February 26, 2018

CS 361: Probability & Statistics

Random variables

The discrete uniform distribution

If every value of a discrete random variable has the same probability, then its distribution is called a **discrete uniform distribution**

We've used this in a number of examples: heads 1, tails 0 with a fair coin. Rolling a fair die, etc.

If there are N possible values, $P(X=x) = 1/N$ if x is one of the allowable values

Poisson distribution

A distribution useful for modeling counts of events (not sets of outcomes, just colloquial events)

The number of calls a call center receives is a random variable. If we know that a call center receives on average 100 calls per hour. What is the probability it will receive 75 or 200 in a given hour?

If whatever we are counting
1) has a fixed rate and
2) the events occur independently

Then the **Poisson distribution** will be a good model

If the known rate of occurrence of the events per time interval is given by λ

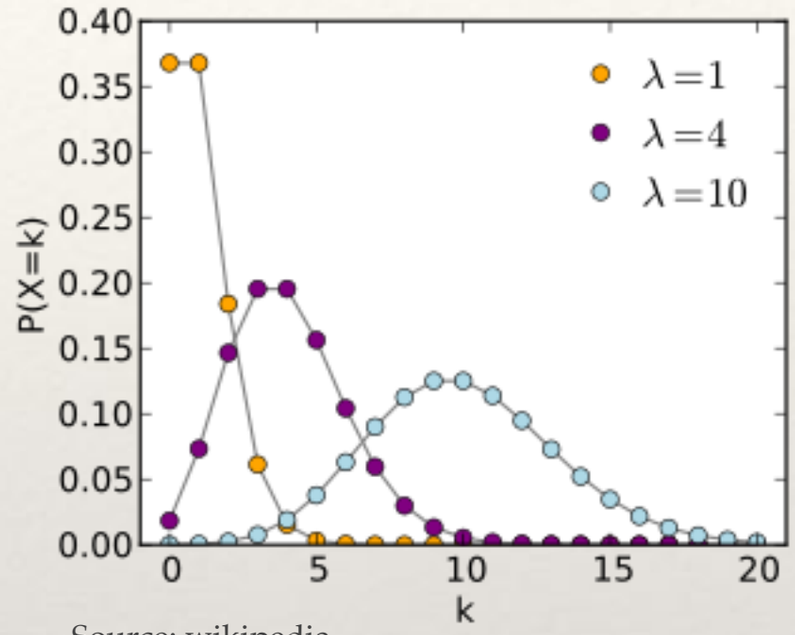
Then the probability that we will see k events in a length of time equal to the interval is given by

$$P(\{X = k\}) = \frac{e^{-\lambda}\lambda^k}{k!}$$

And we say that the count of events is a Poisson random variable with intensity λ

Poisson distribution

$$P(\{X = k\}) = \frac{e^{-\lambda}\lambda^k}{k!}$$



Source: wikipedia

1. The mean of a Poisson distribution with intensity λ is λ.
2. The variance of a Poisson distribution with intensity λ is λ (no, that's not an accidentally repeated line or typo).

Poisson examples

Number of decay events by a radioactive substance in a given time

Number of health insurance claims per month (assuming no disasters or epidemics)

We talked about counts of events in time, could just as easily use counts of events per interval of space

Roadkill per mile of road

Number of genetic mutations per 100 thousand bases as we traverse a strand of DNA

Observation:
If we double, or halve, the length of the interval, we will have to double, or halve, the intensity

Poisson point processes

A Poisson point process with intensity λ is a set of random points with the property that the number of points within an interval of length *s* is a Poisson random variable with intensity λs

Can easily generalize this to points on a plane or in space by saying that such a process in n dimensional space is one with random points in a region *D* and the number of points in any subset *s* of *D* is described by a Poisson random variable with intensity $\lambda m(s)$ where m(s) is the area or volume of *s*

A few continuous random variables

Recall: continuous probability

Continuous probability distributions or continuous random variables, are characterized by density functions

Density functions are non-negative functions that we integrate to compute probabilities.

