

**The Central Limit Theorem (CLT)**  
Stat 1, Spring 2011, Swarthmore College

The Normal distribution is a good approximation to many (not all!) real-life distributions. This is in part due to a remarkable result from probability theory called “the Central Limit Theorem,” or CLT. This theorem states that sums or averages of a large number of independent values, taken from *any* distribution, will have a sampling distribution that is close to Normal (or at least much closer to Normal than the variable’s population distribution).

**This is AMAZING!!!** And it is the reason why most of the sampling distributions we have simulated or looked at have been roughly symmetric and bell-shaped.

**How it works:** A Normal distribution is completely characterized by its mean and standard deviation, so this is all you need to approximate a sampling distribution based on the CLT. For a population distribution with mean  $\mu_x$  and standard deviation  $\sigma_x$ , the average of  $n$  independent values will have mean  $\mu_{\bar{x}} = \mu_x$  and standard deviation  $\sigma_{\bar{x}} = \sigma_x/\sqrt{n}$ :

$$\mu_{\bar{x}} = \mu_x; \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}.$$

The CLT implies that, for large  $n$ , the sampling distribution of  $\bar{x}$  will be approximately Normal with this mean and standard deviation.

**Example:** Suppose that resting pulse rates for Swarthmore students follow a distribution that has mean  $\mu_x = 74$  beats per minute (bpm) and standard deviation of  $\sigma_x = 8$  bpm. Choose a simple random sample of  $n = 16$  students and record  $\bar{x}$ , the average of their pulse rates. If we repeat this many times and make a histogram of the averages, the distribution will be very close to Normal, with mean  $\mu_{\bar{x}} = \mu_x = 74$  bpm and standard deviation  $\sigma_{\bar{x}} = \sigma_x/\sqrt{n} = 8/\sqrt{16} = 2$  bpm.

**Practice problems:**

For parts 1-4, assume the average value  $\bar{x}$  follows a Normal distribution with mean  $\mu_{\bar{x}} = 74.0$  and standard deviation  $\sigma_{\bar{x}} = 2.0$ .

1. About what percent of the time would the average pulse rate be greater than 74 bpm?
  
2. About what percent of the time would the average pulse rate be greater than 76 bpm?
  
3. About what percent of the time would the average be less than 72 bpm?
  
4. About what percent of the time would we see  $|\bar{x} - 74| > 2$  bpm?
  
5. \*Recompute your answers to 1-4 if you were to choose SRSs of  $n = 64$  instead of  $n = 16$ .

### The CLT for sample proportions ( $\hat{p}$ )

Sample proportions are a special kind of average. For example, to find the proportion of times a coin lands heads in  $n$  tosses, we add 1 for each head and 0 for each tail, then divide this total number of “successes” by the number of “trials”  $n$ . Any sample proportion  $\hat{p}$  is the average of a sample of 0’s and 1’s. In baseball, a batting average is the average number of times a batter hits safely in  $n$  “at-bats” (allowable attempts at getting on base safely). Often this proportion is multiplied by 1000 to give values like 300 or 240 ( $\hat{p} = 0.3$  or  $\hat{p} = 0.24$ ).

The mean value for a sample proportion  $\hat{p}$  is the expected proportion of 1’s or the *probability*  $p$  of a “success” on any particular trial - e.g., the probability of a coin landing heads on each coin flip, or of a batter getting a “hit” on each at-bat (apologies to people who are not familiar with baseball).

For fair coin flips, the probability of a coin landing heads is  $p = 0.5$ . Each flip represents either 0 or 1 heads, so every outcome deviates from the mean by exactly 0.5. The standard deviation for a coin flip is  $\sigma = 0.5$ . If we flip a coin  $n$  times, the proportion of heads (or the average number of heads) has standard deviation  $\sigma_{\hat{p}} = 0.5/\sqrt{n}$ . For a random process with arbitrary “success” probability  $p$ , the mean and standard deviation of  $\hat{p}$  are

$$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \leq \frac{0.5}{\sqrt{n}}, \quad 0 \leq p \leq 1.$$

**Example:** Imagine flipping a coin  $n = 100$  times and computing the proportion of heads  $\hat{p}$  – the number of heads divided by 100. The mean of  $\hat{p}$  is  $\mu_{\hat{p}} = 0.5$  and the standard deviation is  $\sigma_{\hat{p}} = \sqrt{0.5(1-0.5)/100} = 0.05$ . The CLT implies that the sampling distribution of  $\hat{p}$  is approximately Normal with this mean 0.5 and standard deviation 0.05.

#### Practice Problems:

For problems 1-4, assume you flip a fair coin  $n = 100$  times and compute  $\hat{p}$ .

1. What is the approximate probability you will get  $\hat{p} > 0.6$ ?  $\hat{p} < 0.4$ ?
2. About what proportion of the time would you get  $|\hat{p} - 0.5| > 0.1$ ?
3. What is the approximate probability you will get  $\hat{p} > 0.55$ ?  $\hat{p} < 0.45$ ?
4. What is the approximate probability you will get  $\hat{p} > 0.51$ ?  $\hat{p} < 0.49$ ?
5. \*Recompute your answers to 1-4 if you flipped the coin  $n = 25$  times instead of  $n = 100$ .