Stat 111 Notes

Elvin Lo

Spring 2024

2024.3.15

Preface

These notes follow the Stat 111 Book by Blitzstein and Shephard, the text accompanying STAT 111 at Harvard College.

Contents

1 Introduction

1.1 Overview

1.2 The big picture

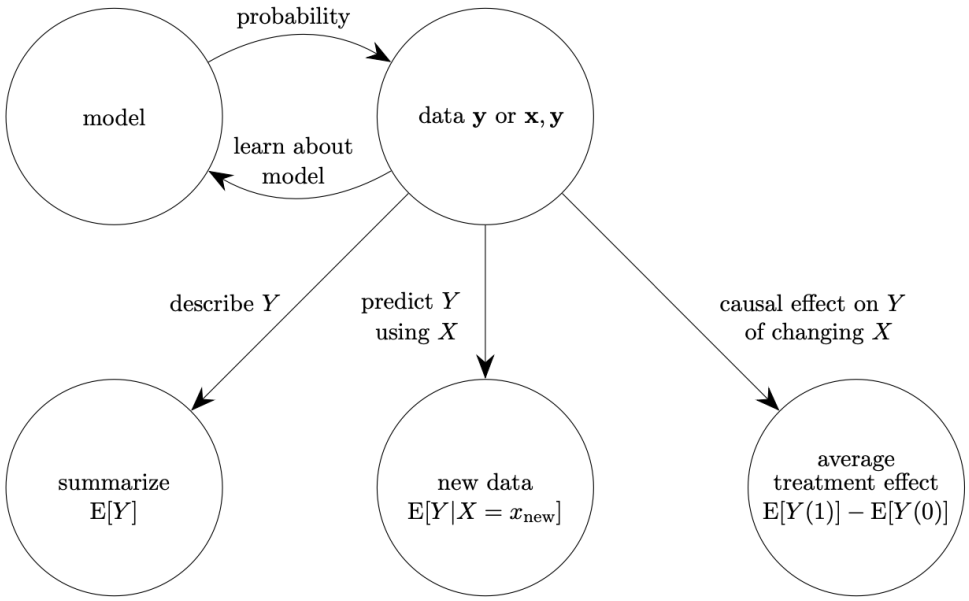

Figure 1.1: Roadmap of the relationships between some of the most fundamental concepts in statistics.

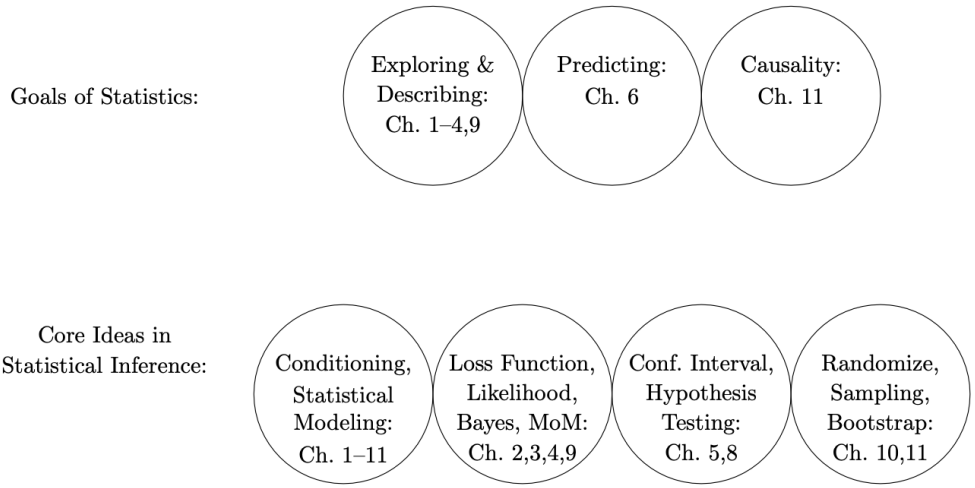

Figure 1.2: The core goals of statistics together with the main tools developed in statistical inference supporting them.

Notation 1.1.
We use capital letters for r.v.s and lowercase letters for the observed data corresponding to those r.v.s. Because we are discussing data in statistics, we generally use the letter Y when we have only one set of r.v.s.

Definition 1.2. Fundamental statistics tasks
Let $Y_1, \ldots, Y_n$ be the r.v.s that will "crystallize" into the observed values $y_1, \ldots, y_n$.

- A *statistical model* is a collection of possible joint distributions for $Y_1, \ldots, Y_n$, possibly indexed by some parameter θ.
- A statistic is a function of the data.
- An *estimand* is a particular quantity that we wish to learn, e.g., the model parameters or the mean of a random variable.
- Based on the model, *probability* lets us determine how likely various events are and what the typical values of our random variables are.
- *Exploring* provides interpretable summaries of the data, either visually or numerically.
- *Describing* goes in the reverse direction to probability, addressing the fact that θ is typically unknown. We can consider questions such as (1) how should we estimate θ given some data, (2) how should we estimate some probability like $P(Y_1 = 0)$, (3) how confident should we be about our estimates, (4) how good are our model's assumptions, and so on. Note that various strategies for estimating an estimand are possible.
- *Prediction* considers how to use observed data to predict not yet observed data.
- *Causality* asks what the effect on a variable will be if we intervene to change another variable.

1.3 Learning and deciding: frequentist and Bayesian inference

Definition 1.3. Learning and deciding
There are two major kinds of tasks for inference of unknown quantities: learning and deciding. Let θ be an estimand, that is an unknown quantity of interest we wish to learn.

1. Learning about θ.
   (a) *Point estimation*: we choose an estimator $\hat{\theta}$, as a function of $Y_1, \ldots, Y_n$. Ideally, we will be able to prove that $\hat{\theta}$ will be close to θ with high probability.
   (b) *Interval estimation*: provide an interval that contains θ with high probability.
2. Deciding about θ. In some applications, the goal is to make a decision,e.g., is it plausible or not that $\theta = 0$, or which of two rival hypotheses $\theta = \theta_0$ and $\theta = \theta_1$ should we choose.

Definition 1.4.
Learning and deciding can be approached from either frequentist or Bayesian perspectives.

- The *frequentist approach* focuses on coming up with procedures that work well in the long run; this requires considering drawing new datasets over and over again.
- The *Bayesian approach* focuses on the data at hand. We model θ as a random variable with some prior distribution, and then use Bayes' rule to update our probabilities for θ based on the data. Once we have the posterior distribution, we might estimate θ using the posterior mean, median, or even create an interval that has, say, 95% chance of containing θ, given the data.

1.4 Exploring and describing Y

We discuss some useful summaries of a dataset $y_1, \ldots, y_n$ (aside from visualizations like histograms and scatter plots).

Definition 1.5. Sample mean, sample standard deviation
The sample mean of $y_1, \ldots, y_n$ is

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j.$$

The sample standard deviation s is the square root of the sample variance

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 .$$

Definition 1.6. Sample covariance, sample correlation, linear regression
The sample covariance is

$$s_{x,y} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}) (y_j - \bar{y})$$

and the sample correlation is

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}.$$

The linear regression of y on x is

$$b_{y \sim x} = \frac{s_{x,y}}{s_x^2},$$

where s_x and s_y are the sample standard deviations of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively.

Definition 1.7. Order statistics
The order statistics of $y_1, y_2, \ldots, y_n$ are the same data points, sorted in increasing order:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

Some examples of quantities based on order statistics are the sample minimum $y_{(1)}$, the sample maximum $y_{(n)}$, the range $y_{(n)} - y_{(1)}$, and the sample median $y_{((n+1)/2)}$ (if n is odd; there are different conventions about what to do if n is even).

Definition 1.8. Quantile function
Let F be a CDF. The quantile function of F is defined by

$$F^{-1}(p) = Q(p) = \min\{y : F(y) \geq p\}.$$

The value $Q(p)$ is called the p-quantile of the distribution. For a random variable Y, we will use Q_Y to denote the quantile function of Y. Note that for a general CDF F (that may not be continuous or invertible), we have

$$F\left(F^{-1}(p)\right) \geq p.$$

Definition 1.9. Sample quantile
The corresponding sample quantity is the p-sample quantile, defined to be a value such that approximately proportion p of the sample is less than or equal to that value. The p-sample quantile of the dataset $y_1, \ldots, y_n$ is the order statistic $y_{(\lceil np \rceil)}$, denoted by $\hat{Q}(p)$:

$$\hat{Q}(p) = y_{(\lceil np \rceil)}.$$

There are different conventions for the sample quantile, though if n is large then it is unlikely to matter which convention is used. We will use a simple convention.

