Chapter 6 Notes and elaborations for STAT 141-Introduction to Statistics

Assignment:
Chapter 6 is also pretty good, so I'll be following the text pretty closely. It is of the upmost importance that you are able to find areas under the normal curve. You will need to be able to do this for the remainder of the semester. Do read the book for sections 6.1 – 6.3. These notes are not comprehensive.

!Section 6.1 and 6.2: p.322, and 334-337!
The bit on determining normality by eye is just plain wrong. One can't just assume that a distribution is normal because it looks bell shaped. There are some hard distributions out there to deal with that look very much like a normal distribution. Some of the other parts are generally correct, but assessing normality is something that is fairly difficult to verify and is beyond the scope of this class. I will not ask you to determine if an unknown data set is normally distributed.

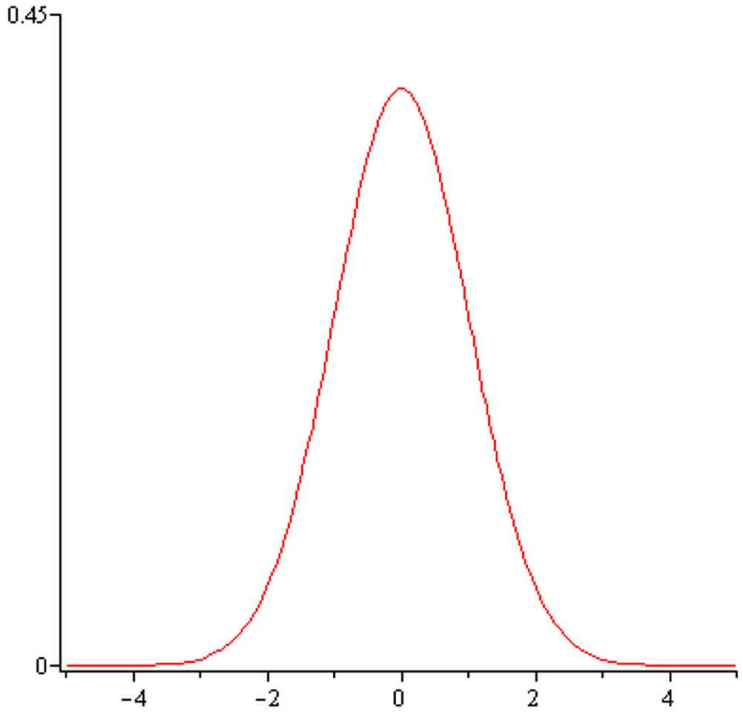
As an interesting footnote, it is speculated by some very bright econometrists that many of our basic assumptions of normality in financial risk are not correct. We love to assume normality because we know so much about the normal curve. But it appears as though many types of risk are actually distributed as some very nasty distributions, e.g., Cauchy and Levy distributions. They can look just like normals to the untrained eye, but they make for very unstable and unreliable estimates.

You do not have to read about the finite population correction factor at the end of 6.3. And we will not cover 6.4.

Do the following exercises:
6.1: 1-5 odd. 7-25 odd (Do 3 of them.) 27-40 odd. (Do 3 of them.) 41-49 odd.
6.2: 3, 5, 9, 11, any of the odds from 13 to 29 are pretty good. Definitely do 31 and 33.
6.3: 1-5 odd, 7, 11, 13, 23.

Chapter 6 concerns the normal distribution, probably the most famous continuous probability distributions. We've already discussed the normal distribution once when working with the Empirical Rule. I'm sure this is not the first time many of you have heard of this particular distribution; it is taught in high school math almost everywhere now.

Here's the distribution function of the "standard normal" distribution. A standard normal has a mean of zero, and a standard deviation of 1.



Some facts about this curve.

- It is a continuous random variable. As with all continuous random variables:
  - Area under the curve gives you probability. If you know X is a continuous random variable, then the probability that X is between say, -2 and 3, is the area under the curve between -2 and 3.
  - The total area under the curve is 100% or 1.0.
- The mean, mode and median are all at zero.
- The curve is perfectly symmetrical.
- The empirical rule is based on this distribution. Recall
  - 68% of the area is between -1 and 1.
  - 95% of the area is between -2 and 2.
  - 99.7% of the area is between -3 and 3.
- Although most of the area is concentrated close to zero, it stretches forever in both directions.

