18.650

Rachel Wu

Spring 2018

These are my lecture notes from 18.650, Fundamentals of Statistics, at the Massachusetts Institute of Technology, taught this semester (Spring 2018) by Professor Victor Brunel[1]. My TA is Josh Pfeffer[2]. These notes owe primarily to their hard work and course materials.

I wrote these lecture notes in LaTeX in real time during lectures, so there may be errors, typos or omissions. I have lovingly pillaged Evan Chen's[3] and Tony Zhang's[4] formatting commands. Should you encounter an error in the notes, wish to suggest improvements, or alert me to a failure on my part to keep the web notes updated, please contact me at rmwu@mit.edu.

This document was last modified 2018-05-24. The permalink to these notes is http://web.mit.edu/rmwu/www.

[1] vebrunel@mit.edu
[2] pfeffer@mit.edu
[3] evanchen@mit.edu
[4] tz@mit.edu

Contents

1 February 6, 2018 **1**
1.1 Administrivia . . . 1
1.2 Introduction . . . 1

2 February 7, 2018 **3**
2.1 Recitation 1 . . . 3

3 February 8, 2018 **5**
3.1 Convergence . . . 5

4 February 13, 2018 **8**
4.1 The Delta method . . . 9

5 February 15, 2018 **12**
5.1 Inference . . . 12
5.2 Parametric inference . . . 12

6 February 22, 2018 **15**
6.1 Confidence intervals . . . 15
6.2 Covariance review . . . 16

7 February 27, 2018 **18**
7.1 Multivariate extensions . . . 18
7.2 Likelihood functions . . . 19

8 March 1, 2018 **20**
8.1 Continuous likelihood . . . 20
8.2 Discrete likelihood . . . 21
8.3 Maximum likelihood estimator . . . 22

9 March 6, 2018 **23**
9.1 Fisher information . . . 23
9.2 Performance of MLE . . . 25

10 March 7, 2018 **26**
10.1 Recitation 5 . . . 26

11 March 8, 2018 **28**
11.1 Fisher information and MLE . . . 28
11.2 Limitations of MLE . . . 29
11.3 Method of moments . . . 29

12 March 14, 2018 **31**
12.1 Recitation 6 31

13 March 20, 2018 **34**
13.1 Method of moments theorem 34
13.2 Generic method of moments 35
13.3 M-estimators 36

14 March 21, 2018 **38**
14.1 Recitation 7 38

15 March 22, 2018 **40**
15.1 M-estimators, ctd. 40
15.2 M-estimator asymptotics 41
15.3 M-estimators for robust statistics 42
15.4 Parametric hypothesis testing 43

16 April 3, 2018 **45**
16.1 Hypothesis testing 45
16.2 Type 1 and type 2 errors 47

17 April 4, 2018 **50**
17.1 Recitation 8 50

18 April 5, 2018 **52**
18.1 Hypothesis testing, ctd. 52
18.2 P-values 53
18.3 Neyman-Pearson's paradigm 55
18.4 Chi-squared distributions 55

19 April 10, 2018 **56**
19.1 Wald's test 56
19.2 Likelihood ratio test 58

20 April 11, 2018 **60**
20.1 Recitation 9 60

21 April 12, 2018 **61**
21.1 Testing implicit hypotheses 61
21.2 Multinomial chi-squared test 62
21.3 Student's distributions 63

22 March 18, 2018 **65**
22.1 Recitation 10 65

23 April 24, 2018 **68**
23.1 Chi-squared test of independence—discrete case 68
23.2 Chi-squared goodness-of-fit test—discrete case 69
23.3 Chi-squared goodness-of-fit test—infinite case 70

24 April 25, 2018 **72**
24.1 Recitation 11 72

25 April 26, 2018 **74**
25.1 Chi-squared degrees of freedom 74
25.2 Cumulative distribution function 74
25.3 Kolmogorov-Smirnov test 76

26 May 1, 2018 **79**
26.1 Bayesian statistics 79

27 May 2, 2018 **81**
27.1 Recitation 12 81

28 May 3, 2018 **83**
28.1 Non-informative priors 83
28.2 Bayesian confidence region 85
28.3 Bayesian estimation 85

29 May 8, 2018 **86**
29.1 Linear regression 86
29.2 Multivariate linear regression 88

30 May 9, 2018 **89**
30.1 Recitation 13 89

31 May 10, 2018 **91**
31.1 Linear regression review 91
31.2 Linear regression with deterministic design 91
31.3 Significance tests 92

32 May 15, 2018 **94**
32.1 Generalized Cochran's theorem 94
32.2 Significance tests, ctd. 95
32.3 Implicit hypotheses (linear) 96

33 May 17, 2018 **98**

33.1 Review 98

34 May 21, 2018 **101**

34.1 Final office hours 101

A Acknowledgements **106**

1 February 6, 2018

1.1 Administrivia

> "For those of you who just came in, I said something really important that you already missed—welcome to this class."—vebrunel

So welcome to the class!

- 11 problem sets, with the lowest grade dropped (30% of grade)
- 2 midterm exams on March 15 and April 19, with lower grade dropped (30% of grade)
- final exam (40% of grade)

1.2 Introduction

Suppose we want to estimate a parameter p associated with a coin, where p is the proportion of the mass of tails side to heads. We flip the coin n times and observe each outcome.

Formally, for $i = 1, 2, \ldots, n$, let $H_i = 1$ if heads shows up and 0 otherwise. The estimate of p is the average

$$\tilde{H_n} = \frac{1}{n} \sum_{i=1}^{n} H_i.$$

We assume that

- each H_i is a Bernoulli random variable with parameter p,
- and $H_1, \ldots, H_n$ are mutually independent.

Definition 1.1 (Population mean)**.** The population mean μ for each random variable H_i is its expected value $\mathbb{E}[H_i]$.

Note that not all random variables have means. For example the Cauchy random variable

$$f(x) = \frac{2}{\pi} \frac{1}{x^2 + 1}, x \in \mathbb{R}$$

has no mean because the integral

$$\int_{\mathbb{R}} x f(x) dx$$

does not exist.

Theorem 1.2 (Law of large numbers)
For i.i.d. random variables $X_1, \ldots, X_n$,

$$\tilde{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu$$

as n approaches ∞.

Theorem 1.3 (Central limit theorem)
For i.i.d. random variables $X_1, \ldots, X_n$,

$$\sqrt{n}\frac{\tilde{X}_n - \mu}{\sigma} \to \mathcal{N}(0,1)$$

as n approaches ∞.

Equivalently, we can write

$$\sqrt{n}(\tilde{X}_n - \mu) \to \mathcal{N}(0, \sigma^2) \tag{1.1}$$

since multiplying a random variable by c increases its variance by c^2.

We can apply these laws to our Bernoulli random variables, for which $\sigma = \sqrt{p(1-p)}$. Let $z \sim \mathcal{N}(0,1)$. Then

$$\Pr\left\{\left|\sqrt{n}\frac{\tilde{H}_n - p}{\sqrt{p(1-p)}}\right| \le t\right\} \to \Pr\{|z| \le t\}.$$

That is, the probability distributions converge point wise. With this inequality, we can bound p as

$$\begin{aligned}\left|\sqrt{n}\frac{\tilde{H}_n - p}{\sqrt{p(1-p)}}\right| &\le t \\ -t \le \sqrt{n}\frac{\tilde{H}_n - p}{\sqrt{p(1-p)}} &\le t \\ \tilde{H}_n - \frac{t\sqrt{p(1-p)}}{\sqrt{n}} \le p &\le \tilde{H}_n + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}\end{aligned}$$

So p is contained within the interval

$$\mathbb{I}_t = \left[\tilde{H}_n - \frac{t\sqrt{p(1-p)}}{\sqrt{n}}, \tilde{H}_n + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

We take $t = 1.96$, so

$$\Pr\{|z| \le 1.96\} = 0.95,$$

or

$$\Pr\{p \in \mathbb{I}_{1.96}\} \approx 95\%.$$

But we are sad! Because $\mathbb{I}$ depends on the true value of p! So we substitute p with $\tilde{H}_n$, to obtain

$$\tilde{\mathbb{I}}_t = \left[\tilde{H}_n - \frac{t\sqrt{\tilde{H}_n(1-\tilde{H}_n)}}{\sqrt{n}}, \tilde{H}_n + \frac{t\sqrt{\tilde{H}_n(1-\tilde{H}_n)}}{\sqrt{n}}\right]$$

The above is valid since

$$\mathbb{I}_t \subset \left[\tilde{H}_n - \frac{t}{2\sqrt{n}}, \tilde{H}_n + \frac{t}{2\sqrt{n}}\right]$$

because $\sqrt{p(1-p)} \le 1/2$ (think about this geometrically).

2 February 7, 2018

2.1 Recitation 1

Recitations focus on problem solving. Without further ado, consider a sequence of i.i.d. Bernoulli random variables $X_1, X_2, \ldots, X_n$ with parameter p.

1. Show that $\sqrt{n}(\tilde{x_n} - p)/(\sqrt{p(1-p)}) \stackrel{d}{=} z$.
   Using the central limit theorem, $\mathbb{E}[X_1] = p$ and $\operatorname{Var} X_1 = p(1-p)$.

2. Prove that $\forall t > 0$,

$$\Pr\{|z| \leq t\} = 2\Pr\{z \leq t\} - 1.$$

   Proof. We can expand $\Pr\{|z| \leq t\}$ as

$$\begin{aligned}
\Pr\{|z| \leq t\} &= \Pr\{-t \leq z \leq t\} \\
&= \Pr\{z \leq t\} - \Pr\{z < -t\} \\
&= \Pr\{z \leq t\} - \Pr\{-z < -t\}, \text{ since } z = -z \text{ for Gaussian } z \\
&= \Pr\{z \leq t\} - \Pr\{z > t\} \\
&= \Pr\{z \leq t\} - (1 - \Pr\{z \leq t\}) \\
&= 2\Pr\{z \leq t\} - 1
\end{aligned}$$

   □

3. For $t > 0$, let

$$\mathbb{I}_t = \left[\tilde{X_n} - \frac{t\sqrt{p(1-p)}}{\sqrt{n}}, \tilde{X_n} + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}\right].$$

   Prove that as $n \to \infty$,

$$\Pr\{\mathbb{I}_t \ni p\} \to 2\Phi(t) - 1$$

   where Φ is the cdf of z.

   Proof.

$$\begin{aligned}
\Pr\{\mathbb{I}_t \ni p\} &= \Pr\left\{\sqrt{n}\frac{\left|\tilde{X_n} - t\right|}{\sqrt{p(1-p)}} \leq t\right\} \\
&= \Pr\left\{\left|\sqrt{n}\frac{\tilde{X_n} - p}{\sqrt{p(1-p)}}\right| \leq t\right\}
\end{aligned}$$

since $\frac{\tilde{X_n}-p}{\sqrt{p(1-p)}} \to z$. So

$$\begin{aligned}\Pr\left\{\sqrt{n}\frac{\tilde{X_n}-p}{\sqrt{p(1-p)}} \le t\right\} &\to \Pr\{|z| \le t\} \\ &= 2\Pr\{z \le t\} - 1 \\ &= 2\Phi(t) - 1\end{aligned}$$

□

4. Solve for $2\Phi(t_0) - 1 = 0.95$. Using the table of Gaussian distributions, we find that $t \approx 1.96$.

5. Prove that for all p, $p(1-p) \le 1/4$.

$$\begin{aligned}0 &\le (2p-1)^2 \\ 0 &\le 4p^2 - 4p + 1 \\ p(1-p) &\le 1/4\end{aligned}$$

6. Find an interval $\mathbb{I}_t$ centered around $\tilde{X_n}$ that does not depend on p but still contains p with probability ≥ 0.95.

   We can plug in $1/4$ for the $p(1-p)$ term, and using $t = 1.98$, we obtain

$$\begin{aligned}\mathbb{I}_{t_0} &\le \left[\tilde{X_n} - \frac{t_0 \cdot \sqrt{0.25}}{\sqrt{n}}, X_n + \frac{t_0 \cdot \sqrt{0.25}}{\sqrt{n}}\right] \\ &= \left[\tilde{X_n} - \frac{0.98}{\sqrt{n}}, X_n + \frac{0.98}{\sqrt{n}}\right].\end{aligned}$$

   What if we know beforehand that $p \le 0.3$?

   We now bound $p(1-p) \le 0.3 \cdot 0.7 = 0.21$. Now we obtain

$$\mathbb{I}_{t_0} \le \left[\tilde{X_n} - \frac{t_0 \cdot \sqrt{0.21}}{\sqrt{n}}, X_n + \frac{t_0 \cdot \sqrt{0.21}}{\sqrt{n}}\right].$$

7. Prove that the statement $\mathbb{I}_t \ni p$ is equivalent to a polynomial inequality of degree 2 in p.

   We just say that

$$\begin{aligned}\sqrt{n}\left|\tilde{X_n} - p\right| &\le t_0\sqrt{p(1-p)} \\ n(\tilde{X_n} - p)^2 &\le t_0^2 p(1-p) \\ (n + t_0^2)p^2 - (2n\tilde{X_n} + t_0^2)p + n{\tilde{X_n}}^2 &\le 0.\end{aligned}$$

   We find that the roots of this polynomial are

$$\frac{2n\tilde{X_n} + t_0^2 \pm \sqrt{t_0^4 + 4t_0^2 + \tilde{X_n}(1-\tilde{X_n})}}{2n + 2t_0^2},$$

   which bound p.

8. Substitute p with our approximation.

3 February 8, 2018

3.1 Convergence

Let T_n for $n \geq 1$ be a sequence of random variables and T a random variable.

Definition 3.1. We say a sequence **converges in probability**, $T_n \xrightarrow[n\to\infty]{\mathbb{P}} T$ if and only if

$$\Pr\{|T_n - T| \geq \epsilon\} \to 0, \forall \epsilon > 0.$$

In the law of large numbers, $\tilde{X_n} \to \mu$.

Definition 3.2. We say a sequence **converges in distribution**, $T_n \xrightarrow[n\to\infty]{d} T$ if and only if

$$\Pr\{T_n \leq x\} \to \Pr\{T \leq x\}$$

for all $x \in \mathbb{R}$ at which the cdf of T is continuous.

Equivalent definitions include

1. $\mathbb{E}[f(T_n)] \underset{n\to\infty}{\to} \mathbb{E}[f(T)]$
2. moment generating function converges point wise $\forall x$

That is, the cdf of T_n converges to the cdf of T wherever it is continuous. In the central limit theorem, $\sqrt{n}(\tilde{X_n} - \mu)/\sigma \xrightarrow[n\to\infty]{d} z$, where $z \sim \mathcal{N}(0,1)$.

Example 3.3
Suppose we have $X_1, \ldots, X_n$ uniformly distributed random variables on the interval $[0,1]$. Then $\max_i X_i \xrightarrow[n\to\infty]{\mathbb{P}} 1$. Exercise to prove this.

The quantity $n(1 - \max_i X_i) \xrightarrow[n\to\infty]{d} z$, where $z \sim \text{Exp}(1)$, or $\Pr\{z \leq x\} = 1 - e^{-x}, \forall x \geq 0$.

Definition 3.4. We say that a sequence converges in L^p for $p \geq 1$

$$T_n \xrightarrow[n\to\infty]{L^p} T \text{ iff } \mathbb{E}[|T_n - T|^p] \underset{n\to\infty}{\longrightarrow} 0.$$

Definition 3.5. We say that a sequence converges in almost surely

$$T_n \xrightarrow[n\to\infty]{a.s.} T \text{ iff } \Pr\left\{T_n \underset{n\to\infty}{\longrightarrow} T\right\} = 1$$

Proposition 3.6

$$\tilde{X_n} \xrightarrow[n\to\infty]{L^2} \mu$$

Proof. We show that

$$\mathbb{E}\left[\left|\tilde{X_n} - \mu\right|^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

We know that $\mathbb{E}\left[\tilde{X_n}\right] = \mu$. By definition, the left hand side is $\operatorname{Var} \tilde{X_n}$, which is expanded as

$$\operatorname{Var} \frac{1}{n}\sum_i X_i = \frac{1}{n^2} \operatorname{Var} \sum_i X_i.$$

Since the X_i are independent,

$$\frac{1}{n^2} \operatorname{Var} \sum_i X_i = \frac{1}{n^2} \sum_i \operatorname{Var} X_i = \frac{1}{n^2} \sum_i \sigma^2 = \frac{\sigma^2}{n} \to 0.$$

□

> "Markov's inequality is just for your culture... It looks like ehhh but we don't care."—vebrunel

What wimps.

Theorem 3.7 (Markov's inequality)

Let z be a positive random variable that has an expectation. Then for all $x > 0$,

$$\Pr\{z > x\} \leq \frac{\mathbb{E}[z]}{x}$$

Proof. Observe that

$$x\mathbb{1}_{z>x} \leq z.$$

If $z > x$, then this event is satisfied, and the left hand side is x. Otherwise, the left hand side is 0. We take the expected value of both sides,

$$x \cdot \mathbb{E}[\mathbb{1}_{z>x}] \leq \mathbb{E}[z].$$

The expectation of the indicator is a probability,

$$x \cdot \Pr\{z > x\} \leq \mathbb{E}[z]$$

$$\Pr\{z > x\} \leq \frac{\mathbb{E}[z]}{x}.$$

□

Recall theorem 1.2 from the previous lecture, or the law of large numbers, restated below for clarity.

Theorem

For i.i.d. random variables $X_1, \ldots, X_n$,

$$\tilde{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

as n approaches ∞.

Proof. We show that for all $\epsilon > 0$,

$$\Pr\left\{\left|\overline{X_n} - \mu\right| > 0\right\} \underset{n\to\infty}{\longrightarrow} 0.$$

Earlier we have shown that this quantity converges in L^2. Note that $\Pr\left\{\left|\overline{X_n} - \mu\right| > 0\right\}$ is equivalent to $\Pr\left\{(\overline{X_n} - \mu)^2 \geq \epsilon^2\right\}$. Using Markov's inequality,

$$\Pr\left\{(\overline{X_n} - \mu)^2 \geq \epsilon^2\right\} \leq \frac{\mathbb{E}\left[(\overline{X_n} - \mu)^2\right]}{\epsilon^2}.$$

□

Example 3.8

Let X be a Bernoulli random variable. Then

$$X/n \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Example 3.9

Let $X_1, \ldots, X_n$ be Bernoulli random variables with $p = 1/2$. Then

$$X_n/n \xrightarrow[n\to\infty]{\text{a.s.}} 0$$

since $0 \leq X_n/n \leq 1/n$.

These ideas extend to multivariate random variables with the Euclidean norm instead of absolute value.

1. Convergence a.s. implies convergence in probability, and the two limits are equal a.s.
2. Convergence in L^p implies convergence in L^q for all $q \leq p$ and in probability, with equivalent limits a.s.
3. If f is a continuous function, convergences $T_n \to T$ imply $f(T_n) \to f(T)$.

The last fact motivates

$$\Pr\left\{\left|\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma}\right| \leq t\right\} \to \Pr\left\{|z| \leq t\right\}.$$

"Is it accepted in the US to say 'pain in the ass'?"—vebrunel

4 February 13, 2018

Recall that last class we discussed the various forms of convergence.

Proposition 4.1

If $\overline{X_n} \xrightarrow[n\to\infty]{\mathbb{P}} p$, then $\overline{X_n}$ is known as a **consistent estimator** for p. For any continuous f, $f(\overline{X_n}) \xrightarrow[n\to\infty]{\mathbb{P}} f(p)$.

Example 4.2

Suppose we would like to estimate the variance $p(1-p)$. If $\overline{X_n} \xrightarrow[n\to\infty]{\mathbb{P}} p$, then $\overline{X_n}(1-\overline{X_n})$ is a consistent estimator for the variance.

One can add and multiply limits almost surely and in probability. That is, if $U_n \xrightarrow[n\to\infty]{\mathbb{P}} U$ and $V_n \xrightarrow[n\to\infty]{\mathbb{P}} V$, then

- $U_n + V_n \xrightarrow[n\to\infty]{\mathbb{P}} U + V$
- $U_n V_n \xrightarrow[n\to\infty]{\mathbb{P}} UV$
- and if $V \neq 0$ a.s., then $U_n/V_n \xrightarrow[n\to\infty]{\mathbb{P}} U/V$.

The same holds for $\xrightarrow{\text{a.s.}}$. However, these rules *do not* apply to convergence in distribution unless the pair (U_n, V_n) convergences in distribution to (U, V).

Example 4.3

Let T_n be a sequence such that $T \xrightarrow[n\to\infty]{d} z$, where z is a standard Gaussian variable. Now consider $-T_n$. If we take $f(x) = -x$, then

$$f(T_n) \xrightarrow[n\to\infty]{d} f(z)$$
$$-T_n \xrightarrow[n\to\infty]{d} -z$$
$$-T_n \xrightarrow[n\to\infty]{d} z$$

since $z = -z$. Note here that we *cannot* say that $T_n - T_n \xrightarrow[n\to\infty]{d} 2z$ because 0 is a deterministic constant.

Note that (U, V) is the joint distribution. It may be possible that the marginals of U_n, V_n converge to the marginals of U, V, but the joint distribution may not converge.

Example 4.4

We observe that the times between arrivals at a call center is $T_1, \ldots, T_n$. We may assume that these times are mutually independent, and

$$T_1, \ldots, T_n \sim \text{Exponential}(\lambda).$$

We may want to estimate λ.

Mutual independence is reasonable because customers are unrelated to each other; they do not coordinate to DDoS the call center at the same time. Furthermore, exponential variables exhibit lack of memory, such that

$$\Pr\{T_1 > t + s \mid T_1 > t\} = \Pr\{T_1 > s\}.$$

Suppose the density of T_1 is

$$f(t) = \lambda e^{-\lambda t}$$

so $\mathbb{E}[T_1] = 1/\lambda$. Thus, a reasonable estimate of $1/\lambda$ is $\overline{T_n} = \frac{1}{n}\sum_i T_i$. By the law of large numbers, $\overline{T_n} \to \mathbb{E}[T_1] = 1/\lambda$, so

$$\frac{1}{\overline{T_n}} \xrightarrow[n\to\infty]{\mathbb{P}} \lambda.$$

4.1 The Delta method

We can use the central limit theorem to provide bounds on this estimate.

$$\sqrt{n}\left(\overline{T_n} - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^{-2})$$

However, we care about λ, not $1/\lambda$. Let $f(x) = 1/x$ and $\theta = 1/\lambda$. From calculus,

$$f(\overline{T_n}) - f(\theta) \approx f'(\theta)(\overline{T_n} - \theta)$$

or the first order approximation. In context,

$$\sqrt{n}(f(\overline{T_n}) - f'(\theta)) \approx f(\theta)\sqrt{n}(\overline{T_n} - \theta) \xrightarrow[n\to\infty]{d} z$$

where $z \sim \mathcal{N}(0, 1/\lambda^2)$. So

$$f'(\theta)z \sim \mathcal{N}\left(0, f'(\theta)^2 \frac{1}{\lambda^2}\right).$$

Definition 4.5. Let $(Z_n)_{n\geq 1}$ be a sequence of random variables that satisfies

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2).$$

Then Z_n is **asymptotically normal** around θ with **asymptotic variance** σ^2.

Note 4.6. The asymptotic variance should be some function of θ that does not depend on n and is not a random variable.

Theorem 4.7

Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable around the point θ. Then $g(Z_n)$ is also asymptotically normal.

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

There are several implications.

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^2)$$

If we rearrange terms,

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\lambda} \xrightarrow[n\to\infty]{d} z$$

where $z \sim \mathcal{N}(0, 1)$. Let $\alpha \in (0, 1)$. Now we want to find an interval $\mathbb{I}$ that only depends on $T_1, \ldots, T_n$ such that $\mathbb{I} \ni \lambda$ with probability approaching $1 - \alpha$ for n sufficiently large.

Since absolute value is a continuous function, $\forall t \geq 0$,

$$\Pr\left\{\left|\sqrt{n}\frac{\hat{\lambda} - \lambda}{\lambda}\right| \leq t\right\} \xrightarrow[n\to\infty]{} \Pr\{|z| < t\}$$

Solving for λ,

$$\Pr\{\mathbb{I}_{\text{dumb}} \ni \lambda\} = \Pr\left\{\hat{\lambda} - \frac{\lambda t}{\sqrt{n}} \leq \lambda \leq \hat{\lambda} + \frac{\lambda t}{\sqrt{n}}\right\}$$

when $\mathbb{I}_{\text{dumb}} = [\hat{\lambda} - \frac{\lambda t}{\sqrt{n}}, \hat{\lambda} + \frac{\lambda t}{\sqrt{n}}]$

How do we choose t? Well $\Pr\{|z| \leq t\} = 1-\alpha$. Solving, $1-\alpha = 1-2(1-\Phi(t))$, so $\Phi(t) = 1 - \alpha/2$. So take $t = q_{1-\alpha/2}$, where q is the quantile of $\mathcal{N}(0, 1)$.

But recall! $\mathbb{I}_{\text{dumb}}$ still contains λ!

1. We can write that $\lambda \ni \mathbb{I}_{\text{dumb}}$ and solve the inequalities.
2. We can also substitute λ with $\hat{\lambda}$.

We prove that the latter works. We need to show that

$$\sqrt{n}\frac{\hat{\lambda} - \lambda}{\hat{\lambda}} \xrightarrow[n\to\infty]{d} z$$

where the denominator gave us the original annoying λ. We know that $\lambda/\hat{\lambda} \to 1$, so if we multiply the original expression by this term, we have our $\hat{\lambda}$!

Theorem 4.8 (Slutsky's theorem)

Let X_n, Y_n be two sequences of random variables such that $X_n \xrightarrow[n\to\infty]{d} X, Y_n \xrightarrow[n\to\infty]{\mathbb{P}} c$, where X is a random variable and $c \in \mathbb{R}$. Then

$$(X_n, Y_n) \xrightarrow[n\to\infty]{d} (X, c).$$

In particular,

$$X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$$
$$X_n Y_n \xrightarrow[n\to\infty]{d} cX$$

This is where the 1 works in.

5 February 15, 2018

5.1 Inference

Often, we have a lot of data and we want to learn something from them.

Suppose we know how the data are distributed, but we do not know the exact parameters of those distributions. We may want to infer parameters from our observations.

