- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

Gwen has 2000 goats. She randomly selects 30 of her goats to measure their weights and finds an average weight of 80 pounds.

What is the sample?

- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

Gwen has 2000 goats. She randomly selects 30 of her goats to measure their weights and finds an average weight of 80 pounds.

What is the sample?

The 30 goats.

What is the population?

- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

Gwen has 2000 goats. She randomly selects 30 of her goats to measure their weights and finds an average weight of 80 pounds.

What is the sample?

The 30 goats.

What is the population?

The 2000 goats.

What is the sample statistic?

- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

Gwen has 2000 goats. She randomly selects 30 of her goats to measure their weights and finds an average weight of 80 pounds.

What is the sample?

The 30 goats.

What is the population?

The 2000 goats.

What is the sample statistic?

The 30 goats' average weight is **80 pounds**.

What is the population parameter?

- Sample statistic - How salty the spoonful is
- Population parameter - How salty the whole pot is
- We often estimate the population parameter from the sample statistic.

Gwen has 2000 goats. She randomly selects 30 of her goats to measure their weights and finds an average weight of 80 pounds.

What is the sample?

The 30 goats.

What is the population?

The 2000 goats.

What is the sample statistic?

The 30 goats' average weight is **80 pounds**.

What is the population parameter?

The unknown average weight of all 2000 goats.

Observational studies vs. experiments

Observational studies vs. experiments

- In observational studies, data are collected only by monitoring what occurs. Observational studies show associations (not causal relationships).

Observational studies vs. experiments

- In observational studies, data are collected only by monitoring what occurs. Observational studies show associations (not causal relationships).
- Experiments require the primary explanatory variable in a study be (randomly) assigned for each subject by the researchers. Experiments can show causal relationships.

Observational studies vs. experiments

- In observational studies, data are collected only by monitoring what occurs. Observational studies show associations (not causal relationships).
- Experiments require the primary explanatory variable in a study be (randomly) assigned for each subject by the researchers. Experiments can show causal relationships.

	Experiment	**Observational study**
representative sample	Best!	No causality
biased sample	Can't generalize to population	useless

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

If we are considering a causal relationship, we are suggesting that by changing the **explanatory variable**, we can expect changes in the **response variable**.

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

If we are considering a causal relationship, we are suggesting that by changing the **explanatory variable**, we can expect changes in the **response variable**.

With alcohol and driving ability, determine the explanatory variable.

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

If we are considering a causal relationship, we are suggesting that by changing the **explanatory variable**, we can expect changes in the **response variable**.

With alcohol and driving ability, determine the explanatory variable.

The explanatory variable would be whether the driver was given alcohol.

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

If we are considering a causal relationship, we are suggesting that by changing the **explanatory variable**, we can expect changes in the **response variable**.

With alcohol and driving ability, determine the explanatory variable.

The explanatory variable would be whether the driver was given alcohol.

With alcohol and driving ability, determine the response variable.

Causality

Experiments can show **causality**. For example, experiments can show alcohol impairs driving ability. How would you run this experiment?

If we are considering a causal relationship, we are suggesting that by changing the **explanatory variable**, we can expect changes in the **response variable**.

With alcohol and driving ability, determine the explanatory variable.

The explanatory variable would be whether the driver was given alcohol.

With alcohol and driving ability, determine the response variable.

The response variable would be performance on some driving task(s).

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Amount of sunscreen used.

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Amount of sunscreen used.

What is the suggested response variable?

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Amount of sunscreen used.

What is the suggested response variable?

Likelihood of skin cancer.

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Amount of sunscreen used.

What is the suggested response variable?

Likelihood of skin cancer.

What is a possible confounding (lurking) variable?

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

What is the suggested explanatory variable?

Amount of sunscreen used.

What is the suggested response variable?

Likelihood of skin cancer.

What is a possible confounding (lurking) variable?

