18.650 Notes

Lecturer: Promit Ghosal

Andrew Liu

Spring 2023

My notes for 18.650, “Statistics for Applications”. The instructor for this course was Promit Ghosal, C.L.E. Moore Instructor (`https://sites.google.com/view/promit-ghosal/home`).

Last updated on Saturday 27th May, 2023.

Contents

1 February 7, 2023 4

1.1 Introduction . . . 4

1.2 Review: Fundamental Theorems . . . 4

1.3 Review: Notions of Convergence . . . 4

2 February 9, 2023 7

2.1 Delta Method . . . 7

2.2 Kissing Experiment (confidence intervals) . . . 8

3 February 14, 2023 10

3.1 Statistical Models and Identifiability . . . 10

3.2 Estimation . . . 11

3.3 More on Confidence Intervals . . . 12

4 February 15, 2023 (R) 14

5 February 16, 2023 17

5.1 Maximum Likelihood Estimation . . . 17

5.2 Relative Entropy . . . 18

5.3 Fisher Information, Cramer Rao . . . 19

6 February 23, 2023 22

6.1 M-estimation . . . 22

6.2 Hypothesis Testing . . . 23

6.3 Errors . . . 24

7 February 28, 2023 26

7.1 p-values . . . 26

7.2 Parametric Hypothesis Testing . . . 26

8 March 7, 2023 28

8.1 t-test . . . 28

8.2 Goodness-of-fit Test . . . 30

9 March 9, 2023 31

9.1 Empirical CDF . . . 31

9.2 Kolmogorov-Smirnov Test 33
9.3 Computing values for KS 33
9.4 Kolmogorov-Lilliefors Test 34
9.5 Linear Regression 34

10 March 16, 2023 35
10.1 Multivariate Linear Regression 35
10.2 Significance Testing 37
10.3 Non-parametric Regression 37

11 March 21, 2023 38
11.1 Nonparametric Regression 38
11.2 Exponential Family 39

12 March 23, 2023 40
12.1 Generalized Linear Models 40

13 April 4, 2023 41
13.1 Principal Component Analysis 41

14 April 11, 2023 43
14.1 Classification 43

15 April 13, 2023 45
15.1 Nearest Neighbors Classifier 45
15.2 Determinant Analysis 46

16 April 18, 2023 47
16.1 Bayesian vs Frequentist statistics 47
16.2 Choices of Priors 48

17 April 25, 2023 49
17.1 Bootstrapping 49

18 April 27 50
18.1 Exponential Distributions 50
18.2 Bayesian Statistics 51

1 February 7, 2023

1.1 Introduction

General goal of this class is to get better at statistical methods, understand their applicability, and their limitations.

Midterm dates: March 3rd, April 6th, and May 4th. They take place during class. There is no final. There is also on data presentation project. Grading is $(MT1 + MT2 + MT3 + Project + HW)/5$.

1.2 Review: Fundamental Theorems

Let $X_1, \ldots,$ be i.i.d. r.v. (independent and identically distributed random variables) with $\mathbb{E}[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2$.

Theorem 1.1 (Strong Law of Large Numbers)

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{a.s.} \mu.$$

The weak law of large numbers says the same thing, with convergence in probability (instead of almost sure convergence).

Theorem 1.2 (Central Limit Theorem)

$$\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

1.3 Review: Notions of Convergence

Definition 1.3 (Almost Surely)

We say that $Y_n \xrightarrow[n\to\infty]{a.s.} Y$ if

$$\mathbb{P}[\omega : Y_n(\omega) \xrightarrow[n\to\infty]{} Y(\omega)] = 1.$$

Definition 1.4 (In Probability)
We say that $Y_n \xrightarrow[n\to\infty]{p} Y$ if

$$\lim_{n\to\infty} \mathbb{P}[|Y_n - Y| > \varepsilon] = 0,$$

for all $\varepsilon > 0$.

Definition 1.5 (In L^p)
We say that $Y_n \xrightarrow[n\to\infty]{L^p} Y$ if

$$\lim_{n\to\infty} \mathbb{E}[|Y_n - Y|^p] = 0.$$

Definition 1.6 (In Distribution)
We say that $Y_n \xrightarrow[n\to\infty]{d} Y$ if

$$\lim_{n\to\infty} \mathbb{P}[Y_n \leq x] = \mathbb{P}[Y \leq x],$$

for all $x \in \mathbb{R}$ where the CDF of Y is continuous.

The following are equivalent to converging in distribution:

- $\lim_{n\to\infty} \mathbb{E}[f(Y_n)] = \mathbb{E}[f(Y)]$ for all continuous bounded functions f.
- $\lim_{n\to\infty} \mathbb{E}[\exp(ixY_n)] = \mathbb{E}[\exp(ixY)]$ for all $x \in \mathbb{R}$.

Relationships between the types of convergence:

$$Y_n \xrightarrow[n\to\infty]{a.s.} Y \implies Y_n \xrightarrow[n\to\infty]{p} Y \implies Y_n \xrightarrow[n\to\infty]{d} Y.$$

Also, if $q \geq p \geq 1$,

$$Y_n \xrightarrow[n\to\infty]{L^q} Y \implies Y_n \xrightarrow[n\to\infty]{L^p} Y.$$

Operations and Convergence:

- If f is a continuous function, then

$$Y_n \xrightarrow[n\to\infty]{} Y \implies f(Y_n) \xrightarrow[n\to\infty]{} f(Y)$$

holds for all three modes of convergence.

- If f is a continuous function, then

  $$(X_n, Y_n) \xrightarrow[n\to\infty]{} (X, Y) \implies f(X_n, Y_n) \xrightarrow[n\to\infty]{} f(X, Y)$$

  holds for all three modes of convergence. For example, if $f(x,y) = ax + by$, or $f(x,y) = xy$, or $f(x,y) = x/y$ with $y \neq 0$.

- $X_n \xrightarrow[n\to\infty]{a.s.} X$ and $Y_n \xrightarrow[n\to\infty]{a.s.} Y$ implies $(X_n, Y_n) \xrightarrow[n\to\infty]{a.s.} (X, Y)$. The same holds for convergence in probability. In fact, any mix of convergences in probability or almost surely will work (i.e., two convergences in probability implies convergence almost surely, etc). However, the same cannot be said if at least one of the single distributions converges in distribution, with the only exception being Slutsky's Theorem.

Theorem 1.7 (Slutsky's Theorem)
Suppose

- $Y_n \xrightarrow[n\to\infty]{d} Y$
- $Z_n \xrightarrow[n\to\infty]{p} c$, where c is a real number.

Then,

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

Example 1.8
Slutsky's theorem implies $Y_n + Z_n \xrightarrow[n\to\infty]{d} Y + c$, $Y_n Z_n \xrightarrow[n\to\infty]{d} cY$, etc.

This is equivalent to showing that $(Y_n, Z_n) \xrightarrow[n\to\infty]{d} (Y, c)$ implies $Y_n + Z_n \xrightarrow[n\to\infty]{d} Y + c$. To see this, take the result of the convergence operations listed above, with $f(X, Y) = X + Y$, $f(X, Y) = XY$, or whatever you want.

2 February 9, 2023

2.1 Delta Method

Last time, we learned about Slutsky's Theorem. Why is it the case that

$$\left.\begin{matrix} Y_n \xrightarrow[n\to\infty]{d} Y \\ Z_n \xrightarrow[n\to\infty]{d} Z \end{matrix}\right\} \not\Rightarrow Y_n + Z_n \xrightarrow[n\to\infty]{d} Y + Z?$$

Consider the following counterexample:

$$Y_n = \sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1) \quad \text{and} \quad Z_n = -Y_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

In this case, $Y_n + Z_n \not\to Y + Z = \mathcal{N}(0,2)$.

Theorem 2.1 (The Delta Method)
Suppose

- $\sqrt{n}(Y_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2)$
- g is continuously differentiable at θ

Then,

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

Note the similarities to the CLT. If we treat $Y_n = \sum_{i=1}^n X_i$, then our θ is $\mathbb{E}[X_i]$. The only difference here is that we're calculating the distribution of $g(Y_n)$ instead of Y_n itself.

Lemma 2.2
If $|Y_n - X_n| \xrightarrow[n\to\infty]{p} 0$ and $X_n \xrightarrow[n\to\infty]{d} X$, then $Y_n \xrightarrow[n\to\infty]{d} X$.

Proof. Let $Z_n = Y_n - X_n$. By Slutsky's theorem, $(X_n, Z_n) \xrightarrow[n\to\infty]{d} (0, X)$, which implies $X_n + Z_n = Y_n \xrightarrow[n\to\infty]{d} X$. □

2.2 Kissing Experiment (confidence intervals)

Let p be the proportion of couples that turn their head to the right when kissing. Observe $n = 124$ couples kissing. It turns out that 80 couples turned to the right. Therefore, we can estimate p with the estimator

$$\hat{p} = \frac{80}{124} = 64.5\%.$$

It seems intuitively true that there is a preference for couples to turn to the right, since $65.5\% > 50\%$. On the other hand, if we observed $n = 3$ couples, and found that 2 of them turned to the right, we would be less convinced that this is necessarily the case. At what threshold are we actually convinced that $p > 50\%$?

Define a sequence of random variables $\{R_i\}_{1 \le i \le n}$, where $R_i = 1$ if the ith couple turns to the right, and $R_i = 0$ otherwise. For the sake of our model, we assume:

- $R_i \sim \textsc{Bern}(p)$. Modelling each R_i as a r.v. is how we deal with the lack of other information. If we knew more, we could use psychology or physics to deduce whether there is a natural tendency to lean right while kissing (in other words, we would not need to use statistics).
- $R_1, \ldots, R_n$ are mutually independent. This is reasonable since the behavior of one couple does not interfere with the behavior of another.

Now, by the strong law of large numbers,

$$\hat{p} = \overline{R_n} \xrightarrow[n\to\infty]{a.s.} p.$$

How do we quantify how confident we are with our estimate when n is not infinitely large? By the CLT,

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{R_n} - p}{\sqrt{p(1-p)}} \le x\right) \xrightarrow[n\to\infty]{} \mathbb{P}(\mathcal{N}(0,1) \le x),$$

for all **quantiles** x. In other words, for large n, we may say

$$\sqrt{n}(\overline{R_n} - p) \approx \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, p(1-p)),$$

which implies

$$\mathbb{P}[|\overline{R_n} - p| \geq a/\sqrt{n}] \approx \mathbb{P}[|\mathcal{N}(0,1)| \geq a/\sigma] = 2 - 2\Phi\left(\frac{a}{\sigma}\right).$$

Let $q_{\alpha/2}$ be the $(1-\alpha/2)$ quantile of $\mathcal{N}(0,1)$, i.e.,

$$1 - \alpha/2 = \Phi\left(q_{\alpha/2}\right).$$

Then,

$$\mathbb{P}[|\mathcal{N}(0,1)| \leq q_{\alpha/2}] = 1 - \alpha.$$

Per the equation earlier, we thus have that, with probability $1-\alpha$,

$$|\overline{R_n} - p| \leq \frac{q_{\alpha/2}\sigma}{\sqrt{n}} \leq \frac{q_{\alpha/2}}{2\sqrt{n}},$$

following from the fact that $\sigma = \sqrt{p(1-p)} \leq 1/2$.

