Probability Distributions

2020-01-16

Preparation

In this set of notes, you will learn about common probability distributions. We will not be using a specific dataset in these notes.

```r
# Load libraries
library(broom)
library(corrr)
library(educate) #Need version 0.1.0.1
library(patchwork)
library(tidyverse)
```

Normal Distribution

The probability distribution of a normal distribution is mathematically defined as:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for $-\infty \leq x \leq \infty$. Consider a normal distribution with a mean (μ) of 50, and a standard deviation (σ) of 10. We can compute the probability density ($p(x)$) for a particular x value by using this equation. For example, the probability density for $x = 65$ can be found using,

$$p(65) = \frac{1}{10\sqrt{2\pi}} \exp\left[-\frac{(65-50)^2}{2 \times 10^2}\right] = 0.01295176$$

Using R, we can carry out the computation,

```r
(1 / (10 * sqrt(2 * pi))) * exp(-(225) / 200)
```

```
[1] 0.01295176
```

There is also a more direct way to compute this using the `dnorm()` function. This function computes the density of `x` from a normal distribution with a specified `mean` and `sd`.

```r
dnorm(x = 65, mean = 50, sd = 10)
```

```
[1] 0.01295176
```

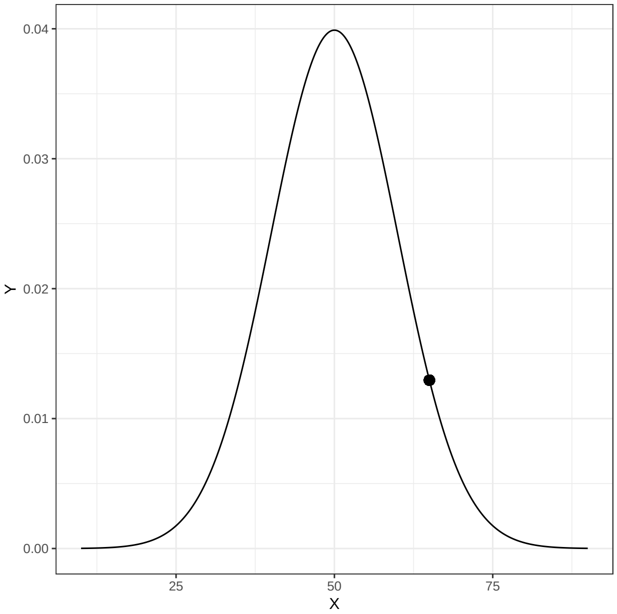
If we compute the density for several x values and plot them, we get the familiar normal shape; the graphical depiction of the mathematical equation.

```
# Create dataset
fig_01 = data.frame(
  X = seq(from = 10, to = 90, by = 0.01)
  ) %>%
  rowwise() %>%
  mutate(
    Y = dnorm(x = X, mean = 50, sd = 10)
    )

# Create plot
ggplot(data = fig_01, aes(x = X, y = Y)) +
  geom_line() +
  theme_bw() +
  geom_point(x = 65, y = 0.01295176, size = 3)
```

Figure 1

Plot of the Probability Density Function (PDF) for a $\mathcal{N}(50, 10)$ Distribution



Note. The density value for $x = 65$, $p(65) = 0.01295176$, is also displayed on the PDF.

Other Useful R Functions for Working with Probability Distributions

There are four primary functions for working with the normal probability distribution:

- `dnorm()` : To compute the probability density (point on the curve)
- `pnorm()` : To compute the probability (area under the PDF)
- `qnorm()` : To compute the x value given a particular probability
- `rnorm()` : To draw a random observation from the distribution

Each of these requires the arguments `mean=` and `sd=`. Let's look at some of them in use.

