Chapter 11

Two Sample Inference

In this chapter, we discuss two-sample inference procedures for the following population parameters:

- The difference of two population means $\mu_1 - \mu_2$
- The ratio of two population variances σ_2^2/σ_1^2
- The difference of two population proportions $p_1 - p_2$
- The difference of two population medians $M_1 - M_2$

Remember that these are population-level quantities (now involving two different populations), so they are unknown. Our goal is to use sample information (now with two samples) to estimate these quantities.

Usefulness: In practice, it is very common to compare the same characteristic (e.g., mean, variance, proportion, etc.) on two different populations. For example, we may wish to compare

- the population mean starting salaries of male and female engineers (compare μ_1 and μ_2). Is there evidence that males have a larger mean starting salary?
- the population variance of sound levels from two indoor swimming pool designs (compare σ_1^2 and σ_2^2). Are the sound-level acoustics of a new design more variable than the standard design?
- the population proportion of defectives produced from two different suppliers (compare p_1 and p_2). Are there differences between the two suppliers?

Note: Our methods in the last chapter are applicable only for a single population (i.e., a population mean μ, a population variance σ^2, and a population proportion p). We therefore extend these methods to two populations. We start with comparing two population means.

11.1 The Difference of Two Population Means

Setting: Suppose that we have two independent random samples:

$$\begin{aligned} &\text{Sample 1:} \quad X_{11}, X_{12}, ..., X_{1n_1} \sim \text{Gaussian}\left(\mu_1, \sigma_1^2\right) \\ &\text{Sample 2:} \quad X_{21}, X_{22}, ..., X_{2n_1} \sim \text{Gaussian}\left(\mu_2, \sigma_2^2\right). \end{aligned}$$

Point estimators: Define the statistics

$$\begin{aligned} \bar{X}_{1\cdot} &= \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} = \text{sample mean for sample 1} \\ \bar{X}_{2\cdot} &= \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} = \text{sample mean for sample 2} \\ S_1^2 &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_{1\cdot})^2 = \text{sample variance for sample 1} \\ S_2^2 &= \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_{2\cdot})^2 = \text{sample variance for sample 2} \end{aligned}$$

Goal: Our goal is to construct a $100(1-\alpha)$ percent confidence interval for the difference of two population means $\mu_1 - \mu_2$.

Important: How we construct this interval depends on our assumptions on the population variances σ_1^2 and σ_2^2. In particular, we consider two cases:

- $\sigma_1^2 = \sigma_2^2$
- $\sigma_1^2 \neq \sigma_2^2$

11.1.1 Independent samples: Equal population variances

Result: Under the assumptions above and when $\sigma_1^2 = \sigma_2^2$, the quantity

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}.$$

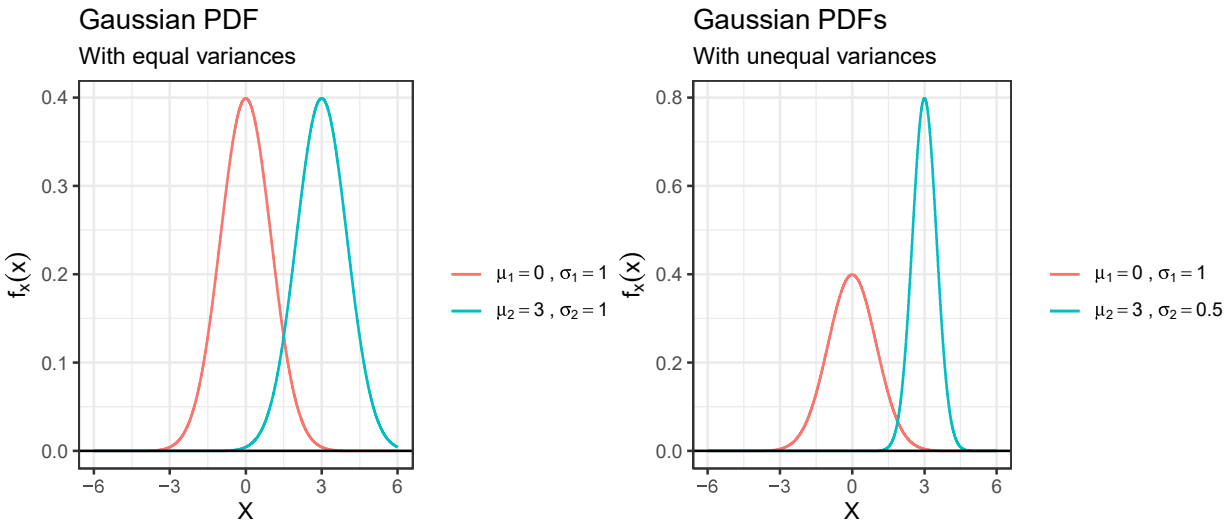

Figure 11.1.1: Two Gaussian distributions with $\sigma_1^2 = \sigma_2^2$ (left) and $\sigma_1^2 \neq \sigma_2^2$ (right).

- For this sampling distribution to hold exactly, we need
  - the two random samples to be independent
  - the two population distributions to be Gaussian (normal)
  - the two population distributions to have the same variance; i.e., $\sigma_1^2 = \sigma_2^2$.

The statistic S_p^2 is called the pooled sample variance estimator of the common population variance, say, σ^2. It is a weighted average of the two sample variances S_1^2 (where the weights are functions of the sample sizes n_1 and n_2).

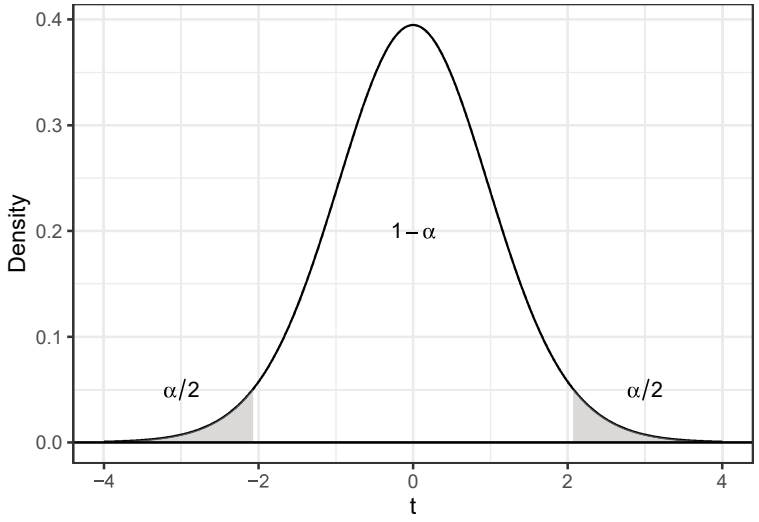

Figure 11.1.2: A t pdf with $n_1 + n_2 - 2$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $t_{n1_+n_2-2,1-\alpha/2}$ (upper) and $t_{n1_+n_2-2,\alpha/2}$ (lower), respectively.

The sampling distribution $t \sim t(n_1 + n_2 - 2)$ suggests that confidence interval quantiles will come from this t distribution; note that this distribution depends on the sample sizes from both samples.

In particular, because $t \sim t(n_1 + n_2 - 2)$, the upper quantile $t_{n_1+n_2-2,1-\alpha/2}$ satisfies

$$P\left(t_{n_1+n_2-2,\alpha/2} < \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2,1-\alpha/2} \right) = 1 - \alpha.$$

This probability equation is seen by examining Figure 11.1.2.

After performing algebraic manipulations (similar to those in the last chapter), we obtain

$$\left(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}\right) \pm t_{n_1+n_2-2,1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

This is a $100(1-\alpha)$ percent confidence interval for the difference of two population means $\mu_1 - \mu_2$.

We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot})} \pm \underbrace{\text{quantile}}_{t_{n_1+n_2-2,1-\alpha/2}} \underbrace{\text{standard error}}_{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

We interpret the interval in the same way.

> "We are $100(1-\alpha)$ percent confident that the population mean difference $\mu_1 - \mu_2$ is in this interval."

Important: In two-sample situations, it is usually of interest to compare the population means μ_1 and μ_2:

- If the confidence interval for $\mu_1 - \mu_2$ includes 0, this does not suggest that the population means μ_1 and μ_2 are different.
- If the confidence interval for $\mu_1 - \mu_2$ does not include 0, this suggests the population means are different.

We can also build a hypothesis test, under the assumption that $\sigma_1^2 = \sigma_2^2$, using this sampling distribution,

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Again, we choose between the three sets of hypotheses,

$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_0 : \mu_1 - \mu_2 = \mu_0$
$H_a : \mu_1 - \mu_2 < \mu_0$	$H_a : \mu_1 - \mu_2 > \mu_0$	$H_a : \mu_1 - \mu_2 \neq \mu_0$.

Often, we take $\mu_0 = 0$

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = 0 & \text{or} & H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 < 0 & & H_a : \mu_1 - \mu_2 > 0 \end{array} \quad \text{or} \quad H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0,$$

so that we can test $\mu_1 < \mu_2$, $\mu_1 > \mu_2$, or $\mu_1 \neq \mu_2$ (from right to left).

The test statistic for this test is

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_0)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

We succinctly summarize the five steps for the t hypothesis test, noting many of the steps are the same as the one-sample version

Step One	$H_a : \mu_1 - \mu_2 < \mu_0$	$H_a : \mu_1 - \mu_2 > \mu_0$	$H_a : \mu_1 - \mu_2 \neq \mu_0$
Step Two	Check Assumptions		
Step Three	$t^* = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_0)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$		
Step Four	$P(T_{n_1+n_2-2} < t^*)$	$P(T_{n_1+n_2-2} > t^*)$	$2P(T_{n_1+n_2-2} < -\lvert t^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

Example 11.1. In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the population mean weight in fish (of the same species) from two locations. Independent samples are taken from each location and the following weights (in ounces) are observed:

Location 1:	21.9	18.5	12.3	16.7	21.0	15.1	18.2	23.0	36.8	26.6
Location 2:	21.0	19.6	14.4	16.9	23.4	14.6	10.4	16.5		

We want to construct a 90 percent confidence interval for the population mean weight difference $\mu_1 - \mu_2$, where the mean weight μ_1 (μ_2) corresponds to location 1 (2).