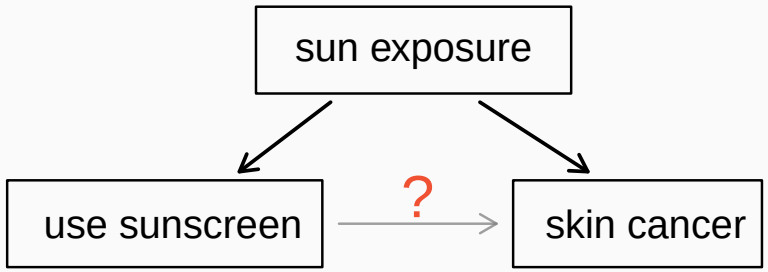
Whether the person lives in a sunny area or whether the person is pale

Confounding variables

A **confounding variable** is a third variable that may cause two other variables to have an association.

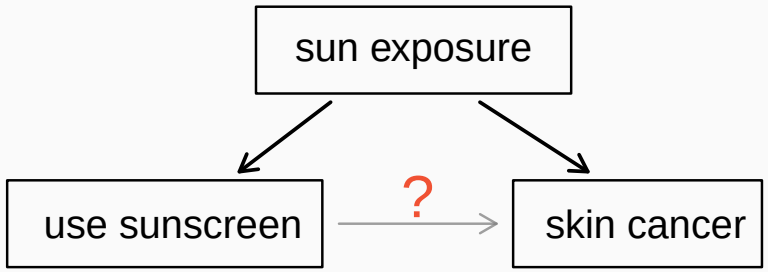
Confounding variables

A **confounding variable** is a third variable that may cause two other variables to have an association.



Confounding variables

A **confounding variable** is a third variable that may cause two other variables to have an association.



Sun exposure is a confounding variable that may explain why more suncreen is associated with more cancer.

Imagine a study finds a positive association between amount of ice cream a community eats and the rate of drowning in that community.

Imagine a study finds a positive association between amount of ice cream a community eats and the rate of drowning in that community.

Do you think ice cream causes drowning?

Imagine a study finds a positive association between amount of ice cream a community eats and the rate of drowning in that community.

Do you think ice cream causes drowning?

Do you think drownings cause ice cream?

Imagine a study finds a positive association between amount of ice cream a community eats and the rate of drowning in that community.

Do you think ice cream causes drowning?

Do you think drownings cause ice cream?

What could be a confounding variable?

Maybe hot temperatures cause more ice cream and more drowning.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.

Therefore, sleeping with one's shoes on causes headache.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.
Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.
Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Young children who sleep with the light on are much more likely to develop myopia in later life.
Therefore, sleeping with the light on causes myopia.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.
Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Young children who sleep with the light on are much more likely to develop myopia in later life.
Therefore, sleeping with the light on causes myopia.

Myopic parents are more likely to leave the light on in child's room.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.

Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Young children who sleep with the light on are much more likely to develop myopia in later life.

Therefore, sleeping with the light on causes myopia.

Myopic parents are more likely to leave the light on in child's room.

People who drink Gatorade are more likely to develop knee injuries.

Therefore, drinking Gatorade causes knee injuries.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.
Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Young children who sleep with the light on are much more likely to develop myopia in later life.
Therefore, sleeping with the light on causes myopia.

Myopic parents are more likely to leave the light on in child's room.

People who drink Gatorade are more likely to develop knee injuries.
Therefore, drinking Gatorade causes knee injuries.

Whether someone is an athlete causes higher chance of both drinking Gatorade and getting knee injuries.

Identify possible confounding variables

Sleeping with one's shoes on is strongly correlated with waking up with a headache.
Therefore, sleeping with one's shoes on causes headache.

Both may be caused by drunkiness

Young children who sleep with the light on are much more likely to develop myopia in later life.
Therefore, sleeping with the light on causes myopia.

Myopic parents are more likely to leave the light on in child's room.

People who drink Gatorade are more likely to develop knee injuries.
Therefore, drinking Gatorade causes knee injuries.

Whether someone is an athlete causes higher chance of both drinking Gatorade and getting knee injuries.

