Math 20, Fall 2017

Edgar Costa

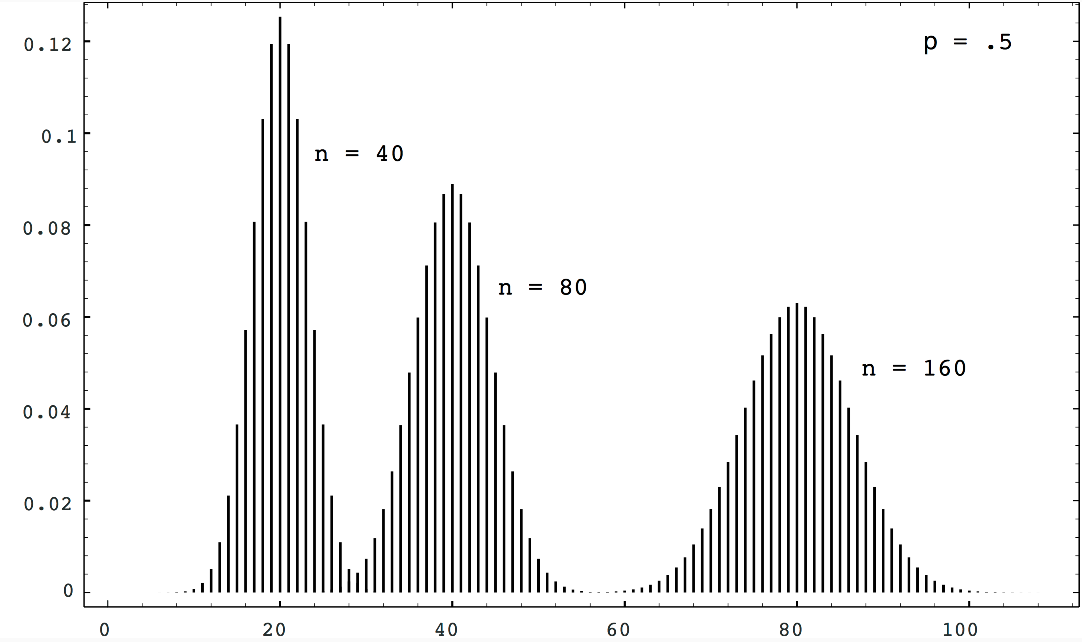
Week 7

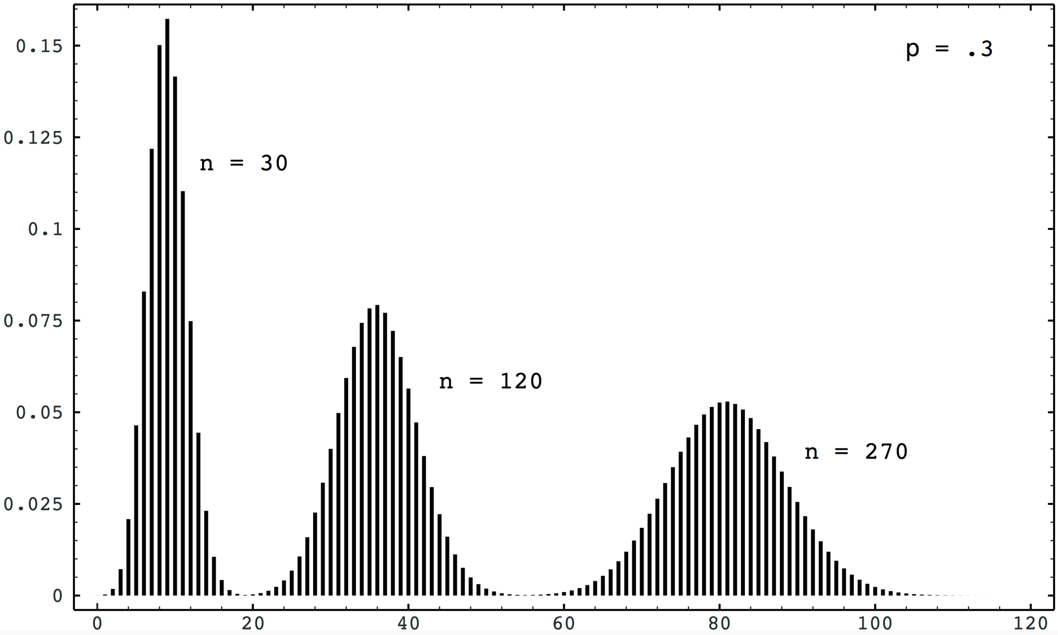
Dartmouth College

Central Limit Theorem

- Consider a Bernoulli trials process with probability p for success, i.e., a series $\{X_i\}$ of i.i.d. Bernoulli trials.
- $X_i = 1$ or 0 if the ith outcome is a success or a 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 it is distributed as a binomial distribution with parameters n and p.

