χ^2 tests

E. Pastucha

October 2024

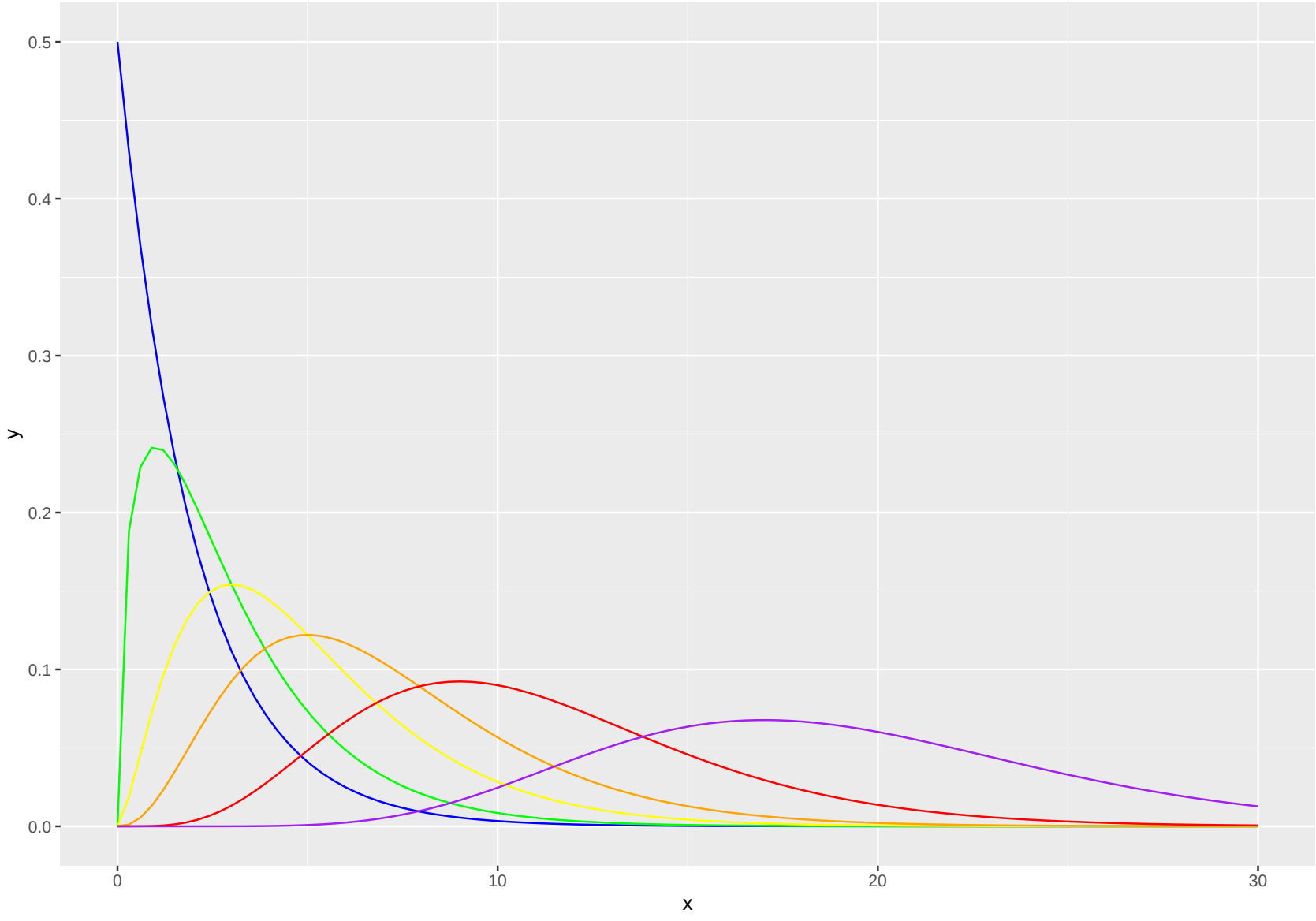
χ^2 goodness of fit test

A sample of of 669 gummy bears was taken from a production line. Within that sample there were:

Flavour	**Count**
strawberry	83
raspberry	142
lemon	100
orange	103
apple	104
pineapple	137

Can we assume equal proportion of all gummy bears flavours?

