(Lane 3, 5)

wk 8 47

Module 3 of 3 – Probability & Statistics

Probability review.

- Probabilities are degrees of certainty $0 \le p \le 1$.
- We can calculate the probability of an event as $Pr(\text{event}) = \frac{|\text{success outcomes}|}{|\text{all outcomes}|}$, if all outcomes are equally likely.
- A conditional probability $Pr(A|B)$ is the probability of A given event B is true. Counting, this is $\frac{|\text{outcomes with A and B true}|}{|\text{outcomes with B true}|}$.

For example, Pr(6-sided die rolls 4) is $1/6$, but Pr(6-sided die rolls 4 | roll is even) is $1/3$ because only 3 outcomes satisfy event B true.

Also, when rolling 2 dice, not all sums are equally likely – we can't say all rolls 2–12 have the same likelihood of $1/11$, because the outcomes aren't equally likely to start. But each (a, b) outcome of (1st roll, 2nd roll) is equally likely, and that's way we can tell a 2 has $1/36$ probability while a 7 has $6/36 = 1/6$ probability $(1,6), (2,5), \ldots (6,1)$. (This fact will lead to many natural distributions where not all outcomes are equally likely – especially the Gaussian distribution.)

We can also have random variables that take on different values instead of being true or false – like a die roll, which takes on the values 1 to 6 with equal likelihood. If a random variable has a uniform distribution, then all outcomes are equally likely. But other distributions, which assign probabilities to outcomes, are possible; if X = \$ won from lottery, that's unlikely to be uniform.

With random variables, we can find expectations. $E[X]$ is the expected value of X. It's what the average value of X

would be if we averaged over a very large number of games.

The definition of $E[X]$, recall, is

$$\sum_{x} X \Pr(X)$$

or in other words, the sum over all outcomes of probability * outcome. We can think of this as a "probability-weighted average." If all outcomes are equally likely, it's just the average. But if they aren't, then we sum $p \cdot$ outcome instead of $\frac{1}{n} \cdot$ outcome to get the answer.

$$E[\text{20-sided die}] = \sum_{x=1}^{20} \frac{1}{20} \cdot X = \frac{1}{20} \sum_{x=1}^{20} X$$

= average roll = 10.5

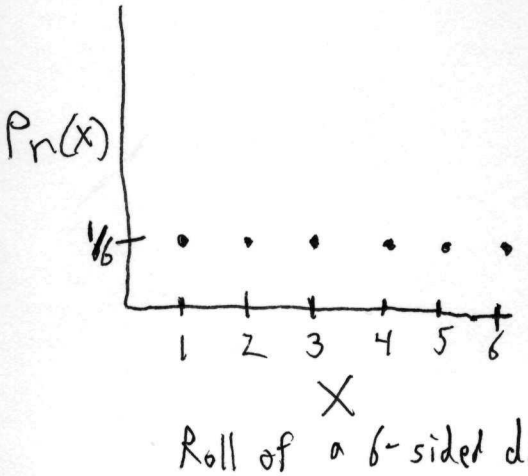
Raffle ticket costs \$5, $\frac{1}{100}$ chance of \$100

