STAT 460/560 Class 5:
Delta Method
&
the big picture

Ben Bloem-Reddy

Reading: Chapter 3.1, 3.4, [van98]. Supplement: Chapter 5, [Was04].

1. Delta Method, intuitively

Suppose we have an estimator T_n for θ, but we're interested in some function $\phi(\theta)$. For example, we might have a sample of n patients who are tested for a particular medical condition. Let $X_1, \dots, X_n$ denote the outcome of those tests, modeled as IID $\text{Bern}(p)$, $p \in (0,1)$, random variables. We know by the CLT that $\sqrt{n}(\bar{X}_n - p) \rightsquigarrow \mathcal{N}(0, p(1-p))$. But maybe we're interested in the *log-odds*,

$$\phi(p) := \ln \frac{p}{1-p} \ .$$

A natural estimator of this is $\phi(\bar{X}_n)$. How does $\phi(\bar{X}_n) - \phi(p)$ behave for large n?

The continuous mapping theorem tells us that since $\bar{X}_n \xrightarrow{\text{p}} p$ (how do we know this?) and $\phi(p)$ is continuous for all $p \in (0,1)$, it is the case that $\phi(\bar{X}_n) \xrightarrow{\text{p}} \phi(p)$.

Great. But what about the limiting distribution? Informally, since ϕ is differentiable, it is reasonable to expect that

$$\sqrt{n}(\phi(\bar{X}_n) - \phi(p)) \approx \phi'(p)\sqrt{n}(\bar{X}_n - p) \sim \mathcal{N}(0, \phi'(p)^2 p(1-p)) \ .$$

The intuition here is based on the linear approximation of ϕ in the vicinity of p: $\phi'(p)h \approx \phi(p+h) - \phi(p)$ for small $\|h\|$. Since $\sqrt{n}(\bar{X}_n - p) \rightsquigarrow Z$ (for normal random variable Z), then for large enough n we expect that $\sqrt{n}(\phi(\bar{X}_n) - \phi(p)) \rightsquigarrow \phi'(p)Z$. We will prove this in generality today.

2. Delta Method, rigorously

Consider a vector-valued statistic, $T_n = (T_{n,1}, \dots, T_{n,k})$, and a function $\phi\colon \mathbb{R}^k \to \mathbb{R}^m$, which we assume is defined at least on a neighborhood of θ. The function ϕ is **differentiable** at $\theta \in \mathbb{R}^k$ if there exists a linear map (matrix) $\phi'_\theta\colon \mathbb{R}^k \to \mathbb{R}^m$ such that

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|) \ , \quad h \to 0 \ .$$

In practice, this gets formed by the matrix (function)

$$\phi'_\theta = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1}(\theta) & \dots & \frac{\partial \phi_1}{\partial x_k}(\theta) \\ \vdots & & \vdots \\ \frac{\partial \phi_m}{\partial x_1}(\theta) & \dots & \frac{\partial \phi_m}{\partial x_k}(\theta) \end{bmatrix} \ ,$$

so that the function $h \mapsto \phi'_\theta(h)$ is just matrix multiplication $\phi'_\theta h$, with $h \in \mathbb{R}^k$. If the dependence of ϕ'_θ on θ is continuous then ϕ'_θ is said to be **continuously differentiable**.

Here's the main result on the delta method.

Theorem 5.1. *Let $\phi\colon \mathbb{R}^k \to \mathbb{R}^m$ be a map defined on a subset of $\mathbb{R}^k$ and differentiable at θ. Let T_n be random vectors that take values in the domain of ϕ. If $r_n(T_n - \theta) \rightsquigarrow T$ for numbers $r_n \to \infty$, then:*

(i) $r_n(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$*; and*

(ii) $r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) \xrightarrow{p} 0$.

The proof uses two results from Chapter 2 of [van98] that we skipped. The first is Prohorov's theorem (Theorem 2.4 in [van98]), which says in part that if $X_n \rightsquigarrow X$ then the sequence X_n is bounded in probability, i.e., $X_n = O_p(1)$. The second result is Lemma 2.12(i), which says the following. Suppose R is a function such that $R(0) = 0$, that X_n takes values in the domain of R, and that $X_n \xrightarrow{p} 0$. If $R(h) = o(\|h\|^p)$ as $h \to 0$ for some $p > 0$, then $R(X_n) = o_p(\|X_n\|^p)$.

We'll also need the following lemma.

