15. Inference in practice

The Practice of Statistics in the Life Sciences
Third Edition
Objectives (PSLS Chapter 15)

Inference in practice

- Confidence intervals in practice
- Significance tests in practice
- Beware of multiple analyses
- Caution about $z$ procedures for a population mean
Confidence intervals in practice

The margin of error does not cover all errors: The margin of error in a confidence interval covers only \textit{random sampling} error.

Undercoverage, nonresponse or other forms of bias are often more serious than random sampling error (e.g., our elections polls). The margin of error does not take these into account at all.

For instance, most opinion polls have very high nonresponse (about 90%). This is not taken into account in the margin of error.
For the same confidence level, narrower confidence intervals can be achieved by using larger sample sizes.
Sample size and experimental design

You may need a certain margin of error (e.g., drug trial, manufacturing specs). In many cases, the population variability ($\sigma$) is fixed, but we can choose the number of measurements ($n$).

Using simple algebra, you can find what sample size is needed to obtain a desired margin of error.

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \iff \quad n = \left( \frac{z^* \sigma}{m} \right)^2$$
Density of bacteria in solution

A measuring equipment gives results that vary Normally with standard deviation $\sigma = 1$ million bacteria/ml fluid.

How many measurements should you make to obtain a margin of error of at most 0.5 million bacteria/ml with a confidence level of 90%?

For a 90% confidence interval, $z^* = 1.645$.

$$n = \left( \frac{z^* \sigma}{m} \right)^2 \Rightarrow n = \left( \frac{1.645 \times 1}{0.5} \right)^2 = 3.29^2 = 10.8241$$

Using only 10 measurements will not be enough to ensure that $m$ is no more than 0.5 million/ml. Therefore, we need at least 11 measurements.
Statistical significance only says whether the effect observed is likely to be due to chance alone because of random sampling.

- Statistical significance doesn’t tell about the magnitude of the effect.
- Statistical significance may not be practically important.
- With a large sample size, even a small effect could be significant.

A drug to lower temperature is found to consistently lower a patient’s temperature by 0.4° Celsius ($P$-value < 0.01). But clinical benefits of temperature reduction require a 1°C decrease or greater.

Plot your results, compare them with a baseline or similar studies.
Sample size affects statistical significance

- Because **large random samples** have small chance variation, very small population effects can be highly significant if the sample is large.

- Because **small random samples** have a lot of chance variation, even large population effects can fail to be significant if the sample is small.
Type I and II errors

Statistical conclusions are not certain.

- A **Type I error** occurs when we reject the null hypothesis but the null hypothesis is actually true (incorrectly reject a true $H_0$).

- A **Type II error** occurs when we fail to reject the null hypothesis but the null hypothesis is actually false (incorrectly keep a false $H_0$).
The probability of making a Type I error (incorrectly rejecting a true $H_0$) is the significance level $\alpha$.

The probability of making a Type II error (incorrectly keeping a false $H_0$) is labeled $\beta$, a computed value that depends on a number of factors. The **power** of a test is defined as the value $1 - \beta$.

A Type II error is not definitive because “failing to reject the null hypothesis” does not imply that the null hypothesis is true.
A regulatory agency checks air quality for evidence of unsafe levels (> 5.0 ppt) of nitrogen oxide (NO$_x$). The agency gathers NO$_x$ concentrations in an urban area on a random sample of 60 different days and calculates a test of significance to assess whether the mean level of NO$_x$ is greater than 5.0 ppt.

**A Type I error here would be to believe that the population mean NO$_x$ level**

A. exceeds 5.0 ppt when it really does.

B. exceeds 5.0 ppt when it really doesn’t.

C. is 5.0 ppt or less when it really is.

D. is 5.0 ppt or less when it really isn’t.
The power of a test

The power of a test of hypothesis is its ability to detect a specified effect size (reject $H_0$ when a given $H_a$ is true) at significance level $\alpha$.

The specified effect size is chosen to represent a biologically/practically meaningful magnitude.

What affects power?

- The size of the specified effect
- The significance level $\alpha$
- The sample size $n$
- The population variance $\sigma^2$
Do poor mothers have smaller babies?

The national average birth weight is 120 oz: \(N(\mu_{\text{natl}} = 120, \sigma = 24 \text{ oz})\).

We want to be able to detect an average birth weight of 114 oz (5% lower than the national average).

What power would we get from an SRS of 100 babies born of poor mothers if we chose a significance level of 0.05?

\[ \rightarrow 80\% \]
Beware of multiple analyses

The probability of incorrectly rejecting $H_0$ (Type I error) is the significance level $\alpha$. If we set $\alpha = 5\%$ and make multiple analyses, we can expect to make a Type I error about 5\% of the time.

- If you run only 1 analysis, this is not a problem.
- If you try the same analysis with 100 random samples, you can expect about 5 of them to be significant even if $H_0$ is true.

You run a test of hypotheses for extra sensory perception on an individual chosen at random. You then run the same test on 19 other individuals also chosen at random. **What’s wrong with that?**

For a significance level $\alpha = 5\%$, you can expect that one individual will have a significant result just by chance even if extrasensory perception doesn’t exist.
Cell phones and brain cancer

Might the radiation from cell phones be harmful to users? Many studies have found little or no connection between using cell phones and various illnesses. Here is part of a news account of one study:

A hospital study that compared brain cancer patients and similar patients without brain cancer found no statistically significant association between cell phone use and brain cancer (glioma). But when 20 types of glioma were considered separately, an association was found between phone use and one rare form of glioma. Puzzlingly, however, this risk appeared to decrease rather than increase with greater cell phone use.

Interpretation?
Caution about \( z \) procedures for a mean

- The data must be a probability sample or come from a randomized experiment. Statistical inference cannot remedy basic design flaws, such as voluntary response samples or uncontrolled experiments.

- The sampling distribution must be approximately Normal. This is not true in all instances (if the population is skewed, you will need a large enough sample size to apply the central limit theorem).

- To use a \( z \) procedure for a population mean, we must know \( \sigma \), the population standard deviation. This is often an unrealistic requisite. We'll see what can be done when \( \sigma \) is unknown in the next chapter.
Mammary artery ligation

Angina is the severe pain caused by inadequate blood supply to the heart. Perhaps we can relieve angina by tying off (“ligation”) the mammary arteries to force the body to develop other routes to supply blood to the heart. Patients reported a statistically significant reduction in angina pain.

Problem?

- This experiment was uncontrolled, so that the reduction in pain might be nothing more than the placebo effect.
- A randomized comparative experiment later found that ligation was no more effective than a placebo. Surgeons abandoned the procedure.
- Statistical significance says that something other than chance is at work, but it doesn’t say what that something is.