Example 5.1

We know that arrival time at a call center is exponentially distributed. Given arrival time data, observed over the course of a week, how is arrival time distributed?

This question relies on the strong assumption that $X_1, \ldots, X_n$ are exponential, and its answer is simple: estimate λ.

A harder problem would be to estimate the distribution of X. Is it Poisson distributed? Exponential? Gaussian? This "testing" problem falls under **non-parametric inference**. Non-parametric estimation also includes density estimation, which is a really hard problem (we'll talk about it in a month!)

Since it's only the second week, we'll focus on the easy problem today. We assume a parametric model and do **parametric inference**.

- Estimation is a subset of inference. We can assume a Gaussian and *estimate* the mean and variance.
- Hypothesis testing is another type of inference. This includes answering a decision problem: is the mean positive?
- Confidence intervals are yet another type. What is a 95% interval that contains the true mean?

5.2 Parametric inference

Definition 5.2. Let the observed outcome of a statistical experiment be a sample $X_1, \ldots, X_n$ of i.i.d. random variables in space $E \subseteq \mathbb{R}$. Let $\mathbb{P}$ be their common distribution. A **statistical model** is a pair $(E, \mathbb{P}_{\theta \in \Theta})$ where

- E is the sample space,
- $\mathbb{P}$ is the family of probability distributions on E, and
- Θ is the parameter set.

We assume $\mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Theta$.

If the last statement does not hold, the statistical model is **misspecified**.

Example 5.3

Here are some sample statistical models.

- For n Bernoulli trials:

$$\left(\{0,1\}, \text{Bernoulli}(p)_{p\in\{0,1\}}\right).$$

- If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \epsilon(\lambda)$, for some unknown $\lambda > 0$:

$$\left(\{0,\infty\}, \epsilon(\lambda)_{\lambda>0}\right).$$

- If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$:

$$\left(\{0,\infty\}, \mathcal{N}(\mu,\sigma^2)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}\right).$$

We introduce standard notation for statistical models.

- We assume that the model is **well-specified**. That is, $\mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Theta$.
- We denote the expectation operator association with $\mathbb{P}_\theta$ as $\mathbb{E}_\theta$.
- θ is known as the unknown **true parameter**.

In this class, we assume that $\Theta \subseteq \mathbb{R}^d$.

> "Let us agree that σ^2 is one symbol. It's not a Greek letter, it's French. It's written as `sigmasquared`."—vebrunel

Definition 5.4. The parameter θ is **identified** iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective. That is,

$$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}.$$

Suppose our model is

$$\left(\{0,1\}, \text{Bernoulli}\left(\Phi(\frac{\mu}{\sigma})\right)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}\right).$$

Then we *cannot* identify the values of μ, σ^2 from this model. That is because multiple settings of μ, σ^2 result in the same distribution.

Definition 5.5. A **statistic** is a measurable function of the sample.

Definition 5.6. A **estimator** of θ is any statistic whose expression does not depend on θ.

Definition 5.7. An estimator $\hat{\theta_n}$ of θ is **consistent** iff

$$\hat{\theta_n} \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

with respect to $\mathbb{P}_\theta$.

Definition 5.8. The **bias** of an estimator $\hat{\theta_n}$ is

$$\mathbb{E}_\theta[\hat{\theta_n}] - \theta.$$

Definition 5.9. The **risk** (or quadratic risk) of an estimator $\hat{\theta_n}$ is

$$\mathbb{E}_\theta\left[\left|\hat{\theta_n} - \theta\right|^2\right].$$

It is not necessarily true that an unbiased estimator is consistent. For example, X_1 is not biased, but it is not consistent.

Theorem 5.10 (Bias-variance decomposition)
If $\Theta \subseteq \mathbb{R}$, then quadratic risk is $\text{bias}^2 + \text{variance}$.

Proof. We expand the risk.

$$\begin{aligned}
\mathbb{E}_\theta\left[\left|\hat{\theta_n} - \theta\right|^2\right] &= \mathbb{E}_\theta\left[\left|\hat{\theta_n} - \mathbb{E}_n[\hat{\theta_n}] + \mathbb{E}_\theta[\hat{\theta_n}] - \theta\right|^2\right] \\
&= \mathbb{E}_\theta\left[(\hat{\theta_n} - \mathbb{E}_\theta(\hat{\theta_n})^2\right] + 0 + (\mathbb{E}_\theta[\hat{\theta_n}] - \theta)^2 \\
&= \operatorname{Var}\hat{\theta_N} + \text{bias}^2
\end{aligned}$$

□

6 February 22, 2018

The problem sets have been graded, and the professor will have them at office hours tomorrow.

> "I know for some people in the room, handwriting is a big problem, but please make an effort—just for your name! Your name!—veb

6.1 Confidence intervals

Let $E, (\Pr_\theta)_{\theta\in\Theta}$ be a statistical model based on observations $X_1, \dots, X_n$ and assume $\Theta \subseteq \mathbb{R}$.

Definition 6.1. Let $\alpha \in (0,1)$. A **confidence interval** of level α for θ is any random interval $\mathcal{I}$ whose boundaries do not depend on θ and

$$\Pr\{\mathcal{I} \ni \theta\} \geq 1-\alpha, \forall\theta \in \Theta.$$

A confidence interval of asymptotic level α is an interval such that

$$\lim_{n\to\infty} \Pr\{\mathcal{I} \ni \theta\} \geq 1-\alpha, \forall\theta \in \Theta.$$

Question 6.2. How big of n is big enough? In most models we care about, $n = 30$ is as good as $n \to \infty$ in practice.

Example 6.3

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

First, assume that σ^2 is known. Since X_i are Gaussian distributed,

$$\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

Note that we do *not* require $n \to \infty$, and we do not use the central limit theorem. So for $z \sim \mathcal{N}(0,1)$,

$$\begin{aligned}\Pr\left\{\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \leq t\right\} &= \Pr\{|z| \leq t\} = 2\Pr\{z \leq t\} - 1 \\ &= \Pr\left\{\left[\overline{X_n} - \frac{t\sigma}{\sqrt{n}}, \overline{X_n} + \frac{t\sigma}{\sqrt{n}}\right] \ni p\right\}\end{aligned}$$

Interval $\mathcal{I}_t$ contains p with probability $2\phi(t)-1$. We take t such that $2\phi(t)-1 = 1-\alpha$, so

$$\phi(t) = 1-\alpha 2, t = \phi^{-1}\left(1-\frac{\alpha}{2}\right) = q_{1-\alpha/2}.$$

Now suppose σ^2 is not known. Recall that

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}\left[E - \mathbb{E}[X]\right]^2.$$

Let $\hat{\sigma^2}$ be the sample variance

$$\hat{\sigma^2} = \frac{1}{n}\sum_i X_i^2 - \left(\frac{1}{n}\sum_i X_i\right)^2 = \frac{1}{n}\sum_i (X_i - \overline{X_i})^2 = \overline{X_n^2} - (\overline{X}_n)^2.$$

We show that $\hat{\sigma^2}$ is a consistent estimator of σ^2. By the law of large numbers,

$$\frac{1}{n}\sum_i X_i^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\left[X_1^2\right]$$
$$\left(\frac{1}{n}\sum_i X_i\right)^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\left[X_1\right]^2$$

Subtracting the two lines, we find that

$$\hat{\sigma^2} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2 = \sigma^2.$$

However. this is a biased estimator. If we replace n with $n-1$, we have the bias-corrected sample variance. Let

$$\mathcal{J} = \left[\overline{X_n} - \frac{t\hat{\sigma}}{\sqrt{n}}, \overline{X_n} + \frac{t\hat{\sigma}}{\sqrt{n}}\right]$$

where $\hat{\sigma} = \sqrt{\hat{\sigma^2}}$ and $t = q_{1-\alpha/2}$. So we can say that

$$\Pr\{\mathcal{J} \ni p\} \xrightarrow[n\to\infty]{} 1-\alpha$$

a la Slutsky's theorem,

$$\sqrt{n}\frac{\overline{X_n} - \mu}{\hat{\sigma}} = \sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \cdot \frac{\sigma}{\hat{\sigma}}.$$

"You're losing 12 precious seconds of your life, I know I know."—veb

Example 6.4

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. What is an upper bound for μ?

Just take $t = q_\alpha$, where t is the threshold.

6.2 Covariance review

Let

$$\hat{\sigma^2} = g(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i X_i^2).$$

The following pair should converge to

$$\begin{pmatrix} \frac{1}{n}\sum_i X_i \\ \frac{1}{n}\sum_i X_i^2 \end{pmatrix} = \frac{1}{n}\sum_i \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix} = \mathcal{N}(0, \Sigma)$$

where Σ is the covariance matrix.

Example 6.5

If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$, then $\overline{X_n} \sim \mathcal{N}(\theta, 1/n)$. So

$$\Pr\left\{\overline{X_n} \leq 0\right\} = \Pr\left\{\sqrt{n}(\overline{X_n} - \theta) \leq -\sqrt{n}\theta\right\} = \Phi(-\sqrt{n}\theta).$$

7 February 27, 2018

7.1 Multivariate extensions

Theorem 7.1 (Multivariate CLT)

Let $X_1, \dots, X_n \in \mathbb{R}^d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$. Then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma).$$

We know that by CLT,

$$\sqrt{n}(\overline{X}_n - \mathbb{E}[X_1]) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1).$$

We rewrite the equation in terms of z,

$$\sqrt{n}(\overline{X}_n - \mathbb{E}\left[X_1^2\right]) \xrightarrow[n\to\infty]{d} z$$

where $z \sim \mathcal{N}(0, \Sigma)$. By the Delta method,

$$\sqrt{n}(\overline{X}_n - \mathbb{E}[X_1]^2) \xrightarrow[n\to\infty]{d} z'$$

where $z' \sim \mathcal{N}(0, 4\mathbb{E}[X_1]^2 \operatorname{Var} X_1)$.

Let U, V be two random variables, and let $Y = \begin{pmatrix} U & V \end{pmatrix}^T$. Then

$$\mathbb{E}[Y] = \begin{pmatrix} \mathbb{E}[U] \\ \mathbb{E}[V] \end{pmatrix}$$

and

$$\operatorname{Var} Y = \begin{pmatrix} \operatorname{Var} U & \operatorname{cov}(U, V) \\ \operatorname{cov}(V, U) & \operatorname{Var} V \end{pmatrix}.$$ [5]

Remark 7.2. Note that we often say $\sigma = \sqrt{\sigma^2}$ is the standard deviation. In multiple dimensions, σ doesn't make sense, so the more general quantity to report is variance.

Example 7.3

Show that $\begin{pmatrix} \overline{X}_n & \overline{X}_n^2 \end{pmatrix}^T$ is asymptotically normal.

We expand as

$$\begin{pmatrix} \overline{X}_n \\ \overline{X}_n^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_i X_i \\ \frac{1}{n}\sum_i X_i^2 \end{pmatrix} = \frac{1}{n} \cdot \begin{pmatrix} \sum_i X_i \\ \sum_i X_i^2 \end{pmatrix} = \frac{1}{n} \cdot \sum_i Y_i$$

where $Y_i = \begin{pmatrix} X_i & X_i^2 \end{pmatrix}^T$.

[5] Recall that $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$.

By the CLT,

$$\sqrt{n}(\overline{Y}_n - \mathbb{E}[Y_1]) \xrightarrow[n\to\infty]{d} \mathcal{N}_2(0, \Sigma).$$

where $\Sigma = \operatorname{Var} Y_1$. We take $g(u, v) = v - u^2$, so $g(\overline{Y}_n) = \overline{X_n^2} - \overline{X}_n^2 = \hat{\sigma^2}$. By the delta method,

$$\sqrt{n}\left(g(\overline{Y}_n) - g(\mathbb{E}[Y_1])\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \nabla g(\mathbb{E}[Y_1])^T \cdot \Sigma \cdot \nabla g(\mathbb{E}[Y_1])\right)$$

Theorem 7.4 (Multivariate Delta method)
Let $(T_n)_{n\geq 1} \in \mathbb{R}^d$ be a sequence of random vectors that satisfies

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, \Sigma)$$

for some $\theta \in \mathbb{R}^d$ and some symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$. Let $g : \mathbb{R}^d \to \mathbb{R}^k, k \geq 1$ be continuously differentiable at θ. Then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \nabla g(\theta)^T \sigma \nabla g(\theta)\right).$$

7.2 Likelihood functions

Let X_1 have a given density f_θ for some unknown $\theta \in \Theta$. How should we estimate θ? We should choose the density that maximizes the likelihood of X_1.

Example 7.5
Let $X_1 \sim \epsilon(\lambda)$. How do we select λ?

Suppose $g_\lambda(x) = \lambda e^{-\lambda x}$, for $x \geq 0$. Let $g_\lambda(X_1) = L(\lambda)$, where we fix X_1. Then the maximum likelihood estimate for λ is

$$\lambda^* = \arg\max_\lambda L(\lambda).$$

We write

$$L'(\lambda) = e^{-\lambda X_1}(1 - \lambda X_1) = 0$$

and find that $\lambda^* = 1/X_1$. We check that this point is indeed a maximum.

More generally, let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$. Then the joint density is $f_\theta(x_1), \ldots, f_\theta(x_n)$, which we evaluate at $x_1 = X_1$, etc.

Remark 7.6. It is standard to use upper case for random variables X and lower case for parameters x.

Definition 7.7. Let $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model associated with a sample of i.i.d. random variables $X_1, \ldots, X_n$. Assume that all the $\mathbb{P}_\theta$ have a density f_θ w.r.t. the Lebesgue measure ($\theta \in \Theta$).

The **likelihood** of the model is map L defined as:

$$L_n : E^n \times \Theta \to \mathbb{R}.$$

8 March 1, 2018

8.1 Continuous likelihood

We continue our discussion of likelihood functions. Recall that if $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$, then $\prod_i f_\theta(X_i)$ is the joint density.

Example 8.1

(x_1, x_2) has density given by

$$f_\theta(x_1) f_\theta(x_2), \forall x_1, x_2 \in E.$$

Note 8.2. Distribution and density are not synonymous. Bernoulli distribution, for instance.

Example 8.3

We have $X_1, \dots, X_{67} \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda)$. At $T = 500$, we observe all the machines and record their lifetimes. We observe $Y_i = \min(X_i, 500), i = 1, \dots, 67$.

The statistical model is

$$((0, 500], \mathbb{P}_\lambda)$$

where $\Pr_\lambda$ is the distribution of $\min(X, 500)$ for any $X \sim \epsilon(\lambda)$.

Example 8.4

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda), \lambda \in (0, 3)$. Find the likelihood function?

The statistical model is

$$\left((0, \infty), \{\epsilon(\lambda)\}_{\lambda \in (0,3)}\right).$$

Then the likelihood function is

$$\begin{aligned} L_n : (0, \infty)^n \times (0, 3) &\to \mathbb{R}. \\ (x_1, \dots, x_n, \lambda) &\mapsto \prod_i \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_i x_i}. \end{aligned}$$

Example 8.5

Suppose $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0, \theta]), \theta > 0$. Find the likelihood function?

The statistical model is

$$\left([0, \infty), \{\text{Uniform}([0, \theta])\}_{\theta > 0}\right).$$

Note that $[0, \theta]$ cannot be the domain because we do not know θ. Thus, we cannot define the sample space in terms of θ. If we know a priori that $\theta < c$, then we could write $[0, c)$ instead.

Then the likelihood function is

$$L_n : [0,\infty)^n \times (0,\infty) \to \mathbb{R}.$$
$$(x_1,\dots,x_n,\theta) \mapsto \prod_i (1/\theta)\mathbb{1}_{x_i\in[0,\theta]} = (1/\theta^n)\mathbb{1}_{\max_i x_i \le \theta}.$$

Example 8.6

Suppose $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$. Find the likelihood function.

It is

$$L_n : \mathbb{R}^n \times (\mathbb{R}\times(0,\infty)) \to \mathbb{R}.$$
$$(x_1,\dots,x_n,\theta) \mapsto \prod_i \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right).$$

8.2 Discrete likelihood

Definition 8.7. The **likelihood** of a discrete model is the map L_θ defined as

$$L_n : E^n \times \Theta \to \mathbb{R}$$
$$(x_1,\dots,x_n,\theta) \mapsto \Pr_\theta(X_1 = x_1,\dots,X_n = x_n).$$

Example 8.8

Suppose $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \lambda > 0$. Find the likelihood function.

The statistical model is

$$\left(\mathbb{N}, \{\text{Poisson}(\lambda)\}_{\lambda>0}\right).$$

The likelihood function is

$$L_n : \mathbb{N}^n \times (0,\infty) \to \mathbb{R}$$
$$(x_1,\dots,x_n,\theta) \mapsto \prod_i e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}.$$

Example 8.9

Suppose $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda), \lambda > 0$. Find the likelihood function.

The statistical model is

$$\left(\mathbb{N}, \{\text{Bernoulli}(p)\}_{p\in(0,1)}\right).$$

The likelihood function is

$$L_n : \{0,1\}^n \times (0,1) \to \mathbb{R}$$
$$(x_1,\dots,x_n,\theta) \mapsto \prod_i p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}.$$

8.3 Maximum likelihood estimator

Definition 8.10. The **maximum likelihood estimator** of θ is defined as

$$\hat{\theta}_n^{\text{MLE}} = \arg\max_{\theta \in \Theta} L_n(X_1, \ldots, X_n, \theta),$$

provided that it exists.

Remark 8.11. In practice, we use the log likelihood estimator

$$\hat{\theta}_n^{\text{MLE}} = \arg\max_{\theta \in \Theta} \log L_n(X_1, \ldots, X_n, \theta).$$

Example 8.12

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), p \in (0, 1)$.

Recall that the likelihood function is

$$L_n : (x_1, \ldots, x_n, \theta) \mapsto p^{\sum_i x_i}(1-p)^{n - \sum_i x_i}.$$

The MLE $\hat{p}^{\text{MLE}}$ is the value of $p \in (0, 1)$ that maximizes $L_n(X_1, \ldots, X_n, p)$. In this case, we find that $\hat{p}^{\text{MLE}} = \overline{X}_n$.

> "I know a lot of you in the physics department take the log of 0 and call it $-\infty$. We're in the math department here, and we do not do such horrible things."—veb

9 March 6, 2018

9.1 Fisher information

Recall that last class, we introduced maximum likelihood estimators. How do we evaluate this estimator? Sometimes it might make intuitive sense, but intuition may lead to useless, or even wrong results in mathematics.

Let us denote the log likelihood as

$$\ell(X, \theta) = \log L_1(X, \theta), \theta \in \Theta$$

where L_1 represents the likelihood for a single sample ($n = 1$). We assume that ℓ is twice differentiable.

Definition 9.1. The **Fisher information** of the statistical model is defined as

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ell}{\partial \theta \partial \theta'}(X, \theta) \right]$$

where θ' represents the transpose.

Note 9.2. If $\theta \in \mathbb{R}^d$, then the Fisher information is a d by d matrix.

Example 9.3

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), p \in (0, 1)$. Compute the Fisher information.

The L_1 likelihood is

$$\begin{aligned} L_1 : \{0, 1\} \times (0, 1) &\to \mathbb{R} \\ (x, p) &\mapsto p^x (1 - p)^{1-x}. \end{aligned}$$

We can define $\ell(X_1, p) = \log L_1(X_1, p)$ since our parameter space is an open interval,[6]

$$\ell(X_1, p) = X_1 \log p + (1 - X_1) \log(1 - p).$$

The second-order derivative is

$$\begin{aligned} \frac{\partial^2 \ell}{\partial p^2}(X_1, p) &= \frac{\partial \ell}{\partial p} \frac{X_1}{p} - \frac{1 - X_1}{1 - p} \\ &= -\frac{X_1}{p^2} - \frac{1 - X_1}{(1 - p)^2}. \end{aligned}$$

Now we take the expectation,

$$-\mathbb{E}_p \left[\frac{\partial^2 \ell}{\partial p^2}(X_1, p) \right] = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

Example 9.4

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda), \lambda > 0$. Compute the Fisher information.

[6] It is very important that the parameter space is an open interval!

The L_1 likelihood is

$$\begin{aligned} L_1 : (0,\infty) \times (0,\infty) &\to \mathbb{R} \\ (x,\lambda) &\mapsto \lambda e^{-\lambda x}. \end{aligned}$$

Since $e^{-\lambda x}$ is always positive,

$$\ell(X_1,\lambda) = \log\lambda - \lambda X_1$$

and we find that

$$\frac{\partial^2 \ell}{\partial \lambda^2}(X_1,\lambda) = -\frac{1}{\lambda^2}.$$

So the Fisher information is

$$I(\lambda) = -\mathbb{E}_\lambda\left[-\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

Example 9.5

Suppose $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$. Compute the Fisher information.

The L_1 likelihood is

$$\begin{aligned} L_1 : \mathbb{R}\times\mathbb{R}\times(0,\infty) &\to \mathbb{R} \\ (x,\lambda,\sigma^2) &\mapsto \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}). \end{aligned}$$

The log likelihood is

$$\ell(X_1,\mu,\sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X_1-\mu)^2.$$

Since our parameter is two-dimensional, we compute the Hessian, whose entries we bash out below.

$$\begin{aligned} \frac{\partial^2\ell}{\partial\mu^2}(X_1,\mu,\sigma^2) &= -\frac{1}{\sigma^2} \\ \frac{\partial\ell}{\partial(\sigma^2)}(X_1,\mu,\sigma^2) &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(X_1-\mu)^2 \\ \frac{\partial^2\ell}{\partial(\sigma^2)^2}(X_1,\mu,\sigma^2) &= \frac{1}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}(X_1-\mu)^2 \\ \frac{\partial^2\ell}{\partial\mu\partial(\sigma^2)}(X_1,\mu,\sigma^2) &= -\frac{X_1-\mu}{(\sigma^2)^2} \end{aligned}$$

Now we compute the expectations.

$$\begin{aligned} -\mathbb{E}_{\mu,\sigma^2}\left[\frac{\partial^2\ell}{\partial\mu^2}(X_1,\mu,\sigma^2)\right] &= \frac{1}{\sigma^2} \\ -\mathbb{E}_{\mu,\sigma^2}\left[\frac{\partial^2\ell}{\partial(\sigma^2)^2}(X_1,\mu,\sigma^2)\right] &= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\cdot\mathbb{E}_{\mu,\sigma^2}\left[(X_1-\mu)^2\right] = \frac{1}{2\sigma^4} \\ -\mathbb{E}_{\mu,\sigma^2}\left[\frac{\partial^2\ell}{\partial\mu\partial(\sigma^2)}(X_1,\mu,\sigma^2)\right] &= 0 \end{aligned}$$

Finally, the Fisher information is

$$I(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

9.2 Performance of MLE

We have good news! The MLE is both consistent and asymptotically normal.

Theorem 9.6

Let $\theta^* \in \Theta$ be the true parameter. Assume the following.

1. The model is identified.
2. $\forall \theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ.
3. θ^* is not on the boundary of Θ.
4. $I(\theta)$ is invertible in a neighborhood of θ^*.
5. A few more unmentioned.

Then $\hat{\theta}^{\text{MLE}}$ satisfies

$$\hat{\theta}^{\text{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^* \text{ w.r.t. } \mathbb{P}_{\theta^*}$$
$$\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^*) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta^*)^{-1}) \text{ w.r.t. } \mathbb{P}_{\theta^*}.$$

Generally, these conditions require that parameter space be open. Further details are irrelevant to this class. Now we provide some notes on the conditions.

- Suppose we considered the support of Bernoulli(p), for $p \in [0,1]$. Including the boundaries, the support is

$$\text{Support} = \begin{cases} \{0,1\} & p \in (0,1) \\ \{0\} & p = 0 \\ \{1\} & p = 1. \end{cases}$$

  Therefore, if we say that $p \in [0,1]$, then this theorem does not apply.

- Suppose we considered the support of Uniform$(0,\theta)$. The support is $[0,\theta]$, while the domain is $\mathbb{R}$, so we cannot apply the theorem here either.

Example 9.7

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

The maximum likelihood estimators are

$$\hat{\mu} = \overline{X}_n$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2.$$

Then we know that

$$\sqrt{n}\left[\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right] \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

10 March 7, 2018

10.1 Recitation 5

Consider a finite space $E = \{a_1, a_2, \ldots, a_r\}$ of size $r \geq 2$ and let X be a random variable taking values in E. For $j \in [r]$, let $p_j^* = \Pr\{X = a_j\}$, where $p_j^* > 0, \forall j$.

Consider a sample of n i.i.d. copies $X_1, \ldots, X_n$ of X. Based on this sample, we would like to estimate the multivariate parameter $p^* = (p_1^*, \ldots, p_r^*)$.

1. The parameter space is
$$\Theta = \left\{ \vec{p} = (p_1, \ldots, p_r) \in (0,1)^r \mid \sum_i p_i = 1 \right\}.$$

2. The likelihood is
$$\begin{aligned} L_n : E^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n, \vec{p}\,) &\mapsto \prod_i \Pr\{X_i = x_i\} \\ &= \prod_i p_1^{\mathbb{1}_{x_i = a_i}} \ldots p_r^{\mathbb{1}_{x_r = a_r}} \\ &= p_1^{\sum_i \mathbb{1}_{x_i = a_1}} \ldots p_r^{\sum_i \mathbb{1}_{x_i = a_r}}. \end{aligned}$$

3. To compute the maximum likelihood estimator $\hat{p}$, we use the method of Lagrange multipliers[7]
$$\mathcal{L}(\vec{p}, \lambda) = \log L_n(X_1, \ldots, X_n, \vec{p}) + \lambda \left(-1 + \sum_i p_i \right).$$
The first-order conditions say that
$$\frac{n_j}{p_j} + \lambda = 0, \forall j,$$
so $\lambda = -n_j/p_j$. So we see that $\hat{p_j} = n_j/n$.
A minor caveat is that we assume $\lambda \neq 0$, which assumes that $n_j > 0, \forall j$. The TA will check with the professor on this.

