University of Washington

STATISTICS

# Elements of Statistical Methods
# Introduction and Preliminaries (Ch 1-2)

Fritz Scholz

Spring Quarter 2010

March 31, 2010

# Experiments

Read Chapter 1 as entertaining introduction. You may skip Section 1.1.3.

Tossing a coin is often used to decide issues ”fairly” with 50/50 chance.

However, there are people who are good at manipulating such tosses.

`http://www-stat.stanford.edu/~susan/papers/headswithJ.pdf`, p. 2.

What about spinning a penny on a smooth flat surface?

Vigorously flick the penny. The starting position could have head or tail facing us.

Does it matter? Do we have 50/50 chance? If not, what is the chance of heads?

These are questions that can be resolved by statistical methods.

# Shaved Dice

Symmetric and unloaded dice have equal chance $\frac{1}{6}$ to show up any of the faces.

Statistician Persi Diaconis has asked:

What are the chances, when one of the faces is shaved off by some amount?

For any such (parallel) shave we can argue by symmetry considerations:

$\frac{1}{6} - \frac{\varepsilon}{4}$ for the 4 faces perpendicular to shaved face and $\frac{1}{6} + \frac{\varepsilon}{2}$ for the other 2 faces.

For a fat shave, leaving only a thin square slice or "coin," the corresponding chances would be $\frac{\delta}{4}$ for the 4 thin edges and $\frac{1}{2} - \frac{\delta}{2}$ for the 2 flats.

How does $\varepsilon$ ($\delta$) depend on the shaving fraction? The answer seems to be unknown.

Much would depend on how the die is "rolled". Dynamic models are very complex.

# Some Comments

While such games of rolling dice and flipping or spinning coins seem frivolous they often can be viewed as useful abstractions for other real world problems:

drug or procedure effectiveness,

differences between different types of drugs or procedures (better than a placebo?)

issues of profiling at airport security checks, and on and on.

While one could relate similarities between these different situations to each other, the varying details tend to distract.

It is easier to relate each to such common models as coin tosses or rolls of dice. Not much explanation needs to take place for the latter.

We can focus on the essentials without application specific distractions.

# Measuring the Speed of Light

The speed of light has had an interesting history (see text for more details).

Estimates have ranged or progressed from

no speed (Aristotle), infinite speed (Bacon, Kepler, Descartes), finite speed (Galileo),

214,000 km/sec (Rømer 1676), 301,000 km/sec (Bradley 1729),

300,267.64 km/sec (Delambre 1809), all from astronomical measurements

Fizeau (1849), Foucault (1851) and Michelson (1879) used terrestrial experiments,

using cog wheels and rotating mirrors. Michelson got 299,944$\pm$ 51 km/sec.

It is now fixed (as of 1974) at $c = 299,792.458$ km/sec, based on a definition

of a second via Caesium-133 oscillations.

The meter was redefined in the International System of Units (SI)

as the distance travelled by light in vacuum in 1/299,792,458 of a second.

# Measurement Errors

While the first measurements were amazingly well in the ball park, they also exhibited measurement error or variation.

For the Michelson experiment the text discussess 13 different sources of measurement error/variation

My highschool physics teacher used a similar technique of rotating mirrors to measure the speed of light. He took great care to avoid vibrations in the building (e.g., setting up PE equipment).

At NIST on a site visit we were shown the room where they measure weight standards. It was impressive to see the temperature change in the enclosed scale apparatus because a group of humans had entered the room.

# Randomization

It seems that random variation (unpredictable outcomes, measurement variation) is something that we have to live with. Understanding it is essential.

But, to understand randomness we need to study probability theory.

Randomness is the bread and butter of statistics.

Myles Hollander: Statistics means never having to say you're certain.

Deliberate randomness or randomization can be used to our advantage, as the next examples show.

# The Lanarkshire Milk Experiment (1930)

School children were either given one of 2 milk supplements or none at all.

Of interest was the weight gain over a four month period (Feb.-June).

The group without milk supplement showed significantly higher weight gains.

In retrospect, it appears that various well meaning forces may have been at work.

There was no randomization of treatment and control assignments to the students.

Possibly beneficial effects of milk were confounded with hidden assignment factors.

Did teachers favor undernourished kids with milk assignment? The clothing factor?

Such confounding could have been avoided by randomized assignments. Why?

# Pre-Election Polls of 1948

Polls before the 1948 election put Thomas Dewey ahead of Harry Truman

Crossley $50\% \searrow 45\%$, Gallup $50\% \searrow 44\%$, and Roper $53\% \searrow 38\%$.

Truman won with slightly less than $50\%$ versus slighty more than $45\%$ for Dewey.

What happened? The culprit was quota sampling.

This attempts to construct samples according to known population characteristics,

such as residential location, sex, age, race, i.e., align samples closely to population.

Even though we allocate appropriate numbers for the various population slices,

it is left open how to select within those slices.

What if some population characteristics (education level) are not considered?

Judicious choices within slices created an imbalance w.r.t. education levels.

Randomization could have avoided this pitfall.

