Foundations of Medical Statistics

Statistical Inference 8: Normal errors

Aims

The aim of this session is to introduce the theory for inference in Normal errors models.

Objectives

At the end of this session you should understand the basis of the F and t-distributions, and the use of the t-distribution for comparing means in independent groups. You should also understand the central role of the χ^2 distribution and its relation to other distributions.

8.1 Two-parameter contexts

In Inference 9 and 10 multi-parameter contexts are considered in more detail. Here we generalise the likelihood function to a simple two-parameter context.

Suppose data $\underline{x} = x_1, \ldots, x_n$ are independent observations generated from a model with probability function f defined in terms of two parameters, θ and ϕ. The likelihood and log-likelihood functions are then:

$$L(\theta, \phi|\underline{x}) = \prod_{i=1}^{n} f(x_i|\theta, \phi) \quad \text{and} \quad l(\theta, \phi|\underline{x}) = \sum_{i=1}^{n} \log f(x_i|\theta, \phi)$$

The parameter MLEs can then be obtained by setting the **partial** derivatives to zero

$$\frac{\partial l}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial l}{\partial \phi} = 0$$

and solving the two simultaneous equations.

Note that when a two-parameter situation arises from two *independent* datasets, $\underline{x_1}$ and $\underline{x_2}$, each generated independently from probability functions $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$ respectively, we describe the likelihood as the **joint** likelihood, conditional on jointly observing both datasets. Suppose the sample sizes are respectively n and m:

$$L\left(\theta_1, \theta_2|\, \underline{x_1}, \underline{x_2}\right) = \prod_{i=1}^{n} f_1(x_{1i}|\theta_1) \cdot \prod_{j=1}^{m} f_2(x_{2i}|\theta_2)$$

The joint log-likelihood will then be the sum of the two separate log-likelihoods:

$$l\left(\theta_1, \theta_2|\, \underline{x_1}, \underline{x_2}\right) = \sum_{i=1}^{n} \log\big(f_1(x_{1i}|\theta_1)\big) + \sum_{j=1}^{m} \log\big(f_2(x_{2i}|\theta_2)\big)$$

The MLEs will then of course be the same as if the two datasets were considered separately, since the contribution from the second dataset will vanish on partial differentiation with respect to the first parameter, and vice-versa. In other words, when partially differentiating with respect to θ_1, we will get no contribution from the $\sum_{j=1}^{m} \log\big(f_2(x_{2i}|\theta_2)\big)$ terms, since these depend only on θ_2.

The invariance property of MLEs extends to the multi-parameter context (see Casella & Berger, p320,321). In some contexts the variance of a function of MLEs is also a simple function of the variance of the individual MLEs (Inference 10 discusses this in more detail).

8.2 Normal context when μ and σ^2 are unknown: estimates for σ^2

In Inference 2 we derived estimators of the population variance (one unbiased, one asymptotically unbiased) from the intuitive idea of the average squared deviation around the mean. Here we will use the maximum likelihood approach.

When $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and both parameters are unknown, the log-likelihood is (see for example Inference 6, Question 1 solution):

$$l\big(\mu, \sigma^2|\underline{x}\big) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

We know that differentiating with respect to μ and setting the derivative to zero gives $\hat{\mu} = \bar{\text{x}}$. Now let us differentiate with respect to σ^2:

$$\frac{\partial l}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

So substituting $\mu = \hat{\mu}$ gives

$$\frac{\partial l}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and setting to zero then gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

However, this is biased, though since it is an MLE it is asymptotically unbiased (see Inference 2). We saw in section 2.4 that when we simply substitute the sample mean in the formula for V_μ (the average squared deviation around the unknown population mean), we ignore the variability of the sample mean around the unknown mean, and obtain a biased estimate. In section 2.4 we obtained the unbiased variance estimator (V_{n-1}) by partitioning the variability of the observations around the unknown mean into their

variability around the sample mean, and that of the sample mean around the unknown mean.

