Chapter 9

One Sample Confidence Intervals

In this chapter, we discuss one-sample inference procedures for four population parameters:

- A population mean μ (Section 9.1)
- A population variance σ^2 (Section 9.2)
- A population proportion p (Section 9.3)
- A population median M (Section 9.4)

Remember that these are population-level quantities, so they are unknown. Our goal is to use sample information to estimate these quantities. The methods we discuss are outlined in Figure 9.0.1

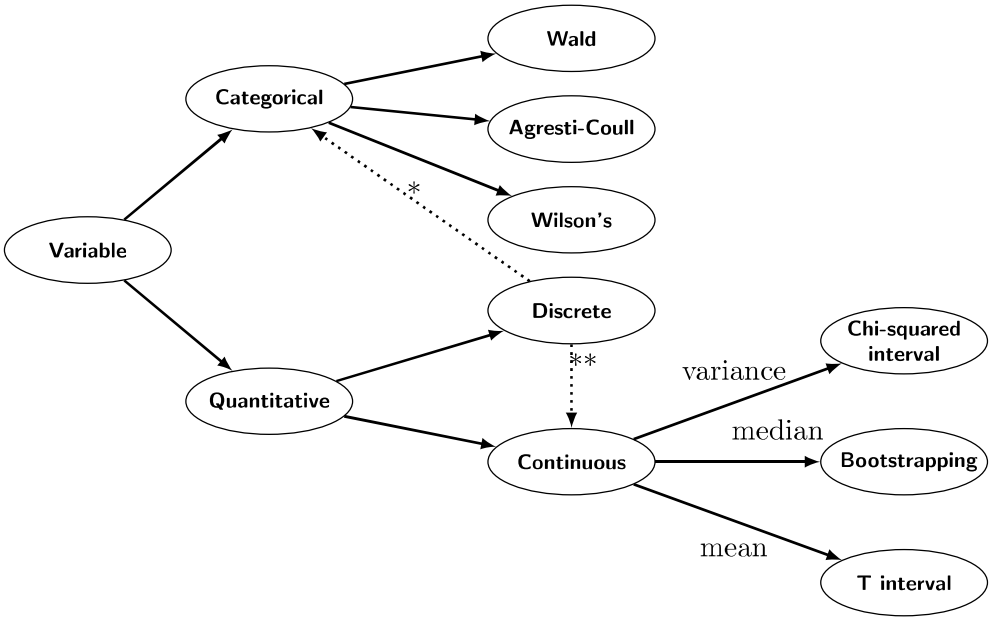

Figure 9.0.1: Flow chart for confidence interval approaches based on datatype. * and ** denote that we can treat a discrete variable as categorical when it has few observable values and as a continuous variable when it has very many observable values.

Relevance: To begin our discussion, suppose that we would like to estimate a population mean μ_x. To do so, suppose we have a random sample $X_1, X_2, ..., X_n$ from a population distribution (e.g., Gaussian, Poisson, etc.). Regardless of what the population distribution is, we know that $\bar{X}$ is an unbiased estimator for μ_x, that is,

$$E(\bar{X}) = \mu_x.$$

However, reporting $\bar{X}$ alone does not acknowledge that there is variability attached to this estimator. In Example 8.6, with the $n = 25$ measured pipes, we reported

$$\bar{x} \approx 1.299\text{in}$$

as an estimate of the population mean μ, but this does not account for the fact that

- the 25 pipes measured were drawn randomly from a population of all pipes, and
- different samples would give different sets of pipes (and different values of $\bar{X}$).

In other words, using $\bar{X}$ only ignores important information; namely, how variable the population of pipes is.

Remedy: To address this problem, we therefore pursue the topic of interval estimation, also known as confidence intervals. The main difference between a point estimate, like $\bar{x} \approx 1.299$, and an interval estimate is that

- a point estimate is a "one-shot guess" at the value of the parameter; this ignores the variability in the estimate.
- an interval estimate (i.e., confidence interval) is an interval of "plausible" values. It is formed by taking the point estimate and then adjusting it downwards and upwards to account for the point estimate's variability. The end result is an "interval estimate."

Terminology: A $100(1-\alpha)\%$ **confidence interval** is an interval that is guaranteed to capture the population parameter in $100(1-\alpha)\%$ of all samples. Usually, calculations are restricted to a 95% level of confidence, where $\alpha = 0.05$.

Note: $(1-\alpha) \times 100$ is a percentage between 0 and 100. We want $(1-\alpha) \times 100$ to be a large percentage; e.g., 80% ($\alpha = 0.20$), 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), and 99% ($\alpha = 0.01$) are common. Smaller choices of α lead to more confidence.

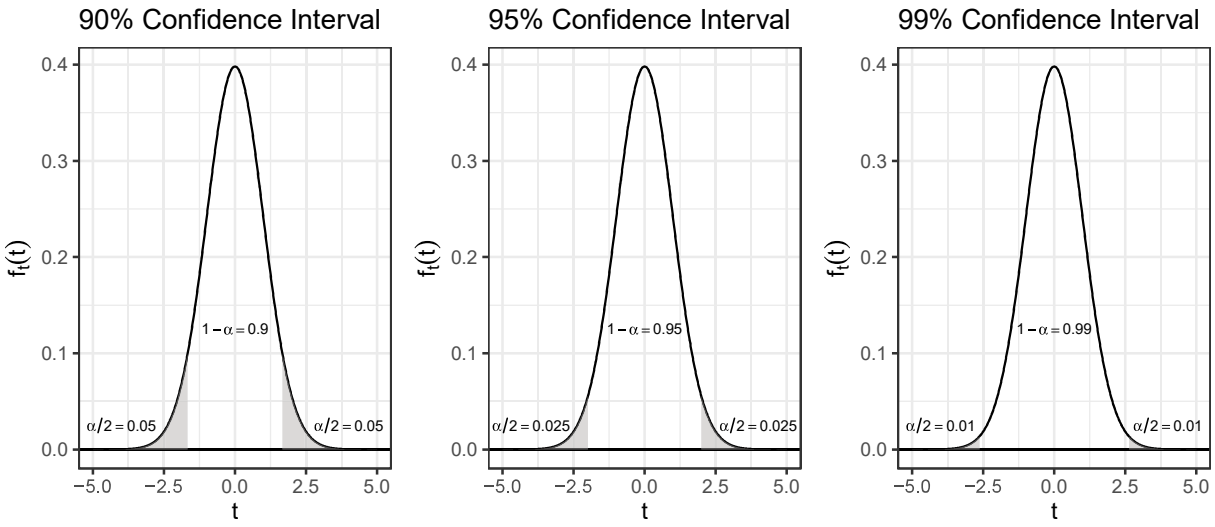

Figure 9.0.2: An illustration of 90%, 95% and 99% confidence intervals; higher confidence of containing the population mean requires a wider interval.

Remark: The more confidence we require, the wider the intervals must be to guarantee this. Think about trying to catch a fish with a net; if you want to me more confident that you'll catch a fish, you need a bigger net.

Figure 9 depicts what we mean by confidence. At 95% confidence, we expect 95% of confidence intervals to contain the true population mean. We can see that most of our intervals for μ_X contain the true value μ_X. However, it is possible for μ_X to be at the ends of the confidence interval or not contained in the interval at all. When we compute a confidence interval, we don't know where μ_X lies in the interval or whether it is contained in the interval at all.

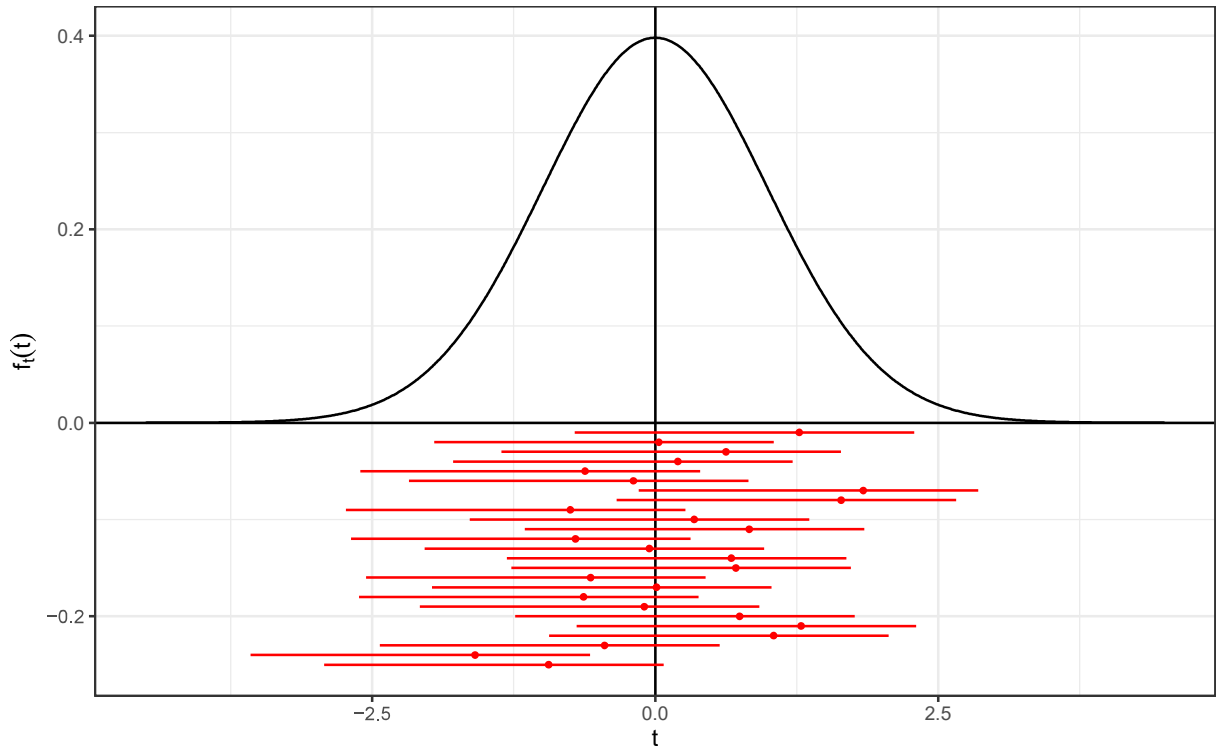

Figure 9.0.3: An illustration of possible confidence intervals based on an underlying sampling distribution.

