Test Statistics and Chi Square Tests

Jared Cross

2022-12-01

Suspicious Dice

Ancient Romans used asymmetrical dice. Let's imagine that we collected data on one of these dice.

Dice Results

side	num_times
1	80
2	103
3	93
4	105
5	96
6	123

Is it Fair?

Is there compelling evidence that this isn't a fair die?

side	num_times
1	80
2	103
3	93
4	105
5	96
6	123

A Little Math

```r
num_times <- c(80, 103, 93, 105, 96, 123)

expected <- rep(100, 6)

differences = num_times-expected

differences
```

```
## [1] -20   3  -7   5  -4  23
```

Are these differences larger than we would expect by chance? How do we capture how big these differences are?

The Sum of the Absolute Differences?

The “test statistic” could be the sum of the absolute differences

```
test_stat = sum(abs(differences))
test_stat
```

```
## [1] 62
```

Is this a larger sum of the absolute differences that we could reasonably expect due to chance alone?

Let's Simulate This

The simulation is based on the null hypothesis. The null hypothesis here could be that all 6 sides are equally likely on every roll.

```
roll600 = sample.int(6, 600, replace=TRUE)

roll600
```

```
##   [1] 1 6 3 3 5 2 5 2 4 2 3 6 3 6 4 3 4 6 2 3 3 1 2 3 2
##  [38] 2 1 6 1 2 6 3 1 3 4 4 6 4 2 6 5 2 3 2 6 5 2 5 4 3
##  [75] 6 5 2 6 5 4 3 5 2 3 4 2 3 6 1 1 6 2 5 5 3 4 4 3 6
## [112] 6 1 2 4 3 3 5 6 2 6 1 3 1 2 6 1 4 3 1 3 6 3 5 4 5
## [149] 4 6 3 4 2 3 3 6 1 6 5 6 2 3 3 2 3 3 6 3 1 4 1 6 2
## [186] 5 5 1 5 3 4 5 3 3 1 5 1 1 3 3 6 1 4 4 2 3 1 4 1 1
## [223] 1 4 4 2 2 5 1 6 3 4 5 4 6 1 6 4 4 4 2 4 1 4 3 6 5
## [260] 2 4 2 2 3 1 1 1 1 6 4 1 4 2 1 2 3 6 6 6 2 3 5 3 6
## [297] 3 2 5 1 6 2 2 6 6 2 3 3 6 2 4 5 3 5 4 3 6 4 3 1 3
## [334] 5 3 2 3 3 2 6 6 6 2 2 2 5 6 3 2 5 1 5 1 6 4 1 4 1
## [371] 6 3 3 3 6 4 2 6 2 2 6 6 5 1 5 2 4 3 1 3 3 6 4 5 3
## [408] 6 3 2 1 3 6 6 1 6 2 4 6 1 2 2 5 6 6 6 4 6 1 4 6 2
```

How'd we do?

```
table(roll600)
```

```
## roll600
##   1   2   3   4   5   6
##  94 101 110  97  94 104
```

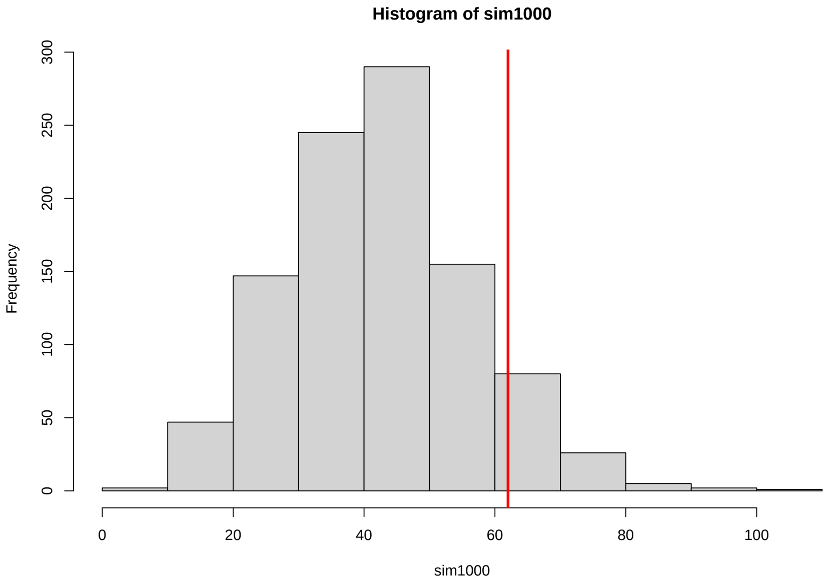
```
sim_differences = table(roll600)-100

sum(abs(sim_differences))
```

```
## [1] 30
```

