Taking Your Class for a Walk, Randomly
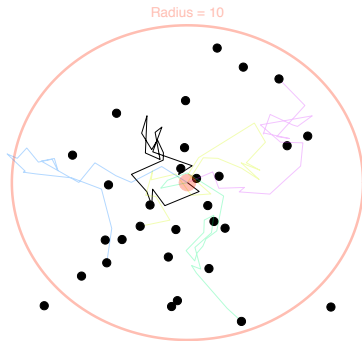
Daniel Kaplan
Macalester College

Oct. 27, 2009

Overview of the Activity

You are going to turn your students into an ensemble of random walkers. They will start at (approximately) the same point, follow different random paths, and spread out (**diffuse**) accordingly. After 20 steps, a bird's eye view of your class will be like this:

Total Distance for n = 20 steps

Philosophical Background I

- ▶ Statistics is largely about variation: how to describe it, how to extract information from it, how to structure it through sampling or experiment.
- ▶ In my classroom activities, I like to have the students each perform the same activity, but randomized in some way so that the class as a whole serves as a sampling distribution. Doing this makes it clear to students that the interpretation of their own, individual result makes the most sense when put in the context of the variability among the students.

Simulations and the Computer

The computer is an indispensable tool for carrying out simulations.

- It makes it easy to undertake statistical computations.
- It allows a random sample easily to be selected from a file.
- It allows repetition to be automated.

Most of my classroom activities involve the computer in some essential way.

For today's activity, I wanted something that doesn't require a computer.

Why No Computer Today?

The reason is that many instructors and students don't have ready access to a computer, or don't use software that makes it easy to randomize or repeat. This is a shame. Randomization and repetition are central to statistical reasoning and students ought to have the tools to carry them out.

I hope that this activity will help to convince instructors that randomization (whether in sampling or in permutations/resampling of existing data) can provide important insight for students. Computation is not just about getting the result in one situation, but being able to place that result in the context of the results that might have happened in a different random realization.

Some Simulations that Use the Computer I

In each of these, I have the students in the class perform some computation based on either a random sample of data or a randomization of the same data.

- ▶ Sampling distribution. Students draw a random sample from a population and all calculate the same sample statistic: perhaps the mean or a regression coefficient. The resulting distribution reflects sampling variability.

- ▶ Resampling distribution. Similar to the above, but the random sample is based on drawing from an existing sample with replacement: bootstrapping.

- ▶ Coverage. Students compute a population parameter. Then they draw a sample from the population and compute a confidence interval on a sample statistic. (A 50% level works well.) How many of the different students' confidence intervals include the population parameter?

Some Simulations that Use the Computer II

▶ Power. Students compute a population parameter, e.g. a difference in group means or a regression coefficient. Then they draw samples of some size n and carry out a hypothesis test on the corresponding sample statistic. How often do they reject the null hypothesis? How does this depend on the sample size n?

▶ Tabulated distributions. Students carry out a hypothesis test (on a random sample from randomized data, so the null is true) and report the test statistic, for instance the t-value or the F value in ANOVA. We then compare this distribution to the tabulated version.

▶ Significance. Like the above, but students look to see how often they can reject the null hypothesis. I like to do this where the data involve multiple groups, and have the students consider the single lowest p-value when deciding whether they have "found something."

Arranging a Random Walk I

1. Provide each of your students with a unique series of about 40 random angles, between 0 and 360 degrees.

2. Go to the "arena," a large open area with a radius of about 10 paces and with a clearly defined boundary.

3. Mark a point in the middle as the starting point.

4. In front of your students, pace out the distance from the starting point to the boundary of the arena. Let's assume that this is 10 steps.

5. Ask your students how many steps it is going to take them to get from the center to the boundary. Remind them that they will be taking a random walk.

   This step is critical in forming student understanding. In my experience, almost all students will assume that the random nature of the walk imposes some fixed inefficiency. So, instead of taking 10 steps to get from the center to the boundary, they think it might take 20 or 30 steps.

