The Central Limit Theorem

Central Limit Theorem

The central limit theorem says: when we generate the **sampling distribution of the means** from a **large number** of **samples** of a **given size**, taken **from any population** – normally distributed or not – this **distribution** will be (approximately) **normally distributed**; with a **mean equal to** the **population mean**. The **standard deviation** of this distribution, known as the **standard error of the mean**, will be **less than** that of the **sampled population**.

$$\mu_{\bar{x}} = \mu, \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

If the population is finite of size N, we use the **FPCF**.

Finding the probability that $\bar{x}$ lies in a given range

Example: A manufacturer of glass tubing suspects that his production machines are not working up to specifications. He knows that if things are running properly, the diameter of the tubes is **normally distributed** with a **mean of 1.5** cm and a **standard deviation of 0.05** cm. His analysts select a sample of **8** tubes to determine whether the production machinery needs expensive tweeking. If the 8 diameters measure

1.57 1.60 1.62 1.52 1.55 1.59 1.48 1.59

what will they conclude?

Solution:

step 1: calculate the sample mean: (statistic in question)

$$\bar{x} = \frac{\sum x}{n} = 1.565\ cm.$$

step 2: define the sampling distribution for $\bar{x}$: state the values of $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

since $\mu_{\bar{x}} = \mu,\ \mu = 1.5\ cm$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.05}{\sqrt{8}} = 0.0177\ cm$

step 3: define the criteria (event of interest):

We want $P(\bar{x} \geq 1.565\ cm)$

(*A high probability indicates it's not due to chance, so the machinery needs tweeking*.)

step 4: standardize the sample statistic $\bar{x}$ with the normal curve (z-value)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.565 - 1.5}{0.0177} = 3.67$$

conclusion: The **probability** for a z-value **this high is not even included** on most normal distribution tables, so the analysts must **conclude** that the **difference** between the sample mean and the population mean **is not due to chance**.

Had the sample **mean** $\bar{x}$ been = 1.52, then the calculated value of z would have been 1.13

$$z = \frac{1.52 - 1.5}{0.0177} = 1.13$$

and $P(z \geq 1.13) = 0.5000 - 0.3708 = 0.1292$

-- making the probability of error due to chance = 12.92 %.

The Central Limit Theorem

Example:

A bank auditor claims that credit card balances are normally distributed, with a mean of $2780 and a standard deviation of $900.

a) What is the probability that a randomly selected credit card holder has a credit card balance less than $2500?

Solution

Since we're asked about ***only one*** credit card holder, we use the mean and SD of the population to perform our z-score:

As 2500 is less than the mean, we will find the probability that $x > 2500$ and subtract it from 0.5.

$$z = \frac{x - \mu}{\sigma} \quad \Rightarrow \quad \frac{2500 - 2780}{900} = -0.3111$$

This z-value tells us P($2500 < x < 2780$) = 0.1217

Therefore P($x < 2500$) = 0.5000 – 0.1217 = 0.3783

So the probability that a cardholder has a balance less than $2500 is **37.83 %**.

b) If we randomly select **25 credit card holders**. What is the probability that their mean credit card balance is **less than $2500**?

Solution

Now we have a ***sampling*** of **more than one credit card holder**, so we use $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ to find our z-score:

$$\mu_{\bar{x}} = \mu = 2780, \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{900}{\sqrt{25}} = 180$$

Now we find our z-score:

$$z = \frac{x - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \quad \Rightarrow \quad \frac{2500 - 2780}{180} = -1.56$$

This z-value tells us P($2500 < x < 2780$) = 0.4406

Therefore P($x < 2500$) = 0.5000 – 0.4406 = 0.0594

Therefore the probability that the mean of 25 cardholders is less than $2500 is **5.94 %**.

The Central Limit Theorem

example: The diameter of ping-pong balls manufactured at a large factory is expected to be approximately normally distributed, with a mean of 33.0 millimetres and a standard deviations 1.0 millimetre.

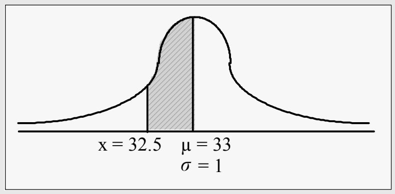
a) What is the probability that the diameter of a randomly selected ping-pong ball will fall between 32.5 and 33.0 millimetres?

solution: again we've got a **single unit** from the population, so we use population parameters. Since 33 = μ, we need only find the probability for $x = 32.5$

To find P(32.5 < x < 33.0), we standardize the value 32.5.

$$z = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad \frac{32.5-33}{1} = -0.5$$

