Normal Probability Distribution

CJ 702: Advanced Criminal Justice Statistics

Thomas Bryan Smith*

February 03, 2025

Contents

1 **Load the USArrest data** 1

2 **Viewing the *frequency distribution* for the Assault variable** 2

3 **The normal *probability distribution*** 2

4 **The problem of non-normal observed distributions** 4

1 Load the USArrest data

First, let's load in the built-in USArrest data, and take a look at the first 10 observations (US states).

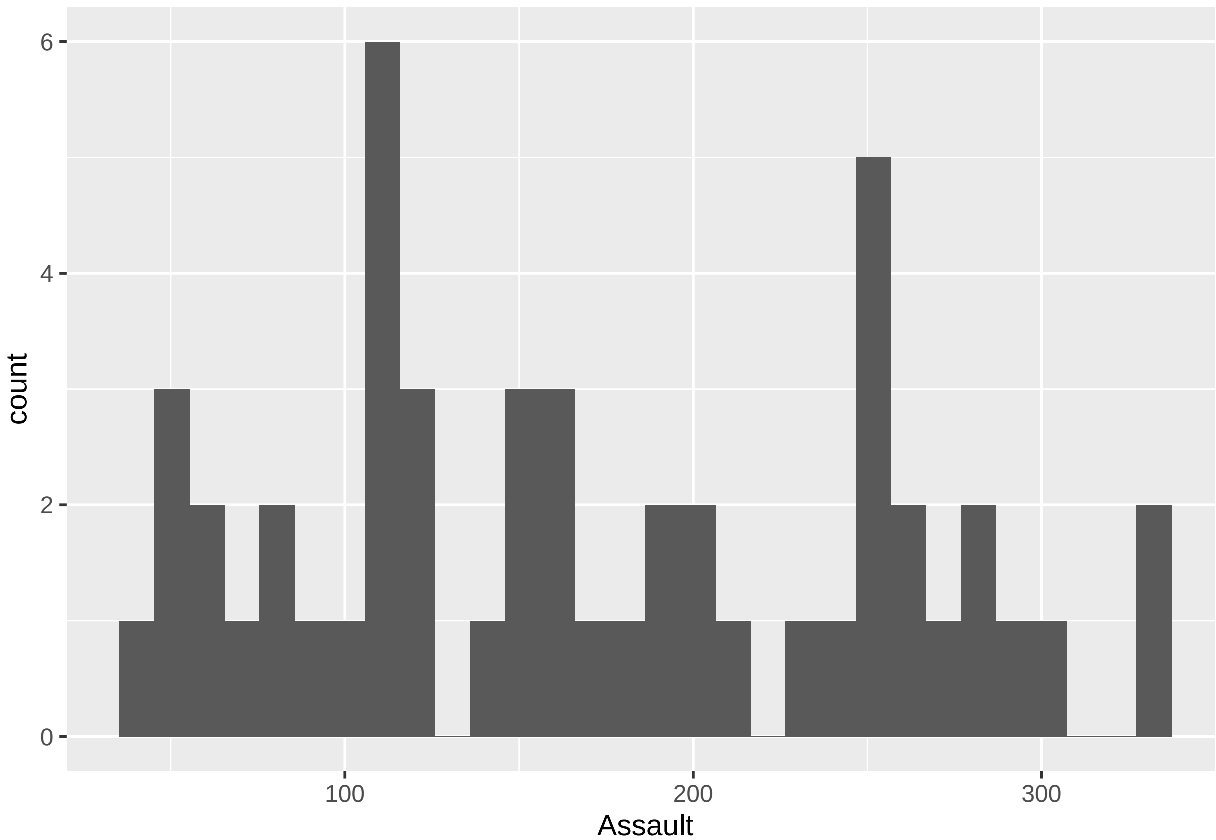
```
data(USArrests)

head(USArrests, 10)
```

```
##             Murder Assault UrbanPop Rape
## Alabama       13.2     236       58 21.2
## Alaska        10.0     263       48 44.5
## Arizona        8.1     294       80 31.0
## Arkansas       8.8     190       50 19.5
## California     9.0     276       91 40.6
## Colorado       7.9     204       78 38.7
## Connecticut    3.3     110       77 11.1
## Delaware       5.9     238       72 15.8
## Florida       15.4     335       80 31.9
## Georgia       17.4     211       60 25.8
```

*University of Mississippi, tbsmit10@olemiss.edu

2 Viewing the *frequency distribution* for the Assault variable

Now, let's visualize the *frequency distribution* for the Assault variable. You know how to do this, we're just going to present it as a histogram.



