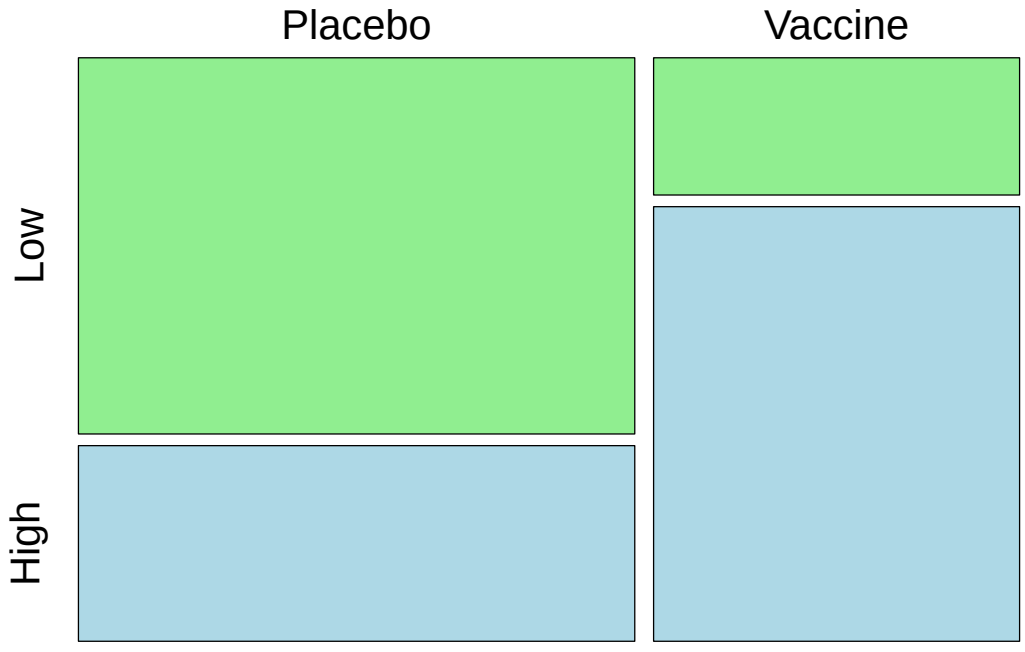
Lecture 20

STA 100 | A. Farris | Spring 2020

E.g. A Flu vaccine trial assigned subjects the trial vaccine or a placebo. After 6 weeks, titre levels of hemagglutinin inhibiting antibodys were recorded as 'low' or 'high'.

A 2x2 contingency table:

	Low	High
Placebo	25	13
Vaccine	6	19


Placebo
Vaccine
Low
High

Population version:

	Low	High
Placebo	p_{11}	p_{12}
Vaccine	p_{21}	p_{22}

To evaluate a claim like "$p_{11} = 0.2, p_{12} = 0.2, p_{21} = 0.3, p_{22} = 0.3$", use a test of goodness of fit.

What about a claim like “having a placebo is independent of having high antibody levels?”

	Low	High	Overall
Placebo	p_{11}	p_{12}	$p_{11} + p_{12}$
Vaccine	p_{21}	p_{22}	$p_{21} + p_{22}$
Overall	$p_{11} + p_{21}$	$p_{12} + p_{22}$	1

	Low	High	Overall
P:	P(Placebo and Low)	P(Placebo and High)	P(Placebo)
V:	P(Vaccine and Low)	P(Vaccine and High)	P(Vaccine)
O:	P(Low)	P(High)	1

Independence:

$$P(\text{Vaccine and High}) = P(\text{Vaccine})P(\text{High})$$

I.e.

$$p_{22} = (p_{21} + p_{22})(p_{12} + p_{22})$$

Back to the sample result:

	Low	High
Placebo	25	13
Vaccine	6	19

What would we expect this to look like under independence?

