1 The Law of Large Numbers

A general rule in probability is that the aggregate behavior of random events become more predictable when we have many independent sources of randomness. Suppose that we run an experiment where we flip a coin n times. If we record the number of heads then we expect that

$$\frac{\#\text{ of heads}}{n} \approx \frac{1}{2}.$$

The more coins we flip, the closer the proportion of heads will be to 0.5. Of course, the outcomes are random so it might never be exactly 0.5, but one expects that somehow the events become less random the more experiments we run. This intuition can be made precise, and this is called the law of large numbers.

1.1 Law of Large Numbers

Consider i.i.d. random variables $X_1, \ldots, X_n$ with (finite) mean μ and variance σ^2.

Definition 1 (Sample Mean)**.** The average of $X_1, \ldots, X_n$ is called the *sample mean*, and is given by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Using the properties of the expected and variance of linear combinations, it follows that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu \quad \text{and} \quad \operatorname{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{\sigma^2}{n}.$$

This suggests that the variance of the sample mean shrinks as $n \to \infty$. The weak law of large number says that the probability the sample mean is not equal to its expected value goes to zero.

Theorem 1 (*Weak Law of Large Numbers*)

Let $X_1, \ldots, X_n$ be i.i.d. *with (finite) mean μ and finite variance. Then for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty.$$

In other words, $\bar{X}_n \to \mu$ in probability*.*

Remark 1. For course, this also means that

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1 - \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 1 \text{ as } n \to \infty.$$

In fact, we have a stronger statement that doesn't even need to assume that the variance is finite. The probability that the sample mean takes the value μ in the limit is exactly 1.

Theorem 2 (*Strong Law of Large Numbers*)

Let $X_1, \ldots, X_n$ be i.i.d. *with (finite) mean μ. Then*

$$\mathbb{P}(\{\omega : \bar{X}_n(\omega) \to \mu\}) = 1$$

In other words, $\bar{X}_n \to \mu$ almost surely*.*

Remark 2. There are many notions of the convergence of random variables. We have seen convergence in distribution already, and we have just introduced the notions of convergence in probability and almost sure convergence. We have that almost sure convergence implies convergence in probability, so it is a stronger notion of convergence. The reverse implication is not true in general.

Remark 3. The law of large numbers, says that long-run averages converges to the mean. This does not imply that if you flip several heads in a row, then things will even out with a long string of tails. This is called the *gambler's fallacy*.

1.2 Connection Between Tail Probabilities and Moments

The laws of large numbers are qualitative results since it guarantees convergence, but it does not tell us how precisely how close the random variables are for finite n. We now introduce some basic quantitative results, often called concentration inequalities.

We saw before that the distribution of the random variable determines its expected value. The expected value does not determine the distribution, but it can be used to bound probabilities of some events. In particular, the moments of the random variable can be used to control the probability that a random variable takes extreme outlying values away from its expected value.

Theorem 3 (*Markov's Inequality*)

For any random variable and constant $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{t}.$$

Remark 4. This is intuitive, because a small expected value means that the probability that $|X|$ is big should be small. Furthermore, the more extreme the outlier is (which is measured by t), the more rare it should be. For example, if X denotes income of a randomly selected person in a city, then

$$\mathbb{P}(X \geq 2\,\mathbb{E}[X]) \leq \frac{1}{2}$$

since we can't have more than half the people in the city making twice the average income.

This is a good first step to controlling the tail events since it only requires some information about the first moment $\mathbb{E}[|X|]$. However, if we have more information such as control of the second moment, then we have better control of the tails.

Corollary 1 (*Chebyshev's Inequality*)

Let X have mean μ. Then for any constant $a > 0$,

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathbb{E}[(X-\mu)^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

Remark 5. Notice that the tail probability in Chebyshev's Inequality is now bounded by t^{-2} instead of t^{-1} so it gives us better control than Markov's inequality.

If we have even more information about the moments, then we can get upgrade the bound to give us exponential control of the tails.

Corollary 2 (*Chernoff Bound*)

For any random variable and constantss $t > 0$ and $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = M_X(t)e^{-ta}.$$

There is an extra parameter t in the statement of the Chernoff bound, so we can minimize the upper bound as a function of t to get the better control of the tail probabilities.

1.3 Example Problems

Problem 1.1. An insurance company sells $1,000$ independent health insurance policies. Suppose that for any policy holder, the expected claim is $5,000$ but the variance is $2,000,000$ since the costs of the treatment can vary a lot.

1. What can the company say about the average claim over all $1,000$ policies?
2. What is a bound on the probability that the insurance company has to pays more than $10,000,000$ in a given year.

Solution 1.1.

Part 1: The law of large number states that the average claim

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_1] = 5000.$$

So, since $n = 1000$ is quite a large number of policies, the average claim will be close to 5000 even though the individual claim amounts can vary widely.

Part 2: Notice that $T = \sum_{i=1}^{1000} X_i$ the total total claim amount satisfies

$$\mathbb{E}(T) = \sum_{i=1}^{1000} \mathbb{E}[X_i] = 5,000,0000 \text{ and } \operatorname{Var}(T) = \sum_{i=1}^{1000} \operatorname{Var}\left(X_i\right) = 2,000,000,000.$$

Therefore, by Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}(T > 10,000,0000) &= \mathbb{P}(T - 5,000,000 > 5,000,000) \\
&\leq \mathbb{P}(|T - 5,000,000| > 5,000,000) \\
&\leq \frac{\operatorname{Var}(T)}{5,000,000^2} = \frac{2,000,000,000}{5,000,000^2} = 0.0008.
\end{aligned}$$

Remark 6. If we used Markov's inequality instead, we would have gotten

$$\mathbb{P}(T > 10,000,0000) \leq \frac{\mathbb{E}(T)}{10,000,000} = \frac{1}{2}$$

which is true, but not as good of a bound as if we used information about the variance.