From the table of Normal Probabilities, we find P(– 0.5 < z < 0.0) = **0.1915.**



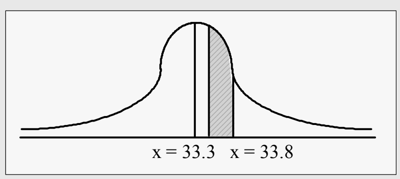
b) What is the probability that the diameter of a randomly selected ping-pong ball will fall between 33.3 and 33.8 millimetres?

solution: Now we standardize both values 33.3 and 33.8 and since they're both above the mean, we will subtract the probabilities.

$$z_1 = \frac{33.3-33}{1} = 0.3 \qquad z_2 = \frac{33.8-33}{1} = 0.8$$

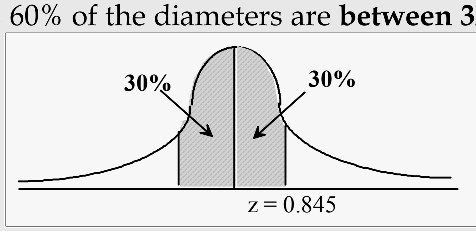
P(33.8 < x < 33.0) = 0.2881 and P(33.3 < x < 33.0) = 0.1179

so P(33.3 < x < 33.8) = 0.2881 – 0.1179 = **0.1702**



c) Between what two values for the diameter will 60% of the ping-pong balls fall?

solution: We must have 30% of the ping pong balls on either side of the mean, so we find the z-score for 30%, that tells us how many standard deviations to add to and subtract from the mean to include 30% of the population.
The z-score for 30% = 0.845 after some <u>interpolation</u> and since $\sigma = 1$, we subtract 0.845 from 33 and add it to 33 to get the range of diameters that includes 60% of the ping pong balls. So, 60% of the diameters are **between 32.155 mm and 33.845** mm.



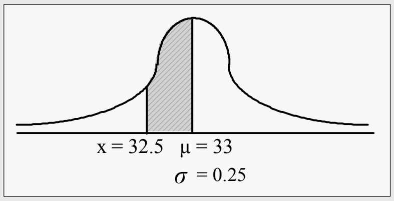
d) If many random samples of 16 ping-pong balls are selected, what will be the values of the population mean and standard error of the mean?

solution: Now we're using a sampling distribution, so we need $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

$$\mu_{\bar{x}} = \mu = 33 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{16}} = 0.25$$

e) If many random samples of 16 ping-pong balls are selected, what proportion of the sample means will be between 32.5 and 33.0 millimetres?

solution: Since $\sigma_{\bar{x}} = 0.25$ we can see that 32.5 is exactly 2 sigmas below the mean, so we need the probability of falling between the mean and 2 sigmas below it making $z = -2$.
P($-2 < z < 0$) = **0.4772**



f) If many random samples of 16 ping-pong balls are selected, what proportion of the sample means will be between 33.3 and 33.8 millimetres?

solution: We find the *z*-scores for 33.3 and 33.8 using the new parameter for $\sigma_{\bar{x}} = 0.25$.

$$z_1 = \frac{33.3 - 33}{0.25} = 1.2 \qquad z_2 = \frac{33.8 - 33}{0.25} = 3.2$$

Now we need the difference between their probabilities.
P($0 < z < 3.2$) = 0.4993, and P($0 < z < 1.2$) = 0 .3849,
Therefore, P($1.2 < z < 3.2$) = 0.4993 – 0 .3849 = **0.1144**

g) Between what two values will 60% of the sample means be?

solution: We must have 30% of the ping pong balls on either side of the mean as before but now we use $\sigma_{\bar{x}} = 0.25$, and since the *z*-score for 30% = 0.845 after some <u>interpolation</u> and since σ = 0.25, we subtract 0.845 × (0.25) = 0.2113 from 33 and add it to 33 to get the range of diameters that includes 60% of the ping pong balls.
So, 60% of the ping pong balls will have diameters **between 32.789 mm and 33.211** mm.

h) Which is most likely to occur – **i)** an individual ball with a diameter greater than 34.0 millimetres, **ii)** a sample mean greater than 33.5 millimetres in a sample of 4 ping pong balls, or **iii)** a sample mean greater than 33.3 millimetres in a sample of 16?

solution: We will find the *z*-score for all 3 situations and compare them:

$$z_1 = \frac{34 - 33}{1} = 1 \qquad z_2 = \frac{33.5 - 33}{\frac{1}{\sqrt{4}}} = 1 \qquad z_3 = \frac{33.3 - 33}{\frac{1}{\sqrt{16}}} = 1.2$$

P($z_1 > 1$) = 0.5000 – 0.3413 = 0.1587 P($z_3 > 1.2$) = 0.5000 – 0.3849 = 0.1151

Since 0.1587 > 0.1151, both **i)** and **ii)** are most likely to occur.

