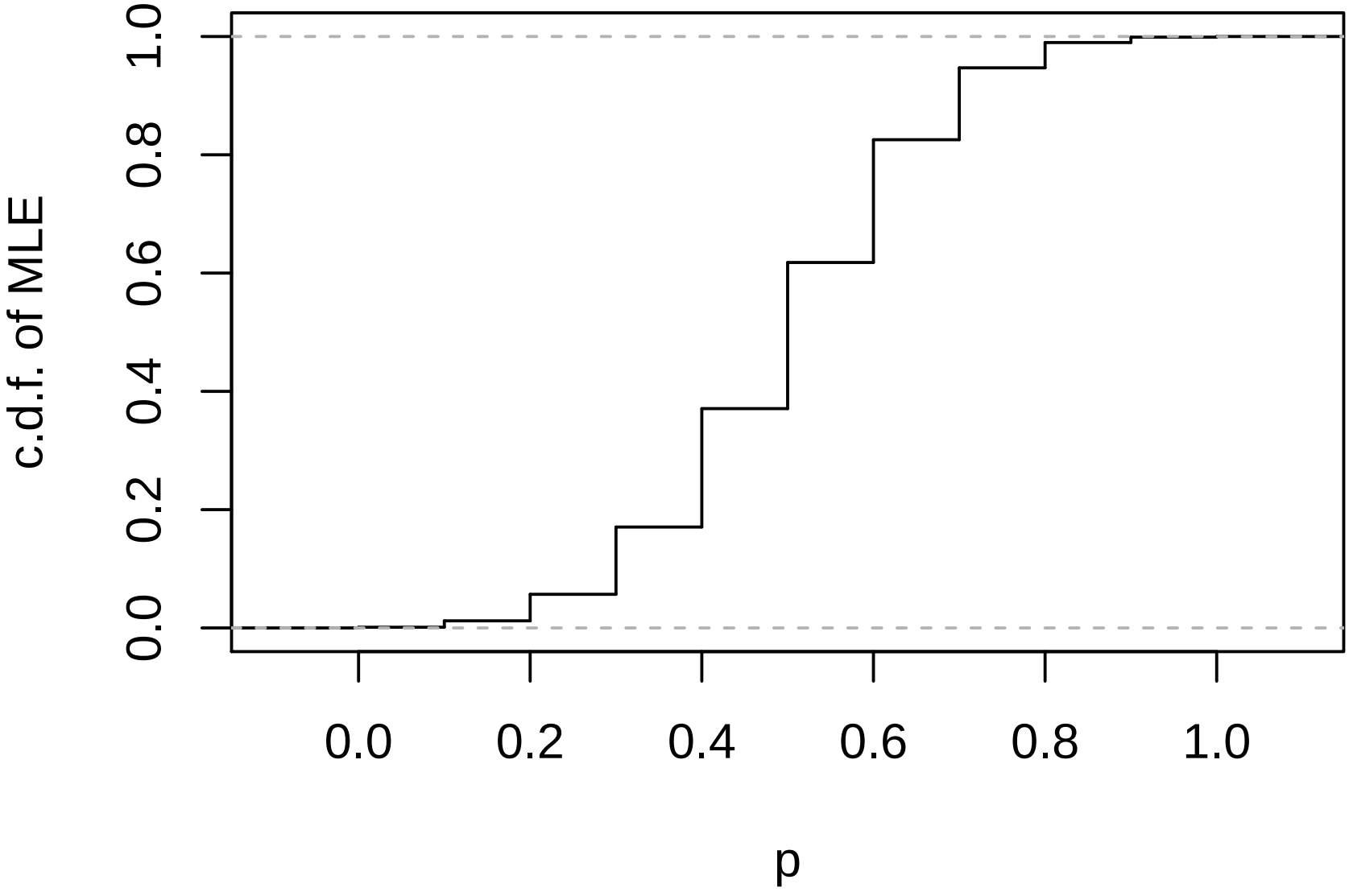
Bernoulli example: p.f. of $\hat{p}$ ($n = 10$, $p = 0.5$)