χ^2 distribution

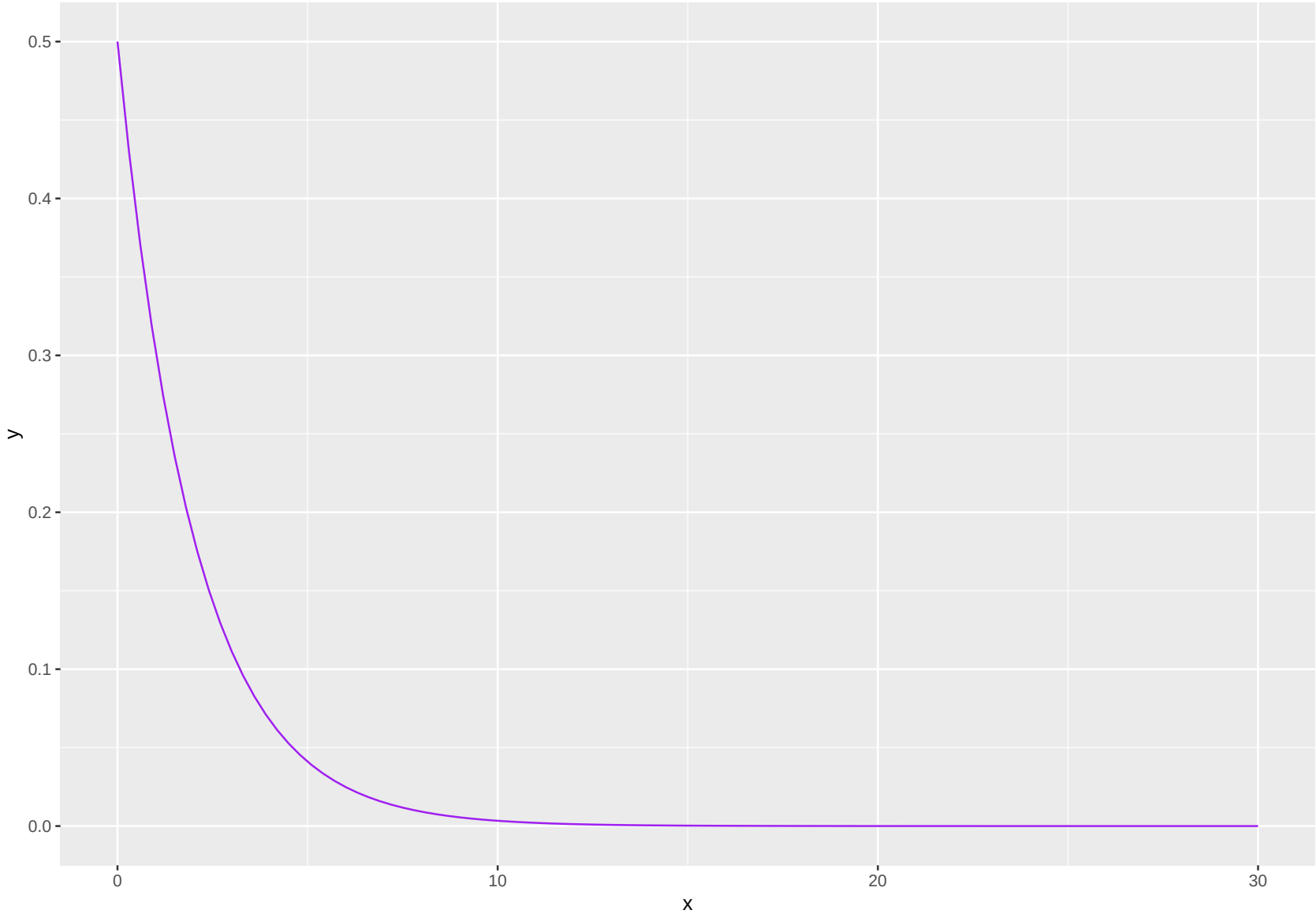


χ^2 distribution

- used to characterize data that is only positive and right skewed
- described by one parameter - Degrees of Freedom
- mean of each distribution is equal to df
- variability increases with df and becomes more symetric

