Introductory Statistics
6: Chi-squared (χ^2) test

Richard Veale

Graduate School of Medicine
Kyoto University

https://youtu.be/EFqDh4_Z6so

Lecture Video at above link

Summary

1) Big Data: Coronary Heart Disease (CHD) and hormone replacement therapy (HRT)

2) How to compute χ^2 (chi-squared) statistic from 2x2

3) What does the chi-squared statistic mean? Where does it come from?

Bigger Data...

Usually, you have more data than in our example...

Are you Man?

Like Ramen?

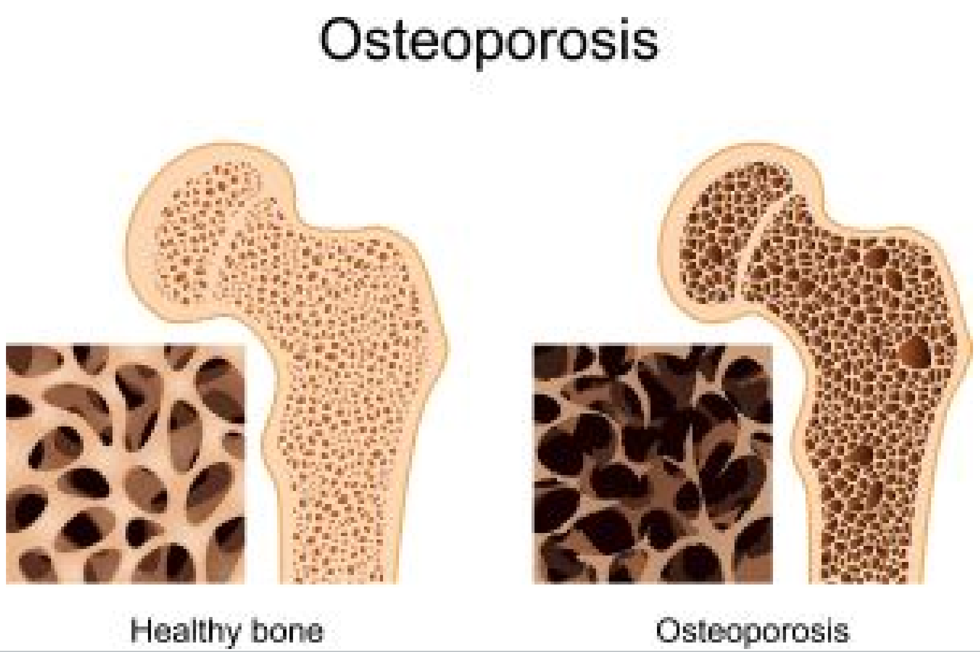
	Yes	No	TOTAL
Yes	5	1	6
No	3	1	4
TOTAL	8	2	10

Person	Like Ramen?	You a man?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes

Postmenopausal Hormone Therapy

Elderly women after menopause (loss of menstruation) exhibit low estrogen levels which leads to:

- hot flashes, vaginal atrophy (short-term effects)
- increased risk of coronary heart disease and osteoporosis (long-term effects)

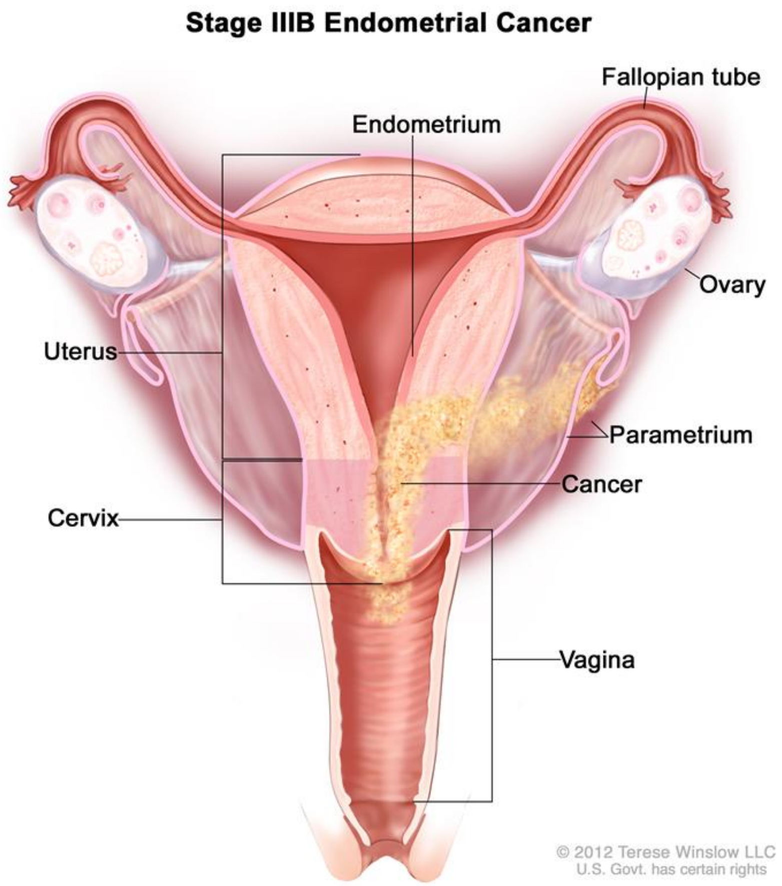


Progestin/Estrogen Therapy

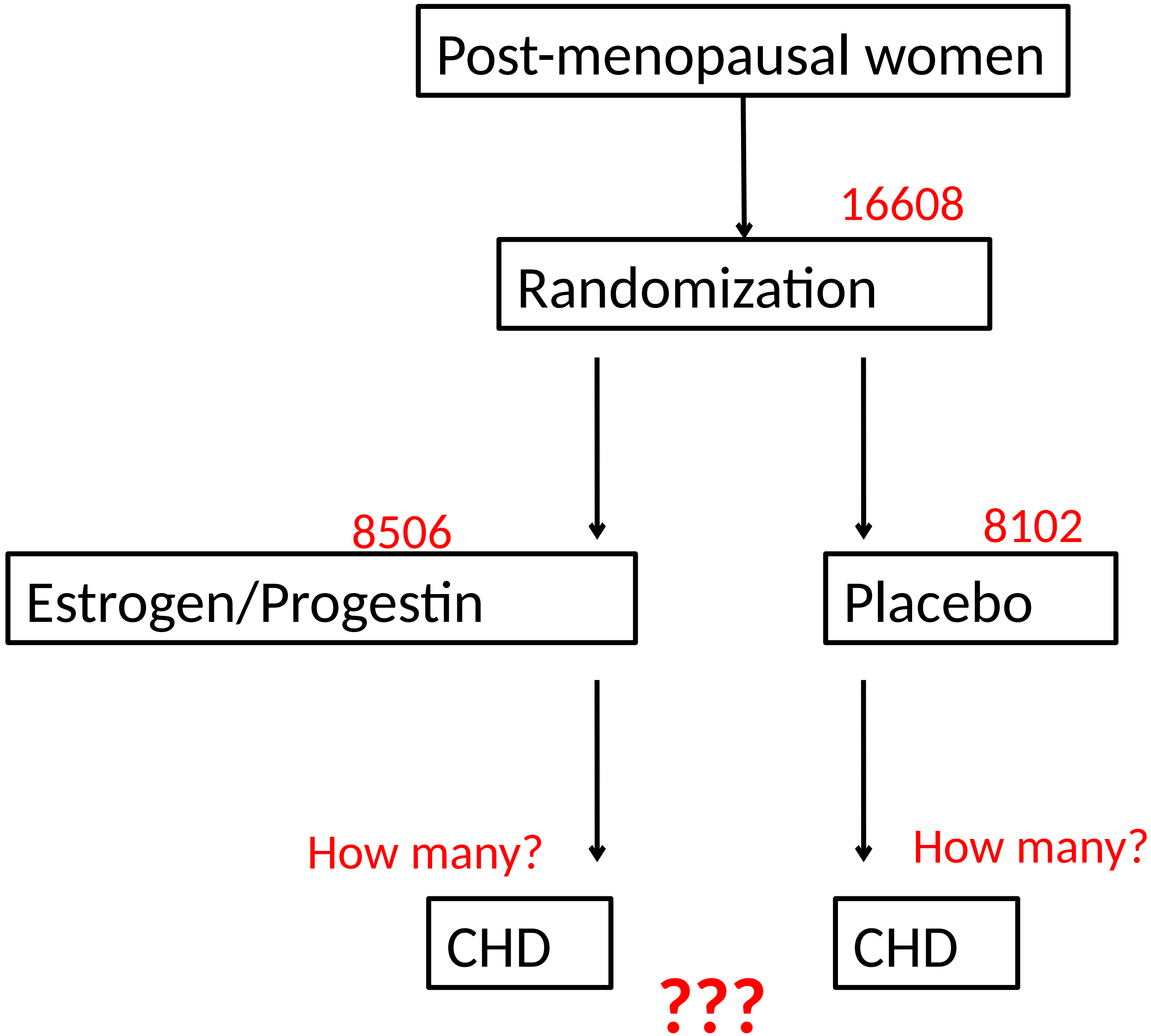
In the 1990s, it seemed reasonable to replace estrogen for postmenopausal women.

However, this sometimes led to carcinoma in the uterus, so estrogen was combined with progestin.

In 1993, the WHO started the study "The Women's Health Initiative (WHI)" to confirm the relationship between hormone therapy and coronary heart disease (CHD).



Design of WHI Study



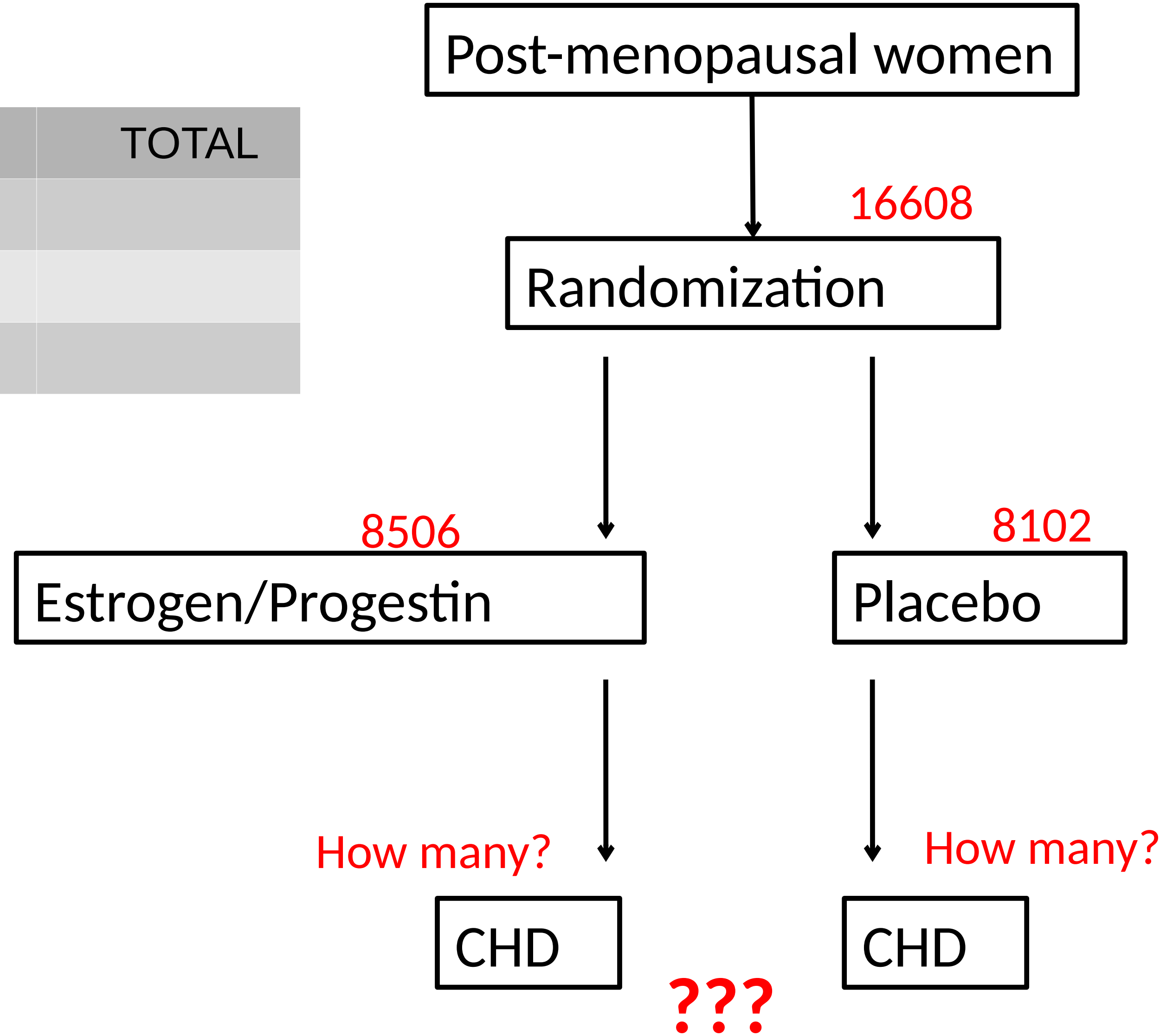
CHD: coronary heart disease

The Women's health initiative investigators. JAMA, 2002

Design of WHI Study

Got CHD?