Finding Cumulative Probability

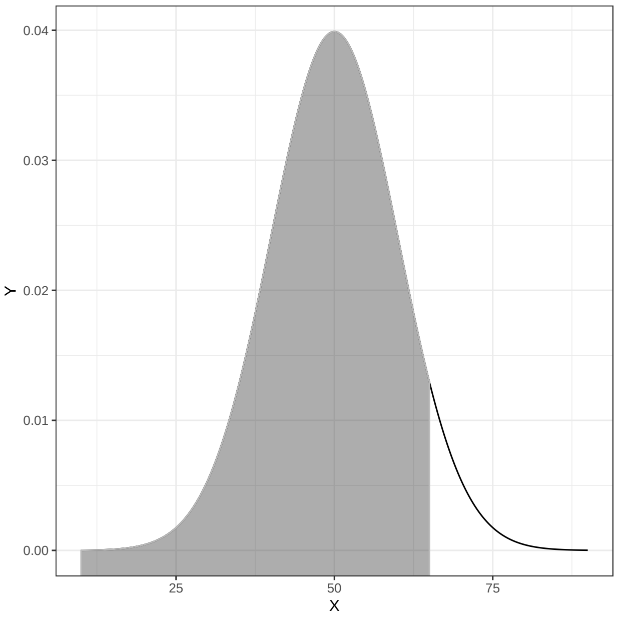
The function `pnorm()` gives the probability x is less than or equal to some quantile value in the distribution; the cumulative probability. For example, to find the probability that $x \leq 65$ we would use,

```r
pnorm(q = 65, mean = 50, sd = 10)
```

```
[1] 0.9331928
```

This is akin to finding the proportion of the area under the normal PDF that is to the left of 65.

Figure 2
Plot of the Cumulative Probability Density for X of 65



Note. The distribution shown is $\mathcal{N}(50, 10)$.

For the mathematically inclined, the grey-shaded area is expressed as an integral

$$\int_{-\infty}^{65} p(x)dx$$

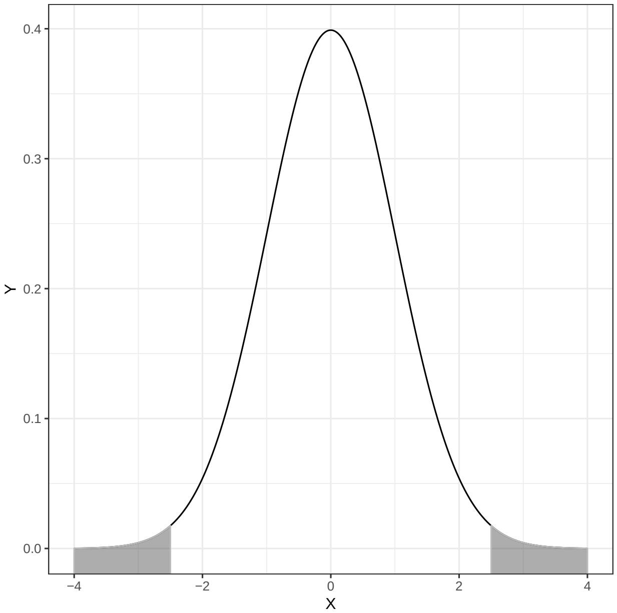
where $p(x)$ is the PDF for the normal distribution.

Cumulative Density and p-Value

This type of computation is used most commonly to find a p-value. The p-value is just the area under the distribution (curve) that is AT LEAST as extreme as some observed value. Consider a hypothesis test of whether a population parameter is equal to 0. Also consider that we observed a statistic (that has been standardized) of $z = 2.5$. Then, the p-value can be graphically displayed in the standard normal distribution as follows:

Figure 3

Cumulative Density Representing the p-Value for a Two-Tailed Test Using an Observed z-Value of 2.5



Note. Plot of the probability density function (PDF) for the standard normal distribution ($M = 0$, $SD = 1$). The cumulative density representing the p-value for a two-tailed test evaluating whether $\mu = 0$ using an observed z-value of 2.5 is also displayed.

In most hypothesis tests, we test whether the parameter IS EQUAL to 0. Thus the values in the standard normal distribution more extreme than 2.5 encompass evidence against the hypothesis; those values greater than 2.5 and also those values less than -2.5. (This is akin to testing a fair coin when both 8 heads OR 8 tails would provide evidence against fairness …we have to consider evidence in both directions).

