R Tutorial-08

T. Scofield

You may click here to access the .qmd file.

In this issue, we investigate

- the behavior of the `var()` and `sd()` commands
- the Chi-Square distributions
- t-distributions

What `var()` and `sd()` calculate

In class, I asserted

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator of the population variance σ^2. To illustrate that this is just what R calculates with its `var()` command, take a small set of values

$$81, 35, 42, 58$$

The mean $\bar{X}$

```
x = c(81, 35, 42, 58)
xbar = sum(x) / 4
xbar      # mean(~x) would have calculated this, too
```

```
[1] 54
```

The values $(X - \bar{X})$ come from

```
x - xbar
```

```
[1]  27 -19 -12   4
```

so we find $\frac{1}{3}\sum(X - \bar{X})^2$ via

```r
sum((x - xbar)^2) / 3
```

```
[1] 416.6667
```

That is S^2, as we have defined it. It is easier just to do

```r
var(~x)
```

```
[1] 416.6667
```

Of course, the point here is that `var()` builds in the division by $(n-1)$, like the formula calls for. So does `sd()`:

```r
s = sd(~x)
s^2
```

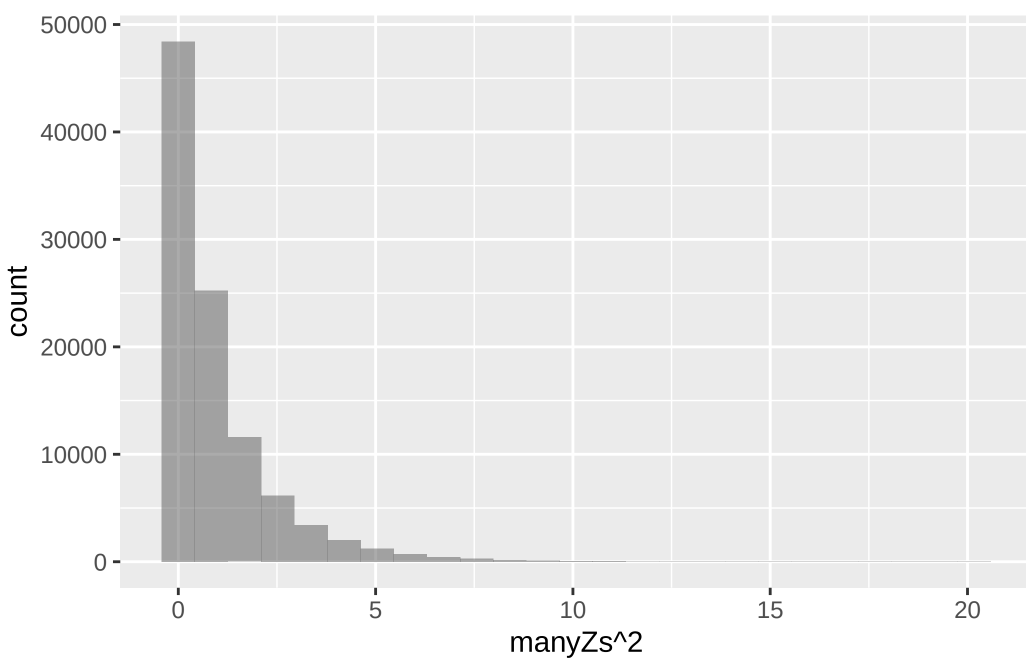
```
[1] 416.6667
```

The Chi-Square distributions

These are wholly new to us. I have described them as "sums of squared standard normal variables".