In a somewhat similar way, when we substitute $\mu = \hat{\mu} = \bar{x}$ into the expression for the MLE $\hat{\sigma}^2$, we are in effect assuming that the sample mean is the true mean: in determining the likelihood, we are in effect ignoring the probability of observing the sample mean, given the true mean. In a similar manner to 2.4 we can correct for this by 'partitioning' the probability of observing the data, conditional on unknown μ and σ^2, into i) the probability of observing the data conditional on the observed sample mean and unknown σ^2, and ii) the probability of observing the sample mean conditional on the two unknown parameters:

$$\text{Prob}\left(\underline{x}|\mu,\sigma^2\right) = \text{Prob}\left(\underline{x}|\bar{X}=\bar{x},\sigma^2\right) \times \text{Prob}(\bar{x}|\mu,\sigma^2)$$

$$\Rightarrow \quad \text{Prob}\left(\underline{x}|\bar{X}=\bar{x},\sigma^2\right) = \frac{\text{Prob}\left(\underline{x}|\mu,\sigma^2\right)}{\text{Prob}(\bar{x}|\mu,\sigma^2)}$$

This of course is a conditional probability. You have met the idea of a conditional distribution in Probability: if X and Y have a joint distribution given by probability function $f(x,y)$, then the conditional distribution of $X|Y=y$ is given by

$$f(x|Y=y) = \frac{f(x,y)}{f(y)}$$

where $f(y)$ is the marginal distribution of Y (i.e. the distribution without restricting the possible values of X).

It turns out that the MLE of σ^2 obtained from the likelihood based on the conditional distribution $\text{Prob}\left(\underline{x}|\bar{X}=\bar{x},\sigma^2\right)$ is unbiased. Theoretically this is because it takes full account of the probability of observing the sample mean (or, from another perspective, takes account of the degrees of freedom). This particular setting forms the basis of the idea of REML [restricted (or residual) maximum likelihood; see eg Garthwaite, Jolliffe & Jones, p59], which is used in more advanced contexts (discussed in Analysis of Hierarchical and Other Dependent Data, Term 2; see also Appendix). But note that this approach is only possible when $\text{Prob}\left(\underline{x}|\bar{X}=\bar{x},\sigma^2\right)$ is free of μ, requiring us to condition on a sufficient statistic for μ, which is not possible in general. We return to this concept in Inference 9.

We can derive the MLE of σ^2 using the conditional distribution as follows:

Denoting the Normal density function as $\phi(x)$,

$$\phi\left(\underline{x}|\bar{X}=\bar{x},\sigma^2\right) = \frac{\phi\left(\underline{x}|\mu,\sigma^2\right)}{\phi(\bar{x}|\mu,\sigma^2)} = \frac{\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{1}{2\sigma^2/n}(\bar{x}-\mu)^2\right)}$$

Taking logs of the above expression gives us the conditional log-likelihood:

$$l(\sigma^2|\underline{x},\bar{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{2}\log\left(\frac{2\pi\sigma^2}{n}\right) + \frac{1}{2\sigma^2/n}(\bar{x}-\mu)^2$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{2}\log\left(\frac{2\pi}{n}\right) + \frac{1}{2}\log(\sigma^2) + \frac{n}{2\sigma^2}(\bar{x}-\mu)^2$$

Neglecting terms not in μ or σ^2:

$$= -\frac{(n-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right)$$

Now let us use the partition from Inference 2:

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(\bar{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

Rearranging:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2$$

Therefore:

$$l(\sigma^2|\underline{x},\bar{x}) = -\frac{(n-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Note that the above is now free of μ.

And now, differentiating our conditional log-likelihood with respect to σ^2:

$$l'(\sigma^2|\underline{x},\bar{x}) = -\frac{(n-1)}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

And setting this equal to zero:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

This is the unbiased estimate of σ^2 that we have seen before, with the loss of one degree of freedom since we have conditioned on the sample mean, rather than the true mean.

8.3 The *F* and *t* distributions

The F distribution is defined as the ratio of two independent χ^2 variables, scaled by their degrees of freedom:

$$F_{p_1,p_2} = \frac{X^2_{p_1}/p_1}{X^2_{p_2}/p_2}$$

where F_{p_1,p_2} is a random variable with an F distribution with p_1, p_2 degrees of freedom, and $X^2_{p_i}$ $(i = 1, 2)$ are random variables that follow $\chi^2_{p_i}$ distributions.
Now consider $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and testing the hypotheses:

$$H_0: \mu = \mu_0 \text{ vs } H_1: \mu > \mu_0$$

Recall from Inference 6 that there is a uniformly most powerful test based on $\bar{x}$, the sample mean.

When σ^2 is **known**, we know the distribution of the appropriate test statistic below, since we know the distribution of the sample mean:

$$H_0 \Rightarrow \left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}\right) \sim N(0,1)$$

or equivalently,

$$\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}\right)^2 = X^2_1 \qquad (1)$$

where again, X^2_1 is a random variable distributed according to a χ^2_1 distribution.

