Introducing Probability

Matt Lawrence

October 9, 2019

Transitioning to Probabilty

Today we'll move from multivariate descriptions to probability, and we'll switch datasets to look at some distributions across counties rather than commuting zones. Load the counties dataset and the usual packages to get started.

BACK TO SLIDES; WE'LL RETURN TO R SOON

Introducing Probability And Z-Scores

A z-score or standardized value is a value's distance from the mean in standard deviations. It is calculated as: $z = \frac{x-\mu}{\sigma}$. In words, the z-score is the difference between the observed value and the sample mean divided by the standard deviation.

After confirming that `teen_birthrate` is approximately normally distributed, let's make a new variable with the standardized values of `teen_birthrate`:

```
counties <- mutate(counties,
     teen_birthrate_z = (teen_birthrate - mean(teen_birthrate)) /
                        sd(teen_birthrate))
```

Z-scores should be normally distributed with a mean of 0 and a standard deviation of 1. Were we successful?

REPLACE THIS LINE WITH YOUR CODE

```
summary(counties$teen_birthrate_z)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.41776 -0.78049 -0.02083  0.00000  0.72741  3.46247
```

```
sd(counties$teen_birthrate_z)
```

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

What is the z-score for Addison County, Vermont's birth rate? (Note that in the counties dataset you have to use the state's full name, not its abbreviation.)

REPLACE THIS LINE WITH YOUR CODE

```
counties$teen_birthrate_z[counties$county=="Addison" &
                               counties$state=="Vermont"]
```

