Machine Learning and Statistics in Genetics and Genomics

III: Introduction to hypothesis testing

Christoph Lippert

Microsoft Research
eScience group

Los Angeles , USA

Microsoft Research

Current topics in computational biology
UCLA
Winter quarter 2014

Outline

Outline

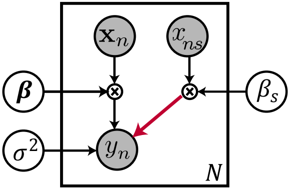
Hypothesis Testing

- Introduction
- P-values and significance
- t-test in linear regression
- Likelihood ratio test
- Multiple Hypothesis Testing
- Model checking - useful heuristics

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $x_{n,s}$: SNP to be tested
- remaining x_n: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



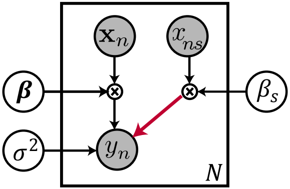
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2\right)$$

- $x_{n,s}$: SNP to be tested
- remaining x_n: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



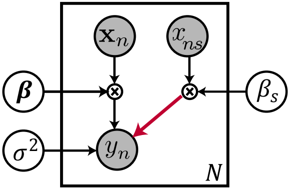
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $x_{n,s}$: SNP to be tested
- remaining $\boldsymbol{x}_n$: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



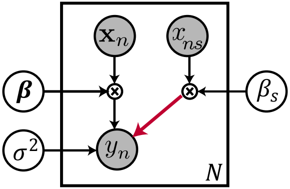
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0$: "The true underlying β_s that generated the data is 0 for the SNP s."
  (true β unknown)
- Use the estimate $\beta_{s\text{ML}}$ as a test statistic.
- **Intuition:** The larger the absolute value of the estimate $\beta_{s\text{ML}}$, the less likely is $\mathcal{H}_0 : \beta_s = 0$.



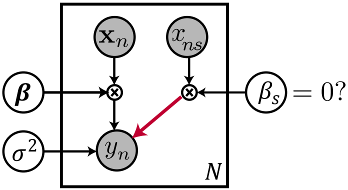
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0$: "The true underlying β_s that generated the data is 0 for the SNP s."
  (true $\boldsymbol{\beta}$ unknown)
- Use the estimate $\beta_{s\mathrm{ML}}$ as a test statistic.
- **Intuition:** The larger the absolute value of the estimate $\beta_{s\mathrm{ML}}$, the less likely is $\mathcal{H}_0 : \beta_s = 0$.



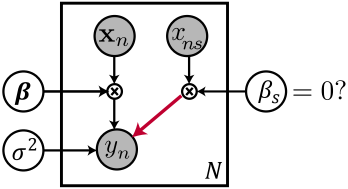
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \,|\, \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0$: "The true underlying β_s that generated the data is 0 for the SNP s."
  (true $\boldsymbol{\beta}$ unknown)
- Use the estimate $\beta_{s\text{ML}}$ as a test statistic.
- **Intuition:** The larger the absolute value of the estimate $\beta_{s\text{ML}}$, the less likely is $\mathcal{H}_0 : \beta_s = 0$.



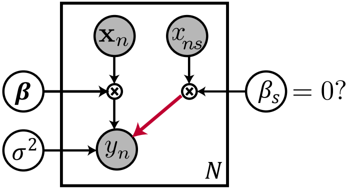
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0$: "The true underlying β_s that generated the data is 0 for the SNP s."
  (true $\boldsymbol{\beta}$ unknown)
- Use the estimate $\beta_{s\text{ML}}$ as a test statistic.
- **Intuition:** The larger the absolute value of the estimate $\beta_{s\text{ML}}$, the less likely is $\mathcal{H}_0 : \beta_s = 0$.



Equivalent graphical model

x_n: regression covariates

Hypothesis Testing

Some definitions

Example:

- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing

Some definitions

Example:

- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing

Some definitions

Example:

- Given a sample $\mathcal{D} = \{x_1, \dots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing
Some definitions

Example:

- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Hypothesis Testing

Some definitions

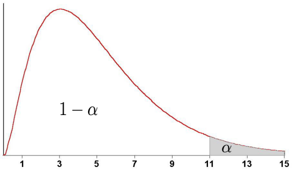
Example:

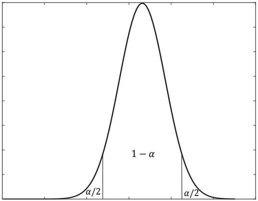
- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Hypothesis Testing

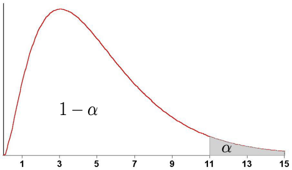
- Given a sample $\mathcal{D} = \{x_1, \dots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** $\mathcal{R}_\alpha$ defines the values of the test statistic that lead to a rejection of the test at significance α.

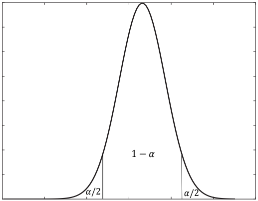




Hypothesis Testing

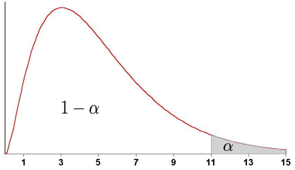
- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** $\mathcal{R}_\alpha$ defines the values of the test statistic that lead to a rejection of the test at significance α.

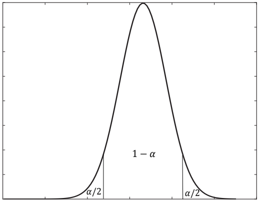




Hypothesis Testing

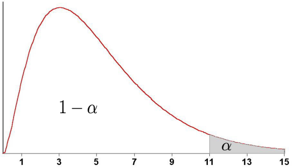
- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** $\mathcal{R}_\alpha$ defines the values of the test statistic that lead to a rejection of the test at significance α.

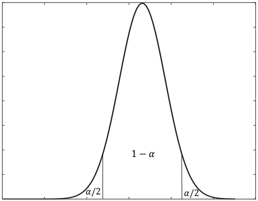




Hypothesis Testing

- Given a sample $\mathcal{D} = \{x_1, \ldots, x_N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** $\mathcal{R}_\alpha$ defines the values of the test statistic that lead to a rejection of the test at significance α.





