Frequentists and Bayesian methods to incorporate recruitment rate stochasticity at the design stage of a clinical trial

---

Master Thesis in Biostatistics (STA495)

by

Pilar Pastor
23-733-975

supervised by

PD Dr. Malgorzata Roos

Zurich, month year

Frequentists and Bayesian methods to incorporate recruitment rate stochasticity at the design stage of a clinical trial

Pilar Pastor

Version March 4, 2025

Contents

Preface

Howdy!

Pilar Pastor
June 2025

Chapter 1

Introduction

Why, what and how...

Chapter 2

Methodology

2.1 Definitions

The **Target Population** is a specific group within the broader population, defined by attributes relevant to the research question. This group is focused on criteria that match the study's goals (Willie, 2024). Defining the target population allows researchers to refine their objectives and recruitment methods to align with the study's aims.

The **Eligibility** criteria are the specific requirements that individuals must meet to participate in a study. Eligible patients will be selected from the target population. Inclusion criteria specify the conditions that allow individuals to participate in the trial, particularly focusing on the medical condition of interest. Any other factors that limit eligibility are classified as exclusion criteria (Van Spall *et al.*, 2007), conditions or circumstances that disqualify potential participants (Food *et al.*, 2018).

In clinical trials, **Enrollment** refers to the formal process of registering participants into a study after they have met all eligibility criteria and provided informed consent. This process includes verifying that each participant satisfies the inclusion and exclusion criteria outlined in the study protocol (National Institute of Allergy and Infectious Diseases, 2021). It is important to distinguish between recruitment and enrollment. Recruitment involves identifying and inviting potential participants to join the study, whereas enrollment occurs after these individuals have been screened, consented, and officially registered into the trial (Frank, 2004).

Once enrolled, participants are assigned to specific treatment groups or interventions as defined by the study design. The most common practice is **Randomization**. In clinical research, randomization is the process of assigning participants to different treatment groups using chance methods, such as random number generators or coin flips (Lim and In, 2019). Randomized controlled trials (RCTs) are considered the most effective method for preventing bias in the evaluation of new interventions, drugs, or devices. (Van Spall *et al.*, 2007).

In clinical research, **Statistical Analysis** involves applying statistical methods to collect, summarize, interpret, and present data derived from clinical studies. This process is essential for evaluating the safety, efficacy, and overall outcomes of medical interventions, ensuring that conclusions drawn are both reliable and valid (Panos and Boeckler, 2023). Not all participants who are randomized may be included in the final statistical analysis due to protocol deviations of patients not adhering to the protocol (Rehman *et al.*, 2020), missing data (Shih, 2002) or loss-to-follow-up, some participants may become unreachable or withdraw consent during the study, resulting in missing outcome data (Nüesch *et al.*, 2009).

The number of patients decreases at each stage of a clinical study, from defining the target population to final statistical analysis, see Figure 2.2. This process is known as patient leakage (Desai, 2014), alternative terms are attrition or retention. Eligibility criteria narrow down participants, and enrollment further reduces numbers as only those meeting strict criteria are registered. Randomization assigns individuals to treatment groups, but some may later be

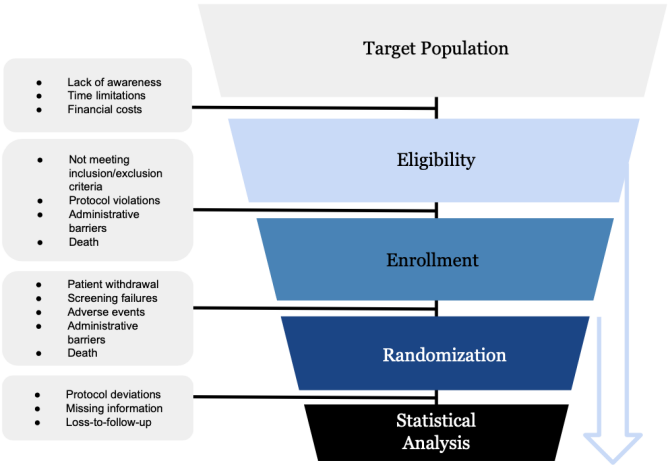

Figure 2.1: Patient leakage at each stage of a clinical study (Piantadosi and Meinert, 2022; Whelan *et al.*, 2018; Bogin, 2022).

excluded due to protocol deviations, missing data, or loss to follow-up.

The general notion of **Recruitment** in this Master Thesis refers to the number of patients (Counts) at the Eligibility, or Enrollment, or Randomization, or Statistical Analysis stage in Figure 2.1. We define **Accrual** as cumulative recruitment.

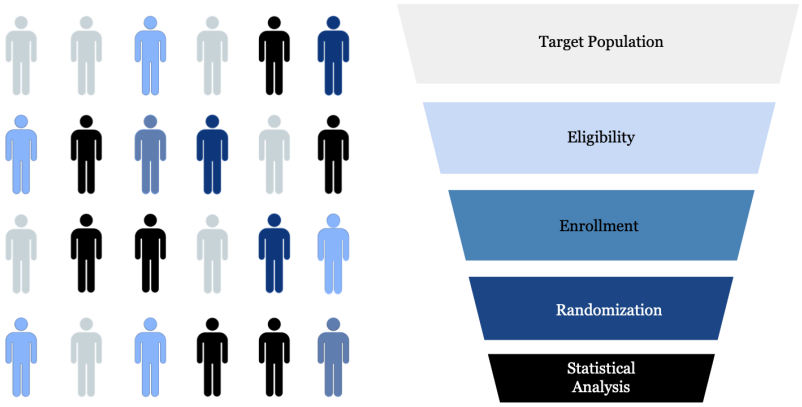

Figure 2.2: Visual representation of patient leakage at each stage of a clinical study (Piantadosi and Meinert, 2022; Whelan *et al.*, 2018; Bogin, 2022).

