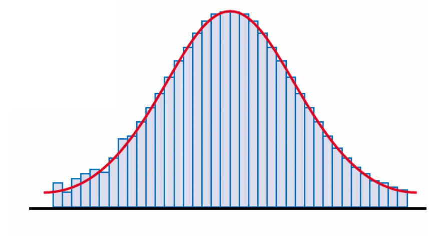
Unit 7: Normal Distribution

Lesson 7.3: Normal Approximation to the Binomial Distribution

We have covered:
- Use z-table to find the area under the normal distribution curve, so that knowing the proportion of data falls into a certain interval.
- Confidence interval with given p or given mean and standard deviation (use sample to estimate population)



All the examples of normal distributions we have seen so far have modelled continuous data. There are many situations, however, where discrete data can also be modelled as normal distributions.

Example 1:

A company produces boxes of candy-coated chocolate pieces. The number of pieces in each box is assumed to be normally distributed with a mean of 48 pieces and a standard deviation of 4.3 pieces. Quality control will reject any box with fewer than 44 pieces. Boxes with 55 or more pieces will result in excess costs to the company.

a) What is the probability that a box selected at random contains exactly 50 pieces?

b) What percent of the production will be rejected by quality control as containing too few pieces?

c) Each filling machine produces 130 000 boxes per shift. How many of these will lie within the acceptable range?

If the discrete data have a binomial probability distribution and certain **simple conditions** are met, the normal distribution makes a very good approximation by using **continuity correction**. This approximation allows the probabilities of value ranges to be calculated more easily than with the binomial formulas.

<u>Condition:</u>

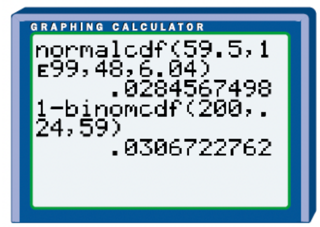
To ensure the shape of the binomial distribution needs to be similar to the shape of the normal distribution, the quantities np and nq must both be greater than five. (i.e., **np > 5 and nq > 5**), the approximation is better if they are both greater than or equal to 10.

Then the binomial can be approximated by the normal distribution with $\boldsymbol{\mu = np \text{ and } \sigma = \sqrt{npq}}$.

In order to get the bets approximation, add 0.5 to x or subtract 0.5 from x (use x + 0.5 or x – 0.5). The number 0.5 is called the continuity correction factor.

Example 2:

A bank found that 24% of its loans to new small businesses become delinquent. If 200 small businesses are selected randomly from the bank's files, what is the probability that at least 60 of them are delinquent? Compare the results from the normal approximation with the results from the calculations using a binomial distribution.



Practice:

QuenCola, a soft-drink company, knows that it has a 42% market share in one region of the province. QuenCola's marketing department conducts a blind taste test of 70 people at the local mall.

a) What is the probability that fewer than 25 people will choose QuenCola?

b) What is the probability that exactly 25 people will choose QuenCola?

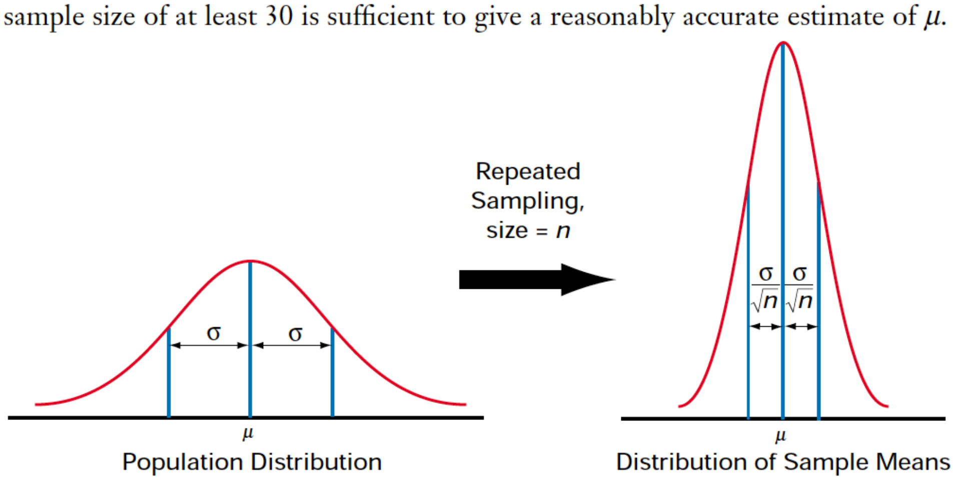
Lesson 7.4: Sampling Distribution (8.5) and Hypothesis Test (8.6)

Part 1: Sampling Distribution (8.5)

When you draw a sample from a population, you often use the sample mean, $\bar{x}$, as an estimate of the population mean, μ, and the sample standard deviation, s, as an estimate of the population standard deviation, σ. However, the statistics for a single sample may differ radically from those of the underlying population. Statisticians try to address this problem by repeated sampling. Do additional samples improve the accuracy of the estimate?

When repeated samples of the same size are drawn from a normal population, the sample means will be normally distributed with a mean equal to the population mean μ.

The distribution of sample means will have a standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, where n is the sample size. Notice that if $n = 1$, the samples are single data points and $\sigma_{\bar{x}} = \sigma$. As n increases, however, $\sigma_{\bar{x}}$ decreases, so the distribution of sample means becomes more tightly grouped around the true mean. Usually a sample size of at least 30 is sufficient to give a reasonably accurate estimate of μ.