Arranging a Random Walk II

6. Have your students crowd together (politely) near the center. Ask them to estimate their standard deviation: the typical distance from the center.

7. On your command, have them turn by their first random angle and take one step. Once this is done, ask for the standard deviation.

8. Repeat: another random turn and a step. Again, ask for the standard deviation.

9. Repeat this several times.

10. Ask them to take steps on their own until they have taken 20 steps altogether. Point out that few people are near the boundary.

11. Ask them to take more steps on their own until they have taken 40 steps altogether. Tease them that most people are not near the boundary. Some are still near the starting point.
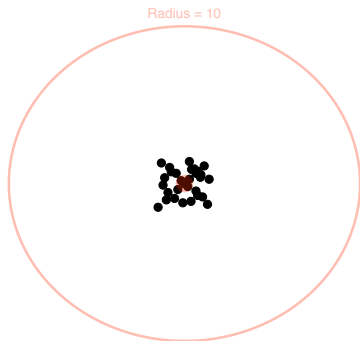
Arranging a Random Walk III

12. Optional: One at a time, ask the students to pace off the direct distance from where they are to the center point. Record this set of distances.
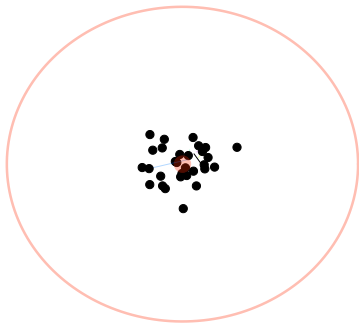
How the walk will progress.

Total Distance for n = 0 steps

Radius = 10

The start.

Note that the students can't fit exactly at the starting point — there is a certain spread.
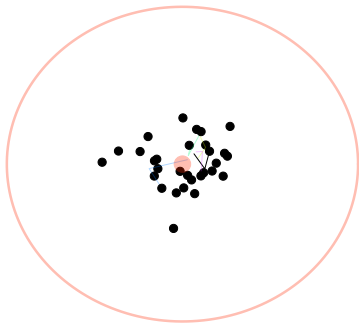
Total Distance for n = 1 steps

Radius = 10

After one step...

The students have spread out quite a lot.
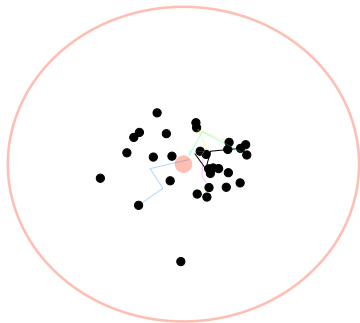
Total Distance for n = 2 steps

Radius = 10

After two steps ...

The spread continues in a discernible way.
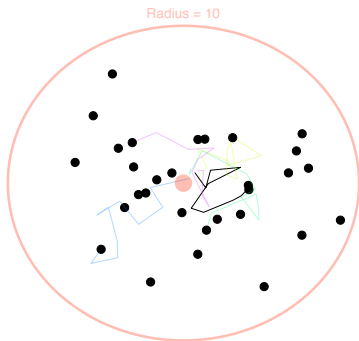
Total Distance for n = 3 steps

Radius = 10

After 3 steps ...

It's not obvious
that there is much
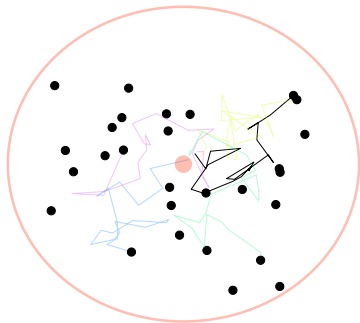more spread with
each step.

Total Distance for n = 10 steps

After 10 steps ...

The students are
only about twice as
far as after 3 steps.

Total Distance for n = 20 steps

After 20 steps …

Only a few students have reached the boundary.

Some Key Questions To Ask Before the Walk

1. How many steps will it take before a typical walker has reached the boundary?
2. How does this depend on the radius of the arena?
3. How to define sensibly a "typical walker?"

