

Section 0.2: Principles of Statistical Inference (Two Sample Problem)

William Christensen and Dennis Tolley

Basic Two Sample Paradigm

- There are two independent random variables X_1 and X_2 with means μ_1 and μ_2 , respectively.
- n_1 observations, $x_{11}, \dots, x_{1n_1}$, of random variable X_1 are observed.
- n_2 observations, $x_{21}, \dots, x_{2n_2}$, of random variable X_2 are observed.
- We obtain $\bar{x}_1$ and $\bar{x}_2$ from these two samples.
- We wish to compare means μ_1 and μ_2 .
- We desire a confidence interval for $\mu_1 - \mu_2$.

Two Sample Statistics

- What is the probability that $(\bar{x}_1 - \bar{x}_2)$ could be this big (just due to “chance”) if $\mu_1 = \mu_2$?
- What do we mean by “chance??”
- The answer is the p-value

Def’n The p-value is the probability of observing a test statistic as or more extreme than the one we observed, given that the null hypothesis, $\mu_1 = \mu_2$, is true.

+ for our example:

$$\text{p-value} = \Pr \left\{ \underbrace{|\bar{X}_1 - \bar{X}_2|}_{\text{Random variable}} \geq \underbrace{|\bar{x}_1 - \bar{x}_2|}_{\text{Observed statistic}} \mid \mu_1 = \mu_2 \right\}$$

↑
“given”

Two Sample Normal Distribution

Normal (Gaussian) distribution — applies to linear combinations of statistics

- Let a_1 and a_2 be two real constants.
- Let X_1 and X_2 be two normally distributed random variables.
- Then

$$Y = a_1 * X_1 + a_2 * X_2 \quad (1)$$

is also normally distributed.

- Therefore the normal distribution can be used to calculate p-values for linear combinations of two normally distributed random variables.

Density function

- A simple function of Y with parameters μ_y and σ_y^2 .

$$f(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{\frac{-1}{2\sigma_y^2}(y-\mu_y)^2} \quad -\infty < x < \infty$$

where $-\infty < \mu_y < \infty$ and $\sigma_y^2 > 0$.

Sampling Distributions and CLT

Often we are able to determine the approximate probability distribution of a linear combination of averages if we know something about the distribution of the population from which we draw our samples to calculate our means. The distribution of a statistic (such as $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$) is called a sampling distribution of the linear combination.

- Imagine all possible samples of size n_1 and n_2 from populations 1 and 2, respectively.
- Find the sample mean $\bar{x}_1$ and $\bar{x}_2$ from each of the two samples.
- Consider the distribution of the values of $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$ over all these possible samples.

-OR-

Just think of $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$ as a continuous random variable

$$-\infty < \bar{x} < \infty$$

The Sampling Distribution for $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$:

- 1 Has mean of μ (i.e., $E\{\bar{X}\} = \mu_y = a_1 * \mu_1 + a_2 * \mu_2$)
- 2 Has variance σ_y^2 that is a function of n_1 , n_2 and the variances and covariance of X_1 and X_2 .
- 3 As $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$, Y has a distribution that converges toward the normal distribution. (i.e.,
 $a_1 * \bar{X}_1 + a_2 * \bar{X}_2 \sim N(\mu_y, \sigma_y^2)$ (for large n_1 and n_2))

Property #3 above is because of the Central Limit Theorem.

- This means that for large n_1 and n_2 , $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$ follows a normal distribution regardless¹ of the distribution of X_1 and X_2 .
 - You're right!! This IS amazing!

¹The distribution of $a_1 * \bar{X}_1 + a_2 * \bar{X}_2$ must have finite mean and variance.

Confidence Intervals

An important implication of the CLT is that we now know what values of $\bar{x}$ are likely, based on the distribution of X . For large n ,

$$Z = \frac{Y - \mu_y}{\sigma_y} \sim N(0, 1) \quad \text{“}\sim\text{”} = \text{“approx. dist'd”}$$

So, for some future random samples ($X_{11}, \dots, X_{1n}$ and $X_{11}, \dots, X_{1n}$) then for $Y = a_1 * \bar{X}_1 + a_2 * \bar{X}$

$$Pr\{-1.96 < \frac{Y - \mu_y}{\sigma_y} < 1.96\} = 0.95$$

↑
2.5th percentile of
 $N(0, 1)$ distribution

↑
97.5th percentile of
 $N(0, 1)$ distribution

Derive C.I. formula for μ :

$$100 \times (1 - \alpha)\% \text{ C.I. for } \mu: \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

NOTE

We need either σ_y or an estimate!!!

