Laboratory 1a: Data Distributions

Introduction

In our everyday lives we are used to hearing simple questions, making measurements to answer these questions, and reporting our results, i.e. “What time is it?”, “How far is it?”, “How much does it weigh?” etc. Unfortunately, this familiarity has made us somewhat cavalier in making scientific measurements in that we frequently ignore the uncertainties involved in the measurements; that is, how precise is the result we report? Consider the single measurement of the length of a rigid rod. If we use an ordinary meter stick, we might measure 10.12 cm where the last digit is an eyeball guess between 10.1 and 10.2 cm. How precise is the measurement? Presumably our eye can interpolate at least one-half a gradation mark or one-half a millimeter; therefore, we might report 10.12 ± .05 cm. This is the best we can do. Now, what if we had taken several measurements? How is the situation treated? Have we improved our precision? Can we learn more about the quantity measured and the error associated with it from the distribution of measured values? The answer is frequently yes. In this laboratory, we will examine the uncertainties associated with making several measurements of a single quantity, how to interpret the results, and how to report the findings.

The Parent Distribution—A theoretical distribution you have no access to

For the purposes of today’s laboratory we will ignore the problem of *systematic error*; i.e., errors due to faulty equipment, inaccurate calibration, or human bias, for these errors may only be corrected if their existence is known. If this is the case, then presumably they have been eliminated or compensated for. If systematic errors are present, then our measurements may be *precise* (determined by the smallest scale of the measuring device), but not *accurate* (determined by how well calibrated the measuring device is).

Our interest is in *random error*, either instrumental or statistical. If we make N measurements of a physical quantity, let’s call it x, then we will expect some variability in the values we find. This list of measured values, x_i (i=1…N), while random, are related to each other (note: some of the x_i may be the same). In the case of instrumental error these are also related to *X*, *the real and true value of our physical quantity if only our instruments and methods would work perfectly*. The more measurements we take (bigger N) the better we will be able to infer how X is related to our actual measurements.

The question we wish to address is this: What is the best numerical way to express X from our N measurements and to assign an error which reflects the precision of the measurements? We can start to make sense of this dataset by sorting the values we found into bins. Each bin (using fewer bins than measurements implies n<N) will hold similarly sized measurements and be labeled according to what it holds, e.g. x_j (with j = 1…n). One sensible way to do this is to find the full range of possible values and chop the range up into n equal size pieces Δx wide. Notice that fuller bins represent more likely outcomes. We can visualize these bins by plotting a histogram, i.e., the frequency distribution of the x_j. This binned tally for an infinite number of observations made using identical experimental conditions (all variations in x are due to chance) is called the *parent distribution*. We can formalize this by defining the frequency, $f(x_j)$, as the frequency of occurrence for a value in the interval around $x_{j.}$ This implies that the sum of the frequencies over all values x_j must equal the total number of measurements, or

$$\sum_{j=1}^{N} f(x_j) = N \tag{1}$$

The probability of a *single* measurement value belonging in bin x_j is

$$P(x_j) = \lim_{N\to\infty} \frac{f(x_j)}{N} \tag{2}$$

Equation 2 just normalizes the probability distribution to have an area of one.

$$\sum_{j=1}^{N} P(x_j) = 1 \tag{3}$$

A fictitious distribution is shown in Figure 1. Note that Eq. 3 implies that the area under this curve must equal one.

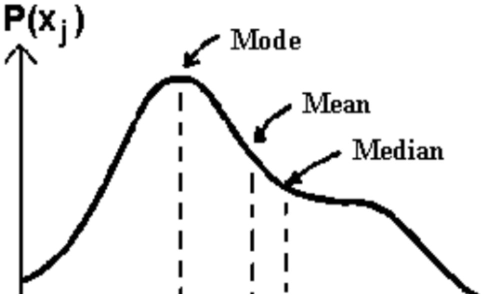


Figure 1 Sample Distribution

Central measures of a distribution

Three quantities are commonly used to describe the center of a distribution.