2.2 Uncertainty and models for counts

There are two types of uncertainty, aleatory and epistemic (O'Hagan, 2006). The **Aleatory Uncertainty** reflects randomness that is inherent, irreducible and unpredictable in nature. **Epistemic Uncertainty** arises primarily from limited or imperfect knowledge about the parameters of a statistical model and can reflect fluctuations of the parameter, see Figure 2.3. Obtaining more or better information about the parameter typically reduces the epistemic uncertainty.

Let us denote

- $T = time$

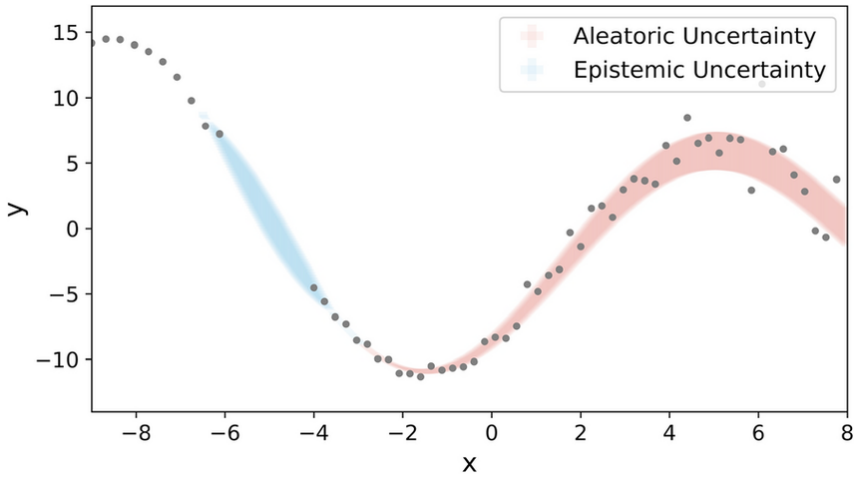

Figure 2.3: Visualization of two types of uncertainty (Yang and Li, 2023).

- $C = counts$
- $\lambda = \frac{C}{T}$

We define the rate ($\lambda = \frac{C}{T}$) at which eligible patients are entered onto a clinical trial, measured as persons per unit of time as **Accrual Rate**. Where **rate** is understood as a ratio in which the numerator and denominator are incremental differences (Piantadosi, 2024).

Methods	Counts	Expectation	Variance	Aleatory	Epistemic
Expectation	$C(t) = \lambda t$	λt	0	No	No
Poisson	$C(t) \sim \text{Po}(\lambda t)$	λt	λt	Yes	No
Negative Binomial	$C(t) \sim \text{Po}(\Lambda t);\ \Lambda \sim \text{G}(\alpha, \beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha(\beta+1)}{\beta^2}$	Yes	Yes

Table 2.1: Moments and aleatory and epistemic uncertainty in accrual covered by different models for counts.

2.3 Counts: Model based on Expectations

If we fix the duration of a study at time T and we expect that we collect C patients until T, we deterministically predict the recruitment rate (without taking into consideration any uncertainty) to be $\hat{\lambda} = \frac{C}{T}$.

2.3.1 Expected recruitment in one unit of time

$C = \text{E}C = \text{E}\lambda = \lambda\ \text{Var}(C) = \text{Var}(\lambda) = 0$

2.3.2 Expected accrual at time point t

$$C(t) = \text{E}(\underbrace{C + \ldots + C}_{t \text{ times}}) = \text{E}(\lambda t) = \lambda t\ \text{Var}(C(t)) = \text{Var}(\underbrace{C + \ldots + C}_{t \text{ times}}) = t\text{Var}(\lambda) = 0$$

Both the expected accrual and its zero-variance are recorded in Table 2.1 and visualized in Figure 2.4 and Figure 2.7.

2.4 Counts: Model based on Poisson Process

The Poisson distribution $C \sim \text{Po}(\lambda)$ allows us to explain the recruitment of patients. It is a discrete variable that expresses the probability of a given number of events (in our case, patient recruitment) occurring in a fixed unit interval of time. We assume that these events occur with a known constant rate λ and are independent of each other.

$$\begin{aligned} \text{P}[C = c] &= \frac{\lambda^c}{c!} e^{-\lambda} \\ c &= 0, 1, 2, \ldots \end{aligned}$$

One important property from the Poisson distribution is that it is infinitely divisible (Held and Bové, 2014). If $X_i \sim \text{Po}(\lambda_i)$ for $i = 1, \ldots, n$ are independent, then, $\sum_{i=1}^{n} X_i \sim \text{Po}\Big(\sum_{i=1}^{n} \lambda_i\Big)$.

