MAT 105 - Summary of Lecture 19
Tuesday, March 22, 2016

> *There emerges the vision of the fair coin, the biased coin. This coin exists in some mental universe and all modern writers on probability theory have access to it. They toss it regularly and they speculate about what they 'observe.* (Davis, Philip and Hersh, Reuben)

Central Limit Theorem (CLT)

Recall that the central limit theorem states that given a large set of data, it will be distributed according to a bell curve, irrelevant of what kind of data it is. For example, we can look at values of a roll of a 6-six sided die, or at how often a six comes up, or how long it takes to hit the target at darts, this data will be distributed according to a normal distribution, once we repeat the experiment enough times. The only difference is that the bell curves might have different mean and standard deviation, in which case we need to normalize the data to arrive at $N(0,1)$ for which we can use the z-table to calculate probabilities. More formally, CLT says

Theorem (Central Limit Theorem): If $X_1, X_2, \dots$ are independent trials of an experiment, each trial having the same distribution with expectation μ (finite) and variance σ^2 with $0 < \sigma^2 < \infty$, and we let $S_n = X_1 + X_2 + \cdots + X_n$, then $S_n \approx N(n\mu, n\sigma^2)$ and

$$P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \leq b\right) \approx \phi(b) - \phi(a), \qquad \text{for } n \text{ very large.}$$

Example: Roll 10 fair 6-sided dice. Approximate the probability that the sum is between 30 and 40.
Let $X_1, X_2, \dots$ be the value on die $1, 2, \dots$. Then they have the same mean and variance which we found to be,

$$E[X_1] = 1 \cdot P(X_1 = 1) + 2 \cdot P(X_1 = 2) + \cdots + 6 \cdot P(X_1 = 6) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

$$\begin{aligned} Var(X_1) &= (1-3.5)^2 \cdot P(X_1 = 1) + (2-3.5)^2 \cdot P(X_1 = 2) + \cdots + (6-3.5)^2 \cdot P(X_1 = 6) \\ &= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + \cdots + (6-3.5)^2 \cdot \frac{1}{6} = \frac{35}{12} \end{aligned}$$

Therefore, by CLT, $S_{10} \approx N(35, 350/12)$ and

$$P\left(30 \leq S_{10} \leq 40\right) = P\left(\frac{30-35}{\sqrt{\frac{350}{12}}} \leq \frac{S_{10}-35}{\sqrt{\frac{350}{12}}} \geq \frac{40-35}{\sqrt{\frac{350}{12}}}\right) = \phi(.92) - \phi(-.92) = .8212 - (1 - .8212) = .6424$$

Confidence intervals

Before we go any further, let us find the following probabilities for $z = N(0,1)$:
(a) $P(-1 \leq z \leq 1) = \phi(1) - \phi(-1) = \phi(1) - (1 - \phi(1)) = 2\phi(1) - 1 = 1(.8413) - 1 = .6826.$

(b) $P(-2 \leq z \leq 2) = \phi(2) - \phi(-2) = \phi(2) - (1 - \phi(2)) = 2\phi(2) - 1 = 1(.9772) - 1 = .9544.$

(c) $P(-3 \leq z \leq 3) = \phi(3) - \phi(-3) = \phi(3) - (1 - \phi(3)) = 2\phi(3) - 1 = 1(.9987) - 1 = .9974.$

Interpretation: For a normal random variable (or a random variable that can be approximated by a normal)

(a) the probability that the outcome is within 1 standard deviation from the mean is about 68%.

(b) the probability that the outcome is within 2 standard deviations from the mean is about 95%.

(c) the probability that the outcome is within 3 standard deviations from the mean is about 99%.

Definition: The 95% **confidence interval** for a normal random variable with mean μ and standard deviation σ is the interval $[\mu - 2\sigma, \mu + 2\sigma]$. The 99% **confidence interval** for a normal random variable with mean μ and standard deviation σ is the interval $[\mu - 3\sigma, \mu + 3\sigma]$.

Intuitively, this means that when sampling a normal random variable, the outcomes fall in the 95% confidence interval 95% of the time, or we can be 95% confident that the outcome falls in this interval. A similar interpretation holds for the 99% confidence interval.

We mentioned above that this also holds for random variables that can be approximated by normal distributions, or in other words, that can be normalized (by subtracting a mean and dividing by standard deviation). This begs the question: can all random variables be normalized? The answer is NO. So what can we normalize?