3 The normal *probability distribution*

However, the *observed frequency distribution* is not the same as the *probability distribution*.

The normal probability distribution (*assumed* by Ordinary Least Squares Regression) is typically indicated with the following expression:

$$N(\mu, \sigma^2)$$

This expression is simply saying that you are working with a normal distribution with a given mean, μ, and standard deviation, σ.

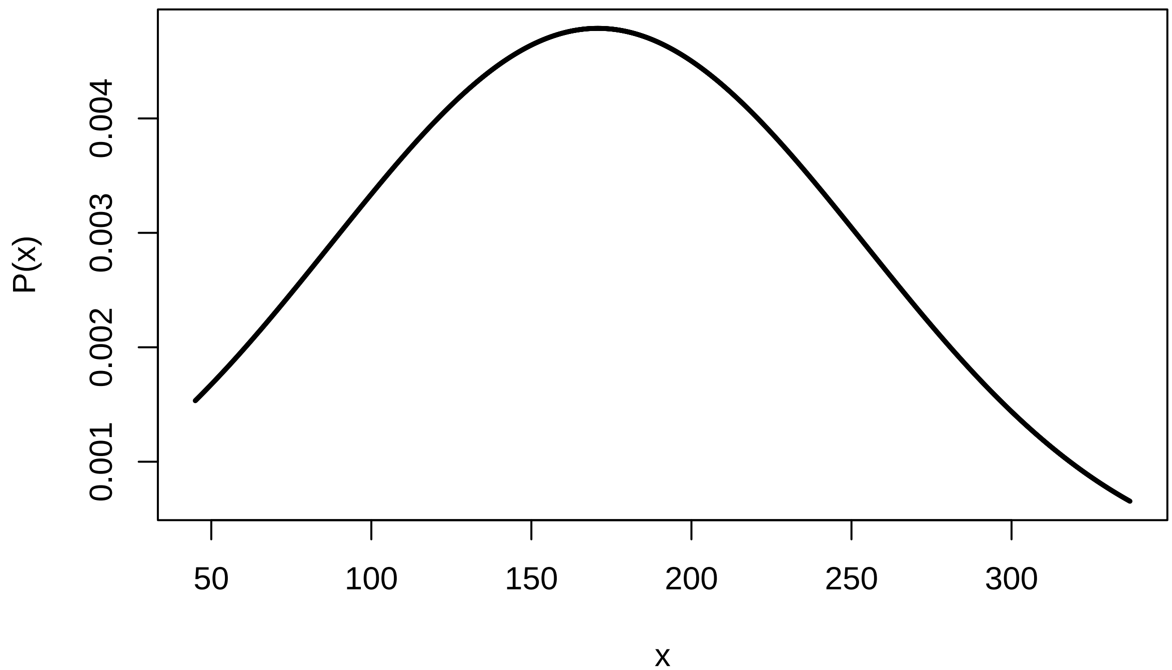
We could visualize this *theoretical* (reads: *assumed*) probability distribution using the following probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

As ever, μ is the mean of your variable, σ is the standard deviation of your variable. As for the letters you may not recognize: π is 3.14159..., and e is Euler's constant, or 2.71828... Finally, x is any possible value that can be assumed by your independent variable. Below, I am going to treat this as a sequence of numbers that starts at the minimum assault rate, 45, ends at the maximum assault rate, $-\infty$, and increases in intervals of 0.01.

As seen below, this equation generates a normal distribution that represents all theoretically possible values of a given *normally distributed continuous variable*. Unlike your *frequency distribution* (i.e., the histogram above), which is beholden to the observations that *actually exist*, this *probability distribution* can *theoretically* take on any value for variable x (here, the assault rate per 100,000).