Definition 1.10. Empirical CDF
The empirical CDF (ECDF) of the data set is the CDF of a r.v. obtained by choosing one of the n data points $y_1, \ldots, y_n$ uniformly at random,

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I\left(y_j \leq y\right).$$

Note that the ECDF is always a step function, jumping every time it reaches one of the data points.

Note that the ECDF converges to the true CDF as the sample size grows. If we have a model under which the data are realizations of i.i.d. r.v.s $Y_1, \ldots, Y_n$ with CDF F, then the strong LLN implies that for each $y \in \mathbb{R}$, we have with probability 1 that

$$\lim_{n \to \infty} \hat{F}(y) = E\left[I\left(Y_1 \leq y\right)\right] = P\left(Y_1 \leq y\right) = F(y).$$

1.5 Predicting Y from X

Remark 1.11. What is prediction?
Think of it is as: having seen some data X, what are the likely values of Y? *Forecasting* is a special case of prediction where X is an aspect of the past and Y is an aspect of the future.

Definition 1.12. Regression models
In a regression model, we have predictor variables $X_1, \ldots, X_k$ and an outcome variable Y, and we try to use the predictor variables to predict the outcome variable. We can think of

$$E\left[Y \mid X_1, \ldots, X_k\right]$$

as the best prediction of Y as a function of $X_1, \ldots, X_k$.

1.6 Causal impact on Y of manipulating X

To help differentiate prediction and causality, let us establish some notation for thinking about causal effects.

In causal studies, X is called the assignment and Y is the outcome. As a simple example, consider a binary treatment with binary outcomes. Then we have $X \in \{0, 1\}$ where the event $X = 1$ is receiving the treatment and the event $X = 0$ is the control, and we might write the outcome Y as 1 for a success and 0 otherwise.

Definition 1.13. Potential outcomes, treatment effect
The pair $\{Y(0), Y(1)\}$ are called the potential outcomes. The random variable

$$\tau = Y(1) - Y(0)$$

is the treatment effect (or causal effect) on the outcome of moving a person from control to treatment. Across a population, moving everyone from control to treatment, the

$$\mathrm{E}[\tau] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$$

is the population's average treatment effect.

Definition 1.14. Counterfactual
A major challenge is that even after the study is over we never see both $Y(0)$ and $Y(1)$: we only see one of them as the individual is either under treatment or control, not both. So we cannot compute τ directly, which makes it tricky to estimate $\mathrm{E}[\tau]$.

To infer $\mathrm{E}[\tau]$, we have the assignment X and an outcome, mathematically written as

$$\begin{aligned} Y &= \begin{cases} Y(1), & \text{if } X = 1 \\ Y(0), & \text{if } X = 0 \end{cases} \\ &= Y(X) \\ &= XY(1) + (1 - X)Y(0). \end{aligned}$$

The potential outcome we do not see, $Y(1 - X)$, is called a counterfactual. It is impossible to observe a counterfactual.

2 Models, Likelihood, Estimation, and Method of Moments

2.1 Statistical models

Definition 2.1. Statistical model
A statistical model views $\mathbf{y}$ as a realization of the random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from their joint cumulative distribution function (CDF) $F_{\mathbf{Y}}$. The model specifies a collection of possibilities for $F_{\mathbf{Y}}$. Before making the observation, we have a random vector $\mathbf{Y}$. After making the observation, $\mathbf{Y}$ crystallizes into the data $\mathbf{y}$. We say that the model generated the data, and often we want to use the data to learn about the model.

Definition 2.2. Estimand
An estimand is an aspect of $F_{\mathbf{Y}}$ that we wish to learn about from the data that we will observe.

Estimands are often denoted by Greek letters.

Definition 2.3. Parametric model
A parametric statistical model is a family of probability distributions for $\mathbf{Y}$, indexed by a finite-dimensional parameter θ. The distributions in a model are usually specified by their joint CDFs or by their joint densities (joint PMFs in the discrete case, joint PDFs in the continuous case). The parameter space, denoted by Θ, is the set of all allowable values of θ. Thus each $\theta \in \Theta$ picks out a single probability distribution for $\mathbf{Y}$.

If in the above definition we instead allow θ to be infinite-dimensional, then we have a non-parametric model.

Notation 2.4. CDF notation
To make explicit the fact that the CDF in a parametric model depends on θ, we sometimes write the CDF as $F_{\mathbf{Y};\theta}$. Then the CDF evaluated at $\mathbf{y}$ is denoted by $F_{\mathbf{Y};\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y};\theta)$. If θ is being modeled as a random variable, as in the Bayesian approach, then ; is replaced by the conditioning bar | in the notation, yielding $F_{\mathbf{Y}|\theta}$ for the CDF (which technically is, from a Bayesian point of view, the conditional CDF of $\mathbf{Y}$ given θ). Then the CDF evaluated at $\mathbf{y}$ is written as $F_{\mathbf{Y}|\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y} \mid \theta)$.

Example 2.5. Data $Y_1, \ldots, Y_n$ are i.i.d.
Often it is scientifically plausible to assume that $Y_1, \ldots, Y_n$ are independent and identically distributed. Under the i.i.d. assumption,

$$Y_j \overset{\text{i.i.d.}}{\sim} F_{Y_1;\theta}, \quad j = 1, \ldots, n,$$

where $F_{Y_1;\theta}$ is the CDF of each Y_j. By independence, the joint CDF is the product of the marginal CDFs:

$$F_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{j=1}^{n} F_{Y_1}(y_j;\theta).$$

2.2 Likelihood

Definition 2.6. Likelihood function
Let $\mathbf{y}$ be the observed value of $\mathbf{Y}$. The function given by

$$L(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \theta),$$

regarded as a function of the parameter, with the data held constant, is called the likelihood function. That is, the likelihood function is the probability or probability density of the data given the parameters, as a function of the parameters. So the likelihood function is a function of θ, with $\mathbf{y}$ treated as fixed.

Notationally, it is conventional to separate θ and y by a semicolon: $L(\theta; \mathbf{y})$. Often we even write $L(\theta)$ for the likelihood function, leaving the $\mathbf{y}$ implicit, to simplify the notation further and to emphasize that the likelihood function is regarded as a function of θ. Also note that two likelihood functions are viewed as equivalent if one is a positive constant times the other. In fact, the "constant" can even be a function of the data (it just cannot depend on the parameter)!

Remark 2.7. Bayesian vs frequentist perspectives
Bayesian perspective: In a Bayesian approach, we have a prior density $\pi(\theta)$ for θ, and use Bayes' rule to obtain the posterior density:

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\theta) f(\mathbf{y} \mid \theta)}{f(\mathbf{y})} \propto \pi(\theta) f(\mathbf{y} \mid \theta) = L(\theta; \mathbf{y}) \pi(\theta),$$

where the proportionality stems from the fact that we are treating $\mathbf{y}$ as fixed (so the denominator $f(\mathbf{y})$, which is the marginal density of $\mathbf{y}$, is viewed as a constant). That is, Bayes' rule says, in words: The posterior is proportional to likelihood times prior.
So the two key ingredients for a Bayesian analysis are the prior distribution $\pi(\theta)$ and the likelihood function $L(\theta; \mathbf{y})$. Combining the likelihood and the prior, we obtain the posterior distribution, which we then base our inferences on.

Frequentist perspective: In a frequentist approach, θ does not have a posterior distribution, but we can use the likelihood function as a surrogate for assessing how plausible various possible values of θ are. One of the most widely used estimation techniques in statistics is maximum likelihood estimation, which says to estimate θ using

$$\hat{\theta} = \arg\max_{\theta} L(\theta; \mathbf{y}),$$

the parameter value that maximizes the likelihood function. This value is called the maximum likelihood estimate (MLE).

