Journal of Statistics Education

Teaching Inference for Randomized Experiments

Michael D. Ernst
St. Cloud State University

Journal of Statistics Education Volume 17, Number 1 (2009), www.amstat.org/v17n1/ernst.html

---

Key Words: ANOVA; Computing; Curriculum; Normal distribution; Randomization distribution; Randomization test; Sampling distribution; Two-sample *t*-test.

Abstract

Nearly all introductory statistics textbooks include a chapter on data collection methods that includes a detailed discussion of both random sampling methods and randomized experiments. But when statistical inference is introduced in subsequent chapters, its justification is nearly always based on principles of random sampling methods. From the language and notation that is used to the conditions that students are told to check, there is usually no mention of randomized experiments until an example that is a randomized experiment is encountered, at which point the author(s) may offer a statement to the effect of "the randomization allows us to view the groups as independent random samples." But a good student (or even an average one) should ask, "Why?"

This paper shows, in a way easily accessible to students, why the usual inference procedures that are taught in an introductory course are often an appropriate approximation for randomized experiments even though the justification (the Central Limit Theorem) is based entirely on a random sampling model.

1. Introduction

Consider how your introductory statistics students would solve the following exercise, which is taken from a previous version of Moore (2004, pg. 448):

> **Is red wine better than white wine?** Observational studies suggest that moderate use of alcohol reduces heart attacks, and that red wine may have special benefits. One reason may be that red wine contains polyphenols, substances that do good things to cholesterol in the blood and so may reduce the risk of heart attacks. In an experiment, healthy men were assigned at random to drink half a bottle of either red or white wine each day for two weeks. The level of polyphenols in their blood was measured before and after the two-week period. Here are the percent changes in level for the subjects in both groups:

Red	3.5	8.1	7.4	4.0	0.7	4.9	8.4	7.0	5.5
White	3.1	0.5	−3.8	4.1	−0.6	2.7	1.9	−5.9	0.1

Is there good evidence that red wine drinkers gain more polyphenols on the average than white wine drinkers?

If you teach from just about any introductory textbook, your students will most likely choose to do a two-sample *t*-test. You hope that they also check the necessary conditions for using the two-sample *t*-test. Moore (2008, pg. 462) gives these conditions along with some advice on how to check them:

1. We have two simple random samples (SRSs), from two distinct populations. The samples are independent.
2. Both populations are normally distributed. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

How would your students evaluate those conditions for the exercise above? Does the fact that the stem of the problem contains the word "random" convince them that the first condition is satisfied? Do your students read right over the word "populations" in the second condition and focus on the "In practice" part? Or, do they correctly recognize that these are not random samples from any type of populations, and so neither of these conditions are even close to being satisfied?

At this point in the course, you have probably already covered, in some detail, both random sampling methods and randomized experiments. You may have even convinced your students that, when possible, randomized experiments are preferable. And yet, when you get to inference, a student could easily get the impression that you cannot use a *t*-test on data from a randomized experiment since these are not random samples from any populations.

So what do we tell our students? If textbook authors say anything at all, they make a statement similar to Moore (2008, pg. 463), that "[b]ecause of the randomization, we are willing to regard the two groups ... as two independent SRSs." My hope is that students (and teachers) would not take such a statement at face value. My fear is that students (and teachers?) interpret that statement to mean that observational and experimental data aren't that different after all.

The problem here comes from the fact that the theory of inference, as it's developed in an introductory textbook, is based on random sampling (sampling distributions), and yet many of the most interesting examples come from randomized experiments. This paper describes, using examples that are accessible to students (and teachers), why the usual random sampling-based inference procedures are often an appropriate approximation for randomized experiments. Section 2 briefly illustrates inference for two random samples. Section 3 introduces the analogous inference for randomized experiments. Section 4 compares and contrasts these two methods and generalizes our findings. Section 5 considers one of the most common methods used for randomized experiments: analysis of variance. Section 6 summarizes the paper and draws some conclusions.

2. Inference for Random Samples

To make the illustration easier for students, I prefer to start with a simpler example (as you will see, by "simpler" I mean "smaller sample size") like the following one taken from Moore (2008, pg. 463):

How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then

dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study buried 10 strips of polyester fabric in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:

2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

The null and alternative hypotheses of interest are H_0: $\mu_2 = \mu_{16}$ and H_1: $\mu_2 > \mu_{16}$, where the alternative hypothesis indicates that there is more decay in 16 weeks. If we are willing to pretend that these data are random samples from two populations, then we should check the normality condition. With such small sample sizes, we are mostly looking for outliers. The histogram in Figure 1 shows no "strong outliers," convincing us that the normality condition is reasonable and we can do a two-sample *t*-test.

