(*) A ***random variable*** is a variable X whose value depends on the outcome of a chance process.

Examples:

- H = number of *heads* in 10 coin flips.
- W = average weight (in lbs) in a simple random sample of 25 feral cats in San Francisco.
- D = number of deaths in a given day in New York City.

(*) The ***range*** of a random variable X (Range(X)) is the set of *possible* values that it may assume.

Examples:

- Range(H) = $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.
- Range(W) = $(8, 12)$ (i.e., 8lbs to 12lbs)
- Range(D) = $\{0, 1, 2, 3, 4, \ldots\}$.

(*) A random variable is ***discrete*** if its range is a *discrete set* of numbers. A random variable is ***continuous*** if its range is a *continuous set* of numbers.

Examples:

- H = number of *heads* in 10 coin flips... Discrete (and finite).
- W = average weight ... Continuous.
- D = number of deaths ... Discrete (and infinite)

(*) The ***probability distribution*** of a *discrete* random variable X gives the probability of observing each value that X can assume. If Range$(X) = \{x_1, x_2, \ldots, x_j, \ldots\}$, then the probability distribution of X gives the probabilities

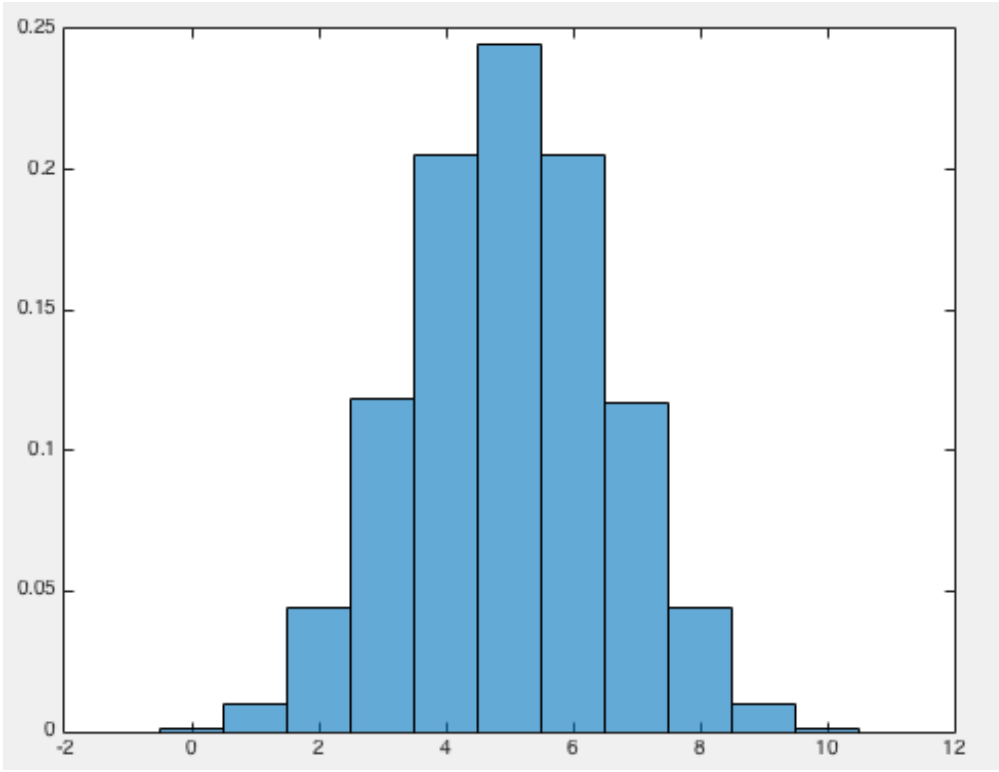
$$P(x_j) = P(X = x_j).$$

Example. H = number of *heads* in 10 coin flips. We can display the probability distribution of H in a table

x	$P(H = x)$
0	0.001
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.246
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.001

(*) Probability distributions can also be displayed as *Histograms* (called *probability histograms*).

Example. H = number of *heads* in 10 coin flips.



(*) Heights of bars are equal to probabilities. Bar widths are all 1...

(*) ...so ***areas of bars*** are also equal to probabilities.

(*) Discrete probability distributions must satisfy two important conditions.

