Robustness

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Robustness

Introduction

- In Psychology 310, we discussed the idea of *robustness* of a statistical test.
- Our discussion was at the level common to textbooks that are common in our field.
- However, robustness in statistical testing has actually advanced considerably in the past 3 decades, and some of the points of view that were common in 1995 have now been questioned.
- In this module, we discuss what robustness is, and why we need it.

Robust Statistics

- In our previous discussions of robustness in Psychology 310 and in the introductory material in this course, we have concentrated on the robustness of estimators to violations of the assumptions of classic parametric statistics.
- For example, many of our test statistics examine hypotheses about the population mean, and assume samples are i.i.d from normally distributed populations with the same variance.
- Our focus in examining robustness was at the level of the test statistic, that is, we asked whether the test statistic maintains its "nominal" behavior when key assumptions like independence, normality, and homogeneity are violated.

Robust Statistics

- This is an important area of statistical research, and people continue to devote attention to it.
- If a statistic is not robust to violations of assumptions, and those assumptions are violated, then the statistic may reject the null hypothesis either very frequently or never. In the former case, the true α may be much higher than the "nominal" α. In the latter case, power may be extremely low along with α.
- Either situation can seriously compromise the enterprize of inferential statistics.
- However, as important as robustness of a statistical test is, in recent years, additional attention has been paid to a question that is possibly even more fundamental: should we be estimating the classic quantities like the population mean? Are these *parameters themselves* robust?

Robust Parameters

- Consider a measure of *location*, or *central tendency.*
- Let X be a random variable with cumulative distribution function F, and let $\theta(X)$ be some descriptive measure of F. Then (see, e.g., Wilcox, 2012) $\theta(X)$ is said to be a *measure of location* if it satisfies the following conditions for constants a and b:
  1. $\theta(aX + b) = a\theta(X) + b$
  2. $X \geq 0$ implies $\theta(X) \geq 0$
  3. Define $F_x(x) = \Pr(X \leq x)$ and $F_y(x) = \Pr(Y \leq x)$. Then X is said to be *stochastically larger than* Y if, for all x, $F_x(x) \leq F_y(x)$, with strict inequality for some value of x. If all the quantiles of X are greater than the corresponding quantiles of Y, then X is stochastically larger than Y. If X is stochastically larger than Y, then if $\theta()$ is to qualify as a measure of location, it should be the case that $\theta(X) \geq \theta(Y)$.

Qualitative Robustness

- Colloquially speaking, a parameter has *qualitative robustness* if the parameter is relatively unaffected by small differences in the cdf, $F(x)$.
- Here we view a statistical parameter as a functional that carries distribution functions, for example F, into numbers. For example the mean is written $(X) = T(F)$, this boils down to a condition of continuity for the functional T.
- It turns out that the population mean $\mu = (X)$ is not a continuous functional, and so, by that criterion, it is ruled out.

Qualitative Robustness

- On the other hand, the γ-trimmed mean is a continuous functional.
- The γ-trimmed mean is the mean of a distribution after it has been transformed in a particular way.
- Specifically, the distribution is truncated at its γ and $1-\gamma$ quantiles.
- In order for the revised distribution to have an area under it equal to 1, the density function is then re-standardized by dividing it by $1-2\gamma$, i.e.,

$$f_\gamma = \frac{1}{1-2\gamma} f(x), \text{for} \quad x_\gamma \leq x \leq x_{1-\gamma} \tag{1}$$

Qualitative Robustness

The 20 % Trimmed Normal

- As an example, consider the standard normal distribution.
- It has the density function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp -x^2/2, \quad \text{for} \quad -\infty \leq x \leq \infty \tag{2}$$

- Since the 80th percentile of the normal distribution is 0.8416, the 20% trimmed normal has the density function

$$f(x) = \frac{1}{0.6} \frac{1}{\sqrt{2\pi}} \exp -x^2/2, \quad \text{for} \quad -0.8416 \leq x \leq 0.8416 \tag{3}$$

Infinitesimal Robustness

- Ideally, small changes in x should not be accompanied by huge differences in $f(x)$.
- One way to impose such a condition is to insist that the derivative of $f(x)$ be bounded.
- In the statistics literature, the derivative of a functional $T(F)$ is called the *influence function* of T at F.
- Roughly, the influence function assesses the degree to which a small change in F produces a large difference in the parameter $T(F)$.

