

Sources of variation

In the real world, you carry out a research study one time. But as a thought experiment, what would happen if you were to repeat your research study?

Suppose we are estimating a quantity θ with an estimator $\hat{\theta}$.

For example, $\theta = EY - EX$ might be the expected difference between treatment and control populations in a medical study. Or θ might be the population correlation coefficient between calorie intake and activity level in 25 year old females.

In a perfect replication, the value of θ doesn't change, but we do expect to get a different value for $\hat{\theta}$.

What affects the variation in $\hat{\theta}$?

Populations

The simplest definition of a population is the set of all units (i.e. people with asthma, rats, cars, middle school students) that are of interest in a research study.

Examples:

1. If we are retrospectively studying voter participation in a past US presidential election, the population is everyone who was eligible to vote in the election.
2. If we are studying factors that were associated with a voter's decision about which candidate to vote for in a past US election, the population would be everyone who actually voted in the election.
3. If we are studying patients' responses to a new type of leukemia therapy, the population is everyone with leukemia (or everyone with leukemia who is in some sense "eligible" for the treatment).
4. If we are studying energy metabolism of Sprague Dawley rats, the population is the set of all Sprague Dawley rats.
5. If we are studying the relationship between CO₂ levels in the Earth's atmosphere and mean global temperature, the population would be ... (this is tricky, see below).

Sample and census

A **census** is a complete enumeration of every member of a population.

A **sample** is a selected subset of a population.

When a census can be performed, it aims to give exact results, not estimates.

- The decennial U.S. census aims to count every person “dwelling in a U.S. residential structure.”
- On “count day” the state of Michigan performs a census of students in all public schools in the state.

The goal of working with a sample is to identify characteristics of the sample that are likely to **generalize** to the population.

Since a sample only allows us to estimate properties of the population, it is critical that we assess our **confidence** that any generalizations we make are correct.

Sampling variation

A sample will differ from the population, and repeated samples will generally differ from each other.

Variation due to the process of sampling is called “sampling variation.”

For example, suppose the population consists of the three values 1, 2, 5, with a population mean of $8/3$. We sample two of the values, and use the sample mean to estimate the population mean. The following table gives the possible results.

Sample	Estimated mean
1,2	1.5
1,5	3.0
2,5	3.5

The variation among 1.5, 3.0, and 3.5 is the sampling variability of the sample mean \bar{X} .

Measurement error

Continuing with the previous example, suppose we are unable to exactly measure the true value.

For example, when we sample a unit with true value 1, suppose we observe $1 + \epsilon$, where ϵ has a normal distribution with mean 0 and standard deviation $\sigma = 0.2$. Then we might get

Ideal sample	Actual sample	Measurement error	Estimated mean
1,2	0.71,2.05	-0.29,0.05	1.38
1,5	1.03,5.13	0.03,0.13	3.08
2,5	1.86,5.31	-0.14,0.31	3.59

The variation among 1.38, 3.08, and 3.59 is a combination of sampling and measurement variability.

Random and systematic measurement error

Measurement error has both a mean and a variance.

If the measurement error has mean zero, the measurement is correct on average, and the measurement error is “purely random.”

If the measurement error has a nonzero mean, there is a systematic component to the measurement error (“measurement bias”).

Some common working assumptions about measurement error:

- The measurement error is independent across units.
- The measurement error is independent of the underlying true value being sampled (e.g. if measuring heights, the measurement error variance is the same for tall and short people).

Sampling and measurement variability in practice

Concrete sources of measurement error in a survey asking about voting behavior:

- Voters may incorrectly remember whether they voted, or who they voted for.
 - Innocent errors of recall may be purely random, but it is also possible that supporters of a particular candidate are more susceptible to such errors (systematic errors).
- Voters may deliberately misstate whether they voted, or who they voted for.
 - These may be treated as random measurement errors if each respondent has his or her own reasons for deciding whether to respond truthfully. But it is more likely that misstatements have a systematic component (over-statement of participation, voting in favor of the winner, or voting in favor of the candidate perceived in hindsight to be the better choice).

Sampling and measurement variability in practice

Concrete sources of measurement error in the study of leukemia treatment might depend on the “endpoint” that is used:

- Measures of tumor extent in the body like tumor cell count are subject to random measurement errors because the instruments used for this purpose have limited accuracy.
- Measures of disease symptoms (pain, anxiety, sleeplessness, etc.) are nearly always measured with considerable random and systematic measurement error because they are subjective or depend on the subjects' recall ability.
- “Hard endpoints” like overall survival (the patient's lifespan after diagnosis) are usually measured without appreciable error.

