M378K Introduction to Mathematical Statistics
Problem Set #11
De Moivre-Laplace.

Problem 11.1. *You are given a TRUE/FALSE exam with 30 questions. Suppose that you need to answer 21 questions correctly in order to pass. You have no idea what the class is about and decide to toss a fair coin to answer all the questions; you circle TRUE if the outcome is tails and you circle FALSE if the outcome is heads. What is your approximation of the probability that you manage to pass the exam using this strategy?*

For $Y \sim Binomial(n,p)$ we know that its probability mass function is:

$$p_Y(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n$$

Moreover, its expectation and its variance are

$$\mathbb{E}[Y] = np \quad \text{and} \quad Var[Y] = np(1-p).$$

Now, consider a sequence of binomial random variables $Y_n \sim Binomial(n,p)$. Note that, while the number of trials n varies, the probability of success in every trial p remains the same for all n. The *normal approximation to the binomial* is a theorem which states that

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{as } n \to \infty.$$

Practically, this means that Y_n is "approximately" normal with mean np and variance $np(1-p)$ for "large" n. The usual rule of thumb is that both $\boxed{np > 10}$ and $\boxed{n(1-p) > 10}$.

Another practical adjustment needs to be made due to the fact that discrete distributions of Y_n are approximated by a continuous (normal) distribution. This adjustment is usually referred to as the **continuity correction**. More specifically, provided that the conditions above are satisfied, for every integer $a < b$, we have that

$$\begin{aligned}
\boxed{\mathbb{P}[a \le Y_n \le b]} &= \mathbb{P}\left[a - \tfrac{1}{2} < Y_n < b + \tfrac{1}{2}\right] \\
&= \mathbb{P}\left[\frac{a - \frac{1}{2} - np}{\boxed{\sqrt{np(1-p)}}} < \frac{Y_n - np}{\sqrt{np(1-p)}} < \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right] \\
&\approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

where Φ, as usual, stands for the cummulative distribution function of the standard normal distribution.

For more about the history of the theorem and ideas for its proof, go to:
Wikipedia: de Moivre-Laplace.

$Y_n \sim b(n,p)$

$\mathbb{P}[Y_n = 50] = \mathbb{P}[\overset{a}{\widehat{50}} \le Y_n \le \overset{b}{\widehat{50}}] = 0$

$\sum_{k=0}^{n} \mathbb{P}[Y_n = k] = \dots = 0$

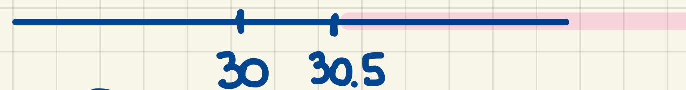
→: $Y \ldots$ # of questions answered correctly

$Y \sim b(\text{\# of trials} = n = 30,\ \text{success prob} = p = 0.5)$

$\mathbb{P}[Y \geq 21] = \ldots$ exactly $\ldots = \sum_{k=21}^{30} \left(\binom{30}{k} (0.5)^k (0.5)^{30-k} \right)$

Using the <u>normal approximation</u>:

$\mathbb{P}[Y \geq 21] = \mathbb{P}[Y > 20.5]$



$\mathbb{E}[Y] = 30(0.5) = 15; \quad Var[Y] = 15(0.5) = 7.5$

$\Rightarrow SD[Y] = \sqrt{7.5}$

$\mathbb{P}[Y > 20.5] = \mathbb{P}\left[\underbrace{\frac{Y-15}{\sqrt{7.5}}}_{\approx N(0,1)} > \frac{20.5-15}{\sqrt{7.5}} \right]$

$\approx \mathbb{P}[\ Z > 2\ ] = 1 - \Phi(2)$

$Z \sim N(0,1)$

$= 1 - 0.9772 = 0.0228$ $\square$

Problem 11.2. *A new addition of Kafka's "Metamorphosis" has 72 pages. The printing press often malfunctions and introduces typos. The number of typos on each page has a Poisson distribution with mean* $\ln(3)$ *and is independent of the number of typos on other pages (or other books). A book is thrown away if it contains typos on more than 32 pages. Use the normal approximation to estimate the proportion of books that get thrown away.*

$e^{-\lambda} \cdot \frac{\lambda^k}{k!}$

$k \geq 0$

→:

q ... a page has no typos

$q = \mathbb{P}[\text{Poisson w/ mean} = \ln(3) \text{ is equal to } 0]$

$= e^{-\ln(3)} = \frac{1}{3}$

p... probab. of @ least one typo on a page

$\boxed{p = \frac{2}{3}}$

Y... # of pages w/ typos

$Y \sim b(72, 2/3)$

$\mathbb{E}[Y] = 72 \cdot \frac{2}{3} = 48$; $\mathrm{Var}[Y] = 48 \cdot \frac{1}{3} = 16 \Rightarrow SD[Y] = 4$

$\mathbb{P}[Y \geq 32] = \mathbb{P}[Y > 31.5]$

$= \mathbb{P}\left[\underbrace{\frac{Y-48}{4}}_{\approx N(0,1) \sim Z} > \frac{31.5-48}{4}\right]$

$= \mathbb{P}[Z > -4.125] \approx 1$ □

M378K Introduction to Mathematical Statistics
Problem Set #12
The Central Limit Theorem (CLT).

