Sampling and Hypothesis Testing

LoPSE

Pavel Janda
pavel.janda.early@gmail.com

LoPSE – http://lopsegdansk.blogspot.com
https://www.paveljandaphil.com/

Sampling

Populations and Samples

1. Suppose you want to learn something about a **target population**.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
    1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
   1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.
2. Often, we cannot investigate the whole population, so we need to collect a sample, examine it, and make a conclusion about the population.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
   1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.
2. Often, we cannot investigate the whole population, so we need to collect a sample, examine it, and make a conclusion about the population.
   2.1 The larger the sample is the more accurate our conclusion will be but, most likely, the more expensive the survey will be. We often need to make a compromise between **economy** and **accuracy**.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
   1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.
2. Often, we cannot investigate the whole population, so we need to collect a sample, examine it, and make a conclusion about the population.
   2.1 The larger the sample is the more accurate our conclusion will be but, most likely, the more expensive the survey will be. We often need to make a compromise between **economy** and **accuracy**.
   2.2 **representative sample** – a pattern in the sample implies a similar pattern in the target population.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
   1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.
2. Often, we cannot investigate the whole population, so we need to collect a sample, examine it, and make a conclusion about the population.
   2.1 The larger the sample is the more accurate our conclusion will be but, most likely, the more expensive the survey will be. We often need to make a compromise between **economy** and **accuracy**.
   2.2 **representative sample** – a pattern in the sample implies a similar pattern in the target population.
   2.3 **random selection (sampling)** – and we say that the members of the sample are selected, or chosen, at random – or that they are randomly chosen.

Populations and Samples

1. Suppose you want to learn something about a **target population**.
   1.1 "population" does not necessarily mean people but could mean businesses, schools, governments, etc.
2. Often, we cannot investigate the whole population, so we need to collect a sample, examine it, and make a conclusion about the population.
   2.1 The larger the sample is the more accurate our conclusion will be but, most likely, the more expensive the survey will be. We often need to make a compromise between **economy** and **accuracy**.
   2.2 **representative sample** – a pattern in the sample implies a similar pattern in the target population.
   2.3 **random selection (sampling)** – and we say that the members of the sample are selected, or chosen, at random – or that they are randomly chosen.
   2.4 There are other methods e.g. systematic random sampling, stratified sampling, or cluster sampling

Sample and Population Values

1. The idea – we collect data from a sample (**sample data** or **sample values**). We want to infer back from the sample values to values for members of the target population as a whole (**population values**).

Sample and Population Values

1. The idea – we collect data from a sample (**sample data** or **sample values**). We want to infer back from the sample values to values for members of the target population as a whole (**population values**).
2. **Problem:** any method of choosing the relatively small sample required can produce a sample that is not very representative of the target population.

Sample and Population Values

1. The idea – we collect data from a sample (**sample data** or **sample values**). We want to infer back from the sample values to values for members of the target population as a whole (**population values**).
2. **Problem:** any method of choosing the relatively small sample required can produce a sample that is not very representative of the target population.
3. We usually obtain results from the collection of all possible samples of a fixed size.

Bigger Sample is More Representative

1. Let us have a population of 1000 people

Response	Rating	Number
Much worse off	1	300
Somewhat worse off	2	100
About the same	3	200
Somewhat better off	4	300
Much better off	5	100
Total		1000

Notice that the median is the rating value 3.

Bigger Sample is More Representative

1. Let us have a population of 1000 people

Response	Rating	Number
Much worse off	1	300
Somewhat worse off	2	100
About the same	3	200
Somewhat better off	4	300
Much better off	5	100
Total		1000

   Notice that the median is the rating value 3.

2. Suppose that we choose a very small sample, of size three, from our target population of size 1000. There are 166167000 possible samples of size three.

