Introductory Statistics
3: Data Collection (Experiments)

Richard Veale

Graduate School of Medicine
Kyoto University

https://youtu.be/uo0aUEteM8I

Lecture Video at above link

Summary

Experiments and Clinical Trials
- Historical Trial: Scurvy
- Historical Trial: Beri-Beri
- Recent Trial: Osteoarthritis Surgery
- Double-blinding
- Trial design
- Ethics

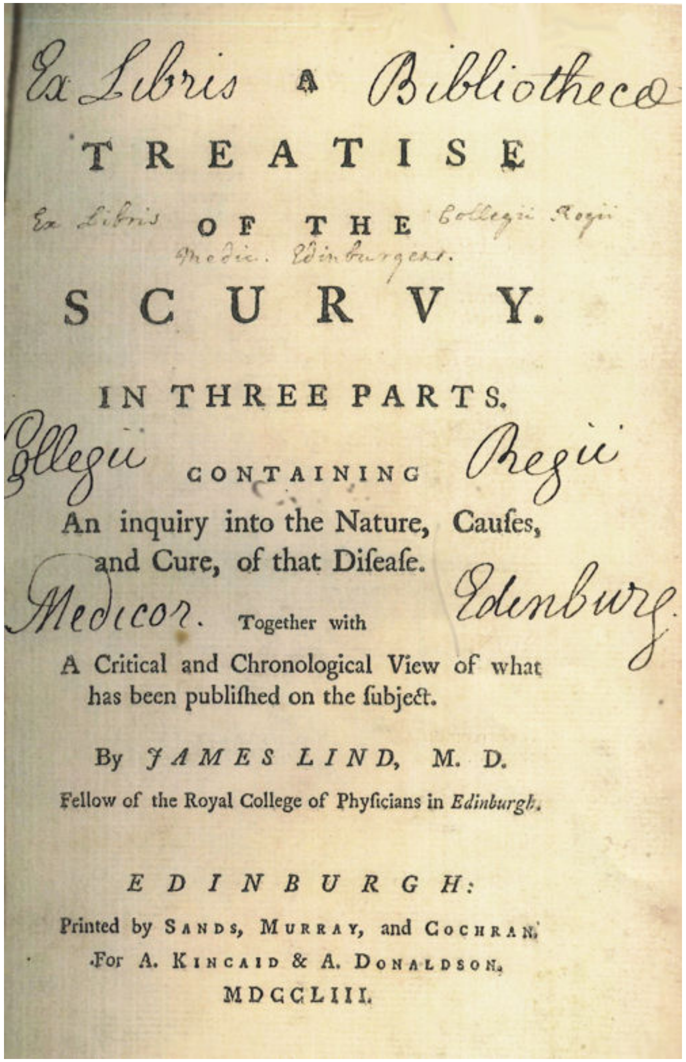
Practice
- Statistical Independence
- Expected frequencies in contingency tables

Scurvy

On long sea journeys (4-6 weeks), sailors would often develop *scurvy*.
- Spongy, bleeding gums
- Bleeding under the skin
- Extreme weakness

JamesLindLibrary.com

Scurvy


Ex Libris Bibliotheca

TREATISE

Ex Libris OF THE Collegii Regii

Medic. Edinburgens.

SCURVY.

IN THREE PARTS.

Collegii CONTAINING Regii

An inquiry into the Nature, Caufes, and Cure, of that Difeafe.

Medicor. Together with Edinburg.

A Critical and Chronological View of what has been publifhed on the fubject.

By *JAMES LIND*, M. D.

Fellow of the Royal College of Phyficians in *Edinburgh*.

EDINBURGH:

Printed by SANDS, MURRAY, and COCHRAN,

For A. KINCAID & A. DONALDSON,

MDCCLIII.

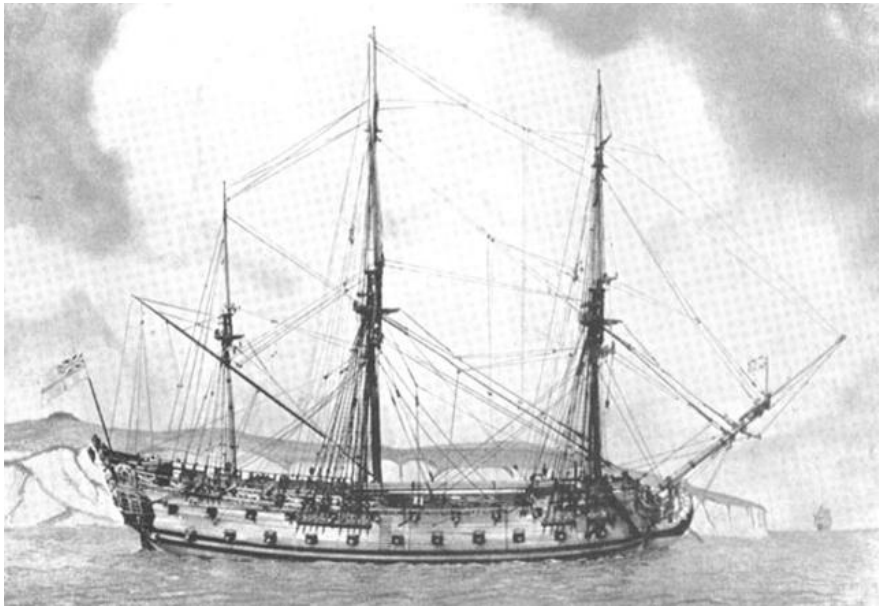
Physician James Lind tried different types of treatment for scurvy on a sea voyage 1747 and published the results 1753.

James Lind, 1716-1794

Jameslindlibrary.com

Scurvy

Sutton, Journal of the Royal Society of Medicine, 2003



HMS Salisbury

Lind separated the patients in six treatment groups.

Treatment	Cider	Vinegar	Sea water	Lemon	Elixir	Garlic
Patients	2	2	2	2	2	2

Scurvy

Two patients in the lemon group improved

Treatment	Cider	Vinegar	Sea water	Lemon	Elixir	Garlic
Patients	2	2	2	2	2	2
Improved	0	0	0	2	0	0

Scurvy

The conſequence was, that the moſt ſudden and viſible good effects were perceived from the uſe of the oranges and lemons; one of thoſe who had taken them, being at the end of ſix days fit for duty. The ſpots were not indeed at that time quite off his body, nor his gums ſound; but without any other medicine, than a gargariſm of *elixir vitriol*, he became quite healthy before we came into *Plymouth*, which was on the 16th of *June*. The other was the beſt recovered of any in his condition; and being now deemed pretty well, was appointed nurſe to the reſt of the ſick.

We now know scurvy was is caused by vitamin C deficiency (vitamin C is found in lemons..).

Beri-Beri

Kanehiro Takaki
(1849-1920)

from Miyazaki prefecture

Medical Officer of the Japanese Navy

1875-1880:
St. Thomas Hospital Medical School, London

Investigated Beri-Beri

http://www.pref.miyazaki.lg.jp

Beri-Beri

かっけ
Beri-Beri (脚気) was a wide-spread disease in the Japanese Imperial Navy in the late 1800s.

“Beri-Beri” from Sinhalese “I cannot, I cannot”

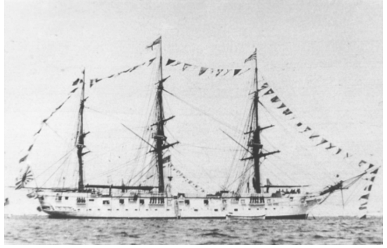
Deficits of the peripheral nervous system, e.g., difficulties with walking, tingling or loss of sensation, loss of tendon reflexes.

Deficits of the heart/blood circulation, increased heart rate, with or without peripheral edema (“wet” vs. “dry” beri-beri).

