New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403

New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403
- Lecturers: Lee Fawcett and Eric Ruto

New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403
- Lecturers: Lee Fawcett and Eric Ruto
- **Compulsory** for Stage 1 Marketers, **optional** for Stage 1 Business Management and Marketing & Management

New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403
- Lecturers: Lee Fawcett and Eric Ruto
- **Compulsory** for Stage 1 Marketers, **optional** for Stage 1 Business Management and Marketing & Management
- **Topics:**
  - Hypothesis testing
  - Forecasting and prediction
  - Dynamic modelling
  - Resource allocation

New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403
- Lecturers: Lee Fawcett and Eric Ruto
- **Compulsory** for Stage 1 Marketers, **optional** for Stage 1 Business Management and Marketing & Management
- **Topics:**
  - Hypothesis testing
  - Forecasting and prediction
  - Dynamic modelling
  - Resource allocation
- Students can take this module in second/third year

New stage 2 module – ACE20**: Applied Statistics

- Designed to follow–on from MAS1403
- Lecturers: Lee Fawcett and Eric Ruto
- **Compulsory** for Stage 1 Marketers, **optional** for Stage 1 Business Management and Marketing & Management
- **Topics:**
  - Hypothesis testing
  - Forecasting and prediction
  - Dynamic modelling
  - Resource allocation
- Students can take this module in second/third year
- Not sure about A&F students etc...

Lecture 4

GOODNESS–OF–FIT TESTS

Goodness–of–fit tests are often used to see if our data follow any pattern, or fit a specified probability distribution.

Goodness–of–fit tests are often used to see if our data follow any pattern, or fit a specified probability distribution.

- We compare **observed** frequencies with **expected** frequencies obtained from the hypothesised distribution;

Introduction

Goodness–of–fit tests are often used to see if our data follow any pattern, or fit a specified probability distribution.

- We compare **observed** frequencies with **expected** frequencies obtained from the hypothesised distribution;
- If there is a large difference between the observed and expected frequencies, then this might cast doubt on our hypothesised distribution.

A simple example: traffic accidents

The following data are the number of traffic accidents involving children within two kilometres of schools:

A simple example: traffic accidents

The following data are the number of traffic accidents involving children within two kilometres of schools:

Day	No. of accidents
Monday	23
Tuesday	18
Wednesday	17
Thursday	19
Friday	23

- In the previous chapter, we might have tested the null hypothesis that the population mean number of accidents is equal to, say, 20 per day;

- In the previous chapter, we might have tested the null hypothesis that the population mean number of accidents is equal to, say, 20 per day;
- In the present chapter, we are interested in the **distribution** of the data. We might ask the following questions:

- In the previous chapter, we might have tested the null hypothesis that the population mean number of accidents is equal to, say, 20 per day;
- In the present chapter, we are interested in the **distribution** of the data. We might ask the following questions:
  - "Are there any **patterns** in the data?"

- In the previous chapter, we might have tested the null hypothesis that the population mean number of accidents is equal to, say, 20 per day;
- In the present chapter, we are interested in the **distribution** of the data. We might ask the following questions:
    - "Are there any **patterns** in the data?"
    - "Do the data follow any **theoretical probability distribution**?"

- In the previous chapter, we might have tested the null hypothesis that the population mean number of accidents is equal to, say, 20 per day;
- In the present chapter, we are interested in the **distribution** of the data. We might ask the following questions:
  - "Are there any **patterns** in the data?"
  - "Do the data follow any **theoretical probability distribution**?"
- We can use **goodness–of–fit tests** to answer such questions!

Do traffic accidents occur uniformly?

Do traffic accidents occur uniformly?

- There seems to be more accidents at the start of the week and the end of the week;

Do traffic accidents occur uniformly?

- There seems to be more accidents at the start of the week and the end of the week;
- The number of accidents seems to dip mid–week;

Do traffic accidents occur uniformly?

- There seems to be more accidents at the start of the week and the end of the week;
- The number of accidents seems to dip mid–week;
- But is this pattern **significant**? Is there *really* a dip mid–week, or are these differences just an effect of sampling variation?

