Chapter 3

Two-Sample Inferences

Learning objectives

After completing this topic, you should be able to:

select graphical displays that meaningfully compare independent populations.

assess the assumptions of the two-sample t-test visually.

decide whether the means between two populations are different.

recommend action based on a hypothesis test.

Achieving these goals contributes to mastery in these course learning outcomes:

1. organize knowledge.

5. define parameters of interest and hypotheses in words and notation.

6. summarize data visually, numerically, and descriptively.

8. use statistical software.

12. make evidence-based decisions.

3.1 Comparing Two Sets of Measurements

Suppose you have collected data on one variable from two (independent) samples and you are interested in "comparing" the samples. What tools are good to use?

Example: Head Breadths In this analysis, we will compare a physical feature of modern day Englishmen with the corresponding feature of some of their ancient countrymen. The Celts were a vigorous race of people who once populated parts of England. It is not entirely clear whether they simply died out or merged with other people who were the ancestors of those who live in England today. A goal of this study might be to shed some light on possible genetic links between the two groups.

The study is based on the comparison of maximum head breadths (in millimeters) made on unearthed Celtic skulls and on a number of skulls of modern-day Englishmen. The data are given below. We have a sample of 18 Englishmen and an independent sample of 16 Celtic skulls.

```
#### Example: Head Breadths
# unstacked data as two vectors
english <- c(141, 148, 132, 138, 154, 142, 150, 146, 155, 158,
             150, 140, 147, 148, 144, 150, 149, 145)
celts   <- c(133, 138, 130, 138, 134, 127, 128, 138, 136, 131,
             126, 120, 124, 132, 132, 125)
english
##  [1] 141 148 132 138 154 142 150 146 155 158 150 140 147 148 144 150
## [17] 149 145
celts
##  [1] 133 138 130 138 134 127 128 138 136 131 126 120 124 132 132 125
```

What features of these data would we likely be interested in comparing? The centers of the distributions, the spreads within each distribution, the distributional shapes, etc.

These data can be analyzed in R as either **unstacked** separate vectors or as **stacked** data where one column contains both samples, with a second column of labels or **subscripts** to distinguish the samples. It is easy to create stacked data from unstacked data and vice-versa. Many comparisons require the plots for the two groups to have the same scale, which is easiest to control when the data are stacked.

```
# stacked data as a vector of values and a vector of labels
HeadBreadth <- c(english, celts)
Group <- c(rep("English", length(english)), rep("Celts", length(celts)))
hb <- data.frame(HeadBreadth, Group)
hb
```

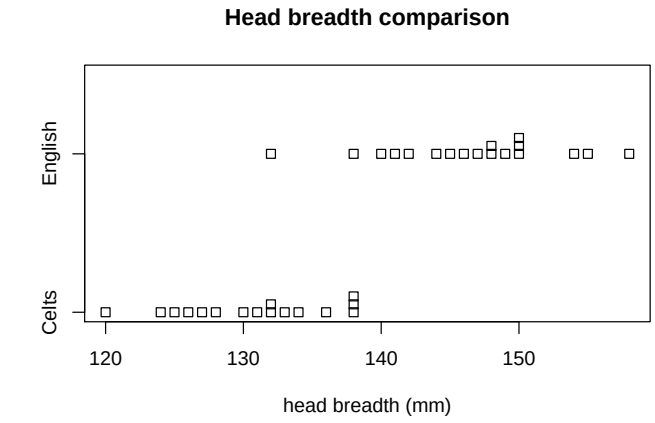
```
##    HeadBreadth   Group
## 1          141 English
## 2          148 English
## 3          132 English
## 4          138 English
## 5          154 English
## 6          142 English
## 7          150 English
## 8          146 English
## 9          155 English
## 10         158 English
## 11         150 English
## 12         140 English
## 13         147 English
## 14         148 English
## 15         144 English
## 16         150 English
## 17         149 English
## 18         145 English
## 19         133   Celts
## 20         138   Celts
## 21         130   Celts
## 22         138   Celts
## 23         134   Celts
## 24         127   Celts
## 25         128   Celts
## 26         138   Celts
## 27         136   Celts
## 28         131   Celts
## 29         126   Celts
## 30         120   Celts
## 31         124   Celts
## 32         132   Celts
## 33         132   Celts
## 34         125   Celts
```

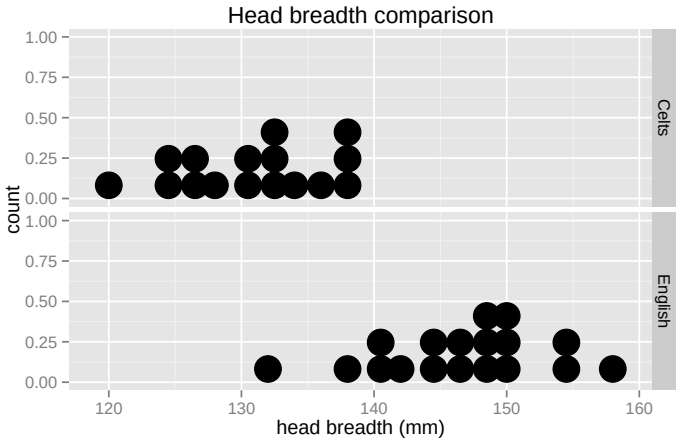
3.1.1 Plotting head breadth data:

1. A dotplot with the same scale for both samples is obtained easily from the stacked data.

```
#### Plotting head breadth data
# stripchart (dotplot) using R base graphics
stripchart(HeadBreadth ~ Group, method = "stack", data = hb,
  main = "Head breadth comparison", xlab = "head breadth (mm)")
```

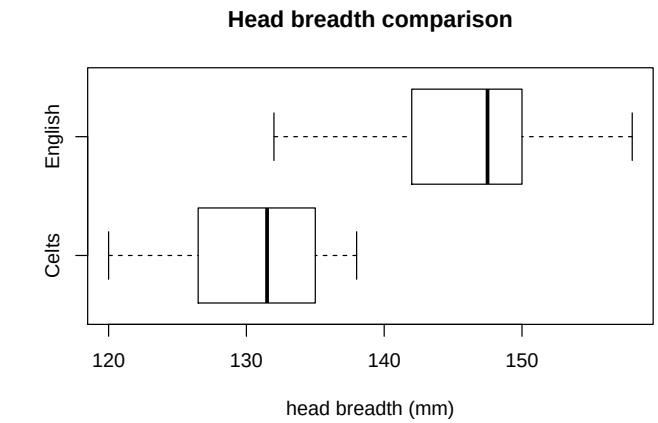
```
# stripchart (dotplot) using ggplot
library(ggplot2)
p <- ggplot(hb, aes(x = HeadBreadth))
p <- p + geom_dotplot(binwidth = 2)
p <- p + facet_grid(Group ~ .)      # rows are Group categories
p <- p + labs(title = "Head breadth comparison") + xlab("head breadth (mm)")
print(p)
```

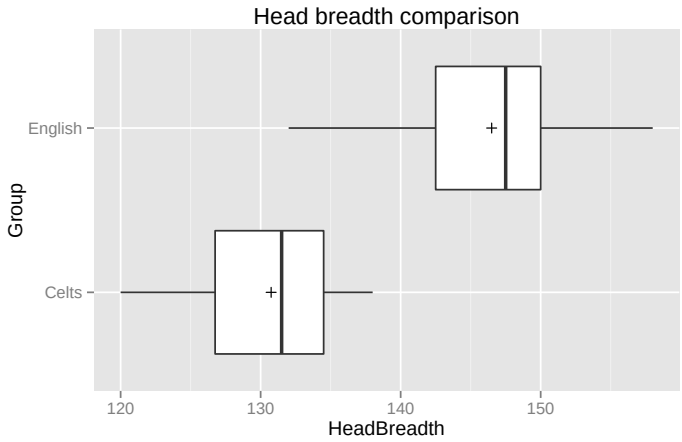




2. Boxplots for comparison are most helpful when plotted in the same axes.

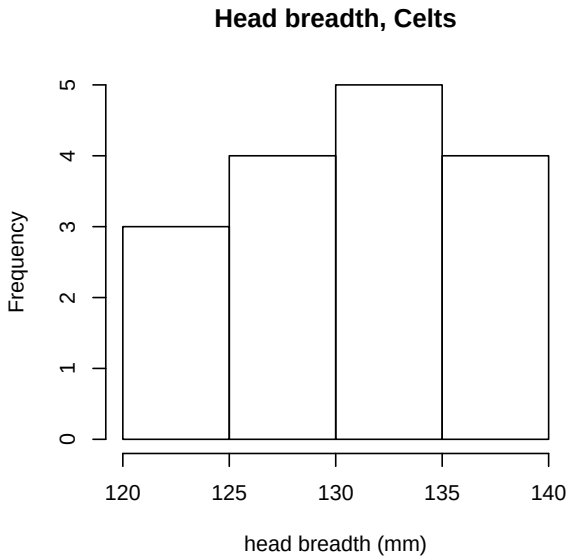
```
# boxplot using R base graphics
boxplot(HeadBreadth ~ Group, method = "stack", data = hb,
  horizontal = TRUE,
  main = "Head breadth comparison", xlab = "head breadth (mm)")


p <- ggplot(hb, aes(x = Group, y = HeadBreadth))
p <- p + geom_boxplot()
# add a "+" at the mean
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
p <- p + coord_flip()
p <- p + labs(title = "Head breadth comparison")
print(p)
```