Adam’s and Victor’s Principles of Neurology

Beri-Beri

Ryujo

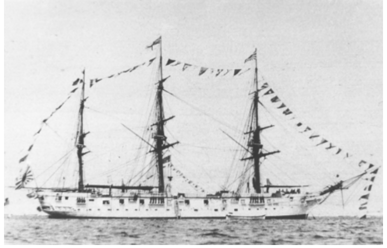


Travel to New Zealand, Hawaii, South America (rations: polished rice)

169 of 376 crew members suffered from Beri-Beri

1882/1883

Tsukuba



Same trip, different diet (rations: protein/nitrogen-rich, vegetable, meat, fish)

14 of 333 crew members suffered from Beri-Beri

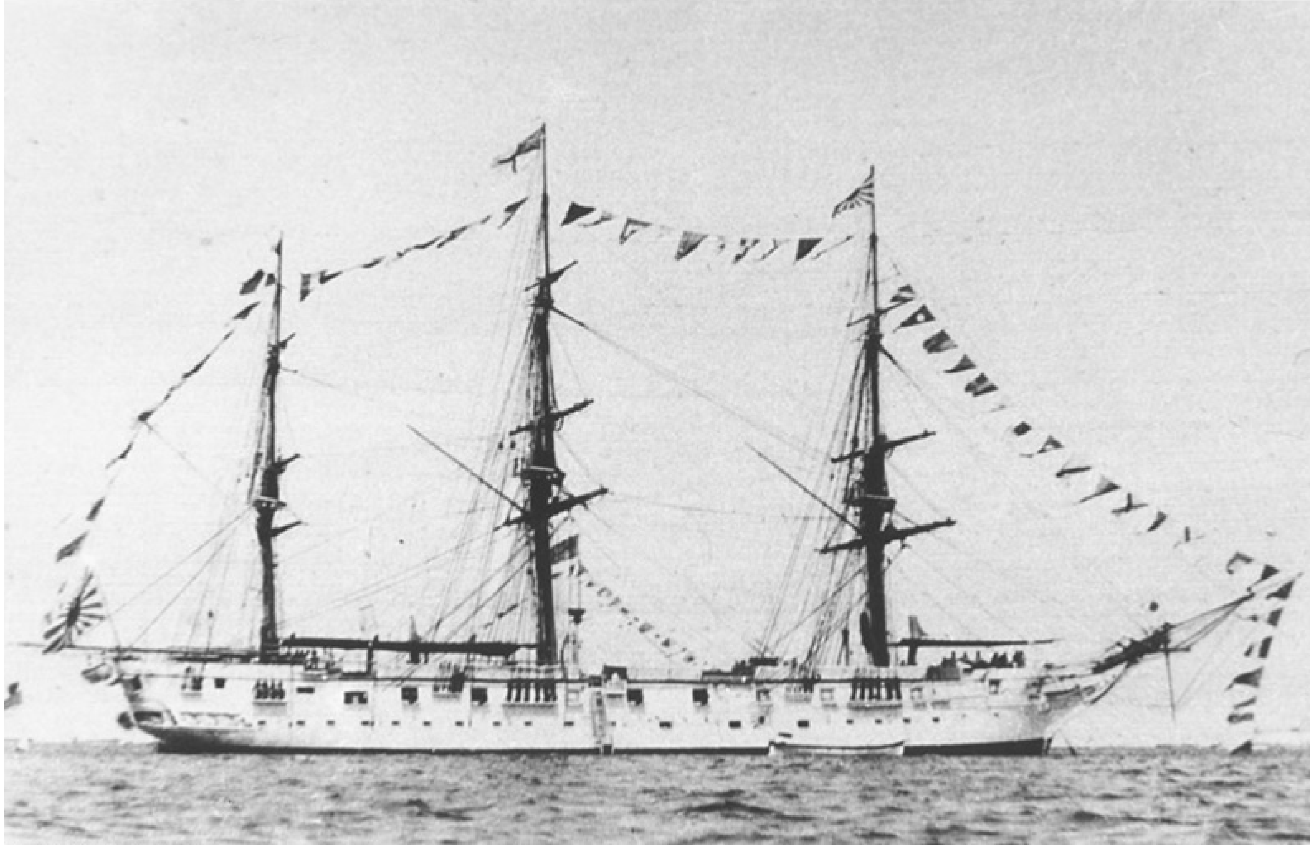
1884

Beri-Beri

Manchuria, 1904, ward with patients of the Japanese army

It took some time for the new diet to be introduced.
During the Russian-Japanese war, 80.000 soldiers were sent home because of Beri-Beri, 10% died. (Hawk, 2006)

Beri-Beri



Sailing ship Tsukuba

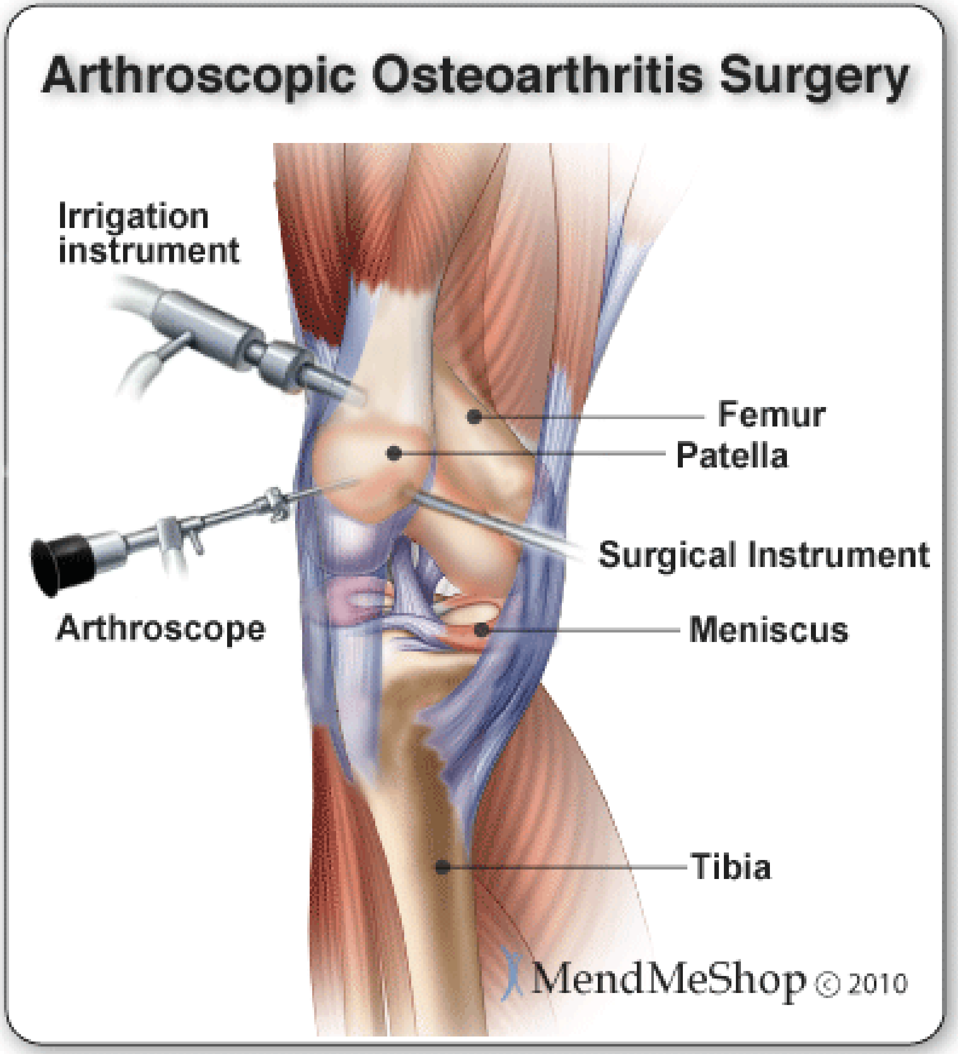
We now know Vitamin B1 (thiamine) deficiency causes Beri-Beri.

