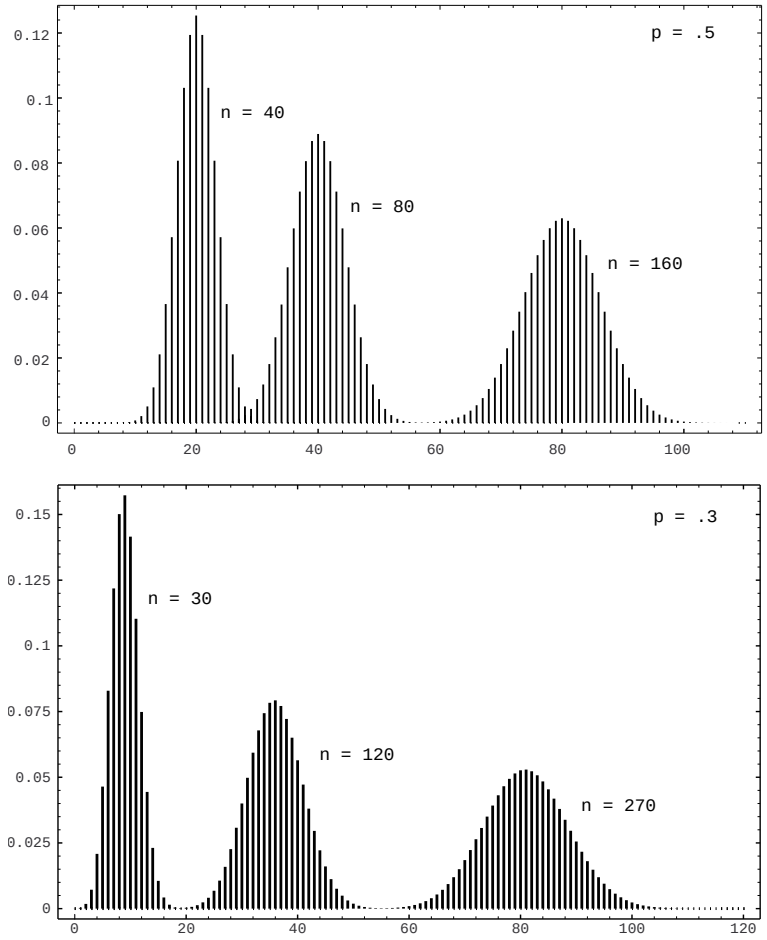
Central Limit Theorem: Bernoulli Trials

May 12, 2006

- Consider a Bernoulli trials process with probability p for success on each trial.
- Let $X_i = 1$ or 0 according as the ith outcome is a success or failure, and let $S_n = X_1 + X_2 + \cdots + X_n$.
- Then S_n is the number of successes in n trials.
- We know that S_n has as its distribution the binomial probabilities $b(n, p, j)$.


p = .5
n = 40
n = 80
n = 160
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
p = .3
n = 30
n = 120
n = 270
0.15
0.125
0.1
0.075
0.05
0.025
0
0
20
40
60
80
100
120

Standardized Sums

- We can prevent the drifting of these spike graphs by subtracting the expected number of successes np from S_n.
- We obtain the new random variable $S_n - np$.
- Now the maximum values of the distributions will always be near 0.
- To prevent the spreading of these spike graphs, we can normalize $S_n - np$ to have variance 1 by dividing by its standard deviation $\sqrt{npq}$

Definition

- The *standardized sum* of S_n is given by

$$S_n^* = \frac{S_n - np}{\sqrt{npq}} \ .$$

- We plot a spike graph with the spikes placed at the possible values of S_n^*: x_0, x_1, . . . , x_n, where

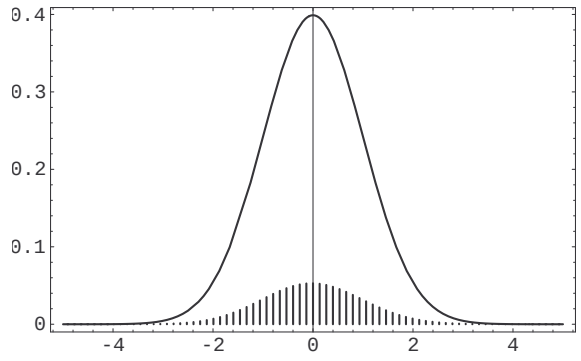
$$x_j = \frac{j - np}{\sqrt{npq}} \ .$$

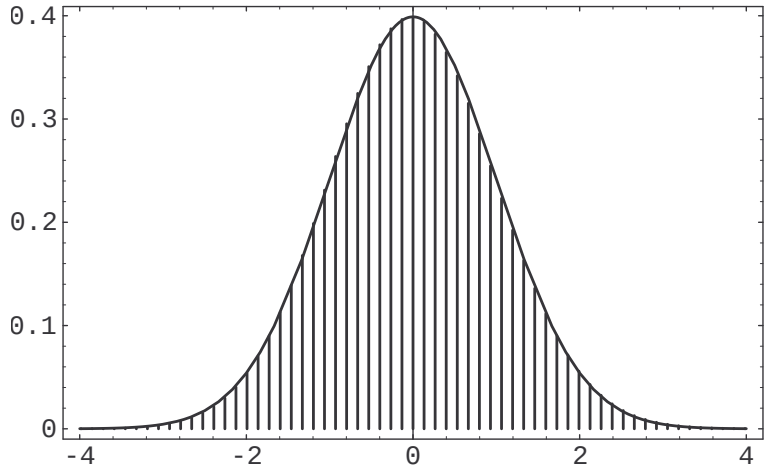
- We make the height of the spike at x_j equal to the distribution value $b(n, p, j)$.

- We plot a spike graph with the spikes placed at the possible values of S_n^*: x_0, x_1, . . . , x_n, where

$$x_j = \frac{j - np}{\sqrt{npq}} \ .$$

- We make the height of the spike at x_j equal to the distribution value $b(n, p, j)$.


0.4
0.3
0.2
0.1
0
-4
-2
0
2
4


0.4
0.3
0.2
0.1
0
-4
-2
0
2
4

- Let us fix a value x on the x-axis and let n be a fixed positive integer.

- Then the point x_j that is closest to x has a subscript j given by the formula

$$j = \langle np + x\sqrt{npq} \rangle \ .$$

- Thus the height of the spike above x_j will be

$$\sqrt{npq}\, b(n, p, j) = \sqrt{npq}\, b(n, p, \langle np + x_j\sqrt{npq} \rangle) \ .$$

Central Limit Theorem for Binomial Distributions

Theorem. *For the binomial distribution $b(n, p, j)$ we have*

$$\lim_{n\to\infty} \sqrt{npq}\, b(n, p, \langle np + x\sqrt{npq}\rangle) = \phi(x) \ ,$$

where $\phi(x)$ is the standard normal density.

Approximating Binomial Distributions

- To find an approximation for $b(n, p, j)$, we set

$$j = np + x\sqrt{npq}$$

- Solve for x

$$x = \frac{j - np}{\sqrt{npq}} \, .$$

$$b(n, p, j) \approx \frac{\phi(x)}{\sqrt{npq}}$$

$$= \quad \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right) .$$

Example

- Let us estimate the probability of exactly 55 heads in 100 tosses of a coin.
- For this case $np = 100 \cdot 1/2 = 50$ and $\sqrt{npq} = \sqrt{100 \cdot 1/2 \cdot 1/2} = 5$.
- Thus $x_{55} = (55 - 50)/5 = 1$ and

$$\begin{aligned} P(S_{100} = 55) \sim \frac{\phi(1)}{5} &= \frac{1}{5}\left(\frac{1}{\sqrt{2\pi}} e^{-1/2}\right) \\ &= .0484\ . \end{aligned}$$

Central Limit Theorem for Bernoulli Trials

Theorem. *Let S_n be the number of successes in n Bernoulli trials with probability p for success, and let a and b be two fixed real numbers. Define*

$$a^* = \frac{a - np}{\sqrt{npq}}$$

and

$$b^* = \frac{b - np}{\sqrt{npq}} \ .$$

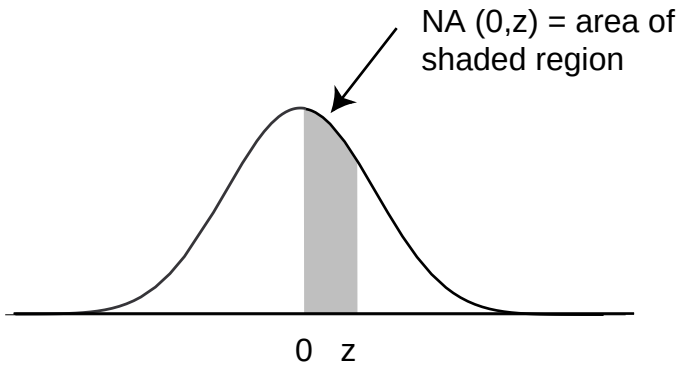
Then

$$\lim_{n \to \infty} P\left(a \leq S_n \leq b\right) = \int_{a^*}^{b^*} \phi(x)\, dx \ .$$

How to use this theorem?

- The integral on the right side of this equation is equal to the area under the graph of the standard normal density $\phi(x)$ between a and b.

- We denote this area by $NA(a^*, b^*)$.



z	NA(z)	z	NA(z)	z	NA(z)	z	NA(z)
.0	.0000	1.0	.3413	2.0	.4772	3.0	.4987
.1	.0398	1.1	.3643	2.1	.4821	3.1	.4990
.2	.0793	1.2	.3849	2.2	.4861	3.2	.4993
.3	.1179	1.3	.4032	2.3	.4893	3.3	.4995
.4	.1554	1.4	.4192	2.4	.4918	3.4	.4997
.5	.1915	1.5	.4332	2.5	.4938	3.5	.4998
.6	.2257	1.6	.4452	2.6	.4953	3.6	.4998
.7	.2580	1.7	.4554	2.7	.4965	3.7	.4999
.8	.2881	1.8	.4641	2.8	.4974	3.8	.4999
.9	.3159	1.9	.4713	2.9	.4981	3.9	.5000

Approximation of Binomial Probabilities

- Suppose that S_n is binomially distributed with parameters n and p.

$$P(i \leq S_n \leq j) \approx NA\left(\frac{i - \frac{1}{2} - np}{\sqrt{npq}}, \frac{j + \frac{1}{2} - np}{\sqrt{npq}}\right) .$$

Example

- Dartmouth College would like to have 1050 freshmen.
- The college cannot accommodate more than 1060.
- Assume that each applicant accepts with probability .6 and that the acceptances can be modeled by Bernoulli trials.
- If the college accepts 1700, what is the probability that it will have too many acceptances?

Exercise

A true-false examination has 48 questions. June has probability 3/4 of answering a question correctly. April just guesses on each question. A passing score is 30 or more correct answers. Compare the probability that June passes the exam with the probability that April passes it.

Applications to Statistics

- Suppose that a poll has been taken to estimate the proportion of people in a certain population who favor one candidate over another in a race with two candidates.

- We pick a subset of the population, called a *sample*, and ask everyone in the sample for their preference.

- Let p be the actual proportion of people in the population who are in favor of candidate A and let $q = 1 - p$.

- If we choose a sample of size n from the population, the preferences of the people in the sample can be represented by random variables $X_1,\ X_2,\ \ldots,\ X_n$, where $X_i = 1$ if person i is in favor of candidate A, and $X_i = 0$ if person i is in favor of candidate B.

- Let $S_n = X_1 + X_2 + \cdots + X_n$.

- If each subset of size n is chosen with the same probability, then S_n is hypergeometrically distributed.

- If n is small relative to the size of the population, then S_n is approximately binomially distributed, with parameters n and p.

- The pollster wants to estimate the value p. An estimate for p is provided by the value $\bar{p} = S_n/n$.

- The mean of $\bar{p}$ is just p, and the standard deviation is

$$\sqrt{\frac{pq}{n}} \, .$$

- The standardized version of $\bar{p}$ is

$$\bar{p}^* = \frac{\bar{p} - p}{\sqrt{pq/n}} \, .$$

- The distribution of the standardized version of $\bar{p}$ is approximated by the standard normal density.

- 95% of its values will lie within two standard deviations of its mean, and the same is true of $\bar{p}$.

$$P\left(p - 2\sqrt{\frac{pq}{n}} < \bar{p} < p + 2\sqrt{\frac{pq}{n}}\right) \approx .954 \ .$$

- The pollster does not know p or q, but he can use $\bar{p}$ and $\bar{q} = 1 - \bar{p}$ in their place

$$P\left(\bar{p} - 2\sqrt{\frac{\bar{p}\bar{q}}{n}} < p < \bar{p} + 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\right) \approx .954 \ .$$

- The resulting interval

$$\left(\bar{p} - \frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}},\ \bar{p} + \frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}}\right)$$

  is called the 95 percent confidence interval for the unknown value of p.

- The pollster has control over the value of n. Thus, if he wants to create a 95% confidence interval with length 6%, then he should choose a value of n so that

$$\frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}} \leq .03\ .$$

Exercise

A restaurant feeds 400 customers per day. On the average 20 percent of the customers order apple pie.

1. Give a range (called a 95 percent confidence interval) for the number of pieces of apple pie ordered on a given day such that you can be 95 percent sure that the actual number will fall in this range.

2. How many customers must the restaurant have, on the average, to be at least 95 percent sure that the number of customers ordering pie on that day falls in the 19 to 21 percent range?