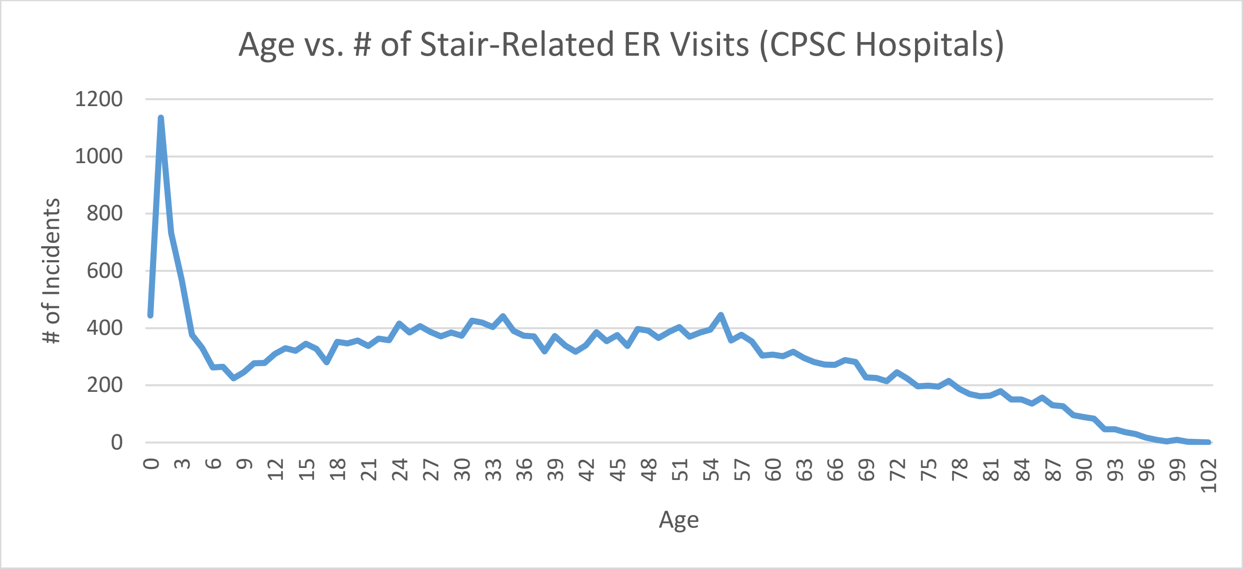
Introduction

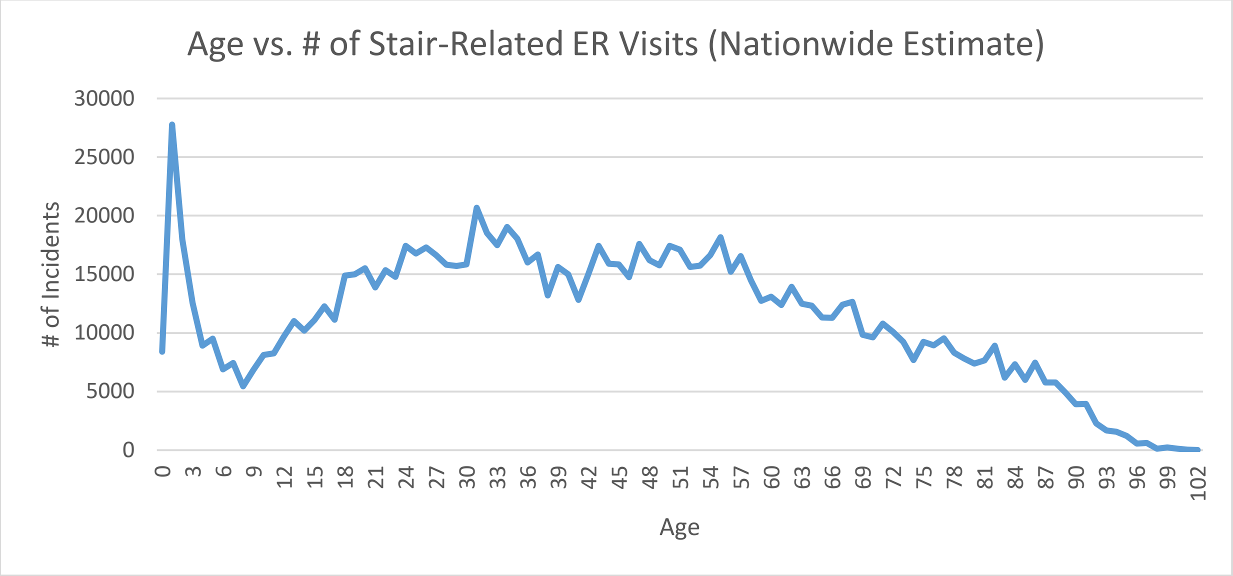
Everyone falls down the stairs once in a while. It's a fact of life that doesn't bother most, but there are those among us that desire to arm themselves with more concrete knowledge. Specifically, I think it would be useful for one to know his or her likelihood of injuring themselves on stairs or steps based on age. Thankfully, the US Consumer Product Safety Commission (CPSC), which tracks emergency room visits at 96 hospitals spread across the US[1], has us covered. Using their data from 2015, we can now delve into the statistics of every stair-related visit.

Data Preparation

First, I'll explain how I used their data to build my own. They record which consumer product caused each ER visit, so I was able to filter the results to show only those visits caused by "Stairs or steps (excluding pull-down and folding stairs)". While this obviously includes some instances of stubbed toes and the like, the vast majority of injuries are caused by tripping while climbing up or falling down the stairs. I used Microsoft Excel to count the number of instances of the age of the each ER visitor and create the dataset shown in the chart below.

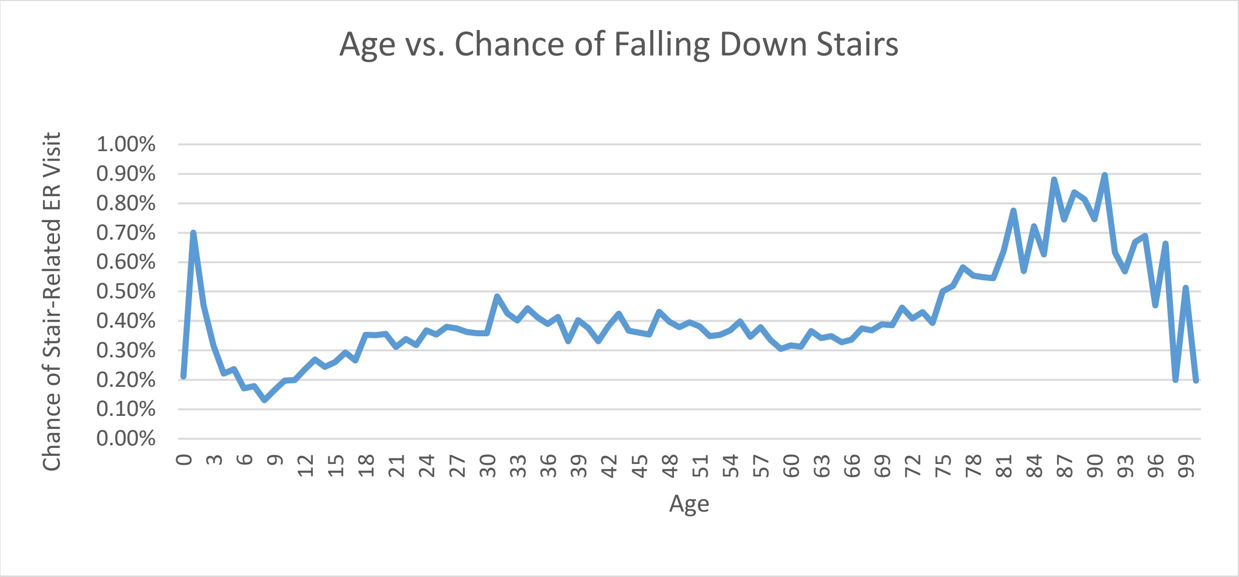


This data gives us a clear indication of what ages are the most likely to injure themselves on the stairs (I'm looking at you, kids who just learned how to walk), but the actual numbers are meaningless. We need to know how the subset of hospitals recorded in this study fit into the overall number of hospitals across the US in order to get the nationwide number of stair-related incidents. Lucky for us, the CPSC has a statistical "weight" field for each injury, which can be used for this exact purpose. By summing up the statistical weights for each injury of each age, I was able to produce this graph, which shows an estimate of nationwide stair-related injuries by age.



You'll notice, that though the shape of the curves are very similar, the right of this graph (ages 9+) is much higher relative to the children (ages 0-3) than it was in the previous graph. This is likely due to the fact that the CPSC monitors a large number of children's hospitals, so their data is heavily skewed towards kids.

By factoring in the weight, we've produced a very accurate representation of the nationwide stair-related ER visits. However, we're looking for probability data here- we need to combine what we have with the 2015 US Census data[2]. By taking the *Estimated Number of Nationwide Injuries* divided by the *Population* for each Age (as of July 1, 2015), I was able to compute the probability of going to the ER for falling down the stairs by age.



This is radically different than our previous findings. We can see that, much like diapers, falling down the stairs is most popular with children and the elderly. I hypothesized early on in my research that because so few people live to be over the age of 80, we would see a spike in injury probability after that, but I did not expect this significant of a change to the graph shape. Now, since we have the simplified dataset best suited to find our desired results, we can move on to analyzing the data.