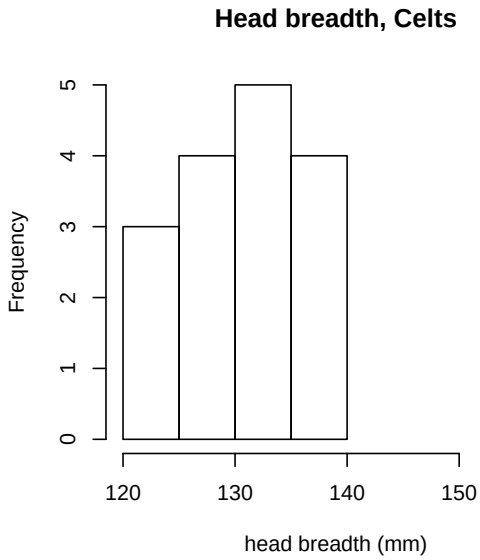


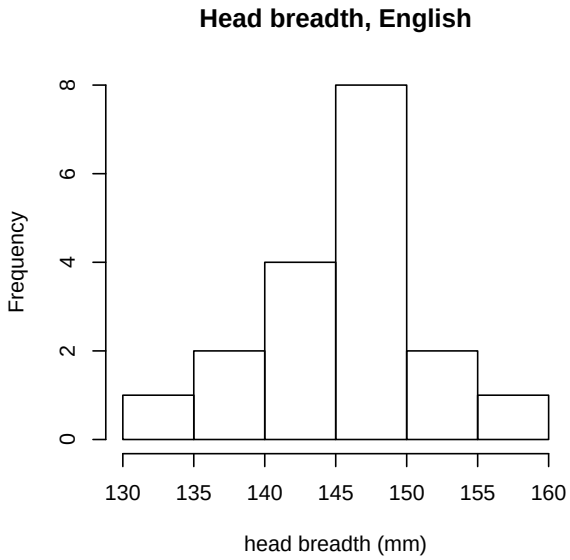


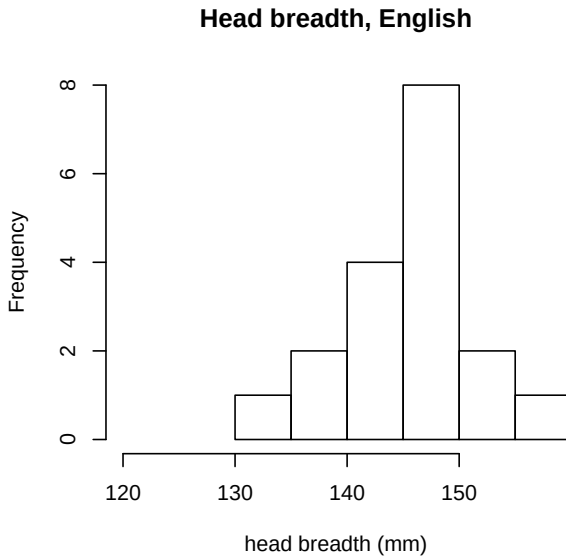
3. Histograms are hard to compare unless you make the scale and actual bins the same for both. Why is the pair on the right clearly preferable?

```r
# histogram using R base graphics
par(mfcol=c(2,2))
hist(hb$HeadBreadth[(hb$Group == "Celts")],
  main = "Head breadth, Celts", xlab = "head breadth (mm)")
hist(hb$HeadBreadth[(hb$Group == "English")],
  main = "Head breadth, English", xlab = "head breadth (mm)")

# common x-axis limits based on the range of the entire data set
hist(hb$HeadBreadth[(hb$Group == "Celts")], xlim = range(hb$HeadBreadth),
  main = "Head breadth, Celts", xlab = "head breadth (mm)")
hist(hb$HeadBreadth[(hb$Group == "English")], xlim = range(hb$HeadBreadth),
  main = "Head breadth, English", xlab = "head breadth (mm)")
```

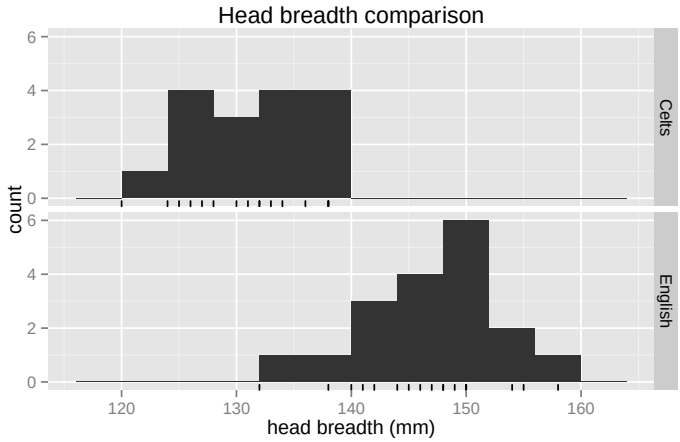




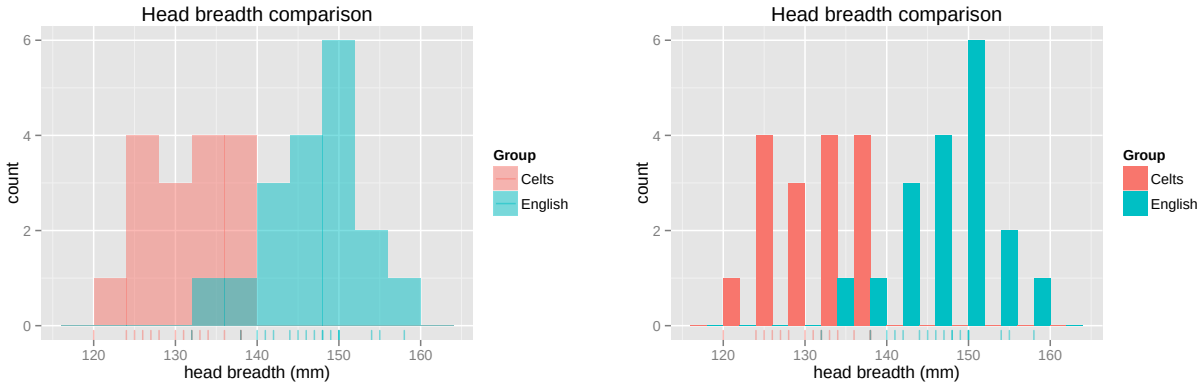





```
# histogram using ggplot
p <- ggplot(hb, aes(x = HeadBreadth))
p <- p + geom_histogram(binwidth = 4)
p <- p + geom_rug()
p <- p + facet_grid(Group ~ .)
p <- p + labs(title = "Head breadth comparison") + xlab("head breadth (mm)")
print(p)
```




```
p <- ggplot(hb, aes(x = HeadBreadth, fill=Group))
p <- p + geom_histogram(binwidth = 4, alpha = 0.5, position="identity")
p <- p + geom_rug(aes(colour=Group), alpha = 1/2)
p <- p + labs(title = "Head breadth comparison") + xlab("head breadth (mm)")
print(p)

p <- ggplot(hb, aes(x = HeadBreadth, fill=Group))
p <- p + geom_histogram(binwidth = 4, alpha = 1, position="dodge")
p <- p + geom_rug(aes(colour=Group), alpha = 1/2)
p <- p + labs(title = "Head breadth comparison") + xlab("head breadth (mm)")
print(p)
```



4. Stem-and-leaf displays for comparisons in R can be pretty useless. The stems are not forced to match (just like with histograms). It is pretty hard to make quick comparisons with the following:

```
stem(english, scale = 2)

##
##   The decimal point is 1 digit(s) to the right of the |
##
##   13 | 2
##   13 | 8
##   14 | 0124
##   14 | 567889
##   15 | 0004
```

```
##   15 | 58
```

```
stem(celts, scale = 2)
```

```
##
##   The decimal point is at the |
##
##   120 | 0
##   122 |
##   124 | 00
##   126 | 00
##   128 | 0
##   130 | 00
##   132 | 000
##   134 | 0
##   136 | 0
##   138 | 000
```

Using the `by()` function, you can get summaries by group.

```
#### summaries by group
# summary for separate vectors
summary(english)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   132.0   142.5   147.5   146.5   150.0   158.0
```

```
summary(celts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   120.0   126.8   131.5   130.8   134.5   138.0
```

```
# comparing spreads, an assumption of equal variances seems reasonable
sd(english)
```

```
## [1] 6.382421
```

```
sd(celts)
```

```
## [1] 5.434458
```

```
IQR(english)
```

```
## [1] 7.5
```

```
IQR(celts)
```

```
## [1] 7.75
```

```
# numerical summary of each column in data.frame hb by Group
by(hb, Group, summary)
```

```
## Group: Celts
##   HeadBreadth        Group
##  Min.   :120.0   Celts  :16
##  1st Qu.:126.8   English: 0
##  Median :131.5
##  Mean   :130.8
```

```
##  3rd Qu.:134.5
##  Max.   :138.0
## ------------------------------------------------
## Group: English
##   HeadBreadth        Group
##  Min.   :132.0   Celts  : 0
##  1st Qu.:142.5   English:18
##  Median :147.5
##  Mean   :146.5
##  3rd Qu.:150.0
##  Max.   :158.0
```

3.1.2 Salient Features to Notice

The dotplots, boxplots, and histograms indicate that the English and Celt samples are slightly skewed to the left. There are no outliers in either sample. It is not unreasonable to operationally assume that the population frequency curves (i.e., the histograms for the populations from which the samples were selected) for the English and Celtic head breadths are normal. Therefore, the sampling distribution of the means will be reasonably normal.

The sample means and medians are close to each other in each sample, which is not surprising given the near symmetry and the lack of outliers.

The data suggest that the typical modern English head breadth is greater than that for Celts. The data sets have comparable spreads, as measured by either the standard deviation or the IQR.

3.2 Two-Sample Methods: Paired Versus Independent Samples

Suppose you have two populations of interest, say populations 1 and 2, and you are interested in comparing their (unknown) population means, μ_1 and μ_2. Inferences on the unknown population means are based on samples from each population. In practice, most problems fall into one of two categories.

Independent samples where the sample taken from population 1 has no

effect on which observations are selected from population 2, and vice versa.

Paired or dependent samples where experimental units are paired based on factors related or unrelated to the variable measured.

The distinction between paired and independent samples may be made clear through a series of examples.

Example The English and Celt head breadth samples are independent.

