t and z tests

Lecture 2 Handout

Statistics 139

Topics

- Hypothesis tests and Confidence Intervals
- Two sample t-tests: derivation, coding, interpretation, and assumptions
- Two sample proportion tests
- Sign test

The material in this lab corresponds to the Lecture 2 Notes.

Note: when performing a hypothesis test, be sure to explicitly state (1) hypotheses, (2) the calculated test statistic (and degrees of freedom if appropriate), (3) the calculated p-value or critical value, and (4) the conclusion in context of the problem along with the scope of inference. Use Type I error rates of $\alpha = 0.05$ and confidence levels of 95% unless explicitly stated otherwise. You can assume all tests are two-sided unless otherwise specified.

Since 1981, the United States Surgeon General has labeled cigarette packages with the warning: 'Smoking by pregnant women may result in fetal injury, premature birth, and low birth weight.' We will use a subset of the Child Health and Development Studies (CHDS) that examined association between smoking status of pregnant women and birthweight. The study was conducted between 1960 and 1967 by Kaiser Foundation Health Plan, Oakland, and was used as part of the evidence for the Congressional bill that led to the Surgeon General warnings.

The following question will be explored in this lab using the 'birthweight.csv' data set:

1. Is birthweight of babies associated with smoking status of the mother?
2. Does any possible relationship between birthweight and mother's smoking hold up after controlling for possible confounder(s)?

Data recorded here were for a random sample of 1236 babies in the study period: baby boys born during one year of the study, survived at least 28 days, and were single births. The variables measured were:

- `bwt`: birthweight of baby, in ounces.
- `gestation`: estimated time in womb based on due date, in days.

- `parity`: an indicator where 0 represents first full-term pregnancy for the mother, and 1 indicates the mother has had previous full-term pregnancies (1 or more).
- `age`: age of mother at birth, in years.
- `height`: height of mother, in inches.
- `weight`: weight of mother, in pounds.
- `smoke`: an indicator where 0 represents a non-smoking mother, 1 represents smoking mothers.

Concept Checks:

a) What is the rigorous interpretation of a 95% confidence interval (say for a population mean μ)? Why is it called a confidence interval and not a probability interval?

95% CI means that we're 95% confident that the interval we get includes the true population mean. It's not a probability interval because it doesn't measure the probability that the CI contains the true population mean (that probability is either 0 or 1, since we're talking about specific values and there is no randomness).

b) A 95% confidence interval (t-based) for the mean was calculated to be (3.0, 11.0) based on a sample size of $n = 61$ observations. Determine the t-test statistic for determining $H_0 : \mu = 0$ based on the same sample of data.

c) When data are paired (twin studies, for example), why should the pairing being taken into account? Justify mathematically. Hint: think about $\text{Var}(\bar{X}_1 - \bar{X}_2)$.

Question 1.

a) The study used 'baby boys born during one year of the study, survived at least 28 days, and were single births.' Present one pro and one con of making this decision.

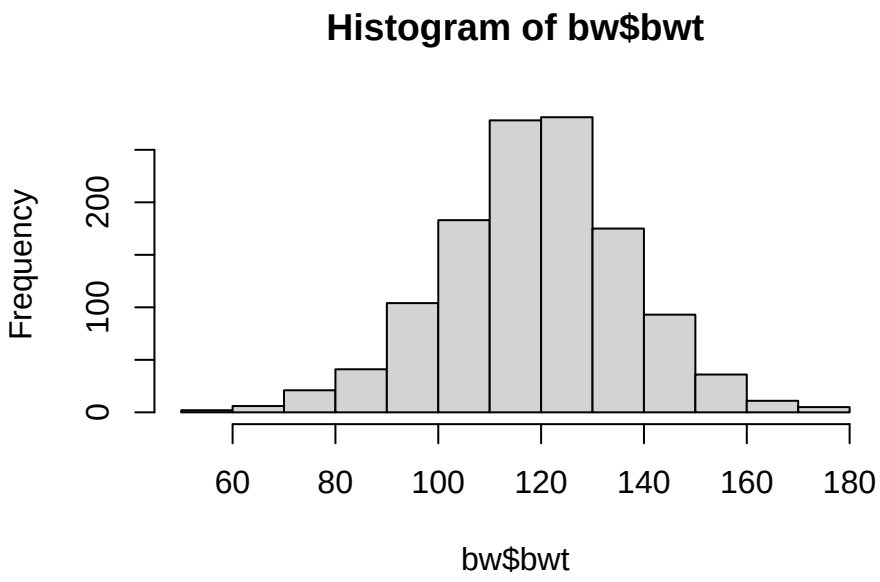
- Pro: standardized time frame (within one year), so able to control for confounding variables (eg. if there's a bad flu season in another year)
- Con: only studying baby boys; girls are not included in the cohort so can't generalize to all babies

b) Begin by reading in the data set and exploring. Be sure to look at (i) summary statistics, (ii) visuals for distributions, and (iii) visuals for relationships of variables with the outcome: birth weight.