For large n_1 and n_2 in the two-sample case when X_1 and X_2 are independent, with $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$\sigma_y^2 = \sigma^2 * \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (2)$$

Recall, also that

$$\mu_y = a_1 * \mu_1 + a_2 * \mu_2 \quad (3)$$

OFTEN, $a_1 = 1$ and $a_2 = -1$ resulting in

$$Y = \overline{X_1} - \overline{X_2}, \quad (4)$$

where

$$\mu_y = \mu_1 - \mu_2 \quad (5)$$

For large n_1 and n_2 in the two-sample (equal variance) case when X_1 and X_2 are independent:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\sigma^2}} \sim N(0, 1)$$

So, for some future random samples $(X_{11}, \dots, X_{1n_1})$ and $(X_{21}, \dots, X_{2n_2})$:

$$Pr \left\{ -1.96 < \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\sigma^2}} < 1.96 \right\} = 0.95$$

and, following similar logic,

$$Pr \left\{ \bar{X}_1 - \bar{X}_2 - Z_{\alpha/2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + Z_{\alpha/2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2} \right\} = 1 - \alpha$$

$100 \times (1 - \alpha)\%$ C.I. for $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm Z_{\alpha/2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2}$$

Unknown Variances

(focusing on the two-sample problem)

It is most often the case that σ^2 is not known and must be estimated.
In these cases, our test statistic is not normally distributed.

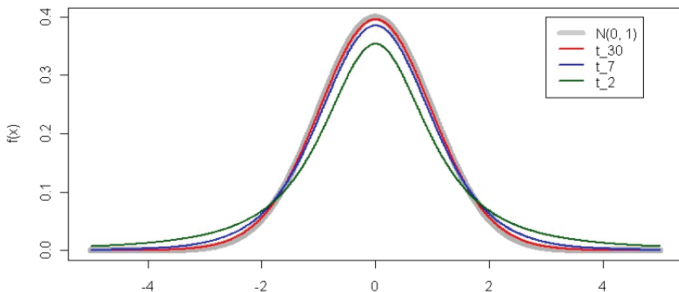
Our test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})s_{p\ell}^2}} \text{ where } s_{p\ell}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- $\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})s_{p\ell}^2}$ is the standard error (“S.E.”) of $\bar{x}_1 - \bar{x}_2$
- “ $p\ell$ ” stands for “pooled”

and

$$\underbrace{t}_{\substack{\uparrow \\ t \text{ statistic}}} \sim \underbrace{t_{n_1+n_2-2}}_{\substack{t \text{ distribution with} \\ n_1 + n_2 - 2 \text{ d.f.}}}$$



t-distributions vs normal

Note that the t distributions reflect the additional uncertainty present

in $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})s_{p\ell}^2}}$ that does not affect $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\sigma^2}}$

$100 \times (1 - \alpha)\%$ C.I. for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{p\ell}^2}$$

Hypothesis Testing

In statistics, a hypothesis is a claim about a parameter of interest. Hypothesis testing (or significance testing) often follows the pattern:

- 1 STATE a question about a parameter and specify the claimed value
- 2 PLAN
 - Specify H_o and H_A and choose α .
 - Select appropriate procedure/statistic.
- 3 SOLVE
 - Collect data and check conditions for inference
 - Calculate test statistics
 - Calculate p-value
- 4 CONCLUDE

EX Shoe Wear Experiment (Two-Sample t-test)

Two different synthetic materials (A and B) were compared by placing each material on the bottom of 10 different shoe soles (total of 20 shoes).

1 STATE: Is there a difference between the mean wear of materials A and B?

2 PLAN:

$$\begin{array}{lll} & H_o : \mu_A = \mu_B & (\text{or } \mu_A - \mu_B = 0) \\ \text{vs.} & H_a : \mu_A \neq \mu_B & (\text{or } \mu_A - \mu_B \neq 0) \end{array}$$

Choose desired significance level α – the probability of falsely rejecting a true H_o .

? What are the issues in choosing α ?

