UCLA STAT 110A
Applied Statistics

- **Instructor:** Ivo Dinov,
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistant:** Helen Hu, UCLA Statistics

University of California, Los Angeles, Spring 2002
http://www.stat.ucla.edu/~dinov/

Examples – Birthday Paradox

- The Birthday Paradox: In a random group of N people, what is the change that at least two people have the same birthday?
- E.x., if N=23, P>0.5. Main confusion arises from the fact that in real life we rarely meet people having the same birthday as us, and we meet more than 23 people.
- The reason for such high probability is that any of the 23 people can compare their birthday with any other one, not just you comparing your birthday to anybody else's.
- There are N-Choose-2 = 20*19/2 ways to select a pair or people. Assume there are 365 days in a year, P(one-particular-pair-same-B-day)=1/365, and
- P(one-particular-pair-failure)=1-1/365 ~ 0.99726.
- For N=20, 20-Choose-2 = 190. E={No 2 people have the same birthday is the event all 190 pairs fail (have different birthdays)}, then P(E) = P(failure)^{190} = 0.99726^{190} = 0.59.
- Hence, P(at-least-one-success)=1-0.59=0.41, quite high.
- Note: for N=42 → P>0.9 …

Significance Testing --
Using Data to Test Hypotheses

- What do we test? Types of hypotheses
- Measuring the evidence against the Null
- Hypothesis testing as decision making tool
- Why tests should be supplemented by intervals?
- Test Statistics & P-values

Was Cavendish's experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased. Goal now is to test whether the Cavendish data below really supports the true mean density of the Earth.

Mean density of the earth: True value = 5.517

Cavendish's data:

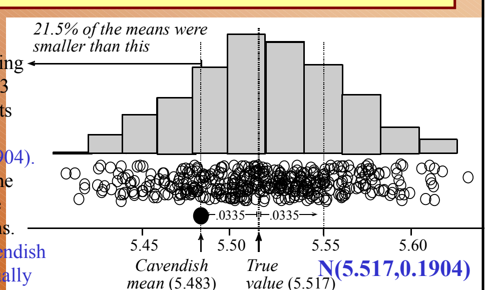
{ 5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.10, 5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85 }

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

Was Cavendish's experiment biased?

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

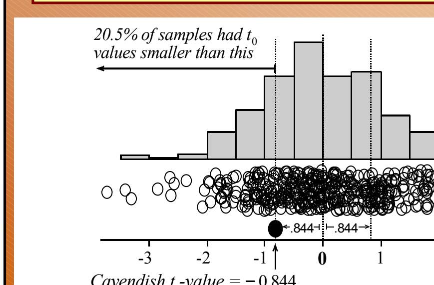
Are the Cavendish values unusually diff. from true mean?



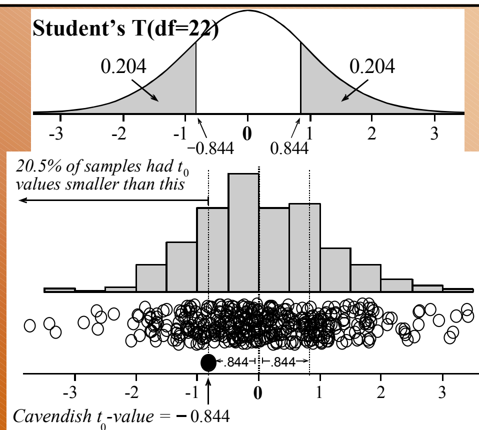
Sample means from 400 sets of observations from an unbiased experiment.

Cavendish: measuring distances in std errors

Cavendish data lies within the central 60% of the distribution



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



Another example – Carbon content in Steel

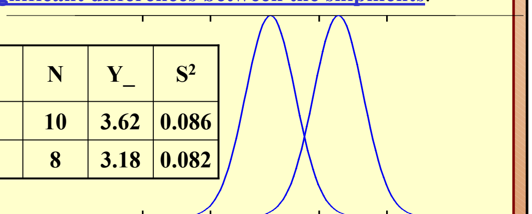
Percentage of ***C*** (Carbon) in 2 random samples taken from 2 steel shipments are measured and summarized below. The question is to **determine if there are statistically significant differences between the shipments**.