Infinitesimal Robustness

- Consider a mixture of two distributions, where one distribution F_1 occurs with probability π and the other, F_2, occurs with probability $1 - \pi$.
- The cdf of this probabilistic mixture is

$$F(x) = \pi F_1(x) + (1 - \pi)F_2(x) \tag{4}$$

  Let's use this formula to characterize the influence function, especially in terms of a single observation.
- Consider the special probability distribution Δ_x, which yields the value x with probability 1.
- Now, consider a mixture which samples randomly from F with probability $1 - \epsilon$, and from Δ_x with probability ϵ.
- From Equation 4, we see that the distribution has cdf

$$F_{x,\epsilon} = (1 - \epsilon)F + \epsilon\Delta_x \tag{5}$$

Infinitesimal Robustness

- If F has mean μ, then $F_{x,\epsilon}$ has mean $(1-\epsilon)\mu + \epsilon x$.
- So, the *difference* between the mean of $F_{x,\epsilon}$ and the mean of F is

$$(1-\epsilon)\mu + \epsilon x - \mu = \epsilon(x-\mu) \tag{6}$$

- The relative influence on $T(F)$ of having the value x occur with probability ϵ is

$$\frac{T(F_{x,\epsilon}) - T(F)}{\epsilon} \tag{7}$$

- So the relative influence of the mean is

$$\frac{\epsilon(x-\mu)}{\epsilon} = x - \mu \tag{8}$$

- The influence function is the limit of the relative influence as ϵ approaches 0 from above.
- This is simply

$$IF(x) = x - \mu \tag{9}$$

which does not depend on F and is not bounded. So the population mean μ does not possess infinitesimal robustness.

Quantitative Robustness

- The way that the quantitative robustness of a parameter is assessed is via the *breakdown point*.
- The general idea is to describe quantitatively the effect of a small change in F on some functional $T(F)$.
- With $F_{x,\epsilon}$, the mean is $(1-\epsilon)\mu + \epsilon x$.
- For any value of $\epsilon > 0$, the mean can go to infinity as x gets large.
- The minimal value of ϵ for which a functional can go to infinity as x increases is called the *breakdown point*.
- Thus, the breakdown point for the mean is zero.
- The γ-trimmed mean has a breakdown point of γ.

False Assumptions about Normality

- Many introductory discussions of robustness examine distributions that, when plotted, are very obviously different from a normal distribution.
- These examples lead many people to believe that they can visually detect substantial departures from normality easily.
- These examples are misleading.

The Contaminated Normal Distribution

- Consider a distribution that is a *probabilistic mixture*.
- With probability $1 - \epsilon$, an observation is drawn from distribution F_1, with probability ϵ, it is drawn from distribution F_2. Means and variances are μ_1, μ_2, σ_1^2, and σ_2^2, respectively.
- Recalling that $\sigma_x^2 = (X^2) - \mu_x^2$, it is easily established that the probabilistic mixture has mean

$$\mu = (1 - \epsilon)\mu_1 + \epsilon\mu_2$$

and variance

$$\sigma^2 = (1 - \epsilon)(\sigma_1^2 + \mu_1^2) + \epsilon(\sigma_2^2 + \mu_2^2) - \mu^2$$

.

- If means of both parts of the mixture are zero, then $\mu = 0$, and $\sigma^2 = (1 - \epsilon)\sigma_1^2 + \epsilon\sigma_2^2$.

The Contaminated Normal Distribution

- A standard normal distribution has cdf $\Phi(x)$. For positive constant K, a normal distribution with mean zero and standard deviation K has cdf $\Phi(x/K)$.
- Consider a contaminated normal distribution in which a standard normal occurs with probability $1 - \epsilon$, and a $N(0, K)$ with probability ϵ. This distribution has cdf