1. The mean, or average, of x_i which can be calculated by

$$\mu = \sum_{j=1}^{n} x_j P(x_j) \tag{4}$$

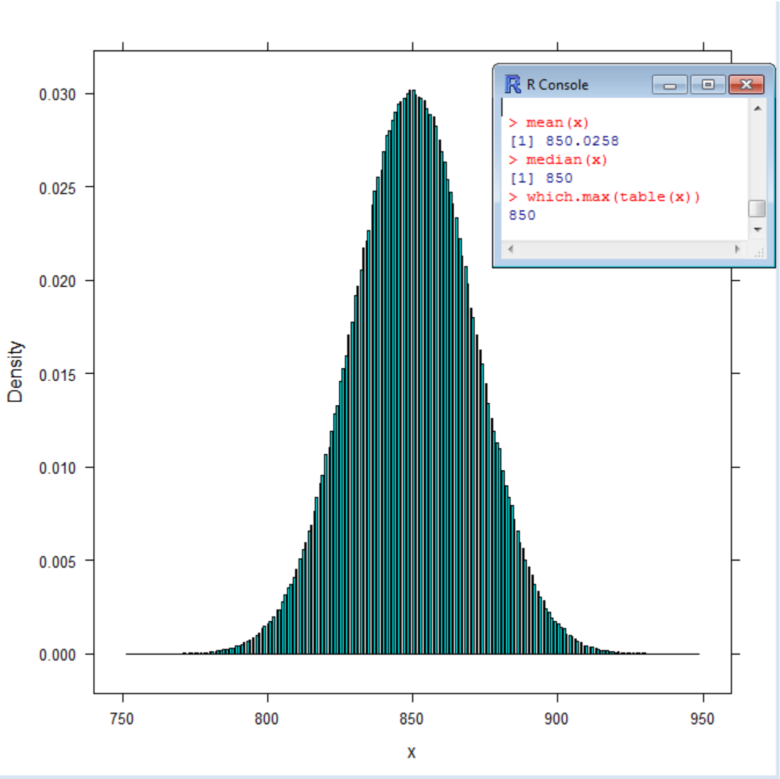
2. The *median,* $\mu_{1/2}$, is the value which splits the parent population in half, that is half of x_i are less than $\mu_{1/2}$ and the other half are greater. This is equivalent to

$$\sum_{x_j < \mu_{1/2}} P(x_j) = \sum_{x_j > \mu_{1/2}} P(x_j) = \frac{1}{2} \tag{5}$$

3. The *mode*, μ_{max}, is the value having the greatest probability of occurring; i.e.,

$$P(\mu_{\max}) \geq P(x_j) \tag{6}$$

Any or all of these parameters might be of use to the experimentalist, although the mean is generally the most quoted value. This is because in many situations the physical quantity we are looking to estimate is not skewed by a small number of extreme events. Conviniently, many distributions we will encounter are symmetric around a central peak, then $\mu = \mu_{1/2} = \mu_{max}$. An example of such a distribution appears in Fig. 2.



```
> mean(x)
[1] 850.0258
> median(x)
[1] 850
> which.max(table(x))
850
```

Figure 2 An approximation to a symmetric parent distribution.

In addition to estimating the center of the distribution it is convenient to be able to describe the spread in the data. One possible choice is the deviations from the mean, $\delta_i = x_i - \mu$. While this can be calculated for each data point its average,$< \delta >$, tells us nothing since

$$< \delta > = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)\right] = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}x_i\right] - \mu = \mu - \mu = 0 \qquad (7)$$

One common, and simple, way to fix this is by calculating the average squared deviation which is called the *variance* σ^2

$$\sigma^2 =< \delta^2 >= \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right] = \sum_{j=1}^{N}P(x_j)(x_j - \mu)^2 \qquad (8)$$

The standard deviation, which is a measure of the spread is then $\sigma = \sqrt{\sigma^2}$. Note that the standard deviation is the root mean square of the squared deviations.

We can now address the question we posed ourselves: given the parent population of x, and thus the parent distribution, what is *X*, the "true value" of x? By convention we call the mean μ the true value of x and the standard deviation σ is a measure of the uncertainty involved in attempting to measure the mean in any *single* measurement.

The Sample Distribution—the distribution of your data

In the previous section we considered a theoretical distribution of a true value *X* found from an infinite set of measurements of x. Clearly cannot collect that much data, nor would we want to. Instead we will have to sample the parent distribution by making a finite number of measurements. Using this smaller set of data, what then are our best estimates of the parent mean and standard deviation?

Our best estimate of the parent mean is the *sample mean*, $< x >$,

$$< x >= \frac{1}{N}\sum_{i=1}^{N}x_i = \frac{1}{N}\sum_{j=1}^{n}x_j\, f(x_j) \qquad (9)$$

This is a good choice of definition for $< x >$ as

- the deviations $\delta_i = x_i -< x >$ sum to zero and
- the sum of δ_i^2 is minimized with respect to the choice of $< x >$.

Our best estimate of the sample standard deviation is given as

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i -< x >)^2} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{n}f\left(x_j -< x >\right)^2} \qquad (10)$$

The primary difference between (10) and (8) is the replacement of the factor N with N - 1. This change is justified in two ways.

1. The number of degrees of freedom, ν, is N, the number of measurements x_i. We use up one degree of freedom determining $< x >$ so $\nu \rightarrow N - 1$ when we determine the sample standard deviation. This makes s slightly larger than if we truly knew *X* to be μ.
2. If N = 1, we have no means of knowing the possible dispersion of the x_i values; thus, (10) is meaningless in this case.

