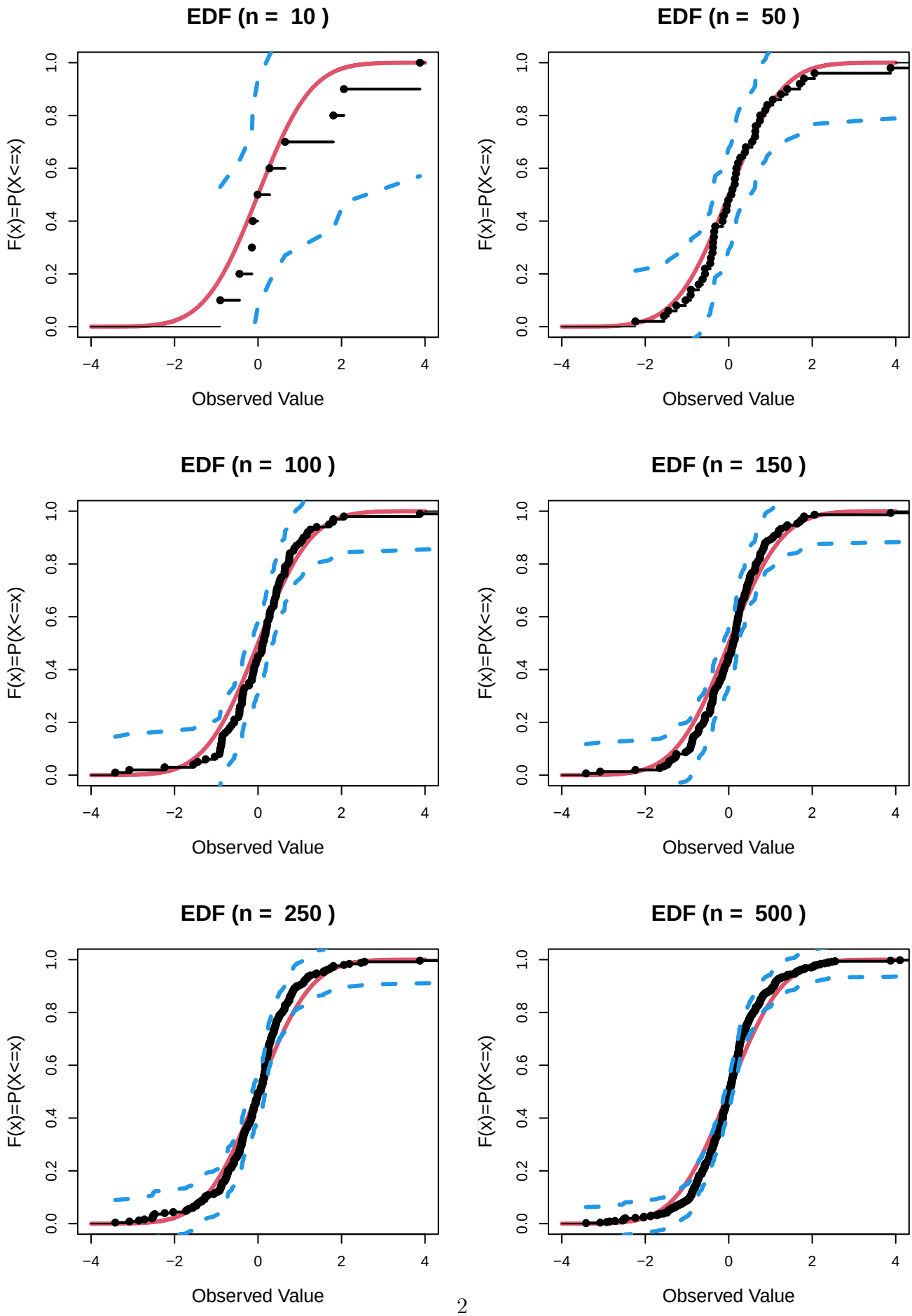
Carson Slater STA 5380 Homework #6

---

1. A SRS, $X_1, X_2, \ldots, X_n$, is taken from a population that follows a t-distribution with 3 degrees of freedom, $X_i \sim t(3)$. Let $Y_i = \frac{X_i}{s_x}$ where s_x is the standard deviation of the x_i's. This standardizes the Y_i's to have variance 1. Roughly, how large of a sample size is needed for the edf in order to determine that the cdf of the Y_i's does not match that of a standard normal cdf? To get started, plot the edf, confidence bands, and the cdf of a standard normal distribution all on the same plot, and repeat this for various values of n. Test the following sample sizes: $n = 10$, $n = 50$, $n = 100$, and $n = 500$. Note that your estimates of the edf and the confidence bands will vary for different samples of the same size. See following page for plots.

   After plotting a sample size of 10, 50, 100, 150, 250, and 500, the EDF begins to significantly deviate from a normal distribution at a sample size of $n = 500$, as the confidence bounds begin to not contain the normal CDF in the regions where the 'lobes' of the normal PDF are. So at around $n = 500$, the CDF of the Y_i's does not really resemble a normal CDF.

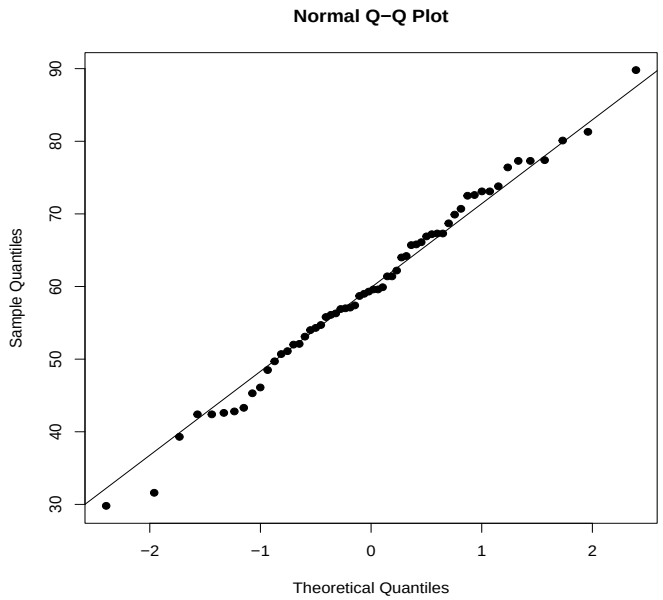
```r
x <- seq(-4, 4, len = 10000); cdf <- pnorm(x)
counts <- c(10, 50, 100, 150, 250, 500)
par(mfrow = c(3,2))
for (i in seq_along(counts)) {
  set.seed(613)
  n <- counts[i]
  t <- rt(n, df = 3)
  y <- t/sqrt(3)
  sort_y <- sort(y)
  edf <- array()
  for(i in 1:length(sort_y)){
    edf[i] <- length(which(y <= sort_y[i])) / length(y)
  }
  plot(x, cdf, type = "l", lwd = 3, ylab = "", xlab = "",col = 2)
  title(paste("EDF (n = ", n,")"), cex.main = 1.5, xlab = "Observed Value", ylab = "
  points(sort_y, edf, pch = 19, xlab = "", ylab = "", ylim = c(0, 1), xlim = c(-3, 3
  for(i in 1:length(sort_y[-1])){
    lines(c(sort_y[i], sort_y[i+1]), c(edf[i], edf[i]), lwd = 2)
  }
  lines(c(-4, min(sort_y)), c(0, 0))
  lines(c(max(sort_y), 4), c(1, 1))
  alphan <- sqrt((1/(2*length(y)))*log(2/0.05))
  flow <- edf - alphan
  fupp <- edf + alphan
  lines(sort_y, flow, lty = 2, col = 4, lwd = 3)
  lines(sort_y, fupp, lty = 2, col = 4, lwd = 3)
}
```


EDF (n = 10)
EDF (n = 50)
EDF (n = 100)
EDF (n = 150)
EDF (n = 250)
EDF (n = 500)
F(x)=P(X<=x)
Observed Value
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4

2. An important quality characteristic of water is the concentration of suspended solid material. In the accompanying file titled "solids.csv" are the concentrations on suspended solids from 60 random locations in a certain lake. A chemist wants to test the claim that the mean concentration of suspended solids exceeds 55% at $\alpha = 0.01$.

