

UCLA STAT 13
**Introduction to Statistical Methods for
the Life and Health Sciences**

• **Instructor:** Ivo Dinov,
Asst. Prof. In Statistics and Neurology

• **Teaching Assistants:**
Ming Zheng, Annie Che
UCLA Statistics

University of California, Los Angeles, Winter 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13, UCLA, Ivo Dinov Slide 1

**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

STAT 13, UCLA, Ivo Dinov Slide 2

ESP (extra sensory perception) or just guessing?

Deck of equal number of Zener/Rhine cards

n=60,000 random draws resulting in 12,489 correct guesses

True value for just guessing (0.200)

Pratt & Woodruff's proportion (0.2082)

Sample proportions from 7 just-guessing games

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

Slide 3 STAT 13, UCLA, Ivo Dinov

ESP or just guessing?

Computer simulation making 60,000 guesses with 20% chance of correct guess.

Figure 9.1.1 Sample proportions from 400 "just-guessing" experiments.

From Chance Encounters by C.J. Wild and G.A.P. Seber, © John Wiley & Sons, 2000.

Slide 4 STAT 13, UCLA, Ivo Dinov

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)

5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.10,
5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85

n = 23, sample mean = 5.483, sample SD = 0.1904

Slide 5 STAT 13, UCLA, Ivo Dinov

Was Cavendish's experiment biased?

21.5% of the means were smaller than this

Simulate taking 400 sets of 23 measurements from N(5.517, 0.1904). Plotted are the results of the sample means.

Are the Cavendish values unusually diff. from true mean?

Figure 9.1.2 Sample means from 400 sets of observations from an unbiased experiment.

Slide 6 STAT 13, UCLA, Ivo Dinov

Cavendish: measuring distances in std errors

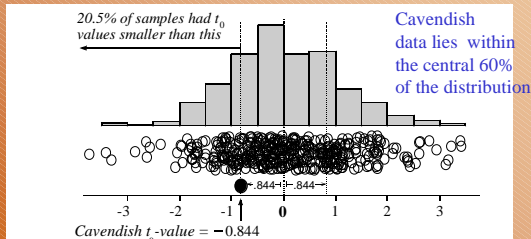


Figure 9.1.3 Sample t_0 -values from 400 unbiased experiments (each t_0 -value is distance between sample mean and 5.517 in std errors).

Slide 7 STAT 13, UCLA, Jon Dineen

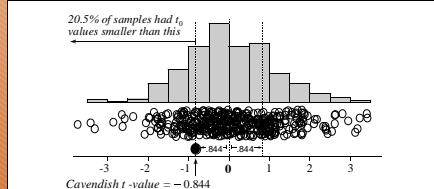


Figure 9.1.3 Sample t_0 -values from 400 unbiased experiments (each t_0 -value is distance between sample mean and 5.517 in std errors).

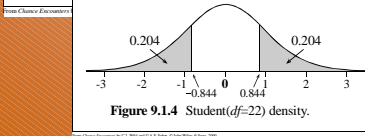


Figure 9.1.4 Student($df=22$) density.

Slide 8 STAT 13, UCLA, Jon Dineen

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.

Slide 9 STAT 13, UCLA, Jon Dineen

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.

Slide 10 STAT 13, UCLA, Jon Dineen

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.)

Slide 11 STAT 13, UCLA, Jon Dineen

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.)

Slide 12 STAT 13, UCLA, Jon Dineen

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

Slide 13 STAT 13, UCLA, Joe Dinger

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_0 : 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!)

Slide 14 STAT 13, UCLA, Joe Dinger

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 $H_a: \mu \neq 5.517$.

Slide 15 STAT 13, UCLA, Joe Dinger

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:
 - $H_0 : \mu_1 - \mu_2 = 0 \rightarrow H_a : \mu_1 - \mu_2 > 0$
 - $H_0 : \mu_{rest} - \mu_{activation} = 0 \rightarrow H_a : \mu_{rest} - \mu_{activation} \neq 0$

Slide 16 STAT 13, UCLA, Joe Dinger

The t-test

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

STEP 1 Calculate the *test statistic*,

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

[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.

Slide 17 STAT 13, UCLA, Joe Dinger

The t-test

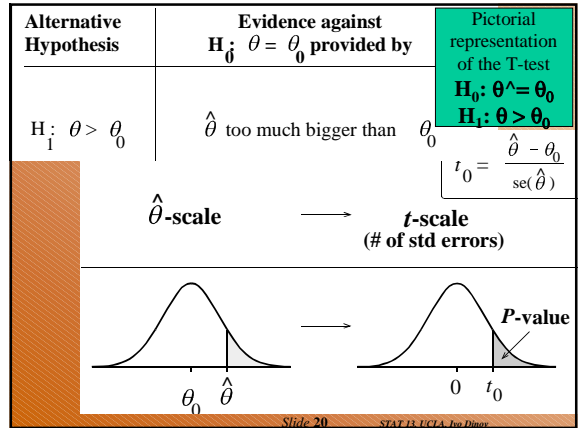
Alternative hypothesis	Evidence against $H_0: \theta = \theta_0$ provided by	P-value
$H_1: \theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0 (i.e., $\hat{\theta} - \theta_0$ too large)	$P = \text{pr}(T \geq t_0)$
$H_1: \theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0 (i.e., $\hat{\theta}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

