

EKT-816 Lecture 2

Probability Review (2)

Jesse Naidoo

University of Pretoria

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

Desirable Properties of Estimators

- *consistency*: $\hat{\theta} \rightarrow_p \theta_0$
- if $E[\hat{\theta}] = \theta_0$ we say $\hat{\theta}$ is *unbiased*
- *efficiency* or *precision*: say we have two estimators $\hat{\theta}$ and $\tilde{\theta}$
 - for now assume both are unbiased
 - if $V[\hat{\theta}] \leq V[\tilde{\theta}]$, say that $\hat{\theta}$ is *more efficient* than $\tilde{\theta}$
- the *mean square error* of $\hat{\theta}$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]$
 - easy to see that $\text{MSE} = V[\hat{\theta}] + \text{bias}^2$
 - often a trade-off between the two criteria
 - typically people seek unbiased estimators, but not always clear they are better in a MSE sense

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 $T(X_1, \dots, X_n)$ 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

- 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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 θ_0
 - in fact, can show the MLE is $\hat{\theta} = \max\{X_1, \dots, X_n\}$
- another estimator would be $\tilde{\theta} = 2\bar{X}_n$
 - 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

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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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 π_i be the probability that unit i is selected
 - in a simple random sample of size $n \ll N$, $\pi_i \approx n/N$, 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 $n \ll N$, 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

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

- 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!

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

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

- 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!

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

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

- 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!

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

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

- 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!

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

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

- 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!

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

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

- 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!

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

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

- 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
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

Stratification

- to understand why stratification is useful, think about trying to measure the national average of some X where there are two cities
 - a fraction p lives in city 1
 - the mean and variance of X (say, income) are μ_1 and σ_1^2 in city 1
 - let $\mu = p\mu_1 + (1 - p)\mu_2$ be the national average
- the 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 $V[X] = E[V[X|S]] + V[E[X|S]]$
- the idea of stratification is to turn one sample into two independent ones
 - by exploiting the known drivers of heterogeneity, you can improve precision

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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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

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

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

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

- 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:

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

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

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

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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:

$$X_{ic} = \mu + \alpha_c + \varepsilon_{ic} \quad (4)$$

- here α and ε are independent of each other, and both have mean zero
- let σ_α^2 be the variance of the cluster-specific mean
- σ_ε^2 be the variance of the idiosyncratic error term ε_{ic}
- the mean of sample from k clusters, with m units per cluster, has variance

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

- 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

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- a *hypothesis* is a subset of the parameter space
 - if this is a single point, we say the hypothesis is *simple*
 - e.g. $H_0 : \mu = 1$
 - otherwise, we say the hypothesis is *complex* or *compound*
 - e.g. $H_1 : \mu \neq 1$
- we often designate one particular hypothesis as the “null hypothesis”
 - then, see if the data provides strong enough evidence against it
- the frequentist approach to hypothesis testing takes parameters as fixed and the data as random
 - thus $P(\mu = 1|X)$ makes no sense, but $P(X|\mu = 1)$ does
 - the Bayesian approach is more intuitive here, but econometrics is overwhelmingly frequentist

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is "small"
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is "small"
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is "small"
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is “small”
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is “small”
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is “small”
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is “small”
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- to perform a *test* of the null hypothesis we form a *test statistic*, say $\hat{S}(X_1, \dots, X_n)$
 - not always the estimator itself, e.g. the *t*-test
 - then, we need to compute (at least approximately) the distribution of \hat{S} under the null
- suppose we know the distribution of \hat{S} under H_0
 - now, we can set a *rejection region* R such that $P(\hat{S} \in R | H_0)$ is “small”
 - given a rejection region R ,

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 | \theta)$$

is called the *size* of the test, or the *significance level*

- we take the maximum over H_0 (in case H_0 is a compound hypothesis) to be conservative
- the number

$$\beta = P(\text{reject } H_0 | \theta)$$

is called the *power* of the test

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

Hypothesis Testing

- consider some extreme cases:
 - could ignore the data and never reject: then you never make Type I errors
 - similarly could always reject: never make Type II errors
- in general there is a tradeoff between size and power
- how well you can do, and how severe the tradeoff is, depends on the problem
 - in some “ill-posed” problems, you cannot beat the trivial test
 - i.e. ignore the data, generate a random number $U \sim U(0, 1)$ and reject if $U < \alpha$
 - these pathologies often arise when there are “nuisance parameters”
 - in those cases, the requirement to keep the size of the test low imposes so many constraints you cannot have nontrivial power

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. doi:10.1146/annurev.economics.050708.142910.
- Weyl, E. Glen. 2019. “Price Theory.” *Journal of Economic Literature* (Forthcoming), no. August.

Table of Contents

Desirable Properties of Estimators

Weighting and Sample Design

Hypothesis Testing