X = net winnings

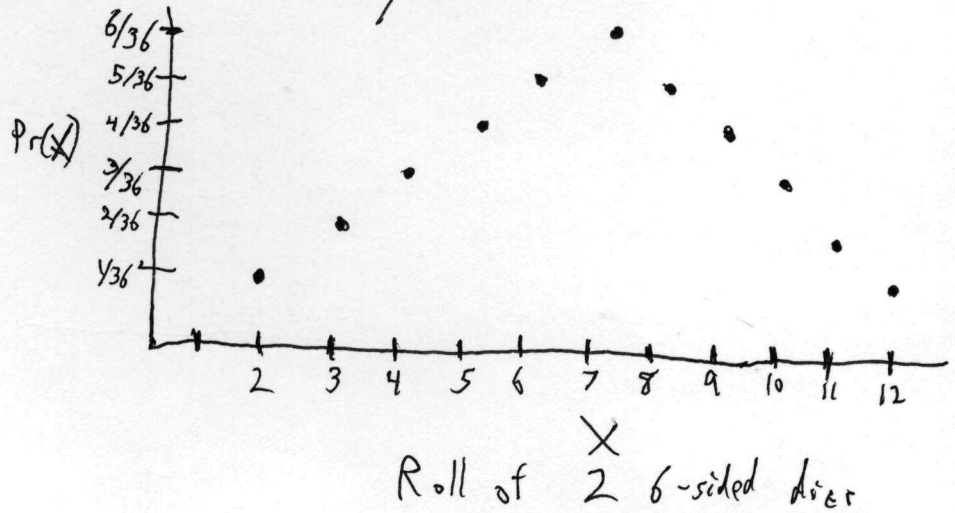
$$E[X] = \frac{99}{100} \cdot -5 + \frac{1}{100} \cdot 95 = -\frac{400}{100} = -\$4$$

We'll bring up other reminders as needed.

We can visualize distributions by graphing outcomes on the x-axis, and probabilities on the y-axis.



Roll of a 6-sided die



Roll of 2 6-sided dice

Something is clearly different here — on the left, there's no particular tendency to give us the expected value, except on average in the long run. But on the right, a 7 is more likely than an 8 even on just one trial. If we somehow "bet" on a non-extreme result, it's a safer bet on the right.

The number that characterizes how much of a tendency there is in the data is called the variance of the distribution.

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right]$$

This is the expected value of the squared difference between the result and expectation. (As with least squares regression we assume we care more about big differences than little ones.) You may sometimes in statistical settings see $E(X)$ written μ if it's just the average value, which is true in some common distributions. Thus:

$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right].$$

to stray from the mean

The square root of the variance is called the "standard deviation" and sometimes the greek letter σ = sigma ("s") is used to denote the standard deviation.

So $\sigma^2 = E\left[(X - E[X])^2\right]$ or $\sigma = \sqrt{E\left[(X - E[X])^2\right]}$.

As with variance, a big σ indicates a spread-out distribution.

Let's take as examples the 1-die and two-die cases, respectively. These have $E[X] = 3.5$ and $E[X+Y] = 7$,

Variance for 1 die is $\frac{1}{6}(-2.5)^2 + \frac{1}{6}(-1.5)^2 + \frac{1}{6}(-0.5)^2 + \frac{1}{6}(0.5)^2 + \frac{1}{6}(1.5)^2 + \frac{1}{6}(2.5)^2$

$= \frac{1}{6}(6.25 + 2.25 + 0.25) * 2 =$ 2.916

So the expected value of the square of the difference from the mean is 2.92. The standard deviation is thus 1.71 or so, and tells us roughly how far from the mean we can expect to be.

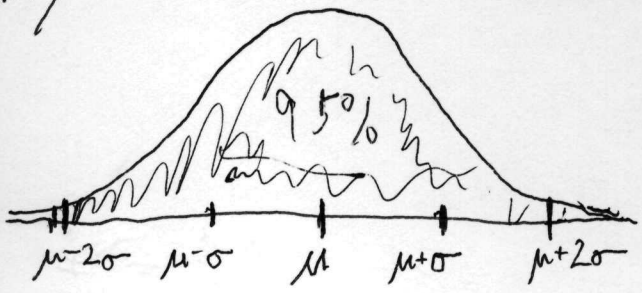
Variance for 2 dice is $\frac{1}{36}(-5)^2 + \frac{2}{36}(-4)^2 + \frac{3}{36}(-3)^2 + \cdots + \frac{6}{36}(0)^2 + \frac{5}{36}(1)^2 + \cdots + \frac{1}{36}(5)^2$

$= 5\frac{5}{6}$ or 5.83.

When we add the results of independent events, the variance sums as well. We can see this here: each die roll had variance 2.916, so the variance of the sum is $2.916 \cdot 2 = 5.83$.

