# 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 $P = \{p_1, p_2, \ldots, p_n\}$.

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**
This assumption allows us to try to understand something about $f$. This pdf $f$ could represent a much larger population (all people who will vote) where the data set $P$ are the results of a poll of $n = 1000$ people.
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, through 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 $\bar{P} = \frac{1}{n}\sum_{i=1}^{n} p_i$.

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

**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 $\{X_1, X_2, \ldots, X_n\}$, 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 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 $X_1, X_2, \ldots, X_n$, where each $X_i \sim f$ for any fixed distribution $f$ with mean $\mu$ and bounded variance $\sigma^2$. 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 $\sigma^2/n$.

Lets highlight some important consequences:

- For any $f$ (that is not too crazy, since $\sigma^2$ is not infinite), then $\bar{X}$ looks like a normal distribution.

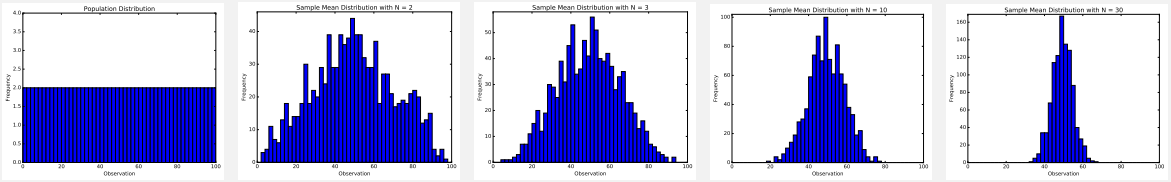
- The mean of the normal distribution, which is the expected value of $\bar{X}$ satisfies $\mathbf{E}[\bar{X}] = \mu$, the mean of $f$. This implies that we can use $\bar{X}$ (and then also $\bar{P}$) as a guess for $\mu$.

- 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 $\{p_1, \ldots, p_n\}$ and their mean $\bar{P}$, then repeat this 1000 times, we can plot the output in a histograms:




We see that starting at $n = 2$, the distributions look vaguely like Gaussians, and that their standard deviation narrows 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 definition of converge is a bit nebulous, and we will not go into it 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 $\varepsilon$ from $\mu$. 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 $\mu$ is more than $\varepsilon$, with probability at most $\delta$. 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:

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

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

I like to think of $\varepsilon$ as the *error tolerance* and $\delta$ 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 $\varepsilon$ and $\delta$. This allows us to trade these two terms off for any fixed known $n$; we can allow less 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 $\delta$ 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 $\alpha > 0$

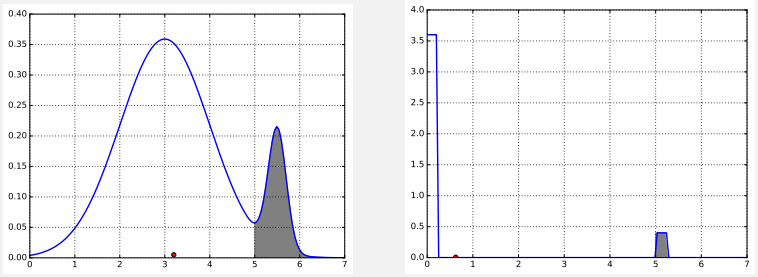
$$\mathbf{Pr}[X > \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$

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: $\mathbf{Pr}[X - \mathbf{E}[X] > \varepsilon] \leq \delta = \mathbf{E}[X]/(\varepsilon + \mathbf{E}[X])$.

This bound can be seen true by considering balancing the pdf of $X$ at $\mathbf{E}[X]$. Then, more than $\alpha$ fraction of its mass cannot be greater than $\mathbf{E}[X]/\alpha$ since the rest is at least 0, and it cannot balance.

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

**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. $\mathbf{Pr}[X \geq 5]$) is the shaded area.



Notice that in both cases that $\mathbf{Pr}[X \geq 5]$ is about 0.1. This is the quantity we want to bound by above by $\delta$. 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 $\delta = 0.12$.

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

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

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

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

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

Consider now that we have input parameters $\varepsilon$ and $\delta$, 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 $\sigma$), then we can solve for how large $n$ needs to be: $n = \sigma^2/(\varepsilon^2\delta)$.

**Chernoff-Hoeffding Inequality.** Following the above example, we can consider a set of $n$ iid random variables $X_1, X_2, \ldots, X_n$ 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 $\varepsilon > 0$

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

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

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

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

Now consider a set of $n = 100$ random variables $X_1, X_2, \ldots, X_n$ 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 $\mu_n = \textbf{E}[\bar{X}] = \mu$ and that $\sigma_n^2 = \textbf{Var}[\bar{X}] = \sigma^2/n = 1/(3n) = 1/300$.

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