

r x k tables

Population	Blood type				
	A	B	AB	O	
Florida	122	117	19	244	502
Iowa	1781	1351	289	3301	6721
Missouri	353	269	60	713	1395
	2256	1737	367	4258	8618

Question: Same distribution of blood types in each population?

Underlying probabilities

Observed data

	1	2	...	k	
1	n_{11}	n_{12}	...	n_{1k}	n_{1+}
2	n_{21}	n_{22}	...	n_{2k}	n_{2+}
⋮	⋮	⋮	...	⋮	⋮
r	n_{r1}	n_{r2}	...	n_{rk}	n_{r+}
	n_{+1}	n_{+2}	...	n_{+k}	n

Underlying probabilities

	1	2	...	k	
1	p_{11}	p_{12}	...	p_{1k}	p_{1+}
2	p_{21}	p_{22}	...	p_{2k}	p_{2+}
⋮	⋮	⋮	...	⋮	⋮
r	p_{r1}	p_{r2}	...	p_{rk}	p_{r+}
	p_{+1}	p_{+2}	...	p_{+k}	1

$$H_0: p_{ij} = p_{i+} \times p_{+j} \quad \text{for all } i, j$$

Expected counts

Observed data

	A	B	AB	O	
F	122	117	19	244	502
I	1781	1351	289	3301	6721
M	353	269	60	713	1395
	2256	1737	367	4258	8618

Expected counts

	A	B	AB	O	
F	131	101	21	248	502
I	1759	1355	286	3321	6721
M	365	281	59	689	1395
	2256	1737	367	4258	8618

Expected counts, under H_0 : $e_{ij} = n_{i+} \times n_{+j}/n$ for all i,j

χ^2 and LRT statistics

Observed data

	A	B	AB	O	
F	122	117	19	244	502
I	1781	1351	289	3301	6721
M	353	269	60	713	1395
	2256	1737	367	4258	8618

Expected counts

	A	B	AB	O	
F	131	101	21	248	502
I	1759	1355	286	3321	6721
M	365	281	59	689	1395
	2256	1737	367	4258	8618

$$\chi^2 \text{ statistic} = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \dots = 5.64$$

$$\text{LRT statistic} = 2 \times \sum \text{obs} \ln(\text{obs}/\text{exp}) = \dots = 5.55$$

Asymptotic approximation

If the sample size is large, the **null distribution** of the χ^2 and likelihood ratio test statistics will approximately follow a

χ^2 distribution with $(r - 1) \times (k - 1)$ d.f.

In the example, $df = (3 - 1) \times (4 - 1) = 6$

$$X^2 = 5.64 \longrightarrow P = 0.46.$$

$$\text{LRT} = 5.55 \longrightarrow P = 0.48.$$

Fisher's exact test

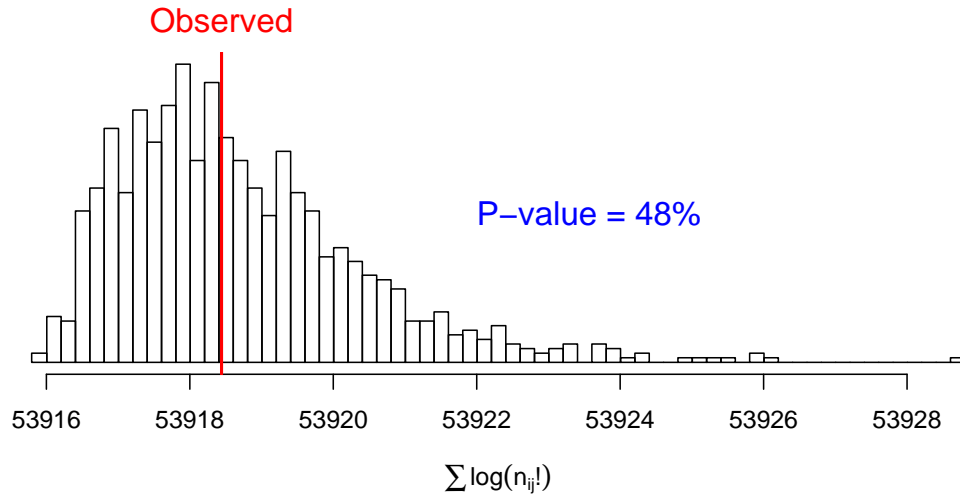
Observed data

	1	2	...	k	
1	n_{11}	n_{12}	\cdots	n_{1k}	n_{1+}
2	n_{21}	n_{22}	\cdots	n_{2k}	n_{2+}
\vdots	\vdots	\vdots	\cdots	\vdots	\vdots
r	n_{r1}	n_{r2}	\cdots	n_{rk}	n_{r+}
	n_{+1}	n_{+2}	\cdots	n_{+k}	n

