Topic 20

t Procedures

> *A curve has been found representing the frequency distribution of values of the means of such samples, when these values are measured from the mean of the population in terms of the standard deviation of the sample....* - William Sealy Gosset. 1908

The z-score is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

taken under the assumption that the population standard deviation is known.

If we are forced to replace the unknown σ^2 with its unbiased estimator s^2, then the statistic is known as t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The term $s/\sqrt{n}$ which estimate the standard deviation of the sample mean is called the **standard error**.

We have previously noted that for independent normal random variables the distribution of the t statistic can be determined **exactly**. Because we approximate σ with s, the t-statistic has a higher level of uncertainty than the corresponding z-statistic. This uncertainty decreases with n, the number of observations. Thus, when using the t distribution to construct a confidence interval for the population mean μ, we saw that the margin of error decreased as the number of observations increased. Typically, we do not use the number of observations n to describe this but rather **degrees of freedom** $n - 1$ to match the division by $n - 1$ in the computation of the sample variance, s^2.

We now turn to using the t-statistic as a test statistic for hypothesis tests of the population mean. As with several other procedures we have seen, the two-sided t test is a likelihood ratio test. We will save showing this result into the last section and instead focus on the applications of this widely used set of procedures.

20.1 Guidelines for Using the t Procedures

- Except in the case of small samples, the assumption that the data are a simple random sample from the population of interest is more important that the population distribution is normal.
- For sample sizes less than 15, use t procedures if the data are close to normal.
- For sample sizes at least 15 use t procedures except in the presence of outliers or strong skewness.
- The t procedures can be used even for clearly skewed distributions when the sample size is large, typically over 40 observations.

These criteria are designed to ensure that $\bar{x}$ is a sample from a nearly normal distribution. When these guidelines fail to be satisfied, then we can turn to alternatives that are not based on the central limit theorem, but rather use the rankings of the data. These alternatives, the Mann-Whitney or Wilcoxon rank sum test and the Wilcoxon signed-ranked test, are discussed at the end of this topic.

20.2 One Sample t Tests

We will later explain that the likelihood ratio test for the two sided hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

based on independent **normal** observations $X_1 \ldots, X_n$ with unknown mean μ and **unknown** variance σ^2 is a t-test.

So, compute the t statistic $T(\mathbf{x})$ from the data $\mathbf{x}$. Then, the critical region

$$C = \{|T(\mathbf{x})| > t_{n-1,\alpha/2}\}.$$

where $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ tail probability of the t distribution with $n-1$ degrees of freedom.

Example 20.1. *Radon is a radioactive, colorless, odorless, tasteless noble gas, occurring naturally as the decay product of uranium. It is one of the densest substances that remains a gas under normal conditions.*

Radon is responsible for the majority of the public exposure to ionizing radiation and is the most variable from location to location. Radon gas from natural sources can accumulate in buildings, especially in confined areas such as attics, and basements. Epidemiological evidence shows a clear link between breathing high concentrations of radon and incidence of lung cancer. According to the United States Environmental Protection Agency, radon is the second most frequent cause of lung cancer, after cigarette smoking, causing 21,000 lung cancer deaths per year in the United States.

*To check the reliability of radon detector, a university placed 12 detectors in a chamber having 105 picocuries of radon. (*1 *picocurie is* 3.7×10^{-2} *decays per second. This is roughly the activity of 1 picogram of the radium 226.)*

The two-sided hypothesis

$$H_0 : \mu = 105 \quad \textit{versus} \quad H_1 : \mu \neq 105,$$

where μ *is the actual amount of radon radiation. In other words, we are checking to see if the detector is biased either upward or downward.*

The detector readings were:

91.9 97.8 111.4 122.3 105.4 95.0 103.8 99.6 96.6 119.3 104.8 101.7

Using R*, we find for an* $\alpha = 0.05$ *level significance test:*

```
> radon<-c(91.9,97.8,111.4,122.3,105.4,95.0,103.8,99.6,96.6,119.3,104.8,101.7)
> hist(radon)
> mean(radon)
[1] 104.1333
> sd(radon)
[1] 9.39742
> length(radon)
[1] 12
> qt(0.975,11)
[1] 2.200985
```

Thus, the t*-statistic is*

$$t = \frac{105 - 104.1333}{9.39742/\sqrt{12}} = -0.3195.$$

Thus, for a 5% significance test, $|t| < 2.200985$*, the critical value and we fail to reject* H_0*.* R *handles this procedure easily.*

```
> t.test(radon,alternative=c("two.sided"),mu=105)

One Sample t-test
```

```
data:  radon
t = -0.3195, df = 11, p-value = 0.7554
alternative hypothesis: true mean is not equal to 105
95 percent confidence interval:
  98.1625 110.1042
sample estimates:
mean of x
 104.1333
```

