Chapter 9: Significance Testing -- Using Data to Test Hypotheses

- Getting Started
- What do we test? Types of hypotheses
- Measuring the evidence against the null
- Hypothesis testing as decision making
- Why tests should be supplemented by intervals

ESP (extra sensory perception) or just guessing?

Deck of equal number of Zener/Rhine cards

- True value for just guessing (0.200)
- Pratt & Woodruff’s proportion (0.2082)

Sample proportions:
- 0.198
- 0.200
- 0.202
- 0.204
- 0.206
- 0.208

Can sampling variations alone account for Pratt & Woodruff’s success rate = 20.82% correct vs. 20% expected.

Was Cavendish’s experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased.

Mean density of the earth

- True value = 5.517
- Cavendish’s data: (from previous Example 7.2.2)
  - n = 23, sample mean = 5.483, sample SD = 0.1904

Simulate taking 400 sets of 23 measurements from N(5.517, 0.1904). Plotted are the results of the sample means. Is the Cavendish value apparently diff. from true mean?
Cavendish: measuring distances in std errors

20.5% of samples had $t$ values smaller than this. Cavendish data lies within the central 60% of the distribution.

Figure 9.1.3 Sample $t$-values from 400 unbiased experiments (each $t$-value is distance between sample mean and 5.517 in std errors).

20.5% of samples had $t$ values smaller than this. Cavendish $t$-value = 0.844

Figure 9.1.4 Student($df=22$) density.

Measuring the distance between the true-value and the estimate in terms of the SE

- Intuitive criterion: Estimate is credible if it’s not far away from its hypothesized true-value!
- But how far is far-away?
- Compute the distance in standard-terms: $T = \frac{\text{Estimator} - \text{True Parameter Value}}{\text{SE}}$
- Reason is that the distribution of $T$ is known in some cases (Student’s $t$, or $N(0,1)$). The estimator (obs-value) is typical/atypical if it is close to the center/tail of the distribution.

Comparing CI’s and significance tests

- These are different methods for coping with the uncertainty about the true value of a parameter caused by the sampling variation in estimates.
- **Confidence interval**: A fixed level of confidence is chosen. We determine a range of possible values for the parameter that are consistent with the data (at the chosen confidence level).
- **Significance test**: Only one possible value for the parameter, called the hypothesized value, is tested. We determine the strength of the evidence (confidence) provided by the data against the proposition that the hypothesized value is the true value.

Review

- What intuitive criterion did we use to determine whether the hypothesized parameter value ($p=0.2$ in the ESP Example 9.1.1, and $\mu = 5.517$ in Example 9.1.2) was credible in the light of the data? (Determine if the data-driven parameter estimate is consistent with the pattern of variation we’d expect get if hypothesis was true. If hypothesized value is correct, our estimate should not be far from its hypothesized true value.)
- Why was it that $\mu = 5.517$ was credible in Example 9.1.2, whereas $p=0.2$ was not credible in Example 9.1.1? (The first estimate is consistent, and the second one is not, with the pattern of variation of the hypothesized true process.)
- What do $t$-values tell us? (Our estimate is typical/atypical, consistent or inconsistent with our hypothesis.)
- What is the essential difference between the information provided by a confidence interval (CI) and by a significance test (ST)? (Both are uncertainty quantifiers. CI’s use a fixed level of confidence to determine possible range of values. ST’s one possible value is fixed and level of confidence is determined.)
Guiding principles
We cannot rule in a hypothesized value for a parameter, we can only determine whether there is evidence to rule out a hypothesized value.

The null hypothesis tested is typically a skeptical reaction to a research hypothesis.

Comments
- Why can’t we (rule-in) prove that a hypothesized value of a parameter is exactly true? (Because when constructing estimates based on data, there’s always sampling and may be non-sampling errors, which are normal, and will effect the resulting estimate. Even if we do 60,000 ESP tests, as we saw earlier, repeatedly we are likely to get estimates like 0.2 and 0.200001, and 0.199999, etc. – non of which may be exactly the theoretically correct, 0.2.)
- Why use the rule-out principle? (Since, we can’t use the rule-in method, we try to find compelling evidence against the observed/data-constructed estimate – to reject it.)
- Why is the null hypothesis & significance testing typically used? (Ho: skeptical reaction to a research hypothesis; ST is used to check if differences or effects seen in the data can be explained simply in terms of sampling variation.)