4. We show that $\hat{p}$ is asymptotically normal. Recall that
$$\hat{p_j} = \frac{n_j}{n} = \frac{1}{n} \sum_i \mathbb{1}_{x_i = a_j}.$$
Let Y_i be the vector of p_j's. By the multivariate central limit theorem,
$$\sqrt{n}(\hat{p} - \mathbb{E}[Y_i]) \xrightarrow[n \to \infty]{d} \mathcal{N}_\sigma(0, \Sigma)$$
where $\mathbb{E}[Y_i] = p$, and $\Sigma = \mathrm{Var}(Y_i)$, computed below.
The covariances are
$$\mathbb{E}\left[\mathbb{1}_{x_i = a_j} \mathbb{1}_{x_i = a_k}\right] - \mathbb{E}\left[\mathbb{1}_{x_i = a_j}\right] \mathbb{E}\left[\mathbb{1}_{x_i = a_k}\right] = -p_j p_k$$

[7] To maximize f subject to $g = 0$, we use $\mathcal{L} = f - \lambda g$.

and the variances are Bernoulli variances, $p(1-p)$. So the covariance matrix is

$$\operatorname{Var} Y_i = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & \\ -p_1p_2 & p_2(1-p_2) & \dots & \\ \dots & & & \\ \dots & & & \\ & & & p_r(1-p_r) \end{pmatrix}.$$

5. Σ is not invertible since it has a nontrivial kernel (vector of ones), so the theorem for maximum likelihood cannot be applied here.

11 March 8, 2018

11.1 Fisher information and MLE

First we make some remarks on supports. Recall that the support is a subset of the sample space.

Example 11.1

Suppose we have variables $X_i \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda)$. We may write the likelihood in two ways, depending on the domain.

If we let $E = R$, that is, the sample space is the real line,

$$\begin{aligned} L_n : \mathbb{R}^n \times (0,\infty) &\to \mathbb{R} \\ (x_1, \dots, x_n, \lambda) &\mapsto \prod_i \lambda e^{-\lambda x_i} \mathbb{1}_{x_i > 0}. \end{aligned}$$

However, if we take $E = (0,\infty)$, then we redefine the likelihood with a different domain, but may omit the indicator.

$$\begin{aligned} L_n : (0,\infty)^n \times (0,\infty) &\to \mathbb{R} \\ (x_1, \dots, x_n, \lambda) &\mapsto \prod_i \lambda e^{-\lambda x_i}. \end{aligned}$$

We see here that if the support does not depend on Θ, we can change the domain to match the support without changing the likelihood.

As a result, in the continuous case we may write

$$\begin{aligned} L_n : E^n \times \Theta &\to \mathbb{R} \\ (x_1, \dots, x_n, \theta) &\mapsto \prod_i f_\theta(x_i) \end{aligned}$$

since $f_\theta(x_i) > 0$ by definition of support.

Proposition 11.2

$I(\theta) = \operatorname{Var}_\theta(\nabla_\theta \ell(X, \theta))$.

That is, the Fisher information is equal to the covariance matrix of the gradient of the log likelihood.

Proof. Without loss of generality (and with gain of clarity), suppose $\Theta \subseteq \mathbb{R}$. Recall that

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log L_1}{\partial \theta^2}(x, \theta) \right].$$

Note that $\int_E L_1(x,\theta)dx = 1, \forall \theta \in \Theta$, so

$$\begin{aligned} \frac{\partial}{\partial \theta} \int_E L_1(x,\theta)dx &= \int_E \frac{\partial L_1}{\partial \theta}(x,\theta)dx \\ &= \int_E \frac{\partial \log L_1}{\partial \theta}(x,\theta) L_1(x,\theta)dx \\ &= \mathbb{E}_\theta \left[\frac{\partial \log L_1}{\partial \theta}(X_1, \theta) \right] = 0. \end{aligned}$$

So we have shown that $\mathbb{E}_\theta\left[\frac{\partial \log L_1}{\partial \theta}(X_1,\theta)\right] = 0, \forall\theta$. Now we differentiate again with respect to θ to obtain

$$\begin{aligned}\int_E \frac{\partial}{\partial\theta}\left(\frac{\partial \log L_1}{\partial\theta}L_1\right)(x,\theta)dx &= \int_E \left(\frac{\partial^2 \log L_1}{\partial\theta^2}L_1 + \frac{\partial \log L_1}{\partial\theta}\frac{\partial L_1}{\partial\theta}\right)(x,\theta)dx \\ &= \mathbb{E}_\theta\left[\frac{\partial^2 \log L_1}{\partial\theta^2}(X_1,\theta)\right] + \mathbb{E}_\theta\left[\left(\frac{\partial \log L_1}{\partial\theta}(X_1,\theta)\right)^2\right] = 0.\end{aligned}$$

Therefore, we can say that

$$\mathbb{E}_\theta\left[\left(\frac{\partial \log L_1}{\partial\theta}(X_1,\theta)\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log L_1}{\partial\theta^2}(X_1,\theta)\right] = -I(\theta).$$

Since the random variable in the first term is centered, the expectation of its square is its variance. Finally,

$$\mathrm{Var}\left[\frac{\partial \log L_1}{\partial\theta}(X_1,\theta)\right] = -I(\theta).$$

□

> "It's a dummy variable. You can call it 'computer' or 'bug' but I call it x because it sounds less silly"—veb

11.2 Limitations of MLE

There are some limitations to consider.

- Often, there is no closed form solution for the MLE. For example, consider $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \text{Cauchy}(\alpha), \forall x \in \mathbb{R}$.

  $$f_\theta = \frac{1}{\pi}\frac{1}{(x-\alpha)^2+1}$$

  But there is no solution for the MLE.

- There are numerical methods, such as the Newton-Raphson algorithm and expectation-maximization (EM). These don't always work.

- The MLE is not always robust (e.g. contaminated samples, outliers).

11.3 Method of moments

We may consider complicated models in which the MLE has no closed form. How can we provide consistent, asymptotically normal estimators? Let's look at some examples.

- $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \lambda > 0$. Then $\hat{\lambda} = \overline{X}_n$.

- $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda), \lambda > 0$. Then $\hat{\lambda} = 1/\overline{X}_n$.

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. For a Gaussian, we need the first moment $\mathbb{E}[X_1] = \mu$ and the second moment $\mathbb{E}\left[X_1^2\right] = \mu^2 + \sigma^2$. Then

$$\begin{aligned}\mu &= \mathbb{E}[X_1] \\ \sigma^2 &= \mathbb{E}\left[X_1^2\right] - \mathbb{E}[X_1]^2.\end{aligned}$$

  From these, we can craft estimators

$$\begin{aligned}\hat{\mu} &= \overline{X}_n \\ \hat{\sigma}^2 &= \overline{X_n^2} - \overline{X}_n^2.\end{aligned}$$

Note that if we convert all parameters into some expectation, then we can provide consistent estimators by the law of large numbers.

Formally, let $X_1, \ldots, X_n$ be an i.i.d. sample associated with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$. We assume that $E \subseteq \mathbb{R}$ and $\theta \subseteq \mathbb{R}^d, d \geq 1$.

Definition 11.3. A **population moment** is $m_k(\theta) = \mathbb{E}_\theta\left[X_1^k\right], 1 \leq k \leq d$.

Definition 11.4. An **empirical moment** is

$$\hat{m}_k(\theta) = \overline{X_n^k} = \frac{1}{n} = \sum_i X_i^k, 1 \leq k \leq d.$$

Let M be a vector of the first d moments,

$$\begin{aligned}M : \Theta &\to \mathbb{R}^d \\ \theta &\mapsto (m_1(\theta), \ldots, m_d(\theta)).\end{aligned}$$

If we assume that M is bijective, we can reconstruct θ as

$$\theta = M^{-1}(m_1(\theta), \ldots, m_d(\theta)).$$

Definition 11.5. The **moment estimator** of θ is

$$\hat{\theta}_n^{\text{MM}} = M^{-1}(\hat{m}_1, \ldots, \hat{m}_d)$$

provided that it exists.

Example 11.6

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), p \in (0, 1)$.

We compute the first moment $\mathbb{E}_p[X_1] = p$, so $p = \mathbb{E}_p[X_1]$, so $\hat{p} = \overline{X}_n$.

12 March 14, 2018

Happy π day! Tomorrow is the exam, let's do a lot of practice problems today.

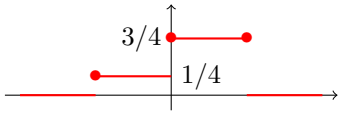
12.1 Recitation 6

Let $X_1, \dots, X_n$ be i.i.d. random variables with density f_θ given by

$$f_\theta(x) = \begin{cases} \theta & \text{if } -1 \le x < 0 \\ 1-\theta & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is unknown.

1. Plot the function $f_{1/4}$.



   f_θ is the density of Uniform($[-1, 1]$) for $\theta = 1/2$.

2. The likelihood is

$$\begin{aligned} L_n : [-1,1]^n \times (0,1) &\to \mathbb{R} \\ (x_1, \dots, x_n, \theta) &\mapsto \prod_i f_\theta(x_i) \\ &= \prod_i \theta^{\mathbb{1}_{x_i<0}} (1-\theta)^{\mathbb{1}_{x_i \ge 0}}. \end{aligned}$$

3. We find the MLE. From above,

$$L_n(x_1, \dots, x_n, \theta) = \theta^{N_-}(1-\theta)^{N_+}$$

   where N_- is the number of negative samples and N_+ is the number of positive samples. We maximize $\log L_n$,

$$\log L_n(x_1, \dots, x_n, \theta) = \log \theta^{N_-} + \log(1-\theta)^{N_+}$$

   whose derivative is set to 0:

$$\frac{\partial}{\partial \theta} \log L_n(x_1, \dots, x_n, \theta) = \frac{N_-}{\theta} - \frac{N_+}{1-\theta} = 0.$$

   So the MLE is equal to the proportion of negative numbers in the sample,

$$\hat{\theta} = \frac{N_-}{n}.$$

4. Using the central limit theorem, show that $\hat{\theta}$ is asymptotically normal.

   We redefine

$$\hat{\theta} = \frac{N_-}{n} = \frac{1}{n} \sum_i \mathbb{1}_{x_i<0}$$

   as a sample average. Then by the central limit theorem,

$$\sqrt{n}(\hat{\theta} - \mathbb{E}\left[\mathbb{1}_{X_1<0}\right]) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \theta(1-\theta))$$

   since the indicator is a Bernoulli with parameter θ.

5. We find the Fisher information. By definition,

$$\ell(X_1, \theta) = \mathbb{1}_{X_1<0} \log \theta + \mathbb{1}_{X_1 \geq 0}(1-\theta).$$

We take the second derivative,

$$\frac{\partial^2}{\partial \theta^2} \ell(X_1, \theta) = \frac{-\mathbb{1}_{X_1<0}}{\theta^2} - \frac{\mathbb{1}_{X_1 \geq 0}}{(1-\theta)^2}.$$

Finally, we find the expectation,

$$-\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ell(X_1, \theta)\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)},$$

which is equal to the Fisher information.

This quantity is equal to the inverse of the asymptotic variance for the MLE, found above.

6. The first moment is

$$\begin{aligned}\mathbb{E}[X_1] &= \int_{\mathbb{R}} x f_\theta(x) dx \\ &= \theta \int_{-1}^{0} x dx + (1-\theta) \int_0^1 x dx \\ &= \frac{1}{2} - \theta\end{aligned}$$

so $\theta = \frac{1}{2} - \mathbb{E}[X_1]$, and our estimator is $\tilde{\theta} = \frac{1}{2} - \overline{X}_n$

7. We prove that $\tilde{\theta}$ is asymptotically normal. By the central limit theorem,

$$\sqrt{n}(\overline{X}_n - \mathbb{E}[X_1]) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1).$$

We calculated the expectation above, and the variance is

$$\operatorname{Var} X_1 = \mathbb{E}\left[X_1^2\right] - \mathbb{E}[X_1]^2 = \frac{1}{12} + \theta(1-\theta).$$

So $\hat{\theta}$ has the smaller asymptotic variance.

8. For $\hat{\theta}$, the quadratic risk is

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] &= \operatorname{Bias}^2(\theta) + \operatorname{Var}(\theta) \\ &= 0 + \frac{1}{n} \cdot \theta(1-\theta)\end{aligned}$$

since $\hat{\theta}$ converges to θ.

For $\tilde{\theta}$, the quadratic risk is

$$\mathbb{E}\left[(\tilde{\theta} - \theta)^2\right] = 0 + \frac{1}{n}\left(\frac{1}{12} + \theta(1-\theta)\right).$$

So the risk for $\hat{\theta}$ is smaller.

9. Now we provide a confidence interval using $\hat{\theta}$.

$$\sqrt{n}\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Since $\hat{\theta}(1-\hat{\theta}) \xrightarrow[n\to\infty]{d} \theta(1-\theta)$, by Slutsky's theorem,

$$\sqrt{n}\frac{\hat{\theta}-\theta}{\sqrt{\hat{\theta}(1-\hat{\theta})}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

So we provide a confidence interval

$$\mathcal{I} = [xxx]$$

13 March 20, 2018

The professor is sad that the graders are too nice, so he will randomly select 10 problem sets every week to grade by himself.

> "Those who are selected will be unlucky, since the grades will be very low—but the problem sets are worth 1, and the midterm is worth 10!"—veb

We had an exam last Thursday and a snow day the previous Tuesday, so now we return to our wonderful method of moments.

13.1 Method of moments theorem

Recall that the moment estimator is

$$\hat{\theta} = M^{-1}\left(\begin{pmatrix} \overline{X}_n & \overline{X_n^2} & \dots & \overline{X_n^d} \end{pmatrix}^T\right),$$

where M is an invertible function. By the central limit theorem, each moment is asymptotically normal,

$$\begin{aligned}
\sqrt{n}\left(\overline{X}_n - m_1\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1) \\
\sqrt{n}\left(\overline{X_n^2} - m_2\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1^2) \\
&\vdots \\
\sqrt{n}\left(\overline{X_n^d} - m_d\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1^d).
\end{aligned}$$

> "I'm not asking you to know the central limit theorem perfectly. I'm asking you to master it, which is even more. You must be able to write these lines with your eyes closed."—veb

However, it is insufficient to state that each of the moments is asymptotically normal. Rather, we should use the multivariate central limit theorem,

$$\sqrt{n} \cdot \left(\begin{pmatrix} \overline{X}_n \\ \overline{X_n^2} \\ \vdots \\ \overline{X_n^d} \end{pmatrix} - \mathbb{E}\left[\begin{pmatrix} X_1 \\ X_1^2 \\ \vdots \\ X_1^d \end{pmatrix}\right]\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\vec{0}, \Sigma\right)$$

where Σ is the covariance matrix $\Sigma = \operatorname{Var}\begin{pmatrix} \overline{X}_n & \overline{X_n^2} & \dots & \overline{X_n^d} \end{pmatrix}^T$ with entries

$$\Sigma_{i,j} = \operatorname{cov}\left(X_1^i, X_1^j\right) = \mathbb{E}\left[X_1^{i+j}\right] - \mathbb{E}\left[X_1^i\right]\mathbb{E}\left[X_1^j\right].$$

Now we can use the multivariate delta method to obtain our theorem.

Theorem 13.1

Let $\hat{\theta}_n^{MM}$ be the estimator obtained by the method of moments. Then

$$\sqrt{n}\left(\hat{\theta}_n^{MM} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Gamma(\theta))$$

where

$$\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta} M(\theta)\right]^T \Sigma(\theta) \left[\frac{\partial M^{-1}}{\partial \theta} M(\theta)\right].$$

13.2 Generic method of moments

Unfortunately, the method of moments does not always for a simple reason: M may not always have an inverse.

Example 13.2

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta(x)$ where

$$f_\theta(x) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}}\exp(-(x-\theta)^2/2) + \frac{1}{\sqrt{2\pi}}\exp(-(x+\theta)^2/2)\right)$$

This is a mixture of two Gaussians.

First, we might want to find the MLE, but will will quickly realize that the likelihood is disgusting. Then, we will try for the method of moments, so we will compute the first moment (since there is only one parameter θ). However, the first moment is 0! That is totally not bijective.

Instead we can use the second moment,

$$\mathbb{E}\left[X_1^2\right] = \theta^2 + 1,$$

so we may estimate that $\hat{\theta} = \sqrt{\overline{X_n^2} - 1}$.

Example 13.3

Suppose we want to determine the distribution of eye colors in Boston, and our sample space is

$$X_i \in \{\text{blue}, \text{green}, \text{brown}\}.$$

If $X_1, \ldots, X_n$ are iid, then we would like to estimate the probability of each eye color.

If we use the old method of moments, we write the first four moments...

$$\mathbb{E}\left[X_1\right] =$$
$$\mathbb{E}\left[X_1^2\right] = \ldots$$

but wait, X_1 is a color. Can we square a color? How about no. Instead, let's look at these friendly moments.

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{X_i=\text{blue}}\right] &= p_{\text{blue}}\\ \mathbb{E}\left[\mathbb{1}_{X_i=\text{green}}\right] &= p_{\text{green}}\\ \mathbb{E}\left[\mathbb{1}_{X_i=\text{brown}}\right] &= p_{\text{brown}}\end{aligned}$$

Then our estimators are very intuitive!

$$\begin{aligned}\hat{p}_{\text{blue}} &= \frac{1}{n}\sum_i \mathbb{1}_{X_i=\text{blue}} = \frac{\#\text{ blue}}{n}\\ \hat{p}_{\text{green}} &= \frac{1}{n}\sum_i \mathbb{1}_{X_i=\text{green}} = \frac{\#\text{ green}}{n}\\ \hat{p}_{\text{brown}} &= \frac{1}{n}\sum_i \mathbb{1}_{X_i=\text{brown}} = \frac{\#\text{ brown}}{n}\end{aligned}$$

Note 13.4. The estimators obtained by the method of moments are often multidimensional, so we cannot apply d separate central limit theorems to show that the *estimator* is asymptotically normal; we should apply the multivariate central limit theorem.

More generally, let $g_1, \ldots, g_d : E \to \mathbb{R}$ be some functions, chosen for convenience.[8] Let

$$\begin{aligned}m_k(\theta) &= \mathbb{E}_\theta\left[g_k(X)\right], \forall k = 1, \ldots, d\\ \Sigma(\theta) &= \text{Var}(g_1(X_1), g_2(X_1), \ldots, g_d(X_1))\end{aligned}$$

Then [TODO]

Compared to the moment estimator, the MLE is often more accurate, but it may also be intractable. Therefore, each has tradeoffs.

13.3 M-estimators

Let $X_1, \ldots, X_n$ be iid for some unknown distribution P in some sample space $E \subseteq \mathbb{R}^d, d \geq 1$. We need not assume a statistical model. Suppose we want to estimate some parameter μ^* associated with P (e.g. mean, variance, etc.).

- The mean and variance are easy. We just take the sample mean or variance, and we can use the central limit theorem to show that they are consistent.
- However, what if we wanted to estimate the median? We could consider the sample median, but it's a bit harder to show consistency.

In general, we want to find a function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values for the known μ^*, such that

$$\mathcal{Q}(\mu) \coloneqq \mathbb{E}\left[\rho(X_1, \mu)\right]$$

achieves its minimum at $\mu = \mu^*$.

[8] Previously, we let $g_k(x) = x^k$, but this need not be the case.

Example 13.5

$X_1, \ldots, X_n$ have mean μ.

We take $\rho(x, \mu) = (x - \mu)^2$, and

$$\mathcal{Q}(u) = \mathbb{E}\left[\rho(X_1, \mu)\right] = \mathbb{E}\left[(X_1 - \mu)^2\right] = \mathbb{E}\left[X_1^2\right] - 2\mu\mathbb{E}\left[X_1\right] + \mu^2.$$

This is a parabola with its minimum at μ. We see that $\mathcal{Q}$ depends on an expectation, but it's very easy to estimate $\mathcal{Q}$,

$$\hat{\mathcal{Q}}(\mu) = \frac{1}{n}\sum_i \rho(X_i, \mu)$$

which converges by the law of large numbers. As desired, the μ that minimizes this function is the sample mean![9]

If we instead take $\rho(x, \mu) = |x - \mu|$, then

$$\mathcal{Q}(u) = \mathbb{E}\left[\rho(X_1, \mu)\right] = \mathbb{E}\left[|X_1 - \mu|\right],$$

which is minimized when μ is a median of X_1.[10]

If we want to estimate the α-quantile instead, take $\rho(x, \mu) = C_\alpha(x - \mu)$, where C_α is the check function

$$C_\alpha = \begin{cases} -(1-\alpha) & \text{if } x < 0 \\ \alpha x & \text{if } x \geq 0. \end{cases}$$

[9] If we have d-dimensional vectors, then we replace this with the Euclidean norm, and the estimator still converges to the mean.

[10] Medians are not always uniquely defined. Consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$.

14 March 21, 2018

14.1 Recitation 7

Let $X_1, \dots, X_n$ be iiid continuous random variables with density

$$f_{\lambda,\theta}(x) = Cx^{\lambda-1}\mathbb{1}_{0\le x\le\theta}$$

for all $x \in \mathbb{R}$, where $\lambda, \theta > 0$ are unknown parameters and C is some positive constant.

1. Find the value of C. Proper densities must integrate to 1.

$$\begin{aligned}\int_0^\theta Cx^{\lambda-1}dx &= 1\\ C\theta^\lambda\lambda^{-1} &= 1\\ C &= \lambda\theta^{-\lambda}\end{aligned}$$

   So our density is

$$f_{\lambda,\theta}(x) = \lambda\theta^{-\lambda}x^{\lambda-1}\mathbb{1}_{0\le x\le\theta}.$$

2. Is the parameter (λ, θ) identified?

   - θ is the rightmost boundary of the support of $f_{\lambda,\theta}$.
   - λ is also identified: consider $f(\theta)$. Then the density evaluates to λ/θ, so we can find λ.

   So the pair is identified.

3. Define the likelihood function.

$$\begin{aligned}&L_n : [0,\infty)^n \times (0,\infty)^2 \to \mathbb{R}\\ &(x_1,\dots,x_n,(\lambda,\theta)) \mapsto \lambda^n\theta^{-n\lambda}\left(\prod_i x_i\right)^{\lambda-1}\mathbb{1}_{\max_i x_i\le\theta,\min_i x_i\ge 0}\end{aligned}$$

4. Compute the maximum likelihood estimator $(\hat\lambda, \hat\theta)$.

   Upon inspection, we see that $\hat\theta = \max_i X_i$ (that's how we showed that θ is identified).

   Without loss of generality, assume $0 \le x_i \le \theta, \forall i$. The density is positive and continuous here, so

$$\log L_n(x_1,\dots,x_n,(\lambda,\theta)) = n\log\lambda - n\lambda\log\theta + (\lambda-1)\sum_i \log x_i.$$

   The derivative with respect to λ is

$$\frac{\partial}{\partial\lambda}\log L_n(x_1,\dots,x_n,(\lambda,\theta)) = n/\lambda - n\log\theta + \sum_i \log x_i.$$

   When we set the derivative to 0,

$$\hat\lambda = \left(-\frac{1}{n}\sum_i \log\frac{x}{\max_i x_i}\right)^{-1}.$$

5. We observe that $\hat{\theta}$ is not asymptotically normal. Intuitively, this is because $\hat{\theta}$ is biased. There's a more mathy way the TA showed but I was busy doing number 7 oops.

6. Show that $\hat{\lambda}$ is asymptotically normal.

   By the central limit theorem,

$$\sqrt{n}\left(\sum_i \log\frac{X_i}{\theta} - \mathbb{E}\left[\sum_i \log\frac{X_i}{\theta}\right]\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V).$$

   Let $Y = \sum_i \log X_i/\theta$. By lots of disgusting calculus,

$$\begin{aligned}\mathbb{E}[Y] &= -\frac{1}{\lambda}\\ \mathbb{E}\left[Y^2\right] &= \frac{2}{\lambda^2}\\ V &= \frac{1}{\lambda^2}.\end{aligned}$$

   Now we might not be too interested in an expression that depends on θ, but we can express this in terms of $\hat{\theta}$ instead.

$$\sqrt{n}\left(\sum_i \log\frac{X_i}{\hat{\theta}} - +\frac{1}{\lambda}\right) = \sqrt{n}\left(\sum_i \log\frac{X_i}{\theta} - +\frac{1}{\lambda}\right) + \sqrt{n}\log\frac{\theta}{\hat{\theta}}.$$

   Since $\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta$, we know that $\log\theta/\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} 0$, so by Slutsky's theorem,

$$\sqrt{n}\left(\sum_i \log\frac{X_i}{\hat{\theta}} - +\frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V).$$

7. Find the estimator $(\tilde{\lambda}, \tilde{\theta})$ obtained by the method of moments.

$$\begin{aligned}\mathbb{E}[X_1] &= \frac{\lambda\theta}{1+\lambda}\\ \mathbb{E}\left[X_1^2\right] &= \frac{\lambda\theta^2}{2+\lambda}\end{aligned}$$

   Solving for $(\tilde{\lambda}, \tilde{\theta})$,

$$\begin{aligned}\tilde{\lambda} &= -1 + \frac{\overline{X_n^2}}{\sqrt{\overline{X_n^2}^2 - \overline{X_n^2}\cdot\overline{X}_n^2}}\\ \tilde{\theta} &= \frac{\overline{X_n^2} + \sqrt{\overline{X_n^2}^2 - \overline{X_n^2}\cdot\overline{X}_n^2}}{\overline{X}_n}.\end{aligned}$$

8. Show that $(\tilde{\lambda}, \tilde{\theta})$ is asymptotically normal.

   I'm lazy.

9. Which of $\hat{\lambda}$ has the smallest asymptotic variance?

   I'd assume $\hat{\lambda}$ but I honestly don't know cuz we didn't get here lol.

10. Confidence interval for λ

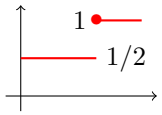
11. I wanna go home already and play phone games omo.

15 March 22, 2018

15.1 M-estimators, ctd.

A few notes from last time.

- Often there may be multiple meadians. For example, consider $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(1/2).



$$\mathcal{Q}(u) = \mathbb{E}\left[|X_1 - \mu|\right] = \frac{1}{2}(1-\mu) + \frac{1}{2}\mu = \frac{1}{2}.$$

Since $\mathcal{Q}$ is constant, any value between $[0, 1)$ is a valid median.

Theorem 15.1

Let $\mathcal{M} = \Theta$ and $\rho(x, \theta) = -\log L_1(x, \theta)$, provided the likelihood is positive everywhere. Then

$$\mu^* = \theta^*$$

where $\mathbb{P} = \mathbb{P}_{\theta^*}$.

