Confidence Interval, Hypothesis Test and P-value

Confidence intervals - a way to combine the point estimate (e.g., the mean of a sample) and its standard error is using a confidence interval. Most confidence intervals we encounter are of the form:

estimate ± table value × standard error

Confidence interval for a population proportion.

- Suppose we are trying to estimate an unknown population proportion p.
- In large samples, the $100(1-\alpha)$% confidence interval for p is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})\right)$$

  where $z_{\alpha/2}$ is that value from the standard normal distribution for which only $\alpha/2$ of the probability is above. (For example $z_{.025} = t_{\infty,.025} = 1.96$, $z_{.05} = t_{\infty,.05} = 1.645$, $z_{.005} = t_{\infty,.005} = 2.576.$)
- Example: To test the effectiveness of a new pain-relieving drug, 80 (randomly selected?) patients at a clinic were given a pill containing a drug, and 56 of the patients showed improvement of their pain symptoms. The estimated probability of showing improvement is $\hat{p} = 56/80 = .7$, and the 90% confididence interval for the true population proportion showing improvement is

$$.7 \pm 1.645\sqrt{.7(1-.7)/80}$$

  or approximately, (.62,.78).

Confidence interval for a population mean.

- Suppose we are interested in estimating the mean μ of a normal population whose variance σ^2 is unknown.
- Take a random sample $X_1, X_2, \ldots, X_n$ from the population.
- If the underlying population is approximately normally distributed, the $100(1-\alpha)$% confidence interval for the mean μ is

  $$\left(\bar{X} - t_{n-1,\alpha/2}s/\sqrt{n}, \bar{X} + t_{n-1,\alpha/2}s/\sqrt{n}\right)$$

  where $t_{n-1,\alpha/2}$ is that value from the t-distribution with $n-1$ degrees of freedom for which only $\alpha/2$ of the probability lies above. (For example $t_{12,.025} = 2.179$, $t_{40,.025} = 2.021$, $t_{\infty,.025} = 1.960$.)
- Example: To assess the level of iron in the blood of children with cystic fibrosis, a random sample is selected from the population of children with CF. There were $n = 13$ children in the sample, having an average iron level $\bar{X} = 11.9\mu$mol/l with sample standard deviation $s = 6.3\mu$mol/l.

  Based on this sample, the 95% confidence interval for the population mean μ is

  $$11.9 \pm 2.179(6.3)/\sqrt{13}$$

  or approximately (8.09, 15.70).

Interpretation of a confidence interval

- The interpretation of a confidence interval is non-intuitive, but very important.
- Before the data are collected, the 95% confidence interval for a population mean μ has probability .95 of containing μ.
- Once the data are collected and we have values for $\bar{X}$ and s, the interval becomes fixed.
- The true mean is either in the interval or not. We don't know and can't assign probability to the mean being in the interval.
- (The population mean μ is not a random variable, but an unknown constant which we are trying to estiamte.)
- We say we are 95% confident that a 95% confidence interval contains the true mean μ, because of the large probability in advance that such an interval contains the mean.
- Confidence intervals get narrower as the sample size increases.
- e.g. A 90% confidence interval for a proportion is narrower than a 95% confidence interval, because $z_{.05}$ is smaller than $z_{.025}$
- (To be more confident of containing the true value we must take a wider interval.)
- Note that a 95% confidence interval does NOT contain 95% of the data values, or 95% of the values in the population.
  - The population standard deviation is σ rather than $\sigma/\sqrt{n}$, and the interval containing 95% of the population is

    $$(\bar{X} - z_{.05}\sigma, \bar{X} + z_{.05}\sigma)$$

    which can be much wider than the confidence interval.

Hypothesis testing

- In hypothesis testing we wish to compare one theory to another.
- In this course we will compare the effects of two drugs or treatments.
- These theories are most often stated in terms of population parameters.
- For example, where μ is the mean iron concentration in the population of children with CF, we may wish to compare

$$H_0 : \mu \leq \mu_0$$

to

$$H_a : \mu > \mu_0$$

where μ_0 is a particular value of interest.

