Basic concepts of Prob. and Stats.

NRES 710

Fall 2020

Download the R code for this lecture!

To follow along with the R-based lessons and demos, right (or command) click on this link and save the script to your working directory

Overview of basic concepts of probability and statistics

Now that we've explored data types, we are going to talk about interpreting data! Specifically, through the lens of (frequentist) probability and (classical) statistical tests!

Probability can be tricky. Humans are not very intuitive about probability!

Yet it's not just laypeople who are confused by concepts of probability and uncertainty. Statisticians themselves argue about probabilities. For example, the hot hand in basketball.

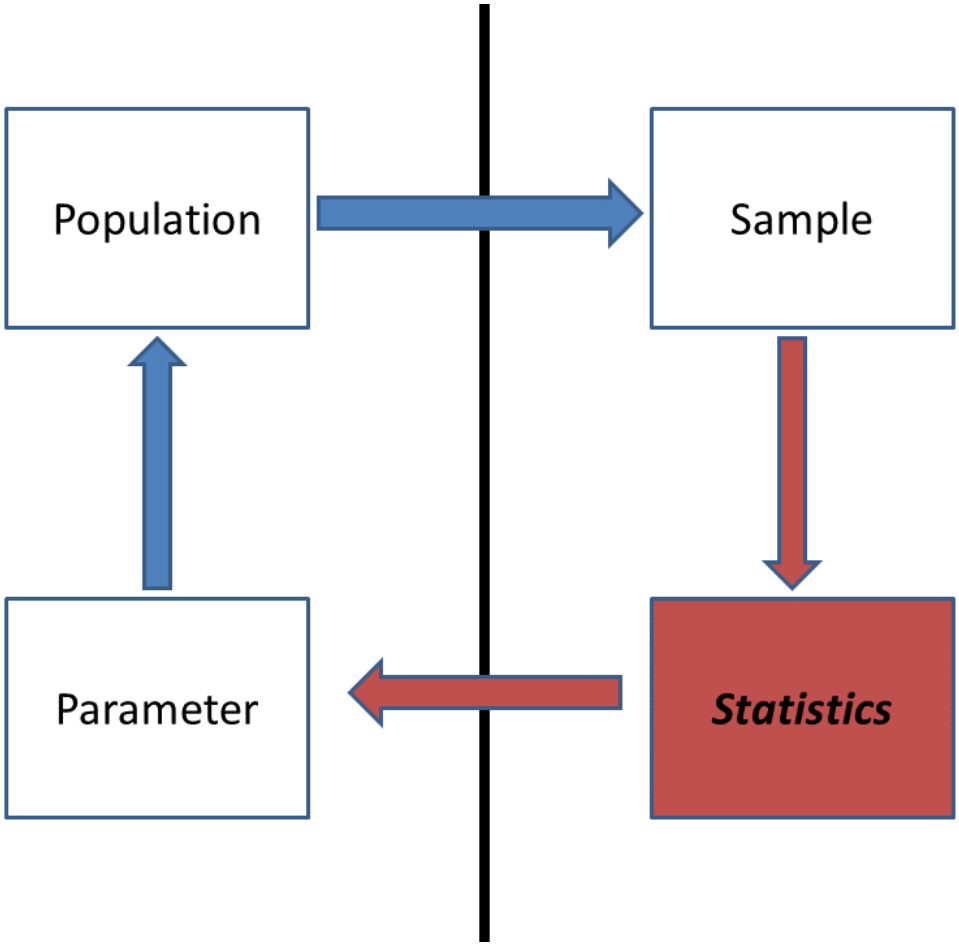
Why does this matter to us? Because we have to use probability to make inferences from our data (which is what statistics is all about)!

Classical Statistics Part 1

For example, say I asked you... Are people with last names beginning A-M taller than those with last names beginning N-Z?

As a scientist, what would we do?

We would want to identify, and then sample, the population!



Q: what is the population here?
Q what is the sample in the above example?
Q what is the parameter in the above example?
Q what is the 'statistic' in the above example?

Population: ???
Parameter: ???
Sample: ???
(summary) Statistic: ???

When we go from the population to the sample, we implicitly make several assumptions about the sample (i.e., that the sample is representative of the population).

When we make *inference* about a population-level property (the *parameter*) from a sample, we use *statistics*.

Classical statistics involves computing a *summary statistic* from the data, and then using a 'sampling distribution' to determine whether or not that summary statistic is meaningful or likely a result of random noise.

Frequentist Probability Part 1

Where does probability come in? Well, samples are random- and by random we mean that every time we sample from a population can result in a different outcome. Therefore, all *statistics* derived from samples are random quantities, and will differ with each sample. When something is random, we use probability to quantify how commonplace or unexpected an outcome is. Probability helps us to quantify things that we can't know with certainty.

If all samples (and the statistics derived from them) differ from one another, how can we be sure that our statistic (derived from a random sample) is actually telling us something about the population of interest? How can we be convinced that our result is meaningful and not simply an *artifact of random sampling*?

Aside: probability measures lack of surprise!

Q: what is the probability of getting three heads out of 4 coin flips?

We assume that heads or tails are equally likely outcomes of a single coin flip and that the four flips are completely independent.

How many possible outcomes are there?

TTTT
TTTH
TTHT
TTHH
THTT
THTH
THHT
THHH
HTTT
HTTH
HTHT
HTHH
HHTT
HHTH
HHHT
HHHH

Are each of these outcomes equally likely?

How many of the possible outcomes meet our criterion (all heads)? (How many ways are there of getting 4 heads?)

What is the probability of getting all heads?

Probability of a specific outcome = (# ways of getting a specific outcome)/(total number of equally plausible outcomes)

Probability measures your lack of surprise at observing a particular outcome from a sampling process. If there is only one way of getting a particular outcome out of hundreds of different, equally possible sampled outcomes, you might be pretty surprised to see that one specific outcome!

Q What is the probability of getting at least three of the same face (heads or tails) out of 4 coin flips?

How surprised would you be to flip 4 coins and observe at least three flips with the same face??

Classical Statistics Part 2 (the null!)

The null hypothesis is simply the notion that the effect you are testing for is absent or inconsequential, or that the question you are asking is testable yet totally false.

The null universe is the universe where the null hypothesis is true. It is this universe that we must dwell in when we practice statistics, at least in classical/frequentist statistics! Therefore in this class we will tend to dwell in the null universe a lot!

Q what is the null hypothesis for the height example above?

The *p value* is a key concept of frequentist statistics, but one that is often misinterpreted and misunderstood.

The p value is simply the probability of obtaining a result at least as extreme as your observed result *in the null universe*. The p value answers the question: "what's the probability that my result (or one even more extreme) could have happened by chance alone?"

Aside: extremeness

When we use the phrase 'at least as extreme as your observed result' we mean a result that would lend even more support for rejecting the null hypothesis than your observed result (a result with even more 'signal' than the observed data). For example, let's imagine the following scenario for the height/last name question:

Sample size: 20, 10 with names between A and M and 10 with names between N and Z.
Observed summary statistic: mean height for A-M is 2 cm greater than mean height for N-Z
Q What are some results that would lend at least as much support for supporting the notion that the null hypothesis is false??
Q Would you be comfortable rejecting the null hypothesis if random sampling in the null universe yielded mean height differences of 2 cm or greater 20% of the time (p = 0.2)?

Rejecting the null

We reject the null hypothesis if the chance of the result occurring by random chance is sufficiently small. For historical reasons, 0.05 is used as the threshold (known as alpha)- if a p value is less than 0.05, we reject the null!

Fisher proposed alpha=0.05 as a nice balance between the probability of mistakenly rejecting a true null hypothesis (Type 1 error), and the probability of failing to reject a false null hypothesis (a Type 2 error). This cutoff value was NEVER intended to be a fixed value to be applied unthinkingly though!!!

Example: Fisher's cups of tea:

A Woman claimed she could tell if milk was added to a cup of tea first or last.

Fisher suggested we give her 8 cups of tea at once, 4 with milk first and 4 with tea first. The woman is asked to identify the four cups that were poured with milk first.

There are 70 possible ways of choosing four cups out of 8. The probability of getting all of them right by random chance (in the null universe) would be 1/70 = 0.014 = 1.4%. That would be kinda surprising, right (in the null universe that is)? So if the lady selected all four cups correctly, are we surprised enough to reject the null?

Here are the possibilities (X=correct id, O=false id), which in the null universe are equally likely (since the lady can't actually tell the difference!). Let's assume cups 5-8 are milk first and 1-4 are milk last. The number of correct choices is in parentheses...

O1/O2/O3/O4 (0) O1/O3/O4/X5 (1) O1/O4/X7/X8 (2) O2/O4/X5/X6 (2) O3/O4/X7/X8 (2)
O1/O2/O3/X5 (1) O1/O3/O4/X6 (1) O1/X5/X6/X7 (3) O2/O4/X5/X7 (2) O3/X5/X6/X7 (3)
O1/O2/O3/X6 (1) O1/O3/O4/X7 (1) O1/X5/X6/X8 (3) O2/O4/X5/X8 (2) O3/X5/X6/X8 (3)
O1/O2/O3/X7 (1) O1/O3/O4/X8 (1) O1/X5/X7/X8 (3) O2/O4/X6/X7 (2) O3/X5/X7/X8 (3)
O1/O2/O3/X8 (1) O1/O3/X5/X6 (2) O1/X6/X7/X8 (3) O2/O4/X6/X8 (2) O3/X6/X7/X8 (3)
O1/O2/O4/X5 (1) O1/O3/X5/X7 (2) O2/O3/O4/X5 (1) O2/O4/X7/X8 (2) O4/X5/X6/X7 (3)
O1/O2/O4/X6 (1) O1/O3/X5/X8 (2) O2/O3/O4/X6 (1) O2/X5/X6/X7 (3) O4/X5/X6/X8 (3)
O1/O2/O4/X7 (1) O1/O3/X6/X7 (2) O2/O3/O4/X7 (1) O2/X5/X6/X8 (3) O4/X5/X7/X8 (3)
O1/O2/O4/X8 (1) O1/O3/X6/X8 (2) O2/O3/O4/X8 (1) O2/X5/X7/X8 (3) O4/X6/X7/X8 (3)
O1/O2/X5/X6 (2) O1/O3/X7/X8 (2) O2/O3/X5/X6 (2) O2/X6/X7/X8 (3) *X5/X6/X7/X8 (4)*
O1/O2/X5/X7 (2) O1/O4/X5/X6 (2) O2/O3/X5/X7 (2) O3/O4/X5/X6 (2)

O1/O2/X5/X8 (2) O1/O4/X5/X7 (2) O2/O3/X5/X8 (2) O3/O4/X5/X7 (2)
O1/O2/X6/X7 (2) O1/O4/X5/X8 (2) O2/O3/X6/X7 (2) O3/O4/X5/X8 (2)
O1/O2/X6/X8 (2) O1/O4/X6/X7 (2) O2/O3/X6/X8 (2) O3/O4/X6/X7 (2)
O1/O2/X7/X8 (2) O1/O4/X6/X8 (2) O2/O3/X7/X8 (2) O3/O4/X6/X8 (2)

Q what if the lady selected 3 of 4 correctly? [there are 17 ways of getting three of the four correct out of 70 total permutations]. Would you be surprised enough to reject the null hypothesis. Is it possible that you and the lady are living in the null universe? How surprised would you have to be to reject the null universe and admit that she can tell the difference??

Statistics Terminology: Type 1 and Type 2 errors

A type 1 error is mistakenly rejecting a true null hypothesis.

A *p value* is the probability that you collect a sample in the null universe with greater evidence against the null (signal) than your observed data.

Alpha is the threshold p value above which we are insufficiently convinced that we can reject the null. We usually set the alpha level prior to performing an experiment or collecting data so we ensure a low probability of committing a type 1 error.

You can reject the null if $p <$ alpha. Conversely you will fail to reject the null hypothesis if p > alpha (you are insufficiently convinced that you are NOT living in the null universe!)

A type 2 error is mistakenly failing to reject a false null hypothesis (you are not living in the null universe but you fail to realize it!)

Beta is the probability of committing a Type 2 error.

Q what would it mean to set the Beta level? Why do we set alpha rather than beta?

Q how many universes are there where the null hypothesis is true? How about where the null hypothesis is false?

Finally *Power* is defined as the probability of correctly rejecting an incorrect null hypothesis. Power is simply 1-Beta!

		Null H is	
		True	False
Decision about H	Fail to reject	OK!	Type II error (beta)
	Reject	Type I error (alpha)	OK!

Quick Review of Basic Concepts

A p value is the probability of obtaining a result at least as extreme (containing at least as much 'signal') as the observed result, given you are dwelling in the null universe.

Usually the 'result' we are referring to is a single 'summary statistic' calculated from our sample – for example, the sample mean, the standardized difference between two sample means (as in a 2-sample t test), some other signal to noise ratio, or the proportion of a sample that meets some criterion (e.g., proportion red M+Ms).

