# EKT-816 Lecture 2

## Probability Review (2)

Jesse Naidoo

University of Pretoria

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \ldots X_n)$ such that the joint density factors as

$$f(X_1, \ldots X_n, \theta) = g(T(X_1, \ldots X_n), \theta) \cdot h(X_1, \ldots X_n)$$

  - e.g. $T = \sum_{i=1}^{n} X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \dots X_n)$ such that the joint density factors as

$$f(X_1, \dots X_n, \theta) = g(T(X_1, \dots X_n), \theta) \cdot h(X_1, \dots X_n)$$

  - e.g. $T = \sum_{i=1}^{n} X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \dots X_n)$ such that the joint density factors as

$$f(X_1, \dots X_n, \theta) = g(T(X_1, \dots X_n), \theta) \cdot h(X_1, \dots X_n)$$

  - e.g. $T = \sum_{i=1}^n X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \dots X_n)$ such that the joint density factors as

$$f(X_1, \dots X_n, \theta) = g(T(X_1, \dots X_n), \theta) \cdot h(X_1, \dots X_n)$$

  - e.g. $T = \sum_{i=1}^n X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \ldots X_n)$ such that the joint density factors as

$$f(X_1, \ldots X_n, \theta) = g(T(X_1, \ldots X_n), \theta) \cdot h(X_1, \ldots X_n)$$

  - e.g. $T = \sum_{i=1}^{n} X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- suppose there is a statistic $T(X_1, \ldots X_n)$ such that the joint density factors as

$$f(X_1, \ldots X_n, \theta) = g(T(X_1, \ldots X_n), \theta) \cdot h(X_1, \ldots X_n)$$

  - e.g. $T = \sum_{i=1}^{n} X_i$ for normal data with known variance
- then, a maximum likelihood estimator must be a function of $T$
- in fact, the *Rao-Blackwell Theorem* says (roughly) that any unbiased estimator which is *not* a function of the sufficient statistic has higher variance than "necessary"
  - more precisely, higher variance than the MLE, which hits the (Cramer-Rao) lower bound
  - intuition: if you base estimates on irrelevant information, you are sacrificing precision

# Sufficient Statistics

- we will not use an explicit likelihood framework much
  - but, the idea of "sufficiency" is still useful
  - in some situations, all of the relevant information can be reduced to some low-dimensional summaries
  - see Chetty (2009) and Weyl (2019) for examples of how this idea connects theory and econometrics
- this idea also comes up in the guise of "selection (only) on observables"

# Sufficient Statistics

- we will not use an explicit likelihood framework much
  - but, the idea of "sufficiency" is still useful
  - in some situations, all of the relevant information can be reduced to some low-dimensional summaries
  - see Chetty (2009) and Weyl (2019) for examples of how this idea connects theory and econometrics
- this idea also comes up in the guise of "selection (only) on observables"

# Sufficient Statistics

- we will not use an explicit likelihood framework much
  - but, the idea of "sufficiency" is still useful
  - in some situations, all of the relevant information can be reduced to some low-dimensional summaries
  - see Chetty (2009) and Weyl (2019) for examples of how this idea connects theory and econometrics
- this idea also comes up in the guise of "selection (only) on observables"

# Sufficient Statistics

- we will not use an explicit likelihood framework much
  - but, the idea of "sufficiency" is still useful
  - in some situations, all of the relevant information can be reduced to some low-dimensional summaries
  - see Chetty (2009) and Weyl (2019) for examples of how this idea connects theory and econometrics
- this idea also comes up in the guise of "selection (only) on observables"

# Sufficient Statistics

- we will not use an explicit likelihood framework much
  - but, the idea of "sufficiency" is still useful
  - in some situations, all of the relevant information can be reduced to some low-dimensional summaries
  - see Chetty (2009) and Weyl (2019) for examples of how this idea connects theory and econometrics
- this idea also comes up in the guise of "selection (only) on observables"