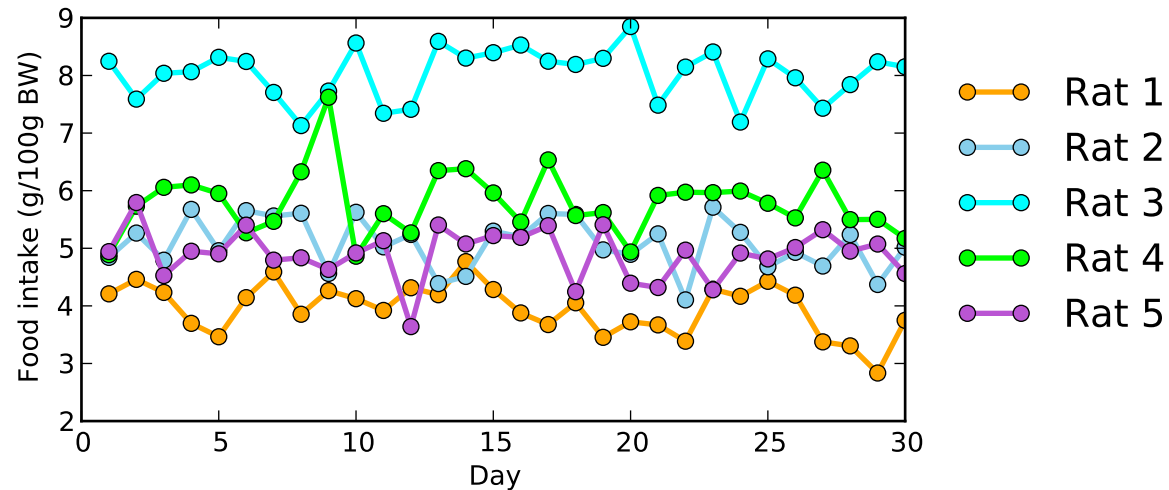
Heterogeneity

Suppose we place 10 Sprague Dawley rats in separate cages, and provide each rat with a fixed mass of food each morning for 30 days.

Every morning we weigh each rat's remaining food to determine its daily intake.

Suppose that for all practical purposes the food can be accurately weighed so there is negligible measurement error.

We might observe something like this:



There are daily fluctuations in each rat's intake, and also substantial differences between the rats.

Stable differences such as occur between the rats are called "subject heterogeneity" (perhaps due to genetic differences, or the life histories of the rats from conception up to the beginning of the study).

The stable differences (heterogeneity) and the day-to-day fluctuations represent distinct sources of variation.

The heterogeneity results from our choice of these particular 5 rats to study as representatives of the population of all Sprague Dawley rats.

Types of heterogeneity

Several common forms of heterogeneity are described below.

Baseline heterogeneity A group of patients with a particular diagnosis who are participating in a treatment trial are likely to differ before the treatment begins. This can be due to genetic or life history differences, variation in disease severity, and variation in comorbidities (other health conditions that are present).

Treatment response heterogeneity Patients who appear highly similar at baseline may respond to a treatment to substantially different degrees.

Adherence heterogeneity If a medical treatment requires some action on the part of the patient (e.g. remembering to take medication, lifestyle modification), people will adhere to the treatment to different degrees, possibly affecting their responses to treatment.

Batch heterogeneity In manufacturing, raw material is supplied in batches that are not perfectly consistent. Also many biological assays rely on reagents that are purchased in batches.

Heterogeneity from secular trends A long term study will be affected by gradual structural changes in the environment (including social, cultural, economic, demographic, technological factors, etc.).

Subpopulation heterogeneity The population of interest may actually be comprised of several distinct subpopulations. For example, if we are doing an observational study on the relationship between physical activity and heart disease, but we didn't collect information on smoking status, the sub-populations of smokers and non-smokers would differ in many relevant characteristics.

What about the geoscience example?

It's difficult to define a population in this case.

Over the history of the Earth there have been numerous periods with relatively high or low CO₂ levels.

But two historical periods with the same CO₂ level may not be directly comparable (e.g. due to different configurations of the continents, the orbital position of the Earth relative to the sun, volcanic and biological activity).

We can still use statistics to account for measurement variation, but there is no clear parent population to which we can generalize from the data.

This situation is common in dynamical systems that evolve over time (e.g. financial markets). We can study historical data in detail, but “emergent” behavior can arise that prevents us from considering the past and the future as being samples from the same system.

Sampling frames

The population is defined as the set of all physical entities of interest (i.e. eligible voters, Sprague Dawley rats, cancer patients).

In practice, it may not be practical to sample from the population of interest, so a different, related but simpler set called the “sampling frame” is defined.

For example, suppose we are interested in eligible voters for a US election that is to take place three weeks from today. Our sampling frame may be “people contactable by telephone who state that they voted in the last election.”

The sampling frame differs from the population in that not all eligible voters can be contacted by telephone, and some eligible voters may not have voted in the last election.

Blocking

Suppose we have enough information about the sources of heterogeneity in the dataset to be able to break our sample into relatively homogeneous subgroups. By focusing our analysis on comparisons within these subgroups, we can end up with a more powerful analysis. This technique is called **blocking**.

The paired t-test is the most basic analysis that uses blocking. If the goal is to analyze the difference in expected values between populations X and Y , the paired t-test can be applied when the sample can be grouped into homogeneous pairs, each consisting of one observation from the X population and one observation from the Y population.