Make this a cohesive story.

In general, we would like to find ρ such that

$$\mathcal{Q}(u) = \mathbb{E}\left[\rho(X_1, \mu)\right]$$

is minimized when $\mu = \theta^*$. We take $\rho(X_1, \theta) = -\log L_1(X_1, \theta)$ if ρ is well-defined (that is, L_1 is positive).

Lemma 15.2

$\mathcal{Q}(\theta) - \mathcal{Q}(\theta^*) \geq 0, \forall \theta \in \Theta$

Proof.

$$\begin{aligned}\mathcal{Q}(\theta) - \mathcal{Q}(\theta^*) &= -\mathbb{E}\left[\log L_1(X_1, \theta)\right] + \mathbb{E}\left[\log L_1(X_1, \theta^*)\right] \\ &= -\mathbb{E}\left[\log \frac{L_1(X_1, \theta)}{L_1(X_1, \theta^*)}\right]\end{aligned}$$

where the latter line follows from linearity of expectation and properties of the logarithm. By Jensen's inequality,

$$\mathbb{E}\left[\log \frac{L_1(X_1, \theta)}{L_1(X_1, \theta^*)}\right] \leq \log \mathbb{E}\left[\frac{L_1(X_1, \theta)}{L_1(X_1, \theta^*)}\right].$$

Let us examine the expectation. In the continuous case, $L_1(X_1, \theta) = f_\theta(X_1)$, where f_θ is the density of $\mathbb{P}_\theta$. So

$$\begin{aligned}\mathbb{E}\left[\frac{L_1(X_1,\theta)}{L_1(X_1,\theta^*)}\right] &= \mathbb{E}\left[\frac{f_\theta(X_1)}{f_{\theta^*}(X_1)}\right] \\ &= \int_E \frac{f_\theta(x)}{f_{\theta^*}(x)} \cdot f_{\theta^*}(x)dx = 1\end{aligned}$$

since f_θ integrates to 1 over its support. Therefore,

$$\mathcal{Q}(\theta) - \mathcal{Q}(\theta^*) = -\mathbb{E}\left[\log \frac{L_1(X_1,\theta)}{L_1(X_1,\theta^*)}\right] \geq 0.$$

□

Let $\hat{\mu}_n$ be a minimizer of the form

$$\mathcal{Q}_n(\theta) \coloneqq \frac{1}{n}\sum_i \rho(X_1, \mu).$$

Note that above, the minimizer for θ^* is

$$\begin{aligned}\min_\theta -\frac{1}{n}\sum_i \log L_1(X_1,\theta) \Leftrightarrow \max_\theta \sum_i \log L_1(X_1,\theta) \\ &= \log \prod_i L_1(X_i,\theta) \\ &= L_n(X_1,\ldots,X_n,\theta).\end{aligned}$$

Therefore, the M-estimator for $\rho = \log L_1$ is simply the MLE!

15.2 M-estimator asymptotics

Now that we have seen that the MLE is just a special M-estimator, we might recall that the MLE converges asymptotically in many cases. Are there similar guarantees for M-estimators in general?

Let

$$\begin{aligned}J(\mu) &= -\frac{\partial^2 \mathcal{Q}}{\partial\mu\partial\mu^T}(\mu) \\ K(\mu) &= \operatorname{Var}\frac{\partial\rho}{\partial\mu}(X_1,\mu)\end{aligned}$$

where J is equivalent to $-\mathbb{E}[xxx]$ under some regularity conditions.

Remark 15.3. In the log-likelihood case, where $\mu = \theta$,

$$J(\theta) = K(\theta) = I(\theta),$$

the Fisher information. In the general case, J might not equal K.

Theorem 15.4

Let $\mu^* \in \mathcal{M}$ be the true parameter. Assume the following:

1. $\mathcal{M}$ is an open set.
2. μ^* is the only minimizer of the function $\mathcal{Q}$.
3. $J(\mu)$ is invertible $\forall \mu \in \mathcal{M}$.
4. A few other unmentioned conditions.

Then $\hat{\mu}_n$ satisfies

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$$
$$\sqrt{n}(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, J(\mu^*)^{-1}K(\mu^*)J(\mu^*)^{-1}).$$

Let's discuss these conditions.

- We have seen throughout the course so far that open sets are love for parameters.
- When we discussed the median, we recognized that the median may not be unique. Thus, the median violates condition (2).

Jensen's inequality only holds if $L_1(X_1, \theta)/L_1(X_1, \theta^*)$ is constant (deterministic), which is only possible if $L_1(x, \theta)/L_1(x, \theta^*)$ is a constant function. That is,

$$L_1(X_1, \theta) = cL_1(X_1, \theta^*), \forall x \in E.$$

However, both are proper densities, so $c = 1$. Then

$$\begin{aligned} &L_1(x, \theta) = L_1(x, \theta^*), \forall x \in E \\ \Leftrightarrow\ &f_\theta(x) = f_\theta(x), \forall x \in E \\ \Leftrightarrow\ &f_\theta = f_{\theta^*} \\ \Leftrightarrow\ &\mathbb{P}_\theta = \mathbb{P}_{\theta^*}. \end{aligned}$$

In the theorem for the MLE, we required that θ be identified. Here, we see that this condition allows us to argue that θ^* is the unique minimizer.

15.3 M-estimators for robust statistics

Example 15.5

Suppose we wanted to know a population's average monthly salary. But people are, after all, human! We may ask for monthly salaries in a survey, and someone's bound to put a weekly or annual salary—but we don't know! They could just be real rich or real poor.

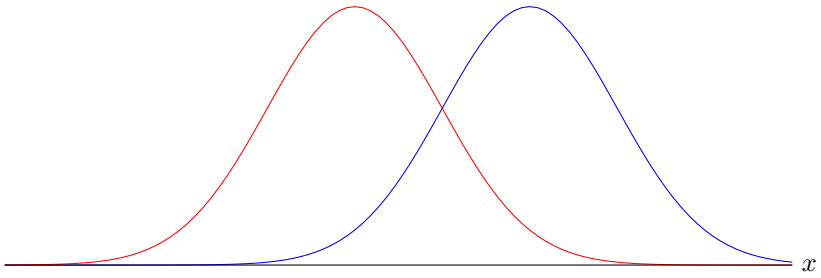
In general, suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot - m)$, where

- f is an unknown, positive, even function (e.g. Cauchy)

- m is a **location parameter** (real number of interest)

how would we estimate m?

We could use M-estimators for empirical mean and median and compare their risks or asymptotic variances. The empirical median is quite robust, especially in the case of corrupted monthly salaries.



15.4 Parametric hypothesis testing

> "I'll tell you about hypothesis testing before you leave for spring break so you don't think you came for nothing. Everyone's heard of hypothesis testing, but you've been here for two months and I've told you nothing about hypothesis testing."—veb

Let's start with some examples to gain intuition.

Example 15.6

We run a medical lab that tests patients for HIV. People with HIV have extra antibodies, so we sample some blood and check the concentration of antibodies.

We should dilute out many samples, where each sample

$$X_i \sim \mathcal{N}(\xi, \sigma^2)$$

and we check if $\xi > \xi_0$, our threshold.

There are tow types of errors we could make:

- **False negative**: The patient is sick and we send them home. This is VERY dangerous! The patient will die.
- **False positive**: The patient is fine and we treat the patient, which isn't great, but not fatal.

Example 15.7

We toss a coin 80 times and obtain 54 heads. Can we conclude that the coin is significantly unfair?

Formally, for $n = 80, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, and $\overline{X}_n = 56/80 = 0.58$. If it were true that $p = 0.5$, then

$$\sqrt{n}\frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(\overline{X}_n)}} \approx \mathcal{N}(0, 1).$$

Of course, this is very handwavy, so we are NOT allowed to write this on the test. Here,

$$\sqrt{n}\frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(\overline{X}_n)}} \approx 3.45,$$

so it seems pretty reasonable to reject the hypothesis $p = 0.5$.

Here,

$$\sqrt{n}\frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(\overline{X}_n)}} \approx 0.77,$$

it *seems* impossible to reject significantly the hypothesis $p = 0.5$.

The only things we can do are invalidate, or say we cannot invalidate. We cannot validate.

16 April 3, 2018

Recall from the first class that we will discuss three ways of statistical inference.

1. Estimation produces an approximate value.
2. Confidence intervals find a range.
3. Hypothesis tests answer a question. For example, "is it true that θ_1 and θ_2 are significantly different?"

Our next chapter, we focus on the last point.

16.1 Hypothesis testing

Consider a sample $X_1, \ldots, X_n$ of i.i.d. random variables and a statistical model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$.

Let Θ_0 and Θ_1 be two disjoint subsets of Θ. Consider the two hypotheses,

$$\begin{cases} H_0 & \theta \in \Theta_0 \\ H_1 & \theta \in \Theta_1 \end{cases}$$

where H_0 is known as the **null hypothesis** and H_1 is the **alternative hypothesis**. We will always be able to invalidate the null hypothesis H_0 or conclude that we cannot invalidate it. In this class we do not care about H_1.

Example 16.1

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. We would like to test if $p = 1/2$.

The hypotheses are

$$H_0 : p = 1/2 \qquad H_1 : p \neq 1/2$$

and the parameter sets are

$$\Theta_0 : \{1/2\} \qquad \Theta_1 : (0,1) \setminus \{1/2\}.$$

Example

Now suppose we want to know if the coin leans either way. The professor wins if the coin is $p \geq 1/2$.

The hypotheses are

$$H_0 : p \geq 1/2 \qquad H_1 : p < 1/2$$

and the parameter sets are

$$\Theta_0 : [1/2, 1) \qquad \Theta_1 : (0, 1/2).$$

Example

Suppose now that our professor is a smart opponent, and he would never play with a coin for which $p < 1/2$.

More reasonable hypotheses are

$$H_0 : p = 1/2 \qquad H_1 : p < 1/2$$

and the parameter sets are

$$\Theta_0 : \{1/2\} \qquad \Theta_1 : (0, 1/2)$$

which do not cover the whole parameter space, but that's fine.

Example 16.2 (HIV testing)

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ represents the concentration of antibodies.

The hypotheses are

$$H_0 : \mu \geq \mu_0 \qquad H_1 : \mu < \mu_0$$

and the parameter sets are

$$\Theta_0 : [\mu_0, \infty) \times (0, \infty) \qquad \Theta_1 : (-\infty, \mu_0) \times (0, \infty).$$

Don't forget to include σ^2 in the parameter space.

There is inherent asymmetry between H_0 and H_1. So how do we assign H_0 and H_1? Recall our discussion about false positives and false negatives.

In the case that H_0 is true, it is very dangerous or costly to reject H_0. For example, false negatives are very dangerous for HIV testing since sick patients are sent home. Therefore, the null hypothesis should be "the patient is sick." We do *not* want to reject this, but it's okay to reject that "the patient is healthy."

> **Question 16.3.** What's the use of defining H_1 if we know nothing about it? Short answer: no use in this class. Long answer: it would still be ideal to control false positives (patient is healthy and we treat them) since there may be nontrivial costs. However, false negatives are much more dangerous, so we focus on them in this class.

Definition 16.4. A **test** is a statistic $\delta \in \{0, 1\}$ such that

- if $\delta = 0$, H_0 is not rejected
- if $\delta = 1$, H_0 is rejected.

As a memorization tool (not entirely proper math), H_δ is the hypothesis that we keep. We still *cannot* conclude that it is true.

Example

Recall the fair coin example, where $H_0 : p = 1/2, H_1 : p \neq 1/2$.

By the central limit theorem, combined with Slutsky's theorem,

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

If H_0 is true, then we can say that

$$\sqrt{n}\frac{\overline{X}_n - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

The left hand side is a statistic we can compute. When H_0 is true, then it is very likely that this quantity will fall between $(-2, 2)$ since the 0.975 quantile is 1.96.

The test is

$$\delta = \mathbb{1}_{\left|\sqrt{n}(\overline{x}_n - 0.5)/\sqrt{\overline{x}_n(1-\overline{x}_n)}\right| > c}$$

for some $c > 0$. If we set $c = 1.96$, then we reject H_0 when this statistic is larger, and it is unlikely that it comes from the standard normal.

Alternatively, if H_0 is true, then we can say that

$$\sqrt{n}\frac{\overline{X}_n - 1/2}{\sqrt{1/2(1-1/2)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

and we can define another test

$$\delta' = \mathbb{1}_{\left|2\sqrt{n}(\overline{x}_n - 0.5)\right| > 1.96}.$$

Note that δ is a function on $X_1, \dots, X_n$ that produces a binary statistic.

Definition 16.5. The **rejection region** of a test δ is defined as

$$R_\delta = \{x \in E^n : \delta(x) = 1\}$$

where $x = (x_1, \dots, x_n)$ and E is the sample space.

That is, the rejection region is the set of all possible outcomes that leads to rejection. For example, in the Bernoulli case,

$$R_\delta = \left\{(x_1, \dots, x_n) \in \{0,1\}^n : \left|\sqrt{n}\frac{\overline{X}_n - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\right| > 1.96\right\}.$$

16.2 Type 1 and type 2 errors

Definition 16.6. A **type 1 error** of a test δ rejects H_0 when it is actually true.

$$\begin{aligned} \alpha_\delta : \Theta_0 &\to \mathbb{R} \\ \theta &\mapsto \mathbb{P}_\theta[\delta = 1] \end{aligned}$$

Definition 16.7. A **type 2 error** of a test δ does not reject H_0 when H_1 is actually true.

$$\begin{aligned} \beta_\delta : \Theta_1 &\to \mathbb{R} \\ \theta &\mapsto \mathbb{P}_\theta[\delta = 0] \end{aligned}$$

Colloquially, we may say that type 1 errors are false negatives and type 2 errors are false positives.

Example 16.8

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown. How do we test if $\mu = 0$ or if $\mu \neq 0$?

We know that

$$\sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, 1)$$

is always true (since the X_i are Gaussian and $\overline{X}_n$ is a linear combination of Gaussian random variables). If H_0 were true, then we could replace μ with 0,

$$\sqrt{n} \cdot \overline{X}_n \sim \mathcal{N}(0, 1).$$

Let

$$\delta = \mathbb{1}_{\left|\sqrt{n} \cdot \overline{X}_n\right| > t}$$

for some $t > 0$. The type 1 error of δ is

$$\begin{aligned} \alpha_\delta : \{0\} &\to \mathbb{R} \\ \mu &\mapsto \Pr_\mu \{\delta = 1\} = \Pr_\theta \left\{\left|\sqrt{n} \cdot \overline{X}_n\right| > t\right\}. \end{aligned}$$

Our Θ_0 is a singleton, so $\forall \mu \in \{0\}$, $\mu = 0$, and

$$\Pr_\theta \left\{\left|\sqrt{n} \cdot \overline{X}_n\right| > t\right\} = 2 - 2\phi(t).$$

Then we choose $t = q_{1-\alpha/2}$, where q is a quantile of the standard normal, so $\alpha_\delta(0) = \alpha$.

Example

Suppose instead we wanted to test if $\mu \geq 0$ or $\mu < 0$.

It is still true that

$$\sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, 1).$$

If H_0 is true,

$$\sqrt{n} \cdot \overline{X}_n \geq \sqrt{n}(\overline{X}_n - \mu)$$

since μ would be non-negative. So we reject when $\sqrt{n} \cdot \overline{X}_n$ is too small,

$$\delta = \mathbb{1}_{\sqrt{n} \cdot \overline{X}_n < -t}$$

where $t > 0$.

For all $\mu > 0$ that satisfy H_0,

$$\begin{aligned} \alpha_\delta(\mu) &= \Pr_\mu \{\delta = 1\} \\ &= \Pr_\mu \left\{\sqrt{n} \cdot \overline{X}_n < -t\right\} \\ &\leq \Pr_\mu \left\{\sqrt{n}(\overline{X}_n - \mu) < -t\right\} = \phi(-t) = 1 - \phi(t). \end{aligned}$$

We take $t = q_{1-\alpha}$ so $\alpha_\delta(\mu) \leq \alpha, \forall \mu \geq 0$. We achieve equality when $\mu = 0$.

We care the most about the type 1 error, so we define the level of a test via the type 1 error. However, we would also like the "1 - type 2 error" to be as large as possible.

Remark. After today we will never see the type 2 error again in this class. This is just "for our culture."

Definition 16.9. The **power** of a test δ is defined as

$$\pi_\theta = \inf_{\theta \in \Theta_1} (1 - \beta_\delta(\theta)).$$

In this class, the definition is just plopped here for completeness and will not be tested on.

Definition 16.10. A test has **level** α if

$$\alpha_\delta(\theta) \leq \alpha, \forall \theta \in \Theta_0.$$

A test has **asymptotic level** α if

$$\lim_{n\to\infty} \alpha_\delta(\theta) \leq \alpha, \forall \theta \in \Theta_0.$$

Remark. If a test has level 5%, then it trivially has level 10% as well. The converse is not true.

In general, a test takes the form

$$\delta = \mathbb{1}_{T_n > c}$$

for some **test statistic** T_n and threshold $c \in \mathbb{R}$. The corresponding rejection region is $R_\delta = \{T_n > c\}$.

Example

We're flipping coins again. We're also flipping the hypotheses.

$$H_0 : p \leq 1/2 \qquad H_1 : p > 1/2$$

As always,

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

No matter what p is, we can apply Slutsky's theorem to obtain

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

If H_0 is true, then

$$\sqrt{n}\frac{\overline{X}_n - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}.$$

The left hand side should not be too large, so we reject when this quantity is large,

$$\delta = \mathbb{1}_{\sqrt{n}\frac{\overline{X}_n - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} > t}.$$

We find t as an exercise left to the reader.

17 April 4, 2018

17.1 Recitation 8

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ for some $\lambda > 0$ and let λ_0 be a fixed (known) positive number.

1. Consider the following hypotheses:

   $$H_0 : \lambda = \lambda_0 \qquad H_1 : \lambda \neq \lambda_0$$

   By the central limit theorem,

   $$\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

   and by Slutsky's theorem,

   $$\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\overline{X}_n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

   Let

   $$S_n = \sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\overline{X}_n}}$$
   $$T_n = \sqrt{n}(\overline{X}_n - \lambda_0)/(\sqrt{\overline{X}_n}).$$

   If H_0 is true, then $T_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$. We know that $|T_n| > 1.96$ is very unlikely if H_0 is true. A test with asymptotic level 5% is

   $$\delta = \mathbb{1}_{|T_n|>1.96}.$$

2. Consider the following hypotheses:

   $$H_0 : \lambda \leq \lambda_0 \qquad H_1 : \lambda > \lambda_0$$

   If H_0 is true, then we would expect $T_n < S_n$ and

   $$\Pr\{T_n > 1.65\} \leq \Pr\{S_n > 1.65\}.$$

   A test with asymptotic level 5% is

   $$\delta = \mathbb{1}_{T_n>1.65}.$$

3. Consider the following hypotheses:

   $$H_0 : \lambda \geq \lambda_0 \qquad H_1 : \lambda < \lambda_0$$

   If H_0 is true, then we would expect $T_n > S_n$. A test with asymptotic level 5% is

   $$\delta = \mathbb{1}_{T_n<-1.65}.$$

4. Consider the following hypotheses:

$$H_0 : |\lambda - 2| \leq 1 \qquad H_1 : |\lambda - 2| > 1$$

Equivalently, $H_0 : 1 \leq \lambda \leq 3$. Let

$$R_n = \sqrt{n}(\overline{X}_n - 1)/(\sqrt{\overline{X}_n})$$
$$L_n = \sqrt{n}(\overline{X}_n - 3)/(\sqrt{\overline{X}_n}).$$

If H_0 is true, $L_n \leq S_n \leq R_n$ and

$$\Pr\{L_n > 1.96 \vee R_n < -1.96\} = \Pr\{S_n > 1.96 \vee S_n < -1.96\} = 0.95.$$

A test with asymptotic level 5% is

$$\delta = \mathbb{1}_{L_n > 1.96 \vee R_n < -1.96}.$$

We are interested in comparing the proportion of people in their 20s who smoke from the US and France.

Suppose we sample n people in both countries, and let N_{US}, N_F be the number of people who smoke in the US and France, respectively. Let p_{US}, p_F be the corresponding proportion of people and let

$$X_i = \begin{cases} 1 & \text{person } i \text{ from US smokes,} \\ 0 & \text{otherwise.} \end{cases} \qquad Y_i = \begin{cases} 1 & \text{person } i \text{ from France smokes,} \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, $N_{US} = \sum_i X_i$ and $N_F = \sum_i Y_i$. X_i and Y_i are independent Bernoulli random variables. The CLT tells us that

$$\frac{(N_{US} - N_F) - n(p_{US} - p_F)}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \operatorname{Var} X_1 - Y_1)$$
$$\sqrt{n}(N_{US} - N_F - (p_{US} - p_F)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p_{US}(1 - p_{US}) + p_F(1 - p_F)).$$

By Slutsky's theorem,

$$T_n = \frac{N_{US} - N_F - n(p_{US} - p_F)}{\sqrt{n}\sqrt{\frac{N_{US}}{n}\left(1 - \frac{N_{US}}{n}\right) + \frac{N_F}{n}\left(1 - \frac{N_F}{n}\right)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

Let

$$H_0 : p_{US} = p_F \qquad H_1 p_{US} \neq p_F.$$

Our test is

$$\mathbb{1}_{|T_n| > q_{1-\alpha/2}}$$

where α is the level of our test.

18 April 5, 2018

This chapter doesn't introduce any new ideas mathematically. Rather, it wraps concepts in new words and vocabulary, so this chapter should be easy—that is, we just need to memorize the definitions.

Here are some remarks about tests in general.

1. Tests should not depend on the true parameter θ.

   For example, when we have a hypothesis like

   $$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

   we *can* include θ_0 because it is a given, but we cannot use θ, which is an unknown parameter.

2. Generally, we can only achieve asymptotic levels α, since we often use the central limit theorem. In practice, this is quite a limitation because we may only have a few data points, say $n = 10$, which is not enough to apply results for $n \to \infty$.

3. Rejecting H_0 does not imply accepting H_1.

   For example, consider $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0, 1)$. If our hypotheses are $p = 1/2$ and $p > 1/2$, they do not cover the whole space, and $p \neq 1/2 \not\Rightarrow p > 1/2$.

18.1 Hypothesis testing, ctd.

Since the 70s, everyone's agreed to take $\alpha = 0.05$ for some reason.

Example 18.1

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ for some unknown $p \in (0, 1)$. We want to test

$$H_0 : p = 1/2 \qquad H_1 : p \neq 1/2$$

with asymptotic level $\alpha \in (0, 1)$.

By the central limit theorem, we know that

$$\sqrt{n} \frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

If H_0 is true, then

$$\sqrt{n} \frac{\overline{X}_n - 1/2}{\sqrt{1/2(1-1/2)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

Let $T_n = \left| \sqrt{n}(\hat{p}_n - 1/2) / \sqrt{1/2(1-1/2)} \right|$, where $\hat{p}_n$ is the MLE and our test is

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha/2}}.$$

Now let's sanity check this. The asymptotic level of δ_α is

$$\Pr_{1/2} \{\delta_\alpha = 1\} = \Pr_{1/2} \left\{ |T_n| > q_{1-\alpha/2} \right\} \to \alpha.$$

Suppose instead that we considered

$$S_n = \left| \sqrt{n} \frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \right|$$

with corresponding test

$$\delta'_\alpha = \mathbb{1}_{S_n > q_{1-\alpha/2}}.$$

This test has the same asymptotic level as δ_α! A minor note though: δ'_α has a better power.

Finally, we can even consider

$$U_n = \sqrt{n} \frac{\overline{X}_n - 1/2}{\sqrt{1/2(1-1/2)}}$$

which is T_n without the absolute value. The corresponding test is

$$\delta''_a = \mathbb{1}_{U_n > q_{1-\alpha}}.$$

This test *also* has asymptotic level α! However, this test has the worst type 2 error. We can rewrite U_n as

$$U_n = \sqrt{n} \frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} + \sqrt{n} \frac{p - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}.$$

If p is small, the second term is very small (negative), and U_n will be super small, so we will not reject.

We conclude with a quick recipe for designing tests.

1. Suppose that H_0 is true.
2. Find a test statistic such that if H_0 is true, we can say something about its (asymptotic) distribution.
3. Reject H_0 if the value of the test statistic is not in the typical range of the distribution.
4. Define such a typical range according to the (asymptotic) level that we want to achieve.

18.2 P-values

The professor begins by declaring that none of us know what a p-value is. In fact, most of his professors didn't know what a p-value was.

Let's recall the previous coin examples from the first lecture on hypothesis testing.

Example

A coin is tossed 80 times, and heads are obtained 54 times. Can we conclude that the coin is significantly unfair?

We reject when

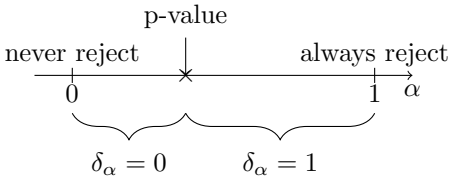
$$\left| \sqrt{n} \frac{\overline{X}_n - 1/2}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \right| > q_{1-\alpha/2}.$$

We may compute that $\overline{X}_n = 54/80 = 0.68$, and the above quantity is 3.45. It seems quite reasonable to reject the hypothesis $p = 0.5$.

If we take $\alpha = 0.1$ or $\alpha = 0.2$, we reject even more values. However, if we take $\alpha = 0.01$, we reject fewer values, and in this case, we no longer reject!

If we choose a small α, then the probability of mistakenly rejecting is small, so we want to be very careful about rejecting.

In the extreme cases, if $\alpha = 1$, we always reject (the 0.5 quantile is 0, the median). If we take $\alpha = 0$, we never reject (the 1 quantile is infinite).



Definition 18.2. The (asymptotic) **p-value** of a test δ_α is the smallest (asymptotic) level α at which δ_α rejects H_0.

The p-value is the smallest value of α at which the indicator becomes 1:

$$T_n = q_{1-\alpha/2} \Leftrightarrow \phi(T_n) = 1 - \alpha/2 \Leftrightarrow \alpha = 2 - 2\phi(T_n).$$

So the p-value is a statistic and a random variable!