Arthroscopic Surgery

1996:
Performed 650,000 times/year in the USA

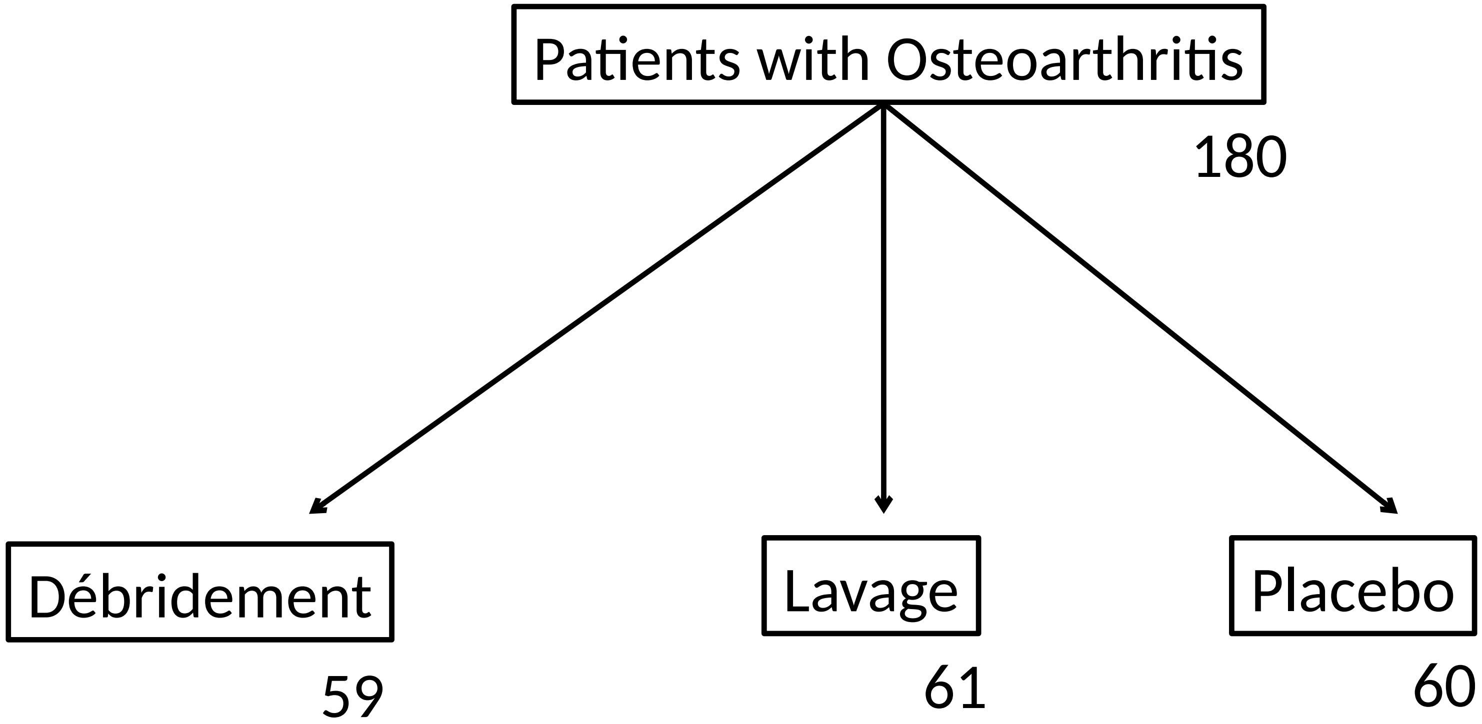
Cost per operation: USD 5000$

Does it really help?



www.mendmyknee.com

Arthroscopic Surgery



Moseley et al. conducted a study to test the advantages of debridement (getting rid of small debris, cutting damaged tissue) and lavage (washing) over placebo (sham treatment).

Moseley et al., New England Journal of Medicine, 2002

Arthroscopic Surgery

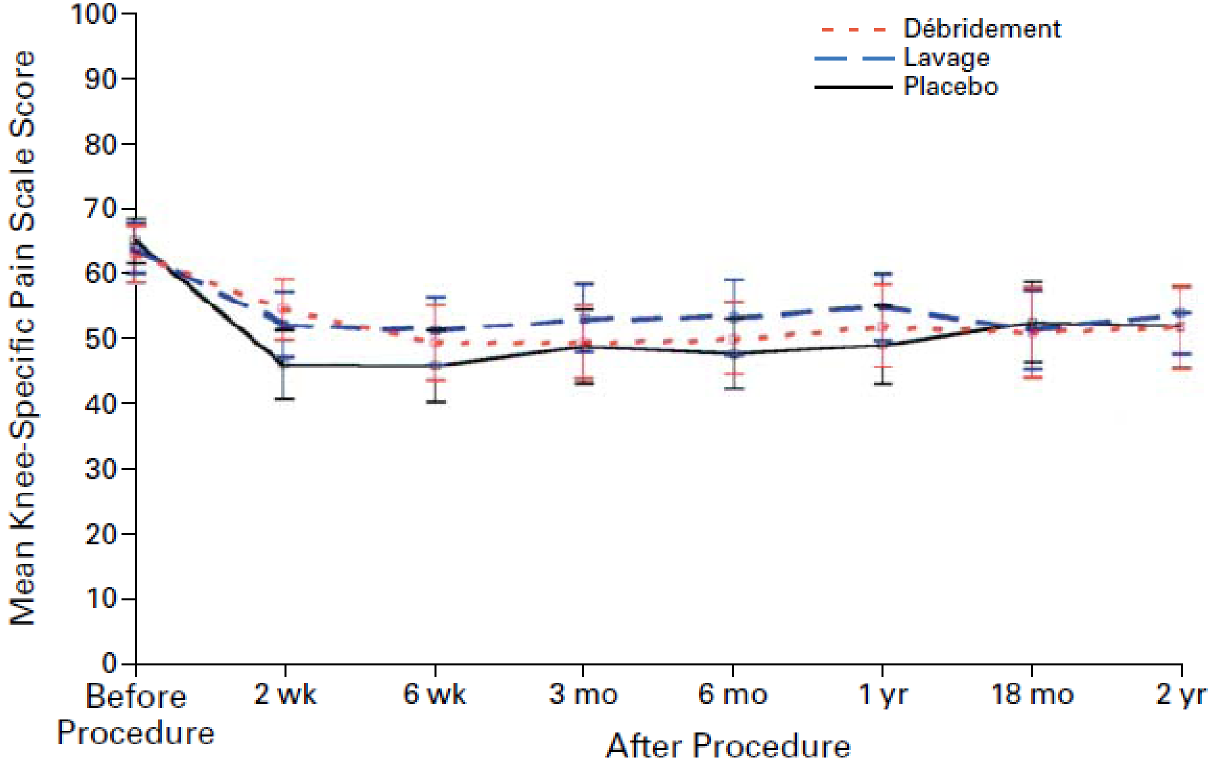
KNEE-SPECIFIC PAIN SCALE.*

ITEM	RANGE OF RESPONSES	MEANING OF RESPONSES
Pain magnitude		
Pain intensity		
1. How much pain are you currently having in your left/right knee?	1–7	Severe pain to no pain
2. At the present time (right now), how intense is your left/right knee pain?	0–10	No pain to bad as could be
3. In the past week, how intense was your worst left/right knee pain?	0–10	No pain to bad as could be
4. In the past week, on the average, how intense was your left/right knee pain?	0–10	No pain to bad as could be
Pain frequency		
5. On about how many days have you had knee pain in the past week in your left/right knee?	0–7	No days to all days
6. On days when you've had knee pain in the past week, how many hours were you usually in pain in your left/right knee?	0–24	No hours to all hours
7. On about how many days in the past week have you been kept from your usual activities (work, school, housework) because of the pain in your left/right knee?	0–7	No days to all days
Pain distastefulness		
8. Compared to other people your age, do you rate your situation regarding pain as . . .	1–5	Very poor to excellent
9. How satisfied are you with your current situation regarding pain?	1–5	Very satisfied to very dissatisfied
10. How pleased are you with your current situation regarding pain?	1–5	Very displeased to very pleased
11. How much of a problem do you have with pain because of your left/right knee?	1–6	None to very severe
12. In the past week, how unpleasant or distressing was your left/right knee pain?	1–10	No pain to bad as could be

*To calculate the total score, subtract the scores for items 1, 8, and 10 from the highest possible scores for those items +1 (to reverse the direction of scores, in keeping with the scores for the other items), rescale all items to a 0-to-10 scale, and add the scores for the items in each group (intensity, frequency, and distastefulness). Then average the scores for intensity and frequency to create the final pain-magnitude score, and average the sums of the pain-magnitude and pain-distastefulness scores to generate a total score. Each patient received a survey regarding his or her study knee; only the word "left" or "right" was included on each survey.