where $T \sim \text{Student}(df)$

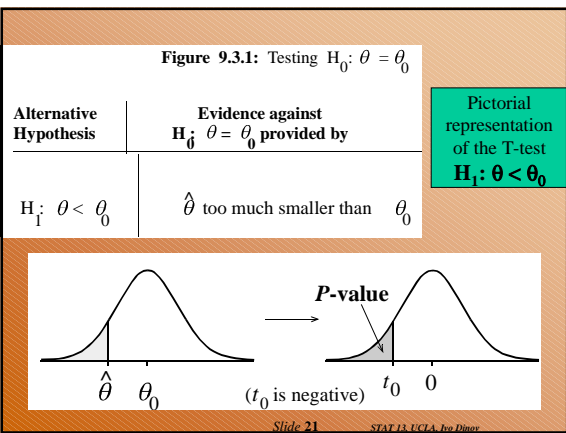
Slide 18 STAT 13, UCLA, Joe Dinger

Interpretation of the p-value	
TABLE 9.3.2 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

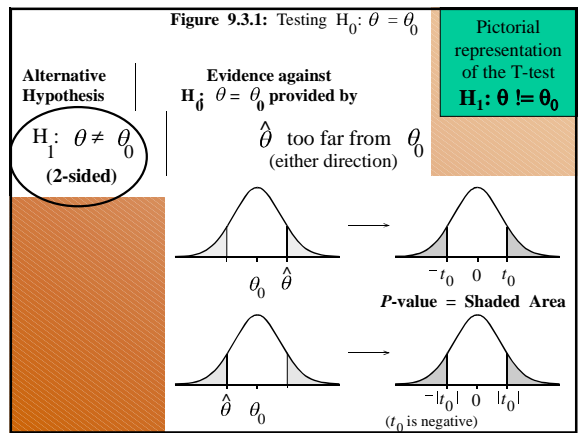
Slide 19 STAT 13, UCLA, Jon Dineen



Slide 20 STAT 13, UCLA, Jon Dineen



Slide 21 STAT 13, UCLA, Jon Dineen



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 H_0 .
(The second and third points are true for significance tests generally, and not just for t-tests.)

Slide 24 STAT 13, UCLA, Jon Dineen

Review

- What does the t-statistic tell us?**
The T-statistics, $t_0 = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$ 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 ($\leftarrow \mu \rightarrow$).

Slide 25 STAT 13, UCLA, Jon Dineen

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 (cf. Fig. 9.3.1). [$P = \Pr(T > t_0)$, $P = \Pr(T < -t_0)$, $P = 2\Pr(T > |t_0|)$.]

Slide 26 STAT 13, UCLA, Joe Dimez

Review

- 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 $\uparrow\downarrow$ the *P*-value, the $\uparrow\downarrow$ 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 $n \rightarrow \infty$)

Slide 27 STAT 13, UCLA, Joe Dimez

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

TABLE 9.3.4 First and Second Births by Sex

First Child	Second Child		Total
	Male	Female	
	3,202	2,776	
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.

Slide 29 STAT 13, UCLA, Joe Dimez

Analysis of the birth-gender data – data summary

Group	Second Child	
	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)

Slide 30 STAT 13, UCLA, Joe Dimez

Hypothesis testing as decision making

TABLE 9.4.1 Decision Making

Decision made	Actual situation	
	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)

Slide 31 STAT 13, UCLA, Joe Dimez

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 \rightarrow use formula 8.5.5.a

$$t_0 = \frac{\text{Estimate} - \text{Hypothesized Value}}{SE} = 5.49986 = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0.5159 - 0.4644}{\sqrt{\frac{0.5159(1-0.5159)}{5412} + \frac{0.4644(1-0.4644)}{5978}}} = 1.9 \times 10^{-8}$$

Slide 32 STAT 13, UCLA, Joe Dimez

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:

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

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

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

Slide 33 STAT 13, UCLA, Jon Dineen

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; $120 \times 5\% = 6$)
- 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 $\beta = P(\text{Type II error}) = P(\text{Accepting } H_0 \text{ as true, when its truly false})$

Slide 36 STAT 13, UCLA, Jon Dineen

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

		HIV Infected (True Case)	
		+	-
ELISA Test	+	9990 (TP)	10 (FP, α)
	-	10 (FN, β)	9990 (TN)
		10,000 (TP+FN)	10,000 (FP+TN)
		Sensitivity = TP/(TP+FN) 9990/(9990+10) = 0.999	Specificity = TN/(FP+TN) 9990/(9990+10) = 0.999

Slide 37 STAT 13, UCLA, Jon Dineen

4 Factors affecting the power

- **Larger:** → **Causes:**
- **Sample size (positive)**
- **Sample variance (negative)**
- **Effect size (positive)**
- **The chosen level for α (positive)**

Slide 38 STAT 13, UCLA, Jon Dineen

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 (true positives -- TP) people who are actually infected and miss 10 (false negative -- FN). Looking at those numbers, we would think that our test is very good because we have detected 9990 out of 10,000 HIV 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 (remember the specificity of the assay is 99.9%), we find that 989,010 are found to be not infected by the ELISA (true negatives -- TN), but we have 990 individuals who are found to be positive by the ELISA (false positives -- FP). 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.