As a function, the general normal distribution with mean μ and standard deviation σ is:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ You will not need this fact, but it looks cool. :-)

Why is the normal distribution is so important? Well, it just appears all over! There's good reason for it too. Basically, if you add up or take an average of "reasonable" independent random variables, the sum (or average) starts to look like a normal distribution. Adding more gets you closer. Let me show you.

Example 6.0.0

Suppose you have a fair die. Let X be the value of the roll of the die. So X is a discrete random variable. We're going to be generating many trials (a trial is a roll) of this random variable, so we need to construct some notation. If you roll the die 30 times, you can label each roll of the die like this:

X_1 = whatever you roll the first time
X_2 = whatever you roll the second time
X_3 = whatever you roll the third time
<keep going>. . .
X_{30} = whatever you roll the 30th time

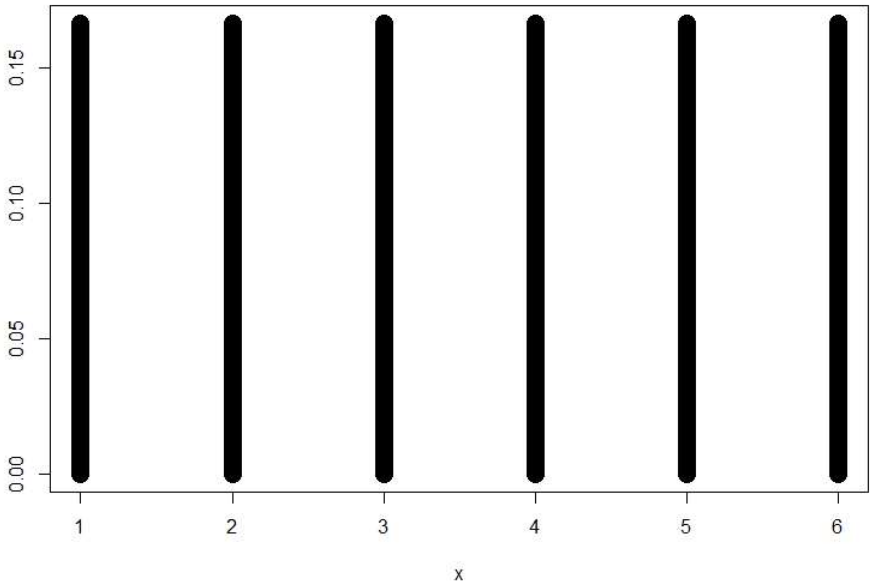
The rolls of a fair die are independent of each other, and each roll of the die is a random variable just like all the rest of the rolls. So we give these "trials" a statistical name. We say the random variables X_1 to X_{30} are *independent and identically distributed*. Think about this: the process of rolling a die does not change from roll to roll and each roll is independent of every other, i.e., the rolls are 100 independent identical experiments. So, it should make intuitive sense that they have identical distributions. We usually just abbreviating this to "iid."

What if we take the sum of all of these values? (We will be doing this a lot for the rest of this class.) Realize that the average is again a random variable. Let's give this random variable a name.

