Statistical Hypothesis Testing in Positive Unlabelled Data: Supplementary Material

Konstantinos Sechidis[1], Borja Calvo[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{sechidik,gavin.brown}@cs.manchester.ac.uk
[2] Department of Computer Science and Artificial Intelligence,
University of the Basque Country, Spain
borja.calvo@ehu.es

1 Proof of Lemma 1

Lemma 1. *Assuming data is selected completely at random, the conditional distribution of x given $y = 1$ is equal to the conditional distribution of x given that it is labelled.*

$$p(x|y^+) = p(x|s^+) \quad \forall \ \ x \in \mathcal{X}.$$

Proof. Using Bayes' theorem we have:

$$p(x|s^+) = \frac{p(x)p(s^+|x)}{p(s^+)} \tag{1}$$

Using the selected completely at random assumption the following expression is proved in [2, Lemma 1].

$$p(y^+|x) = \frac{p(s^+|x)}{p(s^+|y^+)} \tag{2}$$

So from eq. (1) and (2) we have:

$$\begin{aligned} p(x|s^+) = p(x)p(y^+|x)\frac{p(s^+|y^+)}{p(s^+)} = p(x, y^+)\frac{p(s^+, y^+)}{p(s^+)p(y^+)} = \\ = p(x|y^+)p(y^+|s^+) = p(x|y^+). \end{aligned}$$

where in the last step we used the property $p(y^+|s^+) = 1$, since all the labelled examples are positive.

2 Proof of Theorem 1

Theorem 1. *In the positive unlabelled scenario, under the selected completely at random assumption, a variable X is independent of the class label Y if and only if X is independent of S, so it holds $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp S$.*

Proof. To prove $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp S$, we need to prove that

$$p(x,s) = p(x)p(s) \Leftrightarrow p(x,y) = p(x)p(y) \quad \forall\ x \in \mathcal{X},\ \ y \in \mathcal{Y}\ \ and\ \ s \in \mathcal{S}.$$

Since the random variables S and Y are binary it is sufficient to prove this for the two classes. So for the first class we have

$$p(x,s^+) = p(x)p(s^+) \Leftrightarrow p(x|s^+) = p(x) \Leftrightarrow$$
$$p(x|y^+) = p(x) \Leftrightarrow p(x,y^+) = p(x)p(y^+),$$

where in the second step we have used the selected completely at random assumption as expressed in Lemma 1.

Using the above result for the first class, we will prove it also for the second class

$$p(x,s^-) = p(x)p(s^-) \Leftrightarrow p(x) - p(x,s^+) = p(x)(1-p(s^+)) \Leftrightarrow$$
$$p(x,s^+) = p(x)p(s^+) \Leftrightarrow p(x,y^+) = p(x)p(y^+) \Leftrightarrow$$
$$p(x) - p(x,y^-) = p(x)(1-p(y^-)) \Leftrightarrow p(x,y^-) = p(x)p(y^-).$$

3 Proof of Theorem 2

Theorem 2. *In the positive unlabelled scenario, under the selected completely at random assumption, when X and Y are dependent random variables ($X \not\perp\!\!\!\perp Y$) we have $I(X;Y) > I(X;S)$.*

Proof. The mutual information between variables X, Y (for categorical X and binary Y) is

$$I(X;Y) = \sum_{x\in\mathcal{X}} p(x,y^+) \ln \frac{p(x,y^+)}{p(x)p(y^+)} + \sum_{x\in\mathcal{X}} p(x,y^-) \ln \frac{p(x,y^-)}{p(x)p(y^-)}.$$

Using the equation $p(x|y^-) = \frac{p(x)-p(x|y^+)p(y^+)}{1-p(y^+)}$, this can be re-expressed as

$$I(X;Y) = \sum_{x\in\mathcal{X}} p(x|y^+)p(y^+) \ln \frac{p(x|y^+)}{p(x)}$$
$$+ \sum_{x\in\mathcal{X}} (p(x) - p(x|y^+)p(y^+)) \ln \frac{p(x) - p(x|y^+)p(y^+)}{p(x)(1-p(y^+))}.$$

