Week 10 Discussion

Hanying

Announcement

Next week's discussion will be Q&A section. I will appreciate it if you can let me know what question/concept confuse you in advance.

Concept clarification: hw 7 Ex 1 (c)

Power / Type I error(α) / Type II error (β)

(Visualize the type I error and type II error. Use the figure to illustrate how α, population SD σ, sample size n, and the difference between the true population mean and the μ_0 in the hypothesis will influence the power.)

Exercise 1

Question

Possible relationships between mothers' characteristics and birth outcomes have been studied extensively. Using the `lbw` data set, consider what inference tools we could apply to address the following questions. This data was collected as part of a larger study at Bayside Medical Center, Springfield, Massachusetts in 1986. It contains information on 189 births to women that were seen in the obstetrics clinic.

a. Load the lbw data into your environment by reading in the csv file `lbw.csv` to the variable lbw.

I have included vectors of the relevant data at the bottom of this page if you are having issues importing and would rather just define vectors of data.

```
lbw <- read.csv("lbw.csv", header = TRUE)
```

Run `str(lbw)`. Confirm that you see the 11 variables types. We will be focusing on the variables bwt, lwt, and smoke. Make sure R has read the data type for those 3 variables correctly. Set any of those three variables that should be categorical using the as.factor() command. Check that they have been set correctly with str().

```
str(lbw)
lbw$smoke <- as.factor(lbw$smoke)
lbw$low <- as.factor(lbw$low)
str(lbw)
```

We will focus on the continuous variable `bwt` (birthweight) compared across the groups with `smoke = 0` (non smoking) and `smoke = 1` (smoking).

b. Create side by side boxplots and comparative histograms of the variable `bwt` (Birth Weight) between babies with mothers who had smoking status of 0 and 1. Compare the center, variability and shape of the samples' data.

```
SmokeY <- subset(lbw, smoke == "1")
SmokeN <- subset(lbw, smoke == "0")

bwtSmokeY <- SmokeY$bwt
bwtSmokeN <- SmokeN$bwt
```

c. What tools might we apply to perform a t-test or construct a 98% CI for the difference in mean birthweight for babies whose mothers who reported smoking during pregnancy and those who did not? Discuss whether the assumptions of each of the tests are met, based on the plots in part (b).

Equal Variance T Tools?

Welch's T Tools?

Bootstrap T Tools?

d. Calculate the observed t test statistic and p value for the hypothesis test of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ assuming equal population variances. Confirm your findings with t.test and look at the corresponding 98% CI. You can use the summary values below:

Birthweight Summary Values	Smoke: Yes	Smoke: No
mean	2772.297	3054.957
sd	659.8075	752.409
size	74	115

Point Estimate for $\mu_{bwtN} - \mu_{bwtY}$

Approximate SE for $\bar{X}_{bwtN} - \bar{X}_{bwtY}$ *assuming population variances are equal*

Observed Test Statistic

P-value

Confirm findings in R

e. Use `t.test()` on this same data to perform a Welch's T test of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ at a 2% significance level. Also read off the 98% t CI for $\mu_{bwtN} - \mu_{bwtY}$ that results when you allow variances to differ. How do the results compare to (d)? (If you want more practice, you should try to confirm these calculations "by hand" from the summary values given.)