To compute this we use `pnorm()`. Remember, it computes the proportion of the area under the curve TO THE LEFT of a particular value. Here we will compute the are to the left of -2.5 and then double it to produce the actual p-value.

```
2 * pnorm(q = -2.5, mean = 0, sd = 1)
```

```
[1] 0.01241933
```

Finding Quantiles

The `qnorm()` function is essentially the inverse of the `pnorm()` function. The `p` functions find the cumulative probability GIVEN a particular quantile. The `q` functions find the quantile GIVEN a cumulative probability. For example, in the normal distribution we defined earlier, half of the area is below the quantile value of 50 (the mean).

```
qnorm(p = 0.5, mean = 50, sd = 10)
```

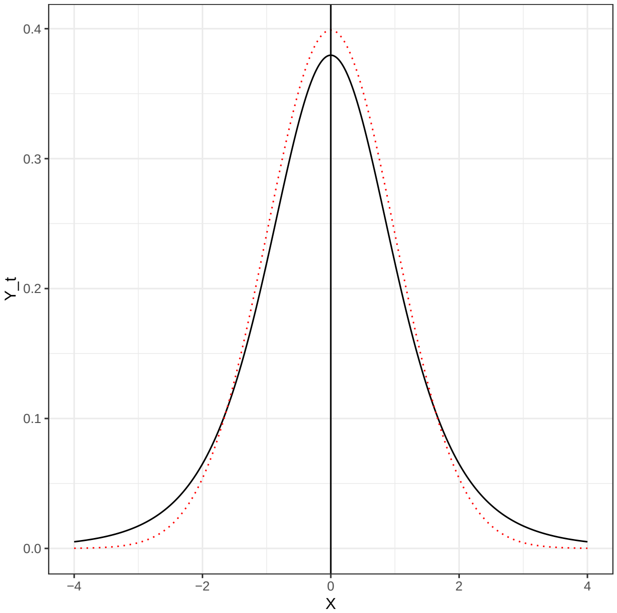
```
[1] 50
```

Student's t-Distribution

Student's t-distribution looks like a standard normal distribution. In the figure below, Student's t-distribution is depicted with a solid, black line and the standard normal distribution ($M = 0$, $SD = 1$) is depicted with a dotted, red line.

Figure 4

Plot of the Probability Density Function (PDF) for the Standard Normal Distribution (Dotted, Red Line) and Student's t(5) Distribution (Solid, Black Line)



Both the standard normal distribution and Student's t-distribution have a mean (expected value) of 0. The standard deviation for Student's t-distribution is larger than the standard deviation for the standard normal distribution ($SD > 1$). You can see this in the distribution because the tails in Student's t-distribution are fatter (more error) than the standard normal distribution.

In practice, we often use Student's t-distribution rather than the standard normal distribution when we are using sample data to estimate the population. This estimation increases the error and thus is typically modeled using Student's t-distribution.

Student's t-distribution constitutes a family of distributions—not just a single distribution. The specific shape (and thus probability density) is defined by the *degrees of freedom*; *df*. The plot below shows the standard normal distribution (purple) and four t-distributions with varying df-values.