P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \mathrm{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \mathrm{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

definition

- P-value of a test statistic x is the largest possible α, such that x is still rejected.

$$P - \text{value}(x) = \inf_{\alpha} (x \in \mathcal{R}_\alpha)$$

- Probability of observing a test statistic at least as extreme as x, given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- For every $u \in [0, 1]$,

$$P_{\mathcal{H}_0}(P - \text{value}(x) \leq u) = P_{\mathcal{H}_0}(x \in \mathcal{R}_u) = u$$

- It follows that under $\mathcal{H}_0$ the P-values are uniformly distributed in the interval $[0, 1]$.

P-value

Permutation procedure

Repeat M times:

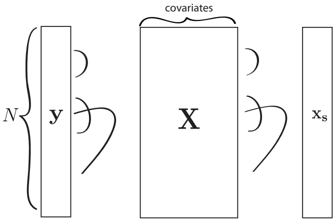
- Permute phenotype y and covariates x jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

P-value

Permutation procedure

Repeat M times:

- Permute phenotype $\boldsymbol{y}$ and covariates $\boldsymbol{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

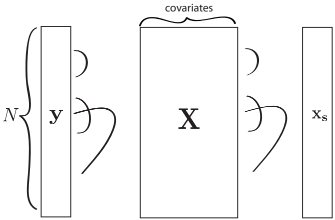


P-value

Permutation procedure

Repeat M times:

- Permute phenotype $\boldsymbol{y}$ and covariates $\boldsymbol{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

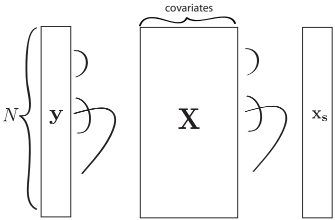


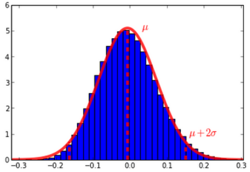
P-value

Permutation procedure

Repeat M times:

- Permute phenotype y and covariates x jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution



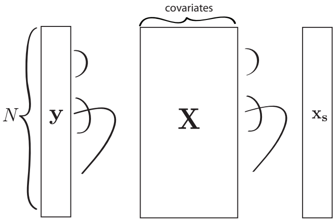


P-value

Permutation procedure

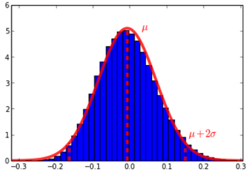
Repeat M times:

- Permute phenotype $\boldsymbol{y}$ and covariates $\boldsymbol{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution



The P-value is the quantile of real test statistic in artificial null distribution.

- The quantile is the fraction of the empirical distribution that is more extreme than the test statistic.



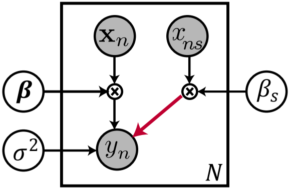
Testing in Linear Regression

Analytic solution

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0$.
- Can we find an analytic solution for the distribution of the estimate $\beta_{s\text{ML}}$ under $\mathcal{H}_0$?
- **Intuition:** The estimate is a linear transformation of a Normal distributed variable, namely $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right)$, where $\boldsymbol{\beta}$ is the value under $\mathcal{H}_0$ (with $\beta_s = 0$).
- $\beta_{\text{ML}} = \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}_{\text{transformation}} \boldsymbol{y}$

$$\beta_{\text{ML}} \sim \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{I}\boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \right)$$



Equivalent graphical model

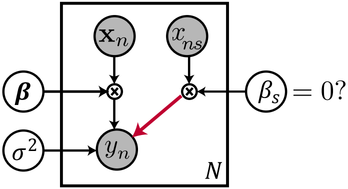
x_n: regression covariates

Testing in Linear Regression

Analytic solution

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0$.
- Can we find an analytic solution for the distribution of the estimate $\beta_{s\text{ML}}$ under $\mathcal{H}_0$?
- **Intuition:** The estimate is a linear transformation of a Normal distributed variable, namely $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right)$, where $\boldsymbol{\beta}$ is the value under $\mathcal{H}_0$ (with $\beta_s = 0$).
- $\beta_{\text{ML}} = \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}_{\text{transformation}} \boldsymbol{y}$



Equivalent graphical model

x_n: regression covariates

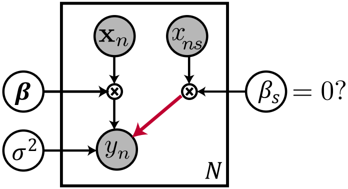
$$\beta_{\text{ML}} \sim \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{I} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

Testing in Linear Regression

Analytic solution

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2\right)$$

- $\mathcal{H}_0 : \beta_s = 0$.
- Can we find an analytic solution for the distribution of the estimate $\beta_{s\text{ML}}$ under $\mathcal{H}_0$?
- **Intuition:** The estimate is a linear transformation of a Normal distributed variable, namely $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}\right)$, where $\boldsymbol{\beta}$ is the value under $\mathcal{H}_0$ (with $\beta_s = 0$).
- $\beta_{\text{ML}} = \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}_{\text{transformation}} \boldsymbol{y}$



Equivalent graphical model

x_n: regression covariates

$$\beta_{\text{ML}} \sim \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{I} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

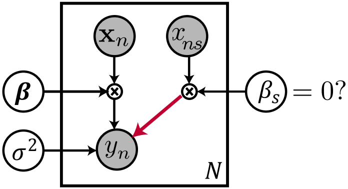
Testing in Linear Regression

Analytic solution

$$p(\boldsymbol{y} \,|\, \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0.$
- Can we find an analytic solution for the distribution of the estimate $\beta_{s\text{ML}}$ under $\mathcal{H}_0$?
- **Intuition:** The estimate is a linear transformation of a Normal distributed variable, namely $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right)$, where $\boldsymbol{\beta}$ is the value under $\mathcal{H}_0$ (with $\beta_s = 0$).
- $\beta_{\text{ML}} = \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}_{\text{transformation}} \boldsymbol{y}$

$$\beta_{\text{ML}} \sim \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{I}\boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$



