FFI Norwegian Defence
Research Establishment

23/02387

FFI-RAPPORT

An analytical review of perforation statistics

– using aluminium as witness plates in perforation experiments

Martin Fonnum Jakobsen

An analytical review of perforation statistics
– using aluminium as witness plates in perforation experiments

Martin Fonnum Jakobsen

Norwegian Defence Research Establishment (FFI) 20 December 2023

Keywords
Statistikk
Perforering
Terminalballistikk
Aluminiumplater

FFI report
23/02387

Project number
1680

Electronic ISBN
978-82-464-3508-4

Approvers
Morten Huseby, *Research Manager*
Halvor Ajer, *Director of Research*

The document is electronically approved and therefore has no handwritten signature.

(U) Summary

This report documents the statistical methods that are used in perforation experiments at the Norwegian Defence Research Establishment (FFI). We also introduce new (or at least unused) statistical methods, which have certain advantages over the ones we currently employ.

A concrete example is the receiver operating characteristic curves (ROC curves) and corresponding classification tables, which provide a much more direct method to evaluate the goodness of fit compared to the more conventional pseudo-R^2 measures. The ROC curves take multiple additional details into account (such as true positives and negatives and false positives and negatives) that R^2 measures do not when evaluating model performance. It is therefore a more reliable goodness of fit metric. We recommend using both R^2 and ROC methods to supplement each other.

As an example of the application of the various statistical methods, we evaluate whether a 0.5 mm aluminium plate is suitable as a witness plate in perforation experiments. We find that the aluminium plate is suitable as a witness plate for shots against skin simulants modelling the abdomen and back. For other target locations, such as the thorax, thigh, or buttocks, we find that it is necessary to use plates that are more easily perforated. Both the R^2 measures and the ROC curves indicate that the experimental model has a high goodness of fit.

(U) Sammendrag

Denne rapporten dokumenterer de nåværende statistiske metodene som blir brukt i perforeringseksperimenter ved Forsvarets forskningsinstitutt. I tillegg introduseres nye (eller i hvert fall ubrukte) statistiske metoder som har visse fordeler sammenlignet med metodene som blir benyttet per dags dato.

Et konkret eksempel er ROC-kurver (opprinnelig «kurver for en mottakers operasjonskarakteristikk» – receiver operating characteristic curves – men nå brukt i overført betydning) med tilhørende klassifikasjonstabeller, som lar oss mer nøyaktig bedømme modellegnethet enn om vi utelukkende hadde brukt mer konvensjonelle pseudo R^2-mål. Selv om ROC-kurvene inneholder mer informasjon (som sanne positiver og negativer og falske positiver og negativer) enn R^2-målene, anbefaler vi at begge metodene brukes for å supplere hverandre.

Som et konkret eksempel på anvendelser av de nye og eldre statistiske metodene studerer vi egenskapene til en 0,5 mm aluminiumsplate for å finne ut om den egner seg som vitneplate i perforeringseksperimenter. De statistiske resultatene tilsier at aluminiumsplaten egner seg godt som vitneplate for skudd mot hudsimulanter som representerer mageregionen og rygg. For andre mål, som brystkassen og lår, trengs det tynnere vitneplater. Både R^2-målene og ROC-kurvene indikerer at den eksperimentelle modellen gir resultater som samsvarer godt med dataene.

Contents

Appendix

1 Introduction

In the ballistic lab at the Norwegian Defence Research Establishment (FFI), perforation experiments are performed routinely. Usually, the goal is to determine in which velocity regime the armour system of interest is perforated or the projectile is stopped. The purpose of this report is to document the statistical methods that can be used to analyze perforation experiments. Some of these methods have been used at FFI before, while some have not been used. In either case, we believe a detailed documentation can elucidate the strengths and weaknesses of the statistical methods we employ, and in particular provide a deeper understanding of how the machinery works.

To provide a concrete example of the statistical methods we have performed a perforation experiment using a 0.5 mm aluminium plate. Such a system is often used passively in perforation experiments as a witness plate, to detect whether a projectile perforated an armour system typically made from steel, ceramics, and ballistic fibres. In our experiment, we have used the aluminium plates we had available which is of the type 1050-H14. We have compared the perforation properties of the aluminium plate to the perforation properties of human skin. We find that there is overlap in the perforation–velocity regime for aluminium plate and human skin, and conclude that the aluminium plate is suitable as a witness plate for shots against the abdomen and back. For shots against the thorax, thigh, or buttocks we find that witness plates that are more easily perforated may be suitable.

2 Perforation statistics

In this chapter we will introduce some of the statistics necessary to analyze perforation experiments. Concretely, we will introduce the various regression models used to describe binary outcomes, how the parameterers of the regression models are determined using the maximum likelihood formulation, and how to define and calculate the relevant confidence intervals.

2.1 Regressional models

In a perforation experiment a projectile is fired at a target, typically an armour system. Right before impact the velocity of the projectile is measured, and after impact it is recorded whether or not the armour was perforated. The set of data is therefore binary in nature. A set of data for n shots can be denoted as $\{(v_1, y_1), (v_2, y_2), \dots (v_n, y_n)\}$, where v_i is the projectile velocity and y_i determines whether the armour is completely perforated $(y_i = 1)$ or not $(y_i = 0)$.

From the experimental data it is possible to approximately determine the probability of perforation as a function of velocity if a suitable regression scheme is utilized. We will denote the desired probability as $P(y = 1|v)$. In an ordinary linear regression one would attempt to write something of the form

$$P(y = 1|v) = \beta_0 + \beta_1 v + \epsilon. \tag{2.1}$$

Here β_0 and β_1 are the regression coefficients and ϵ denotes the error term. In a linear regression the coefficients become constant, and since there is linear dependence on velocity there will be an unphysical regime where the probability is greater than one or smaller than zero. Therefore, as well as for other more technical reasons, the binary nature of the data makes ordinary linear regression unsuitable for this problem. The solution is to instead use a so–called generalized linear model (GLM) of the form

$$P(y = 1|v) = F\left(\beta_0 + \beta_1 v + \epsilon\right) \tag{2.2}$$

where a nonlinear function F (also known as the link function) is used to map the probability back to the physical interval $[0, 1]$. In the context of binary data the most frequent choices of F are the ones described by an S shaped (also called sigmoid) curve. There are several choices of S shaped curves, but the two most frequently used are the so-called probit and logit models. In App. A we provide a derivation of why it is natural to use the probit or logit model. The functional form of the probit and logit is

$$F(z) \equiv \Phi(z) = \int_{-\infty}^{z} \phi(u) \mathrm{d}u, \tag{2.3}$$

and

$$F(z) \equiv \Lambda(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}, \tag{2.4}$$

respectively. Here $\Phi(z)$ and $\Lambda(z)$ denotes the cumulative distribution function (cdf) of the normal and logistic distribution. $\phi(u)$ denotes the normal probability density (pdf) and $z = \beta_0 + \beta_1 v$.

The mean μ and standard deviation σ of the probit and logit model is given by

$$\mu_{\text{Probit}} = -\frac{\beta_0}{\beta_1}, \qquad \sigma_{\text{Probit}} = \frac{1}{\beta_1}, \tag{2.5}$$

and

$$\mu_{\text{Logit}} = -\frac{\beta_0}{\beta_1}, \qquad \sigma_{\text{Logit}} = \frac{\pi}{\sqrt{3}} \frac{1}{\beta_1}, \tag{2.6}$$

respectively.

2.2 Maximum likelihood estimation of regression coefficients

Most programming languages and statistical software (Matlab, python, Mathematica, R) are equipped with statistical packages such that the determination of the regression coefficients become trivial. Nevertheless, it is useful to know how the machinery works, which is what we will succinctly explain in this section.

2.2.1 Coefficients

To determine the statistical regression coefficients the maximum likelihood estimation (MLE) method is used. Given some observed data the likelihood function is maximized so that, under the assumed statistical model, the observed data is the most probable.

Since the data is binary there are only two things which may occur. Either the projectile perforates the armour with probability $P(y = 1|v) = F(z)$ or it doesn't with probability $P(y = 0|v) = 1 - P(y = 1|v) = 1 - F(z)$. This can be succintly expressed on a binomial form as

$$P(y = y_i|v) = [F(z)]^{y_i} [1 - F(z)]^{1-y_i} \tag{2.7}$$

where $y_i = 0, 1$. The likelihood of obtaining a set of experimental data $\{(v_1, y_1), (v_2, y_2), \ldots (v_n, y_n)\}$ can be written as the following product

$$L_n = \prod_{i=1}^{n} P(y = y_i|v) = \prod_{i=1}^{n} [F(z)]^{y_i} [1 - F(z)]^{1-y_i} . \tag{2.8}$$

Directly maximizing the above product can become numerically unwieldy. Instead it is customary to work with the so–called log–likelihood where we take the logarithm to convert the product into the following sum

$$\mathcal{L}_n = \ln L = \sum_{i=1}^{n} y_i \ln [F(z)] + (1 - y_i) \ln [1 - F(z)] , \tag{2.9}$$

where $z = \beta_0 + \beta_1 v$. The maximum likelihood method works by solving the equations

$$\frac{\partial \mathcal{L}_n}{\partial \beta_0} = 0 \quad \frac{\partial \mathcal{L}_n}{\partial \beta_1} = 0 \tag{2.10}$$

numerically. Two concrete numerical schemes often employed when solving Eqs. (2.10) are the gradient descent method or the Newton–Raphson method.