Lemma 5.2. *If the sequence $r_n X_n$, for numbers $r_n \to \infty$, is bounded in probability, then $X_n \xrightarrow{p} 0$.*

Proof. Because $r_n X_n$ is bounded in probability, for every $\epsilon > 0$ there is some M such that

$$\sup_n P(r_n\|X_n\| > M) < \epsilon$$
$$\sup_n P(\|X_n\| > M/r_n) < \epsilon \; .$$

Since M is finite and $r_n \to \infty$, this implies that $X_n \xrightarrow{p} 0$. □

Proof of Theorem 5.1. Because $r_n(T_n - \theta)$ converges in distribution, by Prohorov's theorem it is bounded in probability, i.e.,

$$r_n\|T_n - \theta\| = O_P(1) \; .$$

File this fact away for later. Moreover, by the lemma above, $T_n - \theta \xrightarrow{p} 0$.

Now define the remainder function

$$R(h) = \phi(\theta + h) - \phi(\theta) - \phi'_\theta(h) \; .$$

By the differentiability of ϕ, we have that $R(h) = o(\|h\|)$ as $h \to 0$. Since $T_n - \theta \xrightarrow{p} 0$, we have by Lemma 2.12(i) that

$$R(T_n - \theta) = \phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta) = o_P(\|T_n - \theta\|) \; .$$

We can multiply by sides by r_n, and using the o_P-calculus rules (p. 13 in [van98]) along with the fact that $\phi'_\theta(\,\bullet\,)$ is linear, we get

$$r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) = o_P(r_n\|T_n - \theta\|) = r_n\|T_n - \theta\| o_P(1) = O_P(1)o_P(1) = o_P(1) \; .$$

That proves (ii).

Now, matrix multiplication is continuous, so by the continuous mapping theorem, $\phi'_\theta(r_n(T_n - \theta)) \rightsquigarrow \phi'_\theta(T)$. We can apply Slutsky's lemma to conclude that

$$\begin{aligned} r_n(\phi(T_n) - \phi(\theta)) &= r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) + \phi'_\theta(r_n(T_n - \theta)) \\ &\rightsquigarrow 0 + \phi'_\theta(T) = \phi'_\theta(T) \; . \end{aligned}$$

□

A common situation (illustrated in the next activity) is that $\sqrt{n}(T_n - \theta) \xrightarrow{p} \mathcal{N}_k(\mu, \Sigma)$. Then the delta method (and the above theorem) indicates that

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \rightsquigarrow \mathcal{N}_m(\phi'_\theta\mu, \phi'_\theta\Sigma(\phi'_\theta)^\top) \; .$$

Activity 5.1.

Suppose we have a random sample of n patients who are tested for a particular medical condition. Let $X_1, \ldots, X_n$ denote the outcome of those tests, modeled as IID Bern(p), $p \in (0,1)$, random variables. Suppose that $Y_1, \ldots, Y_n \sim_{\text{IID}} \text{Bern}(q)$, $q > 0$, are the outcome of tests for a different medical condition from an *independent* sample of patients. Show the following:

1. As $n \to \infty$, $\sqrt{n}(\ln(\bar{X}_n) - \ln(p)) \rightsquigarrow \mathcal{N}(0, (1-p)/p)$.
2. As $n \to \infty$, $\sqrt{n}(\ln(\bar{X}_n/\bar{Y}_n) - \ln(p/q)) \rightsquigarrow \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = (1-p)/p + (1-q)/q$.

Solution:

Since $X_1, \ldots, X_n \sim_{\text{IID}} \text{Bern}(p)$, by the CLT,

$$\sqrt{n}(\bar{X}_n - p) \rightsquigarrow \mathcal{N}(0, p(1-p)) \ .$$

In terms of the theorem above, $r_n = \sqrt{n}$, $T_n = \bar{X}_n$, $\theta = p$, and $T \sim \mathcal{N}(0, p(1-p))$. Similarly for $\bar{Y}_n$.

For 1., consider the function $\phi(p) = \ln(p)$, which is continuously differentiable for $p > 0$, i.e., $\phi'_p = 1/p$. Applying the delta method, we find that

$$\sqrt{n}(\ln(\bar{X}_n) - \ln(p)) \rightsquigarrow \mathcal{N}(0, (1-p)/p) \ .$$

Similarly,

$$\sqrt{n}(\ln(\bar{Y}_n) - \ln(q)) \rightsquigarrow \mathcal{N}(0, (1-q)/q) \ .$$

For part 2., since the samples are independent, joint convergence holds, i.e.,

$$\sqrt{n}(\bar{X}_n - p, \bar{Y}_n - q) \rightsquigarrow \mathcal{N}_2(0, \Sigma) \ ,$$

with Σ a diagonal matrix with non-zero entries $p(1-p)$ and $q(1-q)$.