Prospective vs. retrospective studies

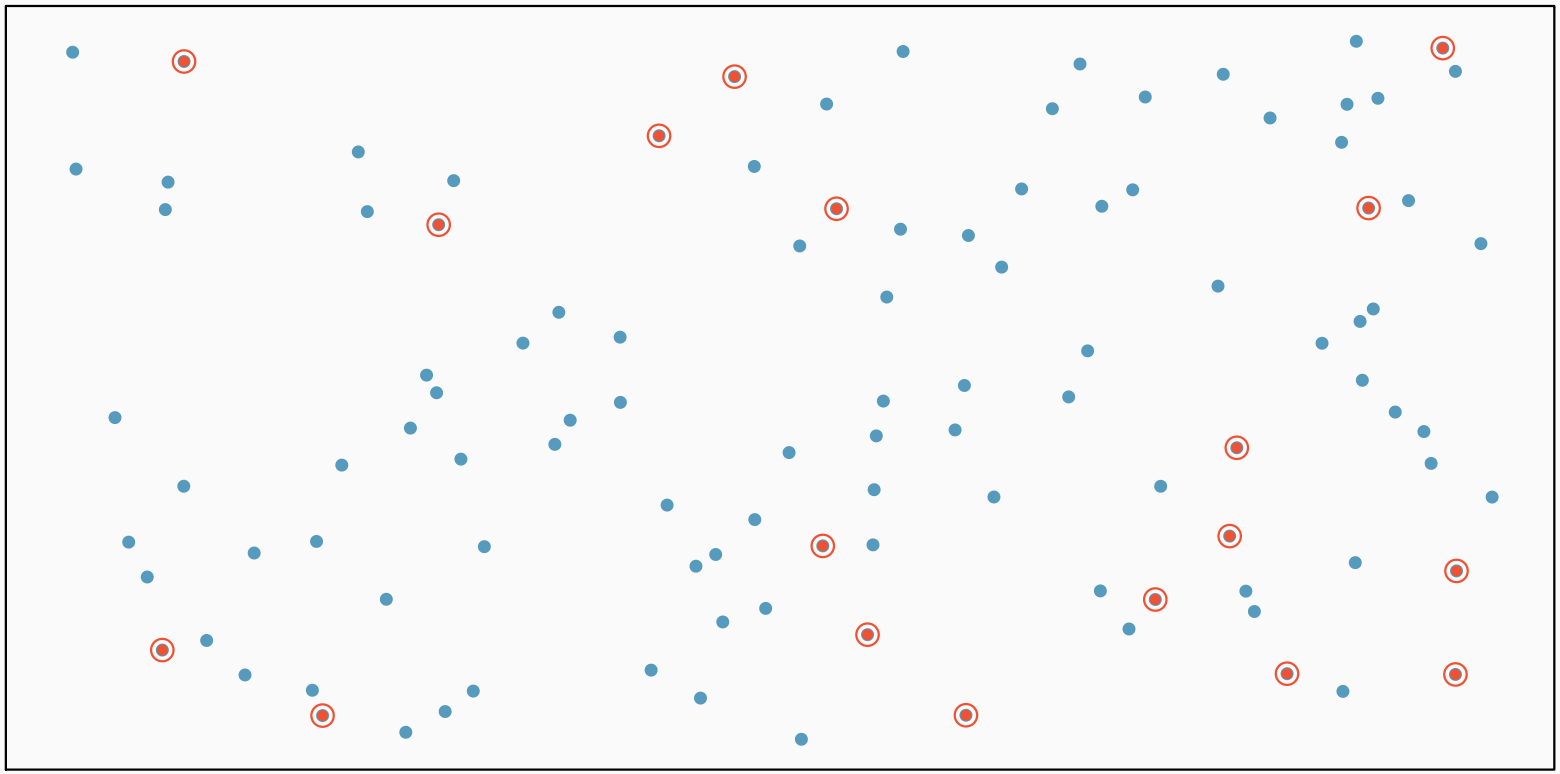
- A *prospective* study identifies individuals and collects information as events unfold.
  - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- *Retrospective studies* collect data after events have taken place.
  - Example: Researchers reviewing past events in medical records.