Definition 2.8. Log-likelihood
It is very common when working with likelihood to work with the log-likelihood.

$$L(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y}).$$

One reason is numerical stability, since extremely small probabilities often come up in likelihood calculations. But also, the log lets us consider sums rather than products, and since log is a continuous, strictly increasing function, maximizing the likelihood is equivalent to maximizing the log-likelihood (for when we study MLE later). If the Y_j are independent, then the likelihood function is

$$L(\theta; \mathbf{y}) = \prod_{j=1}^{n} f_{Y_j}(y_j; \theta)$$

and the log-likelihood function is

$$l(\theta; \mathbf{y}) = \sum_{j=1}^{n} \log f_{Y_j}(y_j; \theta).$$

Theorem 2.9. Invariance of likelihood under transformation of the parameter
The likelihood function is unchanged under reparameterization, in the following sense. Consider a likelihood function $L(\theta; \mathbf{y})$ and let $\psi = g(\theta)$ be a reparameterization, where g is a known one-to-one function. Then

$$L(\psi; \mathbf{y}) = L(\theta; \mathbf{y}).$$

Theorem 2.10. Invariance of likelihood under transformation of the data
Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the observed data, coming from a model with parameter θ. Let h be a known one-to-one function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Use h to transform the data, letting $\mathbf{x} = h(\mathbf{y})$. Then taking the dataset to be $\mathbf{x}$ rather than y has no effect on the likelihood function:

$$L(\theta; \mathbf{x}) = L(\theta; \mathbf{y})$$

Proof. For simplicity, we will only write the proof in the case of a single observation y from a continuous distribution, with h a differentiable, strictly increasing function. Let Y be the r.v. that "crystallizes" to $y, x = h(y)$, and $X = h(Y)$. By the change of variables formula,

$$L(\theta; x) = f_X(x; \theta) = f_Y(y; \theta)\frac{1}{h'(y)}.$$

But $\frac{1}{h'(y)}$ is a multiplicative "constant" (not depending on θ), so it can be dropped. Then we can take the likelihood function for θ, based on the data x, to be

$$L(\theta; x) = f_Y(y; \theta) = L(\theta; y)$$

□

2.3 Statistics, estimators, and estimates

Definition 2.11. Statistic
A statistic is a function of $Y_1, \ldots, Y_n$ (and possibly other known quantities). We can write a statistic as $T(\mathbf{Y})$, where computing the function T must not require knowledge of any unknown parameters.

Definition 2.12. Estimator
Suppose that we use the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ to construct a statistic

$$\widehat{\theta} = T(\mathbf{Y})$$

with the intention that this statistic should estimate an estimand θ. The statistic $\widehat{\theta}$ is called an estimator.

Definition 2.13. Bias
The bias of an estimator $\widehat{\theta}$ for θ is

$$\text{bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta.$$

We say that $\widehat{\theta}$ is unbiased for θ if its bias is 0, i.e., its expected value is θ.
To compute the bias, recall that by LOTUS,

$$\text{E}[\widehat{\theta}] = \int T(\mathbf{y}) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

if $\mathbf{Y}$ is continuous, and

$$\text{E}[\widehat{\theta}] = \sum_{\mathbf{y}} T(\mathbf{y}) P(\mathbf{Y} = \mathbf{y}; \theta),$$

if $\mathbf{Y}$ is discrete, where the integral and sum are over the support of $\mathbf{Y}$. Typically, the bias depends on θ (so we may be able to compute the bias theoretically but may not know the actual value of the bias due to θ being unknown).

Definition 2.14. Standard error
The standard error of an estimator $\hat{\theta}$ for θ is its standard deviation:

$$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

This is a measure of how variable the estimator is.

Lemma 2.15. Sum of squares identity
For any random variables $Y_1, \ldots, Y_n$ and any constant c,

$$\sum_{j=1}^{n} (Y_j - c)^2 = n(\bar{Y} - c)^2 + \sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2.$$

Definition 2.16. Estimate
An estimate is a realization of an estimator. So if our data $\mathbf{y}$ is a realization of $\mathbf{Y}$ and $T(\mathbf{Y})$ is an estimator of some estimand θ, then $T(\mathbf{y})$ is an estimate of θ.

2.4 Sample moments and method of moments

Definition 2.17. Method of moments
Consider the following strategy, the method of moments principle, for obtaining an estimator. Express the estimand in terms of moments, and then replace the estimand with the estimator and the theoretical moments with the sample moments. An estimator obtained in this way is called a method of moments estimator (MoM).

3 Loss Functions, Bias-Variance Tradeoff, and Asymptotics

3.1 Bias and variance of sample p-quantiles

Theorem 3.1. Asymptotic distribution of sample quantile
If $Y_1, \ldots, Y_n$ be i.i.d. Uniform random variables on $[0,1]$ and $p \in (0,1)$. Then as $n \to \infty$,

$$\sqrt{n}\left\{Y_{(\lceil np \rceil)} - Q_{Y_1}(p)\right\} \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Proof. The proof of this theorem uses the change of variables formula and Taylor approximations. The details of the proof get pretty technical, so the proof is in the starred Section 3.8. □

3.2 Bias and variance are sometimes in conflict

Definition 3.2. Kernel density estimator
Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with a CDF $F_{Y_1}(y)$ and PDF $f_{Y_1}(y)$. Let the estimand be $\theta = f_{Y_1}(y)$ for some value of y, and let $h > 0$. Let

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} I\left(Y_j \in (y - h/2, y + h/2]\right).$$

The estimator $\hat{\theta}$ is called the kernel density estimator (KDE) of $f_{Y_1}(y)$, with rectangular kernel. The number h is called a bandwidth. More generally, let K be a nonnegative function, called the kernel function. Then the kernel density estimator for this choice of K is

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} K\left(\frac{Y_j - y}{h}\right).$$

A widely used kernel aside from the rectangular kernel is to let K be the standard Normal PDF. In this book, by default when discussing KDE we will assume we are using the rectangular kernel. In terms of the ECDF, we can write the KDE as

$$\hat{\theta} = \frac{\hat{F}(y + h/2) - \hat{F}(y - h/2)}{h}.$$

Taking the limit as $h \to 0$ on the right-hand side would yield the definition of the derivative of $\hat{F}$ at y. As noted earlier, the derivative of the ECDF is not useful, but here we fix h rather than letting $h \to 0$. Thus, the KDE can be thought of as a way to get an approximate notion of "slope" for the ECDF without actually taking the derivative.

3.3 Loss functions, risk, and mean square error

Definition 3.3. Loss function
A loss function is a function

$$\text{Loss}(\theta, \hat{\theta})$$

interpreted as the loss or cost associated with using the estimate $\hat{\theta}$ when the estimand is θ. We require $\text{Loss}(\theta, \hat{\theta}) \geq 0$ and $\text{Loss}(\theta, \theta) = 0$.

We take a frequentist approach for now, treating θ as an unknown constant (not having a distribution). The estimator $\hat{\theta} = T(\mathbf{Y})$ is, of course, a random variable. So $\text{Loss}(\theta, \hat{\theta})$ is also a random variable, as it is a function of $\hat{\theta}$ (for each fixed θ).

Definition 3.4. Risk function
For the estimator $\hat{\theta} = T(\mathbf{Y})$, the risk function is the expected loss

$$\text{Risk}(\theta) = \text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = \int \text{Loss}(\theta, T(\mathbf{y})) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

where the integral is over the support of $\mathbf{Y}$.

Definition 3.5. MSE
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

is called the squared error loss. Its expected value is the mean square error (MSE) of $\hat{\theta}$ is:

$$\text{MSE}(\hat{\theta}) = \text{E}_\theta \left[(\hat{\theta} - \theta)^2 \right].$$

Sometimes the square root of the MSE is used instead, in order to have the units be the same as that of θ; this is called the root mean square error (RMSE).

Definition 3.6. MAE
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

is called the absolute error loss. Its expected value is the mean absolute error (MAE) of $\hat{\theta}$ is

$$\text{MAE}(\hat{\theta}) = \text{E}_\theta[|\hat{\theta} - \theta|].$$

Absolute error loss punishes small errors more severely and large errors more gently than the RMSE. Which one of the two makes more sense in practice depends, of course, on the application.

Definition 3.7. $0-1$ **loss**
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = I(\hat{\theta} \neq \theta)$$

is called $0-1$ loss. Note that its expected value is $\text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = P_\theta(\hat{\theta} \neq \theta)$.

Definition 3.8. Check loss
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta| \left\{ c_+ I(\hat{\theta} > \theta) + c_- I(\hat{\theta} < \theta) \right\}, \quad c_+, c_- \geq 0,$$

is called the check loss. Unlike the previous few, this last loss function is not symmetric.

3.4 Bias-Variance tradeoff

Theorem 3.9. Bias-variance tradeoff
The mean square error of an estimator $\hat{\theta}$ for the estimand θ is the variance plus the square of the bias:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2.$$

Proof. Let $V = \hat{\theta} - \theta$. Then

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{E}_\theta \left[V^2 \right] \\ &= \text{Var}_\theta(V) + (\text{E}_\theta[V])^2 \\ &= \text{Var}(\hat{\theta}) + \{\text{E}(\hat{\theta}) - \theta\}^2 \\ &= \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2. \end{aligned}$$

□

3.5 Consistency of estimators

Definition 3.10. Consistency
An estimator $\hat{\theta}$ is consistent for the estimand θ if $\hat{\theta}$ converges in probability to θ as the sample size $n \to \infty$, i.e., for every $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) \to 0$$

as $n \to \infty$. This is written in shorthand as

$$\hat{\theta} \xrightarrow{p} \theta.$$

Note that implicitly $\hat{\theta}$ depends on n. Sometimes it is clearer to write the dependence explicitly, with notation such as $\hat{\theta}_n$ that explicitly indicates the sample size n.