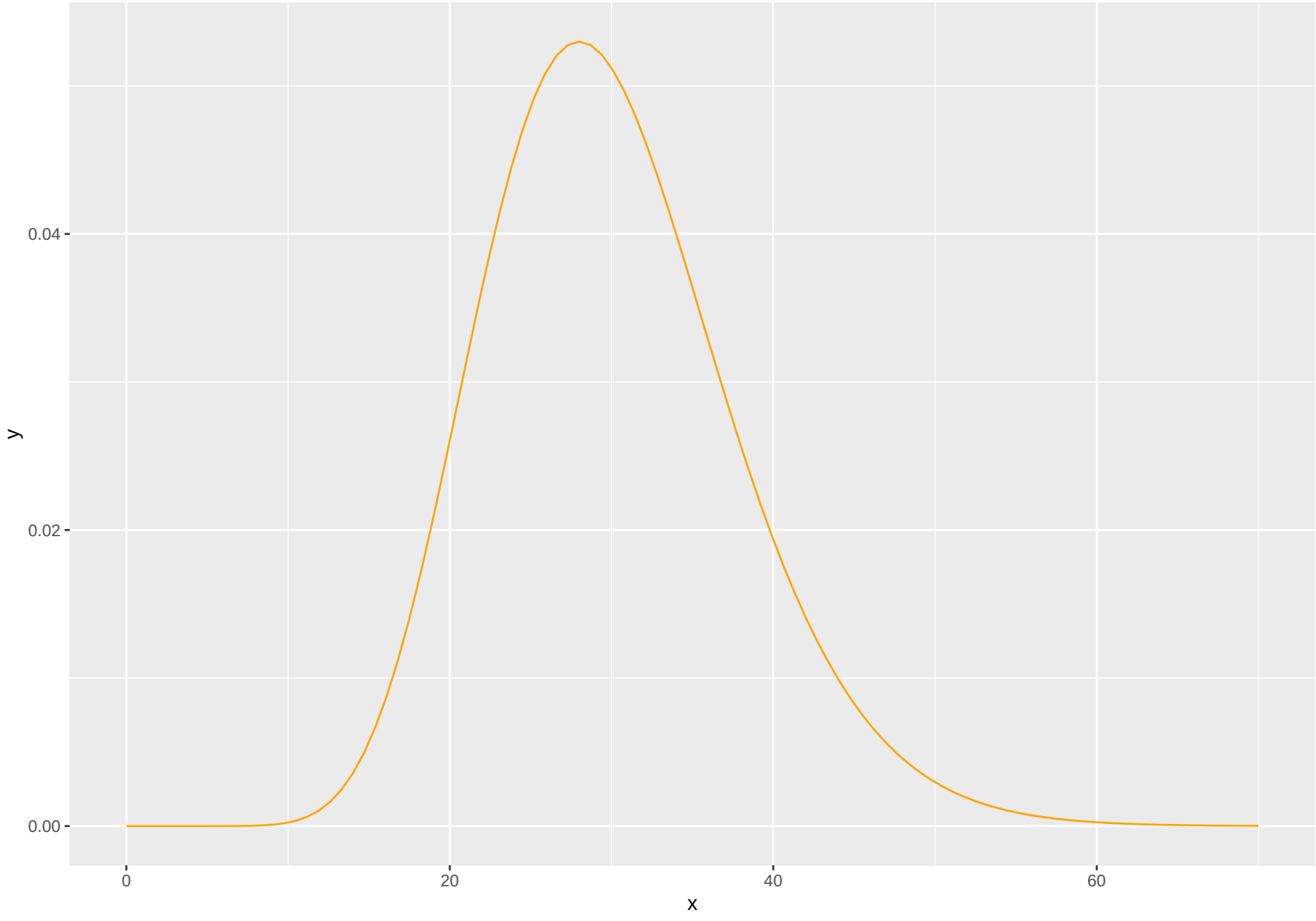
χ^2 distribution

2 df χ^2 distribution

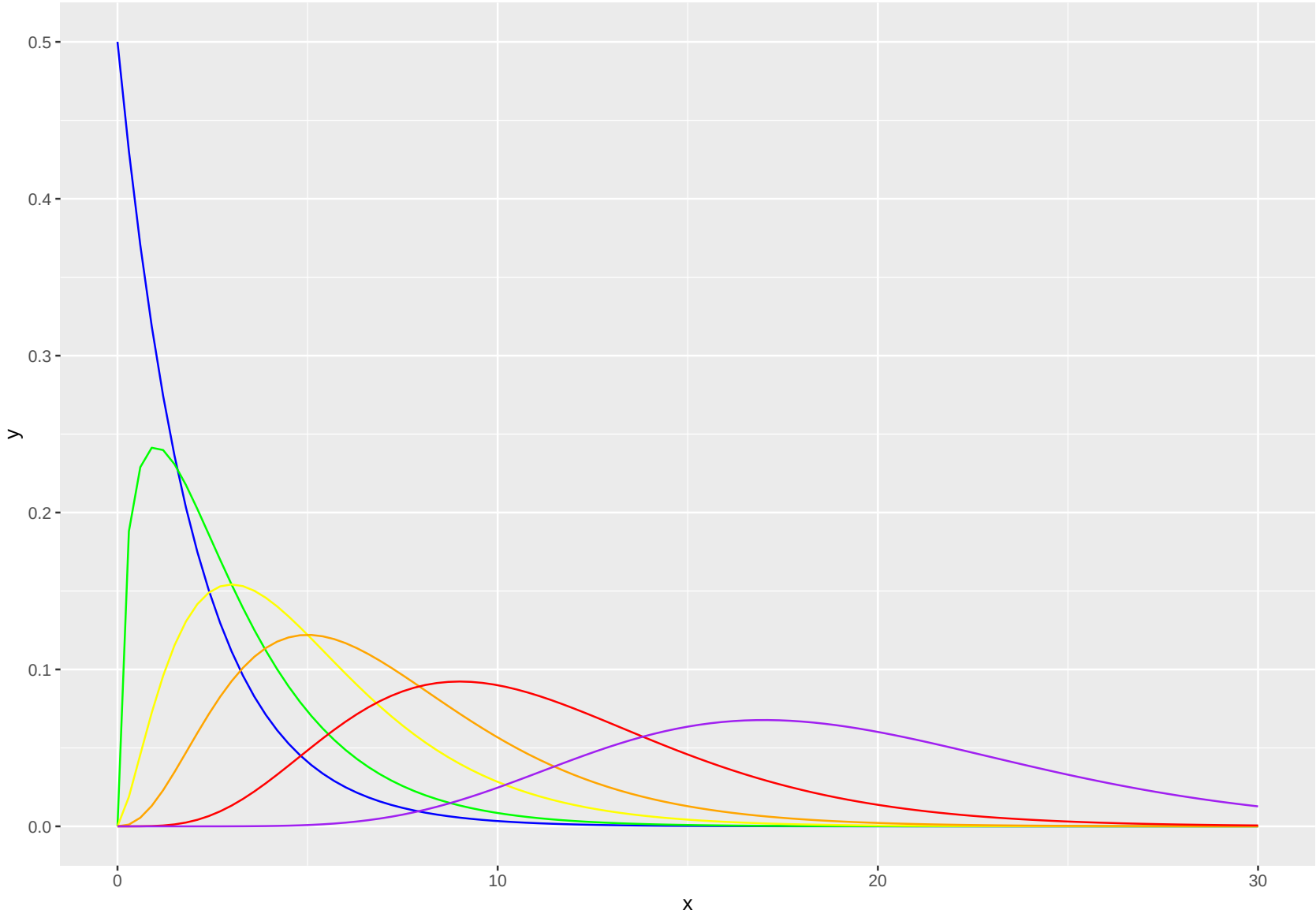


χ^2 distribution

30 df χ^2 distribution



χ^2 distribution



χ^2 goodness of fit test

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

χ^2 goodness of fit test

1. Set up hypothesis
2. Check conditions
3. Get the observed count for each category
4. Get the expected count for each category
5. Calculate test statistic
6. Calculate p - value
7. Form conclusions

χ^2 goodness of fit test

Let's look at a simple example of χ^2 test...