   (a) State the hypotheses.
   Let μ be the mean concentration of suspended solids in a certain lake.
   - H_0: $\mu \leq 55\%$.
   - H_a: $\mu > 55\%$.

   (b) Check the assumptions.



   The sample size is $n > 30$, and the QQ-plot indicates the data was drawn from a normal distribution. Therefore we can safely assume that $\bar{x} \sim \mathcal{N}\left(55, \frac{\sigma}{60}\right)$.

   (c) Find the test statistic.

```
mean(solids); sd(solids)

## [1] 59.86667
## [1] 12.49778
```

$$\text{T.S.} = \frac{59.87 - 55}{\sigma/\sqrt{n}} \longrightarrow \frac{59.87 - 55}{s/\sqrt{60}} \approx \frac{59.87 - 55}{12.50/\sqrt{60}} \sim t_{59}$$

   The test statistic for this hypothesis test is 3.016.

   (d) What are the critical value and rejection region?

```
qt(0.99, df = 59)

## [1] 2.391229
```

The critical value is approximately 2.39, and the rejection region is any number greater than 2.39. In other words, $(2.39, \infty)$.

(e) What is the p-value?

```
pt(3.016, df = 59, lower.tail = FALSE)

## [1] 0.001887313
```

The p-value is approximately 0.002.

(f) What is the decision?
The decision is to reject H_0 in favor of H_a.

(g) What is the conclusion in the context of this problem?
With an α of 0.01, we find that there is sufficient evidence that the true the mean concentration of suspended solids in a certain lake is greater than 55%.

(h) What type of error could have been made, and what are the consequences of such an error in the context of this problem?
In this problem, there exists the possibility of a Type I error. This occurs when we reject H_0 when it is true. In this particular problem, we would believe the true mean suspended solids in a particular lake is greater than it actually is, which might influence the decisions of how environmental protection agencies care for the lake.

(i) Interpret the p-value in the context of this problem.
The probability of observing a sample mean percentage of suspended solids of 59.87% given the true mean percentage of suspended solids is 55% is approximately is approximately 0.002.

(j) If the true mean is $\mu = 60\%$, interpret what it means for the power of the test to be 0.7590 for detecting this alternative.
If the true mean concentration of suspended solids in a certain lake is $\mu = 60\%$, we have a 75.90% chance of drawing a sample whose sample mean leads us to reject the null hypothesis that $\mu = 55\%$.

(k) Compute the appropriate confidence interval that will also test the hypotheses stated in part (a).

```
t.test(solids, alternative = "greater", conf.level = 0.99, mu = 55)$conf.int

## [1] 56.00852      Inf
## attr(,"conf.level")
## [1] 0.99
```

The 95% confidence interval that will test the hypotheses in part (a) is approximately (56.009, ∞).

3. For each of the following situations, state whether the appropriate test would be an Independent samples T-test or a Paired T-test, assuming the conditions are met to run these tests. State the null and alternative hypotheses that would be used to test the claim, and define the parameter of interest in the context of the problem.

   (a) It's claimed that taking Aspirin regularly reduces blood pressure. To investigate this claim, a random sample of 40 identical twin pairs (for a total of 80 people) are recruited, and one member of each pair takes aspirin every day for 2 months, and the other takes a placebo every day for 2 months. For each twin pair, the difference in blood pressure between treatments, calculated as the blood pressure of the twin taking the aspirin minus the blood pressure of the twin taking the placebo is computed.

   In this case, the appropriate test would be a Paired T-test. Taking the difference of the blood pressure measurements for each samples would induce dependence in the sampling process. The parameter of interest in the context of this problem would be μ_d, the mean difference in blood pressure between treatments. The null hypothesis would be that H_0: $\mu_d = 0$, whereas the alternative hypothesis would be H_a: $\mu_d \neq 0$.

   (b) A sociologist wants to test the claim that the number of serious relationships that men and women have before age 21 is different. She takes a SRS of 45 women and a SRS of 50 men between the ages of 21 and 25. The researcher determines the number of serious relationships each individual in her study had before the age of 21.

   The most appropriate T-test in this case would be an independent samples T-test. The parameter of interest to the sociologist is the mean number of serious relationships each gender has had before 21 for each group, μ_1 and μ_2. To test her claim, she must suppose the null hypothesis H_0: $\mu_1 - \mu_2 = 0$ (1 = men, 2 = women), whereas the alternative hypothesis would suppose that H_a: $\mu_1 - \mu_2 \neq 0$.

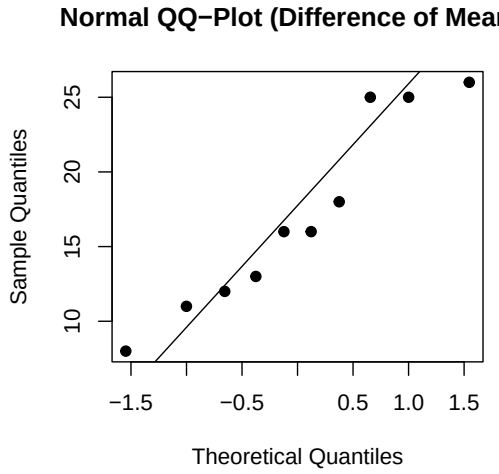
   (c) A study reported in the *Journal of the American Dietetic Association* reported a study testing the claim that vitamin C content of ready to drink brands of orange juice is higher on average when it is newly opened than after it's been open for 4 weeks. To test the claim, the researcher took a random sample of 40 ready to drink orange juice samples, opened them, and immediately measured the vitamin C content of each. After the packages had been open for 4 weeks, they measured the vitamin C content of each orange juice again.

   In this case, the appropriate test would be a Paired T-test. Taking the difference in Vitamin C content from when a bottle is freshly opened and after four weeks would induce dependence in the sampling process. Hence, the parameter of interest in the context of this problem would be μ_d, the mean difference in Vitamin C content from when a bottle is freshly opened and after four weeks. The null hypothesis would be that H_0: $\mu_d \leq 0$, whereas the alternative hypothesis would be H_a: $\mu_d > 0$. If the content of Vitamin C decreases over time as the study claims, then the difference in Vitamin C content between the freshly opened bottle and the same bottle four weeks later should be positive.

(d) You claim there is an association between smoking and coffee consumption and that smokers on average drink more coffee than non-smokers. To test your claim, you take a SRS of 30 smokers and a SRS of 30 non-smokers and determine the daily coffee consumption of each person.

For this experiment, an independent samples T-test would be very appropriate. The parameter of interest would be the mean daily coffee consumption between groups, which are smokers and non-smokers. The null hypothesis would be that H_0: $\mu_1 - \mu_2 \leq 0$ (1 = smokers, 2 = non-smokers), whereas the alternative hypothesis would be H_a: $\mu_1 - \mu_2 > 0$.