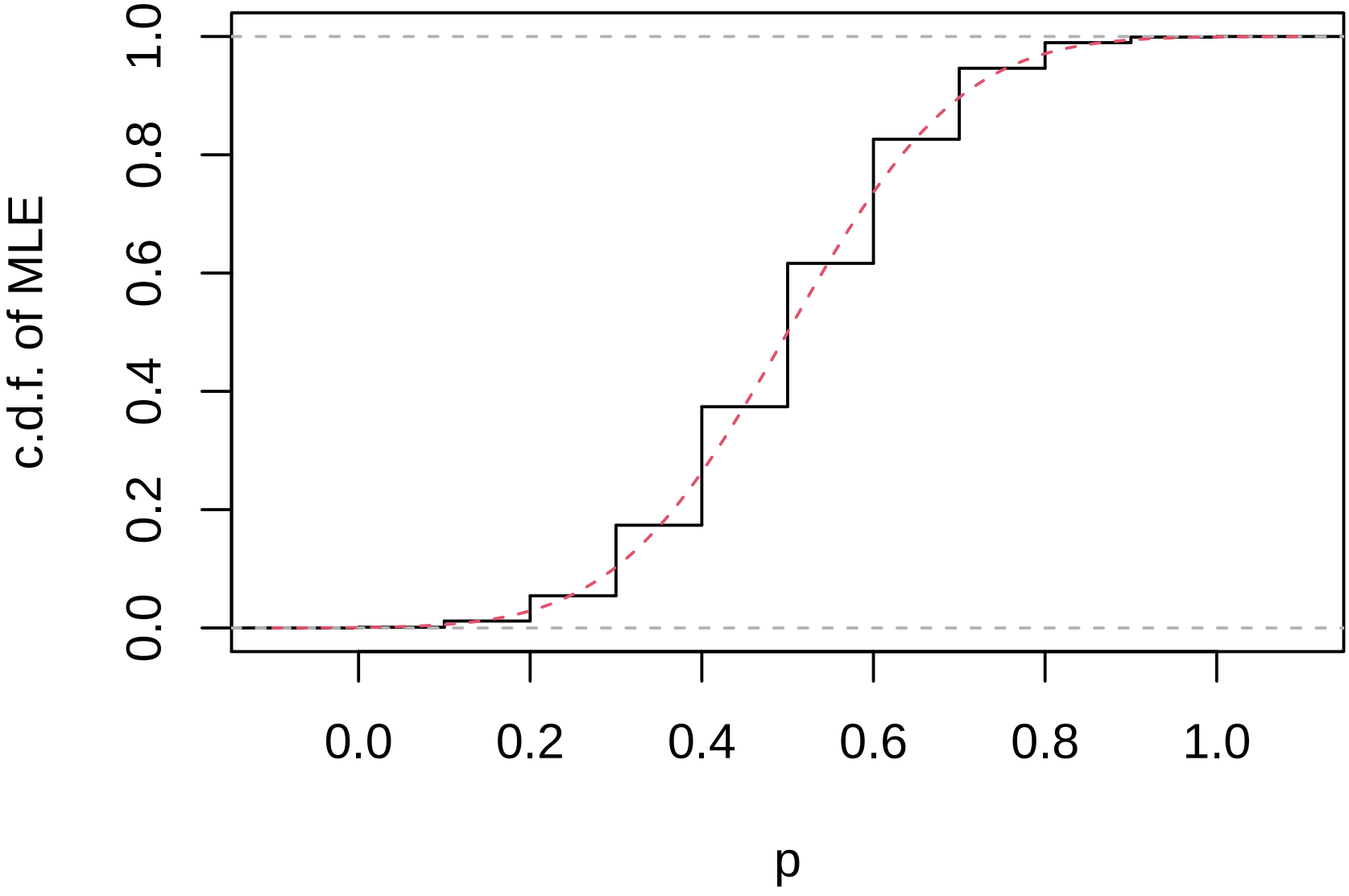


How should we compare these two distributions?