2.4.1 Recruitment in one unit of time

The recruitment of patients in one unit of time follows $C \sim \text{Po}(\lambda)$ and the expectation and variance are:

$$\begin{aligned} \text{E}C &= \lambda \\ \text{Var}(C) &= \lambda \end{aligned}$$

2.4.2 Accrual at time point t

At time point t, the accrual follows $C \sim \text{Po}(\lambda t)$. Using the infinitely divisible property from the Poisson applicable to independent random variables, $\underbrace{\text{Po}(\lambda) + \cdots + \text{Po}(\lambda)}_{t \text{ times}} = \text{Po}(\lambda t)$. We assume that the recruitment of patients at a given time point is independent from another. As we can see in Table 2.1, the expectation and variance are the following:

$$\begin{aligned} \text{E}C(t) &= \lambda t \\ \text{Var}(C(t)) &= \lambda t \end{aligned}$$

For example, if we assume $\lambda = 0.591$ per day and $T = 550$, we can show the accrual of 100 different studies in Figure 2.4 and the histogram at $t = 550$ days. The exact distribution at $T = 550$ is provided in Figure 2.5 and the Cummulative Distribution Function (CDF) in Figure 2.6. The uncertainty bands based on the theoretical quantiles are displayed in Figure 2.7.

2.5 Counts: Negative Binomial model derived from Poisson-Gamma model

There are two different interpretations of the Negative Binomial, Failure-Based and Count-based.

2.5.1 Failure-Based

1. The Negative Binomial $X \sim \text{NBin}(r, \pi)$ models the number of **failures** before achieving a fixed number of **successes** in a sequence of Bernoulli trials.

2. Parametrization:

   - r: Number of successes to be achieved (fixed).

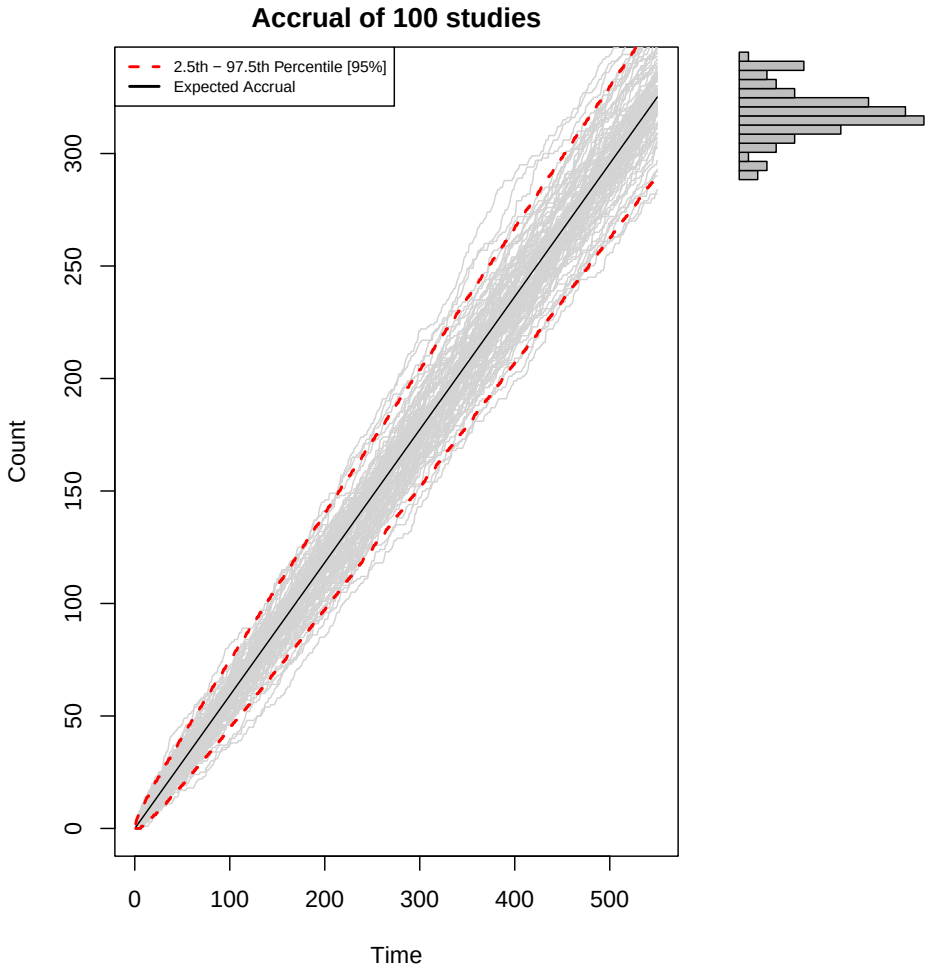

Figure 2.4: Poisson-distributed counts with $\lambda = 0.591$ per day and uncertainty range. The black line represents the point estimate of the expected accrual from section 2.3, while the red dashed lines indicate Poisson's 95% aleatory uncertainty. The histogram illustrates the distribution of observed counts in 100 studies at time $t = 550$ days (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

- π: Probability of success in each trial
- The random variable X represents the number of failures before achieving r successes.

3. Probability Mass Function (PMF):

$$\mathrm{P}(X = k) = \binom{k+r-1}{k} \pi^r (1-\pi)^k,$$
$$k \geq 0$$

where k is the number of failures.

2.5.2 Count-Based

1. The Negative Binomial $X \sim \mathrm{NBin}\left(\alpha, \frac{\beta}{\beta+1}\right)$ can also be seen as a Poisson-Gamma mixture, where the observed count data follows a Poisson distribution with a mean that itself follows a Gamma distribution, $C|\Lambda \sim \mathrm{Po}(\Lambda)$ and $\Lambda \sim \mathrm{G}(\alpha, \beta)$.

