Chapter 3
The Sampling Distribution of the Sample Mean

We begin by establishing a fundamental fact about any normal distribution: about 95% of the probability lies within 2 SD of the mean. If we integrate the area under these curves, between 2 SD below the mean and 2 SD above the mean, we find the following areas, which correspond to the amount of probability within these bounds:

```
> integrate(function(x) dnorm(x, mean = 0, sd = 1),
        -2, 2)

0.9544997 with absolute error < 1.8e-11

> integrate(function(x) dnorm(x, mean = 0, sd = 4),
        -8, 8)

0.9544997 with absolute error < 1.8e-11
```

We can display this fact graphically (see Figure 3.1):

```
> main.title <- "Area within 2 SD of the mean"
> multiplot(1, 2)
> plot(function(x) dnorm(x, mean = 0, sd = 1),
        xlim = c(-3, 3), main = "SD 1", xlab = "x",
        ylab = "", cex = 2)
> segments(-2, 0, -2, 0.4)
> segments(2, 0, 2, 0.4)
> plot(function(x) dnorm(x, mean = 0, sd = 4),
        xlim = c(-12, 12), main = "SD 4", xlab = "x",
        ylab = "", cex = 2)
> segments(-8, 0, -8, 0.1)
> segments(8, 0, 8, 0.1)
```

Suppose that we have a population of people and that we know the age of each individual; let us assume also that distribution of the ages is approximately normal. Finally, let us also suppose that we know that mean age of the population is 60 and the population SD is 4.

S. Vasishth, M. Broe, *The Foundations of Statistics: A Simulation-based Approach*, DOI 10.1007/978-3-642-16313-5_3,

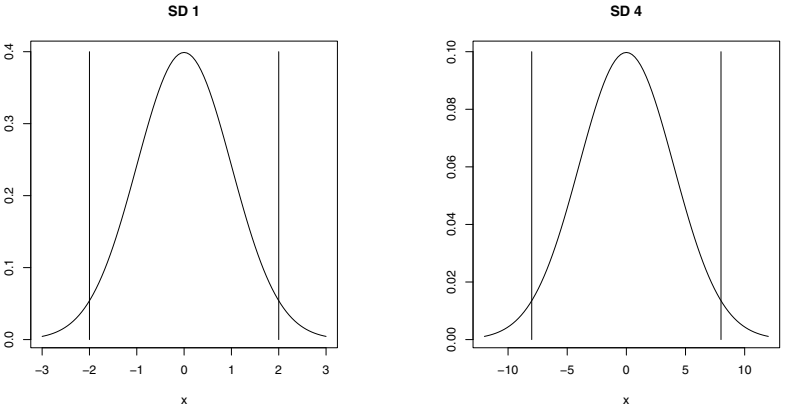

Fig. 3.1 Two normal distributions with SD = 1 (left), SD = 4 (right). The lines delimit the region 2 SD from the mean in each case.

Now suppose that we repeatedly sample from this population: we take samples of 40, a total of 1000 times; and we calculate the mean $\bar{x}$ each time we take a sample. After taking 1000 samples, we have 1000 means; if we plot the distribution of these means, we have the sampling distribution of the sample mean.

```
> sample.means <- rep(NA, 1000)
> for (i in 1:1000) {
       sample.40 <- rnorm(40, mean = 60, sd = 4)
       sample.means[i] <- mean(sample.40)
   }
```

We can calculate the mean and standard deviation of this sampling distribution:

```
> means40 <- mean(sample.means)

[1] 59.98692

> sd40 <- sd(sample.means)

[1] 0.6156489
```

If we plot this distribution of means, we find that it is roughly normal. We can characterize the distribution of means visually, as done in Figure 3.2 below, or in terms of the mean and standard deviation of the distribution. The mean value in the above simulation is 59.99 and the standard deviation of the distribution of means is 0.6156. Note that if you repeatedly run the above simulation code, these numbers will differ slightly in each run.

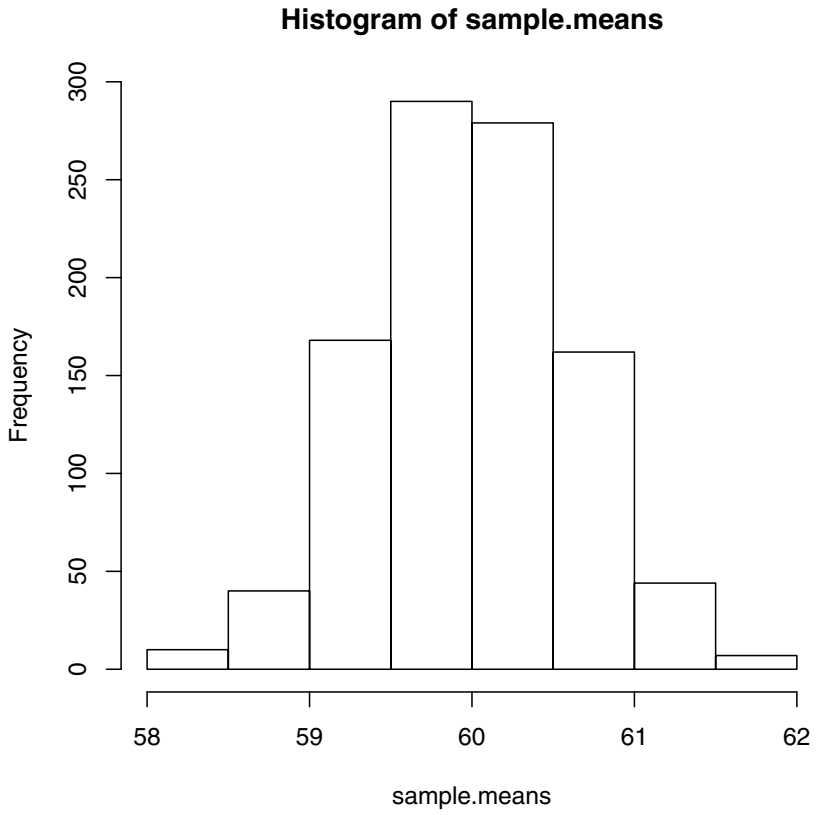

Fig. 3.2 The sampling distribution of the sample mean with 1000 samples of size 40.

```
> hist(sample.means)
```

Consider now the situation where our sample size is 100. Note that the mean and standard deviation of the population ages is the same as above.

```
> sample.means <- rep(NA, 1000)
> for (i in 1:1000) {
       sample.100 <- rnorm(100, mean = 60, sd = 4)
       sample.means[i] <- mean(sample.100)
   }

> means100 <- mean(sample.means)

[1] 59.99521

> sd100 <- sd(sample.means)
```

```
[1] 0.4065139
```

In this particular simulation run, the mean of the means is 60 and the standard deviation of the distribution of means is 0.4065. Let's plot the distribution of the means (Figure 3.3).

```
> hist(sample.means)
```

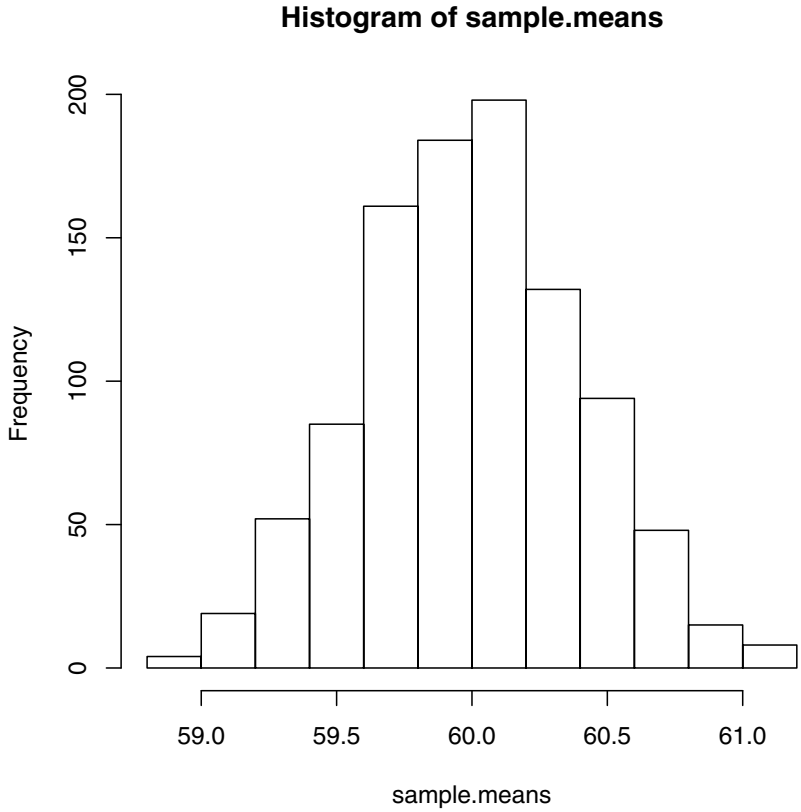

Fig. 3.3 The sampling distribution of the sample mean with samples of size 100.

The above simulations show us several things. First, the standard deviation of the distribution of means gets smaller as we increase sample size. When the sample size is 40, the standard deviation is 0.6156; when it is 100, the standard deviation is 0.4065. Second, as the sample size is increased, the mean of the sample means comes closer and closer to the *population* mean $\mu_{\bar{x}}$. A third point (which is not obvious at the moment) is that there is a

lawful relationship between the standard deviation σ of the population and the standard deviation of the *distribution of means*, which we will call $\sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.1}$$

Here, n is the sample size. It is possible to derive equation 3.1 from first principles. We do this in Appendix A. For now, simply note the important point that n is in the denominator in this equation, so there is an inverse relationship between the sample size and the standard deviation of the sample means. Let's take this equation on trust for the moment and use it to compute $\sigma_{\bar{x}}$ by using the population standard deviation (which we assume we know). Let's do this for a sample of size 40 and another of size 100:

```
> 4/sqrt(40)

[1] 0.6324555

> 4/sqrt(100)

[1] 0.4
```

The above calculation is consistent with what we just saw: $\sigma_{\bar{x}}$ gets smaller and smaller as we increase sample size.

