

3 Convergence

This topic will overview a variety of extremely powerful analysis results that span statistics, estimation theorem, and big data. It provides a framework to think about how to aggregate more and more data to get better and better estimates. It will cover the *Central Limit Theorem* (CLT), Chernoff-Hoeffding bounds, and Probably Approximately Correct (PAC) algorithms.

3.1 Sampling and Estimation

Most data analysis starts with some data set; we will call this data set P . It will be composed of a set of n data points $P = \{p_1, p_2, \dots, p_n\}$.

But underlying this data is almost always a very powerful assumption, that this data comes iid from a fixed, but usually unknown pdf, call this f . Lets unpack this: What does “iid” mean: Identically and Independently Distributed. The “identically” means each data point was drawn from the same f . The “independently” means that the first points have no bearing on the value of the next point.

Example: Polling

Consider a poll of $n = 1000$ likely voters in an upcoming election. If we assume each polled person is chosen iid, then we can use this to understand something about the underlying distribution f , for instance the distribution of all likely voters.

More generally, f could represent the outcome of a process, whether that is a randomized algorithm, a noisy sensing methodology, or the common behavior of a species of animals. In each of these cases, we essentially “poll” the process (algorithm, measurement, thorough observation) having it provide a sample, and repeat many times over.

Here we will talk about estimating the mean of f . To discuss this, we will now introduce a random variable $X \sim f$; a hypothetical new data point. The *mean* of f is the expected value of X : $\mathbf{E}[X]$.

We will estimate the mean of f using the *sample mean*, defined $\bar{P} = \frac{1}{n} \sum_{i=1}^n p_i$. The following diagram represents this common process: from a unknown process f , we consider n iid random variables $\{X_i\}$ corresponding to a set of n independent observations $\{p_i\}$, and take their average $\bar{P} = \frac{1}{n} \sum_{i=1}^n p_i$ to estimate the mean of f .

$$\bar{P} = \frac{1}{n} \sum \{p_i\} \xleftarrow{\text{realize}} \{X_i\} \underset{\text{iid}}{\sim} f$$

Central Limit Theorem. The central limit theorem is about how well the sample mean approximates the true mean. But to discuss the sample mean \bar{P} (which is a fixed value) we need to discuss random variables $\{X_1, X_2, \dots, X_n\}$, and their mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Note that again \bar{X} is a random variable. If we are to draw a new iid data set P' and calculate a new sample mean \bar{P}' it will likely not be exactly the same as \bar{P} ; however, the distribution of where this \bar{P}' is likely to be, is precisely \bar{X} . Arguably, this distribution is more important than \bar{P} itself.

There are many formal variants of the central limit theorem, but the basic form is as follows:

Central Limit Theorem: Consider n iid random variables X_1, X_2, \dots, X_n , where each $X_i \sim f$ for any fixed distribution f with mean μ and bounded variance σ^2 . Then $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ converges to the normal distribution with mean $\mu = \mathbf{E}[X_i]$ and variance σ^2/n .

Lets highlight some important consequences:

