Section 9
Statistical Inference for Two Means

9.1 – Comparing Dependent Outcomes

Introduction

Just as we did last section, we will continue expanding statistical inference for making comparisons and evaluating the validity of the results we obtained. Where we focused on categorical data and proportions last section, we will now focus on numerical data and means.

The first case we will look at is based on the idea of collecting two groups of data that are dependent in some way – that is, there is some association or link across the groups. When this occurs, this is known as a _________________ design or sample.

Repeated observations

The first kind of paired design that we will examine is where you take repeated observations on the same subjects. Because you're observing the same subjects in each group of data you collect, there is reason to believe that the first observation is going to be similar to the second observation for that subject. This means that the two samples you have are not independent of each other – the second observation for a subject is dependent on what you observed the first time.

In terms of the arithmetic or code you would run to analyze this type of data, there's actually very little to learn here based on what we know about doing inference on one mean! Consider the following example.

> *Example*: A company that publishes a dieting program is interested in estimating how effective their program is in terms of weight loss. They have recruited 10 participants to participate in a 2-month program, and their weights before and after the program were recorded in the table on the next page. Conduct a hypothesis test at level $\alpha = 0.05$ to determine if the weight loss program is effective, and evaluate the validity of the results based on the study design.

Weight before	Weight after	Difference
176	166	
251	236	
244	243	
180	173	
150	141	
199	179	
141	136	
289	261	
310	313	
275	250	

Related observations

In the paired case, while our data may have originally been organized into two groups, these groups have a dependency – that is, there is a link between certain subjects in each of the two groups. In this previous example, the link was based on the fact that we took two separate measurements on the same subjects. So while it may appear that there are two different groups of data that we want to compare, in reality, we're really interested in examining the paired differences. Let's see another example:

> *Example*: An experiment to compare the tire wear of two brands of tires is conducted. A sample of 12 used tires is selected from two brands, A and B, and they are each placed on one of the axles of 12 cars. The axle that each tire is attached to is randomly chosen such that an equal amount are assigned to each axle (3 to front left, 3 to rear left, etc.) across the 12 cars. Each car was driven for six weeks, and at the end, the amount of wear in thousandths of an inch was recorded. That information is provided in the table below:

Car	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Mean**	**Std Deviation**
Brand A	100	77	37	66	90	110	76	44	82	66	38	89	72.9	23.8
Brand B	117	91	48	71	97	118	80	48	84	66	36	82	78.2	26.1
Diff = A - B	-17	-14	-11	-5	-7	-8	-4	-4	-2	0	2	7	-5.3	6.8

> Conduct a test at level $\alpha = 0.05$ to determine if there are any differences in wearing with these two brands of tires.

Thus, we can see conducting a hypothesis test on a paired mean of differences is just like conducting a hypothesis test for a single mean, since we are able to just take the difference for each set of observations and use that to conduct the test. There are two main ways we can identify when the two samples we conducted are paired:

1. Observations in each sample are just different measurements of the same subject (pre/post, before/after testing)
2. There is some way in which observations in each of the two groups are linked (parent-child relationship, tires from the same car, married couples)

In these cases, we are simply testing the population mean difference, μ_d. We denote the means with a *d* subscript to make it clear that this is based on the difference in two observations. Thus, the same assumption of normally distributed data or a large enough sample of pairs ($n \geq 25$) still applies to testing in this scenario.

Paired samples in R

To conduct this test in R, we would use the same functions as we learned in Section 7, once the differences were found. However, data may not always be presented in terms of differences. Using the **weightloss.csv** data file, we can compute a new variable for the differences:

```
weightloss$diff = weightloss$before - weightloss$after
```

As there are only 10 observations in this sample, we should check that the differences are normally distributed. Note that it does not matter if the before weights or after weights are normally distributed, since we are not building a confidence interval for those weights.

```
qqnorm(weightloss$diff)
qqline(weightloss$diff)
```

We can see from the plot above that it is reasonable to assume that the data came from a normal distribution. Thus, the test we conduct below is valid:

```
t.test(weightloss$diff, mu=0, alternative=ALT)
```

9.2 – Comparing Independent Samples

Theoretical background

Another way we can draw data into two samples is if they were *independently* drawn from completely different populations, rather than taking two samples from the same group of subjects or from linked subjects. Instead of estimating a mean of differences, μ_d, we would be estimating a difference of means, $\mu_1 - \mu_2$, where each μ represents the mean of two separate populations. Clearly, the best way to estimate this difference in population means is by taking the difference in sample means, as

$$E(\bar{x}_1 - \bar{x}_2) =$$

Now, we just need to derive the variability of a mean difference in order to construct a confidence interval:

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) =$$

And since the CLT guarantees each mean to be normally distributed, we know that the difference in means will also be normally distributed. However, CLT only applies if each of our samples has at least 25 observations, that is, both $n_1 \geq 25$ and $n_2 \geq 25$, otherwise, we would have to assume that both of our populations are normal.

With estimated standard deviations s_1 and s_2 for each mean, we know that the difference in sample means follows a *t*-distribution, as we saw with the case of the single mean. Thus, we get the following test statistic for testing the difference in means:

$$t =$$

Because two means are being estimated here, the degrees of freedom are somewhat complicated to compute compared to the one mean case:

$$df =$$

We can also use the expected value/variance to derive a confidence interval for a difference in means, as shown below:

$$\bar{x}_1 - \bar{x}_2 \pm$$

Testing a difference in means by hand

We will start with an example of a test that uses the equations from above.