# Two Exams

Suppose I want to give a midterm and I construct two versions (A and B), to be given alternately to students sitting next to each other.

How do I adjust for the possible difficulty difference between the two exams?

Exam A has an average score 20 points higher than the average for exam B. What does it tell me? It depends.

If strong and weak students pair up sitting next to each other, but each time the stronger student winds up with exam A, then the 20 point difference could be due to the stronger student alone, or exam A is actually quite tough and the weaker student would have scored on average 40 points lower on exam A, or exam A was easy and the weaker student would have scored only 5 points lower on exam A.

As it is, the exam difficulty and student capability are confounded.

# Randomization Solution

We could still alternate exams A and B, but randomly assign student seating.

A random half of the students are assigned to the exam A seats,

the rest get the exam B seats. This is called simple random sampling.

Strong and weak students will be $\approx$ equally represented among exams A and B.

Any perceived difference in the average score could then be credited to the exam alone. An appropriate adjustment (shift or $\%$ change) can be made to equalize.

To represent freshmen and nonfreshmen equally among both exams, randomly split each group into halves, giving exam A to one half and exam B to the other in each group. This is called stratified random sampling.

You may need to make separate exam score adjustments for both groups.

# The Importance and Role of Probability

For the coin spin one might argue that it is a deterministic process, if all influencing factors had been known or controlled sufficiently. Same with weather.

Unfortunately, the controlling or knowing is not an easy matter.

Chaos theory addresses such phenomena with complex dynamical models. However, tiny changes in starting conditions can have opposite effects (H or T).

A more parsimonious and practical approach is to understand and describe such processes through probability models.

Similarly for measurement processes, even for an absolute like the speed of light. Rather than getting lost in all the what ifs of factors influencing measurement error, describe such measurement errors through probability models.

# Probability and Statistics

A probability model allows us to quantify in terms of chances or probabilities how likely it is to observe some experimental outcome (measurement or randomization).

Probability models may be known in general terms but not in specifics,
e.g., in the coin spinning example: $p = P(\text{tail}) = ?$

By observing several/many spins we can make some inference about $p$:
Estimate $p$, test hypotheses about it $H : p = \frac{1}{2}$?, or get a confidence interval for $p$.

Different people spinning $n = 25$ times will come up with different results.
What is a reasonable estimate for $p$? At what point should I doubt $H$?
A confidence interval is a form of estimation with built in margins of error.

Probability Theory uses logical deduction to obtain chances of specific events.

In Statistics you use inductive reasoning (no air tight proof) to induce or strongly suggest a conclusion about the specific probability model behind the data.

# Mathematical Preliminaries: Sets and Notation

Much of probability theory is expressed in the language/concepts of set theory.

Most of this should not be new, take it as a review, or getting on the same page.

We start with a universe $S$ which consists of all objects that we want to consider.

These objects could be physical objects, distinct events, states of mind, etc.

A set $A$ is a collection of objects in $S$. Any object in $S$ either belongs to $A$ or not.

All objects/elements in $S$ that do not belong to $A$ form the complementary set $A^c$.

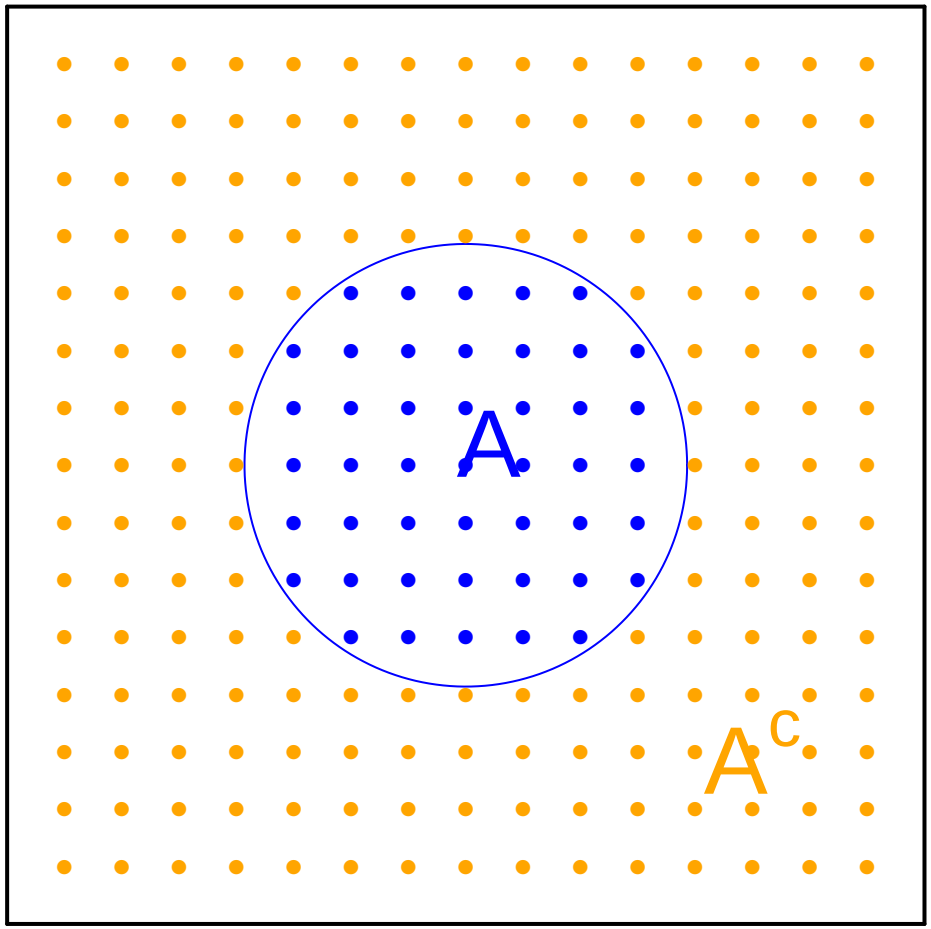
$S^c = \emptyset$ the empty set.