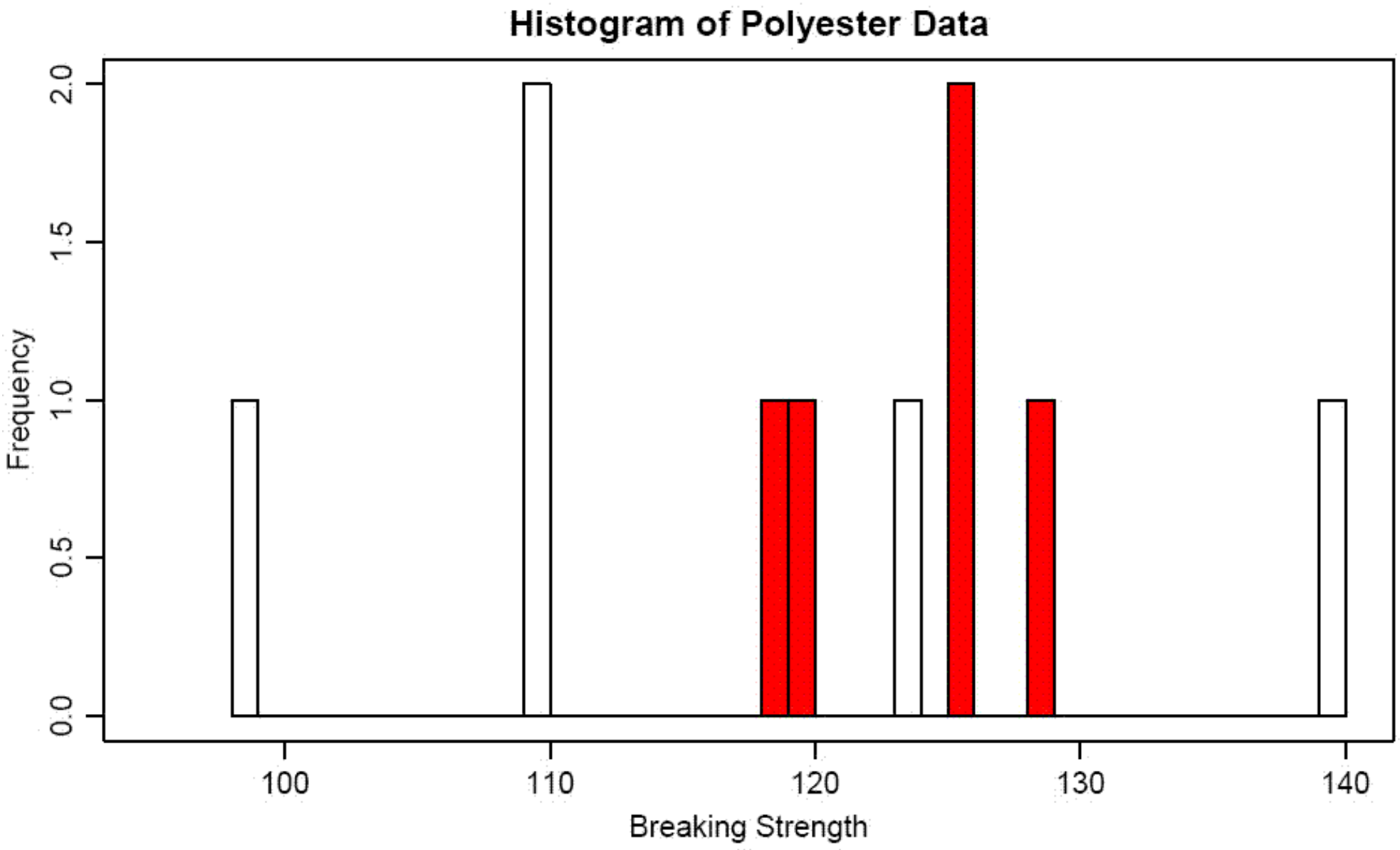

Figure1: Breaking strengths of 10 pieces of polyester buried for 2 weeks (red) and 16 weeks (white).

A few calculations show that t=0.99 and the p-value is 0.1857. Now, think a little harder about what that p-value represents. It is the probability, under H_0, of getting a t-statistic equal to or larger than the observed value. This is the area in the tail of the t distribution to the right of t=0.99. Why do we use the t distribution? Because it is the appropriate sampling distribution. That is, it is the distribution of the t-statistic in all possible random samples from two identical normal distributions.

3. Inference for Randomized Experiments

The random sampling model uses randomness in the sampling scheme. By considering all possible samples in that scheme, the *sampling distribution* is obtained. A randomized experiment uses randomness differently -- in the assignment of treatments to the subjects. So it is natural to consider all possible assignments of the treatments to the subjects. If H_0 is true, the fabric's breaking strength will be the same regardless of the treatment it is assigned. Calculating the *t*-statistic for each of the possible randomizations results in the *randomization distribution*. I illustrate this to students with a simple piece of R code (see Appendix A) that randomizes the 10 pieces of fabric (and their breaking strengths) into two groups of 5 and calculates the *t*-statistic. Repeating this multiple times shows students how each possible randomization gives a different result. Figure 2 shows three such randomizations.

```
> randomize(two,sixteen)
$two
[1] 110 126 120 110 118
$sixteen
[1] 126 129 124  98 140
$t.statistic
       t
-0.87114
```

```
> randomize(two,sixteen)
$two
[1] 124 126 129 110 140
$sixteen
[1] 118 126 120  98 110
$t.statistic
        t
1.669982
```

```
> randomize(two,sixteen)
$two
[1] 120 124 126 110 129
$sixteen
[1] 118 126  98 110 140
$t.statistic
         t
0.4339051
```

Figure 2: Three possible randomizations of the polyester data with *t*-statistics.

