t and z tests

Lecture 2 Handout

Statistics 139

Topics

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

The material in this lab corresponds to the Lecture 3 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 weeks, 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 confidence interval and not a probability interval?

The rigorous interpretation: If many, many samples of size n were taken and a confidence interval was built from each one, we would expect to see 95% of such potential confidence intervals cover the true mean μ (note, the potential bounds of a confidence interval are random variables before the sample is taken). The simple interpretation: It is the range of plausible values for the true unknown parameter, μ. It is called a confidence interval (reflecting the Bayesian perspective on probability) because μ is assumed to be a fixed, unknown constant in the Frequentist paradigm, and thus has no randomness and no probability. Once the interval is calculated, there is nothing random anymore. The sampling of data is what creates the 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.

Zero should certainly be excluded from the interval. The mean of this sample is the center of the interval, 7.0. The critical value for a t-distribution is essentially 2.00, and thus the standard error, $s/\sqrt(n)$, of this CI is also 2.0. And so the mean is 3.5 standard errors away form the null hypothesis value of 0, which is what the t-statistic represents.

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

If observations are paired, then $\text{Var}(\bar{X}_1 - \bar{X}_2)$ will tend to be smaller then $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ since there will be positive correlation between $\bar{X}_1$ and $\bar{X}_2$ (Remember: $\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) - 2\text{Cov}(\bar{X}_1, \bar{X}_2)$). This will lead to larger t-statistics if the true difference in population means is non-zero (better statistical power).

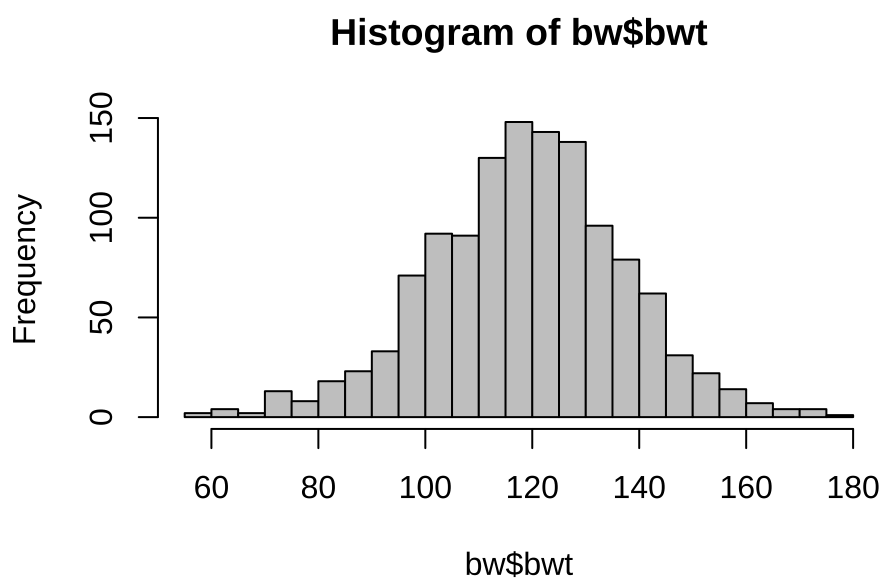
Question 1.

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

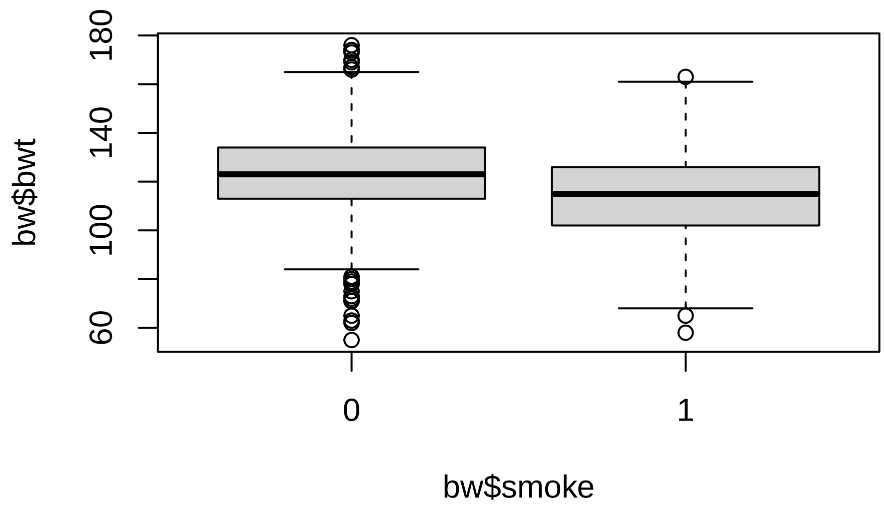
This decision is to have a more precise target population: a pro is that this approach controls for the possible effects of gender, premature birth, and multiple births (like twins) on birthweight (both confounding effects and simply just sources of variability). A con is that the result here may not generalize to other subpopulations excluded (though in this case, is likely to still hold for baby girls, etc.).