Problem 1.2. Suppose a machine produces metal rods whose lengths (in cm) are nonnegative random variables L with mean 100.

1. Use Markov's inequality to bound the probability that a rod exceeds 150 cm.
2. Suppose that we now know that the variance of the rods are 25 cm^2. Use Chebyshev's inequality to bound the probability that a rod exceeds 150 cm.

Solution 1.2.

Part 1: By Markov's inequality (since the lengths of the rods are non-negative)

$$\mathbb{P}(L > 150) \leq \frac{\mathbb{E}[L]}{150} = \frac{2}{3}.$$

Part 2: Since $\{|L - 100| > 50\} = \{L > 150\} \cup \{L < 50\}$, by Chebyshev's inequality,

$$\mathbb{P}(L > 150) \leq \mathbb{P}(|L - 100| > 50) \leq \frac{\text{Var}(L)}{50^2} = \frac{1}{100}.$$

Problem 1.3. Let Y be the average income (in thousands) of 100 independently surveyed people from a population with mean 50 and variance 100. Use Chebyshev's inequality to bound

$$\mathbb{P}(|Y - 50| \geq 10).$$

Solution 1.3. We have that Y is the average of, so

$$\text{Var}(Y) = \text{Var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \frac{100\,\text{Var}(X)}{100^2} = 1$$

Therefore,

$$\mathbb{P}(|Y - 50| \geq 10) \leq \frac{\text{Var}(Y)}{10^2} = \frac{1}{100},$$

which is quite small even though the standard deviation of a single person is 10.

1.4 Proofs of Key Results

Problem 1.4. Prove Theorem 1, the weak law of large numbers.

Solution 1.4. We have that $\text{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$. For any $\epsilon > 0$, Chebyshev's inequality implies that

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n} \to 0$$

as $n \to \infty$.

Problem 1.5. Prove Theorem 3, Markov's inequality.

Solution 1.5. For any $a > 0$, we define $Y = \frac{|X|}{a}$. It suffices to show that $\mathbb{P}(Y > 1) \leq \mathbb{E}[Y]$ since

$$\mathbb{P}(|X| > a) = \mathbb{P}(Y > 1) \leq \mathbb{E}[Y] = \frac{\mathbb{E}[|X|]}{a}.$$

Notice that we always have

$$\mathbb{1}(Y \geq 1) = \begin{cases} 1 & Y \geq 1 \\ 0 & Y < 1 \end{cases} \leq Y.$$

Monotonicity of the expected value implies that

$$\mathbb{E}[\mathbb{1}(Y \leq 1)] \leq \mathbb{E}[Y] \implies \mathbb{P}(Y > 1) \leq \mathbb{E}[Y].$$

Problem 1.6. Prove Corollary 1 and Corollary 2. That is, for any $t > 0$ and $a > 0$, show that

1. $$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$
2. $$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

Solution 1.6. Both of these statements are direct consequences of Markov's inequality.

Part 1: We have

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}(|X - \mu|^2 \geq a^2),$$

so Markov's inequality implies that

$$\mathbb{P}(|X - \mu|^2 \geq a^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

Part 2: We have

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}),$$

so Markov's inequality implies that

$$pP(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = M_X(t)e^{-ta}.$$

2 The Central Limit Theorem

The central limit theorem is the most important result in probability theory. It connects the fields of probability and statistics. It is the fundamental reason why so many of the statistical tests and methods work.

The law of large numbers says that the sample mean $\bar{X}_n$ converges to μ almost surely. However, the $\bar{X}_n$ is a random variable, so it will still fluctuate around its expected value. The distribution of these fluctuation around the expected value are given by the central limit theorem.

Theorem 4 (*Central Limit Theorem*)

Suppose that $X_1, ..., X_n$ i.i.d. random variables with (finite) mean μ and (finite) variance σ^2. Then,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to \mathrm{N}(0, 1),$$

in distribution as $n \to \infty$. In other words, if Φ is the CDF of a standard normally distributed random variable, we have for all x,

$$F_{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}(x) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) \to \Phi(x) \text{ as } n \to \infty.$$

Remark 7. Notice that we don't need to specify the distribution of X. Only the mean and variance are needed to characterize the behavior of the sample mean when n is large. This means the normal distribution is a universal distribution, since it governs the behavior of the sample mean of a very large class of random variables. It is especially powerful in statistics, since it's universal nature means that we can apply statistical tests without knowing the distribution of random variables in practice.

This is an asymptotic result, but it is still useful for approximations. Recall that if X_i are independent $\mathrm{N}(\mu, \sigma^2)$ random variables, then the stability property implies

$$\bar{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1).$$

The central limit theorem says that this holds in an approximate sense even if X_i are not normally distributed. That is, for large n, we have that

$$\bar{X}_n \dot{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

so we can approximate the distribution of $\bar{X}_n$ with a normally distributed random variable with the same mean and variance. In fact, Lindeberg's proof of the CLT (see Problem 2.9) uses the observation that we can replace the terms in an arbitrary sample mean one by one with a normally distributed random variable with the same mean and variance.