The t-test

Using \( \hat{\theta} \) to test \( H_0: \theta = \theta_0 \) versus some alternative \( H_1 \).

**STEP 1** Calculate the test statistic.

\[
T = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}
\]

(standard error)

This tells us how many standard errors the estimate is above the hypothesized value (\( \theta_0 \), positive) or below the hypothesized value (\( \theta_0 \), negative).

**STEP 2** Calculate the \( P \)-value using the following table.

**STEP 3** Interpret the \( P \)-value in the context of the data.

Alternative hypothesis| Evidence against \( H_0: \theta = \theta_0 \) provided by| \( P \)-value
---|---|---
\( H_0: \theta > \theta_0 \)| \( \hat{\theta} \) too much bigger than \( \theta_0 \) (i.e., \( \hat{\theta} \), too large)\( P = pr(T > t_{0}) \)
\( H_0: \theta < \theta_0 \)| \( \hat{\theta} \) too much smaller than \( \theta_0 \) (i.e., \( \hat{\theta} \), too negative)\( P = pr(T < t_{0}) \)
\( H_0: \theta \neq \theta_0 \)| \( \hat{\theta} \) too far from \( \theta_0 \) (i.e., \( |\hat{\theta} - \theta_0| \), too large)\( P = 2 \cdot pr(T \geq T) \)

where \( T \sim \text{Student}(df) \)
### Interpretation of the p-value

**TABLE 9.3.2 Interpreting the Size of a P-Value**

<table>
<thead>
<tr>
<th>Approximate size of P-Value</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 0.12 (12%)</td>
<td>No evidence against $H_0$</td>
</tr>
<tr>
<td>0.10 (10%)</td>
<td>Weak evidence against $H_0$</td>
</tr>
<tr>
<td>0.05 (5%)</td>
<td>Some evidence against $H_0$</td>
</tr>
<tr>
<td>0.01 (1%)</td>
<td>Strong evidence against $H_0$</td>
</tr>
<tr>
<td>0.001 (0.1%)</td>
<td>Very Strong evidence against $H_0$</td>
</tr>
</tbody>
</table>

### P-values from t-tests

- **P-value** is the probability that, if the hypothesis was true, sampling variation would produce an estimate that is further away from the hypothesized value than our data-estimate.
- The **P-value** measures the strength of the evidence against $H_0$.
- The **smaller** the P-value, the **stronger** the evidence against $H_0$.
  (The second and third points are true for significance tests generally, and not just for t-tests.)

### Review

- **What does the t-statistic tell us?**
  The t-statistic, $t_0 = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}$, tells us (in std. units) if the observed value/estimate is typical/consistent and can be explained by the variation in the sampling distribution.
- **When do we use a 2-tailed rather than a 1-tailed test?**
  We use two-sided/two-tailed test, unless there is a prior (knowledge available before data was collected) or a strong reason to believe that the result should go in one particular direction (e.g., $\mu \rightarrow$).
What were the 3 types of alternative hypothesis involving the parameter \( \theta \) and the hypothesized value \( \theta_0 \)? Write them down!

Let’s go through and construct our own \textit{t-Test} Table.

For each alternative, think through what would constitute evidence against the hypothesis and in favor of the alternative. Represent these \( P \)-values on hand-drawn curves (cf. Fig. 9.3.1). \( P = \Pr(T \geq t_0), P = \Pr(T \leq t_0), P = 2\Pr(T \geq |t_0|) \).

What does the \( P \)-value measure? (If \( H_0 \) was true, sampling variation alone would produce an estimate farther than the hypothesized value.)

What do very small \( P \)-values tell us? What do large \( P \)-values tell us? (strength of evidence against \( H_0 \).)