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.

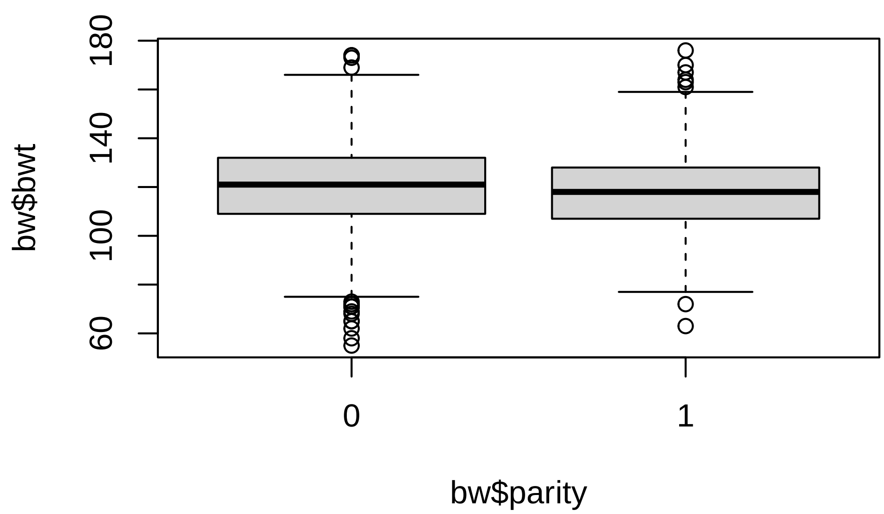
```
bw = read.csv("data/birthweight.csv")
hist(bw$bwt, breaks=20, col = "gray")
```



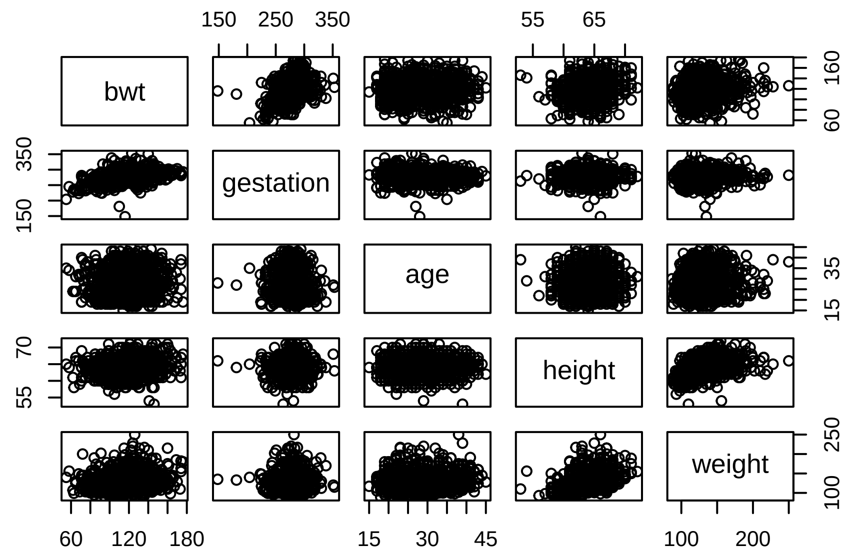

```
boxplot(bw$bwt ~ bw$smoke)
```




```
boxplot(bw$bwt ~ bw$parity)
```




```
pairs(bw[c("bwt","gestation","age","height","weight")])
```



There are a lot of possible ways to explore, a subset is shown here. The general distribution of the outcome, birth weight, is reasonably normally distributed, and its relationship with the other predictors is shown. There is evidence of a relationship of birthweight with gestation age, mother's height and weight, and smoking status, and almost no relationship with mother's age or parity.