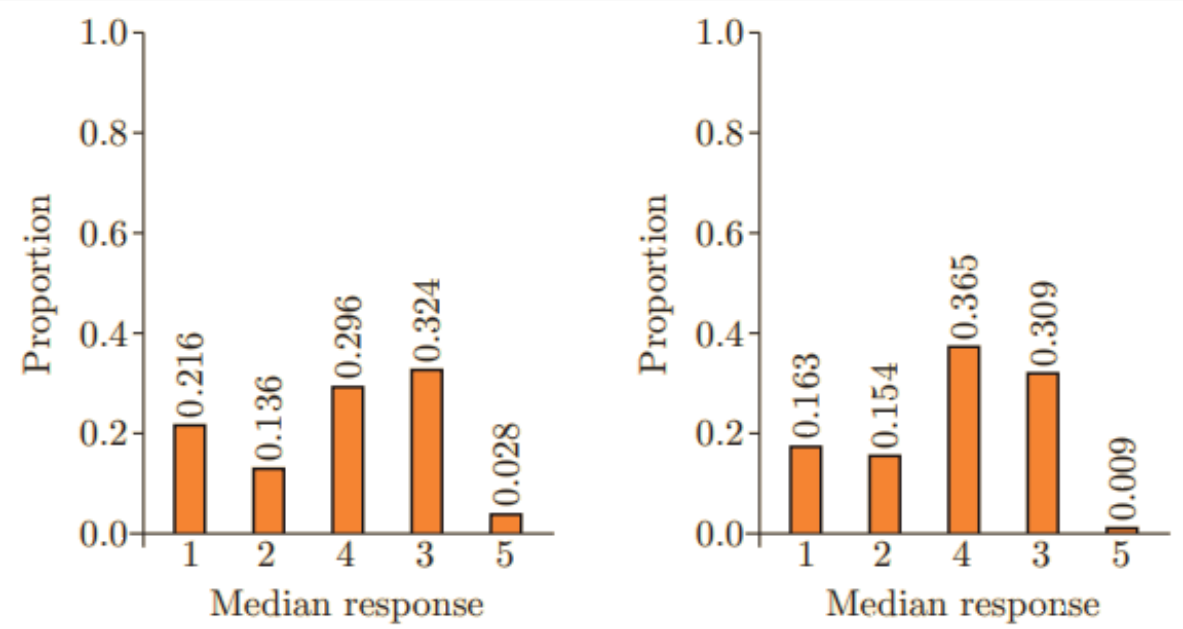
Sample	1st person	2nd person	3rd person
A	4	1	2
B	5	4	1
C	1	4	4

Bigger Sample is More Representative

1. We find rate for each answer in the sample. E.g. the rate for the answer 1 is the fraction $\frac{\text{number of people who asnwered 1}}{1000}$.

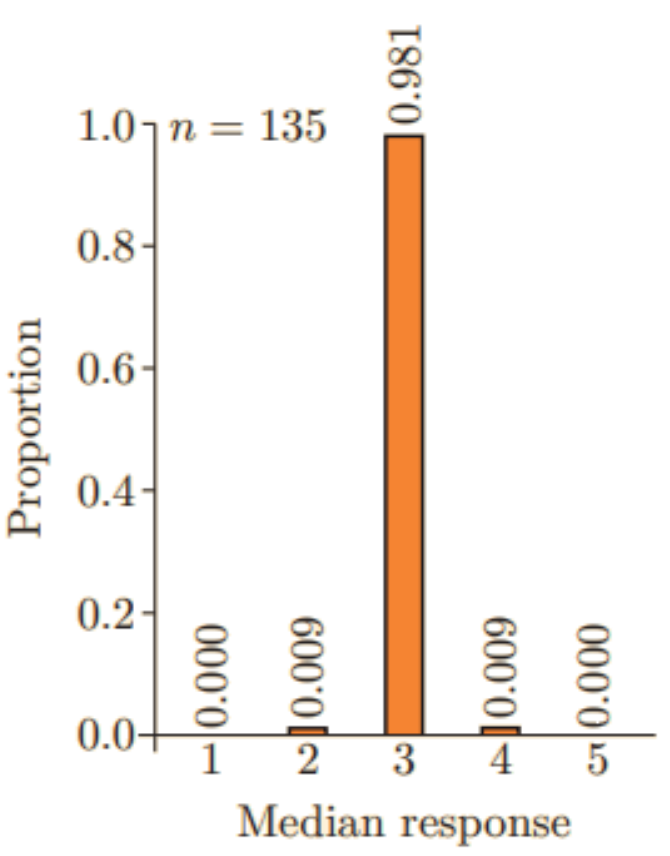
Bigger Sample is More Representative

1. We find rate for each answer in the sample. E.g. the rate for the answer 1 is the fraction $\frac{\text{number of people who asnwered 1}}{1000}$.
2. Suppose we collect two samples. One with the size $n = 3$ (Figure on the left) and the other with $n = 5$ (Figure on the right)