We use upper case letters $A, B, C$ from beginning of alphabet to denote sets.

Lower case letters $a, b, x, y$ denote specific elements in such sets.

We write $x \in A$ to denote that the element $x$ belongs to $A$.

# Set Complement




The universe $S$ consists of all solid dots (blue or orange).

# Some Special Number Universes

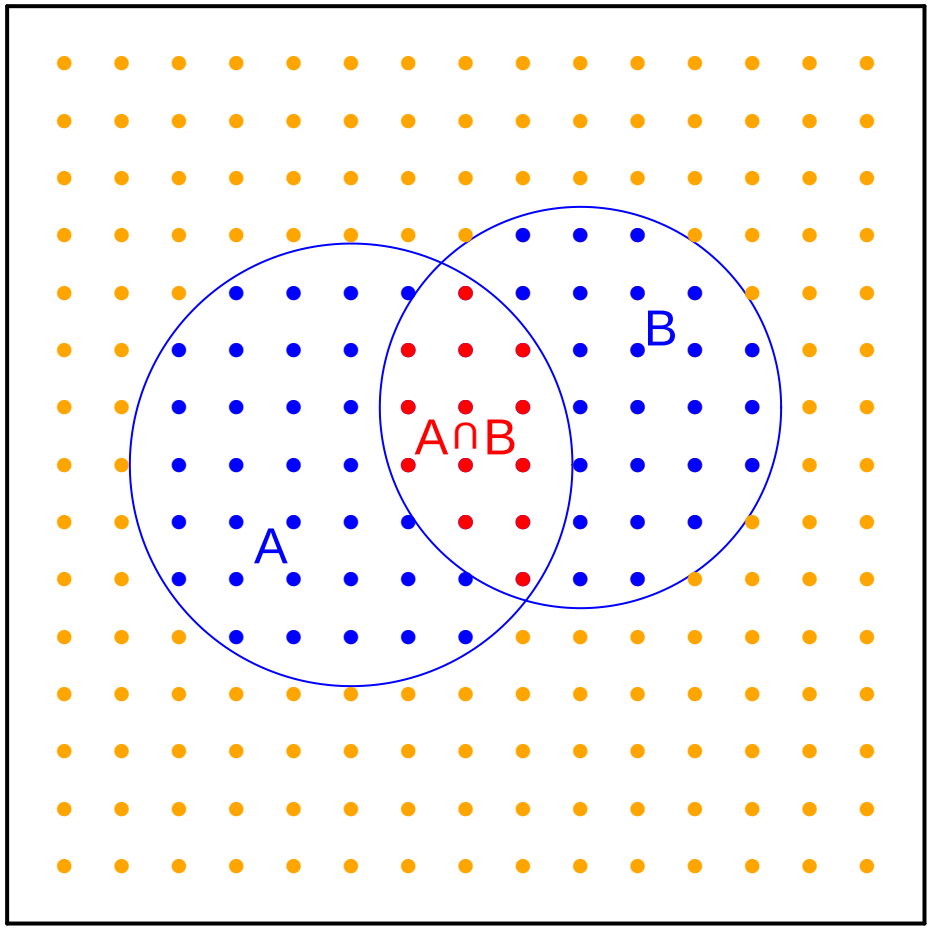
- The set of natural numbers $N = \{1, 2, 3, \ldots\}$
- The set of integers $Z = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
- The set of real numbers $R = (-\infty, \infty)$