In order to visually assess the equal population variance assumption $\sigma_1^2 = \sigma_2^2$ we use violin plots to display the data for each sample; see Figure 11.1.3.

```
> ggdat<-data.frame(weights=c(21.9,18.5,12.3,16.7,21.0,15.1,18.2,23.0,36.8,26.6,
+                             21.0,19.6,14.4,16.9,23.4,14.6,10.4,16.5),
+                   locations=c(rep("Location 1",10),rep("Location 2",8)))
> ggplot(data=ggdat,aes(x=locations, y=weights))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("Sampling Location")+
+   ylab("Weight in Ounces")
```

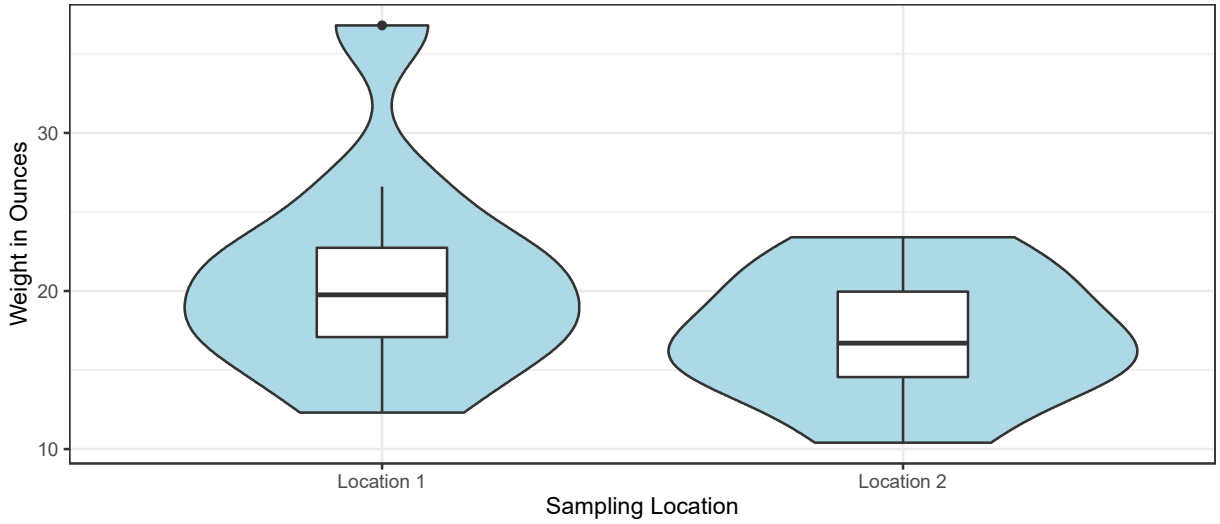

Figure 11.1.3: Violin plots of fish weight data by location.

The equal variance assumption looks reasonable; the spread in each distribution looks roughly the same (save the outlier in Location 1). Later, we will look at formal statistical inference procedures to compare two population variances. For now, we will rely on this rough assessment (based on sample information only; no inference).

The 90% confidence interval is then calculated using the formula, as follows.

```
> alpha<-0.10
> (n.1<-length(which(ggdat1$locations=="Location 1")))
[1] 10
> (n.2<-length(which(ggdat1$locations=="Location 2")))
[1] 8
> (xbar.1<-mean(ggdat1$weights[which(ggdat1$locations=="Location 1")]))
[1] 21.01
> (xbar.2<-mean(ggdat1$weights[which(ggdat1$locations=="Location 2")]))
[1] 17.1
> (sd.1<-sd(ggdat1$weights[which(ggdat1$locations=="Location 1")]))
[1] 6.903212
> (sd.2<-sd(ggdat1$weights[which(ggdat1$locations=="Location 2")]))
[1] 4.140048
> (sd.p<-sqrt(((n.1-1)*sd.1^2+(n.2-1)*sd.2^2)/(n.1+n.2-2)))
[1] 5.856988
> (xbar.1-xbar.2) + c(-1,1)*qt(1-alpha/2,df=n.1+n.2-2)*sqrt(sd.p^2*(1/n.1+1/n.2))
[1] -0.9404376  8.7604376
```

We can use R to calculate the confidence interval directly:

```
> t.test(ggdat1$weights[which(ggdat1$locations=="Location 1")],
+        ggdat1$weights[which(ggdat1$locations=="Location 2")],
+        conf.level=0.90,var.equal=TRUE)
Two Sample t-test

data:  ggdat1$weights[which(ggdat1$locations == "Location 1")] and
       ggdat1$weights[which(ggdat1$locations == "Location 2")]
t = 1.4074, df = 16, p-value = 0.1784
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
-0.9404376  8.7604376
sample estimates:
mean of x mean of y
21.01     17.10
```

A 90 percent confidence interval for the population mean difference $\mu_1 - \mu_2$ is

$$(-0.940, 8.760) \text{ oz.}$$

Interpretation: We are 90 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -0.940 and 8.760 oz. Note that this interval includes "0." Therefore, we do not have sufficient evidence that the population mean fish weights μ_1 and μ_2 are different.

Robustness: Some comments are in order about the robustness properties of the two-sample confidence interval

$$\left(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}\right) \pm t_{n_1+n_2-2,1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

for the difference of two population means $\mu_1 - \mu_2$.

We should only use this confidence interval if there is strong evidence that the population variances σ_1^2 and σ_2^2 are similar. In other words, this confidence interval is not robust to a violation of the equal population variance assumption. Like the one-sample t confidence interval for a single population mean μ, this two sample t confidence interval is robust to mild departures from normality. This means that we can feel comfortable using the interval even if the underlying population distributions are not perfectly Gaussian (normal).

Remark: With such small sample sizes in the last example ($n_1 = 10$ and $n_2 = 8$), it is hard to make any conclusive assessments about the Gaussian population assumption. This uncertainty manifests itself in the qq plots; note how the "bands of uncertainty" are very wide in Figure 11.1.4. The very heavy fish (36.8 oz) in the first sample might be regarded as an outlier, because it is the only observation that falls outside the bands.

```
> library("qqplotr")
> ggplot(data=ggdat,aes(sample=weights))+
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")+
+   facet_grid(. ~ locations)
```

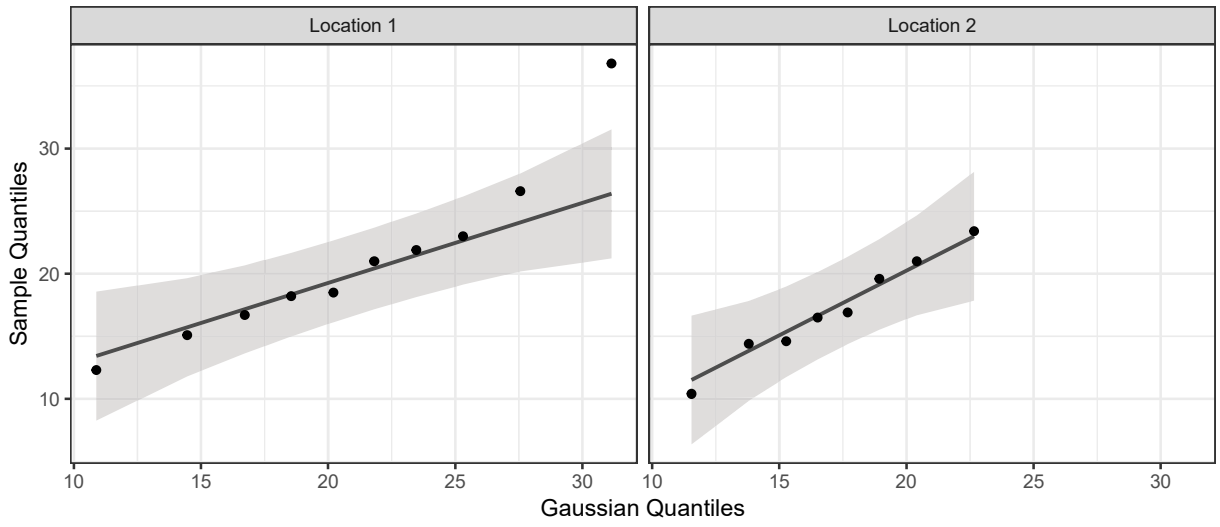

Figure 11.1.4: Quantile-quantile plots for the fish weight data.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.1.5, created in R as follows.

```
> alpha<-0.10 #significance level
> test<-t.test(ggdat$weights[which(ggdat$locations=="Location 1")],
+              ggdat$weights[which(ggdat$locations=="Location 2")],
+              conf.level=0.90,var.equal=TRUE)
> ggdat<-data.frame(t=seq(from=-4,to=4,by=0.01),
+                   f=dt(seq(from=-4,to=4,by=0.01),df=test$parameter))
> ggdat.highlight<-data.frame(x=qt(p=c(alpha/2,1-alpha/2),df=test$parameter),
+                             y=c(0,0))
> axis.labels<-round(c(qt(alpha/2,test$parameter),0,qt(1-alpha/2,test$parameter))*
+                      test$stderr + (-diff(test$estimate)),3)
> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t<=qt(alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   annotate("text",x=-3.1,y=0.05,label= deparse(bquote(alpha/2==0.05)),parse=TRUE,size=3.5)+
+   annotate("text",x=3.1,y=0.05,label= deparse(bquote(alpha/2==0.05)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label= deparse(bquote(1-alpha==0.90)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~., breaks=c(qt(alpha/2,test$parameter),0,
+                                                       qt(1-alpha/2,test$parameter)),
+          labels = axis.labels,name="Difference of Weight in Ounces"))
```

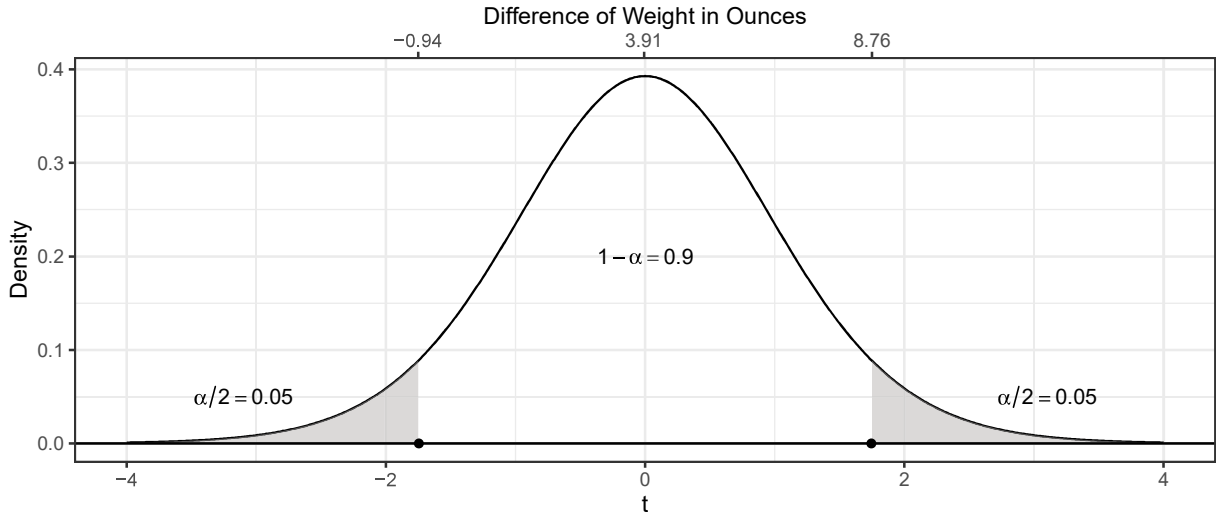

Figure 11.1.5: The approximate sampling distribution of $(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})$ with key values of the confidence interval for the population mean difference highlighted.

