STAT 390 A
Statistical Methods in Engineering and Science
Week 8 Lectures – Part 2 – Spring 2023
Confidence Intervals

Alexander Giessing

Department of Statistics

University of Washington

May 17, 2023

Outline

Example: Gross calorific content of coal

ISO 1928 is a standardized procedure issued by the International Organization for Standardization (ISO) to determine the heat content (aka gross calorific value) of coal in megajoules per kilogram (MJ/kg). When this procedure is carried out the resulting measurement errors are known to be approx. normal with standard deviation 0.1 MJ/kg. Consider the following data obtained from a shipment of coal using IS 1928:

23.870 23.730 23.712 23.760 23.640 23.850 23.840 23.860
23.940 23.830 23.877 23.700 23.796 23.727 23.778 23.740
23.89 23.780 23.678 23.771 23.860 23.6090 23.800

Construct a 95% CI for the expected gross calorific value in MJ/kg of the shipment of coal.

Example: Gross calorific content of coal (Cont.)

Construct a 95% CI for the unknown expected gross calorific value in MJ/kg of the shipment of coal.

- What CI do we want to construct?
  - the data set is small, $n = 23$.
  - the data are (approx.) normal with known standard deviation.

  $\Longrightarrow$ small-sample CI for the mean with known variance!
- We have $\bar{x}_n = 23.788$, $\sigma = 0.1$ and $\alpha = 5\%$.
- We use R to find the critical value

$$z_{0.025} = \texttt{qnorm(0.975)} = 1.96.$$

Hence, a 95% CI for the expected gross calorific value in MJ/kg of the shipment of coal is

$$\left(23.788 - 1.96\frac{0.1}{\sqrt{23}},\ 23.788 + 1.96\frac{0.1}{\sqrt{23}}\right) = (23.747,\ 23.829) \quad MJ/kg.$$

Example: Rutherford-Geiger experiment

In a well-known experiment Rutherforad and Geiger collected observations on the radioactive decay of polonium by counting the number of alpha-particles emitted from a small disk coated with polonium during 2608 intervals of 7.5 seconds each. Their data are summarized in the following table:

Count	0	1	2	3	4
Frequency	57	203	383	525	532
Count	5	6	7	8	9
Frequency	408	273	139	45	27
Count	10	11	12	13	14
Frequency	10	4	0	1	1

Construct a 98% CI for the expected number of alpha-particles per interval.

Example: Rutherford-Geiger experiment (Cont.)

Construct a 98% CI for the expected number of alpha-particles per interval.

- What CI do we want to construct?
  - the data set is large, $n = 2608$.
  - the data are not normal, because count data is discrete and non-negative.
  - we do not know the variance of the data.

  $\Longrightarrow$ large-sample CI for the mean with unknown variance!

- The sample average of alpha-particles is $\bar{x}_n = 3.8715$, the sample standard deviation $s_n = 1.9225$ and $\alpha = 2\%$.

- We use R to find the critical value

$$z_{0.01} = \texttt{qnorm(0.99)} = 2.33.$$

  Hence, a 98% CI for the expected number of alpha-particles per interval is

$$\left(3.8715 - 2.33\frac{1.9225}{\sqrt{2608}},\ 3.8715 + 2.33\frac{1.9225}{\sqrt{2608}}\right) = (3.784,\ 3.959).$$

Example: Gross calorific value (Cont.)

Suppose that there is a different shipment of coal and there are some doubts about whether the stated accuracy of the ISO 1928 method was attained. Therefore you prefer to consider the standard deviation σ unknown. We have the following 22 measurements:

30.990	31.030	31.060	30.921	30.920	30.990	31.024	30.929
31.050	30.991	31.208	30.830	31.330	30.810	31.060	30.800
31.091	31.170	31.026	31.020	30.880	31.125		

1. Construct a 95% CI for the expected gross calorific value in MJ/kg of the shipment of coal.

2. You decide to buy the shipment of coal if you are confident that the gross calorific content exceeds 31.00 MJ/kg. Would you buy the shipment?