Placebo?	Yes	No	TOTAL
Yes			
No			
TOTAL			



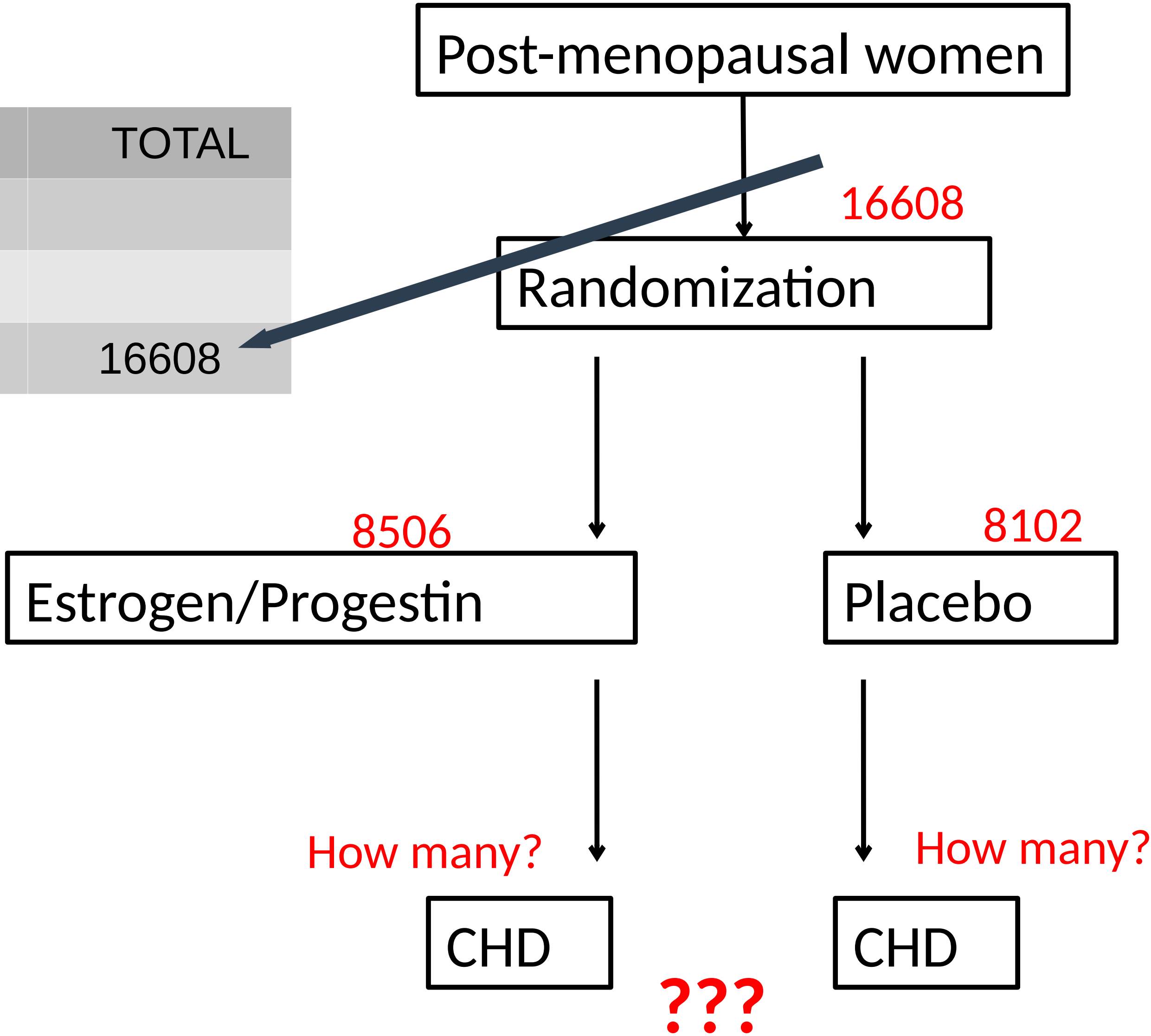
The Women's health initiative investigators. JAMA, 2002

Design of WHI Study

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			
No			
TOTAL			16608

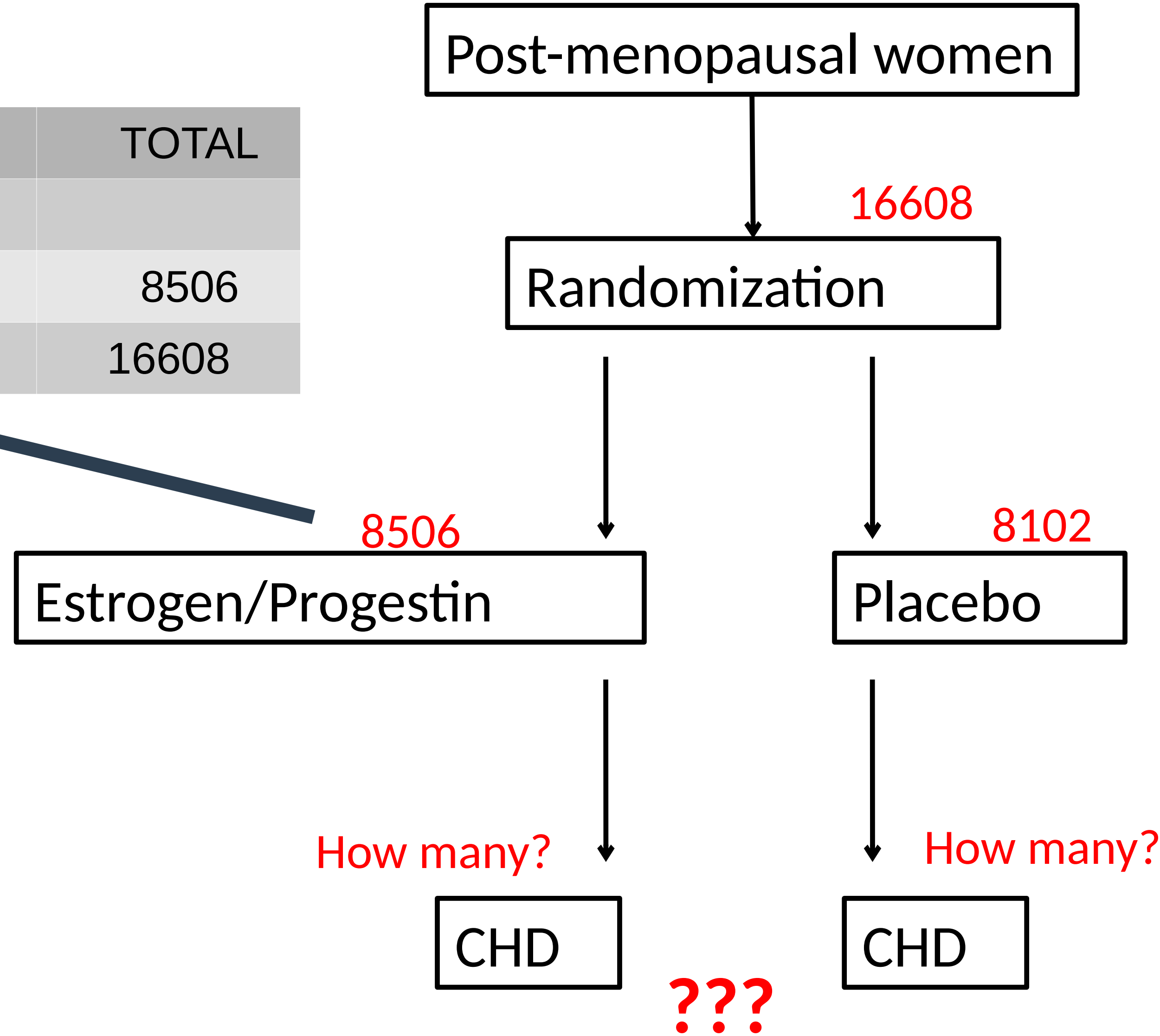


The Women's health initiative investigators. JAMA, 2002

Design of WHI Study

Got CHD?

Placebo?	Yes	No	TOTAL
Yes			
No			8506
TOTAL			16608



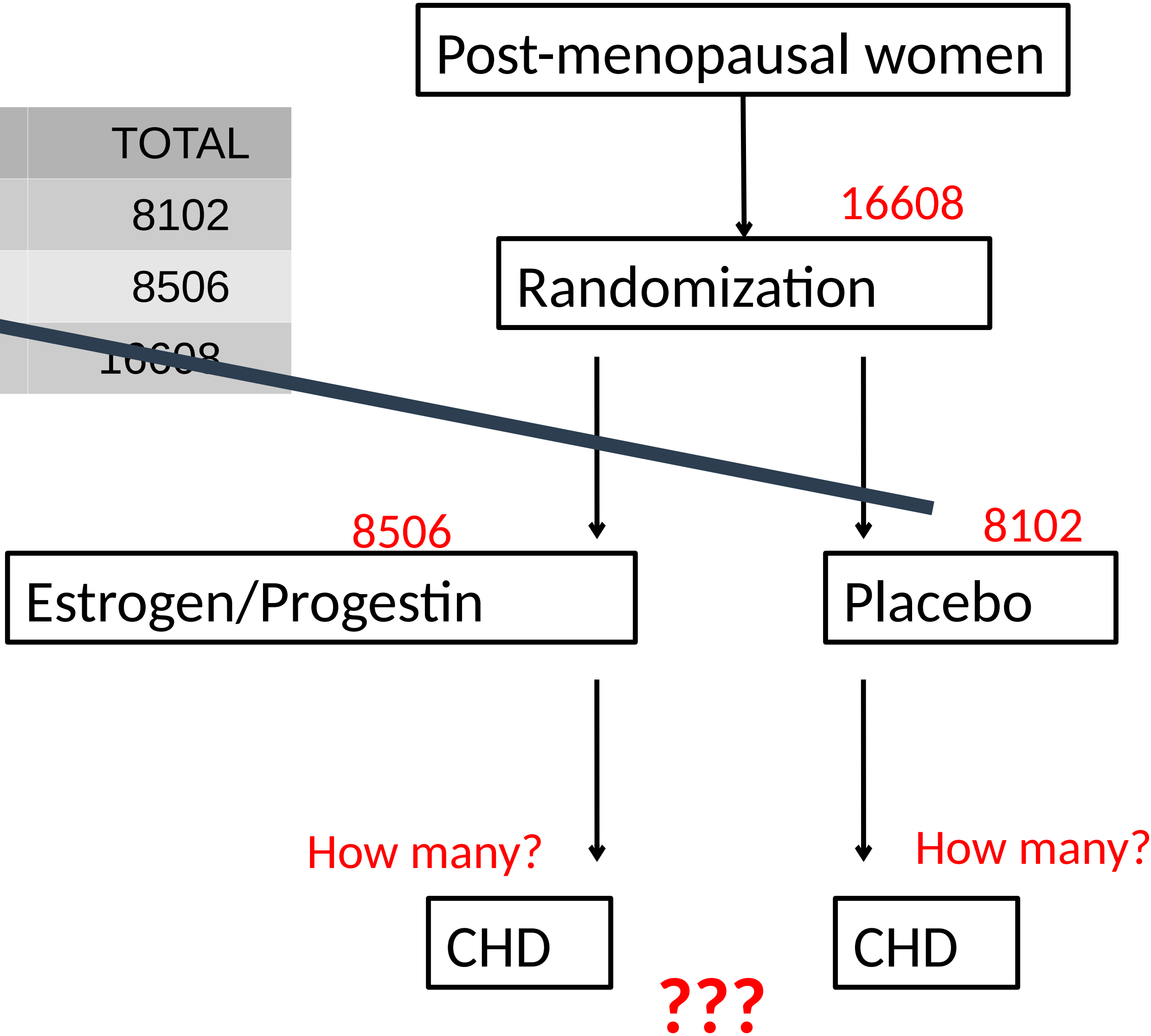
The Women's health initiative investigators. JAMA, 2002

Design of WHI Study

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL			16608

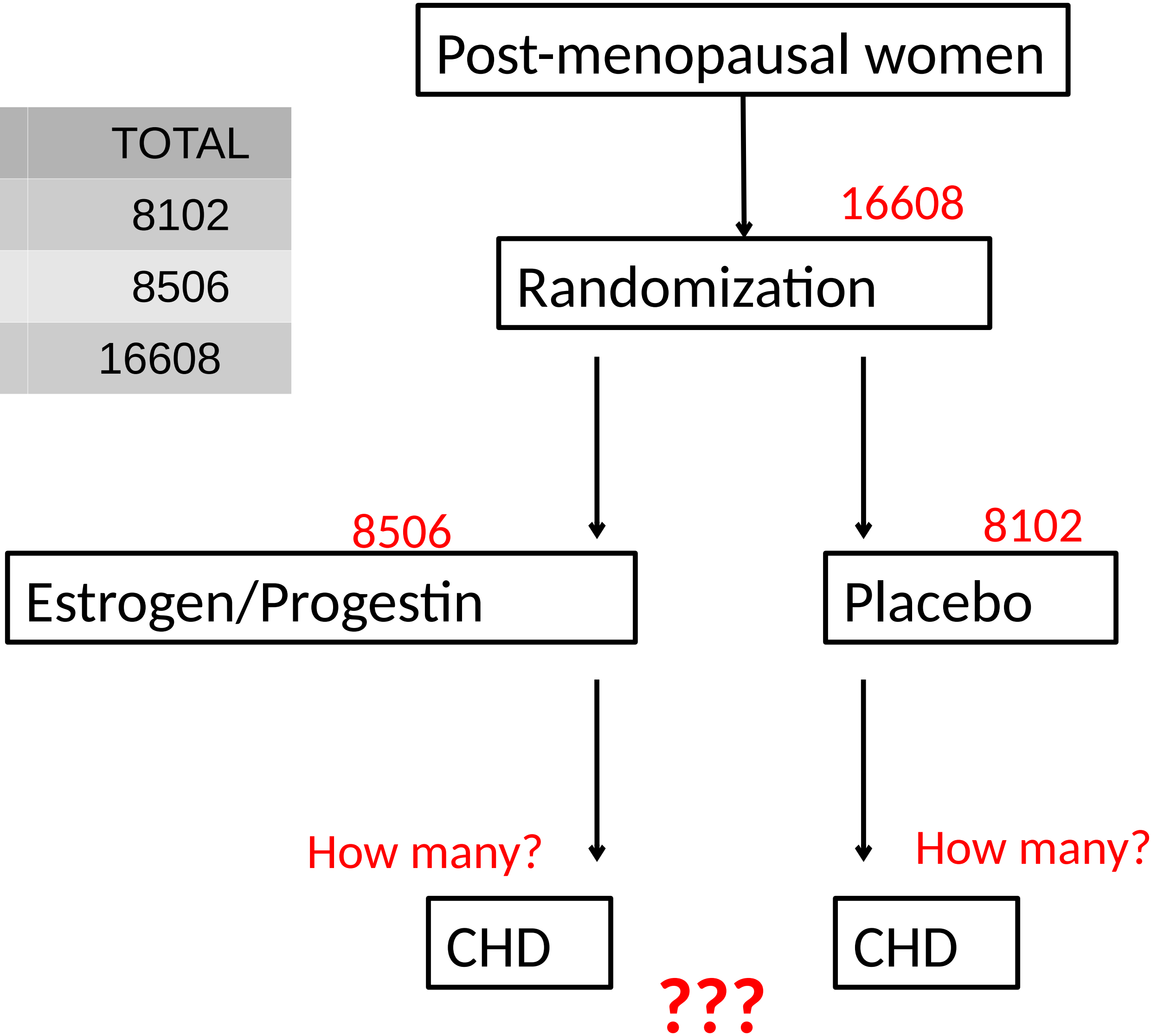


The Women's health initiative investigators. JAMA, 2002

Design of WHI Study

Got CHD?