Example Suppose you are interested in whether the $CaCO_3$ (calcium carbonate) level in the Atrisco well field, which is the water source for Albuquerque, is changing over time. To answer this question, the $CaCO_3$ level was recorded at each of 15 wells at two time points. These data are paired. The two samples are the observations at Times 1 and 2.

Example To compare state incomes, a random sample of New Mexico households was selected, and an independent sample of Arizona households was obtained. It is reasonable to assume independent samples.

Example Suppose you are interested in whether the husband or wife is typically the heavier smoker among couples where both adults smoke. Data are collected on households. You measure the average number of cigarettes smoked by each husband and wife within the sample of households. These data are paired, i.e., you have selected husband wife pairs as the basis for the samples. It is reasonable to believe that the responses within a pair are related, or correlated.

Although the focus here will be on comparing population means, you should recognize that in paired samples you may also be interested, as in the problems above, in how observations compare within a pair. That is, a **paired comparison** might be interested in the **difference** between the two paired samples. These goals need not agree, depending on the questions of interest. Note that with paired data, the sample sizes are equal, and equal to the number of pairs.

■ CLICKERQs — Independent or paired 1, STT.08.02.030 ■

■ CLICKERQs — Independent or paired 2, STT.08.02.040 ■

3.3 Two Independent Samples: CI and Test Using Pooled Variance

These methods assume that the populations have normal frequency curves, with equal population standard deviations, i.e., $\sigma_1 = \sigma_2$. Let $(n_1,\ \bar{Y}_1,\ s_1)$ and $(n_2,\ \bar{Y}_2,\ s_2)$ be the sample sizes, means and standard deviations from the two samples. The standard CI for $\mu_1 - \mu_2$ is given by

$$\text{CI} \ = \ (\bar{Y}_1 - \bar{Y}_2) \pm t_{\text{crit}} SE_{\bar{Y}_1 - \bar{Y}_2}$$

The t-statistic for testing $H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$ against $H_A : \mu_1 - \mu_2 \neq 0$ $(\mu_1 \neq \mu_2)$ is given by

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}.$$

The standard error of $\bar{Y}_1 - \bar{Y}_2$ used in both the CI and the test is given by

$$SE_{\bar{Y}_1 - \bar{Y}_2} = s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Here the **pooled variance estimator**,

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

is our best estimate of the common population variance. The pooled estimator of variance is a weighted average of the two sample variances, with more weight given to the larger sample. If $n_1 = n_2$ then s^2_{pooled} is the average of s_1^2 and s_2^2.

The critical value t_{crit} for CI and tests is obtained in usual way from a t-table with $df = n_1 + n_2 - 2$. For the test, follow the one-sample procedure, with the new t_s and t_{crit}.

The pooled CI and tests are sensitive to the normality and equal standard deviation assumptions. The observed data can be used to assess the reasonableness of these assumptions. You should look at boxplots and histograms to assess normality, and should check whether $s_1 \doteq s_2$ to assess the assumption $\sigma_1 = \sigma_2$. Formal tests of these assumptions will be discussed later.

3.4 Satterthwaite's Method, unequal variances

Satterthwaite's method assumes normality, but does not require equal population standard deviations. Satterthwaite's procedures are somewhat conservative, and adjust the SE and df to account for unequal population variances. Satterthwaite's method uses the same CI and test statistic formula, with a modified standard error:

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Note that $df = n_1 + n_2 - 2$ when $n_1 = n_2$ and $s_1 = s_2$. The Satterthwaite and pooled variance procedures usually give similar results when $s_1 \doteq s_2$.

The df formula for Satterthwaite's method is fairly complex, so when done by hand some use a conservative df formula that uses the minimum of $n_1 - 1$ and $n_2 - 1$ instead.

3.4.1 R Implementation

R does the pooled and Satterthwaite (Welch) analyses, either on stacked or unstacked data. The output will contain a p-value for a two-sided test of equal population means and a CI for the difference in population means. If you include `var.equal = TRUE` you will get the pooled method, otherwise the output is for Satterthwaite's method.

Example: Head Breadths The English and Celts are independent samples. We looked at boxplots and histograms, which suggested that the normality assumption for the t-test is reasonable. The R output shows the English and Celt sample standard deviations and IQRs are fairly close, so the pooled and Satterthwaite results should be comparable. The pooled analysis is preferable here, but either is appropriate.

We are interested in difference in mean head breadths between Celts and English.

1. Define the population parameters and hypotheses in words and notation

Let μ_1 and μ_2 be the mean head breadth for the Celts and English, respectively.

In words: "The difference in population means between Celts and English is different from zero mm."

In notation: $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$.

Alternatively: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$.

2. Calculate summary statistics from sample

Mean, standard deviation, sample size:

```
#### Calculate summary statistics
m1 <- mean(celts)
s1 <- sd(celts)
n1 <- length(celts)
m2 <- mean(english)
s2 <- sd(english)
n2 <- length(english)
c(m1, s1, n1)

## [1] 130.750000   5.434458  16.000000
```

```
c(m2, s2, n2)
```

```
## [1] 146.500000   6.382421  18.000000
```

The pooled-standard devation, standard error, and degrees-of-freedom are:

```
sdpool <- sqrt(((n1 - 1) * s1^2 + (n2 - 1) * s2^2) / (n1 + n2 - 2))
sdpool
```

```
## [1] 5.956876
```

```
SEpool <- sdpool * sqrt(1 / n1 + 1 / n2)
SEpool
```

```
## [1] 2.046736
```

```
dfpool <- n1 + n2 - 2
dfpool
```

```
## [1] 32
```

```
t_pool <- (m1 - m2) / SEpool
t_pool
```

```
## [1] -7.69518
```

The Satterthwaite SE and degrees-of-freedom are:

```
SE_Sat <- sqrt(s1^2 / n1 + s2^2 / n2)
SE_Sat
```

```
## [1] 2.027043
```

```
df_Sat <- (SE_Sat^2)^2 / (s1^4 / (n1^2 * (n1 - 1)) + s2^4 / (n2^2 * (n2 - 1)))
df_Sat
```

```
## [1] 31.9511
```

```
t_Sat <- (m1 - m2) / SE_Sat
t_Sat
```

```
## [1] -7.769937
```

3. Specify confidence level, calculate t-stat, CI limits, p-value

Let us calculate a 95% CI for $\mu_1 - \mu_2$.
Assuming equal variances, using pooled-variance procedure:

```
## Equal variances
#   var.equal = FALSE is the default

# two-sample t-test specifying two separate vectors
t.summary.eqvar <- t.test(celts, english, var.equal = TRUE)
t.summary.eqvar
```

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

```
##
## data:  celts and english
## t = -7.6952, df = 32, p-value = 9.003e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.91906 -11.58094
## sample estimates:
## mean of x mean of y
##    130.75    146.50
```

Not assuming equal variances, Satterthwaite (Welch):

```
# two-sample t-test with unequal variances (Welch = Satterthwaite)
#   specified using data.frame and a formula, HeadBreadth by Group
t.summary.uneqvar <- t.test(HeadBreadth ~ Group, data = hb, var.equal = FALSE)
t.summary.uneqvar
```

```
##
##  Welch Two Sample t-test
##
## data:  HeadBreadth by Group
## t = -7.7699, df = 31.951, p-value = 7.414e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.8792 -11.6208
## sample estimates:
##   mean in group Celts mean in group English
##                130.75                146.50
```

The form of the output will tell you which sample corresponds to population 1 and which corresponds to population 2.

4. **Summarize in words** (Using the pooled-variance results.)

The pooled analysis strongly suggests that $H_0 : \mu_1 - \mu_2 = 0$ is false, given the large t-statistic of -7.7 and two-sided p-value of 9×10^{-9}. Because the p-value < 0.05 we reject the Null hypothesis in favor of the Alternative hypothesis concluding that the difference in population mean head breadths between the Celts and English are different.

We are 95% confident that the difference in population means, $\mu_1 - \mu_2$, is between -19.9 and -11.6 mm. That is, we are 95% confident that the population mean head breadth for Englishmen (μ_2) exceeds the population mean head breadth for Celts (μ_1) by between 11.6 and 19.9 mm.

The CI interpretation is made easier by recognizing that we concluded the population means are different, so the direction of difference must be con-

sistent with that seen in the observed data, where the sample mean head breadth for Englishmen exceeds that for the Celts. Thus, the limits on the CI for $\mu_1 - \mu_2$ tells us how much smaller the mean is for the Celts (that is, between -19.9 and -11.6 mm).

