Stat 102

Introduction to Business Statistics

Class 4

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Comparisons between groups
- Confidence intervals and tests for the difference in means
- Confidence interval and a test for the difference in proportions
- Summary
- Next time

Last time

Main points:

- Confidence intervals
- Hypothesis tests
- P-values

Statistical significance and practical importance

- A highly statistically significant result does not imply a practically important result.
- A very small p-value allows you to strongly reject the null hypothesis.
- But that doesn't tell you if the alternative is a *meaningful* distance from the null.
- In medicine there is an identical concept of a *clinically meaningful* effect.
- But if you just think about practical importance without looking at statistical significance you run the danger of being a perpetual *noise chaser*.
- Good decisions consider both statistical significance and practical importance.

P-values are often misinterpreted: they are **NOT**

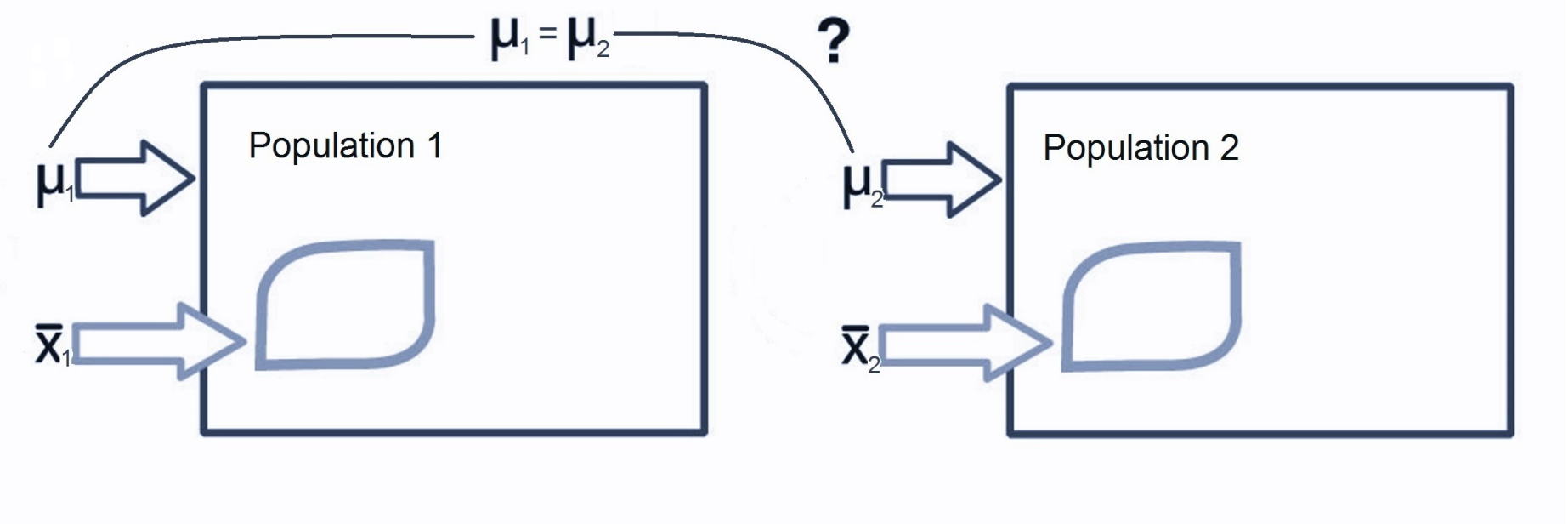
the probability that the null hypothesis is true

Motivation for two-sample comparisons

- A HR manager has heard that an employee is thinking about bringing a sex discrimination law suit against the company on the grounds that men appear to be getting paid systematically more than women.
- You have access to the company's HR database.
- Does a review of the salary records give you cause for concern?
- You want to **compare** the salaries of men to women to determine if there might be an issue.
- Hence the comparative analytics.

The population sample paradigm again, but with two populations

The key question: "is there evidence that $\mu_1 \neq \mu_2$"?'



Thinking about where the data comes from

- You have a **randomized experiment** where the group membership has been randomly assigned. Fantastic if you can do it because you can infer a causal relationship between the group membership variable and the outcome. *Having drug A rather than drug B causes you to live longer.* Randomization removes systematic biases between the two groups by making sure that all other factor are evenly distributed across the two groups.
- You have drawn **random samples** from two populations. You can infer association but not causation. Confounding is the issue – other factors may be driving the differences.
- You just **happen to have** data from two groups and you want to compare them. You have to make a number of assumptions to make conclusions, namely that these samples are representative of the populations they come from.