We're back at Haribo gummy bears.

χ^2 goodness of fit test

A sample of of 669 gummy bears was taken from a production line. Within that sample there were:

Flavour	**Count**
strawberry	83
raspberry	142
lemon	100
orange	103
apple	104
pineapple	137

Can we assume equal proportion of all gummy bears flavours?

χ^2 goodness of fit test

H_0 : Equal proportion of all flavours of gummy bears are produced.

H_A : Unequal proportion of all flavours of gummy bears are produced.

χ^2 goodness of fit test

Conditions:

- Independence
- Sample size - each category must have at least 5 expected cases.

χ^2 goodness of fit test

Check conditions:

Flavour	**Count**	**Null Count**
strawberry	83	111.5
raspberry	142	111.5
lemon	100	111.5
orange	103	111.5
apple	104	111.5
pineapple	137	111.5

χ^2 goodness of fit test

Test statistic - for each category we calculate:

$$Z_n = \frac{\textit{observed count} - \textit{null count}}{\sqrt{\textit{null count}}}$$

then to calculate test statistic:

$$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

χ^2 goodness of fit test

Test statistic:

$$X^2 = \frac{(\text{ obs. count }_1 - \text{ null count }_1)^2}{\text{null count }_1} + \cdots + \frac{(\text{ obs. count }_n - \text{ null count }_n)^2}{\text{null count }_n}$$

χ^2 goodness of fit test

Test statistic- for each category we calculate:

$$\textit{null count} = 669/6 \longrightarrow \textit{null count} = 111.5$$

```
gummies <- gummies %>%
  mutate(Z_n = (count - `null count`)
         /sqrt(`null count`))
```

Flavour	Count	Null Count	Z_n
strawberry	83	111.5	-2.6990282
raspberry	142	111.5	2.8884337
lemon	100	111.5	-1.0890816
orange	103	111.5	-0.8049733
apple	104	111.5	-0.7102706
pineapple	137	111.5	2.4149200

χ^2 goodness of fit test

Test statistic:

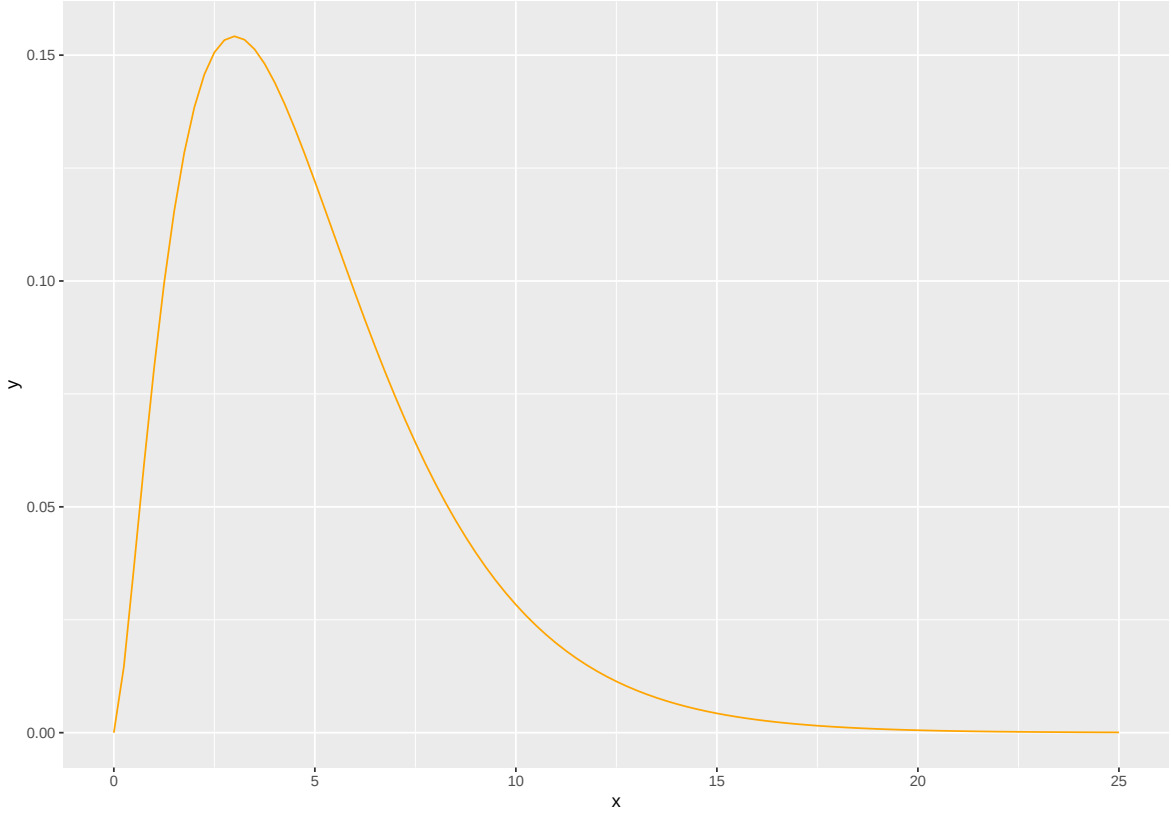
```
(test_statistic <- sum((gummies$Z_n)^2))
```

```
## [1] 23.79821
```