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

```
print(overall <- t.test(bw$bwt ~ bw$smoke))
```

```
##
##  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
```

Based on the output of the `t.test` command above:

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_A: \mu_1 \neq \mu_2$$

$$t_{df \approx 1003} = 8.58$$

$$\text{p-value} < 0.00001$$

Since the 2-sided p-value is less than $\alpha = 0.05$, we can reject the null hypothesis. There is substantial evidence that non-smoking mothers have heavier babies than smoking mothers, on average.

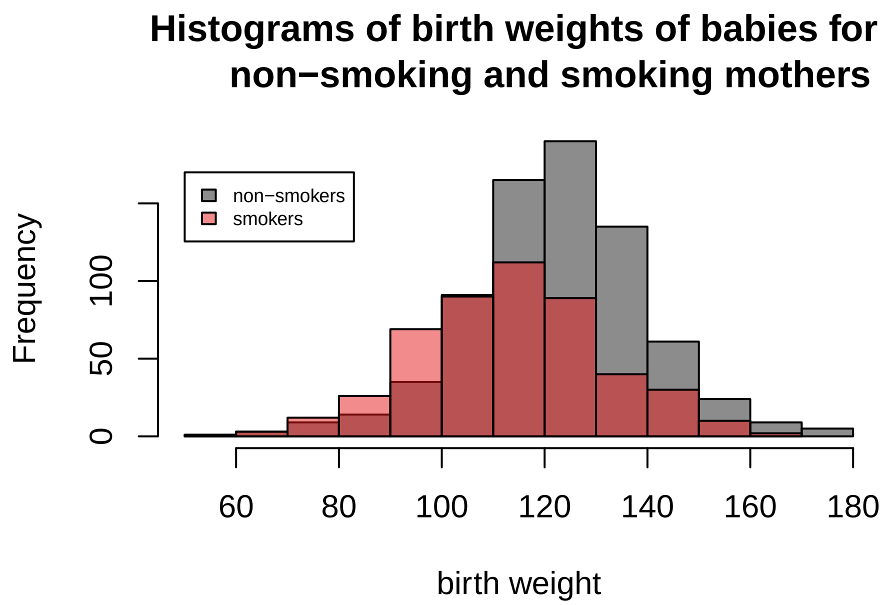
d) Provide a reasonable 95% confidence interval for estimating the 'effect' of smoking on a baby's birth weight. Compare this confidence interval to the hypothesis test results in the previous part. Is this truly an 'effect'?

The 95% confidence interval is calculated for us in R in the `t.test` command: (6.894, 10.982). This is a range of plausible values for the true difference in mean birth weights, in ounces, comparing non-smoking to smoking mothers. The CI agrees with the results of the hypothesis test since the value 0 is not inside of the CI, 0 is not plausible and thus 0 should be rejected as a null hypothesis value. I would hesitate to call this an 'effect' as there are lots of possible confounders involved in this result.

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

The 2 main assumptions are (1) independence between groups and within groups and (2) normally distributed observations in the population. Independence is reasonable since the data were gathered via two separate simple random samples, and the normality assumption is reasonable based on the general symmetry and 'bell-shape' of the distributions in the side-by-side boxplots in part (b). The histograms below also support this normality (though the sample sizes are so large, this may not matter all that much either). Note: the results generalize to a population similar to what is mentioned in part (a).

```
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 confounders (measured and unmeasured) could be affecting these results? How could you incorporate any measured ones into the analysis?

There is a nearly uncountable number of unmeasured confounders (diet, drug use, other unhealthy behaviors, etc.) and measured confounders are those that are at a minimum, associated with the response (birth weight), which includes gestation age and possibly height or parity. These could be incorporated into the analysis as other predictors in a regression model, as variables to match on, or as factors to do subgroups analysis on, as is done in problem 3 below.

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.

The median birth weight is 120 ounces (7.5 pounds), and roughly 49% of observations are below this value (there are a lot of 'ties' at 120 ounces).

```
print(m <- median(bw$bwt[!is.na(bw$smoke)]))
```

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

```r
bw$low_bwt = 1*(bw$bwt < m)

mean(bw$low_bwt[!is.na(bw$smoke)])
```

```
## [1] 0.4893964
```

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

X_i will be Bernoulli distributed with parameter roughly equal to 0.5 (or really 0.489 as estimated in the previous part).

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.

Based on the output below, the two sample proportion test leads to a test statistic of $z = 7.613$ and a p-value < 0.0001 (Note: R provides the square of this value in its `prop.test` command). In any case, there is clear evidence to reject the null hypothesis: smoking mothers truly have a higher proportion of lower birthweight babies than non-smoking mothers: 62.4% vs. 40.2% (be careful: the estimates given in `prop.test` are $1 - \hat{p}$.