This implies a multitude of potential results. Your observed result is just one of an infinite number of possible results from some sampling/measurement process. This variation across possible samples/results is called 'sampling variation'

Imagine drawing 10 M&Ms from some much larger population of M&Ms in a big jar and counting the proportion of red ones. There are practically countless possible samples we could draw (each one potentially associated with a different proportion of reds), even though there is only one true proportion of red M+Ms in the big jar.

the truth is fixed but unknown and we can only approach the truth indirectly and imperfectly via sampling

We often (if performing a test) start off by pretending we live in the null universe, and we start collecting imaginary samples (LOTS of them) in the null universe. We then determine the probability that these imaginary results are at least as extreme as our observed results. This is our p value. We use our p value, along with our pre-determined type I error rate (level of surprise we need to reject the null universe).

Even more generally, the goal of statistics (even when not performing a test) is to infer something about a population from a sample. In the yellow-legged frog example, there is a "true" mean body size of frogs in ponds in the central Sierra Nevada. We just don't know what it is! After collecting the data, statistics might help use to say something about the mean body size in the population – both about what we know AND what we don't know! In fact, all of statistics is about dealing with uncertainty. We don't need statistics if we have complete certainty!

Yellow-Legged frog example:
Population: all yellow-legged frogs in ponds in the central Sierra Nevada
Parameter: mean body size (SVL) of adult yellow-legged frogs in all ponds in the central Sierra Nevada
Sample: 10 ponds- as many frogs are captured and measured as possible
Statistic: Sample mean

However, we assume that our sample mean is representative of the population mean. How? Why? Because of the *Central Limit Theorem*.

The Central Limit Theorem

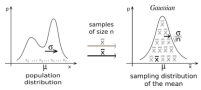
The Central Limit Theorem (CLT) says that if you have a sample with a reasonably large number of observations, and each observation is randomly sampled, then the mean of those values will be similar to the actual population mean: the mean of ALL yellow-legged frogs in ponds in the central Sierra Nevada. And as the sample size gets bigger, the sample mean will become more representative of the true mean (it will converge on the true mean as sample size approaches infinity).

This is useful, but that's not all.

The CLT is the magic wand of statistics. It does enormous amounts of work for us. Why?

The CLT also implies that the distribution of sample means collected from repeated sampling is approximately normally distributed even if the underlying data themselves are not normally distributed.

Did you ever wonder why the normal distribution is so common in statistics? It's because of the CLT- many summary statistics derived from a sample are expected to have a sampling distribution that is approximately normally distributed!



Example

How does this work? Let's use the yellow-legged frog example.

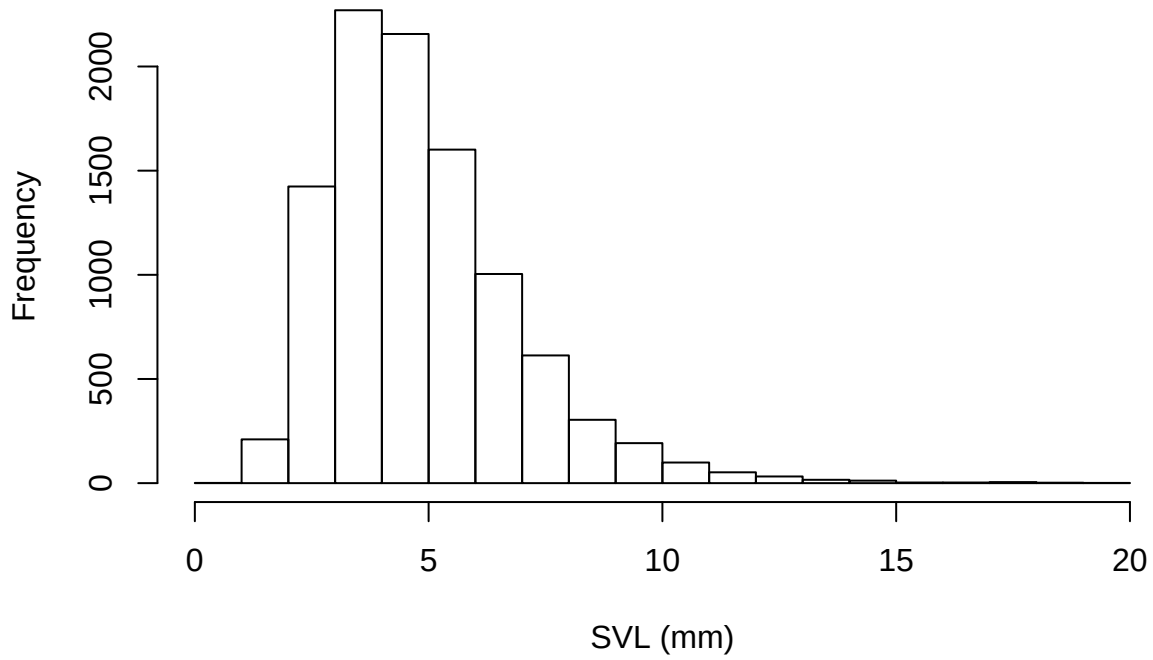
Let's say that we could measure ALL the frogs in ALL the ponds in CA. What would that look like?

Let's simulate it using a lognormal distribution that is strongly right skewed (positively skewed), suggesting that there are a lot of frogs out there that are relatively small-bodied, and a few that are giants! NOTE: this is not biologically realistic, but it makes a point.

First, let's set the population/parameter (the truth about which we hope to make inference but can never know in reality)

Q if we could measure the entire population of interest, do we even need statistics???

```
#### ALL FROGS IN CA

allfrogs.bodysize <- rlnorm(10000,1.5,0.4)        # statistical 'population'
hist(allfrogs.bodysize,main="",xlab="SVL (mm)")   # plot out histogram
```




```
truemean_SVL <- mean(allfrogs.bodysize)           # the 'parameter'
truemean_SVL
```

```
## [1] 4.85296
```

Now let's take a sample!

```
mysample <- sample(allfrogs.bodysize,10)    # take sample of size 10 (10 frogs measured)
mean(mysample)   # compute the sample mean
```

```
## [1] 5.486005
```

And another, this time with n=20

```
mysample <- sample(allfrogs.bodysize,20)    # take sample of size 20 (20 frogs measured)
mean(mysample)   # compute the sample mean
```

```
## [1] 4.93051
```

Since sampling is random, sampling will produce a different result every time.

To get a better picture of the sampling variance, lets' sample many times!

```
lotsofsamples <- list()

for(s in 1:5000){
  lotsofsamples[[paste0("sample",s)]] <- sample(allfrogs.bodysize,30)    # take sample of size 30 (20 f
}

lotsofsamples$sample1
```

```
##  [1]  2.372883  5.597451  2.730395  3.524137  8.256081  5.016583  4.903979
##  [8]  2.072417  3.460012  2.583852  6.436480  8.861281  5.271463  3.700272
## [15]  2.015766  4.416131  6.035820  6.229058  4.280973 10.891733  4.524146
## [22]  5.732792  5.871246  2.808683  3.380603  2.339923  2.971632  3.051538
## [29]  1.853315  5.579736
```

```
lotsofsamples$sample99
```

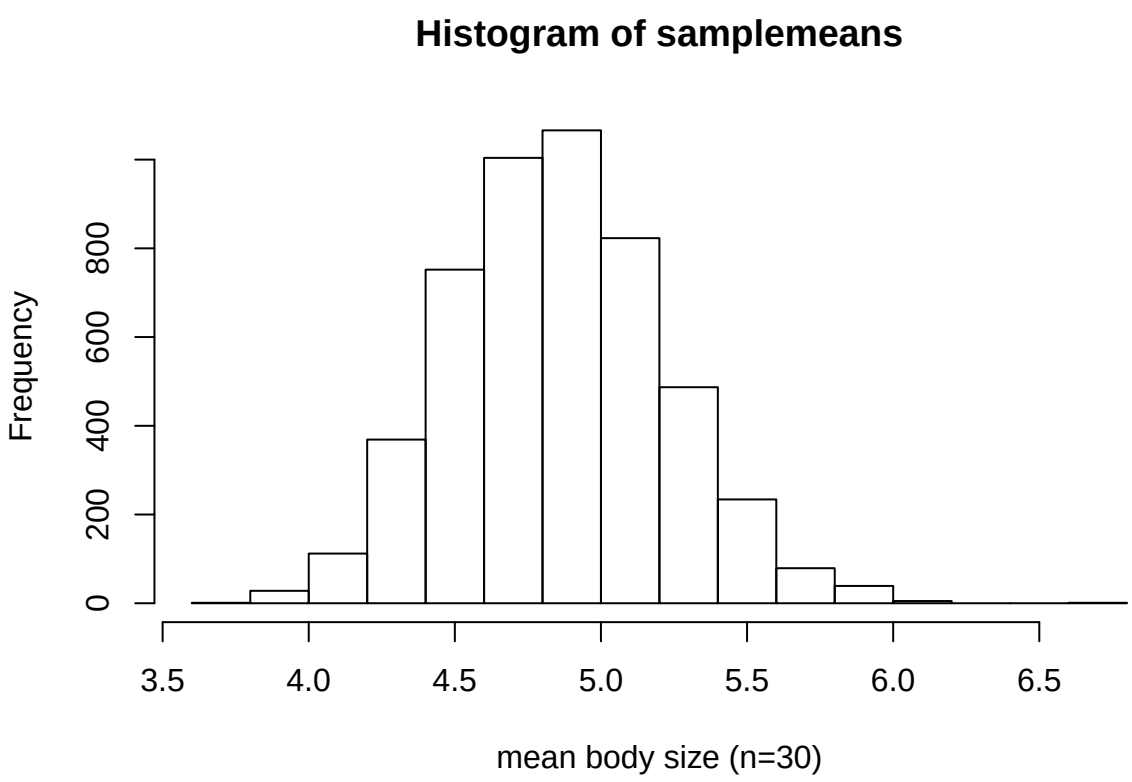
```
##  [1]  3.082741  4.643865  9.087745  2.861233  4.839637  4.620912 10.227511
##  [8]  4.198693  5.007569  2.233877  3.864987  3.217923  5.824509  6.938096
## [15]  4.286399  5.921866 12.639477  5.504835  3.662911  7.404917  3.118482
## [22]  3.968394  2.443898  4.773560  4.047094  5.746137  4.042524  1.662506
## [29]  2.795141  2.645004
```

```
lotsofsamples$sample732
```

```
##  [1] 1.694850 4.451887 2.216197 4.429486 3.705984 2.636826 6.193785
##  [8] 4.588675 2.793727 3.353492 5.526981 3.647781 9.039995 3.498335
## [15] 6.087886 7.626371 3.284825 5.495483 2.198753 4.957695 4.633282
## [22] 2.006623 2.726117 6.983993 3.243971 5.724081 3.481902 3.874363
## [29] 2.491129 3.851397
```

Now we can compute the sample means and the sampling variance for the summary statistic (mean body size)

```
samplemeans <- sapply(lotsofsamples,mean)

hist(samplemeans,xlab="mean body size (n=30)")
```



Interesting- does this look skewed to you? Doesn't it look like a normal distribution??

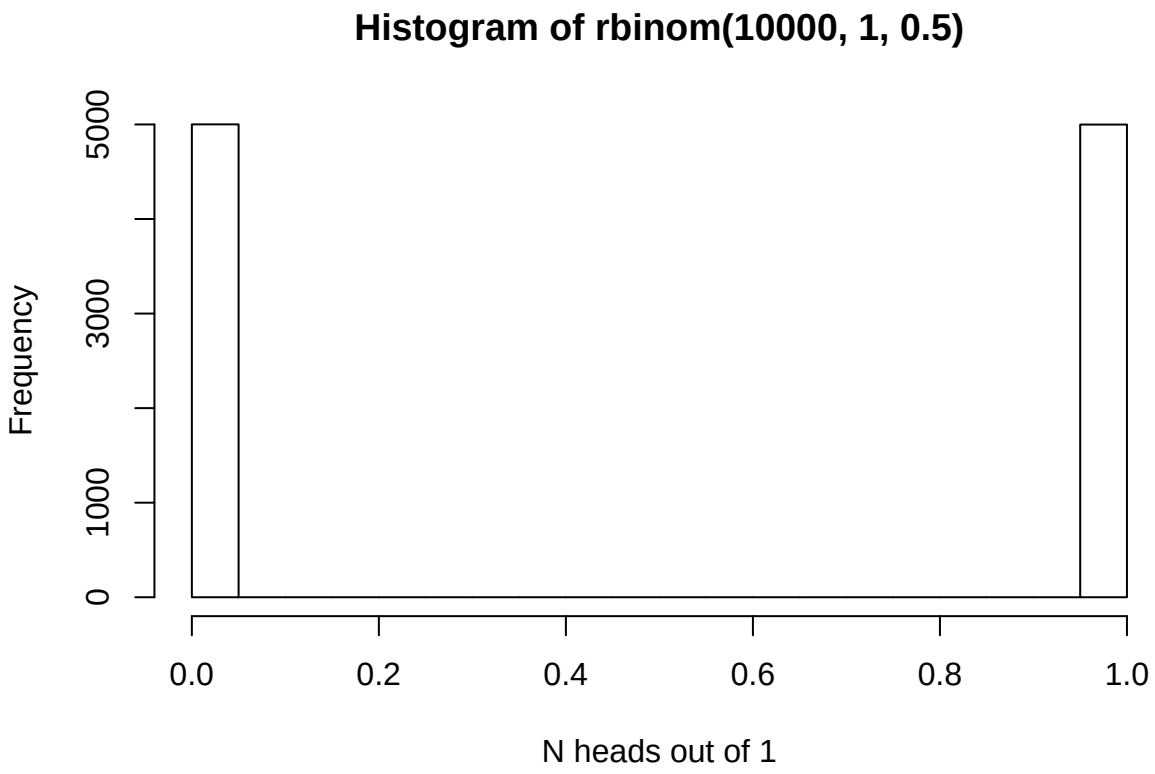
It's the CLT at work!!