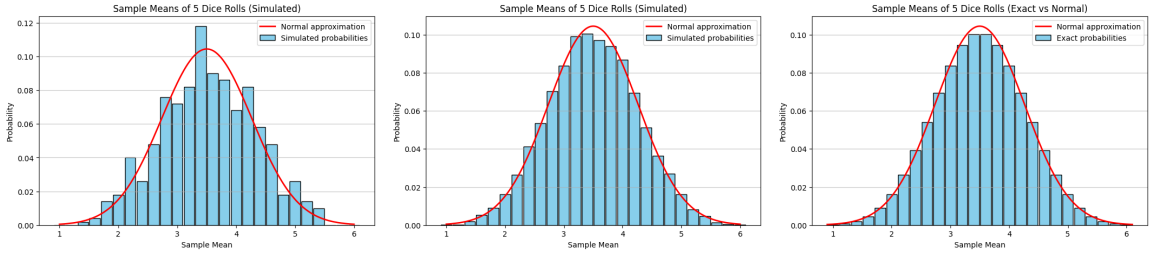


Figure: The histogram of the (re-normalized) sample mean of a roll of 5 dice were simulated and plotted when $n = 500, 10000, \infty$.

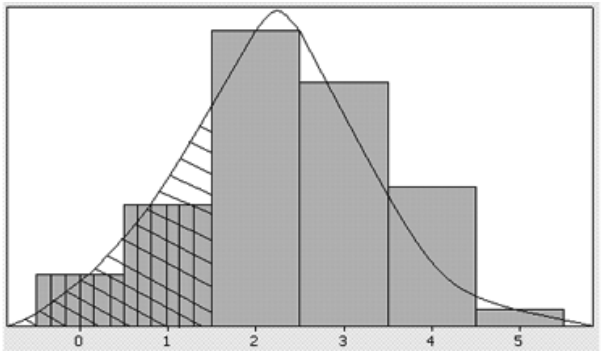
2.1 Continuity Correction for Discrete Approximations

When we approximate a *discrete random variables* taking consecutive *integer* values with a continuous density, we need to apply a *continuity correction* to account for the fact the random variables can't take non-integer values:

$$\text{To approximate } \mathbb{P}(a \leq X \leq b) \text{ we instead compute } \mathbb{P}(a - 0.5 \leq X \leq b + 0.5).$$

The latter will give us a better approximation.When the intervals are wide, the continuity correction has a small effect, but when the intervals are small then there can be a big effect (Problem 2.3 and 2.4).

Example 1. We want to compute $\mathbb{P}(0 \leq X \leq 1)$, but notice that the area under the continuous approximation of the discrete PMF is closer if we integrate from -0.5 to 1.5.



Notice that the continuity correction is half the width of the block in the histogram.

In general, if X does not take consecutive values, then we subtract or add **half the width** of the increments **before standardization to integer valued bounds** a **and** b to our various inequalities when approximating discrete distributions using the CLT. The correction should not be applied when approximating continuous distributions.

Remark 8. Strict inequalities matter when applying the continuity correction. To remember which way to apply the continuity correction, we should always write the probabilities using inequalities before applying our rule to widen the interval.

For example, if T takes values in $\mathbb{Z}$, then for integer valued x

$$\mathbb{P}(T > x) = \mathbb{P}(T \geq x + 1) \overset{\text{correction}}{\Rightarrow} \mathbb{P}(T \geq x + 1 - 0.5) = \mathbb{P}(T \geq x + 0.5) = \mathbb{P}(T > x + 0.5).$$

Other cases are similar, for example for integer valued x

$$\mathbb{P}(T = x) = \mathbb{P}(x \leq T \leq x) \overset{\text{correction}}{\Rightarrow} \mathbb{P}(x - 0.5 \leq T \leq x + 0.5).$$

2.2 Normal Approximations of Important Distributions

We can use the CLT to approximate the distributions of important distributions we have already encountered.

2.2.1 Normal Approximation of the Binomial Distribution

Since $X \sim \text{Bin}(n, p)$ can be written as $X = \sum_{i=1}^{n} X_i$ where the $X_1, \ldots, X_n$ are independent $\text{Bin}(1, p)$, we can apply the CLT to $\sum_{i=1}^{n} X_i$.

Theorem 5

If $X \sim \text{Bin}(n, p)$, then

$$\frac{X - np}{\sqrt{np(1-p)}} \to N(0, 1).$$

in distribution as $n \to \infty$.

2.2.2 Normal Approximation of the Poisson Distribution

If λ is a natural number, then $X \sim \text{Poi}(\lambda)$ can be written as $X = \sum_{i=1}^{\lambda} X_i$ where the $X_1, \ldots, X_n$ are independent Poi(1), we can apply the CLT to $\sum_{i=1}^{\lambda} X_i$.

Theorem 6

If $X \sim \text{Bin}(n, p)$, then

$$\frac{X - np}{\sqrt{np(1-p)}} \to N(0, 1).$$

in distribution as $n \to \infty$.

2.3 Example Problems

2.3.1 Applications

Problem 2.1. A carton of wine consists of 20 winebottles. Suppose we can model the volume of wine in each bottle as independent normal random variables $X_1, \ldots, X_{20}$ where mean 1.05 litres and and standard deviation $\sqrt{0.0004}$. What is the distribution of the total amount of wine in a carton, say T?

