Problem 1. **Midterm** Solve all of the problems on the midterm again (including the ones you got correct).

Solution 1. See the midterm solutions.

Problem 2. **Coupon Collector Convergence**
Recall the coupon collector problem: for a positive integer \( n \), there are \( n \) different coupons, and you are trying to collect them all. Each time you purchase an item, you receive one of the \( n \) coupons uniformly at random. Let \( T_n \) denote the amount of time it takes to collect all \( n \) coupons.

Prove that \( \frac{T_n}{n \ln n} \to 1 \) in probability as \( n \to \infty \).

Solution 2. Recall that from the analysis of the coupon collector problem, we have

\[
\mathbb{E}[T_n] = n \sum_{i=1}^{n} \frac{1}{i} = nH_n,
\]

where \( H_n = \sum_{i=1}^{n} i^{-1} \) is the harmonic sum. We also need to estimate the variance of \( T_n \) for this problem. Let \( X_i \) denote the amount of time required to collect the \( i \)th new coupon, so that \( T_n = \sum_{i=1}^{n} X_i \). The \( X_i \) are independent, so \( \text{var} T_n = \sum_{i=1}^{n} \text{var} X_i \), where \( X_i \) is a geometric random variable with probability \( p = 1 - (i - 1)/n \). Hence, \( \text{var} X_i = (1 - p)/p^2 \leq p^{-2} \), which gives

\[
\text{var} T_n \leq \sum_{i=1}^{n} \left(1 - \frac{i-1}{n}\right)^{-2} = n^2 \sum_{i=1}^{n} \frac{1}{i^2} \leq n^2 \sum_{i=1}^{\infty} \frac{1}{i^2}.
\]

It is a well-known fact that \( \sum_{i=1}^{\infty} i^{-2} = \pi^2/6 \), but we don’t need precise bounds for our purposes; it suffices to note that the summation converges. Notably,

\[
\text{var} \left( \frac{T_n - nH_n}{n \ln n} \right) \leq \frac{n^2}{n^2(\ln n)^2} \sum_{i=1}^{\infty} \frac{1}{i^2} \to 0 \quad \text{as } n \to \infty,
\]

and now Chebyshev’s Inequality gives

\[
\Pr \left( \left| \frac{T_n - nH_n}{n \ln n} \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} \cdot \frac{n^2}{n^2(\ln n)^2} \sum_{i=1}^{\infty} \frac{1}{i^2} \to 0 \quad \text{as } n \to \infty
\]

for every \( \varepsilon > 0 \). Hence, \( (T_n - nH_n)/(n \ln n) \to 0 \) in probability as \( n \to \infty \). To conclude, we note that \( H_n \sim \ln n \) asymptotically. Hence, \( T_n/(n \ln n) \to 1 \) in probability as \( n \to \infty \).
Remark. From our previous analysis of the coupon collector problem, we know that $E[T_n]$ is close to $n \ln n$, so we have shown a result which is similar in spirit to a “weak law of large numbers for the coupon collector problem”: as $n \to \infty$, $T_n$ is “close” to its expected value. However, since we are not dealing with i.i.d. random variables, we cannot use the variant of the WLLN proved in lecture to deal with this problem.

Problem 3. Two-Population Sampling
We are conducting a public opinion poll to determine the fraction $p$ of people who will vote for Mr. Whatshisname as the next president. We ask $N_1$ college-educated and $N_2$ non-college-educated people, where $N_1$ and $N_2$ are positive integers. We assume that the votes in each of the two groups are i.i.d. $Ber(p_1)$ and $Ber(p_2)$, respectively in favor of Whatshisname. In the general population, the percentage of college-educated people is known to be $q$.

(a) What is a 95% confidence interval for $p$, using an upper bound for the variance?

(b) How do we choose $N_1$ and $N_2$ subject to $N_1 + N_2 = N$ to minimize the width of that interval? (You may ignore the constraint that $N_1$ and $N_2$ must be integers.)

Solution 3. (a) If we let $\hat{p}_1$ and $\hat{p}_2$ denote the fraction of people who vote for Mr. Whatshisname in the two groups respectively, then an unbiased estimator for $p$ is $\hat{p} := q\hat{p}_1 + (1-q)\hat{p}_2$. The variance of $\hat{p}$ is

$$\text{var} \hat{p} = \frac{q^2 p_1(1-p_1)}{N_1} + \frac{(1-q)^2 p_2(1-p_2)}{N_2} \leq \frac{1}{4} \left( \frac{q^2}{N_1} + \frac{(1-q)^2}{N_2} \right).$$

So, an approximate 95% confidence interval for $p$, using the CLT, is $\hat{p} \pm \sqrt{\text{var} \hat{p}} = \hat{p} \pm \sqrt{\frac{q^2}{N_1} + \frac{(1-q)^2}{N_2}}$.