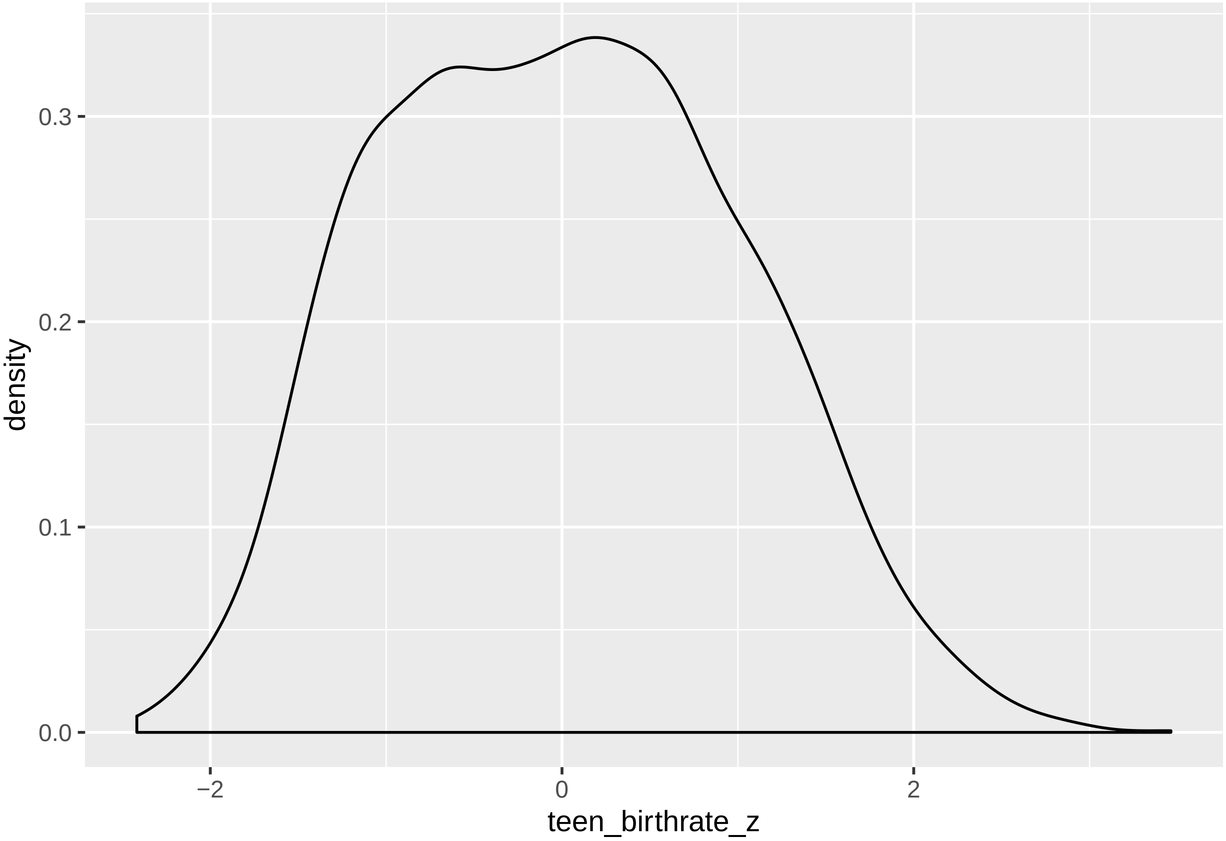
```
## [1] -1.75531
```

When we plot standardized values that are approximately normal, we now know a lot about the proportion of observations falling along different points of the distribution. To see how, make a density plot showing the distribution of the standardized teen birthrates.

REPLACE THIS LINE WITH YOUR CODE

```
birthrate_z_plot <- ggplot(counties, aes(x = teen_birthrate_z))

birthrate_z_plot + geom_density()
```



To find the probability of getting any z-score, use `dnorm()`. Think about this value as the y axis intersection with the density curve for any specific value on the x axis. For example, the probability that a randomly pulled county would have the same birthrate as Addison County, Vermont is:

```
dnorm(-1.75531)
```

```
## [1] 0.0854781
```

Probabilities of specific values are more helpful for descriptives than for inference. Moving forward, what is more helpful is knowing the probability of randomly pulling a value that is greater than or less than an observed value. In other words, we want to add up the probabilities of pulling any value less than Addison County's value.

We get that summed probability by thinking not of the density but of the *cumulative density*. The cumulative density is also the percentile.

If you have the z-value and want the percentile associated with it, use `pnorm()`. For Addison County:

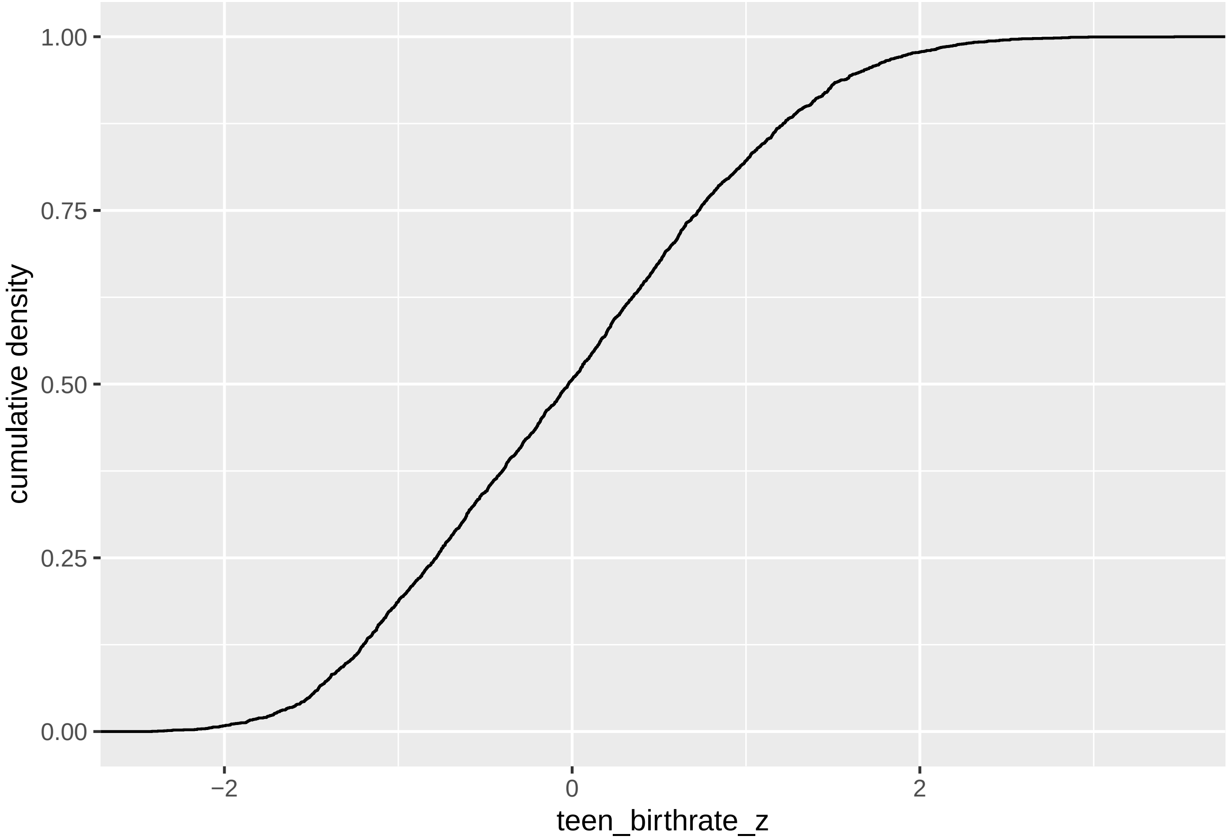
```
pnorm(-1.75531)
```

```
## [1] 0.03960315
```

