Introduction to Probabilities and Normal Distribution

MATH 3512, BCIT

Matrix Methods and Statistics for Geomatics

November 12, 2018

Probabilities

Probabilities are positive real numbers that add up to 1. For example, the probabilities for heads and tails on a coin flip may be

$$P(X = H) = 0.5 \text{ and } P(X = T) = 0.5 \tag{1}$$

If you flip two coins, what are the probabilities for getting two heads ($X = 2$), one head(s) ($X = 1$), or none ($X = 0$)?

Binomial Probabilities I

It turns out that order matters, so that

$$P(X = 2) = \frac{1}{3}, P(X = 1) = \frac{1}{3}, P(X = 0) = \frac{1}{3} \tag{2}$$

is incorrect, while the correct distribution is

$$P(X = 2) = \frac{1}{4}, P(X = 1) = \frac{1}{2}, P(X = 0) = \frac{1}{4} \tag{3}$$

because $P(X = 1) = P(\text{HT}) + P(\text{TH}) = 1/4 + 1/4 = 1/2$.

Binomial Probabilities II

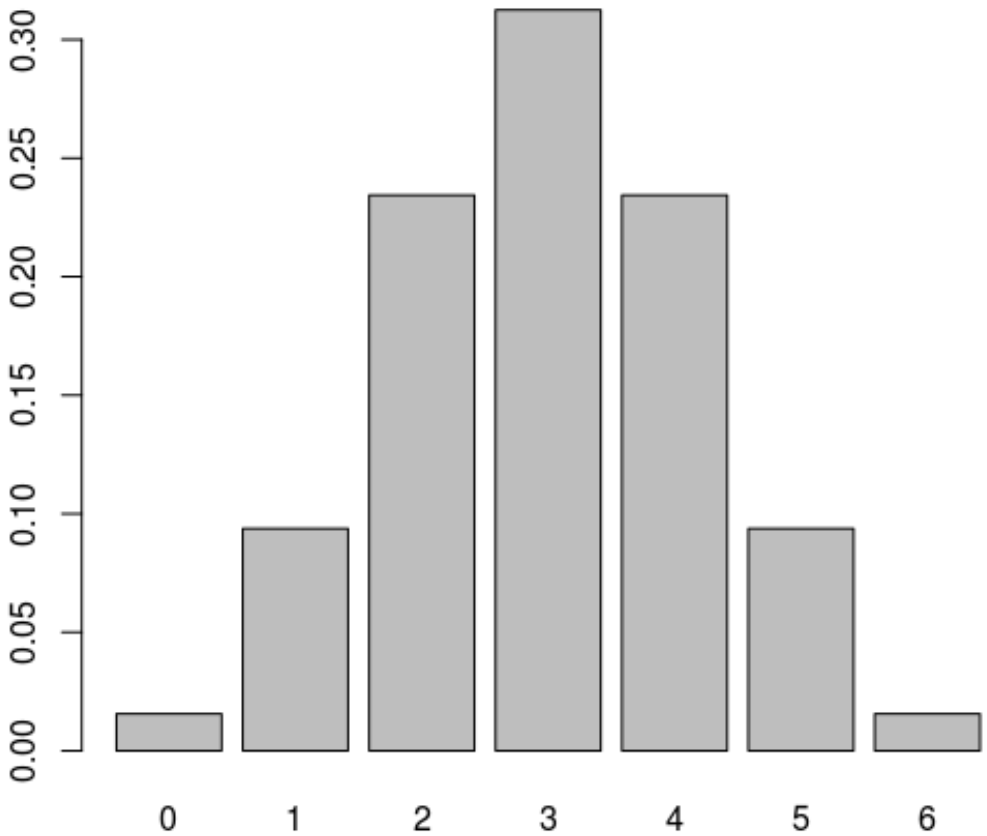
Here is the formula for a binomial setup with n trials (for example, $n = 2$ coin tosses), probability of success p (for example, $p = 0.5$ for the probability of heads), and x number of successes,

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \tag{4}$$

where $0! = 1$ and $(n+1)! = n!(n+1)$, for example $4! = 1 \cdot 2 \cdot 3 \cdot 4$ (say "four factorial"). X is a random variable, the number that the random process spits out.