We have also introduced a notational convention that we will use throughout the book: sample statistics are symbolized by Latin letters $(\bar{x}, s)$; population parameters are symbolized by Greek letters (μ, σ).

3.1 The Central Limit Theorem

We've seen in the previous chapter that the distribution of a sample count (and sample proportion) has the shape of a binomial distribution, which is closely approximated by the normal distribution. Now we see that the *sampling distribution of the sample mean* is also normally distributed. In the above example the means were drawn from a population with normally distributed scores. It turns out that the sampling distribution of the sample mean will be normal even if the population is not normally distributed, as long as the sample size is large enough. This is known as the Central Limit Theorem, and it is so important that we will say it twice:

> Provided the sample size is large enough, the sampling distribution of the sample mean will be close to normal *irrespective of what the population's distribution looks like.*

Let's check whether this theorem holds by testing it in an extreme case, simulating a population which we *know* is not normally distributed. Let's take

our samples from a population (Figure 3.4) whose values are distributed exponentially with the same mean of 60 (the mean of an EXPONENTIAL DISTRIBUTION is the reciprocal of the so-called 'rate' parameter).

```
> sample.100 <- rexp(100, 1/60)
> hist(sample.100)
```

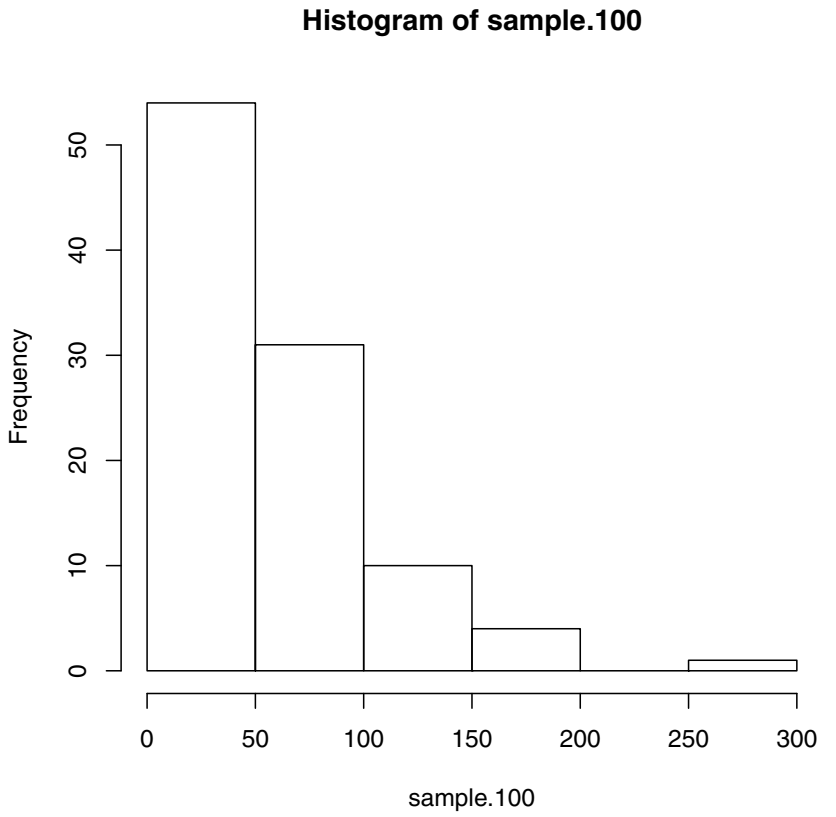

Fig. 3.4 A sample from exponentially distributed population scores.

Now let us plot the sampling distribution of the sample mean. We take 1000 samples of size 100 each from this exponentially distributed population. As shown in Figure 3.5, the distribution of the means is again normal!

```
> sample.means <- rep(NA, 1000)
> for (i in 1:1000) {
      sample.100 <- rexp(100, 1/60)
```

```
        sample.means[i] <- mean(sample.100)
    }
> hist(sample.means)
```

Recall that the mean of each sample is a 'point summary' of the distribution. Some of these samples will have a mean slightly above the true mean, some slightly below, and the sampling distribution of *these* values is roughly normal. Try altering the sample size in this example to get a feel for what happens if the sample size is not 'large enough.'

To summarize:

1. The sampling distributions of various statistics (the sampling distribution of the sample mean, or sample proportion, or sample count) are nearly normal. The normal distribution implies that a sample statistic that is close to the mean has a higher probability than one that is far away.
2. The mean of the sampling distribution of the sample mean is (in the limit) the same as the population mean.
3. It follows from the above two facts that the mean of a sample is more likely to be close to the population mean than not.

3.2 σ and $\sigma_{\bar{x}}$

We saw earlier that the standard deviation of the sampling distribution of the sample mean $\sigma_{\bar{x}}$ gets smaller as we increase sample size. When the sample has size 40, this standard deviation is 0.6156; when it is 100, this standard deviation is 0.4065.

Let's study the relationship between $\sigma_{\bar{x}}$ and σ. Recall that our population mean $\mu = 60$, $\sigma = 4$. The equation below summarizes the relationship; it shouldn't surprise you, since we just saw it above (also see Appendix A):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.2}$$

But note also that the tighter the distribution, the greater the probability that the estimate of the mean based on *a single sample* is close to the population mean. So the $\sigma_{\bar{x}}$ is an indicator of how good our estimate of the population mean is. As we increase the size of a single sample, the smaller the standard deviation of its corresponding sampling distribution becomes, and the higher the probability of its providing a good estimate of the population parameter. Let's quantify exactly how this estimate improves.

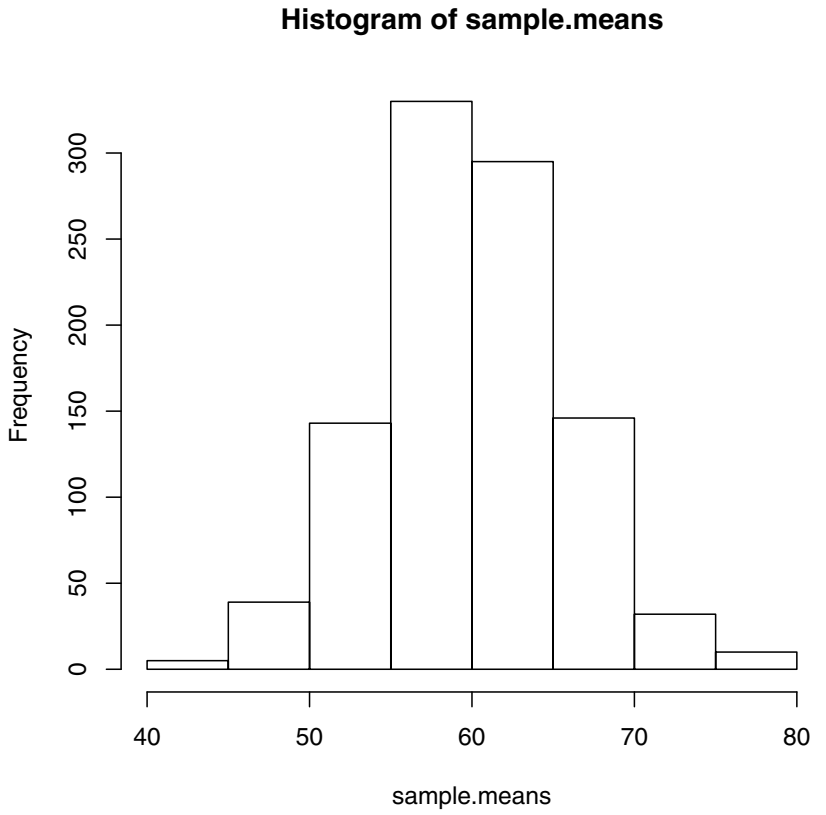

Fig. 3.5 The sampling distribution of sample mean from an exponentially distributed population.

3.3 The 95% Confidence Interval for the Sample Mean

Let's take a sample of 11 heights from a normally distributed population with known mean height $\mu = 60$ and SD $\sigma = 4$ (inches).

```
> sample.11 <- rnorm(11, mean = 60, sd = 4)

 [1] 58.88860 54.30641 59.35683 59.91140 56.21208 56.40233
 [7] 56.67023 63.84889 57.14325 50.89041 61.36143
```

And now let's estimate a population mean from this sample using the sample mean $\bar{x}$, and compute the SD $\sigma_{\bar{x}}$ of the corresponding sampling distribution. Since we know the precise population standard deviation we can get a precise value for $\sigma_{\bar{x}}$.

```
> estimated.mean <- mean(sample.11)

[1] 57.72653

> SD.population <- 4

[1] 4

> n <- length(sample.11)

[1] 11

> SD.distribution <- SD.population/sqrt(n)

[1] 1.206045
```

We know from the Central Limit Theorem that the sampling distribution of the sample mean is roughly normal, and we know that in this case $\sigma_{\bar{x}} = 1.2$. Recall that the probability that the population mean is within 2 $\sigma_{\bar{x}}$ of the sample mean is a bit over 0.95. Let's calculate this range:

$$\bar{x} \pm (2 \times \sigma_{\bar{x}}) = 58 \pm (2 \times 1.206) \tag{3.3}$$

The 0.95 probability region is between 55.3 and 60.1. The number 0.95 is a probability from the point of view of the sampling distribution, and a confidence level from the point of view of parameter estimation. In the latter case it's conventionally expressed as a percentage and is called the 95% confidence interval (CI).

Suppose now that sample size was four times bigger (44). Let's again calculate the sample mean, the standard deviation of the corresponding sampling distribution, and from this information, compute the 95% confidence interval.

```
> sample.44 <- rnorm(44, mean = 60, sd = 4)
> estimated.mean <- mean(sample.44)
> n <- length(sample.44)
> (SD.distribution <- 4/sqrt(n))

[1] 0.6030227
```

Now we get a much tighter 95% confidence interval:

$$\bar{x} \pm 2 \times \sigma_{\bar{x}} = 60 \pm 2 \times 0.603 \tag{3.4}$$

The interval now is between 59.2 and 61.6, smaller than the one we got for the smaller sample size. In fact, it is exactly half as wide. Take a moment to make sure you understand why.

