Frequentist Inference

Muchang Bahng

Winter 2022

Contents

In statistics, we are given some data $\mathcal{D} = \{x_i\}_{i=1}^n$. The simplest thing we can do is summarize this data by extracting some nice characteristics—for example, the mean. This is known as **descriptive statistics**.

In **inferential statistics**, we have much stronger assumptions. We assume that that data are realizations of random variables following a joint probability distribution. Sometimes, we may assume that these **samples** are iid coming from $\mathbb{P}^*$, known as the *true data generating distribution* (and sometimes known as the *population* in survey statistics or causal inference). As the name suggests, we must infer from $\mathcal{D}$ what $\mathbb{P}$ is. This immediately raises some questions: How should we interpret the population? What are we inferring? And how does this process work? Let's establish this confusion with an example.

Example 0.1 (Measurement Problem)

Say we have a dataset consisting of real-valued measurements $x_1, \ldots, x_n$ to estimate some quantity θ. We may try to summarize the mean of this data by computing

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{1}$$

This seems so common and intuitive that we might forget why this specific formula works. Two nice properties are:

1. It minimizes the sum of least squares
$$\bar{x} = \operatorname*{argmin}_a \sum_{i=1}^n (x_i - a)^2 \tag{2}$$
2. The value $\bar{x}$ makes the sum of the residuals to be 0.

These two properties land on the level of descriptive statistics. They describe the mean as a reasonable descriptive measure of the center of the observations, but they cannot justify $\bar{x}$ as an estimate of the true value θ since no explicit assumption has been made connecting the observations x_i with θ.
To do inference, we can furthermore assume that the x_i are observed values of n independent random variables which have a common distribution depending on θ. Which assumptions we make will determine which estimators are reasonable. Here are two cases in which means are not a reasonable estimate.

1. We assume that $x_i = \theta + \epsilon_i$ where ϵ_i satisfies $\mathbb{P}(\epsilon_i < 0) = \mathbb{P}(\epsilon_i > 0)$.
2. *Larger samples may not improve estimate.* If the x_i turns out to have finite variance the variance of the mean is σ^2/n. However, if the x_i's have a Cauchy distribution, then the distribution of $\bar{x}$ is the same as x, so nothing is gained by taking more measurements.

To answer the first question, the population is usually introduced as some finite true distribution of some quantity, but more often it is treated as an abstract data generating distribution. For example, say that we have a large barrel of grains, and we take a random sample of 100 grains and measure their weight. Though we can spend much more effort and time weighing every single grain in the barrel, for practical reasons we want to work with the sample. On the other hand, think of the distribution of facial features of humanity. We may assume that every time a human is born, we can think of it being sampled from some abstract distribution (specified by "God"), and so even taking all humans in the world is still a sample of this population.

As for the second and third questions, the general statistical procedure (both frequentist and Bayesian) goes like this.[1]

1. There exists some true distribution $\mathbb{P}^*$ of the population over some **sample space** $\mathcal{X}$. We want to estimate some (population) **parameter** $\theta^* = T(\mathbb{P}^*)$ for some map T. This parameter of interest θ^* is called the **estimand**.

   There are many types of parameters we can choose from, and generally the form of T can change quite

[1] Thanks to Ed Tam for giving me such a clear bird's eye view.

dramatically depending on the type of inference we are doing.

(a) In **point estimation**, we are interested in vector-valued estimates. For example, the mean $T(\mathbb{P}) = \int x d\mathbb{P}(x)$ is one such candidate.

(b) In **hypothesis testing**, we are given a null hypothesis H_0, and $T(\mathbb{P}) = \mathbb{1}_{H_0}(\mathbb{P}) \in \{0, 1\}$. T is therefore a binary function that indicates whether the null hypothesis is true.

(c) In **confidence sets**, T is still a point estimate. However, instead of estimating θ^* directly, we construct a set-valued estimator $C_n(\mathcal{D}) \subset \Theta$ designed to contain θ^* with high probability. This is useful for uncertainty quantification.

(d) In **density estimation**, $T(\mathbb{P}) = \mathbb{P}$, and we must try to find the distribution P_θ in our model that best fits $\mathbb{P}^*$.

2. We are given a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ that is sampled from $\mathbb{P}^*$. Often, we assume that this is iid.

3. We specify a **model**, which is simply a family of distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$.[2] If $\Theta \subset \mathbb{R}^k$, then we say $\mathcal{P}$ is a parameteric family.

4. We derive an **estimator**, which is a function δ that maps your data to the space that the estimand lives in. The actual value of the function evaluated on your dataset $\hat{\theta} = \delta(\mathcal{D})$ is called the **estimate**. There are several **principles** that guide us into deriving such an estimator.[3]

(a) **Maximum Likelihood** estimators attempt to maximize the likelihood L of the data given a model, and so

$$\hat{\theta} = \delta_{MLE}(\mathcal{D}) = \operatorname*{argmax}_\theta L(\theta \mid \mathcal{D}) \tag{3}$$

(b) **Method of Moments** estimators try to match the moments of the model with that of the data.

(c) **Score matching** attempt to match the score function—the derivative of the log-likelihood.

(d) **M-estimators**.

(e) **Minimax estimators**.

Next, we talk about model selection, focusing on the cross validation and information criteria.

Definition 0.1 (Empirical Distribution)

Now given that we have these iid samples, we can construct the **empirical distribution** $\widehat{X} \sim \widehat{P}$, defined as the discrete distribution that assigns probability $1/n$ to each value x_i for $i \in [n]$. In other words, we have

$$\mathbb{P}(\widehat{X} = x) = \frac{1}{n} \text{ for } x \in \{x_1, \ldots, x_n\} \tag{4}$$

We can write the CDF of the empirical distribution, called the **empirical distribution function**, as the sum of indicators

$$F_X(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[x_i, +\infty)}(x) \tag{5}$$

As expected, we would expect the empirical distribution to converge to the actual distribution.

[2] We may also say a family of functions, but models is more fundamental.

[3] But none of these principles are frequentist in nature. The key is that frequentism is a way to view probabilities and to view decisions. In other words, it is a way to evaluate things. Principles a to e are often principles that frequentist statisticians use to derive estimators, mainly because such estimators enjoy nice frequentist properties. But none of the methods per se are frequentist in nature. abcd can technically all be lumped under the umbrella of m estimators (although that level of generality and abstractness is seldom useful beyond proof purposes).

Theorem 0.1 (Glivenko–Cantelli theorem)

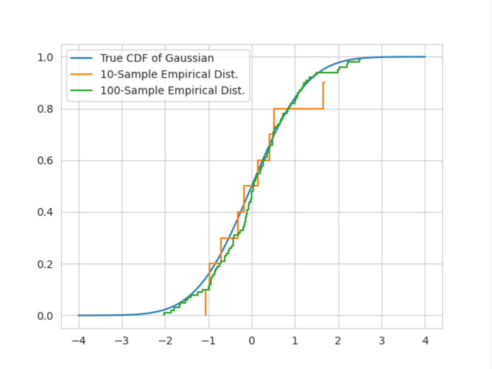
The empirical distribution of iid samples $x_1, \ldots, x_n \sim P_n$ converges almost surely to $X \sim P$ as $n \rightarrow \infty$. More specifically, given that the CDF of X is F and the CDF of P_n is the step function F_n, we have

$$||F_n - F||_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \rightarrow 0 \tag{6}$$

almost surely as $n \rightarrow \infty$.

