

# Examining Algorithms

Nathaniel Forde

2021-03-06



# Contents

<b>1</b>	<b>Preface</b>	<b>5</b>
<b>2</b>	<b>The Average Man</b>	<b>7</b>
2.1	...and Expected Value . . . . .	8
2.2	Probability: Twin Aspects . . . . .	9
2.3	Small Worlds and Statistical Inference . . . . .	13
2.4	Utility Curves . . . . .	26
2.5	Optimising Utility . . . . .	30
<b>3</b>	<b>Literature</b>	<b>33</b>
<b>4</b>	<b>Methods</b>	<b>35</b>
<b>5</b>	<b>Applications</b>	<b>37</b>
5.1	Example one . . . . .	37
5.2	Example two . . . . .	37
<b>6</b>	<b>Final Words</b>	<b>39</b>



# Chapter 1

## Preface

What this book is



## Chapter 2

# The Average Man

In the 1840s the Average man stalked the nightmares of Augustin Cournot. The mathematician was haunted by a melange of imagined parts. A Frankenstein's monster of mismatched limbs, variously soldered, stapled or sown at the joints. He worried that no one but a "physical monstrosity" could exhibit the average, weight, height and other mean attributes in one body. (Stigler, 2016) But fantastical fears were no bar to a useful mathematical fiction. The notion of averaging was a technological breakthrough - applications of averaging multiplied without cease: polling, gambling, forecasting - statistics were recorded everywhere, compounding one on another; averages of averages tenuously tethered to observable facts by layers of abstraction and scales of measurement.

Slowly, doubts began to creep back into the statistics. In the 1970s the psychologist Paul Meehl would worry that such brute approximations had stifled the development of the softer sciences and contributed to the mis-measurement of man. He would go on to elaborate twenty features of psychological science which made such measurement constructs inapt, unreliable and difficult to clearly falsify. This was progress! (Meehl, 1978) He then showed that the methods of validating such constructs were the main culprit and the likely cause of psychology's implausible claims to concrete, replicable results. Around thirty years later some of the same replication issues would come to be called a crisis. Cournot was right to fear.

Seen from the perspective of a patient the difficulties are more urgent and stark. In 2019 Esme Yang would write with hope of the comfort given by a diagnosis, the knowledge that she was not "pioneering an inexplicable experience."<sup>1</sup> To find yourself described in the DSM provides: a framework and the glimmer of a cure, a chance to slot yourself into a category, a community of care and a medicinal regime. Frustrating then when the categories themselves are in flux. Diagnostic

---

<sup>1</sup>All remarks on Schizophrenia in the following draw on the precise and personal essays in (Wang, 2019)

criteria arrayed over page after page attempting to capture the elusive core of a psychological dysfunction. Descriptions are derived from clinical interviews correlated and meshed with genetic markers, then poorly mapped to a family of ailments. Neither clearly bipolar disorder nor manic depression, schizophrenia, psychosis or schizo-affective disorder. But an idiosyncratic presentation (as they all must be), uniquely felt and individually suffered.

Wang’s collection of schizophrenias defy easy taxonomy “because there are just so many different ways in which people can develop a syndrome that looks like schizophrenia ... as we now define it.” The task then becomes one of coping with uncertainty and adjusting expectations. It’s not so exact a science that you can measure twice and cut once. The measurement schemes change and you need multiple cuts. You measure out the impact of treatments and the trade-offs - what’s non-negotiable and how many side-effects are acceptable? What works for you versus what’s advised by the professionals. Sterile cost-benefit considerations become suddenly dramatic and life changing.

## 2.1 ...and Expected Value

There is an algorithm beloved by bureaucrats. An unsung hero of administrators and accountants. An algorithm both ubiquitous and under appreciated. It’s pivotal for nearly every business and informs the actions of tech firms and policy makers the world over. It is only mildly hyperbolic to say that understanding this formula unlocks wealth and power. The algorithm lies at the heart of online A/B testing, all policy analysis, sound strategy and poker play.

$$EV(O)_p = p_1 u(o_1) + p_2 u(o_2) + \dots + p_k u(o_k)$$

The expected financial value of a random process is just the sum of the utility (typically dollar outcomes) weighted by their probabilities. Outcomes can vary from deals of cards, to customer transactions, election results, continued sanity. Pascal can argue sincerely that such considerations compel even belief in God. The infinite downsides of hell at any likelihood ought to compel even the cynical sceptic. But the formula, glossed as a rule for rational action, merits your attention for more mundane wagers too. If you intend to maximise your expected value, the meaning of probability is not an idle concern.<sup>2</sup> While statistics are often tortured to rubber stamp decisions and probabilities are abused to fit policy prescriptions with false precision, the crisp clarity of the rule has an enduring appeal that promises to sift the murky swamps of Big Data. It’s a scalpel that anyone can wield to parse the syntax of statistical jargon and carve answers from an abstract space of probabilities. “What’s my expected return? My likely life-quality?” - a simple question, with a surprisingly complex answer.

---

<sup>2</sup>For a typical example of how *EV* is used to express decisions under uncertainty see chapter 7 in the textbook (Barber, 2012), but the paradigm owes much of its influence to decision theoretic work of Leonard Savage in (Savage, 1954)



Decision theory is an abstract formalism which purports to account for how you ought to reason under uncertainty, conditional on knowing your own mind, what you want and the likelihood of those outcomes. Wrestling with a diagnosis of schizophrenia you weigh your future in the face of sickness while knowing your mind could fail in the act, knowing, perhaps, you're not yourself. Schizophrenics are involuntarily detained if a medical professional deems them to have lost sufficient "insight"<sup>3</sup>, but the stated dysfunction assumes that these kinds of insight should be transparent when the psychiatric crisis is resolved. We'll see that the level of insight assumed by decision theory is, at best, hard won and far from transparent.

There is a secondary complication in that whenever expected utility as used as an explanatory model of rational action, there is a risk of pivot, a point where it is substituted for more obscure black-box optimisation techniques over metrics of profit or overall accuracy. The focus switches from modelling the consumer or the patient to modelling returns based on the consumer and the firm's expected value. Swapping a customer catering model for a customer-as-commodity perspective focused on profit in aggregate and customer acquisition. The tendency is common because profit overwhelms all other priorities, but the loss of understanding typically amounts to a longer term net-loss. Shareholders take comfort from increased profit and algorithms deployed at scale, but it's rare that any single algorithm actually or always maximises the available profit. This dynamic enforces a kind of inescapable see-saw motion where the consumer modeling exercise goes through a constant feedback loop. A good model (informal or formal) of human needs and wants feeds a better a predictive model of individual action. When the latter fails we go back to the utility curves and the algorithm of expected value because it is (if not reliably predictive) a rich and deeply explanatory model of human action under uncertainty.

## 2.2 Probability: Twin Aspects

Probability emerged slowly and with a dual aspect. On one tradition probability refers to the long run tendency of a random process, on another probability is construed as the degree of belief in an outcome. On the first (frequentist) interpretation a probability distribution has certain fixed theoretical characteristics: as in a uniform probability distribution of a fair coin where all outcomes are equally likely, or as with the normal distribution where most outcomes cluster symmetrically about a central average. On the second (Bayesian) reading the characteristics of the probability distribution are learned from the data. The controversy centres around the fact that it's unclear how a frequentist could ascribe probabilities to unique events. Without appeal to a large set of observations (or known theoretical distribution) the claim that an event appears

---

<sup>3</sup>"The mind has been *taken over*. The mind has *lost the ability to make rational decisions* in (Wang, 2019) pg58

frequently or infrequently is not well defined. Consequently, tabulations of probability appear inappropriate for claims of unique or rare events. In contrast the Bayesian is content to ascribe probabilities to any all partial beliefs no matter how specific. For the Bayesian, the probability calculus is a set of edicts about how to rationally manage and modulate your beliefs. So it's acceptable to have a probabilistic belief in rare cases so long as you update those probabilities with new data when available.

