The more we know about a distribution, the stronger concentration inequality we can derive. We know that Markov’s inequality is weak, since we only use the expectation of a random variable to get the probability bound. Chebyshev’s inequality is a bit stronger, because we incorporate the variance into the probability bound. However, as we showed in the example in 6.1, these bounds are still pretty “loose”. (They are tight in some cases though).

What if we know even more? In particular, its MGF? That will allow us to have an even stronger bound. The Chernoff bound is derived using a combination of Markov’s inequality and moment generating functions.

### 6.2.1 The Chernoff Bound for the Binomial Distribution

Here is the idea for the Chernoff bound. We will only derive it for the Binomial distribution, but the same idea can be applied to any distribution.

Let $X$ be any random variable. $e^{tX}$ is always a non-negative random variable. Thus, for any $t > 0$, using Markov’s inequality and the definition of MGF:

$$
P(X \geq k) = \mathbb{P}(e^{tX} \geq e^{tk}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tk}} = \frac{M_X(t)}{e^{tk}} \quad \text{[Markov’s inequality]}$$

(Note that the first line requires $t > 0$, otherwise it would change to $\mathbb{P}(e^{tX} \leq e^{tk})$. This is because $e^t > 1$ for $t > 0$ so we get something like $2^t$ which is monotone increasing. If $t < 0$, then $e^t < 1$ so we get something like $0.3^X$ which is monotone decreasing.)

Now the right hand side holds for uncountably infinitely many $t$. For example, if we plugged in $t = 0.5$ we might get $\frac{M_X(t)}{e^{tk}} = 0.53$ and if we plugged in $t = 3.26$ we might get 0.21. Since $\mathbb{P}(X \geq k)$ has to be less than all the possible values when plugging in different $t > 0$, it in particular must be less than the minimum of all the values.

$$P(X \geq k) \leq \min_{t>0} \left( \frac{M_X(t)}{e^{tk}} \right)$$

This is good - if we can minimize the right hand side, we can get a very tight/strong bound.

We’ll now focus our attention to deriving the Chernoff bound when $X$ has a binomial distribution. Everything above applies generally though.
### Definition 6.2.1.1: Chernoff Bound for Binomial Distribution

Let $X \sim \text{Bin}(n, p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:

**Upper tail bound:**
$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp \left( -\frac{\delta^2 \mu}{3} \right)$$

**Lower tail bound:**
$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp \left( -\frac{\delta^2 \mu}{2} \right)$$

where $\exp(x) = e^x$.

### Proof of Chernoff Bound for Binomial

If $X = \sum_{i=1}^n X_i$ where $X_1, X_2, \ldots, X_n$ are iid variables, then since the MGF of the (independent) sum equals the product of the MGFs:
$$\mathbb{P}(X \geq k) \leq \min_{t > 0} \frac{M_X(t)}{e^{tk}} = \min_{t > 0} \frac{\prod_{i=1}^n M_{X_i}(t)}{e^{tk}}$$

Let’s derive a Chernoff bound for $X \sim \text{Bin}(n, p)$, which has the form $\mathbb{P}(X \geq (1 + \delta)\mu)$ for $\delta > 0$. For example with $\delta = 4$, you may want to bound $\mathbb{P}(X \geq 5\mathbb{E}[X])$.

Recall $X = \sum_{i=1}^n X_i$ where $X_i \sim \text{Ber}(p)$ are iid, with $\mu = \mathbb{E}[X] = np$.

$$M_{X_i}(t) = \mathbb{E}[e^{tX_i}] = e^{t1_pX_i(1) + t0_pX_i(0)} [\text{def of MGF}]$$
$$= pe^t + (1 - p) [\text{LOTUS}]$$
$$= 1 + p(e^t - 1)$$
$$\leq e^{p(e^t - 1)} [1 + x \leq e^x]$$

See here for a pictorial proof that $1 + x \leq e^x$ for any real number $x$ (just plot the two functions). Alternatively, use the Taylor series for $e^x$ to argue this.

---

#### Graph

![Graph](image-url)
Now using the result from earlier and plugging in the MGF for the $Ber(p)$ distribution, we get:

$$P(X \geq k) \leq \min_{t>0} \frac{\prod_{i=1}^{n} M_{X_i}(t)}{e^{tk}}$$

[from earlier]

$$\leq \min_{t>0} \frac{e^{np(t-1)}}{e^{tk}}$$

[MGF of $Ber(p)$, $n$ times]

$$= \min_{t>0} \frac{e^{np(t-1)}}{e^{tk}}$$

[algebra]

$$= \min_{t>0} \frac{e^{np(t-1)}}{e^{tk}}$$

[$\mu = np$]

For our bound, we want something like $P(X \geq (1 + \delta)\mu)$, so our $k = (1 + \delta)\mu$. To minimize the RHS and get the tightest bound, the best bound we get is by choosing $t = \ln(1 + \delta)$ after some terrible algebra. We simply plug in $k$ and our optimal value of $t$ to the above equation:

$$P(X \geq (1 + \delta)\mu) \leq \frac{e^{\mu(\ln(1+\delta)-1)}}{e^{(1+\delta)\mu(\ln(1+\delta))}} = \frac{e^{\mu(\ln(1+\delta)-1)}}{(1+\delta)^{(1+\delta)\mu}} = \left(\frac{e^{\delta}}{(1 + \delta)^{(1+\delta)}}\right)$$

Again, we wanted to choose $t$ that minimizes our upper bound for the tail probability. Taking the derivative with respect to $t$ tells us we should plug in $t = \ln(1 + \delta)$ to minimize that quantity. This would actually be pretty annoying to plug into a calculator.