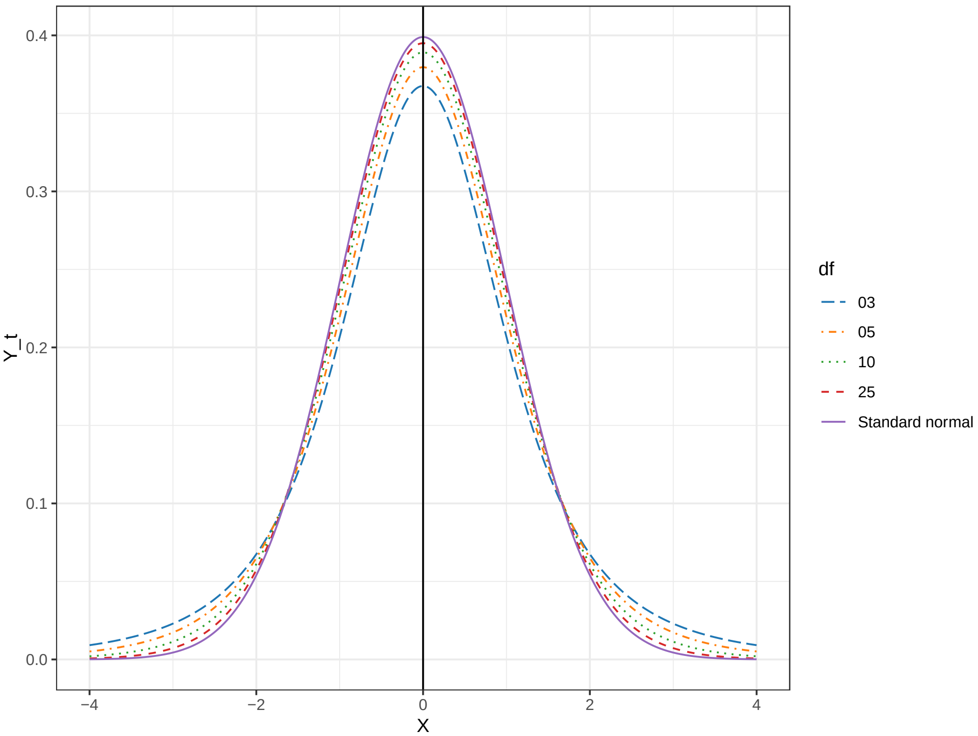
Table 1

Means and Standard Deviations for Four t-Distributions and the Standard Normal Distribution

df	M	SD
03	0	2.00
05	0	1.50
10	0	1.22
25	0	1.08
z	0	1.00

Figure 5

Plot of Several t-Distributions with Differing Degrees of Freedom



If we compare the means and SDs for these distributions, we find that the mean for all the t-distributions is 0, same as the standard normal distribution. All t-distributions are unimodal and symmetric around zero. The SD for every t-distribution is higher than the SD for the standard normal distribution. Student t-distributions with higher *df* values have less variation. It turns out that the standard normal distribution is a t-distribution with ∞ *df*. For the formula for the SD in a t-distribution, see Fox (2009).

There are four primary functions for working with Student's t-distribution:

- `dt()` : To compute the probability density (point on the curve)
- `pt()` : To compute the probability (area under the PDF)
- `qt()` : To compute the x value given a particular probability
- `rt()` : To draw a random observation from the distribution

Each of these requires the arguments `df=`. Let's look at some of them in use.

Comparing Probability Densities

How do the probability densities for a value of X compare across these distributions? Let's examine the X value of 2.

```
# Standard normal distribution
pnorm(q = 2, mean = 0, sd = 1)
```

```
[1] 0.9772499
```

```
# t-distribution with 3 df
pt(q = 2, df = 3)
```

```
[1] 0.930337
```

```
# t-distribution with 5 df
pt(q = 2, df = 5)
```

```
[1] 0.9490303
```

```
# t-distribution with 10 df
pt(q = 2, df = 10)
```

```
[1] 0.963306
```

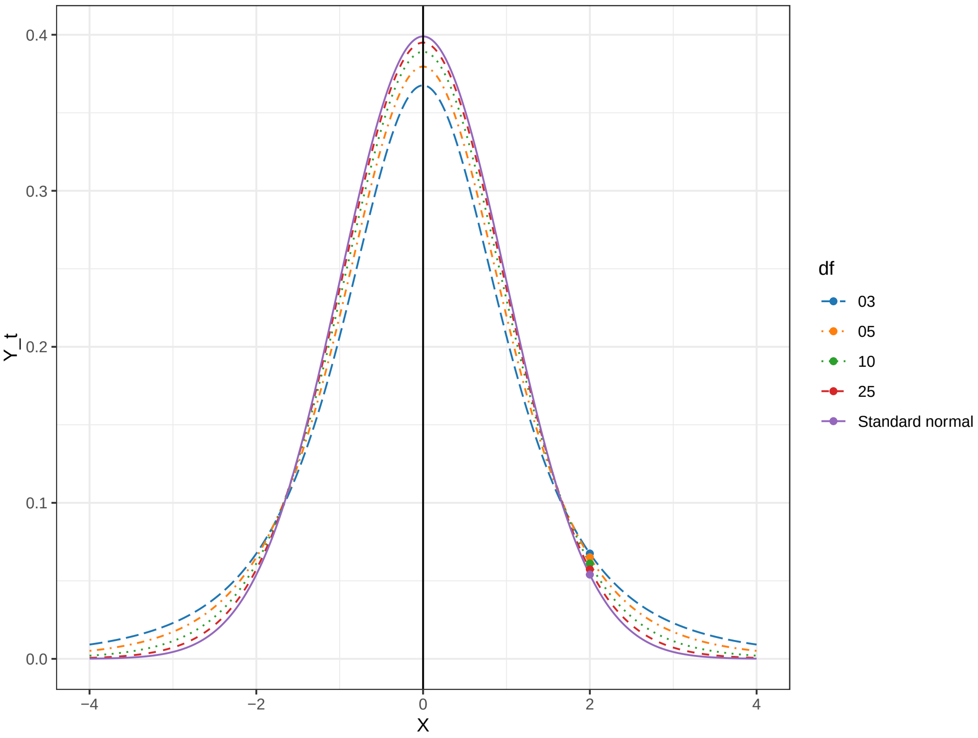
```
# t-distribution with 25 df
pt(q = 2, df = 25)
```

```
[1] 0.971762
```

