CSE 103: Probability and statistics **Fall 2010**

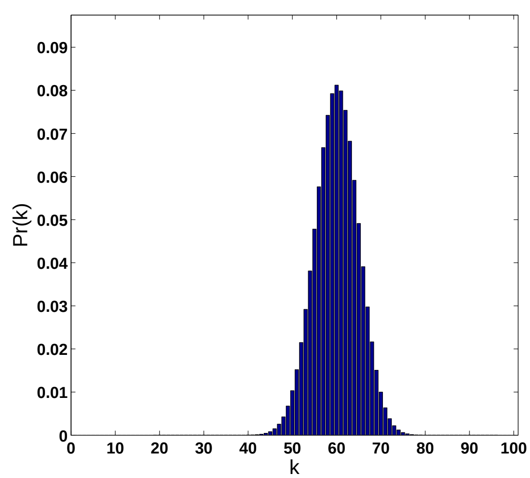
Topic 6 — Sampling, hypothesis testing, and the central limit theorem

6.1 The binomial distribution

Let X be the number of heads when a coin of bias p is tossed n times. The distribution of X is so fundamental to probability and statistics that it merits a special name, the *binomial*(n,p) distribution. Here it is, precisely:

$$\begin{aligned} \Sigma &= \{0,1,\ldots,n\} \\ \Pr(X=k) &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

The figure below shows the binomial$(100, 0.6)$ distribution.



The key property of this distribution is that it is tightly concentrated around $np = 60$. This is the mean of X, as can be seen by writing it in the form $X = X_1 + \cdots + X_n$, where each $X_i \sim B(p)$; in words, X_i is 1 if the ith coin toss comes up heads.

How spread out is the distribution? To quantify this, we need to compute the variance of X. Once again, the representation of X as the sum of X_i's comes in handy, coupled with the following useful fact.

Fact 1. *If $Z_1, \ldots, Z_n$ are independent, then* $\mathrm{var}(Z_1 + \cdots + Z_n) = \mathrm{var}(Z_1) + \cdots + \mathrm{var}(Z_n)$.

Each X_i has variance $p(1-p)$, so X has variance $np(1-p)$ and thus standard deviation $\sqrt{np(1-p)}$.

Fact 2. *If $X \sim$ binomial(n,p), then:*

$$\begin{aligned} \mathbb{E}(X) &= np \\ \mathrm{var}(X) &= np(1-p) \\ \mathrm{stddev}(X) &= \sqrt{np(1-p)} \end{aligned}$$

There's another very useful fact about the binomial that comes up over and over again: *about 95% of the distribution lies within two standard deviations of the mean.* That is to say,

$$\Pr\left(np - 2\sqrt{np(1-p)} \leq X \leq np + 2\sqrt{np(1-p)}\right) \;\approx\; 0.95. \tag{6.1}$$

We'll see how this approximation arises a little later on. As an example of its use, consider the figure shown above for binomial$(100, 0.6)$. The standard deviation is $\sigma = \sqrt{100(0.6)(0.4)} \approx 5$. And indeed almost all the distribution lies in the range $np \pm 2\sigma = 60 \pm 10$.

The remarkable thing about (6.1) is that although X can potentially take on $n+1$ possible values, it effectively stays within a range of size just $O(\sqrt{n})$; this is miniscule compared to $n+1$ for large n.

6.2 Hypothesis testing

6.2.1 Testing a vaccine

Suppose there is a certain disease that cattle can contract; and that any given cow has a 25% chance of contracting it within the course of a year. A new serum is proposed, and we want to test it to see how well it works. So we choose n cows at random, inject them with the serum, and then keep an eye on them over the next year. Let's say K of them remain healthy.

One possibility is that the serum has no effect whatsoever; call this hypothesis H_0. If this hypothesis is true, then the chance of infection is exactly the same for the cows that were injected as it is for the cow population at large, and thus $K \sim$ binomial$(n, 0.75)$.

We hope that the experiment yields a large value of K; this will provide evidence against hypothesis H_0. But how large does K need to be? It depends on how much *statistical confidence* we want in our assertion that H_0 is false. A common goal is a *95% confidence level*.