4. Ten individuals participated in a diet-modification program to stimulate weight loss. Their weight both before and after participation in the program is shown in the lists below. Is there evidence to support the claim that this particular diet modification program is effective in producing a mean weight reduction? Test the claim at $\alpha = 0.05$.




```
before <- c(195, 213, 247, 201, 187, 210, 215, 246, 294, 310)
after <- c(187, 195, 221, 190, 175, 197, 199, 221, 278, 285)
difference <- before - after

shapiro.test(difference)$p.value

## [1] 0.2699747
```

Before conducting a paired difference of means T-test, we observe that the QQ-plots look fine, and the Shapiro-Wilk normality test for the mean difference of weights before and after treatment indicates there is insufficient evidence to reject normality.

```
test <- t.test(before, after, paired = TRUE, conf.level = 0.95)
test$p.value; test$conf.int

## [1] 1.518548e-05
## [1] 12.41328 21.58672
## attr(,"conf.level")
## [1] 0.95
```

The results from this paired T-test indicate there is sufficient evidence to conclude that the mean difference in weights of patients before and after the dietary program is not equal to zero. In other words, there exists evidence to support the claim that this particular diet modification program is effective in producing a mean weight reduction.

5. Studies are made to estimate the number of passengers (other than the driver) per car in urban traffic. In Los Angeles, a study estimated that the proportion of cars with at least 1 passenger is 67%. A researcher believes that the proportion in Denver may be higher than that and takes a random sample of cars passing through a busy intersection between the hours of 10 am and 3 pm on a Wednesday. Of the 200 cars sampled, 139 cars had at least one passenger.

   (a) State the hypotheses.

   - H_0: $p_{\text{Denver}} \leq 0.67$.
   - H_a: $p_{\text{Denver}} > 0.67$.

   (b) Test the researcher's question at the $\alpha = 0.05$ level (include a p-value and a critical value in your answer).

```
# assumptions
n <- 200; p0 <- 0.67
n*p0; n*(1-p0)

## [1] 134
## [1] 66

# test statistic
(ts <- (139/200 - 0.67)/sqrt(0.67*(0.33)/200))

## [1] 0.751901
```

Both np_0 and $n(1-p_0)$ are greater than 10, so we may safely proceed under the assumption that $p \sim \mathcal{N}(0,1)$. The test statistic is

$$\text{T.S.} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.685 - 0.67}{\sqrt{\frac{0.67(0.33)}{200}}} \approx 0.752.$$

```
# p-value
pnorm(ts, lower.tail = FALSE)

## [1] 0.2260553

# critical value
qnorm(0.95)

## [1] 1.644854
```

The p-value is 0.226, and the critical value is 1.645, which indicates there is insufficient evidence to reject H_0, hence the decision is to fail to reject H_0.

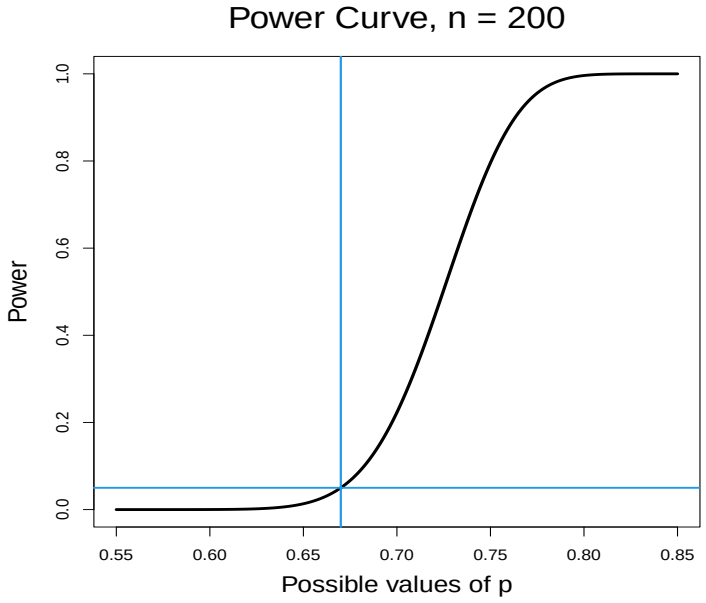
   (c) What conclusion should the researcher draw based on this sample of 200 cars?
   The researcher should conclude at the $\alpha = 0.05$ level, there is insufficient evidence to reject H_0, that the proportion of cars with more than one passenger in Denver is less than or equal to 0.67.

(d) What type of error could the researcher have made?
In this case, the researcher is susceptible to Type II error, which is failing to reject the null hypothesis when the null is indeed false.

(e) Obtain and plot the power curve (similar computations were done on page 161 of the notes for the mean with a "less than" alternative).
To obtain the power curve, we observe that the power is found using the following expression:

$$
\begin{aligned}
&P\left(\frac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > \frac{p_0-p_1+\sqrt{\frac{p_0(1-p_0)}{n}}z_\alpha}{\sqrt{\frac{p_1(1-p_1)}{n}}}\middle| p_1=p\right)\\
&= P\left(z > \frac{(0.67)-p_1+\sqrt{\frac{(0.67)(0.33)}{200}}(1.645)}{\sqrt{\frac{p_1(1-p_1)}{200}}}\middle| p_1=p\right)\\
&= 1-\Phi\left(\frac{(0.67)-p_1+\sqrt{\frac{(0.67)(0.33)}{200}}(1.645)}{\sqrt{\frac{p_1(1-p_1)}{200}}}\middle| p_1=p\right),
\end{aligned}
$$

yielding the following curve.



6. Two different types of injection-molding machines are used to form plastic parts. A part is considered defective if it has excessive shrinkage or is discolored. Two random samples, each of size 300, are selected, and 15 defective parts are found in the sample from machine 1 while 8 defective parts are found in the sample from machine 2. Using $\alpha = 0.01$, is there a significant difference in the proportion of defective parts produced by the two machines?

```
(p_hat1 <- 15/300); (p_hat2 <- 8/300)

## [1] 0.05
## [1] 0.02666667

p_hat <- (15 + 8)/(300 + 300)
```

Given two independent samples, we observed sample proportions from machines 1 and 2 $\hat{p}_1 = 0.05$ and $\hat{p}_2 \approx 0.0267$, respectively. We want to test if the true proportion of defective parts is significantly different between the two machines. We suppose the true proportions of defective parts is equal for both machines as a null hypothesis, H_0: $p_1 - p_2 = 0$. As for the alternative hypothesis, we suggest that there exists difference in the true respective proportions, H_a: $p_1 - p_2 \neq 0$.

```
300*p_hat; 300*(1-p_hat)

## [1] 11.5
## [1] 288.5
```

We know that $n_1 = n_2$, so $n_1\hat{p} = n_2\hat{p} = 11.5$ and $n_1(1-\hat{p}) = n_2(1-\hat{p}) = 288.5$. Because all four of these values are greater than 10, we can safely assume the test statistic is normally distributed. Finding the test statistic, we have it that

```
# test statistic,  critical value
(ts <- (p_hat1 - p_hat2)/sqrt(p_hat*(1-p_hat)*(1/300 + 1/300))); qnorm(0.995)

## [1] 1.488407
## [1] 2.575829
```

$$\text{T.S.} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2}} \approx 1.488.$$

The critical value at the $\alpha = 0.01$ level for a two-sided difference in proportions test is approximately 2.576, which is more extreme than 1.488 on a standard $\mathcal{N}(0, 1)$ distribution. Therefore at the $\alpha = 0.01$ level, there is not sufficient evidence to reject the null hypothesis that there is a difference in true mean proportions of defective parts produced by machines 1 and 2. Hence, we fail to reject H_0 at the $\alpha = 0.01$ level via the critical value approach.