Now the function $\phi(p,q) = \ln(p) - \ln(q)$ is continuously differentiable in the positive quadrant, with $\phi'_{(p,q)} = [1/p, 1/q]$ a row vector (or 1×2 matrix). We find that in this case, $\phi'_\theta \Sigma (\phi'_\theta)^\top) = \sigma^2$ as defined.

The big picture

Reading: Chapter 6, [Was04]; Chapter 1, [EH21].

The point of today's class is to discuss the big picture of what statistical inference (and related ideas) is, and how we construct models and algorithms to help us perform it. At a high level, the whole point of this is to use observations (in the form of data) to infer or learn things about reality.[1] That's it. There are many ways we might (and do) perform this task.

The statistics part comes in when we incorporate uncertainty and mathematics into the process, using probability to specify a mathematical model of some aspect of reality. Wasserman oversimplifies this by saying that statistical inference is "the process of using data to infer the distribution that generated the data." I would emphasize that the "distribution that generated the data" is itself an object in our model, and inferring it is only useful if, through its relationship to reality, we are able to infer something from it about reality. We'll generally assume that is true, but I can't stress this enough—the model matters.

3. Statistical models and estimation

[1] This assumes that there is such a thing as reality and that it is external to each of us; we won't go down this rabbit hole.

A **statistical model** $\mathcal{F}$ is a set of probability distributions (or densities or conditional distributions). A **parametric model** is a set $\mathcal{F}$ that can be parameterized by a finite-dimensional parameter. Abstractly, this typically looks like

$$\mathcal{F} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\} , \tag{5.1}$$

where Θ is the **parameter space**. An example is the family of univariate normal PDFs with unknown mean and variance parameters $\theta = (\mu, \sigma^2)$. Then $\Theta = \mathbb{R} \times (0, \infty)$. If we are only interested in some components of θ, the remaining parameters are called **nuisance parameters**. For example, if we are only interested in μ, then σ^2 is a nuisance parameter.

A **nonparametric model** is one that cannot be parameterized by a finite number of parameters. Often, it will consist of a set of functions, such as $\mathcal{F}_{\text{CDF}} = \{\text{all CDFs}\}$. Another example is nonparametric density estimations, where the model is the set of PDFs that are not "too wiggly," $\mathcal{F}_{\text{PDF}} \cap \mathcal{F}_{\text{SOB}}$, where

$$\mathcal{F}_{\text{SOB}} = \left\{ f : \int (f''(x))^2 dx < \infty \right\} . \tag{5.2}$$

A **semiparametric model** is one that has a parametric part (often the part we're interested in) and a nonparametric part (often a nuisance "parameter").

Given a set of observations assumed to be sampled from *some* distribution in $\mathcal{F}$, **estimation** is the process of selecting *one* of the distributions $F \in \mathcal{F}$. In many cases, we may only be interested in estimating some property of F, which can be thought of as a function of F, $T(F)$. Some examples:

- The mean, $T(F) = \mathbb{E}(X) = \int x dF(x)$.
- The median, $T(F) = F^{-1}(1/2)$.

4. Prediction

In some cases, we observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and wish to predict Y_i from X_i. Here, X is called the **regressor** or **predictor** (or feature or independent variable). Y is called the **outcome** or **response variable** (or dependent variable). At the most general level, the idea is to model and infer

$$\mathbb{P}_\theta(Y \mid X) . \tag{5.3}$$

It is common to model this with a **regression model**, which consists of a regression function $r(x) = \mathbb{E}(Y|X = x)$, and a noise model. For example, an additive noise model is that

$$Y = r(X) + \epsilon , \tag{5.4}$$

where $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}(\epsilon) = 0$. Note that any regression model can be written in the form (5.4), but that ϵ in general will not be independent from X.

5. A word on notation

If θ parameterizes our statistical model, then $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ denote the corresponding probability distribution and expectation with respect to that distribution, respectively. Similarly for the variance, $\mathbb{V}_\theta$.

