Stat 324 – Introduction to Statistics for Engineers

LECTURE 13: SAMPLING DISTRIBUTION, HYPOTHESIS TESTING, AND CONFIDENCE INTERVALS FOR TWO DEPENDENT/PAIRED POPULATIONS

Researchers designed an experiment to study the effects of caffeine withdrawal. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression.

One way to test this is to compare the mean depression score of people who drink caffeine to mean depression score of those that do not.

Hypotheses:

H_0 : $\mu_{placebo} - \mu_{caffeine} = 0$ $\quad$ H_A : $\mu_{placebo} - \mu_{caffeine} > 0$

Where μ_C is the mean depression score of people who use caffeine and
μ_{Plac} is the mean depression score of people who do not use caffeine

Significance Level: 5%

Data Collection: *The researchers recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from caffeine substances for the duration of the experiment and for a wash-out period of 1 week before. During one 2-day period, subjects received their normal caffeine intake in a capsule and during another 2-day period, they took a placebo capsule. The order in which subjects took the caffeine and the placebo was* ***randomized*** *(and there was a 1 week period between treatments). At the end of each 2 – day period, a test for depression was given to all 11 subjects.*

Depression scores for the 11 subjects after the caffeine and placebo treatments are given below.

Subject	1	2	3	4	5	6	7	8	9	10	11
Caffeine	5	5	4	3	8	5	0	0	2	11	1
Placebo	16	23	5	7	14	24	6	3	15	12	0

Sample Statistics and Graphing.

In the past, we summarized the two groups with statistics and graphs separately. This made sense when the groups were independent. Are the two samples and populations independent? caffeine depression scores, placebo depression scores

NO, there are two measurements (Pi, Ci) being taken on each individual

In ***matched pairs*** experiments, randomly assign both treatments to the same (or very similar) subjects so changes in measurement are most likely due to treatment

- Testing a drug on pairs of "twins" to get rid of variability due to age, gender, genetics, etc.
- Testing the same subjects twice, the order of test is randomly assigned and washout time often given.
- We expect Ci and Pi to be relatively large or small together, but taking the differences within each subject filters out the subject/"block" effect.

Instead of looking at the raw data: $\{(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)\}$, we look at a random sample from the population of differences: $\{D_i = X_i - Y_i\}$ or $\{D_i = Y_i - X_i\}$. We could compute differences in either direction, I just did it consistent with my ordering in the original hypotheses. $H_0: \mu_{placebo} - \mu_{caffeine} = 0$

Subject	1	2	3	4	5	6	7	8	9	10	11
Caffeine	5	5	4	3	8	5	0	0	2	11	1
Placebo	16	23	5	7	14	24	6	3	15	12	0
Diff: placebo - caffeine	11	18	1	4	6	19	6	3	13	1	-1

We can do a variety of **one-sample tests** on our new data set of differences, depending on our questions of interest and the assumptions we are willing to make.

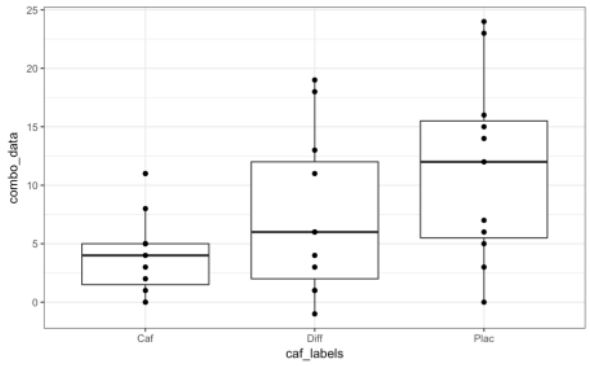
Interested in looking at mean difference:

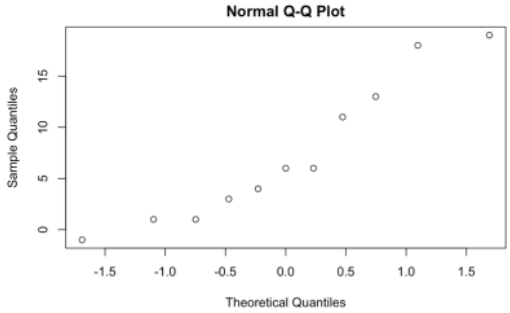
- 1 sample z test if we know pop of differences is normal and σ Known or sample size is large
- 1 sample t test if we know pop of differences is normal, but σ unKnown.
- Bootstrap for non-normal population of differences

Interested in shifts in distributions

- Sign Test or
- Wilcoxon Signed-Rank Test

Start with graphical and numeric summaries (plots of the Difference matter!)