These two 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 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 this graph we have fixed 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 (true) population mean.

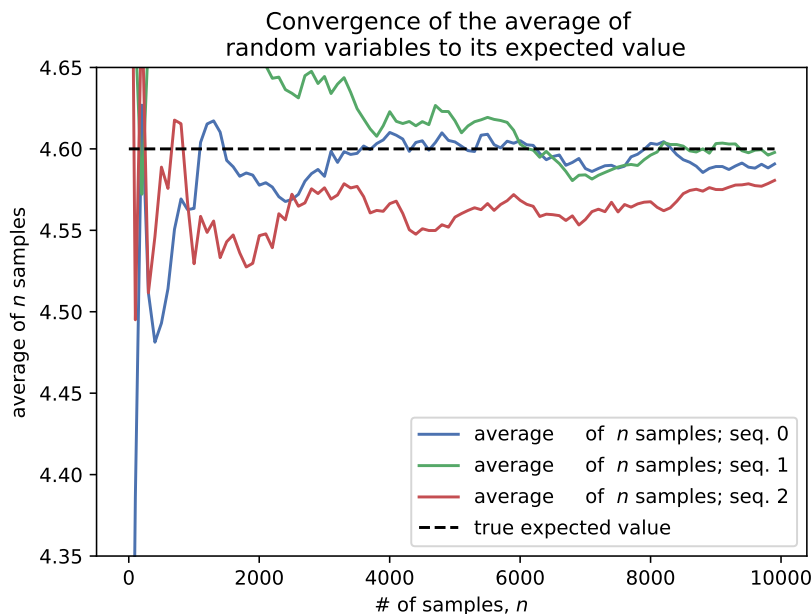
```
# set up the ground truth
np.random.seed(100)

sample_size = 10000
expected_value = lambda_ = 4.6
poi = np.random.poisson
N_samples = range(1, sample_size, 100)

for k in range(3):
    samples = poi(lambda_, sample_size)
    partial_average = [samples[:i].mean() for i in N_samples]
    plt.plot(N_samples, partial_average, lw=1.5, label="average \
of $n$ samples; seq. %d" % k)

plt.plot(N_samples, expected_value * np.ones_like(partial_average),
         ls="--", label="true expected value", c="k")

plt.title("Convergence of the average of \n random variables to its \
expected value")
plt.ylabel("average of $n$ samples")
plt.xlabel("# of samples, $n$")
plt.legend()
lims = plt.ylim(4.35, 4.65)
plt.show()
```



Though common knowledge today, in 1650 “the very concept of averaging is [new]... and most people could not observe an average because they did not take averages.” (Hacking, 2006) Systematically grappling with the implications of observations is a somewhat recent human endeavour - one which is far from perfected. This tendency is now fundamental to the interpretation of probability. Take a game with fixed and fair odds and we see that repeated play will converge over time because of characteristics which govern the process. Dice are a homely example. In the wild we never know the characteristics which cause the observed spread of outcomes, but such is the influence of gambling on probability, that we assume there exists a stable pattern to be gamed. Partially this is pragmatic, the maths is more tractable if we can assume a 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. The distributions cleanly circumscribe and corral likely patterns of events.

```
normal = np.random.normal(0, 1, 1000)
poisson = np.random.poisson(4.5, 1000)
uniform = np.random.uniform(-4, 4, 1000)
binomial = np.random.binomial(10, .8, 1000)

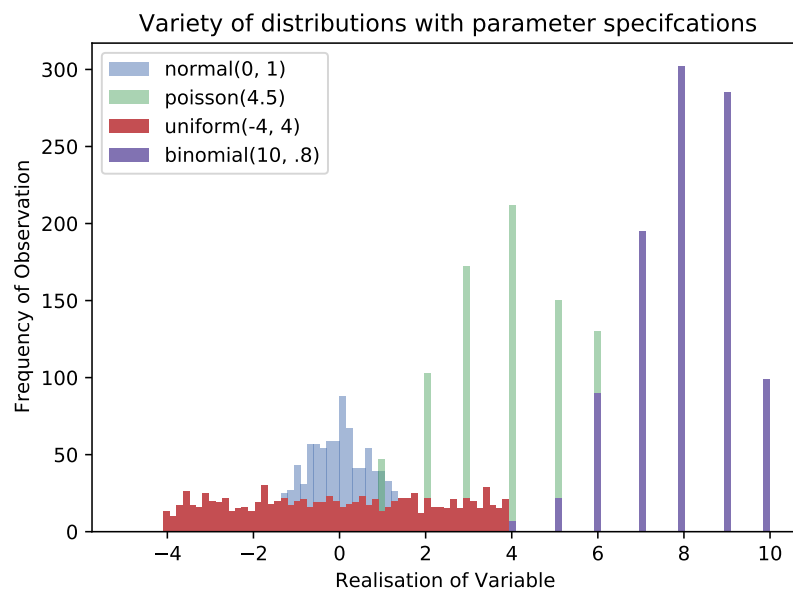
bins = np.linspace(-5, 10, 100)

h1 = plt.hist(normal, bins, alpha=0.5, label='normal(0, 1)')
```

```

h2 = plt.hist(poisson, bins, alpha=0.5, label='poisson(4.5)')
h3 = plt.hist(uniform, bins, label='uniform(-4, 4)')
h4 = plt.hist(binomial, bins, label='binomial(10, .8)')
plt.legend(loc='upper left')
plt.title("Variety of distributions with parameter specifications")
plt.xlabel("Realisation of Variable ")
plt.ylabel("Frequency of Observation")
plt.show()

```



But the gambling 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 resulting from a unknown number of data-generating processes. 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? The Bayesian focus is to try and learn from the new data the expected characteristics of the underlying process, while the Frequentist tries to gauge the accuracy of their assumptions about the underlying process. Both are attempts to validate the structure of the model's theoretical distribution to inform inference. If we can't validate a model, we're better learning what we can from the sample, trusting to wide confidence intervals and worst scenario planning. But in all cases when we need to make a decision, the following questions are inescapable : What are your expectations based on? How do they figure in our choices, and can we use them to improve our outcomes?

## 2.3 Small Worlds and Statistical Inference

Anissa Weier and Morgan Geyser, later diagnosed with schizophty, were in 2017 trailed for the attempted murder of their friend Payton. While believing themselves to be acting at the behest of Slendar Man, they were “willing to forgo even friendship for the sake [their] version of unreality”<sup>4</sup>. Just as in the depth of an obsession reality can be warped by a psychotic focus, a statistical model will accentuates some parts and ignores others.

This narrowness can have devastating effects if deployed without care. But prediction is a visceral need, unavoidable, it precedes the sophistication of probability in any order of analysis. Without some regularity between  $X, Y$  we can only interpret their collisions as timorous noise. But even when there is a better than arbitrary correlation between  $X, Y$  we’ll insist on ascribing a measurable probability to their association. Heuristics will kick in, and confidence will grow out of proportion to the evidence. So whether we view probabilities as a measure of credibility or frequency, the focus is always on a process which in reality reveals a pattern under repetition. Even tenuous connections can be enough to cause havoc. Models, more often than not, are deliberate simplifications of complexity, attempts to formulate a unobserved process within a mathematical structure. Lego-like versions of the world are built and rebuilt, in which we bash parts together to see what reliably sticks. Forget about the pieces we don’t own. Count the pairs of blocks that: match, wedge, smash, click-into-place or break, then draw out the ratio of success and the spread of outcomes. Parse out the details of how reds go with greens, and blues with yellows. This is your sample distribution. The expected gain depends on both this uncertainty, measured in this small world, and the finer points of statistical inference.