We are essentially comparing the height of these distributions at $X = 2$.

Figure 6

Plot of Several t-Distributions with Differing Degrees of Freedom



Note. The probability density for $t = 2$ is also displayed for each of the distributions.

Comparing Cumulative Densities

What if we wanted to look at cumulative density? Consider out hypothesis test of whether a population parameter is equal to 0. Also consider that we observed a statistic (that has been standardized) of 2.5 using a sample size of $n = 15$.

If we can assume that the SAMPLING DISTRIBUTION is normally-distributed then we can use the cumulative density in a normal distribution to compute a p-value:

```
2 * pnorm(q = -2.5, mean = 0, sd = 1)
```

```
[1] 0.01241933
```

If, however, the SAMPLING DISTRIBUTION is t-distributed then we need to use the cumulative density for a t-distribution with the appropriate df to compute a p-value. For example if we use $df = n - 1$, the two-tailed p-value would be:

```
2 * pt(q = -2.5, df = 14)
```

```
[1] 0.02546666
```

The p-value using the t-distribution is larger than the p-value computed based on the standard normal distribution. This is again because of the increased error (uncertainty) we are introducing when we estimate from sample. This added uncertainty makes it harder for us to reject a hypothesis.

Using the t-Distribution in Regression

To illustrate how probability distributions are used in practice, we will will use the *riverview.csv* (see the data codebook for more information about these data) and fit a regression model that uses education level and seniority to predict variation in employee income.

```
# Read in data
city = read_csv(file = "~/Documents/github/epsy-8252/data/riverview.csv")
head(city)
```

```
# A tibble: 6 x 6
  education income seniority gender  male party
      <dbl>  <dbl>     <dbl> <chr>  <dbl> <chr>
1         8  37449         7 male       1 Democrat
2         8  26430         9 female     0 Independent
3        10  47034        14 male       1 Democrat
4        10  34182        16 female     0 Independent
5        10  25479         1 female     0 Republican
6        12  46488        11 female     0 Democrat
```

```
# Fit regression model
lm.1 = lm(income ~ 1 + education + seniority, data = city)

# Coefficient-level output
tidy(lm.1)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    6769.     5373.      1.26 0.218
2 education      2252.      335.      6.73 0.000000220
3 seniority       739.      210.      3.52 0.00146
```

How do we obtain the p-value for each of the coefficients? Recall that the coefficients and SEs for the coefficients are computed directly from the raw data. Then we can compute a test-statistic by dividing the coefficient estimate by the SE. For example, to compute the test-statistic associated with education level:

$$t = \frac{2252}{335} = 6.72$$

Since we are estimating the SE using sample data, our test statistic is likely t-distributed. Which value should we use for *df*? Well, for that, statistical theory tells us that we should use the error *df* value. In our data,

$$\begin{aligned} n &= 32 \\ \text{Total df} &= 32 - 1 = 31 \\ \text{Model df} &= 2 \text{ (two predictors)} \\ \text{Error df} &= 31 - 2 = 29 \end{aligned}$$