Binomial Probabilities III

Here are the binomial probabilities for $n = 6$. One way to conceptualize these numbers is by looking at Pascal's Triangle (next slide).



Pascal's Triangle

$$
\begin{array}{ccccccccccc}
 & & & & & 00 & & & & & \\
 & & & & 10 & & 11 & & & & \\
 & & & 20 & & 21 & & 22 & & & \\
 & & 30 & & 31 & & 32 & & 33 & & \\
 & 40 & & 41 & & 42 & & 43 & & 44 & \\
50 & & 51 & & 52 & & 53 & & 54 & & 55 \\
60 & & 61 & & 62 & & 63 & & 64 & & 65 & & 66 \\
70 & & 71 & & 72 & & 73 & & 74 & & 75 & & 76 & & 77
\end{array}
$$

Pascal's Triangle

$$\binom{0}{0}$$

$$\binom{1}{0} \quad \binom{1}{1}$$

$$\binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{2}$$

$$\binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3}$$

$$\binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4}$$

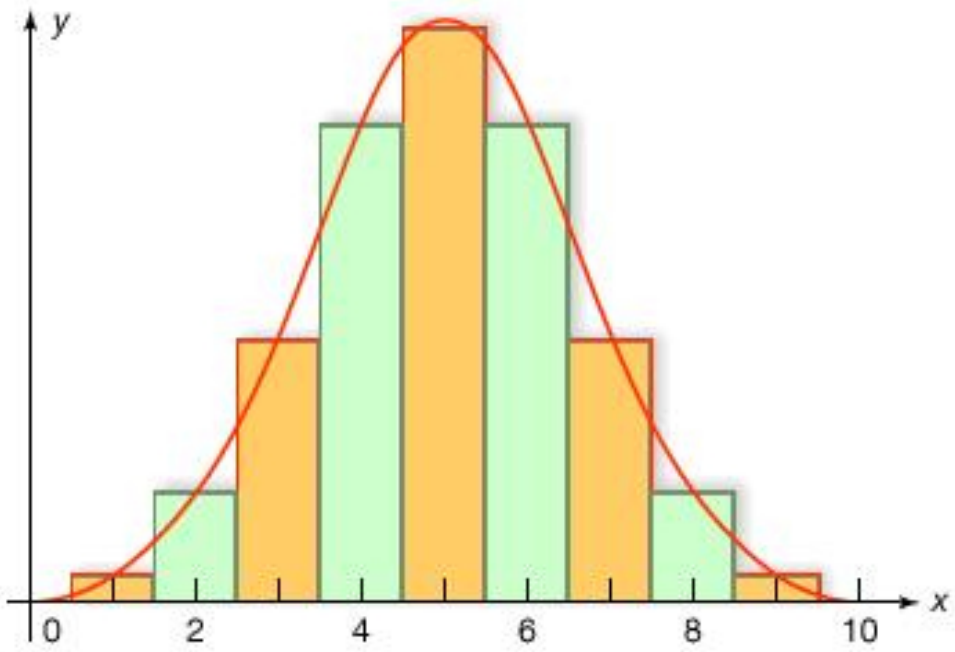
$$\binom{5}{0} \quad \binom{5}{1} \quad \binom{5}{2} \quad \binom{5}{3} \quad \binom{5}{4} \quad \binom{5}{5}$$

$$\binom{6}{0} \quad \binom{6}{1} \quad \binom{6}{2} \quad \binom{6}{3} \quad \binom{6}{4} \quad \binom{6}{5} \quad \binom{6}{6}$$

$$\binom{7}{0} \quad \binom{7}{1} \quad \binom{7}{2} \quad \binom{7}{3} \quad \binom{7}{4} \quad \binom{7}{5} \quad \binom{7}{6} \quad \binom{7}{7}$$

Binomial Probabilities IV

Binomial probabilities are difficult to calculate for high numbers. We approximate the binomial distribution with the normal distribution. Compare the binomial distribution for $n = 20$ with the normal distribution.