Theorem 3.11. Sufficient condition for consistency
If $\hat{\theta}$ is an estimator for the estimand θ and $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent. In particular, since MSE is variance plus squared bias, to show that $\hat{\theta}$ is consistent it suffices to show that both the bias and the variance go to 0 as $n \to \infty$.

Proof. Suppose that the MSE of $\hat{\theta}$ goes to 0 . Then by Markov's inequality, for any $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) = P\left((\hat{\theta} - \theta)^2 \geq \epsilon^2\right) \leq \frac{\mathrm{E}\left[(\hat{\theta} - \theta)^2\right]}{\epsilon^2} = \frac{\text{MSE}(\hat{\theta})}{\epsilon^2} \to 0$$

□

Remark 3.12.
It may be tempting to believe that $\hat{\theta} \xrightarrow{p} \theta$ implies that $\mathrm{E}[\hat{\theta}] \to \theta$, since if two r.v.s are very likely to be very close to each other, then it may seem intuitively that their means should also be close. But this implication is false. For a counterexample, let the data be realizations of independent r.v.s $U, Y_1, \ldots, Y_n$ where $U \sim \text{Unif}(0, 1)$ and $Y_j \sim \text{Bern}(\theta)$. Let

$$\hat{\theta} = \bar{Y} + nI(U \leq 1/n).$$

Then $\hat{\theta} \xrightarrow{p} \theta$, since $\bar{Y} \xrightarrow{p} \theta$ and for n large, the second term in $\hat{\theta}$ is very likely to be 0 . But

$$\mathrm{E}[\hat{\theta}] = \mathrm{E}[\bar{Y}] + nP(U \leq 1/n) = \theta + 1$$

Example 3.13.
As we mentioned earlier, many estimators have $\text{bias}(\hat{\theta}) \approx b/n$ and $\text{Var}(\hat{\theta}) \approx a/n$, which would mean that $\text{MSE}(\hat{\theta}) \approx a/n$. Other estimators have slower rates at which $\text{MSE}(\hat{\theta})$ converges to zero as n gets large. Table 3.1 gives some core examples of this for some descriptive statistics.

Statistics	Approximate MSE	Consistent?
sample mean	$\text{Var}(Y_1)/n$	Yes
sample variance	$\text{Var}\left[\{Y_1 - \mathrm{E}[Y_1]\}^2\right]/n$	Yes
sample covariance	$\text{Var}[\{X_1 - \mathrm{E}[X_1]\}\{Y_1 - \mathrm{E}[Y_1]\}]/n$	Yes
sample p-quantile for uniform	$p(1-p)/n$	Yes
kernel density estimator	$d/n^{4/5}$	Yes
Y_1	$\text{Var}(Y_1)$	No
$\frac{1}{2}(Y_1 + Y_2)$	$\text{Var}(Y_1)/2$	No

Theorem 3.14. Continuous mapping theorem
Let $X, X_1, X_2, \ldots$ be a sequence of r.v.s and let g be a continuous function. If

$$X_n \xrightarrow{p} X$$

then

$$g(X_n) \xrightarrow{p} g(X).$$

Also, if

$$X_n \xrightarrow{d} X,$$

then

$$g(X_n) \xrightarrow{d} g(X).$$

In particular, it follows that if $\hat{\theta}$ is a consistent estimator for θ and g is a continuous function, then $g(\hat{\theta})$ is a consistent estimator for $g(\theta)$.

Theorem 3.15. Properties of convergence in probability
Let $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then

$$X_n + Y_n \xrightarrow{p} X + Y,$$
$$X_n - Y_n \xrightarrow{p} X - Y,$$
$$X_n Y_n \xrightarrow{p} XY,$$

and, if $P(Y_n = 0) = P(Y = 0) = 0$,

$$X_n / Y_n \xrightarrow{p} X/Y.$$

3.6 Large sample (asymptotic) approximations

Definition 3.16. Convergence in distribution
Let $X_1, X_2, \ldots$ be a sequence of random variables and F_{X_n} be the CDF of X_n. Let X be a random variable with CDF F_X. Then X_n converges in distribution to the random variable if

$$F_{X_n}(x) \to F_X(x),$$

for all $x \in \mathbb{R}$ such that F_X is continuous at x. This is written in shorthand as $X_n \xrightarrow{d} X$.

Theorem 3.17.
Let $X_n \xrightarrow{p} X$. Then $X_n \xrightarrow{d} X$. The converse is false in general. But if X is a constant c (i.e., X is a degenerate r.v. that always equals c), then $X_n \xrightarrow{p} X$ is equivalent to $X_n \xrightarrow{d} X$.

Remark 3.18.
Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. It does not follow that $X_n + Y_n \xrightarrow{d} X + Y$. As a simple counterexample, let $X_n = Y_n \sim \mathcal{N}(0,1)$ and X, Y be i.i.d. $\mathcal{N}(0,1)$. Then $X_n + Y_n = 2X_n \sim \mathcal{N}(0,4)$, whereas $X + Y \sim \mathcal{N}(0,2)$. Clearly, the $\mathcal{N}(0,4)$ distribution does not converge to $\mathcal{N}(0,2)$. The problem is that the statement $X_n \xrightarrow{d} X$ is about the marginal distributions of X_n and X, and similarly for the statement $Y_n \xrightarrow{d} Y$, whereas the distribution of $X_n + Y_n$ depends heavily on the joint distribution of X_n and Y_n.

Theorem 3.19. Slutsky's Theorem
If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n - Y_n \xrightarrow{d} X - c$;
- $X_n Y_n \xrightarrow{d} cX$;
- $X_n / Y_n \xrightarrow{d} X/c$, if $c \neq 0$.

Theorem 3.20. Delta method
Suppose that g is a differentiable function and

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \omega^2\right).$$

Then

$$\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} \xrightarrow{d} \mathcal{N}\left(0, \left(g'(\theta)\right)^2 \omega^2\right).$$

As an approximation, this says that

$$g(\hat{\theta}) \dot{\sim} \left(g(\theta), \left(g'(\theta)\right)^2 \frac{\omega^2}{n}\right),$$

for n large.

Proof. If n is large, then $\hat{\theta}$ is close to θ (with high probability). Taylor expand $g(\hat{\theta})$ about θ, yielding the approximation

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

as the higher order terms should be smaller as they involve squares, cubes, etc. of $\{\hat{\theta} - \theta\}$ which is going to zero at rate $n^{-1/2}$. Rearranging,

$$\begin{aligned}\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} &\approx g'(\theta)\sqrt{n}\{\hat{\theta} - \theta\} \\ &\xrightarrow{d} g'(\theta)\omega Z,\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. □

3.7 Multivariate asymptotic approximations*

3.8 A couple of technical proofs*

3.9 Concentration inequalities

4 Maximum Likelihood Estimation

4.1 Defining and finding the maximum likelihood estimate (MLE)

Definition 4.1. Maximum likelihood estimator
The maximum likelihood estimate (MLE) of θ is the value $\hat{\theta}$ that maximizes the likelihood function $L(\theta; \mathbf{y})$. Mathematically, this is written as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{y})$$

4.2 Properties of the MLE

Some of the main properties that the MLE enjoys are as follows. All of these require some technical assumptions known as regularity conditions. Under these assumptions, which we will discuss more later, we have the following.

- The MLE is invariant, which means that if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- The MLE $\hat{\theta}$ is consistent, which means that it converges in probability to the true θ.
- The MLE is asymptotically Normal (so its distribution is approximately Normal if the sample size is large).
- The MLE is asymptotically unbiased (the bias approaches 0 as the sample size grows).
- The MLE is asymptotically efficient (no other asymptotically unbiased estimator will have a lower standard error asymptotically).

Theorem 4.2. Invariance of MLE
Let $\hat{\theta}$ be the MLE of θ, and let g be a one-to-one function. Then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Proof. This result follows from the invariance property of likelihood. Let $\tau = g(\theta)$. Each point on the reparameterized likelihood curve $L(\tau)$ has a corresponding point on the original likelihood function $L(\theta)$, such that the likelihood value for τ is the same as the likelihood value for the corresponding θ. In particular, the value $\hat{\tau}$ that maximizes $L(\tau)$ is $\hat{\tau} = g(\hat{\theta})$. □

Definition 4.3. MLE under a parameter transformation that is not one-to-one
Invariance of the MLE is so convenient, in fact, that we define it to be true even when g is not one-to-one. If $\hat{\theta}$ is the MLE of θ and g is not a one-to-one function, then we define the MLE of $g(\theta)$ to be $g(\hat{\theta})$.

