Lecture 34: Calculus and Statistics

A **random variable** X is a variable that can take one of many values depending on the outcome of some random process.

For example, C could be the random variable that represents the number of heads after flipping a coin twice. Then the probability of getting 0 for C is $P(C=0)=\frac{1}{4}$. Likewise, $P(C=1)=\frac{1}{2}$ and $P(C=2)=\frac{1}{2}$. If a random variable X is discrete, taking on integer values, it must always be true that

$$\sum_k P(X=k)=1;$$

in other words, the sum of the probabilities of all possible outcomes is 1 (100%).

The **expected value** of a random variable X, denoted $E[X]$, is the mean, or our single best guess ("expectation") for what any given value of X might be. For a discrete random variable X, the expectation is

$$\sum_k k \cdot P(X=k),$$

1 For our variable C evaluates to $0\cdot P(C=0)+1\cdot P(C=1)+2\cdot P(C=2)=0\cdot\frac{1}{4}+1\cdot\frac{1}{2}+2\cdot\frac{1}{4}=1$; in other words, if we flip a coin twice many times, on average we'll get 1 head for each pair of flips.

The **variance** of a random variable X, denoted $\mathrm{Var}[X]$, is a measure of how spread apart a distribution is (how likely we are to get values of X that are far from the mean). It is

$$\sum_k (k-E[X])^2 P(X=k),$$

which for our variable C evaluates to $(0-1)^2P(C=0)+(1-1)^2P(C=1)+(2-1)^2P(C=2)=\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$.

Note that we are basically just adding up $(k-E[X])^2$ (which is 0 when k is at the mean of X and gets larger when k gets further away from the mean) and weighting it by the probability of having $X=k$.

Now let's consider a continuous random variable X that could take on any real number.

2 As an example, H might be the height, in inches, of a randomly chosen Harvard student. Note that since H is continuous, $P(H=h)=0$ for any particular height h, so it makes more sense to talk about $P(h\le H<h+\epsilon)$ for some small $\epsilon>0$.

For a random variable, the **probability density function** for that random variable is a function $f(x)$ such that

$$P(a\le X\le b)=\int_a^b f(x)\,dx.$$

You can probably see where we are going with this: instead of the sum of all probabilities being 1, it will instead be the case that

$$\int_{-\infty}^{\infty} f(x)\,dx=1.$$

Similarly, the expected value of a continuous distribution X is

$$E[X]=\int_{-\infty}^{\infty} x f(x)\,dx$$

and the variance is

$$\mathrm{Var}[X]=\int_{-\infty}^{\infty}(x-E[X])^2 f(x)\,dx.$$

1 Our height distribution H would have what is called a **normal probability density function**. (Many natural quantities follow a normal distribution.)

The normal probability density function is

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}.$$

If X has a normal distribution, then

$$E[X]=\int_{-\infty}^{\infty} x f(x)\,dx=\mu \qquad \text{and} \qquad \mathrm{Var}[X]=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx=\sigma^2.$$

2 Not every distribution is normal, though. For example, incomes are not normally distributed: most people have relatively moderate incomes, but no one has a negative income and there are a few people that have very high incomes. Some people (see below) argue that income follows an **exponential distribution**, a distribution with probability distribution function $f(x)=\lambda e^{-\lambda x}$ (where $\lambda>0$ is a constant).

Exponential distributions have mean

$$E[X]=\int_0^{\infty} x\lambda e^{-\lambda x}\,dx=\frac{1}{\lambda}$$

and variance

$$\mathrm{Var}[X]=\int_0^{\infty}\left(x-\frac{1}{\lambda}\right)^2\lambda e^{-\lambda x}\,dx=\frac{1}{\lambda^2}.$$

Another quantity of interest is the **cumulative distribution function**, which is

$$F(x)=P(X\le x)=\int_0^x f(x)\,dx.$$

For our income function, the mean household income in the US in 1997 is a random variable I that is exponentially distributed with mean \$35,200[1], so $\lambda=1/35200$. The probability that a randomly selected person makes \$100,000 or less is

$$P(I\le 100000)=\int_0^{100000}\frac{1}{35200}e^{-t/35200}\,dt=1-e^{-100000/35200}=94\%.$$

3 The probability density function $f(x)=(a-1)/x^a$ represents a **power law distribution**, where $a>1$ is a constant parameter that changes the shape of the distribution.

[1] Dragulescu & Yakovenko (2001). Evidence for the exponential distribution of income in the USA. **The European Physical Journal B, 20, 585-559.**

Homework

1 The **uniform distribution on** $[a, b]$ is a distribution where any real number between a and b is equally likely to occur. The probability density function is $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and 0 elsewhere. Verify that $f(x)$ is a valid probability density function (i.e., check that it integrates to 1).

2 Find the mean of the uniform distribution on $[a, b]$.

3 Explain why the mean you found in problem 2 makes sense intuitively.

4 The **Cauchy distribution** is important in physics. It has a probability density function of

$$f(x) = \frac{1}{\pi} \frac{b}{(x-m)^2 + b^2}.$$

Verify that $f(x)$ is a valid probability density function.

5 Find the cumulative distribution function $F(x)$ for the Cauchy distribution.