Introduction to Statistical Methods SOC-GA 2332

Lecture 3: Comparing Groups and Contingency Tables

Siwei Cheng

NEW YORK UNIVERSITY

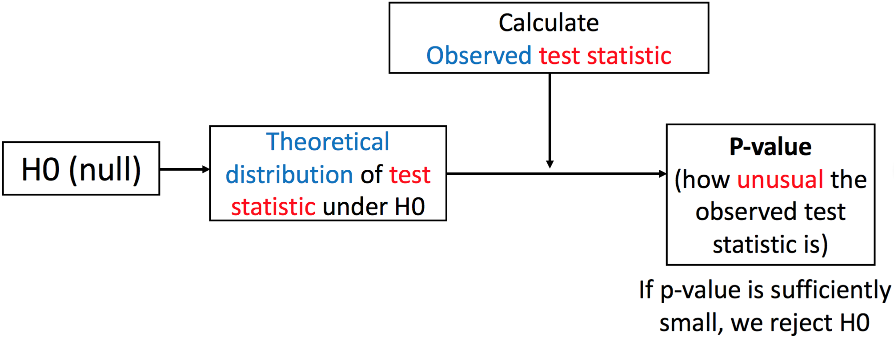
Turn on lecture recording!!

Turn on lecture recording!!
Turn on lecture recording!!
Turn on lecture recording!!

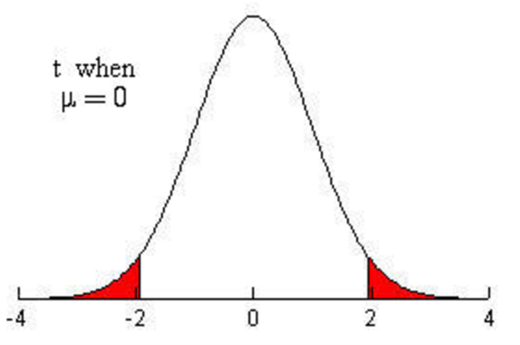
Lecture Outline

- Testing for differences between two group means (independent samples, dependent samples)
- Some notes on P-value
- Introducing the replicaiton project.

Significance Test for the Mean - A Review



Significance Test for the Mean - A Review



Note that in large samples, the test statistic distribution can be approximated by the normal distribution.

One Sample T Test Example

- Test statistics

$$t = \frac{\bar{y} - \mu_0}{se} \quad where \quad se = s/\sqrt{n}$$

- Research question: is the average temporature in February different from 0°C (=32°F)?

- **City A:** $n = 100$, $s = 2$, $\bar{y} = 0.3$.
- **City B**: $n = 100$, $s = 2$, $\bar{y} = 0.5$.
- **City C**: $n = 10000$, $s = 2$, $\bar{y} = 0.3$

Comparing Means between Two Groups

- When we compare means between two groups, we are doing a **bivariate analysis**.
- Basically, we want to test whether there is a statistically significant difference in the means of the outcome variable (response variable, dependent variable) across groups of the explanatory variable (independent variable).
- e.g. Do men and women spend equal amount of hours on housework? Do whites and blacks live in neighborhoods of the same level of socioeconomic status? Are fertility rates the same in developed versus developing countries? Are average income different in blue versus red states?

Comparing Means between Two Groups

Number of hours worked per week

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	950	45	8
Women	1020	35	10

In the sample...does $\bar{y}_1 - \bar{y}_2 = 0$?

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	n_1	$\bar{y}_1$	s_1
Women	n_2	$\bar{y}_2$	s_2

In the population...does $\mu_1 - \mu_2 = 0$?

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	-	μ_1	σ_1
Women	-	μ_2	σ_2

Independent and Dependent Samples

Independent Samples

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	950	45	8
Women	1020	35	10

Dependent Samples (matched sample, same subjects in each sample)

Sex	Sample size	Mean of hours	Sample Std Deviation
September	600	45	8
November	600	35	10

Independent and Dependent Samples

- Why do we distinguish between independent and dependent samples? The standard error formulas are different.
- Sum of **correlated** variables (or the general case):

  $Var(X + Y) = Var(X) + Var(Y) + 2 \cdot \mathbf{Cov(X,Y)}$
- Sum of **uncorrelated** variables:

  $Var(X + Y) = Var(X) + Var(Y)$

Independent and Dependent Samples

- Why do we distinguish between independent and dependent samples? The standard error formulas are different.
- Sum of **correlated** variables (or the general case):

  $Var(X + Y) = Var(X) + Var(Y) + 2 \cdot \mathbf{Cov(X,Y)}$
- Sum of **uncorrelated** variables:

  $Var(X + Y) = Var(X) + Var(Y)$
- In independent samples, the squared standard error of the distribution of test statistic is equal to the sum of the squared error in the two samples.
- But, in dependent samples, matched responses are likely to be associated. So the calculation of the standard error need to account for this association.

