1 Matched pair comparison (p430-)

This has a quite different assumption (matched pair) from the other three methods.

Remember LEC 32 page 1 example:

Consider 4 twins A1 A2 B1 B2 C1 C2 D1 D2 (total 8 people), where A1 and A2 are twin, B1 and B2 are twin etc. We want to test if coke affects sleeping time compared to water. Mike, Nancy, and Tom decided to use following methods:

Nancy’s method: To divide 8 people into two groups, separate twins:

(A1, B1, C1, D1) : group 1 (give them water)

(X_1, \ldots, X_4)

(A2, B2, C2, D2) : group 2 (give them coke)

(Y_1, \ldots, Y_4)

(X_1 and Y_1 are matched, X_2 and Y_2 are matched etc.)

and measure sleeping time for 8 people:

(X_1, X_2, X_3, X_4, Y_1, Y_2, Y_3, Y_4) Define D_i = X_i - Y_i, i = 1, \ldots, 4. (n=4. NOT n=8)

• Assume D_i are a random sample from N(\delta, \sigma_D).
• Estimator of $\delta$: $\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$

• 100(1 − $\alpha$)% CI of $\delta$:

$$\overline{D} \pm t_{df=n-1, \alpha/2} \cdot \frac{S_D}{\sqrt{n}}$$

where

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}.$$

• Hypothesis test with $\alpha$ level (C3):

$H_0 : \delta = \delta_0$

$H_1 : \delta \neq \delta_0$

($\delta_0$ is just a number, eg, $\delta_0 = 3$)

Step1) Get

$$T_{obs} = \frac{\overline{D} - \delta_0}{S_D/\sqrt{n}} \quad (1)$$

If $T_{obs}$ is in rejection region, ”Reject $H_0$.” In other word, if $|T_{obs}| > t_{df=n-1, \alpha/2}$, then ”Reject $H_0$.”

Suppose you do $H_1 : \mu_1 - \mu_2 > \delta_0$ (C1) test.
If $T_{obs} > t_{df=n-1, \alpha}$, then ”Reject $H_0$.”

Suppose you do $H_1 : \mu_1 - \mu_2 < \delta_0$ (C2) test.
If $T_{obs} < -t_{df=n-1, \alpha}$, then ”Reject $H_0$.”
2 Review: Ch10 Two sample comparison

We learned 3+1 method:

M1) Large sample
M2) Small sample Normal dist. equal variance
M3) Small sample Normal dist. unequal variance
M4) Matched pair

M1, M2, and M3 use **randomization**: (p429) Randomly assign treatment 1 or treatment 2 to individuals.

M4 uses **blocking**: (p431) Each pair (twin) receives different treatment treatment 1 and treatment 2. **Within the block, you do randomization.** (Suppose you measure a weight of twins and label a lighter person as group A, and the heavier person as group B, **this is NOT randomization within the block** Hence BAD experimental design.)

Why is blocking important? Suppose there are two twins. A1, A2, B1, B2. A1 and A2 are twin (from family A) and
B1 and B2 are twin (from family B). Suppose you measure the sleeping time using non-matched pair (bad !) design:

Group1 (water) : A1, A2  
Group2 (coke) : B1, B2.

Genetically A1 and A2 are similar. For example, family A has 5 hour sleeping habit and family B has 9 hour sleeping habit. (See picture below)

Then giving water for family A and coke for family B makes result confusing. Lack of blocking.

In this case, we should do blocking (matched pair design) (See p431 figure),

Group1 (water) : A1, B1  
Group2 (coke) : A2, B2
EG1) To compare the crop yield from two strains of wheat, A and B, an experiment was conducted at eight farms located in different parts of a state. At each farm, strain A was grown on one plot and strain B on another; all 16 plots were equal sizes.

<table>
<thead>
<tr>
<th>Farm</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain A</td>
<td>23</td>
<td>39</td>
<td>19</td>
<td>43</td>
<td>33</td>
<td>29</td>
<td>28</td>
<td>42</td>
</tr>
<tr>
<td>Strain B</td>
<td>18</td>
<td>33</td>
<td>21</td>
<td>34</td>
<td>33</td>
<td>20</td>
<td>21</td>
<td>40</td>
</tr>
<tr>
<td>A-B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q) Is there strong evidence that strain A has a higher mean yield than strain B? Test at $\alpha = 0.05$.