Solution 2.1. Since X_i are normally distributed, the total $T = \sum_{i=1}^{2} 0X_i$ is also normally distributed. We need to find the mean and variance of T,

$$\mathbb{E}[T] = \mathbb{E}\left[\sum_{i=1}^{20} X_i\right] = 20\,\mathbb{E}[X_1] = 20 \cdot 1.05 = 21$$

and by independence,

$$\text{Var}[T] = \text{Var}\left[\sum_{i=1}^{20} X_i\right] = \sum_{i=1}^{20} \text{Var}(X_i) = 20 \cdot 0.0004 = 0.008.$$

Therefore, $T \sim \text{N}(21, 0.008)$.

Problem 2.2. Harold is eating a box of chocolate right now (yes, right now). Each box contains 20 cubes, and it is supposed to have a total of 500 grams of chocolate in it. The weight of each chocolate cube varies a little because they are hand-made from Switzerland. The weight W of each cube is a random variable with mean $\mu = 25$ grams, and the standard deviation $\sigma = 0.1$ grams. Find the probability that a box has at least 500 grams of chocolate in it, assuming that the weight of each cube is independent.

Solution 2.2. Let $W_1, \ldots, W_{20}$ be the weights of the cubes and $T = \sum_{i=1}^{20} W_i$ the total weight. The $W_1, \ldots, W_{20}$ are independent with mean $\mu = 25$, $\sigma = 0.1$, and by the CLT, T is approximately $N(20\mu, 20\sigma^2)$. Using Z-tables gives us

$$\begin{aligned}\mathbb{P}(T \geq 500) &= \mathbb{P}\left(\frac{T - 20 \cdot \mu}{\sqrt{20\sigma^2}} \geq \frac{500 - 20 \cdot \mu}{\sqrt{20\sigma^2}}\right) \\ &\approx \mathbb{P}(Z \geq 0) = 1 - \Phi(0) = 0.5\end{aligned}$$

Problem 2.3. In February this year, various Youtubers participated in a '100 cup challenge' related to the Roll Up the Rim to Win promotion at Tim Hortons. The advertised cance to win is 1/6. Participants bought 100 promotional cups, and filmed themselves as they found out how many times they'd won. We want to use the central limit theorem to estimate the probability a participant recorded between 15 and 20 wins (inclusive).

1. Compute using the CLT *without* continuity correction.
2. Compute the probability exactly.
3. Compute using the CLT *with* continuity correction.

Note: think about the assumptions we must make when considering real-world examples!

Solution 2.3.

CLT without Correction: Let $X_i \sim \text{Bin}\left(1, \frac{1}{6}\right)$, $i = 1, ..., 100$ and $T = \sum_{i=1}^{100} X_i$ be the total wins. In this case, we have

$$\mathbb{E}[X_i] = p = \frac{1}{6} \text{ and } \text{Var}(X_i) = p(1-p) = \frac{1}{6}\left(1 - \frac{1}{6}\right) = \frac{5}{36}.$$

The CLT says that $T \sim \text{N}(np, np(1-p)) = \text{N}(100 \cdot \frac{1}{6}, 100 \cdot \frac{5}{36})$. Using Z-tables gives us

$$\begin{aligned}
\mathbb{P}(15 \leq T \leq 20) &= \mathbb{P}(T \leq 20) - \mathbb{P}(T < 15) \\
&\approx \mathbb{P}\left(\frac{T - np}{\sqrt{np(1-p)}} \leq \frac{20 - np}{\sqrt{np(1-p)}}\right) - \mathbb{P}\left(\frac{T - np}{\sqrt{np(1-p)}} < \frac{15 - np}{\sqrt{np(1-p)}}\right) \\
&= \mathbb{P}\left(Z \leq 0.894\right) - P\left(Z < -0.447\right) \\
&= 0.487.
\end{aligned}$$

Exact: We can compute the probability exactly since $T \sim \text{Bin}(100, 1/6)$, and so

$$\begin{aligned}
\mathbb{P}(15 \leq T \leq 20) &= \sum_{x=15}^{20} \binom{100}{x} \left(\frac{1}{6}\right)^x \left(1 - \frac{1}{6}\right)^{100-x} \\
&= 0.561
\end{aligned}$$

which is quite far away from the CLT approximation.

CLT with Correction: If we use correction, then we need to compute

$$\begin{aligned}
\mathbb{P}(14.5 \leq T \leq 20.5) &= \mathbb{P}(T \leq 20.5) - \mathbb{P}(T < 14.5) \\
&\approx \mathbb{P}\left(\frac{T - np}{\sqrt{np(1-p)}} \leq \frac{20.5 - np}{\sqrt{np(1-p)}}\right) - \mathbb{P}\left(\frac{T - np}{\sqrt{np(1-p)}} < \frac{14.5 - np}{\sqrt{np(1-p)}}\right) \\
&= \mathbb{P}\left(Z \leq 1.029\right) - \mathbb{P}\left(Z < -0.581\right) \\
&= 0.568.
\end{aligned}$$

which is very close to the probability of 0.561.

Problem 2.4. Suppose $X \sim \text{Poi}(\mu)$. Use the normal approximation to approximate

$$\mathbb{P}(X > \mu)$$

and compare this approximation with the true value when $\mu = 9$.

