Week 8 Recap

Monday October 9: **Fall break**

Wednesday October 11: First limit theorems.

The law of large number for Bernoulli random variables: Assume that $(X_i)_1^\infty$ is a sequence of independent Bernoulli random variable with parameter p. Then, for any $\epsilon > 0$, $\lim_{n\to\infty} P(|\frac{1}{n}\sum_1^n X_i - p| > \epsilon) = 0$.

The central limit theorem for binomial random variables (when $np(1-p)$ is large enough): Let S_n be a Binomial n, p random variable, e.g., the number of successes in a sequence of n repeated identical independent experiments with probability of success equal to p. If np is large enough (in practice, $np(1-p) \geq 10$), for any fixed $a, b \in \mathbb{R}$

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a)$$

where
$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$. This function is described (approximately) by a **normal table (https://en.wikipedia.org/wiki/Standard_normal_table)** (there is one at the end of the book and you should learn how to use it).

How to to use the binomial CLT in practice: The first step is to identify the problem at hand as being a problem that can be modeled using a binomial distribution. For a real life practical problem (or problems resembling a real life practical problem), this requires to make some assumptions: we are counting the occurrences of a certain event in repeated experiments and we need to feel comfortable assuming that the "experiments" are repetition of a well defined experiment (identical experiments) and that the results of these different experiments are independent of each others.

This first step then tells us that we are looking at a random variable S following a binomial distribution and we need to find the parameters n, p of this binomial distribution. In general, the parameter n is clear from the context once we have identified the exact experiment that is repeated n times. The parameter p can usually be computed from the description of that experiment.

The third step is to check that we are in the the normal approximation regime. For us in this course, we take that to mean that $np(1-p) \geq 10$. If p is smaller than $1-p$, this means that n must be large enough and p not small so that the product np is large enough.

Finally, let's look at what the result says in terms of the random variable S. Clearly

$$P\left(a \leq \frac{S-np}{\sqrt{np(1-p)}} \leq b\right) = P\left(np + a\sqrt{np(1-1)} \leq S \leq np + b\sqrt{np(1-p)}\right)$$

and the given approximation reads

$$P\left(np + a\sqrt{np(1-1)} \leq S \leq np + b\sqrt{np(1-p)}\right) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx = \Phi(b) - \Phi(a).$$

In words, we obtain an acceptable approximation for those events of the type $S \in I$ where the interval I is located relatively close to the mean $E(S) = np$ at the scale $\sqrt{np(1=p)} = \sqrt{\mathrm{Var}(S)}$.

Learn about the continuity correction and how to use it when needed:

A binomial $\mathrm{Bin}(n,p)$ random variable S take integer values and it follows that, for integers $0 \leq k_1 \leq k_2$,

$$P(k_1 \leq S \leq k_2) = P\left(k_1 - \frac{1}{2} \leq S \leq k_2 + \frac{1}{2}\right).$$

Working with the right-hand side expression when using the normal approximation leads to slightly more accurate
results, especially when the the integers k_1, k_2 are close to each other or the variance $np(1-p)$ is not very large.
Switching from $P(k_1 \leq S \leq k_2)$ to $P\left(k_1 - \frac{1}{2} \leq S \leq k_2 + \frac{1}{2}\right)$ before using the normal approximation is called the continuity correction. It is a very natural think to do.