χ^2 goodness of fit test

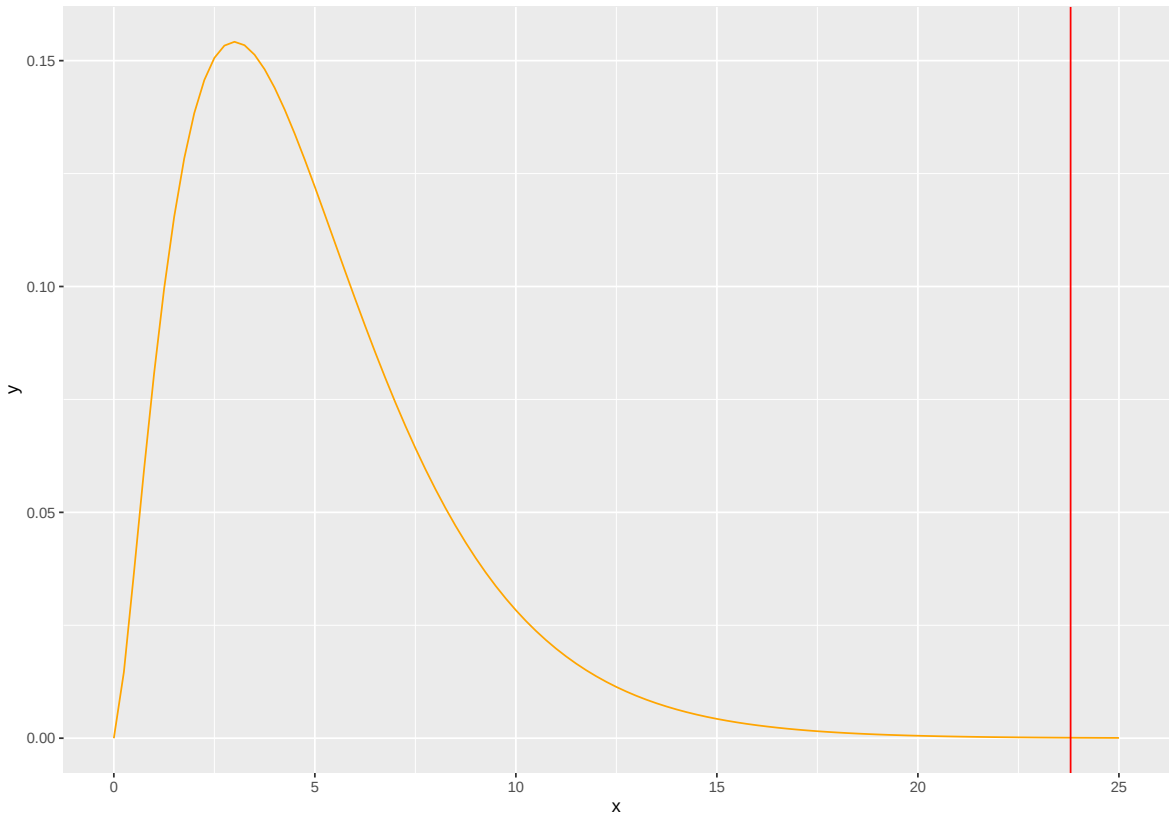
For a simple goodness of fit test degrees of freedom is number of categories - 1.