- The alternative hypothesis H_a is usually the theory we want to prove, and we attempt to do so by showing that the data disagrees with H_0.
- The null hypothesis H_0 is the theory we will stick with, unless we get strong evidence against it. The null hypothesis always includes the $=$ piece, which is the value being tested for.
- The significance probability, or P value, or "p-value", is the probability of getting data as extreme or more extreme than what was observed if H_0 is true.
- A small P value gives evidence against H_0 and in favour of H_a.
- Many test statistics we encounter also combine the estimate and its standard error.

- For example,

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

  is typically used to test the hypotheses above.

- This test statistic measures the difference between the estimate and the value specified in the equality in H_0, on a standardized scale.

- Before we collect the data, the distribution of the test statistic assuming H_0 is true is called the null distribution.

- Often this distribution is the standard normal (when testing for proportions), or something close to the normal, like the t distribution (when testing for means).

- The **p-value** is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. For the above example, the alternative hypothesis is **one-sided** and the p-value is $P(t > t_{obs})$. It is obtained from tables or using a suitable computer program.

- The alternative hypotheses can be **two-sided**, e.g. $H_a : \mu \neq \mu_0$, in which case the p-value is $2P(t \geq |t_{obs}|)$, to allow for values as extreme in either tail of the distribution.

- The p-value is widely used to quantify the evidence against H_0, generally in medical applications, as follows:

p-value	Amount of evidence
$\geq .1$	none
$.05 \leq P < .1$	weak
$.01 \leq P < .05$	strong
$< .01$	very strong

- These cut-off values arose from the use of tables which gave the .01, .05 and .1 percentiles of the distribution.
- If $p - value < \alpha$, where α is a small number (like .05) we can say that the results are statistically significant at the α (say .05) level of significance. Some people use phrases like "reject the null hypothesis at level α", "reject the null hypothesis at the 5% level", etc.
- There are two types of error which can occur with fixed α testing.
- A type I error occurs when we declare statistical significance when H_0 is true.
  - This is considered a bad thing, so α (the probability of a type I error) is chosen to be small.
- A type II error occurs when we fail to declare statistical significance when H_0 is false.
- The power of a test is the probability of correctly declaring statistical significance when H_0 is false.
- The power is increased by increasing the sample size.
- There is an analogy to a court of law.
  - We assume the defendant is not guilty, so H_0 : not guilty.
  - The alternative is that they are guilty, H_a: guilty.
  - Evidence is presented and weighed (test statistic).
  - If the evidence is overwhelming (the test statistic is large and P is very small), the presumption of innocence is overturned (statistical significance is declared).
  - A type I error is to find defendant guilty when they are not (it is considered very bad, for to hang an innocent person, we want the evidence to be overwhelming.)

  - A type II error is to let a guilty man go free. (This is not considered to be too bad.)

- Remember that statistical significance doesn't imply practical significance. A small difference will always give a small p-value if the sample is suitably large. For example, suppose that $\bar{X} = 20.1$, $H_0 : \mu = 20$, $H_a : \mu > 20$ and $\sigma = 1$.

n	t	p-value
9	.3	.38
100	1	.16
900	3	.001

  - It is generally advised to also report a confidence interval when carrying out a hypothesis test.
  - The confidence interval shows us how precisely we are estimating the mean, and gets narrower as the sample size gets bigger.

- Don't ignore lack of significance.
  - The negative result may be unexpected.
  - Was the sample size large enough? i.e. was the study underpowered?

- Beware of searching for significance.
  - Many tests on same data.
  - One test on the most extreme difference.
  - "Almost any large data set will contain some unusual pattern."
  - The hypotheses must be stated first.
- There is a relationship between two-sided fixed α tests and $1 - \alpha$ confidence intervals
  - A level $(1 - \alpha)$100% confidence interval contains all μ_0 for which $H_0 : \mu = \mu_0$ is not statistically significant at level α, when a two-sided alternative is used.
  - REJECT the two sided hypothesis at level α if and only if the $100(1 - \alpha)$% confidence interval for μ DOES NOT contain μ_0.