Chapter 9: Hypothesis Testing

Wong Shujia

Contents

1 Basic Concepts and Notation

1.1 Introduction

The Elements of a Statistical Hypothesis
Null Hypothesis Signifcance Testing (NHST)

1. The **null hypothesis**, denoted by H_0, is usually the nullification of a claim. Unless evidence from the data indicates otherwise, the null hypothesis is assumed to be true.
2. The **alternate hypothesis**, denoted by H_1 (or sometimes denoted by H_a), is customarily the claim itself.
3. The **test statistic**, denoted by TS, is a function of the sample measurements upon which the statistical decision will be based.
4. A **rejection region** (or a *critical region*) is the region that specifies the values of the observed test statistic for which the null hypothesis will be rejected.
5. **Conclusion**: Reject or fail to reject the null hypothesis.
6. **Interpret the results**: State in words what the conclusion means to the problem we started with.

Form of hypothesis tests

- $H_0 : \theta \in \Theta_0$,where Θ_0 is a subset of the *parameter space* Θ.
  - If $\Theta_0 = \{\theta_0\}$, then the hypothesis is said to be **simple**; Otherwise is called **composite**.
- $H_1 : \theta \in \Theta_1$ where $\Theta_1 \subset \Theta$ and $\Theta_0 \cup \Theta_1 = \Theta$

Null Hypothesis	Alternative Hypothesis	Type of Alternative
$H_0 : \theta = \theta_0$	$\theta < \theta_0$	lower one-sided
	$\theta > \theta_0$	upper one-sided
	$\theta \neq \theta_0$	two-sided

1.2 Type I and II Errors

Type I and II Errors

There are two types of errors you can make when testing

- Type I Error: Rejecting a true H_0.
- Type II Error: Not rejecting a false H_0.

Significance Level (also known as the **size of the test**):

$$\alpha = \max_{\theta \in \Theta_0} P(\text{reject } H_0 | H_0 \text{ is true})$$

The probability of committing a type II error is β, where

$$\beta = \text{P(type II error)} = \text{P(accept } H_0 \; | H_0 \text{ is false)}$$

Relationship between type I and type II errors

		Decision (Based on data)	
		Reject H_0	Fail To Reject H_0
H_0	True	Type I Error P(reject $H_0 \vert H_0$)=α (Level of Significance)	Correct Decision P(accept H_0 $\vert H_0$)$= 1-\alpha$ (Confidence Level)
	False	Correct Decision P(accept H_1 $\vert H_1$)$= 1-\beta$ (Power of the Test)	Type II Error P(reject H_1 $\vert H_1$)$= \beta$

1.3 Power Function

Power Function

The **power function** of a test is the probability that the test rejects H_0, when the actual parameter value is θ.

$$\text{Power}(\theta) = P(\text{reject } H_0 | \theta \in \Theta)$$

- When $\theta \in \Theta_1$ that is H_1 is true,

  $$\text{Power}(\theta) = P(\text{reject } H_0 | \theta \in \Theta_1) = 1 - \beta(\theta)$$

  where $\beta(\theta)$ is the probability of a type II error at a given θ.

- When the null hypothesis is simple, $\theta = \theta_0$,

  $$\text{Power}(\theta) = \begin{cases} 1-\beta(\theta) & \theta \in \Theta_1 \\ \alpha & \theta = \theta_0 \end{cases}$$

- In statistical tests, it is generally accepted that power should be 0.8 or greater.

Example: Achievement Test

An achievement test score is assumed $X \sim N(\mu, 6^2)$, to test $H_0 : \mu = 40$ vs $H_1 : \mu \neq 40$.

1. Find the probability of type I error for n = 9.
   - Decision rule: Reject the null hypothesis if the sample mean is less than 36 or greater than 44.
2. Find the probability of type I error for n = 36.
   - Decision rule: Reject the null hypothesis if the sample mean is less than 38 or greater than 42.
3. Plot the power functions for n = 9 and n = 36 for values of μ between 30 and 50.

Solution

Note that $\overline{X} \sim N(\mu, 6^2/\sqrt{n})$

(a) The probability of type I error for n = 9 is

$$\begin{aligned} P(\text{Type I error}) &= P(\overline{X} < 36|\mu = 40) + P(\overline{X} > 44|\mu = 40) \\ &= P(Z < \frac{36-40}{6/\sqrt{9}}) + P(Z > \frac{44-40}{6/\sqrt{9}}) \\ &= \Phi(-2) + 1 - \Phi(2) = 0.0455. \end{aligned}$$

(b) The probability of type I error for n = 36 is

$$\begin{aligned} P(\text{Type I error}) &= P(\overline{X} < 38|\mu = 40) + P(\overline{X} > 42|\mu = 40) \\ &= P(Z < \frac{38-40}{6/\sqrt{36}}) + P(Z > \frac{42-40}{6/\sqrt{36}}) \\ &= \Phi(-2) + 1 - \Phi(2) = 0.0455. \end{aligned}$$