$$P(S_n = j) = \binom{n}{j} p^j (1-p)^{n-j}$$


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.

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.

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 $\sqrt{npq}$. Note: it does not spread as $n \to +\infty$

Standardized Sum: Definition

The *Standardized* sum of S_n is given by

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

Note: S_n^* always has expected value 0 and variance 1.

Standardized Sums

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

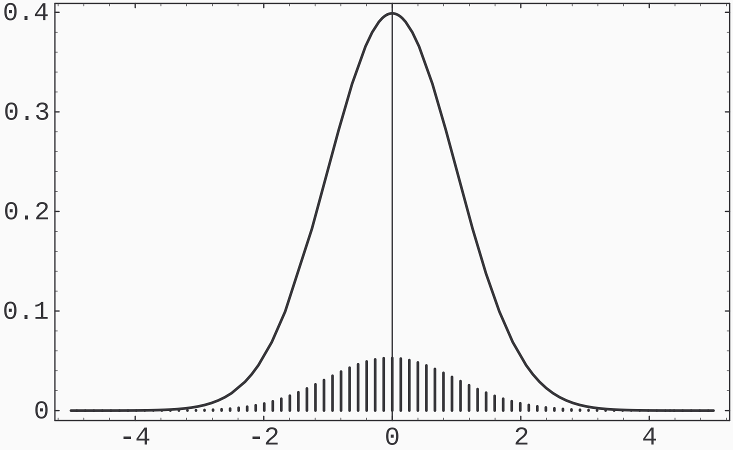
- We plot a spike graph with spikes placed at the possible values $S_n^* : x_0, x_1, \ldots, x_n$, where
$$x_j = \frac{j - np}{\sqrt{npq}}$$

Standardized Sums

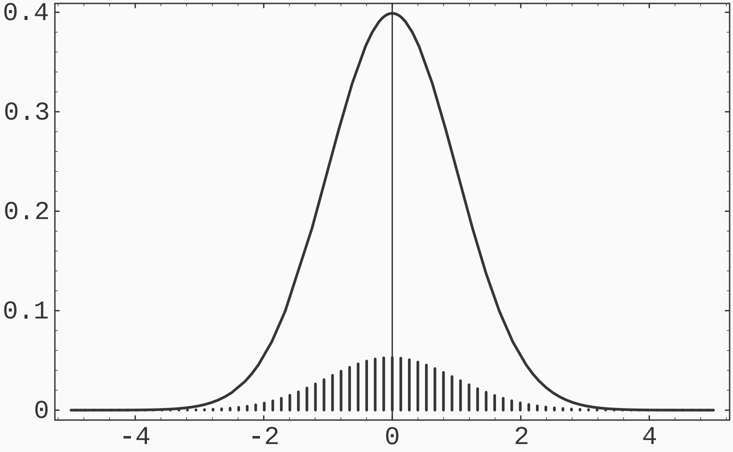
$$S_n^* = \frac{S_n - np}{\sqrt{npq}}.$$

- We plot a spike graph with spikes placed at the possible values $S_n^* : x_0, x_1, \ldots, x_n$, where
$$x_j = \frac{j - np}{\sqrt{npq}}$$
- We make the height of the spikes at x_j equal to the distribution value
$$\binom{n}{j} p^j (1-p)^{n-j}$$

Standardized Sum $n = 270$, $p = 0.3$ VS standard normal density



Standardized Sum $n = 270$, $p = 0.3$ VS standard normal density



Can we make them match?

Can we make them match?

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$g_n(x) = P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

where $j = \operatorname{round}(np + x\sqrt{npq})$

In other words, $x_j = \frac{j-np}{\sqrt{npq}}$ is the closest point of that shape close to x.