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \dots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \dots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \dots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \dots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \dots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \dots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Example: Sufficiency and Comparison of Estimators

- say we have an iid sample of size $n$ from a $U(0, \theta_0)$ distribution
  - the sample maximum is, in this case, sufficient for $\theta_0$
  - in fact, can show the MLE is $\widehat{\theta} = \max\{X_1, \ldots X_n\}$
- another estimator would be $\widetilde{\theta} = 2\overline{X}_n$
  - this is unbiased (show this!)
- which estimator has lower MSE? Which has lower variance?
- to derive the distribution of the sample maximum:
  - use the fact that $\max\{X_1, \ldots X_n\} \leq x$ if and only if each $X_i \leq x$
  - by independence, the CDF of the sample maximum is the product of the individual CDFs

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
    - cross-sectional:
        - units from the population are surveyed once, and all at roughly the same time.
        - a “snapshot” of the population
    - stratified or two-stage designs
        - units are sampled randomly within certain pre-specified groups
        - e.g region, race, sex
    - clustered designs
        - often would be expensive to collect a simple random sample
        - save on transport and labor costs by selecting clusters of units
        - attempt to correct for the resulting correlations (why?)
    - panel or longitudinal designs: repeated observations on the same units
        - rotating panels, where some units are "rotated" in and out of the survey
        - retrospective histories
        - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a “snapshot” of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are “rotated” in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a “snapshot” of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are “rotated” in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a “snapshot” of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are “rotated” in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a “snapshot” of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are “rotated” in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- some types of data you may encounter:
  - cross-sectional:
    - units from the population are surveyed once, and all at roughly the same time.
    - a "snapshot" of the population
  - stratified or two-stage designs
    - units are sampled randomly within certain pre-specified groups
    - e.g region, race, sex
  - clustered designs
    - often would be expensive to collect a simple random sample
    - save on transport and labor costs by selecting clusters of units
    - attempt to correct for the resulting correlations (why?)
  - panel or longitudinal designs: repeated observations on the same units
    - rotating panels, where some units are "rotated" in and out of the survey
    - retrospective histories
    - synthetic panels: aggregate individuals to form a panel at cohort level

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
    - they are expensive to collect
    - differential attrition can be a serious problem
- not all data are survey data
    - e.g. administrative data (e.g. tax records)
    - these sources can be much more complete and data quality can be high
    - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
  - they are expensive to collect
  - differential attrition can be a serious problem
- not all data are survey data
  - e.g. administrative data (e.g. tax records)
  - these sources can be much more complete and data quality can be high
  - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
    - they are expensive to collect
    - differential attrition can be a serious problem
- not all data are survey data
    - e.g. administrative data (e.g. tax records)
    - these sources can be much more complete and data quality can be high
    - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
  - they are expensive to collect
  - differential attrition can be a serious problem
- not all data are survey data
  - e.g. administrative data (e.g. tax records)
  - these sources can be much more complete and data quality can be high
  - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
  - they are expensive to collect
  - differential attrition can be a serious problem
- not all data are survey data
  - e.g. administrative data (e.g. tax records)
  - these sources can be much more complete and data quality can be high
  - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
  - they are expensive to collect
  - differential attrition can be a serious problem
- not all data are survey data
  - e.g. administrative data (e.g. tax records)
  - these sources can be much more complete and data quality can be high
  - for work on developed countries, the current frontier in labor and public economics

# Sample Design

- panel data can overcome some problems related to unobservable variables, but
  - they are expensive to collect
  - differential attrition can be a serious problem
- not all data are survey data
  - e.g. administrative data (e.g. tax records)
  - these sources can be much more complete and data quality can be high
  - for work on developed countries, the current frontier in labor and public economics