Equivalent graphical model

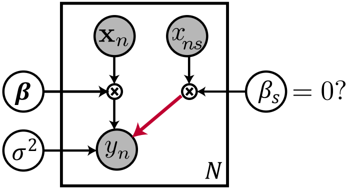
x_n: regression covariates

Testing in Linear Regression

Analytic solution

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0$.
- Can we find an analytic solution for the distribution of the estimate $\beta_{s\text{ML}}$ under $\mathcal{H}_0$?
- **Intuition:** The estimate is a linear transformation of a Normal distributed variable, namely $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right)$, where $\boldsymbol{\beta}$ is the value under $\mathcal{H}_0$ (with $\beta_s = 0$).
- $\boldsymbol{\beta}_{\text{ML}} = \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}_{\text{transformation}} \boldsymbol{y}$



Equivalent graphical model

x_n: regression covariates

$$\boldsymbol{\beta}_{\text{ML}} \sim \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{I} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \right)$$

Testing in Linear Regression

Analytic solution

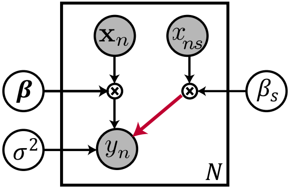
$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0.$

$$\boldsymbol{\beta}_{\mathsf{ML}} \sim \mathcal{N}\left(\boldsymbol{\beta}\,,\, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

- We are only interested in one entry (β_s)
- Use the marginal distribution of $\beta_{s\mathrm{ML}}$.

$$\beta_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\, \sigma^2 \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{s,s}\right),$$



Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

Analytic solution

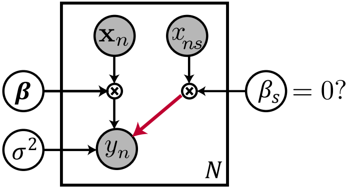
$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2\right)$$

- $\mathcal{H}_0 : \beta_s = 0.$

$$\boldsymbol{\beta}_{\text{ML}} \sim \mathcal{N}\left(\boldsymbol{\beta}\,,\, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

- We are only interested in one entry (β_s)
- Use the marginal distribution of $\beta_{s\text{ML}}$.

$$\beta_{s\text{ML}} \sim \mathcal{N}\left(0\,,\, \sigma^2 \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{s,s}\right),$$



Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression

Analytic solution

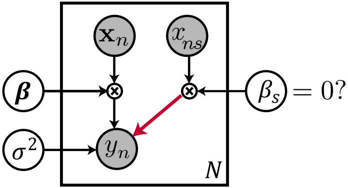
$$p(\boldsymbol{y} \,|\, \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- $\mathcal{H}_0 : \beta_s = 0.$

$$\boldsymbol{\beta}_{\mathsf{ML}} \sim \mathcal{N}\left(\boldsymbol{\beta}\,,\, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

- We are only interested in one entry (β_s)
- Use the marginal distribution of $\beta_{s\mathrm{ML}}$.

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\, \sigma^2 \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{s,s}\right),$$

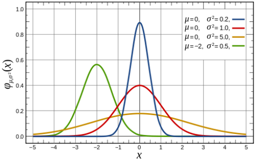


Equivalent graphical model

x_n: regression covariates

Cumulative distribution function

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right)$$



- Now we know the probability distribution of β_s.
- But the P value is the probability of observing something at least as extreme.

- Cumulative distribution function:

$$CDF(x) = P(X <= x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}\,z$$

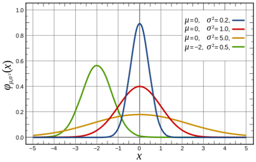
- For the univariate normal distribution with mean μ and variance σ^2:

$$\int_{-\infty}^{x} \mathcal{N}\left(y \mid \mu\,,\,\sigma^2\right)\,\mathrm{d}\,y = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{1}{2}\frac{x-\mu}{\sqrt{\sigma^2}}\right)\right)$$

- $\Rightarrow P = 2\min\left(CDF(\beta_{s\mathsf{ML}}), 1 - CDF(\beta_{s\mathsf{ML}})\right)$

Cumulative distribution function

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right)$$



- Now we know the probability distribution of β_s.
- But the P value is the probability of observing something at least as extreme.
- Cumulative distribution function:

$$CDF(x) = P(X <= x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}\,z$$

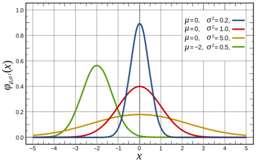
- For the univariate normal distribution with mean μ and variance σ^2:

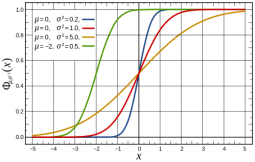
$$\int_{-\infty}^{x} \mathcal{N}\left(y \mid \mu\,,\,\sigma^2\right)\,\mathrm{d}\,y = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{1}{2}\frac{x-\mu}{\sqrt{\sigma^2}}\right)\right)$$

- $\Rightarrow P = 2\min\left(CDF(\beta_{s\mathsf{ML}}), 1 - CDF(\beta_{s\mathsf{ML}})\right)$

Cumulative distribution function

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right)$$





- Now we know the probability distribution of β_s.
- But the P value is the probability of observing something at least as extreme.

- Cumulative distribution function:

$$CDF(x) = P(X <= x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}\,z$$

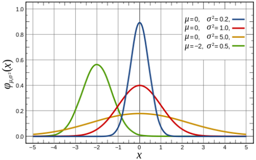
- For the univariate normal distribution with mean μ and variance σ^2:

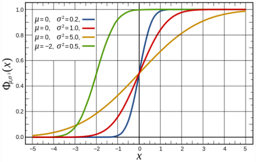
$$\int_{-\infty}^{x} \mathcal{N}\left(y \mid \mu\,,\,\sigma^2\right)\,\mathrm{d}\,y = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{1}{2}\frac{x-\mu}{\sqrt{\sigma^2}}\right)\right)$$

- $\Rightarrow P = 2\min\left(CDF(\beta_{s\mathsf{ML}}), 1 - CDF(\beta_{s\mathsf{ML}})\right)$

Cumulative distribution function

$$\boldsymbol{\beta}_{s\text{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right)$$





- Now we know the probability distribution of β_s.
- But the P value is the probability of observing something at least as extreme.