In order to explore the relationship between $I(X;Y)$ and $I(X;S)$ we introduce the following function:

$$f(\tilde{p}) = \sum_{x\in\mathcal{X}} p(x|s^+)\tilde{p} \ln \frac{p(x|s^+)}{p(x)} + \sum_{x\in\mathcal{X}} (p(x) - p(x|s^+)\tilde{p}) \ln \frac{p(x) - p(x|s^+)\tilde{p}}{p(x)(1-\tilde{p})}. \quad (3)$$

When $\tilde{p} = p(y^+)$, and using Lemma 1, this is exactly $I(X;Y)$. Alternatively, when $\tilde{p} = p(s^+)$, this is exactly $I(X;S)$. So in order to explore the relationship between $I(X;Y)$ and $I(X;S)$, we should explore the monotonicity of the function $f(\tilde{p})$ with the following lemma.

Lemma 2. *Under the selected completely at random assumption f is a non-decreasing function of $\tilde{p}$, and it is strictly increasing when $X \not\perp\!\!\!\perp Y$.*

Proof (of Lemma 2). By taking the first derivative of f with respect to $\tilde{p}$ we have

$$\frac{d}{d\tilde{p}} f(\tilde{p}) = -\sum_{x \in \mathcal{X}} p(x|s^+) \ln \frac{\frac{p(x)}{p(x|s^+)} - \tilde{p}}{1-\tilde{p}}.$$

Applying Jensen's inequality to the strictly convex function $-ln(\cdot)$, we get

$$\begin{aligned}\frac{d}{d\tilde{p}} f(\tilde{p}) &\geq -\ln \sum_{x \in \mathcal{X}} p(x|s^+) \frac{\frac{p(x)}{p(x|s^+)} - \tilde{p}}{1-\tilde{p}} \\ &= -\ln \left(\frac{1}{1-\tilde{p}} \sum_{x \in \mathcal{X}} \big(p(x) - \tilde{p} p(x|s^+)\big) \right) = -\ln \frac{1-\tilde{p}}{1-\tilde{p}} = 0.\end{aligned}$$

So $f(\tilde{p})$ is a non-decreasing function of $\tilde{p}$. Furthermore we will have equality if and only if $\frac{\frac{p(x)}{p(x|s^+)} - \tilde{p}}{1-\tilde{p}}$ is constant for all $x \in \mathcal{X}$. This implies that we will have $\frac{d}{d\tilde{p}} f(\tilde{p}) = 0$ if and only if $p(x) = p(x|s^+) \quad \forall \quad x \in \mathcal{X}$, or in other words when $X \perp\!\!\!\perp Y$. So when $X \not\perp\!\!\!\perp Y$, f is strictly increasing function of $\tilde{p}$. Which finishes the proof of Lemma 2. □

Given Lemma 2, and combining it with the fact that in PU data $p(y^+) > p(s^+)$, when $X \not\perp\!\!\!\perp Y$ we get

$$f(p(y^+)) > f(p(s^+)) \Leftrightarrow I(X;Y) > I(X;S)$$

and this two quantities are equal only when $X \perp\!\!\!\perp Y$. Which finishes the proof of Theorem 2.

4 Proof of Theorem 3

Theorem 3. *The non-centrality parameter of the G-test between X and S takes the form:*

$$\lambda_{G(X;S)} = \kappa \lambda_{G(X;Y)} = \kappa 2NI(X;Y),$$

where $\kappa = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)} = \frac{1-p(y^+)}{p(y^+)} \frac{N_{S^+}}{N-N_{S^+}}$.

Proof. In order to prove that relationship we will use the result of [3] that when we assume local alternatives or contiguous alternatives the χ^2 and the G-test have the same asymptotic power [3, p. 109], in other words their non-centrality parameters converge to a common value as $N \to \infty$ [1, Section 16.3.5].