An In-class Demo of the Walk

It takes pretty long to conduct the actual random walk in the field. An arena of radius 10 steps is the biggest that's usually practicable given the time available. But it's helpful to be able to show longer walks.

- ▶ After carrying out the actual walk, it helps to reprise the walk back in the classroom, using a computer to generate the picture.
- ▶ Then, after showing a walk similar to the one the students took, carry out walks on much larger arenas.

Software

You can do this with software. I provide a simple package that runs under R but requires little or no familiarity with R. Details of installing the software are given as an appendix.

How the Software Works

1. Start up R by clicking on the file **random-walk.Rdata**.
2. At the command prompt, use the startwalk command to create an arena and populate it with walkers.

   ```
   > startwalk( arena=10, nwalkers=100)
   ```

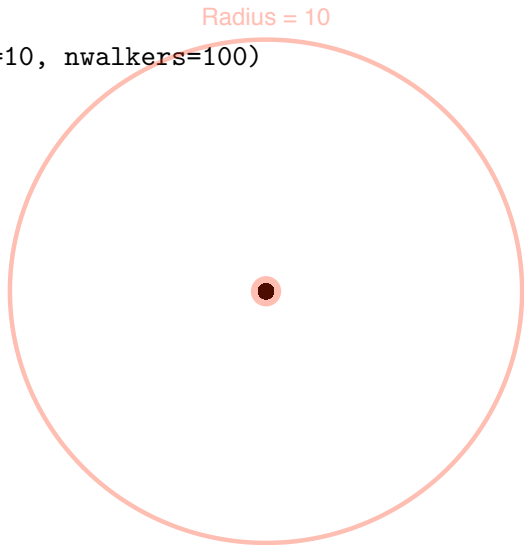   This arena has radius 10 steps. There are 100 "students" walking.
3. Use the walk command to take some steps.

   ```
   > walk(1)
   ```

   The argument, 1 in this case, tells how many steps to take. Each additional walk command will add more steps to the walk. These will accumulate until you start over with a new startwalk command.

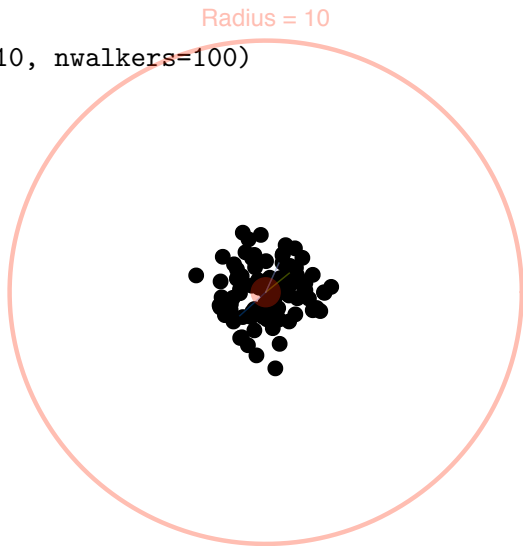# A Reprise of the Walk
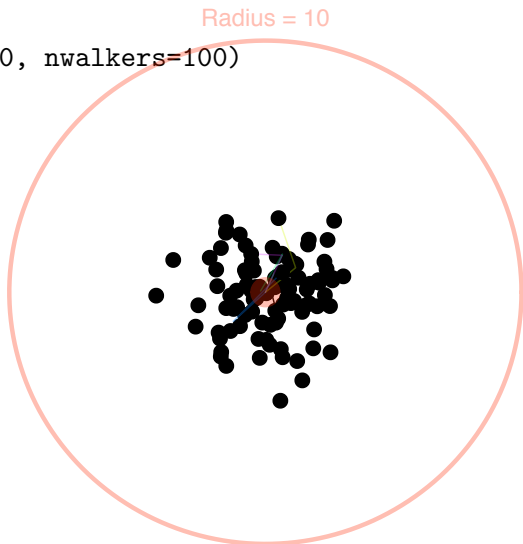
1. `startwalk(arena=10, nwalkers=100)`
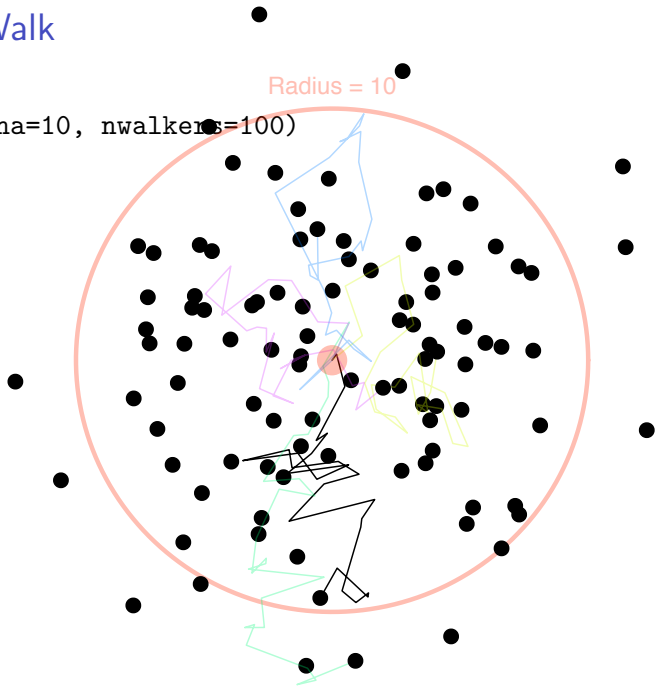