Example: Gross calorific value (Cont.)

1. Construct a 95% CI for the expected gross calorific value in MJ/kg of the shipment of coal.

- What CI do we want to construct?
  - the data set is small, $n = 22$.
  - the data are (approx.) normal with unknown standard deviation.

  $\Longrightarrow$ small-sample CI for the mean with unknown variance!
- We have $\bar{x}_n = 31.012$, $s_n = 0.1294$ and $\alpha = 5\%$.
- Find the critical value from the t-distr. with $n - 1$ degrees of freedom

$$z = t_{21,0.025} = \texttt{qt(0.975, 21)} = 2.080.$$

  Hence, a 95% CI for the expected gross calorific value in MJ/kg is

$$\left(31.012 - 2.080\frac{0.1294}{\sqrt{22}},\ 31.012 + 2.080\frac{0.1294}{\sqrt{22}}\right) = (29.954,\ 31.069)\ \ MJ/kg.$$

- *Verify that the 95% small-sample CI with known variance $\sigma = 0.1294$ is smaller! Intuition? Why?*

Example: Gross calorific value (Cont.)

2. You decide to buy the shipment of coal if you are confident that the gross calorific content exceeds 31.00 MJ/kg. Would you buy the shipment?

- One way of answering this question is to construct a one-sided (lower) CI!
  - the data set is small, $n = 22$.
  - the data are (approx.) normal with unknown standard deviation.

$\Longrightarrow$ small-sample CI for the mean with unknown variance!

- We have $\bar{x}_n = 31.012$, $s_n = 0.1294$ and $\alpha = 5\%$.
- Find the critical value from the t-distr. with $n - 1$ degrees of freedom

$$t_{21,0.05} = \texttt{qt(0.95, 21)} = 1.721$$

The 95% lower CI for the expected gross calorific value in MJ/kg is

$$\left(31.012 - 1.721\frac{0.1294}{\sqrt{22}},\ \infty\right) = (30.964,\ \infty) \quad MJ/kg.$$

- You are 95% confident that the gross calorific value of this shipment is at least 30.964 MJ/ kg. You would only buy it if the lower bound was $31 MJ/kg$ or more.

Outline

Width of confidence intervals and sample size

- The narrower a confidence interval the better. *(Why?)*
- As a general principle, we know that we can make more accurate statements if we have a larger sample:
  - Intuitive: larger sample = more information $\implies$ better inference.
  - Rigorous: large sample n $\implies$ small SE $= \sigma/\sqrt{n}$.
- Sometimes, an accuracy requirement (i.e. quality control) is set before data are collected.
  - Can we make precise recommendations on how much data should be collected in order to achieve the desired accuracy?
  - What factors influence the amount of data that needs to be collected?
  - Can we always achieve the desired accuracy?

Example: Gross calorific value (Cont.)

Suppose that you have to test a shipment of goal. You want to compute a 95% but it should not be wider than 0.05 MJ/kg, i.e.e the lower and upper bound should not differ more than 0.05. How many measurements do you need?

- Assume (for simplicity) that the data are approx. normal with known standard deviation σ. Then, the width of the symmetric $100(1-\alpha)\%$ CI for the mean is

$$2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- Given a desired width w we obtain the following lower bound on the sample size n.

$$2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq w \qquad \Longleftrightarrow \qquad n \geq \left(\frac{2z_{\alpha/2}\sigma}{w} \right)^2.$$

Example: Gross calorific value (Cont.)

- Plugging in the numbers from the example, $z_{0.025} = 1.96$, $\sigma = 0.1$, $w = 0.05$ we learn that

$$n \geq \left(\frac{2 \times 1.96 \times 0.1}{0.05} \right)^2 = 61.4,$$

  i.e. you should at least collect 62 measurements to achieve the desired accuracy.

- How does the sample size requirement depend on σ? Why is this dependence intuitive?

- **Note:** If the standard deviation σ is unknown, we need to guess it. Then, above method gives only a rough indication of the required sample size. The standard deviation as we (afterwards) estimate if from the data may turn out to be quite different from the initial guess. Thus, the obtained CI may be smaller or larger than intended.