We actually can show that the final RHS is $\leq \exp\left(\frac{-\delta^2\mu}{2 + \delta}\right)$ with some more messy algebra. Additionally, if we restrict $0 < \delta < 1$, we can simplify this even more to the bound provided earlier:

$$P(X \geq (1 + \delta)\mu) \leq \exp\left(\frac{-\delta^2\mu}{3}\right)$$

The proof of the lower tail is entirely analogous, except optimizing over $t < 0$ when the inequality flips. It proceeds by taking $t = \ln(1 - \delta)$.

We also get a lower tail bound:

$$P(X \leq (1 - \delta)\mu) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^\mu \leq \left(\frac{e^{-\delta}}{e^{-\delta + \frac{\delta}{2}}}\right)^\mu = \exp\left(\frac{-\delta^2\mu}{2}\right)$$

Wait you may wonder, why are we bounding $P(X \geq (1 + \delta)\mu)$, when we can just sum the PMF of a binomial to get an exact answer? The reason is, it is very computationally expensive to compute the binomial PMF! For example, if $X \sim Bin(n = 20000, p = 0.1)$, then by plugging in the PMF, we get

$$P(X = 13333) = \frac{(20000)!}{13333!(20000 - 13333)!} \cdot 0.1^{13333} \cdot 0.9^{20000 - 13333} = 20000! \cdot 0.1^{13333} \cdot 0.9^{20000 - 13333}$$

(Actually, $n = 20000$ isn’t even that large.) You have to multiply 20,000 numbers on the second two terms, and it multiplies to a number that is infinitesimally small. For the first term (binomial coefficient), computing $20000!$ is impossible - in fact, it is so large you can’t even imagine. You would have to cleverly interleaving multiplying a factorial vs the probability, to keep the value in an acceptable range for the computer. Then, sum up a bunch of these....

This is why we have/need the Poisson approximation, the Normal approximation (CLT), and the Chernoff bound for the binomial!
We have:

\[ E[X] = np = 500 \cdot 0.2 = 150 \]
\[ \text{Var}(X) = np(1 - p) = 500 \cdot 0.2 \cdot 0.8 = 80 \]

Using Markov’s Inequality:

\[ P(X \geq 150) \leq \frac{E[X]}{150} = \frac{100}{150} \approx 0.6667 \]

Using the Chernoff Bound (with \( \delta = 0.5 \)):

\[ P(X \geq 150) = P(X \geq (1 + 0.5) \cdot 100) \leq e^{-\frac{0.5^2 \cdot 100}{3}} \approx 0.00024 \]

Examples

Suppose the number of red lights Alex encounters each day to work is on average 4.8 (according to historical trips to work). Alex really will be late if he encounters 8 or more red lights. Let \( X \) be the number of lights he gets on a given day.

1. Give a bound for \( P(X \geq 8) \) using Markov’s inequality.
2. Give a bound for \( P(X \geq 8) \) using Chebyshev’s inequality, if we also assume \( \text{Var}(X) = 2.88 \).
3. Give a bound for \( P(X \geq 8) \) using the Chernoff bound. Assume that \( X \sim Bin(12, 0.4) \) - that there are 12 traffic lights, and each is independently red with probability 0.4.
4. Compute \( P(X \geq 8) \) exactly using the assumption from the previous part.
5. Compare the three bounds and their assumptions.

1. Since \( X \) is nonnegative and we know its expectation, we can apply Markov’s inequality:

\[ P(X \geq 8) \leq \frac{E[X]}{8} = \frac{4.8}{8} = 0.6 \]

2. Since we know \( X \)’s variance, we can apply Chebyshev’s inequality after some manipulation. We have to do this to match the form required:

\[ P(X \geq 8) \leq P(X \geq 8) + P(X \leq 1.6) = P(|X - 4.8| \geq 3.2) \]

The reason we chose \( \leq 1.6 \) is so it looks like \( P(|X - \mu| \geq \alpha) \). Now, applying Chebyshev’s gives:

\[ \leq \frac{\text{Var}(X)}{3.2^2} = \frac{2.88}{3.2^2} = 0.28125 \]

3. Actually, \( X \sim Bin(12, 0.4) \) also has \( E[X] = np = 4.8 \) and \( \text{Var}(X) = np(1 - p) = 2.88 \) (what a coincidence). The Chernoff bound requires something of the form \( P(X \geq (1 + \delta)\mu) \), so we first need to solve for \( \delta \): \( (1 + \delta)4.8 = 8 \) so that \( \delta = 2/3 \). Now,

\[ P(X \geq 8) = P(X \geq (1 + 2/3)4.8) \leq \exp \left( -\frac{(2/3)^2 \cdot 4.8}{3} \right) \approx 0.4911 \]
4. The exact probability can be found summing the Binomial PMF:

\[
P(X \geq 8) = \sum_{k=8}^{12} \binom{12}{k} 0.4^k 0.6^{12-k} \approx 0.0573
\]

5. Actually it’s usually the case that the bounds are tighter/better as we move down the list Markov, Chebyshev, Chernoff. But in this case Chebyshev’s gave us the tightest bound, even after being weakened by including some additional \( P(X \leq 1.6) \). Chernoff bounds will typically be better for farther tails - 8 isn’t considered too far from the mean 4.8.

It’s also important to note that we found out more information progressively - we can’t blindly apply all these inequalities every time. We need to make sure the conditions for the bound being valid are satisfied.

Even our best bound of 0.28125 was 5-6x larger than the true probability of 0.0573.