f. Also perform a bootstrap t-tesst of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ at a 2% significance level. How do the results compare to your findings in (d) and (e)?

```r
# 1. Calculate data summaries and t_obs
xbar1 <- mean(bwtSmokeN)
s1 <- sd(bwtSmokeN)
n1 <- length(bwtSmokeN)
xbar2 <- mean(bwtSmokeY)
s2 <- sd(bwtSmokeY)
n2 <- length(bwtSmokeY)

t_obs <- (xbar1 - xbar2)/sqrt((s1^2/n1) + s2^2/n2)

# Specify number of bootstrap samples, create a vector to store t_hat values
B <- 5000
t_hat <- numeric(B)

# Bootstrap loop
set.seed(371)
for (i in 1:B) {
    # 2. Draw a SRS of size n1/n2 from data, with replacement
    x1_star <- sample(bwtSmokeN, size = n1, replace = T)
    x2_star <- sample(bwtSmokeY, size = n2, replace = T)

    # 3. Calculate resampled mean and sd
    xbar1_star <- mean(x1_star)
    s1_star <- sd(x1_star)
```

```r
    xbar2_star <- mean(x2_star)
    s2_star <- sd(x2_star)

    # 4. Calculate t_hat, and store it in vector
    t_hat_numer <- (xbar1_star - xbar2_star) - (xbar1 - xbar2)
    t_hat_denom <- sqrt((s1_star^2/n1) + (s2_star^2/n2))

    t_hat[i] <- t_hat_numer/t_hat_denom
}

# Optional: plot approximate sampling distribution and add a line for t_obs
hist(t_hat, main = "Approx. Sampling Distribution of t")
abline(v = t_obs, col = "dodgerblue", lty = 2, lwd = 3)

# Compute 2-sided p-value Count t_hat values above/below t_obs
m_low <- sum(t_hat < t_obs)
m_hi <- sum(t_hat > t_obs)

# Take 2x the minimum proportion
2 * (min(m_low, m_hi)/B)
```

Solution

(a)

```
lbw <- read.csv("lbw.csv", header = TRUE)
# str(lbw)
lbw$smoke <- as.factor(lbw$smoke)
lbw$low <- as.factor(lbw$low)
# str(lbw)
```

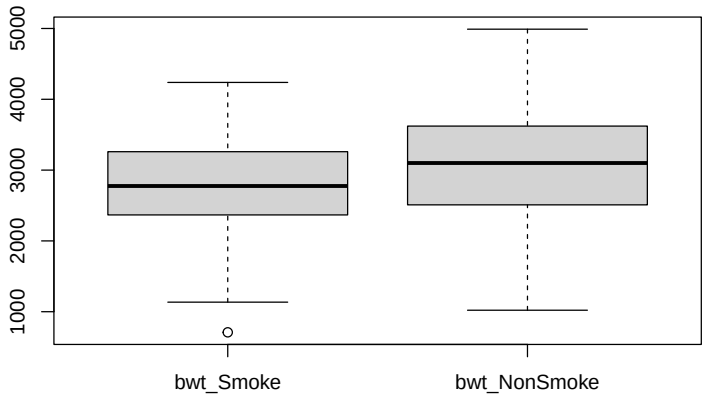
(b)

Create side by side boxplots and comparative histograms of the variable `bwt` (Birth Weight) between babies with mothers who had smoking status of 0 and 1. Compare the center, variability and shape of the samples' data.

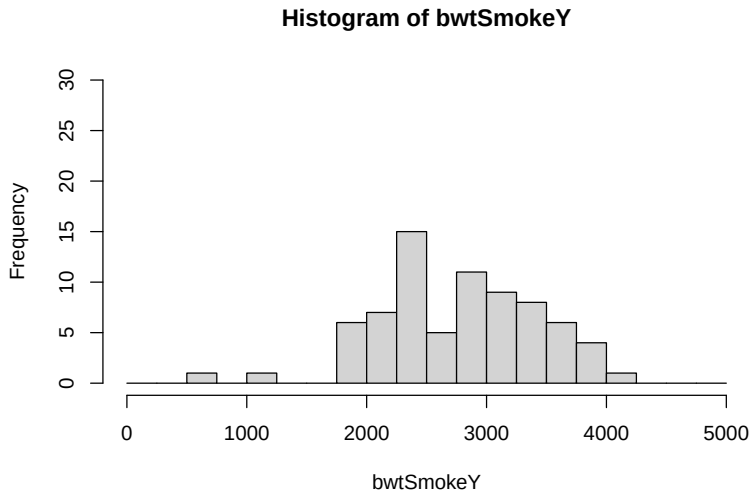
```
SmokeY <- subset(lbw, smoke == "1")
SmokeN <- subset(lbw, smoke == "0")

bwtSmokeY <- SmokeY$bwt
bwtSmokeN <- SmokeN$bwt
```