3.4 Realistic Statistical Inference

Until now we have been sampling from a population whose mean and standard deviation we know. However, we normally don't know the population parameters. In other words, although we know that:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.5}$$

when we take samples in real life, we almost never know σ. After all, it is based on an average of distances from the population mean μ, and that is usually the very thing we are trying to estimate!

What we *do* have, however, is the standard deviation *of the sample itself* (denoted s). This in turn means that we can only get an *estimate* of $\sigma_{\bar{x}}$. This is called the STANDARD ERROR (SE) of the sample mean (or of whatever statistic we are measuring.):

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.6}$$

Pay careful attention to the distinction between s (an estimate of the standard deviation of the population σ) and $SE_{\bar{x}}$ (an estimate of the standard deviation of the sampling distribution, which is in turn based on s).

We saw previously that the size of $\sigma_{\bar{x}}$—a measure of the spread of the sampling distribution—is crucial in determining the size of a 95% confidence interval for a particular sample. Now we only have an estimate of that spread. Moreover, the estimate will change from sample to sample, as the value of s changes. This introduces a new level of uncertainty into our task: it has become an estimate based on an estimate! Intuitively, we would expect the confidence interval to increase in size, reflecting this increase in uncertainty. We will see how to quantify this intuition presently.

First, however, we should explore the pattern of variability in this new statistic we have introduced, s, which (like the sample mean) will vary randomly from sample to sample. Can we safely assume that s is a reliable estimate of σ?

3.5 s is an Unbiased Estimator of σ

Earlier in this chapter we repeatedly sampled from a population of people with mean age 60 and standard deviation 4; then we plotted the distribution of sample means that resulted from the repeated samples. One thing we noticed was that any one sample mean was more likely to be close to the population mean. Let's repeat this experiment, but this time we plot the distribution of the sample's standard deviation (Figure 3.6).

```
> sample.sds <- rep(NA, 1000)
> for (i in c(1:1000)) {
       sample.40 <- rnorm(40, mean = 60, sd = 4)
       sample.sds[i] <- sd(sample.40)
   }
> hist(sample.sds)
```

What we see is that any one sample's standard deviation s is more likely to be close to the population standard deviation σ. This is because the sampling distribution of the standard deviations also has a normal distribution.

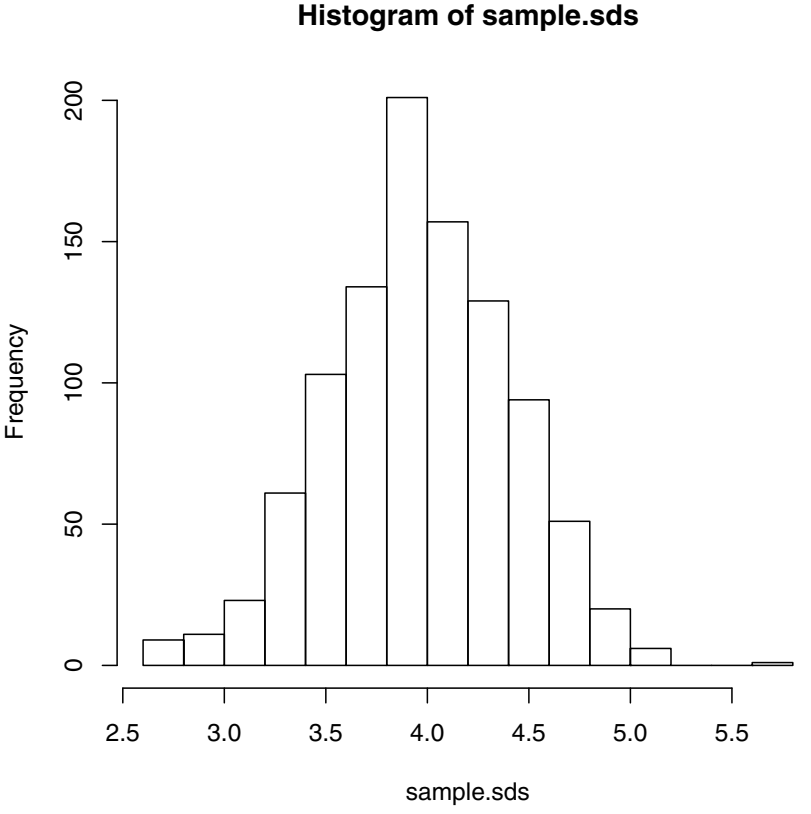

Fig. 3.6 The distribution of the sample standard deviation, sample size 40. The population is normally distributed.

So, if we use s as an estimator of σ we're more likely than not to get close to the right value: we say s is an UNBIASED ESTIMATOR of σ. This is true

even if the population is not normally distributed. Let's check this again for an exponentially distributed population whose SD is 1 (Figure 3.7).

```
> sample.sds <- rep(NA, 1000)
> for (i in c(1:1000)) {
        sample.sds[i] <- sd(rexp(40))
   }
> hist(sample.sds)
```

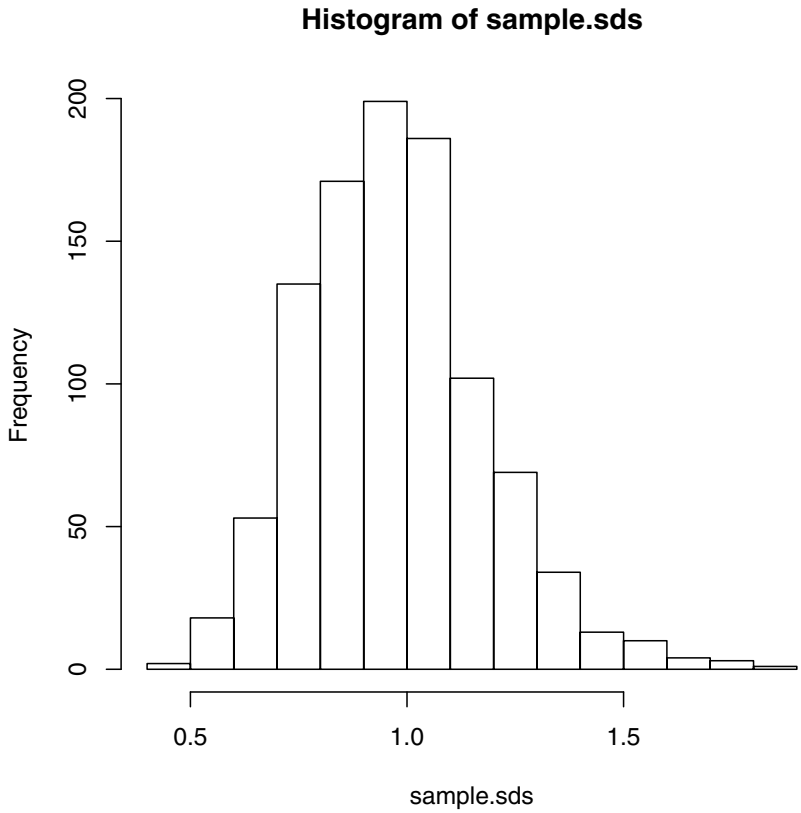

Fig. 3.7 The sampling distribution of sample standard deviations from an exponentially distributed population.

We are now at the point where we can safely use the sample standard deviation s as an estimate of the unknown population standard deviation σ, and this in turn allows us to estimate the standard deviation of the sampling distribution $\sigma_{\bar{x}}$ using the Standard Error $SE_{\bar{x}}$.

Notice that the Standard Error will vary from sample to sample, since the estimate s of the population parameter σ will vary from sample to sample. And of course, as the sample size increases the estimate s becomes more accurate, as does the SE, suggesting that the uncertainty introduced by this extra layer of estimation will be more of an issue for smaller sample sizes.

Our problem now is that the sampling distribution of the sample mean can no longer be modeled by the normal distribution, which is the distribution based ultimately on the *known* parameter σ.

If we were to derive some value v for the SE, and simply plug this in to the normal distribution for the sample statistic, this would be equivalent to claiming that v *really was* the population parameter σ.

What we require is a distribution whose shape has greater uncertainty built into it than the normal distribution. This is the motivation for using the so-called t-DISTRIBUTION, which we turn to next.

3.6 The t-distribution

As discussed above, the distribution needs to reflect greater uncertainty at small sample sizes. There is in fact a family of t-distribution curves whose shapes vary with sample size. In the limit, if the sample were the size of the entire population, the t-distribution would *be* the normal distribution (since then s would *be* σ), so the t-curve becomes more normal as sample size increases. This distribution is formally defined by the DEGREES OF FREEDOM (which is simply sample size minus 1 in this case) and has more of the total probability located in the tails of the distribution. It follows that the probability of a sample mean being close to the true mean is slightly lower when measured by this distribution, reflecting our greater uncertainty. You can see this effect in Figure 3.8 at small sample sizes:

```
> range <- seq(-4, 4, 0.01)
> multiplot(2, 2)
> for (i in c(2, 5, 15, 20)) {
      plot(range, dnorm(range), lty = 1, col = gray(0.5),
          xlab = "", ylab = "", cex.axis = 1.5)
      lines(range, dt(range, df = i), lty = 2,
          lwd = 2)
      mtext(paste("df=", i), cex = 1.2)
  }
```

But notice that with about 15 degrees of freedom, the t-distribution is already very close to normal.