2.2.2 Confidence interval on coefficients

In general, the maximum likelihood estimators have a number of attractive limit properties, which is often referred to as the regularity conditions. For our purposes the most important property is that the estimated parameters $\beta = (\beta_0, \beta_1)$ are themselves normally distributed for sufficiently large sample sizes

$$\hat{\beta} \sim N_2 \left(\beta_T, I_n^{-1}(\beta_T)\right), \quad n \longrightarrow \infty. \tag{2.11}$$

Here β_T is a vector containing the true values of the parameters and the covariance matrix[1] is the inverse of the so–called Fisher information matrix which is defined as

$$I_n(\beta_T) = E\left[\left(\frac{\partial \mathcal{L}_n}{\partial \beta}\right)\left(\frac{\partial \mathcal{L}_n}{\partial \beta}\right)^T\right] = -E\left[\frac{\partial^2}{\partial \beta^2}\mathcal{L}_n\right] \tag{2.12}$$

where the expectation value is taken with respect to the distribution of the dataset $\{(v_1, y_1), \ldots, (v_n, y_n)\}$. In the following we denote the covariance matrix as $V_\beta = I_n^{-1}$, and set $\beta_T = \hat{\beta}$ since we do not actually know the true values of the parameters. If the dataset are identically distributed and independent of each other the Fisher information matrix can be computed using only one observation $I_n = nI_1$. In practice this means that the error for the estimated regression coefficients decreases with the sample size as $1/n$.

Since the estimated parameters β obey a normal distribution with covariance matrix V_β a confidence interval on the parameters is given by

$$\begin{gathered}\left[\beta_0 - Z_{1-\frac{\alpha}{2}}\sqrt{(V_\beta)_{11}}, \beta_0 + Z_{1-\frac{\alpha}{2}}\sqrt{(V_\beta)_{11}}\right], \\ \left[\beta_1 - Z_{1-\frac{\alpha}{2}}\sqrt{(V_\beta)_{22}}, \beta_1 + Z_{1-\frac{\alpha}{2}}\sqrt{(V_\beta)_{22}}\right].\end{gathered} \tag{2.13}$$

The number $Z_{1-\frac{\alpha}{2}}$ is the Z–score[2] of the normal distribution with significance level α. The confidence level is expressed through the significance level as $100 \cdot (1-\alpha)\%$.

2.3 Confidence interval on velocities

In a perforation experiment we are typically interested in the upper or lower quantiles of the velocity which we can estimate from the regression. Equally important we are interested in how much of an error our estimate contains. For example, when comparing various types of armour we are often interested in V_{90}, which is the projectile velocity at which the armour will be perforated 90% of the time. However, depending on the application we could also be interested in other quantiles such as V_{10}, V_{50}, V_{95}, etc. In the following we will derive an expression that can be used to determine the confidence interval on any velocity quantile of interest. The calculations are based on related works from the US Research Laboratory [1, 2].

In general, a confidence interval for an estimated quantity $\hat{O}$ can be written on the form

$$\left[\hat{O} - Z_{1-\frac{\alpha}{2}}\mathrm{SE}(\hat{O}), \hat{O} + Z_{1-\frac{\alpha}{2}}\mathrm{SE}(\hat{O})\right] \tag{2.14}$$

where $\mathrm{SE}(\hat{O})$ is the standard error of $\hat{O}$. If there is only one regression coefficient the standard error is the standard deviation. If there are several regression coefficients the standard error of a linear combination of the coefficients is given by $\sqrt{KVK^T}$, where K is a vector describing the linear combination and V is the covariance matrix of the coefficients themselves.

If we use some programming language or statistical software to perform a logit/probit regression we obtain a covariance matrix of the parameters β_0 and β_1, which we will refer to as V_β. To

[1] A covariance matrix is a square matrix giving the covariance between each pair of regression coefficients. The diagonal elements are coefficient variances, and the off–diagonal elements are the covariances between pairs of coefficients. The covariance matrix is by construction symmetric and positive definite.

[2] Note that if the sample size is small, one can instead use the T–score where the number of degrees of freedom is the sample size minus the number of estimated parameters: d.o.f. $= n - 2$.

estimate the quantiles we need estimates for the mean $\mu = V_{50}$ and slope s of the S shaped regression curve. In the logit/probit model the mean and slope can be expressed as $\mu = -\beta_0/\beta_1$ and $s = 1/\beta_1$ respectively. To simplify the notation, we collect the mean and slope into a single matrix $\theta = [\mu, s]^T$. To determine the covariance matrix V_θ between μ and s we use the matrix equivalent of the chain rule which states that

$$V_\theta = \left(\frac{\mathrm{d}\theta}{\mathrm{d}\beta}\right)^T V_\beta \left(\frac{\mathrm{d}\theta}{\mathrm{d}\beta}\right) \tag{2.15}$$

where

$$\left(\frac{\mathrm{d}\theta}{\mathrm{d}\beta}\right) = \begin{bmatrix} \mathrm{d}\mu/\mathrm{d}\beta_0 & \mathrm{d}s/\mathrm{d}\beta_0 \\ \mathrm{d}\mu/\mathrm{d}\beta_1 & \mathrm{d}s/\mathrm{d}\beta_1 \end{bmatrix} = -\frac{1}{\beta_1^2} \begin{bmatrix} \beta_1 & 0 \\ -\beta_0 & 1 \end{bmatrix}. \tag{2.16}$$

If we denote v_p as the velocity that corresponds to the probability p we can use the inverse of the probit/logit model to express the velocity as a function of probability

$$v_p = \mu + s \cdot Q_0(p). \tag{2.17}$$

Here $Q_0(p)$ is the quantile function which can be written

$$Q_0(p) = \begin{cases} \sqrt{2}\mathrm{erf}^{-1}(2p - 1) & \text{Probit,} \\ \ln \frac{p}{1-p} & \text{Logit,} \end{cases} \tag{2.18}$$

and

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t \tag{2.19}$$

is the error function. The trick is that we can now express Eq. (2.17) as a product of two vectors

$$v_p = [1, Q_0(p)] \begin{bmatrix} \mu \\ s \end{bmatrix} \equiv K\theta \tag{2.20}$$

where we identified the vector $K = [1, Q_0(p)]$ which describes the appropriate linear combination to express a velocity quantile in terms of the mean and slope.

Finally, to obtain the confidence interval, it is necessary to assume that the vector $\theta \sim N(\theta, V_\theta)$ is normally distributed around the true value of the mean and slope. If this is the case, the product $K\theta \sim N(K\theta, KV_\theta K^T)$ will also be normally distributed. Hence, the confidence interval for the velocities $v_p = K\theta$ take the form

$$\left[v_p - Z_{1-\frac{\alpha}{2}} \sqrt{K^T V_\theta K}, v_p + Z_{1-\frac{\alpha}{2}} \sqrt{K^T V_\theta K}\right]. \tag{2.21}$$

In a similar fashion we can determine the confidence interval on the probability. In this case we use the vectors $K = [1, v]$ and $\beta = [\beta_0, \beta_1]$ so that the probability can be expressed as $p = F(K\beta)$. By then assuming that the coefficients are normally distributed around the true values $\beta \sim N(\beta, V_\beta)$, it follows that the product $K\beta \sim N(K\beta, KV_\beta K^T)$ is also normally distributed. The corresponding confidence interval on the probabilities then take the form

$$\left[F\left(K\beta - Z_{1-\frac{\alpha}{2}} \sqrt{K^T V_\beta K}\right), F\left(K\beta + Z_{1-\frac{\alpha}{2}} \sqrt{K^T V_\beta K}\right)\right]. \tag{2.22}$$

3 Experimental method

In our perforation experiment an airgun was used to fire copperhead 2500 premium BB bullets of diameter 4.5 mm and an average mass of 0.33 g at aluminium plates with a thickness of 0.5 mm. The aluminium plates was made from aluminium 1050-H14. A schematic setup of the experiment is illustrated in Fig. 3.1. For each shot, the impact velocity is measured, and it is recorded whether the projectile perforates or is stopped by the aluminium plate. The use of an airgun allows the impact velocity to be varied, either by controlling the pressure inside the airgun or by varying the shooter–target distance. The raw data of the experiment is available in App. B.