	Caf	Plac	Plac-Caf
Sample Size (n)	11	11	11
Mean	4.0	11.36	7.364
Median	4.0	12	6.0
Sample SD	3.38	7.90	6.92

From the graphs and summary measures we see:

Normality of population is not unreasonable based on linearity of qq plot

Differences are almost all positive and look approximately symmetric

linear enough for t-test

Assumptions:

1. Independent Observations if we look at the sample of differences.
2. Normality of population is reasonable

We could answer this by looking at the average difference in depression score

Calculate a Test statistic:

Reword hypotheses:

Let $D = P - C$ the difference in depression scores under placebo and under caffeine for each individual.

$H_o: \mu_D = 0$ and $H_A: \mu_D > 0$

Point estimate for $\widehat{\mu_D}$ is the estimator $\overline{X_D}$. Since σ_D is unknown, we'll use $\widehat{\sigma_D} = S_D$

Our test statistic: $\frac{Obs_{AVG} - EV_{AVG}}{\approx SE_{AVG}}$ $\frac{\overline{X}_D - 0}{S_D/\sqrt{n}}$ ← # of differences

and our observed value is: $t_{obs} = \frac{7.364 - 0}{6.92/\sqrt{11}} = 3.529$

	Diff=Pla-Caf
Sample Size (n)	11
Mean	7.364
Median	6.0
Sample SD	6.92

Calculate p value and draw conclusion in context of question

$H_o: \mu_D = 0$ and $H_A: \mu_D \geq 0$. Where D is the difference in depression scores under placebo and under caffeine for each individual.

$P(T_{10} \geq T_{obs} = 3.529) = pvalue$ and 0.0025 <pvalue< 0.005

There is enough evidence at the 5% level to reject the null. Evidence suggests mean depression scores for caffeine dependent individuals is higher under the placebo than under caffeine treatment.

```
> t.test(PminCaf,  alternative="greater")

   One Sample t-test
  data:  PminCaf
  t = 3.5304, df = 10, p-value = 0.002721
  alternative hypothesis: true mean is greater than 0
```

```
> t.test(placebo,caffeine,alternative="greater", paired=TRUE)

  Paired t-test
  data:  placebo and caffeine
  t = 3.5304, df = 10, p-value = 0.002721
  alternative hypothesis: true difference in means is greater than 0
```

Inferences about the Mean Difference Computed from a Matched Pair

When each measurement in one sample is ***matched*** or ***paired*** with a particular measurement in the other sample, we focus on inferences about the mean difference. Compute the difference in the n pairs of data: $D_i = X_i - Y_i$.

Assumptions for all hypothesis test methods:

1. The D_i's need to be independent (where the pairs randomly chosen?)

Assumptions for ***Matched Pair T Test***

2. The population of D_i's is normal or sample size is large enough that CLT makes sampling distribution of $\overline{X_D}$ approximately normal.

Hypotheses: $H_0: \mu_D = D_0$; $H_A: \mu_D < D_0$, $H_A: \mu_D > D_0$, $H_A: \mu_D \neq D_0$ where D_0 is a difference of interest

Test statistic: $T = \frac{\overline{X_D} - D_0}{\frac{s_D}{\sqrt{n_D}}} \sim \mathbf{T_{n_D - 1}}$

where $\overline{X_D}$, s_D, and n_D are the mean, sample sd, and number of observed differences.

P value: if $H_A: \mu_D < D_0, P(T_{n_D-1} \leq t_{obs})$, $H_A: \mu_D > D_0, P(T_{n_D-1} \geq t_{obs})$, or $H_A: \mu_D \neq D_0, 2 * P(T_{n_D-1} \geq |t_{obs}|)$

(1-α)% CI: $\overline{X_D} \pm t_{\alpha/2} \frac{s_D}{\sqrt{n_D}}$

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers …At the end of each 2 day period, a test for depression was given to all 11 subjects. …**researches wonder whether being deprived of caffeine would lead to an increase in depression.**

We could also answer this question by looking at whether there was a increase in each depression score and performing a Sign test. In the paired case, the sign test will be about whether the median of the differences is equal to a specified value. (If distribution of differences is symmetric, test about mean difference)

Test Statistic: Let $D = P - C$, and B=Number of Increases in Depression. (+)

Under the null of no increase in depression score we'd expect about half of D to be above 0 and half to be below 0, so $H_o: Med_D = 0,\ vs\ H_A: Med_D > 0$ $B \sim Bin(11, 0.5)$ (+ or −), Observed: b=10 (+)

Subject	1	2	3	4	5	6	7	8	9	10	11
Caffeine	5	5	4	3	8	5	0	0	2	11	1
Placebo	16	23	5	7	14	24	6	3	15	12	0
Difference: (P-C)	11	18	1	4	6	19	6	3	13	1	-1
Sign	+	+	+	+	+	+	+	+	+	+	−

…**being deprived of caffeine would lead to an increase in depression.**

B=Number of Increases in Depression.