Next we can consider all possible randomizations by listing them systematically. Students quickly see that there are a lot. Even in this small data set there are $\binom{10}{5} = 252$ possible randomizations, from which a *t*-statistic can be calculated for each. This randomization distribution is pictured in Figure 3.

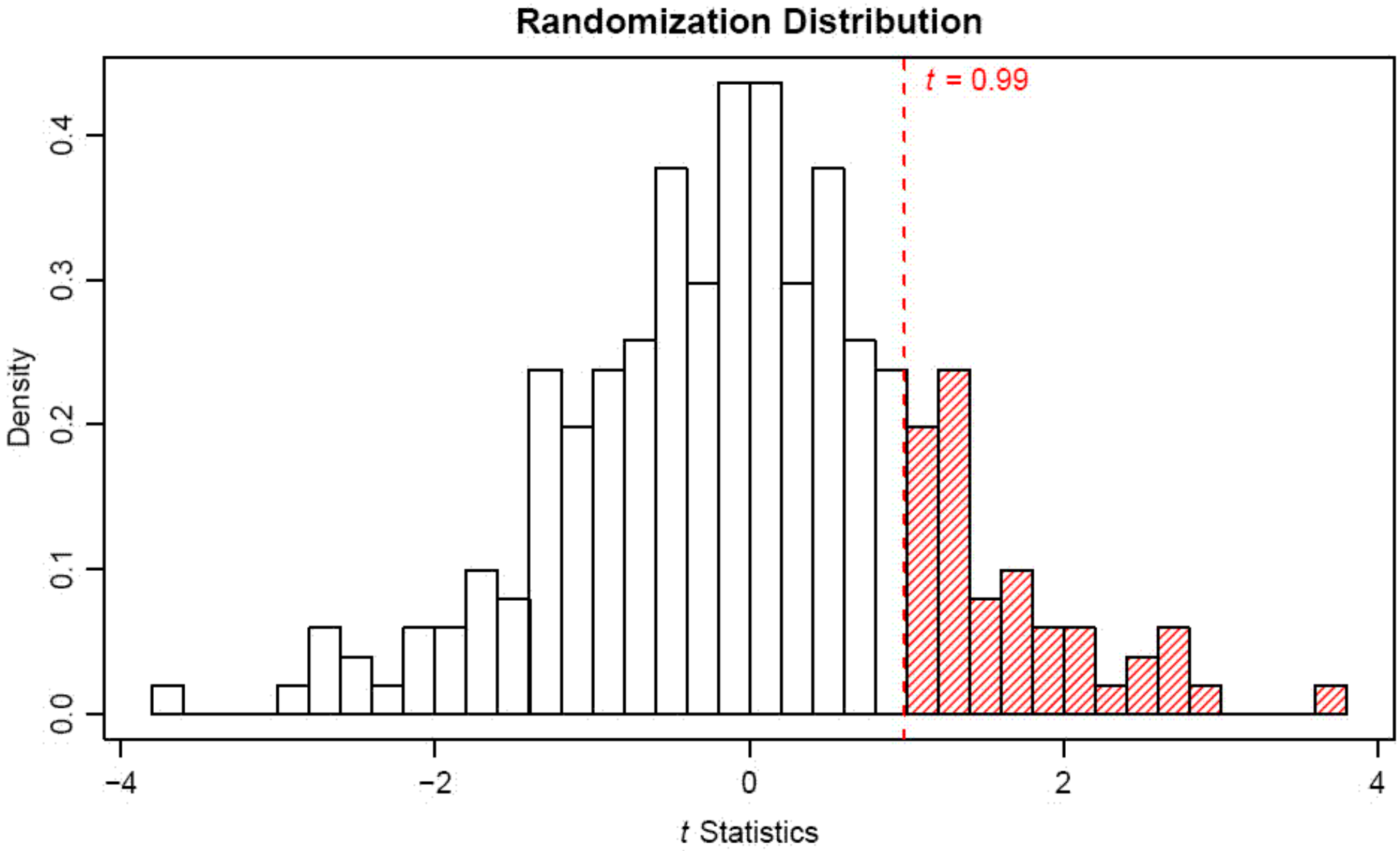

Figure 3: Randomization distribution of the *t*-statistic for the polyester data.

Analogous to the sampling distribution, the *p*-value in a randomization distribution is simply the proportion of *t*-statistics equal to or greater than the observed value (*t*=.99 in this case). The red shaded area in Figure 3 represents this *p*-value, which is 0.1865.

4. When Do They Agree and When Do They Not?

Overlaying the randomization distribution with the *t* distribution makes it clear that the sampling distribution is a good approximation to the randomization distribution.

In Figure 4, the blue shaded area represents the *p*-value calculated from the *t* distribution, which is very close to the correct *p*-value from the randomization distribution.

