Chapter 17 Probability Models

Models are not exact representations of reality. A good model is a useful approximation.

Why use a model? Why not use the actual distribution?
1. A model is compact. Saying that SAT scores are approximately $N(500,100)$ is much more compact than giving the entire set of SAT scores.
2. A model is easier to work with than the raw data and can sometimes lead to simple conclusions.
3. Occasionally, there are theoretical reasons why a model should be the correct one. However, you have to be very careful of what theory says should be and what actually is. For now, we’ll use the normal model only when it appears to be justified from the data.

There are several models for random phenomena that are so common that they have names. These will be discussed here.

**Bernoulli Model:** A Bernoulli trial is a random phenomenon where there are only two possible outcomes (generically called “success” and “failure”). Independent Bernoulli trials are independent repetitions of the random phenomenon where the probability of success (called $p$) stays the same over all the trials. This is the model used for coin tosses. It can be used as a model for the outcome of many betting games where the two outcomes are “win” and “lose.” It can be used for many random phenomena where there are more than two possible outcomes, but where we are only concerned with whether some particular event happens or not. For example, roll a die and see whether or not you get an even number. Randomly choose an adult and determine if they are currently married or not, or intend to vote for Bush or not, or whether their annual income is above $30,000 or not.

Random sampling from a population and observing a binary response variable is not precisely like Bernoulli trials. Even though the draws are independent, the probability of a “success” changes because random sampling is drawing without replacement. If I draw 10 cards without replacement from a deck of cards and observe whether or not each card is an Ace, these are not independent Bernoulli trials. However, if the population is large (like all adults in the U.S.) and the sample relatively small (less than 10% of the population), then the sample can be treated like independent Bernoulli trials without much loss. That is, the Bernoulli trials model will still be an acceptably good model (remember, models are never perfect representations of reality, anyway).

**Binomial Model**
Suppose we have independent Bernoulli trials with constant probability of success $p$. However, suppose now that the number of trials $n$ is fixed and that the random variable $X$ is the number of successes in the $n$ trials. Then $X$ follows a Binomial model. The binomial model has two parameters: $n$ and $p$. It’s denoted by $\text{Binom}(n,p)$.

What are the possible values of $X$?

The binomial probability model is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,\ldots, n,$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{and} \quad q = 1 - p.$$
The mean is $\mu = E(X) = np$ and the standard deviation is $\sigma = SD(X) = \sqrt{npq}$.

Example: Tiger again. Suppose we buy 5 boxes of cereal. What’s the expected number of Tigers? What’s the standard deviation?

What’s the probability that we get exactly one Tiger?

What’s the probability that we get two or fewer Tigers?

![Binomial Distribution](image)

**Deriving the mean and standard deviation of a binomial random variable**

To derive the mean and standard deviation for the binomial model, start with a single Bernoulli trial with probability of success $p$. Let $X$ be the number of successes on this single trial so $X$ is either 1 or 0 and it’s 1 with probability $p$ and 0 with probability $q=1-p$. $X$ is sometimes said to be a Bernoulli random variable.

What are $E(X)$ and $Var(X)$?

Now, suppose we have $n$ independent Bernoulli trials. The number of successes on the $n$ trials is $Y = X_1 + X_2 + \ldots + X_n$ where each of the $X$’s is a Bernoulli trial. So,

$$E(Y) =$$

$$Var(Y) =$$

$$SD(Y) =$$
The Normal model

The normal distribution is an idealized model that is often used for distributions that are unimodal and roughly symmetric: “mound-shaped”. There are often advantages to using a model for a distribution rather than the distribution itself. One is that a normal model is completely characterized by two parameters, $\mu$ and $\sigma$, which are the mean and standard deviation of the normal model. That is, there is a normal model for every possible value of $\mu$ and for every value of $\sigma > 0$. A model is not useful unless it is flexible enough to be used in a variety of situations and, by choosing the values of $\mu$ and $\sigma$, we can use the normal distribution to model SAT scores and heights of U.S. adult women in inches, as long as the distributions are mound-shaped. A normal model is abbreviated $N(\mu, \sigma)$. Thus, SAT math (or verbal) scores may be well-modeled by a $N(500,100)$ model while heights of women (in inches) may be well-modeled by a $N(67,2.5)$ model.