Consider the (null) hypothesis that **traffic accidents occur uniformly**.

Consider the (null) hypothesis that **traffic accidents occur uniformly**.

If this were true, then we'd **expect** there to be the same number of accidents each day.

Consider the (null) hypothesis that **traffic accidents occur uniformly**.

If this were true, then we'd **expect** there to be the same number of accidents each day.

So our **expected frequencies** would be

Day	No. of accidents
Monday	20
Tuesday	20
Wednesday	20
Thursday	20
Friday	20

Do you think there is a substantial difference between the **observed frequencies** and the **expected frequencies**?

Do you think there is a substantial difference between the **observed frequencies** and the **expected frequencies**?

- How close do these observed and expected frequencies have to be for us to say that accidents **do** occur uniformly?

Do you think there is a substantial difference between the **observed frequencies** and the **expected frequencies**?

- How close do these observed and expected frequencies have to be for us to say that accidents **do** occur uniformly?
- Or how far apart do they have to be for us to say there **is** a difference in the number of accidents each day?

Do you think there is a substantial difference between the **observed frequencies** and the **expected frequencies**?

- How close do these observed and expected frequencies have to be for us to say that accidents **do** occur uniformly?
- Or how far apart do they have to be for us to say there **is** a difference in the number of accidents each day?
- A **goodness–of–fit test** will help us decide!

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   $$H_0 \quad : \quad \text{Traffic accidents occur uniformly} \qquad \text{or}$$

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   H_0 : Traffic accidents occur uniformly or

   H_0 : Traffic accidents follow a **Poisson distribution**

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   H_0 : Traffic accidents occur uniformly   or

   H_0 : Traffic accidents follow a **Poisson distribution**

2. State the **alternative hypothesis** (H_1)

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   $$\begin{aligned} H_0 &: \text{Traffic accidents occur uniformly} \qquad \text{or} \\ H_0 &: \text{Traffic accidents follow a } \textbf{Poisson distribution} \end{aligned}$$

2. State the **alternative hypothesis** (H_1)

   As usual, this is just the opposite to the null hypothesis, i.e.

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   H_0 : Traffic accidents occur uniformly or

   H_0 : Traffic accidents follow a **Poisson distribution**

2. State the **alternative hypothesis** (H_1)

   As usual, this is just the opposite to the null hypothesis, i.e.

   H_1 : Traffic accidents do *not* occur uniformly or

Basic framework

The basic framework for a goodness–of–fit hypothesis test is the same as that for the tests we looked at in chapters 2 and 3, i.e.

1. State the **null hypothesis** (H_0)

   In the traffic accidents example, this might be

   H_0 : Traffic accidents occur uniformly   or

   H_0 : Traffic accidents follow a **Poisson distribution**

2. State the **alternative hypothesis** (H_1)

   As usual, this is just the opposite to the null hypothesis, i.e.

   H_1 : Traffic accidents do *not* occur uniformly   or

   H_1 : Traffic accidents *do not* follow a Poisson distribution

3. Calculate the **test statistic**

3. Calculate the **test statistic**

In goodness–of–fit tests, this is

$$X^2 = \sum \frac{(O-E)^2}{E},$$

where O and E are the observed and expected frequencies (respectively).

3. Calculate the **test statistic**

In goodness–of–fit tests, this is

$$X^2 \;=\; \sum \frac{(O-E)^2}{E},$$

where O and E are the observed and expected frequencies (respectively).

For a goodness–of–fit test to be valid, all expected frequencies must be ≥ 5; to achieve this, adjacent categories can be "pooled" (see later).

4. Find your **p–value**

4. Find your **p–value**

In goodness–of–fit tests, we use the chi–squared (χ^2) distribution, with ν degrees of freedom, where

4. Find your **p–value**

In goodness–of–fit tests, we use the chi–squared (χ^2) distribution, with ν degrees of freedom, where

$$\begin{aligned}\nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} \\ &\quad -1.\end{aligned}$$

4. Find your **p–value**

In goodness–of–fit tests, we use the chi–squared (χ^2) distribution, with ν degrees of freedom, where

$$\begin{aligned} \nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} \\ &\quad -1. \end{aligned}$$

As before, we compare our test statistic to the **10%**, **5%** and **1%** critical values from the χ^2 distribution to obtain a range for our *p*–value.