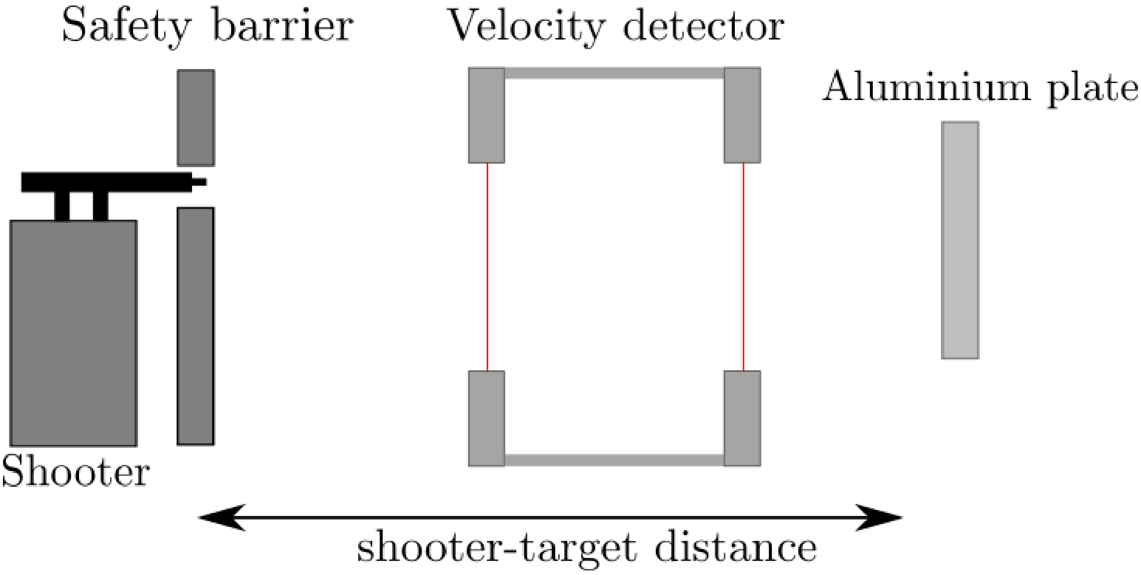
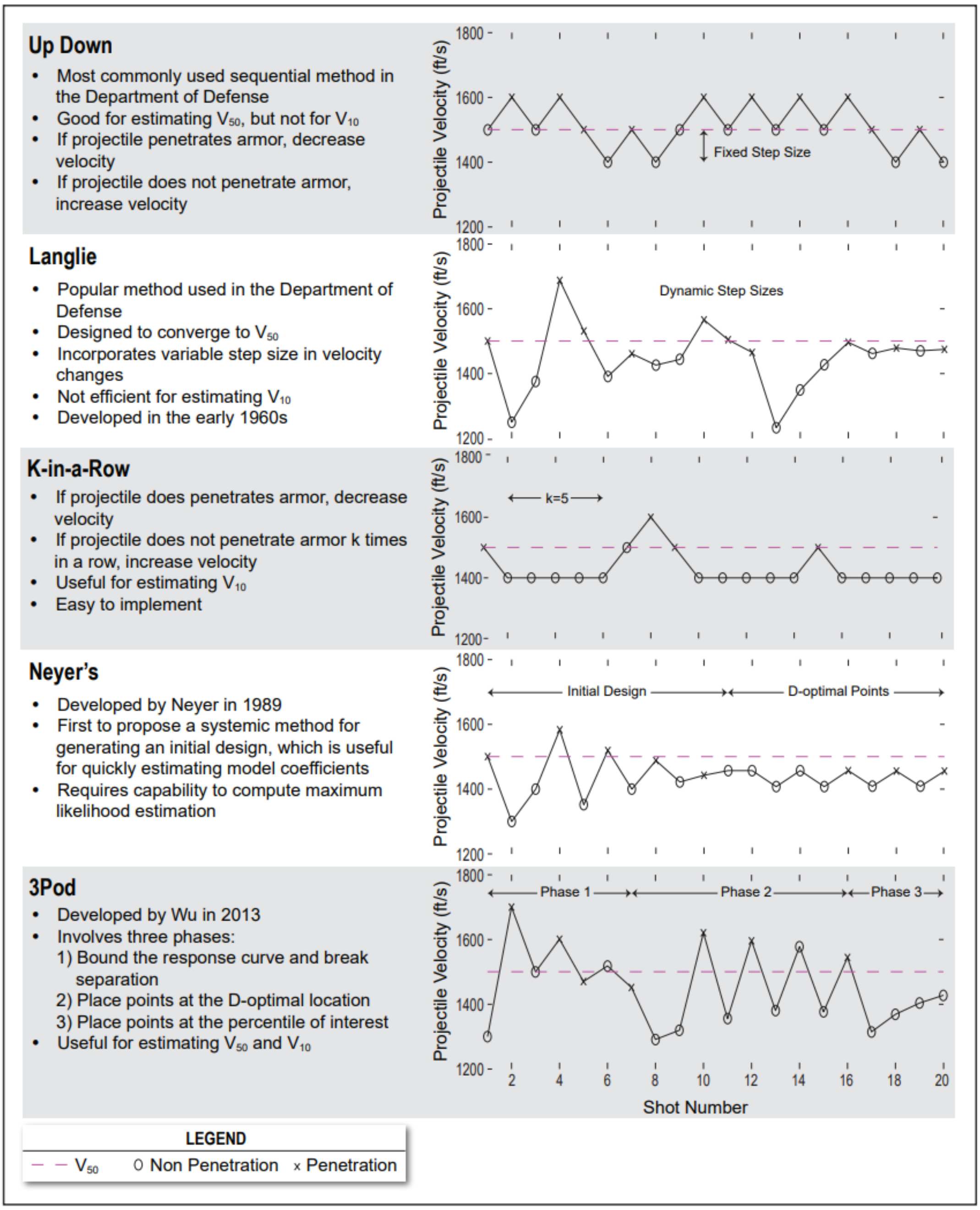


Figure 3.1 A schematic of the experimental setup.

A "good" data set consists of three regions: a high–velocity region of primarily perforations, a low–velocity region of primarily stops, and an overlapping region where there are both perforations and stops. Of most importance is the overlapping region, as it is these data points that provide the most signifcant contributions to the intercept β_0 and slope parameter β_1. In a data set without any overlapping region, the probit or logit method outputs a unit–step function, with infinite slope $\beta_1 = \infty$.

When performing perforation experiments, the shooting method is of importance to obtain data on which a statistical analysis can be performed. There are several choices possible, and the simplest is the up–down method. The up–down method is designed to converge to the V_{50} velocity. Unfortunately, it does not predict the upper or lower quantiles accurately. Therefore alternative shooting methods have been devised. A summary of alternating shooting methods can be found in Fig. 3.2, which is taken from a report written by the American Department of Defence [3].

In the experiment performed here we have utilized the 3POD method [4, 5]. The complete technical details, are provided in the original article. The 3POD method is useful for estimating both V_{50} and an upper velocity quantile, e.g. V_{90}. In simple terms, the estimates are obtained by utilizing different phases, of which there are three in total. In the first phase, the width of the overlap region is crudely estimated, as well as the intercept term β_0 and slope term β_1. In the second phase shots are placed around V_{50} and new estimates of the intercept and slope are provided. In the final and third phase, shots are placed around the upper velocity quantile re–estimating the intercept and slope term. The 3POD method provides better estimates than e.g. up–down, because the shooting is performed around two velocity quantiles, instead of one, which provides a better estimate for the slope.



Up Down

- Most commonly used sequential method in the Department of Defense
- Good for estimating V_{50}, but not for V_{10}
- If projectile penetrates armor, decrease velocity
- If projectile does not penetrate armor, increase velocity

Langlie

- Popular method used in the Department of Defense
- Designed to converge to V_{50}
- Incorporates variable step size in velocity changes
- Not efficient for estimating V_{10}
- Developed in the early 1960s

K-in-a-Row

- If projectile does penetrates armor, decrease velocity
- If projectile does not penetrate armor k times in a row, increase velocity
- Useful for estimating V_{10}
- Easy to implement

Neyer's

- Developed by Neyer in 1989
- First to propose a systemic method for generating an initial design, which is useful for quickly estimating model coefficients
- Requires capability to compute maximum likelihood estimation

3Pod

- Developed by Wu in 2013
- Involves three phases:
  1) Bound the response curve and break separation
  2) Place points at the D-optimal location
  3) Place points at the percentile of interest
- Useful for estimating V_{50} and V_{10}

LEGEND

— — V_{50} O Non Penetration x Penetration

Figure 2. Example Tests for Sequential Methods

Figure 3.2 A comparison of the different shooting methods usually employed in perforation experiments. The figure is taken from [3].

4 Regression analysis of aluminium plate

Having obtained the experimental data we will in this section perform the regression analysis. We use Matlab for the statistical analysis. The logit and probit fit to the experimental data in App. B are shown in Fig. 4.1. We have here used Eq. (2.21) to calculate the 95% confidence interval on the velocities. We note that the probability–velocity relationship provided by the logit and probit models are statistically indistuingishable, since both curves are very similar and lie within each others confidence interval. The curve exhibits the S–shape, and captures the feature that the perforation probability should increase with velocity. Before we can draw conclusions about the aluminium plate, we need to determine whether our model is a good or a bad fit to the experimental data.