Pair the phrases: “the ......... the \( P \)-value, the ........ the evidence for/against the null hypothesis.”

Do large values of \( t_0 \) correspond to large or small \( P \)-values? Why?

What is the relationship between the Student (\( df \)) distribution and Normal(0,1) distribution? (identical as \( \rightarrow \))

Is a second child gender influenced by the gender of the first child, in families with >1 kid?

<table>
<thead>
<tr>
<th>TABLE 9.3.4 First and Second Births by Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Child</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>First Child</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.

Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.

Hypothesis testing as decision making

<table>
<thead>
<tr>
<th>TABLE 9.4.1 Decision Making</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision made</td>
</tr>
<tr>
<td>Actual situation</td>
</tr>
<tr>
<td>Accept ( H_0 ) as true</td>
</tr>
<tr>
<td>( H_0 ) is true</td>
</tr>
<tr>
<td>( H_0 ) is false</td>
</tr>
<tr>
<td>Reject ( H_0 ) as false</td>
</tr>
<tr>
<td>Type I error</td>
</tr>
<tr>
<td>OK</td>
</tr>
</tbody>
</table>

Sample sizes: \( n_1 = 5412, n_2 = 5978 \), Sample proportions (estimates) \( \hat{p}_1 = \frac{2792}{5412} = 0.5159, \hat{p}_2 = \frac{2776}{5978} = 0.4644 \).

\( H_0: p_1 = p_2 \) (skeptical reaction). \( H_a: p_1 > p_2 \) (research hypothesis)

Analysis of the birth-gender data – data summary

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of births</th>
<th>Number of girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Previous child was girl)</td>
<td>5412</td>
<td>2792 (approx. 51.6%)</td>
</tr>
<tr>
<td>2 (Previous child was boy)</td>
<td>5978</td>
<td>2776 (approx. 46.4%)</td>
</tr>
</tbody>
</table>

Let \( p_1 \) = true proportion of girls in mothers with girl as first child, \( p_2 \) = true proportion of girls in mothers with boy as first child. Parameter of interest is \( p_1 - p_2 \).

\( H_0: p_1 - p_2 = 0 \) (skeptical reaction). \( H_a: p_1 - p_2 > 0 \) (research hypothesis)

Analysis of the birth-gender data

\[ t = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE} = \frac{0.5159 - 0.4644}{\sqrt{\frac{0.5159 \times 0.4841}{5412} + \frac{0.4644 \times 0.5356}{5978}}} = 5.49986 \]

\[ P-value = Pr(T \geq t_0) = 1.9 \times 10^{-8} \]
Analysis of the birth-gender data

- We have strong evidence to reject the $H_0$, and hence conclude mothers with first child a girl are more likely to have a girl as a second child.
- How much more likely? A 95% CI:
  
  $CI (\hat{p}_1 - \hat{p}_2) = [0.033; 0.070]$. And computed by:
  
  $\hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{n_1 \hat{p}_1 (1-\hat{p}_1)}{n_2 \hat{p}_2 (1-\hat{p}_2)}}}\right) =
  
  0.051 \pm 1.96 \times 0.0093677 = [0.033; 0.070]$

Review

- If 120 researchers each independently investigated a it true hypothesis, how many researchers would you expect to obtain a result that was significant at the 5% level (just by chance)? (Type I, false-positive; 120*5%=6)
- What was the other type of error described? What was it called? When is the idea useful? (Type II, false-negative)
- Power of statistical test = $1 - \beta$, where
  
  $\beta = P(\text{Type II error}) = P(\text{Accepting } H_0 \text{ as true, when it's truly false})$

Sensitivity vs. Specificity of a Test

An ELISA is developed to diagnose HIV infections. Serum from 10,000 patients that were positive by Western Blot (the gold standard assay) were tested and 9990 were found to be positive by the new ELISA. The manufacturers then used the ELISA to test serum from 10,000 nuns who denied risk factors for HIV infection. 9990 were negative and the 10 positive results were negative by Western Blot.