4.3 Kullback-Leibler divergence

Notation 4.4.
Let θ^* be the estimand, such that the random variables $Y_1, \ldots, Y_n$ are generated by the joint CDF $F_{\mathbf{Y};\theta^*}$. The hope, of course, is that our estimators will be close to θ^*. The distribution $F_{\mathbf{Y};\theta^*}$ is called the data generating process. To summarize our notation:

- $\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta; \mathbf{Y})$ is an estimator,
- θ is the argument in the likelihood function $L(\theta; \mathbf{Y})$, and
- θ^* is the true value or estimand, generating $\mathbf{Y}$ through $F_{\mathbf{Y};\theta^*}$.

Definition 4.5. KL divergence
The Kullback-Leibler divergence (KL divergence) from the CDF F to the CDF G is

$$D_{\mathrm{KL}}(F\|G) = \mathrm{E}\left(\log \frac{f(\mathbf{Y})}{g(\mathbf{Y})}\right) = \mathrm{E}[\log f(\mathbf{Y}) - \log g(\mathbf{Y})] = \int \{\log f(\mathbf{y}) - \log g(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y},$$

where f and g are the PDFs corresponding to F and G (in the case of discrete distributions, we replace the PDFs with PMFs). The expectations are computed with $\mathbf{Y}$ generated according to F, not G.

An important case of KL divergence is when $F = F_{\mathbf{Y};\theta^*}$ and $F = F_{\mathbf{Y};\theta}$. Then

$$D_{\mathrm{KL}}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = \mathrm{E}\left[\log L\left(\theta^*; \mathbf{Y}\right) - \log L(\theta; \mathbf{Y})\right].$$

Intuitively, $D_{\mathrm{KL}}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right)$ measures how much the expected log-likelihood is higher at θ^* than at θ, computing the expectation under the true distribution $F_{\mathbf{Y};\theta^*}$.

Theorem 4.6. Additivity of KL divergence
If we observe independent $Y_1, \ldots, Y_n$, then

$$D_{KL}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = \sum_{j=1}^{n} D_{KL}\left(F_{Y_j;\theta^*} || F_{Y_j;\theta}\right),$$

where $D_{KL}\left(F_{Y_j;\theta^*} \| F_{Y_j;\theta}\right)$ is the Kullback-Leibler divergence for the j th observation:

$$D_{KL}\left(F_{Y_j;\theta^*} \| F_{Y_j;\theta}\right) = \mathrm{E}\left(\log \frac{L\left(\theta^*; Y_j\right)}{L\left(\theta; Y_j\right)}\right)$$

In the i.i.d. case,

$$D_{KL}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = n D_{KL}\left(F_{Y_1;\theta^*} \| F_{Y_1;\theta}\right).$$

Theorem 4.7. Nonnegativity of KL divergence
For any CDFs F and G,

$$D_{KL}(F\|G) \geq 0.$$

This inequality is strict unless $F = G$, that is they are the same distribution functions.

Theorem 4.8. Consistency of MLE
Suppose that the parameter space is finite and that the observations $Y_1, \ldots, Y_n$ are i.i.d. Also assume that for $\theta_1 \neq \theta_2$, the distribution function $F_{Y;\theta_1}$ is different from the distribution of $F_{Y;\theta_2}$ (this is known as identifiability of the model). Then the MLE $\hat{\theta}$ is consistent:

$$\hat{\theta} \xrightarrow{p} \theta^*$$

as the sample size $n \to \infty$.

Score function

Definition 4.9. Score function
The score function is

$$s(\theta; \mathbf{y}) = \frac{\partial \log L(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{L(\theta; \mathbf{y})} \frac{\partial L(\theta; \mathbf{y})}{\partial \theta}$$

Theorem 4.10. Information equality
Under some regularity conditions (mainly that the $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ is a smooth function in θ, the support of $\mathbf{Y}$ does not depend on θ, that the expected values needed below exist, and that we can differentiate under the integral sign when needed below),

$$\begin{aligned} \mathrm{E}\left[s\left(\theta^*; \mathbf{Y}\right)\right] &= 0, \\ \mathrm{Var}\left\{s\left(\theta^*; \mathbf{Y}\right)\right\} &= -\mathrm{E}\left[s'\left(\theta^*; \mathbf{Y}\right)\right]. \end{aligned}$$

The prime in s' denotes taking the partial derivative with respect to θ, that is $s'\left(\theta^*; \mathbf{Y}\right) = \partial s\left(\theta^*; \mathbf{Y}\right) / \partial \theta$.

Fisher information

Definition 4.11. Fisher information
The Fisher information in the sample for a parameter θ in a parametric statistical model $F_{\mathbf{Y};\theta}$ is

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathrm{Var}\left\{s(\theta^*;\mathbf{Y})\right\} = \mathrm{E}\left[s(\theta^*;\mathbf{Y})^2\right],$$

where we compute the variance under the assumption that the true parameter value is θ. Let

$$\mathcal{J}_{\mathbf{Y}}(\theta^*) = -\mathrm{E}\left[s'(\theta^*;\mathbf{Y})\right].$$

Then

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathcal{J}_{\mathbf{Y}}(\theta^*),$$

the information equality. In statistics it is traditional to suppress θ^* and write $\mathcal{I}_{\mathbf{Y}}(\theta)$ and $\mathcal{J}_{\mathbf{Y}}(\theta)$, without the stars, implicitly understanding the role of θ^*.

It is not obvious from looking at the definition why Fisher information is a measure of information. Some intuition for this can be gleaned from thinking about the curvature of $\mathrm{E}\log L(\theta;\mathbf{Y})$. If the expected log-likelihood function has a sharp peak at θ^*, the data can be very informative about θ. If the $\mathrm{E}\log L(\theta;\mathbf{Y})$ is quite flat at θ^*, the data do not seem to be giving us much information that we can use for pinpointing the true parameter value.

Definition 4.12. Fisher information when transforming the parameter
Let $\tau = g(\theta)$, where g is a differentiable function with $g'(\theta) \neq 0$. Then

$$\mathcal{I}_{\mathbf{Y}}(\tau) = \frac{\mathcal{I}_{\mathbf{Y}}(\theta)}{\{g'(\theta)\}^2}.$$

Cramér-Rao lower bound

Theorem 4.13. CRLB
Let $\hat{\theta}$ be an unbiased estimator of θ in a parametric statistical model $F_{\mathbf{Y};\theta}$. Under regularity conditions,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta)}.$$

REVIEW PROOFS IN THIS SECTION

Asymptotic distribution of the MLE

Theorem 4.14. Asymptotic distribution of the MLE
In addition to the good properties we have already seen, such as invariance and consistency, the MLE has excellent asymptotic properties: for large sample size, it is approximately the case that the MLE is Normal, unbiased, and achieves the CRLB. Again we use the θ^* notation at the start, to make the exposition clear.

Let $\hat{\theta}$ be the MLE of a scalar parameter θ, based on i.i.d. observations $Y_1, \ldots, Y_n$ from $F_{\mathbf{Y};\theta^*}$. Under regularity conditions, the asymptotic distribution of $\hat{\theta}$ is given by the following:

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{Y_1}^{-1}(\theta^*)\right),$$

(that is converges in distribution) as the sample size $n \to \infty$. As an approximation, this result says that for large n,

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta^*, \frac{1}{n\mathcal{I}_{Y_1}(\theta^*)}\right)$$

4.4 Likelihoods based on conditional distributions

Suppose the data segments into pieces called $\mathbf{x}$ and $\mathbf{y}$ (e.g., $\mathbf{y}$ are outcomes and $\mathbf{x}$ are predictors), then the statistical model is the joint distribution $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$. Of course, the joint density equals the marginal density times the conditional density:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}).$$

Statisticians often decide to study solely the conditional distribution, either out of convenience (we will see an example of this at the end of this section) or because their scientific focus is on the conditional distribution (our next example). In parametric models the statistical model becomes:

$$f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}; \theta).$$

Then we can define the likelihood for this conditional density

$$L(\theta; \mathbf{y} \mid \mathbf{x}) = f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}; \theta).$$

4.5 Numerical optimization of the likelihood*

4.6 Multiple parameter version*

4.7 Estimation when model approximates the truth*

5 Confidence Intervals

5.1 Introduction

Definition 5.1. Interval estimation
An interval estimate $C(\mathbf{y})$ of a scalar estimand θ based on data $\mathbf{y}$ is an interval $[L(\mathbf{y}), U(\mathbf{y})]$, where the lower bound $L(\mathbf{y})$ and upper bound $U(\mathbf{y})$ are functions of the data, such that $L(\mathbf{y}) \leq U(\mathbf{y})$ for all $\mathbf{y}$. The corresponding random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$, where $\mathbf{Y}$ are the random vectors that give rise to the data, is called an interval estimator, written $C(\mathbf{Y})$. Intuitively, the goal is that the probability should be high that $C(\mathbf{Y})$ contains the estimand.

