The Role of Quantiles in Statistical Learning

Nicola Orsini

Karolinska Institutet

48th Winter Conference in Statistics, Hemavan

March 12, 2024

Outline

- Quantiles to plan
- Quantiles to infer
- Quantiles to distinguish
- Quantiles to impute
- Quantiles to simulate

Quantiles to plan

In design of experimental/observational with simple/complex mechanisms underlying individual outcomes, we focus on

- θ a parameter value
- $d(\hat{\theta})$ a distribution of a parameter estimate

Quantiles to plan

A scientific investigation that aims to distinguish between two parameter values, saying θ_0 and θ_1, leads to two sampling distributions under the two hypothetical parameter values.

The sampling distribution $d(\hat{\theta})$ under the two scenarios, compactly denoted as $Q_p^{\theta_0}$ and $Q_p^{\theta_1}$, is described by all the p-quantiles with $p \in (0, 1)$

$$Q_p^{\theta_0} = Q_p(d(\hat{\theta}); \theta = \theta_0)$$
$$Q_p^{\theta_1} = Q_p(d(\hat{\theta}); \theta = \theta_1)$$

Sample size and power determinations

The specific p-quantiles to choose for $Q_p^{\theta_0}$ and $Q_p^{\theta_1}$ in determining the sample size are up to the investigator and context dependent.

A one-sided test in the upper direction with a type I error of 0.05 and type II error of 0.20 (or power 0.80) would lead to the following equation

$$Q_{0.95}^{\theta_0} = Q_{0.20}^{\theta_1}$$

High statistical power to distinguish between θ_0 and θ_1 is obtained by setting a top quantile under θ_0 to be a bottom quantile under θ_1.