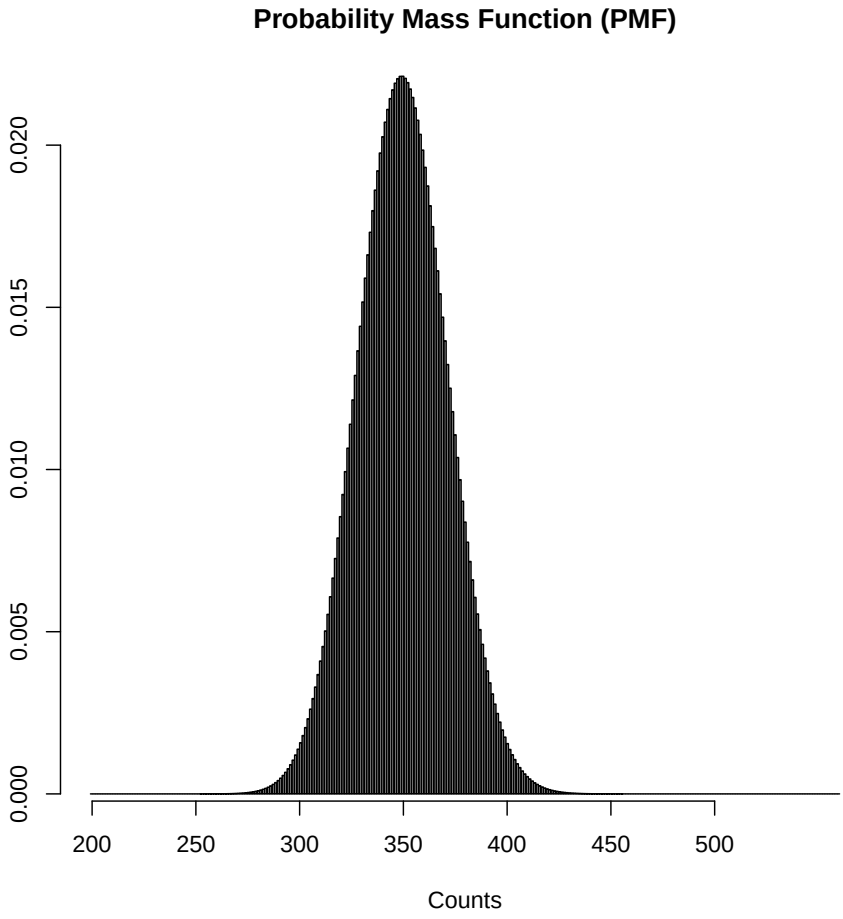

Figure 2.5: Probability Mass Function (PMF) of Poisson-distributed counts: This bar plot represents the probability mass function (PMF) of counts ranging from 200 to 500, using a Poisson distribution $\mathrm{Po}(\lambda t)$ with a rate parameter $\lambda = 0.591$ per day at time $t = 550$ days.

2. Parametrization:

   - $\mu = \frac{\alpha}{\beta}$: Mean of the distribution (expected number of occurrences).
   - α: Dispersion parameter, controlling the variance.

3. Alternative formulation of the PMF:

$$\mathrm{P}(X = c) = \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\mu}{\beta + \mu}\right)^{c} \left(\frac{\alpha}{\alpha + \mu}\right)^{\alpha},$$
$$c \geq 0$$

   where c is the counts.

How do we get to our formulation of the PMF:

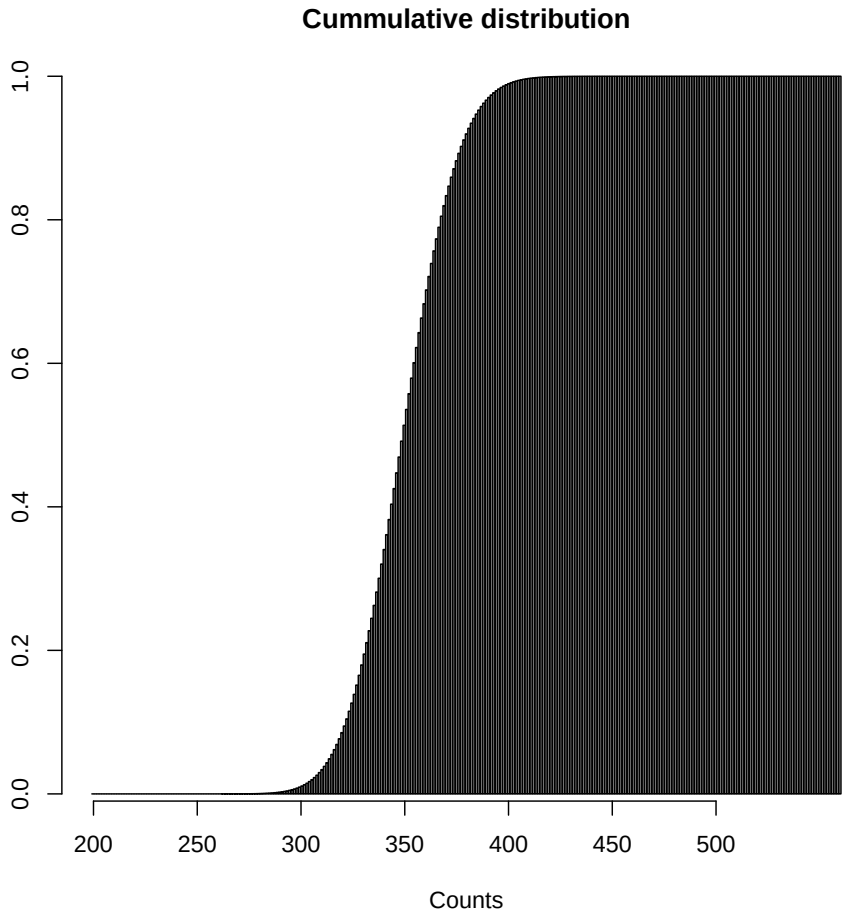

Figure 2.6: Cumulative Distribution Function (CDF) of Poisson-distributed counts: The bar plot illustrates the cumulative probability distribution for counts within the range of 200 to 500, using a Poisson $\mathrm{Po}(\lambda t)$ distribution with a rate parameter $\lambda = 0.591$ per day at time $t = 550$ days.

$$
\begin{aligned}
\mathrm{P}(X = c) &= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\mu}{\beta + \mu} \right)^{c} \left(\frac{\alpha}{\alpha + \mu} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha/\beta}{\alpha + \alpha/\beta} \right)^{c} \left(\frac{\alpha}{\alpha + \alpha/\beta} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha/\beta}{\alpha\beta/\beta + \alpha/\beta} \right)^{c} \left(\frac{\alpha}{\alpha\beta/\beta + \alpha/\beta} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha}{\alpha\beta + \alpha} \right)^{c} \left(\frac{\beta\alpha}{\alpha\beta + \alpha} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{1}{\beta + 1} \right)^{c} \left(\frac{\beta}{\beta + 1} \right)^{\alpha}
\end{aligned}
$$

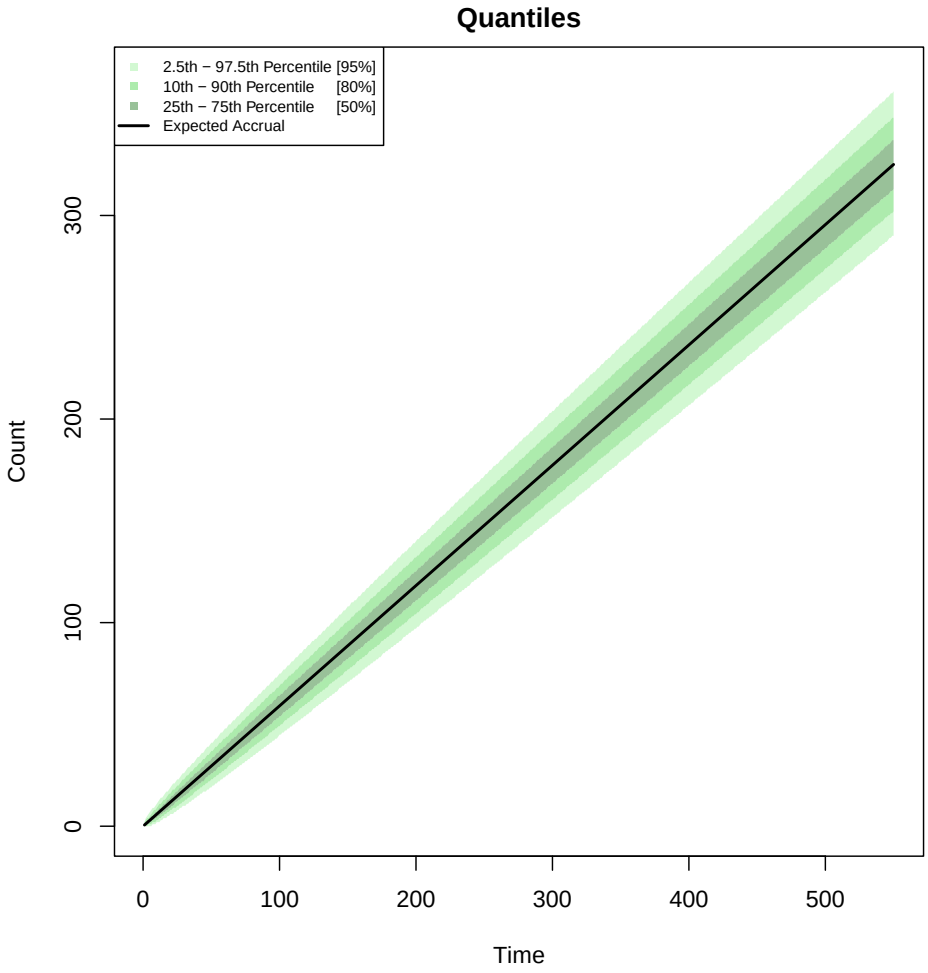

Figure 2.7: Predicted uncertainty bands for Poisson process with $\lambda = 0.591$ per day. The black line represents the expected accrual, while the green shaded regions indicate aleatory uncertainty: the dark green band spans the interquantile range (25th - 75th percentiles), the lighter green band cover the 10th - 90th percentile range and the light green the 2.5th - 97.5th percentile range (Spiegelhalter *et al.*, 2011).