In summary:

$$\text{p-value} \leq \alpha \Leftrightarrow H_0 \text{ is rejected by } \delta_\alpha \text{ at (asymptotic) level } \alpha.$$

The smaller the p-value, the more confidently one can reject H_0.

Consider a test of the form

$$\delta_\alpha = \mathbb{1}_{T_n > t_\alpha}$$

where T_n is a test statistic and t_α is a threshold that ensures (asymptotic) level α. As a sanity check, $t_\alpha \to \infty$ as $\alpha \to 0$ and t_α should decrease as α increases. The p-value is the value of α such that $T_n = t_\alpha$.

Example

What is the p-value of δ''_α from before?

We know that $U_n = q_{1-\alpha} \Leftrightarrow \phi(U_n) = 1 - \alpha$. So the p-value is $1 - \phi(U_n)$.

18.3 Neyman-Pearson's paradigm

For given hypotheses, among all tests of (asymptotic) level α, is is possible to find the one with maximal power?

Above, we found three tests with equivalent level, but δ''_α had terrible power and δ'_α had the highest power. Furthermore, the trivial test $\delta = 0$ has perfect level but no power.

According to Neyman-Pearson, it *is* possible to find the most powerful test with a given level, but in this class, we will only study a few specific cases.

18.4 Chi-squared distributions

Definition 18.3. For a positive integer d, the **χ^2 distribution with d degrees of freedom** is the law of the random variable

$$Z_1^2 + Z_2^2 + \cdots + Z_d^2$$

where $Z_1, \ldots, Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Here are some nice properties.

- If $Z \sim \mathcal{N}_d(0, I_d)$, then $|Z|^2 \sim \chi^2_d$. This is because each dimension is independent of every other.
- Cochran's theorem states that for $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, if S_n is the sample variance, then $nS_n/\sigma^2 \sim \chi^2_{n-1}$.
- $\chi^2_2 = \epsilon(1/2)$.

19 April 10, 2018

We have a test next week! *sadface*

Here are some tips from the professor.

- Tests often have multiple answers, so don't stress over matching the exact answer (to solutions).

  For example, suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Consider the hypotheses

  $$H_0 : p = 1/3 \qquad H_1 : p \neq 1/3.$$

  Let

  $$T_n = \sqrt{n}(\overline{X}_n - 1/3)/\sqrt{1/3(1-1/3)}$$
  $$S_n = \sqrt{n}(\overline{X}_n - 1/3)/\sqrt{\overline{X}_n(1-\overline{X}_n)}.$$

  For T_n we invoke CLT and for S_n we invoke CLT + Slutsky's. Potential tests include the following.

  $$\begin{array}{ll} \delta = \mathbb{1}_{|T_n| > q_{1-\alpha/2}} & \delta = \mathbb{1}_{|S_n| > q_{1-\alpha/2}} \\ \delta = \mathbb{1}_{T_n > q_{1-\alpha}} & \delta = \mathbb{1}_{S_n > q_{1-\alpha}} \\ \delta = \mathbb{1}_{T_n < -q_{1-\alpha}} & \delta = \mathbb{1}_{S_n < -q_{1-\alpha}} \end{array}$$

  There are infinitely many tests!

  We can even give the trivial test 0, and the professor says he'd give us full credit, except he'll say "using the previous parts, come up with a test" so this won't work!

- The homework will be due the week after, but it will be released today.
- Problem sets are available from his office.

19.1 Wald's test

Consider an i.i.d. sample $X_1, \dots, X_n$ with statistical model $(E, \{\Pr_\theta\}_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d, d \geq 1$ and let $\theta_0 \in \Theta$ be fixed and given.

Consider the following hypotheses,

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

Let $\hat{\theta}^{\text{MLE}}$ be the MLE and assume its technical conditions are satisfied (open parameter set, etc.). We know that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta)^{-1}).$$

If H_0 is true, then we can substitute $\theta \leftarrow \theta_0$ and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta_0)^{-1}).$$

We will use this fact to show an elegant result—but first, a proposition.

Proposition 19.1

$I(\theta)$ is positive semi-definite.[a]

[a] That is, it is symmetric and has non-negative eigenvalues.

Proof. $I(\theta)$ is the covariance matrix (symmetric) of the gradient of log-likelihood. It is also a Hessian, which is necessarily symmetric.

Let Σ be a covariance matrix of a random vector $Y \in \mathbb{R}_d$ and let λ be an eigenvalue of Σ. That is, $\exists u \in \mathbb{R}^d \neq 0$ such that $\Sigma\mu = \lambda\mu$. Then $u^T\Sigma u = \lambda u^T u = \lambda |u|^2$. However, we recognize that $u^T\Sigma u = \operatorname{Var} u^T Y$, which is non-negative, so $\lambda \geq 0$. □

Since Σ is real and symmetric, we can diagonalize it as $\Sigma = PDP^{-1}$, where D is a diagonal matrix with eigenvalues on the diagonal.

Now let

$$\Sigma^{1/2} = P \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \searrow & \\ & & \sqrt{\lambda_d} \end{pmatrix} P^{-1}$$

such that $(\Sigma^{1/2})^2 = PDP^{-1} = \Sigma$.

Let's return to our original claim. For any $A \in \mathbb{R}^{d\times d}$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta_0)^{-1})$$
$$\sqrt{n}A(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, AI(\theta_0)^{-1}A^T)$$

Suppose we take $A = I(\theta_0)^{1/2}$. Then

$$\sqrt{n}I(\theta_0)^{1/2}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} z, z \sim \mathcal{N}(0, I).$$

The squared norm is a continuous function, so

$$\left|\sqrt{n}I(\theta_0)^{1/2}(\hat{\theta} - \theta_0)\right|^2 \xrightarrow[n\to\infty]{d} |z|^2$$

where $|z|^2 \sim \chi^2_d$.

> "Did I say something cool was going to happen? Because it hasn't happened yet."—veb

Recall that $|u| = u^T u, \forall u \in \mathbb{R}^d$. So the disgusting norm becomes

$$n(\hat{\theta} - \theta_0)^T (I(\theta_0)^{1/2})^T I(\theta_0)^{1/2}(\hat{\theta} - \theta_0) = n(\hat{\theta} - \theta_0)^T I(\theta_0)(\hat{\theta} - \theta_0).$$

Theorem 19.2 (Wald's test)

Let

$$T_n = n(\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0).$$

Wald's test with asymptotic level $\alpha \in (0, 1)$ is

$$\delta = \mathbb{1}_{T_n > q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of χ^2_d (use tables).

We can replace θ_0 with $\hat{\theta}$ via Slutsky's theorem when substituting $A \leftarrow I(\theta_0^{1/2}$.

- If H_1 is of the form $\theta \neq \theta_0$, Wald's test is the most powerful.
- If H_1 is of the form $\theta > \theta_0$ it is not the most powerful.

Visually, we can think of our desired "range" as a ball with radius $\sqrt{q_{1-\alpha}}$, but this visualization is less convenient since we do not want to compute $I(\theta)^{1/2}$.

19.2 Likelihood ratio test

Consider an i.i.d. sample $X_1, \dots, X_n$ with statistical model $(E, \{\Pr_\theta\}_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d, d \geq 1$ and let $\theta_0 \in \Theta$ be fixed and given.

Suppose the null hypothesis has form

$$H_0 : (\theta_{r+1}, \dots, \theta_d) = (\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)})$$

where we only test some parameters.

Let $\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta)$ be the MLE and let $\hat{\theta}_n^c = \arg\max_{\theta \in \Theta_0} \ell_n(\theta)$ be the constrained MLE, where Θ_0 is the set of parameters we want to test.

Example 19.3

Consider a sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where μ, σ^2 are unknown. Consider the following hypotheses,

$$H_0 : \sigma^2 = 1 \qquad H_1 : \sigma^2 \neq 1.$$

Here, $d = 2, r = 1$ since our θ is two-dimensional.

Example 19.4

Consider a sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where μ, σ^2 are unknown. Consider the following hypotheses,

$$H_0 : \mu = 0 \qquad H_1 : \mu \neq 0.$$

The setup is the same. The MLE is simply $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\overline{X}_n, \overline{X_n^2} - \overline{X}_n^2)$. The constrained MLE can be calculated when we set $\mu = 0$,

$$\log L_n(X_1, \dots, X_n, \mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2$$

$$\log L_n(X_1, \dots, X_n, 0, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i X_i^2$$

so the constrained MLE for σ^2 is $(\hat{\sigma}^2)^c = \overline{X_n^2}$.

Our test statistic is

$$T_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c)\right).$$

Theorem 19.5

Assume H_0 is true and the MLE technical conditions are satisfied. Then

$$T_n \xrightarrow[n\to\infty]{d} \chi^2_{d-r}$$

with respect to $\Pr_\theta$.

The likelihood ratio test with asymptotic level $\alpha \in (0,1)$ is

$$\mathbb{1}_{T_n > q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_{d-r} (again, look this up in a table). The asymptotic p-value of this test is $1 - F_{d-r}(T_n)$, where F_{d-r} is the cdf of χ^2_{d-r}. The same holds for Wald's test.

20 April 11, 2018

20.1 Recitation 9

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for unknown μ, σ^2.

1. We want to test whether the X_i are standard Gaussians, with asymptotic level $\alpha \in (0,1)$.

   (a) The hypotheses that test this fact are

   $$H_0 : (\mu, \sigma^2) = (0,1) \qquad H_1 : (\mu, \sigma^2) \neq (0,1).$$

   (b) Let

   $$T_n = n\hat{\theta}^T I(\hat{\mu}, \hat{\sigma}^2)\theta$$

   where $\theta = \begin{pmatrix} \mu & \sigma^2 \end{pmatrix}^T$. The corresponding Wald's test is $\delta = \mathbb{1}_{T_n > q_{1-\alpha}}$.

   (c)

   $$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2$$
   $$\ell_n(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}\log 2\pi\hat{\sigma}^2 - \frac{n}{2}$$
   $$\ell_n(0,1) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\overline{X_n^2}$$

   Let

   $$S_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c)\right)$$

   and the likelihood ratio test is $\delta' = \mathbb{1}_{S_n > q_{1-\alpha}}$.

   (d) Assume $n = 100$ and the empirical mean of the sample is 0.18 and the empirical variance is 1.12.

   The p-value of δ is

   $$1 - F_{\chi_2^2}(T_n) \approx 1 - F_{\chi_2^2}(3.47) \approx 0.2$$

   and the p-value of δ' is

   $$1 - F_{\chi_2^2}(S_n) \approx 1 - F_{\chi_2^2}(3.91) \approx 0.1.$$

2. If we only wanted to test whether $\sigma^2 = 1$, we could use the likelihood ratio test. The constrained MLE is where we set $\sigma^2 \leftarrow 1$, so $\hat{\mu}^c = \hat{\mu}$.

   The test statistic is

   $$Z_n =$$

3. Now we just want to test if $\mu = 0$. Literally the same thing.

merp i have a headache lol so ima go home and sleep :D took a picture

21 April 12, 2018

Note 21.1. A note about the previous lecture: we can apply the likelihood ratio test wherever we can apply Wald's test, since "all parameters" is a specific case of "some parameters." We simply set $r = 0$.

21.1 Testing implicit hypotheses

Let $X_1, \dots, X_n$ be i.i.d. random variables and let $\theta \in \mathbb{R}^d$ be a parameter associated with the distribution of X_1 (e.g. a moment, parameter, etc.).

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be continuously differentiable, with $k < d$ (else the system of equations is over-specified and may not have a solution). Consider the following hypotheses

$$H_0 : g(\theta) = 0 \qquad H_1 : g(\theta) \neq 0.$$

Example 21.2

The function g can be multidimensional and test a variety of conditions.

- $g(\theta) = \theta_1 - \theta_2$ is a one-dimensional function testing whether $\theta_1 = \theta_2$.
- $g(\theta) = (\theta_1, \theta_2)$ provides a test about the first two coordinates.
- $g(\theta) = (\theta_1, \theta_3 - \theta_2)$ tests whether $\theta_1 = 0$ and $\theta_2 = \theta_3$.

Where else have we seen functions like g? The Delta method!

Suppose an asymptotically normal estimator $\hat{\theta}_n$ is available:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, \Sigma(\theta)).$$

By the Delta method,

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}_k(0, \Gamma(\theta))$$[11]

where $\Gamma(\theta) = \nabla g(\theta)^T \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k\times k}$.

Suppose $\Sigma(\theta)$ is invertible (all eigenvalues strictly positive) and $\nabla g(\theta)$ has rank k.[12] Recall that Γ is diagonalizable and $\Gamma^{-1/2}$ takes the square root of the eigenvalues. Thus, $\Gamma(\theta)$ is invertible and

$$\begin{aligned}
\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}_k(0, \Gamma(\theta)) \\
\sqrt{n}\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}_k\left(0, \Gamma(\theta)^{-1/2}\Gamma(\theta)(\Gamma(\theta)^{-1/2})^T\right) \\
\sqrt{n}\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) &\xrightarrow[n\to\infty]{d} \mathcal{N}_k(0, I_k).
\end{aligned}$$

Honestly linear algebra should be a prerequisite for this class. Notice that this is *almost* what we want. However, we don't know $\Gamma(\theta)$, so we apply Slutsky's

[11] The matrices are $k \times d, d \times d, d \times k$, so the product is $k \times k$.

[12] Only Σ is a square matrix, so we can only guarantee that g has full rank.

theorem and replace it with $\Gamma(\hat{\theta})$,

$$\sqrt{n}\Gamma(\hat{\theta})^{-1/2}\left(g(\hat{\theta}_n)-g(\theta)\right)=\sqrt{n}\underbrace{\Gamma(\hat{\theta})^{-1/2}\Gamma(\theta)^{1/2}}_{\xrightarrow[n\to\infty]{\mathbb{P}} I_k}\underbrace{\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n)-g(\theta)\right)}_{\xrightarrow[n\to\infty]{d}\mathcal{N}(0,I_k)}.$$

Now we have a quantity that only depends on our data,

$$\sqrt{n}\Gamma(\hat{\theta})^{-1/2}\left(g(\hat{\theta}_n)-g(\theta)\right)\xrightarrow[n\to\infty]{d}\mathcal{N}_k(0,I_k).$$

However, we don't want to compute square roots; instead, we'll take the squared norm, which converges to a χ^2_k distribution. If H_0 is true (that is, $g(\theta)=0$),

$$\underbrace{ng(\hat{\theta}_n)^T\Gamma^{-1}(\hat{\theta}_n)g(\hat{\theta}_n)}_{T_n}\xrightarrow[n\to\infty]{d}\chi^2_k.$$

As usual, the p-value of this test is $1-F_k(T_n)$, where F_k is the cdf of χ^2_k.

> "If you don't like tests—and I don't like tests—you'll be bored for the next 40 minutes. But if you like tests, if you like multiplying matrices... and I mean some people do?"—veb

21.2 Multinomial chi-squared test

> "The remaining of the chapter is going to be very boring, but it's only 8 more slides!"—veb

Let $E=\{a_1,\dots,a_K\}$ be a finite space and $(\Pr_p)_{p\in\Delta_k}$ be the family of all probability distributions on E,

$$\Delta_k=\left\{p=(p_1,\dots,p_k)\in(0,1)^K:\sum_{j=1}^K p_j=1\right\}.$$

Since the space is discrete, the pmf suffices to characterize the entire distribution. For $p\in\Delta_k$ and $X\sim\Pr_p$,

$$\Pr_p\{X=a_j\}=p_j, j=1,\dots,K.$$

Remark 21.3. The Bernoulli distribution is a special case of the multinomial, where $K=2$.

Let $X_1,\dots,X_n\overset{\text{i.i.d.}}{\sim}\Pr_p$, for some unknown $p\in\Delta_k$, and let $p^0\in\Delta_k$ be fixed. We want to test

$$H_0:p=p^0\qquad H_1:p\neq p^0$$

with asymptotic level $\alpha\in(0,1)$.

Example 21.4

If we let $p^0=(1/K,1/K,\dots,1/K)$, then we are testing whether $\Pr_p$ is the uniform distribution on E.

The likelihood of the model is

$$L_n(X_1, \dots, X_n, p) = p_1^{N_1} p_2^{N_2} \dots p_K^{N_K}$$

where $N_j = \#\{i = 1, \dots, n : X_i = a_j\}$. The MLE $\hat{p}$ is equal to

$$\hat{p}_j = \frac{N_j}{n} = \frac{\#\{i = 1, \dots, n : X_i = a_j\}}{n}$$

subject to the constraint that $\sum_j p_j = 1$, which is a sample average of $\mathbb{1}_{X_i = a_j}$. We might consider using the central limit theorem

$$\sqrt{n}(\hat{p} - p) \xrightarrow[n\to\infty]{d} \mathcal{N}_k(0, \Sigma(p))$$

but we find that $\Sigma(p)$ is *not* invertible! So unfortunately, we cannot apply Wald's test, which relies on the fact that the covariance matrix is invertible. As an exercise to the reader, it *is* possible to find a matrix such that once we multiply both sides, we can obtain almost the identity:

$$\sqrt{n} A(p)(\hat{p} - p) \xrightarrow[n\to\infty]{d} \mathcal{N}_k \left(0, \begin{pmatrix} 1 & & 0 \\ & \searrow & \\ & 1 & \\ 0 & & 0 \end{pmatrix} \right).$$

If we take the squared norm, it converges to the χ^2_{K-1} distribution!

Theorem 21.5

$$\underbrace{n \sum_{j=1}^{k} \frac{\left(\hat{p}_j - p_j^0\right)^2}{p_j^0}}_{T_n} \xrightarrow[n\to\infty]{d} \chi^2_{K-1}.$$

This equation should go on your equation sheet since it's important but not very intuitive. However, Wald's test is quite intuitive, so you should know it.

The χ^2 test with asymptotic level α is

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of χ^2_{K-1}. The corresponding asymptotic p-value of this test is $1 - F_{K-1}(T_n)$, where F_{K-1} is the cdf of χ^2_{K-1}.

21.3 Student's distributions

> "I want you to understand that the next 4 slides are really *really* useless."—veb

Remark 21.6. "Student" is not a researcher, but rather how they signed the paper. An intern was working with confidential data, so he published anonymously under "student."

Definition 21.7. For $d \in \mathbb{N}$, the **Student's distribution with d degrees of freedom** (denoted by t_d) is the law of the random variable

$$\frac{U}{\sqrt{V/d}}$$

where $U \sim \mathcal{N}(0,1), V \sim \chi^2_d$, and $U \perp\!\!\!\perp V$.

Theorem 21.8 (Cochran's theorem)

For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, if S_n is the sample variance, then

$$\sqrt{n-1}\frac{\overline{X}_n - \mu}{\sqrt{S_n}} \sim t_{n-1}.$$

- This is really weird! It suggests that $\hat{\mu} \perp\!\!\!\perp \hat{\sigma}^2$ if the X_i are Gaussian. That is definitely not true in general (Poisson, Bernoulli, etc.).
- We can rescale for $nS_n/\sigma^2 \sim \chi^2_{n-1}$.
- We know by the CLT that $\sqrt{n}(\overline{X}_n - \mu)/\sigma \sim \mathcal{N}(0,1)$, so with Cochran's theorem,

  $$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \perp\!\!\!\perp \frac{n\hat{\sigma}^2}{\sigma^2}.$$

  Furthermore,

  $$\frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{n\hat{\sigma}^2/\sigma^2}{n-1}}} = \sqrt{n-1} \cdot \frac{\overline{X}_n - \mu}{\sqrt{\hat{\sigma}^2}} \sim t_{n-1}.$$

Example 21.9

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Consider the hypotheses

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$$

We could solve this two weeks ago, with the central limit theorem or now, with the likelihood ratio test. But what if we wanted a non-asymptotic test?

No matter what, however, if H_0 is true, then

$$T_n = \sqrt{n-1}\frac{\overline{X}_n - \mu_0}{\sqrt{\hat{\sigma}^2}} \sim t_{n-1}$$

even when n is small! This does not "converge in distribution" to a student's distribution; it *is* t_{n-1} random variable.

Using this test statistic, we can create a test

$$\delta = \mathbb{1}_{T_n > q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of t_{n-1}.

So why is this useless? It assumes that the sample is perfectly Gaussian. Cochran's theorem is wrong even if the sample is only *slightly* not Gaussian. In practice, you'll never get perfect random variables.

22 March 18, 2018

There's a test tomorrow! *doom*

22.1 Recitation 10

1. Assume $\mathcal{I}$ is a confidence interval for a parameter $\theta \in \mathbb{R}$ with asymptotic level α. Then $\mathbb{1}_{\theta_0 \notin I}$ is a test of asymptotic level α for $H_0 : \theta = \theta_0$ where $\theta_0 \in \mathbb{R}$ is fixed.

   True. By definition, $\Pr\{\theta \notin \mathcal{I}\} \to \alpha$. If H_0 is true, then $\theta = \theta_0$, so $\Pr\{\theta_0 \notin \mathcal{I}\} \to \alpha$.

2. Wald's test is always valid when H_0 is of the form $\theta = \theta_0$ where θ_0 is a fixed value in the parameter space.

   False. We still require the conditions of the theorem for the MLE.

3. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \lambda > 0$. Which of the following have asymptotic level 5%?

   (a) 1—**No** this is always true.

   (b) $\mathbb{1}_{T_n > 1.65}$—**Sure**.

   (c) $\mathbb{1}_{|T_n| > 1.96}$—**Sure**.

   (d) 0—**Sure** this is the trivial test.

   Recall that the test should be true less than 5% of the time.

Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables in $[0, \theta]$ where $\theta > 0$ is unknown. Consider the hypotheses

$$H_0 : \theta \geq 1 \qquad H_1 : \theta > 1.$$

1. (a) The likelihood function is

$$\begin{aligned} L_n : \mathbb{R}^n \times (0, \infty) &\to \mathbb{R} \\ (x_1, \ldots, x_n, \theta) &\mapsto \frac{1}{\theta^n} \mathbb{1}_{\max_i x_i < \theta}. \end{aligned}$$

   (b) The parameter θ is identified since $\forall \theta' \neq \theta, [0, \theta] \neq [0, \theta']$.

   (c) The MLE $\hat{\theta}$ must be greater than the largest X_i, and the likelihood is a strictly decreasing function, so $\hat{\theta} = \max_i X_i$.

   (d) We want to show that

$$S_n = \frac{n(\theta - \hat{\theta})}{\theta} \xrightarrow[n \to \infty]{d} \epsilon(1).$$

   Recall that convergence in distribution implies that the cdfs converge at all values. That is, $\forall t > 0$,

$$\Pr\left\{ \left| \frac{n(\theta - \hat{\theta})}{\theta} \right| \leq t \right\} \to 1 - e^{-t}.$$

We know that $\theta > \hat{\theta}$, so we simply rearrange terms on the left and find that

$$\Pr\left\{\hat{\theta} \geq \theta\left(1 - \frac{t}{n}\right)\right\} = 1 - \left(1 - \frac{t}{n}\right)^n \to 1 - e^{-t},$$

as desired. The last limit is good to remember.

(e) We know that $\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta$, so by Slutsky's theorem,

$$T_n = \frac{n(\theta - \hat{\theta})}{\hat{\theta}} \xrightarrow[n\to\infty]{d} \epsilon(1).$$

If H_0 is true, then $T_n \leq S_n$. That means that T_n should *not* be large. For all $\theta \geq 1$,

$$\Pr\{\delta_\alpha = 1\} = \Pr_\theta\{T_n \geq q_{1-\alpha}\}.$$

A test for the hypotheses with asymptotic level $\alpha \in (0, 1)$ is

$$\mathbb{1}_{T_n \geq q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal.

2. Now we find a test with *non*-asymptotic level α.

(a) The cdf of S_n is

$$\begin{aligned} F(t) &= \Pr\{T_n \leq t\} \\ &= 1 - \left(1 - \frac{t}{n}\right)^n \end{aligned}$$

if $t < n$ or 1 if $t \geq n$.

(b) Let $R_n = n(1 - \hat{\theta})$. If H_0 is true, how does R_n compare to S_n?
If $\theta > 1$, then $S_n \geq R_n$. So we know that R_n should *not* be very large.

(c) Now we find a non-asymptotic test. Let

$$\delta'_\alpha = \mathbb{1}_{R_n \geq c}.$$

where c is the $1 - \alpha$ quantile of S_n. We have the cdf so we solve for c where

$$\alpha = 1 - \left(1 - \frac{c}{n}\right)^n.$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. pairs of random variables where $X_i \sim$ Bernoulli$(p), Y_i \sim$ Bernoulli(q) for unknown $p, q \in (0, 1)$.

1. We are not going to compute this, but

$$\begin{aligned} \hat{p} &= \overline{X}_n \\ \hat{q} &= \overline{Y}_n. \end{aligned}$$

2. We see that $\hat{p}, \hat{q}$ are both sample averages themselves:

$$\begin{pmatrix} \hat{p} \\ \hat{q} \end{pmatrix} = \frac{1}{n} \sum_i \begin{pmatrix} \hat{X}_i \\ \hat{Y}_i \end{pmatrix} \xrightarrow[n\to\infty]{\mathbb{P}} \begin{pmatrix} p \\ q \end{pmatrix}.$$

We can use the CLT out of the box, where the variance is

$$\Sigma = \begin{pmatrix} p(1-p) & \operatorname{cov}(X_1, Y_1) \\ \operatorname{cov}(X_1, Y_1) & q(1-q) \end{pmatrix}$$

3. Now assume that X_1, Y_1 are independent.[13]

   (a) The covariance matrix is

$$\begin{pmatrix} p(1-p) & 0 \\ 0 & q(1-q) \end{pmatrix}.$$

   (b) Consider the hypotheses

$$H_0 : p \geq q \qquad H_1 p < q.$$

Based on $\hat{p}, \hat{q}$, we find a test with asymptotic level α. By the delta method,

$$\sqrt{n}\left((\hat{p} - \hat{q}) - (p - q)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)q(1-q)).$$

By Slutsky's,

$$\sqrt{n}\frac{(\hat{p} - \hat{q}) - (p - q)}{\sqrt{\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

So on a practical note—how do we come up with this parameter $p - q$? We can think of $p \geq q$ as $p - q \geq 0$.

[13] Note that this is different from the pairs being independent. For example, if we take $Y = 2X$, then the pairs are still independent, but X and Y are not.