A [small world] is... completely satisfactory, only if it is actually a microcosm, that is, only if it leads to a probability measure ... that can be written down explicitly pg 88 (Savage, 1954)] .

The danger of shrinking your world comes when you mistake the map for the territory and act on that delusion.

Small worlds are machines for figuring out expected values of a statistical process. We shrink the parameter set to better measure the variance and flux throughout the system. For any hypothetical system there can be multiple plausible approximations of the underlying process which need to be assessed comparatively on their “goodness” of fit. We iterate through new and improved versions, each an attempt to make a conjectural connection between  $X, Y$  mathematically precise. But once built, they embed the errors and assumptions made in their design. McElreath dubs them Golems, primed and then loosed on the world, insensitive to subtlety and context they perform only as instructed. Consider how a basic regression model tries to predict an outcome  $Y$  as linear function of some observed features  $X$ :

---

<sup>4</sup>“The Slender Man, the Nothing and Me” in (Wang, 2019)

$$Y = \text{const} + \beta X + \epsilon$$

where  $\epsilon$  is a random variable representing the error (or noise). A modest notational device for disaster. While  $\text{const}, \beta$  are parameters estimated by an optimisation process to ensure the equation fits the data as neatly as possible. In the plot below 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 and variance of the data makes our basic model a poor fit, i.e. the data no longer exhibits a linear relationship. This presents three examples of error in the modelling process: (i) it's difficult to identify (in the moment) those changepoints in the data which reflect structural change, (ii) the linearity 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.

### ### Piecewise Linear Fits

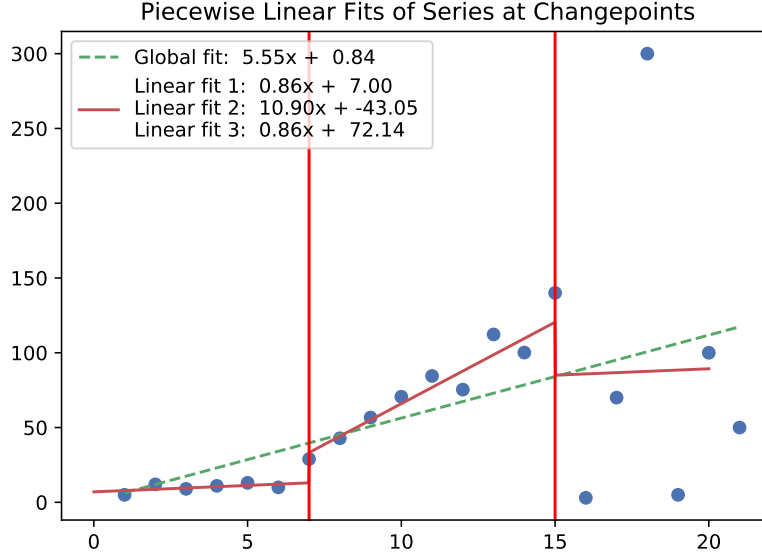
```
x = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21])
y = np.array([5, 12, 9, 11, 13, 10,
              28.92, 42.81, 56.7, 70.59, 84.47, 75.36, 112.25, 100.14, 140.03, 3, 70, 3])

def piecewise_linear(x, b1, a1, b2, a2, b3, a3):
    funcs = [lambda x:b1*x + a1,
              lambda x:b2*x + a2,
              lambda x:b3*x + a3]
    conds = [x < 7, ((x >= 7) & (x < 15)), x > 15]
    return np.piecewise(x, conds, funcs)

p, e = optimize.curve_fit(piecewise_linear, x, y, method="trf")
a, b = polyfit(x, y, 1)
xd = np.linspace(0, 20, 1000)

plt.plot(x, y, "o")
plt.plot(x, a + b * x, '--', label="Global fit: {b: .2f}x + {a: .2f}".format(b=b, a=a))
plt.plot(xd, piecewise_linear(xd, *p), label="Linear fit 1: {b: .2f}x + {a: .2f} \n"
                                         "Linear fit 2: {b1: .2f}x + {a1: .2f} \n"
                                         "Linear fit 3: {b2: .2f}x + {a2: .2f} \n".format(*p))

plt.axvline(x=7, color='red')
plt.axvline(x=15, color='red')
plt.title("Piecewise Linear Fits of Series at Changepoints")
plt.legend()
plt.show()
```



Every model is a guess as to the implicit order in apparent noise. Sometimes there is no order, and other times the patterns is too subtle for a dumb model to capture. In practice you never really know whether a single 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. Imagine now that the data points in Figure 3 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 the parameter fits will converge to the correct values of  $const, \beta$ ; correct in the sense that they can be used to draw the line of best fit for the data. A statistically stable process is one that can be modelled with errors  $\epsilon$  normally distributed around 0, so that the model will be *correct on average* because  $E(\epsilon) = 0$ . Our predictions will overshoot in some cases but on the whole the errors up and down will cancel each other out.

“Typically, the assumptions in a statistical model are quite hard to prove or disprove, and little effort is spent in that direction. The strength of empirical claims made on the basis of such modeling therefore does not derive from the solidity of those assumptions. Equally, these beliefs cannot be justified by the complexity of the calculations... These observations lead to uncomfortable questions” (Freedman, 2009)

Forecasting with the parameters of best fit minimises our forecast errors because the fluctuations are stable about the centre of the line. These are the required

assumptions for a process to exhibit the tendency of regression towards the mean. If they're not met, we will see poor parameter estimates and wild swings away from the linear path. The fundamental statistical assumption here is about the properties of our mistakes! The model is less plausible if our judgements are made in the grip of a delusion.

### 2.3.1 Modeling: Improper Assumptions and Skewed Expectations

Below we build two sampling distributions based on different models of an underlying processes. One in which the errors are independent, normally distributed around 0 and in the other the errors are correlated in a sine-wave like pattern, increasing and decreasing periodically. This is akin to the 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.

```
#### Sampling Distributions of Linear Fits
#### Build True Model
N = 1000
X = np.random.uniform(0, 20, N)
uncorrelated_errors = np.random.normal(0, 10, N)
correlated_errors = 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 + .6*X

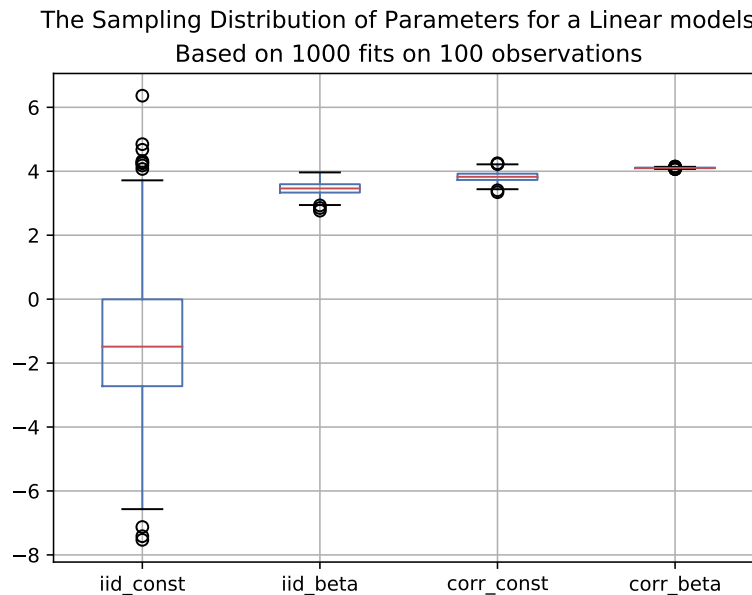
Y_corr = -2 + 3.5 * X + correlated_errors
Y = -2 + 3.5 * X + uncorrelated_errors
population = pd.DataFrame({'X': X, 'Y': Y, 'Y_corr': Y_corr})

fits = pd.DataFrame(columns=['iid_const', 'iid_beta', 'corr_const', 'corr_beta'])
for i in range(0, 1000):
    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()
```



```
plt.suptitle("The Sampling Distribution of Parameters for a Linear models")
plt.title("Based on 1000 fits on 100 observations")
plt.show()
```



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. Statistical models are just algebraic equations where we use regular sampling to solve for  $Y$  from  $X$ . Because  $Y$  is also a random variable the regression model encodes a conditional expectation result.