Placebo?	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL			16608



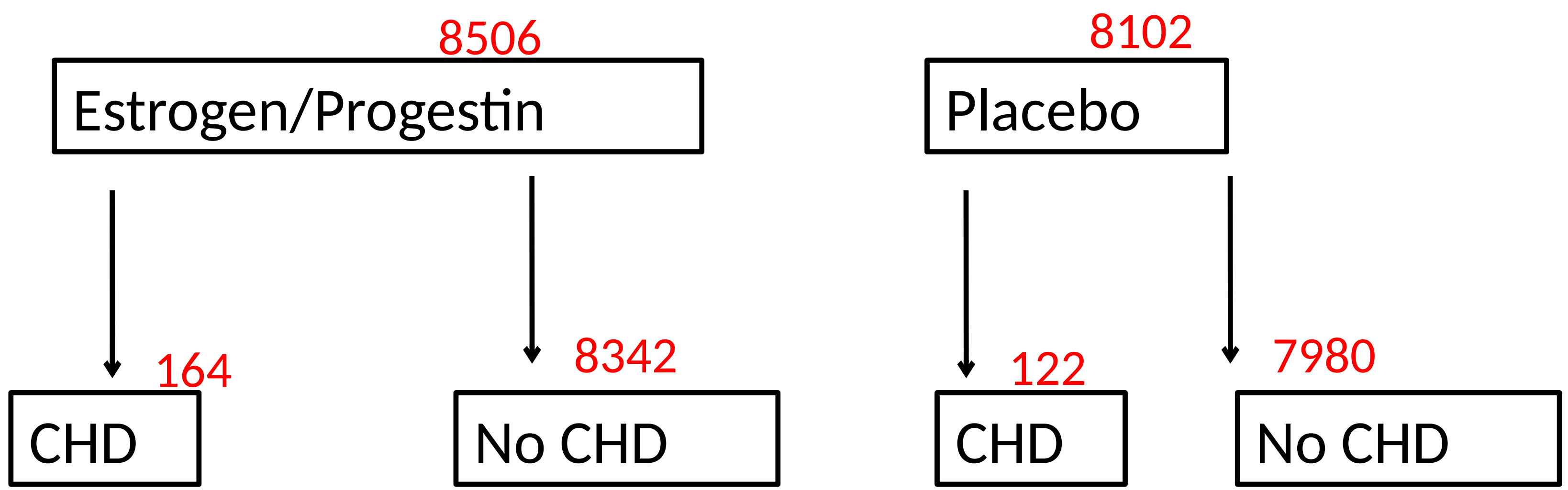
The Women's health initiative investigators. JAMA, 2002

Results

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL			16608

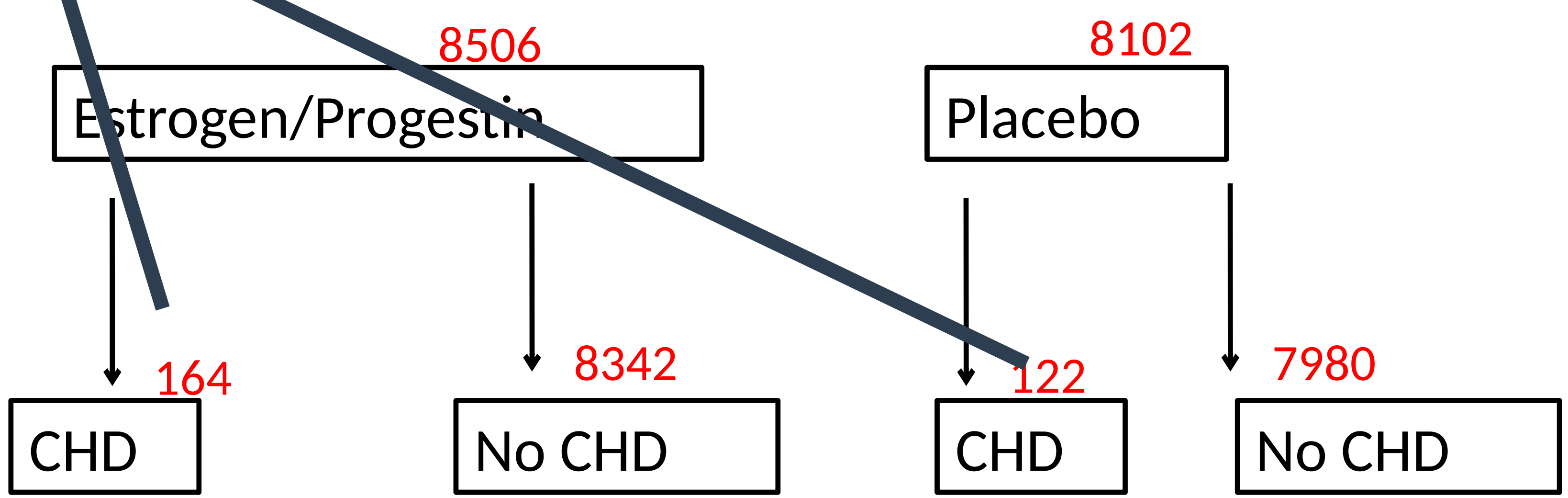


Results

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL	286		16608

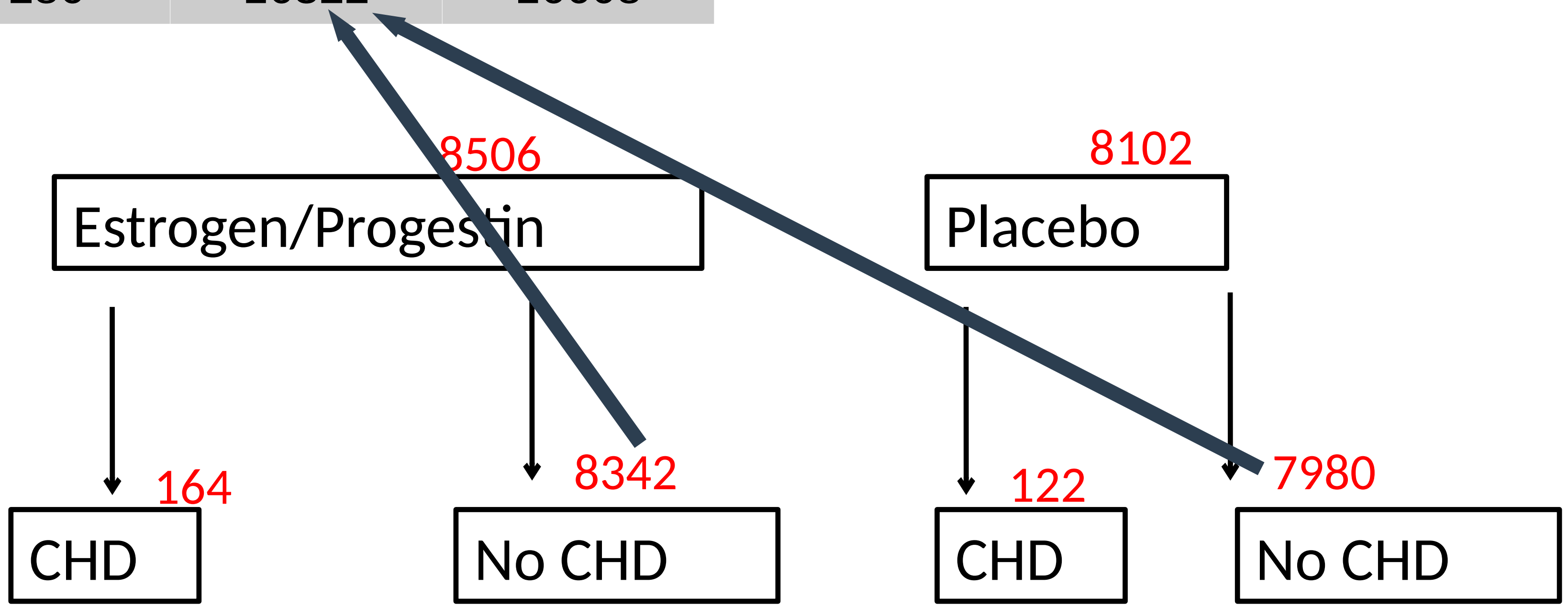


Results

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL	286	16322	16608

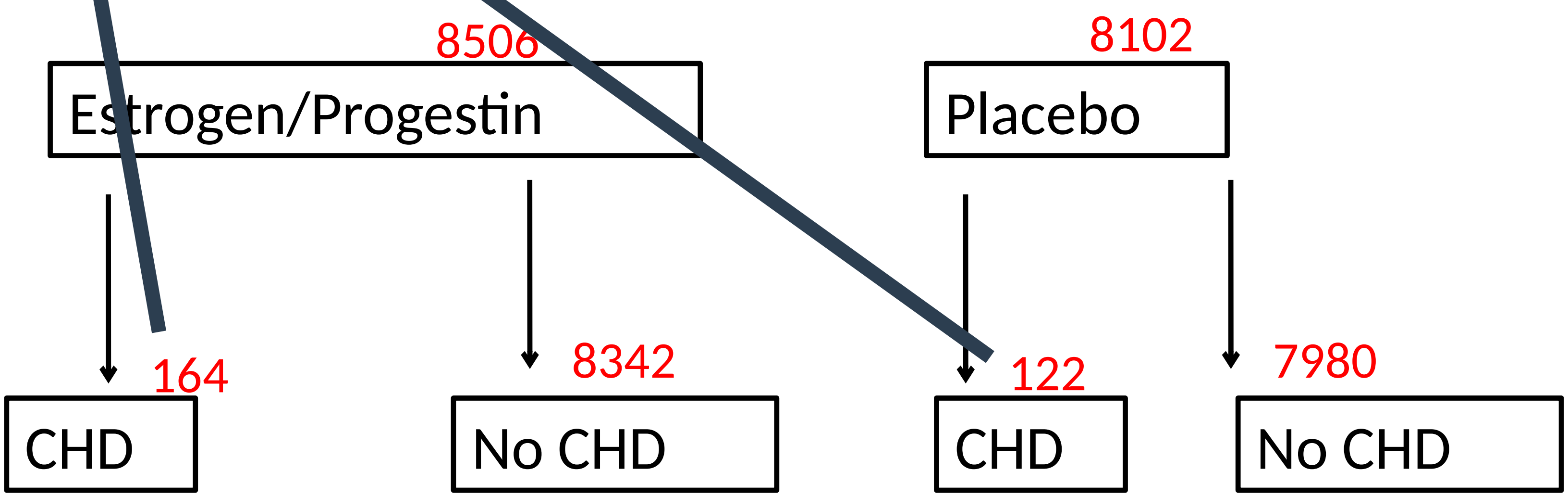


Results

Got CHD?

Placebo?