Example: Planning a study on disease incidence

$\theta_0 = 0.2$

$\theta_1 = 0.3$

$Q_{0.95}^{\theta_0} = Q_{0.20}^{\theta_1}$

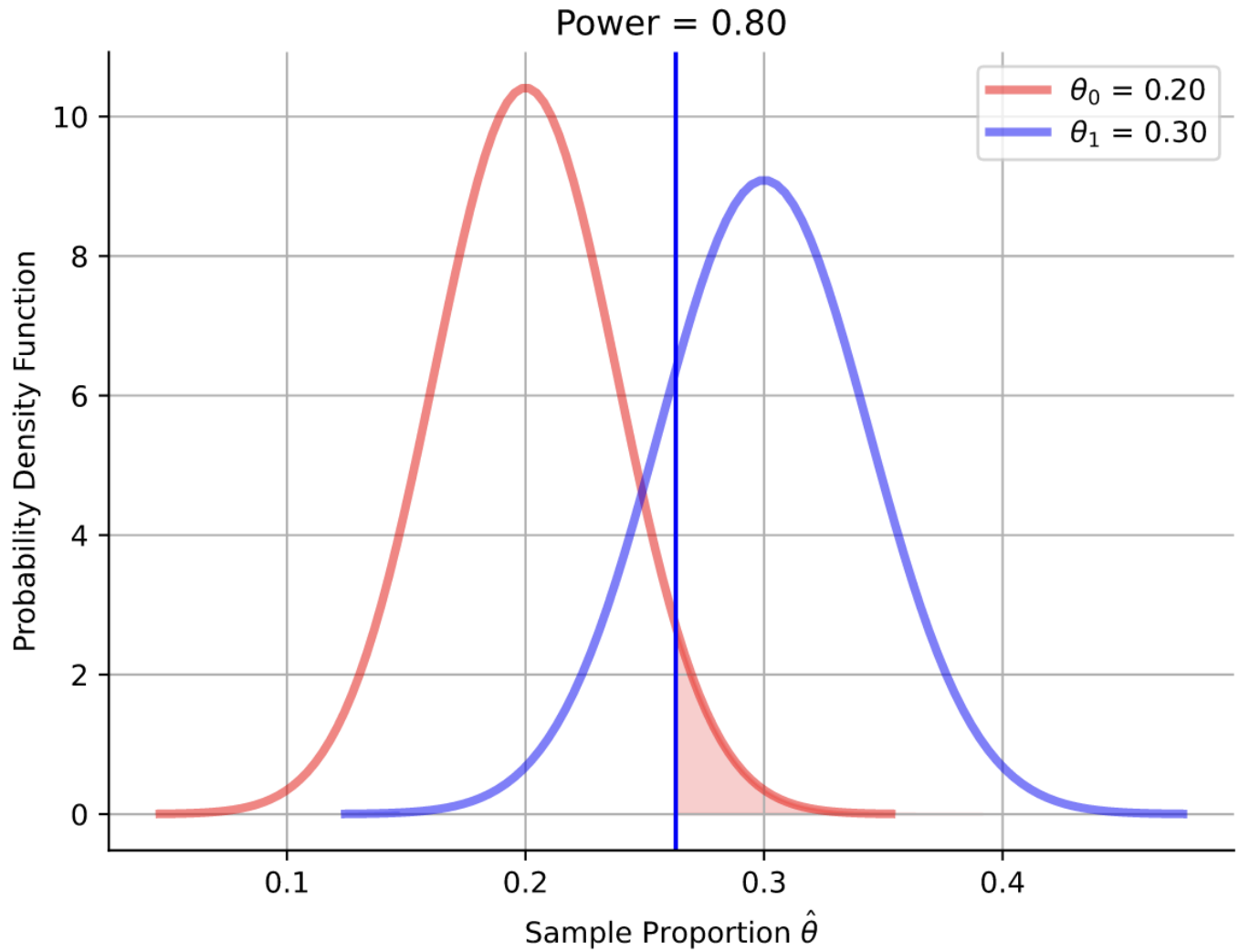
$$\theta_0 + \phi^{-1}(0.95)\sqrt{\theta_0(1-\theta_0)/n}) = \theta_1 + \phi^{-1}(0.20)\sqrt{\theta_1(1-\theta_1)/n})$$

where $\phi^{-1}(p)$ is the p-quantile of a standard normal distribution. Solving the equation for n

$$0.2 + 1.96\sqrt{0.2(1-0.2)/n}) = 0.3 - 0.84\sqrt{0.3(1-0.3)/n})$$

The sample size $n = 109$ makes $Q_{0.95}^{\theta_0} = Q_{0.20}^{\theta_1} = 0.263$.

Theoretical sampling distributions under alternative parameter values



Quantiles to infer - the logic of a test

The test of hypothesis is typically formulated in terms of an inequality statement regarding the unknown parameter θ. In one-sided test in the upper direction, the null and alternative hypothesis are defined as follows

$$H : \theta \leq \theta_0$$
$$\bar{H} : \theta > \theta_0$$

Finding what p quantile $\hat{\theta}$ actually is in the sampling distribution centered about the postulated divisive parameter under θ_0 provides a degree p of epistemic probability for the inequality statement presented in H.

Quantiles to infer - the logic of a test

Consider a study designed such that $Q_{0.95}^{\theta_0} = Q_{0.20}^{\theta_1}$.

The equation $\hat{\theta} = Q_p^{\theta_0}$ is solved for p.

$$\hat{\theta} = \theta_0 + \phi^{-1}(p)\sqrt{\theta_0(1-\theta_0)/n}$$

$$p = \phi\left(\frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$$

$$p_{value} = 1 - p$$

The complement of p defining the quantile of an empirical estimate under H is known as one-side p_{value}. It follows that if $\hat{\theta} > Q_{0.95}^{\theta_0}$ then $p > 0.95$ and $p_{value} < 0.05$.