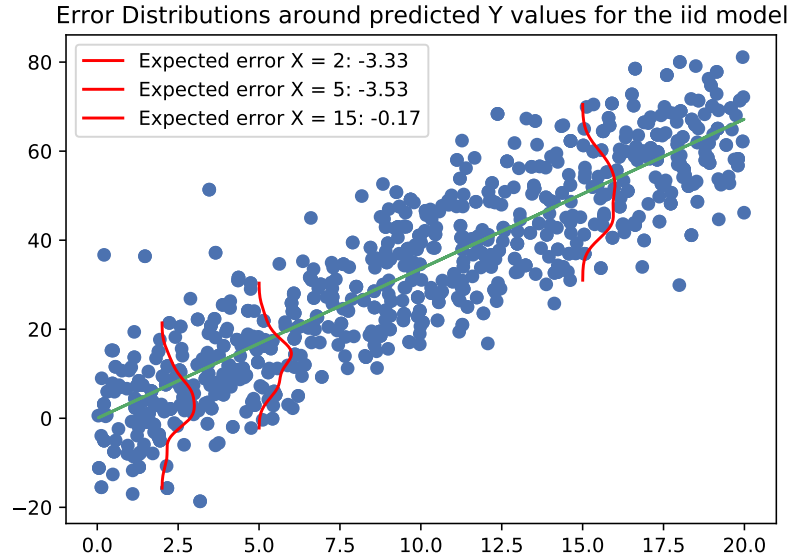
```
def draw_gaussian_at(position, sample, ax_main=None, model='iid', color='k', **kwargs):
    filter_var = round(sample['X'], 0) == position
    avg = sample[filter_var]['predicted_Y_ ' + model].mean()
    min = sample[filter_var]['Y'].min()
    max = sample[filter_var]['Y'].max()
    dist = pd.Series(sample[filter_var]['predicted_error_' + model].values)
    kde = sm.nonparametric.KDEUnivariate(dist)
    kde.fit()
    density = kde.density
    density /= density.max()
    density *= 1
    y_axis = np.linspace(min, max, len(density))
    label = "Expected error X = {x:}: {err: .2f}".format(err = sample[filter_var]['predicted_error_' + model].mean())
```

```

x=position)
ax_main.plot((density + position), y_axis, color=color, label=label)

sample = population.sample(n=1000, replace=True, random_state=100)
true_model = sm.OLS(sample['Y'], sample['X']).fit()
error_model = sm.OLS(sample['Y_corr'], sample['X']).fit()
sample['predicted_Y_iid'] = true_model.predict(sample['X'])
sample['predicted_Y_corr'] = error_model.predict(sample['X'])
sample['predicted_error_iid'] = sample['Y'] - sample['predicted_Y_iid']
sample['predicted_error_corr'] = sample['Y'] - sample['predicted_Y_corr']
fig, ax1 = plt.subplots()
ax1.plot(sample['X'], sample['Y'], 'o')
ax1.plot(sample['X'], sample['predicted_Y_iid'])
for each in [2, 5, 15]:
    d = draw_gaussian_at(position=each, sample=sample, ax_main=ax1, model='iid', color='red')
plt.title("Error Distributions around predicted Y values for the {model:}".format(model='iid'))
plt.legend()
plt.show()

```



$$E(Y_i | X_i = x)$$

For fixed values of  $X$ , the predictions  $Y_i$  can be spread in a pattern enforced by the various ways we can realise the linear function with estimates for  $\beta$  and  $const$ . But the regression model selects the best parameter values to minimise

the squared prediction error and represent the conditional expected distribution of  $Y$ .

“The statement that regression approximates the [Conditional Expectation Function] lines up with our view of empirical work as an effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly” p38 - (Angrist and Pischke, 2008)

The consequent point predictions for  $Y$  are always expected values, skewed by the how the parameters are realised from sample data as much as by poor choices in model design and predictive features. So too then any measures of expected utility based on these models or inferences from these distributions.

### 2.3.2 Frequentism: Inference from Expected Frequency

Making inference from a model is delicate thing. Even simple cases come with controversy. Count the number heads in a series of 5 successive coin flips, then repeat the process 1000 times and you'll arrive at a proportion which characterises that process. If it's a fair coin the long run expected result will be half the number of your coin flips. If the coin is weighted you might have as few as 0. This is the binomial distribution, and it really shines when you're trying to gauge fairness. If a process is biased, the distribution will be skewed. We can use this fact for inference. Consider a dispute over whether the game was rigged.

$$H_0 : \text{true proportion of heads} = 0.4$$

$$H_1 : \text{true proportion of heads} = 0.5$$

```

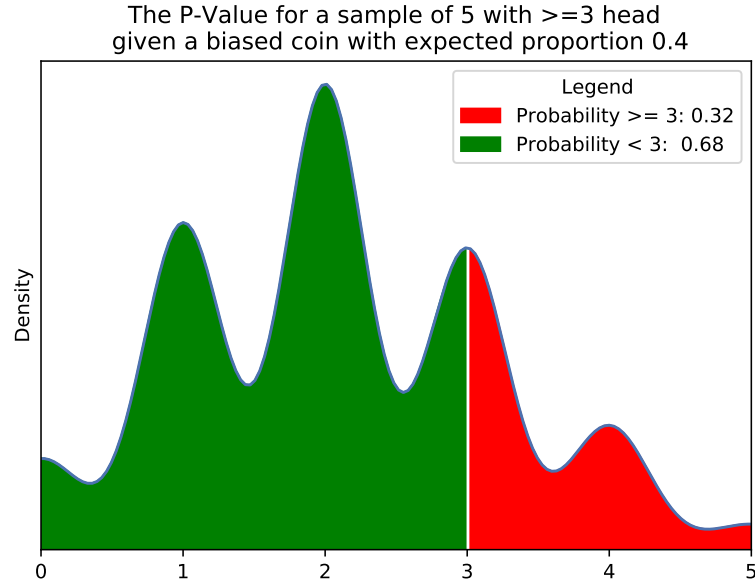
samples = 5
prop = 0.4
successes = 3

X = binom(samples, prop)
x = X.rvs(1000)
points = sns.distplot(x, hist=False, kde=True).get_lines()[0].get_data()
prob = 1 - X.cdf(2)
z = points[0]
y = points[1]

plt.fill_between(z,y, where = z >= successes, color='r', label="Probability >= 3:{p: .2f}".format(
plt.fill_between(z,y, where = z < successes , color='g',
                    label="Probability < 3: {mean: .2f}".format(mean=1-prob))
plt.legend(loc='upper right', title='Legend')
plt.title("The P-Value for a sample of 5 with >=3 head \n given a biased coin with expected propo

```

```
xlim = plt.xlim((0, 5))
ytck = plt.yticks([])
plt.show()
```



Take ( $H_0$ ) as given then if we observe a sequence:

$$(3in5) : H, H, H, T, T$$

what does it say about the possibility that we're being hustled? If the coin is biased, then the count of heads in repeated sampling will reflect a clear bias. For any new data we can check if the data is consistent with the data generated by the biased coin. The pattern of reasoning is straightforward (i) make some assumptions about the structure of the random process under investigation, (ii) tease out the consequences of these assumptions (iii) evaluate the incoming data against these consequences to see if you need to revise your assumptions. The frequentist asks, does the data looks weird given the assumed shape of our probability distribution?