Definition 2.3

The interval given by

$$\left[\overline{R_n} - \frac{|q_{\alpha/2}|}{2\sqrt{n}}, \overline{R_n} + \frac{|q_{\alpha/2}|}{2\sqrt{n}}\right]$$

is called the $1-\alpha$ **Confidence Interval** (C.I.) for p.

Intuition checks:

- This naming makes sense, because the probability that p lies in this interval is $1-\alpha$ (by rearranging the equation we had earlier).
- When $\alpha = 1$, we are 0% confident that p lies in the interval. Indeed, $q_{1/2} = 0$, so the interval has length 0.
- When $\alpha = 0$, we are 100% confident that p lies in this interval. Indeed, $|q_0| = \infty$, so the interval spans all possible values for p.

3 February 14, 2023

3.1 Statistical Models and Identifiability

Definition 3.1

Let Ω be a sample space. A **statistical model** is given by

$$(\Omega, (\mathbb{P}_\theta)_{\theta \in \Theta}),$$

where $\mathbb{P}_\theta$ is a probability distribution on Θ for each $\theta \in \Theta$.

The goal of the statistical model is to estimate the paramter θ. Θ is called the parameter set; as an example, if we are trying to estimate a proportion, we say that the parameter set is $[0,1]$. If θ exists, we say that the model is **well-specified**, and this particular value of θ is called the **true parameter**.

Example 3.2

Consider the kissing experiment from last time.

In this example,

- $\Omega = \{0,1\}$.
- $R_i \sim \textsc{Bern}(p)$, i.e., $\mathbb{P}[R_i = 1] = p$, $\mathbb{P}[R_i = 0] = 1 - p$.
- The statistical model given by each R_i is $\left(\{0,1\}, \textsc{Bern}(p)_{p \in [0,1]}\right)$.
- The statistical model given by a pair (R_1, R_2) is $\left(\Omega^{(1)}, \textsc{Bern}(p) \otimes \textsc{Bern}(p)\right)$, where $\Omega^{(1)} = \{(0,0),(0,1),(1,0),(1,1)\}$.

Definition 3.3

- **Parametric** model: We assume $\Theta \subseteq \mathbb{R}^d$ for finite d
- **Nonparametric** model: Θ can be infinite dimensional
- **Semiparametric** model: $\Theta = \Theta_1 \times \Theta_2$, with one finite-dimensional and the other infinite-dimensional. We won't cover these models in this class.

Example 3.4

Common models for different distributions.

Gaussian Model:

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

Exponential model:

$$\left((0,\infty), (\textsc{Exp}(\lambda))_{\lambda\in(0,\infty)}\right).$$

Binomial model:

$$\left(\{0,1\}, (\textsc{Bern}(p))_{p\in[0,1]}\right).$$

Poisson model:

$$\left(\mathbb{N}, (\textsc{Pois}(\lambda))_{\lambda\in(0,\infty)}\right).$$

Definition 3.5

The parameter θ is **identifiable** if and only if $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective.

In other words, we can identify θ if and only if each θ maps to a unique distribution. As an example where this is not satisfied, suppose $X \sim \mathcal{N}(\mu,\sigma^2)$, and we observe the indicator $Y = 1_{X\geq 0}$. Since

$$\mathbb{P}[Y=1] = \Phi\left(-\frac{\mu}{\sigma}\right),$$

μ and σ^2 are not identifiable, since there are many different combinations that produce the same observed distribution. On the other hand, $\theta = \mu/\sigma$ is identifiable.

3.2 Estimation

Given a statistical model $(\Omega, (\mathbb{P}_\theta)_{\theta\in\Theta})$, and some sequence of i.i.d. $X_1,\ldots,X_n \sim \mathbb{P}_\theta$, we generate some prediction $\hat{\theta}_n$ for θ. We call $\hat{\theta}_n$ an **estimator** for θ.

Definition 3.6

An estimator $\hat{\theta}_n$ of θ is called **consistent** if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{p} \theta.$$

This estimator is called **strongly consistent** if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{a.s.} \theta.$$

This estimator is **asymptotically normal** if

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

Definition 3.7

The **bias** of an estimator $\hat{\theta}_n$ of θ is given by

$$\mathrm{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta.$$

Definition 3.8

The **quadratic risk** of an estimator $\hat{\theta}_n \in \mathbb{R}$ is given by

$$R(\hat{\theta}_n) = \mathbb{E}[|\hat{\theta}_n - \theta|^2] = \mathrm{Var}[\hat{\theta}_n] + \mathrm{bias}^2(\hat{\theta}_n).$$

Recall that θ is a constant, i.e., $\mathbb{E}[\theta] = \theta$, so the secondary definition follows by expansion:

$$\begin{aligned}
\mathbb{E}[|\hat{\theta}_n - \theta|^2] &= \mathbb{E}[\hat{\theta}_n^2 - 2\theta\hat{\theta}_n + \theta^2] \\
&= (\mathbb{E}[\hat{\theta}_n^2] - \mathbb{E}[\hat{\theta}_n]^2) + (\mathbb{E}[\hat{\theta}_n]^2 - 2\theta\mathbb{E}[\hat{\theta}_n] + \theta^2) \\
&= \mathrm{Var}[\hat{\theta}_n] + \mathrm{bias}^2(\hat{\theta}_n).
\end{aligned}$$

3.3 More on Confidence Intervals

Let $(\Omega, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model with observations $X_1, \ldots, X_n$. Say $\Theta \subseteq \mathbb{R}$ and let $\alpha \in (0, 1)$.

- A "confidence interval of level $1 - \alpha$ for θ" is defined as a random interval (based on our observations) $\mathcal{I}$, **which does not depend on** θ, such that

$$\mathbb{P}_\theta[\theta \in \mathcal{I}] \geq 1 - \alpha.$$

- A "confidence interval of asymptotic level $1-\alpha$ for θ" is defined as $\lim_{n\to\infty} \mathcal{I}_n$, where
$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in \mathcal{I}_n) \geq 1 - \alpha.$$

There are a few reasons why we define confidence intervals as inequalities, rather than equalities. The reality is that we can rarely get exact estimates (even in the asymptotic limit) without knowing more information about the true parameter. Consider the following example.

Example 3.9
Recall the kissing experiment from last lecture.

By the central limit theorem, we deduced

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

Then, we said that

$$\lim_{n\to\infty} \mathbb{P}\left(p \in \left[\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

This interval by itself is not a confidence interval, since it depends on p. There are a few ways to resolve this issue.

- **The Conservative Method**: this is the method that we used implicitly last lecture. We know that $p(1-p) \leq 1/4$, so we can remove the dependence on p be relaxing the bounds of our interval. This gives us:
$$\mathcal{I}_{\text{conservative}} = \left[\overline{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}}\right].$$
Indeed,
$$\lim_{n\to\infty} \mathbb{P}(p \in \mathcal{I}_{\text{conservative}}) \geq 1 - \alpha,$$

so this is a valid confidence interval (which is asymptotic). Note that our probability is no longer exact, since we had to relax our interval to account for all possible values of p.

- **Solve for p**: we could also manually solve for p, since our interval is of the form
$$\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \le p \le \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}.$$
Rearranging both inequalities, we seek the roots
$$\left(1+\frac{q_{\alpha/2}^2}{n}\right)p^2 - \left(2\overline{R}_n + \frac{q_{\alpha/2}^2}{n}\right)p + \overline{R}_n^2 = 0.$$
Our desired interval is now:
$$\mathcal{I}_{\text{solve}} = \left[\frac{1}{1+\frac{q_{\alpha/2}^2}{n}}\left(\overline{R}_n + \frac{q_{\alpha/2}^2}{2n}\right) \pm \frac{q_{\alpha/2}}{1+\frac{q_{\alpha/2}^2}{n}}\sqrt{\frac{(\overline{R}_n(1-\overline{R}_n))}{n} + \frac{q_{\alpha/2}^2}{4n^2}}\right].$$

- **Slutsky**: $\hat{p} \xrightarrow[n\to\infty]{a.s.} p$, we can use Slutsky to substitute p for $\hat{p}$, giving us
$$\mathcal{I} = \left[\hat{p} - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right].$$

4 February 15, 2023 (R)

Example 4.1

Let $X_1, \dots, X_n \sim \textsc{Unif}([\theta, 2\theta])$ be a sequence of i.i.d. random variables for some $\theta > 1$. Compute the quadratic risk of $\hat{\theta} = \min\{X_1, \dots, X_n\}$.

This is order statistics. The minimum order is described by the **beta distribution**, which we can use to check our work. (If we know the expected value and variance of the minimum order statistic over $U \sim \textsc{Unif}[0,1]$, we could just plug it in, but we'll go through the whole derivation here anyways for fun). Recall the formula for quadratic risk:

$$R(\hat{\theta}_n) = \mathrm{Var}[\hat{\theta}_n] + \mathrm{bias}^2(\hat{\theta}_n).$$

First, we compute $f_{\hat{\theta}}$. Note

$$\mathbb{P}[\hat{\theta} \geq x] = \mathbb{P}[X_i \geq x]^n = \left(\frac{2\theta - x}{\theta}\right)^n,$$

so

$$f_{\hat{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\theta}\left[1 - \left(\frac{2\theta - x}{\theta}\right)^n\right] = \frac{n}{\theta}\left(\frac{2\theta - x}{\theta}\right)^{n-1}.$$

Now,

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \int_\theta^{2\theta} \frac{n}{\theta}\left(\frac{2\theta - x}{\theta}\right)^{n-1} x\mathrm{d}x \\
&= \left[-x\left(\frac{2\theta - x}{\theta}\right)^n\right]\Bigg|_\theta^{2\theta} + \int_\theta^{2\theta} \left(\frac{2\theta - x}{\theta}\right)^n \mathrm{d}x \\
&= \left[-x\left(\frac{2\theta - x}{\theta}\right)^n + \frac{-\theta}{n+1}\left(\frac{2\theta - x}{\theta}\right)^{n+1}\right]\Bigg|_\theta^{2\theta} = \left(\frac{n+2}{n+1}\right)\theta.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}[\hat{\theta}^2] &= \int_\theta^{2\theta} \frac{n}{\theta}\left(\frac{2\theta - x}{\theta}\right)^{n-1} x^2\mathrm{d}x \\
&= \left[-x^2\left(\frac{2\theta - x}{\theta}\right)^n - 2x\frac{\theta}{n+1}\left(\frac{2\theta - x}{\theta}\right)^{n+1}\right]\Bigg|_\theta^{2\theta} + \int_\theta^{2\theta} \frac{2\theta}{n+1}\left(\frac{2\theta - x}{\theta}\right)^{n+1} \mathrm{d}x \\
&= \left[-x^2\left(\frac{2\theta - x}{\theta}\right)^n - 2x\frac{\theta}{n+1}\left(\frac{2\theta - x}{\theta}\right)^{n+1} + \frac{-2\theta^2}{(n+1)(n+2)}\left(\frac{2\theta - x}{\theta}\right)^{n+2}\right]\Bigg|_\theta^{2\theta} \\
&= \left(\frac{n^2 + 5n + 8}{(n+1)(n+2)}\right)\theta^2.
\end{aligned}$$

Finally,

$$\mathrm{Var}[\hat{\theta}] = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2 = \frac{\theta^2 n}{(n+1)^2(n+2)}.$$

So,

$$R(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] + \mathrm{bias}^2(\hat{\theta}) = \frac{\theta^2 n}{(n+1)^2(n+2)} + \left(\frac{\theta}{n+1}\right)^2 = \frac{2\theta^2}{(n+1)(n+2)}.$$

Example 4.2

Consider a sample of n i.i.d. continuous random variables $X_1, \dots, X_n$ with density

$$f(x) = e^{-(x-a)} 1_{x \geq a}, x \in \mathbb{R},$$

where a is an unknown parameter.