- Cumulative distribution function:

$$CDF(x) = P(X <= x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}\,z$$

- For the univariate normal distribution with mean μ and variance σ^2:

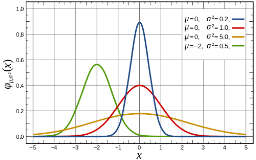
$$\int_{-\infty}^{x} \mathcal{N}\left(y \mid \mu\,,\,\sigma^2\right)\,\mathrm{d}\,y = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{1}{2}\frac{x-\mu}{\sqrt{\sigma^2}}\right)\right)$$

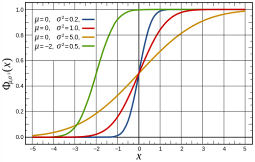
- $\Rightarrow P = 2\min\left(CDF(\beta_{s\text{ML}}), 1 - CDF(\beta_{s\text{ML}})\right)$

Cumulative distribution function

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right)$$

- Now we know the probability distribution of β_s.
- But the P value is the probability of observing something at least as extreme.





- Cumulative distribution function:

$$CDF(x) = P(X <= x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}\,z$$

- For the univariate normal distribution with mean μ and variance σ^2:

$$\int_{-\infty}^{x} \mathcal{N}\left(y \mid \mu\,,\,\sigma^2\right)\,\mathrm{d}\,y = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{1}{2}\frac{x-\mu}{\sqrt{\sigma^2}}\right)\right)$$

- $\Rightarrow P = 2\min\left(CDF(\beta_{s\mathsf{ML}}), 1 - CDF(\beta_{s\mathsf{ML}})\right)$

Caution!

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}\,(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\mathsf{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma}_2$ given the full D-by-1 vector β_{ML}!

$$\bar{\sigma}_2 = \frac{1}{N-D}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(\,0\,,\,1\,)$.

Caution!

$$\boldsymbol{\beta}_{s\mathrm{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\mathrm{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma}_2$ given the full D-by-1 vector β_{ML}!

$$\bar{\sigma}_2 = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\text{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\text{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\, \sigma^2 \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}\left(0\,,\,1\right), \qquad z = \frac{\boldsymbol{\beta}_{s\mathsf{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

Problem:

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}\left(0\,,\,1\right), \qquad z = \frac{\boldsymbol{\beta}_{s\mathsf{ML}}}{\sigma\sqrt{\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

Problem:

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\mathsf{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\mathsf{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

Problem:

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\text{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\text{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ss}}}$$

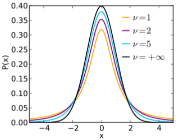
- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

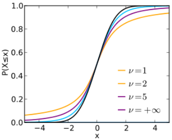
$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

Problem:

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$





- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.

Caution!

$$\boldsymbol{\beta}_{s\mathrm{ML}} \sim \mathcal{N}\left(0\,,\,\sigma^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{ss}\right) \Leftrightarrow z \sim \mathcal{N}(0\,,\,1)\,, \qquad z = \frac{\boldsymbol{\beta}_{s\mathrm{ML}}}{\sigma\sqrt{\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{ss}}}$$

- σ^2 is unknown, or a nuisance parameter.
- In practice we have to use an estimate $\bar{\sigma_2}$ given the full D-by-1 vector $\boldsymbol{\beta}_{\mathrm{ML}}$!

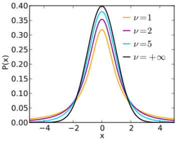
$$\bar{\sigma_2} = \frac{1}{N-D}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{ML}}\right)$$

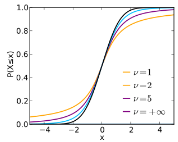
Problem:

- Sampling distribution of the test statistic should not depend on nuisance parameters.
- For large samples this is not an issue, as $\bar{\sigma}^2 \to \sigma^2$.
- For small samples use t-distribution with $\nu = N - D$ degrees of freedom!

$$t = \frac{z\sigma}{\bar{\sigma}} \sim \Gamma(\frac{\nu+1}{2})\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- For $\nu = +\infty$ t-distribution equals $\mathcal{N}(0\,,\,1)$.





Some relationships between distributions

- Normal distribution

$$x_n \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- z-score: Standard normal distribution

$$z_n = \frac{x_n - \mu}{\sigma} \sim \mathcal{N}\left(0\,,\,1\right)$$

- Sum of squares of N iid standard normals: χ^2 distribution with N dof

$$\sum_{n=1}^{N} z_n^2 \sim \chi_N^2$$

- Ratio of a standard normal and an independent χ_N^2 variable

$$t = \frac{z_1}{\sqrt{\frac{\sum_{n=2}^{N+1} z_n^2}{N}}} \sim \text{Student}-t(N)$$

- Ratio of a $\chi_{N_1}^2$ and an independent $\chi_{N_2}^2$: F-distribution with N_1 numerator dof and N_2 denominator dof

$$F = \frac{\sum_{n=1}^{N_1} z_n^2}{\sum_{n=N_1+1}^{N_1+N_2} z_n^2} \sim F(N_1, N_2)$$

Some relationships between distributions

- Normal distribution

$$x_n \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- z-score: Standard normal distribution

$$z_n = \frac{\boldsymbol{x}_n - \mu}{\sigma} \sim \mathcal{N}\left(0\,,\,1\right)$$

- Sum of squares of N iid standard normals: χ^2 distribution with N dof

$$\sum_{n=1}^{N} z_n^2 \sim \chi_N^2$$

- Ratio of a standard normal and an independent χ_N^2 variable

$$t = \frac{z_1}{\sqrt{\frac{\sum_{n=2}^{N+1} z_n^2}{N}}} \sim \mathrm{Student}{-}t(N)$$

- Ratio of a $\chi_{N_1}^2$ and an independent $\chi_{N_2}^2$: F-distribution with N_1 numerator dof and N_2 denominator dof

$$F = \frac{\sum_{n=1}^{N_1} z_n^2}{\sum_{n=N_1+1}^{N_1+N_2} z_n^2} \sim F(N_1, N_2)$$

Some relationships between distributions

- Normal distribution

$$x_n \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- z-score: Standard normal distribution

$$z_n = \frac{\boldsymbol{x}_n - \mu}{\sigma} \sim \mathcal{N}\left(0\,,\,1\right)$$

- Sum of squares of N iid standard normals: χ^2 distribution with N dof

$$\sum_{n=1}^{N} z_n^2 \sim \chi_N^2$$

- Ratio of a standard normal and an independent χ_N^2 variable

$$t = \frac{z_1}{\sqrt{\frac{\sum_{n=2}^{N+1} z_n^2}{N}}} \sim \mathrm{Student}{-}t(N)$$

