# Independence tests

Is there a significant difference of the outcome of a variable between two groups?

# Fisher’s exact test

- Fisher's exact test is a statistical significance test used in the analysis of contingency tables where sample sizes are small.
- It is named after its inventor, R. A. Fisher.

< Ronald Aylmer Fisher >

- One of a class of exact tests, so called because the significance of the deviation from a null hypothesis can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

# Purpose and scope

- The test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification.

Assumptions:

- Small sample size
  - Total sample size is less than about 20~40.
- Independent observations
  - Each subject must be independently selected from the population.

# Example

Some soldiers are being trained as parachutists. One rather windy afternoon 55 practice jumps take place at two localities, dropping zone A and dropping zone B. Of 15 men who jump at dropping zone A, five suffer sprained ankles, and of 40 who jump at dropping zone B, two suffer this injury. The casualty rate at dropping zone A seems unduly high, so the medical officer in charge decides to investigate the disparity.

| Table 9.2 Numbers in table 9.1 rearranged for exact probability tes | | | |
|---|---|---|---|
| | **Injured** | **Uninjured** | **Total** |
| Dropping zone A | 2 | 38 | 40 |
| Dropping zone B | 5 | 10 | 15 |
| **Total** | **7** | **48** | **55** |

# Null Hypothesis

- Is it a difference that might be expected by chance?

- The null hypothesis is that there is no difference in the probability of injury generating the proportion of injured men at each dropping zone.

- I.e are the portion of injured the same in both dropping zone?

# Possible table

| Table 9.2 Numbers in table 9.1 rearranged for exact probability tes | | | |
|---|---|---|---|
| | Injured | Uninjured | Total |
| Dropping zone A | 2 | 38 | 40 |
| Dropping zone B | 5 | 10 | 15 |
| **Total** | **7** | **48** | **55** |

The number of possible tables with these marginal totals is eight, that is, the smallest marginal total plus one.

| | | |
|---|---|---|
| 0 | 40 | 40 |
| 7 | 8 | 15 |
| 7 | 48 Set 0 | 55 |

| | | |
|---|---|---|
| 1 | 39 | 40 |
| 6 | 9 | 15 |
| 7 | 48 Set 1 | 55 |

| | | |
|---|---|---|
| 4 | 36 | 40 |
| 3 | 12 | 15 |
| 7 | 48 Set 4 | 55 |

| | | |
|---|---|---|
| 5 | 35 | 40 |
| 2 | 13 | 15 |
| 7 | 48 Set 5 | 55 |

| | | |
|---|---|---|
| 2 | 38 | 40 |
| 5 | 10 | 15 |
| 7 | 48 Set 2 | 55 |

| | | |
|---|---|---|
| 3 | 37 | 40 |
| 4 | 11 | 15 |
| 7 | 48 Set 3 | 55 |

| | | |
|---|---|---|
| 6 | 34 | 40 |
| 1 | 14 | 15 |
| 7 | 48 Set 6 | 55 |

| | | |
|---|---|---|
| 7 | 33 | 40 |
| 0 | 15 | 15 |
| 7 | 48 Set 7 | 55 |

# Calculating Probability

| | Injured | Uninjur | Total |
|---|---|---|---|
| Zone A | a | b | a+b |
| Zone B | c | d | c+d |
| Total | a+c | b+d | n |

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

※ Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution.

| | | |
|---|---|---|
| 0 | 40 | 40 |
| 7 | 8 | 15 |
| 7 | 48 | 55 |

Set 0

$$P_0 = \frac{40!15!7!48!}{55!0!40!7!8!}$$

$$P_1 = P_0 \times \frac{(b_0 \times c_0)}{(a_1 \times d_1)} = 0.0317107 \times \frac{(40 \times 7)}{(1 \times 9)} = 0.9865551$$

$$P_2 = P_0 \times \frac{(b_1 \times c_1)}{(a_2 \times d_2)} = 0.9865551 \times \frac{(39 \times 6)}{(2 \times 10)} = 11.542694$$

| Set | Probability |
|---|---|
| 0 | 0.0000317 |
| 1 | 0.0009866 |
| 2 | 0.0115427 |
| 3 | 0.0664581 |
| 4 | 0.2049126 |
| 5 | 0.3404701 |
| 6 | 0.2837251 |
| 7 | 0.0918729 |
| Total | 0.9999998 |

