Understanding Variance Estimation: The Bootstrap

Ashley I Naimi

Fall 2024

Contents

Learning Objectives

- Explain when and why it might be important to use the bootstrap for confidence interval and/or standard error estimation.
- Articulate the bootstrap principle, and why the procedure of resampling the empirical data can replicate features of the target population.
- Explain the difference between normal interval, percentile, and bias-corrected and accelerated bootstrap confidence intervals.

1 Introduction

Epidemiologists rely almost exclusively on statistical models to estimate exposure effects. After obtaining a point estimate, standard practice is to quantify its statistical uncertainty. This quantification is usually accomplished via standard errors and/or confidence interval estimates, which are meant to provide information on a plausible range of point estimate values that could be observed under repeated random sampling and/or exposure allocation (i.e., due only to random variation, and not systematic bias). There are several ways to estimate confidence intervals of an estimate. The most common technique is to use maximum likelihood theory, which provides an estimate of the standard error of the parameter of interest directly from the likelihood function (known as the observed Fisher information),[1] which can be used to obtain confidence intervals. In a linear modeling context, this "observed Fisher information" approach is equivalent to the variance equation we observed in the previous section:

$$V(\hat{\beta}) = (X^\intercal X)^{-1} X^\intercal \Sigma X (X^\intercal X)^{-1}$$

Ninety-five percent confidence intervals are computed by assuming the estimate follows a normal distribution and invoking the empirical rule, which states that 95% of the mass of the estimate's distribution falls within $1.96 \times SE(\hat{\beta})$ units of the point estimate.[2] This approach is referred to as the Wald equation or normal-interval confidence intervals.

This approach is based on asymptotic approximations that assume the sample size is sufficiently large.

There are generally three situations in which an alternative is preferred:

[1] Stephen R Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *Am J Epidemiol*, 179(2):252–260, 2013; Yudi. Pawitan. *In all likelihood : statistical modelling and inference using likelihood*. Clarendon Press ; Oxford University Press, Oxford; New York, 2001; and Terry M. Therneau and Patricia M. Grambsch. *Modeling survival data : extending the Cox model*. Springer, New York, 2000

[2] Dennis D. Wackerly, William. Mendenhall, and Richard L. Scheaffer. *Mathematical statistics with applications*. Duxbury, Pacific Grove, CA, 2002

1) when an analytic expression for the standard error or confidence interval of an estimator is unknown;

2) when such an expression exists, but is too complex to implement manually with standard software;

3) or when the sample size is too small to allow large-sample (asymptotic) approximations to be reliably invoked.

A well-known alternative to variance estimation in such situations is the bootstrap.[3]

Several bootstrap estimators exist,[4] however epidemiologists often rely on a few simple versions, namely the normal interval (Wald) bootstrap, and the percentile bootstrap.

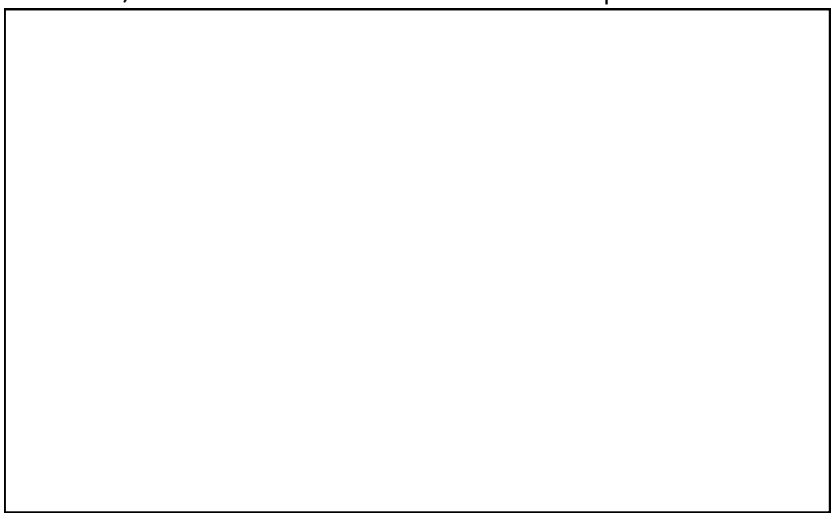
Here, we'll illustrate how to use bootstrap confidence interval estimators. We'll first provide a brief conceptual overview of the bootstrap, focusing on three particular bootstrap estimators. We then show how the bootstrap can be used to obtain measures of uncertainty when modeling the effect of aspirin on pregnancy outcomes in the Effects of Aspirin on Gestation and Reproduction (Aspirin Data) Trial, a large randomized trial conducted in women at high risk of pregnancy loss.