Size of the Sample Matters

1. Suppose we have another, big, sample sample $n = 135$



Hypothesis Testing

Idea

1. We collect samples and use them to test hypothesis. It is best if we consider a concrete example. The sign test is rather for demonstrative purposes. Usually, you would use more sophisticated methods.

Idea

1. We collect samples and use them to test hypothesis. It is best if we consider a concrete example. The sign test is rather for demonstrative purposes. Usually, you would use more sophisticated methods.

Example of Hypothesis Testing

In the example, we looked at data from large schools with over 1000 pupils. The data below show the truancy rates from a random sample of 23 small (secondary) schools (those with fewer than 500 pupils) in the East of England. You are asked to investigate whether the median percentage truancy rate for small schools is the same as the median percentage truancy rate for all schools in the East of England (0.98%).

0.70	0.73	0.16	1.76	0.95	0.80	1.48	0.96	0.64	2.80	0.52	0.96
0.36	0.10	1.21	0.04	0.83	0.64	0.71	0.16	0.71	0.75	0.71	

(a) Write down the hypothesis to be tested.

(b) Record the number of values lying above and the number lying below the assumed median. Hence write down the test statistic.

(c) What is the appropriate critical value at the 5% significance level?

(d) Decide whether you would reject the hypothesis at the 5% significance level.

Solution

1. **(a) Hypothesis**: The median truancy rate for small schools in the East of England is 0.98%.

Solution

1. **(a) Hypothesis**: The median truancy rate for small schools in the East of England is 0.98%.
2. **(b)** Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The **test statistic** is 4, the smaller of these numbers.

Solution

1. **(a) Hypothesis**: The median truancy rate for small schools in the East of England is 0.98%.
2. **(b)** Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The **test statistic** is 4, the smaller of these numbers.
3. **(c)** The critical value for a sample of size 23 at 5% significance level is 6 (it is given – we are now not interested where it comes from). We need it for **(d)**

Solution

1. **(a) Hypothesis**: The median truancy rate for small schools in the East of England is 0.98%.
2. **(b)** Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The **test statistic** is 4, the smaller of these numbers.
3. **(c)** The critical value for a sample of size 23 at 5% significance level is 6 (it is given – we are now not interested where it comes from). We need it for **(d)**
4. **(d)** Since 4 is less than 6, we should reject the hypothesis at the 5% significance level and decide that the median truancy rate for small schools in the East of England is probably not equal to 0.98%.

Solution

1. **(a) Hypothesis**: The median truancy rate for small schools in the East of England is 0.98%.
2. **(b)** Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The **test statistic** is 4, the smaller of these numbers.
3. **(c)** The critical value for a sample of size 23 at 5% significance level is 6 (it is given – we are now not interested where it comes from). We need it for **(d)**
4. **(d)** Since 4 is less than 6, we should reject the hypothesis at the 5% significance level and decide that the median truancy rate for small schools in the East of England is probably not equal to 0.98%.