Shipment	N	Y_bar	S^2
1	10	3.62	0.086
2	8	3.18	0.082

Another example – Carbon content in Steel

Percentage of ***C*** (Carbon) in 2 random samples taken from 2 steel shipments are measured and summarized below. The question is to **determine if there are statistically significant differences between the shipments**.

#	N	Y_	S^2
1	10	3.62	0.086
2	8	3.18	0.082



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

- 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 $\boldsymbol{T}$ 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 against the data. We determine the *strength of the evidence* (confidence) provided by the data against the proposition that the hypothesized value is the true value.

Review

- Why was it that the true mean-density of the Earth, $\mu = 5.517$, was credible in terms of the Cavendish data of 23 observations with **sample mean = 5.483**?
- Since, the two values are only **~0.84 SD's** away from each other!

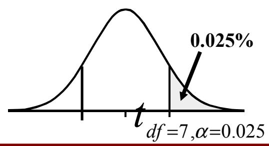
Review

- Are the **carbon** contents in the two steel shipments any <u>different</u>?

#	N	Y_	S²
1	10	3.62	0.086
2	8	3.18	0.082

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{SD(\bar{Y}_1 - \bar{Y}_2)} = \frac{3.62 - 3.18}{\sqrt{0.086^2 + 0.082^2}} = \frac{0.44}{0.12} = 3.7 \sim t_{df=7}$$

$$t_{df=7,\alpha=0.025} = 2.3646$$



Hypotheses

Guiding principles

We <u>cannot</u> **rule in** a hypothesized value for a parameter, we *can only* determine whether there is <u>evidence, provided by the data, *to* **rule out** a hypothesized value</u>.

The ***<u>null hypothesis</u>*** <u>tested is typically a skeptical reaction</u> 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.)

Comments

- 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:μ!=5.517.

Comments

- Typically, the null (skeptical) hypothesis is:

  $H_0 : \mu = \mu_0$. And the alternative is H_a:μ>0.2.
- Other commonly encountered situations are:
  - $H_0 : \mu_1 - \mu_2 = 0$ → $H_a : \mu_1 - \mu_2 > 0$
  - $H_0 : \mu_{rest} - \mu_{activation} = 0$ → $H_a : \mu_{rest} - \mu_{activation}$!=0

The t-test

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

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

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

[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.

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., $\lvert\hat{\theta} - \theta_0\rvert$ too large)	$P = 2\ \text{pr}(T \geq \lvert t_0\rvert)$

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

Interpretation of the p-value

Interpreting the Size of a *P*-Value

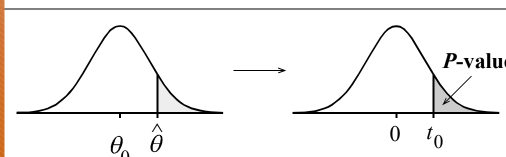
Approximate size of *P*-Value	Translation
> 0.12 (12%)	**No** evidence against H_0
0.10 (10%)	**Weak** evidence against H_0
0.05 (5%)	**Some** evidence against H_0
0.01 (1%)	**Strong** evidence against H_0
0.001 (0.1%)	**Very Strong** evidence against H_0

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: $\theta > \theta_0$

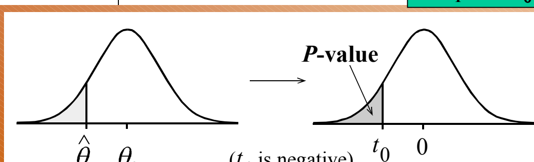
$$t_0 = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

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



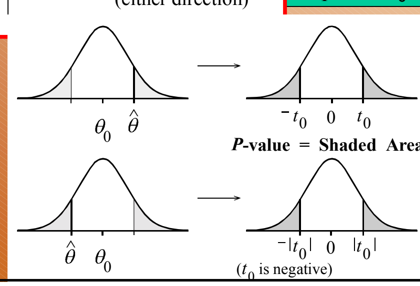
Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0

Pictorial representation of the T-test
H_0: $\theta = \theta_0$
vs.
H_1: $\theta < \theta_0$



Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta \neq \theta_0$ (2-sided)	$\hat{\theta}$ too far from θ_0 (either direction)

Pictorial representation of the T-test
H_0: $\theta = \theta_0$
vs.
H_1: θ != θ_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 <u>strength of the evidence against H_0</u>.
- The ***smaller*** the *P-value*, the ***stronger*** the evidence against $\boldsymbol{H_0}$.