[3] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979

[4] Bradley Efron and Robert Tibshirani. *Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993

1.1 The Bootstrap

Let's revisit the superpopulation construct we introduced in the previous lecture. This time, we'll show the basic idea behind the bootstrap.



This concept is often depicted using the following Figure, first attributed to Efron and Tibshirani (**?**)

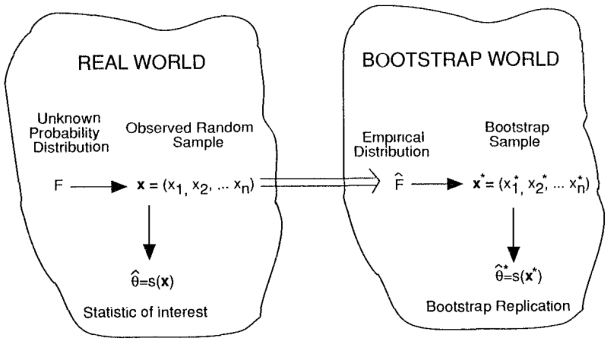


Figure 1: Illustration of the core principle behind the bootstrap.

The bootstrap is a technique that employs the Monte Carlo method and the substitution (or plug-in) principle to obtain standard error or confidence interval estimates (Efron, 2003). The Monte Carlo method is a general approach for solving a broad class of problems using computation to generate random numbers (Metropolis and Ulam, 1949).

This method is used for both **parametric** and **non-parametric** bootstrap estimators.

For the parametric bootstrap, the Monte Carlo method is used to generate residuals from a statistical model defining the relation between the exposure, confounders, and outcome of interest. The estimated parameters from this model are used, along with the exposure and confounder data, and the residuals generated from the Monte Carlo process, to generate a bootstrapped outcome. This process is repeated B times, giving B datasets with a bootstrapped outcome, and the original exposure and confounder data. One then fits a separate model to each of these B datasets to obtain a distribution of bootstrapped parameter estimates that can be use to quantify the uncertainty around the original estimates. In this setting, the bootstrapping process is "parametric' ' in that a parametric (e.g., logistic regression) model is used to generate the bootstrapped outcome data. Violations of the model's assumptions can lead to biased estimates of statistical uncertainty for the original parameter estimates (Carpenter and Bithell, 2000).

The non-parametric bootstrap uses the random numbers generated from the Monte Carlo process to select random samples with replacement from

the original data. The number of bootstrap samples chosen by the researcher is limited only by the available computing power. For each bootstrap sample, one can obtain a "bootstrap replicate' ' of the parameter of interest. With a large enough number of bootstrap replicates, one can use the distribution of bootstrap estimates to obtain information on the degree of uncertainty associated with the point estimate of interest. In effect, one can substitute (or plug-in) the empirical distribution of the estimates from each bootstrap resample for the unknown distribution of the point estimate. This empirical distribution can be used to estimate any feature (such as the standard error or percentiles) of the unknown distribution of the parameter estimate.

It is often stated that this version of the bootstrap is non-parametric in that the data are not assumed to follow a specified parametric model (Carpenter and Bithell, 2000). While true, this does not imply that the non-parametric bootstrap will work equally well when the model used to generate the parameter estimate of interest is itself nonparametric. Indeed, this is the result of the fact that both the parametric and nonparametric bootstrap estimators require that the underlying estimation model meets certain regularity conditions, which are not guaranteed to hold when nonparametric methods are used.

