Statistics for the Biological, Environmental and Health Sciences

STAT 007

Estimating Parameters and Determining the Sample Size

Chapter 7

Point Estimates, Confidence Interval, and Samples Sizes

Sections 7-1 and 7-2

- In this sections we will:
  - Introduce estimates for population parameters (population mean and population proportion).
  - Discuss how to interpret these estimates.
  - Discuss how to determine the sample size to estimate a population parameter.

Determining Sample Size for p

Example

If we were to conduct a survey to determine the proportion of children (older than 1 year) who have received measles vaccinations, how many children must be surveyed in order to be 95% confident that the sample proportion is in error by no more than three percentage points?

a) Assume that a recent survey showed that the proportion of children that have received measles vaccinations is 0.9.

b) Assume that we have no prior information suggesting a possible value of the population proportion.

Estimates for μ

- In what follows we will provide estimates for the population mean, μ.
- We will consider the more realistic case in which the population standard deviation, σ, is also unknown.
- Critical values from a $(1-\alpha)100\%$ confidence interval estimate for μ are obtained from the Student *t* distribution and denoted $t_{\alpha/2}$.
- The Student *t* distribution has a parameter called *degrees of freedom*, denoted *df*.
- When estimating μ, the critical value $t_{\alpha/2}$ separates an area of $\alpha/2$ in the right tail of the Student *t* distribution with $df = n - 1$, where *n* is the size of the sample.

Estimates for μ (unknown σ)

- The sample mean, $\overline{x}$, is the best **point estimate** of the population mean, μ:
  - it is an unbiased estimator
  - is a consistent estimator: the standard deviation of the sampling distribution of $\overline{x}$ is smaller than that of other estimators for μ.
- The $(1-\alpha)100\%$ **confidence interval estimate** for the population mean is

$$\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

  where $s = \sqrt{\sum \frac{(x_i-\overline{x})^2}{n-1}}$ is the sample standard deviation.
- **Requirements** to use the above interval estimate:
  - The sample is a simple random sample.
  - Either or both of these conditions are satisfied: the population is normally distributed or $n > 30$.

Determining Sample Size when estimating μ

- To determine how large the sample size n should be in order to estimate the population mean μ with a $(1-\alpha)100\%$ confidence interval and a desired E:
  - — if σ is known use $n = \left[\frac{z_{\alpha/2}\ \sigma}{E}\right]^2$
  - — When σ is not known (most regular case) use $n = \left[\frac{z_{\alpha/2}\ \sigma}{E}\right]^2$, with
    a) $\sigma \approx range/4$
    b) $\sigma \approx s$
    c) σ a value from results of some other earlier studies.

Estimates for μ

Example

Listed below are weights (in hectograms or hg) of randomly selected girls at birth and corresponding normal quantile plot, based on data from the National Center for Health Statistics. Here are the summary statistics: $n = 15$, $\overline{x} = 30.9$ hg, $s = 2.9$ hg.

Data: `33 28 33 37 31 32 31 28 34 28 33 26 30 31 28`

a) Find a point estimate for the mean birth weight of girls.

b) Based on the quantile plot in next slide, are the requirements to find a confidence interval estimate for μ reasonable?

c) Find the margin of error E that corresponds to a 95% confidence interval for μ.

d) Find the 95% confidence interval estimate of the population mean birth weight of girls.

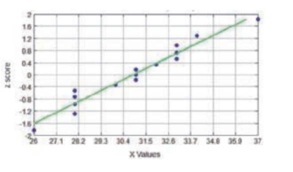
e) Make an interpretation of the confidence interval estimate you found.

f) Make an interpretation of the confidence level of the interval you found.

Estimates for μ

Example

Quantile plot:



Determining Sample Size for μ

Example

Assume that we want to estimate the mean IQ score for the population of adults who smoke, for which the population standard deviation is 15.

How many smokers must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

Practice

Look at the exercises at the end of Section 7-1 in page 295

Specially, look at exercises: 1 to 8, 13, 15, 18, 21, 26, 29, 31, 33

Look at the exercises at the end of Section 7-2 in page 309

Specially, look at exercises: 9, 10, 11, 13, 17, 18, 24, 28, 30, 34

Hypothesis Testing

Chapter 8

Basics of Hypothesis Testing and Tests for proportions

Sections 8-1 and 8-2

Hypothesis Testing

- We use Hypothesis Testing to make claims about population parameters:
  - "The proportion of students that passes a class is greater than 0.8."
  - "The mean score of the test is greater than 22."
  - "The standard deviation of of the test is smaller than 0.7."
- The structure of the claims can be of three types:
  - The population parameter is greater than a specific value.
  - The population parameter is smaller than a specific value.
  - The population parameter is equal to a specific value.
- Estimates (point and interval) and hypothesis tests are both methods of inferential statistics, but they have different objectives.
- In what follows, we will discuss the general aspects of hypothesis test, but will center our attention to the details of hypothesis tests for a proportion.

Hypothesis Testing

- The **first step** in a Hypothesis Test is to identify the claim of interest and its symbolic form:
  — Claim of interest: The proportion of girls born to parents using the XSORT method of gender selection is greater that 0.5
  Symbolic form: $p > 0.5$.

- The **second step** is to provide a symbolic form for the claim that must be true, when the claim of interest is false:
  — If $p > 0.5$ is false, then $p \leq 0.5$ must be true. Symbolic form: $p \leq 0.5$.