Error Propagation

Given a function g of several independent variables, g(a,b,c), then the total differential of g is:

$$dg = \frac{\partial g}{\partial a} da + \frac{\partial g}{\partial b} db + \frac{\partial g}{\partial c} dc \tag{11}$$

Let us identify the differentials with the deviations from the mean values of g, a, b, and c; i.e., in the ith measurement we would find

$$(g_i - < g >) \cong (a_i - < a >)\frac{\partial g}{\partial a} + (b_i - < b >)\frac{\partial g}{\partial b} + (c_i - < c >)\frac{\partial g}{\partial c} \tag{12}$$

According to equation (9), the variance of g for an infinite number of g_i s is

$$\sigma_g^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (g_i - < g >)^2 \tag{13}$$

In truth the g_is are determined by the a_i, b_i, and c_is so it is appropriate to substitute (12) into (13). If the variations in a, b, c are independent, then for an infinity of measurements the cross terms in the squaring of the bracket cancel out, and $\sigma_g^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (g_i - < g >)^2$

$$\sigma_g^2 \cong \sigma_a^2 \left(\frac{\partial g}{\partial a}\right)^2 + \sigma_b^2 \left(\frac{\partial g}{\partial b}\right)^2 + \sigma_c^2 \left(\frac{\partial g}{\partial c}\right)^2 \tag{14}$$

This is a key result and allows one to connect uncertainty in measured quantities (a, b, and c) to uncertainty in desired quantities (g).

The Error of the Mean

We have given estimates for the sample mean, Eq. 9, and sample standard deviation, Eq. 10, which best describe the parent distribution of *X*. We will now determine, using our N measurements, the precision with which we know the sample mean using the fact that each measurement of x is independent of the others.

Now, applying Eq. 14 to the sample mean, $< x >$, where the variables are the measured values x_i, the variance of the sample mean is

$$s_{<x>}^2 = \sum_{i=1}^{N}\left[s_i^2\left(\frac{\partial < x >}{\partial x_i}\right)\right] \tag{15}$$

where s_i^2 is the variance associated with measurement of x_i. If all data points are taken under the same experimental conditions, then $s_i^2 = s^2$, the sample variance. Furthermore, from equation (9) we get :

$$\frac{\partial < x >}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) = \frac{1}{N} \tag{16}$$

thus:

$$s_{<x>}^2 = \frac{s^2}{N} \tag{17}$$

The *error in the sample mean* is then:

$$s_{<x>} = \frac{s}{\sqrt{N}} \tag{18}$$

This is a key distinction, while the standard deviation of the sample distribution is determined by the sources of error in the measurement, we can reduce the error in the mean to any level we choose by increasing the number of measurements we take. There are, however, diminishing returns for this trick. You must take four times as much data to cut the error in the mean in half. For example let $< x >$= 10.02 cm and $s_{<x>}$ =0.13 cm. If the 1 in $s_{<x>}$ is uncertain, then we know nothing about the 3. Thus the number should be reported as 10.0 ± 0.1 cm.

To report the mean of a set of measurements one specifies: $< x > \pm s_{<x>}$

The first significant digit of the error in the mean is the only significant digit.

Distributions

While there are many probability distributions, in the measurements we will make only a few appear consistently. We will study the Gaussian, Poisson, and binomial distributions.

Gaussian or Normal Error Distribution

The Gaussian distribution is the most important probability distribution in the analysis of physical data where the discrepancies from the "true" value are due to *random* experimental errors. The characteristics of this distribution are

1) It is a *continuous* distribution; i.e., one in which an infinity of different measured values are possible.

2) It is symmetric about a central peak; thus the mean, mode, and median are the same.

3) It is *solely determined by a mean* (determined by the distribution) and *standard deviation* (determined by the experimental precision).

The normalized Gaussian probability density function (probability per interval) is defined as

$$\Omega_G(x_i,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \tag{19}$$

We convert this relationship to a dimensionless probability density by multiplying the density by the width σ of the distribution and making it a function of the dimensionless parameter:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{20}$$

which represents some fraction of a standard deviation. Then

$$\mathrm{P}_G(z_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z_i^2}{2}} \tag{21}$$

Figure 3 illustrates the Gaussian probability density function $P_G(z)$ as a function of z from the mean. A third parameter, dependent on the distribution and sometimes quoted, is the full width at half maximum

(FWHM) $\Gamma \cong 2.35\sigma$. *It is important to recognize that the dimensionless representation of Fig. 3 and Equation 21 are valid for any Gaussian distribution.*

This is a continuous distribution defined at every point. But the reason we are interested in this distribution is to determine how likely any one measurement is. We will infer this from the probability distribution function (pdf) but must recognize that the pdf is not directly the probability. To see why, recall that all real experimental devices have a precision Δx determined by the smallest scale division of the device. We don't actually know where in the interval Δx our measurement belongs. So the probability is actually the integral of the density function over the interval from $\left(x_i - \frac{\Delta x}{2}\right)$ to $\left(x_i + \frac{\Delta x}{2}\right)$; as Δx is the limit of our experimental precision. Geometrically, the probability associated with the measurement of x_i is the area under the probability density curve in the interval Δx, as represented in Fig. 4.

