Lecture

INSTITUTE OF ACTUARIAL
& QUANTITATIVE STUDIES

Class: FY BSc

Subject : Probability and Statistics -2

Subject Code: PUSASQ 1.2

Chapter: Unit 2 Chp 1

Chapter Name: Theory of Estimation - 2

Index

Index

1 Introduction

- Previously we used the method of moments and the method of maximum likelihood to obtain estimates for the population parameter(s). For example, we might have the following numbers of claims from a certain portfolio that we receive in 100 different monthly periods:

Claims	0	1	2	3	4	5	6
Frequency	9	22	26	21	13	6	3

- Assuming a Poisson distribution with parameter μ for the number of claims in a month, our estimate of μ using the methods given in the previous chapter would be $\mu = \bar{x} = 2.37$.

- The problem is that this might not be the correct value of μ . In this chapter we look at constructing confidence intervals that have a high probability of containing the correct value. For example, a 95% confidence interval for μ means that there is a 95% probability that it contains the true value of μ.

1 Introduction

- Although point estimation is a common way in which estimates are expressed, it leaves room for many questions. For instance, it does not tell us on how much information the estimate is based, nor does it tell us anything about the possible size of the error. Thus, we might have to supplement a point estimate $\hat{\theta}$ of θ with the size of the sample and the value of var($\widehat{\Theta}$) or with some other information about the sampling
- distribution of $\widehat{\Theta}$. As we shall see, this will enable us to appraise the possible size of the error.

- Alternatively, we might use interval estimation. An interval estimate of θ is an interval of the form $\hat{\theta}_1 < \theta < \hat{\theta}_2$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of appropriate random variables $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$.

2 Confidence Intervals

- A confidence interval provides an 'interval estimate' of an unknown parameter (as opposed to a 'point estimate'). It is designed to contain the parameter's value with some stated probability. The width of the interval provides a measure of the precision accuracy of the estimator involved.

- A $\mathbf{100(1-\alpha)}$% confidence interval for $\theta$is defined by specifying random variables $\boldsymbol{\hat{\theta}_1(\underline{X})}, \boldsymbol{\hat{\theta}_2(\underline{X})}$ such that $\boldsymbol{P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 1 - \alpha}$.

- Rightly or wrongly, $\alpha = 0.05$ leading to a 95% confidence interval, is by far the most common case used in practice and we will tend to use this in most of our illustrations.

- Thus $\boldsymbol{P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 0.95}$ specifies $(\boldsymbol{\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})})$ as a 95% confidence interval for $\boldsymbol{\theta}$. This emphasises the fact that it is the interval and not $\boldsymbol{\theta}$ that is random. In the long run, 95% of the realisations of such intervals will include $\boldsymbol{\theta}$ and 5% of the realisations will not include $\boldsymbol{\theta}$.

2 Confidence Intervals

- Confidence intervals are not unique. In general they should be obtained via the sampling distribution of a good estimator, in particular the maximum likelihood estimator. Even then there is a choice between one-sided and two-sided intervals and between equal-tailed and shortest-length intervals although these are often the same, eg for sampling distributions that are symmetrical about the unknown value of the parameter.

3 Derivation of confidence intervals

The pivotal method

- There is a general method of constructing confidence intervals called the pivotal method.

- This method requires the finding of a pivotal quantity of the form $g(\underline{X}, \theta)$ with the following properties:
  i. it is a function of the sample values and the unknown parameter θ
  ii. its distribution is completely known
  iii. it is monotonic in θ.

- The distribution in condition (2) must not depend on θ. 'Monotonic' means that the function either consistently increases or decreases with θ.

3 Derivation of confidence intervals

- The equation

$$\int_{g_1}^{g_2} f(t)dt = 0.95$$

- (where $f(t)$ is the known probability (density) of $g(\underline{X}, \theta)$) defines two values, g_1 and g_2, such that

$$P\left(g_1 < g\left(\underline{X}, \theta\right) < g_2\right) = 0.95$$

- g_1 and g_2 are usually constants.
- We are assuming here that X has a continuous distribution. We will look at examples based on discrete distributions.