Specific sets can be specified as follows

$$A = \{x \in Z : x^2 < 5\} = \{-2, -1, 0, 1, 2\}$$

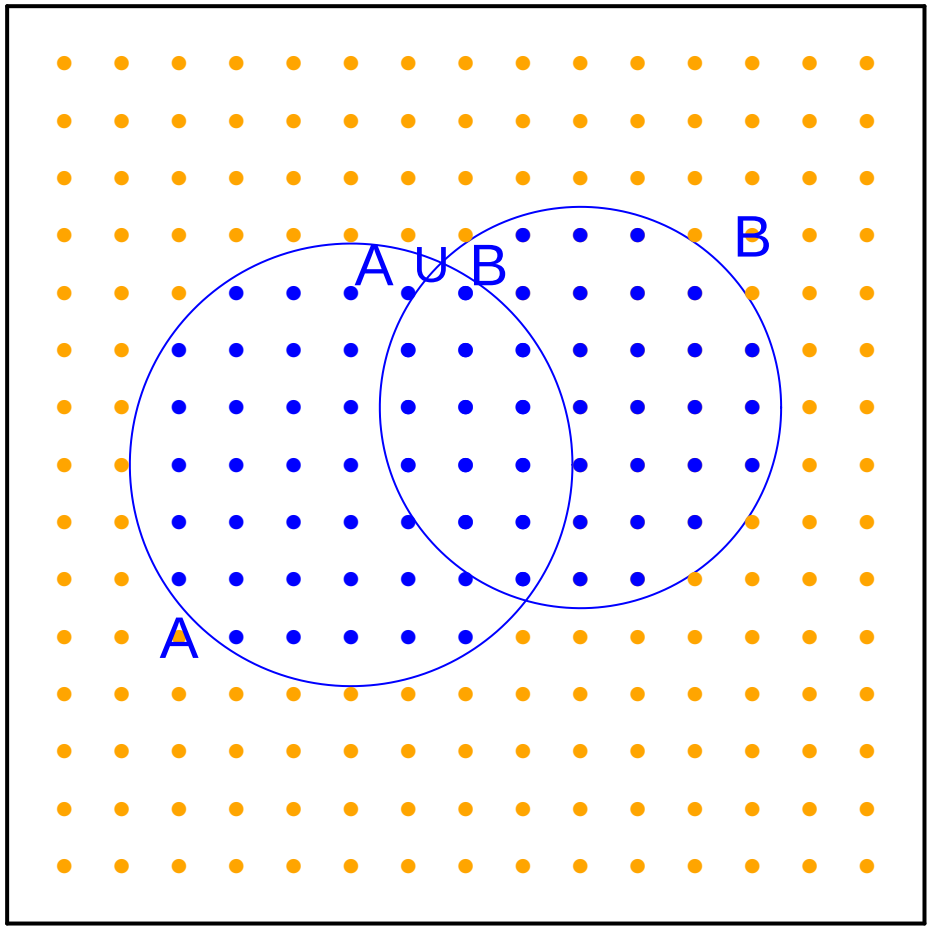
i.e., the set of all integers $x$ with $x^2 < 5$.

# Intersection of $A$ and $B = A \cap B$




$$A \cap B = \{x \in S : x \in A \text{ and } x \in B\} = B \cap A$$

# Union of $A$ and $B = A \cup B$




$$A \cup B = \{x \in S : x \in A \text{ or } x \in B\} = B \cup A$$

# Subsets and Mutually Exclusive Sets

If every element in $A$ is also in $B$ we write $A \subset B$, i.e., $A$ is a subset of $B$.

We also say $A$ implies $B$, i.e., if $x \in A$ then $x \in B$.

For our number universes we have

$$N \subset Z \subset R$$

For general sets $A$ and $B$ we have (see the previous Venn diagrams)

$$\emptyset \subset A \cap B \subset A \subset A \cup B \subset S$$

$$\emptyset \subset A \cap B \subset B \subset A \cup B \subset S$$

If $A \cap B = \emptyset$ then $A$ and $B$ don't share any common element, they are mutually exclusive.

# Union and Intersection of Many Sets

For any collection of subsets of $S$, say $A_k,\ k \in J$, where $J$ is some index set, we can form their union

$$\bigcup_{k \in J} A_k = \{x \in S : x \in A_k \text{ for some } k \in J\}$$

and their intersection

$$\bigcap_{k \in J} A_k = \{x \in S : x \in A_k \text{ for all } k \in J\}$$

A collection of subsets is pairwise disjoint if and only if

$$A_k \cap A_\ell = \emptyset \quad \text{for any } k, \ell \in J \text{ with } k \neq \ell$$

# Distributive Laws

$$B \cap \left( \bigcup_{k \in J} A_k \right) = \bigcup_{k \in J} (B \cap A_k)$$

Note the analogy with distributive law for addition and multiplication of numbers

$$b \times (a_1 + a_2 + \ldots + a_n) = b \times a_1 + b \times a_2 + \ldots + b \times a_n$$

and

$$B \cup \left( \bigcap_{k \in J} A_k \right) = \bigcap_{k \in J} (B \cup A_k)$$

No such number analogy here.

Prove either identity by showing: $x \in$ left side $\Longleftrightarrow x \in$ right side.

# DeMorgan's Laws

$$\left(\bigcup_{k\in J} A_k\right)^c = \bigcap_{k\in J} A_k^c \qquad \text{and} \qquad \left(\bigcap_{k\in J} A_k\right)^c = \bigcup_{k\in J} A_k^c$$

In words:

an element is not in any of the $A_k$ if and only if it is outside of each $A_k$

and

an element is not in every $A_k$ if and only if it is outside at least one of the $A_k$

In words it sounds almost tautological.