Can we make them match?

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$g_n(x) = P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

$$\text{where } j = \text{round}(np + x\sqrt{npq})$$

In other words, $x_j = \frac{j-np}{\sqrt{npq}}$ is the closest point of that shape close to x.

$$\int_{\mathbb{R}} \phi(x)\,\mathrm{d}x = 1 = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}$$

Can we make them match?

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$g_n(x) = P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

where $j = \mathrm{round}(np + x\sqrt{npq})$

In other words, $x_j = \frac{j-np}{\sqrt{npq}}$ is the closest point of that shape close to x.

$$\int_{\mathbb{R}} \phi(x)\,\mathrm{d}x = 1 = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{n} P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

Can we make them match?

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$g_n(x) = P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

where $j = \mathrm{round}(np + x\sqrt{npq})$

In other words, $x_j = \frac{j-np}{\sqrt{npq}}$ is the closest point of that shape close to x.

$$\int_{\mathbb{R}} \phi(x)\, \mathrm{d}x = 1 = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{n} P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

$$= \sum_{j=0}^{n} g_n\left(\frac{j - np}{\sqrt{npq}}\right)$$

Can we make them match?

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad g_n(x) = P\left(S_n^* = \frac{j - np}{\sqrt{npq}}\right)$$

where $j = \mathrm{round}(np + x\sqrt{npq})$

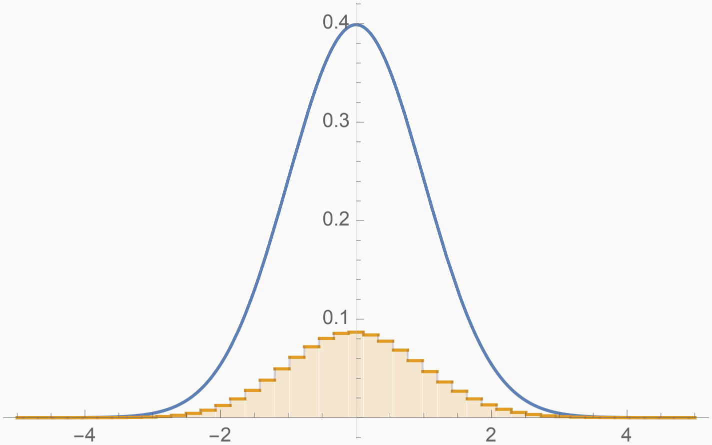
In other words, $x_j = \frac{j-np}{\sqrt{npq}}$ is the closest point of that shape close to x.

$$\int_{\mathbb{R}} \phi(x)\,\mathrm{d}x = 1 = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{n} P\left(S_n^* = \frac{j-np}{\sqrt{npq}}\right)$$

$$= \sum_{j=0}^{n} g_n\left(\frac{j-np}{\sqrt{npq}}\right) \neq \int_{\mathbb{R}} g_n(x)\,\mathrm{d}x$$

The last line is not an approximation for the integral! Why?

Standardized Sum $n = 100$, $p = 0.3$ VS standard normal density



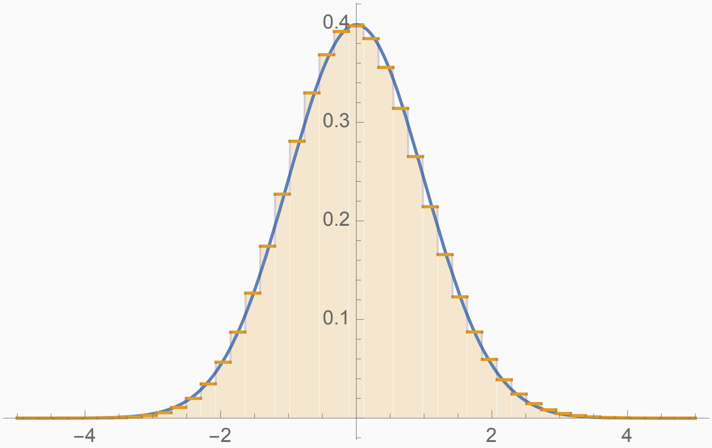
Integrating $g_n(x)$

$$
\begin{aligned}
\int_{\mathbb{R}} g_n(x)\,\mathrm{d}x &= \sum_{j=0}^{n} \frac{1}{\sqrt{npq}} g_n\left(\frac{j-np}{\sqrt{npq}}\right) \\
&= \sum_{j=0}^{n} \frac{1}{\sqrt{npq}} \binom{n}{j} p^j q^{n-j} \\
&= \frac{1}{\sqrt{npq}} \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}
\end{aligned}
$$