9.1 Confidence Interval for a population mean μ_x

We start our discussion by revisiting Definition 8.7. Recall that if $X_1, X_2, ..., X_n$ is a random sample from a Gaussian(μ_x,σ_x^2) distribution, then the quantity

$$t = \frac{\bar{X} - \mu_x}{s_x/\sqrt{n}} \sim t(n-1),$$

a t distribution with $n-1$ degrees of freedom.

Goal: We will use this sampling distribution to create an interval estimate (i.e., a confidence interval) for the population mean μ_x.

Notation: We introduce new notation that identifies quantiles from a t distribution with $n-1$ degrees of freedom. Define

$$t_{n-1,1-\alpha/2} = \texttt{upper } \alpha/2 \texttt{ quantile from } t(n-1) \texttt{ PDF}$$
$$t_{n-1,\alpha/2} = \texttt{lower } \alpha/2 \texttt{ quantile from } t(n-1) \texttt{ PDF}$$

Because the $t(n-1)$ PDF is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 9.1.4, which is created with the following `R` code.

```
> alpha<-0.05
> ggdat<-data.frame(x=seq(from=-4,to=4,by=0.01),
```

```
+                     f=dt(seq(from=-4,to=4,by=0.01),df=25-1))
> ggdat.highlight<-data.frame(x=qt(p=c(alpha/2,1-alpha/2),df = 25-1),
+                     y=c(0,0))
> ggplot(data=ggdat,aes(x=x,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,x<=qt(alpha/2,df=25-1)),aes(ymax=f),ymin=0,
+               fill="grey",colour=NA,alpha=0.5)+ #alpha here refers to color transparancy
+   geom_ribbon(data=subset(ggdat,x>=qt(1-alpha/2,df=25-1)),aes(ymax=f),ymin=0,
+               fill="grey",colour=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   #deparse keeps bquote a string which is processed later
+   annotate("text",x=-3, y=0.05, label= deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=3, y=0.05, label= deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=0, y=0.2, label= deparse(bquote(1-alpha)),parse=TRUE,size=3.5)
```

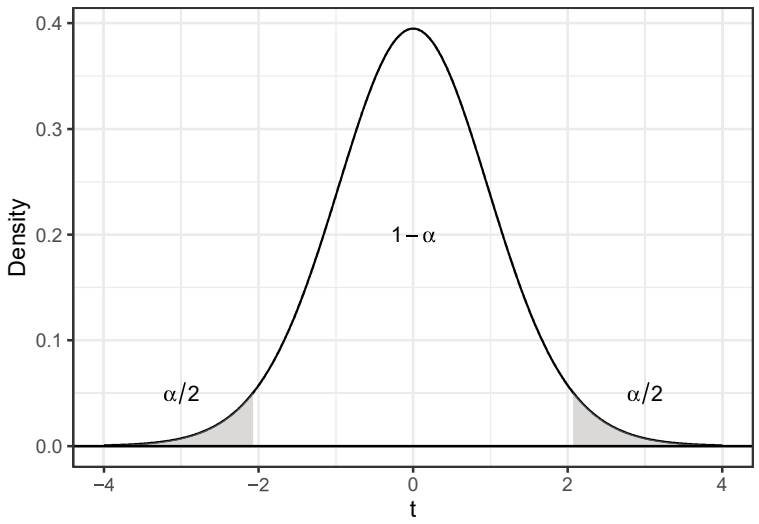

Figure 9.1.4: A t PDF with $n-1$ degrees of freedom, where $n = 25$. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $t_{n-1,1-\alpha/2}$ (upper) and $t_{n-1,\alpha/2}$ (lower), respectively.

Illustration: If $n = 25$ and $\alpha = 0.05$ then

$$t_{n-1,1-\alpha/2} = t_{24,0.025} \approx 2.06$$
$$t_{n-1,\alpha/2} = -t_{24,0.025} \approx -2.06$$

```
> qt(1-alpha/2,25-1) ## upper 0.025 quantile
[1] 2.063899
> qt(alpha/2,25-1) ## lower 0.025 quantile
[1] -2.063899
```

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

$$\begin{aligned}
1-\alpha &= P\left(t_{n-1,\alpha/2} < \frac{\bar{X}-\mu_x}{s_x/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) \\
&= P\left(t_{n-1,\alpha/2}\ \frac{s_x}{\sqrt{n}} < \bar{X}-\mu_x < t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}\right) && ([\times s_x/\sqrt{n}]) \\
&= P\left(-\bar{X}+t_{n-1,\alpha/2}\ \frac{s_x}{\sqrt{n}} < -\mu_x < -\bar{X}+t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}\right) && ([-\bar{x}]) \\
&= P\left(\bar{X}-t_{n-1,\alpha/2}\ \frac{s_x}{\sqrt{n}} > \mu_x > \bar{X}-t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}\right) && ([\times(-1)]) \\
&= P\left(\bar{X}-t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}} < \mu_x < \bar{X}-t_{n-1,\alpha/2}\ \frac{s_x}{\sqrt{n}}\right) && ([\textbf{Rewrite}]) \\
&= P\left(\bar{X}-t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}} < \mu_x < \bar{X}+t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}\right), && ([\textbf{Symmetry}])
\end{aligned}$$

due to symmetry, $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$.

We call

$$\left(\bar{X}-t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}, \bar{X}+t_{n-1,1-\alpha/2}\ \frac{s_x}{\sqrt{n}}\right)$$

a $100(1-\alpha)$ percent confidence interval for the population mean μ_x. This is written more succinctly as

$$\bar{X} \pm t_{n-1,1-\alpha/2}\frac{s_x}{\sqrt{n}}.$$

Discussion: Before we do an example, let's discuss relevant issues about this confidence interval.

Note the form of the interval:

$$\underbrace{\text{point estimate}}_{\bar{X}} \pm \underbrace{\text{quantile}}_{t_{n-1,1-\alpha/2}} \times \underbrace{\text{standard error}}_{s_x/\sqrt{n}}.$$

Many confidence intervals we will study follow this same general form.

Here is how we interpret this interval: We say

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

Unfortunately, the word "confident" does not mean "probability." The term "confidence" means that if we were able to sample from the population over and over again, each time computing a $100(1-\alpha)$ percent confidence interval for μ_x, then we would expect $100(1-\alpha)$ percent of the intervals we would compute to contain the population mean μ_x.

In other words, "confidence" refers to "long term behavior" of many intervals; not probability for a single interval. Because of this, we call $100(1-\alpha)$ the confidence level.

The length of the $100(1-\alpha)$ percent confidence interval

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

is equal to

$$2t_{n-1,1-\alpha/2}\frac{s_x}{\sqrt{n}}.$$

Therefore, other things being equal,

- the larger the sample size n, the smaller the interval length.
- the smaller the population variance σ_x^2, the smaller the interval length. Recall that S_x^2 is an unbiased estimator for σ_x^2.
- the larger the confidence level $100(1-\alpha)$, the larger the interval length.

Remark: Shorter confidence intervals are preferred. They are more informative. Lower confidence levels will produce shorter intervals; however, you pay a price. You have less confidence that your interval contains μ_x.

Question: Why don't we just be 100% confident?
Answer: The only way to be 100% confident is to say "we are 100% confident that the true population mean is between $-\infty$ and ∞." This isn't informative at all. When completing confidence intervals we're exchanging confidence for information; e.g., we're willing to be 95% confident (compared to 100%) to access a more informative interval.

Assumptions: The confidence interval

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

for the population mean μ_x was created based on the following assumptions:

1. $X_1, X_2, ..., X_n$ is a random sample
2. The population distribution is Gaussian(μ_x,σ_x^2).