7. The sugar content in canned peaches is normally distributed, and the variance is thought to be $\sigma^2 = 18$ milligrams^2. A random sample of $n = 10$ cans yields a sample standard deviation of $s = 4.8$ milligrams.

   (a) Test the hypothesis that the variance differs from 18 mg^2 using $\alpha = 0.05$.
   We assume a null hypothesis H_0: $\sigma^2 = 18$ milligrams^2, and test it against an alternative hypothesis, H_a: $\sigma^2 \neq 18$ milligrams^2. Because the sugar content is normally distributed, we can assume that the test statistic,

$$\text{T.S.} = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

```
9*(4.8)^2/18

## [1] 11.52
```

   We observe our test statistic is 11.52. Our p-value would be computed as
   $p\text{-value} = 2 \cdot \min\left(P(\chi^2_9 > \text{T.S.}|\sigma^2 = \sigma_0^2), P(\chi^2_9 < \text{T.S.}|\sigma^2 = \sigma_0^2)\right)$.

```
2*min(pchisq(11.52, df = 9), pchisq(11.52, df = 9, lower.tail = FALSE))

## [1] 0.4834818
```

   We observe a p-value of 0.4835, which indicates we do not have sufficient evidence at the $\alpha = 0.05$ level to reject H_0: $\sigma^2 = 18$ milligrams^2 in favor of H_a: $\sigma^2 \neq 18$ milligrams^2.

   (b) Discuss how part (a) could be answered by constructing a 95% confidence interval for σ. Construct the 95% CI, and show that the same conclusion you reach is part (a) is drawn based on the CI.
   To answer part(a) by constructing a 95% confidence interval for σ, we could form a 95% confidence interval for σ^2, and take the square root of the endpoints of the interval. This would give a 95% confidence interval for σ, and if $s = 4.8$ fell outside of the interval, we can reject H_0: $\sigma^2 = 18$ milligrams^2 in favor of H_a: $\sigma^2 \neq 18$ milligrams^2. We demonstrate this by first creating a confidence interval for σ^2,

```
# 95% CI for variance
lower <- 9*4.8^2/qchisq(0.975, df = 9)
upper <- 9*4.8^2/qchisq(0.025, df = 9)
sqrt(lower); sqrt(upper)

## [1] 3.301609
## [1] 8.762929
```

   The observed $s = 4.8$ falls within the bound of the 95% confidence interval for σ, which is (3.30, 8.76). This indicates that there is sufficient evidence to fail to reject the hypothesis that the true variance of sugar content in canned peaches is 18 mg^2.

8. A procurement specialist has purchased 25 resistors from vendor 1 and 35 resistors from vendor 2. Each resistor's resistance is measured with the following results:

```
vendor1 <- c(96.8, 99.6, 99.7, 99.4, 98.6, 100.0, 99.4, 101.1, 99.8, 100.3,
     99.9, 97.7, 99.1, 98.5, 101.1, 98.6, 99.6, 98.3, 103.7, 101.9, 101.2,
     98.2, 97.7, 101.0, 98.2)
vendor2 <- c(106.8, 103.2, 102.6, 104.0, 104.6, 106.4, 106.8, 103.7, 100.3,
     106.3, 103.5, 106.8, 104.7, 106.8, 104.0, 102.2, 106.3, 104.1, 104.7,
     105.1, 107.0, 102.8, 109.2, 107.1, 108.0, 104.0, 104.3, 104.2, 107.2,
     107.7, 102.2, 106.2, 105.8, 103.4, 105.4)
```

(a) What distributional assumption is needed to test the claim that the variance of resistance of the product from vendor 1 is significantly different from the variance of resistance of the product from vendor 2? Perform both a graphical analysis and a formal hypothesis test to check the validity of this assumption.

In order to test the claim that the variance of resistance of the product from vendor 1 is significantly different from the variance of resistance of the product from vendor 2, the data in the sample must be assumed to have been drawn from a normal distribution. We can test this assumption both through a Shapiro-Wilk normality test, and a graphical QQ-plot analysis, as seen in Figure 1.

```
shapiro.test(vendor1)

##
##  Shapiro-Wilk normality test
##
## data:  vendor1
## W = 0.96503, p-value = 0.5234

shapiro.test(vendor2)

##
##  Shapiro-Wilk normality test
##
## data:  vendor2
## W = 0.97886, p-value = 0.7216
```

Both the hypothesis tests and graphical analysis have given us reasonable and sufficient evidence to safely assume these values were drawn from normal distributions (SW-tests fail to reject normality, QQ-plots look typical for a normal distribution).

(b) Assuming that the distributional assumption is met, test the claim stated in part (a). In order to test if the variance of resistance of the product from vendor 1 is significantly different from the variance of resistance of the product from vendor 2, we must first state our null and alternative hypotheses:

- H_0: $\sigma_1^2 = \sigma_2^2$.
- H_a: $\sigma_1^2 \neq \sigma_2^2$.

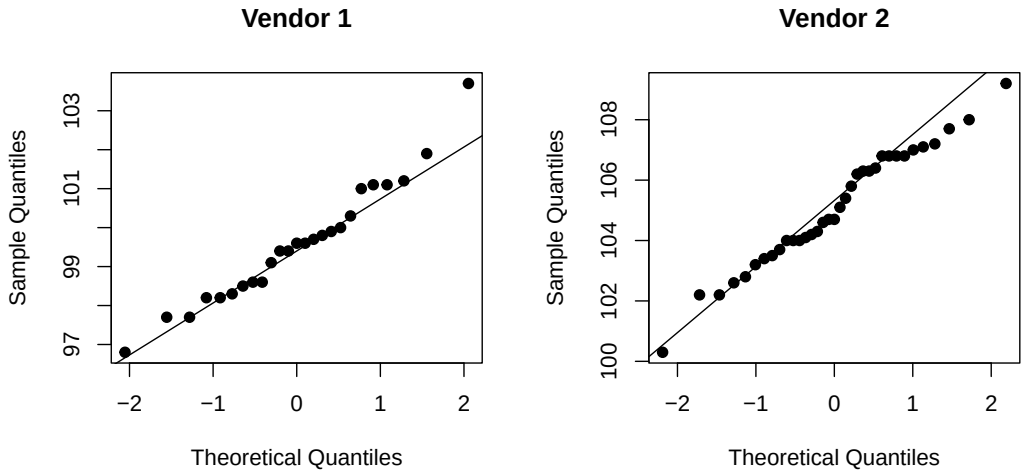


Figure 1: QQ-plots for resistance of the product from vendor 1 and vendor 2.

We obtain the test the statistic,

$$\text{T.S.} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-2} \approx 0.779.$$

```
# test statistic
(ts <- var(vendor1)/var(vendor2))

## [1] 0.6069465

# p-value
2*min(pf(ts, df1 = 24, df2 = 34),
      pf(ts, df1 = 24, df2 = 34, lower.tail = FALSE))

## [1] 0.2053735
```

Testing at the $\alpha = 0.05$ level, the observed test statistic can be used to compute a p-value using the following formula: p-value $= 2 \cdot \min\left(P(F_{34,24,0.05} > \text{T.S.}), P(F_{34,24,0.05} < \text{T.S.})\right) = 0.205$. From this resulting p-value, we fail to reject the null hypothesis, that there exists a significant difference in the variance of resistance of the product from vendor 1 and the variance of resistance of the product from vendor 2.

