MA5701: Statistical Methods

Chapter 5 : Inferences for Two Populations

Kui Zhang, Mathematical Sciences

Inference for Two Parameters

- So far we have performed the statistical inference for one parameter.
- The case of two populations (two parameters) is important too.
  - Many interesting applications involve only two populations, for example, any comparisons involving differences between the two sexes, comparing a drug with a placebo, comparing old versus new, or before and after some events.
  - Some of the concepts underlying comparing several populations are more easily introduced for the two-population case.
  - The comparison of two populations results in a single easily understood statistic: the difference between two sample means.

Methods for Collecting Data

- **Independent Samples, for example,**
  - A sample of migraine sufferers is randomly divided into two groups. The first group is given remedy A while the other is given remedy B, both to be taken at the onset of a migraine attack. The pills are not identified, so patients do not know which pill they are taking.
- **Dependent or Paired Samples, for example,**
  - Each person in a group of migraine sufferers is given two pills, one of which is red and the other is green. The group is randomly split into two subgroups and one is told to take the green pill the first time a migraine attack occurs and the red pill for the next one. The other group is told to take the red pill first and the green pill next.

Introduction

- In this chapter, we will present procedures for:
  - Making inferences on the difference of means of two normally distributed populations where the variances are unknown.
  - Making inferences on the difference of proportions of successes in two binomial populations.

Inferences on the Difference of Means

- We are interested in comparing two populations whose means are $\mu_1$and μ_2 and variances are σ_1^2 and σ_2^2.
- To compare means, we have H_0: $\mu_1 - \mu_2 = \delta_0$ versus H_a: $\mu_1 - \mu_2 \neq \delta_0$.
- In many situations, we set $\delta_0 = 0$.

Inference For Two Independent Samples

- **Example (Packaging of Ground Beef)** – Maxcy and Lowry (1984) looked a packaging process for ground beef over a series of days. An interesting question is whether the true mean amount delivered by this process changes from day to day.
- Denote the population mean and variance of beef packed on one day are μ_1 and σ_1^2 and on another day are μ_2 and σ_2^2, respectively.
- We are interested in the difference of means, $\mu_1 - \mu_2$.

Inference For Two Independent Samples

- If $\mu_1 - \mu_2 > 0$ then the true mean amount of beef packed on one day is larger than the true mean amount of beef packed on the other day.
- If $\mu_1 - \mu_2 = 0$ then the true mean amount of beef packed on one day is no different from the true mean amount of beef packed on the other day.
- If $\mu_1 - \mu_2 < 0$, then the true mean amount of beef packed on one day is less than the true mean amount of beef packed on the other day.

Distribution of A Linear Function of Random Variables

For a set of n independent random variables $Y_1, \cdots, Y_n$, whose means are $\mu_1, \cdots, \mu_n$ and whose variances are $\sigma_1^2, \cdots, \sigma_n^2$. A linear function of these random variables is defined as $L = \sum_{i=1}^{n}(a_i Y_i + b_i)$, where the a_i and b_i are arbitrary constants, then

1. $E(L) = \sum_{i=1}^{n}(a_i \mu_i + b_i)$
2. $Var(L) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.
3. If Y_i are normally distributed, then L is normally distributed.

Sampling Distribution of Difference of Means

- Since sample means are random variables, the difference between two sample means is a linear function of two random variables.
- First, $E(\bar{Y}_1) = \mu_1, Var(\bar{Y}_1) = {}^{\sigma_1^2}/_{n_1}, E(\bar{Y}_2) = \mu_2, Var(\bar{Y}_2) = {}^{\sigma_2^2}/_{n_2}$
- Then $L = \bar{Y}_1 - \bar{Y}_2$, we can get

$$E(L) = \mu_1 - \mu_2; Var(L) = {}^{\sigma_1^2}/_{n_1} + {}^{\sigma_2^2}/_{n_2}$$

- When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $Var(L) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$
- Finally, the central limit theorem states that if the sample sizes are sufficiently large, the sample means are approximately normally distributed; hence generally $\bar{Y}_1 - \bar{Y}_2$ is also normally distributed too.

Inferences with Known Variance

- If the variances are known, we can use the following random variable to make the inference on $\delta = \mu_1 - \mu_2 : (\bar{Y}_1 - \bar{Y}_2)$ and we know that $Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ has a standard normal distribution.
- In this situation, you can consider $(\bar{Y}_1 - \bar{Y}_2)$ as a random variable with $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$, then all formulas introduced in chapter 3 can be used here.
- Again, this method has little practical use since we generally do not know the populations variances.

Variances Unknown but Assumed Equal

- **Solution** to unknown variance - Assume that the two population variances are equal and find an estimate of that variance. The equal variance assumption is actually quite reasonable since in many studies, a focus on means implies that the populations are similar in many respects. However, if the assumption of equal variances cannot be made, then other methods must be used.
- Again, we will use the point estimate of that difference $(\bar{y}_1 - \bar{y}_2)$.

Two Sample *t*-test – Equal Variance

- Two sample *t*-test is used for the inference of a difference of two means with unknown population variance.
- Assumptions for the two sample *t*-test (equal variance) are:
  - Each random sample is selected from the target population.
  - Two samples are independent.
  - The sample mean, $\bar{Y}_1$ and $\bar{Y}_2$ are normal or approximate normal.
  - The variances of two populations are same: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
  - Note that two sample sizes may not be equal.

Two Sample *t*-test – Equal Variance

- The null hypothesis is: $H_0: \mu_1 - \mu_2 = \delta_0$
- δ_0 is the nominal difference two means. In many situations, $\delta_0 = 0$
- The test statistic is:

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{Estiamted\ Standard\ Deviation\ of\ (\bar{y}_1 - \bar{y}_2)}$$

- How to estimate the standard deviation of $(\bar{Y}_1 - \bar{Y}_2)$?
- Note that the variance of $\bar{Y}_1 - \bar{Y}_2$ is $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$ here.
- What is the degrees of freedom of the *t*-distribution here?

Pooled Estimate of σ^2

- n_1: sample size from the first sample;
- n_2: sample size from the second sample;
- s_1^2: sample variance from the first sample;
- s_2^2: sample variance from the second sample;
- The pooled estimate of σ^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$$

Two Sample *t*-test – Equal Variance

- The null hypothesis is: $H_0: \mu_1 - \mu_2 = \delta_0$
- The test statistic is:

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ when } H_0 \text{ is true}$$

- Where $s_p = \sqrt{s_p^2}$ and $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{SS_1+SS_2}{n_1+n_2-2}$.
- The degrees of freedom is $n_1 + n_2 - 2$.