Note: Strictly speaking, the version of the CLT taught in this course does not apply to this problem because the random variables are not i.i.d. (they do not all have the same distribution). However, there are other variants of the CLT (notably Lindeberg’s CLT) which do hold even when the random variables do not all have the same distribution (however, independence is crucial). For this problem, the CLT is being used as an approximation anyway, so you can take the approximation on faith.

(b) Minimizing the width of the interval is equivalent to minimizing the variance. We can explicitly enforce the constraint by writing $N_2 = N - N_1$, and then we have:

$$\frac{d}{dN_1} \left( \frac{q^2}{N_1} + \frac{(1-q)^2}{N - N_1} \right) = -\frac{q^2}{N_1^2} + \frac{(1-q)^2}{(N-N_1)^2}.$$ 

The second derivative is positive so the function is convex, and so the first-order condition tells us the minimizer. Setting the derivative to 0, we find that $q/N_1 = (1-q)/(N-N_1)$. Therefore, the minimizer is $N_1 = qN$, $N_2 = (1-q)N$. 

2
**Problem 4. Confidence Intervals: Chebyshev vs. Chernoff vs. CLT**

Let $X_1, \ldots, X_n$ be i.i.d. $\text{Ber}(q)$ random variables, with common mean $\mu = \mathbb{E}[X_1] = q$ and variance $\sigma^2 = \text{var}(X_1) = q(1-q)$. We want to estimate the mean $\mu$, and towards this goal we use the sample mean estimator

$$X_n \triangleq \frac{X_1 + \cdots + X_n}{n}.$$

Given some confidence level $a \in (0,1)$ we want to construct a confidence interval around $X_n$ such that $\mu$ lies in this interval with probability at least $1-a$.

1. Use Chebyshev’s inequality in order to show that $\mu$ lies in the interval

$$\left( X_n - \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}}, X_n + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)$$

with probability at least $1-a$.

2. Recall that in Homework 4 you showed that

$$\Pr(|\bar{X}_n - q| \geq \epsilon) \leq 2e^{-2n\epsilon^2}, \quad \text{for any } \epsilon > 0.$$

Use this inequality in order to show that $\mu$ lies in the interval

$$\left( \bar{X}_n - \frac{1}{\sqrt{n}} \sqrt{\frac{2 \ln \frac{2}{a}}{a}}, \bar{X}_n + \frac{1}{\sqrt{n}} \sqrt{\frac{2 \ln \frac{2}{a}}{a}} \right)$$

with probability at least $1-a$.

3. Show that if $Z \sim \mathcal{N}(0,1)$, then

$$\Pr(|Z| \geq \epsilon) \leq 2e^{-\epsilon^2}, \quad \text{for any } \epsilon > 0.$$

4. Use the Central Limit Theorem, and Part (c) in order to heuristically argue that $\mu$ lies in the interval

$$\left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \sqrt{\frac{2 \ln \frac{2}{a}}{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \sqrt{\frac{2 \ln \frac{2}{a}}{a}} \right)$$

with probability at least $1-a$.

5. Compare the three confidence intervals.

**Solution 4.** 1. Rewrite the probability that $\mu$ lies in the specified interval as the probability that $X_n$ lies in an interval of the same width around $\mu$:

$$\Pr\left\{ \mu \in \left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right) \right\} = \Pr\left( |\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)$$

$$= 1 - \Pr\left( |\bar{X}_n - \mu| > \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)$$

$$\geq 1 - \frac{\text{var}X_n}{(\sigma^2/n)(1/a)} = 1 - a,$$

because $\text{var}X_n = \sigma^2/n$. 