Quantiles to infer - the coverage of an empirical quantile

Suppose we replicate K samples under the parameter value equal to θ, the fraction of samples in which $\theta \leq \hat{Q}_p^{\theta}$ should be about p. The coverage of the empirical p quantile can be appreciated by computing the following fraction

$$\sum_{i=1}^{K} \frac{I(\theta \leq \hat{Q}_p^{\hat{\theta}})}{K} \approx p$$

If the process underlying a large sample of individual observations is holding then in approximately $p\%$ of the K samples the parameter value θ will be less than or equal $\hat{Q}_p^{\hat{\theta}}$, for any p.

Quantiles to infer - the logic of confidence

The degree of confidence (denoted here as C) or rational belief an investigator can place in the claim that the unknown parameter θ is less than or equal to the empirical p-quantile is actually p.

$$C(\theta \leq \hat{Q}_p^{\hat{\theta}}) = p$$

The empirical p-quantile, $\hat{Q}_p^{\hat{\theta}}$, is obtained by shifting and rescaling the p-quantile of a standard normal distribution. So, expression of confidence is justified by the shape of the sampling distribution

$$C\left(\theta \leq \hat{\theta} + \phi^{-1}(p)\sqrt{\hat{\theta}(1-\hat{\theta})/n}\right) = p$$

Schweder, Tore, and Nils Lid Hjort. *Confidence, likelihood, probability*. Cambridge University Press, 2016.

Duality between test of hypothesis and degree of confidence

The logic of a test consists in locating an empirical estimate in a sampling distribution centered an hypothetical parameter value; that is $\hat{\theta} = Q_p^{\theta}$.

The logic of confidence consists in locating a parameter value in a sampling distribution centered about the sample estimate; that is $\theta = Q_p^{\hat{\theta}}$.

It is the familiarity with the shape (or all quantiles) of the sampling distribution produced by a plausible data generating mechanism underlying individual observations that justifies both ways of drawing statistical inference about the unknown parameter based on a sample of data.

Duality between test of hypothesis and degree of confidence

Denoting with p_t the quantile of a sample estimate $\hat{\theta}$ in the sampling distribution under a parameter value θ and with p_c the quantile of confidence about the unknown parameter θ with respect to $\hat{\theta}$ we have that

$$\hat{\theta} = Q^{\theta}_{p_t}$$

$$p_t = \phi\left(\frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)$$

$$\theta = Q^{\hat{\theta}}_{p_c}$$

$$p_c = \phi\left(\frac{\theta - \hat{\theta}}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}}\right)$$

The connection is that $p_t + p_c \approx 1$.

Example: Inference from one sample

In a sample of $n = 109$ individuals, 24 experienced the adversed health outcome. The proportion $24/109 = 0.22$ is a sample estimate $\hat{\theta}$ of the unknown θ. The hypothesis to be tested is the following

$$H : \theta \leq 0.2$$
$$\bar{H} : \theta > 0.2$$

In the logic of testing, we have that the sample estimate 0.22 is the 0.70 quantile of the sampling distribution under $\theta = 0.2$. Since $0.22 < 0.26$ (or $0.70 < 0.95$), the result of test is compatibility between this sample proportion $\hat{\theta} = 0.22$ and the hypothesis $H : \theta \leq 0.2$.

Example: Inference from one sample

$$p_t = \phi\left(\frac{0.22 - 0.20}{\sqrt{\frac{0.2(1-0.2)}{109}}}\right) = 0.70$$

The one-sided p_{value} corresponding the statistical test is $1 - 0.70 = 0.30$. The divisive parameter value defining the hypothesis H is 0.2. Since 0.2 is the 0.31 quantile of the sampling distribution under the sample estimate 0.22, the degree of confidence that can be placed in the statement $\theta \leq 0.2$ is 0.31.

$$p_c = \phi\left(\frac{0.2 - 0.22}{\sqrt{\frac{0.22(1-0.22)}{109}}}\right) = 0.31$$

The sum of the two proportions (p_t, p_c) defining the two quantiles $(0.70 + 0.31)$ is approximately 1.