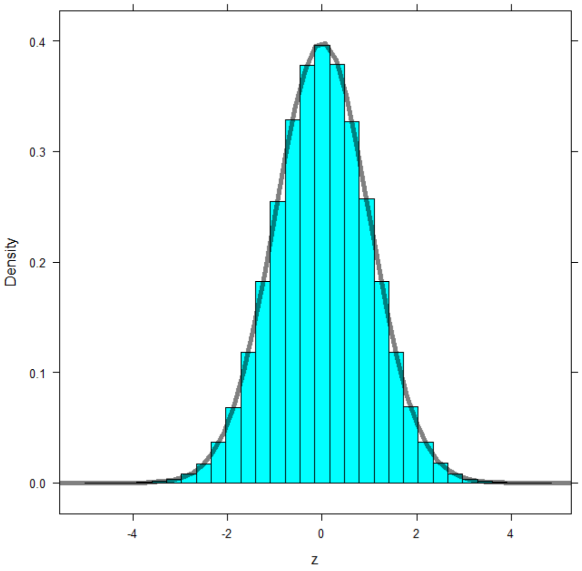


Figure 3 A Gaussian distribution is superimposed on a 1e6 sample distribution histogram.

There are two ways we might evaluate this, by hand or using R. First by let us consider doing this by hand. There are two likely senarios.

1. Variability in the quantity to be measured drives the uncertainty. Here, the precision of our measuring device, Δx, is much better than the standard deviation in our measurements, σ. This means that $\Delta x \ll \sigma$ and the probability density function will be slowly varying over the interval Δx. We can then treat $P_G(z(x_i))$ as being locally constant and the integrated probability over the interval centered at x_i is the central value $P_G(z(x_i))\Delta z$ where $\Delta z = \frac{\Delta x}{\sigma} \ll 1$. This quantity is the probability of measuring x_i in a *single_measurement* if your experimental

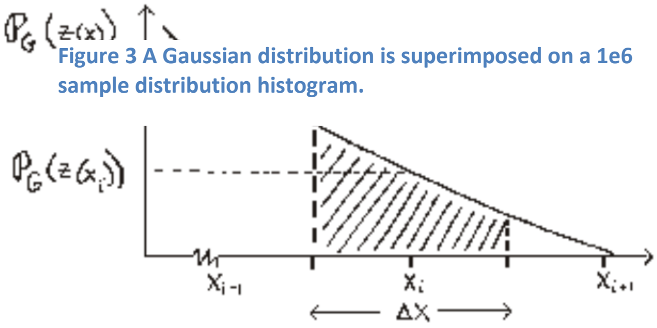


Figure 4: Portion of the Gaussian Distribution

accuracy is Δx.

2. Rapid variation of the probability density, $\Delta x \geq \sigma$, i.e. the precision is on the order or greater than the standard deviation. Here the variation is not linear; and the former procedure is fails. We must instead integrate the function over the interval Δx or the interval between the end points $(x_i - \Delta x/2)$ to $(x_i + \Delta x/2)$. Error functions ensue and it is really much simpler not to go this route.

The Poisson Distribution

The Poisson probability distribution is generally appropriate in counting experiments where the number of observed events is small compared to the number of objects that could cause the event. An example is radioactive decay, where the number of disintegrations in a given time interval is small compared to the number of radioactive atoms present. The dispersion of events measured in this case is not due to random error, but to the *statistical* nature of the physical problem. The characteristics of this distribution are:

1. It is a discrete distribution. You measure integer values.
2. The distribution is asymmetric, that is the mean has a different value than the mode.
3. The distribution (including standard deviation) is *determined solely by the mean*.

The Poisson probability distribution, *normalized to one*, is

$$P_P(x_i, \mu) = \frac{\mu^{x_i} e^{-\mu}}{x_i!} \tag{22}$$

In this case, when the probability distribution is only a function of interval x_i, the variance can be found from equations (8) and (22) with a little manipulation as

$$\sigma^2 = \sum_{x_i=0}^{\infty} (x_i - \mu)^2 P_P(x_i, \mu) = \mu \tag{23}$$

Thus the standard deviation is the square root of the mean.

Figure 6 shows probability distributions for means of 1.67 and 10, respectively.