What do we have available to us to work with now? We have an estimate s of the population SD, and so an estimate $SE_{\bar{x}}$ of the SD of the sampling distribution:

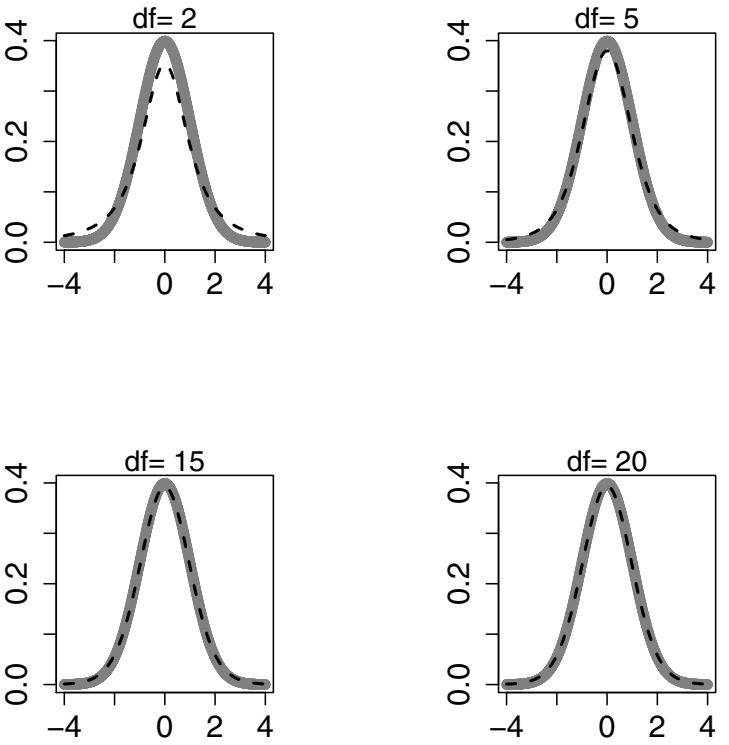

Fig. 3.8 A comparison between the normal (solid gray line) and t-distribution (broken black line) for different degrees of freedom.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.7}$$

We also have a more spread-out distribution than the normal, the t-distribution, defined by the degrees of freedom (in this case, sample size minus 1). We are now ready to do realistic statistical inference.

3.7 The One-sample t-test

How do we build a confidence interval based on this new model of inference?

We need to ask: how many SE's do we need to go to the left and right of the sample mean, within the appropriate t-distribution, to be 95% sure that

the population mean lies in that range? In the pre-computing days, people used to look up a table that told you, for $n-1$ degrees of freedom, how many SE's you need to go around the sample mean to get a 95% CI. Now we can ask R. First we take a sample of size 11 from a population with mean 60 and standard deviation 4.

```
> sample <- rnorm(11, mean = 60, sd = 4)
```

Using this sample, we can ask for the 95% confidence interval:

```
> t.test(sample)$conf.int
```

```
[1] 56.78311 63.56462
attr(,"conf.level")
[1] 0.95
```

Note that all of the information required to perform this inference is contained in the sample itself: the sample mean; the sample size and sample standard deviation s (from which we compute the SE), the degrees of freedom (the sample size minus 1, from which we reference the appropriate t-distribution). Sure enough, if our sample size had been larger, our CI would be narrower:

```
> sample <- rnorm(100, mean = 60, sd = 4)
```

```
> t.test(sample)$conf.int
```

```
[1] 58.62889 60.20964
attr(,"conf.level")
[1] 0.95
```

3.8 Some Observations on Confidence Intervals

There are some subtleties associated with confidence intervals that are often not brought up in elementary discussions, simply because the issues are just too daunting to tackle. However, we will use simulations to unpack some of these subtleties. We hope that the reader will see that the issues are in reality quite simple.

The first critical point to understand is the meaning of the confidence interval. We have been saying up till now that the 95% confidence interval tells you the range within which we are 95% sure that the population mean lies. However, one important point to notice is that the range defined by the confidence interval will vary with each sample even if the sample size is kept constant. The reason is that the sample mean will vary each time, and the standard deviation will vary too. We can check this fact quite easily.

First we define a function for computing 95% CIs:[1]

```
> se <- function(x) {
       y <- x[!is.na(x)]
       sqrt(var(as.vector(y))/length(y))
   }
> ci <- function(scores) {
       m <- mean(scores, na.rm = TRUE)
       stderr <- se(scores)
       len <- length(scores)
       upper <- m + qt(0.975, df = len - 1) * stderr
       lower <- m + qt(0.025, df = len - 1) * stderr
       return(data.frame(lower = lower, upper = upper))
   }
```

Next, we take 100 samples repeatedly from a population with mean 60 and SD 4, computing the 95% CI each time.

```
> lower <- rep(NA, 100)
> upper <- rep(NA, 100)
> for (i in 1:100) {
       sample <- rnorm(100, mean = 60, sd = 4)
       lower[i] <- ci(sample)$lower
       upper[i] <- ci(sample)$upper
   }
> cis <- cbind(lower, upper)
> head(cis)

        lower    upper
[1,] 59.30995 60.84080
[2,] 59.59644 61.10296
[3,] 58.90593 60.61683
[4,] 58.51755 60.22174
[5,] 58.85010 60.50782
[6,] 59.44144 60.87652
```

Thus, the center and the size of any one confidence interval, based on a single sample, will depend on the particular sample mean and standard deviation you happen to observe for that sample. The sample mean and standard deviation are likely to be close to the population mean and standard deviation, but they are ultimately just estimates of the true parameters.

Importantly, because of the normally distributed shapes of the distribution of sample means and sample standard deviations (see Figures 3.5 and 3.7), if we repeatedly sample from a population and compute the confidence intervals

[1] Here, we use a built-in R function called `qt(p,DF)` which, for a given confidence-interval range (say, 0.975), and a given degrees of freedom, DF, tells you the corresponding critical t-value.

each time, in approximately 95% of the confidence intervals the population mean will lie within the ranges specified. In the other 5% or so of the cases, the confidence intervals will not contain the population mean.

This is what the '95%' confidence interval means: it's a statement about hypothetical repeated samples. More specifically, it's a statement about the probability that the hypothetical confidence intervals (that would be computed from the hypothetical repeated samples) will contain the population mean.

Let's check the above statement. We can repeatedly build 95% CIs and determine whether the population mean lies within them. The claim is that it will lie within the CI approximately 95% of the time.

```
> store <- rep(NA, 100)
> pop.mean <- 60
> pop.sd <- 4
> for (i in 1:100) {
       sample <- rnorm(100, mean = pop.mean, sd = pop.sd)
       lower[i] <- ci(sample)$lower
       upper[i] <- ci(sample)$upper
       if (lower[i] < pop.mean & upper[i] > pop.mean) {
           store[i] <- TRUE
       }
       else {
           store[i] <- FALSE
       }
   }
> cis <- cbind(lower, upper)
> store <- factor(store)
> summary(store)

FALSE  TRUE
    6    94
```

So that's more or less true. To drive home the point, we can also plot the confidence intervals to visualize the proportion of cases where each CI contains the population mean (Figure 3.9).

```
> main.title <- "95% CIs in 100 repeated samples"
> line.width <- ifelse(store == FALSE, 2, 1)
> cis <- cbind(cis, line.width)
> x <- 0:100
> y <- seq(55, 65, by = 1/10)
> plot(x, y, type = "n", xlab = "i-th repeated sample",
       ylab = "Scores", main = main.title)
> abline(60, 0, lwd = 2)
> x0 <- x
> x1 <- x
```

```
> arrows(x0, y0 = cis[, 1], x1, y1 = cis[, 2],
       length = 0, lwd = cis[, 3])
```

In this figure, we control the width of the lines marking the CI using the information we extracted above (in the object `store`) to determine whether the population mean is contained in the CI or not: when a CI does not contain the population mean, the line is thicker than when it does contain the mean. You should try repeatedly sampling from the population as we did above, computing the lower and upper ranges of the 95% confidence interval, and then plotting the results as shown in Figure 3.9.

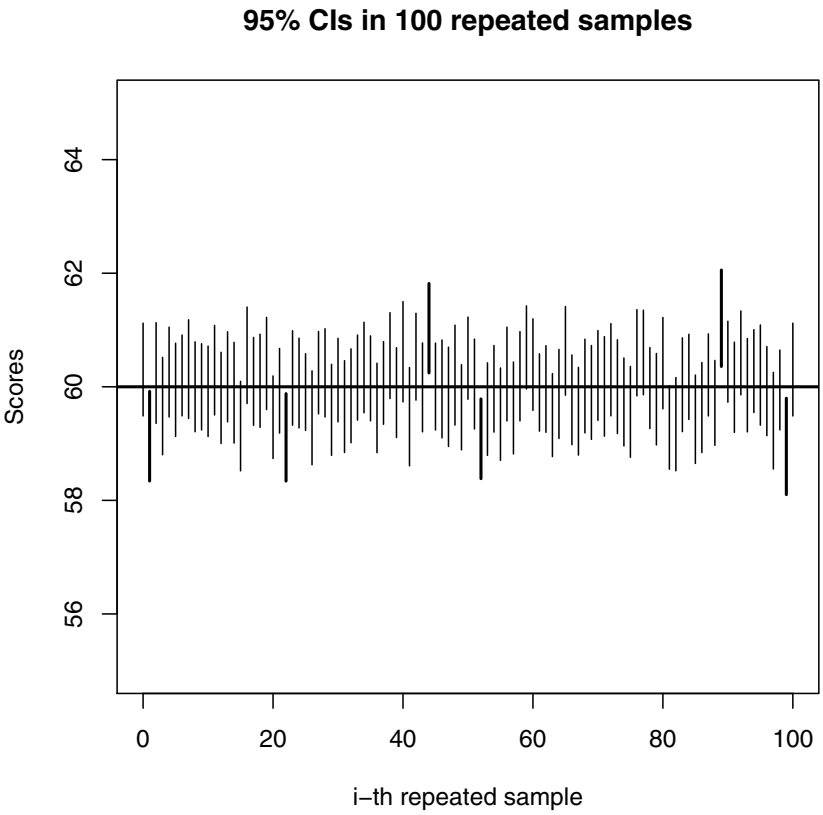

Fig. 3.9 A visualization of the proportion of cases where the population mean is contained in the 95% CI, computed from repeated samples. The CIs that do not contain the population mean are marked with thicker lines.

Note that when we compute a 95% confidence interval for a particular sample, we have only *one* interval. Strictly speaking, that particular interval

does *not* mean that the probability that the population mean lies *within that interval* is 0.95. For that statement to be true, it would have to be the case that the population mean is a random variable, like the heads and tails in a coin are random variables, and 1 through 6 on a die are random variables.

The population mean is a single point value that cannot have a multitude of possible values and is therefore not a random variable. If we relax this assumption, that the population mean is a point value, and assume instead that 'the' population mean is in reality a range of possible values (each value having different probabilities of being the population mean), then we could say that any one 95% confidence interval represents the range within which the population mean lies with probability 0.95. See the book by Gelman and Hill (2007) for more detail on this approach.