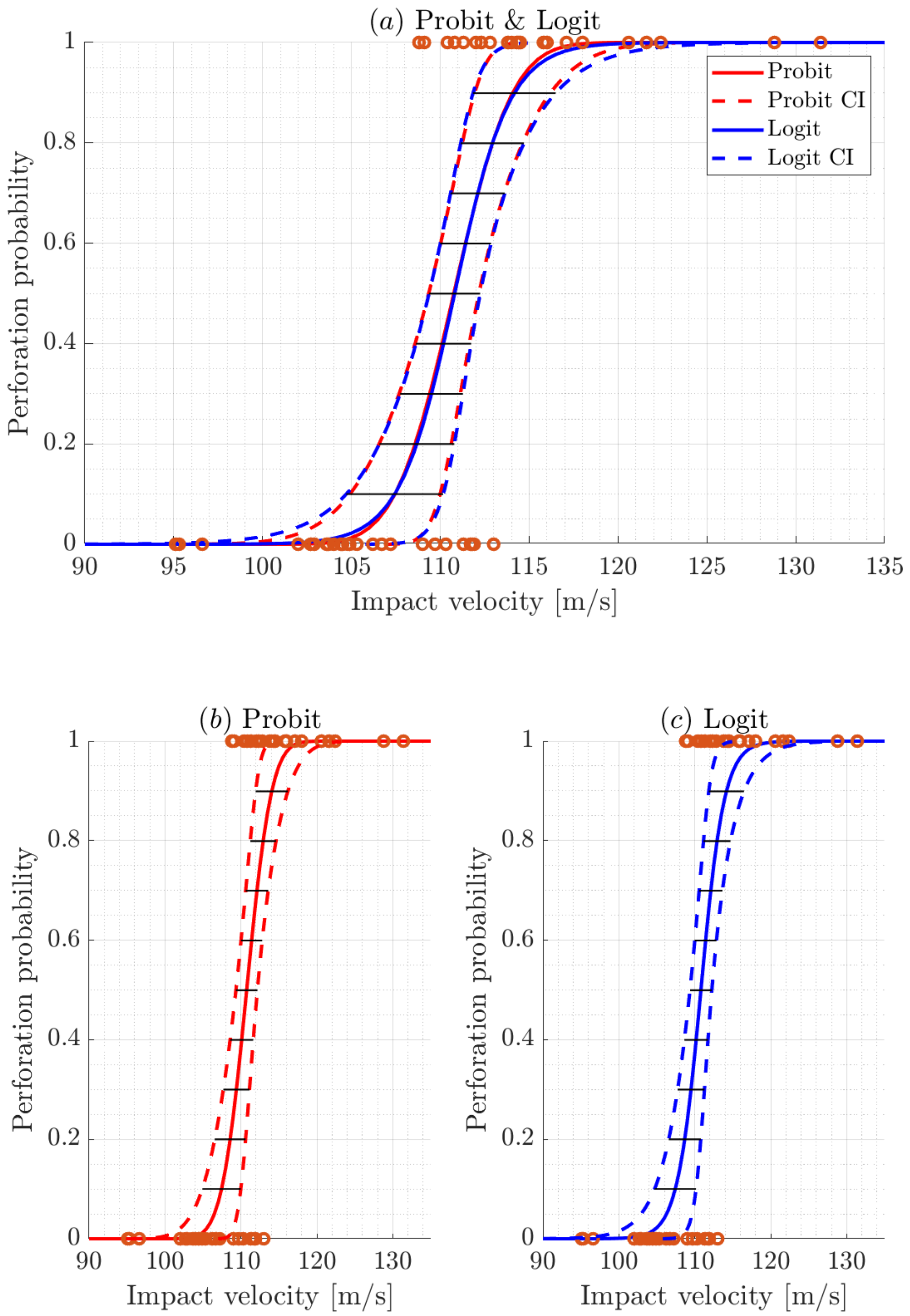


Figure 4.1 A comparison of the probit and logit fit to the experimental data. The dashed lines represent the confidence intervals on the impact velocities.

4.1 Goodness of fit

Next, we discuss the different methods that are useful for evaluating the goodness of fit when employing a logit or probit regression.

4.1.1 Standard statistics of the coefficients

The standard statistics output by Matlab is given in Tab. 4.1. The values and the 95% confidence interval of the intercept (β_0) and slope (β_1) is obtained by maximum likelihood estimation as described in Sec. 2.2. The t–statistic and p–value is obtained from a Wald test as described in App. C.1.

The small p–values indicate that both the intercept (β_0) and slope β_1 are significantly different from zero and should be retained in the regression. The coefficients in the logit and probit are significantly different because the logit and probit have different functional form. Nevertheless, the models are for all purposes practically identical as shown in Fig. 4.1, and we can't learn much more about the goodness of fit from these simple statistics.

Finally, we note that the confidence interval on the coefficients is large. This is a consequence of that the associated standard error is also large. In order to obtain smaller standard errors we would have to fire a much larger number of shots spread more evenly across the velocity range, which for practical purposes are unrealistic. The standard error of the coefficients can nevertheless be used for testing whether the coefficient is significantly different from 0, as we did in App. C.1. The standard errors can also be used to form a confidence interval for the probability according to Eq. (2.22). However, as mentioned before, we are more interested in the probability on the velocities.

Table 4.1 Standard statistics of the logit and probit regression for the aluminium plate.

Model	Intercept β_0			Slope β_1		
	Value (95% CI)	t-statistic	p–value	Value (95% CI)	t-statistic	p–value
Probit	$-42.9\ (-67.1, -18.6)$	-3.46	$0.540 \cdot 10^{-3}$	$0.387\ (0.169, 0.605)$	3.48	$0.511 \cdot 10^{-3}$
Logit	$-72.2\ (-117.1, -27.3)$	-3.15	$1.60 \cdot 10^{-3}$	$0.652\ (0.248, 1.06)$	3.16	$1.60 \cdot 10^{-3}$

4.1.2 Deviance residuals

Deviance is a generalization of the idea of using the sum of squares of residuals (SSR) in ordinary least squares to cases where model-fitting is achieved by maximum likelihood, such as in the logit or probit model. The deviance is defined as the difference in log–likelihood between a saturated (perfect) model and a model of interest. In the saturated model, each data point is connected by a line, thus producing perfect predictions. The deviance can be decomposed into a sum of individual data points. Each term in the sum gives a contribution to the so called deviance residual. Mathematically the deviance is defined and decomposed into a sum as

$$D = 2\left(\mathcal{L}_{\text{saturated}} - \mathcal{L}_{\text{model}}\right) = \sum_{i=1}^{n} d_i, \tag{4.1}$$

and the deviance residual for datapoint i is

$$\text{deviance residual i} = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{d_i}. \tag{4.2}$$

The sign of the deviance residual is determined by whether the datapoint y_i is above or below the predicted value $\hat{\mu}_i$. The size of the deviance residual is determined by the distance between the data point and predicted value, which here is implicitly defined through d_i. A deviance close to zero is thus an indicator of a goodness of fit, and is a useful measure when comparing two different models.

For the saturated model, each datapoint is predicted perfectly. Consequently, the likelihood of the saturated model is one, and its log–likelihood is zero. This leads to the following relationship between the deviance and log–likelihood of the model of interest,

$$D = -2\mathcal{L}_{\text{model}}. \tag{4.3}$$

In Fig. 4.2 we have plotted the deviance residuals for the probit and logit model. The deviance residuals are very similar because the S–shape produced by each curve fit is similar. Naturally both models have small deviance residuals for small and high velocities, where it is the easiest to predict perforation or stop. For intermediate regions, where we have a mix of perforations and stops, the deviance residuals oscillate. The total deviance of the logit and probit model is 28.2 and 27.7 respectively. Naively, this indicates that the probit fits the data slightly better than the logit. We can understand this by noting that in Fig. 4.1 (a) the tail of the logit is slightly heavier than the tail of the probit. We emphasize that the difference in deviances are so small, that we can not conclude that one model is significantly better than the other.

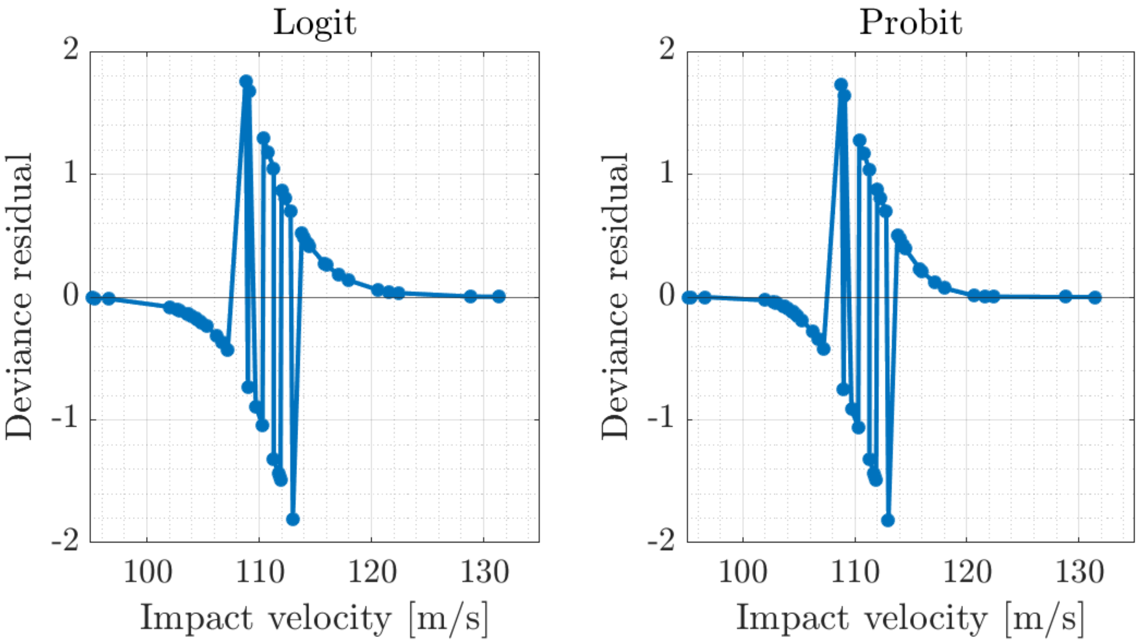


Figure 4.2 A comparison of the probit and logit deviance residuals.

4.1.3 Pseudo R squared

In standard regression techniques the overall fit of a multiple regression model is judged, for example, by a unique well–defined quantity such as R^2 computed from the fitted model. For a logit or probit regression there are several different ways of calculating so–called pseudo R^2 and, unfortunately, there is no consensus on which is best. For the sake of completeness we briefly discuss some of the pseudo R^2 measures that have been proposed. A common feature of many of the various pseudo R^2 are that they are based on comparisons of the predicted values from the fitted model to those from a reduced model. The reduced model is a fit to the data using only the intercept β_0 and fixing the slope to zero, $\beta_1 = 0$. Consequently, the pseudo R^2 only measures if the model is improved by including a dependence on the variable (velocity in our case) and as such do not directly assess the goodness of fit. We think that a better measure for goodness of fit is based strictly on a comparison of observed to expected values from the full fitted model. We will discuss such a goodness of fit measure in the next subsection.

However, the pseudo R^2 can of course provide some useful information when comparing competing models to the same set of data. Mittlböck and Schemper [6] studied the properties of 12 different pseudo R^2 using the following criteria: (i) the R^2 has an intuitive interpretation, (ii) the R^2 is bounded to the interval $[0, 1]$, (iii) the R^2 is well–behaved during the logit or probit transformation. Menard [7] performed a similar analysis, but for different pseudo R^2. Currently, there are several models that are used routinely.

4.1.3.1 McFadden

McFadden's R^2 [8] are calculated by comparing the log–likelihood of the reduced model to the full model,

$$R^2_{\text{McFadden}} = 1 - \frac{\mathcal{L}_{\text{model}}}{\mathcal{L}_{\text{reduced}}}. \tag{4.4}$$

The McFadden R^2 is bounded on the interval $[0, 1]$. The higher the value, the more likely it is that the full model outperforms the reduced model.