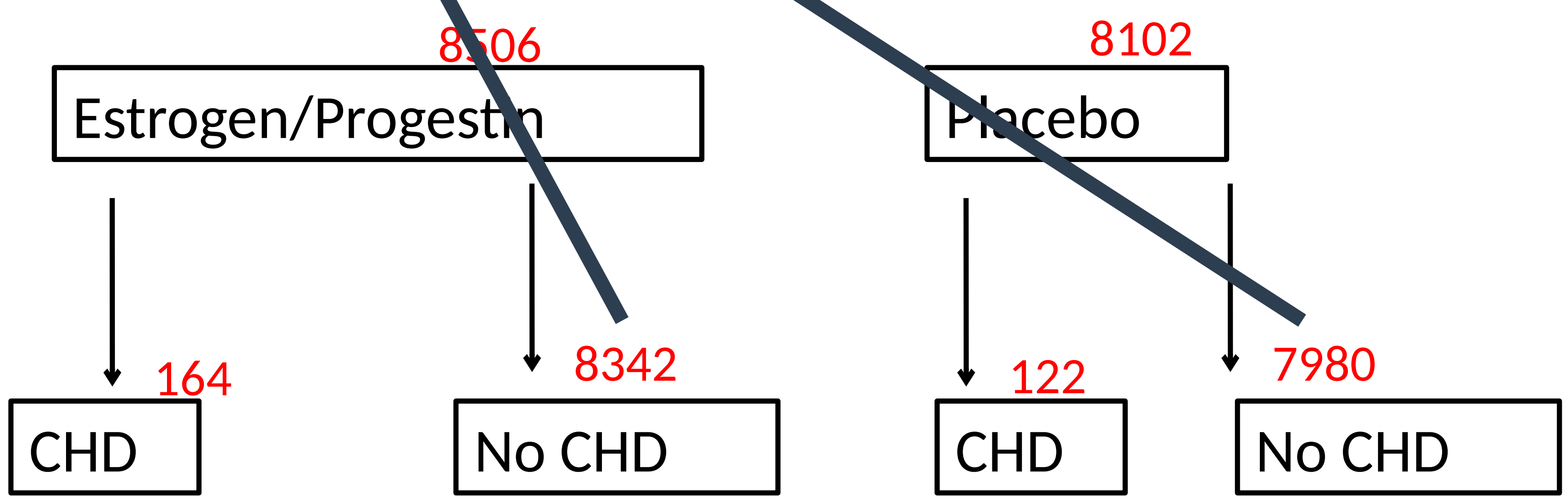
	Yes	No	TOTAL
Yes	122		8102
No	164		8506
TOTAL	286	16322	16608



Results

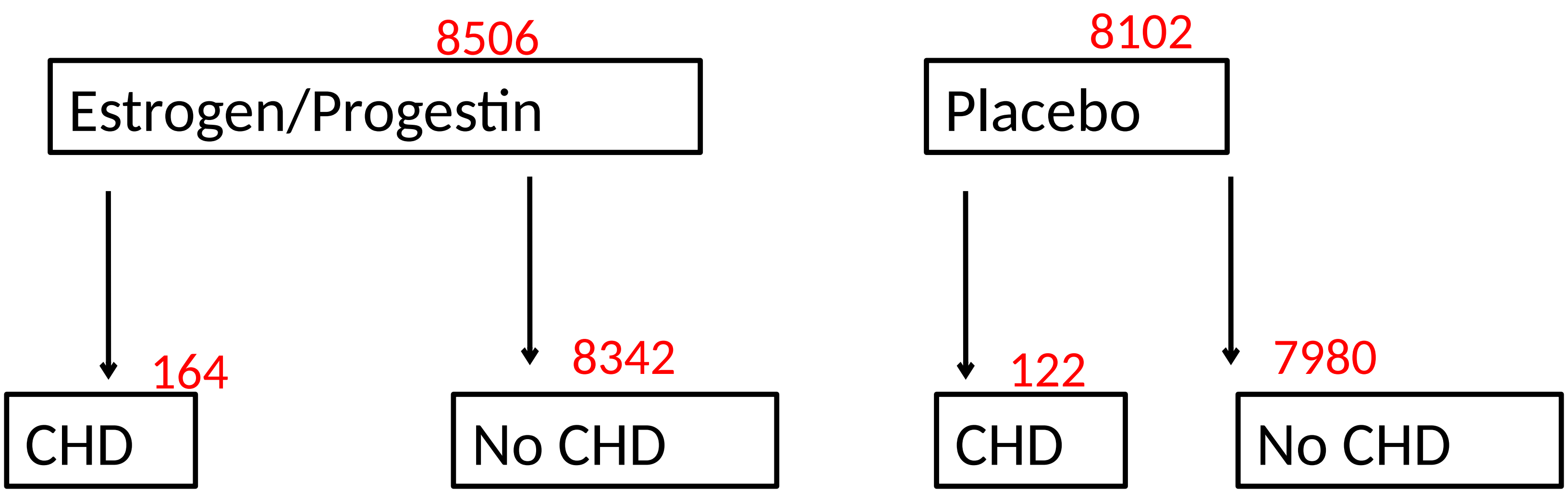
Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608



CHD Results and 2x2 Table

Placebo?	Got CHD?		
	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608



Is it statistically independent?

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Is it statistically independent?

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Expected

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes			8102
No			8506
TOTAL	286	16322	16608

Statistical Independence

How to compute the expected values, assuming statistical independence:

For two statistically independent events A and B:

$P(A|B) = P(A)$;no influence of B on A

$P(B|A) = P(B)$;no influence of A on B

Thus with follows

$P(A|B) = P(A \cap B) / P(B)$

$P(A) = P(A \cap B) / P(B)$

solved for $P(A \cap B)$: $P(A \cap B) = P(A) \cdot P(B)$

and: $N(A \cap B) = N(A) \cdot N(B) / N$; counts

Remember how to compute expected?

Expected:

	CHD	No CHD	SUM
Estrogen/Progestin	146.48	8359.52	8506
Placebo	139.52	7962.48	8102
SUM	286	16322	16608

$N(C \cap E) = N(C) \cdot N(E) / N = 286 \cdot 8506 / 16608 \approx 146.48$

Is it statistically independent?

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Expected

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Is it statistically independent?

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Different by:
17.52

Expected

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Is it statistically independent?

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Is **17.52** a *normal* amount be different?

Different by:
17.52

Expected

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Is it statistically independent?

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Is **17.52** a *normal* amount be different?

By chance?

Different by:
17.52

Expected

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Is it statistically independent?

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

We need to know more about the distribution

Observed

Different by:
17.52

Expected

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Fisher's Exact Test…

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Last time, we learned that one way is to compute the (probability under the): ***hypergeometric distribution***

Probability of a given outcome x:

	Group Y	Group !Y	Total
Group A	x	M-x	M
Group !A	K-x	N-M-K+x	N-M
Total	K	N-K	N

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

$$P(X=x) = \frac{\binom{M}{x} \times \binom{N-M}{K-x}}{\binom{N}{K}}$$

This defines the *hypergeometric distribution* (for 2x2 tables)

Probability of drawing (w/o replacement) “x successes” in K draws where you have M objects of that feature and total population N.

Fisher's Exact Test...

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Last time, we learned that one way is to compute the (probability under the): ***hypergeometric distribution***

But, think about how much data we have...

Think about data...

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Person	Placebo?	CHD?
1	Yes	Yes
2	Yes	Yes
3	Yes	No
4	No	Yes
5	No	No
6	No	Yes
7	Yes	Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Think about data...

Got CHD?

Placebo?

	Yes	No
Yes	122	7980
No	164	8342
TOTAL	286	16322

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
8	Yes	Yes

With only 10 people, we could barely list all the possibilities

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

$$a! = \prod_{x=0}^{a-1} a - x$$

Think about data...

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	
No	164	8342	
TOTAL	286	16322	

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Now we have

<u>16,608</u>

Think about data...

Got CHD?

Placebo?

	Yes	No
Yes	122	7980
No	164	8342
TOTAL	286	16322

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Hint:

10! = 3,628,800
160! = 4.71 E 284

Think about data...

Got CHD?

Placebo?	Yes	No
Yes	122	7980
No	164	8342
TOTAL	286	16322

Atoms in universe
1.00 E 82

160! = 4.71 E 284

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Think about data...

Got CHD?

Placebo?

	Yes	No
Yes	122	7980
No	164	8342
TOTAL	286	16322

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
8	Yes	Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Atoms in universe

1.00 E 82

170! = 7.26 E 306

Think about data...

Got CHD?

Placebo?	Yes	No
Yes	122	7980
No	164	8342
TOTAL	286	16322

Person	Placebo?	CHD?
1	Yes	Yes
		Yes
		No
		Yes
		No
		Yes
		Yes
		Yes
9	No	Yes
10	Yes	Yes
⋮		
16607	No	Yes
16608	Yes	Yes

Atoms in universe
1.00 E 82

180! → your calculator breaks

Probability of a given outcome x:

	Group Y	Group !Y	Total
Group A	x	M-x	M
Group !A	K-x	N-M-K+x	N-M
Total	K	N-K	N

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

$$P(X=x) = \frac{\binom{M}{x} \times \binom{N-M}{K-x}}{\binom{N}{K}}$$

This defines the *hypergeometric distribution* (for 2x2 tables)

Probability of drawing (w/o replacement) “x successes” in K draws where you have M objects of that feature and total population N.

Probability of a given outcome x:

	Group Y	Group !Y	Total
Group A			
Group !A			
Total			

Maybe there is some trick to calculate it?

Terms cancel out...

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

$$P(X=x) = \frac{\binom{M}{x} \times \binom{N-M}{K-x}}{\binom{N}{K}}$$

...efines the ...*eometric* ...*ution* (for ...bles)

...ility of drawing (w/o replacement) "x successes" in K draws where you have M objects of that feature and total population N.

Probability of a given outcome x:

	Group Y	Group !Y	Total
Group A			
Group !A			
Total			

There's got to be a better way...

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

$$P(X=x) = \frac{\binom{M}{x} \times \binom{N-M}{K-x}}{\binom{N}{K}}$$

...efines the ...eometric ...ution (for ...oles)

...ility of drawing (w/o replacement) "x successes" in K draws where you have M objects of that feature and total population N.

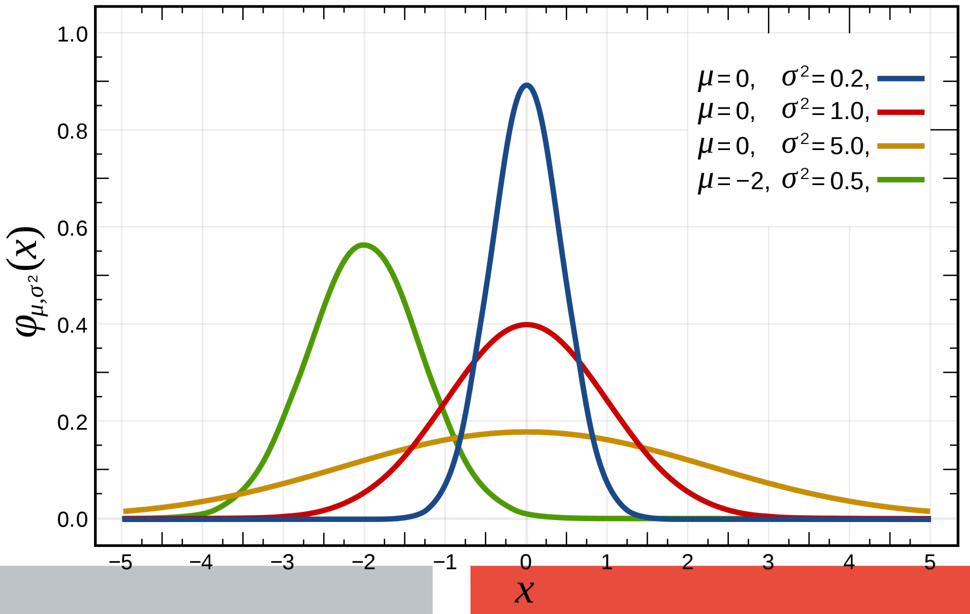
Chi-Squared (χ^2) Distribution

Lots of things in nature follow the "Normal Distribution" (bell curve)

We use it in statistics a lot too…

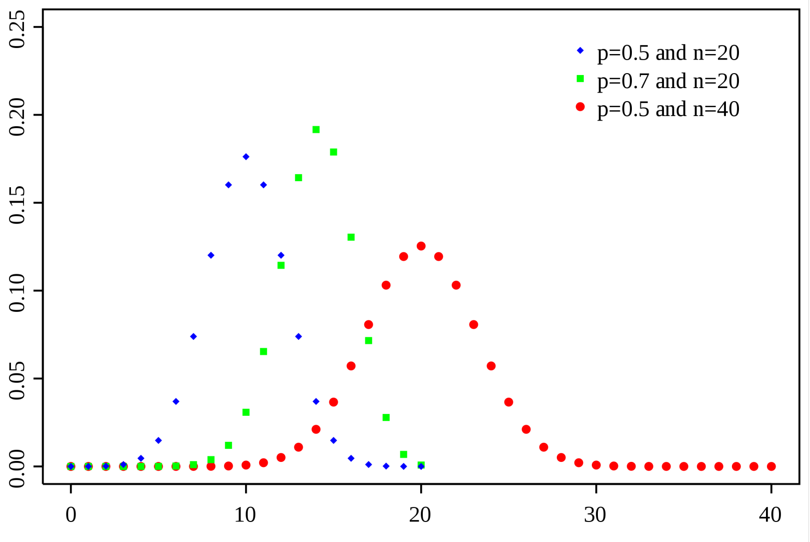
Probability Density Function (PDF):

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

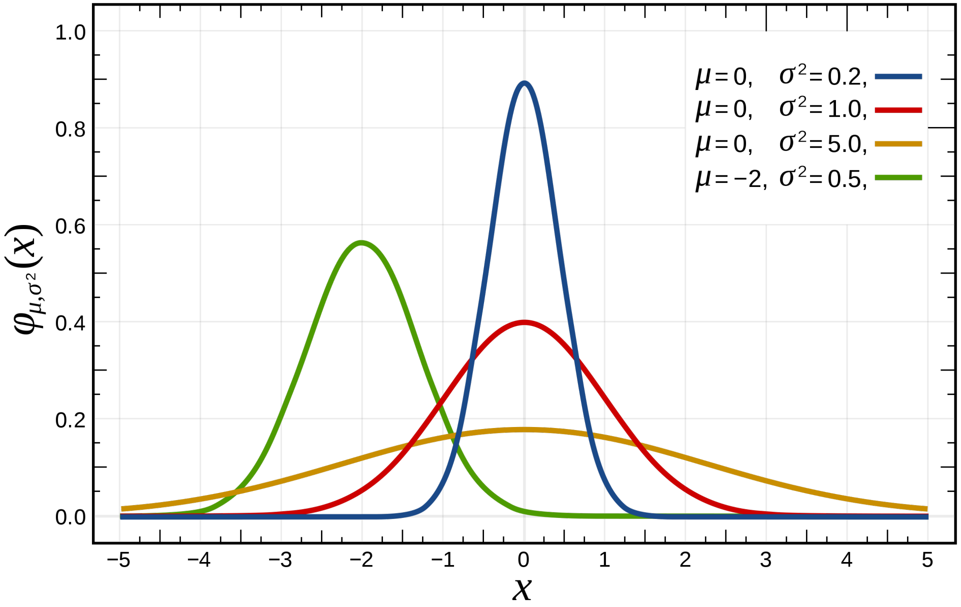


Binomial and Normal…

Binomial Distribution
(what we used to compute
hypergeometric)



Normal Distribution



They look really similar…is it a trick?

