Lecture 15

Type I & Type II Errors; Inference on Two Population Means

Text: Chapters 5 and 6

STAT 8010 Statistical Methods I
October 13, 2020

Whitney Huang
Clemson University

The 2×2 Decision Paradigm for Hypothesis Testing

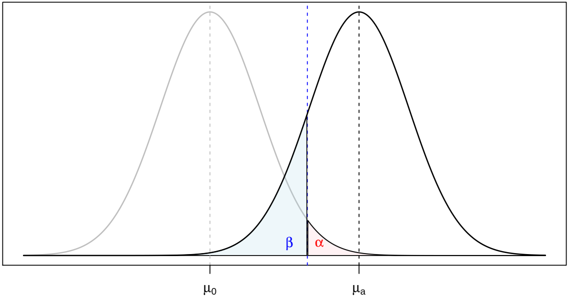
True State	Decision	
	Reject H_0	Fail to reject H_0
H_0 is true	Type I error	Correct
H_0 is false	Correct	Type II error

Errors in Hypothesis Testing

- The probability of a type I error is denoted by α
- The probability of a type II error is denoted by β

Type I & Type II Errors

- Type I error: $\mathrm{P}(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$
- Type II error: $\mathrm{P}(\text{Fail to reject } H_0 | H_0 \text{ is false}) = \beta$



$\alpha \downarrow \beta \uparrow$ and vice versa

Type II Error and Power

- The type II error, β, depends upon the true value of μ (let's call it μ_a)
- We use the formula below to compute β

$$\beta(\mu_a) = \mathrm{P}(z^* \leq z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}})$$

- The power (PWR): $\mathrm{P}(\text{Reject } H_0|H_0 \text{ is false}) = 1 - \beta$. Therefore PWR$(\mu_a) = 1 - \beta(\mu_a)$

Question: What increases Power?

Sample Size Determination

Suppose that we wish to determine what sample size is required to detect the difference between a hypothesized mean and true mean $\mu_0 - \mu_a$, denoted by Δ, with a given power $1 - \beta$ and specified significance level α and known standard deviation σ. We can use the following formulas

$$n = \sigma^2 \frac{(z_\alpha + z_\beta)^2}{\Delta^2} \text{ for a one-tailed test}$$

$$n \approx \sigma^2 \frac{(z_{\alpha/2} + z_\beta)^2}{\Delta^2} \text{ for a two-tailed test}$$

Example

An existing manufacturing process produces, on average, 100 units of output per day. A pilot plant is used to evaluate a possible process change. Suppose the Company CEO wants to know if yield is increased. The CEO uses $\alpha = 0.05$ and the sample mean $(n = 25)$ is 103. Do we have sufficient evidence to conclude that the mean yield exceeds 100 if $\sigma = 10$?

1. $H_0 : \mu = 100$ vs. $H_a : \mu > 100$
2. $z_{obs} = \frac{103-100}{10/\sqrt{25}} = 1.5$
3. The cutoff value of the rejection region is $z_{0.05} = 1.645$. Therefore we do not have enough evidence to conclude that the new process mean yield exceeds 100

Example Cont'd

Suppose the true true mean yield is 104.

- What is the power of the test?

Example Cont'd

Suppose the true true mean yield is 104.

- What is the power of the test?

$$\beta(\mu = 104) = P\left(Z \le z_{0.05} - \frac{|100 - 104|}{10/\sqrt{25}}\right)$$
$$= P(Z \le 1.645 - 4/2) = P(Z \le -0.355)$$
$$= \Phi(-0.355) = 0.3613$$

Therefore, the power is $1 - 0.3613 = 0.6387$

- What sample size is required to yield a power of 0.8 with a significance level of 0.05?

Example Cont'd

Suppose the true true mean yield is 104.

- What is the power of the test?

$$\beta(\mu = 104) = P\left(Z \le z_{0.05} - \frac{|100 - 104|}{10/\sqrt{25}}\right)$$
$$= P(Z \le 1.645 - 4/2) = P(Z \le -0.355)$$
$$= \Phi(-0.355) = 0.3613$$

Therefore, the power is $1 - 0.3613 = 0.6387$

- What sample size is required to yield a power of 0.8 with a significance level of 0.05?

$$n = \sigma^2 \frac{(z_{0.05} + z_{0.2})^2}{\Delta^2} = 10^2 \frac{(1.645 + 0.8416)^2}{4^2} = 38.6324$$

Therefore, the required sample size is 39

Duality of Hypothesis Test with Confidence Interval

There is an interesting relationship between CIs and hypothesis tests. If H_0 is rejected with significance level α then the corresponding confidence interval does not contain the value μ_0 targeted in the hypotheses with the confidence level $(1-\alpha)$, and vice versa

Hypothesis test at α level	$(1-\alpha)\times$ 100% CI
$H_0: \mu = \mu_0$ vs. $H_a: \mu \neq \mu_0$	$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$
$H_0: \mu = \mu_0$ vs. $H_a: \mu > \mu_0$	$\left(\bar{X} - t_{\alpha,n-1}s/\sqrt{n}, \infty\right)$
$H_0: \mu = \mu_0$ vs. $H_a: \mu < \mu_0$	$\left(-\infty, \bar{X} + t_{\alpha,n-1)s/\sqrt{n}}\right)$

Comparing Two Population Means

- We often interested in comparing two groups (e.g.)
  - Does a particular pesticide increase the yield of corn per acre?
  - Do men and women in the same occupation have different salaries?
- The common ingredient in these questions: They can be answered by conducting statistical inferences of two populations using two (independent) samples, one from each of two populations

Notation

- Parameters:
  - Population means: μ_1, μ_2
  - Population standard deviations: σ_1, σ_2
- Statistics:
  - Sample means: $\bar{X}_1, \bar{X}_2$
  - Sample standard deviations: s_1, s_2
  - Sample sizes: n_1, n_2

Statistical Inference for $\mu_1 - \mu_2$

- Point estimate: $\bar{X}_1 - \bar{X}_2$
- Interval estimate: Need to figure out $\sigma_{\bar{X}_1 - \bar{X}_2}$
- Hypothesis Testing:
  - Upper-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 > 0$
  - Lower-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 < 0$
  - Two-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 \neq 0$

Confidence Intervals for $\mu_1 - \mu_2$

If we are willing to **assume** $\sigma_1 = \sigma_2$, then we can "pool" these two (independent) samples together to estimate the common σ using s_p:

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

We can then derive the (estimated) standard error of $\bar{X}_1 - \bar{X}_2$, which takes the following form

$$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

With CLT (assuming sample sizes are sufficiently large), we obtain the $(1 - \alpha) \times 100\%$ CI for $\mu_1 - \mu_2$:

$$\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{point estimate}} \pm \underbrace{t_{\alpha/2, n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{\text{margin of error}}$$

Confidence Intervals for $\mu_1 - \mu_2$: What if $\sigma_1 \neq \sigma_2$?

- We will use s_1^2, s_2^2 as the estimates for σ_1^2 and σ_2^2 to obtain the standard error:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- The formula for the degrees of freedom is somewhat complicated:

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- We can then construct the $(1 - \alpha) \times 100\%$ CI for $\mu_1 - \mu_2$:

$$\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{point estimate}} \pm \underbrace{t_{\alpha/2, \text{ df calculated from above}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{margin of error}}$$

To Pool ($\sigma_1 = \sigma_2$) or Not to Pool ($\sigma_1 \neq \sigma_2$)?

We could perform the following test:

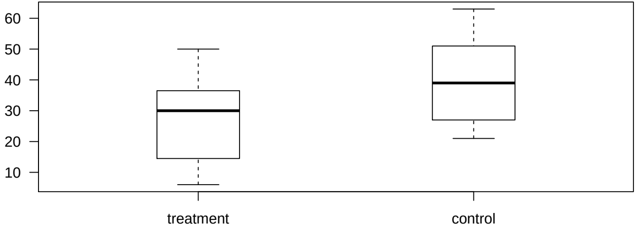
- $H_0 : \sigma_1^2/\sigma_2^2 = 1$ vs. $\sigma_1^2/\sigma_2^2 \neq 1$
- Test statistic: $F^* = s_1^2/s_2^2$. Under H_0, $F^* \sim F(n_1 - 1, n_2 - 1)$
- For a given α, we reject H_0 if the P-value $< \alpha$ (or $F_{obs} > F_\alpha(n_1, n_2)$)
- If we fail to reject H_0, then we will use s_p as an estimate for σ and we have $s_{\bar{X}_1 - \bar{X}_2} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Otherwise, we use $s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Example

An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age and health was randomly divided into two groups. Twelve of the lambs were injected with the drug (treatment group) and the remaining twelve were left untreated (control group). After a 6-month period, the worm counts were recorded:

Treatment	18	43	28	50	16	32	13	35	38	33	6	7
Control	40	54	26	63	21	37	39	23	48	58	28	39

Plot the Two Samples



- $n_1 = n_2 = 12 \Rightarrow$ sample size is perhaps not large enough for CLT to work. But the boxplots suggest the distributions are symmetric with no outliers
- The untreated lambs (control group) appear to have higher average worm counts than the treated lambs (treatment group). But do we have enough evidence ?

Example Cont'd

```
> apply(dat, 2, mean)
treatment   control
 26.58333  39.66667
> apply(dat, 2, sd)
treatment   control
 14.36193  13.85859
> var.test(treatment, control)

        F test to compare two variances

data:  treatment and control
F = 1.074, num df = 11, denom df = 11, p-value = 0.9079
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3091686 3.7306092
sample estimates:
ratio of variances
          1.073959
```

We fail to reject $\sigma_1 = \sigma_2 = \sigma$. Therefore we will use s_p, the pooled standard deviation, as an estimate for σ

Example Cont'd

- Place a 95% confidence interval on $\mu_1 - \mu_2$ to assess the size of the difference in the two population means

- Test whether the mean number of tapeworms in the stomachs of the treated lambs is less than the mean for untreated lambs. Use an $\alpha = 0.05$ test

Another Example

A simple random sample with sample size 37 is taken and are subjected to a treatment ($\bar{X}_1 = 19.45, s_1 = 4.3$). A Simple random sample with sample size 31 is taken and given a placebo ($\bar{X}_2 = 18.2, s_2 = 2.2$). At the 10% level can we say that the means are different between the two groups?

Paired T-Test: Motivating Example

Insurance handlers are concerned about the high estimates they are receiving for auto repairs from garage I compared to garage II. To verify their suspicions, each of 15 cars recently involved in an accident was taken to both garages for separate estimates of repair costs. The estimates from the two garages are given in the following table

Garage I	Garage II	Garage I	Garage II	Garage I	Garage II
17.6	17.3	20.2	19.1	19.5	18.4
11.3	11.5	13.0	12.7	16.3	15.8
15.3	14.9	16.2	15.3	12.2	12.0
14.8	14.2	21.3	21.0	22.1	21.0
16.9	16.1	17.6	16.7	18.4	17.5

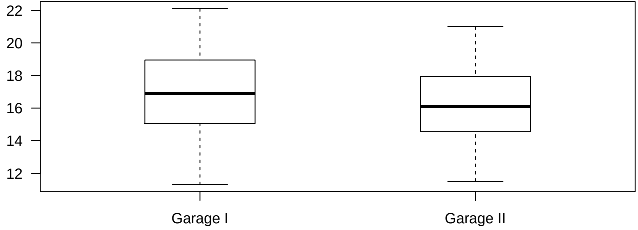
Example Cont'd

Suppose we perform a two-sample test

Sample statistics: $\bar{X}_1 = 16.85, \bar{X}_2 = 16.23, s_1 = 3.20, s_2 = 2.94$

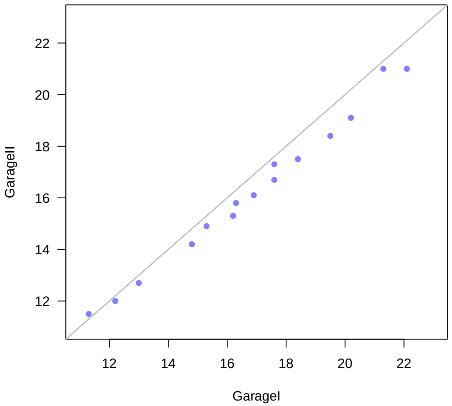
- $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 > 0$
- $t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{16.85 - 16.23}{\sqrt{\frac{3.2^2}{15} + \frac{2.94^2}{15}}} = \frac{0.62}{1.12} = 0.55$
- Critical value for rejection region: $t_{0.05, df=27} = 1.70$
- Since t_{obs} is not in the rejection region. We fail to reject H_0 at 0.05 level.

Boxplots and R Output




```
        Welch Two Sample t-test

data:  GarageI and GarageII
t = 0.54616, df = 27.797, p-value =
0.2947
alternative hypothesis: true difference in means
is greater than 0
95 percent confidence interval:
 -1.29749      Inf
sample estimates:
mean of x mean of y
 16.84667  16.23333
```



For all but one of the 15 cars, the estimates from garage I were higher than that from garage II.

Analyzing Matched Pairs

- Matched pairs are dependent samples where each unit in the first sample is directly linked with a unit in the second sample
- This could occur in several situations, for example, before/after study, study on twins, pairing subjects based on similar characteristics
- We need different strategy for testing two dependent samples $\Rightarrow$ Paired T-Tests

Paired T-Tests

- $H_0 : \mu_d = 0$ vs. $H_a : \mu_d > 0$ (Upper-tailed); $\mu_d < 0$ (Lower-tailed); $\mu_d \neq 0$ (Two-tailed)
- Test statistic: $t^* = \frac{\bar{X}_d - 0}{\frac{s_d}{\sqrt{n}}}$. If $\mu_d = 0$, then $t^* \sim t_{df=n-1}$
- Use rejection region method or P-value method to make a decision

Car Repair Example Revisited

Garage I - Garage II	Garage I - Garage II	Garage I - Garage II
17.6 - 17.3 = 0.3	20.2 - 19.1 = 1.1	19.5 - 18.4 = 1.1
11.3 - 11.5 = -0.2	13.0 - 12.7 = 0.3	16.3 - 15.8 = 0.5
15.3 - 14.9 = 0.4	16.2 - 15.3 = 0.9	12.2 - 12.0 = 0.2
14.8 - 14.2 = 0.6	21.3 - 21.0 = 0.3	22.1 - 21.0 = 1.1
16.9 - 16.1 = 0.8	17.6 - 16.7 = 0.9	18.4 - 17.5 = 0.9

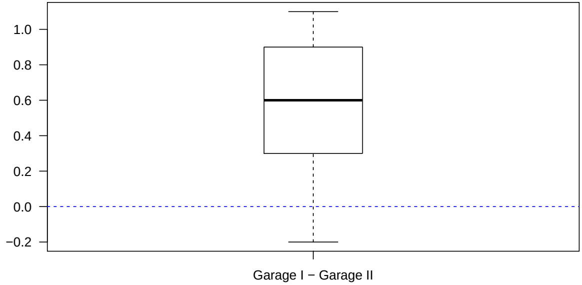
1. First, compute the difference in paired samples
2. Compute the sample mean and standard deviation for the differences
3. Then perform a one sample t-test

Car Repair Example Cont'd

$\bar{X}_d = 0.61, s_d = 0.39$

1. $H_0 : \mu_d = 0$ vs. $H_a : \mu_d > 0$
2. $t_{obs} = \frac{0.61}{\frac{0.39}{\sqrt{15}}} = 6.03$
3. Critical value for rejection region: $t_{0.05,df=14} = 1.76 \Rightarrow$ reject H_0
4. We do have enough evidence that the true mean repair cost difference for the garage I and II is greater than 0

Boxplot and R Output




```
        Paired t-test

data:  GarageI and GarageII
t = 6.0234, df = 14, p-value = 1.563e-05
alternative hypothesis: true difference in means
is greater than 0
95 percent confidence interval:
 0.4339886       Inf
sample estimates:
mean of the differences
              0.6133333
```