Definition 5.2. Coverage
Let $C(\mathbf{Y})$ be an interval estimator for θ and $C(\mathbf{y})$ be the corresponding interval estimate. If $\theta \in C(\mathbf{y})$, we say that the interval covers θ. The probability of the interval estimator covering θ if θ is the true estimand, $P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y}))$, is called the coverage probability of the interval estimator. Note that the coverage probability is a function of θ.

Definition 5.3. Confidence Interval
Fix a number α with $0 < \alpha < 1$. (In practice, the most common choice is $\alpha = 0.05$.) The interval estimator $C(\mathbf{Y}) = [L(\mathbf{Y}), U(\mathbf{Y})]$ is a $(1-\alpha)$ confidence interval (CI) if it has coverage probability $1-\alpha$ for all possible values of θ:

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha.$$

The constant $1-\alpha$ is called the level of the confidence interval. The half-width $0.5\{U(\mathbf{Y}) - L(\mathbf{Y})\}$ is called the margin of error.

Remark 5.4.
Confidence intervals are widely misinterpreted. For example, if the 95% CI $[0.1, 0.4]$ for θ is calculated from the data, a common mistake is to say that we can be 95% confident that θ is between 0.1 and 0.4 , or that the probability is 0.95 that θ is between 0.1 and 0.4. This is a category error since the statement "θ is between 0.1 and 0.4" is deterministic, either true or false: we are currently working in a frequentist setting, so θ is fixed. It is the interval estimator that is random here, not the estimand.

5.2 Constructing confidence intervals

Example 5.5. Ideal case scenario: a Normal estimator
We wish to create a $1-\alpha$ confidence interval for θ. Suppose that we are in the happy situation where $\hat{\theta}$ is Normal. Specifically, let

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \sigma^2\right),$$

with σ^2 known. A simple but powerful approach is then to standardize $\hat{\theta}$ to get a standard Normal random variable:

$$\frac{\hat{\theta}-\theta}{\sigma} \sim \mathcal{N}(0,1).$$

Then

$$P\left(a \leq \frac{\hat{\theta}-\theta}{\sigma} \leq b\right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a),$$

and so letting

$$c_p = Q_{\mathcal{N}(0,1)}(1-p),$$

we derive that

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\sigma, \hat{\theta} + c_{\alpha/2}\sigma\right] = \hat{\theta} \pm c_{\alpha/2}\sigma$$

is a $1-\alpha$ CI for θ with margin of error $c_{\alpha/2}\sigma$, centered at the MLE with $c_{\alpha/2}$ standard errors of slack in each direction.

Definition 5.6. Pivot
A pivotal quantity or pivot is a random variable whose distribution is known. In contrast to a statistic,

- A pivot typically depends on unknown parameters but its distribution cannot depend on unknown parameters.
- A statistic cannot depend on unknown parameters but its distribution typically depends on unknown parameters.

For example, suppose Y_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Then $\bar{Y}$ statistic, but

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$$

is a pivot.

5.3 Asymptotic approximations

Definition 5.7. Approximate Pivot
Suppose that n is large and, based on asymptotics, we know that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

Then

$$\frac{\hat{\theta} - \theta}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

and the expression on the left-hand side is called an approximate pivot
When we have an approximate pivot, we can obtain an approximate CI by proceeding as in Example 5.2.1 (with approximate equalities in place of equalities). It follows that if σ is known then the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\sigma/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\sigma/\sqrt{n}\right]$$

is an approximate $1 - \alpha$ CI. Usually in practice σ is unknown, but can be estimated with some consistent estimator $\hat{\sigma}$. Then it is natural to use the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\hat{\sigma}/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\hat{\sigma}/\sqrt{n}\right]$$

though it is not obvious what effect plugging in $\hat{\sigma}$ for σ has on the coverage probability for fixed n; this may need to be studied via simulation. Asymptotically this substitution is fine though, since if

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

and $\hat{\sigma} \xrightarrow{p} \sigma$, then by the continuous mapping theorem and Slutsky's theorem,

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}}\right) = \sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right)_{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Remark 5.8.
The asymptotic approach only gives approximate confidence intervals. The mathematical statement is that, if $C(\mathbf{Y})$ is the interval estimator, then

$$\lim_{n\to\infty} P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha$$

as the sample size $n \to \infty$. This does not say, for a fixed n, how close the coverage probabilities are to $1 - \alpha$.

Confusingly, people often say "confidence interval" when they mean "approximate confidence interval". Some so-called $1 - \alpha$ confidence intervals are just aspirational, and in reality the coverage probability is far from $1 - \alpha$ for at least some possible values of θ. In such situations it is clearer to call $1 - \alpha$ the nominal level of the interval estimator. The hope is that the coverage probabilities will be close to $1 - \alpha$ for all θ but this may not be true, or it may be true but not yet demonstrated.

5.4 Pivots with non-Gaussian distributions

6 Regression

6.1 Regression

6.2 Predictive regression

Definition 6.1. Predictive regression
The task of estimating the conditional expectation

$$\mu(\mathbf{x}) = \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

is called predictive regression. The variable Y is called the outcome variable, while the $\mathbf{X}$ variables are called predictors, covariates, regressors, or features. Running a statistical method in this setting is sometimes called regressing Y on $\mathbf{X}$.

Definition 6.2. Homoskedasticity and heteroskedasticity
Assume that

$$\sigma^2(\mathbf{x}) = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$$

exists. If $\sigma^2(\mathbf{x})$ does not vary with $\mathbf{x}$, then the predictive regression is called homoskedastic. Otherwise it is heteroskedastic.

Definition 6.3. Regression error
For predictive regression, the regression error is the random variable

$$U(\mathbf{x}) = Y - \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

Theorem 6.4. Signal-noise decomposition
From our definitions, we may decompose Y into signal (the predicted part $\mu(\mathbf{x})$) and noise (the random error $U(\mathbf{x})$):

$$Y = \mu(\mathbf{x}) + U(\mathbf{x}).$$

Broadly speaking, a lot of statistical work is about trying to separate signal from noise.

Theorem 6.5. Regression error: mean 0, uncorrelated with predictors
For a random pair $(\mathbf{X}, Y)$, write the regression error (for $\mathbf{X}$ random) as

$$U(\mathbf{X}) = Y - \mathrm{E}[Y \mid \mathbf{X}].$$

Then

$$\mathrm{E}[U(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}] = 0,$$
$$\mathrm{E}[U(\mathbf{X})] = 0,$$

and for each predictor variable X_j,

$$\mathrm{Cov}\left(U(\mathbf{X}), X_j\right) = 0.$$

Proof. By construction,

$$\mathrm{E}[U(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}] = \mathrm{E}[U(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}] = \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] - \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] = 0,$$

for all $\mathbf{x}$. By Adam's law, we also have $\mathrm{E}[U(\mathbf{X})] = 0$ unconditionally. Again by Adam's law, as long as the covariance exists, for each predictor variable X_j we have

$$\mathrm{Cov}\left(U(\mathbf{X}), X_j\right) = \mathrm{E}\left[X_j U(\mathbf{X})\right] = \mathrm{E}\left[\mathrm{E}\left[X_j U(\mathbf{X}) \mid \mathbf{X}\right]\right] = \mathrm{E}\left[X_j \mathrm{E}[U(\mathbf{X}) \mid \mathbf{X}]\right] = \mathrm{E}\left[0 X_j\right] = 0.$$

□

Proposition 6.6. Variance of Y
It is also important to consider the variance of Y, both conditionally and unconditionally. Recall $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$. Then $\sigma^2(\mathbf{x}) = \mathrm{Var}(U \mid \mathbf{X} = \mathbf{x})$, so by Eve's law,

$$\mathrm{Var}(U) = \mathrm{E}[\mathrm{Var}(U \mid \mathbf{X} = \mathbf{x})] + \mathrm{Var}(\mathrm{E}[U \mid \mathbf{X} = \mathbf{x}]) = \mathrm{E}\left[\sigma^2(\mathbf{X})\right],$$

if this variance exists. Likewise, Eve's law says

$$\mathrm{Var}(Y) = \mathrm{E}\left[\sigma^2(\mathbf{X})\right] + \mathrm{Var}(\mu(\mathbf{X})),$$

so a summary measure of the unconditional effectiveness of the prediction is

$$R^2 = \frac{\mathrm{Var}(\mu(\mathbf{X}))}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)}$$

the share of the variation of Y contributed by the variation in the prediction.