Their outcome ("success") measure was subjective pain perception determined with a standardized questionnaire.

Surgery: Did it work?



No. at Risk								
Placebo	60	59	57	56	57	53	52	55
Lavage	61	59	57	59	59	57	56	55
Débridement	58	59	59	58	56	50	51	53

Figure 1. Mean Values (and 95 Percent Confidence Intervals) on the Knee-Specific Pain Scale.
Assessments were made before the procedure and 2 weeks, 6 weeks, 3 months, 6 months, 12 months, 18 months, and 24 months after the procedure. Higher scores indicate more severe pain.

Surgery: Did it work?

At no point did either arthroscopic-intervention group have greater pain relief than the placebo group (Fig. 1, Table 2, and Supplementary Appendix 2). For example, there was no difference in knee pain between the placebo group and either the lavage group or the débridement group at one year (mean [±SD] KSPS scores, 48.9±21.9, 54.8±19.8, and 51.7±22.4, respectively; P=0.14 for the comparison with the lavage group, and P=0.51 for the comparison with the débridement group) or at two years (mean KSPS scores, 51.6±23.7, 53.7±23.7, and 51.4±23.2, respectively; P=0.64 and P=0.96, respectively). Similarly, there was no significant difference in arthritis pain between the placebo group and the lavage group or the débridement group at one or two years (Table 2).

Blinding

Two things could confound our results:

1) Some patients might feel better ***just because*** they think they got surgery (*placebo* effect).

2) The experimenters may have unconsciously treated the scores of placebo patients lower than the scores of patients who got real surgery.

Blinding

Two things could confound our results:

1) Some patients might feel better ***just because*** they think they got surgery (*placebo* effect).

2) The experimenters may have unconsciously treated the scores of placebo patients lower than the scores of patients who got real surgery.

To remedy this, experiments will ***double-blind.***

-*Double* because you blind both the patients and the experimenters!

Blinding

Two things could confound our results:

1) Some patients might feel better ***just because*** they think they got surgery (*placebo* effect).

2) The experimenters may have unconsciously treated the scores of placebo patients lower than the scores of patients who got real surgery.

If only the patients are unaware of which group they are in, it is *single blind*

Blinding

Researchers "want" their treatment to be effective and are prone to interpret their data such (not necessarily consciously).

This is an **expectation effect**:

1) data is scored and interpreted in a way to support the initial hypothesis.
2) the patients are given (subtle) cues that they are expected to improve.
   (for example, maybe the doctor smiles more at the patients)

"**Double-blinded**" means that not only the patients but also the researchers are unaware of whether a group receives placebo or treatment.

Trial Design

Between-group design

Double-blinded placebo-controlled randomized trial

2 groups are compared with each other (placebo and treatment group)

Disadvantage: care has to be taken that the groups are not to different from each other to begin with

Within-group design

One single group is treated, but with placebo/active treatment at different time-points

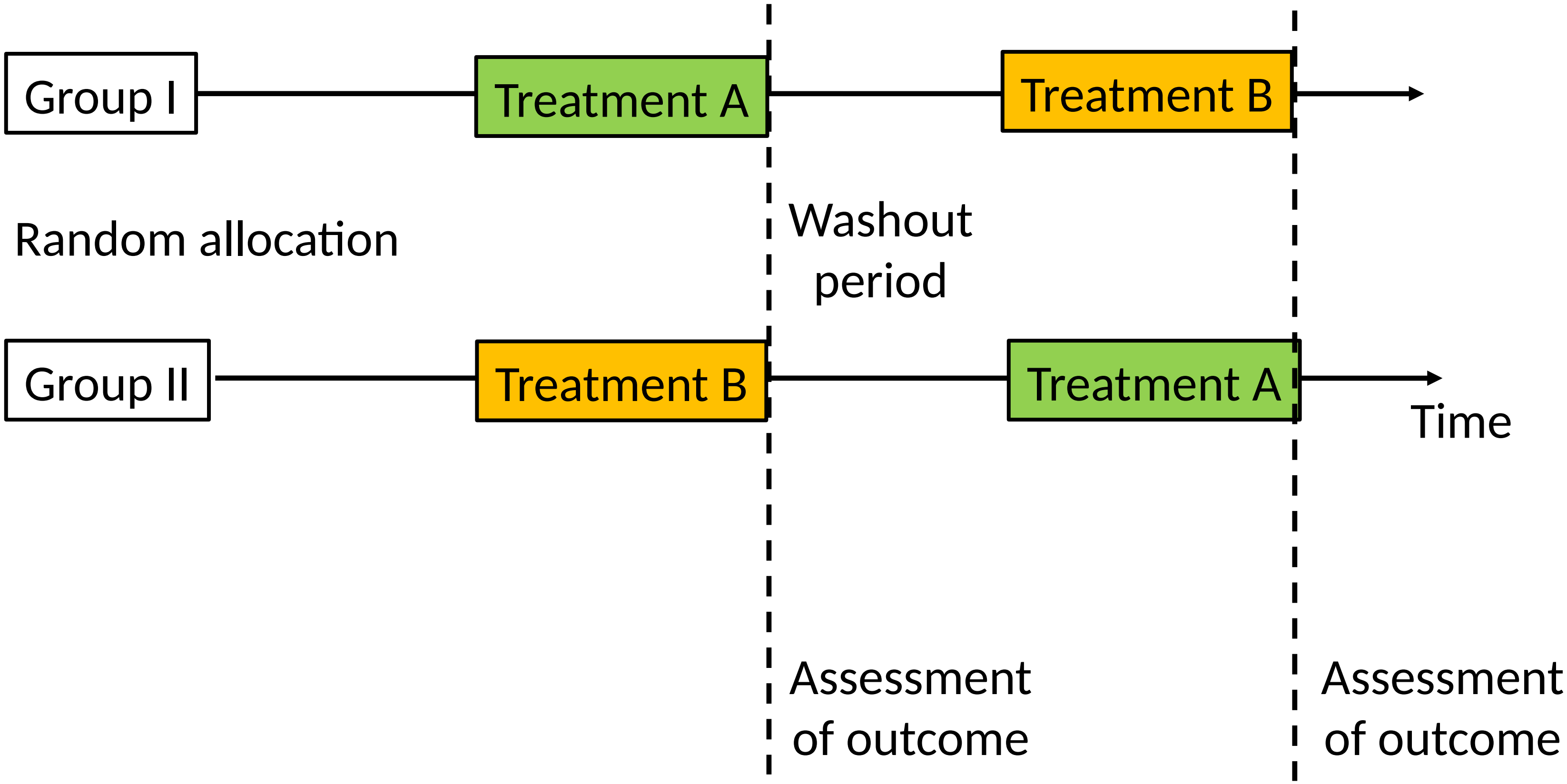
Used for example when treatment should be provided at some point

Not practical for some one-time treatments (e.g., surgeries) or for treatments with long-lasting effect (again, surgeries).

Trial Design

Within-group design: **<u>Cross-over group trial</u>**

→ We need to randomize order of treatments to remove it as a *confound* (*confound* is something that messes up your results)



Trial Design

Example of a within-group design:

Plasma glucose level with 2 different insulin pumps

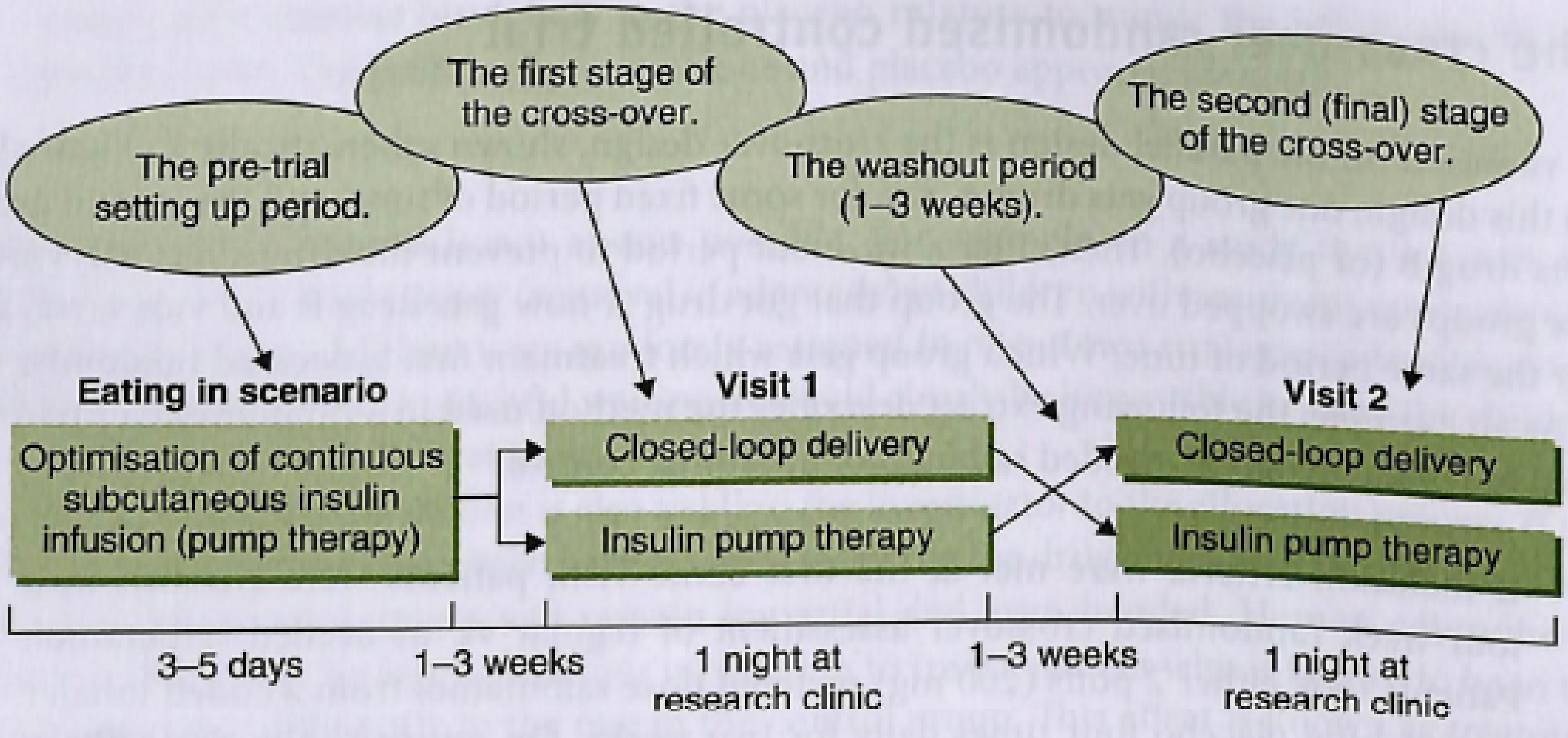


Figure 9.3 Part of a cross-over study comparing closed-loop delivery of insulin with conventional insulin pump therapy. This shows the Eating In scenario only. Source: Hovorka *et al.* (2011). Reproduced by permission of BMJ Publishing Group Ltd

Ethics…

To blind the patients, they had to “trick” the patients and in the placebo group and put them in danger (cut open their skin, anesthetize, etc.).

Ethics...

It is difficult to get ethics approval for such experiments.

But, often they are necessary to prove that a treatment is effective...

→ Maybe easier in situations where the disease is very dangerous (“nothing to lose”)

Placebo Procedure

To preserve blinding in the event that patients in the placebo group did not have total amnesia, a standard arthroscopic débridement procedure was simulated. After the knee was prepped and draped, three 1-cm incisions were made in the skin. The surgeon asked for all instruments and manipulated the knee as if arthroscopy were being performed. Saline was splashed to simulate the sounds of lavage. No instrument entered the portals for arthroscopy. The patient was kept in the operating room for the amount of time required for a débridement. Patients spent the night after the procedure in the hospital and were cared for by nurses who were unaware of the treatment-group assignment.

Postoperatively, there were two minor complications and no deaths. Incisional erythema developed in one patient, who was given antibiotics. In a second patient, calf swelling developed in the leg that had undergone surgery; venography was negative for thrombosis. In no case did a complication necessitate the breaking of the randomization code.

Postoperative care was delivered according to a protocol specifying that all patients should receive the same walking aids, graduated exercise program, and analgesics. The use of analgesics after surgery was monitored; during the two-year follow-up period, the amount used was similar in the three groups.

Cheating in clinical trials

Ben Goldacre bemoans some shady practices in pharma industry:

- Publishing only good trials (unfavorable vanish in “drawers”) .

- Comparison with placebo only, not with other drug (so we don’t know the new drug is better or worse than the currently used one) .

- When comparing to another drug, using a drug dose that is too high or too low (the new drug will look much better than the currently used one).

How do we know if something works?

Now we will start to learn how statistics determines if a surgery works

-Or how we know that smoking causes lung cancer

-Or how we know that men like ramen

We check for ***statistical independence***

→ If the treatment has no effect on the outcome (i.e. outcome is "independent" of treatment), then the treatment *does not work*. "Null hypothesis"

→ Otherwise it *works* (*refuted the null hypothesis*)

Statistical Independence in 2x2 Designs

Are you Man?

Like Ramen?

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

Based on this data, would you say that men like ramen more?

→ To see if it is true, we assume the opposite, i.e. that being a man and liking ramen are *statistically independent*.

If we find they are **not** independent: men like ramen more!

(Actually we have to check that non-men don't like it more first...we'll learn that later)

Statistical Independence in 2x2 Designs

Are you Man?

Like Ramen?

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

These are counts/absolute frequencies: **N()**
→ All groups added together are our total
We will use probability of: **P()**
→ All groups added together are 1.0

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	N(M&L)	N(!M&L)	N(L)
No	N(M&!L)	N(!M&!L)	N(!L)
TOTAL	N(M)	N(!M)	N

Set operations:
!x means "complement of x"

x & y means "intersection of x and y"

x | y means "union of x and y"

→ M is set of men

→ L is set of those who like ramen

So: !M means "set of those not man"

Statistical Independence in 2x2 Designs

Absolute Frequency
COUNT

Are you Man?

Like Ramen?

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

Divide by total (10 in this case) to normalize

Relative Frequency
PROBABILITY

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.5	0.1	0.6
No	0.3	0.1	0.4
TOTAL	0.8	0.2	1.0

Statistical Independence in 2x2 Designs

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.5	0.1	0.6
No	0.3	0.1	0.4
TOTAL	0.8	0.2	1.0

These are counts/absolute frequencies: **N()**
→ All groups added together are our total
We will use probability of: **P()**
→ All groups added together are 1.0

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	P(M&L)	P(!M&L)	P(L)
No	P(M&!L)	P(!M&!L)	P(!L)
TOTAL	P(M)	P(!M)	P

Set operations:
!x means "complement of x"

x & y means "intersection of x and y"

x | y means "union of x and y"