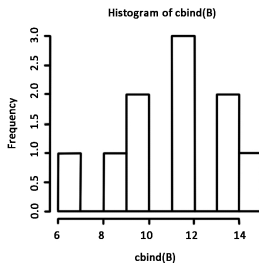
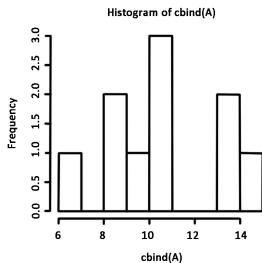
■ Choose $\alpha = 0.05$ for now

3 SOLVE

- Collect data and check assumptions

A: 13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3

B: 14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6



- Normality?
- Equal Variances?
- Independence?

■ Test Statistic:

$$Use : \quad \bar{x}_A = 10.63$$

$$\bar{x}_B = 11.04$$

$$s_A = 2.451$$

$$s_B = 2.518$$

$$n_A = 10$$

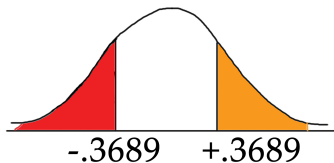
$$n_B = 10$$

$$s_{p\ell}^2 = \frac{(10 - 1)s_A^2 + (10 - 1)s_B^2}{10 + 10 - 2}$$

$$= 6.1758$$

$$df = n_1 + n_2 - 2 = 18$$

$$t = \frac{10.63 - 11.04}{\sqrt{(\frac{1}{10} + \frac{1}{10})6.1758}} = -.3689$$



Using R: `pt(value, df)` gives **left tail**

↗
find area to
the left of value

↖
use df for the
degrees of freedom

Red area:

`pt(-.3689, 18)` gives .3583

Orange area:

`pt(.3689, 18)` gives .6417 $\Rightarrow$ Area = .3583

or `pt(.3689, 18, lower.tail=FALSE)`

p-value = $2 \times .3583 = .7165$

or just use “t.test” in R

```
t.test(A, B, alternative = "two-sided",  
       var.equal=TRUE)
```

Two Sample t-test

data: A and B

t= -0.3689, df=18, p-value = 0.7165

alternative hypothesis: true difference in means
is not equal to 0

95% confidence interval:

-2.744924 1.924924

sample estimates:

mean of x mean of y

10.63 11.04

- What would have been the p-value if H_A was “ $H_A : \mu_A < \mu_B$ ”?
- What would have been the p-value if H_A was “ $H_A : \mu_A > \mu_B$ ”?

4 CONCLUDE

Because the p-value of 0.7165 is greater than $\alpha = 0.05$, we have insufficient evidence to conclude that μ_A is different from μ_B .

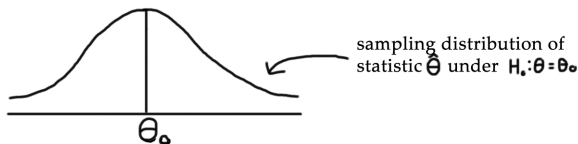
Hypothesis Testing Outcomes

		<u>Decision</u>	
		Fail to Reject H_0	Reject H_0
<u>Reality</u>	H_0 is true	Correct "acceptance" of H_0	Type I Error
	H_0 is false	Type II Error	Correct rejection of H_0

$\alpha = Pr \{ \text{Type I error} \}$ "significance level"

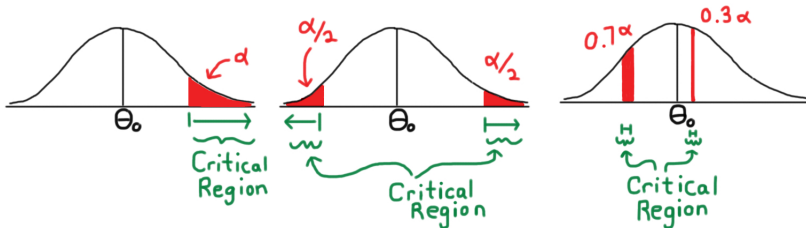
$\beta = Pr \{ \text{Type II error} \}$

$1 - \beta = Pr \{ \text{Correct rejection} \} = \text{"Power"}$



α is an area under the curve that we choose to be associated with the critical region (definition: the set of values for a statistic that would lead to a rejection of H_0)

■ Some choices for critical regions:

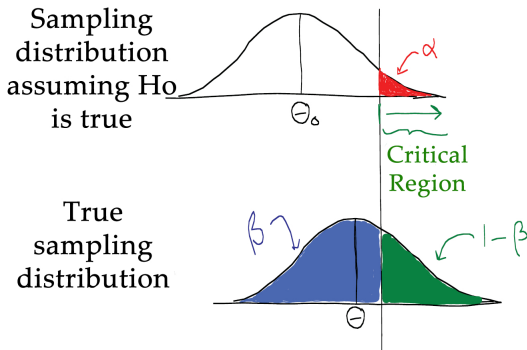


For one-sided
 H_a

For two-sided
 H_a

Legitimate C.R.,
but pretty stupid idea

- We define our critical region using α and assuming H_0 , but what if H_0 is false? What is the probability of rejecting H_0 ?
 - Answer: “power”



- Note that as θ moves further to the right the power ($1 - \beta$) gets larger

- A power curve illustrates the power $(1 - \beta)$ for a test given various values of the parameter of interest.