Example 0.2 (Empirical Distribution of Standard Gaussian)

We expect the empirical distribution of the standard Gaussian to converge. Indeed, numerical results show that for 10 and 100 samples, the empirical CDF does converge to the true CDF.



1 Point Estimation

Let's start with the most fundamental type of estimation. In point-estimation, we are interested in vector-valued estimates. We would like to construct an estimator δ that approximates the true parameter.

$$\delta(\mathcal{D}) \approx \theta^* \tag{7}$$

Definition 1.1 (Sampling Distribution)

The probability distribution of the random variable $\delta(\mathcal{D})$ is called the **sampling distribution**.

1. The standard deviation of $\delta(\mathcal{D})$ is the **standard error**.

The θ^* may be treated as fixed (in frequentist regime) or random (Bayesian). So we must determine what makes an approximation good or bad. Two such measures are

$$\mathbb{P}(\|\theta^* - \delta(\mathcal{D})\| < c) \tag{8}$$

or

$$\mathbb{E}\big[\|\theta^* - \delta(\mathcal{D})\|^p\big] \tag{9}$$

We would like the sampling distribution of our statistic to give us good estimate in two ways. $\delta(\mathcal{D})$ should not be too far off from the actual parameter θ^* (bias is small), and $\delta(\mathcal{D})$ should not fluctuate too widely (variance of θ^* should be small).

Definition 1.2 (Bias on an Estimator)

The **bias** of an estimator is

$$\text{Bias}(\delta) = \big|\theta^* - \mathbb{E}_{\mathcal{D}}[\delta(\mathcal{D})]\big| \tag{10}$$

Definition 1.3 (Variance of an Estimator)

The **variance** of an estimator is

$$\text{Var}(\delta) = \mathbb{E}\big[(\theta^* - \mathbb{E}_{\mathcal{D}}[\delta(\mathcal{D})])^2\big] \tag{11}$$

1.1 Common Estimators

Definition 1.4 (Sample Mean)

The **sample mean** is the estimator

$$\delta(\mathcal{D}) = \frac{x_1 + \ldots + x_n}{n} \tag{12}$$

used to estimate the population mean.

Note that this may or may not be a good estimator, depending on the distribution which we assume the x_i's are coming from, whether they are independent, or other things.

Lemma 1.1 (Mean)

The mean of $\overline{x}_n$ is μ.

$$\mu_{\overline{x}_n} = \mu \tag{13}$$

Proof.

$$\mathbb{E}[\overline{x}_n] = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n x_i\bigg] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i] = \mathbb{E}[x] = \mu \tag{14}$$

Lemma 1.2 (Variance)

If the variance is finite and the samples are iid, then the variance of $\overline{x}_n$ is σ^2/n, i.e. the standard error of $\overline{x}_n$ is $\sigma_{\overline{x}_n} = \sigma/\sqrt{n}$.

$$\sigma_{\overline{x}_n} = \frac{\sigma}{\sqrt{n}} \tag{15}$$

because

$$\sigma^2_{\overline{x}_n} = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(x_i) = \frac{1}{n}\text{Var}(x) = \frac{\sigma^2}{n} \tag{16}$$

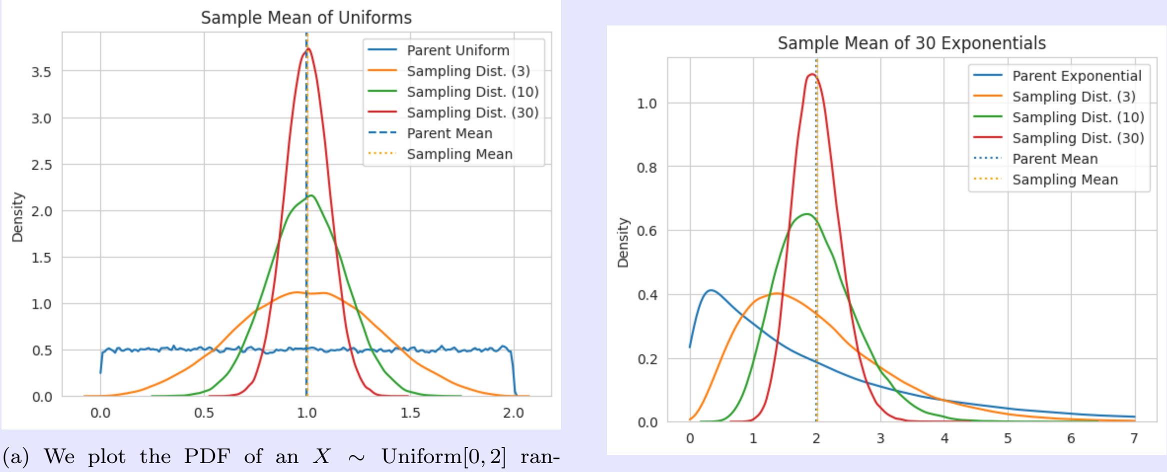
Practically, this tells us that when trying to estimate the value of a population mean, due to the factor of $1/\sqrt{n}$, reducing the error on the estimate by a factor of 2 requires acquiring 4 times as many observations in the sample. But realistically, the true standard deviation σ is unknown, and so the standard error of the mean is usually estimated by replacing σ with the sample standard deviation S instead.

$$\sigma_{\overline{x}_n} \approx \frac{S}{\sqrt{n}} \tag{17}$$

By CLT, $\overline{x}_n$ converges to $\mathcal{N}(\mu, \sigma^2/n)$ in distribution as $n \to +\infty$ (but in practicality, we assume this for $n \geq 30$). The fact that its mean and variance is μ and σ^2/n isn't that impressive. What is really impressive is that no matter what the distribution of x is, the sampling distribution of the mean will be Gaussian.

Example 1.1 (Sample Means)

Here are some figures of sample means. Note that with a uniform parent distribution, the sampling distribution of its mean looks like a Gaussian even without a large n. However, this is not necessarily true for different parent distributions, such as the exponential.



(a) We plot the PDF of an $X \sim \text{Uniform}[0, 2]$ random variable by taking 100k samples. We also take 100k samples from the sampling distribution of the mean $\overline{X}_3, \overline{X}_{10}, \overline{X}_{30}$. We can see that the standard deviation decreases by a factor of $\sqrt{n}$.

(b) We plot the PDF of an $X \sim \text{Exponential}(1.5)$ random variable by taking 100k samples. We also take 100k samples from the sampling distribution of the mean $\overline{X}_3, \overline{X}_{10}, \overline{X}_{30}$.

Figure 1

If the parent distribution is normal, then we don't even need CLT to claim that the sampling distribution of the sample mean is normal, since sums of normals are normal.

Now the variance of the population is defined to be $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$, and by our rule of thumb, we can replace the expectations with sample means, by first setting $\mathbb{E}[X] = \widehat{\mu}$ and averaging out the values $(X - \widehat{\mu})^2$.

Definition 1.5 (Sample Variance)

Given a population X, our estimator for $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$ is simply the average of the squared distances of the n samples $\{(x_i - \widehat{\mu})^2\}_{i=1}^n$.