If $\{x_1, x_2, \ldots, x_n, \ldots\}$ is the range of the distribution, then

(i) $0 \leq P(x_j) \leq 1$ for each x_j in the range

because $P(x_j)$ is a *probability*.

(ii) $\sum P(x_j) = 1$

because the events $(X = x_1)$, $(X = x_2), \ldots, (X = x_n), \ldots$ are *mutually exclusive* and comprise *all the possible outcomes*.

(*) A *probability distribution* can be thought of as a *frequency distribution*, with the probability of each value being the relative frequency of that value. With this analogy in mind, we define the ***mean***, ***variance*** and ***standard deviation*** of a (discrete) probability distribution as follows.

If the range of X is $\{x_1, \ldots, x_n, \ldots\}$ with probabilities $P(x_j)$, then...

(*) The *mean* of the distribution is

$$\mu = \sum x_j P(x_j)$$

(*) The *variance* of the distribution is

$$\sigma^2 = \sum (x_j - \mu)^2 P(x_j) = \overbrace{\left(\sum x_j^2 P(x_j)\right) - \mu^2}^{\text{shortcut}}$$

(*) The *standard deviation* of the distribution is

$$\sigma = \sqrt{\sigma^2}$$

Explanation. Suppose that the process that produce the values of X is repeated a large number, N, of times.

(*) Each value, x_j, will be observed about $P(x_j) \cdot N$ times.

(*) The sum of all the observed values will be about

$$x_1 \cdot (P(x_1) \cdot N) + x_2 \cdot (P(x_2) \cdot N) + \cdots + x_n \cdot (P(x_n) \cdot N) = \sum x_j P(x_j) \cdot N$$

(*) The mean of all the observed values will there be about

$$\frac{\sum x_j P(x_j) \cdot N}{N} = \sum x_j P(x_j) = \mu$$

(*) The mean of the distribution is the (theoretical) *long term mean of the observed values* from the distribution.

(*) The mean of the distribution is also the ***expected value*** of the corresponding random variable X, denoted by $E(X)$.

(*) The same reasoning explains the formulas for the variance and the standard deviation.

Example. To play the game *Red Marble*, you pay \$2.00 and draw two marbles at random *with replacement* from a box containing eight black marbles and two red marbles.

(*) If both marbles are black you win \$0.00

(*) If one marble is red and one is black, you win \$3.00

(*) If both marbles are red, you win \$20.00.

How much can you expect to win/lose in the long run? I.e., $E(X) = ?$

(*) X = profit per game = (winnings − \$1.00).

(*) Probability distribution:

x	$P(x)$
-2	0.64
1	0.32
18	0.04

$E(X) = (-2) \cdot (0.64) + 1 \cdot (0.32) + 18 \cdot (0.04) = -0.24$

The Binomial Distribution. Set up...

(*) A ***trial*** has two outcomes, often labeled *success* and *failure* (S and F).

Note: success is not necessarily good (or bad) and failure is not necessarily bad (or good).

(*) The trial is repeated a certain number (n) of times in such a way that $P(S) = p$ is the same for each trial. (The probability $P(F) = 1 - p = q$ is also the same in each trial as a consequence.)

(*) The random variable X that counts the number of successes in n trials is called a ***binomial random variable***, i.e., it follows the ***binomial distribution***. We write $X \sim B(n, p)$ for short.

Examples.

1. X = the number of *heads* in 10 tosses of a fair coin...

$$X \sim B\left(10, \tfrac{1}{2}\right).$$

2. X = the number of *red marbles* observed when 5 marbles are drawn ***with replacement*** from a jar containing 2 red and 8 black marbles...

$$X \sim B\left(5, \tfrac{1}{5}\right).$$

3. X = the number of sophomores in a simple random sample of 15 UCSC undergraduates...

$$X \sim B(15, p),$$

where p is the proportion of sophomores in the population of UCSC undergraduates.

Comment: Example 3 is technically not a binomial distribution because the sampling is done *without replacement* (the technically correct distribution is called a *hypergeometric* distribution). However when the sample size (15 in this case) is *very* small relative to the population size (e.g., $< 5\%$), the binomial distribution provides a very accurate approximation.

Calculating the binomial probabilities.

