Data Science

```
<link>/</link>
<description>Recent content on Data Science</description>
<generator>Hugo -- gohugo.io</generator>
<language>en-us</language>
<lastBuildDate>Fri, 30 Dec 2016 21:49:57 -0700</lastBuildDate>

<atom:link href="/index.xml" rel="self" type="application/rss+xml" />


<item>
  <title>A Plain Markdown Post</title>
  <link>/2016/12/30/a-plain-markdown-post/</link>
  <pubDate>Fri, 30 Dec 2016 21:49:57 -0700</pubDate>

  <guid>/2016/12/30/a-plain-markdown-post/</guid>
  <description>This is a post written in plain Markdown (*.md) instead of R Markdown (*.Rmd). The major
```

You cannot run any R code in a plain Markdown document, whereas in an R Markdown document, you can embed R code chunks (“‘{r}); A plain Markdown post is rendered through Blackfriday, and an R Markdown document is compiled by rmarkdown and Pandoc. There are many differences in syntax between Blackfriday’s Markdown and Pandoc’s Markdown.

```
<item>
  <title>About</title>
  <link>/about/</link>
  <pubDate>Thu, 05 May 2016 21:48:51 -0700</pubDate>

  <guid>/about/</guid>
  <description>This is a &amp;ldquo;hello world&amp;rdquo; example website for the blogdown package. Th
</item>

<item>
  <title>Hello Test R Markdown</title>
  <link>/2015/07/23/hello-r-markdown/</link>
  <pubDate>Thu, 23 Jul 2015 21:13:14 -0500</pubDate>

  <guid>/2015/07/23/The magic of the central limit theorem:/</guid>
  <description>Sampling, sampling, sampling...
```

As scientists we aim to understand the world around us, not just our immediate environments. In most cases, we don’t have access to **populations**, for one, because they are… large. For example, if you’re studying expectant mothers, it is virtually impossible to collect data from all of the expectant mothers from around the world. Therefore, we make do with random and representative samples of the population to make generalizations – that is, statements – about the population as a whole.

The **central limit theorem (CLT)** states that the larger the sample size collected, the closer the distribution of the sample means will resemble a normal distribution (bell curve), regardless of the population’s distribution. If you were asleep during your stats class or need a refresher, Khan Academy gives a good introduction to CLT.

It’s useful to re-read the statement above, because we’re not saying that we’re assuming that the observations (e.g., one sample of with 100) originate from a normal distribution. We’re saying that the distribution

of sample means (based on taking the mean of many separate samples that originate from the "parent" distribution) will be normally distributed. That is, provided your sample size is "large enough." The theorem works "in the limit", as mathematicians say, but a general rule is that the sample size should be at least 30 for the CLT to hold for almost any data distribution. And, in practice, it will work on smaller samples if they originate from a population that actually has a normal distribution.

Let's refresh ourselves with a few terms. ***Variance*** refers to the amount of variability, or how spread the data are from the mean. The population variance is symbolized as ***sigma squared***, or σ^2. Because variance is a squared term, we tend to look at the square root of variance, or ***standard deviation***, and the population standard deviation is symbolized as σ, or ***sigma***.

We know that as the size of the sample increases, the closer the sampling distribution of the sample mean will resemble a bell curve with a mean that approaches the population mean, μ. How about the standard deviation of the distribution of the sample means? As the sample size increases, the CLT says that the standard deviation will approach $\frac{\sigma}{\sqrt{n}}$. Again, this results holds despite the shape of the population distribution.

A good source of intuitive discussion on the central limit theorem is Mordkoff, J.T. (2016) The Assumption(s) of Normality.

Post originally by Kelly Morrow and edited by L. Pessoa

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com. You can embed an R code chunk like this: summary(cars) ## speed dist ## Min. : 4.0 Min. : 2.00 ## 1st Qu.:12.0 1st Qu.: 26.00 ## Median :15.0 Median : 36.00 ## Mean :15.4 Mean : 42.98 ## 3rd Qu.

```
<item>
  <title>Lorem Ipsum</title>
  <link>/2015/01/01/lorem-ipsum/</link>
  <pubDate>Thu, 01 Jan 2015 13:09:13 -0600</pubDate>

  <guid>/2015/01/01/lorem-ipsum/</guid>
  <description>Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididun
</item>
```