Notice that the variance doesn't really describe the shape per se: the distribution for 2 dice has higher variance despite looking more concentrated (compared to 1 die). The variance just describes the size of differences, so distributions across wider ranges, like 2–12 instead of 1–6, will naturally tend toward higher numbers. If we kept the range the same, then lower variance would tend to imply more concentration toward the mean.

For the very common "normal" or "Gaussian" distribution, standard deviations have a useful rule of thumb associated with them: there is a 95% chance that a sample will fall within 2 standard deviations to either side of the mean. (1.96 to be more precise) Or alternately we can expect 95% of all samples to fall in this range. This is specifically for normal/Gaussian distributions, which we'll cover in more detail later, but it's the most common way to encounter standard deviations.



One last variance fact: $Var(aX) = a^2 Var(X)$. If the original variable is multiplied by a that affects the variance by a factor of a^2. This will come up in some derivations later.

The Binomial Distribution

Some "named" distributions come about because they result naturally from doing some particular process or experiment repeatedly. The binomial distribution is a distribution on the number of coin flips that come up heads after flipping n of them. The coins could

also be "biased" and have a probability of success p instead of being 0.5 for sure. (Flipping biased coins is often used as a model or metaphor for any kind of process with a fixed probability of success.)

You may recall from Discrete Structures that the number of ways to get k heads on n flips is $\binom{n}{k}$ since you're choosing which flips are heads. The probability of getting k successes is the sum over all these sequences of their probabilities. The probability of each <u>specific</u> sequence with the right number of heads is $p^k(1-p)^{n-k}$ — each head in the sequence had probability p, each tails had probability $1-p$, and if we multiply out these terms over all n symbols, we get the above.

Example: $p = 3/4$, $k = 2$

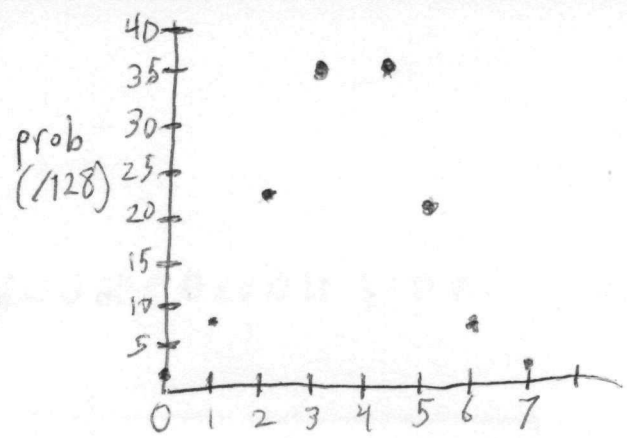
Probability of HTTH: $3/4 \cdot 1/4 \cdot 1/4 \cdot 3/4$

Probability of TTHH: $1/4 \cdot 1/4 \cdot 3/4 \cdot 3/4$ same terms, just rearranged

When we ask "how many sequences have the right k", it's $\binom{n}{k}$, and they'all have the same probability of $p^k(1-p)^{n-k}$, so the total probability of k successes is $\binom{n}{k}p^k(1-p)^k$.

What does this look like? If the coin isn't biased, and $p = (1-p) = 0.5$, then this is $\binom{n}{k}\left(\frac{1}{2}\right)^n$. The overall shape of the distribution can be read off Pascal's triangle, which has the values for $\binom{n}{k}$ in row n.

row 0 → 1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
row 7 → 1 7 21 35 35 21 7 1



This is starting to look a little like a Gaussian or normal, and that's no accident: as any independent random variables are added, the distribution of the sum of their outcomes will start to look more and more Gaussian. (Which also means plotting Pascal's triangle values starts to look more and more Gaussian.)

Two properties that we'll often want to be curious about for a distribution are its mean and its variance.