Let's do this many times!

```r
sim1000 = replicate(1000, {
  roll600 = sample.int(6, 600, replace=TRUE)
  sim_differences = table(roll600)-100
  sum(abs(sim_differences))
})
```

The Distribution of the Test Statistic



A p-value

How often is our simulated test statistic as large or larger than our observed test statistic?

```
mean(sim1000 > test_stat)
```

```
## [1] 0.095
```

A Different Test Statistic

Some test statistics are better than others. The distributions of some test statistics are well understood mathematically.

The test statistic that people would typically use in this case is the sum of the squared differences each divided by the expected number. This test statistic is used often enough that it has a name. It's called a "chi square" value.

$$\chi^2 = \Sigma \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Calculating Chi Square

```
num_times <- c(80, 103, 93, 105, 96, 123)

expected <- rep(100, 6)

differences = num_times-expected

(differences^2)/expected
```

```
## [1] 4.00 0.09 0.49 0.25 0.16 5.29
```

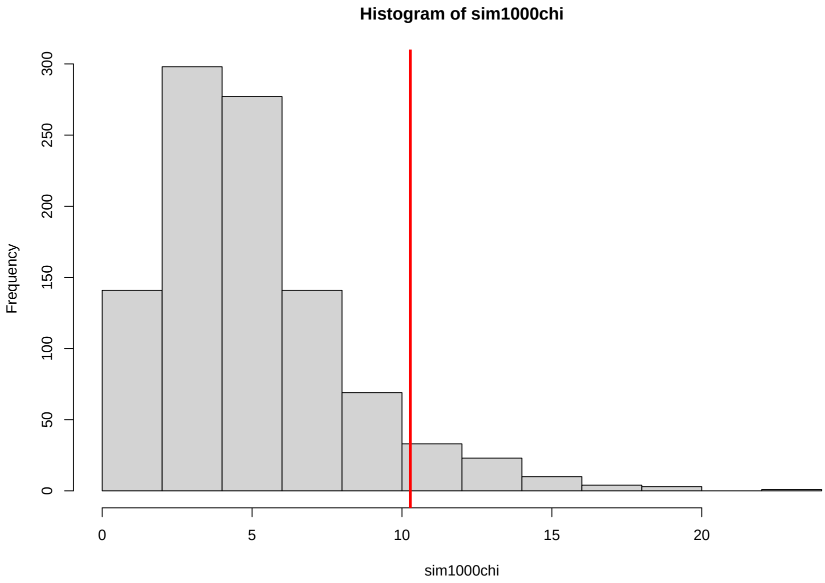
```
chi_square = sum((differences^2)/expected)

chi_square
```

```
## [1] 10.28
```

Now Let's Do Our Simulation Again!

```r
sim1000chi = replicate(1000, {
  roll600 = sample.int(6, 600, replace=TRUE)
  sim_differences = table(roll600)-100
  sum((sim_differences^2)/100)
})
```