Using the t-distribution with 29 *df*,

```
2 * pt(q = -6.72, df = 29)
```

```
[1] 0.0000002257125
```

For seniority (and the intercept), we would use the same t-distribution, but our test statistic would differ:

$$\begin{aligned} t_{\text{Intercept}} &= \frac{6769}{5373} = 1.26 \\ t_{\text{Seniority}} &= \frac{739}{210} = 3.52 \end{aligned}$$

The associated p-values are:

```
# Intercept p-value
2 * pt(q = -1.26, df = 29)
```

```
[1] 0.2177149
```

```
# Seniority p-value
2 * pt(q = -3.52, df = 29)
```

```
[1] 0.001446316
```

Model-Level Inference: The F-Distribution

The model-level inference for regression is based on an F-statistic, which is a standardized measure of R^2.

```
# Model-level output
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1     0.742         0.724 7646.      41.7 2.98e-9     3  -330.  668.  674.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

In this example, the sample R^2 value is 0.742. Computation of the F-statistic relies on two df values—the model degrees of freedom (2) and the error degrees of freedom (29). To compute the F-statistics from R^2 we use:

$$F = \frac{R^2}{1 - R^2} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}}$$

In our example, we compute F as:

$$\begin{aligned} F &= \frac{0.742}{1 - 0.742} \times \frac{29}{2} \\ &= 41.7 \end{aligned}$$

We write this standardization of R^2 as $F(2, 29) = 41.7$.

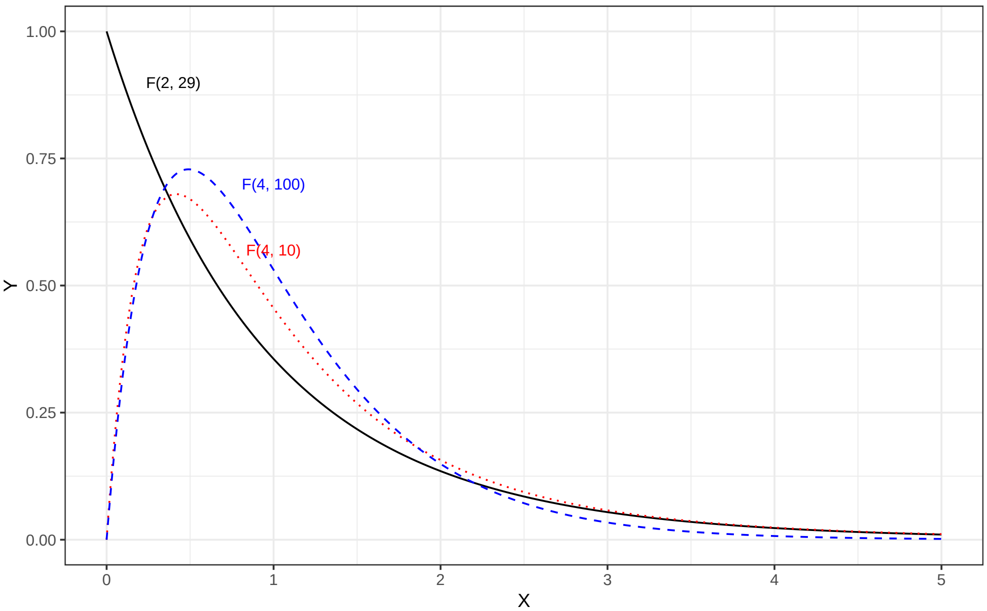
Testing the Model-Level Null Hypothesis

It is often worth testing whether the model explains a statistically significant amount of variation in the population. To do this we test the null hypothesis:

$$H_0 : \rho^2 = 0$$

Similar to the tests of the coefficients, we evaluate our test statistic (F in this case) in the appropriate test distribution, in this case an F-distribution with 2 and 29 degrees of freedom. (The shape of the F-distribution is based on two *df* values.) The figure below, shows the $F(2, 29)$-distribution as a solid, black line.

Figure 7
Plot of Several F-Distributions with Differing Degrees of Freedom



Note. The $F(2,\ 29)$-distribution is shown as a solid, black line.

The F-distribution, like the t-distribution is a family of distributions. They are positively skewed and generally have a lower-limit of 0. Because of this, when we use the F-distribution to compute a p-value, we only compute the cumulative density GREATER THAN OR EQUAL TO the value of the standardized test statistic.

Computing F from the ANOVA Partitioning

We can also compute the model-level F-statistic using the partitioning of variation from the ANOVA table.