(1) Compute $\mathbb{E}[X_1]$.

(2) Determine whether $\overline{X}_n - 1$ is a consistent estimator of a.

(3) Based on $\overline{X_n}$, propose a confidence interval for a with asymptotic level 95%.

(1) The given density function is the same density function as an exponentially distributed r.v. with parameter 1, shifted by a. Therefore, $\mathbb{E}[X_1] = \mathbb{E}[\textsc{Exp}(1)] + a = a + 1$.

(2) By the strong law of large numbers, $\overline{X}_n \xrightarrow[n\to\infty]{a.s.} a + 1$. Let $g(x) = x - 1$. By the continuous mapping theorem,

$$g(\overline{X}_n) = \overline{X}_n - 1 \xrightarrow[n\to\infty]{a.s.} a = g(\mathbb{E}[\overline{X}_n]),$$

so $\overline{X}_n - 1$ is a strongly consistent estimator of a.

(3) By the central limit theorem,

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\sqrt{n}\frac{(\overline{X}_n - a - 1)}{1}\right| \leq q_{0.025}\right] = \mathbb{P}[|\mathcal{N}(0,1)| \leq q_{0.025}] = 95\%.$$

So, a valid asymptotic 95% confidence interval for a is given by

$$a \in \left[\overline{X}_n - 1 - \frac{q_{0.025}}{\sqrt{n}}, \overline{X} - 1 + \frac{q_{0.025}}{\sqrt{n}}\right].$$

5 February 16, 2023

5.1 Maximum Likelihood Estimation

We are given i.i.d. $x_1, \dots, x_n$ data from a statistical model $(\Omega, (\mathbb{P}_\theta)_{\theta \in \Theta})$. Suppose Ω is a discrete probability space. Our goal is to find an estimator $\hat{\theta}$ that maximizes our **likelihood function**

$$L(x_1, \dots, x_n; \theta) = \prod_i \mathbb{P}_\theta(X_i = x_i).$$

This is the same as maximizing the **Log-Likelihood**:

$$\log L(x_1, \dots, x_n; \theta) = \sum_{i=1}^{n} \log \mathbb{P}_\theta(X_i = x_i).$$

Our **maximum likelihood estimator** is given by

$$\hat{\theta} = \arg\max_\theta (\log L(x_1, \dots, x_n; \theta)).$$

In the case where Ω is continuous, all definitions are the same, except we replace $\mathbb{P}_\theta$ with f_θ, the density of $\mathbb{P}_\theta$.

Example 5.1

Compute the maximum likelihood estimator (MLE) given n data points from the statistical model $(\{0,1\}, (\textsc{Bern}(p))_{p \in (0,1)})$.

Note that $\mathbb{P}_\theta(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$. Therefore,

$$\log L(\theta) = n(\overline{X}_n \log \theta + (1 - \overline{X}_n)\log(1-\theta)),$$

and

$$\frac{\partial \log L}{\partial \theta} = n\left(\frac{\overline{X}_n}{\theta} - \frac{1 - \overline{X}_n}{1 - \theta}\right).$$

Setting the derivative to zero, we get $\hat{\theta} = \hat{p} = \overline{X}_n$ as our MLE. This agrees with our intuition.

Example 5.2

Consider the statistical model $(\mathbb{R}, (\mathcal{N}(\mu, \sigma^2))_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)})$.

From the density of the normal distribution, we have

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right),$$

and

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

Maximizing:

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \overline{X}_n)^2.$$

So, our best variance estimator is given by

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{X}_i)^2.$$

5.2 Relative Entropy

The motivation behind the way that we calculate our maximum likelihood estimator is to get our estimated distribution as close as possible to the actual distribution. If we let θ^* be the true parameter, then

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \left(\frac{1}{n}\sum_{i=1}^n \log f_\theta(X_i = x_i)\right) \\ &= \arg\min_\theta \left(\frac{1}{n}\sum_{i=1}^n \log f_{\theta^*}(X_i = x_i) - \frac{1}{n}\sum_{i=1}^n \log f_\theta(X_i = x_i)\right).\end{aligned}$$

By the strong law of large numbers, the quantity inside of the argmin converges almost surely to

$$\mathbb{E}_{\theta^*}\left[\log \frac{f_{\theta^*}(X)}{f_\theta(X)}\right].$$

Definition 5.3

The **relative entropy** between the actual distribution $\mathbb{P}_{\theta^*}$ and some other $\mathbb{P}_\theta$ is given by

$$\mathbb{E}_{\theta^*}\left[\log \frac{f_{\theta^*}(X)}{f_\theta(X)}\right].$$

This quantity is also called **Kullback-Leibler (KL) divergence**. This shows that computing the MLE is the same as minimizing the relative entropy between the actual distribution, and our predicted distribution.

5.3 Fisher Information, Cramer Rao

Let $\ell(x,\theta) = \log L(x,\theta)$, and suppose $\Theta \subseteq \mathbb{R}$.

Definition 5.4

The **Fisher Information** $I(\theta)$ is given by

$$I(\theta) = \operatorname{Var}_\theta\left(\frac{\partial}{\partial\theta}\ell(x,\theta)\right) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial^2\theta}\ell(x,\theta)\right).$$

Note that we are taking the expectation and variance of the inner functions with respect to θ (i.e., averaging over all possible x). It's not immediately clear that these two definitions (expectation and variance) are the same, so we show the proof below:

Proof. First, we have

$$\frac{\partial}{\partial\theta}\ell(x,\theta) = \frac{\partial}{\partial\theta}\log f_\theta(x) = \frac{\partial/(\partial\theta)(f_\theta(x))}{f_\theta(x)},$$

and

$$\frac{\partial^2}{\partial^2\theta}\ell(x,\theta) = \frac{\left[\partial^2/(\partial^2\theta)(f_\theta(x))\right]f_\theta(x) - [\partial/(\partial\theta)(f_\theta(x))]^2}{f_\theta(x)^2}.$$

Now,

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right] &= \int\left(\frac{\partial/(\partial\theta)(f_\theta(x))}{f_\theta(x)}\right)\cdot f_\theta(x)\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int f_\theta(x)\mathrm{d}x = 0,\end{aligned}$$

since $\int f_\theta(x)\mathrm{d}x = 1$ always. Thus,

$$\begin{aligned}\mathrm{Var}\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell(x,\theta)\right)^2\right] - \mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right]^2 \\ &= \int\left(\frac{\partial/(\partial\theta)(f_\theta(x))}{f_\theta(x)}\right)^2 f_\theta(x)\mathrm{d}x \\ &= \int\frac{(\partial/(\partial\theta)(f_\theta(x)))^2}{f_\theta(x)}\mathrm{d}x.\end{aligned}$$

On the other hand,

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ell(x,\theta)\right] &= -\int\left(\frac{\left[\partial^2/(\partial^2\theta)(f_\theta(x))\right]f_\theta(x) - [\partial/(\partial\theta)(f_\theta(x))]^2}{f_\theta(x)^2}\right)f_\theta(x)\mathrm{d}x \\ &= \int\frac{(\partial/(\partial\theta)(f_\theta(x)))^2}{f_\theta(x)}\mathrm{d}x - \frac{\partial}{\partial\theta}\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right] \\ &= \mathrm{Var}\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right],\end{aligned}$$

as desired. □

Recall that the bias of an estimator $\hat{\theta}$ of θ is given by

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

An unbiased estimator is any estimator with no bias.

Definition 5.5

The **Cramer-Rao** lower bound gives a lower bound to the variance of any unbiased estimator. In particular, given unbiased estimator $\hat{\theta}$ for θ,

$$\operatorname{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

Theorem 5.6 (The MLE is consistent and normal)

Let $\theta^* \in \Theta$ be the true parameter, with some technical conditions holding (e.g., the support of f_θ cannot depend on θ). Then,

- $\hat{\theta}_n^{MLE}$ is a consistent estimator, i.e.,

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{p} \theta^*.$$

- It is asymptotically normal:

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

In particular, by Cramer-Rao, this implies that the MLE gives us the best possible variance. This theorem demonstrates that, in theory, the MLE gives us everything we want in an estimator. In reality, it is often difficult to compute the MLE, so we have to resort to more practical estimators that aren't as perfect.

Example 5.7

As usual, let's return to the kissing experiment.

In this case, our statistical model was

$$(\{0,1\}, (\textsc{Bern}(p))_{p\in(0,1)}).$$

We showed here that the MLE for this model is $\hat{\theta}_n^{MLE} = \overline{X}_n$. Moreover,

$$\ell(x,\theta) = x\ln\theta + (1-x)\ln(1-p),$$

and

$$\frac{\partial}{\partial\theta}\ell(x,\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta},$$

and

$$\frac{\partial^2}{\partial^2\theta}\ell(x,\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2},$$

so

$$I(\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

Using our theorem, we can put everything together:

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

which is what we wanted.

6 February 23, 2023

6.1 M-estimation

$X_1, \ldots, X_n$ i.i.d. from $\mathbb{P}_\theta$ drawing from sample space E. In maximum likelihood estimation, we estimate θ with

$$\hat{\theta} = \arg\max_\theta \frac{1}{n}\sum_i \log L(X_i, \theta).$$

In M-estimation, our goal is to find a function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible θ, such that

$$\hat{\theta} = \arg\min_\theta \mathbb{E}[\rho(X_1, \theta)].$$

Note that $\hat{\theta}_{MLE}$ is itself also an M-estimator, by setting $\rho(X,\theta) = -\log L(X,\theta)$.