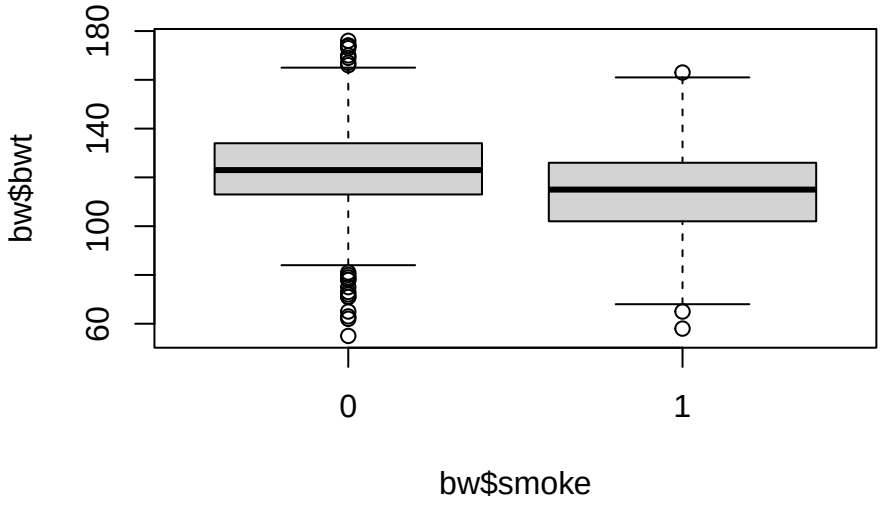
```
bw = read.csv("data/birthweight.csv")

# numerical summary
tapply(bw$bwt, bw$smoke, summary)
```

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      55     113     123     123     134     176
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    58.0   102.0   115.0   114.1   126.0   163.0
```

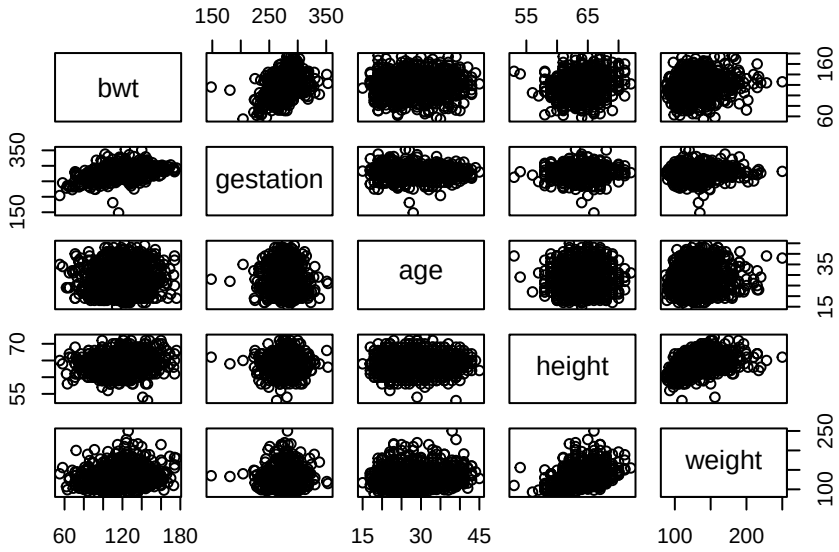
```
# graphical
# histogram of birth weight
hist(bw$bwt)
```




```
# boxplot: birth weight vs smoking status
boxplot(bw$bwt ~ bw$smoke)
```




```
# lots of scatterplots
pairs(bw[c("bwt","gestation","age","height","weight")])
```



c) Perform an appropriate hypothesis test to determine whether birth weight is associated with mother's smoking status.

Hypothesis: * Null: there is no difference between birth weight * Alternative: babies whose mothers don't smoke weight differently at birth

Results: * p-value is smaller than $\alpha = 0.05$. We reject the null hypothesis and conclude that there is substantial evidence that non-smoking mothers have heavier babies.

```
# 2-sample t test
results.2c <- t.test(bw$bwt ~ bw$smoke, alternative = "two.sided", mu = 0)
results.2c
```