Radius = 10

A Reprise of the Walk

1. `startwalk(arena=10, nwalkers=100)`
2. `walk(1)`

Radius = 10

A Reprise of the Walk

1. `startwalk(arena=10, nwalkers=100)`
2. `walk(1)`
3. `walk(1)`

Radius = 10

A Reprise of the Walk

1. `startwalk(arena=10, nwalkers=100)`
2. `walk(1)`
3. `walk(1)`
4. `walk(23)`

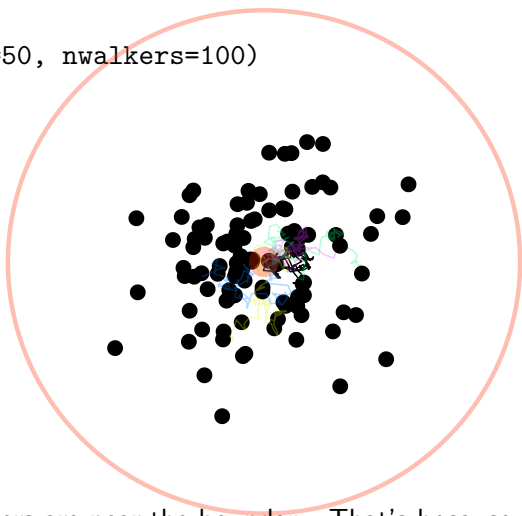Radius = 10

A Longer Walk: Boundary at Radius = 50 Steps

The larger the arena, the longer it takes for a typical walker to reach the boundary. If the arena is m steps in radius, it will take $n = m^2$ steps for the typical walker to reach the boundary. For large m, this can be very large indeed.

1. `startwalk(arena=50, nwalkers=100)`

A Longer Walk: Boundary at Radius = 50 Steps

1. `startwalk(arena=50, nwalkers=100)`
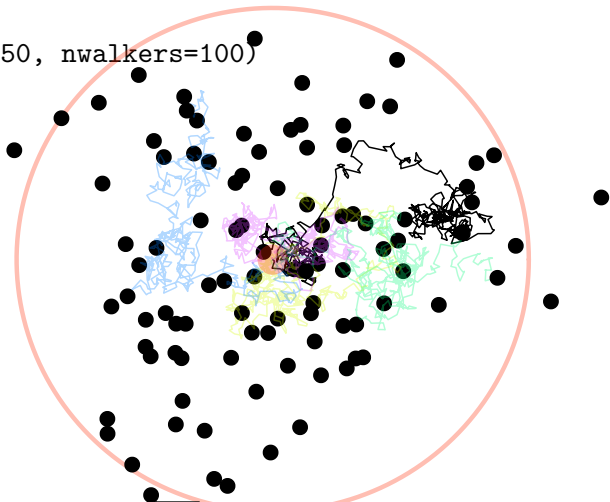2. `walk(100)`