$$P(\{X \text{ takes on a value in the range } [a, b]\}) = \int_a^b p(x)\,dx$$

Continuous uniform distribution

Uniform random variables take on any value within a range with equal probability

Density function for a random variable that's uniformly distributed in the range from l to u

$$p(x) = \begin{cases} 0 & x < l \\ 1/(u-l) & l \leq x \leq u \\ 0 & x > u \end{cases}$$

Exponential distribution

Suppose we have an infinite interval of time or space, with points randomly distributed on it. Assume these points form a Poisson point process with intensity λ

The distance between two consecutive points will be a random variable. X is continuous and described by an **exponential distribution** which has a density given by

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Useful for modeling failures. If failures form a Poisson point process in time and we have just observed a failure, the distribution of times until the next failure will be given by an exponential distribution

Exponential distribution

For an exponential random variable, we have

1. The mean is $\frac{1}{\lambda}$.

2. The variance is $\frac{1}{\lambda^2}$.

So if we are describing call center calls as a Poisson point process with intensity λ per hour. The number of calls per hour is given by the Poisson distribution. The expected number of calls per hour is λ. The time until the next call is given by an exponential distribution and the expected time until the next call is $\frac{1}{\lambda}$

Standard normal distribution

Definition: 6.6 *Standard Normal distribution*

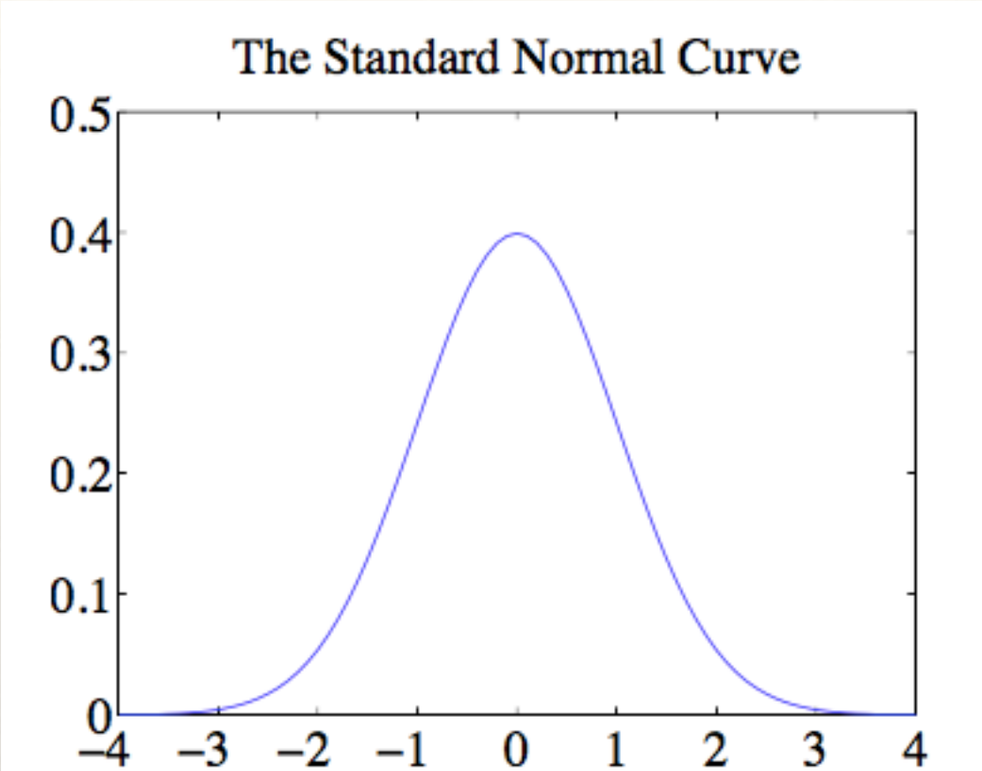
The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right).$$

is known as the **standard normal distribution**

The random variable whose density is given by p is a **standard normal random variable**