For the confidence interval for the population mean μ_x to be meaningful, the random sample assumption must be satisfied. However, recall from the last chapter that the t sampling distribution result

$$t = \frac{\bar{X}-\mu_x}{S_x/\sqrt{n}} \sim t(n-1)$$

does still hold approximately even if the underlying population distribution is not perfectly Gaussian as long as the sample size is large. Therefore, the confidence interval, which was derived from this sampling distribution, is also "robust to normality departures."

This means that even if the population distribution is mildly non-normal, the confidence interval formula

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

can still be used to estimate the population mean μ_x. However, if there is strong evidence that the population distribution is grossly non-Gaussian, then you should exercise caution in using this confidence interval, especially when the sample size n is small. Recall that you can plot the data to assess normality.

Example 9.1. Acute exposure to cadmium produces respiratory distress and kidney and liver damage (and possibly death). For this reason, the level of airborne cadmium dust and cadmium oxide fume in the air, denoted by X (measured in milligrams of cadmium per m3 of air), is closely monitored. A random sample of $n = 35$ measurements from a large factory are given on the next page.

0.044	0.030	0.052	0.044	0.046	0.020	0.066
0.052	0.049	0.030	0.040	0.045	0.039	0.039
0.039	0.057	0.050	0.056	0.061	0.042	0.055
0.037	0.062	0.062	0.070	0.061	0.061	0.058
0.053	0.060	0.047	0.051	0.054	0.042	0.051

Based on past experience, engineers assume a normal population distribution (for the population of all cadmium measurements). Based on the data above, a 99 percent confidence interval for μ_x, the population mean level of airborne cadmium, is

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

We can use R to calculate the sample mean $\bar{x}$ and the sample standard deviation s_x:

```
> data.cadmium<-c(0.044,0.030,0.052,0.044,0.046,0.020,0.066,
+                 0.052,0.049,0.030,0.040,0.045,0.039,0.039,
+                 0.039,0.057,0.050,0.056,0.061,0.042,0.055,
+                 0.037,0.062,0.062,0.070,0.061,0.061,0.058,
+                 0.053,0.060,0.047,0.051,0.054,0.042,0.051)
> mean(data.cadmium)
[1] 0.04928571
> sd(data.cadmium)
[1] 0.0110894
```

For a 99 percent confidence level; i.e., with $\alpha = 0.01$, we use

$$t_{35-1,1-0.01/2} = -t_{35-1,1-0.01/2} \approx 2.728.$$

```
> alpha<-0.01
> qt(1-alpha/2,35-1)
[1] 2.728394
```

A 99 percent confidence interval for the population mean level of airborne cadmium μ_x is

$$\bar{X} \pm t_{n-1,1-\alpha/2}\frac{S_x}{\sqrt{n}}$$
$$0.049 \pm 2.728\left(\frac{0.011}{\sqrt{35}}\right)$$
$$(0.044, 0.054) \quad \text{mg/m}^3.$$

```
> mean(data.cadmium)+c(-1,1)*(sd(data.cadmium)/sqrt(35))*qt(1-alpha/2,35-1)
[1] 0.04417147 0.05439996
```

Interpretation: We are 99 percent confident that the population mean level of airborne cadmium μ is between 0.044 and 0.054 mg/m^3.

It is possible to implement the t interval procedure entirely in R using the `t.test()` function.

```
> t.test(x=data.cadmium,conf.level=0.99)

One Sample t-test

data:  data.cadmium
t = 26.293, df = 34, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
0.04417147 0.05439996
sample estimates:
mean of x
0.04928571
```

A histogram in the data in Figure 9.1.5 does not reveal any serious departures from the normality assumption. We can feel comfortable reporting

$$(0.044, 0.054) \text{ mg/m}^3$$

as a 99 percent confidence interval for the population mean cadmium level μ. We visualize the key values of the t interval in Figure 9.1.5 as follows.

```
> ggdat<-data.frame(cadmium=data.cadmium)
> g1<-ggplot(data=ggdat,aes(x=cadmium))+
+   geom_histogram(aes(y = ..density..),
+                  binwidth=density(ggdat$cadmium)$bw,
+                  fill="lightblue",color="black")+
+   geom_density(color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Level of Cadmium (mg"/m^3*")"))+
+   ylab("Density")
>
> length(data.cadmium) #gives n
[1] 35
> alpha<-0.01 #significance level
> ggdat<-data.frame(t=seq(from=-4,to=4,by=0.01),
+                   f=dt(seq(from=-4,to=4,by=0.01),df=35-1))
> #plot a point at the 0.005 and 0.995 quantiles
> ggdat.highlight<-data.frame(x=qt(p=c(alpha/2,1-alpha/2),df=35-1),
+                             y=c(0,0))
> #Save the cadmium levels corresponding to t=-4,-3,...,3,4
> axis.labels<-round(c(qt(alpha/2,35-1),0,qt(1-alpha/2,35-1))*
+      (sd(data.cadmium)/sqrt(35)) + mean(data.cadmium),3)
> g2<-ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
```

```
+   geom_ribbon(data=subset(ggdat,t<=qt(alpha/2,df=35-1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=35-1)),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("t")+
+   ylab("Density")+
+   annotate("text",x=-3,y=0.05,label= deparse(bquote(alpha/2==0.005)),parse=TRUE,size=3.5)+
+   annotate("text",x=3,y=0.05,label= deparse(bquote(alpha/2==0.005)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label= deparse(bquote(1-alpha==0.99)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~., breaks=c(qt(alpha/2,0,35-1),qt(1-alpha/2,35-1)),
+        labels = axis.labels,name=bquote("Level of Cadmium (mg"/m^3*")")))
> grid.arrange(g1,g2,ncol=2)
```

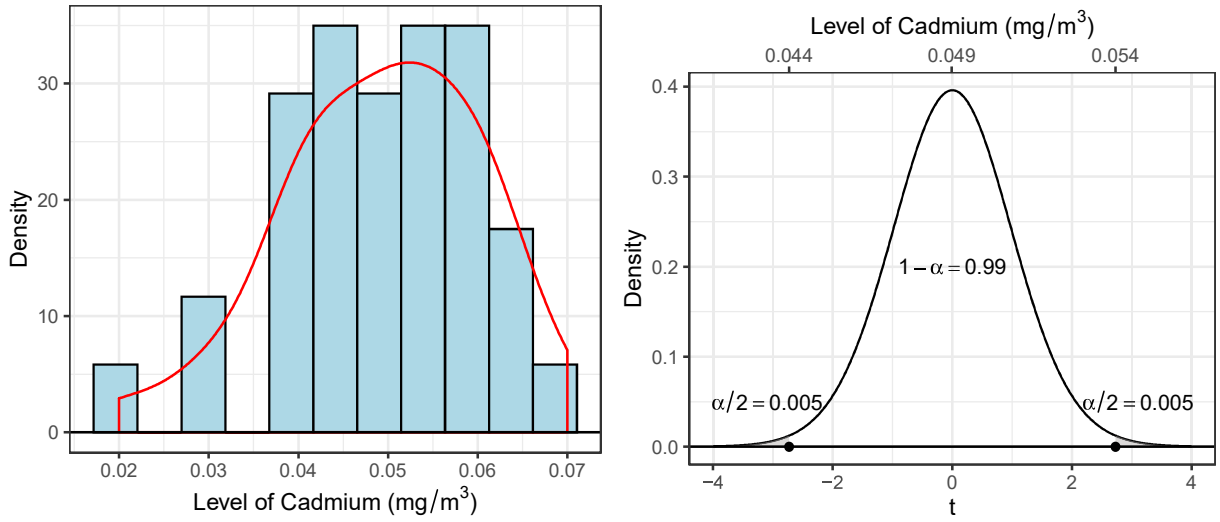

Figure 9.1.5: A histogram for the cadmium data (left) and the approximate sampling distribution of $\bar{X}$ (right) with key values of the confidence interval highlighted.

Remark: As we have just seen, statistical inference procedures are derived from specific assumptions. Going forward, it is important to know what these assumptions are, how critical they are, and how to check them.

9.2 Confidence Interval for a population variance σ_x^2

In many situations, we are concerned not with the mean of a population, but with the variance σ_x^2 instead. If the population variance σ_x^2 is excessively large, this could point to a potential problem with a manufacturing process, for example, where there is too much variation in the measurements produced. Elsewhere,

- in a laboratory setting, engineers might wish to estimate the variance σ_x^2 attached to a measurement system (e.g., scale, caliper, etc.).

- in field trials, agronomists are often interested in comparing the variability levels for different cultivars or genetically-altered varieties.
- in clinical trials, physicians are often concerned if there are substantial differences in the variation levels of patient responses at different clinic sites.

Result: If $X_1, X_2, ..., X_n$ is a random sample from a Gaussian(μ_x, σ_x^2) distribution, then the quantity

$$Q = \frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi^2(n-1),$$

a χ^2 distribution with $n-1$ degrees of freedom.

We will use this sampling distribution to create a confidence interval for the population variance σ^2.

Recall The χ^2 PDF has the following characteristics:

- It is continuous, skewed to the right, and the support is $\mathcal{X} = (0, \infty)$.
- It is indexed by a value v called the degrees of freedom. In practice, v is often an integer (related to the sample size).