```r
anova(lm.1)
```

```
Analysis of Variance Table

Response: income
          Df     Sum Sq    Mean Sq F value         Pr(>F)    
education  1 4147330492 4147330492  70.944 0.000000002781 ***
seniority  1  722883649  722883649  12.366        0.00146 ** 
```

```
Residuals 29 1695313285   58459079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is a ratio of the mean square for the model and the mean square for the error. To compute a mean square we use the general computation

$$\text{MS} = \frac{\text{SS}}{\text{df}}$$

The model includes both the education and seniority predictor, so we combine the SS and df. The MS model is:

$$\begin{aligned}
\text{MS}_{\text{Model}} &= \frac{\text{SS}_{\text{Model}}}{\text{df}_{\text{Model}}} \\
&= \frac{4147330492 + 722883649}{1+1} \\
&= \frac{4870214141}{2} \\
&= 2435107070
\end{aligned}$$

The MS error is:

$$\begin{aligned}
\text{MS}_{\text{Error}} &= \frac{\text{SS}_{\text{Error}}}{\text{df}_{\text{Error}}} \\
&= \frac{1695313285}{29} \\
&= 58459079
\end{aligned}$$

Then, we compute the F-statistic by computing the ratio of these two mean squares.

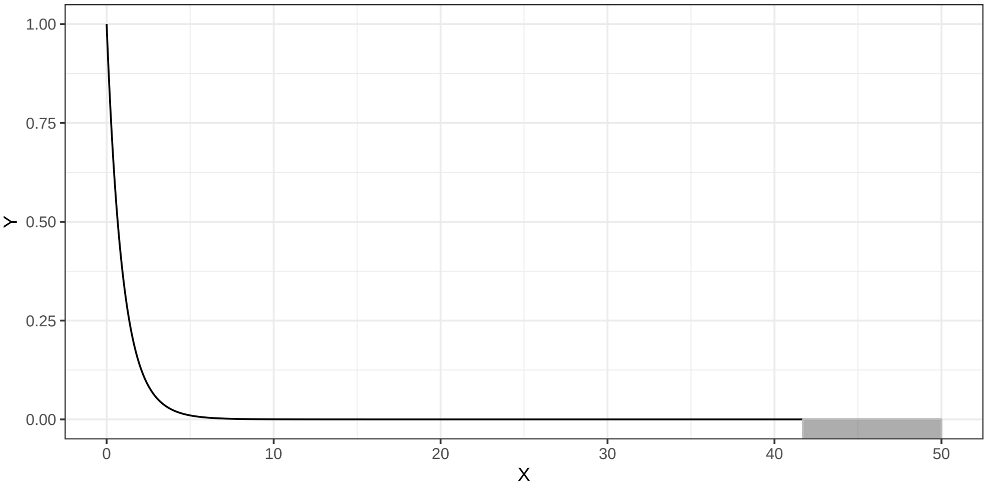
$$\begin{aligned}
F &= \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{Error}}} \\
&= \frac{2435107070}{58459079} \\
&= 41.7
\end{aligned}$$

This is the observed F-statistic for the model. Note that this is an identical computation (although reframed) as the initial computation for F.

$$
\begin{aligned}
F &= \frac{R^2}{1-R^2} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}}{\frac{\text{SS}_{\text{Error}}}{\text{SS}_{\text{Total}}}} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Error}}} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\text{SS}_{\text{Model}}}{\text{df}_{\text{Model}}} \times \frac{\text{df}_{\text{Error}}}{\text{SS}_{\text{Error}}} \\
&= \text{MS}_{\text{Model}} \times \frac{1}{\text{MS}_{\text{Error}}} \\
&= \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{Error}}}
\end{aligned}
$$

To test the null hypothesis, $H_0 : \rho^2 = 0$, we evaluate this observed F-statistic in an F-distribution with the 2 and 29 degrees of freedom.

