Normal distributions with R

9/8/2017

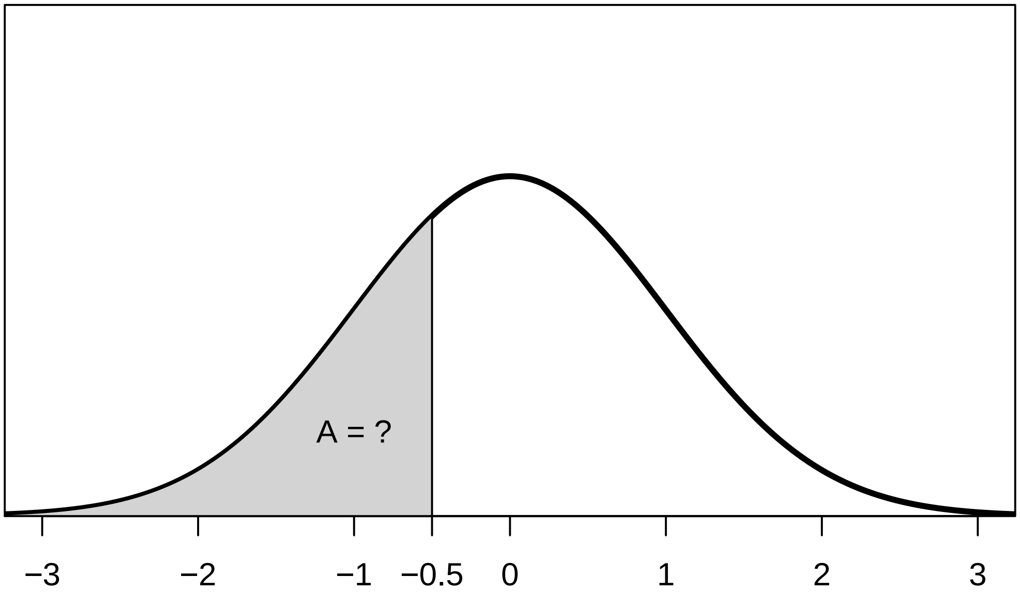

Figure 1: The standard normal

A fundamental question in statistics is - how do we compute and interpret area computations like the one shown above? In this little lab, we'll learn exactly how to do so with R.

The `pnorm` command

Suppose we want to compute the shaded area shown in figure 1 above - that is, the area under the standard normal curve and to the left of the line of the value $Z = -0.5$. It's a simple matter of executing the following command:

```
pnorm(-0.5)
```

Give it a try! Of course, the result should agree with the result obtained from a table. A portion of the our normal distribution table is shown in table 1 below with the relevant row and column highlighted.

Table 1: A few relevant Z values

	0.04	0.03	0.02	0.01	0.00	Z
⋮ ⋮	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
⋮ ⋮	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
⋮ ⋮	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5
⋮ ⋮	0.3300	0.3336	0.3372	0.3409	0.3446	-0.4
⋮ ⋮	0.3669	0.3707	0.3745	0.3783	0.3821	-0.3

A basic example

Let's motivate our further exploration with a simple example: *Suppose the height of third graders at Isaac Dickson is normally distributed with a mean of 50.2 inches and a standard deviation of 2.66 inches. If there are 87 third graders, about how many do we expect to be taller than 54 inches?*

Note that this problem is phrased very much like a quiz or exam problem and we can approach it as such. However, when we get to where we need to compute the final answer, we'll use the R `pnorm` command, rather than a table.

Solution (through the standard normal): We first compute the z-score as usual:

$$Z = \frac{54 - 50.2}{2.66} = 1.428571.$$

Of course, that computation is super easy to type into R as `(54-50.2)/2.66`. The point behind the z-score is that we can now compute the proportion of the kids taller than 54 inches as the area under the standard normal curve and to the *right* of $z = 1.428571$. This is the *white* area under the curve in the top plot of figure 2 shown below. Note that the `pnorm` command computes the so-called *cumulative* probability. Thus, it can compute the shaded gray region shown in the figure. To compute the white region, we'll need to subtract this from one. Taking this into account, our answer can be computed as:

```r
(1 - pnorm(1.428571))*87
```