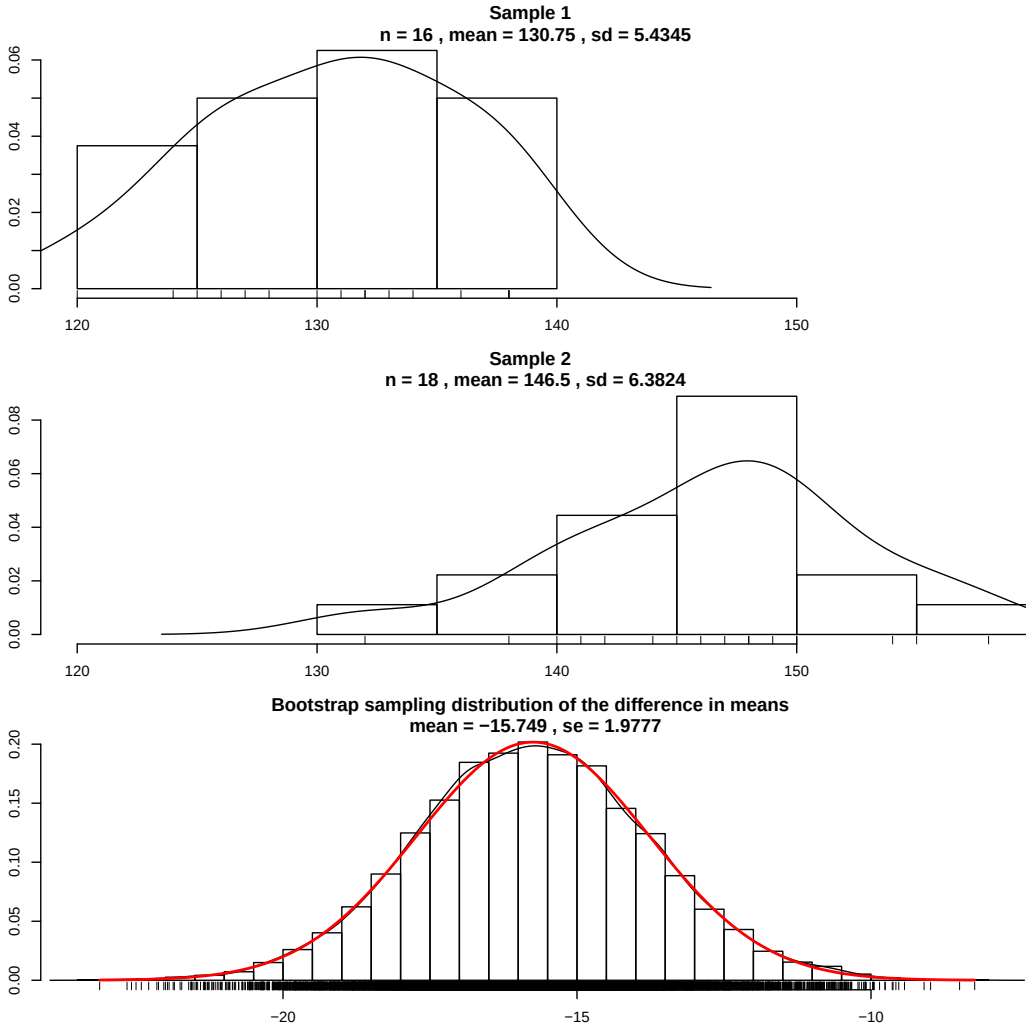
5. Check assumptions

The assumption of equal population variances will be left to a later chapter. We can test the assumption that the distribution of $\bar{Y}_1 - \bar{Y}_2$ is normal using the bootstrap in the following function.

```
#### Visual comparison of whether sampling distribution is close to Normal via Bootstrap
# a function to compare the bootstrap sampling distribution
#   of the difference of means from two samples with
#   a normal distribution with mean and SEM estimated from the data
bs.two.samp.diff.dist <- function(dat1, dat2, N = 1e4) {
  n1 <- length(dat1);
  n2 <- length(dat2);
  # resample from data
  sam1 <- matrix(sample(dat1, size = N * n1, replace = TRUE), ncol=N);
  sam2 <- matrix(sample(dat2, size = N * n2, replace = TRUE), ncol=N);
  # calculate the means and take difference between populations
  sam1.mean <- colMeans(sam1);
  sam2.mean <- colMeans(sam2);
  diff.mean <- sam1.mean - sam2.mean;
  # save par() settings
  old.par <- par(no.readonly = TRUE)
  # make smaller margins
  par(mfrow=c(3,1), mar=c(3,2,2,1), oma=c(1,1,1,1))
  # Histogram overlaid with kernel density curve
  hist(dat1, freq = FALSE, breaks = 6
      , main = paste("Sample 1", "\n"
                     , "n =", n1
                     , ", mean =", signif(mean(dat1), digits = 5)
                     , ", sd =", signif(sd(dat1), digits = 5))
      , xlim = range(c(dat1, dat2)))
  points(density(dat1), type = "l")
  rug(dat1)

  hist(dat2, freq = FALSE, breaks = 6
      , main = paste("Sample 2", "\n"
                     , "n =", n2
                     , ", mean =", signif(mean(dat2), digits = 5)
                     , ", sd =", signif(sd(dat2), digits = 5))
      , xlim = range(c(dat1, dat2)))
  points(density(dat2), type = "l")
  rug(dat2)

  hist(diff.mean, freq = FALSE, breaks = 25
      , main = paste("Bootstrap sampling distribution of the difference in means", "\n"
                   , "mean =", signif(mean(diff.mean), digits = 5)
                   , ", se =", signif(sd(diff.mean), digits = 5)))
  # overlay a density curve for the sample means
  points(density(diff.mean), type = "l")
  # overlay a normal distribution, bold and red
  x <- seq(min(diff.mean), max(diff.mean), length = 1000)
  points(x, dnorm(x, mean = mean(diff.mean), sd = sd(diff.mean))
       , type = "l", lwd = 2, col = "red")
  # place a rug of points under the plot
  rug(diff.mean)
  # restore par() settings
  par(old.par)
}
```

The distribution of difference in means in the third plot looks very close to normal.

```
bs.two.samp.diff.dist(celts, english)
```



■ CLICKER Qs — t-interval, STT.08.02.010 ■

Example: Androstenedione levels in diabetics The data consist of independent samples of diabetic men and women. For each individual, the scientist recorded their androstenedione level (ng/dL) (a hormone, and Mark McGwire's favorite dietary supplement). Let μ_1 = mean androstenedione level for the population of diabetic men, and μ_2 = mean androstenedione level for the population of diabetic women. We are interested in comparing the population means given the observed data.

The raw data and R output are given below. The boxplots suggest that the distributions are reasonably symmetric. However, the normality assumption for the women is unreasonable due to the presence of outliers. The equal population standard deviation assumption also appears unreasonable. The sample standard deviation for men is noticeably larger than the women's standard deviation, even with outliers in the women's sample.

```
#### Example: Androstenedione levels in diabetics
# Data and numerical summaries
men   <- c(217, 123,  80, 140, 115, 135,  59, 126,  70,  63,
           147, 122, 108,  70)
women <- c( 84,  87,  77,  84,  73,  66,  70,  35,  77,  73,
            56, 112,  56,  84,  80, 101,  66,  84)
level <- c(men, women)
sex <- c(rep("men", length(men)), rep("women", length(women)))
andro <- data.frame(level, sex)
andro
```

```
##    level   sex
## 1    217   men
## 2    123   men
## 3     80   men
## 4    140   men
## 5    115   men
## 6    135   men
## 7     59   men
## 8    126   men
## 9     70   men
## 10    63   men
## 11   147   men
## 12   122   men
## 13   108   men
## 14    70   men
## 15    84 women
## 16    87 women
## 17    77 women
## 18    84 women
## 19    73 women
## 20    66 women
## 21    70 women
## 22    35 women
## 23    77 women
## 24    73 women
## 25    56 women
## 26   112 women
## 27    56 women
## 28    84 women
```

```
## 29    80 women
## 30   101 women
## 31    66 women
## 32    84 women
```

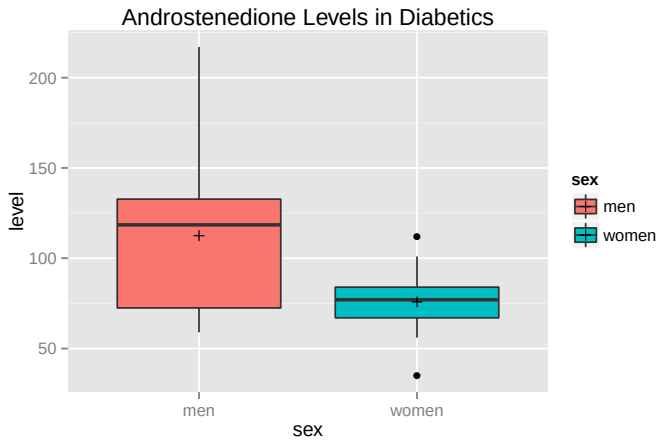
```
# numerical summaries
by(andro, sex, summary)
```

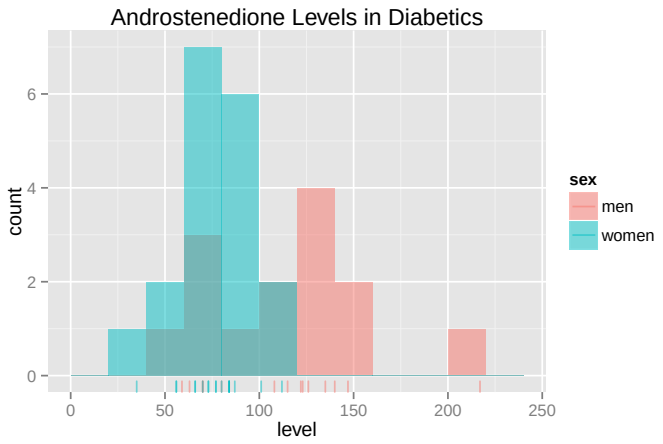
```
## sex: men
##      level          sex
##  Min.   : 59.0   men  :14
##  1st Qu.: 72.5   women: 0
##  Median :118.5
##  Mean   :112.5
##  3rd Qu.:132.8
##  Max.   :217.0
## ------------------------------------------------------
## sex: women
##      level           sex
##  Min.   : 35.00   men  : 0
##  1st Qu.: 67.00   women:18
##  Median : 77.00
##  Mean   : 75.83
##  3rd Qu.: 84.00
##  Max.   :112.00
```

```
c(sd(men), sd(women), IQR(men), IQR(women), length(men), length(women))
```

```
## [1] 42.75467 17.23625 60.25000 17.00000 14.00000 18.00000
```