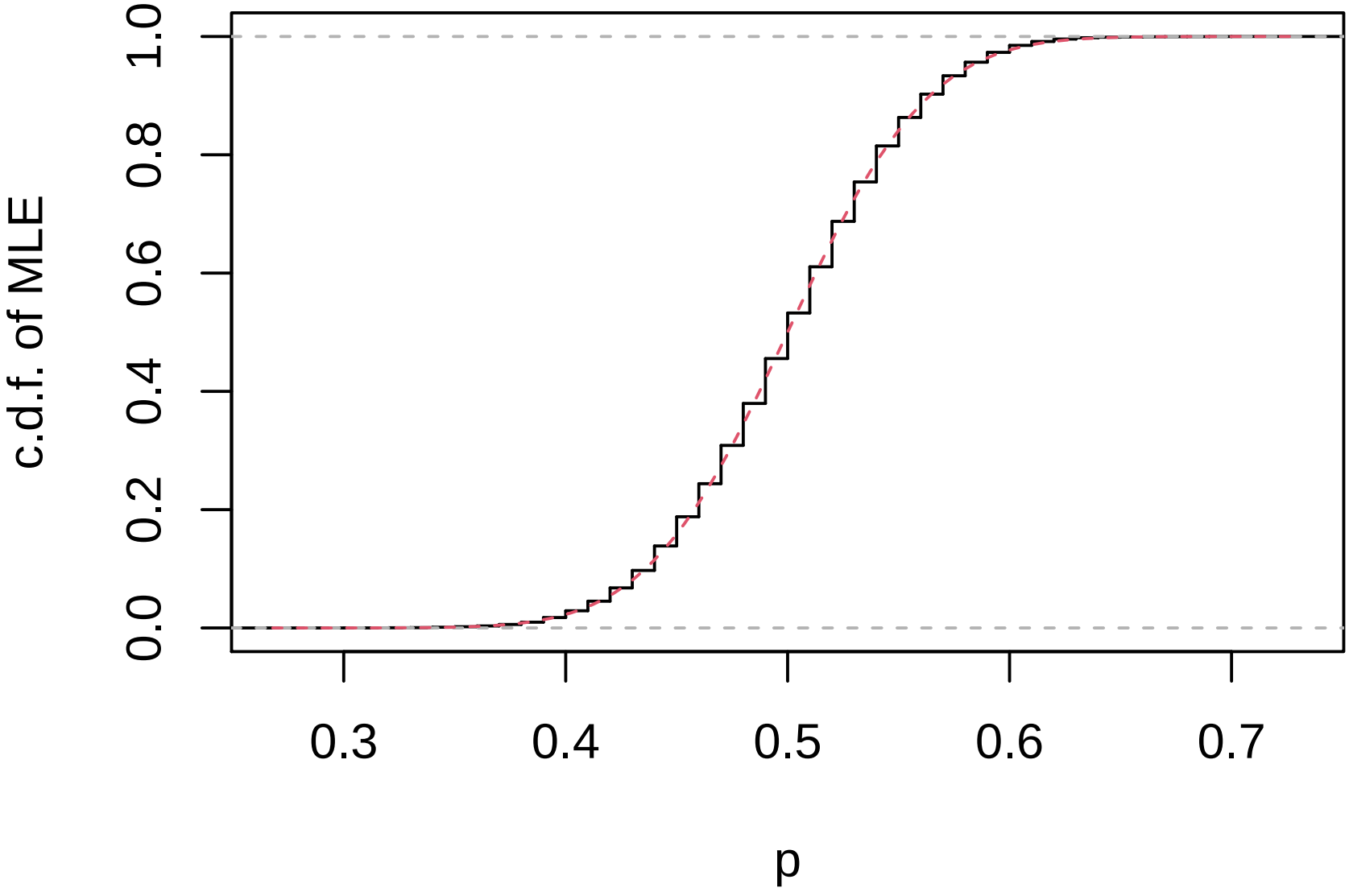
Bernoulli example: c.d.f. of $\hat{p}$ ($n = 10$, $p = 0.45$)



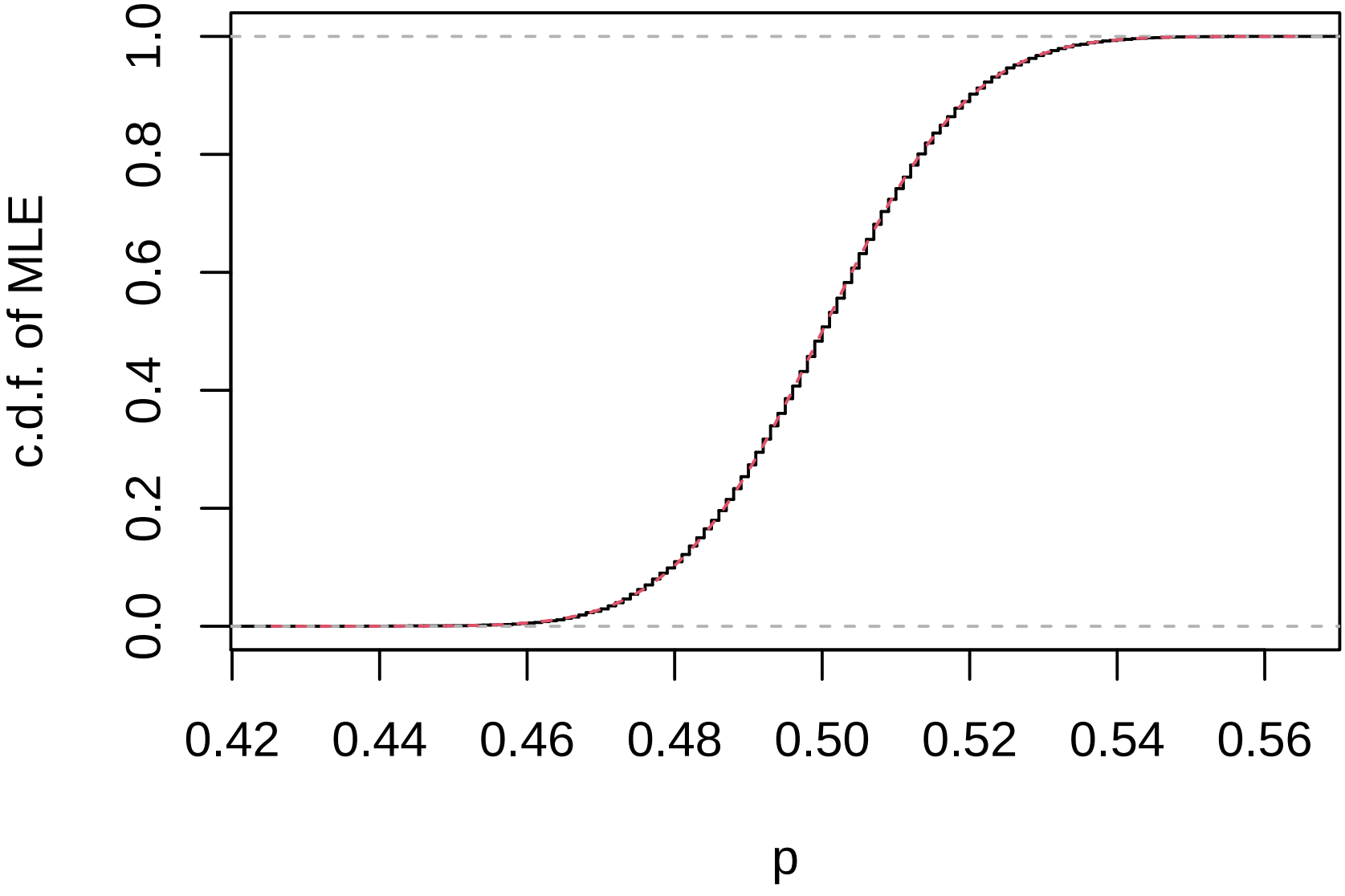
Bernoulli example: c.d.f. of $\hat{p}$ ($n = 10$, $p = 0.45$)