Mean: What *is* the expected number of successes for n flips with probability p? Intuitively, this should be np. For example, if $p = 0.6$ and $n = 10$, we'd expect 6 successes. This is correct and one way to show it is with linearity of expectation. $E[X]$ for one flip is 0.6, so

$$E[X_1 + X_2 + \cdots + X_{10}] = E[X_1] + E[X_2] + \cdots + E[X_{10}] = 6.$$

To find the variance we can use a similar trick; recall $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if the two are independent. So we really just need to find the variance of one flip.

$$\underbrace{(1-p)}_{\text{fail prob}}\underbrace{(-p)^2}_{\substack{\text{fail is} \\ \text{off by} \\ -p}} + \underbrace{p}_{\text{success prob}}\underbrace{(1-p)^2}_{\substack{\text{success} \\ \text{off by } (1-p)}} = p^2 - p^3 + p^3 - 2p^2 + p = p - p^2 = p(1-p)$$

So by our variance rule, the overall variance must be $np(1-p)$.

The Poisson Distribution

A different distribution that is sometimes used for modeling is the Poisson distribution, which captures phenomena as varied as how many stars can be seen in a patch of night sky, or how many bombs hit a particular building, or how many typos can be found on a page in a book. In general, the Poisson should be applied when:

1) We are counting events within a fixed span of time or volume
2) The last success doesn't tell us anything about when the next will occur
3) We know an overall rate for events within the given time interval!

The rate of successes is also the expectation here; for example, if we think there's an average of 1 typo per page, then the expected number of typos on the page is 1. This parameter is often given a name λ.

The distribution, then, is $Pr(k \text{ events}) = e^{-\lambda} \frac{\lambda^k}{k!}$

Notice that there's nonzero probability of an unlimited number of values of k, though the probability goes down quite rapidly.

Example: Suppose we get 1 snow day per spring semester on average, Poisson distributed. $\lambda = 1$.

$Pr(0 \text{ snow days}) = e^{-1} \frac{1}{1} = 0.37$

$Pr(1 \text{ snow day}) = e^{-1} \frac{1}{1} = 0.37$

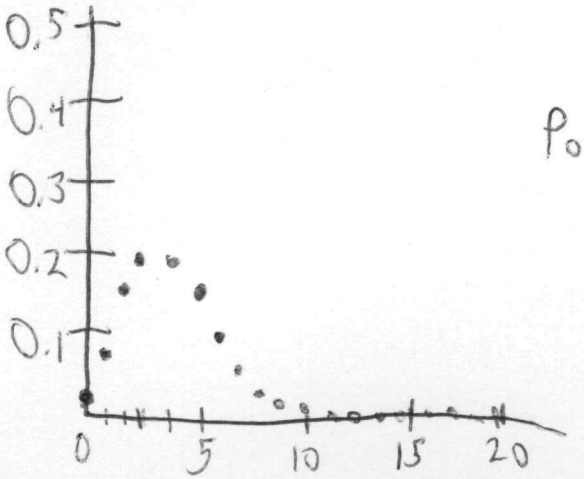
$Pr(2 \text{ snow days}) = e^{-1} \frac{1}{2} = 0.18$

$Pr(3 \text{ snow days}) = e^{-1} \frac{1}{6} = 0.06$

⋮

The variance for a Poisson dist. is λ, too.

We can derive the Poisson by taking the binomial distribution and letting n approach infinity, so that we allow ourselves an infinite number of opportunities, while allowing p to shrink to be relatively small, so that the expected number of successes with these infinite opportunities is still finite, But we won't show the derivation here.



Poisson, $\lambda = 4$

Overall, the Poisson looks like a skewed Binomial or Gaussian distribution - but it remains nonzero going out to infinity (unlike the binomial) and has a hard stop at 0 (unlike the Gaussian). Also unlike the Gaussian, it's discrete - There's no chance of 1.4 events, for example (though λ can be any positive real).

The Gaussian is more well-known in the social sciences, partly because it's useful for A/B sorts of tests for figuring out whether populations differ. But if you want a model for random events happening in time or space, like the number of customers who will use your website in an hour or the number of bug reports filed in a year, the Poisson is the right distribution.