Obtaining good samples

- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- Most commonly used random sampling techniques are *simple*, *stratified*, and *cluster* sampling.

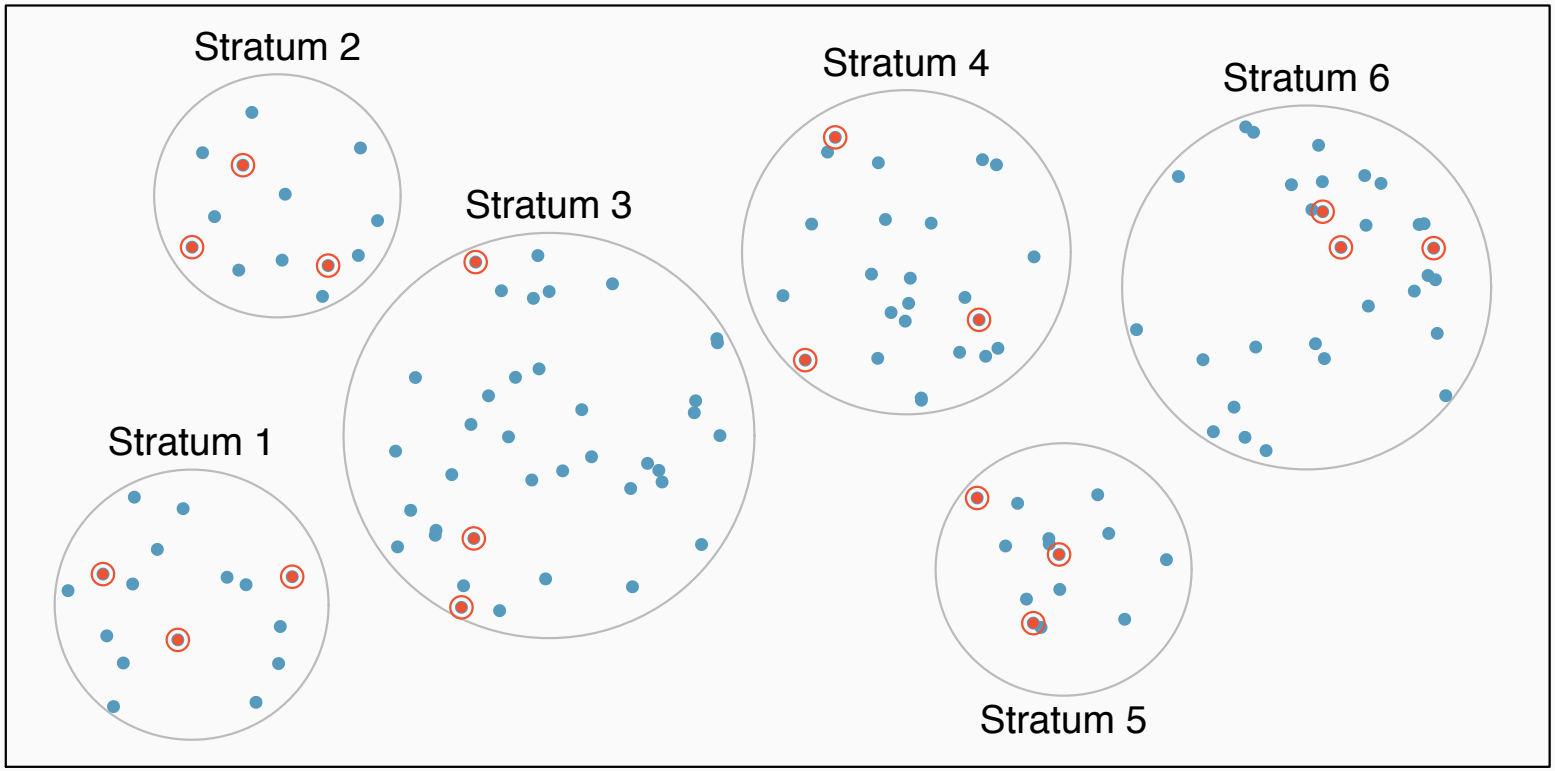
Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



