Confidence Intervals for Poisson data

For an observation from a Poisson distribution, we have $\sigma^2 = \lambda$.

If we observe r events, then our estimate

$$\begin{aligned} \hat{\lambda} &= r \\ &\sim N(\lambda, \lambda) \end{aligned}$$

If r is bigger than 20, we can use this to find an approximate confidence interval.

Example: Geiger counter records 100 radioactive decays in 5 minutes. Find an approximate 95% confidence interval for the number of decays per hour.

We start by finding an interval for the number in a 5 minute period. The estimated standard deviation $s = \sqrt{100} = 10$. So the interval is

$$\begin{aligned} \hat{\lambda} \pm 1.96 \times \sqrt{\lambda} &= 100 \pm 19.6 \\ &\rightarrow \quad (80.4,\ 119.6). \end{aligned}$$

Hence, multiplying by 12, the 95% interval for the hourly rate is

$$(965,\ 1435)$$

Example: BBC news (23 February 2010).

> The homicide rate in Scotland fell last year to 99 from 115 the year before.

Is this reported change really newsworthy?

The actual number of homicides will vary from year to year. If we assume that each homicide is an independent event, then a Poisson distribution could be a reasonable model.

Consider the figures for 2008. A 95% confidence interval for the true homicide rate based on these is:

$$115 \pm 1.96\ \sqrt{115} \quad \rightarrow \quad (94.0,\ 136.0).$$

Thus the true rate could be even lower than 99 per year. It is not reasonable to conclude that there has been a reduction in the true rate.

Note: If some incidents result in more than one death, the assumption of independence will be wrong. The effect of this will be to **widen** the confidence interval.

Confidence Interval for continuous data

Example: A series of 10 water samples gave the following arsenic concentrations (μg As/l).

5.5, 6.9, 3.6, 3.5, 3.5, 9.8, 7.1, 4.7, 3.8, 5.0

The sample mean $\overline{x} = 5.34$

The sample standard deviation $s = 2.059$

We wish to find a 95% confidence interval.

We cannot assume that $s \approx \sigma$ since the sample is too small to ensure this. We allow for this by using a multiplier t which is bigger than 1.96.

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \rightarrow \bar{x} \pm t \frac{s}{\sqrt{n}}$$

The exact value of t depends on n and is given in Lindley & Scott Table 10. We take:

$$\nu = n - 1 = 9 \qquad \Rightarrow \; t = 2.262$$
$$P = 2.5\%$$

Hence, the confidence interval is:

$$\begin{aligned} \bar{x} \pm t \frac{s}{\sqrt{n}} &= 5.34 \pm 2.262 \times \frac{2.059}{\sqrt{10}} \\ &= 5.34 \pm 1.47 \\ &\rightarrow (3.87, 6.81). \end{aligned}$$

Note 1: The parameter ν (nu) that is used in looking up the value of t in Table 10 is called the "degrees of freedom".

Note 2: All of the confidence interval examples above have been of the same form. The interval has been symmetric about the estimate and the width of the interval has been a constant times a value such as $\frac{s}{\sqrt{n}}$. The latter is called the **Standard Error of the Mean**, or more generally the **Standard Error** of the estimate.

Most publications prefer to report their results as estimates and the corresponding standard errors, and assume readers can construct the appropriate confidence intervals if they require them.

Note 3: The method for finding the confidence intervals when n is small will only work well if the data come from a distribution that is roughly Normal. We can check this by plotting the data.

If the sample size is bigger than about 30, the value of t is not much bigger than 1.96. In such cases, the value 2 is often used for convenience.

Confidence Interval for a Difference

(Rees §9.11)

Suppose we have independent random samples from two different populations. We will use the random variable X_1 for observations from the first population and X_2 for the second. Suppose also that:

$$\bar{X}_1 \ \dot{\sim} N(\mu_1, \frac{\sigma^2}{n_1}) \qquad \text{and} \qquad \bar{X}_2 \ \dot{\sim} N(\mu_2, \frac{\sigma^2}{n_2})$$

where n_1 and n_2 are the sample sizes. Note that we have assumed that the two populations have different means but the **same** standard deviation. Then

$$(\bar{X}_1 - \bar{X}_2) \ \dot{\sim} N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Note: We subtract the means but we **add** the variances.