# Hypergeometric distribution

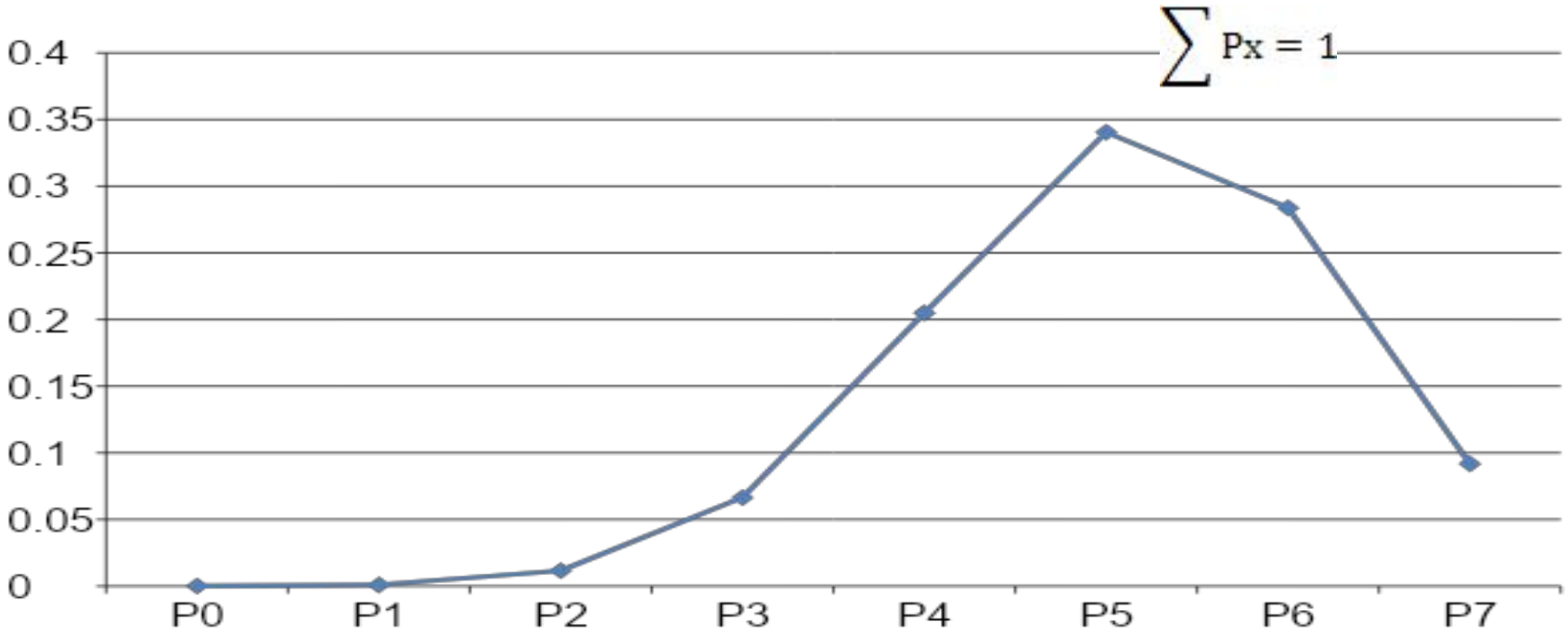
- The hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement, just as the binomial distribution describes the number of successes for draws with replacement.

| | drawn | not drawn | total |
|---|---|---|---|
| successes | $k$ | $m-k$ | $m$ |
| failures | $n-k$ | $N+k-n-m$ | $N-m$ |
| total | $n$ | $N-n$ | $N$ |

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

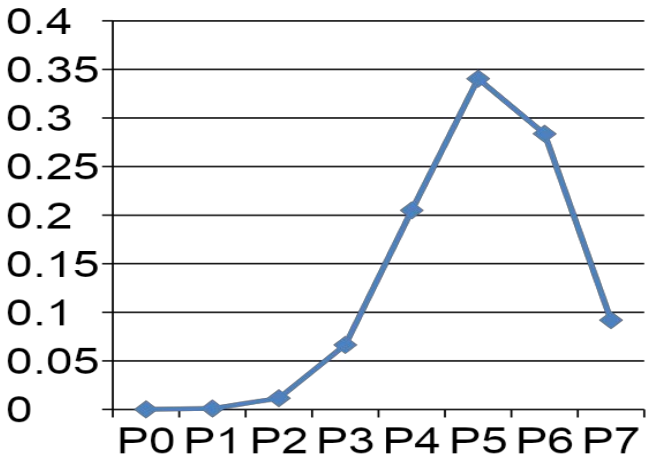
- A random variable X follows the hypergeometric distribution with parameters N, m and n if the probability is given by

# Hypergeometric Distribution



# Calculating P-value

| Set | Probability |
|---|---|
| 0 | 0.0000317 |
| 1 | 0.0009866 |
| 2 | 0.0115427 |
| 3 | 0.0664581 |
| 4 | 0.2049126 |
| 5 | 0.3404701 |
| 6 | 0.2837251 |
| 7 | 0.0918729 |
| Total | 0.9999998 |




- The observed set has a probability of 0.0115427.
- The P value is the probability of getting the observed set, or one more extreme.