If $X \sim B(n, p)$, then

$$\text{Range}(X) = \{0, 1, 2, \ldots, n\}$$

and if $0 \leq x \leq n$, then

$$P(x) = P(X = x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}$$

- $x! = x \cdot (x-1) \cdots 3 \cdot 2 \cdot 1$ if $x \geq 1$ (and x is an integer).
- $0! = 1$.
- The number $\frac{n!}{x!(n-x)!} = \binom{n}{x} = {}_nC_x$ is always an integer and counts the number of different *subsets of size* x you can choose from a set with n objects.

  In particular $\frac{n!}{x!(n-x)!}$ is the number of different sequences n-letters long using only the letters S and F, with exactly x Ss (and $n - x$ Fs).

Examples.

If $X \sim B\left(10, \frac{1}{2}\right)$, then

$$P(X=0) = \frac{10!}{0!10!}(1/2)^0(1/2)^{10} = (1/2)^{10} = 0.0009765625 \approx 0.001.$$

and

$$P(X=6) = \frac{10!}{6!4!}(1/2)^6(1/2)^4 = 210 \cdot (1/2)^{10} = 0.205078125 \approx 0.2051.$$

(Note that in this case, $P(X=6) = P(X=4)$.)

If $Y \sim B\left(10, \frac{1}{5}\right)$, then

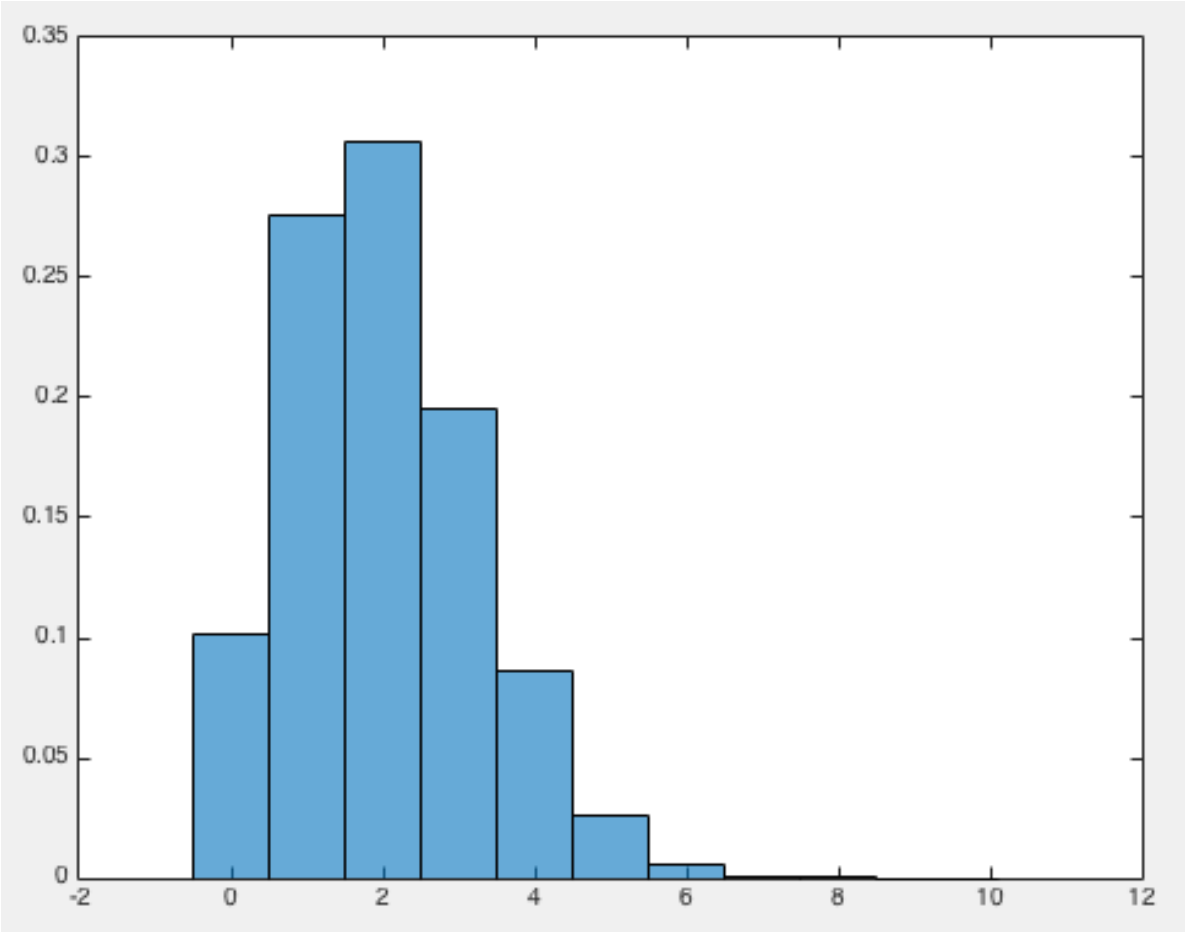
$$P(Y=0) = \frac{10!}{0!10!}(1/5)^0(4/5)^{10} = (4/5)^{10} \approx 0.1074,$$