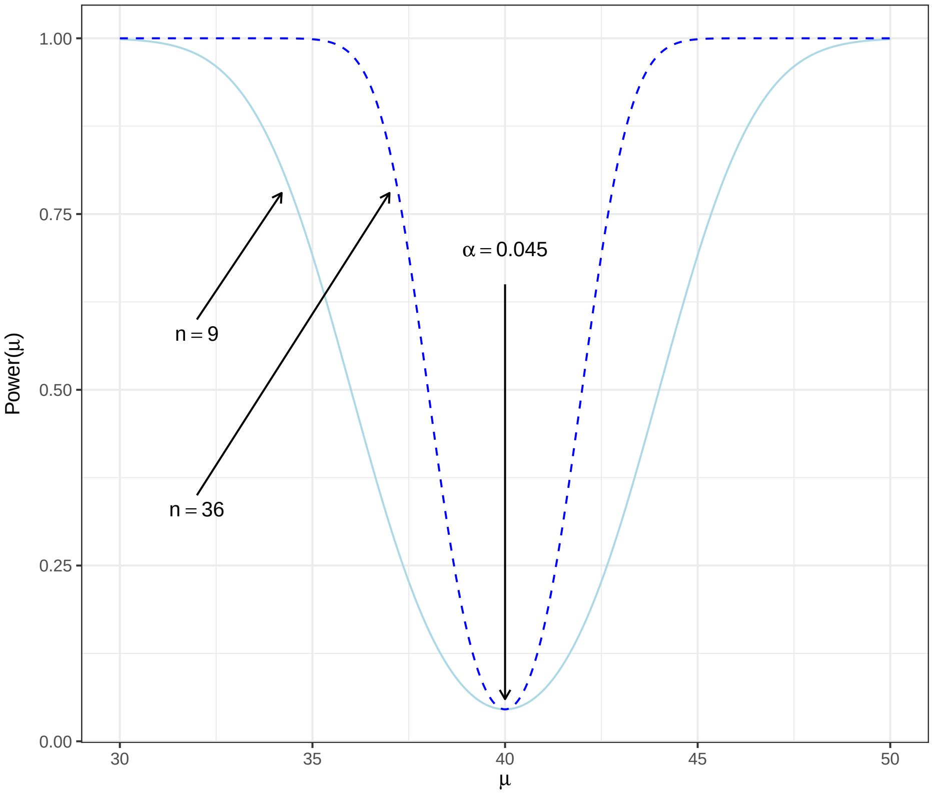
Solution

(c) The power function for n = 9 is

$$\begin{aligned} \text{Power}(\mu) &= P(\overline{X} < 36|\mu \neq 40) + P(\overline{X} > 44|\mu \neq 40) \\ &= P(Z < \frac{36-\mu}{6/\sqrt{9}}) + P(Z > \frac{44-\mu}{6/\sqrt{9}}) \\ &= \Phi(18 - 0.5\mu) + 1 - \Phi(22 - 0.5\mu). \end{aligned}$$

The power function for n = 36 is

$$\begin{aligned} \text{Power}(\mu) &= P(\overline{X} < 38|\mu \neq 40) + P(\overline{X} > 42|\mu \neq 40) \\ &= P(Z < \frac{38-\mu}{6/\sqrt{36}}) + P(Z > \frac{42-\mu}{6/\sqrt{36}}) \\ &= \Phi(38 - \mu) + 1 - \Phi(42 - \mu). \end{aligned}$$



Uniformly Most Powerful (UMP) Test

Definition 1 (Most Powerful Test)**.** To test of the simple hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, if a test that has the largest power(θ) among tests with no larger size α, then the test is called **most powerful**.

Definition 2 (Uniformly Most Powerful Test)**.** To test the simple hypothesis $H_0 : \theta = \theta_0$ versus a composite alternative H_1(e.g. $\theta > \theta_0$), If a test which has the greatest power among all possible tests of a given size α, then the test is called a **uniformly most powerful** test.

Example: Most Powerful Test

Given a $N(\mu, 1)$ population from which one takes a simple random sample of size 1, test the null hypothesis

$$H_0 : \mu = 1 \text{ versus } H_1 : \mu = 2.$$

Determine the significance level and the power of the test for the following rejection regions:

1. $(2.0364, +\infty)$
2. $(1.1000, 1.3000) \cup (2.4617, +\infty)$

Solution

(a) Rejection regions: $(2.0364, +\infty)$

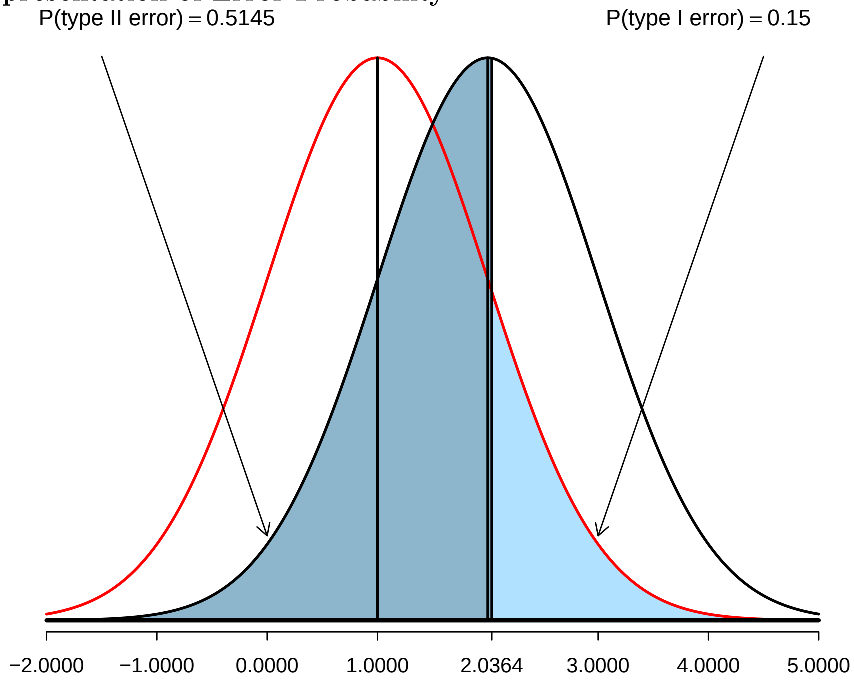
$$\begin{aligned}\alpha &= P(X > 2.0364|N(1,1)) = 1 - \Phi(\frac{2.0364 - 1}{1}) = 0.15 \\ \beta &= P(X > 2.0364|N(2,1)) = 1 - \Phi(\frac{2.0364 - 2}{1}) = 0.5145\end{aligned}$$

The power of the test is $1 - \beta = 1 - 0.5145 = 0.4855$.

(b) Since the rejection region is$(1.1000, 1.3000) \cup (2.4617, \infty)$,

$$\begin{aligned}\alpha &= P(1.1000 < X < 1.3000|N(1,1) + P(X > 2.4617|N(1,1)) \\ &= 0.15 \\ \beta &= P(X \leq 1.1000|N(2,1) + P(1.3000 \leq X \leq 2.4617|N(2,1)) \\ &= 0.6199\end{aligned}$$

Graphical Representation of Error Probability



1.4 P-Value

p-value

The p-value is defined as the probability of observing a difference *as* **extreme or more extreme than** the difference observed under the assumption that the null hypothesis is true.

In other words, the p-value is the **tail area** beyond the observed value of the test statistic.

1. Testing decision rule with p-value
   - Given a significance level α, if the p-value $< \alpha$, reject H_0; otherwise, fail to reject.
2. General rule for computing the p-value:
   - alternative is a "$<$" $\Rightarrow$ left tail area: $P(T \leq t_{obs}|H_0)$
   - alternative is a "$>$" $\Rightarrow$ right tail area: $P(T \geq t_{obs}|H_0)$
   - alternative is a "$\neq$" $\Rightarrow$ $2\times$ smaller tail area:

$$2\min\{P(T \leq t_{obs}|H_0), P(T \geq t_{obs}|H_0)\}$$

Misinterpretation of p-values

1. A p-value is the probability that the null hypothesis is true.
   - True or false: if a p-value was .07, then there was a 93% chance that the alternative hypothesis was correct.
   - In fact, the test assumes that the null hypothesis is correct. It tells us the proportion of trials for which we would receive a result as extreme or more extreme than the one we did if the null hypothesis was correct.
2. A p-value is a measure of the size of an effect.
   - In research papers, it is common to attach phrases like *highly signifcant* and *very highly signifcant* to p-values that are much smaller than .05 (like .01 with "**" and .001 with "***").
   - It is common to interpret p-values such as these, and statements such as these, as signaling a bigger effect than p-values that are only modestly less than .05
   - This is a mistake. this is conflating statistical signifcance with practical signifcance.

