Inference Overview

michaels

October 10, 2017

Contents

1 Inference for a population mean when the population variance is "known"

1.1 Statistical Model

- Data $x_1, \dots, x_n$ modelled as "observed values" or "realisations" of iid random variables $X_1, \dots, X_n$ with $E(X_1) = \mu$ **unknown** and $Var(X_1) = \sigma_0^2$ **known**.
  - We call μ and σ_0^2 the "population mean and variance".
- Under the model, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is our **estimator** (i.e. a random variable whose "observed value" or "realisation" is to be used as an *estimate* of the unknown parameter). We can say that (again, under the model)
  - $E\left(\bar{X}\right) = \mu$
  - $Var\left(\bar{X}\right) = \frac{\sigma_0^2}{n}$, so
  - $SD(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma_0}{\sqrt{n}}$. This is the (true) **standard error**.

- Furthermore, we require that $\bar{X}$ is (at least approximately) normally distributed. Then whatever be the true value of μ, the **pivot**

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1)\,,$$

  (at least approximately):

  - this follows immediately if the X_i's themselves are **exactly** normally distributed;
  - this also follows if the sample size is "large enough" so that the Central Limit Theorem applies:
    * remember "large enough" is a subtle issue, it depends how "non-normal" the X_i's actually are.

- The first step is to provide an estimate $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and quote an associated standard error $\sigma_0/\sqrt{n}$. From here we can think of there being two "different" approaches:

  - Hypothesis test
  - Confidence interval

1.2 Hypothesis test

Is a useful option where there is a known "baseline" or "default" value μ_0 of the parameter, and the scientific question is looking to find evidence that something **different** to the baseline is going on. We formulate a **null hypothesis** $H_0\colon \mu = \mu_0$ which corresponds to "parameter is equal to the baseline value".

- We identify a **test statistic** which is a random variable under the model

  - whose distribution is *known* when H_0 is true;
  - which helps us measure evidence *against* H_0.

- The test statistic may equivalently be identified as either

  - the sample mean $\bar{X} \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right)$ under H_0;
  - the "Z-statistic", obtained by plugging the hypothesised value μ_0 into the pivot above, giving

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1) \text{ under } H_0.$$

- We compute a **p-value** which is the probability, under the model, of a certain event. This event is

  > "Getting a result which constitutes at least as much evidence against H_0 *as was observed*."

- Deciding what constitutes "at least as much evidence against H_0" depends on the context. We can identify a few different possibilities:

  - **One-sided** In this case deviations in a particular direction are anticipated before the data is observed. We express this in terms of an **alternative hypothesis**. There are two sub-cases:

    * $H_1 : \mu > \mu_0$. This would be appropriate if it was anticipated that if any deviations from H_0 occur, they would result in a *larger* sample mean; thus *larger values of the sample mean constitute more evidence against* H_0. In this case the p-value is

      $$P(\bar{X} \geq \bar{x}) \quad \text{where} \quad \bar{X} \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

      or equivalently,

      $$P(Z \geq z)$$

      where

      - $Z \sim N(0, 1)$;
      - $z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}$ is the observed value of the Z-statistic.

    * $H_1 : \mu < \mu_0$. This would be appropriate if it was anticipated that if any deviations from H_0 occur, they would result in a *smaller* sample mean; thus *smaller values of the sample mean constitute more evidence against* H_0. In this case the p-value is

      $$P(\bar{X} \leq \bar{x}) \quad \text{where} \quad \bar{X} \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

      or equivalently,

      $$P(Z \leq z)\,.$$

- **Two-sided** In this case deviations in either direction are equally "interesting". Since the test statistic(s) have *symmetric* distributions under H_0 "at least as much evidence against H_0" can sensibly be translated as either
  - "$\bar{X}$ takes a value *at least as far away from* μ_0 as was observed" or
  - "the Z-statistic is *at least as large in absolute value* as was observed".
- We formulate the alternative hypothesis as $H_1\colon \mu \neq \mu_0$.
  - The p-value is computed as

$$P\left(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|\right) = \begin{cases} 2P(\bar{X} \geq \bar{x}) & \text{if } \bar{x} > \mu_0; \\ 2P(\bar{X} \leq \bar{x}) & \text{if } \bar{x} < \mu_0. \end{cases}$$

    Equivalently,

$$P(|Z| \geq |z|) = 2P(Z \geq |z|)$$

    where $Z \sim N(0, 1)$.

- The **smaller** the p-value, the **more** evidence against H_0, and thus (indirectly) the **stronger the suggestion** that "something interesting is going on".
- If the p-value is **not small** we say that "the data is consistent with H_0". This does not mean H_0 is true; maybe we just didn't collect enough data.

1.3 Confidence intervals

These provide an alternative way of indicating which parameter values might be considered "consistent with the data".

- A confidence interval is the "realisation" of a *random set* which, under the model, contains the true parameter value with a fixed, known, (usually high) probability. We also make the distinction between **one-sided** and **two-sided** intervals.
  - Two-sided intervals are of the form $\bar{x} \pm c\,\frac{\sigma_0}{\sqrt{n}}$, where c is chosen so that the random set this is a realisation of covers μ with probability equal to the given "confidence level" $100(1-\alpha)\%$:

    - the "multiplier" c satisfies

      $$P(-c \leq Z \leq c) = 1 - \alpha$$

      where $Z \sim N(0, 1)$.

  - One sided intervals are of the form
    - $(-\infty, \bar{x} + c\frac{\sigma_0}{\sqrt{n}}]$ or
    - $[\bar{x} - c\frac{\sigma_0}{\sqrt{n}}, \infty)$

    depending on the context.

    - The (finite) endpoints are also called "upper" or "lower" (respectively) confidence limits.
    - The upper confidence limit $\bar{x} + c\frac{\sigma_0}{\sqrt{n}}$ may be interpreted as "the largest parameter value consistent with the data" in some sense, and is often associated with the one-sided alternative hypothesis $H_1\colon \mu < \mu_0$, where it is suspected that the true value is *less* than hypothesised.

- **Connection between the two approaches** Careful consideration of how both confidence intervals and p-values are obtained shows a strong connection between them:

  μ_0 is in the $100(1-\alpha)\%$ C.I. iff p-value of $H_0\colon \mu = \mu_0 > \alpha$.

  e.g. when $\alpha = 0.05$, both occur in the two-sided case iff

  $$\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}} \leq 1.96\,.$$