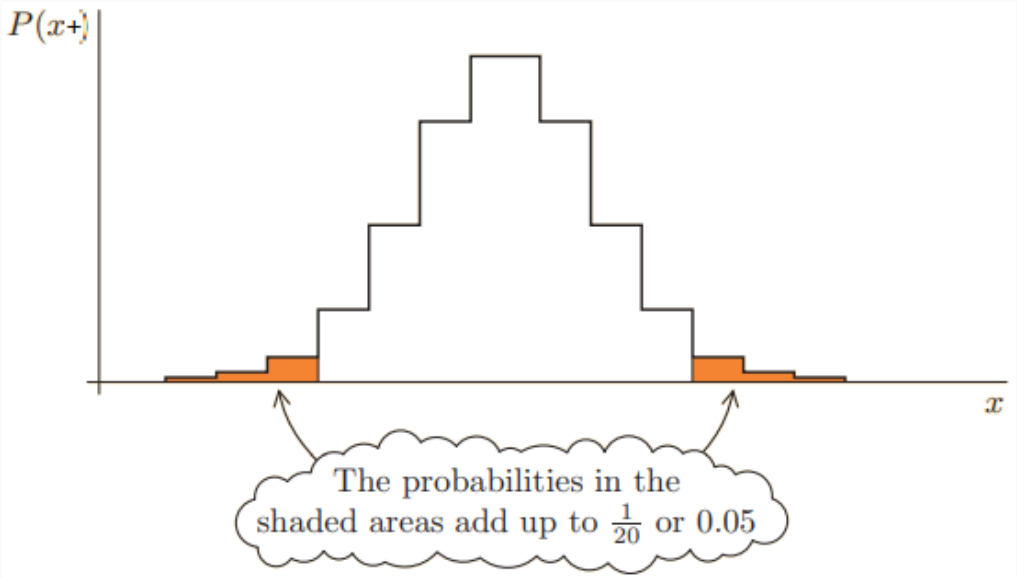
   4.1 If the **test statistic** is less than or equal to the critical value, then we **reject** the hypothesis. Be careful, we **reject or not reject** the hypothesis. **We do not accept a hypothesis or say that is true.**

Significance levels

1. Suppose that our hypothesis about truancy rate 0.98% for small schools in the East of England is true.

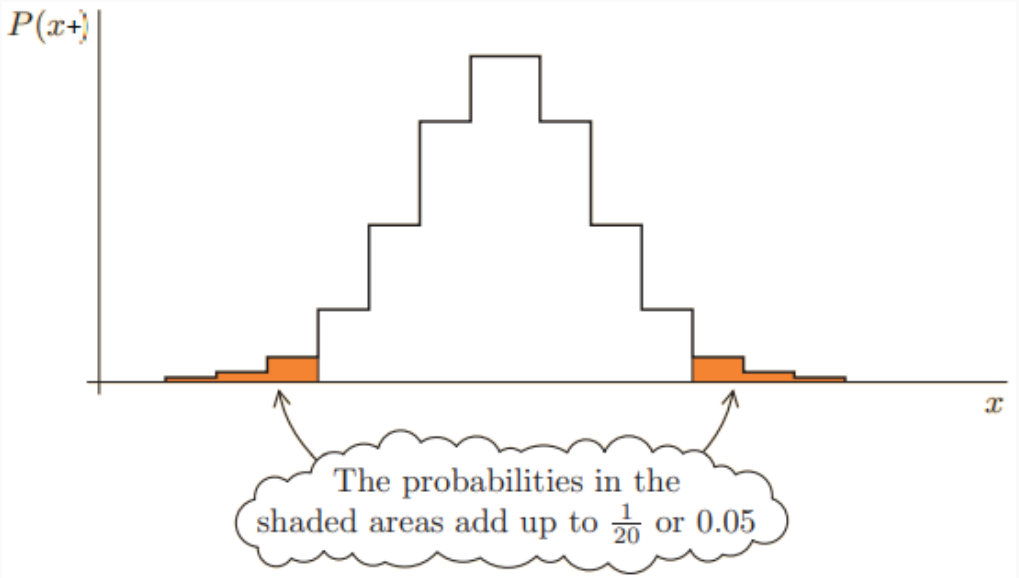
Significance levels

1. Suppose that our hypothesis about truancy rate 0.98% for small schools in the East of England is true.
2. If a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis were true, then we reject the hypothesis at the 5% significance level.