Two Sample *t*-test – Rejection Region

- For two-sided test ($H_a: \mu_1 - \mu_2 \neq \delta_0$), the rejection region is:

$$|t| > t_{n_1+n_2-2,\alpha/2}$$

- For upper-tailed (right-tailed) test ($H_a: \mu_1 - \mu_2 > \delta_0$), the rejection region is:

$$t > t_{n_1+n_2-2,\alpha}$$

- For lower-tailed (left-tailed) test ($H_a: \mu_1 - \mu_2 < \delta_0$), the rejection region is:

$$t < -t_{n_1+n_2-2,\alpha}$$

Two Sample *t*-test – *p*-value Approach

- For two-sided test ($H_a: \mu_1 - \mu_2 \neq \delta_0$):

$$p\text{-value} = 2\Pr(T_{n_1+n_2-2} > |t|)$$

- For upper-tailed (right-tailed) test ($H_a: \mu_1 - \mu_2 > \delta_0$):

$$p\text{-value} = \Pr\left(T_{n_1+n_2-2} > t\right)$$

- For lower-tailed (left-tailed) test ($H_a: \mu_1 - \mu_2 < \delta_0$):

$$p\text{-value} = \Pr(T_{n_1+n_2-2} < t)$$

- Here $T_{n_1+n_2-2}$ represents a random variable with t-distribution of $n_1 + n_2 - 2$ degrees of freedom.

Two Sample *t*-test – Confidence Interval

- For two-sided test ($H_a: \mu_1 - \mu_2 \neq \delta_0$), construct a $100(1-\alpha)\%$ two-sided confidence interval

$$\left(\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{{}^1/_{n_1} + {}^1/_{n_2}}\right).$$

- For upper-tailed test ($H_a: \mu_1 - \mu_2 > \delta_0$), construct a $100(1-\alpha)\%$ lower confidence interval $\left(\bar{y}_1 - \bar{y}_2 - t_{n_1+n_2-2,\alpha} s_p \sqrt{{}^1/_{n_1} + {}^1/_{n_2}}, \infty\right)$.

- For lower-tailed test ($H_a: \mu_1 - \mu_2 < \delta_0$), construct a $100(1-\alpha)\%$ upper confidence interval $\left(-\infty, \bar{y}_1 - \bar{y}_2 + t_{n_1+n_2-2,\alpha} s_p \sqrt{{}^1/_{n_1} + {}^1/_{n_2}}\right)$.

Two Sample *t*-test – Packing of Ground Beef

- **Example (Packing of Ground Beef)** – The data are shown in Tables.

First Day	1397.8	1394.8	1391.7	1400.0	1393.5
First Day	1391.2	1384.0	1391.0	1385.7	1385.3
Second Day	1410.0	1393.9	1405.9	1404.2	1387.3
Second Day	1398.5	1399.9	1392.5	1402.5	1391.8

- **Question**: Is the mean amount of beef delivered on the first day different from the mean amount of beef delivered on the second day? Use 0.05 significance level to perform your test.

Two sample *t*-test – Packing of Ground Beef

- **Step 1:** Specify appropriate H_0, H_a, and an acceptable level of α.

$$H_0: \mu_1 - \mu_2 = 0; H_a: \mu_1 - \mu_2 \neq 0; \alpha = 0.05$$

- **Step 2:** Define a sample-based test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p\sqrt{1/n_1 + 1/n_2}}$$

- **Step 3:** Find the rejection region for the specified H_a and α.

The rejection region is $|t| > t_{n_1+n_2-2,\alpha/2} = t_{18,0.025} = 2.1009$.

Two Sample *t*-test – Packing of Ground Beef

- **Step 4:** Collect the sample data and calculate the test statistic.

$$\bar{y}_1 = 1391.50; s_1^2 = 28.3933;\ \bar{y}_2 = 1398.65; s_2^2 = 51.6361$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 * 28.3933 + 9 * 51.6361}{10 + 10 - 2}$$

$$= 40.0147$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{1391.50 - 1398.65}{\sqrt{40.0147}\sqrt{1/10 + 1/10}} = -2.5274$$

Two Sample *t*-test – Packing of Ground Beef

- **Step 5:** Make a decision to either reject or fail to reject H_0.
  - We have $|t| = |-2.5274| > 2.1009$. At the 5% significance level, the test statistic $t = -2.5274$ falls in the rejection region. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

Packing of Ground Beef – Confidence Interval

- **Step 4:** The 95% confidence interval of mean amount of beef is

$$\left(\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{1/n_1 + 1/n_2}\right)$$

$$= \left(1391.50 - 1398.65 \pm 2.1009 * \sqrt{40.1047} \sqrt{1/10 + 1/10}\right)$$

$$= (-13.1000, -1.2000)$$

Packing of Ground Beef – Confidence Interval

- **Step 5:** Make a decision to either reject or fail to reject H_0.
  - The nominal difference of mean amount of beef 0 does not fall in the 95% confidence interval $(-13.1000, -1.2000)$. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

Packing of Ground Beef – *p*-value Approach

- **Step 4:** we have $t = -2.5274$, so the *p*-value is

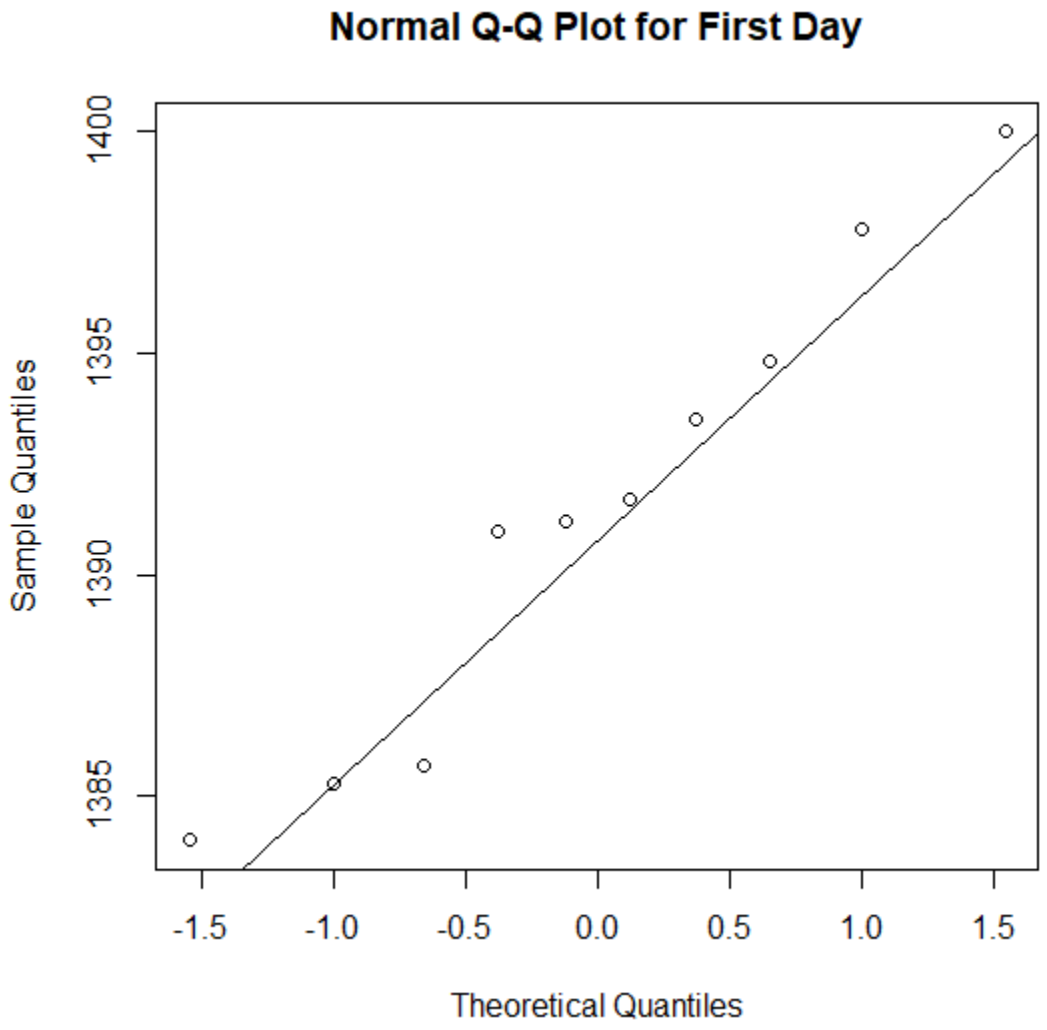
$$p\text{-value} = 2\Pr(T_{18} > |-2.5274|) = 0.0211$$