Is it by accident? (note: Normal has *smooth support*, whereas binomial only has integer support…)

Chi-Squared (χ^2) Distribution

Laplace and de Moivre (two cool dudes) showed asymptotic normality of:

$$\chi = \frac{m - Np}{\sqrt{Npq}}$$

→ N is number of trials
→ p is probability of success
→ q is probability of failure (1-p)
→ m is observed number of successes (there should be N*p)

Chi-Squared (χ^2) Distribution

Laplace and de Moivre (two cool dudes) showed asymptotic normality of:

$$\chi = \frac{m - Np}{\sqrt{Npq}}$$

→ N is number of trials
→ p is probability of success
→ q is probability of failure (1-p)
→ m is observed number of successes (there should be N*p)

In other words, as N gets big, this approaches a normal distribution.

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:
→ N = Np + N(1-p)
→ N = m + (N-m)
→ q = 1 - p

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:
→ N = Np + N(1-p)
→ N = m + (N-m)
→ q = 1 - p

"Failures"

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

"Successes"

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:

→ N = Np + N(1-p)

→ N = m + (N-m)

→ q = 1 - p

Observed successes

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Chi-Squared (χ^2) Test Statistic

Square it…

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:

→ N = Np + N(1-p)

→ N = m + (N-m)

→ q = 1 - p

Expected successes

$$\chi^2 = \frac{(m - \boxed{Np})^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:
→ N = Np + N(1-p)
→ N = m + (N-m)
→ q = 1 - p

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Normalized by expected successes

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:

→ N = Np + N(1-p)

→ N = m + (N-m)

→ q = 1 - p

Observed failures

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(\boxed{N - m} - Nq)^2}{Nq}$$

Chi-Squared (χ^2) Test Statistic

Square it…

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:
→ N = Np + N(1-p)
→ N = m + (N-m)
→ q = 1 - p

Expected Failures

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Chi-Squared (χ^2) Test Statistic

Square it...

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

Since:
→ N = Np + N(1-p)
→ N = m + (N-m)
→ q = 1 - p

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

Normalized by Expected Failures

Chi-Squared (χ^2) Test Statistic

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

For n cells in table.
Observed ↔ Expected

Is it statistically independent?

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

Observed

Different by:
17.52

Expected

Got CHD?

Placebo?	Yes	No	TOTAL
Yes	139.52	7962.48	8102
No	146.48	8359.52	8506
TOTAL	286	16322	16608

Is it statistically independent?

Observed

Got CHD?

Placebo?

	Yes	No	TOTAL
Yes	122	7980	8102
No	164	8342	8506
TOTAL	286	16322	16608

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Expected

Got CHD?

Placebo?

	Yes	No	[illegible]
Yes	139.52	7962.48	[illegible]
No	146.48	8359.52	[illegible]
TOTAL	286	16322	16608

$$\chi^2 = \frac{(122 - 139.52)^2}{139.52} + \frac{(164 - 146.48)^2}{146.48} + \frac{(7980 - 7962.48)^2}{7962.48} + \frac{(8342 - 8359.52)^2}{8359.52}$$

Chi-Squared (χ^2) Test Statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

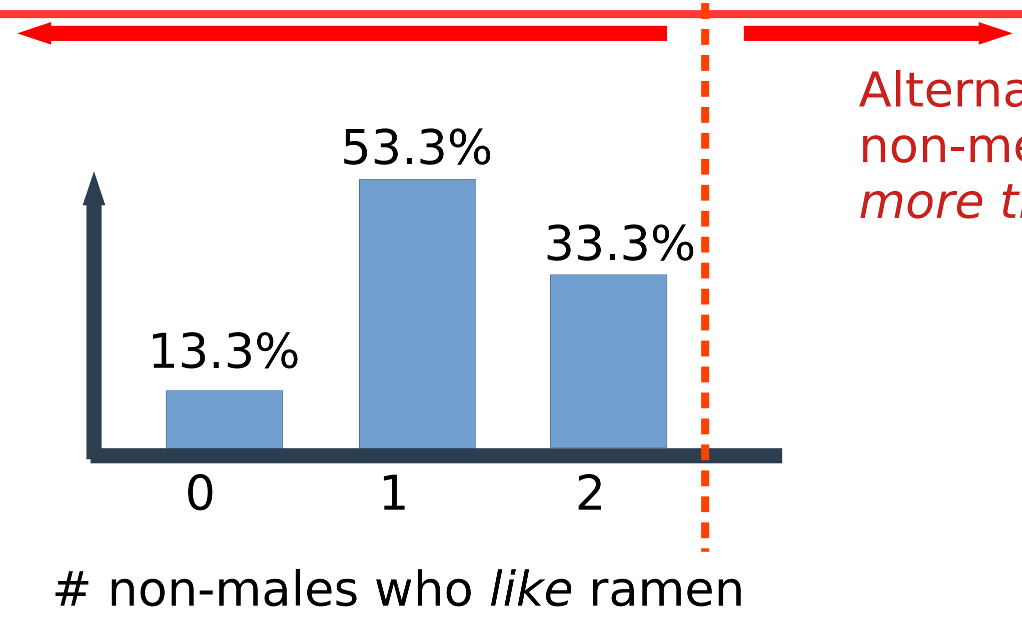
	CHD	No CHD	
Placebo	2.2	0.0385	
Estrogen/Progestin	2.0955	0.0367	
			4.3708

$$\chi^2 = 4.3708$$

Statistic vs Distribution

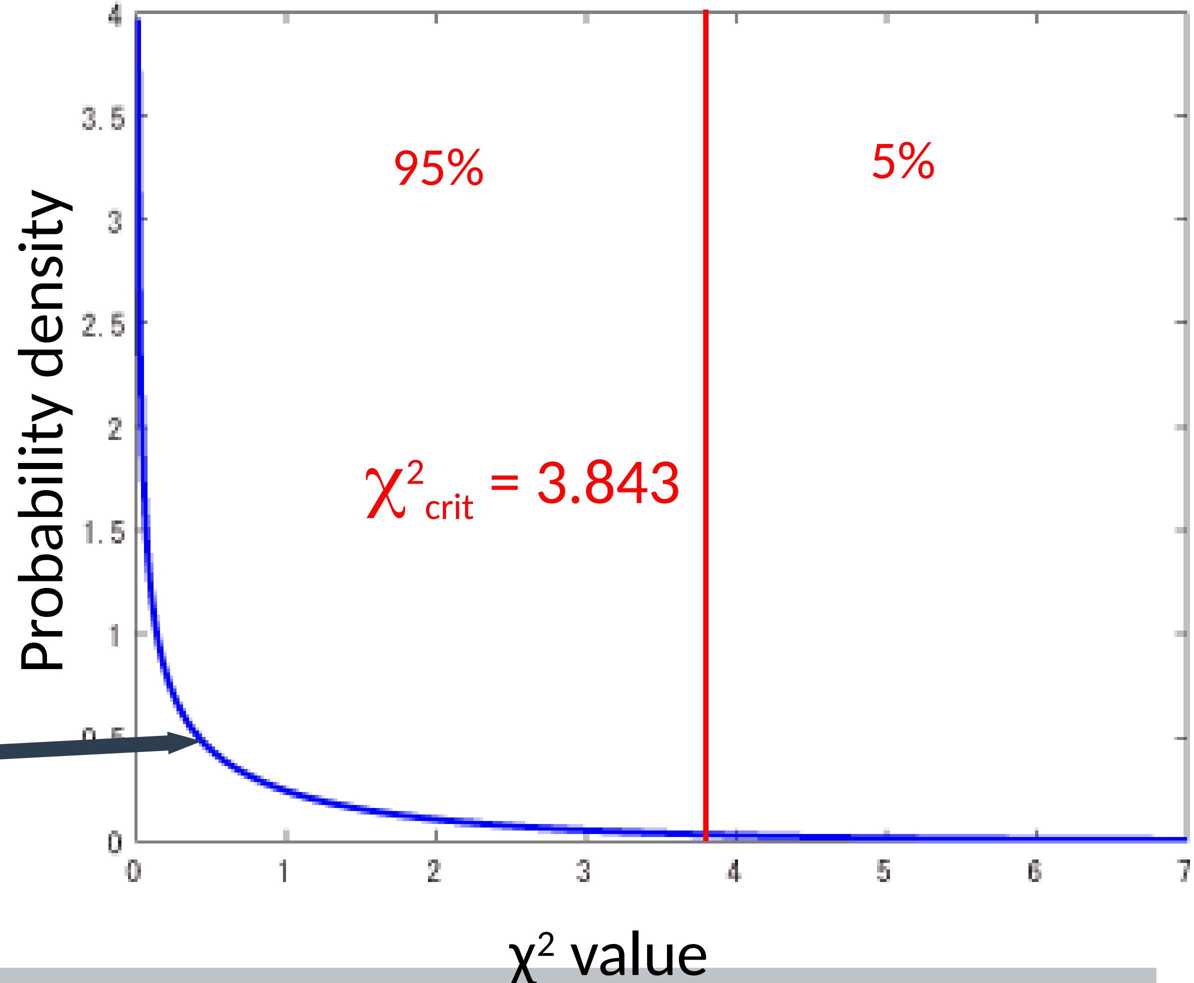
That's just the *statistic*
→ Now, we need to know the *distribution* to calculate **cutoff**