So instead of exploring the relationship of the non-centrality parameters for the G-tests between X, S and X, Y, we can explore the relationship between the

non-centrality parameters of the χ^2-tests between X, S and X, Y. [1, Section 6.6.4] presents the non-centrality parameter ($\lambda_{\chi^2(X;Y)}$) for the χ^2-test

$$\lambda_{\chi^2(X;Y)} = N \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\big(p(x,y) - p(x)p(y)\big)^2}{p(x)p(y)}$$

We will start by re-expressing that non-centrality parameter in the following way

$$\lambda_{\chi^2(X;Y)} = N \sum_{x\in\mathcal{X}} \frac{\big(p(x,y^+) - p(x)p(y^+)\big)^2}{p(x)p(y^+)} + N \sum_{x\in\mathcal{X}} \frac{\big(p(x,y^-) - p(x)p(y^-)\big)^2}{p(x)p(y^-)}$$

$$\lambda_{\chi^2(X;Y)} = N \sum_{x\in\mathcal{X}} \frac{\big(p(x,y^+) - p(x)p(y^+)\big)^2}{p(x)p(y^+)} + N \sum_{x\in\mathcal{X}} \frac{\big(p(x) - p(x,y^+) - p(x)(1-p(y^+))\big)^2}{p(x)(1-p(y^+))}$$

$$\lambda_{\chi^2(X;Y)} = N \sum_{x\in\mathcal{X}} \frac{\big(p(x,y^+) - p(x)p(y^+)\big)^2}{p(x)p(y^+)} + N \sum_{x\in\mathcal{X}} \frac{\big(p(x,y^+) - p(x)p(y^+)\big)^2}{p(x)(1-p(y^+))}$$

$$\lambda_{\chi^2(X;Y)} = p(y^+)N \sum_{x\in\mathcal{X}} \frac{\big(p(x|y^+) - p(x)\big)^2}{p(x)} + \frac{p(y^+)^2}{1-p(y^+)} N \sum_{x\in\mathcal{X}} \frac{\big(p(x|y^+) - p(x)\big)^2}{p(x)}$$

$$\lambda_{\chi^2(X;Y)} = \frac{p(y^+)}{1-p(y^+)} N \sum_{x\in\mathcal{X}} \frac{\big(p(x|y^+) - p(x)\big)^2}{p(x)} \tag{4}$$

Following exactly the same procedure for the non-centrality parameter of the χ^2-test between X and S we get

$$\lambda_{\chi^2(X;S)} = \frac{p(s^+)}{1-p(s^+)} N \sum_{x\in\mathcal{X}} \frac{\big(p(x|s^+) - p(x)\big)^2}{p(x)}$$

Using Lemma 1 this parameter is written as

$$\lambda_{\chi^2(X;S)} = \frac{p(s^+)}{1-p(s^+)} N \sum_{x\in\mathcal{X}} \frac{\big(p(x|y^+) - p(x)\big)^2}{p(x)} \tag{5}$$

So by combining (4) and (5) we get the expression

$$\lambda_{\chi^2(X;S)} = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)} \lambda_{\chi^2(X;Y)}$$

By using the result that the non-centrality parameters for the χ^2 and G-test converge to a common value, we can re-write the above relationship using the non-centrality parameter of the G-test

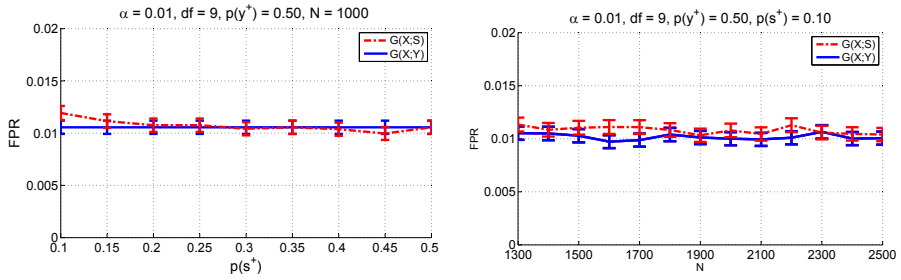
$$\lambda_{G(X;S)} = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)} \lambda_{G(X;Y)} = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)} 2NI(X;Y).$$

so by representing the factor as $\kappa = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)}$ we get