Gummy bears df is 5.



χ^2 goodness of fit test

Test statistic - 23.79821



χ^2 goodness of fit test

P-value:

```
(p_value <- 1 - pchisq(test_statistic, df = 5))
```

```
## [1] 0.0002373797
```

Conclusions:

We reject null hypothesis in favour of the alternative. There is unequall number of gummy bears flavours produced.

χ^2 goodness of fit test

1. Set up hypothesis
2. Check conditions
3. Get the observed count for each category
4. Get the expected count for each category
5. Calculate test statistic
6. Calculate p - value
7. Form conclusions

χ^2 independence tests

Two way table

	Democrats	**Republicans**	**Total**
Listen to POP	100	30	130
Listen to country	5	90	95
Total	105	120	

Corelation?

Two way table

Two way table presents combinations of categories.

	Pepperoni	**Mushrooms**	**Kebab**	**Total**
Kung Fu Panda 4	20	10	5	35
Deadpool & Wolverine	15	12	15	42
Mean Girls	8	13	2	23
Total	43	35	22	

Corelation?

χ^2 independence test

1. Set up hypothesis
2. Check conditions
3. Get the observed count for each category
4. Get the expected count for each category
5. Calculate test statistic
6. Calculate p - value
7. Form conclusions

χ^2 independence test

1. Set up hypothesis
2. Check conditions
3. Get the observed count for each category
4. **Get the expected count for each category**
5. Calculate test statistic
6. **Calculate p - value - degrees of freedom**
7. Form conclusions

χ^2 independence test

Conditions:

- Independence
- Sample size - each category must have at least 5 expected cases.

χ^2 independence test

H0: There is no correlation in between pizza topping choice and movie choice

HA: There is correlation in between pizza topping choice and movie choice

χ^2 independence test

We need expected count:

$$ExpectedCount_{\text{row } i,\ \text{col } j} = \frac{(\text{ row } i \text{ total }) \times (\text{ column } j \text{ total })}{\text{table total}}$$

χ^2 independence test

	Pepperoni	**Mushrooms**	**Kebab**	**Total**
Kung Fu Panda 4	20 (15.05)	10 (12.25)	5 (7.7)	35
Deadpool & Wolverine	15 (18.06)	12 (14.7)	15 (9.24)	42
Mean Girls	8 (9.89)	13 (8.05)	2 (5.06)	23
Total	43	35	22	

χ^2 independence test

Test statistic - for each category we calculate:

$$Z_n = \frac{\textit{observed count} - \textit{null count}}{\sqrt{\textit{null count}}}$$

then to calculate test statistic:

$$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

χ^2 independence test

	Pepperoni	**Mushrooms**	**Kebab**
Kung Fu Panda 4	1.6280731	0.4132653	0.9467532
Deadpool & Wolverine	0.5184718	0.4959184	3.5906494
Mean Girls	0.3611830	3.0437888	1.8505138

Chi-square independence test

```
(test_statistic <- sum(movies$Pepperoni) +
   sum(movies$Mushrooms) +
   sum(movies$Kebab))
```

```
## [1] 12.84862
```