The output also gives the 95% confidence interval

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.025,11}.$$

The power against an alternative of a change in 5 picocuries is

```
> power.t.test(n=12,delta=5,sd=sd(radon),type=c("one.sample"))

     One-sample t test power calculation

              n = 12
          delta = 5
             sd = 9.39742
      sig.level = 0.05
          power = 0.3907862
    alternative = two.sided
```

The `power.t.test` command considers five issues - sample size `n`, the difference between the null and a fixed value of the alternative `delta`, the standard deviation s, the significance level α, and the power. We can use `power.t.test` to drop out any one of these five and use the remaining four to determine the remaining value. For example, if we want to assure an 80% power against an alternative of 110, then we need to make 30 measurements.

```
> power.t.test(power=0.80,delta=5,sd=sd(radon),type=c("one.sample"))

     One-sample t test power calculation

              n = 29.70383
          delta = 5
             sd = 9.39742
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

In these types of application, we often use the terms **specificity** and **sensitivity**. Recall that setting the significance level α is the same as setting the false positive rate or type I error probability. The specificity of the test is equal to $1 - \alpha$, the probability that the test is not rejected when the null hypothesis is true. The sensitivity is the same as the power, one minus the type II error rate, $1 - \beta$.

20.3 Correspondence between Two-Sided Tests and Confidence Intervals

For a two-sided t-test, we have the following list of equivalent conditions:

fail to reject with significance level α.

$$|t| < t_{n-1,\alpha/2}$$

$$\left|\frac{\mu_0 - \bar{x}}{s/\sqrt{n}}\right| < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2} < \frac{\mu_0 - \bar{x}}{s/\sqrt{n}} < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 - \bar{x} < t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

μ_0 is in the $\gamma = 1 - \alpha$ confidence interval

This is displayed in Figure 20.1 with the green $\bar{x}$ and the horizontal green line indicating the γ-level confidence interval containing μ_0. In addition, *reject the hypothesis with significance level* α is equivalent to μ_0 *is not in the confidence interval.* This is displayed in Figure 1 with the red $\bar{x}$ and the horizontal line indicating the γ-level confidence interval that fails to contain μ_0.

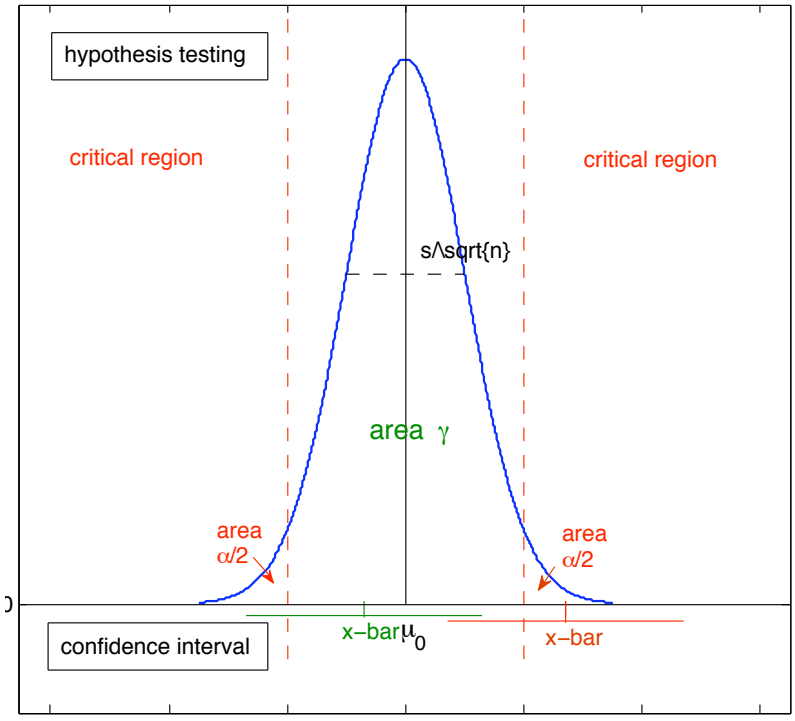


Figure 20.1: γ-level confidence intervals and α level hypothesis tests. $\gamma = 1 - \alpha$. The blue density curve is density of the sampling distribution under the null hypothesis. The red vertical dashes show the critical values for a two-sided test. γ is the area under the density curve between the vertical critical value lines. α is the area under the density curve outside the vertical critical value lines. The green $\bar{x}$ shows the case of *fails to reject* is equivalent to *the confidence interval contains* μ_0. The red $\bar{x}$ show the case *reject* is equivalent to *the confidence interval fails to contain* μ_0.

20.4 Matched Pairs Procedures

A **matched pair procedure** is called for when a pair of quantitative measurements from a simple random sample

$$X_1, X_2, \ldots, X_n, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_n$$

are made on the same subjects. The alternative can be either one-sided or two sided. Underlying this assumption is that the populations are the same under the null hypothesis.

Thus, when H_0 holds and if in addition, if the data are normal, then $\bar{X} - \bar{Y}$ is also normal and so

$$T = \frac{\bar{X} - \bar{Y}}{S_{X-Y}/\sqrt{n}}$$

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

The γ-level confidence interval for the difference in the population means is

$$\bar{x} - \bar{y} \pm \frac{s_{X-Y}}{n} t_{n-1,(1-\gamma)/2,\cdot}$$

Example 20.2. *Researchers are concerned about the impact of vitamin C content reduction due to storage and shipment. To test this, researchers randomly chose a collection of bags of wheat soy blend bound for Haiti, marked them, and measured vitamin C from a sample of the contents. Five months later, the bags were opened and a second sample was measured for vitamin C content. The units are milligrams of vitamin C per 100g of wheat soy blend.*

Factory	Haiti	Factory	Haiti	Factory	Haiti	Factory	Haiti
44	40	45	38	39	43	50	37
50	37	32	40	52	38	40	34
48	39	47	35	45	38	39	38
44	35	40	38	37	38	39	34
42	35	38	34	38	41	37	40
47	41	41	35	44	40	44	36
49	37	40	34	43	35		

Here is the R *output with the 95% confidence interval for $\mu_F - \mu_H$ where*

- *μ_F is the mean vitamin C content of the wheat soy blend at the factory and*
- *μ_H is the mean vitamin C content of the wheat soy blend in Haiti.*