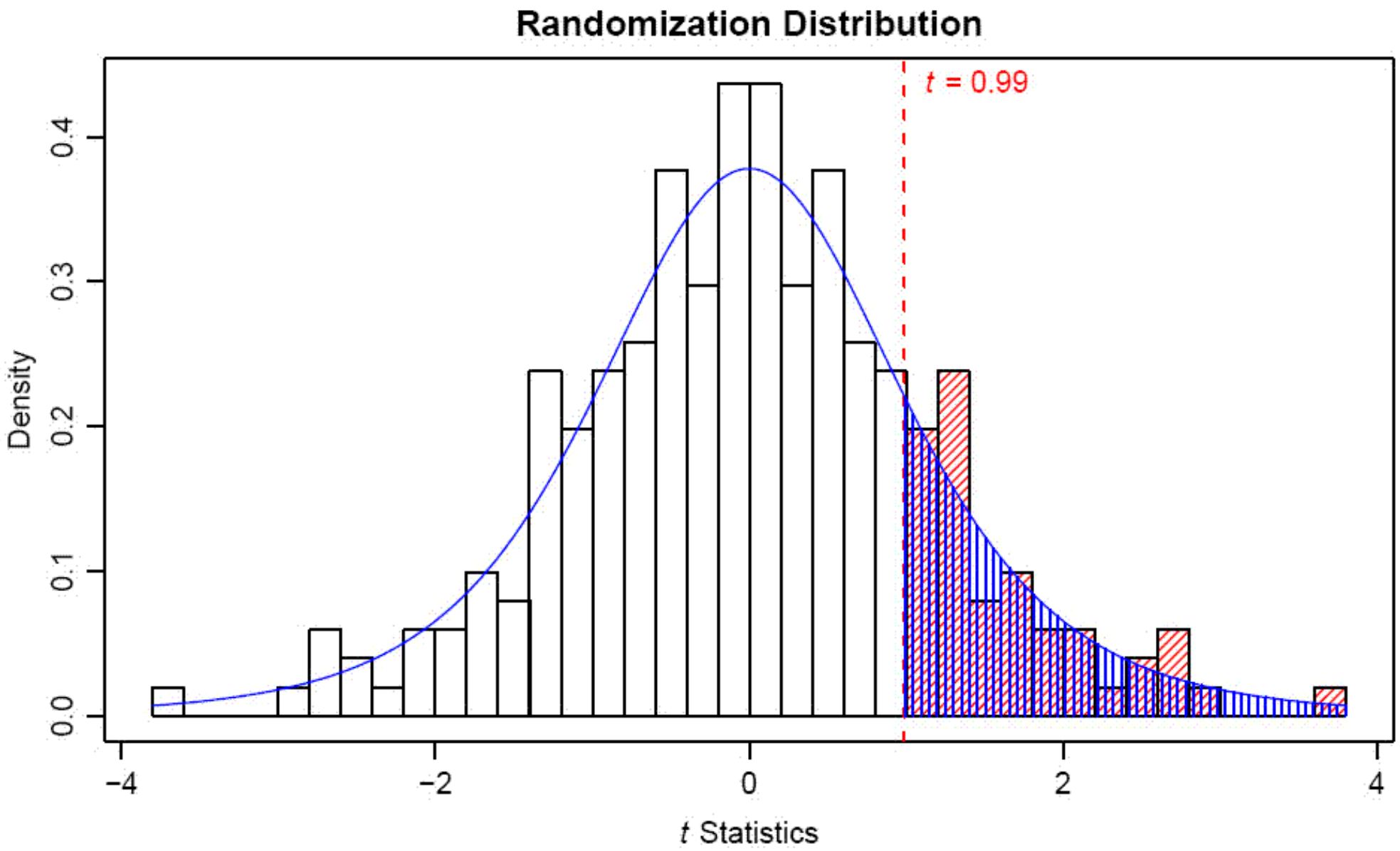

Figure 4: Randomization distribution of the *t*-statistic for the polyester data overlayed with the *t* distribution.

The question remains: When does the *t* distribution approximate the randomization distribution satisfactorily? Students can see the answer to that by looking at several examples. Suppose that the conditions required for the *t*-test were not met and the data had an outlier. In particular, suppose that the smallest breaking strength of 98 pounds was actually even smaller. Figure 5 shows the randomization distributions and *t* distributions for the polyester data when this smallest breaking strength is 88,

78, 68, and 58 pounds.