Example 6.1

Given $X_1, \ldots, X_n$ i.i.d. from some unknown $\mathbb{P}$ in sample space $E \subseteq \mathbb{R}^d$. Estimate $\mathbb{E}[X]$, where X is also distributed as $\mathbb{P}$.

The sample mean $\overline{X}_n = (X_1 + \ldots + X_n)/n$ is a valid M-estimator.

Claim 6.2

If $\rho(X_i, \theta) = (X_i - \theta)^2$, then

$$\arg\min_{\theta} \frac{1}{n} \sum (X_i - \theta)^2 = \overline{X}_n.$$

Example 6.3

Estimate the median of the distribution of X.

The sample median is also a valid M-estimator.

Claim 6.4

If $\rho(X_i, \theta) = |X_i - \theta|$, then

$$\arg\min_{\theta} \frac{1}{n} \sum |X_i - \theta| = \text{sample median.}$$

6.2 Hypothesis Testing

Let $(\Omega, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model. For some partition $\{\Theta_0, \Theta_1\}$ of Θ:

- H_0: $\theta \in \Theta_0$ is the **null hypothesis**
- H_1: $\theta \in \Theta_1$ is the **alternative hypothesis**

For $k \in \{0, 1\}$, we say that Θ_k is a **simple hypothesis** if $\Theta_k = \{\theta_k\}$. It is a **composite hypothesis** if it takes the form $\Theta_k = \{\theta : \theta > \theta_k\}$, $\Theta_k = \{\theta : \theta < \theta_k\}$, or $\Theta_k = \{\theta : \theta \neq \theta_k\}$.

Definition 6.5

A **test** is a function $\Psi : \Theta \to \{0, 1\}$, where $\Psi = 1$ indicates that we reject the null hypothesis, and $\Psi = 0$ means that we fail to reject the null hypothesis. We can define a rejection region R. Then $\Psi = \mathbb{1}(R)$.

Example 6.6

The average waiting time in the emergency room is around 30 minutes. Some people claim that the New-York Presbyterian hospital has a longer waiting time.

Let ω be a random variable modeling the waiting time in the emergency room at the New-York Presbyterian hospital, in minutes. The null hypothesis (H_0) says that $\mathbb{E}[\omega] \leq 30$. The alternate hypothesis (H_1) says that $\mathbb{E}[\omega] > 30$. Both hypotheses in this case are composite hypotheses.

Let $\widehat{\mathbb{E}[\omega]}$ be some estimator that we come up with for $\mathbb{E}[\omega]$ given some data. A typical test might look like

$$\Psi = \mathbb{1}(\widehat{\mathbb{E}[\omega]} > \varepsilon),$$

where ε is a threshold we set for deciding whether or not to reject H_0.

6.3 Errors

	Fail to Reject	Reject
H_0 true ($\theta \in \Theta_0$)	Correct	Type I error
H_1 true ($\theta \in \Theta_1$)	Type II error	Correct

Definition 6.7

The **power function** of a test is a function $\beta : \Theta \to [0,1]$ with

$$\beta(\theta) = \mathbb{P}_\theta(\psi = 1).$$

In words, the power of a test measures how likely it is to reject the null hypothesis. This function can be used to measure the likelihood of each error: if $\theta \in \Theta_0$,

$$\beta(\theta) = \mathbb{P}_\theta[\text{type I error}].$$

In this case, we want the power to be small. If $\theta \in \Theta_1$,

$$\beta(\theta) = 1 - \mathbb{P}_\theta[\text{type II error}].$$

In this case, we want the power to be large.

Example 6.8

Back to the kissing example.

In this example, our test is $\Psi = \mathbb{1}(\hat{p} > c)$, for some value of c. So,

$$\mathbb{P}_\theta(\text{type I error}) = \mathbb{P}(\hat{p} > c) = \mathbb{P}\left(\sqrt{n}\left(\hat{p} - \frac{1}{2}\right) > \sqrt{n}\left(c - \frac{1}{2}\right)\right).$$

(For a type I error to occur, we assume that $p = 1/2$). If we let $c \to \infty$, the probability of a type I error goes to 0. On the other hand,

$$\mathbb{P}_\theta(\text{type II error}) = \mathbb{P}(\hat{p} < c) = \mathbb{P}\left(\sqrt{n}(\hat{p} - p) < \sqrt{n}(c - p)\right).$$

If we want the probability of a type II error to be 0, we need $c \to -\infty$. We cannot do both at the same time.

Definition 6.9

In the **Neyman-Pearson paradigm**,

- First, make sure that $\mathbb{P}_\theta[\text{type I error}] \le \alpha$, where α determines the level of the test (i.e., 5%, 1%, etc.)
- Then, choose c such that $\mathbb{P}_\theta[\text{type II error}]$ is minimized

In the kissing experiment, we therefore want

$$\mathbb{P}_{\theta=1/2}(\hat{p} > c) \le \alpha.$$

In this case, the test $\Psi = \mathbb{1}(\hat{p} > c)$ said to have **level α**. More generally, if we require

$$\lim_{n\to\infty} \mathbb{P}_{\theta=1/2}(\hat{p}_n > c) \le \alpha,$$

then Ψ is an **asymptotic level α** test.

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}_{\theta=1/2}\left(\frac{\sqrt{n}(\hat{p}_n - 1/2)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} > \frac{\sqrt{n}(c - 1/2)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}}\right) &= \mathbb{P}\left(\mathcal{N}(0,1) \ge \frac{\sqrt{n}(c-1/2)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c-1/2)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}}\right) = \alpha.\end{aligned}$$

This gives

$$c = \frac{1}{2} + \frac{q_\alpha}{\sqrt{n}}.$$

7 February 28, 2023

7.1 p-values

From last lecture: for a test $\Psi = \mathbb{1}(\hat{\theta} > c)$, the level of Ψ determines c.

Definition 7.1

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

In general, when we conduct an α-level test, we reject the null hypothesis if we find that the p-value is $\leq \alpha$.

Example 7.2

Back to kissing experiment.

The level $(1-2\beta)$ confidence interval for the kissing interval was given by

$$\mathcal{I} = \left[\overline{R}_n - \frac{q_\beta}{2\sqrt{n}}, \overline{R}_n + \frac{q_\beta}{2\sqrt{n}}\right]$$

The null hypothesis was that $p = 1/2$, and the alternate hypothesis was that $p > 1/2$.

To compute the p-value, we want to compute the first point at which $1/2$ fails to lie in this interval. Since we're looking for $p > 1/2$, we equate the left part of the interval to $1/2$:

$$\frac{1}{2} = \overline{R}_n - \frac{q_\beta}{2\sqrt{n}}.$$

This gives

$$q_\beta = 2\sqrt{n}\left(\overline{R}_n - \frac{1}{2}\right),$$

from which we can solve for β (recall that $\mathbb{P}[\mathcal{N}(0,1) > q_\beta] = \beta$).

7.2 Parametric Hypothesis Testing

Start with a statistical model $(\Omega, (\mathbb{P}_\theta)_{\theta\in\Theta})$.

Example 7.3

Wald's test

The test can come in any of the following forms:

- form 1: $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$
- form 2: $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$
- form 3: $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$

The test statistic is

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}},$$

where $\widehat{\mathrm{Var}}(\hat{\theta})$ is an *estimator* of the variance of $\hat{\theta}$. For example, in the kissing example, we used $\sqrt{(1/2) \cdot (1 - 1/2)/n}$.

Assuming that the test is level-α, our tests are:

- form 1:

  $$\Psi = \mathbb{1}(W > q_\alpha).$$

  When W is normally distributed, the probability of a type 1 error is α, at the right tail of the distribution. For some W_{obs}, our p-value is

  $$\mathbb{P}[\mathcal{N}(0,1) > W_{obs}].$$

- form 2:

  $$\Psi = \mathbb{1}(W < -q_\alpha).$$

  When W is normally distributed, the probability of a type 1 error is α, at the left tail of the distribution. For some W_{obs}, our p-value is

  $$\mathbb{P}[\mathcal{N}(0,1) < W_{obs}].$$

- form 3:

  $$\Psi = \mathbb{1}(|W| > q_{\alpha/2}).$$

  When W is normally distributed, the probability of a type 1 error is $\alpha/2 +$

$\alpha/2 = \alpha$. For some W_{obs}, our p-value is

$$\mathbb{P}[|\mathcal{N}(0,1)| > |W_{obs}|].$$

8 March 7, 2023

8.1 t-test

In a t-test, we have $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$. We assume that everything is independent. Then,

- H_0: $\mu_1 = \mu_2$
- H_1: Any of $\mu_1 \neq \mu_2$, $\mu_1 > \mu_2$, $\mu_2 > \mu_1$

We use estimators $\hat{\mu_1} = \sum_{i=1}^n X_i/n$, and $\hat{\mu_2} = \sum_{i=1}^m Y_i/m$. Then, consider

$$\frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\mathrm{Var}(\hat{\mu_1} - \hat{\mu_2})}},$$

which, like the Wald test, is the quantity that we would like to use as our test statistic. By our independence assumption, $\mathrm{Var}(\hat{\mu_1} - \hat{\mu_2}) = \mathrm{Var}(\hat{\mu_1}) + \mathrm{Var}(\hat{\mu_1}) = \sigma^2/n + \sigma^2/m$. Since we do not explicitly know σ, we need an estimator:

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

This is called the sample variance.

Definition 8.1 (Chi-square distribution)
Let $Z_1 \sim \mathcal{N}(0,1)$. Then, we say that $Z_1^2 \sim \chi_1^2$. More generally, we say that

$$Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2.$$

Theorem 8.2 (Cochran's Theorem)

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Proof. (intuition)

$$\begin{aligned}\frac{(n-1)S_n^2}{\sigma^2} &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}_n}{\sigma}\right)^2\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu-\overline{X}_n+\mu)^2\\ &= \frac{1}{\sigma^2}\sum(X_i-\mu)^2-\frac{n}{\sigma^2}(\overline{X}_n-\mu)^2.\end{aligned}$$

The term on the left is the sum of n squared normal distributions, which is χ_k^2. The term on the right subtracts just enough to turn the final distribution into χ_{n-1}^2. □

Definition 8.3 (t-distribution)
For $Z \sim \mathcal{N}(0,1)$ and $S^2 \sim \chi_k^2$,

$$\Upsilon = \frac{Z}{\sqrt{S^2/k}} \sim t_k.$$

Our test statistic now looks something like this:

$$\frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{\frac{1}{n}\frac{1}{n-1}\sum(X_i-\overline{X}_n)^2+\frac{1}{m}\frac{1}{m-1}\sum(Y_i-\overline{Y}_m)^2}}.$$

Consider the case when $n = m$. Then, this distribution simplifies:

$$\begin{aligned}\frac{(\hat{\mu}_1-\hat{\mu}_2)/\sigma}{\sqrt{\frac{1}{n(n-1)}}\sqrt{\sum\left(\frac{X_i-\overline{X}_n}{\sigma}\right)^2+\sum\left(\frac{Y_i-\overline{Y}_n}{\sigma}\right)^2}} &= \frac{(\hat{\mu}_1-\hat{\mu}_2)/\sigma}{\sqrt{\frac{1}{n(n-1)}}\sqrt{\chi_{n-1}^2+\chi_{n-1}^2}}\\ &= \frac{(\hat{\mu}_1-\hat{\mu}_2)/(\sigma\sqrt{2/n})}{\sqrt{\chi_{2n-2}^2/(2n-2)}} \sim t_{2n-2}.\end{aligned}$$

Example 8.4
Conduct this test for one sample.