Solution 2.4. Since X is discrete, we need to use the normal approximation with continuity correction. We have

$$\mathbb{P}(X > 9) = \mathbb{P}(X \geq 10) \Rightarrow \mathbb{P}(X \geq 10 - 0.5) = \mathbb{P}(X \geq 9.5)$$

If we apply the normal approximation:

$$\frac{X - \mu}{\sqrt{\mu}} \overset{approx}{\sim} Z \sim \mathrm{N}(0, 1)$$

we can see that, with $\mu = 9$:

$$\begin{aligned} \mathbb{P}(X > 9.5) &= \mathbb{P}\left(\frac{X - \mu}{\sqrt{\mu}} > \frac{9.5 - \mu}{\sqrt{\mu}}\right) \\ &\approx \mathbb{P}(Z > 0.17) = 0.432. \end{aligned}$$

Remark 9. We can compute the probability exactly

$$\mathbb{P}(X > 9) = 1 - \mathbb{P}(X \leq \mu) = 1 - \left(e^{-9} + 9e^{-9} + ... + \frac{9^9}{9!}e^{-9}\right) = 0.4126,$$

which is quite close to the normal approximation with correction. If we didn't apply the continuity correction, then

$$\mathbb{P}(X > \mu) = \mathbb{P}\left(\frac{X - \mu}{\sqrt{\mu}} > \frac{\mu - \mu}{\sqrt{\mu}}\right) \approx \mathbb{P}(Z > 0) = 0.5,$$

which is quite far off from the true value.

Problem 2.5. Let p be the proportion of Canadians who think Canada should adopt the US dollar.

(a) Suppose 400 Canadians are randomly chosen and asked their opinion. Let X be the number who say yes. Find the probability that the proportion, $\frac{X}{400}$, of people who say yes is within 0.02 of p, if $p = 0.20$.

(b) Suppose for a future opinion poll we want to determine the number, n, to survey to ensure that there is a 95% $\frac{X}{n}$ lies within 0.02 of p. Suppose $p = 0.20$ is known.

(c) Repeat (b) when the value of p is unknown. (Note that this would be the more realistic situation in the case of conducting an opinion poll.)

Solution 2.5. We want to apply the CLT.

Part (a): We have $X \sim \mathrm{Bin}(n = 400, p = 0.2)$. Notice that

$$\mathbb{E}[X] = np = 80 \text{ and } \mathrm{Var}(X) = np(1 - p) = 64.$$

We want to compute

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0.02\right) = \mathbb{P}\left(n(p - 0.02) \leq X \leq n(p + 0.02)\right) = \mathbb{P}\left(72 \leq X \leq 88\right).$$

Since X is takes integer values and the bounds are also integer valued we need to apply the continuity correction, so we instead compute

$$\mathbb{P}\left(72 - 0.5 \leq X \leq 88 + 0.5\right) = \mathbb{P}\left(71.5 \leq X \leq 96.5\right).$$

Splitting the probabilities, and standardizing gives us

$$\begin{aligned}\mathbb{P}\left(71.5 \le X \le 88.5\right) &= \mathbb{P}\left(X \le 96.5\right) - \mathbb{P}\left(X < 71.5\right) \\ &= \mathbb{P}\left(\frac{X-80}{\sqrt{64}} \le \frac{88.5-80}{\sqrt{64}}\right) - \mathbb{P}\left(\frac{X-80}{\sqrt{64}} < \frac{71.5-80}{\sqrt{64}}\right)\end{aligned}$$

Using the CLT, this is approximately equal to

$$\begin{aligned}&\mathbb{P}\left(Z \le \frac{88.5-80}{\sqrt{64}}\right) - \mathbb{P}\left(Z < \frac{71.5-80}{\sqrt{64}}\right) \\ &= \mathbb{P}(Z \le 1.0625) - \mathbb{P}(Z \le -1.0625) = 2\,\mathbb{P}(Z \le 1.0625) - 1 = 2(0.85543) - 1 = 0.71086.\end{aligned}$$

Remark 10. Notice that we applied the continuity correction to X and not $\frac{X}{n}$, since X is the random variable that takes integer values.

Part (b): It is a bit tricky to apply the continuity correction in this problem because we are solving for n and we don't necessarily know if $n \cdot 0.02$ will be an integer. However, we will see that n is quite large so the effect of the continuity correction will be small.

Let $\bar{X} = \frac{X}{n}$ be the sample mean. We want to find a n such that

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \le 0.02\right) = \mathbb{P}\left(|\bar{X} - p| \le 0.02\right) = 0.95$$

Since

$$\mathbb{E}[\bar{X}] = p \text{ and } \operatorname{Var}(\bar{X}) = \frac{p(1-p)}{n}$$

the CLT implies that

$$\mathbb{P}\left(|\bar{X} - p| \le 0.02\right) = \mathbb{P}\left(\frac{|\bar{X} - p|}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}\right) \approx \mathbb{P}\left(|Z| \le \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}\right).$$

We want the right hand side to be 0.95, so we use quantiles to find the critical value. Since

$$\begin{aligned}\mathbb{P}(|Z| \le x) = \mathbb{P}(-x \le Z \le x) &= \mathbb{P}(Z \le x) - \mathbb{P}(Z \le -x) \\ &= \mathbb{P}(Z \le x) - (1 - \mathbb{P}(Z \le x)) = 2\,\mathbb{P}(Z \le x) - 1\end{aligned}$$

using the quantile table, we have that

$$\mathbb{P}(|Z| \le x) = 0.95 \iff 2\,\mathbb{P}(Z \le x) - 1 = 0.95 \iff \mathbb{P}(Z \le x) = 0.975 \iff x = F_Z^{-1}(0.975) = 1.96.$$

Therefore, we require that (with $p = 0.2$)

$$\frac{0.02}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.02}{\sqrt{\frac{0.2(1-0.2)}{n}}} = 1.96 \implies n = 1536.64$$

so we should take at least $n = 1537$ samples.

