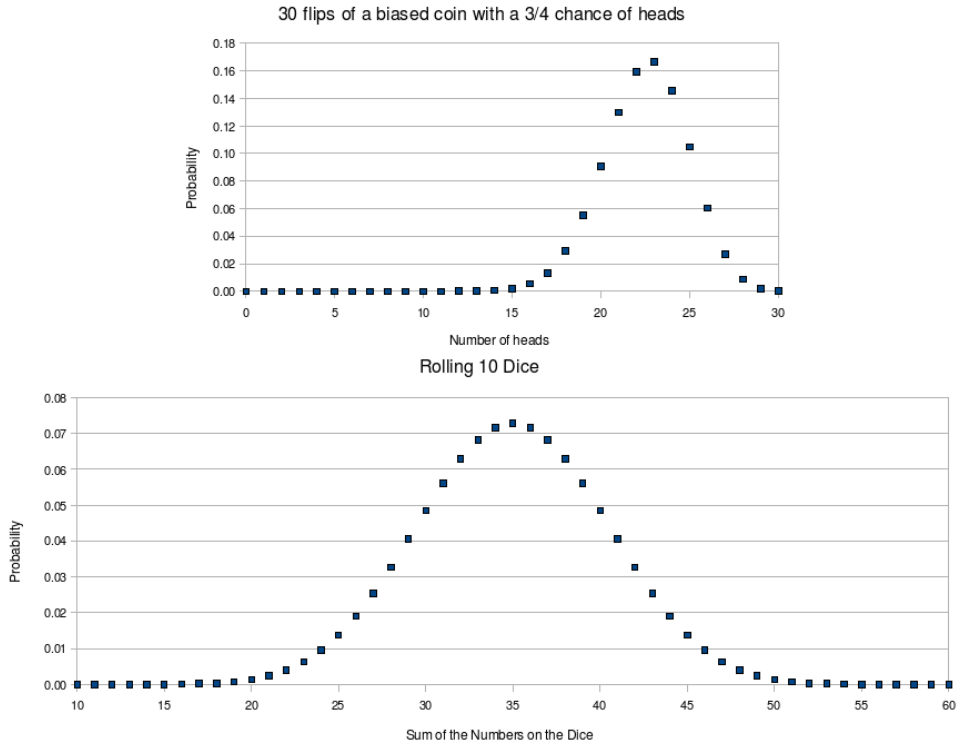
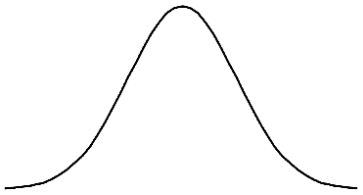
The Central Limit Theorem

The figure below shows the graphs of two random variables.

- The fist random variable is the number of heads obtained after flipping a biased coin 30 times, where the chance of getting heads on a single flip is 3/4.
- The second random variable is the sum of the numbers you get after rolling a (fair six-sided) die 10 times.



Notice how similar these plots look. Of course, these random variables have different expected values and standard deviations, but their basic shape is the same. It's something like this:



This is rather astonishing. Just about the only thing these random variables have in common is that they can be written as the sum of a bunch of independent random variables. What's even more amazing is that this one property: "being the sum of a bunch of independent random variables" is enough to make any other random variable look like this as well. This is essentially what the central limit theorem says.

Before we continue, recall a couple rules about expected value and standard deviation. If a and b are constants, and f is a random variable, then

$$E(af+b) = aE(f)+b \quad \text{and} \quad \sigma(af+b) = a\sigma(f)$$

Now the expected value (which measures where the center of the distribution is) and standard deviation (which measures how spread out the distribution is) of our random variables can be anything. However,

we can remove these differences by considering the random variable $\frac{f-E(f)}{\sigma(f)}$ instead of f. This new random variable has expected value 0, and standard deviation 1: since $E(f)$ and $\sigma(f)$ are constants, by the two rules above,

$$E\left(\frac{f}{\sigma(f)} - \frac{E(f)}{\sigma(f)}\right) = \frac{E(f)}{\sigma(f)} - \frac{E(f)}{\sigma(f)} = 0 \quad \text{and } \sigma\left(\frac{f-E(f)}{\sigma(f)}\right) = \frac{\sigma\left(f-E(f)\right)}{\sigma(f)} = \frac{\sigma(f)}{\sigma(f)} = 1$$

Removing different expected values and standard deviations this way turns out to be exactly the way which makes their distributions almost exactly the same. We're ready to state the central limit theorem:

The (Fuzzy) Central Limit Theorem:[1] If f is a random variable that's the sum of lots of independent random variables: $f = f_1 + f_2 + \cdots + f_n$, then the probability that $\dfrac{f-E(f)}{\sigma(f)}$ is between a and b (for $a < b$) is approximately the same for any such f, and is $\approx \displaystyle\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$. Note that the f_i don't all have to have the same distribution, even though the examples we've given have this property.