1) Write which method you use
2) Write the hypothesis.
3) $S_D =$?
4) Test statistic?
5) Conclusion ?
6) P-value ?
7) Find 95% CI.
3 Ch 11 Analysing Count Data (p452)

In previous chapters, responses are continuous variables. (sleeping time, height, weight, etc). In this chapter responses are categorical.

eg1)

eg2)

eg3) (conversion from continuous to categorical)
3.1 Introduction

Let’s consider p452 Example 2. They want to compare commercial A and B. They did telephone survey to audience. 80 viewers of Commercial A and 70 viewers of Commercial B were interviewed. There are two options:

Option 1: ”Forget key point of the commercial.”
Option 2: ”Remember key point of the commercial.”

Here is the result for total 150 viewers.

<table>
<thead>
<tr>
<th></th>
<th>Forgot Key Point</th>
<th>Remember it</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial A</td>
<td>43</td>
<td>37</td>
<td>80</td>
</tr>
<tr>
<td>Commercial B</td>
<td>52</td>
<td>18</td>
<td>70</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>55</td>
<td>150</td>
</tr>
</tbody>
</table>

Based on this, can you conclude

C1) Commercial A is better than Commercial B?
C2) Commercial A is worth than Commercial B?
C3) Commercial A is different from Commercial B?
We will learn how to do those hypothesis test soon.
3.2 Inferences about a proportion

Use the same example, but we focus on Commercial A. This is one population problem, not comparing two population. Let $p \in [0, 1]$, such that

$$p = \text{success rate} \ (\text{Remember key point})$$  \hspace{1cm} (2)

\begin{center}
\begin{tabular}{l|c|c|c}
 & Forgot Key Point & Remember it & Total \\
\hline
Commercial A & 43 & 37 & 80 \\
\end{tabular}
\end{center}

In other word, $p =$ Population proportion. (p455)

Note: If you have $n$ sample from this Bernoulli trial $Y_1, Y_2, \ldots, Y_n$,
then the total number of success
$X = Y_1 + Y_2 + \ldots + Y_n$
is a random variable, modeled by

\[
\begin{bmatrix}
\end{bmatrix}
\]

In the Commercial example, $n =$
$Y_1 =$
$Y_2 =$
$X =$

What is the relation between population mean/variance for $Y_i$ and $X$? (p454) Express followings, using $n$ and $p$. 
\[ E[Y_1] = \text{(use original formula of Expectation)} = \]

\[ E[X] = \]

\[ E[Y] = \]

\[ Var[Y_1] = \text{(use original formula of Variance)} = \]

\[ Var[X] = \]

\[ Var[Y] = \]

Q) Our main interest is the actual value of the parameter \( p \in [0, 1] \). Which random variable do you use to guess \( p \)?

1) \( Y_1, \ldots, Y_n \)  
2) \( X \)  
3) Both

ANS):
Researchers interest can be one of them:

C1: \( p > 0.5 \)
C2: \( p < 0.5 \)
C3: \( p \neq 0.5 \)

Deja vu! Yes, under the condition

**Large sample: \( n > 30 \)**

we can use similar hypothesis test methods and get \( p \)-value, and CI like we did in Ch8. See LEC25 note.

Here it is. (p459)

<table>
<thead>
<tr>
<th>Case</th>
<th>nick name</th>
<th>( H_0 )</th>
<th>( H_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>One sided hypothesis</td>
<td>( H_0 : p = p_0 )</td>
<td>( H_1 : p &gt; p_0 )</td>
</tr>
<tr>
<td>C2</td>
<td>One sided hypothesis</td>
<td>( H_0 : p = p_0 )</td>
<td>( H_1 : p &lt; p_0 )</td>
</tr>
<tr>
<td>C3</td>
<td>Two sided hypothesis</td>
<td>( H_0 : p = p_0 )</td>
<td>( H_1 : p \neq p_0 )</td>
</tr>
</tbody>
</table>

The method below is for \( C1 \). For \( C2, C3 \), all we need is modify Step1 and Step5.