```
> factory<-c(44,50,48,44,42,47,49,45,32,47,40,38,41,40,39,52,45,37,38,44,43,
+50,40,39,39,37,44)
> haiti<-c(40,37,39,35,35,41,37,38,40,35,38,34,35,34,43,38,38,38,41,40,35,37,
+34,38,34,40,36)
> boxplot(factory,haiti)
> t.test(factory, haiti, alternative = c("two.sided"),mu = 0, paired = TRUE)

Paired t-test

data:  factory and haiti
t = 4.9589, df = 26, p-value = 3.745e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.122616 7.544050
sample estimates:
mean of the differences
               5.333333
```

The input

```
> t.test(factory - haiti, alternative = c("two.sided"),mu = 0)
```

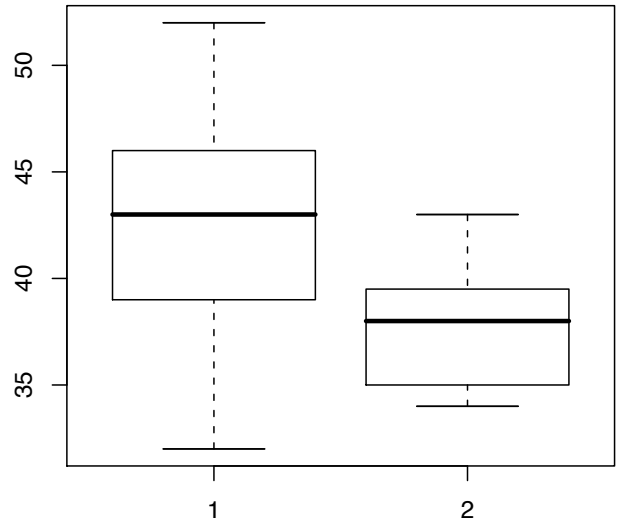


Figure 20.2: Vitamin C content in milligrams per 100 grams, measured at the factory and measured 5 month later in Haiti.

gives essentially the same output.

In addition, the output

```
> t.test(haiti, alternative = c("less"),mu = 40)