Review

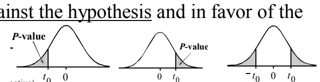
- What does the *t*-statistic tell us?

  The T-statistics, $t_0 = \dfrac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$ 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 (← 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 <u>evidence against the hypothesis</u> 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?

1st Child	Second Child Gender: Male	Female	Total
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. <u>**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**</u>.
- 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. And p_2=true proportion of girls in mothers with boy as first child. <u>Parameter of interest is $p_1 - p_2$</u>.
- H_0: $p_1 - p_2 = 0$ (skeptical reaction). H_a: $p_1 - p_2 > 0$ (research hypothesis)

Hypothesis testing as 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

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

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

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

Analysis of the birth-gender data

- 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.
- **Practical vs. Statistical significance**:
- How much more likely? A 95% CI:

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

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

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

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

Analysis of Carbon in Steel Data

- Percentage of ***C*** (Carbon) in 2 random samples taken from 2 steel shipments are measured and summarized below. The question is to **determine if there are statistically significant differences between the shipments**.

#	N	Y_	S²
1	10	3.62	0.086
2	8	3.18	0.082

$$t_0 = \frac{\text{Est\_1 - Est\_2 - 0}}{SE} =$$

$$\frac{3.62 - 3.18}{SE(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{0.44}{\sqrt{0.086/10 + 0.082/8}} = 3.12$$

Comparing two means for independent samples

Suppose we have 2 samples/means/distributions as follows: $\{\bar{x}_1, N(\mu_1, \sigma_1)\}$ and $\{\bar{x}_2, N(\mu_2, \sigma_2)\}$. We've seen before that to make inference about $\mu_1 - \mu_2$ we can use a T-test for H_0: $\mu_1 - \mu_2 = 0$ with $t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE(\bar{x}_1 - \bar{x}_2)}$

And CI($\mu_1 - \mu_2$) = $\bar{x}_1 - \bar{x}_2 \pm t \times SE(\bar{x}_1 - \bar{x}_2)$

If the 2 samples are independent we use the SE formula

$$SE = \sqrt{s_1^2/n_1 + s_2^2/n_2} \quad \text{with } df = Min(n_1 - 1; n_2 - 1)$$

This gives a conservative approach for hand calculation of an approximation to the what is known as the Welch procedure, which has a complicated exact formula.

Means for independent samples – equal or unequal variances?

Pooled T-test is used for samples with assumed equal variances. Under data Normal assumptions and equal variances of $(\bar{x}_1 - \bar{x}_2 - 0)/SE(\bar{x}_1 - \bar{x}_2)$, where

$$SE = s_p\sqrt{1/n_1 + 1/n_2};\ s_p^2 = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

is exactly Student's t distributed with $df = (n_1 + n_2 - 2)$

Here s_p is called the pooled estimate of the variance, since it pools info from the 2 samples to form a combined estimate of the single variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Comparing two means for independent samples

1. How sensitive is the two-sample *t*-test to non-Normality in the data? (The 2-sample T-tests and CI's are even more robust than the 1-sample tests, against non-Normality, particularly when the shapes of the 2 distributions are similar and $n_1=n_2=n$, even for small n, remember $df= n_1+n_2-2$.
3. Are there nonparametric alternatives to the *two-sample t-test*? (Wilcoxon rank-sum-test, Mann-Witney test, equivalent tests, same P-values.)
4. What difference is there between the quantities tested and estimated by the two-sample *t*-procedures and the nonparametric equivalent? (Non-parametric tests are based on ordering, not size, of the data and hence use median, not mean, for the average. The equality of 2 means is tested and $CI(\tilde{\mu_1} - \tilde{\mu_1})$.