```r
phat1 = mean(bw$low_bwt[bw$smoke==1],na.rm=T)
phat2 = mean(bw$low_bwt[bw$smoke==0],na.rm=T)
n1 = sum(!is.na(bw$low_bwt[bw$smoke==1]))
n2 = sum(!is.na(bw$low_bwt[bw$smoke==0]))

phat.pooled = (n1*phat1+n2*phat2)/(n1+n2)
print(z <- (phat1-phat2) / sqrt(phat.pooled*(1-phat.pooled)*(1/n1+1/n2)))
```

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

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

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

```r
 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?

This will NOT be based on a simple Bernoulli distribution as the observations are not independent (every time you sample an individual below the median, there is now a higher chance the next observation will be above the median). This sampling without replacement is then based on a hypergeometric distribution: $Y \sim HGeom(n_1 + n_2 = 1226, \hat{p}_1 n_1 + \hat{p}_2 n_2 = 600, n_1 = 484)$

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

Since Y is not based on a Binomial, then the Z test statistic will not be approximately standard normal (biggest issue: the standard error is incorrect).

Question 3: controlling for confounders.

a) Let's investigate the effect 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.

Code is shown below. There are 538 younger mothers and 700 older mothers (below 26 vs. 26 and older).

```
younger.mothers = bw[bw$age<=25,]
older.mothers = bw[bw$age>25,]
dim(younger.mothers)
```

```
## [1] 538   8
```

```
dim(older.mothers)
```

```
## [1] 700   8
```

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.

The relationship is similar in these two subgroups: non-smoking mothers have babies 8 to 10 ounces heavier than smoking mothers for both age groups. Note, the p-values are not as significant since we essentially cut the sample sizes in half.

```r
print(subgroup1 <- t.test(bwt~smoke,data=younger.mothers))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 6.6101, df = 476.07, p-value = 1.032e-10
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   6.785514 12.526219
## sample estimates:
## mean in group 0 mean in group 1
##        122.5806        112.9248
```

```r
print(subgroup2 <- t.test(bwt~smoke,data=older.mothers))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 5.5855, df = 516.33, p-value = 3.774e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   5.344024 11.142980
## sample estimates:
## mean in group 0 mean in group 1
##        123.3875        115.1440
```

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

In this approach we use the rough approximate of the median (ignoring missingness in other variables) to split into the subgroups. The association holds up in all groups, with varying degrees of magnitude: between 7.5 and 12.5 ounce differences (non-smoking mothers have heavier children, on average, in all subgroups considered).

```r
# for gestation age
m = median(bw$gestation, na.rm = T)
short.pregnancies = bw[bw$gestation<m,]
long.pregnancies = bw[bw$gestation>=m,]
print(subgroup3 <- t.test(bwt~smoke,data=short.pregnancies))
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  bwt by smoke
## t = 5.8107, df = 546.82, p-value = 1.057e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   5.573849 11.266880
## sample estimates:
## mean in group 0 mean in group 1
##        116.4709        108.0506
```

```
print(subgroup4 <- t.test(bwt~smoke,data=long.pregnancies))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 5.5952, df = 434.53, p-value = 3.906e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   4.860052 10.123270
## sample estimates:
## mean in group 0 mean in group 1
##        128.5320        121.0404
```

```
# for parity
m = median(bw$gestation, na.rm = T)
first.birth = bw[bw$parity==0,]
second.birth = bw[bw$parity==1,]
print(subgroup5 <- t.test(bwt~smoke,data=first.birth))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 7.601, df = 744.3, p-value = 8.858e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   6.937703 11.769260
## sample estimates:
## mean in group 0 mean in group 1
##        123.7226        114.3691
```

```
print(subgroup6 <- t.test(bwt~smoke,data=second.birth))
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  bwt by smoke
## t = 4.0387, df = 257.57, p-value = 7.098e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   4.001187 11.616006
## sample estimates:
## mean in group 0 mean in group 1
##        121.1392        113.3306
```

```
# for mother's weight
m = median(bw$weight, na.rm = T)
skinny.mothers = bw[bw$weight<m,]
other.mothers = bw[bw$weight>=m,]
print(subgroup7 <- t.test(bwt~smoke,data = skinny.mothers))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 6.4078, df = 434.81, p-value = 3.845e-10
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   6.697371 12.623637
## sample estimates:
## mean in group 0 mean in group 1
##        120.5764        110.9159
```