Statistical Significance vs Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in **statistical significance** even for tiny differences between the sample mean and the null value, even when the difference is not **practically significant**.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results–*effect size* (we want observed differences to be *real*, but also large enough *to matter*).
- This inability for the statistical machinery to distinguish between the two types of significance is sometimes called *Lindley's paradox*: With enough data, you'll reject any hypothesis at all!

Hypothesis Testing Procedure

Step 1: Set the hypotheses

Step 2: Select the significance level α

- $\alpha = 0.05$ (the 5% level) is the most commonly used.
- $\alpha = 0.01$ (the 1% level) is prevalent in medical and quality assurance examples.

Step 3: Compute the test statistic (Z or T)

Step 4: Formulate the decision rule (rejection region or p-value)

- For example, reject the Null hypothesis if the p-value< 0.05; or $|Z| > Z_{1-\alpha/2}$

Step 5: Make a decision, and interpret it in context of the research question

Example: Testing of heights

Example 3. A random sample of 36 female college-aged dancers was obtained and their heights (in inches) were measured, the sample information is given by $\bar{x} = 63.6$ inches and $s = 2.13$ inches.The average height of all college-aged females is 64.5 inches. Do these data provide convincing evidence that the average height of female college-aged dancers is *lower* from this value?

Testing Procedure

1. $H_0 : \mu = 64.5 \quad H_1 : \mu < 64.5$
2. Select the significance level $\alpha = 0.05$
3. Test statistic
$$\bar{X} \sim N(\text{mean} = 64.5,\ \text{SE} = \frac{2.13}{\sqrt{36}} = 0.355)$$
$$Z = \frac{63.6 - 64.5}{0.355} = -2.54$$
4. Decision: the rejection region is $Z < Z_{0.025} = -1.96$; the p-value is $P(Z < -2.54) = 0.0055$.
5. Conclusion: Since p-value < 0.05, reject H_0.
6. Intepretation: the data provide convincing evidence that the average height of female college-aged dancers is lower than 64.5 inches.

Duality of Confidence Intervals and Tests of Significance

Confidence intervals and **hypothesis tests** are almost always agree, as long as the two methods use equivalent levels of significance/confidence.

- A two sided hypothesis with significance level α is equivalent to a confidence interval with confidence level $CL = 1 - \alpha$.
  - A one sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - 2\alpha$.
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

Derivation

Consider $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$, and assume $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

The region that fail to reject H_0 with significance level α is

$$\begin{aligned} |z_{obs}| \leq z_{1-\alpha/2} \quad &\Leftrightarrow \quad |\bar{x} - \mu_0| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\Leftrightarrow \quad \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

This implies that μ_0 is in the $(1-\alpha)100\%$ confidence interval.

Question 1

Which of the following is the correct interpretation of the p-value?

1. If in fact the average height of college aged dancers is less than 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.
2. *If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.*
3. If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches is 0.0055.
4. The probability that the average height of college aged dancers is 64.5 inches is 0.0055.

Question 2

What is the equivalent confidence level for this two-sided hypothesis test with $\alpha = 0.05$?

1. 80%
2. *90%*
3. 95%
4. 99.7%
5. 97.5%

Question 3

If we were to calculate a 95% confidence interval for the average height of college-aged dancers, would this interval include the null value (64.5 inches)?

1. Yes
2. No
3. *Cannot tell without calculating the interval*

2 Hypothesis Testing for a Single Parameter

2.1 Testing of Hypotheses for a Population Mean

Testing a Population Mean: Variance Known

A sample $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known.

Hypotheses $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$

Select the significance level α, the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

Rejection region

$$|z_{obs}| > z_{1-\alpha/2} \text{ or } p\text{-value} = P(|Z| > z_{obs}|H_0)$$

- $H_1 : \mu > \mu_0$, Rejection region $z_{obs} > z_{1-\alpha}$
- $H_1 : \mu < \mu_0$, Rejection region $z_{obs} < z_\alpha$

In R PASWR2: *z.test(x, sigma.x, y, sigma.y, alternative = ...).*

Testing a Population Mean: Variance Unknown

A sample $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is unknown.

Hypotheses $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$

Select the significance level α

Test statisti

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t(n-1)$$

Rejection region

$$|t_{obs}| > t_{1-\alpha/2}(n-1) \text{ or } p\text{-value} = P(|T| > t_{obs}|H_0)$$

R function for testing one population mean: *t.test(x,...)*

- $H_1 : \mu > \mu_0$, Rejection region $t_{obs} > t_{1-\alpha}(n-1)$
- $H_1 : \mu < \mu_0$, Rejection region $t_{obs} < t_\alpha(n-1)$

Example: McFarl and Insurance Co.

The claims department at McFarland Insurance reports that the mean cost to process a claim is \$60.

After implementing some cost-cutting measures, McFarland selected a random sample of 26 claims and found that the sample mean was \$57 and the standard deviation was \$10.

At the 1% level should they conclude the cost-cutting measures actually reduced cost?

Set-Up The t-test

Step 1: Null $H_0 : \mu = 60$ versus Alternative $H_1 : \mu < 60$ (One side test)

Step 2: Significance level $\alpha = 0.01$

Step 3: The test statistic is

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where $\mu_0 = 60$, $s = 10$ and $n = 26$.

As $n < 30$, we'll use the t-distribution.

Set-Up The t-test

Step 4: Formulate the decision rule

```
> xbar=57
> mu0=60
> s=10
> n=26
> alpha <- 0.001
> se=s/sqrt(n)
> tobs=(xbar-mu0)/se  # t statistic
> ct <- qt(alpha, df = n-1) # critical value
> pvalue<-pt(tobs,n-1) #p-value,one-sided
> c(ct = ct, tobs = tobs, pvalue = pvalue)

##          ct        tobs      pvalue
## -3.45018873 -1.52970585  0.06932281
```

What conclusions do you draw from this data? ($\alpha = 0.01$)

2.2 Testing of Hypotheses for a Population Proportion

Testing a Proportion

The null hypothesis: $H_0: \; p = p_0$, given a significance level α, for large sample, the appropriate test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \overset{H_0}{\sim} N(0,1)$$

R function: *prop.test(x,n,p,...)*

Rejection region:

- $H_1 : p > p_0$, $z > z_{1-\alpha}$
- $H_1 : p < p_0$, $z < z_\alpha$
- $H_1 : p \neq p_0$, $|z| > z_{1-\alpha/2}$

Example: Graduates' Jobs

A recent report claimed that 20% of all college graduates find a job in their chosen field of study. A random sample of 500 graduates found that 90 obtained work in their field.

