Bioinformatics series

Everything you always wanted to know about statistics*

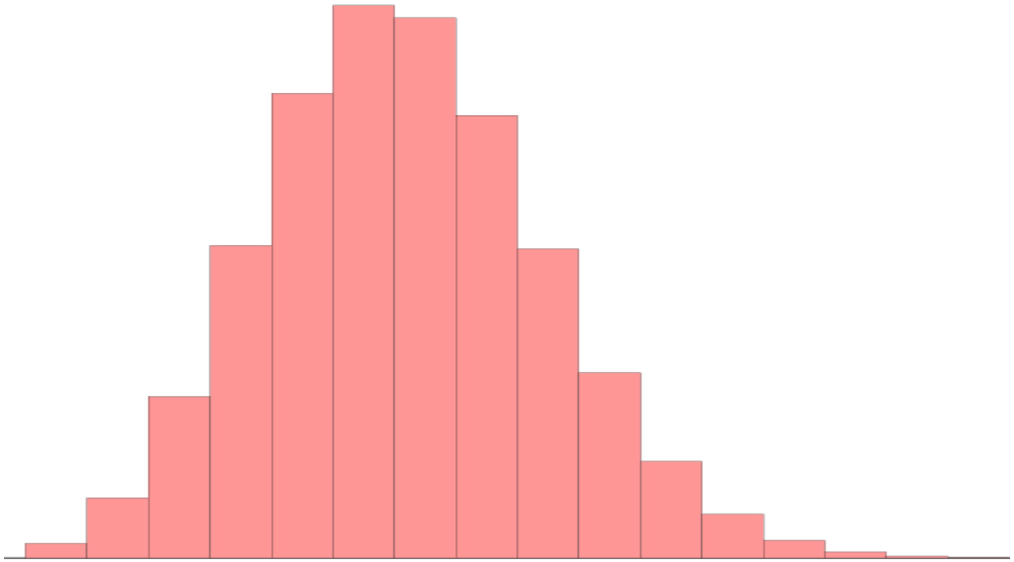
by
Data Analysis Group
Bioinformatics

---

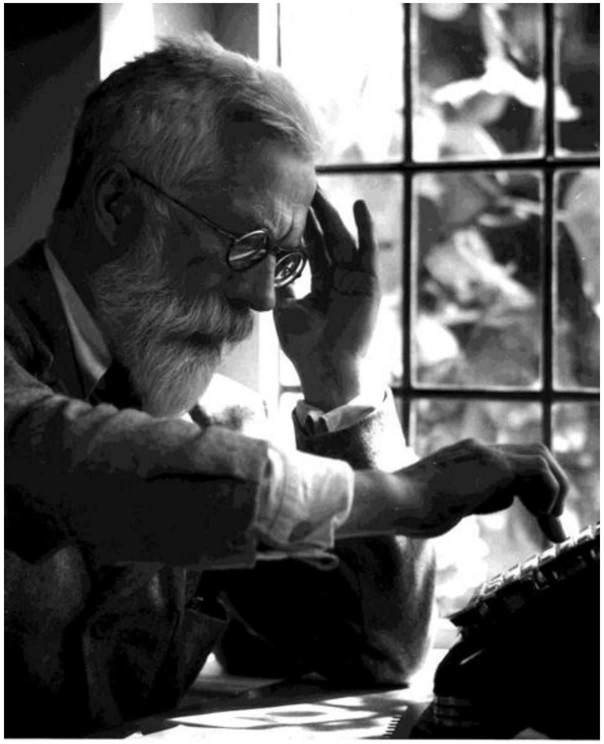
*but were afraid to ask

Tea and independence
or
understanding Fisher’s exact test

Marek Gierliński
Data Analysis Group
Bioinformatics



Ronald Fisher



Sir Ronald Aylmer Fisher
(1890-1962)

Rothamsted Experimental Station
(Hertfordshire)