Remark 11. By the 95% rule we know that approximately 95% of the probability lies within two standard deviations. In the computations above, we carefully showed this result by finding the n such that the $\frac{0.02}{\sqrt{\operatorname{Var}(\frac{X}{n})}} = 1.96$, which is roughly two standard deviations.

Part (c): If p is unknown, then we want to make sure that our sample size is large enough so that we can guarantee a 0.95 accuracy for any p. That is, we want

$$\mathbb{P}\left(|\bar{X} - p| \leq 0.02\right) \geq 0.95$$

for all p. If $p = 0$ or $p = 1$, then the problem is trivial. If $p \in (0, 1)$, then the computations from above implies that

$$\mathbb{P}\left(|\bar{X} - p| \leq 0.02\right) \approx \mathbb{P}\left(|Z| \leq \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

This is larger than 0.95 whenever $\frac{0.02}{\sqrt{\frac{p(1-p)}{n}}} \geq 1.96$. Thus, for this to be larger than 0.95 to hold for all p, we want to find n such that

$$\frac{0.02}{\sqrt{\frac{p(1-p)}{n}}} \geq 1.96$$

for any all p. Notice that $p(1-p)$ is maximized when $p = \frac{1}{2}$, so we have

$$\frac{0.02}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0.02}{\sqrt{\frac{0.5(1-0.5)}{n}}}$$

with equality when $p = 0.5$. Since

$$\frac{0.02}{\sqrt{\frac{0.5(1-0.5)}{n}}} = 1.96 \implies n = 2401$$

we need $n = 2401$ to achieve the desired accuracy for any $p = 0.5$. Therefore, by monotonicity

$$\mathbb{P}\left(|Z| \leq \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}\right) \geq \mathbb{P}\left(|Z| \leq \frac{0.02}{\sqrt{\frac{0.5(1-0.5)}{n}}}\right) = 0.95$$

which gives us the required accuracy for all p.

Remark 12. By the 95% rule we know that approximately 95% of the probability lies within two standard deviations. In the computations above, we found that in the worst case scenario when $p = \frac{1}{2}$ taking n such that the $\frac{0.02}{\sqrt{\text{Var}(\frac{X}{n})}} = 1.96$ achieved the required accuracy. Since n is taken large enough to be close in the worst case scenario, we get the required accuracy in all scenarios.

2.4 Proofs of Key Results

Problem 2.6. If

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} \text{N}(0, 1)$$

show that

$$\bar{X} \overset{approx}{\sim} \text{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } T = n\bar{X} = \sum_{i=1}^{n} X_i \overset{approx}{\sim} \text{N}\left(n\mu, n\sigma^2\right).$$

Solution 2.6. Let $Z \sim \text{N}(0, 1)$. By standardization, we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} Z \implies \bar{X} = \frac{\sigma}{\sqrt{n}} Z + \mu$$

so $\bar{X} \overset{approx}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Likewise, using the fact that $T = n\bar{X}$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} Z \implies T = n\bar{X} = \sqrt{n}\sigma Z + n\mu$$

so $T \overset{approx}{\sim} \mathrm{N}\left(n\mu, n\sigma^2\right)$.

Problem 2.7. Prove the following Normal approximation results

1. If $X \sim \mathrm{Bin}(n, p)$, then for large n
$$\frac{X - np}{\sqrt{np(1-p)}} \overset{approx}{\sim} N(0, 1).$$
2. If $X \sim Poi(\lambda)$, then for large λ
$$\frac{X - \lambda}{\sqrt{\lambda}} \overset{approx}{\sim} N(0, 1).$$

Solution 2.7.

1. Since $X \sim \mathrm{Bin}(n, p)$ can be written as $X = \sum_{i=1}^{n} X_i$ where the $X_1, \ldots, X_n$ are independent $\mathrm{Bin}(1, p)$ (which have mean p and variance $p(1-p)$), we can apply the CLT to $\sum_{i=1}^{n} X_i$,
$$\sum_{i=1}^{n} X_i \sim \mathrm{N}(np, np(1-p))$$
so the result now follows from standardization.
2. If $\lambda = n$ is a natural number, then $X \sim \mathrm{Poi}(n)$ can be written as $X = \sum_{i=1}^{\lambda} X_i$ where the $X_1, \ldots, X_n$ are independent $\mathrm{Poi}(1)$ (which have mean 1 and variance 1), we can apply the CLT to $\sum_{i=1}^{\lambda} X_i$,
$$\sum_{i=1}^{n} X_i \sim \mathrm{N}(n, n)$$
so the result now follows from standardization.

Problem 2.8. Prove the central limit theorem under the additional assumption that the MGF of X is finite for all $t \in [-a, a]$ for some $a > 0$. Suppose that $X_1, ..., X_n$ are i.i.d. random variable such that $\mathbb{E}(X_i) = \mu$, and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then for all $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) \to \Phi(x)$$

as $n \to \infty$.

Solution 2.8. We will show that as $n \to \infty$, the MGF of $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ converges to the MGF of the standard normal distribution $Z \sim \mathrm{N}(0, 1)$,

$$M_Z(t) = e^{\frac{t^2}{2}}.$$

Since the MGF uniquely determines the distribution, this will imply that the CDF of $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ converges to the CDF of a standard normal distribution.