3 Derivation of confidence intervals

- If $\boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta})$ is monotonic increasing in $\boldsymbol{\theta}$, then:
  - $\boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta}) < \boldsymbol{g}_2 \Leftrightarrow \boldsymbol{\theta} < \boldsymbol{\theta}_2$ for some number $\boldsymbol{\theta}_2$
  - $\boldsymbol{g}_1 < \boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta}) \Leftrightarrow \boldsymbol{\theta}_1 < \boldsymbol{\theta}$ for some number $\boldsymbol{\theta}_1$

- and if $\boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta})$ is monotonic decreasing in $\boldsymbol{\theta}$,then:
  - $\boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta}) < \boldsymbol{g}_2 \Leftrightarrow \boldsymbol{\theta}_1 < \boldsymbol{\theta}$
  - $\boldsymbol{g}_1 < \boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta}) \Leftrightarrow \boldsymbol{\theta} < \boldsymbol{\theta}_2$
- resulting in $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ being a 95% confidence interval for $\boldsymbol{\theta}$.

- Fortunately in most practical situations such quantities $\boldsymbol{g}(\underline{\boldsymbol{X}}, \boldsymbol{\theta})$ do exist, although an approximation to the method is needed for the binomial and Poisson cases.

3 Derivation of confidence intervals

- In sampling from a $\boldsymbol{N(\mu,\sigma^2)}$ distribution with known value of $\boldsymbol{\sigma^2}$, a pivotal quantity is:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

- which is N(0,1).

- For example, given a random sample of size 20 from the normal population $\boldsymbol{N(\mu,10^2)}$ which yields a sample mean of 62.75, an equal-tailed 95% confidence interval for μ is:

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = 62.75 \pm 1.96\frac{10}{\sqrt{20}} = 62.75 \pm 4.38$$

- This is a symmetrical confidence interval since it is of the form $\theta \pm \beta$. For symmetrical confidence intervals, we can write down the interval using the ' $\pm$ ' notation, where the two values indicate the upper and lower limits. Alternatively, we can write this confidence interval in the form (58.37,67.13) . Here we are using the pivotal quantity $\frac{\bar{X}-\mu}{10/\sqrt{20}}$, which has a N(0,1) distribution, irrespective of the value of μ.
- The normal mean illustration shows that confidence intervals are not unique.
- Another 95% interval, with unequal tails, is $\left(\bar{X} - 1.8808\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.0537\frac{\sigma}{\sqrt{n}}\right)$.
- However, there would not be much reason to use this one in practice.

4 Confidence limits

- The 95% confidence interval $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is often expressed as:

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

- This is quite informative as it gives the point estimator $\bar{X}$ together with the indication of its accuracy. However, this cannot always be done so simply using a confidence interval.

- Also one-sided confidence intervals correspond to specifying an upper or lower confidence limit only.

5 Sample size

- A very common question asked of a statistician is:
  - 'How large a sample is needed?'

- This question cannot be answered without further information, namely:
  i. the accuracy of estimation required
  ii. an indication of the size of the population standard deviation σ.

- The latter information may not readily be available, in which case a small pilot sample may be needed or a rough guess based on previous studies in similar populations.

- As a consequence of the Central Limit Theorem, a confidence interval that is derived from a large sample will tend to be narrower than the corresponding interval derived from a small sample, since the variation in the observed values will tend to 'average out' as the sample size is increased. Market research companies often need to be confident that their results are accurate to within a given margin (eg $\pm 3\%$). In order to do this, they will need to estimate how big a sample is required in order to obtain a narrow enough confidence interval.

6 Confidence intervals for the normal distribution

The mean

- The previous section dealt with confidence intervals for a normal mean μ in the case where the standard deviation σ was known. In practice this is unlikely to be the case and so we need a different pivotal quantity for the realistic case when σ is unknown.

- Fortunately there is a similar pivotal quantity readily available and that is the t result:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

- where S is the sample standard deviation.