- Assume H_0 is true.
- Condition on the marginal counts
- Then $\text{Pr}(\text{table}) \propto 1 / \prod_{ij} n_{ij}!$

- Consider all possible tables with the observed marginal counts
- Calculate $\text{Pr}(\text{table})$ for each possible table.
- P-value = the sum of the probabilities for all tables having a probability equal to or smaller than that observed.

Fisher's exact test: The example



Since the number of possible tables can be **very large**, we often must resort to **computer simulation**.

Another example

Survival following treatment in five mouse strains

Strain	Survive	
	No	Yes
A	15	5
B	17	3
C	10	10
D	17	3
E	16	4

Question: Is the survival rate the same for all strains?

Results

Observed			Expected under H_0		
Strain	Survive		Strain	Survive	
	No	Yes		No	Yes
A	15	5	A	15	5
B	17	3	B	15	5
C	10	10	C	15	5
D	17	3	D	15	5
E	16	4	E	15	5

$X^2 = 9.07 \rightarrow P = 5.9\%$ [What is the df?]

$LRT = 8.41 \rightarrow P = 7.8\%$

Fisher's exact test: $P = 8.7\%$

All pairwise comparisons

	N	Y
A	15	5
B	17	3

$\rightarrow P=69\%$

	N	Y
B	17	3
C	10	10

$\rightarrow P=4.1\%$

	N	Y
C	10	10
E	16	4

$\rightarrow P=9.6\%$

	N	Y
A	15	5
C	10	10

$\rightarrow P=19\%$

	N	Y
B	17	3
D	17	3

$\rightarrow P=100\%$

	N	Y
D	17	3
E	16	4

$\rightarrow P=100\%$

	N	Y
A	15	5
D	17	3

$\rightarrow P=69\%$

	N	Y
B	17	3
E	16	4

$\rightarrow P=100\%$

	N	Y
A	15	5
E	16	4

$\rightarrow P=100\%$

	N	Y
C	10	10
D	17	3

$\rightarrow P=4.1\%$

Is this a good thing to do?

Two-locus linkage in an intercross

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Are these two loci linked?

General test of independence

Observed data

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Expected counts

	BB	Bb	bb
AA	4.3	14.4	5.3
Aa	7.9	26.4	9.7
aa	5.8	19.2	7.0

χ^2 test: $X^2 = 10.4 \longrightarrow P = 3.5\% \quad [df = 4]$

LRT test: $LRT = 9.98 \longrightarrow P = 4.1\%$

Fisher's exact test: $P = 4.6\%$

A more specific test

Observed data

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Underlying probabilities

	BB	Bb	bb
AA	$\frac{1}{4}(1 - \theta)^2$	$\frac{1}{2}\theta(1 - \theta)$	$\frac{1}{4}\theta^2$
Aa	$\frac{1}{2}\theta(1 - \theta)$	$\frac{1}{2}[\theta^2 + (1 - \theta)^2]$	$\frac{1}{2}\theta(1 - \theta)$
aa	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1 - \theta)$	$\frac{1}{4}(1 - \theta)^2$

$H_0: \theta = 1/2$ versus $H_a: \theta < 1/2$

→ Use a likelihood ratio test.

- Obtain the general MLE of θ .
- Calculate the LRT statistic = $2 \ln \left\{ \frac{\Pr(\text{data} | \hat{\theta})}{\Pr(\text{data} | \theta=1/2)} \right\}$
- Compare this statistic to a $\chi^2(\text{df} = 1)$.

Results

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

MLE: $\hat{\theta} = 0.359$

LRT statistic: $\text{LRT} = 7.74 \rightarrow P = 0.54\% \quad [\text{df} = 1]$

- Here we assume Mendelian segregation, and that deviation from H_0 is “in a particular direction.”
- If these assumptions are correct, we’ll have greater power to detect linkage using this more specific approach.