We can use the same ideas as those used to obtain a confidence interval for a single sample. If we can estimate σ then we will be able to estimate a standard error. Then the confidence interval will be the estimated difference plus or minus a suitable multiple of the standard error.

Example: A further 4 water samples were taken from the same borehole as in an earlier example. The arsenic concentrations were (μg As/l):

5.7, 6.3, 9.6, 7.0

The sample mean $\overline{x} = 7.15$

The sample standard deviation $s = 1.718$

The previous values were $\overline{x} = 5.34, s = 2.059$

How much have arsenic levels increased?

The estimate of σ comes from a weighted average of the squared standard deviations. The weights are the degrees of freedom.

$$\begin{aligned} s &= \sqrt{\frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}} \\ &= \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{(10 - 1) \times 2.059^2 + (4 - 1) \times 1.718^2}{10 + 4 - 2}} \\ &= \sqrt{3.9175} \\ &= 1.979 \end{aligned}$$

The **standard error of the difference** is:

$$\begin{aligned} s &= \sqrt{3.9175\left(\frac{1}{10}+\frac{1}{4}\right)} \\ &= \sqrt{1.3711} \\ &= 1.171 \end{aligned}$$

The sample sizes are small, so we need to use a value of t from NCST Table 10. We use the sum of the degrees of freedom for each sample.

$$\nu = \nu_1 + \nu_2 = n_1 + n_2 - 2$$

$$\nu = (n_1 - 1) + (n_2 - 1) = 12 \qquad \Rightarrow\ t = 2.179$$

$$P = 2.5$$

So the 95% confidence interval is:

$$(7.15 - 5.34) \pm 2.179 \times 1.171 \ \rightarrow\ (-0.74,\ 4.36)$$

Note that this interval includes the value zero. Thus although our estimate suggests that the true concentration has gone up by 1.81 μg/l, it is possible that it has actually decreased.

In the UK, safe levels for arsenic in water are considered to be less than 10 μg/l.

The previous example relied on the two samples being **independent**. If they are not, then the calculations are different.

Example: Sixteen matched pairs of cars, of a variety of makes and ages, were used in a test of a new fuel additive. Within each pair, one car was selected at random to use the additive and the other did not. The results, in miles per gallon were as follows:

Pair	1	2	3	4	5	6	7	8
Add.	31.8	26.0	28.7	23.1	34.8	26.7	37.3	25.7
No A.	29.6	25.4	27.4	20.7	31.1	27.4	35.2	23.1

Pair	9	10	11	12	13	14	15	16
Add.	35.2	41.6	26.9	38.7	31.4	37.1	35.1	28.2
No A.	33.3	42.0	25.8	35.8	29.2	35.4	35.1	27.3

If you do a scatter plot of the pairs, you will find that the 16 points lie close to a line through the origin, showing that the two results for a pair are similar to each other. This means that the two samples are not independent.

To allow for this lack of independence, we find the difference between the two results for each pair. It is important to keep the correct sign with these differences.

Taking the result without the additive from the value with the additive gives the following data values:

```
 2.2   0.6   1.3   2.4   3.7  -0.7   2.1   2.6
 1.9  -0.4   1.1   2.9   2.2   1.7   0.0   0.9
```

This is just a single sample of 16 values, so we can find a 95% confidence interval in the same way as before.

The mean is 1.531, the standard deviation is 1.222, so the 95% confidence interval is:

$$1.531 \pm 2.131 \times \frac{1.222}{\sqrt{16}} \rightarrow (0.880, 2.183)$$

This interval does not come close to including zero, so it is clear that the additive does improve the fuel consumption figures.

Note 1: The last two examples look similar in that there are two sets of values to be compared. However, they **must** be analysed in different ways.

For the car example, each measurement in one set tells us something about one of the values in the other set. This is because the cars in a pair were chosen to be similar to each other.