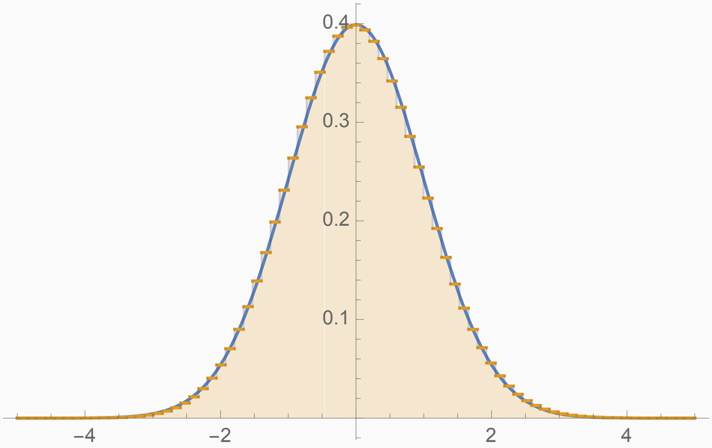
Integrating $g_n(x)$

$$\begin{aligned}\int_{\mathbb{R}} g_n(x)\,\mathrm{d}x &= \sum_{j=0}^{n} \frac{1}{\sqrt{npq}} g_n\left(\frac{j-np}{\sqrt{npq}}\right)\\ &= \sum_{j=0}^{n} \frac{1}{\sqrt{npq}} \binom{n}{j} p^j q^{n-j}\\ &= \frac{1}{\sqrt{npq}} \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}\\ &= \frac{1}{\sqrt{npq}}\end{aligned}$$

rescaled standardized Sum $n = 100$, $p = 0.3$ VS standard normal density



rescaled standardized Sum $n = 270$, $p = 0.3$ VS standard normal density



Theorem

Write $b(n, p, j) := \binom{n}{j} p^j q^{n-j}$. We have

$$\lim_{n \to +\infty} \sqrt{npq}\, b(n, p, \mathrm{round}(np + x\sqrt{npq})) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

We can prove it directly using Stirling's formula $n! \approx \sqrt{2\pi n} n^n e^{-n}$ as $n \to +\infty$.

Theorem

Write $b(n,p,j) := \binom{n}{j} p^j q^{n-j}$. We have

$$\lim_{n \to +\infty} \sqrt{npq}\, b(n,p,\mathrm{round}(np + x\sqrt{npq})) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

We can prove it directly using Stirling's formula $n! \approx \sqrt{2\pi n} n^n e^{-n}$ as $n \to +\infty$.

Challenge: try to carry this out for $x = 0$ and assuming that np is an integer.

Approximating Binomial Distributions

- To find approximations for the values of $b(n, p, j)$, we set

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

- Solve for x

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

$$\begin{aligned} b(n, p, j) &\approx \frac{\phi(x)}{\sqrt{npq}} \\ &= \frac{1}{\sqrt{npq}} \phi\left(\frac{j - np}{\sqrt{npq}}\right) . \end{aligned}$$

Example

$$b(n,p,j) \approx \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right)$$

- Let us estimate the probability of exactly 55 heads in 100 tosses of a coin.

Example

$$b(n,p,j) \approx \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right)$$

- Let us estimate the probability of exactly 55 heads in 100 tosses of a coin.
- For this case $np = 100 \cdot \frac{1}{2} = 50$ and $\sqrt{npq} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{25} = 5.$

Example

$$b(n,p,j) \approx \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right)$$

- Let us estimate the probability of exactly 55 heads in 100 tosses of a coin.
- For this case $np = 100 \cdot \frac{1}{2} = 50$ and $\sqrt{npq} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{25} = 5$.
- Thus $x = \frac{55-50}{5} = 1$ and

$$\begin{aligned} P(S_{100} = 55) &\approx \frac{\phi(1)}{5} \\ &= \frac{1}{5}\frac{1}{\sqrt{2\pi}}e^{-1/2} \\ &= 0.0483941 \end{aligned}$$