→ M is set of men

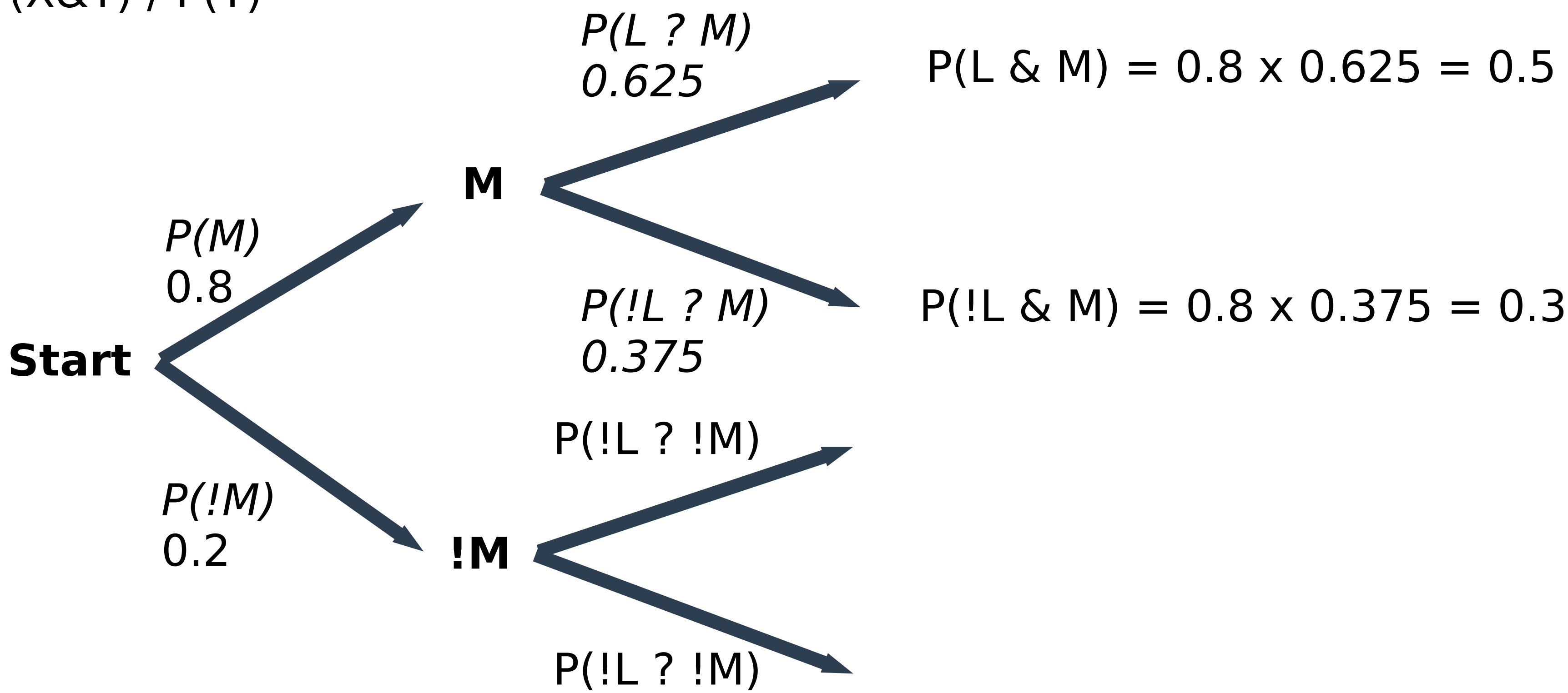
→ L is set of those who like ramen

So: !M means "set of those not man"

As a tree…

<u>Conditional probability:</u>

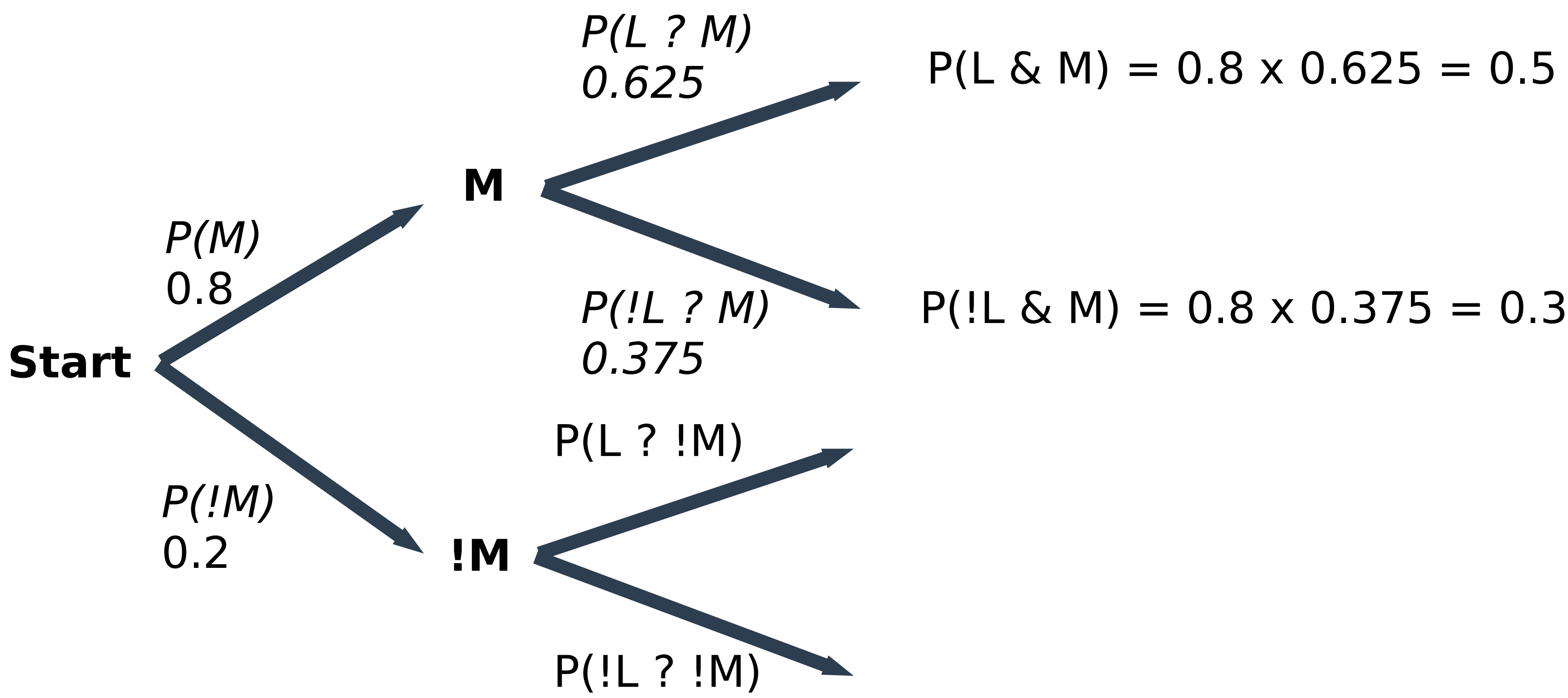
X ? Y means “X given Y is true”
P(L ? M): probability you like ramen *given* you are a male.
P(X?Y) = P(X&Y) / P(Y)



Statistical Independence

If: P(L ? M) == P(L ? !M)
(probability you like it given you are male is same as if you are not male)

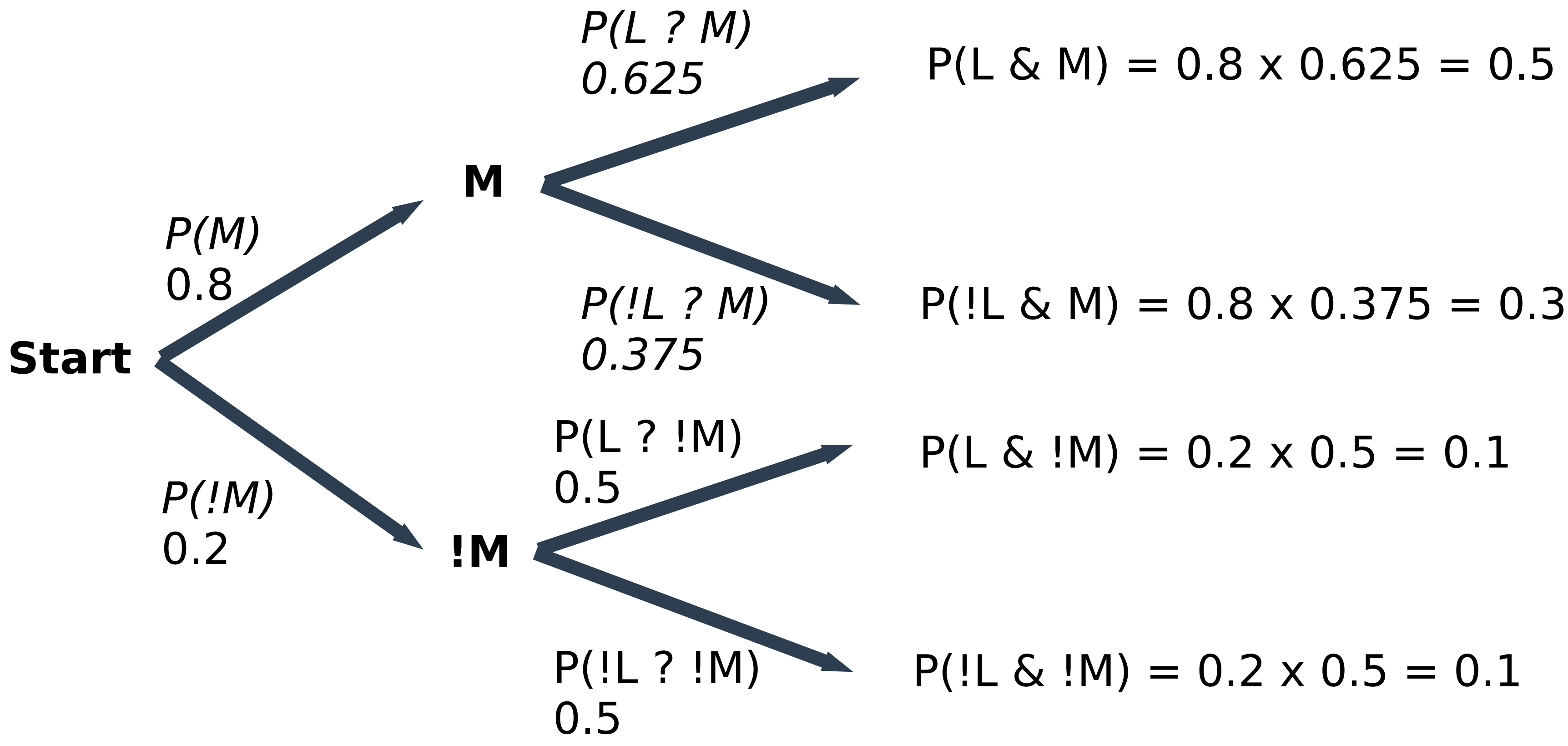
<u>→ Then it is statistically independent.</u>