- For any f (that is not too crazy, since σ^2 is not infinite), then \bar{X} looks like a normal distribution.
- The mean of the normal distribution, which is the expected value of \bar{X} satisfies $\mathbf{E}[\bar{X}] = \mu$, the mean of f . This implies that we can use \bar{X} (and then also \bar{P}) as a guess for μ .
- As n gets larger (we have more data points) then the variance of \bar{X} (our estimator) decreases. So keeping in mind that although \bar{X} has the right expected value it also has some error, this error is decreasing as n increases.

```
# adapted from: https://github.com/mattnedrich/CentralLimitTheoremDemo
import random
import matplotlib as mpl
mpl.use('PDF')
import matplotlib.pyplot as plt

def plot_distribution(distribution, file, title, bin_min, bin_max, num_bins):
    bin_size = (bin_max - bin_min) / num_bins
    manual_bins = range(bin_min, bin_max + bin_size, bin_size)
    [n, bins, patches] = plt.hist(distribution, bins = manual_bins)
    plt.title(title)
    plt.xlim(bin_min, bin_max)
    plt.ylim(0, max(n) + 2)
    plt.ylabel("Frequency")
    plt.xlabel("Observation")
    plt.savefig(file, bbox_inches='tight')
    plt.clf()
    plt.cla()

minbin = 0
maxbin = 100
numbins = 50
nTrials = 1000

def create_uniform_sample_distribution():
    return range(maxbin)
sampleDistribution = create_uniform_sample_distribution()

# Plot the original population distribution
plot_distribution(sampleDistribution, 'output/SampleDistribution.pdf',
    "Population_Distribution", minbin, maxbin, numbins)

# Plot a sampling distribution for values of N = 2, 3, 10, and 30
n_vals = [2, 3, 10, 30]
for N in n_vals:
    means = []
    for j in range(nTrials):
        sampleSum = 0;
        for i in range(N):
            sampleSum += random.choice(sampleDistribution)
```

```

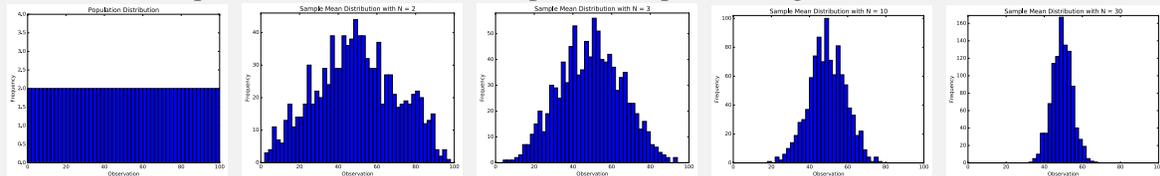
means.append(float(sampleSum) / float(N))

title = "Sample_Mean_Distribution_with_N=%s" % N
file = "output/CLT-demo-%s.pdf" % N
plot_distribution(means, file, title, minbin, maxbin, numbins)

```

Example: Central Limit Theorem

Consider f as a uniform distribution over $[0, 100]$. If we create n samples $\{p_1, \dots, p_n\}$ and their mean \bar{P} , then repeat this 1000 times, we can plot the output in histograms:



We see that starting at $n = 2$, the distributions look vaguely normal (in the technical sense of a normal distribution), and that their standard deviations narrow as n increases.

Remaining Mysteries. There should still be at least a few aspects of this not clear yet: (1) What does “convergence” mean? (2) How do we formalize or talk about this notion of error? (3) What does this say about our data \bar{P} ?

First, *convergence* refers to what happens as some parameter increases, in this case n . As the number of data points increase, as n “goes to infinity” then the above statement (\bar{X} looks like a normal distribution) becomes more and more true. For small n , the distribution may not quite look like a normal, it may be more bumpy, maybe even multi-modal. The statistical definitions of convergence are varied, and we will not go into them here, we will instead replace it with more useful phrasing in explaining aspects (2) and (3).

Second, the error now has two components. We cannot simply say that \bar{P} is at most some distance ε from μ . Something crazy might have happened (the sample is random after all). And it is not useful to try to write the probability that $\bar{P} = \mu$; for equality in continuous distributions, this probability is indeed 0. But we can combine these notions. We can say the distance between \bar{P} and μ is more than ε , with probability at most δ . This is called “probably approximately correct” or PAC.

Third, we want to generate some sort of PAC bound (which is far more useful than “ \bar{X} looks kind of like a normal distribution”). Whereas a frequentist may be happy with a confidence interval and a Bayesian a normal posterior, these two options are not directly available since again, \bar{X} is not exactly a normal. So we will discuss some very common *concentration of measure* tools. These don’t exactly capture the shape of the normal distribution, but provide upper bounds for its tails, and will allow us to state PAC bounds.

3.2 Probably Approximately Correct (PAC)

We will introduce shortly the three most common concentration of measure bounds, which provide increasingly strong bounds on the tails of distributions, but require more and more information about the underlying distribution f . Each provides a PAC bound of the following form:

$$\Pr[|\bar{X} - \mathbf{E}[\bar{X}]| \geq \varepsilon] \leq \delta.$$

That is, the probability that \bar{X} (which is some random variable, often a sum of iid random variables) is further than ε to its expected value (which is μ , the expected value of f where $X_i \sim f$), is at most δ . Note we do not try to say this probability is *exactly* δ , this is often too hard. In practice there are a variety of tools, and a user may try each one, and see which ones gives the best bound.