Null hypothesis: $H_0: \; p = 0.2, n = 500, x = 90$.

```
> prop.test(90,900,0.2,alternative ="less")

##
##  1-sample proportions test with continuity correction
##
## data:  90 out of 900, null probability 0.2
## X-squared = 55.627, df = 1, p-value = 4.381e-14
## alternative hypothesis: true p is less than 0.2
## 95 percent confidence interval:
##  0.0000000 0.1182606
## sample estimates:
##   p
## 0.1
```

What conclusions do you draw from this analysis?

2.3 Testing of Hypotheses for a Population Variance

Testing a Population Variance

A sample $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is unknown. Null hypotheses $H_0 : \sigma^2 = \sigma_0^2$
Select the significance level α , the test statistic is

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1)$$

Rejection region:

- $H_1 : \sigma^2 > \sigma_0^2$, $\chi^2_{obs} > \chi^2_{1-\alpha}(n-1)$
- $H_1 : \sigma^2 < \sigma_0^2$, $\chi^2_{obs} < \chi^2_{\alpha}(n-1)$
- $H_1 : \sigma^2 \neq \sigma_0^2$, $\chi^2_{obs} < \chi^2_{\alpha/2}(n-1)$ or $\chi^2_{obs} > \chi^2_{1-\alpha/2}(n-1)$

Example: Testing the washers diameter variability

The quality control office of a large hardware manufacturer received more than twice the number of complaints it usually receives in reference to the *diameter variability* of its *4 cm* washers.

In light of the complaints, the quality control manager wants to ascertain whether or not there has been an increase in the diameter variability of the company's washers manufactured this month versus last month, where the variance was 0.004 cm^2.

- The manager takes a random sample of 20 washers manufactured this month. The results are stored in the data frame WASHER.
- Conduct an appropriate hypothesis test using a significance level of $\alpha = 0.05$.

Answer

Null hypotheses $H_0 : \sigma^2 = 0.004$, $H_1 : \sigma^2 > 0.004$
The rejection region is $\chi^2_{obs} > \chi^2_{1-\alpha}(n-1)$.

```
> library("PASWR2")
> cv <- qchisq(0.95,19) # Critical Value
> s2 <- var(WASHER$diameter)
> n <- sum(!is.na(WASHER$diameter))
> Chi2Obs <- (n - 1)*s2 / 0.004 #Standardized Test Statistic's Value
> pvalue <- pchisq(Chi2Obs, n-1, lower = FALSE)
>  c(n = n, CriticalValue = cv, Chi2Obs = Chi2Obs, pvalue = pvalue)

##             n CriticalValue       Chi2Obs        pvalue
##    20.0000000    30.1435272    25.2637500     0.1520425
```

What conclusions do you draw from this analysis?

3 Testing of Hypotheses for Two Samples

3.1 Testing of Hypotheses for Mean Differences

Aims of this subsection

We will discuss hypothesis testing for difference of two population means in the following cases:

1. Variance known
2. Variance unknown but equal
3. Variance unknown
4. Paired samples

As an example, we will discuss the stock performance of apple and buffett's Berkshire hathaway in details.

Test Two-Sample Means (Vars Known)

Assume random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are known.

The null hypothesis for testing the difference between two means is

$$H_0 : \mu_X - \mu_Y = \delta_0$$

Test statistic

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\mathrm{SE}(\overline{X} - \overline{Y})} \overset{H_0}{\sim} N(0, 1)$$

where

$$\mathrm{SE}(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

Test Two-Sample Means (Vars Known)

Rejection region

- $H_1 : \mu_X - \mu_Y > \delta_0$,
$$z > z_{1-\alpha}$$
- $H_1 : \mu_X - \mu_Y < \delta_0$,
$$z < z_{\alpha}$$
- $H_1 : \mu_X - \mu_Y \neq \delta_0$,
$$|z| > z_{1-\alpha/2}$$

Test Two-Sample Means (Vars Unknown But Equal)

Assume random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, where $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ unknown.

The null hypothesis for testing the difference between two means is

$$H_0 : \mu_X - \mu_Y = \delta_0$$

Test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\mathrm{SE}(\bar{X} - \bar{Y})} \overset{H_0}{\sim} t(n_X + n_Y - 2)$$

where SE is the standard error and S_p^2 is the pooled sample variance.

$$\mathrm{SE}(\bar{X} - \bar{Y}) = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Test Two-Sample Means (Vars Unknown But Equal)

Rejection region:

- $H_1 : \mu_X - \mu_Y > \delta_0$, $t > t_{1-\alpha}(n_X + n_Y - 2)$
- $H_1 : \mu_X - \mu_Y < \delta_0$, $t < t_{\alpha}(n_X + n_Y - 2)$
- $H_1 : \mu_X - \mu_Y \neq \delta_0$, $|t| > t_{1-\alpha/2}(n_X + n_Y - 2)$

In R: *t.test(x,y,var.equal=T)* or *t.test(z˜w,data,var.equal=T)* , where z is numeric and w is dichotomous.

Test Two-Sample Means (Vars Unknown)

Two independent sample data: $X_1, X_2, \cdots, X_{n_X} \sim N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \cdots, Y_{n_Y} \sim N(\mu_Y, \sigma_Y^2)$
Assume *unknown and unequal variance:* $\sigma_Y^2 \neq \sigma_Y^2$.
To test null hypothesis:

$$H_0 : \mu_X - \mu_Y = \delta_0$$

Welch t-ratio:

$$T = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\mathrm{SE}(\bar{X} - \bar{Y})} \overset{H_0}{\sim} t(df_w)$$

where df_w may not be an integer, and

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

Test Two-Sample Means (Vars Unknown)

Rejection region:

- $H_1 : \mu_X - \mu_Y > \delta_0$, $t > t_{1-\alpha}(df_w)$
- $H_1 : \mu_X - \mu_Y < \delta_0$, $t < t_{\alpha}(df_w)$
- $H_1 : \mu_X - \mu_Y \neq \delta_0$, $|t| > t_{1-\alpha/2}(df_w)$