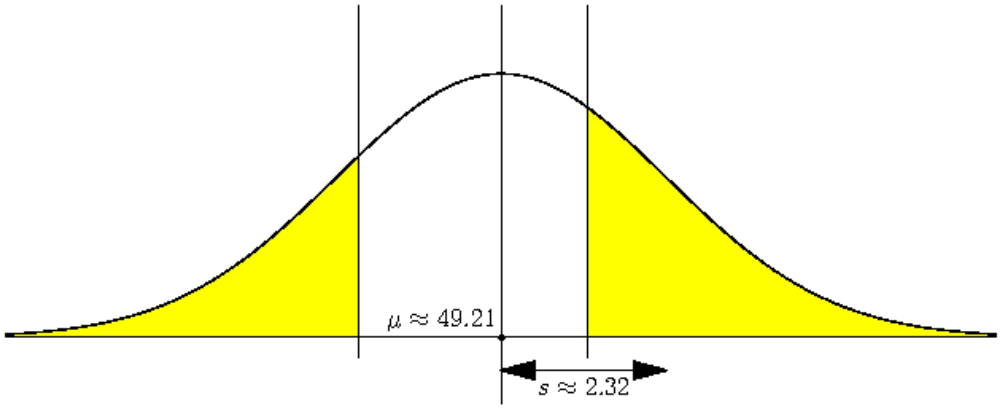


Normal Distribution

There is not just one normal distribution. There are infinitely many, characterized by their mean μ and their standard deviation σ. The formula for the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{5}$$

The area under the curve tells us something about the probability of values in the intervals in which we are interested.



Z Scores

To calculate the area under the curve, we carry around a piece of paper with all the values for the standard normal distribution and then convert to the normal distribution with the relevant μ and σ. The value for the normal distribution is called the x-value and its associate value for the standard normal distribution is called the z-score. Travel back and forth using the following formula,

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

Normal Distribution Example

Here is an example. Men's heights are normally distributed with mean $\mu = 69.5$ inches and standard deviation $\sigma = 2.4$ inches. What percentage of the male population is taller than six feet (72 inches)? Find the z-score, using the formula

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 69.5}{2.4} \approx 1.04 \tag{7}$$

Use your z-score table to find the corresponding p-value, which is the area to the left of the z-score for the standard normal distribution. In this case the p-value is 0.8508. This represents the percentage of the male population that is *shorter* than 72 inches. The answer to our question is therefore, 14.92% of men are taller than six feet.