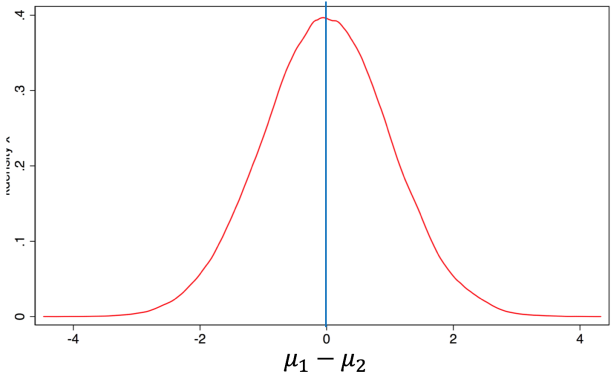
Test for Difference in Means between Two Independent Samples

The information to use is in this table:

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	n_1	$\bar{y}_1$	s_1
Women	n_2	$\bar{y}_2$	s_2

Test for Difference in Means between Two Independent Samples

- H_0: $\mu_1 - \mu_2 = 0$; H_a: $\mu_1 - \mu_2 \neq 0$
- Theoretical distribution of $\bar{y}_1 - \bar{y}_2$ if H_0 is true:



Test for Difference in Means between Two Independent Samples

- Standard error of $\bar{y}_1 - \bar{y}_2$:

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

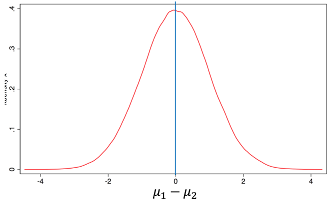
- Note that the standard error of $\bar{y}_1 - \bar{y}_2$ is larger than the standard error for either sample estimate alone.
- But the se for group comparison still decreases as the sample sizes n_1 and n_2 get larger.

Test for Difference in Means between Two Independent Samples

- Test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{se}$$

- In large sample, this statistic should follow normal distribution.
- if H_0 is true, that is, $\mu_1 - \mu_2 = 0$, the distribution of the test statistic should center around zero. And then we can use P-value to determine how unusual the observed t-statistic is.

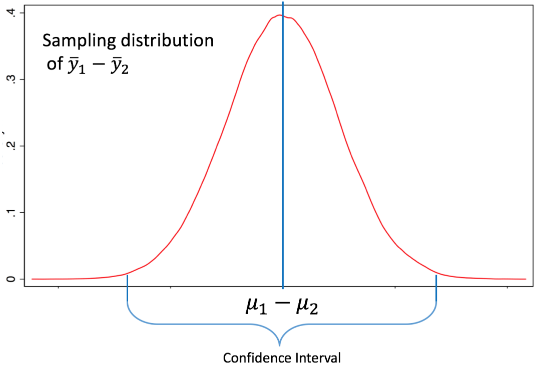


Confidence Interval

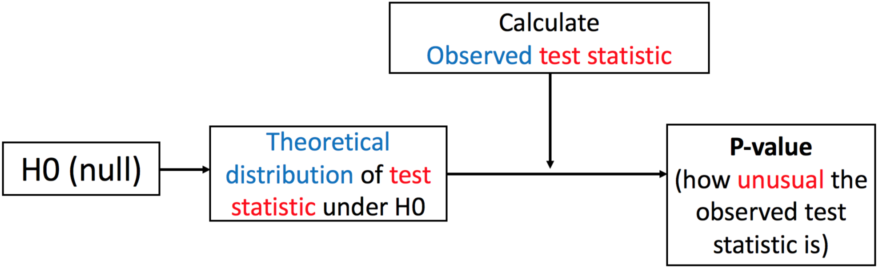
- The 95% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm 1.96 \cdot se$$

- where $se = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$



Significance Test Diagram



If p-value is sufficiently small, we reject H0

Test for Difference in Means between Two Independent Samples

Does mean number of hours worked per week differ by gender?

Sex	Sample size	Mean of hours	Sample Std Deviation
Men	950	45	8
Women	1020	35	10

Test for Difference in Means between Two Independent Samples

Does mean number of hours worked per week differ by gender?