$$H(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/K) \tag{10}$$

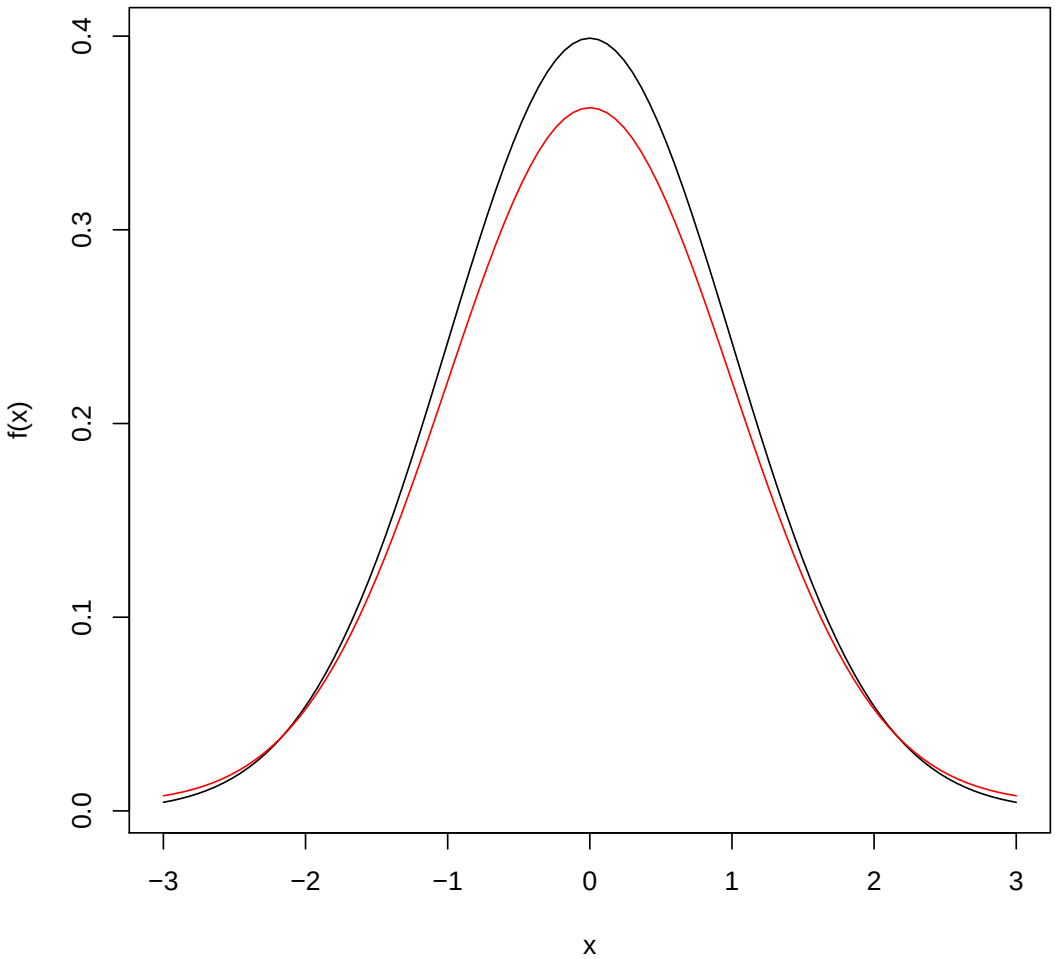
- The contaminated normal has variance $\sigma^2 = \epsilon(K^2 - 1) + 1$. So, for example, the contaminated normal with $K = 10$ and $\epsilon = 0.10$ has variance 10.9.

The Contaminated Normal Distribution

- We know from our work in Psychology 310 that in many circumstances, there are hugely significant practical differences between a normal distribution with a standard deviation of 1, and a normal distribution with a standard deviation of 1.5.
- The picture on the next slide shows two distributions superimposed. One is a standard normal, the other a contaminated normal with $\epsilon = 0.10$, $K = 10$. The variance of the second distribution is 10.9, and so it has a standard deviation of 3.30.
- The picture is not quite what we might expect. The tails of the contaminated distribution (in red) are much heavier than those of the standard normal, but it is difficult to see this since the tail probabilities are low.
- Even though these distributions look very similar, the confidence interval for the sample mean will be much wider for the contaminated normal.

The Contaminated Normal Distribution

```
> curve(dnorm(x),-3,3,ylab="f(x)")
> curve(.9*dnorm(x) + .1*dnorm(x,0,10),-3,3,add=T,col="red")
```