Sample size determination

→ We seek to demonstrate that strains A and B differ in their survival rates following treatment.

How many mice from each group to study?

Generally, our goal is to have 80% power to detect a “meaningful” difference.

Power depends on...

- Structure of the experiment
- Method of analysis
- Sample size
- Chosen significance level (α)
- **The underlying truth**

We usually seek to determine the sample size that will give us 80% power to detect the smallest difference that we consider meaningful.

Calculating power

To determine power, we need:

1. The **null distribution** of the test statistic (so that we can determine the appropriate critical value).
2. The distribution of the test statistic **under the alternative hypothesis**.

For the t-test, there were analytical formulas for these.

For testing independence in a 2 x 2 table, we must resort to computer simulation.

Power in 2 x 2 tables

Suppose we assay 20 individuals from each strain.

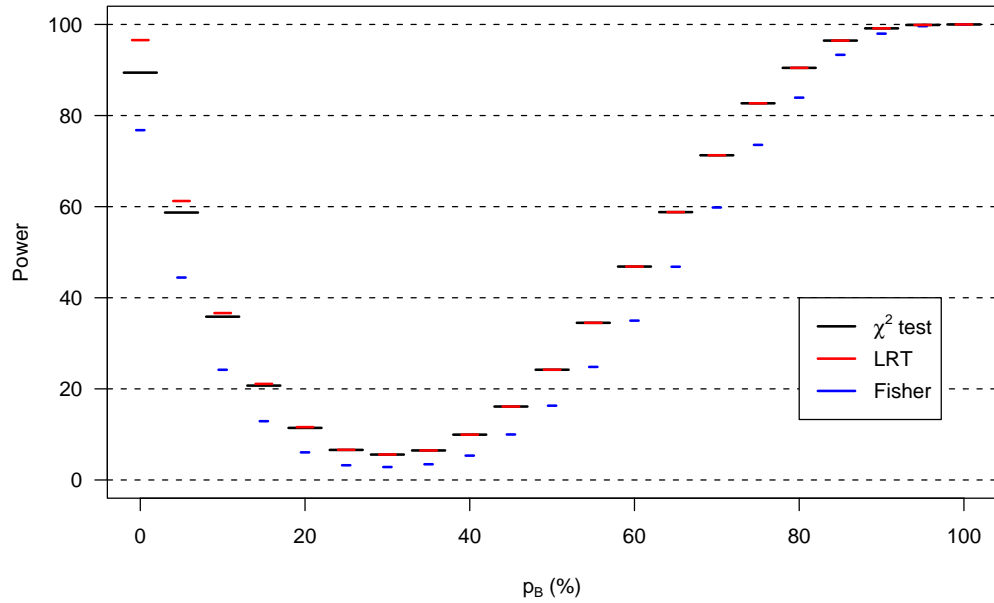
Let $p_A = \Pr(\text{survive treatment} \mid \text{strain A})$ and
 $p_B = \Pr(\text{survive treatment} \mid \text{strain B})$.

To estimate power:

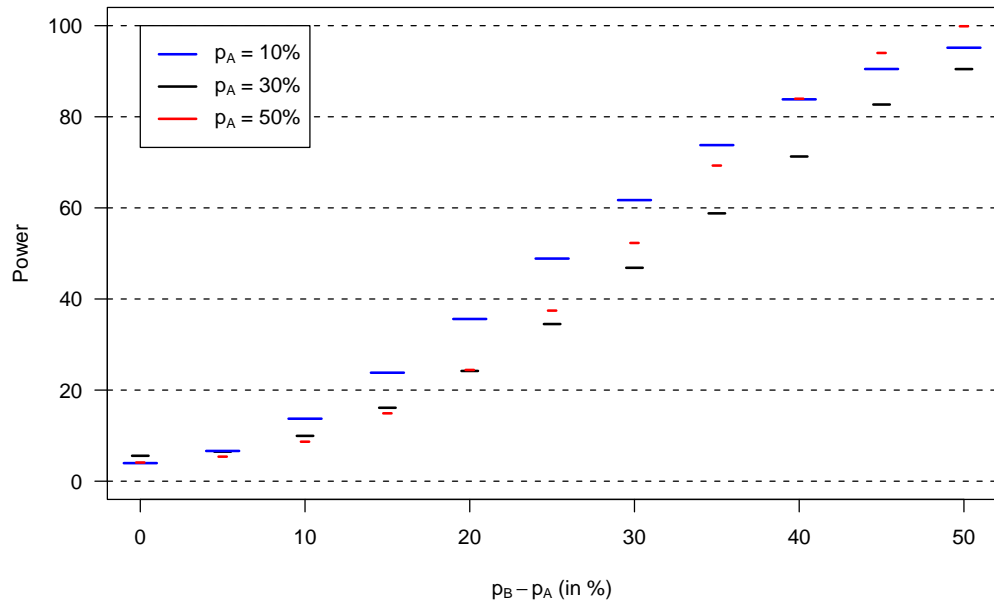
1. Simulate data for some specified p_A and p_B .
2. Calculate the chosen test statistic.
3. Calculate the corresponding P-value.
4. Repeat 1–3 many times (say 250).
5. The estimated power = prop'n of P-values ≤ 0.05

