Summary statistics, distributions of sums and means

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Quantiles

In both empirical distributions and in the underlying distribution, it may help us to know the points where a given fraction of the distribution lies below (or above) that point. In particular:

- The 2.5% point
- The 5% point
- The 25% point (the *first quartile*)
- The 50% point (the *median*)
- The 75% point (the *third quartile*)
- The 95% point (or upper 5% point)
- The 97.5% point (or upper 2.5% point)

Note that if a distribution has a small fraction of very big values far out in one tail (such as the distributions of wealth of individuals or families), the may not be a good “typical” value; the median will do much better. (For a symmetric distribution the median is the mean).

The mean

The mean is the average of points. If the distribution is the theoretical one, it is called the *expectation*, it's the theoretical mean we would be expected to get if we drew infinitely many points from that distribution.

For a sample of points $x_1, x_2, \ldots, x_{100}$ the mean is simply their average

$$\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_{100}) \ / \ 100$$

For a distribution with possible values $0, 1, 2, 3, \ldots$ where value k has occurred a fraction f_k of the time, the mean weights each of these by the fraction of times it has occurred (then in effect divides by the sum of these fractions, which however is actually 1):

$$\bar{x} = 0 \times f_0 + 1 \times f_1 + 2 \times f_2 + \ldots$$

The expectation

For a distribution that has a density function $\mathrm{f(x)}$, we add up the value (x) times its probability of occurring, for the zillions of tiny vertical slices each of width dx which each has a fraction $\mathrm{f(x)\,dx}$ of all points. The result is the integral

$$\mathbb{E}[x] \;=\; \int_{-\infty}^{\infty} x\, f(x)\, dx$$

The expectations are known for the common distributions:

Binomial(N, p)	Np
Geometric(p)	$1/p$
Hypergeometric(N, M, n)	$n(M/N)$
Poisson(λ)	λ
Uniform(0,1)	1/2
Uniform(a,b)	(a+b)/2
Exponential(with rate λ)	$1/\lambda$
Normal($0, 1$)	0, of course

The variance

The most useful measure of spread of a distribution is gotten by taking each point, getting its deviation from the mean, which is $x - \bar{x}$, squaring that, and averaging those:

$$\left[(x_1 - \bar{x})^2 + (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_{100} - \bar{x})^2\right] / 100$$

The result is known as the *variance*.

Note these are squares of how far each point is from the mean. So they do not measure spread in the most straightforward way.

Of course there is an integral version for the true theoretical distribution as well, and we can talk of its variance.

A more unbiased estimate of the variance replaces n in the denominator by $n - 1$. This corrects for the fact that $\bar{x}$ is a bit too close to the n numbers as it is calculated from them.

The standard deviation

The *standard deviation* is the square root of the variance. This gets us back on the original scale. If a distribution is of how many milligrams something weighs, the variance is in units of “square milligrams” which is slightly weird. The standard deviation is just in milligrams.

Sometimes the ratio of the standard deviation to the mean is useful (especially if all values are positive, such as heights). This is called the *coefficient of variation*. Saying that our heights have a standard deviation of (say) 5% of the height conveys the variation in a meaningful way.

- Trick question: why not instead just compute the deviations of each point from the mean (which could be positive or negative) and average that?

The standard deviation

The *standard deviation* is the square root of the variance. This gets us back on the original scale. If a distribution is of how many milligrams something weighs, the variance is in units of “square milligrams” which is slightly weird. The standard deviation is just in milligrams.

Sometimes the ratio of the standard deviation to the mean is useful (especially if all values are positive, such as heights). This is called the *coefficient of variation*. Saying that our heights have a standard deviation of (say) 5% of the height conveys the variation in a meaningful way.

- Trick question: why not instead just compute the deviations of each point from the mean (which could be positive or negative) and average that?
- Less obvious question: why not average the absolute values of deviations from the mean? *(Because its mathematical properties are uglier – even though it seems simpler).*

Distribution of multiples of a variable

Just staring at the formula for calculating the mean you can immediately see that if you multiply all the sampled values (or all the theoretical; values) by, say, 3, the mean is 3 times as big as a result.

The same holds for any constant c, even including negative values and zero. It is also true for the expectation, as the constant comes outside of the integral and that proves this.

Means of sums of variables

When we draw a pair of points (x, y) where the x's come from one distribution and the y's from another, and we compute for each pair their sum, the mean of $x + y$ is simply the mean of the x's plus the mean of the mean of the y's.

That's true for samples of pairs, and its true for theoretical distribution of pairs *even if the two values are not independent*. In fact its true for sums of more quantities as well – the means just add up, and the expectations just add up.

Variances of sums of variables

It will be less obvious that variances also add up. In fact they don't really, but they do for theoretical distributions if the quantity `x` is drawn independently of the quantity `y`. In a sample the variances of sums aren't going to be exactly the sums of variances, even in that case.