And Graph Our Simulated Chi Squares



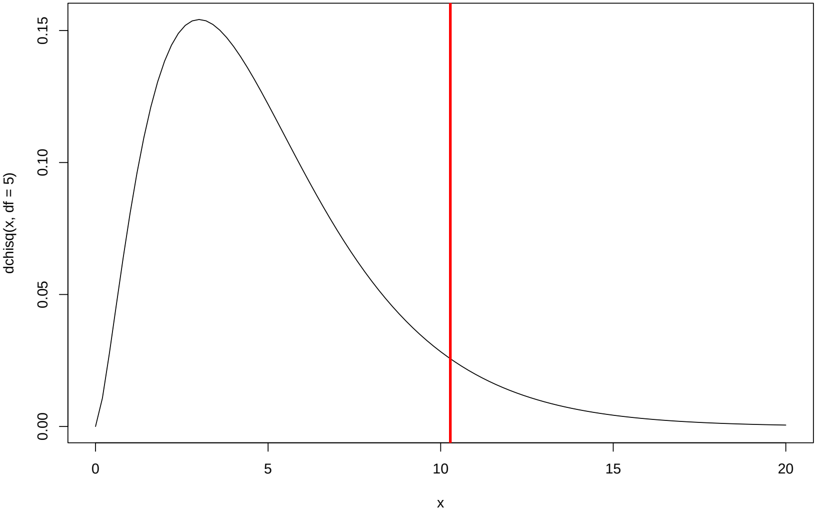
And we can get a new p-value

```
mean(sim1000chi > chi_square)
```

```
## [1] 0.069
```

Chi Square Distribution

Chi square values are understood well enough that we know in advance how many simulations would be distributed:



Degrees of Freedom (generally)

Something's "Degrees of Freedom" is number of independent ways of wiggling it has.

If video games of my youth, characters had two degrees of freedom (up/down, left/right). Today's video game characters might have many more. If they can move in 3D space and rotate around two axes, that's 5 degrees of freedom.

Degrees of Freedom (more specifically)

Our die rolls results have 5 degrees of freedom and not 6, because if you know how many 1's through 5's we rolled, there's no room for wiggle in how many 6's we rolled.

(Here 6 wiggles that have some dependence between them is the same as 5 independent wiggles.)

Finding the p-value the easy way

Chi Square Table

```
1 - pchisq(chi_square, df=5)
```

```
## [1] 0.06767935
```

Example 2: Soccer Birth Months (Again)

Elite soccer players by birth month:

```
births = c(70, 60, 57, 51, 60, 44, 39, 55, 59, 35, 30, 31)
names(births) = month.name

total_births = sum(births)
births
```

```
##   January  February     March     April       May      J
##        70        60        57        51        60
## September   October  November  December
##        59        35        30        31
```

Expected Births

```
days_in_months = c(31, 28.25, 31, 30, 31, 30, 31, 31, 30, 3
)

names(days_in_months) = month.name

expected = total_births*(days_in_months/365.25)

expected
```

```
##   January  February     March     April       May      J
##  50.16016  45.71047  50.16016  48.54209  50.16016  48.54
## September   October  November  December
##  48.54209  50.16016  48.54209  50.16016
```

Finding Chi Square

```
births-expected
```

```
##    January   February      March      April        May
##  19.839836  14.289528   6.839836   2.457906   9.839836
##     August  September    October   November   December
##   4.839836  10.457906 -15.160164 -18.542094 -19.160164
```

```
sum((births-expected)^2/expected)
```

```
## [1] 39.91319
```

Finding p-value

```r
soccer_chi = sum((births-expected)^2/expected)

pchisq(soccer_chi, df=11)
```

```
## [1] 0.999963
```

```r
1-pchisq(soccer_chi, df=11)
```

```
## [1] 3.701896e-05
```

Expected Value of Chi Square Value

Since:

$$\chi^2 = \Sigma \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

$$E[\chi^2] = \Sigma \frac{E[(Observed_i - Expected_i)^2]}{Expected_i}$$

Remember: The variance is the mean square deviation from the mean (which is the numerator above), so we could rewrite this as:

$$E[\chi^2] = \Sigma \frac{Variance_i}{Expected_i}$$

Expected Value of Chi Square Value (continued)

$$E[\chi^2] = \Sigma \frac{Variance_i}{Expected_i}$$

The variance in the number of counts in any category is $np_i(1-p_i)$ where p_i is the chance of being in that category. If we divide this by the expected number of counts in that category np_i, we get $1-p_i$. If we sum up $1-p_i$ over all categories, we get the number of categories minus 1... which is the same as the number of degrees of freedom.

So the expected value of chi square is the same as the number of degrees of freedom.