- Ratio of a $\chi_{N_1}^2$ and an independent $\chi_{N_2}^2$: F-distribution with N_1 numerator dof and N_2 denominator dof

$$F = \frac{\sum_{n=1}^{N_1} z_n^2}{\sum_{n=N_1+1}^{N_1+N_2} z_n^2} \sim F(N_1, N_2)$$

Some relationships between distributions

- Normal distribution

$$x_n \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- z-score: Standard normal distribution

$$z_n = \frac{\boldsymbol{x}_n - \mu}{\sigma} \sim \mathcal{N}\left(0\,,\,1\right)$$

- Sum of squares of N iid standard normals: χ^2 distribution with N dof

$$\sum_{n=1}^{N} z_n^2 \sim \chi_N^2$$

- Ratio of a standard normal and an independent χ_N^2 variable

$$t = \frac{z_1}{\sqrt{\frac{\sum_{n=2}^{N+1} z_n^2}{N}}} \sim \text{Student}-t(N)$$

- Ratio of a $\chi_{N_1}^2$ and an independent $\chi_{N_2}^2$: F-distribution with N_1 numerator dof and N_2 denominator dof

$$F = \frac{\sum_{n=1}^{N_1} z_n^2}{\sum_{n=N_1+1}^{N_1+N_2} z_n^2} \sim F(N_1, N_2)$$

Some relationships between distributions

- Normal distribution

$$x_n \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- z-score: Standard normal distribution

$$z_n = \frac{\boldsymbol{x}_n - \mu}{\sigma} \sim \mathcal{N}\left(0\,,\,1\right)$$

- Sum of squares of N iid standard normals: χ^2 distribution with N dof

$$\sum_{n=1}^{N} z_n^2 \sim \chi_N^2$$

- Ratio of a standard normal and an independent χ_N^2 variable

$$t = \frac{z_1}{\sqrt{\frac{\sum_{n=2}^{N+1} z_n^2}{N}}} \sim \mathrm{Student}{-}t(N)$$

- Ratio of a $\chi_{N_1}^2$ and an independent $\chi_{N_2}^2$: F-distribution with N_1 numerator dof and N_2 denominator dof

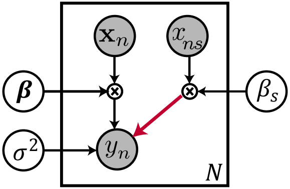
$$F = \frac{\sum_{n=1}^{N_1} z_n^2}{\sum_{n=N_1+1}^{N_1+N_2} z_n^2} \sim F(N_1, N_2)$$

Testing in Linear Regression

Likelihood Ratio Test

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0 : \beta_s = 0$ (rest don't matter)
- The ratio of the likelihood using the **ML estimator** and the **ML$_0$ estimator** restricted to $\mathcal{H}_0$ ($\beta_s = 0$) is another common test statistic.



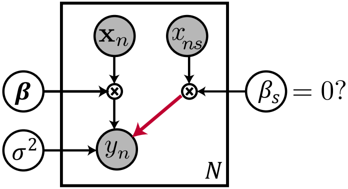
Equivalent graphical model

x_n: regression covariates

Testing in Linear Regression
Likelihood Ratio Test

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0 : \beta_s = 0$ (rest don't matter)
- The ratio of the likelihood using the **ML estimator** and the **ML$_0$ estimator** restricted to $\mathcal{H}_0$ ($\beta_s = 0$) is another common test statistic.



Equivalent graphical model

x_n: regression covariates

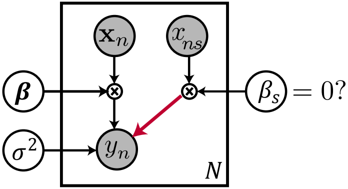
Testing in Linear Regression

Likelihood Ratio Test

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

- Test $\mathcal{H}_0 : \beta_s = 0$ (rest don't matter)
- The ratio of the likelihood using the **ML estimator** and the **ML$_0$ estimator** restricted to $\mathcal{H}_0$ $(\beta_s = 0)$ is another common test statistic.

$$\frac{\prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\mathsf{ML}}, \sigma^2_{\mathsf{ML}} \right)}{\prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\mathsf{ML}_0}, \sigma^2_{\mathsf{ML}_0} \right)}$$



Equivalent graphical model

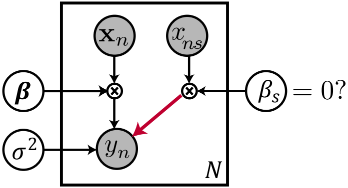
x_n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree-of-freedom χ_1^2. (for $N \to \infty$)
- P-value $= 1 - CDF_{\chi_1^2}(2\text{LR})$.



Equivalent graphical model

x_n: regression covariates

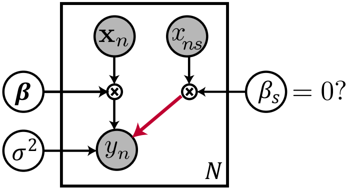
Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\text{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\text{ML}} +, \sigma^2_{\text{ML}} \right)$$
$$- \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\text{ML}_0}, \sigma^2_{\text{ML}_0} \right)$$

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree-of-freedom χ^2_1. (for $N \to \infty$)
- P-value $= 1 - CDF_{\chi^2_1}(2\text{LR})$.



Equivalent graphical model

x_n: regression covariates

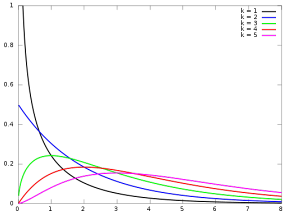
Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\mathsf{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\mathsf{ML}+}, \sigma^2_{\mathsf{ML}} \right) - \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\mathsf{ML}_0}, \sigma^2_{\mathsf{ML}_0} \right)$$

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree-of-freedom χ^2_1. (for $N \to \infty$)
- P-value $= 1 - CDF_{\chi^2_1}(2\mathsf{LR})$.