One Sample t-test

data:  haiti
t = -5.3232, df = 26, p-value = 7.175e-06
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
     -Inf 38.23811
sample estimates:
mean of x
 37.40741
```

shows that we would reject the one sided test

$$H_0 : \mu \geq 40 \quad \textit{versus} \quad H_1 : \mu < 40,$$

based on a goal of having 40mg/100g vitamin C in the wheat soy blend consumed by the Haitians.

We have used R primarily to compute a confidence interval. If the goal of the program is to have reduction in vitamin C be less that a given amount c, then we have the hypothesis

$$H_0 : \mu_F - \mu_H \geq c \quad \text{versus} \quad H_1 : \mu_F - \mu_H < c.$$

We can test this using R by replacing `mu=0` with `mu=c`.

20.5 Two Sample Procedures

Now we consider the situation in which the two samples

$$X_1, X_2, \ldots, X_{n_X}, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_{n_Y}$$

are independent but are not paired. In particular, the number of observations n_X and n_Y in the two samples could be different. If the first sample has common mean μ_X and variance σ_X^2 and the second sample has common mean μ_Y and variance σ_Y^2, then

$$E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y \quad \text{and} \quad \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}.$$

For the two sided hypothesis test

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y,$$

The corresponding t-statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \tag{20.1}$$

with s_X^2 and s_Y^2 the unbiased sample variances. Unlike the match pairs procedures, the test statistic (20.1) does not have a t distribution under the null hypothesis. Indeed, the density and the distribution of this statistic are difficult to compute.

In this circumstance, we now make what is commonly known in statistics as a *conservative* approximation. We replace the actual distribution of the t statistic in (20.1) with one which has slightly bigger tails. Thus, the computed p-value which are just integrals of the density function will be slightly larger. In this way, a conservative procedures is one that does not decrease the type I error probability.

This goal can be accomplished by approximating an ordinary Student's t distribution with the effective degrees of freedom ν calculated using the **Welch-Satterthwaite** equation:

$$\nu = \frac{(s_X^2/n_X + s_Y^2/n_Y)^2}{(s_X^2/n_X)^2/(n_X - 1) + (s_Y^2/n_Y)^2/(n_Y - 1)}. \tag{20.2}$$

As we saw in our discussion on Interval Estimation, this also gives a γ-level confidence interval for the difference in the means μ_x and μ_Y.

$$\bar{x} - \bar{y} \pm t_{(1-\gamma)/2,\nu} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}.$$

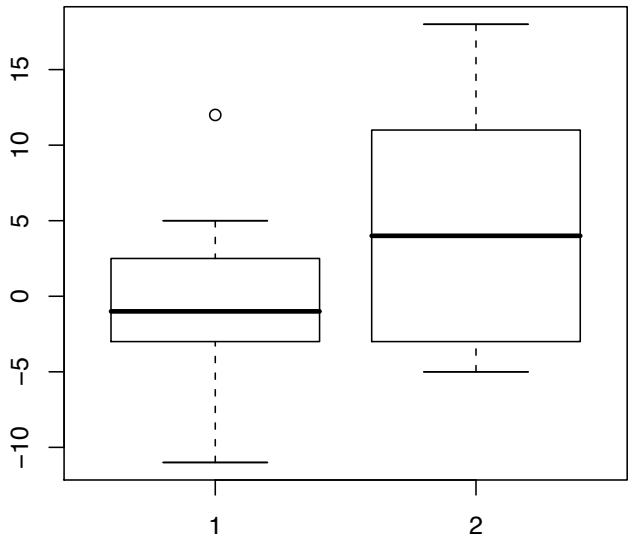
We also learned that the effective degrees of freedom are largest when the two sample variances are nearly equal. In this case the number of degrees of freedom is 2 fewer than the sum of the two sets of observations.

Example 20.3. *To investigate the effect on blood pressure of added calcium in the diet, a researchers conducts a double blind randomized experiment. In the treatment group, each individual receives a calcium supplement. In the control group, the individual takes a placebo. The response variable is the decrease in systolic blood pressure, measured in millimeters of mercury, after 12 weeks. The test subjects are all male.*

```
> calcium<-c(7,-4,18,17,-3,-5,1,10,11,-2)
> mean(calcium)
[1] 5
> sd(calcium)
[1] 8.743251
> placebo<-c(-1,12,-1,-3,3,-5,5,2,-11,-1,-3)
> mean(placebo)
[1] -0.2727273
> sd(placebo)
[1] 5.900693
> boxplot(placebo,calcium)
```

The null hypothesis is that the treatment did not reduce μ_t the mean blood pressure of the treatment any more than it did the mean μ_c for the control group. The alternative is that it did reduce blood pressure more. Formally the hypothesis test is

$$H_0 : \mu_c \leq \mu_t \quad \text{versus} \quad H_1 : \mu_c > \mu_t.$$



The t-statistic is

$$t = \frac{5.000 + 0.273}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.604.$$

```
> t.test(calcium,placebo,alternative = c("greater"))

Welch Two Sample t-test

data:  calcium and placebo
t = 1.6037, df = 15.591, p-value = 0.06442
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.476678       Inf
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Thus. the evidence against the null hypothesis is modest with a p-value of about 6%. Notice that the effective degrees of freedom is $\nu = 15.591$. The maximum possible is value for degrees of freedom is 19.

To see a 90% confidence interval remove the "greater than" alternative" and set the confidence level.