- The resulting confidence interval, in the form of symmetrical 95% confidence limits, is:

$$\bar{X} \pm t_{0.025,n-1}\frac{S}{\sqrt{n}}$$

- $t_{0.025,n-1}$ is used to denote the upper 2.5% point of the t distribution with $n-1$ degrees of freedom, and is defined by:
- $P\left(t_{n-1} > t_{0.025,n-1}\right) = 0.025$

- For example, from the Tables $t_{0.025,10}$, is equal to 2.228.

6 Confidence intervals for the normal distribution

The mean

- This is a small sample confidence interval for μ . For large samples t_{n-1} becomes like N(0,1) and the Central Limit Theorem justifies the resulting interval without the requirement that the population is normal.

- The normality of the population is an important assumption for the validity of the *t* interval especially when the sample size is very small, for example, in single figures. However the *t* interval is quite robust against departures from normality especially as the sample size increases. Normality can be checked by inspecting a diagram, such as a dotplot, of the data. This can also be used to identify substantial skewness or outliers which may invalidate the analysis.

6 Confidence intervals for the normal distribution

The variance

- For the estimation of a normal variance σ^2 , there is again a pivotal quantity readily available:

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

- The resulting 95% confidence interval for the variance σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)S^2}{\chi^2_{0.975,n-1}}\right)$$

- or for the standard deviation σ^2:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{0.025,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{0.975,n-1}}}\right)$$

- Note: Due to the skewness of the χ^2 distribution, these confidence intervals are not symmetrical about the point estimator S^2 , and are also not the shortest-length intervals. So we can't write these using the ' $\pm$ ' notation.
- The above intervals require the normality assumption for the population but are considered fairly robust against departures from normality for reasonable sample sizes.

7 Confidence intervals for binomial & Poisson parameters

- Both these situations involve a discrete distribution which introduces the difficulty of probabilities not being exactly 0.95, and so 'at least 0.95' is used instead. Also when not using the large-sample normal approximations, the pivotal quantity method must be adjusted.

- One approach is to use a quantity h(X) whose distribution involves θ such that:

$$\boldsymbol{P\left(h_1(\theta) < h(\underline{X}) < h_2(\theta)\right) \geq 0.95}$$

- Then if both $h_1(\theta)$ and $h_2(\theta)$ are monotonic increasing (or both decreasing), the inequalities can be inverted to obtain a confidence interval as before.

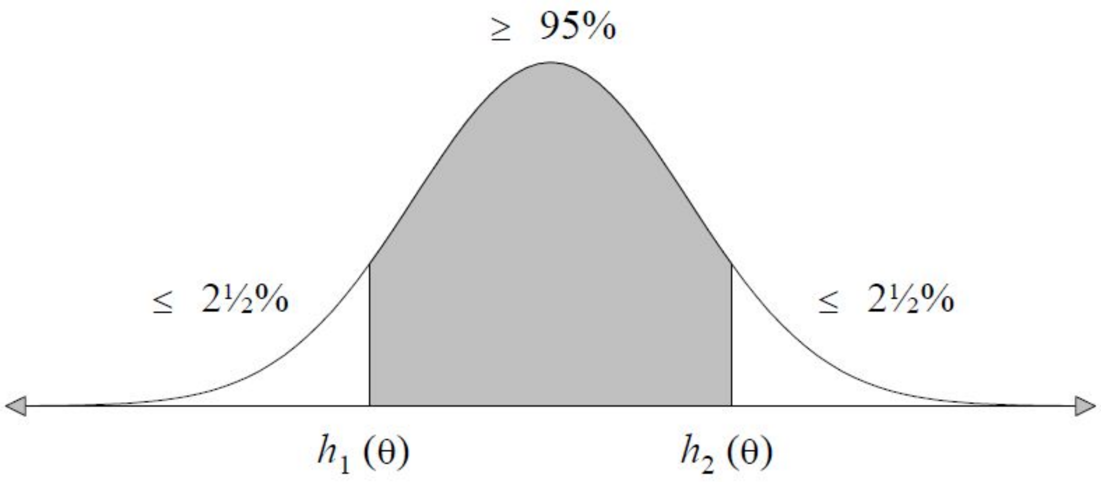
8 Binomial Distribution

- If X is a single observation from Bin(n,θ) , the maximum likelihood estimator is:

$$\hat{\boldsymbol{\theta}} = \frac{\boldsymbol{X}}{\boldsymbol{n}}$$

- What follows is a slight diversion from our aim of obtaining a confidence interval for θ. It is just demonstrating that the method is sound.