[I]n statistical terms  $H_0$  [the null hypothesis] refers to a probability model and the very word 'model' implies idealization. With a very few possible exceptions it would be absurd to think that a mathematical model is an exact representation of a real system... We use the term to mean that in the current state of knowledge it is reasonable to proceed as if the hypothesis was true." pg 31 (Cox, 2006)

In this instance the shape of the binomial distribution defined by a 0.4 biased coin allows for significantly greater than 5% chance for observing the above

sequence. So we do not have enough reason to reject  $(H_0)$  at the traditional threshold. By design the assumed distribution builds in characteristics of long-run variance of the process, and the slim threshold for rejection is designed to minimise incorrect rejections of  $(H_0)$ . We should remain suspicious that we're being conned. However, with a low number of observations the sample distribution is unlikely to be properly representative. This makes even small p-value thresholds unreliable. We cannot blindly take a sample poll to imply the spread or volatility of a population, and with low or un-representative samples it's hard to justify any kind of inference from expectation, since we are not in a position to justify the choice of the null model either! If your hypothesis is both derived and validated against small samples, you risk being swayed by recent observations.

More fundamentally the notion of statistical significance usually cannot falsify the hypothesis under consideration. The sheer number of auxiliary variables that you might need to control for, makes the practical task of definitively rejecting the null almost impossible. There are too many imagined ways in which the auxiliary conditions, sufficiently modified, would have resulted in observations that corroborate the null model. This problem is especially acute in psychological science where the auxiliary contingencies of designing a measurement scale and checking diagnostic criteria are almost always questionable. So the null hypothesis is not confirmed, but not refuted either, it is just preserved in useless stasis.

“[I]t did not get integrated into the total nomological network, nor did it get clearly liquidated as a nothing concept, it did not get killed or resurrected or transformed or solidified; it just kind of dried up and blew away...” pg807 in (Meehl, 1978)

### 2.3.3 Bayesian Inference: Inference to Expected Value

If instead we use probability to calibrate our beliefs, then we can be more explicit in our assessment of  $(H_0), (H_1)$ . Let's assume that our prior beliefs about whether the game is rigged is 50/50. Then we evaluate the two hypothesis using Bayes's rule for incorporating our prior belief and the data. The Bayesian asks whether our hypothesis is a good explanation of the data compared to alternatives. How, upon observing the data, should we view our hypothesis?

$$p(H_i|Data) = \frac{\overset{\text{prior likelihood}}{p(H_i)p(Data|H_i)}}{\sum_{i=1}^{i=K} \underset{\text{evidence}}{p(Data|H_i)p(H_i)}}$$

where  $\$ 1 \leq i \leq K\$$  spans the ways in which the data could have been realised across all competing hypotheses. Then, in our toy example, we have:

$$\frac{p(H_1|3in5)}{p(H_0|3in5)} = \frac{\frac{.5 \cdot .23}{.5 \cdot .32 + .5 \cdot .23}}{\frac{.5 \cdot .32}{.5 \cdot .32 + .5 \cdot .23}} = \frac{.57}{.42}$$

which would lead us to infer that the coin was fair. The really radical move in the Bayesian setting is that you're allowed to ascribe a probability to any event regardless of whether there is any long-run sequence to observe. You may know nothing about your opponent or the coin, but for Bayesians this is no bar to assigning suspicion in the form of expected probability, so long as you act in accordance with the axioms of probability and weigh the probabilities accordingly. In particular it promotes the direct comparison of competing hypotheses conditional on the evidence. It's this free choice of prior which can seem arbitrary and unmotivated or even paradoxical, but in practice probabilities are rarely ascribed entirely without reason and it's frankly irresponsible to ignore those reasons.

“When a piece of evidence E is produced in a court investigating the guilt G or innocence I of the defendant, it is not enough merely to consider the probability of E assuming G; one must also contemplate the probability of E supposing I. In fact, the relevant quantity is the ratio of the two probabilities. Generally if evidence is produced to support some thesis, one must also consider the reasonableness of the evidence were the thesis false. Whenever courses of action are contemplated, it is not the merits or demerits of any course that matter, but only the comparison of these qualities with those of other courses.” in (Lindley, 1993)

Neither the Bayesian or Frequentist analysis ends with these simple calculations, both should continue to probe the limits of each hypothesis. We'd have to consider things like sample size, sensitivity testing, model performance, the cost of errors and appropriateness of the priors. The point is just that there are reasons for dispute. Bayesian inference acts like a logic engine for evidence, whereas the frequentist approach is more focused on diagnosing the possibility of error. In general they are complementary methods, and when they conflict the assumptions should be scrutinised. The frequentist evaluation of our biased coin is very sensitive to the choice of hypotheses, while the Bayesian approach is influenced by the choice of prior. Why set up a significance test against assumed cheating rather than assumed fairness? Why attribute equal weight to both hypotheses? Why use a 5% threshold if you're concerned about systematic cheating? This example shows the heart of the conflict in the dual aspect of probability. There is enough latitude in the manner in which we set up a probability model that the mathematics of inference can yield inconsistent results. Both offer strategies for managing uncertainty, but both approaches come with baggage and in practice not all tests are equally taxing. Consider a more concrete example in the Bayesian spirit.

### 2.3.4 Example: Website Traffic

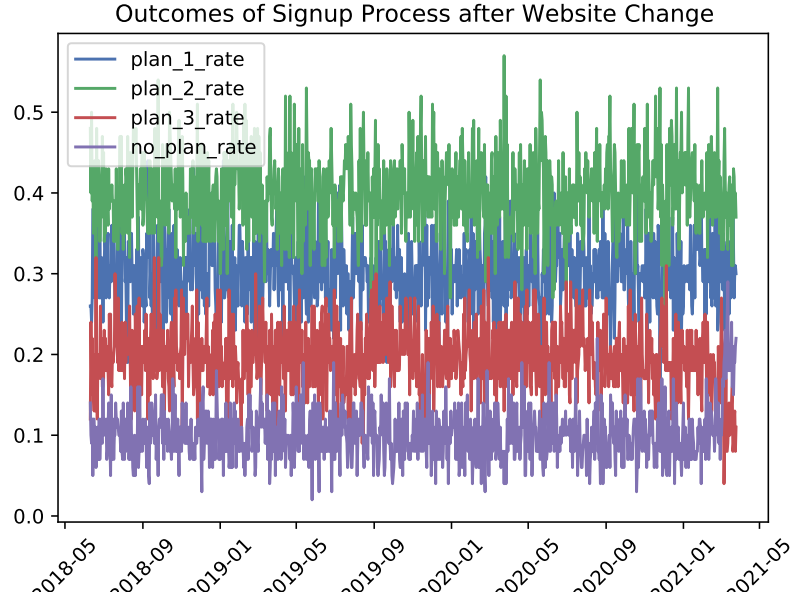
Websites and apps collect traffic and log interactions. Your details are captured and pulled into vast aggregates of consumer data. I can route and re-route your trajectory across an online environment. Applying the same pressures to tens of thousands of others, we can trace out how the topology of particular sites throw up speed bumps on the customer's journey. Imagine we're running a website which aims to funnel customers through to a number of different purchase plans. The historic patterns are relatively stable with only 10% of customers dropping out of our conversion funnel on a daily basis. We can sample actions online (Figure 7) under differing pressures with a view to evaluating expected values of repeated coercive prompts.

```
### Multinomial distribution
m_var = stats.multinomial(n=100, p=[.3, .4, .2, .1])
m_var_sample = m_var.rvs(1000)
m_var_2 = stats.multinomial(n=100, p=[.3, .4, .1, .2])
m_var_2_sample = m_var_2.rvs(20)
base = datetime.datetime.today() - datetime.timedelta(days=1000)

df = pd.DataFrame(m_var_sample, columns=['plan_1', 'plan_2', 'plan_3', 'no_plan'])
df = df.append(pd.DataFrame(m_var_2_sample, columns=['plan_1', 'plan_2', 'plan_3', 'no_plan']), ignore_index=True)

df['totals'] = df['plan_1'] + df['plan_2'] + df['plan_3'] + df['no_plan']
df['plan_1_rate'] = df['plan_1'] / df['totals']
df['plan_2_rate'] = df['plan_2'] / df['totals']
df['plan_3_rate'] = df['plan_3'] / df['totals']
df['no_plan_rate'] = df['no_plan'] / df['totals']
date_list = [base + datetime.timedelta(days=x) for x in range(len(df))]
df.index = date_list
for i, column in enumerate([x for x in df.columns if 'rate' in x]):
    ax = plt.plot(df[column], label=column)

plt.title("Outcomes of Signup Process after Website Change")
plt.legend(loc="upper left")
xticks = plt.xticks(rotation=45)
plt.show()
```



