CATEGORICAL DATA ANALYSIS, 5 credits (intermediate studies), 3.9.–22.10.2015. Literature: Alan Agresti. An Introduction to Categorical Data Analysis, 2. edition. Lecturer: Pekka Pere.

Suggested solutions for the 1st set of exercises

1.

a) Let the number of different arrangements be N. If the a objects could be differentiated from one another they could be arranged in $k!$ different orders and there would be $N \times k!$ different arrangements. If in addition the b were distinguishable, too, they could be arranged in $(n-k)!$ different ways and there would be $N \times k! \times (n-k)!$ arrangements altogether. In that case all the objects would be identifiable and the number of arrangements would be $n!$. Thus it must be the case that

$$N \times k! \times (n-k)! = n!$$

or

$$N = \frac{n!}{k! \times (n-k)!} = \binom{n}{k}.$$

b) By similar reasoning as above the equation

$$N \times n_1! \times n_2! \times \ldots \times n_k! = n!,$$

is obtained. It follows that

$$N = \frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}.$$

2.

a) Let Y be a Bernoulli distributed random variate with success probability π. Suppose that n independent identical Bernoulli experiments (each with probability π) are carried out. The outcome is y realised successes and $n-y$ failures. For example the first three experiments turned out as successes, the next experiment was a failure followed by two successes and two failures so that the total count of successes is y. The probability of the observed sequence is

$$\pi\pi\pi(1-\pi) \times \cdots \times \pi(1-\pi) = \pi^y(1-\pi)^{n-y}.$$

The formulation on the right is obtained by combining terms suitably. The formulation on the right applies regardless of the order of appearances given y.

There are $\binom{n}{y}$ alternative sequences of y successes and $n-y$ failures (exercise 1 a)). Each sequence is a disjoint (combined) event. Probability of disjoint events is the sum of the probabilities of them:

$$\mathsf{P}(Y=y) = \pi^y(1-\pi)^{n-y} + \cdots + \pi^y(1-\pi)^{n-y} = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

b) If a multinomial distribution applies then the probability of each sequence is correspondingly

$$\pi_1^{n_1}\pi_2^{n_2}\dots\pi_c^{n_c}$$

(using the notation of the book). There are

$$\frac{n!}{n_1!n_2!\dots n_c!}$$

alternative sequences (exercise 1 b)). They are disjoint, so the probability for n_1 observations in class 1, n_2 observations in class 2 *etc.*, is

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \dots, N_c = n_c) = \frac{n!}{n_1!n_2!\dots n_c!}\pi_1^{n_1}\pi_2^{n_2}\dots\pi_c^{n_c}$$

($\sum_{i=1}^{c} N_i = n$).

3.

a) Random variates Y_i $(i = 1, 2, \dots, n)$ are independent and follow the Bernoulli distribution with parameter π. The expected value is simple to calculate:

$$\mathsf{E}(Y_i) = \pi \times 1 + (1-\pi) \times 0 = \pi + 0 = \pi \equiv \mu$$

(for all i).

The variance is the expected value of squared deviations from the mean derived above:

$$\begin{aligned}
\mathsf{var}(Y_i) &= \mathsf{E}\left[(Y_i - \mu)^2\right] \\
&= \pi \times (1-\pi)^2 + (1-\pi) \times (0-\pi)^2 \\
&= \pi \times (1 - 2\pi + \pi^2) + \pi^2 - \pi^3 \\
&= \pi - \pi^2 \\
&= \pi(1-\pi).
\end{aligned}$$

Alternatively the variance can be obtained by Steiner's rule

$$\mathsf{var}(Y_i) = \mathsf{E}\left(Y_i^2\right) - \mu^2.$$

It is derived below:

$$\begin{aligned}
\mathsf{var}(Y_i) = \mathsf{E}\left[(Y_i - \mu)^2\right] &= \mathsf{E}\left(Y_i^2 - 2Y_i\mu + \mu^2\right) \\
&= \mathsf{E}\left(Y_i^2\right) - \mathsf{E}\left(2Y_i\mu\right) + \mathsf{E}\left(\mu^2\right) \\
&= \mathsf{E}\left(Y_i^2\right) - 2\mu\mathsf{E}(Y_i) + \mathsf{E}\left(\mu^2\right) \\
&= \mathsf{E}\left(Y_i^2\right) - 2\mu^2 + \mu^2 \\
&= \mathsf{E}\left(Y_i^2\right) - \mu^2.
\end{aligned}$$