Notation: We introduce new notation that identifies quantiles from a χ^2 distribution with $n-1$ degrees of freedom. Define

$$\chi^2_{n-1,1-\alpha/2} = \text{upper } \alpha/2 \text{ quantile from } \chi^2(n-1) \text{ PDF}$$
$$\chi^2_{n-1,\alpha/2} = \text{lower } \alpha/2 \text{ quantile from } \chi^2(n-1) \text{ PDF}$$

Remark: Unlike the $t(n-1)$ distribution, the $\chi^2(n-1)$ PDF is not symmetric about zero, thus the two quantiles are not equal in absolute value; see Figure 9.2.6, which is created with the following R code.

```
> alpha<-0.05
> ggdat<-data.frame(x=seq(from=0,to=30,by=0.01),
+                   f=dchisq(x=seq(from=0,to=30,by=0.01),df=11-1))
> ggdat.highlight<-data.frame(x=qchisq(p=c(alpha/2,1-alpha/2),df=11-1),
+                             y=c(0,0))
> ggplot(data=ggdat,aes(x=x,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,x<=qchisq(alpha/2,df=11-1)),aes(ymax=f),ymin=0,
+               fill="grey",colour=NA,alpha=0.5)+ #alpha here refers to color transparancy
+   geom_ribbon(data=subset(ggdat,x>=qchisq(1-alpha/2,df=11-1)),aes(ymax=f),ymin=0,
+               fill="grey",colour=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(chi^2))+
+   ylab("Density")+
+   annotate("text",x=0,y=0.01,label=deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=9,y=0.05,label=deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=25,y=0.01,label=deparse(bquote(1-alpha)),parse=TRUE,size=3.5)
```

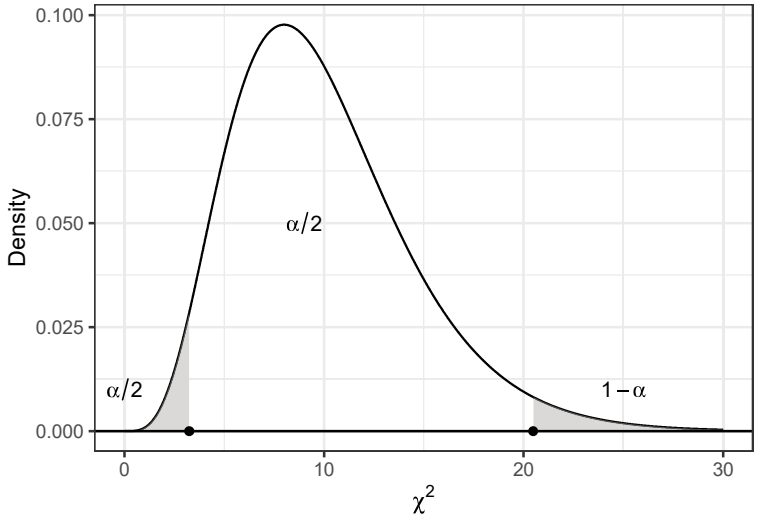

Figure 9.2.6: A χ^2 PDF with $n-1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $\chi^2_{n-1,1-\alpha/2}$ (upper) and $\chi^2_{n-1,\alpha/2}$ (lower), respectively.

Illustration: If $n = 11$ and $\alpha = 0.05$ then

$$\chi^2_{n-1,1-\alpha/2} \approx 20.48$$
$$\chi^2_{n-1,\alpha/2} \approx 3.25$$

```
> qchisq(1-alpha/2,11-1) ## upper 0.025 quantile
[1] 20.48318
> qchisq(alpha/2,11-1) ## lower 0.025 quantile
[1] 3.246973
```

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

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S_x^2}{\sigma_x^2} < \chi^2_{n-1,1-\alpha/2}\right) \\
&= P\left(\frac{1}{\chi^2_{n-1,\alpha/2}} > \frac{\sigma_x^2}{(n-1)S_x^2} > \frac{1}{\chi^2_{n-1,1-\alpha/2}}\right) && ([\textbf{Taking reciprocals}]) \\
&= P\left(\frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}} > \sigma_x^2 > \frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}\right) && ([\times(n-1)S_x^2]) \\
&= P\left(\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma_x^2 < \frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}\right) && ([\textbf{Rewrite}])
\end{aligned}$$

We call

$$\left(\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}\right)$$

a $100(1-\alpha)$ percent confidence interval for the population variance σ_x^2. We interpret the interval in the same way:

"We are $100(1-\alpha)$ percent confident that the population variance σ_x^2 is in this interval."

Note: A $100(1-\alpha)$ percent confidence interval for the population standard deviation σ_x arises from simply taking the square root of the endpoints of the σ_x^2 interval.

That is,

$$\left(\sqrt{\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for the population standard deviation σ_x.

In practice, this interval may be preferred over the σ_x^2 interval, because standard deviation is a measure of variability in terms of the original units (e.g., dollars, inches, days, etc.). The variance is measured in squared units (e.g., dollars^2, in^2, days^2, etc.) and is in general harder to interpret.

Assumptions: The confidence interval

$$\left(\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}\right).$$

for the population variance σ_x^2 was created based on the following assumptions:

1. $X_1, X_2, ..., X_n$ is a random sample
2. The population distribution is Gaussian(μ_x, σ_x^2).

For the confidence interval for the population variance σ_x^2 to be meaningful, the random sample assumption must be satisfied.

Warning: Unlike the t confidence interval for a population mean μ_x, the χ^2 interval for a population variance σ_x^2 is not robust to normality departures. This is true because the sampling distribution

$$Q = \frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2(n-1)$$

depends critically on the Gaussian(μ_x, σ_x^2) population distribution assumption. If the underlying population distribution is non-Gaussian, then the confidence interval formulas for σ_x^2 (and σ_x) are not to be used.

Example 9.2. Industrial engineers at IKEA observed a random sample of $n = 36$ rivet-head screws used in the Billy Bookcase system. The observed diameters of the top of the screws (measured in cm) are given below:

1.206	1.190	1.200	1.195	1.201	1.200	1.198	1.196	1.195	1.202	1.203	1.210
1.206	1.193	1.207	1.201	1.199	1.200	1.199	1.204	1.194	1.203	1.194	1.199
1.203	1.200	1.197	1.208	1.199	1.205	1.199	1.204	1.202	1.196	1.211	1.204

The IKEA manufactured specifications dictate that the population standard deviation diameter for these screws should be no larger than $\sigma_x = 0.003$. Otherwise, there is too much variability in the screws, which could lead to difficulty in construction and hence customer dissatisfaction. Based on the data above, a 95 percent confidence interval for the population variance σ_x^2 is

$$\left(\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}\right).$$

We can ask R to calculate this interval using the "EnvStats" package for R With this function, we calculated

```
> data.diameters<-c(1.206,1.190,1.200,1.195,1.201,1.200,1.198,1.196,1.195,1.202,
+                   1.203,1.210,1.206,1.193,1.207,1.201,1.199,1.200,1.199,1.204,
+                   1.194,1.203,1.194,1.199,1.203,1.200,1.197,1.208,1.199,1.205,
+                   1.199,1.204,1.202,1.196,1.211,1.204)
> varTest(x=data.diameters,conf.level=0.95)

Chi-Squared Test on Variance

data:  data.diameters
Chi-Squared = 0.00082231, df = 35, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
1.545590e-05 3.997717e-05
sample estimates:
variance
2.349444e-05
```

Interpretation: We are 95 percent confident that the population variance σ_x^2 for the screw diameters is between 0.0000155 and 0.0000400 cm^2.

A histogram in the data in Figure 9.2.7 does not reveal any serious departures from the normality assumption. We can feel comfortable reporting

$$(0.0000155, 0.0000400) \text{ cm}$$

as a 95 percent confidence interval for the population variance cadmium level σ_x^2. We visualize the key values of the χ^2 interval in Figure 9.2.7 as follows.

```
> ggdat<-data.frame(diameter=data.diameters)
> g1<-ggplot(data=ggdat,aes(x=diameter))+
+   geom_histogram(aes(y = ..density..),
+                  binwidth=density(ggdat$diameter)$bw,
+                  fill="lightblue",color="black")+
+   geom_density(color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Diameter (cm)")+
+   ylab("Density")
> length(data.diameters) #gives n
[1] 36
> alpha<-0.05 #significance level
> ggdat<-data.frame(chisq=seq(from=0,to=75,by=0.01),
+                   f=dchisq(seq(from=0,to=75,by=0.01),df=36-1))
> #plot a point at the 0.025 and 0.975 quantiles
> ggdat.highlight<-data.frame(x=qchisq(p=c(alpha/2,1-alpha/2),df=36-1),
+                             y=c(0,0))
> #Save the diameterss corresponding to the interval
```

```
> axis.labels<-round(c(qchisq(alpha/2,36-1),0,qchisq(1-alpha/2,36-1))*
+                      (sd(data.diameters)/sqrt(36))+mean(data.diameters),3)
> g2<-ggplot(data=ggdat,aes(x=chisq,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,chisq<=qchisq(alpha/2,df=36-1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,chisq>=qchisq(1-alpha/2,df=36-1)),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(bquote(chi^2))+
+   ylab("Density")+
+   annotate("text",x=8,y=0.005,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=65,y=0.005,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=34,y=0.025,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+       breaks=c(qchisq(alpha/2,36-1),(qchisq(alpha/2,36-1)+qchisq(1-alpha/2,36-1))/2,
+                qchisq(1-alpha/2,36-1)),
+          labels = axis.labels,name="Diameter (cm)"))
>
> grid.arrange(g1,g2,ncol=2)
```