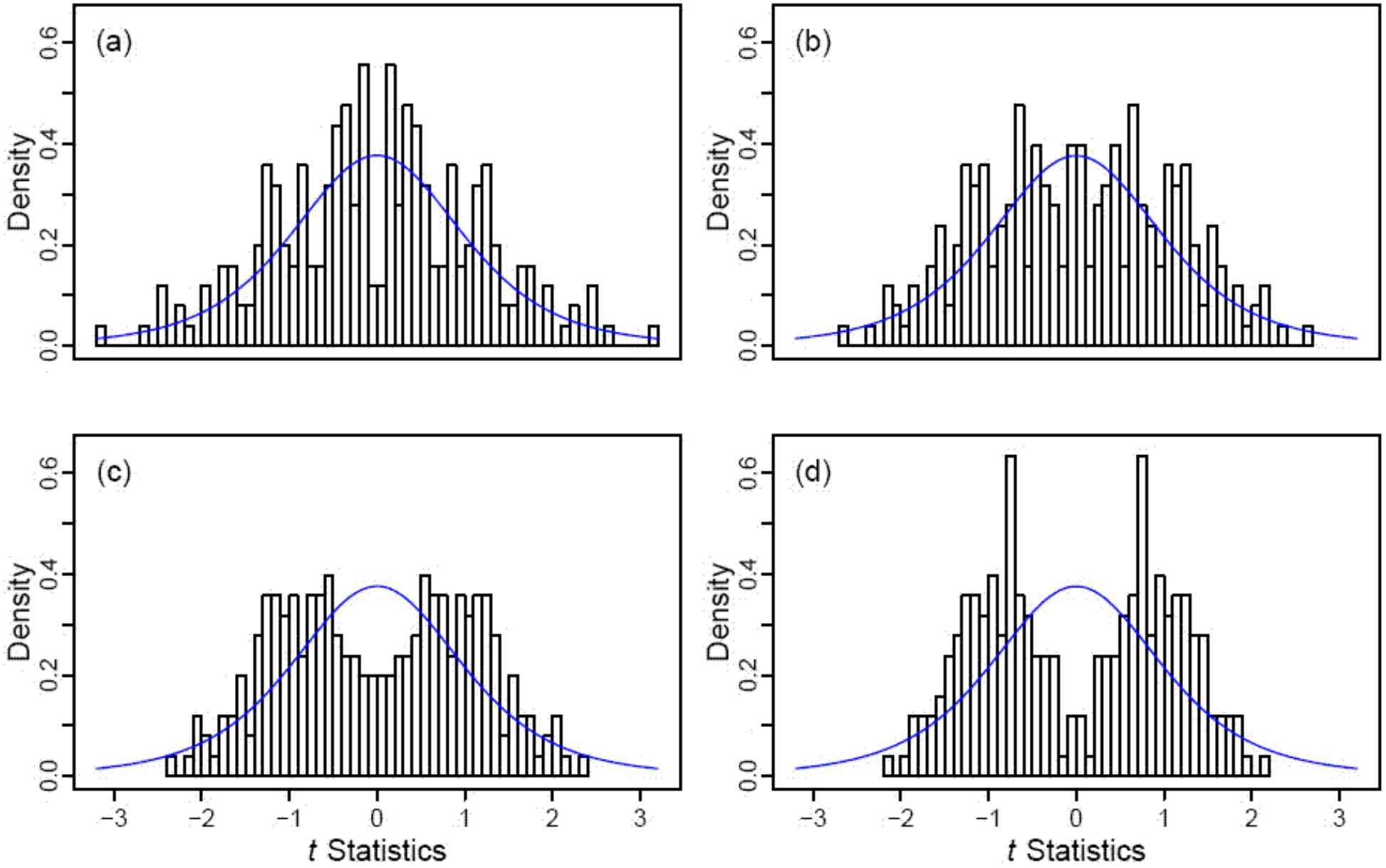

Figure 5: Randomization and *t* distributions for the polyester data when the smallest breaking strength is (a) 88, (b) 78, (c) 68, and (d) 58 pounds.

As this breaking strength becomes more of an outlier, the *t* distribution becomes less and less of a good approximation to the randomization distribution.

If the data had come from two random samples from two normal populations, this outlier would indicate a problem with the normality condition. In fact, we have data

from a randomized experiment, but the outlier is still indicating a problem -- with how well the *t* distribution will approximate the randomization distribution. And so students can see that the check of the normality condition for random sampling is actually relevant for randomized experiments. This helps justify Moore's (2008, pg. 463) statement that "[b]ecause of the randomization, we are willing to regard the two groups of fabric as two independent SRSs."

Returning to the red and white wine example, we can see how good the approximation can be even with moderate sample sizes. Figure 6 shows the randomization distribution for all $\binom{18}{9}$ = 48,620 possible randomizations and the approximate *t* distribution. The randomization *p*-value is 0.00068, while the *t*-test gives a *p*-value of 0.00085.