The alternative derivation of the variance of Y_i is ($\mu = \pi$ is substituted):

$$\begin{aligned}\mathsf{var}(Y_i) &= \mathsf{E}(Y_i^2) - \mu^2 \\ &= \pi \times 1^2 + (1-\pi) \times 0^2 - \pi^2 \\ &= \pi - \pi^2 \\ &= \pi(1-\pi).\end{aligned}$$

b) The results above enable derivation of the mean and variance of $P = \sum_{i=1}^{n} Y_i/n$:

$$\begin{aligned}\mathsf{E}(P) &= \mathsf{E}\left(\sum_{i=1}^{n} \frac{Y_i}{n}\right) = \sum_{i=1}^{n} \mathsf{E}\left(\frac{Y_i}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \pi \\ &= \frac{1}{n} \times n \times \pi = \pi\end{aligned}$$

and

$$\begin{aligned}\mathsf{var}(P) &= \mathsf{var}\left(\sum_{i=1}^{n} \frac{Y_i}{n}\right) = \sum_{i=1}^{n} \mathsf{var}\left(\frac{Y_i}{n}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{var}(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \pi(1-\pi) \\ &= \frac{1}{n^2} \times n \times \pi \times (1-\pi) = \frac{\pi(1-\pi)}{n}.\end{aligned}$$

The second equality in the calculation of the variance follows from the independence of the observations.

P is an unbiased estimator of π because $\mathsf{E}(P) = \pi$.

c) Random variate P can be expressed as follows:

$$P = \sum_{i=1}^{n} \frac{Y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Random variates Y_i are independent and identically distributed so the assumptions of the Central limit theorem apply. As calculated in a), $\mathsf{E}(Y_i) = \pi$ and $\mathsf{var}(Y_i) = \pi(1-\pi)$. Thus P follows approximately the Normal distribution $\mathsf{N}(\pi, \pi(1-\pi)/n)$:

$$P \overset{a}{\sim} \mathsf{N}(\pi, \pi(1-\pi)/n).$$

Above "$\overset{a}{\sim}$" stands for "follows in large samples" or "follows asymptotically".

4. The interpretation below is based on Figures 1–3 in A. Buse (1982): The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *American Statistician*, 36, 153–157. The standard assumptions, which guarantee the usual asymptotic properties of likelihood based statistics, are assumed.

The likelihood ratio test statistic (LR) is based on the vertical distance between $l(\hat{\theta})$ and $l(\theta_0)$, the Wald statistic is based on the horizontal distance between $\hat{\theta}$ and θ_0, and the score statistic (S) is based on the slope of the likelihood function evaluated at the null value of $l(\theta)$ or θ_0 (Figures 1–3). The larger the distances or the absolute value of the slope the larger the statistics tend to be. The Wald and score statistics depend explicitly on the Fisher information for θ.

The Wald and score statistics kind of try to assess the difference $l(\hat{\theta}) - l(\theta_0)$ by exploiting knowledge of the difference $\hat{\theta} - \theta_0$ or of the slope $l'(\theta_0)$ and the observed information

$$j(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}$$

or the curvature around the peak of $l(\theta)$:

- The magnitude of the difference $l(\hat{\theta}) - l(\theta_0)$ cannot be reasoned from the difference $\hat{\theta} - \theta_0$ alone. The former difference can increase even though the latter difference remains unchanged if the peak of $l(\theta)$ steepens (Figure 2). In the close relative

  $$\sqrt{j(\hat{\theta})}(\hat{\theta} - \theta_0)$$

  of the Wald statistic based on the observed information the difference $\hat{\theta} - \theta_0$ is multiplied by $\sqrt{j(\hat{\theta})}$. The statistic increases if the peak steepens or the observed information and the LR statistic increase. In large samples $j(\hat{\theta}) \approx i(\hat{\theta})$ and the close relative statistic and the Wald statistic merge.