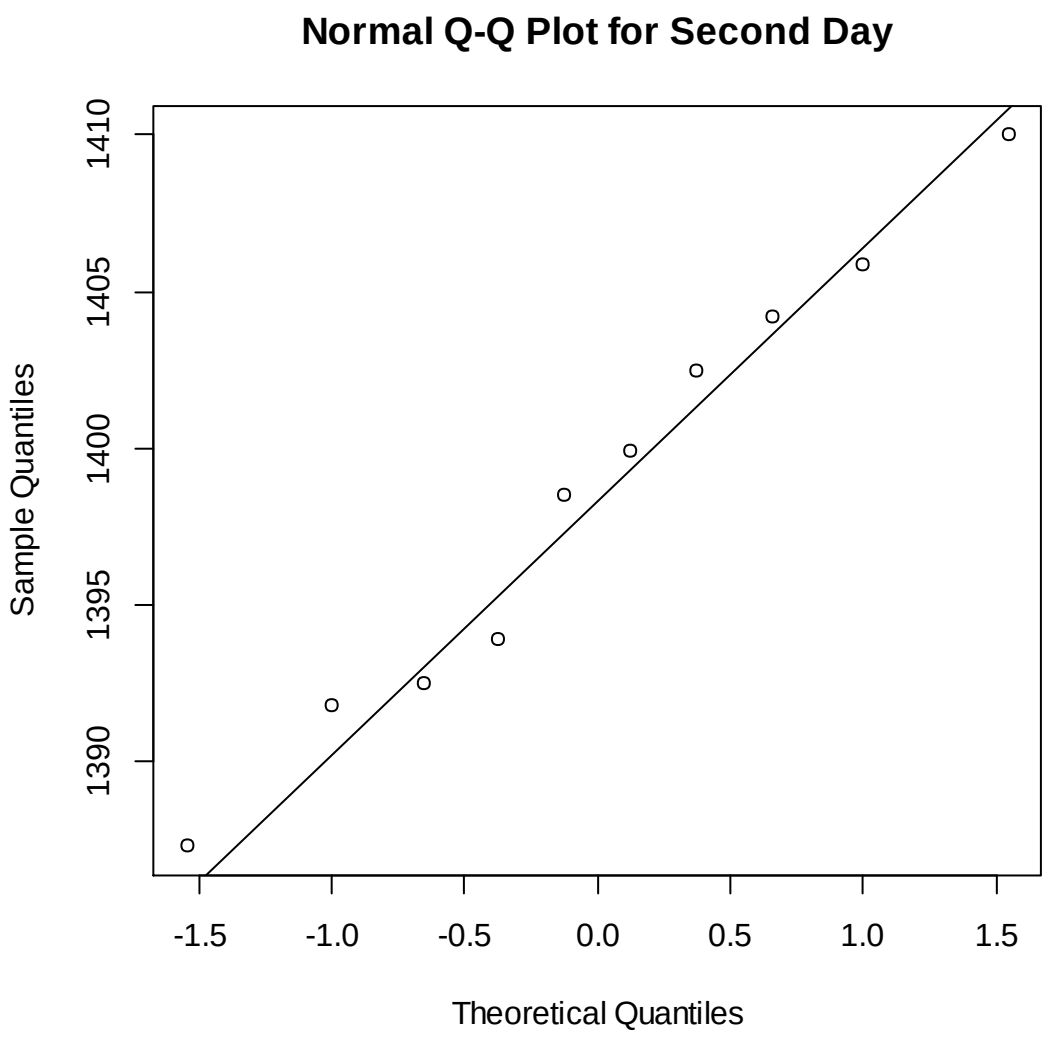
- **Step 5:** Make a decision to either reject or fail to reject H_0.
  - The *p*-value is 0.0211 and is less than the significance level of 0.05. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

Packing of Ground Beef – Check Assumptions

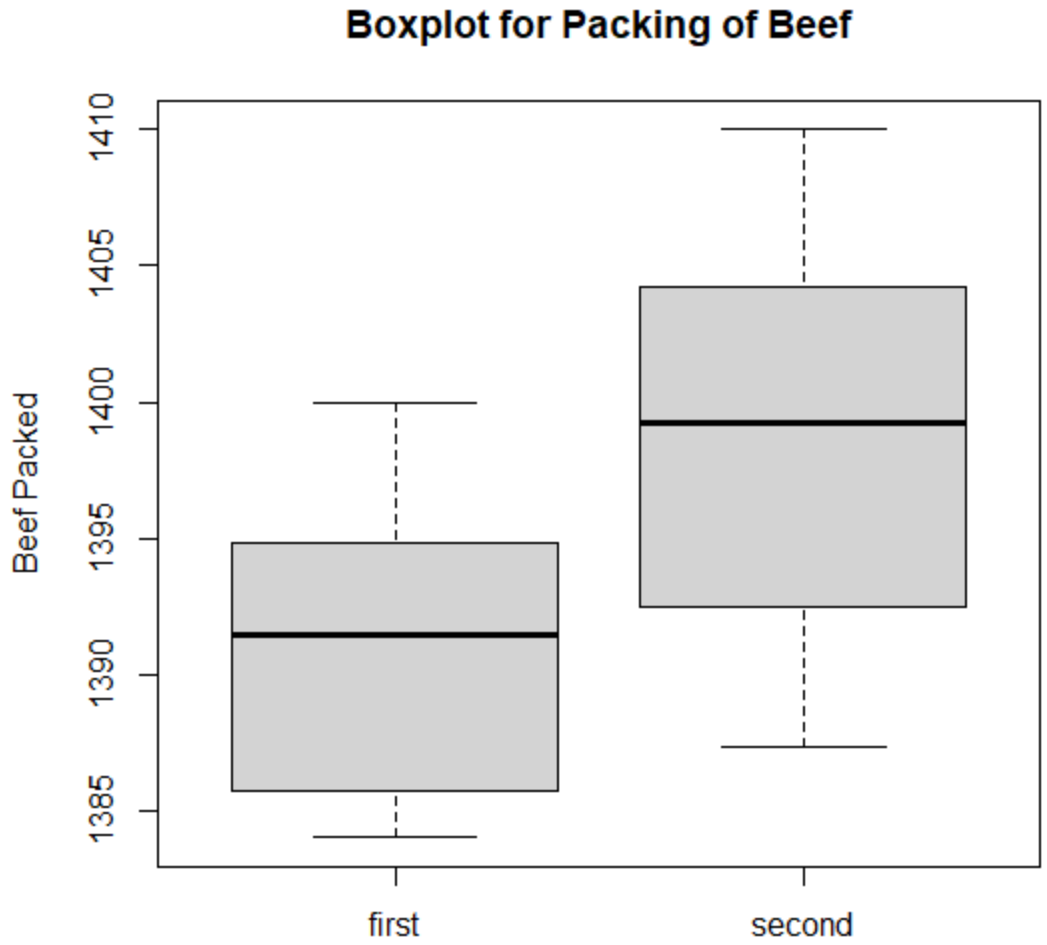
- When the sample size is small, you should check the normality assumption with the stem-and-leaf display and/or Q-Q plot.
- Use a box plot to check the equal variance assumption. There are formal tests for the equal variance assumption. However, they will not be covered in our course.
- From the figures from the next two slides, it seems the data are normally distributed and two samples have the similar variance.