(c) Based on your conclusion in part (b), test the claim that the mean resistance of resistors purchased from vendor 1 is smaller than the mean resistance of resistors purchased from vendor 2 using a parametric hypothesis test.
In order to test if the mean resistance of the product from vendor 1 is significantly smaller from the mean resistance of the product from vendor 2, we must first state our null and alternative hypotheses:

- H_0: $\mu_1 - \mu_2 \geq 0$.
- H_a: $\mu_1 - \mu_2 < 0$.

We can perform a T-test, assuming equal variance.

```r
t.test(vendor1, vendor2, var.equal = TRUE, alternative = "less")

##
##  Two Sample t-test
##
## data:  vendor1 and vendor2
## t = -11.68, df = 58, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.706487
## sample estimates:
## mean of x mean of y
##   99.5760  105.0686
```

The T-test indicates that there exists sufficient evidence to reject H_0: $\mu_1 - \mu_2 \geq 0$ in favor of H_a: $\mu_1 - \mu_2 < 0$. With a p-value very close to zero, there is even sufficient evidence at the $\alpha = 0.05$ level. Therefore we conclude that with an α of 0.05, we find that there is sufficient evidence that the true the difference in mean resistance of the product from vendor 1 and mean resistance of the product from vendor 2 is less than zero, indicating strong evidence that the mean resistance of the product from vendor 1 is significantly smaller from the mean resistance of the product from vendor 2.

(d) Repeat part (c) using a permutation test. Include your code. How does the result compare with the result in (c)?

```r
# permutation test
set.seed(613)

type <- c(rep("v1", length(vendor1)), rep("v2", length(vendor2)))
data <- c(vendor1, vendor2)

P <- 8000

# creating permutations
diff.means <- array()
for(i in 1:P){
  new.data <- sample(data, replace = F)
  diff.means[i] <- mean(new.data[1:35]) - mean(new.data[36:60])
}
obs.mean <- mean(vendor1) - mean(vendor2)
```

```r
library("MASS")

# plotting distribution of permutation test diff.mean dist.
BW <- width.SJ(diff.means, method = "dpi")
plot(density(diff.means, bw = BW, from = -6, to = 6), main = "", lwd = 3, xlim =
abline(v = obs.mean, col = 2, lwd = 2); abline(v = -obs.mean, col = 2, lwd = 2)
title("Distribution of Permuted Differences in Means")
```

```
q <- quantile(diff.means, c(0.025, 0.975))
abline(v = q[1], col = 4, lwd = 2); abline(v = q[2], col = 4, lwd = 2)
legend("topright",
       legend = c("Permutation Test Difference in Means", "Observed Differences"
       lwd = 2,
       col = c("cornflowerblue", "red"))
```

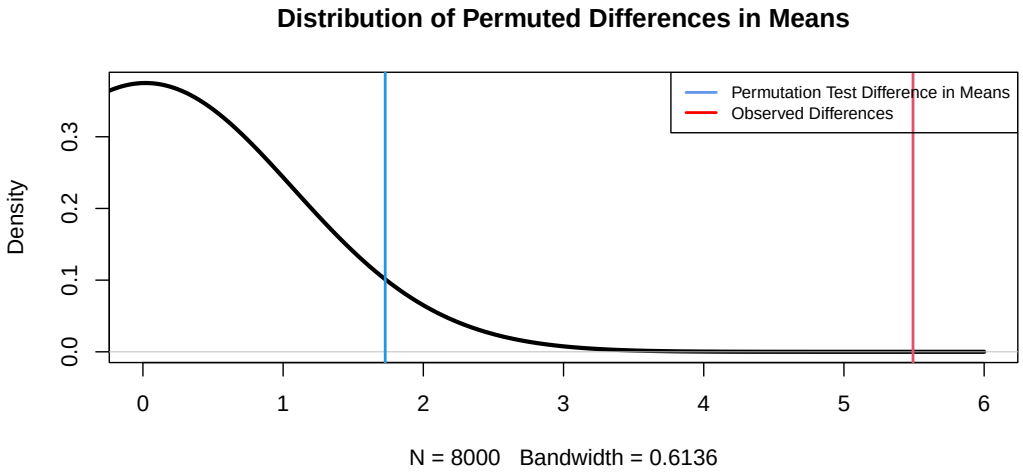


Figure 2: Distribution for difference of means from permutation test.

```
# p-value
length(which(diff.means <= obs.mean)) / P

## [1] 0
```

Like the T-test, this permutation test yields a p-value of 0, which gives sufficient evidence to reject the null hypothesis H_0: $\mu_1 - \mu_2 \geq 0$ in favor of H_a: $\mu_1 - \mu_2 < 0$. Hence, we would conclude that the there is sufficient evidence to suggest the mean resistance of the product from vendor 1 is significantly smaller than the mean resistance of the product from vendor 2.

9. In this problem, you will derive formulas to determine the sample size needed for testing $H_0 : \mu_1 \leq \mu_2$ versus $H_a : \mu_1 > \mu_2$ when the samples are independently drawn from two normal populations, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, and σ_1^2 and σ_2^2 are assumed known for the design purposes. Let n_1 and n_2 be the sample sizes, and let $\bar{x}$ and $\bar{y}$ be the corresponding sample means. The α-level test rejects H_0 if

$$\text{T.S.} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > z_\alpha.$$

(a) Show that the power of the α-level test as a function of $\mu_1 - \mu_2$ is given by

$$\pi(\mu_1 - \mu_2) = \Phi\left[-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right].$$

We know that the power of a test is the probability of correctly rejecting a null hypothesis. This means

$$\begin{aligned}
\pi(\mu_1 - \mu_2) &= P\left(\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > z_\alpha \middle| H_\text{a}\right) \\
&= P\left(\bar{x} - \bar{y} > z_\alpha\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \middle| H_\text{a}\right) \\
&= P\left(\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > \frac{z_\alpha\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \middle| H_\text{a}\right) \\
&= P\left(Z > z_\alpha - \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \middle| H_\text{a}\right) \\
&= P\left(Z < -z_\alpha + \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \middle| H_\text{a}\right) \quad [\text{because } Z \sim \mathcal{N}(0,1)] \\
&= \Phi\left[-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right].
\end{aligned}$$

(b) For detecting a specified difference, $\mu_1 - \mu_2 = \delta > 0$, show that for a fixed total sample size, $n_1 + n_2 = N$, the power is maximized when

$$n_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2}N \quad \text{and} \quad n_2 = \frac{\sigma_2}{\sigma_1 + \sigma_2}N.$$

To find the optimal sample sizes, we can maximize $\pi(\mu_1 - \mu_2)$ with respect to n_1.

$$
\begin{aligned}
0 = \frac{\partial \pi}{\partial n_1} &= \frac{\partial}{\partial n_1} \Phi \left[-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/(N - n_1)}} \right] \\
\implies 0 &= \frac{\partial}{\partial n_1} \left(-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/(N - n_1)}} \right) \quad \text{[because chain rule and } \frac{0}{\Phi'(\cdot)} = 0\text{]} \\
&= 0 - \frac{\mu_1 - \mu_2}{2\left(\sigma_1^2/n_1 + \sigma_2^2/(N - n_1)\right)^{\frac{3}{2}}} \left(\frac{\sigma_2^2}{(N - n_1)^2} - \frac{\sigma_1^2}{n_1^2} \right) \\
&= \left(\frac{\sigma_2^2}{(N - n_1)^2} - \frac{\sigma_1^2}{n_1^2} \right),
\end{aligned}
$$

which yields the equality,

$$
\begin{aligned}
\frac{\sigma_1^2}{n_1^2} &= \frac{\sigma_2^2}{(N - n_1)^2} \\
\frac{\sigma_1}{n_1} &= \frac{\sigma_2}{N - n_1} \\
\sigma_2 n_1 &= \sigma_1 N - \sigma_1 n_1 \\
(\sigma_2 + \sigma_1) n_1 &= \sigma_1 N \\
n_1 &= \frac{\sigma_1}{\sigma_1 + \sigma_2} N.
\end{aligned}
$$