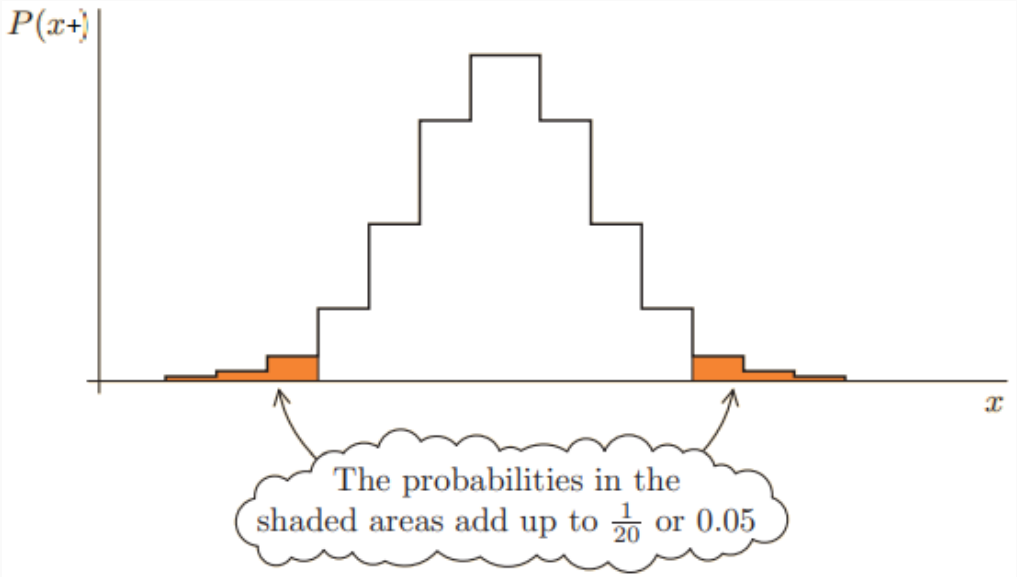


Significance levels

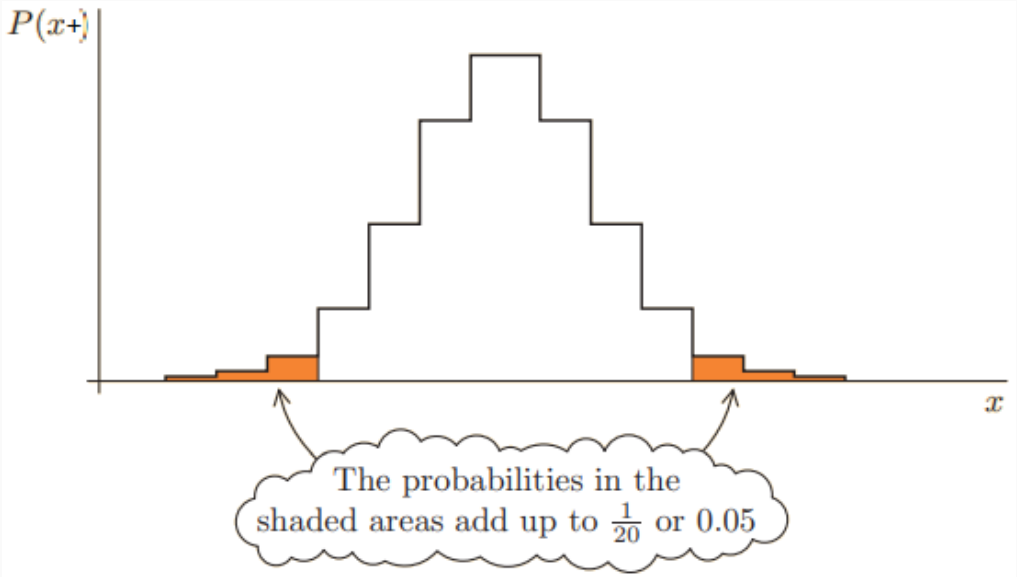
1. Suppose that our hypothesis about truancy rate 0.98% for small schools in the East of England is true.
2. If a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis were true, then we reject the hypothesis at the 5% significance level.
3. We refer to 5% as the significance level of our hypothesis test.