Packing of Ground Beef – Q-Q Plot





Packing of Ground Beef – Boxplot



R Function – t.test

```
t.test( x = beef$first,
        y = beef$second,
        alternative = "two.sided", # two sided is the default
        mu = 0,
        paired = FALSE, # default is FALSE
        var.equal = TRUE,  # default is FALSE
        conf.level = 1 - alpha)
```

R Function – t.test

Two Sample t-test

data: beef$first and beef$second
t = -2.5274, df = 18, p-value = 0.02107
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.093398 -1.206602
sample estimates:
mean of x mean of y
1391.50 1398.65

Exercise – Diet Formulation

- **Exercise (Die Example).** To assess the effectiveness of a new diet formulation, a sample of 8 steers is fed a regular diet and another sample of 10 steers is fed a new diet. The weights of the steers at 1 year are given. Do these results imply that the new diet results in higher weights? (Use $\alpha = 0.05$)
- **Weights from Regular Diet:** 831, 858, 833, 860, 922, 875, 797, 788
- **Weights from New Diet:** 870, 882, 896, 925, 842, 908, 944, 927, 965, 887

Exercise – Diet Formulation

Exercise (Die Example). Some Basic Statistics.

- Regular Diet:

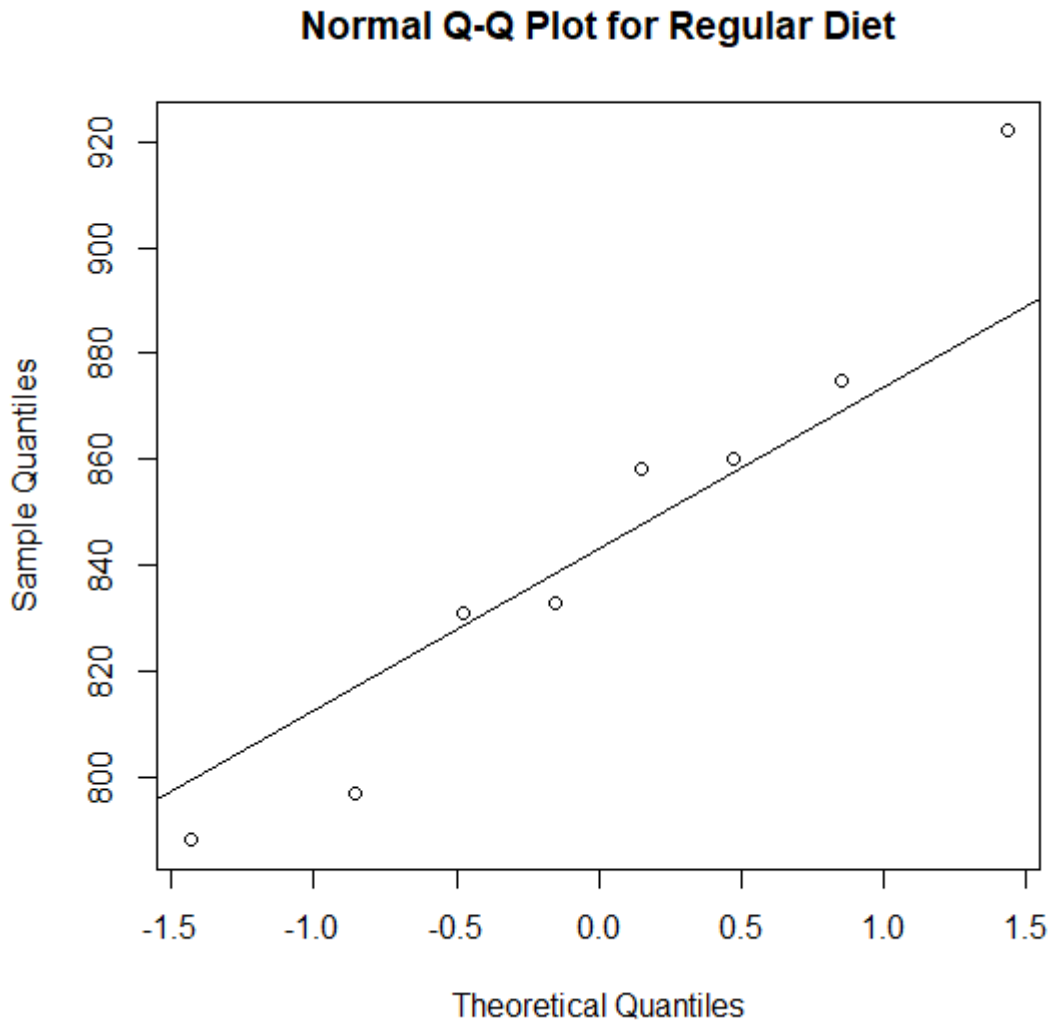
$$n_1 = 8; \bar{y}_1 = 845.5; s_1^2 = 1873.4286$$