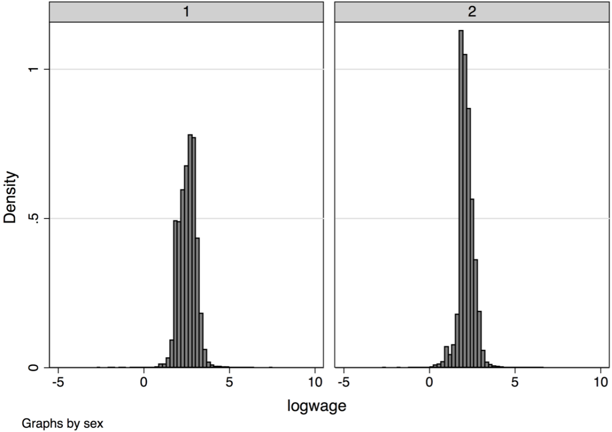
Sex	Sample size	Mean of hours	Sample Std Deviation
Men	950	45	8
Women	1020	35	10

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{(8^2)}{950} + \frac{(10)^2}{1020}} = 0.4067$$

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{se} = \frac{45 - 35 - 0}{0.4067} = 24.59$$

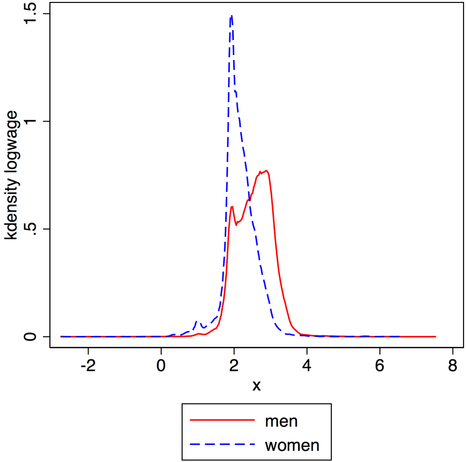
Test for Difference in Means between Two Independent Samples

Gender difference in wage?



Density by Gender

Gender difference in wage?



Test for Gender Difference in Mean Wage

Gender difference in wage?

```
. ttest logwage,by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |  42,402    2.546637    .0024391    .5022617    2.541856    2.551418
       2 |  35,736     2.15712    .0023836    .4505916    2.152448    2.161792
---------+--------------------------------------------------------------------
combined |  78,138    2.368494    .0018499    .5171096    2.364868    2.372119
---------+--------------------------------------------------------------------
    diff |            .3895172     .003442                .3827708    .3962635
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t = 113.1655
Ho: diff = 0                                     degrees of freedom =    78136

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Test for Difference in Dependent Samples

Dependent Samples (matched sample, same subjects in each sample)

Sex	Sample size	Mean of hours	Sample Std Deviation
September	600	45	8
November	600	35	10

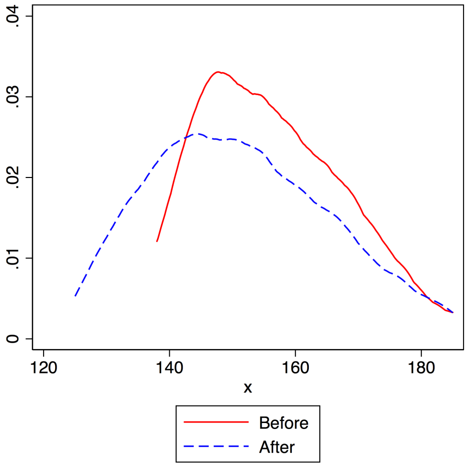
- We first calculate the difference for each paired observation in the sample: $\textit{Diff} = y_1 - y_2$.
- The mean of this difference in the sample is denoted $\bar{y}_d$
- The test statistic for testing $H_0 : \mu_1 = \mu_2$ is:

$$t = \frac{\bar{y}_d - 0}{se_d}$$

- Note that $se_d = s_d / \sqrt{n}$

Test for Difference in Dependent Samples

Blood pressure before and after a treatment...



Test for Difference in Dependent Samples

Blood pressure before and after a treatment...