Problems with the Mean

Heavy Tailed Distributions

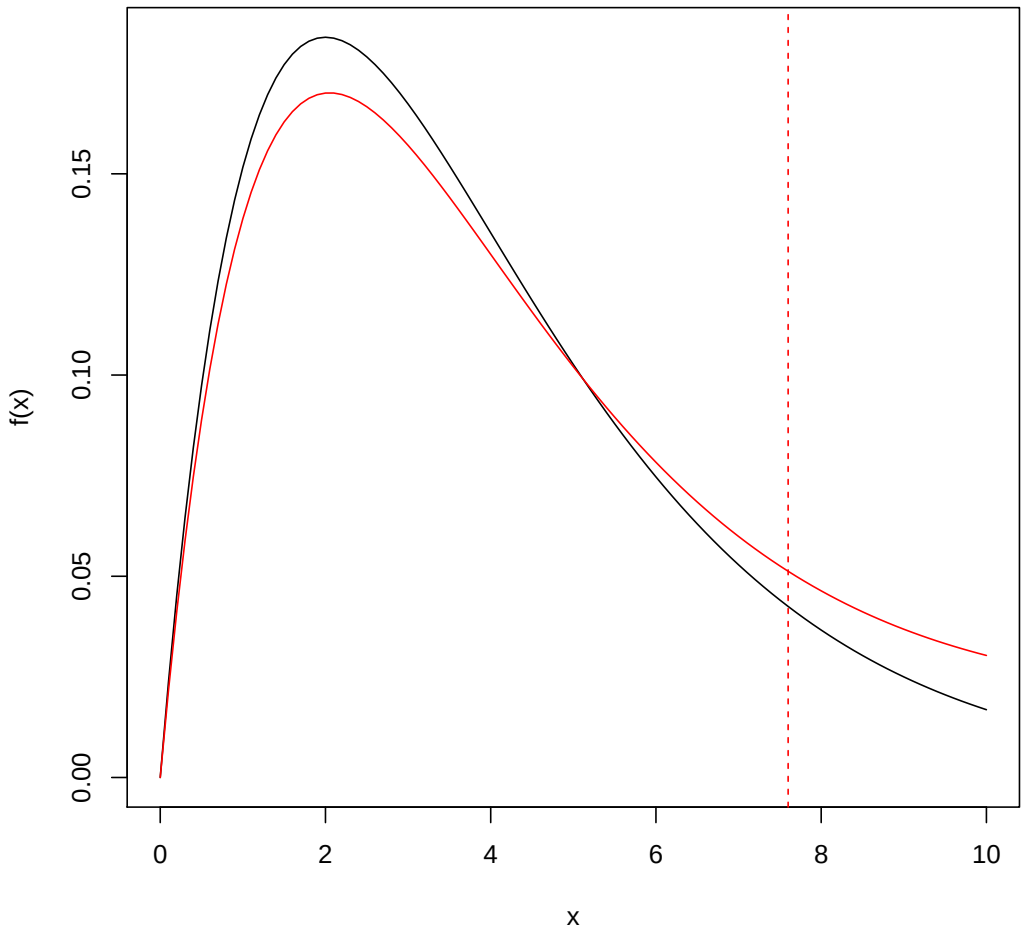
- With a distribution like the one on the previous slide, the standard t test and the "robust" Welch test have a power of only about 0.27 to detect a mean difference of 1.00, with equal sample sizes of $n = 25$ and $\alpha = .05$, 2-tailed.
- What would power be if the distributions were normal? (C.P.)
- Modern robust methods can retain a power of around 0.70 in this situation.

Skewed Distributions

- If a distribution is skewed, the mean may, in an important sense, be unrepresentative of the bulk of observations.
- On the next slide, we see a mixture of two distributions. With probability 0.90, the distribution is χ^2_4, and with probability 0.10 it is 10 times a χ^2_4. The mean of the distribution is 7.6, while the median is only 3.75.
- The contaminated distribution is shown in red, compared with a χ^2_4, shown in black. The latter has a mean of 4, of course, and a median of 3.35. In the case of the red distribution, the bulk of the observations are far removed from the mean, which is indicated with a dotted red vertical line.

Skewed Distributions

```
> curve(dchisq(x,4),0,10,ylab="f(x)")
> curve(.9*dchisq(x,4)+.1*dchisq(x/10,4),0,10,col="red",add=T)
> abline(v=7.6,lty=2,col="red")
```



Examples from Wilcox

- On subsequent slides, we'll see some additional examples of how distribution plots can deceive us into complacency.
- These plots are from a recent workshop presentation by Rand Wilcox, who is one of the leading proponents of the use of robust methods.