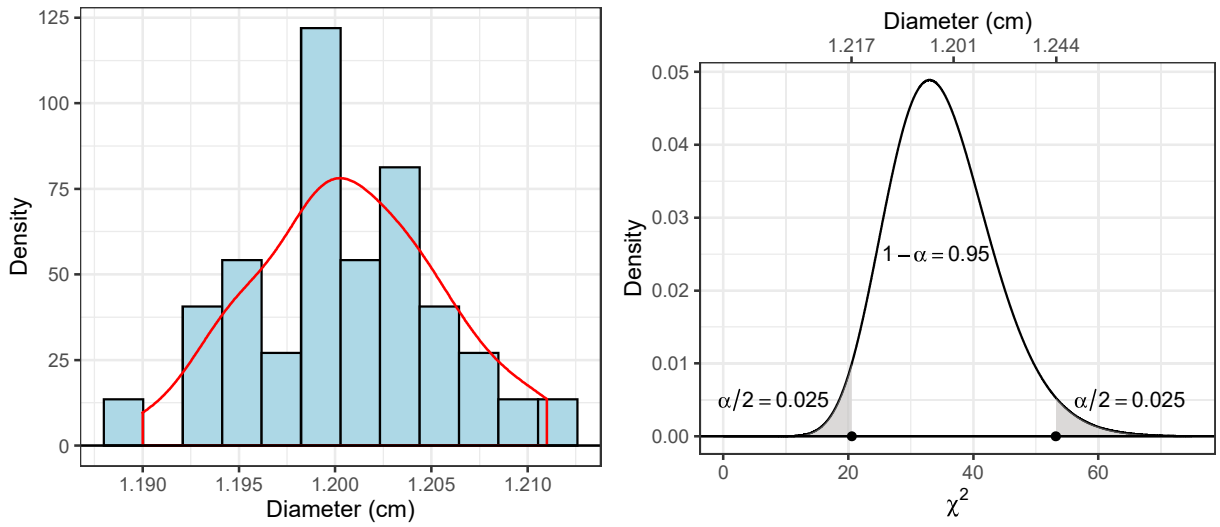

Figure 9.2.7: A histogram for the diameters of the top of the screws (left) and the approximate sampling distribution of S_x^2 (right) with key values of the confidence interval highlighted.

A 95 percent confidence interval for the population standard deviation σ_x, which is what we originally wanted, is

$$\left(\sqrt{\frac{(n-1)S_x^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)S_x^2}{\chi^2_{n-1,\alpha/2}}}\right).$$

We can calculate this in R by taking the square root of the confidence interval for the variance.

```
> sqrt(varTest(x=data.diameters,conf.level=0.95)$conf.int)
```

```
LCL         UCL
0.003931399 0.006322751
attr(,"conf.level")
[1] 0.95
```

Interpretation: We are 95 percent confident that the population standard deviation σ_x for the screw diameters is between 0.0039 and 0.0063 cm. This interval suggests that the population standard deviation is larger than 0.003 cm, which indicates that there is excessive variability in the diameters of the screws.

Remark: Fortunately, the histogram for the IKEA screw diameter data in Figure 9.2.7 shows that there is no cause for concern – the Gaussian assumption for these data is not in doubt. In the presence of non-normality, these confidence intervals may give results which are misleading, and hence potentially dangerous. Therefore, it is very important to check the normality assumption when you construct a confidence interval for a population variance σ_x^2

9.3 Confidence Interval for a population proportion p

We now switch gears and focus on a new population-level parameter: the population proportion p. This parameter is relevant when the characteristic we measure on each individual is binary (i.e., only 2 outcomes possible). Here are some examples:

p = proportion of defective circuit boards
p = proportion of customers who are "satisfied"
p = proportion of payments received on time
p = proportion of HIV positives in SC.

To start our discussion, we need to recall the Bernoulli trial assumptions for each individual in the sample:

1. each individual results in an event of interest ("success") or a ("failure"),
2. the individuals are independent, and
3. the probability of an event of interest ("success") p is the same for every individual.

In our examples above,

event of interest ("success") $\longrightarrow$ circuit board defective
event of interest ("success") $\longrightarrow$ customer satisfied
event of interest ("success") $\longrightarrow$ payment received on time
event of interest ("success") $\longrightarrow$ HIV positive individual.

Recall: If the individual success/failure statuses in the sample adhere to the Bernoulli trial assumptions, then

$$X = \text{the number of successes out of n sampled individuals}$$

follows a binomial distribution, that is,

$$X \sim \text{binomial}(n, p).$$

The statistical problem at hand is to use the information in X to estimate p.

It might seem perplexing that, in the end, we will have four methods - three formal, one informal - for completing the same task. The variety of approaches is thanks to the assortment of solutions for approximating the binomial proportion with a continuous distribution, e.g., the normal distribution; in reality, there are several more approaches.

We will cover three of these approaches that involve approximating the binomial in detail; e.g.,

- Wald
- Agresti-Coull
- Wilson's.

While an exact approach (Clopper-Pearson interval) that uses the binomial distribution exists, it still has issues dealing with the discrete nature of proportions and sometimes reports confidence intervals of higher confidence than asked for; e.g., 99% confidence when calculation is done for 95% confidence.

The Pearson-Clopper interval uses the exact sampling distribution for $\hat{p}$, the binomial. The reason we don't use this interval is because "approximate is better than approximate." This is because the binomial is discrete and thus we sometimes can't ask for the desired level of confidence. Recall the inverse CDF

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} | F(x) \geq \alpha\}.$$

When the random variable is discrete, we may not obtain exactly the α percentile which would mean our confidence could be off!

```
> qbinom(p=0.05,size=50,prob=0.15)
[1] 4
> pbinom(q=4,size=50,prob=0.15)
[1] 0.1121052
```

For this course, we will start with the Wald confidence interval, but we will also cover the Agresti-Coull and Wilson's confidence intervals which are better performing. Our main goal is to be clear that there exist many approaches to solving this problem and if you're interested, you can search for `R` packages and documentation that calculate and outline their differences.

Note: A natural point estimator for p, the population proportion, is

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X},$$

the sample proportion. This statistic is simply the proportion of "successes" in the sample out of n individuals.

Properties: Mathematical arguments can be used to show the following results:

$$E(\hat{p}) = p$$
$$se(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

The first result says that the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p. The second (standard error) result quantifies the precision of $\hat{p}$ as an estimator of p.

Recall that the sampling distribution of $\hat{p}$, which is critical if we are going to formalize statistical inference procedures for p, is conferred by the Central Limit Theorem, which says that

$$\hat{p} \sim \mathcal{AG}\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right),$$

when the sample size n is large. Standardizing $\hat{p}$, we get

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AG}(\mu_z = 0, \sigma_z = 1),$$

an approximate standard Gaussian distribution.

Notation: We introduce new notation that identifies quantiles from a Gaussian$(\mu_z = 0, \sigma_z = 1)$ distribution. Define

$$\begin{aligned} z_{1-\alpha/2} = &\text{ upper } \alpha/2 \text{ quantile from Gaussian}(\mu_z = 0, \sigma_z = 1) \text{ PDF} \\ z_{\alpha/2} = &\text{ lower } \alpha/2 \text{ quantile from Gaussian}(\mu_z = 0, \sigma_z = 1) \text{ PDF} \end{aligned}$$

Because the Gaussian$(\mu_z = 0, \sigma_z = 1)$ PDF is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 9.3.8, which is created with the following R code.

```
> alpha<-0.05
> 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))
> 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",colour=NA,alpha=0.5)+ #alpha here refers to color transparancy
+   geom_ribbon(data=subset(ggdat,z>=qnorm(1-alpha/2,mean=0,sd=1)),aes(ymax=f),ymin=0,
+               fill="grey",colour=NA,alpha=0.5)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("z")+
+   ylab("Density")+
+   #deparse keeps bquote a string which is processed later
+   annotate("text",x=-3,y=0.05,label=deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=3,y=0.05,label=deparse(bquote(alpha/2)),parse=TRUE,size=3.5)+
+   annotate("text",x=0,y=0.2,label=deparse(bquote(1-alpha)),parse=TRUE,size=3.5)
```

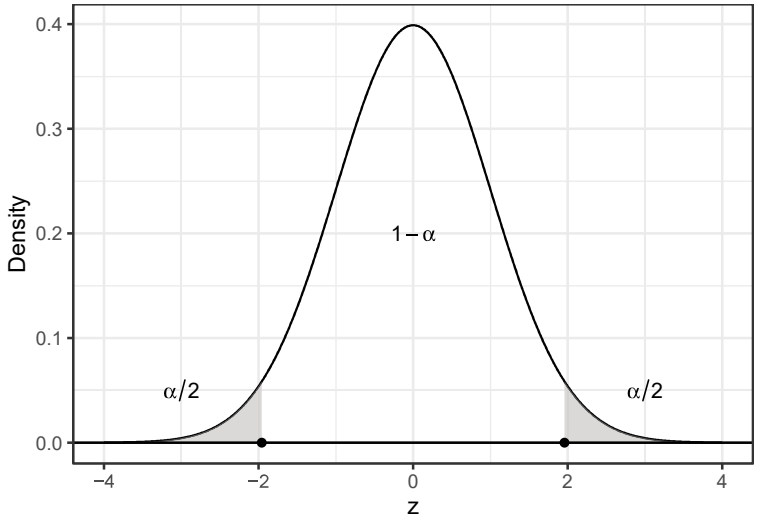

Figure 9.3.8: A Gaussian($\mu = 0$, $\sigma = 1$) PDF. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $z_{1-\alpha/2}$ (upper) and $z_{\alpha/2}$ (lower), respectively.