It's worth repeating the above point about confidence intervals. The meaning of the confidence interval depends crucially on hypothetical repeated samples: the confidence intervals computed in 95% of these repeated samples will contain the population mean. In essence, the confidence interval from a single sample is a random variable just like heads and tails in a coin toss, or the numbers 1 through 6 in a die, are random variables. Just as a fair coin has a 0.5 chance of yielding a heads, and just as a fair die has a 1/6 chance of landing a 1 or 2 etc., a confidence interval in repeated sampling has a 0.95 chance of containing the population mean.

3.9 Sample SD, Degrees of Freedom, Unbiased Estimators

Let's revisit the question: Why do we use $n-1$ in the equation for standard deviation? Recall that the sample standard deviation s is just the root of the variance: the average distance of the numbers in the list from the mean of the numbers:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.8}$$

We can explore the reason why we use $n-1$ in the context of estimation by considering what would happen if we simply used n instead. As we will see, if we use n, then s (which is an estimate of the population variance σ) would be smaller. This smaller s turns out to provide a poorer estimate than when we use $n-1$: it is a BIASED ESTIMATOR. Let's verify this using simulations.

We define new variance and standard deviation functions that use n, and simulate the sampling distribution of this new statistic s' from a population with known standard deviation $\sigma = 1$).

```
> new.var <- function(x) {
        sum((x - mean(x))^2)/length(x)
   }
```

```
> new.sd <- function(x) {
      sqrt(new.var(x))
  }
> correct <- rep(NA, 1000)
> incorrect <- rep(NA, 1000)
> for (i in 1:1000) {
      sample.10 <- rnorm(10, mean = 0, sd = 1)
      correct[i] <- sd(sample.10)
      incorrect[i] <- new.sd(sample.10)
  }
```

As shown below (Figure 3.10), using n gives a biased estimate of the true standard deviation:

```
> multiplot(1, 2)
> hist(correct, main = paste("Mean ", round(mean(correct),
      digits = 2), sep = " "))
> hist(incorrect, main = paste("Mean ", round(mean(incorrect),
      digits = 2), sep = " "))
```

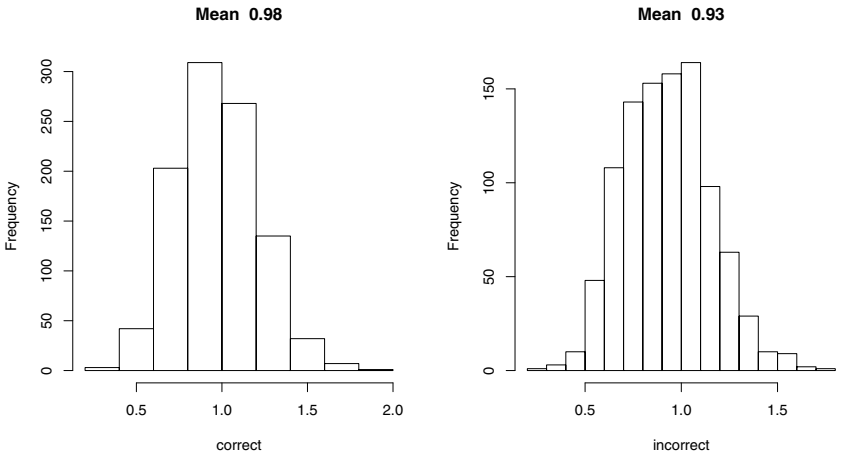

Fig. 3.10 The consequence of taking $n-1$ versus n in the denominator for calculating variance, sample size 10.

3.10 Summary of the Sampling Process

It is useful at this point to summarize the terminology we have been using, and the logic of sampling. First, take a look at the concepts we have covered so far.

We provide a list of the different concepts in Table 3.1 below for easy reference. Here, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$.

Table 3.1 A summary of the notation used.

The sample statistic	is an unbiased estimate of
sample mean $\bar{x}$	population mean μ
sample SD s	population SD σ
standard error $SE_{\bar{x}}$	sampling distribution $\sigma_{\bar{x}}$

1. Statistical inference involves a single sample value but assumes knowledge of the sampling distribution which provides probabilities for all possible sample values.
2. The statistic (e.g., mean) in a random sample is more likely to be closer to the population parameter (the population mean) than not. This follows from the normal distribution of the sample means.
3. In the limit, the mean of the sampling distribution is equal to the population parameter.
4. The further away a sample statistic is from the mean of the sampling distribution, the lower the probability that such a sample will occur.
5. The standard deviation of the sampling distribution $\sigma_{\bar{x}}$ is partially determined by the inherent variability σ in the population, and partially determined by the sample size. It tells us how steeply the probability falls off from the center. If $\sigma_{\bar{x}}$ is small, then the fall-off in probability is steep: random samples are more likely to be very close to the mean, samples are better indicators of the population parameters, and inference is more certain. If $\sigma_{\bar{x}}$ is large, then the fall-off in probability from the center is gradual: random samples far from the true mean are more likely, samples are not such good indicators of the population parameters, and inference is less certain.
6. While we do not know $\sigma_{\bar{x}}$, we can estimate it using $SE_{\bar{x}}$ and perform inference using a distribution that is almost normal, but reflects the increase in uncertainty arising from this estimation: the t-distribution.

3.11 Significance Tests

Recall the discussion of 95% confidence intervals: The sample gives us a mean $\bar{x}$. We compute $SE_{\bar{x}}$ (an estimate of $\sigma_{\bar{x}}$) using s (an estimate of σ) and sample size n. Then we calculate the range $\bar{x} \pm 2 \times SE_{\bar{x}}$. That's the 95% CI.

We don't know the population mean—if we did, why bother sampling? But suppose we have a *hypothesis* that the population mean has a certain value. If we have a hypothesis about the population mean, then we also know what the corresponding sampling distribution would look like: we know the probability of any possible sample given that hypothesis. We then take an actual sample, measure the distance of our sample mean from the hypothesized population mean, and use the facts of the sampling distribution to determine the probability of obtaining such a sample *assuming the hypothesis is true*. This amounts to a *test* of the hypothesis. Intuitively, if the probability of our sample (given the hypothesis) is high, this provides evidence the hypothesis is true. In a sense, this is what our hypothesis predicts. Conversely, if the probability of the sample is low (given the hypothesis), this is evidence against the hypothesis. The hypothesis being tested in this way is termed the NULL HYPOTHESIS. Let's do some simulations to better understand this concept.

Suppose our hypothesis, based perhaps on previous research, is that the population mean is 70, and let's assume for the moment the population $\sigma = 4$. This in turn means that the sampling distribution of the mean, given some sample size, say 11, would have a mean of 70, and a standard deviation $\sigma_{\bar{x}} = 1.2$:

```
> SD.distribution = 4/sqrt(11)

[1] 1.206045
```

Figure 3.11 shows what we expect our sampling distribution to look like if our hypothesis *were in fact* true. This hypothesized distribution is going to be our reference distribution on which we base our test.

```
> range <- seq(55, 85, 0.01)
> plot(range, dnorm(range, mean = 70, sd = SD.distribution),
       type = "l", ylab = "", main = "The null hypothesis")
```

Suppose now that we take an actual sample of 11 from a population whose mean μ is in fact (contra the hypothesis) 60:

```
> sample <- rnorm(11, mean = 60, sd = 4)

> sample.mean <- mean(sample)

[1] 60.76659
```

Inspection of (Figure 3.11) shows that, in a world in which the population mean was really 70, the probability of obtaining a sample whose mean is 61 is extremely low. Intuitively, this sample is evidence against the null hypothesis.

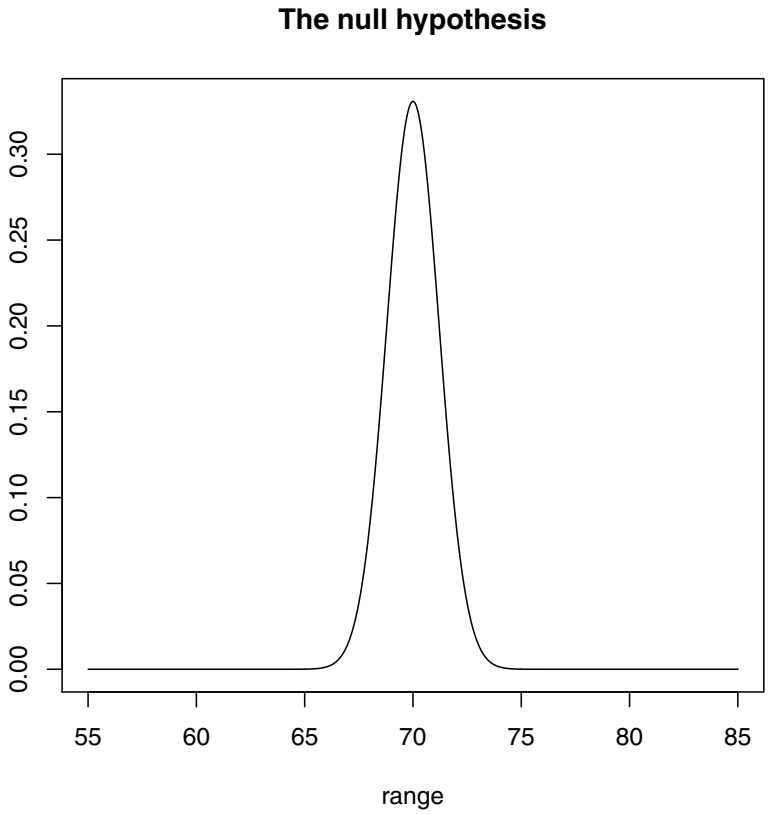

Fig. 3.11 A sampling distribution with mean 70 and $\sigma_{\bar{x}} = 1.206$.

A SIGNIFICANCE TEST provides a formal way of quantifying this insight. The result of such a test yields a probability that indicates exactly how well or poorly the data and the hypothesis agree.

3.12 The Null Hypothesis

While this perfectly symmetrical, intuitive way of viewing things ('evidence for', 'evidence against') is on the right track, there is a further fact about the null hypothesis which gives rise to an asymmetry in the way we perform significance tests.

