CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 1 - Sampling Distributions

Problem 1

To determine whether a bottling machine is working satisfactorily, a production line manager randomly samples ten 12-ounce bottles every hour and measures the amount of beverage in each bottle. The mean $\bar{x}$ of the 10 fill measurements is used to decide whether to readjust the amount of beverage delivered per bottle by the filling machine.

If records show that the amount of fill per bottle is normally distributed, with a standard deviation of .2 ounce, and if the bottling machine is set to produce a mean fill per bottle of 12.1 ounces, what is the approximate probability that the sample mean $\bar{x}$ of the 10 test bottles is less than 12 ounces?

Solution:

The mean and standard deviation are given that $\mu = 12.1$ and $\sigma = 0.2$. To find the probability that $\bar{x}$ is less than 12 ounces, we have

$$P(\bar{x} < 12) = P(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{12 - 12.1}{0.0632}) = P(Z < -1.58) \approx 0.057$$

Problem 2

A college C would like to have 1050 freshmen, and cannot accommodate more than 1060. Assume that each applicant accepts with probability $p = 0.6$ and that the acceptance can be modelled with Binomial distribution. If the college accepts 1700 freshmen, what is the probability that it will have too many acceptances?

Solution:

If this college accepts 1700 students, the expected number of people who will begin the studies is $np = 1700 \cdot 0.6 = 1020$ and the standard deviation for the number that accept is

$$\sqrt{np(1-p)} = \sqrt{1700 \cdot 0.6 \cdot 0.4} \approx 20$$

To find the probability that the number that accept is higher than 1060 we calculate

$$\begin{aligned} P(S_{1700} > 1060) &= P(S_{1700} \geq 1061) \\ &= P(\frac{S_{1700} - 1020}{20} \geq \frac{1060.5 - 1020}{20}) \\ &\approx 0.0214 \end{aligned}$$

Hence the probability is quite small so the college is fairly safe using this policy.

Problem 3

The proportion of individuals with an Rh-positive blood type is 85%. You have a random sample of $n = 500$ individuals. What is the probability that the sample proportion $\hat{p}$ lies between 83% and 88%?

Solution:

For this binomial random variable with $n = 500$ and $p = 0.85$, the mean and standard error of $\hat{p}$ are

$$\mu = p = 0.85 \quad \text{and} \quad SE = \sqrt{\frac{0.85 \cdot 0.15}{500}}$$

We can then compute the probability by

$$\begin{aligned} P(0.83 < \hat{p} < 0.88) &= P\left(\frac{0.83 - \mu}{SE} < Z < \frac{0.88 - \mu}{SE}\right) \\ &= P(Z < 1.88) - P(Z < -1.25) \\ &\approx 0.8643 \end{aligned}$$

Problem 4

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.
What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

Solution:

In order to calculate $P(\bar{X} > 1.4mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim \mathcal{N}(1.3, \frac{0.3}{\sqrt{60}}) = \mathcal{N}(1.3, 0.0387)$$

Thus,

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 \\ &= 0.0049 \end{aligned}$$

Problem 5

Suppose that an insurance company has 10,000 policyholders. The expected yearly claim per policyholder is \$240 with a standard deviation of \$800. What is the approximate probability that the total yearly claims $S_{10,000} >$ \$2.6 million?

Solution:

Note that $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$ by the sum variance law. Since X_i's are identical and independent distributed, we have

$$Var\left(\sum_i X_i\right) = n \cdot \text{Var}(X)$$

and

$$\sigma\left(\sum_i X_i\right) = \sqrt{n} \cdot \sigma(X)$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left(S_{10,000}\right) = 10,000 \times 240 = 2,400,000 \\
&\sigma\left(S_{10,000}\right) = \sqrt{10,000} \times 800 = 80,000 \\
&P\left(S_{10,000} > 2,600,000\right) \\
=&P\left(\frac{S_{10,000} - 2,400,000}{80,000} > \frac{2,600,000 - 2,400,000}{80,000}\right) \\
\approx&P(Z > 2.5) = 0.0062
\end{aligned}$$

Additional Drill Questions (Do not discuss in the tutorial)

Problem 6

Suppose you roll a 6-sided die 10 times. Let X be the total value of all 10 dice, *i.e.*, $X = X_1 + X_2 + \cdots + X_{10}$. You win the game if $X \leq 25$ or $X \geq 45$. Use the central limit theorem to calculate the probability that you win.

