Statistics

General comment: Formulas which have boxes around them are formulas you will have to memorize.

Definitions:

<u>Probability:</u> The probability p that something happens is the fractional chance that it will happen. If you perform a large number N of measurements, and something happens x times, the probability can be estimated as

$$\hat{p} = \frac{x}{N}$$

However, this is only an estimate of the probability. In fact, if N isn't large, this isn't the best estimate, but it is close enough for our purposes. We will discuss this in more detail later.

<u>Distribution:</u> When there are many possible outcomes for a measurement x, you can make a table of all the possibilities $\{x_1, x_2, \ldots, x_k\}$ and their respective probabilities $\{p_1, p_2, \ldots, p_k\}$. Such a table is sometimes called a *distribution*. Knowing the distribution of some random variable x tells you as much as you can know about a random variable.

x_1	p_1
x_2	p_2
x_3	p_3
…	…
x_k	p_k

<u>Median:</u> When you have a distribution, the median is the number such that half the time you get above it and half the time you get below it. If you have a series of measurements, the best guess for the median is the number in the middle. If the measurements of something are 4,5,4,3,9,2,3,4,5,8,6,4,8, the best guess for the median is 4.

<u>Mode:</u> The mode is the number that comes up most often. It is the value with the highest probability. If you make a series of measurements, the best guess for the mode is the number that comes up most often.

<u>Mean:</u> For a probability distribution in which the measured quantity takes on values x_1, x_2, …,x_k with probability values p_1, p_2, …,p_k, the mean is given by

$$\mu = x_1 p_1 + x_2 p_2 + \cdots x_k p_k .$$

For a series of measurements of x, the best guess for the mean is given by

$$\boxed{\hat{\mu} = \frac{x_1 + x_2 + \cdots + x_N}{N}}$$

<u>Average:</u> The mode, median, and mean are all different types of *averages*. The most useful all-around average for a numerical variable is probably the median, but the easiest to use mathematically is the mean.

Standard Deviation: A measure of how spread out the distribution is. Numerically, it is given by the formula

$$\sigma^2 = p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \cdots + p_k(x_k - \mu)^2$$

Another equivalent formula is

$$\sigma^2 = p_1 x_1^2 + p_2 x_2^2 + \cdots + p_k x_k^2 - \mu^2$$

When you have a series of measurements $x_1, x_2, \ldots, x_N$, the best estimate of the standard deviation is

$$\hat{\sigma} = \sqrt{\frac{(x_1 - \hat{\mu})^2 + (x_1 - \hat{\mu})^2 + \cdots + (x_N - \hat{\mu})^2}{N - 1}}$$

An equivalent formula is

$$\hat{\sigma} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_N^2}{N - 1} - \frac{(x_1 + x_2 + \cdots + x_N)^2}{N(N - 1)}}$$

Many calculators, as well as programs such as Excel, will automatically calculate means and standard deviations for you, so it is rare that you would actually use these formulas.

Notation: The mean and standard deviation of a random variable x are commonly written in the form $x = \mu \pm \sigma$. So if you are told the average height of a person is 1.7 ± 0.1 m, the mean is 1.7 m and the standard deviation is 0.1 m.

Math with Random Variables

Multiplying by a constant: Suppose you have a random variable x with mean μ and standard deviation σ. Now suppose you multiply by a constant k. Then the mean will get multiplied by k and the standard deviation will also get multiplied by k. That is, if

$$x = \mu \pm \sigma \qquad \text{then} \qquad \boxed{kx = k\mu \pm k\sigma}$$

For example, when you roll a single fair die, the result you get is $x = 3.5 \pm 1.7$. Now, suppose you are on a game show and you get $20 for every pip on a die you throw. Then the amount of money you would win would be

$$20x = 70 \pm 34 \text{ dollars}$$

Addition: This is a little more complicated, and not very intuitive. Suppose you have two random variable x and y, and you add them, you will get a new random variable $x + y$.

The means will simply add, but the standard deviations add in quadrature (which is explained by the equations below). In other words,

$$\mu_{x+y} = \mu_x + \mu_y \qquad \text{and} \qquad \sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2} \,,$$

Or, to summarize

$$\boxed{x + y = \left(\mu_x + \mu_y\right) \pm \sqrt{\sigma_x^2 + \sigma_y^2}}$$

For example, if you add the results of two dice, each of which has $x, y = 3.5 \pm 1.7$, the result will be $x + y = 7.0 \pm 2.3$.

Subtraction: This is almost the same as addition. Suppose you have two random variables x and y, and you subtract them, then you will get a new random variable. The means will simply subtract, but the standard deviations add in quadrature. In other words,

$$\mu_{x-y} = \mu_x - \mu_y \qquad \text{and} \qquad \sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2} \,,$$

or, to summarize

$$\boxed{x - y = \left(\mu_x - \mu_y\right) \pm \sqrt{\sigma_x^2 + \sigma_y^2}}$$

For example, if you subtract the results of two dice, each of which has $x, y = 3.5 \pm 1.7$, the result will be $x - y = 0.0 \pm 2.3$.