11.1.2 Independent samples: Unequal population variances

When $\sigma_1^2 \neq \sigma_2^2$, we can not use

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

as a basis for inference because the pooled sample variance S_p^2 does not estimate anything meaningful. Constructing a confidence interval for $\mu_1 - \mu_2$ becomes more difficult theoretically. However, we can write an approximate confidence interval.

Result: Under the assumptions above and when $\sigma_1^2 \neq \sigma_2^2$, the quantity

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(v),$$

where the degrees of freedom v is calculated as

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

An approximate $100(1-\alpha)$ percent confidence interval for the difference of two population means $\mu_1 - \mu_2$ is given by

$$(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) \pm t_{v,1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where the degrees of freedom v is calculated as

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

This interval is always approximately valid, as long as

- the two samples are independent
- the two population distributions are approximately Gaussian (normal).

This interval remains valid even when $\sigma_1^2 \neq \sigma_2^2$, but its theoretical properties are not as good as those of the equal population variance interval. No one in their right mind would calculate this interval "by hand" (particularly the formula for v). `R` will produce the interval on request.

We can also build a hypothesis test, under the assumption that $\sigma_1^2 \neq \sigma_2^2$, using this sampling distribution,

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(v),$$

where the degrees of freedom v is calculated as

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Again, we choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = \mu_0 & \text{or} & H_0 : \mu_1 - \mu_2 = \mu_0 \\ H_a : \mu_1 - \mu_2 < \mu_0 & & H_a : \mu_1 - \mu_2 > \mu_0 \end{array} \quad \text{or} \quad \begin{array}{l} H_0 : \mu_1 - \mu_2 = \mu_0 \\ H_a : \mu_1 - \mu_2 \neq \mu_0. \end{array}$$

Often, we take $\mu_0 = 0$

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = 0 & \text{or} & H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 < 0 & & H_a : \mu_1 - \mu_2 > 0 \end{array} \quad \text{or} \quad \begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0, \end{array}$$

so that we can test $\mu_1 < \mu_2$, $\mu_1 > \mu_2$, or $\mu_1 \neq \mu_2$ (from right to left).

The test statistic for this test is

$$t = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_0)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

We succinctly summarize the five steps for the t hypothesis test, noting many of the steps are the same as the one-sample version

Step One	$H_a : \mu_1 - \mu_2 < \mu_0$	$H_a : \mu_1 - \mu_2 > \mu_0$	$H_a : \mu_1 - \mu_2 \neq \mu_0$
Step Two	Check Assumptions		
Step Three	$t^* = \frac{(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) - (\mu_0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$		
Step Four	$P(T_v < t^*)$	$P(T_v > t^*)$	$2P(T_v < -\lvert t^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

Example 11.2. You are part of a recycling project that is examining how much paper is being discarded (not recycled) by employees at two large plants. These data are obtained on the amount of white paper thrown out per year by employees (data are in hundreds of pounds). Samples of employees at each plant were randomly selected.

Plant 1:	3.01	2.58	3.04	1.75	2.87	2.57	2.51	2.93	2.85	3.09
	1.43	3.36	3.18	2.74	2.25	1.95	3.68	2.29	1.86	2.63
	2.83	2.04	2.23	1.92	3.02					
Plant 2:	3.99	2.08	3.66	1.53	4.27	4.31	2.62	4.52	3.80	5.30
	3.41	0.82	3.03	1.95	6.45	1.86	1.87	3.98	2.74	4.81

We want to construct a 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$, where the mean amount discarded μ_1 (μ_2) corresponds to Plant 1 (2).

In order to visually assess the equal population variance assumption $\sigma_1^2 = \sigma_2^2$ we use violin plots to display the data for each sample; see Figure 11.1.6. This figure suggests the equal population variance assumption is doubtful. The spread in the two violin plots is markedly different (again, this is a rough determination based on the sample information).

```
> ggdat<-data.frame(lbs.disgarded=c(3.01,2.58,3.04,1.75,2.87,2.57,2.51,2.93,
+                                   2.85,3.09,1.43,3.36,3.18,2.74,2.25,1.95,
+                                   3.68,2.29,1.86,2.63,2.83,2.04,2.23,1.92,
```

```
+                                3.02,3.99,2.08,3.66,1.53,4.27,4.31,2.62,
+                                4.52,3.80,5.30,3.41,0.82,3.03,1.95,6.45,
+                                1.86,1.87,3.98,2.74,4.81),
+                  plant=c(rep("Plant 1",25),rep("Plant 2",20)))
> ggplot(data=ggdat,aes(x=plant, y=lbs.disgarded))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("Sampling Plant")+
+   ylab("Paper Disgarded (100s of pounds)")
```

Figure 11.1.6: Violin plots of the amount of white paper thrown out per year by employees (data are in hundreds of pounds) by plant.

We can use R to calculate the confidence interval directly.

```
> t.test(x=ggdat$lbs.disgarded[which(ggdat$plant=="Plant 1"],
+        y=ggdat$lbs.disgarded[which(ggdat$plant=="Plant 2")],
+        ,conf.level=0.95,var.equal=FALSE)

        Welch Two Sample t-test

data:  ggdat$lbs.disgarded[which(ggdat$plant == "Plant 1")] and
ggdat$lbs.disgarded[which(ggdat$plant == "Plant 2")]
t = -2.2716, df = 23.614, p-value = 0.03252
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.46179176 -0.06940824
sample estimates:
mean of x mean of y
2.5844    3.3500
```

A 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$ is

$$(-1.461, -0.069) \text{ 100s lbs.}$$

We are 95 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -146.1 and -6.9 lbs. This interval does not include "0" and contains only negative values. Therefore, we have evidence that the population mean amount of discarded paper is smaller for Plant 1 than it is for Plant 2.

Gaussian qq plots for the two samples of white paper data are given in Figure 11.1.7. There is no cause to question the normality assumption.

```
> ggplot(data=ggdat,aes(sample=lbs.disgarded))+
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")+
+   facet_grid(. ~ plant)
```

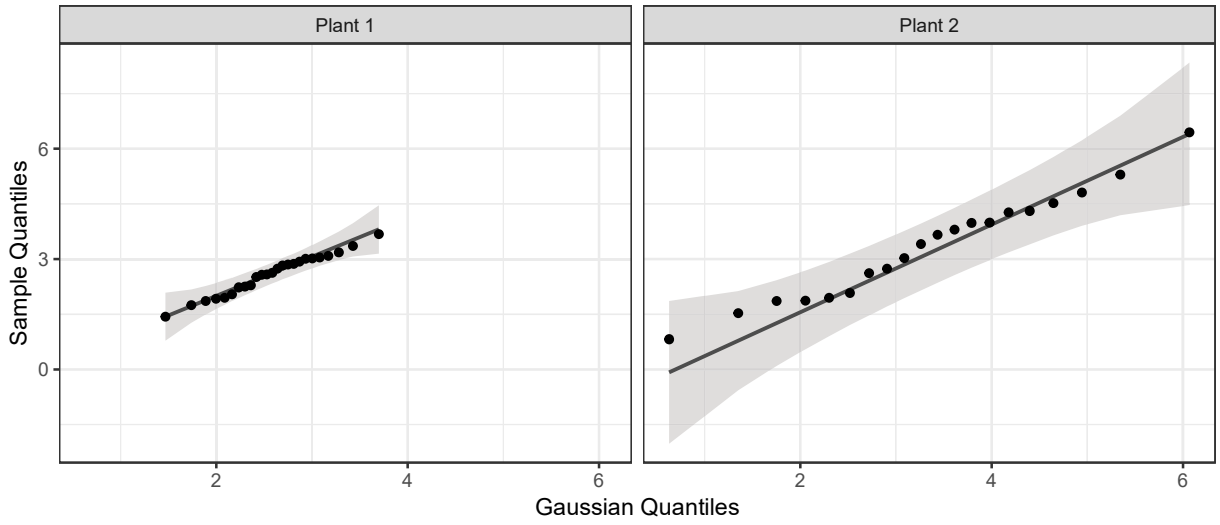

Figure 11.1.7: Quantile-quantile plots for the amount of white paper thrown out per year by employees (data are in hundreds of pounds) by plant.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.1.8, created in R as follows.

```
> alpha<-0.05 #significance level
> test<-t.test(x=ggdat$lbs.disgarded[which(ggdat$plant=="Plant 1")],
+              y=ggdat$lbs.disgarded[which(ggdat$plant=="Plant 2")],
+              conf.level=0.95,var.equal=FALSE)
> ggdat<-data.frame(t=seq(from=-4,to=4,by=0.01),
+                   f=dt(seq(from=-4,to=4,by=0.01),df=test$parameter))
> ggdat.highlight<-data.frame(x=qt(p=c(alpha/2,1-alpha/2),df=test$parameter),
+                             y=c(0,0))
> axis.labels<-round(c(qt(alpha/2,test$parameter),0,qt(1-alpha/2,test$parameter))*
+                    test$stderr + (-diff(test$estimate)),3)
> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t<=qt(alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   annotate("text",x=-3.1,y=0.05,label= deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=3.1,y=0.05,label= deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label= deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~., breaks=c(qt(alpha/2,test$parameter),0,
+                                                       qt(1-alpha/2,test$parameter)),
+           labels = axis.labels,name="Difference of Paper Disgarded (100s of pounds)"))
```

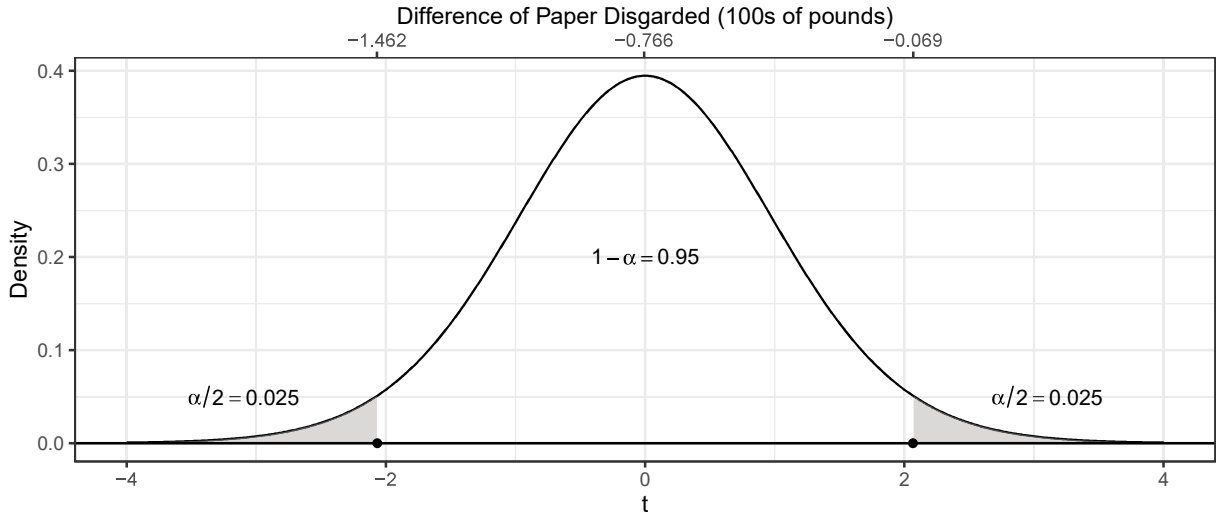

Figure 11.1.8: The approximate sampling distribution of $(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})$ with key values of the confidence interval for the population mean difference highlighted.

