WORKED EXAMPLES 6

INTRODUCTION TO STATISTICAL METHODS

EXAMPLE 1: ONE SAMPLE TESTS

The following data represent the change (in ml) in the amount of Carbon monoxide transfer (an indicator of improved lung function) in smokers with chickenpox over a one week period:

33, 2, 24, 17, 4, 1, -6

Is there evidence of significant improvement in lung function ?

(a) if the data are normally distributed with $\sigma = 10$

(b) if the data are normally distributed with σ unknown

Use a significance level of $\alpha = 0.05$.

SOLUTION: Here we have a sample of size 10 with sample mean $\overline{x} = 10.71$. We want to test

$$\begin{aligned} H_0 &: \mu = 0.0 \\ H_1 &: \mu \neq 0.0 \end{aligned}$$

under the assumption that the data follow a Normal distribution with $\sigma = 10.0$ **known**. Then, we have, in the Z-test,

$$z = \frac{10.71 - 0.0}{10.0/\sqrt{7}} = 2.83$$

which lies in the **critical region**, as the **critical values** for this test are ± 1.96. Therefore we evidence to **reject** H_0. The p-value is given by

$$p = 2\Phi(-2.83) = 0.004 < \alpha$$

(b) The sample variance is $s^2 = 14.19^2$. In the T-test, we have test statistic t given by

$$t = \frac{\overline{x} - 0.0}{s/\sqrt{n}} = \frac{10.71 - 0.0}{14.19/\sqrt{7}} = 2.00$$

The upper critical value C_R is obtained by solving

$$F_{St(n-1)}(C_R) = 0.975$$

where $F_{St(n-1)}$ is the cdf of a Student-t distribution with $n-1$ degrees of freedom; here $n = 7$, so we can use statistical tables or a computer to find that $C_R = 2.447$, and note that, as Student-t distributions are symmetric the lower critical value is $-C_R$.

Thus t lies **between** the critical values, and **not** in the critical region. Therefore we have **no evidence to reject** H_0. The p-value is given by

$$p = 2F_{St(n-1)}(-2.00) = 0.09 > \alpha.$$

EXAMPLE 2: TWO SAMPLE TESTS

The efficacy of a treatment for hypertension (high blood pressure) is to be studied using a small clinical trial. Thirty-eight hypertensive patients were randomly allocated to either Group 0 (placebo control) or Group 1 (treatment) and a three-month follow-up study was carried out. At the end of the study, the difference in blood pressure was measured for patients in each group was recorded. A summary of the results are presented below:

Group	n	$\overline{x}$	s^2
0	21	-0.208	4.101^2
1	17	3.953	4.630^2

Is there evidence of significant improvement in the treatment group ? Use a significance level of $\alpha = 0.05$.

SOLUTION: We will assume that the two data sets are random samples from two normal models with the **same** (unknown) variance, σ^2, that is

$$X_1, \ldots, X_{21} \sim N(\mu_0, \sigma^2)$$

$$Y_1, \ldots, Y_{17} \sim N(\mu_1, \sigma^2)$$

Here we want to test

$$\begin{aligned} H_0 &: \mu_0 = \mu_1 \\ H_1 &: \mu_0 \neq \mu_1 \end{aligned}$$

In the two-sample T-test, we have test statistic t given by

$$t = \frac{\overline{y} - \overline{x}}{s_P \sqrt{\dfrac{1}{n_0} + \dfrac{1}{n_1}}}$$

where s_P^2 is the **pooled estimate** of common variance given by

$$s_P^2 = \frac{(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2}{n_0 + n_1 - 2} = \frac{20 \times 4.101^2 + 16 \times 4.630^2}{36} = 4.344.$$

Thus the test statistic t is given by

$$t = \frac{3.953 - (-0.208)}{4.344\sqrt{\dfrac{1}{21} + \dfrac{1}{17}}} = 2.071$$

The upper critical value C_R is obtained by solving

$$F_{St(n-1)}(C_R) = 0.975$$

where $F_{St(n-1)}$ is the cdf of a Student-t distribution with $n0 + n1 - 2$ degrees of freedom; here $n0 + n1 - 2 = 36$, so we can find that $C_R = 2.028$, and the lower critical value is $-C_R$.

Thus, in this case, t lies in the critical region. Therefore we have **evidence to reject** H_0. The p-value is given by

$$p = 2F_{St(n-1)}(-2.00) = 0.046 < \alpha.$$

However, note here that the result is only just significant, thus before stronger conclusions are claimed, the data should be investigated further.

MAXIMUM LIKELIHOOD ESTIMATION

Suppose a sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted λ, so that

$$f_X(x;\theta) \equiv f_X(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0, 1, 2, ...$$

for some $\lambda > 0$. To estimate λ by maximum likelihood, proceed as follows.

STEP 1 Calculate the likelihood function $L(\lambda)$ for $\lambda \in \Theta = \mathbb{R}^+$

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \right\} = \frac{\lambda^{x_1+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

STEP 2 Calculate the log-likelihood $\log L(\lambda)$.

$$\log L(\lambda) = \sum_{i=1}^{n} x_i \log\lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

STEP 3 Differentiate $\log L(\lambda)$ with respect to λ, and equate the derivative to zero to find the m.l.e..

$$\frac{d}{d\lambda}\{\log L(\lambda)\} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - n = 0 \Rightarrow \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Thus the **maximum likelihood estimate** of λ is $\widehat{\lambda} = \bar{x}$

STEP 4 Check that the second derivative of $\log L(\lambda)$ with respect to λ is negative at $\lambda = \widehat{\lambda}$.

$$\frac{d^2}{d\lambda^2}\{\log L(\lambda)\} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \quad \text{at } \lambda = \widehat{\lambda}$$

EXAMPLE 3. Consider the following Accident Statistics Data that record the counts of the number of accidents in each of 647 households during a one year period. The Poisson distribution model is deemed appropriate for these count data.

We wish to estimate accident rate parameter λ. We have $n = 647$ observations as follows for the frequency with which a given number of accidents occurred in a given time period:

Number of accidents	0	1	2	3	4	5
Frequency	447	132	42	21	3	2

so that the estimate of λ if a Poisson model is assumed is

$$\widehat{\lambda}_{ML} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(447 \times 0) + (132 \times 1) + (42 \times 2) + (21 \times 3) + (3 \times 4) + (2 \times 5)}{647} = 0.465$$

A plot of $\log L(\lambda)$, with the maximum value and ordinate identified, is depicted below: