Lecture 9 Monday January 30

- Lec notes pp. 31-35 Two-sample t example of interpretation (smoking/wrinkling)
- Lecture is recorded, available under Videos

Reminder: The standard two-sample t-test assumes that the two SD's are equal, and goes on to exploit that feature.

We assume equal variances when we pool the two sample variances to compute $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}$$

Applications of the two sample T procedures

The area of A/B testing (e.g. at Google, LinkedIn, Microsoft, etc) consists mainly of two-sample t tests and confidence intervals. (I have put an e-book about this topic on course reserve, and the first two chapters are on the class web site, under References.)

Application of two-sample t procedures

We will now see an illustration with data of a complete application of the two-sample t procedure.

This application is to an observational study.

Setting: In the 1950s and 1960s, epidemiologists argued that smoking was bad for your health.

However, it took decades to convince the public and policy makers of this.

Reason: No controlled, randomized experiments could be done. All existing studies could be criticized for possible confounding.

For example, a famous study by Richard Doll and Bradford Hill in England was carried out in 1948. (See `DollAndHill-BMJ1950-SmokingAndLung.pdf` under References on class web site.)

Eventually tobacco companies were required to print health warnings on cigarette packs.

People still smoked, but numbers of smokers and amount of smoking, declined.

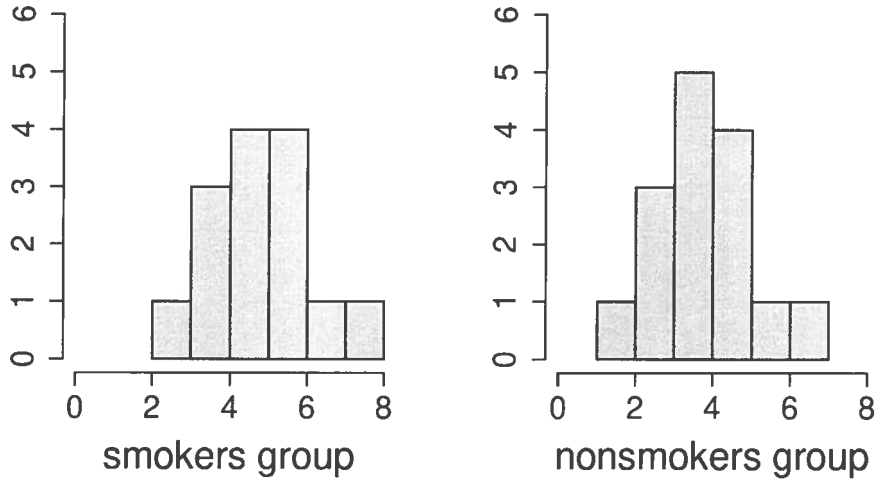
The following study was an effort at persuading smokers to quit smoking for cosmetic reasons.

Example: A physician named A.W. Andrews developed way of scoring a photograph of a face for wrinkles. The scores range from 1 to 10, with 1 indicating no wrinkles, 10 indicating a severe case.

He photographed 29 women aged 45–55, of whom 14 smoked a pack a day and 15 did not smoke. He reported that the 14 who smoked can be considered a random sample from the population of women who smoke a pack a day, and the 15 who didn't smoke can be considered a random sample from the population of women who don't smoke.[1]

Scores:

S:	$2,6,4,6,4,5,5,6,5,6,7,5,4,8$	$n_s = 14$	$\bar{Y}_s = 5.2$	$S_s = 1.48$
NS:	$1,5,3,5,3,4,4,5,4,5,6,4,3,7,4$	$n_{ns} = 15$	$\bar{Y}_{ns} = 4.2$	$S_{ns} = 1.42$



R Output R function `t.test()` was used here; see Lec1.r

Two tests are reported. The first does not assume equal variances; the second is the pooled t-test.

```
Welch Modified Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=26.69, p-value=0.0711
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09,  2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20


Standard Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=27, p-value=0.0706
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09, 2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20
```

The Standard Two Sample t-test is the one based on the assumption of equal variances (assumption 4).

The Welch-corrected procedure does not assume equal variances. Current wisdom is to use the Welch-corrected test, to be safe. The standard two sample t-test can be a lot worse if the two variances are not equal.

The two tests are the same in structure; the difference is in how the standard error in the denominator is estimated.

The Standard test is more similar to what we do in regression analysis, and that is why we focus on that one here.

Comments on the Analysis

- Conclusion:

- The CI is $(-0.09, 2.12)$, which means that we are "95% confident" that $-0.09 < \mu_{\text{s}} - \mu_{\text{ns}} < 2.12$. This interval contains 0.

- Need to do an informal check of normality. You do this by looking at the histograms. The assumption of normality is not really critical, and unless there is something glaring, you usually don't worry about it.

- Side note on Welch's method. The number of df is not necessarily an integer. Actually, there are t-distributions with any number of df. They are not tabulated, but this is not a problem, because you will always do this using a software package anyway.

Conclusion: With $t_{\text{obs}} = 1.88 (P = .0706)$, we fail to reject $H_0 : \mu_Y - \mu_Z = 0$ in a two-sided test. We can't conclude that the average wrinkle score is different for the populations of smokers v. nonsmokers. We note that (i) The observed difference $\bar{Y} - \bar{Z} = 1.0$ is important in the practical sense, and (ii) The result **is** statistically significant at level $\alpha = .10$.

The 95% confidence interval contains 0; this says that the result is not statistical significant at level $\alpha = .05$. We note that the CI **almost** excludes zero. The sample size was small, and perhaps we would have gotten a significant result with larger n_Y and n_Z.

Important remark about the conclusion: Since this is an observational study, we wouldn't be able to infer causation even if the result were found to be statistically significant.

In Lecture 7 we ended by discussing the P-value.
The P-value is often referred to as "the observed significance level." The "significance level" is often denoted α.
Definition: Type I Error If the null hypothesis is true, and we reject H_0, we have made the more serious type of error, which is called *Type I error.*
Definition: Type II Error If the alternative hypothesis is true, and we fail to reject H_0, we have made the less serious type of error, which is called *Type II error.*
In pre-set α significance testing, we set the probability of Type I error in advance, usually to be $\alpha = .05$. This leads to a clear-cut rejection region of our test. For the two-sample pooled t-test, we say we will reject $H_0 : \mu_Y = \mu_Z$ in favor of the two-sided alternative if:

$$|T_{\text{obs}}| > t_{n_Y+n_Z-2,.975}$$

Back to the P-value: The P-value is the smallest significance level at which we would have rejected H_0; thus the name "observed significance level."