Remark: We have presented two confidence intervals for the population mean difference $\mu_1 - \mu_2$. One assumes $\sigma_1^2 = \sigma_2^2$ (equal population variances) and one that assumes $\sigma_1^2 \neq \sigma_2^2$ (unequal population variances). If you are unsure about which interval to use, go with the unequal variance interval. The penalty for using it when $\sigma_1^2 = \sigma_2^2$ is much smaller than the penalty for using the equal variance interval when $\sigma_1^2 \neq \sigma_2^2$. We usually calculate both intervals (easy to do quickly with `R`) and then determine whether the intervals lead to drastically different conclusions.

11.1.3 Dependent samples: Matched pairs

Example 11.3. Ergonomics experts hired by a large company designed a study to determine whether more varied work conditions would have any impact on arm movement. The data on the next page were obtained on a random sample of $n = 26$ employees. Each observation is the amount of time, expressed as a percentage of the total time observed, during which arm elevation was below 30 degrees. This percentage is a surrogate for the percentage of time spent on repetitive tasks. The two measurements from each employee were obtained 18 months apart. During this 18-month period, work conditions were "changed" by the ergonomics team, and subjects were allowed to engage in a wider variety of work tasks. "Before" and "after" measurements are obtained on each of the 26 employees as reported in Table 11.1.1

Individual	Before	After	Individual	Before	After
1	81.3	78.9	14	74.9	58.3
2	87.2	91.4	15	75.8	62.5
3	86.1	78.3	16	72.6	70.2
4	82.2	78.3	17	80.8	58.7
5	90.8	84.4	18	66.5	66.6
6	86.9	67.4	19	72.2	60.7
7	96.5	92.8	20	56.5	65.0
8	73.0	69.9	21	82.4	73.7
9	84.2	63.8	22	88.8	80.4
10	74.5	69.7	23	80.0	78.8
11	72.0	68.4	24	91.1	81.8
12	73.8	71.8	25	97.5	91.6
13	74.2	58.3	26	70.0	74.2

Table 11.1.1: Ergonomics data. Percentage of time arm elevation was less than 30 degrees.

Question: Does the population mean time (during which elevation is below 30 degrees) decrease after the ergonomics team changes the working conditions?

A **matched-pairs design** is an experimental design where one obtains a pair of measurements on each individual (e.g., employee, material, machine, etc.):

- one measurement corresponds to "Treatment 1"
- the other measurement corresponds to "Treatment 2"
- Clearly, the two samples are no longer independent. Each individual contributes a response to both samples.
- If possible, it is important to randomize the order in which treatments are assigned.
- This may eliminate "common patterns" that may be seen when always following, say, Treatment 1 with Treatment 2. In practice, the experimenter could use `R` to determine which treatment is applied first.

This type of design removes variation among the individuals. This allows you to compare the two treatments (e.g., before/after working environment) under more homogeneous conditions where only variation within individuals is present (that is, the variation arising from the difference in the two treatments).

Design	Sources of Variation
Two Independent Samples	among employees, within employees
Matched Pairs	within employees

Table 11.1.2: Ergonomics example. Sources of variation in the two independent sample and matched pairs designs.

When you remove extra variability, this enables you to compare the two experimental conditions (treatments) more precisely. This gives you a better chance of identifying a difference between the

treatment means if one really exists. In a design with two independent samples, the extra variation among individuals may prevent us from being able to identify this difference!

Remark: Sometimes this is not possible for practical or ethical concerns. For example, we couldn't design an experiment in which we randomly assign pregnant mothers to a "non-smoking" or "smoking" condition for their first pregnancy and then the other condition for their next pregnancy. This would be unethical for all we know about the negative effects of smoking, and it may not be practical because we cannot force participants to have another child for our experiment.

Implementation: Data from matched pairs experiments are analyzed by examining the difference in responses of the two treatments. Specifically, compute

$$D_j = X_{1j} - X_{2j},$$

for each individual $j = 1, 2, ..., n$. After doing this, we have essentially created a "one sample problem," where our data are now $D_1, D_2, ..., D_n$, the so-called data differences.

Individual	Before	After	Difference	Individual	Before	After	Difference
1	81.3	78.9	2.4	14	74.9	58.3	16.6
2	87.2	91.4	-4.2	15	75.8	62.5	13.3
3	86.1	78.3	7.8	16	72.6	70.2	2.4
4	82.2	78.3	3.9	17	80.8	58.7	22.1
5	90.8	84.4	6.4	18	66.5	66.6	-0.1
6	86.9	67.4	19.5	19	72.2	60.7	11.5
7	96.5	92.8	3.7	20	56.5	65.0	-8.5
8	73.0	69.9	3.1	21	82.4	73.7	8.7
9	84.2	63.8	20.4	22	88.8	80.4	8.4
10	74.5	69.7	4.8	23	80.0	78.8	1.2
11	72.0	68.4	3.6	24	91.1	81.8	9.3
12	73.8	71.8	2.0	25	97.5	91.6	5.9
13	74.2	58.3	15.9	26	70.0	74.2	-4.2

Table 11.1.3: Ergonomics data. Percentage of time arm elevation was less than 30 degrees. The data differences $D_j = X_{1j} - X_{2j}$ have been added.

The one sample $100(1-\alpha)$ percent confidence interval follows from the one sample case; e.g.,

$$\bar{D} \pm t_{n-1,1-\alpha/2}\frac{s_D}{\sqrt{n}},$$

where $\bar{D}$ and S_D are the sample mean and sample standard deviation of the differences, respectively, is an interval estimate for

$$\begin{aligned}\mu_D &= \mu_1 - \mu_2 \\ &= \text{population mean difference between the 2 treatments.}\end{aligned}$$

The parameter $\mu_D = \mu_1 - \mu_2$ describes the population mean difference for the two treatment groups. If the two population means are then same, then $\mu_D = 0$. Therefore, if the confidence interval for μ_D includes 0, this does not suggest that the two population means are different and if the confidence interval for μ_D does not include 0, this suggests that the two population means are different.

With the ergonomics data, we use `R` to construct a 95 percent confidence interval; note that this is a one-sample confidence interval calculated using the data differences. for $\mu_D = \mu_1 - \mu_2$:

```
> dat.erg<-data.frame(individual=1:26,
+     before=c(81.3,87.2,86.1,82.2,90.8,86.9,96.5,73,84.2,74.5,72,73.8,74.2,
+              74.9,75.8,72.6,80.8,66.5,72.2,56.5,82.4,88.8,80.0,91.1,97.5,70),
+     after=c(78.9,91.4,78.3,78.3,84.4,67.4,92.8,69.9,63.8,69.7,68.4,71.8,58.3,
+             58.3,62.5,70.2,58.7,66.6,60.7,65,73.7,80.4,78.8,81.8,91.6,74.2))
> dat.erg$D<-dat.erg$before-dat.erg$after
> t.test(x = dat.erg$D)

One Sample t-test

data:  dat.erg$D
t = 4.4525, df = 25, p-value = 0.000154
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
3.636034 9.894735
sample estimates:
mean of x
6.765385
```

We are 95 percent confident that the population mean difference $\mu_D = \mu_1 - \mu_2$ is between 3.6 and 9.9 percent. This interval does not include "0" and contains only positive values. Therefore, we have evidence that the population mean percentage of time that arm elevation is below 30 degrees is larger in the "before" condition than in the "after" condition. In other words, there is evidence that the "change" in work conditions implemented by the ergonomics team (in the 18-month interim) did reduce this population mean time.

Assumptions: In matched pairs experiments, the relevant assumptions are

1. The individuals sampled form a random sample.
2. The data differences $D_1, D_2, ..., D_n$ are normally distributed.

A Gaussian qq plot for the data differences is given in Figure 11.1.9. A very picky analyst might pick out the mild departure in the upper tail. However, remember that one-sample t confidence intervals (for means) are generally robust to these mild departures. Therefore, this slight departure (which I don't think is convincingly real) likely does not affect our conclusion.

```
> ggplot(data=dat.erg,aes(sample=D))+
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")
```

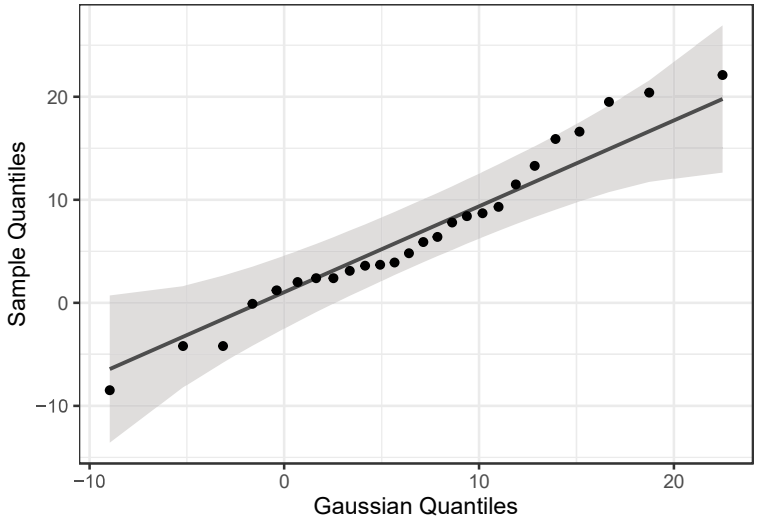

Figure 11.1.9: Gaussian qq plot for the ergonomics data. The observed data differences are plotted versus the theoretical quantiles from a Gaussian distribution. The line added passes through the first and third theoretical quartiles.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.1.10, created in R as follows.

```
> alpha<-0.05 #significance level
> test<-t.test(x = dat.erg$D)
> ggdat<-data.frame(t=seq(from=-4,to=4,by=0.01),
+                   f=dt(seq(from=-4,to=4,by=0.01),df=test$parameter))
> ggdat.highlight<-data.frame(x=qt(p=c(alpha/2,1-alpha/2),df=test$parameter),
+                             y=c(0,0))
> axis.labels<-round(c(qt(alpha/2,test$parameter),0,qt(1-alpha/2,test$parameter))*
+                      test$stderr + test$estimate,3)
> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t<=qt(alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=test$parameter)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   annotate("text",x=-3.1,y=0.05,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=3.1,y=0.05,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis=sec_axis(~.,breaks=c(qt(alpha/2,test$parameter),0,
+                      qt(1-alpha/2,test$parameter)),
+        labels = axis.labels,
+             name="Difference Percentage of Time Arm Elevation Was Less than 30 Degrees"))
```