It is useful to think of ε as the *error tolerance* and δ as the *probability of failure*, i.e., that we exceed the error tolerance. However, often these bounds will allow us to write the required sample size n in terms of ε and δ . This allows us to trade these two terms off for any fixed known n ; we can guarantee a smaller error tolerance if we are willing to allow more probability of failure, and vice-versa.

3.3 Concentration of Measure

We will formally describe these bounds, and give some intuition of why they are true (but not proofs). But what will be the most important is what they imply. If you just know the distance of the expectation from the minimal value, you can get a very weak bound. If you know the variance of the data, you can get a stronger bound. If you know that the distribution f has a small and bounded range, then you can make the probability of failure (the δ in PAC bounds) very very small.

Markov Inequality. Let X be a random variable such that $X \geq 0$, that is it cannot take on negative values. Then for any parameter $\alpha > 0$

$$\Pr[X > \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$

Note this is a PAC bound with $\varepsilon = \alpha - \mathbf{E}[X]$ and $\delta = \mathbf{E}[X]/\alpha$, or we can rephrase this bound as follows: $\Pr[X - \mathbf{E}[X] > \varepsilon] \leq \delta = \mathbf{E}[X]/(\varepsilon + \mathbf{E}[X])$.

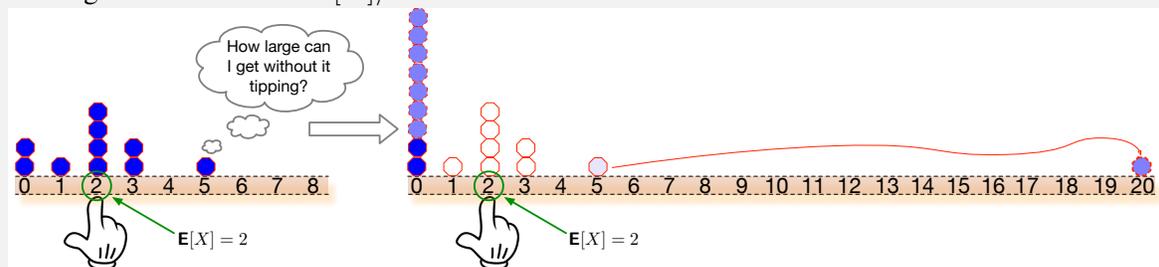
Geometry of the Markov Inequality

Consider balancing the pdf of some random variable X on your finger at $\mathbf{E}[X]$, like a waitress balances a tray. If your finger is not under a value μ so $\mathbf{E}[X] = \mu$, then the pdf (and the waitress's tray) will tip, and fall in the direction of μ – the “center of mass.”

Now for some amount of probability α , how large can we increase its location so we retain $\mathbf{E}[X] = \mu$. For each part of the pdf we increase, we must decrease some in proportion. However, by the assumption $X \geq 0$, the pdf must not be positive below 0. In the limit of this, we can set $\Pr[X = 0] = 1 - \alpha$, and then move the remaining α probability as large as possible, to a location δ so $\mathbf{E}[X] = \mu$. That is

$$\mathbf{E}[X] = 0 \cdot \Pr[X = 0] + \delta \cdot \Pr[X = \delta] = 0 \cdot (1 - \alpha) + \delta \cdot \alpha = \delta \cdot \alpha.$$

Solving for δ we find $\delta = \mathbf{E}[X]/\alpha$.