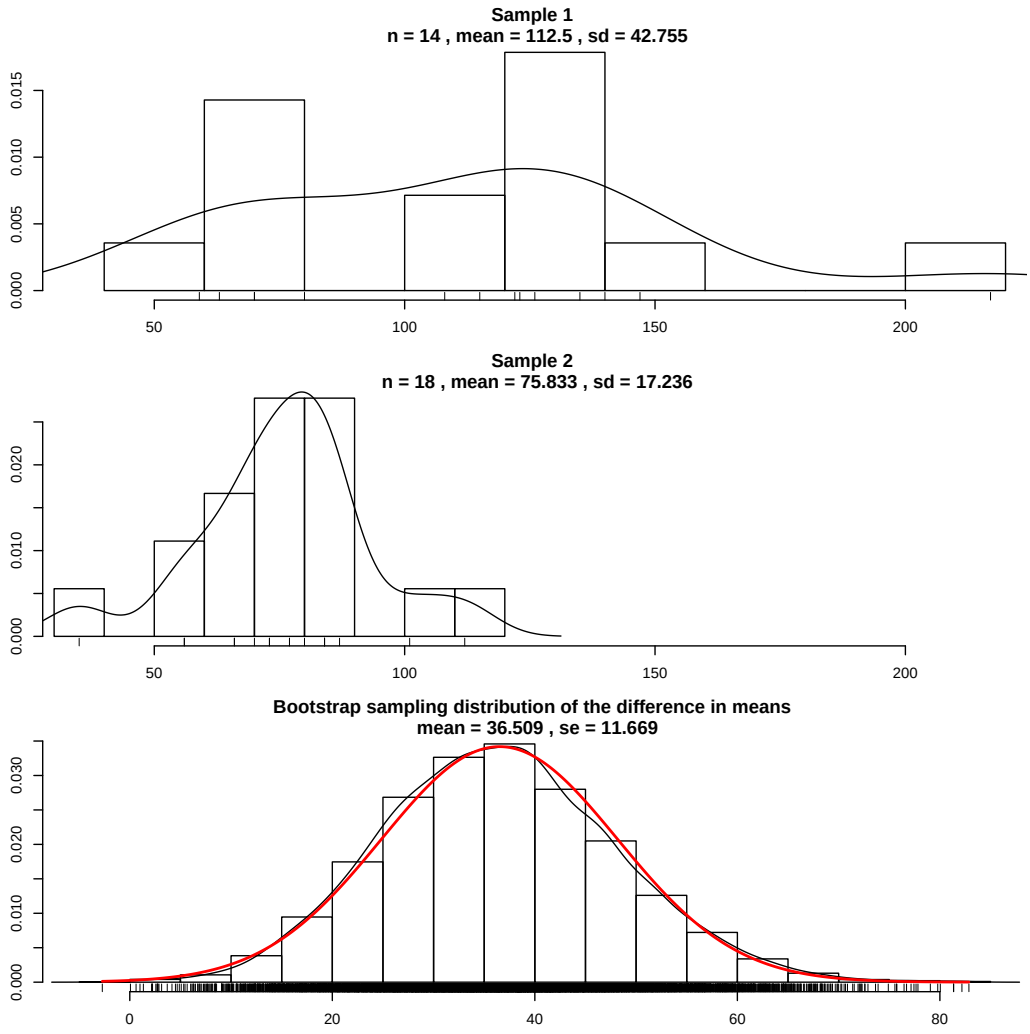
```
p <- ggplot(andro, aes(x = sex, y = level, fill=sex))
p <- p + geom_boxplot()
# add a "+" at the mean
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
#p <- p + coord_flip()
p <- p + labs(title = "Androstenedione Levels in Diabetics")
print(p)

p <- ggplot(andro, aes(x = level, fill=sex))
p <- p + geom_histogram(binwidth = 20, alpha = 0.5, position="identity")
p <- p + geom_rug(aes(colour=sex), alpha = 1/2)
p <- p + labs(title = "Androstenedione Levels in Diabetics")
print(p)
```





Because of the large difference in variances, I will be more comfortable with the Satterthwaite analysis here than the pooled variance analysis. The normality assumption of the difference in means appears to be met using the bootstrap assessment. The distribution of difference in means in the third plot looks very close to normal.

```
bs.two.samp.diff.dist(men, women)
```



1. Define the population parameters and hypotheses in words and notation

Let μ_1 and μ_2 be the mean androstenedione level for diabetic men and women, respectively.

In words: "The difference in population mean androstenedione levels between diabetic men and women is different from zero."

In notation: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$.

2. Calculate summary statistics from sample

(see above)

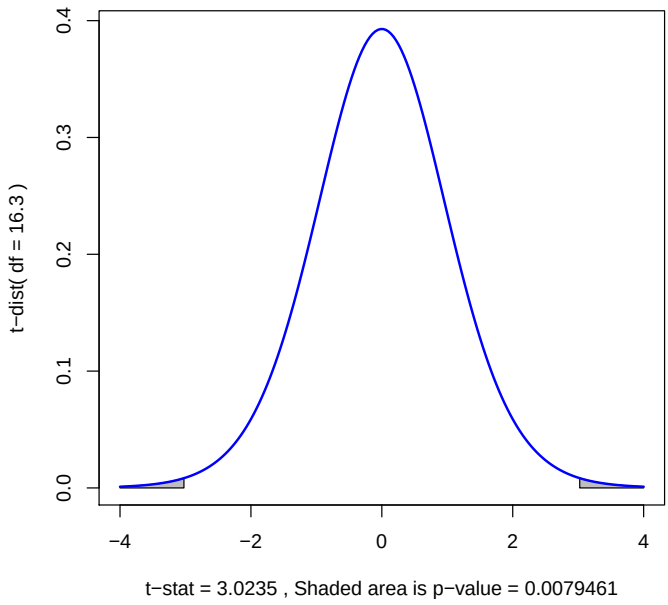
3. Specify confidence level, calculate t-stat, CI limits, p-value

Not assuming equal variances, Satterthwaite (Welch):

```
# two-sample t-test with unequal variances (Welch = Satterthwaite)
#   specified using data.frame and a formula, level by sex
t.summary <- t.test(level ~ sex, data = andro, var.equal = FALSE)
t.summary
```

```
##
##  Welch Two Sample t-test
##
## data:  level by sex
## t = 3.0235, df = 16.295, p-value = 0.007946
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.99555 62.33778
## sample estimates:
##   mean in group men mean in group women
##           112.50000            75.83333
```

```
# plot t-distribution with shaded p-value
t.dist.pval(t.summary)
```



4. Summarize in words The unequal-variance analysis suggests that $H_0 : \mu_1 - \mu_2 = 0$ is false, given the large t-statistic of 3.02 and two-sided p-value of 0.00795. Because the p-value < 0.05 we reject the Null hypothesis in favor of the Alternative hypothesis concluding that the difference in population mean androstenedione levels between diabetic men and women are different.

We are 95% confident confident that the difference in population means, $\mu_1 - \mu_2$, is between 11 and 62.3 ng/dL.

5. Check assumptions

As checked before, while the assumption of equal population variances is not met, the assumption that the distribution of $\bar{Y}_1 - \bar{Y}_2$ is normal using the bootstrap appeared reasonable.

As a comparison, let us examine the output for the pooled procedure (which is inappropriate since variances are unequal). The p-value for the pooled t-test is 0.002, whereas the 95% confidence limits are 14.1 and 59.2. That is, we are 95% confident that the population mean andro level for men exceeds that for women by at least 14.1 but by no more than 59.2. These results are qualitatively similar to the Satterthwaite conclusions.

3.5 One-Sided Tests

One-sided tests for two-sample problems are where the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$ but the alternative is directional, either $H_A : \mu_1 - \mu_2 < 0$ (i.e., $\mu_1 < \mu_2$) or $H_A : \mu_1 - \mu_2 > 0$ (i.e., $\mu_1 > \mu_2$). Once you understand the general form of rejection regions and p-values for one-sample tests, the one-sided two-sample tests do not pose any new problems. Use the t-statistic, with the appropriate tail of the t-distribution to define critical values and p-values. One-sided two-sample tests are directly implemented in R, by specifying the type of test with `alternative = "less"` or `alternative = "greater"`. One-sided confidence bounds are given with the one-sided tests.

3.6 Paired Analysis

With paired data, inferences on $\mu_1 - \mu_2$ are based on the sample of differences within pairs. By taking differences within pairs, two dependent samples are transformed into one sample, which contains the relevant information for inferences on $\mu_d = \mu_1 - \mu_2$. To see this, suppose the observations within a pair are Y_1 and Y_2. Then within each pair, compute the difference $d = Y_1 - Y_2$:

$$\begin{aligned} d_1 &= Y_{11} - Y_{21} \\ d_2 &= Y_{12} - Y_{22} \\ &\vdots \\ d_n &= Y_{1n} - Y_{2n} \end{aligned}$$

If the Y_1 data are from a population with mean μ_1 and the Y_2 data are from a population with mean μ_2, then the ds are a sample from a population with mean $\mu_d = \mu_1 - \mu_2$. Furthermore, if the sample of differences comes from a normal population, then we can use standard one-sample techniques on $d_1, \ldots, d_n$ to test $\mu_d = 0$ (that is, $\mu_1 = \mu_2$), and to get a CI for $\mu_d = \mu_1 - \mu_2$.

Let $\bar{d} = n^{-1} \sum_i d_i = \bar{Y}_1 - \bar{Y}_2$ be the sample mean of the differences (which is also the mean difference), and let s_d be the sample standard deviation of the

differences. The standard error of $\bar{d}$ is $SE_{\bar{d}} = s_d/\sqrt{n}$, where n is the number of pairs. The paired t-test (two-sided) CI for μ_d is given by $\bar{d} \pm t_{\text{crit}} SE_{\bar{d}}$. To test $H_0 : \mu_d = 0$ ($\mu_1 = \mu_2$) against $H_A : \mu_d \neq 0$ ($\mu_1 \neq \mu_2$), use

$$t_s = \frac{\bar{d} - 0}{SE_{\bar{d}}}$$

to compute a p-value as in a two-sided one-sample test. One-sided tests are evaluated in the usual way for one-sample tests on means.

A graphical analysis of paired data focuses on the **sample of differences**, and not on the original samples. In particular, the normality assumption is assessed on the sample of differences.