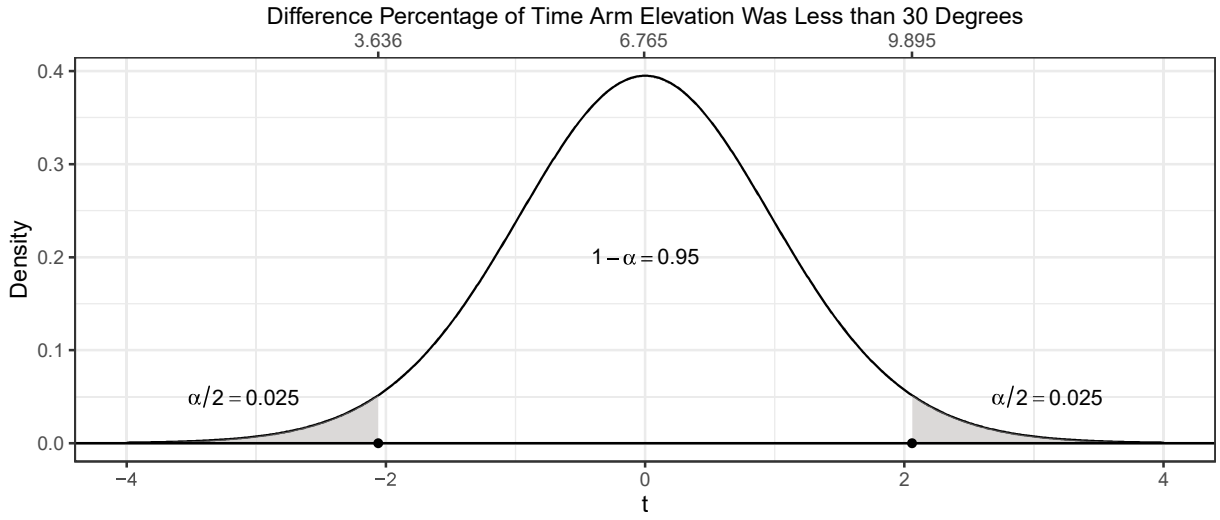

Figure 11.1.10: The approximate sampling distribution of $(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})$ with key values of the confidence interval for the population mean difference highlighted.

11.2 The Ratio of Two Population Variances

Recall that when we wrote a confidence interval for $\mu_1 - \mu_2$, the difference of the population means (with independent samples), we proposed two intervals:

- one interval that assumed $\sigma_1^2 = \sigma_2^2$
- one interval that assumed $\sigma_1^2 \neq \sigma_2^2$

We now propose a confidence interval procedure that can be used to determine which assumption is more appropriate

Suppose that we have two independent random samples:

$$\begin{aligned} &\text{Sample 1:} \qquad X_{11}, X_{12}, ..., X_{1n1} \sim \text{Gaussian}(\mu_1, \sigma_1^2) \\ &\text{Sample 1:} \qquad X_{11}, X_{12}, ..., X_{1n1} \sim \text{Gaussian}(\mu_2, \sigma_2^2) \end{aligned}$$

.

Our goal is to construct a $100(1-\alpha)$ percent confidence interval for the ratio of population variances σ_2^2/σ_1^2.

Under the setting described above,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2},$$

has an F distribution with (numerator) $n_1 - 1$ and (denominator) $n_2 - 1$ degrees of freedom.

Recall: The F pdf has the following characteristics:

- continuous, skewed right, and always positive

- indexed by two degree of freedom parameters v_1 and v_2; these are usually integers and are related to sample sizes
- the mean of an F distribution is close to 1 (regardless of the values of v_1 and v_2)
- The F pdf formula is complicated and is unnecessary for our purposes. R will compute F probabilities and quantiles from the F distribution.

We introduce new notation that identifies quantiles from an F distribution with

$$F_{n_1-1,n_2-1,1-\alpha/2} = \text{upper } \alpha/2 \text{ quantile from } F(n_1-1, n_2-1) \text{ PDF}$$
$$F_{n_1-1,n_2-1,\alpha/2} = \text{lower } \alpha/2 \text{ quantile from } F(n_1-1, n_2-1) \text{ PDF}$$

For example, if $n_1 = 11$, $n_2 = 12$ and $\alpha = 0.05$ then

$$F_{n_1-1,n_2-1,1-\alpha/2} = F_{10,10,0.975} \approx 3.72$$
$$F_{n_1-1,n_2-1,\alpha/2} = F_{10,10,0.025} \approx 0.27$$

These values are calculated in R as follows.

```
> qf(p=0.025,df1=10,df2=10)
[1] 0.2690492
> qf(p=0.975,df1=10,df2=10)
[1] 3.716792
```

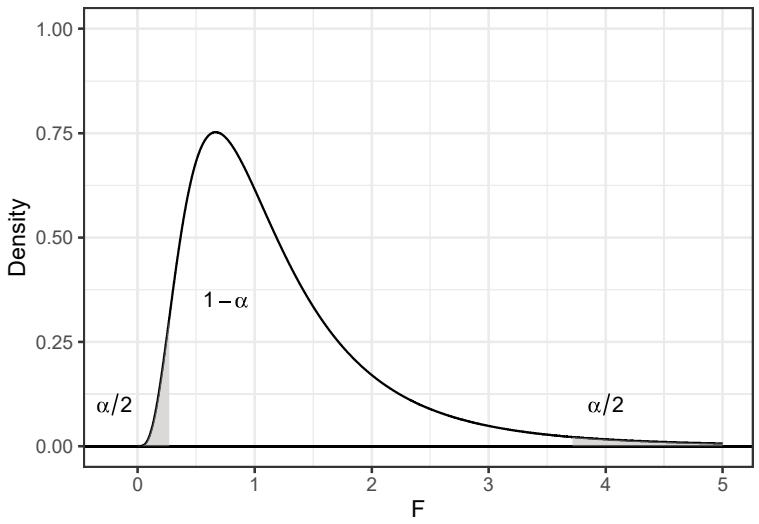

Figure 11.2.11: An F pdf with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $F_{n_1-1,n_2-1,1-\alpha/2}$ (upper) and $F_{n_1-1,n_2-1,\alpha/2}$ (lower), respectively.

Derivation: In general, for any value of $\alpha, 0 < \alpha < 1$, we can write

$$1-\alpha = P\left(F_{n_1-1,n_2-1,\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1,1-\alpha/2}\right)$$
$$= P\left(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2}\right).$$

This argument shows that

$$\left(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}, \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2}\right).$$

is a $100(1-\alpha)$ percent confidence interval for the ratio of the population variances σ_2^2/σ_1^2.

We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the ratio of the population variances σ_2^2/σ_1^2 is in this interval."

If the confidence interval for σ_2^2/σ_1^2 includes 1, this does not suggest that the population variances σ_1^2 and σ_2^2 are different. If the confidence interval for σ_2^2/σ_1^2 does not include 1, this suggests that the population variances σ_1^2 and σ_2^2 are different.

Therefore, this interval can be helpful in selecting an appropriate confidence interval for the difference of the population means $\mu_1-\mu_2$; i.e., between the one that assumes equal population variances and the one that does not. Of course, even if inference for population means is not the objective, this interval is still useful in its own right-it allows you to compare the variances of two populations. This is an important problem if one is concerned about variation.

We can also build a hypothesis test, under the setting described above, using this sampling distribution,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \left(\frac{S_1^2}{S_2^2}\right)\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \sim F(n_1-1, n_2-1),$$

an F distribution with (numerator) n_1-1 and (denominator) n_2-1 degrees of freedom.

Again, we choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = \sigma_0 & \text{or} & H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = \sigma_0 \quad \text{or} \quad \dfrac{\sigma_2^2}{\sigma_1^2} = \sigma_0 \\ H_a : \dfrac{\sigma_2^2}{\sigma_1^2} < \sigma_0 & & H_a : \dfrac{\sigma_2^2}{\sigma_1^2} > \sigma_0 \qquad H_a : \dfrac{\sigma_2^2}{\sigma_1^2} \neq \sigma_0. \end{array}$$

Often, we take $\sigma_0 = 1$

$$\begin{array}{lll} H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = 1 & \text{or} & H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = 1 \quad \text{or} \quad H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = 1 \\ H_a : \dfrac{\sigma_2^2}{\sigma_1^2} < 1 & & H_a : \dfrac{\sigma_2^2}{\sigma_1^2} > 1 \qquad H_a : \dfrac{\sigma_2^2}{\sigma_1^2} \neq 1, \end{array}$$

so that we can test $\sigma_1 > \sigma_2$, $\sigma_1 < \sigma_2$, or $\sigma_1 \neq \sigma_2$ (from right to left).

The test statistic for this test is

$$f = \left(\frac{S_1^2}{S_2^2}\right)\sigma_0 \sim F(n_1-1, n_2-1),$$

We succinctly summarize the five steps for the F Hypothesis Test.

Step One	$H_a : \frac{\sigma_2^2}{\sigma_1^2} < \sigma_0$	$H_a : \frac{\sigma_2^2}{\sigma_1^2} > \sigma_0$	$H_a : \frac{\sigma_2^2}{\sigma_1^2} \neq \sigma_0$
Step Two	Check Assumptions		
Step Three	$f^* = \left(\frac{S_1^2}{S_2^2}\right)\sigma_0$		
Step Four	$P(F_{n_1-1,n_2-1} < f^*)$	$P(F_{n_1-1,n_2-1} > f^*)$	$2P(F_{n_1-1,n_2-1} < -\lvert f^*\rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if f^* is in the rejection region		

Example 11.4. Two automated filling processes are used in the production of automobile paint. The target weight of each process is 128.0 fluid oz (1 gallon). There is little concern about the process population mean fill amounts (no complaints about under/overfilling on average). However, there is concern that the population variation levels between the two processes are different. To test this claim, industrial engineers took independent random samples of $n_1 = 24$ and $n_2 = 24$ gallons of paint and observed the fill amounts.