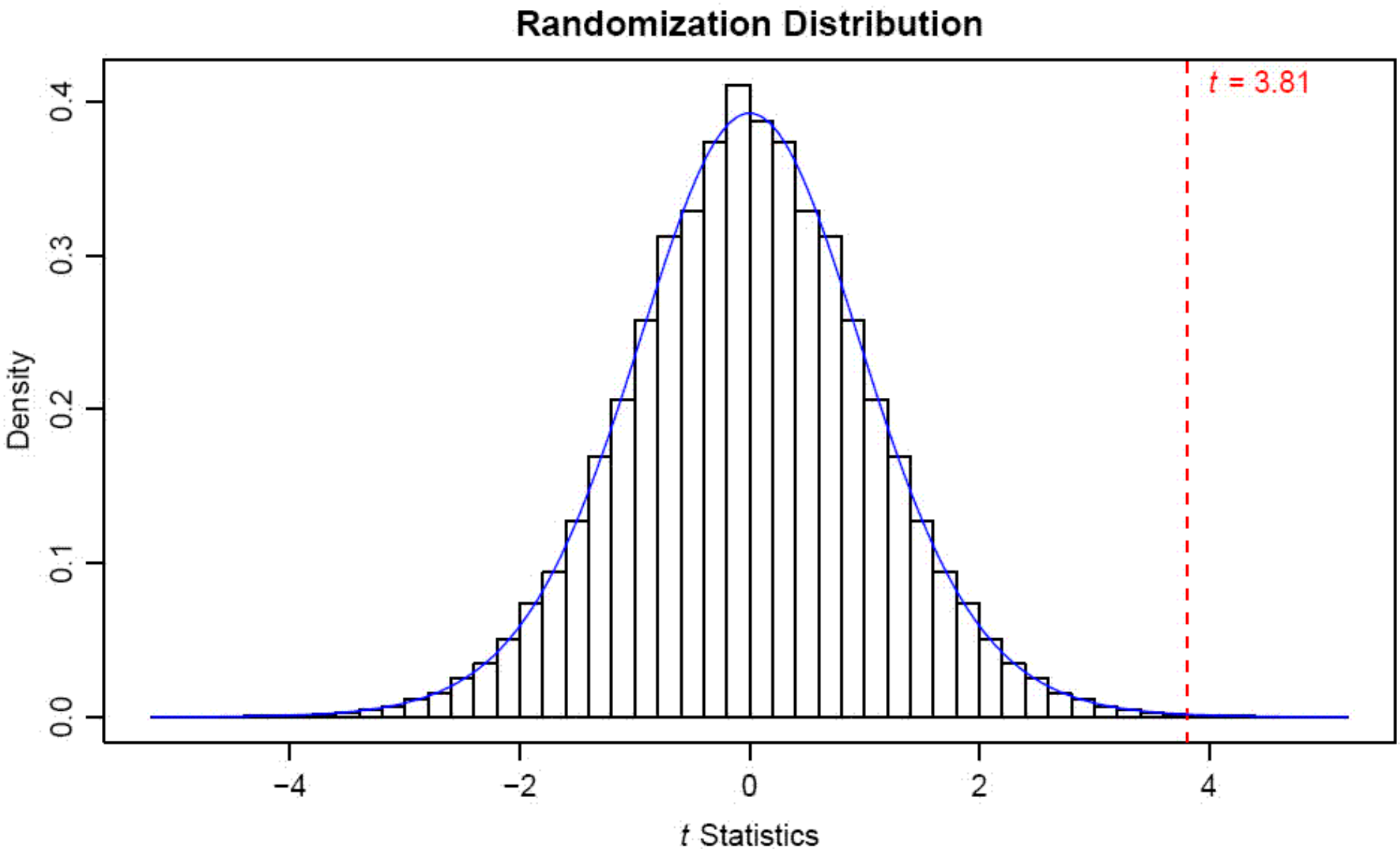

Figure 6: Randomization and *t* distributions for the red and white wine data.

5. Analysis of Variance

We can illustrate how well the normal theory procedure approximates the appropriate randomization test for a randomized experiment in other settings as well. One of the most commonly used procedures to analyze a randomized experiment is the analysis of variance (ANOVA). The theory behind the *F* distribution that is used for ANOVA is based on random sampling from normal populations. In a randomized experiment, the subjects are randomly assigned to the *k* groups. The randomization distribution of the *F*-statistic can be found by calculating the *F*-statistic for each possible randomization. The polyester example described earlier actually had four treatment groups -- two weeks, four weeks, eight weeks, and 16 weeks -- each with five strips of fabric (Moore 2008, pg. 653):

2 weeks	118	126	126	120	129
4 weeks	130	120	114	126	128
8 weeks	122	136	128	146	140
16 weeks	124	98	110	140	110

There are a total of $\binom{20}{5\,5\,5\,5} = 11,732,745,024$ possible randomizations of the 20 pieces of fabric into four equal size groups. This is too many to easily calculate, but we can get an accurate estimate of the randomization distribution by sampling from the possible randomizations (Dwass 1957). Figure 7 shows the randomization distribution of the ANOVA *F*-statistic from 1,000 (random) randomizations of the data. This randomization distribution is overlayed with the *F* distribution with 3 and 16 degrees of freedom that is used by the normal theory approximation. We see that the *F* distribution is a good approximation to the randomization distribution. The randomization *p*-value is 0.083 and the *p*-value from the *F* distribution is 0.084.