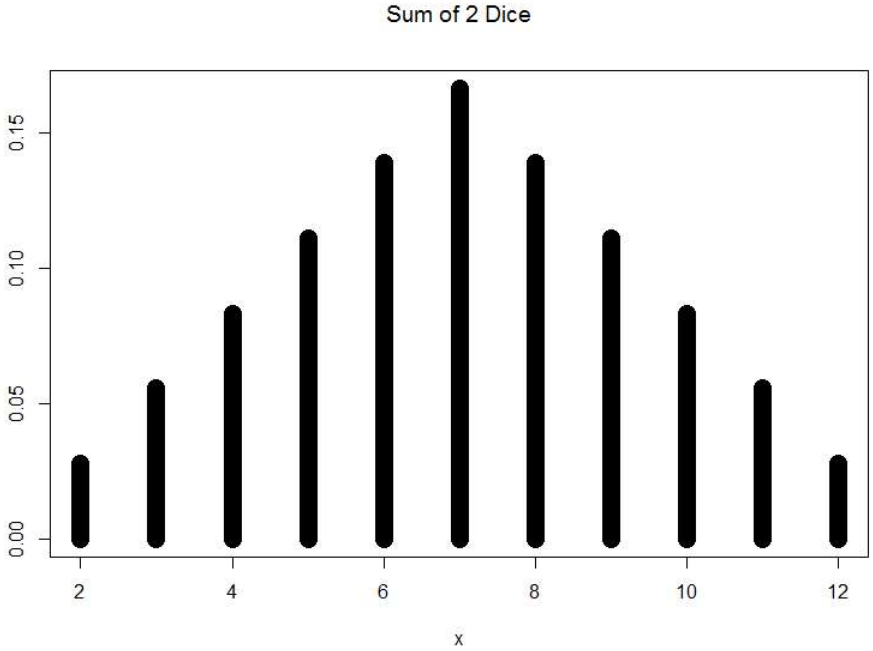
Let Y = the sum of the random variables X_1 up to and including X_{30}

Then Y is a random variable. That is, we don't know what it is equal to unless we know the values of X_1 to X_{30}.

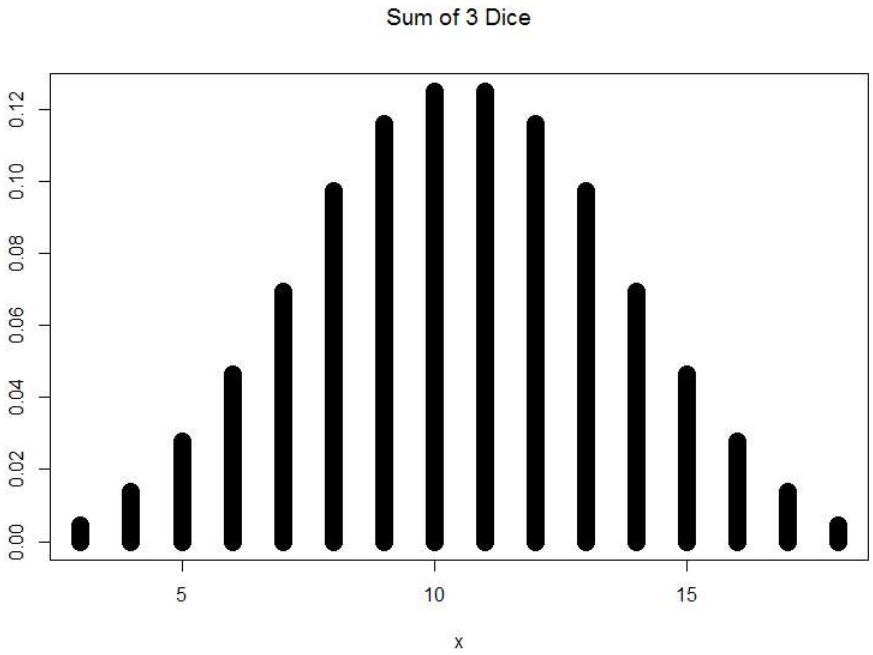
So what do sums of dice look like, distribution-wise?