The $N(0,1)$ model is called the standard normal model (or distribution). If a data distribution follows a $N(\mu, \sigma)$ model, then if we standardize the data values, by $z = \frac{y - \mu}{\sigma}$, then the standardized values will follow a $N(0,1)$ model.
The 68-95-99.7 Rule Revisited (Z-scores)
One of the most useful results for the normal model is that about 68% of the values fall within one standard deviation of the mean, about 95% fall within two standard deviations of the mean, and about 99.7% fall within three standard deviations of the mean.

SAT scores: N(500,100) model

According to the empirical rule, about 68% of SAT scores fall between what two values?

About 95% of scores fall between what two values?

About 99.7% of scores (almost all) fall between what two values?

About what percentage of SAT scores are below 400? Above 700?

What percentile is 400? What percentile is 700?

The Probability Distribution Function for the Normal Model

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Using the Normal Model to Approximate the Binomial Model
If $n$ in the Binomial model is large, then the calculation of the binomial probabilities can get pretty difficult.

Example: Suppose I take a 100-question, multiple-choice test for which there are 4 choices on every problem. What’s the probability I will get more than 30 correct?

Calculating all the binomial probabilities would be tedious and difficult. However, it turns out that for large $n$, the binomial probabilities are well approximated by a normal model. The mean and standard deviation of the normal model are the mean and standard deviation of the binomial distribution you’re trying to approximate.

For the Binom(100,.25) model, $\mu = 100(.25) = 25$ and $\sigma = \sqrt{100(.25)(.75)} = 4.33$. Therefore, if the normal approximation is adequate (we’ll see how to check whether it is, below):

$$P(X > 30) \approx$$

The normal approximation is generally considered adequate if $np$ and $nq$ are both greater than or equal to 10. Check this condition for the example above:

The Poisson model
In the binomial model, what if $n$ is large, but $p$ is so small that the normal approximation to the binomial cannot be used? Another model, the Poisson, can be used as an approximation to the binomial model in this case. The Poisson is commonly used for events that occur at a low rate over a large amount of time or space.

Example: If there are $3 \times 10^9$ basepairs in the human genome and the mutation rate per generation per basepair is $10^{-9}$, what is the probability that a baby will have no mutations? One mutation?

We could use the binomial model for these outcomes. What assumption do we make if we do? What are $n$ and $p$?
Try using the binomial model to compute these probabilities:

The Poisson model has one parameter, which is traditionally called $\lambda$. The probability function is

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The expected value is $E(X) = \lambda$ and the standard deviation is $SD(X) = \sqrt{\lambda}$. To use the Poisson approximation to the binomial, use $\lambda = np$.

Calculate the probabilities of 0, 1, 2, 3 mutations using the Poisson model.

The Poisson model turns out to be the theoretically correct model for events that occur randomly over time or space and we count the number of events in equal-sized intervals of time or space.

Example: If hummingbirds arrive at a flower at a rate of $\lambda$ per hour, how many visits are expected in $x$ hours of observation and what is the variance in this expectation? If significantly more variance is observed than expected, what might this tell you about hummingbird visits?

Example: (From Romano) If bacteria are spread across a plate at an average density of 5000 per square inch, what is the chance of seeing no bacteria in the viewing field of a microscope if this viewing field is $10^{-4}$ square inches? What, therefore, is the probability of seeing at least one cell?
Example:
FLYING-BOMB HITS ON LONDON (576 cells with 537 hits)*

<table>
<thead>
<tr>
<th># of hits</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of cells with # of hits above</td>
<td>229</td>
<td>211</td>
<td>93</td>
<td>35</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td><strong>Poisson Fit</strong></td>
<td>226.7</td>
<td>211.4</td>
<td>98.5</td>
<td>30.6</td>
<td>7.1</td>
<td>1.6</td>
</tr>
</tbody>
</table>

From Time magazine, Sept. 4, 1989

The actual fit of the Poisson for the data is surprisingly good. It is interesting to note that most people believed in a tendency of the points of impact to cluster. If this were true there would be a higher frequency of areas with either no impacts or many impacts and a deficiency in the intermediate classes. The table indicates randomness and homogeneity of the area. We have here an instructive illustration of the fact that to the untrained eye randomness appears as regularity or tendency to cluster.