Bernoulli example: c.d.f. of $\hat{p}$ ($n = 100$, $p = 0.45$)



Bernoulli example: c.d.f. of $\hat{p}$ ($n = 1000$, $p = 0.45$)



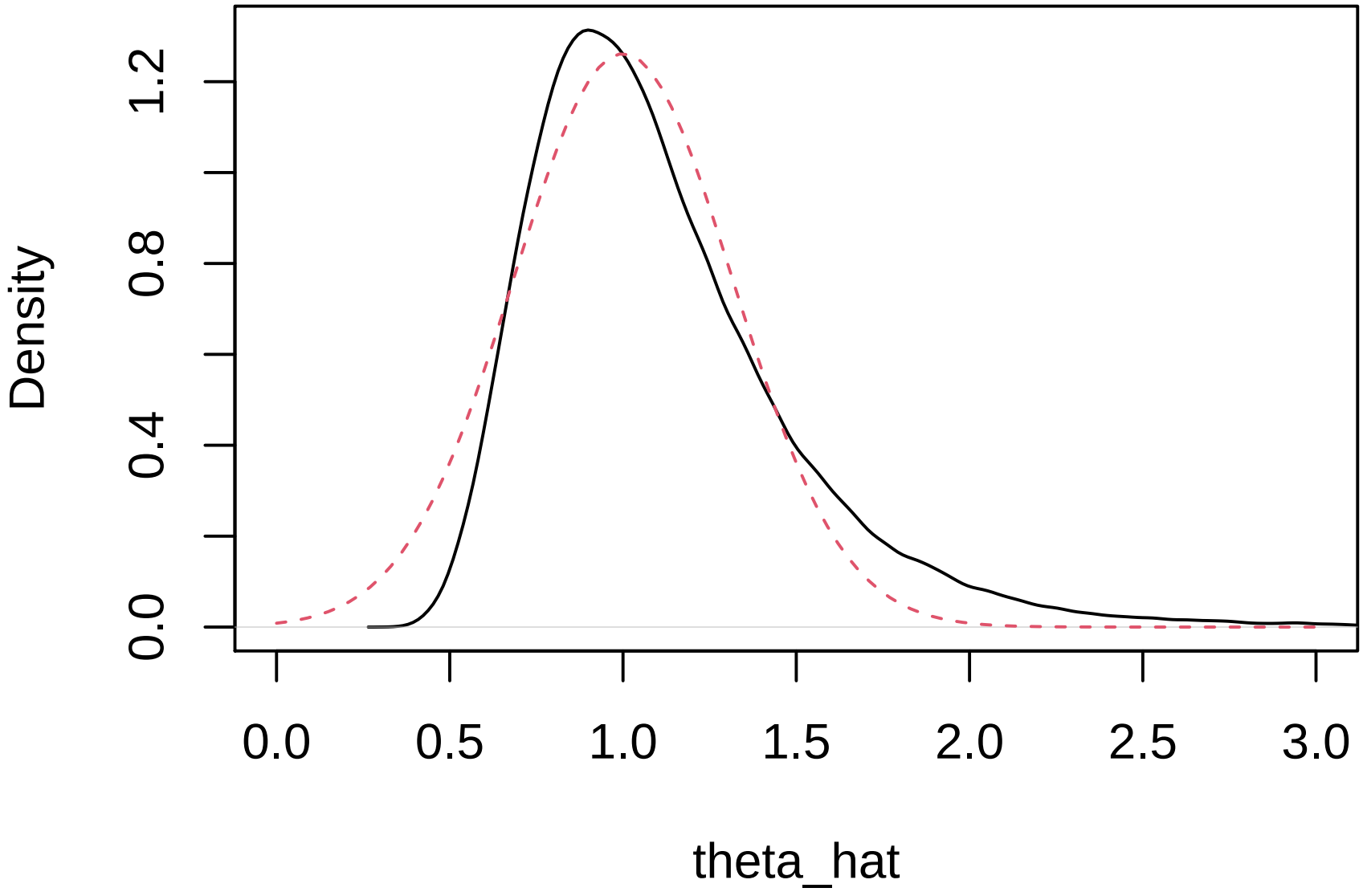
Exponential Examples

If $Y_1, \ldots, Y_n$ are i.i.d. Exponential(θ) random variables then

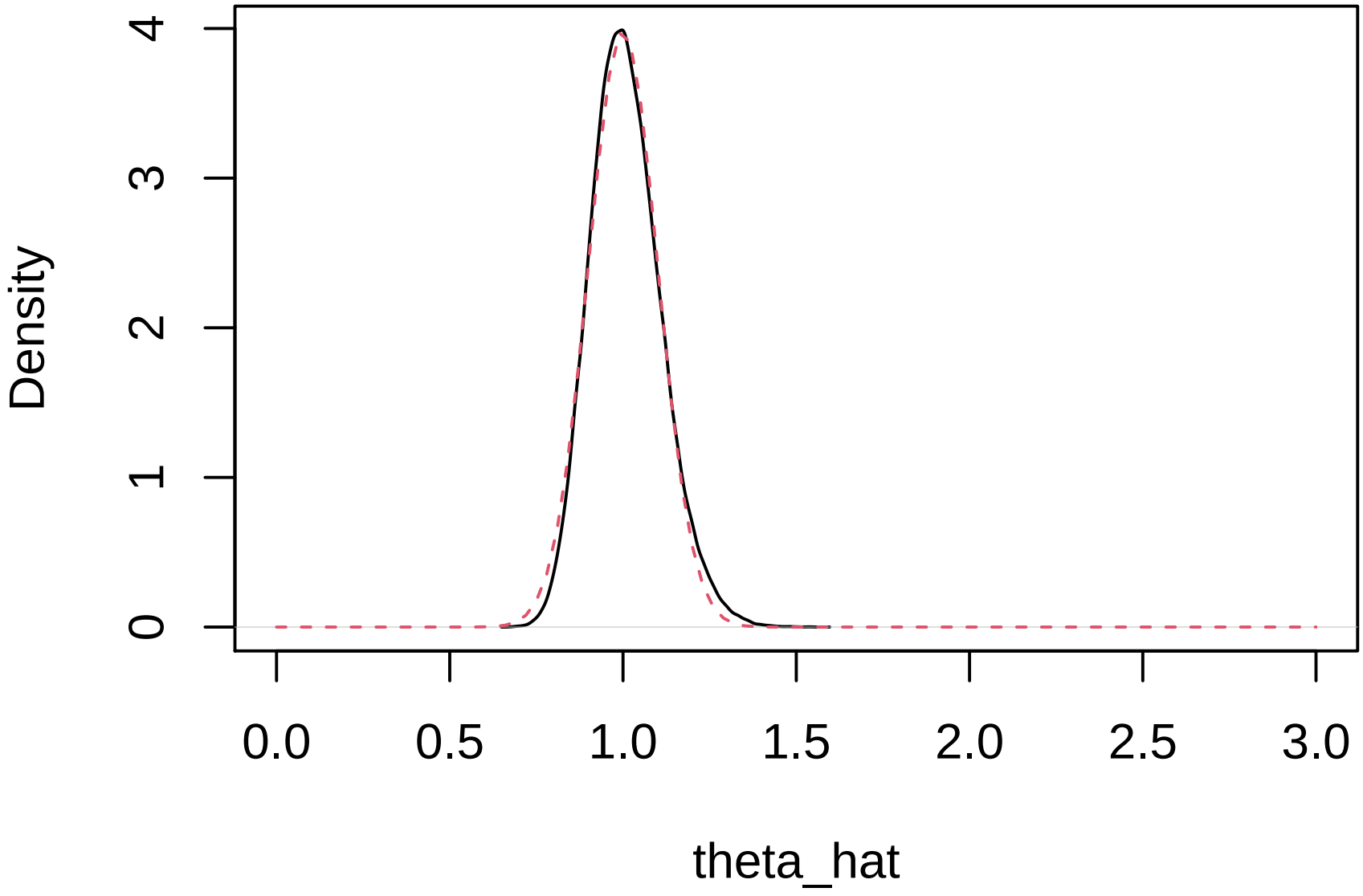
$$\mathcal{I}(\theta) = \frac{n}{\theta^2},$$

so asymptotically $\hat{\theta} = 1/\bar{Y}$ has a $N(\theta, \theta^2/n)$ distribution.

Exponential example: p.d.f. of $\hat{\theta}$ ($n = 10$, $\theta = 1$)



Exponential example: p.d.f. of $\hat{\theta}$ ($n = 100$, $\theta = 1$)



2.3.2 Quantifying uncertainty in parameter estimates

Standard errors

A *standard error* is an estimate of the standard deviation of an estimator.