4.1.3.2 Cox–Snell

The Cox–Snell R^2 [9] compares the likelihood of the full model and the reduced model (McFadden's R^2 uses the log–likelihood). The Cox–Snell R^2 is defined as

$$R^2_{\text{C\&S}} = 1 - \left(\frac{L_{\text{reduced}}}{L_{\text{model}}}\right)^{2/n}. \tag{4.5}$$

It's worth noting that while the Cox–Snell R^2 is similar to McFadden's R^2, the upper limit of Cox–Snell's R^2 is not one. The upper limit can in many cases be much less than one, and is determined by the likelihood of the reduced model.

4.1.3.3 Nagelkerke

Nagelkerke's R^2 [10] can be viewed as an ”adjusted Cox–Snell R^2”, which addresses the problem of the upper limit not being equal to one. This is done by dividing the Cox–Snell R^2 by its largest possible value. Nagelkerke's R^2 is defined as

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{C\&S}}}{\max(R^2_{\text{C\&S}})} = \frac{1 - (L_{\text{reduced}}/L_{\text{model}})^{2/n}}{1 - L^{2/n}_{\text{reduced}}}. \tag{4.6}$$

4.1.3.4 Tjur

Tjur's R^2 [11] has an intuitive definition. For all of the observed 0s in the data, we calculate the mean predicted value π_0. Similarly, for all of the observed 1s in the data, we calculate the mean predicted value π_1. Tjur's R^2 is then the distance between the two means. Thus, a Tjur's R^2 value approaching one indicates that there is clear separation between the predicted values for the 0s and 1s. Additionally, Tjur's R^2 is bounded to the interval $[0, 1]$. Mathematically, Tjur's R^2 is defined as

$$R^2_{\text{Tjur}} = \pi_1 - \pi_0. \tag{4.7}$$

Here we have defined $\pi_i = 1/n_i \sum_j P(y_j = i)$ for $i = 0, 1$. The sum over j indicates that we either sum over the velocities where we obtained a perforation or the velocities where we obtained a stop. Tjur's pseudo R^2 stands out from the others, as it is not based on comparing the full model with the reduced model.

4.1.3.5 Pseudo R^2 for the logit and probit model

In Tab. 4.2 we have calculated the various pseudo R^2 we have just discussed. In all cases, the pseudo R^2 are very similar, but the probit has slightly higher values than the logit. The reason is the same as for the deviance, the logit has slightly heavier tails. The high values of Tjur's R^2 tells us that our model is excellent at distinguishing stops and perforations. The low values of the remaining pseudo R^2's might naively be interpreted as indicating a bad model fit. However, they occur because our data set mostly consists of points in the overlapping region coupled with the fact that the slope is relatively high. If we had performed more shots in the low and high velocity regime the R^2 values would become larger. This discussion illustrates the danger of only relying on one pseudo R^2 value.

Table 4.2 Pseudo R^2 values for the logit and probit model.

Model	R^2_{Tjur}	R^2_{McFadden}	$R^2_{\text{C\&S}}$	$R^2_{\text{Nagelkerke}}$
Probit	0.955	0.592	0.560	0.746
Logit	0.952	0.584	0.555	0.740

4.1.4 Classification tables

In the following the logit and probit analysis produces identical results for our dataset. The goodness of fit can be analyzed through so–called classification tables, which is a much more direct method than computing the pseudo R^2. A detailed description can be found in [12]. A Classification Table (aka a confusion matrix) compares the predicted number of successes with the number of successes actually observed and similarly the predicted number of failures compared to the number actually observed. There are four possible outcomes:

- True positive (TP): The number of cases that were correctly classified as positive. I.e our model predicted a perforation, and we obtained a perforation.

- False positive (FP): The number of cases that were incorrectly classified as positive. I.e our model predicted a perforation, but we obtained a stop.

- True negative (TN): The number of cases that were correctly classified to be negative. I.e our model predicted a stop, and we obtained a stop.

- False negative (FN): The number of cases that were incorrectly classified as negative. I.e our model predicted a stop, but we obtained a perforation.

Utilizing these definitions the classification matrix is generally defined as

$$\text{Classification matrix} = \begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix}. \tag{4.8}$$

The definition of the classification table depends on a so–called probability cutoff p_c. If the predicted probability is larger than the cutoff we expect to obtain a perforation, if it is smaller we expect the projectile not to perforate the target. The most natural choice is a cutoff probability of $p_c = 0.5$, the corresponding classification table for our perforation experiment is then

$$\text{Classification matrix} = \begin{bmatrix} 20 & 5 \\ 4 & 20 \end{bmatrix}. \tag{4.9}$$

The accuracy (inaccuracy) of our model can be defined as the number of correct (wrong) predictions,

$$\text{Model accuracy} = 1 - \text{Model inaccuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = 0.81. \tag{4.10}$$

So with a cutoff value of $p_c = 0.5$, our model gives the correct prediction in 81% of the shots.

As mentioned before the classification table, as well as the accuracy depends on the probability cutoff value p_c. While $p_c = 0.5$ is a natural choice given a large (100-1000) amount of shots, it is not necessarily the best possible choice given our limited number of shots, as is typically (20-50) in the case of perforation experiments. To investigate how our models predictive power depends on the probability cutoff it is customary to use the ROC curves (receiver operating characteristic curve). To define the ROC curves we need to introduce the following rates:

$$\begin{aligned}
\text{True positive rate (TPR)} &= \frac{\text{Number of perforations predicted correctly}}{\text{Total obtained perforations}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{True negative rate (TNR)} &= \frac{\text{Number of stops predicted correctly}}{\text{Total obtained stops}} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \\
\text{False positive rate (FPR)} &= \frac{\text{Number of perforations predicted wrongly}}{\text{Total obtained stops}} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \\
\text{False negative rate (FNR)} &= \frac{\text{Number of stops predicted wrongly}}{\text{Total obtained perforations}} = \frac{\text{FN}}{\text{TP} + \text{FN}}.
\end{aligned} \tag{4.11}$$

Since the number of shots are fixed, we also have the following summation relations,

$$\text{TPR} + \text{FNR} = \text{FPR} + \text{TNR} = 1. \tag{4.12}$$

The ROC curves are calculated by varying the probability cutoff on the interval $p_c \in [0, 1]$ and for each value calculate the rates defined in Eq. (4.11). The result of the calculation is shown in Fig. 4.3. There are several different curves which tells us similar things so we will only focus on Figs. 4.3 (a) and (g).

The easiest curve to understand is 4.3 (g) where the model accuracy is plotted as a function of the probability cutoff value. For almost the entire range of cutoff values, we see that our model predicts perforations or stops with an accuracy of $\approx 80\%$. The high–accuracy is indicative of a high goodness of fit.

Figure 4.3 (a) is what is referred to as the ROC curve in the literature. The true positive rate and false positive rate is plotted for various values of the probability cutoff. The endpoints $(0, 0)$ and $(1, 1)$ are always the same and correspond to $p_c = 1$ and $p_c = 0$ respectively. To understand the plot better it is useful to consider a simple limit case. If the ROC curve was the line $y = x$ (dashed in the figure) the rate of true positives would be equal to the rate of false positives, indicating that the model is no better than guessing the outcome randomly. If the ROC lies below $y = x$ (dashed line) the predictions of the model would be worse than guessing randomly. If the ROC lies above $y = x$ (dashed line) the predictions of the model are much better than guessing randomly, and indicative of a good model fit. This behaviour can be quantified by evaluating the area under the ROC curve, which is commonly referred to as the AUC. Unfortunately, there is no ”magic number”, only general guidelines. In general, we use the following rule of thumb

$$\text{AUC} = \begin{cases} < 0.5 & \text{Bias towards wrong predictions, worse than random guess.} \\ = 0.5 & \text{No discrimination, equivalent to random guess.} \\ (0.5, 0.7) & \text{Poor discrimination, slightly better than random guess.} \\ (0.7, 0.8) & \text{Acceptable discrimination.} \\ (0.8, 0.9) & \text{Good discrimination.} \\ (0.9, 1) & \text{Excellent discrimination.} \end{cases} \tag{4.13}$$

For our case, the ROC curve lies above the line $y = x$ (dashed line), and the area under the ROC curve is 0.95. This indicates that our model has a large true positive rate and a small false positive rate; thus it is useful for predicting the outcome of the perforation experiment.

The remaining curves in Fig. 4.3 $(b) - (f)$ have similar interpretations, so we will not go into more detail. Nevertheless, they are included for completeness to show that they indeed exhibit similar behaviour. Note that $(c) - (d)$ simply expresses that the summation law in Eq. (4.12) is satisfied and serves as a consistency check.

As always when utilizing statistics, no measure of goodness is perfect. The best strategy is to use several measures, while being aware of their limitations. For instance, the classification matrix and ROC curve, utilized here, should be treated with caution. In practice, if more shots were performed at the same plate, the AUC would likely decrease. The rate of decrease depends on whether we have obtained representative data, which again depends on the shooting method. The probability model has an upward bias. The bias arises due to the fact that the same data that were used to fit the model, was used to judge the performance of the model. The model fitted to a specific set of data is, after all, expected to perform well on the same data. The true measure of the performance of the regression model can be obtained by using the same model to predict future observations.