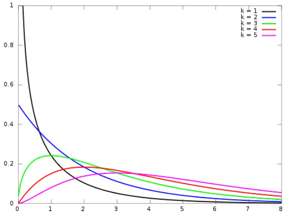
Testing in Linear Regression

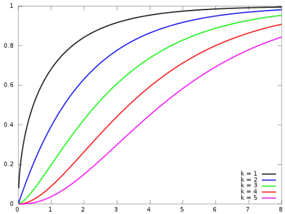
Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\text{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\text{ML}}+, \sigma^2_{\text{ML}}\right)$$
$$-\sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}_{\text{ML}_0}, \sigma^2_{\text{ML}_0}\right)$$

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree-of-freedom χ^2_1. (for $N \to \infty$)
- P-value $= 1 - CDF_{\chi^2_1}(2\text{LR})$.





Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01!$ $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^{S_0} \rightarrow 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01!$ $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^{S_0} \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01$! $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1-\alpha)^{S_0} \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01!$ $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1-\alpha)^{S_0} \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01!$ $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^{S_0} \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- If all tests are independent we would expect 10000 type-1 errors at $\alpha = 0.01!$ $(S = S_0)$
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^{S_0} \rightarrow 1$.
- Individual P-values < 0.01 are not significant anymore.

Need to correct for multiple hypothesis testing!

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i\in\mathcal{H}_0} P_{(i)} \leq \alpha\right) \underbrace{\leq}_{\text{Boole's inequality}} \sum_{i\in\mathcal{H}_0} Pr\left(P_{(i)} \leq \alpha\right)$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

$$\text{FWER} = Pr\left(\cup_{i \in \mathcal{H}_0} P_{(i)} \leq \alpha\right) \underbrace{\leq}_{\text{Boole's inequality}} \sum_{i \in \mathcal{H}_0} Pr\left(P_{(i)} \leq \alpha\right)$$

$$= \alpha \cdot S_0 \leq \alpha \cdot S$$

- Probability of at least one type-2 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \dfrac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

False Discovery Rate (FDR)

- FWER based correction (Bonferroni) leads to very conservative significance thresholds.
- Because of the abundance of tests we might be willing to accept a few false positives.
- Definition of the FDR:
  - $\mathbb{E}\left[\frac{FP}{FP+TP}\right]$
- Note: this can not be bounded when $\mathcal{H}_0$ always true ($FN + TP = 0$). In this case $\mathbb{E}\left[\frac{FP}{FP+TP}\right] = \mathbb{E}\left[\frac{FP}{FP}\right] = 1$

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

False Discovery Rate (FDR)

- FWER based correction (Bonferroni) leads to very conservative significance thresholds.
- Because of the abundance of tests we might be willing to accept a few false positives.
- Definition of the FDR:
  - $\mathbb{E}\left[\frac{FP}{FP+TP}\right]$
- Note: this can not be bounded when $\mathcal{H}_0$ always true $(FN + TP = 0)$. In this case $\mathbb{E}\left[\frac{FP}{FP+TP}\right] = \mathbb{E}\left[\frac{FP}{FP}\right] = 1$

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

False Discovery Rate (FDR)

- FWER based correction (Bonferroni) leads to very conservative significance thresholds.
- Because of the abundance of tests we might be willing to accept a few false positives.
- Definition of the FDR:
  - $\mathbb{E}\left[\frac{FP}{FP+TP}\right]$
- Note: this can not be bounded when $\mathcal{H}_0$ always true $(FN + TP = 0)$. In this case $\mathbb{E}\left[\frac{FP}{FP+TP}\right] = \mathbb{E}\left[\frac{FP}{FP}\right] = 1$

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S - S_0$

False Discovery Rate (FDR)

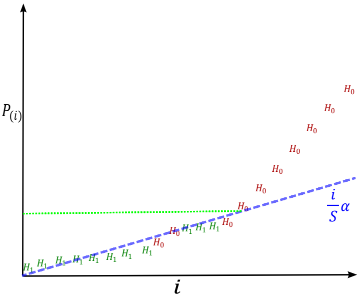
- FWER based correction (Bonferroni) leads to very conservative significance thresholds.
- Because of the abundance of tests we might be willing to accept a few false positives.
- Definition of the FDR:
  - $\mathbb{E}\left[\frac{FP}{FP+TP}\right]$
- Note: this can not be bounded when $\mathcal{H}_0$ always true ($FN+TP=0$). In this case $\mathbb{E}\left[\frac{FP}{FP+TP}\right]=\mathbb{E}\left[\frac{FP}{FP}\right]=1$

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives
	S_0	$S-S_0$

False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

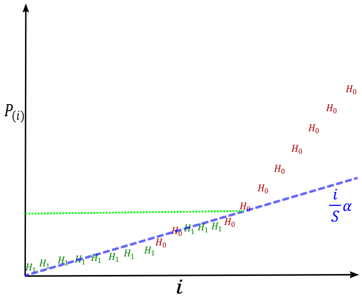
1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}} P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$



False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}} \, P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$

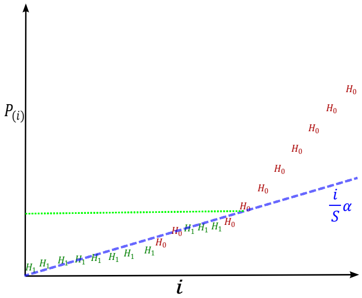


$$\mathrm{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}}$$

False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}} P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$

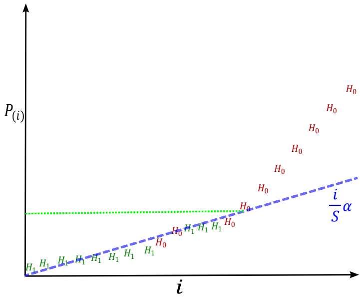


$$\mathrm{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}} = \frac{S_0 \cdot \overbrace{P_{(k)}}^{P_{\mathcal{H}_0} \sim \mathcal{U}[0,1]}}{i}$$

False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}}\, P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$

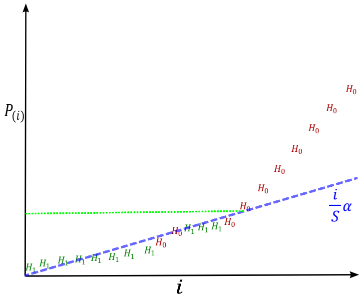


$$\mathrm{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}} = \frac{S_0 \cdot \overbrace{P_{(k)}}^{P_{\mathcal{H}_0} \sim \mathcal{U}[0,1]}}{i} \leq \frac{S \cdot P_{(k)}}{i}$$