$$P(Y=6) = \frac{10!}{6!4!}(1/5)^6(4/5)^4 = 210 \cdot \left(\frac{4^4}{5^{10}}\right) \approx 0.000044.$$

and

$$P(Y=4) = \frac{10!}{4!6!}(1/5)^4(4/5)^6 = 210 \cdot \left(\frac{4^6}{5^{10}}\right) \approx 0.045.$$

Probability histogram for $X \sim B\left(10, \frac{1}{5}\right)$



Where does the formula come from?

(*) The probability of a particular sequence

$$SSFFFSF \ldots SFF$$

with exactly x Ss and $(n-x)$ Fs is

$$p \cdot p \cdot q \cdot q \cdot q \cdot p \cdot q \cdots p \cdot q \cdot q = \overbrace{p \cdots p}^{x} \cdot \overbrace{q \cdots q}^{n-x} = p^x q^{n-x}$$

because the trials are ***independent***.

(*) There are exactly $\frac{n!}{x!(n-x)!}$ (mutually exclusive) sequences with exactly x Ss and $(n-x)$ Fs, so the probability of observing exactly x Ss in n trials is

$$P(x) = \overbrace{p^x q^{n-x} + p^x q^{n-x} + \cdots + p^x q^{n-x}}^{\frac{n!}{x!(n-x)!}} = \frac{n!}{x!(n-x)!} p^x q^{n-x}.$$

Example. Suppose that it is *believed* that 25% of the undergraduate students at UCSC spend 10 hours or more on homework per course per week. A simple random sample of 200 undergraduates are surveyed and 35 of them report that they spend at least 10 hours on homework per course per week. What conclusion might be drawn?

Denote by X the number of *undergraduates etc. etc. per week* observed in a simple random sample of 200 UCSC undergraduates. According to what is *believed*, $X \sim B(200, 0.25)$. So...

$$P(X = 35) = \frac{200!}{35! \cdot 165!}(0.25)^{35}(0.75)^{165} \approx 0.0029.$$

In fact...

$$P(X \leq 35) = \sum_{x=0}^{35} P(X = x) \approx 0.0073.$$

Conclusion: The assumption that 25% of UCSC undergraduates spend 10 or more hours per week per course on homework is likely to be wrong.

Mean, Variance and Standard deviation.

If $X \sim B(n,p)$, then

- $\mu = E(X) = np$.
- $\sigma^2 = npq$.
- $\sigma = \sqrt{npq}$.

  (Where $q = 1 - p$, as usual.)

Example. If $X \sim B\,(1000, 0.2)$, then $\mu = 1000 \cdot 0.2 = 200$ and $\sigma = \sqrt{1000 \cdot 0.2 \cdot 0.8} \approx 12.65$.

Interesting observation: if X is within two standard deviations of 200, then $|X - 200| \leq 25.3$, so in fact

$$175 \leq X \leq 225$$

and

$$P(175 \leq X \leq 225) = \sum_{x=175}^{225} \frac{1000!}{x!(1000-x)!}(0.2)^x(0.8)^{1000-x} \approx 0.9564.$$

The Poisson Distribution.

(*) The random variable X counts the number of occurrences of some event per unit time, per unit area, per unit volume, etc.

Assumptions: (i) The events occur independently of each other; (ii) The events occur with a constant *average rate* per unit (time, area, volume, etc.); (iii) The events occur *at random* — they are equally likely to occur at any point in the time interval, unit of area, unit of length, etc., as in any other (i.e., *no pattern*).