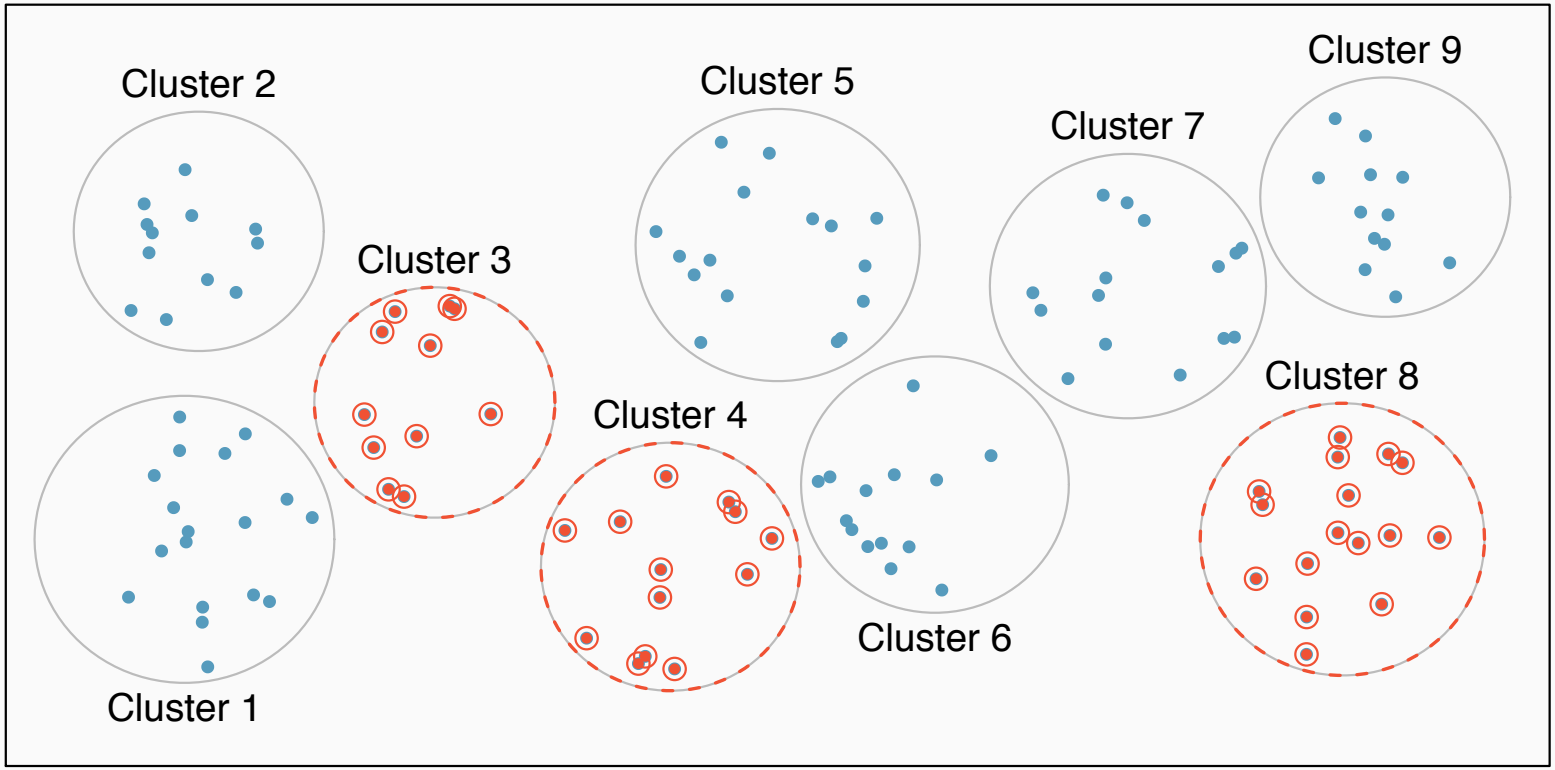
Stratified sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