The pnorm() function will give you the proportion of the distribution to the left of the z score. So about 4% of counties have a teen birth rate less than the teen birth rate of Addison County. Alternative, Addison County is at the 4th percentile in the national distribution of teen birth rates.

We can also plot the entire *empirical cumulative density function*:

```
birthrate_z_plot + geom_step(stat = "ecdf") +
     labs(y = "cumulative density")
```



Since normal distributions are symmetrical, the probability of getting a value that is more extreme of a negative z-score is the same as the probability of getting a value that is more extreme of that positive z-score.

What is the z-score of Hamilton County, Florida?

REPLACE THIS LINE WITH YOUR CODE

```
counties$teen_birthrate_z[counties$county=="Hamilton" &
                              counties$state=="Florida"]
```

```
## [1] 1.751348
```

What is the probability of another county having a birth rate that is higher than that of Hamilton County?

REPLACE THIS LINE WITH YOUR CODE

```
1 - pnorm(1.751348)
```

```
## [1] 0.03994299
```

What proportion of counties have birth rates between those of Addison and Hamilton?

REPLACE THIS LINE WITH YOUR CODE

```
pnorm(1.751348) - pnorm(-1.75531)
```

```
## [1] 0.9204539
```

Exercise

1. Choose a county and find its z-score for the average commute time (variable = `commute`).
2. What is the probability that a randomly pulled county will have an average commute time greater than the county you chose?
3. What is the probability that a randomly pulled county will have an average commute time less than the county you chose?

Exercise Follow Up

Start by calculating z-scores for all values of the `commute` variable:

```
counties <- mutate(counties, commute_z =
     (commute - mean(commute)) / sd(commute))
```

Find the z-score for a specific county:

```
counties$commute_z[counties$county=="Baltimore" & counties$state=="Maryland"]
```

```
## [1] 1.976928
```

Find the probability of a more extreme value than that z-score. In this case, the z-score is negative so we just use pnorm():

```
pnorm(1.961353)
```

```
## [1] 0.9750811
```

For the probability of z-score greater than this one, use:

```
1 - pnorm(1.961353)
```