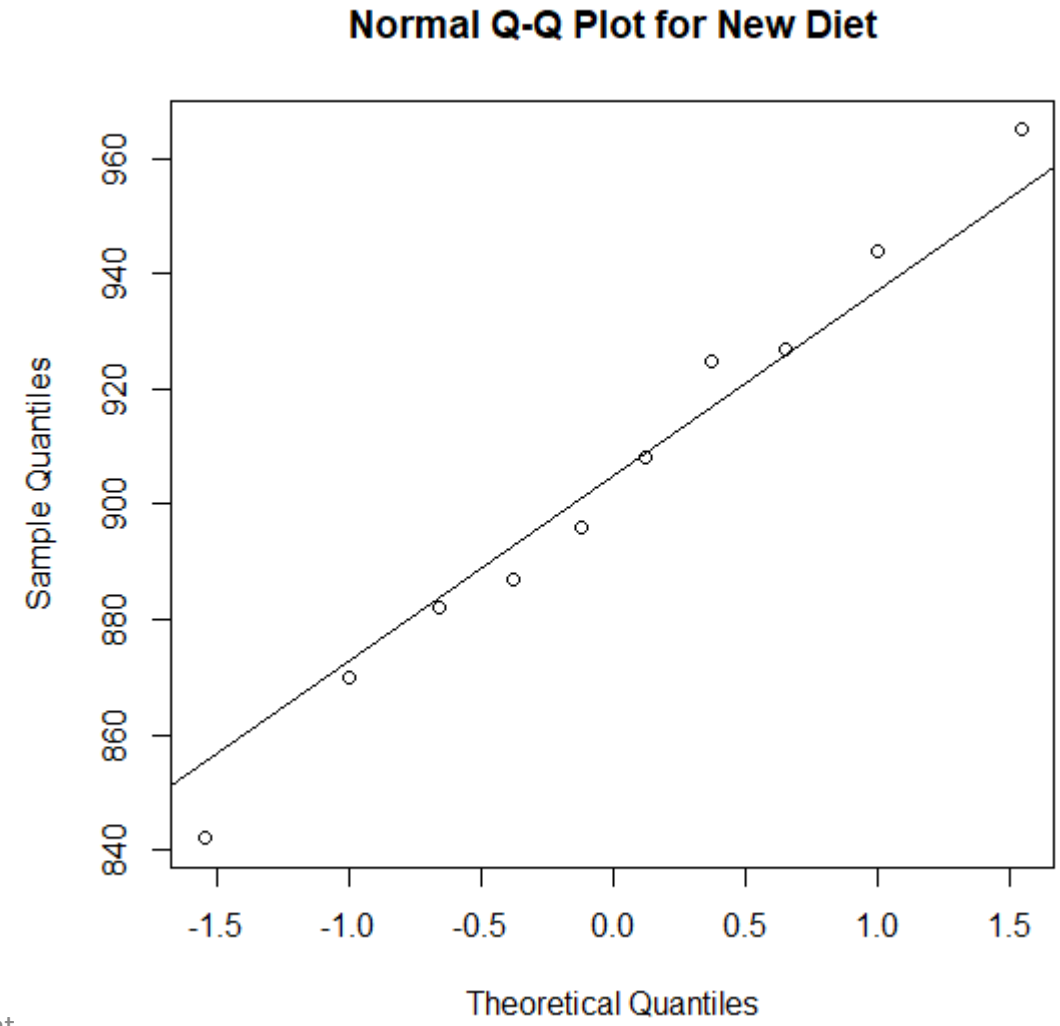
- New Diet:

$$n_2 = 10; \bar{y}_2 = 904.6; s_2^2 = 1348.9333$$

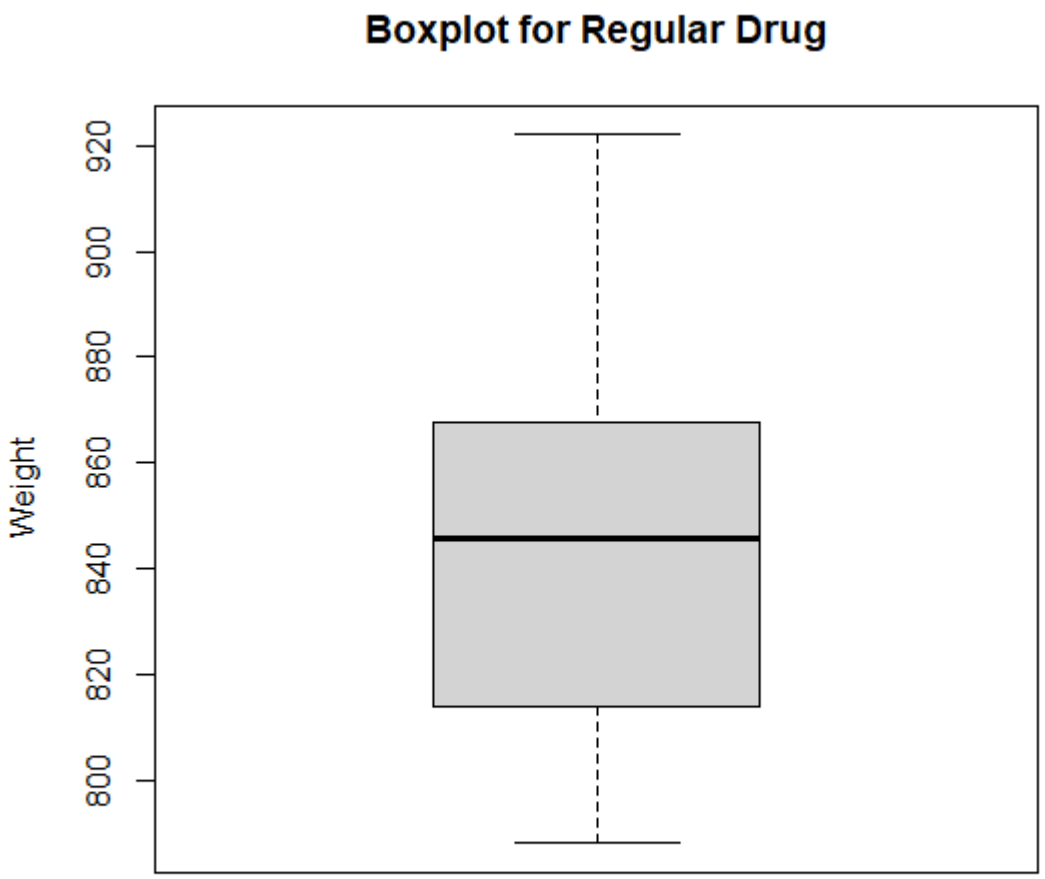
- Now use five steps for hypothesis testing.
- Use the significance level of 0.05.