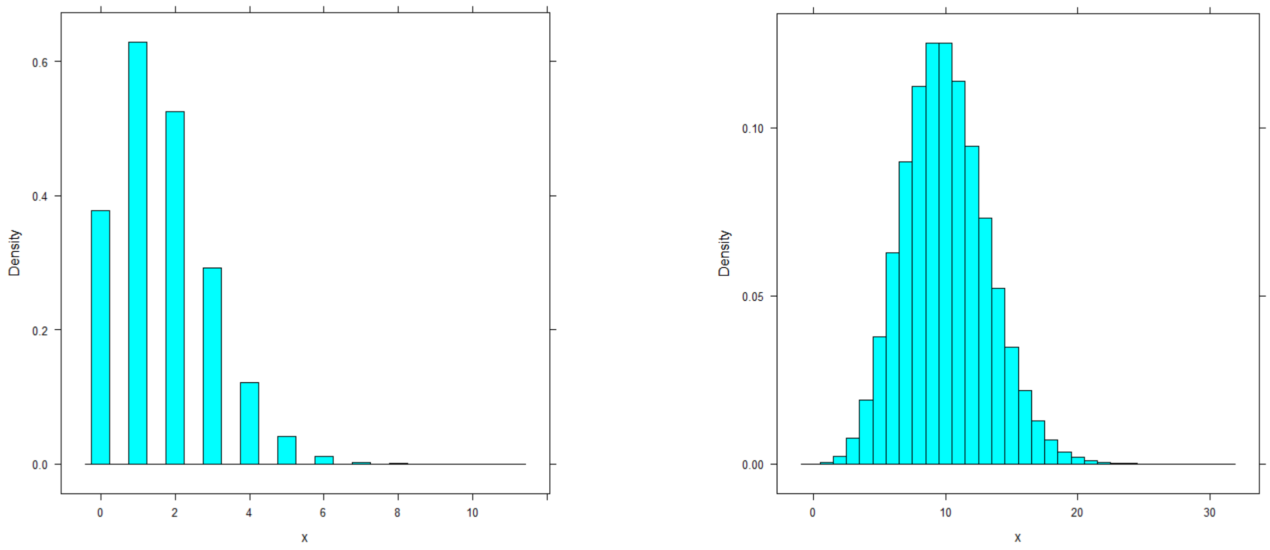


Figure 6: Sample approximations to Poisson distributions with means μ=1.67 and μ=10. Each sample has 1e6 measurements.

You should note three points:

1) While the possible measured values are integers; i.e., 0,1, 2, etc., the mean need not be integer.

2) For $\mu \ll 10$ the probability for no events occurring is non-negligible.

3) For $\mu \gg 10$ the Poisson distribution becomes more symmetrical. I fact, for large μ the distribution becomes Gaussian with the added feature that $\sigma = \sqrt{\mu}$.

The Binomial Distribution

The binomial probability distribution relates the probability of certain combinations of events occurring where each event can have only one of two possible outcomes. To develop this distribution, consider the flip of a single coin. While we cannot predict the result of a single toss, after a large number of tosses we would expect to find 50% of the tosses heads and 50% tails. Suppose now there were two coins. There are four possible distinguishable events occurring over a large number of tosses. Distinguishable means I know which coin is which.

Coin 1	P1	Coin 2	P2	P=P1*P2 Distinguishable	P Indistinguishable
Head	50%	Head	50%	25%	25%
Head	50%	Tail	50%	25%	50%
Tail	50%	Head	50%	25%	
Tail	50%	Tail	50%	25%	25%

Table 3: Two Coins With Equal Probability

Notice that if we cannot distinguish between the coins, the combinational probability for a head-tail pair is 50%.

Now consider the general case of n tossed coins. The probability of any distinguishable toss is just $(½)^n$ since there are 2^n possible ways the coins can land in an ordered fashion. Let us now ask how many of these ordered tosses lead to x heads. The first head could be any one of the n coins. The second head must be one of the other (n - 1) coins, and so on until the x^{th} head must come from one of the $(n - x + 1)^{th}$ coins left. Thus the number of possible permutations of x heads from n coins is

$$P(n,x) = n(n-1)...(n-x+1) = \frac{n!}{(n-x)!} \tag{24}$$

Note that we have kept track of which coin is which. What if we don't care? That is, of the x heads, who cares which one we choose first or last? If not, then there are x! different ways the x heads could have been picked up. Thus, the actual number of indistinguishable combinations of x heads is the number of distinguishable permutations divided by the *degeneracy* x!,

$$C(n,x) = \frac{P(n,x)}{x!} = \frac{n!}{x!(n-x)!} \equiv \binom{n}{x} \qquad (25)$$

The symbol $\binom{n}{x}$ is pronounced n choose x. The probability for observing x heads from a toss of n coins is the number of combinations C(n,x) times the probability of any single combination or C(n,x)(½)^n. Note the distribution is symmetric since C(n,x) = C(n,n-x).

What if our coins are not fair? A more general formulation will allow for unequal probabilities of the two possible states; e.g., p < 1 is the probability of heads, q <1 is the probability of tails, q + p = 1. In this case the binomial probability distribution function*, normalized to one is*,

$$P_B(x,n,p) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (26)$$

The distinctive feature of this binomial probability distribution is that *it is determined solely by the number of objects*, each with two possible states. In situations where p = q the distribution will be symmetric.

Let us consider an example using two fair coins (i.e., n = 2, p = ½, q = ½). We can extract the probability of each outcome from equation (26):

$P_B\left(x,2,\frac{1}{2}\right) = \frac{2!}{x!(2-x)!}\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{2-x} = \frac{2!}{x!(2-x)!}\left(\frac{1}{2}\right)^2$, so:

For x=0 head: $P_B\left(0,2,\frac{1}{2}\right) = \frac{2!}{0!(2-0)!}\left(\frac{1}{2}\right)^2 = \frac{1}{4}$

For x=1 head: $P_B\left(1,2,\frac{1}{2}\right) = \frac{2!}{1!(2-1)!}\left(\frac{1}{2}\right)^2 = \frac{1}{2}$

For x=2 heads: $P_B\left(2,2,\frac{1}{2}\right) = \frac{2!}{2!(2-2)!}\left(\frac{1}{2}\right)^2 = \frac{1}{4}$

This agrees with our expectations in Table 3.

The binomial distribution’s name comes from the fact that when expanding the sum of two numbers raised to the power n, the coefficients of the expansion are given by $\binom{n}{x}$; i.e.,

$$(p+q)^n = \sum_{x=0}^{n}\left[\binom{n}{x} p^x q^{n-x}\right] \tag{27}$$

This is the *binomial theorem*.