5. Reach a **conclusion**

Exactly the same as always – use table 2.1 to form your conclusion!

Back to the traffic accidents example...

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week

Back to the traffic accidents example...

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week

Steps 1 and 2 (hypotheses)

Back to the traffic accidents example...

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week

Steps 1 and 2 (hypotheses)

Our hypotheses are

Back to the traffic accidents example...

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week

Steps 1 and 2 (hypotheses)

Our hypotheses are

H_0 : There are the same number of accidents each day of the week

Back to the traffic accidents example...

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week

Steps 1 and 2 (hypotheses)

Our hypotheses are

$$
\begin{aligned}
H_0 &: \text{There are the same number of accidents each day of the week} \\
H_1 &: \text{There } \textit{aren't} \text{ the same number of accidents each day of the week.}
\end{aligned}
$$

Step 3 (calculating the test statistic)

Step 3 (calculating the test statistic)

This is the hard bit! Remember, the test statistic is

Step 3 (calculating the test statistic)

This is the hard bit! Remember, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

Step 3 (calculating the test statistic)

This is the hard bit! Remember, the test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E}.$$

Drawing up a table usually helps!

Step 3 (calculating the test statistic)

This is the hard bit! Remember, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

Drawing up a table usually helps!

Day	Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
Monday	23	20	0.45
Tuesday	18	20	0.2
Wednesday	17	20	0.45
Thursday	19	20	0.05
Friday	23	20	0.45

So we get

$$\begin{aligned} X^2 &= 0.45 + 0.2 + 0.45 + 0.05 + 0.45 \\ &= 1.6. \end{aligned}$$

So we get

$$\begin{aligned} X^2 &= 0.45 + 0.2 + 0.45 + 0.05 + 0.45 \\ &= 1.6. \end{aligned}$$

Notice we didn't have to pool any categories since all expected frequencies were ≥ 5.

Step 4 (finding the p–value)

Step 4 (finding the p–value)

Using table 4.1, with degrees of freedom

$$\begin{aligned}\nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} - 1 \\ &= 5 - 0 - 1 \\ &= 4,\end{aligned}$$

Step 4 (finding the p–value)

Using table 4.1, with degrees of freedom

$$\begin{aligned}\nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} - 1 \\ &= 5 - 0 - 1 \\ &= 4,\end{aligned}$$

we get the following critical values:

Step 4 (finding the p–value)

Using table 4.1, with degrees of freedom

$$
\begin{aligned}
\nu &= \text{(number of categories after pooling)} \\
&\quad -\text{(number of parameters estimated)} - 1 \\
&= 5 - 0 - 1 \\
&= 4,
\end{aligned}
$$

we get the following critical values:

Significance level	10%	5%	1%
Critical value	7.78	9.49	13.28

Step 4 (finding the p–value)

Using table 4.1, with degrees of freedom

$$\begin{aligned}\nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} - 1 \\ &= 5 - 0 - 1 \\ &= 4,\end{aligned}$$

we get the following critical values:

Significance level	10%	5%	1%
Critical value	7.78	9.49	13.28

Our test statistic lies to the left of the first critical value, and so the p–value is **bigger than 10%**.

Step 5 (conclusion)

Using table 2.1 to interpret our p–value:

Step 5 (conclusion)

Using table 2.1 to interpret our p–value:

- There is **no** evidence against H_0

Step 5 (conclusion)

Using table 2.1 to interpret our p–value:

- There is **no** evidence against H_0
- So we **retain** H_0

Step 5 (conclusion)

Using table 2.1 to interpret our p–value:

- There is **no** evidence against H_0
- So we **retain** H_0
- So accidents **do** occur uniformly! Or at least there's no evidence to suggest otherwise!