# Cartesian Product of Sets

The Cartesian product $A \times B$ of two sets $A$ and $B$ consists of all ordered pairs $(a,b)$, where $a$ comes from $A$ and $b$ comes from $B$, i.e.,

$$A \times B = \{(a,b) : a \in A, b \in B\}$$

Strictly speaking, if $A$ and $B$ are subsets of two respective universes $S_1$ and $S_2$, then we should view $A \times B$ as a subset of its Cartesian product universe

$$S = S_1 \times S_2 = \{(s_1,s_2) : s_1 \in S_1, s_2 \in S_2\}$$

Example

$R^2 = R \times R = \{(x,y) : x \in R, y \in R\}$ the Cartesian coordinate plane

and in extension

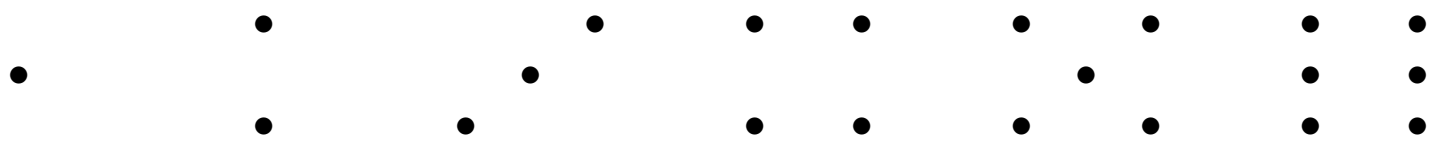
$R^3 = R \times R \times R = \{(x,y,z) : x \in R, y \in R, z \in R\}$ the Cartesian 3-dimensional space

# Counting the Number of Elements in a Set

The number of elements in two sets is the same if and only if (iff) we can pair up each element in one set with uniquely one element from the other, and vice versa. Such a pairing is also called a 1-1 correspondence.

A set $A$ is finite iff for some natural number $n$ it is in 1-1 correspondence with the set $\{1,2,\ldots,n\}$. Its size or cardinality is denoted by $\#(A) = n$.

The sides of a cube can be brought in 1-1 correspondence with the numbers $A = \{1,2,\ldots,6\}$ as it is usually done by using pips, spot, or dots, with $\#(A) = 6$



# Multiplication Principle in Specific Example

While counting by enumeration is the basic step, more efficient counting makes use of the multiplication principle.

Example: A coach has 4 rowers. How many ways can she put two rowers into the two seats of a double when seat order matters?

We have 4 choices for the stroke seat and for each such choice we have 3 choices for the bow seat, giving us $3+3+3+3=4\times 3=12$ boat compositions.

How many ways are there to fill two boats, when seat order and boat matter?

How many crews of two each when seat order matters, but boats do not?

# Multiplication Principle in General

In a situation where we have to make two choices, with the first choice offering us $n_1$ possibilities, and the second one offering us $n_2$ possibilities, then we can make altogether $n_1 n_2$ paired choices.

Note that in the above only the numbers $n_1$ and $n_2$ of choices enter.

The fact that the first choice may alter the possibilities for the second choice does not matter, as long as the number of choices stays fixed.

Suppose we have 4 boxes containing respectively 1, 1, 5, 5 distinct items.
We choose one of the boxes and then one of the items in that chosen box.
How many possible item choices do we have? Do we multiply, why or why not?

# Permutations/Combinations: Specific Case

We have 8 rowers, 4 of which are supposed to fill the seats of a quad.

If the seat order matters, how many ways can this be done?

8 ways to fill the stroke seat (4-seat), 7 ways to fill 3-seat

(i.e., $8 \times 7$ to fill 4- and 3-seat),

then 6 ways to fill 2-seat and 5 ways to fill bow seat or 1-seat,

giving us altogether $8 \times 7 \times 6 \times 5 = 1680$ ways (permutations) to fill the quad.

We applied the multiplication principle how many times?

If the seat order does not matter, there are $4 \times 3 \times 2 \times 1 = 4! = 24$ ways to scramble 4 rowers among the 4 seats (again multiplication principle).

Thus there are $1680/24 = 70$ ways (combinations) to compose teams of 4 taken from a group of 8 without regard to seat order.

# Permutations/Combinations: General Case

The number of permutations (ordered choices) of $r$ objects from $n$ objects is

$$P(n,r) = n \times (n-1) \times \cdots \times (n-r+1)$$

The number of combinations (unordered choices or grabs) of $r$ objects

taken from $n$ objects is

$$C(n,r) = \frac{P(n,r)}{P(r,r)} = \frac{n \times (n-1) \times \cdots \times (n-r+1)}{1 \times 2 \times \cdots \times r}$$

Note that $P(r,r) = 1 \times 2 \times \cdots \times r = r!$ and

$$\begin{aligned} P(n,r) \times P(n-r,n-r) &= n \times (n-1) \times \cdots \times (n-r+1) \times (n-r) \times \cdots \times 2 \times 1 \\ &= P(n,n) \end{aligned}$$

so that

$$C(n,r) = \frac{P(n,r)}{P(r,r)} = \frac{P(n,n)}{P(n-r,n-r) \times P(r,r)} = \frac{n!}{(n-r)!r!} = \binom{n}{r}$$

In all this we use the convention $0! = 1$. Note how it fits into all the above.