Outline

General observation

- Thus far, all CIs that we have considered take the following form:
  - if the variance is known:

$$\left[\begin{pmatrix}\text{point}\\ \text{estimate}\end{pmatrix} \pm \begin{pmatrix}\text{critical}\\ \text{value}\end{pmatrix} \cdot \begin{pmatrix}\text{Standard}\\ \text{Error}\end{pmatrix}\right],$$

  - if the variance is unknown:

$$\left[\begin{pmatrix}\text{point}\\ \text{estimate}\end{pmatrix} \pm \begin{pmatrix}\text{critical}\\ \text{value}\end{pmatrix} \cdot \begin{pmatrix}\text{Estimated}\\ \text{Standard Error}\end{pmatrix}\right],$$

  - the critical values are obtained either from $N(0,1)$ or $t(n-1)$.

- The general definition of CIs is much broader than this.
  - Question 4 in HW 6
  - the example on the following slides
  - ...

Example: A "naive" CI for a Proportion

In a city, a survey organization takes a random sample of 900 families and finds that 187 have an income larger than \$80K (cf. Week 6 Lectures, Slides 36f). Construct a 95% CI for the unknown proportion p of families that have income larger than \$80K.

- large sample size $n = 900$, distribution and variance are unknown.
  $\Longrightarrow$ large-sample CI with unknown variance.

- Let $Y_i = 1$ if the ith draw has income larger than \$80K and 0 otherwise. Then, $\hat{p} = \bar{Y}_n$ is the MLE (and MME) of the unknown parameter p. The estimate is $187/90 = 0.2078$. *Verify this!*

- The standard deviation of Y_i is $\sqrt{(1-p)p}$ and we can estimate this standard deviation using the plug-in estimator $\sqrt{(1-\hat{p})\hat{p}}$. The estimate is $\sqrt{(1-187/900)187/900} = 0.4057$. *Verify this!*

- Therefore, a large-sample 95% CI is given by

$$\left(0.2078 - 1.96\frac{0.4057}{\sqrt{900}},\ 0.2078 + 1.96\frac{0.4057}{\sqrt{900}}\right) = (0.1943,\ 0.2213)$$

Alternative approach to CIs for a Proportion

- Recall the normal approximation to the Binomial distribution: If $X \sim Bin(n, p)$ then, for large n

$$\frac{X - np}{\sqrt{np(1-p)}}$$

can be approximated by the standard normal distribution $N(0, 1)$.

- The Y_i's defined above are independent $Ber(p)$ and, hence,

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\bar{Y} - p)}{\sqrt{p(1-p)}}.$$

- Therefore, for large n,

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\bar{Y} - p)}{\sqrt{p(1-p)}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Alternative Approach to CIs for a Prop. (Cont.)

- We have

$$
\begin{aligned}
-z_{\alpha/2} &< \frac{\sqrt{n}(\bar{Y}-p)}{\sqrt{p(1-p)}} < z_{\alpha/2} \\
&\Leftrightarrow \left(\frac{\sqrt{n}(\bar{Y}_n-p)}{\sqrt{p(1-p)}}\right)^2 < (z_{\alpha/2})^2 \\
&\Leftrightarrow \left(\bar{Y}_n-p\right)^2 - (z_{\alpha/2})^2\frac{p(1-p)}{n} < 0
\end{aligned}
$$

- Plugging in the estimate 0.2078 for $\bar{Y}_n$, 1.96 for $z_{\alpha/2}$, and 900 for n and tedious algebra, yields

$$p^2 - 0.4181p + 0.0430 < 0.$$

- This quadratic form describes a parabola. The values where the parabola intersects the horizontal axis are $p_1 = 0.182$ and $p_2 = 0.2356$. Thus, an alternative 95% CI for the sample proportion is given by

$$(0.1826,\ 0.2356)$$

- **Question:** How does the result change if you apply the continuity correction for the normal approximation of the Binomial distribution?