If we add more (independent) variables, their variances add up. If they aren't independent they don't add up.

(We could test the additivity of the variances for independent draws using R).

Variances of multiples of variables

You can see from the formula for computing the variance that if we multiply all the values x by the same constant, say 7, since the mean is multiplied by 7 too, the squared deviations from the mean are all multiplied by $7^2 = 49$.

So variances are then multiplied by the square of that common multiple (in our case 7).

It follows that the standard deviation is just multiplied by 7, or whatever the quantity is.

Variances of means of variables

Put the preceding two slides together, and you get that (for sums of 12 independent draws from the same distribution) the variance of the mean of the 12 of them is $1/12$ of the variance of one of them.

This is true because:

- The mean is the sum of 12 draws from the distribution, divided by 12.
- The variance of a sum of 12 independent draws is 12 times the variance of the original distribution, and
- The mean is $1/12$ of that sum, so the variance of the mean is $1/(12^2) \times 12$ ...
- ... and that is $1/12$ of the variance of one draw.

Implication: the standard deviation of a mean of `n` independent draws is $1/\sqrt{n}$ as big as the standard deviation of the distribution from which they are drawn.

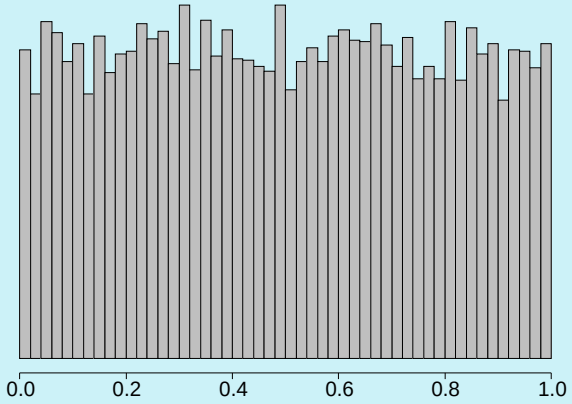
Standard deviation of sum of variables

If they're independent, it's the square root of the sum of their variances, so it's the square root of the sum of squares of their standard deviations.

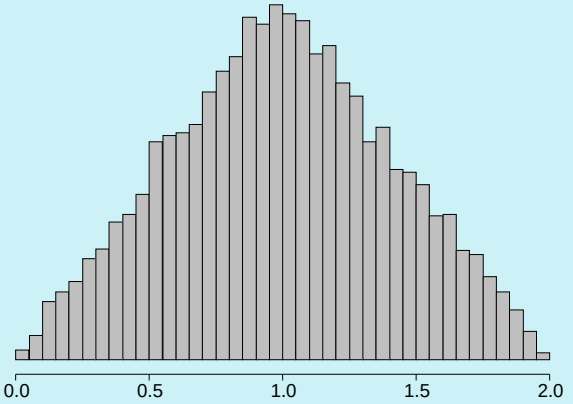
Sums of independent variables get more normal

There is a theorem (the Central Limit Theorem), provable under pretty general conditions, that for independent draws, sums of `n` quantities are more normally distributed that the original quantities are. This happens startlingly quickly:

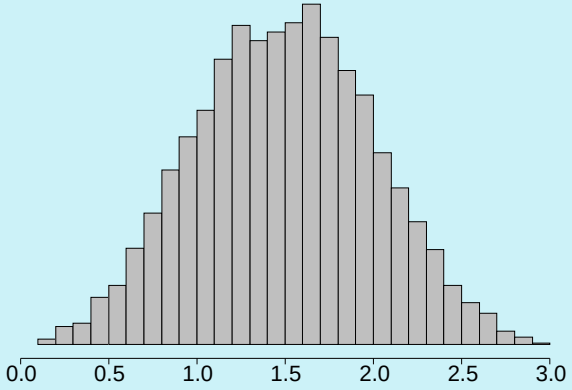
A uniform distribution



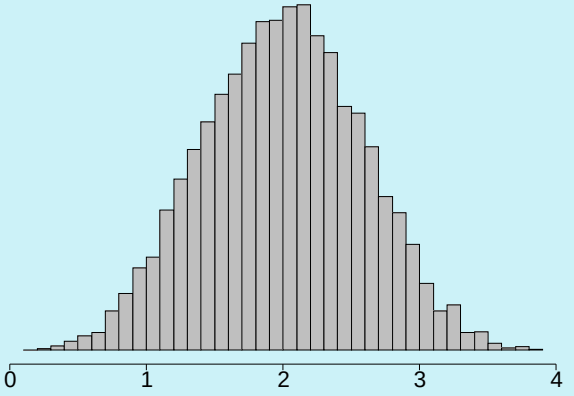
sum of 2 uniform variables



sum of 3



sum of 4



So do means

Means are just scaled versions of sums, so it works for them too: they are distributed nearly normally if there are more than a few.