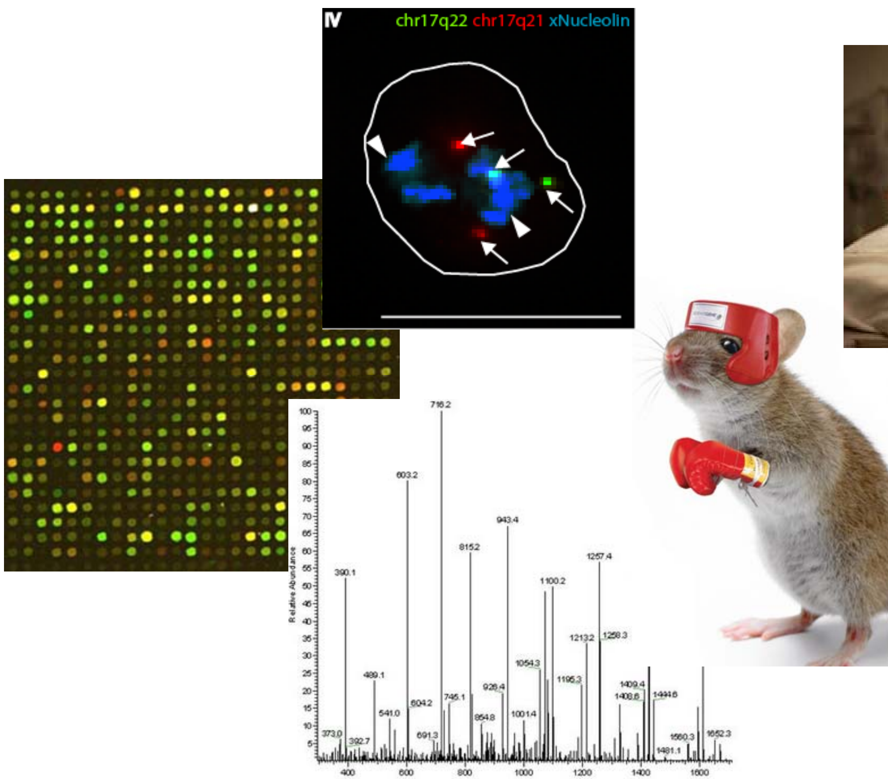
The appreciation of tea

Milk first

Tea first

Typical application

- Dichotomous categorical variable, e.g., success or failure
- Compare observed and estimated success rate
- E.g., enrichment analysis
- Various biological applications: microarray, proteomics, etc.

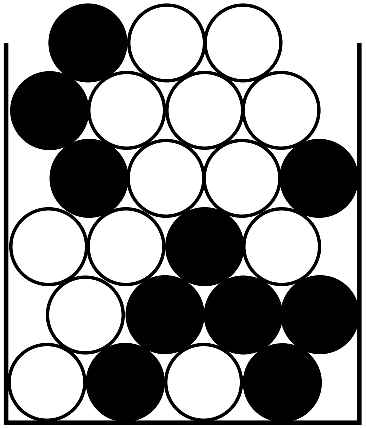


	In cluster	Outside cluster	Total
With GO-term	6	1	7
Without GO-term	38	623	661
Total	44	624	668

Let’s draw some balls

Draw n balls without replacement

removing balls changes probability!



Urn with N balls
m of them white

What is the probability
of finding exactly k white balls?

Binomial coefficient

- “*n* choose *k*”

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- In *combinatorics* it is the number of possible *k*-element subsets of an *n*-element set
- From a 5-element set there are 10 possible 3-element subsets

$$\binom{5}{3} = \frac{5!}{3! \times 2!} = \frac{120}{6 \times 2} = 10$$

- What are your chances of winning the national lottery?

$$\binom{49}{6} = 13{,}983{,}816$$

Set of 5 elements

① ② ③ ④ ⑤

All possible 3-element subsets

① ② ③	① ④ ⑤
① ② ④	② ③ ④
① ② ⑤	② ③ ⑤
① ③ ④	② ④ ⑤
① ③ ⑤	③ ④ ⑤

Count all the possibilities

①②③❹❺ → ○○●

Draw 3 balls. What is the probability of finding exactly 2 whites among them?