To simplify notation, we consider the normalized random variables $Y_i = \frac{X_i - \mu}{\sigma}$. Notice that Y_i satisfies

$$\mathbb{E}[Y_i] = \mathbb{E}\left[\frac{X_i - \mu}{\sigma}\right] = \frac{\mathbb{E}[X_i] - \mu}{\sigma} = 0$$

and

$$\operatorname{Var}(Y_i) = \operatorname{Var}\left[\frac{X_i - \mu}{\sigma}\right] = \frac{1}{\sigma^2}\operatorname{Var}(X_i) = 1.$$

We have

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu)}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i.$$

By the independence property of the MGFs,

$$M_{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}(t) = \mathbb{E}[e^{\frac{t}{\sqrt{n}}\sum_{i=1}^n Y_i}] = \prod_{i=1}^n \mathbb{E}[e^{\frac{t}{\sqrt{n}}Y_i}] = \prod_{i=1}^n M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = M_Y^n\left(\frac{t}{\sqrt{n}}\right). \tag{1}$$

If we take $n \to \infty$, Notice that this is of the indeterminate form 1^∞, so to compute the limit we first take the logarithm and use the fact that $M_Y(0) = 1$, $M_Y'(0) = \mathbb{E}[Y] = 0$, $M_Y''(0) = \mathbb{E}[X^2] = 1$,

$$\begin{aligned}
\lim_{n\to\infty} \ln M_Y^n\left(\frac{t}{\sqrt{n}}\right) &= \lim_{n\to\infty} n \ln M_Y\left(\frac{t}{\sqrt{n}}\right) \\
&= \lim_{m\to 0} \frac{\ln M_Y(mt)}{m^2} && \frac{1}{\sqrt{n}} = m \\
&= \lim_{m\to 0} \frac{tM_Y'(mt)}{2mM_Y(mt)} && \text{L'Hôpital's rule to } \frac{0}{0} \\
&= \lim_{m\to 0} \frac{t^2 M_Y''\left(mt\right)}{2M_Y\left(mt\right) + 2tmM_Y'\left(mt\right)} && \text{L'Hôpital's rule to } \frac{0}{0} \\
&= \frac{t^2}{2}.
\end{aligned}$$

In other words,

$$\lim_{n\to\infty} M_{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}(t) = \lim_{n\to\infty} M_Y^n\left(\frac{t}{\sqrt{n}}\right) = e^{\frac{t^2}{2}} = M_Z(t).$$

Remark 13. We used the assumption that the MGF exists and is finite on an interval around 0 to use the uniqueness and differentiation properties of the MGF. This is also a strong assumption since it essentially assumes the existence of all moments.

Problem 2.9. Prove the central limit theorem under the additional assumption that $\mathbb{E}[|X|^3] < \infty$. That is, suppose that $X_1, ..., X_n$ are i.i.d. random variables such that $\mathbb{E}(X_i) = \mu$, and $\operatorname{Var}(X_i) = \sigma^2 < \infty$. Then for all $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le x\right) \to \Phi(x)$$

as $n \to \infty$.

Solution 2.9. We present the proof using the Lindeberg method. We let $Z_1, ..., Z_n$ be an i.i.d. sequence of standard normal random variables. Notice that by Gaussian stability, $\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \sim \mathrm{N}(0,1)$, so

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \leq x \right) = \Phi(x).$$

In this case, the CLT is exact even at finite n. We will show that this holds in an approximate sense even if Y_i is not Gaussian.

As like in the MGF proof in Problem 2.9 before, it suffices to show that the re-normalized sample mean of the $Y_i = \frac{X_i - \mu}{\sigma}$ converges to CDF of the re-normalized sample mean of the Z_i, i.e.

$$\left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) - \Phi(x) \right| = \left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) - \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \leq x \right) \right| \to 0.$$

Step 1: Recall that

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) = \mathbb{E}\left[\mathbb{1}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) \right].$$

The indicator function $\mathbb{1}(t \leq x)$ as a function of t isn't too nice since it is not differentiable at x. However, this is a minor issue. For any fixed x, we can find a function $\phi(t)$ that does not depend on n so that $\phi(t) \approx \mathbb{1}(t \leq x)$ and $\|\phi'''(t)\|_\infty$ is uniformly bounded. Thus, we have that

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) = \mathbb{E}\left[\mathbb{1}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) \right] \approx \mathbb{E}\left[\phi\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \leq x \right) \right].$$

Therefore, we can analyze expected values of ϕ instead of the indicator random function. A careful construction of this approximation can be done, but it won't add much intuition about why the CLT holds, so we will skip it.

Step 2: It suffices to show that

$$\left| \mathbb{E}\left[\phi\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Y_i \right) \right] - \mathbb{E}\left[\phi\left(\frac{1}{\sqrt{n}} \sum_{i=1}^N Z_i \right) \right] \right|$$

is small as $n \to \infty$. We show this by writing this difference as a telescoping sum, where we replace each term in the sum of Z one by one,

$$\begin{aligned}
S_0 &= \frac{1}{\sqrt{n}}(Z_1 + Z_2 + Z_3 + \cdots + Z_{n-1} + Z_n) \\
S_1 &= \frac{1}{\sqrt{n}}(Y_1 + Z_2 + Z_3 + \cdots + Z_{n-1} + Z_n) \\
S_2 &= \frac{1}{\sqrt{n}}(Y_1 + Y_2 + Z_3 + \cdots + Z_{n-1} + Z_n) \\
&\vdots \\
S_n &= \frac{1}{\sqrt{n}}(Y_1 + Y_2 + Y_3 + \cdots + Y_{n-1} + Y_n)
\end{aligned}$$

so

$$\begin{aligned}\left|\mathbb{E}\left[\phi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{N}Y_i\Big)\right]-\mathbb{E}\left[\phi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{N}Z_i\Big)\right]\right| &= |\,\mathbb{E}[\phi(S_n)]-\mathbb{E}[\phi(0)]| \\ &= \left|\sum_{i=1}^{n}\mathbb{E}[\phi(S_i)]-\mathbb{E}[\phi(S_{i-1})]\right| \\ &\le \sum_{i=1}^{n}|\,\mathbb{E}[\phi(S_i)]-\mathbb{E}[\phi(S_{i-1})]|.\end{aligned}$$

Our goal now is to control the error of each individual term $|\,\mathbb{E}[\phi(S_i)]-\mathbb{E}[\phi(S_{i-1})]|$ and show that they are small enough so that the sum still goes to zero.