One way to think about this is just that there are more ways of getting a sample mean near the real mean than to get one far away from the sample mean. It doesn't matter what the raw distribution is. Of all the samples you could get, there are very few that are all at one end of the distribution. There are a lot more random samples that span the full distribution of values, from low to high. Take the average af all those values, low and high, and you get something in the middle. The normal distribution is humped right in the middle, becuase of the tendency for low and high observations to 'average out' within a sample.

This works with coin flips too!

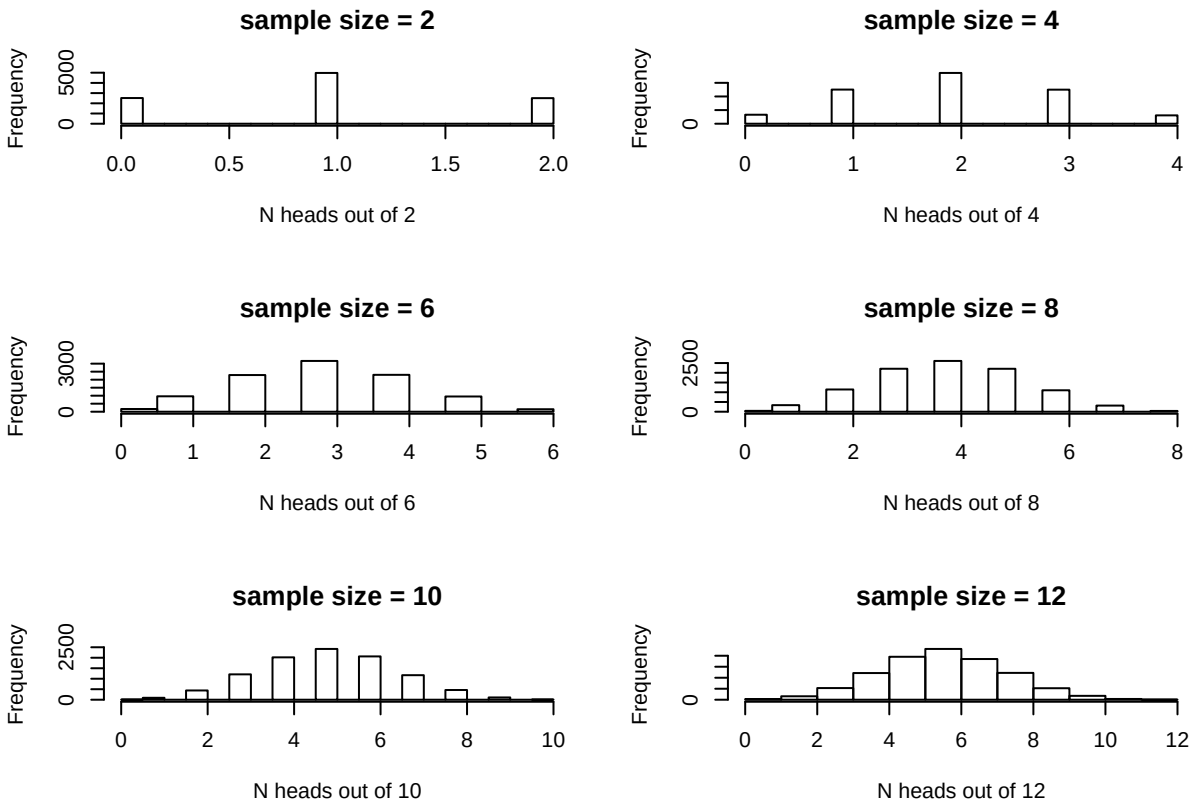
Here's the sampling distribution for the number of heads out of a single coin flip:

```
hist(rbinom(10000,1,.5),xlab="N heads out of 1")
```



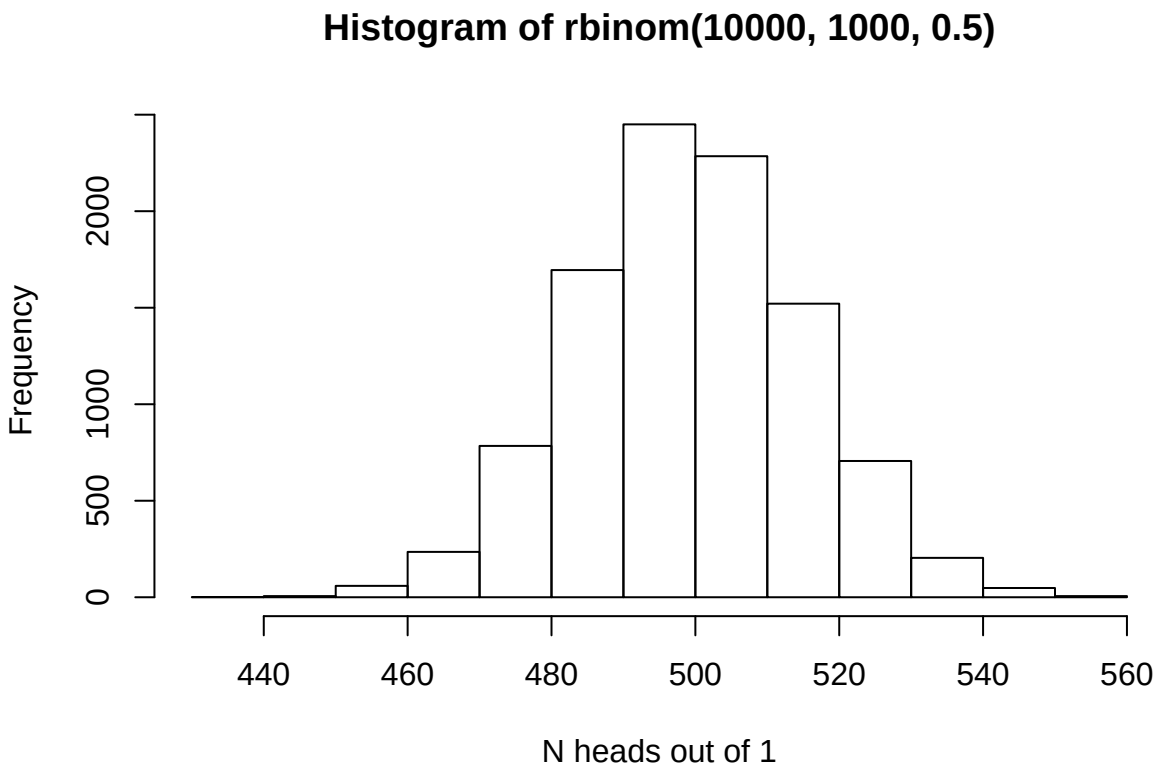
Now let's build up sample size and see how the sampling distribution changes.

```
par(mfrow=c(3,2))
for(i in seq(2,12,2)){
   hist(rbinom(10000,i,.5),main=paste0("sample size = ",i),xlab=sprintf("N heads out of %s",i))
}
```



And with really big sample size:

```
hist(rbinom(10000,1000,.5),xlab="N heads out of 1")
```



Replication

Let's explore replication, and what it truly means.

To start, read the following two popular-press articles, which have to do with scientific replication:

The New Yorker: "The Truth Wears Off". Why does the 'significance' of studies tend to be reduced upon replication? What is scientific truth anyway- is truth itself wearing off??

The New York Times: "When the revolution came for Amy Cuddy". This was a statistically 'valid' study – and yet when truly "replicated", power-posing was shown to be largely ineffective. As ecologists, we are all busy trying to adequately replicate our science. But what does that really mean?

And, while you're on a reading spree, read this highly influential monograph on 'Pseudoreplication' by Stuart Hurlburt.

And here's one more paper for good measure, by Davies and Gray (2015): this is a counter to the Hurlburt article.

Clearly one of the main take-aways is: science is messy!!!

Why does science require replication?

In general, the scientific project is to discover generalizable truths.

Q: How do we know if a result is true, and not a result of just random noise?
A: We use a large enough sample size so that we convince ourselves that random noise could not cause the result!

Q: How do we know if a result is general, and not a result of just some localized or specific phenomenon?
A: We draw our sample randomly from the entire population so that our sample is truly representative!

Find the pseudoreplication!

Population: All little brown bats across the USA
Parameter: Infection rate of white nose syndrome infection in cave hibernacula
Sample: Two cave hibernacula in New York state.
Statistic: Infection rate among sampled bats

Population: All humans
Parameter: Effectiveness of a coronavirus vaccine
Sample: 10,000 humans sampled in Sweden
Statistic: Infection rate in Uppsala (control) vs infection rate in Helsingborg (treatment)

Population: All humans
Parameter: Effectiveness of a coronavirus vaccine
Sample: 10,000 humans sampled in Sweden
Statistic: Infection rate in Uppsala (control) vs infection rate in Helsingborg (treatment)

Population: all yellow-legged frogs in ponds in the central Sierra Nevada
Parameter: mean body size (SVL) of adult yellow-legged frogs in all ponds in the central Sierra Nevada
Sample: 3,000 frogs sampled from a pond in the central Sierra Nevada
Statistic: Sample mean

The takeaway: you can only convince yourself of the generality of a result if the sample is representative of the population. In experimental design, one way to try to ensure generality is to sample randomly from the population of interest.

One part of the scientific endeavor is to poke holes in other people's research. Even more so, we poke holes in our own research. If we can convince ourselves of the truth of our result, only then can we feel comfortable sharing the result with the scientific community. That's not out of meanness or masochism, it's out of a search for truth and generality!

But the search for the truth is messy. In environmental science, we pseudoreplicate all the time - by necessity. For practical reasons our observations are not always completely independent from one another. Are we ever truly replicating perfectly and sampling sufficiently from our generalized target of inference? In many cases we are not. But that shouldn't stop us from trying to find truth- we just need to proceed with caution!

Summary statistics

We often summarize our samples by their centers (e.g., average) and possibly their spread (dispersion)

"Center" statistics: means, medians, geometric mean

Sample Mean (arithmetic mean) = sum of all sampled values divided by the sample size Sample Median (midway point) = 50% quantile. Order the values and select the value at the center.
Sample Geometric mean: product of numbers taken to the nth root. For two numbers, 3 and 4, you'd have the sq root of 3*4 = 3.46

Data spread, or dispersion:

Sample Standard Deviation – sigma for population standard deviation, s for the sample standard deviation Calculated differently for population than sample Sample Variance - Square of the standard deviation (different from sampling variance) Coefficient of variation

Std deviation calculation example

For a complete population: $\sigma = \sqrt{\sum_{n=1}^{i} \frac{(x_i - \mu)^2}{N}}$

For a sample (estimating population variance from a sample): $s = \sqrt{\sum_{n=1}^{i} \frac{(x_i - \bar{x})^2}{(N-1)}}$

For example: compute the variance of 5 numbers: 4, 3, 5, 5, 2

$\mu = (4 + 3 + 5 + 5 + 2)/5 = 3.8$

(4-3.8)^2 = 0.04 (3-3.8)^2 = 0.64 (5-3.8)^2 = 1.44 (5-3.8)^2 = 1.44 (2-2.8)^2 = 3.24

Sum these = 6.8 Divide by 5 = σ = 6.8/5 = 1.36

For sample sd (summary statistic for population variance from a sample):

$\bar{x}$ = (4+3+5+5+2)/5 = 3.8

4-3.8^2 = 0.2^2 = 0.04 3-3.8^2 = 0.64 5-3.8^2 = 1.44 5-3.8^2 = 1.44 2-2.8^2 = 3.24 Sum these = 6.8 Divide by 4 = s = 6.8/4 = 1.7

Aside: degrees of freedom

OK, so why the different estimate of dispersion for population vs. sample?

Which is larger? Which are we less confident in?

Has to do with a concept called *degrees of freedom*.

Sigma is known. Since you have the entire population measured, you can compute a measure of dispersion for the population and that measure is perfect. It is not an estimate, it is a perfect point value that is known without uncertainty. No bias, perfect precision.

The sample standard deviation s, on the other hand, is an imperfect estimate of dispersion for a much larger population!

Q would you expect the sample standard deviation using the first formula (pop stdev) to be biased? Why or why not??

NOTE: the formula for sample stdev *uses the sample mean*. Therefore, the sample data have already been used once to compute the standard deviation. If the true population mean was different than the sample mean, what would happen to the estimate of stdev if we replaced the sample mean with the true mean? Is it possible that the sample mean is not equal to the true mean?

NOTE ALSO: if you know that four of the five sampled values are 4, 3, 5, and 2 AND we know that the sample mean is 3.8, we know that the final sampled data point MUST be 5. There is no *freedom* there- the data point has to be 5. Therfore, even though we have 5 data points, we have only 4 degrees of freedom since the sample mean is included in the formula for the estimate for sample stdev.

By dividing the sum of squared deviations from the sample mean by 4 instead of 5, we are *unbiasing* the estimate of dispersion so that it is a better estimate of the population standard deviation.