Example

$$b(n,p,j) \approx \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right)$$

- Let us estimate the probability of exactly 55 heads in 100 tosses of a coin.
- For this case $np = 100 \cdot \frac{1}{2} = 50$ and $\sqrt{npq} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{25} = 5$.
- Thus $x = \frac{55-50}{5} = 1$ and

$$\begin{aligned} P(S_{100} = 55) &\approx \frac{\phi(1)}{5} \\ &= \frac{1}{5}\frac{1}{\sqrt{2\pi}}e^{-1/2} \\ &= 0.0483941 \end{aligned}$$

- Indeed, $P(S_{100} = 55) = 0.0484743$

Poisson vs Central Limit Theorem

- We derived the Poisson distribution as an approximation to the binomial. It has its own merits and we could have derived independently of the binomial distribution.

Poisson vs Central Limit Theorem

- We derived the Poisson distribution as an approximation to the binomial. It has its own merits and we could have derived independently of the binomial distribution.
- To use it as approximation of the binomial distribution we rely on the limit:

$$(1 - \lambda/n)^{n-k} \to e^{-\lambda}$$

Thus, for it to be a good approximation we better have $p = \frac{\lambda}{n}$ close to 0.

	correct	CLT	Poisson
$k = 55$	0.0484743	0.0483941	0.042164
$k = 50$	0.0795892	0.0797885	0.056325

Poisson vs Central Limit Theorem

- We derived the Poisson distribution as an approximation to the binomial. It has its own merits and we could have derived independently of the binomial distribution.
- To use it as approximation of the binomial distribution we rely on the limit:

$$(1-\lambda/n)^{n-k} \to e^{-\lambda}$$

Thus, for it to be a good approximation we better have $p = \frac{\lambda}{n}$ close to 0.

	correct	CLT	Poisson
$k = 55$	0.0484743	0.0483941	0.042164
$k = 50$	0.0795892	0.0797885	0.056325

- Central Limit Theorem works for any p.

Central Limit Theorem for Bernoulli Trials

Theorem

Let S_n be the number of successes in n independent Bernoulli trials with probability p for success, and let a and b be two fixed real numbers, with $a < b$. Then

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - np}{\sqrt{npq}} \leq b\right) = \int_a^b \phi(x)\, dx \,.$$

Approximation of Binomial Probabilities

Suppose that S_n is binomially distributed with parameters n and p. We know how to estimate a probability of the form

$$P(i \leq S_n \leq j) \approx \sum_{k=i}^{j} \frac{1}{\sqrt{npq}} \phi\left(\frac{k - np}{\sqrt{npq}}\right).$$

Approximation of Binomial Probabilities

Suppose that S_n is binomially distributed with parameters n and p. We know how to estimate a probability of the form

$$P(i \le S_n \le j) \approx \sum_{k=i}^{j} \frac{1}{\sqrt{npq}} \phi\left(\frac{k-np}{\sqrt{npq}}\right).$$

A *slightly* more accurate approximation is given by the area under the standard normal density between the standardized values corresponding to $(i - 1/2)$ and $(j + 1/2)$. Thus,

$$P(i \le S_n \le j) \approx P\left(\frac{i - \frac{1}{2} - np}{\sqrt{npq}} \le N(0,1) \le \frac{j + \frac{1}{2} - np}{\sqrt{npq}}\right).$$

Approximation of Binomial Probabilities

Suppose that S_n is binomially distributed with parameters n and p. We know how to estimate a probability of the form

$$P(i \leq S_n \leq j) \approx \sum_{k=i}^{j} \frac{1}{\sqrt{npq}} \phi\left(\frac{k-np}{\sqrt{npq}}\right).$$

A *slightly* more accurate approximation is given by the area under the standard normal density between the standardized values corresponding to $(i - 1/2)$ and $(j + 1/2)$. Thus,

$$P(i \leq S_n \leq j) \approx P\left(\frac{i - \frac{1}{2} - np}{\sqrt{npq}} \leq N(0,1) \leq \frac{j + \frac{1}{2} - np}{\sqrt{npq}}\right).$$

But remember, at the end of the day, these are all approximations!