Now suppose σ^2 is **unknown**, and *needs to be estimated from the data*. It would seem sensible to use the unbiased estimator derived from the conditional log-likelihood in 8.2:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

However, if we just substitute the square root of this estimator in the left hand side of equation (1) above, to give

$$\left(\frac{\bar{Y} - \mu_0}{S/\sqrt{n}}\right)^2$$

we are ignoring the sampling error of the variance estimator. We therefore underestimate the 'total' uncertainty by treating the variance estimate as the true variance. We need a way to incorporate the uncertainty around the variance into the 'total' uncertainty for (1) when the variance is unknown. The following strategy achieves this.

Recall from Inference 2.5 that

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$$

$$\Rightarrow \frac{S^2}{\sigma^2} = \frac{X_{n-1}^2}{n-1} \quad (2)$$

Now if we divide expression (1) by expression (2) we obtain an expression of the form:

$$\frac{X_1^2}{X_{n-1}^2/(n-1)}$$

which is distributed as an $F_{1,n-1}$.

$$\Rightarrow \frac{\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right)^2}{S^2/\sigma^2} = \frac{(\bar{Y}-\mu_0)^2}{S^2/n} \sim F_{1,n-1}$$

Note that now the unknown variance parameter has been replaced by its estimator, but in such a way that we have incorporated the uncertainty present in both (1) and (2). Thus if we define the statistic

$$T = \frac{\bar{Y}-\mu_0}{S/\sqrt{n}}$$

Then $H_0 \Rightarrow T^2 \sim F_{1,n-1}$, or $H_0 \Rightarrow T \sim \sqrt{F_{1,n-1}}$.

This can be used as the *definition* of the t distribution: so $t_{n-1} \equiv \sqrt{F_{1,n-1}}$.

Note

- Although σ^2 is **unknown**, by dividing (1)/(2) we eliminate the unknown parameter. We get an expression in terms of the sample estimators for both μ and σ^2 with a **known distribution**.
- In the case where σ^2 is known, we saw in Practical 7 that the test statistic derived in example 6.5.1 is equivalent to a log-likelihood ratio test. When σ^2 is unknown, the test statistic can also be derived from a log-likelihood ratio statistic, but in this case a more general multi-parameter form of this statistic must be used. Since the t-statistic may be derived more easily using the method above, the multi-parameter derivation is not included, but may be found in Hogg and Craig p321.

8.4 Comparison of means in two independent groups

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_2, \sigma^2)$. Note that σ^2 is the same in the two groups. We want to test the hypotheses:

$$H_0: \mu_1 = \mu_2 \text{ vs } H_1: \mu_1 > \mu_2$$

An unbiased estimator for σ^2 may be derived from a conditional log-likelihood, as in section 8.2, but now conditioning on both $\bar{x}$ and $\bar{y}$. This estimator is:

$$\hat{\sigma}^2 = S_p^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}\left(Y_j-\bar{Y}\right)^2}{n+m-2} \quad (3)$$

From Inference 2.5 (and also equation (2) above), we have independently,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1} \quad \text{and} \quad \frac{1}{\sigma^2}\sum_{j=1}^{m}\left(Y_j-\bar{Y}\right)^2 \sim \chi^2_{m-1}$$

$$\Rightarrow \quad \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}\left(Y_j-\bar{Y}\right)^2\right) \sim \chi^2_{n+m-2}$$

Substituting in $(n+m-2)S_p^2$ from equation (3):

$$(n+m-2)\frac{S_p^2}{\sigma^2} \sim \chi^2_{n+m-2} \tag{4}$$

Since $\bar{X} \sim N(\mu_1, \sigma^2/n)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/m)$, and the two samples are independent, then under the null hypothesis:

$$H_0: \mu_1 = \mu_2 \Rightarrow \bar{X}-\bar{Y} \sim N\left(0, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$$

$$\Rightarrow \quad \frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim N(0,1) \quad \Rightarrow \frac{(\bar{X}-\bar{Y})^2}{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)} \sim \chi^2_1 \tag{5}$$

Now dividing (5) by (4) gives

$$H_0: \mu_1 = \mu_2 \Rightarrow \frac{(\bar{X}-\bar{Y})^2}{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)} \cdot \frac{\sigma^2}{S_p^2(n+m-2)} = \frac{X_1^2/1}{X^2_{n+m-2}}$$

$$\Rightarrow \quad T^2 = \frac{(\bar{X}-\bar{Y})^2}{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)} = \frac{X_1^2/1}{X^2_{n+m-2}/(n+m-2)} \sim F_{1,n+m-2}$$

or, equivalently,

$$H_0 \Rightarrow T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

which is the standard two-sample t-test, pooled variance version.