1.2 Example Data

To demonstrate implementation, we quantify the effect of aspirin on live birth in a cohort of women from the Aspirin Data Trial. Briefly, the Aspirin Data Trial (2007-2011; Clinical Trials Registration: #NCT00467363) investigated the impact of preconception-initiated low-dose aspirin (LDA) treatment on pregnancy loss and live birth in 1,228 women recruited from four U.S. university medical centers who were trying to become pregnant after experiencing one or two prior pregnancy losses. Details on the trial design and eligibility criteria are provided elsewhere (Schisterman et al., 2013, Schisterman et al. (2014)).

Women were randomized 1:1 to receive 81 mg aspirin per day or placebo (all received 400 mcg folic acid in addition). Follow-up occurred for up to 6 menstrual cycles while attempting to conceive, and if they became pregnant throughout pregnancy (the duration of total follow-up ranged from 1 to 60 weeks, median = 37 weeks). For women who conceived, treatment was continued until 36 weeks gestation. Information on demographic and lifestyle factors was assessed via questionnaire at baseline. Height, weight, and blood pressure

were measured at baseline according to standardized protocols. The primary outcome of interest was live birth, defined as a live born infant as indicated on the medical record.

To estimate the effect of aspirin on live birth, we fit an adjusted logistic regression model, regressing an indicator of live birth against the treatment indicator (low-dose aspirin versus placebo), as well as the following covariates: body mass index (BMI, in kg/m^2), mean arterial pressure (measured in mmHg), eligibility stratum, age (years), and the number of times a women tried to become pregnant prior to entry in the study.

1.3 Example Analysis

For clarity, we refer to these data collectively as $\mathcal{X}$. This set consists of the indicator of whether a live birth occurred (denoted Y), the treatment indicator (denoted T), and the vector of covariates (collectively denoted $\mathbf{C}$). Our regression model can thus be defined as:

$$\operatorname{logit} P(Y = 1 \mid T, \mathbf{C}) = \beta_0 + \theta T + \boldsymbol{\beta}_C \mathbf{C} \tag{1}$$

Because this model was fit in the original sample of all women recruited in Aspirin Data, (under correct parametric modeling assumptions) we may interpret $\hat{\theta}_1$ as the intent to treat (ITT) effect of aspirin on live birth.

Confidence intervals (CIs) for θ may be obtained using the inverse of the Fisher information, which would entail the standard practice of computing the values that result from ± 1.96 times the model-based standard error. However, here we proceed with obtaining CIs via the bootstrap.

Drawing a simple random sample (with replacement) of $N = 1,228$ individuals from $\mathcal{X}$, we obtain a bootstrap sample of our observed data that we denote $\mathcal{X}^*$. This bootstrap re-sample contains the same number of individuals as in the original sample ($N = 1,228$), except that some individuals are selected more than once, while others are not selected. We then re-fit the above logistic model to $\mathcal{X}^*$ to obtain a bootstrap replicate $\hat{\theta}^*$. Repeating the sampling procedure and re-fitting the model B times, we obtain a sample of bootstrap replicates $\left\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\right\}$. This sample of replicates allows us to estimate any feature of the distribution of our point estimate, such as its standard error or percentiles of its distribution.

While many different types of nonparametric bootstrap confidence interval estimators exist, the above features comprise the central characteristics of the approach. Different estimators are distinguished by what information is extracted (and how) from the sample of bootstrap replicates $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$. Three bootstrap confidence interval estimators are arguably best suited to epidemiologic research due to their relative ease of implementation, established theoretical properties, and robustness to violations of a varying range of assumptions (Greenland, 2004). These are the Wald (or normal-interval estimator), percentile, and Bias-Corrected and Accelerated (BC_a) bootstrap estimators.

1.4 Bootstrap Confidence Interval Estimators

1.5 Normal Interval Bootstrap

The Wald, or normal interval, bootstrap estimator is one of the simplest of bootstrap confidence interval estimators. The approach requires the assumption that the parameter estimator $\hat{\theta}$ is normally distributed.