- Using X as the quantity $h(\underline{X})$, it is necessary to find if $h_1(\theta)$ and $h_2(\theta)$ exist such that $P\big(h_1(\theta) < X < h_2(\theta)\big) \geq 0.95$, where with equal tails $P\big(X \leq h_1(\theta)\big) \leq 0.025$ and $P\big(X \geq h_2(\theta)\big) \leq 0.025$.



8 Binomial Distribution

- There is no explicit expression for the pivotal quantity h($\underline{X}$).
- For the Bin (20,0.3) case:

$$P(X \leq 1) = 0.0076 \quad \text{and} \quad \mathrm{P(X} \leq 2) = 0.0355 \qquad \therefore h_1(\theta) = 1$$

- Also:

$$P(X \geq 11) = 0.0171 \ , \qquad \mathrm{P(X} \geq 10) = 0.0480 \qquad \therefore h_2(\theta) = 11$$

- So $h_1(\theta)$ and $h_2(\theta)$ do exist and increase with θ.

- Therefore the inequalities can be inverted as follows:

$$X \leq h_1(\theta) \Rightarrow \theta \geq \theta_1(X)$$
$$X \geq h_2(\theta) \Rightarrow \theta \leq \theta_2(X)$$

- These are the tail probabilities. So the inequalities involving θ_1 and θ_2 are defining the tails. Our confidence interval is the region not covered by these tail inequalities:

- This gives a 95% confidence interval of the form $\theta_2(X) < \theta < \theta_1(X)$.

- Note: The lower limit $\theta_2(X)$ comes from the upper tail probabilities and the upper limit $\theta_1(X)$ from the lower tail probabilities.

8 Binomial Distribution

- However since there are no explicit expressions for $h_1(\theta)$ and $h_2(\theta)$, there are no expressions for $\theta_1(X)$ and $\theta_2(X)$ and they will have to be calculated numerically.

- So, adopting the convention of including the observed x in the tails, θ_1 and θ_2 can be found by solving:

$$\sum_{r=x}^{n} b(r:n,\theta_1) = 0.025 \quad and \quad \sum_{r=0}^{x} b(r;n,\theta_2) = 0.025$$

- Here $b(r;n,\theta)$ denotes $P(X = r)$ when $X \sim Bin(n,\theta)$

- These can be expressed in terms of the distribution function $F(x;\theta)$:

$$1 - F(x-1;\theta_1) = 0.025 \quad and \quad F(x;\theta_2) = 0.025$$

- Note: Equality can be attained as θ has a continuous range (0,1) and the 'discrete' problem does not arise.

9 The normal approximation

- It is no bother for a computer to calculate an exact confidence interval for the binomial parameter p even if n is 'large'. However, on a piece of paper we use the normal approximation to the binomial distribution.
- $\frac{\boldsymbol{X-n\theta}}{\sqrt{\boldsymbol{n\theta(1-\theta)}}}$ can be used as a pivotal quantity.
- Solving the resulting equations for θ would not be easy.
- However $\frac{\boldsymbol{X-n\theta}}{\sqrt{\boldsymbol{n\hat{\theta}(1-\hat{\theta})}}}$, with $\hat{\theta}$ in place of θ (in the denominator only), can be used in a simpler way and yields the standard 95% confidence interval used in practice, namely:

$$\frac{\boldsymbol{X \pm 1.96\sqrt{n\hat{\theta}(1-\hat{\theta})}}}{\boldsymbol{n}}$$

$$\text{or} \quad \boldsymbol{\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}}, \qquad \text{where} \quad \boldsymbol{\hat{\theta} = \frac{X}{n}}$$

10 Poisson Distribution

- The Poisson situation can be tackled in a very similar way to the binomial for both large and small sample sizes.

- If $X_i, i = 1,2,\dots,n$ are independent Poi (λ) variables, that is, a random sample of size *n* from Poi(λ) , then $\sum X_i \sim Poi(n\lambda)$.