- Note: more technically, we can define power as a function of μ

$$1 - \beta(\mu) = \Pr\{\text{reject } H_0 | \mu\}$$

so that we can define the power when $H_0 : \mu = \mu_0$ is true using

$$1 - \beta(\mu_0) = \Pr\{\text{reject } H_0 | \mu_0\} = \alpha$$

Paired Comparisons

Sometimes our two samples are NOT independent. Rather, we have measurements on two units within a block.

Examples:

- A pre-test and a post-test given to the same student before and after training
 - Block: a student
 - Unit: a student **at the time of test**
- Two diets assigned to each athlete (high carb randomly assigned to one month and high protein assigned to the other.)
 - Block: an athlete
 - Unit: an athlete during a month

- Field of corn with two plots which are assigned a fertilizer amount

→ Block: field

→ Unit: Plot within a field

NOTE: In all examples above, we believe that the block effect (i.e., block-to-block differences) will be important.

? When should blocking (e.g., paired comparisons) be used in a design?

? When should it not be used?

EX Shoe leather experiment

? Is there a better approach than the 2-sample design?

Approach: From each pair (x_{1i}, x_{2i}) , calculate

$d_i = x_{2i} - x_{1i}$ and calculate

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i \quad [\text{here, } n \text{ is the number of pairs}]$$

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

Follow one-sample t-test calculations:

$$t_d = \frac{\bar{d} - \mu_{d_0}}{\frac{s_d}{\sqrt{n}}} \sim t_{n-1} \quad \text{for } H_0 : \mu_d = \underbrace{0}_{\substack{\uparrow \\ \mu_{d_0} = 0}}$$

EXAMPLE: Shoe Wear Experiment

Two different synthetic materials (A and B) were compared by placing each material on the bottom of 10 different shoe soles (total of 20 shoes). Each of 10 boys has material A randomly assigned to one foot and B assigned to the other.

- 1 STATE: Is there a difference between the mean wear of materials A and B?
- 2 PLAN: $H_0 : \mu_d = 0$ vs. $H_a : \mu_d \neq 0$ Let $\alpha = 0.10$

3 SOLVE (in R)

$$\bar{d} = -0.41$$

$$S_d = 0.387$$

- normality? $\leftarrow$ of what? X_{1i} ? X_{2i} ? d_i ?
- random sample?

equivalent to $\bar{x}_1 - \bar{x}_2 \downarrow$

$$t = \frac{\bar{d} - \mu_{d0}}{\frac{S_d}{\sqrt{n}}} = \frac{-0.41 - 0}{\frac{0.387}{\sqrt{10}}} = -3.349$$

$$\begin{aligned} p\text{-value} &= 2 * pt(-3.349, 10 - 1) \\ &= 0.0085 \end{aligned}$$



- 4 CONCLUDE Because the p-value of 0.0085 is less than the significance level of 0.10, we reject H_0 and conclude that μ_A is not equal to μ_B (i.e., $\mu_d \neq 0$).

- Note that while the two-sample test showed a non-significant comparison of the materials ($p\text{-value} = 0.7165$), by removing the large boy-to-boy variability via blocking, we obtain a significant test ($p\text{-value} = 0.0085$) when using a paired t-test, **even though the data were exactly the same!!**

- Standard error: (2-sample) $\sqrt{6.1758(\frac{1}{10} + \frac{1}{10})} = 1.111$
- Standard error: (paired) $\frac{0.387}{\sqrt{10}} = 0.122$

Ratio of margin-of-error values for 95% confidence intervals:

$$\frac{t_{.025,18} * 1.11}{t_{.025,9} * 0.122} = \frac{2.101 * 1.111}{2.262 * 0.122} = 8.54$$

- By using the structure in the data to choose the proper analysis, we have over an 8-fold improvement in efficiency of the 95% C.I.