23 April 24, 2018

People did well on the exam—yay! Solutions will be posted later in the week, but the professor suggests that you review your own exam and figure out the mistakes first.

Whenever we encounter a problem in this class, it usually starts with a preamble along the lines of "let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim}$ some distribution." But what if the variables aren't exactly that distribution? This chapter will introduce how we determine whether our model actually fits the data well.

23.1 Chi-squared test of independence—discrete case

Suppose we have a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. pairs[14] on a finite space $\{a_1, \dots, a_K\} \times \{b_1, \dots, b_L\}$. Consider the two hypotheses

$$H_0 : X_1 \perp\!\!\!\perp Y_i \qquad X_1 \not\perp\!\!\!\perp Y_i.$$

For $(k, l) \in \{1, \dots, K\} \times \{1, \dots, L\}$, let

- $p_{k,l} = \Pr\{X_1 = a_k, Y_1 = b_l\}$,
- $p_{k,\cdot} = \Pr\{X_1 = a_k\}$, and
- $p_{\cdot,l} = \Pr\{Y_1 = b_l\}$.

By the law of total probability, $p_{k,\cdot} = \sum_{l=1}^{L} p_{k,l}$, and likewise for $p_{\cdot,l}$.

Now we can rewrite the hypotheses in terms of p.

$$X_1 \perp\!\!\!\perp Y_1 \Leftrightarrow \Pr\{X_1 = a_k, Y_1 = b_l\} = \Pr\{X_1 = a_k\} \Pr\{Y_1 = b_l\}, \forall k, l.$$

In other words, we may say that

$$H_0 : p_{k,l} = p_{k,\cdot} \times p_{\cdot,l}, \forall k, l \qquad H_1 : p_{k,l} \neq p_{k,\cdot} \times p_{\cdot,l}.$$

We may estimate these quantities with empirical frequencies:

$$\hat{p}_{k,l} = \frac{N_{k,l}}{n} \qquad \hat{p}_{k,\cdot} = \frac{N_{k,\cdot}}{n} \qquad \hat{p}_{\cdot,l} = \frac{N_{\cdot,l}}{n}.$$

Note that each (X_1, Y_1) is a multinomial random variable, so these are actually the MLE estimates (we found this in recitation).

If H_0 is true, then $\hat{p}_{k,l} \approx \hat{p}_{k,\cdot}\hat{p}_{\cdot,l}, \forall k, l$. So let

$$T_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(\hat{p}_{k,l} - \hat{p}_{k,\cdot}\hat{p}_{\cdot,l})^2}{\hat{p}_{k,\cdot}\hat{p}_{\cdot,l}}.$$

Thus if H_0 is true, then

$$T_n \xrightarrow[n\to\infty]{d} \chi^2_{(K-1)(L-1)}.$$

So the test of independence with asymptotic level α is

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2_{(K-1)(L-1)}$. The corresponding p-value is $1 - F_{(K-1)(L-1)}(T_n)$.

Remark 23.1. This is very similar to the multinomial χ^2 test.

[14] Note that X_i, Y_i need not be independent of each other.

23.2 Chi-squared goodness-of-fit test—discrete case

Let $X_1, \dots, X_n$ be i.i.d. random variables on a finite space $E = \{a_0, \dots, a_{K-1}\}$ with some probability measure Pr. Let $(\Pr_\theta)_{\theta \in \Theta}$ be a parametric family of probability distributions on E.

For $j = 1, \dots, K$ and $\theta \in \Theta$, set

$$p_j(\theta) = \Pr_\theta \{Y = a_j\}$$

where $Y \sim \Pr_\theta$ and $p_j = \Pr\{X_1 = a_j\}$.

Example 23.2

Suppose we have a set of random variables $\{X_1, \dots, X_n\}$ on sample space $E = \{1, \dots, K-1\}$. Are the X_i's distributed as Binomial$(K-1, \theta)$ for some $\theta \in (0, 1)$?

Let $\hat{p}_k = \frac{1}{n} \sum_i \mathbb{1}_{X_i = k}$ for $k = 0, \dots, K-1$. This is a natural estimator of $p_k = \Pr\{X_1 = k\}$ (it is also the MLE for multinomial random variables). Furthermore, let

$$p_k(\theta) = \binom{K-1}{k} \theta^k (1-\theta)^{K-1}$$

be the pmf of the binomial distribution with the given parameters. Now, we can write our hypothesis as

$$H_0 : p_k = p_k(\theta), \forall k = 0, \dots, K-1$$

for some $\theta \in (0, 1)$.

Suppose we wanted to test if $\theta = \theta_0$. We would use the χ^2 distribution from the previous lecture, where

$$T_n = n \sum_{k=0}^{K-1} \frac{(\hat{p}_k - p_k(\theta_0))^2}{p_k(\theta_0)}.$$

However, this doesn't tell us *if* the random variables are actually binomial. Instead, let's use our best guess for θ and leverage the MLE:

$$T_n = n \sum_{k=0}^{K-1} \frac{\left(\hat{p}_k - p_k(\hat{\theta})\right)^2}{p_k(\hat{\theta})}.$$

Here, the MLE $\hat{\theta}$ is the value of θ we would have computed if H_0 is true; that is, if the data do follow the given distribution.

In the general case, consider the hypotheses

$$H_0 : \Pr \in (\Pr_\theta)_{\theta \in \Theta} \qquad H_1 : \Pr \notin (\Pr_\theta)_{\theta \in \Theta}.$$

H_0 is equivalent to

"the statistical model $(E, (\Pr_\theta)_{\theta \in \Theta})$ fits the data."

Under some technical assumptions, if H_0 is true, then

$$T_n \xrightarrow[n\to\infty]{d} \chi^2_{K-d-1}$$

where d is the size of the parameter θ and $d < K-1$. The test with asymptotic level α is

$$\delta_\alpha = \mathbb{1}_{T_n > 1_{1-\alpha}}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of χ^2_{K-d-1}.

23.3 Chi-squared goodness-of-fit test—infinite case

> "In the two to three months we've been together, we've never talked about binomial distributions, and I've just told you that this test is very useful for binomial distributions!"—veb

Instead, suppose the sample space is infinite (e.g. $E = \mathbb{N}, E = \mathbb{R}$, etc.) We cannot have a infinite χ^2 distribution, but we can partition E into K disjoint bins:

$$E = A_1 \cup \cdots \cup A_k.$$

For example, we can partition $\mathbb{N}$ into

$$\mathbb{N} = \{0\} \cup \{1\} \cup \{2\} \cup (\mathbb{N} \setminus \{0,1,2\}).$$

Now everything is exactly the same as before. For $\theta \in \Theta, j = 1, \ldots, K$:

- $p_j(\theta) = \Pr_\theta \{Y \in A_j\}$ for $Y \sim \Pr_\theta$,
- $p_j = \Pr\{X_1 \in A_j\}$,
- $\hat{p}_j = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A_j}$, and
- $\hat{\theta}$ is the same as before.

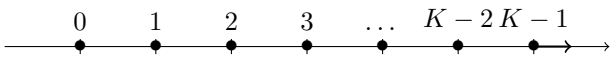
Example 23.3

The professor hates soccer but apparently his hometown's team is doing okay. Suppose the scores are

$$2, 3, 0, 0, 1, 5, 2, 4, 4, \ldots$$

Are the data Poisson-distributed?

Let's take partitions



So our estimates would be

$$\hat{p}_j(\hat{\theta}) = \frac{e^{-\hat{\theta}} \cdot \hat{\theta}^d}{d!}$$

for $j = 1, \ldots, K-2$ and $1 - \sum_{j=1}^{K-2} p_j(\hat{\theta})$ for $j = K-1$. Finally, how do we choose K?

"I can see that half the room already has the wrong answer"—veb

You can't choose K by looking at the data! You cannot say $K = \max X_i$ because K is not a random variable.

If we take K to be too small, the power of the test will suffer, and if we take K to be too large, the level (?) of the test will suffer.

24 April 25, 2018

24.1 Recitation 11

Let X, Y be two Bernoulli random variables and denote by $p = \Pr\{X = 1\}, q = \Pr\{Y = 1\}$, and $r = \Pr\{X = 1, Y = 1\}$.

1. X and Y are independent iff $r = pq$.

   Proof. By definition, if X and Y are independent, then $r = pq$.

   Conversely, if $r = pq$, then there are four cases to check.

   (a) $\Pr\{X = 1, Y = 1\} = r = pq$.

   (b) $\Pr\{X = 1, Y = 0\} = \Pr\{X = 1\} - \Pr\{X = 1, Y = 1\} = p(1 - q)$.

   (c) $\Pr\{X = 0, Y = 1\} = \Pr\{Y = 1\} - \Pr\{X = 1, Y = 1\} = (1 - p)q$.

   (d) $\Pr\{X = 0, Y = 0\} = 1 - \Pr\{11\} - \Pr\{01\} - \Pr\{10\} = (1-p)(1-q)$.

   Ta da! X and Y are independent. □

   "Why am I going through every step here? This is not obvious."
   —pfeffer

2. $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of i.i.d. copies of (X, Y). Let

$$T_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(\hat{p}_{k,l} - \hat{p}_{k,\cdot}\hat{p}_{\cdot,l})^2}{\hat{p}_{k,\cdot}\hat{p}_{\cdot,l}}$$

   where the various p are as defined in lecture. The χ^2 test of independence with asymptotic level α is

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}}$$

   where $q_{1-\alpha}$ is the $1 - \alpha$ quantile of χ^2_1.

3. (a) Let $\hat{p} = \overline{X}_n, \hat{q} = \overline{Y}_n, r = \overline{XY}_n$. By the LLN, these are consistent as sample means of $\mathbb{1}_{X=1}$, etc.

   (b) The vector $(\hat{p}, \hat{q}, \hat{r})$ is asymptotically normal by the CLT with asymptotic variance

$$\Sigma = \begin{pmatrix} \operatorname{Var} X_1 & \operatorname{cov}(X_1, Y_1) & \operatorname{cov}(X_1, X_1Y_1) \\ \operatorname{cov}(X_1, Y_1) & \operatorname{Var} Y_1 & \operatorname{cov}(Y_1, X_1Y_1) \\ \operatorname{cov}(X_1, X_1Y_1) & \operatorname{cov}(Y_1, X_1)\, Y_1 & \operatorname{Var} X_1Y_1 \end{pmatrix} = \begin{pmatrix} p(1-p) & r - pq & r(1-p) \\ r - pq & q(1-q) & r(1-q) \\ r(1-p) & r(1-q) & r(1-r) \end{pmatrix}$$

   verify every cell from photo

   (c) Let $g : \mathbb{R}^3 \to \mathbb{R}$ be the function $(x, y, z) \mapsto z - xy$. By the delta method,

$$n(\hat{r} - \hat{p}\hat{q} - (r - pq)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V)$$

   where $V = \nabla g^T \Sigma \nabla g$.

   If H_0 is true, the the $r - pq$ terms become 0 and our covariance matrix becomes a slight bit nicer to compute.

(d) By Slutsky's theorem,

$$S_n = \sqrt{n}\frac{\hat{r} - \hat{p}\hat{q}}{\sqrt{\hat{V}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

So our test is

$$\delta'_\alpha = \mathbb{1}_{|S_n|} > q_{1-\alpha/2}.$$

(e) We show that $\delta_\alpha = \delta'_\alpha$.
This is just because $X_i = \mathbb{1}_{X_i=1}$.

25 April 26, 2018

"Hello everybody! Happy drop date! Today a piano is going to get dropped. I'm really looking forward to it."—veb

"I know everyone wants an A but not everyone's going to get an A because this is life."—veb

25.1 Chi-squared degrees of freedom

Recall that we encountered three test statistics from last lecture. Can we build some intuition about what these statistics converge to? Let's see...

$$n\sum_{k=1}^{K}\frac{(\hat{p}_k - p_k^0)^2}{p_k^0} \sim \chi^2_{K-1}$$

$$n\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(\hat{p}_{kl} - \hat{p}_{k,\cdot}\hat{p}_{\cdot,l})^2}{\hat{p}_{k,\cdot}\hat{p}_{\cdot,l}} \sim \chi^2_{(K-1)(L-1)}$$

$$n\sum_{k=1}^{K}\frac{(\hat{p}_k - p_k(\hat{\theta}))^2}{p_k(\hat{\theta})} \sim \chi^2_{K-1-d}$$

The degrees of freedom can be thought of as

true number of parameters − number of parameters if H_0 is true.

- In the first, the true number of parameters in a multinomial distribution is not K, but rather $K-1$: the last parameter is redundant as

  $$p_K = 1 - (p_1 + \cdots + p_{K-1}).$$

  Since we know p_k^0, we don't have to estimate any parameters, so we have $K-1$ degrees of freedom.

- For the second, our parameters are $\{1, 2, \ldots, K\} \times \{1, 2, \ldots, L\}$. However, the last parameter is redundant again, and we have $KL - 1$ parameters.

  We must estimate $(K-1) + (L-1)$ parameters, one for each row and column, so we have

  $$KL - 1 - (K + L - 2) = (K-1)(L-1)$$

  degrees of freedom.

25.2 Cumulative distribution function

Let $X_1, \ldots, X_n$ be real random variables. Recall that the cdf of X_1 is defined as

$$F(t) = \Pr\{X_1 \le t\}, \forall t \in \mathbb{R}$$

and it completely characterizes the distribution of X_1.

Example 25.1

The cdf of X is $F(t) = 1 - e^{-4t}, \forall t \geq 0$. What is the distribution of X?

Exponential with parameter 4. Easy. Professor is pleased we included the parameter.

Lemma 25.2

If F is continuous, then $F(X) \sim \mathcal{U}([0,1])$.

Proof. Let X be any continuous random variable with cdf F.

Assume that the cdf is strictly increasing[15], so F is bijective from $\mathbb{R} \to (0,1)$. By definition, the cdf of $F(X)$ is

$$\Pr\{F(X) \leq t\} = \Pr\left\{X \leq F^{-1}(t)\right\} = F(F^{-1}(t)) = t.$$

So the cdf is the function for which $F(t) = t$ on $t \in (0,1)$, so $F(X) \sim \mathcal{U}([0,1])$. □

Remark 25.3. How can we sample from arbitrary probability distributions? Suppose we have a robust random number generator for $u \sim \mathcal{U}([0,1])$. Then we can compute $F^{-1}(u)$ where F is our desired distribution.

We can rewrite the cdf as

$$F(t) = \Pr\{X_1 \leq t\} = \mathbb{E}\left[\mathbb{1}_{X_1 \leq t}\right]$$

where the indicator is a Bernoulli random variable.

Definition 25.4. The **empirical cdf** of sample $X_1, \ldots, X_n$ is defined as

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} = \frac{\#i = 1, \ldots, n : X_i \leq t}{n}, \forall t \in \mathbb{R}.$$

Let Z be a random variable chosen uniformly at random among $X_1, \ldots, X_n$. That is, $Z = X_l, l \sim \mathcal{U}([n])$ and $l \perp\!\!\!\perp (X_1, \ldots, X_n)$. Then $\hat{F}_n$ is the conditional cdf of $Z \mid (X_1, \ldots, X_n)$ and F is the cdf of Z (this should be obvious).

By the law of large numbers, $\forall t \in \mathbb{R}$,

$$\hat{F}_n(t) \xrightarrow[n\to\infty]{\text{a.s.}} F(t).$$

In fact, we can do better.

Theorem 25.5 (Glivenko-Cantelli theorem, Fundamental theorem of statistics)

$$\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

[15] In general, only non-decreasing holds.

A similar result exists for convergence in distribution. By the CLT, $\forall t \in \mathbb{R}$,

$$\sqrt{n}\left(\hat{F}_n(t) - F(t)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, F(t)(1-F(t))\right).$$

This bound is impractical in practice since it depends on F. Again, we have a better result to use.

Theorem 25.6 (Donsker's theorem)

If F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}}\left|\hat{F}_n(t) - F(t)\right| \xrightarrow[n\to\infty]{d} \sup_{0\le t\le 1}|\mathbb{B}(t)|$$

where $\mathbb{B}$ is a Brownian bridge on $[0,1]$.

In the context of this class, we do not care about what a Brownian bridge is, but we should note that the left hand side *always* converges to the same distribution, independent of F.

Remark. We don't need to know these theorems for the final exam.

25.3 Kolmogorov-Smirnov test

Let $X_1, \dots, X_n$ be i.i.d. random variables with unknown cdf F and let F_0 be a continuous cdf. Consider the hypotheses

$$H_0 : F = F_0 \qquad H_1 : F \neq F_0.$$

Let $\hat{F}_n$ be the empirical cdf of the sample $X_1, \dots, X_n$. If H_0 is true, then $\hat{F}_n(t) \approx F_0(t)$ uniformly in $t \in [0,1]$.

Let

$$T_n = \sup_{t\in\mathbb{R}} \sqrt{n}\left|\hat{F}_n(t) - F_0(t)\right|.$$

By Donsker's theorem, $T_n \xrightarrow[n\to\infty]{d} z$, where Z is known (supremum of Brownian bridge, use a table).[16]

Then the **Kolmogorov-Smirnov test** with asymptotic level α is

$$\delta_\alpha^{KS} = \mathbb{1}_{T_n > q_{1-\alpha}}$$

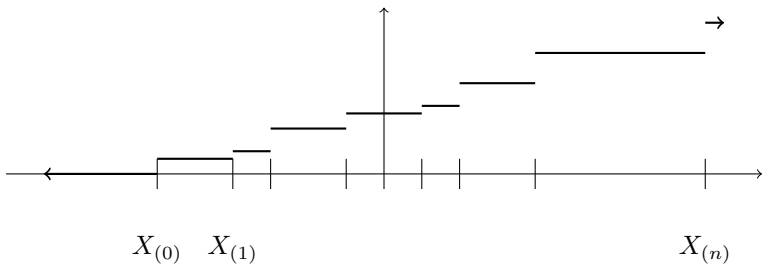
where $q_{1-\alpha}$ is the $1-\alpha$ quantile of Z. The corresponding p-value is $1 - H(T_n)$ where H is the cdf of Z.

Remark 25.7. We reiterate that this test only holds when F_0 is continuous!

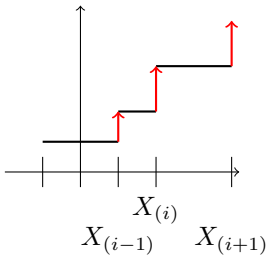
However, we run into some practical issues. How do we compute T_n for *all* values t? We could evaluate it on a lattice, but maybe the supremum falls in between. Instead, we evaluate this value on all samples.

The empirical cdf looks like the following. We observe that F_0 is non-decreasing and $\hat{F}_n$ is piecewise constant, with jumps at $t_i = X_i, i = 1, \dots, n$.

[16] The supremum is the smallest upper bound.



Now let us zoom in around $X_{(i)}$.



For each vertical jump, consider the maximum difference before and after the "jumps" (highlighted in red) and compute

$$\max_{1\le i\le n} \max\left\{\left|F_0(X_{(i)}) - \frac{i-1}{n}\right|, \left|F_0(X_{(i)}) - \frac{i}{n}\right|\right\}$$

so our estimate for T_n is

$$T_n = \sqrt{n}\max_{1\le i\le n} \max\left\{\left|F_0(X_{(i)}) - \frac{i-1}{n}\right|, \left|F_0(X_{(i)}) - \frac{i}{n}\right|\right\}$$

T_n is known as a **pivotal statistic**. If H_0 is true, then the distribution of X_i does not depend on the distribution of the X_i's and it is easy to reproduce in simulations.

When t ranges in $\mathbb{R}$, $F_0^{-1}(s)$ also ranges in $\mathbb{R}$ when s ranges in $(0,1)$ (by definition of a cdf). So we may rewrite our test statistic as

$$\begin{aligned} T_n &= \sqrt{n}\sup_{s\in(0,1)}\left|\hat{F}_n(F_0^{-1}(s)) - F_0(F_0^{-1}(s))\right| \\ &= \sqrt{n}\sup_{s\in(0,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le F_0^{-1}(s)} - s\right| \\ &= \sqrt{n}\sup_{s\in(0,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{F_0(X_i)\le s} - s\right| \end{aligned}$$

However, recall from lemma 25.2 that $F_0(X) \sim \mathcal{U}([0,1])$. So let $U_i = F_0(X_i), i = 1,\dots,n$. Then

$$T_n = \sqrt{n}\sup_{s\in(0,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{U_i\le s} - s\right| = C(U_1,\dots,U_n)$$

where C is some complicated function that might not appear on tables.

In other words, if H_0 is true and $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$,

$$T_n = \sup_{x \in [0,1]} \sqrt{n} \left| \hat{G}_n(x) - x \right|$$

where $\hat{G}_n$ is the empirical cdf of $U_1, \dots, U_n$.

Algorithmically, we can leverage this fact to provide a test with approximate (but non-asymptotic level) α.

1. Sample $U_{i,1}, \dots, U_{i,n} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$.
2. Let $\hat{G}_{i,n}$ be their empirical cdf.
3. Let $T_{i,n} = \sup_{x \in [0,1]} \sqrt{n} \left| \hat{G}_{i,n}(x) - x \right|$.

If H_0 is true, then $T_{1,n}, \dots, T_{M,n} \overset{\text{i.i.d.}}{\sim} T_n$. We can estimate the $1-\alpha$ quantile of T_n by taking the sample $1-\alpha$ quantile of $T_{1,n}, \dots, T_{M,n}$. So our test is

$$\delta_\alpha = \mathbb{1}_{T_n > \hat{q}_{1-\alpha}^{(M,n)}}$$

with approximate p-value

$$\frac{\#\{i = 1, \dots, M : T_{i,n} > T_n\}}{M}.$$

26 May 1, 2018

26.1 Bayesian statistics

So far, we've been studying frequentist statistics—that is, we assume that there is a statistical model with one true parameter. The data were generated randomly by some method, but all based on the hidden parameter.

In contrast, the Bayesian school of thought believes that the data is absolute. We have a prior belief about our parameters, and after seeing the data, we develop a posterior belief.

Example 26.1

Laplace wanted to estimate the probability that a newborn would be a boy or girl, since his impression was that there were more women than men. He was quite sure that the probability $p = \Pr\{\text{woman}\}$ was very close to 0.5, but he believed that $p > 0.5$ with 55% confidence. So he went to the hospital, looked at 100 newborns, and he did *not* update p. Rather, he updated his belief about p. Now, he believed that $p > 0.5$ with 60% confidence.

start making this a cohesive story starting here

Let's formalize this story. Let p be the proportion of women in the population. We sample n people randomly selected from the population and denote their gender as $X_1, \dots, X_n$.

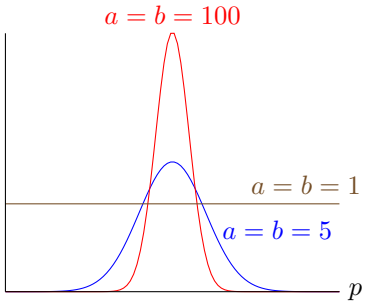
As Bayesian statisticians, we could be 90% sure that $p \in (0.4, 0.6)$ and 95% sure that $p \in (0.3, 0.8)$.

If we fix that the area under the curve sums to 1, then this is a probability density.

Definition 26.2. The Beta distribution $\mathcal{B}(a, b)$ for $a, b > 0$ is a continuous distribution on $(0, 1)$ with density

$$f_{a,b}(x) = cx^{a-1}(1-x)^{b-1}, \forall x \in (0,1)$$

where $c = \left(\int_0^1 x^{a-1}(1-x)^{b-1}\right)^{-1}$ is the normalizing constant.



If we take $a = b = 1$, then we have the uniform distribution on $[0, 1]$.

We can model our prior belief using a distribution for p, as if p were random—which is not true. However, we model our belief as if it were. So we may say

> "$X_1, \dots, X_n$ are assumed to be i.i.d. Bernoulli random variables with parameter p, conditionally on p."

Suppose our prior on p is $p \sim \mathcal{B}(a, a)$. We may say $X_1 \sim \text{Bernoulli}(\mathbb{E}[X_1])$, but we know that $\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1 \mid p]] = \mathbb{E}[p]$. According to our prior, $\mathbb{E}[p] = 1/2$.

Example 26.3

Suppose we were at a casino and we would bet on the outcomes of two independent coins. We observe that the first coin is heads. Do we change our bet for the second coin?

Of course not. The two coins are independent.

Example 26.4

Suppose instead that one coin had $p = 0.99$ and the other coin had $p = 0.01$. We randomly select one coin and toss it twice independently. We observe that the first coin is heads. Do we change our bet for the second toss?

Yes; it is more likely that we have the $p = 0.99$ coin. Given p, the two tosses are independent, but the two tosses are *not* independent.

Definition 26.5. **Probability distribution function** is a probability density function in the continuous case and a probability mass function in the discrete case.

Consider a probability distribution on parameter space Θ with some pdf $\pi(\cdot)$ as the **prior distribution**. Let $X_1, \dots, X_n$ be a sample of n random variables and let $L_n(\cdot \mid \vartheta)$ be the joint pdf of $X_1, \dots, X_n$ conditionally on ϑ where $\vartheta \sim \pi$.

Remark 26.6. $L_n(X_1, \dots, X_n \mid \vartheta)$ is the likelihood in the frequentist approach.

The conditional distribution of ϑ given $X_1, \dots, X_n$ is called the **posterior distribution** with pdf $\pi(\cdot \mid X_1, \dots, X_n)$.

Theorem 26.7 (Bayes' formula)

We know from Bayes' formula that

$$\pi(\theta \mid X_1, \dots, X_n) \propto \pi(\theta) L_n(X_1, \dots, X_n \mid \theta), \forall \in \Theta.$$

In practice, we often omit the normalizing constant since it does not depend on θ.

insert photos.

Suppose our prior was very bad. As n becomes large, however, the prior terms fall away and our new expectation is centered around $\overline{X}_n$! This has two implications:

- if we start with a bad prior, enough points will shift the prior towards the correct value, but
- if we don't have many points, then we're still screwed.

Thus, the posterior is a tradeoff between our data and the prior belief.

27 May 2, 2018

27.1 Recitation 12

Consider n random variables $X_1, \ldots, X_n$. Let $p \sim \mathcal{U}([0,1])$ and assume that conditional on p, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$.