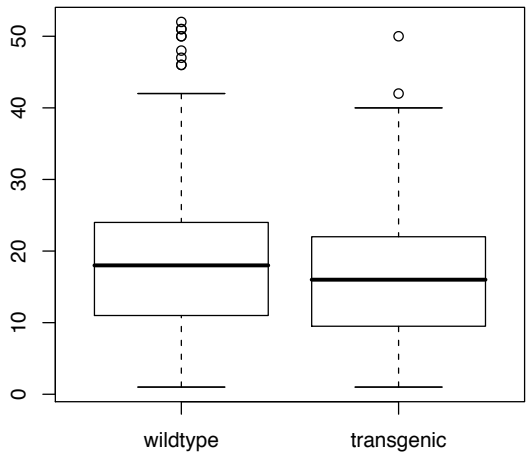
```
> t.test(calcium,placebo,conf.level = 0.9)

Welch Two Sample t-test

data:  calcium and placebo
t = 1.6037, df = 15.591, p-value = 0.1288
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.476678 11.022133
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Example 20.4. *The life span in days of 88 wildtype and 99 transgenic mosquitoes is given in the following data set.*

```
> mosquitoes<-read.delim("http://math.arizona.edu/~jwatkins/mosquitoes.txt")
> boxplot(mosquitoes)
```



The goal is to see if overstimulation of the insulin signaling cascade in the mosquito midgut reduces the μ_t, the mean life span of these transgenic mosquitoes from that of the wild type μ_{wt}.

$$H_0 : \mu_{wt} \leq \mu_t \quad \text{versus} \quad H_1 : \mu_{wt} > \mu_t.$$

```
> wildtype<-mosquitoes[1:88,1]
> transgenic<-mosquitoes[,2]
> t.test(transgenic,wildtype,alternative = c("less"))

Welch Two Sample t-test

data:  transgenic and wildtype
t = -2.4106, df = 169.665, p-value = 0.008497
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
      -Inf -1.330591
sample estimates:
mean of x mean of y
 16.54545  20.78409
```

To determine a 98% confidence interval, we again remove the alternative command.

```
> t.test(transgenic,wildtype,conf.level=0.98)

Welch Two Sample t-test

data:  transgenic and wildtype
t = -2.4106, df = 169.665, p-value = 0.01699
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -8.3680812 -0.1091915
sample estimates:
mean of x mean of y
 16.54545  20.78409
```

Exercise 20.5. *Notice that the 98% confidence interval,* $(-8.3680812, -0.1091915)$ *does not contain 0. What can be said about a two-sided test at the 2% significance level? What can be said about the p-value for a one-sided test?*

Example 20.6 (pooled two-sample t-test)**.** *Sometimes, the two-sample procedure is based on the assumption of a common value* σ^2 *for the variance of the two-samples. In this case, we* **pool** *the data to compute an unbiased estimate for the variance:*

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2\right).$$

Thus, we weight the variance from each of the two samples by the number of degrees of freedom. If we modify the t-statistics above to

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

This indeed has the t-distribution with $n_X + n_Y - 2$ *degrees of freedom. This is accomplished in* R *by adding* `var.equal=TRUE` *command.*

```
> t.test(calcium,placebo,alternative = c("greater"),var.equal=TRUE)

Two Sample t-test

data:  calcium and placebo
t = 1.6341, df = 19, p-value = 0.05935
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3066129        Inf
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Note that this increases the number of degrees of freedom and lowers the P*-value from 6.4% to 5.9%. A natural question to ask at this point is* How do we compare the mean of more that two groups. *As we shall soon learn, this leads us to a test procedure, analysis of variance, that is a generalization of the pooled two-sample procedure.*

20.6 Summary of Tests of Significance

The procedures we have used to perform hypothesis tests are based on some quantity θ generally expressed as a value in a parameter space Θ. In setting the hypothesis, we partition the parameter space into two parts Θ_0 for the null hypothesis, H_0, and Θ_1 for the alternative, H_1. Our strategy is to look for generalizations of the Neyman-Pearson paradigm. For example, the Karlin-Rubin criterion provides a condition for one-sided tests that allows us to say the we have a uniformly most powerful test.

For two-sided tests, we look to the likelihood ratio approach. For this approach, we first maximize the likelihood $L(\theta|\mathbf{x})$ both over Θ_0 and over Θ and then compute the ratio $\Lambda(\mathbf{x})$. If the data, $\mathbf{x}$, lead to a ratio that is sufficiently small, then likelihood for all values of $\theta \in \Theta_0$ are less likely then some values in Θ_1. This leads us to reject the null hypothesis in favor of the alternative. If the number of observations is large, then we can approximate, under the null hypothesis, the distribution of the test-statistic $-2\ln\Lambda(\mathbf{x})$ with a χ^2 distribution. This leads to a critical region $C = \{-2\ln\Lambda(\mathbf{x}) \geq \tilde{k}_\alpha\}$ for an α-level test.