3.6.1 R Analysis

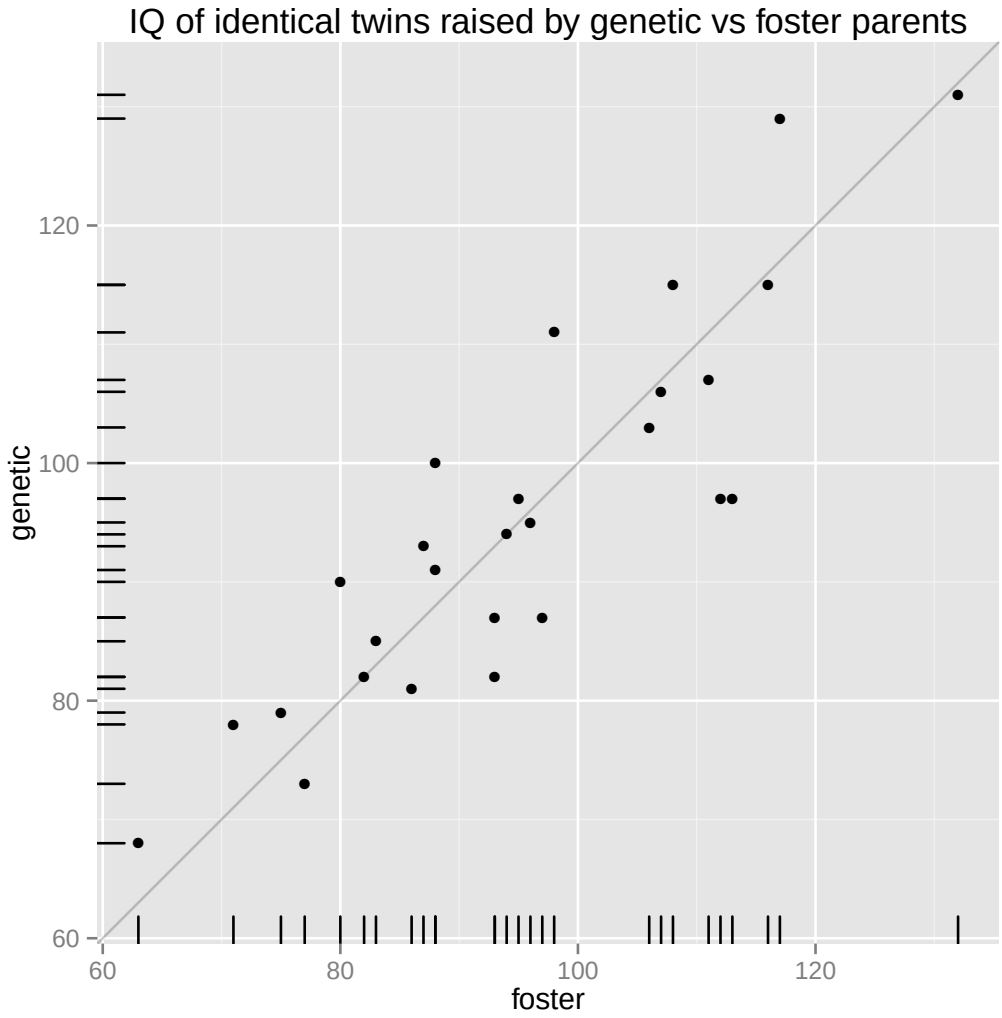
The most natural way to enter paired data is as two columns, one for each treatment group. You can then create a new column of differences, and do the usual one-sample graphical and inferential analysis on this column of differences, or you can do the paired analysis directly without this intermediate step.

Example: Paired Analysis of Data on Twins Burt (1966) presented data on IQ scores for identical twins that were raised apart, one by foster parents and one by the genetic parents. Assuming the data are a random sample of twin pairs, consider comparing the population mean IQs for twins raised at home to those raised by foster parents. Let μ_{f}=population mean IQ for twin raised by foster parents, and μ_{g}=population mean IQ for twin raised by genetic parents.

I created the data in the worksheet (c1=foster; c2=genetic), and computed the differences between the IQ scores for the children raised by the genetic and foster parents (c3=diff=genetic-foster). I also made a scatter plot of the genetic versus foster IQ scores.

```
#### Example: Paired Analysis of Data on Twins
# Data and numerical summaries
foster  <- c(82,  80,  88, 108, 116, 117, 132,  71,  75,  93,
             95,  88, 111,  63,  77,  86,  83,  93,  97,  87,
```

```
             94,  96, 112, 113, 106, 107,  98)
genetic <- c(82,  90,  91, 115, 115, 129, 131,  78,  79,  82,
             97, 100, 107,  68,  73,  81,  85,  87,  87,  93,
             94,  95,  97,  97, 103, 106, 111)
diff <- genetic - foster

axis.lim <- range(c(foster, genetic))

iq <- data.frame(foster, genetic, diff)

# scatterplot of foster and genetic IQs, with 1:1 line
p <- ggplot(iq, aes(x = foster, y = genetic))
# draw a 1:1 line, dots above line indicate "genetic > foster"
p <- p + geom_abline(intercept=0, slope=1, alpha=0.2)
p <- p + geom_point()
p <- p + geom_rug()
# make the axes square so it's a fair visual comparison
p <- p + coord_equal()
p <- p + scale_x_continuous(limits=axis.lim)
p <- p + scale_y_continuous(limits=axis.lim)
p <- p + labs(title = "IQ of identical twins raised by genetic vs foster parents")
print(p)
```

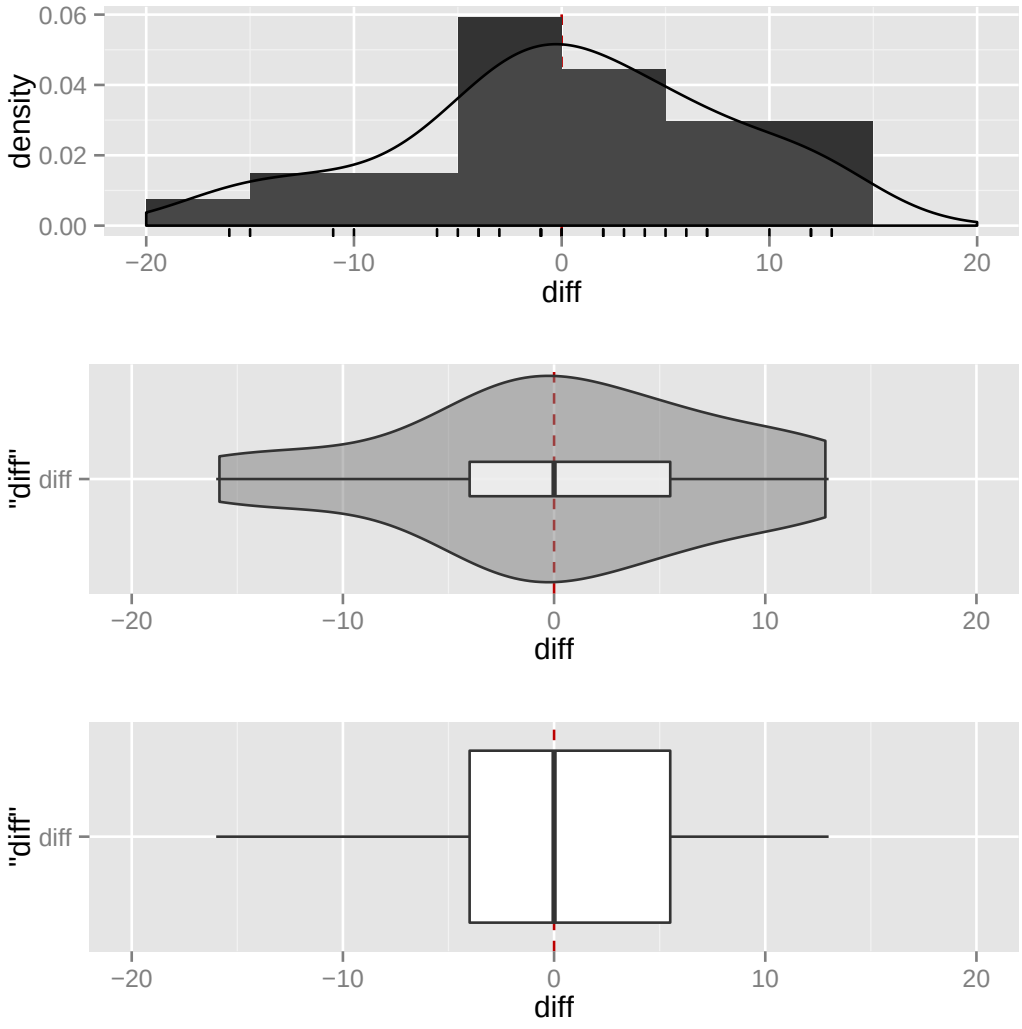


This plot of IQ scores shows that scores are related within pairs of twins. This is consistent with the need for a paired analysis.