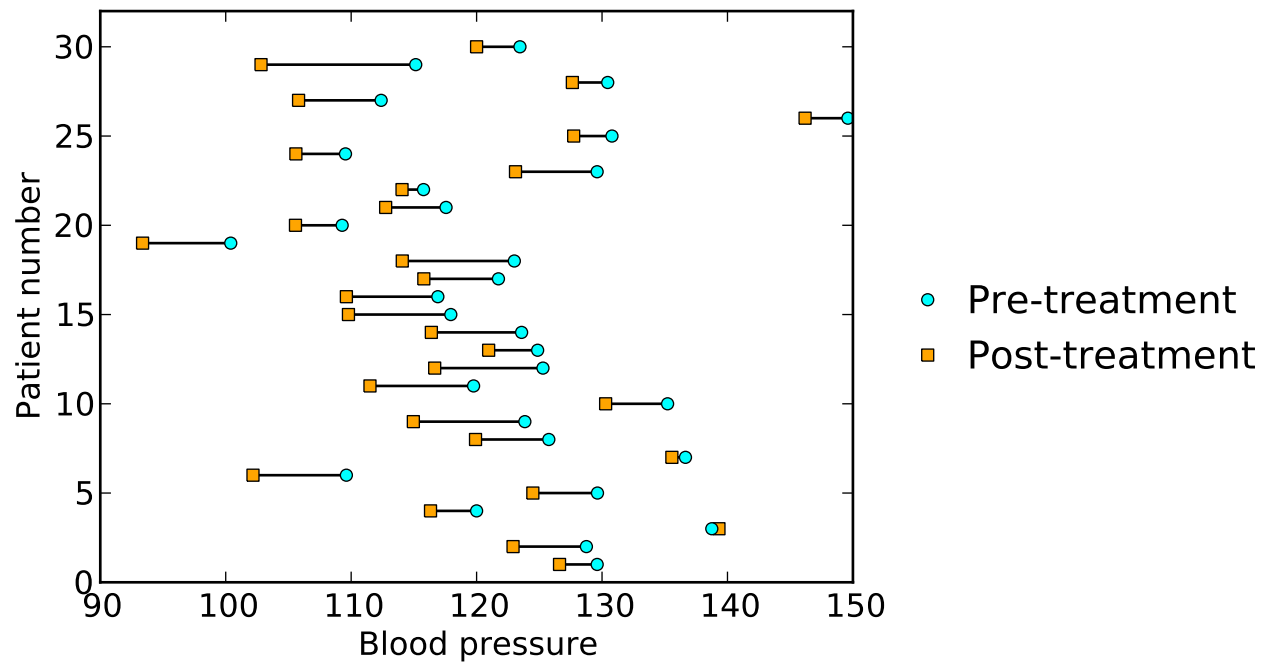
Examples:

- Each individual is his or her own control, e.g. if we measure a person's blood pressure prior to treatment, and again after treatment.
- Closely related individuals are available for comparisons, for example if we sample siblings and randomly treat one sib and give the other sib a placebo.

Paired t-test

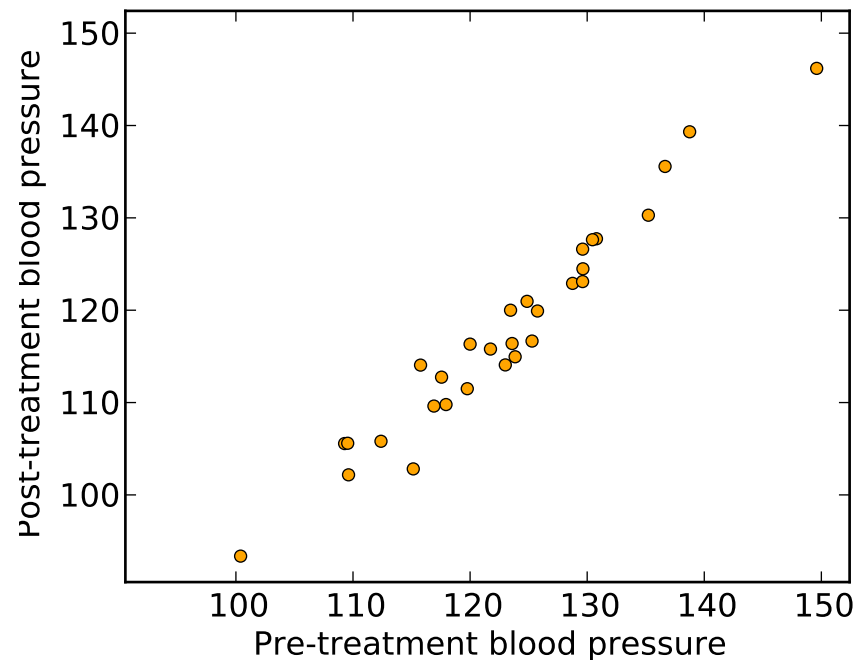
Blocking is most useful if the units within a block are very similar. For example, suppose we are considering the effectiveness of a drug designed to lower blood pressure.