- A one tailed P value
  ⇒ 0.0115427 + 0.0009866 + 0.0000317 = 0.01256

- The two tailed value
  ⇒ P = 0.025
  for the conventional(double the one tailed result)
  ⇒ P = 0.0136
  for the mid P approach.

# One tailed or Two tailed

| | | |
|---|---|---|
| 0 | 40 | 40 |
| 7 | 8 | 15 |
| 7 | 48 | 55 |
| | Set 0 | |

| | | |
|---|---|---|
| 1 | 39 | 40 |
| 6 | 9 | 15 |
| 7 | 48 | 55 |
| | Set 1 | |

| | | |
|---|---|---|
| 4 | 36 | 40 |
| 3 | 12 | 15 |
| 7 | 48 | 55 |
| | Set 4 | |

| | | |
|---|---|---|
| 5 | 35 | 40 |
| 2 | 13 | 15 |
| 7 | 48 | 55 |
| | Set 5 | |

More extreme case than observed table

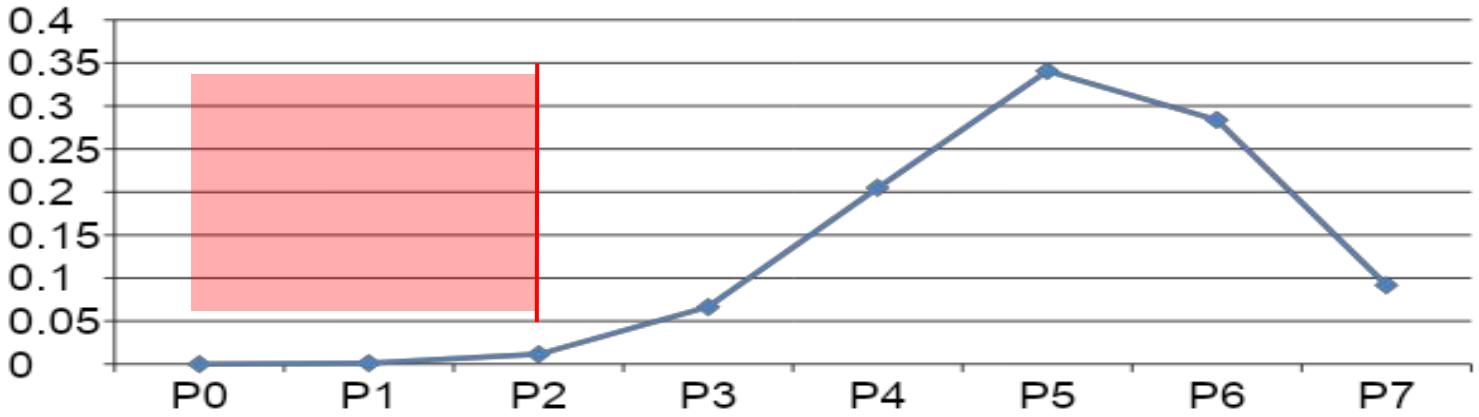
| | | |
|---|---|---|
| 2 | 38 | 40 |
| 5 | 10 | 15 |
| 7 | 48 | 55 |
| | Set 2 | |

Observed table

| | | |
|---|---|---|
| 3 | 37 | 40 |
| 4 | 11 | 15 |
| 7 | 48 | 55 |
| | Set 3 | |

| | | |
|---|---|---|
| 6 | 34 | 40 |
| 1 | 14 | 15 |
| 7 | 48 | 55 |
| | Set 6 | |

| | | |
|---|---|---|
| 7 | 33 | 40 |
| 0 | 15 | 15 |
| 7 | 48 | 55 |
| | Set 7 | |



# Conclusion

- In either case, the P value is less than the conventional 5% level
- The medical officer can conclude that there is a problem in dropping zone A.
- The way to present the results is: Injury rate in dropping zone A was 33%, in dropping zone B 5%; difference 28% (95% confidence interval, P = 0.01256 (Fisher's Exact test - one tail P value).

# Conditional independence tests

# CI-tests

- Fisher’s Z conditional independence test, exact test, categorical variables

- Chi-squared conditional independence test, likelihood, categorical variable, approximate test

- G-squared conditional independence test, likelihood, continues variables

- kernel-based conditional independence test, continues variable, computationally expensive, non-linear

- approximate kernel-based conditional independence test, continues variable, less accurate, computationally feasible