Let $\{X_n, n = 1, 2, 3, \dots\}$ be a sequence of independent identically distributed random variables such that $\mu_X = \mathbb{E}[X_1] < \infty$ and $Var[X] = \sigma_X^2 < \infty$. For every $n = 1, 2, \dots$ define

$$S_n = X_1 + X_2 + \cdots + X_n$$

and

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Problem 12.1. *Find the expected value of S_n and $\bar{X}_n$ for every n.*

$\mathbb{E}[S_n] = n \cdot \mu_X$

$\boxed{\mathbb{E}[\bar{X}_n] = \mu_X}$ accuracy

Problem 12.2. *Find the variance and standard deviation of S_n and $\bar{X}_n$ for every n.*

$Var[S_n] = n \cdot \sigma_X^2 \Rightarrow SD[S_n] = \sigma_X \sqrt{n}$

$Var[\bar{X}_n] = \frac{\sigma_X^2}{n} \Rightarrow \boxed{SD[\bar{X}_n] = \frac{\sigma_X}{\sqrt{n}}}$ precision

Theorem 12.1. The Central Limit Theorem (CLT). *If the above conditions are satisfied, we have that*

$$\frac{S_n - n\mu_X}{\sigma_X \sqrt{n}} = \frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \overset{\mathcal{D}}{\Rightarrow} N(0,1) \quad \text{as } n \to \infty.$$

Practically, for "large enough" n, $\bar{X}_n$ is approximately normal with mean μ_X and variance $\frac{\sigma_X^2}{n}$. The rule of thumb is that we use the theorem for $n \geq 30$. If that is the case, we have that for any real $l < r$,

$$\mathbb{P}[l < S_n \leq r] = \mathbb{P}\left[\frac{l - n\mu_X}{\sigma_X\sqrt{n}} < \underbrace{\boxed{\frac{S_n - n\mu_X}{\sigma_X\sqrt{n}}}}_{\approx N(0,1) \sim Z} \leq \frac{r - n\mu_X}{\sigma_X\sqrt{n}}\right] \approx \Phi\left(\frac{r - n\mu_X}{\sigma_X\sqrt{n}}\right) - \Phi\left(\frac{l - n\mu_X}{\sigma_X\sqrt{n}}\right).$$

Similarly, for any real $a < b$,

$$\mathbb{P}[a < \bar{X}_n \leq b] = \mathbb{P}\left[\frac{a-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} < \frac{\bar{X}_n-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} \leq \frac{b-\mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right] \approx \Phi\left(\frac{b-\mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right) - \Phi\left(\frac{a-\mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right).$$

Problem 12.3. *The* Really Terrible Orchestra[1] *plans a concert at a gazebo in a local park. The orchestra has* 169 *members whose weights are assumed to be independent and identically distributed with mean* 100 *kilos and standard deviation of* 10 *kilos (the weight of the instruments is taken into account here). The gazebo can safely support up to* 17 *tons (each ton is* 1000 *kilos). What is the approximate probability that the gazebo will collapse?*

→: $S = Y_1 + Y_2 + \cdots + Y_{169}$ w/ Y_i has $\mu_Y = 100$ and $\sigma_Y = 10$

$\mathbb{E}[S] = 169 \cdot 100 = 16900$ $\quad SD[S] = 10\sqrt{169} = 130$

$$\mathbb{P}[S > 17\,000] = 1 - \mathbb{P}[S \leq 17\,000] = 1 - \mathbb{P}\Bigg[\underbrace{\boxed{\frac{S-16900}{130}}}_{\approx N(0,1)\sim Z} \leq \underbrace{\frac{17000-16900}{130}}_{0.77}\Bigg]$$

In the tables: $1 - 0.7794 = 0.2206$

Problem 12.4. Source: Sample P exam, Problem #65.
A charity receives 2025 *contributions. Contributions are assumed to be mutually independent and identically distributed with mean* 3125 *and standard deviation* 250*. Calculate the approximate* 90^{th} *percentile for the distribution of the total contributions received.*

→: $S = Y_1 + \cdots + Y_n$ $\quad n = 2025$ w/ Y_i s.t. $\mu_Y = 3125$, $\sigma_Y = 250$

$\mathbb{E}[S] = (2025)(3125) = \mu_S$

$SD[S] = 250\sqrt{2025} = 250(45) = \sigma_S$

$\pi = ?$ such that $\mathbb{P}[S \leq \pi] \approx 0.90$

1^{st} Find the 90^{th} percentile of $Z \sim N(0,1)$

$\boxed{z^* = 1.28}$ from the std normal tables

2^{nd} Apply the linear transform to the above

$\mathbb{P}[Z \leq z^*] = 0.90$

$\mathbb{P}[\boxed{\mu_S + \sigma_S \cdot Z} \leq \mu_S + \sigma_S \cdot z^*] = 0.90$

$\mathbb{P}[\;S\; \leq \boxed{(2025)(3125) + 250(45) \cdot 1.28}\,] \approx 0.90$

π = answer

[1] http://thereallyterribleorchestra.com/wordpress/