All: sets of 3 balls	
①②③	①❹❺
①②❹	②③❹
①②❺	②③❺
①③❹	②❹❺
①③❺	③❹❺

$$N_{\text{all}} = \binom{5}{3} = 10$$

Good: sets of 2 whites and 1 black

2 whites out of 3 / 1 black out of 2

	①②	①③	②③
❹	①②❹	①③❹	②③❹
❺	①②❺	①③❺	②③❺

$$N_{\text{black}} = \binom{2}{1} = 2$$

$$N_{\text{white}} = \binom{3}{2} = 3$$

$$N_{\text{good}} = \binom{2}{1} \times \binom{3}{2} = 6$$

$$P = \frac{N_{\text{good}}}{N_{\text{all}}} = \frac{6}{10} = 0.6$$

Hypergeometric probability

- N balls, m of them white
- Draw n balls
- What is the probability of finding exactly k white balls in the draw?

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

$$= \frac{\begin{pmatrix}\text{Number of ways} \\ \text{for } k \text{ whites}\end{pmatrix}\begin{pmatrix}\text{Number of ways} \\ \text{for } n-k \text{ blacks}\end{pmatrix}}{\begin{pmatrix}\text{Number of ways} \\ \text{for } n \text{ balls}\end{pmatrix}} =$$

$$= \frac{\begin{pmatrix}\text{Number of ways} \\ \text{for } k \text{ whites } \textbf{and } n-k \text{ blacks}\end{pmatrix}}{\begin{pmatrix}\text{Number of ways} \\ \text{for } n \text{ balls}\end{pmatrix}}$$

	Drawn	Not drawn	Total
White	k	$m-k$	m
Black	$n-k$	$N+k-n-m$	$N-m$
Total	n	$N-n$	N

Contingency table

Hypergeometric probability

- 36 balls, 20 of them white
- Draw 10 balls
- What is the probability of finding exactly 8 white balls in the draw?

	Drawn	Not drawn	Total
White	8	12	20
Black	2	14	16
Total	10	26	36

$$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \frac{\binom{20}{8}\binom{16}{2}}{\binom{36}{10}}$$

$$= \frac{\begin{pmatrix}\text{Number of ways}\\ \text{for 8 whites}\end{pmatrix}\begin{pmatrix}\text{Number of ways}\\ \text{for 2 blacks}\end{pmatrix}}{\begin{pmatrix}\text{Number of ways}\\ \text{for 10 balls}\end{pmatrix}} = \frac{125{,}970 \times 120}{254{,}186{,}856} =$$

$$= \frac{\begin{pmatrix}\text{Number of ways}\\ \text{for 8 whites and 2 blacks}\end{pmatrix}}{\begin{pmatrix}\text{Number of ways}\\ \text{for 10 balls}\end{pmatrix}} = \frac{15{,}116{,}400}{254{,}186{,}856} = \underline{\underline{0.059}}$$

Hypergeometric distribution

- If sums are fixed (yellow fields), the cells in the table follow hypergeometric distribution

$$P\begin{bmatrix}0 & 20\\10 & 6\end{bmatrix}=0.000032 \quad P\begin{bmatrix}6 & 14\\4 & 12\end{bmatrix}=0.28$$

$$P\begin{bmatrix}1 & 19\\9 & 7\end{bmatrix}=0.00090 \quad P\begin{bmatrix}7 & 13\\3 & 13\end{bmatrix}=0.17$$