Probability distributions

Recall from **semester 1**:

Probability distributions

Recall from **semester 1**:

- A **probability distribution** of a random variable X is the list of all possible values X can take, with their associated probabilities;

Probability distributions

Recall from **semester 1**:

- A **probability distribution** of a random variable X is the list of all possible values X can take, with their associated probabilities;
- for example, consider

Probability distributions

Recall from **semester 1**:

- A **probability distribution** of a random variable X is the list of all possible values X can take, with their associated probabilities;
- for example, consider X: outcome of a roll of a dice.

Probability distributions

Recall from **semester 1**:

- A **probability distribution** of a random variable X is the list of all possible values X can take, with their associated probabilities;
- for example, consider X: outcome of a roll of a dice. The probability distribution is

x	1	2	3	4	5	6
$\Pr(X = x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Probability distributions

Recall from **semester 1**:

- A **probability distribution** of a random variable X is the list of all possible values X can take, with their associated probabilities;
- for example, consider X: outcome of a roll of a dice. The probability distribution is

x	1	2	3	4	5	6
$\Pr(X = x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

- This is a **discrete probability distribution**, since X can only take **integer** values.

The binomial distribution

The binomial distribution

- Used to model the number of “successes” in a series of n independent trials;

The binomial distribution

- Used to model the number of “successes” in a series of n independent trials;
- Each trial has two possible outcomes – “success” or “failure”;

The binomial distribution

- Used to model the number of “successes” in a series of n independent trials;
- Each trial has two possible outcomes – “success” or “failure”;
- The probability of ”success”, p, is constant across trials;

If the previous statements hold true, then we can use a binomial distribution for our data, where

$$\Pr(X = r) \quad = \quad {}^{n}\mathrm{C}_{r} \times p^{r} \times (1 - p)^{n-r}.$$

Here are some examples where we might assume a binomial distribution for our data:

If the previous statements hold true, then we can use a binomial distribution for our data, where

$$\Pr(X = r) \quad = \quad {}^{n}\mathrm{C}_{r} \times p^{r} \times (1 - p)^{n-r}.$$

Here are some examples where we might assume a binomial distribution for our data:

1. 100 students of equal ability sit an exam. The total number passing is recorded;

If the previous statements hold true, then we can use a binomial distribution for our data, where

$$\Pr(X = r) \quad = \quad {}^{n}\mathrm{C}_{r} \times p^{r} \times (1-p)^{n-r}.$$

Here are some examples where we might assume a binomial distribution for our data:

1. 100 students of equal ability sit an exam. The total number passing is recorded;
2. A fair, six–sided dice is rolled 50 times and the number of sixes obtained is recorded;

If the previous statements hold true, then we can use a binomial distribution for our data, where

$$\Pr(X = r) \quad = \quad {}^{n}C_{r} \times p^{r} \times (1 - p)^{n-r}.$$

Here are some examples where we might assume a binomial distribution for our data:

1. 100 students of equal ability sit an exam. The total number passing is recorded;
2. A fair, six–sided dice is rolled 50 times and the number of sixes obtained is recorded;
3. ten barley seeds are sown in a petri dish. The number of germinating seeds is recorded.

The Poisson distribution

- Used to model data which are counts of events in a certain time interval;

The Poisson distribution

- Used to model data which are counts of events in a certain time interval;
- There is usually no fixed upper limit to the value the random variable can take;

The Poisson distribution

- Used to model data which are counts of events in a certain time interval;
- There is usually no fixed upper limit to the value the random variable can take;
- Events occur at a constant rate, λ;

The Poisson distribution

- Used to model data which are counts of events in a certain time interval;
- There is usually no fixed upper limit to the value the random variable can take;
- Events occur at a constant rate, λ;
- Poisson probabilities take the form

$$\Pr(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

Here are some examples where we might assume a Poisson distribution for our data:

Here are some examples where we might assume a Poisson distribution for our data:

1. Number of radioactive emissions per unit of time;

Here are some examples where we might assume a Poisson distribution for our data:

1. Number of radioactive emissions per unit of time;
2. seedlings per unit area;

Here are some examples where we might assume a Poisson distribution for our data:

1. Number of radioactive emissions per unit of time;
2. seedlings per unit area;
3. knots per cubic foot of wood.

A More Complex Example

Consider the following data:

Number of claims	**Observed frequency** (O)
0	144
1	91
2	32
3	11
4	2
$5+$	0
	280

A More Complex Example

Consider the following data:

Number of claims	**Observed frequency** (O)
0	144
1	91
2	32
3	11
4	2
$5+$	0
	280

The data represents the number of small factories in northern England in which industrial injuries resulted in claims for compensation between April 2003 and March 2004.

