Statistics and Probability

What is statistics?

> We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.
>
> — Sir R. A. Fisher

What is statistics?

⟶ Data exploration and analysis.

⟶ Inductive inference with probability.

⟶ Quantification of evidence and uncertainty.

Example 1

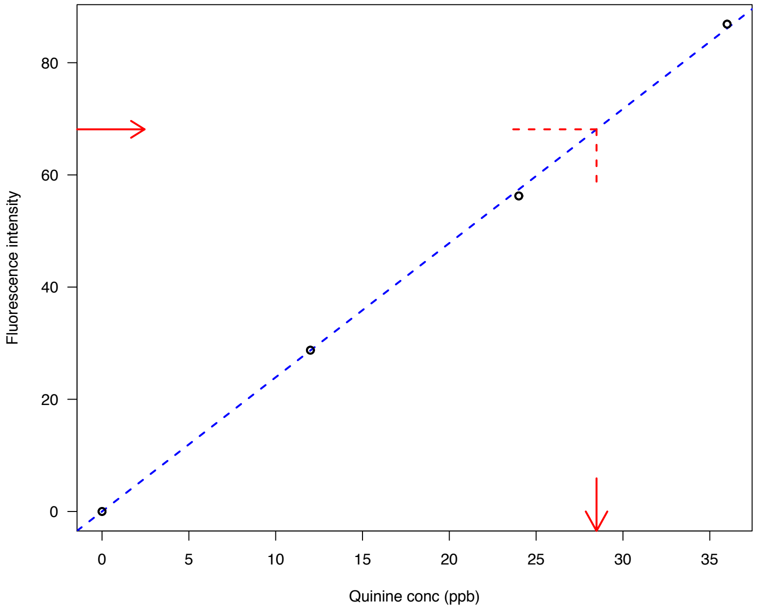
Goal: Determine, by fluoresence, the concentration of quinine in a sample of tonic water.

Method:

1. Obtain a stock solution with known concentration of quinine.
2. Create several dilutions of the stock.
3. Measure fluoresence intensity of each such standard.
4. Measure fluoresence intensity of the unknown.
5. Fit a line to the results for the standards.
6. Use line to estimate quinine concentration in the unknown.

Question: How precise is the resulting estimate?

Example 1



Example 2

Children that have positive response to a pertussis antigen:

○ Vaccinated with DTaP-HBV: 3/38 (8%)

○ History of pertussis infection: 5/21 (24%)

Questions:

⟶ How precisely can we estimate the chance of a positive response given vaccination?

⟶ Are the above rates truly different?

Reference: Esposito et al. (2001) *Infection and Immunity* **69**:4516–20.

Example 3

Place tick on clay island surrounded by water, with two capillary tubes: one treated with deer-gland-substance, and one untreated.

Tick sex	Leg	Deer sex	treated	untreated
male	fore	female	24	5
female	fore	female	18	5
male	fore	male	23	4
female	fore	male	20	4
male	hind	female	17	8
female	hind	female	25	3
male	hind	male	21	6
female	hind	male	25	2

⟶ Is the tick more likely to go to the treated tube?

⟶ Do the sex of the tick or deer, or the leg from which the gland substance was obtained, have an effect?

Reference: Carroll (2001), *J Med Entomol* **38**:114–7.

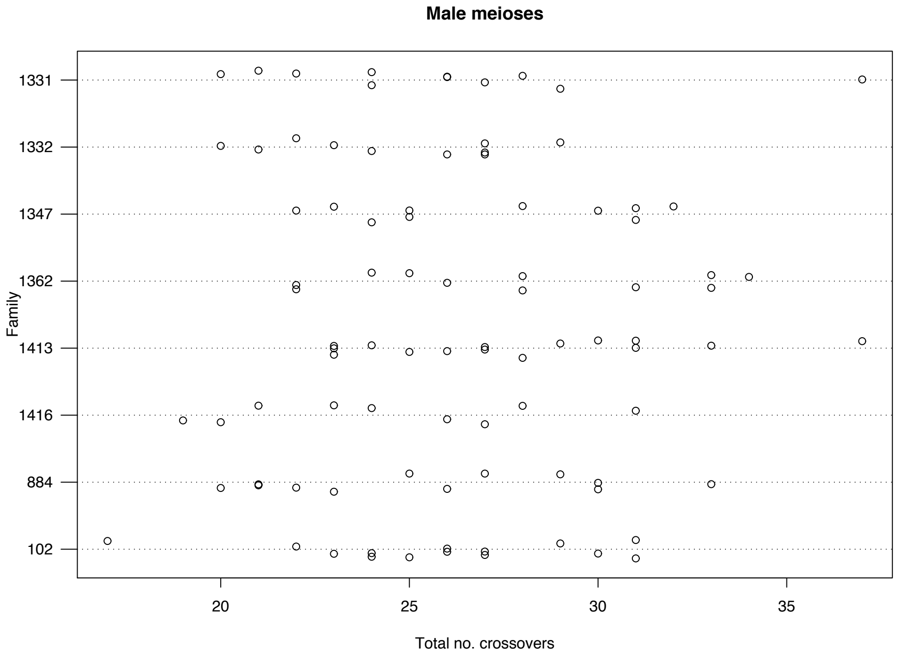
Example 4

For each of 8 fathers and 8 mothers, we observe (estimates of) the number of crossovers, genome-wide, in a set of independent meiotic products.

