Probability Distributions

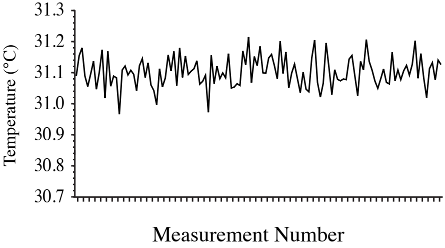
Chapter 2

P. J. Grandinetti

Chem. 4300

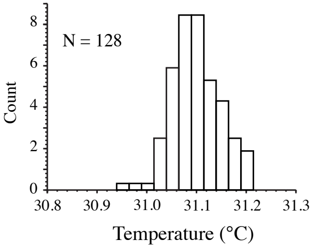
Aug. 25, 2017

Temperature of solution in “constant temperature” bath



What do you report for the solution temperature?

- the mean temperature (31.1 °C)
- histogram of all measured values



histogram,
aka *the sample distribution*

Sample and parent distribution

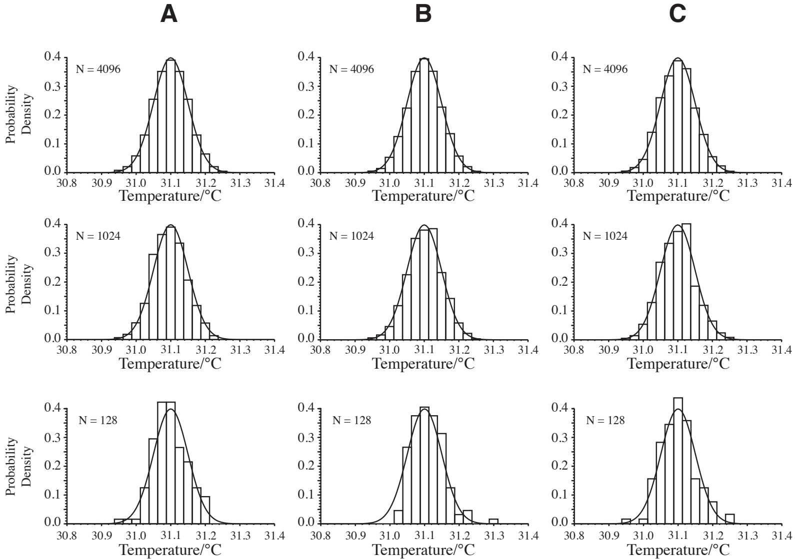
We assume that our histogram of measured values is governed by an underlying probability distribution called the *parent distribution*.

parent distribution

In the limit of an infinite number of measurements our histogram or *sample distribution* becomes the parent distribution.

Sample and parent distribution

Histograms (sample distributions) constructed from the same parent distribution



Probability density

Our main objective in making a measurement is to learn the underlying parent distribution, $p(x)$, that predicts the spread in the measured values.

The parent distribution, $p(x)$, is also called a probability density. A parent distribution is always normalized so the area under the distribution is unity,

$$\int_{\text{all } x} p(x)\, dx = 1.$$

Confidence Limits

The probability that a measured value lies between x_- and x_+ can be calculated from the parent distribution according to

$$P(x_-, x_+) = \int_{x_-}^{x_+} p(x)\, dx.$$

The integral limits x_- and x_+ are called the *confidence limits* associated with a given probability $P(x_-, x_+)$.

Moments of a distribution

When you report confidence limits you lose information concerning the shape of the parent distribution.

There are a few parameters that by convention are often used to describe the parent distribution in part, or sometimes completely.