# Combinations: An Application

Suppose we have 40 subjects and a new drug, promising hours of pain relief.

We will test it by giving the drug to randomly chosen 20 of those 40,

while the remaining 20 get an identical looking placebo.

How many possible splits into two groups are there?

$$\binom{40}{20} = \texttt{choose}(40,20) = 137,846,528,820$$

`choose` is a function in R. Note that $5! = \texttt{factorial}(5) = 120$ in R.

The 40 subjects consist of 30 men and 10 women. To avoid imbalance in the drug assignments we choose 15 men and 5 women to receive the drug.

How many ways? By the multiplication principle we have

$$\binom{30}{15} \times \binom{10}{5} = \texttt{choose}(30,15) * \texttt{choose}(10,5) = 39,089,615,040$$

$$39,089,615,040/137,846,528,820 = 0.2835734$$

or $\approx 28\%$ of our original splits would have resulted in balanced splits.

# Infinite Sets

There are many notions of infinity and many causes to be perplexed.

We call a set denumerable or countably infinite when it can be put in 1-1 correspondence with the set $N$ of natural numbers.

A set is countable when it either finite or denumerable.

A set is uncountable when it not countable.

Thus we already distinguish two forms of infinity.

# Examples of Denumerable Sets

The even natural numbers are denumerable, but comprise only half of $N$.

$$2^{\nearrow 1}, 4^{\nearrow 2}, 6^{\nearrow 3}, 8^{\nearrow 4}, \ldots, 2004^{\nearrow 1002}, \ldots$$

The integers, more than two times the natural ones, are denumerable

$$0^{\nearrow 1}, -1^{\nearrow 2}, 1^{\nearrow 3}, -2^{\nearrow 4}, 2^{\nearrow 5} \ldots, -2004^{\nearrow 4008}, 2004^{\nearrow 4009}, \ldots$$

# Examples of Denumerable Sets

Paired natural numbers are denumerable

$$
\begin{array}{ccccccc}
(1,1)^{\nearrow 1} & (1,2)^{\nearrow 2} & (1,3)^{\nearrow 6} & (1,4)^{\nearrow 7} & (1,5)^{\nearrow 15} & \cdots \\
(2,1)^{\nearrow 3} & (2,2)^{\nearrow 5} & (2,3)^{\nearrow 8} & (2,4)^{\nearrow 14} & (2,5)^{\nearrow 17} & \cdots \\
(3,1)^{\nearrow 4} & (3,2)^{\nearrow 9} & (3,3)^{\nearrow 13} & (3,4)^{\nearrow 18} & (3,5)^{\nearrow 26} & \cdots \\
(4,1)^{\nearrow 10} & (4,2)^{\nearrow 12} & (4,3)^{\nearrow 19} & (4,4)^{\nearrow 25} & (4,5)^{\nearrow 32} & \cdots \\
(5,1)^{\nearrow 11} & (5,2)^{\nearrow 20} & (5,3)^{\nearrow 24} & (5,4)^{\nearrow 33} & (5,5)^{\nearrow 41} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
$$

Thus the positive rational numbers are denumerable.

# Examples of Uncountable Sets

$$
\begin{aligned}
(a,b) &= \{x \in R : a < x < b\} \\
[a,b) &= \{x \in R : a \leq x < b\} \\
(a,b] &= \{x \in R : a < x \leq b\} \\
[a,b] &= \{x \in R : a \leq x \leq b\}
\end{aligned}
$$

Showing that these sets are uncountably infinite is not difficult.

It is done by contradiction, assuming that there is a 1-1 relationship with $N$ and then constructing a number that was missed.

Note double duty of $(a,b)$, as an ordered pair of numbers and as open interval!

# Functions

Let $A \subset S_1$ and $B \subset S_2$ be subsets of possibly different universes $S_1$ and $S_2$.

A function $f$ is a rule that assigns a unique element $y$ in $B$ for each element $x$ in $A$.

We write $f : A \longrightarrow B$ or also $y = f(x)$.

$A$ is called the domain of $f$ and

$C = \{f(x) : x \in A\} = f(A) \subset B$ is called the range of $f$.