Assume the particular values for each plan, then the expected value of customer journey is just:  $p_1\$ (o_1) + p_2\$ (o_2) + p_3\$ (o_3) + p_4\$ (o_4) = .3 * \$10 + .4 * \$7 + .2 * \$12 + .1 * \$0 = \$8.20$  Now imagine there was a change to the website and we observe the following pattern for the next 20 days. The change was made on the hypothesis ( $H+$ ) that it would bring a positive boost to revenue. How much more positive? A slight expected increase makes it harder to conclusively reject the  $H+$  even in the face of contrary indicators. If we observe the above pattern, then what is the new expected value? Have we decisively falsified  $H+$ ? From the frequentist point of view the macro distributional properties haven't significantly changed. But given what we know about the change to the website it would be foolish to accept such a static distributional assumption. Looking only at the small sample of new data, the variance will be large and the estimates of rates of sign-up for each plan will be unstable. Following the Bayesian paradigm we can condition our expectations on the new data, the old data or all the data.

#### Calculate Expected Revenue

```
def expected_revenue(posterior_samples):
    return 10*posterior_samples[:, 0] + 7*posterior_samples[:, 1] + \
        12*posterior_samples[:, 2] + 0*posterior_samples[:, 3]

full_data = df[['plan_1', 'plan_2', 'plan_3', 'no_plan']]
weird_data2 = df[['plan_1', 'plan_2', 'plan_3', 'no_plan']].tail(20)
normal_data = df[['plan_1', 'plan_2', 'plan_3', 'no_plan']].head(1000)
```



```

# Bayesian Posterior
multinomial_posterior_new = np.random.dirichlet(np.array([100, 100, 100, 100]) + np.array(weird_obs))

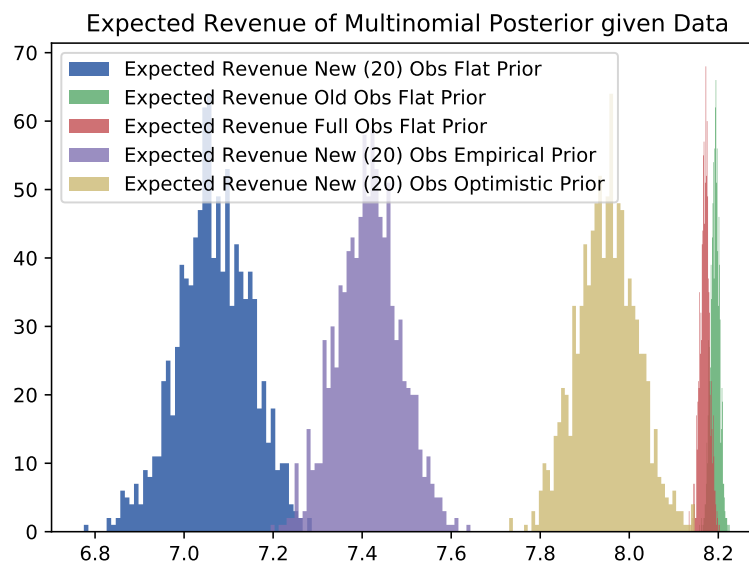
# Empirical Bayes
multinomial_posterior_full = np.random.dirichlet(np.array([100, 100, 100, 100]) + np.array(full_obs))

# Bayesian Posterior Prior Data
multinomial_posterior_old = np.random.dirichlet(np.array([100, 100, 100, 100]) + np.array(normal_obs))
multinomial_posterior_new_empirical_prior = np.random.dirichlet(np.array([300, 400, 200, 100]) + np.array(weird_obs))
multinomial_posterior_crazy_prior = np.random.dirichlet(np.array([100, 400, 500, 0]) + np.array(weird_obs))

expected_value_new = expected_revenue(multinomial_posterior_new)
expected_value_old = expected_revenue(multinomial_posterior_old)
expected_value_full_data = expected_revenue(multinomial_posterior_full)
expected_value_new_bias = expected_revenue(multinomial_posterior_new_empirical_prior)
expected_value_crazy = expected_revenue(multinomial_posterior_crazy_prior)

h1 = plt.hist(expected_value_new, histtype='stepfilled', label="Expected Revenue New (20) Obs Flat Prior")
h2 = plt.hist(expected_value_old, histtype='stepfilled', label="Expected Revenue Old Obs Flat Prior")
h3 = plt.hist(expected_value_full_data, histtype='stepfilled', label="Expected Revenue Full Obs Flat Prior")
h4 = plt.hist(expected_value_new_bias, histtype='stepfilled', label="Expected Revenue New (20) Obs Empirical Prior")
h5 = plt.hist(expected_value_crazy, histtype='stepfilled', label="Expected Revenue New (20) Obs Optimistic Prior")
plt.title("Expected Revenue of Multinomial Posterior given Data")
plt.legend(loc="upper left")
plt.show()

```



The above graph illustrates the spread in values expected revenue calculated on different slices of our data using Bayes Rule. Using a large number of observations, the influence of our priors are minimal and washed out by the data, giving us a strong point estimate with low variance stable around 8.2, but since the recent data involves a step change, we might be better off ignoring the old data. But we can also see that if we condition our expectations only on the new data with different priors drawn from the past data or hope, we can positively bias our expectations. Suppose we're naive and accept either the optimistic prior or retain a frequentist approach, and accept that the website change is associated with a slight drop in revenue, do we revert to the old website or try to explain the dip by contingencies of the market and preserve our test for another thirty days?

This pattern is not rare. Nearly all substantial decisions are made with small samples in circumstances where past behaviour is not a guide. Past behavioural patterns are exactly what we're trying to avoid or change. If you want to know whether the change on your website will drive a material change in financial revenue, you won't have long run patterns to rely on, and it's an open question on how to weight the new data. If you want to judge the long term consequences of a new symptom the same limitation of information applies. All models smuggle in a host of statistical assumptions and these can range from reasonable to absurd. Even when reasonable they're only supported by large sample sizes, and most questions of interest are driven by novelty (or specificity) that short circuits appeal to robust patterns of history. Reasoning from small samples is common, best done with caution and plenty of caveats, but better reason than not. Expectations should be modified accordingly.