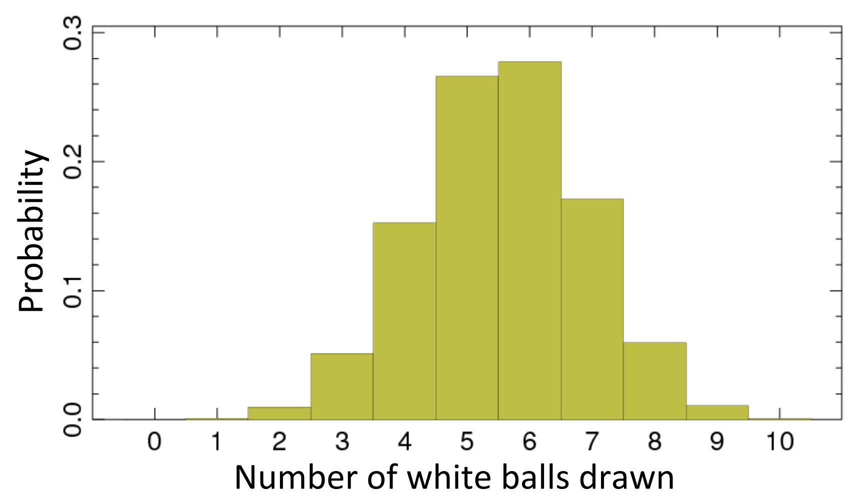
$$P\begin{bmatrix}2 & 18\\8 & 8\end{bmatrix}=0.0096 \quad P\begin{bmatrix}8 & 12\\2 & 14\end{bmatrix}=0.059$$

$$P\begin{bmatrix}3 & 17\\7 & 9\end{bmatrix}=0.051 \quad P\begin{bmatrix}9 & 11\\1 & 15\end{bmatrix}=0.011$$

$$P\begin{bmatrix}4 & 16\\6 & 10\end{bmatrix}=0.15 \quad P\begin{bmatrix}10 & 10\\0 & 16\end{bmatrix}=0.00073$$

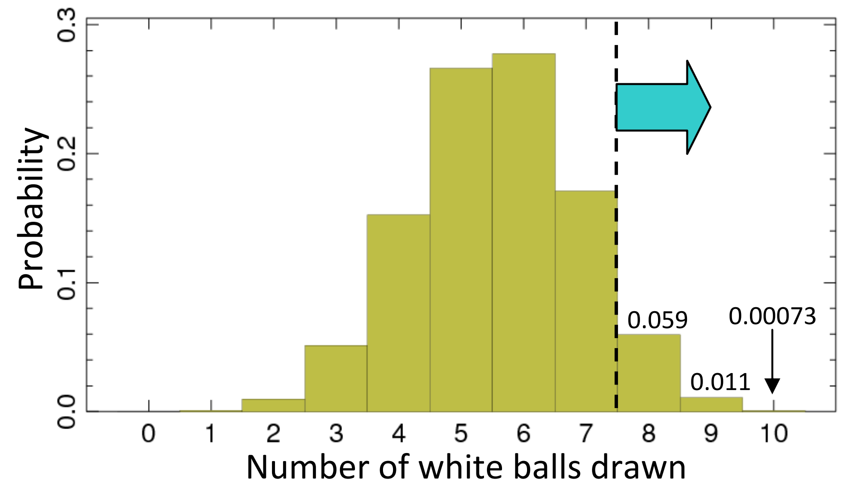
$$P\begin{bmatrix}5 & 15\\5 & 11\end{bmatrix}=0.27$$

	Drawn	Not drawn	Total
White	*a*	*b*	**20**
Black	*c*	*d*	**16**
Total	**10**	**26**	**36**



One-sided test

- Rephrase the question: how unlikely is it to get 8 white balls in a draw?
- Let us consider all cases equally or more extreme
- What is the probability of drawing **8 or more** white balls?
- $P(k \geq 8) = 0.059 + 0.011 + 0.00073 = 0.071$
- *Enrichment*: do we have more than random? (right-sided test)
- *Depletion*: do we have less than random? (left-sided test)