In practice, much of our inference is for population proportions and the population means. In these cases, we often reserve the test for those cases in which the central limit theorem applies and thus the estimates, the sample proportions and the sample means, have approximately a normal distribution. We summarize these procedures below.

20.6.1 General Guidelines

- Hypotheses are stated in terms of a *population parameter*.
- The null hypothesis H_0 is a statement that no effect is present.
- The alternative hypothesis H_1 is a statement that a parameter differs from its null value in a specific direction (one-sided alternative) or in either direction (two-sided alternative).
- A test statistic is designed to assess the strength of evidence against H_0.
- If a decision must be made, specify the significance level α.
- Assuming H_0 is true, the p-value is the probability that the test statistic would take a value as extreme or more extreme than the value observed.
- If the p-value is smaller than the significance level α, then H_0 is rejected and the data are said to be *statistically significant at level* α.

20.6.2 Test for Population Proportions

The design is based on Bernoulli trials . For this we have

- A fixed number of trials n.
- The outcome of each trial is independent of the other trials.
- Each trial has one of two outcomes **success** and **failure**.
- The probability of success p is the same for each trial.

This test statistic is the z-score. and thus is based the applicability of the central limit theorem on using the standard normal distribution. This procedure is considered valid is the sample is small ($< 10\%$) compared to the total population and both np_0 and $n(1-p_0)$ is at least 10. Otherwise, use the binomial distribution directly for the test statistic.

The statistics $\hat{p}$ for a one-proportion procedure and $\hat{p}_1$, $\hat{p}_2$ for a two-sample procedure, is the appropriate proportions of success.

	null hypothesis		
sample proportions	one-sided	two-sided	test statistic
single proportion	$H_0 : p \geq p_0$ $H_0 : p \leq p_0$	$H_0 : p = p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
two proportions	$H_0 : p_1 \geq p_2$ $H_0 : p_1 \leq p_2$	$H_0 : p_1 = p_2$	$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$

The pooled sample proportion $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ where x_i is the number of successes in the n_i Bernoulli trials from group i.

20.6.3 Test for Population Means

- Use the z-statistic when the standard deviations are known.
- Use the t-statistic when the standard deviations are computed from the data.

	null hypothesis	
t or z-procedure	one-sided	two-sided
single sample	$H_0 : \mu \leq \mu_0$ $H_0 : \mu \geq \mu_0$	$H_0 : \mu = \mu_0$
two samples	$H_0 : \mu_1 \leq \mu_2$ $H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 = \mu_2$

The test statistic

$$t = \frac{\text{estimate} - \text{parameter}}{\text{standard error}}.$$

The p-value is determined by the distribution of a random variable having a t distribution with the appropriate number of degrees of freedom. For one-sample and two-sample z procedures, replace the values s with σ and s_1 and s_2 with σ_1 and σ_2, respectively. Use the normal distribution for these tests.

t-procedure	parameter	estimate	standard error	degrees of freedom
one sample	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$	$n-1$
two sample	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	ν in equation (20.2)
pooled two sample	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$n_1 + n_2 - 2$

20.7 A Note on the Delta Method

For a one sample test hypothesizing a value for $g(\mu)$, we use the t statistic

$$t = \frac{g(\bar{x}) - g(\mu_0)}{|g'(\bar{x})|s/\sqrt{n}}$$

and base the test on the t distribution with $n-1$ degrees of freedom.

For a test that compare a function of the mean of a two samples $g(\mu_X)$ and $g(\mu_Y)$ we can use the test statistic

$$t = \frac{g'(\bar{x}) - g(\bar{y})}{\sqrt{\frac{(g'(\bar{x})s_X)^2}{n_X} + \frac{(g'(\bar{y})s_Y)^2}{n_Y}}}$$

The degrees of freedom ν can be computed from the Welch-Satterthwaite equation specialized to this circumstance.

$$\nu = \frac{(g(\bar{x})s_X)^2/n_X + (g'(\bar{y})s_Y)^2/n_Y)^2}{((g'(\bar{x})s_X)^2/n_X)^2/(n_X - 1) + ((g'(\bar{y})s_Y)^2/n_Y)^2/(n_Y - 1)}.$$

20.8 The t Test as a Likelihood Ratio Test

Again, we begin with independent normal observations $X_1 \ldots, X_n$ with unknown mean μ and unknown variance σ^2. We show that the critical region

$$C = \{\mathbf{x}; |T(\mathbf{x})| > t_{n-1,\alpha/2}\}$$

is a consequence of the criterion given by a likelihood ratio test with significance level α

The likelihood function

$$L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

Thus, $\hat{\mu} = \bar{x}$.