We can then perform the same procedure with respect to n_2.

$$
\begin{aligned}
0 = \frac{\partial \pi}{\partial n_2} &= \frac{\partial}{\partial n_2} \Phi \left[-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/(N - n_2) + \sigma_2^2/n_2}} \right] \\
\implies 0 &= \frac{\partial}{\partial n_2} \left(-z_\alpha + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/(N - n_2) + \sigma_2^2/n_2}} \right) \quad \text{[because chain rule and } \frac{0}{\Phi'(\cdot)} = 0\text{]} \\
&= 0 - \frac{\mu_1 - \mu_2}{2\left(\sigma_1^2/(N - n_2) + \sigma_2^2/n_2\right)^{\frac{3}{2}}} \left(\frac{\sigma_1^2}{(N - n_2)^2} - \frac{\sigma_2^2}{n_2^2} \right) \\
&= \left(\frac{\sigma_1^2}{(N - n_2)^2} - \frac{\sigma_2^2}{n_2^2} \right)
\end{aligned}
$$

which yields the equality,

$$
\begin{aligned}
\frac{\sigma_2^2}{n_2^2} &= \frac{\sigma_1^2}{(N - n_2)^2} \\
\frac{\sigma_2}{n_2} &= \frac{\sigma_1}{N - n_2} \\
\sigma_1 n_2 &= \sigma_2 N - \sigma_2 n_1 \\
(\sigma_1 + \sigma_2) n_2 &= \sigma_2 N \\
n_2 &= \frac{\sigma_2}{\sigma_1 + \sigma_2} N.
\end{aligned}
$$

(c) Show that the smallest total sample size, N, required to guarantee at least $1-\beta$ power when $\mu_1 - \mu_2 = \delta > 0$ is given by

$$N = \left[\frac{(z_\alpha + z_\beta)(\sigma_1 + \sigma_2)}{\delta}\right]^2.$$

From the formula for power, $\pi(\mu_1 - \mu_2)$, letting $n_1 = \frac{\sigma_1}{\sigma_1+\sigma_2}N$ and $n_2 = \frac{\sigma_2}{\sigma_1+\sigma_2}N$,

$$\begin{aligned}
\beta &= 1 - \Phi\left[-z_\alpha + \frac{\delta}{\sqrt{\sigma_1(\sigma_1+\sigma_2)/N + \sigma_2(\sigma_1+\sigma_2)/N}}\right] \\
\beta &= \Phi\left[z_{1-\alpha} - \frac{\delta}{\sqrt{(\sigma_1+\sigma_2)^2/N}}\right] \\
z_\beta &= z_{1-\alpha} - \frac{\delta}{\sqrt{(\sigma_1+\sigma_2)^2/N}} \\
z_{1-\alpha} + z_{1-\beta} &= \frac{\delta}{\sqrt{(\sigma_1+\sigma_2)^2/N}} \\
\sqrt{\frac{1}{N}} &= \frac{\delta}{(z_{1-\alpha} + z_{1-\beta})(\sigma_1+\sigma_2)} \\
\implies N &= \left[\frac{(z_\alpha + z_\beta)(\sigma_1+\sigma_2)}{\delta}\right]^2 \quad \text{[because } (z_{1-\alpha} + z_{1-\beta})^2 = (z_\alpha + z_\beta)^2\text{]}.
\end{aligned}$$

(d) Calculate the minimum sample sizes, n_1 and n_2, needed to achieve 90% power when $\alpha = 0.05$, $\delta = 2.0$, $\sigma_1 = 2.0$, and $\sigma_2 = 4.0$

```
z_alpha <- qnorm(0.05); z_beta <- qnorm(0.10)
sigma_1 <- 2; sigma_2 <- 4; delta <- 2

# calculate the smallest total sample size
(N <- (((z_alpha + z_beta)*(sigma_1 + sigma_2))/delta)^2)

## [1] 77.07463

# calculate smallest sample size for each group
(n1 <- (sigma_1/(sigma_1 + sigma_2)*N) |> ceiling())

## [1] 26

(n2 <- (sigma_2/(sigma_1 + sigma_2)*N) |> ceiling())

## [1] 52
```

To achieve a minimum of 90% power, we find that the minimum sample size for group 1 is $n_1 = 26$, and for group 2 it is $n_2 = 52$.

10. Regional climate models produce possible realizations of the atmosphere based on the physics of the system where each observation corresponds to the average of a grid box. The means and standard deviations of five years of 3-hourly regional climate model wind speeds ((5years)×(365 days)×(8 obs/day)=14,600 observations) measured 10 meters above ground level for the area covering Colorado are shown in Figure 3 below.

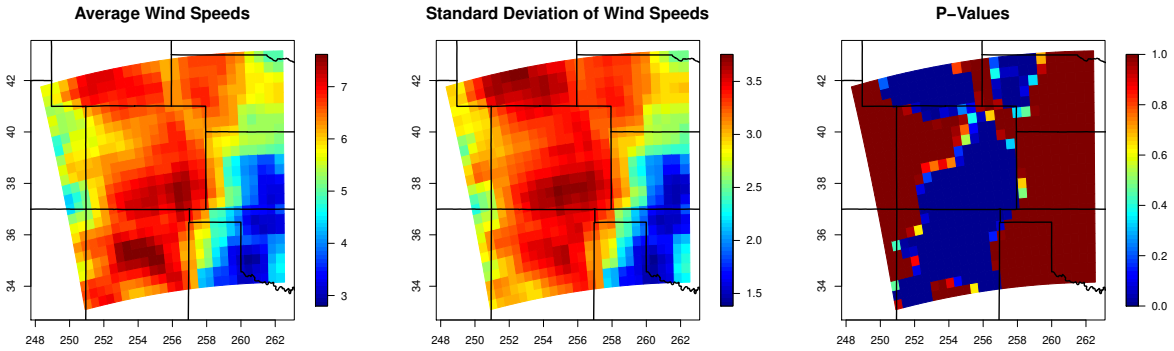


Figure 3: Mean (left) and standard deviation (center) of wind speed (m/s) of 5 years of regional climate model output. The p-values for each grid box to test the hypotheses described below are plotted in the right panel.

One necessary component for wind energy to be economically viable is that the average wind speed must be at least 6.5 m/s, so we can test that for each individual grid box. Thus, for grid box i, the hypotheses and test statistic are

$$H_{0i} : \mu_i \leq 6.5 \quad \text{versus} \quad H_{ai} : \mu_i > 6.5$$

and

$$TS_i = \frac{\bar{X}_i - 6.5}{S_i/\sqrt{14600}},$$

where $\bar{X}_i$ and S_i are the mean and standard deviation of grid box i, respectively. There are 598 grid boxes for which we would like to test these hypotheses, and the p-values are computed as $p_i = P(Z > ts_i)$. The p-values are plotted in the right-hand panel of Figure 3 and are given in the "p_values.csv" file on the class website.