Statistical Independence

If: P(L ? M) == P(L ? !M) (== P(L))
(probability you like it given you are male is same as if you are not male)

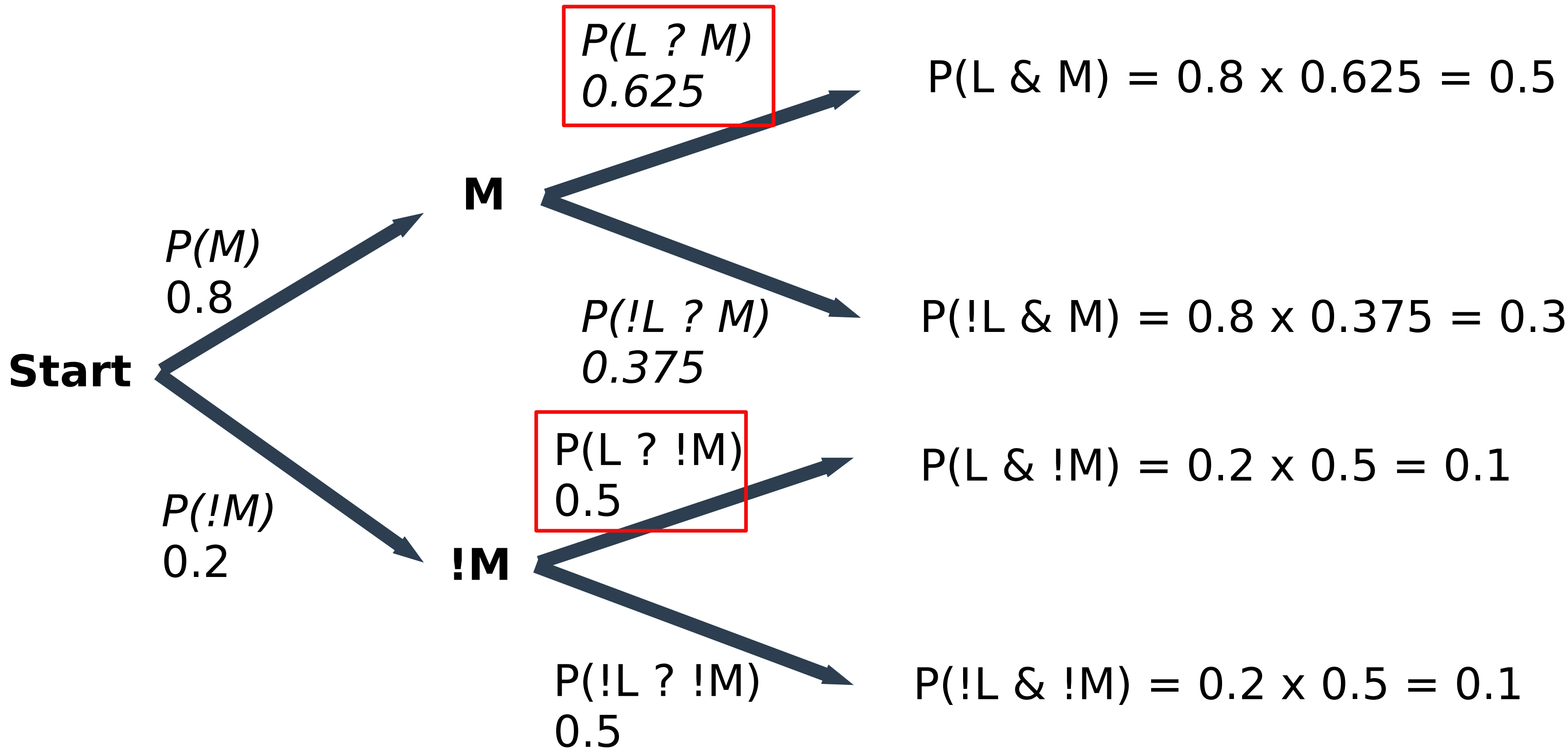
→ Then it is statistically independent.



Statistical Independence

If: P(L ? M) == P(L ? !M) (== P(L))
(probability you like it given you are male is same as if you are not male)

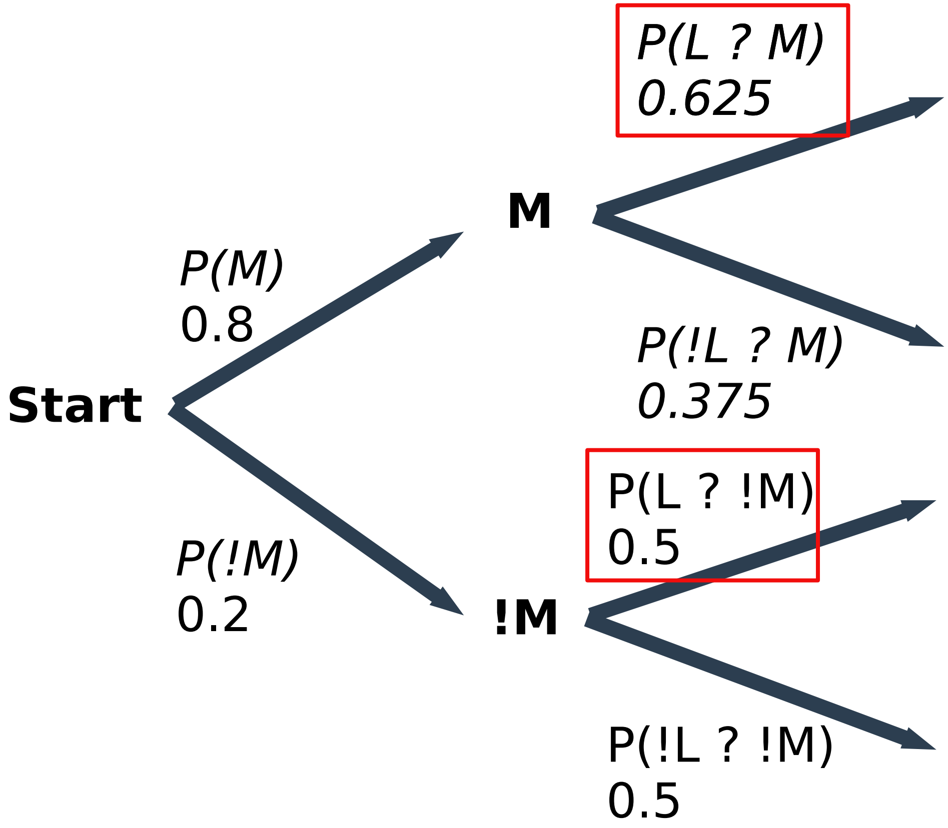
<u>→ Then it is statistically independent.</u>