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$
\begin{aligned}
E\left[\sum_{i=1}^{n} w_i\right] &= E\left[\sum_{i=1}^{N} t_i w_i\right] \\
&\approx \sum_{i=1}^{N} (\pi_i n) w_i = N
\end{aligned}
\qquad (1)
$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
    - let $\pi_i$ be the probability that unit $i$ is selected
    - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
    - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$E\left[\sum_{i=1}^{n} w_i\right] = E\left[\sum_{i=1}^{N} t_i w_i\right] \approx \sum_{i=1}^{N} (\pi_i n) w_i = N \qquad (1)$$

- here $t_i$ is the number of times unit $i$ is included in the sample
- if $n << N$, this is about $\pi_i n$
- so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$E\left[\sum_{i=1}^{n} w_i\right] = E\left[\sum_{i=1}^{N} t_i w_i\right] \approx \sum_{i=1}^{N} (\pi_i n) w_i = N \quad (1)$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$E\left[\sum_{i=1}^{n} w_i\right] = E\left[\sum_{i=1}^{N} t_i w_i\right] \approx \sum_{i=1}^{N} (\pi_i n) w_i = N \quad (1)$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$\begin{aligned} E\left[\sum_{i=1}^{n} w_i\right] &= E\left[\sum_{i=1}^{N} t_i w_i\right] \\ &\approx \sum_{i=1}^{N} (\pi_i n) w_i = N \end{aligned} \tag{1}$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$\begin{aligned} E\left[\sum_{i=1}^{n} w_i\right] &= E\left[\sum_{i=1}^{N} t_i w_i\right] \\ &\approx \sum_{i=1}^{N} (\pi_i n) w_i = N \end{aligned} \tag{1}$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$\begin{aligned} E\left[\sum_{i=1}^{n} w_i\right] &= E\left[\sum_{i=1}^{N} t_i w_i\right] \\ &\approx \sum_{i=1}^{N} (\pi_i n) w_i = N \end{aligned} \tag{1}$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- suppose we have a population of $N$ units, and we sample with unequal probabilities
  - let $\pi_i$ be the probability that unit $i$ is selected
  - in a simple random sample of size $n << N$, $\pi_i \approx n/N$, even if we sample without replacement
  - but we don't always want to sample each unit with equal probability
- let $w_i = (n\pi_i)^{-1}$ be the "design weight"; then the expected sum of the weights is

$$\begin{aligned} E\left[\sum_{i=1}^{n} w_i\right] &= E\left[\sum_{i=1}^{N} t_i w_i\right] \\ &\approx \sum_{i=1}^{N} (\pi_i n) w_i = N \end{aligned} \tag{1}$$

  - here $t_i$ is the number of times unit $i$ is included in the sample
  - if $n << N$, this is about $\pi_i n$
  - so, the weights can be used to estimate the total size of the population

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
    - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

    - exercise: show that $E[\overline{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
    - e.g. white or Indian South Africans
    - this would be true even if response rates were the same across ethnic groups!

# Weighting

- intuition: weights are inversely proportional to the probabilities of inclusion
  - in a 1/100 sample, each included unit represents 100 others
- to estimate the population mean, we weight the observations by $w_i$, forming

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

  - exercise: show that $E[\bar{x}_w] = E[X]$
- we may want to deliberately oversample some groups to improve precision of conditional means
  - e.g. white or Indian South Africans
  - this would be true even if response rates were the same across ethnic groups!

# Stratification

- to understand why stratification is useful, think about trying to measure the national average of some $X$ where there are two cities
- variance of a randomly sampled unit is

$$V[X_i] = p\sigma_1^2 + (1-p)\sigma_2^2 + p(\mu_1 - \mu)^2 + (1-p)(\mu_2 - \mu)^2$$

  - notice that $V[X] = E[V[X|S]] + V[E[X|S]]$

# Stratification

- to understand why stratification is useful, think about trying to measure the national average of some $X$ where there are two cities
- variance of a randomly sampled unit is

$$V[X_i] = p\sigma_1^2 + (1-p)\sigma_2^2 + p(\mu_1 - \mu)^2 + (1-p)(\mu_2 - \mu)^2$$