Comparison of two means

Is there a difference in the mean salary between men and women?
What we are about to do is test the hypothesis:

$$H_0 : (\mu_1 - \mu_2) = 0 \qquad v. \qquad H_1 : (\mu_1 - \mu_2) \neq 0.$$

In general if you had a difference of interest D_0 the hypothesis would be would be

$$H_0 : (\mu_1 - \mu_2) = D_0 \qquad v. \qquad H_1 : (\mu_1 - \mu_2) \neq D_0.$$

Confidence interval and test statistic

The parameter $(\mu_1 - \mu_2)$ is estimated with $(\overline{X_1} - \overline{X_2})$.
The approximate 95% CI is:

$$(\overline{X_1} - \overline{X_2}) \pm 2se(\overline{X_1} - \overline{X_2}).$$

The *t-statistic* for the hypothesis test is again a standard error counter:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{se(\overline{X_1} - \overline{X_2})},$$

where $se(\overline{X_1} - \overline{X_2})$ is the estimated standard error of the difference between sample means.
As before we can find the p-value of the test statistic or see if $|t| > 2$.

The standard error of the difference in means

Notice that $\overline{X_1} - \overline{X_2}$ is a linear combination of two random variables, with weights $+1$ and -1.
If the sample means are independent then we know that:

$$Var(\overline{X_1} - \overline{X_2}) = Var(\overline{X_1}) + Var(\overline{X_2}).$$

Using our knowledge of the variance of the mean of an iid sample we have

$$Var(\overline{X_1} - \overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

and

$$SE(\overline{X_1} - \overline{X_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

so that

$$se(\overline{X_1} - \overline{X_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

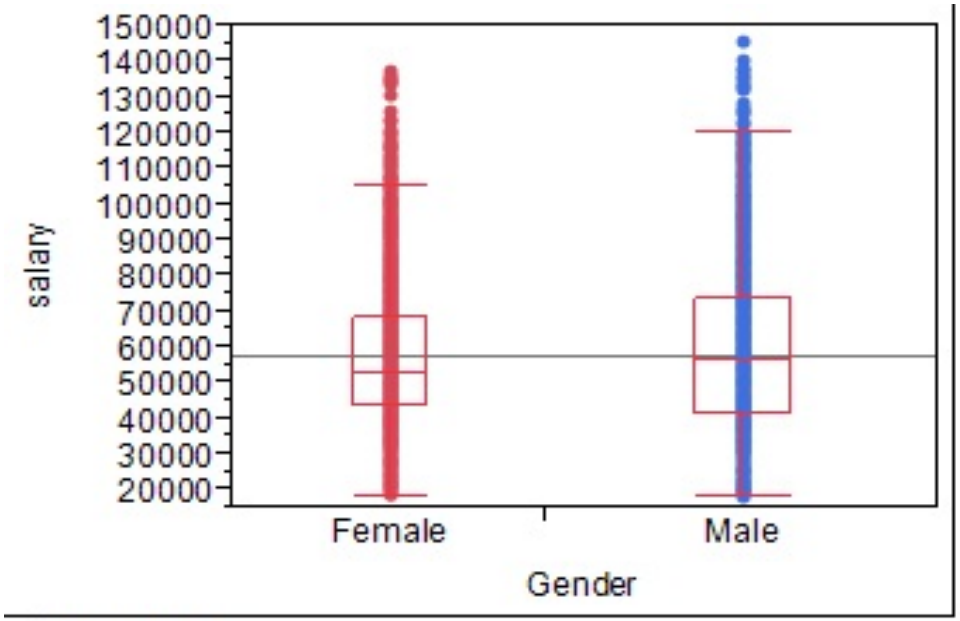
Comparison of two means – the JMP menu

The key menu is

The variable you want to find the mean of goes in as the "Y" variable and the variable you are trying to distinguish between goes in as the "X". Note that "X" should be categorical (in the JMP spreadsheet it is denoted with a red barchart icon).