Examples from Wilcox

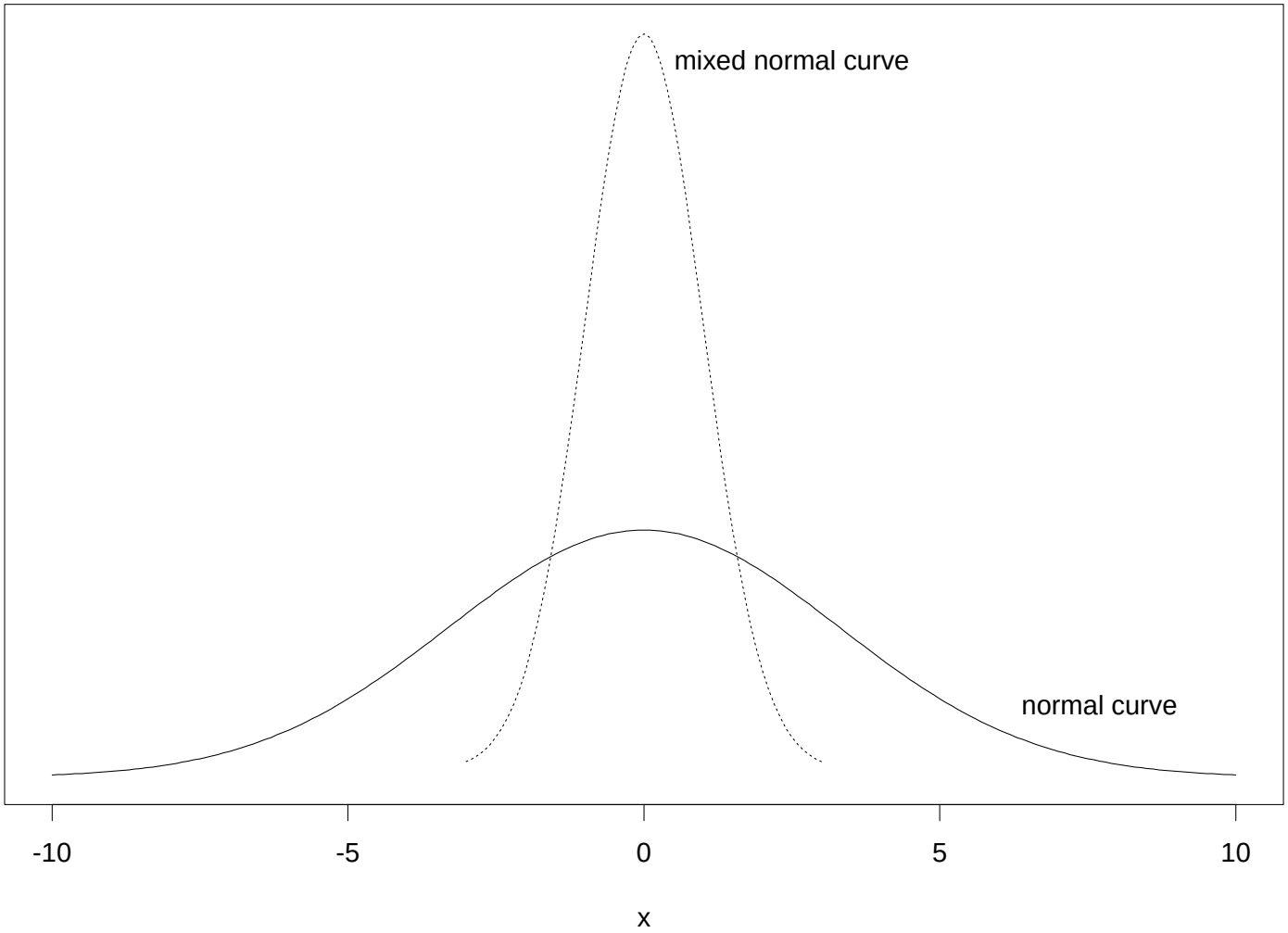


Figure 2: Two probability curves having equal means and variances.