At start, if we have just one standard variable Z, we can simulate the distribution of Z^2:

```r
manyZs = rnorm(100000)
gf_histogram(~ manyZs^2)
```

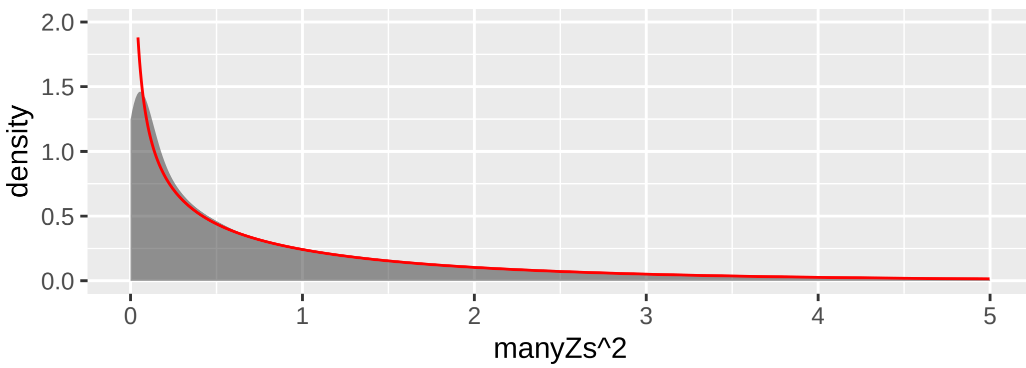


That this is an important (and known) shape/distribution, called the **chi-square distribution with 1 degree of freedom**, comes as something of a surprise.

```
gf_density(~manyZs^2) |> gf_dist("chisq", df=1, color="red") |>
  gf_refine(
    scale_x_continuous(limits = c(0,5)),
    scale_y_continuous(limits = c(0,2)))
```

```
Warning: Removed 2505 rows containing non-finite outside the scale range
(`stat_density()`).
```

```
Warning: Removed 3955 rows containing missing values or values outside the scale range
(`geom_line()`).
```



If we are adding the square of 3 standard normal variables, this is a chi-square distribution with 3 dfs. The command

```
sum(rnorm(3)^2)
```

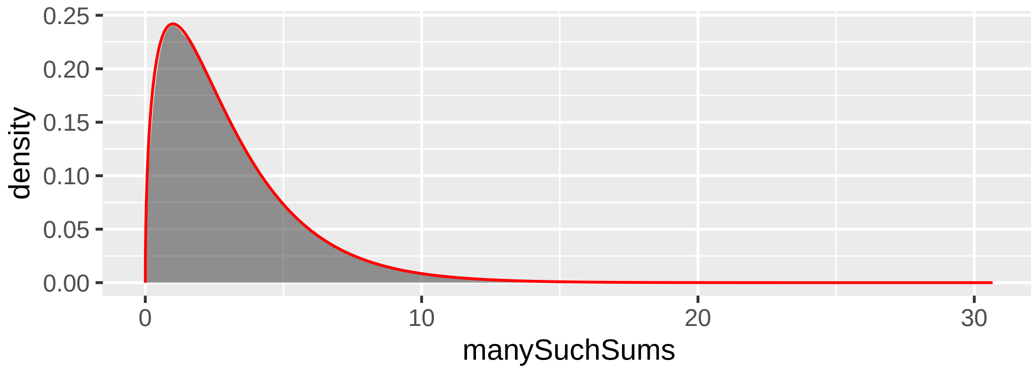
```
[1] 4.394534
```

produces one instance of a sum $Z_1^2 + Z_2^2 + Z_3^2$. To simulate many of them

```
manySuchSums = replicate(100000, sum(rnorm(3)^2))
```

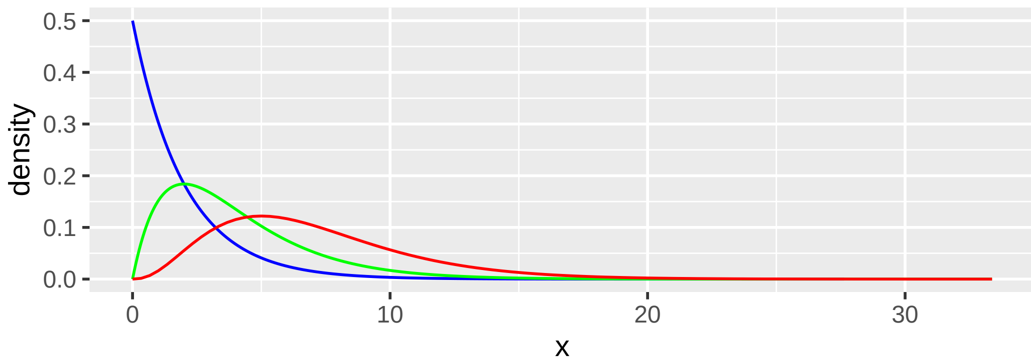
We compare with the chi-square distribution having 3 dfs:

```
gf_density(~manySuchSums) |> gf_dist("chisq", df=3, color="red")
```