- The magnitude of the difference $l(\hat{\theta}) - l(\theta_0)$ cannot be assessed by the slope $l'(\theta_0)$ alone either (Figure 3). A sharp peak or large observed information $j(\theta_0)$ at the null value θ_0 lessens the value of

  $$\frac{l'(\theta_0)}{\sqrt{j(\theta_0)}}$$

  which is a close relative of the score statistic. The difference $l(\hat{\theta}) - l(\theta_0)$ must be smaller and the peak must be closer to θ_0 the sharper the peak. Under the null for larges samples $i(\theta_0) \approx j(\theta_0)$ and hence the score and its relative statistic are essentially the same.

Extra comments. The following presupposes that the model parameter $\boldsymbol{\theta}$ is vector valued (multidimensional).

The LR statistic is often regarded as the most difficult to compute because one has to evaluate the likelihood function at both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ to calculate it. Cases exist where the computation is laborious or leads to theoretical problems. Examples: If the hypothesis

is nonlinear then the Wald statistic may depend on the parameterization chosen. For example denote the odds ratio by θ. Then the Wald statistic takes in general a different value if the null hypothesis is $\theta = 1$ or is $\log \theta = 0$ despite that the null hypotheses are equivalent. In principle for a given data any value of the Wald test statistic and any test outcome is possible by a suitable choice of parametrization.[1] Agresti (2013, pages 174–175)[2] addresses a case in point in the context of logistic regression. If the parameter under test is not identified under the null hypothesis then the score statistic is problematic to evaluate.

The facts below are from R.F. Engle (1984): Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Z. Griliches and M.D. Intriligator: *Handbook of Econometrics II*, 775–826. North-Holland. Amsterdam.

- The asymptotic χ^2 distribution of the LR test statistic was apparently derived first by Wilks 1938.[3]
- The Wald test dates to 1943.[4]
- Rao suggested the score test 1948. The test is often called — in the econometric literature especially — the Lagrange multiplier test. The name derives from a way of deriving the test statistic by solving a restricted maximisation problem by the method of Lagrange multipliers (Aitcheson and Silvey 1958 and Silvey 1959).[5]
- If $l(\boldsymbol{\theta})$ is a quadratic function (with respect to $\boldsymbol{\theta}$) then $W = LR = S$.
- Under the null hypothesis $\boldsymbol{\theta}$ can be approximated (in standard situations) in the neighbourhood of θ_0 by a quadratic function (a Taylor approximation). This is the intuition for the common asymptotic distribution of the three test statistics under the null hypothesis. (In the case under study in the exercise $\chi^2(1)$ or the χ^2 distribution with 1 degrees of freedom.)

---

[1] E.g. A.W. Gregory and M.R. Veall (1985): Formulating Wald Tests of Nonlinear Restrictions. *Econometrica*, 53, 1465–1468. T.S. Breusch and P. Schmidt (1988): Alternative Forms of the Wald Test: How Long Is a Piece of String?. *Communications in Statistics – Theory and Methods*, 17, 2789–2795. F. Lafontaine and K.J. White (1986): Obtaining Any Wald Statistic You Want. *Economics Letters*, 21, 35–40. M.D. Dagenais and J.-M. Dufour (1991): Invariance, Nonlinear Models, and Asymptotic Tests. *Econometrica*, 59, 1601–1615. T.R. Fears, J. Benichou and M.H. Gail (1996): A Reminder of the Fallibility of the Wald Statistic. *The American Statistician*, 50, 226–227. G.C.R. Kemp (2001): Invariance and the Wald Test. *Journal of Econometrics*, 104, 209–217. N.K. Dastoor (2008): A Simple Explanation for the Non-invariance of a Wald Statistic to a Reformulation of a Null Hypothesis. *Economics Bulletin*, 3, 1–10. D.D. Boos and L.A. Stefanski (2013): *Essential Statistical Inference. Theory and Methods*. Springer. New York. Section 3.2.8.