For the arsenic example, the two sets of water samples were taken at different times from the same location. There is no direct or indirect matching of samples. However, if samples had been taken at the same times from two or more different wells, allowance should be made for the lack of independence.

Note 2: For both of these examples, we found a 95% interval for the true value of the difference between the groups. In the first case, the interval included zero which implied that there might not have been a change. This does **not mean that there has been no change**. There might have been a change, but we did not have enough data.

For the second example, zero was well outside the confidence interval, so we were able to conclude that the additive had improved fuel consumption.

There is another way to consider whether a true difference could be zero. This is to assume that there is no difference and then see how likely it is to obtain the difference that was observed, or a more extreme difference. If the probability is small, the conclusion is that there really was a difference between the two groups. This is called a **hypothesis test**.

The calculations behind hypothesis tests here are essentially the same as those used in calculating a confidence interval, so the conclusions will always be the same. If there is a difference, a confidence interval is more informative because it tells us what values of the true difference are likely.

Standard Errors

We have met four types of confidence interval:

- Binomial probability;
- Poisson rate;
- Mean of continuous variable;
- Difference between means of two independent samples from a continuous variable.

Each of these has followed the same pattern:

$$\text{Estimate } \pm t \times \text{ Standard Error of estimate}$$

For a 95% interval, the value of t is 1.96 if the standard deviation is known and a little larger than this if it has to be estimated from the data.

The formula for the standard error depends on the type of estimate:

- Binomial: $\sqrt{\dfrac{p(1-p)}{n}}$
- Poisson: $\sqrt{\lambda}$
- Sample Mean: $\dfrac{\sigma}{\sqrt{n}}$

Standard Errors of Regression Estimates

If a simple regression model is adequate, it is useful to know the accuracy of the parameter estimates and of the predictions that are made from the model. We can do this by calculating standard errors and confidence intervals.

For the slope: Variance of $\hat{\beta} = \dfrac{\sigma^2}{S_{XX}}$

The residual standard deviation s is used as an estimate of σ. To calculate a confidence interval for $\hat{\beta}$, we use t with $n-2$ degrees of freedom.

Example: For the stream data.

$\hat{\beta} = 9.5636$

$s^2 = 1.2122$

$S_{XX} = 0.275$

S.E.$(\hat{\beta}) = \sqrt{\dfrac{1.2122}{0.275}} = 2.100$

For a 95% confidence interval and 9 degrees of freedom, $t = 2.262$, so C.I. for $\hat{\beta}$ is:

$$9.5636 \pm 2.262 \times 2.100 \quad \Rightarrow \quad (4.8,\ 14.3)$$

For the intercept: Variance of $\hat{\alpha} = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{XX}} \right)$

Example: For the stream data.

$\hat{\alpha} = -0.6327$

$\bar{x} = 0.55$

$$\text{S.E.}(\hat{\alpha}) = \sqrt{1.2122\left(\frac{1}{11} + \frac{0.55^2}{0.275}\right)} = 1.2015$$

A 95% confidence interval for $\hat{\alpha}$ is:

$$-0.6327 \pm 2.262 \times 1.2015 \quad \Rightarrow \quad (-3.35,\ 2.09)$$

This interval contains zero. This suggests that we could use the simpler regression model:

Flow $= \beta \times$ Depth $+ \epsilon_i$

This model makes better physical sense.

It is straightforward to fit a regression without an intercept; in the formulae above S_{XX}, S_{XY}, S_{YY} are replaced by $\sum x^2$, $\sum xy$, $\sum y^2$ respectively.

If we wish to predict the y value at some new value x, we just use the regression equation:

$\hat{y} = \hat{\alpha} + \hat{\beta}x$

Then Variance of $\hat{y} = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)$

This can be used to calculate a confidence interval. The variance of the estimate of the intercept $\hat{\alpha}$ that was given earlier is a special case of this; just set $x = 0$ in the formula.

Note: The best predictions are close to the mean of the x values. If we know in advance where predictions will be required, we should centre the x values on this.