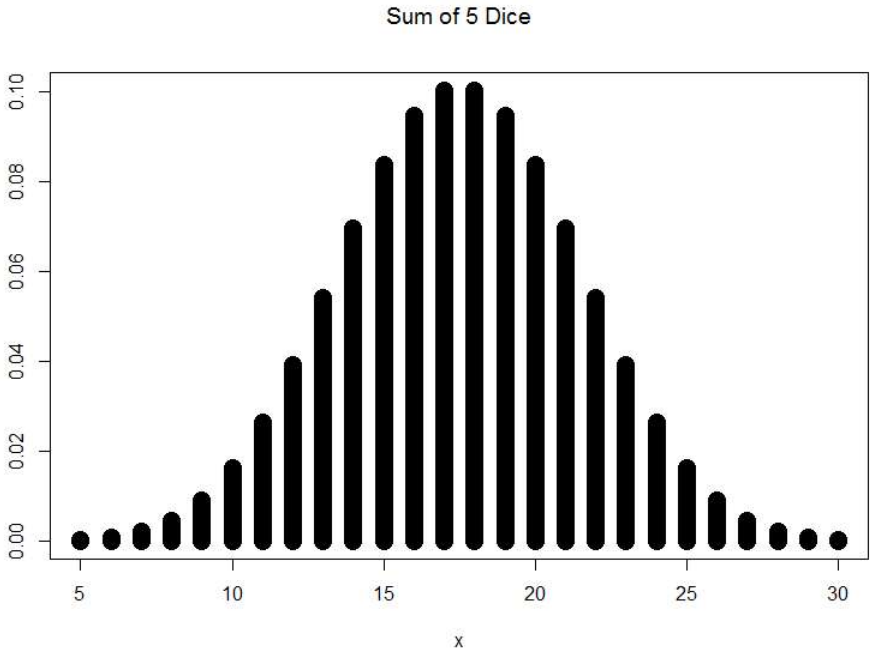
Here's one die. Possibilities are 1 up to 6, each with probability of about 0.167. (So heights are, as usual, the probability of that outcome.)



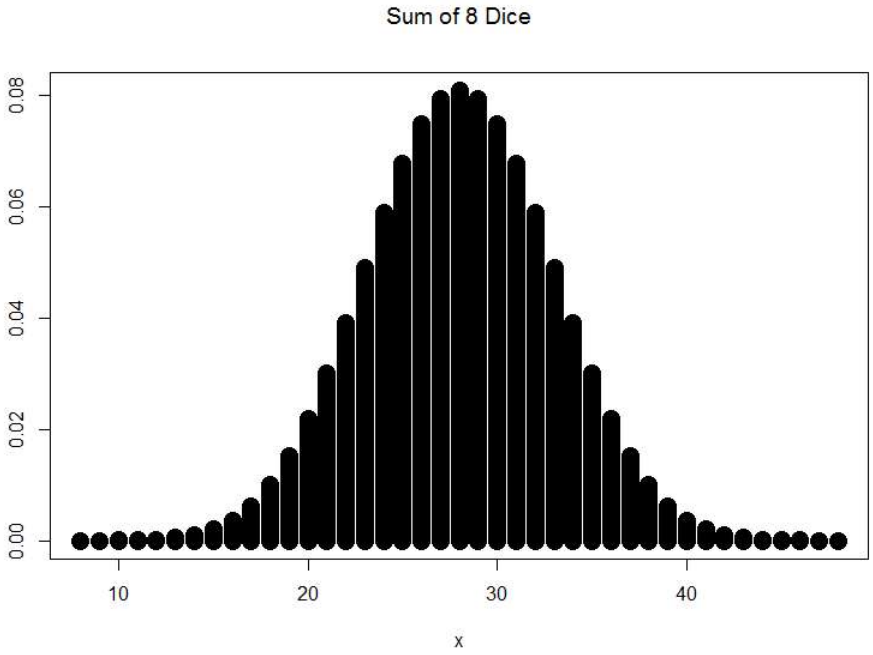
Two dice. Notice that 7 is most likely, 2 and 12 least so. Oh, and it's not flat anymore.