For one sample,

$$\sqrt{n}\frac{\overline{X}_n-\mu}{S_n} \sim t_{n-1},$$

so we can use this as our test statistic. Say we want to test against the null hypothesis $H_0 : \mu = \mu_0$. Then, we use the test statistic

$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n}.$$

- $H_1 : \mu \neq \mu_0$, the T-Test Ψ with level α is given by

$$\Psi = \mathbb{1}(|T| > q_{\alpha/2}^{t_{n-1}}).$$

- The other alternative hypothesis tests have the same form as the Wald test (taken over t_{n-1}).

8.2 Goodness-of-fit Test

Define $(\mathbb{P}_p)_{p\in\Delta_K}$ to be the family of all probability distributions on some sample space $E = \{a_i\}_{i\leq K}$. In other words,

$$\Delta_k = \left\{(p_1,\ldots,p_K) \in (0,1)^K : \sum_{j=1}^{K} p_j = 1\right\},$$

where for any $p \in \Delta_K$ and $X \sim \mathbb{P}_p$,

$$\mathbb{P}_p[X = a_i] = p_i.$$

The setup for the goodness-of-fit test is now as follows. We are given $X_1,\ldots,X_n \sim \mathbb{P}_\theta$ for some $\theta \in \Delta_K$. The goal is to test whether θ is equal to some null hypothesis $\theta^0 \in \Delta_K$, i.e.,

- $H_0 : \theta = \theta^0$.
- $H_1 : \theta \neq \theta^0$.

The pmf has support on E (i.e., the original sample space), and is given by

$$p(x) = \prod_{j=1}^{K} p_j^{\mathbb{1}(x=a_j)}.$$

Let $\theta = (p_1, \dots, p_K)$. The likelihood of the model is given by

$$L_n(X_1, \dots, X_n; \theta) = p_1^{N_1} \dots p_K^{N_K},$$

where N_i is the number of samples equal to a_i. By Jensen's,

$$\sum \frac{N_i}{n} \log p_i \le \log\left(\sum \frac{N_i}{n} p_i\right).$$

Note that the numerator in the right hand expression is $\vec{N} \cdot \vec{p}$, so it is maximized when they are parallel, i.e., when N_i/p_i is constant. This shows that the MLE for θ is

$$\hat{\theta} = \left(\frac{N_1}{n}, \frac{N_2}{n}, \dots, \frac{N_K}{n}\right).$$

(This should intuitively make sense).

Theorem 8.5 (Goodness-of-fit convergence to chi-square)

$$n \sum_{j=1}^{K} \frac{(\hat{\theta}_j - \theta_j^0)^2}{\theta_j^0} \xrightarrow[n \to \infty]{d} \chi^2_{K-1}.$$

This gives us the χ^2 test with level α: $\Psi = \mathbb{1}(T_n > q_\alpha^{\chi^2_{K-1}})$, where T_n is the test statistic (the quantity on the left-hand-side), and $q_\alpha^{\chi^2_{K-1}}$ is the α-quantile of χ^2_{K-1}. The p-value is given by $\mathbb{P}[\chi^2_{K-1} > T_n^{obs}]$.

9 March 9, 2023

9.1 Empirical CDF

Definition 9.1

The **empirical cdf** of a sample $X_1, \dots, X_n$ is defined as

$$F_n(t) = \frac{1}{t} \sum_{i=1}^{n} \mathbb{1}(X_i \le t) = \frac{\#\{i = 1, \dots, n : X_i \le t\}}{n}.$$

Let $Z_i = \mathbb{1}(X_i \le x)$. $\mathbb{P}[Z_i = 1] = \mathbb{P}[X_i \le x] = F(x)$, so $Z_i \sim \textsc{Bern}(F(x))$. This implies that $F_n(t)$ is an unbiased estimator for $F(t)$:

$$\begin{aligned}\mathbb{E}[F_n(x)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}(x_i \le x)] \\ &= \frac{1}{n}(nF(x)) = F(x),\end{aligned}$$

By the strong law of large numbers, $F_n(t)$ is also a consistent estimator:

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Also, the central limit theorem holds:

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, F(x)(1-F(x))).$$

Theorem 9.2 (Glivenko-Cantelli's Theorem)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

This theorem is also known as the **Fundamental Theorem of Statistics**. The next theorem is a generalization of the central limit theorem:

Theorem 9.3 (Donsker's Theorem)
If F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{d} \sup_{0\le t\le 1} |\mathbb{B}(t)|,$$

where $\mathbb{B}$ is a brownian bridge on $[0,1]$.

A brownian bridge on $[0,1]$ is a function modelling Brownian motion which starts and ends at the same point (0).

9.2 Kolmogorov-Smirnov Test

Let $X_1, \dots, X_n$ be i.i.d. real random variables with unknown cdf F and let F^0 be a continuous cdf. Let

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Note that the null hypothesis means that $F(x) = F_0(x) \forall x \in \mathbb{R}$. Let

$$T_n = \sup_x |F_n(x) - F_0(x)|.$$

By Donsker's Theorem, given H_0,

$$\sqrt{n} T_n \xrightarrow[n\to\infty]{d} \sup_{0\le t\le 1} |\mathbb{B}(t)| = Z.$$

Therefore, a level α test is given by

$$\Psi = \mathbb{1}(T_n > \tilde{q}_{1-\alpha}/\sqrt{n}),$$

where $\tilde{q}_{1-\alpha}$ is the $(1-\alpha)$ quantile of Z. The p-value of this test is given by $\mathbb{P}[Z > T_n^{obs}]$. This test is called the **Kolmogorov-Smirnov Test**.

9.3 Computing values for KS

Computing the test statistic:

Let $X_{(1)}, \dots, X_{(n)}$ be the reordered sample. Note that F_0 is non decreasing, while F_n is piecewise constant, such that F_n jumps from $(i-1)/n$ to i/n at $X_{(i)}$. Thus,

$$T_n = \max_{i=1,\dots,n} \left\{ \left| \frac{i-1}{n} - F_0(X_{(i)}) \right|, \left| \frac{i}{n} - F_0(X_{(i)}) \right| \right\}.$$

Computing the quantiles:

Let $U_i \sim F_0(X_i)$. If H_0 is true, then $U_1, \dots, U_n \sim \textsc{Unif}[0,1]$. (Intuitively, since we are drawing from the same distribution, the cdf probabilities should be distributed uniformly). In this case, $T_n = \sup_{0\le x\le 1} |G_n(x) - x|$, where G_n is the emprical cdf of $U_1, \dots, U_n$. Note that G_n does not depend on the distribution of the X_is (as long as the null hypothesis is true). Because of this property, we say that T_n is a **pivotal statistic**.

To compute the quantiles, first sample k batches of $(U_1, \dots U_n) \sim \textsc{Unif}[0,1]$. Then, compute the test statistics $(\tilde{T}_n^1, \dots, \tilde{T}_n^k)$. Estimate the $(1-\alpha)$ quantile $q_{1-\alpha}^{(n)}$ by taking the sample $(1-\alpha)$ quantile $\hat{q}_{1-\alpha}^{(n,k)}$ of our test statistics. The way that sample quantile is defined is by turning our k test statistics into a histogram, and then taking the $(1-\alpha)$ quantile of that histogram.

9.4 Kolmogorov-Lilliefors Test

To test if X is gaussian, we would need to know the exact parameters of the gaussian in order to perform Kolmogorov-Smirnov (since we assumed we knew F_0 precisely). So, if we want to test if it is gaussian in general, we can instead use the test statistic

$$\hat{T}_n = \sup_{t \in \mathbb{R}} |F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t)|,$$

where $\hat{\mu} = \overline{X}_n$ and $\hat{\sigma}^2 = S_n^2$. This is the **Kolmogorov-Lilliefors Test**.

9.5 Linear Regression

Given two random variables (X, Y) we want to compute a regression function to predict Y from X. Our samples are (X_i, Y_i) i.i.d. from an unkonwn joint distribution $\mathbb{P}$. Two ways that this distribution could be described:

- a joint pdf $h(x, y)$
- the marginal density $h(x) = \int h(x, y) \mathrm{d}y$, and a conditional density $h(y|x) = h(x, y)/h(x)$.

So, we may theoretically extract some information about our regression in the following ways:

- $f(x) = \mathbb{E}[Y|X = x] = \int y h(y|x) \mathrm{d}y$. The regression function $f(x)$ is often hard to compute in practice.
- Conditional median:
$$\int_{-\infty}^{m} h(y|x) \mathrm{d}y = \frac{1}{2}.$$
- We can also extract conditional quantiles, using the same idea as above.

In linear regression specifically, we assume $\mathbb{E}[Y|X=x] = a+bx$. The theoretical linear regression is given by

$$(a^*, b^*) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min}\, \mathbb{E}[(Y-a-bX)^2].$$

Setting the partial derivatives equal to 0,

$$(a^*, b^*) = \left(\mathbb{E}[Y] - b^*\mathbb{E}[X], \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}[X]}\right).$$

The random variable $\varepsilon = Y - (a^* + b^*X)$ is called **noise**, and satisfies $\mathbb{E}[\varepsilon] = 0$, $\operatorname{Cov}(X, \varepsilon) = 0$.

10 March 16, 2023

Linear regression setup: we desire

$$\underset{(a,b)}{\arg\min}\, \frac{1}{n}\sum_{i=1}^{n} |Y_i - ax_i - b|^2.$$

If we instead use perpendicular distances instead of vertical distance, we would want to minimize

$$\frac{1}{n}\sum |Y_i - ax_i - b|^2 \sin^2\theta,$$

where θ is the angle between the line and the vertical. Since $\tan\theta = a$ (the slope of the line), $\sin\theta = a/(\sqrt{1-a^2})$, so the expression becomes

$$\frac{1}{n}\sum |Y_i - ax_i - b|^2 \frac{a^2}{1-a^2}.$$

10.1 Multivariate Linear Regression

(this is basically the same content covered in the ML notes)

The setup here is that we have n outputs Y_i, and n $(k+1)$-dimensional explanatory variables $X_i = [1, x_i^{(1)}, \ldots, x_i^{(k)}]^T \in \mathbb{R}^{(k+1)}$, which are linearly related. Our goal is to construct a linear model β between the explanatory variables and the outputs.