Which distribution?

Which distribution?

- The data are **discrete**

Which distribution?

- The data are **discrete**
- So do we use the **binomial** or the **Poisson**?

Which distribution?

- The data are **discrete**
- So do we use the **binomial** or the **Poisson**?
- For the binomial distribution, we need n "trials", each with two outcomes... we don't have that set–up here!

Which distribution?

- The data are **discrete**
- So do we use the **binomial** or the **Poisson**?
- For the binomial distribution, we need n “trials”, each with two outcomes... we don’t have that set–up here!
- The Poisson distribution is often used to model ‘count’ data

Recall that the mean of a Poisson random variable is equal to the rate parameter λ, so

Recall that the mean of a Poisson random variable is equal to the rate parameter λ, so

$$\begin{aligned}\lambda &= \frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} \\ &= \frac{196}{280} \\ &= 0.7.\end{aligned}$$

Recall that the mean of a Poisson random variable is equal to the rate parameter λ, so

$$\begin{aligned}\lambda &= \frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} \\ &= \frac{196}{280} \\ &= 0.7.\end{aligned}$$

Now that we have this we can proceed as before:

- The expected probabilities based on the Poisson distribution will be calculated using the Poisson formula

Recall that the mean of a Poisson random variable is equal to the rate parameter λ, so

$$\begin{aligned}\lambda &= \frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} \\ &= \frac{196}{280} \\ &= 0.7.\end{aligned}$$

Now that we have this we can proceed as before:

- The expected probabilities based on the Poisson distribution will be calculated using the Poisson formula
- These can be converted to expected **frequencies** by multiplying by the sample size.

Steps 1 and 2 (hypotheses)

Since we think the Poisson distribution might be an appropriate model for our data, we test

Steps 1 and 2 (hypotheses)

Since we think the Poisson distribution might be an appropriate model for our data, we test

$$H_0 \quad : \quad \text{Claims follow a Poisson distribution} \qquad \text{against}$$

Steps 1 and 2 (hypotheses)

Since we think the Poisson distribution might be an appropriate model for our data, we test

$$\begin{aligned} H_0 &: \text{Claims follow a Poisson distribution} \qquad \text{against} \\ H_1 &: \text{Claims do } \textit{not} \text{ follow a Poisson distribution.} \end{aligned}$$

Step 3 (calculating the test statistic)

Recall that, for goodness–of–fit tests, the test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E}.$$

Step 3 (calculating the test statistic)

Recall that, for goodness–of–fit tests, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

We already have the O's – these are just the observed frequencies. What we need to calculate are the E's (the expected frequencies).

Step 3 (calculating the test statistic)

Recall that, for goodness–of–fit tests, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

We already have the O's – these are just the observed frequencies. What we need to calculate are the E's (the expected frequencies).

From earlier, we know that Poisson probabilities are found using

$$\Pr(X=r) = \frac{e^{-\lambda}\lambda^r}{r!}.$$

Step 3 (calculating the test statistic)

Recall that, for goodness–of–fit tests, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

We already have the O's – these are just the observed frequencies. What we need to calculate are the E's (the expected frequencies).

From earlier, we know that Poisson probabilities are found using

$$\Pr(X=r) = \frac{e^{-\lambda}\lambda^r}{r!}.$$

We have estimated λ as 0.7; thus, we just need to substitute this into the formula to calculate our probabilities for different values of r.