```
##
##  Welch Two Sample t-test
##
## data:  bw$bwt by bw$smoke
## t = 8.5813, df = 1003.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   6.89385 10.98148
## sample estimates:
## mean in group 0 mean in group 1
##        123.0472        114.1095
```

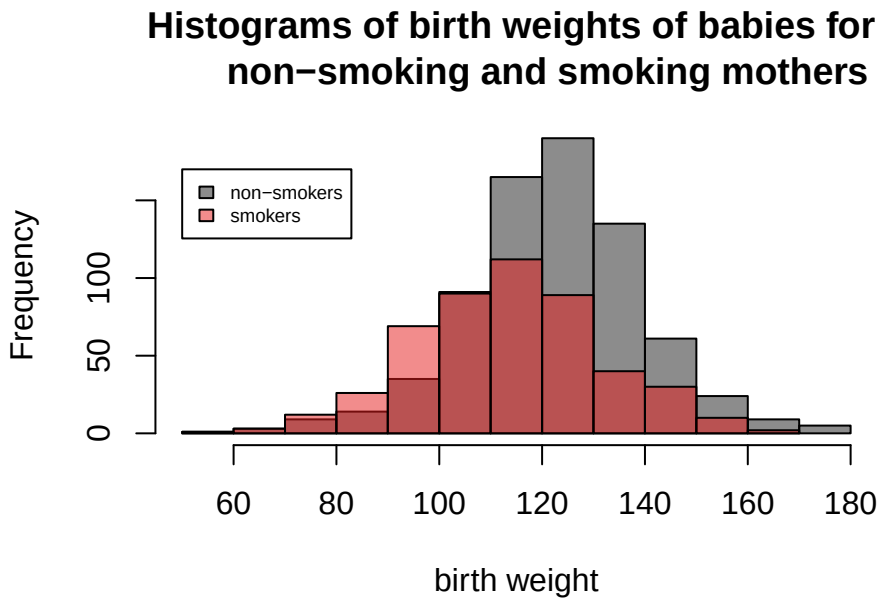
d) Provide a reasonable 95% confidence interval for estimating the 'effect' of smoking on a baby's birth weight. Compare this confidence interval to the hypothesis test results in the previous part. Is this truly an 'effect'?

- 95% CI: 6.8938504, 10.981481. Since 0 is not in the CI, this agrees with the results of the hypothesis test in the previous part. We can't really conclude that there is an effect yet because there are lots of possible confounding variables like mother's gestation.

e) Investigate and comment on the assumptions of your inferential approach in this problem.

Assumptions: * Independence within and between groups: single births ensure independence within group, and different mothers don't affect the weight of different babies * Normally distributed observations: true according to 1b and below

```
hist(bw$bwt[bw$smoke==0], main = "Histograms of birth weights of babies for
      non-smoking and smoking mothers",col=rgb(0.1,0.1,0.1,0.5),xlab="birth weight")
hist(bw$bwt[bw$smoke==1],col=rgb(0.9,0.1,0.1,0.5),add=T)
legend(x=50,y=170,legend=c("non-smokers","smokers"),cex=0.6,
       fill=c(rgb(0.1,0.1,0.1,0.5),rgb(0.9,0.1,0.1,0.5)))
```



f) What possible cofounders (measured and unmeasured) could be affecting these results? How could you incorporate any measured ones into the analysis?

Question 2. One approach to handle violations of the normality assumption in a t-test is to take a non-parametric approach. The simplest non-parametric approach is something called the sign test, which we will implement a simplified version of here.

a) Calculate the median of birth weights for mothers that do not have missing values for smoking status. Create a binary variable `low_bwt` that indicates whether a baby was below this median.

```
# median bwt
med <- median(bw$bwt[!is.na(bw$smoke)], na.rm = T)
med
```

```
## [1] 120
```

```
# create var
bw$low_bwt <- 1*(bw$bwt < med)
```

b) Let X_i be the measurement of `low_bwt` for a randomly sampled baby. What distribution does X_i have?

Bernoulli distribution, parameter = 0.489

c) Perform a hypothesis test to determine whether the proportion of `low_bwt` babies is different comparing smoking mothers to non-smoking mothers. Include the related confidence interval as well.

```
# by hand

# get phat
phat.smoke     <- mean(bw$low_bwt[bw$smoke==1], na.rm=T)
```

```r
phat.non.smoke <- mean(bw$low_bwt[bw$smoke==0], na.rm=T)
n.smoke     <- sum(!is.na(bw$low_bwt[bw$smoke==1]))
n.non.smoke <- sum(!is.na(bw$low_bwt[bw$smoke==0]))

# test statistic
phat <- (phat.smoke*n.smoke + phat.non.smoke*n.non.smoke)/(n.smoke + n.non.smoke)

# conduct prop test
print(z <- (phat.smoke-phat.non.smoke) / sqrt(phat*(1-phat)*(1/n.smoke+1/n.non.smoke)))
```

```
## [1] 7.612786
```

```r
print(pvalue <- 2*(1-pnorm(z)))
```

```
## [1] 2.68674e-14
```

```r
# using prop.test func
prop.test(table(bw$smoke, bw$low_bwt), correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  table(bw$smoke, bw$low_bwt)
## X-squared = 57.955, df = 1, p-value = 2.683e-14
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1666143 0.2780851
## sample estimates:
##    prop 1    prop 2
## 0.5983827 0.3760331
```

d) Let $Y = X_1 + ... + X_{n_1}$ be the total number of `low_bwt` babies in the smoking mother group in this sample. What distribution does Y follow?

Hypergeometric distribution

e) Comments on the assumptions of the test performed in the previous part.

Question 3: controlling for confounders.

a) Let's investigate the affect of confounders on the results from question 1. Create two new data frames in R: one called `younger.mothers` which includes only mothers aged 25 or younger, and one called `older.mothers` which includes mothers aged 26 or older.

b) Perform two separate t-tests to investigate the association of birth weight of babies with smoking status of mothers in these two age subgroups. Comment on what you see.

c) Perform similar subgroup analyses to account for possible confounding of (i) gestation age, (ii) parity, and (iii) weight.

d) Provide a 200 word summary of the results seen in this lab for the Surgeon General (who likely only took Stat 104). Provide a couple of visuals to support your conclusions.