Standard normal distribution



Recall our work with histograms and datasets

This was what our standard normal data looked like when we plotted it

Useful Facts: 6.9 *standard normal distribution*

1. The mean of the standard normal distribution is 0.
2. The variance of the standard normal distribution is 1.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

Normal distribution

If we have a random variable X with a mean of μ and a standard deviation of σ and the random variable we get by taking

$$\frac{X-\mu}{\sigma}$$

is a standard normal random variable, then X is called a **normal random variable** and its density is given by

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Normal distribution

Useful Facts: 6.10 *normal distribution*

The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

has

1. mean μ
2. and variance σ.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

These random variables and distributions are also commonly called **Gaussian** distributions or random variables

Central limit theorem

Normal distributions occur extremely often in real datasets. Anything that behaves like a binomial distribution with a lot of trials will produce a normal distribution. More on this in a minute

Another result which we will not prove is the **central limit theorem** which states that if you add together many independent random variables with the same distribution, no matter the distribution, the resulting sum will be a random variable which has a distribution close to the normal distribution

Things like height, your IQ, etc can be viewed as the sum of bunch of independent random variables which may be why these phenomena usually follow a normal distribution

Normal data

Data that is drawn from a normal distribution tends to be close to the mean in terms of the number of standard deviations. Integrating the standard normal density, we see, that if we observe a bunch of standard normal data:

$$\frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left(-\frac{x^2}{2}\right)dx \approx .68$$

Around 68% of the data will be within 1 standard deviation of the mean

$$\frac{1}{\sqrt{2\pi}}\int_{-2}^{2}\exp\left(-\frac{x^2}{2}\right)dx \approx .95$$

Around 95% of the data will be within 2 standard deviations of the mean

$$\frac{1}{\sqrt{2\pi}}\int_{-3}^{3}\exp\left(-\frac{x^2}{2}\right)dx \approx .99$$