## 2.4 Utility Curves

### 2.4.1 The Stakes: From Utility to Indifference

Our views of probability can flex up and down in response to facts, but it's less clear how our estimates of utility change. Too much of a good thing often tends to the bad. So we dabble, sample and share. In pursuit of variety we swap our goods, shunning stale options in favour of the novel exchange. For a given good we can differ in our appetites but it's relatively straightforward to find the point where - one more donut is one too many. While it can be a bit unclear how we should measure utility, once we've decided on a metric the mathematical characteristics are meaningful. If the scale is donuts, we can infer aspects of your attitudes from your acquisition and enthusiasm for donuts. In most cases we're interested not just in your pursuit of pastries, but how you'd be willing to trade for those pastries.

We seek competitive advantage for our own produce to balance the cost owed to the skills of others. This coordinated compromise lies at the core of maximising

subjective utility in a market, but at the limit some scenarios do not admit any admixture of goods. Not all babies can be cut in half. But in most cases a consumer will try to optimise their bundle of goods over an entire marketplace, preserving enough on one key good; money, to remain liquid. So, to a first approximation our utility estimate would seem to be a multivariate function.

```
def linear_utility(x, a=2, b=4):
    # This works fine on smaller numbers
    return a + b*x

def quadriatic_utility(x, b=0.5):
    return x - (b)*(x**2)

def logarithmic_utility(x, a, b):
    return np.log(a) + b*np.log(x)

def negative_exp_utility(x, c=.5):
    return -np.exp(-(c*x))

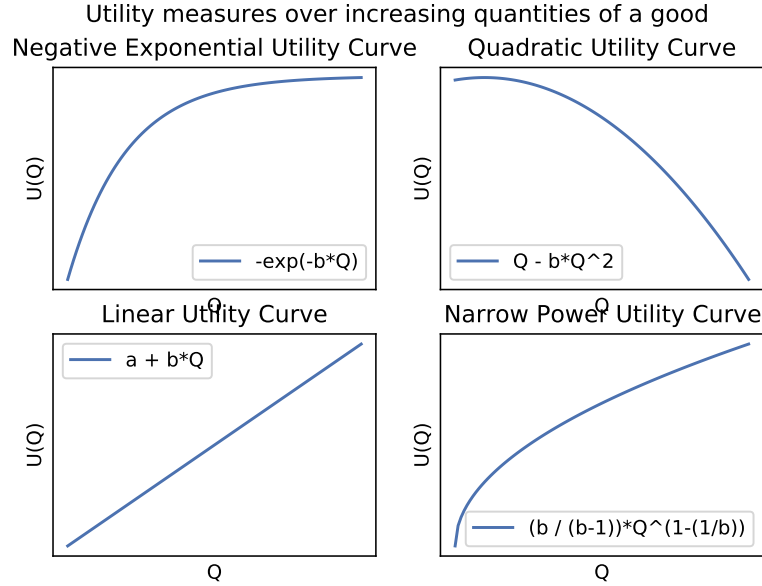
def narrow_power(x, B=2):
    return (B / (B - 1))*(x**(1 - (1/B)))

xdata = np.linspace(0, 10, 100)
negExp = negative_exp_utility(xdata, 0.5)
quad = quadriatic_utility(xdata, 0.5)
narrow_pow = narrow_power(xdata, 2)
lin = linear_utility(xdata, 2, 3)

# Create plot
fig, axes = plt.subplots(2, 2)
labels = ["-exp(-b*Q)", "Q - b*Q^2", "a + b*Q", "(b / (b-1))*Q^(1-(1/b))"]
titles = ["Negative Exponential Utility Curve", "Quadratic Utility Curve",
          "Linear Utility Curve", "Narrow Power Utility Curve"]
y = [negExp, quad, lin, narrow_pow]

for i, ax in enumerate(np.hstack(axes)):
    ax.plot(xdata, y[i], label=labels[i])
    ax.set_title(titles[i])
    tmp = ax.set_xticks([])
    tmp = ax.set_yticks([])
    ax.set_ylabel("U(Q)")
    ax.set_xlabel("Q")
    ax.legend()

fig.suptitle("Utility measures over increasing quantities of a good")
plt.show()
```



$$u(\mathbf{g}) = f(g_0, g_1 \dots g_n)$$

There are number of ways we can specify a utility function as seen in Figure 10, but a typical example is the Cobb-Douglas function.

$$u(\mathbf{g}) = g_0^{\alpha_0} g_1^{\alpha_1} \dots g_n^{\alpha_n} \text{ where } \forall i \sum \alpha_i = 1$$

Then taking the case of two goods  $g_1, g_2$  we can determine an indifference curve where you would be willing to exchange quantities of  $g_1$  for an agreeable amount of  $g_2$ . The task is to express the value of a given good as priced in terms of the other goods. Set

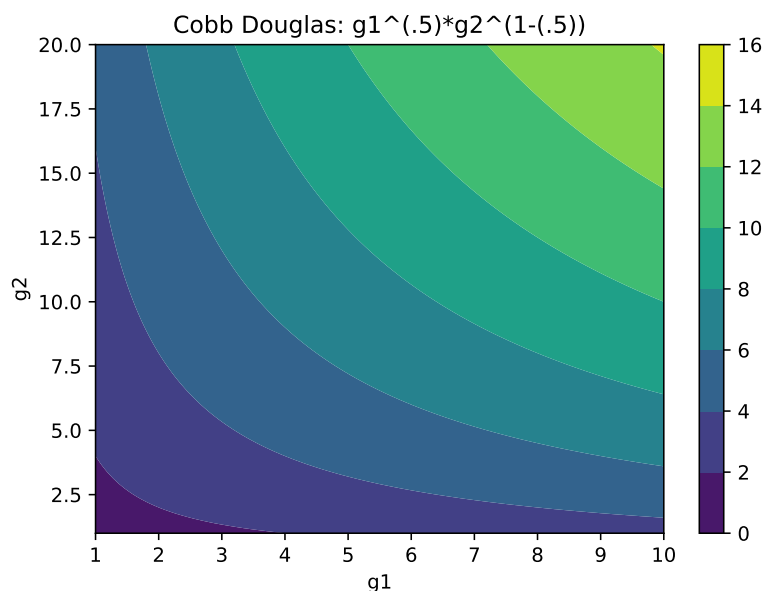
$$\begin{aligned} u(\mathbf{g}) &= k = g_1^{\frac{1}{2}} g_2^{\frac{1}{2}} = (g_1 g_2)^{\frac{1}{2}} = \sqrt{g_1 g_2} \\ \Rightarrow k^2 &= g_1 g_2 \Rightarrow \frac{k^2}{g_2} = g_1 \end{aligned}$$

Using this formula we can express how the quantities of fair exchange vary based on a fixed utility value. This is not to say that these curves represent an actual or objectively fair price, just that when measured in terms of our utility these are mappings of quantities of goods we would be happy to exchange. Your view of a fair price is encoded in your utility theory. It's at this point when utility theory can be said to verge on empirical science. If we can model your preferences as a utility function characteristic of some general attitude toward

acquisition, we might also hope to be able to predict future trades and cater for individual desires.

```
def cobb_douglas(g1, g2, a1, a2):
    return g1**a1 * g2**a2

fig, ax = plt.subplots()
g1 = np.linspace(1, 10, 100)
g2 = np.linspace(1, 20, 100).reshape((100, 1))
contours = ax.contourf(g1, g2.flatten(), cobb_douglas(g1, g2, .5, .5))
tmp = fig.colorbar(contours)
ax.set_xlabel("g1")
ax.set_ylabel("g2")
ax.set_title("Cobb Douglas:  $g1^{(.5)}g2^{(1-(.5))}$ ")
plt.show()
```