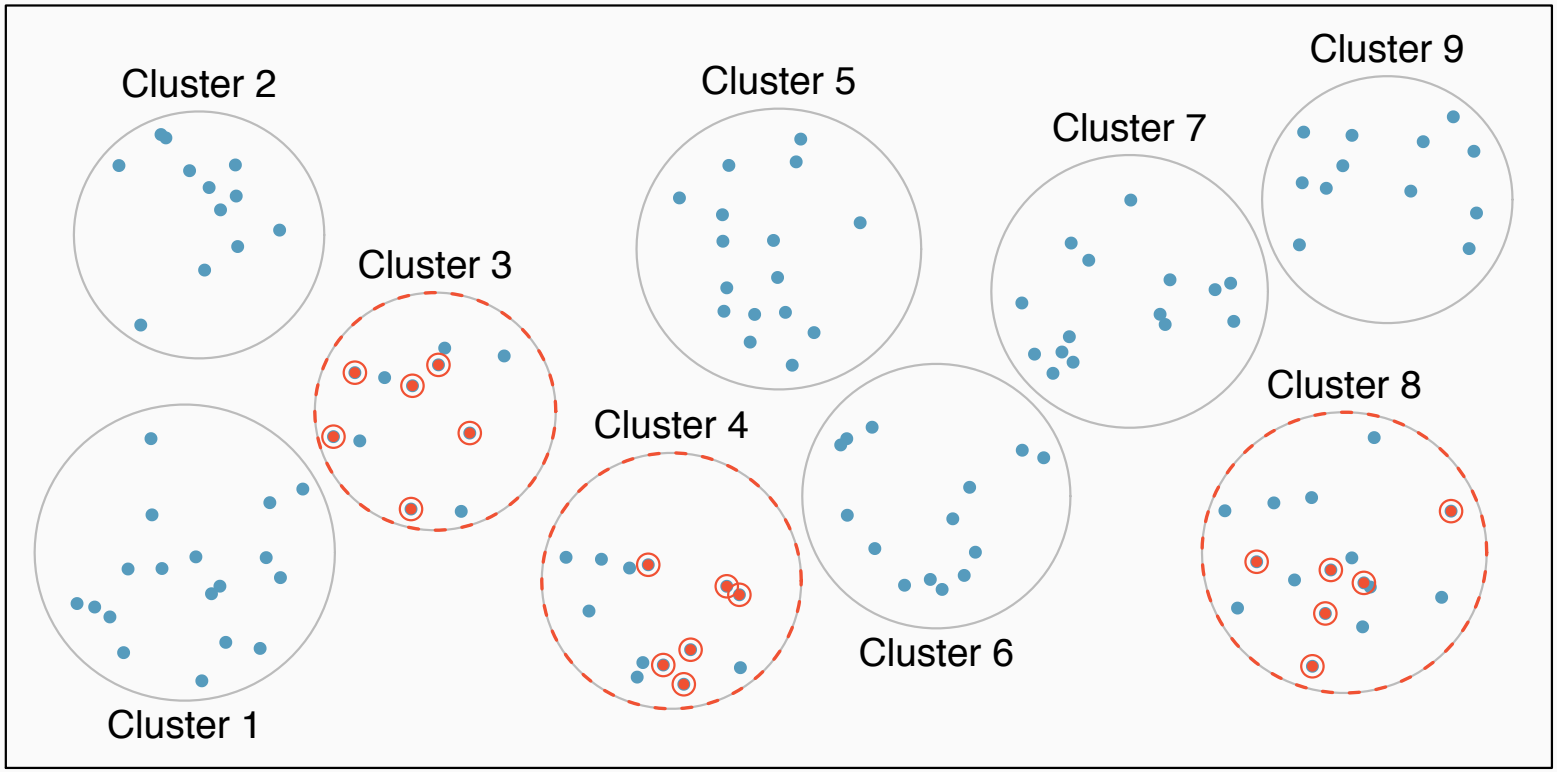
Cluster sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.



Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling

(b) Cluster sampling

(c) Stratified sampling

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling

(b) *Cluster sampling*

(c) Stratified sampling