The Central Limit Theorem

The Sampling Distribution of a Proportion ($\bar{p}$)

A Sampling Distribution of Proportions can be generated in the same way as the Sampling Distribution of the Means. We then use the Standard Normal tables to assess the sampling error and to estimate population parameters.

The Mean and Standard Error of the Sampling Distribution for a Proportion

$$\mu_{\bar{p}} = np \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

Now the **error** in the numerator of the formula for z represents the **difference between** the sample and population **proportions**.

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}}$$

Finding Probability that $\bar{p}$ is in a Given Range:

We standardize (find z-value) the error $\bar{p} - p$, then use the Normal table to find probabilities.

Example:

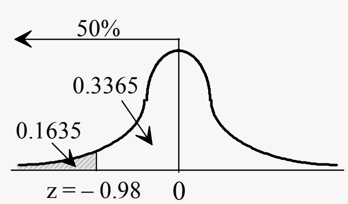
A marketing team claims that 70% of their clients have been with them for 5 years or more. A study showed that 68% of 500 clients had been loyal customers for 5 years or more. With what probability can they expect to get a sample proportion of 68% or less?

Solution:

We find $\sigma_{\bar{p}} = \sqrt{\frac{0.70(0.30)}{500}} = 0.0205$

Now we find z for $\bar{p} = 0.68$: $\quad z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.68 - 0.70}{0.0205} = -0.98$

$P(z < -0.98) = 0.5000 - 0.3365 = 16.35\%$



P(z < – 0.98) = 0.5 – 0.3365 = 0.1635

The Central Limit Theorem

Example:

A Courrier service claims that 80% of their overseas packages are delivered to their destination within 3 days. One of their clients hires this company to deliver 200 packages to destinations overseas, what is the probability that:

a) 150 or more of the packages reach their destination within 3 days?

Solution:

a) From the data, we know: $p = 0.80$

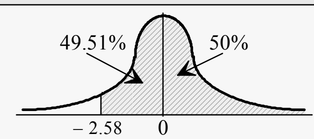
$$\bar{p} = \frac{150}{200} = 0.75 \quad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.8)(0.2)}{200}} = 0.028$$

We find z from the data:

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.75 - 0.80}{0.028} = -2.58$$

We find $P(z > -2.58)$ from the Normal table = 0.5000 + 0.4951 = 99.51%

The company can be nearly 100% sure that 150 or more of the 200 packages will reach their destinations within 3 days



P(z > – 2.58) = 0.5000 + 0.4951 = 0.9951

b) between 150 and 160 of the packages reach their destination within 3 days?

Solution:

Since (0.80)(200) = 160 which is np the mean of the population, we need the probability that z falls between – 2.58 and the mean whic = 0.

$P(-2.58 < z < 0) = 0.4951$ or 49.51%.

The company can be almost 50% sure that 150 to 160 of the packages will reach their destination within 3 days.

Example:

If we sample 1000 items from a population with $p = 0.40$, find the probability that:

a) $\bar{p} < 0.43$

Solution:

a) We find $\sigma_{\bar{p}} = \sqrt{\frac{0.40(0.60)}{1000}} = 0.0155$

Now we find z for $\bar{p} = 0.43$: $z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.43 - 0.40}{0.0155} = 1.94$

We find the probability that $z < 1.94$ from the Normal Distribution table:

$P(z < 1.94) = 0.5000 + 0.4738 = 0.9738$ or **97.38%**

b) $\bar{p} > 0.44$

Solution: Now we find z for $\bar{p} = 0.44$: $z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.44 - 0.40}{0.0155} = 2.58$

We find the probability that $z > 2.58$ from the Normal Distribution table:

$P(z > 2.58) = 0.5000 - 0.4951 = 0.0049$ or **0.49%** very improbable.

The Central Limit Theorem

Practice

1/ In a quality-control experiment, a sample of 500 items is taken from an assembly line. Customarily, 8% of the items produced are defective. What is the probability that more than 50 defective items appear in the sample?

2/ The mean of a random sample of size n = 36 is used to estimate the mean of a normal population with $\sigma = 18$. With what probability can we assert that the error of this estimate will be less than 9 if we use The Central Limit Theorem?

3/ The mean of a random sample of size n = 81 is used to estimate the mean annual growth of certain plants. If **s = 3.6 mm** for such data, use the Central Limit Theorem to find the probabilities that this estimate will be off either way by:

a) less than 1.0 mm b) less than 0.5 mm

4/ A study of 200 families showed they spent an average of $218.67 per week on food with a standard deviation of $14.93. With what probability can we assert that this estimate is "off" by no more than $1.74?