Step 3: Notice that the terms S_i and S_{i-1} only differ in the ith index. Let

$$T_i=\frac{1}{\sqrt{n}}(Y_1+Y_2+\cdots+Y_{i-1}+Z_{i+1}+\cdots+Z_{n-1}+Z_n)$$

denote the terms in the sum with the ith index removed so $S_{i-1}=T_i+\frac{1}{\sqrt{n}}Z_i$ and $S_i=T_i+\frac{1}{\sqrt{n}}Y_i$. We can Taylor expand around T_i up to the second order, using the fact that ϕ''' is uniformly bounded to conclude that there exists (random) numbers ξ_1 and ξ_2

$$\begin{aligned}\mathbb{E}[\phi(S_{i-1})] &= \mathbb{E}\left[\phi(T_i)+\phi'(T_i)\frac{1}{\sqrt{n}}Z_i+\phi''(T_i)\frac{1}{2n}Z_i^2+\frac{\phi'''(\xi_1)}{6n^{\frac{3}{2}}}|Z_i|^3\right] \\ \mathbb{E}[\phi(S_i)] &= \mathbb{E}\left[\phi(T_i)+\phi'(T_i)\frac{1}{\sqrt{n}}Y_i+\phi''(T_i)\frac{1}{2n}Y_i^2+\frac{\phi'''(\xi_2)}{6n^{\frac{3}{2}}}|Y_i|^3\right].\end{aligned}$$

Notice that T_i, Z_i, Y_i are independent since T_i contains summations of the other indices not including i, so

$$\mathbb{E}\left[\phi'(T_i)\frac{1}{\sqrt{n}}Z_i\right]=\mathbb{E}[\phi'(T_i)]\,\mathbb{E}\left[\frac{1}{\sqrt{n}}Z_i\right]=0 \text{ and } \mathbb{E}\left[\phi'(T_i)\frac{1}{\sqrt{n}}Y_i\right]=\mathbb{E}[\phi'(T_i)]\,\mathbb{E}\left[\frac{1}{\sqrt{n}}Y_i\right]=0$$

since both Z_i and Y_i have mean zero. Similarly, for the second order term,

$$\mathbb{E}\left[\frac{\phi''(T_i)Z_i^2}{2n}\right]=\frac{\mathbb{E}[\phi''(T_i)]}{2n}\,\mathbb{E}[Z_i^2]=\frac{\mathbb{E}[\phi''(T_i)]}{2n} \text{ and } \mathbb{E}\left[\frac{\phi''(T_i)Y_i^2}{2n}\right]=\frac{\mathbb{E}[\phi''(T_i)]}{2n}\,\mathbb{E}[Y_i^2]=\frac{\mathbb{E}[\phi''(T_i)]}{2n}.$$

since both Z_i and Y_i have variance 1. Therefore, the zeroth, first and second order terms are equal on average, so by Jensen's inequality (to move the absolute value inside the expected value)

$$|\,\mathbb{E}[\phi(S_i)-\mathbb{E}[\phi(S_{i-1})]|=\left|\mathbb{E}\left[\frac{\phi'''(\xi_2)}{6n^{\frac{3}{2}}}|Z_i|^3\right]+\mathbb{E}\left[\frac{\phi'''(\xi_2)}{6n^{\frac{3}{2}}}|Y_i|^3\right]\right|\le\frac{C_1}{6n^{\frac{3}{2}}}(\mathbb{E}[|Z_i|^3]+\mathbb{E}[|Y_i|^3])\le\frac{C_2}{6n^{\frac{3}{2}}}.$$

We used the fact that ϕ''' was uniformly bounded and that Y has bounded absolute third moment to construct the constants C_1 and C_2.

Step 4: In conclusion, we have shown that

$$\begin{aligned}\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{N}Y_i\le x\right)-\Phi(x)\right| &\approx \left|\mathbb{E}\left[\phi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{N}Y_i\Big)\right]-\mathbb{E}\left[\phi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{N}Z_i\Big)\right]\right| \\ &\le \sum_{i=1}^{n}|\,\mathbb{E}[\phi(S_i)]-\mathbb{E}[\phi(S_{i-1})]|\le n\cdot\frac{C_2}{6n^{\frac{3}{2}}}=\frac{C_2}{6\sqrt{n}}.\end{aligned}$$

This implies that the approximate CDFs are arbitrarily close when $n\to\infty$, so the original CDFs are close.

Remark 14. This is a stronger statement of the CLT since it requires less assumptions on the moments than in the MGF proof in Problem 2.8. We also got that the rate of convergence of the CDF is at most of order $\frac{1}{\sqrt{n}}$. Even stronger versions of the CLT exist, and with a bit more work you can remove the condition on the third moment.