View several chi-square distributions together by piping one on to another (and another, etc):

```
gf_dist("chisq", df=2, color="blue") |>
  gf_dist("chisq", df=4, color="green") |>
  gf_fun(dchisq(x,7)~x, color="red")
```

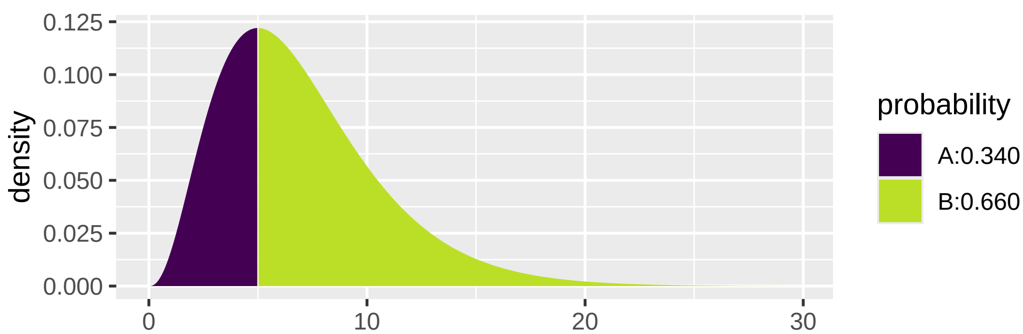


The commands `pchisq()`, `qchisq()`, `dchisq()`, `rchisq()` exist to play rolls similar to those for other distributional families with the p-, q-, d-, and r- prefixes. If X has a chi-square distribution with 7 degrees of freedom (df=7), then the probability $\mathrm{P}(X > 5)$ is found via

```
1 - pchisq(5, df=7)
```

```
[1] 0.6599632
```

```
xpchisq(5, df=7)     # a graphical version
```



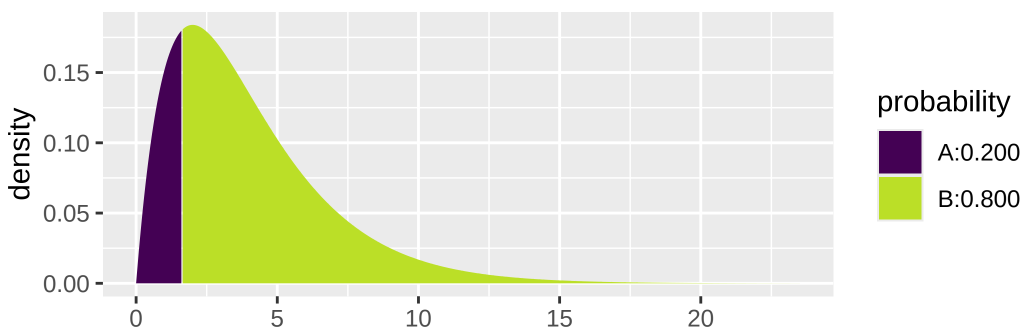
```
[1] 0.3400368
```

And the 20th percentile, the location dividing the lowest 20% of values from the upper 80%, for the region enclosed by the pdf of a chi-square distribution with 4 dfs, is found with

```
qchisq(0.2, df=4)
```

```
[1] 1.648777
```

```
xqchisq(0.2, 4)     # a graphical version
```