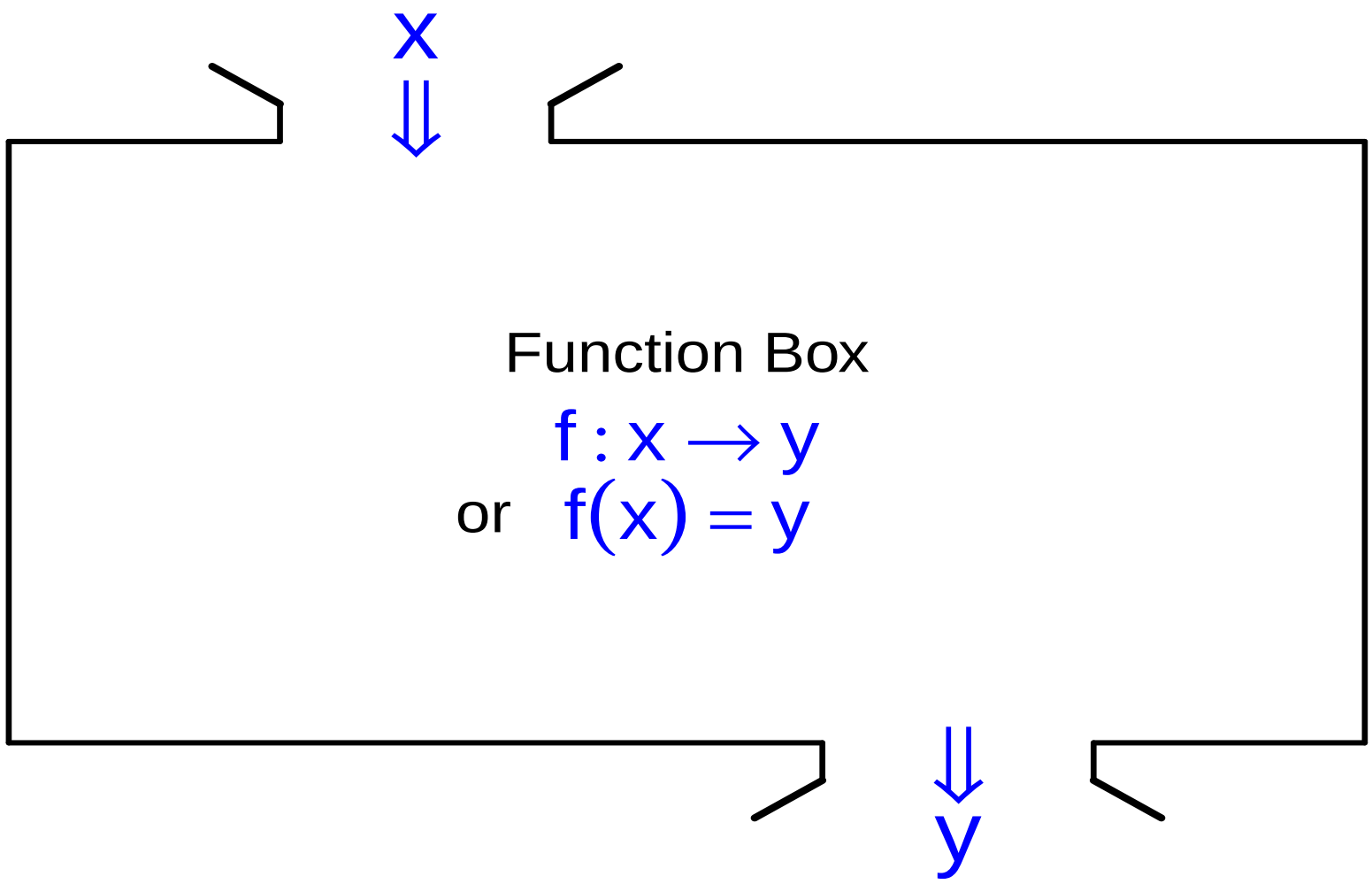
A simple example: the quadratic function $f : x \longrightarrow x^2$ or $y = f(x) = x^2$ with $x \in R$.

We also write: $f : R \longrightarrow R$, depending on what we want to emphasize.