Generalizability of the duality

The almost perfect positive correlation between the degree of confidence $C(\theta \leq \theta_0)$ and the one-sided p_{value} testing $H : \theta \leq \theta_0$ indicates how deeply connected are the two logics of statistical inference.
This connection is holding no matter

- what is the parameter value θ underlying the sample of data
- what is the divisive parameter θ_0 defining the hypothesis H
- what is the minimal distance between θ_0 and θ_1 in the design of the study
- what is the sample size n

Once again, the point of contact between the design of the study and the statistical inference based on one sample of data is the shape of the sampling distribution and not really its location and scale.
The fundamental act of learning, however, may require a continuous indication with respect to what is unknown spanning in either directions about to the actual estimate.

Example: From one quantile of confidence to all of them

Since 0.2 is the 0.31 quantile of the sampling distribution under the sample estimate 0.22, the degree of confidence is $C(\theta \leq 0.2) = 0.31$.
Shifting and rescaling the p-quantile of a standard normal distribution over the (0,1) interval, one is able to express any degree of confidence about the unknown parameter.

$$C\left(\theta \leq 0.22 + \phi^{-1}(p)\sqrt{0.22(1-0.22)/109}\right) = p$$
$$p \in (0,1)$$

The degree of confidence provided by the point estimate is one half.
Therefore, it is very common to derive extreme quantile of confidence in either direction from the point estimate.
For example, parameter values 0.14 and 0.30 are the 0.025 and 0.975 quantiles of sampling distribution under the sample estimate, respectively.

Likely presentation of the results

In a sample of 109 individuals, 24 of them experienced the disease. The incidence of the disease was 0.22 (95% CI = 0.14, 0.30) with no strong evidence against the hypothesis of a disease incidence less or equal than 20% ($p_{value} = 0.30$).

Quantiles to distinguish

The ability of a sample estimate $\hat{\theta}$ to point the investigator toward the correct choice between θ_0 and θ_1 can be evaluated by comparing two random quantiles of $Q_p^{\theta_0}$ and $Q_p^{\theta_1}$.

The intuition is that if the two sampling distributions are easily distinguishable with $\theta_1 > \theta_0$, then a randomly pick quantile from $\theta = \theta_1$ should be larger than a randomly pick quantile from $\theta = \theta_1$ most of the times.

Quantiles to distinguish

Given a pair of random quantiles u_0 and u_1 from a continuous uniform distribution $\mathcal{U}(0,1)$, an investigator would favour $\theta = \theta_1$ if $Q_{u_1}^{\theta_1} > Q_{u_0}^{\theta_0}$ and $\theta = \theta_0$ otherwise.

Replicating this process K times, the fraction of times in which $Q_{u_1}^{\theta_1} > Q_{u_0}^{\theta_0}$ provides a numerical summary of the ability to distinguish between two parameter values in light of a sample proportion.

$$\sum_{i=1}^{K} \frac{I(Q_{u_1}^{\theta_1} > Q_{u_0}^{\theta_0})}{K}$$

The above fraction, ranging from 0.5 (complete overlap) to 1 (complete separation), is strongly related the non-parametric test for the hypothesis of equal sampling distributions $Q_p^{\theta_0} = Q_p^{\theta_1}$.

Quantiles and AUC

It is widely known as the Area Under the Curve (AUC) of a Receiving Operating Characteristic (ROC) curve. The AUC can also be quantified by computing the integral of the function $f(u)$

$$AUC = \int_0^1 f(u)\, du$$

where the function $f(u)$ is the (1-Type II probability) statistical power associated with any u-quantile $Q_u^{\theta_0}$ that might be chosen to discriminate between θ_0 and θ_1

$$f(u) = 1 - \phi((Q_u^{\theta_0} - \theta_1)/\sqrt{\theta_1(1-\theta_1)/n})$$

with u being (1-type I probability) and ϕ is the cumulative distribution function of a standard normal.
In our example, an investigation planned with a type I error of 5% and type II error of 20% to distinguish two proportions $\theta_0 = 0.2$ and $\theta_1 = 0.3$ with a sample size of $n = 109$ has an AUC of about 96%.