Suppose we have a sample of $n = 100$ cows. The binomial$(100, 0.75)$ distribution has mean 75 and standard deviation $\sigma = \sqrt{100(0.25)(0.75)} \approx 4.3$. Using equation (6.1), we can deduce that *if H_0 were true*, we would expect (with 95% confidence) a value of K in the range 75 ± 8.6. If K exceeds this, that is if $K > 83$, then we can reject H_0 with a 95% confidence level.

6.2.2 A blood pressure drug

Suppose a new blood pressure drug is proposed, and to test it, a sample of n patients is selected. Their blood pressure is measured with and without the drug.

Person	B.P. with drug	B.P. without drug
1	x_1	x_1'
2	x_2	x_2'
$\vdots$	$\vdots$	$\vdots$
n	x_n	x_n'

The ith trial is called a success if $x_i < x_i'$; let K be the total number of successes. How should the evidence be evaluated?

Let H_0 be the hypothesis that the drug is useless. Under this hypothesis, $K \sim$ binomial$(n, 0.5)$. Thus K has expected value $0.5n$ and standard deviation $\sqrt{n(0.5)(0.5)} = 0.5\sqrt{n}$. As we noted above, with 95% confidence, we'd expect K to be within two standard deviations of its mean, that is, $K = 0.5n \pm \sqrt{n}$.

So if K exceeds $0.5n + \sqrt{n}$, we can declare that H_0 is invalidated with 95% confidence, implying that the drug is not entirely useless.

6.3 Sampling

We live in an age where the government and the mass media are continuously polling the public. They wish to know the fraction of people who like sushi, think Obama is doing a good job, support Tiger Woods, think God exists, feel pessimistic about the future, smoke, think the war in Iraq is unnecessary, etc. This constant polling affects decision-making at every level. It advises politicians on what to say and do. It helps companies decide how to most effectively advertise their products. It helps investors decide where to entrust their money. And so on.

In each such case, there is an unknown probability p that is sought: the fraction of people who like sushi, for instance. Determining this fraction exactly would require asking *everyone*, which is prohibitively expensive. So instead a sample of n people is chosen at random, and they are each asked whether they like sushi. Let K be the number of positive responses. Then K has the binomial(n, p) distribution, with expected value np and variance $np(1-p)$.

An estimate of the fraction of sushi-lovers is K/n. Notice that

$$\begin{aligned}
\mathbb{E}\left(\frac{K}{n}\right) &= p \\
\mathrm{var}\left(\frac{K}{n}\right) &= \frac{\mathrm{var}(K)}{n^2} = \frac{p(1-p)}{n} \\
\mathrm{stddev}\left(\frac{K}{n}\right) &= \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

A standard rule of thumb, given in equation (6.1) is that the estimate K/n will lie within two standard deviations of its expected value with 95% probability. Thus we can assert with 95% confidence that the true fraction of people who like sushi is

$$\frac{K}{n} \pm 2\sqrt{\frac{p(1-p)}{n}}.$$

But this doesn't make sense, since we don't know what p is. A quick fix is to notice that $p(1-p) \leq 1/4$ (the maximization can be done by calculus, for instance), and thus a valid 95% confidence interval is

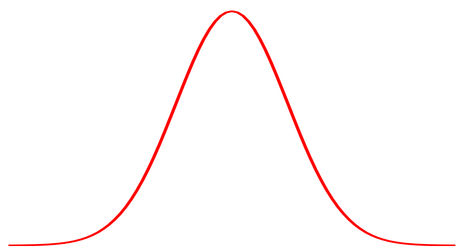
$$\frac{K}{n} \pm \frac{1}{\sqrt{n}}.$$

Thus a sample size of 2,500 gives an estimate that is accurate to within $1/\sqrt{n} = 0.02$ (at 95% confidence), whereas a sample size of 10,000 gives an accuracy of 0.01.