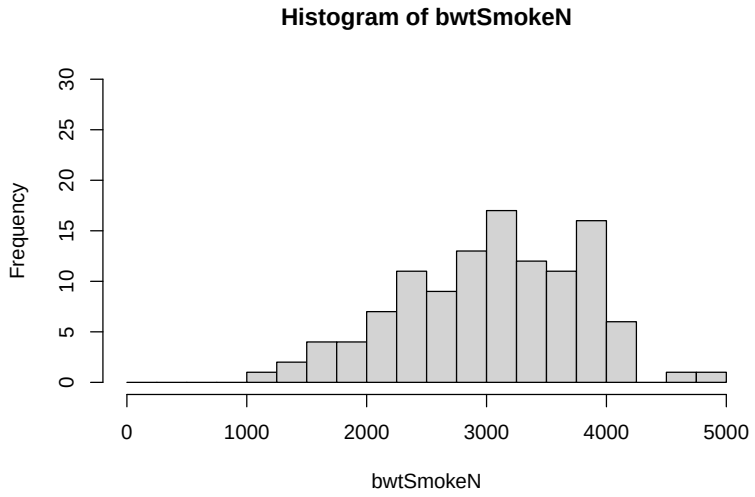
```
boxplot(bwtSmokeY, bwtSmokeN, names = c("bwt_Smoke", "bwt_NonSmoke"))
```




```
hist(bwtSmokeY, breaks = seq(0, 5000, 250), ylim = c(0, 30))
```




```
hist(bwtSmokeN, breaks = seq(0, 5000, 250), ylim = c(0, 30))
```



Notice that we have used the same break points and y-axis range for both histograms for comparison.

- Center: The center of the nonsmoking sample looks slightly higher.
- Variablility: roughly similar (similar SD and range).
- Shape: roughly symmetric and both unimodal.

(c)

What tools might we apply to perform a t-test or construct a 98% CI for the difference in mean birthweight for babies whose mothers who reported smoking during pregnancy and those who did not? Discuss whether the assumptions of each of the tests are met, based on the plots in part (b).

```
sd(bwtSmokeY)
```

```
## [1] 659.8075
```

```
sd(bwtSmokeN)
```

```
## [1] 752.409
```

```
length(bwtSmokeY)
```

```
## [1] 74
```

```
length(bwtSmokeN)
```

```
## [1] 115
```

Equal Variance T Tools?

Welch's T Tools?

Bootstrap T Tools?

Sample size is large, thus $\bar{X}_{bwtN}$ and $\bar{X}_{bwtY}$ are approximately normal. Thus, we can use Welch's T test as it allows for different population variances.

The sample SD is closed to each other. The equal variance assumption is valid. Thus, we can also use Equal variance T test.

We can also use bootstrap. Although as sample size is large enough, we typically go with the inferential t tests.

(d)

Calculate the observed t test statistic and p value for the hypothesis test of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ assuming equal population variances. Confirm your findings with t.test and look at the corresponding 98% CI. You can use the summary values below:

Birthweight Summary Values	Smoke: Yes	Smoke: No
mean	2772.297	3054.957
sd	659.8075	752.409
size	74	115

Point Estimate for $\mu_{bwtN} - \mu_{bwtY}$

Use sample mean as point estimate of population mean:

$$\bar{X}_{bwtN} - \bar{X}_{bwtY} = 3054.957 - 2772.297 = 282.66.$$

Approximate SE for $\bar{X}_{bwtN} - \bar{X}_{bwtY}$ assuming population variances are equal

Estimated pooled variance for the population:

$$
\begin{aligned}
S_p^2 &= \frac{(n_{smokeN}-1)*s^2_{smokeN} + (n_{smokeY}-1)*s^2_{smokeY}}{n_{smokeN}+n_{smokeY}-2} \\
&= \frac{73*659.8075^2 + 114*752.409\text{^}2}{74+115-2} \\
&= 515068.7.
\end{aligned}
$$