If $p = 1$, a standard error of the MLE $\hat{\theta}$ is

$$s.e.(\hat{\theta}) = \frac{1}{\mathcal{I}(\hat{\theta})^{\frac{1}{2}}},$$

and for a vector parameter $\boldsymbol{\theta}$

$$s.e.(\hat{\theta}_i) = [\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}, \quad i = 1, \ldots, p.$$

Constructing large sample confidence intervals

Asymptotically, $\hat{\theta}_i \sim N(\theta_i, [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii})$ and we can find $z_{1-\frac{\alpha}{2}}$ such that

$$P\left(-\boldsymbol{z_{1-\frac{\alpha}{2}}} \leq \frac{\hat{\theta}_i - \theta_i}{[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Therefore

$$P\left(\hat{\theta}_i - z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}} \leq \theta_i \leq \hat{\theta}_i + z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}\right) = 1 - \alpha.$$

Constructing large sample confidence intervals

The endpoints of this interval cannot be evaluated because they also depend on the unknown parameter vector $\boldsymbol{\theta}$. However, if we replace $\mathcal{I}(\boldsymbol{\theta})$ by its MLE $\mathcal{I}(\hat{\boldsymbol{\theta}})$ we obtain the approximate large sample $\mathbf{100(1-\alpha)\%}$ confidence interval

$$[\hat{\theta}_i - z_{1-\frac{\alpha}{2}}[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}, \hat{\theta}_i + z_{1-\frac{\alpha}{2}}[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}].$$

For $\alpha = 0.1, 0.05, 0.01$, $z_{1-\frac{\alpha}{2}} = 1.64, 1.96, 2.58$.

Example (Bernoulli)

If $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables then asymptotically $\hat{p} = \bar{y}$ has a $N(p, p(1-p)/n)$ distribution, and a large sample 95% confidence interval for p is

$$
\begin{aligned}
&[\hat{p} - 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}, \hat{p} + 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}] \\
&= [\hat{p} - 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}, \hat{p} + 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}] \\
&= [\bar{y} - 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}, \bar{y} + 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}].
\end{aligned}
$$

how many heads?

Toss a coin 10 times. How many "heads" did you get?

Can work out $[\bar{y} - 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}, \bar{y} + 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}]$ for each possible number of heads:

```
##          lower     upper
## 0   0.00000000 0.0000000
## 1  -0.08594193 0.2859419
## 2  -0.04792257 0.4479226
## 3   0.01596902 0.5840310
## 4   0.09635811 0.7036419
## 5   0.19009679 0.8099032
## 6   0.29635811 0.9036419
## 7   0.41596902 0.9840310
## 8   0.55207743 1.0479226
## 9   0.71405807 1.0859419
## 10  1.00000000 1.0000000
```

how many heads?

Toss a coin 10 times. How many "heads" did you get?

Can work out $[\bar{y} - 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}, \bar{y} + 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}]$ for each possible number of heads:

```
##          lower     upper
## 0   0.00000000 0.0000000
## 1  -0.08594193 0.2859419
## 2  -0.04792257 0.4479226
## 3   0.01596902 0.5840310
## 4   0.09635811 0.7036419
## 5   0.19009679 0.8099032
## 6   0.29635811 0.9036419
## 7   0.41596902 0.9840310
## 8   0.55207743 1.0479226
## 9   0.71405807 1.0859419
## 10  1.00000000 1.0000000
```

For the number of heads you got, does the interval contain 0.5?

Checking the coverage ($n = 10$, $\alpha = 0.05$)

```r
n <- 10
sum_y <- rbinom(10000, size = n, prob = 0.5)
y_bar <- sum_y / n
I_hat <- y_bar * (1 - y_bar) / n
lower <- y_bar - 1.96 * sqrt(I_hat)
upper <- y_bar + 1.96 * sqrt(I_hat)
coverage <- mean(lower < 0.5 & upper > 0.5)
coverage
```

```
## [1] 0.8906
```

Checking the coverage ($n = 10$, $\alpha = 0.05$)

```r
n <- 10
sum_y <- rbinom(10000, size = n, prob = 0.5)
y_bar <- sum_y / n
I_hat <- y_bar * (1 - y_bar) / n
lower <- y_bar - 1.96 * sqrt(I_hat)
upper <- y_bar + 1.96 * sqrt(I_hat)
coverage <- mean(lower < 0.5 & upper > 0.5)
coverage
```

```
## [1] 0.8906
```

The actual coverage is lower than the nominal 95% level.

Checking the coverage ($n = 100$, $\alpha = 0.05$)

```r
n <- 100
sum_y <- rbinom(10000, size = n, prob = 0.5)
y_bar <- sum_y / n
I_hat <- y_bar * (1 - y_bar) / n
lower <- y_bar - 1.96* sqrt(I_hat)
upper <- y_bar + 1.96* sqrt(I_hat)
coverage <- mean(lower < 0.5 & upper > 0.5)
coverage
```

```
## [1] 0.9404
```

Checking the coverage ($n = 100$, $\alpha = 0.05$)

```
n <- 100
sum_y <- rbinom(10000, size = n, prob = 0.5)
y_bar <- sum_y / n
I_hat <- y_bar * (1 - y_bar) / n
lower <- y_bar - 1.96* sqrt(I_hat)
upper <- y_bar + 1.96* sqrt(I_hat)
coverage <- mean(lower < 0.5 & upper > 0.5)
coverage
```

```
## [1] 0.9404
```

The actual coverage is close to the nominal 95% level. The confidence interval is designed using an approximation which will work well for large n, so this is expected.