- mean : the first moment about the origin
- variance : the second moment about the mean
- skewness : the third moment about the mean
- kurtosis : the fourth moment about the mean

The mean

The Mean describes the average value of the distribution. Given the parent distribution, $p(x)$, the mean is calculated according to

$$\mu = \int_{\text{all } x} x\, p(x)\, dx.$$

From a series of measurements the mean is given by

$$\mu = \lim_{N \to \infty} \frac{1}{N} \sum_i x_i,$$

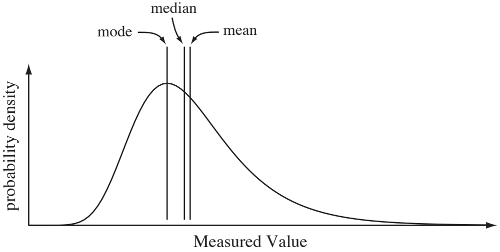
where N corresponds to the number of measurements x_i.

Practically, we cannot make an infinite measurements so the experimental mean, $\overline{x}$, is defined as

$$\overline{x} = \frac{1}{N} \sum_i x_i.$$

The mean, the median, and the mode.

When distribution is not symmetric about the mean



two other parameters used are:

- median: cuts the area of the parent distribution in half,

$$\int_{-\infty}^{x_{\text{median}}} p(x)\, dx = \frac{1}{2}.$$

- mode: most probable value,

$$\frac{dp(x_{\text{mode}})}{dx} = 0, \quad \text{and} \quad \frac{d^2 p(x_{\text{mode}})}{dx^2} < 0.$$

The Variance.

Variance characterizes the width of the distribution and is given by

$$\sigma^2 = \int_{\text{all } x} (x - \mu)^2 p(x) dx.$$

From a series of measurements the variance is obtained through:

$$\sigma^2 = \lim_{N \to \infty} \frac{1}{N} \sum_i^N (x_i - \mu)^2.$$

The experimental variance is defined as:

$$s^2 = \frac{1}{N-1} \sum_i^N (x_i - \overline{x})^2,$$

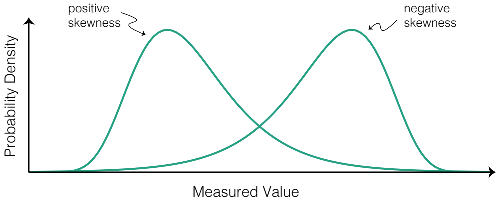
where s^2 is the variance of the experimental parent distribution. σ and s are the *standard deviation* of the parent distribution and experimental parent distribution, respectively.

The Skewness.

The Skewness characterizes the asymmetry of a distribution and is given by

$$\text{skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3.$$

Skewness is dimensionless. A distribution with *positive skewness* has an asymmetric tail extending out more towards $+x$, while a *negative skewness* extends out more toward $-x$.



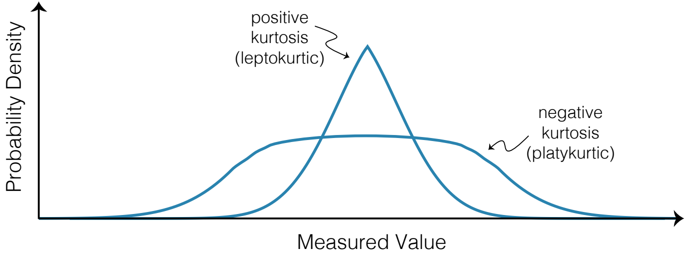
Symmetric distributions have zero skewness (*e.g.*, Gaussian).

The Kurtosis.

The Kurtosis measures the relative peakedness or flatness of a distribution relative to a normal (i.e., Gaussian) distribution. It is defined as:

$$\text{kurtosis} = \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^4 \right] - 3.$$

Subtracting 3 makes the kurtosis zero for a Gaussian distribution. A positive kurtosis is called *leptokurtic*, a negative kurtosis is called *platykurtic*, and in between is called *mesokurtic*.



Probability

$$\text{Probability} = \frac{\text{Number of outcomes that are successful (winning)}}{\text{Total number of outcomes (winning and losing)}}$$

The difficulty lies in counting. In order to count the number of outcomes we appeal to combinatorics.

Permutations

Definition

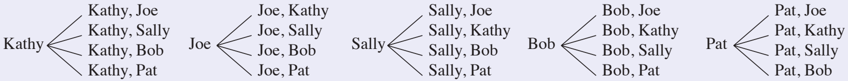
Permutation An arrangement of outcomes in which the order is important.