2.5.3 Recruitment in one unit of time

Let $C|\Lambda \sim \text{Po}(\Lambda)$ and $\Lambda \sim \text{G}(\alpha, \beta)$

$$\begin{aligned}
p(c) &= \int_0^\infty p(c|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^c \exp(-\lambda)}{c!}\left[\lambda^{\alpha-1}\exp(-\beta\lambda)\frac{\beta^\alpha}{\Gamma(\alpha)}\right]d\lambda \\
&= \frac{\beta^\alpha}{c!\Gamma(\alpha)}\int_0^\infty \lambda^{\alpha+c-1}\exp(-\lambda)\exp(-\lambda\beta)d\lambda \\
&= \frac{\beta^\alpha\Gamma(\alpha+c)}{c!\Gamma(\alpha)(\beta+1)^{\alpha+c}}\underbrace{\int_0^\infty \frac{(\beta+1)^{\alpha+c}}{\Gamma(\alpha+c)}\lambda^{\alpha+c-1}\exp(-(\beta+1)\lambda)d\lambda}_{=1} \\
&= \beta^\alpha\binom{\alpha+c-1}{\alpha-1}\left(\frac{1}{\beta+1}\right)^{\alpha+c} \\
&= \binom{\alpha+c-1}{\alpha-1}\left(\frac{1}{\beta+1}\right)^{c}\left(\frac{\beta}{\beta+1}\right)^{\alpha}
\end{aligned}$$

Thus, $C|\Lambda \sim \text{NBin}\left(\alpha, \frac{\beta}{\beta+1}\right)$

Using the expressions of iterated expectation and variance (Held and Bové, 2014) and the expectation and variance from the respective random variables $C|\Lambda \sim \text{Po}(\Lambda)$ and $\Lambda \sim \text{G}(\alpha, \beta)$, we have that:

$$\text{E}C = \text{E}_\Lambda[\text{E}_C(C|\Lambda)] = \text{E}_\Lambda[\Lambda] = \alpha/\beta$$

$$\begin{aligned}
\text{Var}(C) &= \text{Var}_\Lambda[\text{E}_C(C|\Lambda)] + \text{E}_\Lambda[\text{Var}_C(C|\Lambda)] \\
&= \text{Var}_\Lambda[\Lambda] + \text{E}_\Lambda[\Lambda] \\
&= \alpha/\beta^2 + \alpha/\beta = \frac{\alpha(\beta+1)}{\beta^2}
\end{aligned}$$

With respect to the parameters, $r > 0$ represents the number of successes until the experiment is stopped. The success probability in each experiment is represented by $\pi \in [0, 1]$. In R the functions `nbinom(..., size = r, prob = π)` relate to the random variable $X - r$, the number of successes (as opposed to the number of trials) until r successes have been achieved (Held and Bové, 2014).

$$\begin{aligned}
EX &= \frac{r(1-\pi)}{\pi} \\
Var(X) &= \frac{r(1-\pi)}{\pi^2}
\end{aligned}$$

Since we will be using the Count-Based interpretation of the Negative Binomial, our parametrization relates to R with $r = \alpha$ and $\pi = \frac{\beta}{\beta+1}$.

2.5.4 Accrual at time point t

Let $C|\Lambda \sim \text{Po}(\Lambda t)$ and $\Lambda \sim \text{G}(\alpha, \beta)$

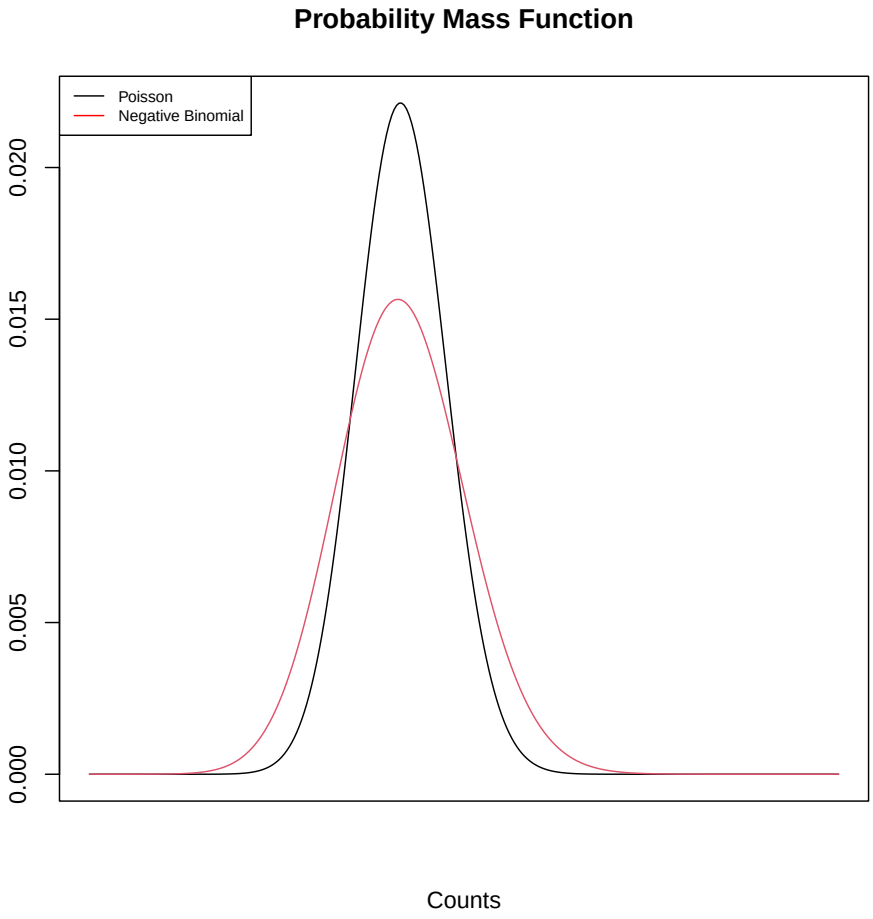

Figure 2.8: Comparison of Probability Mass Function (PMF) between Poisson distribution with $\lambda = 0.591$ and Negative Binomial with $\alpha = 324$ and $\mu = 0.591$.