2.3.3 Comparing statistical models

Comparing statistical models

If we have a set of competing probability models which might have generated the observed data, we may want to determine which of the models is most appropriate.

Suppose that we have two competing alternatives, $f_{\boldsymbol{Y}}^{(0)}$ (model M_0) and $f_{\boldsymbol{Y}}^{(1)}$ (model M_1) for $f_{\boldsymbol{Y}}$, the joint distribution of $Y_1, \ldots, Y_n$.

Often M_0 and M_1 both take the same parametric form, $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ but with $\boldsymbol{\theta} \in \Theta^{(0)}$ for M_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for M_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$.

In the regression setting, we are often interested in determining which of a set of explanatory variables have an impact on the distribution of the response.

2.3.3.1 Hypothesis testing

Hypothesis testing

A hypothesis test provides one way of comparing two competing statistical models.

One hypothesis,

$$H_0\text{: the data were generated from model } M_0,$$

has special status, and is referred to as the *null hypothesis*. The null hypothesis is the reference model, and will be assumed to be appropriate unless the observed data strongly indicate that H_0 is inappropriate, and that

$$H_1\text{: the data were generated from model } M_1,$$

(the *alternative* hypothesis) should be preferred.

The fact that a hypothesis test does not reject H_0 should not be taken as evidence that H_0 is true and H_1 is not, merely that the data does not provide sufficient evidence to reject H_0 in favour of H_1.

Critical region

A hypothesis test is defined by its *critical region* or *rejection region*, which we shall denote by C.

- If $\boldsymbol{y} \in C$, H_0 is rejected in favour of H_1;
- If $\boldsymbol{y} \notin C$, H_0 is not rejected.

Size and power of a test

We define the *size* (or *significance level*) of the test

$$\alpha = \max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\boldsymbol{Y} \in C; \boldsymbol{\theta})$$

This is the maximum probability of erroneously rejecting H_0, over all possible distributions for $\boldsymbol{Y}$ implied by H_0.

We also define the power function

$$\omega(\boldsymbol{\theta}) = P(\boldsymbol{Y} \in C; \boldsymbol{\theta})$$

It represents the probability of rejecting H_0 for a particular value of $\boldsymbol{\theta}$.

A good test will have small size, but large power.

Fixing the size, maximising the power

In general, we fix α to be some small value (often 0.05), so that the probability of erroneous rejection of H_0 is limited. In doing this, we are giving H_0 precedence over H_1.

Given our specified α, we try to choose a test to make $\omega(\boldsymbol{\theta})$ as large as possible for $\boldsymbol{\theta} \in \Theta^{(1)} \setminus \Theta^{(0)}$.

2.3.3.2 Likelihood ratio tests for nested hypotheses

Likelihood ratio test

Suppose that M_0 and M_1 both take the same parametric form, $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta^{(0)}$ for M_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for M_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$.

A *likelihood ratio test* of H_0 against H_1 has a critical region of the form

$$C = \left\{ \boldsymbol{y} : \frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} L(\boldsymbol{\theta})} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\boldsymbol{Y} \in C;\boldsymbol{\theta}) = \alpha.$$

We only reject H_0 if the observed data are much more probable under some distribution in H_1 than any distribution under H_0.

In general, this will not be available to us. However, we can make use of an important asymptotic result.

The log likelihood ratio statistic

First we notice that, as log is a strictly increasing function, the rejection region is equivalent to

$$C = \left\{ \boldsymbol{y} : 2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} L(\boldsymbol{\theta})} \right) > k' \right\}$$

where

$$\max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\boldsymbol{y} \in C; \boldsymbol{\theta}) = \alpha.$$

We call

$$L_{01} \equiv 2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} L(\boldsymbol{\theta})} \right)$$

the *log likelihood ratio statistic*

Asymptotic distribution of the log likelihood ratio statistic

If H_0 is *nested within* H_1, in other words $\Theta^{(0)} \subset \Theta^{(1)}$ ($\Theta^{(0)}$ is a subspace of $\Theta^{(1)}$) then under H_0: $\boldsymbol{\theta} \in \Theta^{(0)}$, asymptotically as $n \to \infty$, L_{01} has a chi-squared distribution with degrees of freedom equal to the difference in the dimensions of $\Theta^{(1)}$ and $\Theta^{(0)}$.

A sketch proof is in the notes, but is not examinable.