Example

Consider a club with 5 members, Joe, Kathy, Sally, Bob, and Pat. In how many ways can we elect a president and a secretary?

One solution

One solution is to make a tree, such as the one below:



Using such a tree diagram we can count that there are a total of 20 possible ways to elect a president and a secretary in a club with 5 members.

Permutations

Example

What if we wanted to elect a president, secretary, and treasurer?

Solution

In this case a tree would be a lot of work. Using the boxes approach we would have $5 \cdot 4 \cdot 3 = 60$ possibilities.

The number of ways r objects can be selected from n objects is

$$_nP_r = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1),$$

or more generally written as

$$\boxed{_nP_r = \frac{n!}{(n-r)!}.}$$

Combinations

Definition

Combination An arrangement of outcomes in which the order is not important. The total number of combinations of n objects taken r at a time is

$$_nC_r = \frac{_nP_r}{r!}, \qquad \text{or} \qquad \frac{n!}{r!(n-r)!}.$$

Example

Consider again our club with 5 members. In how many ways can we form a three member committee?

Solution

Here order is not important. That is, {Joe, Kathy, Sally} = {Kathy, Joe, Sally} = {Kathy, Sally, Joe}. All arrangements are equivalent.

$_5C_3 = \dfrac{5!}{3!2!} = 10$ possible 3 member committees starting with 5 members

Combinations

Definition

Combination An arrangement of outcomes in which the order is not important. The total number of combinations of n objects taken r at a time is

$$ {}_nC_r = \frac{{}_nP_r}{r!}, \qquad \text{or} \qquad \frac{n!}{r!(n-r)!}. $$

${}_nC_r$ is called the binomial coefficient, and is also often written as $\binom{n}{r}$.

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

Calculating probabilities

$$\text{Probability} = \frac{\text{Number of successful outcomes}}{\text{Total number of outcomes}}.$$

Example

The names of 5 members are thrown in a hat and 2 are drawn with the 1st becoming president and the 2nd becoming secretary. What is the probability that Pat becomes president and Kathy secretary?

Solution

There is one successful outcome: Pat as president and Kathy as secretary,

$$\text{Number of successful outcomes} = 1.$$

Total outcomes is the number of permutations of drawing 2 out of a 5

$$\text{Total number of outcomes} = {}_5P_2 = 20.$$

so $$P = 1/20 = 0.05 \text{ or } 5\%.$$

Probabilities involving *independent events with same probability*

Example

If you roll a die ten times, what is the probability that only 3 rolls will come up sixes?

What is the probability of rolling only 3 sixes? e.g., one way it could happen is

$$X, X, 6, X, X, X, 6, 6, X, X$$

where X is a roll that was not 6.

The probability of this particular sequence of independent events is

$$p = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \left(\frac{5}{6}\right)^7 \left(\frac{1}{6}\right)^3 = 1.292044 \times 10^{-3}$$

How many ways can we roll only 3 sixes?

X , X , 6 , X , X , X , 6 , 6 , X , X
X , X , X , 6 , X , X , 6 , 6 , X , X
X , X , X , 6 , X , 6 , X , 6 , X , X

Probabilities involving *independent events with same probability*

Example

If you roll a die ten times, what is the probability that only 3 rolls will come up sixes?

Assuming that all possible combinations are equally probable, then to obtain the overall probability that I will roll only 3 sixes we simply multiply our calculated probability above by the number of combinations that give only 3 sixes. That is,

$$\begin{aligned} \text{P(3 sixes out of 10 rolls)} &= {}_{10}C_3 \left(\frac{5}{6}\right)^7 \left(\frac{1}{6}\right)^3 \\ &= 120 \cdot 1.292044 \times 10^{-3} = 0.15504536, \end{aligned}$$

or roughly a 1 in 6.5 chance.