First, graphically compare the groups

Right click in the title bar and select quantiles



Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
Female	18162	37281.68	43447.75	52162	68266.91	87711.66	136971.8
Male	18162	29880	41377.74	56262	73851.66	91236.48	145162

Now, make a formal comparison

Right click in the title bar and ...
Select ‘‘Means and Std Dev’’
Select ‘‘t-test’’

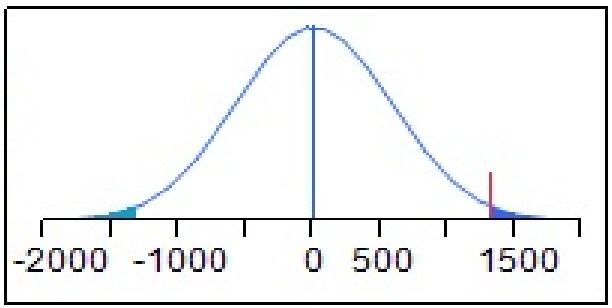
Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Female	2515	57317.9	20071.9	400.24	56533	58103
Male	3132	58631.9	23552.4	420.85	57807	59457

t Test

Male-Female
Assuming unequal variances

Difference	1314.00	t Ratio	2.262474
Std Err Dif	580.78	DF	5625.092
Upper CL Dif	2452.55	Prob > |t|	0.0237*
Lower CL Dif	175.45	Prob > t	0.0119*
Confidence	0.95	Prob < t	0.9881



What the output tells you

- The means are \$1,314 apart.
- The 95% CI for the difference in means is (\$175,\$2,453).
- The p-value for the test is $Prob > |t| = 0.0237$.
- This small (greater than 0.05) p-value says we **reject** the Null.
- That is, there is a statistical evidence of a difference between means and men are being paid more on average than women.
- Be very careful about drawing a definitive conclusion as there could well be a confounding variable issue.