Alternate hypothesis: non-men like ramen *more than* men



Compare statistic against distribution

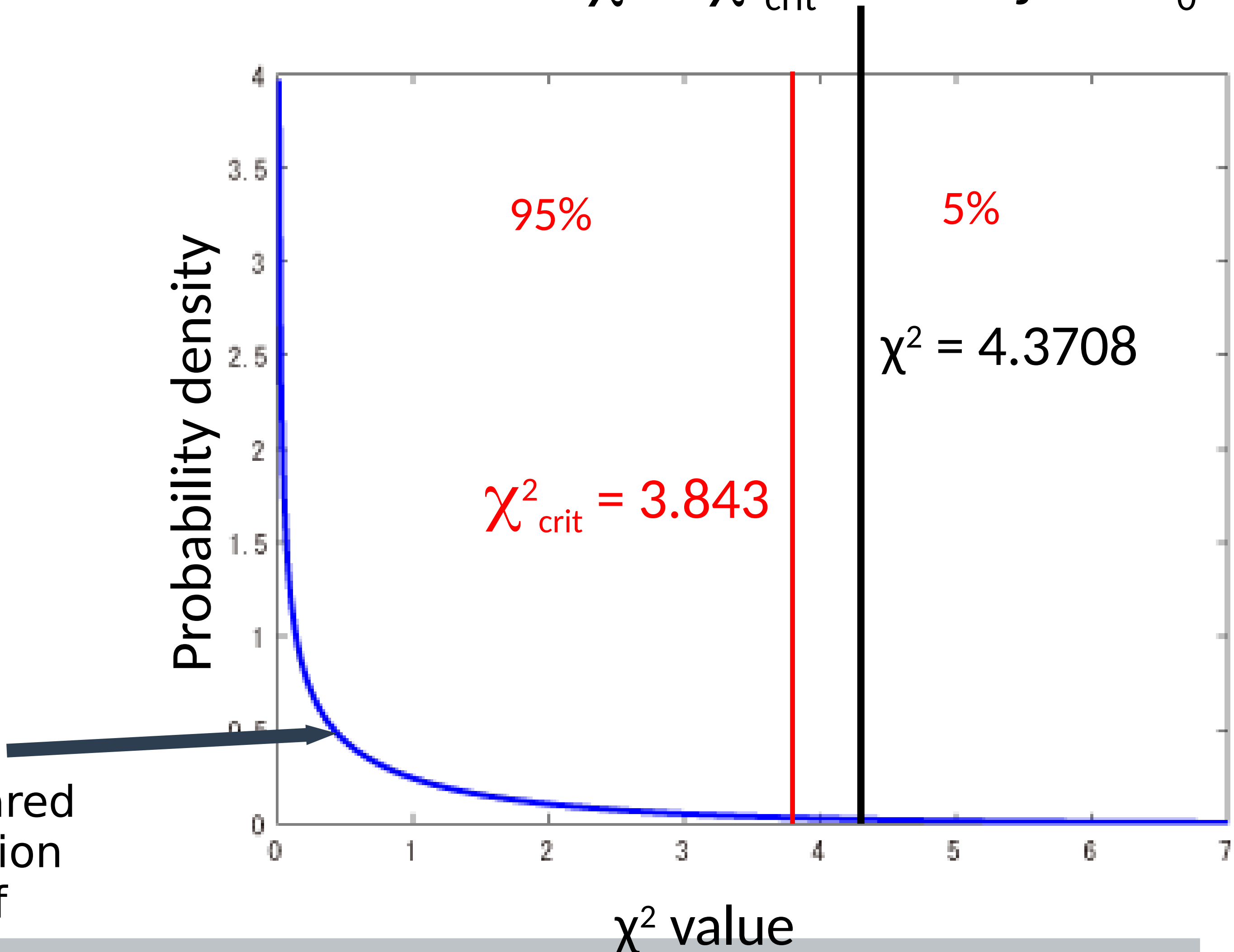
$\chi^2 > \chi^2_{crit}$: we reject H_0



Chi-squared distribution with 1-df

Compare statistic against distribution

$\chi^2 > \chi^2_{crit}$: we reject H_0

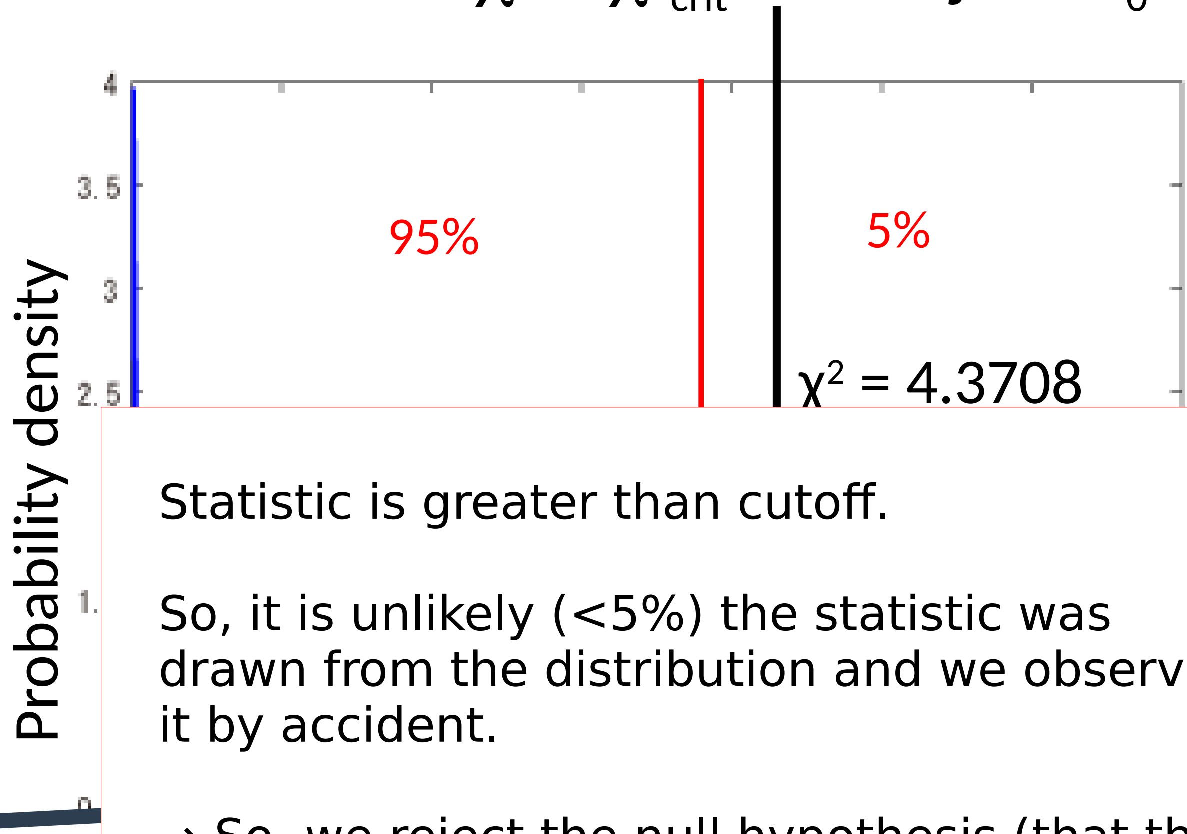


There are tables to tell us cutoff for chi-squared 5%, 1% alpha...

Chi-squared distribution with 1-df

Compare statistic against distribution

$\chi^2 > \chi^2_{crit}$: we reject H_0



There are tables to tell us cutoff for chi-squared … % alpha…

Statistic is greater than cutoff.

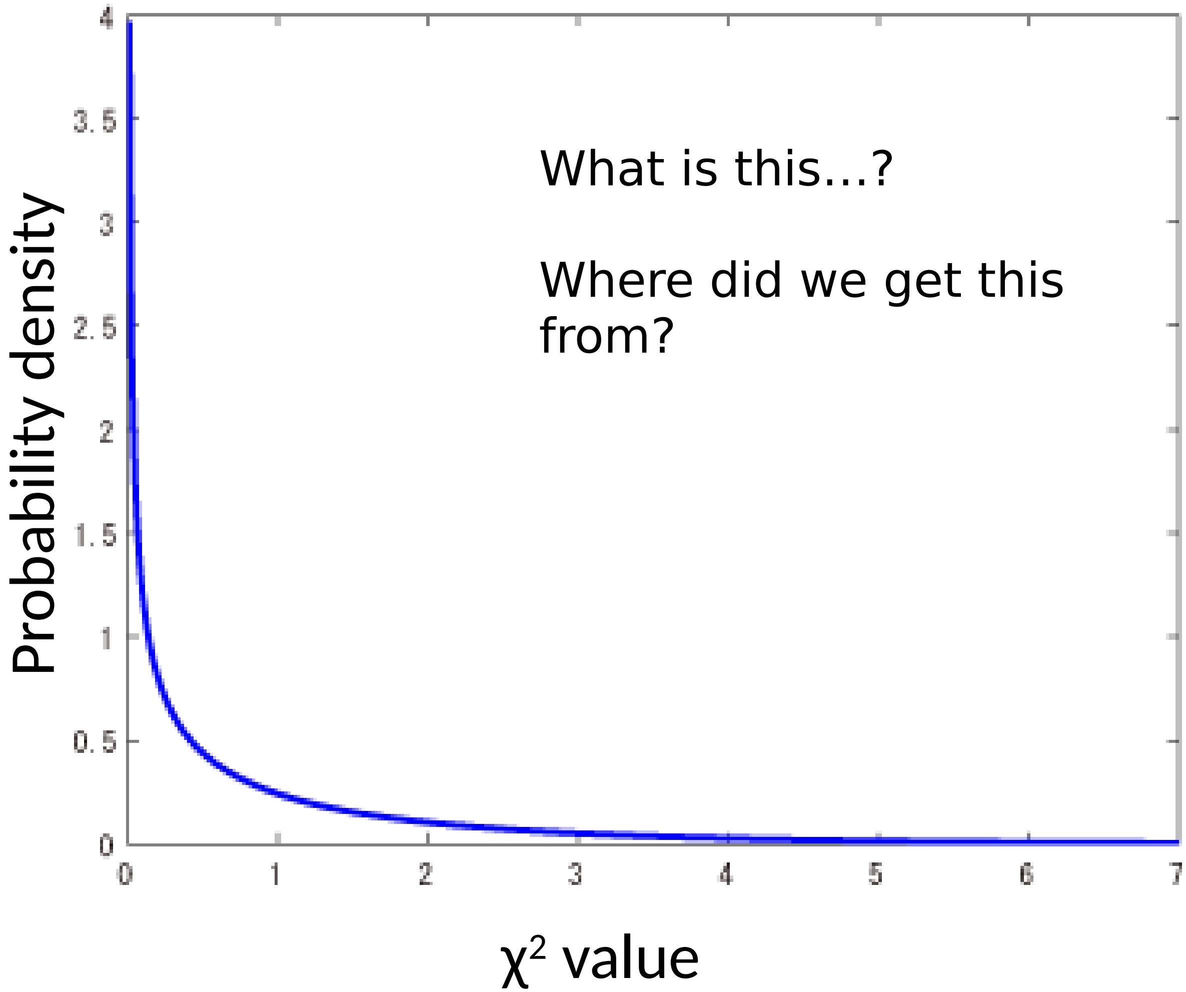
So, it is unlikely (<5%) the statistic was drawn from the distribution and we observed it by accident.

→ So, we reject the null hypothesis (that the statistic is drawn from the distribution).

Chi-squared distribution with 1-df

χ^2 value

Where does chi-square distribution come from?



More intuitive:
Chi-Squared (χ^2)

Simulation:

Marginal totals (row and column sums):

	CHD	No CHD	SUM
Estrogen/Progestin			8506
Placebo			8102
SUM	286	16322	16608

This means:

P(CHD) = 286/16608 = 1.72%

P(Placebo) = 8102/16608 = 48.78%

For each person **X** in 16608 people:

→ put **X** in CHD if a uniform random number generator (RNG) of [0, 1) returns a number < 0.0172 (1.72%)

→ Put **X** in placebo group if RNG returns < 0.4878 (48.78%)

More intuitive:
Chi-Squared (χ^2)

Simulation:

M

Est
Pla
SU

Th

P(

P(

So, I will "simulate" 10,000 new "universes".

In each universe, I am "god" and I will (randomly) choose whether *each person* gets CHD (with 1.72% chance), and whether that person is chosen for the placebo group (48.78% chance).

For each person **X** in 16608 people:

→ put **X** in CHD if a uniform random number generator (RNG) of [0, 1) returns a number < 0.0172 (1.72%)

→ Put **X** in placebo group if RNG returns < 0.4878 (48.78%)

Simulation

1st simulation

	CHD	No CHD	SUM
Estrogen/Progestin	*134*	*8422*	8506
Placebo	*151*	*7901*	8102
SUM	286	16322	16608

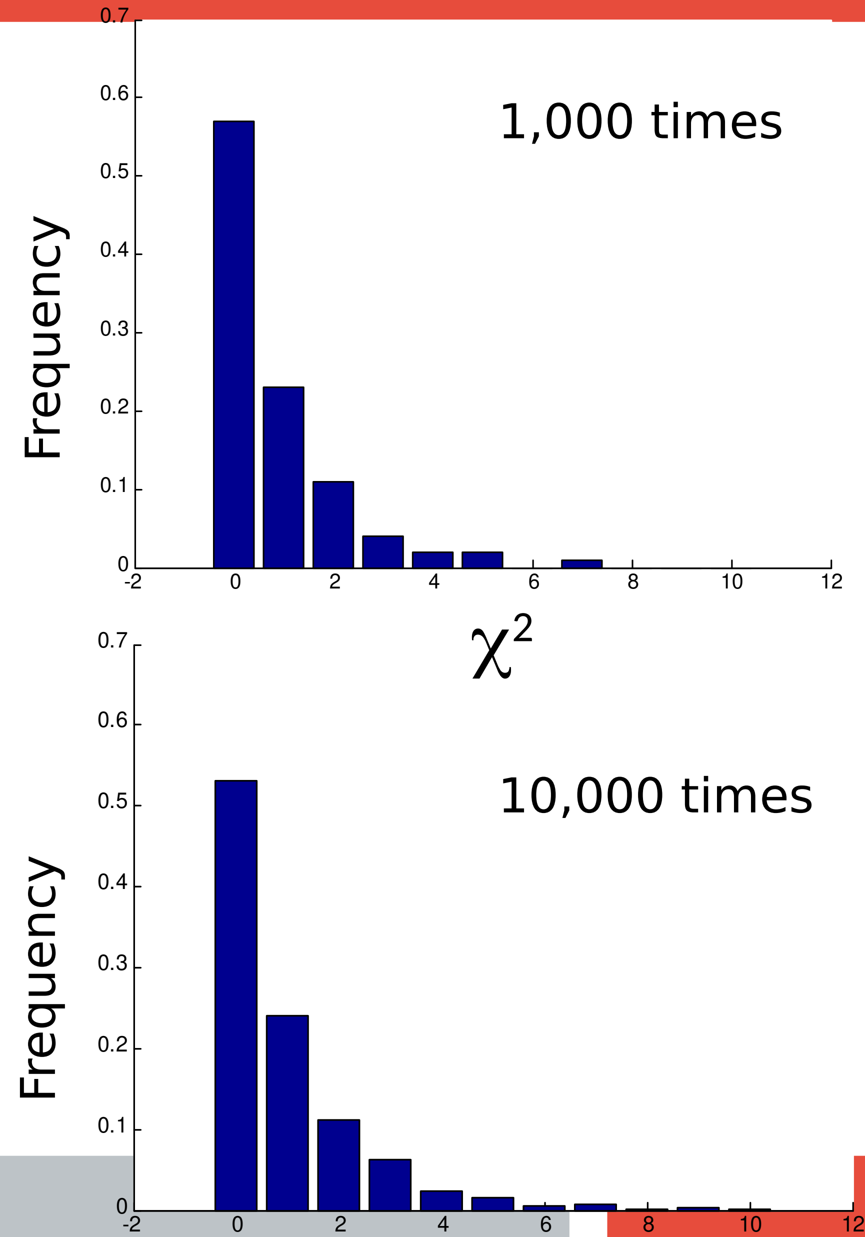
$\chi^2 = 2.3508$

2nd simulation

	CHD	No CHD	SUM
Estrogen/Progestin	*154*	*8403*	8506
Placebo	*147*	*7904*	8102
SUM	286	16322	16608