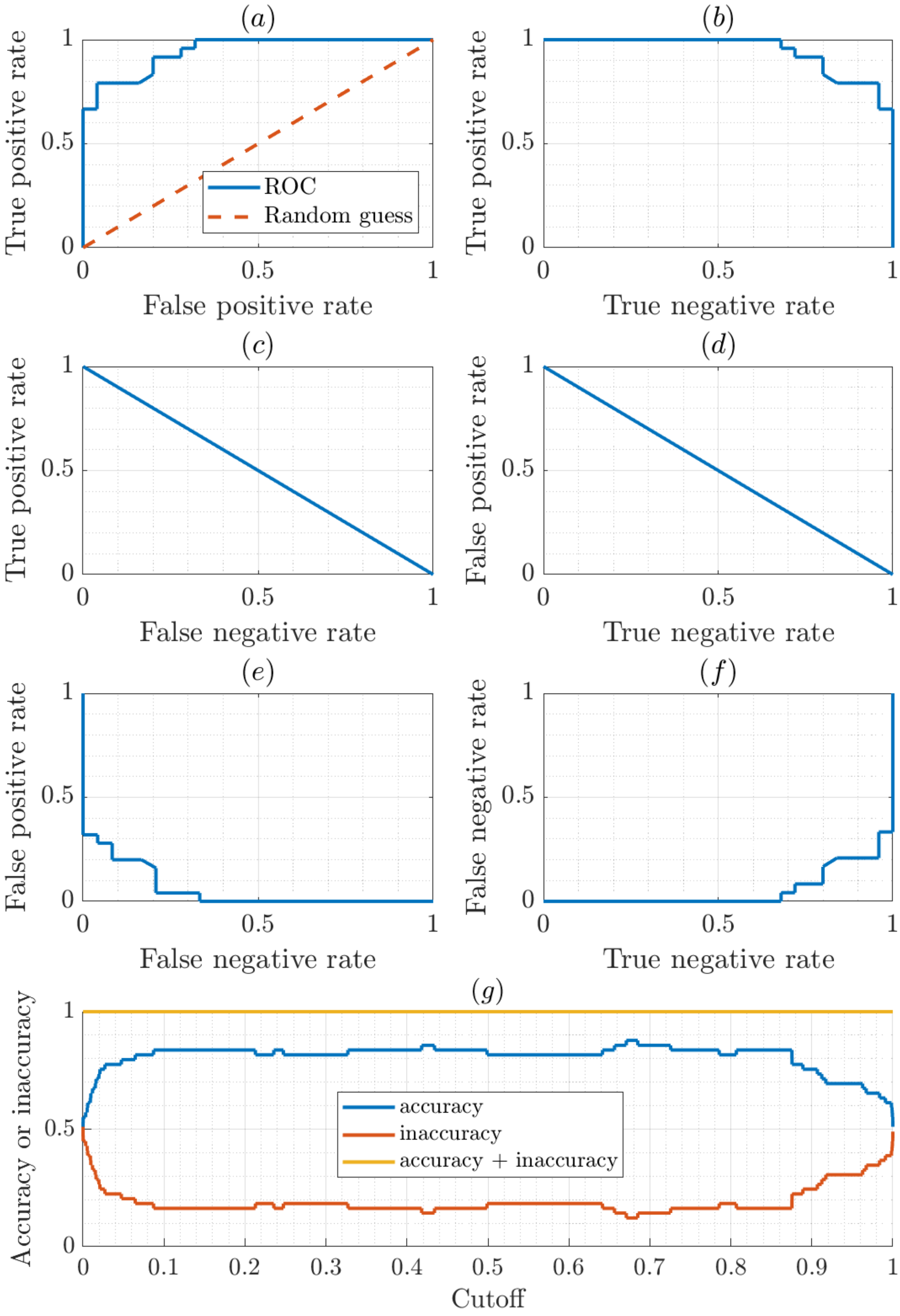


Figure 4.3 The various ROC curves indicating how well our model predicts the data for various values of the probability cutoff.

4.2 The physical perforation properties of the aluminium plate

Based on the discussion of various goodness of fit measures in the previous section, we conclude that both the probit and logit models seems to describe the perforation experiment in an accurate manner. In this section, we will apply the probit model to describe the physics of the aluminium plate and evaluate whether or not it is suitable to be used as a witness plate in perforation experiments.

An extensive literature study of the perforation properties of human skin can be found in G. R. James' thesis [13]. For our purposes the research culminates in one empirical expression for the V_{50} and an expression for the perforation probability as a function of the impact velocity V. The formulas are

$$V_{50} = 153.8\,(\gamma_\alpha \delta_a \epsilon_a \eta_a)\, S^{-0.354(\gamma_b \delta_b \kappa_b \eta_b)} \tag{4.14}$$

and

$$P_{\text{Skin perforation}} = \Phi\left\{2.95 - 2.80\left[\frac{153.8\,(\gamma_\alpha \delta_a \epsilon_a \eta_a)\, S^{-0.354(\gamma_b \delta_b \kappa_b \eta_b)}}{V}\right]\right\}, \tag{4.15}$$

where the projectiles sectional density S is its mass divided by its cross sectional area. Here the empirical constants $\{\gamma_a, \gamma_b\}$ depends on the target type, $\{\delta_a, \delta_b\}$ depends on the target location, ϵ_a depends on the backing type, $\{\eta_a, \eta_b\}$ depends on the projectile shape, and κ_b depends on the storage condition of the target. The empirical parameters are given in Tab. 4.3.

Target type	γ_a	γ_b
Child PMHS	0.898	1.208
Goat	1.053	1.103
Pig	1.226	1.029
PMHS	1.000	1.000
Sheep	0.972	1.007

(a) Target types

Target location	δ_a	δ_b
Abdomen	1.788	1.894
Back	1.225	0.813
Buttocks	0.757	0.719
Thigh	1.000	1.000
Thorax	1.256	1.413

(b) Target locations

Backing type	ϵ_a
Intact	1.000
Isolated	1.2000
Isolated and backed by cork	0.969
Isolated and backed by mipoplast	1.189

(c) Backing types

Projectile shape	η_a	η_b
Blunt	1.345	1.276
Round or pointed	1.000	1.000

(d) Projectile shape

Storage condition	κ_b
Fresh	1.000
Frozen–thawed	1.166
Refrigerated	0.798

(e) Storage condition

Table 4.3 The empirical parameters used in Eqs. (4.14) *and* (4.15). *The tables, with corresponding equations, are adapted from [13].*

In Fig. 4.4 we plot the empirical Eqs. (4.14) and (4.15) together with our statistical estimate for the V_{50} and perforation probability. Concretely, we consider: the target type to be child and adult PMHS, all target locations, intact skin, round or pointed projectile, and fresh storage condition.

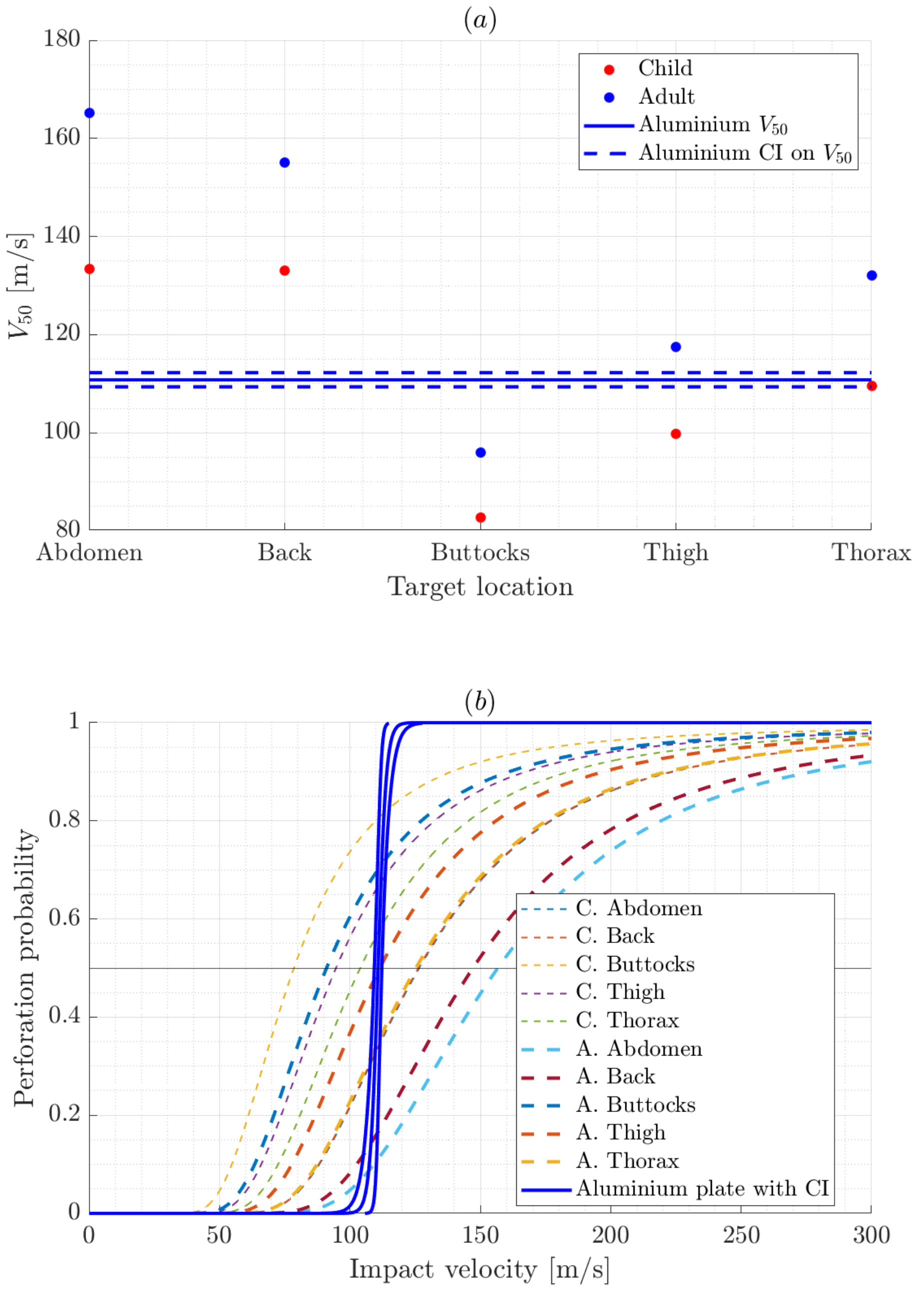


Figure 4.4 A comparison of the perforation statistics of the aluminium plate and human (Child/Adult) skin.