Solution:

Recall that the expected value and variance of X_i are

$$\mathbb{E}(X_i) = \sum_{x=1}^{6} x \cdot P(X_i = x) = 3.5$$

and

$$\text{Var}(X_i) = \frac{1}{6}\sum_{x=1}^{6} (x - \mathbb{E}(X_i))^2 = \frac{35}{12}$$

Thus,

$$\begin{aligned}
P(X \leq 25 \text{ or } X \geq 45) &= 1 - P(25.5 \leq X \leq 44.5) \\
&= 1 - P\left(\frac{25.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \leq \frac{X - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \leq \frac{44.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}}\right) \\
&\approx 0.0786
\end{aligned}$$

Problem 7

Suppose you have a new algorithm and want to test its running time. You have an idea of the variance of the algorithm's run time: $\sigma^2 = 4\,\text{second}^2$ but you want to estimate the mean: $\mu = t\,\text{second}$. You can run the algorithm repeatedly. How many trials do you have to run so that your estimated runtime is $t \pm 0.5$ with 95% certainty?

Solution:

Let X_i be the run time of the i-th run (for $1 \leq i \leq n$). We want to know the smallest n such that

$$P(t - 0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} \leq t + 0.5) \geq 0.95$$

We can rewrite the probability inequality according to the central limit theorem,

$$\begin{aligned} 0.95 &\leq P\left(\frac{-0.5}{2/\sqrt{n}} \leq \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - t}{2/\sqrt{n}} \leq \frac{0.5}{2/\sqrt{n}}\right) \\ &= P\left(\frac{-0.5\sqrt{n}}{2} \leq Z \leq \frac{0.5\sqrt{n}}{2}\right) \end{aligned}$$

Now we can calculate the value of n that makes this inequality hold.

$$\Phi(\frac{\sqrt{n}}{4}) - \Phi(-\frac{\sqrt{n}}{4}) \geq 0.95 \Rightarrow n \geq 61.4,$$

where $\Phi(x)$ represents the CDF of the standard normal distribution, evaluated at the values in x. Thus, it takes at least 62 runs. (You need to use a computer to solve the last step to obtain the final answer.)

Problem 8

Suppose $X_1, X_2, \ldots, X_{30}$ are independent Poisson random variables with mean $\mathbb{E}(X_i) = 2$ and $\text{Var}(X_i) = 2$. Use the central limit theorem to approximate

$$P\left(\sum_{i=1}^{30} X_i > 50\right).$$

Solution:

Since each X_i has mean and variance 2, we have

$$\text{Var}\left(\sum_{i=1}^{30} X_i\right) = \mathbb{E}\left(\sum_{i=1}^{30} X_i\right) = 60$$

By the CTL,

$$P\left(\sum_{i=1}^{30} X_i > 50\right) = P\left(Z > \frac{50 - 60}{\sqrt{60}}\right) = P(Z > -1.291) \approx 0.9016$$

Problem 9 (Not included in quizzes)

Let X_i =weight of car i and Y_i =fuel in gallons to go 100 miles. We use the model $Y_i = \theta X_i + \epsilon_i$ where ϵ_i are independent errors with

$$\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$$

How do we estimate θ from data? We minimize the least squares criterion

$$SS(\theta) = \sum_{i=1}^{n} (Y_i - \theta X_i)^2$$

which is minimized by

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

What is the distribution of $\hat{\theta} - \theta$? (Note X_i is not a random variable in this question.)

Solution:

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i(\theta X_i + \epsilon_i)}{\sum_{i=1}^n X_i^2} = \theta + \frac{\sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2}$$

Let $Z_i = X_i \epsilon_i$. We have $\mathbb{E}(Z_i) = 0$ and $Var(Z_i) = X_i^2 \sigma^2$. Thus, according to CLT

$$\frac{\sum_{i=1}^n (X_i \epsilon_i - 0)}{\sqrt{\sum_{i=1}^n X_i^2 \sigma^2}} \sim \mathcal{N}(0,1)$$

Note that from the first equation, we can rewrite $\hat{\theta} - \theta$ as

$$\hat{\theta} - \theta = \frac{\sum_{i=1}^n X_i \epsilon_i}{\sqrt{\sum_{i=1}^n X_i^2}} \times \frac{1}{\sqrt{\sum_{i=1}^n X_i^2}} \times \frac{\sigma}{\sigma} = \frac{\sum_{i=1}^n X_i \epsilon_i}{\sqrt{\sum_{i=1}^n X_i^2 \sigma^2}} \times \frac{\sigma}{\sqrt{\sum_{i=1}^n X_i^2}}.$$

Thus,

$$\frac{\hat{\theta} - \theta}{\frac{\sigma}{\sqrt{\sum_{i=1}^n X_i^2}}} \sim \mathcal{N}(0,1).$$

Clearly, $\left(\frac{\sigma}{\sqrt{\sum_{i=1}^n X_i^2}} \right)$ and θ are standard deviation and mean of $\hat{\theta}$. Therefore,

$$(\hat{\theta} - \theta) \sqrt{\sum_{i=1}^n X_i^2} \sim \mathcal{N}(0, \sigma^2)$$

and

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{\sum_{i=1}^n X_i^2} \right)$$