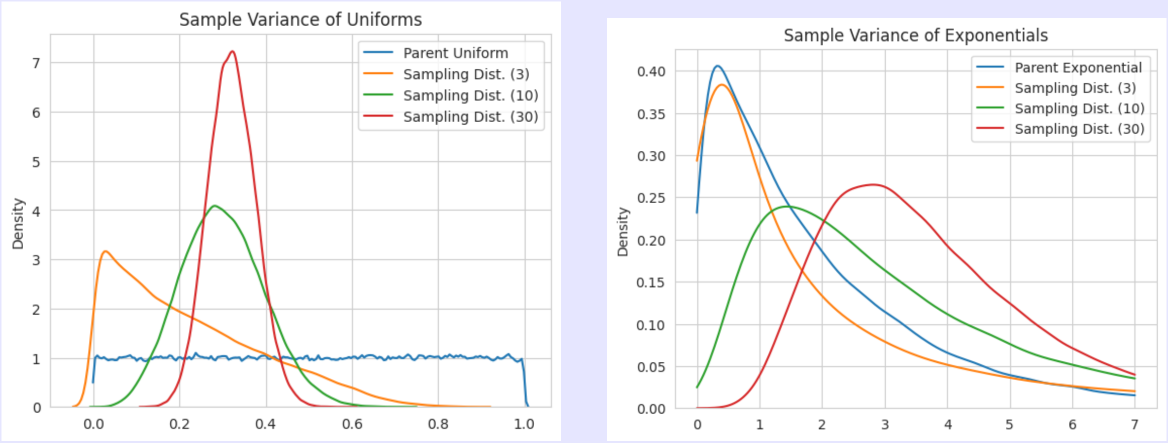
$$S_n^2 = \widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x}_n)^2 \tag{18}$$

The mean and standard deviation of S_n^2 is denoted $\mu_{S_n^2}$ and $\sigma_{S_n^2}$. Note that there is a small difference that the sum for variance is divided by $n-1$ rather than n, since we want it to be unbiased, but we will correct this later.

While the CLT states that the sampling distribution of the sample mean will look approximately Gaussian, we do not have this luxury when looking at the sampling distribution of sample variance.

Example 1.2 (Sample Variance)

Take a look at the following sampling distributions of the sample variance. There does not seem to be strong signs of convergence to a Gaussian. Their means do not align either.



(a)

(b)

Figure 2

1.2 Sampling from Gaussians

Now if we assume that the parent distribution is Gaussian, then we can conclude some extra things and more kinds of distributions arise. Let $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$, with $\overline{x}_n$ the sample mean and S_n^2 the sample variance. Say that we want to find the distribution of $\overline{x}_n$.

1. In the unrealistic case where we know the true σ^2, we don't even need to consider the sample variance. From the basic property of Gaussians, we know that $\overline{x}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, or after standardizing,

$$\frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{19}$$

2. In the realistic case where we don't know the true σ^2, we should replace it with our sample variance S^2, and it turns out that because of this extra uncertainty in the variance, our sampling distribution follows the student-t distribution, which can be interpreted as a mixture of Gaussians with differing variances.
$$\frac{\overline{x}_n - \mu}{S/\sqrt{n}} \sim \text{StudentT}(n-1) \tag{20}$$

Now if we are interested in finding the distribution of S_n^2:

1. In the unrealistic case where the know the true μ, we don't need to consider the sampling distribution of $\overline{x}_n$. We have
$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2 \sim \text{Gamma}\Big(\frac{n}{2}, \frac{n}{2\sigma^2}\Big) \tag{21}$$

2. In the realistic case where we don't know μ, we have
$$\frac{n-1}{\sigma^2} S_n^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \overline{x}_n)^2 \sim \chi^2(n-1) \tag{22}$$

2 Hypothesis Testing

This was done for the first time as early as 1710, when a researcher tried to measure sex ratios in a population [Arb10].

A significance test is a method used to decide whether the data at hand sufficiently supports a particular hypothesis. The hypothesis to be tested is called the **alternative hypothesis**, denoted H_1 or H_a, and the status quo is called the **null hypothesis**, denoted H_0. Assuming that H_0 is true, we compute the likelihood of the data happening. If the sample is not too unlikely (past some significance level), we fail to reject H_0, and if there is strong evidence, we reject H_0. H_0 and H_a can be devised in countless ways.

Example 2.1 ()

There are countless test statistics we can build, but here are some common examples,

1. Proportion: Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards.

$$H_0 : p = 0.10 \text{ versus } H_a : p < 0.10 \tag{23}$$

2. Means: It is known that the average height of boys in KIS is 176cm. Ben claims that the average height is lower than this.

$$H_0 : \mu = 176 \text{ versus } H_a : \mu < 176 \tag{24}$$

3. Difference of Means: If μ_1 and μ_2 denote the true average breaking strengths of the same type of twine produced by two different companies. Jenny claims that the $\mu_1 - \mu_2 > 5$.

$$H_0 : \mu_1 - \mu_2 = 0 \text{ versus } H_a : \mu_1 - \mu_2 > 5 \tag{25}$$

2.1 One Sample Z and T Tests

Let us have some population $X \sim P$ and a null hypothesis that claims $H_0 : \mu = \theta_0$. Since we are interested in the mean, we would like to use CLT or some other theorem to determine what the distribution of the mean of n samples $\overline{x}_n$ looks like (either Normal or Student T centered around θ_0 and scaled down by factor of $\sqrt{n}$). When we actually sample, the value $\overline{x}_n = \hat{\theta}$ is realized, and we would like to see if sampling $\hat{\theta}$ from the distribution centered around θ_0 is likely, usually after normalizing. If it isn't, then we reject H_0.

How do we decide whether to use the z-test or the t-test? It is known that StudentT$(n-1)$ converges to $\mathcal{N}(0,1)$ in distribution as $n \rightarrow +\infty$. Therefore, depending on the context of the problem, at a certain point N (usually $N = 30$ or perhaps higher for skewed distributions), the difference between these two are negligible.

1. Z-test: if we know the population variance σ^2, but it is rarely the case that we actually know σ^2.
2. T-test: if we do not know the population variance σ^2, which we then substitute for the sample variance S^2.
3. Z-test: if we do not know the population variance (which we substitute for S^2), but our sample size is greater than N, then we can approximate the t-distribution with our normal, allowing us to use the Z-test again.

In general, the alternative to the null hypothesis $H_0 : \theta = \theta_0$ will looks like one of the following three assertions:

1. Two-Sided Test: $H_a : \theta \neq \theta_0$
2. One-Sided Test: $H_a : \theta > \theta_0$ (in which case the null hypothesis is $\theta \leq \theta_0$)
3. One-Sided Test: $H_a : \theta < \theta_0$ (in which case the null hypothesis is $\theta \geq \theta_0$)

Now we must still quantify *how* unlikely our sample mean θ must be compared to θ_0 in order to reject the null hypothesis. This is where we specify our **significance level**, denoted by α (common values $0.10, 0.05, 0.01$). This specifies the tail-regions in which θ will land in with probability α. Usually, working with general normal/t distributions is tedious, so we can rescale them and use their z/t-scores.

Definition 2.1 (Z-score)

Given a value x sampled from distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, its **z-score** is defined to be the number of standard deviations away from the mean.

$$z := \frac{x - \mu}{\sigma} \tag{26}$$

Now given a significance level $\alpha \in [0, 1]$, let z_α be the value such that the measure of a standard normal distribution past z_α is $1 - \alpha$ (i.e. the 100α percentile). z_α is called the **critical z-value**.

Definition 2.2 (T-score)

Given a value x sampled from distribution $X \sim \text{StudentT}(n)$, its **t-score** is defined to be the number of standard deviations away from the mean.