After obtaining a sample of bootstrap estimates as outlined above, one can implement the method by simply estimating the standard deviation of this sample to obtain an estimate of the standard error of the point estimate. By assuming that $\hat{\theta}$ follows a normal distribution, one can then add and subtract $1.96 \times SE(\hat{\theta})$ from the point estimate to obtain upper and lower limits, respectively. For the example illustrated with Model 1, 95% Wald-type bootstrap confidence intervals for the estimated odds ratio $\exp(\hat{\theta})$ can be obtained as $\exp\left\{\hat{\theta} \pm 1.96 \times SD(\hat{\theta}^*)\right\}$, where $SD(\hat{\theta}^*)$ is the standard deviation of the bootstrap replicates. The standard deviation of the distribution of bootstrap replicates $SD(\hat{\theta}^*)$ is equivalent to the standard error of the point estimate $SE(\hat{\theta})$ (Altman and Bland, 2005).

Note that this normal-interval bootstrap is NOT transformation respecting. This means that if we are trying to construct confidence intervals for a (e.g.) risk ratio, we cannot transform the log risk ratio to the risk ratio scale, then take the standard deviation, then construct CIs. Instead, we must construct CIs on the log-scale, and then transform all the results (UCL, LCL, estimate). The normal-interval bootstrap is also NOT range respecting. This means that it is possible to estimate upper or lower bounds that are beyond the possible range

of the estimator (e.g., negative risk ratios).

The assumption that $\hat{\theta}$ follows a normal distribution is only valid in reasonably large samples and under other conditions. As a consequence, Wald-type bootstrap confidence intervals can have a less than nominal coverage probability, particularly in small samples (Efron and Tibshirani, 1993, DiCiccio and Efron (1996)). Finally, Efron suggests that between 50 and 200 replicates is sufficient to obtain a good estimate of $SE(\hat{\theta})$ (Efron and Tibshirani, 1993). However, with increases in computing power over the last 30 years, a minimum of 500 replicates is usually expected in the scientific literature.

1.6 Percentile Bootstrap

The percentile bootstrap estimator is as simple to implement as the Wald estimator, but **does not require the assumption that $\hat{\theta}$ follows a normal distribution.** To obtain two-sided percentile bootstrap confidence intervals, one simply selects the bootstrap replicate (i.e., the estimate based on bootstrap resample) corresponding to the $100 \times \alpha/2$ and $100 \times (1 - \alpha/2)$ percentile of the distribution of bootstrap replicates. For example, with 2,000 bootstrap replicates $\{\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_{2,000}\}$ and a nominal coverage of 95%, the 2.5th and 97.5th percentile points representing the lower and upper confidence limits would correspond to $\hat{\theta}^*_{50}$ and $\hat{\theta}^*_{1,950}$, respectively.

The percentile method is both *transformation* and *range respecting*: for example, percentile confidence interval estimates can be obtained on either the log-scale and transformed to the exponential scale or vice versa (Efron and Tibshirani, 1993, (p175)). Moreover, they respect the boundedness of the estimator in that they will not provide confidence interval estimates that fall outside of the allowable range of the parameter estimate. Percentile intervals (when obtained using the non-parametric bootstrap) are completely non-parametric. As a consequence, this method is subject to bias (anti-conservative) in that their coverage probability is usually less than nominal (DiCiccio and Efron, 1996, Greenland (2004),Efron and Tibshirani (1993),Carpenter and Bithell (2000)). Additionally, because percentile-based methods require estimates of the tails of the distribution of bootstrap replicates, more bootstrap resamples are required. Although the specific number may depend on the scenario, 1,000 to 2,000 re-samples is often seen in practice, and more is usually better.

1.7 BCa Intervals

Early recognition of the poor coverage probabilities of the Wald-type and percentile bootstrap confidence intervals led to two modifications of the percentile method (DiCiccio and Efron, 1996). The resulting "bias-corrected and accelerated" confidence intervals are meant to improve the performance of this percentile confidence interval estimator. The BC_a confidence interval estimator is a percentile-based estimator in that the confidence interval end-points selected are percentiles of distribution of bootstrap replicates. This interval estimator is also transformation and range respecting. It differs from the standard percentile method in that the percentiles corresponding to the upper and lower interval estimates are chosen as a function of a bias correction factor and an acceleration factor that are determined by the data.