Sampling variance

Review: we collect a random sample from a population. We want to know something about the population, but any summary statistic we compute from the sample (e.g., mean, median, variance, stdev, whatever!) will be an imperfect estimate of the true population parameter. So what do we do? How can we truly make inference about the population??

First, we need to know something about the sampling variance. We know that the summary statistic will be different for every sample we collect, but *how different, really???*

Statisticians have used probability calculus to work out the sampling distributions (often called sampling variance) for some common summary statistics. For example, the sample mean (super common summary statistics) has a sampling distribution that is centered on the sample mean (or zero, if living in the null universe) and follows a t-distribution (similar to the normal distribution) described by the standard error of the mean and degrees of freedom of N-1. Let's dissect this one!

Standard error of the mean

Standard error of the mean = sample std deviation divided by the square root of the sample size

$se = \frac{s}{\sqrt{N}}$

The standard error of the mean is used to help us describe the sampling distribution for the sample mean (the expected dispersion of sample means if you collected thousands of new samples and computed the mean).

```
#######
# Sampling distribution: the sample mean

mysample <- c(4.1,3.5,3.7,6.6,8.0,5.4,7.3,4.4)
mysample
```

```
## [1] 4.1 3.5 3.7 6.6 8.0 5.4 7.3 4.4
```

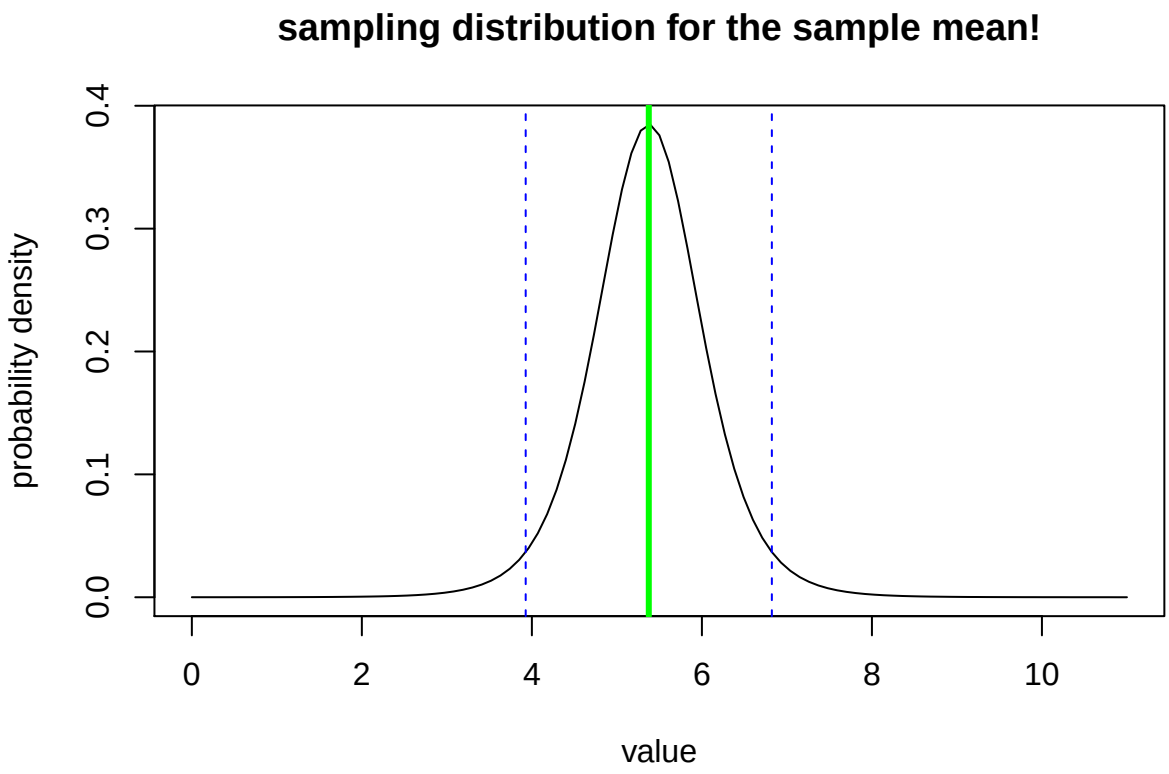
```
n <- length(mysample)    # sample size
sample.mean <- mean(mysample)  # sample mean
sample.stdev <- sd(mysample)   # sample standard deviation (r uses denominator of n-1 by default!)
std.error <- sample.stdev/sqrt(n)

std.error
```

```
## [1] 0.6122995
```

Now we have all the information we need to compute the sampling distribution for the sample mean.

Our sample mean is 5.375. But if we collected different samples of size n=8, we would get different values - even if the true population mean was 5.375. What does this distribution of values look like?

```
sampdist <- function(x){dt((x-sample.mean)/std.error,n-1)}
curve(sampdist,0,11,ylab="probability density",xlab="value",main="sampling distribution for the sample m
abline(v=sample.mean,col="green",lwd=3)
confint <- c(sample.mean+std.error*qt(0.025,n-1),sample.mean+std.error*qt(0.975,n-1))
abline(v=confint,col="blue",lty=2)
```



The vertical blue lines indicate the *confidence interval* around the mean with the *confidence level* set at 95%. We will talk about confidence intervals a lot more, but just know that about 95% of sample means should fall within the 95% confidence interval if the true population mean were equal to the sample mean (Note that we are assuming here that the sample mean is equal to the population mean). The confidence interval helps us visualize what might happen if we repeated our sampling over and over and over- how might our summary statistic change?

Note the use of the *t distribution* in the above code block. The t distribution represents the sampling variation that you would expect to observe for the sample mean if you collected many many samples of the same size as yours (see more details below).

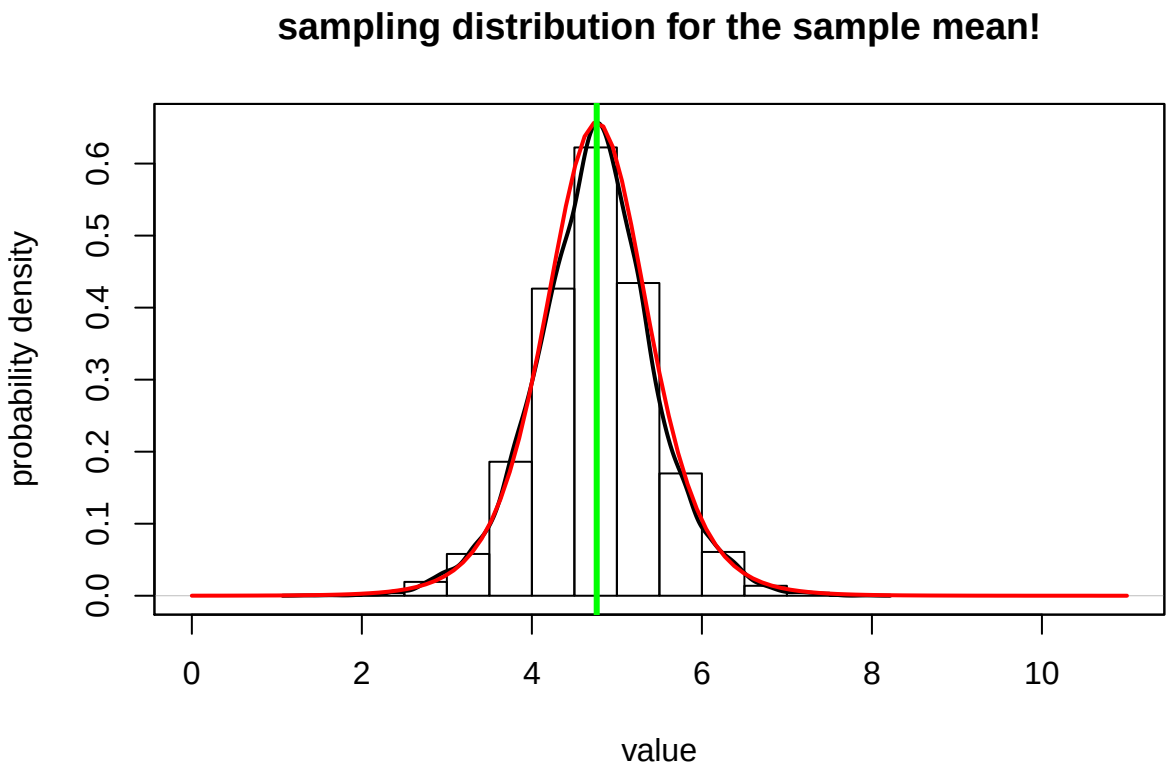
Just to cement this concept, let's compare the distribution above (based on a t distribution) with a brute-force simulation method!

```
#######
# Sampling distribution: the sample mean #2 (brute force simulation version)

mysample <- c(4.1,1.5,3.7,6.6,8.0,4.5,5.3,4.4)
mysample
```

```
## [1] 4.1 1.5 3.7 6.6 8.0 4.5 5.3 4.4
```

```
n <- length(mysample)    # sample size
sample.mean <- mean(mysample)  # sample mean
sample.stdev <- sd(mysample)   # sample standard deviation (r uses denominator of n-1 by default!)
```

```r
simulated.samples <- list()
for(s in 1:10000){
  sd1 <- sqrt(sum((sample(mysample,length(mysample)-1,replace = T)-sample.mean)^2)/(length(mysample)-2))
  simulated.samples[[paste0("sample ",s)]] <- rnorm(n,sample.mean,sd1)
}
sampling.distribution <- sapply(simulated.samples,mean)

plot(density(sampling.distribution),xlim=c(0,11),ylab="probability density",xlab="value",main="sampling
hist(sampling.distribution,add=T,freq=F)
par(new=T)
curve(sampdist,0,11,xlim=c(0,11),xaxt="n",yaxt="n",xlab="",ylab="",col="red",lwd=2)  # official samplin
abline(v=sample.mean,col="green",lwd=3)
```



Not a bad match right? Obviously it's easier and faster to use the t distribution (and more accurate) to approximate the sampling distribution, but I hope this helps to cement the concept!!

NOTE: the t distribution accounts for uncertainty in the sample variance, which is why I did not just use the sample variance in the code block above (if I had, you would see that the t distribution had 'heavier tails' than the brute force distribution)

The t statistic: signal to noise ratio for comparing 1 or 2 means

The *t statistic* (not the same thing as the t distribution) is just another summary statistic that we can calculate from a sample. In this way it is no different from the sample mean or the standard deviation. But, since it has a simple, well described sampling variance, it helps us perform rigorous hypothesis tests and is therefore a very useful summary statistic!

Mathematically, t is the ratio of the departure of the sample mean from its hypothesized value to its standard error:

$t = \frac{(\bar{x} - \mu_0)}{s.e.}$

Let's assume that we are living in a null universe, and that our null hypothesis is that the sample mean is equal to zero.

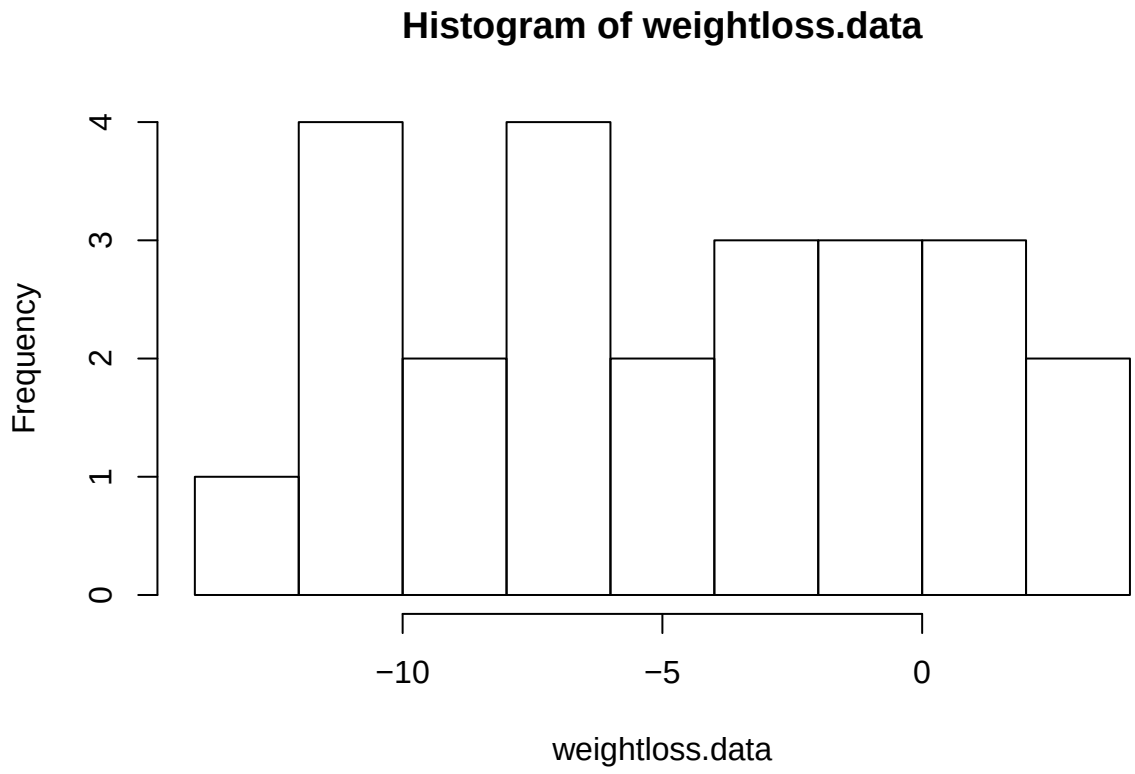
This is almost easier to conceptualize if we imagine a *paired* design in which, say, a set of individuals are monitored at two points in time: before a treatment and after. Let's imagine we are researching the effectiveness of a weight loss drug. We sample 25 overweight patients and weigh them before they start taking the drug and then weigh all of them 1 month after they start taking the drug. Never mind that this isn't the best drug trial!