$$P\begin{bmatrix} 8 & 12 \\ 2 & 14 \end{bmatrix} = 0.059$$

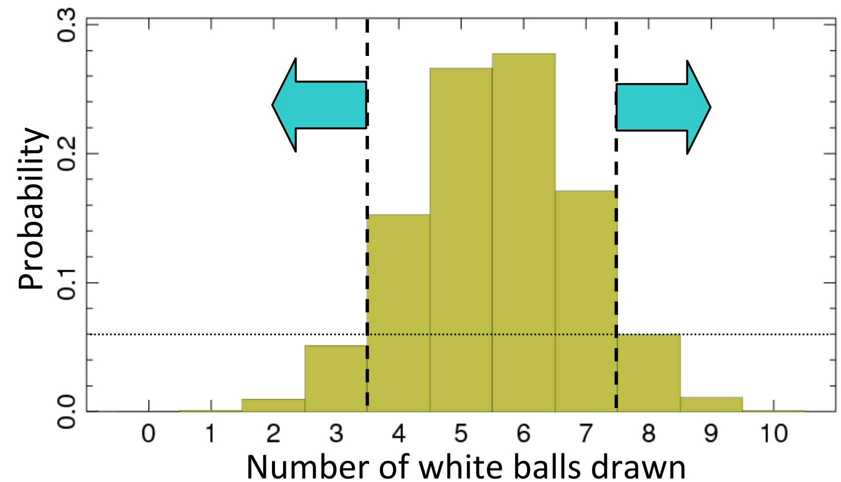
$$P\begin{bmatrix} 9 & 11 \\ 1 & 15 \end{bmatrix} = 0.011$$

$$P\begin{bmatrix} 10 & 10 \\ 0 & 16 \end{bmatrix} = 0.00073$$

Contingency tables considered

Two-sided test

- Two-sided tests ask about any extreme
- Is my result extreme in any way?
- A little practical trick to calculate two-sided probability
- Find all probabilities less or equal $P(k = 8)$ on both sides
- Add them together
- $P(\text{sum}) = 0.000032 + 0.0009 + 0.0096 + 0.051 + 0.059 + 0.011 + 0.00073 = 0.132$



$$P\begin{bmatrix} 0 & 20 \\ 10 & 6 \end{bmatrix} = 0.000032$$

$$P\begin{bmatrix} 1 & 19 \\ 9 & 7 \end{bmatrix} = 0.00090$$

$$P\begin{bmatrix} 2 & 18 \\ 8 & 8 \end{bmatrix} = 0.0096$$

$$P\begin{bmatrix} 3 & 17 \\ 7 & 9 \end{bmatrix} = 0.051$$

$$P\begin{bmatrix} 8 & 12 \\ 2 & 14 \end{bmatrix} = 0.059$$

$$P\begin{bmatrix} 9 & 11 \\ 1 & 15 \end{bmatrix} = 0.011$$

$$P\begin{bmatrix} 10 & 10 \\ 0 & 16 \end{bmatrix} = 0.00073$$

Tea tasting by Muriel Bristol

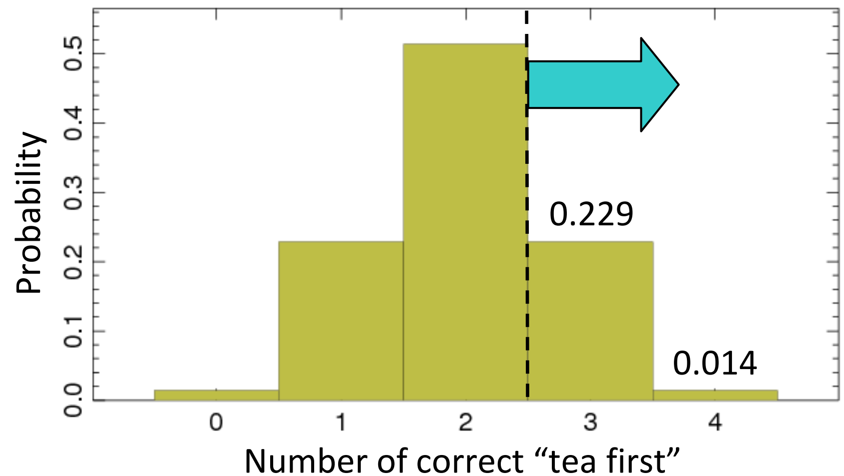
Milk first

Tea first

Tea tasting test

- 8 cups of tea, 4 with tea first, 4 with milk first
- Null hypothesis: Ms Bristol has no ability to tell the difference
- One-sided probability of getting this or more extreme result by chance is $P(k \geq 3) = 0.229 + 0.014 \approx 0.24$
- The null hypothesis cannot be rejected