(i) $X = N(\mu, \sigma^2)$ can be normalized by subtracting μ and dividing by σ (from rules about normal distributions).
$$z = \frac{X - \mu}{\sigma}$$

(ii) $X = Binomial(n, p)$ can be normalized by subtracting np and dividing by $\sqrt{np(1-p)}$ (by normal approximation to the binomial)
$$z = \frac{X - np}{\sqrt{np(1-p)}}$$

(iii) Sums of random variables can be normalized. If $S_n = X_1 + X_2 + \cdots + X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$, then S_n is normalized by subtracting $n\mu$ and dividing by $\sqrt{n}\sigma$ (by CLT)
$$z = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

(iv) Proportions or averages of random variables can be normalized. Again, if $S_n = X_1 + X_2 + \cdots + X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$, then the average $\overline{X}_n = \frac{S_n}{n}$ is normalized by subtracting μ and dividing by $\frac{\sigma}{\sqrt{n}}$ (by CLT)
$$z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

We have seen how to normalize all quantities above, except for (iv). Here is how one derives this normalization. If $S_n = X_1 + X_2 + \cdots + X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$, using the fact that $X_1, X_2, \ldots$ are independent, and recalling properties for expectation and variance of sums, we find the mean and standard deviation of $\overline{X}_n$:

$$E[\overline{X}_n] = E\left[\frac{S_n}{n}\right] = \frac{1}{n}E[X_1 + X_2 + \cdots + X_n] = \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right) = \frac{1}{n}(n\mu) = \mu.$$

$$Var(\overline{X}_n) = Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}Var(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \cdots + Var(X_n)\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

$$StDev(\overline{X}_n) = \sqrt{Var\left(\overline{X}_n\right)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Examples:

1. How does one find the 80% confidence interval for $z = N(0,1)$?

   We want

   $$P(-a \le z \le a) = .8 \;\Leftrightarrow\; \phi(a) - \phi(-a) = .8 \;\Leftrightarrow\; \phi(a) - (1 - \phi(a)) = .8 \;\Leftrightarrow\; \phi(a) = .9 \;\Leftrightarrow\; a = 1.29$$

   Here, we found a by searching the z-score that gives a probability of 0.9 in the z-table. Thus, one needs to be within 1.29 standard deviations from the mean in order to be 80% confident of the results.

2. What is the 95% confidence interval for the proportion of heads in 100 coin flips if our estimated proportion is 0.64?

   We have $\mu = 0.64$ which we assume to be the true proportion for this coin, that is, $p = 0.64$ is the probability that a coin will land on heads. Then if we let $X_1, X_2, \ldots$ be 1 if the 1st, 2nd, etc coins land on heads, $\mu = E[X_i] = p = 0.64$ and $\sigma^2 = Var(X_i) = p(1-p) = (0.64)(0.36)$, so the mean for the proportion is $\mu = 0.64$ and the standard deviation for the proportion is $\dfrac{\sigma}{\sqrt{100}} = \dfrac{.48}{10} = .048$. Therefore, the 95% confidence interval for the proportion of heads is

   $$[0.64 - 2(0.048), 0.64 + 2(0.048)] = [0.544, 0.736].$$

   Note that this tells us that we can be 95% sure that the coin is not fair.

3. What is the probability that the proportion of heads in 100 fair coin flips is less than 60%?

   We compute $\overline{X}_{100} = \dfrac{X_1 + \cdots + X_{100}}{100}$, where $X_1, X_2, \ldots$ are $Bernoulli(0.5)$, that is, they are 1 if the coin lands on heads and zero otherwise. Their mean is $\mu = 0.5$ and variance $\sigma^2 = (0.5)(1 - 0.5) = 0.25$, so $\sigma = 0.5$. Then from above,

   $$P(\overline{X}_{100} < 0.6) = P\left(\frac{\overline{X}_{100} - 0.5}{\frac{0.5}{\sqrt{100}}} < \frac{0.6 - 0.5}{\frac{0.5}{\sqrt{100}}}\right) = \phi(2) = 0.9772.$$

   Alternately, we can simply compute the z-score for 0.6, which is $\dfrac{0.6 - 0.5}{\frac{0.5}{\sqrt{100}}} = 2$ and the probability of less than 2 standard deviations from the mean is $\phi(2) = 0.9772$.

4. What is the 95% confidence interval for the number of heads in 100 tosses of a fair coin?

   Let X count the number of heads in 100 tosses. Then $X = Binomial(100, 0.5)$, so $E[X] = 100(.5) = 50$, $Var(X) = 100(.5)(.5) = 25$ and $StDev(X) = \sqrt{25} = 5$. The 95% confidence interval is

   $$[50 - 2 \times 5, 50 + 2 \times 5] = [40, 60].$$

   This means that 95% of the time, the number of heads in 100 tosses will be between 40 and 60.