Given the row and column totals:

	Low	High	Total
Placebo	25	13	38
Vaccine	6	19	25
Total	31	32	63

Estimated proportions:

	Low	High	Total
Placebo	25/63	13/63	38/63
Vaccine	6/63	19/63	25/63
Total	31/63	32/63	63/63

Given the row and column totals, the expeced counts under independence:

	Low	High	Total
Placebo	$38 \cdot (31/63)$	$38 \cdot (32/63)$	38
Vaccine	$25 \cdot (31/63)$	$25 \cdot (32/63)$	25
Total	$31/63$	$32/63$	

Given the row and column totals, the expeced counts under independence:

	Low	High	Total
Placebo	18.7	19.3	38
Vaccine	12.3	12.7	25
Total	31/63	32/63	

Given the row and column totals:

	Low	High	Total
Placebo	X_{11}	X_{12}	38
Vaccine	X_{21}	X_{22}	25
Total	31	32	63

Given the row and column totals:

	Low	High	Total
Placebo	X_{11}	X_{12}	38
Vaccine	X_{21}	X_{22}	25
Total	31	32	63

If Vaccine and High are independent, then X_{22} has a Hypergeometric distribution with $N = 63, M = 25$, and $n = 32$

Given the row and column totals:

	Low	High	Total
Placebo	X_{11}	X_{12}	38
Vaccine	X_{21}	X_{22}	25
Total	31	32	63

If Vaccine and High are independent, then
$P(X_{22} = 19) \approx 0.00105$

```
dhyper(19, 25, 63-25, 32)
```

```
[1] 0.001046573
```

Given the row and column totals:

	Low	High	Total
Placebo	X_{11}	X_{12}	38
Vaccine	X_{21}	X_{22}	25
Total	31	32	63

If Vaccine and High are independent, then
$P(X_{22} \geq 19) \approx 0.00122$

```
sum(dhyper(19:25, 25, 63-25, 32))
```

```
[1] 0.001221413
```

Fisher's exact test for association between factors ("A" and "B") in a 2x2 contingency table

Fisher's exact test

Null hypothesis: the data is sampled from a population in which A and B are independent

Alternative hypotheses:

- the two categories are positively associated ($P(A|B) > P(A)$) (one-sided)
- the two categories are negatively associated ($P(A|B) < P(A)$) (one-sided)
- the two categories are not associated ($P(A|B) \neq P(A)$) (two-sided)

Fisher's exact test

Test statistic: the number of subjects satisfying both criteria

Null distribution: Hypergeometric(N, M, n) (where N = total subjects, M = row total for the statistic, n = column total for the statistic)

For a test of

H_0 : "Vaccine" and "High antibodies" are independent, i.e.
$P(\text{High antibodies}|\text{Vaccine}) = P(\text{High antibodies})$

vs.

H_A : "High antibodies" are more likely given "Vaccine" (one-sided),
i.e. $P(\text{High antibodies}|\text{Vaccine}) > P(\text{High antibodies})$

given data

	Low	High	Total
Placebo	25	13	38
Vaccine	6	19	25
Total	31	32	63

p-value:

```
sum(dhyper(19:25,25,63-25,32))
```

```
[1] 0.001221413
```

Or:

```
VaccTable <- matrix(c(25,6,13,19),2,2)
fisher.test(VaccTable, alternative="greater")
```

```
    Fisher's Exact Test for Count Data

data:  VaccTable
p-value = 0.001221
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 2.048572      Inf
sample estimates:
odds ratio
  5.898936
```

Or:

```
VaccTable <- matrix(c(25,6,13,19),2,2)
fisher.test(VaccTable, alternative="greater")$p.value
```

```
[1] 0.001221413
```

We can carry out a two-sided test as well:

```
VaccTable <- matrix(c(25,6,13,19),2,2)
fisher.test(VaccTable)$p.value
```

```
[1] 0.001821423
```

Can we prove independence using Fisher's exact test?

E.g. Fisher's tea tasting experiment

	Milk first	Tea first
Guessed Milk	3	1
Guessed Tea	1	3

We can use Fisher's Exact Test for independence to verify that there is evidence to check whether there is evidence that the proportion of positive results truly differs between the two studies.

I.e. $H_0 : P(\text{Guessed milk} \mid \text{Milk first}) = P(\text{Guessed milk})$

against $H_A : P(\text{Guessed milk} \mid \text{Milk first}) \neq P(\text{Guessed milk})$

```
TeaTable <- matrix(c(3,1,1,3),2,2)
fisher.test(TeaTable)$p.value
```

```
[1] 0.4857143
```

against $H_A : P(\text{Guessed milk} \mid \text{Milk first}) > P(\text{Guessed milk})$

```
fisher.test(TeaTable, alternative = "greater")$p.value
```

```
[1] 0.2428571
```

For larger (r x c) contingency tables, use Chi squared test for independence

- Works for $r \geq 2, c \geq 2$
- Uses a large sample approximation for the Null distribution (E.g. expected count in each cell ≥ 5)