In R: *t.test(x,y)* or *t.test(z˜w,data)*

Test Paired-Sample Means

Population mean difference $\mu_D = \mu_X - \mu_Y$ with paired data: $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$. Assume $D_i = X_i - Y_i \sim N(\mu_D, \sigma_D^2)(i = 1, 2, \ldots, n)$.
To test null hypothesis:

$$H_0 : \mu_X - \mu_Y = \delta_0$$

Test statistic

$$T = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}} \overset{H_0}{\sim} t(n-1)$$

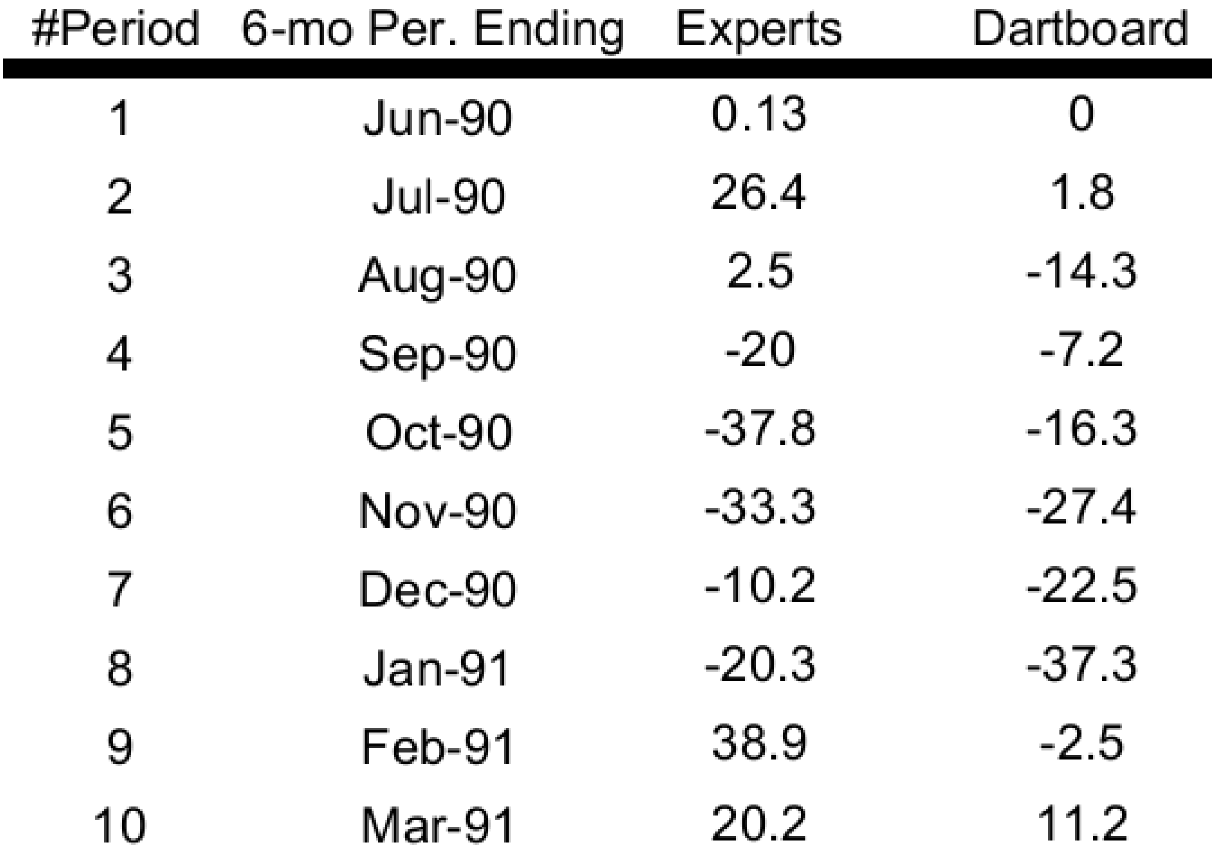
Rejection region:

- $H_1 : \mu_X - \mu_Y > \delta_0$, $t > t_{1-\alpha}(n-1)$
- $H_1 : \mu_X - \mu_Y < \delta_0$, $t < t_{\alpha}(n-1)$
- $H_1 : \mu_X - \mu_Y \neq \delta_0$, $|t| > t_{1-\alpha/2}(n-1)$

R function: *t.test(x, y, paired = TRUE)*

Example: WSJ comparing the performance of stocks

Wall Street Journal runs a contest comparing the performance of stocks. Half are chosen by a experts the others by throwing darts at the stock listings. The first 10 returns on the stocks picked are listed in the following table:



#Period	6-mo Per. Ending	Experts	Dartboard
1	Jun-90	0.13	0
2	Jul-90	26.4	1.8
3	Aug-90	2.5	-14.3
4	Sep-90	-20	-7.2
5	Oct-90	-37.8	-16.3
6	Nov-90	-33.3	-27.4
7	Dec-90	-10.2	-22.5
8	Jan-91	-20.3	-37.3
9	Feb-91	38.9	-2.5
10	Mar-91	20.2	11.2

Testing for No Difference

Matched pairs as each stock can be matched by the date in which its performance is tracked.

For the experts $X = 11.6\%$ and $s_X = 22.4\%$, while for the darts $Y = 6.5\%$ and $s_Y = 22.8\%$. Let $D_i = X_i - Y_i$, compute $\bar{D} = 5.14\%$ and$s_D = 26.48\%$, $n = 115$.

Null hypothesis to be tested $H_0 : \mu_X = \mu_Y$

- For $\alpha = 0.05$, the rejection region is $|t_{obs}| > t_{1-\alpha/2}(n-1) = 1.96$. Now

$$t_{obs} = \frac{\bar{d} - 0}{s_D/\sqrt{n}} = 2.082$$

So we should reject H_0.

Testing for No Difference: Comfidence Interval

A 95% confidence interval for the difference is

$$\bar{D} \pm t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}$$

For our data set, the CI is:*(0.30%, 9.98%)*.

The interval is entirely *positive*, we conclude that we are 95% confident that the experts are better stock pickers than the darts.