- Hypothesis test for \( C1 \)
  
  Step 1) Decide \( H_1 \).

  Researchers (you) want to establish \( p > p_0 \) (for example, \( p > 0.5 \) ). We set

  \( H_1 : p > p_0 \)
where $p_0$ is a number researcher decide. For example, 0.1, 0.5, 0.8 etc. Usually given in a problem. But NEVER -3.2 NOR 100! They are beyond the parameter space $p \in [0, 1]$.

Step 2) Decide $H_0$.

Then automatically, we have

$H_0 : p = p_0$

Step 3) Decide level of significance $\alpha \in (0, 1)$ (= Type 1 error). We control $\alpha$ (Type I Error) not $\beta$ (Type II Error). Hence we (Researcher) decide $\alpha$, not $\beta$.

We usually use $\alpha = 0.05$
Review: Type 1 Error, Type 2 Error. (Same concept for Continuous/Categorical response)
Suppose you observe weight of eggs. \(g=\text{gram}\)

1g 2g 3g 8g 5g, 4g, 12.2g, 15g.

You know they are one of them:
1) Frog egg 2) Chicken egg

Now you make a decision (cut off value) (15g). **The purpose is to find a chicken egg. BECAUSE WE WANT TO EAT CHICKEN EGG!**. This is a fixed value. If an egg’s weight is less than this value, you decide the egg belongs to frogs. Otherwise, it belongs to chickens. The problem here is that whatever the cut off value is, error (wrong decision) always happens.

[Ccase A] If you increase the cut off value (15g → 18g),
Type 1 error (increase/ decrease)
Type 2 error (increase/ decrease)

[Ccase B] If you decrease the cut off value (15g → 10g),
Type 1 error (increase/ decrease)
Type 2 error (increase/ decrease)
So, how can we decide the cut off?

1) Missing Chicken egg (**Type 2 Error**) is not serious error, but Eating Frog’s egg (**Type 1 Error**) is CRITICAL ERROR! (If you are a snake, maybe it does not matter...)

2) You can have only one FIXED cut off value (Critical point). You can not chance it once you decided! (at least in this st301 course.)

3) Hence you look at Type 1 Error (area) and decide the cut off.

Then, **regardless of Chicken egg’s distribution, you can control the percentage of mistakenly eating frog egg.**

[Good way of deciding the cut off]
If you look at Type 2 Error (area) and decide the cut off, depending on frog egg’s distribution, the percentage of mistakenly eating frog egg changes !. In other word, we are not controlling Type 1 error. This is bad !

[Bad way of deciding the cut off]
Step 4) Get $Z_{obs}$ (TEST STATISTIC for $p$).

$$Z_{obs} = \frac{\bar{Y} - p_0}{S}$$ (3)

where $S$= sample sd. i.e.,

$$S =$$ (4)

**Review and comparison:** In Ch8, we had

$$S = \sqrt{\frac{\sum_i(Y_i-\bar{Y})^2}{n-1}}$$

Why is it different? In general, the Test Statistic is of the form

$$Test Statistic = \frac{A - E[A]}{SD(A)}$$ (5)

In Ch8, $A = \bar{X}$. Hence,

$$E[A] = E[\bar{X}] =$$

$$Var[A] = Var[\bar{X}] =$$

Therefore,

$$Test Statistic = Z_{obs} =$$ (6)
In Ch11 \( A = \bar{Y} \). Hence, (look at today’s beginning part) 
\[ E[A] = E[\bar{Y}] = \]

\[ Var[A] = Var[\bar{Y}] = \]
Therefore,

\[ Test\ Statistics = Z_{obs} = \] \hspace{1cm} (7)

Same is true for Ch9, Ch10. Please check them. (Or you can look at p441 summary).