If we take the mean weight of the next 100 vehicles that cross the I-5 bridge, and do this again and again, the numbers we get are going to be nearly normally distributed, even if the original distribution of vehicle weights is strange.

It's even true if the quantities aren't fully independent, provided there is enough lack of correlation of quantities far apart in the sum. (But there are some weird distributions that have "heavy tails" that wont have their averages become more normal).

What about the sum of independent Poissons?

Think about this one: if business calls arrive on a telephone line (say, to a telephone switching center) independently and a Poisson number comes in each hour (with expectation 8.2 calls), and also personal calls arrive in a Poisson process too, but with expectation 17.1 calls per hour ... what is the distribution of the total number of calls in an hour?

- If we just consider them as calls, ignoring business or personal, each little slice of time of width dt is expected to have $8.2\,\mathrm{dt} + 17.1\,\mathrm{dt}$ calls on average.

What about the sum of independent Poissons?

Think about this one: if business calls arrive on a telephone line (say, to a telephone switching center) independently and a Poisson number comes in each hour (with expectation 8.2 calls), and also personal calls arrive in a Poisson process too, but with expectation 17.1 calls per hour ... what is the distribution of the total number of calls in an hour?

- If we just consider them as calls, ignoring business or personal, each little slice of time of width $\mathtt{dt}$ is expected to have $8.2\,\mathtt{dt} + 17.1\,\mathtt{dt}$ calls on average.
- As $\mathtt{dt}$ is infinitesimally small that means they will have 0 or 1 call each with an overall rate of 25.3 calls per hour, and each little slice independent.

What about the sum of independent Poissons?

Think about this one: if business calls arrive on a telephone line (say, to a telephone switching center) independently and a Poisson number comes in each hour (with expectation 8.2 calls), and also personal calls arrive in a Poisson process too, but with expectation 17.1 calls per hour ... what is the distribution of the total number of calls in an hour?

- If we just consider them as calls, ignoring business or personal, each little slice of time of width dt is expected to have $8.2\,\mathrm{dt} + 17.1\,\mathrm{dt}$ calls on average.
- As dt is infinitesimally small that means they will have 0 or 1 call each with an overall rate of 25.3 calls per hour, and each little slice independent.
- ... in short, Poisson with mean 25.3.

... of independent binomials?

If we toss a coin 200 times, with heads probability p, the number of Heads is a draw from a binomial distribution.

Now if we make (say) 300 more tosses, what is the distribution of the total number of Heads over both sets?

I thought you would say that. Now is this true if the second set of tosses is with a different coin with a different probability of Heads? (Think about the extreme case where the first coin almost never can come up Heads, and the second one almost always does. Is that the same as having one coin with, say, 0.6 probability of heads?)

Variances of functions of random variables

If we know the variance of a quantity, what is the variance of (say) its logarithm? This may or may not be calculable exactly. But we can often make an approximation using the *delta method*. If

$$y = \ln(x)$$

then a small change in x will cause a (small) change in y that differs by a factor equal to the slope of y with respect to x.

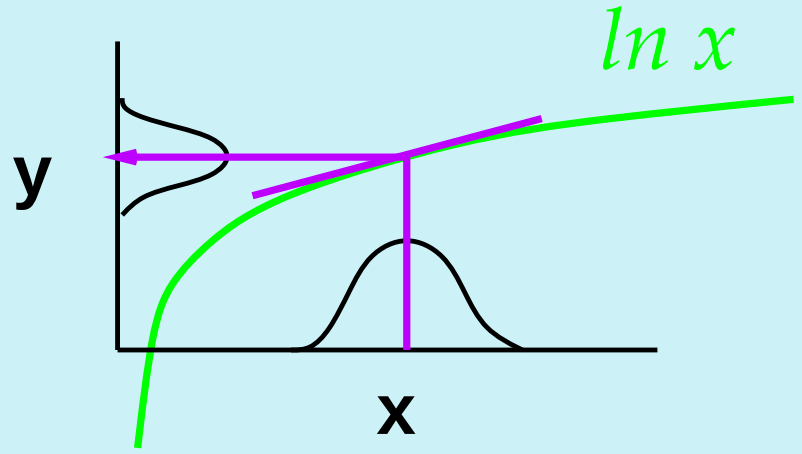
The square of the change in y will be the square of the change in x, multiplied by the square of the slope.

The delta method for the logarithm of a quantity

So for the case of $y = \ln(x)$,

$$\text{Variance}(y) = \text{Variance}(\ln x) \simeq \left(\frac{dy}{dx}\right)^2 \text{Variance}(x) = \left(\frac{1}{\bar{x}}\right)^2 \text{Variance}(x)$$

and we evaluate the derivative at the mean of x. So in that case the variance of $\ln x$ is (to good approximation, if the standard deviation of x is small enough) the variance of x, divided by the square of the mean of x. The logic is shown here:



(Note that one can easily see that the standard deviation of y is almost equal to the standard deviation of x, multiplied by the slope).