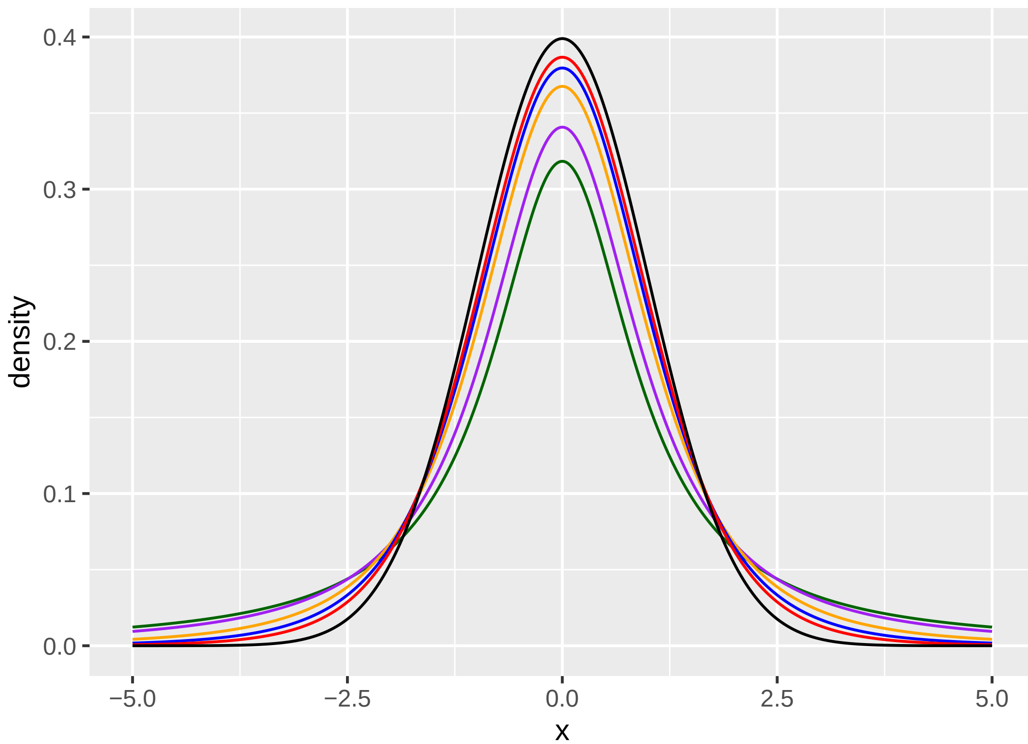

```
[1] 1.648777
```

Student t distributions

Plotted below are the

- t-distribution with $df = 1$ (dark green)
- t-distribution with $df = 1.5$ (purple)
- t-distribution with $df = 3$ (orange)
- t-distribution with $df = 5$ (blue)
- t-distribution with $df = 8$ (red)
- standard normal distribution $\text{Norm}(0, 1)$ (black)

```
gf_dist("t", df=1, xlim=c(-5,5), color="darkgreen") |>
  gf_dist("t", df=1.5, xlim=c(-5,5), color="purple") |>
  gf_dist("t", df=3, xlim=c(-5,5), color="orange") |>
  gf_dist("t", df=5, xlim=c(-5,5), color="blue") |>
  gf_dist("t", df=8, xlim=c(-5,5), color="red") |>
  gf_dist("norm", xlim=c(-5,5), color="black")
```



The Student t-distributions

- are symmetric, "bell"-shaped
- are centered at 0
- are increasingly similar to the (black) standard normal pdf as the number of degrees of freedom grows
- have more area in the tails and less in the middle, when compared with $\text{Norm}(0, 1)$

To illustrate the latter, compare

```
pt(-2.5, df=3)
```

```
[1] 0.04385332
```

```
pnorm(-2.5)
```

```
[1] 0.006209665
```

The first gives the area under orange pdf, the one for a t-distribution with df=3, while the latter value is the area under the standard normal pdf, both computed as integrals from $(-\infty)$ to -2.5. There is more area under the orange curve.