False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \operatorname*{argmax}_i P_{(i)} \leq \dfrac{i}{S}\alpha$
3. Reject all P_s with $P_s < \dfrac{i}{S}\alpha$

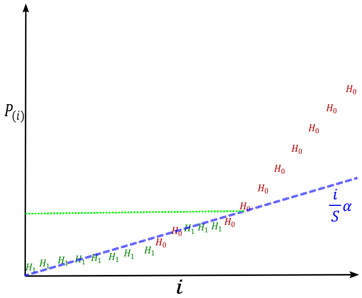


$$\mathrm{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}} = \frac{S_0 \cdot \overbrace{P_{(k)}}^{P_{\mathcal{H}_0} \sim \mathcal{U}[0,1]}}{i} \leq \frac{S \cdot P_{(k)}}{i}$$

False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}} P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$



$$\mathrm{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}} = \frac{S_0 \cdot \overbrace{P_{(k)}}^{P_{\mathcal{H}_0} \sim \mathcal{U}[0,1]}}{i} \leq \frac{S \cdot P_{(k)}}{i}$$

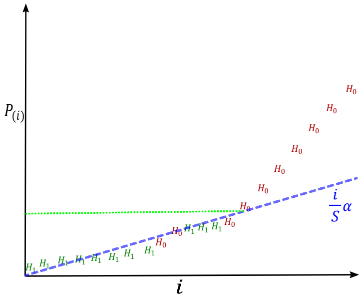
False discovery rates - Benjamini Hochberg procedure

Algorithm for FDR cutoff α:

1. Sort: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(S)}$
2. $k = \underset{i}{\operatorname{argmax}}\, P_{(i)} \leq \frac{i}{S}\alpha$
3. Reject all P_s with $P_s < \frac{i}{S}\alpha$

If tests are independent, then for this procedure:

$$FDR \leq \frac{\overbrace{FP + TN}^{S_0}}{S}\alpha \leq \alpha$$



$$\text{FDR}(\alpha = P_{(i)}) = \frac{\mathbb{E}[FP]}{\underbrace{FP + FN}_{i}} = \frac{S_0 \cdot \overbrace{P_{(k)}}^{P_{\mathcal{H}_0} \sim \mathcal{U}[0,1]}}{i} \leq \frac{S \cdot P_{(k)}}{i}$$

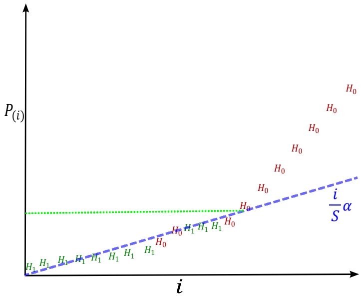
q-values

Definition of a q-value:

$$q(P_{(s)}) = \min_{t \geq P_{(s)}} \text{FDR}(t)$$

"*minimum FDR that can be attained while calling that feature significant*" (Storey and Tibshirani, 2003)

- Using the BH procedure it is possible to transform P values into q-values quite easily



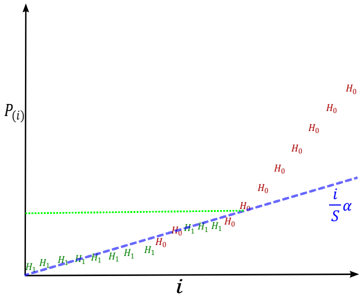
q-values

Definition of a q-value:

$$q(P_{(s)}) = \min_{t \geq P_{(s)}} \text{FDR}(t)$$

"*minimum FDR that can be attained while calling that feature significant*" (Storey and Tibshirani, 2003)

- Using the BH procedure it is possible to transform P values into q-values quite easily



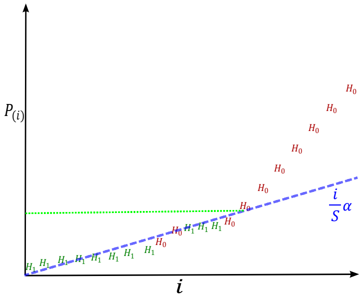
q-values

Definition of a q-value:

$$q(P_{(s)}) = \text{min}_{t \geq P_{(s)}} \text{FDR}(t)$$

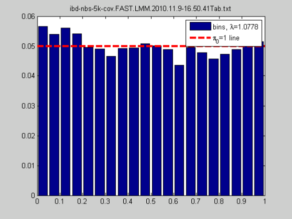
"*minimum FDR that can be attained while calling that feature significant*" (Storey and Tibshirani, 2003)

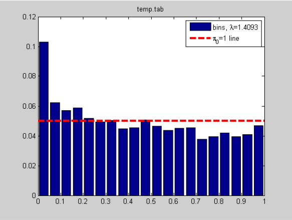
- Using the BH procedure it is possible to transform P values into q-values quite easily



Model Checking

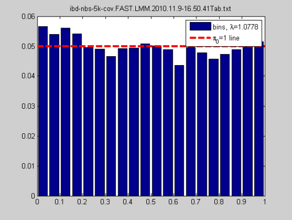
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

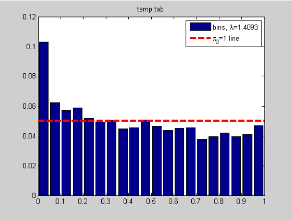




Model Checking

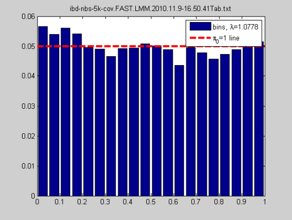
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

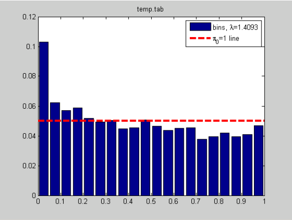




Model Checking

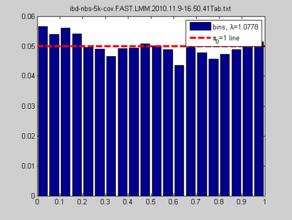
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