The statement being tested in a significance test— the NULL HYPOTHESIS, H_0— is usually formulated in such a way that the statement represents

'no effect,' 'pure chance' or 'no significant difference'—or as Polya called it 'chance, the ever present rival conjecture' (Polya, 1954, 55). Scientists are generally not interested in proving 'no effect.' This is where the asymmetry comes in: we are usually not interested in finding evidence *for* the null hypothesis, conceived in this way. Rather, we are interested in evidence *against* the null hypothesis, since this is evidence for some real statistically significant result. This is what a formal significance test does: it determines if the result provides sufficient evidence against the null hypothesis for us to reject it. Note that if it doesn't provide sufficient evidence, we have *not* proved the contrary—we have not 'proved the null hypothesis.' We simply don't have enough evidence, *based on this single result*, to reject it. We sharpen this idea in chapter 4.

In order to achieve a high degree of skepticism about the interpretation of the data, we require the evidence against the null hypothesis to be very great. In our current example, you might think the result we obtained, 61, was fairly compelling evidence against it. But how do we quantify this? Intuitively, the further away from the mean of the sampling distribution our data lies, the greater the evidence against it. Statistically significant results reside out in the tails of the distribution. How far out? The actual values and ranges of values we get will vary from experiment to experiment, and statistic to statistic. How can we determine a general rule?

3.13 z-scores

We have already seen that, in a normal distribution, about 95% of the total probability falls within 2 SD of the mean, and thus 5% of the probability falls far out in the tails. One way of setting a general rule then, is to say that if an observed value falls far out in the tail of the distribution, we will consider the result extreme enough to reject the null hypothesis (we can set this threshold anywhere we choose: 95% is a conventional setting).

Recall our model: we know the sampling distribution we would see in a world in which the null hypothesis is true, in which the population mean is really 70 (and whose population σ is known to be 4). We also know this distribution is normal. How many SDs from the mean is our observation? Is it more than 2 SDs?

We need to express the difference between our observation $\bar{x}$ and hypothesized mean of the distribution μ_0 in units of the standard deviation of the distribution: i.e., some number z times $\sigma_{\bar{x}}$. We want to know this number z.

$$\bar{x} - \mu_0 = z\sigma_{\bar{x}} \tag{3.9}$$

Solving for z:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \tag{3.10}$$

$$= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{3.11}$$

z is called the STANDARDIZED VALUE or the Z-SCORE. In addition, one could imagine computing this standardized version of the sample mean every time we take a sample, in which case we have effectively defined a new statistic. Viewed in this way, the score is also referred to as a TEST-STATISTIC.

Let's make this concrete. Suppose in our current simulation we draw a sample whose mean is precisely 60: then $\bar{x} = 60, \mu_0 = 70, \sigma = 4, n = 11$. So we get:

$$z = \frac{60 - 70}{4/\sqrt{11}} \tag{3.12}$$

$$= -8.291562 \tag{3.13}$$

We see that this observation is well beyond 2 SDs from the mean, and thus represents statistically significant evidence against the null hypothesis.

Notice there is a natural interpretation of this process in terms of confidence intervals. We know that, on repeated sampling from a population that has mean 70 (the null hypothesis), in 95% of the samples the true population mean (70) would fall within two SDs of the sample mean (which of course would be different for each sample). It follows that in 5% of the samples, this 2 SD confidence interval will *not* contain the population mean. These are precisely the samples which fall out in the tails of the distribution, more than 2 SD from the distribution mean.

z-scores are a quick and accepted way of expressing 'how far away' from the hypothetical value an observation falls, and for determining if that observation is beyond some accepted threshold. Ultimately, however, they take their meaning from the probability corresponding to the value, which is traditionally expressed by rules-of-thumb (2 SD corresponds to 95%), or tables which translate particular scores to particular probabilities. It is this probability we turn to next.

3.14 P-values

It would be nice if we could set a probability threshold like this: 'If the probability of the result is less than 0.05, given the null hypothesis, then we reject the null hypothesis.' But we can't. Recall from Chapter 2 that the probability of obtaining *any particular result* out of all of the possibilities is very low (see page 25), and varies with the number of possibilities under consideration. This value does not generalize from experiment to experiment.

Although we cannot use the actual probability of the observed value, we can usefully ask *how much of the total probability lies beyond the observed value*, out into the tail of the distribution. In the discrete case (binomial distribution) this is a sum of probabilities, in the continuous case (normal distribution) an area under the curve. Call o_1 the observed value, o_2 the next value out, o_3 the next, and so on until we exhaust all the possibilities. The sum of these is the probability of a complex event, the probability of 'observing the value o_1 or a value more extreme.' (Once again, we couch our probability measure in terms of a range of values). This then is a measure, based directly on probability, of 'how far away' from the mean an observed value lies. The smaller this probability, the more extreme the value. We can now say, if this probability is less than 0.05, we reject the hypothesis. The technical name for this measure is the P-VALUE.

In short, the p-value of a statistical test is the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed.

Note that this is a CONDITIONAL PROBABILITY: it is the probability of observing a particular sample mean (or something more extreme) conditional on the assumption that the null hypothesis is true. We can write this conditional probability as $P(\text{Observed mean} \mid H_0)$, or even more succinctly as $P(\text{Data} \mid H_0)$. The p-value does *not* measure the probability of the null hypothesis given the data, $P(H_0 \mid \text{Data})$. There is a widespread misunderstanding that the p-value tells you the probability that the null hypothesis is true (in light of some observation); it doesn't. You can confirm easily that we cannot switch conditional probabilities. The probability of the streets being wet given that rain has fallen $P(\text{Wet Streets} \mid \text{Rain})$ (presumably close to 1) is not at all the same as the probability of rain having fallen given that the streets are wet $P(\text{Rain} \mid \text{Wet Streets})$. There are many reasons why the streets may be wet (street cleaning, burst water pipes, etc.), rain is just one of the possibilities. There is however a technical term that expresses the intuition about the hypothesis in light of the data: if $P(Data \mid H_0)$ is low, we say the LIKELIHOOD of the hypothesis is low.

How do we determine this p-value? We simply integrate the area under the normal curve, going out from our observed value. (Recall that, for the present, we are assuming we *know* the population parameter σ, so it is appropriate to use the normal distribution). We actually have two completely equivalent ways to do this, since we now have two values (the actual observed value and its z-score), and two corresponding curves (the sampling distribution where the statistic is the sample mean, and the sampling distribution where the statistic is the standardized mean, the 'z-statistic'). We have seen what the sampling distribution of the sample mean looks like, assuming the null hypothesis is true (i.e. $\mu_0 = 70$, Figure 3.11). What is the sampling distribution of the z-statistic under this hypothesis? Let's do a simulation to find out.

In Figure 3.12, we repeat the simulation of sample means that we carried out at the beginning of the chapter, but now using the parameters of our

current null hypothesis $\mu_0 = 70$, $\sigma = 4$, sample size $= 11$. But in addition, for each sample we also compute the z-statistic, according to the formula provided above. We also include the corresponding normal curves for reference (recall these represent the limiting case of the simulations). As you can see, the distribution of the z-statistic is normal, with mean $= 0$, and SD $= 1$. A normal distribution with precisely these parameters is known as the STANDARDIZED NORMAL DISTRIBUTION.

```
> sample.means <- rep(NA, 1000)
> zs <- rep(NA, 1000)
> for (i in 1:1000) {
      sample.11 <- rnorm(11, mean = 70, sd = 4)
      sample.means[i] <- mean(sample.11)
      zs[i] <- (mean(sample.11) - 70)/(4/sqrt(11))
  }
> multiplot(2, 2)
> sd.dist <- 4/sqrt(11)
> plot(density(sample.means), xlim = range(70 -
      (4 * sd.dist), 70 + (4 * sd.dist)), xlab = "",
      ylab = "", main = "")
> plot(density(zs), xlim = range(-4, 4), xlab = "",
      ylab = "", main = "")
> plot(function(x) dnorm(x, 70, 4/sqrt(11)), 70 -
      (4 * sd.dist), 70 + (4 * sd.dist), xlab = "",
      ylab = "", main = "")
> plot(function(x) dnorm(x, 0, 1), -4, 4, xlab = "",
      ylab = "", main = "")
```

The crucial thing to note is that the area from either value out to the edge, which is the probability of interest, is precisely the same in the two cases, so we can use either. It is traditional to work with the standardized values, for reasons that will become clear.

Recall the z-score for our actual observation was -8.291562. This is an extreme value, well beyond 2 SDs from the mean, so we would expect there to be very little probability between it and the left tail of the distribution. We can calculate it directly by integration:

```
> integrate(function(x) dnorm(x, mean = 0, sd = 1),
      -Inf, -8.291562)

5.588542e-17 with absolute error < 4.5e-24
```

This yields a vanishingly small probability. We also get precisely the same result using the actual observed sample mean with the original sampling distribution:

```
> integrate(function(x) dnorm(x, mean = 70, sd = 4/sqrt(11)),
      -Inf, 60)
```

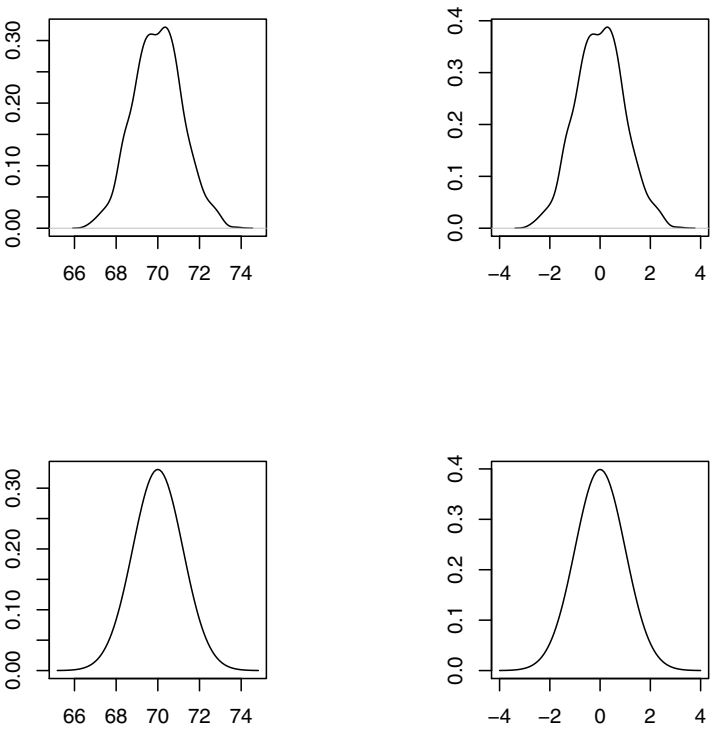

Fig. 3.12 The sampling distribution of the sample mean (left) and its z-statistic (right).

```
5.588543e-17 with absolute error < 6.2e-20
```

Suppose now we had observed a sample mean of 67.58. This is much closer to the hypothetical mean of 70. The standardized value here is almost exactly −2.0:

```
> (67.58 - 70)/(4/sqrt(11))

[1] -2.006558
```

Integrating under the standardized normal curve we find the following probability:

```
> integrate(function(x) dnorm(x, 0, 1), -Inf, -2)

0.02275013 with absolute error < 1.5e-05
```

This accords well with our rule-of-thumb. About 95% of the probability is within 2 SD of the mean. The remainder is split into two, one at each end of the distribution, each representing a probability of about 0.025 (actually a little less).