	127.75	127.87	127.86	127.92	128.03	127.94	127.91	128.10
Process 1:	128.01	128.11	127.79	127.93	127.89	127.96	127.80	127.94
	128.02	127.82	128.11	127.92	127.74	127.78	127.85	127.96
	127.90	127.90	127.74	127.93	127.62	127.76	127.63	127.93
Process 2:	127.86	127.73	127.82	127.84	128.06	127.88	127.85	127.60
	128.02	128.05	127.95	127.89	127.82	127.92	127.71	127.78

In order to visually assess the normality of the fill amounts from both processes, we use violin plots to display the data for each sample; see Figure 11.2.12. We see that there's little concern about the Gaussian assumption as any deviations are slight.

```
> dat.filling<-data.frame(process.1=c(127.75,127.87,127.86,127.92,128.03,127.94,127.91,128.10,
+                                     128.01,128.11,127.79,127.93,127.89,127.96,127.80,127.94,
+                                     128.02,127.82,128.11,127.92,127.74,127.78,127.85,127.96),
+                         process.2=c(127.90,127.90,127.74,127.93,127.62,127.76,127.63,127.93,
+                                     127.86,127.73,127.82,127.84,128.06,127.88,127.85,127.60,
+                                     128.02,128.05,127.95,127.89,127.82,127.92,127.71,127.78))
> ggdat<-data.frame(ounces=c(dat.filling$process.1,dat.filling$process.2),
+                   process=c(rep("Process 1",24),rep("Process 2",24)))
> ggplot(data=ggdat,aes(x=process, y=ounces))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("Process")+
+   ylab("Fluid Ounces")
```

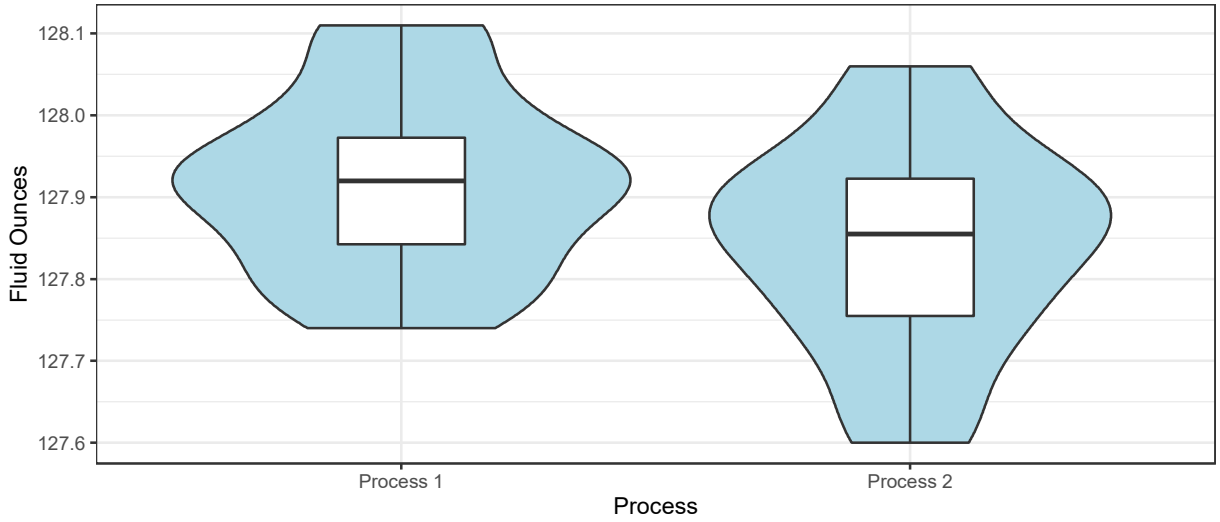

Figure 11.2.12: Violin plots of the paint fill data by process.

We can ask R to calculate the confidence interval for the ratio of two population variances σ_2^2/σ_1^2 directly.

```
> var.test(x=dat.filling$process.2, #top
+          y=dat.filling$process.1, #bottom
+          conf.level=0.95)

        F test to compare two variances

data:  dat.filling$process.2 and dat.filling$process.1
F = 1.3605, num df = 23, denom df = 23, p-value = 0.4662
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.5885236 3.1448830
sample estimates:
ratio of variances
1.360455
```

Gaussian qq plots for the two samples of white paper data are given in Figure 11.2.13. There is no cause to question the normality assumption.

```
> ggplot(data=ggdat,aes(sample=ounces))+
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")+
+   facet_grid(. ~ process)
```

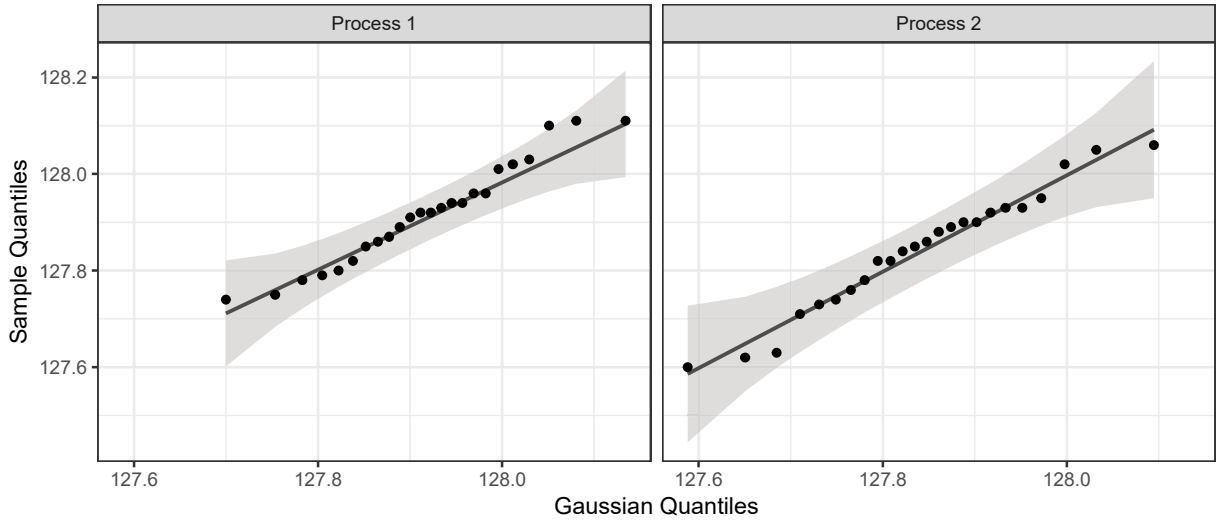

Figure 11.2.13: Gaussian qq plot for the paint fill data. The observed data differences are plotted versus the theoretical quantiles from a Gaussian distribution. The line added passes through the first and third theoretical quartiles.

We are 95 percent confident that the ratio of the population variances σ_2^2/σ_1^2 is between 0.5885 and 3.1449. Because this interval includes "1," we do not have evidence that the population variances σ_1^2 and σ_2^2 are different for the two processes.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.2.14, created in R as follows.

```
> alpha<-0.05 #significance level
> test<-var.test(x=dat.filling$process.2, #top
+                y=dat.filling$process.1, #bottom
+                conf.level=0.95)
> ggdat<-data.frame(f=seq(from=0,to=4,by=0.01),
+                   f1=df(seq(from=0,to=4,
+                        by=0.01),df1=test$parameter[1],
+                             df2=test$parameter[2]))
> ggdat.highlight<-data.frame(x=qf(p=c(alpha/2,1-alpha/2),df1=test$parameter[1],
+                             df2=test$parameter[2]),
+                             y=c(0,0))
> axis.labels<-round(c(qf(alpha/2,df1=test$parameter[1],df2=test$parameter[2]),
+                      qf(1-alpha/2,df1=test$parameter[1],
+                             df2=test$parameter[2]))*test$estimate,3)
> ggplot(data=ggdat,aes(x=f,y=f1))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,f<=qf(alpha/2,df1=test$parameter[1],df2=test$parameter[2])),
+                             aes(ymax=f1),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,f>=qf(1-alpha/2,df1=test$parameter[1],df2=test$parameter[2])),
+                             aes(ymax=f1),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("f")+
+   ylab("Density")+
+   annotate("text",x=0.05,y=0.075,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=3.1,y=0.075,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=1,y=0.2,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+                             breaks=c(qf(alpha/2,df1=test$parameter[1],df2=test$parameter[2]),
+                             qf(1-alpha/2,df1=test$parameter[1],df2=test$parameter[2])),
+                        labels = axis.labels,name="Ratio of Variances of Fill Processes"))
```

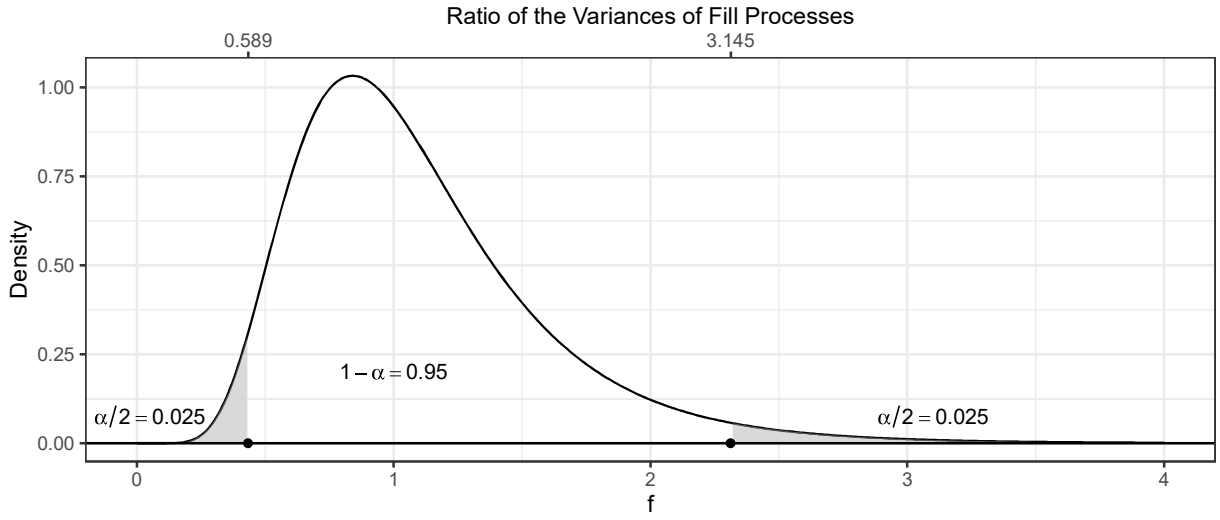

Figure 11.2.14: The approximate sampling distribution of S_2^2/S_1^2 with key values of the confidence interval for the population ratio of variances highlighted.