- The **third step** is to identify the formal *null hypothesis (H_0)* and *alternative hypothesis (H_1)*:
  — **null hypothesis (H_0)** is a symbolic statement that the value of a population parameter is *equal* to some claimed value. This hypothesis indicates *no change, no effect,* or *no difference* and it always uses the symbol: $=$.
  — **alternative hypothesis (H_1)** is a statement that the parameter has a value that somehow differs from H_0. In general, the scientist's claim must be worded so that it becomes H_1 and its symbolic form uses the symbols: $>, <, \neq$.

Hypothesis Testing

- Continuation of **third step**:
  — $H_0 : p = 0.5 \qquad H_1 : p > 0.5$

- The **fourth step** is to select the significance level α. For a hypothesis test, the **significance level,** α**,** is the probability value used as the cutoff for determining when the evidence in the sample data provides significant evidence against the null hypothesis H_0.
  The significant level α is the probability of rejecting the null hypothesis H_0 when it is true:

  $$\text{significance level } \alpha = P(\textit{rejecting } H_0 \textit{ given that } H_0 \textit{ is true}).$$

  Usually α will take the value $0.01, 0.05$, or 0.1.

Hypothesis Testing

- The **fifth step** is to identify the *Test Statistic* relevant for the test of interest and its distribution.
  The **Test Statistic** is a value used to make a decision about H_0 and it is found by converting the sample statistic ($\hat{p}$) to a score (z^{stat}) with the assumption that the null hypothesis is true.
  - For a claim regarding a proportion p:
    we assume that there is a fixed number of trials, the trials are independent, outcomes can be of one of two categories, the probabilities of success remain constant for each trial, $np \geq 5$ and $n(1-p) \geq 5$ under the assumption that H_0 is true (so the normal distribution is a suitable approximation to the binomial distribution).
    sample statistic: $\hat{p}$. Test Statistic $z^{stat} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ follows (approximately) a standard normal distribution under the assumption that H_0 is true.

Hypothesis Testing

- The **sixth step** is to compute the *Test Statistic* and then find either the p-value or the critical value(s).
  - Before finding the p-value or the critical value(s) we need to define the **critical (or rejection) region**, which is the area corresponding to all values of the test statistic that causes us to reject the null hypothesis.
  - A **critical value(s)** separates the critical region from the values of the test statistic that do not lead to rejection of the null hypothesis. Critical values depend on H_0, the sampling distribution, and the significance level.
  - A **p-value** is the probability of getting a value of the test statistic that is at least as extreme as the test statistic computed from the data, assuming that the null hypothesis is true. p-values depend on H_0, the sampling distribution, and the computed test statistic.

Hypothesis Testing

- The proportion of girls born to parents using the XSORT method of gender selection is greater that 0.5: $p > 0.5$.

$$H_0 : p = 0.5 \qquad H_1 : p > 0.5$$

From a sample of 14 babies, 13 were girls. Plot the critical region. For a significance level of 0.05, find the critical value and p-value.

Hypothesis Testing

- The proportion of girls born to parents using the XSORT method of gender selection is greater than 0.5: $p > 0.5$. $H_0 : p = 0.5$ $\quad H_1 : p > 0.5$. From a sample of 14 babies, 13 were girls and $z^{stat} = 3.21$. For a significance level of $\alpha = 0.05$, *critical value* $= 1.65$, critical region is values z^{stat} such that $z^{stat} > 1.65$, and $p - value = 0.0007$.

- The **seventh step** is to make a decision to either reject H_0 or to fail to reject H_0.
  - Decision Criteria for a Critical Value
    - If the test statistic is in the critical region, then reject H_0.
    - If the test statistic is not in the critical region, then fail to reject H_0.
  - Decision Criteria for a p-value
    - If the p-value is equal or smaller than α, then reject H_0.
    - If the p-value is larger than α, then fail to reject H_0.

Hypothesis Testing

- The proportion of girls born to parents using the XSORT method of gender selection is greater than 0.5: $p > 0.5$. $H_0 : p = 0.5$ $\quad H_1 : p > 0.5$. From a sample of 14 babies, 13 were girls and $z^{stat} = 3.21$. For a significance level of $\alpha = 0.05$, $critical\ value = 1.65$, critical region is values z^{stat} such that $z^{stat} > 1.65$, and $p - value = 0.0007$.

- The **eighth step** is to restate the decision Using Simple and Nontechnical Terms

Condition	Conclusion
Original claim does not include equality, and you reject H_0.	"There is sufficient evidence to support the claim that... (original claim)."
Original claim does not include equality, and you fail to reject H_0.	"There is not sufficient evidence to support the claim that... (original claim)."
Original claim includes equality, and you reject H_0.	"There is sufficient evidence to warrant rejection of the claim that... (original claim)."
Original claim includes equality, and you fail to reject H_0.	"There is not sufficient evidence to warrant rejection of the claim that... (original claim)."

Example

A study of sleepwalking or "nocturnal wandering" was described in the journal Neurology. From the study it was observed that 29.2% of 19,136 American adults have sleepwalked and a reporter stated that "fewer than 30% of adults have sleepwalked".

Use a 0.05 significance level to test the reporter's claim.

Practice

Look at the exercises at the end of Section 8-1 in page 351

Specially, look at exercises: 1, 2, 3, 4, 5-8, 17-20, 25-28, 29-32.

Look at the exercises at the end of Section 8-2 in page 362

Specially, look at exercises: 5-8, 9, 10, 13, 14, 16, 17, 18, 21, 22, 25.