3.15 Hypothesis Testing: A More Realistic Scenario

In the above example we were able to use the standard deviation of the sampling distribution $\sigma_{\bar{x}}$, because we were given the standard deviation of the population σ. As we remarked earlier, in the real world we usually do not know σ, it's just another unknown parameter of the population. Just as in the case of computing real world confidence intervals, instead of σ we use the unbiased estimator s; instead of $\sigma_{\bar{x}}$ we use the unbiased estimator $SE_{\bar{x}}$; instead of the normal distribution we use the t-distribution.

Recall the z-score:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \tag{3.14}$$

$$= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{3.15}$$

And recall our formal definition of a statistic: a number that describes some aspect of the sample. Using this definition, the z-score seems to fail as a statistic, since it makes reference to a population *parameter* σ. But if we now replace that parameter with an estimate s derived from the sample itself, we get the so-called t-statistic:

$$t = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \tag{3.16}$$

$$= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{3.17}$$

This then can also be interpreted as yet another sampling statistic, with its own distribution. What does the distribution for *this* statistic look like? This is exactly the question William Gossett posed in 1908, in the paper that introduced the t-distribution (Gossett published under the pseudonym 'Student', and the statistic is often referred to as Student's t). Interestingly, before he was able to develop a precise mathematical formulation, he used simulation to get a feel for the distribution, as he explains (Student, 1908, 13):

> Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals [...] The mea-

> surements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample...

In the spirit of Student then, let's simulate the sampling distribution of the t-statistic and compare it to the distribution of the z-statistic shown above (Figure 3.13). Like Student, we use a small sample size of 4, where the effect of estimating σ by s is more noticeable. And again, we include the limiting-case normal and t-curves for reference. Notice the subtle but real effect of an increase in probability in the tails of the t-distribution. More of the probability is located further from the mean, reflecting the increase in uncertainty due to the fact that we are *estimating* the standard deviation of the population.

```
> zs <- rep(NA, 10000)
> ts <- rep(NA, 10000)
> for (i in 1:10000) {
       sample.4 <- rnorm(4, mean = 70, sd = 4)
       zs[i] <- (mean(sample.4) - 70)/(4/sqrt(4))
       ts[i] <- (mean(sample.4) - 70)/(sd(sample.4)/sqrt(4))
   }
> multiplot(2, 2)
> plot(density(zs), xlim = range(-4, 4), xlab = "z-scores",
       ylab = "", main = "Sampling distribution of z")
> plot(density(ts), xlim = range(-4, 4), xlab = "t-scores",
       ylab = "", main = "Sampling distribution of t")
> plot(function(x) dnorm(x, 0, 1), -4, 4, xlab = "x",
       ylab = "", main = "Limiting case: normal distribution")
> plot(function(x) dt(x, 3), -4, 4, xlab = "x",
       ylab = "", main = "Limiting case: t-distribution")
```

As discussed earlier, there is a family of t-curves corresponding to the different sample sizes; the t-curve is almost identical to the standardized normal curve, especially at larger sample sizes.

Note a rather subtle point: we can have samples with the same mean value, 67.58 say, but different t-scores, since the SD s of the samples may differ. In order to facilitate comparison with the z-score situation, let's suppose that we just happen to find a sample whose SD s is identical to the population SD $\sigma = 4$. In just such a case the t-score would be identical to the z-score, but the probability associated with the score will differ slightly, since we use the t-distribution, not the normal distribution.

The t-score here will be:

```
> (67.58 - 70)/(4/sqrt(11))

[1] -2.006558
```

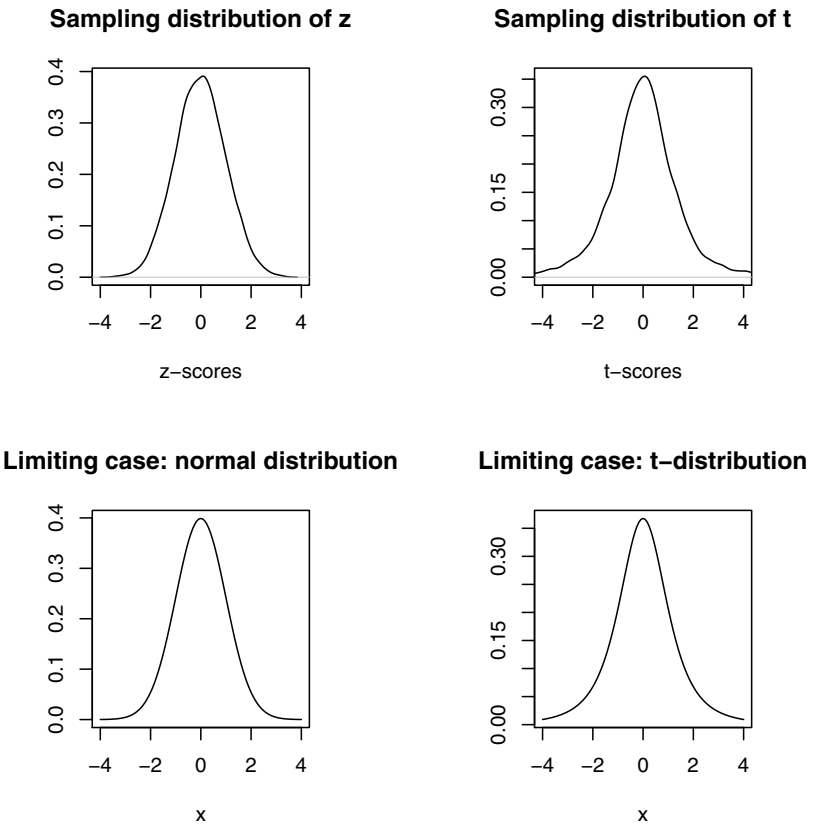

Fig. 3.13 The sampling distribution of the z-statistic (left) and the t-statistic (right).

And so for a sample size of 11 (degrees of freedom 10) we have the corresponding probability:

```
> integrate(function(x) dt(x, 10), -Inf, -2.006558)

0.03629508 with absolute error < 3.8e-06
```

Notice this probability is 'less extreme' than when we use the normal curve. The evidence against the null hypothesis is not as strong, reflecting the fact that we are uncertain of the true SD of the sampling distribution. This is the uncertainty built into the t-curve, with more of its probability in the tails.

Once again we note that, as sample size increases, the estimate s improves, and the more our results resemble the normal curve:

```
> integrate(function(x) dt(x, 20), -Inf, -2.006558)
```

```
0.02925401 with absolute error < 3.8e-06
```

Note that our null hypothesis H_0 was that the observed mean $\bar{x}$ is equal to the hypothesized mean μ_0. Thus, rejecting the null hypothesis amounts to accepting the alternative hypothesis, i.e., that the observed value is less than the mean *or* the observed value is greater than the mean:

$$H_a : \bar{x} < \mu_0 \text{ or } \mu_0 < \bar{x} \tag{3.18}$$

This means that as evidence for rejection of H_0 we will count extreme values on *both* sides of μ. For this reason, the above test is called a TWO-SIDED SIGNIFICANCE TEST (also known as the TWO-TAILED SIGNIFICANCE TEST). Note that if we simply reported the probability corresponding to the t-value t, we would *not* be reporting the probability of 'a value being more than t away' from the mean, but the probability in one direction only. For that reason, in a two-sided test, since the distributions are symmetrical, the p-value will be twice the value of the probability corresponding to the particular t-value we obtain. If the p-value is $\leq \alpha$, we say that the data are significant at level α. Purely by convention, $\alpha = 0.05$.

By contrast, if our null hypothesis were that the observed mean $\bar{x}$ is, say, equal to or less than the hypothesized mean μ_0, then the alternative hypothesis would be:

$$H_a : \mu_0 < \bar{x} \tag{3.19}$$

In this situation, we would use a one-sided significance test, reporting the probability in the relevant direction only.

R does everything required for a t-test of significance as follows, and you can specify (inter alia) what your μ_0 is (note that it need not be zero), whether it is two-sided or not (see the documentation for the `t.test` for details on how to specify this), and the confidence level (the α level) you desire, as follows:

```
> sample.11 <- rnorm(11, mean = 60, sd = 4)
> t.test(sample.11, alternative = "two.sided",
       mu = 70, conf.level = 0.95)

        One Sample t-test

data:  sample.11
t = -8.8974, df = 10, p-value = 4.587e-06
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 56.67570 62.01268
sample estimates:
mean of x
  59.3442
```

Experiment with the above code: change the hypothetical mean, change the mean of the sampled population and its SD, change the sample size, etc. In each case, see how the sample mean, the t-score, the p-value and the confidence interval differ.

It is also instructive to keep the parameters the same and simply repeat the experiment, taking different random samples each time (effectively, REPLICATING the experiment). Watch how the p-values change, watch how they change from replicate to replicate under different parameter settings. Do you ever find you would accept the null hypothesis when it is in fact false? How likely is it that you would make a mistake like that? This is an issue we will return to in more depth in Chapter 4.