```
. ttest bp_before=bp_after

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
bp_bef~e |     120      156.45    1.039746    11.38985    154.3912    158.5088
bp_after |     120    151.3583    1.294234    14.17762    148.7956     153.921
---------+--------------------------------------------------------------------
    diff |     120    5.091667    1.525736     16.7136    2.070557    8.112776
------------------------------------------------------------------------------
     mean(diff) = mean(bp_before - bp_after)                      t =   3.3372
 Ho: mean(diff) = 0                              degrees of freedom =      119

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0011          Pr(T > t) = 0.0006
```

A Look at Real Data

	patient	sex	agegrp	bp_before	bp_after
1	1	Male	30-45	143	153
2	2	Male	30-45	163	170
3	3	Male	30-45	153	168
4	4	Male	30-45	153	142
5	5	Male	30-45	146	141
6	6	Male	30-45	150	147
7	7	Male	30-45	148	133
8	8	Male	30-45	153	141
9	9	Male	30-45	153	131
10	10	Male	30-45	158	125
11	61	Female	30-45	152	149
12	62	Female	30-45	147	142
13	63	Female	30-45	144	146
14	64	Female	30-45	144	138
15	65	Female	30-45	158	131
16	66	Female	30-45	147	145
17	67	Female	30-45	154	134
18	68	Female	30-45	151	135
19	69	Female	30-45	149	131
20	70	Female	30-45	138	135

Statistical Significance versus Practical Significance

- Recall that the test statistic is written as: $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} \approx t_{n-1}$
- So, there are several possible reasons for rejecting H_0:
  1. $\bar{X} - \mu_0$ is large (big difference between sample mean and the mean under null, big effect)
  2. n is large (you have a large data, and so you have a lot of precision).
  3. s is small (the outcome has low variability).
- So, <u>in large samples even tiny effects will be significant</u>, but the results may not be very important substantively.

P-values are random variables!

P-values are random variables!

- The standard definition of the p-value is that it is "the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed." But p-value itself is also a random variable!

P-values are random variables!

- The standard definition of the p-value is that it is "the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed." But p-value itself is also a random variable!
- Recall that the test statistic is written as: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx t_{n-1}$.
- t is a statistic (calculated based on your observed sample).
- What is the p-value? The p-value is just the tail probability associated with this t value. This means that the p-value is, in itself, also a statistic based on the random sample. It is a function of data and thus has a sampling distribution.