- notice that $V[X] = E[V[X|S]] + V[E[X|S]]$

# Stratification

- to understand why stratification is useful, think about trying to measure the national average of some $X$ where there are two cities
- variance of a randomly sampled unit is

$$V[X_i] = p\sigma_1^2 + (1-p)\sigma_2^2 + p(\mu_1 - \mu)^2 + (1-p)(\mu_2 - \mu)^2$$

  - notice that $V[X] = E[V[X|S]] + V[E[X|S]]$

# Stratification

- consider a stratified design where we take two independent samples from the two strata
  - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
  - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \end{aligned} \tag{3}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
  - then

$$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$

  - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
  - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
    - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
    - the variance of the sample mean is

$$
\begin{aligned}
V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\
&= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \qquad (3)
\end{aligned}
$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
    - then

$$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$

    - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
    - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
    - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
    - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \end{aligned} \qquad (3)$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
    - then

$$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$

    - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
    - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
    - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
    - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \end{aligned} \tag{3}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
    - then

$$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$

    - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
    - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
    - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
    - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \end{aligned} \tag{3}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
    - then
    $$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$
    - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
    - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
  - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
  - the variance of the sample mean is

$$V[\overline{x}^{STRAT}] = V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] = \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \tag{3}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
  - then
    $$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$
  - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
  - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
  - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
  - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \end{aligned} \tag{3}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
  - then

$$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$

  - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
  - how is this different to weighting?

# Stratification

- consider a stratified design where we take two independent samples from the two strata
    - size $n_1$ and $n_2$ respectively, with $n_1 + n_2 = n$
    - the variance of the sample mean is

$$\begin{aligned} V[\overline{x}^{STRAT}] &= V\left[n^{-1}\sum_{i=1}^{n_1} x_i + n^{-1}\sum_{i=n_1+1}^{n} x_i\right] \\ &= \left(\frac{n_1}{n}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma_2^2}{n_2} \qquad (3) \end{aligned}$$

- say we choose $n_1/n = p$ and $n_2/n = 1 - p$
    - then
    $$V[\overline{x}^{STRAT}] = n^{-1}\{p\sigma_1^2 + (1-p)\sigma_2^2\} < n^{-1}V[X]$$
    - as long as the means differ ($\mu_1 \neq \mu_2$), stratification improves precision (why?)
- notice that the motivation here is to improve precision of the overall mean
    - how is this different to weighting?

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma^2_\alpha$ be the variance of the cluster-specific mean
  - $\sigma^2_\varepsilon$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma^2_\alpha + \sigma^2_\varepsilon$ is the overall variance of $X$
  - $\rho = \sigma^2_\alpha / \sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} \quad = \quad \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] \quad = \quad \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Clustering

- suppose we have several clusters, indexed by $c$
- the distribution of the variable of interest, $X$, obeys the following:

$$x_{ic} = \mu + \alpha_c + \varepsilon_{ic} \tag{4}$$

  - here $\alpha$ and $\varepsilon$ are independent of each other, and both have mean zero
  - let $\sigma_\alpha^2$ be the variance of the cluster-specific mean
  - $\sigma_\varepsilon^2$ be the variance of the idiosyncratic error term $\varepsilon_{ic}$
- the mean of sample from $k$ clusters, with $m$ units per cluster, has variance

$$V[\overline{x}^{CLUST}] = \frac{\sigma^2}{km}\{(m-1)\rho + 1\} \tag{5}$$

  - here $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ is the overall variance of $X$
  - $\rho = \sigma_\alpha^2/\sigma^2$ is the intercluster correlation coefficient
- positive correlations within a cluster reduce the precision of your estimates
  - effectively less information from each observation from the same cluster
  - in limiting case $\rho = 1$, the effective sample size is $k$, not $km$

# Hypothesis Testing

# References

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1 (1): 451–88. `https://doi.org/10.1146/annurev.economics.050708.142910`.

Weyl, E. Glen. 2019. "Price Theory." *Journal of Economic Literature (Forthcoming)*, no. August.

# Table of Contents