Slide 39 STAT 13, UCLA, Jon Dineen

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 its truly false)

Slide 40 STAT 13, UCLA, Jon Dineen

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 $P\text{-value} = 2\Pr(T > |t_0|) < 0.05$. Where $t_0 = (\text{estimate} - \text{Hypothesized Value}) / SE(\theta) \rightarrow t_0 = (\hat{\theta} - \theta_0) / SE(\hat{\theta})$. Let t be a **threshold** chosen so that $\Pr(T > t) = 0.025$. Now $|t_0|$ tells us how many SE's $\hat{\theta}$ and θ_0 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 .

Slide 41 STAT 13, UCLA, Joe Dimez

“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**.

Slide 42 STAT 13, UCLA, Joe Dimez

“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.

Slide 43 STAT 13, UCLA, Joe Dimez

Review

- What is the relationship between a **95% confidence interval for a parameter θ** and the results of a **two-sided test of $H_0: \theta = \theta_0$** ? (θ_0 is inside the 95% CI(θ), $\leftarrow \rightarrow P\text{-value for the test is } > 0.025$. Conversely, the test is significant, at 5%-level, $\leftrightarrow \theta_0$ is outside the 95% CI(θ).
- If you read, “research shows that is significantly **bigger** than”, what is a likely explanation? (there is evidence that a real effect exists to make the two values different).
- If you read, “research says that ^{drug}..... makes no difference to”, what is a likely explanation? (the data does not have the evidence to reject the skeptical reaction, H_0 or no effects).

Slide 44 STAT 13, UCLA, Joe Dimez

Review

- Is a “significant difference” necessarily large or practically important? Why? (No, significant difference indicates the existence of an effect, practical importance depends on the effect-size.)
- What is the difference between **statistical significance and practical significance**? (stat-significance relates to the strength of the evidence that a real effect exists (e.g., that true difference is not exactly 0); practical significance indicates how important the observed difference is in practice, how large is the effect.)
- What does a P -value tell us about the size of an effect? (P -value says whether the effect is significant, but says nothing about its size.)
- What tool do we use to gauge the size of an effect? (CI(parameter) provides clues to the size of the effect.)

Slide 45 STAT 13, UCLA, Joe Dimez

Review

- 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).
- What is 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(θ) are essentially equal to θ_0 , this is a practical subjective matter decision, not a statistical one.)

Slide 46 STAT 13, UCLA, Joe Dimez

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.

Slide 47 STAT 13, UCLA, Jon Dineen

Chapter 9 Summary

STAT 13, UCLA, Jon Dineen Slide 48

Significance Tests vs. Confidence Intervals

- 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.

Slide 49 STAT 13, UCLA, Jon Dineen

Hypotheses

- 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.

Slide 50 STAT 13, UCLA, Jon Dineen

Example: Is there racial profiling or are there confounding explanatory effects?!

- 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**.

Slide 51 STAT 13, UCLA, Jon Dineen

Hypotheses cont.

- The *null hypothesis* tested is typically a skeptical reaction to the research hypothesis.
- The most commonly tested null hypotheses are of the “it makes no difference” variety.
- Researchers try to demonstrate the existence of real treatment or group differences by showing that the idea that there are no real differences is implausible.
- The *alternative hypothesis*, denoted by H_1 , specifies the type of departure from the null hypothesis, H_0 , that we expect to detect.

Slide 52 STAT 13, UCLA, Jon Dineen

Hypotheses cont.

- 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 θ was on one particular side of θ_0 before the study began. Otherwise a **two-sided alternative**, $H_1: \theta \neq \theta_0$, is used.

Slide 53 STAT 13, UCLA, Jon Dinger

P-values

- 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** .

Slide 54 STAT 13, UCLA, Jon Dinger

P-values cont.

- The **smaller** the P-value, the stronger the evidence against H_0 .
- A large P-value provides no evidence against the null hypothesis.
- A large P-value does **not** imply that the null hypothesis is true.
- 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, **compute a confidence interval**.

Slide 55 STAT 13, UCLA, Jon Dinger

P-values cont.

- 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 $(\hat{\theta} - \theta_0) / se(\hat{\theta})$, 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.

Slide 56 STAT 13, UCLA, Jon Dinger

P-values

- 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.

Slide 57 STAT 13, UCLA, Jon Dinger

Significance

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

Slide 58 STAT 13, UCLA, Jon Dinger

Significance cont.

- 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 .
- 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.

Slide 59 STAT 13, UCLA, Joe Blitz

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*.

Slide 60 STAT 13, UCLA, Joe Blitz