- In fact, when the null hypothesis is true and the underlying random variable is continuous, then the probability distribution of the p-value is **uniform on the interval [0,1]**.

P-values are random variables!

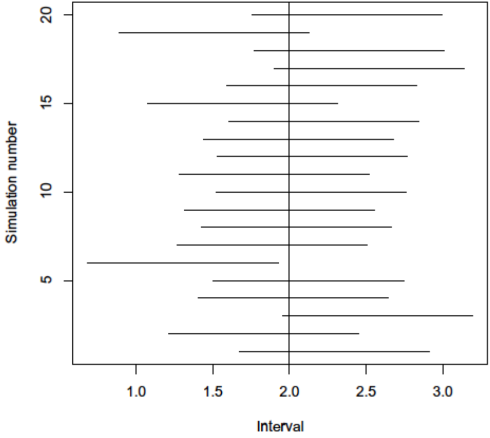


Figure 1. Twenty simulated confidence intervals around a true mean of 2.

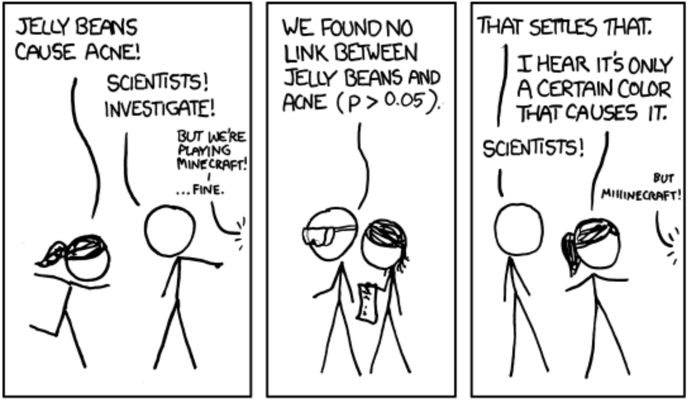
"P-hacking" or Multiple Testing

- If we test the coefficients on a lot of variables separately with a t-test, then we should expect that 5% of them will be significant at 0.05 level just due to random chance.
- Suppose we have 20 independent variables, none of which affects the outcome variable...
- Because we know that $p = 0.05$, which says that out of 20 variables that do not have any effect on the outcome variable of interest, there is, in expectation, one variable $(20 \cdot 0.05)$ that will show up as a false positive!
- We should also expect 2 out of 20 to show up as significant at 0.1 level.

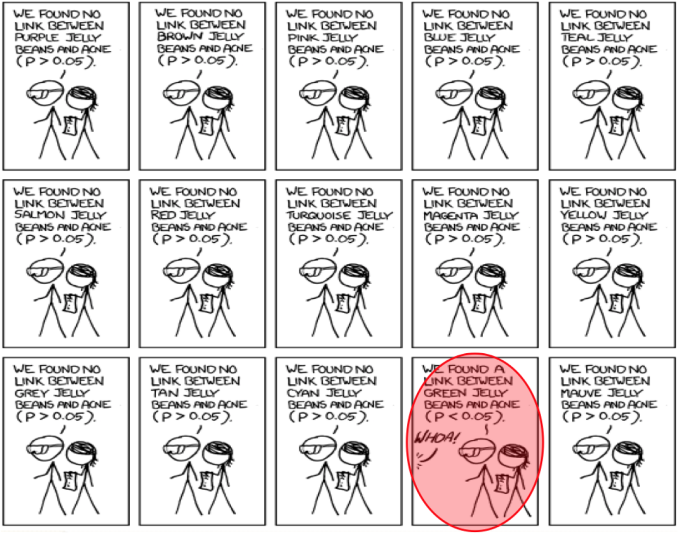
"P-hacking" or Multiple Testing

- The **multiple testing (or multiple comparison) problem** occurs when one considers a set of statistical tests simultaneously.
- Consider m independent hypothesis tests (e.g. control group versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast with k: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That is, even if all null hypotheses are true (i.e. no treatment effect for any of the treatment groups), it is still quite likely we will reject at least one of the null h

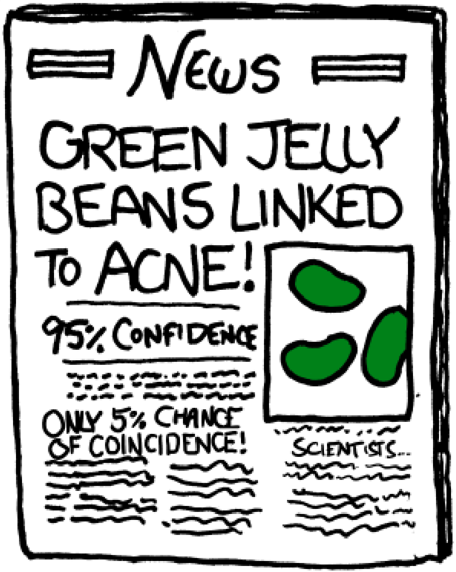
"P-hacking" or Multiple Testing



“P-hacking” or Multiple Testing



"P-hacking" or Multiple Testing

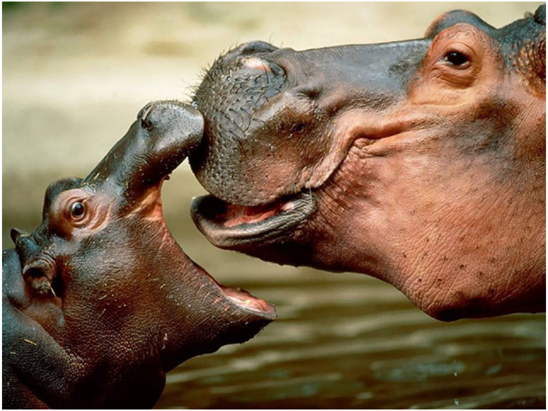


“P-hacking” or Multiple Testing

- Some techniques for adjusting the inflation of overall type I errors for multiple testing:
  1. Bonferroni: for each individual test, we use significance level of $\alpha = \alpha_k / m$.
  2. Sidak: for each individual test, use significance level of $\alpha = 1 - (1 - \alpha_k)/m$

"P-hacking" or Multiple Testing

- Some techniques for adjusting the inflation of overall type I errors for multiple testing:
    1. Bonferroni: for each individual test, we use significance level of $\alpha = \alpha_k / m$.
    2. Sidak: for each individual test, use significance level of $\alpha = 1 - (1 - \alpha_k)/m$
- AS(tatistical)A's statement on p-values: *"Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted."*
- Andrew Gelman proposed changing the sentence to: *"Valid p-values cannot be drawn without knowing, not just what was done with the existing data, but what the choices in data coding, exclusion, and analysis would have been, had the data been different."* (i.e. emphasizing not just "multiple comparisons," but also "multiple potential comparisons")



Introducing the Replication Project