The t-value we see above is indeed the t in equation 3.16; we can verify this by doing the calculation by hand:

```
> (mean(sample.11) - 70)/se(sample.11)

[1] -8.897396
```

3.16 Comparing Two Samples

In one-sample situations our null hypothesis is that there is no difference between the sample mean and the population mean:

$$H_0 : \bar{x} = \mu \tag{3.20}$$

When we compare samples from two different populations, we ask the question: are the population means identical or not? Our goal now is to figure out some way to define our null hypothesis in this situation.

Consider this example of a common scenario in experimental research. Mean reading times and standard deviations are available for children and adults reading English sentences. Let us say that we want to know whether children are faster or slower than adults in terms of reading time. You probably don't need to do an experiment to answer this question, but it will do as an illustration of this type of experiment.

We know that, due to the nature of random sampling, there is bound to be *some* difference in sample means even if the population means are identical. We can reframe the research question as follows: is the difference observed in the two sample means a true difference or just a chance event? The data are shown in Table 3.2.

Notice a few facts about the data. We have different sample sizes in each case. How will that affect our analysis? Notice too that we have different standard deviations in each case: this makes sense, since children exhibit a wider range of abilities than literate adults. But we now know how great an effect the variability of the data has on statistical inference. How will

Table 3.2 Hypothetical data showing reading times for adults and children.

group	sample size n	$\bar{x}$ (secs)	s
children	$n_1 = 10$	$\bar{x}_1 = 30$	$s_1 = 43$
adults	$n_2 = 20$	$\bar{x}_2 = 7$	$s_2 = 25$

we cope with these different SD's? Finally, the mean reading times certainly 'look' significantly different, but are we sure the difference is beyond the realm of chance, and if so, can we say exactly how much?

Such research problems have the properties that (i) the goal is to compare the responses in two groups; (ii) each group is considered a sample from a distinct population (a 'between-subjects' design); (iii) the responses in each group are independent of those in the other group; and (iv) the sample sizes of each group can be different.

The question now is, how can we formulate the null hypothesis?

3.16.1 H_0 in Two-sample Problems

Let us start by saying that the unknown population mean of children is μ_1, and that of adults is μ_2. We can state our null hypothesis as follows:

$$H_0 : \mu_1 = \mu_2 \tag{3.21}$$

Equivalently, we can say that our null hypothesis is that the difference between the two means is zero:

$$H_0 : \mu_1 - \mu_2 = 0 = \delta \tag{3.22}$$

We have effectively created a new population parameter δ:

$$H_0 : \delta = 0 \tag{3.23}$$

We can now define a new statistic $d = \bar{x}_1 - \bar{x}_2$ and use that as an estimate of δ, which we've hypothesized to be equal to zero. But to do this we need a sampling distribution of the difference of the two sample means $\bar{x}_1$ and $\bar{x}_2$.

Let's do a simulation to get an understanding of this approach. For simplicity we will use the sample means and standard deviations from the example above as our population parameters in the simulation, and we will also use the sample sizes above for the repeated sampling. Assume a population with $\mu_1 = 30$, $\sigma_1 = 43$, and another with mean $\mu_2 = 7$, $\sigma_2 = 25$. So we already know in this case that the null hypothesis is false, since $\mu_1 \neq \mu_2$. But let's take 1000 sets of samples of each population, compute the differences in mean in each

set of samples, and plot that distribution of *the differences of the sample mean*:

```
> d <- rep(NA, 1000)
> for (i in 1:1000) {
       sample1 <- rnorm(10, mean = 30, sd = 43)
       sample2 <- rnorm(20, mean = 7, sd = 25)
       d[i] <- mean(sample1) - mean(sample2)
   }
```

Note that the mean of the differences-vector `d` is close to the true difference:

```
> 30 - 7

[1] 23

> mean(d)

[1] 23.09159
```

Then we plot the distribution of `d`; we see a normal distribution (Figure 3.14).

```
> hist(d)
```

So, the distribution of the differences between the two sample means is normally distributed, and centered around the true difference between the two populations. It is because of these properties that we can safely take d to be an unbiased estimator of δ. How accurate an estimator is it? In other words, what is the standard deviation of this new sampling distribution? It is clearly dependent on the standard deviation of the two populations in some way:

$$\sigma_{\bar{x}_1-\bar{x}_2} = f(\sigma_1, \sigma_2) \tag{3.24}$$

(Try increasing one or other or both of the σ in the above simulation to see what happens). The precise relationship is fundamentally additive: instead of taking the root of the variance, we take the root of the sum of variances:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{43^2}{10} + \frac{25^2}{20}} = 14.702. \tag{3.25}$$

```
> newsigma <- round(sqrt((43^2/10) + (25^2/20)),
       digits = 4)
```

In our single sample, $\bar{x}_1 - \bar{x}_2 = 17$. The null hypothesis is $\mu_1 - \mu_2 = 0$. How should we proceed? Is this sample difference sufficiently far away from the hypothetical difference (0) to allow us to reject the null hypothesis? Let's first translate the observed difference 17 into a z-score. Recall how the z-score is calculated:

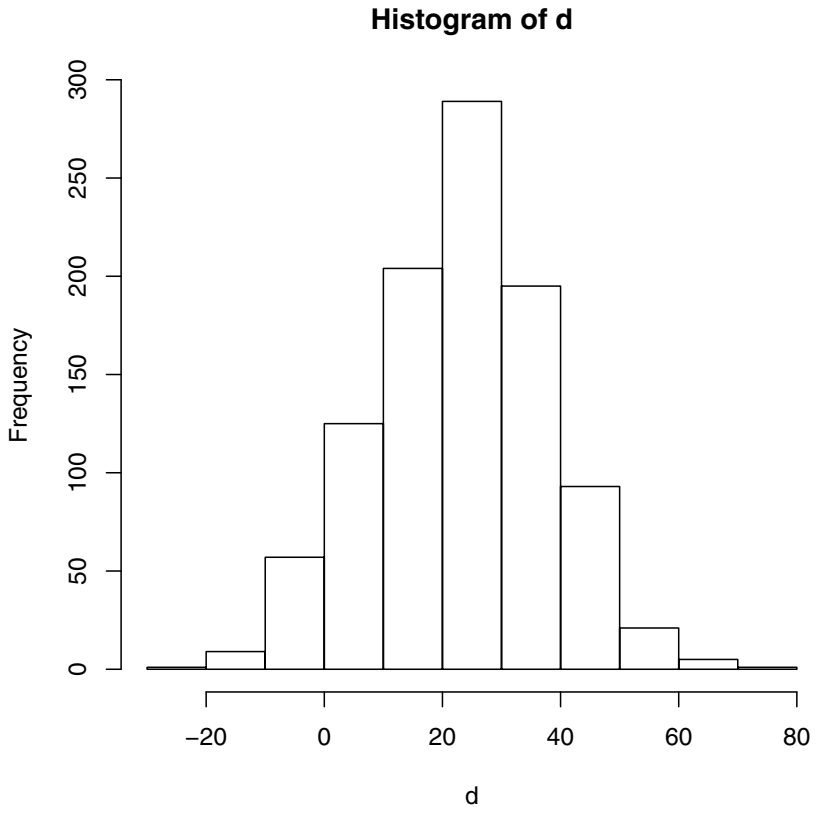

Fig. 3.14 The distribution of the difference of sample means of two samples.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\text{sample mean} - \text{pop. mean}}{\text{sd of sampling distribution}} \tag{3.26}$$

If we replace $\bar{x}$ with d, and the new standard deviation from the two populations' standard deviations, we are ready to work out the answer:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.27}$$

$$= \frac{17 - 0}{14.702} \tag{3.28}$$

$$= 1.1563 \tag{3.29}$$

Using exactly the same logic as previously, because we don't know the population parameters in realistic settings, we replace the σ's with the sample standard deviations to get the t-statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.30}$$

This is the TWO-SAMPLE t-STATISTIC.

So far so good, but we want to now translate this into a p-value, for which we need the appropriate t-curve. The problem we face here is that the degrees of freedom needed for the correct t-distribution are not obvious. The t-distribution assumes that only one s replaces a single σ; but we have two of these. If $\sigma_1 = \sigma_2$, we could just take a *weighted average* of the two sample SDs s_1 and s_2.

In our case the correct t-distribution has $n_1 - 1 + n_2 - 1$ degrees of freedom (the sum of the degrees of freedom of the two sample variances; see Rice, 1995, 422 for a formal proof).

In real life we don't know whether $\sigma_1 = \sigma_2$. One response would be to err on the side of caution, and simply use degrees of freedom corresponding to the smaller sample size. Recall that smaller degrees of freedom reflect greater uncertainty, so the estimate we get from this simple approach will be a conservative one.

However, in a more sophisticated approach, something called Welch's correction corrects for possibly unequal variances in the t-curve. R does this correction for you if you specify that the variances are to be assumed to be unequal (`var.equal=FALSE`).

```
> t.test.result <- t.test(sample1, sample2, mu = 0,
       alternative = "two.sided", conf.level = 0.95,
       var.equal = FALSE)
```

If you print out the contents of `t.test.result`, you will see detailed output. For our current discussion it is sufficient to note that the t-value is 1.35, the degrees of freedom are 12.1 (a value somewhere between the two sample sizes), and the p-value is 0.2. Recall that every time you run the t-test with newly sampled data (you should try this), your results will be slightly different; so do not be surprised if you occasionally fail to find a significant difference between the two groups even though you already know that in reality there is such a difference. We turn to this issue in the next chapter.

Problems

3.1. Choose one answer: 95% confidence intervals describe:

a. The range of individual scores
b. Plausible values for the population mean
c. Plausible values for the sample mean
d. The range of scores within one standard deviation

3.2. A 95% confidence interval has a ?% chance of describing the sample mean:

a. 95%
b. 100%

3.3. For the same data, a 90% CI will be wider than a 95% CI.

a. True
b. False

3.4. True or False?

a. The p-value is the probability of the null hypothesis being true.
b. The p-value is the probability that the result occurred by chance.