Note:
Both the test in this section (two-sample) and the previous section (one-sample) can be derived directly from the best test identified by the log-likelihood ratio statistic, though this requires more algebra. The important point is that the t-test (when σ^2 is unknown), like the Z test (when σ^2 is known), is a **likelihood ratio test**. Note also that the two-sample derivation above requires σ^2 to be the same in the two groups; hence the assumption of the standard t-test (though a variant of the t-test relaxes this assumption: see Analytical Techniques 5).

8.5 Summary of relationships between statistical distributions

The χ^2 distribution is a basic link between the commonly used distributions in statistics. In the summary below, the distribution names are used as informal shorthand to denote random variables distributed accordingly:

$$N(0,1)^2 = \chi_1^2$$

$$\chi_k^2 = \sum_{i=1}^{k} \chi_1^2$$

(provided the summed random variables are independent)

$$F_{k,n} = \frac{\chi_k^2/k}{\chi_n^2/n}$$

(provided the two random variables are independent)

$$t_n^2 = F_{1,n} = \frac{\chi_1^2/1}{\chi_n^2/n}$$

(again, assuming independent random variables).

Appendix (*non-examinable*)

The `.regress` command in Stata uses the correct degrees of freedom to estimate the residual variance in a linear regression model.

When the regression model is the 'null' model, with no predictors, the residual variance is simply the variance around the sample mean. You will see that in this case the residual variance is the same as the standardly obtained sample variance, obtained with denominator $n-1$ (called Vn_1 below).

However, some commands in Stata that perform regressions use the simple maximum likelihood estimation, without using the 'smarter' likelihood conditioning on the sample mean.
Such commands - or the options to use simple `MLE` with them - leads, in the null model, to the variance estimator with denominator n (called Vn below).

```
. clear
. set obs 20
. set seed 1234
. gen y=10*invnorm(uniform()) /*giving Stata some simulated sample data, n=20 */
. summ y,d

Mean              = -2.146366
Variance  (Vn_1)  =   79.6555
```

We also calculate Vn:

```
Vn  = 75.672727
.
```

Now the null model using the familiar regression command:

```
. regress y

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  0,    19) =    0.00
```

```
       Model |           0     0           .           Prob > F      =       .
    Residual |  1513.45455    19  79.6555026           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  1513.45455    19  79.6555026           Root MSE      =   8.925

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -2.146366   1.995689    -1.08   0.296    -6.323391     2.03066
------------------------------------------------------------------------------

. dis e(rmse)^2
79.655503
```

Above is Vn_1, as expected.
Here below is the same null model, but implemented in a regression command designed for more complex data and more complex regressions. It is using simple MLE:

```
.
. sem (y <-)

Log likelihood     = -71.642949
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  y <-       |
       _cons |  -2.146366   1.945157    -1.10   0.270    -5.958804    1.666072
-------------+----------------------------------------------------------------
     var(e.y)|   75.67273   23.92982                      40.71607    140.6413
------------------------------------------------------------------------------
```

No problem with the MLE of the mean, but notice that for the MLE of the variance, it gives us Vn, not Vn_1.

Another regression command, below, designed again for more complex data, has both a simple MLE and a REML option. First the MLE option for the same null model:

```
. mixed y || _all:,var mle nolog nostderr

Log likelihood = -71.642949
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -2.146366   1.945157    -1.10   0.270    -5.958804    1.666072
------------------------------------------------------------------------------
  var(Residual) |   75.67273          .           .           .
------------------------------------------------------------------------------
```

Giving us, above, Vn. (Again, no problem with the MLE of the mean).

And now using the REML option:

```
. mixed y || _all: , var reml emonly nolog nostderr

Log restricted-likelihood = -70.045954
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -2.146366   5.683725    -0.38   0.706    -13.28626    8.993531
------------------------------------------------------------------------------
  var(Residual) |    79.6555          .           .           .
------------------------------------------------------------------------------
```

Now the above is giving us Vn_1.