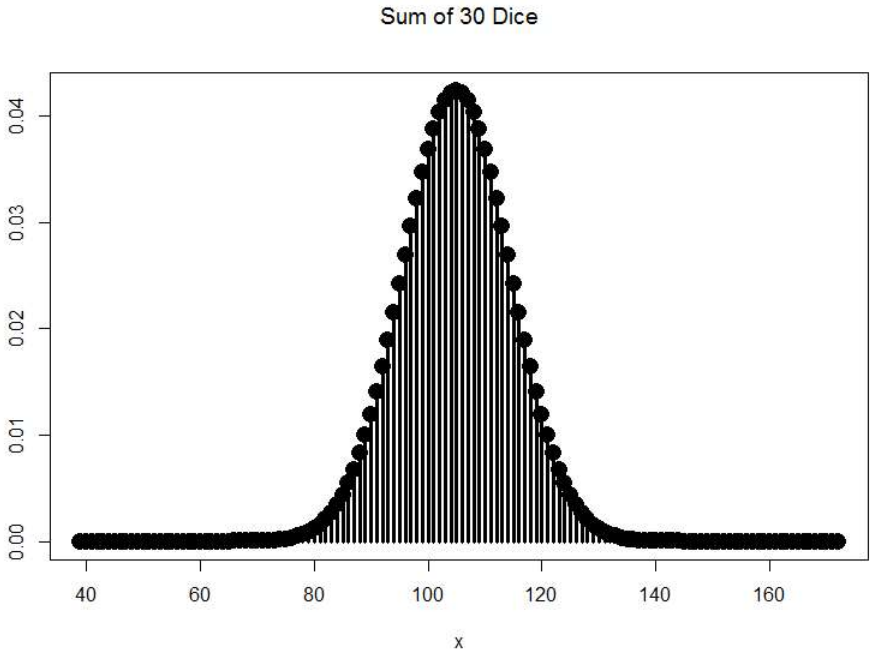
Three dice. Smallest possible roll is 3, largest is 18. The expected value? That would be 3.5 * 3 = 10.5. Note that the histogram is getting curvier.



The sum of 5 dice, $Y = X_1 + X_2 + X_3 + X_4 + X_5$ is really looking normal-shaped.

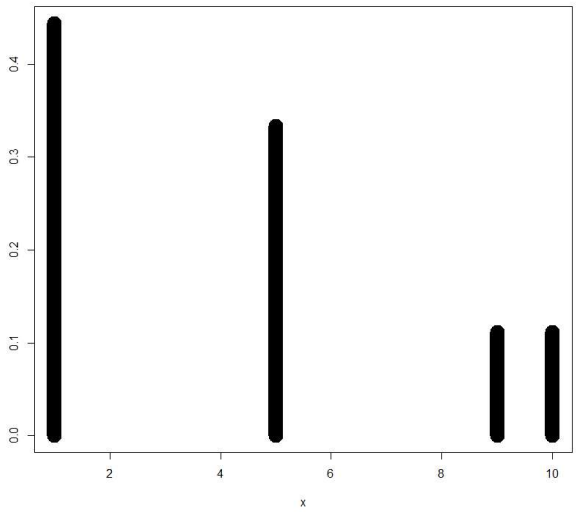


8 Dice. From here on out, the shape is really normal-shaped, but the tails are getting longer.

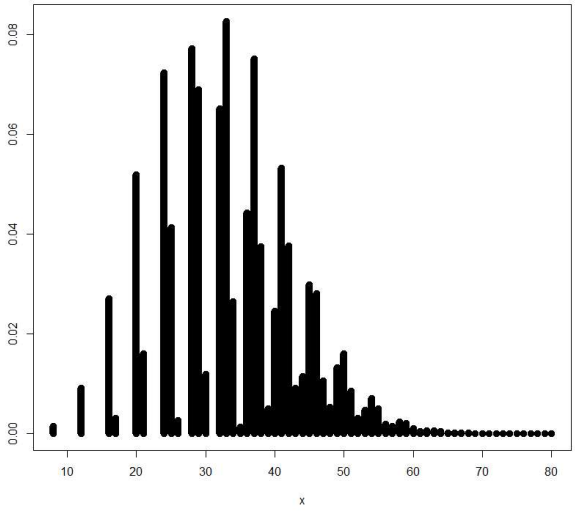


And all 30 dice.