Examples.

1. The number of calls per hour to a call center.
2. The number of photons hitting a telescope per second.
3. The number of mutations on a strand of DNA on a strand of a given length.
4. The number of radioactive atoms that decay per unit time.

If X has the Poisson distribution with mean μ (the average number of occurrences per unit), then Range$(X) = \{0, 1, 2, 3, \ldots, n, \ldots\}$ and

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!},$$

(*) The mean is μ and the standard deviation is $\sigma = \sqrt{\mu}$.

Example. An internet search engine receives an average of 120 search requests per minute (24 hours a day). What is the probability that this search engine will receive 5 search requests in any given second? What about 6 or more search requests in a second?

(*) Assume Poisson distribution.

(*) 120 requests per minute $\implies$ $\mu = 120/60 \approx 2$ requests/second.

(*) $P(5 \text{ requests in a second}) = \dfrac{2^5 \cdot e^{-5}}{5!} \approx 0.036.$

(*) $P(6 \text{ or more requests in a second}) = \displaystyle\sum_{x=5}^{\infty} \frac{2^x \cdot e^{-2}}{x!} \approx 0.017$

Probability distributions for *continuous* random variables.

(*) If X is a continuous random variable, then $P(X = x_0) = 0$ for any fixed value x_0.

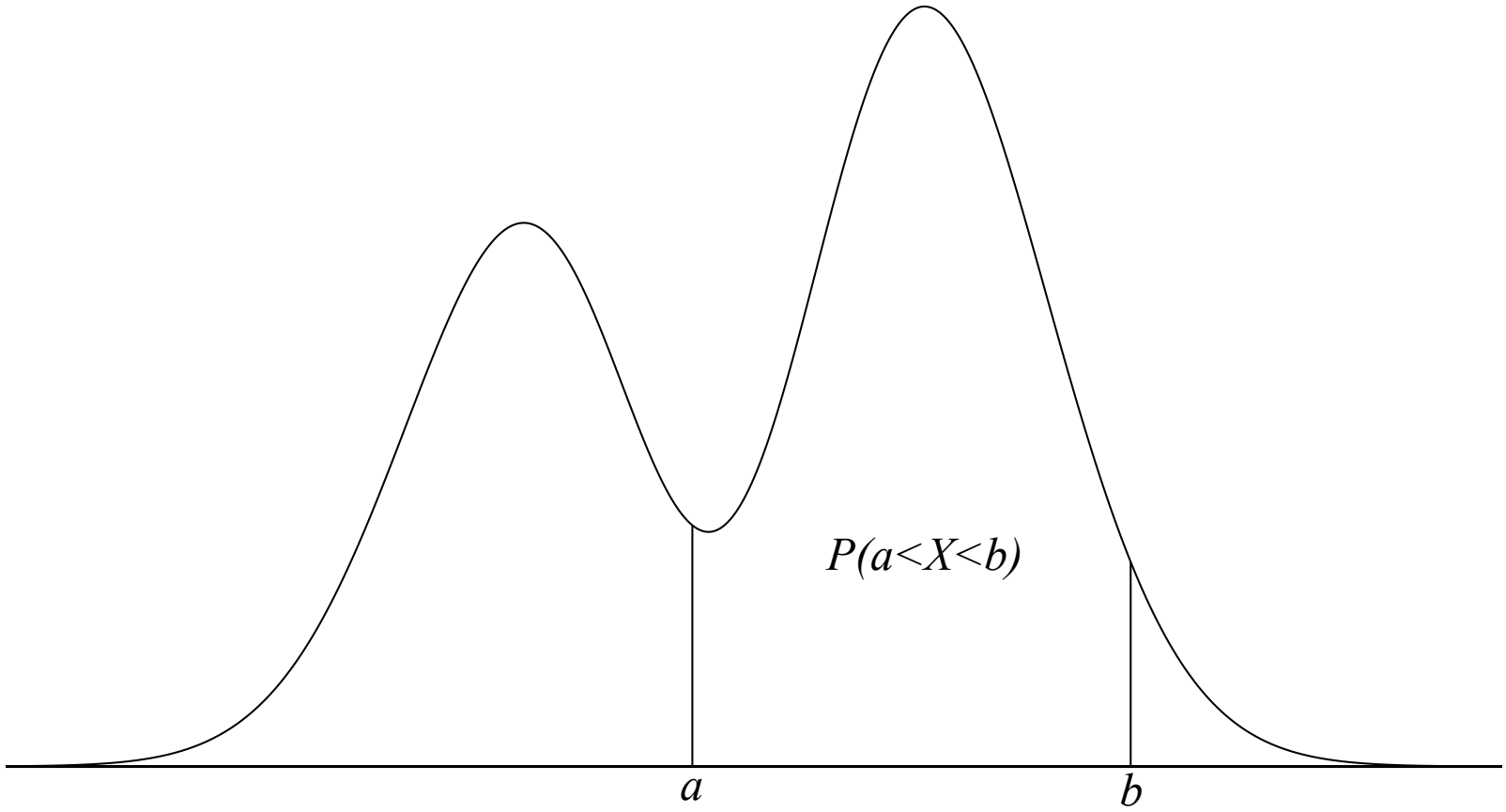
(*) Typical calculations: $P(a < X < b)$, $P(X < b)$ and $P(a < X)$.