Illustration: If $\alpha = 0.05$ then

$$z_{1-\alpha/2} \approx 1.96$$
$$z_{\alpha/2} \approx -1.96$$

```
> qnorm(1-alpha/2,mean=0,sd=1) ## upper 0.025 quantile
[1] 1.959964
> qnorm(alpha/2,mean=0,sd=1) ## lower 0.025 quantile
[1] -1.959964
```

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

$$\begin{aligned}
1-\alpha &\approx P\left(z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\alpha/2}\right) \\
&= P\left(z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p}-p < z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) && ([\times\sqrt{\tfrac{p(1-p)}{n}}]) \\
&= P\left(-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} > p-\hat{p} > -z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) && ([\times(-1)]) \\
&= P\left(-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p-\hat{p} < -z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) && (\textbf{[Rewrite]}) \\
&= P\left(\hat{p}-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) && ([+\hat{p}]) \\
&= P\left(\hat{p}-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p}+z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right), && (\textbf{[Symmetry]})
\end{aligned}$$

due to symmetry, $z_{\alpha/2} = -z_{1-\alpha/2}$.

Remark: We only conduct inference about unknown parameters, so the fact that the calculation of this confidence interval involves p, an unknown parameter, is problematic. Instead, we go forward using $\hat{p}$ in place of p where necessary for calculation.

We call

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

a $100(1-\alpha)$ percent confidence interval for the population proportion p. This is written more succinctly as

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note the form of the interval:

$$\underbrace{\text{point estimate}}_{\hat{p}} \pm \underbrace{\text{quantile}}_{z_{1-\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}.$$

We interpret the interval in the same way:

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

Note: This interval should be used only when the sample size n is "large." Recall the rule of thumb to require

$$\begin{aligned} np &\geq 15 \\ n(1-p) &\geq 15. \end{aligned}$$

Again, this calculation involves p, an unknown parameter, which is problematic. Instead, we continue to use $\hat{p}$ in place of p where necessary for calculation and check

$$\begin{aligned} n\hat{p} &\geq 15 \\ n(1-\hat{p}) &\geq 15. \end{aligned}$$

Under these conditions, the CLT should adequately describe the sampling distribution of $\hat{p}$, thereby making the confidence interval formula above approximately valid.

Example 9.3. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected from the state and the tanks are inspected; 10 stations are found to have at least one leaking tank. A researcher requires 95 percent confidence interval for p, the population proportion of gasoline stations with at least one leaking tank.

In this situation, we interpret

- gasoline station = individual "trial"
- at least one leaking tank = "success"

- p = population proportion of stations with at least one leaking tank.

For 95 percent confidence, we need $z_{1-0.05/2} = z_{0.975} \approx 1.96$.

```
> qnorm(0.975,mean=0,sd=1) ## upper 0.025 quantile
[1] 1.959964
```

The sample proportion of stations with at least one leaking tank is

$$\hat{p} = \frac{10}{74} \approx 0.135.$$

Therefore, an approximate 95 percent confidence interval for p is

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$0.135 \pm 1.96\sqrt{\frac{0.135(1-0.135)}{74}}$$
$$(0.057, 0.213)$$

Interpretation: We are 95 percent confident that the population proportion of stations in Pennsylvania with at least one leaking tank is between 0.057 and 0.213.

CLT approximation check: We have

$$n\hat{p} = 74\left(\frac{10}{74}\right) = 10$$
$$n(1-\hat{p}) = 74\left(1 - \frac{10}{74}\right) = 64.$$

We would report this confidence interval with a warning that we may not feel comfortable because the assumptions for this interval are not quite met.

Note: It is possible to implement the interval procedure for the population proportion entirely in `R` using the `binom.confint()` function from the "binom" library.

```
> library("binom")
> y<-10
> n<-74
> binom.confint(x=y,n=74,conf.level=0.95)
          method  x  n      mean      lower     upper
1  agresti-coull 10 74 0.1351351 0.07314039 0.2331419
2     asymptotic 10 74 0.1351351 0.05724356 0.2130267
3          bayes 10 74 0.1400000 0.06638696 0.2191319
4        cloglog 10 74 0.1351351 0.06929029 0.2229842
5          exact 10 74 0.1351351 0.06675098 0.2345098
6          logit 10 74 0.1351351 0.07427809 0.2332879
7         probit 10 74 0.1351351 0.07201516 0.2284579
8        profile 10 74 0.1351351 0.07014077 0.2249370
9            lrt 10 74 0.1351351 0.07013459 0.2249355
10     prop.test 10 74 0.1351351 0.07017734 0.2390697
11        wilson 10 74 0.1351351 0.07509049 0.2311918
```

The interval discussed so far is the Wald (aysmptotic) interval; e.g.,

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

```
> p.hat<-y/n
> z.quantile<-qnorm(1-alpha/2,mean=0,sd=1)
> se<-sqrt(p.hat*(1-p.hat)/n)
> p.hat+c(-1,1)*z.quantile*se
[1] 0.05724356 0.21302671
```

The **Agresti-Coull** interval has relaxed assumptions and is acceptable to use when $n > 10$ and is calculated as

$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+z^2_{1-\alpha/2}}},$$

where

$$\tilde{p} = \frac{y + 0.5z^2_{1-\alpha/2}}{n+z^2_{1-\alpha/2}}.$$

```
> z.quantile<-qnorm(1-alpha/2,mean=0,sd=1)
> p.tilde<-((y+0.5*z.quantile^2)/(n+z.quantile^2))
> se<-sqrt(p.tilde*(1-p.tilde)/(n+z.quantile^2))
> p.tilde + c(-1,1)*z.quantile*se
[1] 0.07314039 0.23314189
```

The **Wilson interval** is acceptable to use when $n > 5$ and is calculated as

$$\tilde{p} \pm z_{1-\alpha/2}\frac{\sqrt{n\hat{p}(1-\hat{p}) + \frac{z^2_{1-\alpha/2}}{4}}}{n+z^2_{1-\alpha/2}}.$$

```
> p.hat<-y/n
> z.quantile<-qnorm(1-alpha/2,mean=0,sd=1)
> p.tilde<-((y+0.5*z.quantile^2)/(n+z.quantile^2))
> se<-sqrt(n*p.hat*(1-p.hat)+(z.quantile^2)/4)/(n+z.quantile^2)
> p.tilde + c(-1,1)*z.quantile*se
[1] 0.07509049 0.23119179
```

We will generally default to the Wilson interval because it is the least restrictive and tends to provide more reliable inference.

We can feel comfortable reporting

$$(0.075, 0.231)$$

as a 95 percent confidence interval for the population proportion of stations in Pennsylvania with at least one leaking tank. We visualize the key values of the z interval in Figure 9.3.9 as follows.

```
> alpha<-0.05 #significance level
> 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))
> #Save the proportions corresponding to the confidence interval
> p.hat<-y/n
> z.quantile<-qnorm(1-alpha/2,mean=0,sd=1)
> p.tilde<-((y+0.5*z.quantile^2)/(n+z.quantile^2))
> se<-sqrt(n*p.hat*(1-p.hat)+(z.quantile^2)/4)/(n+z.quantile^2)
> axis.labels<-round(c(qnorm(alpha/2,mean=0,sd=1),0,qnorm(1-alpha/2,mean=0,sd=1))*
+                      se+p.tilde,3)
> 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="Proportion of Stations with at Least One Leaking Tank"))
```

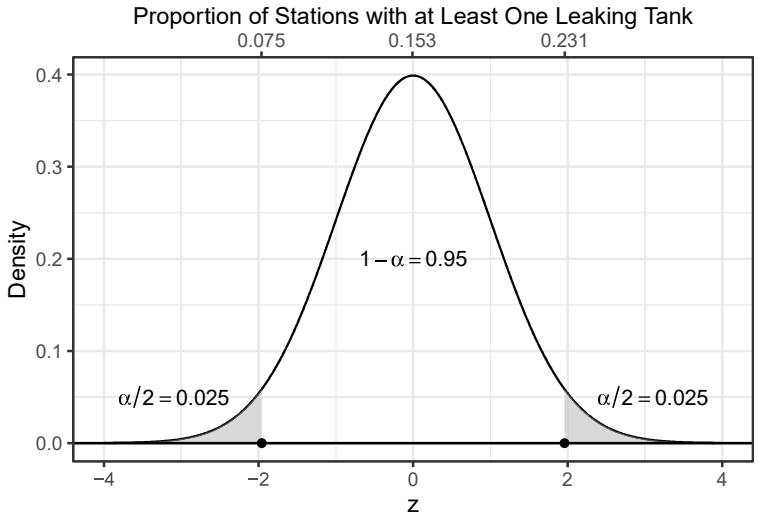

Figure 9.3.9: The approximate sampling distribution of $\hat{P}$ with key values of the Wilson confidence interval highlighted.