[fragile]Example: comparing Apple vs Buffet's Berkshire

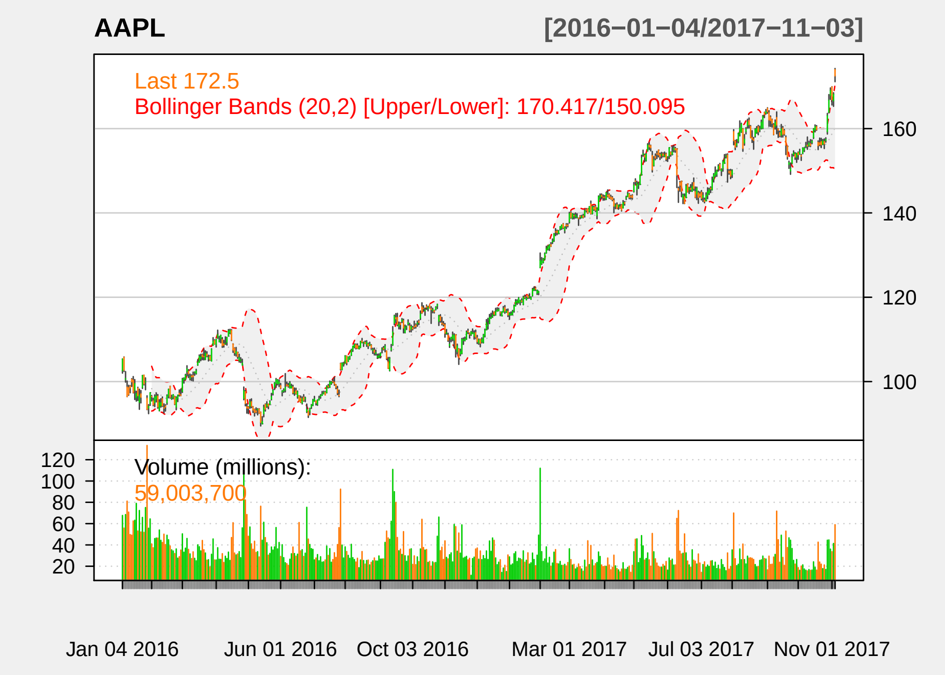
- Prepare package
  - library(”quantmod”) # if installed
- Import daily data (from Yahoo)
  - AAPL <- getSymbols(”AAPL”, from = ”2016-01-01”, to = ”2017-11-5”, auto.assign = FALSE)
  - BUFF <- getSymbols(”BRK-A”, from = ”2016-01-01”, to = ”2017-11-5”, auto.assign = FALSE)
- Plot the stock price
  - chartSeries(AAPL, theme = chartTheme('white'), TA = ”addVo(); addBBands()”)
- Calculate log-returns
  - aapl <- diff(log(AAPL$AAPL.Adjusted))
  - aapl <- aapl[-1]
  - buff <- diff(log(BUFF$'BRK-A.Adjusted'))
  - buff <- buff[-1]

Import the daily data

```
> library("quantmod")
> AAPL <- getSymbols("AAPL", from = "2016-01-01", to = Sys.Date(),
+                    auto.assign = FALSE)
> BUFF <- getSymbols("BRK-A", from = "2016-01-01", to = Sys.Date(),
+                    auto.assign = FALSE)
> head(AAPL)

##            AAPL.Open AAPL.High AAPL.Low AAPL.Close AAPL.Volume
## 2016-01-04    102.61    105.37   102.00     105.35    67649400
## 2016-01-05    105.75    105.85   102.41     102.71    55791000
## 2016-01-06    100.56    102.37    99.87     100.70    68457400
## 2016-01-07     98.68    100.13    96.43      96.45    81094400
## 2016-01-08     98.55     99.11    96.76      96.96    70798000
## 2016-01-11     98.97     99.06    97.34      98.53    49739400
##            AAPL.Adjusted
## 2016-01-04     101.42603
## 2016-01-05      98.88437
## 2016-01-06      96.94924
## 2016-01-07      92.85753
## 2016-01-08      93.34854
## 2016-01-11      94.86006
```

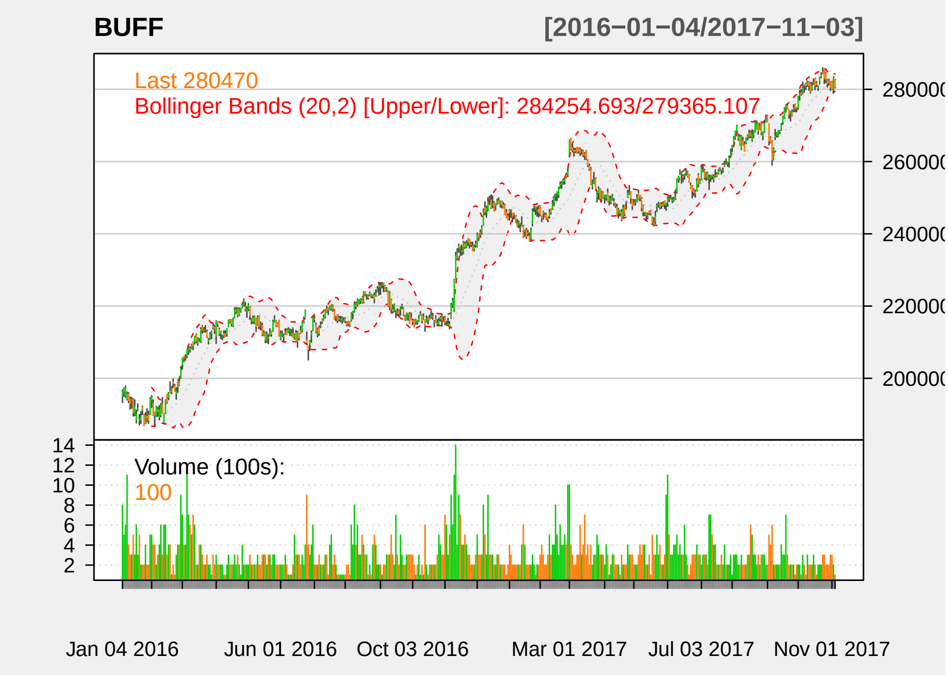
Plot the AAPL Time Series

```
> chartSeries(AAPL,theme=chartTheme('white'),TA="addVo();addBBands()")
```



Plot the BUFF Time Series

```
> chartSeries(BUFF,theme=chartTheme('white'),TA="addVo();addBBands()")
```