Example 2.2 ()

A factory has a machine that dispenses 80mL of fluid in a bottle. An employee believes the average amount of fluid is not 80mL. Using 40 samples, he measures the average amount dispensed by the machine to be 78mL with a sample standard deviation of 2.5.

1. Let the true mean be μ and true standard deviation be σ. The null hypothesis is $H_0 : \mu = 80$ and the alternative is $H_1 : \mu \neq 80$, making this a two-sided test.
2. We don't know the true standard deviation σ, so we must use the sample standard deviation S. This requires us to use the t-test, but since $n > 30$, we can invoke CLT and state that $\overline{x}_{40}$ is (approximately) Gaussian with mean μ and standard deviation $S/\sqrt{n}$. So, we use the z-test.
3. At a 95% confidence level, we have $\alpha = 0.05$, and our rejection region is $(-\infty, z_{0.025}] \cup [z_{0.975}, +\infty)$. Since we are looking at a standard Gaussian, we have by symmetry $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$, and our critical z-value is $z^* = 1.96$.
4. So the z-score for 78 is
$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{78 - 80}{2.5/\sqrt{40}} = -5.06 \tag{27}$$
which is definitely in the reject region. So this tells us that we can reject the null hypothesis with a 95% level of confidence.

Example 2.3 ()

A company manufactures car batteries with an average life span of 2 or more years. An engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 years with a standard deviation of 0.15.

1. Let the true mean be μ and true standard deviation be σ. The null hypothesis is $H_0 : \mu \geq 2$ and the alternative is $H_1 : \mu < 2$, making this a one-sided test.
2. We don't know the true standard deviation σ, so we must use the sample standard deviation S. This requires us to use the t-test, especially since $n = 10$ is not large enough for us to invoke CLT.
3. At a 99% confidence level, we have $\alpha = 0.01$, and our rejection region is $(-\infty, t_{0.01}] = (-\infty, -2.82]$.

4. The t-score for the observed mean value is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}} = -4.22 \tag{28}$$

which is definitely in the reject region. So this tells us that we can reject the null hypothesis with a 99% level of confidence.

We may have to account for errors. There is always a chance that our evidence leads us to an incorrect conclusion, and we have names for this.

Definition 2.3 (Errors)

Given a hypothesis test where we look for evidence supporting our alternative claim,

1. A **type 1 error** is when the null hypothesis is rejected, but it is true (false positive).
2. A **type 2 error** is when we fail to reject the null hypothesis, when it is false (false negative).

2.2 Power of a test

2.3 Common tests (t-test, z-test, chi-square test, F-test)

2.4 Multiple testing problem

3 Method of Moments

4 Maximum Likelihood Estimation

Score function, Fisher information.

5 Bias Variance Decomposition

The likelihood defines a proper loss function. Now let's try to parse the loss a bit more. It turns out that for a lot of popular loss functions, they generally decomposes into

$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise} \tag{29}$$

This decomposition is by no means exact, but it *generally* holds true. This is formalized by deriving an exact decomposition of the MSE loss. This is possible because the MSE loss allows us to get an L^2 space, which allows for orthogonal decompositions. Unfortunately, when you have other losses, this becomes much messier because the inner product structure is not there.

In a more general case, when you take a supremum of risk over a function class, it decomposes into 3 terms.

1. One of which quantifies how big the function class is (more variance).
2. One of which quantifies the distance between the truth and the function class (bias).
3. One is the noise term, which is the irreducible error.

Example 5.1 (Bias and Variance Tradeoff in Polynomial Regression)

Let's motivate this by trying to fit a polynomial on some data.

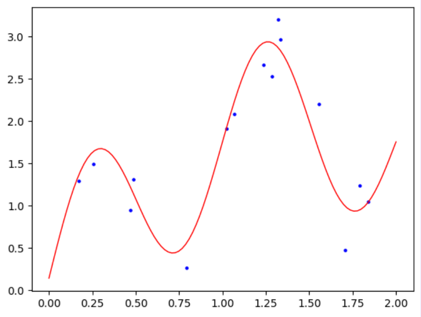
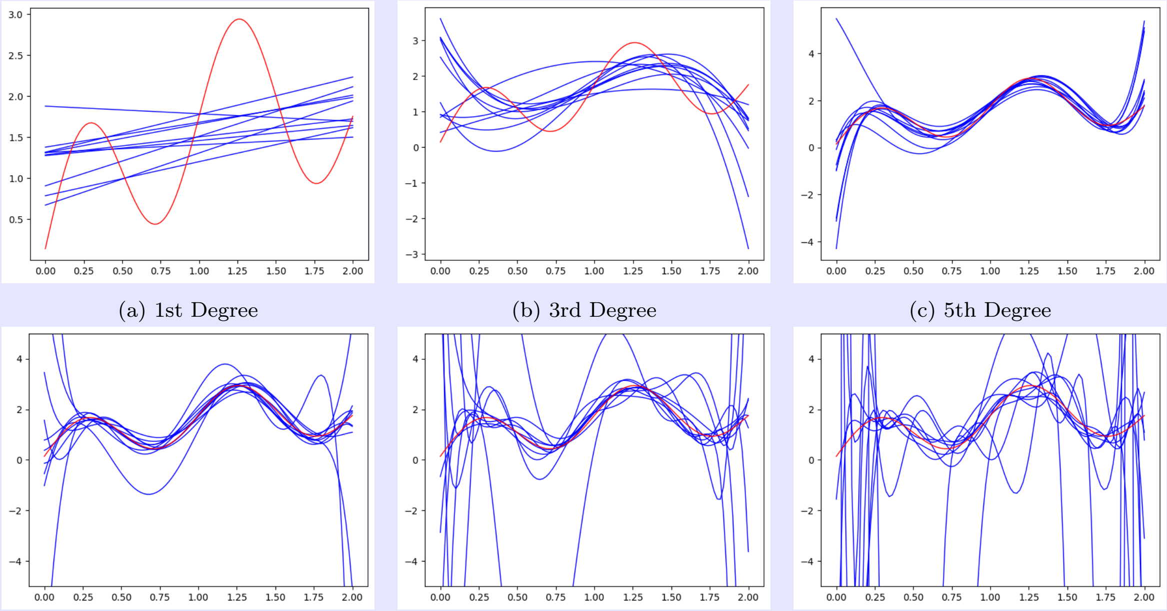


Figure 3: A sample of $|\mathcal{D}| = 15$ data points are generated from the function $f(x) = \sin(2\pi x) + 2\cos(x - 1.5)$ with Gaussian noise $N(0, 0.3)$ on the interval $[0, 1]$.

If we try to fit a polynomial function, how do we know which degree is best? Well the most simple thing is to just try all of them. To demonstrate this even further, I generated 10 different datasets $\mathcal{D}$ of size 15 taken from the same true distribution. The best fitted polynomials for each dataset is shown below.