Paired Comparisons

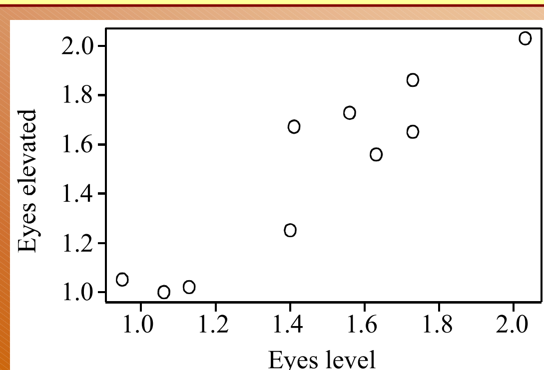
- Sometimes we have two data sets, which are not independent, but rather observations matched in pairs.
- Ex: Kaufman & Rock study of the Moon size illusion. Does the moon size appear different with eyes level and with eyes raised? Does eye position make a difference? Eyes elevated refers to raising the eye from horizontal to zenith position. [10 Subjects are tested under eye-level (control) condition, by physically moving the subject's body from level to zenith position with fixed eye direction – horizontal. Ratios of the Moon size in level and zenith positions, for the two paradigms are given below.]

Moon illusion Data

Subject	Eyes Elevated	Eyes Level	Difference (Elevated - Level)
1	2.03	2.03	0.00
2	1.65	1.73	-0.08
3	1.00	1.06	-0.06
4	1.25	1.40	-0.15
5	1.05	0.95	0.10
6	1.02	1.13	-0.11
7	1.67	1.41	0.26
8	1.86	1.73	0.13
9	1.56	1.63	-0.07
10	1.73	1.56	0.17

Source: Kaufman and Rock [1962].

Plotting Eyes elevated rations vs. eyes level rations



Looking for an effect due to elevating eyes

For ***paired*** data, ***analyze the differences***.

$H_0: \mu_{diff} = 0$

Can't reject H_0, no evidence eye position causes illusion

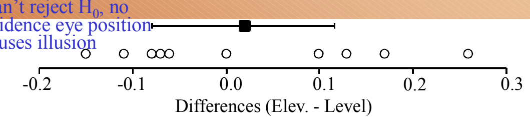


Figure 10.1.7 Dot plot of differences for the moon illusion data (with a 95% CI for the mean difference).

```
Test of mu = 0.0000 vs mu > 0.0000
Variable    N    Mean    StDev   SE Mean   t-stat  P-value
Difference  10   0.0190  0.1371  0.0434    0.44    0.34
                 95% CI ( -0.0791,  0.1171)
```


Review

- For fixed-level hypothesis tests, why are low significance levels chosen? (Large levels would imply we falsely reject the skeptical null hypothesis too often and commit Type I error! On the contrary, if our significance level is too low, preventing to reject H_0, we may end up being unlikely to reject important false hypotheses – Type II error.)
- If you wanted to perform a fixed-level hypothesis test, for what values of the *P*-value would you "reject the null hypothesis at the 1% level" ? (P<0.01)
- Give another name for the significance level. (Probability of making Type I error)

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)

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)

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.

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) ,* ←➔ *P-value for the test is >0.025,. Conversely, the test is significan, at 5%-level,* ←➔ θ_0 *is outside the 95% CI(θ)).*
- If you read, "research shows that ....θ.... is significantly bigger than ....θ0..", 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 disease treatment?", what is a likely explanation? (the data does not have the evidence to reject the skeptical reaction, H_0, or no effects).

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 exact;y 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.)

Review

- If we read that a difference between two proportions is *non-significant*, what does this tell us? What does it not tells 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_0) are essentially equal to θ_0, this is a practical subjective matter decision, not a statistical one.)

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.

Example – Roulette wheels (cont.)

- Roulette has 38 slots 18 **red**, 18 **black**, 2 **neutral**
- 100 random wheel spins → Red=58. Is there evidence of wheel bias? P(Red>=58) =? Where Y=Red ~Binomial(100, 0.47)
  - Before we showed P(Y>=58) = 0.177, using SOCR
- NOW: we use hypothesis testing:
  - H_o: p= 0.47 vs. H_1: p > 0.47
  - Test statistic is sample proportion of Reds: p^ = 0.58
  - Under H_o → p= 0.47, the P-value that P(p >= p^ = 0.58) is:

$$\sum_{k=58}^{100} \binom{100}{y} 0.47^{y} (1-0.47)^{100-y} = 0.0177$$

Summary

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.

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.

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**.

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.

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.

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.

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.*

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.

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

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.

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.

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.

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***.