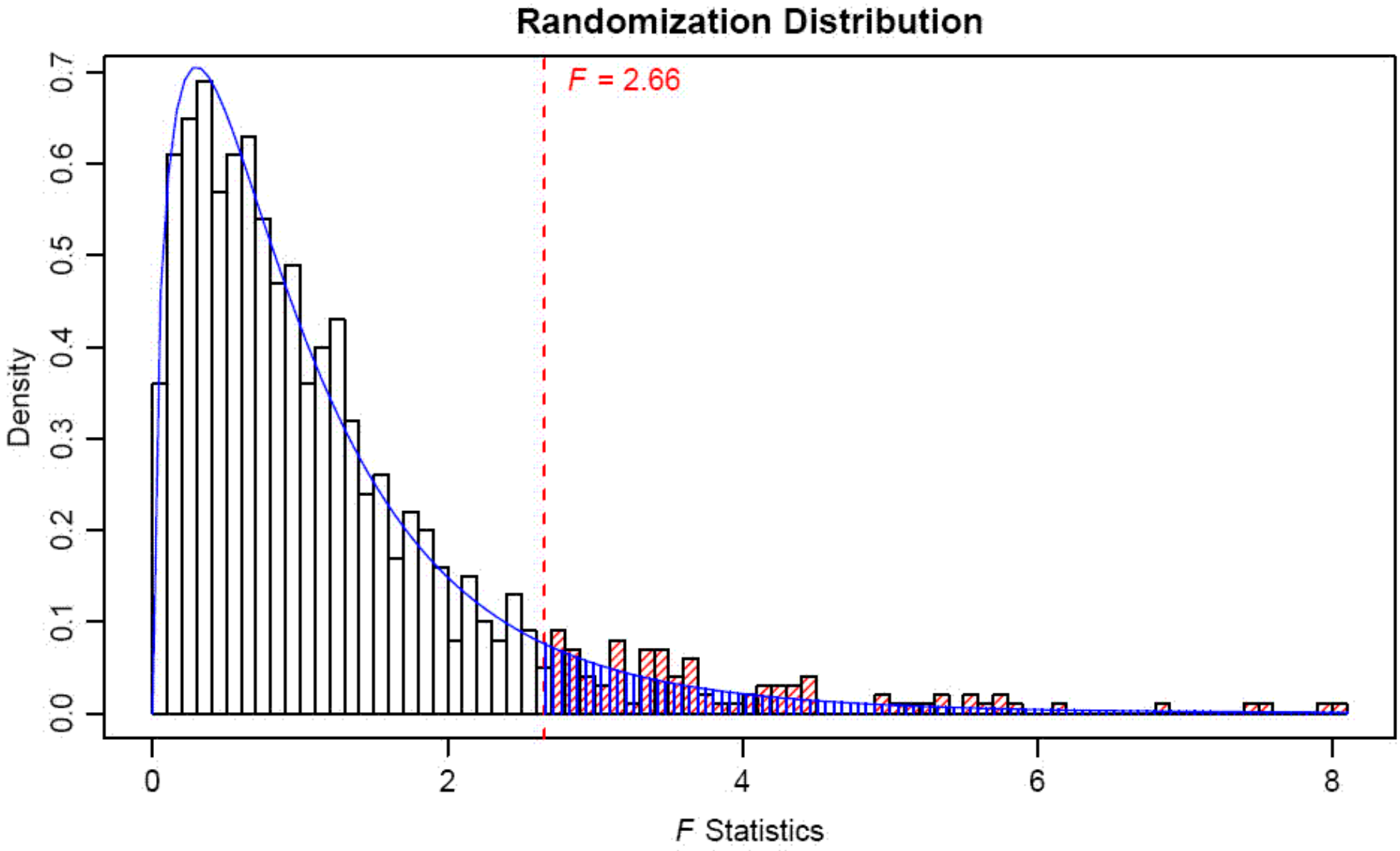

Figure 7: Randomization distribution of the *F*-statistic overlayed with the *F* distribution for the polyester data.

6. Conclusion

Introductory statistics textbooks have come a long way in the last 15 years by including much more information on data collection methods, including both random sampling methods and randomized experiments. But, the way that statistical inference is developed in nearly all introductory textbooks follows directly from random sampling arguments. Randomized experiments are nearly ignored and, at least from some textbooks, students could actually get the impression that the sampling-based methods cannot be used for randomized experiments. This paper gives an easy way to illustrate to students (1) the correct randomization approach to inference, and (2)

why the normal theory methods still work.

Some would argue that we should teach the normal theory methods only as an approximation to (and therefore after) the randomization methods, or possibly not at all (Cobb 2007). While I have some sympathy for such a position, my intent here is simply to provide some ideas to supplement an introductory course as it is most commonly taught. These ideas are easy to implement and not difficult for students to grasp. In addition, they provide a first step to move in the direction suggested by Cobb (2007).

Appendix A

All calculations and graphs in this paper were done in R (R Development Core Team 2008), which is freely available at http://www.R-project.org. This section contains the R code used for the computations in this paper. The function `randomize` produces randomizations of observations into two groups as shown in Figure 2.

```
randomize <- function(x,y){
   z <- c(x,y)
   xx <- z[ind <- sample(length(z),length(x))]
   yy <- z[-ind]
   result <- list(xx,yy,t.test(xx,yy)$statistic)
   names(result) <- c(deparse(substitute(x)),deparse(substitute(y)),"t.statistic")
   return(result)
}
```

