UCLA STAT 251

Statistical Methods for the Life and Health Sciences

- **Instructor: Ivo Dinov,**

Asst. Prof. In Statistics and Neurology

University of California, Los Angeles, Winter 2002

http://www.stat.ucla.edu/~dinov/

Hypothesis Testing

- 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?

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:

$$T = \frac{\text{Estimator} - \text{TrueParameterValue}}{\text{SE}}$$

- 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 do t_0-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.)

Hypotheses

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? (H_o: 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!)

Comments

- How can researchers try to demonstrate that effects or differences seen in their data are real? (Reject the hypothesis that there are no effects)
- How does the alternative hypothesis typically relate to a belief, hunch, or research hypothesis that initiates a study? (H_1=H_a: specifies the type of departure from the null-hypothesis, H_0 (skeptical reaction), which we are expecting (research hypothesis itself).
- In the Cavendish's mean Earth density data, null hypothesis was $H_0 : \mu = 5.517$. We suspected bias, but not bias in any specific direction, hence $H_a: \mu \neq 5.517$.

Comments

- In the ESP Pratt & Woodruff data, (skeptical reaction) null hypothesis was $H_0 : \mu = 0.2$ (pure-guessing). We suspected bias, toward success rate being higher than that, hence the (research hypothesis) $H_a: \mu > 0.2$.
- Other commonly encountered situations are:
  - $H_0 : \mu_1 - \mu_2 = 0 \quad \rightarrow \quad H_a : \mu_1 - \mu_2 > 0$
  - $H_0 : \mu_{rest} - \mu_{activation} = 0 \quad \rightarrow \quad H_a : \mu_{rest} - \mu_{activation} \neq 0$

The t-test

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

STEP 1 Calculate the ***test statistic***,

$$t_0 = \frac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})} = \frac{\text{estimate - hypothesized value}}{\text{standard error}}$$

[This tells us how many standard errors the estimate is above the hypothesized value (t_0 positive) or below the hypothesized value (t_0 negative).]

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

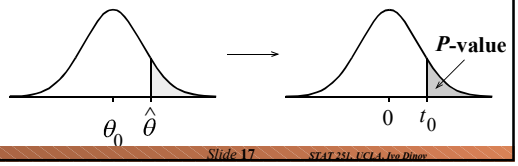
STEP 3 Interpret the *P*-value in the context of the data.

Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0

Pictorial representation of the T-test
H_0: $\theta = \theta_0$
H_1: $\theta > \theta_0$

$$t_0 = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

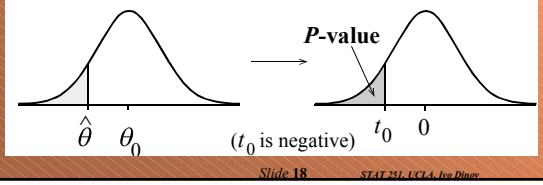
$\hat{\theta}$-scale ⟶ ***t*-scale (# of std errors)**

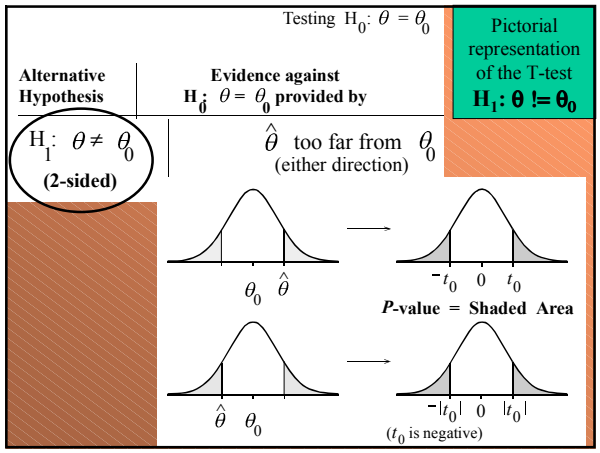


Testing H_0: $\theta = \theta_0$

Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0

Pictorial representation of the T-test
H_1: $\theta < \theta_0$





The t-test

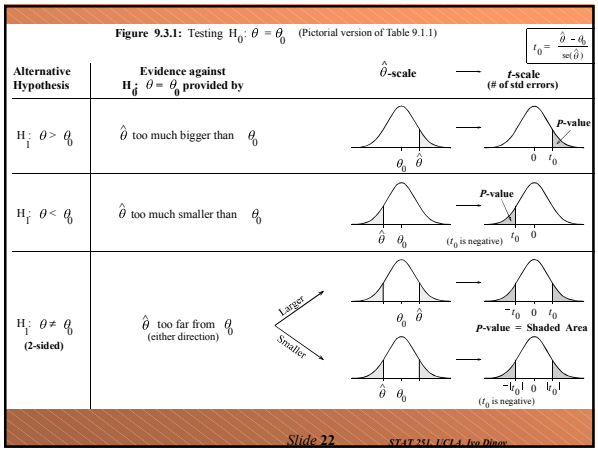
Alternative hypothesis	Evidence against H_0: $\theta > \theta_0$ provided by	*P*-value
H_1: $\theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0 (i.e., $\hat{\theta}$ - θ_0 too large)	$P = \text{pr}(T \geq t_0)$
H_1: $\theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0 (i.e., $\hat{\theta}$ - θ_0 too negative)	$P = \text{pr}(T \leq t_0)$
H_1: $\theta \neq \theta_0$	$\hat{\theta}$ too far from θ_0 (i.e., $\lvert\hat{\theta} - \theta_0\rvert$ too large)	$P = 2\,\text{pr}(T \geq \lvert t_0\rvert)$

where $T \sim$ Student(df)