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Thus,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

For the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

the **likelihood ratio test**

$$\Lambda(x) = \frac{L(\mu_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})}$$

where the value

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2$$

gives the maximum likelihood on the set $\mu = \mu_0$.

$$L(\mu_0, \hat{\sigma}_0^2|\mathbf{x}) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}} \exp -\frac{1}{2\hat{\sigma}_0^2} \sum_{i=1}^{n} (x_i - \mu_0)^2 = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}} \exp -\frac{2}{n},$$

$$L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp -\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}, \exp -\frac{2}{n},$$

and the likelihood ratio is

$$\Lambda(\mathbf{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{-n/2}$$

The critical region $\lambda(\mathbf{x}) \leq \lambda_0$ is equivalent to the fraction in parenthesis above being sufficiently large. In other words for some value c,

$$\begin{aligned} c &\leq \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n \left((x_i - \bar{x}) + (\bar{x} - \mu_0)\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + 2\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu_0) + \sum_{i=1}^n (\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

In a now familiar strategy, we have added and subtracted $\bar{x}$ to decompose the variation. Continuing we find that

$$(c-1)(n-1) \leq \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)} = \frac{(\bar{x} - \mu_0)^2}{s^2/n}$$

or

$$(c-1)(n-1) \leq T(\mathbf{x})^2 \tag{20.3}$$

where

$$T(\mathbf{x}) = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and s is the square root of the *unbiased* estimator of the variance

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

Taking square roots in (20.3), we have the critical region

$$C = \left\{\mathbf{x}; \sqrt{(c-1)(n-1)} \leq |T(\mathbf{x})|\right\}$$

Now take $\sqrt{(c-1)(n-1)} = t_{n-1,\alpha/2}$.generally

20.9 Non-parametric alternatives

When the assumption of the normal distribution or the central limit theorem cannot be said to hold in a two-sample t-procedure, then a second option is to use a test that does not depend on the numerical values of the observations but rather on the *ranks* of the data. Because this test is based on the ranks of the data, we cannot base the hypothesis test on the means of the data, but rather on a statistics that uses only the ranks of the data. For these **non-parametric tests** the hypotheses are stated in terms of medians or, more generally, ranks.

20.9.1 Mann-Whitney or Wilcoxon Rank Sum Test

We will explain this procedure more carefully in the case of the data on the lifetime of wildtype and transgenic mosquitoes. In this case our data are

$$x_1, x_2, \ldots, x_{n_1}$$

are the lifetimes in days for the wildtype mosquitoes and

$$y_1, y_2, \ldots, y_{n_2}$$

are the lifetimes in days for the transgenic mosquitoes.

For the **Wilcoxon signed-rank test**, the hypothesis is based on the median values m_1 and m_2. Here

- m_1 is the median lifetime in days for wildtype mosquitoes, and
- m_2 is the median lifetime in days for transgenic mosquitoes.

$$H_0 : m_2 \geq m_1 \quad \text{versus} \quad H_1 : m_2 < m_1.$$

The following identity will be useful in our discussion.

Exercise 20.7. *The sums of the first m integers*

$$\sum_{j=1}^{m} j = 1 + 2 + \cdots + m = \frac{m(m+1)}{2}$$

Let's look at a small pilot data set to get a sense of the procedure. The values are the lifespan in days.

```
> wildtype
[1] 31 36 18 11 33  9 34
> transgenic
[1]  3  8 21 24 25
```

From these date, we see that the ranks

- transgenic - 1 2 6 7 8
- wildtype - 3 4 5 9 10 11 12

The strategy for the test to to see if there is a significant difference in the ranks of the data. The basic statistic is the sum of the ranks of one of the samples. For the transgenic mosquitoes, this sum is

$$R_1 = 1 + 2 + 6 + 7 + 8 = 24$$

for $n_1 = 5$ observations

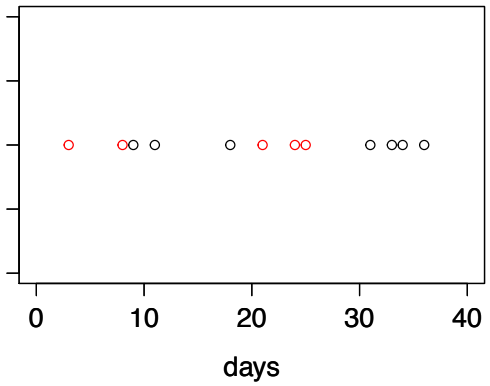


Figure 20.3: Life span in days for 5 transgenic (in red) and 7 wildtype (in black) mosquitoes.