-end wk 8-

The Normal Distribution

Now we'll get to what might be the most well-known probability distribution ~ the "normal" distribution, also known as the "Gaussian" or in the popular press as a "bell curve". This distribution comes about in nature quite a bit, partly because it is the shape you get in the limit when many independent variables are summed. So, for any variable like height that tends to be the result of many contributing factors, the distribution of the final value will tend to look like a bell curve, as each little factor is summed. (That's if the variables are independent; a variable like wealth will tend to look different if, for example, wealth builds on itself in a "takes money to make money" sort of way.)

In particular, measurement error is often well-modeled as a normal distribution - error being the sum of many little factors that don't have anything to do with each other. Systems like radar and sonar systems will try to smooth their error-filled measurements over time using the assumption that the raw sensor readings were corrupted with Gaussian noise.

The mathematical form of the Gaussian, which was originally derived as an approximation to the binomial, is

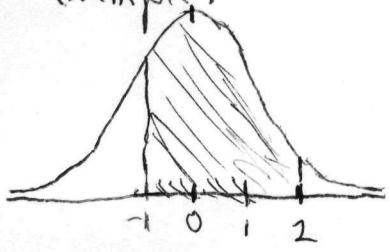
$$PDF(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF: "probability density function"
μ: mean
σ: standard deviation

As a computer scientist, you probably won't need to interact much with this equation directly (except perhaps to follow a proof). Most languages that are in common use have a function that will return this value for you, given the mean and standard deviation. You can draw a sample at random from one of these distributions with another function, and use a still different one to find the likelihood that a sample lies in a particular range.

For example, numpy.random.normal $(\mu, \sigma, \#)$ generates samples from a normal distribution, and a scipy.stats.norm object, representing a normal random variable, has a pdf() method, as well as cdf() (see below).

The CDF is the "cumulative distribution function", returning the integral of the PDF from $-\infty$ to a given point. The probability that a value falls between two points in the distribution is given by the integral of the PDF between those two points, and rather than integrating, we can use the precomputed CDF values. For example:

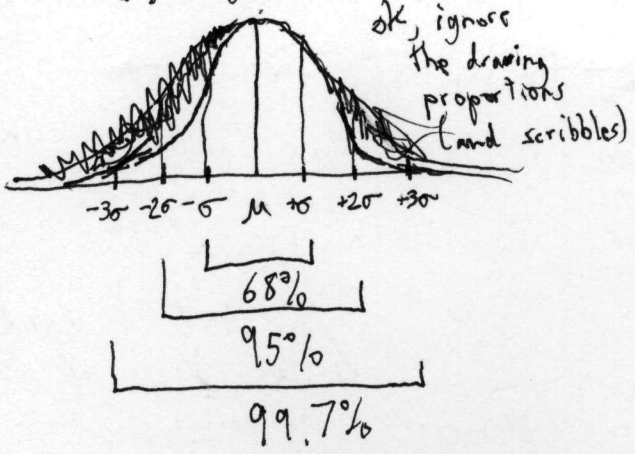


Here $\mu=0$ and $\sigma=1$. Mathematically, the probability a random point will fall in the range $[-1, 2]$ is $\int_{-1}^{2} PDF(x)$ where $PDF(x)$ is the equation given before. But the way we could get this in code would be

norm.cdf(2) - norm.cdf(-1)

For a norm object with $\mu=0$ and $\sigma=1$, this gives a correct probability of 0.816.

In general, 68% of the area under a normal distribution is within one standard deviation of the mean, and 95% is within two standard deviations. Three standard deviations captures all but 0.3%



Law of Large Numbers and Central Limit Theorem

Statistical tests and AI methods assume that things are normally distributed very often, so it's good to have a clear grasp of why we often make that assumption. While people often assume normality when it isn't technically normal, and nothing terrible happens, knowing the theoretical reasoning can help you figure out whether your model is on solid ground.