After 100 Steps

Hardly any of the walkers are near the boundary. That's because after 100 steps, the typical distance is just $\sqrt{100} = 10$, only $1/5$ of the way to the boundary.

A Longer Walk: Boundary at Radius = 50 Steps

Radius = 50

1. `startwalk(arena=50, nwalkers=100)`
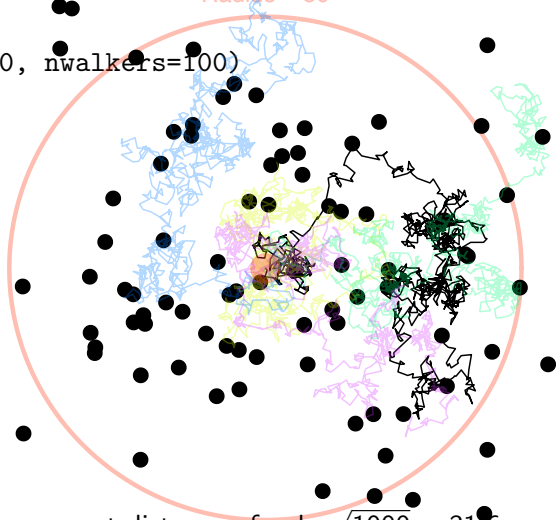2. `walk(100)`
3. `walk(400)`

After 500 Steps

The typical walker has gone only $\sqrt{500} \approx 22.3$. That's only about half the radius of the arena.

A Longer Walk: Boundary at Radius = 50 Steps

1. `startwalk(arena=50, nwalkers=100)`
2. `walk(100)`
3. `walk(400)`
4. `walk(500)`

After 1000 Steps

The typical walker has gone a net distance of only $\sqrt{1000} \approx 31.6$, so a small proportion of the walkers have reached the boundary.

The Theory

The typical distance grows as $\sqrt{n}$.

Another way to think of this is as the mean square distance. This grows as n.

This realization is pretty recent. It was the subject of Einstein's paper on Brownian motion in 1905.

Why the Square Distance?

We're interested in **how far** your walk took you, not in the direction you went. If you were at position (x, y), the distance you travel is $\sqrt{x^2 + y^2}$. Notice that to get this, you take squares of each coordinate.

In doing the calculation of distance algebraically, we first find the **square distance** $x^2 + y^2$ and then take the square root.

Calculating the Average Distance of a Two-Step Walk

- ► Two steps of a walk make the legs of a triangle.
- ► Each leg has length 1 step.
- ► The net distance is the length of the third side of the triangle, opposite the **included angle** α.

Examples:
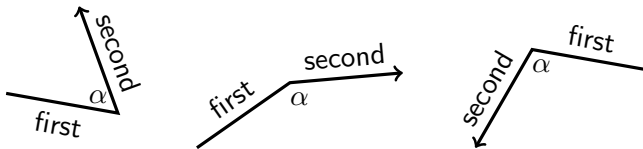
And, showing the net distance:
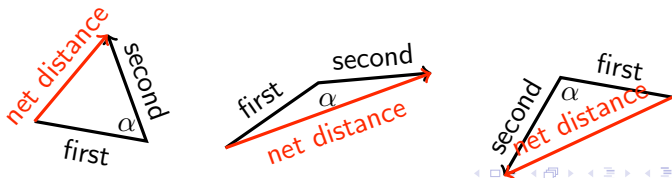
The Included Angle is Random

The included angle is random, anywhere between 0 and 180 degrees.

On average, the included angle is 90 degrees. The typical two step walk is a right triangle!