	Ms Bristol says tea first	Ms Bristol says milk first	Total
Poured tea first	3	1	4
Poured milk first	1	3	4
Total	4	4	8



Test of independence

- Columns and rows in the table, typically groups vs categories
- E.g. treatments vs outcomes
- The *null hypothesis* is that **rows and columns are not associated** and are independent
- If the resulting *p*-value is small, we reject the null hypothesis and conclude that the these are associated
- If sums (yellow fields) are fixed, the numbers in cells follow hypergeometric distribution

Columns

Rows

	Group 1 (Treat. 1)	Group 2 (Treat. 2)	Total
Category 1 (Success)	a	b	$a + b$
Category 2 (Failure)	c	d	$c + d$
Total	$a + c$	$b + d$	$a+b+c+d$

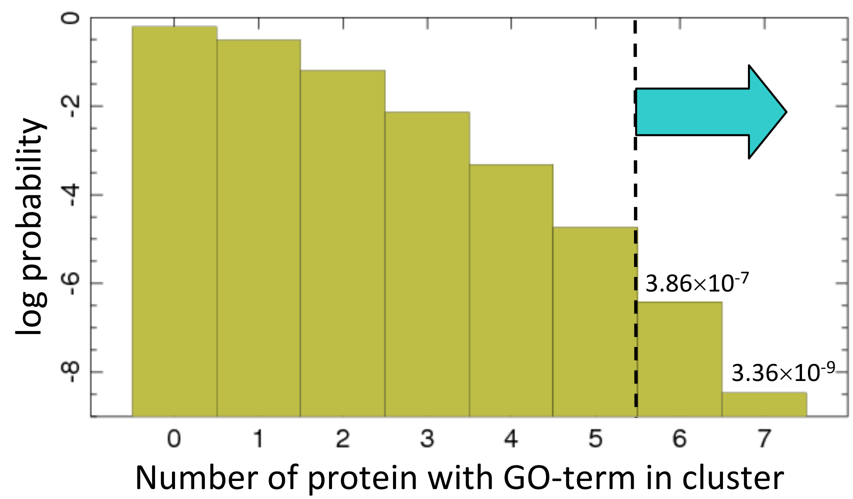
2x2 contingency table

- rows – milk or tea put in first
- columns – Ms Bristol thinks that tea or milk was put in first
- null hypothesis – Ms Bristol's guesses are independent of what was put first (i.e. random)
- *p*-value large – cannot reject null hypothesis

Proteomics example

- There are 668 proteins in an experiment
- 7 of them have an associated Gene Ontology term (GO:00301174, regulation of DNA replication initiation)
- We have a cluster of 44 proteins with similar properties
- 6 of them have this GO term
- Is it significantly enriched?
- $P(k \geq 6) = 3.9\times10^{-7}$