$$\lambda_{G(X;S)} = \kappa \lambda_{G(X;Y)} = \kappa 2NI(X;Y).$$

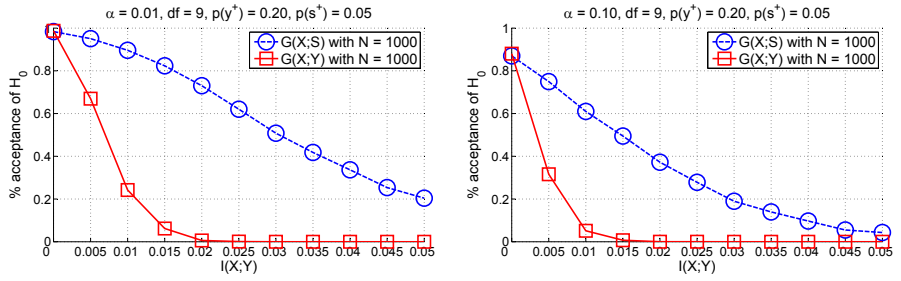
5 Figures for features with $|\mathcal{X}| = 10$

In order to generate the random variables we followed the same procedure as the one presented in the paper. Firstly we generate a random sample a random sample $\mathbf{y} = \{y_{i=1}, .., y_N\}$, where each $y_i \in \{0, 1\}$ and $p(y = 1) = 0.2$. Then we create the sample $\mathbf{x}_{i=1}^N$ as follows: when $y_{i=0}$ we choose uniformly for $y_{x=1}$ an integer random value between 1 to $|\mathcal{X}|/2$, while when $y_{i=1}$ we choose uniformly an integer random value between $|\mathcal{X}|/2+1$ to $|\mathcal{X}|$.We then corrupt this dependency from by picking a random fraction of the examples, and setting a new value for each selected x_i by drawing a uniformly and integer random variable between 1 to $|\mathcal{X}|$. It is clear that by varying the number of examples which are corrupted by noise, we generate random variables with different mutual informations. By taking a large sample estimate, $N = 1,000,000$, we can for example determine that when we corrupt 60% of the examples, the resultant variables have mutual information $I(X;Y) \approx 0.053$.



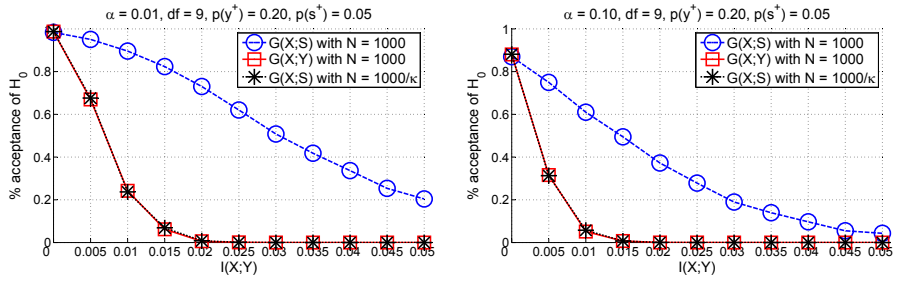
(a) (b)

Fig. 1: Figure for Type-I error. (a) Type-I error changing as a function of the probability of an example being labeled $p(s^+)$, for fixed $\alpha = 0.01$, $N = 1000$. (b) Type-I error changing as a function of N, for fixed $\alpha = 0.01$, $p(s^+) = 0.10$.



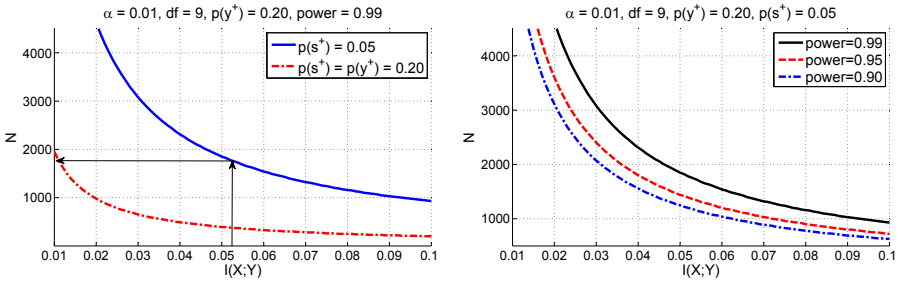
(a) (b)

Fig. 2: Figure for comparing the Type-II error of the two tests using (a) $\alpha = 0.01$ and (b) $\alpha = 0.10$.



(a) (b)

Fig. 3: Figure for comparing the Type-II error of the two tests and the G-test between X and S with corrected sample size , using (a) $\alpha = 0.01$. (b) $\alpha = 0.10$.