Approximating Binomial Probabilities I

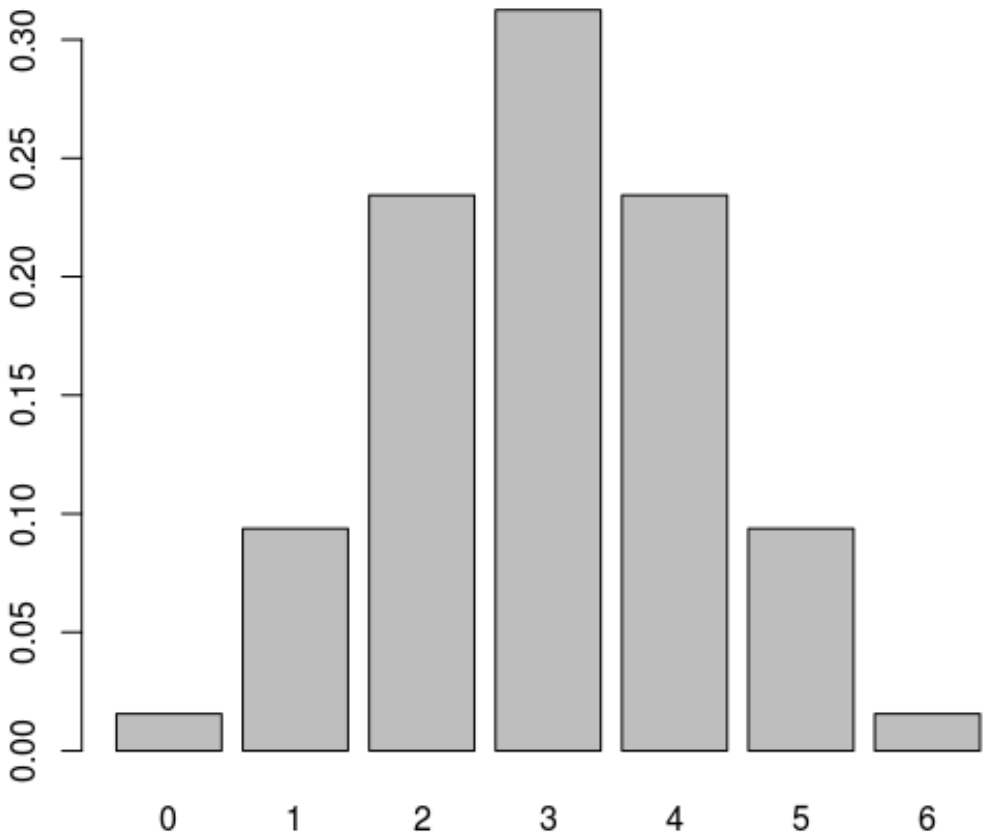
Here is the formula for a binomial setup with n trials (for example, $n = 2$ coin tosses), probability of success p (for example, $p = 0.5$ for the probability of heads), and x number of successes,

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \tag{8}$$

where $0! = 1$ and $(n+1)! = n!(n+1)$, for example $4! = 1 \cdot 2 \cdot 3 \cdot 4$ (say "four factorial"). X is a random variable, the number that the random process spits out.

Approximating Binomial Probabilities II

Here are the binomial probabilities for $n = 6$. One way to conceptualize these numbers is by looking at Pascal's Triangle (next slide).



Pascal's Triangle

$$\binom{0}{0}$$

$$\binom{1}{0} \quad \binom{1}{1}$$

$$\binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{2}$$

$$\binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3}$$

$$\binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4}$$

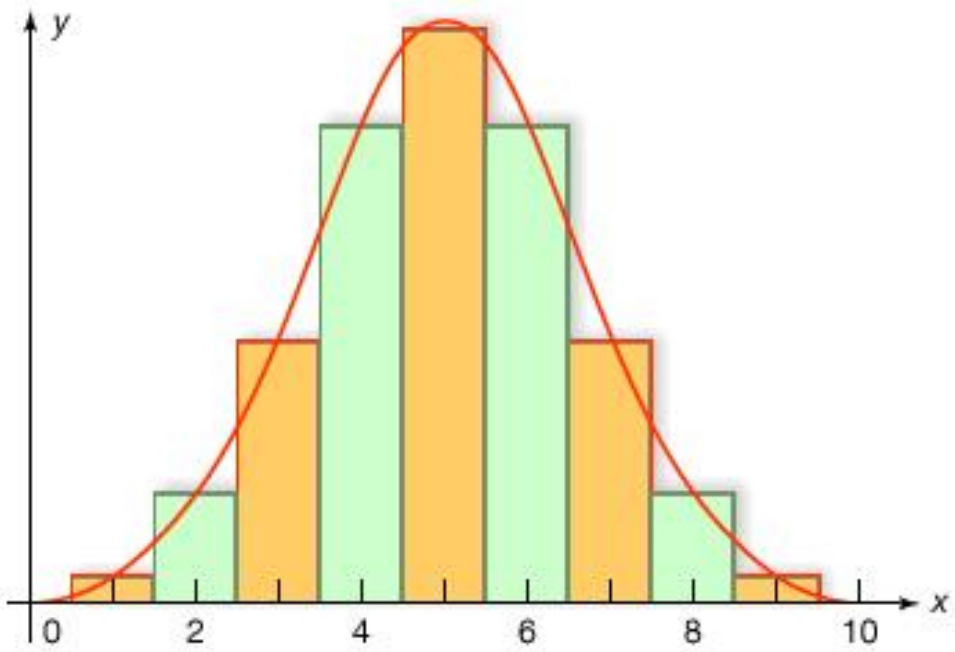
$$\binom{5}{0} \quad \binom{5}{1} \quad \binom{5}{2} \quad \binom{5}{3} \quad \binom{5}{4} \quad \binom{5}{5}$$

