Read M&M §5.1, but ignore two starred parts at the end. Read M&M §5.2. Skip M&M §5.3. Sampling distributions of counts, proportions and averages. Binomial distribution. Normal approximations.

1. Means and variances

Let X be a random variable, defined on a sample space S, taking values $x_1, x_2, \ldots, x_k$ with probabilities $p_1, p_2, \ldots, p_k$. Definitions:

$$\text{mean of } X = \mathbb{E}X = \mu_X = p_1x_1 + p_2x_2 + \ldots + p_kx_k = \sum_i X(s_i)\mathbb{P}\{s_i\}$$

$$\text{variance of } X = \text{var}(X) = \sigma_X^2 = \sum_j p_j(x_j - \mu_X)^2 = \sum_i (X(s_i) - \mu_X)^2\mathbb{P}\{s_i\}$$

where the last sum in each line runs over all outcomes in S. The standard deviation σ_X is the square root of the variance.

Facts

For constants α and β and random variables X and Y:

$\mu_{X+Y} = \mu_X + \mu_Y$,

$\mu_{\alpha+\beta X} = \alpha + \beta\mu_X$,

$\sigma^2_{\alpha+\beta X} = \beta^2\sigma_X^2$.

Particular case $\text{var}(-X) = \text{var}(X)$. Variances cannot be negative.

2. Independent random variables

Two random variables X and Y are said to be ***independent*** if "knowledge of the value of X takes does not help us to predict the value Y takes", and vice versa. More formally, for each possible pair of values x_i and y_j,

$$\mathbb{P}\{Y = y_j \mid X = x_i\} = \mathbb{P}\{Y = y_j\},$$

that is,

$$\mathbb{P}\{Y = y_j \text{ AND } X = x_i\} = \mathbb{P}\{Y = y_j\} \times \mathbb{P}\{X = x_i\} \qquad \text{for all } x_i \text{ and } y_j,$$

and in general, events involving only X are independent of events involving only Y:

$$\mathbb{P}\{\text{something about } X \text{ AND something else about } Y\} = \mathbb{P}\{\text{something about } X\} \times \mathbb{P}\{\text{something else about } Y\}$$

This factorization leads to other factorizations for independent random variables:

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y) \qquad \text{if } X \text{ and } Y \text{ are independent}$$

or in M&M notation:

$$\boxed{\mu_{XY} = \mu_X\mu_Y \qquad \text{if } X \text{ and } Y \text{ are independent}}$$

3. Variances of sums of independent random variables

Standard errors provide one measure of spread for the disribution of a random variable. If we add together several random variables the spread in the distribution increases, in

general. For independent summands the increase is not as large as you might imagine: it is not just a matter of adding together standard deviations. The key result is:

$$\boxed{\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \qquad \text{if } X \text{ and } Y \text{ are independent random variables}}$$

If $Y = -Z$, for another random variable Z, then we get

$$\sigma^2_{X-Z} = \sigma^2_X + \sigma^2_{-Z} = \sigma^2_X + \sigma^2_Z \qquad \text{if } X \text{ and } Z \text{ are independent}$$

Notice the plus sign on the right-hand side: subtracting an independent quantity from X cannot decrease the spread in its distribution.

A simlar result holds for sums of more than two random variables:

$$\boxed{\sigma^2_{X_1+X_2+\ldots+X_n} = \sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n} \qquad \text{for independent } X_1,\ X_2,\ldots}$$

In particular, if each X_i has the same variance, σ^2 then the variance of the sum increases as $n\sigma^2$, and the standard deviation increases as $\sqrt{n}\sigma$. It is this $\sqrt{n}$ rate of growth in the spread that makes a lot of statistical theory work.

4. Concentration of sample means around population means

Suppose a random variable X has a distribution with (population) mean μ_X and (population) variance σ^2_X.

To say that random variables $X_1, \ldots, X_n$ are a ***sample from the distribution of*** X means that the X_i are independent of each other and each has the same distribution as X.

The ***sample mean*** $\overline{X} = (X_1 + \ldots + X_n)$ is also a random variable. It has expectation (that is, the mean of the mean sample mean)

$$\mathbb{E}\overline{X} = \frac{1}{n}\mathbb{E}(X_1 + \ldots + X_n) = \frac{1}{n}(\mu_X + \ldots + \mu_X) = \mu_X$$

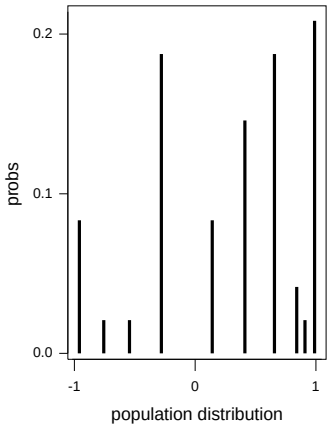
and variance

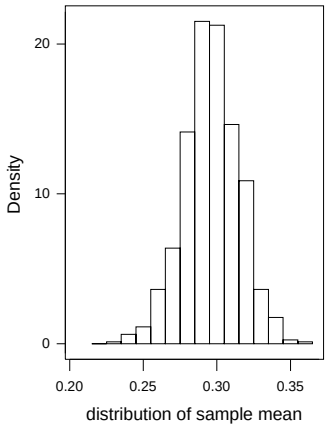
$$\begin{aligned}
\text{var}(\overline{X}) &= \left(\frac{1}{n}\right)^2 \text{var}(X_1 + \ldots + X_n) \\
&= \left(\frac{1}{n}\right)^2 \left(\sigma^2_X + \ldots + \sigma^2_X\right) \qquad \text{by independence of the } X_i \\
&= \frac{\sigma^2_X}{n}
\end{aligned}$$

That is, $\overline{X}$ is centered at the population mean, with spread—as measured by its standard deviation—decreasing like $1/\sqrt{\text{sample size}}$. The sample mean gets more and more concentrated around μ_X as the sample size increases. Compare with the ***law of large nmbers*** (M&M pages 328–332).

The Minitab command **random** with subcommand **discrete** will generate observations from a discrete distribution that you specify. (Menus: Calc→Random Data→Discrete) As an illustration, I repeatedly generated samples of size 1000 from the discrete population distribution shown on the left of the next picture. For each sample, I calculated the sample mean. The histogram (for 800 repetitions of the sampling experiment) gives a good idea of the distribution of the sample mean.

Notice that the distribution for $\overline{X}$ looks quite different from the population distribution. If I were to repeat the experiment another 800 times, the corresponding histogram would be slightly different. As the number of repetitions increases, for fixed sample size, the histogram settles down to a fixed form.





sample size	1000
number of repetitions	800
population mean	0.295766
population standard deviation	0.616357
sample mean	0.295770
sample standard deviation	$0.600818/\sqrt{1000}$

5. The central limit theorem

Not only does the distribution of the sample mean tend to concentrate about the population mean μ_X, with a decreasing standard deviation $\sigma_X/\sqrt{n}$, but also the shape of its distribution settles down, to become more closely normal. Under very general conditions, the distribution of $\overline{X}$ is well approximated $N(\mu_X, \sigma_X/\sqrt{n})$. The distribution of the recentered and rescaled sample mean, $\sqrt{n}(\overline{X} - \mu_X)/\sigma_X$, becomes more closely approximated by the standard normal.

6. The Binomial distribution

Suppose a coin has probability p of landing heads on any particular toss. Let X denote the number of heads obtained from n independent tosses. The random variable X can take values $0, 1, 2 \ldots, n$. It is possible (see M&M pages 387–389) to show that

$$\mathbb{P}\{X = k\} = \frac{n \times (n-1) \times \ldots \times (n-k+1)}{k \times (k-1) \times \ldots \times 1} p^k (1-p)^{n-k} \qquad \text{for } k = 0, 1, 2 \ldots, n$$

Such a random variable is said to have a ***Binomial distribution***, with parameters n and p, or $Bin(n, p)$ for short.

M&M use the abbreviation $B(n, p)$.

Mean and variance of the Binomial distribution

We can write X as a sum $X_1 + X_2 + \ldots + X_n$, where $X_i = \begin{cases} 1 & \text{if } i\text{th toss lands heads} \\ 0 & \text{if } i\text{th toss lands tails} \end{cases}$

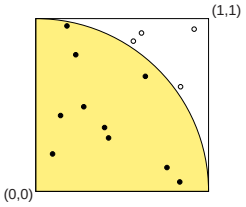
Each X_i takes the value 1 with probability p and 0 with probability $1-p$, giving a mean of $1 \times p + 0 \times (1-p) = p$. Thus $\mu_X = \mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n} = np$. Sound reasonable? Each X_i has variance $p \times (1-p)^2 + (1-p) \times (0-p)^2 = p(1-p)$. By independence of the X_i, the sum of the X_i has variance $\text{var}(X_1) + \ldots + \text{var}(X_n) = np(1-p)$. Notice that the independence between the X_i is **not** used for the calculation of the mean, but it is used for the calculation of the variance.

The proportion of heads in n tosses equals sample mean, $(X_1 + \ldots + X_n)/n$, which, by the central limit theorem, has an approximate $N(p, \sqrt{p(1-p)/n})$ distribution.

7. Monte Carlo

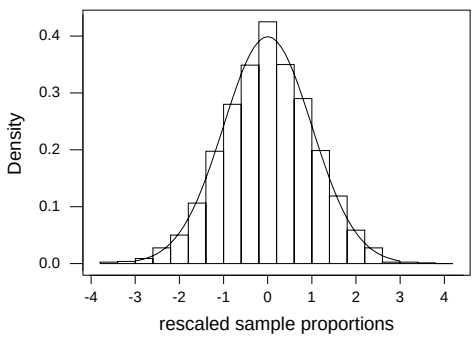
Using a computer and the Binomial distribution, one can determine areas.

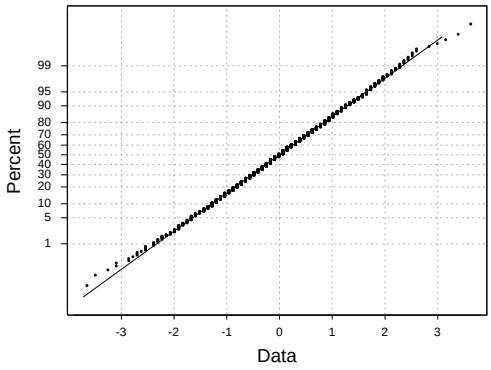
The quarter circle shown in the picture occupies a fraction $p = \pi/4 \approx 0.785397$ of the unit square. I pretended I did not know that fact. I repeatedly generated a large number of points in the unit square using Minitab, and calculated the proportion that landed within the circle. The coordinates of each point came from the Minitab command Random, with subcommand Uniform. I saved a bunch of instructions in a file that I called monte.mac. (You can retrieve a copy of this file from a link on the Syllabus page at the Statistics 101–106 web site.) I was running Minitab from a directory *D:\stat100\Lecture5* on my computer. The file monte.mac was sitting in the directory *D:\stat100\macros*. I had to tell Minitab how to find the ***macro*** file (.. means go up one level). When I typed in a percent sign, followed by the path to my monte.mac file, Minitab excuted the commands in the file.



You might find macro files useful if you want to carry out simulations.

```
 [output slightly edited]
MTB > %..\macros\monte
Executing from file: ..\macros\monte.MAC
Data Display
     sample size          1000
     number of repetitions   2000
     true p              0.785397
     true std.dev         0.0129826
     sample proportion     0.785152  ## mean of all 2000 proportions
     sample std. dev.      0.0126428  ## from the 2000 proportions
```





Left: Histogram of 'standardized' proportions, standard normal density superimposed. Right: normal plot.

The pictures and the output from monte.mac show how the distribution of proportions in 2000 samples, each of size 1000, is concentrated around the true p with an approximately normal shape. For the pictures I subtracted the true p from each sample proportion, then divided by the theoretical standard deviation, $\sqrt{p(1-p)/n}$. The central limit theorem says that the resulting 'standardized' proportions should have approximately a standard normal distribution. What do you think?

In practice, one would take a single large sample to estimate an unknown proportion p, then invoke the normal approximation to derive a measure of precision.