Quantiles to impute

- Distributed data networks that include multiple data-contributing sites are increasingly used for data synthesis to improve evidence-based research and healthcare.
- Data on a key variable of interest can be completely missing in one or more sites.
- Data from different study sites cannot be pooled into a unified file. So conventional imputation approaches become unavailable.
- An imputation method can be based on conditional quantiles

Thiesmeier R, Bottai M, Orsini N. Systematically missing data in individual participant data meta-analysis: multiple imputation when data cannot be pooled. *Statistical Science*. Under Review. 2024.

Steps of Conditional Quantile Imputation

Conditional quantile imputation algorithm

Identify study site(s) with observed data on y_{ij};
Fit quantile regression models of y_{ij} conditional on z_{ij}:
$Q(p|z_{ij}) = z_{ij}\gamma_j(p) \quad p \in \{0.01, 0.02, \ldots, 0.99\}$;
Take the weighted average of $\gamma_j(p)$: $\bar{\gamma}_j(p)$
and distribute it to sites with systematically missing values on y_{ij};
for $m \leftarrow 1$ **to** M **do**
    Impute y_{ij} using $\bar{\gamma}_j(p)$ and z_{ij};
end
Combine estimates of the substantive model across imputations using Rubin's rules.

Step i: Fitting the imputation model

This step entails fitting the imputation model in the j-th study with observed data.

To accomplish the task, the quantile function Q, inverse of the cumulative distribution function, of y_{ij} conditionally on a set of predictors z_{ij} can be derived by estimating the p-quantile of y_{ij} with a quantile regression model where p ranges from 0.01 to 0.99

$$Q(p|z_{ij}) = z_{ij}\gamma_j(p) \quad p \in \{0.01, 0.02, \ldots, 0.99\}$$

Step ii: Collection and transmission

Step ii involves collecting all sets of regression coefficients from step i. At a central study site, the conditional quantiles for the systematically missing values in $j \in B$ are based on an inverse-variance weighted average regression coefficients, defined as $\bar{\gamma}_j(p)$, estimated in studies with no or partially observed missing data ($j \in A$).

To introduce random variability in the estimated regression coefficients of the imputation model of any quantile, a random draw from a normal distribution with mean equal the estimated regression coefficients, $\bar{\gamma}_j(p)$, and standard deviation equal to the corresponding estimated standard error, $\widehat{SE}(\bar{\gamma}_j(p))$. Finally, the vector of regression coefficients, $\bar{\gamma}_j(p)$ is then transmitted to the study sites with systematically missing values to be used for imputation.

Step iii: Imputing the systematically missing values

In MI each missing value y_{ij}, with $i \in M_j$ and $j \in B$, is replaced by M independent imputed values. We denote $y_{ij}^{(m)}$ as the m-th imputation of a missing value in y_{ij}. The following steps define the process underlying the assignment of a single value:

1. Draw a quantile U from a random continuous uniform distribution $\mathcal{U}(0, 1)$.
2. Extract the floor $L = \lfloor U\% \rfloor$ and modulus $W = U - \lfloor U\% \rfloor$.
3. Compute the weighted average of the L and $L + 1$ conditional predicted quantiles $C = (1 - W)\hat{Q}_L(p|z_{ij}) + W\hat{Q}_{L+1}(p|z_{ij})$.
4. Assign $y_{ij}^{(m)} = \arg\min_{k \in V} |C - V|[k]$ where V is a list of unique observed values of y_{ij}.

Quantiles to simulate

Observations from any distribution can be generated using the inverse transformation method, that is, the quantile function.

Monte-Carlo simulations (computer experiments) are useful to obtain sampling distributions of quantities of interest.

In our example, we examine the performance of the proposed imputation method.

To illustrate the use of CQI as an imputation approach when data cannot be pooled, a simplified scenario of an IPD meta-analysis of observational studies with a confounding variable is proposed.