For example, the expected probability of no claims is

$$\begin{aligned}\Pr(X = 0) &= \frac{e^{-0.7} \times 0.7^0}{0!} \\ &= 0.4966.\end{aligned}$$

For example, the expected probability of no claims is

$$\begin{aligned}\Pr(X = 0) &= \frac{e^{-0.7} \times 0.7^0}{0!} \\ &= 0.4966.\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(X = 1) &= \frac{e^{-0.7} \times 0.7^1}{1!} \\ &= 0.3476.\end{aligned}$$

We can do this for **all** our categories and then convert these to **frequencies** by multiplying by the total number of accidents we have observed (280). This gives

We can do this for **all** our categories and then convert these to **frequencies** by multiplying by the total number of accidents we have observed (280). This gives

Number of claims	Expected probability	Expected frequency (E)
0	0.4966	139.048

We can do this for **all** our categories and then convert these to **frequencies** by multiplying by the total number of accidents we have observed (280). This gives

Number of claims	Expected probability	Expected frequency (E)
0	0.4966	139.048
1	0.3476	97.328

We can do this for **all** our categories and then convert these to **frequencies** by multiplying by the total number of accidents we have observed (280). This gives

Number of claims	Expected probability	Expected frequency (E)
0	0.4966	139.048
1	0.3476	97.328
2	0.1217	34.076
3	0.0284	7.952
4	0.0050	1.4
$5+$	0.0007	0.196
		280

For the χ^2 test to be valid, the expected frequencies must be at least 5, so we need to "pool" the last three categories!

For the χ^2 test to be valid, the expected frequencies must be at least 5, so we need to "pool" the last three categories!

Number of claims	Observed frequency (O)	Expected frequency (E)
0	144	139.048
1	91	97.328
2	32	34.076
3+	13	9.548

Now that we have the “O’s” and the “E’s”, we can calculate our test statistic.

Now that we have the "O's" and the "E's", we can calculate our test statistic.

Number of Claims	O	E	$\frac{(O-E)^2}{E}$
0	144	139.048	0.176
1	91	97.328	0.411
2	32	34.076	0.126
3+	13	9.548	1.248

Now that we have the "O's" and the "E's", we can calculate our test statistic.

Number of Claims	O	E	$\frac{(O-E)^2}{E}$
0	144	139.048	0.176
1	91	97.328	0.411
2	32	34.076	0.126
3+	13	9.548	1.248

Thus,

$$\begin{aligned} X^2 &= \sum \frac{(O-E)^2}{E} \\ &= 0.176 + 0.411 + 0.126 + 1.248 \\ &= 1.961. \end{aligned}$$

Step 4 (finding the *p*–value)

We use the χ^2 distribution to obtain our *p*–value. Thus, using table 4.1 with degrees of freedom

Step 4 (finding the p–value)

We use the χ^2 distribution to obtain our p–value. Thus, using table 4.1 with degrees of freedom

$$\begin{aligned}\nu &= \text{(number of categories after pooling)} \\ &\quad -\text{(number of parameters estimated)} - 1 \\ &= 4 - 1 - 1 \\ &= 2,\end{aligned}$$

we obtain the following values:

Step 4 (finding the p–value)

We use the χ^2 distribution to obtain our p–value. Thus, using table 4.1 with degrees of freedom

$$\begin{aligned}\nu &= (\text{number of categories after pooling}) \\ &\quad -(\text{number of parameters estimated}) - 1 \\ &= 4 - 1 - 1 \\ &= 2,\end{aligned}$$

we obtain the following values:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Step 4 (finding the p–value)

We use the χ^2 distribution to obtain our p–value. Thus, using table 4.1 with degrees of freedom

$$\begin{aligned}\nu &= \text{(number of categories after pooling)}\\ &\quad -\text{(number of parameters estimated)} - 1\\ &= 4 - 1 - 1\\ &= 2,\end{aligned}$$

we obtain the following values:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Our test statistic $X^2 = 1.961$ lies to the left of the first critical value, and so our p–value is **bigger than 10%**.

Step 5 (conclusion)

Step 5 (conclusion)

- Using table 2.1, we find that there is **no** evidence against the null hypothesis

Step 5 (conclusion)

- Using table 2.1, we find that there is **no** evidence against the null hypothesis
- Thus we should retain H_0

Step 5 (conclusion)

- Using table 2.1, we find that there is **no** evidence against the null hypothesis
- Thus we should retain H_0
- We can say that it appears that our data **do** follow a Poisson distribution.