$\chi^2 = 0.016$

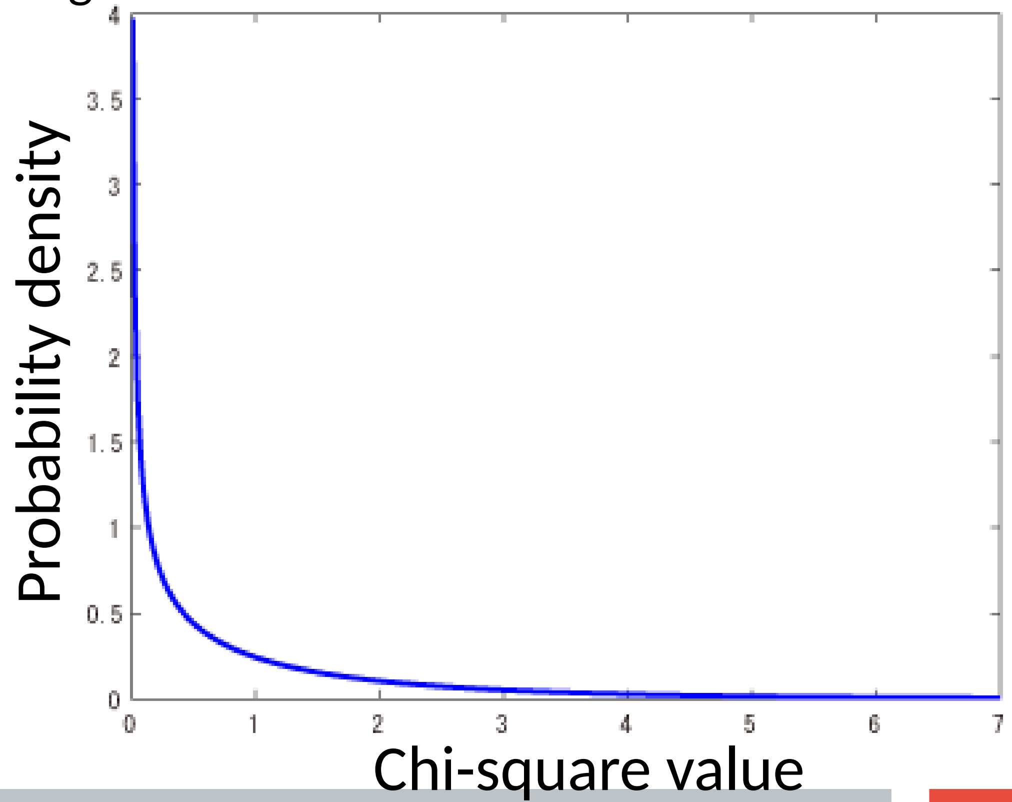
```
Trial 992: Chi2=1.1338232484762085
Act: 142 Exp: 139.03359826589596
Act: 143 Exp: 145.96640173410404
Act: 7960 Exp: 7962.966401734104
Act: 8363 Exp: 8360.033598265896
Chisqr: 0.125733059452652
Trial 993: Chi2=0.125733059452652
Act: 131 Exp: 133.5631021194605
Act: 146 Exp: 143.4368978805395
Act: 7877 Exp: 7874.43689788054
Act: 8454 Exp: 8456.56310211946
Chisqr: 0.09659814006034541
Trial 994: Chi2=0.09659814006034541
Act: 140 Exp: 137.65317919075144
Act: 148 Exp: 150.34682080924856
Act: 7798 Exp: 7800.346820809248
Act: 8522 Exp: 8519.653179190751
Chisqr: 0.07799540816796628
Trial 995: Chi2=0.07799540816796628
Act: 129 Exp: 135.91480009633912
Act: 152 Exp: 145.08519990366088
Act: 7904 Exp: 7897.085199903661
Act: 8423 Exp: 8429.91480009634
Chisqr: 0.6930852539586512
Trial 996: Chi2=0.6930852539586512
Act: 136 Exp: 132.2986512524085
Act: 136 Exp: 139.7013487475915
Act: 7942 Exp: 7945.7013487475915
Act: 8394 Exp: 8390.298651252408
Chisqr: 0.20497670791878034
Trial 997: Chi2=0.20497670791878034
Act: 136 Exp: 126.82875722543352
Act: 121 Exp: 130.17124277456648
Act: 8060 Exp: 8069.171242774566
Act: 8291 Exp: 8281.828757225434
Chisqr: 1.3299329091414815
Trial 998: Chi2=1.3299329091414815
Act: 160 Exp: 144.23169556840077
Act: 135 Exp: 150.76830443159923
Act: 7960 Exp: 7975.768304431599
Act: 8353 Exp: 8337.2316955684
Chisqr: 3.4340352463266792
Trial 999: Chi2=3.4340352463266792
```



Chi-square with df=1

For a 2x2 contingency table, we have df = 1 (1 degree of freedom)

The shape of the χ^2 – distribution depends on the number of random variables that are free to vary. In case of the 2x2 contingency table it is only one cell, because once one cell is fixed, you can compute the values of the other cells from one cell and the marginal totals.



Chi-square distribution

$$Q = \sum_{i=1}^{k} Z_i^2$$

Z_1, Z_2, Z_3...Z_k are independent, standard normal random variables. *K is some positive integer.*

(i.e. drawn from N(0, 1))

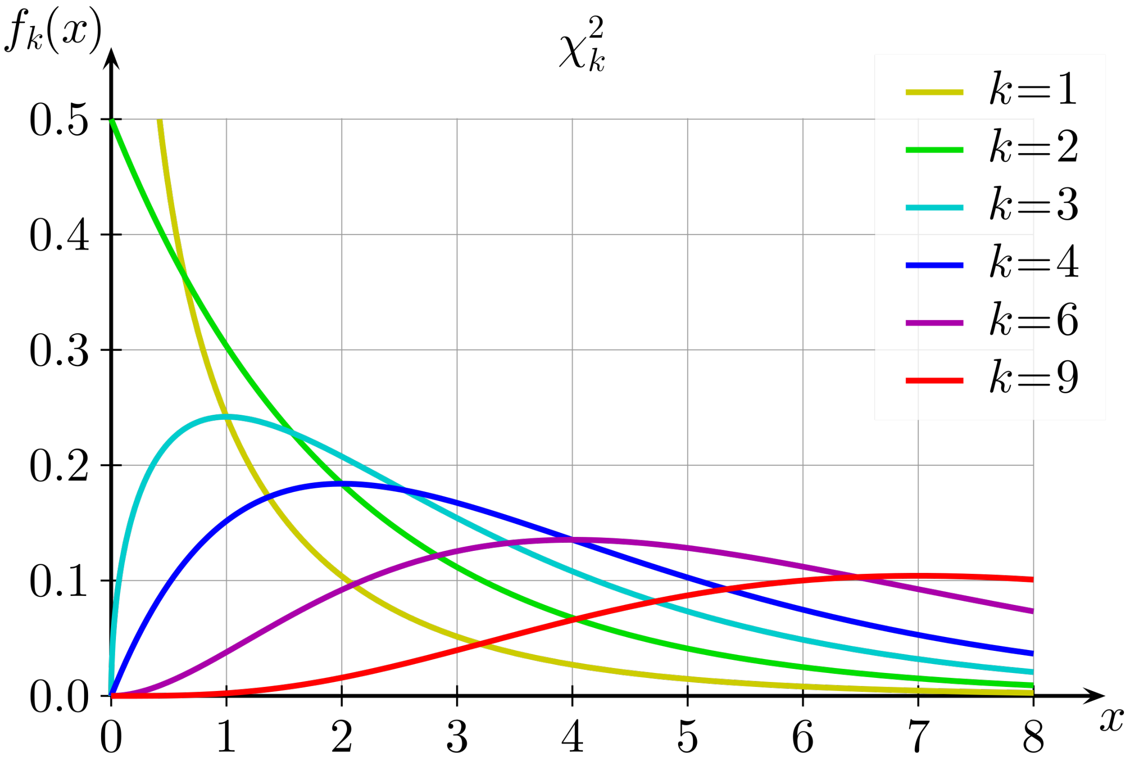
→ N(0,1) means N with mu (mean) of 0, and sigma (standard deviation) of 1.

Then *Q* is distributed according to chi-square distribution with *k* degrees of freedom

Chi-square distribution

$$Q = \sum_{i=1}^{k} Z_i^2$$

$$Pr(x = X) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$



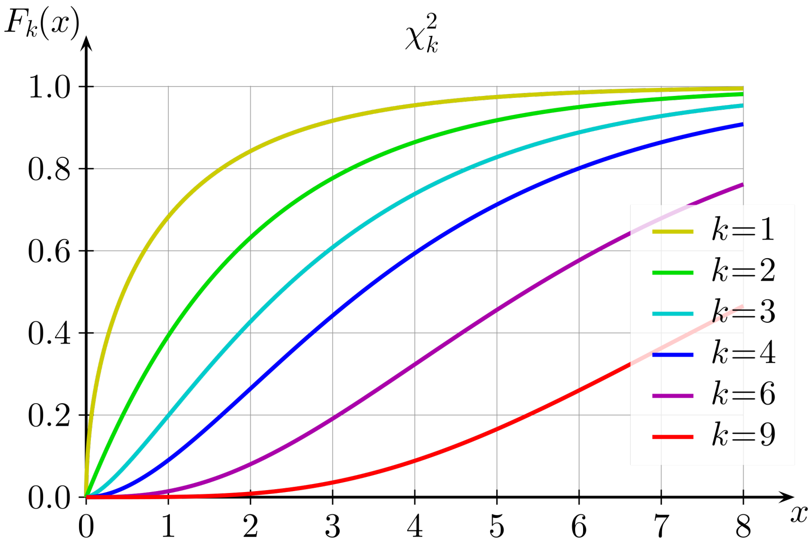
PDF (probability density functions) for various *k*

(stolen from wikipedia)

PDF, CDF

Cumulative Distribution Function

$$Pr\left(x > X\right) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(k/2\right)}$$



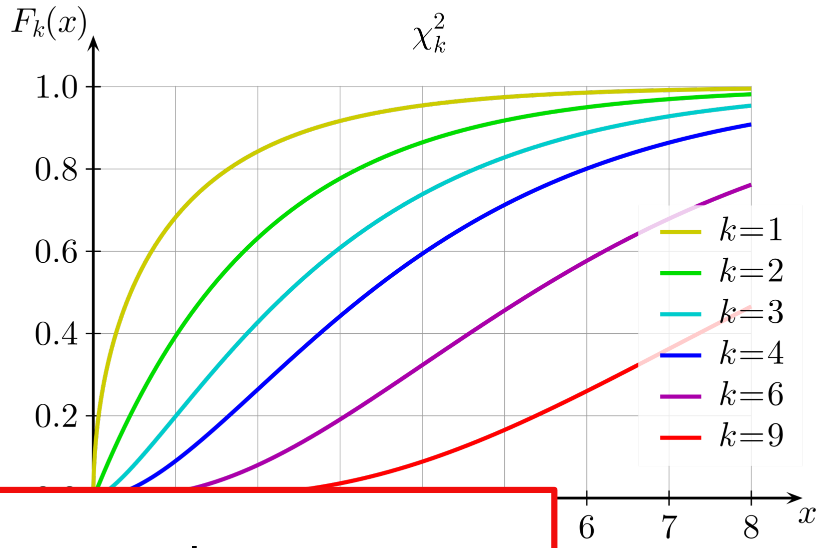
We can “easily” find our cutoff using the (inverse of the) CDF:

For 95%, at what point is Pr(X>x) = 0.95
For 99%, at what point is Pr(X>x) = 0.99

PDF, CDF

Cumulative Distribution Function

$$Pr(x > X) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma(k/2)}$$



Problem with this (and even normal distributions…) is that the PDF and CDF are transcendental functions. (Gamma function)

So, we have to approximate the values numerically.

Fortunately, this is easy since there are tables of pre-computed values.

We can … DF:

For 95…

For 99…

E.g. wikipedia...

They used to sell books with these numbers in them.

And they are usually in text books for common functions.

Degrees of freedom (df)	χ^2 value[19]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Chi-square: what’s the point?

→ Fisher’s exact test is *always* better than chi-square. You *must* use exact test if sparse data (Any expected values < 5).

→ Reason for chi-square in 2x2 is historical (before we had modern computers to easily compute hypergeometric for your specific data).

→ However, when we start to do more complex things than 2x2 categorical, chi-square becomes relevant again (e.g. F-test…)

How to report results

In some fields it is common to report exact p-values:

$(\chi^2[1] = 4.37, p = 0.037)$

In some fields it is common to report exact p-values only for nonsignificant results and otherwise state that p is below the previously set α:

$(\chi^2[1] = 4.37, p < 0.05)$

The value in squared brackets ([]) after χ^2 are the degrees of freedom (df).

How to report results

In some fields it is common to report exact p-values:

(χ^2[1] = 4.37, $p = 0.037$)

In some fields it is common to report exact p-values only for nonsignificant results and otherwise state that p is below the previously set α:

(χ^2[1] = 4.37, $p < 0.05$)

“Informally, the p-value is the probability under a specified statistical model that a statistical summary of the data [...] would be equal to or more extreme than its observed value.” (Wasserstein and Lazar, The American Statistician, 2016)

Here: The p-value is the probability that under the null hypothesis model, χ^2 is equal or larger than the one observed in the data.

How to report results

“We observed a significant association between the preventive intervention (Estrogen/Progestin versus placebo) and later occurrence of coronary heart disease ($\chi^2[1] = 4.37$, $p = 0.037$).”

- The test only tells us about non-independence of two variables, but **does not indicate the direction of this association**.
(i.e. χ^2 is always two-tailed)

Requirements:
1) Groups should be independent, i.e., no repeated measurements.
2) Expected values should be greater than 5.

How to report results

This is how you should write results in homeworks!

"We observed a significant association between the preventive intervention (Estrogen/Progestin versus placebo) and later occurrence of coronary heart disease ($\chi^2[1] = 4.37$, $p = 0.037$)."