Significance levels

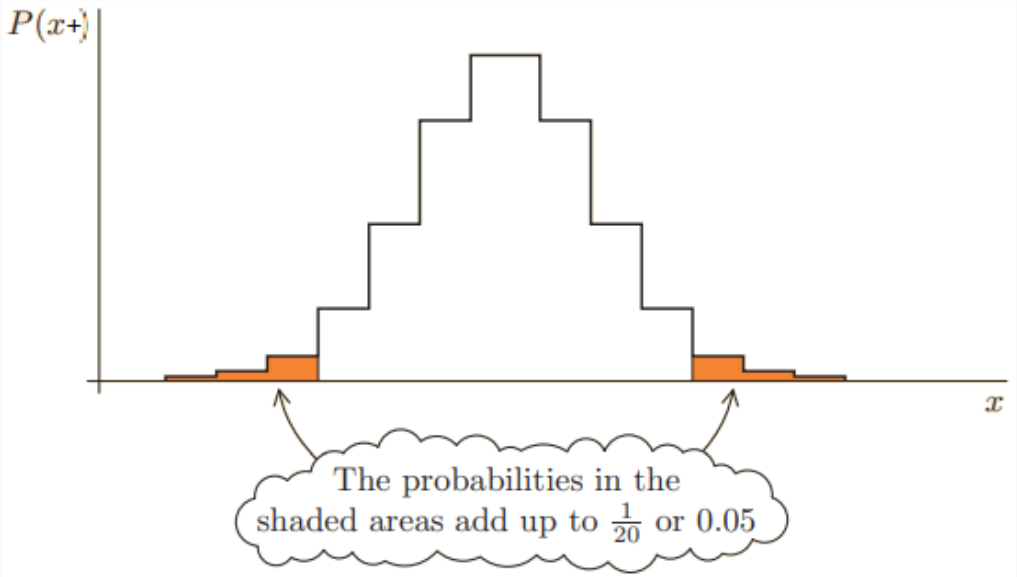


Significance levels



1. The shaded area is called the **critical region** at the 5% significance level – the combined probabilities of the outcomes falling in the region is 0.05 or just less than that.

Significance levels



1. The shaded area is called the **critical region** at the 5% significance level – the combined probabilities of the outcomes falling in the region is 0.05 or just less than that.
2. The critical value at the 5% significance level can be used to determine whether or not an outcome is in the critical region.

p-values

1. **significance probability (p-value)** is the probability of obtaining test results at least as extreme as the results actually observed during the test.

p-values

1. **significance probability (p-value)** is the probability of obtaining test results at least as extreme as the results actually observed during the test.
2. They help us to determine how much evidence we have against our hypothesis

p-values

1. **significance probability (p-value)** is the probability of obtaining test results at least as extreme as the results actually observed during the test.
2. They help us to determine how much evidence we have against our hypothesis
3. A **small p-value** indicates that one of the more unlikely outcomes has occurred or the hypothesis that led to the p-value is false.

p-values

1. **significance probability (p-value)** is the probability of obtaining test results at least as extreme as the results actually observed during the test.
2. They help us to determine how much evidence we have against our hypothesis
3. A **small p-value** indicates that one of the more unlikely outcomes has occurred or the hypothesis that led to the p-value is false.
4. As the p-value decreases, such an outcome becomes less likely and the evidence against the hypothesis increases.

p-values

1. Suppose we want to test that the median truancy rate for large schools in the East of England is 0.98%. We now have the following 12 data points:

p-values

1. Suppose we want to test that the median truancy rate for large schools in the East of England is 0.98%. We now have the following 12 data points:

0.70	0.73	0.16	1.76	0.95	1.48
0.36	0.10	1.21	0.04	0.83	0.71

p-values

1. Suppose we want to test that the median truancy rate for large schools in the East of England is 0.98%. We now have the following 12 data points:

0.70	0.73	0.16	1.76	0.95	1.48
0.36	0.10	1.21	0.04	0.83	0.71

2. What is the p-value?

p-values

1. Suppose we want to test that the median truancy rate for large schools in the East of England is 0.98%. We now have the following 12 data points:

0.70	0.73	0.16	1.76	0.95	1.48
0.36	0.10	1.21	0.04	0.83	0.71

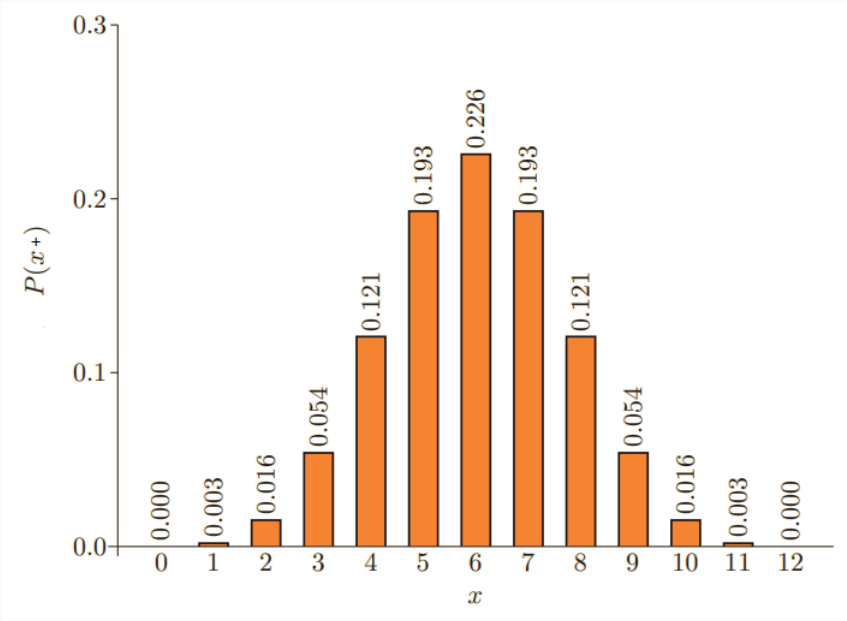
2. What is the p-value?
3. What is the probability of obtaining test results at least as extreme as the results actually observed during the test?

Calculating p-values

1. Our statistics is 3 (only 3 numbers are bellow 0.98%). Below is the probability of the test statistics being above the median.

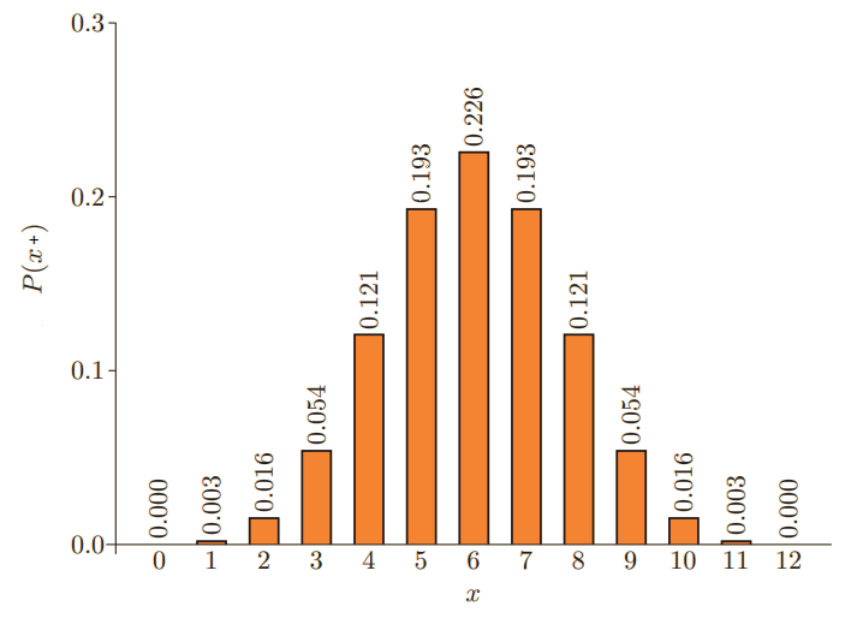
Calculating p-values

1. Our statistics is 3 (only 3 numbers are bellow 0.98%). Below is the probability of the test statistics being above the median.



Calculating p-values

1. Our statistics is 3 (only 3 numbers are bellow 0.98%). Below is the probability of the test statistics being above the median.



$$
\begin{aligned}
&P(0[-]) + P(1[-]) + P(2[-]) + P(3[-]) \\
&+P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) \\
&= 2 \times [P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])] \\
&= 2 \times (0.000 + 0.003 + 0.016 + 0.054) = 0.146.
\end{aligned}
$$

Conclusion

1. Our p-value is 0.146.

Conclusion

1. Our p-value is 0.146.

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

Conclusion

1. Our p-value is 0.146.

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

2. The p-value is 0.146, so there is little evidence against the hypothesis that the median truancy rate in large schools in the East of England is 0.98%.