What if we want a higher level of confidence, like 99%, or even 99.9%? Equation (6.1) is no longer helpful; we need similar approximations for other confidence levels. It turns out that these are easy to obtain, for any desired confidence level, using the *normal approximation to the binomial distribution.*

6.4 The normal distribution

You are probably reading these notes in a library or cafeteria. Take a look at the 100 or so people nearest to you. If you were to plot the heights of all the men, it would look kind of like this:

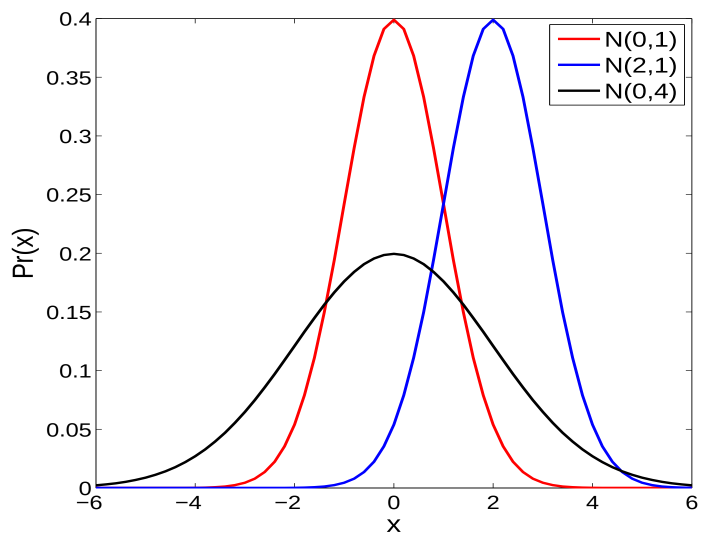


If you were to plot the blood pressure of all the women, it would look much the same. Or if you plotted the velocities of the molecules in the air, once again, same thing. This one distribution is everywhere. For that reason it is called the *normal* distribution.

The normal distribution with mean μ and variance σ^2 is denoted $N(\mu, \sigma^2)$. Unlike most of the examples we've seen so far, it is a *continuous* distribution, a density over the real line. The density at any point x is

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

Don't worry too much about this specific functional form. Instead, let's look at pictures of three different normal distributions.



The red curve (or, if you're reading this in black and white, the tall curve in the middle) is $N(0, 1)$: the normal distribution with mean 0 and variance 1. This is sometimes called the *standard normal*. All other normal distributions are simply scaled and translated versions of it. For instance, the blue curve (the one on the right) is $N(2, 1)$, which is obtained by shifting the standard normal two units to the right. The black curve (the wide one in the middle), on the other hand, is $N(0, 4)$, obtained by scaling the standard normal by a factor of two. Here's a summary of the relationship between different normal distributions.

Fact 3. *If* $X \sim N(\mu, \sigma^2)$ *then* $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

6.4.1 Sums of independent random variables

Here's an interesting fact about the normal distribution.

Fact 4. *If X, Y are independent, with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.*

In words, if a bunch of independent random variables are each normally distributed, and you add them up, the result still has a normal distribution.

In fact, something much more spectacular is true: even if the independent random variables are *not* normally distributed, their sum looks like a normal distribution!

Fact 5 (Central limit theorem)**.** *Suppose $X_1, \ldots, X_n$ are independent with mean μ and variance σ^2. Then for sufficiently large n, the distribution of $(X_1 + \cdots + X_n)/n$ is approximately $N(\mu, \sigma^2/n)$. Equivalently, the distribution of $X_1 + \cdots + X_n$ is approximately $N(n\mu, n\sigma^2)$.*

This astonishing fact helps explain why the normal distribution is observed all over the place.

6.4.2 Tails of the normal

In sampling and hypothesis testing, we ultimately end up adding independent random variables; more on this below. Thus the result is well-approximated by the normal distribution, and by analyzing the tails of this distribution, we can get intervals for any desired level of confidence.

