The Workhorse

There is an algorithm beloved by bureaucrats, the unsung hero of administrators and accountants. An algorithm both ubiquitous and under appreciated. It's pivotal for nearly every project and informs the actions of tech giants and policy makers the world over. It is only mildly hyperbolic to say that understanding this formula unlocks wealth and power. It lies at the heart of online A/B testing, all policy analysis, sound business strategy and poker play. In words: The expected financial value of a repeated process is just the sum of the possible outcomes weighted by their probabilities.

$$EV(O)_p = p_1 v(o_1) + p_2 v(o_2) + ... + p_k v(o_k)$$

Outcomes can vary from deals of cards, to customer transactions and election results. [1] But aside from the mercenary possibilities, the formula merits your attention for the light it sheds on the puzzling topic of probabilities. With real stakes on the line, the meaning of probability is not an idle concern. The formula for expected value is glossed as one which governs rational action. Yet, statistics are tortured and probabilities are abused to fit policy prescriptions and rubber stamp decisions. False precision of risk models is common place, often callous and occasionally insidious. Nevertheless the rule has an enduring appeal that is worth the work it takes to understand it.

Two sides of a Coin

Probability has a dual aspect. On one reading it refers to the long run tendency of a random process, on another the probability is construed as the degree of belief in the an outcome. On the first interpretation the probability is a distribution with certain fixed theoretical characteristics (e.g. a uniform probability distribution is assumed in the roll of die, where all outcomes are equally likely). On the second reading the characteristics of the probability distribution are learned from the data. These approaches are united by the Law of Large numbers which states that as the size of our sample increases our sample average will converge to the expected

---

[1] In the jargon O is a random variable which can be realised as any of the outcomes $o_j : 1 \leq j \leq k$

realisation of the theoretical process.

$$\frac{1}{N} \sum_{i=1}^{N} O_i \text{ converges to } E(O) \text{ as } N \text{ approaches } \infty$$

In the below plot we have stipulated a Poisson distribution with a mean of 4.5 and can see three examples of how consecutive averaging from the increasing sample sizes results in a closer and closer convergence to the population mean.

Convergence of the average of random variables to its expected value
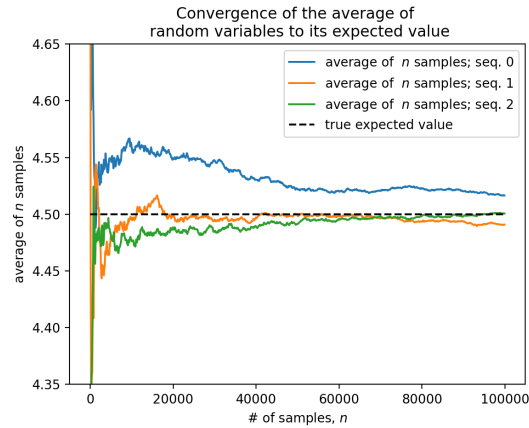
This is fundamental to the long-run tendency interpretation of probability. Given a game with fixed and fair odds we see that repeated play will converge because of characteristics which govern the process. Dice are the paradigm example. In the wild we never know the characteristics which cause the spread of outcomes, but such is the influence of gambling on the consideration of probability, that the default assumes a stable pattern in the generating process. Partially this is pragmatic. The maths are more tractable if we can assume one well behaved underlying process. The results are compelling. The Normal (Bell Curve) distribution, the Poisson distribution the Bernoulli distribution (to name a few) are all rightly famous. Their shapes are characteristics of innumerable random processes. But the paradigm clouds the fact that in practice we start on the left side of the law of large numbers (with samples) and we often start with small numbers.

Variety of distributions with parameter specifcations
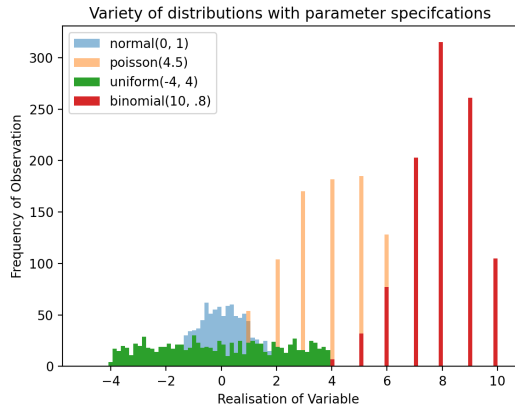
Well behaved probability distributions are rare beasts; a tiny fraction of the world's arbitrary menagerie. The fundamental question in probability is not whether probability is a measure of belief or frequency - it is whether we can safely assume that the underlying process adheres to a known model? If so, we can rely on the structure of the model's theoretical distribution to inform inference. If not we are better learning from the data - trusting to the sampling distribution and worst scenario planning.

Models, Errors and Expectations

When your only tool is a hammer, then everything is a nail. When one model won't do, use two. This is roughly the approach adopted when models fail. The problem is especially vivid in time-series forecasts; projected death rates, body counts and stock prices are all subject to sudden shocks. A basic regression model has following form:

$$Y = const + \beta X + \epsilon$$

where ϵ is a random variable accounting the error between the data and values predicted by the equation. While $const, \beta$ are parameters estimated by an optimisation process to ensure the equation fits the data as neatly as possible. In the below graph we have a series characterised by change. After the first shock we can refit the model so that the line tracks well with the evolving data. After the second shock we try another refit, but the range the and variance of the data makes our basic model a poor fit. This presents three examples of error in the modelling process: (i) forecasts fail for