Step 5-a) Find \textbf{critical point} = c

c = location of \( N(0,1) \) which has right area = \( \alpha \). Use \( Z \)-table. \( ( = Z_\alpha \). See Fig324 Fig)

Step 5-b) Find \textbf{rejection region}

Step 5-c) Find \textbf{p-value} \( \in (0, 1) \)

DEF: \textbf{p-value} = Area right side of \( Z_{obs} \)
Step 6) Conclude your hypothesis test (Only two options!)

Option 1: If $Z_{obs} \in \text{Rejection Region}$,
(or if p-value $\leq \alpha$) then
Conclusion: "Reject $H_0$.”

Option 2: Otherwise,
Conclusion: ”NOT reject $H_0$.”

Note: In step 6, to get a conclusion, all you have to do is just heck one of them:

Check criteria 1: $Z_{obs} \in \text{Rejection Region}$
Check criteria 2: p-value $\leq \alpha$

You should get the same conclusion. I mean,

”Check criteria 1” = Yes $\Leftrightarrow$ ”Check criteria 2” =Yes $\Leftrightarrow$ ”Reject $H_0$.”. 

”Check criteria 1” = No $\Leftrightarrow$ ”Check criteria 2” =No $\Leftrightarrow$ ”NOT Reject $H_0$.”. 
Review: How to decide a conclusion of the hypothesis test? (C1 case)? Use one of them.

Method 1) Rejection Region (not depend of data) and $Z_{obs}$ (data dependent).

Method 2) $\alpha$ (not depend of data) and p-value (data dependent)
Estimator of \( p \equiv \hat{p} = \bar{Y} \) (p453).

\( \hat{p} = \text{Sample proportion.} \)

\( p = \text{Population proportion.} \)

\( S.E(\hat{p})^2 = (p455) \)

100(1 − \( \alpha \))% Error margine=

100(1 − \( \alpha \))% CI of \( p \) (p457) :

Determining the sample size (p458): Given 100(1 − \( \alpha \))% error margin \( d \) (some positive number, say 100(1−0.05)% Error margin \( d=0.02 \)), you try to find a sample size. (Look at Ch8 p326).

\[
z_{0.025} \sqrt{\frac{p(1-p)}{n}} \leq 0.02.
\]

Hence,

\[
n \geq p(1-p)(\frac{z_{0.025}}{0.02})^2 = p(1-p)(\frac{1.96}{0.02})^2 = p(1-p)9604
\]
Well, how about $p(1-p)$? We need an actual number. Can we use $\hat{p}(1-\hat{p})$? (Yes/ No).

Now, Look at Ch8 p326. We used

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$$

where $d$ is given (say 0.02) . $z_{\alpha/2}$ is also given. We needed to know $\sigma$ somehow. You can NOT use sample variance formula $S = \sqrt{\frac{\sum_i(Y_i-\bar{Y})^2}{n-1}}$

Because **we need to decide sample size $n$ before the experimentation. Sample variance can be obtained after the experimentation.**

In p326 Example 3, $\sigma$ is given in the problem.

Hence, in our case,

Case 1) $p$ is given (so that you can get $p(1-p)$ ) in the problem,

Case 2) Use 0.25 for $p(1-p)$. Because this is the biggest number $p(1-p)$ can take.

Proof: Set $f(p) = p(1 - p) = $
Case 2 always works but "conservative". Conservative means, your sample size $n$ is the worst case = biggest $n$. If Researcher use this $n$, they need more effort (time/money) than they actually need. But researcher’s boss will not be angry about the result. (the boss might be angry about the expense...)

$$n \geq (1/4)9604 = 2041. \quad (11)$$

Hence, if you have $n=2041$, you are sure, in this example, that the 95% error margin will be less than $0.02(=d)$.

Really? Yes. If true $p=0.3$, then, sample size you really needed was

$$n \geq 0.3(1 - 0.3)9604 = \quad (12)$$
Commercial example] Researchers interest is whether \( p > 0.5 \) or not.

1) Write the hypothesis.
2) \( S = ? \)
3) Test statistic?
4) Conclusion ?
5) P-value ?
6) Find 95% CI.