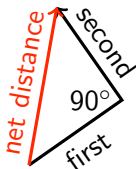# Right Triangles and the Pythagorean Theorem

The Pythagorean theorem says that the length of the average 2-step walk is given by the hypotenuse of the triangle:

$$||\text{two step walk}||^2 = ||\text{step 1}||^2 + ||\text{step 2}||^2.$$

It's square lengths that add for a random walk.
So after two steps, the average square length is $2||\text{step}||^2$. The length itself is $\sqrt{2}||\text{step}||$.
The logic can be extended to three steps and on — just remember to add the square lengths in the Pythagorean way. After n steps, the square length is $n||\text{step}||^2$, so the length itself is $\sqrt{n}||\text{step}||$.

Summary of What We Know

- The net direction is completely random. That's why students were scattered all over the place.
- The distance is also random, but we make some definite statements about it.
  - OBVIOUS: It can't be less than zero. Distances can't be negative!
  - OBVIOUS: It can't be longer than n steps.
  - NOT SO OBVIOUS: The typical distance will be $\sqrt{n}$.

Compare to the Purposeful Walk

When you're trying to get from one place to another, you don't walk at random. You walk in straight lines.

- ▶ The total distance you travel is $n \times$ step size.
- ▶ The average distance, per step, is $\frac{1}{n} \times n \times$ step size, or, just the step size itself. Obvious!

For a random walk, things are different:

- ▶ The total distance you travel is random.
- ▶ Typically, it has length $\sqrt{n} \times$ the step size.
- ▶ The average distance, per step, that each step takes you is $\frac{1}{n}\sqrt{n} \times$ step size, or, more concisely $\frac{1}{\sqrt{n}} \times$ step size.

What's the Point?

Almost everyone has an intuition about **linear growth**. In this setting, that means that the distance you travel — start-to-end distance or **net distance** — grows with the number of steps n in a linear way.

In a random walk, the net distance grows typically as $\sqrt{n}$.

Students start this walk with a conception that it will take 10 or 20 or 30 paces to walk to the boundary of the arena, that they will cover half the distance in the first half of the walk and the other half in the second half of the walk. Obvious. But wrong!

Experiencing the random walk, physically, gives time for the lesson to sink in and helps students see why their intuition is wrong.

A Related Setting: The Stock Market

Stock prices fluctuate, seemingly at random.

Look at minute-to-minute fluctuations. You can easily do this at sites such as http://finance.yahoo.com/ which update prices every 30 seconds or so.

Show the stock price at the beginning of your class. Then five minutes later, and five minutes later, and so on. Extrapolate the growth to the end of the day.

Also note that day-to-day stock price data is readily available over periods of years.

Learning Goals I

1. Disrupt, in a memorable way, the intuition most people have that growth is linear. Students expect quantities to grow linearly with n. But this isn't what happens with random phenomena.

   It's important to get the students, before they undertake the walk, to estimate how many steps it will take them to reach the boundary. Then, when they don't get nearly as far as they expect, they are forced to confront their intuition.

2. Develop an understanding of the reason why $\sqrt{n}$ is related to random sums.

Context for Use

- Requires roughly 20 students or more.
- Can be used at the introductory or the advanced level.
- Time needed:
    - Pre-preparation: Identify a suitable setting for the walk: a large empty space, a field, etc.
    - Preparation in class: 5 minutes to explain the rules and create the random angles.
    - The walk itself: 15 minutes.
    - Reviewing the walk back in class. 15 minutes or more depending on how much depth you go into.

Teaching Materials

- ▶ Handout for angles. See `www.macalester.edu/~kaplan/ISM/RandomWalkSeminar/random-angles.pdf`
- ▶ Software for computer simulation. For those who don't know R, here is an R Workspace at `www.macalester.edu/~kaplan/ISM/RandomWalkSeminar/walk-simulation.Rdata` All you have to do is download this file to your computer, and double click on it to start R. (You must install R first. See `www.r-project.org`.
- ▶ If you already know R, here are the functions themselves, to be "sourced" into R: `www.macalester.edu/~kaplan/ISM/RandomWalkSeminar/random-walks.r`

Generating Random Angles

Each student needs a list of about 40 random angles between 0 and 360 degrees. It suffices to round these, even very coarsely.