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i = X_i^T\beta + \epsilon_i,$$

for all $i \in \{1, \dots, n\}$. β_0 is the **intercept**. Like normal linear regression, ε_i is gaussian noise.

Definition 10.1

The **least squares estimator** (LSE) of β is given by

$$\hat{\beta} = \underset{(\beta \in \mathbb{R}^{k+1})}{\arg\min} \sum_{i=1}^{k} (Y_i - X_i^T \beta)^2.$$

The matrix $\mathbb{X} = [X_1^T, \dots, X_n^T]^T \in \mathbb{R}^{n \times (k+1)}$ is called the **design matrix**. Let $Y = [Y_1, \dots, Y_n]^T \in \mathbb{R}^n$ and $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]^T \in \mathbb{R}^n$. Then, we want to find the best-fit β such that

$$Y = \mathbb{X}\beta + \varepsilon.$$

As before, the LSE is given by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{n+1}}{\arg\min} |Y - \mathbb{X}\beta|_2^2.$$

(The subscript 2 notation says that we are in the L_2 norm, i.e., normal Euclidean distance). **This has an analytic solution:**

$$\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y.$$

Assumptions about this model:

- The model is **homoscedastic**, i.e., all ε_i are i.i.d.
- Noise is gaussian, i.e., $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, for some σ^2.

Other important things to know:

- $\hat{\beta}$ is normally distributed: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$
- $\mathbb{E}[|Y - \mathbb{X}\hat{\beta}|^2] = \sigma^2 (n - k - 1)$
- Unbiased estimator of σ^2: $\hat{\sigma^2} = \dfrac{|Y - \mathbb{X}\beta|_2^2}{n - k - 1}$.
- $\dfrac{|Y - \mathbb{X}\hat{\beta}|_2^2}{\sigma^2} \sim \chi^2_{n-k-1}$. This is true by Cochran's Theorem, since $|Y - \mathbb{X}\hat{\beta}|_2^2 = \hat{\sigma^2}(n - k - 1)$.

10.2 Significance Testing

To test whether the jth explanatory variable is significant in the linear regression, we use

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0.$$

Then, a test with non-asymptotic level α:

$$\Psi = \mathbb{1}\left(\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma^2}\gamma_j}} > q_{\alpha/2}^{(t_{n-k-1})} \right),$$

where γ_j is the jth diagonal coefficient of $(X^T X)^{-1}$.

10.3 Non-parametric Regression

Non-parametric regression is a regression model that does *not* make a parametric assumption about $f(x) = \mathbb{E}[Y_i | X_i = x]$, $x \in \mathbb{R}^{(k+1)}$. Examples of parametric assumptions:

- $f(x) = a + bx$ (linear regression)
- $f(x) = e^{a+bx}$

In parametric cases, we can use LSE and MSE theory to estimate the function f.

Example 10.2

Non-parametric regression: take local averages.

One idea of a non-parametric regression model is to assume that f is very smooth, and take local averages. Let $h > 0$, and $I_x = \{i \in \{1, \ldots, n\} : |X_i - x| < h\}$. Then, we can approximate f with

$$\hat{f}_{n,h}(x) = \begin{cases} \dfrac{1}{|I_x|} \displaystyle\sum_{i \in I_x} Y_i & I_x \neq \emptyset \\ 0 & \text{else.} \end{cases}$$

11 March 21, 2023

11.1 Nonparametric Regression

Recap from last time: non-parametric regression is when we do not make any parametric assumptions on f. For example, a parametric assumption on f would be to assume that $f(x) \in \{a+bx, a+bx+cx^2, e^{a+bx}, \ldots\}$.

One non-parametric regressor is to take **local averages**. This assumes that f is “smooth”, and can thus be well approximated by some piecewise constant function, e.g., $f(t) \approx f(x)$ for t close enough to x.

Let $h > 0$ be the **window size** or **bandwidth**. Then define

$$I_x = \{i = 1, 2, \ldots, n : |X_i - x| < h\}.$$

Our regressor is then given by

$$\hat{f}_{n,h}(x) = \begin{cases} \frac{1}{|I_x|} \sum_{i \in I_x} Y_i & I_x \neq \emptyset \\ 0 & \text{else.} \end{cases}$$

For this to be a good regressor, we need to choose an appropriate bandwidth. As $h \to 0$, the model becomes overfit, and as $h \to \infty$, the model becomes underfit (just a straight horizontal line).

Example 11.1

We will show one way to choose a “smart” value for h. Let $x_i = i/n$ for $i \in \{0\} \cup [n]$. Let $Y_i = f(x_i) + \varepsilon_i$ for standard normally distributed ε_i. Suppose that we know $|f'(x)| \leq L$ for $x \in [0,1]$.

Construct the non-parametric estimator $\hat{f}(x_i) = (\sum_{j \in I_i} Y_j)/|I_i|$, where $I_i = \{j : |j - i| \leq k\}$.

Claim 11.2 (Variance)

$$\mathrm{Var}[\hat{f}(x_i)] \leq \frac{1}{k}.$$

Proof. $\mathrm{Var}[Y_i] = \mathrm{Var}[\varepsilon_i] = 1$. Therefore, $\mathrm{Var}[\hat{f}(x_i)] = \sum_{j \in I_i} \mathrm{Var}[Y_j]/|I_i|^2 \le 1/k$, since $|I_i| \ge k$. □

Claim 11.3 (Bias)

$$|\mathbb{E}[\hat{f}(x_i)] - \mathbb{E}[f(x_i)]| \le \frac{Lk}{n}.$$

Proof. Since $|f'(x_i)| \le L$, the farthest neighboring point (i.e., distance k/n away) differs from $f(x_i)$ by at most Lk/n. Therefore, the average distance from neighboring points differs from $f(x_i)$ is also bounded above by Lk/n. □

Claim 11.4 (Quadratic Risk)

Quadratic risk is the sum of variance and bias:

$$\mathbb{E}[(\hat{f}(x_i) - f(x_i))^2] \le \frac{L^2k^2}{n^2} + \frac{1}{k}.$$

So, we can minimize the quadratic risk by choosing $k = (n/L)^{2/3}$.

11.2 Exponential Family

The family of exponential distributions $\mathbb{P}_\eta$ has density function

$$h(x)\exp(\eta^T t(x) - a(\eta)).$$

- η and t are called sufficient statistics. Note that η is a vector of inputs to the distribution, hence the transpose
- h is called the "underlying measure"

$a(\eta)$ is chosen so that we get a probability distribution:

$$a(\eta) = \log \int h(x)\exp(\eta^T t(x))\mathrm{d}x.$$

Example 11.5

The normal distribution is an exponential family.

Probability density can be expressed in the following way:

$$\begin{aligned}\mathbb{P}_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\\ &= \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma - 0.5\log(2\pi)\right).\end{aligned}$$

- $t(x) = (x, x^2)$
- $\eta = (\mu/\sigma^2, -1/(2\sigma^2))$
- $a(\eta) = \mu^2 + \log\sigma + 0.5\log(2\pi) = -\eta_1^2/4\eta_2 - 0.5\log(-2\eta_2) + 0.5\log(2\pi)$.
- $h(x) = 1$

Claim 11.6

- $\mathbb{E}_{\mathbb{P}_\eta}[t(x)] = \nabla a(\eta)$.
- $\mathrm{Cov}[t(x)] = \nabla^2 a(\eta)$.

12 March 23, 2023

12.1 Generalized Linear Models

This is a special case of the exponential distributions introduced last lecture:

Definition 12.1 (One parameter canonical exponential family)

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right).$$

It is possible to show that $\mathbb{E}[Y/\phi] = b'(\theta)/\phi$ and $\mathrm{Var}[Y/\phi] = b''(\theta)/\phi$, which gives $\mu = \mathbb{E}[Y] = b'(\theta)$ and $\mathrm{Var}[Y] = \phi b''(\theta)$. Here, ϕ acts as a **scale** or **dispersion** parameter, since changing it affects the variance but not the mean of the distribution.

Definition 12.2 (Generalized Linear Model)
A GLM consists of

- A conditional on X, $Y|X$, which is a exponential family.
- A link g:
$$\mu(X) = \mathbb{E}[Y|X] = g^{-1}(X^T\beta).$$

13 April 4, 2023

13.1 Principal Component Analysis

The setup for PCA is that we have some high dimensional data that can be clustered approximately into clouds. We want to project the data into a lower dimension such that the variation between the clouds is still captured. Lower dimensional data saves computational power; we want to sacrifice as little of the depth of the data as possible by doing reducing its dimensionality.

$X_1, \dots, X_n \in \mathbb{R}^d$. $\mathbb{X} = [X_1^T, \dots, X_n^T]^T \in \mathbb{R}^{n \times d}$. The matrix

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

is the sample covariance matrix.

– unclear: the professor says that the usage of $1/n$ instead of $1/(n-1)$ for sample covariance / variance is specific to PCA –

Let $U \in \mathbb{R}^d$. We want $U^TX_1, U^TX_2, \dots, U^TX_n = Y_1, \dots, Y_n$ to have maximal variation. The sample variance of $Y_1, \dots, Y_n$ is

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Claim 13.1

The sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = U^T S U.$$

Proof. We have $Y_i - \overline{Y} = U^T X_i - U^T \overline{X}_n = U^T(X_i - \overline{X}_n)$. Therefore,

$$(Y_i - Y)^2 = (Y_i - \overline{Y}) \cdot (Y_i - \overline{Y}) = U^T(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T U,$$

since $(Y_i - \overline{Y})$ is a scalar and equal to its own transpose. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = U^T\left(\frac{1}{n}\sum(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right)U = U^T S U.$$

□

Definition 13.2 (Positive Definite Matrices)

If A is a symmetric matrix, A is positive definite if all eigenvalues of A are positive. A is positive semidefinite if all eigenvalues of A are nonnegative. Alternatively, A is positive definite if $U^T A U > 0$ for all $U \neq 0$, and A is positive semidefinite if $U^T A U \geq 0$ for all $U \neq 0$.

Claim 13.3

S is positive semidefinite.

Proof. $U^T(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T U = (U^T(X_i - \overline{X}_n))^2 \geq 0$. □

Fix U_0 and define $U_r = rU_0$. Then, $U_r^T S U_r = r^2 U_0^T S U_0$.