We observe that the V_{50} and perforation probability of the aluminium plate lies in a velocity regime representative of the considered targets. However, the slope of the aluminium plate is much steeper than for human skin. This means that the overlap region between perforations and stops is much smaller than for actual human skin. This property makes the aluminium plate a poor skin simulant, but useful as a witness plate. Concretely, the steep slope of the aluminium plate is useful because it allows a definite discrimination between perforations and stops.

Furthermore, our results indicate that it seems reasonable to assume that if a projectile perforates the aluminium plate the potential for perforation or damage of human skin is high. In fact, one might consider employing thinner aluminium with lower V_{50} as a witness plate, to make it easier to detect relatively low–probability (< 0.5) perforations in the lower velocity regime ($50 - 80\,\mathrm{m/s}$). Alternatively, another material could be used, such as the skin simulant developed [14] and utilized [15, 16] at FFI. This would make the perforation experiments even more sensitive to perforations, and therefore possibly avoid a situation where the skin is perforated, while the aluminium plate stops the projectile. An increased sensitivity seems to be of particular importance for experiments where the target is the thorax, thigh, or buttocks. That being said, in experiments testing ballistic wests for adults, where the appropriate target is the abdomen or back, the aluminium plate seems to be sufficiently sensitive.

5 Summary and conclusion

In this report we have introduced various statistical methods that are useful in analyzing perforation experiments. Concretely, we have focused on the probit and logit model to describe data, where the outcome is binary. We have discussed the fundamental properties of the probit and logit and how they are used to perform a nonlinear regression. Importantly, we have derived the confidence interval on the velocity quantiles. The various measures of goodness of fit has also been introduced, which can be used to evaluate how reliable our model actually is in its predictions.

As a concrete example of the application of the statistical methods we have considered the perforation statistics of a 0.5 mm aluminium plate. We find that that the aluminium plate experience perforations in a velocity regime which is a subset of the corresponding velocity regime for human skin. In general, we find that aluminium plate is a good witness plate, but an unrealiable skin simulant because of the steepness of its slope. More specifically, we find that the aluminum plate is a suitable witness plate for shots against the abdomen or back. For shots against the thigh, thorax, or buttocks we recommend to use an aluminum plate that is more easily perforated.

References

[1] J. C. Collins. Quantal response: Practical sensitivity training. Technical Report ARL-TR-6022, Army Research Laboratory (ARL), 2012.

[2] J. C. Collins. Quantal response: Estimation and inference. Technical Report ARL-TR-7088, Army Research Laboratory (ARL), 2014.

[3] T. H. Johnson, L. Freeman, J. Hester, and J. L. Bell. A comparison of ballistic resistance testing techniques in the department of defense. *IEEE Access*, 2:1442–1455, 2014.

[4] C.F. J. Wu and Y. Tian. Three-phase optimal design of sensitivity experiments. *Journal of Statistical Planning and Inference*, 149:1–15, 2014.

[5] D. B. Rahbek. Evaluation of 3pod - a novel procedure for ballistic testing. Technical Report 21/00981, Forsvarets forskningsinstitutt (FFI), 2021.

[6] M. Mittlbock and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15(19):1987–1997, 1996.

[7] S. Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24, 2000.

[8] D. McFadden. *Conditional Logit Analysis of Qualitative Choice Behavior*. Institute of Urban and Regional Development, University of California, 1 edition, 1973.

[9] D. R. Cox and E. J. Snell. *The analysis of binary data*. Chapman & Hall: London, 2 edition, 1989.

[10] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 09 1991.

[11] T. Tjur. Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 63(4):366–372, 2009.

[12] JR D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Inc., 3 edition, 2013.

[13] G. R. James. *Development of models to assess penetrating injury from ballistic projectiles*. PhD thesis, Cranfield university (England), 2013–2020.

[14] S. Bergsrud and M. Huseby. Frangible sikkerhet - delrapport 1 9x 9 mm. Technical Report 12/00834, Forsvarets forskningsinstitutt (FFI), 2012.

[15] M. F. Jakobsen. Comparison of 9x19 mm frangible ammunition. Technical Report 22/02061, Forsvarets forskningsinstitutt (FFI), 2022.

[16] M. F. Jakobsen. Evaluating the recommended safety distance of a recoil booster. Technical Report 22/02059, Forsvarets forskningsinstitutt (FFI), 2022.

A Derivation of the appropriate GLM models

The shape of the appropriate GLM curves can be derived from a latent variable description. We consider the case where there exists some unmeasureable variable y^* which needs to exceed some threshold value in order for a perforation to occur. The value y^* is related to the material characteristics of the armour system. For concreteness we consider the following latent variable model

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{if } y^* < 0, \end{cases} \tag{A.1}$$

where the unobservable variable depends linearly on the velocity plus an error term

$$y^* = \beta_0 + \beta_1 v + \epsilon = z + \epsilon, \tag{A.2}$$

with $z = \beta_0 + \beta_1 v$. Crucially, perhaps surprisingly, it is the distribution of the error term ϵ that determines the appropriate model. Specifically, if the error term is normally distributed, the model is a probit model. If the error term is distributed according to the logistic distribution, we end up with the logistic model. To see this, we simply express the probability of the latent variable to be bigger than 0:

$$\begin{aligned} P(y = 1|v) &= P(y^* > 0|v) \\ &= P(z + \epsilon > 0|v) \\ &= P(\epsilon > -z|v) \\ &= P(\epsilon < z|v) \\ &= F(z) \end{aligned} \tag{A.3}$$

where we in the penultimate line utilized that both the normal and logistic distribution is symmetric around the mean.

Of course it is more natural to think of an error term being normally distributed than for it to obey a logistic distribution. Nevertheless, a logistic distribution is often used to model binary response systems, because its coefficients have direct physical interpretations related to odds and that the equations are not transcendental. In practice, the logit and probit produce very similar results that often are statistically indistinguishable when employing appropriate confidence intervals. Usually this leads to that choosing between logit and probit effectively becomes a matter of taste.

B Experimental data

The raw data from the perforation experiment is given in Tab. B.1.

Table B.1 Experimental data from the perforation experiment. A perforation is indicated by 1 *and a stop is indicated by* 0.

Impact velocity [m/s]	Perforation/stop [1/0]
96.6	0
131.4	1
95.3	0
112.8	1
128.8	1
122.4	1
104.5	0
121.6	1
120.6	1
117.1	1
114.5	1
118.0	1
116.0	1
108.8	1
111.7	0
113.8	1
104.4	0
103.6	0
104.0	0
104.8	0
106.2	0
102.9	0
114.3	1
113.8	1
111.3	1
102.7	0
109.7	0
109.1	1
109.0	0
102.0	0
110.3	0
105.3	0
104.9	0
111.9	0
107.2	0
115.8	1
113.8	1
110.4	1
95.1	0
103.7	0
106.7	0
110.8	1
112.0	1
111.3	0
111.8	0
112.3	1
114.4	1
114.0	1
113.0	0

C Wald test

In logit and probit regression a Wald test can either be used to determine the siginificance of the regression coefficients or compare different experiments with each other.

C.1 Significance of coefficients

In a Wald test for determining if the obtained regression coefficients are significant the zero hypothesis and alternative hypothesis are:

- H_0: $\hat{\beta}_i = 0$,

- H_1: $\hat{\beta}_i \neq 0$,

where $i = 0, 1$. Under the null hypothesis the Wald statistic takes the form

$$W_i = \frac{\hat{\beta_i}^2}{\mathrm{Var}\hat{\beta}_i} \tag{C.1}$$

and obeys a χ_1^2–distribution with one degree of freedom. The t–statistic output by matlab is $\sqrt{W}$. The p–value (probability to obtain a coefficient larger than zero, given the null hypothesis) is determined by the CDF function:

$$\text{p–value} = 1 - \mathrm{CDF}(\chi_1^2, W). \tag{C.2}$$

In our case the p–values are very small, so we reject the null hypothesis and conclude that the coefficients are significant.

C.2 Comparing experiments

Consider that we have performed n experiments (e.g. utilized $n = 2$ different aluminium plates), and performed a probit or logit analysis to obtain the set of estimates $\{\hat{\theta}_i\}_{i=1}^n$. Each estimate can be expressed as

$$\hat{\theta}_i = [\mu_i, s_i]^T \tag{C.3}$$

and is accompanied by a covariance matrix[3]

$$\hat{V}_i = \begin{bmatrix} (V_i)_{11} & (V_i)_{12} \\ (V_i)_{21} & (V_i)_{22} \end{bmatrix}. \tag{C.4}$$

The n estimates and covariance matrices can be collected into a single vector and matrix as

$$X = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n]^T \quad \text{and} \quad V = \mathrm{diag}(V_1, V_2, \dots, V_n) \tag{C.5}$$

respectively. By multiplying the variable matrix X with the so–called contrast matrix K we can construct an hypothesis test as follows:

$$H_0 : KX = 0 \quad \text{and} \quad H_1 : KX \neq 0. \tag{C.6}$$

[3] Note that it has to be the covariance matrix relating the parameters utilized in the vector $\hat{\theta}_i$. The matlab output, which gives the covariance between the coefficients β_0 and β_1 would then need to be transformed according to the chain rule.