1. The distribution of X_1 is $\text{Bernoulli}(\mathbb{E}[X_1]) = \text{Bernoulli}(1/2)$.
2. $X_1, \ldots, X_n$ are not i.i.d. since

$$\begin{aligned}\mathbb{E}[X_1 X_2] &= \mathbb{E}[\mathbb{E}[X_1 X_2 \mid p]] \\ &= \mathbb{E}[\mathbb{E}[X_1 \mid p]\,\mathbb{E}[X_2 \mid p]] \\ &= \mathbb{E}[p \cdot p] = 1/3\end{aligned}$$

   is not equal to $p^2 = 1/4$. The value of $\mathbb{E}\left[p^2\right] = 1/3$ by second moments.

Now we write probability distribution functions up to multiplicative constants. Determine the corresponding distributions.

1. $\pi(x) \propto 1, x \in [-1,1]$. Uniform on $[-1,1]$.
2. $\pi(x) \propto e^{-4x}, x \geq 0$. This is the exponential distribution $\epsilon(4)$.
3. $\pi(x) \propto e^{-4x}, x \geq 4$. This is the shifted exponential $f(x) = 4e^{4(x-4)}$.
4. $\pi(x) \propto \theta^x/x!, n \in \mathbb{N}, \theta > 0$. This is $\text{Poisson}(\theta)$.
5. $\pi(x) \propto \theta^x, x = \{1, 2, \ldots\}, \theta \in (0,1)$. This is geometric with parameter θ.
6. $\pi(x) \propto e^{-ax^2+bx}, x \in \mathbb{R}, a > 0, b \in \mathbb{R}$. This is $\mathcal{N}(b/2a, 1/2a)$. We can find the mean and variance by completing the square.
7. $\pi(x) \propto x(1-x)^2, x \in (0,1)$. This is $\text{Beta}(2,3)$.

Finally let's compute posteriors. We are given priors and conditional distributions. Hint: recall that the Gamma distribution with parameters $q > 0, \lambda > 0$ is the continuous distribution on $\mathbb{R}_{>0}$ whose density is given by

$$f(x) = \frac{\lambda^q x^{q-1} e^{-\lambda x}}{\Gamma(q)}$$

where Γ is the Euler Gamma function and its mean is q/λ.

1. $\pi(\lambda) = 1, \forall \lambda > 0$ and conditional on λ, we know that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \epsilon(\lambda)$. That is, our prior is uniform (improper distribution) and our data is exponential.

$$\begin{aligned}\pi(\lambda \mid X_1, \ldots, X_n) &\propto \pi(X_1, \ldots, X_n \mid \lambda) \cdot \pi(\lambda) \\ &= \lambda^n e^{-\lambda \sum_i X_i} \cdot 1\end{aligned}$$

   Oh look, the hint helps! This is $\text{Gamma}(n+1, \sum_i X_i)$. The conditional expectation is

$$\mathbb{E}[\lambda \mid X_1, \ldots, X_n] = \frac{n+1}{n} \cdot \frac{1}{\overline{X}_n}.$$

2. $\pi(\lambda) = 1/\lambda, \forall \lambda > 0$ and conditional on λ, we know that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$.

$$\begin{aligned}\pi(\lambda \mid X_1, \ldots, X_n) &\propto \pi(X_1, \ldots, X_n \mid \lambda) \cdot \pi(\lambda) \\ &\propto e^{-\lambda n} \lambda^{\sum_i X_i} \cdot \frac{1}{\lambda}\end{aligned}$$

The hint helps again! This is Gamma$(\sum_i X_i, n)$. The conditional expectation is $\overline{X}_n$.

3. $\theta \sim \mathcal{N}(0, 1)$ and conditional on λ, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$. Now for a quite proper prior.

$$\begin{aligned}\pi(\theta \mid X_1, \ldots, X_n) &\propto e^{xxx \sum_i (X_i - \theta)^2} \\ &= \lambda^n e^{-\lambda \sum_i X_i} \cdot 1\end{aligned}$$

The posterior is another Gaussian. $-\theta^2/2 - n\theta^2/2 + \theta \sum_i X_i^2$. The conditional expectation is

$$\mathbb{E}\left[\lambda \mid X_1, \ldots, X_n\right] = \frac{n+1}{n} \cdot \frac{1}{\overline{X}_n}.$$

28 May 3, 2018

28.1 Non-informative priors

Sometimes, we might have a lot of information about the parameter—Laplace was very sure that the ratio of women to men was nearly one to one. At other times, however, we don't know much at all, and we want to impose as little information on the prior as possible. As our professor describes:

> "Suppose we were cavemen who lived in caves our whole life, alone. We know nothing about men or women in the outside world."

A good candidate is simply the constant pdf on Θ, $\pi(\theta) \propto 1$.

- If Θ is bounded, this is simply the uniform distribution.
- If Θ is unbounded, this is *not* a proper distribution! It does not integrate to 1, so we cannot normalize it.

Definition 28.1. An **improper prior** on Θ is a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable.

Fortunately, we can still use improper priors to obtain a proper posterior.

Example 28.2

If $p \sim \text{Uniform}([0,1])$ and conditional on p, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulii}(p)$.

The posterior is

$$\pi(p \mid X_1, \dots, X_n) \propto p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}$$

which is equivalent to $\mathcal{B}(1+\sum_i X_i, 1+n-\sum_i X_i)$.

Example 28.3

Conditioned on θ, we know that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We impose a non-informative prior $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$

Though we have an improper prior, we can still compute the posterior:

$$\begin{aligned}
\pi(\theta \mid X_1, \dots, X_n) &\propto \pi(\theta) L_n(X_1, \dots, X_n \mid \theta) \\
&\propto 1 \cdot \prod_i \frac{1}{\sqrt{2\pi}} e^{-(X_i-\theta)^2/2} \\
&\propto \exp\left(-\frac{1}{2}\sum_i (X_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{n}{2}\theta^2 + \theta \sum_i X_i\right).
\end{aligned}$$

Note that we drop all terms unrelated to θ. We recognize that this expression is equivalent to $\mathcal{N}(\overline{X}_n, 1/n)$.

> "Sometimes you write a nonsense, and then another nonsense that is so bad, but then the nonsense cancels out and you get the correct answer."—veb

Essentially me trying to do arithmetic. This is why Mathematica exists.

> "After that [pset 10] I only meet you four more times! It's horrible. Maybe not for you but it's very sad."—veb

Definition 28.4. **Jeffreys prior** is

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

where $I(\theta)$ is the Fisher information matrix of the statistic model associated with $X_1, \dots, X_n$ in the frequentists approach (if it exists).

Example 28.5

In the previous examples:

- $\pi_J(p) \propto 1/\sqrt{p(1-p)}, p \in (0,1)$ so the prior is $\mathcal{B}(1/2, 1/2)$.
- $\pi_J(p) \propto 1, \theta \in \mathbb{R}$ is improper.

Example 28.6

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(e^{-10\lambda}), \lambda > 0$ What is the Fisher information?

This is a question on the first midterm. Many students simply thought,

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

for Bernoulli, right? But you can't drop the chain rule when taking the derivative with respect to λ.

Here comes the magic. Suppose $\eta = g(\theta)$ for some bijection g.

	θ	$\eta = g(\theta)$
Fisher information	$I(\theta)$	$J(\eta)$
Jeffreys prior	$\sqrt{I(\theta)}$	$\sqrt{J(\eta)}$

We can directly take the square root here, no chain rule or other shenanigans. This is known as **invariance under parametrization**.

We might be concerned that Jeffreys prior is not uniform, and it puts more weight on some values than others. However, by assuming a statistical model, we already assume some values are more likely that others—that is, larger Fisher informations result in smaller asymptotic variances. Thus, Jeffreys prior is "non-informative" in the sense that it does not introduce more information than your choice of model already implies.

28.2 Bayesian confidence region

Definition 28.7. For $\alpha \in (0,1)$, a **Bayesian confidence region** with level α is a random subset $\mathcal{R} \subseteq \Theta$ which depends on the sample $X_1, \dots, X_n$ such that

$$\Pr\{\theta \in \mathcal{R} \mid X_1, \dots, X_n\} = 1 - \alpha.$$

Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

Example 28.8

Suppose we have a distribution. Our interval is

$$\mathcal{I} = \left[\overline{X}_n - \frac{q_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{q_{1-\alpha/2}}{\sqrt{n}}\right].$$

Remark 28.9. Note that *Bayesian confidence region* and *confidence interval* are two **distinct** notions, even if they look exactly the same!

- Bayesian confidence regions have nothing asymptotic about them.
- Here we are concerned with

  $$\Pr\{\theta \in \mathcal{I} \mid X_1, \dots, X_n\} = 1 - \alpha.$$

  Previously, we cared about

  $$\Pr\{\mathcal{I} \ni \theta\} = 1 - \alpha.$$

  The former considers θ as random, while the latter considers our confidence interval as based on random data (and the parameter was never random).

28.3 Bayesian estimation

Why does our professor care about Bayesian statistics if he's a frequentist? We can use the Bayesian framework to estimate underlying parameters!

> "Belief is for statisticians. Knowledge is for frequentists."—veb

Suppose $X_1, \dots, X_n$ is associated with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$. We compute the posterior $\pi(\cdot \mid X_1, \dots, X_n)$ associated with our prior, and we output the **Bayes estimator**

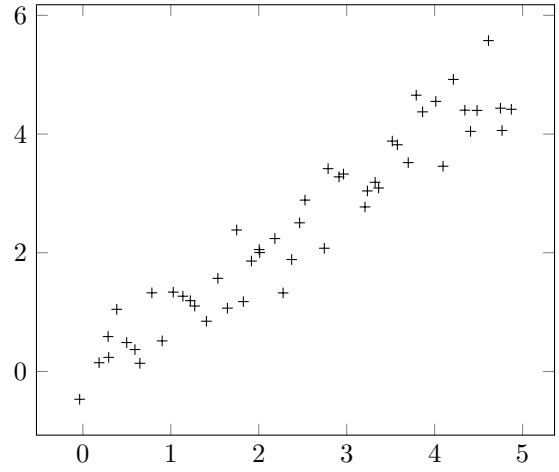
$$\hat{\vartheta}^{(\pi)} = \int_\Theta \theta \, \mathrm{d}\pi(\theta \mid X_1, \dots, X_n)$$

also known as the **posterior mean**.

29 May 8, 2018

29.1 Linear regression

Consider a cloud of i.i.d. random points $(X_i, Y_I), i = 1, \ldots, n$. We want to fit a line that "looks good." How?



For $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d$, all lines have the form

$$Y_I \approx a + X_i^T b, a \in \mathbb{R}, b \in \mathbb{R}^d.$$

Suppose we wanted to minimize

$$f(a, b) = \mathbb{E}\left[(Y - a - bX)^2\right]$$

with respect to a and b. How?

Let's bash out the math and take gradients.

$$\begin{aligned} f(a, b) &= \mathbb{E}\left[(Y - a - bX)^2\right] \\ &= \mathbb{E}\left[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX\right] \\ &= \mathbb{E}\left[Y^2\right] + a^2 + b^2\mathbb{E}\left[X^2\right] - 2a\mathbb{E}\left[Y\right] - 2b\mathbb{E}\left[XY\right] + 2ab\mathbb{E}\left[X\right] \end{aligned}$$

by linearity of expectation. We first compute the gradient and set it to zero:

$$\nabla f(a, b) = \begin{pmatrix} 2a - 2\mathbb{E}\left[Y\right] + 2b\mathbb{E}\left[X\right] \\ 2b\mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[XY\right] + 2a\mathbb{E}\left[X\right] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving with respect to a and b,

$$\begin{aligned} b &= \frac{\operatorname{cov}(X, Y)}{\operatorname{Var} X} \\ a &= \mathbb{E}\left[Y\right] - b\mathbb{E}\left[X\right] \\ &= \mathbb{E}\left[Y\right] - \frac{\operatorname{cov}(X, Y)}{\operatorname{Var} X}\mathbb{E}\left[X\right]. \end{aligned}$$

Now let's evaluate the Hessian at our (a, b) to ensure that (a, b) is a minimum. We obtain

$$\nabla f(a, b) = \begin{pmatrix} 2 & 2\mathbb{E}\left[X\right] \\ \mathbb{E}\left[X\right] & 4\mathbb{E}\left[X^2\right] \end{pmatrix}$$

which is positive definition, since diagonal entries are positive and the determinant is positive. So this function is strictly convex and (a, b) is the unique global minimum.

Now let $\epsilon = Y - (a + bX)$, so that $Y = a + bX + \epsilon$. We know that

$$\operatorname{cov}(\epsilon, X) = \operatorname{cov}(Y, X) - b\operatorname{cov}(X, X) = 0$$

by definition of b, and

$$\mathbb{E}[\epsilon] = \mathbb{E}[Y] - a - b\mathbb{E}[E] = 0$$

by definition of a. Now let's summarize and formalize our arguments.

Definition 29.1. Let X, Y be 2 random variables and assume that $Y = \alpha + \beta X + \eta$, where $\alpha, \beta \in \mathbb{R}$ and η is a random variable such that

$$\begin{cases} \mathbb{E}[\eta] = 0 \\ \operatorname{cov}(X, \eta) = 0. \end{cases}$$

Then $\alpha + \beta X$ is the **theoretical linear regression** of Y on X.

Claim 29.2. The linear regression of Y on X is unique.

Proof. The linear regression of Y on X is given by the two numbers a, b that minimize $f(a, b) = \mathbb{E}\left[(Y - a - bX)^2\right]$. We expand this to

$$\begin{aligned} f(a, b) &= \mathbb{E}\left[(\alpha + \beta X + \eta - a - bX)^2\right] \\ &= \mathbb{E}\left[((\alpha - a) + (\beta - b)X + \eta)^2\right] \\ &= \operatorname{Var}\left[((\beta - b)X + \eta)^2\right] - \mathbb{E}\left[((\alpha - a) + (\beta - b)X + \eta)^2\right]^2 \\ &= (\beta - b)^2 \operatorname{Var} X + \operatorname{Var}\eta + ((\alpha - a) + (\beta - b)\mathbb{E}[X])^2. \end{aligned}$$

Here, both terms are non-negative, so $f(a, b) \geq \operatorname{Var}\eta$ with equality if and only if

$$\begin{aligned} (\beta - b)^2 \operatorname{Var} X &= 0 \\ ((\alpha - a) + (\beta - b)\mathbb{E}[X])^2 &= 0. \end{aligned}$$

This minimum is achieved at exactly $(\alpha, \beta) = (a, b)$. □

Definition 29.3. The **least squared error** (LSE) estimator of (a, b) is the minimizer of the sum of squared errors

$$\sum_{i=1}^{n} (Y_i - a - bX_i)^2.$$

In fact, $(\hat{a}, \hat{b})$ is an M-estimator, and

$$\begin{aligned} \hat{b} &= \frac{\overline{XY} - \overline{XY}}{\overline{X^2} - \overline{X}^2} \\ \hat{a} &= \overline{Y} - \hat{b}\overline{X}. \end{aligned}$$

29.2 Multivariate linear regression

Suppose

$$Y_i = X_i^T \beta + \epsilon_i, i = 1, \ldots, n$$

where $X_i \in \mathbb{R}^d$ is a vector of **explanatory variables** or **covariates** and Y_i is a **dependent variable**. Without loss of generality, we may assume that the first coordinate of X_i is 1, so that

$$\beta = (a, b^T)^T$$

where we add the intercept a as the first term of β.

We can summarize these equations in matrix form,

$$Y = X\beta + \epsilon$$

where $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^n$, and $\epsilon \in \mathbb{R}^n$.[17] The LSE $\hat{\beta}$ satisfies

$$\hat{\beta} = \arg\min_{t \in \mathbb{R}^d} |Y - Xt|^2$$

with respect to $t \in \mathbb{R}^d$.

When t ranges in $\mathbb{R}^d$, then Xt ranges in the image of X. If $\hat{v} = X\hat{\beta}$, then $\hat{v}$ must minimize $|Y - \hat{v}|^2$ with respect to the image of X.

Notice that $|Y - \hat{v}|^2$ is the Euclidean distance between Y and v, so the error $Y - \hat{v}$ is orthogonal to $v, \forall v$ in the image of X. Equivalently, $(Xt)^T(Y - \hat{v}) = 0, \forall t \in \mathbb{R}^d$. We distribute the transpose,

$$t^T X^T (Y - \hat{v}) = 0, \forall t \in \mathbb{R}^d.$$

If this equation holds for all t, then $X^T(Y - \hat{v})$ must be equal to 0. Thus solving, we obtain that

$$X^T Y = X^T \hat{v} = X^T X \hat{\beta}$$

and our $\hat{v}$ is

$$\hat{v} = X(X^T X)^{-1} X^T Y$$

or the orthogonal projection of Y onto the subspace spanned by the columns of X.

We assume the following.

- X is deterministic and has rank d, so $n \geq d$.
- The model is **homoscedastic**—that is, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d.
- The noise vector ϵ is Gaussian,

  $$\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

  for some known or unknown $\sigma^2 > 0$.

[17] The machine learning world seems to enjoy the notation $X^T\beta$ instead.

30 May 9, 2018

"Is he giving you another pset? He can't do that!"—pfeffer

"He's doing the thing professors all do at the end of the semester: rush through a bunch of stuff really quickly."—pfeffer

30.1 Recitation 13

Let $(X_1, Y_1) \dots (X_n, Y_n)$ be i.i.d. pairs where $X_i \in \mathbb{R}^d, Y_i \in \mathbb{R}, d \geq 1$. Furthermore, let $Y_i = X_i^T \beta + \epsilon_i$ where $\beta \in \mathbb{R}^d$ and ϵ_i is a real-valued random variable with $\operatorname{cov}(X_i, \epsilon_i) = 0$.

1. The matrix version is
$$Y = X\beta + \epsilon$$
where $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times d}$ and X_i^T are the rows of X.

2. By definition, the LSE minimizes $f(t) = |Y - Xt|^2$ with respect to $t \in \mathbb{R}^d$. We compute the LSE using two methods.

   (a) We may take the gradient and set it to zero:
   $$f'(t) = -2X^T(Y - Xt) = 0.$$
   Solving, we find that $X^T Y = X^T X t$ and
   $$t = (X^T X)^{-1} X^T Y.$$
   This only holds because we assume X has rank $\geq d$. In class we also proved that this t is a global minimum since its Hessian is positive definite.
   Remark 30.1. If this is "scary," the TA's talking about it isn't going to make it less scary, so he might just hurry up and continue...

   (b) Let $v = Xt$. Rephrasing our problem, we want to minimize $|Y - v|^2$. Observe that this value is simply the Euclidean distance between Y and v, so we are looking for the vector v in the image of X that is closest to Y. Geometrically, the error $Y - X\beta \perp v, \forall v$ in the image of X, so
   $$\begin{aligned} (Xt)^T(Y - X\beta) &= 0, \forall t \in \mathbb{R}^d \\ t^T X^T (Y - X\beta) &= 0, \forall t \in \mathbb{R}^d \\ X^T(Y - X\beta) &= 0 \text{ if above holds for all } t \\ X^T Y &= X^T X \beta \\ \beta &= (X^T X)^{-1} X^T Y, \end{aligned}$$
   which is the same as above.

3. Suppose f is the pdf of X_1 and assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \perp\!\!\!\perp X$. Note that the conditional pdf of $(X_1, Y_1) \dots$ is $g(x, y) = f(x)$???

(a) Since $Y_i = X_i^T\beta + \epsilon_i$, the conditional distribution of Y_i given X_i is $\mathcal{N}(X^T\beta, \sigma^2)$.
We know that the conditional pdf is a Gaussian,

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp$$

The likelihood is

$$L_n : (\mathbb{R}^d \times \mathbb{R})^n \times (\mathbb{R}^d \times (0,\infty)) \to \mathbb{R}$$

$$\left((x_1,y_1),\ldots,(x_n,y_n),(\mu,\sigma^2)\right) \mapsto \prod_i f(x_i)(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_i (Y_i - X_i^T\beta)^2\right)$$

finish w/ photos

$$(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{|Y-X\beta|^2}{2\sigma^2}\right)$$

(b) Hence the MLE $\left(\hat{\beta},\hat{\mu}\right)$ maximizes L_n, which does not depend on f.

(c) Finally, $\hat{\sigma}^2$ maximizes

$$-\frac{n}{2}\log 2\pi\sigma^2 - \frac{|Y-X\beta|^2}{2\sigma^2}$$

and we find that $\hat{\sigma}^2 = |Y-X\beta|^2/n$.

(d) $\hat{\beta} = (X^TX)^{-1}X^Ty$.

insert image

Since ϵ_i is normally distributed, independent of X_i. So the conditional distribution of β given X is

$$\mathcal{N}_d\left(\beta, \sigma^2 AA^T\right)$$

where $AA^T = (X^TX)^{-1}X^TX(X^TX)^{-1} = (X^TX)^{-1}$.

31 May 10, 2018

31.1 Linear regression review

Recall from our previous lecture that linear regression of Y on X takes the form

$$Y = X\beta + \epsilon$$

where $\beta \in \mathbb{R}^d, Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times d}$, and $\epsilon \in \mathbb{R}^n$. In recitation and the previous lecture, we found that the LSE is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

via first order conditions and a geometric explanation. Equivalently, we may say

$$X\hat{\beta} = PY$$

where $P = X(X^T X)^{-1} X^T$, which is an orthogonal projector (easy to check that $P = P^T$).

Now let's translate some linear algebra into statistics. We assumed that X has rank d. In statistical terms, this means that the parameter β is identified. We know that

$$X\beta = X(\beta + t)$$

for all t in the kernel of X. If X has a nontrivial kernel, then β is not unique.

31.2 Linear regression with deterministic design

Example 31.1

We want to investigate if a person's age has anything to do with the number of apples he or she's eaten in a lifetime. How can we sample?

- Maybe we just sample randomly. Okay, then the X_i's are random.
- Suppose instead that we want 100 people who are 10, 20, 30,...,80.

In the latter case, the X_i's are deterministic, but the Y_i's are random.

> "Hmm... how do I do that in the US? Maybe I'll ask the IRS... Wait, I'll just ask Facebook!"—veb

Formally, let the **design matrix** X be deterministic with rank d. Furthermore, we assume that the model is homoscedastic—that is, the $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. and $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ for some known or unknown $\sigma^2 > 0$.

We have $Y_i = X_i^T \beta + \epsilon_i$ where the conditional distribution of ϵ_i given X_i is

$$\epsilon \mid X_i \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

Using this information, we find that

$$\mathbb{E}[\epsilon_i] = \mathbb{E}[\mathbb{E}[\epsilon_i \mid X_i]] = 0.$$

and

$$\begin{aligned}\operatorname{cov}(\epsilon_i, X_i) &= \mathbb{E}[\epsilon_i X_i] - \mathbb{E}[\epsilon_i]\,\mathbb{E}[X_i] \\ &= \mathbb{E}[\mathbb{E}[\epsilon_i X_i \mid X_i]] - 0 \\ &= \mathbb{E}[X_i \mathbb{E}[\epsilon_i \mid X_i]] = 0\end{aligned}$$

since given X_i, X_i itself is fixed. Here, the distribution of $\epsilon_i \mid X_i$ does not depend on X_i, so

- the distribution of ϵ_i is the same as the conditional,
- and ϵ_i is independent of X_i.

Remark 31.2. This model is unrealistic in practice. For example, if someone is really old, they might have Alzheimer's, so their answers might have more error than someone in their prime.

Remark 31.3. Linear regression appears a lot in econometrics. If $\epsilon_1, \dots, \epsilon_n$ are i.i.d., the model is homoscedastic, and otherwise it is heteroscedastic. Statisticians don't care as much.

Recall that $\hat{\beta} = (X^T X)^{-1} X^T Y$, where X is deterministic. So $Y = X\beta + \epsilon, \epsilon \mathcal{N}_n(0, \sigma^2 I)$ and

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I).$$

Let $A = (X^T X)^{-1} X^T$, so that $\hat{\beta} = AY$. Then

$$\hat{\beta} = AY \sim \mathcal{N}_d(AX\beta, A\sigma^2 I A^T) = \mathcal{N}_d(\beta, \sigma^2 (X^T X)^{-1}).$$

Notice that $\hat{\beta}$ is unbiased!

31.3 Significance tests

> "Let's play with this formula! ... It's the end of the semester and my jokes are getting worse and worse."—veb

We can calculate the quadratic risk of $\hat{\beta}$:

$$\begin{aligned}\mathbb{E}\left[\left|\hat{\beta} - \beta\right|^2\right] &= \sum_i \mathbb{E}[(]\,\hat{\beta}_j - \beta_j)^2 \\ &= \sum_i \operatorname{Var} \hat{\beta}_j \\ &= \sum_i \sigma^2 \gamma_i \text{ where } \gamma_i \text{ are the diagonal entries of } (X^T X)^{-1} \\ &= \sigma^2 \operatorname{Tr}(X^T X)^{-1}.\end{aligned}$$

Consider the hypotheses

$$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$$

and assume that σ^2 is known (for now). By the CLT,

$$\frac{1}{\sigma}(X^T X)^{-1/2}(\hat{\beta} - \beta) \sim \mathcal{N}_d(0, I).$$

If H_0 is true, then

$$\frac{1}{\sigma}(X^TX)^{-1/2}\hat{\beta} \sim \mathcal{N}_d(0, I)$$

which is a fine test statistic, except that we have d dimensions. So let's convert it into a chi-squared test:

$$\underbrace{\frac{1}{\sigma^2}\hat{\beta}^T X^T X \hat{\beta}}_{T_n} \sim \chi^2_d.$$

We reject when T_n is greater than some quantile. We may also write that $T_n = \frac{\left|X\hat{\beta}\right|^2}{\sigma^2}$.

Suppose instead that σ^2 is not known. Our likelihood is

$$\begin{aligned} L_n &: \mathbb{R}^n \times \mathbb{R}^d \times (0, \infty) \to \mathbb{R} \\ &(y_1, \dots, y_n, \beta, \sigma^2) \mapsto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - X_i^T\beta)^2}{2\sigma^2}\right) \\ L_n(Y_1, \dots, Y_n, \beta, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (Y_i - X_i^T\beta)^2\right). \end{aligned}$$

Hey look, the right hand side looks a lot like the LSE! So the LSE is actually just the MLE.