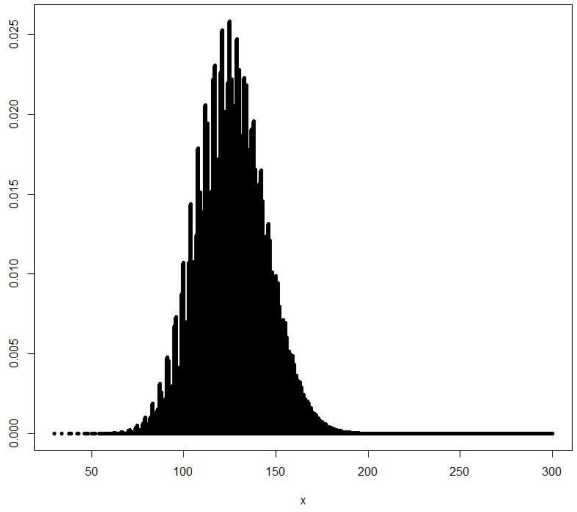
This works for many, many other random variables as well. Pretend you could make a die with any number of faces. Now write any numbers on the faces that you want. Feel free to unbalance the die so that the faces could have any probability of showing. The same thing will happen. So here's a 9-sided die with four 1s, three 5s, one 9, and one 10. With all sides equally likely. (You could think of it as a die with 4 sides where the probability of rolling a 1 is 4/9, rolling a 5 is 1/3, and 1/9 chance for each of the 9 and 10 sides.)



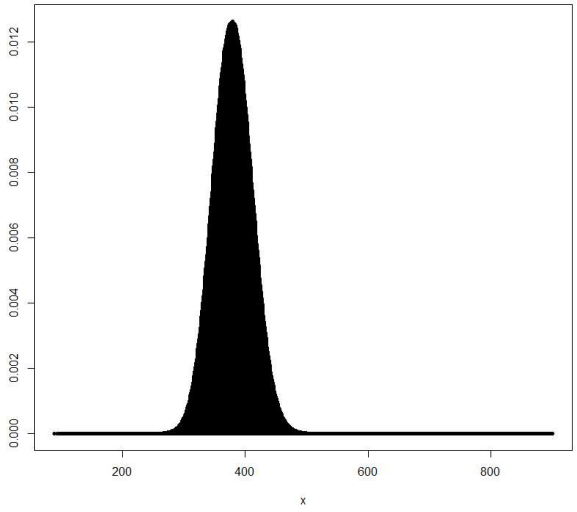
Here's one roll



Eight rolls



30 rolls –the punk normal distribution?



90 rolls

Sometimes it takes more for the sum (or average) to really start looking like the normal distribution.

The Central Limit Theorem

Let X be a random variable with expected value μ and standard deviation σ. Let X_1 to X_n be independent and identically distributed trials of X. Let Y be the average of them. Then Y is approximately a normal random variable for a sufficiently large enough n. The bigger n is, the closer to being normal Y will be. The expected value of Y is μ and standard deviation of Y is $\frac{\sigma}{\sqrt{n}}$.

One could make an argument that this is the single greatest theorem in all of mathematics. Basically, it says that, for most distributions, the distribution of the average of repeated trials approaches a normal distribution. And a bigger sample will make it even closer. This is good because we know a lot about normal distributions, as you will see in the following chapters.

Not every random variable actually has a mean and/or standard deviation. You can think of them as trying to have an infinite standard deviation. The central limit theorem does not work on them. They can and do occur—like when one divides by a random variable that can be close to zero. And newer research shows that they occur in risk assessment and financial markets as well. Categorically, a random variable that doesn't have a standard deviation will always have an infinite sample space and a considerable amount of probability of being arbitrarily large. They are called "heavy-tailed" or "fat-tailed" distributions.

In many situations you can throw out the possibility of one of these monsters. If your random variable is bound below by a number and bound above by a number, it will obey the CLT. So every die will always work. Dice always have a finite number of sides, and that means there will always be a biggest and smallest number on the die.

Finding probabilities of a standard normal in a given range.

The book does a pretty good job of this. You'll need to use the table (Table E). Directions and examples are on pages 316-322. I have created a standard normal calculator (for two or three regions) but you'll still need to be able to use the table in several situations.