The law of large numbers states that, in the limit as the number of samples of random variables with a shared expected value μ goes to infinity, then the mean will always approach the expected value. That is, with enough trials, averaging should get us arbitrarily close to the actual expected value. The variables should be "iid," or "independent and identically distributed" to work but the theorem doesn't require any particular distribution that the variables have in common (not necessarily normal, for example, or uniform). You sometimes hear the Law of Large Numbers casually

invoked to mean that losses will eventually be offset by gains; that's basically true, that deviation from the average eventually reverts itself, but the keyword is "eventually." It's also sometimes used to mean that distributions become normal with enough variables/samples, but that's mostly the central limit theorem, saying that the [redacted] mean tends toward the expectation is just one piece of that.

The central limit theorem is stronger, stating that a sum of iid variables will eventually tend toward a normal distribution. More precisely, the distribution of $\bar{X}_n$, the mean of the first n samples, approaches a normal with mean μ and variance σ^2/n, if [redacted]σ^2 was the variance of each [redacted] variable; the [redacted] sum is then distributed like mean $n\mu$ and variance $n\sigma^2$.

(* assuming finite mean and nonzero finite variance.)

Again, none of the individual variables need to have any particular distribution (such as normal). It just needs to be "iid" independent but identically distributed. And even that requirement of "identically distributed" may not be strictly necessary - there are variations that relax this requirement. If we want to model a variable as having been affected by a bunch of little nudges over time that are roughly independent and roughly identically distributed, the normal makes sense as a model.

Sampling Error

In the last section, we mentioned that the variance in the sampling distribution of the mean is σ^2/N. (We don't really need to have large N for that to be true; the sum of N variables with variance [redacted]σ^2 has variance $N\sigma^2$, and then the mean is a result by $1/N$, hence $\frac{N\sigma^2}{N^2} = \frac{\sigma}{N}$.) If we've performed an experiment where we

took an average, we have some uncertainty about whether our empirical (measured) mean matches the true mean in the wild. But, our variance of σ^2/N combined with the knowledge that means are always normally distributed, can give us an idea as to how far off our sample mean is from the true mean.

The "standard error" of the mean is equal to one standard deviation of its sampling distribution. Since we know the variance is σ^2/N, that makes one standard deviation $\sigma/\sqrt{N}$. The standard error is sometimes used to produce error bars on measured quantities — though, notice you'd need a range of $\overline{X}_N \pm 2\sigma/\sqrt{N}$ to be 95% certain the true mean lies within that interval.

The standard error may take a different form if we're trying to measure something that isn't a mean — for example, the difference between two means. Suppose we measure two populations' means on some measure in the wild — height of Americans versus Canadians, for example — and want to be certain the difference is not 0. We can subtract one mean from the other, but what is our range of uncertainty? As it turns out $\mathrm{Var}(X-Y) =$ [illegible] $\mathrm{Var}(X) + \mathrm{Var}(Y)$, that is, the variance of differences works the same as the variance of sums. We have two uncertain quantities, and that just increases the uncertainty over either alone. So if the variance of each mean is $\frac{\sigma_X^2}{N_1}$ and $\frac{\sigma_Y^2}{N_2}$, the variance of the differences is

$$\frac{\sigma_X^2}{N_1} + \frac{\sigma_Y^2}{N_2},$$

and the "standard error," one standard deviation, for the difference is $\sqrt{\frac{\sigma_X^2}{N_1} + \frac{\sigma_Y^2}{N_2}}$.

(The difference itself has an expectation that is just $\mu_1 - \mu_2$.)

So, for example, let's say we measured the Americans at 175.3 cm (5'9") and Canadians at 175.1 cm (also about 5'9") with 100 people in each sample, and a variance of 10 cm (4 in). What's our range of 95% likely differences?

The standard error in the difference is $\sqrt{\frac{(10)^2}{100} + \frac{(10)^2}{100}}$ $= \sqrt{1+1} = \sqrt{2}$, or 1.41 cm. So we're 95% confident the difference is in the range $[0.2 - 1.41*2, 0.2 + 1.41*2]$ $= [-2.628, 3.028]$. That doesn't look promising for the hypothesis that Americans are generally taller than Canadians, since 0 is in that range. What is the chance from this data that Americans really are taller? That would be $\int_0^\infty PDF(x)$ for the Gaussian with $\mu = 0.2$, $\sigma = \sqrt{2}$ which comes out to ~58%. More likely than not, but not great to bet on.