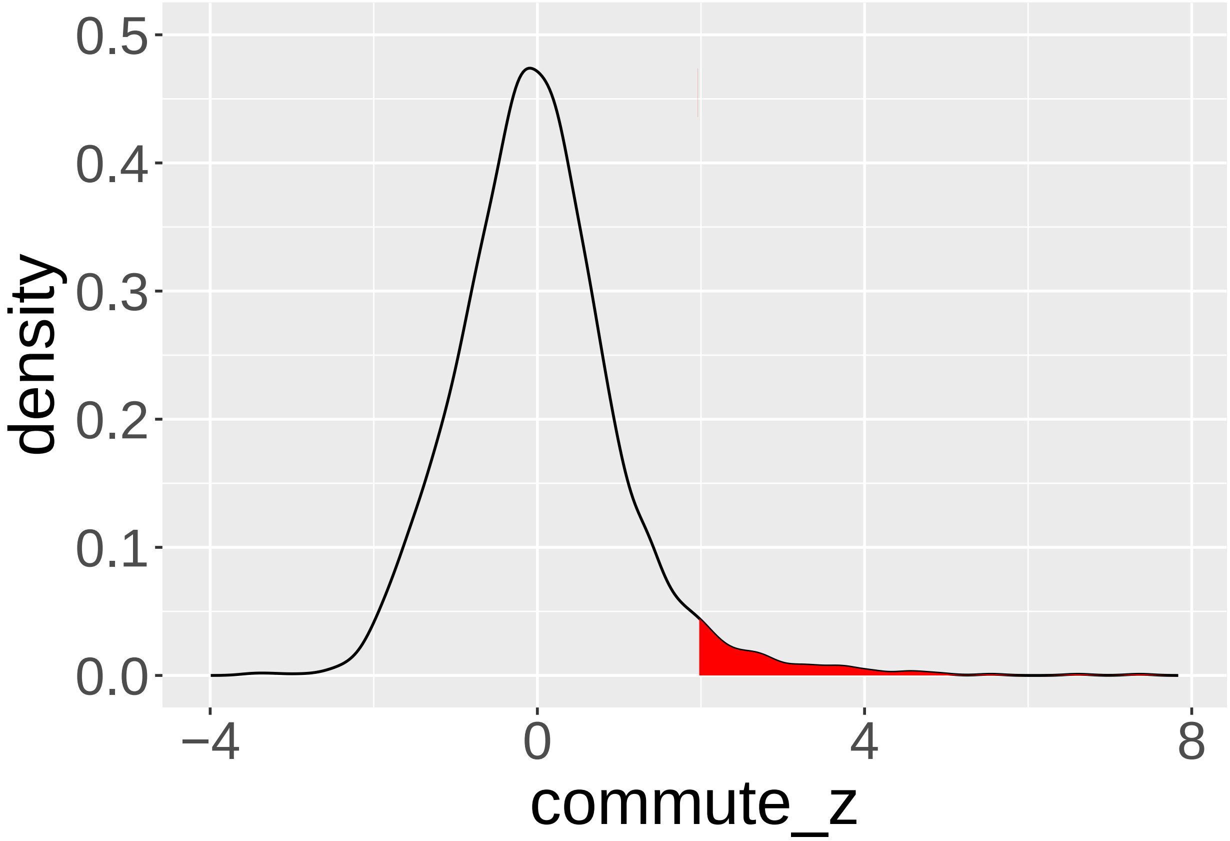
```
## [1] 0.02491893
```

Plot it…

```
commute_z_density <- with(density(counties$commute_z), data.frame(x , y))

commute_cumulative_density_plot <- ggplot(data = commute_z_density,
                                          mapping = aes(x = x, y = y)) +
     geom_line() +
     geom_area(mapping = aes(x = ifelse(x >= 1.961353, x, 1.961353)),
               fill = "red") +
     ylim(c(0, .5)) +
     theme(axis.title = element_text(size = 24),
           axis.text = element_text(size = 20)) +
     labs(x = "commute_z", y = "density")
```

```
commute_cumulative_density_plot
```


density
commute_z