5/ A poll in a magazine stated that on average 36% of 980 customers who bought a new product were dissatisfied with it within a week of purchase.
The poll had a margin of error of ± 3%.

(a) Find the upper and lower values of their estimate for the proportion of customers who were dissatisfied.

(b) What is the probability that the true proportion of dissatisfied customers lies between these values?

The Central Limit Theorem

Solutions

1/ In a quality-control experiment, a sample of 500 items is taken from an assembly line. Customarily, 8% of the items produced are defective. What is the probability that more than 50 defective items appear in the sample?

Solution
It's an assembly line, so we can assume that 500 is less than 5% of the population and we can see that $np \geq 5$ and $n(1-p) \geq 5$ so we can **use the Normal distribution**
We need to find μ and σ so we can calculate z.
Since it is a Binomial distribution, we use the formulas for mean and SD for a binomial distribution.

$\mu = np = 500(0.08) = 40$ and $\sigma = \sqrt{np(1-p)} = \sqrt{(500)(0.08)(0.92)} = 6.066$

Now we find z scores:

$$P(X > 50) = P(Z > \frac{x-\mu}{\sigma}) = P(Z > \frac{50-40}{6.066}) = P(Z > 1.648)$$
$$P(Z > 1.648) = 0.495$$

So the probability of having more than 50 samples defective in the sample is 0.495.

2/ The mean of a random sample of size n = 36 is used to estimate the mean of a normal population with σ = 18. With what probability can we assert that the error of this estimate will be less than 9 if we use The Central Limit Theorem?

Solution
It's an assembly line, so we can assume that 500 is less than 5% of the population and we can see that $np \geq 5$ and $n(1-p) \geq 5$ so we can **use the Normal distribution**
We need to find μ and σ so we can calculate z.
Since it is a Binomial distribution, we use the formulas for mean and SD for a binomial distribution.

$\mu = np = 500(0.08) = 40$ and $\sigma = \sqrt{np(1-p)} = \sqrt{(500)(0.08)(0.92)} = 6.066$

Now we find z scores:

$$P(X > 50) = P(Z > \frac{x-\mu}{\sigma}) = P(Z > \frac{50-40}{6.066}) = P(Z > 1.648)$$
$$P(Z > 1.648) = 0.495$$

So the probability of having more than 50 samples defective in the sample is **0.495**.

The Central Limit Theorem

3/ The mean of a random sample of size n = 81 is used to estimate the mean annual growth of certain plants. If **s = 3.6 mm** for such data, use the Central Limit Theorem to find the probabilities that this estimate will be off either way by:

Solutions:

a) less than 1.0 mm

$s_x = 3.6/9 = 0.4$, so $z = 1/0.4 = 2.5$,

$$P(-2.5 < z < 2.5) = 2(0.4938) = \mathbf{0.9876}$$

b) less than 0.5 mm

$P(E < 0.5) = P(-1.25 < z < 1.25) = 2(0.3944) = \mathbf{0.7888}$

4/ A study of 200 families showed they spent an average of $218.67 per week on food with a standard deviation of $14.93. With what probability can we assert that this estimate is "off" by no more than $1.74?

Solution:

$$z = \frac{E}{\frac{\sigma}{\sqrt{n}}} = \frac{1.74}{\frac{14.93}{\sqrt{200}}} = 1.648.$$

$P(|z| < 1.648) = 2(0.4503) = 90.06\%$

Confidence level is = **90%**. The mean of 90 in 100 samples taken from this population will be "off" by no more than $1.74.

5/ A poll in a magazine stated that on average 36% of 980 customers who bought a new product were dissatisfied with it within a week of purchase. The poll had a margin of error of ± 3%.

Solutions:

(a) Find the upper and lower values of their estimate for the proportion of customers who were dissatisfied.

$\mu = 0.36(980) = 352.8$ and $E = 0.03(980) = 29.4$

$352.8 - 29.4 = \mathbf{323.4}$ and $352.8 + 29.4 = \mathbf{382.2}$

The poll tells us that **between 323 and 383** of 980 customers were dissatisfied.

(b) What is the probability that the true proportion of dissatisfied customers lies between these values?

$E = 0.03$ and $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.36(0.64)}{960}} = 0.0155$

We use the formula for z: $z = \frac{\pm 0.03}{0.0155} = \pm 1.9365.$

From the Normal Distribution table we find

$P(|z| < 1.94) = 2(0.4738) = 0.9476$ which means **95%**.