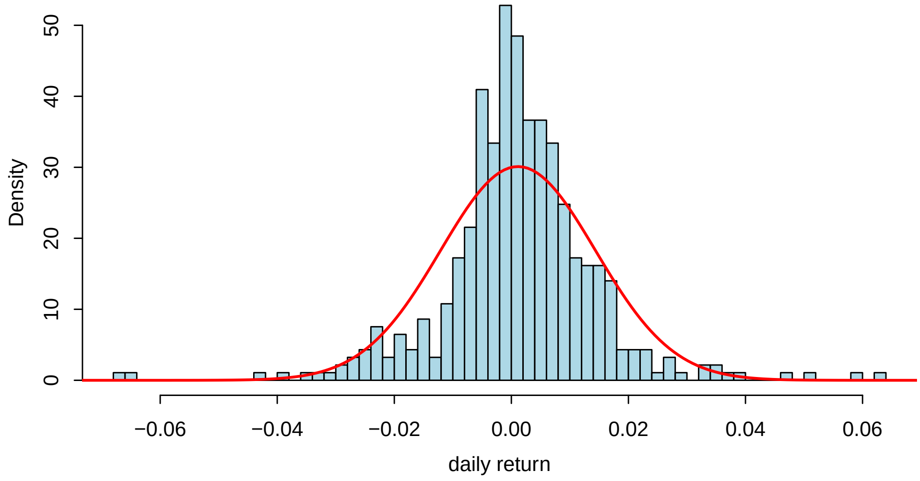


Calculate Daily Returns and moments

```
> aapl <- diff(log(AAPL$AAPL.Adjusted)) # log returns
> aapl <- aapl[-1] # delete the first NA
> buff <- diff(log(BUFF$'BRK-A.Adjusted'))
> buff <- buff[-1]
> aapl.ret <- as.vector(aapl) # Why? aapl is a xts
> buff.ret = as.vector(buff)
> # moments of returns
> library(moments)
> mean(aapl.ret); sd(aapl)

## [1] 0.001096614
## [1] 0.01305313

> skewness(aapl.ret); kurtosis(aapl.ret)

## [1] -0.1249777
## [1] 8.013832
```

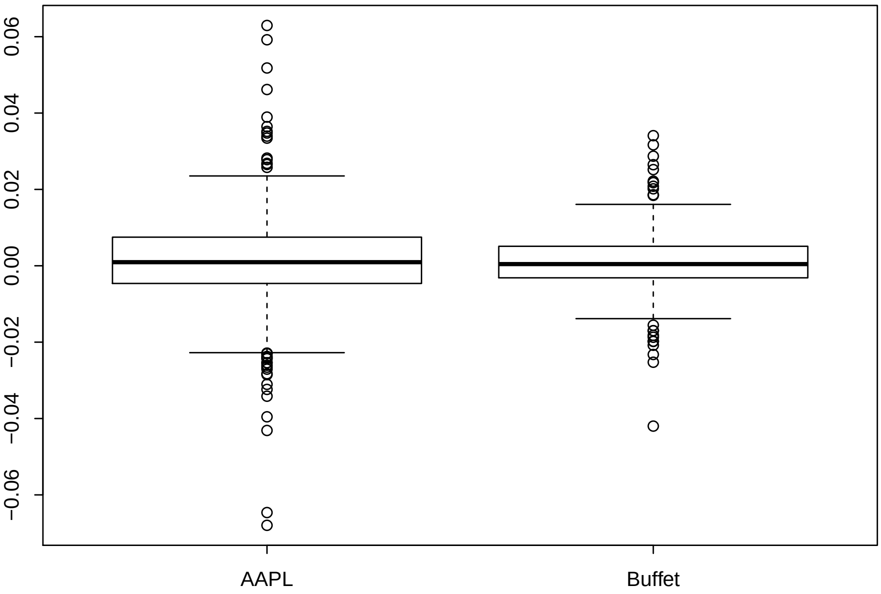
Distribution of Daily Returns of AAPL

```
> hist(aapl.ret,breaks=50,main = "", xlab="daily return",
+      col='lightblue', freq = FALSE)
> y<-seq(-0.08,0.08,length=200)
> lines(y,dnorm(y,mean(aapl.ret),sd(aapl.ret)),type='l',col='red',lwd=2)
```



Returns: AAPL vs Buffet

```
> boxplot(aapl.ret, buff.ret, names=c("AAPL","Buffet"))
```



t-test of mean difference: Apple vs Buffet

```
>  t.test(aapl.ret, buff.ret, var.equal = F)

##
##  Welch Two Sample t-test
##
## data:  aapl.ret and buff.ret
## t = 0.37839, df = 817.14, p-value = 0.7052
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001089040  0.001609192
## sample estimates:
##    mean of x    mean of y
## 0.0010966144 0.0008365385
```

What conclusions do you draw from this analysis?

3.2 Testing of Hypotheses for Proportion Differences

Two Sample Test for Proportions

Consider two populations, $X \sim \text{Bin}(n_X, \pi_X)$ and $Y \sim \text{Bin}(n_Y, \pi_Y)$. If P_X and P_Y are the respective sample proportions of successes.

The null hypothesis to be tested is

$$H_0 : \pi_X - \pi_Y = \delta_0$$

Then the test statistic

$$Z = \frac{(P_X - P_Y) - \delta_0}{\sqrt{\frac{P_X(1-P_X)}{n_X} + \frac{P_Y(1-P_Y)}{n_Y}}} \overset{H_0}{\sim} N(0,1)$$

A 95% confidence interval for a difference in proportions

$$(p_X - p_Y) \pm z_{1-\alpha/2}\sqrt{\frac{p_X(1-p_Y)}{n_X} + \frac{p_X(1-p_Y)}{n_Y}}$$

Two Sample Test for Proportions

Rejection region:

- $H_1 : \mu_X - \mu_Y > \delta_0$, $z > z_{1-\alpha}$
- $H_1 : \mu_X - \mu_Y < \delta_0$, $z < z_{\alpha}$
- $H_1 : \mu_X - \mu_Y \neq \delta_0$, $|z| > z_{1-\alpha/2}$

Two Sample Test for Equality of Proportions

Especially, if $\delta_0 = 0$, the null hypothesis is

$$H_0 : \pi_X = \pi_Y$$

It is standard practice to create a *pooled* estimate of the population proportions estimate of $\pi = \pi_X = \pi_Y$,

$$P = \frac{X+Y}{n_X + n_Y}$$

The test statistic is

$$Z = \frac{P_X - P_Y}{\sqrt{P(1-P)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \overset{H_0}{\sim} N(0,1)$$

Two sample t-test for populations: *prop.test()*

Mammogram check for breast cancer

A mammogram is an X-ray procedure used to check for breast cancer. Whether mammograms should be used is part of a controversial discussion. A 30-year study was conducted with nearly 90,000 female participants[1]. During a 5-year screening period, each woman was randomized to one of two groups:

1. in the first group, women received regular mammograms to screen for breast cancer,
2. and in the second group, women received regular non-mammogram breast cancer exams.

Mammogram check for breast cancer

No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period.

Question: is this study an experiment or an observational study? Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.