Around 99% of the data will be within 3 standard deviations of the mean

Binomial approximation

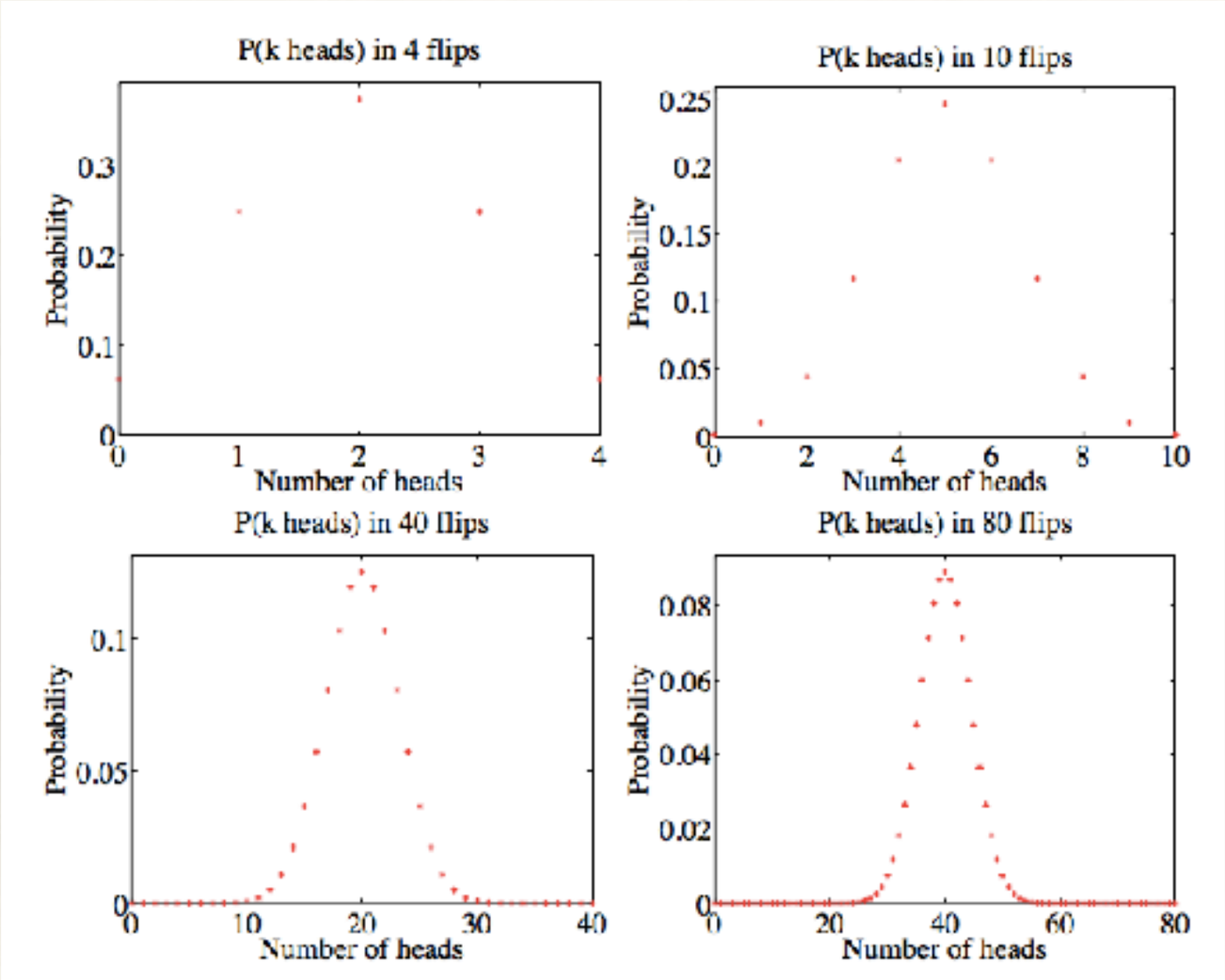
The binomial distribution is pretty straightforward with one caveat. Suppose we have a very large number of trials. If N is the number of trials, p is the probability of a success and $q = 1\text{-}p$ is the probability of a failure, then the probability distribution of X the random variable which counts the number of successes h is

$$P(h) = \frac{N!}{h!(N-h)!} p^h q^{(N-h)}$$

The caveat is that factorials grow rather quickly, so computing this probability can be difficult for large N as it can give us numerical overflows

We will construct an approximation that allows us to evaluate the probability that the number of successes lies within some range

Binomial approximation



Simulating some coin flips or looking at enough real-world binomially distributed data gives the impression that perhaps the binomial distribution looks enough like the normal distribution for that to potentially be useful

Binomial approximation

Finally, recall the definition of a Bernoulli trial and random variable: it takes value 0 with probability 1-p and value 1 with probability p

A Binomial random variable gives the probability of *h* successes in *N* Bernoulli trials or the sum of *N* independent Bernoulli random variables

The central limit theorem told us that the distribution of the sum of independent random variables is approximately normally distributed

All roads point to the normal distribution as being a good approximation for the Binomial distribution

Approximating with a normal

It turns out this is not an accident. So long as p isn't too close to 1 or 0 and N is large, the binomial distribution is well approximated by a normal distribution

If *h* is the number of heads in *N* coin flips where heads occurs with probability *p* and tails occurs with probability *q=1-p* and we write

$$x = \frac{h - Np}{\sqrt{Npq}}$$

Then for large *N*, the probability distribution of *x*, *P(x)*, is well approximated by the standard normal density *p(x)*

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right)$$

in the sense that

$$P(\{x \in [a,b]\}) \approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-u^2}{2}\right) du$$

Inference

Inference

We're going to begin the section on inference. Inference at its most broad is the process of drawing principled conclusions from data

In many of the settings we will consider we will only have data and not necessarily know about the underlying probability distributions that generated the data

In the 4 upcoming chapters we will look at: how to draw general conclusions from a specific subset of data, how to assess the significance of the evidence that exists against a hypothesis about our data, how to tell if two groups are experiencing different effects, and how to infer a probability model from a dataset