Suppose $X \sim N(\mu, \sigma^2)$. If we want a 95% confidence interval, we seek a value of z for which

$$\Pr(\mu - z \leq X \leq \mu + z) = 0.95.$$

The answer, it turns out, is $z = 2\sigma$. If we want a 99% confidence interval, we seek a value of z for which

$$\Pr(\mu - z \leq X \leq \mu + z) = 0.99.$$

In this case, $z = 3\sigma$.

Here is a summary of some key facts about tails of the normal.

Fact 6. *If $X \sim N(\mu, \sigma^2)$, then:*

- *With 66% probability, X lies between $\mu - \sigma$ and $\mu + \sigma$.*
- *With 95% probability, X lies between $\mu - 2\sigma$ and $\mu + 2\sigma$.*
- *With 99% probability, X lies between $\mu - 3\sigma$ and $\mu + 3\sigma$.*

6.5 Sampling revisited

6.5.1 Polls with yes/no answers

From the population at large n people are chosen at random and are asked a particular question ("Do you approve of the corporate bailouts?"). The goal is to estimate the fraction p of the underlying population that would answer yes. Let's say that K of the sampled people say yes; then $K \sim \text{binomial}(n, p)$, and thus a reasonable estimate of p is K/n.

To analyze the situation further, it helps to approximate the binomial by a normal distribution. Notice that

$$K = X_1 + X_2 + \cdots + X_n$$

where the X_i are independent $B(p)$ random variables ("did the ith person say yes?"). By the central limit theorem,

$$\frac{K}{n} \text{ is distributed like } N\left(p,\ \frac{p(1-p)}{n}\right).$$

Writing $\sigma = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}$, we can conclude from Fact 6 that the estimate K/n is accurate within

- $\frac{1}{2\sqrt{n}}$ with at least 66% probability.
- $\frac{1}{\sqrt{n}}$ with at least 95% probability.
- $\frac{3}{2\sqrt{n}}$ with at least 99% probability.

And other confidence intervals are also easy to obtain, from standard tables of the normal distribution.

6.5.2 Polls with numeric answers

Many polls demand numeric rather than yes/no answers. How many glasses of wine do you drink daily? What is your salary? How old are you? And so on.

Suppose, for instance, that we wish to assess the overall educational levels (number of years of schooling) of residents of San Diego. If we were to consider *all* San Diegans of age ≥ 25, this would be a distribution with some unknown mean μ and standard deviation σ. We would like to estimate μ, the average educational level. So we randomly pick n people of age ≥ 25 and ask them how many years they have spent in school. Suppose their answers are $X_1, \ldots, X_n$.

The empirical mean (the mean of the samples) is

$$M = \frac{X_1 + \cdots + X_n}{n}$$

and the empirical standard deviation is

$$S = \sqrt{\frac{(X_1 - M)^2 + \cdots + (X_n - M)^2}{n}}.$$

It is reasonable to use M as an estimate of μ. What kind of confidence interval can we give for it?

This is where the central limit theorem once again comes in. Since the X_i are independent, we can assert that

$$M \text{ is distributed like } N\left(\mu, \frac{\sigma^2}{n}\right).$$

Thus we immediately have the following confidence intervals, from Fact 6.

- With 95% probability, M is lies between $\mu - \frac{2\sigma}{\sqrt{n}}$ and $\mu + \frac{2\sigma}{\sqrt{n}}$.
- With 99% probability, M lies between $\mu - \frac{3\sigma}{\sqrt{n}}$ and $\mu + \frac{3\sigma}{\sqrt{n}}$.

Thus, for instance, when using M to estimate μ, we can be 95% certain that it is accurate within $\pm\frac{2\sigma}{\sqrt{n}}$. But wait: we don't know σ. So instead, it is customary to use the empirical standard deviation instead, and to assert that M is accurate within $\pm\frac{2S}{\sqrt{n}}$, with 95% confidence.

Returning to the schooling example, suppose we poll $n = 400$ people, and find that the empirical average and standard deviation are $M = 11.6$ and $S = 4.1$, respectively. Then we can assert with 95% confidence that the average number of years of schooling of San Diegans is

$$11.6 \pm \frac{2 \times 4.1}{\sqrt{400}} = 11.6 \pm 0.41.$$