So for each patient we compute the change in weight between the first and second measurements. In the null universe, the drug is not effective and we would expect the change to be no greater than expected by random chance. Our task is to convince ourselves that the results (mean weight loss across all patients) are meaningful, allowing us to break free of the null universe.

Remember: - the t statistic is the difference between the sample mean and some hypothesized value, standardized in units of standard error. - the t distribution is the sampling distribution for the t statistic.

Here is a worked example, in R:

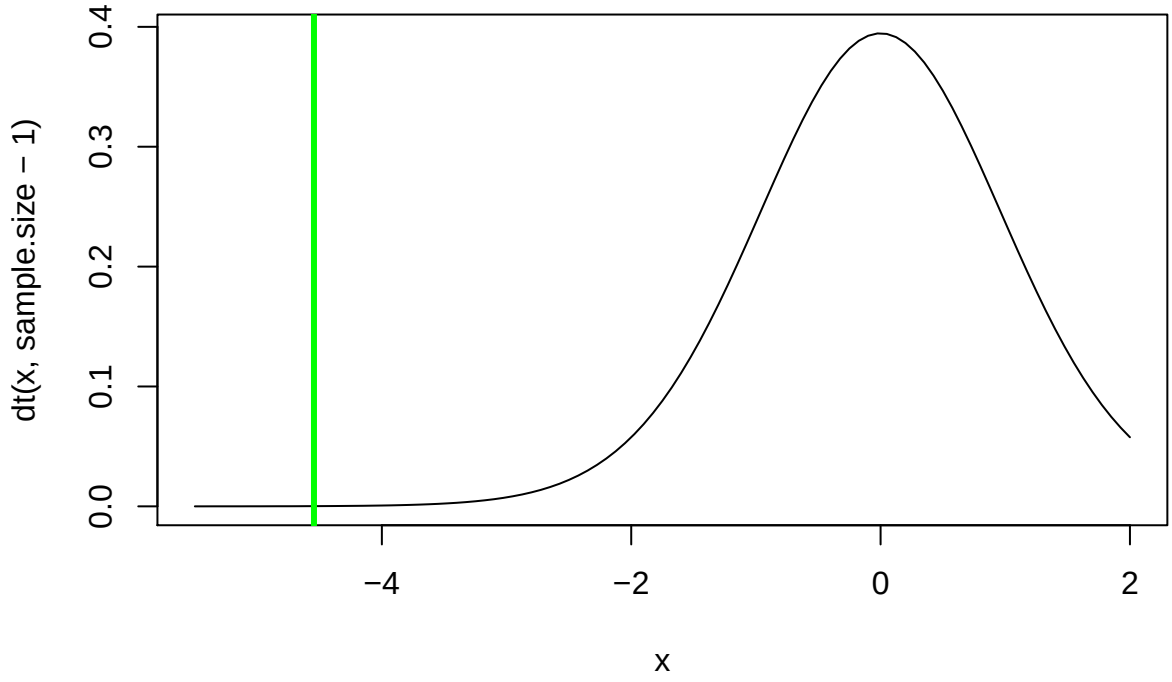
```
## Paired t-test example:

weightloss.data <- c(-10.4,-11.6,3.9,1.5,-0.3,-3.5 -10.0,-6.7,-6.1,-2.4,-6.0,2.3,0.1,-4.1,-3.2, -11.3,-3
hist(weightloss.data,breaks=7)
```



```r
mean.weightloss <- mean(weightloss.data)
null.weightloss <- 0
stdev.weightloss <- sd(weightloss.data)
sample.size <- length(weightloss.data)
std.error <- stdev.weightloss/sqrt(sample.size)

t.statistic <- (mean.weightloss-null.weightloss)/std.error
t.statistic
```

```
## [1] -4.544623
```

```r
curve(dt(x,sample.size-1),-5.5,2)
abline(v=t.statistic,col="green",lwd=3)
```




```r
p=pt(t.statistic,sample.size-1)
p    # this is the p value
```

```
## [1] 7.241049e-05
```

```r
####### Alternative: use R's built in t test

t.test(weightloss.data,alternative = "less")   # should get the same p=value!
```

```
##
##  One Sample t-test
##
## data:  weightloss.data
## t = -4.5446, df = 23, p-value = 7.241e-05
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf -2.966463
## sample estimates:
## mean of x
##   -4.7625
```

Q Does a significant p value here mean that everyone who takes the drug will lose weight? If not, what does it mean?

NOTE: as the df gets large, the t distribution approximates a *standard normal* distribution.

Fun fact: William Sealy Gosset – a student of Pearson, worked at Guinness brewery. He was hired to work out a way to determine the quality of stout. He published anonymously as 'student' because Guinness prevented its employees from publishing. That's why we call it the 'Student's t-test'!

Other statistics and sampling distributions

Pretty much all of classical statistics works this way. We compute summary statistics from data that have well-described sampling distributions. We climb into the null universe and, using the known sampling distribution, compute if our summary statistic (computed from our data) could have been a result of random sampling error. If not (if $p \leq \alpha$) then we remain unconvinced that the null universe is indeed false.

Goodness of fit/ Chi-square test.

This test asks the question "do the data sort into categories as you hypothesized?"

Karl Pearson – Founded the first stats dept, was a socialist and social Darwinist... and became a eugenicist.

Example: Are grad students more likely to be born in a given season?

Assumptions

- Data must be *categorical*
- Samples must be independent
- Data must be randomly sampled from the population
- There must be a sufficient sample size such that the expected number of observations in each categories is at least 5!

The Chi-squared statistic is computed from the data. The sampling distribution for the statistic is called the *Chi-squared distribution.*

The Chi-squared statistic:

$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - exp_i)^2}{exp_i}$

Here, exp_i is the expected number of observations in category i under the null hypothesis. x_i is the observed number in category i. So the Chi-squared statistic basically summarizes the degree to which the number of observations disagree with the expected number of observations across all categories.

Why is the numerator squared? First, this makes all deviations positive- negatives and positives can't cancel each other out. Second, it allows us to use the well-described Chi-squared distribution!

The Chi-squared distribution

The Chi-squared distribution is described (like the t distribution) by a certain degrees of freedom.

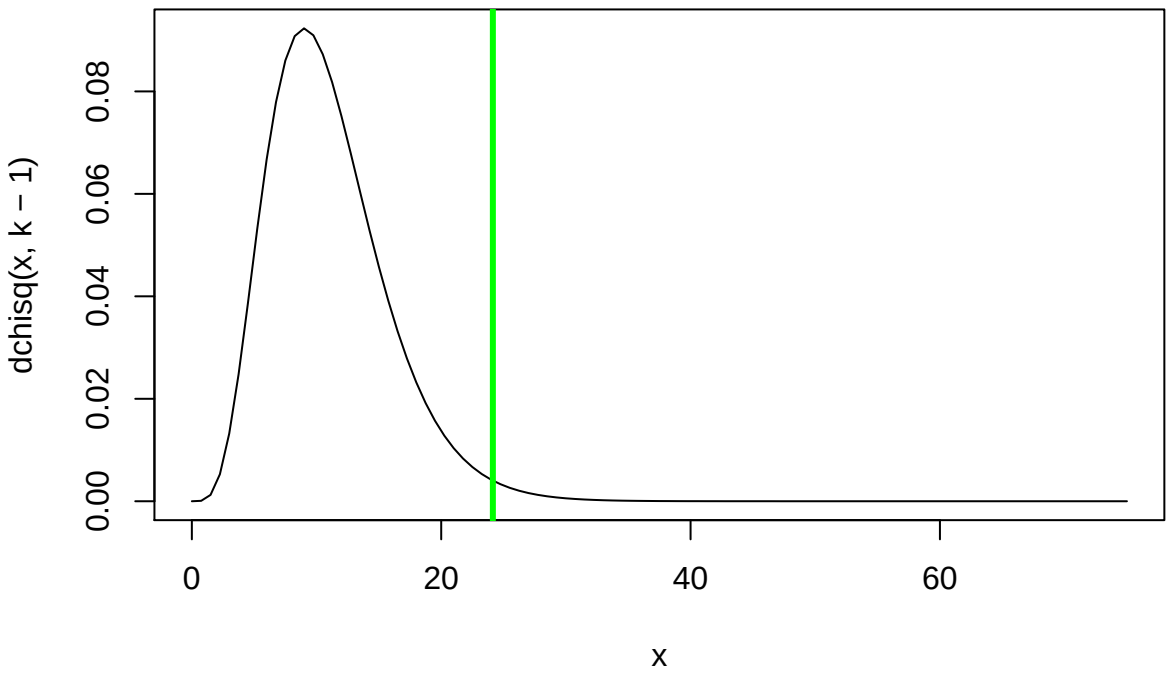
The degrees of freedom for this test is one less than the number of categories.

Example in R

```
## Chi squared example

birthdays.bymonth <- c(40,23,33,39,28,29,45,31,22,34,44,20)
months <- c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
names(birthdays.bymonth) <- months

sample.size <- sum(birthdays.bymonth)
k = length(birthdays.bymonth)   # number of categories
exp.birthdays.bymonth <- sample.size*rep(1/k,times=k)

Chisq.stat <- sum((birthdays.bymonth-exp.birthdays.bymonth)^2/exp.birthdays.bymonth)
Chisq.stat
```

```
## [1] 24.14433
```

```
## View the summary statistic along with its sampling distribution under the null hypothesis

curve(dchisq(x,k-1),0,75)
abline(v=Chisq.stat,col="green",lwd=3)
```




```
p <- 1-pchisq(Chisq.stat,k-1)
p
```

```
## [1] 0.01213825
```

```
### use R's built in chi squared function

chisq.test(birthdays.bymonth)     # should get the same p value!
```

```
##
##  Chi-squared test for given probabilities
##
## data:  birthdays.bymonth
## X-squared = 24.144, df = 11, p-value = 0.01214
```

Z-tests

The Z-test is a simpler version of the t test for when sample size is large enough (n>50) or the population dispersion is known.

The Z statistic is:

$Z = \frac{(\bar{X}-\mu_0)}{s.e.}$

As you can see, this looks very much like the t statistic! Actually it is exactly the same. The difference is that we are now assuming that the sampling variation for this statistic is normally distributed!

Example z-test: Golden State Warriors height

Tall basketball players are often successful. The Golden State Warriors are often successful. How tall are the Golden State Warriors relative to other basketball players in the NBA?

The GSW are 15 players with an average height of 6'8" (80"). A large survey of NBA players suggests they are, on average, 6'7" tall (79") with a standard deviation of 4 inches.

NOTE: the z-test is only appropriate here because we know the population standard deviation.

NOTE: B. Sullivan obtained those values from research into the height of NBA players.

Here, the mean (mu) and the standard deviation (q) are coming from the population (all NBA players) instead of the sample. There are situations where you can use the standard deviation of the sample, but those should be only when the sample size is very large.