Definition 6.7. Linear regression model

$$\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K,$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)^{\mathrm{T}}$. This is called a linear regression model and the elements of $\boldsymbol{\theta}$ are called the regression coefficients. Typically, θ_0 is called the intercept and $\theta_1, \ldots, \theta_K$ are called slopes. Note that linear regression means linear in the parameters θ; the function can be nonlinear in the predictors.

Definition 6.8. Logit function
The logit function is defined by

$$\mathrm{logit}(p) = \log(p/(1-p)),$$

for $0 < p < 1$. The inverse logit function, which is also called the logistic function, sigmoid function, or expit function, is the inverse of the logit function:

$$\mathrm{logit}^{-1}(x) = \frac{e^x}{1 + e^x},$$

for all $x \in \mathbb{R}$.

Definition 6.9. Logistic regression
The logistic regression model assumes that the probability of success, given the predictor variables, is

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \mu(\mathbf{x} \mid \boldsymbol{\theta}) = \text{logit}^{-1}\left(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K\right)$$

6.3 Statistical models of predictive regression

Gaussian linear regression without intercept

The Gaussian linear regression model assumes the scatter around $\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta})$ is Gaussian. For now consider a single predictor and no intercept,

$$Y_j \mid (X_1 = x_1, \ldots, X_n = x_n), \theta \overset{\text{ind.}}{\sim} \mathcal{N}\left(\theta x_j, \sigma^2\right).$$

Our parameter has MLE

$$\hat{\theta} = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j y_j\right).$$

Definition 6.10. Residuals
The value $\hat{\theta} x_j$ is called the fitted value or predicted value of Y_j. The difference between the actual value of y_j and the predicted value is called the residual, and denoted by

$$\hat{U}_j = y_j - x_j \hat{\theta}.$$

Theorem 6.11. Residuals are orthogonal to predictors
With notation as above, we have

$$\sum_{j=1}^{n} x_j \hat{U}_j = 0.$$

Proof. Note that

$$\sum_{j=1}^{n} x_j^2 \hat{\theta} = \hat{\theta} \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j$$

so

$$\sum_{j=1}^{n} x_j \hat{U}_j = \sum_{j=1}^{n} \left(x_j y_j - x_j^2 \hat{\theta}\right) = \sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j^2 \hat{\theta} = 0.$$

□

Theorem 6.12. Properties of the least squares estimator
Assume that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ have conditionally independent outcomes. Inference will condition the random variables $\mathbf{X}$ at the observed values $\mathbf{x} = (x_1, \ldots, x_n)$. Write $\mu_j = \mathrm{E}[Y_j \mid X_j = x_j]$ and $\sigma_j^2 = \mathrm{Var}(Y_j \mid X_j = x_j)$. Then

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j \mu_j\right), \quad \mathrm{Var}(\hat{\theta} \mid \mathbf{X} = \mathbf{x}) = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \left(\sum_{j=1}^{n} x_j^2 \sigma_j^2\right).$$

Proof. Conditioning on the predictors, $\hat{\theta}$ is linear in the outcomes, so (as expectations of sums are sums of expectations)

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \sum_{j=1}^{n} x_j \mathrm{E}[Y_j \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j \mu_j\right).$$

Conditioning on the predictors $\hat{\theta}$ is linear in the conditionally independent outcomes, so

$$\mathrm{Var}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \sum_{j=1}^{n} x_j^2 \mathrm{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \left(\sum_{j=1}^{n} x_j^2 \sigma_j^2\right).$$

□

Lemma 6.13. More properties
Properties of Gaussian linear regression model under successively stronger conditions.

(a) (a) If $\mu_j = \theta x_j$, then

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \theta$$

so then the estimator is conditionally unbiased for θ.

(b) Under homoskedasticity, i.e., $\sigma_j^2 = \sigma^2$,

$$\mathrm{Var}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \sigma^2 \left(\sum_{j=1}^{n} x_j^2\right)^{-1}$$

(c) If $Y_j \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\theta X_j, \sigma^2\right)$, then $\hat{\theta}$ is conditionally unbiased for θ and conditionally achieves the Cramér-Rao lower bound

$$\mathrm{Var}(\hat{\theta} \mid \mathbf{X} = \mathbf{x}) = \mathcal{I}(\theta)^{-1}$$

Furthermore,

$$\hat{\theta} \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\theta, \sigma^2 \left(\sum_{j=1}^{n} x_j^2\right)^{-1}\right).$$

Proof. Special case of Theorem 6.3.5. □

Gaussian linear regression with intercept

Now consider that we have an intercept, yielding the model

$$Y \mid (X = x, \theta_0, \theta_1) \sim \mathcal{N}\left(\theta_0 + \theta_1 x, \sigma^2\right).$$

We may derive the MLE of the parameters

$$\hat{\theta}_1 = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

6.4 Linear regression, method of moments, and least squares

6.5 Linear projection and descriptive regression

Definition 6.14. Linear projection
Assume that the random variables X, Y each have a finite variance. Then the linear projection of Y on X at $X = x$ is defined as

$$\mu_L(x) = \mathrm{E}[Y] + \beta_{Y \sim X}(x - \mathrm{E}[X]).$$

The linear projection $\mu_L(x)$ is not the conditional expectation $\mathrm{E}[Y \mid X = x]$ in general. The conditional expectation $\mathrm{E}[Y \mid X = x]$ is the function of x that best approximates Y, in the sense of minimizing the expected square error; the linear projection $\mu_L(X)$ is the best linear function of x for approximating Y. Writing the linear error as

$$U_L = Y - \mu_L(X),$$

then by construction $\mathrm{E}[U_L] = 0$ and $\mathrm{E}[XU_L] = 0$, due to the derivatives.

7 Exponential Families and Sufficiency

7.1 Natural Exponential Families

Definition 7.1. Natural exponential families (NEFs)
A density $f(y;\theta)$ follows a natural exponential family (NEF) if we can write

$$f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$$

where the nonnegative function h does not depend on θ. The parameter θ is called the natural parameter and may be a reparameterization of how the model was originally specified.
Another way to think of a density of this form is that it factors as a function of y (not involving θ) times a function of θ (not involving y) times a function of both y and θ, where the function of both y and θ takes the simple form $e^{\theta y}$.

Theorem 7.2. Mean and variance in an NEF
Let Y follow the NEF $f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$. Then

$$\mathrm{E}[Y] = \psi'(\theta), \quad \mathrm{Var}(Y) = \psi''(\theta).$$

Proof. Consider that Y is continuous; the discrete case is analogous. Densities must integrate to 1, so

$$\int_{-\infty}^{\infty} e^{\theta y} h(y) dy = e^{\psi(\theta)}.$$

By DUThIS, differentiating both sides with θ, we have

$$\int_{-\infty}^{\infty} y e^{\theta y} h(y) dy = \psi'(\theta) e^{\psi(\theta)}.$$

Therefore,

$$\psi'(\theta) = \int_{-\infty}^{\infty} y e^{\theta y - \psi(\theta)} h(y) dy = \mathrm{E}[Y],$$

by the definition of expectation. For the variance, we can DUThIS again:

$$\psi''(\theta) = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \left(y e^{\theta y - \psi(\theta)} h(y) \right) dy = \int_{-\infty}^{\infty} y \left(y - \psi'(\theta) \right) e^{\theta y - \psi(\theta)} h(y) dy.$$

Let $\mu = \mathrm{E}[Y] = \psi'(\theta)$. By LOTUS, we then have

$$\psi''(\theta) = \mathrm{E}[Y(Y-\mu)] = \mathrm{E}\left[Y^2\right] - \mu \mathrm{E}[Y] = \mathrm{E}\left[Y^2\right] - \mu^2 = \mathrm{Var}(Y),$$

as desired. □

Theorem 7.3. MLE of an NEF
Suppose we have data $Y_1, \ldots, Y_n$ i.i.d. from the NEF $f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$. Let $\mu = \mathrm{E}[Y_1] = \psi'(\theta)$ be the mean parameter (it is a reparameterization of θ). Then

- The MLE of μ is its MoM estimator:

$$\hat{\mu} = \bar{Y}$$
$$\hat{\theta} = \left(\psi'\right)^{-1}(\bar{Y})$$

- The Fisher information per observation is

$$\mathcal{I}_{Y_1}(\theta) = \psi''(\theta) = \mathrm{Var}(Y_1)$$
$$\mathcal{I}_{Y_1}(\mu) = \left(\frac{\partial\theta}{\partial\mu}\right)^2 \mathcal{I}_{Y_1}(\theta) = \frac{1}{\psi''(\theta)} = \frac{1}{\mathrm{Var}(Y_1)}$$

- The asymptotic distribution of the MLE is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \psi''(\theta)^{-1}\right)$$
$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \psi''(\theta)\right).$$

- The MLE of μ achieves the CRLB (with equality, not just asymptotically).