Warning: Like the χ^2 interval for single population variance σ^2, the two-sample F interval for the ratio of two population variances σ_2^2/σ_1^2 is not robust to normality departures. This is true because the sampling distribution

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

depends critically on the Gaussian distribution assumption for both populations. If either underlying population distribution is non-Gaussian (non-normal), then the confidence interval formula for σ_2^2/σ_1^2 is not to be used.

11.3 The Difference of Two Population Proportions

We now extend our confidence interval procedure for a single population proportion p to two populations. Define

$$p_1 = \text{population proportion in Population 1}$$
$$p_2 = \text{population proportion in Population 2.}$$

For example, we might want to compare the proportion of

- defective circuit boards for two different suppliers
- satisfied customers before and after a product design change (e.g., Facebook, etc.)
- on-time payments for two classes of customers
- HIV positives for individuals in two demographic classes.

Point estimators: We assume that there are two independent random samples of individuals, one sample from each population to be compared). Define

$$X_1 = \text{number of "successes" in Sample 1} \sim \text{binomial}(n_1, p_1)$$
$$X_2 = \text{number of "successes" in Sample 2} \sim \text{binomial}(n_2, p_2).$$

The point estimators for p_1 and p_2 are the sample proportions, defined by

$$\hat{p_1} = \frac{X_1}{n_1}$$
$$\hat{p_2} = \frac{X_2}{n_2}$$

We would like to construct a $100(1-\alpha)$ percent confidence interval for $p_1 - p_2$, the difference of two population proportions. We need the following sampling distribution result. When the sample sizes n_1 and n_2 are large,

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1).$$

If this sampling distribution holds approximately, then

$$(\hat{p_1} - \hat{p_2}) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for $p_1 - p_2$.

Note again the form of the interval:

$$\underbrace{\text{point estimate}}_{(\hat{p_1}-\hat{p_2})} \pm \underbrace{\text{quantile}}_{z_{1-\alpha/2}} \underbrace{\text{standard error}}_{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}.$$

We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the population proportion difference $p_1 - p_2$ is in this interval."

The value $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile from the Gaussian($\mu_z = 0$,$\sigma_z = 1$) distribution.

Note: For the Z sampling distribution to hold approximately (and therefore for the interval above to be useful), we need

- the two random samples to be independent
- the sample sizes n1 and n2 to be "large;" common rules of thumb are to require

$$n_i\hat{p_1} \geq 5$$
$$n_i(1-\hat{p_1}) \geq 5$$

for each sample i = 1, 2.

Under these conditions, the Central Limit Theorem should adequately approximate the true sampling distribution of Z, thereby making the confidence interval formula above approximately valid.

In two-sample situations, it is often of interest to see how the population proportions p1 and p2 compare. If the confidence interval for $p_1 - p_2$ includes 0, this does not suggest that the population proportions p_1 and p_2 are different. If the confidence interval for $p_1 - p_2$ does not include 0, this suggests that the population proportions p_1 and p_2 are different.

We can also build a hypothesis test, when the sample sizes n_1 and n_2 are large, using the sampling distribution:

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1).$$

Again, we choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : p_1 - p_2 = p_0 & \text{or} & H_0 : p_1 - p_2 = p_0 \quad \text{or} \quad H_0 : p_1 - p_2 = p_0 \\ H_a : p_1 - p_2 < p_0 & & H_a : p_1 - p_2 > p_0 \quad\quad H_a : p_1 - p_2 \neq p_0. \end{array}$$

Often, we take $p_0 = 0$

$$\begin{array}{lll} H_0 : p_1 - p_2 = 0 & \text{or} & H_0 : p_1 - p_2 = 0 \quad \text{or} \quad H_0 : p_1 - p_2 = 0 \\ H_a : p_1 - p_2 < 0 & & H_a : p_1 - p_2 > 0 \quad\quad H_a : p_1 - p_2 \neq 0. \end{array}$$

so that we can test $p_1 < p_2$, $p_1 > p_2$, or $p_1 \neq p_2$ (from right to left).

The test statistic for this test is

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_0)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1),$$

where p is the pooled proportion

$$p = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}.$$

We succinctly summarize the five steps for the z Hypothesis Test, noting many of the steps are the same as the one-sample version

Step One	$H_a : p_1 - p_2 < p_0$	$H_a : p_1 - p_2 > p_0$	$H_a : p_1 - p_2 \neq p_0$
Step Two	Check Assumptions		
Step Three	$z^* = \frac{(\hat{p_1} - \hat{p_2}) - (p_0)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $p = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$		
Step Four	$P(Z < z^*)$	$P(Z > z^*)$	$2P(Z < -\lvert z^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if z^* is in the rejection region		

Example 11.5. A large public health study was conducted to estimate the prevalence and to identify risk factors of hepatitis B virus (HBV) infection among Irish prisoners. Two independent samples of female ($n_1 = 82$) and male ($n_2 = 555$) prisoners were obtained from five prisons in Ireland:

- 18 out of 82 female prisoners were HBV-positive

- 28 out of 555 male prisoners were HBV-positive.

We want to find a 95 percent confidence interval for $p_1 - p_2$, the difference in the population proportions for the two genders (Female = 1; Male = 2). We can ask for this directly in R.

```
> prop.test(x=c(18,28),n = c(82,555),conf.level = 0.95)

2-sample test for equality of proportions with continuity correction

data:  c(18, 28) out of c(82, 555)
X-squared = 28.007, df = 1, p-value = 1.209e-07
alternative hypothesis: two.sided
95 percent confidence interval:
0.07064269 0.26748080
sample estimates:
prop 1     prop 2
0.21951220 0.05045045
```

We are 95 percent confident the difference of the population proportions $p_1 - p_2$ is between 0.078 and 0.260. This interval does not contain "0" and contains only positive values. This suggests that the population proportion of female prisoners who are HBV positive is larger than the corresponding male population proportion. The sample size conditions on the previous page are satisfied

Remark: When calculating this "by hand" we will notice small differences in our interval compared to the interval calculated by R. This is because R is applying the continuity correction as discussed in Chapter 9.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.3.15, created in R as follows.

```
> alpha<-0.05 #significance level
> test<-prop.test(x=c(18,28),n=c(82,555),conf.level=0.95,correct=FALSE)
> x1<-18
> n1<-82
> p.hat1<-x1/n1
> x2<-28
> n2<-555
> p.hat2<-x2/n2
> se<-sqrt(p.hat1*(1-p.hat1)/n1 + p.hat2*(1-p.hat2)/n2)
> axis.labels<-round(c(qnorm(alpha/2,mean=0,sd=1),0,qnorm(1-alpha/2,mean=0,sd=1))*
+                      se + (-diff(test$estimate)),3)
> ggdat<-data.frame(z=seq(from=-4,to=4,by=0.01),
+                   f=dnorm(seq(from=-4,to=4,by=0.01),mean=0,sd=1))
> #plot a point at the 0.025 and 0.975 quantiles
> ggdat.highlight<-data.frame(x=qnorm(p=c(alpha/2,1-alpha/2),mean=0,sd=1),
+                             y=c(0,0))
> ggplot(data=ggdat,aes(x=z,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,z<=qnorm(alpha/2,mean=0,sd=1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,z>=qnorm(1-alpha/2,mean=0,sd=1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("z")+
+   ylab("Density")+
+   annotate("text",x=-3.1,y=0.05,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=3.1,y=0.05,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~., breaks=c(qnorm(alpha/2,mean=0,sd=1),0,qnorm(1-alpha/2,mean=0,sd=1)),
+                      labels = axis.labels,name="Difference of the proportion of HBV-positive Prisoners"))
```

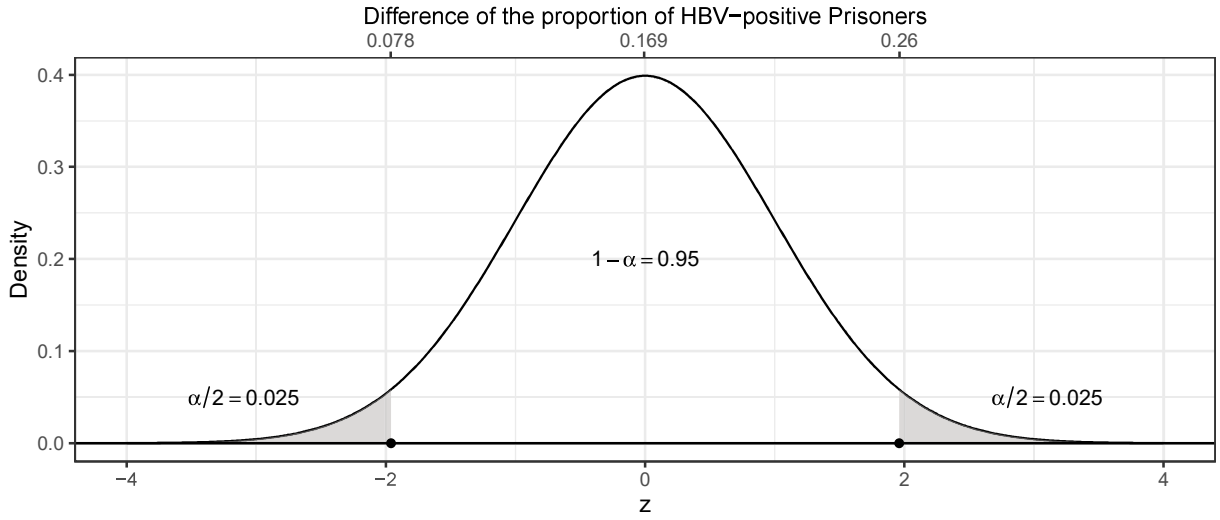

Figure 11.3.15: The approximate sampling distribution of $p_1 - p_2$ with key values of the confidence interval for the population proportion difference highlighted.

11.4 The Difference of Two Population Medians

Setting: Suppose that we have two independent random samples:

$$\begin{aligned} \text{Sample 1:} \quad & X_{11}, X_{12}, ..., X_{1n_1} \sim \text{Some distribtion } f_{x_{1i}} \\ \text{Sample 2:} \quad & X_{21}, X_{22}, ..., X_{2n_1} \sim \text{Some distribtion } f_{x_{2i}}. \end{aligned}$$

Constructing a confidence interval for $M_1 - M_2$ is more difficult theoretically, than the values discussed thus far. However, we can produce an approximate confidence interval using resampling just as we did in the one sample case.

This interval is always approximately valid, as long as

- the two samples are independent
- the two samples are representative of their respective populations.

Recall: A **percentile bootstrap confidence interval** uses estimates about the sampling distribution of $\hat{m}_1 - \hat{m}_2$, the sample difference of medians produced from R repeated random samples from the original data. We take a 95% confidence interval to be:

$$(m_i^*, m_j^*)$$

where

$$\begin{aligned} m_i^* \text{ is the } 2.5^{th} \text{ percentile of the } R \text{ median differences} \\ m_j^* \text{ is the } 97.5^{th} \text{ percentile of the } R \text{ medians differences.} \end{aligned}$$

There is also a hypothesis test for the difference of population medians called **Mood's Median Test**. We don't yet have the information needed to discuss the theoretical details of this test, which could be carried out in `R` using the `mood.medtest()` function from the "RVAideMemoire" package.