(a) If you ignore the effect of conducting 598 hypothesis tests and naïvely compared each p-value to $\alpha = 0.05$, in how many tests is the null hypothesis rejected?

```
pvals <- read.csv("pvalues.csv", header = TRUE)[[1]]
pvals[which(pvals <= 0.05)] |> length()

## [1] 202
```

We observe that if we ignore the effect of conducting 598 hypothesis tests and naïvely compared each p-value to $\alpha = 0.05$, we would reject the null hypothesis in 202 of the 598 individual hypothesis tests.

(b) Compute the Benjamini-Hochberg adjustment, assuming that the 598 hypothesis tests are independent. How many of the null hypotheses are rejected now?

Because all of the p-values are independent, we presume $C_m := 1$.

```r
pvals <- sort(pvals)
m <- length(pvals)

# Assuming p-values are independent
C_m <- 1
r_i <- (1:length(pvals))*0.05 / (C_m * m)
cbind(r_i, pvals) |> invisible()

index <- rep(0, len = m)
for(i in 1:m){
  if(pvals[i] < r_i[i]){
    index[i] <- 1}
}

R <- which.max(which(index == 1))
T.BH <- pvals[R] # our new rejection threshold

pvals[which(pvals <= T.BH)] |> length()

## [1] 194
```

Here we would reject 194 of the 598 null hypotheses, using the Benjamini-Hochberg adjustment assuming the p-values are independent.

(c) In this scenario, does it make sense to assume that the hypothesis tests are independent? Why or why not?

Wind speeds have a geospatial component that induces dependence (average wind speeds are higher in tornado alley than in Los Angeles, because of many environmental factors including temperature). Therefore because the temperature in the areas directly around an area influences the temperature in an area (which impacts wind speeds, and is likely included in region climate models), it would not make sense to assume the p-values are independence, because it is likely that the models use information from the areas around a particular area of interest to calculate their predictions.

(d) How many null hypotheses are rejected when the Benjamini-Hochberg adjustment for dependent hypothesis tests is used?

```r
# assuming p-values are dependent
C_m <- sum(1/(1:m))
r_i <- (1:598)*0.05/(C_m*m)
cbind(r_i, pvals) |> invisible()

index <- rep(0, len = m)
for(i in 1:m){
  if(pvals[i] < r_i[i]){
    index[i] <- 1}
}

R <- which.max(which(index == 1))
```

```r
T.BH <- pvals[R]

pvals[which(pvals <= T.BH)] |> length()

## [1] 181
```

Here we would reject 181 of the 598 null hypotheses, using the Benjamini-Hochberg adjustment assuming the p-values are dependent.

11. Design a simulation study to evaluate the effect of sample size on the power of the Shapiro-Wilks test when sampling from an exponential population with mean 1. Plot the power against various sample sizes. At approximately what sample size is the power of the Shapiro-Wilks test 100% when sampling from this population?

    First, we create a function which returns a vector with the mean values of the test, which are lower than the confidence threshold $\alpha = 0.05$. We test for sample sizes of 3, which is the minimum, up to 75. We then replicate each run 300 times, obtaining the mean value, in order to smooth the results. See Figure 4 for the curve.

```
library("latex2exp")
power_func <- function(reps, alpha, rates, samples) {
  set.seed(613)
  power <- c()
  for (i in 3:samples){
    test <- mean(replicate(reps,(shapiro.test(rexp(i, rates))$p.value<alpha)))
    power <- c(power, test)
  }
  return(power)
}
# number of samples of size n
reps <-  5000
# alpha for power calculation
alpha <-  0.05
# rate of exponential
rates <- c(1)
# sample size from 3 to 75
samples <- 75
df<- data.frame(n = c(3:samples))
for (i in 2:(length(rates)+1)){
  df[,i] <- power_func(reps, alpha, rates[i-1], samples)
}
colnames(df) <- c("n", "rate_1")
long_df <- pivot_longer(df, -n)
colnames(long_df)[2] <- "rate"

plot(long_df$n, long_df$value, type = "l", lwd = 3, xlab = "", ylab = "")
title(latex2exp::TeX(r'(Power vs. Sample Size for Shapiro-Wilk Test, $\[X_1,...,X_n
      xlab = "Sample Size (n)",
      ylab = "Power",
      cex.main = 1, cex.lab = 1)
```

    The power of the Shapiro-Wilk test reaches 100% just after the sample size reaches $n = 51$; however, it fluctuates slightly between 99.5% and 100% after a sample size of $n = 38$, when the power surpasses the 99.5% threshold.

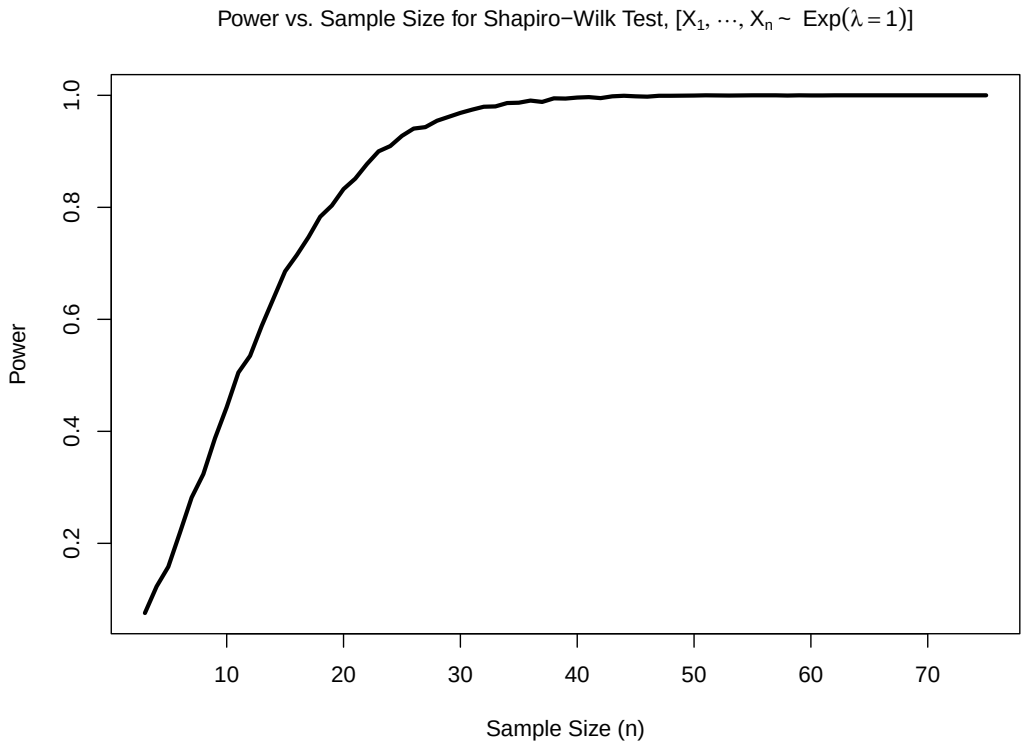


Figure 4: The the effect of sample size on the power of the Shapiro-Wilk test when sampling from an exponential population with mean 1.

12. Use the following data to test the hypothesis that a horse's chances of winning are unaffected by its position on the starting lineup. The data give the starting position of each of 144 winners, where position 1 is closest to the inside rail of the race track.

Starting Position	1	2	3	4	5	6	7	8
Number of Wins	29	19	18	25	17	10	15	11

State the hypotheses, and perform the test at $\alpha = 0.05$.

In this case the two hypotheses are,

- H_0: $p_1 = p_2 = ... = p_8 = 0.125$.
- H_a: At least one $p_i \neq 0.125$.