(a) (b)

Fig. 4: Figures for sample size determination. (a) Contrasting classical power analysis with PU power analysis to determine the minimum sample size. Arrows show that with 5% supervision ($p(s^+) = 0.05$), we need $N \geq 1743$ examples to achieve the desired power in order to observe a supervised effect $I(X;Y) = 0.053$. (b) Sample size determination under the PU constraint. Given a required statistical power, this illustrates the minimum total number of examples (N) needed, assuming we can only label 5% of the instances. For example, if we wish to detect a mutual information as low as 0.04, we need $N \geq 2310$ to have a power of 99%.

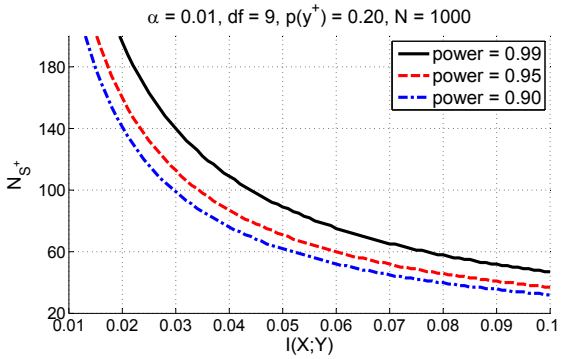
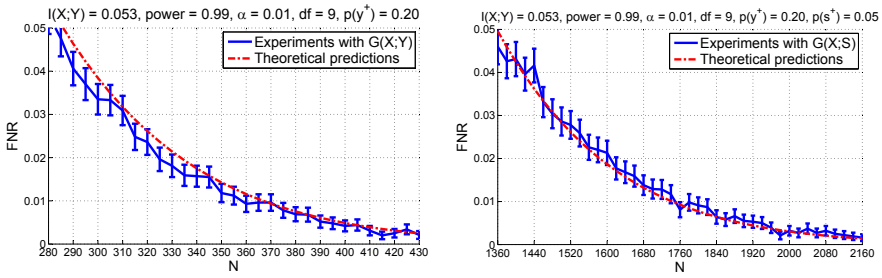


Fig. 5: Determining the required number of labelled examples. This illustrates the required number of labelled examples (N_{S^+}), assuming $N = 1000$. For example, to detect a mutual information dependency as low as 0.02, in order to have a power of 95%, we need labels for 160 examples, which means that we need to label at 80% of the positive examples.



(a) (b)

Fig. 6: Figures for False Negative Rate. (a) Full supervision, when the true mutual information is $I(X;Y) = 0.053$. This verifies the theoretical prediction from Fig. 4a, that the minimum sample size to achieve 99% power is 367. (b) Supervision level $p(s^+) = 0.05$, supporting the predictions of Figure 4a, that the minimum sample size to achieve 99% power is 1743.

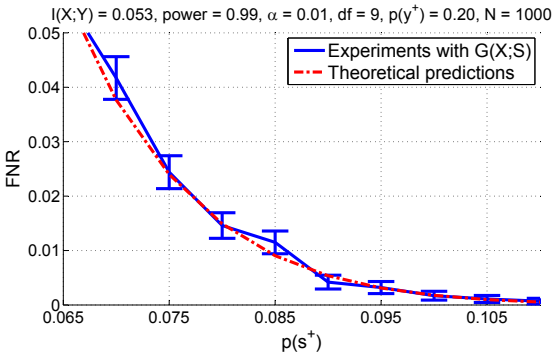


Fig. 7: FNR for varying levels of supervision in the PU constraint, with required power 99%, verifying Figure 5 (solid line), which predicted we would need $p(s^+) \geq 0.085 \Leftrightarrow N_{s^+} \geq 85$ to get $FNR < 0.01$.