Let $\hat{\epsilon} = Y - X\hat{\beta} = X(\beta - \hat{\beta}) + \epsilon$. Plugging in our definition of $\hat{\beta}$,

$$\hat{\epsilon} = -X(X^TX)^{-1}X^T\epsilon + \epsilon = \epsilon - P\epsilon.$$

Recall that $P = X(X^TX)^{-1}X^T$ is an orthogonal projector onto the image of X. Then guess what? $\epsilon - P\epsilon$ is the projection of ϵ onto the orthogonal space. Thus we can write $\hat{\epsilon} = Q\epsilon$ and

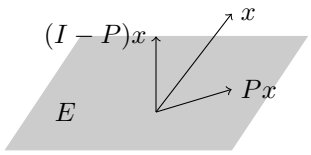
$$\hat{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 QQ^T) = \mathcal{N}_n(0, Q).$$

Since Q has rank $n - d$, $\left|\hat{\epsilon}\right|^2 / \sigma^2 \sim \chi^2_{n-d}$.

32 May 15, 2018

32.1 Generalized Cochran's theorem

Let's start with a bit of linear algebra. Let $U = (U_1 \dots U_n)^T \sim \mathcal{N}(0, I)$, or equivalently, $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let P be an orthogonal projector in $\mathbb{R}^n$.[18] We denote E as the image of P. Consider the matrix $I - P$. The vector $(I-P)x$ is the orthogonal projector onto the kernel of P.



Proposition 32.1

PU follows the properties that

- $PU \perp\!\!\!\perp (I-P)U$ and
- $PU \sim \mathcal{N}(0, P)$.

Proposition 32.2

It follows that

- $|PU|^2 \sim \chi^2_k$ and
- $|(I-P)U|^2 \sim \chi^2_{n-k}$.

Recall that $Y = X\beta + \epsilon$ and the linear regression of Y on X is $\hat{\beta} = (X^TX)^{-1}X^TY$. We denote $X\hat{\beta}$ as the **prediction** of Y.

Theorem 32.3

$\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Proof. We previously solved that $X\hat{\beta} = PY$, where $P = X(X^TX)^{-1}X^T$ is the orthogonal projection of Y onto the image of X.

Consider the quantity

$$X(\hat{\beta} - \beta) = X\hat{\beta} - X\beta = PY - X\beta.$$

By definition, $X\beta$ is in the image of X, and the projection of $X\beta$ into the image of X is $X\beta$. Thus,

$$PY - X\beta = PY - P(X\beta) = P(Y - X\beta) = P\epsilon.$$

[18] By definition, P is symmetric and $PP = P$. For any vector $x \in \mathbb{R}^n$, Px is the closest vector in E to x.

So we may conclude that

$$\frac{1}{\sigma}X(\hat{\beta}-\beta)=PU$$

where $U=\frac{1}{\sigma}\epsilon\sim\mathcal{N}_n(0,I)$. Now consider $(I-P)U$. We expand as

$$\begin{aligned}(I-P)U&=U-PU\\&=\frac{1}{\sigma}\left(\epsilon-X(\hat{\beta}-\beta)\right)\\&=\frac{1}{\sigma}\left(Y-X\hat{\beta}\right).\end{aligned}$$

Since $PU\perp\!\!\!\perp(I-P)U$, we may conclude that

$$X\hat{\beta}-X\beta\perp\!\!\!\perp Y-X\hat{\beta}.$$

First, observe that $X\beta$ is deterministic, so $X\hat{\beta}\perp\!\!\!\perp Y-X\hat{\beta}$. Now note the following:

$$(X^TX)^{-1}(X^TX)\hat{\beta}\perp\!\!\!\perp Y-X\hat{\beta}$$

and we may conclude that $\hat{\beta}\perp\!\!\!\perp Y-X\hat{\beta}$. We take the squared norm of the right side,

$$\hat{\beta}\perp\!\!\!\perp\frac{1}{n}\left|Y-X\hat{\beta}\right|^2=\hat{\sigma}^2$$

where $\hat{\sigma}^2=\frac{1}{n-d}\left|Y-X\beta\right|^2=\frac{n}{n-d}\hat{\sigma}^2_{\text{MLE}}$ is an unbiased estimator.[19] □

Remark 32.4. Recall that long ago, we mentioned that a *corrected* sample variance was

$$\frac{1}{n-1}\sum_i(X_i-\overline{X}_n)^2$$

instead of our standard $1/n$. Finally, we see why.

Corollary 32.5

Multiply by n to obtain

$$\frac{n\hat{\sigma}^2}{\sigma^2}=\left|(I-P)U\right|^2\sim\chi^2_{n-d}$$

where $d=\text{rank }X$ and $\hat{\sigma}^2$ is the MLE.

This looks a lot like Cochran's theorem!

32.2 Significance tests, ctd.

connect with previous lecture

$\hat{\beta}_j=\mathcal{N}\left(\beta_j,[\sigma^2(X^TX)^{-1}]_j\right)$

So by the CLT,

$$\frac{\hat{\beta}_j-\beta_j}{\sigma\sqrt{\gamma_j}}\sim\mathcal{N}(0,1)$$

[19] The MLE is biased as $\mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right]=\frac{\sigma^2(n-d)}{n}$.

where $\gamma_j = [\sigma^2(X^TX)^{-1}]_j$. If σ^2 is known, then let

$$T = \frac{\hat{\beta}_j}{\sigma\sqrt{\gamma_j}}$$

and our test is $\delta = \mathbb{1}_{|T|>q_{1-\alpha/2}}$ where $q_{1-\alpha/2}$ is a quantile of the standard Gaussian.

If σ^2 is unknown, then let

$$T^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2\gamma_J}}$$

where $\hat{\sigma}^2$ is the *unbiased* estimator from the previous section. We may expand $T^{(j)}$ as

$$\begin{aligned} T^{(j)} &= \frac{\hat{\beta}}{\sqrt{\sigma^2\gamma_J}} \cdot \frac{1}{\sqrt{\hat{\sigma}^2/\sigma^2}} \\ &= \frac{\hat{\beta}_j}{\sqrt{\sigma^2\gamma_j}} \cdot \frac{1}{\sqrt{\frac{1}{n-d}\frac{\left|Y-X\hat{\beta}\right|^2}{\sigma^2}}} \end{aligned}$$

If H_0 is true, the first term is a standard Gaussian, and $\frac{\left|Y-X\hat{\beta}\right|^2}{\sigma^2} \sim \chi^2_{n-d}$. Remember what this looks like? A student random variable!

32.3 Implicit hypotheses (linear)

Let G be a $k \times d$ matrix with rank $G = k, k \leq d$ and $\lambda \in \mathbb{R}^k$. Consider the hypotheses

$$H_0 : G\beta = \lambda \qquad H_1 : G\beta \neq \lambda.$$

That is, we only want to test if a linear function on a subset of the coordinates is true.

Example 32.6

If we wanted to test $H_0 : \beta_1 = \beta_2$, when we could say $\beta_1 - \beta_2 - 0$, and $G\beta = 0$ where $G = \begin{pmatrix} 1 & -1 & 0 & \dots \end{pmatrix} \in \mathbb{R}^{1\times p}$.

Example 32.7

If we wanted to test $H_0 : \beta_1 = \beta_2 = \beta_3$, then we have equations $\beta_1 = \beta_2$ and $\beta_2 = \beta_3$. So we may define

$$G = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \end{pmatrix}.$$

Recall that $\hat{\beta} \sim \mathcal{N}_d(\beta, \sigma^2(X^TX)^{-1})$ in the previous problem. We easily see that

$$G\hat{\beta} - \lambda \sim \mathcal{N}_k(G\beta - \lambda, G\sigma^2(X^TX)^{-1}G^T).$$

If H_0 is true, then

$$G\hat{\beta} - \lambda \sim \mathcal{N}_k(0, \sigma^2\Sigma)$$

where $\Sigma = G^T(X^TX)^{-1}G^T$ for "the rest of the semester." We assumed previously that rank $G = k$,[20] so we may apply our usual simplifications

$$\frac{1}{\sigma}\Sigma^{-1/2}(G\hat{\beta} - \lambda) \sim \mathcal{N}_k(0, I).$$

Hence,

$$\frac{1}{\sigma^2}(G\hat{\beta} - \lambda)^T\Sigma^{-1}(G\hat{\beta} - \lambda) \sim \chi^2_k.$$

What is σ^2 is unknown? Use $\hat{\sigma}^2$! Let

$$S_n = \frac{1}{\hat{\sigma}^2}(G\hat{\beta} - \lambda)^T\Sigma^{-1}(G\hat{\beta} - \lambda).$$

If H_0 is true, then $S_n \sim F_{k,n-p}$ where F is the Fisher distribution, defined below.

Definition 32.8. The **Fisher distribution** with p and q degrees of freedom, denoted by $F_{p,q}$ is the distribution of $\frac{U/p}{V/q}$, where

- $U \sim \chi^2_p, V \sim \chi^2_q$, and
- $U \perp\!\!\!\perp V$.

Remark 32.9. Here we cannot use the $1 - \alpha/2$ quantile for S_n, because it is not symmetric! The Fisher distribution is strictly positive.

[20] This makes perfect sense. It means that we don't have redundant equations.

33 May 17, 2018

33.1 Review

Today we will review the main ideas from this course and then discuss some further topics in statistics.

33.1.1 Basics

We started with the building blocks of statistics—you must *own* the law of large numbers and the central limit theorem.

Remark 33.1. The law of large numbers and the central limit theorem are *only* for i.i.d. sample averages.

Example 33.2

Suppose we have $X_1, \dots, X_n$ i.i.d. real valued random variables. We would like to estimate the variance $V = \operatorname{Var} X_1$.

Recall that the variance has two equivalent formulas,

$$\operatorname{Var} X_1 = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^2\right] = \mathbb{E}\left[X_1^2\right] - \mathbb{E}[X_1]^2.$$

The sample average $\hat{V}$ also has two equivalent formulas,

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \overline{X_n^2} - \overline{X}_n^2.$$

Notice that the first formula *is* a sample average, but the random variables $(X_i - \overline{X}_n)^2$ are *not* i.i.d. However, the second formula is the sum of two sample averages (squared), so we *can* use the LLN and show that the sample variance is consistent.

To show asymptotic normality, we *cannot* apply the CLT twice on both terms of the second formula. The difference of two asymptotically normal random variables is not guaranteed to be asymptotically normal. Instead, we should use the Delta method and CLT on the two-dimensional variable $\begin{pmatrix} \overline{X}_n & \overline{X_n^2} \end{pmatrix}$.

Let $\hat{\theta}$ be an estimator of some $\theta \in \mathbb{R}^d$ with

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, \Sigma(\theta))$$

and assume that $\Sigma(\theta)$ is invertible. Then

$$\sqrt{n}\Sigma(\theta)^{1/2}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, I).$$

Consider the hypotheses

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

If H_0 is true, we may write that

$$\underbrace{\sqrt{n}\Sigma(\theta_0)^{1/2}(\hat{\theta} - \theta_0)}_{T_n} \xrightarrow[n\to\infty]{d} z$$

where $z \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, I)$ and T_n is our test statistic. However, it is often more convenient to compute the squared norm $|T_n|^2$,

$$|T_n|^2 = n(\hat{\theta} - \theta_0)^T \Sigma(\theta_0)^{-1}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} \chi^2_d.$$

Finally let's talk Slutsky's theorem. Instead of first assuming that H_0 is true, we may write that

$$\sqrt{n}\Sigma(\hat{\theta})^{1/2}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}_d(0, I).$$

which follows from Slutsky's theorem, where the Σ terms may be expanded as

$$\underbrace{\Sigma(\hat{\theta})^{-1/2}\Sigma(\theta)^{1/2}}_{\xrightarrow[n\to\infty]{\mathbb{P}} I} \underbrace{\sqrt{n}\Sigma(\theta)^{-1/2}(\hat{\theta} - \theta)}_{\xrightarrow[n\to\infty]{d} z}.$$

> "I'm going to write on the left board because it's the first time I'm writing on the left board. It feels great!"—veb

Remark 33.3. If you use Slutsky's theorem, make sure you always know how the terms decompose.

Example 33.4

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. By the CLT,

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

It is clearly false that

$$\sqrt{n}\frac{\overline{X}_n - \overline{X}_n}{\sqrt{\overline{X}_n(1 - \overline{X}_n)}} = 0 \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

We can indeed replace the two p's in the denominator, since we can multiply by the ration of p's to $\overline{X}_n$'s. However, there is no way we can replace the numerator p's. Let's try:

$$\sqrt{n}\frac{\overline{X}_n - \overline{X}_n}{\sqrt{p(1-p)}} = \underbrace{\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}}_{\xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)} + \sqrt{n}\frac{p - \overline{X}_n}{\sqrt{p(1-p)}}.$$

Sure, $p - \overline{X}_n$ converges to 0, but $\sqrt{n}(p - \overline{X}_n)$ doesn't converge to anything!

Example 33.5

The quantity $1/\log n \to 0$, but $\sqrt{n}/\log n \to \infty$.

33.1.2 Parametric estimation

We defined a statistical model as the tuple $\left(E, \{\Pr_\theta\}_{\theta\in\Theta}\right)$ where E is the sample space and $\{\Pr_\theta\}_{\theta\in\Theta}$ is a family of distributions. That is, there exists a function σ that maps $\theta \mapsto \Pr_\theta$. If θ is identified, then σ is injective.

Example 33.6

If $\Theta = (0, \infty)$ and $\Pr_\theta = \mathcal{U}([\theta, 2\theta])$, then θ is identified as the left boundary of the support.

33.1.3 Likelihood

The professor reminds us to remember the domain and codomain.

Example 33.7

Let $X_1, \dots, X_n \sim f_\theta$ with parameter θ.

The likelihood always takes the form

$$\begin{aligned} L_n : E^n \times \Theta &\to \mathbb{R} \\ (x_1, \dots, x_n, \theta) &\mapsto f_\theta(X_1, \dots, X_n), \text{ evaluated at } x_1, \dots, x_n. \end{aligned}$$

Example 33.8

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. pairs where $X_1 \sim f$ on $\mathbb{R}^d$ and conditional on X_1, $Y_1 - X_1^T\beta \sim \mathcal{N}(0, \sigma^2)$ for some unknown $\beta \in \mathbb{R}^d, \sigma^2 > 0$.

Equivalently, we may write $Y_1 = X_1^T\beta + \epsilon_1$ where conditional on X_1, $\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$. The likelihood is

$$\begin{aligned} L_n : (\mathbb{R}^d \times \mathbb{R})^n \times \mathbb{R}^p \times (0, \infty) &\to \mathbb{R} \\ \big((x_1, y_1), \dots, (x_n, y_n), \beta, \sigma^2\big) &\mapsto \prod_{i=1}^n f_{X_i}(x_i) f_{Y_i | X_i = x_i}(y_i). \end{aligned}$$

We can use Baye's formula to find the joint distribution. Conditional on X_1, $Y_1 \sim \mathcal{N}(X_1^T\beta, \sigma^2)$. So our final likelihood is

$$L_n\big(\big((x_1, y_1), \dots, (x_n, y_n), \beta, \sigma^2\big) = \prod_{i=1}^n f(x_i) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^T\right).$$

Remark 33.9. When we refer to "conditional on X, $Y \sim f$", we may *not* write $Y \mid X \sim f$. This notation is wrong. Oops I'm guilty of this. LOL

We move on to Fisher information. Let

$$\begin{aligned} \varphi : \mathbb{R}^d &\to \mathbb{R} \\ t &\mapsto t^T A t + \mu^T t \end{aligned}$$

where $A \in \mathbb{R}^{d \times d}$ is symmetric and $\mu \in \mathbb{R}^d$.

Remark 33.10. It's subtle, but we may write $t = (t_1, t_2, \dots, t_n)$ to refer to a n-dimensional column vector. OMG this would save me so much time.

The gradient of φ is $\nabla\varphi(t) = 2At + \mu$. The Hessian is $\nabla^2\varphi(t) = 2A$. If A is not symmetric, then replace $2A \leftarrow A + A^T$.

34 May 21, 2018

34.1 Final office hours

34.1.1 Slutsky's theorem

We review question 2 from problem set 10.

Example 34.1

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0, \theta])$ given θ, with some prior on θ.

The Bayesian estimator was

$$\hat{\theta} = \frac{n-2}{n-1} \max_i X_i.$$

Is $\hat{\theta}$ asymptotically normal? We cannot apply the theorem for the MLE, but this does *not* imply that $\hat{\theta}$ is not asymptotically normal. Instead, $\hat{\theta}$ is not asymptotically normal because it is biased, and $\sqrt{n}(\hat{\theta} - \theta) \le 0$. Thus $\hat{\theta} - \theta$ does not converge in distribution to anything centered around 0.

Example 34.2

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, given p, $p \sim \text{Beta}(a, b)$.

The Bayesian estimator was

$$\hat{p}^{(i)} = \frac{a + \sum_i X_i}{a + b + n}.$$

How do we show asymptotic normality? Many people divided the top and bottom by n, okay:

$$\hat{p}^{(i)} = \frac{a/n + \overline{X}_n}{(a+b)/n + 1}.$$

Afterwards, people wrote $\frac{a/n + \overline{X}_n}{(a+b)/n+1} \to \overline{X}_n$, which is wrong! The right hand side cannot depend on n. Instead, we should say

$$\frac{a/n + \overline{X}_n}{(a+b)/n+1} \xrightarrow[n\to\infty]{\mathbb{P}} p.$$

Now onto asymptotic normality.

$$\begin{aligned}
\sqrt{n}(\hat{p}^{(i)} - p) &= \sqrt{n}\left(\frac{a/n + \overline{X}_n}{(a+b)/n+1} - p\right) \\
&= \sqrt{n}\left(\frac{a/n + \overline{X}_n}{(a+b)/n+1} - \overline{X}_n + \overline{X}_n - p\right) \\
&= \sqrt{n}\left(\frac{a/n + \overline{X}_n}{(a+b)/n+1} - \overline{X}_n\right) + \underbrace{\sqrt{n}\left(\overline{X}_n - p\right)}_{\xrightarrow[n\to\infty]{d} z, z\sim\mathcal{N}(0,1)}.
\end{aligned}$$

Remember, whenever we use Slutsky's theorem, we need to decompose our quantity of interest into two terms—one that converges to a constant, and one that converges to something else. The right term converges to z, but the left term is more complicated.

$$\begin{aligned}\sqrt{n}\left(\frac{a/n+\overline{X}_n}{(a+b)/n+1}-\overline{X}_n\right) &= \sqrt{n}\left(\frac{a/n+\overline{X}_n-((a+b)/n+1)\overline{X}_n}{(a+b)/n+1}\right)\\ &= \sqrt{n}\left(\frac{a/n-(a+b)\overline{X}_n/n}{(a+b)/n+1}\right)\\ &= \frac{1}{\sqrt{n}}\frac{a-(a+b)\overline{X}_n}{(a+b)/n+1} \xrightarrow[n\to\infty]{\mathbb{P}} 0.\end{aligned}$$

How do we arrive at this result? Our quantity of interest is *almost* just the sample mean:

$$\sqrt{n}(\hat{p}^{(i)}-p) = \underbrace{\sqrt{n}(\overline{X}_n-p)}_{\xrightarrow[n\to\infty]{d}\mathcal{N}(0,p(1-p))} + \underbrace{\text{correction}}_{\xrightarrow[n\to\infty]{\mathbb{P}}0}.$$

34.1.2 Remarks about laziness

Remark 34.3. Read the exam carefully. Sometimes there will be convergence in distribution, followed by non-asymptotic tests. Be careful!

> "In life, I'm a pretty lazy person, but you've never seen me write something like $\sum_i X_i$, and I'm probably more lazy that 50% of the people in this room"—veb

Welp I guess I'm guilty of that.

> "The extreme level of lazy is
>
> $$\sqrt{n}(\overline{X}_n-p) \longrightarrow \mathcal{N}(0,p(1-p)).$$
>
> A little better is
>
> $$\sqrt{n}(\overline{X}_n-p) \xrightarrow{(d)} \mathcal{N}(0,p(1-p)).$$
>
> Best is
>
> $$\sqrt{n}(\overline{X}_n-p) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,p(1-p)).$$
>
> Don't be lazy."—veb

At least I'm not guilty of this!

- Some people will say "approaches." Please don't. Use good notation, ya?
- Don't put an n on the right side of a limit. The n goes to infinity. It's gone.
- Don't drop your $\sqrt{n}$ terms.

  $$\sqrt{n}\left(\overline{X}_n-a/n-p\right) = \sqrt{n}\left(\overline{X}_n-p\right)+a/\sqrt{n}$$

  where we may apply Slutsky's theorem, but

  $$\sqrt{n}\left(\overline{X}_n-a/\log n-p\right)\to\infty$$

  diverges due to the $\sqrt{n}/\log n$.

34.1.3 Notes on chi-squared intuition

Question 34.4. How do we think about χ^2 dimensions intuitively?

Consider the following case.

Example 34.5
If $U = (u_1, u_2, \ldots, u_n) \sim \mathcal{N}_n(0, I)$ and P is an orthogonal projection on a linear subspace of $\mathbb{R}^d$ of dimension d, what is the distribution of $|PU|^2$?

A simple such orthogonal projection takes

$$(u_1, u_2, \ldots, u_n) \to (u_1, u_2, \ldots, u_d, 0, \ldots, 0).$$

Likewise, if we take $I - P$, then we are left with $d+1, d+2, \ldots, n$ as nonzero terms. Hey look, that's why proposition 32.2 has χ^2 degrees of freedom of d and $n - d$.

34.1.4 Positive definiteness and unique minimum

Recall that the ridge estimator was defined as

$$\arg\min_{t \in \mathbb{R}^d} f(t) = |Y - Xt|^2 + \lambda |t|^2 .$$

The Hessian is

$$\nabla^2 f(t) = 2(X^T X + \lambda I).$$

A common pitfall is to say

> the Hessian is positive definite, so it is strictly convex, so it has a unique minimum

which is *wrong*. Consider $g(t) = e^t, t \in \mathbb{R}$. This function is strictly convex but does *not* have a minimum at all.

We may show that the Hessian is positive definite if

$$\forall u \in \mathbb{R}^d \setminus \{0\}, u^T 2(X^T X + \lambda I)u > 0.$$

Let's expand:

$$\begin{aligned} u^T 2(X^T X + \lambda I)u &= 2(u^T X^T X u + \lambda u^T u) \\ &= 2\left((Xu)^T (Xu) + \lambda |u|^2\right) \\ &= 2\left(|Xu|^2 + \lambda |u|^2\right). \end{aligned}$$

Notice, we don't know if X is full rank, and we don't need to. What matters is that λ is positive and u is non-zero, so the whole quantity is strictly positive.

> "You can't just make a right turn because you want to. You have to signal and follow the rules. Why do you write lines in math? You write what you feel like—and you have a car accident! You're not respecting the rule so you're going to jail."—veb

> "Ask anyone who studies algebraic topology and they'll tell you that statistics is not math. Statistics is for mathematicians who fail."—veb

I feel the last quote so much omg lol.

34.1.5 Covariance matrices

Covariance matrices are always positive semi-definite!

Suppose Σ is the covariance matrix of X. Then $u^T\Sigma u$ is the covariance matrix of u^TX, which is always non-negative. If $u \neq 0$ and Σ is positive definite, then we may say that $\operatorname{Var} u^TX > 0$ which is a strict inequality.

34.1.6 Cochran's theorem

> "If you have 'Cochran's theorem' and 'converge' in the same sentence, you're wrong."—veb

One-dimensional	Multi-dimensional
$Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.	$Y_1, \dots, Y_n$ independent with $Y_i \sim \mathcal{N}\left(X^T\beta, \sigma^2\right)$ where $\beta \in \mathbb{R}^d$.
$\hat{\mu} = \overline{Y}_n$	$\hat{\beta} = (X^TX)^{-1}X^TY$
$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu})^2$	$\hat{\sigma}^2 = \frac{1}{n}\left\lvert Y - X\hat{\beta}\right\rvert^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^T\hat{\beta})^2$.
$\hat{\mu} \perp\!\!\!\perp \hat{\sigma}^2$	$\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$
$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$	$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-d}$

In linear regression, the prediction error is

$$\begin{aligned}\mathbb{E}\left[\left\lvert Y - X\hat{\beta}\right\rvert^2\right] &= \sigma^2\mathbb{E}\left[\left\lvert Y - X\hat{\beta}\right\rvert^2/\sigma^2\right] \\ &= \sigma^2 \underbrace{\mathbb{E}\left[n\hat{\sigma}^2/\sigma^2\right]}_{\chi^2_{n-d}} \\ &= \sigma^2(n-d)\end{aligned}$$

The professor will not divulge anything about the exam format. There may be a quiz, there may not, but it's good to be prepared! How many problems? It doesn't matter either! I could give you one big problem. It's a final exam so it covers *everything*. Absolutely everything.

34.1.7 Wald's test review

Let $X_1, \ldots, X_n \sim \epsilon(\lambda)$ and suppose we wanted to compute Wald's test for hypotheses

$$H_0 : \lambda = 1 \qquad H_1 : \lambda \neq 1.$$

First, we show that the MLE is asymptotically normal:

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^2).$$

Next, we rescale to a standard distribution:

$$\sqrt{n}(\hat{\lambda} - \lambda)/\lambda \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

Finally, we apply Slutsky's theorem:

$$\sqrt{n}\frac{\hat{\lambda} - \lambda}{\lambda} \cdot \frac{\lambda}{\hat{\lambda}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

Our test statistic is

$$T_n = \sqrt{n}\frac{\hat{\lambda} - 1}{\hat{\lambda}}$$

and one potential test is $\delta = \mathbb{1}_{|T_n| > q_{1-\alpha/2}}$

Usually we might take the squared norm of T_n, but this is not necessary for one dimension. Wald's test squares T_n to obtain that

$$\delta^W = \mathbb{1}_{T_n^2 > q_{1-\alpha}} = \mathbb{1}_{|T_n| > \sqrt{q_{1-\alpha}}}$$

where $q_{1-\alpha}$ correspond to the $1 - \alpha$ quantile of χ_1^2. Notice here that the square root of χ_1^2 is simply the standard normal, so we don't need to square (and in fact, tables for the standard normal are generally more precise than equivalent tables for the chi-squared).

A Acknowledgements

A big thank you to the following people for pointing out errata and giving suggestions!

- Farrell Eldrian Wu
- Leanne Wang
- Li Wang