- The test only tells us about non-independence of two variables, but **does not indicate the direction of this association**.
(i.e. χ^2 is always two-tailed)

Requirements:
1) Groups should be independent, i.e., no repeated measurements.
2) Expected values should be greater than 5.

How to report results

"We observed a significant association between the preventive intervention (Estrogen/Progestin versus placebo) and later occurrence of coronary heart disease ($\chi^2[1] = 4.37$, $p = 0.037$)."

- The test only tells us about non-independence of two variables, but **does not indicate the direction of this association**.
(i.e. χ^2 is always two-tailed)

Requirements:
1) Groups should be independent, i.e., no repeated measurements.
2) Expected values should be greater than 5.

If any expected values < 5, must use Fisher's Exact!

1) Create a 2x2 contingency table of your data

2) Define the hypotheses:

   H_0: Variables A and B are statistically independent

   H_a: Variables A and B are not statistically independent

3) Calculate the expected values of each cell, assuming independence of the two variables (H_0): $N(A \cap B) = N(A) \cdot N(B) / N$

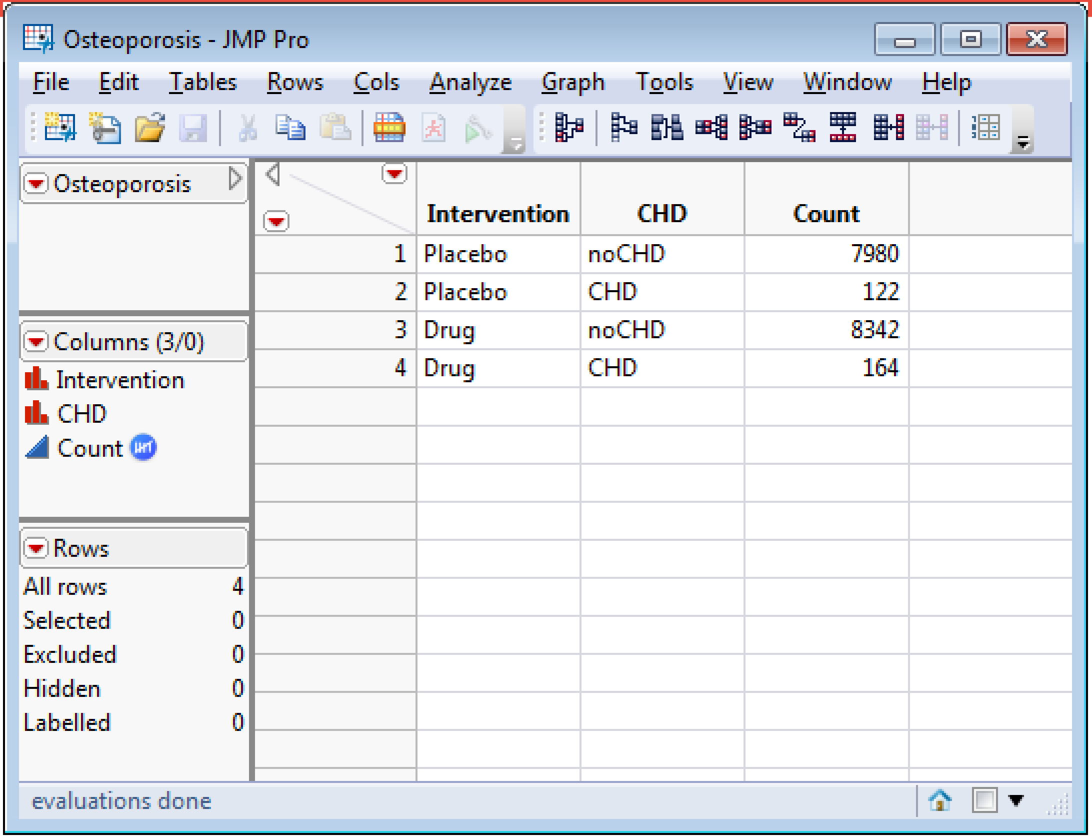
4) Compute the test statistic $$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Compare the observed χ^2 with the critical value χ^2_{crit} which is derived from $P(\chi^2 \geq \chi^2_{crit} | H_0) = \alpha$; with α set to, e.g., 0.05

If $\chi^2 > \chi^2_{crit}$: reject the null hypothesis

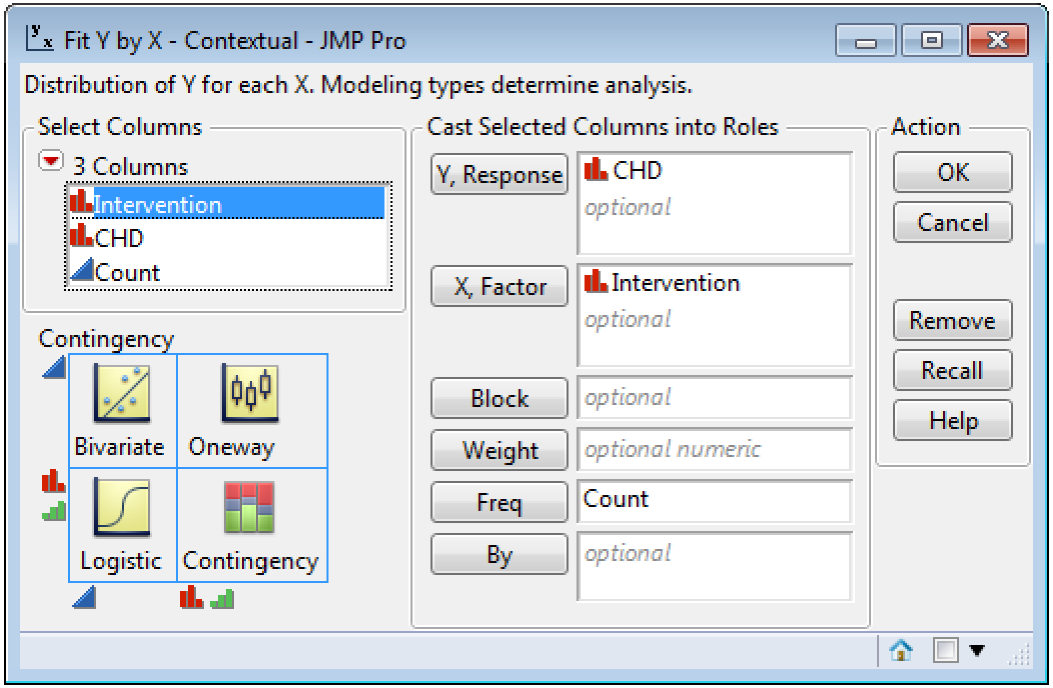
If $\chi^2 \leq \chi^2_{crit}$: do not reject the null hypothesis

Make a contingency table in JMP…



You can directly create a contingency table in JMP
→ Usually rows stand for individual cases/patients/participants.
For a 2x2 contingency we need a third column and do
“Preselect Role”-> “Freq”.

Chi-square test in JMP

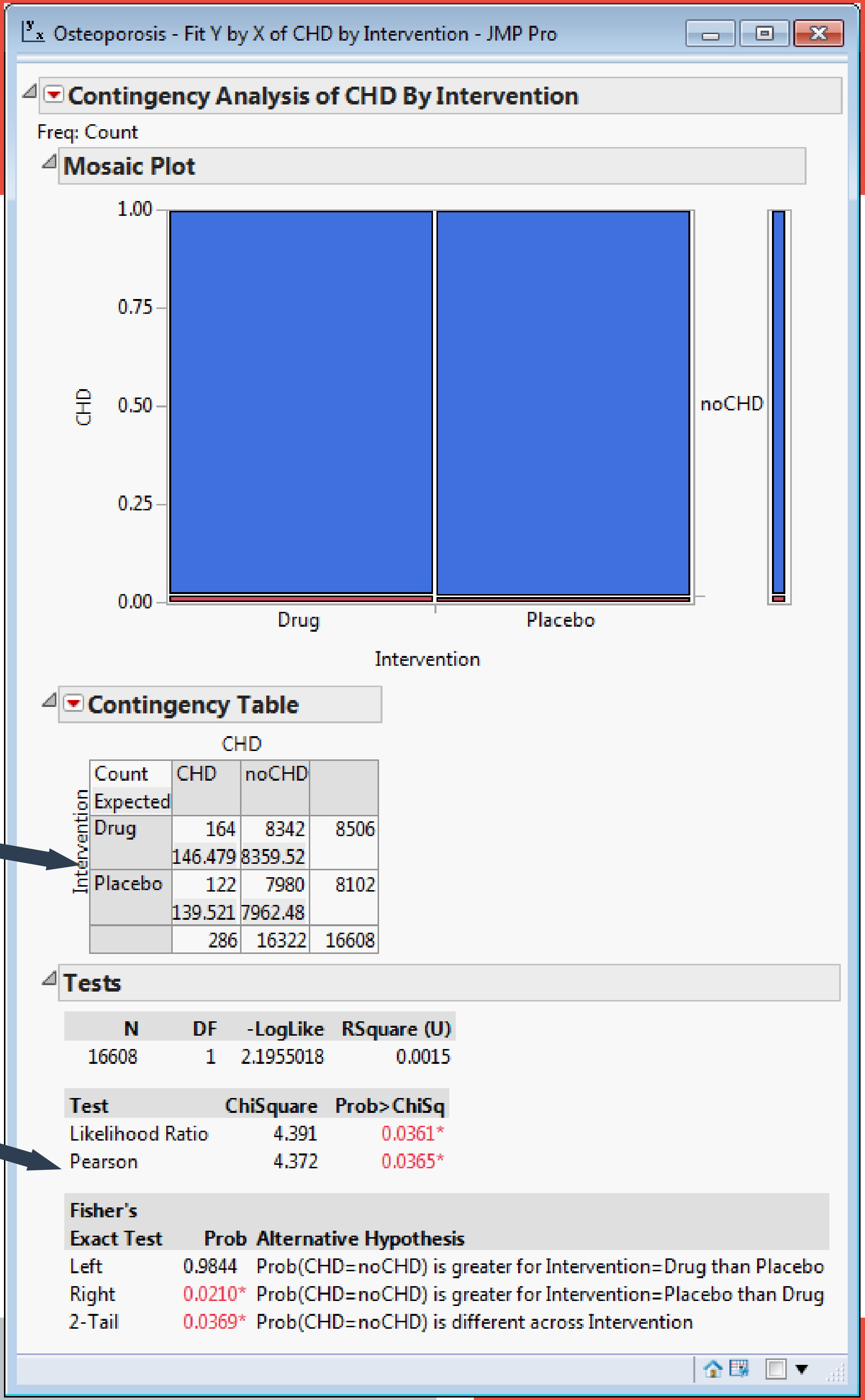


Under "Analyze", we choose "Fit Y by X" and define the roles.

In JMP

We see our contingency table, we can display the expected values.

Below we see the result of the Chi-square test (called Pearson's Chi-square).



Extra Slides

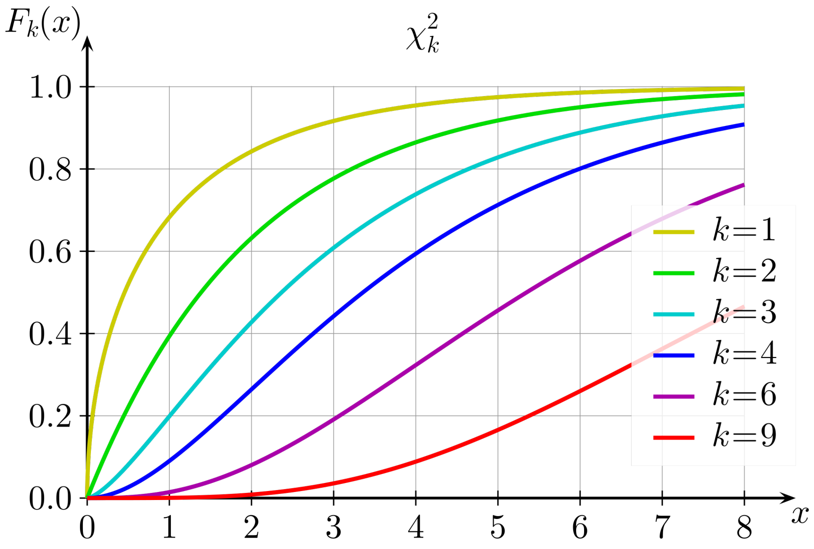
PDF, CDF

Probability Density Function:

$$Pr(x=X)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2}$$

Cumulative Distribution Function

$$Pr(X>x)=\frac{\gamma\left(\frac{k}{2},\frac{x}{2}\right)}{\Gamma(k/2)}$$



Derive χ² k=1 from N(0,1)

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Let f be the pdf of X^2. Then

$$f(x) = \frac{d}{dx}\Pr(X^2 \le x) = \frac{d}{dx}\Pr(-\sqrt{x} \le X \le \sqrt{x})$$

$$= \frac{d}{dx}\frac{1}{\sqrt{2\pi}}\int_{-\sqrt{x}}^{\sqrt{x}} e^{-u^2/2}\,du = \frac{2}{\sqrt{2\pi}}\frac{d}{dx}\int_0^{\sqrt{x}} e^{-u^2/2}\,du$$

$$= \frac{2}{\sqrt{2\pi}} e^{-\sqrt{x}^2/2}\frac{d}{dx}\sqrt{x} = \frac{2}{\sqrt{2\pi}} e^{-x/2}\frac{1}{2\sqrt{x}}$$

$$= \frac{e^{-x/2}}{\sqrt{2\pi x}}.$$

G-test (recommended over chi-square)

G-test:
https://en.wikipedia.org/wiki/G-test

G-test is a likelihood ratio test (maximum likelihood).