Asymptotic distribution of the log likelihood ratio statistic

If H_0 is *nested within* H_1, in other words $\Theta^{(0)} \subset \Theta^{(1)}$ ($\Theta^{(0)}$ is a subspace of $\Theta^{(1)}$) then under H_0: $\boldsymbol{\theta} \in \Theta^{(0)}$, asymptotically as $n \to \infty$, L_{01} has a chi-squared distribution with degrees of freedom equal to the difference in the dimensions of $\Theta^{(1)}$ and $\Theta^{(0)}$.

A sketch proof is in the notes, but is not examinable.

So a *log likelihood ratio test* rejects H_0 if L_{01} exceeds the $100(1-\alpha)\%$ point of the relevant chi-squared distribution.

Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Suppose that we require a size α test of the hypothesis H_0: $p = p_0$ against the general alternative H_1: 'p is unrestricted' where α and p_0 are specified.

Here $\boldsymbol{\theta} = (p)$, $\Theta^{(0)} = \{p_0\}$ and $\Theta^{(1)} = (0, 1)$ and the log likelihood ratio statistic is

$$L_{01} = 2n\bar{y} \log\left(\frac{\bar{y}}{p_0}\right) + 2n(1-\bar{y}) \log\left(\frac{1-\bar{y}}{1-p_0}\right).$$

As $d_1 = 1$ and $d_0 = 0$, under H_0, the log likelihood ratio statistic has an asymptotic χ^2_1 distribution.

Finding the critical value

We reject H_0 if L_{01} is 'too large' to have come from a χ^2_1 distribution.

If $\alpha = 0.05$, then we should reject H_0 if the test statistic is greater than the 95% point of the χ^2_1 distribution:

```
qchisq(0.95, df = 1)
```

```
## [1] 3.841459
```

Quiz: nested or non-nested?

Suppose we have continuous response variables Y_i and explanatory variables x_i, and suppose $\epsilon_i \sim N(0, \sigma^2)$. Consider the following models:

- Model 1: $Y_i = \beta_0 + \epsilon_i$
- Model 2: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Model 3: $Y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i$
- Model 4: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$
- Model 5: $\log Y_i = \beta_0 + \epsilon_i$

Which of the following statements are true? Select as many as apply.

- Model 1 is nested in Model 2
- Model 2 is nested in Model 3
- Model 1 is nested in Model 3
- Model 2 is nested in Model 4
- Model 1 is nested in Model 5

2.3.3.3 Information criteria for model comparison

Comparing non-nested models

We have seen how to use a likelihood ratio test to compare two nested models, but we may also want to compare non-nested models. An alternative approach is to record some criterion measuring the quality of the model for each of a candidate set of models, then choose the model which is the best according to this criterion.

Comparing non-nested models

We have seen how to use a likelihood ratio test to compare two nested models, but we may also want to compare non-nested models. An alternative approach is to record some criterion measuring the quality of the model for each of a candidate set of models, then choose the model which is the best according to this criterion.

It is tempting to to choose the model which has the largest likelihood. However, if we do this we will always end up choosing complicated models, which fit the observed data very closely, but do not meet our requirement of parsimony.

Penalised likelihood approaches

For a given model depending on parameters $\theta \in \mathbb{R}^p$, let $\hat{\ell} = \ell(\hat{\theta})$ be the log-likelihood function for that model evaluated at the MLE $\hat{\theta}$. It is not sensible to choose between models by maximising $\hat{\ell}$ directly, and instead it is common to choose a model to maximise a criteria of the form

$$\hat{\ell} - \text{penalty},$$

where the penalty term will be large for complex models, and small for simple models.

Penalised likelihood approaches

For a given model depending on parameters $\theta \in \mathbb{R}^p$, let $\hat{\ell} = \ell(\hat{\theta})$ be the log-likelihood function for that model evaluated at the MLE $\hat{\theta}$. It is not sensible to choose between models by maximising $\hat{\ell}$ directly, and instead it is common to choose a model to maximise a criteria of the form

$$\hat{\ell} - \text{penalty},$$

where the penalty term will be large for complex models, and small for simple models.

Equivalently, we may choose between models by minimising a criteria of the form

$$-2\hat{\ell} + \text{penalty}.$$

By convention, many commonly-used criteria for model comparison take this form.

AIC and BIC

The Akaike information criterion (AIC) is

$$\text{AIC} = -2\hat{\ell} + 2p,$$

where p is the dimension of the unknown parameter in the candidate model, and the Bayesian information criterion (BIC) is

$$\text{BIC} = -2\hat{\ell} + \log(n)p,$$

where n is the number of observations.

BIC penalises complex models more than AIC. We can choose between models by choosing the one with smaller AIC (or smaller BIC).

Conclusion

- We have seen a sketch proof of the asymptotic distribution of the MLE.
- We have seen how to use this asymptotic distribution to construct (approximate) confidence intervals.
- These approximate confidence intervals will have about the right coverage if the sample size n is sufficiently large.
- We have seen how to use the likelihood ratio test to conduct a hypothesis test to compare two nested models.
- We have introduced AIC and BIC, which can be used to compare models even if those models are not nested.
- You should now be able to attempt all questions in problem sheet 1 and Question 2 in problem sheet 2 (available next week).