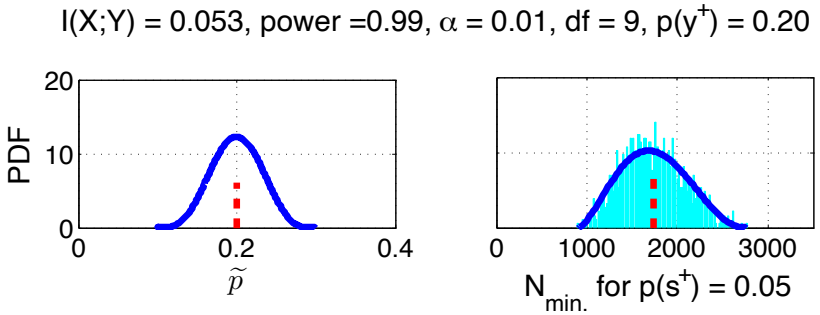


Fig. 8: Sample size determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of $p(y^+)$. The dashed line shows the *true* (but unknown) value in the data. RIGHT: The resultant uncertainty in the required sample size when we have only 5% of the examples being labeled.

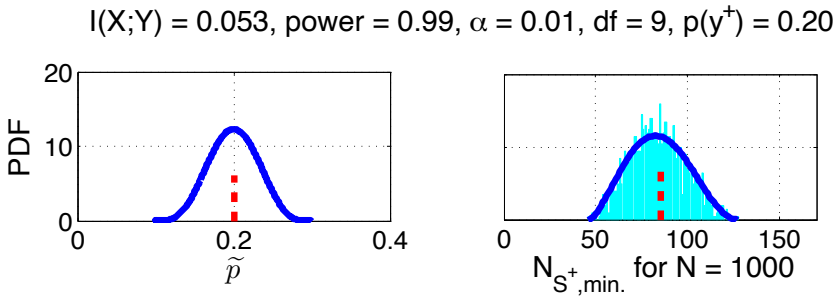


Fig. 9: Supervision determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of $p(y^+)$. The dashed line shows the *true* (but unknown) value in the data. RIGHT: The resultant uncertainty in the minimum number of required labeled examples when we have only $N = 1000$. The dashed line indicates the the true value with no uncertainty in $p(y^+)$.

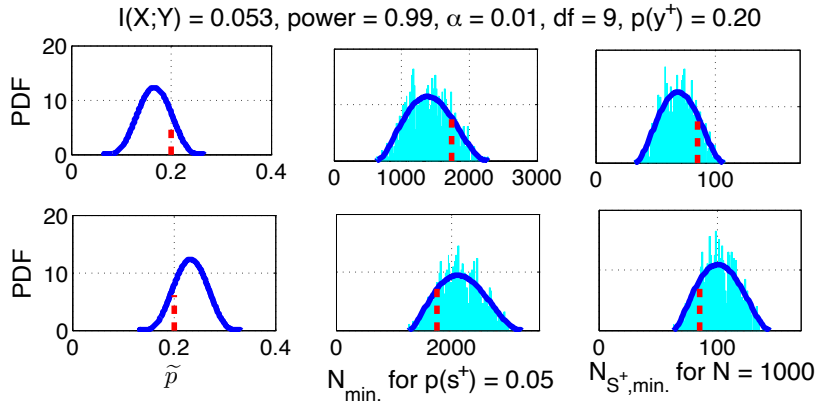


Fig. 10: A-priori power analysis under uncertain prior knowledge, when we underestimate (first row) and overestimate (second row) the prior.

6 Tables for features with $|\mathcal{X}| = 10$

Table 1: Sample size required for $df = 9$ and $\alpha = 0.01$.

(a) Traditional

Power	Effect sizes: Small	Medium	Large
0.70	1830	204	74
0.80	2143	239	86
0.90	2613	291	105
0.95	3031	337	122
0.99	3890	433	156

(b) PU with $p(y^+) = 0.20$, $p(s^+) = 0.05$

Power	Effect sizes: Small	Medium	Large
0.70	8691	966	348
0.80	10179	1131	408
0.90	12409	1379	497
0.95	14395	1600	576
0.99	18474	2053	739

Table 2: Labelled positive examples required for a PU test with $p(y^+) = 0.20$, $N = 5000$, $df = 9$ and $\alpha = 0.01$.

Power	Effect sizes: Small	Medium	Large
0.70	420	51	19
0.80	484	59	22
0.90	578	72	26
0.95	658	83	31
0.99	814	106	39

References

1. A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 3rd edition, 2013.
2. C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, New York, New York, USA, 2008. ACM.
3. S.J. Haberman. *The Analysis of Frequency Data*. Midway reprints. University of Chicago Press, 1974.