[2] A. Agresti (2013): *Categorical Data Analysis, 3rd edition*. CUP. Hoboken, NJ.

[3] S.S. Wilks (1938): The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of Mathematical Statistics*, 9, 60–62.

[4] A. Wald (1943): Tests of Statistical Hypothesis Concerning Several Parameters when the Number of Observations is Large. *Transactions of the Mathematical Society*, 54, 426–482.

[5] C.R. Rao (1948): Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 150–157. J. Aitcheson and S.D. Silvey (1958): Maximum Likelihood Estimation of Parameters Subject to Restraints. *Annals of Mathematical Statistics*, 29, 813–828. D.S. Silvey (1959): The Lagrangean Multiplier Test. *Annals of Mathematical Statistics*, 30, 389–407.

- If the model is the classical linear regression then the likelihood function for $\boldsymbol{\theta}$ is quadratic if the variance of the error term is known. The three test statistics then equal.
- If the error variance of the classical linear regression model needs to be estimated (by the method of maximum likelihood) then the distributions of the three test statistics depend monotonically on the *F* distribution. Hence exact critical values are (in principle) definable and the associated tests based on proper exact critical values yield identical inferences. If the asymptotic distribution is used as the reference distribution then the test outcomes may differ in finite samples. (The test statistics and their distributions are not the same.)
- In the case of the classical linear regression model — including models in which the errors are autocorrelated — it is always the case that

  $$W \geq LR \geq S.$$

  The composer of the exercise adds: Compare the above inequalities with the geometrical interpretation of the Wald and score statistics as approximations of the LR statistic!

5.

a) The probability mass function of a binomially distributed random variable is

$$\binom{n}{k}\pi^y(1-\pi)^{n-y}.$$

The coefficient of the product $\pi^y(1-\pi)^{n-y}$ is irrelevant from the point of view of maximising of the function with respect to π. The likelihood function can hence be defined as

$$\pi^y(1-\pi)^{n-y}.$$

The logarithm of it is

$$l(\pi) = y\log(\pi) + (n-y)\log(1-\pi).$$

b) The first derivative of the log-likelihood function is

$$l'(\pi) = \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y - y\pi - n\pi + \pi y}{\pi(1-\pi)} = \frac{y - n\pi}{\pi(1-\pi)}.$$

c) The maximum likelihood estimator (MLE) is obtained by differentiating the log-likelihood function with respect to π and by solving the root of the equation obtained by setting the derivative equal to zero:

$$\frac{y - n\pi}{\pi(1-\pi)} = 0 \Leftrightarrow y - n\pi = 0 \Leftrightarrow \pi = \frac{y}{n} = n^{-1}\sum_{i=1}^{n} y_i.$$

The MLE of π is

$$\hat{\pi} = p = n^{-1} \sum_{i=1}^{n} y_i.$$

d) The observed information is

$$-\frac{\partial^2 l(\pi)}{\partial \pi^2} = -\frac{\partial}{\partial \pi} \left[\frac{y}{\pi} - \frac{n-y}{1-\pi} \right] = \frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2}.$$

The expected information is

$$\begin{aligned} i(\pi) &= \mathsf{E}[-\frac{\partial^2 l(\pi)}{\partial \pi^2}] \\ &= \mathsf{E}\left[\frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2} \right] \\ &= \frac{n\pi}{\pi^2} + \frac{n-n\pi}{(1-\pi)^2} \\ &= \frac{n}{\pi(1-\pi)}. \end{aligned}$$

The mean $\mathsf{E}(y) = n\pi$ of a binomial random variate is substituted above. The inverse of the expected information is

$$[i(\pi)]^{-1} = \frac{\pi(1-\pi)}{n}.$$

It matches with the variance of $\hat{\pi}$ derived earlier.