Given the sample of differences, I created a boxplot and a stem and leaf display, neither which showed marked deviation from normality. The boxplot is centered at zero, so one would not be too surprised if the test result is insignificant.

```
p1 <- ggplot(iq, aes(x = diff))
p1 <- p1 + scale_x_continuous(limits=c(-20,+20))
# vertical line at 0
p1 <- p1 + geom_vline(xintercept=0, colour="#BB0000", linetype="dashed")
p1 <- p1 + geom_histogram(aes(y=..density..), binwidth=5)
p1 <- p1 + geom_density(alpha=0.1, fill="white")
p1 <- p1 + geom_rug()

# violin plot
p2 <- ggplot(iq, aes(x = "diff", y = diff))
```

```
p2 <- p2 + scale_y_continuous(limits=c(-20,+20))
p2 <- p2 + geom_hline(yintercept=0, colour="#BB0000", linetype="dashed")
p2 <- p2 + geom_violin(fill = "gray50", alpha=1/2)
p2 <- p2 + geom_boxplot(width = 0.2, alpha = 3/4)
p2 <- p2 + coord_flip()


# boxplot
p3 <- ggplot(iq, aes(x = "diff", y = diff))
p3 <- p3 + scale_y_continuous(limits=c(-20,+20))
p3 <- p3 + geom_hline(yintercept=0, colour="#BB0000", linetype="dashed")
p3 <- p3 + geom_boxplot()
p3 <- p3 + coord_flip()

library(gridExtra)
grid.arrange(p1, p2, p3, ncol=1)
```



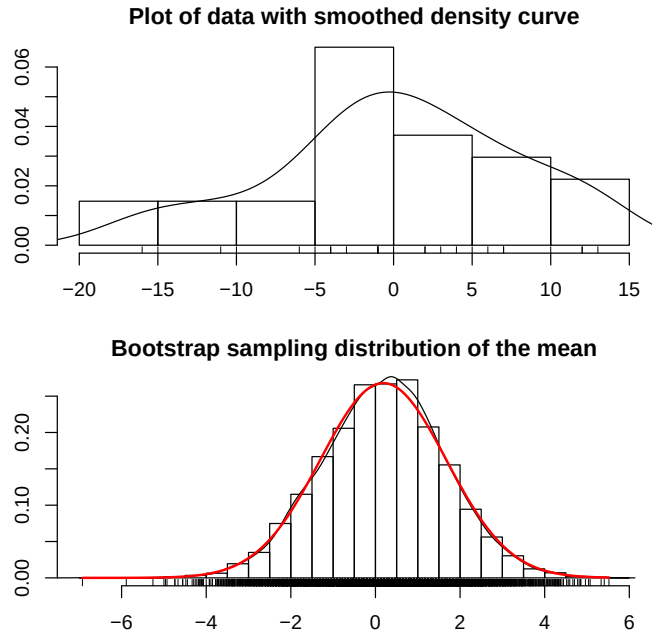
The normality assumption of the sample mean for a one-sample test is satisfied (below, left).

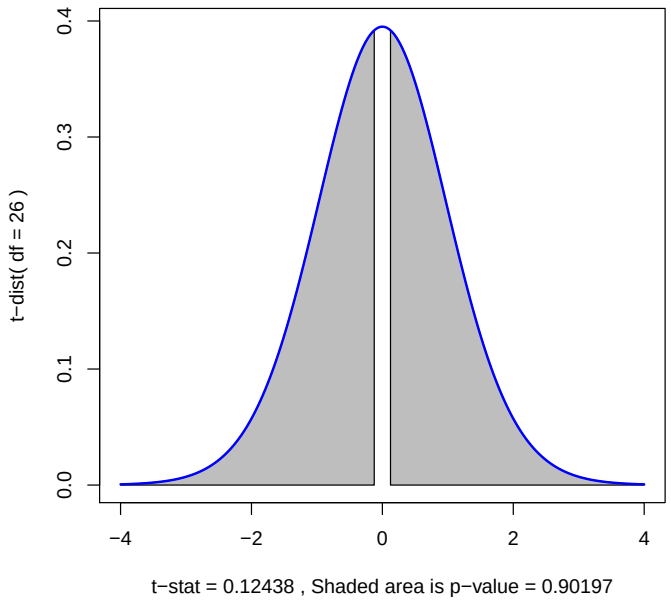
Given the sample of differences, I generated a one-sample CI and test. The hypothesis under test is $\mu_d = \mu_g - \mu_f = 0$. The p-value for this test is large. We do not have sufficient evidence to claim that the population mean IQs for twins raised apart are different. This is consistent with the CI for μ_d given below, which covers zero.

```
bs.one.samp.dist(iq$diff)

# one-sample t-test of differences (paired t-test)
t.summary <- t.test(iq$diff)
t.summary
```

```
##
##  One Sample t-test
##
## data:  iq$diff
## t = 0.12438, df = 26, p-value = 0.902
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.875159  3.245529
## sample estimates:
## mean of x
## 0.1851852
```

```
# plot t-distribution with shaded p-value
t.dist.pval(t.summary)
```





Alternatively, I can generate the test and CI directly from the raw data in two columns, specifying `paired=TRUE`. This gives the following output, which

leads to identical conclusions to the earlier analysis.

```
# two-sample paired t-test
t.summary <- t.test(iq$genetic, iq$foster, paired=TRUE)
t.summary
```

```
##
##  Paired t-test
##
## data:  iq$genetic and iq$foster
## t = 0.12438, df = 26, p-value = 0.902
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.875159  3.245529
## sample estimates:
## mean of the differences
##               0.1851852
```

You might ask why I tortured you by doing the first analysis, which required creating and analyzing the sample of differences, when the alternative and equivalent second analysis is so much easier. (A later topic deals with non-parametric analyses of paired data for which the differences must be first computed.)

Remark: I could have defined the difference to be the foster IQ score minus the genetic IQ score. How would this change the conclusions?

Example: Paired Comparisons of Two Sleep Remedies The following data give the amount of sleep gained in hours from two sleep remedies, A and B, applied to 10 individuals who have trouble sleeping an adequate amount. Negative values imply sleep loss. In 9 of the 10 individuals, the sleep gain on B exceeded that on A.

Let μ_A = population mean sleep gain (among troubled sleepers) on remedy A, and μ_B = population mean sleep gain (among troubled sleepers) on remedy B. Consider testing $H_0 : \mu_B - \mu_A = 0$ or equivalently $\mu_d = 0$, where $\mu_d = \mu_B - \mu_A$.

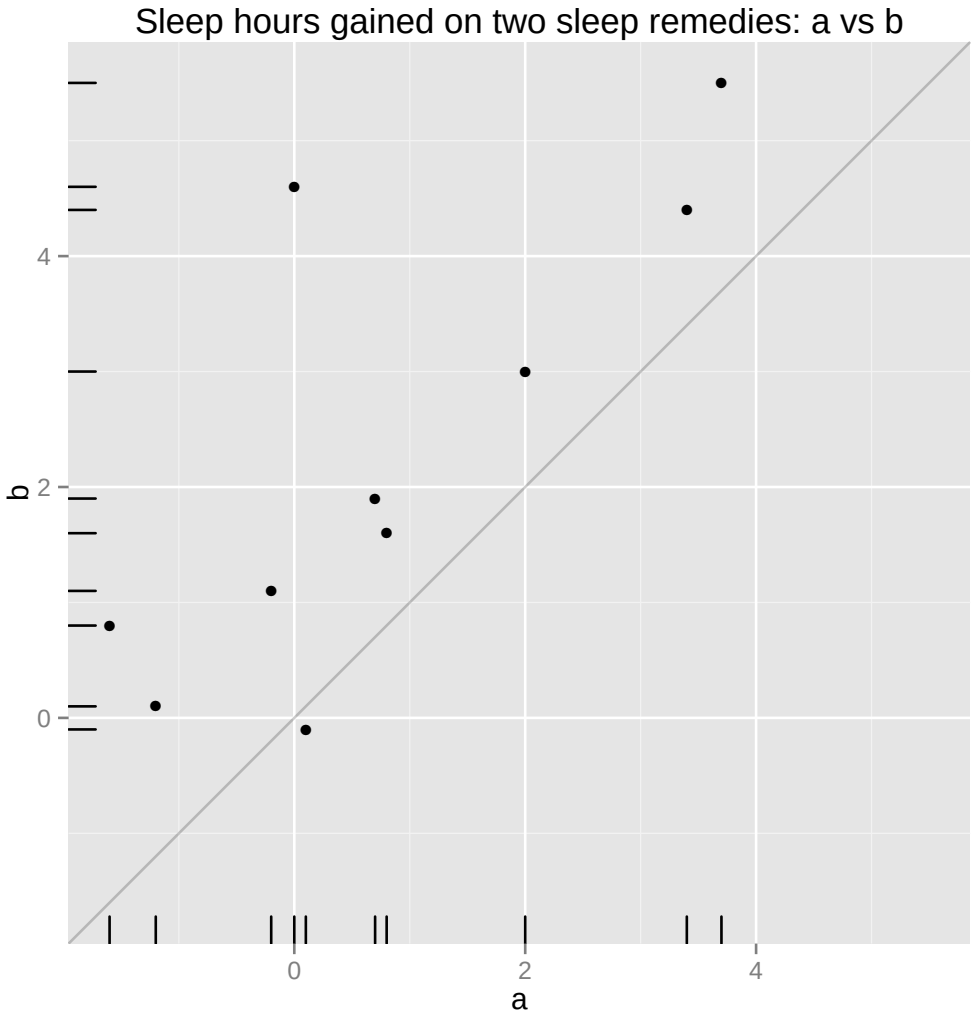
The observed distribution of differences between B and A is slightly skewed to the right, with a single outlier in the upper tail. The normality assumption of the standard one-sample t-test and CI are suspect here. I will continue with

the analysis, anyways.

```
#### Example: Paired Comparisons of Two Sleep Remedies
# Data and numerical summaries
a <- c( 0.7, -1.6, -0.2, -1.2,  0.1,  3.4,  3.7,  0.8,  0.0,  2.0)
b <- c( 1.9,  0.8,  1.1,  0.1, -0.1,  4.4,  5.5,  1.6,  4.6,  3.0)
d <- b - a;
sleep <- data.frame(a, b, d)
sleep
```

```
##       a    b    d
## 1   0.7  1.9  1.2
## 2  -1.6  0.8  2.4
## 3  -0.2  1.1  1.3
## 4  -1.2  0.1  1.3
## 5   0.1 -0.1 -0.2
## 6   3.4  4.4  1.0
## 7   3.7  5.5  1.8
## 8   0.8  1.6  0.8
## 9   0.0  4.6  4.6
## 10  2.0  3.0  1.0
```