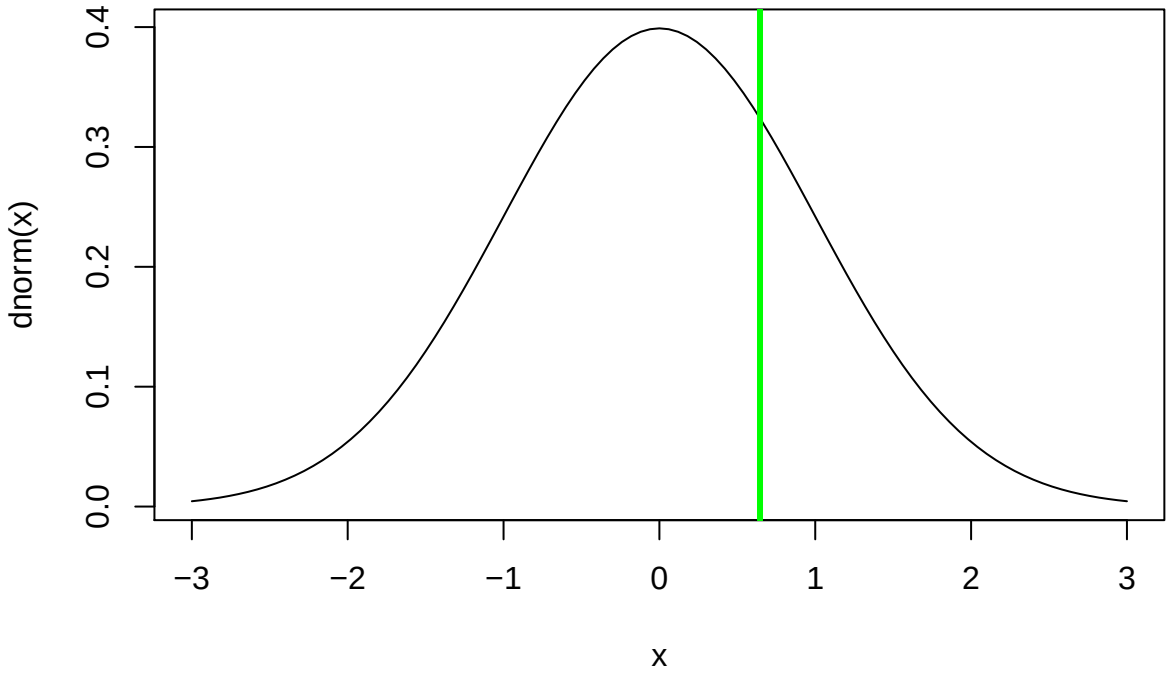
```
## Z test  (Ben Sullivan example)

df <- read.csv("GSW_height.csv")
GSWheight <- df$Height
GSWheight
```

```
##  [1] 75 75 81 79 78 84 79 81 81 79 83 79 83 81 77
```

```
mean.gsw <- mean(GSWheight)
sd.gsw <- sd(GSWheight)
sd.pop <- 4
n <- length(GSWheight)
s.e. <- sd.pop/sqrt(n)



null.height <- 79

z.statistic <- (mean.gsw-null.height)/s.e.
z.statistic
```

```
## [1] 0.6454972
```

```
curve(dnorm(x),-3,3)
abline(v=z.statistic,col="green",lwd=3)
```




```
p <- 1-pnorm(z.statistic)    # is the p value enough evidence to tell you that GSW players are taller t
p
```

```
## [1] 0.2593025
```

```
pnorm(z.statistic)
```

```
## [1] 0.7406975
```

Q That means that 74% of NBA players are shorter than the GSW average height, right?

Wrong- that means that 74% of samples of 15 players drawn randomly from the population of NBA players would have a mean height less than or equal to the mean height of the GSW players. Conversely 26% of samples would have a mean height greater than that of the GSW team (that's the p value).

Q So can you reject the null hypothesis and say that the GSW team is taller than expected by random chance?

F tests (ANOVA, regression)

Fill this in

The p-value: good or evil?

There has been a lot of hating on p-values in recent decades. Largely this is due to frequent misinterpretation of what they mean, unthinking use of the alpha=0.05 threshold, and failure to report other important pieces of information like effect sizes and confidence intervals (reporting only p-values is called the 'naked p-value' phenomenon!).

The American Statistical Association (ASA) released a statement about p-values and their misuse:

Summary of ASA statement
ASA statement

P-values remain very important and widely used and reported. But we need to keep in mind the common pitfalls and try to avoid them!

Probability distributions- the basics (and how to use them in R)

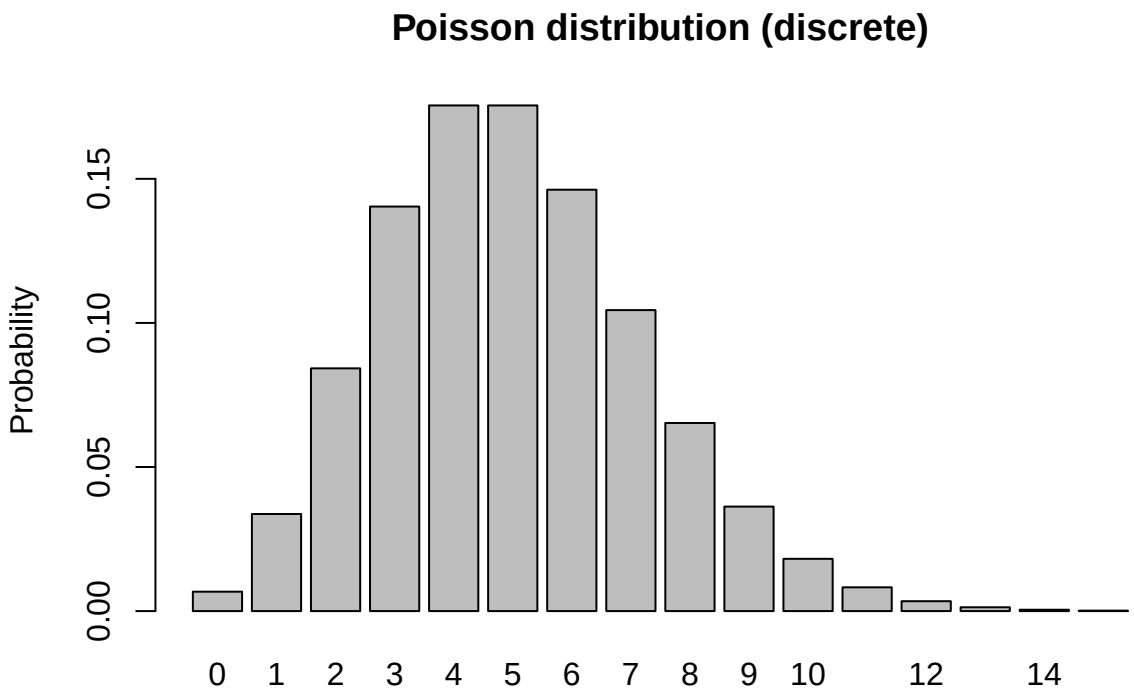
Discrete vs. continuous

In *discrete distributions*, each outcome has a specific probability (like the probability of flipping a coin 10 times and getting 4 heads). For example, let's consider the Poisson distribution

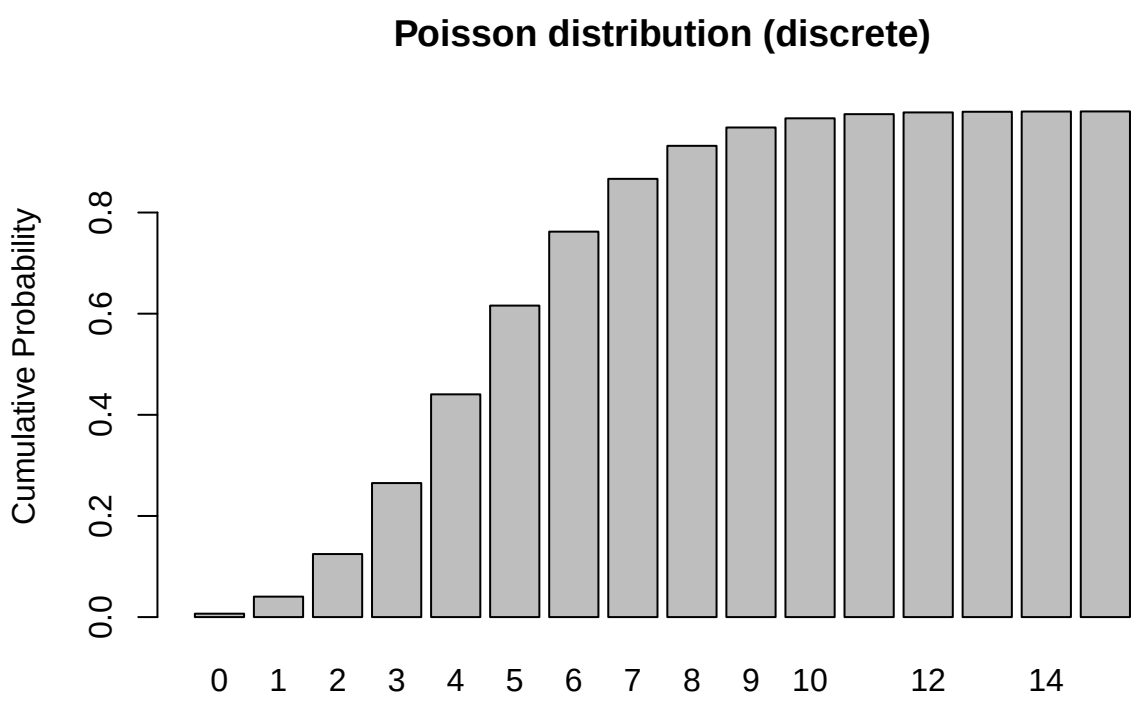
```
#################
# Probability distributions

mean <- 5
rpois(10,mean)     # the random numbers have no decimal component
```

```
##  [1] 3 7 6 5 4 7 4 5 4 6
```

```
             # plot a discrete distribution!
xvals <- seq(0,15,1)
probs <- dpois(xvals,lambda=mean)
names(probs) <- xvals

barplot(probs,ylab="Probability",main="Poisson distribution (discrete)")
```




```
barplot(cumsum(probs),ylab="Cumulative Probability",main="Poisson distribution (discrete)")   # cumulat
```




```
sum(probs)   # just to make sure it sums to 1!  Does it???
```

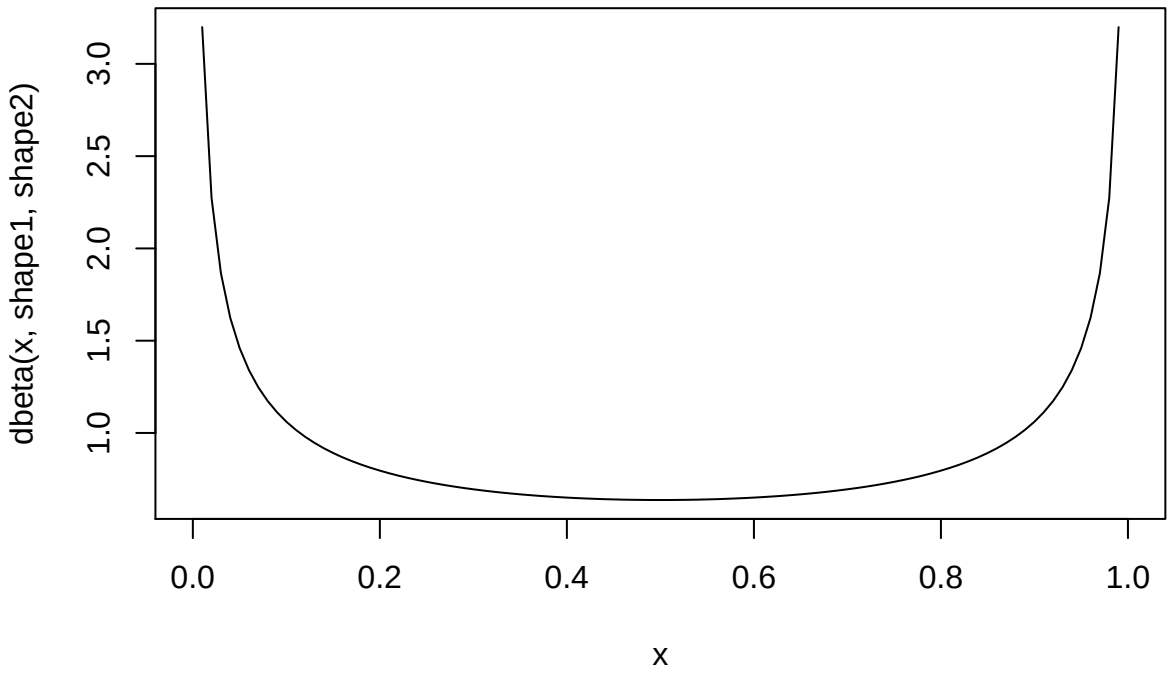
```
## [1] 0.999931
```

In *continuous distributions*, the height of the curve corresponds to *probability density*, $f(x)$, not probability $Prob(x)$. This is because the probability of getting exactly one value in a continuous distribution is effectively zero. This arises from the problem of precision. The sum of the probability distribution must be 1 (there is only 100% of probability to go around). In a continuous distribution, there are an infinite number of possible values of x. So any individual probability is always divided by infinity, which makes it zero. Therefore we have to talk about probability density, unless we want to specify a particular range of values – we can't calculate $Prob(x = 5)$, but we can calculate $Prob(4 < x < 6)$ or $Prob(x > 5)$. Let's consider the beta distribution:

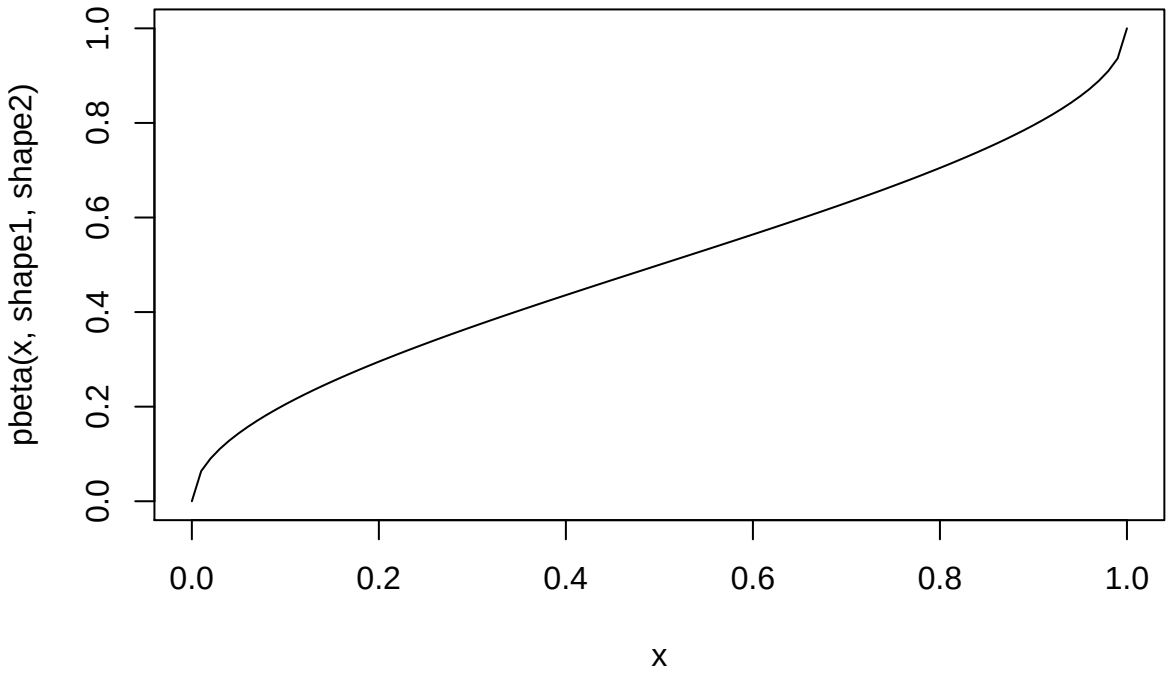
```
#########
# continuous distributions

shape1 = 0.5
shape2 = 0.5

rbeta(10,shape1,shape2)
```

```
##  [1] 0.5276461187 0.0524156878 0.9968286948 0.4472368271 0.0000248312
##  [6] 0.9497981390 0.6738454838 0.8946385237 0.0752431869 0.7166812004
```

```r
curve(dbeta(x,shape1,shape2))   # probability density
```




```r
curve(pbeta(x,shape1,shape2))   # cumulative distribution
```




```
integrate(f=dbeta,lower=0,upper=1,shape1=shape1,shape2=shape2)    # just to make sure it integrates to
```

```
## 1 with absolute error < 3e-06
```

Some other probability distribution terms:

Moments – descriptions of the distribution. For a bounded probability distribution, the collection of all the moments (of all orders, from 0 to infinity) uniquely determines the shape of the distribution.

- The zeroth central moment ($\int (x-\mu)^0 Prob(x)\partial x$) is the total probability (i.e. one),
- The first central moment ($\int (x-\mu)^1 Prob(x)\partial x$) is $\mu - \mu = 0$.
- The second central moment ($\int (x-\mu)^2 Prob(x)\partial x$) is the variance.
- The third central moment ($\int ((x-\mu)/\sigma)^3 Prob(x)\partial x$) is the skewness.
- The fourth central moment is the kurtosis.

Parameters – the values in the probability distribution function, describing the exact shape and location of the distribution. *Parametric statistics* require assuming certain things about distributions & parameters, while *nonparametric stats* do not require these assumptions.

PDF -

CDF -

Quantiles -

Some probability distributions

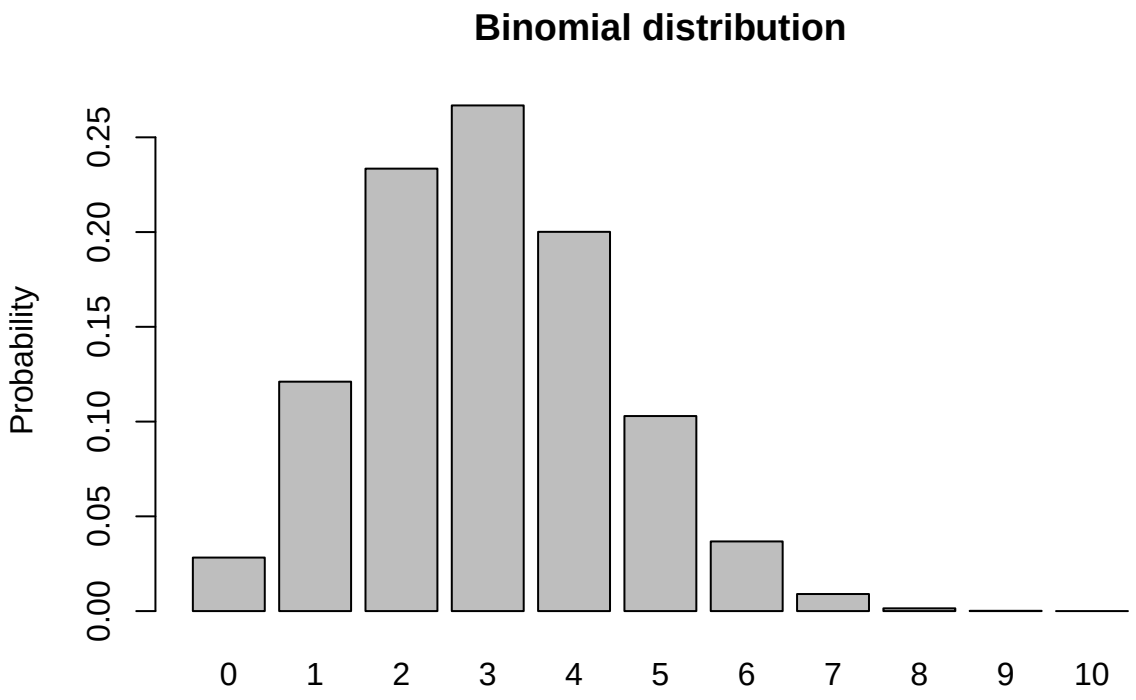
The Bolker book goes through the main distributions we will be using in this course. Pay particular attention to the type of *process* described by each distribution. The key to using these distributions to represent random variables is to figure out which statistical process best matches the ecological process you're studying, then use that distribution. e.g., am I counting independent, random events occurring in a fixed window of time or space (like sampling barnacles in quadrats on an intertidal bench)? Then the distribution of their occurrence is likely to follow a Poisson or Negative Binomial distribution.

Binomial

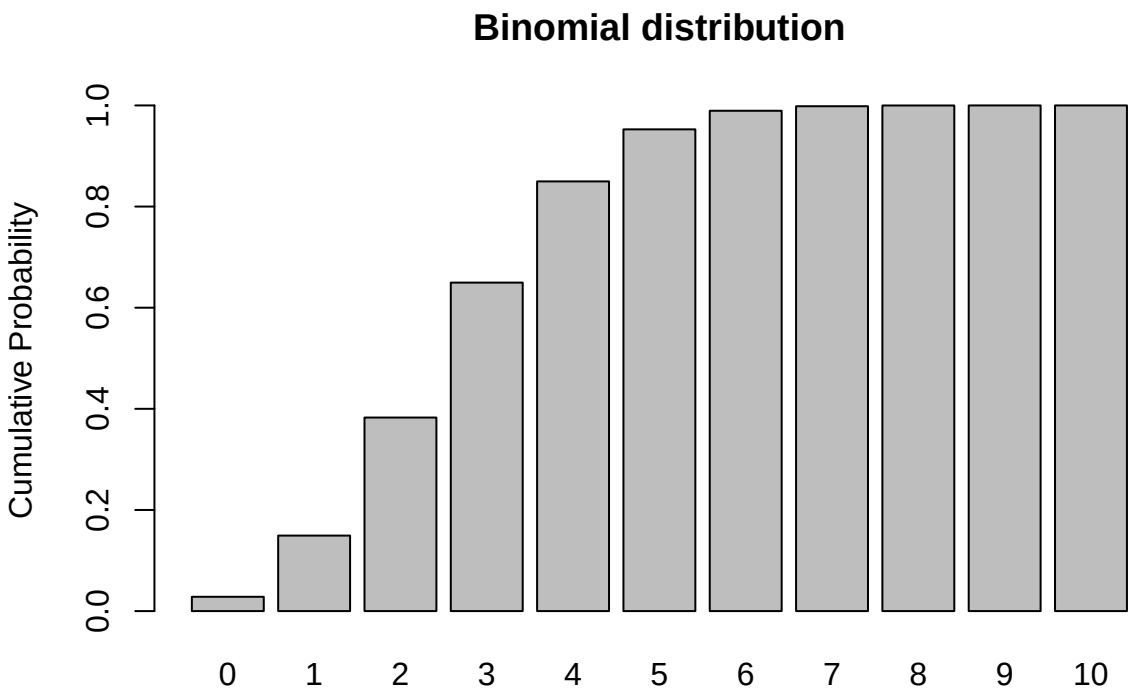
```
##########
# Binomial

size <- 10
prob <- 0.3
rbinom(10,size,prob)
```

```
##  [1] 3 3 4 5 2 3 4 4 5 4
```

```
xvals <- seq(0,size,1)
probs <- dbinom(xvals,size,prob)
names(probs) <- xvals

barplot(probs,ylab="Probability",main="Binomial distribution")
```




```
barplot(cumsum(probs),ylab="Cumulative Probability",main="Binomial distribution")   # cumulative distri
```




```r
sum(probs)   # just to make sure it sums to 1!  Does it???
```

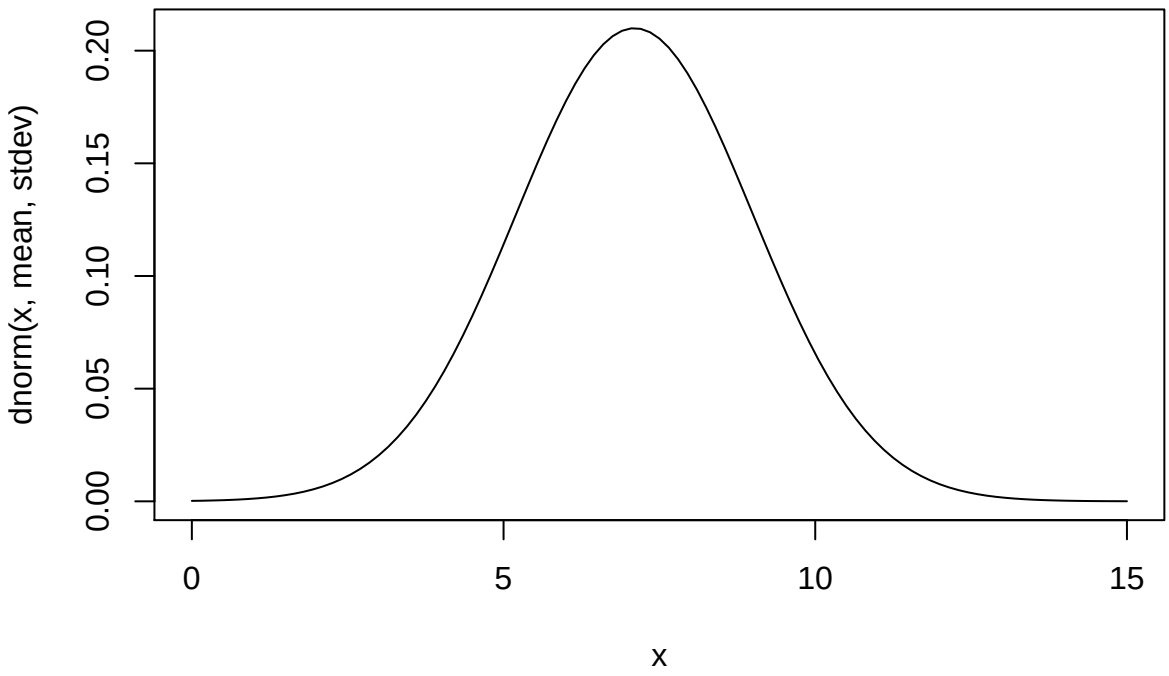
```
## [1] 1
```

Normal

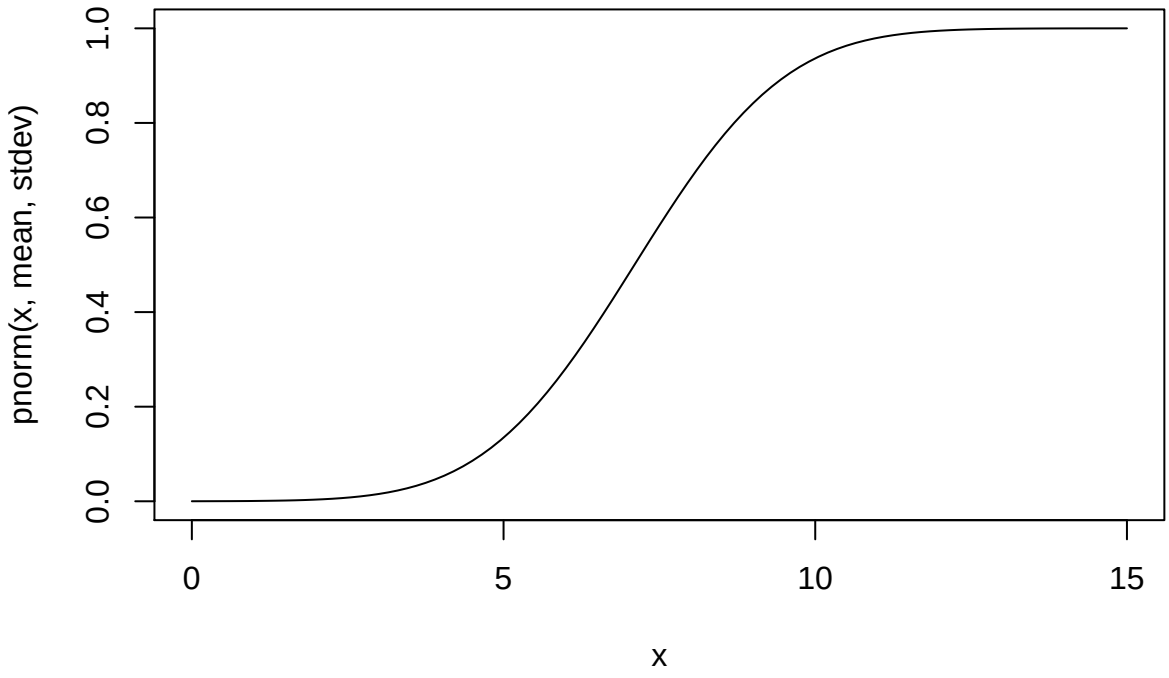
```r
#########
# Gaussian

mean = 7.1
stdev = 1.9

rnorm(10,mean,stdev)
```

```
##  [1]  6.266271  7.482264  6.585611  6.852910 10.959987  6.699947  8.651143
##  [8]  8.291593  5.111104  2.489618
```

```r
curve(dnorm(x,mean,stdev),0,15)   # probability density
```




```
curve(pnorm(x,mean,stdev),0,15)    # cumulative distribution
```




```
integrate(f=dnorm,lower=-Inf,upper=Inf,mean=mean,sd=stdev)    # just to make sure it integrates to 1!!
```