The bias-correction factor is meant to account for the discrepancy between the median of the sample of bootstrap replicates and the point estimate. This factor can be calculated as a function of the number of bootstrap replicates less than the original point estimate:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_i \hat{\theta}_i^* < \hat{\theta}\right)$$

Note that the $\Phi^{-1}(\bullet)$ function is the inverse of the normal CDF, and returns a z-statistic for a given quantile from the normal distribution (e.g., `qnorm()` in R). Note that if exactly 50% of the bootstrap estimates $\hat{\theta}_i^*$ are less than the estimate obtained in the original data ($\hat{\theta}$), then the bootstrap estimator is said to be median unbiased, and the $\hat{z}_0$ value will be zero.

The acceleration factor is meant to compensate for possible heterogeneity in the standard error of the estimator as a function of the true parameter value, and can be calculated using the jackknife procedure (Tukey, 1958):

$$\hat{a} = \frac{\sum_i (\bar{\theta}^{-i} - \hat{\theta}^{-i})^3}{6[\sum_i (\bar{\theta}^{-i} - \hat{\theta}^{-i})^2]^{3/2}}$$

where $\bar{\theta}^{-i}$ denotes average of all estimates with the i^{th} individual removed, and where $\hat{\theta}^{-i}$ is the point estimate in the original data with the i^{th} individual removed.

The bias correction and acceleration factors are then used to select altered percentiles of the distribution of bootstrap replicates. The BCa bounds are

defined as $[\hat{\theta}^*_{(\alpha_1)}, \hat{\theta}^*_{(\alpha_2)}]$, where:

$$\alpha_1 = \left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(\alpha/2)}}{1 - \hat{a}[\hat{z}_0 + z_{(\alpha/2)}]} \right)$$

and

$$\alpha_2 = \left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\alpha/2)}}{1 - \hat{a}[\hat{z}_0 + z_{(1-\alpha/2)}]} \right)$$

1.8 Bootstrap Confidence Intervals in the Aspirin Data

We compared the normal interval, percentile, and bias corrected and accelerated bootstrap CI estimators in Aspirin Data to obtain uncertainty estimates for the effect of aspirin defined in Model 1. In the R programming language, these CI estimators can be easily obtained using the `boot` package.

First, we load the aspirin data:

```
library(AIPW)

data("eager_sim_obs")

set.seed(123)
aspirin <- eager_sim_obs %>%
    slice_sample(n = 1228, replace = T) %>%
    select(-loss_num)

dim(aspirin)
```

```
## [1] 1228    7
```

```
mod_mu <- glm(sim_Y ~ sim_A + eligibility + age + time_try_pregnant +
    BMI + meanAP, data = aspirin, family = binomial("logit"))

summary(mod_mu)$coefficients
```

```
##                      Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)       5.384376116 0.718076562  7.4983315 6.463526e-14
```

```
## sim_A              0.647977705 0.134806560  4.8067224 1.534248e-06
## eligibility       -0.418868878 0.130708108 -3.2046128 1.352443e-03
## age                0.002757795 0.016992055  0.1622991 8.710703e-01
## time_try_pregnant -0.158072511 0.019980827 -7.9112095 2.549002e-15
## BMI               -0.053479513 0.012721290 -4.2039380 2.623108e-05
## meanAP            -0.047307086 0.007575535 -6.2447187 4.245634e-10
```

```r
# log-odds ratio
summary(mod_mu)$coefficients[2, 1]
```

```
## [1] 0.6479777
```

Now we obtain bootstrap confidence intervals for this log-odds ratio of interest. We start with the normal-interval bootstrap:

```r
boot_func <- function(data, index) {

    boot_dat <- data[index, ]

    mod_mu <- glm(sim_Y ~ sim_A + eligibility + age + time_try_pregnant +
        BMI + meanAP, data = boot_dat, family = binomial("logit"))

    res <- summary(mod_mu)$coefficients[2, 1]

    return(res)

}

library(boot)

boot_res <- boot(boot_func, data = aspirin, R = 500)

res_ci_norm <- boot.ci(boot_res, type = "norm")

res_ci_norm
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res, type = "norm")
##
## Intervals :
## Level      Normal
## 95%   ( 0.3923,  0.9248 )
## Calculations and Intervals on Original Scale
```