- Using $\sum X_i$ as a single observation from Poi (nλ) is equivalent to the random sample of size n from Poi (λ). This is similar to the single binomial situation. Recall that a Bin(n,p) distribution arises from the sum of n Bernoulli trials with probability of success *p*.

- Given a single observation X from a Poi (λ) distribution, then $P\big(h_1(\lambda) < X < h_2(\lambda)\big) \geq 0.95$, where $h_1(\lambda)$ and $h_2(\lambda)$ are increasing functions of λ .

- Inverting this gives $P\big(\lambda_1(X) < \lambda < \lambda_2(X)\big) = 0.95$.

- The resulting 95% confidence interval for λ is given by (λ_1, λ_2) where:

$$\boldsymbol{\sum_{r=x}^{\infty} p(r;\lambda_1) = 0.025 \quad \text{and} \quad \sum_{r=0}^{X} p(r;\lambda_2) = 0.025}$$

$$\boldsymbol{or \quad 1 - F(x-1;\lambda_1) = 0.025 \quad and \quad F(x;\lambda_2) = 0.025}$$

9 The normal approximation

- Again, it is easy for a computer to calculate an exact confidence interval for λ even for a large sample from Poisson(λ), or a single observation from Poisson(λ) where λ is large.

- However, on a piece of paper a normal approximation can be used either from $\boldsymbol{\sum X_i \sim Poi(n\lambda) \rightarrow N(n\lambda.\, n\lambda)}$or from the Central Limit Theorem as $\boldsymbol{\bar{X} \rightarrow N\left(\lambda, \frac{\lambda}{n}\right)}$.

- $\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}}$ can then be used as a pivotal quantity yielding a confidence interval. However, as in the binomial case, the standard confidence interval in practical use comes from $\frac{\bar{X}-\lambda}{\sqrt{\hat{\lambda}/n}}$ where $\hat{\lambda} = \bar{X}$.

- This clearly gives $\boldsymbol{\bar{X} \pm 1.96\sqrt{\frac{\bar{X}}{n}}}$ as an approximate 95% confidence interval for λ.

11 Confidence intervals for two-sample problems

- A comparison of the parameters of two populations can be considered by taking independent random samples from each population.

- The importance of the independence is illustrated by noting that:

$$\boldsymbol{var}\,[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- when the samples are independent.

- If the samples are not independent, then a covariance term will be included:

$$\boldsymbol{var}\,[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\mathrm{cov}[\bar{X}_1, \bar{X}_2]$$

- This covariance term can clearly have a substantial effect in the non-independent case.

- The most common form of non-independence is due to paired data.

11 Confidence intervals for two-sample problems

Two normal means

- **Case 1** (known population variance)

- If $\bar{X}_1$ and $\bar{X}_2$ are the means from independent random samples of size n_1 and n_2 respectively taken from normal populations which have known variances σ_1^2 and σ_2^2 respectively, then the equal-tailed $100(1-\alpha)\%$ confidence interval for the difference in the population means is given by:

$$(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \pm \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\boldsymbol{\sigma}_1^2}{\boldsymbol{n}_1} + \frac{\boldsymbol{\sigma}_2^2}{\boldsymbol{n}_2}}$$

- So for example, when $\alpha = 5\%$, we have $\boldsymbol{z}_{\alpha/2} = \boldsymbol{z}_{2.5\%} = \mathbf{1.9600}$.

11 Confidence intervals for two-sample problems

Two normal means

- **Case 2** (unknown population variance)

- If, $\bar{X}_1, \bar{X}_2, S_1$ and S_2 are the means and standard deviations from independent random samples of size n_1 and n_2 respectively taken from normal populations which have equal variances, then the equal-tailed $100(1-\alpha)\%$ confidence interval for the difference in the population means is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}, n_1+n_2-2,} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- where:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

- In any practical situation consideration must be made as to whether n_1 and n_2 are large or small and whether σ_1^2 and σ_2^2 are known or unknown. In the case of the *t* result it should be noted that there is the additional assumption of equality of variances.

- Note: The pooled estimator S_p^2 is based on the maximum likelihood estimator but adjusted to give an unbiased estimator.