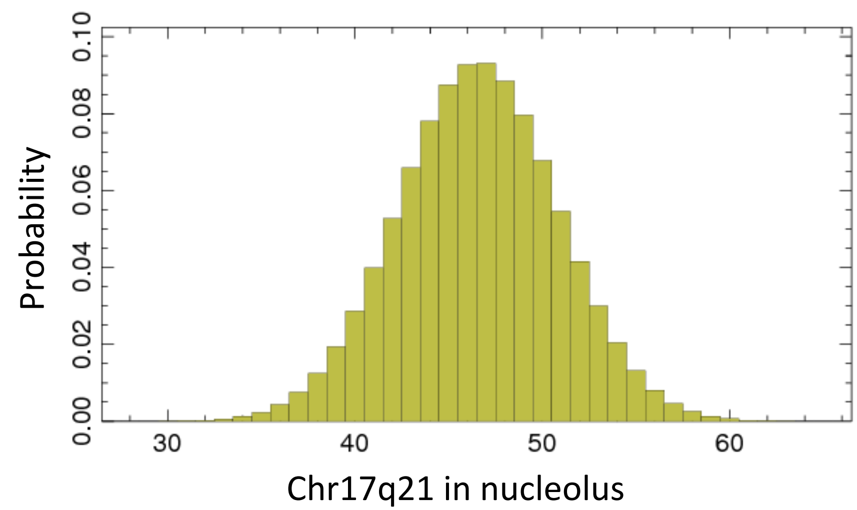
	In cluster	Outside cluster	Total
With GO-term	6	1	7
Without GO-term	38	623	661
Total	44	624	668



In-or-out example

- Using FISH probes
- Observing association of two regions in human chromosome 17 with the nucleolus
- 34 out of 238 (14.3%) probes found in nucleolus for q21
- 56 out of 222 (25.2%) probes found in nucleolus for q22
- Are these significantly different?

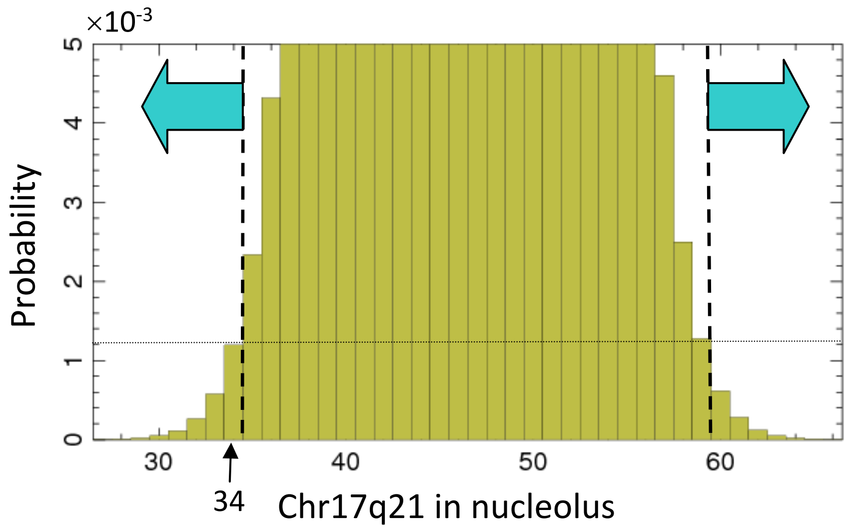
	Chr17q21	Chr17q22	Total
In nucleolus	34	56	90
Outside nucleolus	204	166	370
Total	238	222	460



In-or-out example

- Using FISH probes
- Observing association of two regions in human chromosome 17 with the nucleolus
- 34 out of 238 (14.3%) probes found in nucleolus for q21
- 56 out of 222 (25.2%) probes found in nucleolus for q22
- Are these significantly different?
- No *a priori* assumption about alternative to null hypothesis – either chromosomal locus can be more or less associated with the nucleolus
- Two-sided test: add all the probabilities less or equal $P(k = 34)$
- $P(\text{sum}) = 0.0033$

	Chr17q21	Chr17q22	Total
In nucleolus	34	56	90
Outside nucleolus	204	166	370
Total	238	222	460



Absolute numbers are important!

- A newspaper reports clinical tests on a new drug
- 15% of patients treated with drug A showed improvement
- 30% of patients treated with drug B showed improvement
- So, drug B is 100% better than drug A!

Absolute numbers are important!

- A newspaper reports clinical tests on a new drug
- 15% of patients treated with drug A showed improvement
- 30% of patients treated with drug B showed improvement
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $P(k \leq 3) = 0.306$; insignificant!