Examples from Wilcox

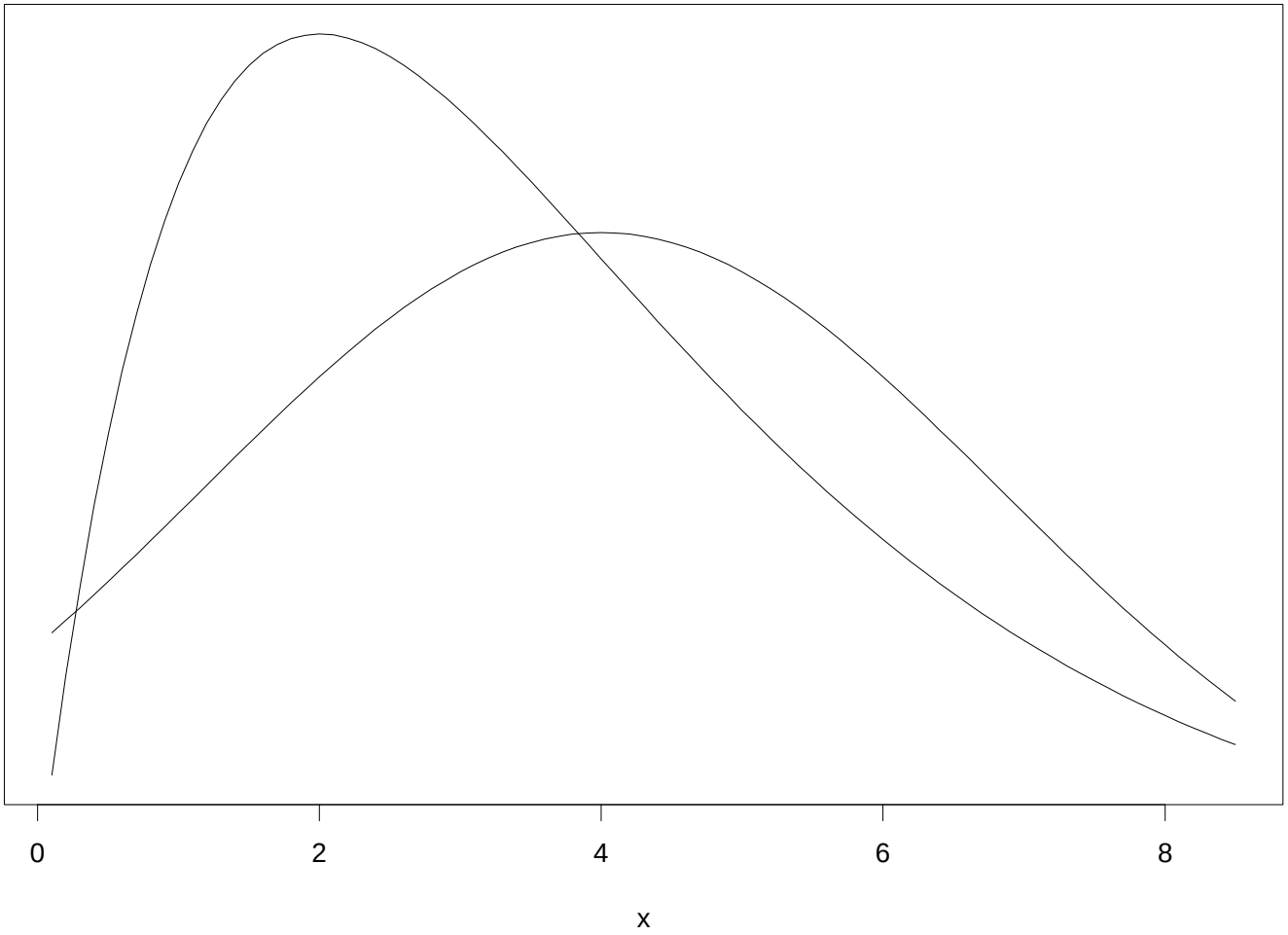


Figure 3: Two probability curves having equal means and variances.