11 Confidence intervals for two-sample problems

Two population variances

- For the comparison of two population variances it is more natural to consider the ratio σ_1^2/σ_2^2 than the difference $\sigma_1^2 - \sigma_2^2$. This follows logically from the concept of variance, but also from a technical point of view there is a pivotal quantity readily available for the ratio of normal variances but not for their difference.

$$\text{It is } \frac{S_1^2/S^2\_2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

- The resulting confidence interval is given by:

$$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_{n_2-1,n_1-1}$$

- It should be said that in practice the estimation of σ_1^2/σ_2^2 is not a common objective.

- However the same *F* result is used for the more common objective of 'testing' whether σ_1^2 and σ_2^2 may be equal, which is relevant for the *t* result for comparing population means. The acceptability of the hypothesis $H_0\colon \sigma_1^2 = \sigma_2^2$ can be determined simply by confirming that the value 1 is included in the confidence interval for σ_1^2/σ_2^2.

12 Two population proportions

- The comparison of population proportions corresponds to comparing two binomial probabilities on the basis of single observations X_1, X_2 from Bin (n_1, θ_1) and Bin(n_2, θ_2) respectively.

- Considering only the case where n_1 and n_2 are large, so that the normal approximation can be used, the pivotal quantity used in practice is:

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \sim N(0,1)$$

- where $\hat{\theta}_1, \hat{\theta}_2$ are the MLEs $\frac{X_1}{n_1}, \frac{X_2}{n_2}$, respectively.

13 Two Poisson parameters

- Considering the comparison of two Poisson parameters (λ_1 and λ_2) when the normal approximation can be used:
- $\bar{X}_i$ is an estimator of λ_i such that $\bar{X}_i \to N\left(\lambda_i, \frac{\hat{\lambda}_i}{n_i}\right)$
- Therefore $\bar{X}_1 - \bar{X}_2$ is an estimator of $\lambda_1 - \lambda_2$ such that:

$$\boldsymbol{\bar{X}_1 - \bar{X}_2 \to N\left(\lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)}$$

- Using $\hat{\lambda}_i = \bar{X}_i$, an approximate 95% confidence interval for $\lambda_1 - \lambda_2$ is given by:

$$\boldsymbol{\bar{X}_1 - \bar{X}_2 \pm 1.96\sqrt{\left(\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}\right)}}$$

- We are assuming that the two samples are independent.

14 Paired data

- Paired data is a common example of comparison using non-independent samples.
- Essentially having paired or matched data means that there is one sample:

$$(X_{11}, X_{21}), (X_{12}, X_{22}), (X_{13}, X_{23}), \dots, (X_{1n}, X_{2n})$$

- rather than two separate samples:

$$(X_{11}, X_{12}, X_{13} \dots, X_{1n})\ and\ (X_{21}, X_{22}, X_{23}, \dots, X_{2n})$$

- The paired situation is really a single sample problem, that is, a problem based on a sample of n pairs of observations. (In the independent two-sample situation the sample sizes need not, of course, be equal.)

- Paired data can arise in the form of 'before and after' comparisons.

14 Paired data

- Investigations using paired data are usually better than two-sample investigations in the sense that the estimation is more accurate.

- Paired data are analysed using the differences $D_i = X_{1i} - X_{2i}$ and estimation of $\mu_D = \mu_1 - \mu_2$ is considered. A z result or a t result can be used, but the latter will be more common as it is unlikely that the variances of the differences will be known. Assuming normality of the population of such differences (but not necessarily the normality of the X_1 and X_2 populations), the pivotal quantity for the t result is:

$$\frac{\bar{\boldsymbol{D}} - \boldsymbol{\mu_D}}{\boldsymbol{S_D}/\sqrt{\boldsymbol{n}}} \sim \boldsymbol{t_{n-1}}$$

- Note that S_D is calculated from the values of D.

- The resulting 95% confidence interval for μ_D will be $\bar{\boldsymbol{D}} \pm \boldsymbol{t_{0.025,n-1}} \frac{\boldsymbol{S_D}}{\sqrt{\boldsymbol{n}}}$.

Thank You