2.6 Important questions when forecasting recruitment at the design-stage of a study

By normal approximation to the Poisson distribution $C \sim \mathrm{N}(\mu = \lambda T, \sigma^2 = (\lambda T)^2)$, we know that the probability of recruiting the desired N participants is 0.5. Which means that the study has 50% chance of obtaining the desired sample size in the suggested T (Carter, 2004). We would also be assuming that the recruitment rate is constant over time.

In fact, we do not need normal approximation to see this...

This raises two questions which will be answered throughout this Master Thesis:

1. **Rate:** If T is fixed, what does the expected rate λ need to be to achieve a certain certainty of enrolling the total sample size N within the time frame T?

2. **Time:** Given a certain rate λ, how long should the recruitment period T be planned to give a confidence above 50% of recruiting the total sample size N? In Machine Learning, this confidence is aimed at 80%. In Carter (2004), at 90%.

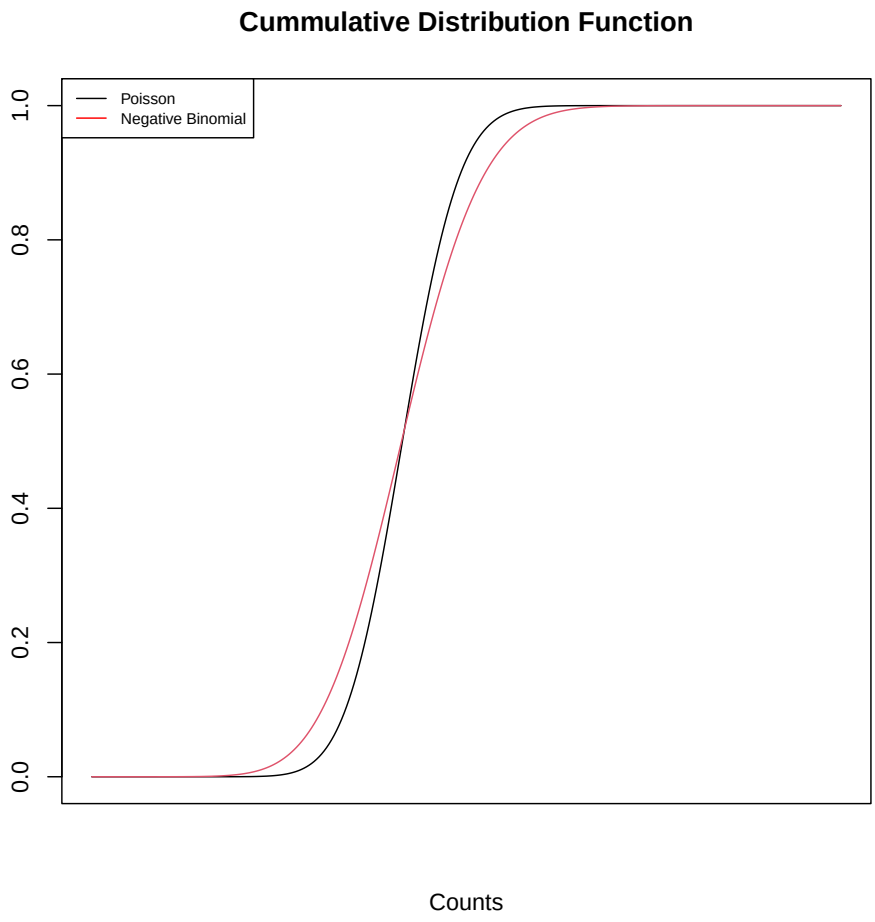

Figure 2.9: Comparison of Cummulative Distribution Function (CDF) between Poisson distribution with $\lambda = 0.591$ and Negative Binomial with $\alpha = 324$ and $\mu = 0.591$.