The mean of the binomial probability distribution can be found from equations 4 and 27 and a little work for the last step to be

$$\mu = \sum_{x=0}^{n} x P_B(x, n, p) = np \tag{28}$$

μ is the parent mean since we know the theoretical probability. For equal probabilities, i.e. for fair coins, p = q = ½, then μ = n/2 and the distribution is symmetric. The standard deviation can be found from equations (8) and (27) and a little work to be

$$\sigma^2 = n\sum_{x=0} (x-\mu)^2 P_B(x, n, p) = np(1-p) \tag{29}$$

If p = q = ½, then $\sigma^2 = \frac{\mu}{2}$ or $\sigma = \sqrt{\frac{\mu}{2}}$. Figure 7 illustrates the theoretical distribution for two possible cases with 10 coins (n = 10) : 10 fair coins (q = p = ½) and 10 trick coins (q = 5/6, p = 1/6).

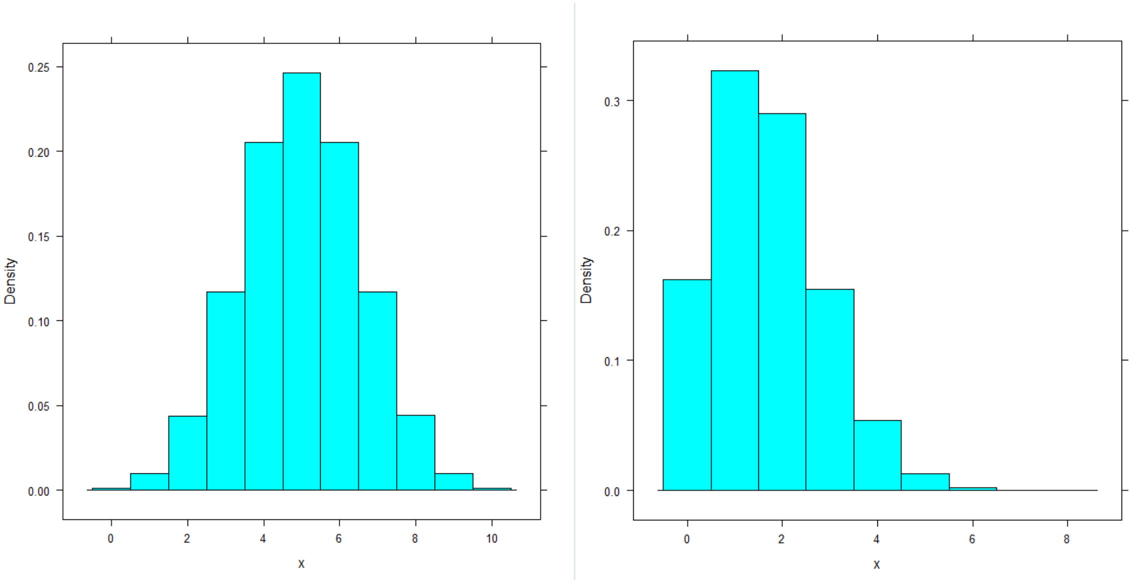


Figure 7: Binomial Distributions for ten coins with 1:1 and 1:5 odds.

Experiments

Gaussian Distribution Part One:

Small Dataset

A follow along example:

The data in Table 1 represents N=5 measurements of the width of a block. The same ruler was used for all measurements and so we expect that both Δx, the uncertainty in the ruler, and σ_i, the uncertainty in the measurement, are constant.

By hand calculate:

1. The sample mean.
2. The sample standard deviation.
3. The error in the sample mean.

Table 1

Width (mm)
20.45
19.91
19.41
19.55
19.86

Using R repeat this calculation.

Start R and load the fastR2 package:

- require(fastR2)

You now need to bring the data into R, for this small amount of data you can type it in as a vector. R uses arrows for assignment, creating a lone variable called width and assigning it the value 3 can be done this way:

- width<-3

or this way:

- 3->width

R builds up vectors through concatenation. A 3D position vector can be defined and filled with a vector (1.1, 1.2, 1.4) using:

- r<-c(1.1,1.2,1.4)

Use this to build a vector called blockWidth holding the block widths in Table 1.

- blockWidth <-c(…)

R will count the number of data points you have, the command length(v) returns the length of vector v.

- length(blockWidth)

R also calculates means and standard deviations of vectors, these commands are mean(v) and sd(v). Use these commands to calculate sample mean, standard deviation.

- mean(blockWidth)
- sd(blockWidth)

R does a good job of giving you an overview of your data showing the min, max, mean, median, as well as the 1st and 3rd Quartiles. The command for this is summary(v).

- summary(blockWidth)

R also does a good job of showing you your data graphically. Here a histogram is useful gf_dhistogram(v).