An observational cohort study with a binary exposure and outcome variable, one continuous predictor of the outcome, and one continuous confounding variable.

Mechanism underlying individual data

The following random variables define a single study:

- V is a binary predictor of the outcome: $V \sim B(\pi_v)$, with a $\pi_v = 0.4$.
- C is a continuous confounding variable: $C \sim \chi^2_d$, where d are the degrees of freedom.
- E is a binary exposure variable: $E \sim B(\pi_e)$ where $\pi_e = \text{expit}(\alpha_0 + \alpha_1 \cdot C)$.
- D is the binary outcome variable: $D \sim B(\pi_d)$ where $\pi_d = \text{expit}(\beta_0 + \beta_1 \cdot E + \beta_2 \cdot V + \beta_3 \cdot C)$.

Of note, the variable C is determining both the probability of being exposed, π_e, and independently of the exposure E, the outcome probability, π_d.

Simulated scenarios

- IPD meta-analyses were generated under a common and heterogeneous confounding mechanism
- 20% of the studies having systematically missing values on the confounding variable, C. That is, for IPD meta-analysis with 5, 10, and 20 studies, the confounder C was set to 100% missing in 1/5, 2/10, and 4/20 studies, respectively.
- The sample size for all scenarios was set to 500 individuals in each study.

Negligible bias of the conditional quantile imputation

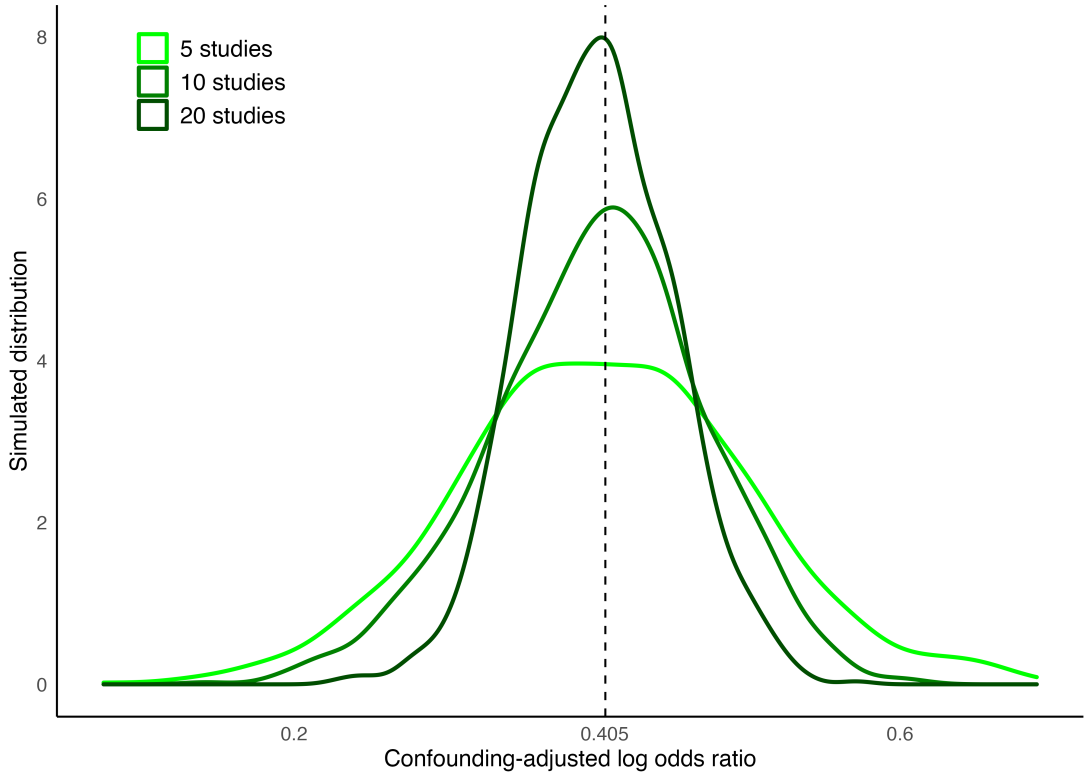


Figure: Sampling distribution of the confounding-adjusted log odds ratio under a common effect after using conditional quantile imputation for systematically missing values. The individual participant data meta-analysis included 5, 10, and 20 studies. The dotted line represents the set parameter of $\beta_1 = 0.405$.