Proof. The log-likelihood function is

$$l(\theta; \mathbf{y}) = n\{\theta\bar{y} - \psi(\theta)\},$$

and the score function is

$$s(\theta; \mathbf{y}) = n\left\{\bar{y} - \psi'(\theta)\right\}.$$

Setting the score equal to 0 , we have that the MLE of θ is as claimed. To check that we have found the maximum, use the second derivative test:

$$\frac{\partial}{\partial\theta} s(\theta; y) = -n\psi''(\theta) < 0,$$

since $\psi''(\theta) = \mathrm{Var}(Y) > 0$. So the log-likelihood function is concave and the Fisher information is as stated in the theorem by the information equality. The MLE is unique as the function ψ' has an inverse since it is continuous (because it is differentiable) and strictly increasing (because $\psi''(\theta) = \mathrm{Var}(Y) > 0$). By invariance, the MLE of $\mu = \psi'(\theta)$ is

$$\hat{\mu} = \psi'(\hat{\theta}) = \psi'\left(\left(\psi'\right)^{-1}(\bar{Y})\right) = \bar{Y}.$$

The asymptotic properties are implied by the standard MLE properties using the Fisher information. To show that $\hat{\mu} = \bar{Y}$ achieves the CRLB, note that $\hat{\mu}$ is unbiased, with

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(Y_1) = \frac{1}{n\mathcal{I}_{Y_1}(\mu)}.$$

□

Definition 7.4. Exponential family (EF)
A density $f(y;\theta)$ follows an exponential family (EF) if we can write

$$f(y;\theta) = e^{\theta T(y) - \psi(\theta)} g(y),$$

where g does not depend on θ. A generalization of NEFs, an EF is obtained by transforming the variable in a natural exponential family. The difference between an NEF and an EF is that the observation appears as itself in the exponent in an NEF, whereas it appears in some transformed form in the exponent in an EF. If $T(y) = y$ then we recover the definition of an NEF.

7.2 Sufficient statistics

Definition 7.5. Sufficient statistic
For $Y_1, \ldots, Y_n$ from the parametric statistical model $F_{\mathbf{Y};\boldsymbol{\theta}}$, a statistic $T(\mathbf{Y})$ is sufficient for θ if the conditional distribution of

$$(Y_1, \ldots, Y_n) \mid T$$

does not depend on θ.

Here the conditional distribution of $\mathbf{Y}$ given T does not involve θ, so once we know T there is no further statistical information to be gained about θ from looking at the entire vector $(Y_1, \ldots, Y_n)$.

- Sufficient statistics are not unique, for any one-to-one transformation preserves sufficiency.
- A sufficient statistic of the smallest dimension possible is called a minimal sufficient statistic.

Theorem 7.6. Factorization criterion
When computing the conditional distribution $\mathbf{Y}$ given T is difficult, we have a simpler criterion. For θ in $F_{\mathbf{Y};\boldsymbol{\theta}}$, the statistic T is sufficient iff we can factor

$$f(\mathbf{y};\theta) = g_\theta(t)h(\mathbf{y}),$$

where t is the observed value of T and the function h does not depend on θ.

Proof. We will prove the factorization criterion in the discrete case. The continuous case is analogous but more technical to prove. Suppose that T is sufficient. Then

$$f(\mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y}, T = t;\theta),$$

since T is a deterministic function of $\mathbf{Y}$. So

$$f(\mathbf{y};\theta) = P(T = t;\theta)P(\mathbf{Y} = \mathbf{y} \mid T = t) = g_\theta(t)h(\mathbf{y}),$$

where $g_\theta(t) = P(T = t;\theta)$ and $h(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} \mid T = t)$. This is a valid choice for h since it does not depend on θ (since the conditional distribution of $\mathbf{Y}$ given T does not involve θ) and since t is a deterministic function of $\mathbf{y}$. Conversely, suppose that

$$f(\mathbf{y};\theta) = g_\theta(t)h(\mathbf{y}).$$

Let $T = s(\mathbf{Y})$. The conditional PMF of $\mathbf{Y} \mid T$ is

$$P(\mathbf{Y} = \mathbf{y} \mid T = t; \theta) = \frac{P(\mathbf{Y} = \mathbf{y}, T = t; \theta)}{P(T = t; \theta)} = \frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{P(T = t; \theta)}$$

for $t = s(\mathbf{y})$, and 0 otherwise. We need to show that this conditional distribution does not depend on θ. To do so, we can expand the denominator based on all possible values of $\mathbf{Y}$ that are compatible with the observed t :

$$\frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{P(T = t; \theta)} = \frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} P(\mathbf{Y} = \mathbf{y}; \theta)} = \frac{g_\theta(t) h(\mathbf{y})}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} g_\theta(t) h(\tilde{\mathbf{y}})} = \frac{h(\mathbf{y})}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} h(\tilde{\mathbf{y}})},$$

which does not depend on θ. Thus, T is a sufficient statistic for θ. □

Theorem 7.7. Likelihood function based on a sufficient statistic
Let t be a sufficient statistic for the model $f(\mathbf{y}; \theta)$. Then the likelihood function can be expressed as a function of t (up to a multiplicative constant that does not depend on θ). In particular, knowing the sufficient statistic suffices for knowing the likelihood function, and we can take the $g(\theta; t)$ appearing in the factorization criterion as our likelihood function.

Proof. Note that if t is sufficient then, with notation as in the factorization criterion,

$$\begin{aligned} L(\theta; \mathbf{y}) &= f(\mathbf{y}; \theta) \\ &= g(\theta; t) h(\mathbf{y}), \end{aligned}$$

where $h(\mathbf{y})$ does not depend on θ. Since for likelihood purposes we can drop multiplicative constants (including functions of the data), we can take $g(\theta; t)$ as our likelihood function. In this function, the data only appears through t. □

Corollary 7.8. MLE depends on data only through sufficient statistic
Suppose that T is a sufficient statistic for θ and that we have factored the likelihood function as

$$L(\theta; \mathbf{y}) = c g(\theta; t).$$

Then the MLE depends upon the data only through T, and the Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \operatorname{Var}\left[\frac{\partial \log g(\theta; T)}{\partial \theta}\right].$$

Corollary 7.9. Posterior depends on data only through sufficient statistic
Suppose that T is a sufficient statistic for θ. Then the posterior for $\theta, \pi(\theta \mid \mathbf{y})$, depends upon the data only through T :

$$\pi(\theta \mid \mathbf{y}) = \pi(\theta \mid t)$$

which depends on the data only through t.

Proof. Let $\pi(\theta)$ be the prior density. Then the posterior density is

$$\pi(\theta \mid \mathbf{y}) \propto g(\theta; t) \pi(\theta).$$

The data enter the right-hand side only through t, and normalizing $g(\theta; t)\pi(\theta)$ so that it integrates to 1 (with respect to θ) does not introduce any additional involvement of the data. □

Theorem 7.10. Rao-Blackwell
Let $\hat{\theta}$ be an estimator for θ and T be a sufficient statistic for θ for the parametric model $F_{\mathbf{Y};\theta}$. Then the Rao-Blackwellized estimator

$$\hat{\theta}_{\mathrm{RB}} = \mathrm{E}[\hat{\theta} \mid T]$$

is better than or equal to $\hat{\theta}$ in MSE. The estimator $\hat{\theta}_{\mathrm{RB}}$ is strictly better than $\hat{\theta}$ in MSE unless $\hat{\theta}$ is already a deterministic function of T.

Proof. The result follows from Adam's law and Eve's law. By Adam's law,

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{RB}}\right] = \mathrm{E}[\mathrm{E}[\hat{\theta} \mid T]] = \mathrm{E}[\hat{\theta}]$$

so the bias of $\hat{\theta}_{\mathrm{RB}}$ is the same as that of $\hat{\theta}$. Therefore, if $\hat{\theta}_{\mathrm{RB}}$ has lower variance than $\hat{\theta}$, then it also has lower MSE. To compare the variances, we can use Eve's law:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}[\mathrm{Var}(\hat{\theta} \mid T)] + \mathrm{Var}(\mathrm{E}[\hat{\theta} \mid T]) = \mathrm{E}[\mathrm{Var}(\hat{\theta} \mid T)] + \mathrm{Var}\left(\hat{\theta}_{\mathrm{RB}}\right) \geq \mathrm{Var}\left(\hat{\theta}_{\mathrm{RB}}\right),$$

with strict inequality unless $\mathrm{Var}(\hat{\theta} \mid T) = 0$ with probability 1 . If $\mathrm{Var}(\hat{\theta} \mid T) = 0$ then, given T, the estimator $\hat{\theta}$ is constant, which means that $\hat{\theta}$ is a deterministic function of T. □

8 Hypothesis Testing