Figure 8
Cumulative Density Representing the p-Value Using an Observed F-Statistic of 41.7



Note. Plot of the probability density function (PDF) for the $F(2,\ 29)$-distribution. The cumulative density representing the p-value for a test evaluating whether $\rho^2 = 0$ using an observed F-statistic of 41.7 is also displayed.

The computation using the cumulative density function, `pf()`, to obtain the p-value is:

```
1 - pf(41.7, df1 = 2, df2 = 29)
```

```
[1] 0.000000002942114
```

Mean Squares are Variance Estimates

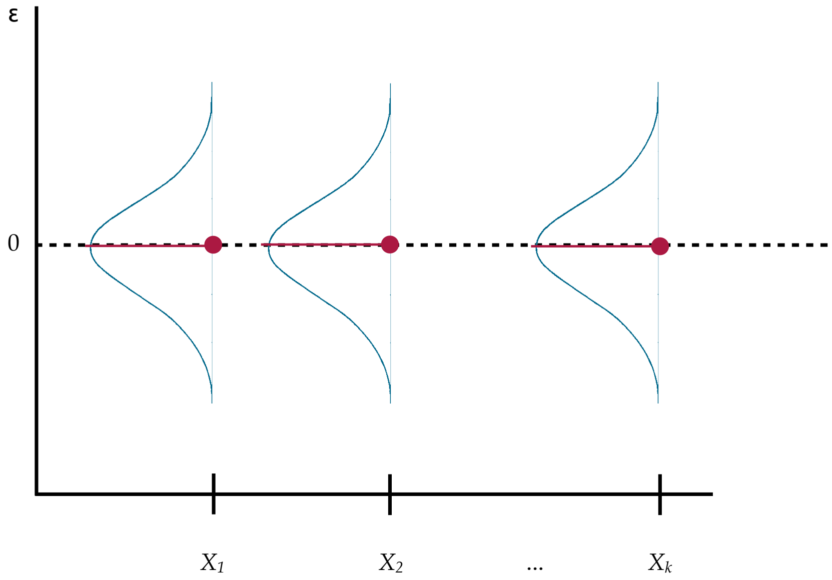
Mean squares are estimates of the variance. Consider the computational formula for the sample variance,

$$\hat{\sigma}^2 = \frac{\sum(Y - \bar{Y})^2}{n - 1}$$

This is the total sum of squares divided by the total *df*. When we compute an F-statistic, we are finding the ratio of two different variance estimates—one based on the model (explained variance) and one based on the error (unexplained variance). Under the null hypothesis that $\rho^2 = 0$, we are assuming that all the variance is unexplained. In that case, our F-statistic would be close to zero. When the model explains a significant amount of variation, the numerator gets larger relative to the denominator and the F-value is larger.

The mean squared error (from the `anova()` output) plays a special role in regression analysis. It is the variance estimate for the conditional distributions of the residuals in our visual depiction of the distributional assumptions of the residuals underlying linear regression.

Figure 9
Visual Depiction of the Distributional Assumptions of the Residuals Underlying Linear Regression



Recall that we made implicit assumptions about the conditional distributions of the residuals, namely that they were identically and normally distributed with a mean of zero and some variance. Based on the estimate of the mean squared error, the variance of each of these distributions is 58,459,079.

While the variance is a mathematical convenience, the standard deviation is a better descriptor of the variation in these distributions. The standard deviation is 7646.

```
sqrt(58459079)
```

```
[1] 7645.854
```

We can also obtain this value from the model-level regression output. Here it is typically referred to as the *Root Mean Squared Error* (RMSE). In the `glance()` output this value is in the `sigma` column.

```r
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1     0.742         0.724 7646.      41.7 2.98e-9     3  -330.  668.  674.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Why is this value important? It gives the expected variation in the distribution. For example, since all of the conditional distributions of the residuals are normally distributed, we would expect that 95% of the residuals would fall between ± 2 standard errors from 0; or, in this case, between -15292 and 15292. Observations with residuals that are more extreme may be regression outliers.

References

Fox, J. (2009). *A mathematical primer for social statistics*. Sage.