Power in 2 x 2 tables

The case $n=20$ per group and $p_A = 30\%$.
[results based on 10,000 simulations]

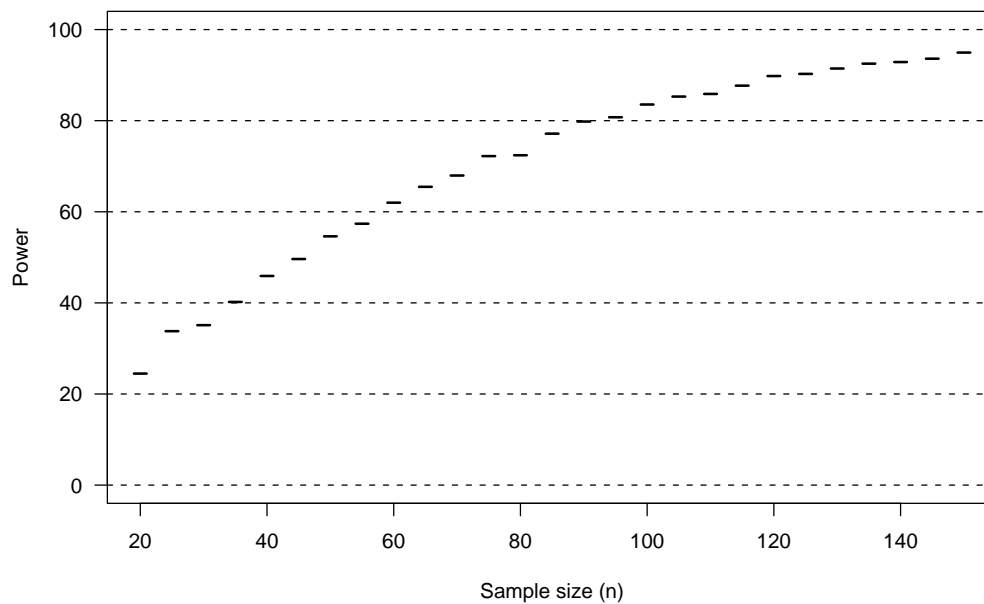


Power of χ^2 test



To get the sample size...

Results χ^2 test for $p_A = 30\%$ and $p_B = 50\%$.



Notes

- There are formulas available for all sorts of different statistical tests and experimental situations.
- Simulations are time-consuming (and require programming), but can be used in virtually any situation.
- 250 simulation replicates is usually enough to get a good estimate of power, but for making power comparisons between different statistical methods, many more replicates are often necessary.
- **Power** is an important criterion in choosing between different statistical tests (such as the χ^2 test versus Fisher's exact test).