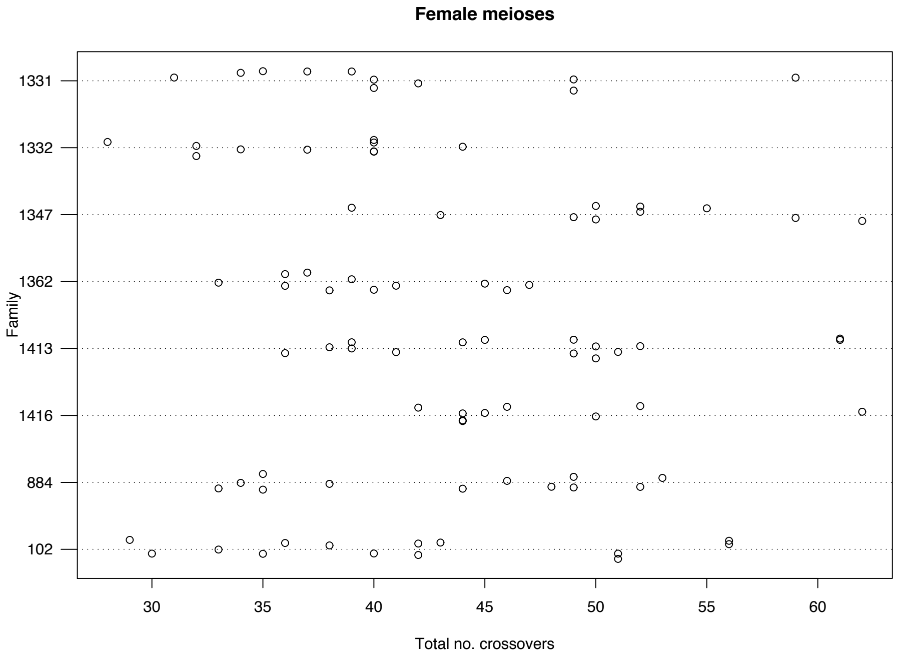
⟶ Do the fathers (or mothers) vary in the number of crossovers they deliver?

⟶ Are there differences between fathers and mothers with regards to the number of crossovers they deliver?

Example 4



Example 4



Example 4

How do we think about this?

If there were no relationship between family ID and number of crossovers in a meiotic product:

⟶ What sort of data would we expect?

⟶ What would be the chance of obtaining data as extreme as what was observed?

What is probability?

⟶ A branch of mathematics concerning the study of random processes.

Note: Random does not mean haphazard!

What do I mean when I say the following?

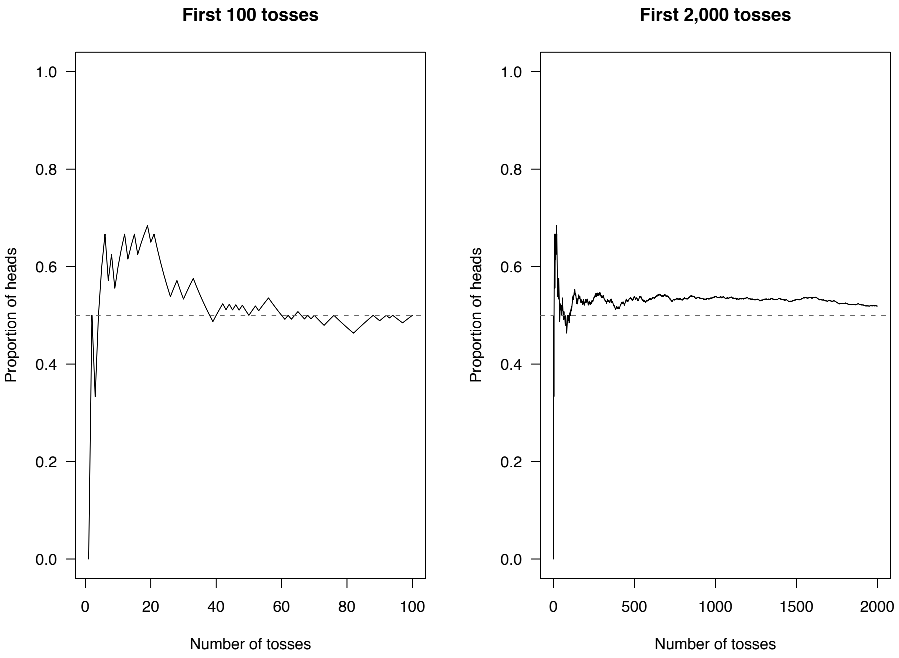
The probability that he is a carrier . . .