(*) The probability distribution of a continuous random variable is characterized by a ***density curve***, $y = f(x)$, with the properties

(i) $f(x) \geq 0$ for all x.

(ii) The total area under the curve is equal to 1.

(*) The probabilities are equal to the areas under the curve between appropriate limits.


P(a<X<b)
a
b

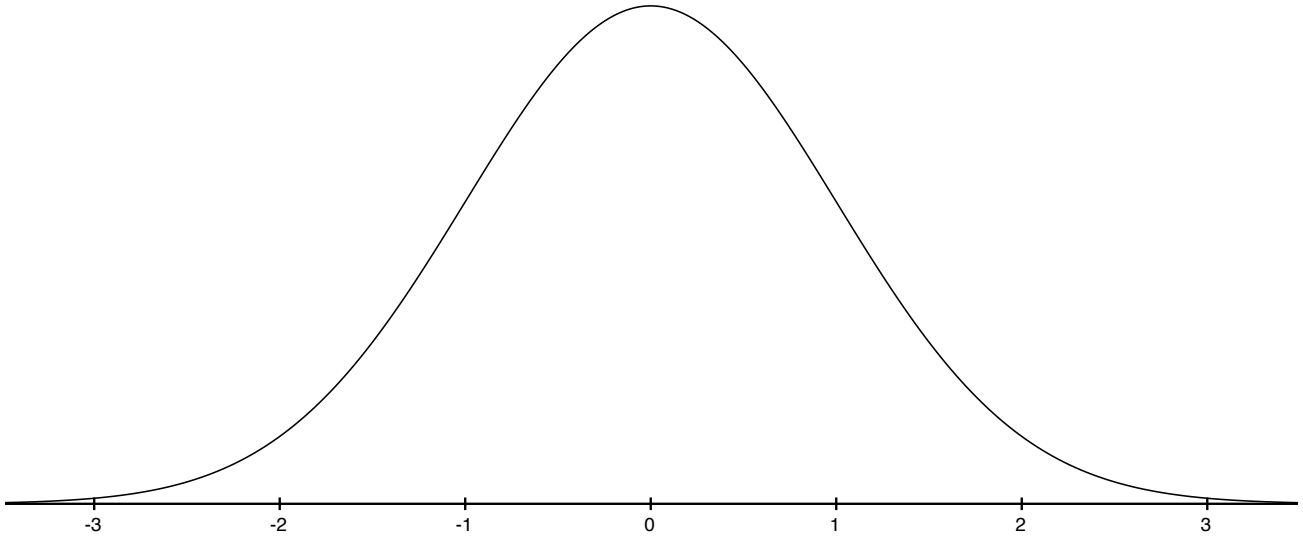
The *standard* normal distribution.

(*) $Z \sim N(0,1)$ has mean $\mu = 1$ and standard deviation $\sigma = 1$.

(*) The standard normal variable has probability density function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

whose graph is the familiar *bell-shaped curve*, centered at $z = 0$:



(*) To calculate probabilities, we use a *Normal Table.*