Another example

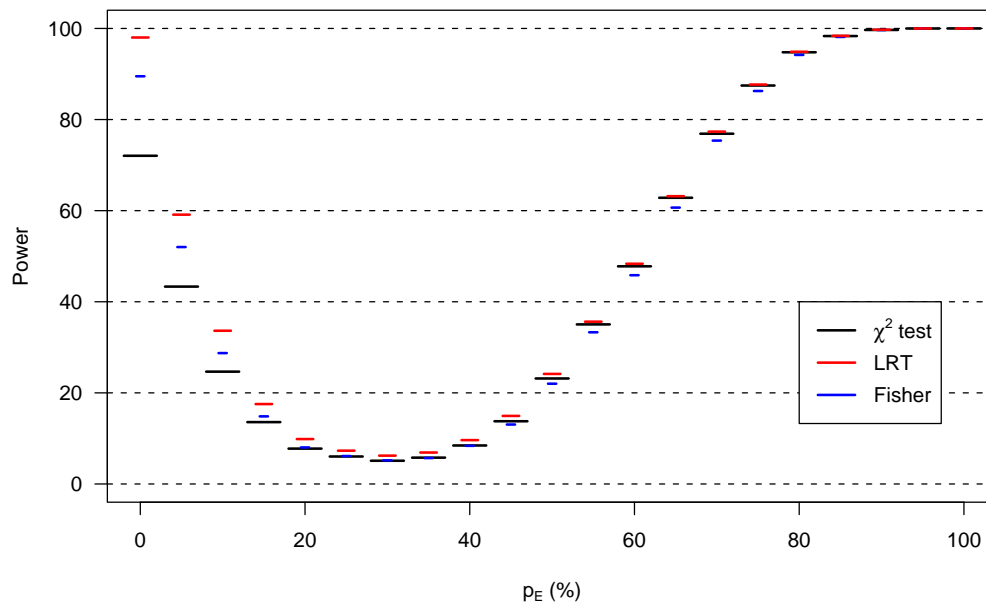
	N	Y
A		
B		
C		
D		
E		

- Survival following treatment in 5 mouse strains.
- Seek to demonstrate that the strains differ.
- Power for the case of 20 individuals per strain?

- We might focus on the case that strains A–D are the same, but strain E is different (the worst possible case).
- We must then specify Let $p_A = \Pr(\text{survive treatment} \mid \text{strain A})$ and $p_E = \Pr(\text{survive treatment} \mid \text{strain E})$.

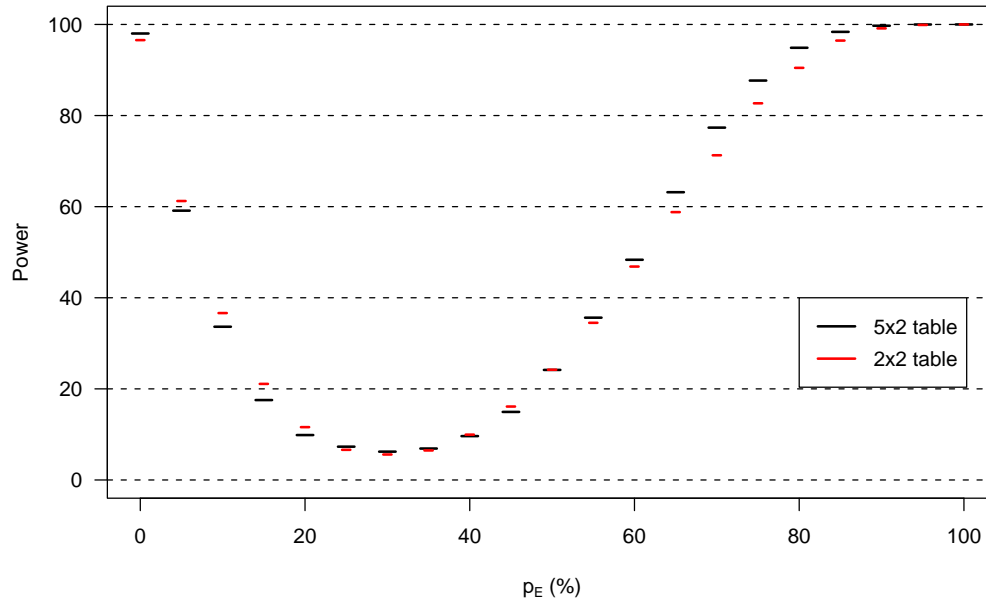
Power for this example

The case $n=20$ per group, and $p_A = p_B = p_C = p_D = 30\%$.



Comparison to 2 x 2 table

Comparing all 5 strains versus comparing just strains A and E.
(Considering just the likelihood ratio test.)



Final points

- Assumptions underlying tests in contingency tables:
 1. Data are a random sample from some population or populations.
 - Two or more independent samples observed with respect to one variable
 - One random sample observed with respect to two variables.
 2. Observations within a sample are independent.
- Ordinal data may require different techniques

	None	Some	A lot
A			
B			