- For any U_0, $U_0^T S U_0 \geq 0$. If $U_0^T S U \neq 0$, then we can increase r to make the variance grow unbounded.

The first principle component is $u_1 = \arg\max_{|U|\leq 1} U^T S U$. The second principle component is $u_2 = \arg\max_{|U|\leq 1, U \perp u_1} U^T S U$. We can continue defining the nth principle component following this general pattern. These vectors represent the vectors among which the variation in the data is maximized, when the data is projected along these vectors.

Theorem 13.4

By the spectral decomposition theorem, any real symmetric matrix A has spectral decomposition

$$A = PDP^T,$$

where P is an orthonormal matrix containing the eigenvectors of A, and D is a diagonal matrix containing the corresponding eigenvalues.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ be the eigenvalues of A. Using the spectral decomposition of A, it is possible to show that the first principle component is given by the eigenvector corresponding to λ_1. Similarly, the second principle component is the second largest eigenvector subject to the constraint that it is orthogonal to the first eigenspace. And so on. (See 701 notes for more detail)

14 April 11, 2023

14.1 Classification

A set of images encoded as vectors $X = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$. Given an input set of data $\{(X_1, Cat), (X_2, Dog), \ldots\}$, the goal is to learn from this data and come up with a classification rule.

Definition 14.1

The classification rule is a function h such that $h(x) \in \{0, 1\}$.

Error/risk function for a classifier:

$$L(h) = \mathbb{E}[\mathbb{1}(h(x) \neq Y)] = \mathbb{P}[h(x) \neq Y].$$

The empirical risk function is

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(x_i) \neq Y_i).$$

Theorem 14.2

Define $h^*(X) = \mathbb{1}(r(X) > 1/2)$ where $r(X) = \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$. The Bayes classifier h^* minimizes the true error among all classifiers, i.e.,

$$\mathbb{P}(h^*(X) \neq y) \leq \mathbb{P}(h(X) \neq y) \quad \forall h.$$

Proof. Since $\mathbb{P}[h(X) \neq y] = \mathbb{E}_X[\mathbb{E}[(h(X) \neq y)|X = x]]$, it suffices to show that

$$\mathbb{P}[h^*(X) \neq y|X = x] - \mathbb{P}[h(X) \neq y|X = x] \leq 0.$$

Fix some hypothesis h. If $h^*(x) = h(x)$, then $\mathbb{P}[h^*(X) \neq y|X = x] = \mathbb{P}[h(X) \neq y|X = x]$, so the result holds.

So, assume $h^*(x) \neq h(x)$. Given that $h(x) \neq y$, we know that $h^*(x) = y$, since $y \in \{0,1\}$, so $\mathbb{P}[h(X) \neq y|X = x] = \mathbb{P}[h^*(X) = y|X = x]$, and it thus suffices to show that

$$\mathbb{P}[h^*(X) \neq y|X = x] \leq \mathbb{P}[h^*(X) = y|X = x].$$

- If $r(x) \leq 1/2$, then $h^*(x) = 0$, so we want to show $\mathbb{P}[Y = 1|X = x] \leq \mathbb{P}[Y = 0|X = x] \iff r(x) \leq 1 - r(x) \iff r(x) \leq 1/2$, which we assumed to be true.
- If $r(x) > 1/2$, then $h^*(x) = 1$, so we want $\mathbb{P}[Y = 0|X = x] \leq \mathbb{P}[Y = 1|X = x] \iff 1 - r(x) \leq r(x) \iff r(x) \geq 1/2$, which we assumed to be true.

□

How do we compute $r(x) = \mathbb{P}[Y = 1|X = x]$? Given:

- $Y \in \{0,1\}$, $Y \in \textsc{Bern}(\pi_0)$
- $f_X(X = x|Y = 1) = f_1(x)$ is the conditional density of X given $Y = 1$
- $f_X(X = x|Y = 0) = f_0(x)$ is the conditional density of X given $Y = 0$.

Then, by Bayes' formula,

$$r(x) = \frac{\pi_0 f_1(x)}{\pi_0 f_1(x) + (1 - \pi_0) f_0(x)}.$$

Given data, we need a way to estimate π_0, f_1, f_0, in order to compute $r(x)$ in this way. Estimator for $\hat{\pi}_0$:

$$\hat{\pi}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i = 1).$$

We can estimate the density functions with **Kernel Density Estimation**:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

where h can be thought of some sort of bandwidth, and $K(x) = e^{-x^2/2}/\sqrt{2\pi}$ is Gaussian.

15 April 13, 2023

15.1 Nearest Neighbors Classifier

From last lecture,

$$r(x) = \mathbb{P}[Y = 1|X = x] = \frac{\pi_0 f_1(x)}{\pi_0 f_1(x) + (1-\pi_0) f_0(x)},$$

where $Y_i \sim \textsc{Bern}(\pi_0)$, and f_1 and f_0 are the conditional densities of x given $Y = 1$ and $Y = 0$, respectively.

The most accurate hypothesis is the Bayes classifier $h(x) = \mathbb{1}(r(x) > 1/2)$. If we assume $\pi_0 = 1/2$, this rearranges to $h(x) = \mathbb{1}(f_1(x) > f_0(x))$.

Definition 15.1

In a nearest neighbor classifier, suppose the data x_i is d-dimensional. Then we use the following kernel density estimator:

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{|x_i - x|}{h}\right),$$

where $K(x) = \mathbb{1}(|x| \le 1)/2$. Then, $\hat{f}_1(x) = \hat{f}(x)\mathbb{1}(Y_i = 1)$ and $\hat{f}_0(x) = \hat{f}(x)\mathbb{1}(Y_i = 0)$.

Recall that last time we used a gaussian kernel, which leads to a different estimator. The nearest neighbors classifier predicts $y = 1$ when $\hat{f}_1(x) > \hat{f}_0(x)$.

If we did not want to naively assume that $\pi_0 = 1/2$, similar logic can be applied with the estimator $\hat{\pi}_0 = \sum_{i=1}^n \mathbb{1}(Y_i = 1)/n$:

$$\begin{aligned} h(x) = \mathbb{1}(r(x) > 1/2) &= \mathbb{1}\left(\frac{\hat{\pi}_0 f_1(x)}{\hat{\pi}_0 f_1(x) + (1-\hat{\pi}_0) f_0(x)} > \frac{1}{2}\right) \\ &= \mathbb{1}(\hat{\pi}_0 \hat{f}_1(x) > (1-\hat{\pi}_0)\hat{f}_0(x)). \end{aligned}$$

(But this is generally how nearest neighbor is implemented).

15.2 Determinant Analysis

Let Σ_i be the covariance matrix for all data classified as $i \in \{0,1\}$. Then, we can use a kernel density estimator with gaussian kernel:

$$f_i(x) = \frac{1}{(2\pi^d \det(\Sigma_i))^{1/2}} \exp\left(-\frac{(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}{2}\right),$$

where we use the fact that $U^T S U$ gives the variance of data point U given covariance matrix S (this was proven two lectures ago).

Here, the threshold for classification $f_1(x) > f_0(x)$ reduces to

$$(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) \leq (x-\mu_0)^T \Sigma^{-1}(x-\mu_0),$$

which assumes that $\Sigma_1 = \Sigma_0 = \Sigma$.

Note **explain this** that

$$x^T(\Sigma^{-1}(\mu_1 - \mu_0)) = \mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0.$$

If μ_0, μ_1, Σ are not known, we may use the estimators

$$\hat{\mu}_i = \frac{\sum_{j=1}^n X_j \mathbb{1}(Y_j = i)}{\sum_{j=1}^n \mathbb{1}(Y_j = i)},$$

and

$$\hat{\Sigma} = \frac{1}{n}\left(\sum_{i=1}^n \mathbb{1}(Y_i = 1)(X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T + \sum_{i=1}^n \mathbb{1}(Y_i = 0)(X_i - \hat{\mu}_0)(X_i - \hat{\mu}_0)^T\right).$$

16 April 18, 2023

16.1 Bayesian vs Frequentist statistics

Consider the kissing example. Suppose $R_1, \ldots, R_n \sim \textsc{Bern}(p)$. We would like to test $H_0 : p = 1/2$, $H_1 : p \neq 1/2$.

Example 16.1
Frequentist analysis.

In frequentist statistics, we assume that parameter p is a true constant, and use estimator $\hat{p} = \mathbb{E}[R_i]$ to estimate the value of this parameter. All of the tools we have discussed so far (CI, Wald test, etc.) would be relevant to conduct this test.

Example 16.2
Bayesian analysis.

We take p to be a random variable with some **prior distribution**. In this case, say $p \sim \textsc{Beta}(a, b)$, where the beta distribution is defined with pdf

$$f_{(a,b)}(x) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{\textsc{beta}(a,b)} & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases},$$

and $\textsc{beta}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. Then, we say that $R_1|p, \ldots, R_n|p \sim \textsc{Bern}(p)$.

The conditional density is

$$f(R_1 = r_1, \ldots, R_n = r_n|p) = p^{\sum_{i=1}^n r_i}(1-p)^{n-\sum_{i=1}^n r_i}.$$

To find the marginal distribution of R_i, we integrate out p:

$$f(R_1 = r_1, \ldots, R_n = r_n) = \int_0^1 p^{\sum_{i=1}^n r_i}(1-p)^{n-\sum_{i=1}^n r_i} \frac{p^{a-1}(1-p)^{b-1}}{\textsc{beta}(a,b)} \mathrm{d}p.$$

Note that this value does not depend on the value of p. It represents the general probability that we observed the given data over all possible values of p.

For any single data point, $f(R_i = r_i|p) = p^{r_i}(1-p)^{1-r_i}$, so $\mathbb{E}[R_i|p] = p$, and by the towering property of conditional expectation,

$$\mathbb{E}[R_i] = \mathbb{E}[\mathbb{E}[R_i|p]] = \mathbb{E}[p] = \int_0^1 p\frac{p^{a-1}(1-p)^{b-1}}{\text{BETA}(a,b)}\mathrm{d}p.$$

The **posterior distribution** is the conditional distribution of p given $R_1,\ldots,R_n$:

$$\begin{aligned}f(p|R_1 = r_1,\ldots,R_n = r_n) &= \frac{f(P=p)f(R_1 = r_1,\ldots,R_n = r|p)}{f(R_1 = r_1,\ldots,R_n = r_n)}\\ &= \frac{p^{\sum_{i=1}^n r_i}(1-p)^{n-\sum_{i=1}^n r_i}\frac{p^{a-1}(1-p)^{b-1}}{\text{BETA}(a,b)}}{\int_0^1 p^{\sum_{i=1}^n r_i}(1-p)^{n-\sum_{i=1}^n r_i}\frac{p^{a-1}(1-p)^{b-1}}{\text{BETA}(a,b)}\mathrm{d}p}.\end{aligned}$$

Claim 16.3

$f(p|R_1 = r_1,\ldots,R_n = r_n)$ is the same density as $\text{Beta}(a+\sum_{i=1}^n r_i, b+n-\sum_{i=1}^n r_i)$.

