Probability distribution explorer teaching notes

Anna Fergusson (a.fergusson@auckland.ac.nz)
Department of Statistics, University of Auckland

I developed the probability distribution explorer as part of my Masters research into teaching probability distribution modelling. The proposed teaching framework and the tool were developed in response to use of data for distribution modelling for AS91586, in particular the need for students to demonstrate use of methods related to the **distribution of true probabilities** versus **distribution of model estimates of probabilities** versus **distribution of experimental estimates of probabilities**.

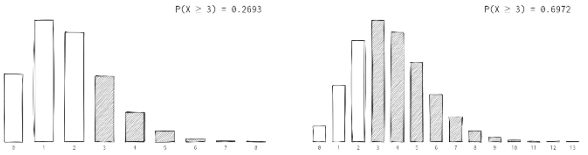
The tool was developed primarily to support comparisons of the "distribution of experimental estimates of probabilities" and "distribution of model estimates of probabilities". When reviewing research literature, I found limited examples of how to teach this comparison using an informal approach i.e. not using a Chi-square goodness-of-fit test. Consequently, I also found a lack of statistically sound criteria to enable drawing of conclusions in such resources as textbooks, workbooks and assessment exemplars.

This led to my research, which involved a small group of New Zealand high school statistics teachers. Focusing on the Poisson distribution, the criteria used by ten Grade 12 teachers for informally testing the fit of a probability distribution model was investigated. I found that criteria currently used by the teachers were unreliable as they could not correctly assess model fit, in particular, **sample size was not taken into account**.

After exploring the goodness-of-fit using my visual inference tool, teachers reported a deeper understanding of model fit. In particular, that the tool had allowed them to take into account sample size when testing the fit of the probability distribution model through the visualization of expected distributional shape variation.

The tool for exploring goodness-of-fit is based on a new framework I developed for teaching modelling in statistics. The tool is designed to reinforce the need to careful to separate the *model world* and the distribution of experimental estimates of probabilities **based on simulated data**, from the *real world* and the distribution of experimental estimates of probabilities **based on observed sample data**. Hence, when exploring goodness-of-fit, the left-hand side is where students work with observed sample data and on the right-hand side is where students work with simulated data (model-generated data).

The tool for exploring features of probability distributions is purposefully situated in the *model-only world*. This tool is designed for students to learn more about the properties of the probability distribution model, such as how the parameters of a probability distribution are linked to its shape and the probabilities of outcomes.



Poisson distribution λ = 1.8 Poisson distribution λ = 3.6

There is a video of my talk about my research and the development of the tool at the 2016 Statistics Teachers' Day using an earlier version of the tool on Census At School. I gave a talk about my research at the New Zealand Statistical Association conference in 2017 and the slides are available here. These slides provide a summary of new and existing research ideas that were used to develop the probability distribution explorer tool.

Dr Michelle Dalrymple has written an article for the Statistics and Data Science Educator which makes use of a previous version of the tool. Her article also provides some more information about my statistical modelling framework on which the tool is designed. You can read this article here: Is the wonky dice fair?. My full Masters thesis is available here. I have a paper based on this research currently submitted for publication and I will make this available once it has been through the peer review process!

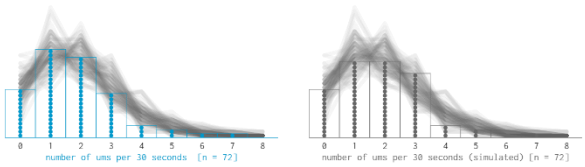
If you are interested in working on software tools like the probability distribution explorer, please get in touch with me to talk more about how can collaborate.

Exploring goodness-of-fit for probablity distribution models

The *Explore goodness-of-fit* feature of the tool allows for an informal visual simulation-based approach to goodness-of-fit tests. The basic idea behind a goodness-of-fit test is to compare what has been observed to what would be expected under the theoretical/model distribution, taking into account how much data is available (sampling/chance variation).

To carry out an informal goodness-of-fit test, students track the over-fitted shape from multiple simulations from the model probability distribution, to build up a visual representation of the model that takes into account sampling variation. The number of values generated from the probability distribution is automatically set to be the same as the number of values in the sample data.

Students can then compare the outcome frequencies for the real sample data with the "fog of uncertainty", to check if there are any sitting well above or below the band for the tracked over-fitted shape as this would suggest a poor fit.



This tool uses individual values for modelling. To create "bins" or class intervals, you can adjust the rounding. The use of dots and bars, and individual values, is purposeful as it reinforces **how much data** is being used in the modelling process. I would suggest using smaller samples to begin with, then moving to larger samples, as the dots get smaller with larger samples.

You do not need to use real data within the *Explore goodness-of-fit* feature of the tool. You could use the "model world" (right-hand) side of the tool to generate different sized samples from probability distributions with different parameters.

Note: This tool is not intended to replace the VIT randomisation test module developed by Chris Wild, which allows students to "test" a single observed value, such as the difference between the mean or median of a variable measured from two treatment groups (for AS91583). Clearly, the visual inference aspect of my tool was based in part on VIT!