Normal Distributions:

Normal distributions come up in many situations. They are also known as Gaussian or bell curves. A normal distribution is described completely by the mean μ and standard deviation σ, and is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} .$$

The probability that x will differ from μ by a certain number of standard deviations σ is given by the chart on the next page.

Central Limit Theorem: The central limit theorem states that if you add together a large number of random variables, the distribution always ends up looking like a normal distribution. For example, if you roll a single 6-sided die, the distribution doesn't look very much like a normal distribution, but if you add together three 6-sided dice, it looks almost exactly like a normal distribution.

z-value: The z-value is the number of standard deviations σ away from the mean μ that some measurement x is. It is given by

$$\boxed{z = \frac{|x - \mu|}{\sigma}}$$

z-test: **This is the fundamental concept you have to learn.** The significance of the z-value is that the bigger the z-value is, the more unlikely it is that something has happened purely by chance. This leads to the idea of the z-test.

Suppose you know the mean μ and standard deviation σ for some random variable, which has a normal distribution. You now make a single measurement x and want to know how "remarkable" it is that x is high or low. You use the z-test. You calculate z. Then you can use the graph below to estimate the probability that such a large z-value occurred. Then you have to make a judgment call – could such a low probability be luck, or must you conclude that there is something else going on?
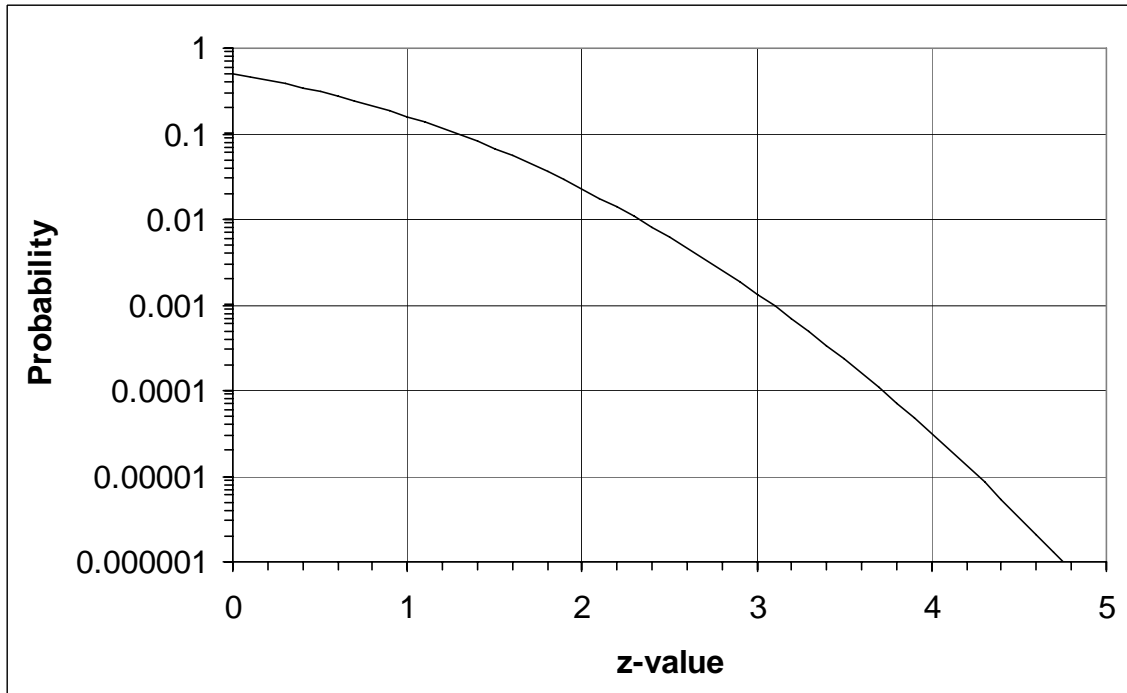
For example, suppose you know the typical height of people is 1.70 ± 0.10 m tall. That is, the mean height is $\mu = 1.70$ m, and the standard deviation is $\sigma = 0.10$ m. Now, suppose you think that someone might be exceptionally tall, and you measure them, and it turns out they are 2.0 m tall. This has a z-value of

$$z = \frac{|2.0 - 1.7|}{0.1} = 3.0$$

The probability that someone is this tall, purely according to chance, is then about 0.0015, or 0.15%. This is pretty unlikely, but not impossibly lucky.

There is no hard and fast rule here. In general, if you get a z value of 2 or less, you haven't proven anything. If you get a z-value in the range of 2 – 5, you probably have a pretty convincing case. If you get a z-value of 5 or more, the case is conclusively proven. If you want more accurate values of probabilities than you can get form the chart above, you can use ERFC(Z/SQRT(2))/2 in Excel, but it is rare that you would need this.

This all assumes that you know the man and standard deviation, and you only make one measurement. This is rarely the case, but it turns out that there are clever ways to get z-values in a variety of other situations. That's what we're going to learn now.