The chance of rain tomorrow . . .

⟶ Degree of belief.

⟶ Long term frequency.

Tossing a coin



The set-up

Experiment

→ A well-defined process with an uncertain outcome.

Draw 2 balls *with replacement* from an urn containing 4 red and 6 blue balls.

Sample space $\mathcal{S}$

→ The set of possible outcomes.

{ RR, RB, BR, BB }

Event

→ A set of outcomes from the sample space (a subset of $\mathcal{S}$).

{the first ball is red} = {RR, RB}

Events are said to occur if one of the outcomes they contain occurs. Probabilities are assigned to events.

Probability rules

$0 \leq$ Pr(A) ≤ 1	for any event A
Pr($\mathcal{S}$) = 1	where $\mathcal{S}$ is the sample space
Pr(A or B) = Pr(A) + Pr(B)	if A and B are *mutually exclusive*
Pr(not A) = 1 - Pr(A)	complement rule

Example

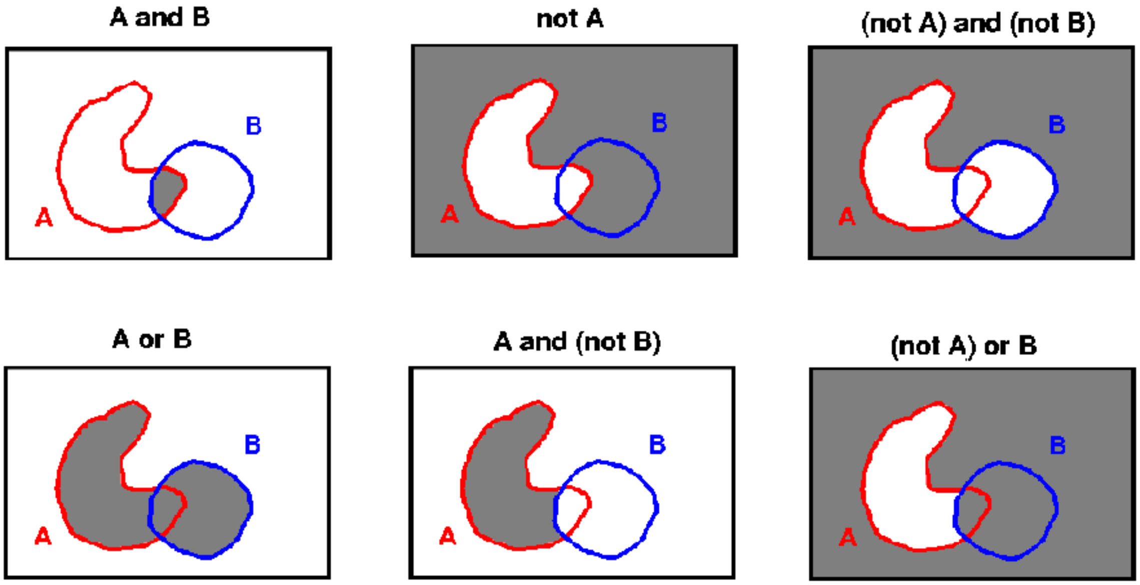
Cage with 10 rats:

- 2 infected with virus X (only)
- 1 infected with virus Y (only)
- 5 infected with both X and Y
- 2 infected with neither

Experiment: Draw one rat at random (each equally likely).

Events:	A = {rat is infected with X}	Pr(A) = 7/10
	B = {rat is infected with Y}	Pr(B) = 6/10
	C = {rat is infected with only X}	Pr(C) = 2/10

Sets



A and B

not A

(not A) and (not B)

A or B

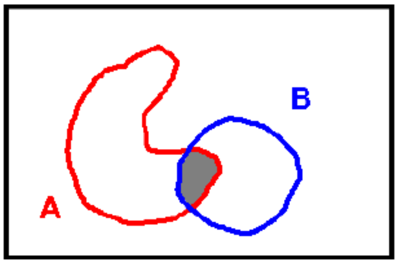
A and (not B)

(not A) or B

Conditional probability

Pr(A | B) = *Probability of A given B* = Pr(A and B) / Pr(B)

Rat example: [2 w/ X only; 1 w/ Y only; 5 w/ both; 2 w/ neither]