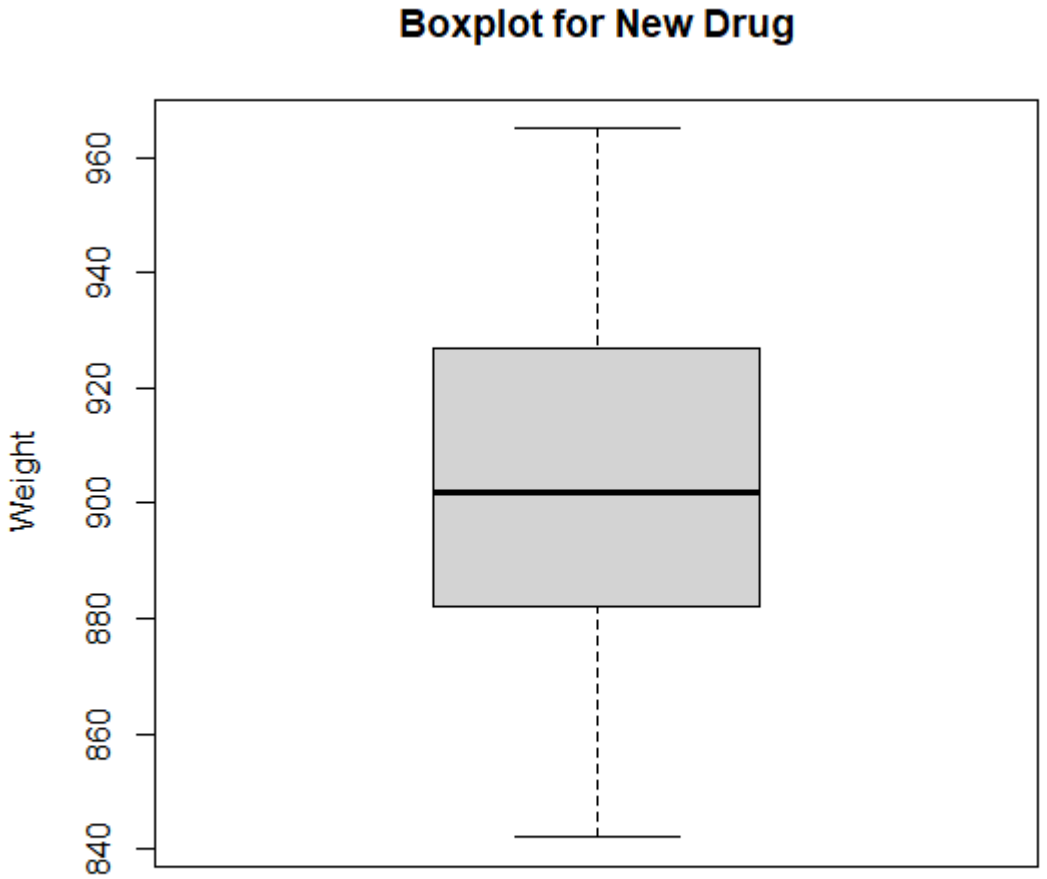
Exercise – Diet Formulation





Exercise – Diet Formulation





Two Sample *t*-test – Unequal Variance

- In **Packing of Ground Beef Example,** we saw that the sample variance of the weights of second day is almost twice that of first day. Therefore we may need to provide a method for comparing means that does not assume equal variances. (A test for equality of variances is presented in Section 5.3 and according to this test these two variances are not significantly different.)
- A test statistic similar to previous ones (next slide) can be use.
- Transformation to make variance equal (e.g., log transformation).

Two Sample *t*-test - Unequal Variance

- We can use the following statistic: $t' = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
- If both sample size are large (both over 30) we can assume a normal distribution and compute the test statistic.

Two Sample *t*-test - Unequal Variance

- We can use the following statistic: $t' = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
- If either sample size is not large but the data come from approximately normally distributed populations, this statistic does have an approximate Student's *t* distribution, but the degrees of freedom cannot be precisely determined.
- A reasonable (and conservative) approximation is to use the degrees of freedom for the smaller sample.
- More precise but complex approximations are available. One such approximation, called Satterthwaite's approximation, is implemented in many statistical packages.

Paired *t*-Test

- **Paired Samples** - pairs of observed values.
- Why paired samples? Using the weight loss of a special diet as an example:
  - Divide samples to two groups – one group with the general diet and the other group with the special diet and then compare the weights of samples. The drawback is that the estimate of the variance is based on the differences in weights among individuals in each sample, and these differences are probably larger than those induced by the special diet. Thus a huge sample size may be needed.
  - Give all samples the special diet but compare their weights before and after the special diet.
- Methods for paired samples – use differences between paired values.

Paired Versus Independent *t*-test

- Paired *t*-test can remove the sampling unit to sampling unit variability, increase the test statistic by decreasing the estimated variance.
- Paired *t*-test reduces the degrees of freedom thus results in the larger critical value, especially for small sample size.
- We should obtain paired data whenever we know the sampling unit to sampling unit variability is large.
- We can use two independent samples when the sampling unit to sampling unit variability is not an issue and the available sample size is small.

Paired t-test

- The paired t-test is essentially the one sample t-test that is applied to the difference of paired data.
- Let $d_i = y_{1,i} - y_{2,i}$ $(i = 1, \cdots, n)$ be the difference of the paired data.
- We perform the one sample t-test to d_i $(i = 1, \cdots, n)$.

Paired *t*-test - Notations

We have the following notations:

- n: number of *pairs* of data
- μ_d: population mean of difference.
- $\bar{d}$: sample mean of difference d_i $(i = 1, \cdots, n)$.
- s_d: sample standard deviation of difference d_i $(i = 1, \cdots, n)$.

Paired *t*-test

- The null hypothesis is: $H_0: \mu_d = \delta_0$
- δ_0 is the nominal difference two means. In many situations, $\delta_0 = 0$.
- The test statistic is:

$$T = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}} \sim t_{n-1} \text{ when } H_0 \text{ is true.}$$

- You can make a decision based on the rejection region, the confidence interval, or the p-value. The procedure is exactly same as the one sample *t*-test.

Paired *t*-Test – Example 5.6

- **Example 5.6 (Baseball Teams)** - For the first 60 years major league baseball consisted of 16 teams, eight each in the National and the American leagues. In 1961 the Los Angeles and the Washington Senators became the first expansion teams in baseball history. It is conjectured that the main reason that the league allowed expansion teams was the fact that total attendance dropped from 20 million in 1960 to slightly over 17 million in 1961.

Data from Example 5.6

Table 5.7 Baseball Attendance (Thousands)

Team	1960	1961	Diff.
1	809	673	−136
2	663	1123	460
3	2253	1813	−440
4	1497	1100	−397
5	862	584	−278
6	1705	1199	−506
7	1096	855	−241
8	1795	1391	−404
9	1187	951	−236
10	1129	850	−279
11	1644	1151	−493
12	950	735	−215
13	1167	1606	439
14	774	683	−91
15	1627	1747	120
16	743	597	−146

Paired *t*-Test - Example 5.6

Example 5.6 (Baseball Teams)

- Sample Size: $n = 16$
- 1960 Samples: $\bar{y}_1 = 1243.8125; s_1^2 = 212016.9625$
- 1961 Samples: $\bar{y}_2 = 1066.1250; s_2^2 = 161046.6500$
- Difference: $\bar{d} = -177.6875; s_d^2 = 86019.0292$
- Note that: $\bar{d} = \bar{y}_2 - \bar{y}_1$ but $s_d^2 \neq s_2^2 - s_1^2$