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 $\sqrt{100 \cdot 1/2 \cdot 1/2} = 5$.

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 $\sqrt{100 \cdot 1/2 \cdot 1/2} = 5$.

$$
\begin{aligned}
P(40 \leq S_n \leq 60) &= P(39.5 \leq S_n \leq 60.5) \qquad\qquad (= 0.9648) \\
&= P\left(\frac{39.5 - 50}{5} \leq S_n^* \leq \frac{60.5 - 50}{5}\right) \\
&= P(-2.1 \leq S_n^* \leq 2.1) \\
&\approx \int_{-2.1}^{2.1} \phi(x)\, \mathrm{d}x = 2\int_0^{2.1} \phi(x)\, \mathrm{d}x \\
&\approx 0.964271
\end{aligned}
$$

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 $\sqrt{100 \cdot 1/2 \cdot 1/2} = 5$.

$$\begin{aligned} P(40 \leq S_n \leq 60) &= P(39.5 \leq S_n \leq 60.5) \qquad (= 0.9648) \\ &= P\left(\frac{39.5 - 50}{5} \leq S_n^* \leq \frac{60.5 - 50}{5}\right) \\ &= P(-2.1 \leq S_n^* \leq 2.1) \\ &\approx \int_{-2.1}^{2.1} \phi(x)\, \mathrm{d}x = 2\int_0^{2.1} \phi(x)\, \mathrm{d}x \\ &\approx 0.964271 \end{aligned}$$

Note $\int_{-2}^{2} \phi(x)\, \mathrm{d}x = 0.9545$

Example

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?

Example

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

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

Example

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

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

Exercise

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.

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. $P(\text{april passes})$ can be approximated in many ways.

Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of *independent* and *identically distributed* random variables with expected value μ and finite variance given by σ^2.

Write $S_n = X_1 + X_2 + \cdots + X_n$.

Then for any $a < b$ two fixed real numbers, we have

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \int_a^b \phi(x)\, dx \,.$$

Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of *independent* and *identically distributed* random variables with expected value μ and finite variance given by σ^2.

Write $S_n = X_1 + X_2 + \cdots + X_n$.

Then for any $a < b$ two fixed real numbers, we have

$$\lim_{n \to \infty} P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \int_a^b \phi(x)\, dx \,.$$

Under some mild assumptions, the result above also holds without requiring the distributions to identically distributed.

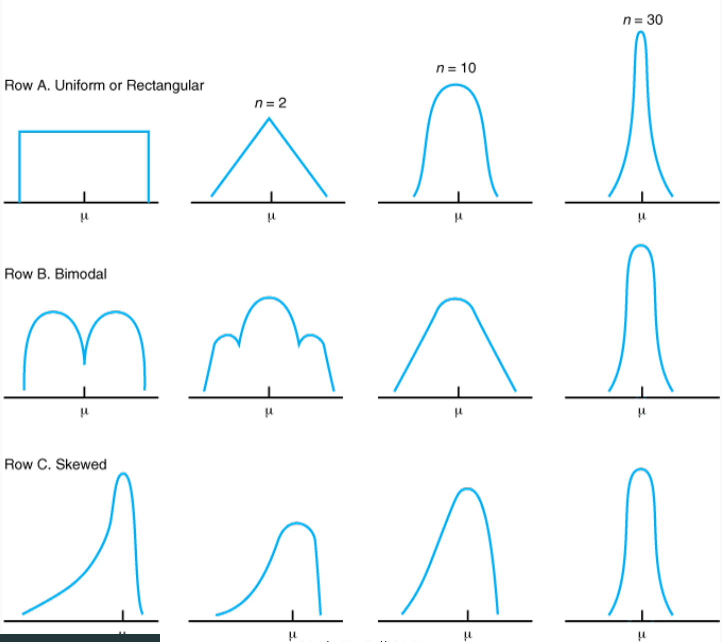
A More General Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of *independent* discrete random variables with finite expected value and variance and let $S_n = X_1 + X_2 + \cdots + X_n$. Assume that there exists a constant A such that $|X_i| \leq A$ and that $V[S_n] \to +\infty$.