	$x_{j\cdot} < \hat{m}$	$x_{j\cdot} > \hat{m}$
Sample 1	O_{11}	O_{12}
Sample 2	O_{21}	O_{22}

Mood's Median test only provides insight for the two sided test where the medians are the same; e.g., $M_1 = M_2 = M$; e.g.,

$$H_0 : M_1 = M_2$$
$$H_a : M_1 \neq M_2.$$

The intuition behind this test is relatively simple and highly related to the Sign test in Chapter 9. The added complexity of considering two samples, instead of one, changes our simple count for the sign test to a table of counts as below, where $\hat{m}$ is the overall sample median. The idea is that if the two populations have the same median, then we would expect half of the observations from both samples to less than or greater than the median; i.e., we would expect $O_{11} \approx O_{12}$ and $O_{21} \approx O_{22}$. We can observe this by making a contingency table for our data as above, but the statistical mechanism for deciding when the data is "different enough," will be discussed in a later chapter.

Example 11.6. Consider data on housing values in suburbs of Boston. It is of interest whether or not houses that border the Charles River, which runs between Boston and Cambridge, are more expensive than those that do not. We load the data from the "MASS" packages as follows.

```
> library("MASS")
> data("Boston")
> Boston$chas<-ifelse(Boston$chas=="0","Doesn't Border Charles River","Borders Charles River")
```

We want to construct a 95 percent confidence interval for the population median difference $M_1 - M_2$, where the median price M_1 (M_2) corresponds to houses that are not along the Charles River (along the Charles River).

We consider the difference of medians due to the skew seen in the violin plots in Figure 11.4.16, which aren't as affected by extreme observations.

```
> ggplot(data=Boston,aes(x=chas, y=medv))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("")+
+   ylab("Median Value of Homes in $1000s")
```

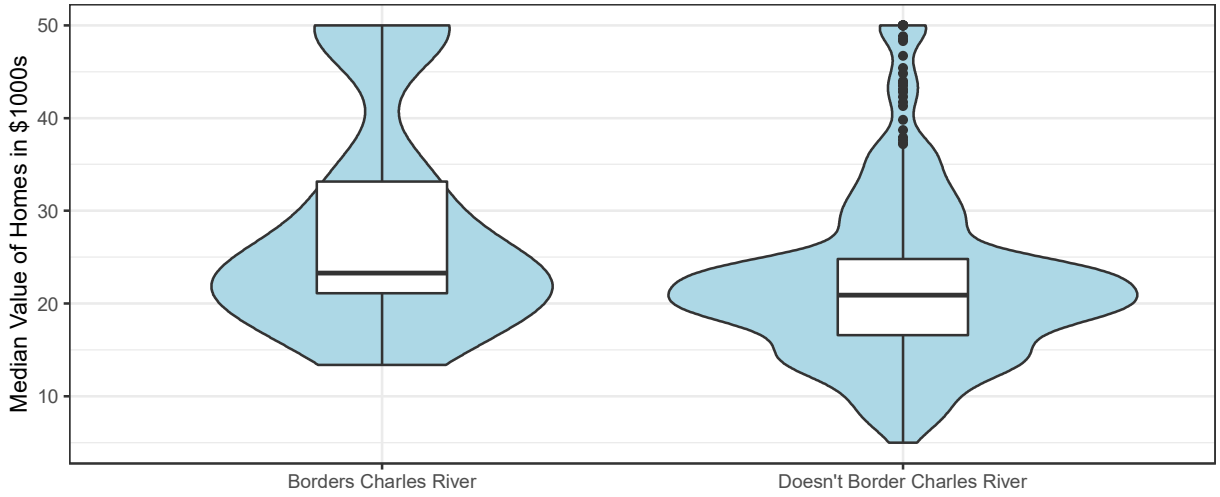

Figure 11.4.16: Violin plots of the median value of owner-occupied homes by whether or not the home borders the Charles River.

We can calculate the confidence interval "by hand" as we recall the bootstrapping technique.

```
> m.hats<-rep(NA,1000)
> for(i in 1:1000){
+   newsamp1<-sample(x=Boston$medv[which(Boston$chas=="Doesn't Border Charles River")],
+                    size=length(Boston$medv[which(Boston$chas=="Doesn't Border Charles River")]),
+                    replace = TRUE)
+   newsamp2<-sample(x=Boston$medv[which(Boston$chas=="Borders Charles River")],
+                    size=length(Boston$medv[which(Boston$chas=="Borders Charles River")]),
+                    replace = TRUE)
+   m.hats[i]<-median(newsamp1)-median(newsamp2)
+ }
> quantile(x=m.hats,probs=c(0.025,0.975))
2.5% 97.5%
-7.8  -0.5
```

We can use `R` to calculate the confidence interval directly.

```
> boot.median<-function(data,indices){
+   d<-data[indices]# allows boot to select sample
+   return(median(d))
+ }
> library("simpleboot")
> boot<-two.boot(sample1=Boston$medv[which(Boston$chas=="Doesn't Border Charles River")],
+                sample2=Boston$medv[which(Boston$chas=="Borders Charles River")],
+                FUN=boot.median, R=1000)
> library("boot")
> (ci<-boot.ci(boot.out=boot,conf=0.95,type="perc"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   (-7.795, -0.500 )
Calculations and Intervals on Original Scale
```

A 95 percent confidence interval for the population median difference $M_1 - M_2$ is

$$(-7.795, -0.5) \text{ 1000s of dollars.}$$

We are 95 percent confident that the population median difference $M_1 - M_2$ is between -$7,795 and -$500. This interval does not include "0" and contains only negative values. Therefore, we have evidence that the population median value of owner-occupied homes is smaller for homes not along the Charles River compared to homes along the Charles River.

We visualize the confidence interval with respect to the sampling distribution in Figure 11.4.17, created in R as follows.

```
> ggdat<-data.frame(m.hats=boot$t)
> lower<-ci$percent[4]
> upper<-ci$percent[5]
> #Start plot
> p<-ggplot(data=ggdat,aes(x=m.hats))+
+   geom_density(color="black")
> #Grab density data from the ggplot
> d <- data.frame(x=ggplot_build(p)$data[[1]]$x,
+                 f=ggplot_build(p)$data[[1]]$density)
> #Finish plot
> ggplot(data=d,aes(x=x,y=f))+
+   geom_line(color="black")+
+   geom_ribbon(data=subset(d,x<lower),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(d,x>upper),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Median Difference in Median Value of Homes")+
+   ylab("Density")+
+   annotate("text",x=-9,y=0.025,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=1,y=0.025,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=-3,y=0.10,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)
```

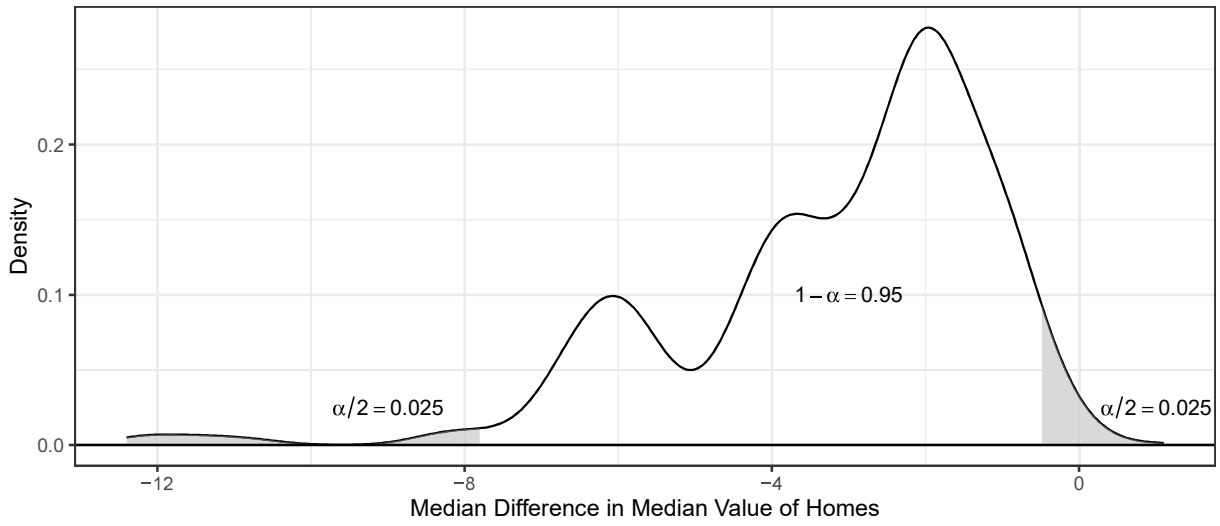

Figure 11.4.17: The approximate sampling distribution of $(\hat{m}_1 - \hat{p}_2)$ with key values of the confidence interval for the population median difference highlighted.

We might, instead, answer this question using a hypothesis test.

$$H0 : M_1 = M_2$$
$$Ha : M_1 \neq M_2$$

where the median price M_1 (M_2) corresponds to houses that are not along the Charles River (along the Charles River).

```
> medv.dat <- Boston$medv #median value data
> medv.group<- Boston$chas
> overall.median <- median(medv.dat) #overall median
> table(medv.group,medv.dat<overall.median)
medv.group                   FALSE TRUE
Borders Charles River           26    9
Doesn't Border Charles River   229  242
> prop.table(table(medv.group,medv.dat<overall.median),margin=1)

medv.group                        FALSE      TRUE
Borders Charles River        0.7428571 0.2571429
Doesn't Border Charles River 0.4861996 0.5138004
```

From the tables we created, we see that there's different behavior. The overall median, which is dominated by the larger sample of homes that don't boarder the Charles River, is a much better fit for that group, where roughly half of those observations are less than (51.4%)or greater than (38.6%) the overall median. For the homes that border the Charles river, we see that the overall median appears to be too small, as only 25.7% of observations are less than and 74.3% are greater than the overall median. This is reflected in a small p-value resulting from Mood's median test.

```
> library(RVAideMemoire)
> mood.medtest(medv~chas,data = Boston)

Mood's median test

data:  medv by chas
X-squared = 8.2719, df = 1, p-value = 0.004026
```

Here, we find evidence that the population median value of owner-occupied homes is different for homes not along the Charles River compared to homes along the Charles River ($\chi^2_1 = 8.2719$, $p = 0.0040$). Considering the violin plots in Figure 11.4.16 we see that this "difference" is that the median value of owner-occupied homes is higher for homes along the Charles River.