Suppose we have the following data:



Paired t-test

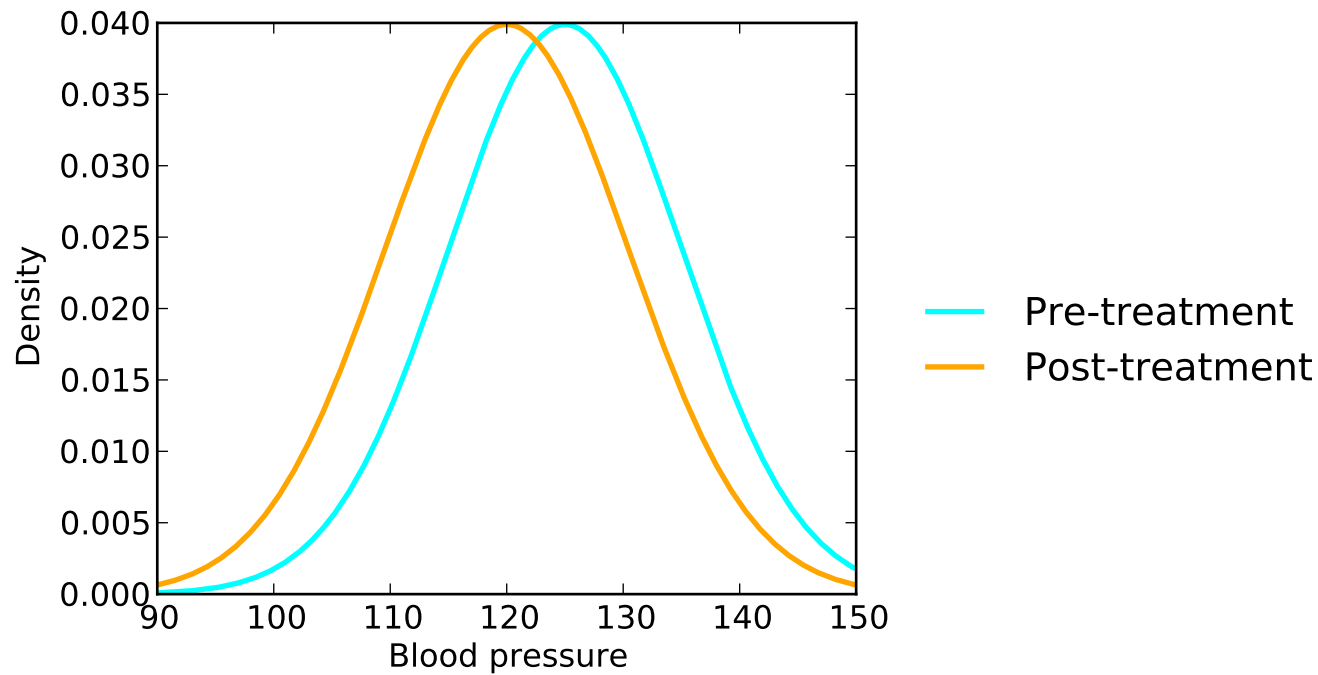
The key feature of these data is that people who have higher than average blood pressure before treatment tend to have higher than average blood pressure after treatment, and similarly for people with lower than average blood pressure:



The similarities between pre-treatment and post-treatment blood pressure are due to other risk factors (diet, exercise, smoking, genetics, ...) that are not affected by the treatment.