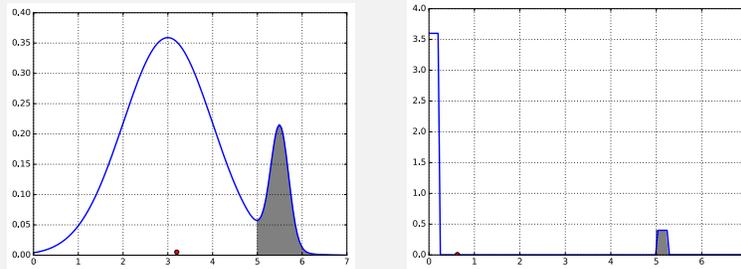


Imagine having 10 α -balls each representing $\alpha = 1/10$ th of the probability mass. As in the figure, if these represent a distribution with $\mathbf{E}[X] = 2$ and this must stay fixed, how far can one ball increase if all others balls must take a value at least 0? One ball can move to 20.

If we instead know that $X \geq b$ for some constant b (instead of $X \geq 0$), then we state more generally $\Pr[X > \alpha] \leq (\mathbf{E}[X] - b)/(\alpha - b)$.

Example: Markov Inequality

Consider the pdf f drawn in blue in the following figures, with $\mathbf{E}[X]$ for $X \sim f$ marked as a red dot. The probability that X is greater than 5 (e.g. $\Pr[X \geq 5]$) is the shaded area.



Notice that in both cases that $\Pr[X \geq 5]$ is about 0.1. This is the quantity we want to bound by above by δ . But since $\mathbf{E}[X]$ is much larger in the first case (about 2.25), then the bound $\delta = \mathbf{E}[X]/\alpha$ is much larger, about 0.45. In the second case, $\mathbf{E}[X]$ is much smaller (about 0.6) so we get a much better bound of $\delta = 0.12$.

Example: Markov Inequality and Coffee

Let C be a random variable describing the number liters of coffee the faculty at Data University will drink in a week. Say we know $\mathbf{E}[C] = 20$.

We use the Markov Inequality to bound the probability that the coffee consumed will be more than 50 liters as

$$\Pr[C \geq 50] \leq \frac{\mathbf{E}[C]}{50} = \frac{20}{50} = 0.4$$

Hence, based on the expected value alone, we can bound the with probability less than 0.4, the faculty at DU will drink less than 50 liters of coffee.

Chebyshev Inequality. Now let X be a random variable where we know $\mathbf{Var}[X]$, and $\mathbf{E}[X]$. Then for any parameter $\varepsilon > 0$

$$\Pr[|X - \mathbf{E}[X]| \geq \varepsilon] \leq \frac{\mathbf{Var}[X]}{\varepsilon^2}.$$

Again, this clearly is a PAC bound with $\delta = \mathbf{Var}[X]/\varepsilon^2$. This bound is typically stronger than the Markov one since δ decreases quadratically in ε instead of linearly.

Example: Chebyshev Inequality and Coffee

Again let C be a random variable for the liters of coffee that faculty at Data University will drink in a week with $\mathbf{E}[C] = 20$. If we also know that the variance is not too large, specifically $\mathbf{Var}[C] = 9$ (liters squared), then we can apply the Chebyshev inequality to get an improved bound.

$$\Pr[C \geq 50] \leq \Pr[|C - \mathbf{E}[C]| \geq 30] \leq \frac{\mathbf{Var}[C]}{30^2} = \frac{9}{900} = 0.01$$

That is, by using the expectation ($\mathbf{E}[C] = 20$) and variance ($\mathbf{Var}[C] = 9$) we can reduce the probability of exceeding 50 liters to at most probability 0.01.

Note that in the first inequality we convert from a one-sided expression $C \geq 50$ to a two-sided expression $|C - \mathbf{E}[C]| \geq 30$ (that is either $C - \mathbf{E}[C] \geq 30$ or $\mathbf{E}[C] - C \geq 30$). This is a bit wasteful, and stricter one-sided variants of Chebyshev inequality exist; we will not discuss these here in an effort for simplicity.

Recall that for an average of random variables $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$, where the X_i s are iid, and have variance σ^2 , then $\mathbf{Var}[\bar{X}] = \sigma^2/n$. Hence

$$\Pr[|\bar{X} - \mathbf{E}[X_i]| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Consider now that we have input parameters ε and δ , our desired error tolerance and probability of failure. If can draw $X_i \sim f$ (iid) for an unknown f (with known expected value and variance σ), then we can solve for how large n needs to be: $n = \sigma^2/(\varepsilon^2\delta)$.