Estimated SE for $\bar{X}_{bwtN} - \bar{X}_{bwtY}$:

$$
\begin{aligned}
SE &= S_p * \sqrt{\frac{1}{n_{smokeN}} + \frac{1}{n_{smokeY}}} \\
&= \sqrt{515068.7} * \sqrt{\frac{1}{74} + \frac{1}{115}} \\
&= 106.9544.
\end{aligned}
$$

Observed Test Statistic

$$
\begin{aligned}
t_{obs} &= \frac{\text{Point estimate}}{SE} \\
&= \frac{282.66}{106.9544} \\
&= 2.642809.
\end{aligned}
$$

P-value

Since it is a two sided test,

```
2 * (1 - pt(2.642809, df = 74 + 115 - 2))
```

```
## [1] 0.008919322
```

We got a very small p-value. If we choose significance level to be 0.02, we would reject the null hypothesis.

Confirm findings in R

```
# t.test(sample1, sample2, conf.level = confidence_level, var.equal = TRUE) conf.level is
# for the confidence level. Foe the test, you will always get the same test statistic and
# p value.  var.equal tells R whether you would assume equal variance.
t.test(bwtSmokeN, bwtSmokeY, conf.level = 0.98, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  bwtSmokeN and bwtSmokeY
## t = 2.6428, df = 187, p-value = 0.00892
## alternative hypothesis: true difference in means is not equal to 0
## 98 percent confidence interval:
##   31.6956 533.6228
## sample estimates:
## mean of x mean of y
##  3054.957  2772.297
```

We got the same test statistic and p-value as our calculation. Besides, the 98% CI doesn't contain 0, which agrees with the results of hypothesis testing.

(e)

Use `t.test()` on this same data to perform a Welch's T test of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ at a 2% significance level. Also read off the 98% t CI for $\mu_{bwtN} - \mu_{bwtY}$ that results when you allow variances to differ. How do the results compare to (d)? (If you want more practice, you should try to confirm these calculations "by hand" from the summary values given.)

```
t.test(bwtSmokeN, bwtSmokeY, conf.level = 0.98, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  bwtSmokeN and bwtSmokeY
## t = 2.7192, df = 170.04, p-value = 0.007224
## alternative hypothesis: true difference in means is not equal to 0
## 98 percent confidence interval:
##   38.53243 526.78602
## sample estimates:
## mean of x mean of y
##  3054.957  2772.297
```

From the output, we can see that the test statistic is 2.7192, the degrees of freedom is 170.04, and the p-value is 0.007224.

The 98% CI is $(38.53243, 526.78602)$.

Compared to (d), the Welch's t test has a slightly more extreme test statistic, slightly smaller df, and a slightly smaller p-value. Besides, the Welch's CI is also slightly narrower.

(f)

Also perform a bootstrap t-tesst of $H_0 : \mu_{bwtN} - \mu_{bwtY} = 0$ vs $H_A : \mu_{bwtN} - \mu_{bwtY} \neq 0$ at a 2% significance level. How do the results compare to your findings in (d) and (e)?

```
# 1. Calculate data summaries and t_obs
xbar1 <- mean(bwtSmokeN)
s1 <- sd(bwtSmokeN)
n1 <- length(bwtSmokeN)
xbar2 <- mean(bwtSmokeY)
s2 <- sd(bwtSmokeY)
n2 <- length(bwtSmokeY)

t_obs <- (xbar1 - xbar2)/sqrt((s1^2/n1) + s2^2/n2)

# Specify number of bootstrap samples, create a vector to store t_hat values
B <- 5000
t_hat <- numeric(B)

# Bootstrap loop
set.seed(371)
```