$$\binom{6}{0} \quad \binom{6}{1} \quad \binom{6}{2} \quad \binom{6}{3} \quad \binom{6}{4} \quad \binom{6}{5} \quad \binom{6}{6}$$

$$\binom{7}{0} \quad \binom{7}{1} \quad \binom{7}{2} \quad \binom{7}{3} \quad \binom{7}{4} \quad \binom{7}{5} \quad \binom{7}{6} \quad \binom{7}{7}$$

Approximating Binomial Probabilities III

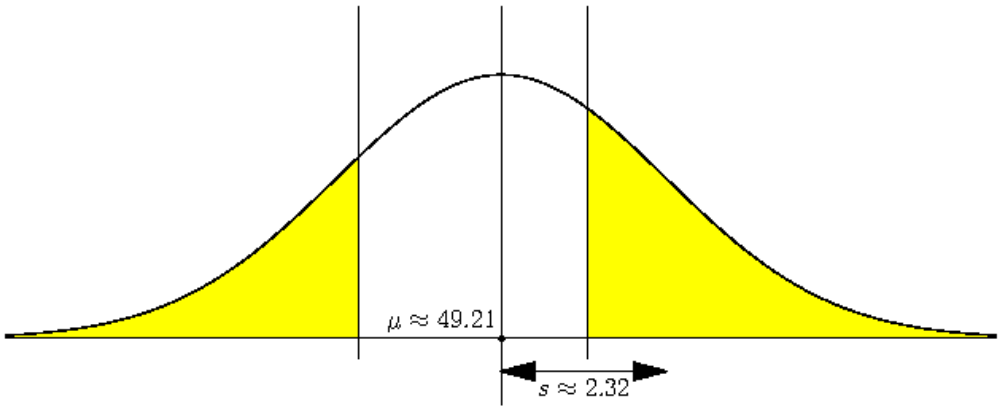
Binomial probabilities are difficult to calculate for high numbers. We approximate the binomial distribution with the normal distribution. Compare the binomial distribution for $n = 10$ with the normal distribution.



Normal Distribution

The normal probabilities distribution is a continuous probabilities distribution. There is not just one normal distribution. There are infinitely many, characterized by their mean μ and their standard deviation σ. The formula for the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{9}$$



Z Scores

To calculate the area under the curve, we carry around a piece of paper with all the values for the standard normal distribution and then convert to the normal distribution with the relevant μ and σ. The value for the normal distribution is called the x-value and its associate value for the standard normal distribution is called the z-score. Travel back and forth using the following formula,

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

Normal Distribution Example Question

Example 1: Men's Heights. The height of adult men is normally distributed with mean $\mu = 69.5$ inches and standard deviation $\sigma = 2.4$ inches. What percentage of the adult male population is taller than six feet (72 inches)?

Normal Distribution Example Answer

Find the z-score, using the formula

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 69.5}{2.4} \approx 1.04 \tag{11}$$

Use your z-score table to find the corresponding p-value, which is the area to the left of the z-score for the standard normal distribution. In this case the p-value is 0.8508. This represents the percentage of the male population that is *shorter* than 72 inches. The answer to our question is therefore, 14.92% of men are taller than six feet.

Normal Distribution Exercises I

Exercise 1: Find the area under the curve for the following sets of z-scores.

$$\{z|z \leq -1.72\} \tag{12}$$

$$\{z|1.96 < z\} \tag{13}$$

$$\{z|-1.55 \leq z \leq -0.81\} \tag{14}$$

Normal Distribution Exercises II

Exercise 2: Find the area under the curve for the following sets of x-values.

$$\{x|x \leq 83\}, \mu = 100, \sigma = 15 \tag{15}$$

$$\{x|0.44 < x\}, \mu = 0.5, \sigma = 0.2 \tag{16}$$

$$\{x|800 \leq x \leq 1200\}, \mu = 911, \sigma = 121 \tag{17}$$

Normal Distribution Exercises III

Exercise 3: The time it takes a student to solve a particular math problem is normally distributed with $\mu = 3$ minutes and 27 seconds and a standard deviation of $\sigma = 59$ seconds. How many students finish the math problem between 3 and 4 minutes?

