

Chi-Square Distribution

Recap

Typically, the probability distribution function for a distribution is regulated by two parameters:

- mean
- standard deviation

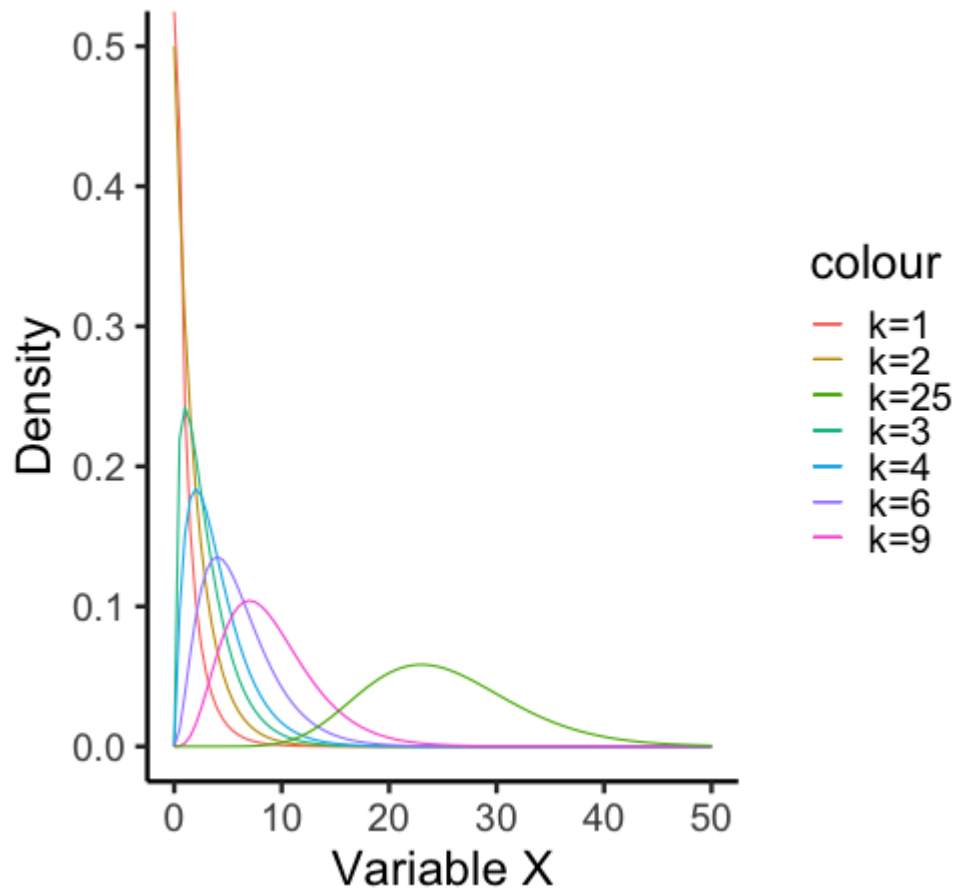
These parameters are sometimes independent (normal, t) and sometimes associated with one another (binomial)

Chi-Square Distribution

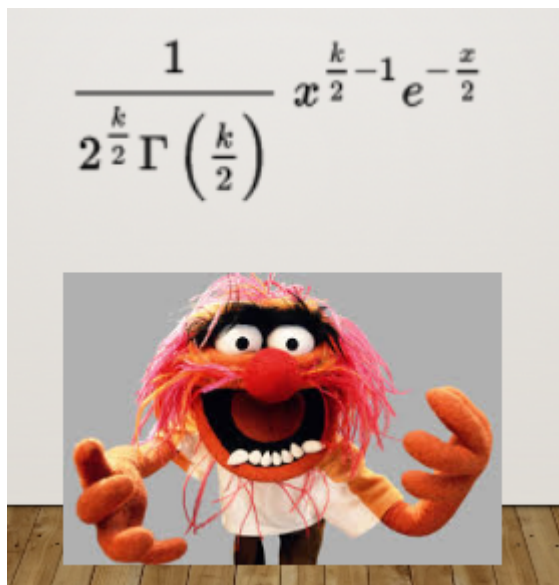
The **chi-square** (χ^2) distribution is very unique in that we really only have one parameter, k

- mean = k
- variance = $2k$

The Chi-Square Distribution



What kind of crazy function would lead to a distribution like this?!?!