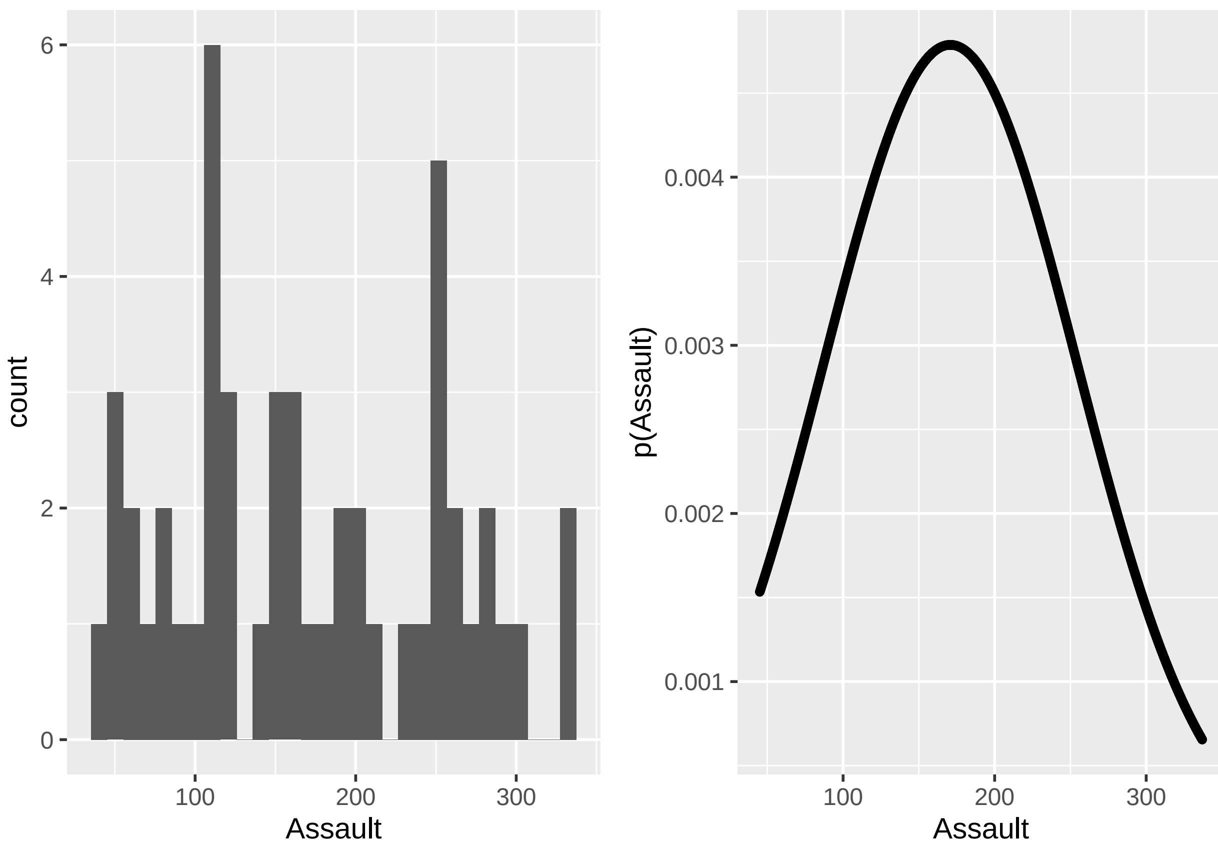
```r
# Find the mean:
mu <- USArrests$Assault %>% mean(na.rm = TRUE)

# Find the standard deviation:
sigma <- USArrests$Assault %>% sd(na.rm = TRUE)

# Generate a sequence of all possible values for x:
min <- USArrests$Assault %>% min(na.rm = TRUE)
max <- USArrests$Assault %>% max(na.rm = TRUE)

x <- seq(min, max, by = 0.1)

# Insert these values into the normal probability density function:
npd <- (1 / (sigma * sqrt(2 * pi))) * exp(1)^(-(x - mu)^2 / (2 * sigma^2))

# Plot the probability distribution:
plot(x, npd,
     ylab = "P(x)",
     cex = 0.2)
```



4 The problem of non-normal observed distributions

This probably all sounds great, but there is a hitch. Let's look at our frequency distribution and our probability distribution side-by-side:

```
plot1 <- ggplot(USArrests, aes(x = Assault)) +
  geom_histogram(bins = 30)

plot2 <- ggplot(data.frame(x, npd), aes(x = x, y = npd)) +
  geom_point(size = 1) +
  ylab("p(Assault)") +
  xlab("Assault")

grid.arrange(plot1, plot2, ncol = 2)
```



Ask yourself: does the observed distribution on the left look much like the probability distribution we have generated on the right? No, not particularly.

However, when your ordinary least squares regression *assumes* that your Assault variable is normally distributed, this exact probability distribution is assumed to be 'true'. So, if your observed frequency distribution looks *nothing* like the probability distribution, you might want to rethink the assumption that your variable is normally distributed!