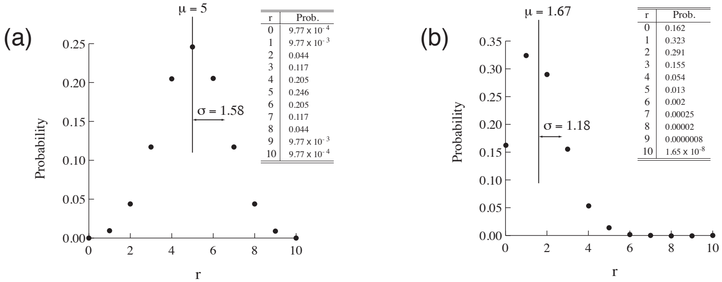
Binomial Distribution

We can generalize this reasoning to the case where the probability of success is p (instead of $1/6$), the probability of failure is $(1-p)$ (instead of $5/6$), the number of trials is n (instead of 10), and the number of successes is r (instead of 3). That is,

$$\boxed{P(r,n,p) = \binom{n}{r} p^r (1-p)^{n-r}}$$

This distribution of probabilities, for $r = 0, 1, 2, \ldots, n$, is called the *binomial distribution*.

Binomial Distribution



The binomial distribution. (a) A symmetric case where $p = 1/2$ and $n = 10$. (b) A asymmetric case where $p = 1/6$ and $n = 10$.

Binomial Distribution

The mean and variance of a discrete distribution is given by

$$\mu_r = \sum_{r=0}^{r_{max}} rP(r), \qquad \text{and} \qquad \sigma_r^2 = \sum_{r=0}^{r_{max}} (r - \mu_r)^2 P(r).$$

The mean of the binomial distribution to be

$$\mu = \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = np,$$

and the variance of the binomial distribution to be

$$\sigma^2 = \sum_{r=0}^{n} \left[(r-\mu)^2 \binom{n}{r} p^r (1-p)^{n-r} \right] = np(1-p).$$

Poisson Distribution

In the limit that $n \to \infty$ and $p \to 0$ such that $np \to$ a finite number the binomial distribution becomes the Poisson Distribution given by

$$P_{\text{Poisson}}(r, n, p) = \frac{(np)^r}{r!} e^{-np}.$$

This distribution often describes the parent distribution for observing independent random events that are occurring at a constant rate, such as photon counting experiments.

The mean of the Poisson distribution is

$$\mu = \sum_{r=0}^{\infty} \left[r \frac{(np)^r}{r!} e^{-np} \right] = np.$$

and the variance of the Poisson distribution is

$$\sigma^2 = \sum_{r=0}^{\infty} \left[(r - np)^2 \frac{(np)^r}{r!} e^{-np} \right] = np.$$

Gaussian Distribution

In the limit of large n when p is not close to zero we can use the *Gaussian distribution* as an approximation for the binomial. That is,

$$P_{\text{Gaussian}}(r, n, p) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{1}{2}\frac{(r-np)^2}{np(1-p)}\right\}.$$

The mean of the Gaussian distribution is

$$\mu = \sum_{r=0}^{\infty}\left[\frac{r}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{1}{2}\frac{(r-np)^2}{np(1-p)}\right\}\right] = np.$$

and the variance of the Gaussian distribution is

$$\sigma^2 = \sum_{r=0}^{\infty}\left[\frac{(r-np)^2}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{1}{2}\frac{(r-np)^2}{np(1-p)}\right\}\right] = np(1-p).$$

Making the substitutions for $\mu = np$ and $\sigma^2 = np(1-p)$ we can rewrite the Gaussian distribution

Gaussian Distribution in the continuous variable limit

Making the substitutions for $\mu = np$ and $\sigma^2 = np(1-p)$ we can rewrite the Gaussian distribution in the form

$$P_{\text{Gaussian}}(r, \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{r-\mu}{\sigma}\right)^2\right\}.$$

Replacing the integer r with a continuous parameter x and get

$$p_{\text{Gaussian}}(x, \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

Gaussian Distribution