Example: For the stream data, suppose $x = 0.75$

$\hat{y} = -0.6327 + 9.5636 \times 0.75 = 6.54$

S.E.$(\hat{y}) = 0.5327$

A 95% C.I. is (5.33, 7.75)

This interval is where we think that the **true mean** flow at a depth of 0.75m is likely to be.

We can also calculate another interval which is where we think that a **new observation** of flow at a depth of 0.75m is likely to be. This is called a **prediction interval**.

The variance that is used to calculate a prediction interval is similar to that used for a confidence interval; we just need to add σ^2. Thus:

Variance $= \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)$

Example: For the stream data at $x = 0.75$

S.E.(Prediction) $= 1.2242$

A 95% P.I. is (3.77, 9.31)

Note: In practice, Confidence Intervals are used much more often than Prediction Intervals.

Significance Tests

The calculation of confidence intervals is based on the Central Limit Theorem – we do not expect an estimate to be far from the true (unknown) value. 'Far' here is relative to the standard error of the estimate.

If we are interested in whether the true value has a particular value, such as zero, we see whether the value lies within the confidence interval.

An alternative method is to standardise by:

$$\frac{\text{Estimate} - \text{Value of Interest}}{\text{S.E.(Estimate)}}$$

This is called a **Test Statistic**.

We use the Normal distribution or t distribution tables to find the probability of getting a value as extreme as this or more extreme, either positive or negative. The result is called a **Significance Probability**.

A small probability means that it is unlikely that the value of the parameter being estimated takes the value specified. A large value means that the value is plausible.

A 95% confidence interval is just the collection of all the plausible values that have a significance probability of greater than 0.05.

Example: Stream Flow data

$\hat{\alpha} = -0.6327$ Value of interest: $\alpha = 0$

S.E.$(\hat{\alpha}) = 1.2015$

Test Statistic: $t = -0.5266$

Use the t tables with 9 degrees of freedom.

From NCST Table 9, $\text{P}(t < 0.5) = 0.6855)$ and $\text{P}(t < 0.6) = 0.7183$, so $\text{P}(t < 0.5266) \approx 0.694$

So Significance Probability $\approx 2(1 - 0.694) = 0.612$

This probability is large, so the intercept could be zero. This suggests that we could use the simpler model: Flow $= \beta \times$ Depth $+ \epsilon_i$

Note: The simpler model will lead to a different estimate of β.

The reasoning that we have used in making a decision from a significance probability is called a **Hypothesis Test**.

Formally we:

1. Have a **Null Hypothesis**.
   (e.g. True value of $\alpha = 0$)
2. Calculate a test statistic.
   $\left(\text{e.g. } \dfrac{\text{Estimate} - 0}{\text{S.E.(Estimate)}}\right)$
3. Calculate a significance probability.
4. Reject the Null Hypothesis if the significance probability is small.

Typically, there is an **Alternative Hypothesis** to the null hypothesis. The test statistic is chosen to give as good a discrimination as possible between the two hypotheses.

There are some difficulties with the hypothesis testing procedure:

Firstly, The failure of **any** of the assumptions can lead to a small Significance Probability, and not just the assumption of particular interest.
(e.g. The test above also assumes that the data are independent observations that come from a Normal distribution.)

Secondly, Acceptance of the Null Hypothesis does **NOT** mean that it is true, although users often treat it as so.

Example: Risks to public health.

Scientists will conclude that there is no evidence of any risk.

Politicians like to treat this as implying that there is no risk.

Journalists like to treat this as implying that there is a risk.

Statisticians would like to say that they can be pretty sure that the risk is less than x cases per million, but the general public is not good at interpreting this. (cf. Rail travel v. car travel.)

Statistical tests are widely (over-)used in scientific journals.
Conventionally, the significance probabilities are graded into:

$P > 0.05$	(Not Significant)
$P < 0.05$	(Statistically Significant)
$P < 0.01$	
$P < 0.001$	(Highly Significant)

Note: Statistical Significance is **not** the same as Practical Significance. If sample sizes are small, a statistical test may not detect an important effect. If sample sizes are large, a statistical test may detect an effect that is too small to be useful.