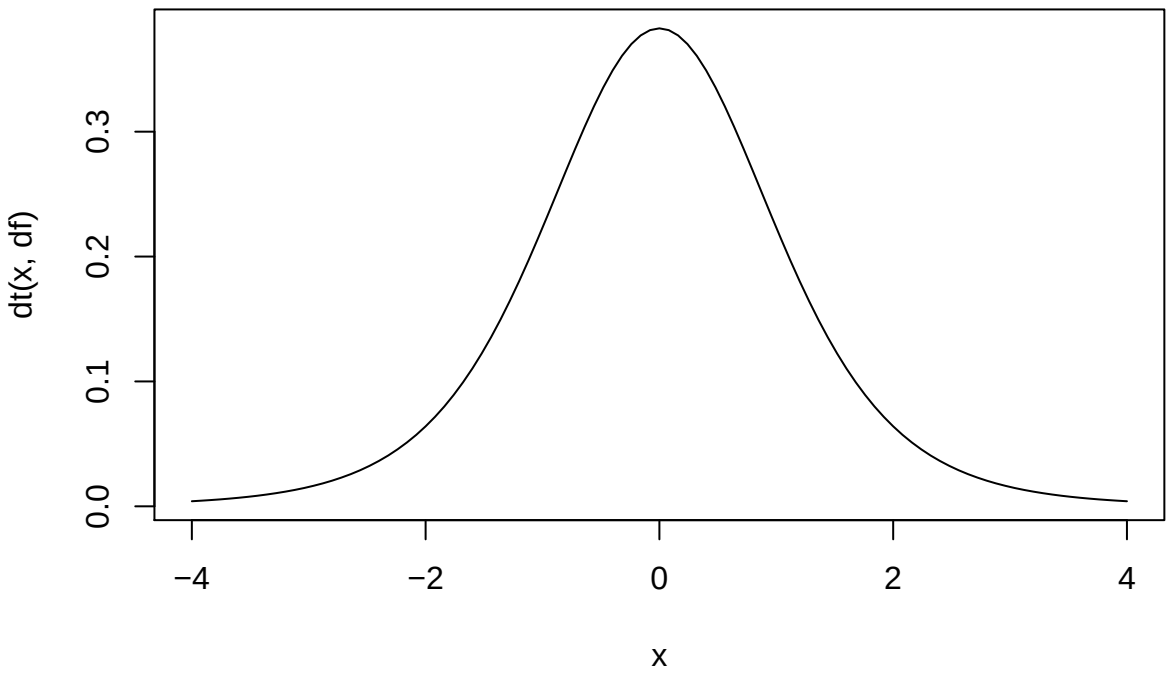
```
## 1 with absolute error < 1.1e-05
```

t distribution

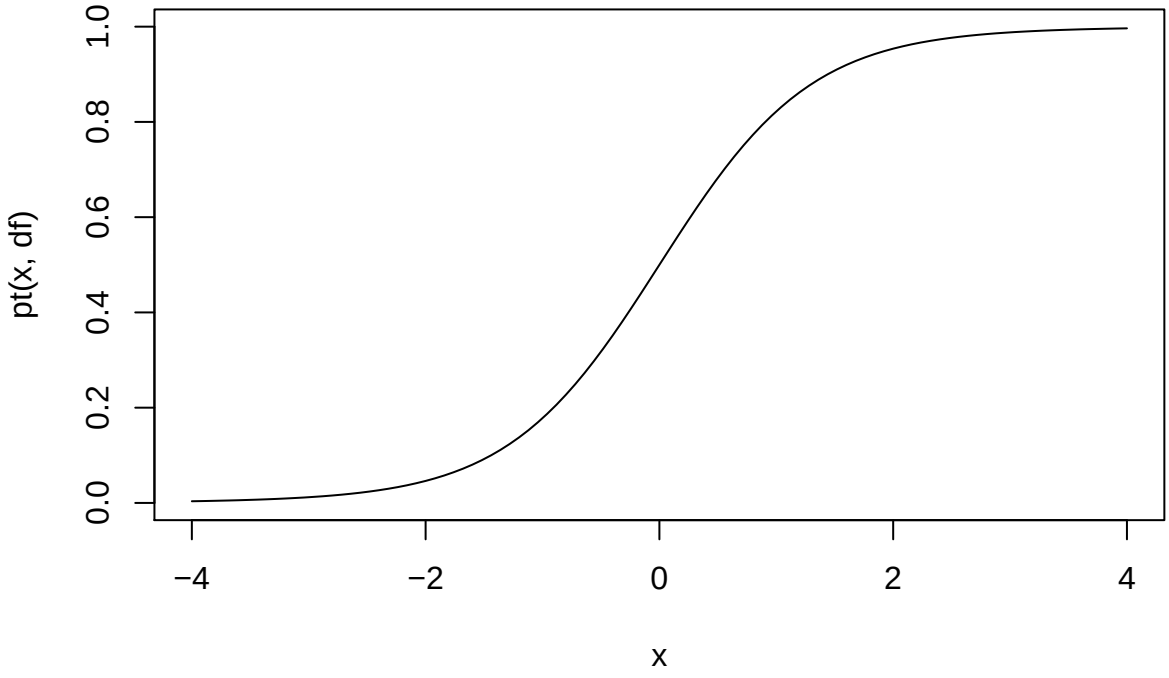
```
#########
# t distribution

df = 6

rt(10,df)     # random numbers from the t distribution (not sure why you would ever want this!)
```

```
##  [1]  0.56265829 -0.61631073  0.16562778 -0.88227661 -0.75621254
##  [6] -0.36654591  0.04490872 -4.34649441  0.37245542  0.55697201
```

```
curve(dt(x,df),-4,4)   # probability density
```




```
curve(pt(x,df),-4,4)    # cumulative distribution
```




```
integrate(f=dt,lower=-Inf,upper=Inf,df=df)    # just to make sure it integrates to 1!!
```

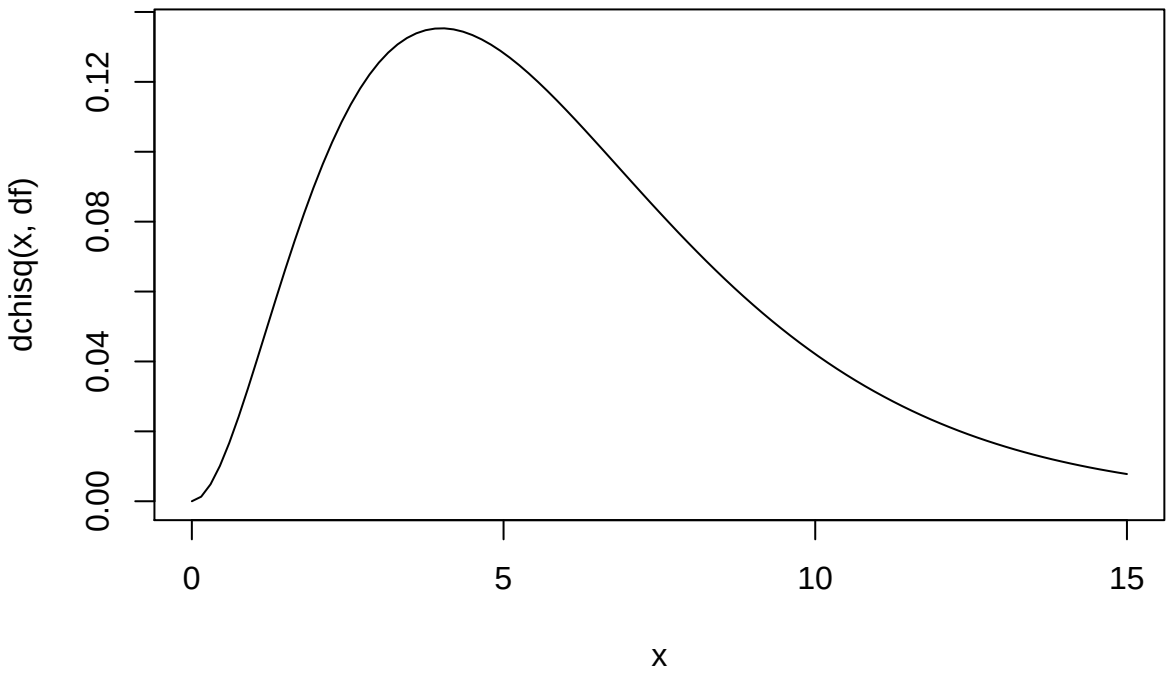
```
## 1 with absolute error < 1.9e-05
```

Chi-squared distribution

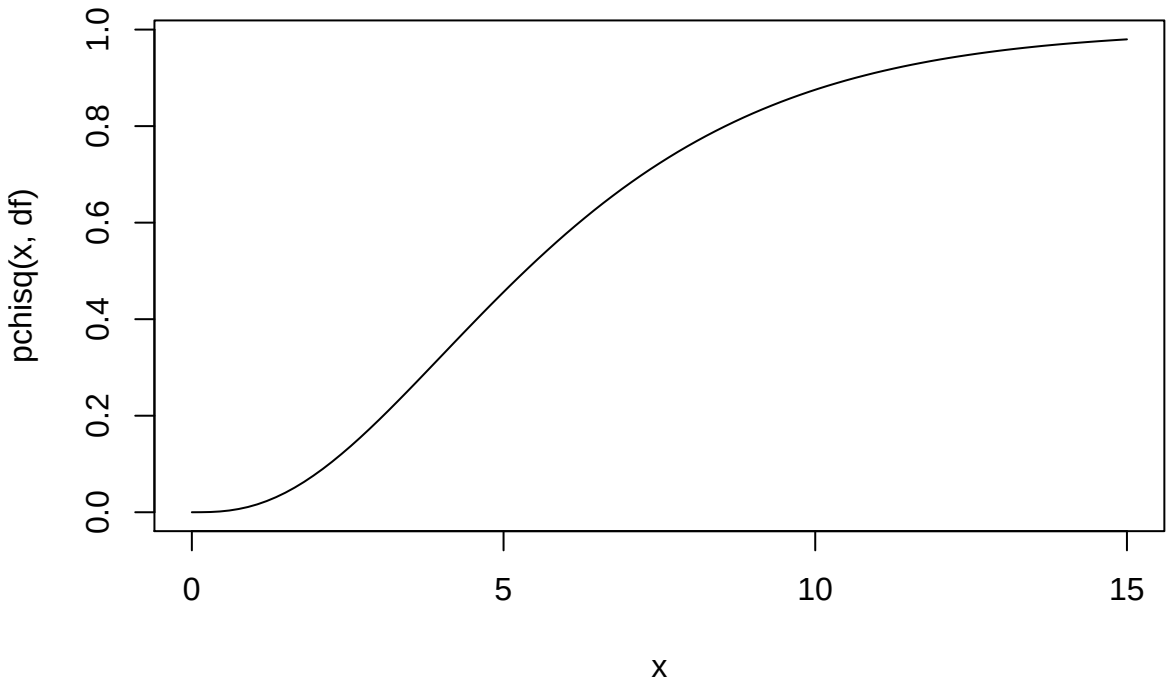
```
#########
# Chi-squared distribution

df = 6

rchisq(10,df)     # random numbers from the t distribution (not sure why you would ever want this!)
```

```
##  [1]  7.662049  7.869946  6.863613  9.557103  2.545120  1.114466  3.393224
##  [8]  6.357000  3.774204 10.238327
```

```
curve(dchisq(x,df),0,15)   # probability density
```




```
curve(pchisq(x,df),0,15)    # cumulative distribution
```




```r
integrate(f=dchisq,lower=0,upper=Inf,df=df)    # just to make sure it integrates to 1!!
```

```
## 1 with absolute error < 2.3e-05
```

Exercise:

Visualize (in R) the following distributions as above: Gamma, Exponential, Lognormal, Negative Binomial.

–go to next lecture–