Since $\mathbf{E}[\bar{X}] = \mathbf{E}[X_i]$ for iid random variables X_1, X_2, \dots, X_n , there is not a similar meaningfully-different extension for the Markov inequality.

Chernoff-Hoeffding Inequality. Following the above extension of the Chebyshev inequality, we can consider a set of n iid random variables X_1, X_2, \dots, X_n where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Now assume we know that each X_i lies in a bounded domain $[b, t]$, and let $\Delta = t - b$. Then for any parameter $\varepsilon > 0$

$$\Pr[|\bar{X} - \mathbf{E}[\bar{X}]| > \varepsilon] \leq 2 \exp\left(\frac{-2\varepsilon^2 n}{\Delta^2}\right).$$

Again this is a PAC bound, now with $\delta = 2 \exp(-2\varepsilon^2 n/\Delta^2)$. For a desired error tolerance ε and failure probability δ , we can set $n = (\Delta^2/(2\varepsilon^2)) \ln(2/\delta)$. Note that this has a similar relationship with ε as the Chebyshev bound, but the dependence of n on δ is exponentially less for this bound.

Relating this all back to the Gaussian distribution in the CLT, the Chebyshev bound only uses the variance information about the Gaussian, but the Chernoff-Hoeffding bound uses all of the “moments”: that it decays exponentially.

These are the most basic and common PAC concentration of measure bounds, but are by no means exhaustive.

Example: Uniform Distribution

Consider a random variable $X \sim f$ where $f(x) = \{\frac{1}{2} \text{ if } x \in [0, 2] \text{ and } 0 \text{ otherwise.}\}$, i.e, the Uniform distribution on $[0, 2]$. We know $\mathbf{E}[X] = 1$ and $\mathbf{Var}[X] = \frac{1}{3}$.

- Using the Markov Inequality, we can say $\mathbf{Pr}[X > 1.5] \leq 1/(1.5) \approx 0.6666$ and $\mathbf{Pr}[X > 3] \leq 1/3 \approx 0.33333$.
or $\mathbf{Pr}[X - \mu > 0.5] \leq \frac{2}{3}$ and $\mathbf{Pr}[X - \mu > 2] \leq \frac{1}{3}$.
- Using the Chebyshev Inequality, we can say that $\mathbf{Pr}[|X - \mu| > 0.5] \leq (1/3)/0.5^2 = \frac{4}{3}$ (which is meaningless). But $\mathbf{Pr}[|X - \mu| > 2] \leq (1/3)/(2^2) = \frac{1}{12} \approx 0.08333$.

Now consider a set of $n = 100$ random variables X_1, X_2, \dots, X_n all drawn iid from the same pdf f as above. Now we can examine the random variable $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. We know that $\mu_n = \mathbf{E}[\bar{X}] = \mu$ and that $\sigma_n^2 = \mathbf{Var}[\bar{X}] = \sigma^2/n = 1/(3n) = 1/300$.

- Using the Chebyshev Inequality, we can say that $\mathbf{Pr}[|\bar{X} - \mu| > 0.5] \leq \sigma_n^2/(0.5)^2 = \frac{1}{75} \approx 0.01333$, and $\mathbf{Pr}[|\bar{X} - \mu| > 2] \leq \sigma_n^2/2^2 = \frac{1}{1200} \approx 0.0008333$.
- Using the Chernoff-Hoeffding bound, we can say that $\mathbf{Pr}[|\bar{X} - \mu| > 0.5] \leq 2 \exp(-2(0.5)^2 n / \Delta^2) = 2 \exp(-100/8) \approx 0.0000074533$, and $\mathbf{Pr}[|\bar{X} - \mu| > 2] \leq 2 \exp(-2(2)^2 n / \Delta^2) = 2 \exp(-200) \approx 2.76 \cdot 10^{-87}$.

Exercises

Q3.1: Consider a pdf f so that a random variable $X \sim f$ has expected value $\mathbf{E}[X] = 3$ and variance $\mathbf{Var}[X] = 10$. Now consider $n = 10$ iid random variables X_1, X_2, \dots, X_{10} drawn from f . Let $\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i$.