Contrast the case of a schizophrenic's utility as they try to weight aspects of their health. Even if you could measure each dimension simultaneously, what scale captures the worth of autonomy and measures the trade off against clarity, delusion and paranoia? How do you plan for children in the fear of what your genetics could seed? Do you fear the harm done through inheritance or the looming lifelong responsibility of care for an affected child? Even adoption foists onto the child the burden of having you as a parent. The disease calibrates the cost, but the equations are hard to solve.<sup>5</sup>

<sup>5</sup>See "The Choice of Children" in (Wang, 2019)

## 2.5 Optimising Utility

A further complication arises when we try to account for a consumer's budget or resources. The shape of the Cobb-Douglas function in Figure 12 shows that the utility surface is constantly increasing with our rate of acquisition. So without any constraints the consumer would not achieve satisfaction, but continue like a glutton consuming forever without cease. Add a budgetary constraint and natural trade-offs between desire and cost mean that we need to find the maximum point at which an indifference curve intersects with our budgetary line. Instead of solving the equation:

$$\text{Find } g_1, g_2 \text{ such that } u(g_1, g_2) = \lambda$$

we need to solve a constrained optimisation problem:

$$\text{maximize } u(g_1, g_2) \text{ subject to } \text{cost}(g_1, g_2) = \lambda$$

This style of problem can be approached with the method of Lagrange multipliers. If we let:

$$L = g_1^{\frac{1}{2}} g_2^{\frac{1}{3}} - \lambda(2g_1 + 3g_2 - 40)$$

where 2 and 3 are the unit cost of the respective goods, and 40 is our total budget. This  $\lambda$  is our Lagrangian multiplier - a term used to re-express the algebra of our equation as a function of the consumer's capacity to spend.

```
g1 = np.linspace(1, 10, 100)
g2 = np.linspace(1, 20, 100).reshape((100, 1))

def g1_indifference(g2, k, alpha=1/3):
    orig = k**(1/(1-alpha)) * g2**(-alpha/(1-alpha))
    return orig

def plot_indifference_curves(ax, alpha=.5):
    k = np.arange(1, 14, 3)
    ax.plot(g2, g1_indifference(g2, k, alpha))
    ax.legend(["U(g1^.5, g2^.5)" + " = {}".format(i) for i in k])
    ax.set_xlabel("g2")
    ax.set_ylabel("g1 ")
    plt.title("Indifference Curves with Budget Constraint")

def budget(g, W=50, price=2):
    return (W - g) / price

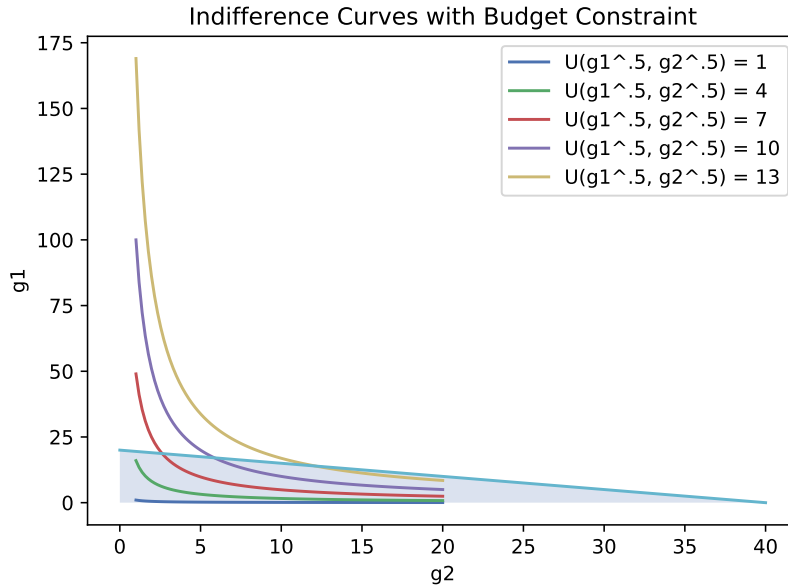
def plot_budget_constraint(ax, W=40, price=2):
```

```

g2 = np.array([0, W])
g1 = budget(g2, W, price)
ax.plot(g2, g1)
ax.fill_between(g2, 0, g1, alpha=0.2)
ax.set_xlabel("g2")
ax.set_ylabel("g1")
return ax

fig, ax = plt.subplots()
plot_indifference_curves(ax)
plot_budget_constraint(ax)

```



We can discover where utility is maximised when the gradient of the “curve” can be set to zero. This is the theory behind the “hill climbing” algorithms of gradient descent. When the curvature of the “slope” has plateaued i.e. is zero, then we’ve reached a maximum or minimum in the multivariate space of the function. As before we want to use this fact to express the implicit function of  $g_1$  in terms of  $g_2$ , but this time including the constraints on our budget.

$$\nabla L = dL/d\mathbf{g} = \left( \frac{\partial u(\mathbf{g})}{\partial g_1}, \frac{\partial u(\mathbf{g})}{\partial g_2} \right) = \left( \frac{1}{2}g_1^{-\frac{1}{2}}g_2^{\frac{1}{2}} - 2\lambda, \frac{1}{2}g_2^{-\frac{1}{2}}g_1^{\frac{1}{2}} - 3\lambda \right) = \mathbf{0}$$

$$\Rightarrow \lambda = \frac{1}{4}g_1^{-\frac{1}{2}}\sqrt{g_2} = \frac{4}{25}g_2^{-\frac{1}{2}}\sqrt{g_1}$$

$$\begin{aligned} \Rightarrow \left(\frac{1}{4}\right)^2 \frac{1}{g_1} g_2 &= \left(\frac{4}{25}\right)^2 \frac{1}{g_2} g_1 \Rightarrow \left(\frac{1}{4}\right)^2 g_2 = \left(\frac{4}{25}\right)^2 \frac{1}{g_2} g_1^2 \Rightarrow \left(\frac{1}{4}\right)^2 g_2^2 = \left(\frac{4}{25}\right)^2 g_1^2 \\ &\Rightarrow g_2 = \frac{16}{25} g_1 \end{aligned}$$

The same pattern holds for cases with more than two goods. We can express the value of given good  $g_n$  in terms of a function  $f(g_1, \dots, g_{n-1})$ . Then substituting this value into our constraint we get:

$$2g_1 + 3\left(\frac{16}{25}\right)g_1 = 40 \Rightarrow 2g_1 + 1.92g_1 = 40 \Rightarrow 3.92g_1 = 40$$

Proving the optimal settings are  $g_1^* = 10.20$  and  $g_2^* = 6.52$  and  $\lambda^* = 0.20$

The above proof shows how we triangulate a consumer's view of any good as expressed through the medium of their utility function. But the method of Lagrangian multipliers is more than a mere algebraic trick. We can interpret the  $\lambda$  term as the rate of change of the consumer's utility as a function of the cost to our resources. How taxing is our treatment, how exhausted is your wallet? The proof is a little more involved, but the significance of this interpretation should be obvious. If we knew our consumers adhered to a particular style of utility function we could model how "price-changes" would impact their returns to utility and select strategies for maximum gain. The challenge lies in deriving a customer's utility profile.



## Chapter 3

# Literature

Here is a review of existing methods.



## Chapter 4

# Methods

We describe our methods in this chapter.



## Chapter 5

# Applications

Some *significant* applications are demonstrated in this chapter.

### 5.1 Example one

### 5.2 Example two



## Chapter 6

# Final Words

We have finished a nice book.





# Bibliography

- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.
- Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.
- Freedman, D. A. (2009). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press.
- Hacking, I. (2006). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press, 2 edition.
- Lindley, D. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46.
- Savage, L. (1954). *The Foundations of Statistics*. Wiley.
- Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press.
- Wang, E. W. (2019). *The Collected Schizophrenias*. Penguin.