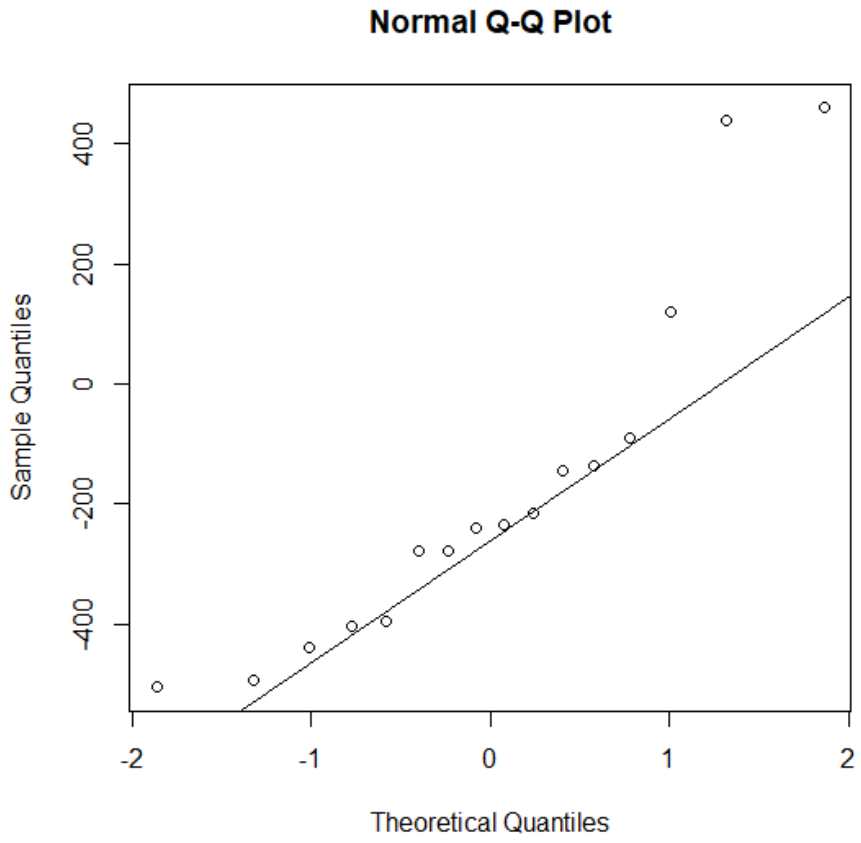
Paired t-Test - Example 5.6

Example 5.6 (Baseball Teams) – Key Results (Again, some steps are skipped here)

- Test statistic: $t = -177.6875/\sqrt{86019.0292/16} = -2.4233$
- Rejection region: $t < -t_{15,0.05} = -1.7531$
- The p-value is: $\Pr(T_{15} < -2.4233) = 0.0142$
- The upper 95% confidence interval is:

$$\left(-\infty, -177.6875 + t_{15,0.05} * \sqrt{\frac{86019.0292}{16}}\right) = (-\infty, -49.1495)$$

Paired *t*-Test - Example 5.6



Paired *t*-test – Baseball Teams

```
t.test(x = baseball$F3, y = baseball$F2, alternative = "less",
       mu = 0, paired = TRUE, conf.level = 1 - alpha)
Paired t-test
data:  baseball$F3 and baseball$F2
t = -2.4234, df = 15, p-value = 0.01425
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
     -Inf -49.14946
sample estimates:
mean difference
     -177.6875
```

Paired *t*-test – Baseball Teams

```
t.test(x = baseball$diff, alternative = "less",
           mu = 0, conf.level = 1 - alpha)
One Sample t-test
data:  baseball$diff
t = -2.4234, df = 15, p-value = 0.01425
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
     -Inf -49.14946
sample estimates:
mean of x
-177.6875
```

Paired *t*-test – Exercise

- **Exercise (BUN in Cat)** - Elevated levels of blood urea nitrogen (BUN) denote poor kidney function. Five elderly cats are placed on a standard high-protein diet. Their BUN is measured both initially and three months after they are placed on a standard high-protein diet.

Cat	1	2	3	4	5	Sample Mean	Sample Variance
Initial BUN	52	41	49	62	39	48.6	85.30
Final BUN	58	41	58	75	44	55.2	183.70
Difference	6	0	9	13	5	6.6	23.30

- **Question**: Is there a significant increase in mean BUN? Use 0.05 as the significance level.

Inference on Two Proportions

- **Purpose** – Compare the probability of success from two binomial distributions.
- Let p_1 and p_2 be the probabilities of success, respectively.
- Let n_1 and n_2 be the sample sizes, respectively.
- Let y_1 and y_2 be the samples (number of observed success), respectively.
- Estimates of proportion of success:

$$\hat{p}_1 = {}^{Y_1}\!/_{n_1}, \hat{p}_2 = {}^{Y_2}\!/_{n_2}$$

Sampling Distribution of the Difference between Two Proportions

- Let $\hat{p}_1 = {}^{Y_1}/_{n_1}$, $\hat{p}_2 = {}^{Y_2}/_{n_2}$. Similarly, we have

$$E(\hat{p}_1) = p_1, \mathrm{Var}(\hat{p}_1) = {}^{p_1(1-p_1)}/_{n_1}$$

$$E(\hat{p}_2) = p_2, \mathrm{Var}(\hat{p}_2) = {}^{p_2(1-p_2)}/_{n_2}$$

- Then $L = \hat{p}_1 - \hat{p}_2$, we can get

$$E(L) = p_1 - p_2$$

$$Var(L) = {}^{p_1(1-p_1)}/_{n_1} + {}^{p_2(1-p_2)}/_{n_2}$$

- **Idea**: use L as the test statistic and use the normal approximation for the calculation.

Inference for Two Proportions

- Assumptions for the test of two proportions are:
  - Each random sample is selected from the target population.
  - Two samples are independent.
  - The sample proportions, $\hat{p}_1$ and $\hat{p}_2$ are approximate normal. That means, the normal approximation to the binomial is appropriate:

$$n_1\hat{p}_1 \geq 5; n_1(1-\hat{p}_1) \geq 5$$
$$n_2\hat{p}_2 \geq 5; n_2(1-\hat{p}_2) \geq 5$$

  - Note that two sample sizes may not be equal.

Inference for Two proportions

- The null hypothesis is: $H_0: p_1 - p_2 = \delta_0$
- δ_0 is the nominal difference two means. In our course, $\delta_0 = 0$
- The test statistic is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{Estiamted\ Standard\ Deviation\ of\ (\hat{p}_1 - \hat{p}_2)}$$

- How to estimate the standard deviation of $(\hat{p}_1 - \hat{p}_2)$?
- Z has a standard normal distribution when H_0 is true.

Pooled Estimate of Proportion

- n_1: sample size from the first sample
- n_2: sample size from the second sample
- $\hat{p}_1 = {}^{y_1}/_{n_1}$: sample proportion from the first sample
- $\hat{p}_2 = {}^{y_2}/_{n_2}$: sample proportion from the second sample
- The pooled estimate of proportion is given by:

$$\hat{p} = \frac{n_1\,\hat{p}_1 + n_2\,\hat{p}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}$$

Inference for Two Proportions

- The null hypothesis is: $H_0: p_1 - p_2 = 0$
- The test statistic is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \sim N(0,1) \text{ when } H_0 \text{ is true}$$

- Where $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$ is the pooled estimate of proportion.