Next, we obtain percentile intervals. Note the larger R value:

```
boot_res <- boot(boot_func, data = aspirin, R = 1000)

res_ci_perc <- boot.ci(boot_res, type = "perc")

res_ci_perc
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 0.3856,  0.9044 )
## Calculations and Intervals on Original Scale
```

Next, we obtain bias-corrected and accelerated intervals. Note the same R value:

```
boot_res <- boot(boot_func, data = aspirin, R = 1000)

res_ci_bca_err <- boot.ci(boot_res, type = "bca")
```

```
## Error in bca.ci(boot.out, conf, index[1L], L = L, t = t.o, t0 = t0.o, : estimated adjustment 'a' is N
```

```
res_ci_bca_err
```

```
## Error in eval(expr, envir, enclos): object 'res_ci_bca_err' not found
```

This implementation of BCa intervals results in an error. This is because the bias corrected and accelerated intervals use the jacknife to estimate the acceleration factor. However, if the number of resamples is less than the sample size, the algorithm for the jacknife in the `boot` package will not work. To remedy this, ensure that $R \geq N$:

```
boot_res <- boot(boot_func, data = aspirin, R = 2000)

res_ci_bca <- boot.ci(boot_res, type = "bca")

res_ci_bca
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.3660,  0.8994 )
## Calculations and Intervals on Original Scale
```

The choice of a given bootstrap estimator will often depend on context. For a computationally intensive estimator, the only practical choice may be the normal interval estimator, due to the need for fewer resamples than the percentile or BCa methods. However, when computationally feasible, these latter estimators should be considered. Between the two, the percentile method is simpler to implement in that in only requires selecting percentile values of the bootstrap distribution of estimates. However, Greenland emphasizes the point that the na"ive percentile method is conceptually backwards: uncorrected bootstrap estimates incorporate a small sample bias and skewness of the original estimators and add bias and skewness in the errors (Efron and Tibshirani, 1993, Greenland (2004) Carpenter and Bithell (2000)). The BCa bootstrap estimator was constructed specifically to overcome these problems.

References

Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *BMJ*, 331(7521):903, 2005.

James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Stat Med*, 19(9): 1141—1164, 2000.

Stephen R Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *Am J Epidemiol*, 179(2): 252–260, 2013.

T.J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Stat Sci*, 11(3): 189—212, 1996.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Bradley Efron. Second thoughts on the bootstrap. *Stat Sci*, 18(2):135—140, 2003.

Bradley Efron and Robert Tibshirani. *Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993.

Sander Greenland. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol*, 33(6):1389–1397, 2004.

N Metropolis and S Ulam. The Monte Carlo method. *J Am Stat Assoc*, 44(247): 335–341, 1949.

Yudi. Pawitan. *In all likelihood : statistical modelling and inference using likelihood*. Clarendon Press ; Oxford University Press, Oxford; New York, 2001.

Enrique F Schisterman, Robert M Silver, Neil J Perkins, Sunni L Mumford, Brian W Whitcomb, Joseph B Stanford, Laurie L Lesher, David Faraggi, Jean Wactawski-Wende, Richard W Browne, Janet M Townsend, Mark White, Anne M Lynch, and Noya Galai. A randomised trial to evaluate the effects of low-dose aspirin in gestation and reproduction: design and baseline characteristics. *Paediatr Perinat Epidemiol*, 27(6):598–609, Nov 2013. DOI: 10.1111/ppe.12088.

Enrique F Schisterman, Robert M Silver, Laurie L Lesher, David Faraggi, Jean Wactawski-Wende, Janet M Townsend, Anne M Lynch, Neil J Perkins, Sunni L Mumford, and Noya Galai. Preconception low-dose aspirin and pregnancy outcomes: results from the eager randomised trial. *Lancet*, 384(9937): 29–36, Jul 2014. DOI: 10.1016/S0140-6736(14)60157-4.

Terry M. Therneau and Patricia M. Grambsch. *Modeling survival data : extending the Cox model*. Springer, New York, 2000.

J.W. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29(2):614–, 1958.

Dennis D. Wackerly, William. Mendenhall, and Richard L. Scheaffer. *Mathematical statistics with applications*. Duxbury, Pacific Grove, CA, 2002.