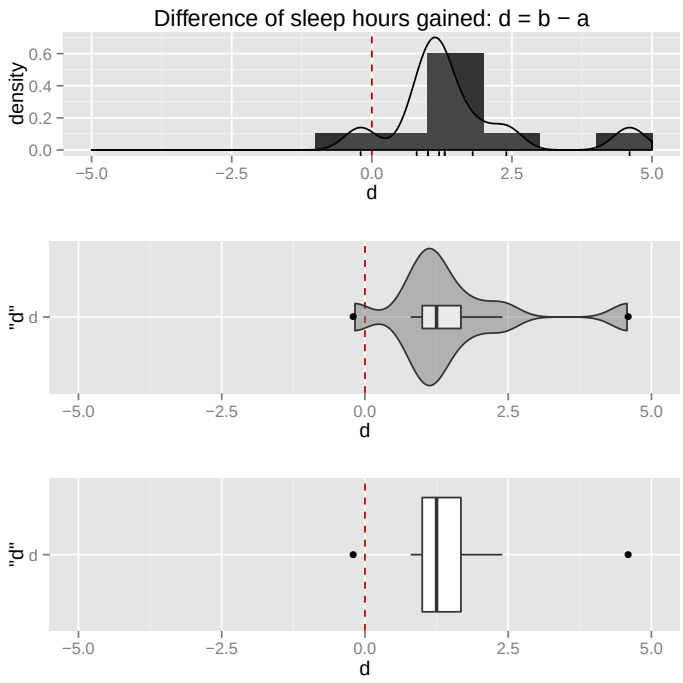
```
axis.lim <- range(c(a, b))

# scatterplot of a and b IQs, with 1:1 line
p <- ggplot(sleep, aes(x = a, y = b))
# draw a 1:1 line, dots above line indicate "b > a"
p <- p + geom_abline(intercept=0, slope=1, alpha=0.2)
p <- p + geom_point()
p <- p + geom_rug()
# make the axes square so it's a fair visual comparison
p <- p + coord_equal()
p <- p + scale_x_continuous(limits=axis.lim)
p <- p + scale_y_continuous(limits=axis.lim)
p <- p + labs(title = "Sleep hours gained on two sleep remedies: a vs b")
print(p)
```

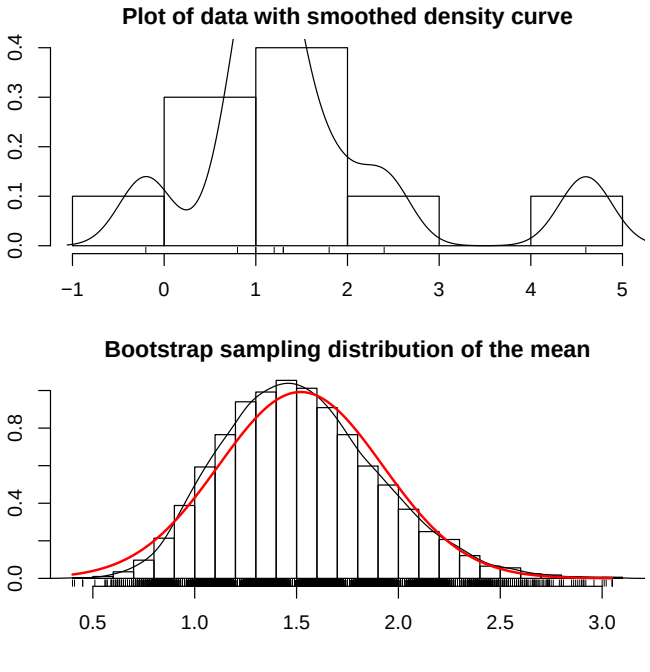


There is evidence here against the normality assumption of the sample mean. We'll continue anyway (in practice we'd use a nonparametric method, instead, in a later chapter).

```
p1 <- ggplot(sleep, aes(x = d))
p1 <- p1 + scale_x_continuous(limits=c(-5,+5))
# vertical line at 0
p1 <- p1 + geom_vline(xintercept=0, colour="#BB0000", linetype="dashed")
p1 <- p1 + geom_histogram(aes(y=..density..), binwidth=1)
p1 <- p1 + geom_density(alpha=0.1, fill="white")
p1 <- p1 + geom_rug()
p1 <- p1 + labs(title = "Difference of sleep hours gained: d = b - a")

# violin plot
p2 <- ggplot(sleep, aes(x = "d", y = d))
p2 <- p2 + scale_y_continuous(limits=c(-5,+5))
p2 <- p2 + geom_hline(yintercept=0, colour="#BB0000", linetype="dashed")
p2 <- p2 + geom_violin(fill = "gray50", alpha=1/2)
p2 <- p2 + geom_boxplot(width = 0.2, alpha = 3/4)
```

```
p2 <- p2 + coord_flip()


# boxplot
p3 <- ggplot(sleep, aes(x = "d", y = d))
p3 <- p3 + scale_y_continuous(limits=c(-5,+5))
p3 <- p3 + geom_hline(yintercept=0, colour="#BB0000", linetype="dashed")
p3 <- p3 + geom_boxplot()
p3 <- p3 + coord_flip()

library(gridExtra)
grid.arrange(p1, p2, p3, ncol=1)


bs.one.samp.dist(sleep$d)
```





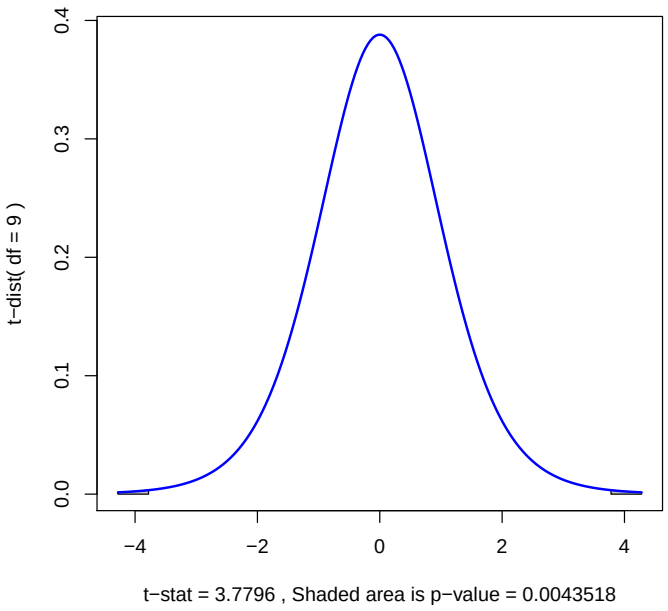
The p-value for testing H_0 is 0.004. We'd reject H_0 at the 5% or 1% level, and conclude that the population mean sleep gains on the remedies are different. We are 95% confident that μ_B exceeds μ_A by between 0.61 and 2.43 hours. Again, these results must be reported with caution, because the normality assumption is unreasonable. However, the presence of outliers tends to make the t-test and CI conservative, so we'd expect to find similar conclusions if we used the nonparametric methods discussed later in the semester.

```
# one-sample t-test of differences (paired t-test)
t.summary <- t.test(sleep$d)
```

```
t.summary
```

```
##
##  One Sample t-test
##
## data:  sleep$d
## t = 3.7796, df = 9, p-value = 0.004352
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.610249 2.429751
## sample estimates:
## mean of x
##      1.52
```

```
# plot t-distribution with shaded p-value
t.dist.pval(t.summary)
```



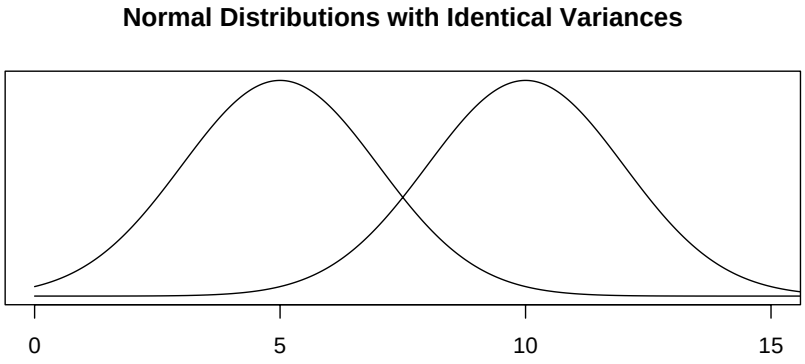
Question: In what order should the remedies be given to the patients?

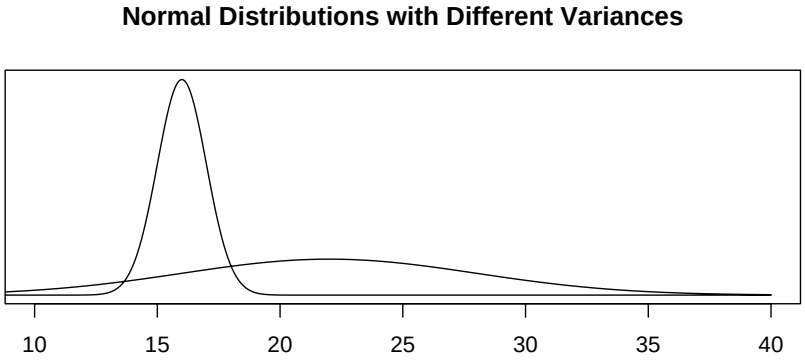
■ ClickerQs — Reporting results, STT.06.03.010 ■

3.7 Should You Compare Means?

The mean is the most common feature on which two distributions are compared. You should not, however, blindly apply the two-sample tests (paired or unpaired) without asking yourself whether the means are the relevant feature to compare. This issue is not a big concern when, as highlighted in the first graph below, the two (normal) populations have equal spreads or standard deviations. In such cases the difference between the two population means is equal to the difference between any fixed percentile for the two distributions, so the mean difference is a natural measure of difference.

Consider instead the hypothetical scenario depicted in the bottom pane below, where the population mean lifetimes using two distinct drugs for a fatal disease are $\mu_1 = 16$ months from time of diagnosis and $\mu_2 = 22$ months from time of diagnosis, respectively. The standard deviations under the two drugs are $\sigma_1 = 1$ and $\sigma_2 = 6$, respectively. The second drug has the higher mean lifetime, but at the expense of greater risk. For example, the first drug gives you a 97.7% chance of living at least 14 months, whereas the second drug only gives you a 90.8% chance of living at least 14 months. Which drug is best? It depends on what is important to you, a higher expected lifetime or a lower risk of dying early.


Normal Distributions with Identical Variances
0
5
10
15


Normal Distributions with Different Variances
10
15
20
25
30
35
40