Paired Data

Paired data often arise when measuring the same individuals twice (before and after a period).

Individual	Weight in 2010	Weight in 2020	Diff
Marion	140	135	-5
Sylvester	190	249	59
Florence	183	183	0
David	90	134	44
Gertrude	208	180	-28
⋮	⋮	⋮	⋮

What would an implied question be?

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Unpaired Data

Two separate random samples would produce unpaired data.

year=2010	
Individual	Weight
Lonzo	140
Rosalia	190
Leora	183
Otis	90
Edward	208
⋮	⋮

year=2020	
Individual	Weight
Henry	310
Harvey	250
Phoebe	210
Donna	150
John	110
⋮	⋮

What would an implied question be?

We will discuss unpaired analysis in Chapter 5.3 (next class).
With paired data, we consider a **mean of differences**.
With unpaired data, we consider a **difference of means**.

Derivation of paired formulas

Let random variable D_i represent the (unknown) difference from a (yet to be) randomly selected individual i.
We want to predict what happens when we find a mean of differences.

$$\bar{D} = \frac{D_1 + D_2 + D_3 + ... + D_n}{n}$$

The central limit theorem still applies!

As $n \to \infty$, $\bar{D}$ becomes normally distributed.

Basically, we can treat these differences just like any other independent and identically distributed random variables.

Note about notation

- I used $\bar{D}$ for the random variable representing an unknown mean of differences.
- I would use $\bar{d}$ for a specific (observed, critical, etc) mean of difference.
- The book uses $\bar{x}_{\text{diff}}$ for both of these concepts. This is misleading, as it looks like a difference of means, not a mean of differences.
- I would at least prefer using $\overline{X_{\text{diff}}}$ and $\overline{x_{\text{diff}}}$ to emphasize we are finding a mean of differences.
- The book's notation of μ_{diff} (for the population's true difference) is useful. We could also use $E(D)$ or μ_D.
- In order to match the book as much as possible, I will now use $x_{\text{diff},i}$ and $\overline{X_{\text{diff}}}$ and $\overline{x_{\text{diff}}}$ and μ_{diff}.

Example problem

A teacher wonders if, on average, a random student will perform about the same on two exams. She decides to run a two-tail t test on a random sample of size $n = 5$ with a signficance level $\alpha = 0.05$. Here are the results of her study:

Student	Exam 1	Exam 2
Norma	98	96
Elliot	15	10
Walton	61	61
Mable	80	79
Loretta	10	8

Perform the t test.

Example problem solution

Find the differences.

i	$x_{1,i}$	$x_{2,i}$	$x_{\text{diff},i}$
1	98	96	-2
2	15	10	-5
3	61	61	0
4	80	79	-1
5	10	8	-2

Find the (differences') sample mean.

$$\overline{x_{\text{diff}}} = \frac{\sum_{i=1}^{n} x_{\text{diff},i}}{n} = \frac{-2-5+0-1-2}{5} = -2$$

Find the (differences') sample standard deviation.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_{\text{diff},i} - \overline{x_{\text{diff}}})^2}{n-1}} = \sqrt{\frac{(0)^2 + (3)^2 + (2)^2 + (1)^2 + (0)^2}{5-1}} = 1.87$$

We are doing a two-tail test with the following:

$$n = 5 \qquad \overline{x_{\text{diff}}} = -2 \qquad s = 1.87 \qquad \alpha = 0.05$$

State the hypotheses.

$$H_0: \quad \mu_{\text{diff}} = 0 \qquad\qquad H_A: \quad \mu_{\text{diff}} \neq 0$$

Determine the critical value, $t^\star$, such that $P(|T| > t^\star) = 0.05$.

$$t^\star = 2.78$$

Find the standard error (the standard deviation of the differences' sampling distribution).

$$SE = \frac{s}{\sqrt{n}} = \frac{1.87}{\sqrt{5}} = 0.837$$

Calculate an observed t score.

$$t_{\text{obs}} = \frac{(-2) - 0}{0.837} = -2.39$$

From the previous slides:

$$n = 5 \qquad \overline{x_{\text{diff}}} = -2 \qquad s = 1.87 \qquad \alpha = 0.05$$
$$t^\star = 2.78 \qquad SE = 0.837 \qquad t_{\text{obs}} = -2.39$$

We can determine a p-value. Remember we are doing a two-tail test, so p-value $= P(|T| > 2.39)$.

$$0.05 \quad < \quad p\text{-value} \quad < \quad 0.1$$

We can compare t_{obs} and $t^\star$. We can also compare p-value and α.

$$|t_{\text{obs}}| < |t^\star|$$

$$p\text{-value} > \alpha$$

Thus, we retain the null hypothesis.

We maintain that maybe students do equally well on both tests.

Practice

The following table has paired data. Test the hypotheses of whether or not the differences have a population average of 0. Use $\alpha = 0.1$.

i	$x_{1,i}$	$x_{2,i}$
1	50	54
2	23	25
3	96	97
4	47	49
5	10	16