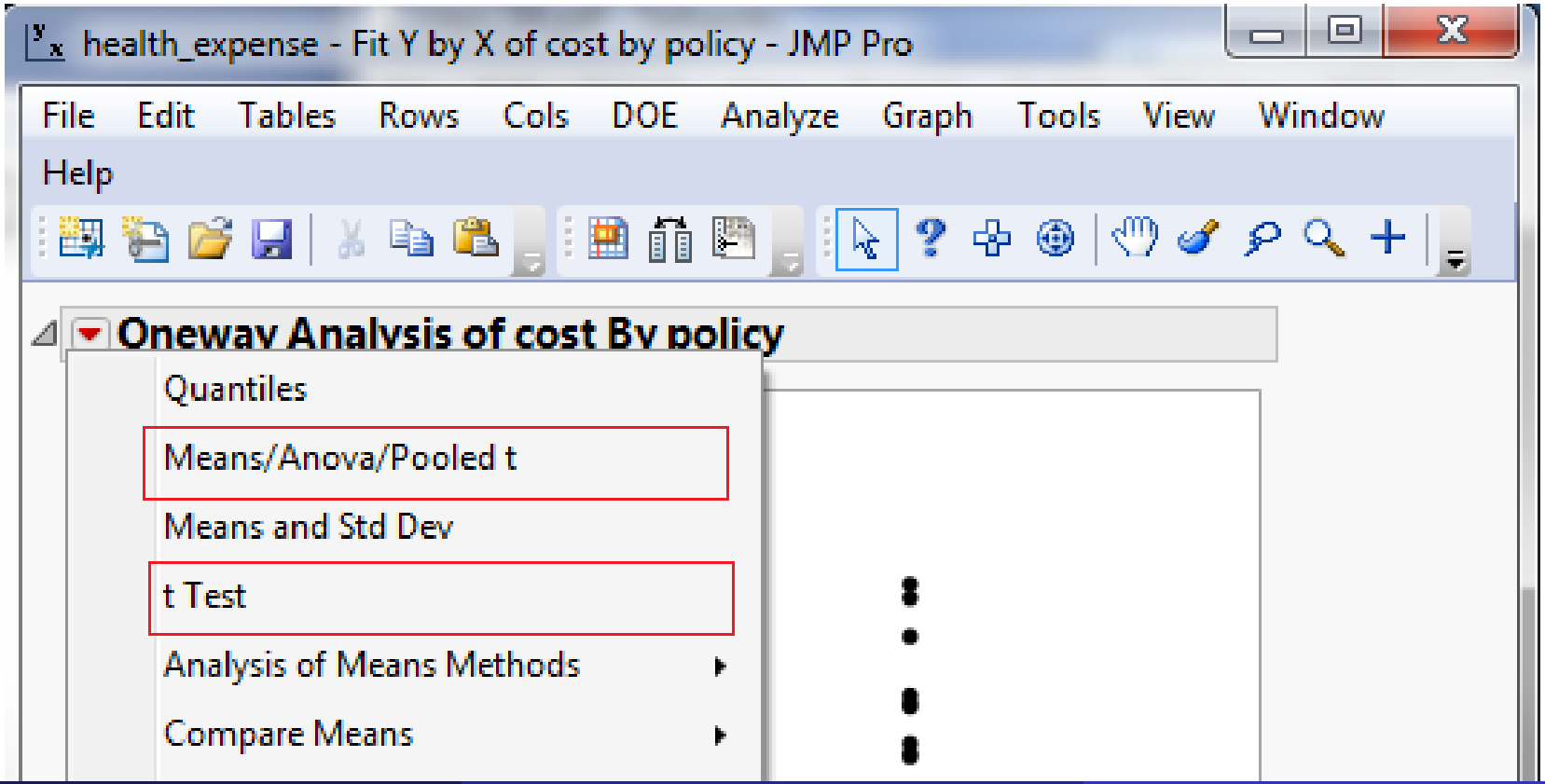
Comments on the two-sample t-test

- We just did a test that allowed the variances of the two groups to be different.
- One nuance: the degrees of freedom for this two sample t-test are complicated and we will use software to calculate them.
- There are assumptions behind the test we just did:
  1. Independence, within and between groups.
  2. Constant variance within each group.
  3. Approximate Normality of the raw data.
- As long as you have decent sample sizes the test does not require the third assumption due to the Central Limit Theorem.
- There is another two-sample t-test that assumes the variances are the same, it is called the *pooled t-test*.
- Regression will be an extension of the pooled t-test (it will assume equality of variances).

The two sorts of 2-sample t-test

1. Assuming equal variances; the pooled t-test. In JMP: `Means/Anova/Pooled t`
2. Not assuming equal variances. In JMP: `t test`

Figure 1: two-sample t-test options



Testing for a difference between two proportions

- Is there a difference between voluntary turnover between men and women?
- The raw data:
  1. Female voluntary turnover = 311 out of 2,515 = 12.37%.
  2. Male voluntary turnover = 316 out of 3,132 = 10.09%.
- Women are more likely to leave than men but is the observed difference statistically significant?

The table of responses

- Make the Turnover variable categorical (it's 0/1 now).
- Create a table.

Analyze→ Fit Y by X
Gender is ‘‘X’’, volturn is ‘‘Y’’

Gender	Count / Total % / Col % / Row %	volturn 0	volturn 1	
	Female	2204 39.03 43.90 87.63	311 5.51 49.60 12.37	2515 44.54
	Male	2816 49.87 56.10 89.91	316 5.60 50.40 10.09	3132 55.46
		5020 88.90	627 11.10	5647

The test for two proportions

$$H_0 : (p_1 - p_2) = 0 \qquad v. \qquad H_1 : (p_1 - p_2) \neq 0.$$

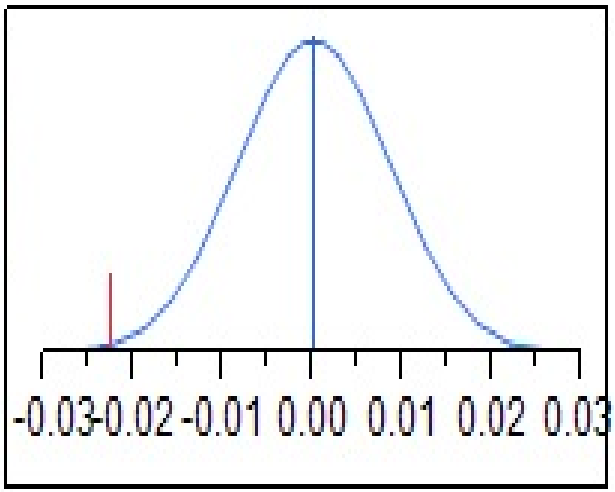
Switch volturn back to continuous
Analyze→ Fit Y by X
Gender is ‘‘X’’, volturn is ‘‘Y’’
Click on t-test

t Test

Male-Female

Assuming unequal variances

Difference	-0.02276	t Ratio	-2.68131
Std Err Dif	0.00849	DF	5158.181
Upper CL Dif	-0.00612	Prob > |t|	0.0074*
Lower CL Dif	-0.03941	Prob > t	0.9963
Confidence	0.95	Prob < t	0.0037*



The formulas for the test statistic and confidence interval, for two-sample proportions

The test statistic is of the form:

$$\frac{\text{Estimate - Null.Hyp. Value}}{\text{std.err(Estimate)}}.$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

The 95% confidence interval is of the form:

$$\text{Estimate } +/- \text{ 2. std.err(Estimate).}$$

In general the $100(1-\alpha)\%$ interval is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Module summary

1. Comparisons between groups
2. Confidence intervals and tests for the difference in means
3. Confidence interval and a test for the difference in proportions

Next time

- Association between two variables: the start of regression.