9.4 Confidence Interval for a population median M

In line with the logic behind Maslow's hammer, many researchers and scientists complete inference on the mean even on highly skewed data when it might be better summarized by the median. The reason for this is that use of the Central Limit Theorem makes inference about the mean relatively simple. Calculating the sampling distribution for something like the median, or any other statistic could require many more semesters of calculus and mathematical statistics.

Here, we started with the theory where we know the sampling distribution, or at least we know it is well approximated by the Gaussian distribution. It was from these assumptions that the t intervals were developed. William Gosset, employed by Guinness, worked to explore inference when the sample size was small - how is the margin of error affected by having a small sample, such as five, compared to a large sample of, say, ten thousand?

The opportunity to discover the t distribution originated when Gosset was asked to find exactly how many observations were needed to be confident that the "degrees saccharine" during production was within 0.50 degrees of the goal.

Gosset had a large set of observations from one batch, which meant by the law of large numbers he knew he could make a good estimate of the "degrees saccharine" for that batch by averaging the observations. He took repeated random samples of size $n = 2$, $n = 3$ and so on from the large set of observations and made inference based on the smaller sample sizes in order to answer the question at hand.

This repeated random sampling from the larger set of observations allowed Gosset to estimate the probability of getting within 0.50 degrees for the smaller sample sizes he cared about. He noted that the chance of being within 0.50 degrees was quite high even for just a handful of observations and as the number of observations grew the probability approached one.

We can thank Guinness for sending Gosset to work with Karl Pearson, another famous mathematician, in addition to making a delightful Irish stout. It was in Pearson's lab that Gosset had formalized his approximation of errors which we now know as Student's t distribution. The reason it is published under the name Student is because Guinness did not want to let competitors gain access to their methodology and increased accuracy.

Definition 9.4. What William Gosset used to approximate errors when using small sample sizes lays the groundwork for **bootstrapping**. Bootstrapping is a process that uses repeated random samples from a know population (which is a representative sample of the population of interest) to estimate variability, and perhaps bias, of the sampling distribution. This gives us a methodology to complete inference without making convenient assumptions that ignores the possibility of closed form mathematical solutions like we've used so far.

The following steps need to be completed to answer a question about a population that we can't measure. We will use the "boot" package (Canty and Ripley, 2019) in `R` to calculate bootstrap confidence intervals.

1. Take a random sample from the population and ask the question.
2. Decide that you are unhappy with restrictive assumptions about the sampling distribution for the statistic being used for inference.
3. Employ resampling to observe the behavior of a statistic.

4. Use the information from the repeated random samples to discuss variability.

Assumptions:

1. the original sample is representative of the population of interest
2. the sample size is reasonably large.

Remark: Bootstrapping uses resampling to create an approximating sampling distribution from which to calculate a confidence interval; e.g., it is how we will calculate confidence intervals for statistics for which we don't know the sampling distribution.

We want to ask a question about a population parameter, but we can't. Usually we leverage the sampling distribution, so far via Central Limit Theorem, to make inference about the population but sometimes we don't have insight about the sampling distribution; this can happen when Central Limit Theorem or some other assumption doesn't apply.

One way that we can learn about "what's going on" in the population is to take samples from the population again and again, ask the question, and see how variable the sample answers tended to be. This is what we might do in a simulation study. Since this isn't possible in practice, we can either make some assumptions about the shape of the population (like Central Limit Theorem), or we can use the information in the sample we have, which is hopefully representative of the sample.

We extract information about how the answer to our question might vary depending on which particular sample we happened to get by repeatedly generating samples of the same size as our sample from the sample we have and asking the same question of those samples. The idea is that we're trying to see "what's going on" in the population through resampling our sample (which is available) with replacement instead of the population (which isn't available).

This is a reasonable approach, because our sample is the best (and only) information we have about "what's going on" in the population. Also, if the sample we have obtained is a good proxy for our population, then the sample will look quite like the population, and finally the resamples (sampled randomly with replacement) will look quite like random samples from the population.

An Illustration: We generally would not use the bootstrap technique for a small sample, but let us do so only once to illustrate the technique and introduce the notation.

A random sample of the population of $n = 10$ U.S. cities in 1920 (in thousands of people) is retrieved in R using the "boot" package (Canty and Ripley, 2019).

```
> library("boot")
> data(city)
> citydat<-city$u  ##save data to object
> citydat
[1] 138 93 61 179 48 37 29 23 30 2
```

This data is right skewed, so we might be interested in the population median (M) city population of U.S. cities in 1920. The sample median ($\widehat{m}$) is calculated in R.

```
> median(citydat)
[1] 42.5
```

We can take repeated random samples of size n with replacement in R as well.

```
> sample(x=citydat,size=10,replace=TRUE)
[1] 138 29 179 179 93 138 30 93 61 2
> sample(x=citydat,size=10,replace=TRUE)
[1] 138 61 29 179 30 61 179 37 30 37
> sample(x=citydat,size=10,replace=TRUE)
[1] 37 37 2 2 2 29 179 29 93 29
> sample(x=citydat,size=10,replace=TRUE)
[1] 30 37 61 93 48 2 48 93 138 37
> sample(x=citydat,size=10,replace=TRUE)
[1] 23 179 61 48 23 138 179 48 23 61
```

As we can see, in some bootstrap samples observations were sampled once, others more than once and others not at all. The resampling process attempts to simulate running many experiments - this is why our original sample must be representative.

Here, the median for each random sample is different, e.g. for the random samples above we have medians of m_1^* =93, m_2^* =49, m_3^* =29, m_4^* =48 and m_5^* =54.5. Our goal is to extract information from the resamples for the population median.

Definition 9.5. A **percentile bootstrap confidence interval** uses estimates about the sampling distribution of M^*, the sample medians produced from R repeated random samples from the original data. We take a 95% confidence interval to be:

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

where

$$m_i^* \text{ is the } 2.5^{th} \text{ percentile of the } R \text{ medians}$$
$$m_j^* \text{ is the } 97.5^{th} \text{ percentile of the } R \text{ medians.}$$

This type of confidence interval can be found in R.

```
> M<-median(citydat)
> R<-1000 #1000 bootstrap samples
> city.boot = c()
> for(i in 1:R){
+   city.boot<-c(city.boot,median(sample(x = citydat, size=10,replace = TRUE)))
+ }
> quantile(x=city.boot,probs=0.025)
[1] 29
> quantile(x=city.boot,probs=0.975)
[1] 99.9
```

Here we say that we are 95% confident that the true population median is between 29 and 99.9 thousand people.

Note: We can also ask R directly for the bootstrapping confidence interval as follows.

```
> boot.median<-function(data,indices){
+   d<-data[indices]# allows boot to select sample
+   return(median(d))
+ }
> cityboot<-boot(citydat,R = 1000,statistic = boot.median)
> boot.ci(boot.out=cityboot,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

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

Intervals :
Level     Percentile
95%   (26.0, 99.5 )
Calculations and Intervals on Original Scale
```

Here we say that we are 95% confident that the true population median is between 26.0 and 99.5 thousand people. Note that the difference in the two intervals is only due to random sampling differences.

Resampling Approximation Check: As long as our sample of cities is representative of all cities, our assumptions are met. This may involve a conversation with the Researcher as to what population this sample is representative of. Checking the information provided through the `R` documentation using `?city`, we see that this is a random sample of 10 U.S. cities in 1920 from the 196 largest cities.

With this information, we would say we are 95% confident that the true median city population for the largest 196 cities in the U.S. is between 26.0 and 99.5 thousand people. This is an inference based on observing 10 of the 196 cities.