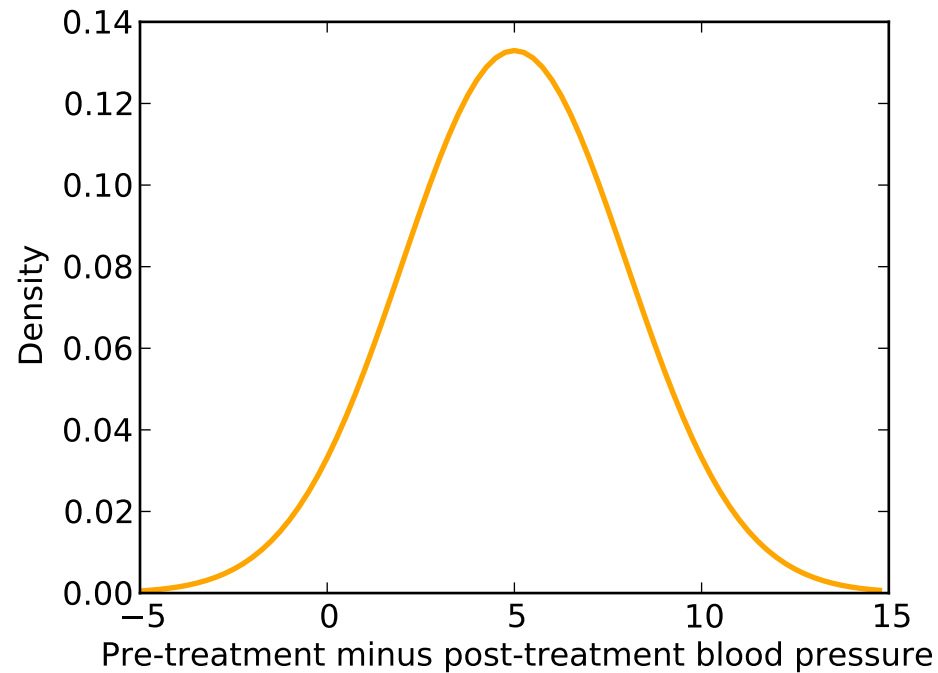
Paired t-test

The overall distributions for pre-treatment and post-treatment data are similar. If we use the standard two-sample comparison, it will require a large sample to have good power for detecting the difference:



Paired t-test

The differences between pre-treatment and post-treatment measures are overwhelmingly positive. This suggests that there may be more information in the data than we realize.



Paired t-test

To carry out the paired t-test, let X_i and Y_i be the pre-treatment and post-treatment measurements for the i^{th} subject, and let

$$D_i = X_i - Y_i$$

be the difference between them. The expected value of D_i is

$$ED_i = E(X_i - Y_i) = EX_i - EY_i = EX - EY.$$

The variance of D_i is

$$\text{var}(D_i) = \text{var}(X_i) + \text{var}(Y_i) - 2\text{cov}(X_i, Y_i),$$

which can be estimated as

$$\hat{\sigma}_D^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\widehat{\text{cov}}(X, Y).$$

Paired t-test

Under the null hypothesis $ED = EX - EY = 0$, so the test statistic

$$\sqrt{n}\bar{D}/\hat{\sigma}_D$$

is large if there is a lot of evidence in the data that EY and EX are different. A p-value can be obtained by comparing the test-statistic to a standard normal (or t) reference distribution.

Note that the $\bar{D} = \bar{X} - \bar{Y}$, so the numerator of the paired and un-paired two-sample statistics are the same. If the unpaired two-sample statistic is used to analyze paired data (ignoring the pairing), we get

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/n}} = \sqrt{n}\bar{D}/\sqrt{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2}.$$

So if $\widehat{\text{cov}}(X, Y) > 0$ (as in the example above), the paired statistic is always larger than the un-paired statistic, and hence has greater power.

Simple random samples

A simple random sample of size m is a subset of a sampling frame that is generated in such a way that any subset of size m is equally likely to be chosen as the sample.

Simple random samples are typically drawn “without replacement” (i.e. the same unit may only appear in a sample once).

It is common to analyze statistical methods as if the sampling were done “with replacement.” This way the data become iid (independent and identically distributed) and are simpler to analyze. For moderate or large samples, the results obtained when treating the sampling as being with or without replacement are very similar.

In practical terms, the only certain way to generate a simple random sample is to use a list of all members of the sampling frame, and draw entries from this list at random.

Systematic sampling

A systematic sample is generated by following a fixed rule for determining which units are in the sample.

For example, suppose we have access to the units in sequence, and every k^{th} unit in the sequence is sampled.

As a concrete example, we could take every 10th person arriving at a hospital emergency room in an ambulance. This would be a systematic sample of all patients who arrive at the hospital's emergency room by ambulance.

A systematic sample may approximate a simple random sample as long as there are no trends or periodicities in the sequential arrangement of the units.

Stratified sampling

Suppose we are studying a treatment regimen for lung cancer, and our goal is to estimate the treatment success rate (the proportion of all treated patients who respond to the treatment).

It may be that the response rates of men and women differ. In this study we aren't interested in this as a research aim, but sex-specific response may affect the precision of our measurement of overall treatment response.

Suppose the treatment population consists of a fraction q_f of females and q_m of males, where $q_f + q_m = 1$.

Stratified sampling

Suppose the treatment response rate for females is p_f , and the treatment response rate for males is p_m .

Let X be a treated patient selected at random from the sampling frame, and code treatment success as $X = 1$ and treatment failure as $X = 0$.

$$\begin{aligned}P(X = 1) &= P(X = 1|X \text{ is male}) \cdot P(X \text{ is male}) + \\ &\quad P(X = 1|X \text{ is female}) \cdot P(X \text{ is female}) \\ &= p_m q_m + p_f q_f \\ &\equiv p.\end{aligned}$$

This is the “overall treatment response rate.”

Stratified sampling

First suppose we draw a simple random sample X_1, \dots, X_n from the sampling frame, and use the sample mean $\bar{X}^{(srs)}$ to estimate p .