For instance if we wish to test if $\mu_1 = \mu_2$ we would choose the contrast matrix $K = [1, 0, -1, 0, 0, \ldots, 0]$. The relevant Wald test statistic then becomes

$$\begin{aligned} W &= (KX)^T \left(KVK^T\right)^{-1} KX \\ &\overset{H_0}{\sim} \chi_r^2 \\ &\overset{H_1}{\sim} \chi_{r,\delta}^2. \end{aligned} \tag{C.7}$$

As indicated above the Wald statistic obeys a central χ_r^2 distribution with $r = \text{rank}(K)$ degrees of freedom under the null hypothesis. Under the alternative hypothesis the Wald statistic is noncentral $\chi_{r,\delta}^2$ distributed with noncentrality parameter $\delta = (KX)^T(KVK^T)^{-1}KX \neq 0$.

Before we proceed we make a comment on what a noncentral distribution is. In general, each of the standard distributions (normal, chi square, student–t) are considered to be central distributions. Each central distribution has a noncentral cousin. The relationship between the central and noncentral distributions is parametrized through a noncentrality parameter. If the noncentrality parameter of a distribution is zero, then the distribution is identical to the corresponding central distribution. For example, when the noncentrality parameter is zero the noncentral normal, chi square, and student–t reduces to the central normal, chi square, or student–t distributions.

In general, the central distribution describes the distribution of a test statistic when the difference tested is null (so H_0 is then true). The noncentral distribution describes the test statistic when the difference tested is nonzero (so H_1 is then true.) Consequently, the central and noncentral distributions are used when calculating the p–value and test power respectively. We can now proceed with the mathematical analysis.

A type–I error means that we reject the null hypothesis H_0 even though it is true. If we say that a type–I error is equal to α, this means that there exists a critical value W_0 of the test statistic W, which satisfies

$$\alpha = P\,(W > W_0|\, H_0). \tag{C.8}$$

It follows automatically that $P\,(W < W_0|\, H_0) = 1 - \alpha$. This tells us that the critical value is $W_0 = \text{CDF}^{-1}(\chi_r^2, 1-\alpha)$. We reject the null hypothesis if the measured value $\hat{W}$ is larger than the critical value, i.e. $\hat{W} > W_0$. In an equivalent fashion, we can also define the p–value of an experiment as

$$p = P\left(W > \hat{W}|H_0\right) = 1 - \text{CDF}(\chi_r^2, \hat{W}). \tag{C.9}$$

We reject the null hypothesis if $p < \alpha$.

A type–II error, means that we keep the null hypothesis, even though it is false. If we say that the type–II error equals β, it means that the conditional probability of not rejecting H_0 is equal to β, i.e.

$$\beta = P\,(W < W_0|H_1) = \text{CDF}\left(\chi_{r,\delta}^2, W_0\right) = \text{CDF}\left[\chi_{r,\delta}^2, \text{CDF}^{-1}(\chi_r^2, 1-\alpha)\right]. \tag{C.10}$$

The power q of the test is then defined as rejecting H_0 when it is false

$$q = 1 - \beta. \tag{C.11}$$

C.2.1 Example 1: Equivalence of velocity

Let us now consider a relevant example. Assume that we want to test whether two V_{50} are statistically equivalent. The null and alternative hypothesis then takes the form

$$H_0 : m_1 = m_2 \quad \text{and} \quad H_1 : m_1 \neq m_2. \tag{C.12}$$

We then use the contrast matrix $K = [1, 0, -1, 0, \dots]$, such that $KX = m_1 - m_2$ and $KVK^T = (V_1)_{11} + (V_2)_{11}$. Since, K only has one independent column, its rank $r = \text{rank}(K) = 1$. Hence, the Wald statistic takes the form

$$\begin{aligned} W &= \frac{(m_1 - m_2)^2}{(V_1)_{11} + (V_2)_{11}} \\ &\overset{H_0}{\sim} \chi_1^2 \\ &\overset{H_1}{\sim} \chi_{1,\delta}^2 . \end{aligned} \tag{C.13}$$

The noncentrality parameter is given by

$$\delta = \frac{\Delta^2}{(V_1)_{11} + (V_2)_{11}}, \tag{C.14}$$

where $\Delta = m_1 - m_2$. If we then have δ, know the covariance matrix, and decide on α we can calculate the p–value and test power from Eqs. (C.9) and (C.11) respectively. Finally, we can then compare the p–value and α to determine whether the two V_{50} are the same.

C.2.2 Example 2: Equivalence of mean and slope

To test for whether two probits (or logits) are equivalent we form the hypothesis test

$$H_0 : \; [m_1, s_1] = [m_2, s_2] \quad \text{and} \quad H_1 : \; [m_1, s_1] \neq [m_2, s_2] \, . \tag{C.15}$$

To this end we use the contrast matrix

$$K = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & \dots \\ 0 & 1 & 0 & -1 & 0 & \dots \end{bmatrix}, \tag{C.16}$$

which has two linearly independent columns and therefore $r = \text{rank}(K) = 2$. If we define the mean and slope differences $\Delta_m = m_1 - m_2$ and $\Delta_s = s_1 - s_2$, the Wald statistic takes the form

$$\begin{aligned} W &= \frac{\Delta_s^2 \left[(V_1)_{11} - (V_2)_{11}\right] - 2\Delta_s\Delta_m \left[(V_1)_{12} - (V_2)_{12}\right] + \Delta_m^2 \left[(V_1)_{22} - (V_2)_{22}\right]}{\left[(V_1)_{11} - (V_2)_{11}\right]\left[(V_1)_{22} - (V_2)_{22}\right] - \left[(V_1)_{12} + (V_2)_{12}\right]^2} \\ &\overset{H_0}{\sim} \chi_2^2 \\ &\overset{H_1}{\sim} \chi_{1,\delta}^2 . \end{aligned} \tag{C.17}$$

The noncentrality parameter is is equal to the Wald statistic, when the differences Δ_m and Δ_s are nonzero. If we then know the necessary parameters, we can proceed as before to determine whether to discard the null hypothesis.

About FFI

The Norwegian Defence Research Establishment (FFI) was founded 11th of April 1946. It is organised as an administrative agency subordinate to the Ministry of Defence.

FFI's mission

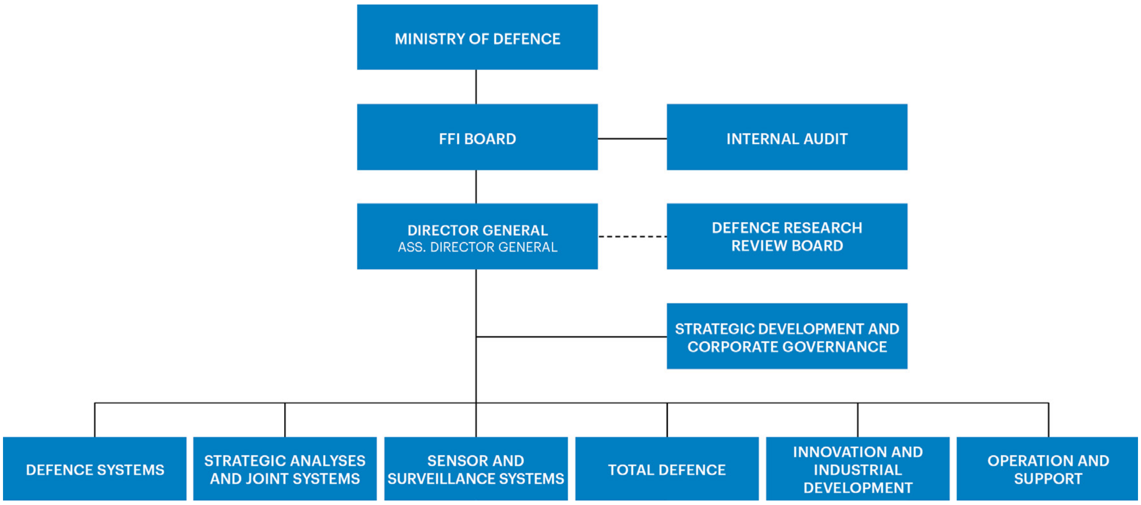
FFI is the prime institution responsible for defence related research in Norway. Its principal mission is to carry out research and development to meet the requirements of the Armed Forces. FFI has the role of chief adviser to the political and military leadership. In particular, the institute shall focus on aspects of the development in science and technology that can influence our security policy or defence planning.

FFI's vision

FFI turns knowledge and ideas into an efficient defence.

FFI's characteristics

Creative, daring, broad-minded and responsible.


MINISTRY OF DEFENCE
FFI BOARD
INTERNAL AUDIT
DIRECTOR GENERAL
ASS. DIRECTOR GENERAL
DEFENCE RESEARCH REVIEW BOARD
STRATEGIC DEVELOPMENT AND CORPORATE GOVERNANCE
DEFENCE SYSTEMS
STRATEGIC ANALYSES AND JOINT SYSTEMS
SENSOR AND SURVEILLANCE SYSTEMS
TOTAL DEFENCE
INNOVATION AND INDUSTRIAL DEVELOPMENT
OPERATION AND SUPPORT

Forsvarets forskningsinstitutt (FFI)
Postboks 25
2027 Kjeller

Besøksadresse:
Kjeller: Instituttveien 20, Kjeller
Horten: Nedre vei 16, Karljohansvern, Horten

Telefon: 91 50 30 03
E-post: post@ffi.no
ffi.no

Norwegian Defence Research Establishment (FFI)
PO box 25
NO-2027 Kjeller
NORWAY

Visitor address:
Kjeller: Instituttveien 20, Kjeller
Horten: Nedre vei 16, Karljohansvern, Horten

Telephone: +47 91 50 30 03
E-mail: post@ffi.no
ffi.no/en