	Drug A	Drug B	Total
Improvement	3	3	6
No improvement	17	7	24
Total	20	10	30

p = 0.306

Absolute numbers are important!

- A newspaper reports clinical tests on a new drug
- 15% of patients treated with drug A showed improvement
- 30% of patients treated with drug B showed improvement
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $P(k \leq 3) = 0.306$; insignificant!
- If we had 80 patients in each group and the same proportions, this would be significant, $P(k \leq 12) = 0.018$
- Moral 1: don’t trust newspapers
- Moral 2: estimate the size of your sample before you do your experiment, so the result is more significant

	Drug A	Drug B	Total
Improvement	3	3	6
No improvement	17	7	24
Total	20	10	30

p = 0.306

	Drug A	Drug B	Total
Improvement	12	24	36
No improvement	68	56	124
Total	80	80	160

p = 0.0182

Chi square test

- χ^2 test is an approximation of Fisher's exact test, which works well for large numbers
- Comparing observed (O_{ij}) with expected (E_{ij}) values
- Expected values are $E_{ij} = Np_i p_j$
  - p_i – proportions in row i
  - p_j – proportions in column j
  - N – total number

	Chr17q21	Chr17q22	Total	Proportion
In nucleolus	34	56	90	19.6%
Outside nucleolus	204	166	370	80.4%
Total	238	222	460	
Proportion	51.7%	48.3%		

Chi square test

- χ^2 test is an approximation of Fisher's exact test, which works well for large numbers
- Comparing observed (O_{ij}) with expected (E_{ij}) values
- Expected values are $E_{ij} = N p_i p_j$
  - p_i – proportions in row i
  - p_j – proportions in column j
  - N – total number

	Chr17q21	Chr17q22	Total	Proportion
In nucleolus	34 46.6	56 43.5	90	19.6%
Outside nucleolus	204 191.2	166 178.6	370	80.4%
Total	238	222	460	
Proportion	51.7%	48.3%		

Observed

Estimated
$460 \times 0.804 \times 0.483$
$= 178.6$

Chi square test

- χ^2 test is an approximation of Fisher's exact test, which works well for large numbers
- Comparing observed (O_{ij}) with expected (E_{ij}) values
- Expected values are $E_{ij} = Np_i p_j$
  - p_i – proportions in row i
  - p_j – proportions in column j
  - N – total number

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Then find p-value from χ^2 distribution (tabulated or computed) for 1 degree of freedom
- Corresponds to two-sided Fisher's test

	Chr17q21	Chr17q22	Total	Proportion
In nucleolus	34 46.6	56 43.5	90	19.6%
Outside nucleolus	204 191.2	166 178.6	370	80.4%
Total	238	222	460	
Proportion	51.7%	48.3%		

Observed

Estimated
$460 \times 0.804 \times 0.483$
$= 178.6$

$$\chi^2 = \frac{(34-46.6)^2}{46.6} + \frac{(56-43.5)^2}{43.5} + \frac{(204-191.2)^2}{191.2} + \frac{(166-178.6)^2}{178.6} = 8.74$$

$$p_{\text{chi square}} = 0.0031$$

$$p_{\text{Fisher}} = 0.0033$$

Summary

- Fisher's exact test is typically used when you have
  - two groups of data
  - two categories in which data fall (success or failure)
- Create a 2×2 contingency table
- Calculate one- or two-sided probability (hypergeometric distribution)
- Null hypothesis:
  - rows and columns in the contingency table are independent
  - or, difference between data groups is due to random sampling
  - or, treatments do not affect outcomes
  - or, my sample is not enriched
- *p*-value small: reject hypothesis, there is something special in the data
- *p*-value large: do not reject hypothesis, data are just random
- When you have more than about 100 counts in the table, use chi squared test instead
- Try estimating significance of your result before you do the experiment

This presentation is available at

http://www.compbio.dundee.ac.uk/user/mgierlinski/fisher.pdf

A good Fisher's test and χ^2 test is available at

http://www.quantitativeskills.com/sisa/statistics/fisher.htm