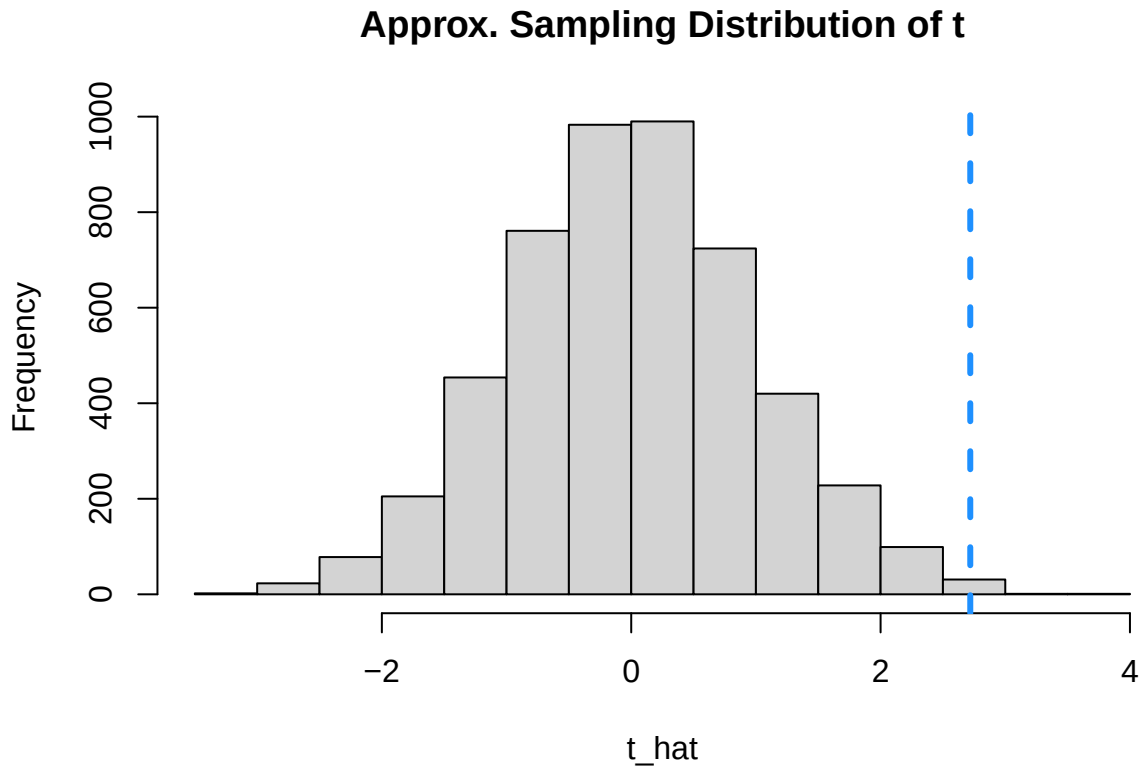
```r
for (i in 1:B) {
    # 2. Draw a SRS of size n1/n2 from data, with replacement
    x1_star <- sample(bwtSmokeN, size = n1, replace = T)
    x2_star <- sample(bwtSmokeY, size = n2, replace = T)

    # 3. Calculate resampled mean and sd
    xbar1_star <- mean(x1_star)
    s1_star <- sd(x1_star)
    xbar2_star <- mean(x2_star)
    s2_star <- sd(x2_star)

    # 4. Calculate t_hat, and store it in vector
    t_hat_numer <- (xbar1_star - xbar2_star) - (xbar1 - xbar2)
    t_hat_denom <- sqrt((s1_star^2/n1) + (s2_star^2/n2))

    t_hat[i] <- t_hat_numer/t_hat_denom
}

# Optional: plot approximate sampling distribution and add a line for t_obs
hist(t_hat, main = "Approx. Sampling Distribution of t")
abline(v = t_obs, col = "dodgerblue", lty = 2, lwd = 3)
```




```r
# Compute 2-sided p-value Count t_hat values above/below t_obs
m_low <- sum(t_hat < t_obs)
m_hi <- sum(t_hat > t_obs)
```

```r
# Take 2x the minimum proportion
2 * (min(m_low, m_hi)/B)
```

```
## [1] 0.0056
```

We get a bootstrap p-value of 0.0056, which is consistent with what we got for the previous two tests. Again we reject H_0. We see that the approximate sampling distribution of T looks very symmetric and is very similar to a t distribution, which makes sense given that our sample sizes are both large. It is probably not necessary to use the bootstrap given that the other two-sample t-tests are also available to us for this data.