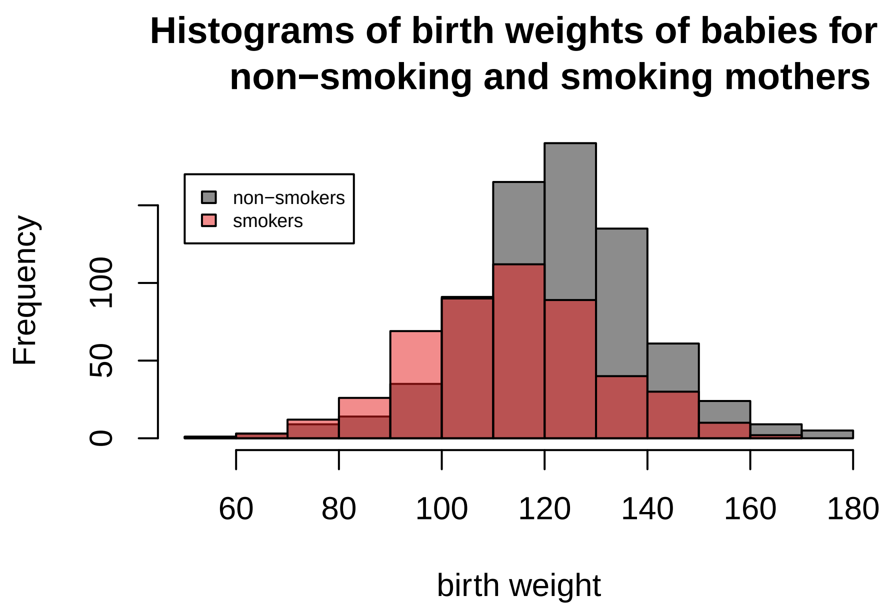
```
print(subgroup8 <- t.test(bwt~smoke,data=other.mothers))
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 5.4218, df = 516.45, p-value = 9.076e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##   5.08120 10.85605
## sample estimates:
## mean in group 0 mean in group 1
##        124.7762        116.8075
```

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 visuals to support your conclusions.

There is overwhelming evidence in the data that smoking is associated with lower birthweights, though a causal link can not be established. Overall, smoking mothers' babies weigh nearly 9 ounces less than non-smoking mothers (95% CI = (6.89, 10.98)). A few subgroups were investigated to see how this association could be modified: mother's age, gestation age, parity (previous births), and mother's weight. And in all analyses of subgroups, this relationship was significant and of similar magnitude (smoking mothers has babies estimated to weigh between 7.5 and 12.5 ounces less than for non-smoking mothers, depending on the subgroup being compared. These relationships are highlighted in the graphs 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)))
```




```
# pull off the CIs and combine into a single matrix
ci.matrix = rbind(overall$conf.int,subgroup1$conf.int,subgroup2$conf.int,
                  subgroup3$conf.int,subgroup4$conf.int,subgroup5$conf.int,
                  subgroup6$conf.int,subgroup7$conf.int,subgroup8$conf.int)
# calculate the midpoints, which are the mean differences
midpoints = apply(ci.matrix,1,mean)

# plot of confidence intervals: called a 'Forest Plot'
plot(x=midpoints,y=1:9,pch="x",xlim=c(0,max(ci.matrix)),ylim=c(0,10),cex=0.5,
col=c("blue",rep("black",8)),ylab="",xlab="smoking effect (in ounces)",yaxt="n",
main="Forest Plot of CIs within Subgroups Considered:\n Overall CI is in Blue")
points(x=ci.matrix[,1],y=1:9,pch="(",cex=0.5,col=c("blue",rep("black",8)))
points(x=ci.matrix[,2],y=1:9,pch=")",cex=0.5,col=c("blue",rep("black",8)))
```

```r
for(i in 1:9){
  lines(x=ci.matrix[i,],y=rep(i,2),col=c("blue",rep("black",8))[i])
}
abline(v=0,lwd=2,col="red")
groups = c("Overall","Mother's Age","Gestation Age","Parity","Mother's Weight")
text(groups,y=c(1,2.5,4.5,6.5,8.5),x=rep(3,5),cex=0.6)
```