Using real data when exploring goodness-of-fit

There are four ways that you get data into the tool:

1. Paste in values from a spreadsheet, then press the update button.
2. Enter each value **individually**, separated by a comma e.g. 2,4,3,5,3,1
3. Enter **frequency** data: Enter the value first, then -, then the frequency e.g. 2,2,2,2,2,2,2 would be entered as 2-7. As above, separate each set by a comma e.g. 2-7,3-9,4-4
4. Create and share a **data link**: Use the formatting rules from 2. and 3. above to prepare your data, ensuring there are no spaces between commas, and add this to the URL for the tool as shown below e.g.

   https://www.stat.auckland.ac.nz/~fergusson/prob_dist_explorer/fit/?data=2-7,3-9,4-4

   or

   https://www.stat.auckland.ac.nz/~fergusson/prob_dist_explorer/fit/?data=2,4,3,5,3,1

You can also add more information to the data link, including the variable name (&var=name+here) and the rounding to be used(&rounding=10). When including the variable name, use "+" for any spaces.

The example below provides a link with data on how many times I said "um" every 30 seconds during 36 minutes of teaching.

https://www.stat.auckland.ac.nz/~fergusson/prob_dist_explorer/fit/?data=4,3,6,3,0,0,1,2,0,0,1,1,3,1,1,1,1,0,0,7,1,0,2,3,2,0,2,3,1,4,3,2,2,2,2,1,1,1,2,0,2,2,3,3,0,0,5,2,5,1,4,1,2,2,2,3,3,2,1,1,1,1,0,1,2,1,3,1,1,2,2,2&var=number+of+ums+per+30+seconds&rounding=1

Teaching ideas and resources

One of the best ways for students to understand **WHAT** they are trying to model, is to collect real data from the underlying process or population. You could check out more information about the SCAMPY tool or consider using a simple data collection tool like my tappity tap tool.

On the Resources for teachers page of the probability explorer tool, I have provided data links that can be used to explore goodness-of-fit. These data have been sourced from the Senior Secondary Guide (SSG), past NZQA exams, and data I have collected.

Some related teaching resources are listed below:

- SSG Lateness: Choice or chance: You could use your own lateness data from a previous class/year.
- SSG A normal search: Some ideas for data that could be collected to investigate the normal distribution as a model. You will need to adjust the rounding to create suitable "bins" to highlight the shape of the distribution.
- SSG A yummy experiment: Some ideas for using food to collect data to investigate the Poisson distribution as a model. Note, be careful about the use of the word experiment for this investigation.
- SSG Am I right? and Have a guess: Some ideas for data for using a multiple choice or true/false test to collect data to investigate the binomial distribution as a model.
- Statistics Teachers' Day 2008 (Auckland) Telepathic or just pathetic: Material I developed ages ago. Note I have not updated these resources so be careful with vocab and terminology, in particular, the incorrect use of the word "resampling" (what I was doing here was randomisation).
- Statistics Teachers' Day 2014 (Auckland)/ Census @School Probability distributions: Material I developed before undertaking this research. In the PowerPoint is more information about the London bombings situation, as well as another example of a multiple choice test. There is also a summary of features of different types of distributions and some old notes about how to fit models. I have not updated these resources so again be careful.
- NZAMT conference 2015 Stimulating simulations: Material I developed in the early stages of this research. In the material is an example of moving from hands on simulation to software driven simulation like this tool. Some more notes are also here on my website teaching statistics is awesome.
- Statistics Teachers' Day 2012 (Auckland)/ Census @School Apply probability distributions in solving problems: Material developed by Dr Marion Steel to support the implementation of the new curriculum and teaching towards AS91586.
- Statistics Teachers' Day 2015 (Christchurch) Making awesome connections between standards: A summary of part of plenary I gave using samples of jelly beans, and a data set for jelly beans.
- Statistics Teachers' Day 2016 (Auckland)/Census @School All models are wrong but some are more wrong than others: Powerpoint slides and video recording of a talk I gave about an earlier version of the tool and my research in this area. Includes a traffic counting activity.
- NZAMT conference 2017/AMA workshops Using data and simulation to teach probability modelling: Slides and teaching notes for AS91585 and AS91586, including using my modelling framework for teaching.

Some relevant considerations for teaching

Below are a few research-informed considerations for teaching probability distribution modelling. Some of these are directly related to using the probability distribution explorer, others are more general ideas related to teaching a modelling perspective.

- Comparisons of data-based estimates for probabilities and model-based estimates for probabilities are difficult to do without taking into account sample size. Just like with confidence intervals, sampling variability needs to be taken into account. The *Explore goodness-of-fit* feature of the probability tool can help students visualise why it matters how much data we have when we are modelling.
- Encourge students to sketch shapes of distributions THROUGH the data, rather than over the top of the data. In this way they are appreciating that the observed proportion for each outcome in the distribution can vary "up and down" from its sample value. The tool can help build this idea because it tracks the over-fitted shape, but the "built up shape" will look like a fuzzy/blurry version of the model distribution shape.
- Students also need to appreciate it is harder to detect the shape of the underlying distribution with small samples – this tool can show how generating simulated data from a normal distribution can give you skewed sample distributions, or how generating simulated data from a skewed triangular distribution can give you symmetric sample distributions.
- More than one version of a model is possible when building a probability distribution from data due to the inferential nature of this process. This can be explored by trying out different parameters for a model for the same sample data and noticing that they all "fit" when building up and comparing the over-fitted shape.
- Remind students that the probability distribution models are being considered as models for the underlying process, system or process that **generated** the data, not the sample data itself. The best model of the sample data, to predict outcomes from within the sample data, is the sample data itself. Be careful to use modelling situations that involve future unknown events with "new data" that involve uncertainty, rather than contrived scenarios like "If we randomly select one of the people in this set of data".
- Try not to fake it! Use real data collected from real situations to explore *goodness-of-fit* informally. Real data is interesting, surprising, and often not as well-behaved as we would like.