Examples from Wilcox

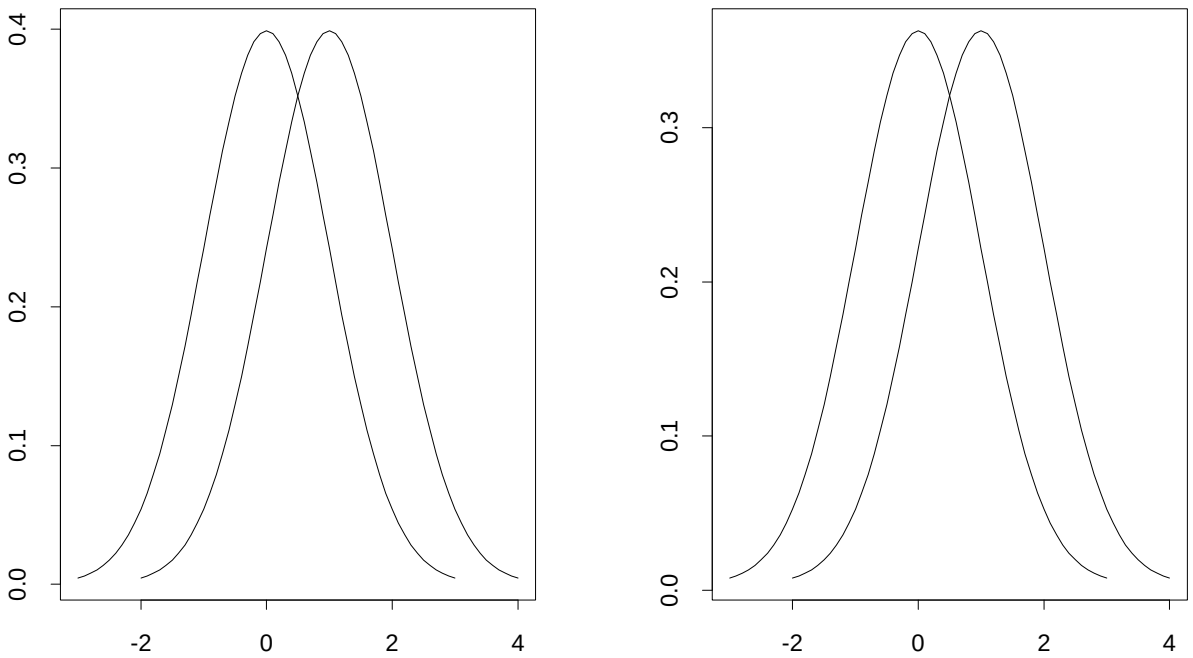


Figure 5: In the left panel, power is .96 based on Student's T, $\alpha = .05$. But in the left panel, power is only .28, illustrating the general principle that slights changes in the distributions being compared can have a large impact on the ability to detect true differences between the population means.

- Wilcox discusses several strategies that one might employ to deal with outliers and heavy-tailed distributions.
- Some seem reasonable, but don't work well. They include
  1. Use transformations.
  2. Discard outliers and use standard methods on the remaining data. The problem with this strategy is that the standard error estimated by the classic methods is radically wrong when applied to trimmed means.
- Overall, the strategy recommended by Wilcox is to use the 20% trimmed mean, *with appropriate adjustments for the standard error*.
- Estimating the standard error of the trimmed mean involves *Winsorizing* a sample.

Winsorizing a Sample

- Winsorization of a random sample consists of first choosing g observations to trim from each end of the sample, and then setting

$$W_i = \begin{cases} X_{(g+1)} & \text{if } X_i \leq X_{(g+1)} \\ X_i & \text{if } X_{(g+1)} < X_i < X_{(n-g)} \\ X_{(n-g)} & \text{if } X_i \geq X_{(n-g)} \end{cases} \tag{11}$$

- The Winsorized sample mean is

$$\bar{X}_w = \frac{1}{n} \sum_{i=1}^{n} W_i \tag{12}$$

- The Winsorized variance is the sample variance of the Winsorized observations.

Winsorizing a Sample

An Example

Example (Self-Awareness Data from Wilcox)

Wilcox (2009, Table 3.2) presents data from a vigilance task experiment by Dana (1990). We'll use these data ($n = 19$) to illustrate several basic robust statistic calculations.

```
> x <- c(77,87,88,114,151,210,219,246,253,262,
+ 296,299,306,376,428,515,666,1310,2611)
> w <- c(114,114,114,114,151,210,219,246,253,262,
+ 296,299,306,376,428,515,515,515,515)
> mean(x)
[1] 448.1053
> median(x)
[1] 262
> mean(w)
[1] 292.7368
> var(w)
[1] 21551.43
> ## Load augmented Wilcox Library Functions
> source("http://www.statpower.net/R311/Rallfun-v27.txt")
> source("http://www.statpower.net/R311/WRS.addon.txt")
```

Winsorizing with $\gamma = 0.20, g = 4$, we find that the Winsorized mean is 292.7368, and the Winsorized variance is 21551.43.

Winsorizing a Sample

An Example

Example (Self-Awareness Data from Wilcox (ctd))

We can compute the Winsorized sample, sample mean, and variance more directly by using functions from the Wilcox library. I've added the function `winsorize()` to generate the winsorized sample directly.

```
> winsorize(x,tr=.2)
 [1] 114 114 114 114 151 210 219 246 253 262 296 299 306 376 428 515 515
[18] 515 515
> win.mean(x,tr=.2)
[1] 292.7368
> win.var(x,tr=.2)
[1] 21551.43
```

The Standard Error of a Trimmed Mean

- Interestingly, the Winsorized sample variance is used to estimate the standard error of the sample trimmed mean.
- Specifically, when using the sample trimmed mean for estimating a population trimmed mean where the proportion trimmed from each tail is γ,we find that the estimated standard error is

$$s_{\bar{X}_t} = \frac{1}{(1-2\gamma)}\frac{s_w}{\sqrt{n}} \tag{13}$$

- Since

$$s_w^2 = \frac{1}{n-1}SS_w = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W})^2 \tag{14}$$

we can rewrite $s_{\bar{X}_t}$ as

$$s_{\bar{X}_t} = \sqrt{\frac{SSw}{(1-2\gamma)^2 n(n-1)}} \tag{15}$$

$$= \sqrt{\frac{SSw}{[(1-2\gamma)n][(1-2\gamma)(n-1)]}} \tag{16}$$

$$= \sqrt{\frac{SSw}{(n-2g)(n-2g-1+2\gamma)}} \tag{17}$$

The Standard Error of a Trimmed Mean

- It should be noted that the formula in Equation 17 for the estimated standard error of a trimmed mean is not the same as
  1. The formula given by Tukey and McLaughlin (1963).
  2. The formula given by Wilcox for use in a two-sample test.
  3. The formula given by SAS for its confidence interval on a single trimmed mean.
- These three other sources all give the same estimator, i.e., the one originally given by Tukey and McLaughlin (1963).
- This latter estimator is