If a tree falls in the forest, and no one is around, does it make a sound?

Trees Example

Let's say a total of 110 trees have fallen. We might *expect* that half of those trees make a sound and half of them don't.

- 55 trees fall and do make a sound
- 55 trees fall and do not make a sound

Trees Example

We expect 55 trees to fall per category (sound and no sound). However, our data show that 10 fell and made a sound whereas 100 fell and did not make a sound.

This let's build a table that shows us what we have:

```
trees = data.frame(TreesFallen = c(10, 100),  
                  Expected = c(55, 55))  
rownames(trees) = c("Registered Noise", "No Registered Noise")  
  
trees
```

##	TreesFallen	Expected
## Registered Noise	10	55
## No Registered Noise	100	55

Trees Example

##	TreesFallen	Expected
## Registered Noise	10	55
## No Registered Noise	100	55

$$\chi_P^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

Trees Example

##	TreesFallen	Expected
## Registered Noise	10	55
## No Registered Noise	100	55

Step 1: Get the ratio across rows:

$$\frac{(10-55)^2}{55} \quad \frac{(100-55)^2}{55}$$

```
noise <- ((10-55)^2)/55  
noise
```

```
## [1] 36.81818
```

```
noNoise <- ((100-55)^2)/55  
noNoise
```

```
## [1] 36.81818
```