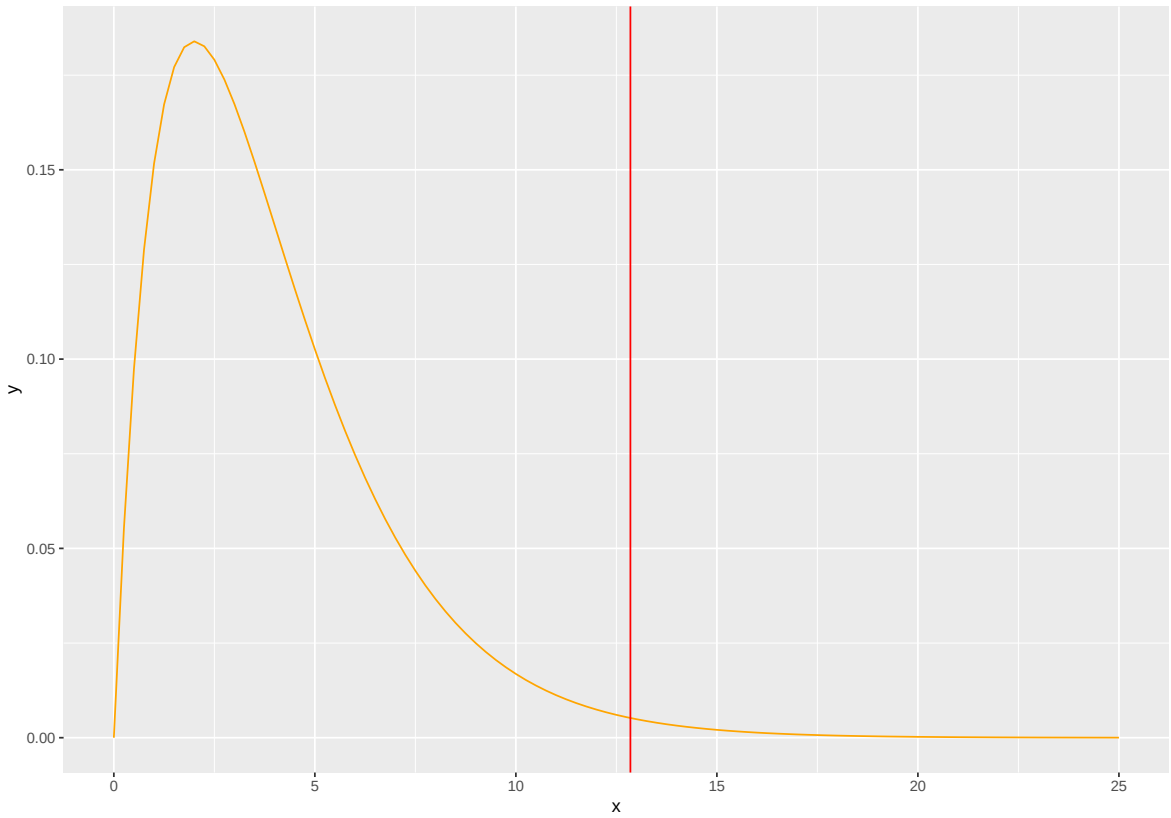
Degrees of Freedom:

$$df = (R - 1) \times (C - 1)$$

$$df = 4$$

Chi-square independence test

Test statistic - 12.8486, df- 4



Chi-square independence test

P-value:

```
(p_value <- 1 - pchisq(test_statistic, df = 4))
```

```
## [1] 0.01203966
```

Conclusions:

We reject null hypothesis in favour of the alternative. There is some correlation in between pizza toppings choice and movie choices.

Question

Who will most likely win the elections in US – Donald Trump or Kamala Harris? You have results of anonymous polls taken from 1000 random citizens from whole U.S. Which test would you use? How would you set up hypothesis? What conditions would you check? How would you determine who is projected to win?

Question

You want to check weather Danish Parliament is a true representation of modern Danish society. You have data from the last census as well as data about parliament members. How would you go about this? Which statistical test would you use? What are the conditions here?

Question

Vaffelhuset Skovsøen recorded for a month each sold scoop of ice cream (flavor of ice cream) as well as if guff was attached to that sale. Now they want to know if guff preference is tied somehow to ice cream flavor? How do you go about checking that?

Question

Lagkagehuset was collecting data from one of their shops for the first quarter of this year. Is there a correlation in between drink choice (coffee, tea, other) and cake (muffin, cream, cookie)? How do you go about checking that?