6. Basic concepts

Point estimation refers to a single "best guess" of the thing we're trying to estimate. It does not attempt to quantify uncertainty about the accuracy of that guess. We denote an estimator of θ by $\hat{\theta}$ or, if the sample size matters, by $\hat{\theta}_n$. It's important to keep in mind that an estimator is a function of data, which appear in our model as random variables, so $\hat{\theta}$ is also a random variable in the model. For example, if we compute $\hat{\theta}$ with a function g of $X_1, \ldots, X_n$, then

$$\hat{\theta} = g(X_1, \ldots, X_n) . \tag{5.5}$$

The **bias** of an estimator is

$$\mathsf{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta \; . \tag{5.6}$$

An estimator is **unbiased** if its bias equals zero. Unbiasedness is not considered as important as it once was (there are good statistical reasons for this), but consistency is still a reasonable requirement. An estimator $\hat{\theta}_n$ is **consistent** if

$$\hat{\theta}_n \xrightarrow{\text{P}} \theta \; . \tag{5.7}$$

The distribution of $\hat{\theta}_n$ is called the **sampling distribution**. The **standard error** of $\hat{\theta}_n$ is

$$\mathsf{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)} \; , \tag{5.8}$$

which typically needs to be estimated, which we denote by $\hat{\mathsf{se}}$.

Exercise 5.1. Let $X_1, \ldots, X_n \sim_{\text{IID}} \text{Pois}(\lambda)$. Show that $\bar{X}_n$ is an unbiased estimator of λ. What is the standard error? How can it be estimated?

Solution: Proof of unbiasedness:

$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}n\lambda = \lambda$, and so the estimator is unbiased.

Computation of Standard Error:

$\mathsf{se}(\bar{X}_n) = \sqrt{\text{Var}(\bar{X}_n)} = \sqrt{\frac{1}{n^2}\text{Var}(\sum_{i=1}^n X_i)}$, and by independence it is $\sqrt{\frac{1}{n^2}n\lambda} = \sqrt{\frac{\lambda}{n}}$.

Estimation of Standard Error: There are two methods.

First method: Since $\bar{X}_n$ is the estimator of λ, $\mathsf{se}(\bar{X}_n)$ can be estimated by $\sqrt{\frac{\bar{X}_n}{n}}$.

Second method: Since λ is also population variance of Poisson distribution, $\mathsf{se}(\bar{X}_n)$ can be estimated by $\sqrt{\frac{s}{n}}$, where s is the sample standard deviation.

The quality of a point estimate is often assessed by a **loss function** $L\colon \Theta \times \Theta \to \mathbb{R}$. One of the most common loss functions is **mean squared error**, or MSE,

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] \; . \tag{5.9}$$

It is a useful fact that

$$\text{MSE}(\hat{\theta}_n) = \mathsf{bias}^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n) \; . \tag{5.10}$$

Because of this, it's easy to prove the following.

Theorem 5.3. *If the MSE of an estimator $\hat{\theta}_n$ converges to zero as $n \to \infty$, then $\hat{\theta}_n$ is consistent.*

Exercise 5.2. Let $X_1, \ldots, X_n \sim_{\text{IID}} \text{Pois}(\lambda)$. Show that $\bar{X}_n + \frac{1}{n}$ is a biased but consistent estimator of λ.

Solution: Since $E(\bar{X}_n + \frac{1}{n}) = \lambda + \frac{1}{n} \neq \lambda$, it's biased with a bias $= \frac{1}{n}$.

Also, since $\text{Var}(\bar{X}_n + \frac{1}{n}) = \text{Var}(\bar{X}_n) = \frac{\lambda}{n}$, we have $\text{MSE}_\lambda(\bar{X}_n + \frac{1}{n}) = \mathsf{bias}^2(\bar{X}_n + \frac{1}{n}) + \text{Var}(\bar{X}_n) = \frac{1}{n^2} + \frac{\lambda}{n}$,

which converges to 0 as $n \to \infty$.

Therefore, $\bar{X}_n + \frac{1}{n}$ is consistent.

An estimator is **asymptotically normal** if

$$\frac{\hat{\theta}_n - \theta}{\mathsf{se}(\hat{\theta}_n)} \rightsquigarrow \mathcal{N}(0,1) \; . \tag{5.11}$$

Many of the estimators we encounter will be asymptotically normal, which is useful for a number of reasons.