- gf_dhistogram(~blockWidth)

If the default choices for the histogram turn out to be poor it can be useful to set either the size of the bins the data are placed in, e.g. making each bin 0.2 units wide can be done with,

- gf_dhistogram(~blockWidth, binwidth=0.2)

or set the number of bins to place the data in, e.g. selecting 3 bins can be done with,

- gf_dhistogram(~blockWidth, bin=3)

You may be interested to see the histogram and to show a probability density function that is fit to it, for this you will need a bit more syntax:

- gf_dhistogram(~blockWidth) %>% gf_dist(“norm”,mean=mean(blockWidth),sd=sd(blockWidth), color=”red”)

(You are stacking two graphs, the histogram and a “normal” distribution with appropriate mean and standard deviation). Note that quotation marks from this document don’t cut and paste properly into R.

While R doesn’t have a built in function to calculate the sample mean error you can build it up from the commands above, if you use the squareroot, sqrt(), function. Using a combination of the commands that you have just learned write a script that does this in a way that doesn’t require you to keep track of the number of data points. **Show me this program before you proceed.**

Calculating integrals:

Since R knows about distributions it can find the area under theoretical curves. For example, the command pnorm:

- pnorm(q,mean,sd)

gives the probability of a measurement yielding a value less than q. You must also provide information about the normal (Gaussian) distribution in question, that it is centered on mean with standard deviation sd. For an interval, the probability that you measurement will be found in the interval is calculated by using pnorm twice. **Explain to me how to do this before you proceed.**

Good habits:

Think of the analysis you do in R as programs and write them in the R editor (under the menu File, New Script) and run them as a whole (under menu Edit, Run All) or parts of them (highlight the part and under Edit, select Run Line or Selection). This helps to make the analysis a cohesive whole rather than a bunch of small bits.

More convenient data entry:

It is nice to see all the data in front of you. Excel does a nice job with this. Here is a recipe for doing this:

1. Open excel and enter the variable name and data from table 1 in a column.
2. Save the spreadsheet as a CSV (Comma delimited) file. It will complain but persevere and remember where you saved it.
3. Open the file using blockA<-read.csv("fullFileName") where
   a. fullFileName is the exact location of the data including the directory and
   b. blockA holds your data in R.

   (R knows about your computer so if you start to type the location and then press TAB it will fill in the rest that it knows how to, and if there is ambiguity it a second TAB will give a list of options.) For example a file called blockie.csv saved to my desktop would have a fullFileName of C:/Users/douglas.armstead/Desktop/blockie.csv. If your computer is online then urls are valid file names for read.csv().

4. blockA holds the data vector and to access the data in the vector use blockA$Width anywhere you used blockWidth before.

Large Dataset

R is even more useful with larger datasets. Let us now use a more extensive set measurements of the block stored in a CSV file on the course website, block.csv. Assign data from the file http://facultyweb.cortland.edu/douglas.armstead/S19/Intermediate/block.csv to blockB. Tip: don't forget to use quotation marks.

The data in blockB is more extensive than the data in blockA. This is because there are multiple variable, i.e., width , length, and depth, each with many data points. The structured set of data you brought in and assigned the name blockB is called a data_frame (block A was too). You can look at the data you brought in a couple of ways. A good place to start is by looking at the top of your data using the command head(data_frame). This will tell you the names of all the variables in the data_frame and there first few values.

- head(blockB)

To display all of the data in your data_frame, type your data_frame's name and hitting enter.

- blockB

Finally you can use the data editor. This is under the Edit tab. You will need to remember the name of your data_frame to do this. This not only gives you a handy way to look at your data, but also a convenient place to edit it if need be.

Some of the commands we used on the small dataset will work directly on a data_frame, e.g., summary(), others require individual vectors. You can access a vector inside a data_frame if you separate their names with a dollar sign. For example:

- blockB$depth

gives you access to the vector depth in the data_frame called blockB.

For the dataset in blockB use R to

1. Calculate the sample mean of the block width.
2. Calculate the sample standard deviation of the block width.
3. Calculate the error in the sample mean of the block width.
4. Draw a frequency histogram of the block width data, include the probability density function.
    a. With a pencil and a ruler label the mean, standard deviation, and FWHM on the histogram.

5. Calculate the probability that your next measured width will fall between 20mm and 20.3mm.

Report your results for blockB and include a copy of the full script you wrote to do the analysis.

Gaussian Distribution Part Two—Real Data:

Use a meter stick to measure the width of four of the lab tables in the room. For each table repeat the measurement at least five times each for a minimum of 20 measurements. Make a realistic estimate for the precision of your measurements considering the precision with which the stick can be read and your ability to make it line up square with the table. From your data:

1. Calculate the sample mean.
2. Calculate the sample standard deviation.
3. Calculate the error in the sample mean.
4. Compare your estimate for the statistical uncertainty (i.e. the standard error in the mean) with your estimate for the measurement uncertainty inherent in your data. Discuss the significance of this finding.

Report your results for the tables and include a copy of the full script you wrote to do the analysis.

Poisson Distribution Part One:

The data in Table 2 represent N = 50 measurements of emission rates from a radioactive source during 1 second intervals.