1. What is $\mathbf{E}[\bar{X}]$?
2. What is $\mathbf{Var}[\bar{X}]$?
3. What is the standard deviation of \bar{X} ?
4. Which is larger $\mathbf{Pr}[X > 4]$ or $\mathbf{Pr}[\bar{X} > 4]$?
5. Which is larger $\mathbf{Pr}[X > 2]$ or $\mathbf{Pr}[\bar{X} > 2]$?

Q3.2: Let X be a random variable that you know is in the range $[-1, 2]$ and you know has expected value of $\mathbf{E}[X] = 0$. Use the Markov Inequality to upper bound $\mathbf{Pr}[X > 1.5]$?
(Hint: you will need to use a change of variables.)

Q3.3: Consider a pdf f so that a random variable $X \sim f$ has expected value $\mathbf{E}[X] = 5$ and variance $\mathbf{Var}[X] = 100$. Now consider $n = 16$ iid random variables X_1, X_2, \dots, X_{16} drawn from f . Let $\bar{X} = \frac{1}{16} \sum_{i=1}^{16} X_i$.

1. What is $\mathbf{E}[\bar{X}]$?
2. What is $\mathbf{Var}[\bar{X}]$?

Assume we know that X is never smaller than 0 and never larger than 20.

3. Use the Markov inequality to upper bound $\mathbf{Pr}[\bar{X} > 8]$.
4. Use the Chebyshev inequality to upper bound $\mathbf{Pr}[\bar{X} > 8]$.
5. Use the Chernoff-Hoeffding inequality to upper bound $\mathbf{Pr}[\bar{X} > 8]$.
6. If we increase n to 100, how will the above three bounds be affected.

Q3.4: Consider a (parked) self-driving car that returns n iid estimates to the distance of a tree. We will model these n estimates as a set of n scalar random variables X_1, X_2, \dots, X_n taken iid from an unknown pdf f , which we assume models the true distance plus unbiased noise. (The sensor can take many iid estimates in rapid fire fashion.) The sensor is programmed to only return values between 0 and 20 feet, and that the variance of the sensing noise is 64 feet squared. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. We want to understand as a function of n how close \bar{X} is to μ , which is the true distance to the tree.

1. Use Chebyshev's Inequality to determine a value n so that $\mathbf{Pr}[|\bar{X} - \mu| \geq 1] \leq 0.5$.
2. Use Chebyshev's Inequality to determine a value n so that $\mathbf{Pr}[|\bar{X} - \mu| \geq 0.1] \leq 0.1$.
3. Use the Chernoff-Hoeffding bound to determine a value n so that $\mathbf{Pr}[|\bar{X} - \mu| \geq 1] \leq 0.5$.
4. Use the Chernoff-Hoeffding bound to determine a value n so that $\mathbf{Pr}[|\bar{X} - \mu| \geq 0.1] \leq 0.1$.

Q3.5: Consider two random variables C and T describing how many coffees and teas I will buy in the coming week; clearly neither can be smaller than 0. Based on personal experience, I know the following summary statistics about my coffee and tea buying habits: $\mathbf{E}[C] = 3$ and $\mathbf{Var}[C] = 1$ also $\mathbf{E}[T] = 2$ and $\mathbf{Var}[T] = 5$.

1. Use Markov's Inequality to upper bound the probability I buy 4 or more coffees, and the same for teas: $\mathbf{Pr}[C \geq 4]$ and $\mathbf{Pr}[T \geq 4]$.

2. Use Chebyshev's Inequality to upper bound the probability I buy 4 or more coffees, and the same for teas: $\Pr[C \geq 4]$ and $\Pr[T \geq 4]$.

Q3.6: The average score on a test is 82 with a standard deviation of 4 percentage points. All tests have scores between 0 and 100.

1. Using Chebyshev's inequality, at least what percentage of the tests have a grade of at least 70 and at most 94?
2. Using Markov's inequality, what is the highest percentage of tests which could have score less than 60?