Normal Distribution Exercises IV

Exercise 4: Scores on a certain intelligence test for children between ages 13 and 15 years are approximately normally distributed with $\mu = 106$ and $\sigma = 15$.

1. What proportion of children aged 13 to 15 years old have scores on this test above 88?
2. Enter the score which marks the lowest 30 percent of the distribution.
3. Enter the score which marks the highest 15 percent of the distribution.

When you want to find an x-value given a p-value, you need to use your table in reverse and then find the x-value given the z-score from the table by using the formula

$$x = z \cdot \sigma + \mu \tag{18}$$

Approximating Binomial Probabilities Example I

For large numbers, even high-powered computers cannot calculate the binomial probabilities. We use the normal distribution to approximate the binomial distribution.

Example 2: Die Rolls. If you roll a die 600 times, what is the probability of rolling a six fewer than 80 times?

It would take very long to calculate this probability using the binomial distribution formula! We use the normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ instead.

Conventionally, it is only acceptable to approximate the binomial distribution by the normal distribution if $np \geq 5$ and $nq \geq 5$. Otherwise, the binomial and the normal distribution are too far apart to provide a useful approximation.

Approximating Binomial Probabilities Example II

We need to make a continuity correction and ask ourselves, what is the probability for the x-value to be 79.5 or less for this normal distribution?

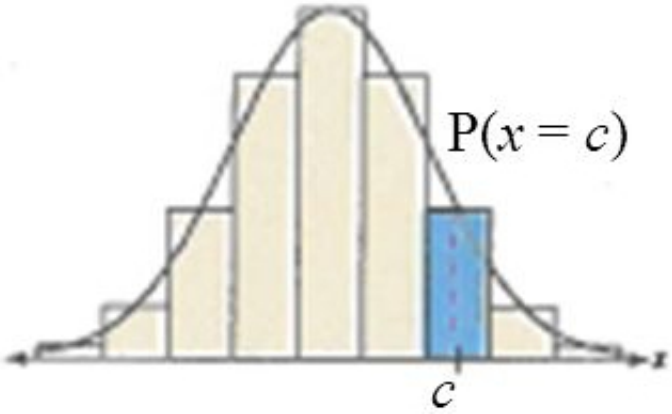
$$z = \frac{x - \mu}{\sigma} = \frac{79.5 - 100}{9.1287} \approx -2.25 \tag{19}$$

The corresponding p-value is 0.0122. There is only a 1.22% probability that you will roll fewer than 80 sixes in 600 rolls.

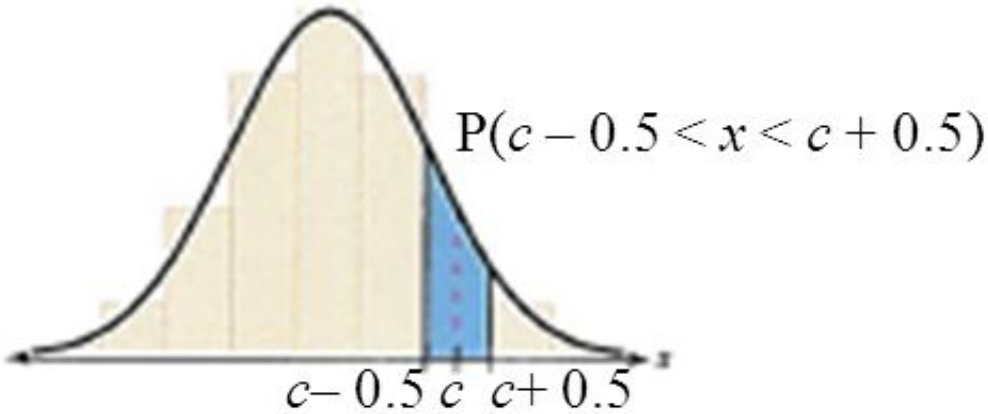
Continuity Correction

When you use a *continuous* normal distribution to approximate a binomial probability, you need to move 0.5 unit to the left and right of the midpoint to include all possible x-values in the interval (**correction for continuity**).