Findings

All of the methods I will use here are described in great detail in *Numerical Analysis*[3] and in online resources such as Wikipedia. Therefore I will not go into detail on how they work- only the results of running my data through them.

While I originally wanted to compare function interpolation methods such as Lagrange and Newton's Divided Differences, I quickly realized that, for my rather large dataset, these methods produced horrendous functions that were hundreds of thousands of characters in length. Instead, I opted to use numerical integration methods to analyze my data for more applicable results. In particular, I'm going to compare the Sum of Rectangles Method, Trapezoidal rule and Simpson's Method for finding the probability that someone will have a stair-related ER visit in a certain period of their life.

Note: I initially planned to compare Romberg Integration as well, but after testing my code for it, I realized it was unfeasible on my computer. With such a large dataset, it slowed to a crawl and eventually ran out of memory entirely due to the calculations of $R(i, 1)$ where it finds the summation of $2^{i-1} - 1$ computations. This gets to be an obscene number of calculations once i gets large.

We will be attempting to estimate the following statement:

$$\int_a^b f(x)\,dx\,, \qquad h \geq 1$$

where a is the starting year, b is the ending year, $f(x)$ is the function returning the probability of a stair-related ER visit for each year and h is the distance (in years) between each evaluated data point. The three programs I have written all take the form $function\_name(f, a, b, h)$ with the same inputs, so it is very easy to evaluate the methods for every possible interval. Please check out the README file with the code if you'd like to try it out yourself.

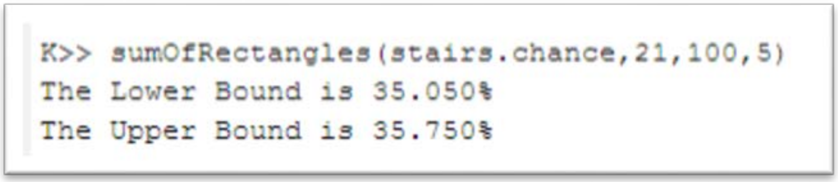
For the purpose of this report, my goal will be to find the probability that I will fall down the stairs between now and my death. I am currently 21 and plan to live to 100 at least, so the statement we will be attempting to estimate with each method is:

$$\int_{21}^{100} f(x)\,dx$$

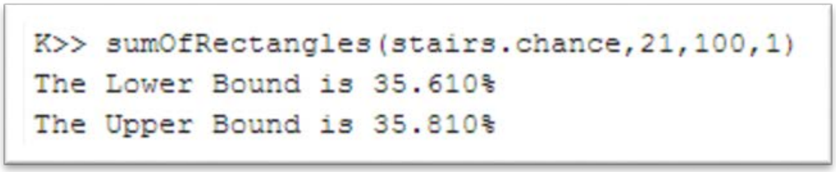
By summing up the values in Excel, we find that $\int_{21}^{100} f(x)\,dx = 35.61\%$. That is to say, that if I live to be 100, there's a 35.61% chance I will go to the ER for a stair-related injury within the remainder of my life. Now let's see which method can get closest to the actual result. In each case, I will evaluate the methods at both $h = 1$ and $h = 5$, in order to provide conclusive results on how well it works for various values of h. A higher h will almost always yield more accurate results since it then accounts for more values from our data ($h = 5$ will only look at every fifth data point).

Sum of Rectangles Method

This method gives the lower and upper bound estimates. Since $f(x)$ is not explicitly increasing or decreasing over the entire interval, I designed my function to determine which quantity was the lower bound and which was the upper bound by simply comparing their values. Using *sumOfRectangles.m*, we find:



```
K>> sumOfRectangles(stairs.chance,21,100,5)
The Lower Bound is 35.050%
The Upper Bound is 35.750%
```

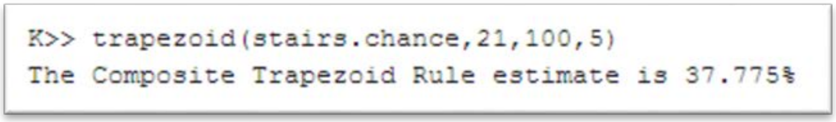


```
K>> sumOfRectangles(stairs.chance,21,100,1)
The Lower Bound is 35.610%
The Upper Bound is 35.810%
```

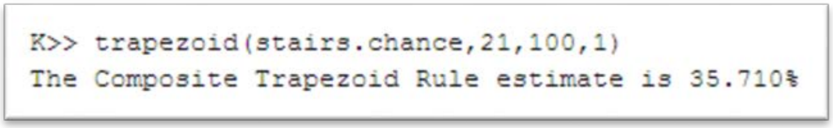
If we assume that the actual estimate is halfway between the Lower Bound and Upper Bound estimates, the estimates for $h = 5$ and $h = 1$ are 35.4% and 35.71%, respectively. The error for $h = 5$ is 0.21% and the error for $h = 1$ is 0.1%. This is impressive, since this method is normally inaccurate for less predictable datasets.