Counts (1/s)	Frequency	Counts (1/s)	frequency
0	0	6	5
1	3	7	6
2	5	8	5
3	7	9	3
4	7	10	1
5	8	11	0

Table 2: Poisson Distribution Data for Radioactive Decay

1. Use excel to enter the data and assign it to a data_frame in R.
    a. Use the variable names counts and freq.
    b. Call the data_frame emission.
2. Calculate the sample mean, this will be more complicated than just using mean() and will instead require the definition of the mean and use of the function sum(). It has the syntax sum(v) where v is a vector, (see Eq. 9).
3. Calculate the sample standard deviation directly from the data, again this requires care.
4. Calculate the error in the sample mean.
5. Use R to draw a frequency histogram of the data and label significant features. Since you don't have the raw data you will need to make use the function gf_point(), it has the syntax gf_point(y~x,data=data_frame).
6. Calculate the theoretical standard deviation from equation (23) and compare to what you obtained in step 2. Does this tell you anything?

Report your results for the emission data and include a copy of the full script you wrote to do the analysis.

Poisson Distribution Part Two:

Import the data set from the url:

http://facultyweb.cortland.edu/douglas.armstead/S16/Intermediate/RadiationCountData.csv

This data represents N = 5,000 measurements of radiation counts from a uranyl nitrate source taken over one second intervals. We can use this large-N data set of radioactive decay data to explore how the value of N effects your results. We will take three samples from the data to produce subsets with 50, 500, and 5,000 data points. To sample the dataset in R with the sample() function, the syntax is sample(originalData, # of samples). Assign your sampled data to new data_frames for example

- r50<-sample(originalData$variable, 50)

1. Calculate the sample mean.
2. Calculate the sample standard deviation directly from the data.
3. Graph a frequency histogram of the data overlaid with the theoretically expected Poisson distribution using the mean and standard deviation calculated from the complete 5,000 point data set in each case. This can be done in a similar way as with a normal distribution swaping pois in for norm and lambda in for mean.

- gf_dhistogram(~v, data=df) %>% gf_dist("pois",lambda=mean(df$v), color="red")
- source("http://facultyweb.cortland.edu/douglas.armstead/S16/Intermediate/showRad.R")

4. Calculate the theoretical standard deviation from the mean found in step 1 and compare it with the standard deviation you obtained in step 2.

Perform the above analysis on each of the three data subsets (i.e. N=50, 500, or 5,000) and discuss the significance of the number of data points collected to the applicability of the theoretical Poisson distribution. Include a copy of the full script you wrote to do the analysis.

Binomial Distribution

You will now perform two experiments to look at the difference between fair coins and unfair coins. Since we don't have any unfair coins at our disposal you will use dice as a proxy for the coins in both experiments.

- In the fair coin experiment you will interpret
  - 1, 2 and 3 as heads
  - 4, 5, and 6 as tails.
- In the unfair coin experiment you will interpret
  - 1, 2, 3, and 4 as heads
  - 5 and 6 as tails.

Roll nine dice 100 times (N = 100 is the number of replicates of the experiment). As there is nothing to be gained from using separate rolls for the fair and unfair experiments (i.e., rolling your dice 200 times), extract one fair and one unfair data point from each roll. It will simplify (and expedite) your data taking to record 3 pieces of information from each roll: the number of dice with values 1-3, the number of dice with value 4, the number of dice with values 5-6. (Note that recording the three types of outcomes provides a check on your data recording, how?)

Record your data in a spreadsheet and bring it into R.

For both interpretations of the experiment use R to:

1. Calculate the sample mean number of heads.
2. Calculate the sample standard deviation number of heads.
3. Calculate the error in the sample mean number of heads.
4. Plot a histogram of your data using the following syntax and an adaptating v, df, x, p to the situation at hand.
   - gf_dhistogram(~v, data=df) %>% gf_dist("binom",size=x, prob=p, color="red")
5. Use equations (28) and (29) to calculate the theoretical mean and standard deviation. Compare to your results to the results of steps 1 and 2. Does this tell you anything?

Report your results for the coin data and include a copy of the full script you wrote to do the analysis.

Conclusions

While there are more than three important distributions, the Gaussian, Poisson, and Binomial distributions occur time and time again in experimental situations and should be understood and appreciated. Note further that there are several complications which occur frequently occur, which we have not discussed, but which deserve careful consideration; to name three: background, unresolved or partially resolved distributions, and unequal uncertainties. Nonetheless, this introduction should provide you with sufficient information to indicate your best estimate of a measured value and the error to be associated with it for each of the remaining experiments in this course.

I make one final plea to stress the need to

- making several measurements of the unknown quantity, and
- plotting your results.

Imagine you are to measure a certain unknown voltage several times. You expect the errors inherent in such a measurement to be experimental, not due to counting, and so expect the distribution of measured values to be Gaussian. With results in hand, you plot a histogram of your data. If you find the distribution to be skewed, then you would know immediately that something is amiss, either you are not collecting the data that you expected (and so have a systematic experimental error) or there is a flaw in your reasoning. You can infer this because the Gaussian distribution should be symmetrical. The skew in your data's distribution is only detectable if you both collect enough data, and plot it. From one or two measurements you might never suspect that anything was wrong.

WHEN YOU FINISH: LEAVE THINGS AS YOU FOUND THEM!