– end wk 9 –

(note that the above example assumes we somehow magically know the true variance – see next lecture for an issue arising from needing to estimate it)

Degrees of Freedom, Bias, & Bessel's correction

Bias in an estimation process is a tendency to produce results that are off in one way or another from the true result. It's contrasted with variability, which is just a tendency to be inconsistent. If our yardstick has the feet markings too far apart, that'll produce a bias toward shorter heights from the true height, but it will be very consistent. On the other hand, if we took measurements on a trampoline, we'd get readings all over the place, but we wouldn't necessarily expect a bias.

It might be surprising, but I've effectively already taught a method with bias - at least, you would probably produce biased results if you tried to estimate the variance from samples, given what you know now. Suppose we want to apply this formula:

$$Var(X) = E[(X-\mu)^2]$$

If we have N samples, it seems like I should be able to average them to estimate μ - call the average $\overline{X}$ - and then find the expected deviation by summing over $(X-\overline{X})^2$ and dividing by N (thus assuming each sample value was equally likely):

$$\text{Biased } \sigma^2 = \frac{\Sigma(X-\overline{X})^2}{N}$$

This seems reasonably straightforward, so where's the bias? This method will generally underestimate the true variance, thus throwing off our idea of the standard deviation size and so forth. The reason is that we've behaved as if we have N independent samples after taking the mean - but the mean itself was computed from those samples. We have lost a degree of freedom in the data. The Nth data point won't be a surprise if we know the mean $\overline{X}$ and the previous $N-1$ points.

When computing an expectation over our data points in statistics, we'll often divide by the degrees of freedom instead of N. This will counter the bias that would be otherwise introduced by letting each point contribute "a little more than once" to the metric.

The easiest derivation for this unbiased estimator doesn't rely on the degrees of freedom idea, but does show in a different way how we missed a source of variance.

We can think of a random variable Z as $Z = X + Y$, where X is a random variable giving our sample mean, and Y is a random variable giving the difference from our sample mean. $Var(Y)$ is our biased estimator of the variance, but the uncaptured variance is the variance of the sample of the mean, $\frac{\sigma^2}{n}$.

So $\sigma^2 = \frac{\sigma^2}{n} + Var(Y)$, using the rule that variances sum.

Rearranging terms gives $Var(Y) = \sigma^2\left(1 - \frac{1}{n}\right) = \sigma^2\left(\frac{n-1}{n}\right)$

which means to find the true variance σ^2, we multiply by $\left(\frac{n}{n-1}\right)$. ↑ "Bessel's correction"

So $\sigma^2_{unbiased} = \frac{\Sigma(X-\bar{X})^2}{N-1}$.

Software doing statistics will generally do the unbiased calculation. You'd only ever need the other if you knew for a fact what the true mean was, and didn't need a sample mean.

In general, the statistics community seems more concerned with avoiding bias than the ML community, though this is likely to change. Lots of the discussion that does happen has focused on good sampling methods that represent more diverse populations. But even in simple calculations, bias can be introduced if you aren't careful.

Degrees of freedom, linear algebra style:

$$\text{data} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \bar{x} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$$

mean (1 df) residuals (must sum to 0) (N-1 df)

The degrees of freedom are the dimension of the space of vectors of residuals, which has one constraint equation, $a + b + \ldots + z = 0$ infinite sols, $n-1$ dims

Confidence Intervals and the t-distribution

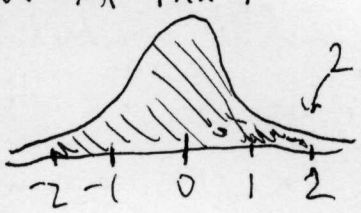
If we knew the true variance for the distribution of the mean estimate, then we could figure out a 95% confidence interval relatively easily; calculate the variance of the sample of the mean $\frac{\sigma^2}{N}$, figure a standard deviation from that $\frac{\sigma}{\sqrt{N}}$, and assume 95% likelihood that the true mean is within 2 of those (technically 1.96) from our estimated mean. (Recall that the central limit theorem guarantees our estimate of the mean is Gaussian, even if the underlying distribution is not.)