Under $H_o: Med_D = 0,\quad B \sim Bin(11, .5)\qquad H_A: Med_D > 0$

Pvalue: $P(B \geq b_{obs}) = P(B \geq 10) = P(B = 10) + P(B = 11) = 0.0054 + 0.0005$

$= 0.0059$

B	0	1	2	3	4	5	6	7	8	9	10	11
P(B=b)	0.0005	0.0054	0.0269	0.0806	0.1611	0.2256	0.2256	0.1611	0.0806	0.0269	0.0054	0.0005

Conclusions: we have strong evidence against the null evidence suggests the median of the differences > 0. Equivalently evidence suggest caffeine withdrawl leads to increased depression

...being deprived of caffeine would lead to an increase in depression?

The **Wilcoxon Signed-Rank test** makes use of the sign **and** the magnitude of the rank between pairs and is an alternative to the paired t test when the population distribution of differences is nonnormal.

Assumptions:
1. Independence of Differences.
2. Population of Differences symmetric about unknown median M .

Signed rank (one sample)

rank sum (two sample)

Hypotheses:
1. H_0: distribution of differences is symmetric about M_0
2. H_A: distribution of differences tend to be greater or less or shifted from M_0

Compute Test Statistic
1. Calculate differences in the n pairs of observations (Let $D_i = P_i - C_i$)
2. Subtract D_0 (if interested in difference other than 0)
3. Only consider nonzero differences
4. Rank absolute values of the ranks (assigning the average for ties)
5. The test statistic is the sum of the ranks of the negative (wlog positive values). This is then compared to the distribution of rank sums that would be generated if the signs had been randomly assigned.

...being deprived of caffeine would lead to an increase in depression?

ranks: $\frac{1+2+3}{3} = 2$ tied so give AVG

H_0 distribution of differences in depression scores between placebo and caffeine is symmetric about $M_0 = 0$
H_A: patients tend to have higher depression scores on placebo compared to caffeine (Differences P-C tend to be shifted above $M_0 = 0$)

Subject	1	2	3	4	5	6	7	8	9	10	11
Difference: (P-C)	11	18	1	4	6	19	6	3	13	1	-1
Abs value rank for Diff-0	8	10	2	5	6.5	11	6.5	4	9	2	2

Computation is complicated, so we leave it to R (Section 6.5 in book if interested):

Thus we reject H_0 at the 5% level.

Evidence suggests that placebo treatment tended to result in higher depression scores for volunteers with caffeine dependency

```
> wilcox.test(x=caffeine, y=placebo, alternative="less", paired=TRUE)
cannot compute exact p-value with ties
        Wilcoxon signed rank test with continuity correction

data:  caffeine and placebo
V = 2, p-value = 0.003279
alternative hypothesis: true location shift is less than 0
```

Paired T test vs

Sign. Test vs

Wilcoxon Signed-Rank test vs

1 Sample Bootstrap

Sign test, Wilcoxon Signed-Rank, Bootstrap tests do not require Normality assumption.

When the distribution of differences is normal, however the t test has greater power than the sign or signed rank test to determine a difference in means or distribution. This is logical, because the t test (and 1 sample bootstrap) uses the magnitudes of the observations rather than just their relative magnitudes(ranks).

When the population distribution has severe skewness, extreme outliers, or very heavy-tailed, the ranks and therefore Sign and Wilcoxon are better able to detect a shift in the population distributions. Bootstrap and paired T tests will be more likely affected by outliers.

If conclusions of tests agree and there are no blatant violations of required conditions, it doesn't matter which test you do!

2 Sample Paired or Independent Analysis?

1.A hypothesis of ongoing clinical interest is that vitamin C prevents the common cold. In a study involving 20 volunteers, 10 are randomly assigned to receive vitamin C capsules and 10 are randomly assigned to receive placebo capsules. The number of colds over a 12 month period is recorded. independent + 2 sample

2. A topic of current interest in ophthalmology is whether or not spherical refraction is different between the left and right eyes. To examine this issue, refraction is measured in both eyes of 17 people.

paired

17 → L R

compute 17 differences, then do 1 sample tests

3.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 and the remaining twelve were left untreated.

2 independent treatment groups

4.To determine whether cardiologists and pharmacists are equally knowledgeable about how nutrition and vitamin K affect anticoagulation therapy (to prevent clotting), an investigator has 10 cardiologists and 10 pharmacists complete a questionnaire to measure what they know. She contacts the administrators at 10 hospitals and asks the administrator to select a cardiologist and pharmacist at random from the hospital's staff to complete the questionnaire.

Paired

Hospital → Cardiologist, Pharmacist