Likelihood when you know the probabilities:

Suppose you know that the probability of something happening is supposed to be p. You do an experiment N times and discover how many times x this something happens. Is the result consistent with the probability p?

The mean number of times you expect the event to happen, in N trials, is pN:

$$\boxed{\mu = pN}$$

However, it will not usually have exactly this many. It will generally differ from this by about one standard deviation, which is given by

$$\boxed{\sigma = \sqrt{Np(1-p)}}$$

The actual number you get is called x. When you want to know is whether x is close enough to μ that you shouldn't be surprised. To figure this out, you calculate z, using the formulas we had before.

$$z = \frac{|x - \mu|}{\sigma}$$

You then proceed as before, and conclude whether the results are consistent with chance or not.

For example, suppose you flip a fair coin 100 times. The probability of it coming out heads should be $p = 0.5$. Therefore, the number of times you should get heads should be about $\mu \pm \sigma = 50 \pm 5$. If you try it and get heads 70 times, this would result in a z-value of 4, which has a probability of about 3×10^{-5}. So if a coin comes out heads $x = 70$ times in $N = 100$ trials, it's probably not a fair coin.

Extracting and comparing probabilities:

Suppose you don't know what the probability is. You'd like to estimate it. All you know is that you did an experiment N times and got a certain result x times. What is the probability? Well, naively, the probability is just x/N, but what is the error in this estimate? The formula for these two quantities is[*]

$$\boxed{\hat{p} = \frac{x}{N}} \qquad \text{and} \qquad \boxed{\hat{\sigma}_p = \sqrt{\frac{x(N-x)}{N^3}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}$$

The probability, then, is simply given by $p = \hat{p} \pm \hat{\sigma}_p$.

Now, when you do an experiment of this type, it is most often useful to compare some group to a control group. For each of the two groups, we measure the probability

---

[*] For accurate work, these formulas should be modified to $\hat{p} = (x+1)/(N+2)$ and

$\hat{\sigma}_p = \sqrt{\hat{p}(1-\hat{p})/(N+3)}$, but if you are working with small enough x or N for this to make a difference, you probably should learn more statistics than you can learn here.

and the error in the probability, and then we want to know if they are different. Let's say our two groups have measured probabilities (with errors) of

$$\hat{p}_1 \pm \hat{\sigma}_1 \qquad \text{and} \qquad \hat{p}_2 \pm \hat{\sigma}_2$$

To analyze whether these numbers are the same or not, we simply use

$$z = \frac{\left| \hat{p}_1 - \hat{p}_2 \right|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$$

Small probabilities:

If you don't know the sample size N, but you do know how many x in that sample have some property, and you do know that N is much larger than x, then the expected number μ that you should get if you could do it repeatedly will be

$$\hat{\mu} = x \pm \sqrt{x}$$

To compare two groups to see if an increase in x is really significant or not, we simply use[*]

$$z = \frac{\left| x_1 - x_2 \right|}{\sqrt{x_1 + x_2}}.$$

---

[*] Again, these formulas are inaccurate if x_1 and x_2 are small. In these cases, $\hat{\mu} = (x+1) \pm \sqrt{x+1}$ and $z = \left| x_1 - x_2 \right| / \sqrt{x_1 + x_2 + 2}$.

|z|	Probability	|z|	Probability	Probability	|z|
0	50%	2.5	0.621%	50%	0.000
0.1	46.0%	2.6	0.466%	20%	0.842
0.2	42.1%	2.7	0.347%	10%	1.282
0.3	38.2%	2.8	0.256%	5%	1.645
0.4	34.5%	2.9	0.187%	2%	2.054
0.5	30.9%	3.0	0.135%	1%	2.326
0.6	27.4%	3.1	0.0968%	0.5%	2.576
0.7	24.2%	3.2	0.0687%	0.2%	2.878
0.8	21.2%	3.3	0.0483%	0.1%	3.090
0.9	18.4%	3.4	0.0337%	0.05%	3.291
1.0	15.9%	3.5	0.0233%	0.02%	3.540
1.1	13.6%	3.6	0.0159%	0.01%	3.719
1.2	11.5%	3.7	0.0108%	0.005%	3.891
1.3	9.68%	3.8	0.00723%	0.002%	4.108
1.4	8.08%	3.9	0.00481%	0.001%	4.265
1.5	6.68%	4.0	0.00317%	0.0005%	4.417
1.6	5.48%	4.1	0.00207%	0.0002%	4.611
1.7	4.46%	4.2	0.00133%	0.0001%	4.753
1.8	3.59%	4.3	0.00085%		
1.9	2.87%	4.4	0.00054%		
2.0	2.28%	4.5	0.00034%		
2.1	1.79%	4.6	0.00021%		
2.2	1.39%	4.7	0.00013%		
2.3	1.07%	4.8	0.00008%		
2.4	0.820%	4.9	0.00005%		
		5.0	0.00003%		