The function $\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ is called a Gaussian, and this is the function that's graphed on the previous page. Finding an antiderivative of this function is impossible (unless we use functions defined by integrals) which is why we've left $\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$ unsimplified. To do these integrals, you have to approximate them. You can do this on a calculator, in spreadsheets[2], computer algebra systems[3], from tables, at webpages (like the one in the class notes), and many other places.

Let's look at some basic properties. First, a "sanity check": the probability that $\frac{f-E(f)}{\sigma(f)}$ is between $-\infty$ and ∞ should be 1, and indeed $\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = 1$ (you can actually do this particular integral, but it involves some tricky vector calculus). Next, notice that $\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ is even (it's symmetric around 0). That means that the integrals $\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = 1$ and $\int_{-b}^{-a} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = 1$ will be equal. So, for instance,

$$P(\frac{f-E(f)}{\sigma(f)} \geq r) \approx P(\frac{f-E(f)}{\sigma(f)} \leq -r) \approx \int_r^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

In particular $\int_0^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} = \frac{1}{2}$, thus, the chance of f being greater than its expected value is approximately 1/2. Also, by adding $P(\frac{f-E(f)}{\sigma(f)} \geq r)$ and $P(\frac{f-E(f)}{\sigma(f)} \leq -r)$ for $r \geq 0$, we get

$$P(|\frac{f-E(f)}{\sigma(f)}| \geq r) \approx 2\int_r^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

In the lecture notes, the central limit theorem was stated using this approximation.

Now let's do a few examples. As we'll see, the central limit theorem has huge significance as a calculational tool. It allows us to accurately estimate probabilities that would almost impossible to find otherwise; we don't have any other tools for estimating the probability that a random variable is between two numbers except for direct calculation.

---

[1] Of course, what's stated above isn't actually a theorem. What the central limit theorem really says is that as we add more and more independent random variables to $f = f_1 + f_2 + \cdots + f_n$, then the limit as $n \to \infty$ of $P\left(a \leq \frac{f-E(f)}{\sigma(f)} \leq b\right)$ is exactly $\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$. What we've written above is a consequence of the central limit theorem that we'll be using. Note also that the central limit theorem has some modest requirements on the f_i that we haven't stated (but which will be true in the problems we do). For instance, they can't conspire to make f attain one value with nonzero probability in the limit $n \to \infty$.

[2] For instance, in Excel, `normdist(x,0,1,1)` does the integral from $-\infty$ to x

[3] For instance, in Matlab, `normcdf(x,0,1)` does the integral from $-\infty$ to x

1. What's the approximate probability of getting a sum between 25 and 40 when you roll 10 dice?

   Let's call the random variable giving the sum of the dice f. The expected value of our random variable is $E(f) = 35$, and the standard deviation is $\sigma(f) = \sqrt{\frac{350}{12}}$. Now

   $$25 \leq f \leq 40 \quad \leftrightarrow \quad -10 \leq f - 35 \leq 5 \quad \leftrightarrow \quad \frac{-10}{\sqrt{\frac{350}{12}}} \leq \frac{f-35}{\sqrt{\frac{350}{12}}} \leq \frac{5}{\sqrt{\frac{350}{12}}}$$

   So we need to estimate the probability that $-1.852 \leq \frac{f-E(f)}{\sigma(f)} \leq .926$. By the central limit theorem this is $\approx \int_{-1.852}^{.926} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \approx .79$. The actual probability is about .8182, so this is a fair estimate.

2. On the last quiz, we considered flipping a fair coin labeled "-1" and "1", 10,000 times. Let f be the random variable giving the sum of all the numbers we get. Then $E(f) = 0$, and $\sigma(f) = 100$. Approximate the probability that $|f| \geq 1,000$, and compare it to the upper bound you found using Chebyshev's theorem.

   By the values of $E(f)$ and $\sigma(f)$ given above,

   $$|f| \geq 1{,}000 \quad \leftrightarrow \quad \left|\frac{f - E(f)}{\sigma(f)}\right| \geq 10$$

   By the central limit theorem, $P(\left|\frac{f-E(f)}{\sigma(f)}\right| \geq 10) \approx 2\int_{10}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \approx 1.524 \times 10^{-23}$. The probability is extremely small! Much much smaller than the upper bound of 0.01 we got using Chebyshev's theorem.