SCM Lecture 8 - teaching notes

Jeanne Wilson

October 22, 2010

1 Gaussian Probability Intervals

We can ask:
What is the probability that a value lies within 1σ of the mean? - ie. $|\frac{x-\mu}{\sigma}| < 1$
We could work this out by integrating the area under the gaussian:

$$P\left(|\frac{x-\mu}{\sigma}| < 1\right) = \int_{-\sigma}^{+\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{1}$$

But the gaussian distribution approximates many different naturally occurring distributions and because it is so useful you can actually look up gaussian probability tables in standard text books - for example *Barlow*.
Show slide
The way you read these tables is to calculate the value $|\frac{x-\mu}{\sigma}|$ - the distance away from the mean in terms of the standard deviation. And then use the left hand and top columns to find the right value. The left hand column is in tenths of a sigma, and the top column is in hundredths of a sigma, so for example, if I wanted the probability of being within $\pm 0.76\sigma$ of the mean, I would locate this value here. If I wanted the probabililty of being within ± 0.765 of the mean I would have to interpolate between two values.

So if we look up $|\frac{x-\mu}{\sigma}| = 1$ we see the probability is 68.27% ie. the area under the curve between $\mu \pm \sigma$ is $\approx$ 68%. This is what we usually use as the most probable interval and what we assume you mean when you quote a value as $A \pm dA$. ie. That there is 68% chance that a value lies within one quoted error of the true mean.

Looking up $|\frac{x-\mu}{\sigma}| = 2$ we find $\approx$95%, and $\frac{x-\mu}{\sigma} = 3$ gives $\approx$99%.
For example we measure a length to be 63 ± 1 cm.

- There is 68% probability that the true value is within 62–64 cm
- There is 95% probability that the true value is within 61–65 cm
- There is 99% probability that the true value is within 60–66 cm

1.1 One tailed integrals

Sometimes we are only interested in *one sided probabilities*. For example, I am trying to place a limit on some quantity. I'll use an example from my research. We are looking for very rare types of particle decay - processes that have never yet been observed but we know they have half-lives greater than 10^{20} years. The fact that we have never observed the decay doesn't mean it doesn't happen but we can calculate a lower-limit for the half-life.

So we observe a certain amount of isotope for a certain period of time and see no decays and from that we calculate a value for the shortest possible half-life consistent with our data along with the uncertainty on that value:

$$T_{1/2}^{\text{limit}} = (9.3 \pm 1) \times 10^{22} \text{ years}$$

But in this case the upper error doesn't really mean anything - we are saying that $T_{1/2}$ is bigger than our measurement but we are not making any claims about how much bigger - this is just a limit and it would be more useful to tell people the limit and our confidence in it - we want to say that the half-life is greater than X with 99% confidence. ie. There is only 1% chance that the true value is less than this. So we look in a *one tailed integral* table for 99% confidence level and find that relates to $\approx 2.3\sigma$ so we quote

$$T_{1/2} > 7 \times 10^{22} \text{ years (99\% C.L.)}$$

Some notes of caution when using these tables.

- Remember that the Gaussain distribution is symmetric. So if you want the probability of being within $+1\sigma$ of the mean - ie in the range $\mu - \mu + \sigma$ you can take the two sided table and simple divide the value by 2.
- Ask yourself whether you are interested in the centre of the distribution, or the tails. For example, if I want to reject only the 5% of values in the tails of a distribution I need to look up the limits from the two sided table for $\frac{x-\mu}{\sigma} = 0.975$, not $\frac{x-\mu}{\sigma} = 0.025$!

2 Central Limit Theorem

The sum of a large number of independent observations from the same probability distribution (not necessarily Gaussian) has approximately a Gaussian distribution.

So far we have considered a number of different distribution functions:

- Binomial
- Poisson
- Gaussian
- Uniform

They all define the probability of one quantity, X, in event space.
Now we consider 2 variables - lets call them X, which has probability $P_X(X)$ and Y, which has probability $P_Y(Y)$.
Now $Z = X + Y$ - what is $P_Z(Z)$?
Well - that is 2 observation, that hardly satisfies my definition of a "large number of independent observations". Lets extend that to the general situation: What if $Z = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + ...a_NX_X$, where N is large - what is $P_Z(Z)$?
The central limit theorem tells us that for $N \to \infty$ $P_Z(Z)$ is a Gaussian.

Lets look at some examples. *slide 1 - sum of two dice.* If I throw a die there is 1/6 probability of getting each number 1–6, ie, uniform across the allowed values. If I throw two dice, the sum can be in the range 2–12 but the probability distribution for the sum value isn't uniform - it approximates to a Gaussian.

slide 2 - 3 uniform distributions Consider three uniform distributions - same thing. *2 more example slides* Just to prove - it works for non uniform distributions too. N doesn't have to be very large - Do 1, 2, 5, 10, 100.

Remember in the last lecture - I told you that in the limit of large N, the binomial distribution approximates to a Gaussian and in the limit of large λ the Poisson distribution approximates to a Gaussian. This makes sense in the context of Central Limit Theorem - break up your Poison or Binomial into many sub samples - sum them all together and central limit tells you the result should be Gaussian.