Composite Trapezoidal Rule

Next up is the Composite Trapezoidal Rule, which approximates the area under $f(x)$ as a trapezoid. Using *trapezoid.m*, we find:



```
K>> trapezoid(stairs.chance,21,100,5)
The Composite Trapezoid Rule estimate is 37.775%
```

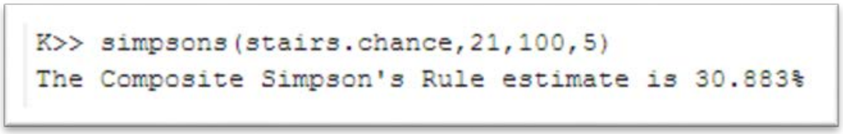


```
K>> trapezoid(stairs.chance,21,100,1)
The Composite Trapezoid Rule estimate is 35.710%
```

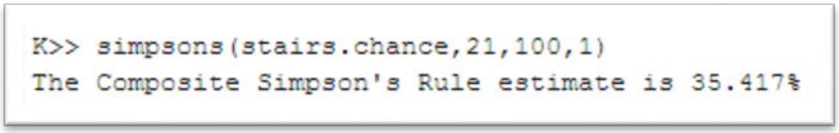
The error for $h = 5$ is 2.165% and the error for $h = 1$ is 0.1%. The best case error for this is the same as the Sum of Rectangles Method, which isn't surprising, given that these methods usually yield similarly accurate results.

Composite Simpson's Rule

Finally, I used the Composite Simpson's Rule, which is an alteration of the Simpson's Rule designed for larger datasets. Using *simpsons.m*, we find:



```
K>> simpsons(stairs.chance,21,100,5)
The Composite Simpson's Rule estimate is 30.883%
```



```
K>> simpsons(stairs.chance,21,100,1)
The Composite Simpson's Rule estimate is 35.417%
```

The error for $h = 5$ is 4.727% and the error for $h = 1$ is 0.193%. The error for this method seems unusually high, since Composite Simpson's is often more accurate than either of the previous methods.

Accuracy

CPSC notes that their national estimates have a coefficient of variation (CV) of 0.08. The CV, as described in the CPSC's documentation, "is the standard error of the estimate divided by the estimate", and "is a measure of sampling variability (errors that occur by chance because observations are made only on a population sample)" [4]. In terms of my findings, this means that the actual range of probability would be from $35.61\% * 0.92 = 32.7612\%$ to $35.61\% * 1.08 = 38.4588\%$.

Since the census data grouped all people 100 and older into the age group 100+, I did the same with all people aged 100, 101 and 102 from the CPSC data in order to combine the two datasets. Fortunately, there are very few people of that age alive in the country, so it doesn't affect the accuracy much.

Summary

In most cases, I would recommend Romberg Integration to solve this type of problem due to its incredible accuracy, but since this dataset was too large to use it effectively, I recommend the Sum of Rectangles Method. Its error for $h = 1$ was 0.1%, which is rather low and is the same error from the Trapezoidal Rule. However, the Sum of Rectangles Method consistently gets much smaller error for $1 \leq h \leq 10$, as we saw with its error of only 0.21% for $h = 5$ (in comparison to the Trapezoidal Rule's error of 2.165% for $h = 5$).

The error of each implemented method is higher than I had hoped for, but being accurate to two significant digits is still quite good. This dataset is an interesting case where approximation is not quite necessary, since it's so easy to calculate the integral directly using a summation of the data points. Still, this report shows that these numerical integration methods work remarkably well- even on large, real-world datasets. Thanks to this study, I hope you've learned to take extra precautions around stairs and maybe even learned something of value.

References

[1] "National Electronic Injury Surveillance System (NEISS)." National Electronic Injury Surveillance System (NEISS) | CPSC.gov. N.p., n.d. Web. 12 Dec. 2016.

[2] Bureau, US Census. "Census.gov." US Census Bureau. N.p., n.d. Web. 12 Dec. 2016.

[3] Burden, Richard L., and J. Douglas. Faires. Numerical Analysis. 9th ed. Pacific Grove, CA: Brooks/Cole Pub., 2010. Print.

[4] "NEISS Data Highlights - 2015." World Humanitarian Data and Trends World Humanitarian Data and Trends 2015 (2016): n. pag. Web. 12 Dec. 2016.