The function `plot.rand` produces the randomization distribution of the two-sample *t*-statistic as shown in Figure 3, Figure 4, Figure 5 and Figure 6, as well as the randomization and *t* distribution *p*-values.

```
plot.rand <- function(x,y,shade=T,t.dist=T,t.shade=T,t.obs=T,
      breaks=50,pause=F,xlab=bquote(paste(italic(t)," Statistics")),
      main="Randomization Distribution",...){
m <- length(x)
n <- length(y)
z <- c(x,y)
S <- sum(c(x,y))
SS <- sum(c(x,y)^2)

calc.t <- function(x){
   Sx <- sum(x)
   Sy <- S - Sx
   SSx <- sum(x^2)
   SSy <- SS - SSx
   return((Sx/m-Sy/n)/sqrt((SSx-(Sx^2)/m)/(m*(m-1))+(SSy-(Sy^2)/n)/(n*(n-1))))
}
```

```
t.stat <- combn(z,m,calc.t)
t.test.obs <- t.test(x,y,alternative="greater")

hist.out <- hist(t.stat,probability=T,breaks=breaks,xlab=xlab,main=main,...)
box()
if(t.obs){
   abline(v=t.stat[1],lty=2,col=2)
   mtext(bquote(paste(italic(t)," = ",.(round(t.stat[1],2)))),
      3,-1,at=t.stat[1],adj=-.2,cex=.8,col=2)
}

if(shade){
   no.shade <- sum(hist.out$mids <= t.stat[1])
   if(pause) null <- locator(1)
   par(new=T)
   hist(t.stat,probability=T,breaks=hist.out$breaks,
      col=c(rep(0,no.shade),rep(2,length(hist.out$breaks)-no.shade-1)),
      axes=F,ann=F,add=T,density=30,border=1)
   if(t.obs) abline(v=t.stat[1],lty=2,col=2)
}

if(t.dist){
   r.br <- range(hist.out$breaks)
   xx <- seq(r.br[1],r.br[2],length=150)
   if(pause) null <- locator(1)
   lines(xx,dt(xx,t.test.obs$parameter),col=4)
   if(t.shade & t.dist){
     if(pause) null <- locator(1)
     yy <- xx[xx>t.stat[1]]
     yt <- dt(xx[xx>t.stat[1]],t.test.obs$parameter)
     lines(yy,yt,type="h",col=4)
   }
}

p.values <- c(mean(t.stat >= t.stat[1]),t.test.obs$p.value)
names(p.values) <- c("Randomization","t-test")
return(p.values)
}
```

The function `plot.rand.anova` produces the randomization distribution of the one-way analysis of variance *F*-statistic as shown in Figure 7, as well as the randomization and *F* distribution *p*-values.

```
plot.rand.anova <- function(x,grp,shade=T,F.dist=T,F.shade=T,F.obs=T,
```

```
      breaks=50,pause=F,B=1000,xlab=bquote(paste(italic(F)," Statistics")),
      main="Randomization Distribution",...){
if(length(x) != length(grp))
   stop("Response and factor vectors must be the same length")

if(!is.factor(grp)) grp <- as.factor(grp)
F.stat <- NULL
F.stat[1] <- anova(lm(x ~ grp))$F[1]
F.p.value.obs <- anova(lm(x ~ grp))$P[1]
F.df <- anova(lm(x ~ grp))$Df
for(i in 2:B){
   F.stat[i] <- anova(lm(sample(x) ~ grp))$F[1]
}

hist.out <- hist(F.stat,probability=T,breaks=breaks,xlab=xlab,main=main,...)
box()
if(F.obs){
   abline(v=F.stat[1],lty=2,col=2)
   mtext(bquote(paste(italic(F)," = ",.(round(F.stat[1],2)))),
     3,-1,at=F.stat[1],adj=-.2,cex=.8,col=2)
}

if(shade){
   no.shade <- sum(hist.out$mids <= F.stat[1])
   if(pause) null <- locator(1)
   par(new=T)
   hist(F.stat,probability=T,breaks=hist.out$breaks,
     col=c(rep(0,no.shade),rep(2,length(hist.out$breaks)-no.shade-1)),
     axes=F,ann=F,add=T,density=30,border=1)
   if(F.obs) abline(v=F.stat[1],lty=2,col=2)
}

if(F.dist){
   r.br <- range(hist.out$breaks)
   xx <- seq(r.br[1],r.br[2],length=150)
   if(pause) null <- locator(1)
   lines(xx,df(xx,F.df[1],F.df[2]),col=4)
   if(F.shade & F.dist){
     if(pause) null <- locator(1)
     yy <- xx[xx>F.stat[1]]
     yt <- df(xx[xx>F.stat[1]],F.df[1],F.df[2])
     lines(yy,yt,type="h",col=4)
   }
}
```

```
p.values <- c(mean(F.stat >= F.stat[1]),F.p.value.obs)
names(p.values) <- c("Randomization","F-test")
return(p.values)
}
```

Acknowledgements

I would like to thank David Moore for his excellent groundbreaking textbooks. The ideas in this article came when I was teaching out of his book. Nearly all introductory textbooks treat inference in the same manner as David Moore and the fact that I quote from his book should in no way be seen as a criticism of him or his books, but simply as an illustration of how the best textbooks available currently treat the subject.

References

Cobb, G. W. (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?," *Technology Innovations in Statistics Education*, Vol. 1, No. 1, Article 1, http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1.

Dwass, M. (1957). "Modified Randomization Tests for Nonparametric Hypotheses," *The Annals of Mathematical Statistics*, 28, 181-187.

Moore, D. S. (2004). *The Basic Practice of Statistics*, 3rd ed., New York: W. H. Freeman.

Moore, D. S. (2008). *The Basic Practice of Statistics*, 4th ed., New York: W. H. Freeman.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Michael D. Ernst
Associate Professor
Department of Business Computer Information Systems
G. R. Herberger College of Business
St. Cloud State University
720 Fourth Avenue South
St. Cloud, MN 56301-4498
mdernst@stcloudstate.edu

Volume 17 (2009) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications