Central Limit Theorem (cont'd)

11/03/2005

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.*

- Recall: The standardized sum

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

- Then

$$P(a \leq S_n \leq b) = P\left(\frac{a - np}{\sqrt{npq}} \leq S_n^* \leq \frac{b - np}{\sqrt{npq}} \right) .$$

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

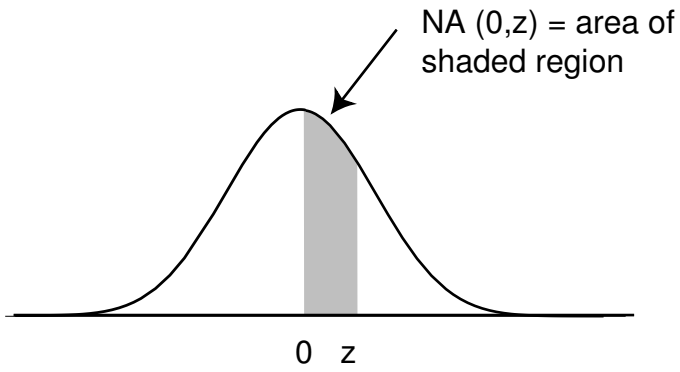
$$\lim_{n \to \infty} P\left(a \leq S_n \leq b\right) = \int_{a^*}^{b^*} \phi(x)\, dx \ .$$

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 $NA(a^*, b^*)$.

- Unfortunately, there is no simple way to integrate the function $e^{-x^2/2}$.



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

A coin is tossed 100 times. Estimate the probability that the number of heads lies between 40 and 60.

Example

A coin is tossed 100 times. Estimate the probability that the number of heads lies between 40 and 60.

The expected number of heads is $100 \cdot 1/2 = 50$, and the standard deviation for the number of heads is $\sqrt{100 \cdot 1/2 \cdot 1/2} = 5$; $n = 100$ is reasonably large

$$
\begin{aligned}
P(40 \leq S_n \leq 60) &\approx P\left(\frac{39.5-50}{5} \leq S_n^* \leq \frac{60.5-50}{5}\right) \\
&= P(-2.1 \leq S_n^* \leq 2.1) \\
&\approx NA(-2.1, 2.1) \\
&= 2NA(0, 2.1) \\
&\approx .9642 \ .
\end{aligned}
$$

Dartmouth College would like to have 1050 freshmen. This 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?

Dartmouth College would like to have 1050 freshmen. This 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?

If it accepts 1700 students, the expected number of students who matriculate is $.6 \cdot 1700 = 1020$. The standard deviation for the number that accept is $\sqrt{1700 \cdot .6 \cdot .4} \approx 20$. Thus we want to estimate the probability

$$P(S_{1700} > 1060) = P(S_{1700} \geq 1061)$$

$$
\begin{aligned}
P(S_{1700} > 1060) &= P(S_{1700} \geq 1061) \\
&= P\left(S^*_{1700} \geq \frac{1060.5 - 1020}{20}\right) \\
&= P(S^*_{1700} \geq 2.025) \ .
\end{aligned}
$$

Exercise

Let S_{100} be the number of heads that turn up in 100 tosses of a fair coin. Use the Central Limit Theorem to estimate

1. $P(S_{100} \leq 45)$.

2. $P(45 < S_{100} < 55)$.

3. $P(S_{100} > 63)$.

4. $P(S_{100} < 57)$.

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?

Central Limit Theorem for Discrete Independent Trials

- Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n independent discrete random variables of an independent trials process with common distribution function $m(x)$ defined on the integers, with mean μ and variance σ^2.

- **Standardized Sums**

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \ .$$

- This standardizes S_n to have expected value 0 and variance 1.

- If $S_n = j$, then S_n^* has the value x_j with

$$x_j = \frac{j - n\mu}{\sqrt{n\sigma^2}} \ .$$

Approximation Theorem

- Let X_1, X_2, $\ldots$, X_n be an independent trials process and let $S_n = X_1 + X_2 + \cdots + X_n$. Assume that the greatest common divisor of the differences of all the values that the X_j can take on is 1. Let $E(X_j) = \mu$ and $V(X_j) = \sigma^2$. Then for n large,

$$P(S_n = j) \sim \frac{\phi(x_j)}{\sqrt{n\sigma^2}} \ ,$$

where $x_j = (j - n\mu)/\sqrt{n\sigma^2}$, and $\phi(x)$ is the standard normal density.

Central Limit Theorem for a Discrete Independent Trials Process

- Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n discrete independent random variables with common distribution having expected value μ and variance σ^2. Then, for $a < b$,

$$\lim_{n \to \infty} P\left(a < \frac{S_n - n\mu}{\sqrt{n\sigma^2}} < b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \ .$$

Example

A die is rolled 420 times. What is the probability that the sum of the rolls lies between 1400 and 1550?

Example

A die is rolled 420 times. What is the probability that the sum of the rolls lies between 1400 and 1550?

The sum is a random variable

$$S_{420} = X_1 + X_2 + \cdots + X_{420} \ .$$

We have seen that $\mu = E(X) = 7/2$ and $\sigma^2 = V(X) = 35/12$.

Thus, $E(S_{420}) = 420 \cdot 7/2 = 1470$, $\sigma^2(S_{420}) = 420 \cdot 35/12 = 1225$, and $\sigma(S_{420}) = 35$.

$$
\begin{aligned}
P(1400 \leq S_{420} \leq 1550) &\approx P\left(\frac{1399.5-1470}{35} \leq S^*_{420} \leq \frac{1550.5-1470}{35}\right) \\
&= P(-2.01 \leq S^*_{420} \leq 2.30) \\
&\approx NA(-2.01, 2.30) = .9670 \ .
\end{aligned}
$$