Confidence Intervals

If we have a box of numbered tickets and know the distribution of the numbers, then we can use the expected value and standard error, together with the Central Limit Theorem, to estimate the chance that a random sample of size n will have sum (or average) in a specified range.

Example: A box contains 500 $\boxed{1}$s, 500 $\boxed{9}$s and 280 $\boxed{5}$s. We take a random sample of size $n = 100$.

a) The sum of the random sample will be around _____ give or take about _____.

There are 1280 tickets in the box. We can use R to get the mean and SD.

```
> x = c(rep(1, 500), rep(9, 500), rep(5, 280))
> mean(x)
[1] 5
> sd(x)*sqrt(1279/1280)
[1] 3.53
```

The expected value of the sum is $\text{EV}_{\text{sum}} = n \cdot \text{AV}_{\text{box}} = 100 \times 5 = 500$. The standard error for the sum is $\text{SE}_{\text{sum}} = \sqrt{n} \cdot \text{SD}_{\text{box}} = \sqrt{100} \times 3.53 = 35.3$.

b) What is the chance that the sum will be between 470 and 530?

The z scores are $z_1 = (470 - 500)/35.3 = -0.85$ and $z_2 = (530 - 500)/35.3 = 0.85$. By the Central Limit Theorem, we can assume that the sample sum has a normal distribution. The area under the normal curve between ± 0.85 is $\text{pnorm}(0.85) - \text{pnorm}(-0.85) = 60.5\%$.

c) The average of the random sample will be around _____ give or take about _____.

The expected value of the average is $\text{EV}_{\text{av}} = \text{AV}_{\text{box}} = 5$. The standard error for the average is $\text{SE}_{\text{av}} = \text{SD}_{\text{box}}/\sqrt{n} = 3.53/\sqrt{100} = 0.35$.

b) What is the chance that the average will be between 4.5 and 5.5?

The z-scores are $z_1 = (4.5 - 5)/0.35 = -1.43$ and $z_2 = (5.5 - 5)/0.35 = 1.43$. So the chance is about $\text{pnorm}(1.43) - \text{pnorm}(-1.43) = 84.7\%$.

What if we don't know the contents of the box? That is usually the case in real life and the reason why we take random samples. A *confidence interval* gives an estimate of the average of the numbers on the tickets. Suppose we take a random sample of size n. Let M be the average for the sample and let S be the sample standard deviation. A 95% confidence interval on the true average of the tickets is

$$M \pm 2\,S/\sqrt{n}$$

1. A box has 100,000 tickets. 40,000 are have a 1 on them and the rest have a 0 on them. A simple random sample of 500 is taken. Calculate the expected value of the percentage of 1s in the sample and the standard error of the percentage.

2. In a certain city, there are 100,000 persons age 18 to 24. A simple random sample of 500 people is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in that city who are currently enrolled in college. Put a give-or-take number on the estimate.

3. In a simple random sample of 100 graduates from a certain college, 48 were earning $50,000 a year or more. Estimate the percentage of all graduates of that college earning $50,000 a year or more. Put a give-or-take number on the estimate.

4. Use the data in #2 to find a 95% confidence interval on the percentage of persons age 18 to 24 who are currently enrolled in college.

5. Use the data in #3 to find a find a 95% confidence interval on the percentage of all graduates from that college earning $50,000 a year or more.

6. Suppose the sample in #2 had turned up 204 persons currently enrolled in college. Find a new 95% confidence interval on the percentage of persons age 18 to 24 who are currently enrolled in college.

7. Suppose the sample in #3 had turned up 45 graduates earning $50,000 or more. Find a new 95% confidence interval on the percentage of all graduates from that college earning $50,000 a year or more.