3 Expectation Value of the Sum of Random Variables

Lets go back to our 2 variables example - $Z = X + Y$ - N is not large so Z isn't Gaussian but we want to know the expectation value of Z.

$$E[Z] = \int_{-\infty}^{+\infty} (x + y) f(x, y) dx dy$$

If x and y are independent we can write $f(x, y) = h(x)g(y)$ leading to

$$\begin{aligned} E[Z] &= \int_{-\infty}^{+\infty} (x + y) h(x) g(y) dx dy \\ &= \int_{-\infty}^{+\infty} x h(x) dx . g(y) dy + \int_{-\infty}^{+\infty} y g(y) dy . h(x) dx \end{aligned}$$

and

$$\begin{aligned} \int_{-\infty}^{+\infty} h(x) dx &= 1 \\ \int_{-\infty}^{+\infty} g(y) dy &= 1 \end{aligned}$$

so

$$\begin{aligned} E[Z] &= \int_{-\infty}^{+\infty} xh(x)dx + \int_{-\infty}^{+\infty} yg(y)dy \\ E[Z] &= E[X] + E[Y] \end{aligned}$$

In the more general case, for any number of independent terms:

$$Z = \sum_{i=1}^{N} X_i \quad \rightarrow \quad E[Z] = \sum_{i=1}^{N} E[X_i]$$

4 Variance of the Sum of Random Variables

What about the variance then

$$\text{var}[Z] = E[Z^2] - (E[Z])^2)$$

We've just worked out E[Z] so taking the first term.

$$\begin{aligned} E[Z^2] &= \int_{-\infty}^{+\infty} (x+y)^2 h(x)g(y)dxdy \\ E[Z^2] &= \int_{-\infty}^{+\infty} (x^2+y^2+2xy)h(x)g(y)dxdy \\ E[Z^2] &= \int_{-\infty}^{+\infty} (x^2)hxdx + \int_{-\infty}^{+\infty} y^2 g(y)dy + 2\int_{-\infty}^{+\infty} xh(x)dx \int_{-\infty}^{+\infty} yg(y)dy \\ E[Z^2] &= E[X^2] + E[Y^2] + 2E[X]E[Y] \end{aligned}$$

we've seen $E[X^2]$ before in

$$\begin{aligned} \text{var}[X] &= E[X^2] - (E[X])^2 \\ E[X^2] &= \text{var}[X] + (E[X])^2 \end{aligned}$$

so plugging back in $E[Z^2]$:

$$E[Z^2] = \text{var}[X] + (E[X])^2 + \text{var}[Y] + (E[Y])^2 + 2E[X]E[Y]$$

And plugging this back into var[Z]:

$$\begin{aligned} \text{var}[Z] &= \text{var}[X] + (E[X])^2 + \text{var}[Y] + (E[Y])^2 + 2E[X]E[Y] - (E[X] + E[Y])^2 \\ \text{var}[Z] &= \text{var}[X] + (E[X])^2 + \text{var}[Y] + (E[Y])^2 + 2E[X]E[Y] - (E[X])^2 - (E[Y])^2 - 2E[X]E[Y] \\ \text{var}[Z] &= \text{var}[X] + \text{var}[Y] \end{aligned}$$

Does this surprise us? It shouldn't because we know that for uncorrelated variables we add the uncertainties in quadrature and variance = standard deviation squared.

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

5 Expectation and Variance of the Mean

What is the expectation of a mean value.

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ E[\bar{x}] &= \frac{1}{N} E\left[\sum_{i=1}^{N} x_i\right] \\ E[\bar{x}] &= \frac{1}{N}\sum_{i=1}^{N} E\left[x_i\right] \\ E[\bar{x}] &= \frac{N}{N} E\left[x\right] = E[x] \end{aligned} \tag{2}$$

Average = sum of variables with same E[x].

What about the variance?

We need to use the fact that

$$\mathrm{var}[aX] = a^2 \mathrm{var}[X]$$

This is a general statement that we can prove:

$$\begin{aligned} \mathrm{var}[aX] &= E[(aX)^2] - (E[aX])^2 \\ &= \int a^2 X^2 f(x) dx - \left(\int aX f(x) dx\right)^2 \\ &= a^2 \int X^2 f(x) dx - a^2 \left(\int X f(x) dx\right)^2 \\ &= a^2 (E[X^2] - E[X]^2) \\ \mathrm{var}[aX] &= a^2 \mathrm{var}[X] \end{aligned}$$

so we can write:

$$\begin{aligned} \mathrm{var}[\bar{x}] &= \mathrm{var}\left[\frac{1}{N}\sum x_i\right] \\ \mathrm{var}[\bar{x}] &= \frac{1}{N^2}\sum \mathrm{var}\left[x_i\right] \\ \mathrm{var}[\bar{x}] &= \frac{N}{N^2}\mathrm{var}[x] = \frac{1}{N}\mathrm{var}[x] \end{aligned} \tag{3}$$

And you should recognise this from our formula for standard deviation of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$