(a) 1st Degree (b) 3rd Degree (c) 5th Degree

(d) 7th Degree (e) 9th Degree (f) 11th Degree

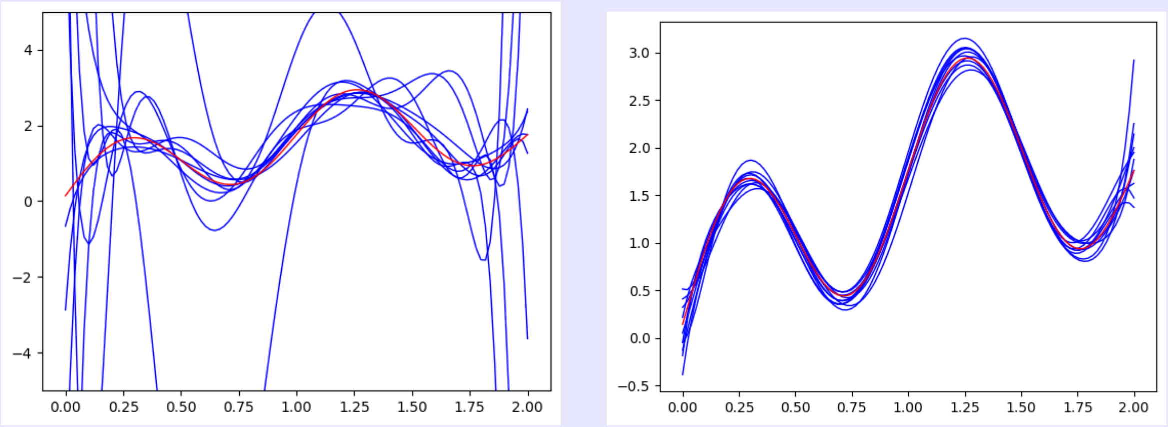
Figure 4: Different model complexities (i.e. different polynomial degrees) lead to different fits of the data generated from the true distribution. The lower degree best fit polynomials don't have much variability in their best fits but have high bias, while the higher degree best fit polynomials have very high variability in their best fits but have low bias. The code used to generate this data is here.

We already know that the 5th degree approximation is most optimal, and the lower degree ones are **underfitting** the data, while the higher degree ones are **overfitting**. As mentioned before, we can describe the underfitting and overfitting phenomena through the bias variance decomposition.

1. If we underfit the data, this means that our model is not robust and does not capture the patterns inherent in the data. It has a high bias since the set of function it encapsulates is not large enough to model $\mathbb{E}[Y \mid X]$. However, it has a low variance since if we were to take different samples of the dataset $\mathcal{D}$, the optimal parameters would not fluctuate.
2. What overfitting essentially means is that our model is too complex to the point where it starts to fit to the *noise* of the data. This means that the variance is high, since different samples of the dataset $\mathcal{D}$ would cause huge fluctuations in the optimal trained parameters $\boldsymbol{\theta}$. However, the function set would be large, and thus it would be close to $\mathbb{E}[Y \mid X]$, leading to a low bias.

Example 5.2 (Polynomial Regression Continued)

Another way to reduce the overfitting problem is if we have more training data to work with. That is, if we were to fit a 9th degree polynomial on a training set of not $N = 15$, but $N = 100$ data points, then we can see that this gives a much better fit. This makes sense because now the random variable $\mathcal{D}$, as a function of more random variables, has lower variance. Therefore, the lower variance in the dataset translates to lower variance in the optimal parameter.



(a) $M = 9, N = 15$

(b) $M = 9, N = 100$

Figure 5: Increasing the number of data points helps the overfitting problem. Now, we can afford to fit a 9th degree polynomial with reasonable accuracy.

5.1 MSE Loss

Definition 5.1 (Mean Squared Error Loss)

The **MSE loss** is defined

$$L(y, x) = (y - f(x))^2 \tag{30}$$

It is a well known fact that the true regressor—which may not be linear at all—that minimizes this loss is

$$f^*(x) = \mathbb{E}[Y \mid X = x] \tag{31}$$

which is the conditional expectation of Y given X. This is the true regressor function, which is the best approximation of Y over the σ-algebra generated by X. Therefore, if we consider a function class of linear predictors, we can decompose our risk, which is the distance between our estimated linear regressor and Y, as the sum of the distance between our estimator and the best regressor plus the distance between the best regressor and Y.

Theorem 5.1 (Pythagorean's Theorem)

The expected square loss over the joint measure $\mathbb{P}_{x,y}$ can be decomposed as

$$\mathbb{E}_{x,y}[(y - f(x))^2] = \mathbb{E}_{x,y}\big[\big(y - \mathbb{E}[y \mid x]\big)^2\big] + \mathbb{E}_x\big[\big(\mathbb{E}[y \mid x] - h(x)\big)^2\big] \tag{32}$$

That is, the squared loss decomposes into the squared loss of $\mathbb{E}[y \mid x]$ and $g(x)$, which is the intrinsic misspecification of the model, plus the squared difference of y with its best approximation $\mathbb{E}[y \mid x]$, which is the intrinsic noise inherent in y beyond the σ-algebra of x.

Proof.

We can write

$$\begin{aligned}
\mathbb{E}_{x,y}\big[\big(y - f(x)\big)^2\big] &= \mathbb{E}_{x,y}\big[\big((y - \mathbb{E}[y \mid x]) + (\mathbb{E}[y \mid x] - f(x))\big)^2\big] && (33)\\
&= \mathbb{E}_{x,y}\big[\big(y - \mathbb{E}[y \mid x]\big)^2\big] + \mathbb{E}_{x,y}[\{y - \mathbb{E}[y \mid x]\}\,\{\mathbb{E}[y \mid x] - f(x)\}] && (34)\\
&\qquad + \mathbb{E}_{x}\big[\big(\mathbb{E}[y \mid x] - f(x)\big)^2\big] && (35)\\
&= \mathbb{E}_{x,y}\big[\big(y - \mathbb{E}[y \mid x]\big)^2\big] + \mathbb{E}_{x}\big[\big(\mathbb{E}[y \mid x] - f(x)\big)^2\big] && (36)
\end{aligned}$$

where the middle term cancels out due to the tower property.

Note that since $\mathbb{E}[\big(\mathbb{E}[Y \mid X] - g(X)\big)^2]$ is the misspecification of the model, we cannot change this (positive) constant, so $\mathbb{E}\big[\big(Y - g(X)\big)^2\big] \geq \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2]$, with equality achieved when we perfectly fit g as $\mathbb{E}[Y \mid X]$ (i.e. the model is well-specified). Therefore, denoting $\mathcal{F}$ as the set of all $\sigma(X)$-measurable functions, then the minimum of the loss is attained when

$$\operatorname*{argmin}_{g \in \mathcal{F}} \mathbb{E}[L] = \operatorname*{argmin}_{g \in \mathcal{F}} \mathbb{E}\big[\big(Y - g(X)\big)^2\big] = \mathbb{E}[Y \mid X] \tag{37}$$

Essentially, we have decomposed our risk to a part that we can optimize and a part that we cannot, i.e. the intrinsic noise.

Corollary 5.1 (Sufficient to Estimate Conditional Expectation)

Minimizing the prediction risk is equivalent to minimizing the risk of our estimator to the conditional distribution.

$$\operatorname*{argmin}_{f \in \mathcal{F}} R(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{x,y}\left[(\mathbb{E}[y|x] - f(x))^2\right] \tag{38}$$

Even though this example is specific for the mean squared loss, this same decomposition, along with the bias variance decomposition, exists for other losses. It just happens so that the derivations are simple for the MSE, which is why this is introduced first. However, the derivations for other losses are much more messy, and sometimes may not hold rigorously. However, the general intuition that more complex models tend to overfit (higher variance) still hold true.