Hypothesis

$$H_0 : \pi_M = \pi_C \text{ vs } H_1 : \pi_M \neq \pi_C$$

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

[1] Miller AB. 2014. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ 2014;348:g366

Using prop.test() to test

```
> prop.test(x=c(500,505),n=c(44425,44405),alternative="two.sided")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(500, 505) out of c(44425, 44405)
## X-squared = 0.017976, df = 1, p-value = 0.8933
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.001531197  0.001295859
## sample estimates:
##     prop 1     prop 2
## 0.01125492 0.01137259
```

What conclusions do you draw from this analysis?

3.3 Testing of Hypotheses for Equality of Variances

Test for Equality of Variances

Assume random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are unknown.

The null hypothesis is

$$H_0 : \sigma_X^2 = \sigma_Y^2$$

Test statistic

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \overset{H_0}{\sim} F(n_X - 1, n_Y - 1)$$

Rejection region:

- $H_1 : \sigma_X^2 > \sigma_Y^2$, $F_{obs} > F_{1-\alpha}(n_X - 1, n_Y - 1)$
- $H_1 : \sigma_X^2 < \sigma_Y^2$, $F_{obs} < F_{\alpha}(n_X - 1, n_Y - 1)$
- $H_1 : \sigma_X^2 \neq \sigma_Y^2$, $F_{obs} < F_{\alpha/2}(n_X - 1, n_Y - 1)$ or $F_{obs} > F_{1-\alpha/2}(n_X - 1, n_Y - 1$

In R: *var.test()*

Testing the Equality of Risks: AAPL vs Buffet

```
> var.test(aapl.ret, buff.ret)

##
##  F test to compare two variances
##
## data:  aapl.ret and buff.ret
## F = 2.6851, num df = 494, denom df = 494, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.250450 3.203582
## sample estimates:
## ratio of variances
##           2.685052
```

What conclusions do you draw from this analysis?

Example: satisfaction of the school graduates

A questionnaire is devised by the Board of Governors to measure the level of satisfaction for graduates from two competing state schools.

The questionnaire is randomly administered to 11 students from State School X and 15 students from State School Y (in dataset STCHOOL).

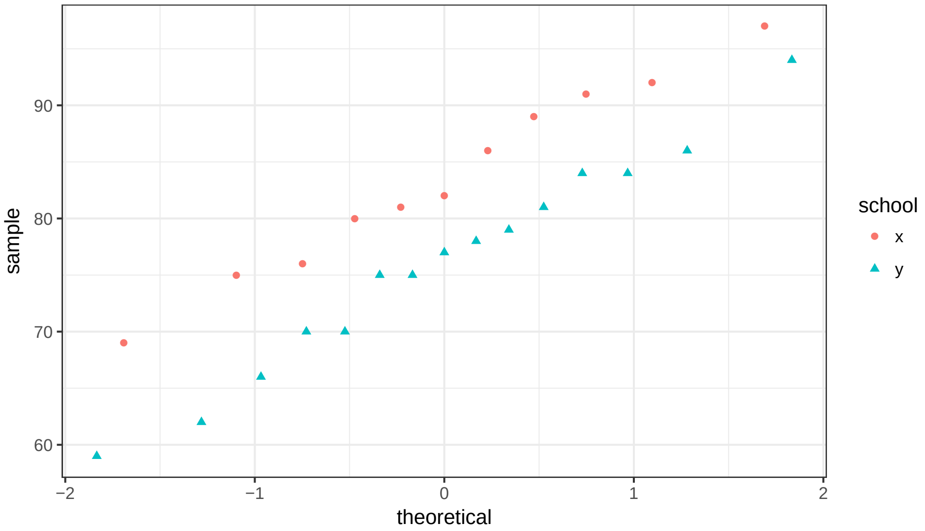
School X: 69 75 76 80 81 82 86 89 91 92 97

School Y: 59 62 66 70 70 75 75 77 78 79 81 84 84 86 94

Test to see if there are significant differences between the mean satisfaction levels for graduates of the two competing state schools using a significance level of 5%.

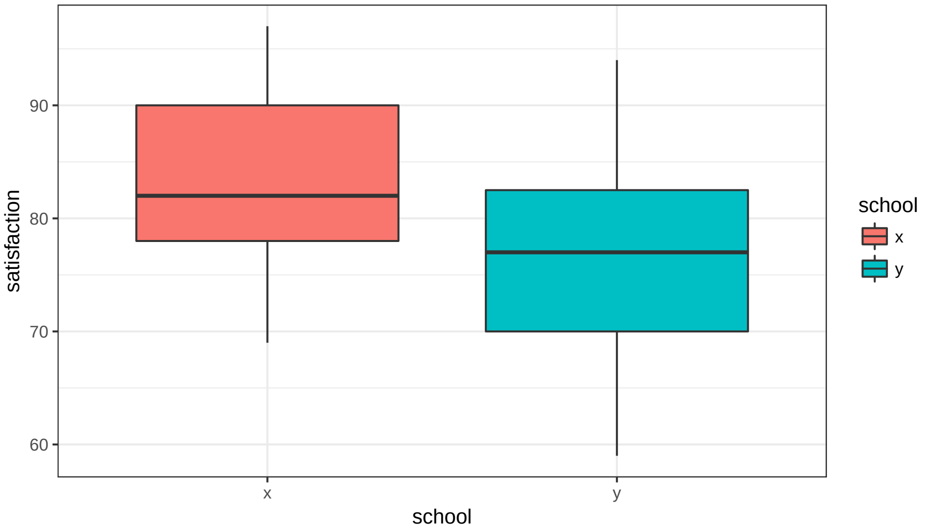
Check normality of data

```
> ggplot(data = STSCHOOL, aes(sample = satisfaction, shape = school,
+                color = school)) +    theme_bw() + stat_qq()
```



Descriptive comparison: boxlots

```
> ggplot(data=STSCHOOL,aes(x=school,y=satisfaction,fill=school)) +
+ geom_boxplot() + theme_bw()
```



Are the two variances equal?

```
> var.test(STSCHOOL$x, STSCHOOL$y)

##
##  F test to compare two variances
##
## data:  STSCHOOL$x and STSCHOOL$y
## F = 0.79153, num df = 10, denom df = 14, p-value = 0.7225
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2515314 2.8102720
## sample estimates:
## ratio of variances
##          0.7915345
```

What conclusions do you draw?

Are the two schools have significant differences in satisfaction?

```
>  t.test(STSCHOOL$x, STSCHOOL$y, var.equal = T)

##
##  Two Sample t-test
##
## data:  STSCHOOL$x and STSCHOOL$y
## t = 2.0798, df = 24, p-value = 0.0484
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.05691592 14.85217499
## sample estimates:
## mean of x mean of y
##  83.45455  76.00000
```

What conclusions do you draw from this analysis?