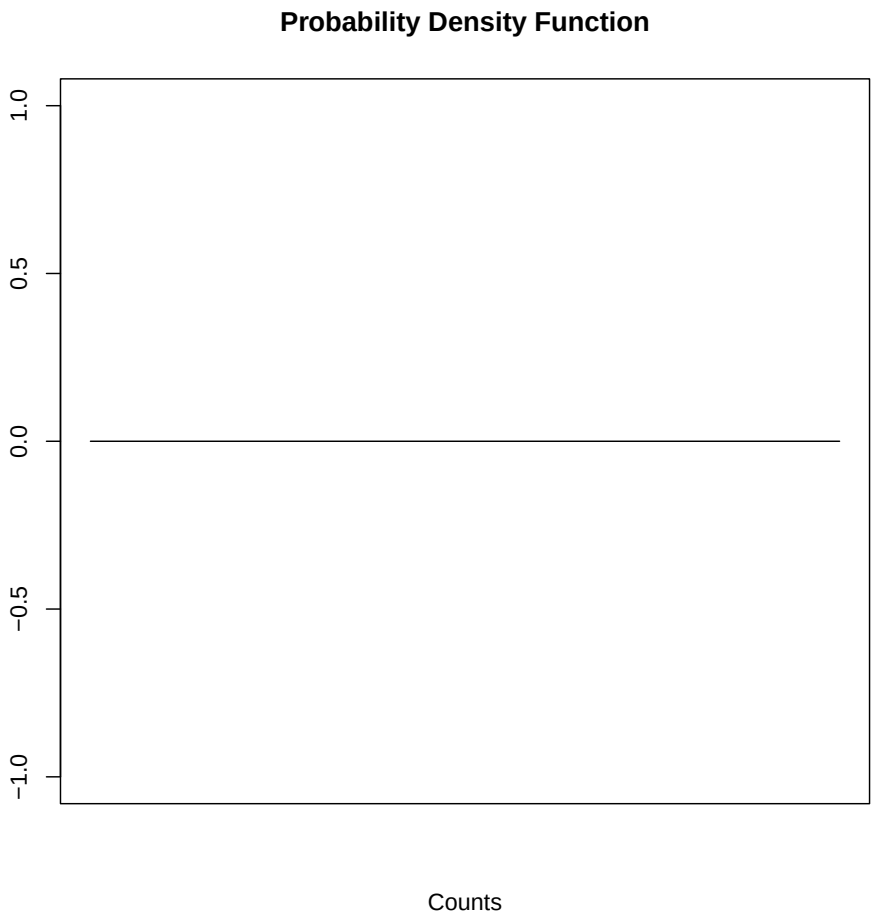

Figure 2.10: Gamma density function with $\alpha = 324$ and $\beta = 1.5 \cdot 365$.

Chapter 3

Results

Chapter 4

Discussion and Outlook

Chapter 5

Conclusions

5.0.1 Personal Statement

NOT AI: This thesis was not written by any generative AI. It was written independently and without assistance from third parties. All sources utilized in this thesis are appropriately cited in the references

AI: During the preparation of this Master Thesis, I used [NAME OF TOOLS AND SERVICES] in order to [REASON]. After using this tool/service, I reviewed and edited the content as needed and I take full responsibility for the content of the Master Thesis.

Bibliography

Bogin, V. (2022). Lasagna's law: A dish best served early. *Contemporary Clinical Trials Communications*, **26**, 100900. 4

Carter, R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, **25**, 429–436. 12

Desai, R. (2014). *Preventing patient volume leakage in healthcare systems*. PhD thesis, University of Pittsburgh. 3

Food, Administration, D.,*et al.* (2018). Evaluating inclusion and exclusion criteria in clinical trials. In *Workshop Report. The National Press Club, Washington DC*. 3

Frank, G. (2004). Current challenges in clinical trial patient recruitment and enrollment. *SoCRA Source*, **2**, 30–38. 3

Held, L. and Bové, D. S. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 16. 6, 11

Lim, C.-Y. and In, J. (2019). Randomization in clinical studies. *Korean Journal of Anesthesiology*, **72**, 221–232. 3

Liu, J., Jiang, Y., Wu, C., Simon, S., Mayo, M. S., Raghavan, R., and Gajewski, B. J. (2023). *accrual: Bayesian Accrual Prediction*. R package version 1.4. 7

National Institute of Allergy and Infectious Diseases (2021). Screening, enrollment, and unblinding of participants. https://www.niaid.nih.gov/sites/default/files/screening-enrollment-unblinding-of-participants.pdf. Accessed: 2025-02-10. 3

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W., Bürgi, E., Scherer, M., Altman, D. G., and Jüni, P. (2009). The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ*, **339**, . 3

O'Hagan, A. e. a. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons. 4

Panos, G. D. and Boeckler, F. M. (2023). Statistical analysis in clinical and experimental medical research: Simplified guidance for authors and reviewers. 3

Piantadosi, S. (2024). *Clinical trials: a methodologic perspective*. John Wiley & Sons. 5

Piantadosi, S. and Meinert, C. L. (2022). *Principles and Practice of Clinical Trials*. Springer Nature. 4

Rehman, A. M., Ferrand, R., Allen, E., Simms, V., McHugh, G., and Weiss, H. A. (2020). Exclusion of enrolled participants in randomised controlled trials: what to do with ineligible participants? *BMJ Open*, **10**, e039546. 3

Shih, W. J. (2002). Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine*, **3**, 1–7. 3

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, **333**, 1393–1400. 7, 10

Van Spall, H. G., Toren, A., Kiss, A., and Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA*, **297**, 1233–1240. 3

Whelan, J., Le Deley, M.-C., Dirksen, U., Le Teuff, G., Brennan, B., Gaspar, N., Hawkins, D. S., Amler, S., Bauer, S., Bielack, S.,*et al.* (2018). High-dose chemotherapy and blood autologous stem-cell rescue compared with standard chemotherapy in localized high-risk ewing sarcoma: Results of euro-ewing 99 and ewing-2008. *Journal of Clinical Oncology*, **36**, 3110–3119. 4

Willie, M. M. (2024). Population and target population in research methodology. *Golden Ratio of Social Science and Education*, **4**, 75–79. 3

Yang, C.-I. and Li, Y.-P. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, **15**, 13. 5