Let's try to decompose this even more. In frequentist inference, we take a dataset $\mathcal{D}$ and optimize $\hat{f}$ that minimizes this empirical risk. Therefore, for a given $\mathcal{D}$, $\hat{f} = \hat{f}(\mathcal{D})$ is determined, and if $\mathcal{D} = (x^{(i)}, y^{(i)})^n$ is a random variable, then $\hat{f}$ is also a random variable, which we will denote as $\hat{f}_{\mathcal{D}}$ for clarity. It is useful to think of $\mathcal{D}$ as a random variable because by seeing how $\hat{f}_{\mathcal{D}}$ varies as the dataset changes, we can measure the uncertainty in our estimate of $\hat{f}_{\mathcal{D}}$ through $\mathcal{D}$.[4]

Lemma 5.1 (Conditional Prediction Risk)

Our conditional prediction risk is

$$r(\mathcal{D}) = \mathbb{E}_{x,y}\left[(\mathbb{E}[y \mid x] - \hat{f}(x))^2 \mid \mathcal{D}\right] \tag{39}$$

If $\mathcal{D}$ is fixed, then this is a real number. If $\mathcal{D}$ is a random variable, then this is a real-valued random variable.

Ideally, we would like two things.

[4] If this didn't make sense to you, consider the following thought experiment. Suppose we had a large number of datasets each of size N and each drawn independently from the joint distribution $X \times Y$. For any given dataset $\mathcal{D}$, we can run our learning algorithm and obtain our best fit function $\hat{f}_{\mathcal{D}}$. Different datasets from the ensemble will give different functions and consequently different values of the squared loss.

1. *Low Bias.* The average prediction we get over all $\hat{f}_\mathcal{D}$ trained on all possible samples of dataset $\mathcal{D}$ should be similar to our best regressor. That is,
$$\mathbb{E}_\mathcal{D}\left[\mathbb{E}_{x,y}\left[(\mathbb{E}[y \mid x] - \hat{f}_\mathcal{D}(x))^2\right]\right] \tag{40}$$
should be as low as possible.

2. *Low Variance.* The variance of our conditional prediction risk
$$\mathrm{Var}_\mathcal{D}\left[\mathbb{E}_{x,y}\left[(\mathbb{E}[y \mid x] - \hat{f}_\mathcal{D}(x))^2\right]\right] \tag{41}$$
should be as low as possible. That is, we may get very low bias for one dataset $\mathcal{D}$, but if we sampled a different dataset, we should not expect the bias to explode.

Unfortunately, having both low bias *and* low variance is not possible, and we wish to show that now.

Theorem 5.2 (Bias Variance Decomposition Under MSE Loss)

The expected optimal MSE loss decomposes to
$$\mathbb{E}_\mathcal{D}\left[(\mathbb{E}[y \mid x] - \hat{f}_\mathcal{D}(x))^2\right] = \underbrace{\big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)^2}_{\text{(bias of } \hat{f}_\mathcal{D})^2} + \underbrace{\mathbb{E}_\mathcal{D}\big[\big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big)^2\big]}_{\text{variance of } \hat{f}_\mathcal{D}} \tag{42}$$

Proof.

Consider the term $\big(\mathbb{E}[y \mid x] - \hat{f}_\mathcal{D}(x)\big)^2$ above, which models the discrepancy in our optimized hypothesis and the best approximation. We take $\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]$[a] So we can split the term into

$$\begin{aligned}
\big(\mathbb{E}[y \mid x] - \hat{f}_\mathcal{D}(x)\big)^2 &= \big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big) + \big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big)\big]^2 && (43)\\
&= \big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)^2 + \big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big)^2 && (44)\\
&\quad + 2\big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)\big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big) && (45)
\end{aligned}$$

Now take the expectation over $\mathcal{D}$, and for the third term, note that $\big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)$ is constant with respect to $\mathbb{D}$ anyways, so we can take it out of the expectation. Therefore,

$$\begin{aligned}
\mathbb{E}_\mathcal{D}\big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)\big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big)\big]& && (46)\\
= \big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big)\,\mathbb{E}_\mathcal{D}\big[\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big]& && (47)\\
= \big(\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]\big) \cdot 0 = 0& && (48)
\end{aligned}$$

Let's parse these terms a bit more.

1. The bias $\mathbb{E}[y \mid x] - \mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]$ is a random variable of x that measures the difference between the average of our learned predictor $\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]$ and the true regressor $\mathbb{E}[y \mid x]$.

2. The variance $\mathbb{E}_\mathcal{D}\big[\big(\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)] - \hat{f}_\mathcal{D}(x)\big)^2\big]$ is a random variable of x that measures the variability of our learned functions $\hat{f}_\mathcal{D}$ around our mean $\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}(x)]$.

Therefore, we can substitute this back into our Pythagoras decomposition, where we must now take the expected bias and the expected variance over x to get a form like
$$\text{Expected Loss} = (\text{Expected Bias})^2 + \text{Expected Variance} + \text{Noise} \tag{49}$$

[a] Over all datasets $\mathcal{D}$, there will be a function $h_{\boldsymbol{\theta};\mathcal{D}}$, and averaged over all datasets $\mathcal{D}$ is $\mathbb{E}_\mathcal{D}[\hat{f}_\mathcal{D}]$.

Corollary 5.2 (Bias Variance Decomposition of Expected MSE Loss)

The expected optimal MSE loss decomposes to

$$\mathbb{E}_{\mathcal{D}}\mathbb{E}_{x,y}\big[(y - \hat{f}_{\mathcal{D}}(x))^2\big] = \mathbb{E}_x\big[\underbrace{\big(\mathbb{E}[y \mid x] - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\big)^2}_{\text{(expected bias)}^2}\big] + \underbrace{\mathbb{E}_{\mathcal{D}}\big[\mathbb{E}_x\big[\big(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\big)^2\big]\big]}_{\text{expected variance}} \tag{50}$$

$$+ \underbrace{\mathbb{E}_{x,y}[(y - \mathbb{E}[y \mid x])^2]}_{\text{noise}} \tag{51}$$

Proof.

By taking the expectation over x and swapping the expectations (since x and $\mathcal{D}$ are independent), and finally substituting back to Pythagoras decomposition, we get the following.

5.2 MAE Loss

6 Confidence Intervals

Recall that the central limit theorem says that given a sequence of iid random variables $x_1, \ldots, x_n$ coming from a random variable with true mean μ and variance σ^2, the sample mean is similar to a $\mathcal{N}(\mu, \sigma^2/n)$ random variable. That is, the sample mean converges in distribution

$$\overline{X}_n \xrightarrow{dist} \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big) \tag{52}$$

as $n \rightarrow \infty$. Another way to state it is that the normalized sample mean is similar to a standard Gaussian.

$$\frac{\overline{x}_n - \mu}{\sigma_{\overline{x}_n}} = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{dist} \mathcal{N}(0,1) \tag{53}$$

So, given that we have enough samples, I will perfectly understand its fluctuations. Now let's introduce some definitions that will allow us to unify some ideas into simpler notation: the realized value x, the number of standard deviations it is away from the mean, and the probability that it takes that value (or more extreme).