What are the expected value and variance of $\bar{X}^{(srs)}$? Note that

$$E\bar{X}^{(srs)} = (EX_1 + \dots + EX_n)/n,$$

and

$$\text{var}(\bar{X}^{(srs)}) = (\text{var}(X_1) + \dots + \text{var}(X_n))/n^2.$$

The “double expectation theorem” tells us that

$$EX_i = E_{\text{sex}}E(X_i|\text{sex}).$$

Since $E(X_i|\text{sex} = \text{female}) = p_f$ and $E(X_i|\text{sex} = \text{male}) = p_m$, $EX_i = q_f p_f + q_m p_m = p$. Therefore $E\bar{X} = p$, $\text{var}X_i = p(1 - p)$, and $\text{var}(\bar{X}) = p(1 - p)/n$.

Stratified sampling

You should also know a different way to get the variance:

The “law of total variation” tells us that

$$\text{var}(X_i) = \text{var}E(X_i|\text{sex}) + E\text{var}(X_i|\text{sex}).$$

Since $E(X_i|\text{sex})$ is p_f with probability q_f and p_m with probability q_m , its variance is $q_f(p_f - p)^2 + q_m(p_m - p)^2$.

Since $\text{var}(X_i|\text{sex} = \text{female}) = p_f(1 - p_f)$ and $\text{var}(X_i|\text{sex} = \text{male}) = p_m(1 - p_m)$, its mean value is $q_f p_f(1 - p_f) + q_m p_m(1 - p_m)$.

Therefore

$$\begin{aligned}\text{var}(X_i) &= q_f(p_f - p)^2 + q_m(p_m - p)^2 + q_f p_f(1 - p_f) + q_m p_m(1 - p_m) \\ &= p(1 - p).\end{aligned}$$

Stratified sampling

Now suppose that instead of a simple random sample, we randomly sample $q_f n$ females from the set of all females in the sampling frame, and $q_m n$ males from the set of all males in the sampling frame. Let $\bar{X}^{(\text{strat})}$ denote the mean of this “stratified random sample.”

What are the mean and variance of $\bar{X}^{(\text{strat})}$?

For notation, let $m_f = q_f n$, $m_m = q_m n$, $X_1^f, \dots, X_{m_f}^f$ be the female data, and $X_1^m, \dots, X_{m_m}^m$ be the male data.

$$\begin{aligned} E\bar{X}^{(\text{strat})} &= (EX_1^f + \dots + EX_{m_f}^f + EX_1^m + \dots + EX_{m_m}^m)/n \\ &= (m_f p_f + m_m p_m)/n \\ &= q_f p_f + q_m p_m \\ &= p. \end{aligned}$$

Stratified sampling

$$\begin{aligned}\text{var}\bar{X}^{(\text{strat})} &= (\text{var}X_1^f + \dots + \text{var}X_{m_f}^f + \text{var}X_1^m + \dots + \text{var}X_{m_m}^m)/n^2 \\ &= (m_f p_f (1 - p_f) + m_m p_m (1 - p_m))/n^2 \\ &= (q_f p_f (1 - p_f) + q_m p_m (1 - p_m))/n.\end{aligned}$$

With some algebra you will see that

$$\text{var}\bar{X}^{(\text{srs})} - \text{var}\bar{X}^{(\text{strat})} = \frac{q_f (p_f - p)^2 + q_m (p_m - p)^2}{n}.$$

Therefore stratified sampling is always equal to or better than simple random sampling in terms of variance. The more different the strata are (in terms of the difference between p_f and p_m), the greater the advantage.

Stratified sampling

In the preceding example, the proportion of males in the sample was the same as the proportion of males in the treatment population ($m_m = q_m n$), and also with females ($m_f = q_f n$). This is called “proportionate allocation.”

Suppose that we sample fractions w_f of females and w_m of males, where $w_f + w_m = 1$. So if the total sample size is n we sample $w_f n$ females and $w_m n$ males. If $w_f \neq q_f$ and $w_m \neq q_m$, the study is using “disproportionate allocation.”

Stratified sampling

Let \bar{X}_f and \bar{X}_m denote the sample proportions of female and male responses.

In order to unbiasedly estimate the overall response rate we need to use a weighted average of the sex-specific rates:

$$q_f \bar{X}_f + q_m \bar{X}_m.$$

The variance of this estimate is

$$q_f^2 p_f (1 - p_f) / m_f + q_m^2 p_m (1 - p_m) / m_m = n^{-1} (q_f^2 p_f (1 - p_f) / w_f + q_m^2 p_m (1 - p_m) / w_m)$$

Since $w_m = 1 - w_f$ the sampling design is determined by the value of w_m .

Exercise: Use calculus to show that the variance is minimized when $w_m = q_m S_m / (q_m S_m + q_f S_f)$ and $w_f = q_f S_f / (q_m S_m + q_f S_f)$, $S_f = \sqrt{p_f (1 - p_f)}$, $S_m = \sqrt{p_m (1 - p_m)}$.

Cluster sampling

Suppose the sampling frame is dispersed across a number of “clusters.” For example:

- We are interested in the mean mathematics test score for all 3rd grade students in the state of Michigan. The students are clustered by their classrooms.
- We are interested in the rate of serious infections among patients in intensive care units in United States hospitals. The patients are clustered by hospital.
- We are interested in the purity of pills synthesized in a pharmaceutical factory. The pills are clustered by the batch of raw materials and reagents used in the chemical synthesis.
- We are interested in whether US adults plan to make a major purchase in the next month. The adults are clustered by census tract.