the reason that's it's also difficult to identify (in the moment) those changepoints in the data which reflect structural change, (ii) the fundamental assumptions that go into the model are sound, but the parameters need be re-estimated based on the new data
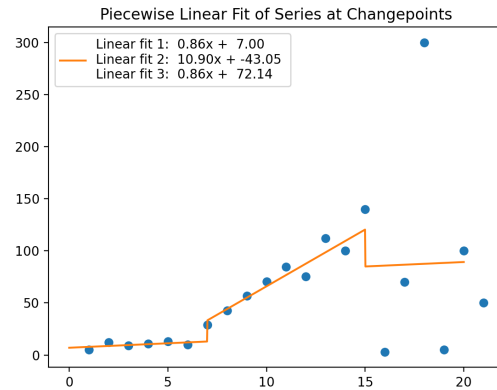
and (iii) the third linear model is simply a terrible match for the pattern in the data. In practice you never really know whether a new error stems from a misfit but appropriate model or an entirely inappropriate model. As we increase our number of sample fits we hope to better approximate the true linear process (if any) generating the data. This warrants a note about errors. Imagine now that the data points on our previous graph are repeatedly re-speckled over the canvas. We can refit a new model for each set of scattered data points and each refit gives us a new sample values for $const, \beta$. If the underlying data generating process is stable, then on average the parameter fits will converge to the characteristics of the stable process[2] Our sample estimates will overshoot in some cases but on the whole the errors up and down will cancel each other out if the errors themselves are normally distributed and the expected realisation of the error ϵ term is 0. Forecasting with the expected parameters derived from our sampling distribution minimises our forecast errors. These are assumptions required for a process to exhibit the tendency of

[2]the correct values of $const, \beta$. For example if the we're predicting height based on weight, our values for $const$ should be above 4.ft and the value for β should be calibrated to make up the average difference when multiplying the value for weight.

2

regression towards the mean.

The code below builds two sampling distributions based on different underlying processes. One in which the errors are independent, normally distributed around 0 the other in which the errors are correlated in a sine-wave like pattern, increasing and decreasing periodically. This is akin to difference between measuring error when predicting the heights of randomly sampled people versus predicting the sales volumes on randomly selected days of the week. A random sample of daily sales risks clumping weekends together and skewing the expected values. No such risk exists when sampling from independent individuals.

```
### Build True Models
N = 100000
X = np.random.uniform(0, 20, N)
independent_err =
np.random.normal(0, 10, N)
corr_err =
np.random.uniform (0, 10) +
np.sin(np.linspace(0, 10*np.pi, N)) +
np.sin(np.linspace(0, 5*np.pi, N))**2 +
np.sin(np.linspace(1, 6*np.pi, N))**2

Y_corr = −2 + 3.5 * X + corr_err
Y = −2 + 3.5 * X + independent_err
population = pd.DataFrame({ 'X': X,
 'Y': Y,
'Y_corr': Y_corr})

### Sample from Data
### and build smaller models

fits = DataFrame(columns=['iid_const',
 'iid_beta', 'corr_const',
 'corr_beta'])

for i in range(0, 10000):
    sample = population.sample(n=100,
    replace=True)
    Y = sample['Y']; X = sample['X']
    Y_corr = sample['Y_corr']
    X = sm.add_constant(X)
    iid_model = sm.OLS(Y, X)
    results = iid_model.fit()
    corr_model = sm.OLS(Y_corr, X)
    results_2 = corr_model.fit()
    row = [results.params[0],
```
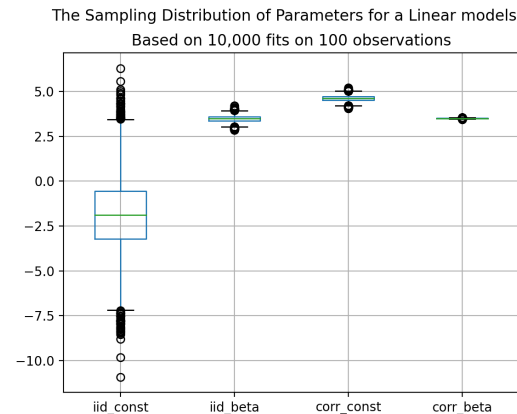
```
    results.params[1],
    results_2.params[0],
    results_2.params[1]]
    fits.loc[len(fits)] = row

fits.boxplot()
```


The Sampling Distribution of Parameters for a Linear models
Based on 10,000 fits on 100 observations

In the case with independent errors the expected value for our parameter estimates match almost exactly the true values of the process. In the second model with correlated errors the parameter estimate for our constant is 4.9 which is significantly different from the true value of -2, and will lead to systematically skewed predictions.

Small World of the Model

When does it make sense to use models? In what sense should expected values guide our action? These are questions which arise in any sustained investigation of a random process. From where do we estimate the values for

$$p_i : 1 \leq i \leq k$$

in the formula (EV) if models are mere heuristics. Models are deliberate simplifications of complexity, a toy version of the world. Recent innovations of in machine learning means that we can fit models with hundreds, thousands and even billions of parameters but this increase scale does not change fact that models are blind to all factors outside their samples. A good model predicts the variance in future data, in part, by ignoring the genuine complexity of reality. This is the only practical test that matters.

3

If the prediction accuracy holds over successive periods we should become increasingly sure that the probabilities in our model are a good measure of risk. But we can never be certain the model holds over all successive periods and is robust to changes in external factors. We can however quantify the degree of uncertainty in the model by observing the spread of the sampling distribution.

Styles of Inference

Prediction precedes probability in the order of analysis. Without a reliable regularity underwriting our observations, we assume them to be noise. Only when there is a better than arbitrary correlation between X, Y will we even think to ascribe a measurable probability to their association. It's at this point that we can debate the finer points of statistical inference.