Example 3:

The tires on a rental-car fleet have lifetimes that are normally distributed with a mean life of 64 000 km and a standard deviation of 4800 km. Every week a mechanic checks the tires on ten randomly selected cars.

a) What are the mean and standard deviation of these samples?

b) How likely is the mechanic to find a sample mean of 62 500 km or less?

Part 2: Hypothesis Testing (8.5)

Statisticians use the standard deviation of the sample mean to quantify the reliability of statistical studies and conclusions. Often, it is not possible to determine with certainty whether a statement is true or false. However, it is possible to test the strength of the statement, based on a sample. This procedure is called a **hypothesis test**.

Consider this scenario. A large candy manufacturer produces chocolate bars with labels stating "45 g net weight." A company spokesperson claims that the masses are normally distributed with a mean of 45 g and a standard deviation of 2 g. A mathematics class decides to check the truth of the 45 g label. They purchase and weigh 30 bars. The mean mass is 44.5 g. Is this evidence enough to challenge the company's claim?

You can construct a hypothesis test to investigate the 45-g net-weight claim, using these steps.

- *Step 1* State the hypothesis being challenged, **null hypothesis** (null means no change in this context). The null hypothesis is usually denoted H_0. So, for the chocolate bars, H_0: $\mu = 45$.
- *Step 2* State the alternative hypothesis H_1 (sometimes called H_a). You suspect that the mean mass may be lower than 45 g; so, H_1: $\mu < 45$.
- *Step 3* Establish a decision rule. How strong must the evidence be to reject the null hypothesis? If, for a normal distribution with a population mean of 45 g, the probability of a sample with a mean of 44.5 g is *very small*, then getting such a sample would be strong evidence that the actual population mean is not 45 g. The **significance level**, α, is the probability threshold that you choose for deciding whether the observed results are rare enough to justify rejecting H_0. For example, if $\alpha = 0.05$, you are willing to be wrong 5% of the time.
- *Step 4* Conduct an experiment. For the chocolate bars, you weigh the sample of 30 bars.
- *Step 5* Assume H_0 is true. Calculate the probability of obtaining the results of the experiment given this assumption. If the standard deviation for chocolate bar weights is 2 g, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$$

and

$$P(\bar{x} < 44.5) = P\left(Z < \frac{44.5 - 45}{0.365}\right) \doteq P(Z < -1.37) = 0.0853$$

The probability of a sample mean of 44.5 or less, given a sample of size 30 from an underlying normal distribution with a mean of 45, is 8.5%.

- *Step 6* Compare this probability to the significance level, α: 8.5% is greater than 5%.
- *Step 7* Accept H_0 if the probability is greater than the significance level. Such probabilities show that the sample result is not sufficiently rare to support an alternative hypothesis. If the probability is less than the significance level reject H_0 and instead accept H_1. So, for the chocolate bars, you would accept H_0.
- *Step 8* Draw a conclusion. How reliable is the company's claim? In this case, the statistical evidence is slightly too weak to refute the company's claim.

If $P(\bar{x} < 44.5) < \alpha$, then the evidence is strong enough to reject null hypothesis, so the population mean = 45 g is not true.

If $P(\bar{x} < 44.5) > \boldsymbol{\alpha}$, then the evidence is NOT strong enough to reject the null hypothesis, so to **accept $\boldsymbol{H_o}$**, the population mean = 45 g is true.

Example 4:

An automobile company is looking for fuel additives that might increase gas mileage. Without additives, their cars are known to average 19 mpg (miles per gallons) with a standard deviation of $\frac{19}{6}$ mpg on a road trip from Muscat to Sur. The company now asks whether a new additive increases this value. In a study, forty cars are sent on a road trip from Muscat to Sur. Suppose it turns out that the forty cars averaged x= 19.75 mpg with the additive. Can we conclude from this result that the additive is effective?

With a significance level of 1%.
With a significance level of 10%

Practice:

An automobile company is looking for fuel additives that might increase gas mileage. Without additives, their cars are known to average 25 mpg (miles per gallons) with a standard deviation of 2.4 mpg on a road trip from London to Edinburgh. The company now asks whether a particular new additive increases this value. In a study, thirty cars are sent on a road trip from London to Edinburgh. Suppose it turns out that the thirty cars averaged $\overline{x} = 25.5$ mpg with the additive. Can we conclude from this result that the additive is effective?

Practice: Drug effectiveness

A drug company tested a new drug on 250 pigs with swine flu. Historically, 20% of pigs contracting swine flu die from the disease. Of the 250 pigs treated with the new drug, 215 recovered. Make a hypothesis test of the drug's effectiveness, with a significance level of $\alpha = 1\%$.

a) Determine whether you can model this study with a normal distribution.

b) Set up a null hypothesis and an alternative hypothesis for the value of μ in the normal approximation.

c) Which hypothesis corresponds to the drug being effective? Explain.

d) Conduct the hypothesis test.

e) Can the company claim that its new drug is effective?

More practice:

Application A newspaper stated that 70% of the population supported a particular candidate's position on health care. In a random survey of 50 people, 31 agreed with the candidate's position. Test the significance of this result with a confidence level of 90%. Should the newspaper print a correction?

Inquiry/Problem Solving A student-loan program claims that the average loan per student per year is \$7500. Dana investigated this statement by asking 50 students about this year's student loan. The mean of the results was \$5800. What additional information does Dana need to test the significance of this result? What significance level would be appropriate here? Why?