But typically we don't really know the variance, and must estimate it from data using fewer degrees of freedom than N because of the aforementioned bias issues. There's a distribution that is designed for this very common case, but it's not the normal distribution; it's the t-distribution.

First, let's go back to creating confidence intervals with a normal distribution. The distribution of

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is normal: the numerator is the difference between our measured mean and the true mean, and the denominator is the standard error. Dividing by the bottom effectively converts our units to standard deviations, so if the value is -1, we got a value 1 standard error to the left of the mean.

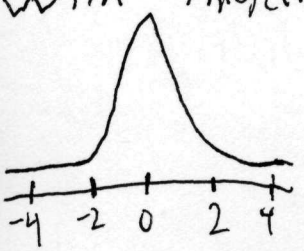


We already know that we're 95% likely to be within ± 1.96 standard errors, so this doesn't tell us much.

The t-distribution is similar to this, but takes into account the uncertainty in our estimate of the variance:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \quad \text{where} \quad S = \frac{\Sigma(X-\bar{X})^2}{N-1} \leftarrow \text{degrees of freedom}$$

It's taking into account that bias. That means that as N gets large, we should expect the t-distribution to look more and more like a normal distribution, because the difference between N and $N-1$ doesn't matter much for big N. But in the meantime, for the values of N that tend to be used in scientific experiments, there's a noticeable difference from the normal, with thicker tails.



A parameter to the t-distribution is the degrees of freedom ($N-1$ if N is the number of data points). This determines how normal-ish vs thick-tailed the distribution is and it feeds into the distribution exactly where the Bessel-correction for bias goes.

The use of the t-distribution, then is two fold. First if we don't know the true variance we should use it to calculate 95% confidence intervals. The traditional way to do this is to look up your degrees of freedom and desired confidence level in a table; it'll tell you how many standard errors you need to have the desired confidence. For example for 20 degrees of freedom ($N=21$), I find 95% confidence is achieved with 2.086 standard errors, so if my mean was 1 and my standard error was 2, my 95% confidence interval is $1 \pm 2*2.086 \approx [-3, 5]$. (In practice, you'll probably call functions to get these numbers but you'll need to know what the degrees of freedom are and what the t-distribution is telling you.)

Second, we can use the t-distribution to figure out whether two means really are likely to be different — a process known as a "t-test." We'll cover statistical significance soon, but first we can look at how to calculate confidence intervals for differences between populations. After all, if the 95% confidence interval doesn't include "0," we can have some confidence something is going on.

Recall that when we subtract $X-Y$ then $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$. So when we're trying to estimate a standard error for a difference of <u>two</u> variables, we can compute

$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{N_1} + \frac{\sigma_Y^2}{N_2}}$ which, if we assume $N_1 = N_2 = N$ and $\sigma_X = \sigma_Y$, becomes $\sqrt{\frac{2\sigma^2}{N}}$

We can estimate σ^2 by averaging the individual variance estimates. Then we just need degrees of freedom for the t-distribution; the individual variance estimates had $N-1$ each, so $2(N-1)$ degrees of freedom in all.

So if we had two groups with 21 people each, and one group got an average of 10 points on a test with variance estimate 2, and the other got 12 points with variance estimate 1 our overall variance estimate is 1.5 the standard error is $\sqrt{\frac{3}{21}} = \sqrt{\frac{1}{7}}$, and the t-distribution 95% confidence is at 2.021 since the df are $20+2=40$. The confidence interval is $2 \pm \sqrt{\frac{1}{7}} \cdot 2.021$ and we can be reasonably certain there's a difference of more than one point.

$= 2 \pm 0.764 = [1.24, 2.764]$

—end wk 10—