A = {infected with X}

B = {infected with Y}

Pr(A | B) = (5/10) / (6/10) = 5/6

Pr(B | A) = (5/10) / (7/10) = 5/7

More rules and a definition

Multiplication rule:

$\longrightarrow$ $\mathrm{Pr}(\mathrm{A\ and\ B}) = \mathrm{Pr}(\mathrm{A}) \times \mathrm{Pr}(\mathrm{B} \mid \mathrm{A})$

A and B are independent if $\mathrm{Pr}(\mathrm{A\ and\ B}) = \mathrm{Pr}(\mathrm{A}) \times \mathrm{Pr}(\mathrm{B})$

If A and B are independent:

$\longrightarrow$ $\mathrm{Pr}(\mathrm{A} \mid \mathrm{B}) = \mathrm{Pr}(\mathrm{A})$

$\longrightarrow$ $\mathrm{Pr}(\mathrm{B} \mid \mathrm{A}) = \mathrm{Pr}(\mathrm{B})$

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Sensitivity → Pr (positive test | disease)

Specificity → Pr (negative test | no disease)

Positive Predictive Value → Pr (disease | positive test)

Negative Predictive Value → Pr (no disease | negative test)

Accuracy → Pr (correct outcome)

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Sensitivity → TP / (TP+FN)

Specificity → TN / (FP+TN)

Positive Predictive Value → TP / (TP+FP)

Negative Predictive Value → TN / (FN+TN)

Accuracy → (TP+TN) / (TP+FP+FN+TN)

Diagnostics

Assume that some disease has a 0.1% prevalence in the population. Assume we have a test kit for that disease that works with 99% sensitivity and 99% specificity. What is the probability of a person having the disease given the test result is positive, if we randomly select a subject from

⟶ the general population?

⟶ a high risk sub-population with 10% disease prevalence?

Diagnostics

		DISEASE	
		+	−
TEST	+	99	999
	−	1	98901

Diagnostics

		DISEASE	
		+	−
TEST	+	99	999
	−	1	98901

Sensitivity	$\rightarrow$ 99 / (99+1) = 99%
Specificity	$\rightarrow$ 98901 / (999+98901) = 99%
Positive Predictive Value	$\rightarrow$ 99 / (99+999) $\approx$ 9%
Negative Predictive Value	$\rightarrow$ 98901 / (1+98901) $>$ 99.9%
Accuracy	$\rightarrow$ (99+98901) / 100000 = 99%

Diagnostics

		DISEASE	
		+	−
TEST	+	9900	900
	−	100	89100

Diagnostics

		DISEASE	
		+	−
TEST	+	9900	900
	−	100	89100

Sensitivity	→ 9900 / (9900+100) = 99%
Specificity	→ 89100 / (900+89100) = 99%
Positive Predictive Value	→ 9900 / (9900+900) ≈ 92%
Negative Predictive Value	→ 89100 / (100+89100) ≈ 99.9%
Accuracy	→ (9900+89100) / 100000 = 99%

Bayes rule

$$\longrightarrow \quad \Pr(\text{A and B}) = \Pr(\text{A}) \times \Pr(\text{B} \mid \text{A}) = \Pr(\text{B}) \times \Pr(\text{A} \mid \text{B})$$

$$\longrightarrow \quad \Pr(\text{A}) = \Pr(\text{A and B}) + \Pr(\text{A and not B})$$
$$= \Pr(\text{B}) \times \Pr(\text{A} \mid \text{B}) + \Pr(\text{not B}) \times \Pr(\text{A} \mid \text{not B})$$

$$\longrightarrow \quad \Pr(\text{B}) = \Pr(\text{B and A}) + \Pr(\text{B and not A})$$
$$= \Pr(\text{A}) \times \Pr(\text{B} \mid \text{A}) + \Pr(\text{not A}) \times \Pr(\text{B} \mid \text{not A})$$

$$\longrightarrow \quad \Pr(\text{A} \mid \text{B}) = \Pr(\text{A and B}) \ / \ \Pr(\text{B})$$
$$= \Pr(\text{A}) \times \Pr(\text{B} \mid \text{A}) \ / \ \Pr(\text{B})$$

Bayes rule

Pr(A | B) =

Pr(A) × Pr(B | A) / Pr(B) =

Pr(A) × Pr(B | A) / { Pr(A) × Pr(B | A) + Pr(not A) × Pr(B | not A) }

Let A denote disease, and B a positive test result!

- ⟶ Pr(A | B) is the probability of disease given a positive test result.
- ⟶ Pr(A) is the prevalence of the disease.
- ⟶ Pr(not A) is 1 minus the prevalence of the disease.
- ⟶ Pr(B | A) is the sensitivity of the test.
- ⟶ Pr(not B | not A) is the specificity of the test.
- ⟶ Pr(B | not A) is 1 minus the specificity of the test.