Exact binomial probability



Normal approximation



Binomial Approximation Exercise

Exercise 5: 29% of a country's population is blue-eyed. What is the probability that a random sample of 1,000 persons contains between 200 and 300 blue-eyed persons? Approximate the binomial distribution by the normal distribution. Calculate $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ for this normal distribution after checking the conditions $np \geq 5, n(1-p) \geq 5$. Then calculate the z-scores for the x-values $x = 199.5$ and $x = 300.5$. Lastly, determine the area under the curve between these z-scores. Provide a complete sentence to answer the question.

Word Problems I

The national mortality rate for a particular type of heart surgery is 12%, so you would expect six deaths per 50 operations. You are a health administrator, and one of your doctors has had eleven deaths in 72 operations. Should you fire her? What is the probability that an average surgeon (whose mortality rate is the national average) will have twelve or more deaths in 72 operations?

You desperately need a person with blood type AB (incidence in Canada: 3%). If there are 49 people in the room, what is the probability that at least one of them is AB?

Word Problems I

The national mortality rate for a particular type of heart surgery is 12%, so you would expect six deaths per 50 operations. You are a health administrator, and one of your doctors has had eleven deaths in 72 operations. Should you fire her? What is the probability that an average surgeon (whose mortality rate is the national average) will have twelve or more deaths in 72 operations?

You desperately need a person with blood type AB (incidence in Canada: 3%). If there are 49 people in the room, what is the probability that at least one of them is AB?

Word Problems II

18. Washing Hands Based on *observed* males using public restrooms, 85% of adult males wash their hands in a public restroom (based on data from the American Society for Microbiology and the American Cleaning Institute). In a survey of 523 adult males, 518 *reported* that they wash their hands in a public restroom. Assuming that the 85% observed rate is correct, find the probability that among 523 randomly selected adult males, 518 or more wash their hands in a public restroom. What do you conclude?

19. Voters Lying? In a survey of 1002 people, 701 said that they voted in a recent presidential election (based on data from ICR Research Group). Voting records show that 61% of eligible voters actually did vote. Given that 61% of eligible voters actually did vote, find the probability that among 1002 randomly selected eligible voters, at least 701 actually did vote. What does the result suggest?

20. Cell Phones and Brain Cancer In a study of 420,095 cell phone users in Denmark, it was found that 135 developed cancer of the brain or nervous system. Assuming that the use of cell phones has no effect on developing such cancers, there is a 0.000340 probability of a person developing cancer of the brain or nervous system. We therefore expect about 143 cases of such cancers in a group of 420,095 randomly selected people. Estimate the probability of 135 or fewer cases of such cancers in a group of 420,095 people. What do these results suggest about media reports that cell phones cause cancer of the brain or nervous system?

Word Problems III

21. Smoking Based on a recent Harris Interactive survey, 20% of adults in the United States smoke. In a survey of 50 statistics students, it is found that 6 of them smoke. Find the probability that should be used for determining whether the 20% rate is correct for statistics students. What do you conclude?

23. Online TV In a Comcast survey of 1000 adults, 17% said that they watch prime-time TV online. If we assume that 20% of adults watch prime-time TV online, find the probability that should be used to determine whether the 20% rate is correct or whether it should be lower than 20%? What do you conclude?

24. Internet Access Of U.S. households, 67% have Internet access (based on data from the Census Bureau). In a random sample of 250 households, 70% are found to have Internet access. Find the probability that should be used to determine whether the 67% rate is too low. What do you conclude?

End of Lesson

Next Lesson: Axioms and Theorems of Probability