Definition 6.1 (z-score)

Given a $\mathcal{N}(\mu, \sigma^2)$ distribution, the **z-score** of a number $x \in \mathbb{R}$ is defined to be the number of standard deviations away from the mean.

$$z = \frac{x - \mu}{\sigma} \tag{54}$$

Definition 6.2 (Percentile)

Given $X \sim \mathcal{N}(0,1)$ and significance level $\alpha \in [0,1]$, let us define $q_\alpha \in \mathbb{R}$ as the point where

$$\mathbb{P}(X \geq q_\alpha) = \alpha \tag{55}$$

i.e. the 100αth percentile of the standard normal. Note that given $X \sim \mathcal{N}(0,1)$, we have

$$\mathbb{P}(|X| > q_{\alpha/2}) = \alpha \tag{56}$$

Now given $x_1, \ldots, x_n$ from a population X with mean μ and standard deviation σ, let $\overline{x}_n$ be the sampling distribution of the mean. By virtue of the central limit theorem, we can write

$$\mathbb{P}\bigg(\bigg|\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\bigg| \geq q_{\alpha/2}\bigg) \approx \alpha \iff \mathbb{P}\bigg(\bigg|\frac{\overline{X}_n - \mu}{\sigma\sqrt{n}}\bigg| \leq q_{\alpha/2}\bigg) \approx 1 - \alpha \tag{57}$$

which implies that with probability $1 - \alpha$, we have

$$\overline{X}_n \in \bigg[\mu - q_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + q_{\alpha/2}\frac{\sigma}{\sqrt{n}}\bigg] \iff \mu \in \bigg[\overline{X}_n - q_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{\alpha/2}\frac{\sigma}{\sqrt{n}}\bigg] \tag{58}$$

This is how we construct a confidence interval. In other words, as n becomes large (ideally at least 30), the probability that an interval around our sample mean contains the actual mean μ can be approximated by a Gaussian. But note that CI requires to know the actual standard deviation σ. There are three ways to deal with this:

1. This may actually be known from the start, especially if we are working with calibrated devices with standard devices that have been experimentally verified.
2. We can simply bound σ, depending on what kind of random variable we are working with. For example, given $X \sim$ Bernoulli(p), its standard deviation is bounded by $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$, so we can create a confidence interval that is larger than any other confidence interval we can make if we had known the true σ.

$$p \in \bigg[\overline{X}_n - q_{\alpha/2}\,\frac{1}{2\sqrt{n}}, \overline{X}_n + q_{\alpha/2}\,\frac{1}{2\sqrt{n}}\bigg] \tag{59}$$

3. We can approximate σ with the sample standard deviation S, which turns out to be an unbiased estimator.

Example 6.1 (Proportion of Right-Side Kissers)

We have observed 80 out of 124 right-side kisses, resulting in a sample estimate of $\widehat{p} = 0.645$. Given that we want a confidence interval of 95%, we want an $\alpha = 0.05$, implying a the value $q_{\alpha/2} = q_{0.025} = 1.96$. So, with probability 0.95, we have

$$p \in \bigg[0.645 - \frac{1.96}{2\sqrt{124}}, 0.645 + \frac{1.96}{2\sqrt{124}}\bigg] = [0.56, 0.73] \tag{60}$$

If we had, say 3 observations, rather than 124, we would have a 95% confidence interval of $p \in [0.10, 1.23]$, which is terrible, but in this case even CLT is not valid.

Example 6.2 (Proportion of Voters)

Given that we sample $n = 100$ people from a city's population to ask whether they support candidate A or B, we have 54 people who support candidate A, so $\widehat{p} = 0.54$. Say that we want a 95% confidence interval, which leads to $q_{\alpha/2} = q_{0.025} = 1.96$. So, with probability 0.95, we have

$$p \in \bigg[0.54 - 1.96\,\frac{\sigma}{\sqrt{100}}, 0.54 + 1.96\,\frac{\sigma}{\sqrt{100}}\bigg] \tag{61}$$

and by substituting σ for $S = \sqrt{0.54(1-0.54)} \approx 0.5$, we get

$$p \in \bigg[0.54 - 1.96\,\frac{0.284}{\sqrt{100}}, 0.54 + 1.96\,\frac{0.284}{\sqrt{100}}\bigg] = [0.44, 0.64] \tag{62}$$

An interpretation of confidence intervals is that if you keep on sampling $\overline{x}$ or $\widehat{p}$ and construct 95% CIs, then 95% of the time these intervals will contain the true mean μ or proportion p (or more if we had bounded the CI with a bigger interval).

Example 6.3 ()

We survey 6250 teachers to ask whether they think computers are essential for teaching. 250 were randomly selected and 142 felt that they were essential. Let's construct a 99% confidence interval for the proportion of teachers who felt that computers were essential. We would like to construct a CI for the true $\mu = p$, and we have $\overline{x} = 142/250 = 0.568$.

1. 99% confidence corresponds to $\alpha = 0.01$, which corresponds to a z-score of $q_{\alpha/2} = 2.576$.
2. The parent distribution is Bernoulli(p), with $\mu = p$ and $\sigma = \sqrt{p(1-p)}$. The sampling distribution of $\overline{x}$ has $\mu_{\overline{x}} = p$ also and $\sigma_{\overline{x}} = \sigma/\sqrt{n}$.
3. We need to know the details of the sampling distribution, but we don't know σ, which is needed to calculate $\sigma_{\overline{x}}$. However, we can estimate it using the sample standard deviation $S = \sqrt{0.568(1-0.568)} = 0.5$.
4. Our sampling distribution has standard deviation $\sigma_{\overline{x}} \approx S/\sqrt{n} = 0.5/\sqrt{250} = 0.031$, and our z-score was 2.576, so our 99% confidence interval is 2.576 standard deviations from our mean. That is, with probability 0.99,

$$p \in \big[0.568 - 2.576 \cdot 0.031, 0.568 + 2.576 \cdot 0.031\big] = \big[0.488144, 0.647856\big] \tag{63}$$

6.1 CIs for means, proportions, and variances

6.2 Bootstrap confidence intervals

7 Cross Validation

We have understood the theoretical foundations of overfitting and underfitting with the bias variance decomposition. But in practice, we don't have an ensemble of datasets; we just have one. Therefore, we don't actually know what the bias, the variance, or the noise is at all. Therefore, how do we actually *know* in practice when we are underfitting or overfitting? Easy. We just split our dataset into 2 different parts: the training set and testing sets.

$$\mathcal{D} = \mathcal{D}_{train} \sqcup \mathcal{D}_{test} \tag{64}$$

What we usually have is a **training set** that allows us to train the model, and then to check its performance we have a **test set**. We would train the model on the training set, where we will always minimize the loss, and then we would look at the loss on the test set. Though we haven't made a testing set, since we know the true model let us just generate more data and use that as our testing set. For each model, we can calculate the optimal $\boldsymbol{\theta}$, which we will denote $\boldsymbol{\theta}^*$, according to the **root mean squared loss**

$$h_{\boldsymbol{\theta}^*} = \operatorname*{argmin}_{h_{\boldsymbol{\theta}}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})\big)^2} \tag{65}$$

where division of N allows us to compare different sizes of datasets on equal footing, and the square root ensures that this is scaled correctly. Let us see how well these different order models perform on a separate set of data generated by the same function with Gaussian noise.

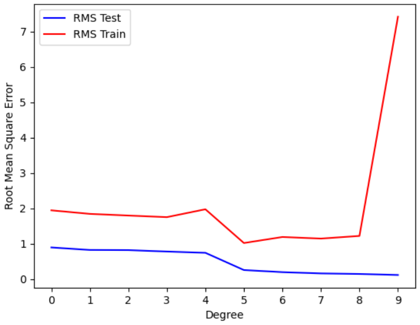


Figure 6: We can see that the RMS decreases monotonically on the training error as more complex functions become more fine-tuned to the data. However, when we have a 9th degree polynomial the RMS for the testing set dramatically increases, meaning that this model does not predict the testing set well, and performance drops.