Similar performance in common and heterogeneous confounding effect

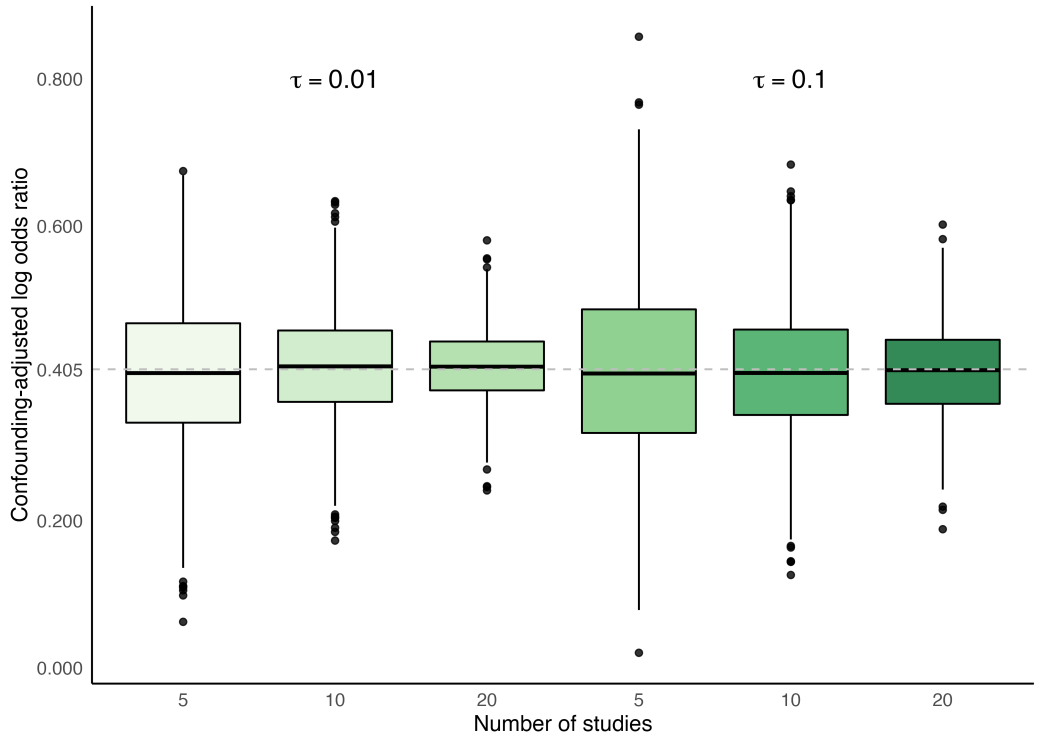


Figure: Distribution of the confounding-adjusted log odds ratio under a heterogeneous effect with $\tau = 0.01$ and $\tau = 0.1$ for 5, 10, and 20 studies included in the individual participant data meta-analysis. Distributions are shown after the use of conditional quantile imputation to impute systematically missing values. The dotted line represents the set parameter of $\beta_1 = 0.405$

Final remarks

- Quantiles are used to calibrate and plan (sample size, error probabilities) a scientific study
- Quantiles are crucial to express a degree of confidence about an inequality statement involving an unknown parameter
- Quantiles are used to measure the ability to discriminate between two continuous distributions (ROC-AUC)
- Quantiles are used to impute variables partially or completely missing in single or multiple studies
- Quantiles are used to generate individual data according to any plausible mechanism