Then for any $a < b$ two fixed real numbers, we have

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - E[S_n]}{\sqrt{V[S_n]}} \leq b\right) = \int_a^b \phi(x)\, dx \,.$$


Row A. Uniform or Rectangular
n = 2
n = 10
n = 30
μ
Row B. Bimodal
Row C. Skewed

Exercise

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

Exercise

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 $\mu = E[X_i] = 7/2$ and $\sigma^2 = V[X_i] = 35/12$.

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

Exercise

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 $\mu = E[X_i] = 7/2$ and $\sigma^2 = V[X_i] = 35/12$.

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

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

Application to Statistics

Application 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.

Application 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 factor 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.

Application to Statistics

- Let $S_n = X_1 + X_2 + \cdots + X_n$.
- If each subset of size n is chose with the same probability, then S_n is hypergeometric distribution.

Application to Statistics

- Let $S_n = X_1 + X_2 + \cdots + X_n$.
- If each subset of size n is chose with the same probability, then S_n is hypergeometric distribution.
- If n is small relative to the size of the population, then S_n is approximately binomially distributed, with parameters n and p.

Application to Statistics

- Let $S_n = X_1 + X_2 + \cdots + X_n$.
- If each subset of size n is chose with the same probability, then S_n is hypergeometric distribution.
- 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 $\overline{p} = S_n/n$.
- What is the mean of $\overline{p}$? and its variance?

Application to Statistics

- Let $S_n = X_1 + X_2 + \cdots + X_n$.
- If each subset of size n is chose with the same probability, then S_n is hypergeometric distribution.
- 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 $\overline{p} = S_n/n$.
- What is the mean of $\overline{p}$? and its variance?
- The standardized version of $\overline{p}$ is

$$\overline{p}^* = \frac{\overline{p} - p}{\sqrt{pq/n}}$$

Application to Statistics

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

$$P\left(p - 2\sqrt{\frac{pq}{n}} < \bar{p} < p + 2\sqrt{\frac{pq}{n}}\right) \approx 0.954$$

Application to Statistics

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

$$P\left(p - 2\sqrt{\frac{pq}{n}} < \bar{p} < p + 2\sqrt{\frac{pq}{n}}\right) \approx 0.954$$

- The pollster does not know p or q, but he can use $\overline{p}$ and $\overline{q} = 1 - \overline{p}$ in their places without too much danger. (Why?)

$$P\left(\bar{p} - 2\sqrt{\frac{\bar{p}\bar{q}}{n}} < p < \bar{p} + 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\right) \approx 0.954\ .$$

Application to Statistics

- 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.

Application to Statistics

- 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.
- 19 times out of 20, that interval will contain the true value of p.

Application to Statistics

- 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.
- 19 times out of 20, that interval will contain the true 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
$$\frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}} \leq .03\ .$$

- 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.
- 19 times out of 20, that interval will contain the true 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
$$\frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}} \leq .03\ .$$
- We can make this independent of $\bar{p}$
$$\frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}} \leq .03 \Rightarrow n \geq 1111$$

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?

Exercise

A bank accepts rolls of pennies and gives 50 cents credit to a customer without counting the contents. Assume that a roll contains 49 pennies 30 percent of the time, 50 pennies 60 percent of the time, and 51 pennies 10 percent of the time.

(a) Find the expected value and the variance for the amount that the bank loses on a typical roll.

(b) Estimate the probability that the bank will lose more than 25 cents in 100 rolls.

(c) Estimate the probability that the bank will lose exactly 25 cents in 100 rolls.

(d) Estimate the probability that the bank will lose any money in 100 rolls.

(e) How many rolls does the bank need to collect to have a 99 percent chance of a net loss?

Exercise

A bank accepts rolls of pennies and gives 50 cents credit to a customer without counting the contents. Assume that a roll contains 49 pennies 30 percent of the time, 50 pennies 60 percent of the time, and 51 pennies 10 percent of the time.

(a) Find the expected value and the variance for the amount that the bank loses on a typical roll.

(b) Estimate the probability that the bank will lose more than 25 cents in 100 rolls.

(c) Estimate the probability that the bank will lose exactly 25 cents in 100 rolls.

(d) Estimate the probability that the bank will lose any money in 100 rolls.

(e) How many rolls does the bank need to collect to have a 99 percent chance of a net loss?

(a) EV is .2 cents and the variance is .36. ; (b) .2024 ; (c) .047 ; (d) .9994 ; (e) 54