- ▶ I have the students generate these on their own computers.
- ▶ If students have calculators, they can use the calculators and generate the random angles as they go. (It's harder to find mistakes here, so watch out!)
- ▶ The page `www.macalester.edu/~kaplan/ISM/RandomWalkWebinar/random-angles.pdf` has several sets of random angles. If the students are skeptical about the randomness, tell them to start in the middle and work backwards, or to pick whatever angle they want and cross it off as they go, or something similar.

Exercise 1

Over a period of 100 trading days, you keep track of the Dow Jones Industrial Average (DJIA) — a stock market index. During that time, it went up by $500. This is an increase of $5 per day on average. If the DJIA is a random walk, without any real trend, then the $500 overall change is just the result of summing up random day-to-day variation. Assuming this is the case, and that the $500 overall change is a typical outcome from the random walk, answer the following questions:

1. What is a good estimate of the typical dollar amount that the DJIA goes up or down randomly on any one trading day?
   1 5 10 25 50 100 150 200 $\boxed{\text{NA-1}}$

2. If the market is random, it's equally like to go up or down. Suppose you plan to watch the market for 200 trading days. Based on the $500 change over 100 days, by how many dollars would you expect the market typically to change — either up or down — in the next 200 days? Choose the closest answer:
   250 500 750 1000 1250 1500 $\boxed{\text{NA-2}}$

Exercise 2

You take 100 individual steps in a random walk and cover altogether a net distance of 25 meters. (The net distance is the straight-line distance from starting point to ending point, as opposed to the length of the path you followed.)

If this particular walk were typical of such random walks, what is the length of each individual step?

$\boxed{\text{A}}$ $25/100 = 0.25$ meters

$\boxed{\text{B}}$ $25/\sqrt{100} = 2.50$ meters

$\boxed{\text{C}}$ $\sqrt{25}/100 = 0.05$ meters

$\boxed{\text{D}}$ $\sqrt{25}/\sqrt{100} = 0.50$ meters

Install R

To use the random-walk simulation software, you need to install the R statistics package. This is easy and doesn't change the settings on your computer. There are versions for Windows, Mac, and Linux.

▶ To install R, you download one file using your web browser. Links to the current version are given below.

  ▶ **Windows** To install R, execute the downloaded file. You will be prompted with several questions; you can accept all of the default settings. Link: `http://streaming.stat.iastate.edu/CRAN/bin/windows/base/release.htm`

  ▶ **Mac OS X** The file is a "disk image" and will appear in the Finder as such. To install R simply double-click on icon of the multi-package "R.mpkg" contained in the R-2.9.0.dmg disk image. Link: `http://streaming.stat.iastate.edu/CRAN/`

More generally you can access the latest version of the software through the Download/CRAN link on the main R web page. You will want the "base" distribution, which is distributed as a "binary" file appropriate for your operating system.
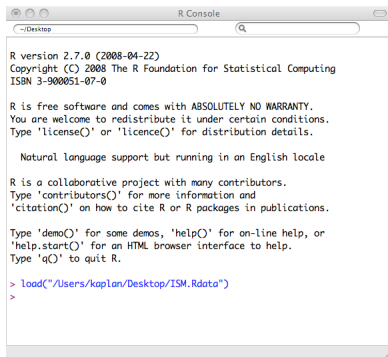
Install the software.

All of the used in this webinar is contained in a single file available on the web:

- ▶ Data/software file: `http://www.macalester.edu/~kaplan/ISM/RandomWalkWebinar/walk-simulation.Rdata`

Use your browser to download this file onto your computer.

Starting R

To start R, find the `walk-simulation.Rdata` file that you installed in the previous step. Double-click on it. This should start R and read in the software and data sets. The window should look something like this:

At this point, you can give the commands as indicated in the body of the talk.