Trees Example

Step 2: Sum these up

```
noise + noNoise
```

```
## [1] 73.63636
```

Trees Example

Step 3: Determine Degrees of Freedom

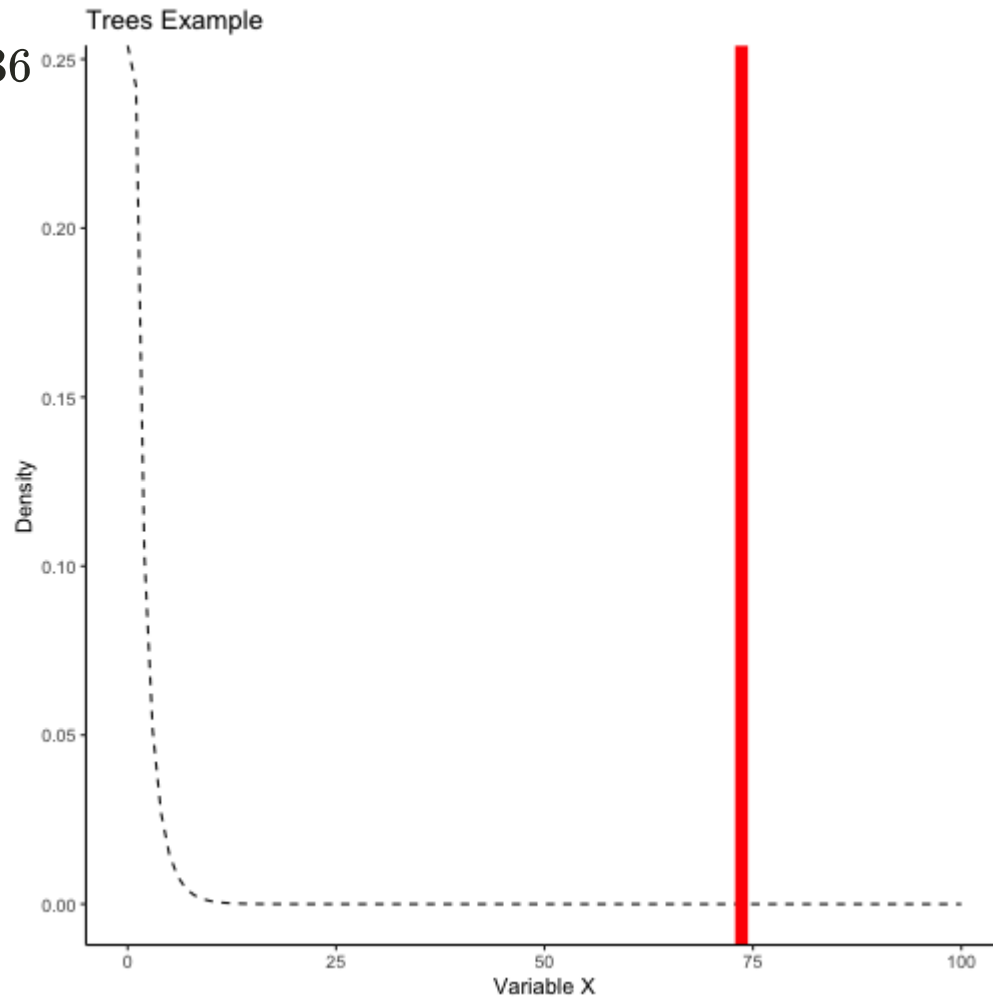
$$df = (r - 1)(c - 1)$$

$$df = (2 - 1)(2 - 1)$$

$$df = 1$$

Trees Example

$$\chi^2(1) = 73.636$$



What did we learn?

The χ^2 distribution gives us the probability to find the data (or those more extreme) given a theory (expected values)

Comparing that against some threshold for determining whether that probability is within the range of expectations or not

```
pchisq(q = 73.636, df = 1, lower.tail = FALSE)
```

```
## [1] 9.393574e-18
```

Our calculation, like with the z-score, reflects our value minus some expectation, to understand how far away it is from that expectation.

When to use?

- Classes of observations
- Non-continuous data

Assumptions

- Each observation is in only one category
- Observations are independent
- N is large (or expected N is large)
 - 2 groups greater than 10 per group
 - More than 2 groups greater than 5 per group

Remember Independence?

$$p(A)p(B) = p(A \cap B)$$

Let's expand... Tests of Association (or Independence)

- aka Pearson Chi-square test
- Developed to see how close we are to the estimated distribution
 - nominal data (ordinal and greater we can use correlations)

Same Old Same

- If we are interested in knowing whether X is contingent (or dependent) upon Y, or whether X and Y are **independent**, we're going to need a bigger table.
- And we need to calculate our expected frequencies differently, where...
- Expected value = $R_i C_j / N$

**Is attrition on a longitudinal study
related to educational attainment?**

Attrition Example

Our data...

##	StayedIn	DroppedOut
## Failed to complete high school	20	20
## High school degree	25	15
## College degree	30	10

Attrition Example

##	StayedIn	DroppedOut
## Failed to complete high school	20	20
## High school degree	25	15
## College degree	30	10

It's the same general formula, but we need to calculate our expected values a little differently...

$$\chi_P^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

$$\text{Expected Value} = R_i C_j / N$$

Attrition Example

The OG Data

##	StayedIn	DroppedOut
## Failed to complete high school	20	20
## High school degree	25	15
## College degree	30	10

The Expected Value Calculations

##	StayedIn	DroppedOut
## Failed to complete high school	$40 \times 75 / 120$	$40 \times 45 / 120$
## High school degree	$40 \times 75 / 120$	$40 \times 45 / 120$
## College degree	$40 \times 75 / 120$	$40 \times 45 / 120$

Expected frequency for staying in = 25

Expected frequency for dropping out = 15

Attrition Example

Plug this in for "expected frequencies"

```
##                               StayedIn  DroppedOut
## Failed to complete high school (20-25)^2/25 (20-15)^2/15
## High school degree             (25-25)^2/25 (15-15)^2/15
## College degree                 (30-25)^2/25 (10-15)^2/15
```

So we get...

```
##                               StayedIn  DroppedOut
## Failed to complete high school      1      1.666667
## High school degree                  0      0.000000
## College degree                      1      1.666667
```

Sum them up:

```
1 + 0 + 1 + 1.666667 + 0 + 1.666667
```

```
## [1] 5.333334
```


Attrition Example

$$\chi^2(?) = 5.33$$

What is our degrees of freedom?