Cluster sampling

For practical reasons, it is easier to construct a sample by randomly selecting a subset of the clusters, then including all units from those clusters in the sample.

This is called “cluster sampling,” and the clusters are called “primary sampling units” (PSU’s).

How does this affect the performance of the estimate?

Units in the same PSU tend to be more similar to each other than units in different PSU’s. We will see that for this reason, a cluster sample of size n gives less precision than a simple random sample of size n . In contrast, we saw above that a stratified sample of size n has more precision than a simple random sample of size n .

Cluster sampling

One way to think about cluster sampling is to model cluster i as having its own mean value μ_i , which is not directly observed. Let Y_{ij} denote the j^{th} observation in cluster i . We can model $Y_{ij}|\mu_i$ as having mean μ_i and variance σ^2 .

The μ_i values are modeled as being independent random values that follow their own distribution with mean μ and variance τ^2 .

The mean and variance of the data are described by:

$$\begin{aligned} E(Y_{ij}|\mu_i) &= \mu_i \\ \text{var}(Y_{ij}|\mu_i) &= \sigma^2 \end{aligned}$$

Cluster sampling

The goal of analysis is to estimate the overall mean μ .

The unconditional mean and variance are:

$$EY_{ij} = E_{\mu_i}E(Y_{ij}|\mu_i) = E_{\mu_i}\mu_i = \mu,$$

using the double expectation theorem.

The unconditional variance is:

$$\text{var}Y_{ij} = E_{\mu_i}\text{var}(Y_{ij}|\mu_i) + \text{var}_{\mu_i}E(Y_{ij}|\mu_i) = \sigma^2 + \tau^2.$$

Thus if we had a simple random sample of size n (i.e. sampling randomly from all units in all clusters), the mean of \bar{Y} is μ and the variance is $(\sigma^2 + \tau^2)/n$.

Cluster Sampling

To calculate the variance for cluster sampling we will need the following result first. Take two values Y_{ij} and $Y_{i'j'}$. The covariance between these values is

$$\begin{aligned}\text{cov}(Y_{ij}, Y_{i'j'}) &= E\text{cov}(Y_{ij}, Y_{i'j'} | \mu_i, \mu_{i'}) + \\ &\quad \text{cov}(E(Y_{ij} | \mu_i), E(Y_{i',j'} | \mu_{i'})) \\ &= \sigma^2 \delta_{ii'} \delta_{jj'} + \text{cov}(\mu_i, \mu_{i'}) \\ &= \sigma^2 \delta_{ii'} \delta_{jj'} + \tau^2 \delta_{ii'}.\end{aligned}$$

where δ_{ij} is 1 if $i = j$ and is 0 otherwise.

Cluster Sampling

If there are n_i observations in the i^{th} cluster, the covariance matrix for the cluster looks like this:

$$\begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 & \dots & \dots & \tau^2 & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \dots & \dots & \tau^2 & \tau^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \tau^2 & \tau^2 & \dots & \dots & \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \dots & \dots & \tau^2 & \sigma^2 + \tau^2 \end{pmatrix}$$

All between-cluster covariances are zero.

Cluster Sampling

Recall that if Y has covariance matrix Σ , then the variance of \bar{Y} is $\sum_{ij} \Sigma_{ij}/n^2$.

Suppose we sample n_i people from the i^{th} of q clusters, where $n_1 + \dots + n_q = n$.

The contribution to $\sum_{ij} \Sigma_{ij}$ from the i^{th} cluster is $n_i(\sigma^2 + \tau^2) + n_i(n_i - 1)\tau^2 = n_i\sigma^2 + n_i^2\tau^2$. Thus the variance of \bar{Y} is

$$n^{-2} \sum_i n_i\sigma^2 + n_i^2\tau^2 = \sigma^2/n + \tau^2 \sum_i (n_i/n)^2.$$

The variance of cluster sampling is greater than or equal to the variance of a simple random sample:

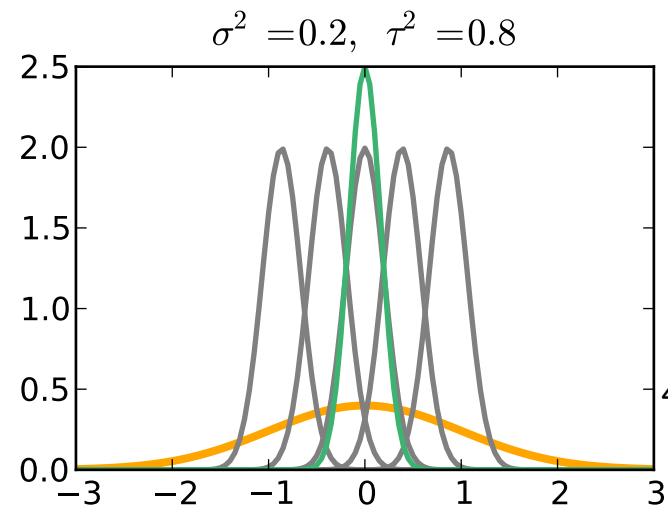
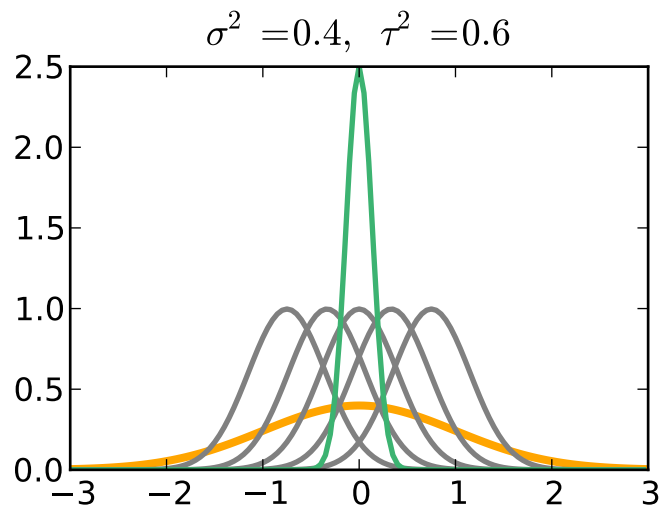
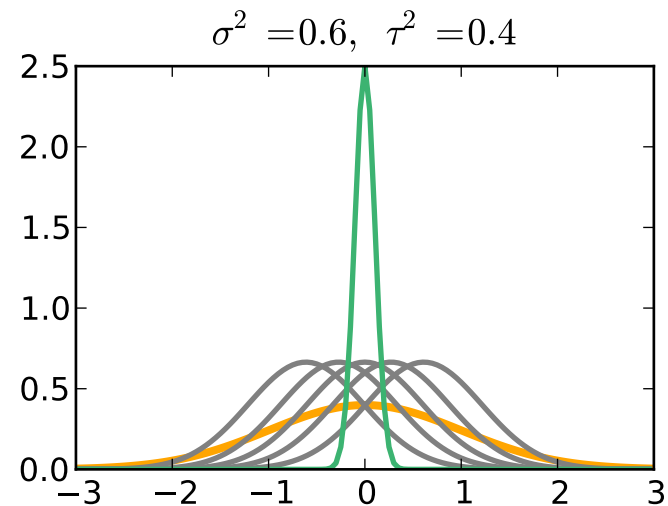
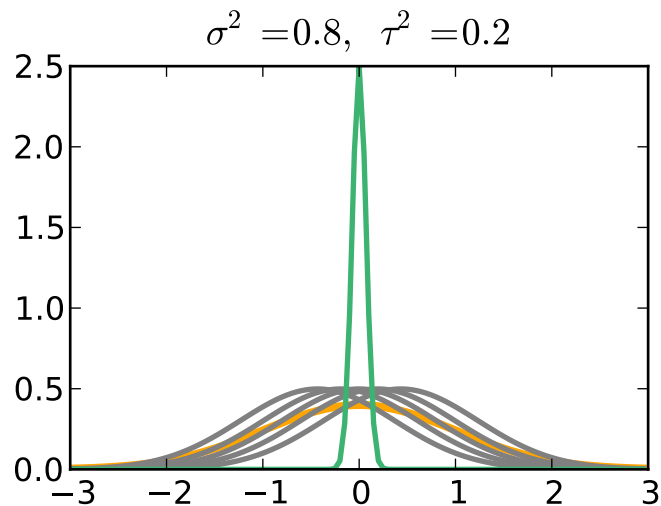
$$\sigma^2/n + \tau^2 \sum_i (n_i/n)^2 - (\sigma^2 + \tau^2)/n = \tau^2 \left(\left(\sum_i (n_i/n)^2 \right) - 1/n \right).$$

Cluster sampling and simple random sampling have the same variance in two situations:

- $\tau^2 = 0$, so all the μ_i are the same
- Every unit is in its own cluster, so every $n_i = 1$.

Otherwise, cluster sampling has greater variance than simple random sampling.

Cluster sampling with 5 clusters when the variance of one observation is $\sigma^2 + \tau^2 = 1$. The orange line is the population of X_i , the grey lines are the distributions of X_i within clusters, and the green line (not to scale) is the distribution of \bar{X} based on $n = 25$:



Summary of sampling designs

Stratified sampling Most precise of the three approaches listed here; requires that we identify and measure factors that contribute to variation in the response; somewhat more complex to analyze than a SRS.

Simple random sampling Simple to analyze, intermediate in precision among the three approaches listed here; may be difficult to carry out if the sampling frame is dispersed over a large area.

Cluster sampling Least precise of the three approaches listed here, but may be significantly easier to implement than the others if the sampling frame is naturally organized into clusters; somewhat more complex to analyze than a SRS.