Interpretation of the p-value

Interpreting the Size of a *P*-Value

Approximate size of *P*-Value	Translation
> 0.12 (12%)	**No** evidence against H_0
0.10 (10%)	**Weak** evidence against H_0
0.05 (5%)	**Some** evidence against H_0
0.01 (1%)	**Strong** evidence against H_0
0.001 (0.1%)	**Very Strong** evidence against H_0



P-values from *t*-tests

- 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 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 $\boldsymbol{H_0}$.

(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-statistics, $t_0 = \dfrac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})}$ 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 (← μ →).

Review

- What were the 3 types of alternative hypothesis involving the parameter θ and the hypothesized value θ_0? Write them down!
- Let's go through and construct our own *t-Test* Table.
  - For each alternative, think through what would constitute evidence against the hypothesis and in favor of the alternative.
  - Then write down the corresponding *P*-values in terms of t_0 and represent these *P*-values on hand-drawn curves [P=Pr(T>=t_0), P=Pr(T<=t_0), P=2Pr(T>=|t_0|) .]

Review

- What does the *P*-value measure? (If H_0 was true, sampling variation alone would produce an estimate farther then 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 $n \to \infty$)

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

First and Second Births by Sex

		Second Child		
		Male	Female	Total
First Child	Male	3,202	2,776	5,978
	Female	2,620	2,792	5,412
	Total	5,822	5,568	11,390

- 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.

Analysis of the birth-gender data – data summary

	Second Child	
Group	Number of births	Number of girls
1 (Previous child was girl)	5412	2792 (approx. 51.6%)
2 (Previous child was boy)	5978	2776 (approx. 46.4%)

- 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)

Hypothesis testing as decision making

TABLE 9.4.1 Decision Making

	Actual situation	
Decision made	H_0 is true	H_0 is false
Accept H_0 as true	OK	Type II error
Reject H_0 as false	Type I error	OK

- Sample sizes: n_1=5412, n_2=5978, Sample proportions (estimates) $\hat{p}_1 = 2792/5412 \approx 0.5159, \hat{p}_2 = 2776/5978 \approx 0.4644$,
- H_0: $p_1 - p_2 = 0$ (skeptical reaction). H_a: $p_1 - p_2 > 0$ (research hypothesis)

Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally diff. mothers they are independent → use formula 8.5.5.a

$$t_0 = \frac{\text{Estimate} - \text{HypothesizedValue}}{SE} = 5.49986 = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE\left(\hat{p}_1 - \hat{p}_2\right)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} =$$

$$P - value = \Pr(T \geq t_0) = 1.9 \times 10^{-8}$$

Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally diff. mothers they are independent → use formula

		Second Child		
		Male	Female	Total
First Child	Male	3,202	2,776	5,978
	Female	2,620	2,792	5,412
	Total	5,822	5,568	11,390

$p_1 - p_2$

$$t_0 = \frac{\text{Estimate} - \text{HypothesizedValue}}{SE} = 5.49986 =$$

$$P - value = \Pr(T \geq t_0) = 1.9 \times 10^{-8}$$

Analysis of the birth-gender data

- We have strong evidence to reject the H_0, and hence conclude mothers with first child a girl a more likely to have a girl as a second child.
- How much more likely? A 95% CI:

CI ($p_1 - p_2$) =[0.033; 0.070]. And computed by:

$$\text{estimate} \pm z \times \text{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$0.0515 \pm 1.96 \times 0.0093677 = [3\%; 7\%]$$

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-β, where

β =P(Type II error) = P(Accepting Ho as true, when its truly false)

Tests and confidence intervals

A ***two-sided*** test of H_0: $\theta = \theta_0$ is ***significant*** at the 5% level **if and only if** θ_0 lies ***outside*** a 95% confidence interval for θ.

A ***two-sided*** test of H_0: $\theta = \theta_0$ gives a result that is significant at the 5% level **if** the P-value=2Pr(T >=$|t_0|$) < 0.05. Where t_0 =(estimate-Hypoth'dValue)/SE(θ) ➔ $t_0 = (\hat{\theta} - \theta_0)/SE(\theta)$. Let **t** be a **threshold** chosen so that Pr(T>= **t**) = 0.025. Now $|t_0|$ tells us now many SE's $\hat{\theta}$ *and* θ are apart (without direction in their diff.) If $|t_0| >$ **t**, then θ_0 is more than **t** SE's away from $\hat{\theta}$ and hence lies outside the 95% CI for θ.

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

"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.

General ideas of "test statistic" and "*p*-value"

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.