- $df = (r - 1)(c - 1)$
- $df = (3 - 1)(2 - 1)$
- $df = 2 * 1$
- $df = 2$

$$\chi^2(2) = 5.33$$

```
pchisq(5.33, df = 2, lower.tail = FALSE)
```

```
## [1] 0.06959935
```

What's up with these df?

We have 3 constraints:

- The cell frequencies must sum to the overall sample size
- The row totals must sum to the overall sample size
- The column totals must sum to the overall sample size

Assumptions

Independent observations

Each observation is in one and only one category

N is large (usually taken to mean that expected N is at least 5 in each cell)

Effect Sizes for Chi-Square

Odds Ratio OR = number experiencing event divided by number who did not experience event.

##	StayedIn	DroppedOut
## Failed to complete high school	20	20
## High school degree	25	15
## College degree	30	10

- $p(\text{Dropped Out} | < \text{high school}) = 20/40$
- $p(\text{Dropped Out} | \text{high school}) = 15/40$

Odds Ratio

OR = number experiencing event divided by number who did not experience event.

- Odds(Dropped Out | < High School = $20/40 = .5$
- Odds(Dropped Out | High School = $15/40 = .375$
- Odds Ratio = $.5 / .375 = 1.33$

The odds of dropping out of the study rather than remaining in the study when they did not complete high school are 1.33 times the odds of dropping out if they completed high school, but not college

Rules of thumb

Cohen (1988) provided the following advice for interpreting odds ratios:

- 1.5 small
- 2.5 medium
- 4.3 large

Phi Correlation

- Pearson correlation between two dichotomous variables is ϕ
- This doesn't quite work with our attrition example, so let's look at another contingency table...
- Rosenstein & Horowitz (1996): Adolescent attachment and psychopathology
- Researchers were interested in whether attachment to mothers was associated with having a conduct disorder or affective disorder (clinical sample)
- Attachment:
 - Preoccupied: appear confused and entangled by attachment relationships
 - Dismissive: dismiss the importance or influence of attachment figure
- Disorder:
 - Conduct: persistent antisocial behavior that violates norms
 - Affective: major or recurrent depressive symptoms, mania, and/or mood disorders

Attachment

##	AnyConduct	JustAffective
## Dismissive	1	5
## Preoccupied	3	14

Practice:

- Work through the full chi-square test on your own time. You should get .003. See if you're right!
- Calculate the effect size (odds ratio). You can choose what you're calculating, but note that your interpretation will change!

Phi Correlation

```
##                AnyConduct JustAffective
## Dismissive 1 (Cell 11)    5 (Cell 12)
## Preoccupied 3 (Cell 21)  14 (Cell 22)
```

$$\phi = \frac{Cell_{11}Cell_{22} - Cell_{12}Cell_{21}}{\sqrt{(Cell_{11} + Cell_{12})(Cell_{21} + Cell_{22})(Cell_{11} + Cell_{21})(Cell_{12} + Cell_{22})}}$$

Phi Correlation

$$\phi = \frac{Cell_{11}Cell_{22} - Cell_{12}Cell_{21}}{\sqrt{(Cell_{11} + Cell_{12})(Cell_{21} + Cell_{22})(Cell_{11} + Cell_{21})(Cell_{12} + Cell_{22})}}$$

$$\phi = \frac{(1 * 14) - (5 * 3)}{\sqrt{(1 + 5)(3 + 14)(1 + 3)(5 + 14)}}$$

$$\phi = .01$$

OR

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Next time

Comparing means with all the t -tests