Now we know that a more complex model (i.e. that captures a greater set of functions) is not necessarily the best due to overfitting. Therefore, researchers perform **cross-validation** by taking the training set $(\mathcal{X}, \mathcal{Y})$. We divide it into S equal pieces

$$\bigcup_{s=1}^{S} D_s = (\mathcal{X}, \mathcal{Y}) \tag{66}$$

Then, we train the model $\mathcal{M}$ on $S-1$ pieces of the data and then test it across the final piece, and do this S times for every test piece, averaging its perforance across all S test runs. Therefore, for every model $\mathcal{M}_k$, we must train it S times, for all K models, requiring KS training runs. If data is particularly scarce, we set $S = N$, called the **leave-one-out** technique. Then we just choose the model with the best average test performance.

Code 7.1 (Minimal Example of Train Test Split in scikit-learn)

To implement this in scikit-learn, we want to use the `train_test_split` class. We can also set a random state parameter to reproduce results.

```
from sklearn.model_selection import train_test_split

# Split into training (80\%) and test (20\%) data
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2,
    random_state=66)
```

7.1 Concentration Bounds

The following result shows that cross-validation (data splitting) leads to an estimator with risk nearly as good as the best model in the class.

Theorem 7.1 (Gyorfi, Kohler, Krzyak, Walk (2002))

Let $\mathcal{M} = \{m_h\}$ be a finite class of regression estimators indexed by a parameter h, with m being the true risk minimizer, $m_{\hat{h}}$ being the empirical risk minimizer over the whole dataset $\mathcal{D}$, and m_H being the empirical risk minimizer over the test set $\mathcal{D}_{\text{test}}$ for ordinary least squares loss.

$$m_H = \operatorname*{argmin}_{m_h} \frac{1}{N} \sum_{i \in \mathcal{D}_{\text{test}}} (y_i - m_h(x_i))^2 \tag{67}$$

$$m_{\hat{h}} = \operatorname*{argmin}_{m_h} \frac{1}{N} \sum_{i \in \mathcal{D}} (y_i - m_h(x_i))^2 \tag{68}$$

If the data Y_i and estimators are bounded by L, then for any $\delta > 0$, we have

$$\mathbb{E} \int |m_H(x) - m(x)|^2 \, d\mathbb{P}(x) \leq (1+\delta) \mathbb{E} \int |m_{\hat{h}}(x) - m(x)|^2 \, d\mathbb{P}(x) + \frac{C(1 + \log|M|)}{n} \tag{69}$$

where $c = L^2(16/\delta + 35 + 19\delta)$.

Proof.

Then

$$\begin{aligned} \mathbb{E}\left(\int |m_H - m|^2 dP(x) | D \right) &= \mathbb{E}\left(\int |Y - m_H|^2 dP(x) | D \right) - \mathbb{E}|Y - m(X)|^2 & (70) \\ &= T_1 + T_2 & (71) \end{aligned}$$

where

$$T_1 = \mathbb{E}\left(\int |Y - m_H|^2 dP(x) | D \right) - \mathbb{E}|Y - m(X)|^2 - T_2 \tag{72}$$

and

$$\begin{aligned} T_2 &= (1+\delta) \frac{1}{n} \sum_{D'} (|Y_i - m_H(X_i)|^2 - |Y_i - m(X_i)|^2) & (73) \\ &\leq (1+\delta) \frac{1}{n} \sum_{D'} (|Y_i - m_{\hat{h}}(X_i)|^2 - |Y_i - m(X_i)|^2) & (74) \end{aligned}$$

and so

$$\mathbb{E}[T_2|D] \leq (1+\delta)\left(\mathbb{E}(|Y - m_{\hat{h}}(X)|^2|D) - \mathbb{E}|Y - m(X)|^2\right) \tag{75}$$
$$= (1+\delta)\int |m_{\hat{h}}(x) - m(x)|^2 dP(x). \tag{76}$$

The second part of the proof involves some tedious calculations. We will bound $P(T_1 \geq s|D)$. The event $T_1 \geq s$ is the same as

$$(1+\delta)\left(\mathbb{E}(|m_H(X) - Y|^2|D) - \mathbb{E}|m(X) - Y|^2 - \frac{1}{n}\sum_{D'}(|Y_i - m_H(X_i)|^2 - |Y_i - m(X_i)|^2)\right) \tag{77}$$
$$\geq s + \delta\left(\mathbb{E}|m_H(X) - Y|^2 - \mathbb{E}|m(X) - Y|^2\right). \tag{78}$$

This has probability at most $|\mathcal{M}|$ times the probability that

$$(1+\delta)\left(\mathbb{E}(|m_h(X) - Y|^2|D) - \mathbb{E}|m(X) - Y|^2 - \frac{1}{n}\sum_{D'}(|Y_i - m_H(X_i)|^2 - |Y_i - m(X_i)|^2)\right) \tag{79}$$
$$\geq s + \delta\left(\mathbb{E}|m_h(X) - Y|^2 - \mathbb{E}|m(X) - Y|^2\right) \tag{80}$$

for some h, that is

$$\mathbb{E}[Z|D] - \frac{1}{n}\sum_i Z_i \geq \frac{s + \delta\mathbb{E}[Z|D]}{1+\delta} \tag{81}$$

for some h, where $Z = |m_h(X) - Y|^2 - |m(X) - Y|^2$. Now

$$\sigma^2 = \mathrm{Var}(Z|D) \leq \mathbb{E}[Z^2|D] \leq 16L^2 \int |m_h(x) - m(x)|^2 dP(x) = 16L^2\mathbb{E}[Z|D]. \tag{82}$$

Using this, and Bernstein's inequality,

$$P\left(\mathbb{E}[Z|D] - \overline{Z} \geq \frac{s + \delta\mathbb{E}[Z|D]}{1+\delta}|D\right) \tag{83}$$
$$\leq P\left(\mathbb{E}[Z|D] - \overline{Z} \geq \frac{s + \delta\sigma^2/(16L^2)}{1+\delta}|D\right) \tag{84}$$
$$\leq e^{-nA/B} \tag{85}$$

where

$$A = \frac{1}{(1+\delta)^2}\left(s + \frac{\delta\sigma^2}{16L^2}\right) \tag{86}$$

and

$$B = 2\sigma^2 + \frac{2\cdot 8L^2}{3(1+\delta)}\left(s + \frac{\delta\sigma^2}{16L^2}\right). \tag{87}$$

Now $A/B \geq s/c$ for $c = L^2(16/\delta + 35 + 196)$. So

$$P(T_1 \geq s|D) \leq |\mathcal{M}|e^{-ns/c}. \tag{88}$$

Finally

$$\mathbb{E}[T_1|D] \leq u + \int_u^\infty P(T_1 > s|D) \leq u + \frac{c|\mathcal{M}|}{n}e^{-nu/c}. \tag{89}$$

The result follows by setting $u = c\log|\mathcal{M}|/n$. □

7.2 Leave 1 Out Cross Validation

7.2.1 Generalized (Approximate) Cross Validation

7.2.2 Cp Statistic

7.3 K Fold Cross Validation

8 Information Criteria

In general, cross-validation requires a lot of training runs and therefore may be computationally infeasible. Therefore, various *information criterion* has been proposed to efficiently select a model.

References

[Arb10] J. Arbuthnot. An argument for divine providence. *Philosophical Transactions*, 27:186–190, 1710.