We visualize this interval by plotting the nonparametric kernel density estimate and shading the lower and upper $\alpha/2$ quantiles; see Figure 9.4.10, which is created with the following `R` code.

```
> ggdat<-data.frame(population=citydat)
> g1<-ggplot(data=ggdat,aes(x=population))+
+   geom_histogram(aes(y = ..density..), #plots the density
+                  binwidth=density(ggdat$population)$bw, #sets bin width
+                  fill="lightblue",color="black")+ #color the histogram
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("City Population (in 1000's)")+
+   ylab("Density")+
+   xlim(-40,200)
> ggdat<-data.frame(m.hats=cityboot$t)
> lower<-boot.ci(boot.out=cityboot,conf=0.95,type="perc")$percent[4]
> upper<-boot.ci(boot.out=cityboot,conf=0.95,type="perc")$percent[5]
> #Start second 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
> g2<-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(bquote("Median City Population"~(hat(m))))+
+   ylab("Density")+
+   xlim(-40,200)
+   annotate("text",x=-15,y=0.002,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=155,y=0.002,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=52,y=0.003,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)
> grid.arrange(g1,g2,ncol=2)
```

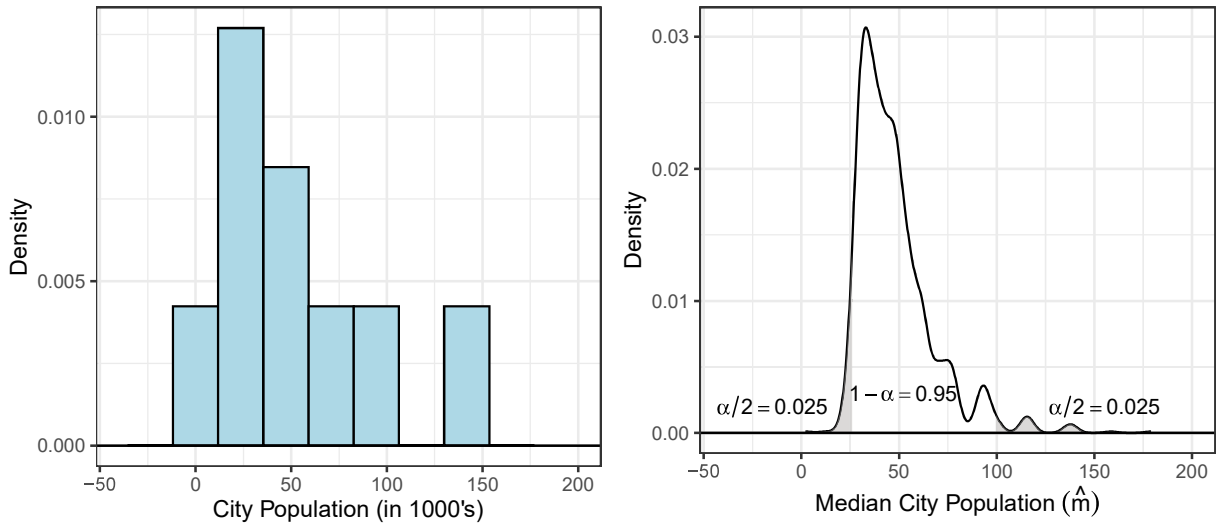

Figure 9.4.10: The histogram of the $n = 49$ city populations (in 1000s) (left) and the approximate sampling distribution of $\hat{M}$ with key values of the bootstrapping confidence interval for the population median highlighted.

Example 9.6. Arsenic is a chemical element (As) found naturally in ground water. Excessive levels may result from contamination caused by hazardous waste or industries that make or use arsenic. A random sample of $n = 102$ water wells in Texas from 1976 as reported by Nicholas et al. (1976).

As mentioned as motivation for bootstrap confidence intervals, we might be more interested in the population median instead of the mean. The histogram in Figure 9.4.11 shows the population distribution might be right skewed, thus motivating our choice to make inference about the population median.

Since the Central Limit theorem only applies to the sampling distribution of the sample mean and we don't yet know the advanced techniques required to mathematically calculate the distribution of

the sample median we can use bootstrapping to approximate the sampling distribution, very similar to what William Gosset did when approximating what would eventually become the t distribution.

We can calculate the percentile bootstrap confidence interval in R.

```
> data.arsenic<-c(17.6,10.4,13.5,4,19.9,16,12,12.2,11.4,12.7,3,10.3,21.4,
+                 19.4,9,6.5,10.1,8.7,9.7,6.4,9.7,63,15.5,10.7,18.2,7.5,
+                 6.1,6.7,6.9,0.8,73.5,12,28,12.6,9.4,6.2,15.3,7.3,10.7,
+                 15.9,5.8,1,8.6,1.3,13.7,2.8,2.4,1.4,2.9,13.1,15.3,9.2,
+                 11.7,4.5,1,1.2,0.8,1,2.4,4.4,2.2,2.9,3.6,2.5,1.8,5.9,2.8,
+                 1.7,4.6,5.4,3,3.1,1.3,2.6,1.4,2.3,1,5.4,1.8,2.6,3.4,1.4,
+                 10.7,18.2,7.7,6.5,12.2,10.1,6.4,10.7,6.1,0.8,12,28.1,9.4,
+                 6.2,7.3,9.7,62.1,15.5,6.4,9.5)
> arsenicboot<-boot(data.arsenic,R=R,statistic=boot.median)
> boot.ci(boot.out=arsenicboot,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

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

Intervals :
Level     Percentile
95%   ( 6.150,  9.449 )
Calculations and Intervals on Original Scale
```

We say that we are 95% confidence that the true population median arsenic concentration was between 6.150 and 9.449 ppb.

We visualize this interval by plotting the nonparametric kernel density estimate and shading the lower and upper $\alpha/2$ quantiles; see Figure 9.4.11, which is created with the following R code.

```
> ggdat<-data.frame(arsenic=data.arsenic)
> g1<-ggplot(data=ggdat,aes(x=arsenic))+
+   geom_histogram(aes(y = ..density..), #plots the density
+                  binwidth=density(ggdat$arsenic)$bw, #sets bin width
+                  fill="lightblue",color="black")+ #color the histogram
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Arsenic Concentration (ppb)")+
+   ylab("Density")
> ggdat<-data.frame(m.hats=arsenicboot$t)
> lower<-boot.ci(boot.out=arsenicboot,conf=0.95,type="perc")$percent[4]
> upper<-boot.ci(boot.out=arsenicboot,conf=0.95,type="perc")$percent[5]
> #Start plot
> p<-ggplot(data=ggdat,aes(x=m.hats))+
+   geom_density(color="black")
> #Grab density data from the ggplot
> d <- data.frame(x=ggplot_build(p)$data[[1]]$x,
+                 f=ggplot_build(p)$data[[1]]$density)
```

```
> #Finish plot
> g2<-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(bquote("Median City Population"~(hat(m))))+
+   ylab("Density")+
+   xlim(0,15)+
+   annotate("text",x=3.5,y=0.025,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=12.5,y=0.025,label=deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text",x=7.75,y=0.075,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)
> grid.arrange(g1,g2,ncol=2)
```

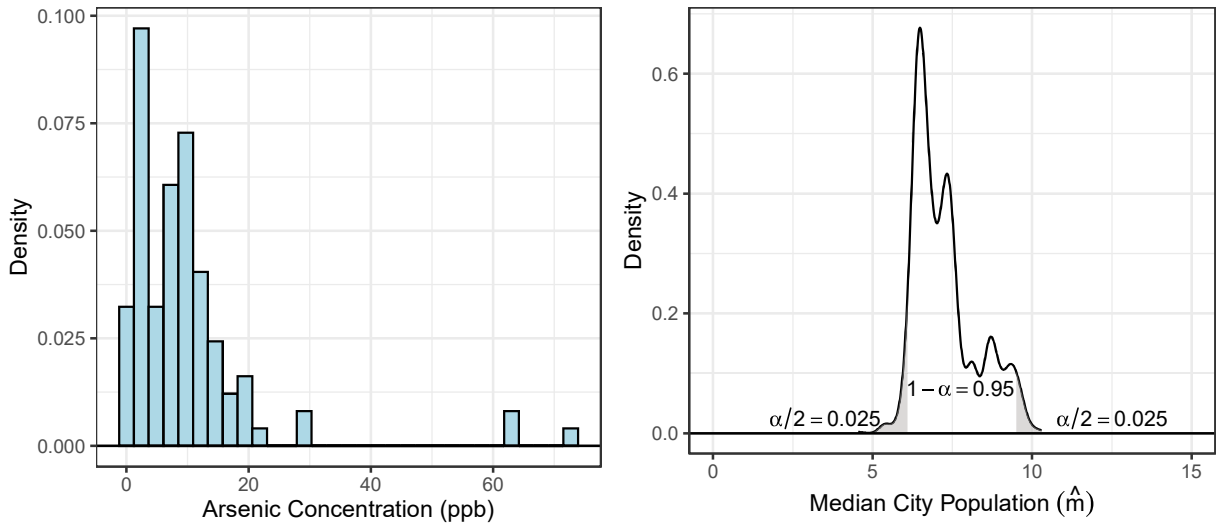

Figure 9.4.11: The histogram of $n = 102$ arsenic concentration observations (left) and the approximate sampling distribution of $\hat{m}$ with key values of the bootstrapping confidence interval for the population median highlighted.

Resampling Approximation Check: As long as our sample of wells is representative of all wells in Texas, our assumptions are met. We might have a conversation with the researcher and ask to which population the sample is generalizable to – is it all wells in Texas? Is it all wells in a certain area of Texas?