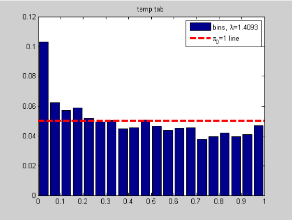




Model Checking

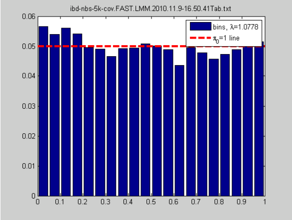
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

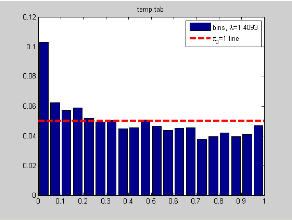




Model Checking

- Do my estimated P-values match the true null distribution?
    - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution





Model Checking
QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statistics against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

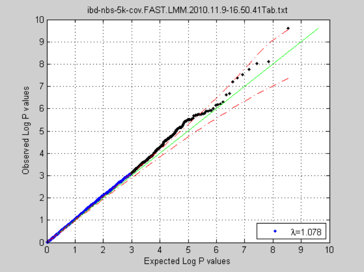
- Sort test statistics
- Plot test statistics against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statistics against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

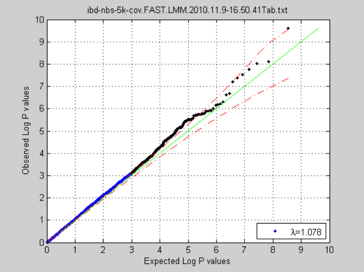


Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

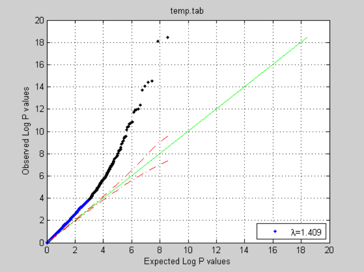
- Sort test statistics
- Plot test statistics against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.



Model Checking
QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

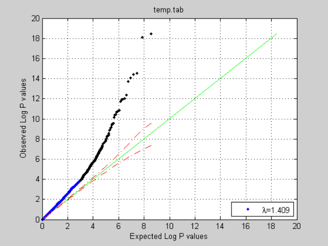
- Sort test statistics
- Plot test statistics against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.



Correction for inflation

Genomic control (λ_{GC})

- Ratio of the 50% quantiles between theoretical distribution and test-statistics known as the genomic inflation factor λ_{GC}.
- **Assumption:** λ_{GC} should be close to 1.
- Estimate degree of inflation (deflation) from this ratio.
- Adjust for degree of inflation by dividing all statistics by ratio of the median (50%-quantile).
- This procedure yields conservative estimates of the P-value distribution null-distribution.

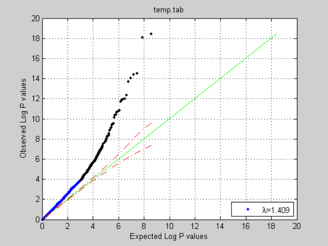


- GC does not make P-values uniform, but only matches one quantile!
- Assumption that 50% quantile of P-values is null-only does not need to hold in practice.
- Example: human height with thousands of causal SNPs

Correction for inflation

Genomic control (λ_{GC})

- Ratio of the 50% quantiles between theoretical distribution and test-statistics known as the genomic inflation factor λ_{GC}.
- **Assumption:** λ_{GC} should be close to 1.
- Estimate degree of inflation (deflation) from this ratio.
- Adjust for degree of inflation by dividing all statistics by ratio of the median (50%-quantile).
- This procedure yields conservative estimates of the P-value distribution null-distribution.

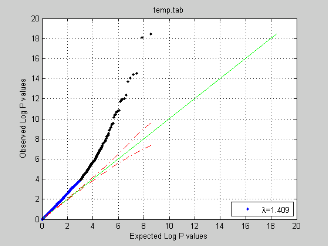


- GC does not make P-values uniform, but only matches one quantile!
- Assumption that 50% quantile of P-values is null-only does not need to hold in practice.
- Example: human height with thousands of causal SNPs

Correction for inflation

Genomic control (λ_{GC})

- Ratio of the 50% quantiles between theoretical distribution and test-statistics known as the genomic inflation factor λ_{GC}.
- **Assumption:** λ_{GC} should be close to 1.
- Estimate degree of inflation (deflation) from this ratio.
- Adjust for degree of inflation by dividing all statistics by ratio of the median (50%-quantile).
- This procedure yields conservative estimates of the P-value distribution null-distribution.

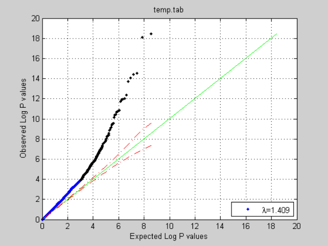


- GC does not make P-values uniform, but only matches one quantile!
- Assumption that 50% quantile of P-values is null-only does not need to hold in practice.
- Example: human height with thousands of causal SNPs

Correction for inflation

Genomic control (λ_{GC})

- Ratio of the 50% quantiles between theoretical distribution and test-statistics known as the genomic inflation factor λ_{GC}.
- **Assumption:** λ_{GC} should be close to 1.
- Estimate degree of inflation (deflation) from this ratio.
- Adjust for degree of inflation by dividing all statistics by ratio of the median (50%-quantile).
- This procedure yields conservative estimates of the P-value distribution null-distribution.

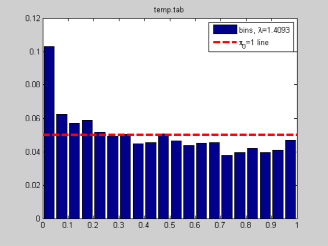


- GC does not make P-values uniform, but only matches one quantile!
- Assumption that 50% quantile of P-values is null-only does not need to hold in practice.
- Example: human height with thousands of causal SNPs

Correction for inflation

Genomic control (λ_{GC})

- Ratio of the 50% quantiles between theoretical distribution and test-statistics known as the genomic inflation factor λ_{GC}.
- **Assumption:** λ_{GC} should be close to 1.
- Estimate degree of inflation (deflation) from this ratio.
- Adjust for degree of inflation by dividing all statistics by ratio of the median (50%-quantile).
- This procedure yields conservative estimates of the P-value distribution null-distribution.



- GC does not make P-values uniform, but only matches one quantile!
- Assumption that 50% quantile of P-values is null-only does not need to hold in practice.
- Example: human height with thousands of causal SNPs