Statistical Independence

If: P(L ? M) == P(L ? !M) (== P(L))
(probability you like it given you are male is same as if you are not male)

<u>→ Then it is statistically independent.</u>



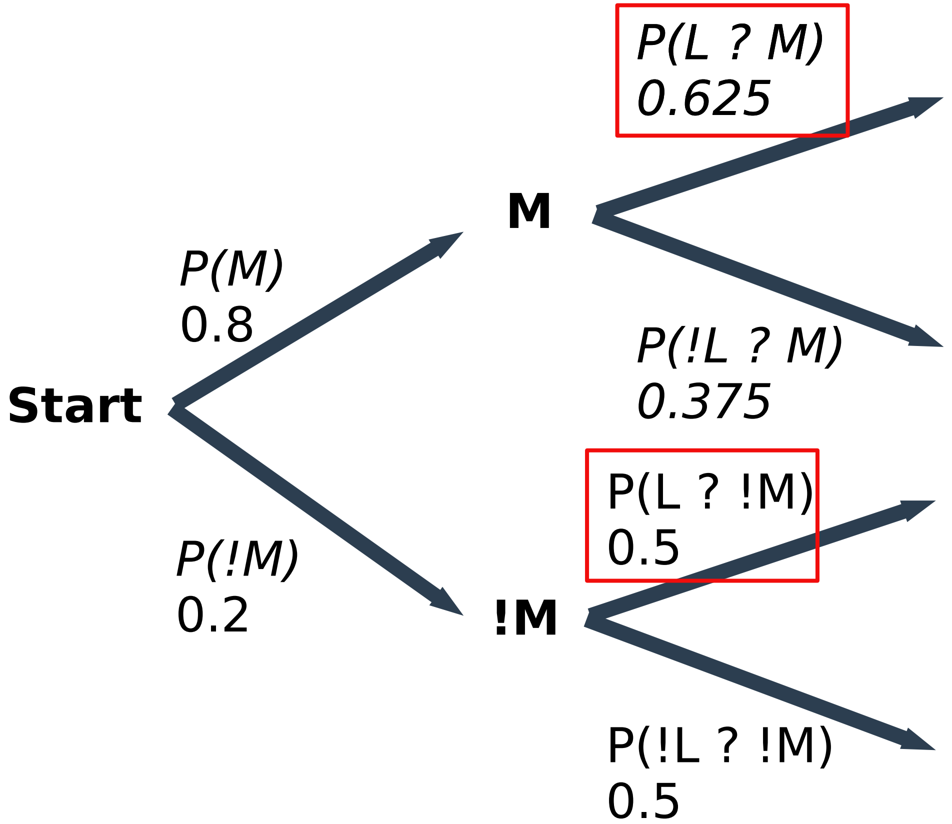
If liking it was statistically independent of being a man (given our 10 samples), these would be the same

They are not...

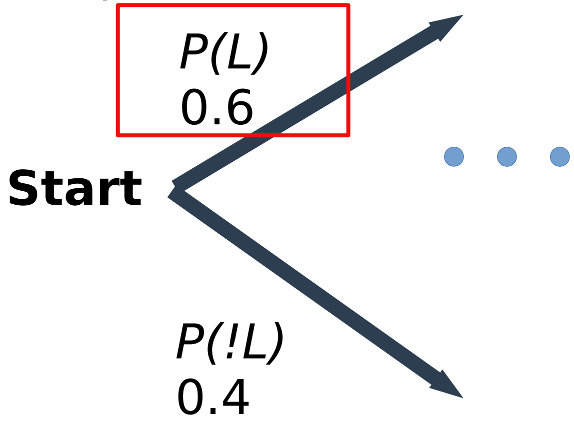
Statistical Independence

If: P(L ? M) == P(L ? !M) (== P(L))
(probability you like it given you are male is same as if you are not male)

→ Then it is statistically independent.



If they were the same, they would also be the same as this automatically (check it yourself)



But it's not.

Expected Probabilities

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.5	0.1	0.6
No	0.3	0.1	0.4
TOTAL	0.8	0.2	1.0

Expected Probabilities

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes			0.6
No			0.4
TOTAL	0.8	0.2	1.0

We can’t change how many men/non-men we have...

Expected Probabilities

Let's "force" statistical independence.

This is called "expected probabilities" (i.e. what we'd expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes			0.6
No			0.4
TOTAL	0.8	0.2	1.0

...Or how many people say they like/dislike ramen

We can't change how many men/non-men we have...

Expected Probabilities

Let's "force" statistical independence.

This is called "expected probabilities" (i.e. what we'd expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes			0.6
No			0.4
TOTAL	0.8	0.2	1.0

We **can** play around with the relative numbers of people who are in each category though!

Expected Probabilities

Let's "force" statistical independence.

This is called "expected probabilities" (i.e. what we'd expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes			0.6
No			0.4
TOTAL	0.8	0.2	1.0

If we want to "equally distribute" the probability, we will just multiply the marginals…

i.e. probably of man AND like ramen = probability man times probability likes ramen

Expected Probabilities

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.6 x 0.8	0.6 x 0.2	0.6
No	0.4 x 0.8	0.4 x 0.2	0.4
TOTAL	0.8	0.2	1.0

This is what we **“expect”** if there is no statistical dependence of any variable on any other...

Expected Probabilities

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

These still all add up to 1.0 (otherwise we made a mistake…)

Are you Man?

Like Ramen?	Yes	No	TOTAL
Yes	0.48	0.12	0.6
No	0.32	0.08	0.4
TOTAL	0.8	0.2	1.0

These are **expected probabilities**

For our specific set of people (10 people), we can go back to absolute frequencies (un-normalize).
→ remember we just divided by 10 to get probabilities in the first place!

Expected Counts

Let's "force" statistical independence.

This is called "expected probabilities" (i.e. what we'd expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.48	0.12	0.6
No	0.32	0.08	0.4
TOTAL	0.8	0.2	1.0

If it adds up to 1 now, and we want to make it add up to 10, what do we do?

→ multiply by 10!

Expected Counts

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	0.48 x 10	0.12 x 10	0.6 x 10
No	0.32 x 10	0.08 x 10	0.4 x 10
TOTAL	0.8 x 10	0.2 x 10	1.0 x 10

If it adds up to 1 now, and we want to make it add up to 10, what do we do?

→ multiply by 10!

Expected Counts

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	4.8	1.2	6
No	3.2	0.8	4
TOTAL	8	2	10

If it adds up to 1 now, and we want to make it add up to 10, what do we do?

→ multiply by 10!

Expected Counts

Let’s “force” statistical independence.

This is called “expected probabilities” (i.e. what we’d expect if they were independent)

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	4.8	1.2	6
No	3.2	0.8	4
TOTAL	8	2	10

Whoa…we got “fractional people”…

Don’t worry, that will usually happen!

Expected vs Observed

Expected

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	4.8	1.2	6
No	3.2	0.8	4
TOTAL	8	2	10

Observed

Are you Man?

Like Ramen?

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

Expected vs Observed

Expected

Are you Man? / Like Ramen?

	Yes	No	TOTAL
Yes	4.8	1.2	6
No	3.2	0.8	4
TOTAL	8	2	10

Observed

Are you Man? / Like Ramen?

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

Those are pretty close…

Only off by 0.2 in all the squares…

Expected vs Observed

Expected

Are you Man?

Like Ramen?

	Yes	No	TOTAL
Yes	4.8	1.2	6
No	3.2	0.8	4
TOTAL	8	2	10

Observed

Are you Man?

Like Ramen?

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

Those are pretty close…

..But are they close enough to say that we observed statistically independent results?

Chi-squared / Fischer's Exact Test

Find out next week!

-We need to know more about the distribution of "possible" errors to know if this is really "chance" (luck) or not.

(i.e. is expecting 0.2 differences "normal" if we have 10 people like we do?)

→ Fischer's exact test uses hypergeometric distribution

→ Chi-squared test assumes chi-squred distribution...

JMP

JMP will automatically compute expected and observed counts for you!

See next week.

→ I will post the 1st homework (about making contingency tables)

Due: Friday *5 June* (2 weeks)