Inference for Two Proportions – Rejection Region

- For two-sided test ($H_a: p_1 - p_2 \neq 0$), the rejection region is:

$$|z| > z_{\alpha/2}$$

- For upper-tailed (right-tailed) test ($H_a: p_1 - p_2 > 0$), the rejection region:

$$z > z_{\alpha}$$

- For lower-tailed (left-tailed) test ($H_a: p_1 - p_2 < 0$), the rejection region is:

$$z < -z_{\alpha}$$

Inference for Two Proportions – Confidence Interval

- For two-sided test ($H_a: p_1 - p_2 \neq 0$), construct a $100(1-\alpha)\%$ two-sided confidence interval $\left(\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{{\hat{p}_1(1-\hat{p}_1)}/{n_1} + {\hat{p}_2(1-\hat{p}_2)}/{n_2}}\right)$.
- For upper-tailed test ($H_a: p_1 - p_2 > 0$), construct a $100(1-\alpha)\%$ lower confidence interval $\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha}\sqrt{{\hat{p}_1(1-\hat{p}_1)}/{n_1} + {\hat{p}_2(1-\hat{p}_2)}/{n_2}}, 1\right)$.
- For lower-tailed test ($H_a: p_1 - p_2 < 0$), construct a $100(1-\alpha)\%$ upper confidence interval $\left(-1, \hat{p}_1 - \hat{p}_2 + z_{\alpha}\sqrt{{\hat{p}_1(1-\hat{p}_1)}/{n_1} + {\hat{p}_2(1-\hat{p}_2)}/{n_2}}\right)$.

Inference for Two Proportions – Confidence Interval

- Note that when we construct the confidence interval for $p_1 - p_2$, we can not assume that $p_1 = p_2$.
- The estimate of the variance of $\hat{p}_1 - \hat{p}_2$ is

$$\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2.$$

Inference for Two Proportions – *p*-value Approach

- For two-sided test ($H_a: p_1 - p_2 \neq 0$):

$$p\text{-value} = 2\Pr(Z > |z|)$$

- For upper-tailed (right-tailed) test ($H_a: p_1 - p_2 > 0$):

$$p\text{-value} = \Pr(Z > z)$$

- For lower-tailed (left-tailed) test ($H_a: p_1 - p_2 < 0$):

$$p\text{-value} = \Pr(Z < z)$$

Inference on Two Proportions – Example 5.8

- **Example 5.8** - A candidate for political office wants to determine whether he is more popular in women than in men. He conducts a sample survey of 250 men and 250 women, of which 105 men and 128 women favor his candidacy. Do these values indicate the candidate is more popular in women than in men?
- Use a significance level of 0.04.
- Basic Statistics:

$$n_1 = n_2 = 250; \hat{p}_1 = 0.42; \hat{p}_2 = 0.512$$

Inference on Two Proportions – Example 5.8

Example 5.8 – Key Results (Again, some steps are skipped)

- Pooled estimate of $\hat{p} = \frac{105+128}{250+250} = 0.4660$
- Test statistic: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}}$
- Test statistic: $z = \frac{0.42-0.512}{\sqrt{0.466*(1-0.466)/250+0.466*(1-0.466)/250}} = -2.0620$
- Rejection Region: $z < -z_{0.04} = -1.7507$
- The p-value is: $\Pr(Z < -2.0620) = 0.0196$
- The upper 96% CI is $(-1, -0.0142)$

Inference on Two Proportions – Example 5.8

```
prop.test(x = c(y1, y2), n = c(n1, n2),
          alternative = "less", conf.level = 1 - alpha, correct = FALSE)
2-sample test for equality of proportions without continuity correction
data:  c(y1, y2) out of c(n1, n2)
X-squared = 4.2517, df = 1, p-value = 0.01961
alternative hypothesis: less
96 percent confidence interval:
 -1.00000000 -0.01422098
sample estimates:
prop 1 prop 2
 0.420  0.512
```

Inference Two Proportions - Exercise

- **Exercise (Newly Hatched Larvae) –** A researcher studied newly hatched larvae of the common carp for exposure to copper or lead during the embryonic development. They were interested in if the percentage of defective larvae differed for the two metal solutions. They put 100 eggs copper and 80 eggs in lead. The number of defective larvae was 38 with copper and 18 with lead.
- **Question:** Did the percentage of defective larvae differ for the two metal solutions? Use the significance of 0.04 for the test.

Comparing Portions Using Paired Samples

- **Example 5.9** - In an experiment for evaluating a new headache remedy, 80 chronic headache sufferers are given a standard remedy and a new drug on different days, and the response is whether their headache was relieved. In the experiment 56, or 70%, were relieved by the standard remedy and 64, or 80%, by the new drug. Do the data suggest that the new drug is better to relieve the headache.

Table 5.9 Data on Headache Remedy

	STANDARD REMEDY		
	Headache	**No Headache**	**Totals**
New drug			
Headache	10	6	16
No Headache	14	50	64
Totals	24	56	80

Comparing Portions Using Paired Samples

- **Example 5.9** – The method here is to look at the pairs that are different and test if the proportion is 0.5.
- The two numbers are 6 and 14.
- 6: number of persons had headache after new drug but did not have headache after standard drug.
- 14: number of persons did not had headache after new drug but had headache after standard drug.
- Note these 20 samples are independent.
- Use 6 as the observed success to perform the test:

$$H_0: p = 0.5 \text{ versus } H_a: p < 0.5$$

Comparing Portions Using Paired Samples

- **Example 5.9** – Key Results (Again, some steps are skipped)
- $\hat{p} = \frac{6}{20} = 0.30$
- Test statistic: $z = \frac{\hat{p} - 0.50}{\sqrt{0.5 * 0.5/20}} = \frac{0.30 - 0.50}{\sqrt{0.5 * 0.5/20}} = -1.7889$
- Rejection region: $z < -z_{0.05} = 1.6449$
- The p-value is: $\Pr(Z < -1.7889) = -0.0368$
- The upper 95% CI is:
- $\left(0, \hat{p} + z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \left(0, 0.30 + 1.6449 * \sqrt{\frac{0.3 * 0.7}{20}}\right) = (0, 0.4685)$

Assumptions

- **Pooled t statistic:** (a) Two sets of samples are independent; (b) Distributions are normal or approximate normal (large sample size); (c) Variances are equal.
- **Paired t statistic:** (a) samples are independent; (b) Observations are paired; (c) Distributions are normal or approximate normal.
- **Inferences on binomial populations:** (a) Observations are independent (for McNemar's test pairs are independent); (b) Probability of success is constant for all observations; (c) Large sample sizes for normal approximation.

Assumptions

- When assumptions are not fulfilled - analysis is not appropriate and/or the significance levels (p-values) are not as advertised.
- Violation of distributional assumptions may be detected by the exploratory data analysis methods described in Chapter 1, which should be routinely applied to all data.
- When assumptions are not fulfilled or not clear-cut.
  - For the t statistics, minor violations are not particularly serious because these statistics are relatively robust.
  - It will be necessary to investigate other analysis strategies.

Chapter Summary

- **Inferences on means based on independent samples where the variances can be assumed equal** - use a single pooled estimate of the common variance.
- **Inferences on means based on independent samples where the variances cannot be assumed equal** - use the estimated variances for large samples. For small samples an approximation must be used.
- **Inferences on means based on dependent (paired) samples** - use differences between the pairs.

Chapter Summary

- **Inferences on proportions from independent samples** - use the normal approximation of the binomial to compute a statistic similar to that for inferences on means when variances are assumed known.
- **Inferences on proportions from dependent samples** - use a statistic based on information only on pairs whose responses differ between the two groups.
- **If assumptions are violated** – use alternative methods.