We have that $e_i = np_{i0} = 144\frac{18}{144} = 18$. Each of the cell counts are greater than 5, so we can safely proceed with the χ^2 goodness of fit test.

```
nwins <- c(29, 19, 18, 25, 17, 10, 15, 11)
e_nwins <- (nwins - 18)^2/18
(TS <- sum(e_nwins))

## [1] 16.33333

# critical value
qchisq(0.95, 7)

## [1] 14.06714

# p-value
1 - pchisq(TS, 7)

## [1] 0.02223948
```

Here we obtain a test statistic of 16.333, and a p-value of 0.022. This is less than our α value of 0.05. Therefore we have significant evidence to reject H_0 in favor of H_a. So we have sufficient evidence to reject the notion that a horse's chances of winning are unaffected by its position on the starting lineup.

13. In a study of the effect of Vitamin B on learning, 12 matched pairs of children were randomly divided into two groups. One child in each pair received a vitamin B tablet (treatment) every day, while the other child received a placebo tablet and served as a control. The table below shows the gain in IQ over the six weeks of the study.

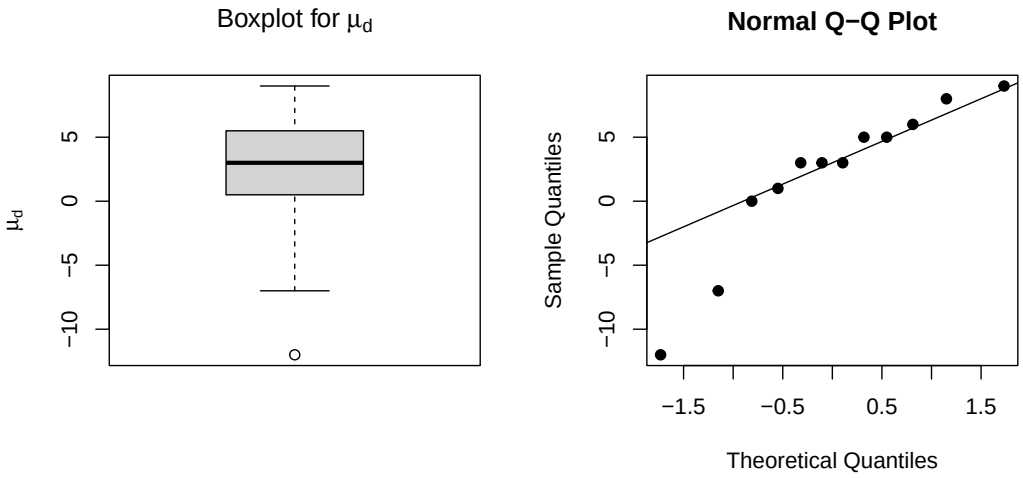
Pair	1	2	3	4	5	6	7	8	9	10	11	12
Treated	14	26	2	4	−5	14	3	−1	1	6	3	4
Control	8	18	−7	−1	2	9	0	−4	13	3	3	3

(a) Is there any evidence that a T-test might not be the best test to use?

```
treated <- c(14, 26, 2, 4, -5, 14, 3, -1, 1, 6, 3, 4)
control <- c(8, 18, -7, -1, 2, 9, 0, -4, 13, 3, 3, 3)
difference <- treated - control

shapiro.test(difference)

##
##  Shapiro-Wilk normality test
##
## data:  difference
## W = 0.86671, p-value = 0.05937
```



There is a low (but not significant at the $\alpha = 0.05$ level) p-value coming from the Shapiro-Wilk normality test, and the QQ-plot looks very skewed. This is grounds for concern as to why a T-test might not be the best test to use, especially given a sample size of $n = 12$.

(b) Find the exact p-value for the sign test to determine if vitamin B improves the IQ.

Here the hypotheses could be listed as:

- H_0: $\tilde{\mu}_d \leq 0$.
- H_a: $\tilde{\mu}_d > 0$.

Proceeding with the sign test, we use `R`.

```
library("BDSA")
SIGN.test(difference, md = 0, alternative = "greater")$p.value
```

This sign test yielded a p-value of 0.0327. At the $\alpha = 0.05$ level, we would have sufficient evidence to reject the null hypothesis that the mean difference between the treatement and control group is less than or equal to zero, in favor of the alternative hypothesis that vitamin B does improve IQ.

Something to note is that we observed a differenced value that was a tie with the median. The `BDSA` package handled the value by throwing away the observation, hence losing information. In the following script, we randomly assign the zero and recompute the p-value.

```
# my sign test
set.seed(613)
(random <- runif(1, 0, 1))

## [1] 0.9786377

splus <- difference[which(difference > 0)]
sminus <- difference[which(difference < 0)]

if (random > 0.5) {
  splus <- c(splus, 0)
} else {
  sminus <- c(sminus, 0)
}
pbinom(length(splus), 12, p = 0.5, lower.tail = FALSE)

## [1] 0.003173828
```

This sign test yielded a p-value of 0.0317, which is very similar to the one computed in the package. So we would conclude the same result at the $\alpha = 0.05$ level. Thus, we would have sufficient evidence to reject the null hypothesis that the mean difference between the treatement and control group is less than or equal to zero, in favor of the alternative hypothesis that vitamin B does improve IQ.

(c) Repeat part (b) using the Wilcoxon signed rank test. Why is a less significant result obtained in this case?

```
# we took out continuity correction
wilcox.test(x = treated, y = control,
            paired = TRUE, alternative = "greater",
            mu = 0, correct = FALSE)
```

```
## Warning in wilcox.test.default(x = treated, y = control, paired = TRUE,
:  cannot compute exact p-value with ties
## Warning in wilcox.test.default(x = treated, y = control, paired = TRUE,
:  cannot compute exact p-value with zeroes

##
##  Wilcoxon signed rank test
##
## data:  treated and control
## V = 47, p-value = 0.106
## alternative hypothesis: true location shift is greater than 0
```

This Wilcoxon signed rank test yielded a p-value of 0.106. At the $\alpha = 0.05$ level, we would have insufficient evidence to reject the null hypothesis that the mean difference between the treatement and control group is less than or equal to zero, in favor of the alternative hypothesis that vitamin B does improve IQ.

Again, this test also struggles (and warns us) to compute a definitive p-value due to rank test. Since we do not know how to compute the p-value for this test, we will program it ourselves, randomly assigning the ties with the hypothesized value, 0. We still struggle to know what to do with the ties, but luckily all ties had value that corresponded to a difference greater than zero. We can calculate the test statistic W_+, and use the critical value approach.

```
# wilcoxon sign-rank test
difference <- difference[order(abs(difference))]

plus <- 0
minus <- 0

for (i in seq_along(difference)) {
  if (difference[i] > 0) plus <- plus + i
  else if (difference[i] < 0) minus <- minus + i
  else {if (runif(1, 0, 1) > 0.5) {plus <- plus + i}
    else {minus <- minus + i}
  }
}
# our new test statistic
plus

## [1] 57
```

We can conclude that we also believe, given difference test statistics between our test and the `stats::wilcox.test()` function, that the `stats::wilcox.test()` does something different with the zero values. We observed a more extreme critical value.

Compared to the sign test, this test yields a less significant p-value due to the consideration of the magnitude of differences between the treatment and control group. Additionally, the ties in differences found when differencing the treatment and con-

trol group induce more uncertainty in the test. Specifically, observations 4 and 6 for treatment and control groups both equal 5, which is a tie. Additionally, there were three differences with a magnitude equal to three.

Basing our decision off of the output of the `stats::wilcox.test()` function, we fail to reject the null hypothesis that the true mean difference between the treatment and control group's increase in IQ is greater than zero.