3
2. Use the same idea as the previous part, but using the stronger tail inequality.

\[ \Pr\left\{ \mu \in \left( \bar{X}_n - \frac{1/2}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}}, \bar{X}_n + \frac{1/2}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \right\} \\
= \Pr\left( |\bar{X}_n - \mu| \leq \frac{1}{2\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \\
= 1 - \Pr\left( |\bar{X}_n - \mu| > \frac{1}{2\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \geq 1 - 2 \exp\left( - \ln \frac{2}{a} \right) = 1 - a. \]

3. For any \( t > 0 \) we have that

\[ \Pr(Z \geq \epsilon) = \Pr(tZ \geq t\epsilon) \\
= \Pr(e^{tZ} \geq e^{t\epsilon}) \\
\leq \frac{\mathbb{E}[e^{tZ}]}{e^{t\epsilon}} \\
= e^{\frac{1}{2}t^2 - t\epsilon}. \]

Optimizing over \( t > 0 \), yields

\[ \Pr(Z \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2}}. \]

The final result follows by a union bound.

4. From the CLT and the previous part we have that

\[ \Pr\left( \left| \frac{\bar{X}_n - \mu}{\sigma} \right| \geq \epsilon \right) \approx \Pr(|Z| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2}}. \]

We are going to set \( \epsilon \) to be such that \( a = 2e^{-\frac{\epsilon^2}{2}} \), which yields \( \epsilon = \sqrt{2\ln \frac{2}{a}} \).

Plugging in this value of \( \epsilon \) we have that

\[ \Pr\left( \left| \bar{X}_n - \mu \right| \geq \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \leq a, \]

or equivalently

\[ \Pr\left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} < \mu < \bar{X}_n + \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \\
= \Pr\left( -\frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} < \mu - \bar{X}_n < \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \\
= \Pr\left( |\bar{X}_n - \mu| < \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \geq 1 - a. \]

5. We can see that Chebyshev’s inequality and the CLT produce confidence intervals with standard deviation term \( \sigma \) present, while on the other hand using the Chernoff bound the standard deviation is replaced by \( 1/2 \), which is only an upper bound on \( \sigma \), since \( \sigma^2 = \sigma^2(q) = q(1-q) \leq 1/4. \)
Chebyshev’s inequality is able to capture the standard deviation term, but on the other hand it has a poor dependence of the form $1/\sqrt{a}$ on the confidence level $a$. Chernoff’s inequality and the CLT have a way better dependence on $a$ of the form $\sqrt{\ln \frac{2}{a}}$.

Finally, while the confidence intervals derived via Chebyshev’s and Chernoff’s inequality, are true/provable confidence intervals, we can only argue heuristically about the interval derived via the CLT.

**Problem 5.** [Bonus] The CLT Implies the WLLN

1. Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables. Show that if $X_n$ converges to $c$ in distribution, where $c$ is a constant, then $X_n \xrightarrow{P} c$.

2. Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of i.i.d. random variables, with mean $\mu$ and finite variance $\sigma^2$. Show that the CLT implies the WLLN.

**Solution 5.**

1. Since $X_n \xrightarrow{d} c$, we can deduce that for any $\epsilon > 0$, we have
   
   $$\lim_{n \to \infty} F_{X_n}(c - \epsilon) = 0,$$
   $$\lim_{n \to \infty} F_{X_n}\left(c - \frac{\epsilon}{2}\right) = 1.$$

   Using this fact we have that
   
   $$\lim_{n \to \infty} \Pr(|X_n - c| \geq \epsilon) = \lim_{n \to \infty} [\Pr(X_n \leq c - \epsilon) + \Pr(X_n \geq c + \epsilon)]$$
   $$= \lim_{n \to \infty} \Pr(X_n \leq c - \epsilon) + \lim_{n \to \infty} \Pr(X_n \geq c + \epsilon)$$
   $$= \lim_{n \to \infty} F_{X_n}(c - \epsilon) + \lim_{n \to \infty} \Pr(X_n \geq c + \epsilon)$$
   $$\leq 0 + \lim_{n \to \infty} \Pr(X_n > c + \frac{\epsilon}{2})$$
   $$= 1 - \lim_{n \to \infty} F_{X_n}\left(c + \frac{\epsilon}{2}\right)$$
   $$= 0.$$

   Therefore $\lim_{n \to \infty} \Pr(|X_n - c| \geq \epsilon) = 0$, for all $\epsilon > 0$ which means that $X_n \xrightarrow{P} c$.

2. From the CLT we know that
   
   $$\frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

   In addition $\frac{\sigma}{\sqrt{n}} \to 0$, so
   
   $$\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \xrightarrow{d} 0$$

   or stated another way
   
   $$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{d} \mu.$$
Finally using Part (a) we can conclude that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$