We can now compare this sum of ranking to all $\binom{12}{5} = 792$ possible rankings of the data. This is accomplished using the `wilcox.test` command in R.

```
> wilcox.test(transgenic,wildtype,alternative=c("less"))

Wilcoxon rank sum test

data:  transgenic and wildtype
W = 9, p-value = 0.101
alternative hypothesis: true location shift is less than 0
```

Thus, the 10.1% of the ranks below the given value of 24 give us the p-value. This small amount of data gives a hint that the transgenic mosquito may have a shorter lifespan. The U statistic is related to the sum of the ranks by subtracting the minimum possible value as shown in Exercise 20.6.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}.$$

R uses another variant W of the $R_!$ statistic.

If the entire data set is used, then we cannot carry out all of the comparisons without a long computational time. However, we have a version of the central limit theorem that gives the mean and standard deviation of the R_1, U_1 or W-statistic. Thus, we can use the normal distribution to determine a p-value. As with the binomial distribution, R uses a continuity correction to deal with the fact that the test statistic W is a discrete random variable.

```
> wilcox.test(transgenic,wildtype,alternative=c("less"))

Wilcoxon rank sum test with continuity correction

data:  transgenic and wildtype
W = 3549.5, p-value = 0.0143
alternative hypothesis: true location shift is less than 0
```

Notice the value $W = 3549.5$ is not an integer. This is a result of the fact that ties are resolved by giving fractional values to ties. For example, if the third and fourth values are equal, they are both given the rank 3 1/2. If the seventh, eighth, and ninth values are equal. they are all given the rank (7+8+9)/3 = 8.

Exercise 20.8. *Define R_2 to be the sum of the ranks for the n_2 observations in the second sample and set*

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

Then, $U_1 + U_2 = n_1 n_2$.

20.9.2 Wilcoxon Signed-Rank Test

For the **Wilcoxon signed-rank test**, the hypothesis is based on the median values m_x and m_y for an experimental procedure in which pairs are matched. This gives an alternative to the measurement of the amount of vitamin C in wheat soy blend at the factory and 5 months later in Haiti. Our data are

$$x_1, x_2, \ldots, x_n$$

are the measurements of vitamin C content of the wheat soy blend at the factory and

$$y_1, y_2, \ldots, y_n$$

are the measurements of vitamin C content of the wheat soy blend in Haiti.

For the hypothesis test to see if vitamin C content decreases due to shipping and shelf time, set

- m_F is the median vitamin C content of the wheat soy blend at the factory and
- m_H is the median vitamin C content of the wheat soy blend in Haiti.

To perform the test

$$H_0 : m_F \leq m_H \quad \text{versus} \quad H_1 : m_F > m_H, ,$$

we use a test statistic based on both the sign of the difference $y_i - x_i$ in the paired observations and in the ranks of $|y_i - x_i|$ Here is the R command and output. Note the choice `paired=TRUE` for the signed-rank test.

```
> wilcox.test(factory, haiti, alternative = c("greater"),paired=TRUE)

Wilcoxon signed rank test with continuity correction

data:  factory and haiti
V = 341, p-value = 0.0001341
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(factory, haiti, alternative = c("greater"),  :
  cannot compute exact p-value with ties
```

20.10 Answers to Selected Exercises

20.5. This means that we can rejec the null hypothesis that transgenic and wildtype mosquitoes have the seem mean lifetime. For a one-sided test, the p-value is 0.01699/2 = 0.008549.

20.6. By the correspondence between two-sided hypothesis tests and confidence intervals, the fact that 0 is not in the confidence interval, indicates that the test *is* significant at the 2% level and thus the p-value < 0.02. Notice that the output shows a p-value of 0.01699. For a one-sided test, we know that the p-value is half that of a two-sided test and thus below 0.01. Notice that the R give a p-value of 0.008497 for a one-sided test.

20.7. We prove this using mathematical induction. For the case $m = 1$, we have $1 = \frac{1(1+1)}{2}$ and the identity holds true. Now assume that the identity holds for $m = k$. We then check that it also holds for $m = k + 1$

$$1 + 2 + \cdots + k + (k + 1) = \frac{k(k+1)}{2} + (k + 1) = \left(\frac{k}{2} + 1\right)(k + 1) = \frac{k+2}{2}(k + 1) = \frac{(k + 1)(k + 2)}{2}.$$

So, by the principle of mathematical induction, we have the identity for all non-negative integers.

20.8. By Exercise 20.6, the sum of the ranks

$$R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

Thus,

$$\begin{aligned} U_1 + U_2 &= \left(R_1 - \frac{n_1(n_1 + 1)}{2}\right) + \left(R_2 - \frac{n_2(n_2 + 1)}{2}\right) \\ &= \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - \frac{n_1(n_1 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} \\ &= \frac{1}{2}(n_1(n_1 + 1) + n_1n_2 + n_2(n_2 + 1) + n_1n_2 - n_1(n_1 + 1) - n_2(n_2 + 1)) = n_1n_2. \end{aligned}$$