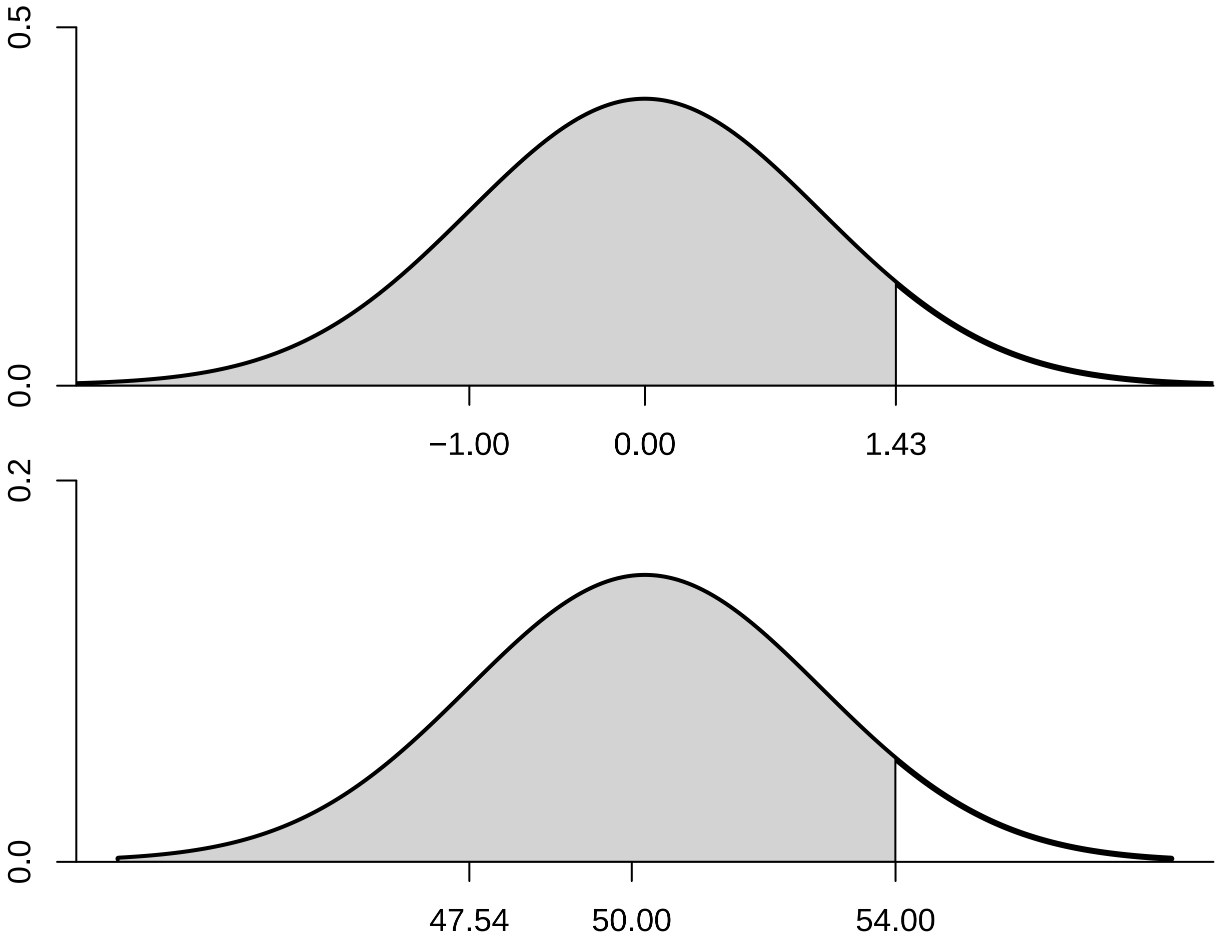

Figure 2: Two normal curves for our example problem

Solution (through another normal): Using R, we can actually skip the standardization step altogether. The reason is that the `pnorm` command accepts more arguments allowing us to specify the mean and standard deviation that we'd like to use. Thus, another command that computes the answer is:

```r
(1 - pnorm(54, 50.2, 2.66))*87
```

Give it a try! The result should agree with your previous computation.

Problem 1

Suppose that cholesterol levels of an adult can be described by a normal model with a mean of 190 mg/dL and a standard deviation of 20.

- What percent of adults do you expect to have cholesterol levels over 200 mg/dL?
- What percent of adults do you expect to have cholesterol levels between 170 mg/dL and 200 mg/dL?

The `qnorm` command

The `qnorm` command is essentially an inverse of the `pnorm` command - that is, given a quantile `q`, `qnorm(q)` returns the z-value at that quantile. For example, can you interpret the following result?

```r
qnorm(0.5)
```

You can also specify the mean `m` and standard deviation `s`. For example, we can compute the theoretical interquartile range for our Isaac Dickson heights (where the mean and standard deviation were 50.2 and 2.66) as follows:

```r
qnorm(0.75, 50.2, 2.66) - qnorm(0.25, 50.2, 2.66)
```

Problem 2

Find the theoretical interquartile range for the cholesterol levels of problem 1.

Working with data

Let's grab some real world data and use `qqnorm` to generate a normal probability plot to see if it looks normally distributed.

```r
cdc = read.csv("https://marksmath.org/data/cdc.csv")
heights = subset(cdc, gender == 'm')$height
qqnorm(heights)
```

Looks good! If we want to model this data with a normal distribution, we'll need to know the mean and standard deviation.

```r
m = mean(heights)
s = sd(heights)
c(m,s)
```

We can now ask the question - what percentage of men are taller than 6 feet (according to the model)? We can answer that question with the following computation.

```r
1-pnorm(72, m, s)
```

We can compare this to the actual value for this data set.

```
100 - quantile(heights, 0.72)
```

Problem 3

Find an appropriate model for the heights of women in the CDC dataset and use it to estimate the proportion of women who are between 5’ 2’‘and 5’6” tall.