Given $p \sim \text{Beta}(a,b)$, it can be shown that $\mathbb{E}[p] = a/(a+b)$. By the above claim,

$$\mathbb{E}[p|R_1 = r,\ldots,R_n = r] = \frac{a+\sum_{i=1}^n r_i}{b+n-\sum_{i=1}^n r_i + a + \sum_{i=1}^n r_i} = \frac{a/n + 1/n\sum_{i=1}^n r_i}{a/n+b/n+1}.$$

As $n \to \infty$, $\mathbb{E}[p|R_1 = r_1,\ldots,R_n = r_n]$ approaches $\mathbb{E}[R_i]$.

16.2 Choices of Priors

- **Non-informative priors** give no preference to any particular point. For example, if $a = b = 1$, then $\text{Beta}(1,1) = \text{Unif}[0,1]$.

- **Jeffrey's Prior** $\pi(\theta) \propto \sqrt{I(\theta)}$.

17 April 25, 2023

17.1 Bootstrapping

Definition 17.1

Bootstrapping is a simulation based method to assess the variability of any estimator.

Consider some

$$\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \dots, \hat{\theta}_n^{(k)} \sim P.$$

A reasonable estimator for the mean: $\hat{\theta}_n^* = \sum_{i=1}^k \hat{\theta}_n^{(i)}/k$. By the strong law of large numbers, this is an unbiased estimator, i.e., it converges almost surely to $\mathbb{E}[\hat{\theta}_n]$. A reasonable estimator for variance: $v_{boot} = \sum_{i=1}^k (\hat{\theta}_n^{(i)} - \hat{\theta}_n^*)^2$. By the strong law of large numbers, this also converges to $\mathrm{Var}[\hat{\theta}_n]$.

Main idea: split the data into some number of blocks, and then compute effectively i.i.d. estimators:

$$\begin{aligned} X^{(1)} &= \{X_1^{(1)}, \dots, X_n^{(1)}\} \to \hat{\theta}_n^{(1)} \\ X^{(2)} &= \{X_1^{(2)}, \dots, X_n^{(2)}\} \to \hat{\theta}_n^{(2)} \\ &\vdots \end{aligned}$$

To construct confidence intervals:

Example 17.2

Constructing normal confidence intervals from bootstrapped samples.

Given $\hat{\theta}$ asymptotically normal, i.e., $\sqrt{n}(\hat{\theta} - \theta_n) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V(\hat{\theta}_n))$, then

$$C.I._{boot} = \left[\hat{\theta}_n - Z_{\alpha/2}\sqrt{v_{boot}}, \hat{\theta}_n + Z_{\alpha/2}\sqrt{v_{boot}}\right].$$

Example 17.3

Constructing **pivotal intervals** from bootstrapped samples.

This method is generally more popular than the first method. The goal is to find t such that

$$\mathbb{P}[-t \le \hat{\theta} - \theta \le t] \approx 1 - \alpha,$$

so that $[\hat{\theta} - t, \hat{\theta} + t]$ is a $(1 - \alpha)$-level confidence interval.

We can compute some estimator for this value t_α s.t.

$$\frac{1}{B} \sum_{i=1}^{B} \mathbb{1}(-t_\alpha \le \hat{\theta}_n - \hat{\theta}^{(i)} \le t_\alpha) \approx 1 - \alpha.$$

Then, our interval is

$$C.I._{boot} = [\hat{\theta}_n - t_\alpha, \hat{\theta} + t_\alpha].$$

Example 17.4

Histogram based confidence intervals.

Compute $\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(B)}$. Then, construct empirical quantiles by plotting these values in a histogram, and take

$$C.I._{boot} = [q^*_{\alpha/2}, q^*_{1-\alpha/2}].$$

18 April 27

Review session.

18.1 Exponential Distributions

$$\mathbb{P}_\eta(x) = h(x) \exp(\eta^T t(x) - a(\eta)) \mathrm{d}x,$$

where $a(\eta)$ is a normalizing factor so that $\int \mathbb{P}_\eta(x) \mathrm{d}x = 1$.

Example 18.1

Show that $\partial a / (\partial \eta_i) = \mathbb{E}[t_i(x)]$.

Given

$$1 = \int h(x)\exp(\eta^T t(x) - a(\eta))\mathrm{d}x,$$

we know

$$\exp(a(\eta)) = \int h(x)\exp(\eta_1 t_1(x) + \eta_2 t_2(x) + \ldots + \eta_n t_n(x))\mathrm{d}x.$$

Differentiate both sides with respect to η_i:

$$\partial a_i/\partial\eta_i \cdot (\exp(a(\eta))) = \int t_i(x)h(x)\exp(\eta^T t(x))\mathrm{d}x,$$

which implies

$$\partial a_i/\partial\eta_i = \int t_i(x)\mathbb{P}_\eta(x)\mathrm{d}x = \mathbb{E}[t_i(x)].$$

Theorem 18.2

More generally, $\nabla a(\eta) = \mathbb{E}[t(x)]$ and $\nabla^2 a(\eta) = \mathrm{Cov}(t(x))$.

18.2 Bayesian Statistics

Let $X_1, \ldots, X_n \sim \textsc{Exp}(\lambda)$.

- classical: make an asymptotic statement about λ, create confidence intervals, etc.
- Bayesian: assume λ has a distribution, use the data to refine the distribution.

Bayesian statistics starts with a **prior** $\pi(\lambda)$, which is the initial distribution that we guess that λ has. The model creates a **posterior** $\pi(\lambda|X_1, \ldots, X_n)$, which is the updated distribution after considering the data.

Bayes rule: posterior is proportional to the prior times the likelihood of the data:

$$\pi(\lambda|X_1, \ldots, X_n) \propto \pi(\lambda) \cdot \mathbb{P}(X_1, \ldots, X_n|\lambda).$$

Estimators:

- MAP (maximum a posterior): the value of λ maximizing the posterior
- Bayes: the mean of $\pi(\lambda|X_1, \ldots, X_n)$.

Note: given a uniform prior, maximizing the posterior is the same as maximizing the likelihood, so the MAP is the same as the MLE.

Example 18.3

Given data $X_1, \dots, X_n \sim \textsc{Pois}(\lambda)$ and prior $\lambda \sim \textsc{Unif}[0,10]$, compute the MAP and Bayes estimators.

We know

$$\pi(\lambda | X_1, \dots, X_n) \propto \pi(\lambda)\mathbb{P}(X_1, \dots, X_n | \lambda).$$

The density of $\textsc{Pois}(\lambda)$ is given by $f_\lambda(x) = \lambda^x e^{-\lambda}/x!$, so

$$\pi(\lambda | X_1, \dots, X_n) = c \cdot \frac{1}{10}\mathbb{1}(\lambda \in [0,10])\lambda^{\sum X_i} e^{-\lambda n},$$

for some normalizing constant c. From here, asssume $X_1, \dots, X_n = 0$ to save some computation. First, we compute the constant:

$$1 = \int_0^{10} \frac{1}{10} c e^{-\lambda n} \mathrm{d}\lambda = \frac{c(1 - e^{-10n})}{10n} \implies c = \frac{10n}{1 - e^{-10n}},$$

which gives us the posterior:

$$\pi(\lambda | X_1, \dots, X_n) = \frac{n}{1 - e^{-10n}} e^{-n\lambda} \mathbb{1}(\lambda \in [0,10]).$$

Now, the Bayes estimator is given by

$$\hat{\lambda} = \mathbb{E}[\pi(\lambda | X_1, \dots, X_n)] = \int_0^{10} \frac{n\lambda}{1 - e^{-10n}} e^{-n\lambda} \mathrm{d}\lambda.$$

To find the MAP estimator, differentiate (we no longer assume the values of X_i since this is an easier computation):

$$\frac{\partial}{\partial \lambda}\left(\frac{1}{10}\lambda^{\sum X_i} e^{-nx}\right) = 0$$
$$\implies \lambda = \left(\sum X_i\right)/n.$$

With consideration to the constraint on λ, this gives us the MAP estimator $\hat{\lambda} = \min((\sum X_i)/n, 10)$.

Example 18.4

Compute Jeffrey's prior for $\textsc{Pois}(\lambda)$ and $\mathcal{N}(0,\theta)$.

- $\textsc{Pois}(\lambda)$: For a poisson distribution,

  $$f_\lambda(x) = \frac{\lambda^x}{x!}e^{-\lambda},$$

  so the total log likelihood is

  $$LL(X_i|\lambda) = \sum (X_i \log \lambda - \log x! - \lambda).$$

  Differentiating, we get

  $$\frac{\partial LL(X_i|\lambda)}{\partial \lambda} = \sum\left(\frac{X_i}{\lambda} - 1\right) \quad \text{and} \quad \frac{\partial^2 LL(X_i|\lambda)}{\partial \lambda^2} = \sum -\frac{x}{\lambda^2}.$$

  Thus the Fisher information is

  $$I(\lambda) = -\mathbb{E}_\lambda\left[\sum \frac{1}{2\lambda^2} - \frac{3x^2}{2\lambda^3}\right] = -\frac{n}{2\lambda^2} + \frac{3n}{2\lambda^3}\cdot\lambda = \frac{n}{\lambda^2},$$

  using the fact that $\mathbb{E}[x] = \lambda$. Finally, Jeffreys prior is

  $$\pi(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1}.$$

- $\mathcal{N}(0,\theta)$: For a normal distribution with mean 0 and variance θ,

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

  So, the total log likelihood is

  $$LL(X_i|\theta) = \sum_{i=1}^{n}\left(-\log\sqrt{2\pi} - \frac{1}{2}\log\theta - \frac{X_i^2}{2\theta}\right).$$

  Differentiating,

  $$\frac{\partial LL}{\partial \theta} = \sum_{i=1}^{n} -\frac{1}{2\theta} + \frac{X_i^2}{2\theta^2} \quad \text{and} \quad \frac{\partial^2 LL}{\partial \theta^2} = \sum_{i=1}^{n}\frac{1}{2\theta^2} - \frac{3X_i^2}{2\theta^3}.$$

Thus, the fisher information is

$$I(\theta) = -E_\theta\left[\sum_{i=1}^{n} \frac{1}{2\theta^2} - \frac{3X_i^2}{2\theta^3}\right] = -\frac{n}{2\theta^2} + \frac{3n}{2\theta^3}\theta = \frac{n}{\theta^2},$$

where we have used the fact that $\mathbb{E}[X^2] = \mathrm{Var}[X] = \theta$, since the mean is fixed at zero. Finally, Jeffreys prior is

$$\pi(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1}.$$