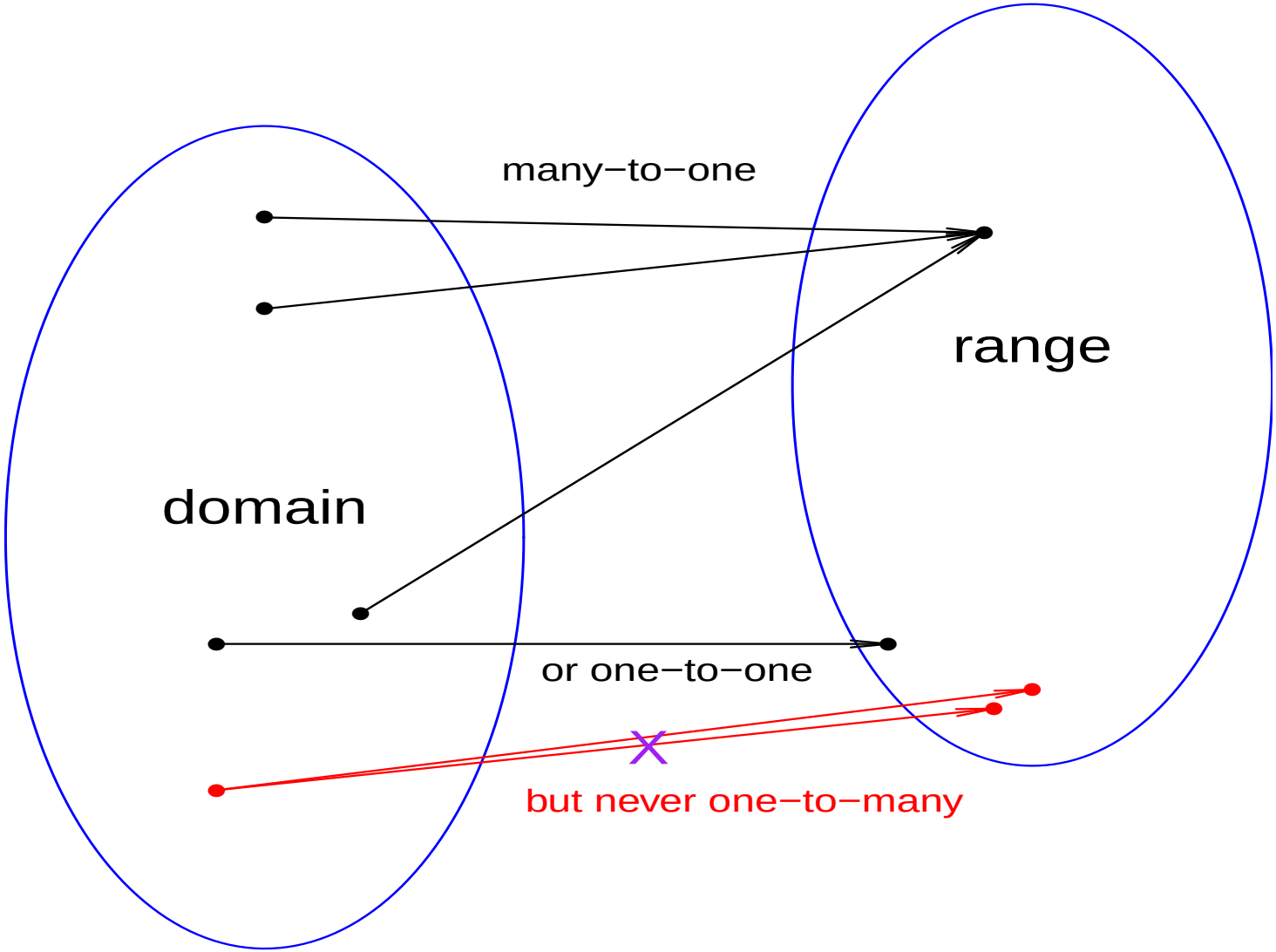
Note that $-5$ and $5$ get assigned the same value $25 = 5^2 = (-5)^2$.

While $f : R \longrightarrow R$ here, the range is $[0, \infty)$.

# Function Box or Black Box



# Function Diagram



# Function Inverse

For any $b \in B$ we define

$$f^{-1}(b) = \{a \in A : f(a) = b\}$$

In the case of the square function we have

$f^{-1}(25) = \{-5, 5\}$ and more generally $f^{-1}(b) = \{-\sqrt{b}, \sqrt{b}\}$ for any $b \geq 0$

For $b < 0$ there is no $a \in R$ for which $a^2 = b$. In that case $f^{-1}(b) = \emptyset$.

The above notion of $f^{-1}$ can be extended to subsets $D \subset B$, i.e.,

$$f^{-1}(D) = \{a \in A : f(a) \in D\} = \text{set of } a \in A \text{ that get mapped into } D.$$

For the square function: $f^{-1}([-3, 4]) = [-2, 2]$ and $f^{-1}([4, 9]) = [-3, -2] \cup [2, 3]$.

# Limits of Numbers

We encountered denumerable sets of real numbers.

By definition they were in 1-1 correspondence with $N = \{1, 2, 3, \ldots\}$,

e.g., $\{a_1, a_2, a_3, a_4, \ldots\} = \{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$

We can ask how this sequence behaves as we progress through it.

In the given example the numbers become arbitrarily small,

they are said to converge to zero or have limit zero

(without zero ever being one of its values).

The series $\sum_{i=1}^{\infty} a_i$ is defined as the limit of the sequence consisting of the finite partial sums $\sum_{i=1}^{n} a_i$:

$$\sum_{i=1}^{1} a_i = a_1, \ \sum_{i=1}^{2} a_i = a_1 + a_2, \ \sum_{i=1}^{3} a_i = a_1 + a_2 + a_3, \ \ldots, \sum_{i=1}^{n} a_i = a_1 + \ldots + a_n, \ \ldots$$

# Limits of Sets

For a denumerable sequence of sets, indexed $A_1, A_2, A_3, \ldots,$

we may ask how it behaves. Does it approach a limit set?

The notion of approach here is not so obvious.

We constrain ourselves to monotone sequences of sets, e.g.,

$$\bigcup_{i=1}^{1} A_i = A_1 \subset \bigcup_{i=1}^{2} A_i = A_1 \cup A_2 \subset \bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3 \ldots$$

$$\ldots \subset \bigcup_{i=1}^{n} A_i = A_1 \cup \ldots \cup A_n \longrightarrow \bigcup_{i=1}^{\infty} A_i$$

where the limit union is simply the set of elements that are in at least one of the $A_i$.

Similarly define $\cap_{i=1}^{\infty} A_i$ to consist of the elements that are in all $A_i$.