<table>
<thead>
<tr>
<th>HIV Infected (True Case)</th>
<th>ELISA Test</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>9990 (TP)</td>
<td>99.9%</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>10 (FP, $\alpha$)</td>
<td>10 (FN, $\beta$)</td>
<td>99.9%</td>
</tr>
<tr>
<td>10,000 (TP+FN)</td>
<td>10,000 (FP+TN)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity = $\frac{TP}{TP+FN}$
Specificity = $\frac{TN}{FP+TN}$

3 Factors affecting the power

- Larger: $\rightarrow$ Causes:
  - Sample size (positive)
  - Sample variance (negative)
  - Effect size (positive)
- The chosen level for $\alpha$ (positive)

Review

- With a sensitivity of 99.9% and a specificity of 99.9%, the ELISA appears to be an excellent test. Let's apply this test to a million people where 1% are infected with HIV. Of the million people, 10,000 would be infected with HIV. Since our ELISA is 99.9% sensitive, the test will detect 9,990 of the infected people. But there is another side to the test. Of our original one million, 990,000 are not infected. If we look at the test results on the HIV negative population, we find that 990 individuals who are found to be positive by the ELISA (false positives -- FN). If you released these test results without confirmatory tests (our gold standard Western Blot), you would have told 990 people or approximately 0.1% of the population that they are HIV infected when in reality, they are not.

Review

- Why is the expression “accept the null hypothesis” dangerous? (Data can not really provide all the evidence that a hypothesis is true, however, it can provide support that it is false. That’s why better lingo is “we can’t reject $H_0$”)
- What is meant by the word non-significant in many research literatures? (P-value > fixed-level of significance)
- In fixed-level testing, what is a Type I error? What is a Type II error? (Type I, false-positive, reject $H_0$ as false, when it’s true in reality; Type II, false-negative, accepting $H_0$ as true, when it’s truly false)
A two-sided test of $H_0: \theta = \theta_0$ is significant at the 5% level if and only if $\theta_0$ lies outside a 95% confidence interval for $\theta$.

Tests and confidence intervals

A two-sided test of $H_0: \theta = \theta_0$ gives a result that is significant at the 5% level if
\[ P-value = 2\Pr(T > |t_0|) < 0.05 \]
where
\[ t_0 = \frac{\text{estimate - Hypoth'd Value}}{\text{SE}(\theta)} \]
Let $t$ be a threshold chosen so that
\[ \Pr(T \geq t) = 0.025 \]
Now $|t_0|$ tells us how many SE’s $\theta$ and $\theta_0$ are apart (without direction in their diff.). If $|t_0| > t$, then $\theta_0$ is more than $t$ SE’s away from $\theta$ and hence lies outside the 95% CI for $\theta$.

“Significance”

Statistical significance relates to the strength of the evidence of existence of an effect.
The practical significance of an effect depends on its size – how large is the effect.
A small $P$-value provides evidence that the effect exists but says nothing at all about the size of the effect.

“Significance” cont.

A non-significant test does not imply that the null hypothesis is true (or that we accept $H_0$).
It simply means we do not have (this data does not provide) the evidence to reject the skeptical reaction, $H_0$.

To prevent people from misinterpreting your report: Never quote a $P$-value about the existence of an effect without also providing a confidence interval estimating the size of the effect.

To estimate the size of an effect (its practical significance), compute a confidence interval.

“Significance” cont.

If we read that a difference between two proportions is non-significant, what does this tell us? What does it not tell us? (Do not have evidence proportions are different, based on this data. Doesn’t mean accept $H_0$).

The closest you can get to showing that a hypothesized value is true and how could you go about it? (Suppose $H_0: \theta = \theta_0$, and our test is not significant. To show $\theta = \theta_0$, we need to show that all values in the CI $\theta_0$ are essentially equal to $\theta_0$. This is a practical subjective matter decision, not a statistical one.)
A test statistic is a measure of discrepancy between what we see in data and what we would expect to see if $H_0$ was true.

The $P$-value is the probability, calculated assuming that the null hypothesis is true, that sampling variation alone would produce data which is more discrepant than our data set.

The main use of significance testing is to check whether apparent differences or effects seen in data can be explained away simply in terms of sampling variation. The essential difference between confidence intervals and significance tests is as follows:

- **Confidence interval**: A range of possible values for the parameter are determined that are consistent with the data at a specified confidence level.
- **Significance test**: Only one possible value for the parameter, called the hypothesized value, is tested. We determine the strength of the evidence provided by the data against the proposition that the hypothesized value is the true value.

The null hypothesis, denoted by $H_0$, is the (skeptical reaction) hypothesis tested by the statistical test.

Principle guiding the formulation of null hypotheses: We cannot rule a hypothesized value in; we can only determine whether there is enough evidence to rule it out. Why is that?

Research (alternative) hypotheses lay out the conjectures that the research is designed to investigate and, if the researchers' hunches prove correct, establish as being true.

The book by Best (Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists, Joel Best) shows how we can test for racial bias in police arrests. Suppose we find that among 100 white and 100 black youths, 10 and 17, respectively, have experienced arrest. This may look plainly discriminatory. But suppose we then find that of the 80 middle-class white youths 4 have been arrested, and of the 50 middle-class black youths 2 arrested, whereas the corresponding numbers of lower-class white and black youths arrested are, respectively, 6 of 20 and 15 of 50. These arrest rates correspond to 5 per 100 for white and 4 per 100 for black middle-class youths, and 30 per 100 for both white and black lower-class youths. Now, better analyzed, the data suggest effects of social class, not race as such.
The alternative hypothesis, typically corresponds to the research hypothesis.

We use one-sided alternatives (using either: $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$) when the research hypothesis specifies the direction of the effect, or more generally, when the investigators had good grounds for believing the true value of $\theta$ was on one particular side of $\theta_0$ before the study began. Otherwise a two-sided alternative, $H_1: \theta \neq \theta_0$, is used.

Differences or effects seen in data that are easily explainable in terms of sampling variation do not provide convincing evidence that real differences or effects exist.

The $P$-value is the probability that, if the hypothesis was true, sampling variation would produce an estimate that is further away from the hypothesized value than the estimate we got from our data.

The $P$-value measures the strength of the evidence against $H_0$.

Never quote a $P$-value about the existence of an effect without also providing a confidence interval estimating the size of the effect.

Suggestions for verbal translation of $P$-values are given in Table 9.3.2.

Computation of $P$-values: Computation of $P$-values for situations in which the sampling distribution of $(\hat{\theta} - \theta_0)/\text{se}(\hat{\theta})$ is well approximated by a Student($df$) distribution or a Normal(0,1) distribution is laid out in Table 9.3.1.

The $t$-test statistic tells us how many standard errors the estimate is from the hypothesized value.

Examples given in this chapter concerned means and differences between means, proportions and differences between proportions.

In general, a test statistic is a measure of discrepancy between what we see in the data and what we would have expected to see if $H_0$ was true.

If, whenever we obtain a $P$-value less than or equal to 5%, we make a decision to reject the null hypothesis, this procedure is called testing at the 5% level of significance.

The significance level of such a test is 5%.

If the $P$-value $\leq \alpha$, the effect is said to be significant at the $\alpha$-level.

If you always test at the 5% level, you will reject one true null hypothesis in 20 over the long run.
Significance cont.

- A two-sided test of $H_0 : \theta = \theta_0$ is significant at the 5% level if and only if $\theta_0$ lies outside a 95% confidence interval for $\theta$.
- In reports on research, the word “significant” used alone often means “significant at the 5% level” (i.e. $P$-value $< 0.05$). “Non-significant”. “does not differ significantly” and even “is no different” often mean $P$-value $> 0.05$.
- A non-significant result does not imply that $H_0$ is true.

Significance cont.

- A Type I error (false-positive) is made when one concludes that a true null hypothesis is false.
- The significance level is the probability of making a Type I error.
- Statistical significance relates to having evidence of the existence of an effect.
- The practical significance of an effect depends on its size.