$$s_{\bar{X}_t,SAS} = \sqrt{\frac{SSw}{(n-2g)(n-2g-1)}} \tag{18}$$

- Although Wilcox comments obliquely on this difference, he never provides a justification for it.
- Since the exact distribution of the trimmed mean is intractable, which standard error estimate is "better" is an empirical question that may indeed by highly situation specific.

The Standard Error of a Trimmed Mean

An Example

Example (Standard Error of a Trimmed Mean)

Using Equation 13, we calculate

```
> n=length(x)
> (1/(1 - 2 * 0.20))*sqrt(win.var(x))/sqrt(n)
[1] 56.13193
```

More directly, we could use the Wilcox function

```
> trimse(x,tr=0.20)
[1] 56.13193
```

Significance Test for a Trimmed Mean

- In the previous section, we saw how we can estimate the standard error of a trimmed mean.
- As you might expect, this leads to a test of the hypothesis that $H_0 : \mu_t = \mu_0$.
- The test statistic is

$$t_{n-2g-1} = \frac{\bar{X}_t - \mu_0}{s_{\bar{X}_t}} \tag{19}$$

$$= \frac{(1-2\gamma)\sqrt{n}(\bar{X}_t - \mu_0)}{s_w} \tag{20}$$

Confidence Interval for a Trimmed Mean

- As you might suspect, the t statistic of Equation 19 can be rearranged to yield a $1 - \alpha$ confidence interval.
- Specifically, the endpoints of the interval are

$$\bar{X}_t \pm t_{1-\alpha/2,n-2g-1} s_{\bar{X}_t} \tag{21}$$

Significance Test for a Trimmed Mean

An Example

Example (Confidence Interval for a Trimmed Mean)

Consider the data from Dana(1990) that we examined earlier. The classic confidence interval for the mean μ is

```
> df = length(x)-1
> c(mean(x) - qt(.975,df)*sd(x)/sqrt(length(x)),
+ mean(x) + qt(.975,df)*sd(x)/sqrt(length(x)))
[1] 161.5030 734.7075
```

The method of Equation 21 gives a confidence interval for the population trimmed mean.

```
> tr=0.20
> df<-length(x)-2*floor(tr*length(x))-1
> c(mean(x,trim=.2) - qt(.975,df)*trimse(x),
+ mean(x,trim=.2) + qt(.975,df)*trimse(x))
[1] 160.3913 404.9933
```

Confidence Interval for a Trimmed Mean

An Example

Example (Confidence Interval for a Trimmed Mean (ctd))

We can do the calculation directly with the WRS function `trimci()`

```
> trimci(x,tr=0.20,alpha=0.05)
[1] "The p-value returned by the this function is based on the"
[1] "null value specified by the argument null.value, which defaults to 0"
$ci
[1] 160.3913 404.9933

$estimate
[1] 282.6923

$test.stat
[1] 5.036212

$se
[1] 56.13193

$p.value
[1] 0.0002911395

$n
[1] 19
```

Confidence Interval for a Trimmed Mean

An Example

Example (Confidence Interval for a Trimmed Mean (ctd))

To emulate the SAS calculation, we need the slightly altered estimate of the standard error.

```
> trimseSAS<-function(x,tr=.2,na.rm=FALSE){
+ #
+ #  Estimate the standard error of the gamma trimmed mean
+ #  Using the original Tukey-McLaughlin (1963) formula
+ #  Used by SAS
+ #  The default amount of trimming is tr=.2.
+ #
+ if(na.rm)x<-x[!is.na(x)]
+ n <- length(x)
+ g <- floor(tr*n)
+ SSw = (n-1)*winvar(x,tr)
+ den = (n-2*g)*(n-2*g-1)
+ trimse<-sqrt(SSw/den)
+ trimse
+ }
> trimciB <- function(x,tr=.2){
+ n <- length(x)
+ df<-n-2*floor(tr*n)-1
+ m<-mean(x,trim=tr)
+ t<-qt(.975,df)
+ se <- trimseSAS(x,tr)
+ c(m - t*se,m + t*se)
+ }
> trimciB(x)
[1] 174.0418 391.3428
```