> *Example*: A clinical trial for high blood pressure is conducted, where 70 subjects are randomly assigned to a control and treatment groups. The control group is given a placebo drug, and the treatment group is given a real drug designed to lower blood pressure. The systolic blood pressure is measured of each subject after 2 weeks of regularly taking the drug/placebo. The summary statistics for this experiment are given in the table below:

	Mean	Standard Deviation	Sample Size
Treatment	140	8	35
Control	150	10	35

> Conduct a hypothesis test at level $\alpha = 0.01$ to determine if this new medication is effective in reducing blood pressure.

Unlike the paired mean case previously, we do not usually represent data from two populations in two columns where each column represents a different population. Typically, each row in a data set represents one person or observation, and using this paired setup would put two separate observations in the same row. Our data for this example might have been stored as follows:

Patient Number	Group	Blood Pressure
1	Treatment	136
2	Control	151
3	Treatment	145
⋮	⋮	⋮

This format for data storage is important, as when comparing two populations, it is possible to have groups of unequal sizes to compare. While this example did have 35 in each group, it is not a requirement to have equal sized groups, and by placing one patient per line, you do not restrict yourself to equal groups.

R code for testing a difference in means
Let's try an example in R where we have data stored in a similar fashion.

> *Example*: How do cash offers for used car trade-ins differ for people of differing ages? A random sample of trade-ins at a dealership for cars of a similar type was taken, and the seller's age group (young vs. middle aged) was recorded along with the cash offer in hundreds of dollars in the data set **cashoffer.csv**. Conduct a hypothesis test at level $\alpha = 0.05$ to determine if the age of a seller makes a significant difference in the cash offer given.

```
t.test(offer ~ ageGroup, data = cashoffer)
```

Using simulation to conduct a difference in means test
In class this week, we conducted a study in class on memorization. Some of you were given the memorization test with bad chunking, and others with good chunking of letters, which likely had a noticeable effect on how many letters you memorized. In class, you were randomly assigned into one of the two chunking groups, so we sampled each variable without replacement to emulate the sampling structure. Let's think about the validity of the results from this study based on this type of design:

> *Example:* Can we generalize the results of our in-class memorization study to a larger population?
>
> Can we conclude that there is a cause-and-effect relationship between the chunking of letters and memorization?

Since the data for this study changes every class I teach it, let's try re-creating the simulation with the previous example on the offers for car trade-ins. Re-read the design of that study, and let's consider these questions again:

> *Example:* What kind of random process was used in this study? How would we implement that in terms of the type of replacement we would use in a sampler?
>
> Can we generalize the results of the car trade-in study to a larger population?
>
> Can we conclude that there is a cause-and-effect relationship between your age and the trade-in offer received?

The responses above show that this study represents an observational study. This is very similar to the blood donations study we simulated in the last section, but now our response variable is numeric instead of categorical. Otherwise, we can use a similar simulation process as before, but this time, we collect on the difference in means instead of percentages! We'll work with the `cashoffer` data set directly below once loaded into R. Remember that when we worked with this data above, we found a difference in means of 2.25, or $2,250.

```
age_rand = sample(cashoffer$ageGroup, 24, replace=F)
offer_rand = sample(cashoffer$offer, 24, replace=T)
car_rand = data.frame(age_rand, offer_rand)
mu_1 = aggregate(offer_rand~age_rand, data=car_rand,
     FUN=mean)$offer[1]
mu_2 = aggregate(offer_rand~age_rand, data=car_rand,
     FUN=mean)$offer[2]
mu_1-mu_2
```

Based on what we've learned in class already, we could also use the `subset()` function to create a data set for the young and middle aged people and take means there too. I used `aggregate()` instead to take means across age groups directly.

```
diffs = rep(0, 1000)

for (i in 1:1000) {
  age_rand = sample(cashoffer$ageGroup, 24, replace=F)
  offer_rand = sample(cashoffer$offer, 24, replace=T)
  car_rand = data.frame(age_rand, offer_rand)
  mu_1 = aggregate(offer_rand~age_rand, data=car_rand,
       FUN=mean)$offer[1]
  mu_2 = aggregate(offer_rand~age_rand, data=car_rand,
       FUN=mean)$offer[2]
  diffs[i] = mu_1-mu_2
}

hist(diffs)
abline(v=2.25, col="red")
abline(v=-2.25, col="red")
mean(diffs >= 2.25 | diffs <= -2.25)
```

We should find here that the *p*-value we get by taking the percentage of all differences in means that were equal to or greater than the 2.25 or less than or equal to -2.25, as we observed a difference of means of 2.25 in the real, non-simulated data. How does this compare to what we found with `t.test` earlier?

9.3 – Additional Practice

Example: From 1968 to 1972, a series of weather modification experiments were conducted in south Florida. These experiments were designed to test a hypothesis called "dynamic seeding" in which it is postulated that massive silver iodide seeding in cumulus clouds leads to increased precipitation.

To test this hypothesis, the researchers selected 44 isolated, growing cumulus clouds, over the course of the study. The clouds were as identical to one another as possible. These clouds were randomly assigned to either a treatment (seeding with silver iodide), or control (no seeding) condition. A calibrated radar was then used to measure the total rain volume falling from the base of the cloud. The resulting set of data to be analyzed here consists of the measurements (in acre-ft) for the 22 control clouds and the 22 seeded clouds. One acre-foot is the amount of water covering 1 acre to a depth of 1 foot. The data for these experiments can be found in **cloudseeding.csv**.

Conduct a hypothesis test at significance level $\alpha = 0.02$ to determine if cloud seeding is effective. Calculate a 96% confidence interval for the difference in the mean rain volume between seeded and unseeded clouds, and explain how this also verifies the results of your test.