7. Confidence sets

Let $C_n = (a, b)$, where a and b are functions of data, be a **random interval** of $\mathbb{R}$. A $1-\alpha$ **confidence interval** is a random interval C_n such that

$$\mathbb{P}_\theta(\theta \in C_n) = \mathbb{P}_\theta(a(X_1, \ldots, X_n) \leq \theta \leq b(X_1, \ldots, X_n)) \geq 1 - \alpha \; , \tag{5.12}$$

for all $\theta \in \Theta$. Note that in the probability statement above, θ is fixed, and the interval is random. So this is really a probability statement about *our procedure for estimating* C_n: if we were to repeat the experiment (or different experiments with different parameters) and construct a $1-\alpha$ confidence interval for each one, then at least $(1-\alpha) \times 100\%$ of the confidence intervals would contain the "true"[2] parameter. *This is not a probability statement about* θ, which is not random.

If $\theta \in \mathbb{R}^k$ then C_n will be a random subset (e.g., a hyperrectangle or hypersphere) of $\mathbb{R}^k$.

We saw an example of a confidence interval for estimating the parameter of the Bernoulli distribution by inverting a probability bound from Hoeffding's inequality. A more common scenario is that $\hat{\theta}_n$ is asymptotically normal, in which case an approximate $1-\alpha$ confidence interval is

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{\mathsf{se}}, \hat{\theta}_n + z_{\alpha/2}\hat{\mathsf{se}}) \; , \tag{5.13}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2) = \mathbb{P}(Z > \alpha/2)$ is the $\alpha/2$-quantile of the standard normal distribution, $Z \sim \mathcal{N}(0,1)$. Then (assuming that $\hat{\mathsf{se}}$ is a consistent estimator of se) as $n \to \infty$,

$$\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha \; . \tag{5.14}$$

Exercise 5.3. Let $X_1, \ldots, X_n \sim_{\text{IID}} \text{Pois}(\lambda)$. Argue that $(\bar{X}_n - \lambda)/\mathsf{se}(\bar{X}_n)$ is asymptotically normal, and construct an approximate 95% confidence interval for λ.

Solution:

For X_i, $i = 1, \ldots, n$, $E(X_i) = \lambda$ and $\text{Var}(X_i) = \lambda$, thus by the CLT we have $Z_n = \frac{\bar{X}_n - \lambda}{\mathsf{se}(\bar{X}_n)} \rightsquigarrow \mathcal{N}(0,1)$. Therefore, $\bar{X}_n$ is distributed approximately as $\mathcal{N}(\lambda, \frac{\lambda}{n})$, and $\mathsf{se}(\bar{X}_n)$ is estimated by $\sqrt{\frac{\bar{X}_n}{n}}$.

The approximate 95% confidence interval is $(\bar{X}_n - z_{0.975}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{0.975}\sqrt{\frac{\bar{X}_n}{n}})$

Exercise 5.4. For $\lambda = 5$ and each of $n \in \{10, 100, 1000\}$, simulate $m = 1000$ repetitions of the experiment and approximate confidence interval from the previous activity. How accurate is the approximate confidence interval for each n?

Solution: Based on our simulations, the coverage rate is 0.943 when n=10, 0.954 when n=100, and 0.954 when n=1000.

[2] I use quotes here because it's rare that such a thing as a true parameter exists because our models tend to be over-simplifications of whatever it is that we're modeling.

Exercise 5.5. Let $X_1, \ldots, X_n \sim_{\text{IID}} \text{Unif}(0, \theta)$ and let $\hat{\theta}_n = 2\bar{X}_n$. Find the bias, standard error, and MSE of this estimator. Is it consistent?

Solution: $X_1, \ldots, X_n \sim_{\text{IID}} \text{Unif}(0, \theta)$. Then $E(X_i) = \frac{\theta}{2}$, and $\text{Var}(X_i) = \frac{\theta^2}{12}$.

Let $\hat{\theta}_n = 2\bar{X}_n$. Then

$$E_\theta(\hat{\theta}_n) = E_\theta(2\bar{X}_n) = 2E_\theta(\bar{X}_n) = \theta \; . \tag{5.15}$$

So it is unbiased.

For the variance,

$$\text{Var}(\hat{\theta}_n) = 4\text{Var}(\bar{X}_n) = 4\frac{\text{Var}(X_1)}{n} = \frac{\theta^2}{3n} \; . \tag{5.16}$$

Therefore, $\text{MSE}_\theta(2\bar{X}_n) = 0 + \frac{\theta^2}{3n} \to 0$ as $n \to \infty$, so the estimator is consistent.

8. What's next?

Many courses on statistical inference would at this point focus on classical methods of parametric inference, and the well established theory that comes with them. We'll follow Wasserman and first study some very generally applicable nonparametric methods (estimating CDFs, and bootstrap procedures for estimating uncertainty) that have less (and less accessible) theory but that require fewer assumptions.

References

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.